В.М. Вишневский, А.Н. Дудин, В.И. Клименок

СТОХАСТИЧЕСКИЕ СИСТЕМЫ С КОРРЕЛИРОВАННЫМИ ПОТОКАМИ. ТЕОРИЯ И ПРИМЕНЕНИЕ В ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЯХ

МОСКВА, 2018

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ

АКТЦМ	асимптотически квазитеплицева цепь Маркова
КТЦМ	квазитеплицева цепь Маркова
ПЛС	преобразование Лапласа – Стилтьеса
ПФ	производящая функция
СМО	система массового обслуживания
ТМО	теория массового обслуживания
ЦМ	цепь Маркова
$BMAP$	групповой марковский входной поток (Batch Markov Arrival Process)
A_n	квадратная матрица A порядка $n \times n$
$\det A$	определитель матрицы A
$(A)_{i,k} = a_{ik}$	элемент матрицы A, стоящий на пересечении i-й строки и k-го столбца
A_{ik}	алгебраическое дополнение элемента a_{ik} матрицы $A = (a_{ik})$
$\mathrm{Adj}A = (A_{ki})$	матрица алгебраических дополнений
$\mathrm{diag}\{a_1, \ldots, a_M\}$	диагональная матрица порядка M с диагональными элементами $\{a_1, \ldots, a_M\}$
$\mathrm{diag}^+\{a_1, \ldots, a_M\}$	квадратная матрица порядка $M+1$ с наддиагональными элементами $\{a_1, \ldots, a_M\}$, все остальные элементы матрицы равны 0
$\mathrm{diag}^-\{a_1, \ldots, a_M\}$	квадратная матрица порядка $M+1$ с поддиагональными элементами $\{a_1, \ldots, a_M\}$, все остальные элементы матрицы равны 0
$\delta_{i,j}$	символ Кронекера, равный 1, если $i = j$, и 0 в противном случае
$\mathbf{e}\,(\mathbf{0})$	вектор-столбец (вектор-строка), состоящий из единиц (нулей). При необходимости порядок вектора определяется нижним индексом
$\hat{\mathbf{e}}$	вектор-строка $(1, 0, \ldots, 0)$
I	тождественная матрица. При необходимости порядок матрицы определяется нижним индексом
O	нулевая матрица. При необходимости порядок матрицы определяется нижним индексом
$\tilde{I}$	матрица $\mathrm{diag}\{0, 1, \ldots, 1\}$

^{T}	символ транспонирования матрицы
$\otimes$	символ кронекерова произведения матриц
$\oplus$	символ кронекеровой суммы матриц
$\bar{W} = W + 1$	
MAP	марковский входной поток (Markovian Arrival Process)
$MMAP$	маркированный марковский входной поток (Marked Markovian Arrival Process)
$MMPP$	марковский модулированный пуассоновский входной поток (Markov Modulated Poisson Process)
PH	распределение фазового типа (Phase Type Distribution)
QBD	векторный процесс гибели и размножения (Quasi-Birth-and-Death Process)
SM	полумарковский процесс (Semi-Markovian Process)
$f^{(j)}(x)$	производная j-го порядка функции $f(x)$, $j \geq 0$
$f'(x)$	производная функции $f(x)$
$\square$	символ конца доказательства

Полужирные строчные латинские и греческие буквы обозначают вектор-строки (кроме специально оговоренных случаев, например $\mathbf{e}$).

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Математические модели сетей и систем массового обслуживания (СМО) широко применяются для исследования и оптимизации различных технических, физических, экономических, производственных, административных, медицинских, военных и других систем. Объектом исследования в теории СМО являются ситуации, когда имеется некоторый ресурс и множество запросов на удовлетворение потребности в этом ресурсе. Ограниченность ресурса и стохастический характер потока запросов приводят к отказу или задержке в удовлетворении запросов. Стремление уменьшить вероятность этих отказов и длительность задержек и явилось побуждающим мотивом развития теории СМО. Особенно важное значение теория СМО имеет в телекоммуникационной отрасли для решения задач оптимального распределения телекоммуникационных ресурсов между многочисленными абонентами, генерирующими запросы в случайные моменты времени.

Отправной точкой развития теории СМО послужили работы датского математика и инженера А. К. Эрланга, опубликованные в 1909–1922 гг. Указанные работы явились следствием его усилий по решению задач проектирования телефонных сетей. В этих работах в качестве модели потока запросов, поступающих в узел сети, рассматривалась модель пуассоновского потока. Этот поток определяется как стационарный ординарный поток без последействия или как рекуррентный поток с экспоненциальным распределением длин интервалов между моментами поступления. В дополнение к этому А. К. Эрланг предположил, что и время обслуживания запроса имеет показательное распределение. При выполнении этих предположений в силу отсутствия “памяти” у показательного распределения число запросов в рассмотренных А. К. Эрлангом системах описывалось одномерным марковским процессом, что позволило легко найти его стационарное распределение. Предположение о показательном виде распределения длин интервалов между моментами поступления выглядит довольно сильным и искусственным. Тем не менее теоретические результаты А. К. Эрланга хорошо согласовывались с результатами практических измерений в реальных телефонных сетях. Позднее этот факт оказался объясненным, благодаря работам А. Я. Хинчина, Г. Г. Ососкова и Б. И. Григелиониса, которые доказали, что при требовании равномерной малости интенсивностей потоков суперпозиция большого числа произвольных рекуррентных

потоков сходится к стационарному пуассоновскому потоку. Этот результат является своеобразным аналогом центральной предельной теоремы теории вероятностей, утверждающей, что при условии равномерной малости случайных величин их нормированная сумма в пределе (при числе слагаемых, стремящемся к бесконечности) по распределению сходится к случайной величине, имеющей стандартное нормальное распределение. Потоки запросов, поступающих в узел телефонной станции, являются суперпозицией большого числа потоков малой интенсивности, поступающих от индивидуальных абонентов сети. В силу этого модель стационарного пуассоновского потока достаточно хорошо описывала реальные потоки в телефонных сетях.

Появление компьютерных сетей передачи данных, в которых использовался эффективный метод коммутации пакетов, в отличие от метода коммутации каналов, применявшегося в телефонных сетях, потребовало разработки нового математического аппарата для их оптимального проектирования. Так же, как в работах А. К. Эрланга, первоначальные исследования в области компьютерных сетей (Л. Клейнрока [1], М. Шварца [2], Г. П. Башарина [3], В. М. Вишневского [4, 5] и др.) базировались на упрощенных предположениях о характере информационных потоков (пуассоновские потоки) и экспоненциальных распределениях времени трансляции пакетов. Это позволило использовать хорошо разработанный к моменту появления компьютерных сетей аппарат теории очередей для оценки производительности, синтеза топологической структуры, управления маршрутизацией и сетью в целом, выбора оптимальных параметров сетевых протоколов и т.д.

Дальнейший импульс развития теоретических исследований связан с появлением цифровых сетей интегрального обслуживания – ЦСИО (Integrated Service Digital Networks – ISDN), являющихся усовершенствованием ранних сетей пакетной коммутации. Характерной особенностью этих сетей является то, что единые аппаратно-программные средства используются для совместной передачи разнообразной мультимедийной информации — речи, интерактивных данных, больших массивов информации, факсимиле, видео и т.д. В силу большой разнородности потоки в таких сетях являются существенно нестационарными (так называемый “bursty traffic” — взрывной трафик) и коррелированными. Количества запросов, поступающих за непересекающиеся интервалы времени, могут быть зависимыми, причем эта зависимость может сохраняться даже для далеко расположенных друг от друга интервалов. Учет этих факторов особенно

важен при моделировании современных широкополосных сетей 4G [6,62] и интенсивно разрабатываемых перспективных сетей нового поколения 5G, внедрение которых ожидается к 2020 г. (см., например, [7–10]).

Описание сложного характера потоков в современных телекоммуникационных сетях может быть осуществлено использованием так называемых самоподобных (selfsimilar) потоков (см., например, [11,12]). Главным недостатком модели самоподобных потоков с точки зрения использования при аналитическом моделировании процессов передачи информации в телекоммуникационных сетях является следующий. Модель входного потока – это лишь один из структурных элементов СМО. Поэтому, имея в виду перспективу аналитического исследования СМО, необходимо стремиться иметь как можно более простую модель потока. Самоподобный поток сам по себе является сложным объектом, и даже наиболее удачные способы его задания, например через суперпозицию большого числа on/off источников с распределениями on и off периодов, имеющими "тяжелые хвосты", предполагают использование асимптотик. Поэтому аналитическое исследование не только самого потока, но и СМО, в которую он поступает, представляется мало реальным.

Наиболее простой моделью потока является стационарный пуассоновский поток. Простота стационарного пуассоновского потока состоит в том, что он задается одним параметром, который имеет смысл интенсивности потока (математического ожидания числа запросов, поступающих в единицу времени) или параметра показательного распределения интервалов между моментами поступления запросов. Но наличие только одного параметра предопределяет и очевидные недостатки модели стационарного пуассоновского потока. В частности, коэффициент вариации распределения длин интервалов между моментами поступления равен 1, а коэффициент корреляции длин соседних интервалов равен нулю. Поэтому если при моделировании некоторого реального объекта результаты статистической обработки данных о входном потоке показывают, что коэффициент вариации существенно отличен от 1, а коэффициент корреляции существенно отличен от 0, заведомо ясно, что стационарный пуассоновский поток не будет приемлемой моделью реального потока. Таким образом, на повестке дня стояло введение в рассмотрение модели входного потока, которая позволяла бы описывать потоки с коэффициентом вариации, существенно отличным от 1, и коэффициентом корреляции, существенно отличным от 0. В то же время эта модель должна была позволять относительно просто

получать результаты исследования СМО с коррелированным потоком как прозрачные аналоги результатов исследования соответствующих СМО со стационарным пуассоновским потоком. Исследования по разработке такой модели велись, не зависимо друг от друга, научным коллективом Г.П. Башарина в Советском Союзе (соответствующие потоки были названы МС-потоками или потоками, управляемыми цепью Маркова – Markov Chain) и научным коллективом М. Ньютса (M. Neuts) в США. Название этих потоков в США эволюционировало от "разносторонних (versatile) потоков" [13] через N-потоки (потоки Ньютса) [14] до марковских входных потоков (MAP – Markovian Arrival Process) и их обобщений – групповых марковских входных потоков ($BMAP$ – Batch Markovian Arrival Process). Эволюция от стационарного пуассоновского потока к $BMAP$-потоку проходила постепенно. Сначала - через рассмотрение IPP (Interrupted Poisson Process)- потока, в котором интервалы поступления стационарного пуассоновского потока перемежались интервалами, длина которых имеет показательное распределение, когда запросы не поступали. Затем пришли к рассмотрению SPP (Switched Poisson Process)- потока, в котором интервалы поступления стационарного пуассоновского потока одной интенсивности альтернировали с интервалами поступления стационарного пуассоновского потока другой интенсивности. Наконец, допустив, что имеется не две, а конечное число интенсивностей, пришли к модели $MMPP$ (Markov Modulated Poisson Process)- потока, от которой перешли к модели $BMAP$-потока.

Важность учета корреляции во входном потоке обусловлена тем, что как результаты измерений на реальных сетях, так и результаты расчета показывают, что наличие положительной корреляции существенно ухудшает характеристики СМО (увеличивает среднее время ожидания, вероятность отказа и т.д.). Например, при одной и той же средней скорости поступления запросов, одинаковом распределении времени обслуживания и одинаковой емкости буфера вероятность потери запроса может отличаться на несколько порядков [15]. Таким образом, если результаты исследования реального потока запросов свидетельствуют о том, что этот поток коррелированный, применение хорошо изученных моделей с рекуррентным потоком, а тем более со стационарным пуассоновским потоком, является неприемлемым. По этой причине развитая в работах М. Ньютса [13, 16], Д. Лукантони [17–19], В. Рамасвами [14, 20], С. Чакраварти [21–24], А. Н. Дудина [25–33], В. И. Клименок [34–40] и др. теория СМО с коррелирован-

ными входными потоками нашла широкое применение при исследовании телекоммуникационных сетей, в частности широкополосных беспроводных сетей, функционирующих под управлением протоколов IEEE802.11 и 16 (см., например, [41–45]), гибридных высокоскоростных систем связи на базе лазерной и радиотехнологий (см., например, [46–49]), сотовых сетей европейского стандарта UMTS и его последней версии LTE (Long Term Evolution) (см., например, [50–55]), беспроводных сетей с линейной топологией (см., например, [56–61]) и т.д.

Активные исследования в области теории очередей с коррелированными потоками, проводимые уже более четверти века, нашли отражение в многочисленных статьях, авторами которых являются A. S. Alfa, S. Asmussen, A. D. Banik, D. Baum, L. Breuer, H. Bruneel, S. R. Chakravarthy, M. L. Chaudhry, B. D. Choi, A. N. Dudin, D. Efrosinin, U. S. Gupta, Q. M. He, C. S. Kim, A. Krishnamoorty, V. I. Klimenok, G. Latoche, H. W. Lee, Q-L .Li , L. Lakatos, F. Machihara, V. A. Naumov, M. Neuts, S. Nishimura, V. Ramaswami, Y. Takahashi, T. Takine, M. Telek, D. Lucantoni, Y. Q. Zhao, V. M. Vishnevsky и др. Краткое описание некоторых вопросов теории очередей с коррелированными потоками приведены в отдельных разделах монографии по теории очередей и ее приложений [62–67].

Однако систематизированное изложение теории очередей с коррелированными потоками в мировой литературе в настоящее время отсутствует. Достаточная завершенность математических результатов, полученных в последние годы, и практические потребности разработчиков современных телекоммуникационных сетей обусловили целесообразность написания предлагаемой монографии, которая закрывает этот пробел.

В основу содержания книги положены оригинальные результаты авторов, ссылки на которые приведены в разделе "литература". Использованы также материалы курсов лекций, прочитанных студентам и аспирантам Белорусского государственного университета и Московского физико-технического института.

Книга включает введение, 9 глав и приложение.

В главе 1, имеющей вводный характер, дано краткое описание истории развития телекоммуникационных технологий - от первых телефонных сетей до современных широкополосных беспроводных сетей, сети Интернет и разрабатываемых перспективных сетей пятого поколения 5G, внедрение которых ожидается к 2020 году. Показана связь между становлением и развитием теории очередей, которая стала одним из основных математи-

ческих аппаратов оценки производительности и проектирования компьютерных сетей и прогрессом в области сетевых технологий.

В главе 2 изложены основы классической теории массового обслуживания. Приводятся основные понятия этой теории. Дано описание марковских и полумарковских случайных процессов, преобразований Лапласа – Стилтьеса и производящих функций, играющих важную роль при исследовании систем массового обслуживания. Рассмотрены классические методы оценки характеристик однолинейных, многолинейных и многофазных СМО с ограниченными или бесконечными буферами, различными дисциплинами поступления и функциями распределения времени обслуживания заявок.

В третьей главе (раздел 3.1) описывается групповой марковский поток ($BMAP$ - Batch Markov Arrival Process), подробно обсуждаются его свойства и дается связь с некоторыми более простыми видами потоков, известными в научной литературе. Кратко определяется маркированный марковский поток, являющийся расширением понятия $BMAP$-потока на случай разнородных заявок, и полумарковский поток (SM - Semi-Markovian), в котором интервалы между моментами поступления запросов могут быть зависимыми и распределенными по произвольному закону. Описываются распределения фазового (PH) типа.

В разделе 3.2 описываются многомерные (или векторные) процессы гибели и размножения (Quasi-Birth-and-Death Processes), являющиеся простейшим и наиболее хорошо изученным классом многомерных цепей Маркова (ЦМ). Приводится критерий эргодичности цепи и формулы для вычисления векторов стационарных вероятностей состояний цепи в так называемом матрично-геометрическом виде. В качестве примера применения векторных процессов гибели и размножения анализируется СМО типа $MAP/PH/1$, т.е., однолинейная СМО с буфером бесконечного объема, марковским входным потоком и распределением времени обслуживания, имеющим фазовый тип. В дальнейшем предполагается, что читатель знаком с символикой Дж. Кендалла для кодирования СМО, и расшифровка типа СМО опускается. Для полноты изложения для упомянутой системы получено также (в терминах преобразования Лапласа – Стилтьеса) распределение времени ожидания запроса в системе. Кратко описан спектральный подход к анализу векторных процессов гибели и размножения.

В разделе 3.3 приводятся результаты для многомерных цепей Маркова типа $G/M/1$. В качестве примера применения многомерных цепей

Маркова типа $G/M/1$ анализируется вложенная по моментам поступления запросов цепь для СМО типа $G/PH/1$. Для полноты изложения для упомянутой системы получено также распределение вероятностей состояний системы в произвольный момент времени и распределение времени ожидания произвольного запроса в системе.

В разделе 3.4 приводятся результаты для многомерных цепей Маркова типа $M/G/1$ (или квазитеплицевых верхне-хессенберговых цепей Маркова). Доказано необходимое и достаточное условие эргодичности цепи. Описаны два метода для нахождения стационарного распределения вероятностей состояний цепи: метод, использующий векторные производящие функции и соображения аналитичности их в единичном круге комплексной плоскости, и метод М. Ньютса и его модификация, полученная на основе теории сенсорных цепей Маркова. Обсуждаются сильные и слабые стороны обоих методов.

В разделе 3.5 кратко описывается методика нахождения стационарного распределения вероятностей состояний многомерных цепей Маркова типа $M/G/1$ с конечным пространством состояний.

В разделе 3.6 приводятся результаты для так называемых асимптотически квазитеплицевых ЦМ (АКТЦМ) с дискретным временем, а в разделе 3.7 — результаты для асимптотически квазитеплицевых ЦМ с непрерывным временем. Доказываются достаточные условия эргодичности и неэргодичности АКТЦМ. Выводится алгоритм для вычисления стационарных вероятностей АКТЦМ на основе аппарата теории сенсорных ЦМ.

В качестве примера применения результатов, приведенных для многомерных цепей Маркова типа $M/G/1$, в главе 4 изучены системы $BMAP/G/1$ и $BMAP/SM/1$. Для системы $BMAP/G/1$ в разделе 4.1 приведен пример расчета стационарного распределения вероятностей состояний цепи Маркова, вложенной по моментам окончания обслуживания запросов. Кратко изложен вывод уравнений для стационарного распределения вероятностей состояний системы в произвольный момент времени на основе использования результатов теории процессов марковского восстановления. В терминах преобразования Лапласа – Стилтьеса получено распределение виртуального и реального времени ожидания и пребывания запроса в системе. Для более общей системы $BMAP/SM/1$ с полумарковским обслуживанием в разделе 4.2 выведено в простом виде условие эргодичности цепи Маркова, вложенной по моментам окончания обслуживания запросов, и обсуждена проблема нахождения блочных матриц пере-

ходных вероятностей этой цепи. Установлена связь стационарного распределения вероятностей состояний системы в произвольный момент времени со стационарным распределением вероятностей состояний цепи Маркова, вложенной по моментам окончания обслуживания запросов. Раздел 4.3 посвящен анализу системы $BMAP/SM/1/N$ с различными дисциплинами принятия группы запросов в ситуации, когда размер группы превышает число свободных мест в буфере в момент поступления группы.

В разделе 4.4 исследована многолинейная система типа $BMAP/PH/N$, а в разделе 4.5 рассмотрена многолинейная система типа $BMAP/PH/N/0$ с потерей запросов, являющаяся обобщением модели Эрланга $M/M/N/0$, с исследования которой ведет свою историю теория очередей. Рассмотрены различные дисциплины принятия группы запросов на обслуживание в ситуации, когда размер группы превышает число свободных приборов в момент поступления группы. Численно проиллюстрировано, что известное свойство инвариантности стационарного распределения вероятностей состояний относительно распределения времени обслуживания (при фиксированном математическом ожидании времени обслуживания), присущее системе $M/G/N/0$, не выполняется, когда входной поток не стационарный пуассоновский, а более общий $BMAP$-поток.

В главе 5 (раздел 5.1) с использованием результатов, полученных для асимптотически квазитеплицевых цепей Маркова, рассматривается система $BMAP/SM/1$ с повторными вызовами. Рассмотрены классическая стратегия повторов (включая обобщенную – линейную стратегию) и стратегия с постоянной интенсивностью повторов. Приведены условия существования стационарного распределения вероятностей состояний системы в моменты окончания обслуживания запросов. Получено дифференциально-функциональное уравнение для векторной производящей функции этого распределения. Установлена связь стационарного распределения вероятностей состояний системы в произвольный момент времени со стационарным распределением вероятностей состояний цепи Маркова, вложенной по моментам окончания обслуживания запросов. В разделе 5.2 рассматривается система $BMAP/PH/N$ с повторными вызовами. Рассмотрены стратегия повторов с бесконечно возрастающей интенсивностью потока повторов с орбиты с ростом числа запросов на орбите и стратегия с постоянной интенсивностью повторов. Поведение системы описано многомерной цепью Маркова, компонентами которой являются число за-

просов на орбите, число занятых приборов, состояние управляющего процесса $BMAP$-потока и состояния PH процесса обслуживания в каждом из занятых приборов. Приведены условия существования стационарного распределения вероятностей состояний системы. Получены формулы для вычисления важнейших характеристик производительности системы. Рассмотрен также случай неабсолютно настойчивых запросов, в котором после каждой неудачной попытки попасть на обслуживание запрос с орбиты навсегда уходит из системы. В разделе 5.3 дано описание алгоритмов расчета характеристик системы $BMAP/PH/N$ с повторными вызовами. Приводятся результаты численных расчетов, иллюстрирующие работоспособность предложенных алгоритмов и зависимость основных характеристик производительности системы от ее параметров. Проиллюстрирована возможность иного описания динамики функционирования системы в терминах многомерной цепи Маркова, блоки инфинитезимального генератора которой имеют в случае большого числа приборов существенно меньшую размерность, чем у исходной цепи.

В главе 6 приводится анализ характеристик гибридных систем связи на базе лазерной и радиотехнологий в терминах двухлинейных или трехлинейных СМО с ненадежными приборами. Исследуются основные архитектуры беспроводных систем этого класса, обеспечивающих высокоскоростной и надежный доступ к информационным ресурсам в существующих и перспективных сетях нового поколения. В разделах 6.1-6.3 описаны модели и методы анализа характеристик двухлинейных СМО с ненадежными обслуживающими приборами и входящими $BMAP$-потоками, адекватно описывающих функционирование гибридных систем связи, в которых лазерный канал резервируется широкополосным радиоканалом (холодный и горячий резерв). В разделе 6.4 приводятся методы и алгоритмы оценки характеристик производительности обобщенной архитектуры гибридной системы: атмосферный лазерный и радиоканал миллиметрового диапазона и радиоволн (71-76 Ггц, 81-86 Ггц), работающие параллельно, резервируются широкополосным каналом IEEE 802.11 сантиметрового диапазона радиоволн. Для всех моделей, СМО рассмотренных в данной главе, дано описание многомерной цепи Маркова, условий существования стационарного режима, алгоритмов вычисления стационарных распределений и основных характеристик производительности.

В главе 7 обсуждаются многофазные СМО с коррелированными потоками и их применение для оценки производительности широкополосных

беспроводных сетей с линейной топологией. В разделе 7.1 приведен краткий обзор работ по многофазным СМО с коррелированными входными потоками. Раздел 7.2 посвящен исследованию двухфазной СМО, первая фаза которой представляет собой систему $BMAP/G/1$, а вторая фаза является многолинейной системой без буфера, с показательным распределением времени обслуживания. В разделе 7.3 результаты предыдушего раздела обобщаются на случай, когда времена обслуживания запросов на первой фазе являются зависимыми случайными величинами, и описываются полумарковским (SM) процессом осблуживания.

В разделах 7.4 и 7.5 рассмотрены методы анализа стационарных характеристик фвухфазных систем $BMAP/G/1 \to \overline{M}/\overline{N}/\overline{R}$ и $BMAP/G/1 \to /M/N/O$ с повторными вызовами и групповым занятием приборов второй фазы.

В разделе 7.6 рассмотрена тандемная система, состоящая из произвольного конечного числа фаз, представленных многолинейными СМО без буферов, в которую поступает MAP-поток запросов. Времена обслуживания на фазах тандема имеют PH распределение, позволяющее адекватно описывать реальные процессы обслуживания. Доказана теорема о том, что выходящий поток каждой фазы тандема принадлежит классу MAP-потоков. Описан метод точного вычисления маргинальных стационарных вероятностей фрагментов тандема, а также всего тандема, и соответствующих вероятностей потерь запросов. С использованием подхода Рамасвами – Луконтини предложен алгоритм расчета основных характеристик многофазной СМО при больших значениях числа обслуживающих приборов на фазах и пространства состояний управляющего процесса PH обслуживания.

В разделе 7.7 рассмотрена более общая, чем в предыдущем разделе, многофазная СМО, на каждую фазу которой поступает коррелированный кросс-трафик, дополнительно с трафиком из предыдущей фазы тандема.

Раздел 7.8 посвящен исследованию тандемной системы с произвольным конечным числом фаз, представленных однолинейными СМО с конечными буферами, адекватно описывающей функционирование широкополосной беспроводной сети с линейной топологией.

Глава 8 посвящена исследованию новых моделей стохастического поллинга, адекватно описывающих функционирование широкополосных беспроводных сетей с централизованным механизмом управления.

В разделе 8.1 приведен краткий обзор публикаций по системам стоха-

стического поллинга и их применения для оценки характеристик производительности сетей сантиметрового и миллиметрового диапазонов радиоволн - IEEE 802.11ac и IEEE 802.11ad.

В разделах 8.2 и 8.3 приводится описание новых моделей и методов исследования систем адаптивного динамического поллинга, являющихся развитием результатов, опубликованных в предыдущих публикациях авторов книги.

Глава 9 посвящена анализу математической модели соты сети мобильной связи с процедурой обеспечения непрерывности соединения (handover).

В изложении математического материала в книге существенно используются некоторые сведения из теории матриц и теории функций от матриц, не входящие в программу стандартных курсов по матричному анализу. Для удобства ссылок они приводятся в приложении А.

Авторы выражают благодарность сотрудникам Белорусского государственного университета и Института проблем управления Российской академии наук, принимавшим участие в подготовке книги. Особая признательность выражается рецензентам книги профессору С. Чакраварти(США) профессору Л. Лакатошу(Венгрия),а также Российскому научному фонду за финансовую поддержку в рамках гранта №16-49-02021.

ГЛАВА 1
ИСТОРИЧЕСКИЙ ОЧЕРК РАЗВИТИЯ СЕТЕВЫХ ТЕХНОЛОГИЙ

В настоящей главе дано краткое описание истории развития телекоммуникационных сетей - от первых телефонных сетей до сети Интернет и перспективных широкополосных сетей пятого поколения 5G. Показана связь между становлением и развитием теории очередей и прогрессом в области сетевых технологий.

Исторически телекоммуникационные технологии зародились еще в позапрошлом веке. Родоначальником всех электронных сетей передачи данных считают американского художника Самуэля Финли Бриза Морзе. В 1837г. он разработал свою систему электросвязи по металлическому проводу и дал ей название "телеграф". Годом позже он дополнил ее знаменитой азбукой Морзе, т.е. механизмом кодирования источника, обязательным элементом всех современных сетей. 24 мая 1844 г. между Балтимором и Вашингтоном состоялся первый публичный сеанс телеграфной связи.

В 1874 г. французский инженер Жан Морис Эмиль Бодо (Baudot) изобрел телеграфный мультиплексор, позволяющий по одному проводу передавать до шести телеграфных каналов. Значимость этого изобретения и авторитет Бодо были столь высоки, что когда в 1877 г. другой французский инженер, Томас Муррэй, разработал первый в истории символьный телеграфный код с фиксированным размером символа (5 бит на символ), он назвал его кодом Бодо. Известный также под названием телексный код, он с незначительными изменениями применяется и сегодня (наиболее распространенная версия – стандартизированный Международный алфавит №2). В честь Бодо названа и единица измерения скорости передачи телекоммуникационных символов (бод).

Следующий шаг сделали изобретатели телефона – профессор физиологии органов речи Бостонского университета Александр Грэйхем Белл при участии Томаса Ватсона (1875 г., приоритет от 14 февраля 1876 г.) и независимо от них – Элайша Грей в Чикаго. Последнему также принадлежит немалая роль в развитии сетевых технологий. Именно он в 1888 г. запантетовал Telautograph – первое устройство передачи факсимильных сообщений. Но это были лишь предпосылки сетей, а именно способы фор-

мирования канала связи и работы в нем. Сеть – это совокупность многих каналов, которыми необходимо управлять (коммутировать). В первых сетях, начиная с 1880 г., этим занимались телефонистки методом установки штекеров в коммутационном поле.

С 1889 г. начался новый этап в развитии сетевых технологий – владелец бюро похоронных услуг из Канзас-Сити Элмон Браун Строуджер разработал систему автоматической коммутации каналов. Именно ему принадлежит приоритет в создании шагового искателя и декадно-шаговых АТС. Предание гласит, что Строуджер столкнулся с промышленной диверсией – жена его конкурента по цеху в Канзас-Сити работала телефонисткой и все звонки направляла своему мужу. Видимо, это был один из первых в мире случаев электронного шпионажа. Это стимулировало Строуджера к разработке АТС, что позволяло освободиться от услуг телефонисток. Изобретение Строуджера оказалось столь удачным, что в 1891 г. он основал компанию Strowger Automatic Exchange (с 1901 г. – Automatic Electric, сегодня – отделение компании General Telephone and Electronics, GTE). Первая АТС этой компании емкостью 99 номеров была запущена в коммерческую эксплуатацию в 1892 г. (Ла-Порт, шт. Индиана). Примечательно, что на первых телефонных аппаратах для работы с АТС номер набирался посредством кнопок. В 1897 г. компания Строуджера представила прототип первого телефонного аппарата с дисковым номеронабирателем.

В 1885 г. произошло еще одно ключевое для сетевых технологий событие. Первые АТС обеспечивали одновременное соединение всех возможных пар абонентов. Очевидно, что при росте номерной емкости коммутационные матрицы становились невероятно дорогими и сложными. Впервые возникла проблема доступа к ограниченному коммутационному ресурсу. Ее разрешил российский инженер М. Ф. Фрейденберг, показавший, что для 10 тыс. абонентов достаточно обеспечить возможность одновременного соединения любых 500 пар. Отметим, что результат Фрейденберга справедлив и сегодня для современных АТС: на 10 тыс. номеров допустимая вероятность предоставления соединения составляет 0,125. В 1895 г. М. Ф. Фрейденберг совместно с другим русским инженером С. М. Бердичевским-Апостоловым разработал и запатентовал в Великобритании АТС с предыскателем, выбиравшим свободный комплект линейных искателей при снятии абонентом трубки. Предыскатель и его принцип свободного поиска стал основой для проектирования всех будущих АТС. Примерно с 1910 г. (к окончанию срока действия патента Строуджера) началось массовое внедрение электроме-

ханических АТС. Работу, начатую М. Ф. Фрейденбергом, до логического завершения довел датский математик А. К. Эрланг, опубликовавший в 1909 г. ставшую классической работу "Теория вероятностей и телефонные переговоры"("The Theory of Probabilities and Telephone Conversations"), в которой предложил формулы для вычисления числа абонентов АТС, желающих одновременно вести разговоры.

Работы А. К. Эрланга положили начало новому научному направлению – теории очередей (теории массового обслуживания), широко используемой первоначально для расчетов в телефонии, а затем при проектировании сетей передачи информации. Значительный вклад в развитие теории очередей внес выдающийся российский математик Александр Яковлевич Хинчин (Математическая теория стационарной очереди: Математический сборник, 1932, т. 39, № 4. О формулах Эрланга в теории массовго обслуживания. Теория вероятностей и ее применения, 1962, т. 7, вып. 3.), выполнивший ряд оригинальных исследований для Московской телефонной сети.

В 1909 г. генерал-майор корпуса связи США, доктор философии Джордж Оуэн Скваер изобрел способ посылки по телефонной линии нескольких радиограмм одновременно – родился метод частотного разделения каналов.

В 1928 г. американский физик-электрик и изобретатель Гарри Найквист в статье "Некоторые вопросы теории телеграфной передачи"("Certain Topics in Telegraph Transmission Theory") изложил принципы преобразования аналоговых сигналов в цифровые и сформулировал знаменитую теорему Найквиста. В СССР ее называли теоремой В.А. Котельникова, хотя Владимир Александрович опубликовал аналогичные результаты через пять лет после Найквиста. Но история все нивелирует – основополагающая теорема Клода Элвуда Шеннона о пропускной способности канала (1948 г.) была сформулирована В. А. Котельниковым в его докторской диссертации годом раньше, в 1947 г. Однако у нас ее называют теоремой Шеннона.

В 1938 г. американец А. Х. Риверс патентует метод преобразования сигнала из аналоговой формы в цифровую для коммутации и передачи, названный импульсно-кодовой модуляцией (ИКМ). Этот метод впервые был практически реализован учеными из Bell Laboratories Клодом Шенноном, Джоном Р. Пирсом и Бернардом М. Оливером в быстродействующей цифровой передающей системе, позволившей транслировать несколько телефонных разговоров по одному каналу с высоким качеством – появилась

система с временным разделением (уплотнение) каналов.

Начиная с 1950-х годов, сетевые и беспроводные технологии начали сближаться настолько тесно, что зачастую грань между ними провести уже трудно. Беспроводные технологии также зарождались в XIX веке. Вплотную к идее беспроводной связи подошли такие ученые, как Г. Герц, О. Лодж, Э. Бранли. В 1892 г. английский ученый Вильям Крукс теоритически показал возможность и описал принципы радиосвязи. В 1893 г. сербский ученый Никола Тесла в США продемонстрировал передачу сигналов на расстоянии. Тогда это событие не вызвало должного резонанса, возможно, потому, что Н. Тесла, работы которого существенно опережали время, интересовался беспроводной передачей на расстояние не информации, а энергии.

С 1878 г. над проблемой беспроводной связи работал преподаватель минных классов в Кронштадте Александр Степанович Попов. В 1884 г. он изобрел первую приемную антенну, создал прибор для регистрации грозовых разрядов на основе когерера – стеклянной трубки, заполненной металлическими опилками. Под воздействием электромагнитного поля проводимость этой трубки резко возрастала. 7 мая 1895 г. на заседании физического отделения Российского физико-химического общества состоялся его исторический доклад "Об отношении металлических порошков к электрическим колебанием". Тогда А. С. Попов продемонстрировал свой прибор для регистрации грозовых разрядов ("грозоотметчик") и высказал мысль о возможности его применения для беспроводной связи. Первая публичная демонстрация прототипа всех беспроводных систем состоялась 24 марта 1896 г. на заседании того же физико-химического общества. А. С. Попов передал на рсстояние 250 м, возможно, первую в мире радиограмму, состоявшую из двух слов "Генрих Герц".

С 1894 г. успешно экспериментировал с физическими приборами для генерации и регистрации электромагнитных колебаний двадцатилетний итальянский юноша Гульельмо Маркони, будущий нобелевский лауреат. В 1895 г. он установил связь на расстоянии порядка двух миль, в 1896 г. запатетовал свое изобретение, в 1901-м установил радиосвязь через Атлантиду.

В 1906 г. Ли де Форест создал первую электронную лампу (триод) – появилась возможность строить электронные усилители сигналов. С тех пор беспроводная связь развивалась и продолжает по сей день – семимильными шагами, главным образом, благодаря достижениям электроники. От-

метим лишь основные вехи.

С 1920-х годов началось коммерческое радиовещание (посредством амплитудной модуляции). В 1933 г. Эдвин Ховард Армстронг изобрел частотную модуляцию (ЧМ), с 1936 г. началось коммерческое ЧМ-радиовещание. В 1946 г. компании AT& T и Bell System приступили к эксплуатации систем подвижной телефонной связи (MTS) для абонентов с автомобильными радиотелефонами. Для полудуплексной связи использовалось шесть каналов шириной по 60 кГц на частоте 150 МГц, однако из-за межканальной интерференции число каналов вскоре сократили до трех. Система позволяла соединяться с городской телефонной сетью.

12 августа 1960 г. был выведен на орбиту высотой 1500 км первый спутник связи – американский космический аппарат (КА) "Эхо-1"(Echo-1). Это был надувной шар с металлизированной оболочкой диаметром 30 м, выполнявший функции пассивного ретранслятора. Через два года, 10 июля и 13 декабря 1962 г., в США на низкие орбиты были запущены, соответственно, КА Telstar I и Relay-1 – первые спутники с активными ретрансляторами. Мощность их передатчиков не превышала 2 Вт. 19 августа 1964 г. впервые спутник связи был выведен на геостанционарную орбиту. Это был также американский Syncom-3 (первые две попытки вывода в 1963 г. были неудачными). На следующий день был создан международный консорциум спутниковой связи Intelsat (International Telecommunications Satellite Organization), который стал крупнейшей международной организацией в области спутниковой связи. Сегодня ее услугами пользуются более чем в 200 странах, причем в начале 2001 г. 2/3 всего международного трафика передавалось через спутники Intelsat. 23 апреля 1965 г. был выведен на орбиту и начал успешно работать первый отечественный спутник связи "Молния-1"(также с третьей попытки). Мир вступил в эру спутниковой связи.

В истории сетевых технологий очередной этап начался в 1960-е годы и связан с массовым появлением компьютеров. Возникла потребность в передаче большого объема данных, зародилось понятие локальной вычислительной сети (ЛВС). Был разработан механизм коммутации сообщений (пакетов). В 1960-е годы над построением сети с коммутацией пакетов работали (параллельно, практически ничего не зная друг о друге) специалисты в трех организациях: в Массачусетском технологическом институте (MTI), корпорации RAND (центр стратегических исследований ВВС США, с 1948 г. – независимая компания) и Национальной британ-

ской физической лаборатории (NPL). Пионерской работой в этой области явилась диссертация Леонарда Клейнрока на соискание степени доктора философии в MIT "Информационный поток в больших коммуникационных сетях"(Information Flow in Large Communication Nets 1961). В 1964 г. была опубликована работа сотрудника корпорации RAND Пола Барана "О распределенных коммуникациях"("On Distributed Communications"). В ней были сформулированы принципы избыточной коммуникативности и показаны различные модели формирования коммуникационной системы, способной успешно функционировать при наличии значительных повреждений. В 1964 г. Лоуренс Робертс из MIT совместно с Томасом Меррилом связал компьютер TX-2 в Массачусетсе с ЭВМ Q-32 в Калифорнии по низкоскоростной коммутируемой телефонной линии. Так была создана первая нелокальная компьютерная сеть. Она убедительно продемонстрировала, что сеть с коммутацией соединений (каналов) неприемлема для таких задач.

В 1962 г. в журнале "Коммунист"(№ 12) появилась статья академика АН СССР Александра Александровича Харкевича "Информация и техника". В ней впервые в мире были сформулированы основные принципы создания единой сети связи (ЕСС), указана важность передачи и коммутации различных видов информации в цифровой форме. ЕСС, по мнению А. А. Харкевича, должна представлять собой крупнейший инженерный комплекс, объединяющий все существующие сети связи и развивающийся путем планомерного его наращивания в органическом взаимодействии с системой вычислительных, управляющих и справочных центров.

Знаковыми для сетевых технологий стали 1967-1968 гг. В NPL заработала первая ЛВС с пакетной коммутацией, во многом благодаря ее директору Дональду Дэвису. Сеть работала с пиковой скоростью – до 768 кбит/с (в начале 1970-х гг. она объединяла порядка 200 компьютеров со скоросью обмена до 250 кбит/с). В том же 1968 г. сотрудник шведского отделения компании IBM Олаф Содерблюм разработал сеть Token Ring. МО США одобрило версию первого в мире стандарта на ЛВС – MIL-STD-1553 (протокол обмена данными по общему последовательному каналу посредством манчестерского линейного кода с выделенным контроллером (отечественный аналог – ГОСТ 26765.52-87). Этот стандарт после ряда модификаций до сих пор применяется в бортовых системах.

Но самое главное – в октябре 1967 г. был представлен начальный план сети ARPANET, развитием которой занимался департамент мето-

дов обработки информации IPTO (Information Processing Techniques Office) агентства перспективных исследовательских проектов ARPA (Advanced Research Projects Agency) МО США. В декабре 1968 г. группа во главе с Фрэнком Хартом из компании Bolt, Beranek и Newman (BBN) выиграла конкурс ARPA на создание так называемого интерфейсного процессора сообщений (Interface Message Processor). В 1969 г. в рамках программы ARPANET в Калифорнийском университете в Лос-Анджелесе "отец"пакетной коммутации Леонард Клейнрок построил первый узел ARPANET – прообраз грядущего интернета. В том же году компания BBN установила в Калифорнийском университете первый компьютер. Второй узел был образован в Стэнфордском Исследовательском институте (SRI). Двумя следующими узлами ARPANET стали Калифорнийский университет в Санта-Барбаре и Университет штата Юта. Эмбрион интернета начал делиться.

В 1970 г. появилась первая пакетная радиосеть передачи данных (через спутник) – знаменитая ALOHA (aloha – приветствие в гавайском диалекте английского языка). Ее разработал и построил Норман Абрамсон (совместно с Франком Куо и Ричардом Биндером) из Гавайского университета. Сеть связывала различные университетские учреждения, разбросанные по отдельным островам Гавайского архипелага. В 1972 г. ALOHA соединили с сетью ARPANET. В ALOHA был реализован принцип подтверждения и повторной посылки пакетов (ARQ), а также механизм множественного доступа к каналу с контролем несущей CSMA. Тогда же начали развиваться проекты создания пакетных радиосетей, в том числе спутниковых.

В октябре 1972 г. известный специалист из компании BBN Роберт Кан на международной конференции по компьютерным коммуникациям впервые публично продемонстрировал работу сети ARPANET. В 1974 г. появляется статья Вирта Серфа (сотрудника Стэнфордского исследовательского института) и Роберта Куна (Cerf V. G., Kanh R. E. A protocol for packet netwotk interconnection // IEEE Trans. Comm. Tech. Vol. COM-22. V. 5. May 1974. P. 627-641), в которой впервые была описана концепция протокола TCP/IP. В том же году компания BBN запустила первую открытую службу пакетной передачи данных (коммерческая версия ARPANET-Telnet).

В 1973 г. сотрудник исследовательского центра компании Xerox в Пал-Альто Роберт Метклаф, до прихода в Xerox защитивший в MIT докторскую диссертацию в области теории пакетной передачи информации и участвовавший в создании сети ARPANET, представил своему руко-

водству докладную записку, в которой впервые появилось слово Ethernet (эфирная сеть). В том же году Метклаф совместно с Дэвидом Боггсом построил первую Ethernet-ЛВС, связывавшую два компьютера со скоростью 2,944 Мбит/с. В основу технологии Ethernet был положен усовершенствованный принцип CSMA/CD с обнаружением коллизий. Через шесть лет, в 1979 г., при активном участии Р. Метклафа три ведущие в своих областях компании США – Xerox, Intel и Digital Equipment (DEC) – начали процесс стандартизации протокола Ethernet, успешно завершившийся через год. В том же 1979 г. Метклаф при участии DEC основал знаменитую компанию 3COM для выпуска Ethernet-совместимого оборудования.

В 1976 г. CCITT выпустила рекомендацию X.25, которая стала первым и чрезвычайно успешным стандартом сети с пакетной передачей данных по выделенному каналу (Interface between DTE and DCE for Terminal Operations in Packet Mode and Connected to Public Data Network by Dedicated Circuit). Массовая пакетная коммуникация стала реальностью.

В 1977 г. будущий вице-президент компании Sony Марио Токорои и другой японский ученый Киичироу Тамару предложили метод адаптации технологии Ethernet к передаче данных через радиоканал посредством механизма подтверждений (Acknowledging Ethernet). Эта работа заложила основу будущих беспроводных ЛВС (IEEE 802.11 и IEEE 802.15).

В 1977 г. Деннис Хайес основал компанию Hayes Microcomputer Products и выпустил на рынок первый массовый модем Micromodem II для персональных компьютеров (Apple II). Он работал со скоростью 110/300 бит/с и стоил 280 долл. В 1979 г. в Женеве CCITT утверждает первую модемную рекомендацию V.21, определяющую стандартный протокол модуляции на скорости 300 бит/с.

Новый этап начался в 1980 г., когда стек протоколов TCP/IP был принят в качестве военного стандарта США. Годом раньше пакетная радиосеть заработала на военной базе США Форт-Брэгг. В 1983 г. сеть ARPANET была переведена на протокол TCP/IP взамен действовавшего изначально NCP. Из ARPANET, которую вскоре стали называть интернетом, выделилась сеть MILNET, обслуживающая оперативные нужды МО США.

В дальнейшем сетевые технологии непрерывно развивались в сторону повышения быстродействия и надежности сетей передачи информации, возможности интегрированной передачи данных, голоса и видеоинформации. Так, в области локальных сетей было создано семейство технологий

Ethernet-Fast Ethernet-Gigabit Ethernet, обеспечивающих иерархию скоростей 10/100/1000 Мбит/с. В глобальных сетях произошел переход от технологии X.25 к технологии Frame Relay, использованию стека протоколов TCP/IP, ATM и Gigabit Ethernet.

Важно отметить, что и в СССР также работало немало выдающихся ученых и специалистов в области систем связи, в том числе и беспроводной. Уже в 1970-1980-х годах проектировались и строились современные сети связи, например система цифровой телефонной связи "Кавказ-5", многочисленные ведомственные сети связи. Хорошо известны системы "Сирена"(первая в СССР гражданская сеть пакетной коммутации) и "Экспресс"для автоматизации бронирования и продажи авиа- и железнодорожных билетов соответственно. Но, видимо, закрытость как самих работ, так и общества никак не согласовывалась с концепцией открытых сетей. Возможно, именно поэтому изначально созданная на деньги МО США открытая сеть Интернет завоевала весь мир, породила множество сетевых технологий, стимулировала развитие науки и смежных отраслей, прежде всего разработку соответствующей аппаратуры и элементной базы для нее.

Конец 1980-х годов ознаменовался появлением и последующим бурным ростом локальных беспроводных сетей (WLAN). Простота развертывания таких сетей ограничена только необходимостью оформления разрешительной документации (в тех странах, где это требуется). По пропускной способности они не уступают выделенным медным линиям. Помехоустойчивость, надежность и защищенность современных протоколов передачи данных сделали WLAN явлением повсеместным, а оборудование для них – массовым продуктом. Отметим, что понятие "локальные сети передачи информации"достаточно условно. Как правило, имеются в виду системы, локализованные в радиусе сотни метров. Однако технологии локальных сетей с успехом применяют и на расстояниях до нескольких десятков километров.

Первые устройства для беспроводных локальных сетей появились в начале-середине 1980-х годов. Но уже в 2000 году объем продаж оборудования WLAN достиг одного миллиарда долларов. Высокие темпы продаж этого оборудования сохраняются до настоящего времени.

Работы над единым стандартом локальных беспроводных сетей начались в 1989 году, когда была организована рабочая группа 11-го комитета IEEE 802. В июле 1997 года в результате работы этой груп-

пы был опубликован стандарт IEEE 802.11 "Спецификация физического уровня и уровня контроля доступа к каналу передачи беспроводных локальных сетей"("Wireless LAN Medium Access Control and Physical Layer Specifications"). Он определял архитектуру сети и вытекающие из этого требования к функциям устройств, принципы доступа устройств к каналам связи, формат пакетов, способы аутентификации и защита данных.

Стандарт IEEE 802.11 непрерывно совершенствуется и развивается в направлении предоставления пользователям новых сервисов, повышении скорости и качества передачи информации. В 2007 г. был выпущен обобщенный стандарт IEEE 802.11-2007, в который вошли все стандарты, завершенные к июню 2007 г., затем стандарт IEEE 802.11-2012 и, наконец, обобщенный стандарт IEEE 802.11-2016. В стандарт IEEE 802.11-2016 вошли все новые дополнения и стандарты, под управлением которых функционируют беспроводные сети сантиметрового диапазона радиоволн (до 6 ГГц), а также дополнение IEEE 802.11ad, регламентирующее работу локальной сети в миллиметровом диапазоне радиоволн 60 ГГц. Дальнейшее расширение этого дополнения – IEEE 802.11ay обеспечит скорость передачи информации до 100 Гбит/с. В 2019 г. ожидается появление дополнения IEEE 802.11ax, позволяющего в сантиметровом диапазоне радиоволн достигать скорости до 10 Гбит/с. Отметим, что в рамках проектируемых сетей 5G, высокоскоростные сети миллиметрового диапазона (60 ГГц-100ГГц), включая самоорганизующиеся MESH-сети, наряду с традиционными сетями сантиметрового диапазона станут одним из основных средств доступа к информационным ресурсам.

Одновременно с беспроводными сетями IEEE 802.11 интенсивно развивались мобильные сотовые технологии – одно из революционных достижений в области беспроводной связи, ставшее обыденным за последние двадцать лет. Роль этой технологии в 1990-х годах столь же велика, как бум персональных компьютеров в 1980-е годы. Мобильный телефон превратился в привычный предмет обихода, по стоимости приближающийся к обычному телефонному аппарату, а по распространенности значительно превзошедший число телефонных аппаратов фиксированной связи. В исторически короткий период происходит стремительная смена поколений сотовой связи: от сетей первого поколения до сегодняшних сетей четвертого поколения, а уже к 2020 году планируется внедрение сотовых сетей 5G.

Системы первого поколения (1G-first generation) были развернуты в

середине 1980-х годов (первые коммерческие сети – в конце 1970-х: 1978 год, Бахрейн, 1979 год – Япония). Системы 1G поддерживали только передачу голоса, обладали аналоговым радиотрактом и охватывали территорию отдельных стран, являясь несовместимыми друг с другом. Цифровые мобильные системы второго поколения 2G появились в конце 90-х годов прошлого столетия. Они обеспечивали не только качественную передачу речи, но и низкоскоростную передачу данных (до 14,4 Мбит/с). Обладая совместимостью, мобильные сети 2G охватили все страны мира. В наиболее известной системе 2G – глобальной системе мобильной связи GSM (Global System for mobile communication) – миллиардный абонент был зарегестрирован в 2004 году. На американском континенте и ряде азиатских стран широкое распространение получила другая система второго поколения – cdmaOne (IS-95), базирующаяся на технологии CDMA.

Важным этапом на пути развития мобильных сетей явилась разработка системы пакетной радиосвязи общего пользования GPRS (General Paket Radio Service). В отличие от GSM, где речь и данные передаются по коммутируемым каналам, GPRS обеспечивала пользователям возможность получать и отправлять данные с большой скоростью (до 50 кбит/с), используя технологию коммутации пакетов. GPRS относят к системам 2,5G, подчеркивая промежуточное положение между 2G и 3G.

Системы третьего поколения мобильной связи 3G были реализованы на базе новой радиотехнологии, обеспечивающей высокую скорость передачи мультимедийной информации и беспроводной доступ в интернет, не уступающий сервису провайдеров стационарной сети Интернет. В Европе для систем 3G используют термин UMTS – универсальная мобильная телекоммуникационная система (Universal Mobile Telecommunication System). Внедрялась и система третьего поколения мобильной связи cdma2000, представляющая собой дальнейшее развитие стандарта IS-95 cdmaOne. Таким образом, системы мобильной связи 3G развивались по двум направлениям – UMTS и cdma2000. В рамках UMTS обеспечивались преемственность GSM и GPRS; разрабатывались технологии повышения пропускной способности нисходящего (к абоненту) и восходящего (к базовой станции) направлений передачи информации: технологии HSDPA и HSUPA соответсвенно. В 2009 году в рамках UMTS была завершена разработка первых версий новой технологии Super 3G или Long Term Evolution (LTE), которую позиционировали как систему 3,9G.

В настоящее время во всем мире, включая Российскую Федерацию, уже

широко используются сети четвертого поколения на базе технологии LTE Advances (стандарт 3GPP, начиная с релиза 10), обеспечивающие в нисходящем радиоканале скорость передачи информации до 1000 Мбит/с, а в восходящем канале до 500 Мбит/с. Однако с появлением таких технологий, как интернет вещей (технологии LPWAN - SigFox и LoRa), виртуальная и дополнительная реальность (Virtual/augmented reality) и т.д., наблюдается экспоненциальный рост трафика, появляются его дополнительные источники, большинство из которых подключается к сети интернет посредством беспроводной связи. На трафик оказывает влияние также рост числа и качества передаваемых мультимедийных данных (видео высокого разрешения) и новые сервисы передачи таких данных от социальных сетей до различных сервисов передачи потокового видео, а также данных межмашинного взаимодействия (Machine-to-Machine,M2M).

В связи с указанными тенденциями и ограниченностью частотного спектра в сетях 4G уже в ближайшее время может возникнуть проблема нехватки мощностей для передачи огромных объемов данных с различными ограничениями на задержку, вероятность потерь, вариацию задержки и другие параметры. Разрабатывающиеся сети 5G ориентированы на решение этой проблемы за счет более эффективного использования существующего частотного спектра, привлечения дополнительного спектра и новых технологий радиодоступа в миллиметровом диапазоне радиоволн, технологий D2D (device-to-device), мультивещания и широковещания и т.д.

Быстрая смена поколений сетевых технологий, направленная на предоставление пользователям новых сервисов, повышение быстродействия и качества передачи и обработки информации, обеспечила проникновение компьютерных сетей во все сферы человеческой деятельности, включая экономику, науку, культуру, образование, промышленность и т.д. Однако стремительное развитие сетевых технологий и повсеместное внедрение компьютерных сетей было бы невозможным без использования и развития методов и алгоритмов математического моделирования сетевых технологий для сравнительного анализа, оценки производительности и выбора оптимальных параметров компьютерных сетей. Стохастический характер поступления пакетов и недетерминированная обработка их в узлах коммутации и при трансляции по каналам связи предопределили использование теории массового обслуживания в качестве одного из основных математических аппаратов проектирования телекоммуникационных сетей.

Теория массового обслуживания, зарождение которой связано с рабо-

тами датского ученого А. К. Эрланга и российского математика А. Я. Хинчина в области телефонных сетей, интенсивно развивается в последнее столетие. Начиная с конца 20-го века в теории очередей начались исследования нового направления – системы массового обслуживания с коррелированными потоками. Развитие этого направления стимулировалось практическими потребностями исследования современных телекоммуникационных сетей, в которых разнородные информационные потоки являются существенно нестационарными и коррелированными. Количество запросов, поступающих в базовые станции и узлы коммутации в непересекающиеся интервалы времени, могут быть зависимыми, причем указанная зависимость сохраняется даже для далеко расположенных друг от друга интервалов. Это обосновывает важность учета перечисленных факторов при моделировании современных широкополосных сетей 4G и, особенно, при проектировании перспективных сетей пятого поколения. Благодаря пионерским работам Л. Клейнрока в области сетей пакетной коммуникации и теоретическим исследованиям М. Ньюта, Г. П. Башарина и других ученых в области создания моделей теории очередей с коррелированными потоками, эта теория, описание которой приводится в следующих главах книги, стала мощным математическим фундаментом проектирования телекоммуникационных сетей.

ГЛАВА 2
МАТЕМАТИЧЕСКИЕ МЕТОДЫ ИССЛЕДОВАНИЯ КЛАССИЧЕСКИХ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ

2.1 Введение

Теория массового обслуживания (англоязычное название - queueing theory – теория очередей) возникла в начале 20-го века. Ее основоположником считается датский ученый А.К. Эрланг, работавший в шведской телефонной компании и занимавшийся вопросами проектирования телефонных сетей. В дальнейшем теория получила интенсивное развитие и применение в различных областях науки, техники, экономики, производства. Это объясняется тем, что эта теория изучает широко распространенные в человеческой практике ситуации, когда имеется некоторый ограниченный ресурс и множество (поток) запросов на его использование, следствием чего являются задержки или отказы в обслуживании некоторых запросов. Стремление понять объективные причины этих задержек или отказов и по возможности уменьшить их воздействие является побудительным мотивом развития теории массового обслуживания.

Как правило, поступление запросов (или их групп) происходит в случайные моменты времени и для их удовлетворения требуется случайная часть ограниченного ресурса (или случайное время его использования). Поэтому изучение процесса удовлетворения потребности в ресурсе (процесса обслуживания) обычно проводится в рамках теории случайных процессов как специальной области теории вероятностей. Иногда исследование процесса обслуживания требует применения достаточно тонких математических методов и серьезного математического аппарата. Это делает полученные результаты практически недоступными инженеру, потенциально заинтересованному в их применении к исследованию реального объекта. Что, в свою очередь, лишает автора математического результата обратной связи, важной для правильного выбора направления для дальнейшего обобщения результатов и объектов исследования. Эта серьезная проблема подмечена в обзоре [174] известного специалиста Р. Сиски, отмечающего опасность возможности распада единой теории массового обслу-

живания на абстрактную и инженерную. Прямым следствием этой проблемы при написании статьи или книги обычно является вопрос выбора языка и соответствующего уровня строгости изложения результатов. Данная книга ориентирована как на специалистов в области теории массового обслуживания, так и на специалистов в области ее приложения к исследованию реальных объектов (в первую очередь, телекоммуникационных сетей). Поэтому в данной главе приведем краткий обзор методов анализа систем массового обслуживания на среднем, компромиссном, уровне строгости. Предполагается знакомство читателя с теорией вероятностей в рамках курса для технического вуза. При необходимости некоторые сведения приводятся непосредственно в тексте.

Важным этапом в применении теории массового обслуживания для исследования реального объекта является формальное описание функционирования этого объекта в терминах той или иной системы массового обслуживания (СМО). СМО считается заданной, если полностью описаны следующие ее компоненты:

- входящий поток запросов (заявок, требований, сообщений, вызовов);
- количество и типы обслуживающих устройств (приборов);
- емкости накопителей (буферов), где запросы, заставшие все приборы занятыми, ожидают начала обслуживания;
- времена обслуживания запросов на приборах;
- дисциплина обслуживания (она определяет порядок обработки запроса в системе, начиная с момента его поступления в систему и до момента, когда он покидает СМО).

Согласно символике Дж. Кендалла, введенной в 1953 году, в теории массового обслуживания принято кодирование (краткое описание) основных СМО в виде совокупности четырех символов, разделенных вертикальными чертами: $A/B/n/m$. Символ $n, n \geq 1$ задает число идентичных параллельных обслуживающих устройств. Символ $m, m \geq 0$ задает число мест для ожидания в буфере. Если $m = \infty$, то четвертый символ в описании СМО может отсутствовать. Символ A описывает входящий поток запросов, а символ B – распределение времени обслуживания запросов. Некоторые возможные значения этих символов будут приведены и пояснены в следующем параграфе.

2.2 Входящий поток, время обслуживания

Входящий поток во многом определяет характеристики производительности функционирования СМО. Поэтому правильное описание потока запросов, поступающих в случайные моменты времени в реальную систему, и идентификация его параметров являются весьма важной задачей. Строгое решение этой задачи лежит в русле теории точечных случайных процессов и находится за пределами данной книги. Здесь мы приводим только краткие сведения из теории однородных случайных потоков, необходимые для понимания последующих результатов.

Во входящем (случайном) потоке запросы поступают в систему в некоторые случайные моменты времени $t_1, t_2, \ldots, t_n, \ldots$. Будем обозначать $\tau_k = t_k - t_{k-1}$ длину интервала между моментами поступления $(k-1)$-го и k -го запросов, $k \geq 1$ (t_0 полагается равным 0) и x_t - число моментов t_k, лежащих на временной оси левее точки $t, t \geq 0$.

Случайный поток считается заданным, если задано совместное распределение величин $\tau_k, k = 1, \ldots, n$ для любого $n, n \geq 1$ или задано совместное распределение величин x_t для всех значений $t, t \geq 0$.

Определение 2.1. Случайный поток называется стационарным, если для любого целого числа m и любых неотрицательных чисел $u_1, \ldots, u_m$ совместное распределение величин $(x_{t+u_k} - x_t),\ k = 1, \ldots, m$ не зависит от величины t.

На содержательном уровне это означает, что распределение числа запросов, поступивших на некотором интервале времени, зависит от длины этого интервала, но не зависит от расположения этого интервала на временной оси.

Определение 2.2. Случайный поток называется ординарным, если для любого t имеет место соотношение

$$\lim_{\Delta \to 0} \frac{P\{x_{t+\Delta} - x_t > 1\}}{\Delta} = 0.$$

На содержательном уровне это означает, что вероятность поступления более одной заявки за малый интервал времени есть величина более высокого порядка малости по сравнению с длиной интервала. Грубо говоря, это означает практическую невозможность одновременного поступления двух и более запросов.

Определение 2.3. Говорят, что случайный поток является потоком без последействия, если числа заявок, поступивших на непересекающихся интервалах времени, являются независимыми в совокупности случайными величинами.

Определение 2.4. Случайный поток называется потоком с ограниченным последействием, если величины $\tau_k, k \geq 1$ независимы в совокупности.

Определение 2.5. Случайный поток называется рекуррентным потоком, если поток является потоком с ограниченным последействием и величины $\tau_k, k \geq 1$ одинаково распределены.

Их функцию распределения будем обозначать $A(t) = P\{\tau_k < t\}$. Функция $A(t)$ полностью характеризует рекуррентный поток.

Если распределение $A(t)$ – показательное: $A(t) = 1 - e^{-\lambda t}$, то в обозначениях Кендалла первый символ принимает значение M.

Если распределение $A(t)$ вырожденное, т.е. запросы поступают через равные промежутки времени, то в обозначениях Кендалла первый символ принимает значение D.

Если распределение $A(t)$ – гиперэкспоненциальное, т.е.

$$A(t) = \sum_{k=1}^{n} q_k(1 - e^{-\lambda_k t}),$$

где $\lambda_k \geq 0$, $q_k \geq 0, k = 1, \ldots, n, \sum\limits_{l=1}^{n} q_l = 1$, то в обозначениях Кендалла первый символ принимает значение HM_n.

Если распределение $A(t)$ – эрланговское с параметрами (λ, k), т.е.

$$A(t) = \int_0^{\infty} \lambda \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} dt,$$

то первый символ принимает значение E_k. Параметр k называют порядком распределения Эрланга.

Более общим классом распределений, включающим гиперэкспоненциальное и эрланговское как частные случаи, является так называемое распределение фазового типа. В обозначениях Кендалла оно кодируется как PH. Подробную информацию о PH-распределении, его свойствах и содержательной интерпретации можно найти в [69].

Если о виде функции распределения $A(t)$ не делается никаких предположений, то в качестве первого символа в обозначениях Кендалла используются буквы G (General) или GI (General Independent). Строго говоря,

использование символа G не предполагает даже требования рекуррентности потока, а символ GI означает именно рекуррентный поток. Но в литературе иногда не делается различия между этими символами.

Определение 2.6. Интенсивностью λ стационарного случайного потока называется математическое ожидание (среднее значение) числа запросов, поступающих в потоке в единицу времени:

$$\lambda = M\{x_{t+1} - x_t\} = \frac{M\{x_{t+T} - x_t\}}{T}, T > 0.$$

Определение 2.7. Параметром α стационарного случайного потока называется положительная величина, определяемая соотношением:

$$\alpha = \lim_{\Delta \to 0} \frac{P\{x_{t+\Delta} - x_t \geq 1\}}{\Delta}.$$

Определение 2.8. Стационарный ординарный поток без последействия называется простейшим.

Справедливы следующие утверждения.

Утверждение 2.1. *Для того чтобы поток был простейшим, необходимо и достаточно, чтобы он был стационарным пуассоновским, т.е.*

$$P\{x_{t+u} - x_t = k\} = \frac{(\lambda u)^k}{k!} e^{-\lambda u}, k \geq 0.$$

Утверждение 2.2. *Для того чтобы поток был простейшим, необходимо и достаточно, чтобы он был рекуррентным с показательным распределением длин интервалов между моментами поступления запросов:* $A(t) = 1 - e^{-\lambda t}$.

Утверждение 2.3. *Если известно, что на интервале длины T поступило n запросов простейшего потока, то вероятность поступления фиксированного запроса на части этого интервала длиной τ не зависит от того, когда поступали другие запросы, и от расположения этой части внутри интервала и равна τ/T.*

Утверждение 2.4. *Для простейшего потока параметр потока и его интенсивность совпадают. Среднее число запросов, поступающих на интервале длины T, равно λT.*

Утверждение 2.5. *Поток, полученный в результате суперпозиции (наложения) двух независимых простейших потоков, имеющих интенсивности λ_1, λ_2, соответственно, является простейшим потоком интенсивности $\lambda_1 + \lambda_2$.*

Утверждение 2.6. *Поток, полученный из простейшего потока интенсивности λ в результате применения простейшей процедуры рекуррентного просеивания: произвольный запрос потока с вероятностью p включается в просеянный поток, а с вероятностью $1-p$ – игнорируется, является простейшим потоком интенсивности $p\lambda$.*

Утверждение 2.7. *Поток, полученный в результате суперпозиции n независимых рекуррентных потоков, имеющих равномерно малую интенсивность, при увеличении числа n сходится к простейшему потоку.*

Наиболее хорошо изученными системами массового обслуживания являются системы, в которых входящий поток является простейшим. Во многом это объясняется Утверждением 2.2 и известным свойством отсутствия последействия у показательного распределения.

Это свойство для показательной случайной величины ν в терминах условных вероятностей записывается следующим образом:

$$P\{\nu \geq t+\tau | \nu \geq t\} = P\{\nu \geq \tau\}.$$

Это равенство доказывается следующим образом. Из определения условной вероятности с учетом того, что $P\{\nu < x\} = 1 - e^{-\lambda x}$, имеем:

$$P\{\nu \geq t+\tau | \nu \geq t\} = \frac{P\{\nu \geq t+\tau, \nu \geq t\}}{P\{\nu \geq t\}} =$$

$$= \frac{P\{\nu \geq t+\tau\}}{P\{\nu \geq t\}} = \frac{e^{-\lambda(t+\tau)}}{e^{-\lambda t}} = e^{-\lambda\tau} = P\{\nu \geq \tau\}.$$

Из этого свойства вытекает, что распределение времени от произвольного момента до момента поступления следующего запроса из простейшего потока не зависит от того, когда поступил предыдущий запрос. Этот факт существенно упрощает анализ соответствующей СМО.

Утверждение 2.7 объясняет тот факт, что простейшие потоки часто имеют место в практических системах (так, поток запросов, поступающий в АТС, является суммой большого числа независимых малых потоков, поступающих от отдельных абонентов телефонной сети, и поэтому близок к простейшему) и поэтому использование простейшего потока для моделирования реального потока не только облегчает исследование СМО, но и оправдано.

Полезными сведениями о рекуррентных потоках являются следующие. Пусть зафиксирован произвольный момент времени и нас интересует функция $F_1(t)$ распределения интервала времени, прошедшего к данному

моменту с момента поступления последнего запроса рекуррентного потока, и функция $F_2(t)$ распределения интервала времени с данного момента по момент поступления следующего запроса. Можно показать, что

$$F_1(t) = F_2(t) = F(t) = a_1^{-1} \int_0^t (1 - A(u))du,$$

где величина a_1 есть средняя длина интервала между моментами поступления запросов: $a_1 = \int_0^\infty (1 - A(t))dt$. Величина a_1 и интенсивность потока λ связаны соотношением: $a_1\lambda = 1$.

Функция $F(t)$ совпадает с функцией $A(t)$ тогда и только тогда, когда поток является простейшим.

Определение 2.9. Мгновенной интенсивностью $\mu(t), \mu(t) > 0$ рекуррентного потока называется величина, определяемая как:

$$P\{x_{t+\Delta} - x_t \geq 1 | E_t\} = \mu(t)\Delta + o(\Delta),$$

где E_t - случайное событие, заключающееся в том, что в момент 0 поступил запрос и за время $(0, t]$ запросов не поступило, а $o(\Delta)$ означает величину более высокого порядка малости (при $\Delta \to 0$), чем Δ.

Мгновенная интенсивность $\mu(t)$ связана с функцией распределения $A(t)$ следующим образом:

$$A(t) = 1 - e^{-\int_0^t \mu(u)du}, \mu(t) = \frac{\frac{dA(t)}{dt}}{1 - A(t)}.$$

В случае простейшего потока мгновенная интенсивность потока совпадает с его интенсивностью.

В современных интегральных цифровых сетях связи (в отличие от традиционных телефонных сетей) потоки информации уже не представляют собой суперпозицию большого числа равномерно малых независимых рекуррентных потоков. В результате эти потоки часто являются не только не простейшими, но и не рекуррентными. Для описания таких потоков Д. Лукантони предложен [159] формализм групповых марковских потоков. Для обозначения их в символике Кендалла используется аббревиатура BMAP (Batch Markovian Arrival Process). Более подробную информацию о BMAP потоках и соответствующих системах обслуживания можно найти в [76], [159].

Отметим, что в телефонии популярен также так называемый поток от конечного источника (примитивный поток, пуассоновский поток второго рода), определяемый следующим образом.

Пусть имеется конечное число n-объектов, порождающих запросы независимо от других и называемых источниками запросов. Источник может находиться в занятом состоянии некоторое случайное время, в течение которого он не может генерировать запросы. После окончания пребывания в занятом состоянии источник переходит в свободное состояние. Находясь в этом состоянии, источник может сгенерировать новый запрос через показательно распределенное с параметром λ время. Сгенерировав запрос, источник немедленно переходит в занятное состояние.

Примитивным потоком от n-источников запросов называется поток с ограниченным последействием, интенсивность λ_i которого в данный момент времени прямо пропорциональна числу свободных в данный момент источников: $\lambda_i = \lambda(n - i)$, где i – число занятых в данный момент источников.

Относительно процесса обслуживания запросов в системе обычно предполагается, что обслуживание – рекуррентное, т.е. времена обслуживания последовательных запросов являются независимыми одинаково распределенными случайными величинами. Их функцию распределения будем обозначать $B(t)$. Обычно предполагается, что $B(+0) = 0$, т.е. исключается возможность мгновенного обслуживания запроса. Также обычно предполагается, что распределение $B(t)$ имеет нужное число конечных начальных моментов распределения.

Для задания типа распределения времени обслуживания в символике Кендалла используются практически те же символы, что и при задании типа потоков. Так, символ G означает либо отсутствие каких-либо предположений о процессе обслуживания, либо он отождествляется с символом GI, означающим рекуррентный процесс обслуживания, символ M означает предположение, что распределение времени обслуживания – показательное, т.е., $B(t) = 1 - e^{-\mu t}, \mu > 0, t > 0$, символ D означает детерминированные времена обслуживания и т.д.

Относительно недавно в обиход вошел символ SM (Semi-Markovian) см., например, [?], означающий более общий, чем рекуррентный, процесс обслуживания, при котором времена обслуживания последовательных запросов являются последовательными временами пребывания в своих состояниях некоторого полумарковского процесса с конечным пространством состоя-

ний и фиксированным ядром.

2.3 Марковские случайные процессы

Марковские случайные процессы играют важную роль при исследовании СМО. Поэтому приведем некоторые сведения из теории таких процессов, которые будут использоваться в дальнейшем изложении.

Определение 2.10. Случайный процесс $y_t,\ t \geq 0$, заданный на некотором вероятностном пространстве и принимающий значения в некотором числовом множестве Y, называется марковским, если для любого натурального числа n, любых $y, u, u_n, \ldots, u_1 \in Y$ и любых $\tau, t, t_n, \ldots, t_1$, упорядоченных следующим образом: $\tau > t > t_n > \cdots > t_1$, выполняется соотношение:

$$P\{y_\tau < y | y_t = u, y_{t_n} = u_n, \ldots, y_{t_1} = u_1\} = P\{y_\tau < y | y_t = u\}. \qquad (2.1)$$

Параметр t процесса будем рассматривать как время. Если τ – будущий момент времени, t – настоящий момент времени, $t_k, k = 1, \ldots, n$ – прошлые моменты времени, то условие (2.1) можно трактовать следующим образом: будущее поведение марковского процесса полностью определяется его состоянием в настоящий момент времени. У немарковского процесса будущее поведение зависит также от состояний процесса в прошлом.

В случае, если пространство состояний Y марковского процесса $y_t, t \geq 0$ является конечным или счетным, марковский процесс называется цепью Маркова. Если параметр t принимает значения только в дискретном множестве, то цепь Маркова называется цепью с дискретным временем. Если же параметр t принимает значения в некотором непрерывном множестве, то цепь Маркова называется цепью с непрерывным временем.

Важным частным случаем цепи Маркова с непрерывным временем является так называемый процесс гибели и размножения.

2.3.1 Процессы гибели и размножения

Определение 2.11. Случайный процесс $i_t,\ t \geq 0$, называется процессом гибели и размножения, если он удовлетворяет условиям:

- пространство состояний процесса есть множество неотрицательных целых чисел (или его некоторое подмножество);
- время пребывания процесса в состоянии i имеет показательное распределение с параметром $\gamma_i, \gamma_i > 0$ и не зависит от предыдущего поведения процесса;

- после завершения пребывания процесса в состоянии i он переходит в состояние $i-1$ с вероятностью $q_i, 0 < q_i < 1$ и в состояние $i+1$ с вероятностью $p_i = 1 - q_i$. Вероятность p_0 полагается равной 1.

Состояние процесса $i_t, t \geq 0$ в момент времени t можно трактовать как размер некоторой популяции в этот момент времени. Переход из состояния i в состояние $i+1$ трактуется как рождение нового члена популяции, а переход в состояние $i-1$ – как гибель члена популяции. Такая трактовка процесса и объясняет его название.

Обозначим $P_i(t)$ вероятность того, что в момент t процесс i_t находится в состоянии i.

Утверждение 2.8. *Вероятности* $P_i(t) = P\{i_t = i\}, i \geq 0$ *удовлетворяют следующей системе линейных дифференциальных уравнений:*

$$P_0^{'}(t) = -\lambda_0 P_0(t) + \mu_1 P_1(t), \tag{2.2}$$

$$P_i^{'}(t) = \lambda_{i-1} P_{i-1}(t) - (\lambda_i + \mu_i) P_i(t) + \mu_{i+1} P_{i+1}(t), i \geq 1, \tag{2.3}$$

где $\lambda_i = \gamma_i p_i, \mu_i = \gamma_i q_i, i \geq 0$.

Для доказательства применим так называемый Δt - метод. Этот метод широко используется при анализе цепей Маркова и марковских процессов с непрерывным временем.

Сущность этого метода состоит в следующем. Фиксируется некоторый момент времени t и некоторое малое приращение времени Δt. Распределение вероятностей состояний марковского процесса в момент времени $t+\Delta t$ выражается через его распределение вероятностей в момент t и вероятности возможных переходов процесса за время Δt. В результате получается система разностных уравнений для вероятностей $P_i(t)$. Деля обе части этих уравнений на Δt и устремляя величину Δt к нулю, получаем систему дифференциальных уравнений для искомых вероятностей.

Применяем этот метод для вывода уравнений (2.3). Обозначим через $R(i, j, t, \Delta t)$ вероятность перехода процесса i_t из состояния i в состояние j за интервал времени $(t, t + \Delta t)$. Поскольку распределение времени пребывания процесса i_t в состоянии i имеет показательное распределение, обладающее свойством отсутствия последействия, то время до окончания пребывания в состоянии i, начиная с момента времени t, также распределено по показательному закону с параметром γ_i. Вероятность того, что процесс i_t осуществит переход из состояния i за время $(t, t + \Delta t)$, есть вероятность

того, что за время Δt показательно распределенное с параметром γ_i время закончилось. По определению функции распределения эта вероятность равна

$$1 - e^{-\gamma_i \Delta t} = \gamma_i \Delta t + o(\Delta t). \tag{2.4}$$

Т.е. вероятность изменения состояния процесса за время Δt есть величина порядка Δt. Отсюда следует, что вероятность того, что за время Δt произойдет два или более перехода процесса i_t, есть величина порядка $o(\Delta t)$. Учитывая то, что процесс гибели и размножения осуществляет переходы за один шаг только в соседние состояния, заключаем, что $R(i, j, t, \Delta t) = o(\Delta t)$ для $|i - j| > 1$.

Используя приведенные рассуждения и формулу полной вероятности, мы получаем соотношения:

$$P_i(t + \Delta t) = P_{i-1}(t) R(i - 1, i, t, \Delta t) + P_i(t) R(i, i, t, \Delta t) + \tag{2.5}$$

$$P_{i+1}(t) R(i + 1, i, t, \Delta t) + o(\Delta t).$$

Из описания процесса и формулы (2.4) следует, что:

$$R(i - 1, i, t, \Delta t) = \gamma_{i-1} \Delta t p_{i-1} + o(\Delta t),$$

$$R(i, i, t, \Delta t) = (1 - \gamma_i \Delta t) + o(\Delta t), \tag{2.6}$$

$$R(i + 1, i, t, \Delta t) = \gamma_{i+1} \Delta t q_{i+1} + o(\Delta t).$$

Подставляя соотношения (2.6) в (2.5) и используя введенные обозначения λ_i, μ_i, переписываем (2.5) в виде:

$$P_i(t + \Delta t) - P_i(t) = P_{i-1}(t) \lambda_{i-1} \Delta t - P_i(t)(\lambda_i + \mu_i) \Delta t$$

$$+ P_{i+1}(t) \mu_{i+1} \Delta t + o(\Delta t). \tag{2.7}$$

Деля обе части этого уравнения на Δt и устремляя Δt к нулю, получаем уравнение (2.3). Уравнение (2.2) выводится аналогично.

Утверждение 2.8 доказано.

Для решения бесконечной системы дифференциальных уравнений (2.2), (2.3), путем перехода к преобразованиям Лапласа (см. ниже) $p_i(s)$ вероятностей $P_i(t)$ ее сводят к бесконечной системе линейных алгебраических уравнений. Однако и эта система может быть решена в явном виде только в некоторых случаях, например когда трехдиагональная матрица

этой системы имеет дополнительную специфику (например, $\lambda_i = \lambda,\ i \geq 0$, $\mu_i = \mu,\ i \geq 1$). Ситуация, когда полученная система уравнений для вероятностей $P_i(t)$ не может быть решена в явном виде, достаточно типична в теории массового обслуживания. Поэтому, несмотря на то, что значительный практический интерес иногда представляют именно эти вероятности, зависящие от t и характеризующие (при известном начальном состоянии процесса i_0 или известном распределении вероятностей начального состояния) динамику рассматриваемого процесса, обычно приходится довольствоваться так называемым стационарным распределением вероятностей процесса:

$$\pi_i = \lim_{t\to\infty} P_i(t), i \geq 0. \tag{2.8}$$

Положительные предельные (стационарные) вероятности π_i могут существовать не всегда, и условия их существования обычно устанавливаются с помощью так называемых эргодических теорем.

Для рассматриваемого нами процесса гибели и размножения можно доказать следующий результат.

Утверждение 2.9. *Стационарное распределение вероятностей (2.8) рассматриваемого процесса гибели и размножения существует, если сходится ряд*

$$\sum_{i=1}^{\infty} \rho_i < \infty, \tag{2.9}$$

где

$$\rho_i = \prod_{l=1}^{i} \frac{\lambda_{l-1}}{\mu_l}, i \geq 1, \rho_0 = 1,$$

и расходится ряд

$$\sum_{i=1}^{\infty} \prod_{l=1}^{i} \frac{\mu_l}{\lambda_l} = \infty. \tag{2.10}$$

При этом стационарные вероятности $\pi_i, i \geq 0$ вычисляются следующим образом:

$$\pi_i = \pi_0 \rho_i, i \geq 1, \pi_0 = \left(\sum_{i=0}^{\infty} \rho_i \right)^{-1}. \tag{2.11}$$

Последняя часть утверждения доказывается элементарно. Предполагаем, что условия (2.9) и (2.10) выполняются и пределы (2.8) существуют. Устремляем в (2.2), (2.3) t к бесконечности. При этом производные

$P_i'(t)$ стремятся к нулю. Существование пределов этих производных следует из существования пределов в правой части системы (2.2), (2.3). Равенство пределов производных нулю следует из того, что предположение о том, что пределы ненулевые, противоречит ограниченности вероятностей: $0 \leq P_i(t) \leq 1$.

В результате, из (2.2), (2.3) получаем систему линейных алгебраических уравнений для распределения $\pi_i, i \geq 0$:

$$-\lambda_0 \pi_0 = \mu_1 \pi_1, \tag{2.12}$$

$$\lambda_{i-1}\pi_{i-1} - (\lambda_i + \mu_i)\pi_i + \mu_{i+1}\pi_{i+1} = 0, i \geq 1. \tag{2.13}$$

Введя обозначение $x_i = \lambda_{i-1}\pi_{i-1} - \mu_i \pi_i, i \geq 1$, систему (2.12),(2.13) можно переписать в виде:

$$x_1 = 0, x_i - x_{i+1} = 0,$$

откуда следует, что $x_i = 0, i \geq 1$, что влечет выполнение соотношений

$$\lambda_{i-1}\pi_{i-1} = \mu_i \pi_i, i \geq 1. \tag{2.14}$$

Отсюда следует,что $\pi_i = \rho_i \pi_0, i \geq 1$. Формула для вероятности π_0 следует из условия нормировки.

2.3.2 Метод диаграмм интенсивностей переходов

Отметим, что существует эффективный метод получения уравнений типа (2.12), (2.13) (так называемых уравнений равновесия) для цепей Маркова с непрерывным временем (и процессов гибели и размножения в частности), без использования уравнений для нестационарных вероятностей.

Этот альтернативный метод называется методом диаграмм интенсивностей переходов. Сущность метода заключается в следующем. Поведение цепи Маркова с непрерывным временем описывается ориентированным графом. Узлы графа соответствуют возможным состояниям цепи. Дуги графа соответствуют возможным одношаговым переходам между состояниями цепи. Каждая дуга снабжается числом, равным интенсивности соответствующего перехода. Для получения системы линейных алгебраических уравнений для стационарных вероятностей используют так называемый принцин сохранения потока. Этот принцип заключается в следующем. Если сделать разрез графа, т.е. удалить некоторые дуги таким

образом, что получится несвязный граф, то поток из одной части разрезанного графа в другую равен потоку в обратном направлении. Под потоком понимается сумма (по всем удаляемым дугам) произведений стационарной вероятности узла, из которого идет удаляемая дуга, на интенсивность соответствующего перехода. Приравнивая таким образом потоки по всем разрезам, выбранным соответствующим образом, мы и получаем искомую систему уравнений.

Заметим, что мы получаем в точности уравнения равновесия, полученные посредством применения "Δt-метода" (называемые уравнениями глобального равновесия), если разрез в графе делается путем сечения всех дуг вокруг узла графа. Иногда за счет выбора более удачных сечений в графе удается получить существенно более простую для решения систему уравнений (называемых уравнениями локального баланса). В частности, если изобразить поведение рассматриваемого в данном параграфе процесса гибели и размножения в виде ленточного графа и сделать разрез не вокруг узла, соответствующего состоянию i (при этом мы получим уравнения (2.12),(2.13)), а между узлами, соответствующими состояниям i и $i+1$, то мы получим сразу более простые уравнения (2.14), из которых автоматически следуют формулы (2.11).

2.3.3 Цепи Маркова с дискретным временем

Приведем необходимые нам краткие сведения из теории таких цепей. Более подробная информация о цепях Маркова с дискретным временем может быть почерпнута, например, в [88].

Без ограничения общности будем считать, что пространством состояний цепи является множество неотрицальных целых чисел (или его некоторое подмножество).

Определение 2.12. Однородная цепь Маркова $i_k,\ k \geq 1$, с дискретным временем считается заданной, если:

- задано начальное распределение вероятностей состояний цепи:

$$r_i = P\{i_0 = i\}, i \geq 0;$$

- задана матрица P вероятностей одношаговых переходов цепи, состоящая из элементов $p_{i,j}$, определенных следующим образом:

$$p_{i,j} = P\{i_{k+1} = j | i_k = i\}, i, j \geq 0.$$

Матрица P одношаговых переходных вероятностей $p_{i,j}$ является стохастической, т.е. ее элементы являются неотрицательными числами и сумма элементов любой строки равна единице. Матрица вероятностей переходов цепи за m шагов является m-й степенью матрицы P.

Обозначим $P_i(k) = P\{i_k = i\}, k \geq 1, i \geq 0$ вероятность того, что после k-го шага цепь Маркова находится в состоянии i. Потенциально вероятности $P_i(k)$, характеризующие нестационарное поведение цепи, могут быть весьма интересны при решении задач нахождения характеристик объекта, описываемого данной цепью Маркова. Однако задача нахождения таких вероятностей является сложной. Поэтому обычно анализируют так называемые стационарные вероятности состояний цепи Маркова:

$$\pi_i = \lim_{k \to \infty} P_i(k), i \geq 0. \tag{2.15}$$

Иначе эти вероятности называют предельными, финальными, эргодическими. Мы будем касаться только неприводимых непериодических цепей, для которых положительные пределы в (2.15) существуют, при этом перечисленные альтернативные названия стационарных вероятностей выражают практически одни и те же свойства цепей: существование пределов (2.15), не зависящих от начального распределения вероятностей ее состояний, и существование единственного положительного решения следующей системы линейных алгебраических уравнений (уравнений равновесия) для стационарных вероятностей:

$$\pi_j = \sum_{i=0}^{\infty} \pi_i p_{i,j}, \tag{2.16}$$

$$\sum_{i=0}^{\infty} \pi_i = 1. \tag{2.17}$$

Существует целый ряд результатов (теоремы Феллера, Фостера, Мустафы, Твиди и т.д.), позволяющих по конкретному виду переходных вероятностей $p_{i,j}$ определить, существуют стационарные вероятности (пределы (2.15)) или нет. Если в результате исследования переходных вероятностей $p_{i,j}$ установлено, что при некоторых условиях на параметры цепи пределы (2.15) существуют, считают эти условия выполненными и решают систему уравнений равновесия (2.16),(2.17), после чего цепь Маркова считается исследованной.

В случае, когда пространство состояний цепи не является конечным, проблема решения уравнений (2.16), (2.17) является весьма сложной и эффективное ее решение возможно только в случае, если матрица одношаговых переходных вероятностей имеет какую-либо специфику.

2.4 Преобразования Лапласа и Лапласа – Стилтьеса. Производящая функция

В теории массового обслуживания интенсивно используется аппарат преобразований Лапласа и Лапласа – Стилтьеса и производящих функций. В частности, выше мы уже упомянули о возможности использования преобразований Лапласа для сведения задачи решения системы линейных дифференциальных уравнений к решению системы линейных алгебраических уравнений. Приведем основные сведения об этих функциях и преобразованиях.

Определение 2.13. Преобразованием Лапласа – Стилтьеса распределения $B(t)$ будем называть функцию $\beta(s)$, определяемую следующим образом:

$$\beta(s) = \int_0^\infty e^{-st} dB(t),$$

а преобразованием Лапласа – функцию $\phi(s)$, определяемую как:

$$\phi(s) = \int_0^\infty e^{-st} B(t) dt.$$

Если s есть чисто мнимая переменная, преобразование Лапласа – Стилтьеса совпадает с характеристической функцией распределения $B(t)$. Областью определения функций $\beta(s), \phi(s)$ обычно считается правая полуплоскость комплексной плоскости. Однако без существенного ограничения общности в рамках данной главы и книги можно рассматривать s как действительное положительное число.

Отметим некоторые из свойств преобразования Лапласа – Стилтьеса.

Свойство 2.1. Если оба преобразования $\beta(s)$ и $\phi(s)$ существуют (т.е. соответствующие несобственные интегралы сходятся), то они связаны между собой следующим образом: $\beta(s) = s\phi(s)$.

Свойство 2.2. Если две независимые случайные величины имеют преобразования Лапласа – Стилтьеса $\beta_1(s)$ и $\beta_2(s)$ их функций распределения,

то преобразованием Лапласа – Стилтьеса функции распределения суммы этих величин является произведение $\beta_1(s)\beta_2(s)$.

Свойство 2.3. Преобразованием Лапласа – Стилтьеса производной $B'(t)$ функции $B(t)$ является $s\beta(s) - sB(+0)$.

Свойство 2.4.

$$\lim_{s\to 0}\beta(s) = \lim_{t\to\infty} B(t).$$

Свойство 2.5. Пусть b_k есть k-й начальный момент распределения: $b_k = \int\limits_0^\infty t^k dB(t), k \geq 1$. Он вычисляется через преобразование Лапласа – Стилтьеса следующим образом:

$$b_k = (-1)^k \frac{d^k\beta(s)}{ds^k}|_{s=0}. \tag{2.18}$$

Свойство 2.6. Преобразованию Лапласа – Стилтьеса $\beta(s)$ может быть придан вероятностный смысл следующим образом. Считаем, что $B(t)$ есть функция распределения длины некоторого интервала времени и в этом интервале времени поступает простейший поток "катастроф" с параметром $s > 0$. Тогда легко видеть, что $\beta(s)$ есть вероятность того, что за интервал не наступит ни одна "катастрофа".

Свойство 2.7. Преобразование Лапласа – Стилтьеса $\beta(s)$, рассматриваемое как функция действительной переменной $s > 0$, является вполне монотонной функцией, т.е. оно имеет производные $\beta^{(n)}(s)$ всех порядков и $(-1)^n\beta^{(n)}(s) \geq 0, s > 0$.

Определение 2.14. Производящей функцией распределения вероятностей $q_k, k \geq 0$, дискретной случайной величины ξ называется функция

$$Q(z) = Mz^\xi = \sum_{k=0}^{\infty} q_k z^k, |z| < 1.$$

Перечислим основные свойства этой функции.

Свойство 2.8.

$$|Q(z)| \leq 1, Q(0) = q_0, Q(1) = 1.$$

Свойство 2.9. Для того чтобы случайная величина ξ имела m-й начальный момент $M\xi^m$, необходимо и достаточно, чтобы существовала конечная левосторонняя производная $Q^{(m)}(1)$ производящей функции $Q(z)$ в точке

$z = 1$, и начальные моменты легко подсчитываются через факториальные моменты

$$M\xi(\xi - 1)\dots(\xi - m + 1) = Q^{(m)}(1).$$

В частности, $M\xi = Q^{'}(1)$.

Свойство 2.10. В принципе производящая функция $Q(z)$ позволяет вычислить (произвести) вероятности q_i по следующей формуле:

$$q_i = \frac{1}{i!}\frac{d^i Q(z)}{dz^i}|_{z=0},\ i \geq 0. \tag{2.19}$$

Свойство 2.11. Производящей функции $Q(z)$ можно придать вероятностный смысл следующим образом. Интерпретируем случайную величину ξ как число запросов, пришедших за некоторый промежуток времени. Каждый приходящий запрос с вероятностью $z, 0 \leq z \leq 1$ окрашиваем в красный цвет, а с дополнительной вероятностью – в синий. Тогда из формулы полной вероятности следует, что $Q(z)$ есть вероятность того за этот промежуток времени пришли только запросы красного цвета.

Таким образом, зная производящую функцию $Q(z)$ распределения вероятностей $q_k, k \geq 0$, мы легко можем вычислить моменты этого распределения и в принципе можем вычислить сами вероятности $q_k, k \geq 0$. Если непосредственный подсчет по формуле (2.19) затруднителен, можно воспользоваться методом обращения производящей функции путем разложения ее на простые дроби или численными методами (см., например, [113], [166]). При решении практических задач можно пытаться аппроксимировать это распределение путем сглаживания по заданному числу совпадающих моментов распределения.

2.5 Однолинейные марковские системы массового обслуживания

Наиболее хорошо исследованными являются однолинейные системы массового обслуживания с простейшим входящим потоком или (и) показательным распределением времени обслуживания. Это объясняется тем, что процессы, которые в первую очередь интересуют исследователей, а именно: число i_t запросов в системе в момент t, время ожидания w_t запроса, который поступит (или может поступить) в момент времени t и др. – являются одномерными и, кроме того, либо являются марковскими, либо легко подвергаются марковизации за счет рассмотрения их только

во вложенные моменты времени или расширения фазового пространства процесса.

2.5.1 Система типа $M/M/1$

Рассмотрим систему $M/M/1$, т.е. однолинейную СМО с ожиданием (буфером неограниченной емкости), в которую поступает простейший поток запросов интенсивности λ, а время обслуживания запросов имеет показательное распределение с параметром μ.

Анализируя поведение этой системы, мы легко устанавливаем, что процесс i_t – число запросов в системе в момент t – является процессом гибели и размножения с параметрами:

$$\gamma_0 = \lambda, \gamma_i = \lambda + \mu, i \geq 1,$$

$$p_i = \int_0^\infty e^{-\mu t} \lambda e^{-\lambda t} dt = \frac{\lambda}{\lambda + \mu}, i \geq 1.$$

Поэтому для него справедливы Утверждения 2.8 и 2.9. При этом в формулировках этих утверждений мы должны задать параметры λ_i, μ_i как: $\lambda_i = \lambda, i \geq 0, \mu_i = \mu, i \geq 1$. Таким образом, величина ρ_i в формулировке Утверждения 2.9 определяется как $\rho_i = \rho^i$, где $\rho = \lambda/\mu$.

Параметр ρ, характеризующий соотношение интенсивности входящего потока и интенсивности обслуживания и называемый коэффициентом загрузки системы, имеет важную роль в теории очередей.

Проверяя условие существования стационарного распределения процесса $i_t,\ t \geq 0$, данное в формулировке Утверждения 2.9, мы легко убеждаемся, что стационарное распределение числа запросов в рассматриваемой системе существует, если выполняется условие:

$$\rho < 1. \tag{2.20}$$

Будем далее считать это условие выполненным.

Отметим, что для большинства однолинейных систем массового обслуживания условие существования стационарного распределения числа запросов в системе также имеет вид (2.20), что хорошо согласуется с интуитивными соображениями: для того, чтобы в системе не накапливалась бесконечная очередь, необходимо, чтобы в среднем запросы в системе обслуживались быстрее, чем они туда поступают.

Итак, мы можем сформулировать следующее следствие Утверждения 2.9.

Утверждение 2.10. *Стационарное распределение* $\pi_i,\ i \geq 0$, *числа запросов в системе* $M/M/1$ *определяется следующим образом:*

$$\pi_i = \rho^i(1-\rho),\ i \geq 0. \tag{2.21}$$

Отсюда следует, что вероятность π_0 того, что в произвольный момент времени система простаивает, равна $1-\rho$, а среднее число L запросов в системе определяется формулой

$$L = \sum_{i=0}^{\infty} i\pi_i = \frac{\rho}{1-\rho}. \tag{2.22}$$

Средняя длина L_o очереди определяется формулой:

$$L_o = \sum_{i=1}^{\infty}(i-1)\pi_i = L - \rho = \frac{\rho^2}{1-\rho}. \tag{2.23}$$

В ситуациях, когда распределение интервалов во входящем потоке и распределение времени обслуживания неизвестны, а известны только их средние значения, формулы (2.22) и (2.23) иногда используют для (грубой) оценки среднего числа запросов в системе и средней длины очереди в произвольный момент времени.

Как отмечалось выше, интересной характеристикой СМО является также распределение времени ожидания w_t (т.е. времени с момента поступления в систему до момента начала обслуживания) запроса, поступившего в момент t.

Обозначим $W(x)$ стационарное распределение процесса w_t :

$$W(x) = \lim_{t\to\infty} P\{w_t < x\}, x \geq 0.$$

Предполагаем, что запросы обслуживаются в порядке их поступления в систему. Иногда такая дисциплина выбора из очереди для краткости кодируется как FIFO (First In – First Out: первым пришел – первым обслужен) или, что означает то же самое, FCFS (First Came – First Served).

Утверждение 2.11. *Стационарное распределение* $W(x)$ *времени ожидания запроса в системе* $M/M/1$ *определяется следующим образом:*

$$W(x) = 1 - \rho e^{(\lambda-\mu)x}. \tag{2.24}$$

Доказательство. Время ожидания начала обслуживания произвольным запросом зависит от числа запросов, присутствующих в системе в момент его прихода. Для данной системы $M/M/1$ распределение числа запросов в системе в произвольный момент поступления запроса и в произвольный момент времени совпадают и задаются формулой (2.21). Запрос, заставший систему свободной (вероятность этого есть π_0), имеет нулевое время ожидания. Запрос, заставший в системе i запросов (вероятность этого есть π_i) ждет в течение времени, имеющего эрланговское распределение с параметрами (μ, i). Последний факт следует из того, что, во-первых, в силу свойства отсутствия последействия у показательного распределения оставшееся к моменту прихода время обслуживания обслуживаемого запроса имеет то же показательное распределение с параметром μ, что и полное время обслуживания, а во-вторых, в силу того, что сумма i независимых показательно распределенных с параметром μ случайных величин есть эрланговская случайная величина с параметрами (μ, i).

Из приведенных рассуждений и (2.21) следует, что:

$$W(x) = 1 - \rho + \sum_{i=1}^{\infty} \rho^i (1-\rho) \int_0^x \mu \frac{(\mu t)^{i-1}}{(i-1)!} e^{-\mu t} dt =$$

$$= 1 - \rho + (1-\rho)\lambda \int_0^x e^{(\lambda-\mu)t} dt.$$

Отсюда непосредственно следует (2.24). □

Среднее время ожидания W запроса в системе вычисляется следующим образом:

$$W = \int_0^{\infty} (1 - W(x)) dx = \lambda^{-1} \frac{\rho^2}{1-\rho}. \tag{2.25}$$

Среднее время V пребывания запроса в системе (т.е. времени с момента поступления в систему до момента окончания обслуживания на приборе) задается формулой:

$$V = W + \mu^{-1} = \lambda^{-1} \frac{\rho}{1-\rho}. \tag{2.26}$$

Сравнивая выражение (2.25) для среднего времени ожидания W и формулу (2.23) для средней длины L_o очереди в системе, а также формулу

(2.26) для среднего времени V пребывания запросов с формулой (2.22) для среднего числа L запросов в системе, видим, что:

$$L_o = \lambda W, L = \lambda V. \tag{2.27}$$

Отметим, что эти формулы справедливы и для многих более общих, чем рассматриваемая система $M/M/1$, систем массового обслуживания и называются формулами Литтла. Практическая значимость этих формул состоит в том, что они избавляют от необходимости непосредственного вычисления величин W, V при известном значении величин L_o, L и наоборот.

2.5.2 Система типа $M/M/1/n$

Рассмотрим теперь систему массового обслуживания $M/M/1/n$, т.е. однолинейную СМО с буфером ограниченной емкости. Запрос из входящего потока, заставший прибор занятым, ожидает начала обслуживания в буфере, если в нем имеется свободное место. Если же все n мест для ожидания заняты, запрос покидает систему необслуженным (теряется).

Обозначим $i_t, t \geq 0$ число запросов в системе в момент t. Этот процесс может принимать значения во множестве $\{0, 1, \ldots, n\}$. Нетрудно убедиться, что процесс $i_t, t \geq 0$ является процессом гибели и размножения и ненулевые параметры λ_i, μ_i определяются следующим образом: $\lambda_i = \lambda, 0 \leq i \leq n-1, \mu_i = \mu, 1 \leq i \leq n$. Тогда из формулы для стационарных вероятностей процесса гибели и размножения следует, что стационарные вероятности числа запросов в рассматриваемой системе имеют вид:

$$\pi_i = \rho^i \frac{1-\rho}{1-\rho^{n+1}}, 0 \leq i \leq n. \tag{2.28}$$

Одной из важнейших характеристик систем, в которых возможна потеря запросов, является вероятность P_{loss} того, что произвольный запрос будет потерян. Для рассматриваемой СМО можно показать, что вероятность потери произвольного запроса совпадает с вероятностью того, что в произвольный момент времени все места для ожидания заняты, т.е. справедлива формула:

$$P_{loss} = \rho^n \frac{1-\rho}{1-\rho^{n+1}}. \tag{2.29}$$

Формула (2.29) может использоваться для планирования необходимого размера буфера в зависимости от загрузки системы и значения допустимой вероятности потери запроса в системе.

Отметим, что в отличие от системы $M/M/1$, стационарное распределение числа запросов в данной системе существует при любых конечных значениях коэффициента загрузки ρ. При $\rho = 1$ вычисления по формулам (2.28), (2.29) можно выполнить, используя правило Лопиталя.

2.5.3 Система с конечным числом источников

В подразделе 2.2 введено понятие потока от конечного числа источников запросов. Рассмотрим кратко модель СМО, обслуживающую такой поток. Впервые эта модель была исследована Т. Энгсетом.

Имеется однолинейная СМО с буфером размера m, на вход которой поступают запросы от m идентичных источников. Любой источник находится в занятом состоянии (и, следовательно, не может генерировать запросы), пока его предыдущий запрос не обслужен прибором. Время обслуживания любого запроса от любого источника имеет показательное распределение с параметром μ. В свободном состоянии источник может сгенерировать следующий запрос через показательно распределенное с параметром λ время, после чего он переходит в занятое состояние.

Обозначим $i_t, t \geq 0$ число запросов в системе (на приборе и в накопителе) в момент t. Этот процесс может принимать значения во множестве $\{0, 1, \ldots, m\}$. Нетрудно убедиться, что этот процесс является процессом гибели и размножения и ненулевые параметры: интенсивности рождения λ_i и интенсивности гибели μ_i – определяются следующим образом:

$$\lambda_i = \lambda(m - i), 0 \leq i \leq m - 1, \mu_i = \mu, 1 \leq i \leq m.$$

Из формулы (2.11) для стационарных вероятностей процесса гибели и размножения очевидным образом получаем следующие выражения для стационарных вероятностей $\pi_i, i = 0, \ldots, m$ числа запросов в рассматриваемой системе:

$$\pi_i = \pi_0 \rho^i \frac{m!}{(m-i)!}, 1 \leq i \leq m, \tag{2.30}$$

где вероятность π_0 находится из условия нормировки:

$$\pi_0 = \left(\sum_{j=0}^{m} \rho^j \frac{m!}{(m-j)!} \right)^{-1}. \tag{2.31}$$

Используя формулы (2.30), (2.31), легко подсчитать среднее число запросов в системе и в очереди. Также можно подсчитать так называемый

стационарный коэффициент k_R готовности источника (вероятность того, что в произвольный момент времени источник готов сгенерировать запрос):

$$k_R = \sum_{i=0}^{m-1} \frac{m-i}{m} \pi_i = \frac{\mu(1-\pi_0)}{\lambda m}.$$

2.6 Полумарковские однолинейные системы массового обслуживания и методы их анализа

Как отмечалось в предыдущем параграфе, процесс $i_t, t \geq 0$, – число запросов в системе $M/M/1$ в момент t – является процессом гибели и размножения, т.е. частным случаем цепи Маркова с непрерывным временем. Аналогичный процесс для систем обслуживания типа $M/G/1$ с распределением времени обслуживания, отличным от показательного, и типа $GI/M/1$ с входящим потоком, отличным от простейшего, уже не является марковским.

Очевидной причиной этого является тот факт, что поведение процесса после некоторого фиксированного момента времени t не определяется, вообще говоря, полностью состоянием этого процесса в этот момент, а зависит также от того, сколь долго уже обслуживается находящийся на приборе в настоящий момент запрос или как давно поступил последний перед данным моментом запрос.

Тем не менее, исследование процесса $i_t, t \geq 0$, может быть сведено к исследованию марковских процессов. Первый способ "марковизации" – так называемый метод вложенных цепей Маркова – будет проиллюстрирован на примере системы $M/G/1$ в подразделе 2.6.1, и на примере системы $GI/M/1$ – в подразделе 2.6.2. Одна из разновидностей другого способа "марковизации" – метода введения дополнительной переменной – будет проиллюстрирована в подразделе 2.6.3. В подразделе 2.6.4 будет кратко описан и проиллюстрирован на примерах еще один мощный метод исследования СМО – метод введения дополнительного события.

2.6.1 Метод вложенных цепей Маркова в приложении для системы $M/G/1$

Рассмотрим однолинейную СМО с ожиданием, на вход которой поступает простейший поток интенсивности λ, а время обслуживания запроса

имеет произвольное распределение с функцией распределения $B(t)$, преобразованием Лапласа – Стилтьеса $\beta(s)$ и конечными начальными моментами $b_k,\ k = 1, 2$.

Как было отмечено, процесс $i_t,\ t \geq 0$, – число запросов в рассматриваемой системе в момент t – не является марковским, поскольку мы не можем описать поведение процесса после произвольного момента времени, не оглядываясь в прошлое. Вместе с тем, очевидно, что если мы знаем состояние $i, i > 0$, процесса i_t в момент t_k окончания обслуживания k -го запроса, то мы можем предсказать значение процесса i_t в момент окончания обслуживания $(k + 1)$ -го запроса, который произойдет через случайное время, u, имеющее распределение $B(t)$. За это время может поступить случайное (распределенное по закону Пуассона с параметром λu) число запросов и один запрос уйдет из системы.

Фактически мы пришли к идее метода вложенных цепей Маркова. В общем случае, этот метод заключается в следующем. Для немарковского процесса $i_t,\ t \geq 0$, ищется последовательность моментов времени $t_k, k \geq 1$, такая, что процесс $i_{t_k}, k \geq 1$, образует цепь Маркова. Методами теории цепей Маркова исследуют стационарное распределение вложенной цепи и затем по этому распределению восстанавливают стационарное распределение вероятностей исходного процесса. Это обычно делается с использованием теории процессов восстановления или процессов марковского восстановления.

Пусть t_k – момент окончания обслуживания в системе $M/G/1$ k-го запроса. Процесс $i_{t_k}, k \geq 1$, является однородной цепью Маркова с дискретным временем.

Выше отмечалось, что эффективное исследование цепи Маркова с дискретным временем и счетным пространством состояний возможно только в случае, если матрица ее одношаговых переходов имеет какую-либо специальную структуру. Матрица P одношаговых переходных вероятностей $p_{i,j}$ рассматриваемой вложенной цепи Маркова $i_{t_k}, k \geq 1$, такую структуру имеет. Найдем элементы матрицы P.

Пусть в некоторый момент окончания обслуживания запроса t_k число запросов i_{t_k} в системе равно $i, i > 0$. Поскольку в момент t_k число запросов в системе претерпевает скачок, для определенности будем считать, что $i_{t_k} = i_{t_k+0}$, т.е. уходящая в данный момент заявка уже не учитывается. Поскольку $i > 0$, то на обслуживание немедленно выбирается следующая заявка, которая покинет систему в следующий момент окончания обслу-

живания запроса t_{k+1}. Поэтому для того, чтобы в момент t_{k+1} в системе осталось j запросов, необходимо, чтобы за интервал времени (t_k, t_{k+1}) в систему поступило $j-i+1$ запросов. Вероятность этого события есть f_{j-i+1}, где величины f_l задаются формулой

$$f_l = \int_0^\infty \frac{(\lambda t)^l}{l!} e^{-\lambda t} dB(t), l \geq 0. \tag{2.35}$$

Итак, переходная вероятность $p_{i,j}, i > 0, j \geq i-1$, определяется формулой:

$$p_{i,j} = f_{j-i+1},\ i > 0,\ j \geq i-1. \tag{2.36}$$

Пусть теперь в момент t_k окончания обслуживания запроса число i_{t_k} запросов в системе равно 0. Очевидно, что система остается пустой до ближайшего момента поступления запроса. Начиная с этого момента, система ведет себя точно так же, как и после момента окончания обслуживания запроса, в который в системе остался один запрос. Поэтому $p_{0,j} = p_{1,j}$, откуда следует, что:

$$p_{0,j} = f_j,\ j \geq 0.$$

Таким образом, мы полностью описали ненулевые элементы матрицы одношаговых вероятностей переходов вложенной цепи. Эта матрица P имеет специальную структуру:

$$P = \begin{pmatrix} f_0 & f_1 & f_2 & f_3 & \dots \\ f_0 & f_1 & f_2 & f_3 & \dots \\ 0 & f_0 & f_1 & f_2 & \dots \\ 0 & 0 & f_0 & f_1 & \dots \\ 0 & 0 & 0 & f_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{2.37}$$

Наличие такой структуры существенно облегчает исследование данной цепи. Используя известные критерии эргодичности, несложно убедиться, что рассматриваемая вложенная цепь Маркова имеет стационарное распределение тогда и только тогда, когда

$$\rho < 1, \tag{2.38}$$

где коэффициент загрузки ρ равен λb_1.

Уравнения равновесия (2.16) с учетом вида (2.37) матрицы вероятностей одношаговых переходов можно переписать здесь в виде:

$$\pi_j = \pi_0 f_j + \sum_{i=1}^{j+1} \pi_i f_{j-i+1}, j \geq 0. \tag{2.39}$$

Для решения бесконечной системы линейных алгебраических уравнений (2.39) воспользуемся аппаратом производящих функций. Вводим в рассмотрение производящие функции

$$\Pi(z) = \sum_{j=0}^{\infty} \pi_j z^j, F(z) = \sum_{j=0}^{\infty} f_j z^j, |z| < 1.$$

Учитывая явный вид (2.35) вероятностей $f_l, l \geq 0$, можно получить явное выражение для производящей функции $F(z)$

$$F(z) = \int_0^{\infty} e^{-\lambda(1-z)t} dB(t) = \beta(\lambda(1-z)). \tag{2.40}$$

Умножая уравнения системы (2.39) на соответствующие степени z и суммируя их, получаем:

$$\Pi(z) = \pi_0 F(z) + \sum_{j=0}^{\infty} z^j \sum_{i=1}^{j+1} \pi_i f_{j-i+1}.$$

Меняя порядок суммирования, переписываем это соотношение в виде:

$$\Pi(z) = \pi_0 F(z) + \sum_{i=0}^{\infty} \pi_i z^{i-1} \sum_{j=i-1}^{\infty} f_{j-i+1} z^{j-i+1} =$$

$$= \pi_0 F(z) + (\Pi(z) - \pi_0) F(z) z^{-1} \tag{2.41}.$$

Отметим, что нам удалось свернуть двойную сумму в (2.41), благодаря специфике матрицы (2.37) вероятностей переходов, а именно благодаря тому, что переходные вероятности $p_{i,j}$ вложенной цепи Маркова для $i > 0$ зависят только от величины $j - i$ и не зависят от i и j отдельно. Это свойство матрицы называют квазитеплицевостью. Существенно использовано также то, что все элементы матрицы ниже ее поддиагонали равны нулю.

Учитывая (2.40), можно переписать формулу (2.41) в следующем виде:

$$\Pi(z) = \pi_0 \frac{(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z)) - z}. \tag{2.42}$$

Формула (2.42) определяет искомую производящую функцию стационарного распределения вероятностей вложенной цепи Маркова с точностью до значения неизвестной пока вероятности π_0 того, что рассматриваемая СМО пуста в произвольный момент окончания обслуживания запроса. Для нахождения этой вероятности вспомним, что система уравнений равновесия содержит еще уравнение (2.17) (условие нормировки). Из условия нормировки следует, что $\Pi(1) = 1$. Поэтому для нахождения вероятности π_0 мы должны подставить в (2.42) $z = 1$. Однако простая подстановка не дает результата, поскольку и числитель, и знаменатель (2.42) обращаются в нуль.

Для раскрытия неопределенности можно использовать правило Лопиталя. Однако при вычислении величины L среднего числа запросов в системе в моменты окончания обслуживания запросов потребуется вычислить величину $\Pi^{'}(1)$, для чего придется применять правило Лопиталя дважды, при вычислении дисперсии числа запросов в системе придется применять правило Лопиталя трижды и т.д. Во избежание многократного применения правила Лопиталя можно рекомендовать заранее разложить числитель и знаменатель дроби в правой части (2.42) в ряд Тэйлора по степеням $(z-1)$ (если мы заинтересованы в вычислении k-го начального момента распределения числа запросов, то разложение нужно провести с точностью до $o(z-1)^{k+1}$), сократить числитель и знаменатель на $(z-1)$ и только потом выполнять операции взятия производных и подстановки значения $z = 1$. Отметим, что если требуется вычислить моменты высокого порядка, при этом можно использовать возможности символьных вычислений, например с помощью пакета "Mathematica".

Используя приведенные соображения, мы получаем следующие выражения для вероятности π_0, среднего числа L запросов в системе и средней длины L_o очереди в системе в моменты окончания обслуживания запросов:

$$\pi_0 = 1 - \rho, \tag{2.43}$$

$$L = \Pi^{'}(1) = \rho + \frac{\lambda^2 b_2}{2(1-\rho)}, L_o = L - \rho. \tag{2.44}$$

Подставляя выражение (2.43) в формулу (2.42), получаем:

$$\Pi(z) = (1-\rho)\frac{(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z)) - z}. \tag{2.45}$$

Формула (2.45) называется формулой Полячека – Хинчина для производящей функции распределения числа запросов в системе $M/G/1$.

Отметим, что в выражения для величин L, L_o входит второй начальный момент b_2 распределения времени обслуживания. Поэтому при одинаковых средних временах обслуживания длины очередей в системах могут существенно отличаться. Так, при показательном распределении времени обслуживания (в этом случае $b_2 = \frac{2}{\mu^2}$) средняя длина L_o очереди равна $\frac{\rho^2}{1-\rho}$, а при детерминированном времени обслуживания ($b_2 = \frac{1}{\mu^2}$) средняя длина L_o очереди в два раза меньше. При эрланговском распределении времени обслуживания средняя длина L_o очереди принимает промежуточное значение. А при гиперэкспоненциальном обслуживании она может принимать существенно большие значения. Следовательно, оценивание вида распределения времени обслуживания в реальной модели имеет весьма важное значение. Учет только среднего значения может привести к значительной погрешности в оценке характеристик производительности системы.

Итак, проблема нахождения стационарного распределения вложенной цепи Маркова решена. Следует, однако, вспомнить, что нас интересует не эта цепь Маркова, а немарковский процесс $i_t,\ t \geq 0$, – число запросов в системе в произвольный момент времени. Введем в рассмотрение стационарное распределение этого процесса:

$$p_i = \lim_{t\to\infty} P\{i_t = i\}, i \geq 0.$$

Из теории процессов марковского восстановления (см., например, [112]) следует, что это распределение существует при тех же условиях, что и вложенное распределение (т.е. при выполнении условия (2.38)), и вычисляется через вложенное распределение следующим образом:

$$p_0 = \tau^{-1}\pi_0 \int_0^\infty e^{-\lambda t}dt, \tag{2.46}$$

$$p_i = \tau^{-1}\left[\pi_0 \int_0^\infty \int_0^t e^{-\lambda v}\lambda dv \frac{(\lambda(t-v))^{i-1}}{(i-1)!} e^{-\lambda(t-v)}(1-B(t-v))dt + \right.$$

$$\left. + \sum_{l=1}^{i} \pi_l \int_0^\infty \frac{(\lambda t)^{i-l}}{(i-l)!} e^{-\lambda t}(1-B(t))dt\right], i \geq 1. \tag{2.47}$$

Здесь τ есть средняя длина интервала между моментами ухода запросов из системы. Для нашей системы (без потери запросов) $\tau = \lambda^{-1}$.

Введем в рассмотрение производящую функцию $P(z) = \sum\limits_{i=0}^{\infty} p_i z^i$.

Умножая соотношения (2.46),(2.47) на соответствующие степени z и суммируя, получаем:

$$P(z) = \tau^{-1}\Bigg\{\pi_0\Bigg[\lambda^{-1} + \int\limits_0^{\infty}\int\limits_0^{t} e^{-\lambda v}\lambda dv \sum_{i=1}^{\infty}\frac{(\lambda(t-v))^{i-1}z^i}{(i-1)!}e^{-\lambda(t-v)}(1-B(t-v))dt\Bigg] +$$

$$+\sum_{i=1}^{\infty} z^i \sum_{l=1}^{i}\pi_l \int\limits_0^{\infty}\frac{(\lambda t)^{i-l}}{(i-l)!}e^{-\lambda t}(1-B(t))dt\Bigg\}.$$

Меняя порядок интегрирования в двойном интеграле, порядок суммирования в двойной сумме и подсчитывая известные суммы, получаем:

$$P(z) = \tau^{-1}\Bigg\{\pi_0\Bigg[\lambda^{-1} + z\lambda\int\limits_0^{\infty} e^{-\lambda v}dv\int\limits_v^{\infty} e^{-\lambda(1-z)(t-v)}(1-B(t-v))dt\Bigg] +$$

$$+\sum_{l=1}^{\infty}\pi_l z^l \int\limits_0^{\infty} e^{-\lambda(1-z)t}(1-B(t))dt\Bigg\}.$$

Делая замену переменной интегрирования $u = t - v$, учитывая равенство $\tau^{-1} = \lambda$ и связь между преобразованиями Лапласа и Лапласа – Стилтьеса, отсюда имеем:

$$P(z) = \pi_0[1 + \frac{z}{1-z}(1-\beta(\lambda(1-z)))] + (\Pi(z) - \pi_0)\frac{1}{1-z}(1-\beta(\lambda(1-z))).$$

Подставляя сюда выражение (2.42) для производящей функции $\Pi(z)$, после элементарных преобразований получаем:

$$P(z) = \pi_0\frac{(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z)) - z}.$$

В результате мы убедились в справедливости соотношения:

$$P(z) = \Pi(z). \tag{2.48}$$

Таким образом, для рассматриваемой системы $M/G/1$ распределения вероятностей числа запросов в системе в моменты окончания обслуживания запросов и произвольные моменты времени совпадают. А.Я. Хинчин [89] назвал это утверждение основным законом стационарной очереди.

Затронем проблему нахождения стационарного распределения времени ожидания и пребывания запросов в системе. Предполагаем, что запросы обслуживаются в порядке их поступления в систему (дисциплина выбора из очереди FIFO). Пусть w_t есть время ожидания, а v_t есть время пребывания в системе запроса, поступившего в нее в момент времени t. Обозначим:

$$W(x) = \lim_{t\to\infty} P\{w_t < x\}, V(x) = \lim_{t\to\infty} P\{v_t < x\} \tag{2.49}$$

и

$$w(s) = \int_0^\infty e^{-sx} dW(x), v(s) = \int_0^\infty e^{-sx} dV(x).$$

Условием существования пределов (2.49) является выполнение неравенства (2.38). Поскольку время пребывания запроса в системе равно сумме его времени ожидания и времени обслуживания, а время обслуживания запросов в классических моделях СМО предполагается независимым от состояния системы (и от времени, в течение которого запрос ожидал в очереди), то из свойства 2.2 преобразования Лапласа – Стилтьеса следует, что:

$$v(s) = w(s)\beta(s). \tag{2.50}$$

Популярным методом получения выражения для преобразований Лапласа – Стилтьеса $w(s), v(s)$ является вывод интегро-дифференциального уравнения Такача для распределения виртуального времени ожидания (т.е. времени, в течение которого ждал бы начала обслуживания запрос, если бы он поступил в систему в данный момент времени), см., например, [73].

Получим эти выражения другим, более простым способом. Нетрудно видеть, что при дисциплине FIFO число запросов, остающихся в системе в момент окончания обслуживания в ней некоторого запроса, совпадает с числом запросов, пришедших в систему за время пребывания в ней уходящего запроса. Отсюда следуют равенства:

$$\pi_i = \int_0^\infty \frac{(\lambda x)^i}{i!} e^{-\lambda x} dV(x), i \geq 0. \tag{2.51}$$

Умножая соотношения (2.51) на соответствующие степени z и суммируя, получаем:

$$\Pi(z) = \int_0^\infty e^{-\lambda(1-z)x} dV(x) = v(\lambda(1-z)).$$

Подставляя в это соотношение явный вид (2.45) производящей функции $\Pi(z)$, используя (2.50) и делая замену переменной: $s = \lambda(1-z)$, получаем следующую формулу:

$$w(s) = \frac{1-\rho}{1-\lambda\frac{1-\beta(s)}{s}}. \tag{2.52}$$

Формула (2.52) называется формулой Полячека – Хинчина для преобразования Лапласа – Стилтьеса распределения времени ожидания в системе $M/G/1$.

Используя формулу (2.18), из (2.52) получаем следующее выражение для величины среднего времени W ожидания запроса в системе:

$$W = \frac{\lambda b_2}{2(1-\rho)}.$$

Среднее время V пребывания запроса в системе находится как:

$$V = b_1 + \frac{\lambda b_2}{2(1-\rho)}. \tag{2.53}$$

Сравнивая формулы (2.44) и (2.53), снова получаем формулу Литтла:

$$L = \lambda V.$$

Если имеется настоятельная необходимость нахождения вида функции распределения времени ожидания $W(x)$, а не ее преобразования Лапласа – Стилтьеса, обращение этого проебразования, заданного формулой (2.52), проводится разложением правой части (2.52) на простые дроби (если это возможно) или численными методами (см., например, [113], [166]). Полезной может оказаться также так называемая формула Бенеша:

$$W(x) = (1-\rho)\sum_{i=0}^{\infty}\rho^i\tilde{B}_i(x), \tag{2.54}$$

где $\tilde{B}_i(x)$ есть свертка i-го порядка функции распределения

$$\tilde{B}(x) = b_1^{-1}\int_0^x(1-B(u))du,$$

а операция свертки определяется рекуррентным образом:

$$\tilde{B}_0(x) = 1, \tilde{B}_1(x) = \tilde{B}(x),$$

$$\tilde{B}_i(x) = \int_0^x \tilde{B}_{i-1}(x-u)d\tilde{B}(u), i \geq 2.$$

2.6.2 Метод вложенных цепей Маркова в приложении для системы $GI/M/1$

Кратко опишем применение метода вложенных цепей Маркова для анализа системы $GI/M/1$ – у которой рекуррентным является входящий поток, а показательное распределение имеет время обслуживания запроса. Пусть $A(t)$ – функция распределения интервалов между моментами поступления запросов, $\alpha(s)$ – ее преобразование Лапласа – Стилтьеса, а $\lambda = a_1^{-1}$ – интенсивность потока. Интенсивность показательно распределенного времени обслуживания будем обозначать μ.

Случайный процесс $i_t,\ t \geq 0$, – число запросов в системе в произвольный момент времени - здесь также является немарковским, поскольку поведение процесса после произвольного момента времени $t, t \geq 0$, не определяется полностью его состоянием в этот момент, а зависит также от уже прошедшего к данному моменту времени с момента поступления последнего запроса. В качестве вложенных моментов времени $t_k,\ k \geq 1$, возьмем моменты поступления запросов в системе. Поскольку в эти моменты происходят скачки значений процесса $i_t,\ t \geq 0$, для определенности будем считать, что $i_{t_k} = i_{t_k - 0}, k \geq 1$, т.е. запрос, который поступает в систему в данный момент, не включается в число запросов в системе в этот момент.

Нетрудно видеть, что случайный процесс $i_{t_k},\ k \geq 1$, является дискретной цепью Маркова с пространством состояний, совпадающим со множеством неотрицательных целых чисел. Вероятности одношаговых переходов вычисляются следующим образом:

$$p_{i,j} = \int_0^\infty \frac{(\mu t)^{i-j+1}}{(i-j+1)!} e^{-\mu t} dA(t), 1 \leq j \leq i+1,\ i \geq 0, \tag{2.55}$$

$$p_{i,0} = 1 - \sum_{j=1}^{i+1} p_{i,j}\ , i \geq 0, \tag{2.56}$$

$$p_{i,j} = 0, j > i+1,\ i \geq 0. \tag{2.57}$$

Формула (2.55) получается из следующих соображений. Поскольку $j \geq 1$, то между вложенными моментами постоянно шло обслуживание запросов и число обслужившихся за это время запросов равно $i - j + 1$. Так как время обслуживания имеет показательное распределение с пара-

метром μ, то поток моментов окончания обслуживания является простейшим (см. Утверждение 2.2), поэтому (см. Утверждение 2.1) вероятность того, что за время t будет обслужено $i-j+1$ запросов есть: $\frac{(\mu t)^{i-j+1}}{(i-j+1)!}e^{-\mu t}$. Усредняя по всевозможным значениям длины интервала между моментами поступления, получаем формулу (2.55). Формула (2.56) следует из условия нормировки.

Используя известные критерии эргодичности, можно убедиться, что необходимым и достаточным условием существования стационарных распределений

$$p_i = \lim_{t\to\infty} P\{i_t = i\}, r_i = \lim_{k\to\infty} P\{i_{t_k} = i\},\ i \geq 0,$$

будет уже известное нам условие

$$\rho < 1, \tag{2.58}$$

где коэффициент загрузки ρ определяется как $\rho = \frac{\lambda}{\mu}$. Будем далее считать это условие выполненным.

Уравнения равновесия для стационарных вероятностей $r_i, i \geq 0$, распределения вложенной цепи с учетом (2.55)-(2.57) выписываются в виде:

$$r_j = \sum_{i=j-1}^{\infty} r_i \int_0^{\infty} \frac{(\mu t)^{i-j+1}}{(i-j+1)!} e^{-\mu t} dA(t),\ j \geq 1. \tag{2.59}$$

Решение системы линейных алгебраических уравнений (2.59) будем искать в виде:

$$r_j = C\sigma^j,\ j \geq 0. \tag{2.60}$$

Подставляя (2.60) в (2.59), получаем, что неизвестное число σ удовлетворяет уравнению:

$$\sigma = \int_0^{\infty} e^{-\mu(1-\sigma)t} dA(t) = \alpha(\mu(1-\sigma)). \tag{2.61}$$

Покажем, что при выполнении условия (2.58) уравнение (2.61) имеет единственный действительный корень $\sigma, 0 < \sigma < 1$.

Обозначим $y(\sigma) = \sigma - \alpha(\mu(1-\sigma))$. Нам необходимо убедиться, что уравнение $y(\sigma) = 0$ имеет единственный действительный корень $\sigma, 0 < \sigma < 1$. Для этого исследуем свойства функции $y(\sigma)$:

$$y(0) = -\alpha(\mu) < 0;$$

$$y(1) = 0;$$
$$y^{'}(\sigma) = 1 + \mu\alpha^{'}(\mu(1-\sigma));$$
$$y^{'}(1) = 1 - \rho^{-1} < 0 \text{ в силу (2.58)};$$
$$y^{''}(\sigma) = -(\mu)^2\alpha^{''}(\mu(1-\sigma)) \leq 0.$$

Последнее неравенство справедливо в силу свойства 2.7 преобразования Лапласа – Стилтьеса.

Таким образом, функция $y(\sigma)$ – вогнутая, отрицательная в точке $\sigma = 0$, нулевая и убывающая в точке $\sigma = 1$. Отсюда следует доказываемая единственность корня.

Неизвестная константа C в (2.60) легко находится из условия нормировки: $\sum\limits_{i=0}^{\infty} r_i = 1$ и имеет вид: $C = 1 - \sigma$.

Таким образом, мы получили следующее выражение для стационарных вероятностей $r_i, i \geq 0$ распределения вложенной цепи Маркова:

$$r_i = \sigma^i(1-\sigma),\ i \geq 0, \tag{2.62}$$

где константа σ является корнем уравнения (2.61).

Найдем теперь стационарные вероятности p_i, $i \geq 0$, наличия i запросов в системе в произвольный момент времени. Фиксируем произвольный момент времени. В данный момент в системе может находиться i, $i \geq 1$, запросов, если в последний перед этим моментом момент поступления запроса в системе было j, $j \geq i-1$, запросов и за время u, прошедшее с момента поступления, $i - j + 1$ запросов ушли из системы, закончив обслуживание. Учитывая, что время u имеет функцию распределения $F(t) = \lambda \int\limits_0^t (1 - A(y))dy$ (см. подраздел 2.2), заключаем, что:

$$p_i = \sum_{j=i-1}^{\infty} r_j \lambda \int_0^{\infty} \frac{(\mu t)^{j-i+1}}{(j-i+1)!} e^{-\mu t}(1 - A(t))dt,\ i \geq 1. \tag{2.63}$$

Подставляя в это соотношение вероятности r_i в виде (2.62) и суммируя, получаем:

$$p_i = (1-\sigma)\sigma^{i-1}\lambda \int_0^{\infty} e^{-\mu(1-\sigma)t}(1 - A(t))dt,$$

откуда, учитывая связь преобразований Лапласа и Лапласа – Стилтьеса и уравнение (2.61), получаем:

$$p_i = \rho(1-\sigma)\sigma^{i-1},\ i \geq 1. \tag{2.64}$$

Вероятность p_0 находим из условия нормировки в виде:

$$p_0 = 1 - \rho. \tag{2.65}$$

Средние число L запросов в системе и число L_o запросов в очереди в произвольный момент времени находятся как:

$$L = \sum_{i=1}^{\infty} i p_i = \frac{\rho}{1-\sigma}, L_o = \sigma\frac{\rho}{1-\sigma}. \tag{2.66}$$

Найдем теперь стационарное распределение $W(x)$ времени ожидания произвольного запроса в системе. Запрос, заставший систему пустой (вероятность этого есть r_0), имеет нулевое время ожидания. Поскольку сумма i независимых показательных случайных величин с параметром μ есть эрланговская случайная величина порядка i, то запрос, заставший в системе $i, i \geq 1$, запросов, ждет в течение времени, распределенного по закону Эрланга с параметрами (i, μ). Отсюда получаем:

$$W(x) = 1 - \sigma + (1-\sigma)\sum_{i=1}^{\infty}\sigma^i \int_0^x \frac{\mu(\mu t)^{i-1}}{(i-1)!}e^{-\mu t}dt =$$

$$= 1 - \sigma e^{-\mu(1-\sigma)x}. \tag{2.67}$$

Среднее время W ожидания равно $\frac{\sigma}{\mu(1-\sigma)}$. Сравнивая это выражение с (2.66), видим, что формула Литтла справедлива и для данной СМО.

2.6.3 Метод введения дополнительной переменной

Сущность этого метода исследования немарковских процессов состоит в расширении пространства состояний процесса за счет введения в его описание некоторых дополнительных компонент с тем, чтобы полученный многомерный процесс был марковским. Если удается провести исследование этого марковского процесса (например, с помощью "Δt метода"), то затем распределение исходного немарковского процесса получается, как правило, элементарным образом.

Для иллюстрации снова, как и в подразделе 2.6.1, рассмотрим систему $M/G/1$, т.е. однолинейную СМО с ожиданием, на вход которой поступает простейший поток интенсивности λ, а время обслуживания запроса имеет произвольное распределение с функцией распределения $B(t)$, ее преобразованием Лапласа – Стилтьеса $\beta(s)$ и конечными начальными моментами $b_k,\ k = 1, 2$.

Мы уже отметили, что в случае этой системы процесс $i_t,\ t \geq 0$, – число запросов в системе в момент t – не является марковским. Проанализировав причину немарковости процесса $i_t,\ t \geq 0$, мы видим, что если включить в описание процесса дополнительную компоненту $\nu_t,\ t \geq 0$, имеющую смысл либо времени обслуживания, прошедшего к данному моменту времени t, либо времени, оставшегося до окончания обслуживания этого запроса, то полученный двумерный случайный процесс $\{i_t, \nu_t\},\ t \geq 0$, является марковским. Оба варианта марковизации примерно одинаково популярны в литературе. Изложим здесь второй вариант.

Итак, рассмотрим двумерный марковский случайный процесс $\{i_t, \nu_t\},\ t \geq 0$, где $\nu_t, \nu_t \geq 0$, – время до окончания обслуживания запроса, находящегося на приборе в момент t. Отметим, что при значении компоненты i_t, равном нулю, нет необходимости введения дополнительной переменной, т.к. в данный момент система не обслуживает запросы.

Введем в рассмотрение функции:

$$\varphi_t(0) = P\{i_t = 0\},$$

$$\varphi_t(i, x) = P\{i_t = i, \nu_t < x\}, i \geq 1, x > 0.$$

Утверждение 2.12. *Функции $\varphi_t(0), \varphi_t(i, x), i \geq 1, x > 0$ удовлетворяют следующей системе уравнений:*

$$\frac{\partial \varphi_t(0)}{\partial t} = -\lambda \varphi_t(0) + \frac{\partial \varphi_t(1, x)}{\partial x}|_{x=0}, \tag{2.68}$$

$$\frac{\partial \varphi_t(1, x)}{\partial t} - \frac{\partial \varphi_t(1, x)}{\partial x} = -\lambda \varphi_t(1, x) - \frac{\partial \varphi_t(1, x)}{\partial x}|_{x=0} + \\ + \frac{\partial \varphi_t(2, x)}{\partial x}|_{x=0} B(x) + \lambda \varphi_t(0) B(x), \tag{2.69}$$

$$\frac{\partial \varphi_t(i, x)}{\partial t} - \frac{\partial \varphi_t(i, x)}{\partial x} = -\lambda \varphi_t(i, x) - \frac{\partial \varphi_t(i, x)}{\partial x}|_{x=0} + \\ + \frac{\partial \varphi_t(i+1, x)}{\partial x}|_{x=0} B(x) + \lambda \varphi_t(i-1, x), i \geq 2. \tag{2.70}$$

Доказательство состоит в применении формулы полной вероятности и анализе возможных переходов процесса за время Δt и вероятностей соответствующих переходов. В результате приходим к следующей системе разностных уравнений для интересующих нас функций:

$$\varphi_{t+\Delta t}(0) = \varphi_t(0)(1-\lambda\Delta t) + \varphi_t(1,\Delta t) + o(\Delta t),$$

$$\varphi_{t+\Delta t}(1,x) = (\varphi_t(1,x+\Delta t) - \varphi_t(1,\Delta t))(1-\lambda\Delta t) + \varphi_t(2,\Delta t)B(x) + \quad (2.71)$$
$$+\varphi_t(0)\lambda\Delta t B(x) + o(\Delta t),$$
$$\varphi_{t+\Delta t}(i,x) = (\varphi_t(i,x+\Delta t) - \varphi_t(i,\Delta t))(1-\lambda\Delta t) + \varphi_t(i+1,\Delta t)B(x) +$$
$$+\varphi_t(i-1,x+\Delta t)\lambda\Delta t + o(\Delta t), i \geq 2.$$

Деля обе части уравнений (2.71) на Δt и устремляя Δt к нулю, получаем систему (2.68) - (2.70).

Выше мы уже отмечали, что задача нахождения нестационарного (зависящего от времени t) распределения вероятностей состояний СМО решается аналитически только в довольно редких случаях. Поэтому переходим к нахождению стационарного распределения процесса $\{i_t, \nu_t\}$:

$$\varphi(0) = \lim_{t\to\infty} \varphi_t(0), \varphi(i,x) = \lim_{t\to\infty} \varphi_t(i,x), i \geq 1, x > 0. \quad (2.72)$$

Условием существования пределов (2.72) является выполнение неравенства:

$$\rho = \lambda b_1 < 1.$$

Будем далее считать это условие выполненным.

Переходя в (2.67) - (2.70) к пределу при $t \to \infty$, получаем следующую систему уравнений для стационарного распределения вероятностей процесса $\{i_t, \nu_t\},\ t \geq 0,$:

$$\lambda\varphi(0) = \frac{\partial\varphi(1,x)}{\partial x}|_{x=0}, \quad (2.73)$$

$$\frac{\partial\varphi(1,x)}{\partial x} - \frac{\partial\varphi(1,x)}{\partial x}|_{x=0} - \lambda\varphi(1,x) + \frac{\partial\varphi(2,x)}{\partial x}|_{x=0}B(x) +$$
$$+\lambda\varphi(0)B(x) = 0, \quad (2.74)$$
$$\frac{\partial\varphi(i,x)}{\partial x} - \frac{\partial\varphi(i,x)}{\partial x}|_{x=0} - \lambda\varphi(i,x) + \frac{\partial\varphi(i+1,x)}{\partial x}|_{x=0}B(x) + \quad (2.75)$$
$$+\lambda\varphi(i-1,x) = 0, i \geq 2.$$

Для решения данной бесконечной системы уравнений применим аппарат производящих функций. Введем в рассмотрение производящую функцию:

$$\Phi(z,x) = \varphi(0) + \sum_{i=1}^{\infty} \varphi(i,x) z^i, |z| < 1.$$

Умножая уравнения системы (2.73) - (2.75) на соответствующие степени z и суммируя, получаем следующее уравнение для производящей функции $\Phi(z,x)$:

$$\frac{\partial \Phi(z,x)}{\partial x} - \lambda(1-z)\Phi(z,x) = \frac{\partial \Phi(z,x)}{\partial x}|_{x=0}\left(1 - \frac{B(x)}{z}\right) - \lambda\varphi(0)(1-z)(1-B(x)). \tag{2.76}$$

Для решения дифференциально-функционального уравнения (2.76) введем в рассмотрение преобразование Лапласа:

$$\phi(z,s) = \int_0^{\infty} e^{-sx}\Phi(z,x)dx, Re\ s > 0.$$

Применяя преобразование Лапласа к обеим частям уравнения (2.76) и используя сведения о связи преобразований Лапласа и Лапласа – Стилтьеса, а также свойство 2.3 преобразования Лапласа – Стилтьеса, получаем уравнение вида:

$$(s - \lambda(1-z))\phi(z,s) = \tag{2.77}$$
$$= \frac{1}{s}\left[\frac{\partial \Phi(z,x)}{\partial x}|_{x=0}\left(1 - \frac{\beta(s)}{z}\right) - \lambda\varphi(0)(1-z)(1-\beta(s))\right] + \varphi(0).$$

Известно, что производящая функция является аналитической (т.е. представимой в виде сходящегося степенного ряда) функцией при $|z| < 1$, а преобразование Лапласа аналитично в области $Re\ s > 0$. Поэтому для любого $z, |z| < 1$ при $s = \lambda(1-z)$ левая часть соотношения (2.77) обращается в нуль. Следовательно, при таком s обращается в нуль и правая часть (2.77). Из этого условия после несложных преобразований получаем:

$$\frac{\partial \Phi(z,x)}{\partial x}|_{x=0} = \varphi(0)\lambda z \frac{\beta(\lambda(1-z))(1-z)}{\beta(\lambda(1-z)) - z}. \tag{2.78}$$

Подставляя (2.78) в (2.77), получаем:

$$(s - \lambda(1-z))\phi(z,s) = \tag{2.79}$$

$$= \frac{\lambda\varphi(0)(1-z)}{s}\left[\frac{z\beta(\lambda(1-z))\lambda(1-z)}{\beta(\lambda(1-z))-z}\left(1-\frac{\beta(s)}{z}\right)-1+\beta(s)\right]+\varphi(0).$$

Формула (2.79) дает вид искомого стационарного распределения с точностью до значения вероятности $\varphi(0)$. Сейчас уместно вспомнить, что мы рассматриваем двумерный марковский процесс $\{i_t, \nu_t\},\ t \geq 0$, вынужденно, т.к. интересующий нас процесс $i_t,\ t \geq 0$, – число запросов в системе в момент t является немарковским. Несложно видеть, что стационарные распределения процессов $\{i_t, \nu_t\},\ t \geq 0$, и $i_t,\ t \geq 0$, связаны соотношениями:

$$p_0 = \lim_{t\to\infty} P\{i_t = 0\} = \varphi(0),$$
$$p_i = \lim_{t\to\infty} P\{i_t = i\} = \lim_{x\to\infty} \varphi(i,x), i \geq 1.$$

Поэтому производящая функция $P(z) = \sum_{i=0}^{\infty} p_i z^i$ определяется как:

$$P(z) = \lim_{x\to\infty} \Phi(z,x).$$

Вспоминая связь преобразований Лапласа и Лапласа – Стилтьеса, а также свойство 2.4 преобразования Лапласа – Стилтьеса, получаем:

$$P(z) = \lim_{x\to\infty} \Phi(z,x) = \lim_{s\to 0} s\phi(z,s),$$

откуда с учетом (2.79) получаем:

$$P(z) = p_0 \frac{(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z))-z}. \tag{2.80}$$

Вычисляя из условия нормировки $P(1) = 1$ константу p_0 в виде $p_0 = 1-\rho$, мы окончательно получаем уже известную нам формулу Полячека – Хинчина:

$$P(z) = (1-\rho)\frac{(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z))-z}. \tag{2.81}$$

Отметим, что из нее элементарно следует уже известная нам формула для стационарного распределения числа запросов в системе $M/M/1$:

$$p_i = (1-\rho)\rho^i,\ i \geq 0.$$

Для системы $M/D/1$ с постоянным временем обслуживания запросов явные выражения для стационарных вероятностей следующие:

$$p_0 = 1-\rho, p_1 = (1-\rho)(e^\rho - 1),$$
$$p_i = (1-\rho)\sum_{k=1}^{i}(-1)^{i-k}e^{k\rho}\left[\frac{(k\rho)^{i-k}}{(i-k)!}+\frac{(k\rho)^{i-k-1}}{(i-k-1)!}\right], i \geq 2.$$

2.6.4 Метод введения дополнительного события

Этот метод, предложенный Данцигом, Кестеном и Ранненбергом (метод коллективных меток – method of collective marks) и развитый затем Г.П. Климовым (метод "катастроф"), позволяет легко получить аналитические результаты в ситуациях, когда другие известные методы приводят к трудоемким выкладкам. Особенно эффективен он оказался при анализе ненадежных и приоритетных систем массового обслуживания.

Сущность этого метода заключается в следующем. Пусть требуется найти некоторое распределение, характеризующее функционирование СМО. Производящей функции этого распределения (если распределение дискретное) или его преобразованию Лапласа – Стилтьеса придается вероятностный смысл за счет "раскрашивания" запросов или введения в рассмотрение потока "катастроф". Затем вводится в рассмотрение некоторое (дополнительное) случайное событие и вероятность его подсчитывается в терминах производящей функции или преобразованию Лапласа – Стилтьеса искомого распределения двумя различными способами. В результате получается уравнение, решением которого является функция, которая интересует исследователя.

Проиллюстрируем этот метод, применив его для нахождения вероятностных характеристик системы $M/G/1$. Важной характеристикой производительности многих реальных систем является распределение периода занятости системы. Период занятости есть интервал времени с момента поступления запроса в пустую систему до момента, когда система впервые вновь окажется пустой. Знание периода занятости позволяет решать задачи, связанные, например, с планированием проведения в системе профилактических работ, исследованием возможности дополнительной загрузки прибора выполнением некоторой второстепенной "фоновой" работы и т.д.

Обозначим $\Pi(t),\ t \geq 0$, функцию стационарного распределения длины периода занятости в рассматриваемой системе, а $\pi(s),\ s > 0$, – ее преобразование Лапласа – Стилтьеса.

Считаем, что выполняется условие:

$$\rho < 1, \tag{2.81}$$

гарантирующее существование стационарного распределения длины периода занятости рассматриваемой СМО.

Утверждение 2.13. *Преобразование $\pi(s)$ Лапласа – Стилтьеса распределения длины периода занятости рассматриваемой СМО удовлетворя-*

ет следующему функциональному уравнению:

$$\pi(s) = \beta(s + \lambda(1 - \pi(s))). \tag{2.82}$$

Доказательство. Легко видеть, что распределение длины периода занятости системы не зависит от того, в каком порядке обслуживаются запросы. Для облегчения анализа структуры периода занятости предположим, что запросы обслуживаются в инверсионном порядке, т.е. на обслуживание всегда выбирается запрос, пришедший в систему последним. Такая дисциплина выбора из очереди кодируется как LIFO (Last In – First Out) или LCFS (Last Came – First Served). При такой дисциплине выбора из очереди каждый запрос как бы порождает период занятости системы запросами, пришедшими в систему после него. Причем структура и, следовательно, распределение длины периода занятости, порожденного некоторым запросом, такие же, как структура и распределение длины периода занятости системы. Используя эти рассуждения, мы приходим к пониманию того, что период занятости системы состоит из времени обслуживания первого запроса, с которого начался период занятости, и случайного числа периодов занятости, порожденных запросами, пришедшими в систему за время обслуживания первого запроса.

Теперь предположим, что независимо от функционирования данной системы поступает простейший поток катастроф интенсивности s. Введем в рассмотрение (дополнительное) событие A, состоящее в том, что за данный период занятости не поступили катастрофы.

Напомним, что согласно вероятностной трактовке преобразования Лапласа – Стилтьеса, величина $h(s) = \int\limits_0^\infty e^{-st} dH(t), s > 0$ есть вероятность того, что не произойдет ни одной катастрофы за случайное время, имеющее функцию распределения $H(t)$. Поэтому легко понять, что вероятность события A определяется следующим образом:

$$P(A) = \pi(s). \tag{2.83}$$

Найдем теперь вероятность этого же события иначе. Назовем произвольный запрос “плохим”, если за период занятости, порожденный им, наступает катастрофа. Используя достигнутое нами понимание структуры периода занятости, нетрудно убедиться, что для того, чтобы запрос, с которого начался период занятости, был неплохим (вероятность этого есть $P(A)$), необходимо и достаточно, чтобы за время его обслуживания не поступили события из суммарного потока катастроф и потока плохих запросов.

Поток катастроф является простейшим потока интенсивности s. Поток плохих запросов получается из исходного простейшего потока интенсивности λ в результате применения простейшей процедуры рекуррентного просеивания (произвольный запрос включается в просеянный поток с вероятностью $1-P(A)=1-\pi(s)$ независимо от других запросов). Поэтому, согласно Утверждению 2.6, просеянный поток является простейшим потоком интенсивности $\lambda(1-\pi(s))$. Согласно Утверждению 2.5, суммарный поток катастроф и плохих запросов является простейшим потоком интенсивности $s+\lambda(1-\pi(s))$.

Таким образом, используя еще раз вероятностную трактовку преобразования Лапласа – Стилтьеса, мы получаем следующую формулу для вероятности события A :

$$P(A)=\beta(s+\lambda(1-\pi(s))). \tag{2.84}$$

Сравнивая выражения (2.83) и (2.84), мы убеждаемся в справедливости формулы (2.82). Утверждение 2.13 доказано.

Уравнение (2.82), полученное Дж. Кендаллом в 1951 году, имеет единственное решение в области $Res>0$, такое, что $|\pi(s)|\leq 1$.

В случае, если распределение времени обслуживания показательное, $B(t)=1-e^{-\mu t}$, рассматриваемая система есть $M/M/1$ и преобразование Лапласа – Стилтьеса распределения времени обслуживания $\beta(s)$ имеет вид: $\beta(s)=\frac{\mu}{\mu+s}$. При этом функциональное уравнение (2.82) переходит в квадратное уравнение для неизвестного преобразования Лапласа – Стилтьеса $\pi(s)$:

$$\rho\pi^2(s)-(s\mu^{-1}+\rho+1)\pi(s)+1=0. \tag{2.85}$$

Решая уравнение (2.85), получаем:

$$\pi(s)=\frac{1+s\mu^{-1}+\rho\pm\sqrt{(1+s\mu^{-1})^2-4\rho}}{2\rho}. \tag{2.86}$$

В этой формуле выбираем только знак "-", чтобы полученное решение удовлетворяло условию $|\pi(s)|\leq 1$. Обращая теперь преобразование Лапласа – Стилтьеса $\pi(s)$, получаем следующее выражение для производной функции $\Pi(t)$ распределения длины периода занятости системы $M/M/1$:

$$\Pi'(t)=\frac{1}{t\sqrt{\rho}}e^{-(\lambda+\mu)t}I_1(2t\sqrt{\lambda\mu}), \tag{2.87}$$

где функция $I_1(x)$ есть модифицированная функция Бесселя первого рода.

В общем случае уравнение (2.82) можно решать методом итераций, снабдив функцию $\pi(s)$ индексом $n+1$ в левой части уравнения и индексом n – в правой части. Эта процедура имеет геометрическую скорость сходимости последовательности $\pi_n(s), n \geq 1$, к значению $\pi(s)$ при фиксированном значении аргумента s.

Кроме того, путем последовательного дифференцирования уравнения (2.82) с последующей подстановкой аргумента $s = 0$ и учета свойства 2.5 преобразования Лапласа – Стилтьеса можно получить рекуррентную последовательность формул для вычисления начальных моментов распределения длины периода занятости. Так, среднее значение π_1 длины периода занятости и второй начальный момент π_2 ее распределения определяются формулой:

$$\pi_1 = \frac{b_1}{1-\rho}, \quad \pi_2 = \frac{b_2}{(1-\rho)^3}. \tag{2.88}$$

Как и следовало ожидать, с ростом коэффициента загрузки ρ и приближением его значения к единице среднее значение периода занятости стремится к бесконечности.

Рассмотрим теперь другую характеристику функционирования системы $M/G/1$ – число ξ запросов, обслуженных за период занятости.

Обозначим $\gamma_i = P\{\xi = i\},\ i \geq 1$, и $\Gamma(z) = \sum_{i=1}^{\infty} \gamma_i z^i$.

Утверждение 2.14. *Производящая функция* $\Gamma(z), |z| < 1$, *удовлетворяет следующему функциональному уравнению:*

$$\Gamma(z) = z\beta(\lambda(1-\Gamma(z))). \tag{2.89}$$

Доказательство. Производящей функции $\Gamma(z)$ придадим вероятностный смысл следующим образом. Каждый из запросов независимо от других назовем красным с вероятностью z и синим с дополнительной вероятностью. Произвольный запрос назовем темно-красным, если он сам красный и за период занятости, порожденный им, в системе обслуживались только красные запросы. Введем событие A, состоящее в том, что запрос, с которого начинается период занятости, является темно-красным. Найдем вероятность этого события. С одной стороны, очевидно, что

$$P(A) = \Gamma(z). \tag{2.90}$$

С другой стороны, из проделанного выше анализа структуры периода занятости ясно, что для того, чтобы запрос был темно-красным, необходимо

и достаточно, чтобы он сам был красным (вероятность этого равна z) и за время его обслуживания могли поступать только темно-красные запросы. Так как поток запросов – простейший с параметром λ, а произвольный запрос является темно-красным с вероятностью $\Gamma(z)$, то поток не темно-красных вызовов (как это следует из Утверждения 2.6) является простейшим с параметром $\lambda(1-\Gamma(z))$. Вспоминая вероятностную интерпретацию преобразования Лапласа – Стилтьеса, из приведенных рассуждений выводим следующую альтернативную формулу для вероятности события A :

$$P(A) = z\beta(\lambda(1-\Gamma(z))). \tag{2.91}$$

Сравнивая формулы (2.90) и (2.91), убеждаемся в справедливости (2.89). □

Уравнение (2.89) определяет единственную аналитическую в области $|z| < 1$ функцию, такую, что $|\Gamma(z)| < 1$.

Следствие 2.1. *Среднее число $M\xi$ запросов, обслуженных в системе $M/G/1$ за один период занятости, задается формулой:*

$$M\xi = \frac{1}{1-\rho}.$$

Приведем еще одно доказательство формулы Полячека – Хинчина для производящей функции распределения вероятностей числа запросов в системе $M/G/1$ в моменты окончания обслуживания. Каждый из запросов, приходящих в систему, независимо от других назовем красным с вероятностью z и синим с дополнительной вероятностью. Введем событие A, состоящее в том, что запрос, уходящий в данный момент окончания обслуживания из системы, сам красный, и все запросы, остающиеся в системе в этот момент, тоже красные.

Из вероятностной интерпретации производящей функции очевидно следует, что:

$$P(A) = z\Pi(z), \tag{2.92}$$

где $\Pi(z)$ есть искомая производящая функция распределения вероятностей числа запросов в системе $M/G/1$ в моменты окончания обслуживания.

С другой стороны, для того, чтобы произошло событие A, необходимо и достаточно, чтобы все запросы, которые находились в системе в предыдущий момент окончания обслуживания (если система была непуста), были красными и за время обслуживания не пришли синие запросы, а если система была пуста, то первый пришедший запрос должен быть красным и за

время его обслуживания не пришли синие запросы. Из этих рассуждений следует, что:

$$P(A) = (\Pi(z) - \pi_0)\beta(\lambda(1-z)) + \pi_0 z\beta(\lambda(1-z)).$$

Из соотношений этого соотношения и (2.92) очевидным образом следует формула Полячека – Хинчина:

$$\Pi(z) = \pi_0 \frac{(1-z)\beta(\lambda(1-z))}{\beta(\lambda(1-z)) - z},$$

полученная нами ранее с помощью метода вложенных цепей Маркова.

В заключение подраздела найдем характеристики системы $M/G/1$ с дисциплиной LIFO.

Выше отмечалось, что распределение периода занятости системы $M/G/1$ не зависит от дисциплины обслуживания. Поэтому уравнение (2.82) определяет преобразование Лапласа – Стилтьеса распределения периода занятости для всех дисциплин. Кроме того, несложно видеть, что и распределения числа запросов в системе $M/G/1$ при дисциплинах FIFO и LIFO совпадают и задаются формулой (2.81).

Распределение времени ожидания запроса при дисциплинах FIFO и LIFO различно. При дисциплине FIFO преобразование Лапласа – Стилтьеса $w(s)$ стационарного распределения времени ожидания задается формулой (2.52).

Утверждение 2.15. *При дисциплине LIFO преобразование Лапласа – Стилтьеса $w(s)$ имеет следующий вид:*

$$w(s) = 1 - \rho + \frac{\frac{\lambda}{s}(1-\pi(s))}{1 + \frac{\lambda}{s}(1-\pi(s))}, \tag{2.93}$$

где функция $\pi(s)$ является решением уравнения (2.82).

Доказательство. Введем поток катастроф и понятие "плохого" запроса, как это было сделано при доказательстве Утверждения 2.13. При этом функция $w(s)$ есть вероятность того, что за время ожидания данного запроса не наступит катастрофа, а функция $\pi(s)$ есть вероятность того, что произвольный запрос не является "плохим", т.е. катастрофа не наступает за период занятости, порожденный этим запросом.

Учитывая сущность дисциплины LIFO и рассуждения, использованные при доказательстве Утверждения 2.13, получаем формулу:

$$w(s) = p_0 + (1-p_0)\tilde{\beta}(s + \lambda - \lambda\pi(s)), \tag{2.94}$$

где $\tilde{\beta}(s)$ есть преобразование Лапласа – Стилтьеса распределения остаточного (после момента поступления запроса, время ожидания которого мы исследуем) времени обслуживания запроса, находящегося на приборе. По аналогии с функцией распределения $F(t)$ остаточного времени до момента поступления запроса в рекуррентном потоке, приведенной в подразделе 2.2, для функции распределения $\tilde{B}(t)$ остаточного времени обслуживания имеем формулу:

$$\tilde{B}(t) = b_1^{-1} \int_0^t (1 - B(u))du. \tag{2.95}$$

Отметим, что эта функция уже использовалась в формуле Бенеша (2.54). Легко видеть, что преобразование Лапласа – Стилтьеса этой функции имеет вид:

$$\tilde{\beta}(s) = \frac{1 - \beta(s)}{sb_1}.$$

Учитывая это, а также уравнение (2.82) и формулу $p_0 = 1 - \rho$, из (2.94) получаем соотношение (2.93). □

Замечание 2.1. Сравнивая формулы (2.52) и (2.93), заключаем, что распределения времени ожидания в системах с дисциплинами FIFO и LIFO – различные. При этом средние времена ожидания совпадают.

2.7 Многолинейные системы массового обслуживания

В предыдущем параграфе мы не касались общей однолинейной СМО $G/G/1$, поскольку для нее не удается получить точных аналитических результатов даже для средних значений длины очереди и времени ожидания запросов в системе. Для этих величин получен лишь ряд оценок снизу и сверху, позволяющих приближенно вычислить их значение. Возможна довольно точная аппроксимация характеристик этой системы с помощью характеристик системы $PH/PH/1$, которая поддается аналитическому исследованию.

По аналогичной причине мы не касаемся систем типа $G/G/n$ и $M/G/n$. Отметим, что средние характеристики последней системы, как правило, оценивают путем суммирования с некоторыми весами соответствующих известных средних характеристик для систем обслуживания типа $M/M/n$

и $M/D/n$. Система типа $GI/M/n$ поддается аналитическому исследованию довольно легко при помощи метода вложенных цепей Маркова и результаты имеют форму, близкую к полученным для системы $GI/M/1$ в предыдущем параграфе. Поэтому мы также не затрагиваем ее.

В подразделе 2.7.1 мы исследуем систему типа $M/M/n$ и систему $M/M/n/m$. В подразделе 2.7.2 приведем результаты для системы типа $M/M/n/0$ (системы Эрланга) и ее обобщения – системы $M/G/n/0$. В последнем подразделе 2.7.3 изучается система типа $M/M/\infty$.

2.7.1 Системы $M/M/n$ и $M/M/n/m$

Пусть имеется n параллельных идентичных обслуживающих устройств (каналов) и бесконечный буфер для ожидания. Входящий поток является простейшим с интенсивностью λ, а время обслуживания запроса в канале имеет показательное распределение с параметром μ. Запрос, пришедший в систему и заставший хотя бы один канал свободным, немедленно занимает любой из свободных каналов и начинает обслуживаться. Если все каналы в момент поступления запроса заняты, он присоединяется к очереди. Из очереди запросы выбираются на обслуживание согласно дисциплине $FIFO$.

Рассмотрим случайный процесс i_t – число запросов в рассматриваемой системе в момент t, $i_t \geq 0,\ t \geq 0$. Легко убедиться, что процесс i_t является процессом гибели и размножения с параметрами:

$$\gamma_0 = \lambda, \gamma_i = \lambda + i\mu, 1 \leq i \leq n, \gamma_i = \lambda + n\mu, i > n,$$

$$p_i = \frac{\lambda}{\lambda + i\mu}, 1 \leq i \leq n, p_i = \frac{\lambda}{\lambda + n\mu}, i > n.$$

Поэтому распределение вероятностей состояний процесса $i_t,\ t \geq 0$, удовлетворяет системе дифференциальных уравнений (2.2),(2.3).

Интенсивности размножения λ_i и гибели μ_i в данном случае определяются следующим образом:

$$\lambda_i = \lambda,\ i \geq 0, \mu_i = i\mu,\ 1 \leq i \leq n,\ \mu_i = n\mu,\ i > n.$$

Тогда величины ρ_i имеют вид

$$\rho_i = \frac{(\rho n)^i}{i!}, 0 \leq i \leq n, \rho_i = \frac{n^n}{n!}\rho^i, i > n,$$

где $\rho = \frac{\lambda}{n\mu}$.

Параметр ρ, характеризующий соотношение интенсивности входящего потока и суммарной интенсивности обслуживания всеми приборами, является коэффициентом загрузки системы.

Проверяя условие существования стационарного распределения процесса $i_t,\ t \geq 0$, данное в Утверждении 2.9, легко убедиться, что стационарное распределение числа запросов в рассматриваемой системе

$$\pi_i = \lim_{t\to\infty} P\{i_t = i\}, i \geq 0,$$

существует, если выполняется условие:

$$\rho < 1.$$

Будем далее считать это условие выполненным.

Утверждение 2.16. *Стационарные вероятности $\pi_i, i \geq 0$, определяются следующим образом:*

$$\pi_i = \begin{cases} \pi_0 \frac{(n\rho)^i}{i!}, & 1 \leq i \leq n, \\ \pi_0 \frac{n^n}{n!}\rho^i, & i > n, \end{cases} \tag{2.96}$$

где

$$\pi_0 = \left[\sum_{j=0}^{n-1} \frac{(n\rho)^j}{j!} + \frac{(n\rho)^n}{n!(1-\rho)}\right]^{-1}. \tag{2.97}$$

Справедливость формул (2.96), (2.97) следует непосредственно из Утверждения 2.9.

Следствие 2.2. *Среднее число L запросов, находящихся в системе в произвольный момент времени, определяется следующим образом:*

$$L = \pi_0 \left[\sum_{j=1}^{n-1} \frac{(n\rho)^j}{(j-1)!} + \frac{(n\rho)^n}{n!}\frac{n(1-\rho)+\rho}{(1-\rho)^2}\right]. \tag{2.98}$$

Следствие 2.3. *Среднее число L_o запросов, находящихся в очереди в произвольный момент времени, определяется следующим образом:*

$$L_o = \pi_0 \frac{(n\rho)^n \rho}{n!(1-\rho)^2}. \tag{2.99}$$

Следствие 2.4. *Вероятность P_o того, что произвольный запрос вынужден ждать обслуживания в очереди, вычисляется по формуле:*

$$P_o = \frac{(n\rho)^n}{n!(1-\rho)}\left[\sum_{j=0}^{n-1} \frac{(n\rho)^j}{j!} + \frac{(n\rho)^n}{n!(1-\rho)}\right]^{-1}. \tag{2.100}$$

Формулу (2.100) иногда называют C - формулой Эрланга.

Функция $W(x)$ стационарного распределения времени ожидания для данной системы определяется следующим образом:

$$W(x) = 1 - P_o e^{-n\mu(1-\rho)x}, x \geq 0. \tag{2.101}$$

Вывод последней формулы аналогичен выводу формулы (2.24). Среднее время ожидания W имеет вид:

$$W = \frac{P_o}{n\mu(1-\rho)}. \tag{2.102}$$

Сравнивая формулы (2.99) и (2.102), замечаем, что для данной системы справедлив следующий вариант формулы Литтла:

$$L_o = \lambda W.$$

Рассмотрим теперь кратко систему $M/M/n/m$, в которой размер очереди ограничен числом $m, m \geq 1$. Запрос, заставший все приборы и все места в очереди занятыми, покидает систему навсегда, не оказывая никакого влияния на ее дальнейшее функционирование.

Стационарные вероятности $\pi_i, 0 \leq i \leq n+m$ наличия в произвольный момент времени i запросов в системе определяются формулами (2.96), в которых вероятность π_0 вычисляется следующим образом:

$$\pi_0 = \left[\sum_{j=0}^{n-1} \frac{(n\rho)^j}{j!} + \frac{(n\rho)^n}{n!}\frac{1-\rho^{m+1}}{1-\rho}\right]^{-1}. \tag{2.103}$$

Величины L, L_o, P_o определяются по формулам:

$$L = \sum_{i=1}^{n+m} i\pi_i, L_o = \sum_{i=n+1}^{n+m} (i-n)\pi_i, P_o = \sum_{i=n}^{n+m} \pi_i.$$

Функция $W(x)$ стационарного распределения времени ожидания в данной системе вычисляется по схеме, приведенной в параграфе 2.3, и представляет собой смесь эрланговских распределений со скачком в нуле.

2.7.2 Системы $M/M/n/0$ и $M/G/n/0$

Пусть имеется n параллельных идентичных обслуживающих устройств (каналов), буфер для ожидания отсутствует. Входящий поток является

простейшим с интенсивностью λ, а время обслуживания запроса в канале характеризуется функцией распределения $B(t)$ с конечным средним. Сначала рассматриваем случай, когда распределение $B(t)$ – показательное распределение с параметром μ.

Запрос, пришедший в систему и заставший хотя бы один канал свободным, немедленно занимает любой из свободных каналов и начинает обслуживаться. Если все каналы в момент поступления запроса заняты, он теряется, т.е. покидает систему навсегда, не оказывая никакого влияния на ее дальнейшее функционирование.

С исследования этой модели А.К. Эрлангом и ведет свой отсчет теория СМО. Практическая важность модели обусловлена тем, что она довольно адекватно описывает функционирование пучка телефонных каналов, на который поступает поток запросов на установление соединения.

Рассмотрим случайный процесс i_t – число запросов в рассматриваемой системе в момент t, $0 \leq i_t \leq n,\ t \geq 0$. Нетрудно убедиться, что процесс i_t является процессом гибели и размножения с параметрами:

$$\gamma_0 = \lambda, \gamma_i = \lambda + i\mu, 1 \leq i \leq n-1, \gamma_i = n\mu, i = n,$$

$$p_i = \frac{\lambda}{\lambda + i\mu}, 1 \leq i \leq n-1, p_n = 0.$$

Параметр ρ (в отличие от предыдущего подраздела) определим здесь как соотношение интенсивности входящего потока и интенсивности обслуживания одним прибором: $\rho = \frac{\lambda}{\mu}$.

Можно показать, что в силу конечности пространства состояний процесса $i_t,\ t \geq 0$, стационарное распределение числа запросов в рассматриваемой системе

$$\pi_i = \lim_{t\to\infty} P\{i_t = i\}, 0 \leq i \leq n$$

существует при любых конечных значениях интенсивностей входящего потока и обслуживания запросов.

Утверждение 2.17. *Стационарные вероятности $\pi_i, 0 \leq i \leq n$ определяются следующим образом:*

$$\pi_i = \frac{\frac{\rho^i}{i!}}{\sum\limits_{j=0}^{n} \frac{\rho^j}{j!}}, 0 \leq i \leq n. \tag{2.104}$$

Справедливость этого утверждения следует из Утверждения 2.9.

Следствие 2.5. *Вероятность P_{loss} потери произвольного запроса имеет вид*

$$P_{loss} = \pi_n = \frac{\frac{\rho^n}{n!}}{\sum\limits_{j=0}^{n} \frac{\rho^j}{j!}}. \tag{2.105}$$

Эта формула, называемая B-формулой Эрланга, играет весьма важную роль в телефонии. С ее помощью можно вычислить вероятность потери запросов (блокировки каналов) при фиксированном числе каналов, интенсивности входящего потока и интенсивности обслуживания запросов. Можно решать также двойственные задачи, например расчет необходимого числа каналов или допустимого потока запросов, исходя из заданной максимально допустимой вероятности потери запроса.

В процессе использования B-формулы Эрланга было замечено, что вероятность отказа, вычисленная по формуле (2.105), очень хорошо согласовывалась со значением этой же вероятности, вычисленной как средняя доля потерянных запросов в реально функционирующей системе. Это казалось несколько странным, поскольку вероятность (2.105) подсчитывается в предположении, что входящий поток является простейшим, а распределение времени обслуживания – показательное. И если хорошее качество аппроксимации потоков информации в телефонных сетях простейшим потоком может быть объяснено с учетом Утверждения 7, то нечувствительность вероятности отказа к виду распределения времени обслуживания вызывала вопрос. Наиболее вероятное значение показательно распределенной случайной величины есть 0, что плохо согласуется с реальной статистикой длительности телефонных разговоров.

Поэтому усилия многих специалистов в области СМО были направлены на доказательство инвариантности вида (2.105) вероятности потери запроса относительно вида функции распределения времени обслуживания при фиксированном значении среднего времени обслуживания. Строгое доказательство этого факта принадлежит Б.А. Севастьянову, который установил, что распределение вероятностей состояний системы $M/G/n/0$ действительно инвариантно относительно распределения времени обслуживания запросов.

2.7.3 Система $M/M/\infty$

Пусть система имеет бесконечное число параллельных идентичных обслуживающих устройств (каналов), т.е. любой запрос, пришедший в систему, немедленно начинает обслуживаться. Входящий поток является простейшим с интенсивностью λ, а время обслуживания запроса в канале имеет показательное распределение с параметром μ.

Исследование такой системы представляет интерес с нескольких точек зрения. Во-первых, при небольшой интенсивности входящего потока и большом (но конечном) числе каналов практически невозможна одновременная занятость всех каналов и поэтому число каналов можно считать бесконечным. Поэтому данная модель может служить для аппроксимации модели с большим числом каналов. Во-вторых, эта модель является одной из немногих, для которых нестационарное распределение вероятностей числа запросов в системе удается получить в относительно простой форме, благодаря чему эту систему иногда называют простейшей СМО.

Рассмотрим случайный процесс i_t – число запросов в рассматриваемой системе в момент t, $i_t \geq 0,\ t \geq 0$. Этот процесс является процессом гибели и размножения с параметрами:

$$\gamma_0 = \lambda, \gamma_i = \lambda + i\mu, i \geq 1,$$

$$p_i = \frac{\lambda}{\lambda + i\mu}, i \geq 1.$$

Обозначим $P_i(t) = P\{i_t = i\}, i \geq 0$. Учитывая, что интенсивности размножения λ_i и интенсивности гибели μ_i для рассматриваемого процесса $i_t,\ t \geq 0$, имеют вид: $\lambda_i = \lambda, i \geq 0, \mu_i = i\mu, i \geq 1$, можно записать систему линейных дифференциальных уравнений (2.2),(2.3) для вероятностей $P_i(t), i \geq 0$, в следующем виде:

$$P_0^{'}(t) = -\lambda P_0(t) + \mu P_1(t), \tag{2.106}$$

$$P_i^{'}(t) = \lambda P_{i-1}(t) - (\lambda + i\mu)P_i(t) + (i+1)\mu P_{i+1}(t), i \geq 1. \tag{2.107}$$

Поскольку нас интересует нестационарные распределения вероятностей, мы должны зафиксировать начальное состояние СМО. Предположим, что в момент времени 0 в системе обслуживалось k запросов. Тогда начальное условие для системы (2.106), (2.107) имеет вид

$$P_k(0) = 1, P_i(0) = 0, i \neq k. \tag{2.108}$$

Для решения системы (2.106), (2.107) введем в рассмотрение производящую функцию $\Pi(z,t) = \sum\limits_{i=0}^{\infty} P_i(t) z^i, |z| < 1$.

Умножая уравнения системы (2.106),(2.107) на соответствующие степени z и суммируя, получаем следующее уравнение в частных производных первого порядка:

$$\frac{\partial \Pi(z,t)}{\partial t} = \lambda(z-1)\Pi(z,t) - \mu(z-1)\frac{\partial \Pi(z,t)}{\partial z}. \tag{2.109}$$

Начальное условие (2.108) переходит в условие

$$\Pi(z,0) = z^k. \tag{2.110}$$

Кроме того, из условия нормировки получаем следующее краевое условие:

$$\Pi(1,t) = 1,\ t \geq 0. \tag{2.111}$$

Непосредственной подстановкой можно убедиться, что решением системы (2.106), (2.107) с начальным условием (2.110) и краевым условием (2.111) является функция:

$$\Pi(z,t) = \left[1 + (z-1)e^{-\mu t}\right]^k e^{\rho(z-1)(1-e^{-\mu t})}, \tag{2.112}$$

где $\rho = \frac{\lambda}{\mu}$.

Таким образом, нами доказано следующее.

Утверждение 2.18. *Производящая функция нестационарного (зависящего от t) распределения вероятностей состояний рассматриваемой СМО при начальном состоянии системы $i_0 = k$ задается формулой (2.112).*

Следствие 2.6. *Среднее число L запросов в системе в момент времени t при начальном состоянии системы $i_0 = k$ задается формулой:*

$$L = \rho + (k - \rho)e^{-\mu t}.$$

Следствие 2.7. *Стационарное распределение вероятностей рассматриваемой СМО*

$$\pi_i = \lim_{t \to \infty} P_i(t), i \geq 0$$

имеет вид:

$$\pi_i = \frac{\rho^i}{i!} e^{-\rho}, i \geq 0, \tag{2.113}$$

т.е. оно подчиняется закону Пуассона с параметром ρ.

Доказательство. Устремляя в (2.112) t к ∞, получаем выражение $\Pi(z) = e^{\rho(z-1)}$. Разлагая эту функцию в ряд Маклорена, получаем (2.113).

Следствие 2.8. *Если начальное состояние СМО не зафиксировано, а выбирается случайно в соответствии с некоторым распределением вероятностей и в качестве этого распределения взято стационарное распределение (2.113), то*

$$\Pi(z,t) = e^{\rho(z-1)}$$

для любого значения t.

Приведем теперь два обобщения Утверждения 2.18.

Пусть $Q_k(i,j,t)$ – вероятность того, что в момент t в системе находится i запросов и j запросов уже обслужилось за интервал времени $(0,t)$ при условии, что в момент времени 0 в системе находилось k запросов. Обозначим

$$Q_k(z,y,t) = \sum_{j=0}^{\infty} \sum_{i=\max(0,k-j)}^{\infty} Q_k(i,j,t) z^i y^j, |z| < 1, |y| < 1.$$

Утверждение 2.19. *Производящая функция $Q_k(z,y,t)$ определяется формулой:*

$$Q_k(z,y,t) = \left[y + (z-y)e^{-\mu t} \right]^k e^{\lambda t(y-1) + \rho(z-y)(1-e^{-\mu t})}.$$

Пусть теперь время обслуживания запросов имеет произвольное распределение $B(t)$ с конечным средним значением b_1.

Утверждение 2.20. *Производящая функция нестационарного (зависящего от t) распределения вероятностей состояний СМО $M/G/\infty$ при начальном состоянии системы $i_0 = k$ задается формулой:*

$$\Pi(z,t) = \left[1 + (z-1)(1 - \tilde{B}(t)) \right]^k e^{\rho(z-1)\tilde{B}(t)},$$

где $\rho = \lambda b_1$, а функция $\tilde{B}(t)$ имеет вид (2.95).

Вывод этой формулы принадлежит Риордану и Бенешу.

В заключение исследуем более общую бесконечно линейную СМО, у которой интенсивность λ_n входящего потока зависит от числа n запросов, находящихся в системе, $n \geq 0$. Будем считать, что $\lambda_0 = \lambda, \lambda_n = n\lambda, n \geq 1$, а функция распределения времени обслуживания имеет плотность $b(x)$,

определяемую как: $B(x) = \int\limits_0^x b(u)du$. При поступлении нового запроса, когда в системе уже обслуживается n запросов, производится перенумерация занятых каналов, причем новый запрос будет обслуживаться в канале с номером $i, i = 1, \ldots, n+1$ с вероятностью $\frac{1}{n+1}$. При завершении обслуживания запроса в канале с номером j происходит перенумерация каналов с номерами, большими j, путем уменьшения их номера на единицу.

Пусть i_t есть число запросов в системе в момент времени t, а $\xi_t^{(k)}$ есть остаточное время обслуживания в канале с номером $k, k = 1, \vdots i_t$. Процесс $\{i_t, \xi_t^{(1)}, \ldots, \xi_t^{(i_t)}, t \geq 0\}$ является марковским.

Обозначим

$$q_0 = \lim_{t\to\infty} P\{i_t = 0\},$$

$$Q_i(x_1, \ldots, x_i) = \lim_{t\to\infty} P\{i_t = i, \xi_t^{(1)} < x_1, \ldots, \xi_t^{(i)} < x_i\}, x_l > 0, l = 1, \ldots, i, i \geq 1.$$

Через $q_i(x_1, \ldots, x_i)$ обозначим плотность распределения $Q_i(x_1, \ldots, x_i)$:

$$Q_i(x_1, \ldots, x_i) = \int\limits_0^{x_1} \cdots \int\limits_0^{x_i} q_i(u_1, \ldots, u_i)du_1 \ldots du_i.$$

Используя "Δt - метод," можно получить следующую систему уравнений для вероятности q_0 и плотностей $q_i(x_1, \ldots, x_i)$:

$$\lambda_0 q_0 = q_1(0),$$

$$-\left(\frac{\partial q_i(x_1, \ldots, x_i)}{\partial x_1} + \cdots + \frac{\partial q_i(x_1, \ldots, x_i)}{\partial x_i}\right) = -\lambda_i q_i(x_1, \ldots, x_i) +$$

$$+\sum_{j=1}^{i+1} q_{i+1}(x_1, \ldots, x_{j-1}, 0, x_{j+1}, \ldots, x_{i+1}) +$$

$$+\frac{\lambda_{i-1}}{i}\sum_{j=1}^{i} q_{i-1}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_i)b(x_j), i \geq 1.$$

Непосредственной подстановкой можно убедиться, что решение этой системы имеет вид:

$$q_i(x_1, \ldots, x_i) = \frac{\lambda_0\lambda_1 \ldots \lambda_{i-1}}{i!}\prod_{j=1}^{i}(1 - B(x_j))q_0 =$$

$$= \frac{\lambda^i}{i} \prod_{j=1}^{i} (1 - B(x_j)) q_0, i \geq 1.$$

Стационарные вероятности $q_i = \lim_{t \to \infty} P\{i_t = i\}$ системы определяются как:

$$q_i = \int_0^\infty \cdots \int_0^\infty q_i(x_1, \ldots, x_i) dx_1 \ldots dx_i = \frac{\rho^i}{i} q_0, i \geq 1,$$

$\rho = \lambda b_1$, а вероятность q_0 находится из условия нормировки.

2.7.4 Система $M/G/1$ с дисциплиной равномерного распределения процессора и дисциплиной LIFO с прерыванием обслуживания

До сих пор мы рассматривали только дисциплины обслуживания FIFO и LIFO, при которых в момент начала обслуживания из очереди выбирается запрос, который пришел в систему первым или последним соответственно.

Важной дисциплиной обслуживания, также используемой в реальных системах, является дисциплина равномерного распределения процессора PS (Processor Sharing), при которой прибор обслуживает одновременно и с одинаковой скоростью (обратно пропорциональной числу запросов) все находящиеся в системе запросы.

Поскольку скорость обслуживания запроса может изменяться, при использовании дисциплины PS полезным является введение понятия длины и остаточной длины запроса. При этом функция $B(x)$ понимается как функция распределения длины произвольного запроса. При нахождении в системе i запросов за время Δt остаточная длина каждого из них уменьшается на величину $\frac{\Delta t}{i}$. При достижении остаточной длиной значения 0 соответствующий запрос покидает систему.

Пусть i_t есть число запросов в системе в момент времени t. Для исследования этого процесса будем использовать метод случайной замены времени. Кроме реального времени t будем рассматривать "фиктивное" время τ, связанное с реальным соотношением $i_t d\tau = dt$. Это означает, что если в системе находится i запросов, то фиктивное время течет в i раз быстрее, чем реальное. Если система пуста или в ней находится один запрос, то фиктивное время совпадает с реальным.

Несложно видеть, что исходная система в фиктивном времени эквивалентна бесконечно линейной СМО $M/G/\infty$, в которой запросы обслуживаются с постоянной скоростью $i\frac{1}{i} = 1$ в каждом из i занятых в данный момент каналов. При этом функция распределения длины запроса $B(t)$ совпадает с функцией распределения времени обслуживания запроса.

При переходе к фиктивному времени следует пересчитать также интенсивность входящего потока. Поскольку фиктивное время течет быстрее, то интенсивность λ_i потока в системе $M/G/\infty$ при нахождении в ней i запросов определяется как: $\lambda_0 = \lambda, \lambda_i = i\lambda, i \geq 1$.

Очевидно, что полученная бесконечно линейная СМО в фиктивном времени полностью совпадает со СМО, изученной в конце предыдущего подпункта. Поэтому стационарное распределение числа запросов в ней задается формулой (2.114).

Поскольку (для эргодических процессов) стационарная вероятность состояния процесса может трактоваться как средняя доля времени, проводимого процессом в данном состоянии, а реальное время течет в i раз медленнее, чем фиктивное, то стационарная вероятность p_i наличия i запросов в системе $M/G/1$ с дисциплиной PS определяется как: $p_0 = q_0, p_i = iq_i, i \geq 1$, где вероятности q_i заданы формулой (2.114).

Отсюда окончательно получаем, что:

$$p_i = (1-\rho)\rho^i, i \geq 0, \rho = \lambda b_1,$$

т.е. стационарное распределение числа запросов в системе $M/G/1$ с дисциплиной PS инвариантно относительно распределения времени обслуживания и является геометрическим. Условием существования этого распределения является выполнение неравенства $\rho < 1$.

Другой известной дисциплиной обслуживания в системе $M/G/1$ является дисциплина LIFO с прерыванием. При этой дисциплине прибывающий в непустую систему запрос вытесняет запрос, находящийся на приборе, во главу очереди, организованной по принципу стека, в которой размещаются прерванные запросы для ожидания дальнейшего дообслуживания.

С использованием метода введения дополнительных переменных можно показать, что и в случае такой дисциплины распределение числа запросов в системе является геометрическим с параметром ρ. Выходящий из системы поток при обеих дисциплинах является простейшим независимо от вида распределения $B(x)$ (см., например, [90]).

При дисциплине LIFO с прерыванием время ожидания начала обслу-

живания – нулевое. Поэтому интерес представляет нахождение распределения времени пребывания запроса в системе. Анализ поведения системы позволяет легко понять, что преобразование Лапласа – Стилтьеса $v(s)$ этого распределения совпадает с преобразованием Лапласа – Стилтьеса $\pi(s)$ распределения периода занятости системы $M/G/1$, не зависящим от дисциплины обслуживания и задаваемым уравнением (2.82).

2.8 Приоритетные однолинейные системы массового обслуживания

Во всех рассмотренных выше СМО предполагалось, что все запросы, поступающие в систему – однородные, т.е. они имеют один и тот же закон распределения времени обслуживания и обслуживаются в системе согласно общей дисциплине выбора из очереди. Однако во многих реальных системах запросы, поступающие в систему, неоднородны как по распределению времени обслуживания, так и по их ценности для системы и, следовательно, праву претендовать на первоочередное обслуживание в момент освобождения прибора. Такие модели исследуются в рамках теории приоритетных СМО. Эта теория довольно хорошо развита и ее изложению посвящено немало монографий (см., например, [74], [75], [70] и т.д.). Здесь мы ограничимся кратким описанием приоритетных систем и рассмотрим одну систему.

Рассмотрим однолинейную СМО с ожиданием. На вход системы поступают r независимых простейших потоков. k-й поток имеет интенсивность $\lambda_k, k = 1, \ldots, r$. Будем обозначать $\Lambda_k = \sum_{i=1}^{k} \lambda_i$.

Времена обслуживания запросов из k-го потока характеризуются функцией распределения $B_k(t)$ с преобразованием Лапласа – Стилтьеса $\beta_k(s)$ и конечными начальными моментами $b_m^{(k)} = \int_0^\infty t^m dB_k(t), m = 1, 2, k = 1, \ldots, r$. Запросы из k-го потока назовем запросами приоритета k.

Считаем, что запросы из i-го потока более приоритетны, чем запросы из j-го потока, если $i < j$. Приоритетность проявляется в том, что в момент окончания обслуживания следующим на обслуживание выбирается из очереди запрос, имеющий максимальный приоритет. Запросы, имеющие один и тот же приоритет, выбираются согласно установленной дисциплине обслуживания, например согласно дисциплине FIFO.

Рассматриваются различные варианты поведения системы в ситуации, когда во время обслуживания запроса некоторого приоритета в систему поступает запрос более высокого приоритета. Система называется СМО с *относительным приоритетом*, если поступление такого запроса не прерывает обслуживание запроса. Если же такое прерывание происходит, то система называется СМО с *абсолютным приоритетом*. В этом случае, однако, требуется уточнить дальнейшее поведение запроса, обслуживание которого оказалось прерванным. Различают следующие варианты: прерванный запрос уходит из системы и теряется; прерванный запрос возвращается в очередь и продолжает обслуживание с места прерывания после ухода из системы всех запросов, имеющих более высокий приоритет; прерванный запрос возвращается в очередь и начинает обслуживание заново после ухода из системы всех запросов, имеющих более высокий приоритет. Прерванный запрос обслуживается прибором после ухода из системы всех запросов, имеющих более высокий приоритет, в течение времени, имеющего прежнее или некоторое другое распределение. Возможен вариант, когда требуемое время обслуживания в последующих попытках идентично времени, которое требовалось для полного обслуживания данного запроса в первой попытке.

Таким образом, имеется достаточно большое число вариантов поведения системы с приоритетом, с которыми можно ознакомиться в вышеупомянутых книгах. Общим в анализе всех систем с приоритетами является использование понятия периода занятости системы запросами приоритета k и выше. При этом основным методом исследования этих систем является метод введения дополнительного события, кратко описанный в параграфе 2.6.

Проиллюстрируем особенности нахождения характеристик систем с приоритетами на примере системы, описанной в начале параграфа. Будем считать, что это система с относительным приоритетом и найдем стационарное распределения времени $w_k(t)$ ожидания запроса приоритета $k, k = 1, \ldots, r$ если бы он поступил в систему в момент времени t (так называемого виртуального времени ожидания), для системы с относительными приоритетами.

Обозначим

$$W_k(x) = \lim_{t\to\infty} P\{w_k(t) < x\}, x > 0.$$

Условием существования этих пределов является выполнение неравенства

$$\rho < 1, \tag{2.115}$$

где величина ρ вычисляется по формуле: $\rho = \sum_{k=1}^{r} \lambda_k b_1^{(k)}$.

Обозначим также $w_k(s) = \int_0^\infty e^{-sx} dW_k(x)$.

Утверждение 2.21. *Преобразование Лапласа – Стилтьеса $w_k(s)$ стационарного распределения виртуального времени ожидания запроса приоритета k определяется следующим образом:*

$$w_k(s) = \frac{(1-\rho)\mu_k(s) + \sum_{j=k+1}^{r} \lambda_j (1 - \beta_j(\mu_k(s)))}{s - \lambda_k + \lambda_k \beta_k(\mu_k(s))}, \tag{2.116}$$

где функции $\mu_k(s)$ задаются формулой:

$$\mu_k(s) = s + \Lambda_{k-1} - \Lambda_{k-1}\pi_{k-1}(s), \tag{2.117}$$

а функции $\pi_l(s), l = 1, \ldots, r$ находятся как решения функциональных уравнений:

$$\Lambda_l \pi_l(s) = \sum_{i=1}^{l} \lambda_i \beta_i(s + \Lambda_l - \Lambda_l \pi_l(s)), l = 1, \ldots, r. \tag{2.118}$$

Доказательство. Заметим, что функция $\pi_l(s)$ представляет собой преобразование Лапласа – Стилтьеса распределения длины периода занятости системы запросами приоритета l и выше (т.е. интервала времени с момента поступления в пустую систему запроса приоритета l и выше и до первого после этого момента, когда система окажется свободной от присутствия запросов приоритета l и выше). Доказательство того, что функция $\pi_l(s)$ удовлетворяет уравнению (2.118), почти дословно повторяет доказательство Утверждения 2.13. Отметим лишь, что величина $\frac{\lambda_i}{\Lambda_l}$ есть вероятность того, что период занятости системы запросами приоритета l и выше начинается с прихода запроса приоритета $i, i = 1, \ldots, l$, а величина $\beta_i(s + \Lambda_l - \Lambda_l \pi_l(s))$ трактуется как вероятность ненаступления катастрофы и поступления запросов приоритета l и выше, за время обслуживания запроса приоритета i, начавшего данный период занятости.

Сначала вместо процесса $w_k(t),\ t \geq 0$, рассмотрим существенно более простой вспомогательный процесс $\bar{w}_k(t),\ t \geq 0$, – время, в течение которого

ожидал бы начала обслуживания запрос приоритета $k, k = 1, \ldots, r$ если бы он поступил в систему в момент времени t и после этого в систему не поступало запросов более высокого приоритета.

Пусть $\bar{w}_k(s,t) = Me^{-s\bar{w}_k(t)}$ – преобразование Лапласа – Стилтьеса распределения случайной величины $\bar{w}_k(t)$. Покажем, что функция $\bar{w}_k(s,t)$ определяется следующим образом:

$$\bar{w}_k(s,t) = e^{\varphi_k(s)t}\Bigg[1 - s\int\limits_0^t P_0(x)e^{-\varphi_k(s)x}dx -$$

$$- \sum_{j=k+1}^{r}(1-\beta_j(s))\int\limits_0^t \bar{P}_j(x)e^{-\varphi_k(s)x}dx\Bigg], \qquad (2.119)$$

где

$$\varphi_k(s) = s - \sum_{i=1}^{k}\lambda_i(1-\beta_i(s)),$$

$P_0(x)$ – вероятность того, что система пуста в момент времени x, а $\bar{P}_j(x)dx$ – вероятность того, что в интервале $(x, x+dx)$ началось обслуживание запроса приоритета j.

Для доказательства (2.119) применим метод введения дополнительного события. Пусть независимо от работы системы поступает простейший поток катастроф интенсивности s. Каждый запрос назовем "плохим," если во время его обслуживания поступает катастрофа, и "хорошим" – в противном случае. Как следует из утверждений 2.5 и 2.6, поток плохих запросов приоритета k и выше является простейшим с интенсивностью $\sum\limits_{i=1}^{k}\lambda_i(1-\beta_i(s))$.

Введем событие $A(s,t)$ - за время t в систему не поступали плохие запросы приоритета k и выше. В силу утверждения 1 вероятность этого события подсчитывается как:

$$P(A(s,t)) = e^{-\sum\limits_{i=1}^{k}\lambda_i(1-\beta_i(s))t}.$$

Подсчитаем эту вероятность иначе. Событие $A(s,t)$ является объединением трех несовместных событий $A_l(s,t), l = 1,2,3$.

Событие $A_1(s,t)$ состоит в том, что катастрофы не поступили ни за время t, ни за время $\bar{w}_k(t)$. При этом, естественно, за время t в систему

поступали только хорошие запросы приоритета k и выше. Вероятность события $A_1(s,t)$, очевидно, равна $e^{-st}\bar{w}_k(s,t)$.

Событие $A_2(s,t)$ состоит в том, что катастрофа поступила в интервале $(x, x+dx), x < t$, но в момент поступления система была пуста, а за время $t-x$ не поступило плохих запросов приоритета k и выше. Вероятность события $A_2(s,t)$ вычисляется как:

$$P(A_2(s,t)) = \int_0^t P_0(x) e^{-\sum_{i=1}^{k} \lambda_i (1-\beta_i(s))(t-x)} s e^{-sx} dx.$$

Событие $A_3(s,t)$ состоит в том, что катастрофа поступила в интервале $(x, x+dx), 0 \le x < t$, но в момент ее поступления в системе обслуживался запрос приоритета ниже k, который начал обслуживаться в интервале $(u, u+du), 0 \le u < x$, а за время $t-u$ не поступило плохих запросов приоритета k и выше. Вероятность события $A_3(s,t)$ определяется следующим образом:

$$P(A_3(s,t)) = \sum_{j=k+1}^{r} \int_0^t \bar{P}_j(u) e^{-\sum_{i=1}^{k} \lambda_i (1-\beta_i(s))(t-u)} du \int_u^\infty (1 - B_j(x-u)) s e^{-sx} dx.$$

Поскольку событие $A(s,t)$ есть сумма трех несовместных событий, то его вероятность есть сумма вероятностей этих событий. Поэтому

$$P(A(s,t)) = e^{-st}\bar{w}_k(s,t) + \int_0^t P_0(x) e^{-\sum_{i=1}^{k} \lambda_i (1-\beta_i(s))(t-x)} s e^{-sx} dx +$$

$$+ \sum_{j=k+1}^{r} \int_0^t \bar{P}_j(u) e^{-\sum_{i=1}^{k} \lambda_i (1-\beta_i(s))(t-u)} du \int_u^\infty (1 - B_j(x-u)) s e^{-sx} dx.$$

Приравнивая два полученных выражения для вероятности $P(A(s,t))$ и умножая обе части равенства на e^{st}, после несложных преобразований получаем (2.119)

Очевидно, что для того, чтобы за время $w_k(t)$ ожидания запроса, поступившего в момент t, не поступило катастрофы, необходимо и достаточно, чтобы за время $\bar{w}_k(t)$ не поступило катастроф и запросов приоритета $k-1$ и выше, таких, что за периоды занятости (запросами приоритета $k-1$ и выше), порожденные ими, наступает катастрофа. Из этих рассуждениий и

вероятностной трактовки преобразования Лапласа – Стилтьеса получаем формулу, дающую связь преобразований $\bar{w}_k(s,t)$ и $w_k(s,t) = Me^{-sw_k(t)}$ в очевидной форме:

$$w_k(s,t) = \bar{w}_k(s + \Lambda_{k-1} + \Lambda_{k-1}\pi_{k-1}(s), t) = \bar{w}_k(\mu_k(s), t). \qquad (2.120)$$

Переходим в (2.120) с учетом (2.119) к пределу при $t \to \infty$.

Можно показать, что вероятность $P_0(x)$ удовлетворяет соотношению

$$\int_0^\infty e^{-sx} P_0(x)dx = [s + \Lambda_1 - \Lambda_1\pi_1(s)]^{-1},$$

откуда в силу свойства 2.4 преобразования Лапласа – Стилтьеса следует, что

$$p_0 = \lim_{t\to\infty} P_0(t) = 1 - \rho.$$

Кроме того, устремляя в (2.119) t к ∞, с учетом ограниченности функции $\bar{w}_k(s,t)$ для всех действительных $s > 0$ и $t \geq 0$, можно получить рекуррентную процедуру для вычисления интегралов $\int_0^\infty e^{-sx}\bar{P}_j(x)dx$ в виде:

$$1 = s\int_0^\infty P_0(x)e^{-\varphi_k(s)x}dx +$$

$$-\sum_{j=k+1}^{r}(1-\beta_j(s))\int_0^\infty \bar{P}_j(x)e^{-\varphi_k(s)x}dx, k = 0,\ldots,r-1.$$

Из этой процедуры можно получить рекуррентные формулы и для величин

$$\bar{p}_j = \lim_{t\to\infty}\bar{P}_j(t) = \lim_{s\to 0} s\int_0^\infty \bar{P}_j(x)e^{-sx}dx, j = 2,\ldots,r.$$

С учетом полученных выражений для величин $p_0, \bar{p}_j, j = 2,\ldots,r$, в результате предельного перехода в (2.120) получаем доказываемое соотношение (2.116). Утверждение 2.21 доказано.

Обозначим $W_1^{(k)}$ математическое ожидание виртуального времени ожидания в системе запроса приоритета k.

Следствие 2.9. *Величины* $W_1^{(k)}, k = 1, \ldots, r$ *высчитываются следующим образом:*

$$W_1^{(k)} = \frac{\sum\limits_{i=1}^{r} \lambda_i b_2^{(i)}}{2\left(1 - \sum\limits_{i=1}^{k} \lambda_i b_1^{(i)}\right)\left(1 - \sum\limits_{i=1}^{k-1} \lambda_i b_1^{(i)}\right)}. \tag{2.121}$$

Доказательство следует из (2.116) с использованием свойства 2.5 преобразования Лапласа – Стилтьеса.

Отметим, что из систем с абсолютным приоритетом наиболее легко исследуется система с двумя потоками запросов и дообслуживанием прерванных запросов. В этой системе характеристики процесса обслуживания приоритетного потока совершенно не зависят от наличия второго потока и вычисляются по обычным формулам для системы $M/G/1$ при интенсивности входящего потока, равной λ_1, и распределении $B_1(x)$ времени обслуживания запроса. В свою очередь, характеристики процесса обслуживания неприоритетного потока вычисляются по формулам для ненадежной системы $M/G/1$ при интенсивности входящего потока, равной λ_2, и распределении $B_2(x)$ времени обслуживания запроса. Приход приоритетного запроса здесь трактуется как поломка прибора, а время ремонта прибора распределено как период занятости системы, обслуживающей приоритетные запросы.

2.9 Многофазные системы массового обслуживания

В предыдущих параграфах были рассмотрены модели СМО, в которых обслуживание запросов производится одним прибором либо одним из нескольких параллельных идентичных приборов. Вместе с тем, процесс обработки запросов во многих реальных системах состоит из их последовательной обработки в нескольких обслуживающих устройствах. Системы обслуживания такого вида явились прототипом сетей массового обслуживания и получили название многофазных СМО.

Многофазные СМО принято кодировать в виде последовательности символов типа:

$$A_1/B_1/n_1/m_1 \to /B_2/n_2/m_2 \to \cdots \to /B_r/n_r/m_r.$$

Здесь символ A_1 описывает входящий поток на вход цепочки из r последовательных обслуживающих устройств, символы $B_k, n_k, m_k, k = 1, \ldots, r$ описывают, соответственно, распределение времени обслуживания, число параллельных каналов и число мест в буфере для ожидания в k-м звене цепочки. Символы $A_1, B_k, k = 1, \ldots, r$ принимают значение в тех же множествах, как и соответствующие символы в описании однофазных систем, изученных нами выше. Символ $\rightarrow$ означает переход запроса на вход следующего обслуживающего устройства по окончании его обслуживания в предыдущем. В некотором смысле этот символ является заменителем символа, задающего вид входящего потока в соответствующую систему обслуживания, поскольку входящий поток в данную СМО определяется выходящим потоком запросов из предыдущей СМО.

Если какой – либо символ m_k принимает конечное значение, то возникает вопрос о поведении многофазной СМО в ситуации, когда буфер перед соответствующей однофазной СМО уже полон, а на вход этой СМО поступает следующий запрос. Обычно рассматриваются два варианта: поступающий запрос теряется и поступающий запрос остается на приборе, где он закончил обслуживание, временно блокируя дальнейшую работу этого прибора.

Известные результаты для широкого круга многофазных СМО довольно исчерпывающе описаны в справочнике [125]. Здесь мы кратко коснемся трех простых многофазных СМО.

Первая из них – это многофазная СМО с бесконечным буфером перед первой фазой, на вход которой поступает простейший поток интенсивности λ, а время обслуживания в каждой из r фаз имеет показательное распределение с параметром μ. Предполагается, что одновременно в системе может обслуживаться только один запрос. Только по завершении прохождения запросом всей цепочки приборов на обслуживание может быть выбран следующий запрос.

Вспоминая, что эрланговская случайная величина с параметрами (μ, r) есть сумма r независимых случайных величин с параметром μ, приходим к выводу, что все характеристики рассматриваемой СМО легко получить из характеристик соответствующей однофазной СМО типа $M/E_r/1$. Анализ такой СМО можно провести с использованием так называемого метода фаз Эрланга. Соответствующие результаты можно получить также из формул для системы $M/G/1$.

Так, для производящей функции $P(z) = \sum\limits_{i=0}^{\infty} p_i z^i$ распределения $p_i, i \geq 0$, числа запросов в системе из формулы Полячека – Хинчина (см., например, формулу (2.45)) с учетом равенства: $\beta(s) = \left(\frac{\mu}{\mu+s}\right)^r$ получаем:

$$P(z) = (1-\rho)\frac{(1-z)}{1 - z(1+\frac{\lambda}{\mu}(1-z))^r}, \tag{2.122}$$

где коэффициент загрузки ρ определяется как $\rho = \frac{r\lambda}{\mu}$.

Обозначим $z_i, i = 1, \ldots, r$ - корни уравнения

$$1 - z(1 + \frac{\lambda}{\mu}(1-z))^r = 0 \tag{2.123}$$

по модулю больше единицы.

Разлагая правую часть уравнения (2.122) на простые дроби и используя свойства производящей функции, получаем:

$$p_0 = 1 - \rho, \tag{2.124}$$

$$p_i = (1-\rho)\sum_{l=1}^{r} K_l z_l^{-ir}\frac{1 - z_l^r}{1 - z_l}, i \geq 1, \tag{2.125}$$

где коэффициенты $K_m, m = 1, \ldots, r$ определяются следующим образом:

$$K_m = \prod_{l=1, l\neq m}^{r} \frac{1}{1 - \frac{z_m}{z_l}}, m = 1, \ldots, r. \tag{2.126}$$

Рассмотрим теперь следующую многофазную СМО:

$$M/M/n_1/\infty \to /M/n_2/\infty \to \cdots \to /M/n_r/\infty,$$

т.е. многофазную СМО, состоящую из r последовательных многоканальных СМО с бесконечным буфером. Пусть λ – интенсивность входящего потока, μ_k – интенсивность обслуживания в каждом из этих каналов k-й системы, $k = 1, \ldots, r$.

Обозначим $i_k(t)$ число запросов в k-й системе в момент времени t, $t \geq 0, k = 1, \ldots, r$ и $p(i_1, \ldots, i_r)$ – стационарную вероятность состояния $\{i_1, \ldots, i_r\}$ рассматриваемой СМО, т.е.

$$p(i_1, \ldots, i_r) = \lim_{t\to\infty} P\{i_1(t) = i_1, \ldots, i_r(t) = i_r\}, i_k \geq 0, k = 1, \ldots, r. \tag{2.127}$$

Можно показать, что условием существования пределов (2.127) является выполнение неравенств

$$\rho_i = \frac{\lambda}{n_i \mu_i} < 1, i = 1, \ldots, r. \tag{2.128}$$

Далее считаем эти неравенства выполненными. Используя "Δt - метод", можно получить систему дифференциальных уравнений для вероятностей $P\{i_1(t) = i_1, \ldots, i_r(t) = i_r\}$, откуда, переходя к пределу при $t \to \infty$, получим следующую систему уравнений для стационарных вероятностей $p(i_1, \ldots, i_r)$:

$$\left(\lambda + \sum_{k=1}^{r} \alpha_k(i_k)\mu_k\right) p(i_1, \ldots, i_r) = \lambda p(i_1 - 1, i_2, \ldots, i_r)(1 - \delta_{i_1,0}) + \tag{2.129}$$

$$+ \sum_{k=1}^{r-1} p(i_1, \ldots, i_k + 1, i_{k+1} - 1, i_{k+2}, \ldots, i_r)\alpha_k(i_k + 1)(1 - \delta_{i_{k+1},0})\mu_k +$$

$$+ p(i_1, \ldots, i_{r-1}, i_r + 1)\alpha_r(i_r + 1)\mu_r, i_k \geq 0, k = 1, \ldots, r,$$

где

$$\alpha_k(i) = \begin{cases} i, & 0 \leq i \leq n_k, \\ n_k, & i > n_k, \end{cases}$$

$$\delta_{i,j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

Непосредственной подстановкой несложно убедиться, что решение системы (2.129) имеет следующий вид:

$$p(i_1, \ldots, i_r) = p(0, \ldots, 0) \prod_{k=1}^{r} c_k(i_k), \tag{2.130}$$

где

$$c_k(i) = \begin{cases} \frac{(n_k \rho_k)^i}{i!}, & 0 \leq i \leq n_k, \\ \frac{n_k^{n_k}}{n_k!}\rho_k^i, & i > n_k. \end{cases}$$

Вероятность $p(0, \ldots, 0)$ находится из условия нормировки:

$$\sum_{i_1=0}^{\infty} \cdots \sum_{i_r=0}^{\infty} p(i_1, \ldots, i_r) = 1 \tag{2.131}$$

и имеет вид

$$p(0,\ldots,0)=\prod_{k=1}^{r}(\sum_{i_k=0}^{\infty}c_k(i_k))^{-1}. \tag{2.132}$$

Из (2.130), (2.131) и формул (2.96),(2.97) видим, что стационарные вероятности $p(i_1,\ldots,i_r)$ рассматриваемой СМО представимы в мультипликативном виде:

$$p(i_1,\ldots,i_r)=\prod_{k=1}^{r}\lim_{t\to\infty}P\{i_k(t)=i_k\}, \tag{2.133}$$

т.е. совместная вероятность того, что в произвольный момент времени на k-й фазе находится i_k запросов, $k=1,\ldots,r$, равна произведению вероятностей того, что i_k запросов в данный момент находится на k-й фазе независимо от числа запросов на других фазах, $k=1,\ldots,r$.

Этот факт, позволяющий рассчитывать распределение вероятностей состояний многофазной системы как произведение вероятностей состояний однофазных СМО, образующих данную СМО, следует из теоремы Берка ([105]), которая формулируется следующим образом.

Теорема. В системе с c параллельными каналами простейшим входящим потоком интенсивности λ и одинаковым для каждого канала показательным распределением времени обслуживания с параметром μ в стационарном состоянии выходящий поток является простейшим потоком интенсивности λ.

Таким образом, обслуживание простейшего потока запросов на каждой фазе многофазной СМО с показательным распределением времени обслуживания не изменяет характера потока и в результате совместное распределение вероятностей числа запросов в соответствующей многофазной системе имеет мультипликативный вид.

В заключение параграфа кратко рассмотрим многофазную СМО типа $M/G/1/\infty\to M/n/0$.

Интенсивность входящего потока обозначим λ, функцию распределения времени обслуживания – $B(t)$, интенсивность обслуживания любым прибором на второй фазе обозначим μ.

Предполагаем, что в случае занятости всех приборов на второй фазе в момент окончания обслуживания запроса на первой фазе с вероятностью $\theta, 0\le\theta\le 1$ запрос уходит из системы недообслуженным (теряется), а с дополнительной вероятностью первый прибор блокируется и не обслуживает следующий запрос, пока не освободится прибор на второй фазе. Как

крайние случаи при $\theta = 0$ мы имеем систему с блокировкой прибора, при $\theta = 1$ – систему с потерями.

Будем рассматривать двумерный процесс $\{i_{t_k}, \nu_{t_k}, k \geq 1, \}$, где t_k есть k-й момент окончания обслуживания запроса на первой фазе, $i_{t_k}, i_{t_k} \geq 0$, – число запросов на первой фазе, $\nu_{t_k}, \nu_{t_k} = 0, \ldots, n$ – число запросов на второй фазе в момент времени $t_k + 0$. Несложно видеть, что этот процесс является двумерной цепью Маркова с дискретным временем.

Обозначим одношаговые вероятности переходов этой цепи $P\{i_{t_{k+1}} = l, \nu_{t_{k+1}} = \nu^{'} | i_{t_k} = i, \nu_{t_k} = \nu\} = P\{(i, \nu) \to (l, \nu^{'})\}, i > 0$.

Введем производящую функцию

$$R_{\nu,\nu'}(z) = \sum_{l=i-1}^{\infty} P\{(i, \nu) \to (l, \nu^{'})\} z^{l-i+1}.$$

Аналогично вложенной цепи для системы $M/G/1$, изученной в параграфе 6, вероятности переходов $P\{(i, \nu) \to (l, \nu^{'})\}$ зависят от значения $l - i$, но не зависят от i и l отдельно. Это делает введенное определение производящей функции $R_{\nu,\nu'}(z)$ корректным.

Анализируя переходы двумерной цепи Маркова, можно убедиться, что производящие функции $R_{\nu,\nu'}(z)$ определяются следующим образом:

$$R_{\nu,\nu'}(z) = \begin{cases} \Gamma(z, n, n)\left(\theta + (1-\theta)\frac{n\mu}{n\mu + \lambda(1-z)}\right), & \nu = n, \nu^{'} = n, \\ \Gamma(z, \nu, \nu^{'} - 1) & , \nu^{'} \leq \nu \leq n, \nu^{'} \neq n, \end{cases}$$

где

$$\Gamma(z, \nu, \nu^{'}) = \int_0^{\infty} e^{-\lambda(1-z)t} C_{\nu}^{\nu^{'}} e^{-\mu\nu^{'}t} (1 - e^{-\mu t})^{\nu - \nu^{'}} dB(t),$$

$$0 \leq \nu^{'} \leq \nu \leq n.$$

Составим из производящих функций $R_{\nu,\nu'}(z)$ матричную производящую функцию $R(z)$ и введем в рассмотрение матрицу Δ, состоящую из элементов

$$\Delta_{\nu,\nu'} = \int_0^{\infty} \lambda e^{-\lambda t} C_{\nu}^{\nu^{'}} e^{-\mu\nu^{'}t} (1 - e^{-\mu t})^{\nu - \nu^{'}} dt,$$

$$0 \leq \nu^{'} \leq \nu \leq n.$$

Матрица Δ характеризует вероятности переходов числа занятых каналов на второй фазе системы за время, когда прибор на первой фазе простаивает, ожидая прихода запроса.

Обозначим

$$\pi(i,\nu) = \lim_{k\to\infty} P\{i_{t_k} = i, \nu_{t_k} = \nu\}, \tag{2.134}$$

$$\vec{\pi}(i) = (\pi(i,0), \pi(i,1), \ldots, \pi(i,n)),$$

$$\vec{\Pi}(z) = \sum_{i=0}^{\infty} \vec{\pi}(i) z^i, |z| < 1.$$

Утверждение 2.22. *Векторная производящая функция* $\vec{\Pi}(z)$ *удовлетворяет матричному функциональному уравнению:*

$$\vec{\Pi}(z)(R(z) - I) = \vec{\Pi}(0)(I - \Delta z)R(z). \tag{2.135}$$

Здесь I *– тождественная матрица.*

Условия существования пределов (2.134) и алгоритмы решения уравнения (2.135) можно найти, например, в [76], [16].

Распределение вероятностей состояний системы в произвольные моменты времени можно найти, используя теорию процессов марковского восстановления по аналогии с тем, как это сделано в подпункте 2.6.1.

ГЛАВА 3
МЕТОДЫ ИССЛЕДОВАНИЯ СМО С КОРРЕЛИРОВАННЫМИ ПОТОКАМИ

3.1 МАРКОВСКИЙ ВХОДНОЙ ПОТОК (*BMAP*). РАСПРЕДЕЛЕНИЕ ФАЗОВОГО ТИПА

3.1.1 Определение группового марковского входного потока

Поступление запросов в $BMAP$-потоке происходит под управлением некоторой неприводимой ЦМ ν_t, $t \geq 0$, с непрерывным временем и конечным пространством состояний $\{0, ..., W\}$. Время пребывания цепи ν_t в некотором состоянии ν имеет показательное распределение с параметром $\lambda_\nu, \nu = \overline{0, W}$. После того как время пребывания процесса в этом состоянии истекло, с вероятностью $p_k(\nu, \nu')$ процесс ν_t переходит (перескакивает) в некоторое состояние ν' и генерируется группа из k запросов, $k \geq 0$. При этом $p_0(\nu, \nu) = 0$ и

$$\sum_{\nu'=0\,\nu' \neq \nu}^{W} p_0(\nu, \nu') + \sum_{k=1}^{\infty} \sum_{\nu'=0}^{W} p_k(\nu, \nu') = 1,\ \nu = \overline{0, W}.$$

3.1.2 Матричная считающая функция потока

Главной характеристикой любого стационарного случайного потока запросов является считающая функция потока $p(n, t)$ – вероятность того, что за время t поступит n запросов. В случае стационарного пуассоновского потока считающая функция потока имеет вид:

$$p(n, t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, n \geq 0.$$

Для $BMAP$-потока невозможно определить считающую функцию аналогично – как безусловную вероятность, поскольку число запросов, поступивших за интервал времени длиной t, зависит от состояния управляющего процесса ν_t, $t \geq 0$, в момент начала этого интервала. Но можно вычислить

вероятности

$$P_{\nu,\nu'}(n,t) =$$

$$= P\{\text{за время } t \text{ поступит } n \text{ запросов и } \nu_t = \nu' \text{ при условии } \nu_0 = \nu\}.$$

Выведем систему уравнений для этих вероятностей. Будем использовать для этого широко известный Δt-метод, заключающийся в выводе уравнений для распределения вероятностей состояний некоторого случайного процесса в момент времени t путем вычисления распределения вероятностей состояний этого процесса в момент $t + \Delta t$ через распределение вероятностей состояний этого процесса в момент t и вероятности его переходов за малый промежуток времени $(t, t + \Delta t)$. В результате очевидным образом получаем следующую систему разностных уравнений:

$$P_{\nu,\nu'}(n, t + \Delta t) = P_{\nu,\nu'}(n,t)e^{-\lambda_{\nu'}\Delta t} + \sum_{k=0}^{n}\sum_{r=0}^{W} P_{\nu,r}(k,t)(1 - e^{-\lambda_r \Delta t})p_{n-k}(r,\nu') +$$

$$+o(\Delta t),\ n \geq 0, \nu, \nu' = \overline{0, W}.$$

Стандартным образом приведем эту систему к системе дифференциальных уравнений:

$$P'_{\nu,\nu'}(n,t) = -\lambda_{\nu'}P_{\nu,\nu'}(n,t) + \sum_{k=0}^{n}\sum_{r=0}^{W} P_{\nu,r}(k,t)\lambda_r p_{n-k}(r,\nu'),\ n \geq 0, \nu, \nu' = \overline{0, W}. \tag{3.1}$$

Введем в рассмотрение квадратную матрицу $P(n,t) = (P_{\nu,\nu'}(n,t))_{\nu,\nu'=\overline{0,W}}$ порядка $\bar{W}$, где $\bar{W} = W + 1$. Нетрудно видеть, что система уравнений (3.1) может быть переписана в виде

$$P'(n,t) = \sum_{k=0}^{n} P(k,t)D_{n-k},\ n \geq 0. \tag{3.2}$$

Здесь D_k – квадратные матрицы порядка $\bar{W}$, элементы которых определяются следующим образом:

$$(D_0)_{\nu,\nu'} = \begin{cases} -\lambda_\nu, & \nu = \nu', \\ \lambda_\nu p_0(\nu,\nu'), & \nu \neq \nu', \end{cases}$$

$$(D_k)_{\nu,\nu'} = \lambda_\nu p_k(\nu,\nu'), k \geq 1,$$

и могут быть интерпретированы как мгновенные интенсивности $BMAP$-потока вследствие того, что величина $\lambda_\nu p_k(\nu,\nu')\Delta t + o(\Delta t)$ есть вероятность того, что за время Δt управляющий процесс $BMAP$-а ν_t перейдет

из состояния ν в состояние ν' и при этом сгенерируется группа из k запросов, а величина $1 - \lambda_\nu \Delta t + o(\Delta t)$ есть вероятность того, что за время Δt процесс ν_t не совершит скачок и не поступят запросы при условии, что в начале интервала Δt процесс ν_t находился в состоянии ν.

Система дифференциальных уравнений (3.2) состоит из бесконечного числа матричных уравнений. Для ее решения введем в рассмотрение матричную ПФ

$$P(z,t) = \sum_{n=0}^{\infty} P(n,t) z^n, \; |z| \le 1.$$

Умножая уравнения системы (3.2) на соответствующие степени z и суммируя, получаем матричное дифференциальное уравнение:

$$P'(z,t) = P(z,t) D(z), \tag{3.3}$$

где

$$D(z) = \sum_{k=0}^{\infty} D_k z^k.$$

Очевидно, что решение линейного матричного дифференциального уравнения (3.3) имеет следующий вид:

$$P(z,t) = C(z) e^{D(z)t},$$

где $C(z)$ – некоторая матрица, не зависящая от t.

Учитывая очевидное начальное условие

$$P(z,0) = P(0,0) = I,$$

получаем, что $C(z) = I$ для любого z, где I – единичная матрица. Следовательно,

$$P(z,t) = e^{D(z)t}. \tag{3.4}$$

Отметим, что для стационарного пуассоновского потока ПФ $P(z,t) = e^{\lambda(z-1)t}$, что согласуется с тем фактом, что для этого потока $D(z) = \lambda(z-1)$.

Из (3.4) следует, в частности, что матрица $P(0,t) = \left(P_{\nu,\nu'}(0,t)\right)_{\nu,\nu'=\overline{0,W}}$ вероятностей поступления 0 запросов за время t имеет вид

$$P(0,t) = e^{D_0 t}. \tag{3.5}$$

Матрицы $P(n,t)$, $n \ge 1$, в принципе, могут быть вычислены через матричную ПФ $P(z,t)$, имеющую вид (3.4), следующим образом:

$$P(n,t) = \frac{1}{n!} \frac{\partial^n P(z,t)}{\partial z^n}\bigg|_{z=0}, \; n \ge 0.$$

Однако вычисление матриц $P(n,t)$ в аналитическом виде по этой формуле возможно только в редких случаях.

В общем случае они вычисляются с помощью процедуры, основанной на идее униформизации марковского процесса, которая состоит в следующем.

Если H – инфинитезимальный генератор ЦМ с непрерывным временем, то справедливо представление:

$$e^{Ht} = e^{ht(L-I)} = e^{-ht}e^{htL} = \sum_{j=0}^{\infty} e^{-ht}\frac{(ht)^j}{j!}L^j, \tag{3.6}$$

где

$$h = \max_{i=\overline{0,W}}(-H)_{ii},\ L = I + h^{-1}H.$$

Матрица L является матрицей одношаговых переходных вероятностей некоторой вспомогательной ЦМ с дискретным временем. Весьма полезным свойством представления (3.6) является мультипликативность членов суммы, представление их в виде произведения скалярной функции от аргумента t и матрицы, не зависящей от t.

Нетрудно видеть, что аналогичное разложение, с поправкой на то, что L – субстохастическая матрица, может быть применено к матричной экспоненте в случае, когда H– матрица, у которой недиагональные элементы неотрицательны, диагональные – отрицательны, и суммы по строкам – отрицательны или равны нулю. В частности, такое разложение справедливо для матричной экспоненты $e^{D_0 t}$.

Обозначим $\tilde{\theta} = \max\limits_{i=\overline{0,W}}(-D_0)_{ii}$. Тогда $P(0,t) = e^{D_0 t}$ можно представить в виде

$$P(0,t) = \sum_{j=0}^{\infty} e^{-\tilde{\theta}t}\frac{(\tilde{\theta}t)^j}{j!}K_0^{(j)},$$

где

$$K_0^{(j)} = (I + \tilde{\theta}^{-1}D_0)^j.$$

По аналогии, будем искать остальные матрицы $P(n,t),\ n \geq 1$, в виде

$$P(n,t) = \sum_{j=0}^{\infty} e^{-\tilde{\theta}t}\frac{(\tilde{\theta}t)^j}{j!}K_n^{(j)},\ n \geq 1, \tag{3.7}$$

где $K_n^{(j)},\ n \geq 1, j \geq 0$, – некоторые матрицы.

Подставляя выражение для $P(0,t)$ и выражения (3.7) в систему дифференциальных уравнений (3.2) и приравнивая коэффициенты при одинаковых степенях t, убеждаемся, что представление (3.7) действительно справедливо, если матрицы $K_n^{(j)},\ n \geq 0, j \geq 0$, удовлетворяют следующей системе рекуррентных соотношений:

$$K_0^{(0)} = I,\ K_n^{(0)} = O,\ n \geq 1,$$
$$K_0^{(j+1)} = K_0^{(j)}(I + \tilde{\theta}^{-1}D_0), \tag{3.8}$$
$$K_n^{(j+1)} = \tilde{\theta}^{-1}\sum_{i=0}^{n-1} K_i^{(j)}D_{n-i} + K_n^{(j)}(I + \tilde{\theta}^{-1}D_0),\ n \geq 1,\ j \geq 0.$$

Формулы (3.7) и (3.8) задают численно устойчивую процедуру для вычисления матриц $P(n,t),\ n \geq 1$. Усечение бесконечной суммы в (3.7) можно произвести после того, как член суммы окажется по норме меньше наперед заданного малого положительного числа. Потенциально полезным при выборе порога усечения является использование формулы

$$K(z,y) = \sum_{n=0}^{\infty}\sum_{j=0}^{\infty} K_n^{(j)} y^j z^n = \Big(I - y(I + \tilde{\theta}^{-1}D(z))\Big)^{-1}.$$

3.1.3 Некоторые свойства и интегральные характеристики $BMAP$-потока

Из определения $BMAP$-потока следует, что:

- матрица, (ν,ν')-й элемент которой есть вероятность того, что при условии, что в момент времени t состояние управляющего процесса потока было равным ν, первый запрос поступит в интервале $(t, t+dt)$ в составе группы из k запросов и состояние управляющего процесса потока в момент $t+dt$ будет равно ν', определяется следующим образом:
$$D_k dt,\ k \geq 1,\ t \geq 0.$$
- матрица, (ν,ν')-й элемент которой есть вероятность того, что при условии, что в момент времени 0 состояние управляющего процесса потока было равным ν, первый запрос поступит в интервале $(t, t+dt)$ в составе группы из k запросов и состояние управляющего процесса потока в момент после поступления будет равно ν', определяется следующим образом:
$$e^{D_0 t}D_k dt,\ k \geq 1.$$

- матрица, (ν, ν')-й элемент которой есть вероятность того, что первый запрос после момента 0 поступит в составе группы из k запросов и состояние управляющего процесса потока в момент после поступления будет равно ν' при условии, что в момент времени 0 оно было равным ν, определяется следующим образом:

$$\int_0^\infty e^{D_0 t} D_k dt = (-D_0)^{-1} D_k,\ k \geq 1.$$

  Существование интегралов, присутствующих в формуле, и обратной матрицы следует из определения функции от матриц и Следствия А1 в Приложении А.
- матрица $D(1)$ является инфинитезимальным генератором процесса ν_t, $t \geq 0$.
- вектор-строка $\boldsymbol{\theta}$ стационарного распределения управляющего процесса ν_t, $t \geq 0$, является единственным решением системы линейных алгебраических уравнений

$$\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1. \tag{3.9}$$

Назовем средней скоростью (интенсивностью) поступления запросов в $BMAP$ – потоке величину λ, задаваемую формулой

$$\lambda = \boldsymbol{\theta} D'(1)\mathbf{e}, \tag{3.10}$$

где $D'(1)$ есть производная матричной производящей функции $D(z)$ в точке $z = 1$.

Понятие средней (фундаментальной) скорости поступления запросов, введенное М. Ньютсом, можно понимать как среднее число запросов, поступающих в единицу времени. Обоснование того, что величина λ имеет смысл среднего числа запросов, поступающих в единицу времени, можно провести, используя следующую формулу:

$$\sum_{k=1}^{\infty} kP(k,t)\mathbf{e} = D'(1)t\mathbf{e} + (e^{D(1)t} - I - D(1)t)(D(1) - \mathbf{e}\boldsymbol{\theta})^{-1} D'(1)\mathbf{e}. \tag{3.11}$$

Вывод этой формулы осуществляется следующим образом:

$$\sum_{k=1}^{\infty} kP(k,t)\mathbf{e} = \left(e^{D(z)t}\right)'\Big|_{z=1}\mathbf{e} = \left(\sum_{k=0}^{\infty} \frac{(D(z)t)^k}{k!}\right)'\Big|_{z=1}\mathbf{e} =$$

$$= \sum_{k=1}^{\infty} \frac{t^k}{k!} (D(1))^{k-1} D'(1)\mathbf{e}.$$

Здесь мы проэксплуатировали тот факт, что матрица $D(1)$ является генератором и, следовательно, $D(1)\mathbf{e} = \mathbf{0}^T$.

Далее мы хотим выразить получившийся ряд снова в терминах матричной экспоненты, но сделать это непосредственно не удается, поскольку матрица $D(1)$ является вырожденной. На помощь приходит Лемма А9 Приложения А. Перефразируя эту лемму, сформулированную для стохастических матриц, в терминах генераторов, можно утверждать, что поскольку вектор $\boldsymbol{\theta}$ является решением системы уравнений (3.9), то матрица $\mathbf{e}\boldsymbol{\theta} - D(1)$ является невырожденной. Используя этот факт и систему (3.9), можно показать, что для $k \geq 2$ выполняется соотношение

$$(D(1))^{k-1} = -(\mathbf{e}\boldsymbol{\theta} - D(1))^{-1}(D(1))^k.$$

Теперь мы можем свернуть ряд и в результате получить формулу (3.11).

Умножим равенство (3.11) на вектор $\boldsymbol{\theta}$. Учитывая вероятностный смысл матриц $P(k,t)$, слева получим среднее число запросов, поступающих за время t, а справа выражение λt.

Разделив полученное соотношение на t и устремляя t к бесконечности, получаем, что среднее число запросов, поступающих в единицу времени, равно величине λ.

Аналогичным образом можно показать, что средняя скорость λ_g поступления групп запросов вычисляется как

$$\lambda_g = \boldsymbol{\theta}(D(1) - D_0)\mathbf{e} = -\boldsymbol{\theta} D_0 \mathbf{e}.$$

Используя понятия средней скорости λ_g поступления групп и средней скорости λ поступления запросов, можно привести еще несколько полезных свойств $BMAP$-потока:

- распределение вероятностей состояний управляющего процесса потока $\nu_t,\ t \geq 0$, в момент после поступления группы из k запросов задается вектором
$$\frac{\boldsymbol{\theta} D_k}{\lambda_g},\ k \geq 1.$$
- вероятность P_k того, что произвольный запрос поступит в группе размера k, вычисляется следующим образом:
$$P_k = \frac{\boldsymbol{\theta} k D_k \mathbf{e}}{\boldsymbol{\theta} \sum\limits_{l=1}^{\infty} l D_l \mathbf{e}} = k \frac{\boldsymbol{\theta} D_k \mathbf{e}}{\boldsymbol{\theta} D'(1)\mathbf{e}} = k \frac{\boldsymbol{\theta} D_k \mathbf{e}}{\lambda},\ k \geq 1.$$

- вероятность Q_k того, что произвольная поступившая в $BMAP$-потоке группа имеет размер k, вычисляется следующим образом:

$$Q_k = \frac{\boldsymbol{\theta} D_k \mathbf{e}}{\boldsymbol{\theta} \sum\limits_{l=1}^{\infty} D_l \mathbf{e}} = -\frac{\boldsymbol{\theta} D_k \mathbf{e}}{\boldsymbol{\theta} D_0 \mathbf{e}} = \frac{\boldsymbol{\theta} D_k \mathbf{e}}{\lambda_g},\ k \geq 1.$$

- распределение вероятностей состояний управляющего процесса потока сразу после момента поступления группы запросов задается вектором:

$$\boldsymbol{\theta} \frac{\sum\limits_{k=1}^{\infty} D_k}{\lambda_g} = \boldsymbol{\theta} \frac{(D(1) - D_0)}{\lambda_g} = -\boldsymbol{\theta} \frac{D_0}{\lambda_g}.$$

Средняя длина T_g интервала между моментами поступления групп запросов рассчитывается как

$$T_g = \frac{\boldsymbol{\theta}(-D_0) \int\limits_0^{\infty} e^{D_0 t} dt \mathbf{e}}{\lambda_g} = {\lambda_g}^{-1} \tag{3.12}$$

или

$$T_g = \frac{\boldsymbol{\theta}(-D_0) \int\limits_0^{\infty} t e^{D_0 t} \sum\limits_{k=1}^{\infty} D_k dt \mathbf{e}}{\lambda_g} = \frac{\boldsymbol{\theta}(-D_0)^{-1}(D(1) - D_0)\mathbf{e}}{\lambda_g} = {\lambda_g}^{-1}.$$

При выводе этих двух формул мы использовали две различные формулы для вычисления математического ожидания $M\xi$ неотрицательной случайной величины ξ, имеющей функцию распределения $F_\xi(x)$:

$$M\xi = \int\limits_0^{\infty} (1 - F_\xi(x)) dx$$

или

$$M\xi = \int\limits_0^{\infty} x dF_\xi(x),$$

а также свойства генератора:

$$D(1)\mathbf{e} = (D_0 + \sum_{k=1}^{\infty} D_k)\mathbf{e} = \mathbf{0}^T,\ \boldsymbol{\theta} D(1) = \mathbf{0}.$$

Начальный момент $T_g^{(m)}$ порядка m распределения длин интервалов между моментами поступления групп запросов рассчитывается как

$$T_g^{(m)} = \frac{\boldsymbol{\theta}(D(1)-D_0)\int\limits_0^\infty t^m e^{D_0 t}\sum\limits_{k=1}^{\infty} D_k dt\mathbf{e}}{\lambda_g} = m!\frac{\boldsymbol{\theta}(-D_0)^{-m+1}\mathbf{e}}{\lambda_g},\ m \geq 1.$$

Дисперсия v длин интервалов между моментами поступления групп запросов рассчитывается как

$$v = T_g^{(2)} - T_g^2 = \frac{2\lambda_g \boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1}{{\lambda_g}^2}. \tag{3.13}$$

Коэффициент корреляции c_{cor} длин двух последовательных интервалов между моментами поступления групп запросов определяется формулой

$$c_{cor} = ({\lambda_g}^{-1}\boldsymbol{\theta}(-D_0)^{-1}(D(1)-D_0)(-D_0)^{-1}\mathbf{e} - {\lambda_g}^{-2})/v. \tag{3.14}$$

Поясним вывод формулы (3.14). Обозначим ξ_1, ξ_2 длины двух последовательных интервалов между моментами поступления групп запросов в $BMAP$-потоке. Поскольку случайные величины ξ_1, ξ_2 имеют одинаковое математическое ожидание T_g и дисперсию v, коэффициентом корреляции c_{cor} этих случайных величин является величина $\rho(\xi_1, \xi_2)$, задаваемая формулой

$$\rho(\xi_1, \xi_2) = \frac{M\xi_1\xi_2 - T_g^2}{v}. \tag{3.15}$$

Из приведенных выше свойств $BMAP$-потока следует, что совместная плотность $f_{\xi_1,\xi_2}(x,y)$ случайных величин ξ_1, ξ_2 имеет вид:

$$f_{\xi_1,\xi_2}(x,y) = \frac{\boldsymbol{\theta}(D(1)-D_0)e^{D_0 x}(D(1)-D_0)e^{D_0 y}(D(1)-D_0)\mathbf{e}}{\lambda_g}.$$

Рассмотрим случайную величину $\eta = \xi_1\xi_2$. Очевидно, что:

$$P\{\eta > z\} = \int\limits_0^\infty \int\limits_{\frac{z}{x}}^\infty f_{\xi_1,\xi_2}(x,y)dxdy =$$

$$= \frac{\boldsymbol{\theta}(D(1)-D_0)\int\limits_0^\infty \int\limits_{\frac{z}{x}}^\infty e^{D_0 x}(D(1)-D_0)e^{D_0 y}(D(1)-D_0)dxdy\mathbf{e}}{\lambda_g} =$$

$$= \frac{\boldsymbol{\theta}(D(1)-D_0)\int\limits_0^\infty e^{D_0 x}(D(1)-D_0)e^{D_0\frac{z}{x}}dx\mathbf{e}}{\lambda_g}.$$

Поэтому

$$M\xi_1\xi_2 = M\eta = \int\limits_0^\infty P\{\eta > z\}dz =$$

$$= \frac{\boldsymbol{\theta}(D(1)-D_0)\int\limits_0^\infty e^{D_0 x}(D(1)-D_0)\int\limits_0^\infty e^{D_0\frac{z}{x}}dzdx\mathbf{e}}{\lambda_g} =$$

$$= \frac{\boldsymbol{\theta}(D(1)-D_0)\int\limits_0^\infty xe^{D_0 x}(D(1)-D_0)(-D_0)^{-1}dx\mathbf{e}}{\lambda_g} =$$

$$= \frac{\boldsymbol{\theta}(-D_0)^{-1}(D(1)-D_0)(-D_0)^{-1}\mathbf{e}}{\lambda_g}.$$

Отсюда и из (3.15) следует формула (3.14).

При анализе СМО с $BMAP$-потоком, особенно при выводе условий существования стационарного режима, возникает необходимость вычисления величины

$$\boldsymbol{\theta}\frac{d\int\limits_0^\infty e^{D(z)t}dB(t)}{dz}\mid_{z=1}\mathbf{e},$$

где $B(t)$ – некоторая функция распределения с математическим ожиданием $b_1 = \int\limits_0^\infty tdB(t)$.

Покажем, что эта величина равна λb_1. Используя определение матричной экспоненты через степенной ряд, запишем

$$\int\limits_0^\infty e^{D(z)t}dB(t) = \int\limits_0^\infty \sum_{k=0}^\infty \frac{(D(z)t)^k}{k!}dB(t).$$

Производная первого порядка от матрицы $(D(z))^j$ считается следующим образом:

$$\frac{d(D(z))^j}{dz} = \sum_{l=0}^{j-1}(D(z))^{j-l-1}\frac{dD(z)}{dz}(D(z))^l.$$

Учитывая, что $\boldsymbol{\theta}D(1) = \mathbf{0}$, $D(1)\mathbf{e} = \mathbf{0}^T$, заключаем, что производная первого порядка от матрицы $(D(z))^j$, $j \geq 1$, в точке $z = 1$, домноженная

на вектор-строку $\boldsymbol{\theta}$ слева и на вектор-столбец $\mathbf{e}$ справа, равна 0 для всех $j,\ j > 1$, и равна λ при $j = 1$, получаем, что

$$\boldsymbol{\theta}\frac{d\int\limits_0^\infty e^{D(z)t}dB(t)}{dz}|_{z=1}\mathbf{e} = \lambda b_1, \tag{3.16}$$

что и требовалось доказать.

3.1.4 Частные случаи $BMAP$-потока

Самым известным частным случаем $BMAP$-потока является стационарный пуассоновский поток, широко используемый в ТМО, начиная с работ А. К. Эрланга, и кодируемый в обозначениях Дж. Кендалла как M. Этот поток получается из $BMAP$-потока, если положить размерность управляющего процесса равной единице, матрицу D_0 равной скаляру $-\lambda$, матрицу D_1 равной скаляру λ, а матрицы $D_k,\ k \geq 2$, равными нулю.

Другие потоки, перечисляемые ниже в порядке их введения в рассмотрение, можно рассматривать как вехи на пути эволюции от рассмотрения стационарного пуассоновского потока к введению $BMAP$-потока как модели потоков информации в современных телекоммуникационных сетях.

- IPP (Interrupted Poisson Process) – прерывающийся пуассоновский поток. Такой поток описывается следующим образом. В течение времени, имеющего показательное распределение с параметром φ_0, поступления запросов не происходит. Затем в течение времени, имеющего показательное распределение с параметром φ_1, поступает стационарный пуассоновский поток запросов интенсивности λ_1. Далее описанный сценарий поступления повторяется.

  Этот поток является частным случаем $BMAP$-потока, описываемым двумя ненулевыми матрицами:

$$D_1 = \mathrm{diag}\{0, \lambda_1\},\ \ D_0 = -D_1 + \begin{pmatrix} -\varphi_0 & \varphi_0 \\ \varphi_1 & -\varphi_1 \end{pmatrix}.$$

- SPP (Switched Poisson Process) – переключающийся пуассоновский поток. Такой поток описывается следующим образом. В течение времени, имеющего показательное распределение с параметром φ_0, поступает стационарный пуассоновский поток запросов интенсивности λ_0. Затем в течение времени, имеющего показательное распределение с параметром φ_1, поступает стационарный пуассоновский поток

запросов интенсивности λ_1. Далее описанный сценарий поступления повторяется.

Этот поток является частным случаем $BMAP$-потока, описываемым двумя ненулевыми матрицами:

$$D_1 = \mathrm{diag}\{\lambda_0, \lambda_1\}, \quad D_0 = -D_1 + \begin{pmatrix} -\varphi_0 & \varphi_0 \\ \varphi_1 & -\varphi_1 \end{pmatrix}.$$

- $MMPP$ (Markovian Modulated Poisson Process) – марковский модулированный пуассоновский поток. Этот поток является естественным расширением SPP-потока на случай произвольного числа состояний управляющего процесса. С практической точки зрения, $MMPP$ есть наиболее важный частный случай $BMAP$-потока, у которого матрицы D_k, $k \geq 0$, определяются следующим образом:

$$D_0 = \Phi(P - I) - \Lambda, \quad D_1 = \Lambda, \quad D_k = 0, \; k \geq 2,$$

где $\Lambda = \mathrm{diag}\{\lambda_0, \ldots, \lambda_W\}$, $\Phi = \mathrm{diag}\{\varphi_0, \ldots, \varphi_W\}$, P — стохастическая матрица.

Этот частный случай $BMAP$-потока имеет прозрачную физическую трактовку, которая заключается в следующем. Имеется $W + 1$ возможных уровней интенсивности входного потока. На уровне номер ν поток ведет себя как обычный стационарный пуассоновский поток интенсивности λ_ν, $\nu = \overline{0, W}$. Уровень ν потока сохраняется в течение времени, имеющего показательное распределение с параметром φ_ν. После этого с вероятностью $p_{\nu,\nu'}$ поток перескакивает на ν'-й уровень. Здесь $p_{\nu,\nu'} = (P)_{\nu,\nu'}$ — (ν, ν')-й элемент матрицы P, $\nu, \nu' = \overline{0, W}$.

- $BMMPP$ (Batch Markovian Modulated Poisson Process) – групповой марковский модулированный пуассоновский поток. Этот поток является групповым аналогом $MMPP$-потока. Для его описания необходимо дополнительно задать множество вероятностей $\delta_k^{(\nu)}$, $k \geq 1$, $\nu = \overline{0, W}$, где $\delta_k^{(\nu)}$ есть вероятность того, что при поступлении группы запросов из ν-го стационарного группового пуассоновского потока эта группа состоит ровно из k запросов. Таким образом, этот поток задается совокупностью матриц

$$D_0 = \Phi(P - I) - \Lambda, \quad D_k = \mathrm{diag}\{\delta_k^{(0)}, \ldots, \delta_k^{(W)}\}\Lambda, \; k \geq 1.$$

- MAP (Markovian Arrival Process) – марковский входной поток. Этот поток является ординарным аналогом $BMAP$-потока. Он задается двумя матрицами: D_0 и D_1.

- PH (Phase type) – поток фазового типа. Это рекуррентный поток, являющийся частным случаем MAP-потока. Времена между моментами поступления запросов в PH-потоке являются независимыми одинаково распределенными случайными величинами, имеющими распределение фазового типа, описанное в следующем подразделе.

3.1.5 Распределение фазового типа – PH-распределение

3.1.5.1 Определение PH-распределения. Чтобы определить распределение фазового типа, введем некоторые обозначения.

Пусть $\boldsymbol{\beta}$ является стохастическим вектором порядка M, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$, где $\beta_m \geq 0$, $m = \overline{1, M}$, $\boldsymbol{\beta}\mathbf{e} = 1$, а S является субгенератором порядка M, то есть квадратной матрицей порядка M, имеющей отрицательные диагональные элементы $S_{m,m}$ и неотрицательные недиагональные элементы $S_{m,m'}$, $m' \neq m$, $m = \overline{1, M}$, причем $\sum\limits_{m'=1}^{M} S_{m,m'} \leq 0$ для любого m и как минимум для одного из m эта сумма строго меньше нуля.

Говорят, что случайная величина ξ имеет распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}, S)$ (матрица $S + \mathbf{S}_0\boldsymbol{\beta}$ является неприводимой), если она определяется следующим образом.

Пусть имеется ЦМ с непрерывным временем η_t, $t \geq 0$. Эта ЦМ имеет пространство состояний $\{1, \ldots, M, M+1\}$, причем состояние $M+1$ является единственным поглощающим состоянием. В начальный момент времени $t = 0$ процесс η_t, $t \geq 0$, принимает значение m, $m \in \{1, \ldots, M\}$ с вероятностью β_m. Время пребывания процесса η_t, $t \geq 0$, в состоянии m имеет показательное распределение с параметром $(-S)_{m,m}$. Затем с интенсивностью $S_{m,m'}$ процесс переходит в состояние m', $m' \in \{1, \ldots, M\}$, $m' \neq m$. При этом величина $-\sum\limits_{m'=1}^{M} S_{m,m'}$ является интенсивностью перехода процесса η_t из состояния m в поглощающее состояние $M + 1$. В момент попадания процесса η_t, $t \geq 0$, в поглощающее состояние истекает длительность случайной величины ξ, имеющей распределение фазового типа. Обозначим $\mathbf{S}_0 = -S\mathbf{e}$. Вектор $\mathbf{S}_0$ неотрицателен и имеет как минимум одну положительную компоненту. $(\mathbf{S}_0)_m$ – m-ая компонента вектора $\mathbf{S}_0$ – задаёт интенсивность перехода в поглощающее состояние из состояния m, $m \in \{1, \ldots, M\}$.

Время, имеющее распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}, S)$, можно интерпретировать следующим образом. Вирту-

альная заявка поступает в сеть, состоящую из M узлов, и осуществляет в ней переходы. Начальный узел выбирается из множества $\{1, \ldots, M\}$ случайным образом в соответствии с распределением, заданным компонентами вектора $\boldsymbol{\beta}$. Попав в узел сети с номером m, виртуальная заявка пребывает в нем в течение времени, имеющего экспоненциальное распределение с параметром $-S_{m,m}$, после чего она с вероятностью $-\frac{S_{m,m'}}{S_{m,m}}$ переходит в некоторое состояние $m' \neq m,\ m' = \overline{1, M}$, и продолжает свои переходы, а с вероятностью $1 + \sum\limits_{m'=1,\, m' \neq m}^{M} \frac{S_{m,m'}}{S_{m,m}}$ она переходит в поглощающее состояние. Время до попадания этой виртуальной заявки в поглощающее состояние имеет распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}, S)$.

ФР $A(x)$ случайной величины, имеющей распределение фазового типа, имеет вид:

$$A(x) = 1 - \boldsymbol{\beta} e^{Sx} \mathbf{e}.$$

Очевидно, что если неприводимое представление $(\boldsymbol{\beta}, S)$ известно, то легко можно определить функцию распределения $A(x)$. Обратная задача (определение неприводимого представления $(\boldsymbol{\beta}, S)$ по виду ФР $A(x)$) может иметь множество решений.

Преобразование Лапласа – Стилтьеса (ПЛС) $\alpha(s) = \int\limits_0^\infty e^{-st} dA(t)$, $\operatorname{Re} s \geq 0$, для распределения фазового типа имеет вид

$$\alpha(s) = \boldsymbol{\beta}(sI - S)^{-1} \mathbf{S}_0.$$

Среднее время (первый момент) между моментами поступления запросов в PH-потоке (среднее время до достижения поглощающего состояния процессом $\eta_t,\ t \geq 0$) определяется как:

$$a_1 = \boldsymbol{\beta}(-S)^{-1} \mathbf{e}.$$

Начальный момент m-го порядка этого распределения определяется как:

$$a_m = \int\limits_0^\infty t^m dA(t) = m! \boldsymbol{\beta}(-S)^{-m} \mathbf{e},\ m \geq 2.$$

Известно, что множество PH распределений всюду плотно, то есть для любой ФР $A(t)$ неотрицательной случайной величины можно подобрать распределение фазового типа сколь угодно близкое к распределению $A(t)$ в смысле слабой сходимости.

3.1.5.2 Частные случаи PH распределения. Многие традиционно используемые в ТМО распределения являются частными случаями PH распределения. Например:

- Распределение Эрланга k-го порядка (E_k). Его ФР имеет вид:

$$A(t) = \int_0^t \lambda \frac{(\lambda u)^{k-1}}{(k-1)!} e^{-\lambda u} du,\ t \geq 0.$$

Это распределение определяется вектором $\boldsymbol{\beta} = (1, 0, 0, ..., 0)$ и субгенератором вида

$$S = \begin{pmatrix} -\lambda & \lambda & 0 & \dots & 0 \\ 0 & -\lambda & \lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda \end{pmatrix}.$$

- Гиперэкспоненциальное распределение k-го порядка (HM_k). Его ФР имеет вид

$$A(t) = \sum_{l=1}^{k} q_l (1 - e^{-\lambda_l t}),$$

где $q_l \geq 0$, $l = \overline{1,k}$, $\sum\limits_{l=1}^{k} q_l = 1$.

В этом случае

$$\boldsymbol{\beta} = (q_1, ..., q_k),\ S = \mathrm{diag}\{-\lambda_1, ..., -\lambda_k\}.$$

3.1.5.3 Распределение минимума и максимума случайных величин, имеющих PH распределение

Лемма 3.1. *Пусть ξ_k есть независимые случайные величины, имеющие распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}_k, S_k)$, $k = \overline{1,K}$, и пусть*

$$\xi = \min_{k=\overline{1,K}} \xi_k.$$

Тогда случайная величина ξ имеет распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}_1 \otimes \dots \boldsymbol{\beta}_K, S_1 \oplus \dots \oplus S_K)$.

Для простоты изложения проведем доказательство для случая $K = 2$. На общий случай оно переносится очевидным образом. Очевидно, что

$$P\{\xi > t\} = P\{\xi_1 > t\}P\{\xi_2 > t\} = \boldsymbol{\beta}_1 e^{S_1 t}\mathbf{e}\boldsymbol{\beta}_2 e^{S_2 t}\mathbf{e}.$$

Поскольку для скалярных величин операции произведения и кронекерова произведения эквивалентны, с учетом правила смешанного произведения матриц и определения кронекеровой суммы матриц (см. Приложение А), получаем

$$P\{\xi > t\} = \boldsymbol{\beta}_1 e^{S_1 t}\mathbf{e} \otimes \boldsymbol{\beta}_2 e^{S_2 t}\mathbf{e} = (\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2)(e^{S_1 t} \otimes e^{S_2 t})(\mathbf{e} \otimes \mathbf{e}) =$$

$$= (\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2)e^{(S_1 \oplus S_2)t}\mathbf{e}.$$

А это и означает, что случайная величина ξ имеет распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2, S_1 \oplus S_2)$.

Лемма 3.2. *Пусть ξ_k есть независимые случайные величины, имеющие распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}, S)$, где размерность вектора и матрицы равна M, и пусть*

$$\eta_n = \max_{k=\overline{1,n}} \xi_k,\ n \geq 2.$$

Тогда случайная величина η_n имеет распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}^{(n)}, S^{(n)})$, которое определяется рекуррентным образом:

$$\boldsymbol{\beta}^{(n)} = \left(\begin{array}{c|c|c} \boldsymbol{\beta} \otimes \boldsymbol{\beta}^{(n-1)} & \mathbf{0}_{M_{n-1}} & \mathbf{0}_M \end{array} \right),$$

$$S^{(n)} = \left(\begin{array}{c|c|c} S \oplus S^{(n-1)} & \mathbf{S}_0 \otimes I_{M_{n-1}} & I_M \otimes \mathbf{S}_0^{(n-1)} \\ \hline O & S^{(n-1)} & O \\ \hline O & O & S \end{array} \right)$$

с начальным условием

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta},\ \ S^{(1)} = S,$$

где размерность M_n вектора $\boldsymbol{\beta}^{(n)}$ определяется как $M_n = (M+1)^n - 1$, $n \geq 1$.

Доказательство. Из интерпретации распределения фазового типа в терминах времени блуждания виртуальной заявки в сети до выхода в поглощающее состояние можно понять, что максимум двух независимых случайных величин, имеющих распределение фазового типа с неприводимыми представлениями $(\boldsymbol{\gamma}^{(k)}, \Gamma^{(k)})$, $k = 1, 2$, где размерность вектора $\boldsymbol{\gamma}^{(k)}$ есть $M^{(k)}$, имеет распределение фазового типа с неприводимыми представлениями $(\boldsymbol{\gamma}, \Gamma)$, определенным следующим образом:

$$\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}^{(1)} \otimes \boldsymbol{\gamma}^{(2)} \;\middle|\; \mathbf{0}_{M^{(2)}} \;\middle|\; \mathbf{0}_{M^{(1)}} \right),$$

$$\Gamma = \left(\begin{array}{c|c|c} \Gamma^{(1)} \oplus \Gamma^{(2)} & \boldsymbol{\Gamma}_0^{(1)} \otimes I_{M^{(2)}} & I_{M^{(1)}} \otimes \boldsymbol{\Gamma}_0^{(2)} \\ \hline O & \Gamma^{(2)} & O \\ \hline O & O & \Gamma^{(1)} \end{array} \right),$$

где $\boldsymbol{\Gamma}_0^{(k)} = -\Gamma^{(k)}\mathbf{e}$. Утверждение леммы теперь очевидным образом следует из этого результата. □

Из данного доказательства нетрудно заключить, что утверждение леммы очевидным образом обобщается на случай, когда случайные величины ξ_k могут иметь неприводимые представления, зависящие от k.

Лемма 3.3. *Пусть ξ_k, $k = \overline{1, K}$, есть независимые одинаково распределенные случайные величины, имеющие распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}, S)$, и пусть*

$$\eta = \max_{k=\overline{1,K}} \xi_k.$$

Тогда ФР случайной величины η имеет вид

$$P\{\eta < t\} = \sum_{k=0}^{K} C_K^k (-1)^k \boldsymbol{\beta}^{\otimes k} e^{S^{\oplus k} t} \mathbf{e},$$

где

$$\boldsymbol{\beta}^{\otimes k} \stackrel{def}{=} \underbrace{\boldsymbol{\beta} \otimes \ldots \otimes \boldsymbol{\beta}}_{k}, \quad k \geq 1,$$

$$S^{\oplus k} \stackrel{def}{=} \underbrace{S \oplus \ldots \oplus S}_{k}, \quad k \geq 1, \quad S^{\oplus 0} \stackrel{def}{=} 0.$$

Доказательство проводится по индукции с учетом опыта доказательства леммы для минимума случайных величин, основываясь на формуле

$$P\{\eta < t\} = \left(1 - \boldsymbol{\beta} e^{St}\mathbf{e}\right)^K.$$

3.1.6 Вычисление вероятностей поступления фиксированного числа запросов $BMAP$-потока за случайное время

При исследовании многих систем, например, системы типа $BMAP/G/1$, возникает следующая вспомогательная задача. Пусть имеется интервал времени, длина которого является случайной величиной ξ, имеющей ФР $B(t)$. Пусть имеется $BMAP$ - поток, заданный последовательностью матриц $D_k,\ k \geq 0$. Матрицы $P(n,t) = (P_{\nu,\nu'}(n,t))_{\nu,\nu'=\overline{0,W}}$, $n \geq 0$, задают условные вероятности поступления n запросов в потоке за время t при соответствующих переходах за это время управляющего процесса потока. Требуется вычислить матрицы $Y_l,\ l \geq 0$, задающие условные вероятности поступления l запросов в потоке за время ξ при соответствующих переходах управляющего процесса потока.

Из формулы полной вероятности следует очевидная формула для расчета матриц $Y_l,\ l \geq 0$:

$$Y_l = \int_0^\infty P(l,t)dB(t).$$

Поскольку матрицы $P(n,t),\ n \geq 0$, в общем случае не вычисляются в явном виде, проблема непосредственного расчета матриц $Y_l,\ l \geq 0$, является довольно сложной.

В общем случае может быть использовано соотношение (3.7). В результате получается следующая формула:

$$Y_l = \sum_{j=0}^{\infty} \gamma_j K_l^{(j)},\ l \geq 0,$$

где

$$\gamma_j = \int_0^\infty e^{-\tilde{\theta}t}\frac{(\tilde{\theta}t)^j}{j!}dB(t),\ j \geq 0,$$

а $K_l^{(j)},\ l \geq 0, j \geq 0$, – матрицы, заданные рекурсией (3.8).

Подчеркнем достоинство этой формулы, состоящее в следующем. Матрицы $K_l^{(j)}$ зависят от входящего потока и не зависят от распределения времени обслуживания. В свою очередь, вероятности γ_j зависят от распределения времени обслуживания, а от входящего потока зависят только через скалярную величину $\tilde{\theta}$.

Вычисление матриц $Y_l,\ l \geq 0$, существенно упрощается, когда распределение $B(t)$ является распределением фазового типа.

Лемма 3.4. *Если распределение $B(t)$ является распределением фазового типа, задаваемым неприводимым представлением $(\boldsymbol{\beta}, S)$, то матрицы $Y_l,\ l \geq 0$, рассчитываются следующим образом:*

$$Y_l = Z_l(I_{\bar{W}} \otimes \mathbf{S}_0),\ l \geq 0, \tag{3.18}$$

где

$$Z_0 = -(I_{\bar{W}} \otimes \boldsymbol{\beta})(D_0 \oplus S)^{-1}, \tag{3.19}$$

$$Z_l = -\sum_{i=0}^{l-1} Z_i(D_{l-i} \otimes I_M)(D_0 \oplus S)^{-1},\ l \geq 1. \tag{3.20}$$

Доказательство. Справедлива следующая цепочка равенств:

$$Y_l = \int_0^\infty P(l,t)dB(t) = \int_0^\infty P(l,t)\boldsymbol{\beta}e^{St}\mathbf{S}_0 dt = \int_0^\infty P(l,t)I_{\bar{W}} \otimes \boldsymbol{\beta}e^{St}\mathbf{S}_0 dt =$$

$$= \int_0^\infty (P(l,t) \otimes \boldsymbol{\beta}e^{St})(I_{\bar{W}} \otimes \mathbf{S}_0)dt = Z_l(I_{\bar{W}} \otimes \mathbf{S}_0),$$

где

$$Z_l = \int_0^\infty P(l,t) \otimes \boldsymbol{\beta}e^{St}dt,\ l \geq 0.$$

Здесь были использованы вид ФР $B(t)$ фазового типа, искусственная замена обычного произведения на скалярную величину $\boldsymbol{\beta}e^{St}\mathbf{S}_0$ на кронекерово произведение и правило смешанного произведения для кронекеровых произведений, см. Приложение А. Таким образом, формула (3.18) получена и осталось вывести формулы (3.19) и (3.20) для матриц $Z_l,\ l \geq 0$. Интегрируя по частям, получаем:

$$Z_l = [P(l,t) \otimes (\boldsymbol{\beta}e^{St}S^{-1})]|_0^\infty - \int_0^\infty P'(l,t) \otimes (\boldsymbol{\beta}e^{St}S^{-1})dt.$$

Можно убедиться, что справедливы соотношения: $P(l,t) \to O$ при $t \to \infty$ и $P(n,0) = \delta_{n,0} I_{\bar{W}}$. Учитывая их и матричное дифференциальное уравнение (3.2), получаем:

$$Z_l = -\delta_{l,0}(I_{\bar{W}} \otimes \boldsymbol{\beta} S^{-1}) - \int_0^\infty \sum_{i=0}^{l} (P(i,t) D_{l-i}) \otimes (\boldsymbol{\beta} e^{St} S^{-1}) dt =$$

$$= -\delta_{l,0}(I_{\bar{W}} \otimes \boldsymbol{\beta} S^{-1}) - \sum_{i=0}^{l} Z_i (D_{l-i} \otimes S^{-1}).$$

Отсюда следует формула

$$Z_l (I_{\bar{W}M} + D_0 \otimes S^{-1}) = -\delta_{l,0}(I_{\bar{W}} \otimes \boldsymbol{\beta} S^{-1}) - \sum_{i=0}^{l-1} Z_i (D_{l-i} \otimes S^{-1}). \quad (3.21)$$

Несложно убедиться, что обратная матрица к матрице $I_{\bar{W}M} + D_0 \otimes S^{-1}$ существует и задается следующим образом:

$$(I_{\bar{W}M} + D_0 \otimes S^{-1})^{-1} = (I_{\bar{W}} \otimes S)(D_0 \oplus S)^{-1}.$$

Учитывая эту формулу, из (3.21) получаем формулы (3.19), (3.20). □

3.1.7 Суперпозиция и просеивание $BMAP$-потоков

Класс $BMAP$-потоков имеет некоторое сходство со стационарным пуассоновским потоком и по отношению к операциям суперпозиции и просеивания.

Известно, что суперпозиция (наложение) двух независимых стационарных пуассоновских потоков с интенсивностями λ_1 и λ_2 также является пуассоновским потоком с интенсивностью $\lambda_1 + \lambda_2$.

Аналогично суперпозиция двух независимых $BMAP$-потоков, характеризующихся матричными ПФ $D_1(z)$ и $D_2(z)$ также является $BMAP$-потоком с матричной ПФ $D(z) = D_1(z) \oplus D_2(z)$. Средняя скорость суперпозиции двух $BMAP$-потоков равна сумме средних скоростей налагающихся потоков.

Также известно, что если к стационарному пуассоновскому потоку интенсивности λ применить простейшую рекуррентную процедуру просеивания, при которой с заданной вероятностью p заявка принимается в просеянный поток, а с дополнительной вероятностью отвергается, то просеянный поток также будет стационарным пуассоновским, а его интенсивность равна $p\lambda$. Поскольку $BMAP$-поток является групповым потоком,

возможны два прямых аналога этой процедуры просеивания. Один из них (вариант а): прибывшая группа принимается в просеянный поток с вероятностью q_a и отвергается с дополнительной вероятностью. Другой вариант (вариант б): каждый из запросов прибывшей группы, независимо от других запросов, принимается в просеянный поток с вероятностью вероятностью q_b и отвергается с дополнительной $1 - q_b$.

Лемма 3.5. *Если к $BMAP$-потоку с управляющим процессом ν_t, $t \geq 0$, имеющим пространство состояний $\{0, \ldots, W\}$ и матричную ПФ $D(z)$, применить один из двух (вариант а или вариант б) описанных процедур просеивания, то просеянный поток является также $BMAP$-потоком. Его управляющий процесс имеет то же пространство состояний, что и управляющий процесс ν_t, $t \geq 0$, исходного потока. А матричная ПФ этого потока имеет вид:*

$$q_a D(z) + (1 - q_a)D(1) \text{ в варианте а}, D(zq_b + (1 - q_b)) \text{ в варианте б}.$$

Средняя скорость потоков при этом равна λq_a или λq_b, где λ — средняя скорость исходного потока.

Таким образом, семейство $BMAP$-потоков является замкнутым относительно операций суперпозиции и простейших процедур просеивания, что аналогично свойству стационарных пуассоновских потоков.

3.1.8 Групповой маркированный марковский входной поток (*BMMAP*)

$BMAP$-поток, описанный в предыдущем разделе, является потоком однородных запросов. Многие потоки в реальных системах являются неоднородными. Например, потоки информации в современных телекоммуникационных сетях являются смесью (суперпозицией) потоков речевых сообщений, передаваемых в цифровой форме, потоков интерактивных данных, потоков данных, не требующих передачи в режиме он-лайн, но требующих высокой надежности передачи, потоков видео-, аудио- и мультимедиа информации. Для моделирования таких неоднородных потоков в случае, если они являются кореллированными, используется так называемый маркированный марковский входной поток ($MMAP$ – Marked Markov Arrival Process) или его групповой аналог $BMMAP$ – Batch Marked Markov Arrival Process.

Процесс поступления запросов в $BMMAP$-потоке, являющемся суперпозицией L потоков запросов, отличающихся типом, происходит под управлением некоторой неприводимой ЦМ ν_t, $t \geq 0$, с непрерывным временем и конечным пространством состояний $\{0, ..., W\}$. Время пребывания цепи ν_t в некотором состоянии ν имеет показательное распределение с параметром $\lambda_\nu, \nu = \overline{0, W}$. После того как время пребывания процесса в этом состоянии истекло, с вероятностью $p_0(\nu, \nu')$ процесс переходит в некоторое состояние $\nu',\ \nu' \neq \nu$ без генерации запросов, либо с вероятностью $p_k^{(l)}(\nu, \nu')$ переходит в некоторое состояние ν' и при этом генерируется группа из k запросов l-го типа, $k \geq 1,\ l = \overline{1, L}$. При этом

$$p_0(\nu, \nu) = 0,\ \sum_{l=1}^{L} \sum_{k=1}^{\infty} \sum_{\nu'=0}^{W} p_k^{(l)}(\nu, \nu') + \sum_{\nu'=0}^{W} p_0(\nu, \nu') = 1,\ \nu = \overline{0, W}.$$

Параметры, характеризующие $BMMAP$-поток, удобно хранить в квадратных матрицах D_0, $D_k^{(l)}$, $k \geq 1$,$l = \overline{1, L}$, порядка $\bar{W} = W + 1$, определенных их элементами следующим образом:

$$(D_0)_{\nu,\nu} = -\lambda_\nu,\ (D_0)_{\nu,\nu'} = \lambda_\nu p_0(\nu, \nu'),\ \nu \neq \nu',$$

$$(D_k^{(l)})_{\nu,\nu'} = \lambda_\nu p_k^{(l)}(\nu, \nu'),\ \nu, \nu' = \overline{0, W},\ k \geq 1,\ l = \overline{1, L}.$$

Обозначим

$$D(1) = D_0 + \sum_{l=1}^{L} \sum_{k=1}^{\infty} D_k^{(l)},\ \hat{D}_k^{(l)} = \sum_{i=k}^{\infty} D_i^{(l)},$$

$$\mathcal{D}_k^{(l)} = \sum_{i=k}^{\infty} \sum_{\bar{l}=1, \bar{l} \neq l}^{L} D_i^{(\bar{l})},\ k \geq 1,\ l = \overline{1, L}.$$

Вектор $\boldsymbol{\theta}$ стационарного распределения вероятностей состояний ЦМ ν_t, $t \geq 0$, является единственным решением системы линейных алгебраических уравнений

$$\boldsymbol{\theta} D(1) = \mathbf{0}, \boldsymbol{\theta}\mathbf{e} = 1.$$

Средняя скорость λ_l поступления запросов l-го типа вычисляется по формуле

$$\lambda_l = \boldsymbol{\theta} \sum_{k=1}^{\infty} k D_k^{(l)} \mathbf{e}, l = \overline{1, L}.$$

Средняя скорость $\lambda_l^{(b)}$ поступления групп запросов l-го типа вычисляется по формуле

$$\lambda_l^{(b)} = \boldsymbol{\theta}\hat{D}_1^{(l)}\mathbf{e},\ \ l = \overline{1, L}.$$

Дисперсия v_l длин интервалов между моментами поступления групп запросов l-го типа вычисляется по формуле

$$v_l = \frac{2\boldsymbol{\theta}(-D_0 - \mathcal{D}_1^{(l)})^{-1}\mathbf{e}}{\lambda_l^{(b)}} - \left(\frac{1}{\lambda_l^{(b)}}\right)^2,\ l = \overline{1, L}.$$

Коэффициент корреляции $C_{cor}^{(l)}$ длин двух соседних интервалов между моментами поступления групп запросов l-го типа вычисляется по формуле

$$C_{cor}^{(l)} = \left[\frac{\boldsymbol{\theta}(-D_0 - \mathcal{D}_1^{(l)})^{-1}}{\lambda_l^{(b)}}\hat{D}_1^{(l)}(-D_0 - \mathcal{D}_1^{(l)})^{-1}\mathbf{e} - \left(\frac{1}{\lambda_l^{(b)}}\right)^2\right]v_l^{-1},\ l = \overline{1, L}.$$

3.1.9 Полумарковский входной поток (SM)

Такой поток является более общим, чем MAP-поток и задается следующим образом. Пусть имеется регулярный эргодический полумарковский случайный процесс $m_t,\ t \geq 0$, характеризующийся своим пространством состояний $\{1, 2, \ldots, M\}$ и полумарковским ядром $B(x)$ с компонентами $B_{m,m'}(x)$. Функция $B_{m,m'}(x)$ интерпретируется как вероятность того, что время пребывания процесса в текущем состоянии не превысит величину x и следующий переход произойдет в состояние m' при условии, что текущим состоянием процесса является состояние $m,\ m, m' = \overline{1, M}$. Матрица $B(\infty)$ является матрицей переходных вероятностей для ЦМ, вложенной по всем моментам скачков процесса $m_t,\ t \geq 0$. Предполагаем, что эта вложенная ЦМ неприводимая и, следовательно, существует вектор $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_M)$ ее стационарного распределения, удовлетворяющий уравнениям:

$$\boldsymbol{\delta}B(\infty) = \boldsymbol{\delta},\ \boldsymbol{\delta}\mathbf{e} = 1.$$

Также предполагается существование интеграла $\int\limits_0^\infty tdB(t)$.

Полумарковским входным потоком называется случайный поток, интервалы между моментами запросов в котором являются последовательными временами пребывания полумарковского процесса $m_t,\ t \geq 0$, в своих состояниях. k-й начальный момент длины интервала между моментами поступления запросов вычисляется как $b_k = \boldsymbol{\delta}\int\limits_0^\infty t^k dB(t)\mathbf{e},\ k \geq 1$.

Важным частным случаем полумарковского входного потока является случай, когда полумарковское ядро $B(t)$ может быть представлено в виде:

$$B(t) = \mathrm{diag}\{B_1(t), \ldots, B_M(t)\}P,$$

где $B_m(t)$ – ФР времени пребывания процесса $m_t,\ t \geq 0$, в состоянии $m,\ m = \overline{1, M}$, а $P = (p_{i,j})_{i,j=\overline{1,M}} = B(\infty)$.

Из определения полумарковского входного потока следует, что, вообще говоря, интервалы между моментами поступления запросов могут быть коррелированными. В упомянутом выше частном случае полумарковского потока формула для вычисления коэффициента корреляции длин соседних интервалов между моментами поступления запросов SM-потока имеет вид:

$$c_{cor} = \frac{\sum\limits_{i=1}^{M}\sum\limits_{j=1}^{M}\delta_i b_1^{(i)} p_{i,j} b_1^{(j)} - (\sum\limits_{i=1}^{M}\delta_i b_1^{(i)})^2}{\sum\limits_{i=1}^{M}\delta_i b_2^{(i)} - (\sum\limits_{i=1}^{M}\delta_i b_1^{(i)})^2},$$

где $b_k^{(i)} = \int\limits_0^\infty t^k dB_i(t),\ k = 1, 2,\ i = \overline{1, M}$.

3.2 МНОГОМЕРНЫЕ ПРОЦЕССЫ ГИБЕЛИ И РАЗМНОЖЕНИЯ

При изучении классических марковских СМО полезным является использование известных результатов для так называемых процессов гибели и размножения.

Процесс гибели и размножения $i_t,\ t \geq 0$, есть частный случай однородной ЦМ с непрерывным временем, принимающей значения в конечном или счетном пространстве состояний. Рассмотрим процесс гибели и размножения со счетным пространством состояний $\{0, 1, 2, \ldots\}$. Матрица инфинитезимальных коэффициентов (генератор) Q этого процесса имеет трехдиагональную структуру

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \ldots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \ldots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Параметр λ_i называется интенсивностью размножения, а параметр μ_i – интенсивностью гибели в состоянии i.

Вероятности $p_i(t) = P\{i_t = i\},\ i \geq 0$, скомпонованные в вектор-строку $\mathbf{p}(t) = (p_0(t), p_1(t), p_2(t), \dots)$, удовлетворяют бесконечной системе уравнений

$$\frac{d\mathbf{p}(t)}{dt} = \mathbf{p}(t)Q$$

с условием нормировки $\mathbf{p}(t)\mathbf{e} = 1$.

Если начальное распределение вероятностей $\mathbf{p}(0)$ известно, то решение этой системы уравнений имеет вид

$$\mathbf{p}(t) = \mathbf{p}(0)e^{Qt}.$$

Обозначим $\rho_k = \frac{\prod\limits_{l=0}^{k-1} \lambda_l}{\prod\limits_{l=1}^{k} \mu_l},\ k \geq 0.$

При выполнении условий сходимости ряда $\sum\limits_{k=0}^{\infty} \rho_k$ и расходимости ряда $\sum\limits_{k=1}^{\infty} \prod\limits_{i=1}^{k} \frac{\mu_i}{\lambda_i}$ существуют стационарные вероятности процесса гибели и размножения $i_t,\ t \geq 0$,

$$p_i = \lim_{t \to \infty} p_i(t),\ i \geq 0,$$

задаваемые выражениями

$$p_0 = (\sum_{k=0}^{\infty} \rho_k)^{-1},\ p_i = p_0\rho_i,\ i \geq 1.$$

3.2.1 Определение многомерного процесса гибели и размножения и его стационарное распределение

При исследовании многих СМО, например систем с MAP-потоком и (или) процессом обслуживания типа PH, систем, функционирующих в случайной среде, тандемных систем и т.д., возникает необходимость анализа стационарного поведения двумерного или многомерного марковского процесса $\xi_t = \{i_t, \mathbf{v}_t\}, t \geq 0$, с пространством состояний

$$\mathcal{S} = \{(i, \mathbf{v}), \mathbf{v} \in \mathcal{V}_i,\ i \geq 0\},$$

где $\mathcal{V}_i$ – некоторое конечное множество, если марковский процесс $\xi_t = \{i_t, \mathbf{v}_t\}, t \geq 0$, двумерный, или некоторое конечное множество конечномерных векторов, если марковский процесс $\xi_t = \{i_t, \mathbf{v}_t\}, t \geq 0$, – многомерный.

Перенумеруем состояния процесса $\xi_t = \{i_t, \mathbf{v}_t\}, t \geq 0$, в лексикографическом порядке и объединим состояния, имеющие значение i компоненты i_t, в макро-состояние i, иногда называемое также уровнем процесса $\xi_t = \{i_t, \mathbf{v}_t\}, t \geq 0$. Соответственно такой нумерации, генератор Q этого процесса запишется в блочном виде $Q = (Q_{i,l})_{i,l\geq 0}$, где $Q_{i,l}$ есть матрица, образованная интенсивностями $q_{(i,\mathbf{r});(l,\mathbf{v})}$ перехода ЦМ $\xi_t, t \geq 0$, из состояния $(i, \mathbf{r}), \mathbf{r} \in \mathcal{V}_i$, в состояние $(l, \mathbf{v}), \mathbf{v} \in \mathcal{V}_l$. Диагональные элементы матрицы $Q_{i,i}$ определены как $q_{(i,\mathbf{r});(i,\mathbf{v})} = -\sum\limits_{(j,\mathbf{v})\in\mathcal{S}\setminus(i,\mathbf{r})} q_{(i,\mathbf{r});(j,\mathbf{v})}$.

Процесс $\xi_t = \{i_t, \mathbf{v}_t\}, t \geq 0$, называется многомерным (векторным) процессом гибели и размножения (QBD – Quasi-Birth-and-Death process), если неприводимый блочный генератор $Q = (Q_{i,l})_{i,l\geq 0}$ этого процесса представим в виде

$$Q = \begin{pmatrix} Q_{0,0} & Q_0 & O & O & \dots \\ Q_2 & Q_1 & Q_0 & O & \dots \\ O & Q_2 & Q_1 & Q_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{3.22}$$

Из вида (3.22) генератора Q следует, что множества $\mathcal{V}_i$ для различных значений i, $i > 0$, здесь совпадают.

Теорема 3.1. *Необходимым и достаточным условием существования стационарного распределения вероятностей ЦМ ξ_t, $t \geq 0$, является выполнение неравенства*

$$\mathbf{y}Q_2\mathbf{e} > \mathbf{y}Q_0\mathbf{e}, \tag{3.23}$$

где вектор-строка $\mathbf{y}$ является единственным решением системы линейных алгебраических уравнений

$$\mathbf{y}(Q_0 + Q_1 + Q_2) = \mathbf{0},\ \mathbf{y}\mathbf{e} = 1. \tag{3.24}$$

Доказательство этого условия приведено в [163]. Это условие также автоматически получается из условия (3.67),(3.68), доказанного ниже, в разделе 3.4, для цепей Маркова типа $M/G/1$.

Далее предполагаем условие (3.23) выполненным. Тогда существуют следующие пределы

$$\pi(i, \mathbf{v}) = \lim_{t\to\infty} P\{i_t = i,\ \mathbf{v}_t = \mathbf{v}\},\ \mathbf{v} \in \mathcal{V}_i,\ i \geq 0,$$

называемые стационарными вероятностями состояний ЦМ.

Соответственно введенному лексикографическому упорядочиванию состояний ЦМ, сгруппируем эти вероятности в векторы-строки

$$\boldsymbol{\pi}_i = (\pi(i, \mathbf{v}),\ \mathbf{v} \in \mathcal{V}_i),\ i \geq 0.$$

Теорема 3.2. *Векторы стационарных вероятностей состояний ЦМ* $\xi_t,\ t \geq 0$, *вычисляются следующим образом:*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \mathcal{R}^i,\ i \geq 0, \tag{3.25}$$

где матрица $\mathcal{R}$ *является минимальным неотрицательным решением матричного уравнения*

$$\mathcal{R}^2 Q_2 + \mathcal{R} Q_1 + Q_0 = O, \tag{3.26}$$

а вектор $\boldsymbol{\pi}_0$ *является единственным решением следующей системы линейных алгебраических уравнений*

$$\boldsymbol{\pi}_0 (Q_{0,0} + \mathcal{R} Q_2) = \mathbf{0}, \tag{3.27}$$

$$\boldsymbol{\pi}_0 (I - \mathcal{R})^{-1} \mathbf{e} = 1. \tag{3.28}$$

Доказательство. Известно, что вектор $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)$ является решением системы уравнений равновесия

$$\boldsymbol{\pi} Q = \mathbf{0}, \tag{3.29}$$

с условием нормировки

$$\boldsymbol{\pi} \mathbf{e} = 1. \tag{3.30}$$

Предположим, что решение системы (3.29) имеет вид (3.25). Подставляя векторы $\boldsymbol{\pi}_i,\ i \geq 1$, в форме (3.25) в уравнения

$$\boldsymbol{\pi}_{i-1} Q_0 + \boldsymbol{\pi}_i Q_1 + \boldsymbol{\pi}_{i+1} Q_2 = \mathbf{0},\ i \geq 1,$$

системы (3.29), легко убедиться, что эти уравнения обращаются в тождества, если матрица $\mathcal{R}$ удовлетворяет матричному уравнению (3.26). Подставляя векторы $\boldsymbol{\pi}_i,\ i \geq 1$, в форме (3.25) в уравнения

$$\boldsymbol{\pi}_0 Q_{0,0} + \boldsymbol{\pi}_1 Q_2 = \mathbf{0},$$

системы (3.29), получаем, что вектор $\boldsymbol{\pi}_0$ удовлетворяет уравнению (3.27).

Известно, см. [163], что условие эргодичности (3.23) выполняется тогда и только тогда, когда спектральный радиус $\varrho(\mathcal{R})$ матрицы $\mathcal{R}$ строго меньше единицы. При выполнении этого условия ряд $\sum_{l=0}^{\infty}\mathcal{R}^l$ сходится и равен $(I-\mathcal{R})^{-1}$. Уравнение (3.28) теперь следует из (3.25) и условия нормировки $\sum_{i=0}^{\infty}\boldsymbol{\pi}_i\mathbf{e}=1$. □

Отметим, что матрица $\mathcal{R}$ представима в виде $\mathcal{R}=Q_0\mathcal{N}$, где элементы $\mathcal{N}_{\mathbf{v},\mathbf{v}'}$ матрицы $\mathcal{N}$ характеризуют среднее время пребывания ЦМ $\xi_t=\{i_t,\mathbf{v}_t\}, t\geq 0$, в некотором макросостоянии i до первого попадания на уровень $i-1$ при переходе компоненты $\mathbf{v}_t$ за это время из состояния $\mathbf{v}$ в состояние $\mathbf{v}'$.

Для решения уравнения (3.26) используются различные методы. Простейший из них – метод последовательных приближений. В качестве начального приближения $\mathcal{R}_0$ для матрицы $\mathcal{R}$ берется нулевая матрица. Формула для вычисления последовательных приближений очевидным образом следует из (3.26) в виде:

$$\mathcal{R}_{k+1}=(\mathcal{R}_k^2Q_2+Q_0)(-Q_1)^{-1},\ k\geq 0. \tag{3.31}$$

Отметим, что обратная матрица в (3.31) существует, поскольку матрица Q_1 является неприводимым субгенератором. В [163] показано, что последовательность матриц $\{\mathcal{R}_k,\ k\geq 0\}$, заданная рекурсией (3.31), сходится к минимальному неотрицательному решению матричного уравнения (3.26).

Другой алгоритм, так называемый алгоритм с логарифмической редукцией, использует следующую связь матрицы $\mathcal{R}$, являющейся решением уравнения (3.26),

$$\mathcal{R}=Q_0(-Q_1-Q_0\mathcal{G})^{-1}$$

с матрицей $\mathcal{G}$, являющейся решением матричного уравнения

$$Q_0\mathcal{G}^2+Q_1\mathcal{G}+Q_2=O.$$

В свою очередь, матрица $\mathcal{G}$ находится методом последовательных приближений на основе соотношения

$$\mathcal{G}=(-Q_1)^{-1}Q_0\mathcal{G}^2+(-Q_1)^{-1}Q_2.$$

Изложенные выше результаты для векторных процессов гибели и размножения легко переносятся на случай, когда блочный генератор процесса

имеет вид

$$Q = \begin{pmatrix} \tilde{Q}_{0,0} & \tilde{Q}_0 & O & O & \dots \\ \tilde{Q}_2 & Q_1 & Q_0 & O & \dots \\ O & Q_2 & Q_1 & Q_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где, в отличие от (3.22), матрицы $\tilde{Q}_{0,0},\ \tilde{Q}_0,\ \tilde{Q}_2$ могут иметь размерность, отличную от размерности блоков $Q_0,\ Q_1,\ Q_2$, и блоки $\tilde{Q}_0,\ \tilde{Q}_2$ – не квадратные матрицы.

В этом случае необходимое и достаточное условие существования стационарного распределения процесса $\xi_t,\ t \geq 0$, также задается неравенством (3.23), где вектор $\mathbf{y}$ является решением системы (3.24). Формулы для расчета векторов стационарных вероятностей здесь имеют следующий вид:

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_1 \mathcal{R}^{i-1},\ i \geq 1,$$

где матрица $\mathcal{R}$ является минимальным неотрицательным решением матричного уравнения (3.26), а векторы $\boldsymbol{\pi}_0$ и $\boldsymbol{\pi}_1$ являются единственным решением следующей системы линейных алгебраических уравнений

$$\boldsymbol{\pi}_0 \tilde{Q}_{0,0} + \boldsymbol{\pi}_1 \tilde{Q}_2 = \mathbf{0},$$

$$\boldsymbol{\pi}_0 \tilde{Q}_0 + \boldsymbol{\pi}_1 (Q_1 + \mathcal{R} Q_2) = \mathbf{0},$$

$$\boldsymbol{\pi}_0 \mathbf{e} + \boldsymbol{\pi}_1 (I - \mathcal{R})^{-1} \mathbf{e} = 1.$$

Аналогичным образом приведенные выше результаты переносятся на случай векторных процессов гибели и размножения, у которых не одна, а некоторое конечное число J блочных строк блочного генератора процесса имеет вид, отличный от вида в (3.22).

Для иллюстрации применим изложенные сведения для векторных процессов гибели и размножения к исследованию СМО типа $MAP/PH/1$.

3.2.2 Применение результатов для векторного процесса гибели и размножения к исследованию системы обслуживания $MAP/PH/1$

Имеется однолинейная СМО с бесконечным буфером. Входной поток является MAP-потоком, задаваемым управляющим процессом – неприводимой ЦМ ν_t, $t \geq 0$, с непрерывным временем и конечным пространством

состояний $\{0, 1, ..., W\}$ – и матрицами D_0 и D_1. Процесс обслуживания фазового типа задается ЦМ с непрерывным временем $\eta_t,\ t \geq 0$, с пространством непоглощающих состояний $\{1, \ldots, M\}$ и неприводимым представлением $(\boldsymbol{\beta}, S)$, где $\boldsymbol{\beta}$ – стохастический вектор-строка, а S – субгенератор.

Поведение этой системы описывается трехмерной ЦМ $\xi_t = \{i_t, \mathbf{v}_t\},\ t \geq 0$, где i_t – число запросов в системе в момент t, а двумерный процесс $\mathbf{v}_t = (\nu_t, \eta_t),\ t \geq 0$, описывает состояние управляющих процессов поступления и обслуживания запросов, $i_t \geq 0,\ \nu_t = \overline{0, W},\ \eta_t = \overline{1, M}$.

Заметим, что в моменты времени, когда $i_t = 0$, то есть запросов в системе нет, вообще говоря, состояние компоненты η_t не определено. Однако с целью избежать усложнения обозначений и результатов можно условиться, что в момент, когда система становится пустой, происходит, в соответствии с вектором $\boldsymbol{\beta}$, вероятностный выбор состояния процесса $\eta_t,\ t \geq 0$, которое считается "замороженным" до момента прихода запроса и начала его обслуживания.

Нетрудно убедиться, что ЦМ $\xi_t = \{i_t, \mathbf{v}_t\} = \{i_t, \nu_t, \eta_t\},\ t \geq 0$, является векторным процессом гибели и размножения и ее блочный генератор имеет вид (3.22) с блоками размера $\bar{W}M$ вида

$$Q_{0,0} = D_0 \otimes I_M,\ Q_0 = D_1 \otimes I_M,\ Q_1 = D_0 \oplus S,\ Q_2 = I_{\bar{W}} \otimes (\mathbf{S}_0 \boldsymbol{\beta}).$$

Здесь $\otimes$ и $\oplus$ – операции так называемых кронекерова произведения и суммы матриц, определяемых следующим образом. Если имеются две матрицы, $A = (a_{i,j})$ и $C = (c_{i,j})$, то их кронекеровым произведением является матрица, обозначаемая как $A \otimes C$ и состоящая из блоков $(a_{i,j}C)$. Кронекеровой суммой этих матриц является матрица, обозначаемая как $A \oplus C$ и задаваемая формулой $A \oplus C = A \otimes I_C + I_A \otimes C$, где I_A и I_C есть тождественные матрицы с порядками, совпадающими с порядками матриц A и C соответственно. Полезность использования этих операций при рассмотрении ЦМ с несколькими конечными компонентами предопределяется следующим. Если A и C есть матрицы одношаговых переходных вероятностей независимых цепей Маркова $\xi_n^{(1)}$ и $\xi_n^{(2)}$ соответственно, то матрицей переходных вероятностей двумерной цепи Маркова с дискретным временем $\{\xi_n^{(1)}, \xi_n^{(2)}\}$ является матрица $A \otimes C$. Аналогичное свойство справедливо для генератора двумерной цепи Маркова с непрерывным временем. Для более полного ознакомления со свойствами кронекерова произведения и суммы матриц рекомендуются книги [69] и [128]. Полезная информация о них приводится также в приложении. Как одно из наиболее ценных свойств

кронекерова произведения отметим так называемое правило смешанного произведения (mixed product rule):

$$(AB) \otimes (CD) = (A \otimes C)(B \otimes D).$$

Поскольку в последующих разделах придется неоднократно выписывать и более сложные генераторы, для более простого понимания кратко поясним приведенный вид блоков данного генератора. Блок $Q_{0,0}$ генератора задает интенсивности переходов процессов $\{\nu_t, \eta_t\}$ без изменения состояния 0 процесса i_t. Такие переходы возможны только при переходах процесса ν_t без генерации запросов. Интенсивности таких переходов задаются недиагональными элементами матрицы D_0. Диагональные элементы матрицы D_0 отрицательны и задают, с точностью до знака, интенсивности выхода процесса ν_t из соответствующих состояний. Поскольку за бесконечно малое время только один из процессов $\{\nu_t, \eta_t\}$ может осуществить переход, переход процесса ν_t влечет, что процесс η_t не осуществляет никакого перехода, т.е. матрица вероятностей его переходов есть I_M. Таким образом, получаем формулу $Q_{0,0} = D_0 \otimes I_M$. Форма блока $Q_0 = D_1 \otimes I_M$ объясняется аналогично.

Этот блок задает интенсивность перехода процессов $\{\nu_t, \eta_t\}$ с изменением состояния i процесса i_t на $i + 1$. Такие переходы возможны только при переходах процесса ν_t с генерацией запроса. Интенсивности таких переходов задаются элементами матрицы D_1. При этом снова процесс η_t не осуществляет никакого перехода, т.е. матрица вероятностей его переходов есть I_M. Форма $Q_2 = I_{\bar{W}} \otimes (\mathbf{S}_0\boldsymbol{\beta})$ блока, задающего интенсивность перехода процессов $\{\nu_t, \eta_t\}$ с изменением состояния i процесса i_t на $i - 1$, легко объясняется следующим образом. Такой переход возможен при окончании обслуживания запроса прибором. Интенсивности таких переходов процесса η_t задаются компонентами вектора-столбца $\mathbf{S}_0$. После этого перехода мгновенно устанавливается начальное состояние процесса η_t для описания обслуживания следующего запроса. Выбор начального состояния процесса η_t осуществляется с вероятностями, заданными компонентами вектора $\boldsymbol{\beta}$. Управляющий процесс потока ν_t при этом осуществлять переходов не может, т.е. матрица вероятностей его переходов есть $I_{\bar{W}}$. Форма блока Q_2 объяснена. Наконец, блок Q_1 генератора задает интенсивности переходов процессов $\{\nu_t, \eta_t\}$ без изменения состояния i процесса i_t. Такие переходы возможны только при переходах процесса ν_t без генерации запросов или процесса η_t внутри множества непоглощающих состояний. В результате

этих рассуждений получаем формулу $Q_1 = D_0 \otimes I_M + I_{\bar{W}} \otimes S = D_0 \oplus S$.

Выведем условие существования стационарного распределения ЦМ $\xi_t = \{i_t, \mathbf{v}_t\},\ t \geq 0$.

Матрица $Q_0 + Q_1 + Q_2$ здесь имеет вид

$$(D_0 + D_1) \oplus (S + \mathbf{S}_0\boldsymbol{\beta}) = (D_0 + D_1) \otimes I_M + I_{\bar{W}} \otimes (S + \mathbf{S}_0\boldsymbol{\beta}).$$

Поэтому учитывая правило смешанного произведения для кронекерова произведения матриц, легко убедиться, что вектор $\mathbf{y}$, являющийся решением системы уравнений (3.24), имеет вид

$$\mathbf{y} = \boldsymbol{\theta} \otimes \boldsymbol{\psi}, \tag{3.32}$$

где $\boldsymbol{\theta}$ – вектор стационарных вероятностей управляющего процесса MAP-потока, удовлетворяющий условиям $\boldsymbol{\theta}(D_0 + D_1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1$, а вектор $\boldsymbol{\psi}$ удовлетворяет уравнениям

$$\boldsymbol{\psi}(S + \mathbf{S}_0\boldsymbol{\beta}) = \mathbf{0},\ \boldsymbol{\psi}\mathbf{e} = 1. \tag{3.33}$$

Несложно проверить, что решением системы уравнений (3.33) является вектор

$$\boldsymbol{\psi} = \mu\boldsymbol{\beta}(-S)^{-1}, \tag{3.34}$$

где величина μ – средняя скорость обслуживания (величина, обратная к среднему времени обслуживания):

$$\mu^{-1} = b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}.$$

Подставляя вектор $\mathbf{y}$ в виде (3.32) в неравенство (3.23), с учетом выражения (3.34) получаем неравенство

$$\mu > \lambda. \tag{3.35}$$

Таким образом, нами доказано следующее утверждение.

Теорема 3.3. *Необходимым и достаточным условием существования стационарного распределения ЦМ $\xi_t = \{i_t, \mathbf{v}_t\},\ t \geq 0$, является выполнение неравенства (3.35).*

Заметим, что неравенство (3.35) интуитивно очевидно: для существования стационарного распределения ЦМ, описывающей рассматриваемую систему, необходимо и достаточно, чтобы средняя скорость обслуживания запросов была больше средней скорости их поступления.

Предполагая неравенство (3.35) выполненным, стационарное распределение вероятностей состояний рассматриваемой СМО находим в виде (3.25), где матрица $\mathcal{R}$ находится как решение уравнения (3.26), а вектор $\boldsymbol{\pi}_0$ – как решение системы линейных алгебраических уравнений (3.27), (3.28).

Далее найдем стационарное распределение виртуального времени ожидания произвольного запроса в рассматриваемой системе в терминах преобразования Лапласа – Стилтьеса. Пусть $W(x)$ – функция стационарного распределения времени ожидания, $w(s) = \int\limits_0^\infty e^{-sx} dW(x), \mathrm{Re}\ s \geq 0$, – ее ПЛС. Известно, что $w(s)$ при действительном s можно трактовать, как вероятность ненаступления за время ожидания катастроф из некоторого воображаемого стационарного пуассоновского потока катастроф с параметром s. Соответственно, вероятность ненаступления катастроф за время обслуживания, имеющее распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}, S)$, есть $\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0$. m-ая компонента $((sI-S)^{-1}\mathbf{S}_0)_m$ векторной функции $(sI-S)^{-1}\mathbf{S}_0$ задает вероятность ненаступления катастроф за остаточное время обслуживания при условии, что в данный момент управляющий процесс обслуживания находится в состоянии m, $m = \overline{1, M}$.

Учитывая такую вероятностную интерпретацию ПЛС, из формулы полной вероятности получаем формулу:

$$w(s) = \boldsymbol{\pi}_0\mathbf{e} + \sum_{i=1}^{\infty} \boldsymbol{\pi}_i(\mathbf{e}_{\bar{W}} \otimes I_M)(sI - S)^{-1}\mathbf{S}_0(\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0)^{i-1} =$$

$$= \boldsymbol{\pi}_0\mathbf{e} + \boldsymbol{\pi}_0\mathcal{R}(I - \mathcal{R}\beta(s))^{-1}(\mathbf{e}_{\bar{W}} \otimes I_M)(sI - S)^{-1}\mathbf{S}_0,$$

где

$$\beta(s) = \boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0.$$

В случае, если распределение времени обслуживания – экспоненциальное с параметром μ, эта формула упрощается:

$$w(s) = \boldsymbol{\pi}_0 \sum_{i=0}^{\infty} \mathcal{R}^i \mathbf{e}(\frac{\mu}{\mu + s})^i = \boldsymbol{\pi}_0(I - \mathcal{R}\frac{\mu}{\mu + s})^{-1}\mathbf{e}.$$

Отметим, что существует еще один подход (так называемый спектральный подход) к решению проблемы вычисления векторов $\boldsymbol{\pi}_i$, $i \geq 0$, стационарных вероятностей состояний векторного процесса гибели и размножения ξ_t, $t \geq 0$. Кратко изложим сущность этого подхода.

3.2.3 Спектральный подход для анализа векторного процесса гибели и размножения

Рассмотрим матричный полином

$$Q(\lambda) = Q_0 + Q_1\lambda + Q_2\lambda^2,$$

где квадратные матрицы $Q_0,\ Q_1,\ Q_2$ порядка N являются элементами блочного генератора (3.22). Пусть λ_k – собственные числа этого полинома, то есть λ_k являются корнями уравнения

$$\det Q(\lambda) = 0. \tag{3.36}$$

При выполнении условия эргодичности рассматриваемого процесса существует ровно N корней уравнения (3.36), лежащих в единичном круге комплексной плоскости. Будем рассматривать только эти корни. Для простоты изложения предположим, что все эти корни – простые. Обозначим через $\boldsymbol{\psi}_k$ левый собственный вектор полинома $Q(\lambda)$, соответствующий собственному числу λ_k, то есть вектор $\boldsymbol{\psi}_k$ удовлетворяет уравнению

$$\boldsymbol{\psi}_k Q(\lambda_k) = \mathbf{0}. \tag{3.37}$$

Система уравнений равновесия для векторов $\boldsymbol{\pi}_i,\ i \geq 0$, стационарных вероятностей векторного процесса гибели и размножения имеет вид:

$$\boldsymbol{\pi}_0 Q_{0,0} + \boldsymbol{\pi}_1 Q_2 = \mathbf{0}, \tag{3.38}$$

$$\boldsymbol{\pi}_{j-1} Q_0 + \boldsymbol{\pi}_j Q_1 + \boldsymbol{\pi}_{j+1} Q_2 = \mathbf{0},\ j \geq 1. \tag{3.39}$$

Прямой подстановкой в уравнения системы (3.39) с учетом (3.37) можно убедиться, что решением системы будет множество векторов

$$\boldsymbol{\pi}_j = \sum_{k=1}^{N} x_k \boldsymbol{\psi}_k \lambda_k^j,\quad j \geq 0, \tag{3.40}$$

где $x_k, k = \overline{1, N}$, – некоторые константы. Отметим, что некоторые корни λ_k могут быть комплексными, а не действительными. При этом корнем уравнения (3.36) является и сопряженное к λ_k число. Соответствующие собственные векторы $\boldsymbol{\psi}_k$ тоже будут комплексными и сопряженными. При

этом комплексными будут и константы x_k, но векторы $\boldsymbol{\pi}_j$, вычисленные по формулам (3.40), будут действительными.

Для нахождения констант $x_k, k = \overline{1,N}$, используем систему уравнений (3.38) и условие нормировки.

Таким образом, мы убеждаемся в справедливости следующего утверждения.

Теорема 3.4. *Векторы стационарных вероятностей состояний ЦМ* $\xi_t,\ t \geq 0$, *вычисляются по формуле (3.40), где константы* $x_k, k = \overline{1,N}$, *являются единственным решением системы линейных алгебраических уравнений*

$$\sum_{k=1}^{N} x_k \boldsymbol{\psi}_k (Q_{0,0} + Q_2 \lambda_k) = \mathbf{0},$$

$$\sum_{k=1}^{N} x_k \boldsymbol{\psi}_k \mathbf{e} \frac{1}{1-\lambda_k} = 1.$$

Описание и примеры использования спектрального подхода для исследования СМО и обсуждение его достоинств и недостатков по сравнению с подходом, заданным теоремой 3.2, можно найти в [161].

3.3 ЦЕПИ МАРКОВА ТИПА $G/M/1$

Многомерные процессы гибели и размножения, описанные в предыдущем разделе, являются простейшим и наиболее хорошо изученным случаем многомерных ЦМ типа $G/M/1$ и $M/G/1$. В данном и следующем разделах рассматриваются такие цепи. В отличие от многомерных процессов гибели и размножения, которые являются ЦМ с непрерывным временем, описание ЦМ типа $G/M/1$ и $M/G/1$ и результаты их исследования будет приведено для случая дискретного времени.

3.3.1 Определение цепи Маркова типа $G/M/1$ и ее стационарное распределение

Рассмотрим многомерный марковский процесс $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, с пространством состояний

$$\mathcal{S} = \{(i, \mathbf{v}), \mathbf{v} \in \mathcal{V}_i,\ i \geq 0\},$$

где $\mathcal{V}_i$ – некоторое конечное множество, если марковский процесс $\xi_n, n \geq 1$, двумерный, или некоторое конечное множество конечномерных векторов, если марковский процесс $\xi_n, n \geq 1$, многомерный.

Перенумеруем состояния процесса $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, в лексикографическом порядке и объединим состояния, имеющие значение i компоненты i_n, в макросостояние i, иногда называемое уровнем процесса $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$. Соответственно этой нумерации, матрица P вероятностей одношаговых переходов этого процесса запишется в блочном виде $P = (P_{i,l})_{i,l \geq 0}$, где $P_{i,l}$ есть матрица, образованная вероятностями одношаговых переходов $p_{(i,\mathbf{r});(l,\boldsymbol{\nu})}$ перехода ЦМ $\xi_n, n \geq 1$, из состояния $(i, \mathbf{r}),\ \mathbf{r} \in \mathcal{V}_i$, в состояние $(l, \boldsymbol{\nu}), \boldsymbol{\nu} \in \mathcal{V}_l$.

Процесс $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, называется ЦМ типа $G/M/1$, если матрица P вероятностей одношаговых переходов имеет следующую структуру:

$$P = \begin{pmatrix} B_0 & A_0 & O & O & O & \dots \\ B_1 & A_1 & A_0 & O & O & \dots \\ B_2 & A_2 & A_1 & A_0 & O & \dots \\ B_3 & A_3 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{3.41}$$

Из вида (3.41) матрицы P следует, что множества $\mathcal{V}_i$ для различных значений $i,\ i > 0$, здесь совпадают. Далее будем использовать обозначение $\mathcal{V}_i = \mathcal{V},\ i \geq 0$.

Название этих цепей объясняется тем, что структуру, аналогичную (3.41), имеет матрица переходных вероятностей вложенной по моментам прихода запросов ЦМ для СМО $G/M/1$. Отличие (3.41) от такой матрицы состоит в том, что элементы матрицы (3.41) являются не числами, а матрицами.

Матрицу, имеющую структуру (3.41), называют блочной нижнехессенберговой. Величина скачка компоненты i_n за один шаг вправо не превосходит единицы. В англоязычной литературе это свойство называется skip-free to the left.

Предположим, что ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, с матрицей переходных вероятностей P является неприводимой и апериодической, а матрицы $A = \sum\limits_{i=0}^{\infty} A_i$ и $B = \sum\limits_{i=0}^{\infty} B_i$ являются стохастическими и неприводимыми.

Обозначим через $\boldsymbol{\delta}$ вектор-строку, являющуюся единственным решением системы уравнений

$$\boldsymbol{\delta} = \boldsymbol{\delta} A,\ \boldsymbol{\delta}\mathbf{e} = 1,$$

а через $\boldsymbol{\gamma}$ – вектор-столбец $\boldsymbol{\gamma} = \sum\limits_{i=1}^{\infty} iA_i\mathbf{e}$.

Теорема 3.5. *Для того чтобы ЦМ* $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, *имела стационарное распределение, необходимо и достаточно, чтобы выполнялось неравенство*

$$\boldsymbol{\delta}\boldsymbol{\gamma} > 1. \tag{3.42}$$

Доказательство теоремы приведено в [163]. Далее считаем условие (3.42) выполненным. Тогда существует стационарное распределение вероятностей ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$:

$$\pi(i, \mathbf{v}) = \lim_{n\to\infty} P\{i_n = i, \mathbf{v}_n = \mathbf{v}\},\ \mathbf{v} \in \mathcal{V}_i,\ i \geq 0.$$

Упорядочим стационарные вероятности для каждого i в лексикографическом порядке и введем векторы

$$\boldsymbol{\pi}_i = (\pi(i, \mathbf{v}),\ \mathbf{v} \in \mathcal{V}),\ i \geq 0.$$

Теорема 3.6. *Векторы стационарных вероятностей состояний ЦМ* $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, *вычисляются следующим образом:*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0\mathcal{R}^i,\ i \geq 0, \tag{3.43}$$

где матрица $\mathcal{R}$ *является минимальным неотрицательным решением матричного уравнения*

$$\mathcal{R} = \sum_{j=0}^{\infty} \mathcal{R}^j A_j, \tag{3.44}$$

а вектор $\boldsymbol{\pi}_0$ *является единственным решением следующей системы линейных алгебраических уравнений*

$$\boldsymbol{\pi}_0 \sum_{j=0}^{\infty} \mathcal{R}^j B_j = \boldsymbol{\pi}_0, \tag{3.45}$$

$$\boldsymbol{\pi}_0 (I - \mathcal{R})^{-1}\mathbf{e} = 1. \tag{3.46}$$

Доказательство. Вектор $\boldsymbol{\pi} = (\boldsymbol{\pi}_0,\ \boldsymbol{\pi}_1,\ \boldsymbol{\pi}_2, \dots)$ стационарного распределения является решением системы уравнений равновесия

$$\boldsymbol{\pi} P = \boldsymbol{\pi}, \tag{3.47}$$

дополненной условием нормировки

$$\boldsymbol{\pi}\mathbf{e} = 1. \tag{3.48}$$

Предположим, что решение системы (3.47) имеет вид (3.43). Подставляя векторы $\boldsymbol{\pi}_i,\ i \geq 1$, в форме (3.43) в уравнения

$$\boldsymbol{\pi}_j = \sum_{i=j-1}^{\infty} \boldsymbol{\pi}_i A_{i-j+1},\ j \geq 1,$$

системы (3.47), легко убедиться, что эти уравнения обращаются в тождества, если матрица $\mathcal{R}$ удовлетворяет матричному уравнению (3.44). Подставляя векторы $\boldsymbol{\pi}_i,\ i \geq 1$, в форме (3.43) в уравнение

$$\boldsymbol{\pi}_0 = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i B_i \tag{3.49}$$

системы (3.47), получаем, что вектор $\boldsymbol{\pi}_0$ удовлетворяет уравнению (3.45).

Известно, см. [163], что если условие эргодичности (3.41) выполняется, то спектральный радиус $\varrho(\mathcal{R})$ матрицы $\mathcal{R}$ строго меньше единицы. При выполнении этого условия ряд $\sum\limits_{l=0}^{\infty} \mathcal{R}^l$ сходится и равен $(I-\mathcal{R})^{-1}$. Уравнение (3.46) теперь следует из (3.43) и условия нормировки $\sum\limits_{i=0}^{\infty} \boldsymbol{\pi}_i \mathbf{e} = 1$. □

Нелинейное матричное уравнение (3.44) для матрицы $\mathcal{R}$ можно решить с помощью метода последовательных приближений. Взяв в качестве начального приближения матрицу $\mathcal{R}_0 = O$, последующие приближения можно вычислить по формуле

$$\mathcal{R}_{k+1} = \sum_{j=0}^{\infty} \mathcal{R}_k^j A_j,\ k \geq 0,$$

или

$$\mathcal{R}_{k+1} = \sum_{j=0,\, j\neq 1}^{\infty} \mathcal{R}_k^j A_j (I - A_1)^{-1},\ k \geq 0.$$

В [163] доказано, что последовательные приближения, вычисленные в этих итерационных процедурах, образуют последовательность матриц с монотонно возрастающими элементами, сходящуюся к минимальному неотрицательному решению уравнения (3.44).

Изложенные выше результаты для ЦМ типа $G/M/1$ легко переносятся на случай, когда блочная матрица вероятностей переходов имеет вид

$$P = \begin{pmatrix} B_0 & C_0 & O & O & O & \dots \\ B_1 & A_1 & A_0 & O & O & \dots \\ B_2 & A_2 & A_1 & A_0 & O & \dots \\ B_3 & A_3 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Заметим, что здесь множества $\mathcal{V}_i$ совпадают для значений $i \geq 1$. Множество же $\mathcal{V}_0$ может отличаться как по составу, так и по мощности.

В этом случае необходимое и достаточное условие существования стационарного распределения процесса $\xi_n,\ n \geq 1$, также задается неравенством (3.41), дополненным условием, что стохастическая матрица

$$\mathcal{T} = \begin{pmatrix} B_0 & C_0 \\ \sum\limits_{j=1}^{\infty} \mathcal{R}^{j-1} B_j & \sum\limits_{j=1}^{\infty} \mathcal{R}^{j-1} A_j \end{pmatrix}$$

имеет положительный левый собственный вектор.

Формулы для расчета векторов стационарных вероятностей здесь имеют следующий вид:

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_1 \mathcal{R}^{i-1},\ i \geq 1,$$

где матрица $\mathcal{R}$ является минимальным неотрицательным решением матричного уравнения (3.44), а векторы $\boldsymbol{\pi}_0$ и $\boldsymbol{\pi}_1$ являются единственным решением следующей системы линейных алгебраических уравнений

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1)\mathcal{T} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1),$$

$$\boldsymbol{\pi}_0 \mathbf{e} + \boldsymbol{\pi}_1 (I - \mathcal{R})^{-1} \mathbf{e} = 1.$$

Аналогичным образом приведенные выше результаты для ЦМ типа $G/M/1$ переносятся на случай, когда матрица переходных вероятностей процесса имеет вид

$$P = \begin{pmatrix} B_0^{(0)} & \dots & B_0^{(J)} & C_0 & O & O & O & \dots \\ B_1^{(0)} & \dots & B_1^{(J)} & A_1 & A_0 & O & O & \dots \\ B_2^{(0)} & \dots & B_2^{(J)} & A_2 & A_1 & A_0 & O & \dots \\ B_3^{(0)} & \dots & B_3^{(J)} & A_3 & A_2 & A_1 & A_0 & \dots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где J – некоторое конечное целое число.

Для иллюстрации применим изложенные выше сведения для ЦМ типа $G/M/1$ к исследованию СМО $G/PH/1$.

3.3.2 Применение результатов для исследования системы обслуживания $G/PH/1$

Имеется однолинейная СМО с бесконечным буфером. Входной поток является рекуррентным с ФР $H(t)$ длин интервалов между моментами поступления запросов. Интенсивность потока обозначим $\lambda = (\int\limits_0^\infty t dH(t))^{-1}$.

Время обслуживания имеет распределение фазового типа, которое задается ЦМ с непрерывным временем $\eta_t,\ t \geq 0$, с пространством непоглощающих состояний $\{1, \ldots, M\}$ и неприводимым представлением $(\boldsymbol{\beta}, S)$, где $\boldsymbol{\beta}$ – стохастический вектор-строка, а S – субгенератор. Обозначим через μ среднюю интенсивность обслуживания: $\mu^{-1} = b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$.

Процесс $i_t,\ t \geq 0$, – число запросов в системе в момент t – не является марковским. Для его исследования применим метод вложенных ЦМ, см., например, [77]. С этой целью вначале будем рассматривать поведение СМО в моменты поступления запросов в систему.

Пусть t_n – момент поступления в систему n-го запроса, $n \geq 1$. Несложно убедиться, что двумерный процесс $\xi_n = \{i_{t_n-0}, \eta_{t_n+0}\},\ n \geq 1$, где η_{t_n+0} – состояние управляющего процесса обслуживания в момент t_n+0, является ЦМ, $i_{t_n-0} \geq 0,\ \eta_{t_n+0} = \overline{1, M}$.

Пусть $P(k,u),\ k \geq 0$, есть матрица, элемент $(P(k,u))_{\eta,\eta'}$ которой есть вероятность того, что за интервал времени $[0,u)$ поступит k событий в рекуррентном потоке событий с распределением интервалов между моментами поступления фазового типа с неприводимым представлением $(\boldsymbol{\beta}, S)$ и состояние управляющего процесса $\eta_t,\ t \geq 0$, поступления событий в момент u будет η' при условии, что в момент 0 оно было η. Иначе говоря, элементы матрицы $P(k,u)$ задают вероятность обслуживания k запросов и соответствующих переходов состояния управляющего процесса обслуживания за время u при условии, что процесс обслуживания в интервале $[0,u)$ не прерывался из-за отсутствия запросов в системе.

Поскольку процесс наступления событий в рекуррентном потоке с распределением интервалов между моментами поступления фазового типа с неприводимым представлением $(\boldsymbol{\beta}, S)$ является частным случаем MAP-

потока, матрицы $P(k,u),\ k \geq 0$, могут быть найдены из разложения:

$$\sum_{k=0}^{\infty} P(k,u)z^k = e^{(S+\mathbf{S}_0\boldsymbol{\beta}z)u},\ u \geq 0,\ |z| \leq 1.$$

Тогда элементы матрицы $\Omega_k = \int\limits_0^\infty P(k,u)dH(u),\ k \geq 0$, задают вероятность обслуживания k запросов и соответствующих переходов управляющего процесса обслуживания за время между двумя последовательными моментами поступления запросов в систему при условии, что процесс обслуживания в интервале не прерывался из-за отсутствия запросов в системе. Обозначим

$$\Omega(z) = \sum_{k=0}^{\infty} \Omega_k z^k = \int\limits_0^\infty e^{(S+\mathbf{S}_0\boldsymbol{\beta}z)u}dH(u).$$

Перенумеруем состояния ЦМ $\xi_n, n \geq 1$, в лексикографическом порядке и объединим состояния, имеющие значение i компоненты i_n в макросостояние i. Соответственно этой нумерации, матрица P вероятностей одношаговых переходов этого процесса запишется в блочной форме $P = (P_{i,l})_{i,l\geq 0}$, где $P_{i,l}$ есть матрица, образованная вероятностями одношаговых переходов ЦМ $\xi_n, n \geq 1$, из состояния (i,η) в состояние $(l,\eta'),\ \eta,\eta' = \overline{1,M}$.

Нетрудно убедиться, что матрицы $P_{i,l},\ i,\ l \geq 0$, задаются следующим образом:

$$P_{i,l} = \begin{cases} O, & l > i+1, \\ \Omega_{i+1-l}, & 0 < l \leq i+1, \\ \sum\limits_{k=i+1}^{\infty} \Omega_k \mathbf{e}\boldsymbol{\beta}, & l = 0. \end{cases}$$

Поэтому ЦМ $\xi_n = \{i_{t_n-0}, \eta_{t_n+0}\},\ n \geq 1$, является ЦМ типа $G/M/1$ с матрицей переходных вероятностей вида (3.41) с блоками

$$A_j = \Omega_j,\ B_j = \sum_{k=j+1}^{\infty} \Omega_k \mathbf{e}\boldsymbol{\beta},\ j \geq 0.$$

Теорема 3.7. *Стационарное распределение вероятностей рассматриваемой системы существует тогда и только тогда, когда выполняется неравенство:*

$$\lambda < \mu. \tag{3.50}$$

Доказательство. Из теоремы 3.1 следует, что необходимым и достаточным условием существования стационарного распределения вероятностей рассматриваемой системы является выполнение неравенства типа (3.42), где вектор $\boldsymbol{\delta}$ является единственным решением уравнения

$$\boldsymbol{\delta} = \boldsymbol{\delta}\Omega(1),\ \boldsymbol{\delta}\mathbf{e} = 1,$$

а $\boldsymbol{\gamma} = \frac{d\Omega(z)}{dz}|_{z=1}\mathbf{e}$. Непосредственной подстановкой можно убедиться, что для ЦМ, описывающей рассматриваемую СМО, вектор $\boldsymbol{\delta}$ имеет вид

$$\boldsymbol{\delta} = \mu\boldsymbol{\beta}(-S)^{-1}.$$

Учитывая, что $\boldsymbol{\delta}(S+\mathbf{S}_0\boldsymbol{\beta}) = \mathbf{0}$ и $(S+\mathbf{S}_0\boldsymbol{\beta})\mathbf{e} = \mathbf{0}$, по аналогии с доказательством формулы (1.16) можно показать, что неравенство (3.42) переходит в неравенство $\mu\lambda^{-1} > 1$, что эквивалентно (3.50). □

В предположении, что условие существования стационарного распределения (3.50) выполняется, стационарное распределение ищется в матрично-геометрическом виде по алгоритму, заданному Теоремой 3.2. Наличие связи

$$B_j = \sum_{k=j+1}^{\infty} A_k\mathbf{e}\boldsymbol{\beta},\ j \geq 0, \tag{3.51}$$

между матрицами B_j и A_j позволяет получить для векторов стационарных вероятностей более простые формулы, чем формулы для произвольной ЦМ типа $G/M/1$. А именно, вектор $\boldsymbol{\pi}_0$, который в общем случае ищется численно как решение системы линейных алгебраических уравнений (3.45), (3.46), здесь находится в явном виде:

$$\boldsymbol{\pi}_0 = c\boldsymbol{\beta}, \tag{3.52}$$

$$c = [\boldsymbol{\beta}(I-\mathcal{R})^{-1}\mathbf{e}]^{-1}. \tag{3.53}$$

Действительно, подставляя выражение (3.51) в уравнение (3.49), получаем:

$$\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0\sum_{j=0}^{\infty}\mathcal{R}^j B_j = \boldsymbol{\pi}_0\sum_{j=0}^{\infty}\mathcal{R}^j\sum_{k=j+1}^{\infty}\Omega_k\mathbf{e}\boldsymbol{\beta} = \boldsymbol{\pi}_0\sum_{k=1}^{\infty}\sum_{j=0}^{k-1}\mathcal{R}^j\Omega_k\mathbf{e}\boldsymbol{\beta} =$$

$$= \boldsymbol{\pi}_0\sum_{k=1}^{\infty}(I-\mathcal{R})^{-1}(I-\mathcal{R}^k)\Omega_k\mathbf{e}\boldsymbol{\beta} = \boldsymbol{\pi}_0(I-\mathcal{R})^{-1}(\Omega(1)-\sum_{k=0}^{\infty}\mathcal{R}^k\Omega_k)\mathbf{e}\boldsymbol{\beta} =$$

$$= \boldsymbol{\pi}_0(I-\mathcal{R})^{-1}(\Omega(1)-\mathcal{R})\mathbf{e}\boldsymbol{\beta} = \boldsymbol{\pi}_0\mathbf{e}\boldsymbol{\beta}.$$

Решая полученное уравнение $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0 \mathbf{e}\boldsymbol{\beta}$ с условием нормировки (3.46), получаем формулы (3.52), (3.53).

Следствие 3.1. *Среднее число L_1 запросов в системе непосредственно перед моментами поступления запросов вычисляется по формуле*

$$L_1 = c\boldsymbol{\beta}(I - \mathcal{R})^{-2}\mathbf{e} - 1.$$

Рассмотрим теперь проблему нахождения стационарного распределения вероятностей состояний СМО в произвольный момент времени.

Изучим процесс $\{i_t,\ \eta_t\},\ t \geq 0$, где i_t – число запросов в системе, η_t – состояние управляющего процесса обслуживания в момент времени t, $i_t \geq 0,\ \eta_t = \overline{1, M}$.

Обозначим

$$p(0) = \lim_{t\to\infty} P\{i_t = 0\},\ p(i,\eta) = \lim_{t\to\infty} P\{i_t = i, \eta_t = \eta\},\ i \geq 1,\ \eta = \overline{1, M},$$

$$\mathbf{p}_i = (p(i,1), p(i,2), \ldots, p(i,M)),\ i \geq 1.$$

Можно показать, что стационарная вероятность $p(0)$ и векторы стационарных вероятностей $\mathbf{p}_i,\ i \geq 1$, существуют при выполнении условия (3.50) существования стационарного распределения вероятностей вложенной ЦМ $\xi_n, n \geq 1$.

Теорема 3.8. *Стационарная вероятность $p(0)$ и векторы стационарных вероятностей $\mathbf{p}_i,\ i \geq 1$, вычисляются следующим образом:*

$$p(0) = 1 - \rho, \tag{3.54}$$

$$\mathbf{p}_i = c\lambda\boldsymbol{\beta}\mathcal{R}^{i-1}(\mathcal{R} - I - \mathcal{R}\mathbf{e}\boldsymbol{\beta})S^{-1},\ i \geq 1, \tag{3.55}$$

где $\rho = \frac{\lambda}{\mu}$ – коэффициент загрузки системы, а константа c задана формулой (3.53).

Доказательство. Техника перехода к распределению вероятностей состояний процесса в произвольный момент времени через распределение вероятностей состояний этого процесса в некоторые вложенные моменты времени будет кратко описана в разделе 4.1 и может быть использована для доказательства данной теоремы. Однако здесь можно обойтись и более простым математическим аппаратом. Из теории рекуррентных потоков известно, что если имеется стационарный рекуррентный поток запросов с функцией распределения $H(t)$ длин интервалов между моментами поступления запросов и если выбран произвольный момент времени, то время с

момента поступления последнего запроса до произвольного момента времени имеет ФР $\tilde{H}(t) = \lambda \int\limits_0^t (1 - H(u))du$, где λ – интенсивность потока, вычисляемая как $\lambda = (\int\limits_0^\infty (1 - H(u))du)^{-1}$.

Используя эти сведения и формулу полной вероятности, получаем следующую связь между распределениями вероятностей состояний рассматриваемой СМО в произвольные и вложенные моменты:

$$\mathbf{p}_i = \sum_{j=i-1}^{\infty} \boldsymbol{\pi}_j \lambda \int\limits_0^\infty P(j+1-i,t)(1 - H(t))dt,\ i \geq 1.$$

Учитывая формулу (3.43), это соотношение можно переписать в виде:

$$\mathbf{p}_i = \boldsymbol{\pi}_0 \mathcal{R}^{i-1} \lambda \sum_{m=0}^{\infty} \mathcal{R}^m \int\limits_0^\infty P(m,t)(1 - H(t))dt,\ i \geq 1. \tag{3.56}$$

Обозначим

$$\Psi(m) = \lambda \int\limits_0^\infty P(m,t)(1 - H(t))dt,\ m \geq 0,$$

$$\Psi(\mathcal{R}) = \sum_{m=0}^{\infty} \mathcal{R}^m \Psi(m) = \lambda \sum_{m=0}^{\infty} \mathcal{R}^m \int\limits_0^\infty P(m,t)(1 - H(t))dt$$

и покажем, что

$$\Psi(\mathcal{R}) = \lambda(\mathcal{R} - \mathcal{R}\mathbf{e}\boldsymbol{\beta} - I)S^{-1}. \tag{3.57}$$

Для этого привлечем введенные ранее матрицы $\Omega_m = \int\limits_0^\infty P(m,t)dH(t)$, $m \geq 0$, и систему (3.2) матричных дифференциальных уравнений для матриц $P(n,t),\ n \geq 0$. Поскольку при отсутствии интервалов простоя прибора процесс обслуживания является частным случаем MAP-потока с матрицами $D_0 = S,\ D_1 = \mathbf{S}_0\boldsymbol{\beta}$, из системы (3.2) следует, что матричные функции $P(n,t),\ n \geq 0$, удовлетворяют системе дифференциальных уравнений

$$\frac{dP(0,t)}{dt} = P(0,t)S,\ \frac{dP(n,t)}{dt} = P(n,t)S + P(n-1,t)\mathbf{S}_0\boldsymbol{\beta},\ n \geq 1. \tag{3.58}$$

Используя формулу интегрирования по частям, имеем:

$$\Omega_m = \int\limits_0^\infty P(m,t)dH(t) = -\int\limits_0^\infty P(m,t)d(1-H(t)) =$$

$$= -P(m,t)(1-H(t))|_0^\infty + \int\limits_0^\infty \frac{dP(m,t)}{dt}(1-H(t)).$$

С учетом введенных обозначений и системы (3.58) отсюда получаем:

$$\Omega_m = \delta_{m,0}I + \lambda^{-1}\Psi(m)S + (1-\delta_{m,0})\lambda^{-1}\Psi(m-1)\mathbf{S}_0\boldsymbol{\beta},\ m \geq 0.$$

Переходя здесь (путем умножения слева на матрицу $\mathcal{R}^m$ и суммирования по m) к матричным производящим функциям и учитывая, что в силу (3.44) $\sum\limits_{m=0}^{\infty}\mathcal{R}^m\Omega_m = \mathcal{R}$, получаем:

$$\mathcal{R} = I + \lambda^{-1}\Psi(\mathcal{R})S + \lambda^{-1}\mathcal{R}\Psi(\mathcal{R})\mathbf{S}_0\boldsymbol{\beta}. \tag{3.59}$$

Умножая соотношение (3.59) на вектор $\mathbf{e}$, с учетом стохастичности вектора $\boldsymbol{\beta}$ и связи матрицы S с вектором-столбцом $\mathbf{S}_0$, получаем:

$$(I-\mathcal{R})\mathbf{e} = \lambda^{-1}(I-\mathcal{R})\Psi(\mathcal{R})\mathbf{S}_0.$$

Так как спектральный радиус $\varrho(\mathcal{R})$ матрицы $\mathcal{R}$ строго меньше единицы, существует матрица $(I-\mathcal{R})^{-1}$, поэтому справедливы соотношения

$$\mathbf{e} = \lambda^{-1}\Psi(\mathcal{R})\mathbf{S}_0$$

и

$$\mathbf{e}\boldsymbol{\beta} = \lambda^{-1}\Psi(\mathcal{R})\mathbf{S}_0\boldsymbol{\beta}.$$

Подставляя последнее выражение в (3.59), получаем выражение

$$\lambda^{-1}\Psi(\mathcal{R})S = \mathcal{R} - I - \mathcal{R}\mathbf{e}\boldsymbol{\beta},$$

откуда, учитывая невырожденность субгенератора S, и получаем формулу (3.57). Формула (3.55) очевидным образом следует из формул (3.56) и (3.57).

Для доказательства формулы (3.54) используем условие нормировки. Согласно ему

$$p(0) = 1 - \sum_{i=1}^{\infty}\mathbf{p}_i\mathbf{e}.$$

Используя формулы (3.52) и (3.54), отсюда получаем:

$$p(0) = 1 - c\lambda\boldsymbol{\beta}(I - \mathcal{R})^{-1}(\mathcal{R} - \mathcal{R}\mathbf{e}\boldsymbol{\beta} - I)S^{-1}\mathbf{e}.$$

Учитывая, что среднее время обслуживания b_1 задается формулой $b_1 = \boldsymbol{\beta}(-S)^{-1}\mathbf{e}$, и формулу (3.53), получаем:

$$p(0) = 1 + c\lambda b_1(-1 - \boldsymbol{\beta}(I - \mathcal{R})^{-1}\mathcal{R}\mathbf{e}) =$$

$$= 1 - \lambda b_1(\boldsymbol{\beta}(I - \mathcal{R})^{-1}\mathbf{e})^{-1}(1 + \boldsymbol{\beta}(I - \mathcal{R})^{-1}\mathcal{R}\mathbf{e}) = 1 - \lambda b_1.$$

Здесь учтено, что

$$1 + \boldsymbol{\beta}(I - \mathcal{R})^{-1}\mathcal{R}\mathbf{e} = 1 + \boldsymbol{\beta}(I - \mathcal{R})^{-1}(\mathcal{R} - I + I)\mathbf{e} = \boldsymbol{\beta}(I - \mathcal{R})^{-1}\mathbf{e}.$$

□

Следствие 3.2. *Среднее число запросов L_2 в системе в произвольный момент времени вычисляется по формуле*

$$L_2 = \rho L_1 - c\lambda\boldsymbol{\beta}(I - \mathcal{R})^{-1}S^{-1}\mathbf{e}.$$

Рассмотрим теперь распределение времени ожидания в системе. Пусть $W(t)$ – функция стационарного распределения времени ожидания произвольного запроса, $w(s) = \int\limits_0^\infty e^{-st}dW(t),\ Re\ s \geq 0.$

Учитывая вероятностную интерпретацию ПЛС, из формулы полной вероятности получаем формулу:

$$w(s) = \boldsymbol{\pi}_0\mathbf{e} + \sum_{i=1}^{\infty}\boldsymbol{\pi}_i(sI - S)^{-1}\mathbf{S}_0(\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0)^{i-1}.$$

Подставляя сюда выражения для $\boldsymbol{\pi}_i$ из (3.54), (3.55), приходим к выражению

$$w(s) = c\boldsymbol{\beta}\left(\mathbf{e} + \sum_{i=1}^{\infty}\mathcal{R}^i(sI - S)^{-1}\mathbf{S}_0(\boldsymbol{\beta}(sI - S)^{-1}\mathbf{S}_0)^{i-1}\right).$$

Следствие 3.3. *Среднее время ожидания V_{wait} произвольного запроса вычисляется по формуле*

$$V_{wait} = \lambda^{-1}L_2 - b_1.$$

3.4 ЦЕПИ МАРКОВА ТИПА $M/G/1$

3.4.1 Определение цепи Маркова типа $M/G/1$ и критерий эргодичности

Рассмотрим многомерный марковский процесс $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, с пространством состояний

$$\mathcal{S} = \{(i, \mathbf{v}), \mathbf{v} \in \mathcal{V}_i,\ i \geq 0\},$$

где $\mathcal{V}_i$ – некоторое конечное множество, если марковский процесс $\xi_n, n \geq 1$, двумерный, или некоторое конечное множество конечномерных векторов, если марковский процесс $\xi_n, n \geq 1$, многомерный.

Объединим состояния ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, имеющие значение i компоненты i_n, в макросостояние i, процесса $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, и перенумеруем уровни в лексикографическом порядке. Состояния внутри уровня могут быть перенумерованы в произвольном порядке. Без ограничения общности будем считать, что это также лексикографический порядок. Соответственно этой нумерации матрица P вероятностей одношаговых переходов этого процесса запишется в блочной форме $P = (P_{i,l})_{i,l \geq 0}$, где $P_{i,l}$ есть матрица, образованная вероятностями одношаговых вероятностей $p_{(i,\mathbf{r});(l,\boldsymbol{\nu})}$ переходов ЦМ $\xi_n, n \geq 1$, из состояния $(i, \mathbf{r}), \mathbf{r} \in \mathcal{V}_i$, в состояние $(l, \boldsymbol{\nu}), \boldsymbol{\nu} \in \mathcal{V}_l$.

Процесс $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, называется ЦМ типа $M/G/1$, или многомерной квазитеплицевой цепью Маркова (КТЦМ), если матрица P вероятностей одношаговых переходов имеет следующую структуру:

$$P = \begin{pmatrix} V_0 & V_1 & V_2 & V_3 & \dots \\ Y_0 & Y_1 & Y_2 & Y_3 & \dots \\ O & Y_0 & Y_1 & Y_2 & \dots \\ O & O & Y_0 & Y_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{3.60}$$

Из структуры матрицы P видно, что все множества $\mathcal{V}_i$, $i \geq 1$, совпадают. В дальнейшем будем предполагать, что мощность этих множеств равна K.

Название цепи объясняется тем, что структуру, аналогичную (3.60), имеет квазитеплицева матрица переходных вероятностей вложенной по моментам окончания обслуживания запросов ЦМ для СМО $M/G/1$. Отличие

(3.60) от такой матрицы состоит в том, что элементы матрицы (3.60) являются не числами, а матрицами.

Матрицу, имеющую структуру (3.60), называют блочной верхнехессенберговой. Величина скачка компоненты i_n за один шаг влево не превосходит единицы. В англоязычной литературе это свойство называется skip-free to the left.

Введем матричные ПФ $Y(z) = \sum\limits_{i=0}^{\infty} Y_i z^i$ и $V(z) = \sum\limits_{i=0}^{\infty} V_i z^i,\ |z| \le 1$.

Теорема 3.9. *Пусть ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \ge 1$, с матрицей переходных вероятностей P является неприводимой и апериодической, матрицы $Y(1)$ и $V(1)$ являются стохастическими и неприводимыми и выполняются неравенства $Y^{'}(1) < \infty$ и $V^{'}(1) < \infty$. Для того чтобы ЦМ ξ_n, $n \ge 1$, была эргодичной, необходимо и достаточно, чтобы выполнялось неравенство*

$$[\det(zI - Y(z))]'_{z=1} > 0. \tag{3.61}$$

Доказательство. Для нахождения условий эргодичности ЦМ часто используют понятие среднего одношагового сдвига. Для одномерной ЦМ $i_n,\ i_n \ge 0,\ n \ge 1$, и произвольной неотрицательной тестовой функции x_i, заданной на пространстве состояний цепи $i_n,\ n \ge 1$, средний одношаговый x-сдвиг γ_i^x цепи i_n в состоянии i определяется следующим образом:

$$\gamma_i^x = E(x_{i_{n+1}} - x_{i_n} | i_n = i), i \ge 0,$$

где E – символ математического ожидания или, если $p_{ij},\ i,\ j \ge 0$ – переходные вероятности цепи, то

$$\gamma_i^x = \sum_{j=0}^{\infty} p_{ij} x_j - x_i,\ i \ge 0.$$

Тогда в качестве аналогов величин $x_i,\ \gamma_i^x$ для цепи $\xi_n = \{i_n, \mathbf{v}_n\}, n \ge 1$, можно рассматривать векторы $\mathbf{x}_i,\ \mathbf{\Gamma}_i^X,\ i \ge 1$, образованные величинами $x_{i,\mathbf{v}}$ и $\gamma_{i,\mathbf{v}}^x$ соответственно, упорядоченными в лексикографическом порядке возрастания индексов $\mathbf{v}$. Нетрудно показать, что векторы $\mathbf{\Gamma}_i^X,\ i \ge 0$, имеют вид

$$\mathbf{\Gamma}_i^X = \sum_{j=\max\{0,i-1\}} P_{ij}\mathbf{x}_j - \mathbf{x}_i,\ i \ge 0,$$

где

$$P_{ij} = \begin{cases} V_j,\ i = 0,\ j \ge 0, \\ Y_{j-i+1},\ j \ge i-1,\ i > 0. \end{cases}$$

Назовем векторы $\mathbf{\Gamma}_i^X,\ i \geq 0$, векторами средних одношаговых X-сдвигов цепи $\xi_n,\ n \geq 1$, в состояниях, соответствующих значению i счетной компоненты.

Пусть векторная тестовая функция $\mathbf{x}_i,\ i \geq 0$, заданная на пространстве состояний цепи $\xi_n,\ n \geq 1$, имеет вид

$$\mathbf{x}_i = \begin{cases} \mathbf{0},\ \text{если}\ \ i < J, \\ \\ (i+1)\mathbf{e} + \boldsymbol{\alpha},\ i \geq J, \end{cases} \tag{3.62}$$

где $\boldsymbol{\alpha}$ — некоторый вещественнозначный вектор порядка K, $J = [\alpha_{\max}]$, $\alpha_{\max}$ – максимальный из модулей элементов вектора $\boldsymbol{\alpha}$. При тестовой функции вида (3.62) векторы $\mathbf{\Gamma}_i^X$ средних сдвигов при $i > J$ имеют вид

$$\mathbf{\Gamma}_i^X = \sum_{j=i-1}^{\infty} Y_{j-i+1}[(j+1)\mathbf{e} + \boldsymbol{\alpha}] - [(i+1)\mathbf{e} + \boldsymbol{\alpha}],\ i > J.$$

Эти векторы нетрудно выразить в терминах ПФ $Y(z)$:

$$\mathbf{\Gamma}_i^X = [(Y(z) - zI)]'_{z=1}\mathbf{e} + [Y(1) - I]\boldsymbol{\alpha},\ i > J.$$

Тогда по лемме А10 (см. Приложение А) существует вектор $\boldsymbol{\Delta}$ такой, что

$$\mathbf{\Gamma}_i^X = -\boldsymbol{\Delta},\ i > J, \tag{3.63}$$

где знак каждой из компонент вектора $\boldsymbol{\Delta}$ совпадает со знаком производной $[\det(zI - Y(z))]'_{z=1}$ (или $\boldsymbol{\Delta} = 0$, если эта производная равна 0).

Докажем достаточность выполнения неравенства (3.61) для эргодичности цепи $\xi_n,\ n \geq 1$. Для этого будем использовать теорему Мустафы. Сформулируем аналог этой теоремы для нашей цепи.

Для того чтобы неприводимая непериодическая ЦМ $\xi_n,\ n \geq 1$, была эргодичной, достаточно существования $\varepsilon > 0$, натурального числа J и набора неотрицательных векторов $\mathbf{x}_i,\ i \geq 0$, размерности K таких, что выполняются следующие неравенства:

$$\sum_{j=0}^{\infty} P_{ij}\mathbf{x}_j - \mathbf{x}_i < -\varepsilon\ \mathbf{e},\ i > J, \tag{3.64}$$

$$\sum_{j=0}^{\infty} P_{ij}\mathbf{x}_j < +\infty,\ i = \overline{0, J}. \tag{3.65}$$

По определению, левая часть (3.64) представляет собой (с учетом того, что $P_{ij} = O$, $j < i-1$) вектор $\mathbf{\Gamma}_i^X$. Векторы $\mathbf{\Gamma}_i^X$, $i > J$, при тестовой функции вида (3.62) имеют вид (3.63). Тогда выполнение неравенства (3.61) гарантирует наличие вектора $\boldsymbol{\alpha}$, определяющего тестовую функцию, при котором $\mathbf{\Gamma}_i^X < 0$, $i > J$, то есть выполняются условия (3.64) теоремы Мустафы. Нетрудно показать также, что неравенства $Y^{'}(1) < \infty$ и $V^{'}(1) < \infty$ обеспечивают выполнение условий (3.65) теоремы Мустафы. Из сказанного следует, что выполнение неравенства (3.61) достаточно для эргодичности рассматриваемой цепи.

Доказательство необходимости выполнения неравенства (3.61) для эргодичности рассматриваемой цепи здесь опускаются.

□

Условие эргодичности (3.61) имеет абстрактный математический вид. Кроме того, вычисление левой части (3.61) сопряжено со значительными вычислительными затратами. Поэтому мы приведем другой, более легко проверяемый, вид условия (3.61).

Следствие 3.4. *Неравенство (3.61) эквивалентно следующему неравенству:*

$$\mathbf{x}Y'(1)\mathbf{e} < 1, \tag{3.66}$$

где вектор $\mathbf{x}$ *удовлетворяет системе линейных алгебраических уравнений*

$$\begin{cases} \mathbf{x}Y(1) = \mathbf{x}, \\ \\ \mathbf{x}\mathbf{e} = 1. \end{cases} \tag{3.67}$$

Доказательство. Вычислим левую часть (3.61) следующим образом. Разложим определитель $\det(zI - Y(z))$ по какому-либо столбцу (без ограничения общности будем считать, что это первый столбец) и, дифференцируя соответствующее выражение, получим

$$(\det(zI - Y(z)))'|_{z=1} = \boldsymbol{\nabla}(zI - Y(z))'|_{z=1}\mathbf{e}, \tag{3.68}$$

где вектор $\boldsymbol{\nabla}$ — вектор-строка алгебраических дополнений первого столбца определителя $\det(I - Y(1))$. Как упоминается в доказательстве леммы А10, для стохастической неприводимой матрицы $Y(1)$ вектор $\boldsymbol{\nabla}$ положителен. Кроме того, известно, что вектор $\boldsymbol{\nabla}$ с точностью до постоянного множителя представляет решение $\mathbf{x}$ системы линейных алгебраических уравнений

$$\mathbf{x}Y(1) = \mathbf{x},$$

то есть

$$\mathbf{x} = c\boldsymbol{\nabla}.$$

Тогда единственное решение системы (3.67) имеет вид:

$$\mathbf{x} = (\boldsymbol{\nabla}\mathbf{e})^{-1}\boldsymbol{\nabla},\ \boldsymbol{\nabla} > 0.$$

Из последнего равенства, (3.61) и (3.68) следует (3.67). □

Далее предполагаем, что условие (3.61) выполнено, и изучим стационарное распределение вероятностей ЦМ $\xi_n,\ n \geq 1$:

$$\pi(i, \mathbf{v}) = \lim_{n\to\infty} P\{i_n = i, \mathbf{v}_n = \mathbf{v}\},\ \mathbf{v} \in \mathcal{V}_i,\ i \geq 0.$$

Для упрощения изложения далее будем предполагать, что $\mathcal{V}_i = \mathcal{V},\ i \geq 0$, и множество $\mathcal{V}$ имеет мощность $K, K \geq 1$.

Упорядочим стационарные вероятности в лексикографическом порядке и введем векторы

$$\boldsymbol{\pi}_i = (\pi(i, \mathbf{v}),\ \mathbf{v} \in \mathcal{V}),\ i \geq 0,$$

порядка K.

В литературе существуют три различных подхода к вычислению векторов стационарных вероятностей $\boldsymbol{\pi}_i,\ i \geq 0$. Опишем эти подходы.

3.4.2 Метод производящих функций для нахождения стационарного распределения вероятностей состояний цепи

Векторы стационарных вероятностей $\boldsymbol{\pi}_l,\ l \geq 0$, цепи Маркова $\xi_n,\ n \geq 1$, удовлетворяют системе уравнений равновесия

$$\boldsymbol{\pi}_l = \boldsymbol{\pi}_0 V_l + \sum_{i=1}^{l+1} \boldsymbol{\pi}_i Y_{l-i+1},\ l \geq 0. \tag{3.69}$$

Система (3.69) получена путем использования формулы полной вероятности и матрицы переходных вероятностей цепи вида (3.60).

Стационарное распределение $\boldsymbol{\pi}_l,\ l \geq 0$, полностью определяется матричной ПФ

$$\boldsymbol{\Pi}(z) = \sum_{l=0}^{\infty} \boldsymbol{\pi}_l z^l,$$

которая аналитична в области $|z| < 1$ и непрерывна на ее границе $|z| = 1$. Для этой функции справедливо следующее утверждение.

Теорема 3.10. *Векторная ПФ* $\mathbf{\Pi}(z)$ *стационарного распределения* $\boldsymbol{\pi}_l$, $l \geq 0$, *цепи* ξ_n, $n \geq 1$, *удовлетворяет матричному функциональному уравнению*

$$\mathbf{\Pi}(z)(zI - Y(z)) = \boldsymbol{\pi}_0 H(z),\ |z| \leq 1, \tag{3.70}$$

где

$$H(z) = (zV(z) - Y(z)).$$

При этом $\mathbf{\Pi}(z)$ *является единственным аналитическим в круге* $|z| < 1$ *и непрерывным на границе* $|z| = 1$ *решением уравнения (3.70) таким, что* $\mathbf{\Pi}(1)\mathbf{e} = 1$.

Доказательство. Формулу (3.70) получаем, умножая l-е уравнение системы (3.69) на z^l и суммируя по всем $l \geq 0$.

Единственность решения докажем методом от противного. Пусть $\mathbf{\Pi}(z)$ не единственное решение уравнение (3.70), удовлетворяющее условиям теоремы. Тогда существует другая функция $\mathbf{P}(z)$, аналитическая в области $|z| < 1$ и непрерывная на границе $|z| = 1$, которая удовлетворяет равенству $\mathbf{P}(1)\mathbf{e} = 1$ и является решением уравнения (3.70), то есть

$$\mathbf{P}(z)(zI - Y(z)) - \mathbf{P}(0)H(z) = 0,\ |z| \leq 1. \tag{3.71}$$

Разложив в ряд по степеням z^l, $l \geq 0$, левую часть уравнения (3.71) и приравнивая к нулю коэффициенты при z^l, $l > 0$, получим систему уравнений (3.69) для векторов $\mathbf{p}_l,\ l \geq 0$, определяемых разложением

$$\mathbf{P}(z) = \sum_{l=0}^{\infty} \mathbf{p}_l z^l. \tag{3.72}$$

Так как решение системы уравнений равновесия определяется с точностью до постоянного множителя, то векторы $\mathbf{p}_l$, $l \geq 0$, связаны со стационарным распределением $\boldsymbol{\pi}_l$, $l \geq 0$, следующим образом:

$$\mathbf{p}_l = c\boldsymbol{\pi}_l,\ l \geq 0. \tag{3.73}$$

Учитывая вид (3.73) коэффициентов $\mathbf{p}_l,\ l \geq 0$, можно констатировать, что ряд (3.72) сходится на границе области $|z| < 1$, и, следовательно, для непрерывной на границе функции $\mathbf{P}(z)$ остается справедливым представление (3.73) в точках, принадлежащих границе $|z| = 1$.

В частности,

$$\mathbf{P}(1) = \sum_{i=0}^{\infty} \mathbf{p}_i = c \sum_{i=0}^{\infty} \boldsymbol{\pi}_i,$$

и из того, что $\mathbf{P}(1)\mathbf{e} = 1$, следует, что $c = 1$.

Тогда, вследствие (3.73), коэффициенты $\boldsymbol{\pi}_l$, $\mathbf{p}_l$, $l \geq 0$ разложений функций $\boldsymbol{\Pi}(z)$ и $\mathbf{P}(z)$ в ряды по степеням z совпадают. Но эти ряды однозначно определяют функции $\boldsymbol{\Pi}(z)$ и $\mathbf{P}(z)$ в круге $|z| \leq 1$, из чего следует, что $\mathbf{P}(z) = \boldsymbol{\Pi}(z)$ при $|z| \leq 1$.

□

Из теоремы 3.10 следует, что найдя решение функционального уравнения (3.70), удовлетворяющее условиям теоремы, мы решим проблему нахождения искомого стационарного распределения.

Заметим, что в случае, когда многомерная КТЦМ $\xi_n,\ n \geq 1$, описывает поведение некоторой СМО, уравнение (3.70), по существу, является матричным аналогом формулы Полачека – Хинчина для стационарного распределения длины очереди в системе $M/G/1$. Действительно, для такой системы с интенсивностью входного потока λ и распределением времени обслуживания $B(t)$ мы имеем: $Y(z) = V(z) = \beta(\lambda(1-z))$, где $\beta(s) = \int\limits_0^\infty e^{-st} dB(t)$, $\operatorname{Re} s \geq 0$, и уравнение (3.70) приобретает вид

$$\Pi(z) = \pi_0 \frac{(z-1)\beta(\lambda(1-z))}{z - \beta(\lambda(1-z))}.$$

Но если в случае системы $M/G/1$ вывод такой формулы практически полностью решает задачу нахождения стационарного распределения, поскольку вероятность π_0 легко находится из условия нормировки, то в матричном случае эта задача гораздо сложнее. Здесь также уравнение можно считать решенным, если определен вектор $\boldsymbol{\pi}_0$. Нахождение этого вектора является ключевой проблемой при решении уравнения (3.70). Для простоты изложения сделаем два предположения:

Предположение 3.1. *Пусть матрица $Y(1)$ неразложимая.*

Предположение 3.2. *Среди элементов матрицы $Y(z)$ найдется хотя бы один элемент $Y_{k,k'}(z)$ такой, что коэффициенты $Y_l(k,k')$ его разложения $Y_{k,k'}(z) = \sum\limits_{l=0}^{\infty} Y_l(k,k') z^l$ в ряд по степеням z удовлетворяют условию: существует l, $l \geq 0$, такое, что*

$$Y_l(k,k') Y_{l+1}(k,k') > 0.$$

При выполнении этих предположений и условия эргодичности рассматриваемой ЦМ можно доказать, что уравнение

$$\det(zI - Y(z)) = 0 \tag{3.74}$$

имеет ровно K корней (с учетом их кратности) в круге $|z| \le 1$, одним из которых является простой корень $z = 1$, а все остальные корни лежат внутри единичного круга. Обозначим через $\tilde{K}$ число различных корней z_k, таких, что $|z_k| < 1$, а их кратности — через n_k, $n_k \ge 1$, $k = \overline{1, \tilde{K}}$, $\sum\limits_{k=1}^{\tilde{K}} n_k = K - 1$.

Перепишем уравнение (3.70) в виде

$$\mathbf{\Pi}(z) = \frac{\mathbf{\Pi}(0)H(z)\mathrm{Adj}(zI - Y(z))}{\det(zI - Y(z))}. \tag{3.75}$$

Следствие 3.5. *ПФ* $\mathbf{\Pi}(z)$ *стационарного распределения* $\boldsymbol{\pi}_i$, $i \ge 0$, *ЦМ* ξ_n, $n \ge 1$, *является решением функционального уравнения (3.70) тогда и только тогда, когда система линейных алгебраических уравнений для компонент вектора* $\mathbf{\Pi}(0)$

$$\mathbf{\Pi}(0)\frac{d^n}{dz^n}[H(z)Adj(zI - Y(z))]_{z=z_k} = 0,\ n = \overline{0, n_k - 1},\ k = \overline{1, \tilde{K}}, \tag{3.76}$$

$$\mathbf{\Pi}(0)\frac{d}{dz}[H(z)Adj(zI - Y(z))]_{z=1}\mathbf{e} = \frac{d}{dz}[\det(zI - Y(z))]_{z=1} \tag{3.77}$$

имеет единственное решение.

При этом $\mathbf{\Pi}(z)$ *— единственное аналитическое в круге* $|z| < 1$ *и непрерывное на границе* $|z| = 1$ *решение уравнения (3.75) такое, что* $\mathbf{\Pi}(1)\mathbf{e} = 1$.

Доказательство. Пусть $\mathbf{\Pi}(z)$ является решением (3.70). Как отмечено выше, это — единственное решение в рассматриваемом классе функций. Так как $\mathbf{\Pi}(z)$ — аналитическая в круге $|z| < 1$ и непрерывная на его границе, то числитель (3.75) и его $n_k - 1$ производных обращаются в нуль в каждой из точек z_k, $k = \overline{1, \tilde{K}}$, то есть выполняются соотношения (3.76). Значение $\mathbf{\Pi}(z)$ в точке $z = 1$ можно вычислить, применив правило Лопиталя к правой части (3.75). Умножив полученное при этом соотношение на вектор $\mathbf{e}$, получим уравнение (3.77).

Неоднородная система линейных уравнений (3.76), (3.77) имеет единственное решение. Предположим противное. Тогда либо не существует вектора $\mathbf{\Pi}(0)$, при котором уравнение (3.70) имеет своим решением ПФ $\mathbf{\Pi}(z)$

(нарушаются условия аналитичности решения в области $|z| < 1$, непрерывности на границе $|z| = 1$ и нормировки — все одновременно либо некоторые из них), либо существуют два или более линейно-независимых вектора $\mathbf{\Pi}(0)$ и уравнение (3.70) имеет не единственное решение в рассматриваемом классе функций, что противоречит предположению, сделанному в начале доказательства.

Пусть теперь система уравнений (3.76), (3.77) имеет единственное решение. Тогда из построения этой системы следует, что подставив ее решение $\mathbf{\Pi}(0)$ в (3.75), мы определим функцию $\mathbf{\Pi}(z)$, аналитическую в области $|z| < 1$, непрерывную на границе $|z| = 1$ и удовлетворяющую условию нормировки $\mathbf{\Pi}(1)\mathbf{e} = 1$. Эта функция является также решением уравнения (3.70). Тогда в силу единственности решения уравнения (3.70) в рассматриваемом классе функций, определенная нами функция и является ПФ стационарного распределения цепи $\xi_n,\ n \geq 1$. □

Сформулированные результаты лежат в основе алгоритма нахождения стационарного распределения многомерной КТЦМ $\xi_n,\ n \geq 1$, состоящего из следующих шагов.

Шаг 1. Проверка условия существования стационарного распределения.

Условия $Y^{'}(1) < \infty,\ V^{'}(1) < \infty$ практически всегда выполняются для цепей, описывающих функционирование СМО, поскольку их нарушение означает возможность для счетной компоненты цепи совершить с положительной вероятностью скачок бесконечно большого размера вправо.

Решаем систему линейных алгебраических уравнений (3.67) и проверяем выполнение неравенства (3.66). Если это неравенство не выполняется, то рассматриваемая ЦМ не имеет стационарного распределения. Останавливаем работу алгоритма. Если оно выполняется, то переходим на шаг 2.

Стоит заметить, что довольно часто при исследовании конкретных систем массового обслуживания и наличии достаточного опыта решение системы (3.67) можно угадать (с последующей проверкой путем подстановки угаданного решения в систему). В таких случаях условие эргодичности удается получить в более или менее простом аналитическом виде.

Отметим также, что находя значения параметров системы массового обслуживания, при которых левая часть (3.66) обращается в единицу, мы можем найти некоторые практически важные характеристики функциони-

рования системы, такие, как пропускная способность (то есть предельное значение средней интенсивности потока в систему), минимально допустимая скорость обслуживания требований, при которой в системе не накапливается бесконечная очередь и т. п.

Шаг 2. Нахождение вектора $\mathbf{\Pi}(0)$.

Вектор $\mathbf{\Pi}(0)$ является единственным решением системы линейных алгебраических уравнений (3.76), (3.77). Поэтому первым шагом при вычислении $\mathbf{\Pi}(0)$ является формирование матрицы коэффициентов при неизвестных в системе (3.76), (3.77).

Решаем уравнение (3.74) и обозначаем, как и выше, $z_1, \ldots, z_{\tilde{K}}$ — корни кратностей $n_1, \ldots, n_{\tilde{K}}$ внутри единичного круга, $z_K = 1$. При формировании матрицы системы (3.76), (3.77) полезны следующие сведения об этой системе.

Ранг системы (3.76) равен $K-1$, в то время как количество уравнений равно $K(K-1)$ (каждый корень порождает K уравнений). В случае простых корней $z_1, \ldots, z_{K-1}$ $(\tilde{K} = K-1)$ можно показать, что ранг подсистемы, порожденной каждым из корней, равен 1. В этом случае система (3.76), (3.77) приобретает вид

$$\mathbf{\Pi}(0)[H(z)\mathrm{Adj}(zI - Y(z))]_{z=z_k}\mathbf{e}^{(1)} = 0,\ k = \overline{1, K-1}, \tag{3.78}$$

$$\mathbf{\Pi}(0)\frac{d}{dz}[H(z)\mathrm{Adj}(zI - Y(z))]_{z=1}\mathbf{e} = [\det(zI - Y(z))]'_{z=1}, \tag{3.79}$$

где $\mathbf{e}^{(1)}$ — вектор-столбец, первая компонента которого равна единице, остальные — нулю.

В случае кратных корней аналитически выделить $K-1$ линейно-независимых уравнений из $K(K-1)$ уравнений (3.76) не удается. В этом случае система заведомо избыточная и для нахождения линейно независимых уравнений требуются соответствующие программные средства. Сформировав матрицу системы (3.76), (3.77), решаем систему линейных алгебраических уравнений и находим вектор $\mathbf{\Pi}(0)$.

Шаг 3. Вычисление стационарных вероятностей.

Вычислив вектор $\mathbf{\Pi}(0) = \boldsymbol{\pi}_0$, можно считать проблему нахождения стационарного распределения вероятностей цепи $\xi_n,\ n \geq 1$, решенной. Используя уравнение равновесия (3.69), значения любого наперед заданного

числа $\tilde{L}$ векторов $\boldsymbol{\pi}_l,\ l = \overline{0, \tilde{L}}$, можно вычислить по рекуррентным формулам

$$\boldsymbol{\pi}_{l+1} = [\boldsymbol{\pi}_l - \boldsymbol{\pi}_0 V_l - \sum_{i=1}^{l} \boldsymbol{\pi}_i Y_{l-i+1}] Y_0^{-1},\ l \geq 0. \tag{3.80}$$

Препятствием для этого может быть вырожденность матрицы Y_0. Заметим, однако, что предположение о невырожденности матрицы Y_0 является достаточно неограничительным, когда речь идет о цепи, описывающей функционирование однолинейных систем массового обслуживания с групповым марковским входным потоком. Другим препятствием может явиться численная неустойчивость рекурсии (3.80) из-за потери точности (исчезновения порядка), вызванная многократным использованием операции вычитания. Если это происходит или если матрица Y_0 — вырожденная, то для нахождения векторов $\boldsymbol{\pi}_l,\ l \geq 1$, следует использовать другую рекурсию, выведенную в рамках матрично-аналитического подхода М. Ньютса, описанного ниже, в подразделе 4.2.

3.4.3 Вычисление факториальных моментов

Введем обозначения для факториальных моментов:

$$\mathbf{\Pi}^{(m)} = \frac{d^m}{dz^m} \mathbf{\Pi}(z)|_{z=1},\ m \geq 1.$$

Пусть также $\mathbf{\Pi}^{(0)} = \mathbf{\Pi}(1)$.

В случае, когда распределение $\boldsymbol{\pi}_l,\ l \geq 0$ имеет не очень тяжелый "хвост" (то есть вероятностная масса распределения не сосредоточена в состояниях с большим значением первой компоненты ЦМ), при не очень больших значениях $\tilde{L}$ величина $\sum_{l=0}^{\tilde{L}} \boldsymbol{\pi}_l \mathbf{e}$ близка к единице, и вычислить факториальные моменты распределения можно непосредственно по формуле

$$\mathbf{\Pi}^{(m)} = \sum_{i=m}^{\infty} \frac{i!}{(i-m)!} \boldsymbol{\pi}_i,\ m \geq 1.$$

Если же "хвост" распределения тяжелый, то сумма $\sum_{l=0}^{\tilde{L}} \boldsymbol{\pi}_l \mathbf{e}$ медленно сходится к единице с ростом величины $\tilde{L}$ и еще медленнее сходятся суммы вида $\sum_{i=m}^{\tilde{L}} \frac{i!}{(i-m)!} \boldsymbol{\pi}_i$. Тогда для расчета факториальных моментов будем

использовать следующее утверждение, доказательство которого основано на использовании функционального уравнения (3.71).

Следствие 3.6. *Пусть* $Y^{(m)}(1) < \infty$, $V^{(m)}(1) < \infty$, $m = \overline{1, M+1}$, *где* M *– некоторое заданное целое число. Тогда векторы* $\mathbf{\Pi}^{(m)}$, $m = \overline{0, M}$, *находятся рекуррентно из формулы*

$$\mathbf{\Pi}^{(m)} = \left[\left(\mathbf{\Pi}(0)H^{(m)}(1) - \sum_{l=0}^{m-1}\left(\begin{array}{c} m \\ l \end{array}\right)\mathbf{\Pi}^{(l)}(zI - Y(z))^{(m-l)}|_{z=1}\right)\tilde{I} + \right.$$

$$\left. + \frac{1}{m+1}\left(\mathbf{\Pi}(0)H^{(m+1)}(1) - \sum_{l=0}^{m-1}\left(\begin{array}{c} m+1 \\ l \end{array}\right)\mathbf{\Pi}^{(l)}(zI - Y(z))^{(m+1-l)}|_{z=1}\right)\mathbf{e}\hat{\mathbf{e}}\right]\tilde{\mathbf{A}}^{-\mathbf{1}}, \tag{3.81}$$

где

$$\tilde{\mathbf{A}} = (I - Y(1))\tilde{I} + (I - Y'(1))\mathbf{e}\hat{\mathbf{e}},$$

$$\hat{\mathbf{e}} = (1, 0, \ldots, 0),$$

$\tilde{I} = \mathrm{diag}\{0, 1, \ldots, 1\}$.

Доказательство. Для краткости выкладок обозначим $\mathbf{A}(z) = zI - Y(z)$, $\mathbf{b}(z) = \boldsymbol{\pi}_0(zV(z) - Y(z)) = \boldsymbol{\pi}_0 H(z)$. Последовательно дифференцируя уравнение (3.70), получаем соотношения

$$\mathbf{\Pi}^{(m)}(z)\mathbf{A}(z) = \mathbf{b}^{(m)}(z) - \sum_{l=0}^{m-1}\left(\begin{array}{c} m \\ l \end{array}\right)\mathbf{\Pi}^{(l)}(z)\mathbf{A}^{(m-l)}(z),\ m \geq 0. \tag{3.82}$$

Матрица $\mathbf{Y}(1)$ является стохастической, поэтому матрица $\mathbf{A}(1)$ является вырожденной и не представляется возможным вывести рекурсию для вычисления векторов факториальных моментов $\mathbf{\Pi}^{(m)}$, $m \geq 0$, непосредственно из (3.82). Поэтому модифицируем уравнения (3.82) следующим образом. Полагая в (3.82) $z = 1$, подставляя вместо m число $m+1$ и умножая обе части полученого уравнения на $\mathbf{e}$, с учетом очевидного равенства $\mathbf{A}(1)\mathbf{e} = \mathbf{0}^T$ получаем уравнение

$$(m+1)\mathbf{\Pi}^{(m)}\mathbf{A}^{(1)}(1)\mathbf{e} = \mathbf{b}^{(m+1)}(1)\mathbf{e} - \sum_{l=0}^{m-1}\left(\begin{array}{c} m+1 \\ l \end{array}\right)\mathbf{\Pi}^{(l)}\mathbf{A}^{(m+1-l)}(1)\mathbf{e}. \tag{3.83}$$

Можно показать, что правая часть (3.83) не равна нулю. Заменяя одно из уравнений системы (3.82) при $z = 1$ (без ограничения общности будем заменять первое уравнение) уравнением (3.83), получаем следующую систему уравнений относительно компонент вектора $\mathbf{\Pi}^{(m)}$:

$$\mathbf{\Pi}^{(m)}\tilde{\mathbf{A}} = \left[\left(\mathbf{b}^{(m)}(1) - \sum_{l=0}^{m-1}\begin{pmatrix} m \\ l \end{pmatrix}\mathbf{\Pi}^{(l)}\mathbf{A}^{(m-l)}(1)\right)\tilde{I}+\right.$$

$$\left.+\frac{1}{m+1}\left(\boldsymbol{b}^{(m+1)}(1) - \sum_{l=0}^{m-1}\begin{pmatrix} m+1 \\ l \end{pmatrix}\mathbf{\Pi}^{(l)}\mathbf{A}^{(m+1-l)}(1)\right)\mathbf{e}\hat{\mathbf{e}}\right]. \qquad (3.84)$$

Уравнение (3.81) непосредственно получается из (3.84), если мы покажем, что матрица $\tilde{\mathbf{A}}$ невырождена. Для этого вычислим определитель этой матрицы. Можно убедиться, что $\det\tilde{\mathbf{A}} = c[\det\mathbf{A}(z)]'_{z=1} = c[\det(zI - Y(z)]'_{z=1}$, где $c \neq 0$. В силу неравенства (3.61) заключаем, что искомый определитель отличен от нуля.

□

Описанные выше алгоритмы решают проблему нахождения вектора $\mathbf{\Pi}(0) = \boldsymbol{\pi}_0$, наперед заданного числа векторов $\boldsymbol{\pi}_l$, $l = \overline{1, \tilde{\tilde{L}}}$, и наперед заданного числа факториальных моментов $\mathbf{\Pi}^{(k)}$, $k = \overline{0, M}$, по заданному виду матричных ПФ $Y(z)$, $V(z)$, описывающих переходы многомерной КТЦМ.

Отметим, что уравнение (3.81) можно использовать для получения неоднородного уравнения в системе линейных алгебраических уравнений (3.78), (3.79) для компонент вектора $\mathbf{\Pi}(0)$, имеющего более конструктивный вид, чем уравнение (3.79).

Из уравнения (3.81) при $m = 0$ мы получаем

$$\mathbf{\Pi}(1) = \mathbf{\Pi}(0)S, \qquad (3.85)$$

где

$$S = (H(1)\tilde{I} + H^{(1)}(1)\mathbf{e}\hat{\mathbf{e}})\tilde{\mathbf{A}}^{-1}. \qquad (3.86)$$

Поэтому из условия нормировки следует, что уравнение (3.79) может быть заменено уравнением

$$\mathbf{\Pi}(0)S\mathbf{e} = 1.$$

Еще одна возможность получения соотношения между векторами $\mathbf{\Pi}(1)$ и $\mathbf{\Pi}(0)$ и последующего получения из условия нормировки уравнения для

вектора $\mathbf{\Pi}(0)$, альтернативного неоднородному уравнению (3.79), следующая.

Как следует из Леммы А9, если $\boldsymbol{\pi}$ есть стохастический левый собственный вектор неприводимой матрицы $Y(1)$, то есть

$$\boldsymbol{\pi} Y(1) = \boldsymbol{\pi},\ \boldsymbol{\pi}\mathbf{e} = 1, \tag{3.87}$$

то матрица $I - Y(1) + \mathbf{e}\boldsymbol{\pi}$ является невырожденной. Учитывая, что вектор $\mathbf{\Pi}(1)$ является стохастическим и $\mathbf{\Pi}(1)\mathbf{e}\boldsymbol{\pi} = \boldsymbol{\pi}$, суммируя вектор $\mathbf{\Pi}(1)\mathbf{e}\boldsymbol{\pi}$ к обеим частям уравнения (3.70) при $z = 1$ и умножая на матрицу $(I - Y(1) + \mathbf{e}\boldsymbol{\pi})^{-1}$, получаем:

$$\mathbf{\Pi}(1) = \left(\mathbf{\Pi}(1)\mathbf{e}\boldsymbol{\pi} + \mathbf{\Pi}(0)H(1)\right)(I - Y(1) + \mathbf{e}\boldsymbol{\pi})^{-1},$$

откуда с учетом (3.87) получаем:

$$\mathbf{\Pi}(1) = \boldsymbol{\pi} + \mathbf{\Pi}(0)H(1)(I - Y(1) + \mathbf{e}\boldsymbol{\pi})^{-1}. \tag{3.88}$$

Уравнение (3.88) задает связь между векторами $\mathbf{\Pi}(1)$ и $\mathbf{\Pi}(0)$ (с участием известного вектора $\boldsymbol{\pi}$), альтернативную выведенному нами соотношению (3.85).

3.4.4 Матрично-аналитический метод для нахождения стационарного распределения вероятностей состояний цепи (метод М. Ньютса)

Существенным этапом метода векторных ПФ и решения функционального уравнения (3.70) с использованием соображений аналитичности векторной ПФ является процедура нахождения корней определителя матрицы $Y(z) - zI$ в единичном круге комплексной плоскости. Теоретически эта процедура довольно проста, и существует множество методов ее решения. Однако при реализации этой процедуры на практике могут возникать проблемы, связанные с плохой отделимостью корней и с недостаточностью разрядной сетки персональных компьютеров.

Побудительным мотивом разработки алгоритма М. Ньютса [**?**] было желание найти неизвестный вектор $\mathbf{\Pi}(0)$, избежав поиска корней функции комплексной переменной. Отметим, что уравнения равновесия (3.69) содержат всю доступную информацию о поведении ЦМ типа $M/G/1$, и тем не менее соотношение (3.70) для ПФ $\mathbf{\Pi}(z)$, полученное на основе уравнений

равновесия, определяет эту функцию лишь с точностью до неизвестного постоянного вектора $\mathbf{\Pi}(0)$. Основываясь на уравнениях равновесия, не удается получить какой-либо дополнительной информации об этом векторе. Выходом из положения является использование дополнительных соображений о векторной ПФ $\mathbf{\Pi}(z)$ или векторе $\mathbf{\Pi}(0)$. В предыдущем параграфе мы использовали дополнительную информацию о том, что ПФ вероятностей является аналитической в единичном круге комплексной плоскости. М. Ньютс использовал другую известную из теории ЦМ информацию: стационарная вероятность состояния цепи обратно пропорциональна математическому ожиданию времени до первого возвращения в это состояние. Для однородной цепи с дискретным временем это время измеряется числом переходов до возвращения в состояние. Детальное изложение подхода М. Ньютса требует знания читателем теории марковских процессов восстановления. Поэтому мы ограничимся лишь его схематичным изложением.

Нас интересует значение вектора $\mathbf{\Pi}(0)$, ν-я компонента $\pi(0,\nu)$ которого обратно пропорциональна среднему числу переходов, которое совершит цепь, выйдя из состояния $(0,\nu)$, до первого возвращения в него. Поэтому мы должны определить распределение числа переходов цепи между двумя соседними попаданиями в состояние с нулевым значением счетной компоненты. При этом, однако, мы должны "отслеживать" и переходы конечной компоненты за это время.

Краеугольным понятием в подходе М. Ньютса является понятие фундаментального периода. Фундаментальный период — это интервал времени с момента, когда значение счетной компоненты равно i, до первого момента, когда значение этой компоненты станет равным $i-1$, $i \geq 1$. Из определения ЦМ типа $M/G/1$ (квазитеплицевой цепи) следует, что длина фундаментального периода не зависит от значения i.

Обозначим $G(z)$ матричную ПФ, (ν,ν')-й элемент которой есть ПФ числа переходов, осуществленных ЦМ за фундаментальный период, который закончился, когда значение конечной компоненты есть ν' при условии, что в момент начала фундаментального периода значение конечной компоненты равнялось ν; $\nu,\nu' = \overline{0,W}$. Используя факты из теории марковских процессов восстановления, можно показать, что матрица $G(z)$ удовлетворяет уравнению

$$G(z) = z\sum_{j=0}^{\infty} Y_j G^j(z). \tag{3.89}$$

Это уравнение легко вывести и используя известную вероятностную интер-

претацию ПФ в терминах раскрашенных запросов, см., например, [76], [77].

Структура времени между соседними попаданиями счетной компоненты в нулевое состояние ясна: на первом шаге осуществляется переход счетной компоненты из нулевого состояния в некоторое состояние j, $j \geq 0$, (вероятности переходов конечной компоненты за это время задаются матрицей V_j), а затем следует j фундаментальных периодов (если $j \geq 1$), за которые значение счетной компоненты, последовательно уменьшаясь на единицу, станет равным нулю.

Обозначим через $K(z)$ матричную ПФ, (ν, ν')-й элемент которой есть ПФ числа переходов цепи за время до возвращения счетной компоненты в нулевое состояние с состоянием конечной компоненты ν' в момент возвращения при условии, что в момент выхода из нулевого состояния эта компонента находилась в состоянии ν, $\nu, \nu' = \overline{0, W}$.

Анализируя описанную выше структуру времени между моментами попадания в нулевое состояние счетной компоненты, легко убедиться, что

$$K(z) = z \sum_{j=0}^{\infty} V_j G^j(z).$$

Обозначим $\boldsymbol{\kappa}$ и $\boldsymbol{\kappa}^*$ вектор-строку и вектор-столбец, задаваемые соотношениями

$$\boldsymbol{\kappa} K(1) = \boldsymbol{\kappa}, \ \boldsymbol{\kappa}\mathbf{e} = 1,$$

$$\boldsymbol{\kappa}^* = K'(1)\mathbf{e}.$$

Компонента κ_ν вектора $\boldsymbol{\kappa}$ есть стационарная вероятность нахождения процесса ν_t в состоянии ν в моменты выхода счетной компоненты из нулевого состояния, $\nu = \overline{0, W}$. Компонента κ^*_ν вектора $\boldsymbol{\kappa}^*$ есть среднее число переходов квазитеплицевой цепи за время до возвращения счетной компоненты в нулевое состояние при условии, что в момент выхода из нулевого состояния конечная компонента находилась в состоянии ν, $\nu = \overline{0, W}$.

Используя вышеупомянутую интерпретацию стационарной вероятности в терминах среднего числа переходов до возвращения и сведения из теории марковских процессов восстановления, получаем итоговую формулу, задающую значение вектора $\boldsymbol{\Pi}(0)$:

$$\boldsymbol{\Pi}(0) = \frac{\boldsymbol{\kappa}}{\boldsymbol{\kappa}\boldsymbol{\kappa}^*}. \tag{3.90}$$

Кроме возможности получения значения вектора $\boldsymbol{\Pi}(0)$ без использования корней уравнения (3.74), матрично-аналитический подход позволяет полу-

чить рекуррентную процедуру вычисления векторов стационарных вероятностей $\boldsymbol{\pi}_l,\ l \geq 1$, альтернативную рекурсии (3.80).

Выше отмечалось, что рекурсия (3.80) неприменима, когда матрица Y_0 является вырожденной. Кроме этого, рекурсия (3.80) имеет следующий существенный недостаток. Если распределение имеет тяжелый "хвост", то чтобы вычислить достаточное число векторов $\boldsymbol{\pi}_l,\ l \geq 0$, приходится выполнить много шагов в рекуррентной процедуре (3.80), которая содержит операции вычитания. Как показывает практика, рекуррентная процедура, содержащая операцию вычитания, обладает способностью накапливать ошибки вычислений, что при современных возможностях компьютерной техники может приводить к недопустимым вычислительным погрешностям. В такой ситуации разработка альтернативного алгоритма, лишенного указанных недостатков, является актуальной.

Для разработки альтернативного алгоритма будем использовать понятие сенсорной ЦМ, см., например, [80]. Для простоты его понимания опишем сначала понятие сенсорной ЦМ для случая одномерной цепи.

Пусть имеется ЦМ $i_n,\ i_n \geq 0\ n \geq 1$, с одношаговыми переходными вероятностями $P_{i,l}$. Зафиксируем некоторое целое число $N,\ N \geq 0$. Сенсорной цепью $i_n^{(N)},\ n \geq 1$, по отношению к цепи $i_n,\ n \geq 1$, с уровнем сенсорирования N назовем случайный процесс, траектории которого получаются из траекторий ЦМ $i_n,\ n \geq 1$, путем удаления участков траекторий, на которых $i_n > N$, и склеивания оставшихся частей.

Нетрудно видеть, что полученный в результате такой операции случайный процесс также будет ЦМ. Эта ЦМ $i_n^{(N)}, n \geq 1$, имеет пространство состояний $\{0, 1, 2, ..., N\}$ и вероятности переходов $P_{i,l}^{(N)}$, которые вычисляются как $P_{i,l}^{(N)} = P_{i,l}, l < N$, и $P_{i,N}^{(N)} = \sum\limits_{l=N}^{\infty} P_{i,l}$.

Известно, что стационарные вероятности исходной цепи и стационарной вероятности сенсорной цепи совпадают на множестве состояний сенсорной цепи с точностью до мультипликативной константы

Рассмотрим теперь многомерную ЦМ $\xi_n = \{i_n, \mathbf{v}_n\},\ n \geq 1$, с пространством состояний $\mathcal{S} = \{(i, \mathbf{v}), \mathbf{v} \in \mathcal{V}\}$. Построим сенсорную по отношению к ней многомерную ЦМ $\xi_n^{(N)} = \{i_n^{(N)}, \mathbf{v}_n^{(N)}\}, n \geq 1$, с уровнем сенсорирования $N,\ N \geq 0$, путем удаления участков траектории, на которых $i_n > N$.

Обозначим матрицы переходных вероятностей из состояний $(i, \mathbf{v})$ в состояния $(l, \mathbf{v}')$ через $P_{i,l}^{(N)},\ i, l = \overline{0, N}$. Легко видеть, что для $l < N$ $P_{i,l}^{(N)} = P_{i,l}$. Матрицы $P_{i,N}^{(N)}$ в этом случае рассчитываются существенно

сложнее, чем в случае одномерной ЦМ, поскольку требуется отслеживать переходы конечной компоненты $\mathbf{v}_n^{(N)},\ n \geq 1$, ЦМ $\xi_n^{(N)} = \{i_n^{(N)}, \mathbf{v}_n\}, n \geq 1$, на "вырезанных" участках траектории.

В случае многомерной ЦМ с блочной матрицей вероятностей переходов произвольной структуры задача расчета вероятностей переходов конечной компоненты $\mathbf{v}_n,\ n \geq 1$, на "вырезанных" участках траектории очень сложна. Поэтому далее мы будем использовать специфику блочной матрицы (3.60) вероятностей переходов рассматриваемых в данной главе ЦМ типа $M/G/1$. Эта специфика состоит, во-первых, в том, что невозможно уменьшение счетной компоненты i_n ЦМ $\xi_n,\ n \geq 1$, за один шаг более чем на единицу: $P_{i,l} = O,\ l < i-1$. Во-вторых, матрица вида (3.60) является блочной квазитеплицевой, то есть $P_{i,i+l-1} = Y_i,\ l \geq 0$. Иначе говоря, вероятность перехода из состояния со значением i счетной компоненты в состояние со значением k этой компоненты зависит только от разности $i-k$, но не зависит от i и k по отдельности. Учет этой специфики существенно упрощает задачу вычисления матриц $P_{i,N}^{(N)}$.

Проанализируем поведение ЦМ $\xi_n = \{i_n, \mathbf{v}_n\},\ n \geq 1$, на "вырезанном" участке траектории между моментом, когда значение первой компоненты было $i,\ i \leq N$, на следующем шаге стало больше N, и моментом, когда оно впервые станет равным N. Либо между моментом, когда значение первой компоненты было $i,\ i < N$, а на следующем шаге стало равно N.

Это поведение таково. Либо на первом шаге ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, перейдет в состояние со значением первой компоненты равным N. Либо на первом шаге она перейдет в состояние со значением первой компоненты $l, l > N$. После этого цепь может осуществить один или конечное множество переходов до тех пор, пока значение первой компоненты не станет равным $l-1$. Конечность числа таких переходов обеспечена сделанным нами предположением, что исходная ЦМ является эргодической. Затем она может осуществить еще один или несколько переходов до тех пор, пока значение первой компоненты не станет равным $l-2$, и так далее до тех пор, пока значение этой компоненты не станет равным N.

Обозначим через $G^{(i)}$ матрицу вероятностей переходов компоненты $\mathbf{v}_n$ за время первого достижения первой компонентой i_n значения (уровня) i, стартуя со значения $i+1$.

Используя формулу полной вероятности, нетрудно убедиться, что мат-

рицы $G^{(i)},\ i \geq 0$, удовлетворяют уравнениям

$$G^{(i)} = P_{i+1,i} + \sum_{l=i+1}^{\infty} P_{i+1,l} G^{(l-1)} G^{(l-2)} \dots G^{(i)},\ i \geq 0. \tag{3.91}$$

Учитывая квазитеплицевость рассматриваемой ЦМ, можно заключить, что матрицы $G^{(i)},\ i \geq 0$, не зависят от i. Пусть все они равны некоторой матрице G. Тогда из (3.91) можно заключить, что матрица G, удовлетворяет уравнению

$$G = \sum_{j=0}^{\infty} Y_j G^j. \tag{3.92}$$

Отметим, что уравнение (3.92) автоматически следует из уравнения (3.89), поскольку из определения матриц G и $G(z)$ следует, что $G(1) = G$.

Нелинейное матричное уравнение (3.92) для матрицы G можно решить с помощью метода последовательных приближений аналогично тому, как это делалось при решении нелинейного матричного уравнения (3.4) для матрицы $\mathcal{R}$ в разделе 4.3. Взяв в качестве начального приближения матрицу $G_0 = O$, или $G = I$, или G равно произвольной стохастической матрице, последующие приближения можно вычислить по формуле

$$G_{k+1} = \sum_{j=0}^{\infty} Y_j G_k^j,\ k \geq 0,$$

или

$$G_{k+1} = \sum_{j=0,\, j\neq 1}^{\infty} (I - Y_1)^{-1} Y_j G_k^j,\ k \geq 0.$$

Последовательность матриц $G_k,\ k \geq 0$ сходится к решению уравнения (3.92).

Вычислив матрицы $G^{(i)} = G,\ i \geq 0$, и учитывая описанное выше поведение ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}$, $n \geq 1$, между моментом, когда состояние первой компоненты сенсорной цепи $\xi_n^{(N)} = \{i_n^{(N)}, \mathbf{v}_n^{(N)}\}$, $n \geq 1$, было $i,\ i \leq N$, и моментом, когда оно впервые станет равным N, легко убедиться в правильности соотношения

$$P_{i,N}^{(N)} = P_{i,N} + \sum_{l=N+1}^{\infty} P_{i,l} G^{(l-1)} G^{(l-2)} \dots G^{(N)} = P_{i,N} + \sum_{l=N+1}^{\infty} P_{i,l} G^{l-N}.$$

Обозначим через $\boldsymbol{\pi}_i,\ i \geq 0$, векторы стационарных вероятностей исходной ЦМ $\xi_n,\ n \geq 1$, и через $\boldsymbol{\pi}_i^{(N)},\ i = \overline{0, N}$, векторы стационарных вероятностей сенсорной ЦМ $\xi_n^{(N)}$, $n \geq 1$.

Векторы $\boldsymbol{\pi}_i^{(N)},\ i=\overline{0,N}$, удовлетворяют уравнениям Чепмена – Колмогорова (уравнениям равновесия)

$$\boldsymbol{\pi}_i^{(N)}=\sum_{l=0}^{i+1}\boldsymbol{\pi}_l^{(N)}P_{l,i},\ i=\overline{0,N-1},$$

$$\boldsymbol{\pi}_N^{(N)}=\sum_{l=0}^{N}\boldsymbol{\pi}_l^{(N)}P_{l,N}^{(N)}. \tag{3.93}$$

Выше отмечалось, что известным является факт, что стационарные вероятности исходной цепи и стационарной вероятности сенсорной цепи совпадают на множестве состояний сенсорной цепи с точностью до мультипликативной константы:

$$\boldsymbol{\pi}_i^{(N)}=c\boldsymbol{\pi}_i,\ i=\overline{0,N}.$$

Поэтому из (3.93) следует, что векторы $\boldsymbol{\pi}_i,\ i\geq 0$, стационарных вероятностей исходной цепи удовлетворяют уравнению

$$\boldsymbol{\pi}_N=\sum_{l=0}^{N}\boldsymbol{\pi}_l P_{l,N}^{(N)}. \tag{3.94}$$

Поскольку уровень сенсорирования N выбирался произвольным, $N\geq 0$, из (3.94) следует, что векторы $\boldsymbol{\pi}_i,\ i\geq 0$, стационарных вероятностей исходной цепи удовлетворяют уравнениям

$$\boldsymbol{\pi}_j=\sum_{i=0}^{j}\boldsymbol{\pi}_i\bar{P}_{i,j},\ i\geq 0, \tag{3.95}$$

где $\bar{P}_{i,j}\stackrel{def}{=}P_{i,j}^{(j)}$ и матрицы $\bar{P}_{i,j}$ задаются формулами

$$\bar{P}_{i,j}=P_{i,j}+\sum_{l=j+1}^{\infty}P_{i,l}G^{l-j},\ j\geq i. \tag{3.96}$$

Из (3.95) следует, что векторы $\boldsymbol{\pi}_i,\ i\geq 0$, стационарных вероятностей можно представить в виде

$$\boldsymbol{\pi}_i=\boldsymbol{\pi}_0\Phi_i,\ i\geq 0, \tag{3.97}$$

где матрицы $\Phi_i,\ i \geq 0$, удовлетворяют рекуррентным соотношениям

$$\Phi_0 = I,\ \Phi_l = \left(\bar{P}_{0,l} + \sum_{i=1}^{l-1} \Phi_i \bar{P}_{i,l}\right)(I - \bar{P}_{l,l})^{-1},\ l \geq 1. \tag{3.99}$$

Невырожденность матрицы $I - \bar{P}_{l,l}$ и неотрицательность матрицы $(I - \bar{P}_{l,l})^{-1},\ l \geq 1$, следуют из легко проверяемой строгой субстохастичности матрицы $\bar{P}_{l,l}$, теоремы О. Таусски и Леммы А6. Неотрицательность матриц $(I - \bar{P}_{l,l})^{-1},\ l \geq 1$, влечет то, что в рекурсии (3.99) не используется операция вычитания, что обеспечивает вычислительную устойчивость этой рекурсии.

Соотношения (3.97) определяют векторы $\boldsymbol{\pi}_i,\ i \geq 0$, стационарных вероятностей с точностью до неизвестного пока вектора $\boldsymbol{\pi}_0$. Этот вектор можно вычислить через аппарат векторных ПФ или по формуле (3.90). Однако приведенные выше результаты позволяют предложить еще одну процедуру для подсчета вектора $\boldsymbol{\pi}_0$.

Из уравнений (3.95) при $j = 0$ получаем:

$$\boldsymbol{\pi}_0(I - \bar{P}_{0,0}) = \mathbf{0}. \tag{3.100}$$

Домножением соотношения (3.96) справа на вектор $\mathbf{e}$ несложно убедиться, что матрица $\bar{P}_{0,0}$ — стохастическая. По построению эта матрица является неприводимой. Поэтому ранг системы (3.100)) на единицу меньше размерности вектора $\boldsymbol{\pi}_0$. Значит, система (3.100)) определяет вектор $\boldsymbol{\pi}_0$ с точностью до некоторой константы. Следовательно, если нам удастся получить еще одно, неоднородное, уравнение для компонент вектора $\boldsymbol{\pi}_0$, то полученнная система будет иметь единственное решение. Такое уравнение легко получается из (3.97) и условия нормировки и имеет вид:

$$\boldsymbol{\pi}_0 \sum_{i=0}^{\infty} \Phi_i \mathbf{e} = 1. \tag{3.101}$$

Отметим, что при выполнении условия существования стационарного распределения рассматриваемой ЦМ выполняются следующие предельные соотношения для норм матриц $\Phi_l,\ l \geq 1 : ||\Phi_l|| \to 0$ при $l \to \infty$. Поэтому вычисление бесконечной суммы в левой части уравнения (3.101) можно прекратить, когда $||\Phi_l|| < \varepsilon$, где ε — желаемая точность вычисления.

Подводя итог, сформулируем следующий алгоритм нахождения векторов $\boldsymbol{\pi}_i,\ i \geq 0$, стационарных вероятностей.

Шаг 1. Находим матрицу G как решение нелинейного матричного уравнения (3.92).

Шаг 2. Вычисляем матрицы $\bar{P}_{i,l}$ по формулам (3.96).

Шаг 3. Вычисляем матрицы Φ_l по рекуррентным формулам (3.99).

Шаг 4. Находим вектор $\boldsymbol{\pi}_0$ как единственное решение системы (3.100), (3.101).

Шаг 5. Находим желаемое число векторов $\boldsymbol{\pi}_i$ по формулам (3.97).

Отличием этого алгоритма нахождения векторов $\boldsymbol{\pi}_i,\ i \geq 0$, стационарных вероятностей от алгоритма, основанного на использовании рекурсии (3.80), кроме наличия возможности вычисления вектора $\boldsymbol{\pi}_0$, является его вычислительная устойчивость. Заметим, что В. Рамасвами (V. Ramaswami) предложил такой алгоритм для нахождения векторов $\boldsymbol{\pi}_i,\ i \geq 0$, стационарных вероятностей в предположении, что вектор $\boldsymbol{\pi}_0$ известен и определяется формулой (3.90).

Сопоставляя подход к нахождению стационарного распределения вероятностей состояний ЦМ, основанный на использовании векторных ПФ, с подходом М. Ньютса, можно констатировать следующее. Оба подхода имеют свои сильные и слабые стороны.

Слабыми сторонами аналитического подхода являются следующие:

– необходимость поиска заданного числа корней функции комплексной переменной в единичном круге комплексной плоскости. Теоретически проблема поиска не является сложной. На практике она достаточно сложна, особенно в случае, когда корни плохо отделимы или корни кратные;

– необходимость выполнения дифференцирования (аналитически или с помощью компьютера) матричных функций в (3.76) в случае наличия кратных корней уравнения (3.74). Заметим, что наличие комплексных корней не усложняет задачу и не требует привлечения комплексной арифметики при расчетах на компьютере, поскольку каждый такой корень дает два уравнения с действительными коэффициентами, получающиеся отдельным приравниванием нулю действительной и мнимой части уравнений;

– необходимость поиска заданного числа линейно независимых уравнений в системе (3.76), (3.77);

– использование численно неустойчивой рекурсии (3.80) для вычисле-

ния векторов $\boldsymbol{\pi}_i,\ i \geq 1$.

Слабыми сторонами подхода М. Ньютса являются следующие:

– необходимость решения нелинейного матричного уравнения (3.92);

– необходимость вычисления бесконечных сумм (см. формулы (3.96), (3.101);

– возникновение проблем с вычислением факториальных моментов распределения при "тяжелых хвостах" распределения.

Сильной стороной аналитического подхода является наличие простых рекуррентных формул для подсчета факториальных моментов распределения.

Сильными сторонами подхода М. Ньютса является следующие:

– наличие численно устойчивой рекурсии для вычисления векторов $\boldsymbol{\pi}_i,\ i \geq 1$, стационарных вероятностей;

– возможность его эффективного переноса на случай пространственно неоднородных ЦМ (см. раздел 3.6 ниже).

Заметим, что при решении задачи нахождения факториальных моментов иногда рационально использовать сочетание двух подходов: аналитического и подхода М. Ньютса. При помощи подхода М. Ньютса находим вектор $\boldsymbol{\pi}_0$, а факториальные моменты вычисляем при помощи формулы (3.81). Таким образом удается преодолеть трудности в вычислении факториальных моментов по вероятностям $\boldsymbol{\pi}_i,\ i \geq 1$, в случае "тяжелого" хвоста.

3.5 ЦЕПИ МАРКОВА ТИПА $M/G/1$ С КОНЕЧНЫМ ПРОСТРАНСТВОМ СОСТОЯНИЙ

Если первая компонента i_n цепи Маркова типа $M/G/1$ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, принимает только конечное множество значений, например $i_n = \overline{0, N}$, то ее стационарное распределение может быть найдено с помощью алгоритма, являющегося модификацией алгоритма, полученного в разделе 3.5.4 на основе построения семейства сенсорных цепей Маркова.

Этот алгоритм для нахождения векторов $\boldsymbol{\pi}_i,\ i = \overline{0, N}$, стационарных вероятностей состоит из следующих шагов.

Шаг 1. Находим матрицы $G_i,\ i = \overline{0, N-1}$, как решение следующей матричной обратной рекурсии:

$$G_{N-1} = (I - P_{N,N})^{-1} P_{N,N-1},$$

$$G_i = (I - \sum_{l=i+1}^{N} P_{i+1,l} G_{l-1} G_{l-2} \cdot \ldots \cdot G_{i+1})^{-1} P_{i+1,i},\ i = \overline{0, N-2}.$$

Шаг 2. Вычисляем матрицы $\bar{P}_{i,l}$ по формулам

$$\bar{P}_{i,l} = P_{i,l} + \bar{P}_{i,l+1} G_l,\ i = \overline{0, N},\ l = \overline{i, N},\ \bar{P}_{i,N+1} = O.$$

Шаг 3. Вычисляем матрицы Φ_l по формулам

$$\Phi_0 = I,\ \Phi_l = \sum_{i=0}^{l-1} \Phi_i \bar{P}_{i,l} (I - \bar{P}_{l,l})^{-1},\ l = \overline{1, N}.$$

Шаг 4. Находим вектор $\boldsymbol{\pi}_0$ как единственное решение системы уравнений

$$\boldsymbol{\pi}_0 (I - \bar{P}_{0,0}) = \mathbf{0},\ \boldsymbol{\pi}_0 \sum_{l=0}^{N} \Phi_l \mathbf{e} = 1.$$

Шаг 5. Находим векторы $\boldsymbol{\pi}_i,\ i = \overline{0, N}$, по формулам

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \Phi_i,\ i = \overline{0, N}.$$

3.6 АСИМПТОТИЧЕСКИ КВАЗИТЕПЛИЦЕВЫ ЦЕПИ МАРКОВА С ДИСКРЕТНЫМ ВРЕМЕНЕМ

Рассмотрим многомерный марковский процесс $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, с пространством состояний

$$\mathcal{S} = \{(i, \mathbf{v}), \mathbf{v} \in \mathcal{V}_i,\ i \geq 0\},$$

где $\mathcal{V}_i$ – некоторое конечное множество, если марковский процесс $\xi_n, n \geq 1$, двумерный, или некоторое конечное множество конечномерных векторов, если марковский процесс $\xi_n, n \geq 1$, многомерный.

Объединим состояния ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, имеющие значение i компоненты i_n, в макро-состояние $\mathbf{i}$, иногда называемое уровнем процесса $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, и перенумеруем уровни в лексикографическом порядке. Состояния внутри уровня могут быть перенумерованы в произвольном

порядке. Без ограничения общности будем считать, что это также лексикографический порядок. Соответственно этой нумерации, матрица P одношаговых вероятностей переходов этого процесса запишется в блочном виде $P = (P_{i,l})_{i,l\geq 0}$, где $P_{i,l}$ есть матрица, образованная вероятностями $p_{(i,\mathbf{r});(l,\boldsymbol{\nu})}$ одношаговых переходов перехода ЦМ $\xi_n, n \geq 1$, из состояния $(i,\mathbf{r}), \mathbf{r} \in \mathcal{V}_i$, в состояние $(l,\boldsymbol{\nu}), \boldsymbol{\nu} \in \mathcal{V}_l$. Число состояний, входящих в уровень $\mathbf{i}$, является мощностью множества $\mathcal{V}_i$ и обозначается как $K_i,\ i \geq 0$.

Дополнительно предполагаем, что существует неотрицательное целое число i^* и множество $\mathcal{V}$ такие, что $\mathcal{V}_i = \mathcal{V}$, если $i > i^*$. Мощность множества $\mathcal{V}_i$ обозначим через K.

Определение 3.1. Неприводимая непериодическая ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$, называется асимптотически квазитеплицевой цепью Маркова (АКТЦМ), если блоки $P_{i,l}$ матрицы P одношаговых вероятностей переходов удовлетворяют условиям:

а) $P_{i,l} = O$, если $l < i-1, i > 1$,

б) Существуют матрицы $Y_k, k \geq 0$, такие, что

$$\lim_{i\to\infty} P_{i,i+k-1} = Y_k,\ k \geq 0, \tag{3.102}$$

и матрица $\sum\limits_{k=0}^{\infty} Y_k$ является стохастической.

Замечание 3.1. Под сходимостью последовательности матриц здесь подразумевается поэлементная сходимость. Поскольку рассматриваются только последовательности неотрицательных матриц конечного порядка, это эквивалентно сходимости по норме.

Замечание 3.2. Матрицы Y_k, $k \geq 0$, несут в себе информацию о предельном поведении ЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$. Их можно рассматривать как матрицы одношаговых вероятностей переходов некоторой квазитеплицевой цепи, предельной по отношению к АКТЦМ.

Далее приведем условия эргодичности и неэргодичности таких АКТЦМ и разработаем алгоритм нахождения их эргодического (стационарного) распределения.

3.6.1 Условия эргодичности и неэргодичности асимптотически квазитеплицевой цепи Маркова

Для получения условий эргодичности и неэргодичности многомерных АКТЦМ будет использован подход, основанный на анализе средних сдвигов (или подход А.А. Ляпунова), см., например, [86], [162], и результаты

работы [170], в которой получены достаточные условия неэргодичности для одномерных цепей Маркова.

Сформулируем прямые аналоги теоремы 4 из [170] и теоремы Мустафы из [162] для многомерной АКТЦМ $\xi_n = \{i_n, \mathbf{v}_n\}, n \geq 1$.

Пусть

- $x_{i,\mathbf{r}}$ – неотрицательная функция, заданная на пространстве состояний $\mathcal{S}$ ЦМ $\xi_n, n \geq 1$;
- $\gamma^x_{i,\mathbf{r}} = \sum\limits_{(j,\boldsymbol{\nu})\in\mathcal{S}} p_{(i,\mathbf{r})(j,\boldsymbol{\nu})}(x_{j,\boldsymbol{\nu}} - x_{i,\mathbf{r}})$ есть x-сдвиг (см. [170]) в состоянии $(i, \mathbf{r})$;
- $\mathbf{X}_i$, $\mathbf{\Gamma}^x_i$ – векторы, составленные из компонент $x_{i,\mathbf{r}}$ и $\gamma^x_{i,\mathbf{r}}$, соответственно, перенумерованных в лексикографическом порядке, $i \geq 0$.

Очевидно, что

$$\mathbf{\Gamma}^x_i = \sum_{j=\max\{0,i-1\}}^{\infty} P_{i,j}\mathbf{X}_j - \mathbf{X}_i, i \geq 0. \tag{3.103}$$

Следуя [170], будем говорить, что ЦМ $\xi_n, n \geq 1$, удовлетворяет обобщенному условию Каплана, если существует положительная константа B, целое число $L > 0$ и константа c, $0 \leq c < 1$, такие, что $\Psi_{i,\mathbf{r}}(z) \geq -B$ для $(i, \mathbf{r}) \in \mathcal{S}$, $i > L$ и $z \in [c, 1)$, где обобщенная функция Каплана $\Psi_{i,\mathbf{r}}(z)$ задана следующим образом:

$$\Psi_{i,\mathbf{r}}(z) = -\sum_{(j,\boldsymbol{\nu})\in\mathcal{S}} p_{(i,\mathbf{r})(j,\boldsymbol{\nu})}(z^{x_{j,\boldsymbol{\nu}}} - z^{x_{i,\mathbf{r}}})/(1-z),\ (i,\mathbf{r}) \in \mathcal{S}.$$

Следующее утверждение является прямым аналогом теоремы 4 из [170].

Утверждение 3.1. *Предположим, что ЦМ $\xi_n, n \geq 1$, удовлетворяет обобщенному условию Каплана, $\mathbf{\Gamma}^x_i < \infty, i \geq 0$, и пространство состояний $\mathcal{S} = \{(i, \mathbf{r}), \mathbf{r} \in \mathcal{R}_i, i \geq 0\}$ разбито на два множества $\mathcal{S}_1, \mathcal{S}_2$ такие, что $\mathcal{S}_1$ – конечное множество, $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2, \mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, и выполняются следующие неравенства:*

$$\min_{(i,\mathbf{r})\in\mathcal{S}_2} x_{i,\mathbf{r}} > \max_{(i,\mathbf{r})\in\mathcal{S}_1} x_{i,\mathbf{r}}, \quad \gamma^x_{i,\mathbf{r}} \geq 0,\ (i,\mathbf{r}) \in \mathcal{S}_2.$$

Тогда ЦМ $\xi_n, n \geq 1$, неэргодична.

Утверждение 3.2. *(аналог теоремы Мустафы из [162]) Для того чтобы ЦМ $\xi_n, n \geq 1$, была эргодична, достаточно, чтобы существовали положительное число ε, положительное целое число J_0, $J_0 \geq i^*$ и множество*

неотрицательных векторов $\mathbf{X}_i, i \geq 0$, *порядка* K *таких, что выполняются следующие неравенства:*

$$\sum_{j=0}^{\infty} P_{i,j}\mathbf{X}_j - \mathbf{X}_i < -\varepsilon\mathbf{e}, \quad i > J_0, \tag{3.104}$$

$$\sum_{j=0}^{\infty} P_{i,j}\mathbf{X}_j < \infty, \quad i = \overline{0, J_0}. \tag{3.105}$$

Известно, что эргодичность или неэргодичность ЦМ определяется ее поведением при больших значениях счетной компоненты. Поэтому интуитивно понятно, что эргодичность или неэргодичность АКТЦМ определяется тем, эргодична или неэргодична предельная для нее КТЦМ. Ниже это будет показано строго.

Пусть $Y(z) = \sum\limits_{k=0}^{\infty} Y_k z^k$, $|z| \leq 1$. Будем различать случаи, когда матрица $Y(1)$ является неприводимой и приводимой.

В случае неприводимой матрицы $Y(1)$ условие эргодичности определяется знаком величины $\beta = [\det(zI - Y(z))]'_{z=1}$.

Для доказательства условия эргодичности понадобится следующая лемма. В ней символ sign в применении к вектору означает, все компоненты вектора имеют один и тот же знак, определенный символом sign.

Лемма 3.6. *Пусть* $Y'(1) < \infty$. *Тогда существует вектор-столбец* $\boldsymbol{\Delta}$ *такой, что* $\mathrm{sign}\boldsymbol{\Delta} = \mathrm{sign}\beta$, *или* $\boldsymbol{\Delta} = \mathbf{0}$, *если* $\beta = 0$, *и система линейных алгебраических уравнений относительно компонент вектора* $\boldsymbol{\alpha}$

$$(I - Y(1))\boldsymbol{\alpha} = (Y(z) - zI)'|_{z=1}\mathbf{e} + \boldsymbol{\Delta} \tag{3.106}$$

имеет бесконечное множество решений.

Доказательство. Так как матрица $Y(1)$ – неприводимая и стохастическая, то ранг матрицы $I - Y(1)$ равен $K - 1$ (см., например, [72]). Пусть $D_n(\boldsymbol{\Delta})$, $n = \overline{1, K}$, есть определитель матрицы Y_n^*, полученной из матрицы $I - Y(1)$ путем замены ее n-го столбца вектором в правой части системы (3.106). Из линейной алгебры известно, что система уравнений (3.106) имеет бесконечное множество решений, если существует вектор $\boldsymbol{\Delta}$, удовлетворяющий системе уравнений

$$D_n(\boldsymbol{\Delta}) = 0, n = \overline{1, K}. \tag{3.107}$$

Установим факт существования такого вектора $\boldsymbol{\Delta}$.

Разлагая определитель $D_n(\boldsymbol{\Delta})$ матрицы Y_n^* по элементам n-го столбца, получаем следующую систему линейных алгебраических уравнений для компонент вектора $\boldsymbol{\Delta}$

$$\mathbf{u}_n\boldsymbol{\Delta} = \mathbf{u}_n(zI - Y(z))'|_{z=1}\mathbf{e},\ n = \overline{1,K}. \tag{3.108}$$

Здесь $\mathbf{u}_n$ – вектор - строка алгебраических дополнений к элементам n-го столбца матрицы $I - Y(1)$.

Из теории матриц известно (см., например, [68], [72]), что все векторы $\mathbf{u}_n,\ n = \overline{1,K}$, для неприводимой стохастической матрицы $Y(1)$ совпадают и положительны:

$$\mathbf{u}_n = \mathbf{u} > 0,\ n = \overline{1,K}. \tag{3.109}$$

Таким образом, система (3.108) эквивалентна уравнению

$$\mathbf{u}\boldsymbol{\Delta} = \mathbf{u}(zI - Y(z))'|_{z=1}\mathbf{e}. \tag{3.110}$$

Правая часть уравнения (3.110) равна $(\det(zI - Y(z)))'|_{z=1} = \beta$. Чтобы убедиться в этом, прибавим все столбцы матрицы $zI - Y(z)$ к произвольно выбранному столбцу, разложим определитель полученной матрицы по элементам этого столбца и продифференцируем его в точке $z = 1$. Как результат, мы получим требуемое соотношение

$$\beta = \mathbf{u}(zI - Y(z))'|_{z=1}\mathbf{e}. \tag{3.111}$$

Таким образом, уравнение (3.110) эквивалентно уравнению $\mathbf{u}\boldsymbol{\Delta} = \beta$.

Поскольку согласно (3.109) $\mathbf{u} > 0$, существует вектор $\boldsymbol{\Delta}$ такой, что $\mathrm{sign}\boldsymbol{\Delta} = \mathrm{sign}\beta$, или $\boldsymbol{\Delta} = \mathbf{0}$, если $\beta = 0$, удовлетворяющий этому уравнению. По построению этот же вектор является решением системы (3.107). Подставляя это решение в (3.106), получаем систему линейных алгебраических уравнений для компонент вектора $\boldsymbol{\alpha}$, которая имеет бесконечное множество решений. □

Теперь можем сформулировать и доказать условия эргодичности и неэргодичности АКТЦМ.

Теорема 3.11. *Пусть матрица $Y(1)$ – неприводимая. Предположим, что ряд $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}\mathbf{e}$ сходится для $i = \overline{0,i^*}$ и ряд $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}$ сходится для любого $i > i^*$ и существует целое число $j_1,\ j_1 > i^*$, такое, что второй ряд равномерно сходится в области $i \geq j_1$.*

Тогда справедливы следующие утверждения:

- *Если* $\beta > 0$, *то АКТЦМ* ξ_n, $n \geq 1$, *эргодична.*
- *Если* $\beta < 0$ *то АКТЦМ* ξ_n, $n \geq 1$, *неэргодична.*

Доказательство. Предположим, что $\beta \neq 0$. Выполнение условий теоремы 3.11 влечет выполнение условия $Y'(1) < \infty$ леммы 3.6. Тогда можно найти вектор $\boldsymbol{\Delta}$ такой, что $\text{sign}\boldsymbol{\Delta} = \text{sign}\beta$ и уравнение (3.106) имеет решение $\boldsymbol{\alpha} = (\alpha_{\boldsymbol{r}})_{\boldsymbol{r}\in\mathcal{R}}$. Возьмем векторную тестовую функцию $\boldsymbol{X}_i, i \geq 0$, вида

$$\boldsymbol{X}_i = \begin{cases} \mathbf{0}, & i < J, \\ (i+1)\mathbf{e} + \boldsymbol{\alpha}, & i \geq J, \end{cases} \tag{3.112}$$

где $J = \max\limits_{\mathbf{r}\in\mathcal{R}}\{j_1, [|\alpha_{\mathbf{r}}|] + 1\}$.

Очевидно, что эта функция неотрицательная. Можно показать, что для $i > J$ векторы $\boldsymbol{\Gamma}_i^x$, определенные формулой (3.103), имеют вид

$$\boldsymbol{\Gamma}_i^x = (\sum_{k=1}^{\infty} kP_{i,i+k-1} - I)\mathbf{e} + (\sum_{k=0}^{\infty} P_{i,\,i+k-1} - I)\boldsymbol{\alpha},\ i > J. \tag{3.113}$$

Из леммы 3.6 следует, что

$$0 = (Y(z) - zI)'|_{z=1}\mathbf{e} + (Y(1) - I)\boldsymbol{\alpha} + \boldsymbol{\Delta}. \tag{3.114}$$

Вычитая (3.114) из (3.113), получаем

$$\boldsymbol{\Gamma}_i^x = -\boldsymbol{\Delta} + \boldsymbol{\epsilon}_i,\ i > J, \tag{3.115}$$

где $\boldsymbol{\epsilon}_i = [\sum\limits_{k=1}^{\infty} kP_{i,i+k-1} - Y'(1)]\mathbf{e} + [\sum\limits_{k=0}^{\infty} P_{i,i+k-1} - Y(1)]\boldsymbol{\alpha},\ i > J.$

Из пункта б) определения 3.1 для АКТЦМ и равномерной сходимости рядов $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}$ для больших значений i следует, что $\boldsymbol{\epsilon}_i$ стремится к $\mathbf{0}$ при $i \to \infty$. Поэтому существует положительное целое число $J_0 \geq \max\{j_1, J\}$ такое, что

$$\text{sign}\,\boldsymbol{\Gamma}_i^x = \text{sign}\,(-\boldsymbol{\Delta}),\ i > J_0. \tag{3.116}$$

Пусть $\beta > 0$.

Из равенства (3.103) и пункта а) определения 3.1 для АКТЦМ следует, что левая часть неравенства (3.104) равна $\boldsymbol{\Gamma}_i^x$. Векторы $\boldsymbol{\Gamma}_i^x$, $i > J$, имеют вид (3.115) и для $i > J_0 \geq J$ удовлетворяют соотношению (3.116). В рассматриваемом случае $\boldsymbol{\Delta} > 0$, поскольку $\beta > 0$. Поэтому из (3.116) следует,

что $\mathbf{\Gamma}_i^x < 0,\ i > J_0$. Это означает, что условие (3.104) утверждения 3.2 выполняется.

Легко проверить, что сходимость рядов $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}\mathbf{e}, i = \overline{0, i^*}$, и рядов $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}, i > i^*$, влечет выполнение условие (3.105) утверждения 3.2.

Таким образом, мы показали, что если $\beta > 0$, то выполняются условия утверждения 3.2. Поэтому ЦМ $\xi_n,\ n \geq 1$, эргодична. Другими словами, выполнение неравенства $\beta > 0$ является достаточным условием эргодичности ЦМ $\xi_n,\ n \geq 1$.

Пусть теперь $\beta < 0$.

Путем несложных преобразований можно убедиться, что для векторной тестовой функции вида (3.112) цепь $\xi_n,\ n \geq 1$, удовлетворяет обобщенному условию Каплана. Принимая во внимание сходимость рядов $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}\mathbf{e}, i = \overline{0, i^*}$, и $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}, i > i^*$, легко видеть, что вектора средних сдвигов $\mathbf{\Gamma}_i^x$ удовлетворяют неравенствам $\mathbf{\Gamma}_i^x < \infty,\ i \geq 0$.

Декомпозируем пространство состояний $\mathcal{S} = \{(i, \mathbf{r}), \mathbf{r} \in \mathcal{R}_i, i \geq 0\}$ ЦМ на два множества $\mathcal{S}_1, \mathcal{S}_2$ следующим образом. Множество $\mathcal{S}_1$ включает состояния $(i, \mathbf{r}), i \leq J_0$. Оставшиеся состояния образуют множество $\mathcal{S}_2$. Теперь переместим все состояния $(j, \boldsymbol{\nu}) \in \mathcal{S}_2$, такие что $x_{j,\boldsymbol{\nu}} \leq \max\limits_{(i,\mathbf{r})\in\mathcal{S}_1} x_{i,\mathbf{r}}$ из множества $\mathcal{S}_2$ в множество $\mathcal{S}_1$. В результате получим конечное множество $\mathcal{S}_1$ и счетное множество $\mathcal{S}_2$ такие, что $\min\limits_{(i,\mathbf{r})\in\mathcal{S}_2} x_{i,\mathbf{r}} > \max\limits_{(i,\mathbf{r})\in\mathcal{S}_1} x_{i,\mathbf{r}}$.

Поскольку множество $\mathcal{S}_2$ включает только состояния $(i, \mathbf{r}), i > J_0$, и $\mathbf{\Delta} < 0$, в силу неравенства $\beta < 0$ из (3.116) следует, что $\gamma_{i,\mathbf{r}}^x > 0, (i, \mathbf{r}) \in \mathcal{S}_2$.

Таким образом, мы показали, что если $\beta < 0$, то выполнены все условия утверждения 3.1. Поэтому ЦМ $\xi_n,\ n \geq 1$, неэргодична, т.е., неравенство $\beta < 0$ является достаточным условием неэргодичности ЦМ $\xi_n,\ n \geq 1$. $\square$

Условия $\beta = [\det(zI - Y(z))]'_{z=1} = \mathbf{u}(zI - Y(z))'|_{z=1}\mathbf{e} > 0 (< 0)$ не очень удобны для проверки. Получим более удобные условия.

Следствие 3.7. *АКТЦМ $\xi_n,\ n \geq 1$, является эргодической, если выполняется неравенство*

$$\mathbf{y}Y'(1)\mathbf{e} < 1 \tag{3.117}$$

и неэргодической, если выполняется неравенство

$$\mathbf{y}Y'(1)\mathbf{e} > 1,$$

где вектор-строка $\mathbf{y}$ *является единственным решением системы уравнений*

$$\mathbf{y}Y(1) = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1. \tag{3.118}$$

Доказательство. Из теории матриц известно (см., например, [72]), что все решения уравнения $\mathbf{y}Y(1) = \mathbf{y}$ имеют вид $\mathbf{y} = c\mathbf{u}$, где $\mathbf{u}$ – вектор-столбец алгебраических дополнений, введенный при доказательстве леммы 3.6, а c – константа. Тогда единственное решение системы (3.118) имеет вид

$$\mathbf{y} = (\mathbf{u}\mathbf{e})^{-1}\mathbf{u}. \tag{3.119}$$

Отметим, что $(\mathbf{u}\mathbf{e})^{-1} > \mathbf{0}$ в силу неравенства (3.109). Из соотношений (3.111) и (3.119) следует, что неравенство $\beta > 0$ ($\beta < 0$) эквивалентно неравенству $\mathbf{y}Y'(1)\mathbf{e} < 1$ ($\mathbf{y}Y'(1)\mathbf{e} > 1$). Поэтому утверждение доказываемого следствия эквивалентно утверждению теоремы 3.11. □

Пусть теперь матрица $Y(1)$ является приводимой. Тогда матричная ПФ $Y(z)$ также приводимая и имеет ту же структуру, что и матрица $Y(1)$. Без ограничения общности предположим, что матрица $Y(z)$ уже представлена в канонической нормальной форме (определение см., например, в приложении А или в [72], [?]), то есть она имеет следующую структуру:

$$Y(z) = \tag{3.120}$$

$$= \begin{pmatrix} Y^{(1)}(z) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & Y^{(2)}(z) & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & Y^{(m)}(z) & 0 & \cdots & 0 \\ Y^{(m+1,1)}(z) & Y^{(m+1,2)}(z) & \cdots & Y^{(m+1,m)}(z) & Y^{(m+1)}(z) & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ Y^{(s,1)}(z) & Y^{(s,2)}(z) & \cdots & Y^{(s,m)}(z) & Y^{(s,m+1)}(z) & \cdots & Y^{(s)}(z) \end{pmatrix},$$

где $Y^{(1)}(z), \ldots, Y^{(m)}(z)$ – неприводимые квадратные матрицы порядков $r_1, \ldots, r_m$ соответственно, $\sum\limits_{n=1}^{m} r_n \leq K$, и в каждой строке $Y^{(n,1)}(z), \ldots, Y^{(n,n-1)}(z)$, $n = m+1, \ldots, s$ есть по крайней мере одна ненулевая матрица.

Пусть $\beta_l = [\det(zI - Y^{(l)}(z))]'_{z=1}$, $l = \overline{1,m}$.

Лемма 3.7. *Существует вектор-столбец* $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_m, \boldsymbol{\Delta}_{m+1})$, *где* $\boldsymbol{\Delta}_l$ *– вектор порядка* r_l, *удовлетворяющий условиям* $\operatorname{sign} \boldsymbol{\Delta}_l = \operatorname{sign} \beta_l$, *если* $\beta_l \neq 0$, $\boldsymbol{\Delta}_l = \mathbf{0}$, *если* $\beta_l = 0$, $l = \overline{1, m}$, *такой, что система линейных уравнений (3.106) имеет бесконечное множество решений.*

Доказательство Леммы 3.7. аналогично доказательству Леммы 3.6.

Теорема 3.12. *Пусть матрица* $Y(1)$ *приводимая с канонической нормальной формой (3.120). Если ряды* $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}\mathbf{e}$, $i = \overline{0, i^*}$, *и* $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}$, $i > i^*$, *удовлетворяют условиям теоремы 6.1, то справедливы следующие утверждения:*

- *Если* $\beta_l > 0$, $l = \overline{1, m}$, *то АКТЦМ* ξ_n, $n \geq 1$, *эргодична.*
- *Если* $\beta_l < 0$, $l = \overline{1, m}$, *то АКТЦМ* ξ_n, $n \geq 1$, *неэргодична.*

Доказательство теоремы 3.12 аналогично доказательству теоремы 3.11 с использованием леммы 3.7 вместо леммы 3.6.

Следствие 3.8. *АКТЦМ* ξ_n, $n \geq 1$, *эргодична, если*

$$\mathbf{y}_l \frac{dY^{(l)}(z)}{dz}|_{z=1}\mathbf{e} < 1,\ l = \overline{1, m}, \tag{3.121}$$

и неэргодична, если

$$\mathbf{y}_l \frac{dY^{(l)}(z)}{dz}|_{z=1}\mathbf{e} > 1,\ l = \overline{1, m}.$$

Здесь вектор-строки $\mathbf{y}_l$, $l = \overline{1, m}$, *являются единственными решениями систем уравнений*

$$\mathbf{y}_l Y^{(l)}(1) = \mathbf{y}_l,\ \mathbf{y}_l\mathbf{e} = 1. \tag{3.122}$$

Замечание 3.3. Если условие б) в определении 3.1 выполняется в более сильной форме, а именно, существует такое целое число $N, N \geq 1$, что $P_{i,i+k-1} = Y_k, k \geq 0$, для всех $i \geq N$, то АКТЦМ является многомерной КТЦМ (или ЦМ типа $M/G/1$) с N граничными состояниями.

В этом случае неравенства (3.117) и (3.121) задают необходимое и достаточное условие эргодичности ЦМ.

3.6.2 Алгоритм для вычисления стационарных вероятностей асимптотически квазитеплицевой цепи Маркова

Предположим, что условия эргодичности для АКТЦМ ξ_n, $n \geq 1$, выполняются, и введем в рассмотрение стационарные вероятности состояний АКТЦМ

$$\pi(i, \mathbf{r}) = \lim_{n \to \infty} P\{i_n = i, \mathbf{r}_n = \mathbf{r}\}.$$

Составим из вероятностей состояний АКТЦМ, имеющих значение i счетной компоненты и перенумерованных в лексикографическом порядке, векторы строки $\boldsymbol{\pi}_i,\ i \geq 0$.

Уравнения Чепмена – Колмогорова для этих векторов имеют вид

$$\boldsymbol{\pi}_j = \sum_{i=0}^{j+1} \boldsymbol{\pi}_i P_{i,j}, j \geq 0. \tag{3.123}$$

При исследовании многомерных КТЦМ в разделе 3.4 уже возникала проблема решения аналогичной системы уравнений (система (3.69). Там при решении системы применялись два подхода: подход, основанный на использовании векторных ПФ, и подход, восходящий к М. Ньютсу. В случае АКТЦМ подход, основанный на использовании векторных ПФ, применим только в редких случаях. Один из таких случаев будет рассмотрен ниже при изучении СМО $BMAP/SM/1$ с повторными вызовами и постоянной интенсивностью повторов. Здесь же основное внимание будет уделено второму подходу.

Как уже отмечалось в разделе 3.4, этот подход базируется на использовании сенсорных ЦМ, см. [137]. В разделе 3.4 идея такого подхода была объяснена на примере одномерных ЦМ, а затем применена для многомерных КТЦМ. Здесь начнем сразу со случая многомерных цепей.

Зафиксируем натуральное число $N,\ N \geq 0$. Обозначим через $\xi_n^{(N)},\ n \geq 1$, сенсорную ЦМ, имеющую пространство состояний $(i, \mathbf{r}),\ \mathbf{r} \in \mathcal{V}_i,\ i = \overline{0, N}$. Обозначим через $P^{(N)} = (P_{i,l}^{(N)})_{i,l=\overline{0,N}}$, матрицу вероятностей переходов сенсорной ЦМ $\xi_n^{(N)},\ n \geq 1$. В общем случае, если представить матрицу вероятностей переходов исходной ЦМ в блочном виде

$$P = \begin{pmatrix} T & U \\ R & Q \end{pmatrix},$$

где T, U, R, Q – матрицы соответствующих порядков, то матрица вероят-

ностей переходов сенсорной ЦМ задается следующим образом:

$$P^{(N)} = T + U\Psi R,$$

где $\Psi = \sum\limits_{i=0}^{\infty} Q^i$.

Полезным свойством сенсорной ЦМ $\xi_n^{(N)}$, $n \geq 1$, является то, что ее стационарные вероятности совпадают с точностью до нормирующего множителя со стационарными вероятностями соответствующих состояний исходной ЦМ.

Поскольку согласно свойству а) АКТЦМ блочная матрица ее вероятностей переходов является блочно-верхне-хессенберговой, матрица $P^{(N)} = (P_{i,l}^{(N)})_{i,l=\overline{0,N}}$, переходных вероятностей сенсорной ЦМ определяется проще, чем в общем случае. В частности, легко видеть, что

$$P_{i,l}^{(N)} = P_{i,l}, i = \overline{0,N}, l = \overline{0,N-1}.$$

Несколько более сложно получается выражение для матриц $P_{i,N}^{(N)}$, поскольку при их вычислении требуется учитывать возможные переходы исходной ЦМ на вырезанных участках ее траектории. При этом учитываем, что любая траектория переходов исходной цепи из состояния, имеющего значение i счетной компоненты цепи, в состояние, имеющее значение j этой компоненты, $i > j \geq 0$, должна как минимум единожды посетить все промежуточные состояния со значениями $i-1, i-2, \ldots, j+1$ счетной компоненты. Эта особенность АКТЦМ в силу ее свойства а) открывает следующий путь для вычисления матриц $P_{i,N}^{(N)}$.

Введем в рассмотрение матрицы

$$G_i = (g_i(\mathbf{r}; \boldsymbol{\nu}))_{\mathbf{r} \in \mathcal{V}_{i+1}, \boldsymbol{\nu} \in \mathcal{V}_i}, i \geq N,$$

элемент $g_i(\mathbf{r}; \boldsymbol{\nu})$ которых есть условная вероятность того, что ЦМ ξ_n, $n \geq 1$, достигнет множества состояний $\{(i, \mathbf{s}), \mathbf{s} \in \mathcal{V}_i\}$, попав в состояние $(i, \boldsymbol{\nu})$, при условии, что она стартует из состояния $(i+1, \mathbf{r})$. Матрицы G_i были введены в рассмотрение М. Ньютсом, см. [**?**], и уже использовались нами в разделе 3.4.

Анализируя поведение ЦМ ξ_n, $n \geq 1$, и используя формулу полной вероятности, можно убедиться, что матрицы $G_i, i \geq N$, удовлетворяют следующим рекуррентным соотношениям:

$$G_i = P_{i+1,i} + \sum_{l=i+1}^{\infty} P_{i+1,l} G_{l-1} G_{l-2} \ldots G_i, i \geq N.$$

Теперь, с учетом рассуждений о необходимости посещения всех промежуточных уровней, легко видеть, что

$$P_{i,N}^{(N)} = P_{i,N} + \sum_{l=1}^{\infty} P_{i,N+l} G_{N+l-1} G_{N+l-2} \dots G_N, i = \overline{0,N}.$$

Таким образом, проблема вычисления блоков матрицы $P^{(N)}$ полностью решена. Векторы стационарных вероятностей $\boldsymbol{\pi}_i^{(N)},\ i = \overline{0,N}$, удовлетворяют системе линейных алгебраических уравнений

$$\boldsymbol{\pi}_l^{(N)} = \sum_{i=0}^{l+1} \boldsymbol{\pi}_i^{(N)} P_{i,l}^{(N)}, l = \overline{0,N-1},$$

$$\boldsymbol{\pi}_N^{(N)} = \sum_{i=0}^{N} \boldsymbol{\pi}_i^{(N)} P_{i,N}^{(N)}. \tag{3.124}$$

Из сказанного выше о том, что стационарные вероятности состояний сенсорной цепи совпадают с точностью до нормирующего множителя со стационарными вероятностями соответствующих состояний исходной ЦМ, следует, что векторы стационарных вероятностей $\boldsymbol{\pi}_i,\ i = \overline{0,N}$, исходной ЦМ также удовлетворяют уравнению (3.124), то есть

$$\boldsymbol{\pi}_N = \sum_{i=0}^{N} \boldsymbol{\pi}_i P_{i,N}^{(N)},$$

или

$$\boldsymbol{\pi}_N = \sum_{i=0}^{N-1} \boldsymbol{\pi}_i P_{i,N}^{(N)} (I - P_{N,N}^{(N)})^{-1}. \tag{3.125}$$

Невырожденность матрицы $I - P_{N,N}^{(N)}$ при $N \geq 1$ следует из того, что матрица $P_{N,N}^{(N)}$ является неприводимой субстохастической.

Поскольку уравнения (3.125) справедливы для любого числа $N,\ N \geq 0$, они задают альтернативную системе (3.123) систему уравнений для вычисления векторов $\boldsymbol{\pi}_j, j \geq 1$.

Из (3.125) следует, что можно выразить все вектора $\boldsymbol{\pi}_l, l \geq 1$, через вектор $\boldsymbol{\pi}_0$ как:

$$\boldsymbol{\pi}_l = \boldsymbol{\pi}_0 \Phi_l,\ l \geq 1, \tag{3.126}$$

где матрицы $\Phi_l,\ l \geq 1$, вычисляются по рекуррентным формулам

$$\Phi_0 = I,\ \ \Phi_l = \sum_{i=0}^{l-1} \Phi_i P_{i,l}^{(l)} (I - P_{l,l}^{(l)})^{-1},\ l \geq 1.$$

Для нахождения неизвестного вектора $\boldsymbol{\pi}_0$ положим в (3.124) $N = 0$. Получаем следующую однородную систему линейных алгебраических уравнений:

$$\boldsymbol{\pi}_0(I - P_{0,0}^{(0)}) = \mathbf{0}. \tag{3.127}$$

Можно показать, что матрица $P_{0,0}^{(0)}$ является неприводимой стохастической. Поэтому система (3.127) определяет вектор $\boldsymbol{\pi}_0$ с точностью до константы, которая определяется из условия нормировки.

Таким образом, с теоретической точки зрения проблема нахождения стационарного распределения АКТЦМ ξ_n, $n \geq 1$, решена. Алгоритм решения состоит из следующих принципиальных шагов:

- Вычисляем матрицы G_i из рекуррентных соотношений

$$G_i = P_{i+1,i} + \sum_{l=i+1}^{\infty} P_{i+1,l} G_{l-1} G_{l-2} \dots G_i, i \geq 0, \tag{3.128}$$

  как

$$G_i = (I - \sum_{l=i+1}^{\infty} P_{i+1,l} G_{l-1} G_{l-2} \dots G_{i+1})^{-1} P_{i+1,i},\ i \geq 0.$$

- Вычисляем матрицы $\bar{P}_{i,l}$ по формулам:

$$\bar{P}_{i,l} = P_{i,l} + \sum_{n=l+1}^{\infty} P_{i,n} G_{n-1} G_{n-2} \dots G_l,\ l \geq i, i \geq 0. \tag{3.129}$$

- Вычисляем матрицы $\Phi_l, l \geq 1$, из рекурсии:

$$\Phi_0 = I,\ \ \Phi_l = \sum_{i=0}^{l-1} \Phi_i \bar{P}_{i,l} (I - \bar{P}_{l,l})^{-1},\ l \geq 1. \tag{3.130}$$

- Вычисляем вектор $\boldsymbol{\pi}_0$ как единственное решение системы уравнений

$$\boldsymbol{\pi}_0(I - \bar{P}_{0,0}) = \mathbf{0}, \tag{3.131}$$

$$\boldsymbol{\pi}_0 \sum_{l=0}^{\infty} \Phi_l \mathbf{e} = 1. \tag{3.132}$$

- Вычисляем векторы $\boldsymbol{\pi}_l,\ l \geq 1$, по формуле (3.126).

Замечание 3.4. Отметим, что все обратные матрицы, фигурирующие в алгоритме, существуют и неотрицательны. Это обусловливает численную устойчивость предложенного алгоритма.

Однако практическая реализация описанного алгоритма сталкивается с некоторыми трудностями. Наиболее существенная из них состоит в организации вычислений матриц G_i из рекурсии (3.128). Дело в том что это – бесконечная обратная рекурсия и матрица G_i может быть вычислена из нее только если известны все матрицы $G_m, m > i$. Чтобы преодолеть эту трудность, проэксплуатируем вероятностный смысл этих матриц и асимптотические свойства АКТЦМ. Напомним, что АКТЦМ при больших значениях компоненты i_n ведет себя аналогично квазитеплицевой цепи с матрицами переходных вероятностей $P_{i,i+k-1} = Y_k,\ k \geq 0$. В разделе 3.4 отмечалось, что для КТЦМ матрицы G_i для всех $i,\ i \geq 0$, совпадают и равны матрице G, которая является минимальным неотрицательным решением нелинейного матричного уравнения

$$G = \sum_{k=0}^{\infty} Y_k G^k \tag{3.133}$$

В разделе 3.4 обсуждались пути решения этого уравнения.

Отсюда следует, что при компьютерной реализации предложенного алгоритма можно использовать рекурсию (3.128), где положено $G_i = G$ для всех $i \geq \tilde{i}$, $\tilde{i}$ – достаточно большое число. Приведем более строгое обоснование такого решения проблемы вычисления по обратной рекурсии.

Обозначим через $R_i = G - \sum\limits_{k=i}^{\infty} P_{i+1,k} G^{k-i}$ невязку в (3.128), где все матрицы $G_m, m \geq i$, заменены матрицей G. Норма матрицы R_i определяет степень близости решения $\{G_i, G_{i+1}, \dots\}$ рекурсии (3.128) к набору матриц $\{G, G, \dots\}$.

Теорема 3.13. *Норма матрицы R_i стремится к нулю при i, стремящемся к бесконечности.*

Доказательство. В качестве нормы $||A||$ произвольной матрицы $A = (a_{i,j})$ будем понимать величину $\max\limits_i \sum\limits_j |a_{i,j}|$. Нетрудно видеть, что для неотрицательных матриц $P_{i,j}$ и стохастической матрицы G выполняется соотношение

$$||\sum_{k=k_0}^{\infty} P_{i+1,i+k} G^k|| = ||\sum_{k=k_0}^{\infty} P_{i+1,i+k}||,\ i \geq 0, k_0 \geq 0. \tag{3.134}$$

Выше было предположено, что условия эргодичности АКТЦМ $\xi_n, n \geq 1$, выполняются. Одним из этих условий является равномерная сходимость

рядов $\sum\limits_{k=1}^{\infty} kP_{i+1,i+k}$ при больших значениях i. Тогда ряд $\sum\limits_{k=0}^{\infty} P_{i+1,i+k}$ также равномерно сходится при больших значениях i. Учитывая равенство (3.134) и замечание 3.1, приходим к выводу, что ряды $\sum\limits_{k=0}^{\infty} P_{i+1,i+k}G^k$ также равномерно сходятся при больших значениях i.

Из этого и (3.102) следует справедливость следующих соотношений:

$$\lim_{i\to\infty}\sum_{k=0}^{\infty} P_{i+1,i+k}G^k = \sum_{k=0}^{\infty}\lim_{i\to\infty} P_{i+1,i+k}G^k = \sum_{k=0}^{\infty} Y_k G^k.$$

Эти соотношения означают, что последовательность матриц

$$R_i = G - \sum_{k=i}^{\infty} P_{i+1,k}G^{k-i} = \sum_{k=0}^{\infty} Y_k G^k - \sum_{k=0}^{\infty} P_{i+1,i+k}G^k,\ i \geq 0,$$

поэлементно сходится к нулевой матрице. Это влечет равенство

$$\lim_{i\to\infty} ||R_i|| = 0.$$

□

Таким образом, численная реализация бесконечной обратной рекурсии (3.128) может быть выполнена следующим образом. Фиксируем некоторое большое число $\tilde{i},\ \tilde{i} \geq i^*$. Это число зависит от скорости сходимости матриц $P_{i,i+k-1}$ к соответствующим матрицам $Y_k, k \geq 0$. Число $\tilde{i}$ должно быть выбрано как минимальное из чисел i, для которых норма $||R_i||$ невязки меньше наперед заданного малого числа ε. Полагаем $G_i = G$ для $i \geq \tilde{i} \geq i^*$ и вычисляем матрицы $G_i, i = \overline{0, \tilde{i}-1}$, из обратной рекурсии (3.128).

Отметим, что значение $\tilde{i}$ может быть найдено и путем сравнения с ε нормы матрицы $G_{\tilde{i}-1} - G$, где $G_{\tilde{i}-1}$ – матрица, вычисленная на первом шаге описанной рекурсии. Если эта норма меньше, чем ε, считаем, что число $\tilde{i}$ действительно является достаточно большим. В противном случае увеличиваем число $\tilde{i}$ и снова делаем один шаг обратной рекурсии. Повторяем эту процедуру до тех пор, пока не получим, что норма матрицы $G_{\tilde{i}-1} - G$ меньше, чем ε. То, что рано или поздно мы найдем подходящее число $\tilde{i}$, вытекает из утверждения теоремы 3.13.

При описанном способе организации вычислений по обратной рекурсии (3.128) формула (3.129) может быть переписана в виде:

$$\bar{P}_{i,l} = P_{i,l} + \sum_{n=l+1}^{\infty} P_{i,n}G^{\max\{0,n-\max\{\tilde{i},l\}\}}G_{\min\{\tilde{i},n\}-1}\dots G_l,\ l \geq i,\ i \geq 0.$$

Описанная процедура выбора значения $\tilde{i}$ может рассматриваться как эвристическая. Но эта процедура неулучшаемая в том смысле, что более точной компьютерной процедуры не может быть построено в принципе, если мы возьмем в качестве ε так называемое компьютерное "эпсилон", то есть минимальное по модулю число, с которым может оперировать данный компьютер.

Отметим, что на последнем шаге описанного выше алгоритма требуется провести усечение бесконечного ряда в (3.132). Это можно сделать путем прекращения вычислений в (3.132) после того, как норма очередной матрицы Φ_l окажется меньше некоторого наперед заданного малого числа ε_1. То, что рано или поздно это произойдет, следует из эргодичности рассматриваемой ЦМ.

Для упомянутых выше многомерных КТЦМ с N граничными состояниями описанный алгоритм упрощается, поскольку здесь все матрицы $G_i, i \geq N$, равны матрице G, вычисленной из уравнения (3.133).

3.7 АСИМПТОТИЧЕСКИ КВАЗИТЕПЛИЦЕВЫ ЦЕПИ МАРКОВА С НЕПРЕРЫВНЫМ ВРЕМЕНЕМ

3.7.1 Определение АКТЦМ с непрерывным временем

В данном подразделе вводятся в рассмотрение АКТЦМ с непрерывным временем. Исследование этих цепей будет существенно опираться на результаты раздела 3.6, в котором изучены АКТЦМ с дискретным временем. При этом будет использоваться тот факт, что процесс с дискретным временем, описывающий переходы ЦМ с непрерывным временем в моменты изменения ее состояний (вложенная цепь, или jump Markov chain), является ЦМ с дискретным временем и связь между этими цепями следующая. Пусть A – генератор ЦМ с непрерывным временем, P – матрица вероятностей переходов ее вложенной цепи, и T – диагональная матрица, диагональные элементы которой совпадают с диагональными элементами генератора A, взятыми с противоположным знаком. Тогда имеет место соотношение

$$P = T^{-1}A + I.$$

Итак, пусть $\xi_t = \{i_t, \mathbf{r}_t\}, t \geq 0$, есть регулярная неприводимая ЦМ

с непрерывным временем. Предполагаем, что эта ЦМ имеет то же пространство состояний $\mathcal{S}$, что и АКТЦМ с дискретным временем, изученная в разделе 3.6, а именно,

$$\mathcal{S} = \{(i, \mathbf{r}), \mathbf{r} \in \mathcal{R}_i, i = 0, 1, \ldots, i^*; (i, \mathbf{r}), \mathbf{r} \in \mathcal{R}, i > i^*\}.$$

Относительно нумерации состояний делаем те же предположения, что и в разделе 3.6: состояния $(i, \mathbf{r})$ перенумерованы в порядке возрастания компоненты i, а при фиксированном значении i состояния $(i, \mathbf{r})$, $\mathbf{r} \in \mathcal{R}_i$, перенумерованы в лексикографическом порядке, $i \geq 0$.

Представим генератор A ЦМ $\xi_t, t \geq 0$, в блочном виде: $A = (A_{i,l})_{i,l \geq 0}$, где $A_{i,l}$ есть матрица размера $K_i \times K_l$, образованная интенсивностями $a_{(i,\mathbf{r});(l,\boldsymbol{\nu})}$ переходов цепи из состояния $(i, \mathbf{r}), \mathbf{r} \in \mathcal{R}_i$, в состояние $(l, \boldsymbol{\nu}), \boldsymbol{\nu} \in \mathcal{R}_l$. Отметим, что для $i, l \geq i^*$ матрицы $A_{i,l}$ являются квадратными матрицами порядка K. Диагональные элементы матрицы $A_{i,i}$ определены как $a_{(i,\mathbf{r});(i,\mathbf{r})} = -\sum\limits_{(j,\boldsymbol{\nu}) \in \mathcal{S} \setminus (i,\mathbf{r})} a_{(i,\mathbf{r});(j,\boldsymbol{\nu})}$.

Обозначим через $T_i, i \geq 0$, диагональную матрицу, имеющую диагональные элементы $-a_{(i,\mathbf{r});(i,\mathbf{r})}$.

Определение 3.2. Регулярная неприводимая ЦМ с непрерывным временем $\xi_t, t \geq 0$, называется асимптотически квазитеплицевой, если

1^0. $A_{i,l} = O$ при $l < i - 1$, $i > 0$.

2^0. Существуют такие матрицы $Y_k, k \geq 0$, что

$$Y_k = \lim_{i \to \infty} T_i^{-1} A_{i,i+k-1}, k = 0, 2, 3, \ldots, \tag{3.135}$$

$$Y_1 = \lim_{i \to \infty} T_i^{-1} A_{i,i} + I, \tag{3.136}$$

и матрица $\sum\limits_{k=0}^{\infty} Y_k$ является стохастической.

3^0. Вложенная ЦМ $\xi_n, n \geq 1$, является непериодической.

Замечание 3.5. Условия $[2^0]$ автоматически выполняются, если справедливы следующие условия:

2^*. Существует матрица T такая, что $\lim\limits_{i \to \infty} T_i^{-1} = T$.

3^*. Существуют целые числа $i_1, k_1 \geq 0$, такие, что матрицы $A_{i,i+k}$ не зависят от i при $i \geq i_1, k \geq k_1$.

4^*. Существуют пределы $\lim\limits_{i \to \infty} T_i^{-1} A_{i,i+k}$, $k = -1, 0, \ldots, k_1 - 1$.

Условия $[2^*]$-$[4^*]$ более ограничительны, чем условия $[2^0]$. Мы приводим их, поскольку они выполняются для ЦМ, описывающих многие многолинейные СМО с повторными вызовами (см., например, систему $BMAP/PH/N$ с повторными вызовами, рассмотренную ниже в разделе 5.2), в то время как они проверяются существенно проще, чем условия $[2^0]$.

Замечание 3.6. Если существует такое целое число $N, N \geq 1$, что $A_{i,i+k-1} = \tilde{A}_k, k \geq 0$, для всех $i \geq N$, то условие $[2^0]$ в определении 3.2 заведомо выполняется. ЦМ с непрерывным временем такого вида будем называть многомерными КТЦМ с непрерывным временем с N-граничными уровнями.

3.7.2 Условия эргодичности асимптотически квазитеплицевой цепи Маркова с непрерывным временем

Для исследования АКТЦМ с непрерывным временем будем рассматривать ее вложенную цепь $\xi_n = \{i_n, \mathbf{r}_n\},\ n \geq 1$, которая имеет пространство состояний $\mathcal{S}$ и матрицы вероятностей одношаговых переходов $P_{i,l},\ i, l \geq 0$, заданные следующим образом:

$$P_{i,l} = \begin{cases} O, & l < i-1,\ i > 0; \\ T_i^{-1} A_{i,l}, & l \geq \max\{0, i-1\},\ l \neq i; \\ T_i^{-1} A_{i,i} + I, & l = i,\ i \geq 0. \end{cases} \tag{3.137}$$

Справедливо следующее утверждение.

Лемма 3.8. *ЦМ $\xi_n,\ n \geq 1$, имеющая матрицы вероятностей одношаговых переходов $P_{i,l},\ i, l \geq 0$, вида (3.137), принадлежит классу АКТЦМ с дискретным временем.*

Доказательство. Покажем, что ЦМ $\xi_n, n \geq 1$, удовлетворяет определению 3.1.

Цепь $\xi_n,\ n \geq 1$, неприводимая, то есть все ее состояния являются сообщающимися, поскольку таковой является ЦМ $\xi_t,\ t \geq 0$, и непериодическая по пункту $[3^0]$ определения 3.2.

Условия а) и б) выполняются вследствие (3.135)-(3.137).

Таким образом, ЦМ $\xi_n,\ n \geq 1$, удовлетворяет всем условиям определения 3.1 и, следовательно, она является АКТЦМ. □

Обозначим через $Y(z)$ ПФ матриц $Y_k,\ k \geq 0$, заданных формулами (3.135) и (3.136):

$$Y(z) = \sum_{k=0}^{\infty} Y_k z^k,\ |z| \leq 1.$$

Достаточное условие эргодичности АКТЦМ $\xi_t,\ t \geq 0$, задается в терминах этой матричной ПФ. Так же, как и в случае АКТЦМ с дискретным временем, будем различать случаи, когда матрица $Y(1)$ является неприводимой и приводимой.

Теорема 3.14. *Пусть матрица $Y(1)$ – неприводимая. Предположим, что:*

(а) ряды $\sum\limits_{k=1}^{\infty} kA_{i,i+k-1}\mathbf{e}$ сходятся при $i = \overline{0, i^*}$;

(б) ряды $\sum\limits_{k=1}^{\infty} kA_{i,i+k-1}$ сходятся для всех $i > i^*$ и существует целое число $j_1 > i^*$, такое, что эти ряды сходятся равномерно при $i \geq j_1$;

(в) последовательность T_i^{-1}, $i \geq 0$, ограничена сверху.

Тогда достаточным условием эргодичности АКТЦМ $\xi_t,\ t \geq 0$, является выполнение неравенства

$$(\det(zI - Y(z)))'|_{z=1} > 0. \tag{3.138}$$

Доказательство. Прежде всего покажем, что при выполнении условий теоремы неравенство (3.138) определяет достаточное условие эргодичности вложенной цепи $\xi_n,\ n \geq 1$. Для этого будем использовать теорему 3.11.

В соответствии с этой теоремой необходимо доказать, что ряды $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}\mathbf{e}$ сходятся при $i = \overline{0, i^*}$, ряды $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}$ сходятся при $i > i^*$ и сходятся равномерно при больших значениях i.

Сходимость рядов $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}\mathbf{e}$ следует из (3.137) и условия (а) доказываемой теоремы.

Ряды $\sum\limits_{k=1}^{\infty} kP_{i,i+k-1}$ при $i > i^*$ представимы в виде

$$\sum_{k=1}^{\infty} kP_{i,i+k-1} = T_i^{-1} \sum_{k=1}^{\infty} kA_{i,i+k-1} + I. \tag{3.139}$$

Сходимость рядов (3.139) при $i > i^*$ следует из сходимости при $i > i^*$ рядов $\sum\limits_{k=1}^{\infty} kA_{i,i+k-1}$. Равномерная сходимость рядов (3.139) следует из рав-

номерной сходимости рядов $\sum\limits_{k=1}^{\infty} kA_{i,i+k-1}$ и равномерной ограниченности матриц T_i^{-1} при больших значениях i.

Таким образом, мы показали, что выполнение неравенства (3.138) является достаточным условием эргодичности ЦМ $\xi_n, n \geq 1$.

Представим эргодическое (стационарное) распределение ЦМ $\xi_n,\ n \geq 1$, в форме $(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots)$, где вектор-строка $\boldsymbol{\pi}_i$ есть вектор вероятностей состояний, имеющих значение i счетной (первой) компоненты. Нетрудно видеть, что для любой константы c вектор-строка

$$(\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2, \dots),$$

где

$$\boldsymbol{p}_i = c\boldsymbol{\pi}_i T_i^{-1},\ i \geq 0,$$

удовлетворяет системе уравнений Чепмена – Колмогорова (уравнениям равновесия) для стационарного распределения вероятностей исходной ЦМ с непрерывным временем $\xi_t,\ t \geq 0$. Поскольку эта цепь является регулярной и неприводимой, то в соответствии с теоремой Фостера, см., например, [86], достаточным условием ее эргодичности является отличие от нуля константы c имеющей вид:

$$c = (\sum_{i=0}^{\infty} \boldsymbol{\pi}_i T_i^{-1}\mathbf{e})^{-1}. \tag{3.140}$$

Учитывая равномерную ограниченность матриц T_i^{-1} при больших значениях i, нетрудно видеть, что ряд в правой части (3.140) сходится в некоторому положительному числу. Поэтому константа, определенная формулой (3.140), конечна и положительна. Следовательно, вектор-строка $(\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{p}_2, \dots)$ при константе вида (3.140) задает стационарное распределение ЦМ $\xi_t, t \geq 0$. □

Следствие 3.9. *Если АКТЦМ удовлетворяет условиям* $[2^*]-[3^*]$ *в замечании 3.5, то условие (в) теоремы 3.14 может быть опущено, а условие (б) сводится к условию:*

(б′) ряды $\sum\limits_{k=1}^{\infty} kA_{i,i+k-1}$ *сходятся при* $i = \overline{i^*+1, i_1-1}$.

Следствие 3.10. *Для АКТЦМ, определенной в замечании 3.6, условие (в) теоремы 3.14 может быть опущено, а условие (б) сводится к условию:*

(б″) ряды $\sum\limits_{k=1}^{\infty} kA_{i,i+k-1}$ *сходятся при* $i = \overline{i^*+1, N}$.

Следствие 3.11. *Неравенство (3.138) эквивалентно неравенству вида*

$$\mathbf{y}Y'(1)\mathbf{e} < 1, \tag{3.141}$$

где вектор-строка $\mathbf{y}$ *является единственным решением системы уравнений*

$$\mathbf{y}Y(1) = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1.$$

Пусть теперь матрица $Y(1)$ приводимая.

Теорема 3.15. *Пусть матрица* $Y(1)$ *приводимая и приведена к нормальной форме вида (3.120). Если выполняются условия (а)-(в) теоремы 3.14, то достаточным условием эргодичности АКТЦМ* $\xi_t,\ t \geq 0$, *является выполнение неравенств*

$$(\det(zI - Y^{(l)}(z)))'|_{z=1} > 0,\ l = \overline{1, m}. \tag{3.142}$$

Доказательство теоремы выполняется по аналогии с доказательством теорем 3.14 и 3.11.

Следствие 3.12. *Неравенства (3.142) эквивалентны неравенствам вида (3.121).*

Проверка условия эргодичности в случае приводимой матрицы $Y(1)$ требует предварительного приведения этой матрицы к нормальной форме. Это приведение производится путем согласованной перестановки строк и столбцов матрицы. Нам не известно формальных процедур для такого приведения, поэтому при рассмотрении ЦМ, описывающих конкретные СМО, требуются определенный опыт и искусство. Случай приводимой матрицы $Y(1)$ достаточно часто имеет место при рассмотрении многолинейных СМО. При этом часто приводимая матрица $Y(1)$ имеет специальный вид, что позволяет при проверке условия эргодичности рассматривать только часть матричной ПФ $Y(z)$. Как правило, эта часть является диагональным блоком матрицы $Y(z)$, имеющим гораздо меньшую размерность, чем вся матрица $Y(z)$. В таком случае могут быть полезными результаты следующих леммы и теоремы.

Лемма 3.9. *Пусть* $Y(1)$ *– приводимая стохастическая матрица порядка* K, *которая может быть представлена в виде*

$$Y(1) = \begin{pmatrix} Y_{11}(1) & Y_{12}(1) \\ O & Y_{22}(1) \end{pmatrix},$$

где $Y_{11}(1)$ *и* $Y_{22}(1)$ *– квадратные матрицы размеров* L_1 *и* L_2 *соответственно,* $0 \le L_1 \le K-1$, $L_1 + L_2 = K$. *Предположим, что все диагональные и поддиагональные элементы матрицы* $Y_{11}(1)$ *равны нулю.*

Пусть также $Y_{22}^{(l)}(1)$, $l = \overline{1,m}$, *являются неприводимыми стохастическими блоками нормальной формы* $Y_{22}^{\{N\}}(1)$ *матрицы* $Y_{22}(1)$.

Тогда эти блоки также являются неприводимыми стохастическими блоками нормальной формы $Y^{\{N\}}(1)$ *матрицы* $Y(1)$, *и матрица* $Y^{\{N\}}(1)$ *не имеет других стохастических блоков.*

Доказательство. Путем согласованной перестановки строк и столбцов матрица $Y(1)$ может быть приведена к виду

$$\hat{Y} = \begin{pmatrix} Y_{22}(1) & O \\ \hat{Y}_{12} & \hat{Y}_{11} \end{pmatrix},$$

где все диагональные и наддиагональные элементы матрицы $\hat{Y}_{11}$ равны нулю.

Приведем блок $Y_{22}(1)$ к его нормальной форме $Y_{22}^{\{N\}}(1)$ путем согласованной перестановки строк и столбцов матрицы $\hat{Y}$. В результате получаем нормальную форму $Y^{\{N\}}(1)$ матрицы $Y(1)$:

$$Y^{\{N\}}(1) = \begin{pmatrix} Y_{22}^{\{N\}}(1) & O \\ \tilde{Y}_{12} & \hat{Y}_{11} \end{pmatrix}.$$

Все неприводимые стохастические диагональные блоки матрицы $Y^{\{N\}}(1)$ содержатся в матрице $Y_{22}^{\{N\}}(1)$, так как неприводимые диагональные блоки матрицы $\hat{Y}_{11}$ являются субстохастическими матрицами размера 1×1, точнее, каждый из этих блоков является скаляром, равным нулю. □

Теорема 3.16. *Пусть выполняются все условия теоремы 3.15 и матрица* $Y(1)$ *удовлетворяет условиям леммы 3.9. Тогда достаточным условием эргодичности АКТЦМ* ξ_t, $t \ge 0$, *является выполнение неравенств*

$$(\det(zI - Y_{22}^{(l)}(z)))|'_{z=1} > 0,\ l = \overline{1,m}.$$

Доказательство теоремы следует из леммы 3.9 и теоремы 3.15.

Замечание 3.7. В случае КТЦМ с непрерывным временем с N граничными уровнями (см. замечание 3.6) матричная ПФ $Y(z)$ вычисляется как

$Y(z) = zI + T^{-1}\tilde{A}(z)$, где $\tilde{A}(z) = \sum\limits_{k=0}^{\infty} \tilde{A}_k z^k$, а T является диагональной матрицей с диагональными элементами, совпадающими с модулями соответствующих диагональных элементов матрицы $\tilde{A}_1$. В этом случае неравенства (3.141) в случае неприводимой матрицы $Y(1)$ или неравенства (3.121) в случае приводимой матрицы $Y(1)$ задают не только достаточное, но и необходимое условие соответствующих ЦМ.

3.7.3 Алгоритм нахождения стационарного распределения вероятностей состояний

Алгоритм нахождения стационарного распределения вероятностей состояний АКТЦМ с непрерывным временем можно получить на основе соответствующего алгоритма для АКТЦМ с дискретным временем, изложенного в разделе 3.6. При этом используется соотношение (3.137) между блоками генератора цепи с непрерывным временем и блоками матрицы переходных вероятностей вложенной ЦМ, а также соотношение

$$\boldsymbol{\pi}_i = c^{-1}\mathbf{p}_i T_i,\ i \geq 0,$$

между векторами стационарных вероятностей $\boldsymbol{\pi}_i$ вложенной ЦМ и $\mathbf{p}_i$ исследуемой ЦМ с непрерывным временем. Здесь константа c определена формулой (3.140).

В результате получается следующая процедура для вычисления векторов стационарных вероятностей $\mathbf{p}_i$, $i \geq 0$:

- Вычисляем матрицы G_i из обратной рекурсии

$$G_i = (-\sum_{n=i+1}^{\infty} A_{i+1,n} G_{n-1} G_{n-2} \dots G_{i+1})^{-1} A_{i+1,i},\ i \geq 0. \qquad (3.143)$$

- Вычисляем матрицы $\bar{A}_{i,l}$, $l \geq i$, $i \geq 0$, по формулам

$$\bar{A}_{i,l} = A_{i,l} + \sum_{n=l+1}^{\infty} A_{i,n} G_{n-1} G_{n-2} \dots G_l,\ \ l \geq i,\ i \geq 0. \qquad (3.144)$$

- Вычисляем матрицы F_l, $l \geq 0$, используя рекуррентные соотношения

$$F_0 = I,\ F_l = \sum_{i=0}^{l-1} F_i \bar{A}_{i,l} (-\bar{A}_{l,l})^{-1},\ l \geq 1. \qquad (3.145)$$

- Вычисляем вектор $\mathbf{p}_0$ как единственное решение системы линейных алгебраических уравнений

$$\mathbf{p}_0(-\bar{A}_{0,0}) = \mathbf{0},\ \mathbf{p}_0 \sum_{l=0}^{\infty} F_l \mathbf{e} = 1.$$

- Вычисляем векторы $\mathbf{p}_l$, $l \geq 1$, по формуле

$$\mathbf{p}_l = \mathbf{p}_0 F_l,\ l \geq 1. \tag{3.146}$$

Отметим, что обратные матрицы в (3.143), (3.145) существуют и неотрицательны. Проблема нахождения терминального условия для рекурсии (3.143) решается аналогично тому, как она решалась в разделе 3.6 для ЦМ с дискретным временем.

ГЛАВА 4
СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ С ОЖИДАНИЕМ С КОРРЕЛИРОВАННЫМИ ПОТОКАМИ И ИХ ПРИМЕНЕНИЕ ДЛЯ ОЦЕНКИ ПРОИЗВОДИТЕЛЬНОСТИ СЕТЕВЫХ СТРУКТУР

СМО с ожиданием, т.е. системы, в которых имеется входной буфер и запрос, поступивший, когда все приборы заняты, помещается в этот буфер и обслуживается позже, когда появится свободный прибор, являются адекватными математическими моделями многих реальных систем, включая проводные и беспроводные телекоммуникационные системы. В данной главе будут изложены результаты анализа нескольких таких систем в предположении, что входной поток является $BMAP$-потоком.

4.1 СИСТЕМА $BMAP/G/1$

Система $BMAP/G/1$ — это однолинейная СМО с бесконечным буфером, на вход которой поступает $BMAP$-поток запросов. Как было описано выше, $BMAP$ характеризуется управляющим процессом $\nu_t, t \geq 0$, с пространством состояний $\{0, 1, \ldots, W\}$ и матричной ПФ $D(z)$. Времена обслуживания заявок взаимно независимы и характеризуются ФР $B(t)$ с ПЛС $\beta(s)$ и конечными начальными моментами $b_k = \int\limits_0^\infty t^k dB(t),\ k = 1, 2.$

Представляет интерес исследование процессов i_t — числа запросов в системе в момент t, и w_t — виртуального времени ожидания в системе в момент $t,\ t \geq 0$. Предметом исследования являются стационарные распределения этих процессов и характеристики производительности системы, вычисляемые на основе этих распределений.

Процессы i_t, w_t не являются марковскими, поэтому непосредственно найти их стационарное распределение не представляется возможным. Для решения задачи требуется осуществить их марковизацию.

Для марковизации будем использовать сочетание двух методов марковизации – метода вложенных ЦМ и расширение пространства состояний путем введения дополнительных переменных.

Сначала рассмотрим проблему существования и нахождения стацио-

нарного распределения специально построенной вложенной ЦМ. А затем опишем технику перехода от стационарного распределения этой цепи к стационарному распределению процессов i_t и w_t.

4.1.1 Стационарное распределение вложенной цепи Маркова

Поскольку распределение числа запросов, поступающих в $BMAP$-потоке, зависит от состояний управляющего процесса $BMAP$-потока $\nu_t,\ t \geq 0$, естественно сделать вывод, что построение марковского процесса не обойдется без учета состояния этого процесса. Сначала рассмотрим двумерный процесс $\{i_t, \nu_t\},\ t \geq 0$. Поскольку и этот двумерный процесс не является марковским, для его исследования применим метод вложенных ЦМ.

Пусть $t_k,\ k \geq 1$, есть k-й момент окончания обслуживания запроса, i_{t_k} – число запросов в системе в момент $t_k + 0$, ν_{t_k} – состояние управляющего процесса в момент t_k. Нетрудно видеть, что двумерный процесс $\xi_k = \{i_{t_k}, \nu_{t_k}\},\ k \geq 1$, является цепью Маркова.

Ее одношаговые переходные вероятности

$$P\{(i,\nu) \to (l,\nu')\} = P\left\{i_{t_{k+1}} = l, \nu_{t_{k+1}} = \nu' | i_{t_k} = i, \nu_{t_k} = \nu\right\},$$

$$i > 0,\ l \geq i-1,\ \nu,\nu' = \overline{0,W},$$

упорядоченные в лексикографическом порядке, объединим в матрицы размера $\bar{W} \times \bar{W}$

$$\Omega_{l-i+1} = \int_0^\infty P(l-i+1,t)dB(t),\ i > 0,\ l \geq i-1.$$

Матрицы Ω_m, $m \geq 0$, играют важную роль в анализе систем с $BMAP$-потоком. Вопросы вычисления матриц $P(n,t),\ n \geq 0$, были обсуждены выше, в разделе 3.1, см., например, формулу (3.7). ПФ этих матриц определяется формулой (3.4). Из этой формулы следует, что ПФ матриц Ω_m, $m \geq 0$, является матрица

$$\sum_{m=0}^{\infty} \Omega_m z^m = \beta(-D(z)) = \int_0^\infty e^{D(z)t} dB(t).$$

Сопоставляя вид одношаговых переходных вероятностей рассматриваемой цепи с определением ЦМ типа $M/G/1$, мы видим, что данная цепь

принадлежит классу ЦМ типа $M/G/1$. Таким образом, задача нахождения ее стационарного распределения (и условий его существования) может быть эффективно решена с использованием изложенных в предыдущей главе результатов, если мы зададим явный вид матричных ПФ $Y(z)$ и $V(z)$.

Матрица $Y(z)$ является ПФ вероятностей переходов цепи из состояний с ненулевым значением счетной компоненты i. У рассматриваемой цепи $\{i_{t_k}, \nu_{t_k}\}$, $k \geq 1$, эти вероятности переходов описываются матрицами Ω_m, $m \geq 0$. Таким образом, получаем:

$$Y(z) = \sum_{m=0}^{\infty} \Omega_m z^m = \beta(-D(z)) = \int_0^{\infty} e^{D(z)t} dB(t), \tag{4.1}$$

при этом, естественно, матрицы Y_m в разложении $Y(z) = \sum\limits_{m=0}^{\infty} Y_m z^m$ совпадают с матрицами Ω_m, $m \geq 0$.

Для нахождения вида матричной ПФ $V(z)$ вероятностей переходов цепи из состояний с нулевым значением счетной компоненты проанализируем поведение СМО после попадания ее в состояние, когда $i = 0$, то есть очередь в системе в момент окончания обслуживания пуста. В этом случае система остается пустой до тех пор, пока в нее не поступит некоторая группа запросов. При этом переходы состояний управляющего процесса BMAP за время ожидания, завершающегося приходом группы, состоящей из k, $k \geq 1$, запросов, характеризуются, как отмечалось в разделе 3.1, матрицей $e^{D_0 t} D_k dt$ при условии, что это время равняется t. Отсюда с учетом того, что дальнейшие переходы цепи осуществляются так же, как и переходы, начинающиеся с моментов окончания обслуживания запросов, в которые в системе находится k, $k \geq 1$, запросов, а также с использованием формулы полной вероятности, получаем следующее выражение для матрицы V_m:

$$V_m = \sum_{k=1}^{m+1} \int_0^{\infty} e^{D_0 t} D_k dt Y_{m+1-k} = -D_0^{-1} \sum_{k=1}^{m+1} D_k Y_{m+1-k}. \tag{4.2}$$

Соответственно, ПФ $V(z) = \sum\limits_{m=0}^{\infty} V_m z^m$ имеет вид:

$$V(z) = (-D_0)^{-1} \frac{1}{z} (D(z) - D_0) \beta(-D(z)). \tag{4.3}$$

Подставляя полученные выражения для матричных ПФ в (4.1), получаем уравнение для векторной ПФ $\mathbf{\Pi}(z)$ в виде:

$$\mathbf{\Pi}(z)\left(zI-\beta(-D(z))\right)=\mathbf{\Pi}(0)(-D_0)^{-1}D(z)\beta(-D(z)). \tag{4.4}$$

Получим условие существования стационарного распределения цепи ξ_k, $k\geq 1$, основываясь на Следствии 3.4. Поскольку матричная ПФ $Y(z)$ имеет вид (4.1), несложно убедиться, что вектор $\mathbf{x}$, являющийся решением уравнения (3.67), равен вектору $\boldsymbol{\theta}$. Учитывая приведенную в разделе 3.1 формулу для вычисления величины

$$\boldsymbol{\theta}\frac{d\int\limits_0^\infty e^{D(z)t}dB(t)}{dz}|_{z=1}\mathbf{e},$$

можно убедиться, что необходимым и достаточным условием существования стационарного распределения вложенной ЦМ является выполнение неравенства

$$\rho=\lambda b_1<1, \tag{4.5}$$

где λ — интенсивность $BMAP$-потока, определенная в разделе 3.1, а $b_1=\int\limits_0^\infty(1-B(t))dt$ — среднее время обслуживания запроса. Далее будем считать это условие выполненным.

Для применения алгоритма нахождения стационарного распределения вложенной ЦМ, описанного выше, требуется указать способ вычисления величин $Y^{(m)}(1)$ и $V^{(m)}(1)$ – m-й производной матричных ПФ $Y(z)$ и $V(z)$ в точке $z=1$, $m\geq 0$. Несмотря на относительно простой вид (4.1), (4.3) этих функций, задача нахождения матриц $Y^{(m)}(1)$ и $V^{(m)}(1)$, $m\geq 0$, является довольно сложной даже в относительно простых случаях.

Приведем два примера.

Пример 1. Если входной поток является $MMPP$-потоком, задаваемым матрицами $D_1=\Lambda$, $D_0=-\Lambda+H$, где $H=\Phi(P-I)$, $D_k=0$, $k\geq 2$, $\Lambda=\mathrm{diag}\{\lambda_0,\ldots,\lambda_W\}$, то есть диагональная матрица с диагональными элементами λ_ν, $\nu=\overline{0,W}$, $\Phi=\mathrm{diag}\{\varphi_0,\ldots,\varphi_W\}$, P — стохастическая матрица, а время обслуживания имеет гиперэрланговское распределение, то есть

$$B(t)=\sum_{i=1}^{k}q_i\int_0^t\frac{\gamma_i(\gamma_i\tau)^{h_i-1}}{(h_i-1)!}e^{-\gamma_i\tau}d\tau,$$

$$q_i \geq 0,\ \sum_{i=1}^{k} q_i = 1,\ \gamma_i > 0,\ h_i \geq 1,\ i = \overline{1,k},\ k \geq 1,$$

то

$$\int_0^\infty e^{D(z)t} dB(t) = \sum_{i=1}^{k} q_i (\gamma_i)^{h_i} (\gamma_i I - H + \Lambda - \Lambda z)^{-h_i}.$$

Тогда

$$\Omega_l = \sum_{i=1}^{k} q_i \gamma_i^{h_i} D_l^{(i)},\ l \geq 0,$$

где

$$D_l^{(i)} = \sum_{(n_1,\ldots,n_{h_i}) \in N_{h_i}{}^{(l)}} \Gamma_{n_{h_i}}^{(i)} \prod_{r=1}^{h_i - 1} \Gamma_{n_{h_i} - n_{h_i - r + 1}}^{(i)},$$

$$\Gamma_l^{(i)} = (S_i \Lambda)^l S_i,\ S_i = (\gamma_i I - H + \Lambda)^{-1},$$

$$N_{h_i}^{(l)} = \{(n_1, \ldots, n_{h_i}) : l = n_1 \geq n_2 \geq \ldots \geq n_{h_i} \geq 0\}.$$

В этом случае формулы для вычисления матриц $Y^{(m)}(1)$, $m = \overline{1,2}$, следующие:

$$Y^{(1)}(1) = \sum_{i=1}^{k} q_i \left[\sum_{r=1}^{h_i} \gamma_i^{h_i - r} \Lambda \bar{S}_i^{h_i + 1 + r} - I \right],$$

$$Y^{(2)}(1) = \sum_{i=1}^{k} q_i \left[\sum_{r=1}^{h_i} \sum_{l=0}^{h_i - r} \gamma_i^{r+l-1} \Lambda \bar{S}_i^{l+1} \Lambda \bar{S}_i^{r} \right],$$

где

$$\bar{S}_i = (\gamma_i I - H)^{-1}.$$

Пример 2. Если время обслуживания запросов имеет экспоненциальное распределение с параметром γ, а входной поток является $BMMPP$-потоком с геометрическим распределением размера группы, имеющим параметр q_ν во время пребывания управляющего процесса ν_t в состоянии ν, $\nu = \overline{0, W}$, то

$$D_0 = -\Lambda + H,\ D(z) = D_0 + \Lambda\theta(z),$$

где

$$\theta(z) = \mathrm{diag}\{\theta_0(z), \ldots, \theta_W(z)\},$$

$$\theta_\nu(z) = z \frac{1 - q_\nu}{1 - q_\nu z},\ 0 < q_\nu < 1,\ \nu = \overline{0, W}.$$

При этом

$$\int_0^\infty e^{D(z)t}dB(t) = \gamma(\gamma I - H + \Lambda - \Lambda\theta(z))^{-1}$$

и

$$\Omega_0 = \gamma S,\ S = (\gamma I - H + \Lambda)^{-1},$$

$$\Omega_l = \gamma((S(\Lambda + Q\tilde{S}^{-1}))^l - Q(S(\Lambda + Q\tilde{S}^{-1}))^{l-1}),\ l \geq 1,$$

$$\tilde{S} = (\gamma I - H)^{-1},\ Q = \mathrm{diag}\{q_0, \ldots, q_W\}.$$

В этом случае формулы для вычисления матриц $Y^{(m)}(1)$, $m = \overline{1,2}$, следующие:

$$Y^{(1)}(1) = \tilde{Q}\Lambda\tilde{S} - I,$$

$$Y^{(2)}(1) = \tilde{Q}\Lambda\tilde{S}\tilde{Q}\Lambda\tilde{S} + \tilde{Q}^2\Lambda\tilde{S},$$

где $\tilde{Q} = \mathrm{diag}\left\{\frac{1}{1-q_0}, \ldots, \frac{1}{1-q_W}\right\}$, $\tilde{S} = (\gamma I - H)^{-1}$.

В случае произвольного $BMAP$-потока и произвольного распределения $B(t)$ времени обслуживания величины $Y^{(m)}(1)$, $m \geq 1$, можно вычислить с использованием формулы (3.7) следующим образом:

$$\Omega_m = \int_0^\infty \sum_{j=0}^\infty \frac{(\tilde{\theta}t)^j}{j!} e^{-\tilde{\theta}t} K_m^{(j)} dB(t),\ m \geq 0,$$

$$Y^{(m)}(1) = \sum_{l=m}^\infty l(l-1)\ldots(l-m+1)\Omega_l,\ m \geq 1.$$

Матрицы $V^{(m)}(1)$ находятся через матрицы $Y^{(m)}(1)$, $m \geq 0$, на основе соотношения (4.3).

Для иллюстрации работы алгоритма вычисления стационарного распределения КТЦМ, основанного на методе производящих функций и описанного выше, расмотрим численный пример вычисления стационарного распределения рассматриваемой СМО.

Пусть управляющий процесс ν_t, $t \geq 0$, $BMAP$-потока имеет пространство состояний $\{0, 1, 2, 3\}$, то есть $W = 3$, а поведение потока описывается матрицами D_0, D_1, D_2 вида

$$D_0 = \begin{pmatrix} -1,45 & 0,2 & 0,15 & 0,1 \\ 0,2 & -2,6 & 0,1 & 0,3 \\ 0,2 & 0,1 & -3,7 & 0,4 \\ 0,1 & 0,05 & 0,15 & -4,3 \end{pmatrix},$$

$$D_1 = D_2 = \mathrm{diag}\{0,5;\ 1;\ 1,5;\ 2\}.$$

Времена пребывания управляющего процесса $\nu_t,\ t \geq 0$, в состояниях $\{0, 1, 2, 3\}$ имеют экспоненциальное распределение с параметрами $\{1,45;\ 2,6;\ 3,7;\ 4,3\}$ соответственно. В моменты скачков процесса либо происходит его "холостой" переход в другое состояние (то есть переход без генерации заявок), либо с равной вероятностью поступает одна или две заявки и процесс $\nu_t,\ t \geq 0$, сохраняет свое состояние до следующего момента скачка.

Инфинитезимальный генератор процесса $\nu_t,\ t \geq 0$, имеет вид

$$D(1) = D_0 + D_1 + D_2 = \begin{pmatrix} -0,45 & 0,2 & 0,15 & 0,1 \\ 0,2 & -0,6 & 0,1 & 0,3 \\ 0,2 & 0,1 & -0,7 & 0,4 \\ 0,1 & 0,05 & 0,15 & -0,3 \end{pmatrix},$$

а вычисленный вектор-строка $\boldsymbol{\theta}$ его стационарного распределения имеет вид:

$$\boldsymbol{\theta} = (0,238704;\ 0,144928;\ 0,167945;\ 0,448423).$$

Интенсивность $BMAP$-потока $\lambda \approx 4,23913$, то есть в среднем в единицу времени в систему поступает $4,239$ запросов.

Предположим, что время обслуживания запросов имеет распределение Эрланга с параметрами (3, 15), то есть

$$B(t) = \int_0^t \frac{15(15\tau)^2}{2!} e^{-15\tau} d\tau,\ \beta(s) = \left(\frac{15}{15+s}\right)^3,\ b_1 = \frac{3}{15} = 0,2.$$

Матричное ПЛС $\beta(-D(z))$ имеет вид:

$$\beta(-D(z)) = 15^3(15I - D_0 - D_1 z - D_2 z^2)^{-3}.$$

Как уже отмечалось выше, для рассматриваемой цепи условие существования стационарного распределения имеет вид (4.5) и нет необходимости выполнения шага 1 алгоритма. Коэффициент загрузки ρ здесь равен 0.8478, поэтому цепь обладает стационарным распределением.

Опишем выполнение шагов алгоритма, на которых вычисляется это распределение.

Уравнение (3.74) имеет три простых корня в единичном круге комплексной плоскости:

$$z_1 = 0,737125,\ z_2 = 0,787756,\ z_3 = 0,063899,$$

и простой корень $z = 1$.

Матрица $A = Y(1) - I$ имеет вид:

$$A = \begin{pmatrix} -0,082987 & 0,035370 & 0,026676 & 0,020941 \\ 0,036077 & -0,109504 & 0,018607 & 0,054820 \\ 0,035864 & 0,018279 & -0,125627 & 0,071493 \\ 0,019084 & 0,009720 & 0,026836 & -0,055641 \end{pmatrix}.$$

Матрица $\tilde{\mathbf{A}}$ имеет вид:

$$\tilde{\mathbf{A}} = \begin{pmatrix} -0,669415 & 0,035370 & 0,026676 & 0,020941 \\ -0,381314 & -0,109504 & 0,018607 & 0,054820 \\ -0,103749 & 0,018270 & -0,125627 & 0,071493 \\ 0,179084 & 0,009720 & 0,026836 & -0,055641 \end{pmatrix}.$$

Матрица $H(1) = Y(1) - V(1)$ имеет вид:

$$\begin{pmatrix} 0,263048 & -0,091349 & -0,077829 & -0,093871 \\ -0,048899 & 0,191608 & -0,037651 & -0,105058 \\ -0,034798 & -0,024618 & 0,152498 & -0,093082 \\ -0,017330 & -0,011891 & -0,026394 & 0,055615 \end{pmatrix}.$$

Матрица $H(1)\tilde{I} + H^{(1)}(1)\mathbf{e}\hat{\mathbf{e}}$ имеет вид:

$$\begin{pmatrix} -1,663108 & -0,091349 & -0,077829 & -0,093871 \\ -1,562058 & 0,191608 & -0,037651 & -0,105058 \\ -1,499574 & -0,024618 & 0,152498 & -0,093082 \\ -1,468385 & -0,011891 & -0,026394 & 0,055615 \end{pmatrix}.$$

Здесь $\tilde{I} = \mathrm{diag}\{0, 1, \ldots, 1\}$, $\hat{\mathbf{e}} = (1, 0, \ldots, 0)$.

Матрица S имеет вид:

$$\begin{pmatrix} 2,956759 & 3,557284 & 4,301839 & 11,832143 \\ 3,551012 & 0,462866 & 2,630622 & 7,060740 \\ 2,576184 & 1,617995 & 0,658583 & 5,082831 \\ 2,016273 & 1,248130 & 1,425717 & 2,820937 \end{pmatrix}.$$

Матрица системы линейных алгебраических уравнений (3.76), (3.77) имеет вид:

$$\begin{pmatrix} 22,648025 & -0,000001 & 0,000026 & 0,000000 \\ 13,705240 & -0,000015 & -0,000037 & -0,000000 \\ 9,935593 & 0,000011 & -0,000003 & -0,000001 \\ 7,511057 & 0,000006 & -0,000002 & 0,000001 \end{pmatrix}.$$

Вектор $\mathbf{\Pi}(0)$, являющийся решением системы уравнений (3.76), (3.77), имеет вид:

$$\mathbf{\Pi}(0) = (0,0239881;\ 0,0150405;\ 0,0113508;\ 0,0183472).$$

Векторы $\boldsymbol{\pi}_l$, вычисленные по формулам (3.80), имеют вид:

$$\boldsymbol{\pi}_1 = (0,0205056;\ 0,0150091;\ 0,0133608;\ 0,0244394),$$

$$\boldsymbol{\pi}_2 = (0,0129573;\ 0,0110476;\ 0,0121506;\ 0,0256954),$$

$$\boldsymbol{\pi}_3 = (0,010361;\ 0,00883173;\ 0,0110887;\ 0,0259778),\ldots$$

В силу довольно большого значения коэффициента загрузки системы $\rho = 0,8478$ "хвост" распределения вероятностей $\boldsymbol{\pi}_l,\ l \geq 0$, довольно "тяжелый", то есть удельный вес вероятностей с большими номерами l довольно велик. Так, сумма всех компонент векторов $\boldsymbol{\pi}_l,\ l = \overline{0,14}$, равна $0,668468$, то есть с вероятностью $0,331532$ число запросов в системе превышает 14. В этой ситуации вычисление факториальных моментов непосредственно по их определению $\mathbf{\Pi}^{(m)} = \sum\limits_{i=m}^{\infty} \frac{i!}{(i-m)!}\boldsymbol{\pi}_i,\ m \geq 1$, через стационарные вероятности $\boldsymbol{\pi}_l,\ l \geq 0$, неэффективен. Для их вычисления следует использовать формулу (3.81). Задавшись целью вычислить факториальные моменты $\mathbf{\Pi}^{(m)}$ для $m = \overline{0,3}$, мы последовательно получаем:

$$\mathbf{\Pi}^{(0)} = \mathbf{\Pi}(1) = (0,190571;\ 0,133559;\ 0,176392;\ 0,499478),$$

$$\mathbf{\Pi}^{(1)} = (2,34708;\ 1,56858;\ 2,3965;\ 7,76206),$$

$$\mathbf{\Pi}^{(2)} = (67,9937;\ 44,6851;\ 69,0566;\ 229,087),$$

$$\mathbf{\Pi}^{(3)} = (2994,5;\ 1963,99;\ 3034,92;\ 10103,2).$$

Заметим, что если мы изменим значение 15 интенсивности фазы эрланговского распределения на значение 25, то есть коэффициент загрузки ρ уменьшится до величины $0,5087$, то сумма всех компонент векторов $\boldsymbol{\pi}_l$, $l = \overline{0,14}$, составит уже $0,9979135$, то есть "хвост" распределения числа запросов в системе будет "легкий". При этом значения факториальных моментов, вычисленные непосредственно по определению через векторы стационарных вероятностей, очень близки к вычисленным по формуле (3.81).

4.1.2 Стационарное распределение вероятностей состояний системы в произвольный момент времени

В предыдущем параграфе исследована вложенная ЦМ $\xi_k = \{i_{t_k}, \nu_{t_k}\}$, $k \geq 1$, посредством метода векторных ПФ. Но, как уже отмечалось выше, эта цепь имеет вспомогательный характер. Более интересным с точки зрения практики является процесс $\zeta_t = \{i_t, \nu_t\}$, $i \geq 0$. Нетрудно видеть, что в отличие от вложенной ЦМ $\{i_{t_k}, \nu_{t_k}\}$, $k \geq 1$, процесс $\{i_t, \nu_t\}$, $t \geq 0$, не является марковским. Вместе с тем, интуитивно понятно, что распределение цепи $\{i_{t_k}, \nu_{t_k}\}$ несет в себе информацию о распределении процесса $\{i_t, \nu_t\}$. Привлечь эту информацию для нахождения стационарного распределения процесса $\{i_t, \nu_t\}$ позволяет подход, основанный на использовании вложенных процессов марковского восстановления [112].

В качестве вложенного процесса марковского восстановления (ПМВ) для нашего процесса $\zeta_t, t \geq 0$, естественно рассматривать ПМВ $\{i_{t_k}, \nu_{t_k}, t_k\}$, $k \geq 1$. Полумарковское ядро $Q(t)$ процесса $\{i_{t_k}, \nu_{t_k}, t_k\}$, $k \geq 1$, имеет блочную структуру вида

$$Q(t) = \begin{pmatrix} \tilde{V}_0(t) & \tilde{V}_1(t) & \tilde{V}_2(t) & \cdots \\ \tilde{Y}_0(t) & \tilde{Y}_1(t) & \tilde{Y}_2(t) & \cdots \\ O & \tilde{Y}_0(t) & \tilde{Y}_1(t) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где блоки $\tilde{V}_l(t)$, $\tilde{Y}_l(t)$, $l \geq 0$, порядка $(W+1) \times (W+1)$ определяются следующим образом:

$$\tilde{Y}_l(t) = \int_0^t P(l,u)dB(u),$$

$$\tilde{V}_l(t) = \sum_{k=1}^{l+1} \int_0^t e^{D_0(t-u)} D_k \tilde{Y}_{l+1-k}(u)du.$$

Обозначим через $H(t)$ матричную функцию восстановления процесса $\{i_{t_k}, \nu_{t_k}, t_k\}$, $k \geq 1$. Матрица бесконечного размера $H(t)$ имеет блочную структуру: $H(t) = (H_{i,j}(t))_{i,j \geq 0}$, где блок $H_{i,j}(t)$ имеет вид: $H_{i,j}(t) = \left(h_{i,j}^{\nu,r}(t)\right)_{\nu,r=\overline{0,W}}$. Элемент $h_{i,j}^{\nu,r}(t)$ матрицы $H(t)$ есть математическое ожидание числа посещений цепью $\{i_{t_k}, \nu_{t_k}\}$ состояния (j,r) на интервале $(0,t)$ при условии, что в начале этого интервала цепь находилась в состоянии

(i,ν). Можно показать, что величина $dh_{i,j}^{\nu,r}(u)$ есть элемент условной вероятности того, что на интервале $(u, u+du)$ цепь $\{i_{t_k}, \nu_{t_k}\}$ посетит состояние (j,r). Зафиксируем некоторое состояние (i_0,ν_0) процесса $\zeta_t,\ t \geq 0$. Для удобства дальнейших выкладок обозначим $\mathbf{h}_j(t),\ j \geq 0$, вектор-строку с номером ν_0 матрицы $H_{i_0,j}(t)$, то есть $\mathbf{h}_j(t) = \left(h_{i_0,j}^{\nu_0,0}(t), \ldots, h_{i_0,j}^{\nu_0,W}(t)\right)$.

Пусть $p(i,\nu,t),\ i \geq 0,\ \nu = \overline{0,W}$, есть распределение вероятностей состояний процесса $\{i_t, \nu_t\},\ t \geq 0$, в момент времени t. Введем в рассмотрение векторы $\mathbf{p}(i,t) = (p(i,0,t), \ldots, p(i,W,t)),\ i \geq 0$. Без потери общности будем считать, что момент времени $t = 0$ совпадает с моментом окончания обслуживания запроса, и в этот момент цепь $\{i_{t_k}, \nu_{t_k}\}$ находилась в состоянии (i_0,ν_0). Используя вероятностную трактовку величин $dh_{i_0,j}^{\nu_0,r}(t)$ и формулу полной вероятности, получим следующие выражения для векторов $\mathbf{p}(i,t)$:

$$\mathbf{p}(0,t) = \int_0^t d\mathbf{h}_0(u) e^{D_0(t-u)},$$

$$\mathbf{p}(i,t) = \int_0^t d\mathbf{h}_0(u) \int_0^{t-u} e^{D_0 v} \sum_{k=1}^{i} D_k dv P(i-k, t-u-v)(1-B(t-u-v)) +$$

$$+ \sum_{k=1}^{i} \int_0^t d\mathbf{h}_k(u) P(i-k, t-u)(1-B(t-u)),\ i > 0. \qquad (4.6)$$

Далее нам понадобится понятие фундаментального среднего ПМВ. Величина τ фундаментального среднего ПМВ равна математическому ожиданию длины интервала между соседними моментами восстановления. Ее можно вычислить как скалярное произведение инвариантного вектора матрицы $Q(\infty)$ и столбца $\int_0^\infty t dQ(t)\mathbf{e}$ математических ожиданий сумм элементов строк матрицы $Q(t)$. Для рассматриваемой системы величина τ определяется как математическое ожидание длины интервала между соседними моментами окончания обслуживания запросов, и для нее получается следующее интуитивно понятное выражение:

$$\tau = b_1 + \boldsymbol{\pi}_0 (-D_0)^{-1} \mathbf{e}. \qquad (4.7)$$

Первое слагаемое есть среднее время обслуживания запроса, а второе задает среднее время простоя прибора после произвольного момента окончания обслуживания. Отметим, что при работе системы в стационарном

режиме величина τ равна величине, обратной средней интенсивности λ поступления запросов в $BMAP$-потоке, то есть

$$\tau = \lambda^{-1}. \tag{4.8}$$

Найдем теперь стационарное распределение процесса $\{i_t, \nu_t\}$, $t \geq 0$. Заметим, что условие существования такого распределения задается неравенством (4.5). Чтобы найти само это распределение, перейдем в (4.6) к пределу при $t \to \infty$. К правым частям (4.6) можно применить узловую теорему марковского восстановления [112], которая в данном случае имеет вид

$$\lim_{t\to\infty} \sum_{k=0}^{\infty} \int_0^t d\mathbf{h}_k(u) G_i(k, t-u) = \frac{1}{\tau} \sum_{k=0}^{\infty} \boldsymbol{\pi}_k \int_0^{\infty} G_i(k,u)du, \ i \geq 0. \tag{4.9}$$

Здесь через $G_i(k, t-u)$ обозначены подинтегральные выражения правых частей соотношений (4.6), то есть

$$G_0(0, t-u) = e^{D_0 t - u},$$

$$G_i(0, t-u) = \int_0^{t-u} e^{D_0 v} \sum_{k=1}^{i} D_k dv P(i-k, t-u-v)(1 - B(t-u-v)),$$

$$G_i(k, t-u) = P(i-k, t-u)(1 - B(t-u)), \ 1 \leq k \leq i, \ i \geq 1,$$

$$G_i(k, u) = O, \ k > i \geq 0, \ u \geq 0.$$

После предельного перехода в (4.6) используем формулу (4.9) и получим следующие выражения для векторов $\mathbf{p}(i) = \lim\limits_{t\to\infty} \mathbf{p}(i,t)$, $i \geq 0$, стационарного распределения вероятностей состояний процесса $\{i_t, \nu_t\}$, $t \geq 0$:

$$\mathbf{p}(0) = \lambda \boldsymbol{\pi}_0 \int_0^{\infty} e^{D_0 u} du = \lambda \boldsymbol{\pi}_0 (-D_0)^{-1},$$

$$\mathbf{p}(i) = \lambda \boldsymbol{\pi}_0 \int_0^{\infty} e^{D_0 v} \sum_{k=1}^{i} D_k dv \int_0^{\infty} P(i-k, u)(1 - B(u))du +$$

$$+ \lambda \sum_{k=1}^{i} \boldsymbol{\pi}_k \int_0^{\infty} P(i-k, u)(1-B(u))du = \lambda \sum_{k=1}^{i} (\boldsymbol{\pi}_0 (-D_0)^{-1} D_k + \boldsymbol{\pi}_k) \tilde{\Omega}_{i-k}, \ i > 0, \tag{4.10}$$

где $\tilde{\Omega}_m = \int\limits_0^\infty P(m,u)(1-B(u))du,\ m \geq 0.$

Отметим, что матрицы $\tilde{\Omega}_m$ легко рассчитываются рекуррентно через матрицы Ω_m, проблема вычисления которых обсуждалась в разделе 3.1, а именно:

$$\tilde{\Omega}_0 = (\Omega_0 - I)D_0^{-1},$$

$$\tilde{\Omega}_m = (\Omega_m - \sum_{k=0}^{m-1} \tilde{\Omega}_k D_{m-k})D_0^{-1},\ m \geq 1.$$

Умножая равенства (4.10) на соответствующие степени z, суммируя и учитывая равенство

$$\int\limits_0^\infty e^{D(z)t}(1-B(t))dt = (-D(z))^{-1}(I - \beta(-D(z)))$$

и уравнение (4.4), после алгебраических преобразований убеждаемся в справедливости следующего утверждения.

Теорема 4.1. *Векторная ПФ* $\mathbf{p}(z) = \sum\limits_{i=0}^{\infty} \mathbf{p}(i)z^i$ *стационарного распределения* $\mathbf{p}(i)$, $i \geq 0$, *состояний рассматриваемой системы в произвольный момент времени выражается через ПФ* $\mathbf{\Pi}(z)$ *стационарного распределения состояний системы в моменты окончания обслуживания запросов следующим образом:*

$$\mathbf{p}(z)D(z) = \lambda(z-1)\mathbf{\Pi}(z). \tag{4.11}$$

Разлагая соотношение (4.11) в ряд Маклорена и приравнивая коэффициенты при одинаковых степенях z, получаем рекуррентные формулы, позволяющие последовательно рассчитывать наперед заданное число векторов $\mathbf{p}(i)$:

$$\mathbf{p}(0) = -\lambda\boldsymbol{\pi}_0 D_0^{-1}, \tag{4.12}$$

$$\mathbf{p}(i+1) = \left[\sum_{j=0}^{i} \mathbf{p}(j)D_{i+1-j} - \lambda(\boldsymbol{\pi}_i - \boldsymbol{\pi}_{i+1})\right](-D_0)^{-1},\ i \geq 0. \tag{4.13}$$

Из (4.7), (4.8) и (4.12) вытекает равенство

$$\mathbf{p}(0)\mathbf{e} = 1 - \rho. \tag{4.14}$$

Таким образом, вероятность того, что система $BMAP/G/1$ в произвольный момент времени пуста, равна величине $1-\rho$, так же как и в случае системы $M/G/1$.

Факториальные моменты $\mathbf{P}^{(m)},\ m \geq 0$, распределения $\mathbf{p}(i),\ i \geq 0$, рассчитываются аналогично факториальным моментам $\mathbf{\Pi}^{(m)}$. Момент 0-го порядка $\mathbf{P}^{(0)}$ совпадет с вектором $\boldsymbol{\theta}$ стационарных вероятностей управляющего процесса $BMAP$-потока, вычисляемым как решение системы уравнений (1.9).

Последующие моменты $\mathbf{P}^{(m)}$ вычисляются рекуррентно:

$$\mathbf{P}^{(m)} = \left[\left(m\lambda\mathbf{\Pi}^{(m-1)} - \sum_{l=0}^{m-1} C_m^l \mathbf{P}^{(l)} D^{(m-l)}\right)\tilde{I} + \right.$$

$$\left. + \left(\lambda\mathbf{\Pi}^{(m)}\mathbf{e} - \frac{1}{m+1}\sum_{l=0}^{m-1} C_{m+1}^l \mathbf{P}^{(l)} D^{(m+1-l)}\mathbf{e}\right)\hat{\mathbf{e}}\right]\hat{A}^{-1}, \qquad (4.15)$$

где

$$\hat{A} = D(1)\tilde{I} + D^{(1)}\mathbf{e}\hat{\mathbf{e}}.$$

Продолжим рассмотрение численного примера, описанного в предыдущем разделе. Имея вычисленные значения стационарных вероятностей $\boldsymbol{\pi}_i,\ i \geq 0$, и факториальных моментов $\mathbf{\Pi}^{(m)}$ вложенной цепи и формулы (4.12)-(4.15), легко можем найти вероятности $\mathbf{p}(i),\ i \geq 0$, и факториальные моменты $\mathbf{P}^{(m)}$.

Так, значения векторов $\mathbf{p}(i),\ i = \overline{0,2}$, следующие:

$$\mathbf{p}(0) = (0,0786320; 0,0317218; 0,0180150; 0,0238051),$$

$$\mathbf{p} = (0,0219269; 0,0146647; 0,0116901; 0,0196986),$$

$$\mathbf{p}(2) = (0,0179116; 0,0137114; 0,0127269; 0,0240296).$$

Значения факториальных моментов $\mathbf{P}^{(m)},\ m = \overline{0,3}$, следующие:

$$\mathbf{P}^{(0)} = (0,238704; 0,144928; 0,167945; 0,448423),$$

$$\mathbf{P}^{(1)} = (2,10506; 1,41165; 2,13705; 6,87336),$$

$$\mathbf{P}^{(2)} = (60,2861; 39,6502; 61,2431; 202,954),$$

$$\mathbf{P}^{(3)} = (2653,26; 1740,33; 2689,32; 8951,38).$$

Вероятность p_0 того, что система пуста в произвольный момент, вычисляется как $p_0 = \mathbf{p}(0)\mathbf{e} = 1-\rho = 0,1521739$.

Среднее число запросов в системе L легко вычисляется следующим образом:

$$L = \mathbf{P}^{(1)}\mathbf{e} = 12,52712.$$

Отметим, что при аппроксимации рассматриваемого $BMAP$-потока стационарным пуассоновским потоком с такой же интенсивностью $\lambda = 4,239$ величина среднего числа запросов в системе $\hat{L}$ для соответствующей системы $M/G/1$, вычислилась бы как

$$\hat{L} = \rho + \frac{\lambda^2 b_2}{2(1-\rho)} = 3,99613,$$

где $b_2 = \int_0^\infty t^2 dB(t)$.

Если теперь сопоставить значение L, подсчитанное точно, и его приближенное значение $\hat{L}$, то можно констатировать, что ошибка аппроксимации превышает 300 %.

Небезинтересно отметить также тот факт, что если использовать еще более грубую аппроксимацию системы $BMAP/G/1$ — аппроксимацию системой $M/M/1$, то есть предположить дополнительно, что время обслуживания запросов имеет экспоненциальное, а не эрланговское распределение с тем же математическим ожиданием, то получаем приближенное значение $\tilde{L}$ как

$$\tilde{L} = \frac{\rho}{1-\rho} = 5,5703.$$

То есть приближенная формула дает погрешность примерно в $2,24$ раза. Более грубая аппроксимация дает меньшую ошибку, чем менее грубая, видимо, за счет взаимного погашения двух погрешностей.

4.1.3 Распределение виртуального и реального времени ожидания в системе

Напомним, что w_t означает виртуальное время ожидания в момент времени t, $t \geq 0$. Пусть $\mathbf{W}(x)$ есть вектор-строка, ν-й элемент которой есть стационарная вероятность того, что в произвольный момент времени управляющий процесс $BMAP$-потока находится в состоянии ν, $\nu = \overline{0, W}$, а виртуальное время ожидания не превысит величину x:

$$\mathbf{W}(x) = \lim_{t\to\infty} \{P\{w_t < x, \nu_t = 0\}, \ldots, P\{w_t < x, \nu_t = W\}\}.$$

Обозначим через $\mathbf{w}(s) = \int\limits_0^\infty e^{-sx} d\mathbf{W}(x)$ вектор-строку, состоящую из ПЛС компонент вектора $\mathbf{W}(x)$.

Теорема 4.2. *Векторное ПЛС* $\mathbf{w}(s)$ *определяется следующим образом:*

$$\mathbf{w}(s)\,(sI + D(\beta(s))) = s\mathbf{p}(0),\ Re\ s > 0, \tag{4.16}$$

где $D(\beta(s)) = \sum\limits_{k=0}^{\infty} D_k(\beta(s))^k$, $\beta(s) = \int\limits_0^\infty e^{-st} dB(t)$.

Доказательство. Очевидно, что $\mathbf{W}(x) = \mathbf{W}(+0) + \mathbf{T}_0(x) + \mathbf{T}_1(x)$, где $\mathbf{W}(+0)$ есть вектор вероятностей того, что виртуальный запрос поступит в пустую систему, т.е. не будет ждать, а немедленно начнет обслуживаться, $\mathbf{T}_0(x)$ есть вектор вероятностей того, что такой запрос прибудет во время первого обслуживания в периоде занятости системы и будет ждать время, не большее x, $\mathbf{T}_1(x)$ есть вектор вероятностей того, что виртуальный запрос прибудет во время обслуживания второго, третьего и т.д. запроса в периоде занятости системы и будет ждать не больше x единиц времени.

Анализируя поведение системы и используя опыт доказательства теоремы в предыдущем параграфе, нетрудно получить следующее выражение для векторной функции $\mathbf{T}_0(x)$:

$$\mathbf{T}_0(x) = \lambda\boldsymbol{\pi}_0 \sum_{i=0}^{\infty} \int\limits_0^\infty dt \sum_{k=1}^{\infty} \int\limits_0^t e^{D_0 v} D_k P(i, t-v) dv \times \tag{4.17}$$

$$\times \int\limits_0^x dB(t+u-v) B^{(i+k-1)}(x-u).$$

Здесь $B^{(m)}(y)$ – свертка m-го порядка ФР $B(y)$, являющаяся ФР суммы m независимых случайных величин, каждая из которых имеет ФР $B(y)$.

При выводе формулы (4.17) переменная t ассоциируется с произвольным моментом времени (моментом поступления виртуального запроса), v — предшествующий ему момент, когда в пустую систему поступила группа из k запросов и один из запросов начал обслуживаться, u — оставшееся к моменту t время обслуживания этого запроса. Отсчет времени начинается с момента окончания предыдущего периода занятости (момента, когда система оказалась пустой в момент окончания обслуживания).

Меняя местами порядок интегрирования по t и v, формулу (4.17) перепишем в виде:

$$\mathbf{T}_0(x) = \mathbf{p}(0)\sum_{i=0}^{\infty}\sum_{k=1}^{\infty}\int_0^{\infty} dyD_kP(i,y)\int_0^{x} dB(y+u)B^{(i+k-1)}(x-u). \quad (4.18)$$

Аналогично выписываем выражение для векторной функции $\mathbf{T}_1(x)$:

$$\mathbf{T}_1(x) = \lambda\sum_{i=1}^{\infty}\sum_{k=1}^{i}\boldsymbol{\pi}_k\int_0^{\infty} dtP(i-k,t)\int_0^{x} dB(t+u)B^{(i-1)}(x-u). \quad (4.19)$$

От совместных векторных ФР $\mathbf{T}_m(x)$, переходим к их ПЛС:

$$\mathbf{T}_m^*(s) = \int_0^{\infty} e^{-sx}d\mathbf{T}_m(x),\ m = 0,1.$$

В результате получаем формулы

$$\mathbf{T}_0^*(s) = \mathbf{p}(0)\sum_{i=0}^{\infty}\sum_{k=1}^{\infty}\int_0^{\infty}\int_0^{\infty} e^{-su}D_kP(i,t)dB(t+u)dt\beta^{i+k-1}(s), \quad (4.20)$$

$$\mathbf{T}_1^*(s) = \lambda\sum_{i=1}^{\infty}\sum_{k=1}^{i}\boldsymbol{\pi}_k\int_0^{\infty}\int_0^{\infty} e^{-su}P(i-k,t)dB(t+u)dt\beta^{i-1}(s).$$

Умножая соотношения (4.20) справа на матрицу $(sI + D(\beta(s)))$, получаем:

$$T_0^*(s)\,(sI + D(\beta(s))) = sT_0^*(s)+$$

$$+\mathbf{p}(0)\sum_{i=0}^{\infty}\sum_{k=1}^{\infty}\int_0^{\infty}\int_0^{\infty} e^{-su}D_kP(i,y)dB(y+u)dy(\beta(s))^{i+k-1}\sum_{r=0}^{\infty}D_r(\beta(s))^r =$$

$$= sT_0^*(s) + \mathbf{p}(0)\sum_{k=1}^{\infty}D_k\mathcal{A}_k(s),$$

где через $\mathcal{A}_k(s)$ обозначено выражение

$$\mathcal{A}_k(s) = \int_0^{\infty}\int_0^{\infty} e^{-su}\sum_{i=0}^{\infty}P(i,y)dB(y+u)dy(\beta(s))^{i+k-1}\sum_{r=0}^{\infty}D_r(\beta(s))^r.$$

Делая замену в этом выражении переменной суммирования $i' = i + r$, получаем:

$$\mathcal{A}_k(s) = \sum_{i=0}^{\infty}\sum_{r=0}^{\infty}\int\limits_0^{\infty}\int\limits_0^{\infty} P(i,y)D_r e^{-su}dB(y+u)dy(\beta(s))^{i+k+r-1} =$$

$$= \sum_{i'=0}^{\infty}\sum_{r=0}^{i'}\int\limits_0^{\infty}\int\limits_0^{\infty} P(i'-r,y)D_r e^{-su}dB(y+u)dy(\beta(s))^{i'+k-1}.$$

Учитывая теперь матричное дифференциальное уравнение (3.2), записанное в виде

$$P'(i,y) = \sum_{r=0}^{i} P(i-r,y)D_r,$$

получаем:

$$\mathcal{A}_k(s) = \sum_{i=0}^{\infty}\int\limits_0^{\infty}\int\limits_0^{\infty} P'(i,y)e^{-su}dB(y+u)dy(\beta(s))^{i+k-1}.$$

Отсюда следует справедливость формулы:

$$T_0^*(s)D(\beta(s)) = \mathbf{p}(0)\sum_{k=1}^{\infty} D_k \int\limits_0^{\infty}\sum_{i=0}^{\infty}\int\limits_0^{\infty} P'(i,y)e^{-su}dB(y+u)dy(\beta(s))^{i+k-1}.$$

Упростим эту формулу. Внутренний интеграл в ее правой части подсчитаем по частям:

$$\int\limits_0^{\infty}\int\limits_0^{\infty} P'(i,y)e^{-su}dB(y+u)dy = \begin{bmatrix} v = y+u \\ u = v-y \end{bmatrix} =$$

$$= \int\limits_0^{\infty} e^{-sv}\int\limits_0^{v} P'(i,y)e^{sy}\,dy\,dB(v) = \begin{bmatrix} u = e^{sy}, & du = se^{sy} \\ dv = P'(i,y)dy, & v = P(i,y) \end{bmatrix} =$$

$$= \int\limits_0^{\infty} e^{-sv}\left((P(i,y)e^{sy})\big|_0^v - s\int\limits_0^{v} P(i,y)e^{sy}\,dy\right)dB(v) =$$

$$= \int\limits_0^{\infty}\left(P(i,v) - P(i,0)e^{-sv} - e^{-sv}s\int\limits_0^{v} P(i,y)e^{sy}\,dy\right)dB(v).$$

$(4.20),, sT_0^*(s) = s\mathbf{p}(0)\sum_{i=0}^{\infty}\sum_{k=1}^{\infty} D_k \int\limits_0^{\infty} e^{-sv} \int\limits_0^{v} P(i,y)e^{sy}\,dy\,dB(v)(\beta(s))^{i+k-1}.T_0^*(s)\,(sI +$

$\mathbf{p}(0)\sum_{k=1}^{\infty} D_k \int\limits_0^{\infty}\sum_{i=0}^{\infty}(P(i,v) - P(i,0)e^{-sv})\,dB(v)(\beta(s))^{i+k-1}.Y_l =$

$\int\limits_0^{\infty} P(l,v)dB(v)P(i,0) = \begin{cases} I, & i=0, \\ 0, & i \neq 0, \end{cases} T_0^*(s)\,(sI + D(\beta(s))) =$

$\mathbf{p}(0)\sum_{k=1}^{\infty} D_k \sum_{i=0}^{\infty} Y_i(\beta(s))^{i+k-1} - \mathbf{p}(0)\sum_{k=1}^{\infty} D_k(\beta(s))^k.\text{l=i+k-}$

$1,(4.2)(4.12), T_0^*(s)\,(sI + D(\beta(s))) = -\mathbf{p}(0)\sum_{k=1}^{\infty} D_k(\beta(s))^k +$

$\lambda\boldsymbol{\pi}_0\sum_{k=0}^{\infty} V_k(\beta(s))^k.T_1^*(s),(4.20) :T_1^*(s)\,(sI + D(\beta(s))) =$

$sT_1^*(s)++\lambda\sum_{l=1}^{\infty}\boldsymbol{\pi}_l\sum_{i=0}^{\infty}\int\limits_0^{\infty} P(i,y)\int\limits_0^{\infty} e^{-su}\,dB(y +$

$u)dy(\beta(s))^{i+l-1}\sum_{r=0}^{\infty} D_r(\beta(s))^r == sT_1^*(s) +$

$\lambda\sum_{l=1}^{\infty}\boldsymbol{\pi}_l\sum_{i=0}^{\infty}\int\limits_0^{\infty}\int\limits_0^{\infty}\dot{P}(i,y)e^{-su}dB(y + u)dy(\beta(s))^{i+l-1} ==$

$\lambda\sum_{l=1}^{\infty}\boldsymbol{\pi}_l\sum_{i=0}^{\infty}\int\limits_0^{\infty}(P(i,v) - P(i,0)e^{-sv})dB(v)(\beta(s))^{i+l-1} ==$

$-\lambda\sum_{l=1}^{\infty}\boldsymbol{\pi}_l(\beta(s))^l + \lambda\sum_{l=1}^{\infty}\boldsymbol{\pi}_l\sum_{i=0}^{\infty} Y_i(\beta(s))^{i+l-1}.\mathbf{W}(x) = \mathbf{W}(+0) +$

$\mathbf{T}_0(x) + \mathbf{T}_1(x), \mathbf{w}(s)(sI + D(\beta(s))) = (W(+0) + T_0^*(s) +$

$T_1^*(s))(sI + D(\beta(s))) == \mathbf{p}(0)sI + \mathbf{p}(0)\sum_{k=0}^{\infty} D_k(\beta(s))^k -$

$\mathbf{p}(0)\sum_{k=1}^{\infty} D_k(\beta(s))^k + \lambda\boldsymbol{\pi}_0\sum_{k=0}^{\infty} V_k(\beta(s))^k - -\lambda\sum_{l=1}^{\infty}\boldsymbol{\pi}_l(\beta(s))^l +$

$\lambda\sum_{l=1}^{\infty}\boldsymbol{\pi}_l\sum_{i=0}^{\infty} Y_i(\beta(s))^{i+l-1}.\sum_{i=0}^{\infty}\boldsymbol{\pi}_i(\beta(s))^i = \boldsymbol{\pi}_0\sum_{i=0}^{\infty} V_i(\beta(s))^i +$

$\sum_{i=0}^{\infty}\sum_{k=1}^{i+1}\boldsymbol{\pi}_k Y_{i-k+1}(\beta(s))^i == \boldsymbol{\pi}_0\sum_{i=0}^{\infty} V_i(\beta(s))^i +$

$\sum\limits_{l=1}^{\infty}\boldsymbol{\pi}_l\sum\limits_{i=0}^{\infty} Y_i(\beta(s))^{i+l-1},(3.69)\boldsymbol{\pi}_i, \quad i \geq 0,$ упрощаем правую часть уравнения для векторной функции $\mathbf{w}(s)$ и приводим его к виду

$$\mathbf{w}(s)(sI + D(\beta(s))) = s\mathbf{p}(0),$$

что и требовалось доказать.

□

Следствие 4.1. *Вектор* $\mathbf{W}_1$, *ν-я компонента которого является средним значением виртуального времени ожидания в момент, когда состояние управляющего процесса BMAP-потока равно* ν, $\nu = \overline{0, W}$, *вычисляется следующим образом:*

$$\mathbf{W}_1 = (-(\mathbf{p}(0) + \boldsymbol{\theta}(D'(1)b_1 - I))\tilde{I} - \boldsymbol{\theta}\frac{1}{2}(D''(1)b_1^2 + D'(1)b_2)\mathbf{e}\hat{\mathbf{e}})\times$$

$$\times(D(1)\tilde{I} + (D'(1)b_1 - I)\mathbf{e}\hat{\mathbf{e}})^{-1}. \qquad (4.21)$$

Доказательство. Введем разложения

$$\mathbf{w}(s) = \mathbf{w}(0) + \mathbf{w}'(0)s + \mathbf{w}''(0)\frac{s^2}{2} + o(s^2),$$

$$D(\beta(s)) = D(1) - D'(1)b_1 s + (D''(1)b_1^2 + D'(1)b_2)\frac{s^2}{2} + o(s^2).$$

Подставляя их в формулу (4.16) и приравнивая коэффициенты в левой и правой частях при одинаковых степенях s, получаем следующие уравнения:

$$\mathbf{w}(0)D(1) = \mathbf{0}, \tag{4.22}$$

$$\mathbf{w}'(0)D(1) + \mathbf{w}(0)(I - D'(1)b_1) = \mathbf{p}(0), \tag{4.23}$$

$$\frac{1}{2}\mathbf{w}''(0)D(1) + \mathbf{w}'(0)(I - D'(1)b_1) + \mathbf{w}(0)(D''(1)b_1^2 + D'(1)b_2)\frac{1}{2} = \mathbf{0}. \tag{4.24}$$

Из уравнения (4.22) очевидным образом следует соотношение $\mathbf{w}(0) = \boldsymbol{\theta}$. Соотношение (4.21) выводится из формул (4.23) и (4.24) по аналогии с доказательством Следствия 3.6 с учетом того, что $\mathbf{W}_1 = -\mathbf{w}'(0)$. □

Следствие 4.2. *Пусть* $\mathbf{V}(x)$ *есть вектор-строка,* ν*-й элемент которой равен вероятности того, что в произвольный момент управляющий процесс BMAP-потока находится в состоянии* ν*, а виртуальное время пребывания этого запроса в системе не превосходит* x *и* $\mathbf{v}(s) = \int\limits_0^\infty e^{-sx} d\mathbf{V}(x)$. *Тогда*

$$\mathbf{v}(s)\,(sI + D(\beta(s))) = s\mathbf{p}(0)\beta(s). \tag{4.25}$$

Формула (4.25) элементарно следует из (4.21) с учетом связи времени пребывания со временем ожидания и временем обслуживания.

Далее рассмотрим распределение реального время ожидания в системе. Пусть $w_t^{(a)}$ означает реальное время ожидания запроса, если он поступит в момент времени t, $t \geq 0$. Пусть $W^{(a)}(x)$ есть стационарная вероятность того, что реальное время ожидания не превысит величину x, $W^{(a)}(x) = \lim\limits_{t\to\infty} P\{w_t^{(a)} < x\}$, и $w^{(a)}(s) = \int\limits_0^\infty e^{-sx} dW^{(a)}(x)$, $\mathrm{Re}\ s \geq 0$, есть ПЛС ФР $W^{(a)}(x)$. Предполагаем, что произвольный запрос, поступивший в группе размера k, с вероятностью $\frac{1}{k}$ будет обслужен j-м из запросов этой группы, $j = \overline{1,k}$.

Теорема 4.3. *ПЛС* $w^{(a)}(s)$ *распределения реального времени ожидания запроса имеет вид*

$$w^{(a)}(s) = -\lambda^{-1}\mathbf{w}(s)(1 - \beta(s))^{-1}D(\beta(s))\mathbf{e}. \tag{4.26}$$

Доказательство. Реальное время ожидания произвольного помеченного запроса состоит из виртуального времени ожидания, которое начинается в момент поступления группы, которой принадлежит данный запрос, и времени обслуживания запросов этой группы, которые будут обслужены раньше помеченного запроса. Вектор $\mathbf{w}(s)$ задает вероятность ненаступления катастрофы из стационарного пуассоновского потока катастроф с параметром s за виртуальное время ожидания при соответствующих состояниях управляющего процесса ν_t $BMAP$ в момент поступления упомянутой группы. При фиксированном распределении вероятностей состояний управляющего процесса в произвольный момент времени, вероятности того, что помеченный запрос поступает в составе группы размера k, задаются вектором $\frac{kD_k\mathbf{e}}{\lambda}$ (см. подраздел 3.1). Как было предположено выше, с вероятностью $\frac{1}{k}$ этот запрос будет обслужен j-м из запросов этой группы, $j = \overline{1,k}$. При этом вероятность ненаступления катастрофы за время обслуживания $j-1$ запросов, которые будут обслужены раньше помеченного, равна $(\beta(s))^{j-1}$. Учитывая приведенные рассуждения, из формулы полной вероятности выводится следующая формула:

$$w^{(a)}(s) = \lambda^{-1}\mathbf{w}(s)\sum_{k=1}^{\infty} kD_k\mathbf{e}\sum_{j=1}^{k}\frac{1}{k}(\beta(s))^{j-1}.$$

Отсюда формула (4.26) следует очевидным образом. □

Следствие 4.3. *Среднее время ожидания $W_1^{(a)}$ запроса в системе задается следующим образом:*

$$W_1^{(a)} = \lambda^{-1}[\mathbf{W}_1 D'(1) + \frac{1}{2}b_1\boldsymbol{\theta}D''(1)]\mathbf{e}. \tag{4.27}$$

Следствие 4.4. *Среднее время пребывания $V_1^{(a)}$ запроса в системе задается следующим образом:*

$$V_1^{(a)} = \lambda^{-1}[\mathbf{W}_1 D'(1) + \frac{1}{2}b_1\boldsymbol{\theta}D''(1)]\mathbf{e} + b_1. \tag{4.28}$$

Отметим, что можно показать, что для данной системы выполняется формула Литтла в следующем виде:

$$V_1^{(a)} = \lambda^{-1}L,$$

где L — среднее число запросов в системе в произвольный момент времени.

4.2 СИСТЕМА $BMAP/SM/1$

В СМО $BMAP/G/1$ предполагалось, что времена обслуживания запросов являются независимыми одинаково распределенными (с функцией распределения $B(t)$) случайными величинами. Во многих реальных системах времена обслуживания последовательных запросов существенно зависимы и могут быть распределены по разным законам. В литературе для моделирования таких процессов обслуживания было предложено использовать формализм SM — процессов, описанный в разделе 3.1. Кратко повторим описание.

Имеется полумарковский случайный процесс $m_t,\ t \geq 0$, обладающий стационарным распределением вероятностей. Известно, что такой процесс полностью характеризуется своим пространством состояний и полумарковским ядром. Считаем, что пространством состояний процесса m_t является множество $\{1, 2, \ldots, M\}$, а ядром является матрица $B(t)$ с компонентами $B_{m,m'}(t)$. Функция $B_{m,m'}(t)$ интерпретируется как вероятность того, что время пребывания процесса в текущем состоянии не превысит величину t и следующий переход произойдет в состояние m' при условии, что текущим состоянием процесса является состояние m, $m, m' = \overline{1, M}$.

Процесс обслуживания является SM-процессом, если времена обслуживания последовательных заявок задаются последовательными временами пребывания полумарковского процесса m_t в своих состояниях. Система $BMAP/SM/1$ впервые была рассмотрена Д. Лукантони и М. Ньютсом на основе аналитического подхода М. Ньютса. Кратко опишем результаты исследования этой системы на основе такого подхода.

4.2.1 Стационарное распределение вероятностей вложенной ЦМ

Используя опыт исследования системы $BMAP/G/1$, исследование системы $BMAP/SM/1$ начнем с изучения вложенного процесса $\{i_n, \nu_n, m_n\}$, $n \geq 1$, где i_n — число запросов в системе в момент t_n+0 (t_n — момент окончания обслуживания n-го запроса), ν_n — состояние управляющего процесса $BMAP$-потока в момент t_n, m_n — состояние управляющего процесса SM-процесса обслуживания в момент $t_n + 0$.

Анализируя поведение этого процесса, несложно понять, что он является трехмерной КТЦМ. Матричные ПФ $Y(z)$ и $V(z)$, характеризующие

одношаговые вероятности переходов этой цепи, имеют следующий вид:

$$Y(z) = \int_0^\infty e^{D(z)t} \otimes dB(t) = \hat{\beta}(-D(z)), \tag{4.29}$$

$$V(z) = \frac{1}{z}(-\tilde{D}_0)^{-1}(\tilde{D}(z) - \tilde{D}_0)\hat{\beta}(-D(z)),$$

где $\tilde{D}_k = D_k \otimes I_M$, $\tilde{D}(z) = D(z) \otimes I_M$. По виду эти матричные ПФ совпадают с аналогичными ПФ для системы $BMAP/G/1$, но скалярная операция умножения под знаком интеграла заменяется на операцию кронекерова произведения матриц.

Теорема 4.4. *Необходимым и достаточным условием существования стационарного распределения ЦМ* $\{i_n, \nu_n, m_n\}$, $n \geq 1$, *является выполнение неравенства*

$$\rho = \lambda b_1 < 1, \tag{4.30}$$

где λ *— средняя интенсивность поступления запросов в* $BMAP$*-потоке,* $\lambda = \boldsymbol{\theta} D'(1)\mathbf{e}$, b_1 *— среднее время обслуживания,* $b_1 = \boldsymbol{\delta} \int\limits_0^\infty t dB(t)\mathbf{e}$, $\boldsymbol{\delta}$ *— инвариантный вектор стохастической матрицы* $B(\infty)$*, то есть вектор, удовлетворяющий системе уравнений*

$$\boldsymbol{\delta} B(\infty) = \boldsymbol{\delta}, \ \boldsymbol{\delta}\mathbf{e} = 1.$$

Доказательство. Нетрудно видеть, что условия эргодичности (3.66), (3.67) для квазитеплицевой цепи $\{i_n, \nu_n, m_n\}$, $n \geq 1$, эквивалентны условиям:

$$\mathbf{x}(\hat{\beta}(-D(z))'|_{z=1}\mathbf{e} < 1, \tag{4.31}$$

где вектор $\mathbf{x}$ удовлетворяет системе линейных алгебраических уравнений:

$$\mathbf{x}(I - \hat{\beta}(-D(1))) = \mathbf{0}, \ \mathbf{x}\mathbf{e} = 1. \tag{4.32}$$

Используя правило смешанного произведения, нетрудно убедиться, что единственное решение системы (4.32) имеет вид:

$$\mathbf{x} = \boldsymbol{\theta} \otimes \boldsymbol{\delta}. \tag{4.33}$$

Подставляя в (4.31) матричную функцию $\hat{\beta}(-D(z))$ в виде:

$$\hat{\beta}(-D(z)) = \int_0^\infty \sum_{l=0}^\infty \frac{(D(z)t)^l}{l!} \otimes dB(t)$$

и вектор $\mathbf{x}$ в виде (4.33), после несложных преобразований получаем неравенство:

$$(\boldsymbol{\theta} D'(1)\mathbf{e})(\boldsymbol{\delta} \int_0^\infty t dB(t)\mathbf{e}) < 1,$$

которое эквивалентно неравенству (4.30). □

Далее будем считать условие (4.30) выполненным. Тогда существуют стационарные вероятности

$$\pi(i,\nu,m) = \lim_{n\to\infty} P\{i_n = i, \nu_n = \nu, m_n = m\},\ i \geq 0,\ \nu = \overline{0,W},\ m = \overline{1,M}.$$

Обозначим:

$$\boldsymbol{\pi}(i,\nu) = (\pi(i,\nu,1),\ldots,\pi(i,\nu,M)),$$
$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0),\ldots,\boldsymbol{\pi}(i,W)),$$
$$\boldsymbol{\Pi}(z) = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i z^i,\ |z| < 1.$$

Из общего уравнения (3.70) для векторной ПФ многомерной КТЦМ с учетом (4.29) вытекает справедливость следующего утверждения.

Теорема 4.5. *Векторная ПФ* $\boldsymbol{\Pi}(z)$ *стационарного распределения вложенной цепи* $\{i_n, \nu_n, m_n\}$, $n \geq 1$, *удовлетворяет матричному функциональному уравнению:*

$$\boldsymbol{\Pi}(z)(\hat{\beta}(-D(z)) - zI) = \boldsymbol{\Pi}(0)\tilde{D}_0^{-1}\tilde{D}(z)\hat{\beta}(-D(z)). \qquad (4.34)$$

Поскольку вид матричных ПФ $Y(z)$ и $V(z)$ известен, можно использовать алгоритмы, описанные в предыдущей главе, для нахождения неизвестного вектора $\boldsymbol{\Pi}(0)$ и ПФ $\boldsymbol{\Pi}(z)$. При этом не возникает никаких дополнительных принципиальных трудностей по сравнению с реализацией этих алгоритмов для системы $BMAP/G/1$. В пояснении нуждается только способ вычисления матриц Y_l, $l > 0$.

Опишем процедуру вычисления матриц Y_l для наиболее важного и интересного с практической точки зрения случая, когда полумарковское ядро имеет вид:

$$B(t) = \mathrm{diag}\{B_1(t),\ldots,B_M(t)\}P = \mathrm{diag}\{B_m(t),\ m = \overline{1,M}\}P, \qquad (4.35)$$

где P — стохастическая матрица, $B_m(t)$, $m = \overline{1,M}$ — некоторые ФР.

Вид (4.35) полумарковского ядра означает следующее. Время пребывания полумарковского процесса m_t в состоянии m зависит только от номера этого состояния. Будущее состояние определяется в момент завершения пребывания в текущем состоянии в соответствии с матрицей вероятностей переходов P.

Используя правило смешанного произведения (см. Приложение А) для кронекерова произведения матриц, получаем следующую формулу для матричного ПЛС $\hat{\beta}(-D(z))$:

$$\hat{\beta}(-D(z)) = \left(\int_0^\infty e^{D(z)t} \otimes \mathrm{diag}\{dB_m(t),\ m = \overline{1, M}\} \right) (I_{\bar{W}} \otimes P).$$

Хорошо известно, что путем домножения матрицы слева или справа на элементарные матрицы можно переставлять ее строки и столбцы. Путем непосредственных вычислений несложно убедиться, что

$$\hat{\beta}(-D(z)) = \left(Q\mathrm{diag}\{ \int_0^\infty e^{D(z)t} dB_m(t),\ m = \overline{1, M}\} Q^T \right) (I_{\bar{W}} \otimes P). \quad (4.36)$$

Здесь Q — некоторое произведение элементарных матриц. В случае $\bar{W} = M$ явный вид этих матриц следующий:

$$Q = \prod_{i=0}^{M-2} \prod_{j=i+2}^{M} S_{Mi+j, M(j-1)+i+1},$$

где $S_{l,k}$ — элементарная матрица, полученная из единичной матрицы I путем перемещения единицы в l-й строке с диагонали в k-й столбец и в k-й строке с диагонали в l-й столбец, $l, k = \overline{1, M}$. В случае $\bar{W} \neq M$ матрицы Q легко строятся алгоритмически.

Обозначим через $Y_l^{(m)}$ матрицы, являющиеся коэффициентами разложений

$$\sum_{l=0}^{\infty} Y_l^{(m)} z^l = \int_0^\infty e^{D(z)t} dB_m(t),\ m = \overline{1, M}.$$

Способ вычисления этих матриц описан нами в главе 1. Тогда из (4.36) очевидным образом следует, что матрица Y_l, являющаяся членом разложения $\sum\limits_{l=0}^{\infty} Y_l z^l = \int\limits_0^\infty e^{D(z)t} \otimes dB(t)$, определяется формулой

$$Y_l = (Q\mathrm{diag}\{Y_l^{(m)},\ m = \overline{1, M}\} Q^T)(I_{W+1} \otimes P),\ l \geq 0.$$

Замечание 4.1. Проблемы, связанной с необходимостью перестановки строк и столбцов, обсужденной выше, можно было бы избежать, если бы изначально вместо рассмотрения ЦМ $\{i_n, \nu_n, m_n\}$, $n \geq 1$, где i_n — число запросов в системе в момент $t_n + 0$ (t_n — момент окончания обслуживания n-го запроса), ν_n — состояние управляющего процесса $BMAP$-потока в момент t_n, m_n — состояние управляющего процесса SM-процесса обслуживания в момент t_n+0, мы бы рассматривали ЦМ $\{i_n, m_n, \nu_n\}$, $n \geq 1$. При этом анализ бы проводился совершенно аналогично проведенному выше путем изменения обозначений. Например, функция $Y(z)$, заданная формулой (4.29), в новых обозначениях имеет вид

$$Y(z) = \hat{\beta}(-D(z)) = \int_0^\infty dB(t) \otimes e^{D(z)t}.$$

4.2.2 Стационарное распределение вероятностей состояний системы в произвольный момент времени

Рассмотрим теперь вопрос о распределении вероятностей состояний процесса $\{i_t, \nu_t, m_t\}$, $t \geq 0$ в произвольный момент времени.

Пусть

$$p(i, \nu, m) = \lim_{t \to \infty} P\{i_t = i, \nu_t = \nu, m_t = m\},$$

$$\mathbf{p}(i, \nu) = (p(i, \nu, 1), \ldots, p(i, \nu, M)), \quad \mathbf{p}(i) = (\mathbf{p}(i, 0), \ldots, \mathbf{p}(i, W)),$$

$$\mathbf{P}(z) = \sum_{i=0}^{\infty} \mathbf{p}(i) z^i, \ |z| < 1.$$

Тогда справедливо следующее утверждение.

Теорема 4.6. *ПФ $\mathbf{P}(z)$ стационарного распределения процесса $\{i_t, \nu_t, m_t\}$ в произвольный момент времени и ПФ $\mathbf{\Pi}(z)$ распределения в моменты окончания обслуживания запросов связаны следующим образом:*

$$\mathbf{P}(z)\tilde{D}(z) = \lambda \mathbf{\Pi}(z)(z(\hat{\beta}(-D(z)))^{-1}\nabla^*(z) - I), \tag{4.37}$$

где $\nabla^*(z) = \int\limits_0^\infty e^{D(z)t} \otimes d\nabla_B(t)$, $\nabla_B(t)$ — диагональная матрица с диагональными элементами $[B(t)\mathbf{e}]_j$, $j = \overline{1, M}$.

Доказательство. Используя подход, основанный на использовании вложенных ПМВ, описанный в предыдущем параграфе, получим следующие формулы, связывающие векторы $\mathbf{p}(i)$, $i \geq 0$, с векторами $\boldsymbol{\pi}_i$, $i \geq 0$:

$$\mathbf{p}(0) = \lambda \boldsymbol{\pi}_0 \int_0^\infty (e^{D_0 t} \otimes I_M) dt = \lambda \boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1},$$

$$\mathbf{p}(i) = \lambda \boldsymbol{\pi}_0 \int_0^\infty \int_0^t (e^{D_0 v} \otimes I_M) \sum_{k=1}^{i} (D_k \otimes I_M) dv \times$$

$$\times [P(i-k, t-v) \otimes (I_M - \nabla_B(t-v))] dt + \lambda \sum_{k=1}^{i} \boldsymbol{\pi}_k \int_0^\infty [P(i-k, t) \otimes (I_M - \nabla_B(t))] dt =$$

$$= \lambda \boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1} \sum_{k=1}^{i} \int_0^\infty [D_k P(i-k, t) \otimes (I_M - \nabla_B(t))] dt +$$

$$+ \lambda \sum_{k=1}^{i} \boldsymbol{\pi}_k \int_0^\infty [P(i-k, t) \otimes (I_M - \nabla_B(t))] dt, \ i > 0. \tag{4.38}$$

Умножая уравнения (4.38) на соответствующие степени z и суммируя, получим следующее выражение для векторной ПФ $\mathbf{P}(z)$:

$$\mathbf{P}(z) = \lambda \left(\boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1} + (\boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1} \tilde{D}(z) + \boldsymbol{\Pi}(z)) \int_0^\infty e^{\tilde{D}(z)t} \otimes (I - \nabla_B(t)) dt \right). \tag{4.39}$$

Вычислим интеграл в формуле (4.39):

$$\int_0^\infty e^{\tilde{D}(z)t} \otimes (I - \nabla_B(t)) dt = (\tilde{D}(z))^{-1} e^{\tilde{D}(z)t} \otimes (I - \nabla_B(t))|_0^\infty +$$

$$+ (\tilde{D}(z))^{-1} \int_0^\infty e^{\tilde{D}(z)t} \otimes d\nabla_B(t) = -(\tilde{D}(z))^{-1} + (\tilde{D}(z))^{-1} \int_0^\infty e^{\tilde{D}(z)t} \otimes d\nabla_B(t) =$$

$$= (\tilde{D}(z))^{-1} (\nabla^*(z) - I).$$

Подставляя полученное выражение в (4.39), после преобразований с использованием уравнения (4.34) для ПФ $\boldsymbol{\Pi}(z)$ получим соотношение (4.37). □

Следствие 4.5. *Скалярная ПФ распределения длины очереди в системе* $BMAP/SM/1$ *имеет вид:*

$$\mathbf{P}(z)\mathbf{e} = \lambda(z-1)\mathbf{\Pi}(z)(\tilde{D}(z))^{-1}\mathbf{e}.$$

Доказательство следует из того, что матрицы $z\hat{\beta}^{-1}(-D(z))\nabla^*(z) - I$ и $(\tilde{D}(z))^{-1}$ перестановочны и имеет место равенство $\nabla^*(z)\mathbf{e} = \hat{\beta}(z)\mathbf{e}$.

Следствие 4.6. *Если полумарковское ядро* $B(t)$ *имеет вид (4.35), то формула (4.37) может быть записана в виде:*

$$\mathbf{P}(z)\tilde{D}(z) = \lambda\mathbf{\Pi}(z)(zI \otimes P^{-1} - I).$$

4.3 СИСТЕМА $BMAP/SM/1/N$

Обе модели, рассмотренные в предыдущих подразделах данного раздела, предполагают наличие в системе бесконечного буфера. Вместе с тем, актуальным является и изучение систем с конечным буфером. Одной из причин этого является быстрое развитие телекоммуникационных сетей. Рост числа пользователей и видов передаваемой информации приводит к увеличению нагрузки на серверы сети, и, хотя емкость буферов составляет сотни-сотни тысяч пакетов, возникают ситуации, когда буфер переполняется. При этом, при коэффициенте загрузки канала (отношение интенсивности поступающего потока к скорости передачи пакетов) в диапазоне $0.9 - 1.1$ вероятность потери пакета весьма чувствительна как к изменению загрузки, так и к изменению размера буфера. Поэтому для успешного решения задачи выбора оптимального объема буфера необходимо уметь точно рассчитывать вероятность отказа при фиксированном виде входного потока, процесса обслуживания и фиксированном размере буфера. Таким образом, задача расчета характеристик системы $BMAP/SM/1/N$ является весьма актуальной.

Значительный вклад в изучение систем с конечным буфером внес П. П. Бочаров, которым интенсивно исследовались системы с конечным буфером и ординарным MAP-потоком. Предположение, что поток является групповым $BMAP$-потоком, ведет к некоторым усложнениям анализа, вызванным необходимостью рассмотрения различных сценариев поведения системы в ситуации, когда в момент поступления группы запросов в системе имеются свободные места, но этих мест не достаточно для принятия

всех запросов группы. В литературе популярны следующие три дисциплины принятия запросов:

- Частичное принятие (partial admission – PA) предполагает, что часть запросов группы, соответствующая числу свободных мест, принимается в систему, а остальные запросы теряются;
- Полное принятие (complete admission – CA) предполагает, что все запросы группы принимаются в систему (например, путем использования расширенной памяти);
- Полный отказ (complete rejection – CR) предполагает, что все запросы группы теряются. Например, такая дисциплина разумна в ситуации, когда все запросы группы принадлежат одному сообщению и потеря хотя бы одного запроса равносильна неуспешной передаче сообщения.

Проведем сначала анализ системы $BMAP/SM/1/N$ с дисциплиной частичного принятия, самой простой в плане сложности анализа из вышеупомянутых дисциплин. Затем рассмотрим оставшиеся две дисциплины. Кратко упомянем гибридную дисциплину, предусматривающую рандомизированный выбор одной из трех перечисленных дисциплин в каждый момент поступления группы запросов. Будем изучать распределение числа запросов в системе. В частности, получим одну из важнейших ее характеристик — вероятность потери произвольного запроса.

4.3.1 Анализ системы с дисциплиной частичного принятия

4.3.1.1 Переходные вероятности вложенной ЦМ Будем рассматривать трехмерный процесс $\{i_n, \nu_n, m_n\}$, $n \geq 1$, где $i_n = i_{t_n}$ — число запросов в системе в момент $t_n + 0$, $i_n = \overline{0, N}$; $\nu_n = \nu_{t_n}$ — состояние управляющего процесса $BMAP$-потока в момент t_n, $\nu_n = \overline{0, W}$; $m_n = m_{t_n}$ — состояние управляющего процесса SM-обслуживания в момент $t_n + 0$, $m_n = \overline{1, M}$.

Пусть $P\{(i, \nu, m) \to (j, \nu', m')\} = P\{i_{n+1} = j, \nu_{n+1} = \nu', m_{n+1} = m' \,|\, i_n = i, \nu_n = \nu, m_n = m\}$, $n \geq 1$ — одношаговые переходные вероятности процесса $\{i_n, \nu_n, m_n\}$, $n \geq 1$. Считаем, что состояния цепи $\{i_n, \nu_n, m_n\}$, $n \geq 1$, упорядочены в лексикографическом порядке и введем обозначение для блочных матриц переходных вероятностей

$$P_{i,j} = \|P\{(i, \nu, m) \to (j, \nu', m')\}\|_{\nu,\nu'=\overline{0,W},\, m,m'=\overline{1,M}} \quad .$$

Используя стандартные рассуждения, несложно убедиться в справедливости следующего утверждения.

Теорема 4.7. *Матрицы переходных вероятностей $P_{i,j}$ определяются следующим образом:*

$$P_{i,j} = Y_{j-i+1}, \quad 0 < i < N, \quad j \geq i-1,$$

$$P_{0,j} = V_j, \quad j = \overline{0, N-1},$$

$$P_{0,N} = V(1) - \sum_{l=0}^{N-1} V_l, \; P_{i,N} = Y(1) - \sum_{l=0}^{N-i} Y_l, \quad i = \overline{1, N},$$

где матрицы Y_l, V_l находятся как коэффициенты разложений:

$$\sum_{l=0}^{\infty} Y_l z^l = Y(z) = \int_0^{\infty} e^{D(z)t} \otimes dB(t),$$

$$\sum_{l=0}^{\infty} V_l z^l = V(z) = \frac{1}{z}(-\tilde{D}_0)^{-1}(\tilde{D}(z) - \tilde{D}_0)Y(z).$$

Доказательство. Вероятности $P_{i,j}$, $j < N$, совпадают с соответствующими вероятностями для системы $BMAP/SM/1$ с бесконечным буфером. Вид вероятностей $P_{i,N}$ очевидным образом следует из условия нормировки и вероятностного смысла матриц $Y(1)$ $(V(1))$, описывающих переходы конечных компонент $\{\nu_n, m_n\}$ за время обслуживания одного запроса (за время с момента окончания обслуживания запроса, в который система оказалась пуста, до момента окончания первого в периоде занятости запроса). □

4.3.1.2 Прямой алгоритм нахождения стационарного распределения вложенной ЦМ Поскольку пространство состояний ЦМ $\{i_n, \nu_n, m_n\}$, $n \geq 1$, конечно, при сделанном выше предположении о неприводимости управляющих процессов потока и обслуживания при любых значениях параметров системы существуют стационарные вероятности

$$\pi(i, \nu, m) = \lim_{n \to \infty} P\{i_n = i, \nu_n = \nu, m_n = m\}, \; i = \overline{0, N}, \; \nu = \overline{0, W}, \; m = \overline{1, M}.$$

В соответствии с принятым выше лексикографическим упорядочением компонента цепи, введем в рассмотрение векторы $\boldsymbol{\pi}_i$, $i = \overline{0, N}$, стационарных вероятностей $\pi(i, \nu, m)$, соответствующих значению i компоненты i_n.

Учитывая вид матриц переходных вероятностей, приведенный выше, легко получить систему уравнений равновесия для вероятностных векторов $\boldsymbol{\pi}_i$:

$$\boldsymbol{\pi}_l = \boldsymbol{\pi}_0 V_l + \sum_{i=1}^{l+1} \boldsymbol{\pi}_i Y_{l+1-i}, \quad l = \overline{0, N-1}, \tag{4.40}$$

$$\boldsymbol{\pi}_N = \boldsymbol{\pi}_0 (V(1) - \sum_{l=0}^{N-1} V_l) + \sum_{i=1}^{N} \boldsymbol{\pi}_i (Y(1) - \sum_{l=0}^{N-i} Y_l). \tag{4.41}$$

Анализируя структуру системы (4.40) – (4.41), нетрудно увидеть, что, полагая в (4.40) $l = 0, 1, .., N-1$, можно последовательно выразить все векторы $\boldsymbol{\pi}_l$, $l = \overline{1, N}$ через вектор $\boldsymbol{\pi}_0$. Для нахождения последнего можно использовать уравненние (4.41) (где все векторы $\boldsymbol{\pi}_l$ уже известны с точностью до $\boldsymbol{\pi}_0$) и условие нормировки.

В результате можно убедиться в справедливости следующего утверждения.

Теорема 4.8. *Стационарные вероятности рассматриваемой цепи Маркова определяются следующим образом:*

$$\boldsymbol{\pi}_l = \boldsymbol{\pi}_0 F_l, \quad l = \overline{0, N},$$

где матрицы F_l определяются рекуррентным образом:

$$F_0 = I, \;\; F_{l+1} = (-V_l F_l - \sum_{i=1}^{l} F_i Y_{l+1-i}) Y_0^{-1}, \quad l = \overline{0, N-1}, \tag{4.42}$$

а вектор $\boldsymbol{\pi}_0$ находится как решение системы линейных алгебраических уравнений

$$\boldsymbol{\pi}_0 \big[-F_N V(1) - \sum_{l=0}^{N-1} V_l \sum_{i=1}^{N} F_i (Y(1) - \sum_{l=0}^{N-i} Y_l) \big] = \mathbf{0}, \tag{4.43}$$

$$\boldsymbol{\pi}_0 \sum_{l=0}^{N} F_l \, \mathbf{e} = 1. \tag{4.44}$$

Замечание 4.2. Одной из важных практических задач, которая может быть решена с помощью полученных результатов, является выбор размера N буфера, обеспечивающего (при фиксированных характеристиках потока и процесса обслуживания) приемлемый уровень вероятности потери произвольного запроса. В процессе решения этой задачи процедура вычисления

стационарного распределения может применяться при нескольких последовательных значениях N. Достоинством процедуры нахождения стационарного распределения, заданной данной теоремой, является тот факт, что при увеличении величины N не требуется повторять процедуру сначала. Следует лишь вычислить необходимое число дополнительных матриц F_l по рекуррентным формулам (4.42) и решить систему линейных алгебраических уравнений

$$(4.43) \; - \; (4.44)$$

с обновленной матрицей.

4.3.1.3 Модифицированный алгоритм нахождения стационарного распределения вложенной цепи Маркова Рекурсия, заданная формулой (4.42), не является численно устойчивой, поскольку в ней производится вычитание матриц, что ведет к исчезновению порядка и накоплению ошибок округления, особенно в случае большой емкости N входного буфера. Используя общий алгоритм для многомерных ЦМ с конечным пространством состояний, описанный в разделе 3.5, можно получить другую рекурсию, заданную следующей теоремой.

Теорема 4.9. *Векторы* $\boldsymbol{\pi}_l, \quad l = \overline{0, N}$, *стационарных вероятностей вложенной цепи Маркова* $\{i_n, \nu_n, m_n\}$, $n \geq 1$, *вычисляются следующим образом:*

$$\boldsymbol{\pi}_l = \boldsymbol{\pi}_0 A_0, \quad l \geq 0, \tag{4.45}$$

где матрицы A_l *удовлетворяют рекуррентным соотношениям*

$$A_0 = I, \; A_l = (\overline{V}_l - \sum_{i=1}^{l-1} A_i \overline{Y}_{l+1-i}^{(l)})(I - \overline{Y}_1^{(l)})^{-1}, \quad l = \overline{1, N-1}, \tag{4.46}$$

$$A_N = (\overline{V}_N - \sum_{i=1}^{N-1} A_i \tilde{Y}_{N+1-i})(I - \tilde{Y}_1)^{-1}, \tag{4.47}$$

где

$$\tilde{Y}_l = Y(1) - \sum_{i=0}^{l-1} Y_i, \quad l = \overline{1, N},$$

а матрицы $\overline{V}_k, \quad \overline{Y}_l^{(k)}$ *определяются следующим образом:*

$$\overline{V}_N = \tilde{V}_N = V(1) - \sum_{i=0}^{N-1} V_i, \tag{4.48}$$

$$\overline{V}_l = V_l + \overline{V}_{l+1} G_{l+1}, \quad l = \overline{0, N-1},$$

$$\overline{Y}_i^{(N)} = \tilde{Y}_i, \quad i = \overline{1, N}, \tag{4.49}$$

$$\overline{Y}_i^{(k)} = Y_i + \overline{Y}_{i+1}^{(k+1)} G_{k+1}, \quad i = \overline{1, k}, \quad k = \overline{1, N-1}, \tag{4.50}$$

где матрицы G_l определяются рекуррентным образом:

$$G_N = (I - \tilde{Y}_1)^{-1} Y_0,$$

$$G_l = Y_0 \sum_{i=1}^{N-l} Y_i G_{l+i-1} \cdot \ldots \cdot G_l + \tilde{Y}_{N-l+1} G_N \cdot \ldots \cdot G_l, \quad l = \overline{1, N-1} \tag{4.51}$$

или формулой

$$G_l = (I - \overline{Y}_1^{(l)})^{-1} Y_0, \quad l = \overline{1, N}. \tag{4.52}$$

Вектор $\boldsymbol{\pi}_0$ определяется как решение системы линейных алгебраических уравнений

$$\boldsymbol{\pi}_0 (I - \overline{V}_0) = \mathbf{0}, \tag{4.53}$$

$$\boldsymbol{\pi}_0 \sum_{l=0}^{N} A_l \mathbf{e} = 1. \tag{4.54}$$

Доказательство. Как отмечалось выше, формулы (4.46), (4.47) получаются путем построения сенсорной цепи для цепи $\{i_n, \nu_n, m_n\}$, $i_n = \overline{0, N}$, $\nu_n = \overline{0, W}$, $m_n = \overline{1, M}$, $n \geq 1$. Заметим, что матрица G_i описывает переходы процесса $\{\nu_n, m_n\}$ за время, пока компонента i_n, стартуя из состояния i, впервые достигнет уровня $i-1$. В отличие от системы с бесконечным буфером, эта матрица не инвариантна относительно i. Другой путь доказательства теоремы состоит в использовании уравнений равновесия (4.40),(4.41). Исключая из этих уравнений $\boldsymbol{\pi}_N$, $\boldsymbol{\pi}_{N-1}$, .., $\boldsymbol{\pi}_1$ последовательно, мы приходим к уравнениям (4.53), (4.54) для вероятностного вектора $\boldsymbol{\pi}_0$, где матрицы, входящие в (4.53), (4.54), задаются соотношениями (4.46) – (4.52). Заметим, что все матрицы, которые приходится обращать в этих соотношениях, являются невырожденными в силу субстохастичности матриц, вычитаемых из матрицы I. □

Замечание 4.3. Рекурсия (4.46), (4.47) включает только неотрицательные матрицы, поэтому она существенно более устойчива численно, чем рекурсия (4.42).

4.3.1.4 Стационарное распределение вероятностей состояний системы в произвольный момент времени Как уже упоминалось ранее, интересующий нас процесс $\{i_t, \nu_t, m_t\}$, $t \geq 1$ не является марковским. Одним из подходов для нахождения его стационарного распределения является подход, использующий распределение вероятностей состояний этого процесса во вложенные моменты времени. В нашем случае в качестве вложенных моментов времени мы рассматриваем моменты $t = t_n, n \geq 1$, окончания обслуживания запросов. Найдя стационарное распределение вложенной цепи $\{i_n, \nu_n, m_n\}$, $n \geq 1$, мы в состоянии теперь найти стационарное распределение вероятностей процесса $\{i_t, \nu_t, m_t\}$, $t \geq 0$.

Обозначим

$$p(i, \nu, m) = \lim_{t \to \infty} P\{i_t = i,\, \nu_t = \nu,\, m_t = m\},$$

$$i = \overline{0, N+1},\, \nu = \overline{0, W},\, m = \overline{1, M}.$$

Упорядочив эти вероятности в лексикографическом порядке, мы получим векторы вероятностей $\mathbf{p}_i$, $i = \overline{0, N+1}$. Предположим, что полумарковское ядро $B(t)$ имеет вид (4.35).

Теорема 4.10. *Стационарное распределение вероятностей* $\mathbf{p}_i$, $i = \overline{0, N}$, *состояний системы в произвольный момент времени определяется следующим образом:*

$$\mathbf{p}_0 = \Lambda \boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1}, \tag{4.55}$$

$$\mathbf{p}_i = \left[\sum_{j=0}^{i-1} \mathbf{p}_j \tilde{D}_{i-j} - \Lambda(\boldsymbol{\pi}_{i-1}(I_{W1} \otimes P)^{-1} - \boldsymbol{\pi}_i) \right] (-\tilde{D}_0)^{-1}, i = \overline{1, N}, \tag{4.56}$$

$$\mathbf{p}_{N+1}\mathbf{e} = 1 - \sum_{i=0}^{N} \mathbf{p}_i \mathbf{e}, \tag{4.57}$$

$$\text{где} \quad \Lambda = \tau^{-1},\, \tau = \boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1} \mathbf{e} b_1. \tag{4.58}$$

Доказательство. Стандартным методом нахождения стационарного распределения в произвольный момент времени по распределению вложенной цепи является применение аппарата процессов марковского восстановления [112]. Применяя этот метод, мы получим следующие соотношения:

$$\mathbf{p}_0 = \Lambda \boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1}, \tag{4.59}$$

$$\mathbf{p}_i = \Lambda \boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1} \sum_{k=1}^{i} \int_0^{\infty} (\tilde{D}_k P(i-k, t)) \otimes (I - \tilde{B}(t)) dt +$$

$$+\Lambda\sum_{k=1}^{i}\boldsymbol{\pi}_k\int_0^\infty P(i-k,t)\otimes((I-\tilde{B}(t))dt, \tag{4.60}$$

$$\mathbf{p}_N=\Lambda\boldsymbol{\pi}_0(-\tilde{D}_0)^{-1}\Big[(\tilde{D}(1)-\tilde{D}_0)\int_0^\infty e^{D(1)t}\otimes(I-\tilde{B}(t))dt-$$

$$-\sum_{k=1}^{N}\int_0^\infty(\tilde{D}_k\sum_{l=0}^{N-k-1}P(l,t))\otimes(I-\tilde{B}(t)dt\Big]+ \tag{4.61}$$

$$+\Lambda\sum_{k=1}^{N}\boldsymbol{\pi}_k\Big[\int_0^\infty e^{D(1)t}\otimes(I-\tilde{B}(t))dt-\sum_{l=0}^{N-k-1}P(l,t))\otimes(I-\tilde{B}(t)dt\Big].$$

Здесь величина Λ задается формулой (4.58) и определяет интенсивность ухода из системы обслуженных запросов. Матрица $P(i,t)$ задает вероятности переходов процесса ν_t за время $(0,t)$, за которое поступило i запросов из $BMAP$-потока. Матрица $\tilde{B}(t)$ является диагональной. Ее диагональные элементы задаются вектором $B(t)\mathbf{e}$.

Уравнение (4.59) совпадает с (4.55). Анализируя оставшееся уравнение системы (4.60) – (4.61), куда входят матрицы $P(i,t)$, которые в явном виде неизвестны и задаются своей производящей функцией

$$\sum_{i=0}^{\infty}P(i,t)z^i=e^{D(z)t},$$

мы не видим простого прямого пути вывода из них формул (4.56), (4.57). Вспоминаем, что в случае системы с бесконечным буфером производящие функции векторов $\boldsymbol{\pi}_i$ и $\mathbf{p}_i$ связаны следующим образом:

$$\sum_{i=0}^{\infty}\mathbf{p}_iz^i=\lambda\sum_{i=0}^{\infty}\boldsymbol{\pi}_iz^i(z\hat{\beta}^{-1}(-D(z))\bigtriangledown^*(z)-I)(\tilde{D}(z))^{-1}. \tag{4.62}$$

Здесь

$$\hat{\beta}^{-1}(-D(z))=\int_0^\infty e^{D(z)t}\otimes dB(t),\quad \bigtriangledown^*(z)=\int_0^\infty e^{D(z)t}\otimes d\tilde{B}(t).$$

При виде (4.35) полумарковского ядра $B(t)$ матричные производящие функции $\hat{\beta}^{-1}(-D(z))$ и $\bigtriangledown^*(z)$ связаны следующим образом:

$$\hat{\beta}^{-1}(-D(z))=\bigtriangledown^*(z)(I_{W1}\otimes P),$$

поэтому формула (4.62) имеет вид:

$$\sum_{i=0}^{\infty} \mathbf{p}_i z^i = \lambda \sum_{i=0}^{\infty} \boldsymbol{\pi}_i z^i (z(I \otimes P)^{-1} - I) \qquad (4.63).$$

Рассмотрим теперь систему (4.59), (4.60) при $0 \leq i \leq \infty$. Переходя к производящим функциям, мы получим соотношение (4.63) с точностью до множителя λ, который в данном случае заменяется на Λ. Разлагая это соотношение в ряд, мы получаем (4.49) для $i = \overline{1, N}$. Величина $\mathbf{p}_{N+1}\mathbf{e}$ находится из условия нормировки. □

Одной из важнейших вероятностных характеристик системы с конечным буфером является P_{loss} потери запроса из-за заполненности буфера.

Теорема 4.11. *Вероятность P_{loss} потери произвольного запроса в данной системе задается следующим образом:*

$$P_{loss} = 1 - \frac{1}{\lambda} \sum_{i=0}^{N} \mathbf{p}_i \sum_{k=0}^{N+1-i} (k - (N+1) + i) \tilde{D}_k \mathbf{e} = 1 - \frac{1}{\lambda \tau}. \qquad (4.64)$$

Замечание 4.4. Наличие двух разных формул для вычисления вероятности P_{loss} потери произвольного запроса является полезным на стадии компьютерной реализации полученных алгоритмов для вычисления стационарных вероятностей и вероятности P_{loss}.

Алгоритмы расчета распределения вероятностей состояний системы в произвольные и вложенные моменты времени реализованы в виде программы на языке C под управлением системной оболочки пакета прикладных программ "СИРИУС", разработанного в Белгосуниверситете, см. [115]. Результаты численных экспериментов, проведенных с использованием этих программных средств, целью которых было выяснение областей устойчивого счета прямого и дополнительного алгоритмов в зависимости от размера буфера N, получение зависимости вероятности P_{loss} потери пакета от величины буфера N и важность учета корреляции во входном потоке, можно найти в [78]. В частности, выяснено, что использование стационарного пуассоновского потока в качестве модели реального потока данных, имеющего даже относительно небольшую (порядка 0.2) корреляцию длин соседних интервалов может дать значение вероятности потери произвольного запроса в 100000 раз меньшее, чем реальное значение этой вероятности.

4.3.2 Анализ системы с дисциплинами полного принятия и полного отказа

4.3.2.1 Переходные вероятности вложенной цепи Маркова

Отметим, что множество значений процесса i_t – число запросов в системе в момент времени t, $t > 0$, есть конечное множество $\{0, 1, \ldots, N+1\}$ для дисциплин PA и CR. В случае дисциплины CA это множество значений бесконечное. Процесс i_t немарковский и для его изучения мы сначала рассматриваем вложенную цепь Маркова $\xi_n = \{i_n, \nu_n, m_n\}$, $n \geq 1$. Рассмотрение начинаем с вывода матриц $P_{i,l}$, состоящих из переходных вероятностей

$$P\{i_{n+1} = l, \nu_{n+1} = \nu', m_{n+1} = m' | i_n = i, \nu_n = \nu\, m_n = m\},$$

$$\nu, \nu' = \overline{0, W}, m, m' = \overline{1, M}.$$

Для вычисления матриц $P_{i,l}$, необходимо сначала вычислить матрицы $P^{(j)}(n,t)$, (ν, ν')-й элемент которых есть условная вероятность того, что n запросов будет принято в систему за интервал времени $(0,t]$ и состояние управляющего процесса $BMAP$ потока ν_t в момент t будет ν' при условии, что состояние этого процесса в момент 0 было ν и как максимум j запросов могло быть принято в систему (из-за ограниченности буфера) интервале времени $(0,t]$, $n = \overline{0, j}$.

В случае дисциплины PA матрицы $P^{(j)}(n,t)$ легко вычислялись по очевидным формулам:

$$P^{(j)}(n,t) = \begin{cases} P(n,t), & n < j, \\ \sum\limits_{l=j}^{\infty} P(l,t), & n = j, \end{cases} \tag{4.65}$$

где матрицы $P(n,t)$, задающие вероятности поступления n запросов в интервале времени $(0,t]$, заданы как коэффициенты в матричном разложении

$$e^{D(z)t} = \sum_{n=0}^{\infty} P(n,t) z^n. \tag{4.66}$$

Проблема вычисления этих матриц была обсуждена в разделе 3.1.6.

В случае дисциплины CA и CR, матрицы $P^{(j)}(n,t)$ не могут быть легко вычислены через матрицы $P(n,t)$, по аналогии с (4.65), (4.66) что и является основной причиной относительно слабой изученности этих дисциплин.

Напомним, что численная процедура вычисления матриц $P(n,t)$, приведенная в разделе 3.1.6, состоит в следующем. Пусть ψ определено как $\psi = \max\limits_{\nu=\overline{0,W}}(-D_0)_{\nu,\nu}$. Тогда

$$P(n,t) = e^{-\psi t}\sum_{i=0}^{\infty}\frac{(\psi t)^i}{i!}U_n^{(i)}, \tag{4.67}$$

где матрицы $U_n^{(i)}$ вычисляются рекуррентно как:

$$U_n^{(0)} = \begin{cases} I, & n = 0, \\ 0, & n > 0, \end{cases}$$

$$U_n^{(i+1)} = U_n^{(i)}(I + \psi^{-1}D_0) + \psi^{-1}\sum_{l=0}^{n-1}U_l^{(i)}D_{n-l},\ i \geq 0,\ n \geq 0.$$

Эти формулы получены с использованием системы матричных дифференциальных уравнений (3.1), которая, в свою очередь, была получена на основе системы матричных разностных уравнений для матриц $P(n,t),\ n \geq 0$.

Действуя аналогично для дисциплин CA и CR, можно убедиться в справедливости следующего утверждения.

Лемма 4.1. *Матрицы $P^{(j)}(n,t)$ для дисциплин CA и CR вычисляются по формулам:*

$$P^{(j)}(n,t) = e^{-\psi t}\sum_{i=0}^{\infty}\frac{(\psi t)^i}{i!}U_n^{(i)}(j),$$

где матрицы $U_n^{(i)}(j)$ вычисляются по формулам:

$$U_n^{(0)}(j) = \begin{cases} I, & n = 0, \\ 0, & n > 0, \end{cases}$$

для обеих дисциплин и

$$U_n^{(i+1)}(j) = U_n^{(i)}(j)\left(I + \psi^{-1}(D_0 + \hat{D}_{j+1-n})\right) + +\psi^{-1}\sum_{l=0}^{n-1}U_l^{(i)}(j)D_{n-l},$$
$$i \geq 0,\ j = \overline{0, N+1},\ n = \overline{0, j},$$
$$\hat{D}_l = \sum_{m=l}^{\infty}D_m$$

для дисциплины CR и

$$U_n^{(i+1)}(j) = U_n^{(i)}(j)(I + \psi^{-1}D_0) + +\psi^{-1}\sum_{l=0}^{n-1} U_l^{(i)}(j)D_{n-l},$$
$$i \geq 0,\ j = \overline{0, N+1},\ n = \overline{0, j-1},$$
$$U_n^{(i+1)}(j) = U_n^{(i)}(j)\left(I + \psi^{-1}D(1)\right) + +\psi^{-1}\sum_{l=0}^{j-1} U_l^{(i)}(j)D_{n-l},$$
$$i \geq 0,\ j = \overline{0, N+1},\ n \geq j$$

для дисциплины CA.

Лемма 4.2. *Матрицы переходных вероятностей $P_{i,l}$ вычисляются по формулам:*

$$P_{0,l} = -(D_0 + \hat{D}_{N+2})^{-1}\sum_{k=1}^{l+1} D_k \int_0^\infty P^{(N+1-k)}(l+1-k,t)dB(t),\ l = \overline{0, N},$$

$$P_{i,l} = \int_0^\infty P^{(N+1-i)}(l+1-i,t)dB(t), i = \overline{1, N},\ l = \overline{i-1, N},$$
$$P_{i,l} = 0,\ i = \overline{i, N},\ l < i-1,$$

для дисциплины CR и

$$P_{0,l} = (-D_0)^{-1}\sum_{k=1}^{l+1} D_k \int_0^\infty P^{(N+1-k)}(l+1-k,t)dB(t),\ l = \overline{0, N},$$
$$P_{0,l} = (-D_0)^{-1}\left(D_{l+1}G + \sum_{k=1}^{N} D_k \int_0^\infty P^{(N+1-k)}(l+1-k,t)dB(t)\right),\ l > N,$$

$$P_{i,l} = \int_0^\infty P^{(N+1-i)}(l+1-i,t)dB(t),\ i = \overline{1, N},\ l \geq i-1,$$
$$P_{i,l} = O,\ i > N,\ l \neq i-1\ u\ \ i > 0,\ l < i-1,$$
$$P_{i,l} = \int_0^\infty e^{D(1)t} \otimes dB(t),\ i > N,\ l = i-1,$$

для дисциплины CA.

Доказательство данной леммы достаточно очевидно, если принять в рассмотрение вероятностный смысл фигурирующих в ней матриц. В частности, элементы матрицы

$$\int_0^\infty P^{(N+1-k)}(l+1-k,t)dB(t)$$

задают вероятность принятия в систему $l+1-k$ запросов и соответствующих переходов процесса $\nu_t, t \geq 0$, за время обслуживания время одного запроса при условии, что в системе в момент начала обслуживания запроса было $N+1-k$ свободных мест.

В случае дисциплины CR элементы матрицы

$$-(D_0+\hat{D}_{N+2})^{-1}D_k = \int_0^\infty e^{(D_0+\hat{D}_{N+2})t}D_k dt$$

задают вероятность того, что пребывание системы в пустом состоянии завершится приходом группы из k запросов и соответствующих переходов процесса $\nu_t, t \geq 0$, за время пребывания системы в пустом состоянии, $k = \overline{1, N+1}$. Отметим, что в выводе этого выражения использовалась устойчивость матрицы $D_0 + \hat{D}_{N+2}$.

Используя полученные выражения для переходных вероятностей вложенной цепи Маркова и алгоритмы для нахождения стационарных вероятностей цепи Маркова, изложенные в разделах 3.4 и 3.5, можно подсчитать векторы стационарных вероятностей вложенной цепи Маркова.

4.3.2.2 Расчет векторов стационарных вероятностей вложенной цепи Маркова.

Теорема 4.12. *Для дисциплины CR векторы $\boldsymbol{\pi}_i$, $i = \overline{0, N}$, вычисляются по формулам*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \Phi_i,\ i = \overline{0, N},$$

где матрицы Φ_i вычисляются по рекуррентным формулам

$$\Phi_0 = I,\ \Phi_l = \sum_{i=0}^{l-1} \Phi_i \bar{P}_{i,l}(I - \bar{P}_{l,l})^{-1},\ l = \overline{1, N},$$

а вектор $\boldsymbol{\pi}_0$ является единственным решением системы

$$\boldsymbol{\pi}_0(I - \bar{P}_{0,0}) = 0,\ \boldsymbol{\pi}_0 \sum_{l=0}^{N} \Phi_l \mathbf{e} = 1.$$

Здесь матрицы $\bar{P}_{i,l}$ вычисляются по рекуррентным формулам

$$\bar{P}_{i,l} = P_{i,l} + \bar{P}_{i,l+1}G_l,\ i = \overline{0, N},\ l = \overline{i, N}, \bar{P}_{i,N+1} = O,$$

а матрицы G_i вычисляются по рекуррентным формулам

$$G_{N-1} = (I - P_{N,N})^{-1} P_{N,N-1},$$
$$G_i = (I - \sum_{l=i+1}^{N} P_{i+1,l} G_{l-1} G_{l-2} \cdot \ldots \cdot G_{i+1})^{-1} P_{i+1,i},\ i = \overline{0, N-2}.$$

Теорема 4.13. *Для дисциплины CA стационарное распределение вероятностей состояний системы существует тогда и только тогда, когда выполняется условие $\lambda b_1 < 1$. При выполнении этого условия векторы стационарных вероятностей $\boldsymbol{\pi}_i$, $i \geq 0$, вычисляются по формулам*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \Phi_i,\ i \geq 0,$$

где матрицы Φ_i и вектор $\boldsymbol{\pi}_0$ определяются соответствующими формулами в предыдущей теореме, если в них положить $N = \infty$, а матрицы G_i для $i \geq N$ равны матрице G, определенной в условии теоремы.

4.3.2.3 Распределение вероятностей состояний системы в произвольный момент времени и вероятность потери запроса.

Подсчитав стационарное распределение вложенной цепи Маркова, можно найти стационарное распределение вероятностей состояний системы и в произвольный момент времени. Обозначим

$$p(i, \nu, m) = \lim_{t \to \infty} P\{i_t = i, \nu_t = \nu, m_t = m\},\ \nu = \overline{0, W}, m = \overline{1, M}.$$

Пусть $\mathbf{p}_i$, $i \geq 0$, есть векторы этих вероятностей, перенумерованных в лексикографическом порядке.

Теорема 4.14. *В случае дисциплины CR векторы $\mathbf{p}_i$, $i = \overline{0, N+1}$, вычисляются следующим образом:*

$$\mathbf{p}_0 = \tau^{-1} \boldsymbol{\pi}_0 (-1) \tilde{D}^{-1},$$

где

$$\tilde{D} = (D_0 + \hat{D}_{N+2}) \otimes I_M,$$

$$\mathbf{p}_i = \tau^{-1} \Bigg(\boldsymbol{\pi}_0 (-1) \tilde{D}^{-1} \sum_{k=1}^{i} \int_0^{\infty} \tilde{D}_k P^{(N+1-k)}(i-k, t) \otimes (I - B(t)) dt +$$

$$+ \sum_{k=1}^{\min\{i,N\}} \boldsymbol{\pi}_k \int_0^{\infty} P^{(N+1-k)}(i-k, t) \otimes (I - B(t)) dt \Bigg), i = \overline{1, N+1},$$

среднее время τ между моментами окончания обслуживания задается формулой

$$\tau = b_1 + \boldsymbol{\pi}_0(-1)\tilde{D}^{-1}\mathbf{e}.$$

Теорема 4.15. *В случае дисциплины CA векторы $\mathbf{p}_i$, $i \geq 0$, вычисляются следующим образом:* $\mathbf{p}_0 = \tau^{-1}\boldsymbol{\pi}_0(-\tilde{D}_0)^{-1}$,

$$\mathbf{p}_i = \tau^{-1}\big(\boldsymbol{\pi}_0(-\tilde{D}_0)^{-1} + \sum_{k=1}^{i}\int_0^\infty \tilde{D}_k P^{(N+1-k)}(i-k,t)\otimes(I-B(t))dt+$$

$$+\sum_{k=1}^{i}\boldsymbol{\pi}_k\int_0^\infty P^{(N+1-k)}(i-k,t)\otimes(I-B(t))dt\big),\ i=\overline{1,N+1},$$

$$\mathbf{p}_{N+l} = \tau^{-1}\big(\boldsymbol{\pi}_0(-\tilde{D}_0)^{-1}\sum_{k=1}^{N+1}\int_0^\infty \tilde{D}_k P^{(N+1-k)}(N+l-k,t)\otimes(I-B(t))dt+$$

$$+\sum_{k=1}^{N}\boldsymbol{\pi}_k\int_0^\infty P^{(N+1-k)}(N+l-k,t)(\otimes(I-B(t))dt+$$

$$+\big(\boldsymbol{\pi}_0(-\tilde{D}_0)^{-1}\tilde{D}_{N+l}+\boldsymbol{\pi}_{N+l}\big)\int_0^\infty e^{D(1)t}\otimes(I-B(t))dt\big), l>1,$$

среднее время τ между моментами окончания обслуживания задается формулой

$$\tau = b_1 + \boldsymbol{\pi}_0(-\tilde{D}_0)^{-1}\mathbf{e}.$$

Теорема 4.16. *Вероятность P_{loss} потери произвольного запроса вычисляется по формуле*

$$P_{loss} = 1 - \lambda^{-1}\sum_{i=0}^{N}\sum_{k=1}^{N+1-i} k\mathbf{p}_i\tilde{D}_k\mathbf{e} = 1-(\tau\lambda)^{-1}$$

при дисциплине CR и по формуле

$$P_{loss} = 1 - \lambda^{-1}\sum_{i=0}^{N}\sum_{k=1}^{\infty} k\mathbf{p}_i\tilde{D}_k\mathbf{e} = 1-(\tau\lambda)^{-1}$$

при дисциплине CA.

Численные результаты для этих дисциплин, включая их сравнение между собой и с дисциплиной PA, приведены в [116]. В работе [117] результаты распространены на систему, в которой стратегия доступа группы запросов в каждый момент поступления группы запросов при нехватке мест в буфере выбирается из множества дисциплин PA, CA и CR рандомизированно.

4.4 СИСТЕМА $BMAP/PH/N$

Все системы, рассмотренные выше в данной главе, являлись однолинейными, имеющими единственный обслуживающий прибор. Многие реальные системы имеют несколько обслуживающих приборов. В данном разделе рассмотрим многолинейную систему с бесконечным буфером и $BMAP$-потоком. Для однолинейных систем исследование удается провести для случая, когда времена обслуживания последовательных запросов являются независимыми одинаково распределенными случайными величинами (см. раздел 4.1), и даже для случая, когда времена обслуживания последовательных запросов определяются полумарковским процессом (см. раздел 4.2). Это сделано выше посредством построения многомерной цепи Маркова, вложенной по моментам окончания обслуживания запросов прибором. В случае, когда число приборов больше, чем один, построить такую цепь не удается. Поэтому при рассмотрении многолинейных систем наиболее общими моделями, анализ которых удается провести аналитически, являются модели, в которых времена обслуживания последовательных запросов являются независимыми одинаково распределенными случайными величинами, имеющими распределение фазового типа (распределение PH), см. подраздел 3.1.5.

Будем считать, что все приборы системы являются идентичными и распределение времени обслуживания последовательного запроса управляется ЦМ с непрерывным временем η_t, $t \geq 0$. Эта ЦМ имеет пространство состояний $\{1, \ldots, M, M+1\}$, причем состояние $M+1$ является единственным поглощающим состоянием. Начальное состояние этой ЦМ в момент начала обслуживания выбирается случайным образом в соответствии с вероятностным распределением, заданным компонентами вектора $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$, где $\beta_m \geq 0$, $m = \overline{1, M}$, $\boldsymbol{\beta}\mathbf{e} = 1$. Переходы между состояниями внутри множества $\{1, \ldots, M\}$ происходят с интенсивностями, заданными элементами субгенератора $S = S_{m,m'}$, $m, m' = \overline{1, M}$. Пере-

ходы из состояний внутри множества $\{1, \ldots, M\}$ в поглощающее состояние $M+1$ происходят с интенсивностями, являющимися компонентами вектора-столбца $\mathbf{S}_0 - S\mathbf{e}$. Переход ЦМ η_t в поглощающее состояние $M+1$ влечет окончание обслуживания запроса. Предполагаем, что матрица $S + \mathbf{S}_0\boldsymbol{\beta}$ является неприводимой.

Опишем поведение числа запросов в рассматриваемой системе $BMAP/PH/N$ многомерной марковской цепью с непрерывным временем. Кроме собственно процесса i_t – число запросов в рассматриваемой системе в момент t, а также управляющего процесса ν_t поступления запросов в $BMAP$-потоке, эта цепь должна полностью описывать текущее состояние процессов обслуживания в приборах этой системы. Отметим, что существует несколько способов задания текущего состояние процессов обслуживания в приборах.

Наиболее простой способ задания – это задать состояние ЦМ η_t, управляющей процессом обслуживания в каждом приборе. Если прибор свободен, состояние этого процесса можно положить равным 0. Этот способ является, видимо, единственным возможным, если приборы системы являются неидентичными. Но пространство состояний такого процесса состоит из $(M+1)^N$ элементов и может быть очень большим, что может повлечь существенные проблемы на этапе численной реализации полученных аналитических результатов.

Другой способ задания управляющих процессов обслуживания, существенно эксплуатирующий идентичность приборов и имеющий меньшую размерность, состоит в следующем. Задается состояние ЦМ η_t, управляющей процессом обслуживания в каждом *занятом* в данный момент времени приборе. При этом необходимо договориться о системе динамической нумерации занятых приборов. Хорошей нумерацией является, например, такая. В каждый момент времени минимальный номер имеет прибор, обслуживание на котором длится к данному моменту наиболее долго,а максимальный номер имеет прибор, обслуживание на котором длится к данному моменту наименее долго. Если какой-то прибор закончил обслуживание и очереди в системе нет, то прибор теряет свой номер. Номера всех приборов, которые больше номера прибора, завершившего обслуживание, уменьшаются на 1. Если в системе есть очередь, то прибор, закончивший обслуживание, не меняет свой номер и снова начинает обслуживание. При начале обслуживания запроса, пришедшего в систему, когда в ней были свободные приборы, один из этих приборов получает номер, на единицу

больший, чем максимальный номер приборов, уже осуществляющих обслуживание, и начинает обслуживание запроса.

Таким образом, при данном способе задания управляющих процессов обслуживания (только для занятых приборов) при i, $i < N$, занятых приборах пространство состояний процессов, управляющих обслуживанием, состоит из M^i элементов. При i, $i \geq N$, занятых приборах пространство состояний процессов, управляющих обслуживанием, состоит из M^N элементов. Таким образом, данный способ задания управляющих процессов обслуживания более экономный, чем первый способ. Существует еще более экономный способ, но он будет описан и использован несколько позже.

Несложно видеть, что поведение рассматриваемой системы $BMAP/PH/N$ описывается многомерной цепью Маркова

$$\zeta_t = \{i_t, \nu_t, \eta_t^{(1)}, \ldots, \eta_t^{(\min\{i_t, N\})}\},\ t \geq 0.$$

Будем называть уровнем i множество состояний процесса ζ_t, у которых значение первой компоненты i_t равно i, а остальные компоненты перенумерованы в лексикографическом порядке. Будем обозначать $Q_{i,j}$ матрицу, состоящую из интенсивностей переходов ЦМ ζ_t с уровня i на уровень j, $j \geq \min\{i-1, 0\}$. Поскольку запросы в системе обслуживаются по одному, очевидно, что $Q_{i,j} = O$, если $j < i-1$, т.е. блочная матрица Q, составленная из блоков $Q_{i,j}$, является верхне-хессенберговой.

Лемма 4.3. *Инфинитезимальный генератор Q ЦМ ζ_t имеет следующую структуру:*

$$Q = (Q_{i,j})_{i,j\geq 0} = \tag{4.68}$$

$$\begin{pmatrix}
Q_{0,0} & Q_{0,1} & Q_{0,2} & \ldots & Q_{0,N} & Q_{0,N+1} & Q_{0,N+2} & Q_{0,N+3} & \ldots \\
Q_{1,0} & Q_{1,1} & Q_{1,2} & \ldots & Q_{1,N} & Q_{1,N+1} & Q_{1,N+2} & Q_{1,N+3} & \ldots \\
O & Q_{2,1} & Q_{2,2} & \ldots & Q_{2,N} & Q_{2,N+1} & Q_{2,N+2} & Q_{2,N+3} & \ldots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots \\
O & O & O & \ldots & Q_{N-1,N} & Q_{N-1,N+1} & Q_{N-1,N+2} & Q_{N-1,N+3} & \ldots \\
O & O & O & \ldots & R^0 & R_1^+ & R_2^+ & R_3^+ & \ldots \\
O & O & O & \ldots & R^- & R^0 & R_1^+ & R_2^+ & \ldots \\
O & O & O & \ldots & O & R^- & R^0 & R_1^+ & \ldots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix},$$

где

$$Q_{i,i} = D_0 \oplus S^{\oplus i},\ i = \overline{0, N-1}, \tag{4.69}$$

$$Q_{i,i-1} = I_{\bar{W}} \otimes \mathbf{S}_0^{\oplus i},\ i = \overline{1, N}, \tag{4.70}$$

$$Q_{i,i+k} = \begin{cases} D_k \otimes I_{M^i} \otimes \boldsymbol{\beta}^{\otimes i},\ i = \overline{0, N-1},\ i+k \le N, \\ D_k \otimes I_{M^i} \otimes \boldsymbol{\beta}^{\otimes(i+k-N)},\ i = \overline{0, N-1},\ i+k > N, \end{cases} \tag{4.71}$$

$$R^0 = D_0 \oplus S^{\oplus N}, \tag{4.72}$$

$$R^- = I_{\bar{W}} \otimes (\mathbf{S}_0 \boldsymbol{\beta})^{\oplus N}, \tag{4.73}$$

$$R_k^+ = D_k \otimes I_{M^N}, \tag{4.74}$$

где

$$\boldsymbol{\beta}^{\otimes k} \stackrel{def}{=} \underbrace{\boldsymbol{\beta} \otimes \ldots \otimes \boldsymbol{\beta}}_{k}, k \ge 1, \boldsymbol{\beta}^{\otimes 0} \stackrel{def}{=} 1, \tag{4.75}$$

$$S^{\oplus i} \stackrel{def}{=} \underbrace{S \oplus \ldots \oplus S}_{i}, i \ge 1, S^{\oplus 0} \stackrel{def}{=} 0, \tag{4.76}$$

$$S_0^{\oplus i} \stackrel{def}{=} \sum_{m=0}^{i-1} I_{M^m} \otimes S_0 \otimes I_{M^{i-m-1}}, i \ge 1. \tag{4.77}$$

Доказательство. Доказательство данной леммы легко проводится по аналогии с объяснением вида генератора трехмерной ЦМ, описывающей поведение системы $MAP/PH/1$, в разделе 3.2.2. Система $MAP/PH/1$ является частным случаем модели $BMAP/PH/N$, рассматриваемой нами в данном в разделе, в котором число приборов $N = 1$, а входной поток является ординарным MAP, а не групповым $BMAP$ потоком. В разделе 3.2.2 мы пояснили важность и эффективность использования при анализе многомерных ЦМ операций кронекерова произведения и суммы матриц. В силу групповости входного потока и наличия $N \ge 1$ приборов роль этих операций еще более возрастает.

Кратко прокомментируем вид (4.69)-(4.74) блоков генератора (4.68). Как уже отмечалось в разделе 3.2.2, все элементы генератора имеют смысл интенсивности переходов между соответствующими состояниями ЦМ за исключением диагональных элементов, которые отрицательны и по модулю равны интенсивности выхода ЦМ из соответствующего состояния. Для избежания излишних повторений текста здесь и в нескольких разделах ниже мы больше не будем отдельно упоминать особую роль диагональных элементов и будем говорить обо всех элементах генератора как об интенсивностях.

Переход процесса ζ_t с уровня i на этот же уровень за интервал времени, имеющий бесконечно малую длину, возможен при переходе управляющего

процесса ν_t потока без генерации запросов (соответствующие интенсивности переходов заданы элементами матрицы D_0), либо при переходе управляющего процесса η_t обслуживания в одном из занятых приборов без попадания в поглощающее состояние. Введенное обозначение (4.76) как раз задает интенсивности таких переходов управляющего процесса в одном из i приборах. Отсюда для $i = \overline{0, N-1}$ получаем:

$$Q_{i,i} = D_0 \otimes I_{M^i} + I_{\bar{W}} S^{\oplus i} = D_0 \oplus S^{\oplus i},$$

формула (4.69) доказана. Матрица R^0 имеет смысл матрицы $Q_{i,i}$ при $i \geq N$. Для $i \geq N$ только N запросов находится на обслуживании, поэтому

$$Q_{i,i} = D_0 \oplus S^{\oplus N} = R^0, \; i \geq N.$$

Формула (4.72) доказана.

Переход процесса ζ_t с уровня i на уровень $i-1$ за интервал времени, имеющий бесконечно малую длину, возможен при переходе управляющего процесса η_t обслуживания в одном из занятых приборов в поглощающее состояние. Введенное обозначение (4.77) как раз задает интенсивности переходов управляющего процесса в одном из i приборах в поглощающее состояние. Отсюда для $i = \overline{1, N}$ получаем:

$$Q_{i,i-1} = I_{\bar{W}} \otimes \mathbf{S}_0^{\oplus i}, \; i = \overline{1, N},$$

формула (4.70) доказана. Матрица R^- имеет смысл матрицы $Q_{i,i-1}$ при $i > N$. Для $i > N$ только N запросов находится на обслуживании, кроме того, поскольку в системе есть очередь, после окончания обслуживания одним из приборов на этом приборе немедленно устанавливается начальная фаза для процесса обслуживания следующего запроса с вероятностями, заданными компонентами вектора $\boldsymbol{\beta}$. поэтому

$$Q_{i,i-1} = R^- = I_{\bar{W}} \otimes (\mathbf{S}_0 \boldsymbol{\beta})^{\oplus N}, \;\; i \geq N.$$

Формула (4.73) доказана.

Переход процесса ζ_t с уровня i на уровень $i+k$, $k \geq 1$, за интервал времени, имеющий бесконечно малую длину, возможен при переходе управляющего процесса ν_t потока с генерацией группы из k запросов (соответствующие интенсивности переходов заданы элементами матрицы D_k). При этом, если $i \geq N$, то управляющие процессы обслуживания в N занятых приборах не осуществляют никаких переходов. Поэтому для $i \geq N$

$$Q_{i,i+k} = R_k^+ = D_k \otimes I_{M^N},$$

т.е. получаем формулу (4.74). Если $i < N$ и $i + k \leq N$, то приход группы размера k влечет необходимость установления начальной фазы для процесса обслуживания следующего запроса с вероятностями, заданными компонентами вектора $\boldsymbol{\beta}$, в k приборах. Вектор $\boldsymbol{\beta}^{\otimes k}$, заданный формулой (4.75), как раз задает такое установления начальной фазы для процесса обслуживания в k приборах. Если же $i < N$ и $i+k > N$, то приход группы размера k влечет необходимость установления начальной фазы для процесса обслуживания следующего запроса только в $i + k - N$ приборах, оставшиеся запросы этой группы не поступают на обслуживание, а становятся в очередь. Формула (4.71) доказана. □

Несложно видеть, что ЦМ ζ_t принадлежит классу АКТЦМ, результаты для которого приведены в разделе 3.7. Причем, начиная с уровня N, вид матриц $Q_{i,i+l}$ не зависит от i, а зависит только от l, $l \geq -1$. Поэтому снимаются проблемы в алгоритме расчета векторов стационарных вероятностей, связанные с заданием конечного условия для обратной рекурсии для матриц G_i, $i \geq 0$.

В частности, несложно убедиться, что в рассматриваемой системе существует стационарный режим, если выполняется условие

$$\lambda < Nb_1,$$

где b_1 есть среднее время обслуживания запроса.

Если же входной поток является ординарным MAP потоком, то генератор (4.68) становится блочно трехдиагональным и расчет векторов стационарных вероятностей процесса ζ_t можно осуществить по аналогии с пунктом 3.2.1. Анализ стационарного распределения времени ожидания произвольного запроса практически идентичен анализу этого распределения для системы $MAP/PH/1$, имея ввиду тот факт, что при всех занятых приборах совокупность N приборов имеет производительность, совпадающую с производительностью одного прибора, у которого субгенератор управляющего процесса PH обслуживания есть $S^{\oplus N}$.

4.5 СИСТЕМА $BMAP/PH/N/0$

В некоторых реальных системах мест для ожидания запросов, поступивших, когда все приборы заняты, нет, и поступивший запрос, заставший все приборы занятыми, теряется. Поскольку входной поток предполагается групповым, могут возникнуть ситуации, когда свободные приборы в

системе есть, но число запросов в поступившей группе превышает число свободных приборов. Так же, как и в разделе 4.3, будем рассматривать различные дисциплины принятия запросов: частичное принятие (PA), полное принятие (CA) и полный отказ (CR).

Важность исследования системы $BMAP/PH/N/0$ объясняется, в частности, тем, что она является естественным существенным обобщением системы $M/M/N/0$, с исследования которой А.К. Эрлангом в начале 20-го века ведет свою историю теория массового обслуживания. Результаты исследования этой системы А.К. Эрлангом были применены к расчету телефонных сетей и расчетные значения, например, вероятности отказа в соединении хорошо совпадали с результатами измерений на реальных сетях. Это было несколько неожиданным, поскольку распределение продолжительности телефонного разговора не является экспоненциальным, как это предполагает модель Эрланга. Это хорошее совпадение было позже объяснено фактом наличия свойства инвариантности стационарного распределения числа запросов в системе $M/M/G/0$ с произвольным распределением времени обслуживания запросов относительно вида этого распределения. Т.е. стационарное распределение числа запросов в системе $M/M/G/0$ такое же, как в системе $M/M/N/0$ при одинаковых значениях среднего времени обслуживания. Этот факт был строго доказан Б.А. Севастьяновым.

Численные эксперименты, осуществленные на базе алгоритмических результатов, приведенных в данном разделе, свидетельствуют, что свойство инвариантности стационарного распределения числа запросов в системе с $BMAP$-потоком несправедливо. Поэтому необходим анализ системы $BMAP/PH/N/0$ с распределением времени обслуживания, близким к выборочному распределению, полученному в результате наблюдения за работой системы, а не существенно более простой системы $BMAP/M/N/0$ с экспоненциальным распределением времени обслуживания.

4.5.1 Стационарное распределение вероятностей состояний системы при дисциплине PA

Напомним, что эта дисциплина предполагает, что если в систему поступает группа из k запросов, когда число свободных приборов равно m, то $min\{k, m\}$ запросов начинают обслуживание. Если $k > m$, то m запросов начинают обслуживание, а $k - m$ запросов теряются.

Нетрудно видеть, что поведение данной системы описывается цепью

Маркова

$$\xi_t = \{i_t, \nu_t, \eta_t^{(1)}, .., \eta_t^{(i_t)}\}, t \geq 0,$$

где i_t – число запросов в системе, ν_t – состояние управляющего процесса $BMAP$ потока, $\eta_t^{(i)}$ – состояние управляющего процесса PH обслуживания в i-м приборе, $i_t = \overline{0, N}, \nu_t = \overline{0, W}, \eta_t^{(i)} = \overline{1, M}, i = \overline{1, N}, t \geq 0$.

Лемма 4.4. *Инфинитезимальный генератор Q ЦМ $\xi_t, t \geq 0$, имеет следующую структуру:*

$$Q = (Q_{m,m'})_{m,m'=\overline{0,N}} = \tag{4.78}$$

$$\begin{pmatrix} D_0 & \mathcal{T}_1^1(0) & \mathcal{T}_2^2(0) & ... & \mathcal{T}_{N-1}^{N-1}(0) & \sum\limits_{k=0}^{\infty} \mathcal{T}_{N+k}^N(0) \\ I_{\overline{W}} \otimes S_0^{\oplus 1} & D_0 \oplus S^{\oplus 1} & \mathcal{T}_1^1(1) & ... & \mathcal{T}_{N-2}^{N-2}(1) & \sum\limits_{k=0}^{\infty} \mathcal{T}_{N+k-1}^{N-1}(1) \\ 0 & I_{\overline{W}} \otimes S_0^{\oplus 2} & D_0 \oplus S^{\oplus 2} & ... & \mathcal{T}_{N-3}^{N-3}(2) & \sum\limits_{k=0}^{\infty} \mathcal{T}_{N+k-2}^{N-2}(2) \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & I_{\overline{W}} \otimes S_0^{\oplus N} & \sum\limits_{k=0}^{\infty} D_k \oplus S^{\oplus N} \end{pmatrix}.,$$

где

$$\mathcal{T}_k^r(m) = D_k \otimes I_{M^m} \otimes \boldsymbol{\beta}^{\otimes r},\ k \geq 1,\ m = \overline{0, N},\ r = \overline{0, N}.$$

Поскольку пространство состояний данной ЦМ конечное и цепь является неприводимой, то при любых значениях параметров системы существует стационарное распределение вероятностей состояний системы

$$p(i, \nu, m^{(1)}, \ldots, m^{(i)}) = \lim_{t \to \infty} P\{i_t = i, \nu_t = \nu, \eta_t^{(1)} = m^{(1)}, \ldots, \eta_t^{(i)} = m^{(i)}\}.$$

Вводя лексикографическое упорядочивание состояний ЦМ ξ_t, формируем векторы $\mathbf{p}_i,\ i = \overline{0, N}$ вероятностей состояний ЦМ, имеющих значение i первой компонент. Обозначим также $\mathbf{p} = (\mathbf{p_0}, .., \mathbf{p_N})$.

Вектор $\mathbf{p}$ удовлетворяет системе уравнений

$$\mathbf{p}Q = \mathbf{0},\ \mathbf{pe} = 1, \tag{4.79}$$

где генератор Q ЦМ ξ_t задан формулой (4.78).

Доказательство леммы 4.4 аналогично приведенному выше доказательству леммы 4.3.

В случае, когда размерность вектора $\mathbf{p}$ невелика, конечная система уравнений (4.79) с матрицей Q заданной формулой (4.78) легко решается на компьютере. Но размер матрицы Q равен $K = \bar{W}\frac{M^{N+1}-1}{M-1}$ и может быть

довольно велик. Например, если $\bar{W} = 2, M = 2, N = 5$, то размер равен 126, при $N = 6$ он уже равен 254 и т.д. Таким образом, при большом числе N приборов, типичном для современных телекоммуникационных сетей, решение системы (4.79) может быть долгим или вообще невыполнимым из-за нехватки оперативной памяти компьютера.

К счастью, матрица Q системы имеет специальную структуру (она является верхней хессенберговой). Это позволяет разработать более эффективные процедуры для решения системы (4.79). Одна из возможных процедур состоит в последовательном исключении блочных компонент вектора $\mathbf{p}_i,\ i = 0, \ldots, N-1$, неизвестного вектора $\mathbf{p}$. А именно, сначала исключаем вектор $\mathbf{p}_0$ из первого уравнения системы (4.79), затем вектор $\mathbf{p}_1$ из второго уравнения системы (4.79) и т.д., вектор $\mathbf{p}_{N-1}$ из N-го уравнения. $(N+1)$-е уравнение системы (4.79) дает уравнение для компонент вектора $\mathbf{p}_N$. Эта система относительно $\bar{W}M^N$ компонент вектора $\mathbf{p}_N$. Ее ранг на единицу меньше. Заменяя одно из уравнений этой системы на уравнение, получаемое из условия нормировки, вычисляем единственное решение $\mathbf{p}_N$ этой системы. Затем последовательно вычисляем все векторы $\mathbf{p}_i, i = N-1, \ldots, 1, 0$. Эта процедура легко выполняется на компьютере. Но она не является достаточно устойчивой численно из-за присутствия операции последовательного вычитания матриц или векторов. Альтернативный алгоритм, устойчивый при компьютерной реализации, основанный на использовании вероятностного смысла вектора $\mathbf{p}$, описанный в разделе 3.5, приводит к процедуре расчета векторов $\mathbf{p_i},\ i = 0, \ldots, N$, заданный следующим утверждением.

Теорема 4.17. *Векторы стационарных вероятностей* $\mathbf{p}_i,\ i = 0, \ldots, N$, *вычисляются по формуле*

$$\mathbf{p}_l = \mathbf{p}_0 F_l, l = \overline{1, N},$$

где матрицы F_l *вычисляются по рекуррентным формулам:*

$$F_l = (\bar{Q}_{0,l} + \sum_{i=1}^{l-1} F_i \bar{Q}_{i,l})(-\bar{Q}_{l,l})^{-1},\ l = \overline{1, N-1}, \tag{4.80}$$

$$F_N = (Q_{0,N} + \sum_{i=1}^{N-1} F_i Q_{i,N})(-Q_{N,N})^{-1}, \tag{4.81}$$

в которых матрицы $\overline{Q}_{i,N}$ *задаются обратной рекурсией*

$$\overline{Q}_{i,N} = Q_{i,N}, i = \overline{0, N},$$

$$\overline{Q}_{i,l} = Q_{i,l} + \overline{Q}_{i,l+1} G_l, i = \overline{0,l}, l = N-1, N-2, \ldots, 0,$$

а матрицы $G_i, i = \overline{0, N-1}$ *находятся из обратной рекурсии*

$$G_i = (-Q_{i+1,i+1} - \sum_{l=1}^{N-i-1} Q_{i+1,i+1+l} G_{i+l} G_{i+l-1} \ldots G_{i+1})^{-1} Q_{i+1,i},$$

$$i = N-1, N-2, \ldots, 0,$$

а вектор $\mathbf{p}_0$ *является единственным решением системы линейных алгебраических уравнений*

$$\mathbf{p}_0 \overline{Q}_{0,0} = \mathbf{0}, \tag{4.82}$$

$$\mathbf{p}_0 (\sum_{l=1}^{N} F_l \mathbf{e} + \mathbf{e}) = 1. \tag{4.83}$$

Замечание 4.5. Процедуру вычисления векторов $\mathbf{p}_i,\ i = 0, \ldots, N$, можно несколько упростить с целью экономии необходимой для ее выполнения памяти компьютера. Для этого сначала решаем систему линейных алгебраических уравнений (4.82), которая имеет единственное решение с точностью до константы. Обозначим это решение как $\tilde{\mathbf{p}}_0$, а константу обозначим как γ. Теперь вместо матриц $F_l,\ l = \overline{0, N}$, можно работать с векторами $\mathbf{f}_l$, заданными формулами $\mathbf{f}_l = \gamma \tilde{\mathbf{p}}_0 F_l$. Т.е. вместо рекуррентного расчета матриц $F_l,\ l = \overline{0, N}$, по формулам (4.80) и (4.81) рекуррентно рассчитываем векторы $\mathbf{f}_l,\ l = \overline{0, N}$, по формулам

$$\mathbf{f}_l = (\tilde{\mathbf{p}}_0 \bar{Q}_{0,l} + \sum_{i=1}^{l-1} \mathbf{f}_i \bar{Q}_{i,l})(-\bar{Q}_{l,l})^{-1}, l = \overline{1, N-1},$$

$$\mathbf{f}_N = (\tilde{\mathbf{p}}_0 Q_{0,N} + \sum_{i=1}^{N-1} \mathbf{f}_i Q_{i,N})(-Q_{N,N})^{-1}.$$

Значение неизвестной константы γ находим теперь из условия нормировки

$$\sum_{l=0}^{N} \mathbf{f}_l \mathbf{e} = \gamma,$$

а векторы $\mathbf{p}_i,\ i = 0, \ldots, N$, вычисляем по формулам $\mathbf{p}_i = \frac{\mathbf{f}_i}{\gamma},\ i = 0, \ldots, N$.

Вычислив вектор $\mathbf{p}$, можно рассчитать значение различных характеристик производительности рассматриваемой системы. Например, одна из основных характеристик – вероятность P_{loss} потери произвольного запроса – рассчитывается на основе следующего утверждения.

Теорема 4.18. *Вероятность P_{loss} потери произвольного запроса рассчитывается следующим образом:*

$$P_{loss} = 1 - \lambda^{-1} \sum_{i=0}^{N-1} \mathbf{p}_i \sum_{k=0}^{N-i} (k + i - N) \tilde{D}_k^{(i)} \mathbf{e}, \tag{4.84}$$

где $\tilde{D}_k^{(i)} = D_k \otimes I_{M^i}, k \geq 0$.

Краткое доказательство этого утверждения следующее. Согласно формуле полной вероятности, вероятность P_{loss} вычисляется как

$$P_{loss} = 1 - \sum_{i=0}^{N-1} \sum_{k=1}^{\infty} P_k P_i^{(k)} R^{(i,k)}, \tag{4.85}$$

где P_k есть вероятность того, что произвольный запрос прибывает в группе, состоящей из k запросов; $P_i^{(k)}$ есть вероятность того, что i занято в момент прихода группы, состоящей из k запросов; $R^{(i,k)}$ есть вероятность того, что произвольный запрос не будет потерян при условии, что он прибывает в группе, состоящей из k запросов и i занято в момент его прихода.

С использованием свойств $BMAP$-потока, описанных в подразделе 3.1.3, можно показать, что

$$P_i^{(k)} = \frac{\mathbf{p}_i \tilde{D}_k^{(i)} \mathbf{e}}{\boldsymbol{\theta} D_k \mathbf{e}}, i = \overline{0, N-1}, k \geq 1, \tag{4.86}$$

$$P_k = \frac{k \boldsymbol{\theta} D_k \mathbf{e}}{\boldsymbol{\theta} \sum\limits_{l=1}^{\infty} l D_l \mathbf{e}} = k \frac{\boldsymbol{\theta} D_k \mathbf{e}}{\lambda}, k \geq 1, \tag{4.87}$$

$$R^{(i,k)} = \begin{cases} 1 & , k \leq N - i, \\ \frac{N-i}{k}, & k > N - i, i = \overline{0, N-1}. \end{cases} \tag{4.88}$$

Подставляя (4.86)-(4.88) в (4.85), после некоторых алгебраических преобразований получаем (4.84).

4.5.2 Стационарное распределение вероятностей состояний системы при дисциплине CR

Согласно определению этой дисциплины, произвольная группа запросов полностью теряется, если в момент ее прихода число свободных приборов меньше, чем число запросов в группе.

Стационарное поведение рассматриваемой системы с такой дисциплиной доступа описывается многомерной ЦМ с непрерывным временем $\xi_t, t \geq 0$, имеющей такое же пространство состояний, что и цепь, изученная в предыдущем подразделе. Но генератор этой ЦП Маркова несколько другой.

Лемма 4.5. *Инфинитезимальный генератор Q ЦМ $\xi_t, t \geq 0$, имеет следующую блочную структуру:*

$$Q = (Q_{m,m'})_{m,m'=\overline{0,N}} = \tag{4.89}$$

$$\begin{pmatrix} D_0 + \sum\limits_{k=N+1}^{\infty} D_k & \mathcal{T}_1^1(0) & \mathcal{T}_2^2(0) & \dots & \mathcal{T}_N^N(0) \\ I_{\overline{W}} \otimes S_0^{\oplus 1} & (D_0 + \sum\limits_{k=N}^{\infty} D_k) \oplus S^{\oplus 1} & \mathcal{T}_1^1(1)... & \dots & \mathcal{T}_{N-1}^{N-1}(1) \\ 0 & I_{\overline{W}} \otimes S_0^{\oplus 2} & (D_0 + \sum\limits_{k=N-1}^{\infty} D_k) \oplus S^{\oplus 2} & \dots & \mathcal{T}_{N-2}^{N-2}(2) \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sum\limits_{k=0}^{\infty} D_k \oplus S^{\oplus N} \end{pmatrix}.$$

Сравнивая генераторы, заданные формулами (4.78) и (4.89), видим, что они различаются только видом диагональных блоков и последнего блочного столбца.

Вектор $\mathbf{p}$ стационарных вероятностей можно рассчитать, используя алгоритм, заданный теоремой 4.17.

Теорема 4.19. *Вероятность P_{loss} потери произвольного запроса для системы с дисциплиной доступа CR рассчитывается следующим образом:*

$$P_{loss} = 1 - \lambda^{-1} \sum_{i=0}^{N-1} \mathbf{p}_i \sum_{k=0}^{N-i} k \tilde{D}_k^{(i)} \mathbf{e}.$$

Доказательство полностью совпадает с доказательством теоремы 4.18 за исключением того, что условная вероятность $R^{(i,k)}$ здесь имеет другой вид:

$$R^{(i,k)} = \begin{cases} 1, & k \leq N - i, \\ 0, & k > N - i, \ i = \overline{0, N-1}. \end{cases}$$

4.5.3 Стационарное распределение вероятностей состояний системы при дисциплине CA

Согласно определению этой дисциплины, произвольная группа запросов полностью принимается, если в момент ее прихода есть хотя бы один

свободный прибор. Часть запросов, соответствующая числу свободных приборов, немедленно начинает обслуживание. Остальные ожидают в очереди и обслуживаются по мере освобождения приборов. меньше, чем число запросов в группе.

Стационарное поведение рассматриваемой системы описывается многомерной ЦМ с непрерывным временем $\xi_t = \{i_t, \nu_t, \eta_t^{(1)}, ..., \eta_t^{(min\{N,i_t\})}\},\ t \geq 0$. В отличие от процессов, описывающих поведение системы в случае дисциплин PA и CR, если размер групп не ограничен, то пространство состояний процесса $\xi_t, t \geq 0$, бесконечно.

Лемма 4.6. *Инфинитезимальный генератор Q ЦМ $\xi_t, t \geq 0$, имеет следующую блочную структуру:*

$$Q_{m,m'} = \begin{cases} D_{m'} \otimes \boldsymbol{\beta}^{\otimes m'}, m = 0, m' = \overline{0, N}, \\ D_{m'} \otimes \boldsymbol{\beta}^{\otimes N}, m = 0, m' > N, \\ I_{\bar{W}} \otimes \mathbf{S}_0^{\oplus(m'+1)}, m = \overline{1, N}, m' = m-1, \\ I_{\bar{W}} \otimes (\mathbf{S}_0\boldsymbol{\beta})^{\oplus N}, m > N, m' = m-1, \\ D_0 \oplus S^{\oplus m}, m = \overline{1, N-1}, m' = m, \\ D(1) \oplus S^{\oplus N}, m \geq N, m' = m, \\ D_{m'-m} \otimes I_{M^m} \otimes \boldsymbol{\beta}^{\otimes(m'-m)}, m = \overline{1, N-1}, m < m' < N, \\ D_{m'-m} \otimes I_{M^m} \otimes \boldsymbol{\beta}^{\otimes(N-m)}, m = \overline{1, N-1}, m' \geq N, \\ 0, \textit{ в других случаях.} \end{cases}$$

Отметим, что матрицы $Q_{i,j}$ имеют размер $\bar{W}M^i \otimes \bar{W}M^j$ при $i \leq N, j \leq N$ и размер $\bar{W}M^N \otimes \bar{W}M^N$ при $i > N, j > N$.

Обозначим стационарные вероятности состояний ЦМ $\xi_t, t \geq 0$, как

$$p(i, \nu, m^{(1)}, .., m^{min\{(i,N)\}}) =$$

$$= \lim_{t\to\infty} P\{i_t = i, \nu_t = \nu, \eta_t^{(1)} = m^{(1)}, .., \eta_t^{min\{(i,N)\}}\}$$

и введем векторы- строки $\mathbf{p}_i, i \geq 0$, этих вероятностей, перенумерованных в лексикографическом порядке. Векторы $\mathbf{p}_i, i = \overline{0, N-1}$, имеют размер $\bar{W}M^i$, а векторы $\mathbf{p}_i, i \geq N$, имеют размер $\bar{W}M^N$.

Поскольку пространство состояний процесса $\xi_t, t \geq 0$, бесконечно, существование введенных стационарных вероятностей неочевидно. Но при стандартных предположениях о $BMAP$-потоке и обслуживании фазового типа process оно может быть довольно легко доказано с использованием результатов раздела 3.4, поскольку несложно убедиться, что ЦМ $\xi_t, t \geq 0$ принадлежит классу цепей типа $M/G/1$ или квазитеплицевых ЦМ.

Теорема 4.20. *Векторы стационарных вероятностей* $\mathbf{p}_i$, $i \geq 0$, *вычисляются по формуле*

$$\mathbf{p}_l = \mathbf{p}_0 F_l, l \geq 1,$$

где матрицы F_l *рекуррентно вычисляются по формулам*

$$F_l = (\bar{Q}_{0,l} + \sum_{i=1}^{l-1} F_i \bar{Q}_{i,l})(-\bar{Q}_{l,l})^{-1}, l \geq 1,$$

где матрицы $\bar{Q}_{i,l}$ *задаются формулами*

$$\bar{Q}_{i,l} = Q_{i,l} + \sum_{k=1}^{\infty} Q_{i,l+k} G^{max\{0,l+k-N\}} \cdot G_{min\{N,l+k\}-1} \cdot G_{min\{N,l+k\}-2} \cdot .. \cdot G_l,$$

$$i = 0, \ldots, l, l \geq 1,$$

матрица G *имеет вид*

$$G = I_{\bar{W}} \otimes (S_0 \boldsymbol{\beta})^{\oplus N} (-D(1) \oplus S^{\oplus N})^{-1},$$

а матрицы $G_i, i = \overline{0, N-1}$, *вычисляются из обратной рекурсии*

$$G_i = -(Q_{i+1,i+1} + \sum_{l=i+2}^{\infty} Q_{i+1,l} G^{max\{0,l-N\}} G_{min\{N,l\}-1} \times$$

$$\times G_{min\{N,l\}-2} \cdot \cdots \cdot G_{i+1})^{-1} Q_{i+1,i}, i = N-1, N-2, .., 0,$$

а вектор $\mathbf{p}_0$ *является единственным решением системы*

$$\mathbf{p}_0 \bar{Q}_{0,0} = \mathbf{0},$$

$$\mathbf{p}_0 (\sum_{l=1}^{\infty} F_l \mathbf{e} + \mathbf{e}) = 1.$$

Теорема 4.21. *Вероятность* P_{loss} *потери произвольного запроса для системы с дисциплиной доступа* CA *рассчитывается следующим образом:*

$$P_{loss} = 1 - \lambda^{-1} \sum_{i=0}^{N-1} \mathbf{p}_i \sum_{k=1}^{\infty} k \tilde{D}_k^{(i)} \mathbf{e}.$$

Доказательство полностью совпадает с доказательством теоремы 4.18 за исключением того, что условная вероятность $R^{(i,k)}$ здесь имеет другой вид:

$$R^{(i,k)} = \begin{cases} 1, i \leq N-1 \\ 0, i > N-1. \end{cases}$$

Численные примеры, иллюстрирующие расчет стационарных вероятностей состояний системы и вероятностей потери запроса, демонстрирующие эффект корреляции во входном потоке, дисперсии распределения времени, дисциплины доступа запросов, приведены в статье [152]. Как упоминалось выше, свойство инвариантности вероятности потери произвольного запроса относительно распределения времени обслуживания, справедливое при стационарном пуассоновском потоке, не выполняется, когда поток является $BMAP$-потоком. Это следует учитывать при проектировании реальных систем, в которых входной поток не является стационарным пуассоновским, в том числе, современных телекоммуникационных сетей.

ГЛАВА 5
СМО С ПОВТОРНЫМИ ВЫЗОВАМИ С КОРРЕЛИРОВАННЫМИ ВХОДНЫМИ ПОТОКАМИ

Важным разделом ТМО является теория СМО с повторными вызовами. В таких системах запрос, поступивший в систему и заставший прибор занятым, не становится в очередь неограниченной или ограниченной длины, как в системах с ожиданием, и не уходит из системы навсегда, как в системах с потерей заявок (с отказами). Этот запрос покидает систему на некоторое случайное время (говорят, что он уходит на орбиту), а затем повторяет попытку попасть на обслуживание. Будем предполагать, что запрос повторяет попытки до тех пор, пока не поступит на обслуживание.

Важность этого раздела теории обусловлена его широкими практическими приложениями. Область приложений лежит в оценивании производительности и проектировании телефонных сетей, локальных вычислительных сетей с протоколами случайного множественного доступа, широковещательных радиосетей, мобильных сотовых радиосетей. Явление повторных попыток является неотъемлемой чертой этих и многих других реальных систем и игнорирование этого явления (например, аппроксимация соответствующих систем СМО с отказами) может привести к значительным погрешностям при принятии инженерных решений.

Вообще говоря, случайные процессы, описывающие поведение систем с повторными вызовами, существенно сложнее аналогичных процессов в системах с ожиданием и отказами. Этим объясняется тот факт, что теория систем с повторными вызовами гораздо менее развита.

5.1 Система $BMAP/SM/1$ с повторными вызовами

Описание системы $BMAP/SM/1$ с ожиданием дано в разделе 4.2. Входной поток характеризуется матричной ПФ $D(z)$, а обслуживание — полумарковским ядром $B(t)$. В отличие от той системы, предполагаем, что запрос, поступивший в рассматриваемую систему и заставший прибор занятым, не становится в очередь, а идет на орбиту, откуда совершает попытки попасть на обслуживание через случайные интервалы времени.

Здесь предположим, что суммарный поток запросов, идущих на прибор с орбиты в промежуток времени, когда на орбите имеется i запросов, является стационарным пуассоновским с интенсивностью α_i, $i \geq 0$, $\alpha_0 = 0$. Далее рассмотрим несколько частных видов зависимости α_i от i. Пока же исследуем общий случай.

5.1.1 Стационарное распределение вложенной цепи Маркова

Изучение системы начнем с рассмотрения вложенной по моментам t_n окончания обслуживания запросов ЦМ $\{i_n, \nu_n, m_n\}$, $n \geq 1$, где i_n — число запросов на орбите в момент $t_n + 0$, $i_n \geq 0$, ν_n — состояние управляющего процесса ν_t $BMAP$-потока в момент t_n, m_n — состояние управляющего обслуживанием полумарковского процесса m_t в момент $t_n + 0$.

Пусть $P\{(i, \nu, m) \to (l, \nu', m')\}$ — одношаговые вероятности переходов ЦМ $\{i_n, \nu_n, m_n\}$, $n \geq 1$. Упорядочим состояния цепи в лексикографическом порядке и составим из вероятностей переходов матрицы

$$P_{i,l} = (P\{(i, \nu, m) \to (l, \nu', m')\})_{\nu,\nu'=\overline{0,W},\ m,m'=\overline{1,M}}.$$

Лемма 5.1. *Матрицы $P_{i,l}$ определяются следующим образом:*

$$P_{i,l} = \alpha_i(\alpha_i I - \tilde{D}_0)^{-1} Y_{l-i+1} + (\alpha_i I - \tilde{D}_0)^{-1} \sum_{k=1}^{l-i+1} \tilde{D}_k Y_{l-i-k+1}, \qquad (5.1)$$

$$l \geq \max\{0, i-1\}, \quad i \geq 0,$$

$$P_{i,l} = O,\ l < i-1,$$

где $\tilde{D}_i = D_i \otimes I_M$, $i \geq 0$, а матрицы Y_l определяются как коэффициенты разложения

$$\sum_{l=0}^{\infty} Y_l z^l = \hat{\beta}(-D(z)) = \int_0^{\infty} e^{D(z)t} \otimes dB(t).$$

Доказательство леммы состоит в использовании формулы полной вероятности с учетом того, что матрицы

$$\alpha_i(\alpha_i I - \tilde{D}_0)^{-1} = \alpha_i \int_0^{\infty} e^{\tilde{D}_0 t} e^{-\alpha_i t} dt$$

и

$$(\alpha_i I - \tilde{D}_0)^{-1}\tilde{D}_k = \int_0^\infty e^{\tilde{D}_0 t}e^{-\alpha_i t}\tilde{D}_k dt$$

задают вероятности переходов процесса $\{\nu_n, m_n\}$ за время с момента окончания обслуживания некоторого запроса до момента начала обслуживания следующего запроса. Причем первая из этих вероятностей соответствует случаю, когда следующим запросом будет один из i запросов, находящихся на орбите, а вторая описывает случай, когда этим запросом будет один из новых (первичных) запросов, поступивших в систему в составе группы размера k, $k \geq 1$.

Анализируя вид (5.1) матриц вероятностей переходов $P_{i,l}$, можно увидеть, что в случае, когда интенсивность α_i не зависит от i, то есть $\alpha_i = \gamma$, $i \geq 1$, $\gamma > 0$, эти матрицы являются функциями разности $i - l$. В этом и только в этом случае ЦМ i_n, ν_n, m_n, $n \geq 1$, принадлежит классу трехмерных КТЦМ. С точки зрения практики, этот случай можно интерпретировать двояко. Первый вариант: при наличии i запросов на орбите каждый запрос получает право совершать повторные попытки независимо от других запросов через экспоненциально распределенное (с параметром $\frac{\gamma}{i}$) время. Второй вариант: только одному из запросов разрешается совершать повторные попытки через экспоненциально распределенные (с параметром γ) интервалы времени. Ниже мы вернемся к этому случаю. Однако чаще в литературе по системам обслуживания с повторными запросами рассматривается случай, когда каждый из запросов совершает повторные попытки независимо от других через экспоненциально распределенное с параметром α время. В этом случае мы имеем: $\alpha_i = i\alpha$, $\alpha > 0$. В литературе рассматривается также следующий вид зависимости: $\alpha_0 = 0$, $\alpha_i = i\alpha + \gamma$, $i > 0$, комбинирующий обе возможности, описанные выше.

В обоих случаях ($\alpha_i = i\alpha$, $\alpha_i = i\alpha + \gamma$) матрицы переходных вероятностей $P_{i,l}$ зависят как от разности $l - i$ величин l и i, что присуще КТЦМ, так и от величины i непосредственно. Последнее обстоятельство приводит к тому, что рассматриваемая цепь не принадлежит классу многомерных КТЦМ. Устремляя в (5.1) i к бесконечности и предполагая, что $\alpha_i \to \infty$ (что выполняется, например, в вышеупомянутых случаях $\alpha_i = i\alpha$ и $\alpha_i = i\alpha + \gamma$), мы видим, что

$$\lim_{i\to\infty} P_{i,i+l-1} = Y_l,\ l \geq 0,$$

и $\sum\limits_{l=0}^{\infty} Y_l$ – стохастическая матрица.

Отсюда следует, что цепь $\xi_n = \{i_n, \nu_n, m_n\}$, $n \geq 1$, принадлежит классу АКТЦМ.

Установим условия эргодичности ЦМ ξ_n, $n \geq 1$. Будем предполагать, что предел $\lim\limits_{i\to\infty} \alpha_i$ существует, и различать два случая:

$$\lim_{i\to\infty} \alpha_i = \infty$$

и

$$\lim_{i\to\infty} \alpha_i = \gamma < \infty.$$

Теорема 5.1. *В случае* $\lim\limits_{i\to\infty} \alpha_i = \infty$ *стационарное распределение ЦМ* ξ_n, $n \geq 1$, *существует, если выполняется неравенство*

$$\rho < 1, \tag{5.2}$$

где $\rho = \lambda b_1$, λ *– интенсивность BMAP-потока,* b_1 *– среднее время обслуживания.*

В случае $\lim\limits_{i\to\infty} \alpha_i = \gamma$ *стационарное распределение цепи* ξ_n, $n \geq 1$, *существует, если выполняется неравенство*

$$\mathbf{x}(\hat{\beta}'(1) + (\gamma I - \tilde{D}_0)^{-1}\hat{\beta}(1)\tilde{D}'(1))\mathbf{e} < 1, \tag{5.3}$$

где вектор $\mathbf{x}$ *удовлетворяет системе линейных алгебраических уравнений*

$$\mathbf{x}\left(I - \hat{\beta}(1) - (\gamma I - \tilde{D}_0)^{-1}\hat{\beta}(1)\tilde{D}(1)\right) = \mathbf{0}, \;\; \mathbf{xe} = 1. \tag{5.4}$$

Доказательство. В случае $\lim\limits_{i\to\infty} \alpha_i = \infty$ матричная ПФ $Y(z)$,

$$Y(z) = \hat{\beta}(-D(z)) = \int_0^{\infty} e^{D(z)t} \otimes dB(t).$$

Поэтому доказательство формулы (5.2) полностью повторяет доказательство теоремы 4.4, приведенное выше.

В случае $\lim\limits_{i\to\infty} \alpha_i = \gamma$ ПФ $Y(z)$ переходных вероятностей предельной квазитеплицевой цепи для АКТЦМ ξ_n, $n \geq 1$, имеет следующий вид:

$$Y(z) = \hat{\beta}(-D(z)) + (\gamma I - \tilde{D}_0)^{-1}\tilde{D}(z)\hat{\beta}(-D(z)),$$

и формулы (5.3), (5.4) очевидно получаются из следствия 3.7. $\square$

Далее предполагаем, что параметры рассматриваемой системы удовлетворяют условиям (5.2) или (5.3), (5.4). Тогда существуют стационарные вероятности состояний ЦМ $\xi_n,\ n \geq 1$,

$$\pi(i,\nu,m) = \lim_{n\to\infty} P\{i_n = i, \nu_n = \nu, m_n = m\},\ i \geq 0, \nu = \overline{0,W}, m = \overline{1,M}.$$

Скомпонуем упорядоченные в лексикографическом порядке вероятности состояний, имеющих значение i компоненты i_n, в векторы $\boldsymbol{\pi}_i,\ i \geq 0$.

В случае произвольной зависимости суммарной интенсивности повторных вызовов α_i от числа i этих вызовов на орбите, нахождение векторов $\boldsymbol{\pi}_i,\ i \geq 0$, возможно только на основе алгоритма, изложенного в разделе 3.6.2.

Попытаемся рассмотреть подробнее частный случай, когда $\alpha_i = i\alpha + \gamma,\ \alpha \geq 0,\ \gamma \geq 0,\ \alpha + \gamma \neq 0$, с помощью аппарата производящих функций. Обозначим $\mathbf{\Pi}(z) = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i z^i,\ |z| < 1$.

Теорема 5.2. *Векторная ПФ $\mathbf{\Pi}(z)$ стационарного распределения вложенной ЦМ $\xi_n,\ n \geq 1$, в случае $\alpha > 0$ удовлетворяет линейному матричному дифференциально-функциональному уравнению*

$$\mathbf{\Pi}'(z) = \mathbf{\Pi}(z)\tilde{S}(z) + \boldsymbol{\pi}_0 \gamma \alpha^{-1} z^{-1} \Phi(z), \tag{5.5}$$

а в случае $\alpha = 0$ — линейному матричному функциональному уравнению

$$\mathbf{\Pi}(z)\left(zI - \hat{\beta}(-D(z)) - (\gamma I - \tilde{D}_0)^{-1}\tilde{D}(z)\hat{\beta}(-D(z))\right) =$$

$$= \boldsymbol{\pi}_0\left((-\tilde{D}_0)^{-1} - (\gamma I - \tilde{D}_0)^{-1}\right)\tilde{D}(z)\hat{\beta}(-D(z)). \tag{5.6}$$

Здесь

$$\tilde{S}(z) = \Phi^{-1}(z)\Phi'(z) + z^{-1}\Phi^{-1}(z)A\Phi(z) - z^{-1}\alpha^{-1}\tilde{D}_0\Phi(z),$$

$$\Phi(z) = \tilde{D}_0^{-1}\tilde{D}(z)\hat{\beta}(-D(z))(\hat{\beta}(-D(z)) - zI)^{-1},\ A = (\tilde{D}_0 - \gamma I)\alpha^{-1}. \tag{5.7}$$

Заметим, что матричная функция $\Phi(z)$ задает функциональное уравнение

$$\mathbf{\Pi}(z) = \mathbf{\Pi}(0)\Phi(z)$$

для СМО $BMAP/SM/1$ с бесконечным буфером, см. формулу (4.34).

Доказательство. Уравнения Чепмена – Колмогорова для векторов стационарных вероятностей $\boldsymbol{\pi}_i,\ i \geq 0$, имеют вид

$$\boldsymbol{\pi}_j = \sum_{i=0}^{j+1} \boldsymbol{\pi}_i P_{i,j}, j \geq 0. \tag{5.8}$$

Подставляя в (5.8) матрицы переходных вероятностей в виде (5.1), умножая уравнения системы (5.8) на соответствующие степени z и суммируя, получаем уравнение

$$\mathbf{\Pi}(z)(zI - \hat{\beta}(-D(z))) = \sum_{i=0}^{\infty} \boldsymbol{\pi}_i z^i (\alpha_i I - \tilde{D}_0)^{-1} \tilde{D}(z) \hat{\beta}(-D(z)). \tag{5.9}$$

Подставляем в (5.9) явный вид $\alpha_i = i\alpha + \gamma,\ i > 0$, интенсивности α_i. В случае $\alpha = 0$ функциональное уравнение (5.6) получается тривиальным образом. Рассмотрим теперь случай $\alpha > 0$. Используя разложение матричной функции на спектре матрицы $\tilde{D}_0$ (см. приложение А), можно доказать справедливость соотношения

$$\sum_{i=0}^{\infty} \boldsymbol{\pi}_i z^i (\alpha_i I - \tilde{D}_0)^{-1} = \alpha^{-1} \int \mathbf{\Pi}(z) z^{-A-I} dz z^A, \tag{5.10}$$

где матрица A определена в (5.7).

Подставляя (5.10) в (5.9), получаем интегральное уравнение для векторной ПФ $\mathbf{\Pi}(z)$, из которого путем дифференцирования получаем дифференциально-функциональное уравнение (5.5). В случае когда $\gamma = 0$, это уравнение является линейным матричным дифференциальным уравнением. □

В случае $\alpha = 0$ функциональное уравнение (5.6) решается с помощью алгоритма, изложенного в разделе 3.4.2.

В случае $\alpha > 0$ уравнение (5.5) решается легко, если входящий поток является стационарным пуассоновским. Решение уравнения (5.5) в общем случае представляется проблематичным, поскольку матрицы $\tilde{S}(z)$ и $\int \tilde{S}(z)dz$, вообще говоря, не коммутируют и решение не имеет вида матричной экспоненты. Проблема решения уравнения усугубляется двумя факторами: начальное условие $\boldsymbol{\pi}_0$ неизвестно и матрица $\tilde{S}(z)$ имеет сингулярности в единичном круге комплексной плоскости (это видно из выражения матрицы $\tilde{S}(z)$ через $\Phi(z)$ в (5.7) и известного нам факта о наличии W точек сингулярности в единичном круге у матрицы $\hat{\beta}(-D(z)) - zI$. Поэтому

так же, как и в случае произвольной зависимости интенсивности α_i от i, можно рекомендовать нахождение векторов стационарных вероятностей с помощью алгоритма, изложенного в разделе 3.4.2. Заметим, что уравнение (5.5) может быть полезным при рекуррентном вычислении факториальных моментов распределения. В частности, из (5.5) вытекает соотношение

$$\mathbf{\Pi}'(1)\mathbf{e} = \left[\boldsymbol{\pi}_0\gamma\alpha^{-1}\Phi_0 + \mathbf{\Pi}(1)\Phi_0^{-1}(\Phi_1 + (A - \Phi_0\alpha^{-1}\tilde{D}_0)\Phi_0)\right]\mathbf{e}, \qquad (5.11)$$

где матрицы Φ_0 и Φ_1 являются коэффициентами в разложении функции $\Phi(z)$, заданной в (5.7):

$$\Phi(z) = \Phi_0 + \Phi_1(z-1) + o(z-1).$$

Соотношение (5.11) может быть использовано для контроля точности вычисления стационарного распределения с помощью алгоритма, представленного в разделе 3.4.2.

5.1.2 Стационарное распределение вероятностей состояний системы в произвольный момент времени

Рассмотрим теперь вопрос о распределении вероятностей состояний процесса $\{i_t, \nu_t, m_t\}$, $t \geq 0$, в произвольный момент времени.

Пусть

$$p(i,\nu,m) = \lim_{t\to\infty} P\{i_t = i, \nu_t = \nu, m_t = m\},$$

$$\mathbf{p}(i,\nu) = (p(i,\nu,1),\ldots,p(i,\nu,M)), \quad \mathbf{p}_i = (\mathbf{p}(i,0),\ldots,\mathbf{p}(i,W)),$$

$$\mathbf{P}(z) = \sum_{i=0}^{\infty}\mathbf{p}_i z^i, \ |z| < 1.$$

Справедливо следующее утверждение.

Теорема 5.3. *ПФ $\mathbf{P}(z)$ стационарного распределения вероятностей состояний системы в произвольный момент времени выражается через ПФ $\mathbf{\Pi}(z)$ распределения вероятностей состояний системы в моменты окончания обслуживания запросов следующим образом:*

$$\mathbf{P}(z) = \lambda\mathbf{\Pi}(z)\left[z\hat{\beta}^{-1}(-D(z))\nabla^*(z) - I\right](\tilde{D}(z))^{-1}, \qquad (5.12)$$

где $\nabla^*(z) = \int\limits_0^\infty e^{D(z)t}\otimes d\nabla_B(t)$, а матрица $\nabla_B(t)$ определена в теореме 4.6.

Доказательство. Используя подход, основанный на вложенных ПМВ, получим следующие выражения для стационарного распределения $\mathbf{p}_i$, $i \geq 0$, через стационарное распределение $\boldsymbol{\pi}_i$, $i \geq 0$:

$$\mathbf{p}_0 = \lambda \boldsymbol{\pi}_0 \int_0^\infty \left[e^{D_0 t} \otimes I_M\right] dt = \lambda \boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1},$$

$$\mathbf{p}_i = \lambda \Bigg[\boldsymbol{\pi}_i \int_0^\infty \left[\left(e^{(D_0 - \alpha_i I)t} \right) \otimes I_M \right] dt + \tag{5.13}$$

$$+ \sum_{l=1}^{i} \boldsymbol{\pi}_l \int_0^\infty \int_0^t \left[\left(e^{(D_0 - \alpha_l I)v} \alpha_l \right) \otimes I_M \right] dv \left[P(i-l, t-v) \otimes (I - \nabla_B(t-v)) \right] dt +$$

$$+ \sum_{l=0}^{i-1} \boldsymbol{\pi}_l \int_0^\infty \int_0^t \sum_{k=1}^{i-l} \left[\left(e^{(D_0 - \alpha_l I)v} \right) \otimes I_M \right] \left[(D_k \otimes I_M) dv \right] \times$$

$$\times \left[P(i-l-k, t-v) \otimes (I - \nabla_B(t-v)) \right] dt \Bigg] =$$

$$= \lambda \Bigg[\boldsymbol{\pi}_i \left[(\alpha_i I - D_0)^{-1} \otimes I_M \right] +$$

$$+ \sum_{l=1}^{i} \boldsymbol{\pi}_l \left[\alpha_l (\alpha_l I - D_0)^{-1} \otimes I_M \right] \int_0^\infty \left[P(i-l, t) \otimes (I - \nabla_B(t)) \right] dt +$$

$$+ \sum_{l=0}^{i-1} \boldsymbol{\pi}_l \left[(\alpha_l I - D_0)^{-1} \otimes I_M \right] \left[\sum_{k=1}^{i-l} D_k \otimes I_M \right] \int_0^\infty \left[P(i-l-k, t) \otimes (I - \nabla_B(t)) \right] dt \Bigg], \; i > 0.$$

Умножая полученные соотношения на соответствующие степени z и суммируя, получим следующее выражение для производящей функции $\mathbf{P}(z)$:

$$\mathbf{P}(z) = \lambda \Bigg[\boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1} + \sum_{l=1}^{\infty} \boldsymbol{\pi}_l z^l (\alpha_l I - \tilde{D}_0)^{-1} + \tag{5.14}$$

$$+ \left[\boldsymbol{\pi}_0 (-\tilde{D}_0)^{-1} (\tilde{D}(z) - \tilde{D}_0) + \sum_{l=1}^{\infty} \boldsymbol{\pi}_l z^l + \sum_{l=1}^{\infty} \boldsymbol{\pi}_l z^l (\alpha_l I - \tilde{D}_0)^{-1} \tilde{D}(z) \right] \times$$

$$\times \int_0^\infty \left[e^{D(z)t} \otimes (I - \nabla_B(t)) \right] dt \Bigg].$$

Интеграл в (5.14) после интегрирования по частям вычисляется как

$$\int_0^\infty \left[e^{D(z)t} \otimes (I - \nabla_B(t))\right] dt = (-\tilde{D}(z))^{-1}(I - \nabla^*(z)).$$

Подставляя полученное выражение в (5.14), после преобразований получим:

$$\mathbf{P}(z) = \lambda \left[-\mathbf{\Pi}(z)(\tilde{D}(z))^{-1} + [\mathbf{\Pi}(z)(\tilde{D}(z))^{-1} + \sum_{l=0}^{\infty} \boldsymbol{\pi}_l z^l (\alpha_l I - \tilde{D}_0)^{-1}]\nabla^*(z) \right].$$

Отсюда с учетом формулы (5.9) получаем формулу (5.12). □

Следствие 5.1. *Соотношение (5.12) между ПФ* $\mathbf{P}(z)$ *и* $\mathbf{\Pi}(z)$ *для системы* $BMAP/SM/1$ *с повторными вызовами имеет такой же вид, как и аналогичное соотношение для системы* $BMAP/SM/1$ *с ожиданием.*

Следствие 5.2. *Скалярная ПФ* $p(z) = \mathbf{P}(z)\mathbf{e}$ *распределения числа запросов в системе* $BMAP/SM/1$ *с повторными вызовами имеет вид*

$$p(z) = \lambda(z-1)\mathbf{\Pi}(z)(\tilde{D}(z))^{-1}\mathbf{e}.$$

5.1.3 Характеристики производительности системы

Вычислив распределение вероятностей состояний системы во вложенные моменты времени и в произвольный момент времени, можно найти различные характеристики проиводительности системы. Приведем некоторые из них.

- Вектор $\mathbf{q}_0(i)$, $(\nu(W+1)+m)$-я компонента которого равна вероятности того, что в произвольный момент времени прибор простаивает, в то время как на орбите находится i запросов, управляющий процесс $BMAP$ находится в состоянии ν, управляющий процесс обслуживания — в состоянии m, имеет вид

$$\mathbf{q}_0(i) = \lambda\boldsymbol{\pi}(i)(\alpha_i I - \tilde{D}_0)^{-1},\ i \geq 0.$$

- Вектор $\mathbf{q}_0 = \sum_{i=0}^{\infty} \mathbf{q}_0(i)$ вычисляется как

$$\mathbf{q}_0 = \lambda\mathbf{\Pi}(1)\Phi_0^{-1}\tilde{D}_0^{-1}.$$

- Условная вероятность $q_0^{(i)}$ того, что в произвольный момент прибор простаивает при условии, что на орбите находится i запросов, вычисляется по формуле

$$q_0^{(i)} = \frac{\mathbf{q}_0(i)\mathbf{e}}{\mathbf{p}_i\mathbf{e}},\ i \geq 0.$$

- Вероятность $p_0^{(a)}$ того, что произвольный запрос попадает на обслуживание немедленно в момент поступления, имеет вид

$$p_0^{(a)} = -\frac{\mathbf{q}_0\tilde{D}_0\mathbf{e}}{\lambda}.$$

- Вероятность $p_0^{(b)}$ того, что первый запрос в группе попадает на обслуживание сразу в момент поступления группы, имеет вид

$$p_0^{(b)} = \frac{-\mathbf{q}_0 D_0\mathbf{e}}{\lambda_g},\ \lambda_g = -\boldsymbol{\theta}\tilde{D}_0\mathbf{e},$$

  где $\boldsymbol{\theta}$ — вектор стационарных вероятностей процесса ν_t.
- Вероятность p_0 того, что в произвольный момент времени система пуста, вычисляется как

$$p_0 = \mathbf{p}_0\mathbf{e}.$$

- Среднее число запросов L в системе в произвольный момент времени вычисляется как

$$L = \mathbf{P}'(1)\mathbf{e}.$$

5.2 Система $BMAP/PH/N$ с повторными вызовами

Важность исследования СМО с повторными вызовами уже отмечалась выше и в предыдущем разделе мы рассмотрели однолинейную систему с повторными вызовами типа $BMAP/SM/1$. Однолинейные системы часто применяются для моделирования процессов передачи в локальных сетях связи, в которых передача информации между станциями производится через один общий канал, например шину или кольцо. При моделировании других реальных систем, в которых присутствуют повторные вызовы, например мобильных сетей связи или контакт-центров, необходимо уметь рассчитывать характеристики многолинейных систем с повторными вызовами. Однако задача исследования многолинейных систем с повторными

вызовами очень сложна и не имеет аналитического решения даже в простейших предположениях о входном потоке (стационарный пуассоновский поток) и процессе обслуживания (времена обслуживания являются независимыми одинаково распределенными случайными величинами, имеющими экспоненциальное распределение). Результаты, полученные в разделе 3.7 для АКТЦМ с непрерывным временем, позволяют рассчитывать характеристики многолинейных СМО с весьма общими предположениями о входном потоке и процессе обслуживания.

В качестве входного потока будем рассматривать $BMAP$. Заметим, что в случае однолинейных систем с $BMAP$-потоком их аналитическое исследование возможно путем применения метода вложенных ЦМ или метода введения дополнительной переменной даже для такого общего процесса обслуживания, как полумарковский процесс (SM). Для многолинейных систем оба этих метода оказываются бессильными. Поэтому максимально общим процессом обслуживания, при котором возможно аналитическое исследование многолинейных систем, является рекуррентный процесс, у которого времена обслуживания запросов являются независимыми одинаково распределенными случайными величинами, имеющими распределение фазового типа, также описанное в разделе 3.1.

Таким образом, наиболее сложной многолинейной системой с повторными вызовами, поддающейся аналитическому исследованию, на теперешний день может считаться система типа $BMAP/PH/N$, которой и будет посвящен данный раздел.

5.2.1 Описание системы

Система имеет N идентичных обслуживающих приборов. В нее поступает $BMAP$-поток запросов с управляющим процессом ν_t, $t \geq 0$, принимающим значения в множестве $\{0, 1, \ldots, W\}$. Поведение $BMAP$ описывается квадратными матрицами D_k, $k \geq 0$, порядка $W + 1$ или их ПФ $D(z) = \sum_{k=0}^{\infty} D_k z^k$, $|z| \leq 1$.

Если поступившая группа запросов застала свободные приборы, соответствующее число приборов начинают обслуживание запросов. Если свободных приборов нет или их меньше, чем число запросов в группе, то вся группа или запросы, которым не хватило приборов, идут в некоторую виртуальную область, называемую орбитой. Емкость орбиты предполагается неограниченной. Запросы, находящиеся на орбите, делают попытки

попасть на обслуживание позже. Предполагаем, что при нахождении на орбите i запросов вероятность повторной попытки с орбиты в интервале $(t, t+\Delta t)$ равна $\alpha_i \Delta t + o(\Delta t)$, $i > 0$, $\alpha_0 = 0$. В разделе 5.1 было отмечено, что популярными среди исследователей являются следующие зависимости α_i от i : $\alpha_i = i\alpha$, $\alpha > 0$, $\alpha_i = \gamma$, $i > 0$, $\alpha_i = i\alpha + \gamma$. В данном разделе будем различать случай постоянной интенсивности повторов ($\alpha_i = \gamma$, $i > 0$) и общий случай бесконечно возрастающей интенсивности повторов ($\lim\limits_{i\to\infty} \alpha_i = \infty$), при котором явная зависимость от i не фиксируется.

Если в момент совершения каким-либо запросом с орбиты попытки попасть на обслуживание имеются свободные приборы, один из них начинает обслуживание данного запроса. Если же в момент совершения попытки свободных приборов нет, запрос возвращается на орбиту. Каждый из запросов повторяет попытки попасть на обслуживание до тех пор, пока не попадет на обслуживание.

Время обслуживания произвольного запроса имеет распределение фазового типа, которое задается вероятностным вектором $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$ и генератором S размера $M \times M$.

5.2.2 Цепь Маркова, описывающая функционирование системы

Пусть

- i_t – число запросов на орбите, $i_t \geq 0$;
- n_t – число занятых приборов, $n_t = \overline{0, N}$;
- $m_t^{(j)}$ – состояние ЦМ, управляющей процессом обслуживания в j-м занятом приборе, $m_t^{(j)} = \overline{1, M}$, $j = \overline{1, n_t}$ (мы предполагаем, что занятые приборы нумеруются в порядке их занятия, то есть прибор, начинающий обслуживание, когда занято k приборов, получает номер k; при окончании обслуживания каким-либо прибором он теряет свой номер, а оставшиеся приборы соответственно перенумеровываются);
- ν_t – состояние ЦМ, управляющей поступлением запросов в $BMAP$-потоке, $\nu_t = \overline{0, W}$,

в момент t, $t \geq 0$.

Рассмотрим многомерный процесс с непрерывным временем

$$\xi_t = (i_t,\, n_t,\, \nu_t,\, m_t^{(1)}, \ldots, m_t^{(n_t)}),\, t \geq 0.$$

Несложно убедиться, что зная состояние данного процесса в момент времени t, можно описать будущее поведение данного процесса, то есть данный процесс является ЦМ с непрерывным временем. Также нетрудно убедиться, что эта цепь является регулярной и неприводимой.

Уровнем i ЦМ ξ_t будем называть множество состояний, у которых компонента i_t принимает значение i, а другие компоненты упорядочены в лексикографическом порядке. На уровне i находится $K = (W+1)\frac{M^{N+1}-1}{M-1}$ состояний. Отметим, что число K может быть довольно большим. Например, если пространства состояний цепей Маркова, управляющих поступлением и обслуживанием запросов, состоят только из двух элементов ($W = 1$, $M = 2$), а число приборов N равно 5, то $K = 126$, если $N = 6$, то $K = 254$ и т.д. Быстрый рост числа K состояний в уровне влечет определенные трудности при реализации алгоритма вычисления стационарного распределения в случае более-менее больших значений N.

Вместе с тем, описание функционирования системы в терминах выбранной ЦМ довольно наглядное и полезное с точки зрения простоты изложения материала. Несколько ниже мы кратко опишем другой способ выбора ЦМ, описывающей функционирование рассматриваемой СМО. При этом выборе число состояний, входящих в уровень, может быть существенно меньшим.

Лемма 5.2. *Инфинитезимальный генератор ЦМ ξ_t имеет следующий вид:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & \dots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \dots \\ O & O & Q_{3,2} & Q_{3,3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{5.15}$$

где блоки $Q_{i,j}$ задают интенсивности перехода цепи с уровня i на уровень j и имеют следующий вид:

$$Q_{i,i-1} = \alpha_i \begin{pmatrix} 0 & I_{\bar{W}} \otimes \boldsymbol{\beta} & 0 & \dots & 0 \\ 0 & 0 & I_{\bar{W}M} \otimes \boldsymbol{\beta} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & I_{\bar{W}M^{N-1}} \otimes \boldsymbol{\beta} \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \; i \geq 1, \tag{5.16}$$

$$Q_{i,i+k} = \begin{pmatrix} 0 & \dots & 0 & D_{k+N} \otimes \boldsymbol{\beta}^{\otimes N} \\ 0 & \dots & 0 & D_{k+N-1} \otimes I_M \otimes \boldsymbol{\beta}^{\otimes(N-1)} \\ 0 & \dots & 0 & D_{k+N-2} \otimes I_{M^2} \otimes \boldsymbol{\beta}^{\otimes(N-2)} \\ \vdots & \ddots & \vdots & \\ 0 & \dots & 0 & D_k \otimes I_{M^N} \end{pmatrix}, \ k \geq 1, \tag{5.17}$$

а блоки $(Q_{i,i})_{n,n'}$ *матрицы* $Q_{i,i}$ *задаются следующим образом:*

$$(Q_{i,i})_{n,n'} = \begin{cases} O, \, n' < n-1, \, n = \overline{2,N}, \\ \\ I_{\bar{W}} \otimes \mathbf{S}_0^{\oplus n}, \, n' = n-1, \, n = \overline{1,N}, \\ \\ D_0 \oplus S^{\oplus n} - \alpha_i(1-\delta_{n,N}) I_{\bar{W}M^n}, \, n' = n, \, n = \overline{0,N}, \\ \\ D_l \otimes I_{M^n} \otimes \boldsymbol{\beta}^{\otimes l}, \, n' = n+l, \, l = \overline{1,N-n}, \, n = \overline{0,N}, \end{cases} \tag{5.18}$$

$$i \geq 0.$$

Здесь $\delta_{n,N} = \begin{cases} 1, \, n = N, \\ 0, \, n \neq N \end{cases}$ *есть символ Кронекера,* $\boldsymbol{\beta}^{\otimes r} \stackrel{def}{=} \underbrace{\boldsymbol{\beta} \otimes \dots \otimes \boldsymbol{\beta}}_{r}$,

$r \geq 1, \ \mathbf{S}_0^{\oplus l} \stackrel{def}{=} \sum\limits_{m=0}^{l-1} I_{M^m} \otimes \mathbf{S}_0 \otimes I_{M^{l-m-1}}, \quad l \geq 1.$

Доказательство. Доказательство проводится путем анализа вероятностей переходов ЦМ ξ_t за бесконечно малый интервал времени. При этом используется прозрачный вероятностный смысл матриц, присутствующих в формулировке теоремы. Матрица A, заданная формулой (5.15), является блочно-верхне-хессенберговой, то есть все блоки этой матрицы ниже первой поддиагонали – нулевые. Это объясняется тем, что уменьшение числа запросов за малый интервал времени Δt более чем на один запрос возможно только с вероятностью порядка $o(\Delta t)$.

Поясним вероятностный смысл блоков $Q_{i,j}, \ i \geq 0, \, j \geq \min\{0, i-1\}$. Каждый из этих блоков является блочной матрицей с блоками $(Q_{i,j})_{n,n'}$, состоящими из интенсивностей переходов ЦМ ξ_t из состояний уровня i в состояния уровня j, при которых значение компоненты n_t изменяется с n на $n', \ n, n' = \overline{0,N}$.

Матрицы $Q_{i,i-1}$ задают интенсивности переходов из состояний, входящих в уровень i, в состояния, входящие в уровень $i-1$. Интенсивности

таких переходов равны интенсивности α_i осуществления повторной попытки. Это объясняет наличие сомножителя α_i в правой части (5.16). Если в момент повторной попытки все N приборов заняты, то повторная попытка не увенчивается успехом и запрос возвращается на орбиту. В этом случае переход на уровень $i-1$ невозможен. Поэтому последняя блочная строка в матрице $Q_{i,i-1}$ нулевая. Если число занятых приборов в момент повторной попытки равно $n, n < N$, то происходит переход на уровень $i-1$ и увеличение числа занятых приборов на единицу. Этим объясняется тот факт, что блоки $(Q_{i,i-1})_{n,n+1}$, стоящие на первой наддиагонали, отличны от нуля, в то время как все остальные блоки этой матрицы – нулевые.

Повторная попытка, произошедшая в малом интервале времени, исключает возможность переходов управляющих процессов потока и обслуживания в этом интервале. Это объясняет наличие сомножителя $I_{\bar{W}M^n}$ в первом наддиагональном блоке в n-й блочной строке матрицы $Q_{i,i-1}$. Этот сомножитель кронекерово умножается на вектор $\boldsymbol{\beta}$, поскольку в момент занятия прибора (пусть это будет j-й прибор) запросом, поступившим с орбиты, происходит розыгрыш (в соответствии с вектором $\boldsymbol{\beta}$) начального состояния ЦМ $m_t^{(j)}$, которая будет управлять процессом обслуживания на данном приборе. Заметим, что операция кронекерова произведения матриц вообще очень полезна при описании интенсивностей совместных переходов отдельных независимых компонент многомерных ЦМ.

Матрицы $Q_{i,i+k}$ задают интенсивности переходов из состояний, входящих в уровень i, в состояния, входящие в уровень $i+k$. Такие переходы происходят, когда в систему поступает группа, размер которой равен числу свободных приборов в момент поступления плюс k запросов. Если число занятых приборов в момент поступления равно n, $n \leq N$, то интенсивности поступления такой группы задаются элементами матрицы D_{k+N-n}. k запросов поступившей группы идут на орбиту, остальные – занимают все свободные $N-n$ приборов, и число занятых приборов становится равным N. Этим объясняется тот факт, что все блоки матрицы $Q_{i,i+k}$, кроме блоков $(Q_{i,i+k})_{n,N}$ последнего блочного столбца, – нулевые. При занятии свободных приборов в каждом из $N-n$ этих приборов происходит розыгрыш (в соответствии с вектором $\boldsymbol{\beta}$) начального состояния ЦМ, которая будет управлять процессом обслуживания на данном приборе. Результат одновременного розыгрыша в $N-n$ приборах определяется вектором $\boldsymbol{\beta}^{\otimes(N-n)}$. Состояния же управляющих процессов обслуживания в n занятых приборах не меняются. Из сказанного следует легко "читаемая" формула для

блока $(Q_{i,i+k})_{n,N}$

$$(Q_{i,i+k})_{n,N} = D_{k+N-n} \otimes I_{M^n} \otimes \boldsymbol{\beta}^{\otimes(N-n)}.$$

Перейдем теперь к пояснению выражений, задающих матрицу $Q_{i,i}$.

Недиагональные элементы матриц $Q_{i,i}$ задают интенсивности переходов ЦМ ξ_t из состояний, входящих в уровень i, в состояния, входящие в этот же уровень. Диагональные элементы матриц $Q_{i,i}$ есть взятые со знаком минус интенсивности выхода из соответствующих состояний. Вид блоков $(Q_{i,i})_{n,n'}$ в формуле (5.18) объясняется следующим образом.

• За малый промежуток времени может произойти не более одного окончания обслуживания запроса. Поэтому $(Q_{i,i})_{n,n'} = O$, $n' < n-1$.

• Блок $(Q_{i,i})_{n,n-1}$ соответствует ситуации, когда за малый промежуток времени окончилось обслуживание в одном из n занятых каналов. Интенсивности окончания обслуживания в произвольном канале при заданном состоянии управляющего процесса обслуживания задаются вектором $\mathbf{S}_0$. Интенсивности окончания обслуживания в одном из n занятых каналов при заданном состоянии управляющего процесса обслуживания задаются вектором $\mathbf{S}_0 \otimes I_{M^{n-1}} + I_M \otimes \mathbf{S}_0 I_{M^{n-2}} + \ldots I_{M^{n-1}}\mathbf{S}_0$, что соответствует обозначению $\mathbf{S}_0^{\oplus n}$.

• Блок $(Q_{i,i})_{n,n}$ – диагональный и соответствует ситуации, когда за малый промежуток времени не изменилось ни число запросов на орбите, ни число занятых приборов. Недиагональные элементы этого блока определяют либо интенсивности переходов управляющего процессом потока (без переходов управляющих процессов обслуживания) и задаются соответствующими элементами матрицы $D_0 \otimes I_{M^n}$, либо интенсивности перехода управляющего процесса обслуживания в одном из n занятых каналов (без переходов переходов управляющего процесса потока) и задаются соответствующими элементами матрицы $I_{\bar{W}} \otimes S^{\oplus n}$. Диагональные элементы блока $(Q_{i,i})_{n,n}$ определяют интенсивности выхода ЦМ из текущего состояния за счет переходов управляющего процесса потока, управляющих процессов обслуживания. При этом должна отсутствовать повторная попытка с орбиты в случае $n \neq N$. С учетом того, что $D_0 \otimes I_{M^n} + I_{\bar{W}} \otimes S^{\oplus n} = D_0 \oplus S^{\oplus n}$, получаем формулу для блока $(Q_{i,i})_{n,n}$ в (5.18).

• Блок $(Q_{i,i})_{n,n+l}$, $l \geq 1$, соответствует ситуации, когда за малый промежуток времени поступила группа из l запросов и все эти запросы немедленно пошли на обслуживание. Вектор $\boldsymbol{\beta}^{\otimes l}$ задает начальное распределение процессов обслуживания этих запросов на соответствующих приборах. □

5.2.3 Условие эргодичности системы

Как и прежде, будем различать случаи постоянной интенсивности повторов ($\alpha_i = \gamma,\ i > 0$) и бесконечно возрастающей интенсивности повторов ($\lim\limits_{i\to\infty} \alpha_i = \infty$).

В первом случае несложно убедиться, что рассматриваемая ЦМ является квазитеплицевой с N блочными граничными условиями и необходимое и достаточное условие ее эргодичности легко получается из результатов раздела 3.4 в следующем виде.

Теорема 5.4. *В случае постоянной интенсивности повторов необходимое и достаточное условие эргодичности рассматриваемой СМО имеет вид*

$$\mathbf{y}((D^\star(z))'|_{z=1} - \gamma\hat{I})\,\mathbf{e} < 0, \tag{5.19}$$

где вектор $\mathbf{y}$ *является единственным решением системы линейных алгебраических уравнений*

$$\mathbf{y}(I - Y(1)) = \mathbf{0},\ \mathbf{ye} = 1. \tag{5.20}$$

Здесь

$$Y(z) = zI + (C + \gamma\hat{I})^{-1}(\gamma\tilde{I}_\beta - \gamma\hat{I}z + zD^\star(z)),$$

$$C = \mathrm{diag}\,\{\,\mathrm{diag}\,\{\,\lambda_\nu,\ \nu = \overline{0, W}\,\} \oplus [\,\mathrm{diag}\,\{\,s_m,\ m = \overline{1, M}\,\}\,]^{\oplus r},\ r = \overline{0, N}\,\},$$

$$D^\star(z) =$$

$$\begin{bmatrix}
D_0 & D_1 \otimes \boldsymbol{\beta}^{\otimes 1} & D_2 \otimes \boldsymbol{\beta}^{\otimes 2} & \dots & D_{N-1} \otimes \boldsymbol{\beta}^{\otimes(N-1)} & \delta_{N-1}(z, \boldsymbol{\beta}) \\
I_{\bar{W}} \otimes \mathbf{S}_0^{\oplus 1} & D_0 \oplus S^{\oplus 1} & D_1 \otimes I_M \otimes \boldsymbol{\beta}^{\otimes 1} & \dots & D_{N-2} \otimes I_M \otimes \boldsymbol{\beta}^{\otimes(N-2)} & \delta_{N-2}(z, \boldsymbol{\beta}) \\
O & I_{\bar{W}} \otimes \mathbf{S}_0^{\oplus 2} & D_0 \oplus S^{\oplus 2} & \dots & D_{N-3} \otimes I_{M^2} \otimes \boldsymbol{\beta}^{\otimes(N-3)} & \delta_{N-3}(z, \boldsymbol{\beta}) \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
O & O & O & \dots & I_{\bar{W}} \otimes \mathbf{S}_0^{\oplus N} & D(z) \oplus S^{\oplus N}
\end{bmatrix},$$

$$\Delta_m(z, \boldsymbol{\beta}) = z^{-m+1}(D(z) - \sum_{k=0}^{m} D_k z^k) \otimes I_{M^{N-m-1}} \otimes \boldsymbol{\beta}^{\otimes(m+1)}, m = \overline{0, N-1},$$

$$\hat{I} = \begin{pmatrix}
I_{\bar{W}} & O & \dots & O & O \\
O & I_{\bar{W}M} & \dots & O & O \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
O & O & \dots & I_{\bar{W}M^{N-1}} & O \\
O & O & \dots & O & O_{\bar{W}M^N}
\end{pmatrix},$$

$$\tilde{I}_{\boldsymbol{\beta}} = \begin{pmatrix} O & I_{\bar{W}} \otimes \boldsymbol{\beta} & O & \dots & O \\ O & O & I_{\bar{W}M} \otimes \boldsymbol{\beta} & \dots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \dots & I_{\bar{W}M^{N-1}} \otimes \boldsymbol{\beta} \\ O & O & O & \dots & O \end{pmatrix},$$

$\bar{I} = I - \hat{I}$.

Рассмотрим теперь случай бесконечно возрастающей интенсивности повторов ($\lim\limits_{i\to\infty} \alpha_i = \infty$).

Справедливо следующее утверждение.

Теорема 5.5. *В случае бесконечно возрастающей интенсивности повторов достаточное условие эргодичности СМО $BMAP/PH/N$ с повторными вызовами имеет вид:*

$$\rho = \lambda/\bar{\mu} < 1, \tag{5.21}$$

где λ – средняя скорость $BMAP$-потока, а величина $\bar{\mu}$ задается формулой

$$\bar{\mu} = \mathbf{y}\mathbf{S}_0^{\oplus N}\,\mathbf{e}_{M^{N-1}}, \tag{5.22}$$

где вектор $\mathbf{y}$ является единственным решением системы линейных алгебраических уравнений

$$\mathbf{y}(S^{\oplus N} + \mathbf{S}_0^{\oplus N}(I_{M^{N-1}} \otimes \boldsymbol{\beta})) = \mathbf{0},\ \mathbf{y}\mathbf{e} = 1. \tag{5.23}$$

Доказательство. В нашем случае матрицы $\hat{T}_i$, участвующие в определении АКТЦМ, имеют вид:

$$\hat{T}_i = C + \alpha_i \hat{I}.$$

Вычислим предельные матрицы $Y_k,\ k \geq 0$, и найдем их ПФ $Y(z)$. В результате получим формулу

$$Y(z) = \tilde{I}_{\boldsymbol{\beta}} + z\bar{I} + zC^{-1}\bar{I}D^{\star}(z). \tag{5.24}$$

Подставляя в (5.24) явный вид матриц $\bar{I}, \tilde{I}_{\boldsymbol{\beta}}$, C и $D^{\star}(z)$, легко убедиться, что матрица $Y(z)$ является приводимой, причем нормальная форма матрицы $Y(1)$ имеет только один неприводимый стохастический диагональный блок $\tilde{Y}(1)$. Соответствующий блок $\tilde{Y}(z)$ матрицы $Y(z)$ имеет вид

$$\tilde{Y}(z) =$$

$$\begin{pmatrix} (\Lambda \oplus S_N)^{-1}(D(z) \oplus S^{\oplus N})z + zI & (\Lambda \oplus S_N)^{-1}(I_{\bar{W}} \otimes \mathbf{S}_0^{\oplus N})z \\ I_{\bar{W}M^{N-1}} \otimes \boldsymbol{\beta} & O_{\bar{W}M^{N-1}} \end{pmatrix}. \quad (5.25)$$

Здесь $\Lambda \oplus S_N \stackrel{def}{=} \mathrm{diag}\,\{\,\lambda_\nu,\, \nu = \overline{0, W}\,\} \oplus [\,\mathrm{diag}\,\{\,s_m,\, m = \overline{1, M}\,\}\,]^{\oplus N}$.

Из теоремы 3.14 следует, что достаточным условием эргодичности ЦМ $\xi_t, t \geq 0$, является выполнение неравенства

$$(\,\det\,(zI - \tilde{Y}(z)\,)'\,|_{z=1} > 0. \quad (5.26)$$

Учитывая блочную структуру определителя $\det\,(zI - \tilde{Y}(z))$, его можно записать в виде

$$\det\,(zI - \tilde{Y}(z)) = \det\,(\Lambda \oplus S_N)^{-1} z^{\bar{W}M^{N-1}} \det R(z), \quad (5.27)$$

где

$$R(z) = -z(D(z) \oplus S^{\oplus N}) - (I_{\bar{W}} \otimes S_0^{\oplus N})\,(I_{\bar{W}M^{N-1}} \otimes \beta).$$

Легко проверить, что матрица $R(1)$ является неприводимым инфинитезимальным генератором. Это свойство учитывается в дальнейших выкладках. Дифференцируя (5.27) в точке $z = 1$, учитывая, что $\det R(1) = 0$, и подставляя в (5.26), получим следующее неравенство:

$$(\det R(z))'\,|_{z=1} > 0, \quad (5.28)$$

что, в свою очередь, как было показано выше, эквивалентно неравенству

$$\mathbf{x}\, R'(1)\, \mathbf{e} > 0, \quad (5.29)$$

где $\mathbf{x}$ является единственным решением системы линейных алгебраических уравнений

$$\mathbf{x}\, R(1) = \mathbf{0},\ \mathbf{x}\mathbf{e} = 1. \quad (5.30)$$

Представляя вектор $\mathbf{x}$ в виде $\mathbf{x} = \boldsymbol{\theta} \otimes \mathbf{y}$, можно убедиться, что вектор $\mathbf{x}$ является решением системы (5.30), если вектор $\mathbf{y}$ является решением системы (5.23). Такое решение существует и единственно, так как матрица $S^{\oplus N} + \mathbf{S}_0^{\oplus N}(I_{\bar{W}M^{N-1}} \otimes \boldsymbol{\beta})$ является неприводимым инфинитезимальным генератором. □

Предположим, что условия эргодичности (5.19)-(5.20) или (5.21)-(5.23) выполняются. Тогда существуют стационарные вероятности состояний ЦМ $\xi_t,\ t \geq 0$,

$$p(i, n, \nu, m^{(1)}, \ldots, m^{(n)}) =$$

$$= \lim_{t\to\infty} P\{\, i_t = i,\, n_t = n,\, \nu_t = \nu,\, m_t^{(1)} = m^{(1)}, \ldots, m_t^{(n)} = m^{(n)}\},$$

$$i \geq 0,\ \nu = \overline{0, W},\ m^{(j)} = \overline{1, M},\ j = \overline{1, n},\ n = \overline{0, N}.$$

Введем векторы стационарных вероятностей $\mathbf{p}_i$, составленные из упорядоченных в лексикографическом порядке вероятностей $p(i, n, \nu, m^{(1)}, \ldots, m^{(n)})$, соответствующих значению i счетной компоненты, $i \geq 0$.

Известно, что векторы $\mathbf{p}_i$, $i \geq 0$, удовлетворяют уравнениям Чепмена – Колмогорова

$$(\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \ldots)Q = \mathbf{0},\ (\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \ldots)\mathbf{e} = 1, \tag{5.31}$$

где генератор Q ЦМ $\xi_t,\ t \geq 0$, задан формулами (5.15)-(5.18).

Для решения бесконечной системы уравнений (5.31) может быть использован численно устойчивый алгоритм для нахождения стационарного распределения АКТЦМ, основанный на выводе альтернативной системы уравнений и изложенный в разделе 3.7. Этот алгоритм успешно реализован в рамках пакета прикладных программ „СИРИУС++“, разработанного в НИЛ прикладного вероятностного анализа Белгосуниверситета, см., например, [115].

5.2.4 Характеристики производительности системы

Вычислив векторы $\mathbf{p}_i, i \geq 0$, можно вычислить различные характеристики производительности СМО $BMAP/PH/N$ с повторными вызовами. Приведем некоторые из них.

- Совместная вероятность того, что в произвольный момент времени на орбите находится i запросов и n приборов заняты

$$q_n(i) = [\mathbf{p}_i]_n \mathbf{e},\ n = \overline{0, N},\ i \geq 0.$$

Здесь обозначение $[\boldsymbol{x}]_n$ имеет следующий смысл. Пусть вектор $\boldsymbol{x}$ размера K представлен в виде $\boldsymbol{x} = (\boldsymbol{x}_0, \ldots, \boldsymbol{x}_N)$, где вектор $\boldsymbol{x}_n$ имеет размерность $(W+1)M^n$, $n = \overline{0, N}$. Тогда $[\boldsymbol{x}]_n \stackrel{def}{=} \boldsymbol{x}_n$. То есть, $[\mathbf{p}_i]_n$ есть часть вектора $\mathbf{p}_i$, соответствующая состояниям СМО, при которых на орбите находится i запросов и занято n приборов, $i \geq 0,\ n = \overline{0, N}$.

- Вероятность того, что в произвольный момент времени заняты n приборов, при условии, что в данный момент на орбите находится i запросов

$$q_n^{(i)} = \frac{q_n(i)}{\mathbf{p}_i\mathbf{e}},\ n = \overline{0,N}.$$

- Вероятность того, что в произвольный момент времени заняты n приборов

$$q_n = [\mathbf{p}(1)]_n\mathbf{e},\ n = \overline{0,N},\ \mathbf{P}(1) = \sum_{i=0}^{\infty}\boldsymbol{p}_i.$$

- Среднее число занятых приборов в произвольный момент времени

$$\hat{n} = \sum_{n=1}^{N} nq_n.$$

- Среднее число запросов на орбите в произвольный момент времени

$$L_{orb} = \sum_{i=1}^{\infty} i\mathbf{p}_i\mathbf{e}.$$

- Стационарное распределение числа занятых приборов в момент поступления группы, состоящей из k запросов

$$P_n^{(k)} = \frac{[\mathbf{p}(1)]_n(D_k \otimes I_{M^n})\mathbf{e}}{\boldsymbol{\theta}}D_k\mathbf{e},\ n = \overline{0,N},\ k \geq 1.$$

- Вероятность того, что произвольный запрос поступит на обслуживание сразу по поступлении в систему

$$P_{imm} = \frac{1}{\lambda}\sum_{n=1}^{N}[\mathbf{P}(1)]_{N-n}(\sum_{k=0}^{n}(k-n)D_k \otimes I_{M^{N-n}})\mathbf{e}. \qquad (5.32)$$

Доказательство. Как отмечалось в разделе 3.1, при фиксированном состоянии управляющего процесса $BMAP$-потока вероятность того, что произвольный запрос поступит в группе из k запросов, вычисляется следующим образом:

$$\frac{kD_k\mathbf{e}}{\boldsymbol{\theta}\sum\limits_{l=1}^{\infty} lD_l\mathbf{e}} = \frac{kD_k\mathbf{e}}{\lambda}.$$

Произвольный запрос поступит на обслуживание сразу по поступлении в систему, если он поступает, когда занято m приборов, $m = \overline{0, N-1}$, в группе из k запросов $k = \overline{1, N-m}$, или в группе из k запросов $k \geq N-m+1$, но он оказывается в числе первых $N-m$ запросов этой группы. Мы предполагаем, что произвольный запрос, поступивший в группе из k запросов, является l-м в группе $l, l = \overline{1, k}$, с вероятностью $\frac{1}{k}$. Используя формулу полной вероятности, получаем

$$P_{imm} = \frac{1}{\lambda}\sum_{m=0}^{N-1}[\mathbf{P}(1)]_m(\sum_{k=1}^{N-m} kD_k \otimes I_{M^m} + \sum_{k=N-m+1}^{\infty} k\frac{N-m}{k}D_k \otimes I_{M^m})\mathbf{e}.$$

Принимая во внимание формулу

$$(\sum_{k=N-m+1}^{\infty} D_k \otimes I_{M^{N-n}})\mathbf{e} = -(\sum_{k=0}^{N-m} D_k \otimes I_{M^{N-n}})\mathbf{e}$$

и меняя индекс суммирования m на $n = N-m$, получаем формулу (5.32). □

- Вероятность того, что произвольная группа запросов поступит на обслуживание сразу по поступлении в систему

$$P^b_{imm} = \frac{1}{\lambda_b}\sum_{m=1}^{N}[\mathbf{P}(1)]_{N-m}\sum_{k=1}^{m}(D_k \otimes I_{M^{N-m}})\mathbf{e},$$

где λ_b есть средняя скорость поступления групп, $\lambda_b = -\boldsymbol{\theta} D_0 \mathbf{e}$.

5.2.5 Случай ненастойчивых запросов

Выше мы предположили, что запрос, попавший на орбиту, является абсолютно настойчивым, то есть он повторяет попытки попасть на обслуживание до тех пор, пока не застанет свободный прибор. Сейчас рассмотрим более общий случай, когда запрос, совершивший повторную попытку и заставший все приборы занятыми, с вероятностью p возвращается на орбиту, а с вероятностью $1-p$ $(0 \leq p \leq 1)$ уходит из системы навсегда (считается потерянным).

В этом случае компоненты ЦМ ξ_t, $t \geq 0$, описывающей функционирование системы, имеют тот же смысл, что и соответствующие компоненты ЦМ, описывающей СМО с абсолютно настойчивыми запросами, а инфинитезимальный генератор также имеет структуру вида (5.15). При этом

блоки $Q_{i,i+k},\ k \geq 1$, имеют вид (5.17), блоки $Q_{i,i-1}$ получаются путем замены в (5.16) нулевой матрицы в последней строке и столбце на матрицу $(1-p)I_{\bar{W}M^N}$, а блоки $Q_{i,i}$ – путем замены в (5.18) матрицы $\alpha_i(1-\delta_{n,N})I_{\bar{W}M^n}$ на матрицу $\alpha_i(1-p\delta_{n,N})I_{\bar{W}M^n}$.

Существенное отличие систем с абсолютно настойчивыми и ненастойчивыми запросами состоит в следующем. Если вероятность $p = 1$, то есть запросы абсолютно настойчивы, система не всегда является эргодичной. Достаточные условия эргодичности даны в теоремах 5.4 и 5.5. Если же $p < 1$, система является эргодичной при любых ее параметрах. Это легко проверяется с помощью результатов для АКТЦМ, изложенных в разделе 3.7. При этом предельная матрица Y_0 в определении АКТЦМ является стохастической, а все матрицы $Y_k,\ k \geq 1$, – нулевыми. Следовательно, $Y'(1) = O$ и неравенство (3.7.7) выполняется автоматически при любых параметрах системы.

Другое отличие системы с неабсолютно настойчивыми запросами состоит в возможности потери запросов. Поэтому одной из важных характеристик производительности системы является вероятность потери произвольного запроса. Она вычисляется следующим образом:

$$P_{loss} = 1 - \frac{1}{\lambda}\sum_{n=1}^{N}[\mathbf{P}(1)]_n\mathbf{S}_0^{\oplus n}\mathbf{e}. \tag{5.33}$$

Дадим краткий вывод формулы (5.33). В стационарном режиме функционирования системы средняя скорость окончания обслуживания запросов в системе задается как

$$\sum_{n=1}^{N}[\mathbf{P}(1)]_n\boldsymbol{S}_0^{\oplus n}\mathbf{e}.$$

Величина λ есть средняя скорость поступления запросов. Отношение этих двух скоростей задает вероятность того, что произвольный запрос не будет потерян. Дополнительная к ней вероятность задает вероятность потери произвольного запроса.

5.2.6 Численные результаты

Для численной иллюстрации полученных результатов в данном подразделе приведем графики, показывающие зависимость среднего числа запро-

сов на орбите L_{orb}, вероятности потери произвольного запроса P_{loss} и вероятности того, что произвольный запрос посетит орбиту $P_{orb} = 1 - P_{imm}$ от вероятности p возвращения на орбиту после неудачной попытки, а также от загрузки системы, коэффициентов корреляции и вариации интервалов между моментами поступления запросов.

Будем рассматривать четыре различных $BMAP$-потока, имеющих среднюю скорость поступления запросов $\lambda = 12$. Первый из этих потоков будем обозначать Exp. Он является стационарным пуассоновским с параметром $\lambda = 12$. Коэффициенты корреляции и вариации интервалов между моментами поступления запросов равны, соответственно, 0 и 1. Другие три $BMAP$-потока имеют коэффициент вариации интервалов между моментами поступления, равный 2. Эти потоки заданы матрицами $D_0 = \bar{c}\hat{D}_0$ и D_k, $k = \overline{1,4}$, которые определены как $D_k = \bar{c}Dq^{k-1}(1-q)/(1-q^4)$, $k = \overline{1,4}$, где $q = 0,8$ и D – подобранная соответствующим образом матрица, а скалярный коэффициент $\bar{c}$ взят равным $2,4$, чтобы обеспечить среднюю скорость поступления запросов $\lambda = 12$.

$BMAP$-поток, обозначаемый как $BMAP_1$, определен матрицами

$$\hat{D}_0 = \begin{pmatrix} -13,334625 & 0,588578 & 0,617293 \\ 0,692663 & -2,446573 & 0,422942 \\ 0,682252 & 0,414363 & -1,635426 \end{pmatrix},$$

$$D = \begin{pmatrix} 11,546944 & 0,363141 & 0,218669 \\ 0,384249 & 0,865869 & 0,080851 \\ 0,285172 & 0,04255 & 0,211089 \end{pmatrix}.$$

Коэффициент корреляции длин соседних интервалов между моментами поступления групп запросов в этом потоке равен 0,1.

$BMAP$-поток, обозначаемый как $BMAP_2$, имеет коэффициент корреляции, равный 0,2, и задан матрицами

$$\hat{D}_0 = \begin{pmatrix} -15,732675 & 0,606178 & 0,592394 \\ 0,517817 & -2,289674 & 0,467885 \\ 0,597058 & 0,565264 & -1,959665 \end{pmatrix},$$

$$D = \begin{pmatrix} 14,1502 & 0,302098 & 0,081805 \\ 0,107066 & 1,03228 & 0,164627 \\ 0,08583 & 0,197946 & 0,513566 \end{pmatrix}.$$

$BMAP$-поток, обозначаемый как $BMAP_3$, имеет коэффициент корреляции, равный 0,3, и задан матрицами

$$\hat{D}_0 = \begin{pmatrix} -25,539839 & 0,393329 & 0,361199 \\ 0,145150 & -2,232200 & 0,200007 \\ 0,295960 & 0,387445 & -1,752617 \end{pmatrix},$$

$$D = \begin{pmatrix} 24,242120 & 0,466868 & 0,076323 \\ 0,034097 & 1,666864 & 0,186082 \\ 0,009046 & 0,255481 & 0,804685 \end{pmatrix}.$$

Также рассмотрим три $BMAP$-потока, имеющих среднюю скорость поступления запросов $\lambda = 12$, коэффициент корреляции соседних интервалов между моментами поступления, равный 0,3, но различный коэффициент вариации интервалов между моментами поступления. Эти $BMAP$-потоки заданы матрицами D_k, $k = \overline{0,4}$, полученными по матрицам $\hat{D}_0$ и D тем же путем, как описано выше.

$BMAP$-поток, обозначаемый как $BMAP^2$ и имеющий коэффициент вариации, равный 2, идентичен потоку $BMAP_3$, введенному выше.

$BMAP$-поток, обозначаемый как $BMAP^5$ и имеющий коэффициент вариации, равный $4,68$, задается матрицами

$$\hat{D}_0 = \begin{pmatrix} -16,196755 & 0,090698 & 0,090698 \\ 0,090698 & -0,545154 & 0,090698 \\ 0,090698 & 0,090699 & -0,313674 \end{pmatrix},$$

$$D = \begin{pmatrix} 15,949221 & 0,066138 & 0 \\ 0,033069 & 0,297622 & 0,033069 \\ 0 & 0,013228 & 0,119049 \end{pmatrix}.$$

$BMAP$-поток, обозначаемый как $BMAP^7$ и имеющий коэффициент вариации, равный $7,34404$, задается матрицами

$$\hat{D}_0 = \begin{pmatrix} -16,268123 & 0,040591 & 0,040591 \\ 0,040591 & -0,223595 & 0,040591 \\ 0,040591 & 0,040591 & -0,132969 \end{pmatrix},$$

$$D = \begin{pmatrix} 16,161048 & 0,025893 & 0 \\ 0,012947 & 0,116519 & 0,012947 \\ 0 & 0,005179 & 0,046608 \end{pmatrix}.$$

Считаем, что число приборов N в рассматриваемой СМО равно 3.

Будем рассматривать три различные PH-распределения времени обслуживания, имеющие один и тот же коэффициент вариации $1,204$, но различные интенсивности. Эти PH-распределения определены вектором $\boldsymbol{\beta} = (0,5; 0,5)$ и матрицами S, заданными через матрицу $S^{(0)} = \begin{pmatrix} -4 & 2 \\ 0 & -1 \end{pmatrix}$ следующим образом.

PH распределение, обозначаемое как PH_2, задается матрицей $S = 17,5S^{(0)}$. При этом интенсивность обслуживания μ равна 20, а коэффициент загрузки системы $\rho = \frac{\lambda}{N\mu}$ равен 0,2.

PH распределение, обозначаемое как PH_5, задается матрицей $S = 7S^{(0)}$. Здесь $\mu = 8$, $\rho = 0,5$.

PH распределение, обозначаемое как PH_8, задается матрицей $S = 4,375S^{(0)}$. Здесь $\mu = 5$, $\rho = 0,8$.

Рисунки 5.1-5.3 иллюстрируют зависимость вероятностей P_{loss}, P_{orb} и среднего числа запросов на орбите L_{orb} от вероятности p возвращения на орбиту после неудачной попытки и от загрузки системы при потоке $BMAP_2$.

Рисунки 5.4-5.5 иллюстрируют зависимость вероятности P_{loss} и среднего числа запросов на орбите L_{orb} от вероятности p возвращения на орбиту после неудачной попытки и корреляции $BMAP$-потока при распределении времени обслуживания PH_8.

Из рисунков 5.4-5.5 можно сделать вывод, что корреляция во входном потоке существенно влияет на характеристики производительности системы и эти характеристики ухудшаются с ростом корреляции.

Рисунки 5.6-5.7 иллюстрируют зависимость вероятности P_{loss} и среднего числа запросов на орбите L_{orb} от вероятности p возвращения на орбиту после неудачной попытки и вариации $BMAP$-потока при распределении времени обслуживания PH_8.

Рисунки 5.6-5.7 показывают, что вариация во входном потоке также существенно влияет на характеристики производительности системы, и эти характеристики ухудшаются с ростом вариации.

Таким образом, из проделанных численных экспериментов можно сделать вывод о том, что аппроксимация реального потока стационарным пуассоновским потоком может привести к слишком оптимистическому предсказанию характеристик производительности системы, если коэффициент корреляции интервалов между моментами поступления не близок к нулю, а коэффициент вариации не близок к единице.

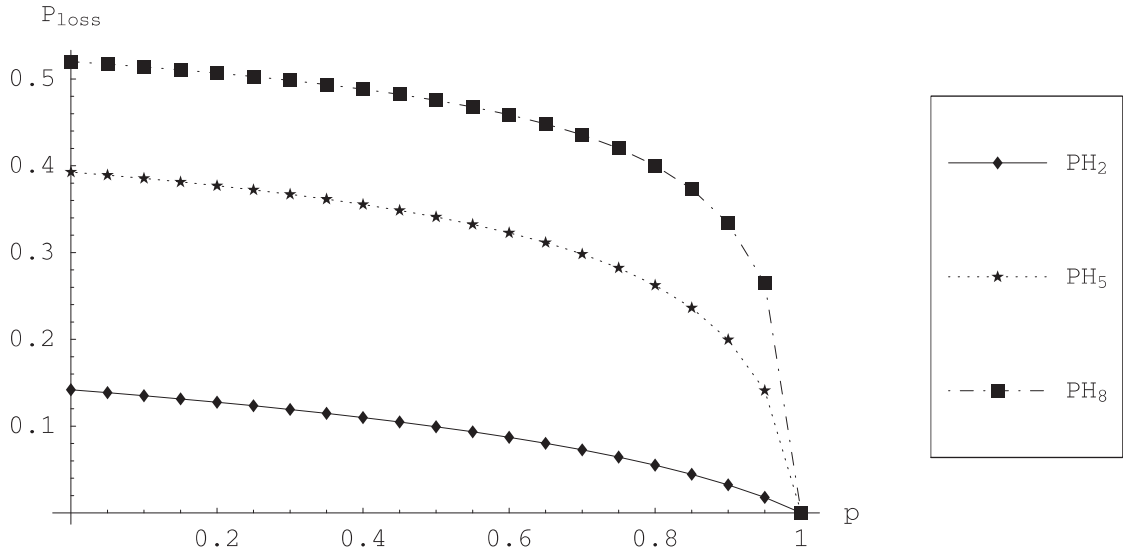

Рисунок 51: Вероятность P_{loss} как функция от вероятности p при различной загрузке системы

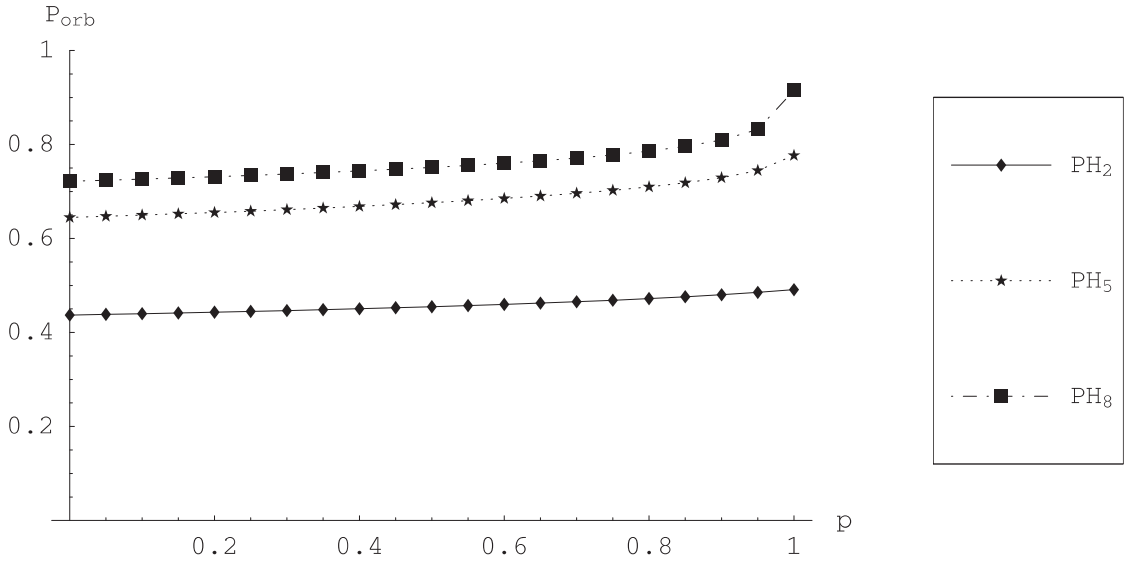

Рисунок 52: Вероятность P_{orb} как функция от вероятности p при различной загрузке системы

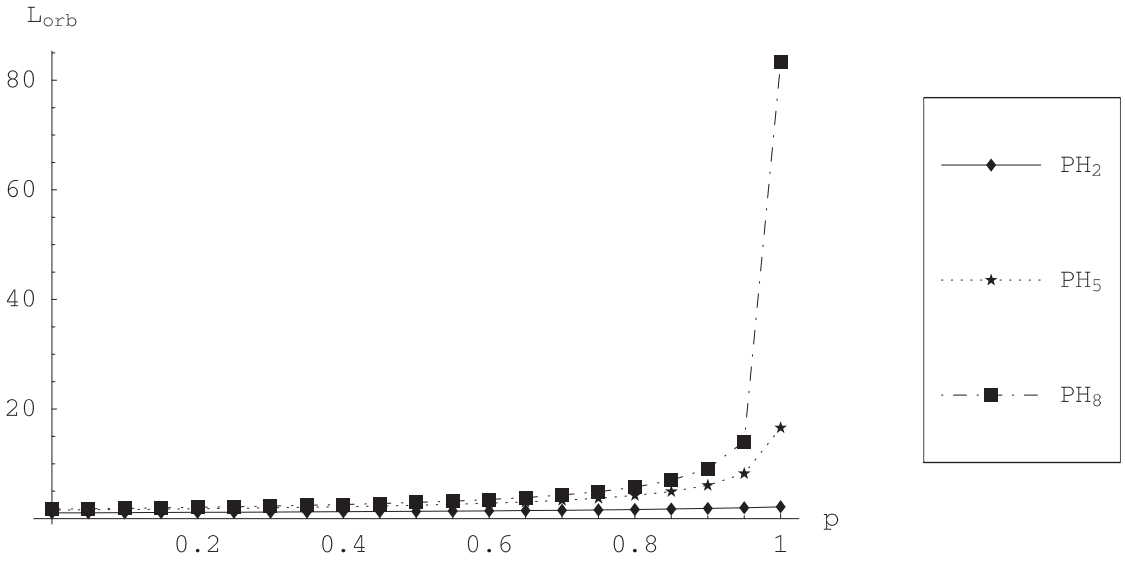

Рисунок 53: Среднее число запросов на орбите L_{orb} как функция от вероятности p при различной загрузке системы

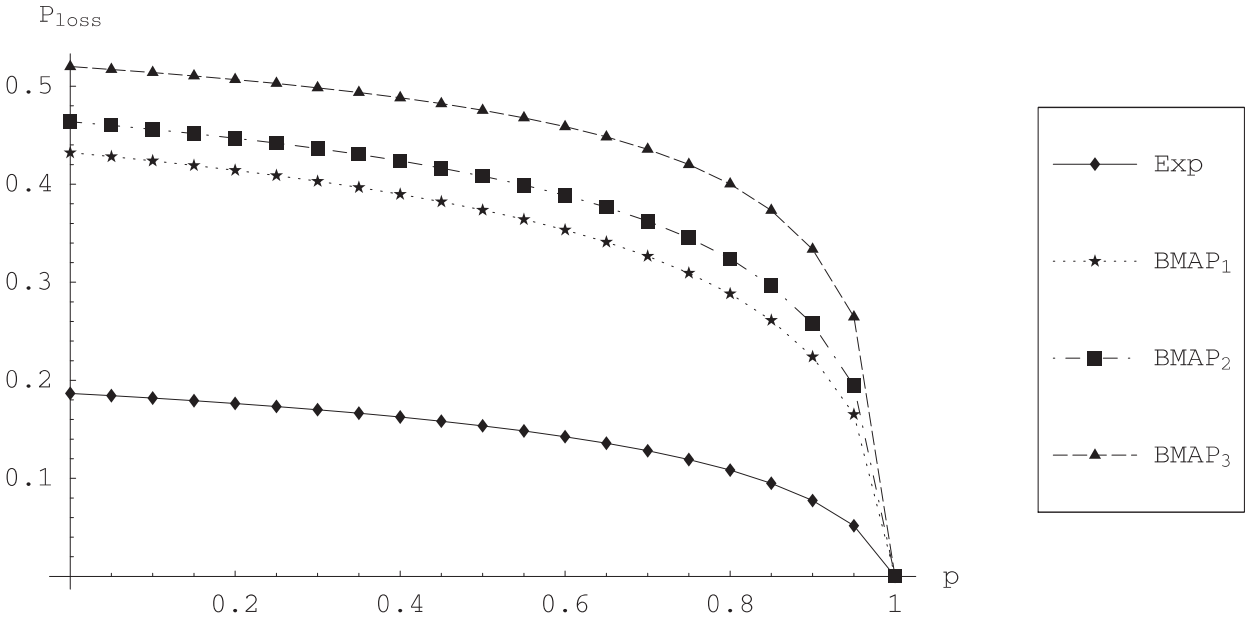

Рисунок 54: Вероятность P_{loss} как функция от вероятности p для потоков с разной корреляцией

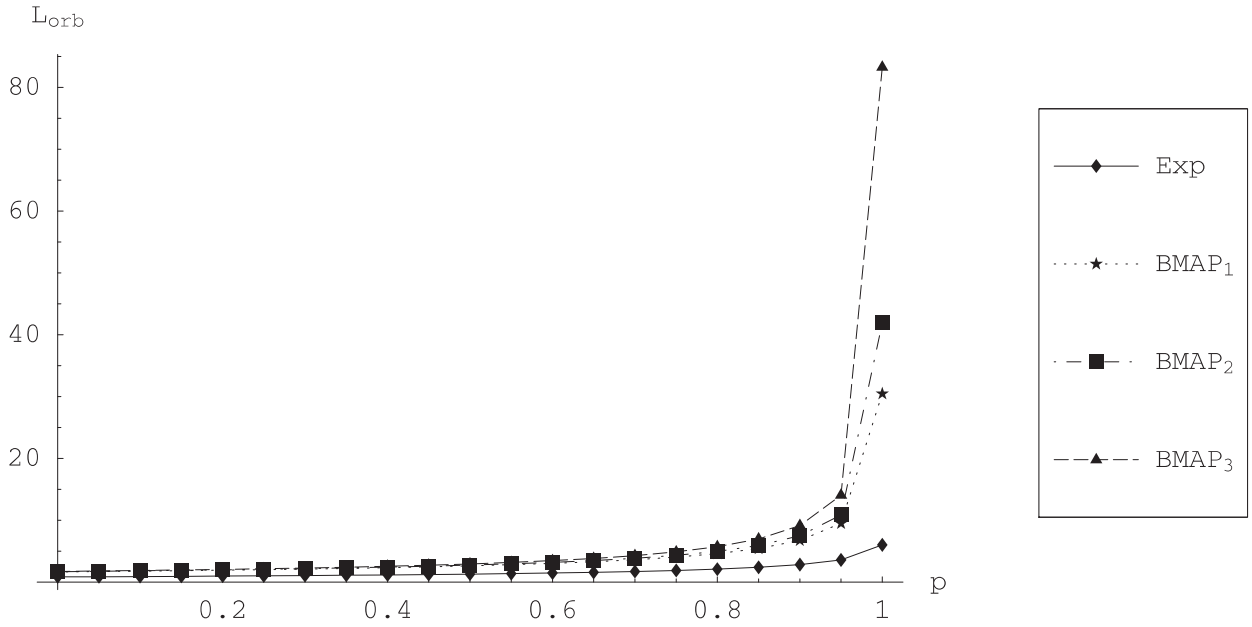

Рисунок 55: Среднее число запросов на орбите L_{orb} как функция от вероятности p для потоков с разной корреляцией

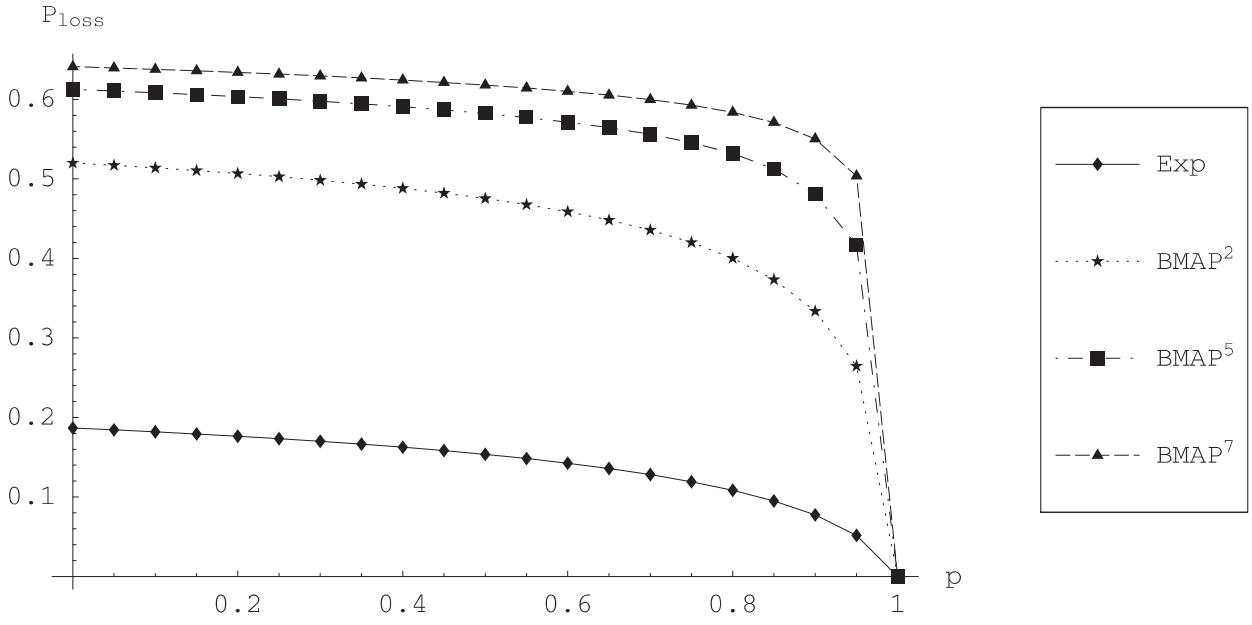

Рисунок 56: Вероятность P_{loss} как функция от вероятности p для потоков с разной вариацией

Проиллюстрируем возможное применение полученных результатов для решения несложной задачи оптимизации. Предположим, что система несет потери из-за ухода необслуженных запросов из системы и из-за долгого пребывания запросов на орбите. Взвешенный критерий качества работы системы имеет вид

$$I = c_1 \lambda P_{loss} + c_2 L_{orb}. \tag{5.34}$$

Здесь c_1 и c_2 – заданные стоимостные коэффициенты. Коэффициент c_1 имеет смысл штрафа за досрочный уход одного запроса, а c_2 – штраф за пребывание запроса на орбите в единицу времени.

Предположим, что имеется возможность выбора вероятности p возвращения запроса на орбиту после неудачной попытки попасть на обслуживание.

Рисунок 5.8 иллюстрирует зависимость критерия качества (5.34) от вероятности p при различных соотношениях между стоимостными коэффициентами c_1 и c_2. Входной поток описывается как $BMAP_3$, а процесс обслуживания – распределением PH_5.

Из рисунка 5.8 видно, что при большом штрафе за досрочный уход запросов управление параметром p может дать существенный выигрыш в значении критерия качества (5.34).

Рассмотрим еще один численный пример. Предположим, что модель $BMAP/PH/N$ применяется для принятия решения о выборе числа каналов, обеспечивающих мобильную связь, например в терминале аэропорта. Один физический канал можно использовать для организации восьми логических каналов. Один-два логических канала из каждого физического канала резервируются для управления работой системы. Поэтому при числе физических каналов, равном 1, 2 и 3, число логических каналов, доступных пользователям сети мобильной связи, равно $N = 7, 14$ и 22 соответственно. Рассчитаем основные характеристики системы при выборе числа физических каналов, равном 1, 2 или 3, и различных предположениях о характере входного потока в систему. Чтобы снизить размерность блоков генератора, предположим здесь, что время обслуживания запросов распределено по показательному закону с параметром (интенсивностью) $\mu = 10$. Предположим также, что стратегия повторов с орбиты – классическая, и каждый запрос, находящийся на орбите, генерирует повторные попытки через интервалы времени, имеющие показательное распределение с параметром (интенсивностью) $\alpha = 20$.

Будем рассчитывать следующие характеристики производительности

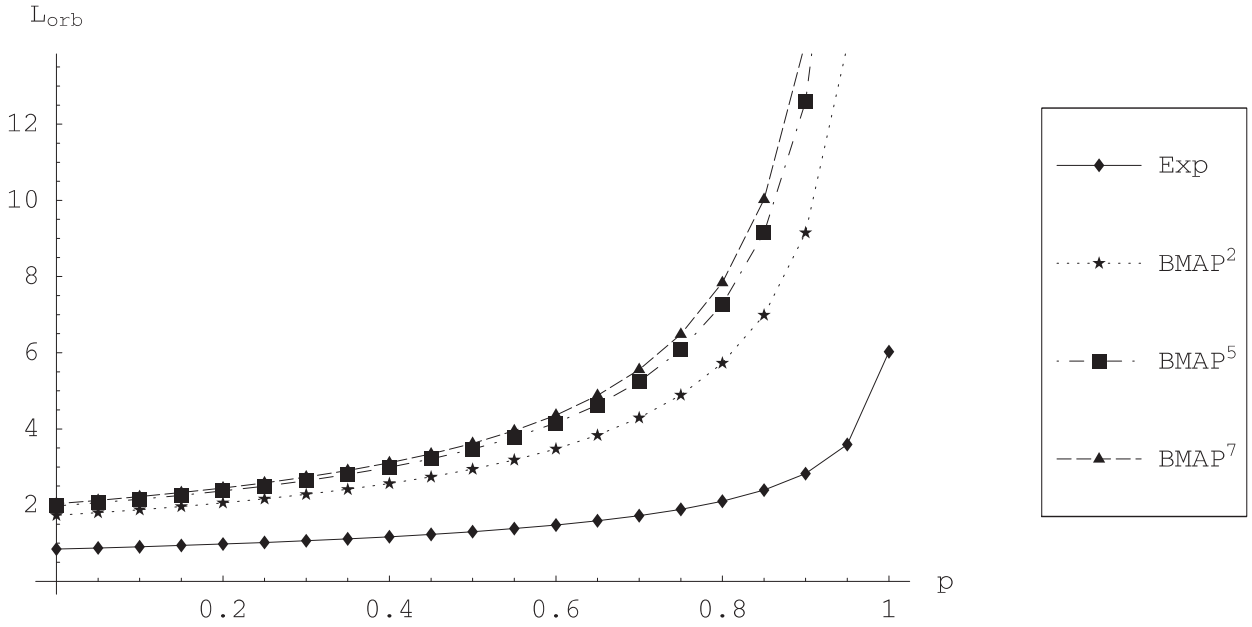

Рисунок 57: Среднее число запросов на орбите L_{orb} как функция от вероятности p для потоков с разной вариацией

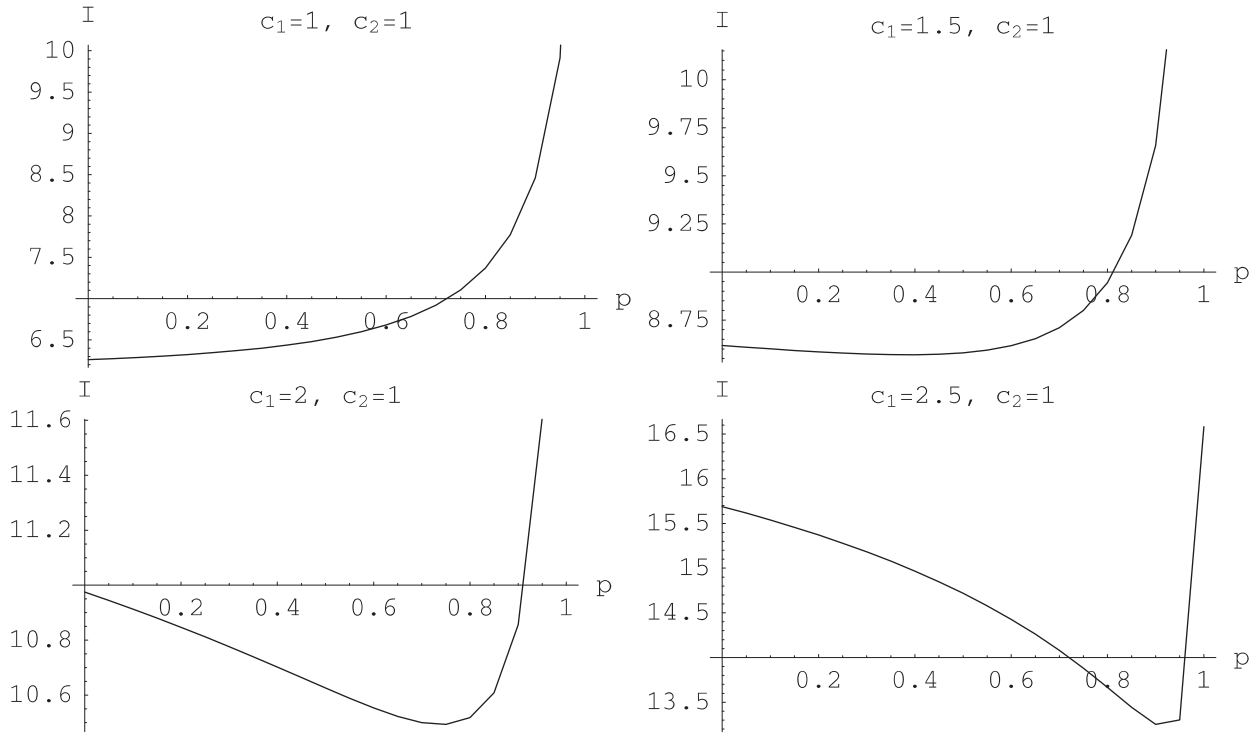

Рисунок 58: Критерий качества (5.34) как функция от вероятности p при различных стоимостных коэффициентах c_1 и c_2

системы: среднее число запросов на орбите L_{orb}, вероятность немедленного доступа на обслуживание P_{imm} и вероятность того, что орбита пуста P_0. Выше был проиллюстрирован тот факт, что при одной и той же средней скорости поступления запросов характеристики прозводительности системы могут существенно зависеть от корреляции соседних интервалов между моментами поступления запросов и вариации интервалов между моментами поступления запросов. Поэтому будем рассматривать пять различных $BMAP$-потоков. Все они имеют одну и ту же среднюю скорость поступления запросов $\lambda = 27,5$. Первые четыре потока имеют коэффициент вариации длин интервалов, равный $12,2733$, а их коэффициенты корреляции равны $c_{cor} = 0,1; 0,2; 0,3; 0,4$.

Первый из этих $BMAP$-потоков (будем обозначать его как $BMAP_1$) имеет коэффициент корреляции, равный 0,1, и задается матрицами

$$D_0 = \begin{pmatrix} -5,874 & 0 \\ 0 & -0,129 \end{pmatrix} \quad \text{и} \quad D_k = \begin{pmatrix} 0,586 & 0,002 \\ 0,010 & 0,003 \end{pmatrix}, \ k = 1, ..., 10.$$

Второй $BMAP$-поток (будем обозначать его как $BMAP_2$) имеет коэффициент корреляции, равный 0,2, и задается матрицами

$$D_0 = \begin{pmatrix} -6,745 & 0 \\ 0 & -0,219 \end{pmatrix} \quad \text{и} \quad D_k = \begin{pmatrix} 0,670 & 0,004 \\ 0,012 & 0,010 \end{pmatrix}, \ k = 1, ..., 10.$$

Третий $BMAP$-поток ($BMAP_3$) имеет коэффициент корреляции, равный 0,3, и задается матрицами

$$D_0 = \begin{pmatrix} -8,771 & 0 \\ 0 & -0,354 \end{pmatrix} \quad \text{и} \quad D_k = \begin{pmatrix} 0,867 & 0,010 \\ 0,012 & 0,024 \end{pmatrix}, \ k = 1, ..., 10.$$

Четвертый $BMAP$-поток ($BMAP_4$) имеет коэффициент корреляции, равный 0,4, и задается матрицами

$$D_0 = \begin{pmatrix} -17,231 & 0 \\ 0,002 & -0,559 \end{pmatrix} \quad \text{и} \quad D_k = \begin{pmatrix} 1,705 & 0,018 \\ 0,006 & 0,049 \end{pmatrix}, \ k = 1, ..., 10.$$

Пятый $BMAP$-поток является $BIPP$ (Batch Interrupted Poisson Process) потоком. Он задается матрицами

$$D_0 = \begin{pmatrix} -1 & 1 \\ 0 & -10 \end{pmatrix} \quad \text{и} \quad D_k = \begin{pmatrix} 0 & 0 \\ 0,1 & 0,9 \end{pmatrix}, \ k = 1, ..., 10.$$

Этот поток имеет нулевую корреляцию, но характеризуется нерегулярностью, заключающейся в том, что промежутки времени, когда запросы поступают интенсивно, перемежаются промежутками, когда запросы не приходят вообще.

Приведем три таблицы, характеризующие величины L_{orb}, P_{imm} и P_0 при различных входных потоках и числе логических каналов, равном $N = 7, 14$ и 22 соответственно.

Таблица 5.1: Зависимость среднего числа запросов на орбите L_{orb} от числа приборов N и поступающего потока

L_{orb}	$BIPP$	$BMAP_1$	$BMAP_2$	$BMAP_3$	$BMAP_4$
$N = 7$	$3,1$	$1,7$	$2,1$	$3,7$	$47,9$
$N = 14$	$0,3$	$0,1$	$0,2$	$0,3$	$1,3$
$N = 22$	$0,02$	0	0	0	$0,2$

Таблица 5.2: Зависимость вероятности P_{imm} немедленного поступления на обслуживание от числа приборов N и поступающего потока

P_{imm}	$BIPP$	$BMAP_1$	$BMAP_2$	$BMAP_3$	$BMAP_4$
$N = 7$	$0,39$	$0,53$	$0,48$	$0,35$	$0,09$
$N = 14$	$0,87$	$0,94$	$0,92$	$0,87$	$0,59$
$N = 22$	$0,99$	1	$0,99$	$0,99$	$0,92$

Таблица 5.3: Зависимость вероятности P_0 того, что орбита пуста в произвольный момент времени, от числа приборов N и поступающего потока

P_0	$BIPP$	$BMAP_1$	$BMAP_2$	$BMAP_3$	$BMAP_4$
$N = 7$	$0,64$	$0,67$	$0,66$	$0,63$	$0,64$
$N = 14$	$0,92$	$0,95$	$0,95$	$0,92$	$0,85$
$N = 22$	$0,99$	1	1	$0,99$	$0,96$

Основываясь на результатах расчетов, приведенных в таблицах 5.1-5.3, можно сделать следующие выводы:

• корреляция имеет существенное влияние на характеристики функционирования системы. Чем больше корреляция, тем хуже характеристики. Поэтому при исследовании реальных систем необходимо строить адекватную модель входного потока, учитывающую наличие или отсутствие корреляции в потоке. Это поможет избежать ошибок в проектировании.

Например, если при проектировании была поставлена цель сделать вероятность немедленного доступа произвольного пользователя к ресурсам мобильной сети большей, чем 0,9, то при корреляции потока порядка 0.1-0.2 достаточно организовать 2 физических канала (14 логических каналов). А при корреляции порядка 0,3 и выше требуется организовать 3 физических канала. Если же требуемая вероятность немедленного доступа произвольного пользователя к ресурсам мобильной сети порядка 0,9, то при корреляции потока 0,4 и выше даже 3 физических канала не могут обеспечить требуемое качество обслуживания.

• наряду с корреляцией, нерегулярность поступления запросов в потоке (в англоязычной литературе это называется словом burstyness), характерная, например, для узлов, обеспечивающих мобильную связь в терминалах транспортных систем, также негативно влияет на качество обслуживания пользователей.

Интересно также исследовать эффект группового поступления запросов в потоке. Рассмотрим такую характеристику производительности системы как $P_{imm}^{(k)}$ – вероятность того, что произвольный запрос по поступлению в систему немедленно пойдет на обслуживание, если он поступил в группе, состоящей из k, $k \geq 1$, запросов. Несложно убедиться, что эта вероятность вычисляется следующим образом:

$$P_{imm}^{(k)} = \sum_{m=0}^{N-k} P_m^{(k)} + \sum_{m=\max\{0,N-k+1\}}^{N-1} \frac{N-m}{k} P_m^{(k)}, \quad k \geq 1,$$

где $P_m^{(k)}$ есть вероятность того, что в произвольный момент поступления группы из k, $k \geq 1$, запросов в системе занято m приборов $m = \overline{0, N}$,

$$P_m^{(k)} = \frac{[\mathbf{P}(1)]_m (D_k \otimes I_{M^m})\mathbf{e}}{\boldsymbol{\theta} D_k \mathbf{e}}, \quad m = \overline{0, N}, k \geq 1.$$

Результаты вычисления вероятности $P_{imm}^{(k)}$ для входных данных, описанных в предыдущем эксперименте (в качестве входного потока взят $BMAP_1$), как функции числа приборов и размера группы приведены на рисунке 5.5.

Из рисунка 5.9 видно, что при небольшом числе приборов значения вероятности $P_{imm}^{(k)}$ сильно зависят от k. Так, при $N = 7$ и $k = 1$ эта вероятность близка к 0,8. Усредненная по размеру группы вероятность (соответствующая линия на рисунке нарисована жирно и помечена буквой а) равна

приблизительно 0,55. А при $k = 10$ эта вероятность меньше 0,4. С ростом числа приборов N эта разница начинает нивелироваться. Тем не менее эффект группового поступления также следует принимать во внимание при вынесении проектных решений.

5.3 Система $BMAP/PH/N$ с повторными вызовами в случае распределения фазового типа времени обслуживания и большого числа приборов

5.3.1 Выбор ЦМ для анализа рассматриваемой системы

При построении ЦМ $\xi_t,\ t \geq 0$, описывающей функционирование системы $BMAP/PH/N$ с повторными вызовами, важную роль играет способ учета состояний цепей Маркова, управляющих процессами обслуживания фазового типа на приборах системы. С целью большей наглядности и простоты изложения при введении ЦМ $\xi_t,\ t \geq 0$, нами использовался способ учета, предусматривающий динамическую нумерацию приборов. Номера назначаются только занятым приборам. Эти приборы нумеруются в порядке их занятия, то есть прибор, начинающий обслуживание, когда занято n приборов, получает номер $n + 1$; при окончании обслуживания на каком-либо из приборов и отсутствии очереди последний теряет свой номер, а оставшиеся приборы (имевшие большие номера) соответственно перенумеровываются. Компонентами ЦМ $\xi_t,\ t \geq 0$, являются состояния цепей Маркова, управляющих процессами обслуживания на занятых в настоящий момент времени приборах. Этот способ учета управляющих процессов ослуживания достаточно естественный и удобный. При таком способе размерность блоков генератора ЦМ $\xi_t,\ t \geq 0$, равна $K = \bar{W}\frac{M^N-1}{M-1}$.

Другой возможный простой способ учета состоит в том, что для каждого из приборов системы, независимо от того, занят ли он или свободен, отслеживается его состояние. Это состояние предполагается равным m, если состояние ЦМ, управляющей процессом обслуживания на этом приборе равно $m,\ m = \overline{1, M}$. Если же в данный момент времени этот прибор свободен, в качестве состояния ЦМ, управляющей процессом обслуживания, можно искусственно взять, например, 0 или $M + 1$. При этом способе описания нет необходимости в постоянной перенумерации приборов, но необходимо проводить рандомизированный выбор прибора, на который пойдет на обслуживание запрос, в момент поступления которого есть несколько

свободных приборов. При этом способе учета размерность блоков генератора ЦМ равна $K_0 = \bar{W}(M+1)^N$. Очевидно, что $K_0 > K$. Например, если N равно 5, а пространства состояний цепей Маркова, управляющих поступлением и обслуживание запросов, состоят только из двух элементов ($W = 1$, $M = 2$), то $K = 126$, а $K_0 = 486$.

Существует, однако, еще один способ отслеживания состояний цепей Маркова, управляющих процессами обслуживания фазового типа на приборах системы. Этот способ особенно эффективен по сравнению с двумя вышеупомянутыми в ситуациях, когда число приборов N велико, а число M состояний ЦМ, управляющей процессом обслуживания, мало. Идея этого способа проста: вместо отслеживания текущей фазы обслуживания на каждом из приборов отслеживать, сколько приборов в данный момент находятся на той или иной фазе обслуживания. То есть для каждого возможного состояния m ЦМ, управляющей процессом обслуживания, учитывать, сколько приборов в данный момент находятся в данном состоянии, $m = \overline{1, M}$. При таком способе учета размерность блоков генератора ЦМ $\xi_t,\ t \geq 0$, описывающей процесс функционирования системы $BMAP/PH/N$ с повторными вызовами, равна $K_1 = \bar{W}\binom{N+M}{M}$. Нетрудно видеть, что $K_1 < K$. Например, в упомянутом выше примере, где $K = 126$, имеем $K_1 = 42$.

Приведем кратко результаты исследования ЦМ, описывающей рассматриваемую СМО $BMAP/PH/N$ с повторными вызовами и построенной путем использования только что описанного способа учета текущих фаз обслуживания.

Вместо ЦМ

$$\xi_t = (i_t,\, n_t,\, \nu_t,\, m_t^{(1)}, \ldots, m_t^{(n_t)}),\ t \geq 0,$$

рассмотренной в подразделе 5.2, будем рассматривать многомерную ЦМ

$$\zeta_t = \{i_t, n_t, \nu_t, h_t^{(1)}, \ldots, h_t^{(M)}\},\ t \geqslant 0,$$

где компоненты i_t, n_t, ν_t имеют тот же смысл, что и соответствующие компоненты ЦМ ξ_t, $t \geq 0$, а $h_t^{(m)}$ есть число приборов, на которых в момент t процесс обслуживания находится на фазе m. Очевидно, что $h_t^{(m)} \in \{0, \ldots, n_t\}$, $m = \overline{1, M}$, $\sum\limits_{m=1}^{M} h_t^{(m)} = n_t$, $n_t = \overline{0, N}$.

Следуя [20], [167], при формировании уровней будем перенумеровывать состояния, входящие в уровень i в лексикографическом порядке значений

компонент n_t, ν_t и обратном лексикографическом порядке значений компонент $h_t^{(m)},\ m = \overline{1, M}$.

При записи инфинитезимального генератора ЦМ ζ_t будут использоваться матрицы, которые были введены в [20], [167] для описания интенсивностей переходов N параллельно существующих ЦМ. Приведем перечень этих матриц с описанием их вероятностного смысла.

Матрицы $L_{N-n}(N, \tilde{S})$ описывают интенсивности переходов компонент $\{h_t^{(1)}, \ldots, h_t^{(M)}\}$ ЦМ ζ_t, которые происходят при окончании обслуживания в одном из n занятых приборов (уменьшение на единицу значения компоненты, соответствующей состоянию, из которого произошел переход в поглощающее состояние ЦМ, описывающей процесс обслуживания в одном из приборов системы).

Матрицы $A_n(N, S)$ описывают интенсивности переходов компонент $\{h_t^{(1)}, \ldots, h_t^{(M)}\}$ цепи Маркова ζ_t, которые происходят при переходе одной из этих компонент из одного состояния в другое непоглощающее состояние, что не влечет изменения числа занятых приборов в системе.

Матрицы $P_n(\boldsymbol{\beta})$ описывают вероятности переходов компонент $\{h_t^{(1)}, \ldots, h_t^{(M)}\}$ ЦМ ζ_t, которые происходят при инициализации процесса обслуживания в приборе системы, когда число занятых приборов было равно n.

Подробное описание матриц $P_n(\boldsymbol{\beta}),\ A_n(N, S),\ L_{N-n}(N, \tilde{S})$ и рекурсивные алгоритмы для их вычисления приведены в статьях [20], [167], где эти алгоритмы были предложены, а также в статье [145], где они описаны в более понятном виде, и, для полноты изложения, в конце данного подраздела.

5.3.2 Генератор выбранной ЦМ и условие ее эргодичности

Лемма 5.3. *Инфинитезимальный генератор ЦМ ζ_t имеет блочную структуру вида (5.1), где блоки $Q_{i,j}$ имеют следующий вид:*

$$(Q_{i,i})_{n,n'} =$$

$$= \begin{cases} I_{\bar{W}} \otimes L_{N-n}(N, \tilde{S}), & n' = n-1, \\ D_0 \oplus A_n(N, S) + I_{\bar{W}} \otimes \triangle^{(n)} - i\alpha(1 - \delta_{n,N}) I_{\bar{W}K(n)}, & n' = n, \\ D_{n'-n} \otimes P_{n,n'}(\boldsymbol{\beta}), & n' = \overline{n+1, N}, \end{cases}$$

$$(Q_{i,i-1})_{n,n'} = i\alpha I_{\bar{W}} \otimes P_{n,n'}(\boldsymbol{\beta}), n' = n+1, n = \overline{0, N-1},$$

$$(Q_{i,i+k})_{n,n'} = \begin{cases} D_{N+k-n} \otimes P_{n,N}(\boldsymbol{\beta}), & n' = N, \quad n = \overline{0, N-1}, \\ D_k \otimes I_{K(N)}, & n' = n = N, \end{cases} \quad k \geqslant 1,$$

где

- $\triangle^{(n)},\ n = \overline{0, N}$, – *диагональные матрицы, которые обеспечивают выполнение равенств* $A\mathbf{e} = \mathbf{0}$,
- $\tilde{S} = \begin{pmatrix} \mathbf{0} & O \\ \mathbf{S_0} & S \end{pmatrix}$,
- $K(n) = \binom{n+M-1}{M-1},\ n = \overline{0, N}$,
- $P_{n,n'}(\boldsymbol{\beta}) = P_n(\boldsymbol{\beta})P_{n+1}(\boldsymbol{\beta}) \ldots P_{n'-1}(\boldsymbol{\beta}),\ 0 \leqslant n < n' \leqslant N$.

Справедливо следующее утверждение.

Теорема 5.6. *В случае бесконечно возрастающей интенсивности повторов достаточное условие эргодичности СМО* $BMAP/PH/N$ *с повторными вызовами имеет вид (5.7), где* λ *– средняя скорость* $BMAP$*-потока, а величина* $\bar{\mu}$ *задается формулой*

$$\bar{\mu} = \mathbf{y} L_0(N, \tilde{S})\mathbf{e}, \tag{5.35}$$

где вектор $\mathbf{y}$ *является единственным решением системы линейных алгебраических уравнений*

$$\mathbf{y}\Big(A_N(N, S) + \triangle^{(N)} + L_0(N, \tilde{S})P_{N-1,N}(\boldsymbol{\beta})\Big) = \mathbf{0},\ \mathbf{ye} = 1. \tag{5.36}$$

Доказательство теоремы проводится с использованием результатов, приведенных в разделе 3.7, и аналогично доказательству теоремы 5.2.

Замечание 5.1. Поскольку неравенство (5.15) с величиной $\bar{\mu}$, заданной формулами (5.22)-(5.23), задает то же условие, что и неравенство с величиной $\bar{\mu}$, заданной формулами (5.35)-(5.36), то разумно предположить, что величины $\bar{\mu}$, заданные формулами (5.22)-(5.23) и (5.35)-(5.36), совпадают. Аналитически показать этот факт не удается. Но компьютерная реализация этих формул при различных параметрах входного потока, процесса обслуживания и различном числе приборов показывает, что эти величины действительно совпадают. Более того, они совпадают с интуитивно ожидаемой величиной $\bar{\mu} = N\mu$.

5.3.3 Численные примеры

Поскольку ЦМ $\zeta_t,\ t \geq 0$, является АКТЦМ, ее стационарное распределение также находится с использованием алгоритма для АКТЦМ, изложенного в разделе 3.7. Поскольку при больших значениях числа каналов N

размер K_1 блоков $Q_{i,j}$ генератора ЦМ $\zeta_t,\ t \geq 0$, существенно меньше размера K соответствующих блоков $Q_{i,j}$ генератора цепи Маркова $\xi_t,\ t \geq 0$, то вычисление стационарного распределения ЦМ $\zeta_t,\ t \geq 0$, удается выполнить для существенно больших значениях числа приборов N.

Таблица 5.4 иллюстрирует размеры блоков генератора для обеих ЦМ и время вычисления для тестовой системы с размерностью пространства состояний управляющих процессов входного потока и процесса обслуживания, равной двум. Вычисления производились на персональном компьютере PC AMD Athlon(tm) 64 3700+, 2.21 GHz, RAM 2 Gb, под управлением операционной системы Microsoft Windows XP. Обозначение типа $XXhYYmZZs$ означает XX часов, YY минут и ZZ секунд. Из таблицы видно, что при $N = 7$ вычисления для ЦМ $\zeta_t,\ t \geq 0$, занимают менее трех минут, в то время как вычисления для ЦМ $\xi_t,\ t \geq 0$, занимают более 32 часов. При $N > 7$ вычисления для ЦМ $\xi_t,\ t \geq 0$, провести не удается из-за невозможности выделения требуемого объема оперативной памяти. Для ЦМ $\zeta_t,\ t \geq 0$, вычисления удается провести до $N = 22$. Информация о размере K_1 соответствующих блоков и времени счета приведена в таблице 5.5. Интересно отметить следующий факт. Размер блоков K равен 510 при $N = 7$. Он почти равен размеру K_1 ($K_1 = 506$) блоков при $N = 21$. А время вычисления (32h16m26s) намного больше (при $N = 21$ оно равно 21h58m24s). Это объясняется тем, что в случае ЦМ $\zeta_t,\ t \geq 0$, последовательность матриц G_i быстрее сходится к предельной матрице G в алгоритме.

Таблица 5.4: Зависимость размерности блоков и времени счета от числа приборов N для ЦМ ξ_t и ζ_t

Число приборов N	K_1	Время вычисления для цепи ξ_t	K	Время вычисления для цепи ζ_t
1	6	0h0m0s	6	0h0m0s
2	12	0h0m1s	14	0h0m2s
3	20	0h0m5s	30	0h0m14s
4	30	0h0m13s	62	0h2m9s
5	42	0h0m37s	126	0h19m13s
6	56	0h1m22s	254	2h46m52s
7	72	0h2m52s	510	32h16m26s

Таблица 5.5: Зависимость размерности блоков и времени счета от числа приборов N для ЦМ ζ_t

Число приборов N	K_1	Время вычисления
7	72	0h2m52s
8	90	0h5m35s
9	110	0h10m17s
10	132	0h17m48s
11	156	0h29m13s
12	182	0h48m15s
13	210	1h17m51s
14	240	1h45m47s
15	272	2h43m33s
16	306	3h52m45s
17	342	5h21m16s
18	380	8h45m59s
19	420	12h26m18s
20	462	16h35m1s
21	506	21h58m24s

5.3.4 Алгоритм для вычисления матриц $P_n(\boldsymbol{\beta})$, $A_n(N,S)$ и $L_{N-n}(N,\tilde{S})$

5.3.4.1 Алгоритм для вычисления матриц $P_n(\boldsymbol{\beta})$, $A_n(N,S)$ и $L_{N-n}(N,\tilde{S})$ Предполагается, что зафиксировано число приборов N, вектор $\boldsymbol{\beta}$ и матрица S, характеризующие процесс обслуживания фазового типа.

Сначала вычисляем матрицы $L_z(N,\tilde{S}),\ z=\overline{0,N-1}$.

Пусть $\tau^{(k)}(S)$ есть матрица, полученная из матрицы S путем удаления ее первых k строк и k столбцов. По определению $\tau^{(0)}(S)=S$.

Начинаем цикл по $j,\ j=\overline{1,M-2}$.

Шаг 1. Вычисление вспомогательных матриц $T(j)$

Вычисляем матрицы $T(j)=\tau^{M-2-j}(S)$. Далее аргумент j будет опускаться. Через $t_{i,l}$ обозначаем элементы матрицы T. Через r обозначаем число строк матрицы T.

Шаг 2. Вычисление вспомогательных матриц $U_z(N,T),\ z=\overline{1,N}$

Пусть $U_z^{(0)}=t_{1,r},\ z=\overline{1,N}$.

Вычисляем в цикле по $w = \overline{1, r-2}$:

$$U_z^{(w)} = \begin{pmatrix} t_{1,r-w}I & U_N^{(w-1)} & 0 & \cdots & 0 & 0 \\ 0 & t_{1,r-w}I & U_{N-1}^{(w-1)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & t_{1,r-w}I & U_z^{(w-1)} \end{pmatrix}, \ z = \overline{1, N}.$$

$U_z(N,T) = zU_z^{(r-2)}, \ z = \overline{1, N}.$

Шаг 3. Вычисление матриц $L_z(N,T), \ z = \overline{0, N-1}$

Пусть $L_z^{(0)} = (N-z)t_{r,1}, \ z = \overline{0, N-1}.$

Вычисляем в цикле по $w = \overline{1, r-2}$:

$$L_z^{(w)} = \begin{pmatrix} (N-z)t_{r-w,1}I & 0 & \cdots & 0 \\ L_{N-1}^{(w-1)} & (N-z-1)t_{r-w,1}I & \cdots & 0 \\ 0 & L_{N-2}^{(w-1)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_{r-w,1}I \\ 0 & 0 & \cdots & L_z^{(w-1)} \end{pmatrix}, \ z = \overline{0, N-1}.$$

Вычисляем матрицы $L_z(N,T) = L_z^{(r-2)}, \ z = \overline{0, N-1}.$

Заканчиваем цикл по $j, \ j = \overline{1, M-2}.$

5.3.4.2 Вычисление матриц $A_m(N,S), \ m = \overline{0, N}$

Пусть

$$A_m^{(0)} = \begin{pmatrix} 0 & mS_{M-1,M} & 0 & \cdots & 0 & 0 \\ S_{M,M-1} & 0 & (m-1)S_{M-1,M} & \cdots & 0 & 0 \\ 0 & 2S_{M,M-1} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & S_{M-1,M} \\ 0 & 0 & 0 & \cdots & mS_{M,M-1} & 0 \end{pmatrix},$$

$$m = \overline{1, N}.$$

Для $j = \overline{1, M-2}$, рекуррентно вычисляем матрицы

$$A_m^{(j)} =$$

$$= \begin{pmatrix} 0 & \frac{m}{N}U_N(N,T) & 0 & \cdots & 0 & 0 \\ L_{N-1}(N,T) & A_1^{(j-1)} & \frac{m-1}{N-1}U_{N-1}(N,T) & \cdots & 0 & 0 \\ 0 & L_{N-2}(N,T) & A_2^{(j-1)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_{m-1}^{(j-1)} & \frac{1}{N-m+1}U_{N-m+1}(N,T) \\ 0 & 0 & 0 & \cdots & L_{N-m}(N,T) & A_m^{(j-1)} \end{pmatrix},$$

$$m = \overline{1, N}.$$

Вычисляем матрицы $A_m(N,S) = A_m^{(M-2)},\ m = \overline{1,N},\ A_0(N,S) = 0.$

5.3.4.3 Вычисление матриц $P_m(\boldsymbol{\beta}),\ m = \overline{1, N-1}$ Вычисляем матрицы размерности $(m+1) \times (m+2)$

$$P_m^{(0)} = \begin{pmatrix} \beta_{M-1} & \beta_M & 0 & \cdots & 0 & 0 \\ 0 & \beta_{M-1} & \beta_M & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \beta_{M-1} & \beta_M \end{pmatrix}, m = \overline{1, N-1}.$$

Для $j = \overline{1, M-2}$, полагаем $\mathbf{a} = \big(\beta_{M-j}, \ldots, \beta_M\big)$, и вычисляем матрицы

$$P_m^{(j)} = \begin{pmatrix} \beta_{M-j-1} & \mathbf{a} & 0 & 0 & \cdots & 0 & 0 \\ 0 & \beta_{M-j-1}I & P_1^{(j-1)} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \beta_{M-j-1}I & P_2^{(j-1)} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \beta_{M-j-1}I & P_m^{(j-1)} \end{pmatrix},$$

$$m = \overline{1, N-1}.$$

Вычисляем матрицы $P_m(\boldsymbol{\beta}) = P_m^{(M-2)},\ m = \overline{1, N-1},\ P_0(\boldsymbol{\beta}) = \boldsymbol{\beta}.$

ГЛАВА 6
МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ ИССЛЕДОВАНИЯ ГИБРИДНЫХ СЕТЕЙ СВЯЗИ НА ОСНОВЕ ЛАЗЕРНОЙ И РАДИОТЕХНОЛОГИЙ

Быстрое и непрерывное увеличение количества пользователей в сети интернет, повышение объема и качества передаваемой информации в широкополосных беспроводных сетях требует резкого увеличения производительности каналов передачи мультимедийной информации. В связи с этим в последние годы в рамках разработки сетей следующего поколения (next generation networks) ведутся интенсивные исследования по повышению производительности беспроводной связи. Одним из направлений создания сверхвысокоскоростных (до 10 Гбит/с) и надежных беспроводных средств связи является разработка гибридных систем на базе лазерной и радиотехнологий.

Технология лазерных атмосферных оптических линий связи или FSO (Free Space Optics) получила широкое распространение в последнее время. Указанная технология основывается на передаче данных модулированным излучением в инфракрасной (или видимой) части спектра через атмосферу и их последующим детектированием оптическим фотоприемным устройством.

К основным преимуществам атмосферных оптических линий связи относятся:

• высокая пропускная способность и качество цифровой связи. Современные FSO-решения могут обеспечить скорость передачи цифровых потоков до 10 Гбит/с при показателе битовых ошибок 10-12, что в настоящее время невозможно достичь при использовании любых других беспроводных технологий;

• высокая защищенность канала от несанкционированного доступа и скрытность. Ни одна беспроводная технология передачи не может предложить такую конфиденциальность связи, как лазерная. Отсутствие ярко выраженных внешних признаков (в основном, это электромагнитное излучение) позволяет скрыть не только передаваемую информацию, но и сам факт информационного обмена. Поэтому лазерные системы часто применяются для разнообразных приложений, где требуется высокая кон-

фиденциальность передачи данных, включая финансовые, медицинские и военные организации;

• высокий уровень помехоустойчивости и помехозащищенности. FSO-оборудование невосприимчиво к радиопомехам и само их не создает;

• скорость и простота развертывания FSO-сети.

Наряду с основными преимуществами беспроводных оптических систем известны и их главные недостатки: зависимость доступности канала связи от погодных условий и необходимость обеспечения прямой видимости между излучателем и приемником. Неблагоприятные погодные условия, такие как снег, туман, могут значительно снизить эффективный диапазон работы лазерных атмосферных линий связи. Так, затухание сигнала в оптическом канале при сильном тумане может доходить до критических 50-100 дБ/км. Поэтому для достижения операторских значений надежности FSO-канала связи необходимо прибегать к использованию гибридных решений.

Гибридное радиооптическое оборудование основывается на использовании резервных радиоканалов (сантиметрового и/или миллиметрового диапазона радиоволн) совместно с оптическим каналом. Отметим, что функционирование радиоканала сантиметрового диапазона радиоволн практически не зависит от погодной среды. На производительность беспроводного канала миллиметрового диапазона не оказывает влияние туман; в то же время соотношение сигнал/шум, определяющее качество функционирования канала, сильно снижается при густом дожде. Такое взаимодополняющее поведение оптических и широкополосных радиоканалов позволило выдвинуть концепцию гибридных систем операторского класса, надежно функционирующих при любых погодных условиях.

В силу высокой потребности в высокоскоростных и надежных каналах связи для решения проблемы «последней мили» в настоящее время широко используются следующие архитектуры гибридных систем [181–188]: высокоскоростной лазерный канал резервируется широкополосным радиоканалом, функционирующим под управлением протокола IEEE 802.11n в сантиметровом диапазоне радиоволн (холодный или горячий резерв); FSO-канал резервируется радиоканалом миллиметрового E-диапазона радиоволн (71-76 ГГц, 81-86 ГГц); FSO-канал и радиоканал миллиметрового диапазона работают параллельно и резервируются каналом IEEE 802.11n, находящемся в холодном резерве.

Практические потребности стимулировали теоретические исследова-

ния по оценке производительности и выбору оптимальных режимов функционирования гибридных систем с использованием моделей теории очередей с ненадежными каналами обслуживания, чередованием их функционирования или параллельной работой, а также учетом случайного времени переключения основного канала на резервный (и наоборот). Первоначально в этих работах [189–191] использовались упрощенные предположения о пуассоновском характере входного потока, экспоненциальном распределении времени обслуживания пакетов, времени безотказной работы и восстановления каналов связи.

В настоящей главе рассмтрен ряд СМО, которые могут быть использованы для моделирования гибридных коммуникационных систем с одним или двумя ненадежными каналами передачи данных и различными схемами резервирования (холодный и горячий резерв). Эти системы, в отличие от упоминавшихся выше, рассматриваются при более общих предположениях о природе входящего потока, потока отказов и распределений длительностей обслуживания и восстановления отказавших приборов.

Некоторые из этих систем были рассмотрены в работах авторов настоящей книги [192,193], посвященных исследованию ненадежных СМО с Марковскими входящими потоками ($BMAP$ и MAP потоки) и PH-распределением времени трансляции пакетов по каналам связи. Хотя указанные предположения усложняют исследование моделей, адекватно описывающих функционирование гибридных систем, но позволяют учитывать нестационарный, коррелированный характер информационных потоков в современных и разрабатываемых перспективных сетях 5G.

6.1 Анализ характеристик процесса передачи информации при архитектуре горячего резервирования высокоскоростного FSO-канала беспроводным широкополосным радиоканалом

В данном разделе рассмотрена однолинейная система массового обслуживания с ненадежным прибором и наличием абсолютно надежного резервного обслуживающего прибора, функционирующего в горячем резерве. Данная система может использоваться при математическом моделировании гибридной сети связи, состоящей из ненадежного FSO-канала и надежного радиоканала, функционирующего под управлением протокола

IEEE 802.11n.

6.1.1 Описание системы

Рассматривается СМО с ожиданием, состоящая из двух приборов, один из которых (основной, прибор номер 1) является ненадежным, а другой (резервный, прибор номер 2) – абсолютно надежным. Последний находится в так называемом "горячем" резерве. Интерпретация: ненадежный прибор – это лазерный, или FSO-канал, а надежный – это радиоканал миллиметрового E-диапазона радиоволн 71-76 ГГц, 81-86 ГГц. При "горячем" резервировании информация передается параллельно по двум каналам. Однако под влиянием смога, тумана, снега FSO канал может выходить из строя, тогда он сразу начинает восстанавливаться, а информация во время восстановления (ремонта) передается по резервному каналу. После восстановления основного прибора он немедленно подключается к передаче информации.

Опишем математическую модель рассматриваемой СМО.

Запросы поступают в систему в соответствии с $BMAP$-потоком (см. раздел 3), который задается управляющим процессом $\nu_t, t \geq 0$, с пространством состояний $\{0, 1, \ldots, W\}$ и матрицами $D_k, k \geq 0$, порядка $W + 1$. Интенсивность поступления запросов, интенсивность поступления групп, коэффициенты вариации и корреляции в $BMAP$ обозначаются как λ, λ_b, c_{var}, c_{cor} соответственно и вычисляются по формулам, приведенным в разделе 3.

Запрос, приходящий на обслуживание, когда приборы являются свободными, немедленно начинает обслуживание на обоих приборах. Если же приборы являются занятыми в момент прихода запроса, запрос становится в очередь, длина которой неограничена, и выбирается на обслуживание позже согласно стратегии FIFO (первым пришел – первым обслужен).

Прибор 1 является более высокоскоростным, но является ненадежным. Поломки поступают на этот прибор в марковском потоке (MAP), который задается управляющим процессом $\eta_t, t \geq 0$, с конечным пространством состояний $\{0, 1, \ldots, V\}$ и матрицами H_0 и H_1. Интенсивность потока поломок задается формулой $h = \boldsymbol{\vartheta} H_1 \mathbf{e}$, где вектор-строка $\boldsymbol{\phi}$ является единственным решением системы уравнений $\boldsymbol{\vartheta}(H_0 + H_1) = \mathbf{0}$, $\boldsymbol{\vartheta}\mathbf{e} = 1$.

Приход поломки вызывает выход из строя прибора 1 независимо от того, занят прибор или нет. Время, необходимое для ремонта прибора, име-

ет PH-распределение с неприводимым представлением $(\boldsymbol{\tau}, T)$. Управляющий процесс $\vartheta_t,\ t \geq 0$, времени ремонта имеет несущественные состояния $\{1, \ldots, R\}$ и поглощающее состояние $R+1$. Вектор-столбец интенсивностей переходов в поглощающее состояние определяется как $\boldsymbol{T}_0 = -T\mathbf{e}$. Интенсивность ремонта вычисляется как $\tau = -(\boldsymbol{\tau} T^{-1}\mathbf{e})^{-1}$.

Предполагаем, что поломка, поступившая, когда прибор еще не восстановился после предыдущей поломки, игнорируется.

Время обслуживания запроса прибором с номером $k,\ k = 1, 2$, имеет PH-распределение с неприводимым представлением $(\boldsymbol{\beta}^{(k)}, S^{(k)})$ и имеет управляющий процесс $m_t^{(k)},\ t \geq 0$, в несущественными состояниями $\{1, \ldots, M^{(k)}\}$ и поглощающим состоянием $M^{(k)} + 1$. Векторы интенсивностей переходов с поглощающее состояние определяются как $\mathbf{S}_0^{(k)} = -S^{(k)}\mathbf{e}, k = 1, 2$. Интенсивности обслуживания задаются как $\mu^{(k)} = -[\boldsymbol{\beta}^{(k)}(S^{(k)})^{-1}\mathbf{e}]^{-1},\ k = 1, 2$.

При взятии запроса на обслуживание, если оба прибора исправны, они одновременно начинают его обслуживание. В момент окончания обслуживания данного запроса каким-либо прибором его обслуживание другим прибором немедленно прекращается.

Как отмечалось выше, приход поломки вызывает прекращение обслуживания запроса (если таковое имеет место) прибором 1, который уходит на ремонт, в то время как прибор 2 продолжает обслуживание запроса. Если при взятии запроса на обслуживание исправен только прибор 2, он проводит обслуживание запроса. Если во время этого обслуживания прибор 1 восстанавливается, он немедленно начинает обслуживание этого же запроса. Запрос будет обслужен, когда его обслуживание одним из приборов будет завершено. При этом другой прибор прекращает обслуживание данного запроса.

6.1.2 Цепь Маркова, описывающая функционирование системы

Пусть в момент времени t

• i_t – число запросов в системе, $i_t \geq 0$;

• $r_t = 0$, если прибор 1 на ремонте в момент времени t и $r_t = 1$, если прибор 1 исправен;

• $m_t^{(k)}$ – состояние управляющего процесса обслуживания на k-м занятом приборе, $m_t^{(k)} = \overline{1, M^{(k)}},\ k = 1, 2$;

• ϑ_t – состояние управляющего процесса ремонта прибора 1, $\vartheta_t = \overline{1, R}$, if $r_t = 0$;

• ν_t and η_t – состояние управляющих процессов $BMAP$-потока запросов и MAP-потока поломок соответственно, $\nu_t = \overline{0, W}$, $\eta_t = \overline{0, V}$.

Тогда процесс изменения состояний системы описывается регулярной регулярной неприводимой цепью Маркова с непрерывным временем $\xi_t, t \geq 0$, с пространством состояний

$$X = \{(0, 0, \vartheta, \eta, \nu)\} \bigcup \{(0, 1, \eta, \nu)\} \bigcup \{(i, 0, \vartheta, m^{(2)}, \eta, \nu)\} \bigcup$$
$$\{(i, 1, m^{(1)}, m^{(2)}, \eta, \nu)\},\ i \geq 1,\ \vartheta = \overline{1, R},\ m^{(k)} = \overline{1, M^{(k)}}, k = 1, 2,$$
$$\eta = \overline{0, V},\ \nu = \overline{0, W}.$$

Далее будем предполагать, что состояния цепи упорядочены следующим образом. При фиксированном значении пары (i, r) состояния цепи упорядочим в лексикографическом порядке. Обозначим полученное множество как $\Omega_{i,r}$. Все состояния цепи из множества X упорядочим, расположив множества $\Omega_{i,r}$, следующим образом:

$$\Omega_{0,0},\ \Omega_{0,1},\ \Omega_{1,0},\ \Omega_{1,1},\ \Omega_{2,0},\ \Omega_{2,1},\ \Omega_{3,0},\ \Omega_{3,1}, \ldots.$$

Пусть Q_{ij}, $i, j \geq 0$, – матрицы, образованные интенсивностями переходов цепи из состояний, соответствующих значению i счетной компоненты i_n, в состояния, соответствующие значению j этой компоненты. Справедливо следующее утверждение.

Лемма 6.1. *Инфинитезимальный генератор Q цепи Маркова ξ_t, $t \geq 0$, имеет следующую блочную структуру:*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \cdots \\ Q_{1,0} & Q_1 & Q_2 & Q_3 & \cdots \\ O & Q_0 & Q_1 & Q_2 & \cdots \\ O & O & Q_0 & Q_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где ненулевые блоки Q_{ij} имеют вид

$$Q_{0,0} = \begin{pmatrix} T \oplus (H_0 + H_1) \oplus D_0 & \mathbf{T}_0 \otimes I_a \\ \boldsymbol{\tau} \otimes H_1 \otimes I_{\bar{W}} & H_0 \oplus D_0 \end{pmatrix},$$

$$Q_{0,k} = \begin{pmatrix} I_R \otimes \boldsymbol{\beta}^{(2)} \otimes I_{\bar{V}} \otimes D_k & O \\ O & \boldsymbol{\beta}^{(1)} \otimes \boldsymbol{\beta}^{(2)} \otimes I_{\bar{V}} \otimes D_k \end{pmatrix},\ k \geq 1,$$

$$Q_{1,0} = \begin{pmatrix} I_R \otimes \mathbf{S}_0^{(2)} \otimes I_a & O \\ O & (\mathbf{S}_0^{(1)} \otimes \mathbf{e}_{M^{(2)}} + \mathbf{e}_{M^{(1)}} \otimes \mathbf{S}_0^{(2)}) \otimes I_a \end{pmatrix},$$

$$Q_0 = \begin{pmatrix} I_R \otimes \mathbf{S}_0^{(2)} \boldsymbol{\beta}^{(2)} \otimes I_a & O \\ O & \tilde{S} \otimes I_a \end{pmatrix},$$

$$Q_1 = \begin{pmatrix} T \oplus S^{(2)} \oplus (H_0 + H_1) \oplus D_0 & \mathbf{T}_0 \boldsymbol{\beta}^{(1)} \otimes I_{M^{(2)}a} \\ \mathbf{e}_{M^{(1)}} \boldsymbol{\tau} \otimes I_{M^{(2)}} \otimes H_1 \otimes I_{\bar{W}} & S^{(1)} \oplus S^{(2)} \oplus H_0 \oplus D_0 \end{pmatrix},$$

$$Q_k = \begin{pmatrix} I_{RM^{(2)}\bar{V}} \otimes D_{k-1} & O \\ O & I_{M^{(1)}M^{(2)}\bar{V}} \otimes D_{k-1} \end{pmatrix}, \; k \geq 2.$$

Здесь $\tilde{S} = \mathbf{S}_0^{(1)} \boldsymbol{\beta}^{(1)} \otimes \mathbf{e}_{M^{(2)}} \boldsymbol{\beta}^{(2)} + \mathbf{e}_{M^{(1)}} \boldsymbol{\beta}^{(1)} \otimes \mathbf{S}_0^{(2)} \boldsymbol{\beta}^{(2)}$, $\bar{W} = W + 1$, $\bar{V} = V + 1$, $a = \bar{W}\bar{V}$.

Доказательство леммы выполняется путем вычисления вероятностей переходов цепи Маркова ξ_t, $t \geq 0$, на бесконечно малом интервале времени, извлечения из выражений для этих вероятностей интенсивностей переходов и дальнейшего объединения интенсивностей в блоки $Q_{i,j}$.

Анализируя вид генератора Q, легко видеть, что он имеет блочную верхне-хессенбергову структуру и блоки $Q_{i,j}$, сформированные интенсивностями переходов цепи ξ_t, $t \geq 0$, из состояний, соответствующих значению $i > 1$ счетной компоненты, в состояния, соответствующие значению j этой компоненты, зависят от i, j только через разность $j - i$. Это означает, что рассматриваемая цепь Маркова принадлежит классу квазитеплицевых цепей Маркова, который был исследован в [150].

Следствие 6.1. *Цепь Маркова* $\xi_t, t \geq 0$, *принадлежит классу квазитеплицевых цепей Маркова с непрерывным временем.*

В дальнейшем полезно будет иметь выражения для матричных производящих функций $\tilde{Q}(z) = \sum\limits_{k=1}^{\infty} Q_{0,k} z^k$, $Q(z) = \sum\limits_{k=0}^{\infty} Q_k z^k$, $|z| \leq 1$.

Следствие 6.2. *Матричные производящие функции* $\tilde{Q}(z)$, $Q(z)$ *имеют следующий вид:*

$$\tilde{Q}(z) = z \begin{pmatrix} I_R \otimes \boldsymbol{\beta}^{(2)} \otimes I_{\bar{W}} & O \\ O & \boldsymbol{\beta}^{(1)} \otimes \boldsymbol{\beta}^{(2)} \otimes I_a \end{pmatrix} + \operatorname{diag}\{I_{R\bar{V}} \otimes (D(z) - D_0), \; I_{\bar{V}} \otimes (D(z) - D_0)\}, \tag{6.1}$$

$$Q(z) = Q_0 + z \begin{pmatrix} T \oplus S^{(2)} \oplus (H_0 + H_1) \otimes I_{\bar{W}} & T_0 \boldsymbol{\beta}^{(1)} \otimes I_{M^{(2)}} \otimes I_a \\ \mathbf{e}_{M^{(1)}} \boldsymbol{\tau} \otimes I_{M^{(2)}} \otimes H_1 \otimes I_{\bar{W}} & S^{(1)} \oplus S^{(2)} \oplus H_0 \otimes I_{\bar{W}} \end{pmatrix}$$

$$+z\mathrm{diag}\{I_{RM^{(2)}\bar{V}} \otimes D(z),\ I_{M^{(1)}M^{(2)}\bar{V}} \otimes D(z)\}. \qquad (6.2)$$

6.1.3 Условие существования стационарного режима в системе. Стационарное распределение

Теорема 6.1. *Необходимым и достаточным условием существования стационарного режима в системе является выполнение неравенства*

$$\lambda < \mathbf{x}\ \mathrm{diag}\{\mathbf{e}_R \otimes \mathbf{S}_0^{(2)} \otimes \mathbf{e}_{\bar{V}}, (\mathbf{S}_0^{(1)} \otimes \mathbf{e}_{M^{(2)}} + \mathbf{e}_{M^{(1)}} \otimes \mathbf{S}_0^{(2)}) \otimes \mathbf{e}_{\bar{V}}\}\mathbf{e}, \qquad (6.3)$$

где вектор $\mathbf{x}$ *есть единственное решение системы линейных алгебраических уравнений*

$$\mathbf{x}\Gamma = \mathbf{0},\ \mathbf{x}\mathbf{e} = 1, \qquad (6.4)$$

а матрица Γ *имеет вид*

$$\Gamma = \begin{pmatrix} T \oplus S^{(2)} \oplus (H_0 + H_1) + I_R \otimes \mathbf{S}_0^{(2)}\boldsymbol{\beta}^{(2)} \otimes I_{\bar{V}} & \mathbf{T}_0\boldsymbol{\beta}^{(1)} \otimes I_{M^{(2)}} \otimes I_{\bar{V}} \\ \mathbf{e}_{M^{(1)}}\boldsymbol{\tau} \otimes I_{M^{(2)}} \otimes H_1 & \tilde{S} \otimes I_{\bar{V}} + S^{(1)} \oplus S^{(2)} \oplus H_0 \end{pmatrix}.$$

Доказательство. Условие существования стационарного режима в системе находится как условие существования стационарного распределения цепи Маркова $\xi_t, t \geq 0$, описывающей процесс функционирования системы. В силу неприводимости цепи условие существования ее стационарного распределения совпадает с условием эргодичности. При выводе условия эргодичности воспользуемся результатами для квазитеплицевых цепей Маркова, изложенными [150]. Согласно [150], необходимое и достаточное условие эргодичности квазитеплицевой цепи Маркова ξ_t, $t \geq 0$, может быть сформулировано в терминах производящей функции $Q(z)$ и имеет вид неравенства

$$\mathbf{y}Q^{'}(1)\mathbf{e} < 0, \qquad (6.5)$$

где вектор-строка $\mathbf{y}$ есть единственное решение системы линейных алгебраических уравнений

$$\mathbf{y}Q(1) = \mathbf{0},\ \mathbf{y}\mathbf{e} = 1. \qquad (6.6)$$

Представим вектор $\mathbf{y}$ в виде

$$\mathbf{y} = \mathbf{x} \otimes \boldsymbol{\theta}, \qquad (6.7)$$

где $\boldsymbol{\theta}$ – вектор стационарного распределения управляющего процесса ν_t, $t \geq 0$, $BMAP$-потока, а $\mathbf{x}$ – некоторый стохастический вектор. Подставляя выражение (6.7) в (6.6), убеждаемся в том, что $\mathbf{y}$ есть единственное решение системы (6.6), если вектор $\mathbf{x}$ удовлетворяет системе (6.4).

Подставляя выражение (6.7) для вектора $\mathbf{y}$ в (6.5) и используя формулу (6.2) для вычисления производной $Q'(1)$, сводим неравенство (6.5) к виду (6.3). □

Замечание 6.1. Условие эргодичности (6.3) становится интуитивно понятным, если учесть следующее. Левая часть неравенства (6.3) есть интенсивность входящего потока, а правая часть этого неравенства – интенсивность выходящего потока в условиях перегрузки системы. Очевидно, что в стационарном режиме последняя интенсивность должна быть меньше интенсивности входящего потока.

Замечание 6.2. Коэффициент загрузки системы ρ находится как отношение левой части (6.3) к правой части, т.е.

$$\rho = \frac{\lambda}{\mathbf{x}\,\mathrm{diag}\{\mathbf{e}_R \otimes \mathbf{S}_0^{(2)} \otimes \mathbf{e}_{\bar{V}}, (\mathbf{S}_0^{(1)} \otimes \mathbf{e}_{M^{(2)}} + \mathbf{e}_{M^{(1)}} \otimes \mathbf{S}_0^{(2)}) \otimes \mathbf{e}_{\bar{V}}\}\mathbf{e}}.$$

Следствие 6.3. *В случае стационарного пуассоновского потока поломок и экспоненциальноro распределения времен обслуживания и ремонтов условие (6.3)-(6.4) существования стационарного режима в системе сводится к следующему неравенству:*

$$\lambda < \mu_2 + \frac{\tau}{\tau + h}\mu_1. \tag{6.8}$$

Доказательство. В рассматриваемом случае вектор $\mathbf{x}$ состоит из двух компонент, пусть $\mathbf{x} = (x_1, x_2)$. Легко видеть, что неравенство (6.3) сводится к виду

$$\lambda < x_1\mu_2 + x_2(\mu_1 + \mu_2). \tag{6.9}$$

Система (6.4) запишется как

$$\begin{cases} -x_1\tau + x_2 h = 0, \\ x_1\tau - x_2 h = 0, \\ x_1 + x_2 = 1. \end{cases} \tag{6.10}$$

Решая систему (6.10) и подставляя решение в (6.9), получим искомое неравенство (6.8). □

В дальнейшем будем предполагать, что условие существования стационарного режима, заданное теоремой 6.1, выполняется.

Введем обозначения для стационарных вероятностей цепи Маркова ξ_t, $t \geq 0$:

$$p(0,0,\vartheta,\eta,\nu),\ p(0,1,\eta,\nu),\ \ p(i,0,\vartheta,m^{(2)},\eta,\nu),$$
$$p(i,1,m^{(1)},m^{(2)},\eta,\nu),\ i \geq 1,\ \vartheta = \overline{1,R},\ m^{(k)} = \overline{1,M^{(k)}}, k = 1,2,$$
$$\eta = \overline{0,V},\ \nu = \overline{0,W}.$$

Перенумеруем стационарные вероятности в лексикографическом порядке и сформируем векторы-строки $\mathbf{p}_i$ стационарных вероятностей, соответствующих значению i счетной компоненты цепи, $i \geq 0$.

Чтобы вычислить векторы $\mathbf{p}_i$, $i \geq 0$, используется численно устойчивый алгоритм, разработанный в [150] для вычисления стационарного распределения многомерных квазитеплицевых цепей Маркова с непрерывным временем. Этот алгоритм основан на использовании сенсорных цепей Маркова и состоит из следующих основных шагов.

Алгоритм 6.1.

1. Вычисляем матрицу G как единственное минимальное неотрицательное решение нелинейного матричного уравнения

$$\sum_{n=0}^{\infty} Q_n G^n = O.$$

2. Вычисляем матрицу G_0 из уравнения

$$Q_{1,0} + \sum_{n=1}^{\infty} Q_{1,n} G^{n-1} G_0 = O,$$

откуда

$$G_0 = -(\sum_{n=1}^{\infty} Q_{1,n} G^{n-1})^{-1} Q_{1,0}.$$

3. Вычисляем матрицы $\bar{Q}_{i,l}$, $l \geq i, i \geq 0$, используя формулы

$$\bar{Q}_{i,l} = \begin{cases} Q_{0,l} + \sum\limits_{n=l+1}^{\infty} Q_{0,n} G_{n-1} G_{n-2} \ldots G_l,\ i = 0,\ l \geq 0, \\ Q_{l-i} + \sum\limits_{n=l+1}^{\infty} Q_{n-i} G_{n-1} G_{n-2} \ldots G_l,\ i \geq 1,\ l \geq i, \end{cases}$$

где $G_i = G$, $i \geq 1$.

6. Вычисляем матрицы Φ_i по рекуррентным формулам

$$\Phi_0 = I_{R\bar{V}\bar{W}}, \Phi_i = \sum_{l=0}^{i-1} \Phi_l \bar{Q}_{l,i}(-\bar{Q}_{i,i})^{-1},\ i \geq 1.$$

5. Вычисляем вектор $\mathbf{p_0}$ как единственное решение системы линейных алгебраических уравнений

$$\boldsymbol{p}_0(-\bar{Q}_{0,0}) = \mathbf{0},\ \boldsymbol{p}_0(\mathbf{e}_{R\bar{V}\bar{W}} + \sum_{i=1}^{\infty} \Phi_i \mathbf{e} = 1.$$

6. Вычисляем векторы $\mathbf{p_i},\ \mathbf{i \geq 1}$, как

$$\boldsymbol{p}_i = \boldsymbol{p}_0 \Phi_i,\ i \geq 1.$$

Предложенный алгоритм является численно устойчивым, так как все вовлеченные в него матрицы являются неотрицательными.

6.1.4 Векторная производящая функция стационарного распределения. Характеристики производительности

Вычислив стационарное распределение $\mathbf{p}_i$, $i \geq 0$, можно вычислить также ряд важных стационарных характеристик производительности системы. При вычислении характеристик производительности, особенно в случае когда стационарное распределение имеет "тяжелый хвост", очень полезным является следующий результат.

Лемма 6.2. *Векторная производящая функция* $\mathbf{P}(z) = \sum_{i=1}^{\infty} \mathbf{p}_i z^i$, $|z| \leq 1$, *удовлетворяет следующему уравнению:*

$$\mathbf{P}(z)Q(z) = z[\mathbf{p}_1 Q_0 - \mathbf{p}_0 \tilde{Q}(z)]. \tag{6.11}$$

Формула (6.11) может использоваться, в частности, при нахождении значений производящей функции $\mathbf{P}(z)$ и ее производных в точке $z = 1$ без вычисления бесконечных сумм. Как будет видно далее, вычислив указанные производные, нетрудно вычислить многие характеристики производительности системы. Вместе с тем, задача нахождения значений функции $\mathbf{P}(z)$ и ее производных в точке $z = 1$ из уравнения (6.11) не является тривиальной, как это может показаться на первый взгляд. Причина этого в

том, что матрица $Q(z)$ является вырожденной в точке $z = 1$. Так, например, непосредственно из (6.11) следует неоднородная система линейных алгебраических уравнений для элементов вектора $\mathbf{P}(1)$, но матрица этой системы, $Q(1)$, является вырожденной. В такой ситуации вспоминаем, что матрица $Q(1)$ является неприводимым генератором, т.е. имеет ранг на единицу меньший ее размерности, и пытаемся заменить одно из уравнений этой системы на некоторое другое уравнение таким образом, чтобы матрица полученной системы была невырожденной. Эта идея лежит в основе приведенного ниже следствия (6.4).

Прежде чем сформулировать следствие, обозначим через $f^{(n)}(z)$ n-ю производную функции $f(z)$, $n \geq 1$, и положим $f^{(0)}(z) = f(z)$.

Следствие 6.4. *m-я, $m \geq 0$, производная векторной производящей функции $\mathbf{P}(z)$ в точке $z = 1$ (так называемый m-й векторный факториальный момент) вычисляется по рекуррентной формуле*

$$\mathbf{P}^{(m)}(1) = \left[\left(\mathbf{b}^{(m)}(1) - \sum_{l=0}^{m-1} C_m^l \mathbf{P}^{(l)}(1) Q^{(m-l)}(1)\right)\tilde{I} + \right.$$

$$\left. \frac{1}{m+1}\left[\mathbf{b}^{(m+1)}(1) - \sum_{l=0}^{m-1} C_{m+1}^l \mathbf{P}^{(l)}(1) Q^{(m+1-l)}(1)\right)\mathbf{e}\hat{\mathbf{e}}\right]\mathcal{Q}^{-1}, \qquad (6.12)$$

где

$$\mathcal{Q} = Q(1)\tilde{I} + Q'(1)\mathbf{e}\hat{\mathbf{e}},\ \hat{\mathbf{e}} = (1, 0, \ldots, 0)^T,$$

$$\mathbf{b}^{(m)}(1) = \begin{cases} \mathbf{p}_1 Q_0 - \mathbf{p}_0 \tilde{Q}(1), & m = 0, \\ \mathbf{p}_1 Q_0 - \mathbf{p}_0 \tilde{Q}(1) - \mathbf{p}_0 \tilde{Q}'(1), & m = 1, \\ -\mathbf{p}_0 [m\tilde{Q}^{(m-1)}(1) + Q^{(m)}(1)], & m > 1, \end{cases}$$

$\tilde{I}$ – диагональная матрица с диагональными элементами $(0, 1, \ldots, 1)$, а матричные производные $Q^{(m)}(1)$, $\tilde{Q}^{(m)}(1)$ вычисляются с использованием формул (6.1)-(6.2).

Доказательство. Обозначим вектор, стоящий в правой части уравнения (6.11) как $\mathbf{b}(z)$. Дифференцируя m раз уравнение (6.11), получим следующую неоднородную систему линейных алгебраических уравнений для компонент вектора $\mathbf{P}^{(m)}(1)$:

$$\mathbf{P}^{(m)}(1)Q(1) = \mathbf{b}^{(m)}(1) - \sum_{l=0}^{m-1} C_m^l \mathbf{P}^{(l)}(1) Q^{(m-l)}(1). \qquad (6.13)$$

Как упоминалось выше, матрица $Q(1)$ этой системы вырождена и имеет ранг на единицу меньший ее порядка. Продифференцируем (6.13) еще раз и умножим обе части полученного уравнения на вектор $\mathbf{e}$. Учитывая, что $Q(1)\mathbf{e} = \mathbf{0}^T$, получим еще одно уравнение для компонент вектора $\mathbf{P}^{(m)}(1)$

$$\mathbf{P}^{(m)}(1)Q'(1)\mathbf{e} = \frac{1}{m+1}[\mathbf{b}^{(m+1)}(1) - \sum_{l=0}^{m-1} C_{m+1}^{l}\mathbf{P}^{(l)}(1)Q^{(m+1-l)}(1)]\mathbf{e}, \quad (6.14)$$

Теперь заменим одно из уравнений системы (6.13) (без ограничения общности заменим первое уравнение) на уравнение (6.14). В итоге получим неоднородную систему линейных алгебраических уравнений для компонент вектора $\mathbf{P}^{(m)}(1)$

$$\mathbf{P}^{(m)}(1)\mathcal{Q} = \left[\left(\mathbf{b}^{(m)}(1) - \sum_{l=0}^{m-1} C_{m}^{l}\mathbf{P}^{(l)}(1)Q^{(m-l)}(1)\right)\tilde{I} + \right.$$

$$\left.\frac{1}{m+1}\left[\mathbf{b}^{(m+1)}(1) - \sum_{l=0}^{m-1} C_{m+1}^{l}\mathbf{P}^{(l)}(1)Q^{(m+1-l)}(1)\right)\mathbf{e}\hat{\mathbf{e}}\right], \quad (6.14)$$

которая имеет единственное решение (6.12), если матрица $\mathcal{Q}$ невырожденная.

Вычислим определитель этой матрицы. Из ее определения и того факта, что $Q(1)\mathbf{e} = \mathbf{0}^T$, следует, что $\det\mathcal{Q} = [\det Q(z)]'_{z=1}$. В свою очередь, можно показать (см. [150]), что условие (6.3) существования стационарного режима в системе эквивалентно условию $[\det Q(z)]'_{z=1} < 0$. Поскольку предполагается, что условие (6.3) выполняется, то $\det\mathcal{Q} = [\det Q(z)]'_{z=1} < 0$, т.е. матрица $\mathcal{Q}$ – невырожденная. Из этого следует, что система (6.14) имеет единственное решение (6.12). □

Далее приведем некоторые важные характеристики производительности рассматриваемой системы.

- Пропускная способность системы (максимальное число запросов, которые могут обслужиться в единицу времени) определяется как правая часть неравенства (6.3).
- Среднее число запросов в системе

$$L = \mathbf{P}^{(1)}(1)\mathbf{e}.$$

- Дисперсия числа запросов в системе

$$V = \mathbf{P}^{(2)}(1)\mathbf{e} + L - L^2.$$

- Доля времени, в течение которого прибор 1 исправен

$$P_1 = \mathbf{P}(1)\mathrm{diag}\{O_{RM^{(2)}a}, I_{M^{(1)}M^{(2)}a}\}\mathbf{e} + \mathbf{p}_0\mathrm{diag}\{O_{Ra}, I_a\}\mathbf{e}.$$

- Доля времени, в течение которого прибор 1 не исправен (находится на ремонте)

$$P_0 = \mathbf{P}(1)\mathrm{diag}\{I_{RM^{(2)}a}, O_{M^{(1)}M^{(2)}a}\}\mathbf{e} + \mathbf{p}_0\mathrm{diag}\{I_{Ra}, O_a\}\mathbf{e}.$$

6.1.5 Распределение времени пребывания в системе

Пусть $V(x)$ – функция стационарного распределения времени пребывания произвольго запроса в системе, $v(s) = \int\limits_0^\infty e^{-sx}dV(x),\ Re\ s \geq 0$, – преобразование Лапласа – Стилтьеса этой функции.

Теорема 6.2. *Преобразование Лапласа – Стилтьеса стационарного распределения времени пребывания запроса в системе вычисляется как*

$$v(s) = \lambda^{-1}\{\mathbf{p}_0 \sum_{k=1}^{\infty} Q_{0,k} \sum_{l=1}^{k} \Psi^l(s) + \sum_{i=1}^{\infty} \mathbf{p}_i \sum_{k=2}^{\infty} Q_k \sum_{l=1}^{k-1} \Psi^{i+l}(s)\}\mathbf{e}, \qquad (6.15)$$

где

$$\Psi(s) = (sI - \hat{Q})^{-1}Q_0,\ \ \hat{Q} = Q(1) - Q_0.$$

Доказательство. Доказательство основано на вероятностном смысле ПЛС. Предполагается, что независимо от функционирования системы поступает стационарный пуассоновский поток так называемых катастроф с интенсивностью s, $s > 0$. Тогда ПЛС $v(s)$ можно интерпретировать как вероятность того, что за время пребывания запроса в системе не произойдет катастрофа. Такая интерпретация позволяет вывести выражение для $v(s)$ путем вероятностных рассуждений.

Предполагаем, что в начале обслуживания запроса первоначальная фаза времени обслуживания уже установлена. Тогда матрица вероятностей того, что за время обслуживания не произойдет катастрофа и произойдут соответствующие переходы конечных компонент цепи Маркова ξ_t, $t \geq, 0$ вычисляется как

$$\tilde{\Psi}(s) = \int_0^\infty e^{(-sI+\hat{Q})t} Q_{1,0} dt = (sI - \hat{Q})^{-1} Q_{1,0},$$

если в момент окончания обслуживания нет очереди, и как

$$\Psi(s) = \int_0^\infty e^{(-sI+\hat{Q})t} Q_0 dt = (sI - \hat{Q})^{-1} Q_0,$$

если в момент окончания обслуживания есть очередь.

Заметим, что $\tilde{\Psi}(s)\mathbf{e} = \Psi(s)\mathbf{e}$ так как $Q_{1,0}\mathbf{e} = Q_0\mathbf{e}$.

Предположим, что произвольный запрос, поступающий в группе размера k, размещается на j-м месте в группе с вероятностью $1/k$. Заметим, что

$$\tilde{\Psi}(s)\mathbf{e} = \Psi(s)\mathbf{e}, \tag{6.16}$$

так как $Q_{1,0}\mathbf{e} = Q_0\mathbf{e}$. Тогда, используя (6.16) и формулу полной вероятности, получим следующее выражение для $v(s)$:

$$v(s) = \mathbf{p}_0 \sum_{k=1}^{\infty} \frac{k}{\lambda} Q_{0,k} \sum_{l=1}^{k} \frac{1}{k} \Psi^l(s)\mathbf{e} +$$

$$+ \sum_{i=1}^{\infty} \mathbf{p}_i \sum_{k=2}^{\infty} \frac{k-1}{\lambda} Q_k \sum_{l=1}^{k-1} \frac{1}{k-1} \Psi^{i+l}(s)\mathbf{e}. \tag{6.17}$$

Формула (6.15) немедленно следует из формулы (6.17).

□

Используя теорему 6.2, можно находить моменты произвольного порядка времени пребывания путем дифференцирования (6.15) в точке $s = 0$. Момент k-го порядка вычисляется как $v^{(k)} = \frac{d^k v(s)}{ds^k}|_{s=0}$, $k \geq 1$. В частности имеет место

Следствие 6.5. *Среднее время пребывания произвольного запроса в системе вычисляется как*

$$\bar{v} = -\lambda^{-1}[\mathbf{p}_0 \sum_{k=1}^{\infty} Q_{0,k} \sum_{l=1}^{k} \sum_{m=0}^{l-1} \Psi^m(0) +$$

$$+ \sum_{i=1}^{\infty} \mathbf{p}_i \sum_{k=2}^{\infty} Q_k \sum_{l=1}^{k-1} \sum_{m=0}^{i+l-1} \Psi^m(0)]\Psi'(0)\mathbf{e}, \tag{6.18}$$

где

$$\Psi'(0) = -[\hat{Q}(1)]^{-2} Q_0.$$

Доказательство. Чтобы получить формулу (6.18), дифференцируем (6.15) в точке $s = 0$ и затем преобразуем полученное выражение, используя тот факт, что матрица $\Psi(0)$ является стохастической. □

6.2 Анализ характеристик процесса передачи информации при архитектуре холодного резервирования высокоскоростного FSO-канала беспроводным широкополосным радиоканалом

Данный раздел посвящен исследованию однолинейной системы массового обслуживания с ненадежным прибором и "холодным" резервированием. Эта система отличается от системы с "горячим" резервированием, рассмотренной в предыдущем разделе, тем, что резервный прибор не передает информацию параллельно с основным прибором, а подключается к передаче только во время ремонта основного прибора.

6.2.1 Описание системы

Рассматривается система массового обслуживания, состоящая из двух приборов, один из которых (основной) является ненадежным, а другой (резервный) – абсолютно надежным. Главное отличие от системы, исследованной в предыдущем разделе, состоит в том, что здесь используется схема "холодного" резервирования, при котором резервный прибор подключается к обслуживанию запросов только во время ремонта основного прибора. Кроме того, здесь предполагается, что переключение с основного прибора на резервный не мгновенное, оно требует времени.

Опишем более подробно сценарий взаимодействия основного и резервного приборов (прибора номер 1 и прибора номер 2 соответственно). Если основной прибор исправен, то обслуживание заявок производится им. Если произошла поломка основного прибора, то одновременно запускаются механизмы ремонта этого прибора и переключения на резервный прибор. Если ремонт закончится раньше, чем закончится переключение, то переключение прекращается, и основной прибор возобновляет обслуживание, прерванное приходом поломки, или берет первую заявку из очереди, если

в момент прихода поломки прибор простаивал, а за время ремонта в буфер пришли заявки. Если прибор был пуст как в момент прихода поломки, так и в момент окончания ремонта, он будет ждать поступления новой заявки. Если же переключение закончилось раньше, чем ремонт, то резервный прибор обслуживает заново заявку, обслуживание которой было прервано приходом поломки, а затем обслуживает заявки из буфера и вновь поступающие заявки до тех пор, пока не закончится ремонт основного прибора. С момента окончания ремонта обслуживание заявок немедленно вновь переносится на основной прибор. Обслуживание заявки, которая находилась на обслуживании резервным прибором, если таковая имелась, начинается на основном приборе заново.

Входной поток запросов, поток поломок, процессы обслуживания на обоих приборах и процесс ремонтов основного прибора такие же, как аналогичные процессы, описанные выше, в разделе 6.2.1. Время переключения с основного на резервный прибор имеет PH-распределение с неприводимым представлением $(\boldsymbol{\alpha}, A)$. Процесс переключения происходит под управлением цепи Маркова $l_t,\ t \geq 0$, с пространством состояний $\{1, \ldots, L, L{+}1\}$, где $L+1$ есть поглощающее состояние. Интенсивности переходов в поглощающее состояние задаются вектором $\boldsymbol{A}_0 = -A\mathbf{e}$. Интенсивность переключения вычисляется как $\alpha = -[\boldsymbol{\alpha}A^{-1}\mathbf{e}]^{-1}$.

6.2.2 Цепь Маркова, описывающая функционирование системы

Пусть в момент t

i_t - число запросов в системе, $i_t \geq 0$,

$$n_t = \begin{cases} 0, & \text{если основной прибор (прибор номер 1) исправен;} \\ 1, & \text{если основной прибор на ремонте;} \end{cases}$$

$$r_t = \begin{cases} 0, & \text{если момент } t \text{ не принадлежит периоду переключения;} \\ 1, & \text{если в момент } t \text{ идет переключение;} \end{cases}$$

$m_t^{(j)}$ – состояние управляющего процесса PH – обслуживания на j-м занятом приборе, $j = 1, 2, m_t^{(j)} = \overline{1, M^{(j)}}$;

l_t – состояние управляющего процесса PH-времени переключения с прибора номер 1 на прибор номер 2 , $l_t = \overline{1, L}$;

ϑ_t – состояние управляющего процесса PH-времени ремонта, $\vartheta_t = \overline{1, R}$;

ν_t и η_t – состояния управляющих процессов входящего $BMAP$-потока

и MAP-потока поломок соответственно, $\nu_t = \overline{0, W}$, $\eta_t = \overline{0, V}$.

Процесс функционирования системы описывается регулярной неприводимой цепью Маркова ξ_t, $t \geq 0$, с пространством состояний

$$X = \{(i, n, \nu, \eta),\, i = 0, n = 0, \nu = \overline{0, W}, \eta = \overline{0, V}\} \bigcup$$

$$\{(i, n, \nu, \eta, \vartheta),\, i = 0, n = 1, \nu = \overline{0, W}, \eta = \overline{0, V}, \vartheta = \overline{0, R}\} \bigcup$$

$$\{(i, n, r, \nu, \eta, m^{(1)}),\, i > 0, n = 0,\, r = 0, \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(1)} = \overline{1, M^{(1)}}\} \bigcup$$

$$\{(i, n, r, \nu, \eta, m^{(2)}, \vartheta),\, i > 0, n = 1,\, r = 0, \nu = \overline{0, W}, \eta = \overline{0, V}, \vartheta = \overline{1, R},$$

$$m^{(2)} = \overline{1, M^{(2)}}\} \bigcup \{(i, n, r, \nu, \eta, \vartheta, l),\, i > 0, n = 1,\, r = 1, \nu = \overline{0, W}, \eta = \overline{0, V},$$

$$\vartheta = \overline{1, R}, l^{(2)} = \overline{1, L}\}.$$

Далее будем предполагать, что состояния цепи $\xi_t, t \geq 0$, внутри каждого из обозначенных подмножеств упорядочены в лексикографическом порядке и таким образом упорядоченные подмножества упорядочены в том порядке, в котором они перечислены выше. Обозначим через $Q_{i,j}$ матрицу интенсивностей переходов цепи их состояний, соответствующих значению i первой (счетной) компоненты в состояния, соответствующие значению j этой компоненты, $i, j \geq 0$.

Лемма 6.3. *Инфинитезимальный генератор Q цепи Маркова ξ_t, $t \geq 0$, имеет следующую блочную структуру*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \cdots \\ Q_{1,0} & Q_1 & Q_2 & Q_3 & \cdots \\ O & Q_0 & Q_1 & Q_2 & \cdots \\ O & O & Q_0 & Q_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

где ненулевые блоки имеют следующий вид:

$$Q_{0,0} = \begin{pmatrix} D_0 \oplus H_0 & I_{\bar{W}} \otimes H_1 \otimes \boldsymbol{\tau} \\ I_a \otimes T_0 & D_0 \oplus H \oplus T \end{pmatrix},$$

$$Q_{0,k} = \begin{pmatrix} D_k \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} & O & O \\ O & D_k \otimes I_{\bar{V}} \otimes I_R \otimes \boldsymbol{\beta}^{(2)} & O \end{pmatrix},\ k \geq 1,$$

$$Q_{1,0} = \begin{pmatrix} (I_a \otimes \boldsymbol{S}_0^{(1)} & O \\ O & I_{aR} \otimes \boldsymbol{S}_0^{(2)} \\ O & O \end{pmatrix},$$

$$Q_0 = \begin{pmatrix} I_a \otimes \boldsymbol{S}_0^{(1)} \boldsymbol{\beta}^{(1)} & O & O \\ O & I_{aR} \otimes \boldsymbol{S}_0^{(2)} \boldsymbol{\beta}^{(2)} & O \\ O & O & O \end{pmatrix},$$

$$Q_1 =$$

$$= \begin{pmatrix} (D_0 \oplus H_0 \oplus S^{(1)} & O & I_{\bar{W}} \otimes H_1 \otimes \mathbf{e}_{M^{(1)}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\alpha} \\ I_a \otimes \boldsymbol{T}_0 \otimes \mathbf{e}_{M^{(2)}} \otimes \boldsymbol{\beta}^{(1)} & D_0 \oplus H \oplus T \oplus S^{(2)} & O \\ I_a \otimes \boldsymbol{T}_0 \otimes \mathbf{e}_L \otimes \boldsymbol{\beta}^{(1)} & I_a \otimes I_R \otimes \boldsymbol{A}_0 \otimes \boldsymbol{\beta}^{(2)} & D_0 \oplus H \oplus T \oplus A \end{pmatrix},$$

$$Q_k = diag\{D_{k-1} \otimes I_{\bar{V}} \otimes I_{M^{(1)}},\ D_{k-1} \otimes I_{\bar{V}} \otimes I_R \otimes I_{M^{(2)}},\ D_{k-1} \otimes I_{\bar{V}} \otimes I_R \otimes I_L\},\ , k \geq 2,$$

где $H = H_0 + H_1$.

Доказательство леммы проводится путем анализа поведения цепи на бесконечно малом интервале времени.

Следствие 6.6. *Цепь Маркова* $\xi_t, t \geq 0$, *принадлежит классу квазитеплицевых цепей Маркова с непрерывным временем.*

Доказательство следует из вида генератора, заданного леммой 6.1, и определения квазитеплицевых цепей Маркова, данного в [150].

В дальнейшем нам понадобятся выражения для производящих функций $\tilde{Q}(z) = \sum\limits_{k=1}^{\infty} Q_{0,k} z^k$ и $Q(z) = \sum\limits_{k=0}^{\infty} Q_k z^k,\ |z| \leq 1$.

Следствие 6.7. Матричные производящие функции $\tilde{Q}(z),\ Q(z)$ имеют следующий вид:

$$\tilde{Q}(z) =$$

$$= \begin{pmatrix} (D(z) - D_0) \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} & O & O \\ O & (D(z) - D_0) \otimes I_{\bar{V}_R} \otimes \boldsymbol{\beta}^{(2)} & O \end{pmatrix}, \quad (6.19)$$

$$Q(z) = Q_0 + \mathcal{Q}z + +zdiag\{D(z) \otimes I_{\bar{V}M^{(1)}},\ D(z) \otimes I_{\bar{V}R},\ D(z) \otimes I_{\bar{V}RL}\}, \quad (6.20)$$

где матрица $\mathcal{Q}$ имеет вид

$$\mathcal{Q} = \begin{pmatrix} I_{\bar{W}} \otimes H_0 \oplus S^{(1)} & O & I_{\bar{W}} \otimes H_1 \otimes \mathbf{e}_{M^{(1)}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\alpha} \\ I_a \otimes \boldsymbol{T}_0 \otimes \mathbf{e}_{M^{(2)}} \otimes \boldsymbol{\beta}^{(1)} & I_{\bar{W}} \otimes H \oplus T \oplus S^{(2)} & O \\ I_a \otimes \boldsymbol{T}_0 \otimes \mathbf{e}_L \otimes \boldsymbol{\beta}^{(1)} & I_{aR} \otimes \boldsymbol{A}_0 \otimes \boldsymbol{\beta}^{(2)} & I_{\bar{W}} \otimes H \oplus T \oplus A \end{pmatrix}.$$

6.2.3 Условие существования стационарного режима в системе. Характеристики производительности

Теорема 6.3. *Необходимым и достаточным условием существования стационарного режима в системе является выполнение неравенства*

$$\lambda < \boldsymbol{\pi}_1 \mathbf{S}_0^{(1)} + \boldsymbol{\pi}_2 \mathbf{S}_0^{(2)}, \tag{6.21}$$

где

$$\boldsymbol{\pi}_1 = \mathbf{x}_1(\mathbf{e}_{\bar{V}M^{(1)}}),\ \boldsymbol{\pi}_2 = \mathbf{x}_2(\mathbf{e}_{\bar{V}RM^{(2)}}), \tag{6.22}$$

а вектор $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ *является единственным решением системы линейных алгебраических уравнений*

$$\mathbf{x}\Gamma = 0,\ \mathbf{x}\mathbf{e} = 1. \tag{6.23}$$

Здесь

$$\Gamma = \begin{pmatrix} I_{\bar{V}} \otimes \boldsymbol{S}_0^{(1)}\boldsymbol{\beta}^{(1)} & O & H_1 \otimes \mathbf{e}_{M^{(1)}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\alpha} \\ I_{\bar{V}} \otimes \boldsymbol{T}_0 \otimes \mathbf{e}_{M^{(2)}} \otimes \boldsymbol{\beta}^{(1)} & I_{\bar{V}R} \otimes \boldsymbol{S}_0^{(2)}\boldsymbol{\beta}^{(2)} & O \\ I_{\bar{V}} \otimes \boldsymbol{T}_0 \otimes \mathbf{e}_L \otimes \boldsymbol{\beta}^{(1)} & I_{\bar{V}R} \otimes \boldsymbol{A}_0 \otimes \boldsymbol{\beta}^{(2)} & O \end{pmatrix} +$$

$$+diag\{H_0 \oplus S^{(1)}, H \oplus T \oplus S^{(2)}, H \oplus T \oplus A\}.$$

Доказательство. Доказательство производится путем рассуждений, аналогичных приведенным выше, при доказательстве теоремы 6.1. В данном случае представляем вектор $\mathbf{y}$, присутствующий в соотношениях (6.5)-(6.6), в виде

$$\mathbf{y} = (\boldsymbol{\theta} \otimes \mathbf{x_1}, \boldsymbol{\theta} \otimes \mathbf{x_2}, \boldsymbol{\theta} \otimes \mathbf{x_3},), \tag{6.24}$$

где $\mathbf{x} \stackrel{def}{=} (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ – стохастический вектор, а векторы $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ имеют размерности $\bar{V}M^{(1)}, \bar{V}RM^{(2)}, \bar{V}RL$ соответственно.

Тогда, учитывая, что $\boldsymbol{\theta} \sum_{k=0}^{\infty} D_k = 0$, система линейных алгебраических уравнений (6.24) сводится к виду (6.23).

Далее, подставляя в неравенство (6.21) вектор $\mathbf{y}$ в виде (6.24), выражение для $Q^{'}(1)$, вычисленное с помощью формулы (6.20), и учитывая, что $\boldsymbol{\theta} D'(1)\mathbf{e} = \lambda$, сводим это неравенство к виду

$$\lambda + \mathbf{x}\mathcal{Q}^{-}\mathbf{e} < 0, \tag{6.25}$$

где

$$\mathcal{Q}^- = \begin{pmatrix} H_0 \oplus S^{(1)} & O & H_1 \otimes \mathbf{e}_{M^{(1)}} \otimes \boldsymbol{\tau} \otimes \boldsymbol{\alpha} \\ I_{\bar{V}} \otimes \boldsymbol{T}_0 \otimes \boldsymbol{\beta}^{(1)} & H \oplus T \oplus S^{(2)} & O \\ I_{\bar{V}} \otimes \boldsymbol{T}_0 \otimes \mathbf{e}_L \otimes \boldsymbol{\beta}^{(1)} & I_{\bar{V}} \otimes I_R \otimes \boldsymbol{A}_0 \otimes \boldsymbol{\beta}^{(2)} & H \oplus T \oplus A \end{pmatrix}.$$

Используя соотношения $H\mathbf{e} = (H_0 + H_1)\mathbf{e} = \mathbf{0}$, $T\mathbf{e} + \boldsymbol{T}_0 = \mathbf{0}$, $A\mathbf{e} + \boldsymbol{A}_0 = \mathbf{0}$, сводим неравенство (6.25) к виду

$$\lambda < \mathbf{x}_1(\mathbf{e}_{\bar{V}} \otimes I_{M^{(1)}})\mathbf{S}_0^{(1)} + \mathbf{x}_2(\mathbf{e}_{\bar{V}} \otimes \mathbf{e}_R \otimes I_{M^{(2)}})\mathbf{S}_0^{(2)}.$$

После использования обозначений (6.22) это неравенство приобретает вид (6.21).

□

Замечание 6.3. При физической интерпретации условия эргодичности (6.3) учитываем, что данное условие отражает процесс обслуживания в системе в условиях перегрузки и векторы $\boldsymbol{\pi}_n$, $n = 1, 2$, имеют следующий смысл: компонента $\boldsymbol{\pi}_1(m^{(1)})$ вектора $\boldsymbol{\pi}_1$ есть вероятность того, что прибор 1 исправен и обслуживает заявку на фазе $m^{(1)}$, $m^{(1)} = \overline{1, M^{(1)}}$, компонента $\boldsymbol{\pi}_2(m^{(2)})$ вектора $\boldsymbol{\pi}_2$ есть вероятность того, что прибор 1 на ремонте, переключение закончилось, прибор 2 обслуживает заявку на фазе $m^{(2)}$, $m^{(2)} = \overline{1, M^{(2)}}$. Тогда правая часть неравенства (6.3) выражает суммарную интенсивность выходящего потока обслуженных заявок. Очевидно, что для существования стационарного режима в системе необходимо и достаточно, чтобы интенсивность входного потока λ была меньше интенсивности выходящего потока.

Следствие 6.8. *В случае стационарного пуассоновского потока поломок и экспоненциального распределения времен обслуживания и ремонтов условие существования стационарного режима сводится к следующему неравенству:*

$$\lambda < \pi_1\mu_1 + \pi_2\mu_2, \tag{6.26}$$

где

$$\pi_1 = \frac{\alpha + \tau}{h}[1 + \frac{\alpha}{\tau} + \frac{\alpha + \tau}{h}]^{-1}, \;\; \pi_2 = \frac{\alpha}{\tau}[1 + \frac{\alpha}{\tau} + \frac{\alpha + \tau}{h}]^{-1}. \tag{6.27}$$

Предполагаем, что условие существования стационарного режима, заданное теоремой 6.1, что гарантирует существование стационарных вероятности состояний системы, задаваемых как

$$p_0^{(0)}(\nu,\eta)=\lim_{t\to\infty}P\{i_t=0,n_t=0,\nu_t=\nu,\eta_t=\eta\},\ \nu=\overline{0,W},\eta=\overline{0,V},$$

$$p_0^{(1)}(\nu,\eta)=\lim_{t\to\infty}P\{i_t=0,n_t=1,\nu_t=\nu,\eta_t=\eta,\vartheta_t=\vartheta\},$$

$$\nu=\overline{0,W},\eta=\overline{0,V},\vartheta=\overline{1,R},$$

$$p_i^{(0,0)}\{(\nu,\eta,m^{(1)})\}=\lim_{t\to\infty}P\{i_t=i,n_t=0,r_t=0,\nu_t=\nu,\eta_t=\eta,m_t^{(1)}=m^{(1)}\},$$

$$i>0,\nu=\overline{0,W},\eta=\overline{0,V},m^{(1)}=\overline{1,M^{(1)}},$$

$$p_i^{(1,0)}\{(\nu,\eta,m^{(2)},\vartheta)\}=\lim_{t\to\infty}P\{i_t=i,n_t=1,r_t=0,\nu_t=\nu,\eta_t=\eta,\vartheta_t=\vartheta,$$

$$m_t^{(2)}=m^{(2)}\},i>0,\nu=\overline{0,W},\eta=\overline{0,V},\vartheta=\overline{1,R},m^{(2)}=\overline{1,M^{(2)}},$$

$$p_i^{(1,1)}\{(\nu,\eta,\vartheta,l)\}=\lim_{t\to\infty}P\{i_t=i,n_t=1,r_t=1,\nu_t=\nu,\eta_t=\eta,\vartheta_t=\vartheta,l_t=l\},$$

$$i>0,\nu=\overline{0,W},\eta=\overline{0,V},\vartheta=\overline{1,R},l=\overline{1,L}.$$

Внутри каждой выделенной группы упорядочим вероятности в лексикографическом порядке компонент и сформируем векторы этих вероятностей

$$\mathbf{p}_0^{(0)},\ \mathbf{p}_0^{(1)},\ \mathbf{p}_i^{(0,0)},\ \mathbf{p}_i^{(1,0)},\ \mathbf{p}_i^{(1,1)}.$$

Порядки этих векторов равны, соответственно, a, aR, $aM^{(1)}$, $aRM^{(2)}$, aRL.

Далее сформируем векторы стационарных вероятностей, соответствующих значениям счетной компоненты как

$$\mathbf{p}_0=(\mathbf{p}_0^{(0)},\ \mathbf{p}_0^{(1)}),\ \mathbf{p}_i=(\mathbf{p}_i^{(0,0)},\ \mathbf{p}_i^{(1,0)},\ \mathbf{p}_i^{(1,1)}),\ i>0.$$

Векторы $\mathbf{p}_i, i\geq 0$, образующие стационарное распределение, вычисляются при помощи алгоритма, приведенного выше, в разделе 6.2.3.

Вычислив векторы стационарных вероятностей $\mathbf{p}_i$, $i\geq 0$, можно вычислить также различные характеристики производительности системы. Для их вычисления будет использоваться векторная производящая функция $\mathbf{P}(z)=\sum\limits_{i=1}^{\infty}\mathbf{p}_i z^i$, $|z|\leq 1$. Эта функция, так же как и аналогичная функция в разделе 6.2.4, удовлетворяет уравнению вида (6.11), где производящие функции $Q(z),\tilde{Q}(z)$ вычисляются по формулам (6.1), (6.2) соответственно. По этому уравнению вычисляются векторные факториальные моменты

$\mathbf{P}^{(m)}(1)$, $m \geq 0$, числа запросов в системе. При этом используется формула (6.12).

Приведем формулы для некоторых важных характеристик производительности системы.

- Пропускная способность системы

$$\varrho = \boldsymbol{\pi}_1 \mathbf{S}_0^{(1)} + \boldsymbol{\pi}_2 \mathbf{S}_0^{(2)}.$$

- Среднее число запросов в системе

$$L = P^{(1)}(1)\mathbf{e}.$$

- Дисперсия числа запросов в системе

$$V = P^{(2)}(1)\mathbf{e} + L - L^2.$$

- Вероятность того, что система пуста и прибор 1 в исправном состоянии

$$P_0^{(0)} = \mathbf{p}_0^{(0)}\mathbf{e}.$$

- Вероятность того, что система пуста и прибор 1 в неисправном состоянии

$$P_0^{(1)} = \mathbf{p}_0^{(1)}\mathbf{e}.$$

- Вероятность того, что прибор 1 обслуживает запрос

$$P_0^{(0,0)} = \mathbf{P}(1)\mathrm{diag}\{I_{aM^{(1)}},\, 0_{aR(M^{(2)}+L)}\}\mathbf{e}.$$

- Вероятность того, что прибор 1 в неисправном состоянии, а прибор 2 обслуживает запрос

$$P_0^{(1,0)} = \mathbf{P}(1)\mathrm{diag}\{0_{aM^{(1)}},\, I_{aRM^{(2)}},\, 0_{aRL}\}\mathbf{e}.$$

- Вероятность того, что прибор 1 в неисправном состоянии, и идет переключение с этого прибора на прибор 2

$$P_0^{(1,1)} = \mathbf{P}(1)\mathrm{diag}\{0_{a(M^{(1)}+RM^{(2)})},\, I_{aRL)}\}\mathbf{e}.$$

- Доля времени, в течение котрого прибор 1 находится в исправном состоянии

$$P_0 = \mathbf{p}_0^{(0)}\mathbf{e} + \mathbf{P}(1)\mathrm{diag}\{I_{aM^{(1)}},\, 0_{aR(M^{(2)}+L)}\}\mathbf{e}.$$

- Доля времени, в течение котрого прибор 1 находится в неисправном состоянии

$$P_1 = \mathbf{p}_0^{(1)}\mathbf{e} + \mathbf{P}(1)\mathrm{diag}\{0_{aM^{(1)}},\, I_{aR(M^{(2)}+L)}\}\mathbf{e}.$$

Пусть $V(x)$ – функция стационарного распределения времени пребывания произвольного запроса в системе, $v(s) = \int\limits_0^\infty e^{-sx}dV(x),\ Re\ s \geq 0$, – преобразование Лапласа – Стилтьеса этой функции.

Преобразование Лапласа – Стилтьеса $v(s)$ и его производные в нуле, определяющие моменты времени пребывания, можно найти по формулам, приведенным в разделе 6.2.6. В частности, среднее время пребывания в системе вычисляется по формуле (6.18).

6.2.4 Случай системы с экспоненциальным видом описывающих ее распределений

В данном разделе исследуется система массового обслуживания, которая является частным случаем системы, исследованной выше, в подразделах 6.3.1-6.3.3. Здесь предполагается, что входной поток заявок и поток поломок являются стационарными пуассоновскими, а времена обслуживания на основном и резервном приборах и время ремонта основного прибора распределены по экспоненциальному закону. Кроме того, считаем, что переключение с основного прибора на резервный (с прибора 1 на прибор 2) происходит мгновенно. Заметим, что здесь, вследствие свойства отсутствия последействия экспоненциального распределения, можем считать, что при переходе с прибора на прибор заявка не начинает обслуживаться сначала, а продолжает свое обслуживание с другой скоростью.

Особенностью исследования, выполненного в данном подразделе, является то, что здесь получены явные формулы для производящих функций стационарного распределения состояний системы и найдено в простом виде преобразование Лапласа-Стилтьеса распределения времени пребывания произвольной заявки в системе.

Процесс функционирования системы описывается неприводимой цепью Маркова с непрерывным временем

$$\xi_t = \{i_t, n_t\},\ t \geq 0,$$

где

i_t - число заявок в системе в момент времени $i_t \geq 0$;

$$n_t = \begin{cases} 1, & \text{если в момент времени } t \text{ основной прибор исправен;} \\ 2, & \text{если основной прибор на ремонте.} \end{cases}$$

Лемма 6.4. Инфинитезимальный генератор цепи ξ_t можно представить в блочном виде

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \dots \\ Q_{-1} & Q_0 & Q_1 & O & \dots \\ O & Q_{-1} & Q_0 & Q_1 & \dots \\ O & O & Q_{-1} & Q_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где

$$Q_{0,0} = \begin{pmatrix} -(\lambda + h) & h \\ \tau & -(\lambda + \tau) \end{pmatrix},\ Q_{0,1} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix},$$

$$Q_{-1} = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix},\ Q_0 = \begin{pmatrix} -(\lambda + \mu_1 + h) & h \\ \tau & -(\lambda + \mu_2 + \tau) \end{pmatrix},\ Q_1 = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}.$$

Из вида генератора следует, что исследуемая цепь принадлежит классу квазитеплицевых цепей Маркова, см. [150].

Теорема 6.4. Необходимым и достаточным условием существования стационарного распределения в рассматриваемой системе является выполнение неравенства

$$\lambda < \frac{\tau}{h + \tau}\mu_1 + \frac{h}{h + \tau}\mu_2. \tag{6.28}$$

Доказательство может быть выполнено аналогично доказательству теоремы 6.3. Более простой путь состоит в использовании следствия 6.8. Искомое условие следует из (6.26)-(6.27) при $\alpha \to \infty$. В этом случае рассматриваемая система эквивалентна системе, для которой сформулировано следствие 6.8. □

Далее будем предполагать, что условие (6.28) выполняется. Обозначим стационарные вероятности состояний цепи Маркова ξ_t как

$$\alpha_i = \lim_{t \to \infty} P\{i_t = i, n_t = 1\},\ \beta_i = \lim_{t \to \infty} P\{i_t = i, n_t = 2\},\ i \geq 0.$$

Уравнения равновесия для этих вероятностей запишутся следующим образом:

$$\alpha_0 (\lambda + h) = \beta_0 \tau + \alpha_0 \mu_1, \tag{6.29}$$

$$\beta_0 (\lambda + \tau) = \beta_0 \mu_2 + \alpha_0 h, \tag{6.30}$$

$$\alpha_i (\lambda + h + \mu_1) = \alpha_{i+1} \mu_1 + \beta_i \tau + \alpha_{i-1} \lambda,\ i > 0, \tag{6.31}$$

$$\beta_i (\lambda + \tau + \mu_2) = \beta_{i+1} \mu_2 + \alpha_{i+1} h + \beta_{i-1} \lambda,\ i > 0. \tag{6.32}$$

Уравнения (6.29)-(6.32) вместе с уравнением нормировки

$$\sum_{i=0}^{\infty} (\alpha_i + \beta_i) = 1 \tag{6.33}$$

определяют единственное стационарное распределение цепи Маркова ξ_t.

Для нахождения стационарных вероятностей воспользуемся методом производящих функций. Введем производящие функции

$$\alpha (z) = \sum_{i=0}^{\infty} \alpha_i z^i,\ \beta (z) = \sum_{i=0}^{\infty} \beta_i z^i,\ |z| \le 1.$$

Теорема 6.5. Производящие функции $\alpha(z)$ и $\beta(z)$ имеют следующий вид:

$$\alpha (z) = \frac{\mu_1 \alpha_0 (1 - z) - \tau \beta (z) z}{\lambda z^2 - (\lambda + \mu_1 + h) z + \mu_1}, \tag{6.34}$$

$$\beta(z) = \frac{\alpha_0 h \mu_1 z - \beta_0 \mu_2 \left(\lambda z^2 - z (h + \lambda + \mu_1) + \mu_1\right)}{M_3(z)}, \tag{6.35}$$

где

$$M_3(z) = \lambda^2 z^3 - (h\lambda + \lambda^2 + \lambda\mu_1 + \lambda\mu_2 + \lambda\tau) z^2 +$$
$$+ (h\mu_2 + \lambda\mu_1 + \lambda\mu_2 + \mu_1\mu_2 + \mu_1\tau) z - \mu_1\mu_2,$$

вероятности α_0 и β_0 выражаются следующим образом

$$\alpha_0 = \frac{\left(\lambda\sigma^2 - (\lambda + \mu_1 + h)\sigma + \mu_1\right) (h\mu_2 - \lambda(h + \tau) + \mu_1\tau)}{\mu_1 (1 - \sigma) (h + \tau) (\mu_1 - \lambda\sigma)}, \tag{6.36}$$

$$\beta_0 = \frac{\alpha_0 h \mu_1 \sigma}{\lambda\sigma^2 - (\lambda + \mu_1 + h)\sigma + \mu_1}, \tag{6.37}$$

а σ есть единственный действительный корень уравнения

$$M_3(z) = 0$$

в области $|z| < 1$.

Доказательство. Легко проверить, что система уравнений (6.29)-(6.31) эквивалентна уравнению (6.34), а система (6.30)-(6.32) – уравнению (6.35) для производящих функций $\alpha(z)$, $\beta(z)$.

Теперь, для того чтобы получить производящие функции в явном виде, требуется найти вероятности α_0, β_0. При нахождении этих неизвестных вероятностей будем использовать то обстоятельство, что при выполнении неравенства (6.28) производящие функции $\alpha(z), \beta(z)$ являются аналитическими внутри единичного круга $|z| < 1$.

Рассмотрим функцию $\beta(z)$. Легко проверить, что числитель и знаменатель соответствующей дроби делятся на $z - 1$. Поделив, получим (6.35) выражения (6.35) для этой функции:

Можно показать, что многочлен третьей степени, стоящий в знаменателе (6.35), имеет единственный и действительный корень внутри единичного круга. Обозначим этот корень как σ. Тогда, чтобы функция $\beta(z)$ была аналитической в области $|z| < 1$, необходимо, чтобы числитель (6.35) в точке $z = \sigma$ также обращался в нуль. Используя равенство нулю числителя, получим следующее выражение для β_0 :

$$\beta_0 = \frac{\alpha_0 h \mu_1 \sigma}{\lambda\sigma^2 - (\lambda + \mu_1 + h)\sigma + \mu_1}. \tag{6.38}$$

Из (6.34) следует соотношение

$$h \sum_{i=0}^{\infty} \alpha_i = \tau \sum_{i=0}^{\infty} \beta_i.$$

Используя это соотношение в уравнении нормировки (6.33), получаем

$$\beta(1) = \sum_{i=0}^{\infty} \beta_i = \frac{h}{\tau + h}.$$

Подставляя это выражение для $\beta(1)$ и выражение (6.38) для β_0 в (6.35), где полагаем $z = 1$, получим уравнение для α_0, которое имеет решение (6.36).

Подставляя (6.36) в (6.38), получим выражение (6.37) для β_0. □

Найдем теперь преобразование Лапласа – Стилтьеса стационарного распределения времени пребывания произвольного запроса в системе.

Особенностью данной системы является то, что время пребывания запроса в системе существенно зависит от поступления поломок и продолжительности ремонтов прибора 1. В течение этого времени запрос может многократно переходить с прибора 1 на прибор 2 и наоборот.

Пусть $V(t)$ – функция распределения времени пребывания запроса в системе при работе в стационарном режиме, $v(s) = \int\limits_0^\infty e^{-st} dV(t), Res \geq 0$, – ее преобразование Лапласа – Стилтьеса.

Теорема 6.6. Преобразование Лапласа – Стилтьеса распределения времени пребывания запроса в системе вычисляется по формуле

$$v\left(s\right) = \sum_{i=0}^{\infty} \mathbf{p}_i \mathbf{v}_i^T(s), \tag{6.39}$$

где

$$\mathbf{p}_i = (\alpha_i, \beta_i),\ i \geq 0,$$

а вектор-строки $\mathbf{v}_i(s)$ вычисляются по формуле

$$\mathbf{v}_i\left(s\right) = \mathbf{v}_0\left(s\right)\left[A\left(s\right)\left[I - B\left(s\right)\right]^{-1}\right]^i,\ i \geq 1, \tag{6.40}$$

где вектор-строка $\mathbf{v}_0(s)$ имеет вид

$$\mathbf{v}_0(s) = \left(\frac{\mu_1\left(\mu_2 + \tau + s\right) + h\mu_2}{\left(\mu_1 + h + s\right)\left(\mu_2 + \tau + s\right) - h\tau},\right.$$

$$\left.\frac{1}{\mu_2 + \tau + s}[\mu_2 + \tau\,\frac{\mu_1\left(\mu_2 + \tau + s\right) + h\mu_2}{\left(\mu_1 + h + s\right)\left(\mu_2 + \tau + s\right) - h\tau}]\right), \tag{6.41}$$

а матрицы $A(s)$ и $B(s)$ определяются как

$$A(s) = \begin{pmatrix} \frac{\mu_1}{\mu_1 + h + s} & 0 \\ 0 & \frac{\mu_2}{\mu_2 + \tau + s} \end{pmatrix},\quad B(s) = \begin{pmatrix} 0 & \frac{\tau}{\mu_2 + \tau + s} \\ \frac{h}{\mu_1 + h + s} & 0 \end{pmatrix}. \tag{6.42}$$

Доказательство. Пусть $v_i(s, n)$ – преобразование Лапласа – Стилтьеса распределения времени пребывания в системе запроса, заставшего систему в состоянии (i, n), т.е. с i запросами и с прибором 1 в состоянии n, $i \geq 0, n = 1, 2$.

Используя метод введения дополнительного события, составим следующую систему линейных алгебраических уравнений для $v_0(s, 1)$ и $v_0(s, 2)$:

$$v_0\left(s, 1\right) = \int\limits_0^\infty e^{-st} e^{-ht} e^{-\mu_1 t}\,\mu_1\,dt + \int\limits_0^\infty e^{-st} e^{-ht} e^{-\mu_1 t}\,h\,dt\,v_0\left(s, 2\right),$$

$$v_0\left(s, 2\right) = \int\limits_0^\infty e^{-st} e^{-\tau t} e^{-\mu_2 t}\,\mu_2\,dt + \int\limits_0^\infty e^{-\tau t} e^{-st} e^{-\mu_2 t}\,\tau\,dt\,v_0\left(s, 1\right).$$

В результате интегрирования получаем систему

$$v_0(s,1) = \frac{\mu_1}{\mu_1+h+s} + \frac{h}{\mu_1+h+s} v_0(s,2),$$
$$v_0(s,2) = \frac{\mu_2}{\mu_2+\tau+s} + \frac{\tau}{\mu_2+\tau+s} v_0(s,1).$$

Решая эту систему, получим явные выражения для функций $v_0(s,1)$ и $v_i(s,2)$

$$v_0(s,1) = \frac{\mu_1(\mu_2+\tau+s)+h\mu_2}{(\mu_1+h+s)(\mu_2+\tau+s)-h\tau}, \tag{6.43}$$

$$v_0(s,2) = \frac{1}{\mu_2+\tau+s}\left[\mu_2 + \tau\frac{\mu_1(\mu_2+\tau+s)+h\mu_2}{(\mu_1+h+s)(\mu_2+\tau+s)-h\tau}\right]. \tag{6.44}$$

Снова используя метод введения дополнительного события, составим систему линейных алгебраических уравнений для функций $v_i(s,1)$ и $v_i(s,2)$ при $i \geq 1$.

$$v_i(s,1) = \frac{\mu_1}{\mu_1+h+s} v_{i-1}(s,1) + \frac{h}{\mu_1+h+s} v_i(s,2),$$
$$v_i(s,2) = \frac{\mu_2}{\mu_2+\tau+s} v_{i-1}(s,1) + \frac{\tau}{\mu_2+\tau+s} v_i(s,1), i \geq 1. \tag{6.45}$$

Введем в рассмотрение векторы

$$\mathbf{v}_i(s) = (v_i(s,1), v_i(s,2)), i \geq 0.$$

Тогда систему (6.45) можно записать в следующем виде:

$$\mathbf{v}_i(s) = \mathbf{v}_{i-1}(s)\ A(s) + \mathbf{v}_i(s)\ B(s), i \geq 1, \tag{6.46}$$

где матрицы $A(s), B(s)$ имеют вид (6.42).

Из (6.46) получаем рекуррентную формулу для вычисления векторов $v_i(s)$

$$\mathbf{v}_i(s) = \mathbf{v}_{i-1}(s) A(s) [I - B(s)]^{-1}, i \geq 1,$$

используя которую, выразим все векторы $\mathbf{v}_i(s), i > 0$, через известный (см. формулы (6.43)-(6.44)) вектор $\mathbf{v}_0(s)$. В результате получаем формулу (6.40).

Тогда по формуле полной вероятности искомое преобразование Лапласа – Стилтьеса $\mathbf{v}(s)$ запишется в виде (6.39).

Следствие 6.9. Среднее время пребывания запроса в системе вычисляется по формуле

$$Ev = -\mathbf{v}'(0) = -\sum_{i=0}^{\infty} \mathbf{p}_i \frac{d\mathbf{v}_i^T(s)}{ds}|_{s=0}. \tag{6.47}$$

6.3 Методы и алгоритмы для оценки характеристик производительности двухлинейной системы с ненадежными обслуживающими приборами

В данном разделе исследуется двухлинейная система с ненадежными неоднородными обслуживающими приборами, которая может использоваться при математическом моделировании гибридной сети связи, состоящей из ненадежных FSO (Free Space Optics) -канала и миллиметрового радиоканала (71-76 GHz, 81-86 GHz). Предполагается, что погодные условия, неблагоприятные для одного из каналов, не являются таковыми для другого канала. FSO-канал не может передавать данные в условиях тумана или пасмурной погоды, а миллиметровый радиоканал не может осуществлять передачу во время осадков (дождь, снег и т.д.). Таким образом, гибридная система связи способна передавать данные практически при любых погодных условиях, используя тот или иной канал связи.

6.3.1 Описание системы

Рассматривается двухлинейная СМО с ненадежными обслуживающими приборами. Запросы поступают в систему в MAP-потоке, который является ординарным аналогом $BMAP$. Заявка, приходящая на обслуживание, когда оба прибора являются свободными и исправными, немедленно начинает обслуживание на обоих приборах. Если же приборы заняты в момент прихода запроса, заявка становится в очередь, длина которой неограничена, и выбирается на обслуживание позже согласно стратегии FIFO (первым пришел – первым обслужен).

Полагаем, что процесс обслуживания на k-м, $k = 1, 2$, приборе имеет PH-распределение с неприводимым представлением $(\boldsymbol{\beta}_k, S_k)$ и управляющим процессом $m_t^{(k)}$, $t \geq 0$, с пространством состояний $\{1, \ldots, M_k, M_k+1\}$, где состояние $M_k + 1$ является поглощающим. Среднее время обслуживания на k-м приборе вычисляется по формуле $b_1^{(k)} = \boldsymbol{\beta}_k(-S_k)^{-1}\mathbf{e}$, интенсивность обслуживания $\mu_k = (b_1^{(k)})^{-1}$, $k = 1, 2$.

Оба прибора являются ненадежными. Как было сказано выше, поломки приборов зависят от погодных условий. Считаем, что поломки приходят в MAP-потоке с управляющим процессом η_t, $t \geq 0$, принимающем значения в множестве $\{0, 1, ..., V\}$. Этот MAP задается $(V + 1) \times (V + 1)$

матрицами H_0 и H_1. Интенсивность потока поломок задается формулой $h = \boldsymbol{\vartheta} H_1 \mathbf{e}$, где вектор-строка $\boldsymbol{\vartheta}$ является единственным решением системы уравнений $\boldsymbol{\vartheta}(H_0 + H_1) = \mathbf{0}$, $\boldsymbol{\vartheta}\mathbf{e} = 1$. Поступившая поломка (неблагоприятные погодные условия) направляется на один из приборов. Любую поломку с вероятностью p классифицируем как дождь, и тогда она направляется на прибор 1, а с дополнительной вероятностью $1 - p$ – как туман, и тогда она направляется на прибор 2.

Сразу после поступления поломки на k-й прибор на нем начинается ремонт (период неблагоприятных погодных условий). Время, необходимое для ремонта k-го прибора, имеет PH-распределение с неприводимым представлением $(\boldsymbol{\tau}^{(k)}, T^{(k)})$ и интенсивностью τ_k, $k = 1, 2$. Предполагаем, что поломка, поступившая, когда какой-либо из приборов еще не восстановился после предыдущей поломки, игнорируется, т.е. не может быть ситуаций, когда оба прибора не работают по причине поломки.

При взятии заявки на обслуживание, если оба прибора исправны, они одновременно начинают ее обслуживание. В момент окончания обслуживания данной заявки каким-либо прибором его обслуживание другим прибором немедленно прекращается. Если при взятии запроса на обслуживание только k-й прибор, $k = 1, 2$, исправен, он проводит обслуживание запроса. Если во время этого обслуживания другой прибор восстанавливается, он начинает обслуживание этой же заявки заново.

Как отмечалось выше, приход поломки вызывает прекращение обслуживания заявки (если таковое имеет место) прибором, на который направляется поломка. Этот прибор уходит на ремонт, в то время как другой прибор продолжает обслуживание заявки. Если во время этого обслуживания другой прибор восстанавливается, он снова начинает обслуживание этой же заявки, но заново. Заметим, что в случае экспоненциальных распределений времен обслуживания можно считать, что восстановившийся прибор присоединяется к обслуживанию упомянутой заявки, начиная с текущего момента ее обслуживания.

6.3.2 Цепь Маркова, описывающая функционирование системы

Пусть в момент времени t

• i_t – число заявок в системе, $i_t \geq 0$,

• $n_t = \begin{cases} 0, & \text{если оба прибора исправны;} \\ 1, & \text{если прибор 1 на ремонте;} \\ 2, & \text{если прибор 2 на ремонте;} \end{cases}$

• $m_t^{(j)}$ – состояние управляющего процесса -обслуживания на j-м занятом приборе, $j = 1, 2, m_t^{(j)} = \overline{1, M_j}$;

• $r_t^{(j)}$ – состояние управляющего процесса -времени ремонта на приборе j, $r_t^{(j)} = \overline{1, R_j}$;

• ν_t и η_t – состояния управляющих процессов MAP потока запросов и MAP потока поломок соответственно, $\nu_t = \overline{0, W}$, $\eta_t = \overline{0, V}$.

Процесс функционирования системы описывается неприводимой цепью Маркова $\xi_t, t \geq 0$, с пространством состояний

$$X = \{\{(i, 0, \nu, \eta),\ i = 0, n = 0, \nu = \overline{0, W}, \eta = \overline{0, V}\} \bigcup$$

$$\{(i, 1, \nu, \eta, r^{(1)}),\ i = 0, n = 1,\ \nu = \overline{0, W}, \eta = \overline{0, V}, r^{(1)} = \overline{1, R_1}\} \bigcup$$

$$\{(i, 2, \nu, \eta, r^{(2)}),\ i = 0, n = 2,\ \nu = \overline{0, W}, \eta = \overline{0, V}, r^{(1)} = \overline{1, R_2}\} \bigcup$$

$$\{(i, 0, \nu, \eta, m^{(1)},\ m^{(2)}),\ i > 0, n = 0, \nu = \overline{0, W}, \eta = \overline{0, V},\ m^{(k)} = \overline{1, M_k}, k = 1, 2\}$$

$$\bigcup \{(i, 1, \nu, \eta, m^{(2)}, r^{(1)}),\ i > 0, n = 1,\ \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(2)} = \overline{1, M_2}, r^{(1)} =$$

$$\overline{1, R_1}\} \bigcup \{(i, 2, \nu, \eta, m^{(1)}, r^{(2)}),\ i > 0, n = 2,\ \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(1)} = \overline{1, M_1},$$

$$r^{(1)} = \overline{1, R_2}\}.$$

Далее будем предполагать, что состояния цепи $\xi_t, t \geq 0$, внутри каждого из приведенных подмножеств упорядочены в лексикографическом порядке и таким образом упорядоченные подмножества упорядочены в том порядке, в котором они перечислены выше. Обозначим через $Q_{i,j}$ матрицу интенсивностей переходов цепи из состояний, соответствующих значению i первой (счетной) компоненты в состояния, соответствующие значению j этой компоненты, $i, j \geq 0$.

Лемма 6.5 Инфинитезимальный генератор Q цепи Маркова ξ_t, $t \geq 0$, имеет блочную трехдиагональную структуру

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \cdots \\ Q_{1,0} & Q_1 & Q_2 & O & \cdots \\ O & Q_0 & Q_1 & Q_2 & \cdots \\ O & O & Q_0 & Q_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

где ненулевые блоки имеют следующий вид:

$$Q_{0,0} = \begin{pmatrix} D_0 \oplus H_0 & I_{\bar{W}} \otimes pH_1 \otimes \boldsymbol{\tau}^{(1)} & I_{\bar{W}} \otimes (1-p)H_1 \otimes \boldsymbol{\tau}^{(2)} \\ I_a \otimes \boldsymbol{T}_0^{(1)} & D_0 \oplus H \oplus T^{(1)} & O \\ I_a \otimes \boldsymbol{T}_0^{(2)} & O & D_0 \oplus H \oplus T^{(2)} \end{pmatrix},$$

$$Q_{0,1} =$$

$$\begin{pmatrix} D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes \boldsymbol{\beta}^{(2)} & O & O \\ O & D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(2)} \otimes I_{R_1} & O \\ O & O & D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes I_{R_2} \end{pmatrix},$$

$$Q_{1,0} = \begin{pmatrix} I_a \otimes \tilde{S}_0 & O & O \\ O & I_a \otimes \boldsymbol{S}_0^{(2)} \otimes I_{R_1} & O \\ O & O & I_a \otimes \boldsymbol{S}_0^{(1)} \otimes I_{R_2} \end{pmatrix},$$

$$Q_0 = \begin{pmatrix} I_a \otimes \tilde{S}_0(\boldsymbol{\beta}^{(1)} \otimes \boldsymbol{\beta}^{(2)}) & O & O \\ O & I_a \otimes \boldsymbol{S}_0^{(2)}\boldsymbol{\beta}^{(2)} \otimes I_{R_1} & O \\ O & O & I_a \otimes \boldsymbol{S}_0^{(1)}\boldsymbol{\beta}^{(1)} \otimes I_{R_2} \end{pmatrix},$$

$$Q_1 =$$

$$= \begin{pmatrix} D_0 \oplus H_0 \oplus S^{(1)} \oplus S^{(2)} & I_{\bar{W}} \otimes pH_1 \otimes \mathbf{e}_{M_1} \otimes I_{M_2} \otimes \boldsymbol{\tau}_1 & I_{\bar{W}} \otimes \bar{p}H_1 \otimes I_{M_1} \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\tau}_2 \\ I_a \otimes \beta^{(1)} \otimes I_{M_2} \otimes \boldsymbol{T}_0^{(1)} & D_0 \oplus H \oplus S^{(2)} \oplus T^{(1)} & O \\ I_a \otimes I_{M_1} \otimes \beta^{(2)} \otimes \boldsymbol{T}_0^{(2)} & O & D_0 \oplus H \oplus S^{(1)} \oplus T^{(2)} \end{pmatrix},$$

$$Q_2 = \begin{pmatrix} D_1 \otimes I_{\bar{V}M_1M_2} & O & O \\ O & D_1 \otimes I_{\bar{V}M_2R_1} & O \\ O & O & D_1 \otimes I_{\bar{V}M_1R_2} \end{pmatrix},$$

где $\bar{p} = 1-p$, $H = H_0 + H_1$, $\tilde{S}_0 = -(S_1 \oplus S_2)\mathbf{e}$, $\bar{W} = W+1$, $\bar{V} = V+1$, $a = \bar{W}\bar{V}$.

Доказательство леммы проводится путем анализа поведения цепи на бесконечно малом интервале времени.

Следствие 6.10. Цепь Маркова $\xi_t, t \geq 0$, принадлежит классу векторных процессов гибели и размножения.

Доказательство следует из вида генератора, заданного леммой 6.4, и определения векторных процессов гибели и размножения, см., например, [163].

6.3.3 Условие существования стационарного распределения. Характеристики производительности

Теорема 6.7. Необходимым и достаточным условием существования стационарного режима в системе является выполнение неравенства

$$\lambda < \boldsymbol{\delta}_0 \tilde{S}_0 + \boldsymbol{\delta}_1 (\boldsymbol{S}_0^{(2)} \otimes \mathbf{e}_{R_1}) + \boldsymbol{\delta}_2 (\boldsymbol{S}_0^{(1)} \otimes \mathbf{e}_{R_2}), \tag{6.48}$$

где вектор $\boldsymbol{\delta} = (\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ является единственным решением системы линейных алгебраических уравнений

$$\boldsymbol{\delta}\Gamma = 0, \ \boldsymbol{\delta}\mathbf{e} = 1, \tag{6.49}$$

где

$$\Gamma = \begin{pmatrix} I_{\bar{V}} \otimes \tilde{S}_0(\boldsymbol{\beta}^{(1)} \otimes \boldsymbol{\beta}^{(2)}) & pH_1\mathbf{e}_{M_1} \otimes I_{M_2} \otimes \boldsymbol{\tau}_1 & (1-p)H_1 \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\tau}_2 \\ I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes I_{M_2} \otimes \boldsymbol{T}_0^{(1)} & I_{\bar{V}} \otimes \boldsymbol{S}_0^{(2)}\boldsymbol{\beta}^{(2)} \otimes I & O \\ I_{\bar{V}M_1} \otimes \boldsymbol{\beta}_2 \otimes \boldsymbol{T}_0^{(2)} & O & I_{\bar{V}} \otimes \boldsymbol{S}_0^{(1)}\boldsymbol{\beta}^{(1)} \otimes I \end{pmatrix} +$$

$$+diag\{H_0 \oplus S^{(1)} \oplus S^{(2)}, \ H \oplus S^{(2)} \oplus T^{(1)}, \ H \oplus S^{(1)} \oplus T^{(2)}\}$$

Доказательство. Условие существования стационарного режима в системе находится как условие существования стационарного распределения цепи Маркова $\xi_t, t \geq 0$, описывающей процесс функционирования системы. В силу неприводимости цепи условие существования ее стационарного распределения совпадает с условием эргодичности. Согласно [163], необходимое и достаточное условие эргодичности рассматриваемой цепи выражается неравенством

$$\mathbf{x}Q_2\mathbf{e} < \mathbf{x}Q_0\mathbf{e}, \tag{6.50}$$

где вектор $\mathbf{x}$ является единственным решением системы

$$\mathbf{x}(Q_0 + Q_1 + Q_2) = \mathbf{0}, \tag{6.51}$$

$$\mathbf{x}\mathbf{e} = 1. \tag{6.52}$$

Представим вектор $\mathbf{x}$ в виде

$$\mathbf{x} = (\boldsymbol{\theta} \otimes \boldsymbol{\delta}_0, \boldsymbol{\theta} \otimes \boldsymbol{\delta}_1, \boldsymbol{\theta} \otimes \boldsymbol{\delta}_3), \tag{6.53}$$

где векторы $\boldsymbol{\delta}_0$, $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_3$ имеют порядки $\bar{V}M_1M_2$, $\bar{V}M_2R_1$ и $\bar{V}M_1R_2$ соответственно.

Подставив вектор $\mathbf{x}$ в виде (6.53) в неравенство (6.50) и систему (6.51)-(6.52) после некоторых алгебраических преобразований получим неравенство (6.48) и систему (6.49) соответственно. □

Следствие 6.11. В случае стационарного пуассоновского потока поломок и экспоненциальных распределений времен обслуживания и ремонтов необходимое и достаточное условие существования стационарного режима в системе имеет вид

$$\lambda < \frac{\tau_1\tau_2}{\tau_1\tau_1 + ph\tau_2 + (1-p)h\tau_1}(\mu_1 + \mu_2 + \frac{ph}{\tau_1}\mu_2 + \frac{(1-p)h}{\tau_2}\mu_1). \qquad (6.54)$$

Доказательство. Учитывая, что в рассматриваемом случае $S^{(n)} = -\boldsymbol{S}_0^{(n)} = \mu_n, T^{(n)} = -\boldsymbol{T}_0^{(n)} = \tau_n, n = 1, 2$, неравенство (6.48) сводится к виду

$$\lambda < \delta_0(\mu_1 + \mu_2) + \delta_1\mu_2 + \delta_2\mu_1, \qquad (6.55)$$

а система (6.49) сводится к виду

$$\begin{cases} \delta_0 h - \delta_1\tau_1 - \delta_0\tau_2 = 0, \\ \delta_0 ph - \delta_1\tau_1 = 0, \\ \delta_0(1-p)h - \delta_2\tau_2 = 0, \\ \delta_0 + \delta_1 + \delta_2 = 1. \end{cases}$$

Решение последней системы имеет следующий вид:

$$\delta_0 = \frac{\tau_1\tau_2}{\tau_1\tau_1 + ph\tau_2 + (1-p)h\tau_1},\ \delta_1 = \delta_0\frac{ph}{\tau_1},\ \delta_2 = \delta_0\frac{(1-p)h}{\tau_2}. \qquad (6.56)$$

Подставляя (6.56) в (6.55), убеждаемся, что условие эргодичности имеет вид (6.54). □

Замечание 6.4. Неравенство (6.54) легко интерпретировать, если заметить, что величины $\delta_0, \delta_1, \delta_2$ есть вероятности того, что в условиях перегрузки системы либо оба прибора исправны, либо первый находится на ремонте либо второй находится на ремонте, соответственно. В таком случае правая часть (6.54) есть интенсивность обслуживания в данной системе в условиях перегрузки. Условие существования стационарного режима заключается в том, что эта интенсивность должна быть больше, чем интенсивность входного потока λ.

Заметим, что аналогично интерпретируется условие (6.48) существования стационарного режима в случае MAP-потока поломок и PH-распределений времен обслуживания и ремонтов.

Далее будем предполагать, что условие (6.48) существования стационарного режима выполняется. Введем обозначения для стационарных вероятностей состояний системы:

$$p_0^{(0)} = \lim_{t\to\infty} P\{i_t = 0, n_t = 0, \nu_t = \nu, \eta_t = \eta\},$$

$$p_0^{(1)} = \lim_{t\to\infty} P\{i_t = 0, n_t = 1, \nu_t = \nu, \eta_t = \eta,\ r_t^{(1)} = r^{(1)}\},$$

$$p_0^{(2)} = \lim_{t\to\infty} P\{i_t = 0, n_t = 2, \nu_t = \nu, \eta_t = \eta,\ r_t^{(2)} = r^{(2)}\},$$

$$p_i^{(0)} = \lim_{t\to\infty} P\{i_t = i, n_t = 0, \nu_t = \nu, \eta_t = \eta, m_t^{(1)} = m^{(1)},\ m_t^{(2)} = m^{(2)}\},$$

$$p_i^{(1)} = \lim_{t\to\infty} P\{i_t = i, n_t = 1, \nu_t = \nu, \eta_t = \eta, m_t^{(2)} = m^{(2)},\ r_t^{(1)} = r^{(1)}\},$$

$$p_i^{(2)} = \lim_{t\to\infty} P\{i_t = i, n_t = 2, \nu_t = \nu, \eta_t = \eta, m_t^{(1)} = m^{(1)},\ r_t^{(2)} = r^{(2)}\},$$

$$i > 0, \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(k)} = \overline{1, M_k}, r^{(k)} = \overline{1, R_k},\ k = 1, 2.$$

Внутри каждой из выделенных групп упорядочим вероятности в лексикографическом порядке компонент и сформируем векторы-строки этих вероятностей

$$\mathbf{p}_i^{(0)},\ \mathbf{p}_i^{(1)},\ \mathbf{p}_i^{(2)},\ i \geq 0.$$

Далее сформируем векторы стационарных вероятностей, соответствующих значениям счетной компоненты как

$$\mathbf{p}_i = (\mathbf{p}_i^{(0)},\ \mathbf{p}_i^{(1)},\ \mathbf{p}_i^{(2)}),\ i \geq 0.$$

Векторы $\mathbf{p}_i$, $i \geq 0$, вычисляются по адаптированной версии алгоритма вычисления стационарного распределения квазитеплицевой цепи Маркова, описанного в разделе 6.2.3. Основанием для применения этого алгоритма является то, что векторный процесс гибели и размножения, которым описывается функционирование рассматриваемой системы, является частным случаем квазитеплицевой цепи Маркова. В общем случае генератор имеет блочную верхне-хессенбергову структуру, а в нашем частном случае – блочную трехдиагональную структуру.

Описание алгоритма, адаптированного под наш случай, задается следующим образом.

Алгоритм 6.2. Векторы стационарных вероятностей $\mathbf{p}_i,\ i \geq 0$, вычисляются следующим образом.

1. Вычисляется матрица G как стохастическое решение нелинейного матричного уравнения $\sum\limits_{k=0}^{2} Q_k G^k = O$.
2. Вычисляется матрица G_0 из уравнения

$$Q_{1,0} + Q_1 G_0 + Q_2 G G_0 = O,$$

откуда

$$G_0 = -(Q_1 + Q_2 G)^{-1} Q_{1,0}.$$

3. Вычисляются матрицы $\bar{Q}_{i,l},\ l = i, i+1, i \geq 0$, по формулам

$$\bar{Q}_{i,l} = \begin{cases} Q_{0,0} + Q_{0,1} G_0,\ i = 0,\ l = 0, \\ Q_{0,1},\ i = 0,\ l = 1, \\ Q_0 + Q_1 G_i,\ l = i,\ i \geq 1, \\ Q_1,\ l = i+1,\ , i \geq 1, \end{cases}$$

где $G_i = G,\ i \geq 1$.

4. Вычисляются матрицы $\Phi_i,\ i \geq 0$, по рекуррентным формулам

$$\Phi_0 = I_{a(1+R_1+R_2)}), \Phi_i = \Phi_{i-1} \bar{Q}_{i-1,i} (-\bar{Q}_{i,i})^{-1},\ i \geq 1.$$

5. Вычисляется вектор $\mathbf{p}_0$ как единственное решение системы линейных алгебраических уравнений

$$\mathbf{p}_0 (-\bar{Q}_{0,0}) = \mathbf{0},\ \mathbf{p}_0 (\mathbf{e}_{a(1+R_1+R_2)}) + \sum_{i=1}^{\infty} \Phi_i \mathbf{e} = 1.$$

6. Вычисляются векторы стационарных вероятностей $\mathbf{p}_i,\ i \geq 1$, по формуле

$$\mathbf{p}_i = \mathbf{p}_0 \Phi_i,\ i \geq 1.$$

Существует и другой устойчивый алгоритм (см. раздел 3.2.1) для вычисления стационарного распределения, основанный на матрично-геометрическом представлении векторов $\mathbf{p}_i,\ i > 0$.

Алгоритм 6.3. Векторы стационарных вероятностей $\mathbf{p}_i,\ i \geq 0$, вычисляются следующим образом.

1. Вычисляется субстохастическая матрица R как минимальное неотрицательное решение матричного уравнения

$$R^2 Q_0 + R Q_1 + Q_2 = O.$$

2. Вычисляется вектор $\mathbf{p}_1$ как единственное решение системы линейных алгебраических уравнений

$$\mathbf{p}_1[Q_1 + Q_{1,0}(-Q_{0,0})^{-1}Q_{0,1} + RQ_0] = \mathbf{0},$$
$$\mathbf{p}_1[\mathbf{e} + Q_{1,0}(-Q_{0,0})^{-1}\mathbf{e} + R(I-R)^{-1}\mathbf{e}] = 1.$$

3. Векторы $\mathbf{p}_0$, $\mathbf{p}_i$, $i \geq 2$, вычисляются по формулам

$$\mathbf{p}_0 = \mathbf{p}_1 Q_{1,0}(-Q_{0,0})^{-1}, \quad \mathbf{p}_i = \mathbf{p}_1 R^{i-1},\ i \geq 2.$$

Вычислив векторы стационарных вероятностей $\mathbf{p}_i$, $i \geq 0$, можно вычислить также различные характеристики производительности системы. Ниже приводятся наиболее важные из них.

- Пропускная способность системы

$$\varrho = \boldsymbol{\delta}_0 \tilde{S}_0 + \boldsymbol{\delta}_1[\boldsymbol{S}_0^{(2)} \otimes \mathbf{e}_{R_1}] + \boldsymbol{\delta}_2[\boldsymbol{S}_0^{(1)} \otimes \mathbf{e}_{R_2}].$$

- Среднее число заявок системе $L = \sum\limits_{i=1}^{\infty} i\mathbf{p}_i\mathbf{e}$.
- Дисперсия числа заявок в системе $V = \sum\limits_{i=1}^{\infty} i^2\mathbf{p}_i\mathbf{e} - L^2$.
- Вероятность того, что в системе находится $i, i > 0$, заявок и оба прибора исправны (обслуживают заявку)

$$P_i^{(0)} = \mathbf{p}_i \begin{pmatrix} \mathbf{e}_{aM_1M_2} \\ \mathbf{0}_{a(M_2R_1+M_1R_2} \end{pmatrix}.$$

- Вероятность того, что в системе находится $i, i > 0$, заявок, прибор 1 находится на ремонте, а прибор 2 обслуживает заявку

$$P_i^{(1)} = \mathbf{p}_i \begin{pmatrix} \mathbf{0}_{aM_1M_2} \\ \mathbf{e}_{aM_2R_1} \\ \mathbf{0}_{aM_1R_2} \end{pmatrix}.$$

- Вероятность того, что в системе находится $i, i > 0$, заявок, прибор 2 находится на ремонте, а прибор 1 обслуживает заявку

$$P_i^{(2)} = \mathbf{p}_i \begin{pmatrix} \mathbf{0}_{aM_1M_2} \\ \mathbf{0}_{aM_2R_1} \\ \mathbf{e}_{aM_1R_2} \end{pmatrix}.$$

- Вероятность того, что в произвольный момент времени система не пустая и приборы находятся в состоянии n

$$P^{(n)} = \sum_{i=1}^{\infty} P_i^{(n)},\ n = 0, 1, 2.$$

- Вероятность того, что поступившая заявка найдет в системе $i, i \geq 0$, заявок и оба прибора исправными

$$P_{i,0}^{(arrival)} = \lambda^{-1} \left[\delta_{0,i} \mathbf{p}_0 \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}} \\ O_{a(R_1+R_2)\times\bar{W}} \end{pmatrix} + (1-\delta_{0,i}) \mathbf{p}_i \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_1M_2} \\ O_{a(M_2R_1+M_1R_2)\times\bar{W}} \end{pmatrix} \right] D_1,$$

где $\delta_{i,j}$ – символ Кронекера.

- Вероятность того, что поступившая заявка найдет в системе $i, i \geq 0$, заявок и прибор 1 на ремонте

$$P_{i,1}^{(arrival)} = \lambda^{-1} \left[\delta_{0,i} \mathbf{p}_0 \begin{pmatrix} O_{a\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}R_1} \\ O_{aR_2\times\bar{W}} \end{pmatrix} + (1-\delta_{0,i}) \mathbf{p}_i \begin{pmatrix} O_{aM_1M_2\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_2R_1} \\ O_{aM_1R_2\times\bar{W}} \end{pmatrix} \right] D_1.$$

- Вероятность того, что поступившая заявка найдет в системе i, $i \geq 0$, заявок и прибор 2 на ремонте

$$P_{arrival}^{(2,i)} = \lambda^{-1} \left[\delta_{0,i} \mathbf{p}_0 \begin{pmatrix} O_{a(1+R_1)\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}R_2} \end{pmatrix} + (1-\delta_{0,i}) \mathbf{p}_i \begin{pmatrix} O_{a(M_1M_2+M_2R_1)\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_1R_2} \end{pmatrix} \right] D_1.$$

- Вероятность $P_{break}^{(k)}$ того, что поступившая поломка найдет в системе i, $i > 0$, заявок, оба прибора исправными и будет направлена на k-й прибор

$$P_{break}^{(k)} = h^{-1} \left[\delta_{1,k} p + \delta_{2,k} (1-p) \right] \mathbf{p}_i \begin{pmatrix} \mathbf{e}_{\bar{W}} \otimes I_{\bar{V}} \otimes \mathbf{e}_{M_1M_2} \\ O_{a(M_2R_1+M_1R_2)\times\bar{V}} \end{pmatrix} H_1,\ k = 1, 2.$$

6.4 Методы и алгоритмы для оценки характеристик производительности системы массового обслуживания с двумя ненадежными обслуживающими приборами и резервным прибором, функционирующим в холодном резерве

В данном разделе исследуется двухлинейная система с ненадежными неоднородными обслуживающими приборами и резервным надежным прибором, которая может использоваться при математическом моделировании гибридной сети связи, состоящей из трех каналов: FSO канала, миллиметрового радиоканала и широкополосного радиоканала, функционирующего под управлением протокола IEEE 802.11 и используемого как резервный канал для передачи информации. Основные каналы – FSO и миллиметрового радиоканала – подвержены влиянию погодных условий и могут выходить из строя при неблагоприятных условиях. FSO канал не может передавать данные в условиях плохой видимости (тумана или пасмурной погоды), а миллиметровый радиоканал не может осуществлять передачу во время осадков (дождь, снег и т.д.). В случае, когда выходят из строя оба канала в условиях плохой видимости и осадков, информация передается по резервному широкополосному радиоканалу, который является абсолютно надежным, но обладает гораздо меньшей скоростью передачи по сравнению с основными каналами. Таким образом, гибридная система связи способна передавать данные практически при любых погодных условиях, используя тот или иной канал связи.

6.4.1 Описание системы

Рассматривается система массового обслуживания с двумя основными ненадежными приборами, которые моделируют FSO и миллиметровый каналы, и одним резервным надежным прибором (широкополосный радиоканал). Будем называть FSO канал как прибор 1, миллиметровый радиоканал – как прибор 2, и широкополосный радиоканал – как прибор 3.

Запросы поступают в систему в соответствии с $BMAP$-потоком (см. раздел ?), который задается управляющим процессом ν_t, $t \geq 0$, с пространством состояний $\{0, 1, \ldots, W\}$ и матрицами $D_k, k \geq 0$, порядка

$W+1$. Интенсивность поступления запросов, интенсивность поступления групп, коэффициенты вариации и корреляции в $BMAP$ обозначаются как λ, λ_b, c_{var}, c_{cor}, соответственно, и вычисляются по формулам, приведенным в разделе ?.

Запрос, приходящий на обслуживание, когда прибор 1 является исправным и свободным, немедленно начинает обслуживаться на этом приборе. Если этот прибор занят или находится на ремонте, запрос начинает обслуживаться на приборе 2. Если один из основных приборов находится на ремонте, а второй занят обслуживанием, то запрос становится в очередь, длина которой неограничена, и выбирается на обслуживание позже согласно стратегии FIFO (первым пришел – первым обслужен). Если оба прибора заняты в момент поступления запроса, то последний становится в очередь. Если оба прибора находятся на ремонте в момент прихода запроса, он начинает обслуживаться на приборе 3.

Полагаем, что процесс обслуживания на k-м, $k = 1, 2, 3$, приборе имеет PH-распределение с неприводимым представлением $(\boldsymbol{\beta}_k, S_k)$ и управляющим процессом $m_t^{(k)}$, $t \geq 0$, с пространством состояний $\{1, \ldots, M_k, M_k+1\}$, где состояние $M_k + 1$ является поглощающим. Среднее время обслуживания на k-м приборе вычисляется по формуле $b_1^{(k)} = \boldsymbol{\beta}_k(-S_k)^{-1}\mathbf{e}$, интенсивность обслуживания $\mu_k = (b_1^{(k)})^{-1}$, $k = 1, 2, 3$.

Как было сказано выше, поломки основных приборов зависят от погодных условий. Считаем, что поломки приходят в $MMAP$ (*Marked Markovian Arrival Process*)-потоке под управлением цепи Маркова (управляющего процесса) η_t, $t \geq 0$, принимающей значения в множестве $\{0, 1, ..., V\}$. Этот $MMAP$ задается $(V+1)\times(V+1)$ матрицами H_0, H_1, H_2. Матрица $H = H_0+H_1+H_2$ является инфинитезимальным генератором цепи Маркова η_t, $t \geq 0$. Интенсивность потока поломок задается формулой $h = \boldsymbol{\vartheta}(H_1 + H_2)\mathbf{e}$, где вектор-строка $\boldsymbol{\vartheta}$ является единственным решением системы уравнений $\boldsymbol{\vartheta}H = \mathbf{0}$, $\boldsymbol{\vartheta}\mathbf{e} = 1$. Приходящие поломки (неблагоприятные погодные условия) могут быть двух типов. Поломки первого типа интерпретируются как факторы, вызывающие плохую видимость, и направляются на прибор 1. Поломки второго типа интерпретируются как осадки и направляются на прибор 2. Интенсивности поступлений поломок n-го типа задаются элементами матрицы $H_n, n = 1, 2$. Недиагональные элементы матрицы H_0 задают интенсивности перехода среды, не сопровождающиеся ухудшением видимости или осадками.

Сразу после поступления поломки на k-й прибор на нем начинается

ремонт (период неблагоприятных погодных условий). Время, необходимое для ремонта k-го прибора, имеет PH-распределение с неприводимым представлением $(\boldsymbol{\tau}^{(k)}, T^{(k)})$ и интенсивностью $\tau_k,\ k = 1, 2$.

Предполагаем, что поломка, поступившая на прибор, когда тот еще не восстановился после предыдущей поломки, игнорируется.

Как отмечалось выше, приход поломки вызывает прекращение обслуживания запроса (если таковое имеет место) основным прибором, на который направляется поломка. Этот прибор уходит на ремонт, в то время как другой основной прибор, если он не на ремонте и не занят, начинает обслуживание запроса заново. Если этот прибор выходит из строя, в то время как другой основной прибор еще не восстановился, то запрос переходит на резервный прибор (прибор 3), где начинает обслуживаться заново. Если во время обслуживания запроса резервным прибором какой-либо из основных приборов восстанавливается, запрос немедленно переходит на этот прибор и начинает обслуживаться заново (т.е. не может быть ситуаций, когда какой-либо из основных приборов свободен, а резервный прибор занят).

Заметим, что в случае экспоненциальных распределений времен обслуживания можно считать, что обслуживание текущего запроса при переходе его с прибора на прибор продолжается, начиная с текущего момента его обслуживания.

6.4.2 Цепь Маркова, описывающая функционирование системы

Пусть в момент времени t

i_t – число заявок в системе, $i_t \geq 0$,

$$n_t = \begin{cases} 0, & \text{если оба основных прибора исправны} \\ & \text{(оба заняты или оба свободны);} \\ 0_k, & \text{если оба основных прибора исправны, } k\text{-й прибор обслуживает запрос, а другой основной прибор свободен, } k = 1, 2; \\ 1, & \text{если прибор 1 на ремонте;} \\ 2, & \text{если прибор 2 на ремонте;} \\ 3, & \text{если оба прибора на ремонте;} \end{cases}$$

$m_t^{(j)}$ – состояние управляющего процесса PH-обслуживания на j-м занятом приборе, $m_t^{(j)} = \overline{1, M_j}, j = 1, 2, 3$;

$r_t^{(j)}$ – состояние управляющего процесса PH-времени ремонта на j-м приборе, $j = 1, 2$, $r_t^{(j)} = \overline{1, R_j}$;

ν_t и η_t – состояния управляющих процессов $BMAP$ потока заявок $MMAP$ потока поломок соответственно, $\nu_t = \overline{0, W}$, $\eta_t = \overline{0, V}$.

Процесс функционирования системы описывается неприводимой цепью Маркова $\xi_t, t \geq 0$, с пространством состояний

$$X = \{(0, n, \nu, \eta),\ i = 0, n = \overline{0, 3},\ \nu = \overline{0, W},\ \eta = \overline{0, V}\} \bigcup$$

$$\{(i, 0_k, \nu, \eta, m^{(k)}),\ i = 1, k = 1, 2, n = 0_k, \nu = \overline{0, W}, \eta = \overline{0, V},\ m^{(k)} = \overline{1, M_k}\} \bigcup$$

$$\{(i, 0, \nu, \eta, m^{(1)}, m^{(2)}),\ i > 1, n = 0, \nu = \overline{0, W}, \eta = \overline{0, V},\ m^{(1)} = \overline{1, M_1}, m^{(2)} =$$

$$= \overline{1, M_2}\} \bigcup \{(i, 1, \nu, \eta, m^{(2)}, r^{(1)}),\ i > 0, n = 1,\ \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(2)} = \overline{1, M_2},$$

$$r^{(1)} = \overline{1, R_1}\} \bigcup \{(i, 2, \nu, \eta, m^{(1)}, r^{(2)}),\ i > 0, n = 2,\ \nu = \overline{0, W}, \eta = \overline{0, V}, m^{(1)} =$$

$$= \overline{1, M_1}, r^{(1)} = \overline{1, R_2}\} \bigcup \{(i, 3, \nu, \eta, m^{(3)}, r^{(1)}, r^{(2)}),\ i > 0, n = 3,\ \nu = \overline{0, W}, \eta =$$

$$= \overline{0, V}, m^{(3)} = \overline{1, M_3}, r^{(k)} = \overline{1, R_k}, k = 1, 2\}.$$

Далее будем предполагать, что состояния цепи $\xi_t, t \geq 0$, внутри каждого из приведенных подмножеств упорядочены в лексикографическом порядке и таким образом упорядоченные подмножества упорядочены в том порядке, в котором они перечислены выше. Обозначим через $Q_{i,j}$ матрицу интенсивностей переходов цепи из состояний, соответствующих значению i счетной компоненты, в состояния, соответствующие значению j этой компоненты, $i, j \geq 0$.

Лемма 6.6. *Инфинитезимальный генератор Q цепи Маркова ξ_t, $t \geq 0$, имеет блочную структуру*

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & Q_{0,4} & \cdots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & Q_{1,4} & \cdots \\ O & Q_{2,1} & Q_1 & Q_2 & Q_3 & \cdots \\ O & O & Q_0 & Q_1 & Q_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

где ненулевые блоки имеют следующий вид:

$$Q_{0,0} =$$

$$= \begin{pmatrix} D_0 \oplus H_0 & I_{\bar{W}} \otimes H_1 \otimes \boldsymbol{\tau}^{(1)} & I_{\bar{W}} \otimes H_2 \otimes \boldsymbol{\tau}^{(2)} & O \\ I_a \otimes \boldsymbol{T}_0^{(1)} & D_0 \oplus (H_0 + H_1) \oplus T^{(1)} & O & I_{\bar{W}} \otimes H_2 \otimes I_{R_1} \otimes \boldsymbol{\tau}^{(2)} \\ I_a \otimes \boldsymbol{T}_0^{(2)} & O & D_0 \oplus (H_0 + H_2) \oplus T^{(2)} & I_{\bar{W}} \otimes H_1 \otimes I_{R_2} \otimes \boldsymbol{\tau}^{(1)} \\ O & I_a \otimes I_{R_1} \otimes \boldsymbol{T}_0^{(2)} & I_a \otimes \boldsymbol{T}_0^{(1)} \otimes I_{R_2} & D_0 \oplus H \oplus T^{(1)} \oplus T^{(2)} \end{pmatrix},$$

$$Q_{0,1} = diag\{D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)}, D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(2)} \otimes I_{R_1}, D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes I_{R_2},$$
$$D_1 \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}_3 \otimes I_{R_1 R_2}\},$$

$$Q_{0,k} = diag\{D_k \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes \boldsymbol{\beta}^{(2)}, D_k \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(2)} \otimes I_{R_1}, D_k \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes I_{R_2},$$
$$D_k \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}_3 \otimes I_{R_1 R_2},$$
$$k > 1,$$

$$Q_{1,0} = \begin{pmatrix} I_a \otimes \boldsymbol{S}_0^{(1)} & O & O & O \\ I_a \otimes \boldsymbol{S}_0^{(2)} & O & O & O \\ O & I_a \otimes \boldsymbol{S}_0^{(2)} \otimes I_{R_1} & O & O \\ O & O & I_a \otimes \boldsymbol{S}_0^{(1)} \otimes I_{R_2} & O \\ O & O & O & I_a \otimes \boldsymbol{S}_0^{(3)} \otimes I_{R_1 R_2} \end{pmatrix},$$

$$Q_{1,1} = \begin{pmatrix} Q_{1,1}^{(1,1)} & Q_{1,1}^{(1,2)} \\ Q_{1,1}^{(2,1)} & Q_{1,1}^{(2,2)} \end{pmatrix},$$

$$Q_{1,1}^{(1,1)} = \begin{pmatrix} D_0 \oplus H_0 \oplus S^{(1)} & O & I_{\bar{W}} \otimes H_1 \otimes \mathbf{e}_{M_1} \otimes \boldsymbol{\beta}^{(2)} \otimes \boldsymbol{\tau}_1 \\ O & D_0 \oplus H_0 \oplus S^{(2)} & I_{\bar{W}} \otimes H_1 \otimes I_{M_2} \otimes \boldsymbol{\tau}_1 \\ O & I_a \otimes I_{M_2} \otimes \boldsymbol{T}_0^{(1)} & D_0 \oplus (H_0 + H_1) \oplus S^{(2)} \oplus T_1 \end{pmatrix},$$

$$Q_{1,1}^{(1,2)} = \begin{pmatrix} I_{\bar{W}} \otimes H_2 \otimes I_{M_1} \otimes \boldsymbol{\tau}_2 & O \\ I_{\bar{W}} \otimes H_2 \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\beta}^{(1)} \otimes \boldsymbol{\tau}_2 & O \\ O & I_{\bar{W}} \otimes H_2 \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\beta}^{(3)} \otimes \boldsymbol{\tau}_2 \end{pmatrix},$$

$$Q_{1,1}^{(2,1)} = \begin{pmatrix} I_a \otimes I_{M_1} \otimes \boldsymbol{T}_0^{(2)} & O & O \\ O & O & I_a \otimes \boldsymbol{\beta}^{(2)} \otimes \mathbf{e}_{M_3} \otimes I_{R_1} \otimes \boldsymbol{T}_0^{(2)} \end{pmatrix},$$

$$Q_{1,1}^{(2,2)} = \begin{pmatrix} D_0 \oplus (H_0 + H_2) \oplus S^{(1)} \oplus T_2 & I_{\bar{W}} \otimes H_1 \otimes \mathbf{e}_{M_1} \otimes \boldsymbol{\beta}^{(3)} \otimes \boldsymbol{\tau}_1 \\ I_a \otimes \boldsymbol{\beta}^{(1)} \otimes \mathbf{e}_{M_3} \otimes \boldsymbol{T}_0^{(1)} \otimes I_{R_2} & D_0 \oplus H \oplus S^{(3)} \oplus T_1 \oplus T_2 \end{pmatrix},$$

$$Q_{1,k} =$$

$$= \begin{pmatrix} D_{k-1} \otimes I_{\bar{V}M_1} \otimes \boldsymbol{\beta}^{(2)} & O & O & O \\ D_{k-1} \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes I_{M_2} & O & O & O \\ O & D_{k-1} \otimes I_{\bar{V}M_2R_1} & O & O \\ O & O & D_{k-1} \otimes I_{\bar{V}M_1R_2} & O \\ O & O & O & D_{k-1} \otimes I_{\bar{V}M_3R_1R_2} \end{pmatrix},$$

$$k \geq 2,$$

$$Q_{2,1} =$$

$$= \begin{pmatrix} I_{aM_1} \otimes \boldsymbol{S}_0^{(2)} & I_a \otimes \boldsymbol{S}_0^{(1)} \otimes I_{M_2} & O & O & O \\ O & O & I_a \otimes \boldsymbol{S}_0^{(2)} \boldsymbol{\beta}^{(2)} \otimes I_{R_1} & O & O \\ O & O & O & I_a \otimes \boldsymbol{S}_0^{(1)} \boldsymbol{\beta}^{(1)} \otimes I_{R_2} & O \\ O & O & O & O & I_a \otimes \boldsymbol{S}_0^{(3)} \boldsymbol{\beta}_3 \otimes I_{R_1R_2} \end{pmatrix},$$

$$Q_0 = diag\{I_a \otimes (\boldsymbol{S}_0^{(1)} \boldsymbol{\beta}^{(1)} \oplus \boldsymbol{S}_0^{(2)} \boldsymbol{\beta}^{(2)}), I_a \otimes \boldsymbol{S}_0^{(2)} \boldsymbol{\beta}^{(2)} \otimes I_{R_1}, I_a \otimes \boldsymbol{S}_0^{(1)} \boldsymbol{\beta}^{(1)} \otimes I_{R_2},$$
$$I_a \otimes \boldsymbol{S}_0^{(3)} \boldsymbol{\beta}_3 \otimes I_{R_1R_2}\},$$

$$Q_1 = \begin{pmatrix} Q_1^{(1,1)} & Q_1^{(1,2)} \\ Q_1^{(2,1)} & Q_1^{(2,2)} \end{pmatrix},$$

$$Q_1^{(1,1)} = \begin{pmatrix} D_0 \oplus H_0 \oplus S^{(1)} \oplus S^{(2)} & I_{\bar{W}} \otimes H_1 \otimes \mathbf{e}_{M_1} \otimes I_{M_2} \otimes \boldsymbol{\tau}_1 \\ I_a \otimes \boldsymbol{\beta}^{(1)} \otimes I_{M_2} \otimes \boldsymbol{T}_0^{(1)} & D_0 \oplus (H_0 + H_1) \oplus S^{(2)} \oplus T_1 \end{pmatrix},$$

$$Q_1^{(1,2)} = \begin{pmatrix} I_{\bar{W}} \otimes H_2 \otimes I_{M_1} \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\tau}_2 & O \\ O & I_{\bar{W}} \otimes H_2 \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\beta}^{(3)} \otimes I_{R_1} \otimes \boldsymbol{\tau}_2 \end{pmatrix},$$

$$Q_1^{(2,1)} = \begin{pmatrix} I_a \otimes I_{M_1} \otimes \boldsymbol{\beta}^{(2)} \otimes \boldsymbol{T}_0^{(2)} & O \\ O & I_a \otimes \boldsymbol{\beta}^{(2)} \otimes \mathbf{e}_{M_3} \otimes I_{R_1} \otimes \boldsymbol{T}_0^{(2)} \end{pmatrix},$$

$$Q_1^{(2,2)} = \begin{pmatrix} D_0 \oplus (H_0 + H_2) \oplus S^{(1)} \oplus T_2 & I_{\bar{W}} \otimes H_1 \otimes \mathbf{e}_{M_1} \otimes \boldsymbol{\beta}^{(3)} \otimes I_{R_2} \otimes \boldsymbol{\tau}_1 \\ I_a \otimes \boldsymbol{\beta}^{(1)} \otimes \mathbf{e}_{M_3} \otimes \boldsymbol{T}_0^{(1)} \otimes I_{R_2} & D_0 \oplus H \oplus S^{(3)} \oplus T_1 \oplus T_2 \end{pmatrix},$$

$$Q_{k+1} = diag\{D_k \otimes I_{\bar{V}M_1M_2}, D_k \otimes I_{\bar{V}M_2R_1}, D_k \otimes I_{\bar{V}M_1R_2}, D_k \otimes I_{\bar{V}M_3R_1R_2}\},\ k \geq 1,$$

где $\bar{W} = W + 1$, $\bar{V} = V + 1$, $a = \bar{W}\bar{V}$.

Доказательство леммы проводится путем анализа поведения цепи на бесконечно малом интервале времени.

Следствие 6.12. *Цепь Маркова* $\xi_t, t \geq 0$, *принадлежит классу квазителицевых цепей с непрерывным временем.*

Введем обозначения для производящих функций

$$Q^{(n)}(z) = \sum_{k=2}^{\infty} Q_{n,k} z^k,\ n = 0, 1,\ Q(z) = \sum_{k=0}^{\infty} Q_k z^k,\ \ |z| \leq 1.$$

Следствие 6.13. *Матричные производящие функции* $Q(z)$, $Q^{(n)}(z)$, $n = 0, 1$, *имеют следующий вид:*

$$Q^{(0)}(z) = diag\{\bar{D}(z) - D_1 z) \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes \boldsymbol{\beta}^{(2)},\ (\bar{D}(z) - D_1 z) \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(2)} \otimes I_{R_1},$$

$$(\bar{D}(z) - D_1 z) \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes I_{R_2},\ (\bar{D}(z) - D_1 z) \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}_3 \otimes I_{R_1} \otimes I_{R_2}\}, \quad (6.57)$$

$$Q^{(1)}(z) = \quad (6.58)$$

$$= z \begin{pmatrix} \bar{D}(z) \otimes I_{\bar{V}M_1} \otimes \boldsymbol{\beta}^{(2)} & O & O & O \\ \bar{D}(z) \otimes I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes I_{M_2} & O & O & O \\ O & \bar{D}(z) \otimes I_{\bar{V}M_2R_1} & O & O \\ O & O & \bar{D}(z) \otimes I_{\bar{V}M_1R_2} & O \\ O & O & O & \bar{D}(z) \otimes I_{\bar{V}M_3R_1R_2} \end{pmatrix},$$

$$Q(z) = Q_0 + \mathcal{Q}(z) + zdiag\{D(z) \otimes I_{\bar{V}M_1M_2},\ D(z) \otimes I_{\bar{V}M_2M_3R_1},$$
$$D(z) \otimes I_{\bar{V}M_1M_3R_2}, D(z) \otimes I_{\bar{V}M_3R_1R_2}\}, \quad (6.59)$$

где

$\bar{D}(z) = D(z) - D_0$, *а матрица* $\mathcal{Q}$ *имеет вид*

$$\mathcal{Q} = \begin{pmatrix} \mathcal{Q}^{(1,1)} & \mathcal{Q}^{(1,2)} \\ \mathcal{Q}^{(2,1)} & \mathcal{Q}^{(2,2)} \end{pmatrix},$$

$$\mathcal{Q}^{(1,1)} = \begin{pmatrix} I_{\bar{W}} \otimes H_0 \oplus S^{(1)} \oplus S^{(2)} & I_{\bar{W}} \otimes H_1 \otimes \mathbf{e}_{M_1} \otimes I_{M_2} \otimes \boldsymbol{\tau}_1 \\ I_a \otimes \boldsymbol{\beta}^{(1)} \otimes I_{M_2} \otimes \boldsymbol{T}_0^{(1)} & I_{\bar{W}} \otimes (H_0 + H_1) \oplus S^{(2)} \oplus T_1 \end{pmatrix},$$

$$\mathcal{Q}^{(1,2)} = \begin{pmatrix} I_{\bar{W}} \otimes H_2 \otimes I_{M_1} \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\tau}_2 & O \\ O & I_{\bar{W}} \otimes H_2 \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\beta}^{(3)} \otimes I_{R_1} \otimes \boldsymbol{\tau}_2 \end{pmatrix},$$

$$\mathcal{Q}^{(2,1)} = \begin{pmatrix} I_a \otimes I_{M_1} \otimes \boldsymbol{\beta}^{(2)} \otimes \boldsymbol{T}_0^{(2)} & O \\ O & I_a \otimes \boldsymbol{\beta}^{(2)} \otimes \mathbf{e}_{M_3} \otimes I_{R_1} \otimes \boldsymbol{T}_0^{(2)} \end{pmatrix},$$

$$\mathcal{Q}^{(2,2)} = \begin{pmatrix} I_{\bar{W}} \otimes (H_0 + H_2) \oplus S^{(1)} \oplus T_2 & I_{\bar{W}} \otimes H_1 \otimes \mathbf{e}_{M_1} \otimes \boldsymbol{\beta}^{(3)} \otimes I_{R_2} \otimes \boldsymbol{\tau}_1 \\ I_a \otimes \boldsymbol{\beta}^{(1)} \otimes \mathbf{e}_{M_3} \otimes \boldsymbol{T}_0^{(1)} \otimes I_{R_2} & I_{\bar{W}} \otimes H \oplus S^{(3)} \oplus T_1 \oplus T_2 \end{pmatrix}.$$

6.4.3 Условие существования стационарного режима в системе. Алгоритм вычисления стационарного распределения

Теорема 6.8. *Необходимым и достаточным условием существования стационарного режима в системе является выполнение неравенства*

$$\lambda < \boldsymbol{\pi}_0(\mathbf{S}_0^{(1)} \oplus \mathbf{S}_0^{(2)})\mathbf{e} + \boldsymbol{\pi}_1\mathbf{S}_0^{(2)} + \boldsymbol{\pi}_2\mathbf{S}_0^{(1)} + \boldsymbol{\pi}_3\mathbf{S}_0^{(3)}, \tag{6.60}$$

где

$$\boldsymbol{\pi}_0 = \mathbf{x}_0(\mathbf{e}_{\bar{V}} \otimes I_{M_1M_2}),\ \boldsymbol{\pi}_1 = \mathbf{x}_1(\mathbf{e}_{\bar{V}} \otimes I_{M_2} \otimes \mathbf{e}_{R_1}),$$
$$\boldsymbol{\pi}_2 = \mathbf{x}_2(\mathbf{e}_{\bar{V}} \otimes I_{M_1} \otimes \mathbf{e}_{R_2}),\ \boldsymbol{\pi}_3 = \mathbf{x}_3(\mathbf{e}_{\bar{V}} \otimes I_{M_3} \otimes \mathbf{e}_{R_1R_2}),$$

а вектор $\mathbf{x} = (\mathbf{x}_0,\ \mathbf{x}_1,\ \mathbf{x}_2,\ \mathbf{x}_3)$ *является единственным решением системы линейных алгебраических уравнений*

$$\mathbf{x}\Gamma = 0,\ \mathbf{x}\mathbf{e} = 1. \tag{6.61}$$

Здесь

$$\Gamma = \begin{pmatrix} \Gamma^{(1,1)} & \Gamma^{(1,2)} \\ \Gamma^{(2,1)} & \Gamma^{(2,2)} \end{pmatrix},$$

$$\Gamma^{(1,1)} =$$
$$= \begin{pmatrix} I_{\bar{V}} \otimes (\boldsymbol{S}_0^{(1)}\boldsymbol{\beta}^{(1)} \oplus \boldsymbol{S}_0^{(2)}\boldsymbol{\beta}^{(2)}) & H_1 \otimes \mathbf{e}_{M_1} \otimes I_{M_2} \otimes \boldsymbol{\tau}_1 \\ I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes I_{M_2} \otimes \boldsymbol{T}_0^{(1)} & +I_{\bar{V}} \otimes \boldsymbol{S}_0^{(2)}\boldsymbol{\beta}^{(2)} \otimes I_{R_1} \end{pmatrix} +$$

$$+diag\{H_0 \oplus S^{(1)} \oplus S^{(2)},\ (H_0 + H_1) \oplus S^{(2)} \oplus T_1\},$$

$$\Gamma^{(1,2)} = \begin{pmatrix} H_2 \otimes I_{M_1} \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\tau}_2 & O \\ O & H_2 \otimes \mathbf{e}_{M_2} \otimes \boldsymbol{\beta}^{(3)} \otimes I_{R_1} \otimes \boldsymbol{\tau}_2 \end{pmatrix},$$

$$\Gamma^{(2,1)} = \begin{pmatrix} I_{\bar{V}} \otimes I_{M_1} \otimes \boldsymbol{\beta}^{(2)} \otimes \boldsymbol{T}_0^{(2)} & O \\ O & I_{\bar{V}} \otimes \boldsymbol{\beta}^{(2)} \otimes \mathbf{e}_{M_3} \otimes I_{R_1} \otimes \boldsymbol{T}_0^{(2)} \end{pmatrix},$$

$$\Gamma^{(2,2)} =$$

$$= \begin{pmatrix} I_{\bar{V}} \otimes \boldsymbol{S}_0^{(1)} \boldsymbol{\beta}^{(1)} \otimes I_{R_2} & H_1 \otimes \mathbf{e}_{M_1} \otimes \boldsymbol{\beta}^{(3)} \otimes I_{R_2} \otimes \boldsymbol{\tau}_1 \\ I_{\bar{V}} \otimes \boldsymbol{\beta}^{(1)} \otimes \mathbf{e}_{M_3} \otimes \boldsymbol{T}_0^{(1)} \otimes I_{R_2} & I_{\bar{V}} \otimes \boldsymbol{S}_0^{(3)} \boldsymbol{\beta}_3 \otimes I_{R_1 R_2} \end{pmatrix} +$$

$$+diag\{(H_0 + H_2) \oplus S^{(1)} \oplus T_2,\ H \oplus S^{(3)} \oplus T_1 \oplus T_2\}.$$

Доказательство производится путем рассуждений, аналогичных приведенным выше, при доказательстве теоремы 6.1. В данном случае представляем вектор $\mathbf{y}$, присутствующий в соотношениях (6.5)-(6.6), в виде

$$\mathbf{y} = (\boldsymbol{\theta} \otimes \mathbf{x_0}, \boldsymbol{\theta} \otimes \mathbf{x_1}, \boldsymbol{\theta} \otimes \mathbf{x_2}, \boldsymbol{\theta} \otimes \mathbf{x_3}), \tag{6.62}$$

где $\mathbf{x} = (\mathbf{x}_0,\ \mathbf{x}_1,\ \mathbf{x}_2,\ \mathbf{x}_3)$ – стохастический вектор.

Тогда, учитывая, что $\boldsymbol{\theta} \sum\limits_{k=0}^{\infty} D_k = \mathbf{0}$, система (6.6) сводится к виду (6.61).

Далее, подставляя в неравенство (6.5) вектор $\mathbf{y}$ в виде (6.62), выражение для $Q^{'}(1)$, вычисленное с помощью формулы (6.59), и учитывая, что $\boldsymbol{\theta} D'(1)\mathbf{e} = \lambda$, сводим это неравенство к виду

$$\lambda + \mathbf{x}\mathcal{Q}^{-}\mathbf{e} < 0, \tag{6.63}$$

где матрица $\mathcal{Q}^{-}$ получена из матрицы $\mathcal{Q}$ формально путем удаления множителя $I_{\bar{W}}\otimes$ и замены множителя I_a на множитель $I_{\bar{V}}$.

Используя соотношения $H\mathbf{e} = (H_0 + H_1 + H_2)\mathbf{e} = \mathbf{0}$, $S_n\mathbf{e} + \boldsymbol{S}_0^{(n)} = \mathbf{0}$, $T_n\mathbf{e} + \boldsymbol{T}_0^{(n)} = \mathbf{0},\ n = 1, 2$, сведем неравенство (6.63) к виду

$$\lambda < \mathbf{x}_0(\mathbf{e}_{\bar{V}} \otimes I_{M_1 M_2})(\mathbf{S}_0^{(1)} \oplus \mathbf{S}_0^{(2)})\mathbf{e} + \mathbf{x}_1(\mathbf{e}_{\bar{V}} \otimes I_{M_2} \otimes \mathbf{e}_{R_1})\mathbf{S}_0^{(2)} +$$

$$+\mathbf{x}_2(\mathbf{e}_{\bar{V}} \otimes I_{M_1} \otimes \mathbf{e}_{R_2})\mathbf{S}_0^{(1)} + \mathbf{x}_3(\mathbf{e}_{\bar{V}} \otimes I_{M_3} \otimes \mathbf{e}_{R_1 R_2})\mathbf{S}_0^{(3)}.$$

После использования обозначений, введенных в теореме, это неравенство приобретает вид (6.60). □

Замечание 6.5. При физической интерпретации условия эргодичности (6.60) учитываем, что данное условие отражает процесс обслуживания в системе в условиях перегрузки. Рассмотрим физический смысл первого слагаемого в правой части неравенства (6.60). Компонента $\boldsymbol{\pi}_0(m^{(1)}, m^{(2)})$ вектора-строки $\boldsymbol{\pi}_0$ есть вероятность того, что приборы 1 и 2 исправны и обслуживают запросы на фазах $m^{(1)}$ и $m^{(2)}$ соответственно. Соответствующая компонента вектора-столбца $(\mathbf{S}_0^{(1)} \oplus \mathbf{S}_0^{(2)})\mathbf{e}$ есть суммарная интенсивность обслуживания запросов 1-м и 2-м приборами при условии, что обслуживание на этих приборах находится в фазах $m^{(1)}$ и $m^{(2)}$ соответственно. Тогда произведение $\boldsymbol{\pi}_0(\mathbf{S}_0^{(1)} \oplus \mathbf{S}_0^{(2)})\mathbf{e}$ представляет собой интенсивность выходящего потока в периоды, когда запросы обслуживаются 1-м и 2-м приборами. Аналогично трактуются остальные слагаемые суммы в правой части неравенства (6.60): второе слагаемое есть интенсивность выходящего потока при обслуживании запросов 2-м прибором (1-й прибор находится на ремонте), третье слагаемое – интенсивность выходящего потока при обслуживании запросов 1-м прибором (2-й прибор находится на ремонте), четвертое слагаемое – интенсивность выходящего потока при обслуживании запросов 3-м прибором (1-й и 2-й приборы находятся на ремонте). Тогда правая часть неравенства (6.60) выражает суммарную интенсивность выходящего потока запросов в условиях перегрузки. Очевидно, что для существования стационарного режима в системе необходимо и достаточно, чтобы интенсивность входного потока λ была меньше интенсивности выходящего потока.

Следствие 6.13. В случае стационарных пуассоновских потоков поломок, экспоненциальных распределений времен обслуживания и ремонтов стационарный режим в системе существует тогда и только тогда, когда выполняется неравенство

$$\lambda < \pi_0(\mu_1 + \mu_2) + \pi_1\mu_2 + \pi_2\mu_1 + \pi_0\mu_3,$$

где вектор $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2, \pi_3)$ есть единственное решение системы линейных алгебраических уравнений

$$\boldsymbol{\pi} \begin{pmatrix} -(h_1+h_2) & h_1 & h_2 & 0 \\ \tau_1 & -(h_2+\tau_1) & 0 & h_2 \\ \tau_2 & 0 & -(h_1+\tau_2) & h_1 \\ 0 & \tau_2 & \tau_1 & -(\tau_1+\tau_2) \end{pmatrix} = \mathbf{0},$$

$$\boldsymbol{\pi}\mathbf{e} = 1.$$

Далее предполагаем, что условие (6.57) выполняется. Тогда существуют стационарные вероятности состояний системы, которые обозначим как

$$p_0^{(n)}(\nu,\eta)=\lim_{t\to\infty}P\{i_t=0,n_t=n,\nu_t=\nu,\eta_t=\eta\},\ n=\overline{0,3},\ \nu=\overline{0,W},\eta=\overline{0,V};$$

$$p_1^{(0_n)}\{(\nu,\eta,m^{(k)})\}=\lim_{t\to\infty}P\{i_t=i,n_t=0_n,\nu_t=\nu,\eta_t=\eta,m_t^{(n)}=m^{(n)}\},$$

$$n=1,2,\ \nu=\overline{0,W},\ \eta=\overline{0,V},\ m^{(n)}=\overline{1,M^{(n)}};$$

$$p_1^{(1)}(\nu,\eta,m^{(2)})=\lim_{t\to\infty}P\{i_t=1,n_t=1,\ \nu_t=\nu,\eta_t=\eta,\ m_t^{(2)}=m^{(2)},\ r_t^{(1)}=r^{(1)}\},$$

$$\nu=\overline{0,W},\ \eta=\overline{0,V},\ m^{(2)}=\overline{1,M^{(2)}},\ r^{(1)}=\overline{1,R^{(1)}};$$

$$p_1^{(2)}(\nu,\eta,m^{(1)})=\lim_{t\to\infty}P\{i_t=1,n_t=2,\ \nu_t=\nu,\eta_t=\eta,\ m_t^{(1)}=m^{(1)},\ r_t^{(2)}=r^{(2)}\},$$

$$\nu=\overline{0,W},\ \eta=\overline{0,V},\ m^{(1)}=\overline{1,M^{(1)}},\ r^{(2)}=\overline{1,R^{(2)}};$$

$$p_1^{(3)}(\nu,\eta,m^{(3)})=\lim_{t\to\infty}P\{i_t=1,n_t=3,\ \nu_t=\nu,\eta_t=\eta,\ m_t^{(3)}=m^{(3)},\ r_t^{(1)}=$$

$$=\overline{1,R^{(1)}},r_t^{(2)}=\overline{1,R^{(2)}}\},\ \nu=\overline{0,W},\ \eta=\overline{0,V},\ m^{(3)}=\overline{1,M^{(3)}},\ r^{(1)}=\overline{1,R^{(1)}},$$

$$r^{(2)}=\overline{1,R^{(2)}};$$

$$p_i^{(1)}(\nu,\eta,m^{(2)})=\lim_{t\to\infty}P\{i_t=i,n_t=1,\ \nu_t=\nu,\eta_t=\eta,\ m_t^{(2)}=m^{(2)},\ m_t^{(3)}=m^{(3)},$$

$$r_t^{(1)}=\overline{1,R^{(1)}}\},\ i\geq 2,\nu=\overline{0,W},\ \eta=\overline{0,V},\ m^{(2)}=\overline{1,M^{(2)}},\ m^{(3)}=\overline{1,M^{(3)}},$$

$$r^{(1)}=\overline{1,R^{(1)}};$$

$$p_i^{(2)}(\nu,\eta,m^{(1)})=\lim_{t\to\infty}P\{i_t=i,n_t=2,\ \nu_t=\nu,\eta_t=\eta,\ m_t^{(1)}=m^{(1)},\ m_t^{(3)}=m^{(3)},$$

$$r_t^{(2)}=\overline{1,R^{(2)}}\},\ i\geq 2,\ \nu=\overline{0,W},\ \eta=\overline{0,V},\ m^{(1)}=\overline{1,M^{(1)}},\ m^{(3)}=\overline{1,M^{(3)}},$$

$$r^{(2)}=\overline{1,R^{(2)}};$$

$$p_i^{(3)}(\nu,\eta,m^{(3)})=\lim_{t\to\infty}P\{i_t=i,n_t=3,\ \nu_t=\nu,\eta_t=\eta,\ m_t^{(3)}=m^{(3)},\ r_t^{(1)}=$$

$$=\overline{1,R^{(1)}},r_t^{(2)}=\overline{1,R^{(2)}}\},\ i\geq 2,\ \nu=\overline{0,W},\ \eta=\overline{0,V},\ m^{(3)}=\overline{1,M^{(3)}},$$

$$r^{(1)}=\overline{1,R^{(1)}},r^{(2)}=\overline{1,R^{(2)}};$$

$$p_i^{(0)}(\nu,\eta,m^{(1)},m^{(2)})=\lim_{t\to\infty}P\{i_t=i,n_t=0,\ \nu_t=\nu,\eta_t=\eta,$$

$$m_t^{(1)}=m^{(1)},\ m_t^{(2)}=m^{(2)}\},i>1,\ \nu=\overline{0,W},\ \eta=\overline{0,V},\ m^{(n)}=\overline{1,M^{(n)}}.$$

Внутри каждой выделенной группы упорядочим вероятности в лексикографическом порядке компонент и сформируем векторы этих вероятностей

$$\mathbf{p}_0^{(n)}, n = \overline{0,3};\ \mathbf{p}_1^{(0_1)},\ \mathbf{p}_1^{(0_2)},\ \mathbf{p}_1^{(n)}, n = \overline{1,3};\ \mathbf{p}_i^{(n)}, n = \overline{0,3},\ i \geq 2.$$

Далее сформируем векторы стационарных вероятностей, соответствующие значениям счетной компоненты:

$$\mathbf{p}_0 = (\mathbf{p}_0^{(0)},\ \mathbf{p}_0^{(1)},\ \mathbf{p}_0^{(2)},\ \mathbf{p}_0^{(3)}),$$
$$\mathbf{p}_1 = (\mathbf{p}_1^{(0_1)},\ \mathbf{p}_1^{(0_2)},\ \mathbf{p}_1^{(1)},\ \mathbf{p}_1^{(2)},\ \mathbf{p}_1^{(3)}),$$
$$\mathbf{p}_i = (\mathbf{p}_i^{(0)},\ \mathbf{p}_i^{(1)},\ \mathbf{p}_i^{(2)},\ \mathbf{p}_i^{(3)}),\ i \geq 2.$$

Векторы стационарных вероятностей $\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \dots)$ вычисляются по алгоритму, который разработан аналогично алгоритму 6.1, описанному в разделе 6.2.3, на основе применения техники сенсорных цепей Маркова (см. [150]). Отличие этих алгоритмов обусловлено тем, что пространственная однородность цепи раздела 6.2 нарушается при $i = 0$, в то время как пространственная однородность цепи, рассматриваемой в данном разделе, нарушается и при $i = 0$, и при $i = 1$. Это отражается в структуре генераторов и, соответственно, в алгоритмах вычисления стационарных распределений. В данном случае алгоритм имеет следующий вид.

Алгоритм 6.4.

1. Вычисляем матрицу G как минимальное неотрицательное решение матричного уравнения

$$\sum_{n=0}^{\infty} Q_n G^n = O.$$

2. Вычисляем матрицу G_1, используя следующее уравнение

$$Q_{2,1} + \sum_{n=1}^{\infty} Q_n G^{n-1} G_1 = O,$$

откуда следует, что

$$G_1 = -(\sum_{n=1}^{\infty} Q_n G^{n-1})^{-1} Q_{2,1}.$$

3. Вычисляем матрицу G_0, используя следующее уравнение

$$Q_{1,0} + (Q_{1,1} + \sum_{n=2}^{\infty} Q_{1,n} G^{n-2} G_1) G_0 = O,$$

откуда следует, что

$$G_0 = -(Q_{1,1} + \sum_{n=2}^{\infty} Q_{1,n} G^{n-2} G_1)^{-1} Q_{1,0},$$

4. Вычисляем матрицы

$$\bar{Q}_{i,l} = \begin{cases} Q_{i,l} + \sum\limits_{n=l+1}^{\infty} Q_{i,n} G_{n-1} G_{n-2} \dots G_l, \; i = 0, 1, \; l \geq i; \\ Q_{l-i+1} + \sum\limits_{n=l+1}^{\infty} Q_{n-i+1} G_{n-1} G_{n-2} \dots G_l, \; i > 1, \; l \geq i, \end{cases}$$

где $G_i = G$, $i \geq 2$.

5. Вычисляем матрицы Φ_l используя рекуррентную формулу

$$\Phi_0 = I, \; \Phi_i = (\bar{Q}_{0,i} + \sum_{l=1}^{i-1} \Phi_l \bar{Q}_{l,i})(-\bar{Q}_{i,i})^{-1}, \; i \geq 1.$$

6. Вычисляем вектор $\mathbf{p}_0$ как единственное решение системы

$$\begin{cases} \mathbf{p}_0 \bar{Q}_{0,0} = \mathbf{0}, \\ \mathbf{p}_0 \mathbf{e} + \boldsymbol{p}_0 \Phi_1 \mathbf{e} + \mathbf{p}_0 \sum\limits_{i=2}^{\infty} \Phi_i \mathbf{e} = 1. \end{cases}$$

7. Вычисляем векторы $\mathbf{p}_i$ следующим образом: $\mathbf{p}_i = \mathbf{p}_0 \Phi_i$, $i \geq 1$.

6.4.4 Векторная производящая функция стационарного распределения. Характеристики производительности системы

Вычислив векторы стационарных вероятностей $\mathbf{p}_i$, $i \geq 0$, можно вычислить также различные характеристики производительности системы. При этом будет полезным следующий результат.

Лемма 6.8. Векторная производящая функция $\boldsymbol{P}(z) = \sum\limits_{i=2}^{\infty} \mathbf{p}_i z^i$, $|z| \leq 1$, удовлетворяет следующему уравнению:

$$\boldsymbol{P}(z) Q(z) = \mathcal{B}(z), \tag{6.12}$$

где

$$\mathcal{B}(z) =$$

$$= z\{\mathbf{p}_2[Q_0 - Q_{2,1}] + \mathbf{p}_1[Q_{1,0} + Q_{1,1} - Q^{(1)}(z)] + \mathbf{p}_0[Q_{0,0} + Q_{0,1} - Q^{(0)}(z)]\}. \quad (6.13)$$

Формула (6.12), в частности, может быть использована для вычисления значений функции $\boldsymbol{P}(z)$ и ее производных в точке $z = 1$ без вычисления бесконечных сумм. Полученные значения позволят найти моменты числа заявок в системе и ряд других характеристик системы. Заметим, что непосредственно вычислить величину $\boldsymbol{P}(z)$ и ее производных в точке $z = 1$ из уравнения (6.12) не удается, так как матрица $Q(1)$ вырожденная. Преодолеть эту трудность можно путем использования рекуррентных формул, приведенных ниже, вследствии (6.14).

Обозначим через $f^{(m)}(z)$ m-ю производную функции $f(z)$, $m \geq 1$, и $f^{(0)}(z) = f(z)$.

Следствие 6.14. Производные векторной производящей функции $\boldsymbol{P}(z)$, $|z| \leq 1$, в точке $z = 1$ (факториальные моменты) вычисляются рекуррентно как решения следующих систем линейных алгебраических уравнений:

$$\begin{cases} \boldsymbol{P}^{(m)}(1)Q(1) = \mathcal{B}^{(m)}(1) - \sum\limits_{l=0}^{m-1} C_m^l \boldsymbol{P}^{(l)}(1)Q^{(m-l)}(1), \\ \boldsymbol{P}^{(m)}(1)Q'(1)\mathbf{e} = \frac{1}{m+1}[\mathcal{B}^{(m+1)}(1) - \sum\limits_{l=0}^{m-1} C_{m+1}^l \boldsymbol{P}^{(l)}(1)Q^{(m+1-l)}(1)]\mathbf{e}, \end{cases}$$

где производные $\mathcal{B}^{(m)}(1)$ вычисляются с использованием формул (6.13), (6.57)-(6.59).

Доказательство следствия аналогично доказательству следствия 6.4.

Вычислив векторы стационарных вероятностей $\mathbf{p}_i$, $i \geq 0$, можно вычислить также различные характеристики производительности системы. Некоторые из них приведены ниже.

1. Пропускная способность системы

$$\varrho = \boldsymbol{\pi}_0(\mathbf{S}_0^{(1)} \oplus \mathbf{S}_0^{(2)})\mathbf{e} + \boldsymbol{\pi}_1\mathbf{S}_0^{(2)} + \boldsymbol{\pi}_2\mathbf{S}_0^{(1)} + \boldsymbol{\pi}_3\mathbf{S}_0^{(3)}.$$

   В случае экспоненциальных распределений времен обслуживания и ремонтов пропускная способность вычисляется как

$$\varrho = \pi_0(\mu_1 + \mu_2) + \pi_1\mu_2 + \pi_2\mu_1 + \pi_0\mu_3.$$

2. Среднее число запросов в системе $L = [\mathbf{p}_1 + \boldsymbol{P}'(1)]\mathbf{e}$.
3. Дисперсия числа запросов в системе $V = \boldsymbol{P}^{(2)}(1)\mathbf{e} + L - L^2$.

4. Вероятность того, что в системе находится i запросов $P_i = \mathbf{p}_i\mathbf{e}$.
5. Вероятность $P_i^{(0)}$ того, что в системе находится i запросов и оба прибора исправны

$$P_0^{(0)} = \mathbf{p}_0 \begin{pmatrix} \mathbf{e}_a \\ \mathbf{0}_{a(R_1+R_2+R_1R_2)} \end{pmatrix}, \quad P_1^{(0)} = \mathbf{p}_1 \begin{pmatrix} \mathbf{e}_{a(M_1+M_2)} \\ \mathbf{0}_{a(M_2R_1+M_1R_2+M_3R_1R_2)} \end{pmatrix},$$

$$P_i^{(0)} = \mathbf{p}_i \begin{pmatrix} \mathbf{e}_{aM_1M_2} \\ \mathbf{0}_{a(M_2R_1+M_1R_2+M_3R_1R_2} \end{pmatrix}, \; i \geq 2.$$

6. Вероятность $P_i^{(1)}(P_i^{(2)})$ того, что в системе находится i запросов и прибор 1 (прибор 2) на ремонте

$$P_0^{(1)} = \mathbf{p}_0 \begin{pmatrix} \mathbf{0}_a \\ \mathbf{e}_{aR_1} \\ \mathbf{0}_{aR_2} \\ \mathbf{0}_{aR_1R_2} \end{pmatrix}, \quad P_0^{(2)} = \mathbf{p}_0 \begin{pmatrix} \mathbf{0}_{a(1+R_1)} \\ \mathbf{e}_{aR_2} \\ \mathbf{0}_{aR_1R_2} \end{pmatrix},$$

$$P_1^{(1)} = \mathbf{p}_1 \begin{pmatrix} \mathbf{0}_{a(M_1+M_2)} \\ \mathbf{e}_{aM_2R_1} \\ \mathbf{0}_{aM_1R_2} \\ \mathbf{0}_{aM_3R_1R_2} \end{pmatrix}, \quad P_0^{(2)} = \mathbf{p}_1 \begin{pmatrix} \mathbf{0}_{a(M_1+M_1+M_2R_1)} \\ \mathbf{e}_{aM_1R_2} \\ \mathbf{0}_{aM_3R_1R_2} \end{pmatrix},$$

$$P_i^{(1)} = \mathbf{p}_i \begin{pmatrix} \mathbf{0}_{aM_1M_2} \\ \mathbf{e}_{aM_2R_1} \\ \mathbf{0}_{aM_1R_2} \\ \mathbf{0}_{aM_3R_1R_2} \end{pmatrix}, \quad P_0^{(2)} = \mathbf{p}_0 \begin{pmatrix} \mathbf{0}_{a(M_1M_2+M_2R_1)} \\ \mathbf{e}_{aM_1R_2} \\ \mathbf{0}_{aM_3R_1R_2} \end{pmatrix}, \; i \geq 2.$$

7. Вероятность $P_i^{(3)}$ того, что в системе находится i запросов и оба основных прибора на ремонте

$$P_0^{(3)} = \mathbf{p}_0 \begin{pmatrix} \mathbf{0}_{a(1+R_1+R_2)} \\ \mathbf{e}_{aR_1R_2} \end{pmatrix}, \quad P_1^{(3)} = \mathbf{p}_1 \begin{pmatrix} \mathbf{0}_{a(M_1+M_2+M_2R_1+M_1R_2)} \\ \mathbf{e}_{aM_3R_1R_2} \end{pmatrix},$$

$$P_i^{(3)} = \mathbf{p}_i \begin{pmatrix} \mathbf{0}_{a(M_1M_2+M_2R_1+M_1R_2)} \\ \mathbf{e}_{aM_3R_1R_2} \end{pmatrix}, \; i \geq 2.$$

8. Вероятность того, что в произвольный момент времени приборы находятся в состоянии n

$$P^{(n)} = \sum_{i=0}^{\infty} P_i^{(n)}, \; n = \overline{0,3}.$$

9. Вероятность $P_{i,k}^{(0)}$ того, что поступившая группа размера k найдет в системе i запросов и оба прибора исправными

$$P_{0,k}^{(0)} = \lambda^{-1}\mathbf{p}_0 \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}} \\ O_{a(R_1+R_2+R_1R_2)\times\bar{W}} \end{pmatrix} D_k,$$

$$P_{1,k}^{(0)} = \lambda^{-1}\mathbf{p}_1 \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}(M_1+M_2)}) \\ O_{a(M_2R_1+M_1R_2+M_3R_1R_2)\times\bar{W}} \end{pmatrix} D_k,$$

$$P_{i,k}^{(0)} = \lambda^{-1}\mathbf{p}_i \begin{pmatrix} I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_1M_2} \\ O_{a(M_2R_1+M_1R_2+M_3R_1R_2)\times\bar{W}} \end{pmatrix} D_k,\ i \geq 2.$$

10. Вероятность $P_{i,k}^{(1)}(P_{i,k}^{(2)})$ того, что поступившая группа размера k, $k \geq 1$, найдет в системе i запросов и прибор 1 (прибор 2) на ремонте

$$P_{0,k}^{(1)} = \lambda^{-1}\mathbf{p}_0 \begin{pmatrix} O_{a\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}R_1} \\ O_{a(R_2+R_1R_2)\times\bar{W}} \end{pmatrix} D_k,$$

$$P_{1,k}^{(1)} = \lambda^{-1}\mathbf{p}_1 \begin{pmatrix} O_{a(M_1+M_2)\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_2R_1} \\ O_{a(M_1R_2+M_3R_1R_2)\times\bar{W}} \end{pmatrix} D_k,$$

$$P_{i,k}^{(1)} = \lambda^{-1}\mathbf{p}_i \begin{pmatrix} O_{aM_1M_2\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_2R_1} \\ O_{a(M_1R_2+M_3R_1R_2)\times\bar{W}} \end{pmatrix} D_k,\ i \geq 2,$$

$$P_{0,k}^{(2)} = \lambda^{-1}\mathbf{p}_0 \begin{pmatrix} O_{a(1+R_1)\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}R_2} \\ O_{aR_1R_2\times\bar{W}} \end{pmatrix} D_k,$$

$$P_{1,k}^{(2)} = \lambda^{-1}\mathbf{p}_1 \begin{pmatrix} O_{a(M_1+M_1+M_2R_1)\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_1R_2} \\ O_{aM_3R_1R_2\times\bar{W}} \end{pmatrix} D_k,$$

$$P_{i,k}^{(2)} = \lambda^{-1}\mathbf{p}_0 \begin{pmatrix} O_{a(M_1M_2+M_2R_1)\times\bar{W}} \\ \mathbf{e}_{\bar{V}M_1R_2} \\ O_{aM_3R_1R_2\times\bar{W}} \end{pmatrix} D_k,\ i \geq 2.$$

11. Вероятность $P_{i,k}^{(3)}$ того, что поступившая группа размера k, $k \geq 1$, найдет в системе i запросов и оба основных прибора на ремонте

$$P_{0,k}^{(3)} = \lambda^{-1}\mathbf{p}_0 \begin{pmatrix} O_{a(1+R_1+R_2)\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}R_1R_2} \end{pmatrix} D_k,$$

$$P_{1,k}^{(3)} = \lambda^{-1}\mathbf{p}_1 \begin{pmatrix} O_{a(M_1+M_2+M_2R_1+M_1R_2)\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_3R_1R_2} \end{pmatrix} D_k,$$

$$P_{i,k}^{(3)} = \lambda^{-1}\mathbf{p}_i \begin{pmatrix} O_{a(M_1M_2+M_2R_1+M_1R_2)\times\bar{W}} \\ I_{\bar{W}} \otimes \mathbf{e}_{\bar{V}M_3R_1R_2} \end{pmatrix} D_k,\ i \geq 2.$$

6.5 Численные эксперименты

Численные примеры носят иллюстративный характер и состоят из графиков зависимостей характеристик производительности системы от ее параметров. Ниже приведены результаты пяти экспериментов.

Эксперимент 6.6. Зависимость среднего числа L заявок в системе от интенсивности входного потока λ при различных интенсивностях потока поломок h.

Определим входные данные для этого эксперимента.

$BMAP$ поток запросов задается матрицами D_0 и $D_k, k = \overline{1,3}$. Матрица D_0 имеет вид

$$D_0 = \begin{pmatrix} -1.349076 & 1.09082 \times 10^{-6} \\ 1.09082 \times 10^{-6} & -0.043891 \end{pmatrix}.$$

Матрицы D_k определяются следующим образом: $D_k = Dq^{k-1}(1-q)/(1-q^3), k = \overline{1,3}, q = 0.8$, где матрица D определяется как

$$D = \begin{pmatrix} 1.340137 & 0.008939 \\ 0.0244854 & 0.0194046 \end{pmatrix}.$$

Этот MAP имеет коэффициент вариации $c_{var}^2 = 9.6$ и коэффициент корреляции $c_{cor} = 0.41$.

В ходе эксперимента интенсивность $BMAP$-потока λ изменяется путем нормирования матриц $D_k, k = \overline{0,3}$.

$MMAP$ -поток поломок характеризуется матрицами

$$H_0 = \begin{pmatrix} -8.110725 & 0 \\ 0 & -0.26325 \end{pmatrix},$$

$$H_1 = \frac{1}{3}\begin{pmatrix} 8.0568 & 0.053925 \\ 0.146625 & 0.116625 \end{pmatrix}, H_2 = \frac{2}{3}\begin{pmatrix} 8.0568 & 0.053925 \\ 0.146625 & 0.116625 \end{pmatrix}$$

Для этого $MMAP$ $c_{cor} = 0,2$, $c^2_{var} = 12,3$.

PH-распределения времен обслуживания на трех приборах будем обозначать как $PH_1^{(serv)}, PH_2^{(serv)}, PH_3^{(serv)}$.

$PH_1^{(serv)}$ – распределение Эрланга 2-го порядка с $c^2_{var} = 0.5$ – характеризуется следующим вектором и матрицей:

$$\boldsymbol{\beta}^{(1)} = (1, 0),\ S^{(1)} = \begin{pmatrix} -20 & 20 \\ 0 & -20 \end{pmatrix}.$$

$PH_2^{(serv)}$ – распределение Эрланга 2-го порядка с $c^2_{var} = 0.5$ – характеризуется следующим вектором и матрицей:

$$\boldsymbol{\beta}^{(2)} = (1, 0),\ S^{(2)} = \begin{pmatrix} -15 & 15 \\ 0 & -15 \end{pmatrix}.$$

$PH_3^{(serv)}$ – распределение Эрланга 2-го порядка с $c^2_{var} = 0.5$- характеризуется следующим вектором и матрицей:

$$\boldsymbol{\beta}^{(3)} = (1, 0),\ S^{(3)} = \begin{pmatrix} -4 & 4 \\ 0 & -4 \end{pmatrix}.$$

PH-распределения времен ремонтов приборов 1 и 2 будем обозначать как $PH_1^{(repair)}$ и $PH_2^{(repair)}$. Будем считать, что эти распределения совпадают, являются гиперэкспонентами порядка 2 с $c^2_{var} = 25$ и характеризуются следующими векторами и матрицами:

$$\boldsymbol{\tau}^{(1)} = \boldsymbol{\tau}^{(2)} = (0.05,\ 0.95),\ T^{(1)} = T^{(2)} = \begin{pmatrix} -0.003101 & 0 \\ 0 & -0.245 \end{pmatrix}.$$

На рисунке 6.6 изображен график зависимости среднего числа L заявок в системе от интенсивности входного потока λ при различных интенсивностях поступления поломок: $h = 0.0001, h = 0.001, h = 0.001$.

Как и следовало ожидать, при увеличении интенсивности поступления запросов и поломок величина L возрастает, причем скорость возрастания увеличивается при росте загрузки системы. Числа, соответствующие этому графику, приведены в таблице 6.16.

Таблица 6.16. Среднее число запросов в системе L и коэффициент загрузки ρ как функции интенсивности входного потока λ при различных интенсивностях h $MMAP$- потока поломок

λ	1.0	3.0	5.0	7.0	9.0	11.0	13.0	15.0
$h = 0.0001$								
ρ	0.057	0.172	0.286	0.400	0.515	0.629	0.744	0.858
L	0.148	1.191	13.645	37.297	72.677	130.285	239.641	525.405
$h = 0.001$								
ρ	0.058	0.173	0.288	0.404	0.519	0.635	0.750	0.865
L	0.154	1.540	14.879	40.671	81.659	152.103	291.168	673.258
$h = 0.01$								
ρ	0.063	0.189	0.314	0.439	0.566	0.691	0.817	0.943
L	0.418	7.015	34.682	97.493	231.646	517.529	1267.291	5959.399

Эксперимент 6.7. Зависимость среднего числа запросов L и дисперсии V числа запросов от интенсивности входного потока λ при различных коэффициентах корреляции c_{cor} во входном потоке.

Рассмотрим три различных $BMAP$-потока с разными коэффициентами корреляции и одинаковой интенсивностью.

Эти $BMAP$-потоки будем обозначать как $BMAP_1, BMAP_2, BMAP_3$.

Первый $BMAP$ есть стационарный пуассоновский поток. Он кодируется как $BMAP_1$ и имеет $c_{cor} = 0$, $c_{var} = 1$.

$BMAP_1$ – это стационарный пуассоновский поток, у которого, как известно, $c_{cor} = 0$, $c_{var} = 1$.

$BMAP_2$ задан матрицами D_0 и $D_k = Dq^{k-1}(1-q)/(1-q^3), k = \overline{1,3}$, где $q = 0.8$,

$$D_0 = \begin{pmatrix} -6.34080 & 1.87977 \times 10^{-6} \\ 1.87977 \times 10^{-6} & -0.13888 \end{pmatrix},$$

$$D = \begin{pmatrix} 6.32140 & 0.01939 \\ 0.10822 & 0.03066 \end{pmatrix}.$$

Для этого $BMAP$ $c_{var} = 3.5$, $c_{cor} = 0.1$.

$BMAP_3$ – это $BMAP$ из эксперимента 6.6. Для этого $BMAP$ $c_{var}^2 = 9,6$, $c_{cor} = 0,41$.

Из первого эксперимента также возьмем $MMAP$-поток поломок и PH-распределения времен обслуживания и ремонтов.

На рисунке 6.7-6.8 изображены графики зависимости среднего числа запросов L и дисперсии D числа запросов от интенсивности входного потока λ при различных коэффициентах корреляции c_{cor} во входном потоке.

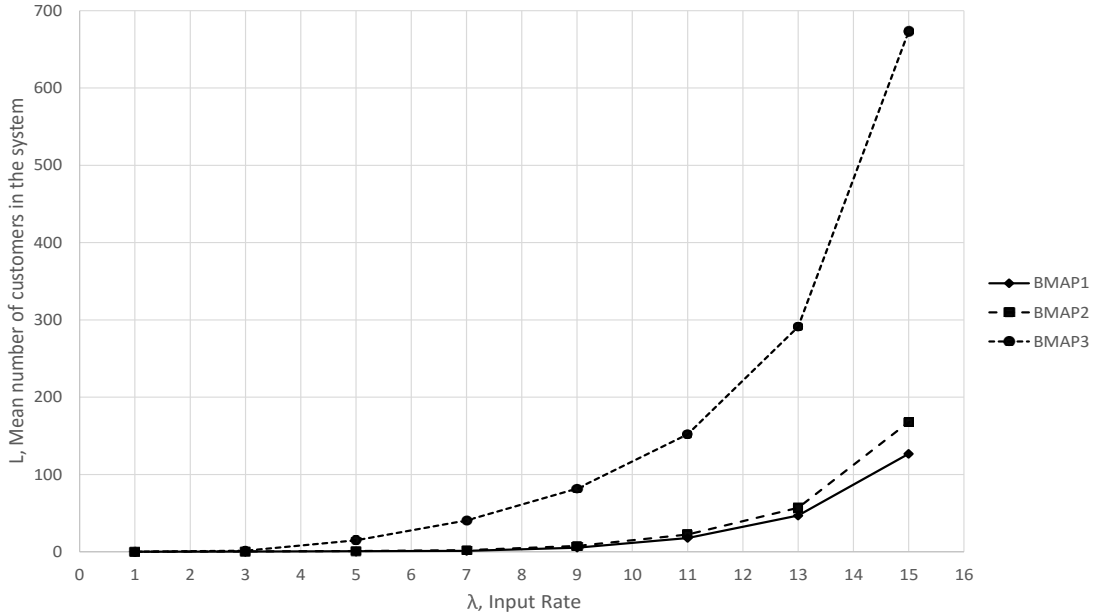

Рисунок 6.7: Зависимость L от λ для $BMAP$-потоков с различными коэффициентами корреляции

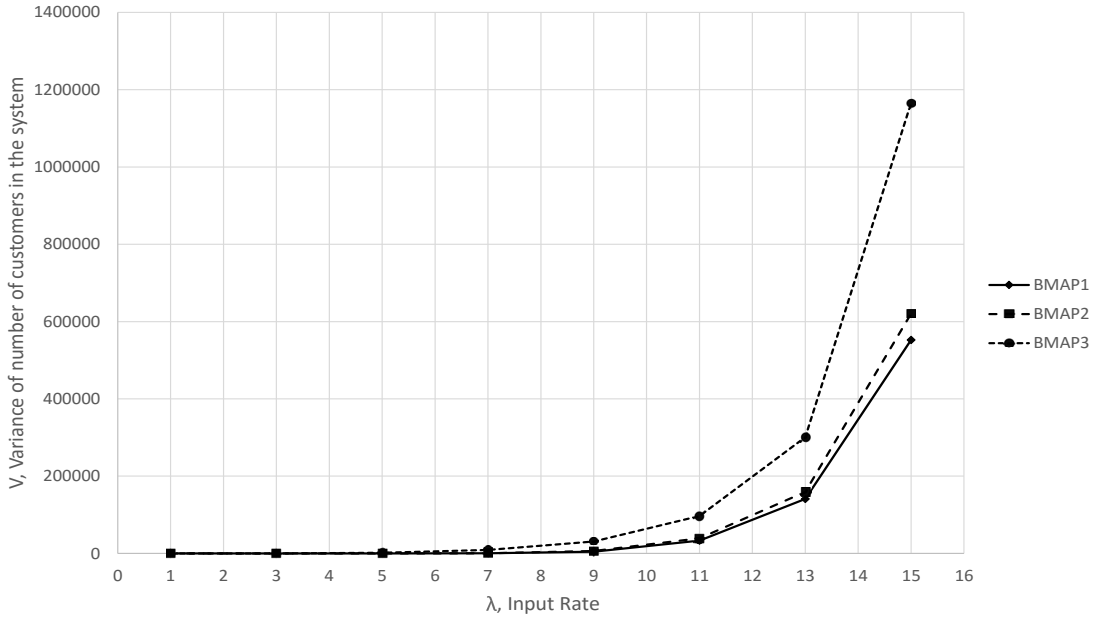

Рисунок 6.8: Зависимость V от λ для $BMAP$-потоков с различными коэффициентами корреляции

Из рисунков видно, что среднее число запросов в системе и дисперсия являются возрастающими функциями интенсивности входного потока λ. Более интересный факт заключается в том, что при одинаковой интенсивности λ, обе характеристики принимают большее значение при больших значениях коэффициента корреляции. Это свидетельствует о том, что кор-

реляция во входном потоке может существенно влиять на характеристики системы и должна учитываться при расчете ее характеристик и оценке производительности.

Эксперимент 6.8. Зависимость вероятности $P^{(n)}$ того, что приборы находятся в состоянии $n, n = \overline{0,3}$, от интенсивности поломок h.

Напомним, что $P^{(n)}$ — это вероятность того, что в произвольный момент времени приборы находятся в состоянии n, где состояние $n = 0$ означает, что оба прибора исправны, $n = 1$ – что только первый прибор на ремонте, $n = 2$ – что только второй прибор на ремонте, а $n = 3$ – что оба прибора находятся на ремонте.

В данном эксперименте будем использовать следующие входные данные

Входной $BMAP$ строится следующим образом: матрица D_0 имеет вид

$$D_0 = \begin{pmatrix} -8.110725 & 0 \\ 0 & -0.26325 \end{pmatrix},$$

Максимальное число запросов в группе равно трем, матрицы D_k вычисляются как $D_k = Dq^{k-1}(1-q)/(1-q^3), k = \overline{1,3}$, где $q = 0.8$,

$$D = \begin{pmatrix} 8.0568 & 0.053925 \\ 0.146625 & 0.116625 \end{pmatrix}.$$

Нормируем эти матрицы, чтобы получить $\lambda = 10$. Для такого $BMAP$ $c_{cor} = 0,2$, $c_{var}^2 = 12$.

$MMAP$ -поток поломок и PH - распределения времён обслуживания и ремонтов возьмем из эксперимента 6.6.

Для наглядности результатов эксперимента отложим на оси абсцисс значения h, которые отличаются друг от друга на порядок. При этом будем использовать на этой оси логарифмическую шкалу, где логарифм понимается как десятичный.

На рисунках 6.9-6.11 приведены графики для $P^{(0)}, P^{(1)}, P^{(2)}, P^{(3)}$ соответственно.

Из рисунка 6.9 видно, что при значениях h, близких к нулю, вероятность $P^{(0)}$ стремится к единице. Начиная приблизительно с $h = 0.005$, вероятность $P^{(0)}$ начинает убывать, в то время как вероятности $P^{(n)}, n = \overline{1,3}$ начинают возрастать.

Эксперимент 6.9. Зависимость среднего числа запросов в системе L от интенсивности поломок h при различных коэффициентах вариации c_{var} времени ремонта.

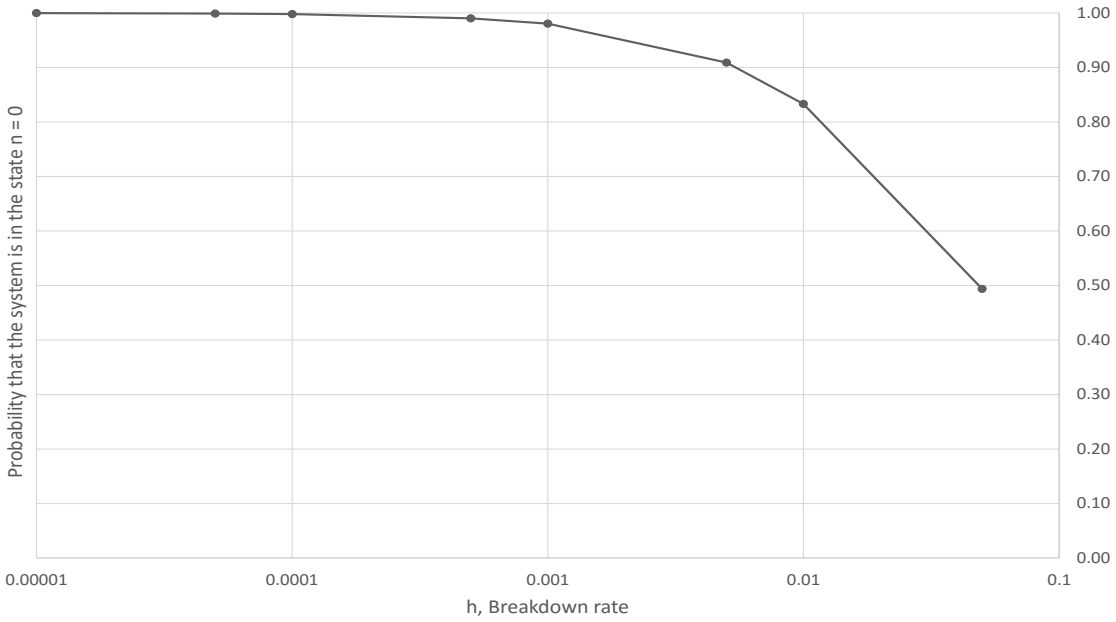

Рисунок 6.9: Зависимость вероятности $P^{(0)}$ того, что оба основных прибора исправны от интенсивности поломок h

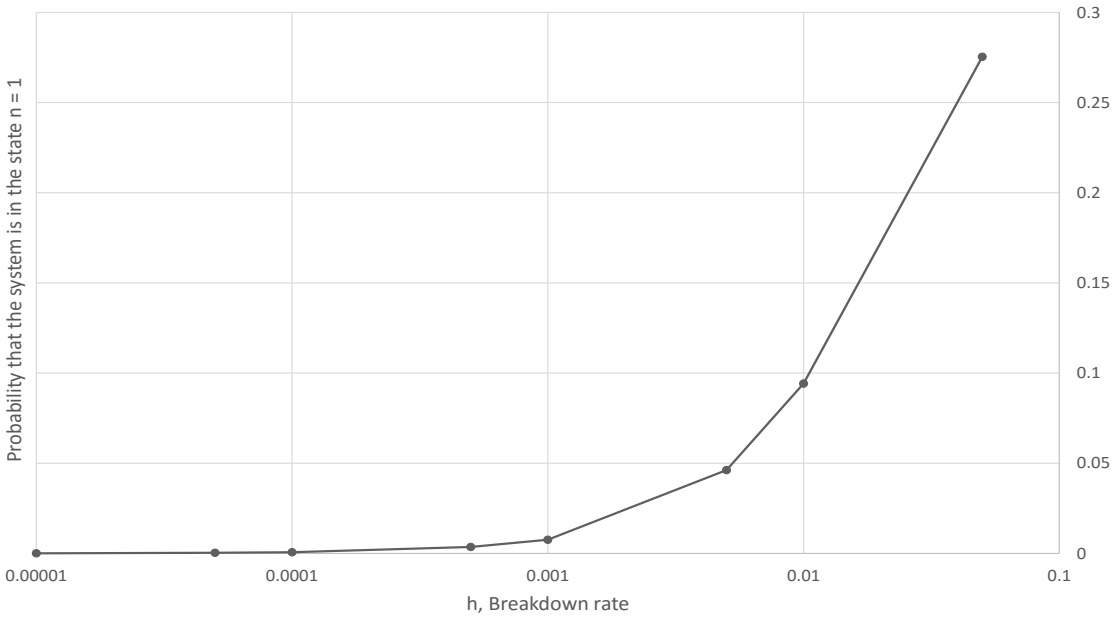

Рисунок 6.10: Зависимость вероятности $P^{(1)}$ того, что 1 прибор на ремонте, от интенсивности поломок h

$BMAP$-поток запросов, $MMAP$-поток поломок и PH-распределения времен обслуживания возьмем такие же, как в эксперименте 6.6.

Для простоты в этом эксперименте будем считать, что времена ремонтов на приборе 1 и приборе 2 имеют одинаковые PH-распределения и расмотрим три таких распределения с различными коэффициентами ва-

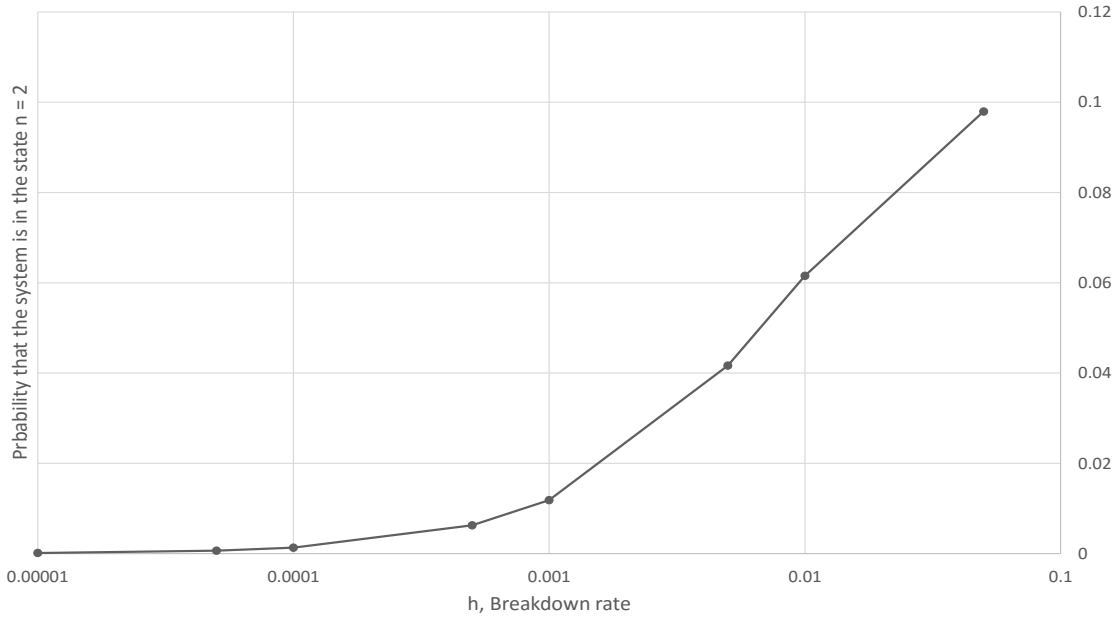

Рисунок 6.11: Зависимость вероятности $P^{(2)}$ того, что 2 прибор на ремонте, от интенсивности поломок h

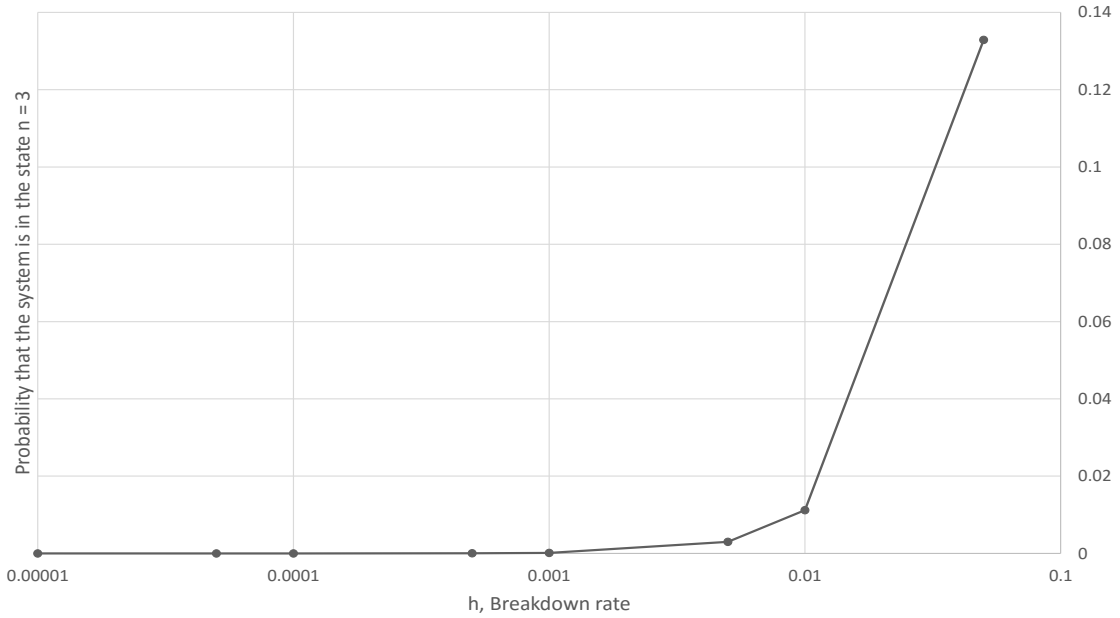

Рисунок 6.12: Зависимость вероятности $P^{(3)}$ того, что оба основных прибора на ремонте, от интенсивности поломок h

риации. Обозначим их как PH_1, PH_2, PH_3.

PH_1 – экспоненциальное распределение с интенсивностью 0.05. У него $c_{var} = 1$.

PH_1 – гиперэкспоненциальное распределение 2 порядка, заданное следующими вектором и матрицей:

$$\boldsymbol{\tau} = (0.05,\, 0.95),\ T = \begin{pmatrix} -0.0031012652092457 52 & 0 \\ 0 & -0.2450002015316405 \end{pmatrix}.$$

У этого распределения $c_{var} = 5$.

PH_2 – гиперэкспоненциальное распределение 2 порядка, заданное следующими вектором и матрицей:

$$\boldsymbol{\tau} = (0.05,\, 0.95),\ T = \begin{pmatrix} -100 & 0 \\ 0 & -0.0002 \end{pmatrix}.$$

У этого распределения $c_{var} = 9.9$.

Пронормируем матрицу T в обоих последних случаях так, чтобы получить интенсивность ремонта 0.05.

В этом эксперименте, так же, как и в предыдущем, мы выбираем интенсивности ремонта такими, что следующее значение отличается от предыдущего на порядок, и используем логарифмическую шкалу по оси абсцисс.

Зависимость среднего числа запросов в системе L от интенсивности поломок h при различных коэффициентах вариации $c_{var} = 1, 5, 9.9$ времени ремонта изображена на рисунке 6.12.

Рисунок 6.12, а также таблица 6.17 показывают, что вариация времени ремонта приборов существенно влияет на среднее число запросов в системе при больших значениях h, при которых коэффициент загрузки системы $\rho > 0.56$.

Таблица 6.17. Среднее число запросов в системе L и коэффициент загрузки ρ как функции интенсивности потока поломок h для PH-распределений времен ремонтов с различными коэффициентами вариации ($c_{var} = 1, 5, 9.9$)

h	0.00001	0.0001	0.001	0.01	0.05
$PH_1 : c_{var=1}$					
ρ	0.514	0.515	0.519	0.561	0.707
L	71.815	72.085	74.807	104.515	348.21
$PH_2 : c_{var=5}$					
ρ	0.514	0.515	0.519	0.566	0.754
L	71.873	72.677	81.659	231.646	1926.314
$PH_3 : c_{var=9.9}$					
ρ	0.514	0.515	0.519	0.569	0.7909
L	71.994	73.972	101.003	631.539	7000.653

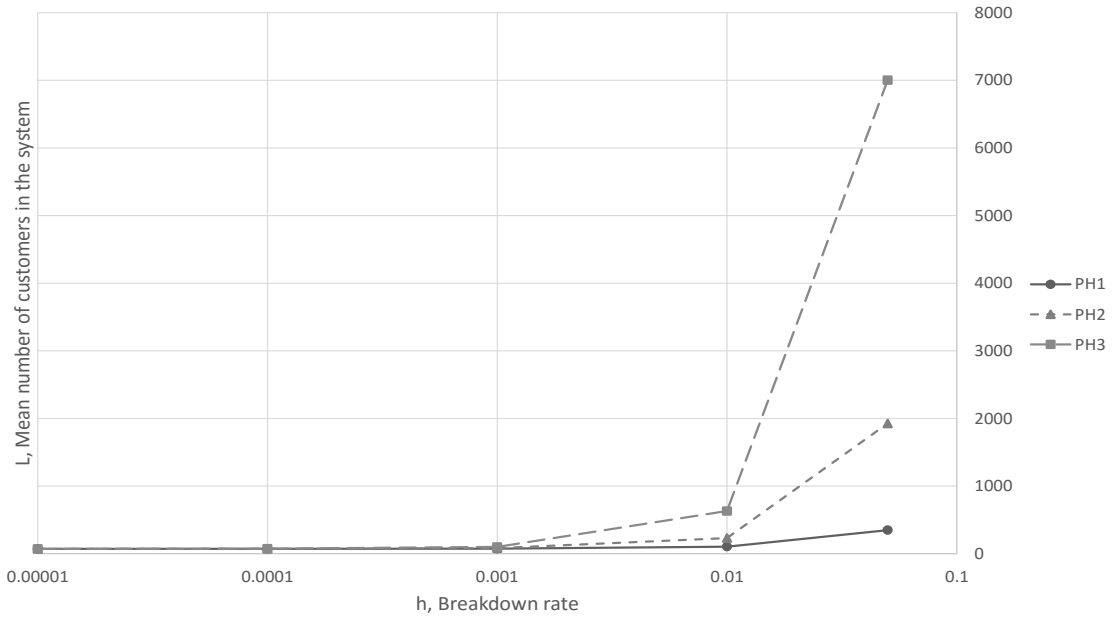

Рисунок 6.13: Зависимость L от h для PH-распределений времен ремонтов с различными коэффициентами вариации ($c_{var} = 1, 5, 9.9$)

Эксперимент 6.10. Зависимость пропускной способности ϱ от интенсивности поломок h при различных интенсивностях ремонтов.

В данном эксперименте будем использовать те же входные данные, что и в эксперименте 6.6. Будем также считать, что интенсивности ремонтов на приборе 1 и приборе 2 совпадают. Обозначим их общую интенсивность как τ. Значения h на оси абсцисс откладываются в соответствии с логарифмической (логарифм по основанию 10) шкалой.

На рисунке 6.13 изображена зависимость пропускной способности от интенсивности поступления поломок при разных интенсивностях ремонтов.

Из рисунка видно, что когда поломки поступают очень редко (h близко к нулю) и ремонт происходит очень быстро ($\tau = 10^{-6}$), то $\varrho \approx \mu_1 + \mu_2$, где μ_n — интенсивность обслуживания n-м прибором. Этот результат интуитивно понятен при анализе работы системы в режиме перегрузки и подтверждает выводы эксперимента 6.8 о том, что в таком случае запросы обслуживаются основными приборами и почти никогда не требуют участия в обслуживании резервного прибора. Из рисунка также можно заметить, что при увеличении интенсивности поступления поломок и уменьшении интенсивности ремонтов все кривые стремятся к горизонтальной асимптоте $\varrho = \mu_3$, где μ_3 — интенсивность обслуживания на резервном приборе, что совпадает с ожидаемым поведением системы в этом случае, так как основные

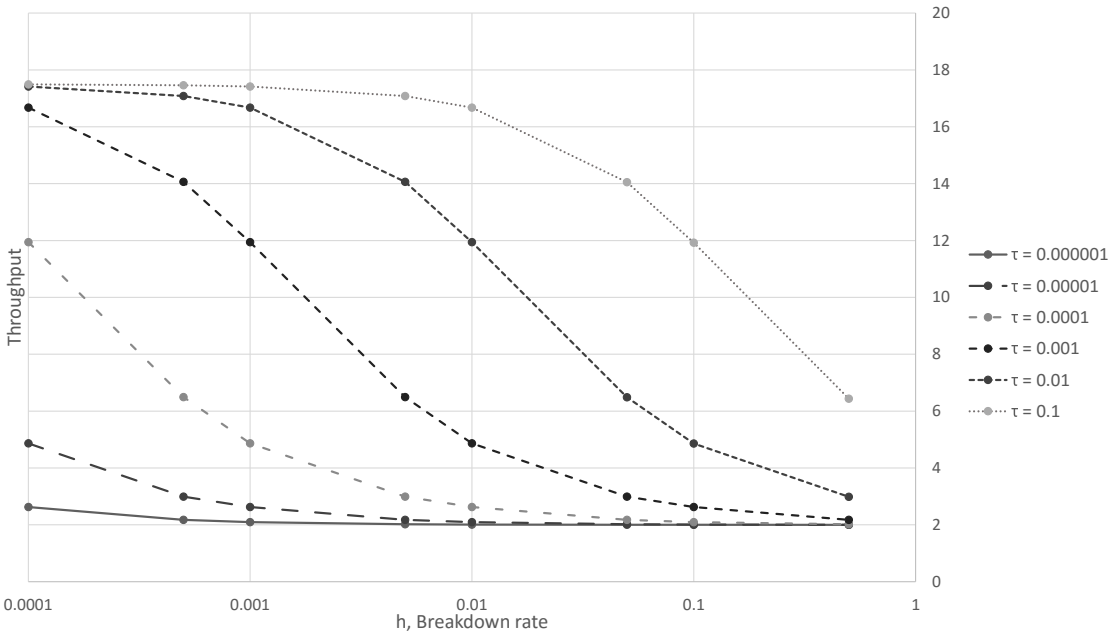

Рисунок 6.14. Зависимость пропускной способности системы ϱ от h при различных интенсивностях ремонтов $\tau(\tau = 10^{-n}, n = \overline{1,6})$

приборы почти всегда находятся на ремонте, а запросы обслуживаются на резервном третьем приборе. Также можно заметить, что пропускная способность не может превысить горизонтальную асимптоту $\varrho = \mu_1 + \mu_2$.

ГЛАВА 7
МНОГОФАЗНЫЕ СМО С КОРРЕЛИРОВАННЫМИ ВХОДНЫМИ ПОТОКАМИ И ИХ ПРИМЕНЕНИЕ ДЛЯ ОЦЕНКИ ПРОИЗВОДИТЕЛЬНОСТИ СЕТЕВЫХ СТРУКТУР

7.1 Краткий обзор работ по многофазным СМО с коррелированными потоками

Многофазные (тандемные) СМО предполагают наличие нескольких последовательных фаз (стадий, станций) обслуживания. В таких системах после получения обслуживания на некоторой фазе запрос переходит на обслуживание на следующую фазу обслуживания. Такие СМО адекватно моделируют функционирование широкополосных беспроводных сетей с линейной топологией (см. например [58–60]). Наиболее хорошо изучены в литературе дуальные тандемные системы, т.е. системы, состоящие из двух последовательных фаз обслуживания. Аналитическое исследование таких систем сводится к исследованию многомерного случайного процесса. В наиболее простых, но довольно редко встречающихся в реальных системах случаях стационарное распределение вероятностей состояний этого процесса имеет мультипликативную форму. При отсутствии стационарного распределения в такой форме и бесконечном размере буферов на обоих станциях тандема, даже в простейших предположениях о потоке и обслуживании (стационарный пуассоновский поток запросов, показательное распределение времен обслуживания на фазах тандема) проблема нахождения стационарного распределения вероятностей состояний тандема сводится к сложному анализу случайного блуждания в четверть-плоскости, который, как правило, проводится путем решения функционального уравнения для двумерной производящей функции распределения. Если как минимум один из двух буферов тандема конечен, входной поток является MAP-потоком, а распределения времен обслуживания имеют фазовый тип, то процесс, описывающий состояние тандема, принадлежит классу изученных в разделе 3.2 многомерных (векторных процессов гибели и размножения). Поэтому необходимость исследования тандемных систем, на-

ряду с исследованием однофазных систем с MAP-потоком и распределением времени обслуживания фазового типа, и явилась мотивацией введения в рассмотрение и исследования векторных процессов гибели и размножения, проведенного М. Ньютсом.

В работе [158] рассмотрен тандем двух систем с бесконечными буферами и экспоненциальным распределением времен обслуживания, на вход которого поступает MAP поток запросов. Функционирование системы описывается векторным процессом гибели и размножения, генератор которого является матрицей с блоками бесконечно возрастающей размерности. Первой компонентой векторного процесса гибели и размножения является суммарное число запросов на обоих фазах, а вторая компонента задает всевозможные варианты распределения этого числа между двумя фазами. Поскольку теория векторных процессов гибели и размножения строго развита только для блоков фиксированной размерности, результаты [158] являются в известной степени эвристическими. Поэтому мы будем рассматривать только дуальные тандемные системы с произвольным распределением времени обслуживания на первой фазе, в которых как минимум один буфер является конечным, входной поток являются MAP-потоком или $BMAP$-потоком, а распределение времени обслуживания на второй фазе тандема является распределением фазового типа.

Более просто исследуются системы, в которых поток является MAP, но не $BMAP$, потоком, произвольное распределение времени обслуживания допускается на второй фазе, а буфер первой фазы конечен. Такой анализ выполнен в серии работ А. Гомез-Коррала (см., например, [126, 127]). Относительная простота исследования объясняется тем, что первая фаза описывается СМО типа $MAP/PH/1/N$. Для такой СМО выходящий поток (являющийся входящим потоком на вторую фазу) является MAP-потоком, матрицы, характеризующие который, несложно выражаются через матрицы, задающие входной поток и процесс обслуживания.

Если же произвольное распределение времени обслуживания допускается именно на первой фазе, анализ системы усложняется, поскольку при описании вложенной по моментам окончания обслуживания на первой фазе цепи Маркова как цепи типа $M/G/1$ существенно более громоздким является описание динамики конечных компонент цепи. Поэтому значительная часть существующих работ предполагает декомпозицию тандема и приближенный анализ выходящего потока из первой фазы, который в этом случае не является $BMAP$-потоком. К числу работ, предлагающих

такой подход, относятся [121, 122, 132, 134, 172].

Точный аналитический анализ тандемных систем с системой типа $BMAP/G/1$ или $BMAP/G/1/N$ на первой фазе проведен на основе использования метода вложенных цепей Маркова с последующим применением аппарата процессов марковского восстановления в следующих работах. В работе [153] исследованы тандемные системы типа $BMAP/G/1 \rightarrow PH/1/M-1$ и $BMAP/G/1/N \rightarrow PH/1/M-1$, в которых при полной занятости конечного буфера второй фазы, описываемой однолинейной системой с распределением времени обслуживания запросов фазового типа, в момент окончания обслуживания запроса на первой фазе обслуженный запрос теряется. В статье [85] системы такого вида изучены в предположении, что обслуженный запрос не теряется, а блокирует первый прибор до тех пор, пока не появится свободное место на второй фазе. В [100] изучена модель с MAP-потоком, конечными буферами, произвольным распределением времени обслуживания на обеих фазах и потерей запросов после первой фазы при заполненности буфера второй фазы. В [101] изучена аналогичная модель с блокировкой первого прибора. В обоих этих работах для анализа используется метод вложенных цепей Маркова и аппарат процессов марковского восстановления. Вложенный процесс строится по моментам окончания обслуживания запросов на первой фазе. В силу предположения об общем виде распределения времени обслуживания на второй фазе, для обеспечения марковости вложенного процесса дополнительно предполагается, что в момент окончания обслуживания запроса на первой фазе время обслуживания на второй фазе начинается заново. В [154] рассмотрена система, аналогичная изученной в [153], но с обратной связью (feedback). Обратная связь по существующей в теории массового обслуживания терминологии предполагает, что после получения обслуживания на второй фазе тандема запрос может возвращаться для обслуживания на первую фазу. В [138] рассмотрена модель, аналогичная изученной в [153], но предполагающая отсутствие буфера на первой фазе. Запросы, заставшие прибор занятым, совершают повторные попытки попасть на обслуживание. В работе [84] рассмотрена модель, первая фаза которой описывается системой типа $BMAP/G/1$. Вторая фаза описана многолинейной системой без буфера, с показательным распределением времени обслуживания. Запрос, получивший обслуживание на первой фазе, требует для своего обслуживания случайного числа приборов второй фазы. При отсутствии необходимого числа приборов запрос либо теряется, либо временно блоки-

рует работу прибора на первой фазе обслуживания. Статья [139] содержит более развернутое исследование такой системы, включая анализ времен пребывания запроса на каждой из фаз тандема и в тандеме в целом. Результаты получены в простом и компактном виде. Они будут приведены ниже в разделе 6.2. В разделе 6.3 будут приведены результаты анализа аналогичной системы с повторными вызовами. Результаты подразделов 6.2 и 6.3 существенно используют результаты монографии [83].

В [140] проанализирован тандем, в котором на вторую фазу поступает дополнительный поток (кросс-траффик), для обслуживания которого проводится резервирование каналов. В [141] рассмотрен тандем, в котором первая фаза описывается системой типа $BMAP/G/1$, а вторая фаза описана произвольной цепью Маркова с конечным пространством состояний. В работах [79, 118, 142, 144, 148], рассмотрены дуальные тандемные СМО с $BMAP$ или MAP-потоком, в которых обе фазы являются многолинейными СМО. Большая часть из этих работ содержит анализ соответствующих тандемных СМО как моделей тех или иных реальных систем (систем технической поддержки, контакт-центров и т.д.).

7.2 Система $BMAP/G/1 \to \cdot/M/N/0$ с групповым занятием приборов второй фазы

7.2.1 Математическая модель

Рассматривается двухфазная система $BMAP/G/1 \to \cdot/M/N/0$. Первую фазу этой системы образует однолинейная СМО с ожиданием. В систему поступает $BMAP$-поток запросов, заданный управляющим процессом ν_t, $t \geq 0$, с пространством состояний $\{0, ..., W\}$ и матричной ПФ $D(z) = \sum_{k=0}^{\infty} D_k z^k, |z| \leq 1$. Если группа запросов размера $k \geq 1$ поступает в систему и застает обслуживающий прибор первой фазы свободным, то один из запросов начинает обслуживаться, а остальные располагаются в очереди неограниченной длины и ожидают обслуживания. Если прибор занят в момент поступления группы, то все запросы этой группы становятся в конец очереди.

Времена обслуживания запросов на первой фазе являются независимыми случайными величинами с функцией распределения (ФР) $B(t)$, ПЛС $\beta(s) = \int_0^{\infty} e^{-st} dB(t)$ и конечным первым моментом $b_1 = \int_0^{\infty} t dB(t)$.

После обслуживания на первой фазе запрос поступает на вторую фазу рассматриваемой системы, которую образуют N независимых идентичных приборов.

Для обслуживания на второй фазе запросу требуется случайное число приборов ϕ. Здесь ϕ – целочисленная случайная величина с распределением

$$q_m = P\{\phi = m\},\ q_m \geq 0,\ m = \overline{0, N},\ \sum_{m=0}^{N} q_m = 1.$$

Отметим, что запросы, поступившие в одной группе, могут потребовать для своего обслуживания различное число приборов. Предполагаем, что $q_0 \neq 1$, так как в этом случае рассматриваемая СМО сводится к классической однофазной системе $BMAP/G/1$, исследованной в разделе 4.1. После занятия запросом приборов второй фазы каждый из приборов независимо от других обслуживает этот запрос в течение экспоненциально распределенного времени с параметром μ.

Если в момент окончания обслуживания запроса прибором первой фазы необходимое число свободных приборов на второй фазе отсутствует, то с вероятностью p, $0 \leq p \leq 1$, запрос уходит из системы недообслуженным (теряется), а с дополнительной вероятностью ждет, пока освободится нужное число приборов. Период ожидания сопровождается блокировкой прибора первой фазы. Как крайние случаи при $p = 0$ мы имеем систему с блокировкой, при $p = 1$ – систему с потерями.

7.2.2 Стационарное распределение вложенной цепи Маркова

Пусть t_n – n-й момент окончания обслуживания на первой фазе, $n \geq 1$. Рассмотрим процесс $\xi_n = \{i_n, r_n, \nu_n\}$, $n \geq 1$, где i_n – число запросов на первой фазе (не включая запрос, вызвавший блокировку) в момент времени $t_n + 0$; r_n – число занятых приборов на второй фазе в момент времени $t_n - 0$; ν_n – состояние $BMAP$ в момент времени t_n, $i_n \geq 0, r_n = \overline{0, N}, \nu_n = \overline{0, W}$.

Процесс ξ_n, $n \geq 1$, является неприводимой ЦМ. Обозначим через $P\{(i, r, \nu) \to (j, r', \nu')\}$ вероятности переходов этой цепи. Перенумеруем все состояния в лексикографическом порядке и сформируем матрицы вероятностей переходов следующим образом:

$$P\{(i, r) \to (j, r')\} = (P\{(i, r, \nu) \to (j, r', \nu')\})_{\nu, \nu' = \overline{0, W}},$$

$$P_{i,j} = (P\{(i,r) \to (j,r')\})_{r,r'=\overline{0,N}},\ i,j \geq 0.$$

Матрицы $P_{i,j}$, $i,j \geq 0$, задают вероятности переходов цепи из состояний, соответствующих значению i счетной компоненты, в состояния, соответствующие значению j этой компоненты.

Лемма 7.1. *Матрица вероятностей переходов ЦМ $\xi_n, n \geq 1$, имеет блочную структуру вида (3.60), где*

$$V_i = \sum_{k=1}^{i+1}[-\hat{Q}(\Delta \oplus D_0)^{-1}\tilde{D}_k + (1-p)\mathcal{F}_k\tilde{Q}_3]\Omega_{i-k+1}, \tag{7.1}$$

$$Y_i = \bar{Q}\Omega_i + (1-p)\sum_{k=0}^{i}\mathcal{F}_k\tilde{Q}_3\Omega_{i-k}, \tag{7.2}$$

$$\Omega_n = \int_0^\infty e^{\Delta t} \otimes P(n,t)dB(t), \qquad \mathcal{F}_n = \int_0^\infty dF(t) \otimes P(n,t),\ n \geq 0,$$

$$\Delta = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ \mu & -\mu & 0 & \cdots & 0 & 0 \\ 0 & 2\mu & -2\mu & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & N\mu & -N\mu \end{pmatrix}.$$

Здесь $\tilde{D}_k = I_{N+1} \otimes D_k$, $k \geq 0$; $P(n,t)$, $n \geq 0$, – *матрицы, которые являются коэффициентами разложения* $e^{D(z)t} = \sum\limits_{n=0}^{\infty} P(n,t)z^n$; $F(t) = (F_{r,r'}(t))_{r,r'=\overline{0,N}}$, *где* $F_{r,r'}(t) = 0$, *если* $r \leq r'$ *и, если* $r > r'$, $F_{r,r'}(t)$ – *ФР, определенная ее ПЛС* $f_{r,r'}(s) = \prod\limits_{l=r'+1}^{r} l\mu(l\mu+s)^{-1}$; $\tilde{Q}_m = Q_m \otimes I_{\bar{W}}$, $m = \overline{1,3}$; Q_m, $m = \overline{1,3}$, – *квадратные матрицы порядка* $N+1$:

$Q_2 = \mathrm{diag}\{\sum\limits_{m=N-r+1}^{N} q_m, r = \overline{0,N}\}$,

$$Q_1 = \begin{pmatrix} q_0 & q_1 & \dots & q_N \\ 0 & q_0 & \dots & q_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & q_0 \end{pmatrix},\ Q_3 = \begin{pmatrix} 0 & \dots & 0 & q_N \\ 0 & \dots & 0 & q_{N-1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & q_0 \end{pmatrix};$$

$\bar{Q} = \tilde{Q}_1 + p\tilde{Q}_2$, $\hat{Q} = \tilde{Q}_1 + p\tilde{Q}_2 + (1-p)\int\limits_0^\infty (dF(t) \otimes e^{D_0 t})\tilde{Q}_3$.

Доказательство. Запишем матрицы переходных вероятностей V_i, Y_i в блочном виде $V_i = (V_i^{(r,r')})_{r,r'=\overline{0,N}}$, $Y_i = (Y_i^{(r,r')})_{r,r'=\overline{0,N}}$, где блоки $V_i^{(r,r')}$, $Y_i^{(r,r')}$ соответствуют переходам процесса r_n числа занятых приборов на второй фазе из состояния r в состояние r'.

Пусть $\delta_{r,r'}(t)$ определяет вероятность того, что за время t число запросов на второй фазе изменится с r на r' при условии, что новые запросы на эту фазу не поступают.

Поясним вероятностный смысл матриц, которые будут использоваться в дальнейшем.

(ν,ν')-й элемент матрицы $P(n,t)$ есть вероятность того, что на интервале $(0,t]$ поступит n запросов в $BMAP$, управляющий процесс $\nu_t = \nu'$ при условии $\nu_0 = \nu$, $\nu,\nu' = \overline{0,W}$.

(ν,ν')-й элемент матрицы $\int\limits_0^\infty \delta_{r,r'}(t)P(n,t)dB(t)$ есть вероятность того, что в течение времени обслуживания запроса прибором первой фазы поступит n запросов, число занятых приборов на второй фазе изменится с r на r' и состояние управляющего процесса $BMAP$ изменится с ν на ν'.

Пусть в некоторый момент времени процесс (ν_t, r_t) находится в состоянии (ν, r). Тогда (ν,ν')-й элемент матрицы $\int\limits_0^\infty \delta_{r,r'}(t)e^{D_0 t}D_k dt$ есть вероятность того, что первой в $BMAP$-потоке поступит группа размера k, процесс ν_t окажется в состоянии ν' непосредственно после момента поступления группы и число запросов на второй фазе равно r' в этот момент.

$F_{r,r'}(t)$ – ФР времени, в течение которого число занятых приборов на второй фазе изменится с r на r' при условии, что новые запросы на вторую фазу не поступают.

(ν,ν')-й элемент матрицы $\int\limits_0^\infty P(n,t)dF_{r,r'}(t)$ есть вероятность того, что за время, распределенное по закону $F_{r,r'}(t)$, в систему поступит n запросов и процесс ν_t перейдет из состояния ν в состояние ν'.

Принимая во внимание вероятностную интерпретацию матриц и анализируя поведение системы между двумя соседними моментами окончания обслуживания на первой фазе, получим следующие выражения для матриц $V_i^{(r,r')}$, $Y_i^{(r,r')}$, $i \geq 0$:

$$V_i^{(r,r')} = \sum_{m=0}^{N-r} q_m \sum_{l=r'}^{r+m} \int\limits_0^\infty \delta_{r+m,l}(t)e^{D_0 t}dt \sum_{k=1}^{i+1} D_k \int\limits_0^\infty \delta_{l,r'}(t)P(i-k+1,t)dB(t)+$$

$$+p\sum_{m=N-r+1}^{N} q_m \sum_{l=r'}^{r} \int_0^\infty \delta_{r,l}(t) e^{D_0 t} dt \sum_{k=1}^{i+1} D_k \int_0^\infty \delta_{l,r'}(t) P(i-k+1,t) dB(t) +$$

$$+(1-p)\sum_{m=N-r+1}^{N} q_m \Bigg[\int_0^\infty e^{D_0 t} dF_{r,N-m}(t) \sum_{l=r'}^{N} \int_0^\infty \delta_{N,l}(t) e^{D_0 t} dt \times$$

$$\times \sum_{k=1}^{i+1} D_k \int_0^\infty \delta_{l,r'}(t) P(i-k+1,t) dB(t) +$$

$$+\sum_{k=1}^{i+1} \int_0^\infty P(k,t) dF_{r,N-m}(t) \int_0^\infty \delta_{N,r'}(t) P(i-k+1,t) dB(t) \Bigg], \qquad (7.3)$$

$$Y_i^{(r,r')} = \sum_{m=0}^{N-r} q_m \int_0^\infty P(i,t) \delta_{r+m,r'}(t) dB(t) +$$

$$+\sum_{m=N-r+1}^{N} q_m \Bigg[p \int_0^\infty \delta_{r,r'}(t) P(i,t) dB(t) +$$

$$+(1-p)\sum_{k=0}^{i} \int_0^\infty P(k,t) dF_{r,N-m}(t) \int_0^\infty \delta_{N,r'}(t) P(i-k,t) dB(t) \Bigg]. \qquad (7.4)$$

Для того чтобы получить выражения (7.1) и (7.2) из (7.3) и (7.4), используем обозначения для матриц, введенные выше, и соотношение $(\delta_{r,r'}(t))_{r,r'=\overline{0,N}} = e^{\Delta t}$. Последнее соотношение следует из того, что, в случае, когда запросы на вторую фазу не поступают, процесс r_t, который описывает эволюцию числа занятых приборов на второй фазе, является процессом гибели с генератором Δ.

□

Следствие 7.1. *Процесс $\xi_n, n \geq 1$, является КТЦМ.*

Доказательство. ЦМ $\xi_n, n \geq 1$, является неприводимой и непериодической, матрицы переходных вероятностей $P_{i,j}$, $i > 0$, которой представимы как функции от разности $j - i$. Поэтому эта цепь относится к классу многомерных КТЦМ или ЦМ типа $M/G/1$. □

Следствие 7.2. *Матричные ПФ $V(z)$ и $Y(z)$ имеют вид*

$$V(z) = \frac{1}{z}[-\hat{Q}(\Delta \oplus D_0)^{-1}(\tilde{D}(z) - \tilde{D}_0) + (1-p)(\mathcal{F}(z) - \mathcal{F}_0)\tilde{Q}_3]\Omega(z), \quad (7.5)$$

$$Y(z) = [\bar{Q} + (1-p)\mathcal{F}(z)\tilde{Q}_3]\Omega(z), \quad (7.6)$$

где

$$\Omega(z) = \int_0^\infty e^{\Delta t} \otimes e^{D(z)t} dB(t),\ \mathcal{F}(z) = \int_0^\infty dF(t) \otimes e^{D(z)t},$$

$$\tilde{D}(z) = \sum_{k=0}^{\infty} \tilde{D}_k z^k,\ |z| \le 1.$$

Доказательство. Умножая (7.1) на z^i и суммируя по $i \ge 0$, после алгебраических преобразований получим выражение (7.5) для $V(z)$. Аналогично, используя (7.2), получим (7.6). □

Теорема 7.1. *Необходимым и достаточным условием существования стационарного распределения ЦМ ξ_n, $n \ge 1$, является выполнение неравенства*

$$\rho < 1, \quad (7.7)$$

где

$$\rho = \lambda[b_1 + (1-p)\sum_{r=1}^{N} y_r \sum_{m=N-r+1}^{N} q_m \sum_{l=N-m+1}^{r} (l\mu)^{-1}].$$

Здесь $(y_1, ..., y_N)$ – часть вектора $\mathbf{y} = (\mathbf{y_0}, \mathbf{y_1}, ..., \mathbf{y_N})$, который является единственным решением СЛАУ

$$\mathbf{y}\mathcal{Q}\mathcal{B}^*(0) = \mathbf{y}, \quad \mathbf{ye} = \mathbf{1}, \quad (7.8)$$

где

$$\mathcal{Q} = Q_1 + pQ_2 + (1-p)EQ_3,\ \mathcal{B}^*(s) = \int_0^\infty e^{-st} e^{\Delta t} dB(t).$$

Доказательство. Матрица $Y(1)$ отражает эволюцию $BMAP$ и числа занятых приборов на второй фазе между соседними моментами окончания обслуживания на первой фазе. Учитывая, что процесс ν_t, $t \ge 0$, является неприводимым и по предположению $q_0 \ne 1$, нетрудно показать, что матрица $Y(1)$ является неприводимой. Тогда необходимым и достаточным

условием эргодичности КТЦМ ξ_n является выполнение неравенства (3.66), где вектор $\mathbf{x}$ является единственным решением СЛАУ (3.67).

Покажем, что неравенство (3.66) эквивалентно неравенству (7.7).

Перепишем систему (7.8) в виде

$$\mathbf{y}[Q_1 + pQ_2 + (1-p)\int\limits_0^\infty dF(t)Q_3]\int\limits_0^\infty e^{\Delta t}dB(t) = \mathbf{y}, \quad \mathbf{ye} = \mathbf{1},$$

а систему (3.67) как

$$\mathbf{x}[(Q_1+pQ_2)\otimes I_{\bar{W}}+(1-p)\int\limits_0^\infty dF(t)Q_3\otimes e^{D(1)t}]\int\limits_0^\infty e^{\Delta t}\otimes e^{D(1)t}dB(t)=\mathbf{x},\ \mathbf{xe}=1. \tag{7.9}$$

Анализируя две последние СЛАУ, легко убедиться, что вектор

$$\mathbf{x} = \mathbf{y} \otimes \boldsymbol{\theta} \tag{7.10}$$

является единственным решением системы (7.9) и, следовательно, эквивалентной ей системы (3.67).

Дифференцируя (7.6) в точке $z = 1$ и подставляя полученное выражение для $Y'(1)$ и вектор $\mathbf{x}$ вида (7.10) в неравенство (3.66), получим

$$\lambda[b_1 + (1-p)\mathbf{y}\int\limits_0^\infty tdF(t)Q_3\mathbf{e}] < \mathbf{1}. \tag{7.11}$$

Учитывая, что $\int\limits_0^\infty tdF(t)$ – это матрица, у которой (r, r')-й элемент равен $\sum\limits_{l=r'+1}^{r}(l\mu)^{-1}$ при $r > r'$, а все остальные элементы равны нулю, неравенство (7.11) легко свести к неравенству (7.7). □

Замечание 7.1. Вектор $\mathbf{y}$ задает стационарное распределение числа занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе при условии, что прибор первой фазы работает без остановки. Тогда выражение $(1-p)\sum\limits_{r=1}^{N} y_r \sum\limits_{m=N-r+1}^{N} q_m \sum\limits_{l=N-m+1}^{r}(l\mu)^{-1}$ определяет среднее время блокировки прибора первой фазы в условиях перегрузки системы, а величина ρ является коэффициентом загрузки.

В дальнейшем будем предполагать, что неравенство (7.7) выполняется. Тогда существует стационарное распределение ЦМ ξ_n, $n \geq 1$, совпадающее с ее эргодическим распределением

$$\pi(i,r,\nu) = \lim_{n\to\infty} P\{i_n = i, r_n = r, \nu_n = \nu\},\ i \geq 0,\ r = \overline{0,N},\ \nu = \overline{0,W}.$$

Введем обозначения для векторов стационарных вероятностей

$$\boldsymbol{\pi}(i,r) = (\pi(i,r,0), \pi(i,r,1), \ldots, \pi(i,r,W)),$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1), \ldots, \boldsymbol{\pi}(i,N)),\ i \geq 0.$$

Для вычисления векторов $\boldsymbol{\pi}_i$, $i \geq 0$, используется алгоритм, приведенный в разделе 3.4.

7.2.3 Стационарное распределение в произвольный момент времени

Определим состояние системы в произвольный момент времени t как $(i_t,\ k_t,\ r_t,\ \nu_t)$, где i_t – число запросов на первой фазе (включая запрос, вызвавший блокировку); k_t – величина, принимающая значения 0, 1, 2 в зависимости от того, свободен, обслуживает запрос или блокирован прибор первой фазы; r_t – число занятых приборов на второй фазе; ν_t – состояние $BMAP$-потока в момент времени t, $t \geq 0$.

Рассмотрим процесс $\zeta_t = \{i_t, k_t, r_t, \nu_t\}$, $t \geq 0$. Этот процесс не является марковским, но его стационарное распределение удается выразить через стационарное распределение вложенной ЦМ ξ_n, $n \geq 1$, используя результаты для процессов марковского восстановления и полурегенерирующих процессов, см. [112].

Пусть $p(i,k,r,\nu),\ i \geq 0,\ k = 0,1,2,\ r = \overline{0,N},\ \nu = \overline{0,W}$, – стационарные вероятности процесса ζ_t, $t \geq 0$. Определим векторы этих вероятностей как

$$\mathbf{p}(i,k,r) = (p(i,k,r,0), ..., p(i,k,r,W)),\ \mathbf{p}_i(k) = (\mathbf{p}(i,k,0), ..., \mathbf{p}(i,k,N)).$$

Теорема 7.2. *Ненулевые векторы стационарных вероятностей* $\mathbf{p}_i(k)$, $i \geq 0$, $k = 0,1,2$, *выражаются через векторы стационарного распределения* $\boldsymbol{\pi}_i$, $i \geq 0$, *вложенной ЦМ* ξ_n, $n \geq 1$, *следующим образом:*

$$\mathbf{p}_0(0) = -\tau^{-1}\boldsymbol{\pi}_0\hat{Q}(\Delta \oplus D_0)^{-1},$$

$$\mathbf{p}_i(1) = \tau^{-1}\{\boldsymbol{\pi}_0 \sum_{k=1}^{i}[-\hat{Q}(\Delta \oplus D_0)^{-1}\tilde{D}_k + (1-p)\mathcal{F}_k\tilde{Q}_3]\tilde{\Omega}_{i-k}+$$

$$+\sum_{l=1}^{i}\boldsymbol{\pi}_l[\bar{Q}\tilde{\Omega}_{i-l} + (1-p)\sum_{k=0}^{i-l}\mathcal{F}_k\tilde{Q}_3\tilde{\Omega}_{i-k-l}]\},$$

$$\mathbf{p}_i(2) = \tau^{-1}(1-p)\sum_{l=0}^{i-1}\boldsymbol{\pi}_l\sum_{k=0}^{i-l-1}(\mathcal{F}_k+\hat{\delta}_{0,k}I)\tilde{Q}_2\int_0^\infty e^{-\mu\mathcal{N}t}\otimes P(i-k-l-1,t)dt,\ i\geq 1,$$

где τ – средняя величина интервала между двумя соседними моментами окончания обслуживания на первой фазе:

$$\tau = b_1 + \boldsymbol{\pi}_0\hat{Q}(-\tilde{D}_0)^{-1}\mathbf{e} + (1-p)\boldsymbol{\Pi}(1)(I_{N+1}\otimes\mathbf{e})\int_0^\infty tdF(t)Q_3\mathbf{e}, \qquad (7.12)$$

$$\tilde{\Omega}_n = \int_0^\infty e^{\Delta t}\otimes P(n,t)(1-B(t))dt,\ n\geq 0.$$

Доказательство. Согласно определению, данному в [112], процесс ζ_t, $t \geq 0$, является полурегенерирующим с вложенным процессом марковского восстановления $\{\xi_n, t_n\}$, $n \geq 1$.

Стационарное распределение $p(i,k,r,\nu)$, $i \geq 0$, $k = 0,1,2$, $r = \overline{0,N}$, $\nu = \overline{0,W}$, процесса ζ_t существует, если процесс $\{\xi_n, t_n\}$ является неприводимым непериодическим и выполняется неравенство $\tau < \infty$. В нашем случае все эти условия выполняются, если вложенная ЦМ ξ_n, $n \geq 1$, эргодична.

Сначала найдем векторы $\mathbf{p}(i,k,r)$, $i \geq 0$, $k = 0,1,2$, $r = \overline{0,N}$, стационарных вероятностей $p(i,k,r,\nu)$. Чтобы вычислить эти векторы, будем использовать предельную теорему для полурегенерирующих процессов, приведенную в [112]. Согласно этой теореме, стационарное распределение процесса ζ_t, $t \geq 0$, можно выразить через стационарное распределение вложенной ЦМ $\xi_n, n \geq 1$.

Введем в рассмотрение $\bar{W}\times\bar{W}$-матрицы

$$K_{l,r}(i,k,r',t),\ l,i\geq 0,\ k=0,1,2,\ r,r'=\overline{0,N}. \qquad (7.13)$$

(ν,ν')-й элемент этих матриц определяет вероятность того, что некоторый запрос закончит обслуживание позже момента времени t и процесс ζ_t, $t \geq$

0, будет находиться в состоянии (i, r', ν', k) в момент t при условии, что предыдущий запрос закончил обслуживание в момент времени $t = 0$ и в этот момент вложенная ЦМ ξ_n находилась в состоянии (l, r, ν).

По предельной теореме векторы $\mathbf{p}(i, k, r)$ выражаются через стационарное распределение $\boldsymbol{\pi}(i, r),\ i \geq 0,\ r = \overline{0, N}$, цепи ξ_n следующим образом:

$$\mathbf{p}(i, k, r') = \tau^{-1} \sum_{l=0}^{\infty} \sum_{r=0}^{N} \boldsymbol{\pi}(l, r) \int_0^{\infty} K_{l,r}(i, k, r', t)dt, i{\geq}0, k{=}0, 1, 2, r'{=}\overline{0, N}. \tag{7.14}$$

Для дальнейшего доказательства нужно найти выражения для матриц (7.13) и величины τ.

Средняя длина τ интервала между двумя соседними моментами окончания обслуживания на первой фазе определяется формулой

$$\tau = \sum_{i=0}^{\infty} \sum_{r=0}^{N} \sum_{\nu=0}^{W} \boldsymbol{\pi}(i, r, \nu)\tau_{i,r,\nu},$$

где $\tau_{i,r,\nu}$ – это средняя длина интервала между двумя соседними моментами окончания обслуживания на первой фазе при условии, что в начале этого интервала $\xi_n = (i, r, \nu)$. Путем алгебраических преобразований можно убедиться, что величина τ вычисляется по формуле (7.12).

Далее, анализируя поведение системы между двумя соседними моментами окончания обслуживания на первой фазе и принимая во внимание вероятностный смысл функций и матриц, приведенный в доказательстве леммы 7.1, получим следующие выражения для ненулевых матриц (7.13):

$$K_{0,r}(0, 0, r', t) = \sum_{m=\max\{0, r'-r\}}^{N-r} q_m e^{D_0 t}\delta_{r+m,r'}(t)+$$

$$+ \sum_{m=N-r+1}^{N} q_m[pe^{D_0 t}\delta_{r,r'}(t) + (1-p)\int_0^t e^{D_0 s}dF_{r,N-m}(s)e^{D_0(t-s)}\delta_{N,r'}(t-s)],$$

$$K_{0,r}(i, 1, r', t) = \sum_{m=\max\{0, r'-r\}}^{N-r} q_m \sum_{l=r'}^{r+m} \int_0^t e^{D_0 s} \times$$

$$\times\delta_{r+m,l}(s) \sum_{k=1}^{i} D_k ds\delta_{l,r'}(t-s)P(i-k, t-s)(1-B(t-s))+$$

$$+\sum_{m=N-r+1}^{N} q_m\{p\sum_{l=r'}^{r}\int_0^t e^{D_0 s}\delta_{r,l}(s)\sum_{k=1}^{i} D_k ds\delta_{l,r'}(t-s)P(i-k,t-s)(1-B(t-s))+$$

$$+(1-p)[\int_0^t e^{D_0 s}dF_{r,N-m}(s)\sum_{l=r'}^{N}\int_0^{t-s} e^{D_0 x}\delta_{N,l}(x)\sum_{k=1}^{i} D_k dx\times$$

$$\times\delta_{l,r'}(t-s-x)P(i-k,t-s-x)(1-B(t-s-x))+$$

$$+\sum_{k=1}^{i}\int_0^t P(k,s)dF_{r,N-m}(s)\delta_{N,r'}(t-s)P(i-k,t-s)(1-B(t-s))]\},$$

$$K_{l,r}(i,1,r',t)=\sum_{m=\max\{0,r'-r\}}^{N-r} q_m\int_0^t \delta_{r+m,r'}(s)P(i-l,s)(1-B(s))ds+$$

$$+\sum_{m=N-r+1}^{N} q_m[p\int_0^t \delta_{r,r'}(s)P(i-l,s)(1-B(s))ds+$$

$$+(1-p)\sum_{k=0}^{i-l}\int_0^t P(k,s)dF_{r,N-m}(s)\delta_{N,r'}(t-s)\times$$

$$\times P(i-l-k,t-s)(1-B(t-s))], i\geq l\geq 1,$$

$$K_{l,r}(i,2,r',t)=(1-p)\sum_{m=N+1-\min\{r,r'\}}^{N} q_m\sum_{k=0}^{i-l-1}\int_0^t P(k,s)dF_{r,r'}(s)\times$$

$$\times P(i-k-l-1,t-s)e^{-r'\mu(t-s)},\ i\geq 1,\ l\geq 0,\ r,r'=\overline{0,N}.$$

Подставляя полученные выражения для матриц $K_{l,r}(i,k,r',t)$ в (7.14), после ряда алгебраических преобразований, включающих изменение порядка интегрирования, получим следующие формулы:

$$\mathbf{p}(0,0,r')=\tau^{-1}\sum_{r=0}^{N}\boldsymbol{\pi}(0,r)[\sum_{m=\max\{0,r'-r\}}^{N-r} q_m\int_0^\infty \delta_{r+m,r'}(t)e^{D_0 t}dt+$$

$$+p\sum_{m=N-r+1}^{N} q_m\int_0^\infty \delta_{r,r'}(t)e^{D_0 t}dt+$$

$$+(1-p)\sum_{m=N-r+1}^{N} q_m \int_0^{\infty} e^{D_0 t} dF_{r,N-m}(t) \int_0^{\infty} \delta_{N,r'}(t) e^{D_0 t} dt],$$

$$\mathbf{p}(i,1,r') = \tau^{-1}\{\sum_{r=0}^{N} \boldsymbol{\pi}(0,r)\{\sum_{k=1}^{i}[\sum_{m=\max\{0,r'-r\}}^{N-r} q_m \sum_{l=r'}^{r+m} \int_0^{\infty} \delta_{r+m,l}(t) e^{D_0 t} dt+$$

$$+p\sum_{m=N-r+1}^{N} q_m \sum_{l=r'}^{r} \int_0^{\infty} \delta_{r,l}(t) e^{D_0 t} dt+$$

$$+(1-p)\sum_{m=N-r+1}^{N} q_m \int_0^{\infty} e^{D_0 t} dF_{r,N-m}(t) \sum_{l=r'}^{N} \int_0^{\infty} \delta_{N,l}(t) e^{D_0 t} dt]\times$$

$$\times D_k \int_0^{\infty} \delta_{l,r'}(t) P(i-k,t)(1-B(t))dt+$$

$$+(1-p)\sum_{m=N-r+1}^{N} q_m \sum_{k=1}^{i} \int_0^{\infty} P(k,t) dF_{r,N-m}(t)\times$$

$$\times \int_0^{\infty} \delta_{N,r'}(t) P(i-k,t)(1-B(t))dt\}+$$

$$+\sum_{l=1}^{i}\sum_{r=0}^{N} \boldsymbol{\pi}(l,r)[\sum_{m=\max\{0,r'-r\}}^{N-r} q_m \int_0^{\infty} \delta_{r+m,r'}(t) P(i-l,t)(1-B(t))dt+$$

$$+p\sum_{m=N-r+1}^{N} q_m \int_0^{\infty} \delta_{r,r'}(t) P(i-l,t)(1-B(t))dt+$$

$$+(1-p)\sum_{m=N-r+1}^{N} q_m \sum_{k=0}^{i-l} \int_0^{\infty} P(k,t) dF_{r,N-m}(t)\times$$

$$\times \int_0^{\infty} \delta_{N,r'}(t) P(i-l-k,t)(1-B(t))dt]\},$$

$$\mathbf{p}(i,2,r') = \tau^{-1}(1-p)\sum_{l=0}^{i-1}\sum_{r=0}^{N} \boldsymbol{\pi}(l,r) \sum_{m=N+1-\min\{r,r'\}}^{N} q_m \times$$

$$\times \sum_{k=0}^{i-l-1} \int_0^\infty P(k,t)dF_{r,r'}(t) \int_0^\infty P(i-k-l-1,t)e^{-r'\mu t}dt, i \geq 1,\ r' = \overline{0,N}.$$

Очевидно, что $\mathbf{p}(0,1,r') = \mathbf{p}(0,2,r') = \mathbf{p}(i,0,r') = \mathbf{0}$, $i \geq 1$.

Объединяя векторы $\mathbf{p}(i,k,r')$, $r' = \overline{0,N}$, в векторы $\mathbf{p}_i(k)$ для всех $i \geq 0$, $k = 0,1,2$, и используя матричные обозначения, введенные ранее, получим формулы для векторов $\mathbf{p}_i(k)$. □

Следствие 7.3. *Векторы стационарных вероятностей* $\mathbf{p}_i$, $i \geq 0$, *процесса* $\{i_t, r_t, \nu_t\}$, $t \geq 0$, *вычисляются следующим образом:*

$$\mathbf{p}_i = \sum_{k=0}^{2} \mathbf{p}_i(k),\ i \geq 0.$$

Следствие 7.4. *ПФ* $\mathbf{P}(z) = \sum_{i=0}^{\infty} \mathbf{p}_i z^i$, $|z| \leq 1$, *выражается через ПФ* $\mathbf{\Pi}(z)$ *следующим образом:*

$$\mathbf{P}(z)(\Delta \oplus D(z)) = \tau^{-1}\mathbf{\Pi}(z)\{z[I - (1-p)\mathcal{F}(z)\tilde{Q}_2 \times$$

$$\times(\mu\mathcal{N} \oplus (-D(z)))^{-1}(\Delta \oplus D(z))] - [\bar{Q} + (1-p)\mathcal{F}(z)\tilde{Q}_3]\}. \qquad (7.15)$$

Для краткости перепишем (7.15) в виде

$$\mathbf{P}(z)\mathcal{A}(z) = \boldsymbol{b}(z), \qquad (7.16)$$

где

$$\mathcal{A}(z) = \Delta \oplus D(z),$$

$$\boldsymbol{b}(z) = \tau^{-1}\mathbf{\Pi}(z)\{z[I - (1-p)\mathcal{F}(z)\tilde{Q}_2(\mu\mathcal{N} \oplus (-D(z)))^{-1}(\Delta \oplus D(z))] -$$

$$-[\bar{Q} + (1-p)\mathcal{F}(z)\tilde{Q}_3]\}.$$

Используя соотношение (7.16), можно вычислить векторные факториальные моменты $\mathbf{P}^{(m)} = \frac{d^m\mathbf{P}(z)}{dz^m}|_{z=1}$, $m = \overline{0,M}$, по векторным факториальным моментам $\mathbf{\Pi}^{(m)} = \frac{d^m\mathbf{\Pi}(z)}{dz^m}|_{z=1}$, $m = \overline{0,M+1}$. Векторы $\mathbf{\Pi}^{(m)}$ можно вычислить непосредственно по формуле $\mathbf{\Pi}^{(m)} = \sum_{i=m}^{\infty} \frac{i!}{(i-m)!}\boldsymbol{\pi}_i$, $m \geq 1$, зная стационарное распределение $\boldsymbol{\pi}_i$, $i \geq 0$, либо для расчета этих моментов можно применить реккурентную процедуру.

Теорема 7.3. *Пусть* $\mathbf{\Pi}^{(m)} < \infty$, $m = \overline{0, M+1}$. *Векторы* $\mathbf{P}^{(m)}$, $m = \overline{0, M}$, *вычисляются рекуррентно*

$$\mathbf{P}^{(m)}(1) = \left[\left(\boldsymbol{b}^{(m)}(1) - \sum_{l=0}^{m-1} C_m^l \mathbf{P}^{(l)}(1)\mathcal{A}^{(m-l)}(1)\right)\tilde{I} + \right. \tag{7.17}$$

$$\left. + \frac{1}{m+1}\left(\boldsymbol{b}^{(m+1)}(1) - \sum_{l=0}^{m-1} C_{m+1}^l \mathbf{P}^{(l)}(1)\mathcal{A}^{(m+1-l)}(1)\right)\mathbf{e}\hat{\mathbf{e}}\right]\tilde{\mathcal{A}}^{-1},$$

где

$$\tilde{\mathcal{A}} = \mathcal{A}(1)\tilde{I} + \mathcal{A}'(1)\mathbf{e}\hat{\mathbf{e}}.$$

Доказательство. Последовательно дифференцируя (7.16), получим:

$$\mathbf{P}^{(m)}(1)\mathcal{A}(1) = \boldsymbol{b}^{(m)}(1) - \sum_{l=0}^{m-1} C_m^l \mathbf{P}^{(l)}(1)\mathcal{A}^{(m-l)}(1). \tag{7.18}$$

Матрица $\mathcal{A}(1)$ является вырожденной, поэтому не представляется возможным получить рекурсию для вычисления векторных факториальных моментов непосредственно из (7.18).

Модифицируем (7.18) следующим образом. Умножим справа обе части выражения (7.18) для $m+1$ на $\mathbf{e}$. Далее заменим одно из уравнений системы (7.18) (без ограничения общности первое) полученным уравнением, в результате имеем

$$\mathbf{P}^{(m)}(1)\tilde{\mathcal{A}} = \left(\boldsymbol{b}^{(m)}(1) - \sum_{l=0}^{m-1} C_m^l \mathbf{P}^{(l)}(1)\mathcal{A}^{(m-l)}(1)\right)\tilde{I} + \tag{7.19}$$

$$+ \frac{1}{m+1}\left(\boldsymbol{b}^{(m+1)}(1) - \sum_{l=0}^{m-1} C_{m+1}^l \mathbf{P}^{(l)}(1)\mathcal{A}^{(m+1-l)}(1)\right)\mathbf{e}\hat{\mathbf{e}}.$$

Рекуррентная процедура (7.17) получается непосредственно из (7.19), если матрица $\tilde{\mathcal{A}}$ является невырожденной. Докажем, что $\det\tilde{\mathcal{A}} \neq 0$.

Можно показать, что $\det\tilde{\mathcal{A}} = \det\mathcal{D}\prod_{n=1}^{N}\det(n\mu I - D(1))$, где $\mathcal{D} = D(1)\tilde{I} + D'(1)\mathbf{e}\hat{\mathbf{e}}$. Определители $\det(n\mu I - D(1))$, $n = \overline{1, N}$, не равны нулю в силу того, что диагональные элементы матриц $n\mu I - D(1)$, $n = \overline{1, N}$, доминируют по строкам.

Анализируя структуру матрицы $\mathcal{D}$, можно убедиться, что $\det\mathcal{D} = \boldsymbol{\nabla}D'(1)\mathbf{e}$, где $\boldsymbol{\nabla}$ – вектор алгебраических дополнений первого столбца матрицы $D(1)$. Так как $\boldsymbol{\theta}$ – инвариантный вектор матрицы $D(1)$, то $\boldsymbol{\nabla} = c\boldsymbol{\theta}$, $c \neq 0$. Тогда $\det\mathcal{D} = c\boldsymbol{\theta}D'(1)\mathbf{e} = c\lambda$. Следовательно, $\det\tilde{\mathcal{A}} \neq 0$. □

7.2.4 Характеристики производительности системы

Среднее число запросов на первой фазе в момент окончания обслуживания на этой фазе и в произвольный момент времени:

$$L = \mathbf{\Pi}'(1)\mathbf{e}, \quad \tilde{L} = \mathbf{P}'(1)\mathbf{e}.$$

Векторы стационарного распределения числа занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$\boldsymbol{r} = \mathbf{\Pi}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}}), \qquad \tilde{\boldsymbol{r}} = \mathbf{P}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}}).$$

Среднее число занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$N_{busy} = \boldsymbol{r}\mathcal{N}\mathbf{e}, \qquad \tilde{N}_{busy} = \tilde{\boldsymbol{r}}\mathcal{N}\mathbf{e}.$$

Вероятность того, что произвольный запрос покинет систему (вызовет блокировку) после обслуживания на первой фазе:

$$P_{loss} = p\mathbf{\Pi}(1)\tilde{Q}_2\mathbf{e}, \quad P_{block} = (1-p)\mathbf{\Pi}(1)\tilde{Q}_2\mathbf{e}.$$

Вероятности P_{idle}, P_{serve}, P_{block} того, что прибор первой фазы свободен, занят обслуживанием или заблокирован:

$$P_{idle} = \tau^{-1}\boldsymbol{\pi}_0\hat{Q}(-\tilde{D}_0)^{-1}\mathbf{e}, \; P_{serve} = \tau^{-1}b_1, \; P_{block} = 1 - P_{idle} - P_{serve}.$$

7.2.5 Стационарное распределение времени пребывания

В данном подразделе рассматриваются задачи нахождения ПЛС стационарного распределения виртуального и реального времен пребывания в системе.

Виртуальное время пребывания. Виртуальное время пребывания в системе состоит из виртуального времени пребывания на первой фазе и времени пребывания на второй фазе. Полагаем, что запросы обслуживаются согласно дисциплине $FIFO$ (first-in-first-out).

Виртуальное время пребывания на первой фазе состоит из:

– *времени дообслуживания*, начинающегося в произвольный момент (момент поступления виртуального запроса) и заканчивающегося в ближайший момент окончания обслуживания на первой фазе;

– *обобщенных времен обслуживания* запросов, ожидающих обслуживания на первой фазе в момент поступления виртуального запроса;

– *обобщенного времени обслуживания* самого виртуального запроса.

Примечание 2 – Обобщенное время обслуживания запроса состоит из времени обслуживания этого запроса прибором первой фазы и возможного времени блокировки прибора предыдущим запросом.

Изучим сначала *время дообслуживания*. Для этого рассмотрим процесс $\chi_t = \{i_t,\, k_t,\, r_t,\, \nu_t,\, \tilde{v}_t\}$, $t \geq 0$, где i_t – число запросов на первой фазе (включая заблокированный), k_t – величина, принимающая значения 0, 1, 2 в зависимости от того, свободен, обслуживает запрос или блокирован прибор первой фазы, ν_t – состояние $BMAP$-потока в момент t; r_t – число занятых приборов на второй фазе перед моментом окончания обслуживания на первой фазе, следующим за моментом t; $\tilde{v}_t$ – время дообслуживания с момента времени t до упомянутого момента окончания обслуживания.

Согласно определению, данному в [112], процесс χ_t является полурегенерирующим с вложенным процессом марковского восстановления $\{\xi_n, t_n\}$, $n \geq 1$.

Обозначим стационарное распределение процесса χ_t, $t \geq 0$, как

$$\tilde{V}(i,k,r,\nu,x) = \lim_{t\to\infty} P\{i_t = i, k_t = k, r_t = r, \nu_t = \nu, \tilde{v}_t < x\}, \qquad (7.20)$$

$$i \geq 0,\ k = 0,1,2,\ r = \overline{0,N}, \nu = \overline{0,W},\ x \geq 0.$$

Согласно [112], пределы (7.20) существуют, если процесс $\{\xi_n, t_n\}$, $n \geq 1$, является неприводимым непериодическим и выполняется неравенство $\tau < \infty$. В нашем случае все эти условия выполнены, если выполнено неравенство (7.7).

Пусть $\tilde{\boldsymbol{V}}(i,k,x)$ – вектор-строки стационарных вероятностей $\tilde{V}(i,k,r,\nu,x)$, упорядоченных в лексикографическом порядке компонент (r,ν), и $\tilde{\boldsymbol{v}}(i,k,s) = \int\limits_0^\infty e^{-sx} d\tilde{\boldsymbol{V}}(i,k,x)$, $i \geq 0$, $k = 0,1,2$, – соответствующие векторные ПЛС.

Лемма 7.2. *Векторные ПЛС* $\tilde{\boldsymbol{v}}(i,k,s)$ *вычисляются следующим образом:*

$$\tilde{\boldsymbol{v}}(0,1,s) = \boldsymbol{0},\ \tilde{\boldsymbol{v}}(i,0,s) = \boldsymbol{0},\ \ i > 0,\ \tilde{\boldsymbol{v}}(0,2,s) = \boldsymbol{0}, \qquad (7.21)$$

$$\tilde{\boldsymbol{v}}(0,0,s) = -\tau^{-1}\boldsymbol{\pi}_0 \hat{Q}(\Delta \oplus D_0)^{-1}[\mathcal{B}^*(s) \otimes I_{\bar{W}}], \qquad (7.22)$$

$$\tilde{\boldsymbol{v}}(i,1,s) = \tau^{-1}\{\boldsymbol{\pi}_0 \sum_{k=1}^{i}[-\hat{Q}(\Delta \oplus D_0)^{-1}\tilde{D}_k + (1-p)\mathcal{F}_k\tilde{Q}_3]\times$$

$$\times \int_0^\infty (e^{\Delta u} \otimes I_{\bar{W}}) \int_0^u I_{N+1} \otimes P(i-k,y) e^{-s(u-y)} dy dB(u) +$$

$$+ \sum_{j=1}^{i} \boldsymbol{\pi}_j [(\tilde{Q}_1 + p\tilde{Q}_2) \int_0^\infty (e^{\Delta u} \otimes I_{\bar{W}}) \int_0^u I_{N+1} \otimes P(i-j,y) e^{-s(u-y)} dy dB(u) +$$

$$+ (1-p) \sum_{k=0}^{i-j} \mathcal{F}_k \tilde{Q}_3 \int_0^\infty e^{\Delta u} \otimes P(i-k-j,y) e^{-s(u-y)} dy dB(u)]\}, \qquad (7.23)$$

$$\tilde{\boldsymbol{v}}(i,2,s) = \tau^{-1}(1-p) \sum_{j=0}^{i-1} \boldsymbol{\pi}_j (\int_0^\infty dF(u) \otimes I_{\bar{W}}) \tilde{Q}_3 \times$$

$$\times \int_0^u e^{-s(u-y)} [I_{N+1} \otimes P(i-j-1,y)] dy [\mathcal{B}^*(s) \otimes I_{\bar{W}}],\ i > 0. \qquad (7.24)$$

Доказательство. Пусть $\kappa_j^{(i,k,x,t)}(r,\nu;\, r',\nu')$ определяет вероятность того, что некоторый запрос закончит обслуживание позже момента времени t, дискретные компоненты процесса χ_t примут значения $(i,\, k,\, r',\, \nu')$ в момент времени t, а непрерывная компонента $\tilde{v}_t < x$, при условии, что предыдущий запрос закончил обслуживание в момент времени $t = 0$ и в этот момент вложенная ЦМ ξ_n находилась в состоянии (j, r, ν).

Упорядочим вероятности $\kappa_j^{(i,k,x,t)}(r,\nu;\, r',\nu')$ при фиксированных значениях $i,\, j,\, k$, в лексикографическом порядке компонент $(r,\nu;\, r',\nu')$ и сформируем квадратные матрицы

$$\tilde{K}_j(i,k,x,t) = (\kappa_j^{(i,k,x,t)}(r,\nu;\, r',\nu'))_{\nu,\nu'=\overline{0,W};r,r'=\overline{0,N}},\ i \geq 0,\ k = 0,1,2.$$

Тогда, используя предельную теорему для полурегенерирующих процессов, см. [112], векторы $\tilde{\boldsymbol{V}}(i,k,x)$ можно выразить через стационарное распределение $\boldsymbol{\pi}_j$, $j \geq 0$, вложенной ЦМ ξ_n, $n \geq 1$, следующим образом:

$$\tilde{\boldsymbol{V}}(i,k,x) = \tau^{-1} \sum_{j=0}^{\infty} \boldsymbol{\pi}_j \int_0^\infty \tilde{K}_j(i,k,x,t) dt,\ i \geq 0,\ k = 0,1,2.$$

Соответствующие векторные ПЛС $\tilde{\boldsymbol{v}}(i,k,s)$ определяются формулой

$$\tilde{\boldsymbol{v}}(i,k,s) = \tau^{-1} \sum_{j=0}^{\infty} \boldsymbol{\pi}_j \int_0^\infty \tilde{K}_j^*(i,k,s,t) dt,\ i \geq 0,\ k = 0,1,2, \qquad (7.25)$$

где $\tilde{K}_j^*(i,k,s,t) = \int\limits_0^\infty e^{-sx} d\tilde{K}_j(i,k,x,t)$.

Формулы (7.21) следуют непосредственно из (7.25), учитывая, что $\tilde{K}_j^*(i,k,s,t) = 0$ в областях $\{j \geq 0,\ i = 0,\ k = 1,2\}$ и $\{j \geq 0,\ i > 0,\ k = 0\}$. Вывод выражений для остальных матриц $\tilde{K}_j^*(i,k,s,t)$ основан на использовании их вероятностного смысла, вероятностного смысла ПЛС (см., например, [86]) и внимательном анализе различных сценариев поведения рассматриваемой СМО на интервале $[0,t]$, который начинается в момент окончания обслуживания некоторого запроса, а заканчивается раньше момента окончания обслуживания следующего запроса.

Пусть, например, $k = 1, i > 0$. Тогда матрицы $\tilde{K}_j^*(i,1,s,t)$ имеют вид:

$$\tilde{K}_0^*(i,1,s,t) = \bar{Q} \int\limits_0^t [I_{N+1} \otimes e^{D_0 x} \sum_{k=1}^{i} D_k dx P(i-k, t-x)] \times$$

$$\times \int\limits_0^\infty e^{-su} e^{\Delta(t+u)} \otimes I_{\bar{W}} dB(t-x+u) +$$

$$+(1-p)\{\int\limits_0^t [dF(x) \otimes \sum_{k=1}^{i} P(k,x) P(i-k, t-x)] \tilde{Q}_3 \times$$

$$\times \int\limits_0^\infty e^{-su} e^{\Delta(t-x+u)} \otimes I_{\bar{W}} dB(t-x+u) +$$

$$+ \int\limits_0^t \int\limits_0^y [dF(x) \otimes (e^{D_0 y} \sum_{k=1}^{i} D_k dy P(i-k, t-y))] \tilde{Q}_3 \times$$

$$\times \int\limits_0^\infty e^{-su} e^{\Delta(t-x+u)} \otimes I_{\bar{W}} dB(t-y+u)\},\ i > 0,$$

$$\tilde{K}_j^*(i,1,s,t) = \bar{Q}[I_{N+1} \otimes P(i-j,t)] \int\limits_0^\infty e^{-su} e^{\Delta(t+u)} \otimes I_{\bar{W}} dB(t+u) +$$

$$+(1-p) \int\limits_0^t [dF(x) \otimes \sum_{k=0}^{i-j} P(k,x) P(i-k-j, t-x)] \tilde{Q}_3 \times$$

$$\times \int_0^\infty e^{-su} e^{\Delta(t-x+u)} \otimes I_{\bar{W}} dB(t-x+u),\ j>0,\ i>0.$$

Подставляя эти выражения в формулу (7.25), после алгебраических преобразований, включающих изменение порядка интегрирования, получим формулу (7.23) для векторов $\tilde{\boldsymbol{v}}(i,1,s)$, $i>0$.

Аналогично выводятся формулы (7.22) и (7.24).

□

Далее найдем ПЛС распределения *обобщенного времени обслуживания* на первой фазе. Пусть $\hat{B}(x) = (\hat{B}(x)_{r,r'})_{r,r'=\overline{0,N}}$, где $\hat{B}(x)_{r,r'} = P\{t_{n+1} - t_n < x, r_{n+1} = r' \mid r_n = r, i_n \neq 0\}$, и $\mathcal{B}(s) = \int_0^\infty e^{-st} d\hat{B}(t)$. В дальнейшем для краткости будем называть матрицы $\hat{B}(x)$ и $\mathcal{B}(s)$ матричной ФР и матричным ПЛС распределения *обобщенного времени обслуживания*. Аналогичные термины будем применять для других матриц (векторов), состоящих из ФР и ПЛС некоторого распределения.

Лемма 7.3. *Матричное ПЛС распределения обобщенного времени обслуживания на первой фазе имеет вид:*

$$\mathcal{B}(s) = [Q_1 + pQ_2 + (1-p)F^*(s)Q_3]\mathcal{B}^*(s), \tag{7.26}$$

где

$$F^*(s) = \int_0^\infty e^{-st} dF(t).$$

Доказательство. Проанализируем структуру обобщенного времени обслуживания, то есть интервала времени между двумя соседними моментами окончания обслуживания на первой фазе, принадлежащими периоду занятости.

Обобщенное время обслуживания помеченного запроса равно времени обслуживания запроса на первой фазе, если предыдущий запрос не вызвал блокировку этой фазы. В этом случае матричное ПЛС распределения обобщенного времени пребывания имеет вид $(Q_1 + pQ_2)\mathcal{B}^*(s)$. Если же происходит блокировка, обобщенное время обслуживания состоит из времени, в течение которого прибор блокирован предыдущим запросом, и времени обслуживания помеченного запроса. Тогда соответствующее ПЛС имеет вид $(1-p)\int_0^\infty e^{-st} dF(t) Q_3 \mathcal{B}^*(s)$. □

Теперь можем получить уравнение для векторного ПЛС $\boldsymbol{v}_1(s)$ распределения виртуального времени пребывания на первой фазе. Пусть $v_1(r,\nu,x)$ – вероятность того, что в произвольный момент времени $BMAP$ находится в состоянии ν, виртуальное время пребывания на первой фазе меньше, чем x, и число занятых приборов на второй фазе перед моментом окончания виртуального времени пребывания равно r. Тогда $\boldsymbol{v}_1(s)$ определяется как вектор преобразований Лапласа – Стилтьеса $v_1(r,\nu,s) = \int\limits_0^\infty e^{-sx}dv_1(r,\nu,x)$, записанных в лексикографическом порядке.

Теорема 7.4. *Векторное ПЛС $\boldsymbol{v}_1(s)$ стационарного распределения виртуального времени пребывания на первой фазе удовлетворяет уравнению*

$$\boldsymbol{v}_1(s)A(s) = \boldsymbol{\pi}_0\Phi(s), \tag{7.27}$$

где

$$A(s) = sI + \sum_{r=0}^{\infty}\mathcal{B}^r(s)\otimes D_r, \tag{7.28}$$

$$\Phi(s) = \tau^{-1}\hat{Q}(\Delta\oplus D_0)^{-1}(\Delta\otimes I_{\bar{W}} - sI)[\mathcal{B}^*(s)\otimes I_{\bar{W}}].$$

Доказательство. Как было отмечено выше, виртуальное время пребывания на первой фазе в момент времени t состоит из времени дообслуживания текущего запроса, обобщенного времени обслуживания запросов, которые ожидали обслуживания в момент t, и обобщенного времени обслуживания виртуального запроса. Отметим, что:

– время дообслуживания и обобщенные времена обслуживания запросов являются независимыми случайными величинами при условии, что состояния ЦМ ξ_n, $n \geq 1$, фиксированы;

– обобщенные времена обслуживания являются одинаково распределенными при фиксированном числе занятых приборов на концах соответствующих временных интервалов;

– после произвольного времени t входной поток не влияет на виртуальное время пребывания, начавшееся в момент t.

Принимая во внимание эти пояснения и используя формулу полной вероятности, получим следующее выражение для векторного ПЛС $\boldsymbol{v}_1(s)$:

$$\boldsymbol{v}_1(s)=\tilde{\boldsymbol{v}}(0,0,s)+\sum_{i=1}^{\infty}\tilde{\boldsymbol{v}}(i,1,s)[\mathcal{B}^i(s)\otimes I_{\bar{W}}]+\sum_{i=1}^{\infty}\tilde{\boldsymbol{v}}(i,2,s)[\mathcal{B}^{i-1}(s)\otimes I_{\bar{W}}]. \tag{7.29}$$

Умножим (7.29) на матрицу $sI+\sum\limits_{r=0}^{\infty}\mathcal{B}^r(s)\otimes D_r$ и, после ряда трудоемких алгебраических преобразований, получим соотношение

$$\boldsymbol{v}_1(s)(sI+\sum_{r=0}^{\infty}\mathcal{B}^r(s)\otimes D_r)=\tau^{-1}\{\boldsymbol{\pi}_0\sum_{i=0}^{\infty}[V_i+\sum_{j=1}^{i+1}\boldsymbol{\pi}_jY_{i-j+1}][\mathcal{B}^{i+1}(s)\otimes I_{\bar{W}}]+$$

$$+\boldsymbol{\pi}_0\hat{Q}[-(\Delta\oplus D_0)^{-1}(sI+\tilde{D}_0)+I][\mathcal{B}^*(s)\otimes I_{\bar{W}}]-\sum_{j=0}^{\infty}\boldsymbol{\pi}_j[\mathcal{B}^{j+1}(s)\otimes I_{\bar{W}}]\}. \quad (7.30)$$

Для дальнейших преобразований (7.30) используем уравнение баланса (3.69). Умножая i-е уравнение (3.69) на $\mathcal{B}^{i+1}(s)\otimes I_{\bar{W}}$ и суммируя по i, получим

$$\sum_{i=0}^{\infty}\boldsymbol{\pi}_i[\mathcal{B}^{i+1}(s)\otimes I_{\bar{W}}]=\boldsymbol{\pi}_0\sum_{i=0}^{\infty}V_i[\mathcal{B}^{i+1}(s)\otimes I_{\bar{W}}]+\sum_{i=0}^{\infty}\sum_{j=1}^{i+1}\boldsymbol{\pi}_jY_{i-j+1}[\mathcal{B}^{i+1}(s)\otimes I_{\bar{W}}]. \quad (7.31)$$

Используя (7.31), чтобы упростить (7.30), получим (7.27). □

Таким образом, мы получили уравнение (7.27) для ПЛС распределения времени пребывания на первой фазе. Найдем теперь ПЛС распределения времени пребывания на второй фазе.

Пусть $\boldsymbol{v}_2(s)$ – вектор-столбец ПЛС условных распределений времени пребывания на второй фазе. r-й элемент этого вектора является ПЛС распределения времени пребывания запроса на второй фазе при условии, что число занятых приборов на второй фазе равно r перед моментом окончания времени пребывания этого запроса на первой фазе.

Теорема 7.5. *Векторное ПЛС стационарного распределения времени пребывания на второй фазе имеет вид:*

$$\boldsymbol{v}_2(s)=[Q_4(F^*(s)+I)\hat{I}+pQ_2+$$

$$+(1-p)F^*(s)\mathrm{diag}\{f_{r,0}(s),r=N,N-1,\ldots,0\}Q_3]\mathbf{e}, \quad (7.32)$$

где

$$Q_4=\begin{pmatrix} q_0 & q_1 & \ldots & q_{N-1} & q_N \\ q_0 & q_1 & \ldots & q_{N-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ q_0 & 0 & \ldots & 0 & 0 \end{pmatrix}.$$

Доказательство. Время пребывания на второй фазе запроса, для обслуживания которого требуется m приборов этой фазы, состоит из:

а) времени обслуживания самого запроса, когда по крайней мере m приборов второй фазы доступно сразу после обслуживания на первой фазе;

б) нулевого времени, если требуемое число приборов недоступно и запрос покидает систему;

в) времени блокировки и времени обслуживания запроса, если нужное число приборов недоступно и этот запрос ждет их освобождения.

Предполагается, что обслуживание запроса m приборами выполняется независимо друг от друга и заканчивается, когда все m приборов закончат обслуживание. Распределение этого времени обслуживания определяется ПЛС $f_{m,0}(s),\ m = \overline{1,N}$.

Принимая во внимание последнее предположение и пункты (а)-(в), получим выражение (7.32) для вектора $\boldsymbol{v}_2(s)$. В этом выражении первое слагаемое относится к случаю (а), второе и третье слагаемые задают ПЛС в случаях (б) и (в) соответственно. □

Теорема 7.6. *ПЛС стационарного распределения виртуального времени пребывания в системе имеет вид:*

$$v(s) = \boldsymbol{v}_1(s)(I_{N+1} \otimes \mathbf{e}_{\bar{W}})\boldsymbol{v}_2(s). \tag{7.33}$$

Доказательство. Формула (7.33) следует из структуры виртуального времени пребывания в системе, которое состоит из виртуального времени пребывания на первой фазе и времени пребывания на второй. □

Реальное время пребывания. Пусть $v_1^{(a)}(s)$ и $v^{(a)}(s)$ – ПЛС распределения реального времени пребывания на первой фазе и в системе в целом соответственно.

Теорема 7.7. *ПЛС стационарного распределения реальноого времени пребывания на первой фазе вычисляется следующим образом:*

$$v_1^{(a)}(s) = \lambda^{-1}\boldsymbol{v}_1(s)\sum_{k=0}^{\infty}[\mathcal{B}^k(s)(\mathcal{B}(s) - I)^{-1} \otimes D_k]\mathbf{e}. \tag{7.34}$$

Доказательство. Реальное время пребывания на первой фазе произвольного помеченного запроса, который поступил в группе размера k и располагался на j-й позиции внутри группы, состоит из:

а) реального времени пребывания на первой фазе первого запроса в группе, которое совпадает с виртуальным временем пребывания на первой фазе;

б) обобщенных времен обслуживания на первой фазе $j-2$ запросов в группе, которые поступили вместе с помеченным запросом;

в) обобщенного времени обслуживания помеченного запроса на первой фазе.

Векторное ПЛС распределения времени пребывания на первой фазе первого запроса из группы размера k, которая содержит помеченный запрос, определяется вектором $\boldsymbol{v}_1(s)(I_{N+1} \otimes \frac{kD_k\mathbf{e}}{\lambda})$. Полагая, что произвольный помеченный запрос, поступивший в группе размера k, располагается на j-й позиции с вероятностью $1/k$, и используя формулу полной вероятности, получим:

$$v_1^{(a)}(s) = \sum_{k=1}^{\infty} \boldsymbol{v}_1(s)(I_{N+1} \otimes \frac{kD_k\mathbf{e}}{\lambda}) \sum_{j=1}^{k} \frac{1}{k} \mathcal{B}^{j-1}(s)\mathbf{e},$$

или

$$v_1^{(a)}(s) = \lambda^{-1}\boldsymbol{v}_1(s) \sum_{k=1}^{\infty} (I_{N+1} \otimes D_k\mathbf{e}) \sum_{j=1}^{k} \mathcal{B}^{j-1}(s)\mathbf{e}. \qquad (7.35)$$

Путем алгебраических преобразований (7.35) сводится к виду (7.34). □

Теорема 7.8. *ПЛС распределения реального времени пребывания в системе в целом вычисляется следующим образом:*

$$v^{(a)}(s) = \lambda^{-1}\boldsymbol{v}_1(s) \sum_{k=0}^{\infty} [\mathcal{B}^k(s)(\mathcal{B}(s) - I)^{-1} \otimes D_k](I_{N+1} \otimes \mathbf{e}_{\bar{W}})\boldsymbol{v}_2(s).$$

7.2.6 Моменты времени пребывания

Обозначим через $\boldsymbol{v}_1^{(m)}(s)$ m-ю производную вектора $\boldsymbol{v}_1(s)$, $m \geq 1$. Пусть также $\boldsymbol{v}_1^{(0)}(s) = \boldsymbol{v}_1(s)$.

Теорема 7.9. *Пусть* $\int\limits_0^{\infty} t^m dB(t) < \infty$, $m = \overline{1, M+1}$. *Тогда векторы* $\boldsymbol{v}_1^{(m)}(0)$, $m = \overline{1, M}$, *вычисляются рекуррентно*

$$\boldsymbol{v}_1^{(m)}(0) = \left[\left(\boldsymbol{\pi}_0 \Phi^{(m)}(0) - \sum_{l=0}^{m-1} C_m^l \boldsymbol{v}_1^{(l)}(0) A^{(m-l)}(0) \right) \tilde{I} + \right. \qquad (7.36)$$

$$+\frac{1}{m+1}\left(\boldsymbol{\pi}_0\Phi^{(m+1)}(0)-\sum_{l=0}^{m-1}C_{m+1}^{l}\boldsymbol{v}_1^{(l)}(0)A^{(m+1-l)}(0)\right)\mathbf{e}\hat{\mathbf{e}}\Big]\tilde{A}^{-1},$$

с начальным условием

$$\boldsymbol{v}_1^{(0)}(0)=\boldsymbol{v}_1(0)=[\tau^{-1}\boldsymbol{\pi}_0\hat{Q}(\Delta\oplus D_0)^{-1}(\Delta\mathcal{B}^*(0)\otimes I_{\bar{W}})\tilde{I}+P_{idle}\hat{\mathbf{e}}]\tilde{A}^{-1}, \quad (7.37)$$

где

$$\tilde{A}=A(0)\tilde{I}+A'(0)\mathbf{e}\hat{\mathbf{e}}.$$

Доказательство. Дифференцируя уравнение (7.27), получим

$$\boldsymbol{v}_1^{(m)}(0)A(0)=\boldsymbol{\pi}_0\Phi^{(m)}(0)-\sum_{l=0}^{m-1}C_m^l\boldsymbol{v}_1^{(l)}(0)A^{(m-l)}(0),\ m\geq 0. \quad (7.38)$$

Из (7.28) следует, что $A(0)=\sum\limits_{r=0}^{\infty}\mathcal{B}^r(0)\otimes D_r$, где $\mathcal{B}(0)$ – неприводимая стохастическая матрица. Тогда, как нетрудно видеть, $A(0)$ является неприводимым инфинитезимальным генератором, из чего следует, что $A(0)$ – вырожденная матрица. Вследствии этого невозможно получить рекуррентную процедуру для вычисления векторов $\boldsymbol{v}_1^{(m)}(0)$, $m\geq 0$, непосредственно из (7.38). Преобразуем систему (7.38), чтобы получить систему с невырожденной матрицей. Для этого умножим обе части выражения (7.38) для $m+1$ справа на $\mathbf{e}$. Принимая во внимание то, что $A(0)\mathbf{e}=\mathbf{0}^T$, получим

$$\boldsymbol{v}_1^{(m)}A'(0)\mathbf{e}=\frac{1}{m+1}[\boldsymbol{\pi}_0\Phi^{(m+1)}(0)-\sum_{l=0}^{m-1}C_{m+1}^{l}\boldsymbol{v}_1^{(l)}(0)A^{(m+1-l)}(0)]\mathbf{e}. \quad (7.39)$$

Можно показать, что правая часть выражения (7.39) не равна нулю. Она положительна, если $m=2k$, и отрицательна, если $m=2k+1$, $k\geq 0$. Заменив одно из уравнений системы (7.38) (без потери общности первое) уравнением (7.39), получим следующую неоднородную СЛАУ для вектора $\boldsymbol{v}_1^{(m)}(0)$:

$$\boldsymbol{v}_1^{(m)}(0)\tilde{A}=[\boldsymbol{\pi}_0\Phi^{(m)}(0)-\sum_{l=0}^{m-1}C_m^l\boldsymbol{v}_1^{(l)}(0)A^{(m-l)}(0)]\tilde{I}+$$

$$+\frac{1}{m+1}[\boldsymbol{\pi}_0\Phi^{(m+1)}(0)-\sum_{l=0}^{m-1}C_{m+1}^{l}\boldsymbol{v}_1^{(l)}(0)A^{(m+1-l)}(0)]\mathbf{e}\hat{\mathbf{e}},\ m\geq 0.$$

Эта система имеет единственное решение, если матрица $\tilde{A}$ является невырожденной. Докажем это, показав, что $\det\tilde{A}\neq 0$.

Вычислим $\det\tilde{A}$ следующим образом

$$\det\tilde{A} = \boldsymbol{\nabla} A'(0)\mathbf{e}, \tag{7.40}$$

где $\boldsymbol{\nabla}$ – вектор алгебраических дополнений первого столбца матрицы $A(0)$. Так как $A(0)$ – неприводимый инфинитезимальный генератор, то вектор $\boldsymbol{\nabla}$ пропорционален решению системы

$$\mathbf{x}A(0) = \mathbf{0}, \tag{7.41}$$

то есть

$$\boldsymbol{\nabla} = c\mathbf{x},\ c \neq 0. \tag{7.42}$$

Пусть вектор $\mathbf{y}$ является единственным решением СЛАУ

$$\mathbf{y}\mathcal{B}(0) = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1.$$

Тогда вектор $\mathbf{x} = \mathbf{y} \otimes \boldsymbol{\theta}$ является решением системы (7.41). Этот факт доказывается прямой подстановкой этого вектора в (7.41).

Из (7.42) следует, что

$$\boldsymbol{\nabla} = c(\mathbf{y} \otimes \boldsymbol{\theta}). \tag{7.43}$$

Подставляя выражение (7.43) для $\boldsymbol{\nabla}$ в (7.40), получим

$$\det\tilde{A} = c(\mathbf{y} \otimes \boldsymbol{\theta})A'(0)\mathbf{e} = c(\mathbf{y} \otimes \boldsymbol{\theta})(sI + \sum_{r=0}^{\infty} \mathcal{B}^r(s) \otimes D_r)'|_{s=0}\mathbf{e} =$$

$$= c + c\sum_{r=1}^{\infty} r\mathbf{y}\mathcal{B}'(0)\mathbf{e} \otimes \boldsymbol{\theta}D_r\mathbf{e} = c(1 + \lambda\mathbf{y}\mathcal{B}'(0)\mathbf{e}). \tag{7.44}$$

При дальнейшем анализе $\det\tilde{A}$ используем условие эргодичности, полученное в теореме 7.1.

Полагая в (7.26) $s = 0$ и учитывая, что, по обозначению, $Q_1 + pQ_2 + (1-p)F^*(0)Q_3 = \mathcal{Q}$, получим $\mathcal{B}(0) = \mathcal{Q}\mathcal{B}^*(0)$, то есть $\mathcal{B}(0)$ совпадает с матрицей системы (7.8). Поэтому вектор $\mathbf{y}$ является единственным решением системы (7.8). Легко показать, что

$$\mathbf{y}\mathcal{B}'(0)\mathbf{e} = -[b_1 + (1-p)\mathbf{y}\int_0^{\infty} tdF(t)Q_3\mathbf{e}]. \tag{7.45}$$

Умножив (7.45) на λ и сравнив полученное выражение с выражением для ρ в формулировке теоремы 7.1, получим

$$\lambda \mathbf{y}\mathcal{B}'(0)\mathbf{e} = -\rho. \tag{7.46}$$

Из (7.44) и (7.46) следует, что $\det\tilde{A} = c(1-\rho)$. Так как условие эргодичности $\rho < 1$ выполнено и $c \neq 0$, то $\det\tilde{A} \neq 0$. □

В дальнейшем полагаем, что $\int\limits_0^\infty t^k dB(t) < \infty$, $k = \overline{1,2}$.

Следствие 7.5. *Среднее значение виртуального времени пребывания на первой фазе вычисляется по формуле*

$$\overline{v}_1 = \{[\tau^{-1}\boldsymbol{\pi}_0\hat{Q}(\Delta \oplus D_0)^{-1}((\mathcal{B}^*(0) - \Delta\mathcal{B}^{*\prime}(0)) \otimes I_{\bar{W}}) + \boldsymbol{v}_1(0)A'(0)]\tilde{I} +$$

$$+[P_{idle}b_1 + \frac{1}{2}\boldsymbol{v}_1(0)A''(0)\mathbf{e}]\hat{\mathbf{e}}\}\tilde{A}^{-1}\mathbf{e}, \tag{7.47}$$

где вектор $\boldsymbol{v}_1(0)$ определяется формулой (7.37).

Доказательство. Доказательство следует из соотношения $\overline{v}_1 = -\boldsymbol{v}_1'(0)\mathbf{e}$ и формулы (7.36). □

Теорема 7.10. *Среднее значение виртуального времени пребывания в системе вычисляется по формуле*

$$\overline{v} = \overline{v}_1 + \boldsymbol{v}_1(0)(I_{N+1} \otimes \mathbf{e}_{\bar{W}})\overline{\boldsymbol{v}}_2.$$

Здесь $\boldsymbol{v}_1(0)$ и $\overline{v}_1$ определены в (7.37) и (7.47) соответственно, а $\overline{\boldsymbol{v}}_2$ – вектор условных средних времен пребывания на второй фазе, который вычисляется следующим образом:

$$\overline{\boldsymbol{v}}_2 = -Q_4 \int\limits_0^\infty t dF(t)\hat{\mathbf{e}}^T - (1-p)[\int\limits_0^\infty t dF(t) +$$

$$+E\mathrm{diag}\,\{\sum_{l=1}^{r}(l\mu)^{-1}, r = N, N-1, \ldots, 0\}]Q_3\mathbf{e}. \tag{7.48}$$

Доказательство. Продифференцируем (7.33). Подставив $s = 0$ и поменяв знак, получим

$$\overline{v} = -\boldsymbol{v}_1'(0)(I_{N+1} \otimes \mathbf{e}_{\bar{W}})\boldsymbol{v}_2(0) - \boldsymbol{v}_1(0)(I_{N+1} \otimes \mathbf{e}_{\bar{W}})\boldsymbol{v}_2'(0). \tag{7.49}$$

Подставив $s = 0$ в (7.32), получим

$$\boldsymbol{v}_2(0) = [Q_4(E + I)\hat{I} + pQ_2 + (1 - p)EQ_3]\mathbf{e} = \mathbf{e}.$$

Это означает, что первое слагаемое в правой части (7.49) равно $-\boldsymbol{v}_1'(0)\mathbf{e} = \overline{v}_1$. Используя соотношение $\overline{\boldsymbol{v}}_2 = -\boldsymbol{v}_2'(0)\mathbf{e}$ и дифференцируя (7.32) в точке $s = 0$, получим, что $\overline{\boldsymbol{v}}_2$ имеет вид (7.48). □

Теорема 7.11. *Среднее значение реального времени пребывания на первой фазе вычисляется по формуле*

$$\overline{v}_1^{(a)} = -\lambda^{-1}\{\boldsymbol{v}_1'(0)(\mathbf{e} \otimes \sum_{k=1}^{\infty} kD_k\mathbf{e}) +$$

$$+\boldsymbol{v}_1(0)\sum_{k=1}^{\infty}(I_{N+1} \otimes D_k\mathbf{e})[\sum_{n=1}^{k-1}\sum_{l=0}^{n-1}\mathcal{B}^l(0)\mathcal{B}'(0)\mathbf{e}]\}.$$

Доказательство. Доказательство проводится на основе соотношения $\overline{v}_1^{(a)} = -\frac{dv_1^{(a)}(s)}{ds}|_{s=0}$ и формулы (7.34). □

Следствие 7.6. *Среднее значение реального времени пребывания в системе вычисляется по формуле*

$$\overline{v}^{(a)} = \overline{v}_1^{(a)} + \lambda^{-1}\boldsymbol{v}_1(0)\sum_{k=1}^{\infty}(I_{N+1} \otimes D_k\mathbf{e})\sum_{n=0}^{k-1}\mathcal{B}^n(0)\overline{\boldsymbol{v}}_2.$$

7.2.7 Численные примеры

Эксперимент 7.1. В этом эксперименте исследуется влияние коэффициента корреляции во входном потоке на основные характеристики производительности системы.

Ниже представлены четыре MAP-потока, которые определяются матрицами D_0 и $D_1 = D$. Все эти MAP-потоки имеют одинаковую интенсивность $\lambda = 5$ и различные коэффициенты корреляции.

MAP_1 – это стационарный пуассоновский поток, который определяется матрицами $D_0 = -5$ и $D = 5$, имеет коэффициент корреляции $c_{cor} = 0$ и коэффициент вариации $c_{var} = 1$.

MAP_2 имеет коэффициент корреляции $c_{cor} = 0.1$ и характеризуется матрицами

$$D_0 = \begin{pmatrix} -13.3346 & 0.5886 & 0.6173 \\ 0.6927 & -2.4466 & 0.4229 \\ 0.6823 & 0.4144 & -1.6354 \end{pmatrix}, D = \begin{pmatrix} 11.5469 & 0.3631 & 0.2187 \\ 0.3842 & 0.8659 & 0.0809 \\ 0.2852 & 0.0425 & 0.2111 \end{pmatrix}.$$

MAP_3 имеет коэффициент корреляции $c_{cor} = 0.2$ и характеризуется матрицами

$$D_0 = \begin{pmatrix} -15.7327 & 0.6062 & 0.5924 \\ 0.5178 & -2.2897 & 0.4679 \\ 0.5971 & 0.5653 & -1.9597 \end{pmatrix}, D = \begin{pmatrix} 14.1502 & 0.3021 & 0.0818 \\ 0.1071 & 1.032 & 0.1646 \\ 0.0858 & 0.1979 & 0.5136 \end{pmatrix}.$$

MAP_4 имеет коэффициент корреляции $c_{cor} = 0.3$ и характеризуется матрицами

$$D_0 = \begin{pmatrix} -25.5398 & 0.3933 & 0.3612 \\ 0.1452 & -2.2322 & 0.2000 \\ 0.2960 & 0.3874 & -1.7526 \end{pmatrix}, D = \begin{pmatrix} 24.2421 & 0.4669 & 0.0763 \\ 0.0341 & 1.6668 & 0.1861 \\ 0.0090 & 0.2555 & 0.8047 \end{pmatrix}.$$

MAP_2-MAP_4 имеют коэффициент вариации $c_{var} = 2$.

На основе этих MAP-потоков построим $BMAP$-потоки с максимальным размером групп, равным пяти. Каждый из $BMAP$-потоков определяется матрицами $D_k, k = \overline{0,5}$, полученными следующим образом: матрица D_0 совпадает с соответствующей матрицей для MAP-потока, а остальные матрицы определяются следующим образом: $D_k = D\varkappa^{k-1}(1-\varkappa)/(1-\varkappa^5)$, $k = \overline{1,5}$, где $\varkappa = 0.8$. Затем все матрицы D_k, $k = \overline{0,5}$, умножим на положительную константу, чтобы получить $BMAP$ со средней интенсивностью $\lambda = 5$. $BMAP$-поток, полученный из MAP_n, будем в дальнейшем обозначать как $BMAP_n$, $n = \overline{1,4}$. Заметим, что коэффициенты корреляции таким образом построенных $BMAP$-потоков будут совпадать с соответствующими коэффициентами исходных MAP-потоков.

Время обслуживания на первой фазе имеет распределение Эрланга третьего порядка с интенсивностью 20. Среднее время обслуживания $b_1 = 0.15$ и коэффициент вариации в квадрате $c_{var}^2 = 1/3$.

Число приборов на второй фазе $N = 5$, вероятность потери запроса после обслуживания на первой фазе $p = 0.5$. Параметры обслуживания на второй фазе определяются интенсивностью $\mu = 5$ и вероятностями $q_0 = 0.1$, $q_1 = q_2 = 0.3$, $q_3 = q_4 = q_5 = 0.1$.

Рисунки 7.1 и 7.2 показывают зависимость среднего виртуального $\bar{v}$ и среднего реального $\bar{v}_a$ времен пребывания, вероятности потери P_{loss} и среднего числа N_{busy} занятых приборов на второй фазе от коэффициента загрузки системы ρ. Изменение величины ρ происходит за счет изменения средней интенсивности λ, которая, в свою очередь, изменяется путем нормирования элементов матриц D_k, $k = \overline{0,5}$. Отметим, что при этом коэффициенты вариации и корреляции $BMAP$-потоков не меняются.

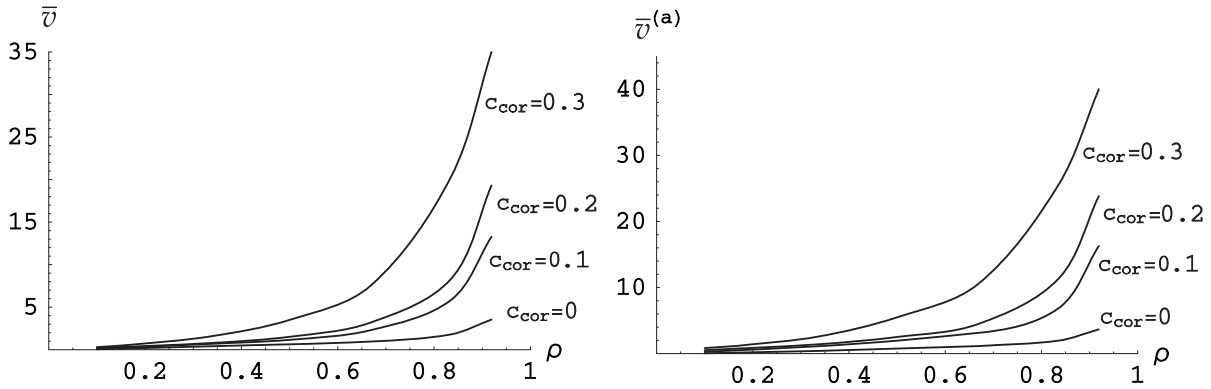

Рисунок 7.1. Зависимость среднего виртуального и реального времен пребывания от загрузки системы для $BMAP$-потоков с различным коэффициентом корреляции

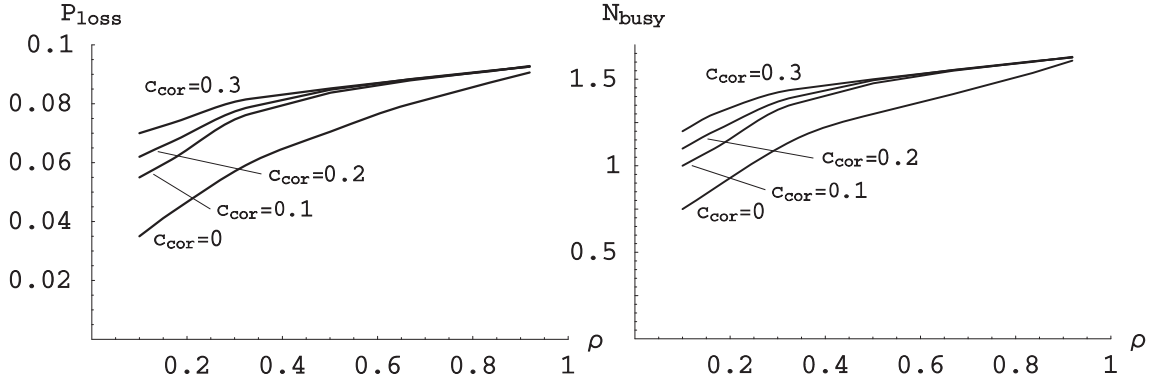

Рисунок 7.2. Зависимость вероятности потери и среднего числа занятых приборов на второй фазе от загрузки системы для $BMAP$-потоков с различным коэффициентом корреляции

Из рисунков 7.1 и 7.2 видно, что при одном и том же значении загрузки системы значения этих характеристик существенно отличаются для $BMAP$-потоков с разной корреляцией. С увеличением коэффициента корреляции все приведенные характеристики возрастают, что свидетельствует об ухудшении качества обслуживания и подтверждает необходимость учета влияние корреляции во входном потоке при оценке производительности рассматриваемой СМО.

Эксперимент 7.2. Исследуем влияние коэффициента вариации процесса обслуживания на первой фазе на основные характеристики производительности.

Рассмотрим четыре распределения времени обслуживания с одинаковым средним временем обслуживания $b_1 = 0.1$, но различными коэффициентами вариации: D – детерминированное распределение; M – экспоненциальное распределение; $HM_2^{(1)}$ и $HM_2^{(2)}$– гиперэкспоненциальные распределения второго порядка с разными параметрами. Гиперэкспоненциальные распределения определяются вероятностями (0.05, 0.95) и интенсивностями 0.62025, 48.9998 в случае $HM_2^{(1)}$ и вероятностями (0.98, 0.02) и интенсивностями 10000, 0.2 – в случае $HM_2^{(2)}$. Коэффициенты вариации процессов D, M, $HM_2^{(1)}$ и $HM_2^{(2)}$ равны 0, 1, 5, 9.95 соответственно.

Входящий поток отличается от $BMAP_2$ из предыдущего эксперимента средней интенсивностью λ. Здесь матрицы $D_k, k = \overline{0,5}$, нормированы так, что $\lambda = 1$. Интенсивность обслуживания μ на второй фазе равна 0.8. Остальные параметры системы полагаем такими же, как и в предыдущем эксперименте.

Будем изменять среднее время обслуживания b_1 для всех процессов в интервале $[0.1, 0.95]$. Отметим, что при этом коэффициент вариации остается неизменным.

Рисунки 7.3 и 7.4 показывают зависимость основных характеристик производительности рассматриваемой СМО от среднего времени обслуживания на первой фазе.

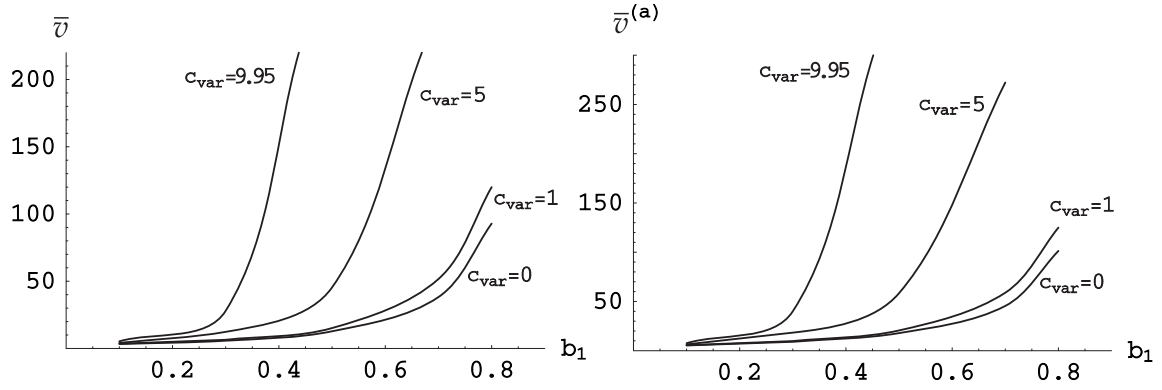

Рисунок 7.3. Зависимость среднего виртуального и реального времен пребывания от среднего времени обслуживания для процессов обслуживания с различными коэффициентами вариации

Из рисунков 7.3 и 7.4 видно, что все характеристики существенно за-

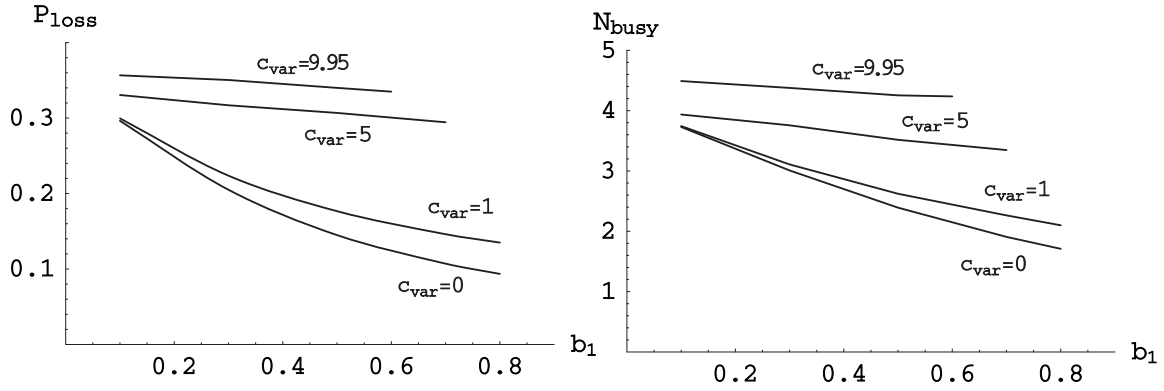

Рисунок 7.4. Зависимость вероятности потери и среднего числа занятых приборов на второй фазе от среднего времени обслуживания для процессов обслуживания с различными коэффициентами вариации

висят от коэффициента вариации процесса обслуживания. Были также проведены эксперименты, в которых рассматривались равномерное распределение и распределение Эрланга. Так как коэффициенты вариации этих распределений находятся в интервале $(0, 1)$, то соответствующие линии, как и ожидалось, располагаются между двумя нижними линиями на рисунках 7.3 и 7.4.

Для процессов $HM_2^{(1)}$ и $HM_2^{(2)}$ при значениях b_1, превышающих 0.7 и 0.6 соответственно, условие эргодичности $\rho < 1$ не выполняется.

Таблица 7.1 содержит значения коэффициента загрузки ρ в этом эксперименте при различных значениях b_1 и c_{var}. Значения ρ, при которых нарушается условие существования стационарного режима, выделены полужирным шрифтом.

В таблице 7.2 представлены значения среднего виртуального и реального времен пребывания в этом эксперименте.

Таблица 7.1. Значения коэффициента загрузки системы при различных значениях среднего времени обслуживания на первой фазе и коэффициента вариации

ρ	$c_{var} = 0$	$c_{var} = 1$	$c_{var} = 5$	$c_{var} = 9.95$
$b_1 = 0.1$	0.42754	0.43016	0.45557	0.47940
$b_1 = 0.2$	0.47460	0.48228	0.53677	0.57784
$b_1 = 0.3$	0.53173	0.54515	0.62243	0.67725
$b_1 = 0.4$	0.59698	0.61590	0.71010	0.77695
$b_1 = 0.5$	0.66879	0.69247	0.79902	0.87678

$b_1 = 0.6$	0.74579	0.77339	0.88883	0.97667
$b_1 = 0.7$	0.82687	0.85763	0.97937	**1.07663**
$b_1 = 0.8$	0.91116	0.94440	**1.07051**	**1.17655**

Таблица 7.2. Значения среднего реального и виртуального времен пребывания при различных значениях среднего времени обслуживания и коэффициента вариации

		$c_{var} = 0$	$c_{var} = 1$	$c_{var} = 5$	$c_{var} = 9.95$
$b_1 = 0.1$	$\bar{v}$	3.31836	3.35900	4.14502	5.31900
	$\bar{v}^{(a)}$	5.34117	5.56531	6.42475	7.61900
$b_1 = 0.3$	$\bar{v}$	5.78467	6.44377	12.36466	28.03929
	$\bar{v}^{(a)}$	9.06977	9.79000	18.36466	40.03929
$b_1 = 0.5$	$\bar{v}$	12.78808	15.18521	45.44273	307.20933
	$\bar{v}^{(a)}$	17.88462	20.60316	58.44273	383.20933
$b_1 = 0.7$	$\bar{v}$	38.23352	51.20479	252.27566	–
	$\bar{v}^{(a)}$	45.47547	58.89069	272.27566	–
$b_1 = 0.8$	$\bar{v}$	92.87134	120.01298	–	–
	$\bar{v}^{(a)}$	101.358218	125.03418	–	–

Проанализировав рисунки 7.2 и 7.4, можно заметить, что величины P_{loss} и N_{busy} возрастают с ростом загрузки системы ρ, но убывают с ростом среднего времени обслуживания b_1. На первый взгляд, это кажется неправдоподобным, потому что увеличение b_1 влечет увеличение загрузки системы, которое обычно негативно сказывается на качестве обслуживания. Поясним полученное. На рисунке 7.2, значение ρ увеличивается за счет увеличения интенсивности λ. Очевидно, что увеличение интенсивности входного потока приводит к росту P_{loss} и N_{busy}. Когда возрастает среднее время обслуживания на первой фазе b_1, то происходит сглаживание входного потока и уменьшается влияние корреляции на вторую фазу, что положительно влияет на величины P_{loss} и N_{busy}.

Эксперимент 7.3. В данном эксперименте исследуем численно задачу выбора оптимального числа приборов на второй фазе. Найдем число приборов N на второй фазе, при котором достигается минимум следующего критерия качества (средний штраф в единицу времени):

$$J = J(N) = aN + c_1 \lambda P_{loss} + c_2 \bar{v}^{(a)}.$$

Здесь a – стоимость содержания одного прибора второй фазы в единицу времени, c_1 – штраф за потерю одного запроса после обслуживания на первой фазе в единицу времени, c_2 – плата за нахождение (время пребывания) одного запроса в системе в единицу времени.

Здесь и далее при решении задач целочисленной оптимизации использовался метод прямого перебора.

Используя MAP, который характеризуется матрицами

$$D_0 = \begin{pmatrix} -6.74538 & 5.45412 \times 10^{-6} \\ 5.45412 \times 10^{-6} & -0.219455 \end{pmatrix}, \; D = \begin{pmatrix} 6.700685 & 0.044695 \\ 0.122427 & 0.097023 \end{pmatrix},$$

построим $BMAP$ с матрицами $D_k, k = \overline{0,5}$, аналогичным способом, как и в первом эксперименте, и нормируем матрицы таким образом, чтобы получить интенсивность $\lambda = 3$. Полученный $BMAP$ имеет коэффициент корреляции $c_{cor} = 0.2$ и $c^2_{var} = 12.2732$.

Время обслуживания на первой фазе имеет распределение Эрланга третьего порядка с интенсивностью 20. Полагаем, что $p = 0.5$, $q_0 = 0.1$, $q_1 = 0.9$, $q_m = 0$, $m = \overline{2,N}$.

Стоимостные коэффициенты возьмем следующими: $a = 5, c_1 = 50, c_2 = 3$.

Зависимость критерия качества J от числа приборов N на второй фазе при различных интенсивностях обслуживания μ представлена на рисунке 7.5.

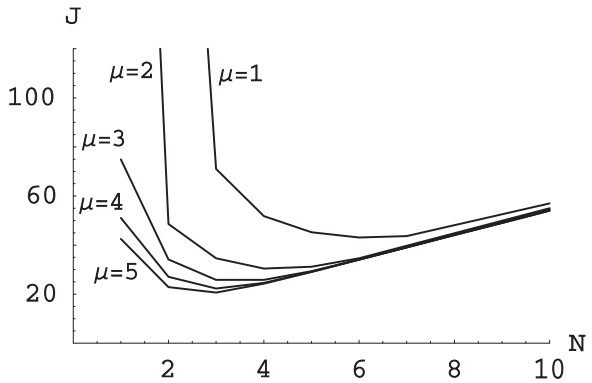

Рисунок 7.5. Зависимость критерия качества от числа приборов на второй фазе при различных интенсивностях обслуживания

Таблица 7.3 содержит значения критерия качества в этом эксперименте. Оптимальные значения J^* критерия качества для каждой интенсивности обслуживания выделены полужирным шрифтом.

Таблица 7.3. Значения критерия качества при различных значениях числа приборов второй фазы и интенсивности обслуживания на второй фазе

J	$\mu = 1$	$\mu = 2$	$\mu = 3$	$\mu = 4$	$\mu = 5$
$N = 1$	∞	457.0346	74.9351	51.0311	42.5146
$N = 2$	353.7378	48.6349	34.1130	27.0826	22.9120
$N = 3$	71.0480	34.6565	25.8441	**22.2656**	**20.6657**
$N = 4$	51.7822	**30.4068**	**25.8202**	24.6008	24.2393
$N = 5$	45.2556	31.2080	29.3925	29.1141	29.0630
$N = 6$	**43.0750**	34.6437	34.0980	34.0535	34.0486
$N = 7$	43.6867	39.0715	39.0533	39.0504	39.0476
$N = 8$	47.2564	44.0527	44.0514	44.0495	44.0474
$N = 9$	51.9854	49.0521	49.0511	49.0491	49.0473
$N = 10$	57.0056	55.0458	54.5001	54.0489	54.0472

Из рисунка 7.5 и таблицы 7.3 видно, что при уменьшении интенсивности обслуживания с 5 до 1, оптимальное число приборов N^* увеличивается с 3 до 6.

Относительный выигрыш при использовании оптимального числа приборов N^* по сравнению с произвольным числом N приборов на второй фазе определяется формулой $g(N) = \frac{J(N)-J^*}{J^*}100\%$.

Более подробно остановимся на случае, когда $\mu = 5$. Как видно из рисунка 7.5 и таблицы 7.3, оптимальное значение критерия качества J^* равно 20.6657, а оптимальное число приборов $N^* = 3$. Следует отметить, что в этом случае минимальный относительный выигрыш составляет более 10%, если установить оптимальное число приборов $N^* = 3$ вместо 2 приборов, а максимальный относительный выигрыш – более 161%, если использовать $N^* = 3$ вместо 10.

7.3 Система $BMAP/SM/1 \to \cdot/M/N/0$ с групповым занятием приборов второй фазы

7.3.1 Математическая модель

Рассматривается двухфазная система $BMAP/SM/1 \to \cdot/M/N/0$ с полумарковским процессом обслуживания. В модели $BMAP/G/1 \to \cdot/M/N/0$, описанной в разделе 7.2, предполагалось, что времена обслуживания запросов на первой фазе являются независимыми одинаково распределенными случайными величинами. Однако во многих реальных системах времена обслуживания запросов могут быть зависимыми и распределенными по разным законам. Как отмечалось выше, для моделирования таких процессов обслуживания в литературе предложен формализм полумарковских (SM) процессов обслуживания. В нашем случае предполагается, что времена обслуживания последовательных запросов задаются последовательными временами пребывания в своих состояниях регулярного эргодического полумарковского процесса m_t, $t \geq 0$, с пространством состояний $\{1, 2, \ldots, M\}$ и полумарковским ядром $B(t) = (B_{m,m'}(t))_{m,m'=\overline{1,M}}$. Функция $B_{m,m'}(t)$ интерпретируется как вероятность того, что время пребывания процесса в текущем состоянии не превысит величину t и следующий переход произойдет в состояние m' при условии, что текущим состоянием процесса является состояние m, $m, m' = \overline{1, M}$. Среднее время обслуживания b_1 вычисляется как

$$b_1 = \boldsymbol{\delta} \int_0^\infty t dB(t) \mathbf{e}, \tag{7.50}$$

где $\boldsymbol{\delta}$ – инвариантный вектор стохастической матрицы $B(\infty)$, то есть вектор, удовлетворяющий уравнению:

$$\boldsymbol{\delta} B(\infty) = \boldsymbol{\delta}, \ \boldsymbol{\delta}\mathbf{e} = 1.$$

7.3.2 Стационарное распределение вложенной цепи Маркова

Исследование системы $BMAP/SM/1 \to \cdot/M/N/0$ начнем с введения вложенного процесса $\xi_n = \{i_n, r_n, \nu_n, m_n\}$, $n \geq 1$, где i_n – число запросов на первой фазе (не включая запрос, вызвавший блокировку) в момент времени $t_n + 0$ (t_n – n-й момент окончания обслуживания на первой фазе);

r_n – число занятых приборов на второй фазе в момент времени $t_n - 0$; ν_n – состояние $BMAP$ в момент времени t_n; m_n – состояние SM-процесса обслуживания в момент $t_n + 0$, $i_n \geq 0, r_n = \overline{0, N}, \nu_n = \overline{0, W}$, $m_n = \overline{1, M}$.

Лемма 7.4. *Процесс $\xi_n, n \geq 1$, является КТЦМ. Матрица вероятностей переходов этой цепи имеет блочную структуру вида (2.2), где*

$$V_i = \sum_{k=1}^{i+1}[-\hat{H}(\Delta \oplus \bar{D}_0)^{-1}\hat{D}_k + (1-p)\bar{\mathcal{F}}_k\tilde{H}_3]\bar{\Omega}_{i-k+1},$$

$$Y_i = \bar{H}\bar{\Omega}_i + (1-p)\sum_{k=0}^{i}\bar{\mathcal{F}}_k\tilde{H}_3\bar{\Omega}_{i-k},\ i \geq 0,$$

$$\bar{\Omega}_n = \int_0^\infty e^{\Delta t} \otimes P(n,t) \otimes dB(t), \qquad \bar{\mathcal{F}}_n = \int_0^\infty dF(t) \otimes P(n,t) \otimes I_M,\ n \geq 0.$$

Здесь

$$\bar{D}_k = D_k \otimes I_M,\ \hat{D}_k = I_{N+1} \otimes \bar{D}_k,\ k \geq 0,$$

$$\tilde{H}_m = Q_m \otimes I_{\bar{W}M},\ m = \overline{1,3}, \bar{H} = \tilde{H}_1 + p\tilde{H}_2,$$

$$\hat{H} = \tilde{H}_1 + p\tilde{H}_2 + (1-p)\int_0^\infty (dF(t) \otimes e^{D_0 t} \otimes I_M)\tilde{H}_3,$$

а матрицы Δ, Q_m, $m = \overline{1,3}$, совпадают с введенными ранее в разделе 7.2.

Доказательство проводится аналогично доказательству леммы 7.1.

Следствие 7.7. *Матричные ПФ $V(z)$, $Y(z)$ имеют вид*

$$V(z) = \frac{1}{z}[-\hat{H}(\Delta \oplus \bar{D}_0)^{-1}(\hat{D}(z) - \hat{D}_0) + (1-p)(\bar{\mathcal{F}}(z) - \bar{\mathcal{F}}_0)\tilde{H}_3]\bar{\Omega}(z),$$

$$Y(z) = [\bar{H} + (1-p)\bar{\mathcal{F}}(z)\tilde{H}_3]\bar{\Omega}(z), \tag{7.51}$$

где

$$\hat{D}(z) = \sum_{k=0}^{\infty}\hat{D}_k z^k,$$

$$\bar{\Omega}(z) = \int_0^\infty e^{\Delta t} \otimes e^{D(z)t} \otimes dB(t),\ \bar{\mathcal{F}}(z) = \int_0^\infty dF(t) \otimes e^{D(z)t} \otimes I_M, |z| \leq 1.$$

Теорема 7.12. *Необходимым и достаточным условием существования стационарного распределения ЦМ ξ_n, $n \geq 1$, является выполнение неравенства*

$$\rho < 1, \tag{7.52}$$

где

$$\rho = \lambda[b_1 + (1-p)\mathbf{y}\int_0^\infty (tdF(t)Q_3 \otimes I_M)\mathbf{e}],$$

а вектор $\mathbf{y}$ *является единственным решением СЛАУ*

$$\mathbf{y}(\mathcal{Q} \otimes I_M)\int_0^\infty e^{\Delta t} \otimes dB(t) = \mathbf{y}, \;\; \mathbf{ye} = 1. \tag{7.53}$$

Здесь матрицы Q_3, $\mathcal{Q}$ введены в разделе 7.2, b_1 – среднее время обслуживания на первой фазе – определяется формулой (7.50).

Доказательство. Матрица $Y(1)$, определенная формулой (7.51), является неприводимой, поэтому, в терминах матрицы $Y(z)$, необходимое и достаточное условия существования стационарного распределения имеет вид (3.66), (3.67).

Пусть в (3.67) вектор $\mathbf{x}$ имеет вид

$$\mathbf{x} = \mathbf{y}_1 \otimes \boldsymbol{\theta} \otimes \mathbf{y}_2, \tag{7.54}$$

где $\mathbf{y}_1$ и $\mathbf{y}_2$ – некоторые векторы с неотрицательными компонентами размера $N+1$ и M, соответсвенно, и такие, что $\mathbf{y} = \mathbf{y}_1 \otimes \mathbf{y}_2$ – единственное решение системы (7.53).

При помощи прямой подстановки можно убедиться, что вектор $\mathbf{x}$ вида (7.54) является единственным решением системы (3.67).

Продифференцируем (7.51) в точке $z = 1$ и подставим полученное выражение для $Y'(1)$ и вектор $\mathbf{x}$ вида (7.54) в неравенство (3.66). Затем, используя соотношение $\mathbf{y}(\mathbf{e}_{N+1} \otimes I_M) = \boldsymbol{\delta}$, путем алгебраических преобразований получим неравенство (7.52). □

Пусть неравенство (7.52) выполняется и пусть $\pi(i, r, \nu, m)$, $i \geq 0$, $r = \overline{0, N}$, $\nu = \overline{0, W}$, $m = \overline{1, M}$, – стационарное распределение ЦМ ξ_n, $n \geq 1$. Обозначим через $\boldsymbol{\pi}_i$ вектор упорядоченных в лексикографическом порядке компонент (r, ν, m) вероятностей $\pi(i, r, \nu, m)$, $i \geq 0$. Для нахождения векторов $\boldsymbol{\pi}_i$, $i \geq 0$, используем один из алгоритмов, приведенных в разделе 3.4.

7.3.3 Стационарное распределение в произвольный момент времени

Рассмотрим процесс $\zeta_t = \{i_t,\, k_t,\, r_t,\, \nu_t,\, m_t\},\ t \geq 0$, состояний системы в произвольный момент времени, где i_t – число запросов на первой фазе (включая запрос, вызвавший блокировку); k_t – величина, принимающая значения 0, 1, 2 в зависимости от того, свободен, обслуживает запрос или блокирован прибор первой фазы; r_t – число занятых приборов на второй фазе; ν_t – состояние $BMAP$-потока; m_t – состояние SM-процесса обслуживания в момент времени t, $t \geq 0$.

Пусть $p(i,k,r,\nu,m), i \geq 0,\ k = 0,1,2,\ r = \overline{0,N},\ \nu = \overline{0,W}, m = \overline{1,M},$ – стационарные вероятности процесса ζ_t, а $\mathbf{P}_i(k)$ – векторы этих вероятностей, упорядоченных в лексикографическом порядке компонент (r,ν,m).

Теорема 7.13. *Ненулевые векторы стационарных вероятностей* $\mathbf{p}_i(k)$, $i \geq 0, k = 0,1,2$, *выражаются через векторы стационарного распределения* $\boldsymbol{\pi}_i$, $i \geq 0$, *вложенной ЦМ* ξ_n, $n \geq 1$, *следующим образом:*

$$\mathbf{p}_0(0) = -\tau^{-1}\boldsymbol{\pi}_0\hat{H}(\Delta \oplus \bar{D}_0)^{-1},$$

$$\mathbf{p}_i(1) = \tau^{-1}\{\boldsymbol{\pi}_0 \sum_{k=1}^{i}[-\hat{H}(\Delta \oplus \bar{D}_0)^{-1}\hat{D}_k + (1-p)\bar{\mathcal{F}}_k\tilde{H}_3]\hat{\Omega}_{i-k} +$$

$$+ \sum_{l=1}^{i}\boldsymbol{\pi}_l[\bar{H}\hat{\Omega}_{i-l} + (1-p)\sum_{k=0}^{i-l}\bar{\mathcal{F}}_k\tilde{H}_3\hat{\Omega}_{i-k-l}]\},$$

$$\boldsymbol{p}_i(2) = \tau^{-1}(1-p)\sum_{l=0}^{i-1}\boldsymbol{\pi}_l \sum_{k=0}^{i-l-1}(\bar{\mathcal{F}}_k + \hat{\delta}_{0,k}I)\tilde{H}_2 \times$$

$$\times \int_0^\infty e^{-\mu\mathcal{N}t} \otimes P(i-k-l-1,t) \otimes I_M dt,\ i \geq 1,$$

где

$$\tau = b_1 + \boldsymbol{\pi}_0\hat{H}(-\hat{D}_0)^{-1}\mathbf{e} + (1-p)\boldsymbol{\Pi}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}M}) \int_0^\infty t dF(t) Q_3\mathbf{e},$$

$$\hat{\Omega}_n = \int_0^\infty e^{\Delta t} \otimes P(n,t) \otimes (I_M - \nabla_B(t))dt, n \geq 0,$$

$$\nabla_B(t) = \mathrm{diag}\{(B(t)\mathbf{e})_j, j = \overline{1,M}\}.$$

Доказательство проводится аналогично доказательству теоремы 7.2.

Следствие 7.8. *Векторы стационарных вероятностей* $\mathbf{p}_i$, $i \geq 0$, *процесса* $\{i_t, r_t, \nu_t, m_t\}$, $t \geq 0$, *вычисляются следующим образом:*

$$\mathbf{p}_i = \sum_{k=0}^{2} \mathbf{p}_i(k), \ i \geq 0.$$

7.3.4 Характеристики производительности системы

Среднее число запросов на первой фазе в момент окончания обслуживания на этой фазе и в произвольный момент времени:

$$L = \mathbf{\Pi}'(1)\mathbf{e}, \quad \tilde{L} = \mathbf{P}'(1)\mathbf{e}.$$

Векторы стационарного распределения числа занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$\boldsymbol{r} = \mathbf{\Pi}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}M}), \qquad \tilde{\boldsymbol{r}} = \mathbf{P}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}M}).$$

Среднее число занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$N_{busy} = \boldsymbol{r}\mathcal{N}\mathbf{e}, \qquad \tilde{N}_{busy} = \tilde{\boldsymbol{r}}\mathcal{N}\mathbf{e}.$$

Вероятность того, что произвольный запрос покинет систему (вызовет блокировку) после обслуживания на первой фазе:

$$P_{loss} = p\mathbf{\Pi}(1)\tilde{H}_2\mathbf{e}, \quad P_{block} = (1-p)\mathbf{\Pi}(1)\tilde{H}_2\mathbf{e}.$$

Вероятности P_{idle}, P_{serve}, P_{block} того, что прибор первой фазы свободен, занят обслуживанием или заблокирован:

$$P_{idle} = \tau^{-1}\boldsymbol{\pi}_0\hat{H}(-\hat{D}_0)^{-1}\mathbf{e}, \ P_{serve} = \tau^{-1}b_1, \ P_{block} = 1 - P_{idle} - P_{serve}.$$

7.3.5 Численные примеры

Исследуем влияние коэффициента корреляции во входном потоке на основные стационарные характеристики производительности системы при различных значениях интенсивности входного потока и интенсивности обслуживания на второй фазе.

Рассмотрим четыре $BMAP$-потока, описанные в первом эксперименте подраздела 7.2.7. Эти потоки имеют одинаковую среднюю интенсивность $\lambda = 5$ и различные коэффициенты корреляции $c_{cor} = 0, 0.1, 0.2, 0.3$.

Полумарковское ядро имеет следующую структуру

$$B(t) = \mathrm{diag}\{B_1(t), B_2(t)\}P,$$

где матрица переходов $P = \begin{pmatrix} 0.6 & 0.4 \\ 0.35 & 0.65 \end{pmatrix}$, $B_1(t)$ и $B_2(t)$ – ФР Эрланга третьего порядка со средними интенсивностями 20 и 50 соответственно. Среднее время обслуживания $b_1 = 0.102$.

Остальные параметры системы полагаем следующими: $N = 5$, $p = 0.5$, $q_0 = 0.1$, $q_1 = 0.3$, $q_2 = 0.3$, $q_3 = q_4 = q_5 = 0.1$ и $\mu = 5$.

Рисунок 7.6 показывает зависимость среднего числа запросов L на первой фазе в моменты окончания обслуживания на этой фазе и вероятности потери P_{loss} от интенсивности λ.

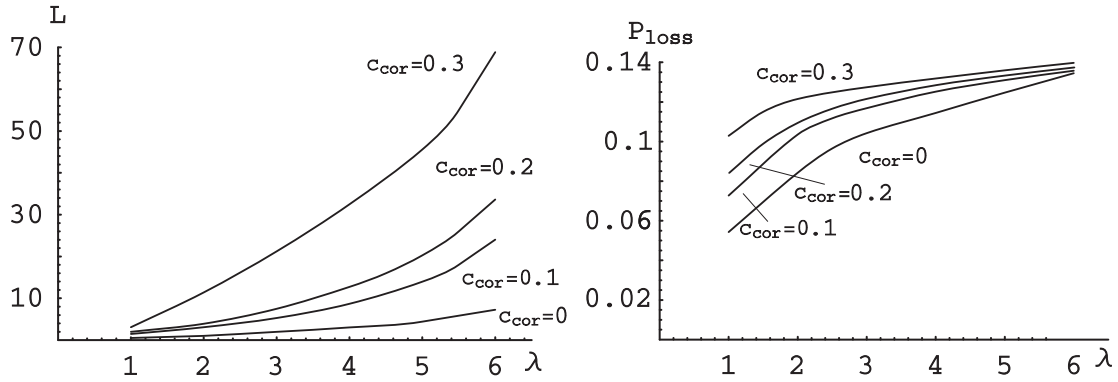

Рисунок 7.6. Зависимость L и P_{loss} от интенсивности λ для $BMAP$-потоков с различным коэффициентом корреляции

Рисунок 7.7 иллюстрирует зависимость среднего числа запросов на первой фазе L и среднего числа занятых приборов на второй фазе N_{busy} от интенсивности обслуживания на второй фазе μ.

На основании приведенных рисунков можно сделать вывод, что коэффициент корреляция является важной характеристикой входного потока, игнорирование которой может привести к существенным ошибкам при оценке производительности рассматриваемой СМО.

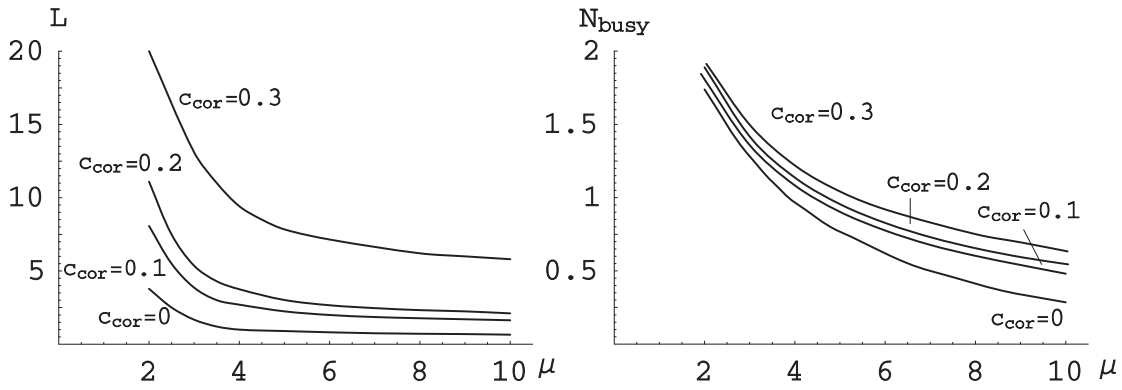

Рисунок 7.7. Зависимость L и N_{busy} от интенсивности обслуживания на второй фазе для $BMAP$-потоков с различным коэффициентом корреляции

7.4 Система $BMAP/G/1 \to \cdot/M/N/R$

7.4.1 Математическая модель

Рассматривается система $BMAP/G/1 \to \cdot/M/N/R$. Данная СМО отличается от системы, исследованной в разделе 7.2, наличием промежуточного буфера размера $R < \infty$. В случае занятости всех приборов второй фазы запрос ожидает обслуживания в буфере. При отсутствии свободного места в буфере с вероятностью p, $0 \le p \le 1$, запрос уходит из системы недообслуженным (теряется), а с вероятностью $1-p$ прибор первой фазы блокируется и не обслуживает следующий запрос до ближайшего окончания обслуживания на второй фазе. Предполагается, что для обслуживания на второй фазе запросу требуется один прибор.

7.4.2 Стационарное распределение вложенной цепи Маркова

Пусть t_n – n-й момент окончания обслуживания на первой фазе, $n \ge 1$. Рассмотрим процесс $\xi_n = \{i_n, l_n, \nu_n\}$, $n \ge 1$, где i_n – число запросов на первой фазе (не включая запрос, вызвавший блокировку) в момент времени t_n+0; l_n – число запросов на второй фазе в момент времени t_n-0; ν_n – состояние $BMAP$ в момент времени t_n, $i_n \ge 0, l_n = \overline{0, N+R}, \nu_n = \overline{0, W}$.

Лемма 7.5. *Процесс $\xi_n, n \ge 1$, является КТЦМ. Матрица вероятностей переходов этой цепи имеет блочную структуру вида (3.60), где*

$$V_i = \sum_{k=1}^{i+1}[-\hat{Q}(\bar{\Delta} \oplus D_0)^{-1}\tilde{D}_k + (1-p)N\mu Q_2 \otimes \int_0^\infty P(k,t)e^{-N\mu t}dt]\Omega_{i-k+1},$$

$$Y_i = \bar{Q}\Omega_i + (1-p)N\mu(Q_2 \otimes \sum_{k=0}^{i} \int_0^{\infty} P(k,t)e^{-N\mu t}dt)\Omega_{i-k},\ i \geq 0.$$

Здесь $\bar{Q} = \tilde{Q}_1 + p\tilde{Q}_2$, $\hat{Q} = \tilde{Q}_1 + p\tilde{Q}_2 + (1-p)N\mu Q_2 \otimes (N\mu I - D_0)^{-1}$, *где* $\tilde{Q}_m = Q_m \otimes I_{\bar{W}}$, $m = 1,2$, Q_1 – *квадратная матрица порядка* $N+R+1$, *у которой элементы* $(Q_1)_{j,j+1}$ *равны единице, а остальные – нулю,* Q_2 – *квадратная матрица порядка* $N+R+1$, *у которой все элементы нулевые, кроме последнего диагонального элемента, равного единице.* $\bar{\Delta}$ – *квадратная блочно-двухдиагональная матрица порядка* $N+R+1$, *диагональные блоки которой определяются следующим образом:* $-\min\{i,N\}\mu, i = \overline{0,N+R}$, *а поддиагональные блоки как* $\min\{i,N\}\mu, i = \overline{1,N+R}$, $\tilde{D}_k = I_{N+R+1} \otimes D_k, k \geq 0$.

Доказательство проводится аналогично доказательству леммы 7.1.

Следствие 7.9. *Матричные ПФ* $V(z)$, $Y(z)$ *имеют вид*

$$V(z) = \frac{1}{z}\{-\hat{Q}(\bar{\Delta} \oplus D_0)^{-1}(\tilde{D}(z) - \tilde{D}_0) +$$
$$+(1-p)N\mu Q_2 \otimes [(N\mu I - D(z))^{-1} - (N\mu I - D_0)^{-1}]\}\Omega(z),$$
$$Y(z) = [\bar{Q} + (1-p)Q_2 N\mu \otimes (N\mu I - D(z))^{-1}]\Omega(z). \qquad (7.55)$$

Теорема 7.14. *Необходимым и достаточным условием существования стационарного распределения ЦМ* ξ_n, $n \geq 1$, *является выполнение неравенства*

$$\rho < 1, \qquad (7.56)$$

где

$$\rho = \lambda[b_1 + (1-p)y_{N+R}(N\mu)^{-1}].$$

Здесь y_{N+R} – *последняя компонента вектора* $\mathbf{y} = (y_0, y_1, ..., y_{N+R})$, *который является единственным решением СЛАУ*

$$\mathbf{y}(Q_1 + Q_2) \int_0^{\infty} e^{\bar{\Delta}t}dB(t) = \mathbf{y}, \ \ \mathbf{ye} = 1. \qquad (7.57)$$

Доказательство. Так как матрица $Y(1)$, определенная формулой (7.55), является неприводимой, то необходимым и достаточным условием эргодичности ЦМ ξ_n, $n \geq 1$, является выполнение неравенства (3.66), где вектор $\mathbf{x}$ является единственным решением СЛАУ (3.67).

Пусть вектор $\mathbf{x}$ имеет вид

$$\mathbf{x} = \mathbf{y} \otimes \boldsymbol{\theta}. \qquad (7.58)$$

Принимая во внимание соотношение $\boldsymbol{\theta} e^{D(1)t} = \boldsymbol{\theta}$, прямой подстановкой можно убедиться, что вектор $\mathbf{x}$ вида (7.58) является решением системы (3.67). Подставляя $\mathbf{x}$ вида (7.58) в (3.66), путем алгебраических преобразований получим неравенство (7.56). □

Далее будем считать, что неравенство (7.56) выполняется. Пусть $\pi(i,l,\nu),\ i \geq 0,\ l = \overline{0, N+R},\ \nu = \overline{0, W}$, – стационарное распределение ЦМ ξ_n, $n \geq 1$. Вектор-строки стационарных вероятностей $\boldsymbol{\pi}_i,\ i \geq 0$, вычисляются с помощью алгоритма, описанного в разделе 3.4.

7.4.3 Стационарное распределение в произвольный момент времени

Рассмотрим процесс $\zeta_t = \{i_t,\ k_t, l_t, \nu_t\},\ t \geq 0$, состояний системы в произвольный момент времени, где i_t – число запросов на первой фазе, k_t– величина, принимающая значения 0, 1, 2 в зависимости от того, свободен, обслуживает запрос или блокирован прибор первой фазы, l_t – число запросов на второй фазе, ν_t – состояние $BMAP$ в момент времени $t,\ t \geq 0$.

Пусть $p(i,k,l,\nu) = \lim\limits_{t\to\infty} P\{i_t = i,\ k_t = k,\ l_t = l,\ \nu_t = \nu\},\ i \geq 0,\ k = 0,1,2,\ l = \overline{0, N+R},\ \nu = \overline{0, W}$, – стационарное распределение процесса $\zeta_t,\ t \geq 0$, а $\mathbf{p}_i(k),\ i \geq 0,\ k = 0,1,2$, – векторы этих вероятностей.

Теорема 7.15. *Ненулевые векторы стационарных вероятностей* $\mathbf{p}_i(k),\ i \geq 0,\ k = 0,1,2$, *выражаются через векторы стационарного распределения* $\boldsymbol{\pi}_i,\ i \geq 0$, *вложенной ЦМ* $\xi_n,\ n \geq 1$, *следующим образом:*

$$\mathbf{p}_0(0) = -\tau^{-1}\boldsymbol{\pi}_0\hat{Q}(\bar{\Delta} \oplus D_0)^{-1},$$

$$\mathbf{p}_i(1) = \tau^{-1}\{\boldsymbol{\pi}_0 \sum_{k=1}^{i}[-\hat{Q}(\bar{\Delta}\oplus D_0)^{-1}\tilde{D}_k + (1-p)N\mu Q_2 \otimes \int_0^\infty P(k,t)e^{-N\mu t}dt]\tilde{\Omega}_{i-k} +$$

$$+ \sum_{l=1}^{i}\boldsymbol{\pi}_l[\bar{Q}\tilde{\Omega}_{i-l} + (1-p)N\mu(Q_2 \otimes \sum_{k=0}^{i-l}\int_0^\infty P(k,t)e^{-N\mu t}dt)\tilde{\Omega}_{i-k-l}]\},$$

$$\boldsymbol{p}_i(2) = \tau^{-1}(1-p)\sum_{l=0}^{i-1}\boldsymbol{\pi}_l(Q_2 \otimes \int_0^\infty e^{-N\mu t}P(i-l-1,t)dt),$$

где

$$\tau = b_1 + \boldsymbol{\pi}_0\hat{Q}(-\tilde{D}_0)^{-1}\mathbf{e} + (1-p)\boldsymbol{\Pi}(1)\tilde{Q}_2\mathbf{e}(N\mu)^{-1}.$$

Доказательство проводится аналогично доказательству теоремы 7.2.

Следствие 7.10. *Векторы стационарных вероятностей* $\mathbf{p}_i, i \geq 0$, *процесса* $\{i_t, l_t, \nu_t\}, t \geq 0$, *вычисляются следующим образом:*

$$\mathbf{p}_i = \sum_{k=0}^{2} \mathbf{p}_i(k), i \geq 0.$$

7.4.4 Характеристики производительности системы

Среднее число запросов на первой фазе в момент окончания обслуживания на этой фазе и в произвольный момент времени:

$$L_1 = \mathbf{\Pi}'(1)\mathbf{e}, \quad \tilde{L}_1 = \mathbf{P}'(1)\mathbf{e}.$$

Среднее число запросов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$L_2 = \mathbf{\Pi}(1)(I_{N+R+1} \otimes \mathbf{e}_{\bar{W}})R_1\mathbf{e}, \qquad \tilde{L}_2 = \mathbf{P}(1)(I_{N+R+1} \otimes \mathbf{e}_{\bar{W}})R_1\mathbf{e},$$

где $R_1 = \mathrm{diag}\{r,\ r = \overline{0, N+R}\}$.

Дисперсия числа запросов на первой и второй фазах в произвольный момент времени:

$$D_1 = \mathbf{P}''(1)\mathbf{e} - (\tilde{L}_1)^2,\ D_2 = \mathbf{P}''(1)(I_{N+R+1} \otimes \mathbf{e}_{\bar{W}})R_1\mathbf{e} - (\tilde{L}_2)^2.$$

Среднее число занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$N_{busy} = \mathbf{\Pi}(1)(I_{N+R+1} \otimes \mathbf{e}_{\bar{W}})R_2\mathbf{e}, \qquad \tilde{N}_{busy} = \mathbf{P}(1)(I_{N+R+1} \otimes \mathbf{e}_{\bar{W}})R_2\mathbf{e},$$

где $R_2 = \mathrm{diag}\{r, r = 0, 1, \ldots, N, N, \ldots, N\}$.

Среднее число занятых мест в буфере в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$N_{buffer} = \mathbf{\Pi}(1)(I_{N+R+1} \otimes \mathbf{e}_{\bar{W}})R_3\mathbf{e}, \tilde{N}_{buffer} = \mathbf{P}(1)(I_{N+R+1} \otimes \mathbf{e}_{\bar{W}})R_3\mathbf{e},$$

где $R_3 = \mathrm{diag}\{0, \ldots, 0, 1, \ldots, R\}$.

Вероятность того, что произвольный запрос покинет систему (вызовет блокировку) после обслуживания на первой фазе:

$$P_{loss} = p\mathbf{\Pi}(1)\tilde{Q}_2\mathbf{e}, \quad P_{block} = (1-p)\mathbf{\Pi}(1)\tilde{Q}_2\mathbf{e}.$$

Вероятности P_{idle}, P_{serve}, $P_{block}^{(server)}$ того, что прибор первой фазы свободен, занят обслуживанием или заблокирован:

$$P_{idle} = \tau^{-1}\boldsymbol{\pi}_0\hat{Q}(-\tilde{D}_0)^{-1}\mathbf{e},\ P_{serve} = \tau^{-1}b_1,\ P_{block}^{(server)} = 1 - P_{idle} - P_{serve}.$$

7.4.5 Численные примеры

Эксперимент 7.4. В этом эксперименте исследуется влияние коэффициента корреляции входного потока и коэффициента загрузки на стационарные характеристики производительности системы.

Рассмотрим три $BMAP$-потока, описанные в эксперименте 7.1. $BMAP_1$ – это групповой пуассоновский поток с нулевым коэффициентом корреляции. $BMAP_2$ и $BMAP_3$ имеют коэффициенты корреляции $c_{cor} = 0.1$ и $c_{cor} = 0.2$ соответственно.

Время обслуживания на первой фазе постоянное и равно $b_1 = 0.1$. Число приборов на второй фазе $N = 6$, размер буфера $R = 3$, вероятность блокировки после обслуживания на первой фазе $p = 0.6$, интенсивность обслуживания на второй фазе $\mu = 2$.

На рисунке 2.8 представлены графики зависимости среднего числа запросов на первой фазе L_1 и на второй фазе L_2 от коэффициента загрузки системы ρ. Рисунок 2.9 показывает зависимость вероятностей P_{loss} и P_{block}, а рисунок 2.10 – зависимость дисперсий числа запросов на первой фазе D_1 и на второй фазе D_2 от коэффициента загрузки ρ. Изменение величины ρ происходит за счет изменения средней интенсивности λ.

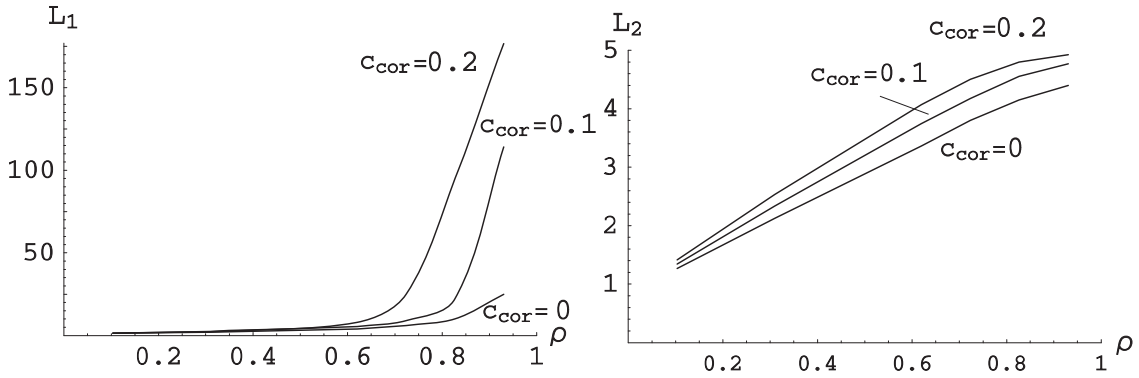

Рисунок 7.8. Зависимость среднего числа запросов на первой и второй фазах от загрузки системы для $BMAP$-потоков с различным коэффициентом корреляции

Из рисунков видно, что при росте ρ дисперсия D_2 возрастает до некоторого значения, а затем начинает убывать. Это объясняется невозможностью дальнейшего роста дисперсии при увеличении числа запросов на второй фазе вследствие ограниченного числа мест на этой фазе.

Эксперимент 7.5. Исследуем влияние коэффициента корреляции входного потока и размера промежуточного буфера на стационарные характеристики производительности системы.

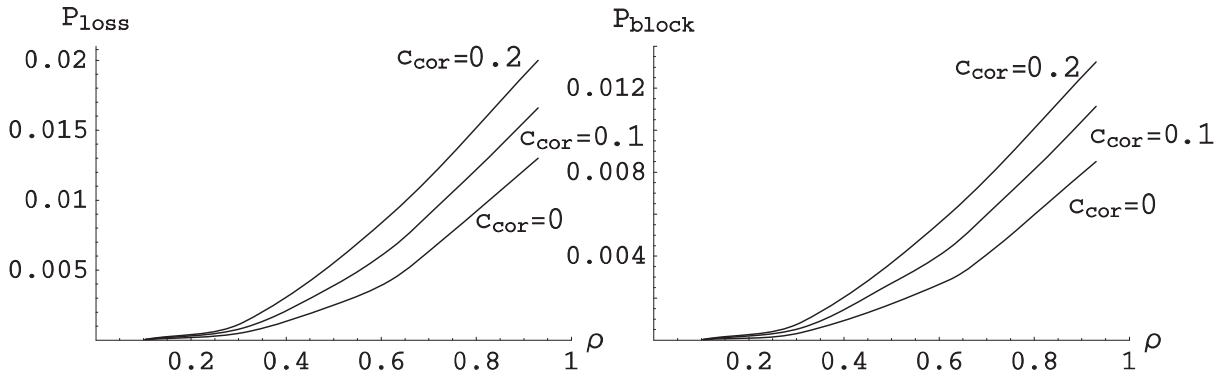

Рисунок 7.9. Зависимость вероятностей отказа и блокировки от загрузки системы для $BMAP$-потоков с различным коэффициентом корреляции

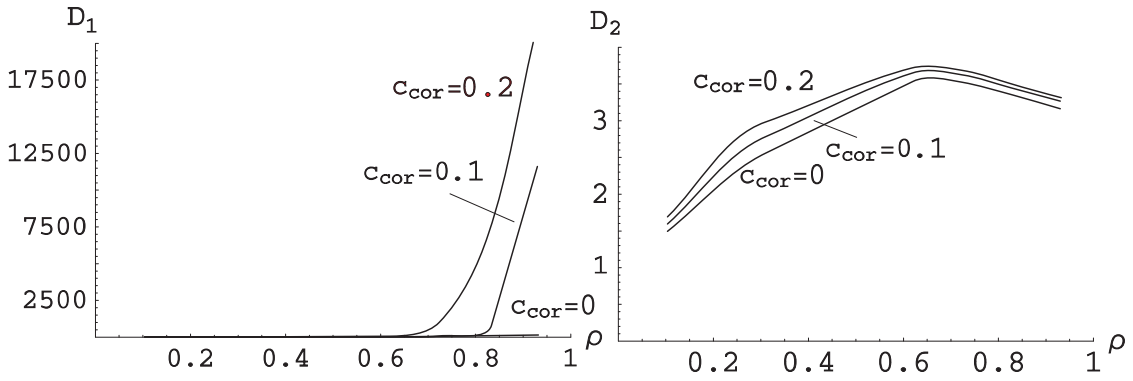

Рисунок 7.10. Зависимость дисперсий числа запросов на первой и второй фазах от загрузки системы для $BMAP$-потоков с различным коэффициентом корреляции

Рассмотрим четыре $BMAP$-потока, описанные в эксперименте 7.1, с максимальным размером групп $k = 3$, одинаковой средней интенсивностью $\lambda = 2.5$ и различными коэффициентами корреляции $c_{cor} = 0, 0.1, 0.2, 0.3$. Время обслуживания на первой фазе постоянное и равно $b_1 = 0.2$, $N = 3$, $p = 0.5$, $\mu = 1$.

Рисунки 7.11-7.13 иллюстрируют зависимость основных стационарных характеристик производительности системы от размера буфера R для $BMAP$-потоков с различным коэффициентом корреляции.

На основании приведенных рисунков можно сделать вывод, что значения характеристик $\tilde{L}_1$, $\tilde{N}_{busy}$, $P_{block}^{(server)}$ уменьшаются, а значения $\tilde{L}_2$, $\tilde{N}_{buffer}$, P_{idle} возрастают с ростом размера R промежуточного буфера для всех $BMAP$-потоков. Очевидно также, что при фиксированном значении размера буфера R рост коэффициента корреляции входного потока негативно влияет на процесс функционирования как первой фазы, так и второй и

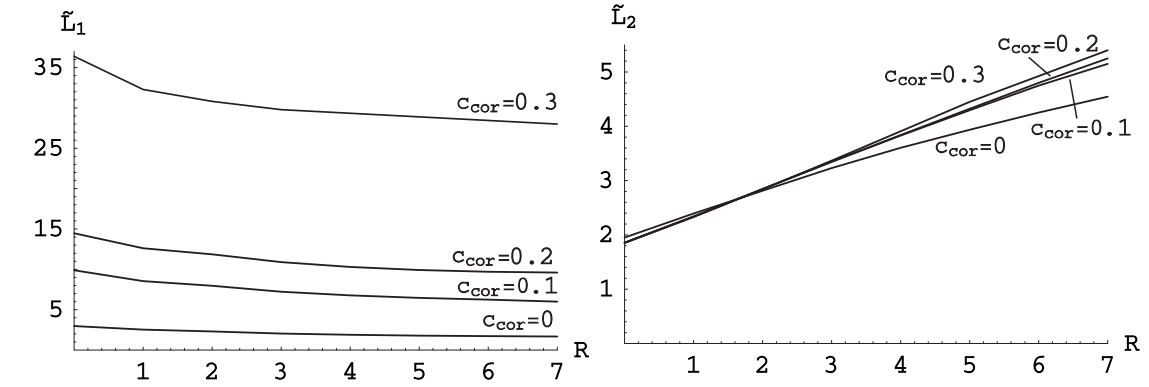

Рисунок 7.11. Зависимость среднего числа запросов на первой и второй фазах от размера буфера для $BMAP$-потоков с различным коэффициентом корреляции

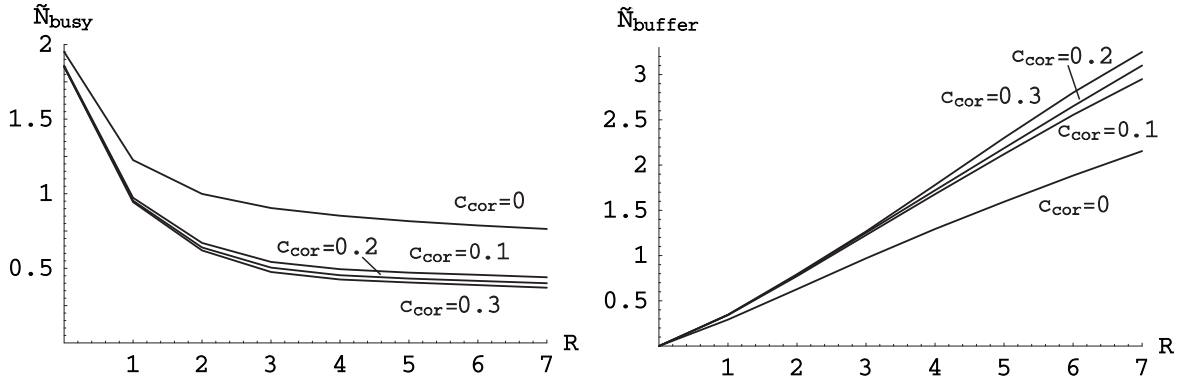

Рисунок 7.12. Зависимость среднего числа занятых приборов на второй фазе и занятых мест в буфере от размера буфера для $BMAP$-потоков с различным коэффициентом корреляции

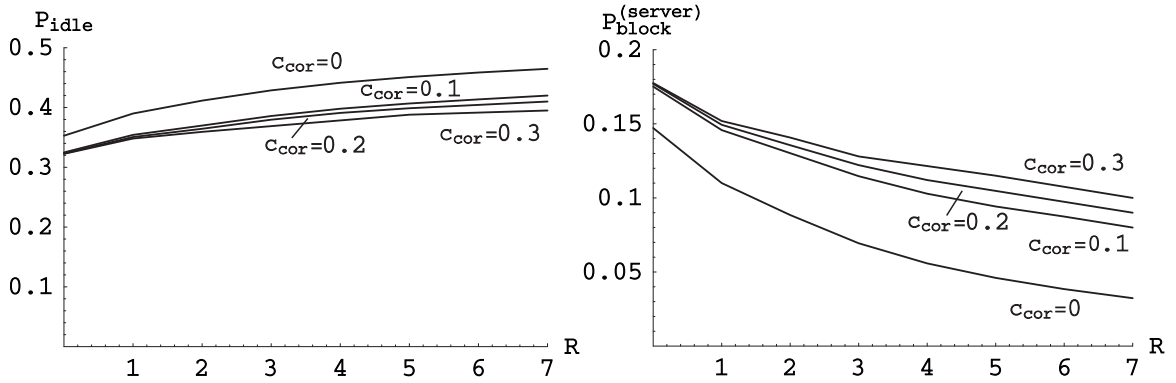

Рисунок 7.13. Зависимость вероятностей P_{idle} и $P_{block}^{(server)}$ от размера буфера для $BMAP$-потоков с различным коэффициентом корреляции

приводит к ухудшению качества обслуживания системы в целом. Таким образом, проведенные эксперименты подтверждают важность исследования

двухфазных систем с $BMAP$-потоком для точной оценки характеристик реальных систем.

7.5 Система $BMAP/G/1 \to \cdot/M/N/0$ с повторными вызовами и с групповым занятием приборов второй фазы

7.5.1 Математическая модель

Рассматриваемая в данном разделе система отличается от системы $BMAP/G/1 \to \cdot/M/N/0$ с ожиданием, исследованной в разделе 7.2, тем, что запрос, поступивший в систему и заставший прибор первой фазы занятым, не становится в очередь, а покидает систему на некоторое случайное время, иначе говоря, уходит на орбиту неограниченного объема, а затем повторяет попытки попасть на обслуживания. Более подробно: если поступившая группа первичных запросов застает обслуживающий прибор первой фазы свободным, то один из запросов группы начинает обслуживаться, а остальные идут на орбиту. Если прибор занят в момент поступления группы, то все запросы этой группы идут на орбиту. Каждый из запросов, находящихся на орбите, независимо от других запросов, повторяет свои попытки попасть на обслуживание через случайные интервалы времени. Если в момент времени t число запросов на орбите равно i, $i > 0$, то вероятность повторной попытки с орбиты в интервале $(t, t + \Delta t)$ равна $\alpha_i \Delta t + o(\Delta t)$. Относительно функции α_i, задающей суммарную интенсивность повторных попыток, предполагается, что $\lim\limits_{i\to\infty} \alpha_i = \infty$, $\alpha_0 = 0$. Такая функция, в частности, описывает классическую стратегию повторных попыток ($\alpha_i = i\alpha$) и линейную стратегию ($\alpha_i = i\alpha + \gamma,\ \alpha > 0$).

7.5.2 Стационарное распределение вложенной цепи Маркова

Пусть t_n – n-й момент окончания обслуживания на первой фазе. Рассмотрим процесс $\xi_n = \{i_n,\, r_n,\, \nu_n\}$, $n \geq 1$, где i_n – число запросов на орбите в момент t_n; r_n – число занятых приборов на второй фазе в момент $t_n - 0$; ν_n – состояние $BMAP$ в момент t_n, $i_n \geq 0, r_n = \overline{0, N}, \nu_n = \overline{0, W}$.

Процесс ξ_n, $n \geq 1$, является ЦМ. Перенумеруем ее состояния в лексикографическом порядке и сформируем матрицы $P_{i,j}$, $i, j \geq 0$, вероятностей

переходов цепи из состояний, соответствующих значению i счетной компоненты, в состояния, соответствующие значению j этой компоненты.

Анализируя поведение системы между соседними моментами окончания обслуживания на первой фазе, можно убедиться в справедливости следующего утверждения.

Лемма 7.6. *Матрица вероятностей переходов ЦМ ξ_n, $n \geq 1$, имеет блочную структуру $P = (P_{i,j})_{i,j\geq 0}$, где*

$$P_{i,j} = O,\ j < i-1,\ P_{i,j} = \bar{Q}A_i[\alpha_i\Omega_{j-i+1} + \sum_{k=1}^{j-i+1}\tilde{D}_k\Omega_{j-i-k+1}]+$$

$$+(1-p)[\sum_{n=0}^{j-i+1}\mathcal{F}_n\tilde{Q}_3A_{i+n}\alpha_{i+n}\Omega_{j-i-n+1}+$$

$$+\sum_{n=0}^{j-i}\mathcal{F}_n\tilde{Q}_3A_{i+n}\sum_{k=1}^{j-i-n+1}\tilde{D}_k\Omega_{j-i-n-k+1}],\ j \geq \max\{0, i-1\},\ i \geq 0, \quad (7.59)$$

$$A_i = \int_0^\infty e^{-\alpha_i t}e^{(\Delta\oplus D_0)t}dt = (\alpha_i I - \Delta\oplus D_0)^{-1},\ i \geq 0.$$

Здесь матрицы Ω_n, $\mathcal{F}_n$, $n \geq 0$, Δ, $\bar{Q}$, $\tilde{Q}_3$ и $\tilde{D}_k$, $k \geq 0$, совпадают с ранее введенными в разделе 7.1.

Доказательство. Доказательство проводится путем анализа возможных переходов ЦМ ξ_n, $n \geq 1$, за один шаг с учетом вероятностного смысла матриц, входящих в правую часть (7.59).

Матрица $\bar{Q}$ определяет вероятности того, что запрос, обслужившийся на первой фазе, либо застанет на второй фазе нужное для его обслуживания число свободных приборов и, следовательно, сразу начнет обслуживаться на этих приборах, либо не застанет нужного числа свободных приборов и уйдет из системы недообслуженным.

Матрица Δ является инфинитезимальным генератором процесса гибели, который описывает эволюцию числа занятых приборов на второй фазе между двумя соседними моментами окончания обслуживания на первой фазе.

Матрица $A_i\alpha_i$ определяет вероятности того, что обслуживание на первой фазе инициируется запросом с орбиты при условии, что в предыдущий момент окончания обслуживания на орбите находилось i запросов.

Матрица $A_i\tilde{D}_k$ имеет аналогичный вероятностный смысл с единственным отличием, что обслуживание на первой фазе инициируется первичным запросом, поступившим в группе размера k.

Матрицы Ω_n и $\mathcal{F}_n$ задают вероятности того, что за время обслуживания на первой фазе и за время блокировки прибора первой фазы, соответственно, в систему поступит n запросов в $BMAP$-потоке.

Матрица $(1-p)\mathcal{F}_n\tilde{Q}_3$ определяет вероятности того, что запрос, обслужившийся на первой фазе, не застанет нужного числа свободных приборов на второй фазе и будет ждать их освобождения, заблокировав прибор первой фазы, и за время блокировки в систему поступит n запросов.

Отметим, что все упомянутые матрицы, кроме матрицы Δ, имеют порядок $(N+1)\bar{W}\times(N+1)\bar{W}$ и элемент, стоящий на пересечении $(r\bar{W}+\nu)$-й строки и $(r'\bar{W}+\nu')$-го столбца каждой из этих матриц, задает вероятность события, озвученного при описании вероятностного смысла матрицы в целом и сопровождающегося переходом числа занятых приборов на второй фазе из состояния r в состояние r' и переходом управляющегося процесса $BMAP$ из состояния ν в состояние ν'.

Принимая во внимание приведенные выше пояснения и используя формулу полной вероятности, получаем искомое выражение (7.59) для вероятностей переходов рассматриваемой ЦМ. □

Как видно из (7.59), зависимость матриц вероятностей переходов $P_{i,j}$ от значений i и j счетной компоненты не сводится к зависимости от $j-i$, как в случае КТЦМ, или ЦМ типа $M/G/1$, см. [16, 150], описывающих многие СМО с ожиданием. В то же время зависимость от i ослабевает при $i\to\infty$ и матрицы $P_{i,j}$ в пределе зависят от значений i,j только через их разность $j-i$, то есть существуют пределы

$$\tilde{Y}_k=\lim_{i\to\infty}P_{i,i+k-1},\ k\geq 0. \tag{7.60}$$

Данное обстоятельство, дополненное тем, что рассматриваемая цепь ξ_n, $n\geq 1$, является неприводимой и непериодической, позволяет отнести эту цепь к классу АКТЦМ (см. [150]).

Пусть цепь η_n, $n\geq 1$, является предельной по отношению к цепи ξ_n, $n\geq 1$, и пусть $\tilde{Y}(z)$, $|z|\leq 1$, – ПФ матриц $\tilde{Y}_k$, $k\geq 0$, вероятностей переходов этой предельной цепи. Используя (7.60) и лемму 7.1, получим следующее утверждение.

Следствие 7.11. *ЦМ ξ_n, $n\geq 1$, принадлежит классу АКТЦМ. ПФ матриц вероятностей переходов ее предельной цепи имеет следующий*

вид:

$$\tilde{Y}(z) = [\bar{Q} + (1-p)\mathcal{F}(z)\tilde{Q}_3]\Omega(z), \tag{7.61}$$

где

$$\Omega(z) = \sum_{n=0}^{\infty} \Omega_n z^n,\ \mathcal{F}(z) = \sum_{n=0}^{\infty} \mathcal{F}_n z^n,\ |z| \le 1.$$

Теорема 7.16. *Достаточным условием эргодичности ЦМ* ξ_n, $n \ge 1$, *является выполнение неравенства*

$$\rho < 1, \tag{7.62}$$

где

$$\rho = \lambda[b_1 + (1-p)\mathbf{y}\int_0^{\infty} tdF(t)Q_3\mathbf{e}].$$

Здесь вектор $\mathbf{y}$ *является единственным решением СЛАУ*

$$\mathbf{y}\mathcal{Q}\int_0^{\infty} e^{\Delta t}dB(t) = \mathbf{y},\ \ \mathbf{ye} = 1. \tag{7.63}$$

Доказательство. Матрица $\tilde{Y}(1)$, определенная формулой (7.61), является неприводимой, поэтому воспользуемся видом (3.141), (3.142) достаточного условия существования стационарного распределения многомерных АКТЦМ.

Используя выражение (7.61) для матрицы $\tilde{Y}(z)$, перепишем систему (3.142) в виде

$$\mathbf{x}[(Q_1+pQ_2)\otimes I_{\bar{W}}+(1-p)\int_0^{\infty} dF(t)Q_3\otimes e^{D(1)t}]\int_0^{\infty} e^{\Delta t}\otimes e^{D(1)t}dB(t) = \mathbf{x}, \tag{7.64}$$

$$\mathbf{xe} = 1.$$

Подставляя вектор $\mathbf{x}$ вида $\mathbf{x} = \mathbf{y}\otimes\boldsymbol{\theta}$ в систему (7.64), легко убедиться, что такой вектор является решением этой системы и, соответственно, системы (3.142). Заметим, что $\mathbf{y}$ – единственное решение системы (7.63), так как матрица $\mathcal{Q}\int_0^{\infty} e^{\Delta t}dB(t)$ является стохастической неприводимой.

Подставляя теперь $\mathbf{x} = \mathbf{y}\otimes\boldsymbol{\theta}$ и выражение для $\tilde{Y}'(1)$, вычисленное с использованием (7.61), в неравенство (3.141), сведем это неравенство к виду (7.62) путем ряда алгебраических преобразований. □

Замечание 7.1. Вектор $\mathbf{y}$ задает стационарное распределение числа занятых приборов на второй фазе в момент окончания обслуживания на первой фазе при условии, что прибор первой фазы работает без остановки. Тогда выражение $(1-p)\mathbf{y}\int\limits_0^\infty tdF(t)Q_3\mathbf{e}$ определяет среднее время блокировки прибора первой фазы в условиях перегрузки системы, а величина ρ является коэффициентом загрузки.

В дальнейшем будем предполагать, что неравенство (7.62) выполняется, и найдем стационарное распределение цепи ξ_n, $n \geq 1$. Пусть $\boldsymbol{\pi}_i$, $i \geq 0$, – векторы, задающие искомое стационарное распределение. Для нахождения их используется алгоритм, приведенный выше для АКТЦМ.

7.5.3 Стационарное распределение в произвольный момент времени

Рассмотрим процесс $\zeta_t = \{i_t, r_t, \nu_t, k_t\}$, $t \geq 0$, состояний системы в произвольный момент времени t, где i_t – число запросов на орбите, r_t – число занятых приборов на второй фазе, ν_t – состояние управляющего процесса $BMAP$-потока, k_t – случайная величина, принимающая значения 0, 1, 2 в зависимости от того, свободен, занят или заблокирован прибор первой фазы, в момент времени t.

Пусть

$$p(i,r,\nu,k) = \lim_{t\to\infty} P\{i_t = i,\ r_t = r,\ \nu_t = \nu,\ k_t = k\}, \tag{7.65}$$

$$i \geq 0,\ r = \overline{0,N},\ \nu = \overline{0,W},\ k = 0,1,2,$$

стационарное распределение вероятностей процесса ζ_t, $t \geq 0$. Пусть также $\mathbf{p}_i(k)$ – вектор-строка вероятностей $p(i,r,\nu,k)$, упорядоченных в лексикографическом порядке компонент (r,ν), $i \geq 0$, $k = 0,1,2$.

Теорема 7.17. *Ненулевые векторы $\mathbf{p}_i(k)$ стационарных вероятностей процесса ζ_t, $t \geq 0$, выражаются через стационарное распределение $\boldsymbol{\pi}_i$, $i \geq 0$, вложенной ЦМ ξ_n, $n \geq 1$, следующим образом:*

$$\mathbf{p}_0(0) = \tau^{-1}\boldsymbol{\pi}_0[\bar{Q} + (1-p)\mathcal{F}_0\tilde{Q}_3]A_0, \tag{7.10}$$

$$\mathbf{p}_i(0) = \tau^{-1}[\boldsymbol{\pi}_i\bar{Q}A_i + (1-p)\sum_{l=0}^{i}\boldsymbol{\pi}_l\mathcal{F}_{i-l}\tilde{Q}_3A_i],\ i \geq 1, \tag{7.67}$$

$$\mathbf{p}_i(1) = \tau^{-1}\{\sum_{l=0}^{i} \boldsymbol{\pi}_l[\bar{Q}A_l \sum_{k=1}^{i-l+1} \tilde{D}_k\tilde{\Omega}_{i-l-k+1}+$$

$$+(1-p)\sum_{k=0}^{i-l} \mathcal{F}_k\tilde{Q}_3A_{l+k} \sum_{m=1}^{i-l-k+1} \tilde{D}_m\tilde{\Omega}_{i-l-k-m+1}]+$$

$$+\sum_{l=0}^{i+1} \boldsymbol{\pi}_l[\bar{Q}A_l\alpha_l\tilde{\Omega}_{i-l+1} + (1-p)\sum_{k=0}^{i-l+1} \mathcal{F}_k\tilde{Q}_3A_{l+k}\alpha_{l+k}\tilde{\Omega}_{i-l-k+1}]\}, \; i \geq 0, \quad (7.68)$$

$$\mathbf{p}_i(2) = \tau^{-1}(1-p)\sum_{l=0}^{i} \boldsymbol{\pi}_l \sum_{k=0}^{i-l}(\mathcal{F}_k+\hat{\delta}_{0,k}I)\tilde{Q}_2 \int_0^\infty e^{-\mu\mathcal{N}t}\otimes P(i-l-k,t)dt,$$

$$i \geq 0, \quad (7.69)$$

где τ – средняя величина интервала между двумя соседними моментами окончания обслуживания на первой фазе:

$$\tau = b_1 + (1-p)\sum_{i=0}^{\infty} \boldsymbol{\pi}_i(I_{N+1}\otimes \mathbf{e}_{\bar{W}}) \int_0^\infty tdF(t)Q_3\mathbf{e}+$$

$$+\sum_{i=0}^{\infty} \boldsymbol{\pi}_i[\bar{Q}(\mathbf{e}_{N+1}\otimes I_{\bar{W}})(\alpha_i I - D_0)^{-1}+$$

$$+(1-p)\sum_{n=0}^{\infty} \mathcal{F}_n\tilde{Q}_3(\mathbf{e}_{N+1}\otimes I_{\bar{W}})(\alpha_{i+n}I - D_0)^{-1}]\mathbf{e}_{\bar{W}}.$$

Доказательство. Согласно определению, данному в [112], процесс ζ_t, $t \geq 0$, является полурегенерирующим с вложенным процессом марковского восстановления $\{\xi_n, t_n\}$, $n \geq 1$. Пределы (7.65) существуют, если процесс $\{\xi_n, t_n\}$ является неприводимым непериодическим и выполняется неравенство $\tau < \infty$. В нашем случае все эти условия выполняются, если вложенная ЦМ ξ_n, $n \geq 1$, эргодична. Поэтому достаточным условием существования пределов (7.65) является выполнение неравенства (7.62). Выражения (7.66)-(7.69) для векторов, определяющих стационарное распределение, выводятся с использованием предельной теоремы для полурегенерирующих процессов, приведенной в [112], и вероятностного смысла входящих в эти выражения матриц (см. доказательство леммы 7.4). □

Следствие 7.12. *Векторы* $\mathbf{p}_i$, $i \geq 0$, *стационарных вероятностей процесса* $\{\tilde{i}_t, r_t, \nu_t\}, t \geq 0$, *где* $\tilde{i}_t$ *– число запросов на первой фазе (на орбите и на приборе, включая заблокированный), вычисляются следующим образом:*

$$\mathbf{p}_0 = \mathbf{p}_0(0),\ \mathbf{p}_i = \mathbf{p}_i(0) + \sum_{k=1}^{2} \mathbf{p}_{i-1}(k),\ i \geq 1.$$

7.5.4 Характеристики производительности системы

Полученные результаты позволяют рассчитывать стационарное распределение рассматриваемой СМО и всевозможные характеристики ее производительности. Наиболее важные из них приведены ниже.

Среднее число запросов на первой фазе в моменты окончания обслуживания и в произвольный момент времени:

$$L = \mathbf{\Pi}'(1)\mathbf{e},\ \tilde{L} = \mathbf{P}'(1)\mathbf{e}.$$

где $\mathbf{\Pi}(z) = \sum\limits_{i=0}^{\infty} \boldsymbol{\pi}_i z^i,\ \mathbf{P}(z) = \sum\limits_{i=0}^{\infty} \mathbf{p}_i z^i,\ |z| \leq 1.$

Векторы стационарного распределения числа занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$\mathbf{r} = \mathbf{\Pi}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}}),\ \tilde{\mathbf{r}} = \mathbf{P}(1)(I_{N+1} \otimes \mathbf{e}_{\bar{W}}).$$

Среднее число занятых приборов на второй фазе в моменты окончания обслуживания на первой фазе и в произвольный момент времени:

$$N_{busy} = \mathbf{r}\mathcal{N}\mathbf{e},\ \tilde{N}_{busy} = \tilde{\mathbf{r}}\mathcal{N}\mathbf{e}.$$

Вероятность того, что произвольный запрос покинет систему (вызовет блокировку) после обслуживания на первой фазе:

$$P_{loss} = p\mathbf{\Pi}(1)\tilde{Q}_2\mathbf{e},\quad P_{block} = (1-p)\mathbf{\Pi}(1)\tilde{Q}_2\mathbf{e}.$$

Вероятности P_{idle}, P_{serve}, P_{block} того, что прибор первой фазы свободен, занят обслуживанием или заблокирован:

$$P_{idle} = \sum_{i=0}^{\infty} \mathbf{p}_i(0)\mathbf{e},\ P_{serve} = \tau^{-1}b_1,\ P_{block} = 1 - P_{idle} - P_{serve}.$$

Вероятность того, что произвольный запрос, поступивший в систему, сразу пойдет на обслуживание (не посетив орбиту):

$$P_{imm} = -\lambda^{-1} \sum_{i=0}^{\infty} \mathbf{p}_i(0)(\mathbf{e}_{N+1} \otimes D_0\mathbf{e}). \tag{7.70}$$

Коротко поясним вывод формулы (7.70). Произвольный запрос попадет на обслуживание сразу, если в момент его поступления прибор первой фазы свободен. Вероятность того, что произвольный запрос поступит в группе размера k и в момент поступления прибор будет свободен, вычисляется по формуле $\sum\limits_{i=0}^{\infty} \mathbf{p}_i(0)(I_{N+1} \otimes \sum\limits_{k=1}^{\infty} \frac{kD_k}{\lambda})\mathbf{e}$. Предполагаем, что произвольный запрос, поступивший в группе размера k, помещается на первое место в группе с вероятностью $1/k$. Тогда по формуле полной вероятности

$$P_{imm} = \sum_{i=0}^{\infty} \mathbf{p}_i(0)(I_{N+1} \otimes \sum_{k=1}^{\infty} \frac{kD_k}{\lambda} \frac{1}{k})\mathbf{e}. \tag{7.71}$$

Преобразовывая (7.71) с учетом соотношения $\sum\limits_{k=1}^{\infty} D_k\mathbf{e} = -D_0\mathbf{e}$, получим (7.70).

Вероятность того, что произвольный запрос, поступивший в систему, успешно обслужится на обеих фазах (не попадет на орбиту, не потеряется и не вызовет блокировку прибора первой фазы):

$$P_{success} = -\lambda^{-1} \sum_{i=0}^{\infty} \mathbf{p}_i(0)(I_{N+1} \otimes D_0\mathbf{e}) \int_0^{\infty} e^{\Delta t} dB(t) Q_1\mathbf{e}. \tag{7.72}$$

Формула (7.72) становится понятной, если принять во внимание то, что вектор-строка $-\lambda^{-1} \sum\limits_{i=0}^{\infty} \mathbf{p}_i(0)(I_{N+1} \otimes D_0\mathbf{e})$ определяет распределение числа занятых приборов на второй фазе в момент поступления произвольного помеченного запроса, заставшего прибор первой фазы свободным, а вектор-столбец $\int\limits_0^{\infty} e^{\Delta t} dB(t) Q_1\mathbf{e}$ определяет вероятности того, что помеченный запрос после обслуживания на первой фазе застанет на второй фазе нужное для его обслуживания число свободных приборов.

7.5.5 Численные примеры

Эксперимент 7.6. В этом эксперименте исследуется влияние коэф-

фициента корреляции входного потока на основные стационарные характеристики производительности.

Рассмотрим три MAP-потока, которые определяются матрицами D_0 и $D_1 = D$, имеют одинаковую среднюю интенсивность $\lambda = 7$, но различные коэффициенты корреляции.

MAP_1 – это стационарный пуассоновский поток, который определяется матрицами $D_0 = -7$ и $D = 7$ и имеет коэффициент корреляции $c_{cor} = 0$.

MAP_2 имеет коэффициент корреляции $c_{cor} = 0.1$ и определяется матрицами

$$D_0 = \begin{pmatrix} -8.22458 & 8.22458 \times 10^{-6} \\ 8.22458 \times 10^{-6} & -0.180152 \end{pmatrix}, D = \begin{pmatrix} 8.199422 & 0.02515 \\ 0.140373 & 0.039771 \end{pmatrix}.$$

MAP_3 имеет коэффициент корреляции $c_{cor} = 0.2$ и определяется матрицами

$$D_0 = \begin{pmatrix} -9.443532 & 7.63574 \times 10^{-6} \\ 7.63574 \times 10^{-6} & -0.307237 \end{pmatrix}, D = \begin{pmatrix} 9.380959 & 0.062565 \\ 0.171397 & 0.135832 \end{pmatrix}.$$

На основе этих MAP-потоков построим $BMAP$-потоки с максимальным размером групп, равным трем. Каждый из $BMAP$-потоков определяется матрицами $D_k, k = \overline{0,3}$, полученными следующим образом: матрица D_0 совпадает с соответствующей матрицей MAP-потока, а остальные матрицы вычисляются по формуле $D_k = D\varkappa^{k-1}(1-\varkappa)/(1-\varkappa^3), k = \overline{1,3}$, где $\varkappa = 0.8$. Затем все матрицы $D_k, k = \overline{0,3}$, умножим на некоторую положительную константу, чтобы получить $BMAP$-поток со средней интенсивностью $\lambda = 7$. $BMAP$-поток, полученный из MAP_n, будем в дальнейшем обозначать как $BMAP_n$, $n = 1, 2, 7$.

Время обслуживания на первой фазе имеет распределение Эрланга порядка 3 с параметром 60. Среднее время обслуживания $b_1 = 0.05$, квадрат коэффициента вариации $c_{var}^2 = 1/3$.

Полагаем, что стратегия повторных вызовов – классическая, то есть функция α_i имеет вид $\alpha_i = i\alpha$, $i \geq 0$.

В рамках данного эксперимента $\alpha = 10$. Число приборов на второй фазе $N = 7$. Вероятность того, что запрос, не получивший немедленного доступа на вторую фазу, будет потерян, равна $p = 0.6$. Параметры обслуживания на второй фазе определяются интенсивностью $\mu = 3$ и вероятностями $q_0 = 0$, $q_1 = 0.9$, $q_2 = q_3 = 0.05$.

На рисунке 7.14 представлены графики зависимости среднего числа запросов на первой фазе $\tilde{L}$ и вероятности потери P_{loss} произвольного запроса

после обслуживания на первой фазе от коэффициента загрузки системы ρ.

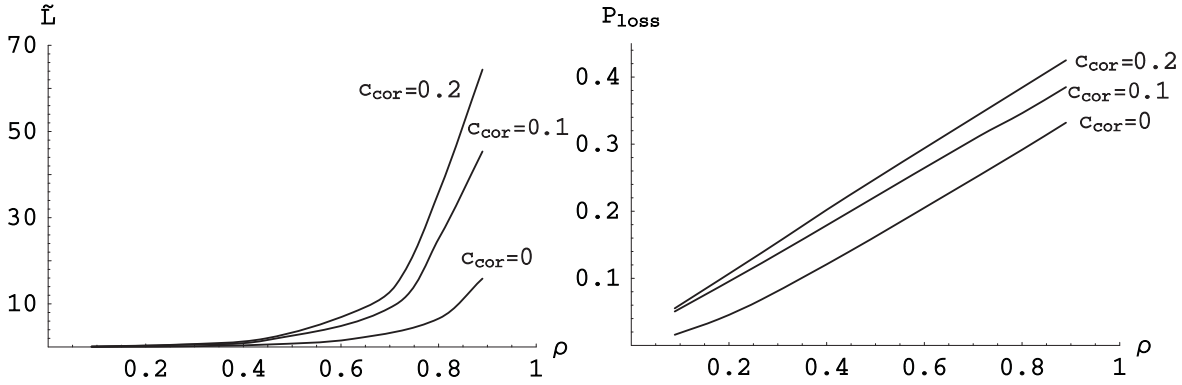

Рисунок 7.14. Зависимость среднего числа запросов на первой фазе и вероятности потери от загрузки системы для $BMAP$-потоков с различными коэффициентами корреляции

Рисунок 7.15 иллюстрирует зависимость вероятности P_{imm} немедленного доступа на прибор первой фазы и вероятности $P_{success}$ успешного обслуживания запроса на обеих фазах коэффициента загрузки системы ρ. Изменение величины ρ происходит за счет изменения интенсивности λ, которая изменяется путем нормирования элементов матриц $D_k, k = \overline{0,3}$.

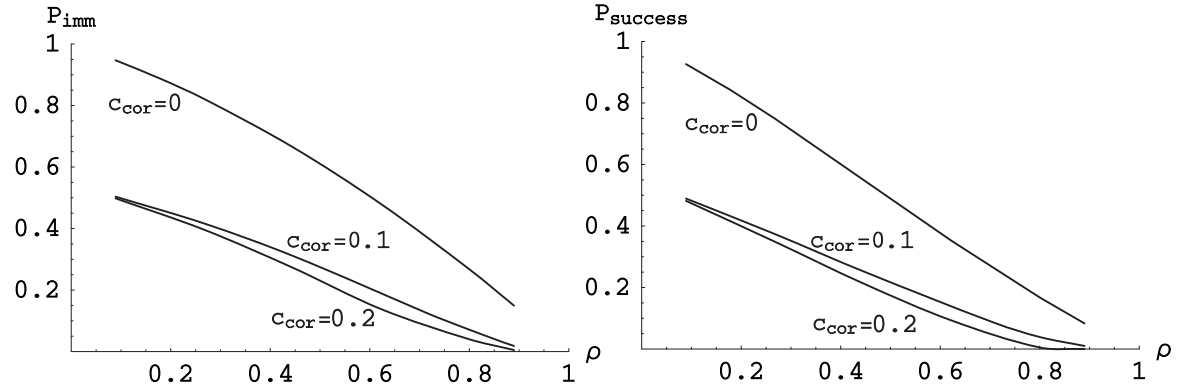

Рисунок 7.15. Зависимость вероятностей P_{imm} и $P_{success}$ от загрузки системы для $BMAP$-потоков с различным коэффициентом корреляции

Из приведенных графиков видно, что при одном и том же значении загрузки ρ значения характеристик системы значительно отличаются для $BMAP$-потоков с разной корреляцией. Рост коэффициента корреляции вызывает увеличение $\tilde{L}$, P_{loss} и приводит к уменьшению вероятностей P_{imm} и $P_{success}$, то есть негативно сказывается на качестве обслуживания в рассматриваемой СМО.

Проведенный эксперимент подтверждает необходимость учета влияния корреляции при прогнозировании поведения системы, ибо игнорирование этой важной характеристики входного потока может привести к значительным ошибкам при проектировании и оценке производительности системы.

Эксперимент 7.7. В данном эксперименте исследуем численно задачу оптимального выбора параметра p – вероятности потери запроса после первой фазы в случае отсутствия необходимого для его обслуживания числа свободных приборов на второй фазе. Целесообразность такой задачи вытекает из следующего. Для того чтобы улучшить качество обслуживания в рассматриваемой системе, необходимо уменьшить вероятность потери запросов P_{loss}. В частности это можно сделать за счет уменьшения вероятности p, однако при этом увеличится вероятность блокировки прибора первой фазы, что приведет к увеличению числа запросов на орбите и негативно скажется на качестве обслуживания.

В качестве стоимостного критерия качества обслуживания рассмотрим средний штраф в единицу времени

$$J = J(p) = c_1\tilde{L} + c_2\lambda P_{loss},$$

где c_1 – плата за нахождение одного запроса на орбите в течение единицы времени, c_2 – штраф за потерю одного запроса в единицу времени.

Задача состоит в нахождении оптимального значения вероятности p, которое доставляет минимум критерию.

В качестве входного потока рассмотрим $BMAP_2$ из предыдущего эксперимента. Вероятности q_n возьмем в виде $q_0 = 0$, $q_1 = 1$, $q_2 = q_3 = 0$, то есть каждому запросу для обслуживания на второй фазе требуется один прибор. Остальные параметры полагаем такими же, как и в предыдущем эксперименте.

Стоимостные коэффициенты возьмем следующими: c_1=2, c_2=10.

На рисунке 7.16 представлен график функции $J(p)$ при различных значениях интенсивности α повторных вызовов: $\alpha = 5, 10, 15, 50$.

Как видно из рисунка, при фиксированной величине p уменьшение интенсивности повторных вызовов приводит к увеличению значения критерия J. Наибольшая разница в значениях критерия наблюдается в области малых значений p. В то же время точки минимума очень близки и располагаются в интервале $(0.315, 0.405)$. Интересно отметить, что точки минимума не сдвигаются при увеличении загрузки системы, которая уменьшается

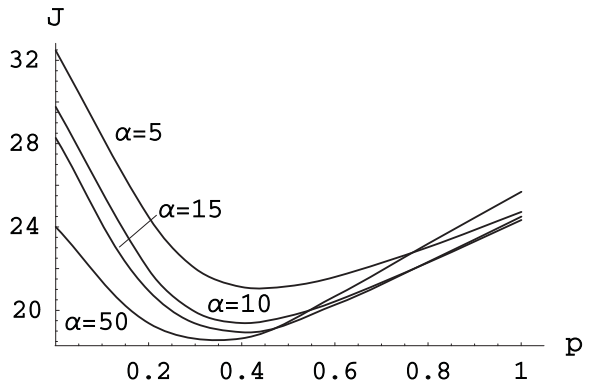

Рисунок 7.16. Зависимость критерия качества J от p при различных значениях интенсивности повторных вызовов α

с ростом p, как это показано в таблице 7.1.

Таблица 7.1. Коэффициент загрузки ρ при различных значениях p

p	0	0.2	0.3	0.4	0.5	0.6	0.8	1
ρ	0.8195	0.7256	0.6786	0.6317	0.5847	0.5378	0.4439	0.3500

При фиксированной интенсивности повторов α относительный выигрыш, полученный при выборе оптимального параметра p^* по сравнению с произвольным значением p этого параметра, вычисляется по формуле $g(p) = \frac{J(p)-J^*}{J^*}100\%$.

Приведем значения относительного выигрыша при $\alpha = 10$. В этом случае минимум достигается при $p^* = 0.4$ и равен $J^* = 19.3844$. Таблица 7.2 содержит значения относительного выигрыша $g(p)$ для этого случая.

Таблица 7.2. Относительный выигрыш $g(p)$ при различных значениях p

p	0	0.2	0.3	0.4	0.5	0.6	0.8	1
$g(p)$	54%	13%	3%	0%	2%	5%	15%	26%

Как видно из таблицы, даже при близких к оптимальному значениях p величина относительного выигрыша составляет 2-13%, в то время как максимальный относительный выигрыш может достигать 54%.

7.6 Многофазный тандем многолинейных СМО без буферов

Теория тандемных (многофазных) систем массового обслуживания представляет собой связывающее звено между теорией систем массового обслуживания и теорией сетей массового обслуживания. Тандемная система может рассматриваться как простейший случай сети массового обслуживания с линейной топологией. Тандемные системы являются популярной темой для исследований, см., например, работы [99, 125, 132–135, 165].

Большинство работ, посвященных тандемным системам, ограничиваются рассмотрением двухфазных тандемов со стационарным пуассоновским входящим потоком и экспоненциально распределенным временем обслуживания. В данном разделе рассматривается тандемная система, состоящая из произвольного конечного числа фаз (станций), представленных многолинейными системами без буферов, в которую поступает MAP поток запросов. Предположение, что процесс поступления запросов определяется как MAP, позволяет учесть коррелированный нестационарный характер информационных потоков в современных телекоммуникационных сетях, см., например, [136, 149].

Времена обслуживания на станциях тандема имеют PH-распределение, которое, как известно, является значительно более общим, чем экспоненциальное, и позволяет хорошо моделировать реальные процессы обслуживания.

Запрос, поступающий в рассматриваемую систему, должен получить последовательное обслуживание на всех станциях тандема. Однако в силу отсутствия мест для ожидания на станциях запросы могут быть потеряны на каждой из них. Качество обслуживания в системе определяется, в основном, вероятностью успешного обслуживания произвольного запроса на всех станциях. Вместе с тем, для оценки эффективности тандема, а также для обнаружения и предотвращения так называемых узких мест в тандемной сети, существенно используются вероятности потерь на различных участках и подсистемах тандема.

В качестве релевантных работ стоит отметить работы [103, 104]. В этих работах рассматриваются системы с однолинейными станциями, рекуррентными входящими потоками и экспоненциально распределенными временами обслуживания. Поскольку интервалы между поступлениями запросов в рекуррентном потоке являются независимыми случайными вели-

чинами, распределенными по произвольному закону, то этот поток можно рассматривать как более общий, чем MAP. Вместе с тем, и MAP можно считать более общим потоком по сравнению с рекуррентным, поскольку в нем времена между моментами поступления запросов могут быть зависимыми.

Частный случай рассматриваемой в этом разделе системы был исследован в работе [151]. Здесь предполагалось, что времена обслуживания запросов на станциях имеют экспоненциальное распределение.

7.6.1 Описание системы

Рассматривается тандемная система массового обслуживания, состоящая из $R,\ R > 1$, станций. В терминах обозначений Кендалла эта система может быть описана как

$$MAP/PH^{(1)}/N_1/N_1 \to \bullet/PH^{(2)}/N_2/N_2 \to \cdots \to \bullet/PH^{(R)}/N_R/N_R.$$

Станция номер $r,\ r = 1, \ldots, R$, представлена системой из N_r приборов без буфера. Приборы, принадлежащие одной и той же станции, независимые и идентичные. На вход первой станции поступает MAP-поток, который характеризуется пространством состояний $\{0, 1, ..., W\}$ управляющего процесса $\nu_t, t \geq 0$, и матрицами D_0, D_1. Время обслуживания запроса на r-й станции имеет PH-распределение с неприводимым представлением $(\boldsymbol{\beta}^{(r)}, S^{(r)}), r = \overline{1, R}$.

Если запрос, поступающий на первую станцию или переходящий на r-ю, $r = 2, \ldots, R$, станцию после обслуживания на $(r-1)$-й станции, застает все приборы занятыми, то он покидает тандем навсегда.

Целью данного раздела является изучение выходящих потоков со станций тандема, расчет стационарного распределения тандема и его фрагментов, а также вычисление вероятностей потерь, ассоциированных с тандемом.

7.6.2 Потоки, выходящие из станций

Процесс изменения состояний системы описывается в терминах неприводимой многомерной цепи Маркова с непрерывным временем $\xi_t, t \geq 0$,

$$\xi_t = \{n_t^{(R)}, m_{1,t}^{(R)}, \ldots, m_{n^{(R)},t}^{(R)};\ n_t^{(R-1)}, m_{1,t}^{(R-1)}, \ldots,$$

$$m^{(R-1)}_{n^{(R-1)},t}; \ldots, n^{(1)}_t, m^{(1)}_{1,t}, \ldots, m^{(1)}_{n^{(1)},t};\ \nu_t\},\ t \geq 0,$$

где

- $n^{(r)}_t$ число занятых приборов на r-й станции, $n^{(r)}_t = \overline{0, N^{(r)}},\ r = \overline{1, R}$;
- $m^{(r)}_{l,t}$ –фаза -процесса обслуживания на l-м занятом приборе r-й станции, $m^{(r)} = l, t = \overline{1, M_r},\ r = \overline{1, R},\ l = \overline{1, n^{(r)}}$;
- $\nu_t,\ \nu_t \in \{0, \ldots, W\}$ – состояние управляющего процесса MAP-потока в момент времени t.

Вектор-строка $\mathbf{p}$ стационарных вероятностей состояний цепи имеет размерность

$$(W+1)\prod_{r=1}^{R}\frac{M_r^{N_{r+1}}-1}{M_r-1}$$

и вычисляется как единственное решение системы линейных алгебраических уравнений

$$\mathbf{p}Q = \mathbf{0},\ \mathbf{pe} = 1, \tag{7.73}$$

где матрица Q является инфинитезимальным генератором цепи Маркова $\xi_t,\ t \geq 0$.

Построение этой матрицы может быть выполнено с помощью стандартной методики, используемой в теории сетей массового обслуживания и является не слишком трудной задачей. Но эта работа в случае более или менее большого значения R является достаточно трудоемкой. Вследствие этого представляется интересным найти способ для расчета стационарных распределений состояний тандема в целом или его частей (фрагментов) или маргинальных стационарных распределений состояний любой станции, не записывая явное выражение генератора Q. Стоит отметить, что функционирование фрагмента тандема, состоящего из любого количества станций и расположенного в начале тандема, не зависит от состояний остальных станций. Таким образом, интуитивно ясно, что какая-либо декомпозиция может быть применена для расчета стационарного распределения тандема и его фрагментов без полного построения генератора.

В данном исследовании мы разрабатываем простой, точный и удобный метод вычисления маргинальных стационарных распределений вероятностей фрагментов тандема, а также всего тандема и соответствующих вероятностей потерь. Этот метод опирается на результаты исследования потоков, выходящих из станций тандема. Эти потоки принадлежат классу марковских потоков и описываются следующей теоремой.

Теорема 7.18. *Выходящий поток из r-й станции (входящий поток на $(r+1)$-ю станцию), $r \in \{1,2,\ldots,R-1\}$, принадлежит классу MAP-потоков. Этот MAP-поток задается матрицами $D_0^{(r+1)}$ и $D_1^{(r+1)}$, которые вычисляются по следующим рекуррентным формулам:*

$$D_0^{(r+1)} = diag\{(S^{(r)})^{\oplus n},\, n = \overline{0,N_r}\} \otimes I_{K_r} + \tag{7.74}$$

$$+\begin{pmatrix} D_0^{(r)} & \boldsymbol{\beta}^{(r)} \otimes D_1^{(r)} & 0 & \ldots & 0 & 0 \\ 0 & I_{M_r} \otimes D_0^{(r)} & I_{M_r} \otimes \boldsymbol{\beta}^{(r)} \otimes D_1^{(r)} & \ldots & 0 & 0 \\ 0 & 0 & I_{M_r^2} \otimes D_0^{(r)} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & I_{M_r^{N_r-1}} \otimes D_0^{(r)} & I_{M_r^{N_r-1}} \otimes \boldsymbol{\beta}^{(r)} \otimes D_1^{(r)} \\ 0 & 0 & 0 & \ldots & 0 & I_{M_r^{N_r}} \otimes (D_0^{(r)} + D_1^{(r)}) \end{pmatrix},$$

$$D_1^{(r+1)} =$$

$$= \begin{pmatrix} 0_{1\times 1} & 0 & \ldots & 0 & 0 & 0 \\ (\boldsymbol{S}_0^{(r)})^{\oplus 1} & 0_{M_r\times M_r} & \ldots & 0 & 0 & 0 \\ 0 & (\boldsymbol{S}_0^{(r)})^{\oplus 2} & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & (\boldsymbol{S}_0^{(r)})^{\oplus (N_r-1)} & 0_{M_r^{N_r-1}\times M_r^{N_r-1}} & 0 \\ 0 & 0 & \ldots & 0 & (\boldsymbol{S}_0^{(r)})^{\oplus N_r} & 0_{M_r^{N_r}\times M_r^{N_r}} \end{pmatrix} \otimes I_{K_r}, \tag{7.75}$$

$$r = 1,2,\ldots,R-1,$$

с начальным условием

$$D_0^{(1)} = D_0,\ D_1^{(1)} = D_1,$$

где величина K_r вычисляется как $K_r = (W+1)\prod\limits_{r'=1}^{r-1} \frac{M_r^{N_{r+1}}-1}{M_{r-1}-1}$.

Доказательство. Пусть $r = 1$. Очевидно, что процесс $\{n_t^{(1)}, m_{1,t}^{(1)}, m_{2,t}^{(2)}, \ldots, m_{N_1,t}^{(N_1)}, \nu_t\}$, $t \geq 1$, описывающий работу первой станции, является цепью Маркова. Перенумеруем состояния этой цепи в лексикографическом порядке. Тогда, как легко видеть, переходы цепи, которые не приводят к завершению обслуживания на первой станции (и поступлению запроса на вторую станцию), определяются матрицей $D_0^{(2)}$, которая рассчитывается по формуле (7.74). Переходы цепи, ведущие к завершению обслуживания на первой станции (и поступлению запроса на вторую станцию), определяются матрицей $D_1^{(2)}$. Согласно определению MAP это означает, что процесс поступления запросов на вторую

станцию является MAP-потоком, который определяется матрицами $D_0^{(2)}$ и $D_1^{(2)}$. Дальнейшее доказательство для $r = 2, \ldots, R$ осуществляется по индукции.□

Замечание 7.2. В дальнейшем будем обозначать MAP– входящий поток на r-ю станцию как $MAP^{(r)}$, $r = \overline{1, R}$. Выходящий поток из R-й станции будет обозначать как $MAP^{(R+1)}$. Матрицы $D_0^{(R+1)}$ и $D_1^{(R+1)}$, определяющие $MAP^{(R+1)}$, задаются формулами (7.74) и (7.75) при $r = R + 1$.

Используя результаты теоремы 7.18, можно рассчитать маргинальное стационарное распределение r-й станции тандема как стационарное распределение системы массового обслуживания $MAP^{(r)}/PH^{(r)}/N_r/N_r$, $r = \overline{1, R}$.

Для удобства читателя в следующем разделе кратко описываются алгоритм вычисления стационарного распределения такого типа систем. Для краткости будем опускать индекс r в обозначении матриц, описывающих MAP-поток и PH процесс обслуживания.

7.6.3 Стационарное распределение вероятностей состояний системы $MAP/PH/N/N$

Функционирование системы $MAP/PH/N/N$ описывается цепью Маркова $\zeta_t = \{i_t, \nu_t, m_t^{(1)}, .., m_t^{(i_t)}\}$ где i_t – число занятых приборов, $m_t^{(n)}$, $m_t^{(n)} = \overline{0, M}$, – фаза обслуживания на n-м занятом приборе, ν_t, $\nu_t = \overline{0, W}$, – состояние управляющего процесса MAP в момент времени t.

Перенумеруем состояния цепи в лексикографическом порядке. Тогда инфинитезимальный генератор этой цепи определяется как

$$A = (A_{i,j})_{i,j=\overline{0,N}} =$$

$$\begin{pmatrix}
D_0 & D_1 \otimes \boldsymbol{\beta} & O & ... & O & O \\
I_{\overline{W}} \otimes \boldsymbol{S}_0^{\oplus 1} & D_0 \oplus S^{\oplus 1} & D_1 \otimes I_M \otimes \boldsymbol{\beta} & ... & O & O \\
0 & I_{\overline{W}} \otimes \boldsymbol{S}_0^{\oplus 2} & D_0 \oplus S^{\oplus 2} & ... & O & O \\
... & ... & ... & ... & ... & ... \\
0 & 0 & 0 & ... & D_0 \oplus S^{\oplus N-1} & D_1 \otimes I_{M^{N-1}} \otimes \boldsymbol{\beta} \\
0 & 0 & 0 & ... & I_{\overline{W}} \otimes \boldsymbol{S}_0^{\oplus N} & (D_0 + D_1) \oplus S^{\oplus N}
\end{pmatrix}.$$

Пусть $\mathbf{q}$ является вектором-строкой стационарного распределения вероятностей состояний цепи ζ_t. Этот вектор определяется как единственное

решение системы линейных алгебраических уравнений

$$\mathbf{q}A = \mathbf{0},\ \mathbf{q}\mathbf{e} = 1.$$

В случае большой размерности данной системы для ее решения целесообразно использовать специальные алгоритмы. Наиболее известный численно устойчивый алгоритм предложен в [152]. Он описан ниже.

Представим вектор $\mathbf{q}$ как $\mathbf{q} = (\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_N)$, где вектор $\mathbf{q}_l$ имеет порядок $(W+1)M^l,\ l = 0, \ldots, N$.

Алгоритм 7.1. Векторы $\mathbf{q}_l$, $l = 0, \ldots, N$, вычисляются как

$$\mathbf{q}_l = \mathbf{q}_0 F_l,\ l = \overline{0, N},$$

где матрицы F_l вычисляются рекуррентно по следующим формулам:

$$F_l = (\bar{A}_{0,l} + \sum_{i=1}^{l-1} F_i \bar{A}_{i,l})(-\bar{A}_{l,l})^{-1}, l = \overline{1, N-1},$$

$$F_N = (A_{0,N} + \sum_{i=1}^{N-1} F_i A_{i,N})(-A_{N,N})^{-1},$$

матрицы $\overline{A}_{i,N}$ вычисляются по обратной рекурсии:

$$\overline{A}_{i,N} = A_{i,N}, i = \overline{0, N},$$

$$\overline{A}_{i,l} = A_{i,l} + \overline{A}_{i,l+1} G_l, i = \overline{0, l}, l = N-1, N-2, \ldots, 0,$$

матрицы $G_i, i = \overline{0, N-1}$ вычисляются по обратной рекурсии:

$$G_i = (-A_{i+1,i+1} - \sum_{l=1}^{N-i-1} A_{i+1,i+1+l} G_{i+l} G_{i+l-1} \ldots G_{i+1})^{-1} A_{i+1,i},$$

$$i = N-1, N-2, \ldots, 0,$$

вектор $\mathbf{q}_0$ является единственным решением системы

$$\mathbf{q}_0 \overline{A}_{0,0} = 0,\ \ \mathbf{q}_0 (\sum_{l=1}^{N} F_l \mathbf{e} + \mathbf{e}) = 1.$$

Более подробную информацию об этом алгоритме можно найти в [152]. Обратим внимание, что операции вычитания не присутствуют в данном алгоритме, а все обратные матрицы существуют и неотрицательны. Таким образом, алгоритм является численно устойчивым.

7.6.4 Стационарное распределение тандема и его фрагментов

В данном разделе представляются методы расчета стационарного распределения тандема и его фрагментов на основе результатов исследования выходящих потоков, представленных в теореме 7.18.

Пусть $\langle r, r+1, \ldots, r' \rangle$ обозначает фрагмент тандема, состоящий из r-й, $(r+1)$-й,..., r'-й станций, $1 \leq r \leq r' \leq R$.

Теорема 7.19. Стационарное распределение фрагмента $\langle r, r+1, \ldots, r' \rangle$ тандема может быть рассчитано как стационарное распределение тандема

$$MAP^{(r)}/PH^{(r)}/N_r/N_r \to \bullet/PH^{(r+1)}/N_{r+1}/N_{r+1} \to \ldots \to \bullet/PH^{(r')}/N_{r'}/N_{r'},$$

где $MAP^{(r)}$ определяется формулами (7.74)-(7.75).

Следствие 7.13. Маргинальное стационарное распределение r-й станции тандема вычисляется как стационарное распределение системы массового обслуживания $MAP^{(r)}/PH^{(r)}/N_r/N_r$, $r = \overline{1, R}$.

Заметим, что каждое из этих маргинальных распределений может быть вычислено с использованием алгоритма 7.1, представленного в предыдущем подразделе.

Теорема 7.20. Совместное стационарное распределение $\mathbf{p}^{(1,\ldots,r)}$ вероятностей состояний первых r станций тандема может быть рассчитано как стационарное распределение управляющего процесса $MAP^{(r+1)}$, т.е., имеет место формула

$$\mathbf{p}^{(1,\ldots,r)} = \boldsymbol{\theta}^{(r+1)},$$

где вектор $\boldsymbol{\theta}^{(r+1)}$ – единственное решение системы

$$\boldsymbol{\theta}^{(r+1)}(D_0^{(r+1)} + D_1^{(r+1)}) = \mathbf{0}, \;\; \boldsymbol{\theta}^{(r+1)}\mathbf{e} = 1, \; r = \overline{1, R}, \tag{7.76}$$

а матрицы $D_0^{(r+1)}$, $D_1^{(r+1)}$ вычисляются по рекуррентным формулам (7.74)-(7.75).

Эта система может быть успешно решена с помощью устойчивого алгоритма, разработанного в [152].

Очевидно, что вектор $\mathbf{p}$ стационарного распределения всего тандема совпадает с вектором $\mathbf{p}^{(1,\ldots,R)}$.

Следствие 7.14. Вектор стационарного распределения тандема вычисляется как единственное решение системы линейных алгебраических

уравнений

$$\mathbf{p}(D_0^{(R+1)} + D_1^{(R+1)}) = \mathbf{0}, \ \ \mathbf{pe} = 1. \tag{7.77}$$

Как следует из системы (7.77), матрица $D_0^{(R+1)} + D_1^{(R+1)}$ совпадает с инфинитезимальным генератором Q цепи Маркова $\xi_t, t \geq 0$, описывающей функционирование тандема, т.е.

$$Q = D_0^{(R+1)} + D_1^{(R+1)}.$$

Таким образом, используя результаты анализа потоков, выходящих из станций тандема, мы построили матрицу Q, избежав трудоемкой работы, требующейся при прямом подходе к построению этой матрицы.

7.6.5 Вероятности потерь запросов

Рассчитав стационарные распределения вероятностей состояний тандема и его фрагментов, можно найти ряд важных стационарных характеристик производительности системы. Важнейшими из них являются различные вероятности потерь в тандеме. Согласно эргодической теореме для цепей Маркова, см., например, [169], вероятность потери запроса в фрагменте тандема может быть вычислена как отношение разности интенсивности входного потока в фрагмент и выходящего потока из этого фрагмента к интенсивности входящего потока. Таким образом, справедлива следующая теорема.

Теорема 7.21. Вероятность потери произвольного запроса в фрагменте $\langle r, r+1, \ldots, r' \rangle$ тандема рассчитывается как

$$P_{loss}^{(r,\ldots,r')} = \frac{\lambda_r - \lambda_{r'+1}}{\lambda_r},$$

где λ_l – интенсивность поступления запросов на l-ю станцию в $MAP^{(l)}$-потоке,

$$\lambda_l = \boldsymbol{\theta}^{(l)} D_1^{(l)} \mathbf{e},$$

где вектор $\boldsymbol{\theta}^{(l)}$ – единственное решение системы

$$\boldsymbol{\theta}^{(l)}(D_0^{(l)} + D_1^{(l)}) = \mathbf{0}, \ \ \boldsymbol{\theta}^{(l)}\mathbf{e} = 1, \ l = \overline{1, R}.$$

Следствие 7.15. Вероятность потери произвольного запроса на r-й станции тандема вычисляется как

$$P_{loss}^{(r)} = \frac{\lambda_r - \lambda_{r+1}}{\lambda_r}.$$

Следствие 7.16. Вероятность потери произвольного запроса в тандеме в целом вычисляется по формуле

$$P_{loss} = \frac{\lambda_1 - \lambda_{R+1}}{\lambda_1}.$$

7.6.6 Исследование системы на основе построения цепи Маркова с использованием подхода Рамасвами – Лукантони

Полученные результаты дают простой, элегантный способ для расчета стационарного распределения тандема без промежуточных буферов. Однако этот способ не приводит к уменьшению размера системы линейных алгебраических уравнений, которую необходимо решить для вычисления стационарного распределения некоторых фрагментов и целого тандема. В частности, система (7.77) имеет тот же ранг K_R, что и система (7.73).

Как видно из формул (7.74)-(7.77), размерность систем линейных уравнений, определяющих стационарные распределения фрагментов тандема и всего тандема в целом, существенно зависит от числа N_r обслуживающих приборов на станциях и от размерности M_r пространства состояний управляющего процесса PH обслуживания. При больших значениях N_r и(или) M_r система (7.73) или тождественная ей система (7.77) не поддается аналитическому решению из-за большой размерности. Эта размерность преимущественно зависит от размерностей фазовых пространств процессов $\{m_{1,t}^{(r)}, \ldots, m_{n^{(r)},t}^{(r)}\}$, $r = \overline{1,R}$, описывающих состояния PH процессов обслуживания на приборах станций тандема. Размерность фазового пространства процесса $\{m_{1,t}^{(r)}, \ldots, m_{n^{(r)},t}^{(r)}\}$ равна $M_r^{N_r}$ и может быть очень большой. Чтобы уменьшить данную размерность, мы воспользуемся другим подходом к описанию PH процесса обслуживания на приборах r-й станции тандема. Этот подход был предложен в статьях [20], [167] американских авторов В. Рамасвами и Д. Лукантони и основан на особенностях

функционирования нескольких марковских процессов в параллели. При этом подходе вместо процесса $\{m_{1,t}^{(r)},\dots,m_{n^{(r)},t}^{(r)}\}$ рассматривается процесс $\{\vartheta_{1,t}^{(r)},\dots,\vartheta_{M_r,t}^{(r)}\}$, где $\vartheta_{l,t}^{(r)}$ – число приборов r-й станции, находящихся на l-й фазе обслуживания, $\vartheta_{l,t}^{(r)} = \overline{1,n^{(r)}}$, $r = \overline{1,R}$, $l = \overline{1,M_r}$. Размерность фазового пространства процесса $\{\vartheta_{1,t}^{(r)},\dots,\vartheta_{M_r,t}^{(r)}\}$ равна $\frac{(n^{(r)}+M_r-1)!}{n^{(r)}!(M_r-1)!}$ и может быть гораздо меньше размерности фазового пространства процесса $\{m_{1,t}^{(r)},\dots,m_{n^{(r)},t}^{(r)}\}$. Например, если $N_r = 10$ и $M_r = 2$, то соответствующие размерности равны 2^{10} и 11 соответственно.

Используя подход Рамасвами – Лукантони, мы описываем процесс изменения состояний системы в терминах неприводимой многомерной цепи Маркова с непрерывным временем

$$\tilde{\xi}_t = \{n_t^{(R)},\vartheta_{1,t}^{(R)},\dots,\vartheta_{M_R,t}^{(R)};\ n_t^{(R-1)},\vartheta_{1,t}^{(R-1)},\dots,\vartheta_{M_{R-1},t}^{(R-1)};\ \dots;$$

$$n_t^{(1)},\vartheta_{1,t}^{(1)},\dots,\vartheta_{M_1,t}^{(1)};\ \nu_t\},\ t\geq 0,$$

где

- $n_t^{(r)}$ число занятых приборов на r-й станции, $n_t^{(r)} = \overline{0,N^{(r)}}$, $r = \overline{1,R}$;
- $\vartheta_{l,t}^{(r)}$ – число приборов r-й станции, находящихся на l-й фазе обслуживания, $\vartheta_{l,t}^{(r)} = \overline{1,n^{(r)}}$, $r = \overline{1,R}$, $l = \overline{1,M_r}$;
- ν_t, $\nu_t \in \{0,\dots,W\}$ – состояние управляющего процесса MAP-потока в момент времени t.

В этом случае вектор-строка стационарных вероятностей состояний цепи имеет размерность

$$(W+1)\prod_{r=1}^{R}(1+\sum_{n=1}^{N_r}C_{n+M_r-1}^{M_r-1}).$$

Чтобы сформулировать теорему, аналогичную теореме 7.18, введем в рассмотрение матрицы $P_{n^{(r)}}(\boldsymbol{\beta}^{(r)})$, $A_{n^{(r)}}(N_r,S^{(r)})$ и $L_{N_r-n^{(r)}}(N_r,S^{(r)}$, которые описывают интенсивности переходов процесса $\vartheta_t^{(r)} = \{\vartheta_t^{(r)}(1),\dots,\vartheta_t^{(r)}(M_r)\}$, $r = \overline{1,R}$.

Дадим краткое объяснение значений этих матриц.

Положим $L^{(k)}(j) = C_{j+k-1}^{k-1}$. Тогда

• $L_{N_r-n^{(r)}}(N_r,S^{(r)})$ – матрица порядка $L^{(M_r)}(n^{(r)})\times L^{(M_r)}(n^{(r)}-1)$, которая содержит интенсивности переходов процесса $\vartheta_t^{(r)}$, приводящих к освобождению одного из $n^{(r)}$ занятых приборов;

• $P_{n^{(r)}}(\boldsymbol{\beta}^{(r)})$ – матрица порядка $L^{(M_r)}(n^{(r)}) \times L^{(M_r)}(n^{(r)}+1)$, которая содержит интенсивности переходов процесса $\vartheta_t^{(r)}$, приводящих к увеличению числа занятых приборов с $n^{(r)}$ до $n^{(r)}+1$;

• $A_{n^{(r)}}(N_r, S^{(r)})$ – матрица порядка $L^{(M_r)}(n^{(r)}) \times L^{(M_r)}(n^{(r)})$, которая содержит интенсивности переходов процесса $\vartheta_t^{(r)}$ в его пространстве состояний без увеличения или уменьшения числа $n^{(r)}$ занятых приборов.

В дальнейшем мы предполагаем, что $L_{N_r}(N_r, S^{(r)}) = A_0(N_r, S^{(r)}) = 0_{1\times 1}$. Остальные матрицы $P_{n^{(r)}}(\boldsymbol{\beta}^{(r)})$, $A_{n^{(r)}}(N_r, S^{(r)})$ и $L_{N_r-n^{(r)}}(N_r, S^{(r)}$, вычисляются по алгоритму, приведенному, например, в [119].

Теорема 7.22. *Выходящий поток из r-й станции (входящий поток на $(r+1)$-ю станцию), $r \in \{1, 2, \ldots, R-1\}$, принадлежит классу MAP-потоков. Этот MAP-поток задается матрицами $D_0^{(r+1)}$ и $D_1^{(r+1)}$, которые вычисляются по следующим рекуррентным формулам:*

$$D_0^{(r+1)} = diag\{A_n(N_r, S^{(r)}),\ n = \overline{0, N_r}\} \otimes I_{K_r} +$$

$$+ \begin{pmatrix} D_0^{(r)} & D_1^{(r)} \otimes P_0(\beta^{(r)}) & 0 & \ldots & 0 & 0 \\ 0 & I_{C_{M_r}^{M_r-1}} \otimes D_0^{(r)} & I_{C_{M_r}^{M_r-1}} \otimes D_1^{(r)} \otimes P_1(\beta^{(r)}) & \ldots & 0 & 0 \\ 0 & 0 & I_{C_{M_r+1}^{M_r-1}} \otimes D_0^{(r)} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & I_{C_{N_r+M_r-2}^{M_r-1}} \otimes D_0^{(r)} & I_{C_{N_r+M_r-2}^{M_r-1}} \otimes D_1^{(r)} \otimes P_{N_r-1}(\beta^{(r)}) \\ 0 & 0 & 0 & \ldots & 0 & I_{C_{N_r+M_r-1}^{M_r-1}} \otimes (D_0^{(r)} + D_1^{(r)}) \end{pmatrix},$$

$$D_1^{(r+1)} =$$

$$\begin{pmatrix} 0_1 & 0 & \ldots & 0 & 0 & 0 \\ L_{N_r-1}(1, S^{(r)}) & 0_{C_{M_r}^{M_r-1}} & \ldots & 0 & 0 & 0 \\ 0 & L_{N_r-2}(2, S^{(r)}) & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & L_1(N_r-1, S^{(r)}) & 0_{C_{N_r+M_r-2}^{M_r-1}} & 0 \\ 0 & 0 & \ldots & 0 & L_0(N_r, S^{(r)}) & 0_{C_{N_r+M_r-1}^{M_r-1}} \end{pmatrix} \otimes I_{K_r},$$

$$r = \overline{1, R-1},$$

с начальным условием

$$D_0^{(1)} = D_0,\ \ D_1^{(1)} = D_1,$$

где величина K_r вычисляется как $K_r = (W+1)\prod_{l=1}^{r-1}(1 + \sum_{n=1}^{N_l} C_{n+M_l-1}^{M_l-1})$.

Теорема 7.22 дает возможность снизить порядок матриц $D_0^{(r+1)}, D_1^{(r+1)}$, $r \in \{1, 2, \ldots, R-1\}$, и использовать их для вычисления стационарных

распределений фрагментов тандема и всего тандема, а также для вычисления всевозможных вероятностей потерь по формулам, приведенным ранее в данном разделе.

Если даже метод Рамасвами – Лукантони построения генератора и приведенный выше алгоритм вычисления стационарного распеделения, эффективно использующий разреженную структуру генератора, не дают возможности решить системы линейных алгебраических уравнений для стационарного распределения фрагментов тандема и тандема в целом, то могут быть использованы своего рода эвристики. Например, когда рекуррентно вычисляются матрицы $D_0^{(r)}$ и $D_1^{(r)}$, определяющие MAP-поток на r-ю станцию, и порядок этих матриц становится слишком большим для некоторого $r \in \{1, \ldots, R-1\}$, то можно попытаться аппроксимировать этот MAP MAP-потоком с той же интенсивностью, некоторым количеством совпадающих моментов длин интервалов между поступлениями запросов (по крайней мере, с тем же коэффициентом вариации) и тем же коэффициентом корреляции, но с гораздо меньшей размерностью пространства состояний управляющего процесса, см., например, [92,133,135]. После этого можно продолжать рекуррентную процедуру расчета матриц $D_0^{(r')}$ и $D_1^{(r')}$, $r' > r$, исходя из матриц, которые определяют этот аппроксимирующий MAP.

7.7 Многофазный тандем многолинейных СМО без буферов с кросс-трафиком

Рассматриваемая в данном разделе тандемная система массового обслуживания является более общей по сравнению с системой, исследованной в предыдущем разделе, так как в ней учитывается наличие коррелированного кросс-трафика, который поступает на каждую станцию тандема дополнительно с трафиком, передающимся из предыдущей станции тандема.

7.7.1 Описание системы

Рассматривается тандемная система массового обслуживания, состоящая из R, $R > 1$, станций. В терминах обозначений Кендалла эта система может быть описана как

$$MAP^{(1)}/PH^{(1)}/N_1/N_1 \to \bullet, MAP_2/PH^{(2)}/N_2/N_2 \to \cdots \to \bullet, MAP_R/PH^{(R)}/N_R/N_R.$$

Станция номер r, $r = 1, \ldots, R$, представлена системой из N_r приборов без буфера. Приборы, принадлежащие одной и той же станции, являются независимыми и идентичными. На вход первой станции поступает MAP-поток, который обозначим как $MAP^{(1)}$. Он характеризуется пространством состояний $\{0, 1, ..., W_1\}$ управляющего процесса $\nu_t^{(1)}, t \geq 0$, и матрицами D_0, D_1. Целью каждого запроса из этого потока является получение обслуживания на всех станциях тандема.

Кроме потока запросов, поступающих на r-ю, $r > 1$, станцию из $(r-1)$-й станции, на нее поступает дополнительный MAP-поток запросов, который обозначим как MAP_r. Этот поток характеризуется пространством состояний $\{0, 1, ..., W_r\}$ управляющего процесса $\nu_t^{(r)}, t \geq 0$, и матрицами $H_0^{(r)}$, $H_1^{(r)}$. Интенсивность поступления запросов в MAP_r-потоке обозначим как h_r. Целью каждого запроса из этого потока является получение обслуживания на r-й и всех последующих станциях тандема.

Время обслуживания запроса на r-й станции имеет PH-распределение с неприводимым представлением $(\boldsymbol{\beta}^{(r)}, S^{(r)}), r = \overline{1, R}$.

Если запрос, поступающий на станцию тандема из предыдущей станции или из дополнительного потока, застает все приборы занятыми, то он покидает тандем навсегда.

7.7.2 Потоки, выходящие из станций, и потоки, входящие на станции тандема. Стационарное распределение тандема и его фрагментов

Процесс изменения состояний системы описывается в терминах неприводимой многомерной цепи Маркова с непрерывным временем

$$\xi_t = \{n_t^{(R)}, m_{1,t}^{(R)}, \ldots, m_{n^{(R)},\,t}^{(R)}, \nu_t^{(R)};\ n_t^{(R-1)}, m_{1,t}^{(R-1)}, \ldots, m_{n^{(R-1)},\,t}^{(R-1)}, \nu_t^{(R-1)};$$

$$\ldots, n_t^{(1)}, m_{1,t}^{(1)}, \ldots, m_{n^{(1)},\,t}^{(1)}, \nu_t^{(1)}\},\ t \geq 0,$$

где

- $n_t^{(r)}$ число занятых приборов на r-й станции, $n_t^{(r)} \in \{0, \ldots, N^{(r)}\},\ r \in \{1, \ldots, R\}$;
- $m_{l,t}^{(r)}$ – фаза PH – процесса обслуживания на l-м занятом приборе r-й станции, $m_{l,t}^{(r)} = \overline{1, M_r}$, $r = \overline{1, R}$, $l = \overline{1, n^{(r)}}$;
- $\nu_t^{(r)}$, $\nu_t^{(r)} \in \{0, \ldots, W_r\}$ – состояние управляющего процесса MAP_r-потока

  в момент времени t.

Пространство состояний рассматриваемой цепи имеет размерность

$$\prod_{r=1}^{R} \frac{(W_r+1)(M_r^{N_r+1}-1)}{M_r-1}.$$

Теорема 7.23. *Выходящий поток из r-й станции $r \in \{1,2,\ldots,R\}$, принадлежит классу MAP-потоков. Этот MAP-поток задается матрицами $D_0^{(r)}$ и $D_1^{(r)}$, которые вычисляются по следующим рекуррентным формулам:*

$$D_0^{(r)} = diag\{(S^{(r)})^{\oplus n},\, n=\overline{0,N_r}\} \otimes I_{K_r} + \tag{7.78}$$

$$+\begin{pmatrix} \tilde{D}_0^{(r)} & \boldsymbol{\beta}^{(r)} \otimes \tilde{D}_1^{(r)} & 0 & \ldots & 0 & 0 \\ 0 & I_{M_r} \otimes \tilde{D}_0^{(r)} & I_{M_r} \otimes \boldsymbol{\beta}^{(r)} \otimes \tilde{D}_1^{(r)} & \ldots & 0 & 0 \\ 0 & 0 & I_{M_r^2} \otimes \tilde{D}_0^{(r)} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & I_{M_r^{N_r-1}} \otimes \tilde{D}_0^{(r)} & I_{M_r^{N_r-1}} \otimes \boldsymbol{\beta}^{(r)} \otimes \tilde{D}_1^{(r)} \\ 0 & 0 & 0 & \ldots & 0 & I_{M_r^{N_r}} \otimes (\tilde{D}_0^{(r)} + \tilde{D}_1^{(r)}) \end{pmatrix},$$

$$D_1^{(r)} = \tag{7.79}$$

$$= \begin{pmatrix} 0_{1\times 1} & 0 & \ldots & 0 & 0 & 0 \\ (\boldsymbol{S}_0^{(r)})^{\oplus 1} & 0_{M_r\times M_r} & \ldots & 0 & 0 & 0 \\ 0 & (\boldsymbol{S}_0^{(r)})^{\oplus 2} & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & (\boldsymbol{S}_0^{(r)})^{\oplus (N_r-1)} & 0_{M_r^{N_r-1}\times M_r^{N_r-1}} & 0 \\ 0 & 0 & \ldots & 0 & (\boldsymbol{S}_0^{(r)})^{\oplus N_r} & 0_{M_r^{N_r}\times M_r^{N_r}} \end{pmatrix} \otimes I_{K_r},$$

$$r = 1,2,\ldots,R,$$

где

$$\tilde{D}_0^{(r)} = D_0^{(r-1)} \oplus H_0^{(r)},\ \ \tilde{D}_1^{(r)} = D_1^{(r-1)} \oplus H_1^{(r)}, r = \overline{2,R},$$

начальные условия устанавливаются как

$$\tilde{D}_0^{(1)} = D_0,\ \ \tilde{D}_1^{(1)} = D_1,$$

а значение K_r вычисляется как

$$K_r = (W_r+1)\prod_{l=1}^{r-1} \frac{(W_l+1)(M_l^{N_l+1}-1)}{M_l-1},\ r = 1,2,\ldots,R.$$

Доказательство. Доказательство аналогично доказательству теоремы 7.18 для аналогичного тандема без кросс-трафика. Единственным отличием является то, что в данном случае выходящий поток из $(r-1)$-й станции не совпадает с входящим потоком в r-ю станцию. Последний является суперпозицией выходящего потока из $(r-1)$-й станции и дополнительного MAP-потока на r-ю станцию. Эта суперпозиция снова является MAP-потоком, который определяется матрицами $\tilde{D}_0^{(r)}, \tilde{D}_1^{(r)}$. □

Замечание 7.3. В дальнейшем будем обозначать MAP– входящий поток на r-ю станцию как $MAP^{(r)}$, $r = 1, 2, \ldots, R$.

Следствие 7.17. *Входящий поток на r-ю станцию, $MAP^{(r)}$, определяется матрицами*

$$\tilde{D}_0^{(r)} = D_0^{(r-1)} \oplus H_0^{(r)}, \;\; \tilde{D}_1^{(r)} = D_1^{(r-1)} \oplus H_1^{(r)}, r = \overline{2, R}. \tag{7.80}$$

Используя результаты теоремы 7.23 и следствия 7.15, можно рассчитать маргинальное стационарное распределение r-й станции тандема как стационарное распределение системы массового обслуживания $MAP^{(r)}/PH^{(r)}/N_r/N_r$, $r = 1, 2, \ldots, R$. Соответствующий алгоритм приведен в подразделе 7.7.3.

Используя уже изложенные результаты, также можно рассчитать стационарное распределение всего тандема и его фрагментов, а также вероятности потерь, ассоциированных с тандемом. При этом справедливы утверждения, аналогичные приведенным в разделе 7.7 утверждениям для тандема без кросс-трафика. Для удобства читателя приводим эти утверждения ниже.

Теорема 7.24. *Стационарное распределение фрагмента $\langle r, r+1, \ldots, r' \rangle$ рассматриваемого тандема может быть вычислено как стационарное распределение тандема*

$$MAP^{(r)}/PH^{(r)}/N_r/N_r \to \bullet, MAP_{r+1}/PH^{(r+1)}/N_{r+1}/N_{r+1} \to \ldots \to \bullet, MAP_{r'}/PH^{(r')}/N_{r'}/N_{r'},$$

где $MAP^{(r)}$ определяется матрицами (7.80).

Теорема 7.25. *Совместное стационарное распределение $\mathbf{p}^{(1,\ldots,r)}$ вероятностей состояний первых r станций тандема может быть вычислено как стационарное распределение управляющего процесса MAP-потока, выходящего из r-й станции, т.е. имеет место формула*

$$\mathbf{p}^{(1,\ldots,r)} = \boldsymbol{\theta}^{(r)},$$

где вектор $\boldsymbol{\theta}^{(r)}$ *– единственное решение системы*

$$\boldsymbol{\theta}^{(r)}(D_0^{(r)} + D_1^{(r)}) = \mathbf{0}, \;\; \boldsymbol{\theta}^{(r)}\mathbf{e} = 1, \; r = 1, 2, \ldots, R,$$

а матрицы $D_0^{(r)}$, $D_1^{(r)}$ *вычисляются по рекуррентным формулам (7.78)-(7.79).*

Следствие 7.18. *Вектор стационарного распределения всего тандема вычисляется как единственное решение системы линейных алгебраических уравнений*

$$\mathbf{p}(D_0^{(R)} + D_1^{(R)}) = \mathbf{0}, \;\; \mathbf{pe} = 1.$$

Матрица $D_0^{(R)} + D_1^{(R)}$ *совпадает с инфинитезимальным генератором* Q *цепи Маркова* $\xi_t, t \geq 0$.

7.7.3 Вероятности потерь

Теорема 7.26. *Вероятность потери запроса в фрагменте* $\langle r, r+1, \ldots, r' \rangle$ *тандема рассчитывается как*

$$P_{loss}^{(r,\ldots,r')} = \frac{\tilde{\lambda}_r - \lambda_{r'}}{\tilde{\lambda}_r},$$

где $\tilde{\lambda}_r$ *– интенсивность входящего* $MAP^{(r)}$ *потока на* r*-ю станцию,* λ_r *– интенсивность выходящего потока из* r'*-й станции.*

Интенсивность $\tilde{\lambda}_r$ *вычисляется как*

$$\tilde{\lambda}_r = \tilde{\boldsymbol{\theta}}^{(r)} \tilde{D}_1^{(r)} \mathbf{e},$$

где вектор $\tilde{\boldsymbol{\theta}}^{(r)}$ *– единственное решение системы*

$$\tilde{\boldsymbol{\theta}}^{(r)}(\tilde{D}_0^{(r)} + \tilde{D}_1^{(r)}) = \mathbf{0}, \;\; \tilde{\boldsymbol{\theta}}^{(r)}\mathbf{e} = 1,$$

интенсивность λ_r *вычисляется по формуле*

$$\lambda_r = \boldsymbol{\theta}^{(r)} D_1^{(r)} \mathbf{e},$$

где вектор $\boldsymbol{\theta}^{(r)}$ *– единственное решение системы*

$$\boldsymbol{\theta}^{(r)}(D_0^{(r)} + D_1^{(r)}) = \mathbf{0}, \;\; \boldsymbol{\theta}^{(r)}\mathbf{e} = 1, \; r = 1, 2, \ldots, R.$$

Следствие 7.19. *Вероятность потери* $P_{loss}^{(r)}$ *произвольного запроса на* r*-й станции вычисляется по формуле*

$$P_{loss}^{(r)} = \frac{\tilde{\lambda}_r - \lambda_r}{\tilde{\lambda}_r}. \tag{7.81}$$

Следствие 7.20. *Вероятность потери P_{loss} произвольного запроса в тандеме в целом вычисляется по формуле*

$$P_{loss} = \frac{\tilde{\lambda}_1 - \lambda_R}{\tilde{\lambda}_1}.$$

Теорема 7.27. *Вероятность потери произвольного запроса на 1-й станции тандема вычисляется как*

$$P_{loss/1}^{(1)} = \frac{\boldsymbol{\theta} D_1 \mathbf{e}}{\lambda},$$

Вероятность потери на r-й станции тандема произвольного запроса, поступившего с $(r-1)$-й станции, вычисляется как

$$P_{loss/1}^{(r)} = \frac{\boldsymbol{\theta}^{(r)} diag\{O_{\frac{M_r^{N_r}-1}{M_r-1}K_r},\ I_{M_r^{N_r}} \otimes D_1^{(r-1)} \otimes I_{W_r+1}\}\mathbf{e}}{\lambda_{r-1}},\ r = \overline{2,R}. \tag{7.82}$$

Вероятность потери на r-й станции тандема произвольного запроса из дополнительного MAP-потока вычисляется по формуле

$$P_{loss/2}^{(r)} = \frac{\boldsymbol{\theta}^{(r)} diag\{O_{\frac{M_r^{N_r}-1}{M_r-1}K_r},\ I_{M_r^{N_r}} \otimes I_{\frac{K_r}{W_r+1}} \otimes H_1^{(r)}\}\mathbf{e}}{h_r},\ r = \overline{2,R}. \tag{7.83}$$

Доказательство. Числитель правой части (7.82) есть интенсивность потока запросов, поступающих на r-ю станцию из $(r-1)$-й станции и застающих все приборы r-й станции занятыми, а знаменатель – интенсивность потока всех запросов, поступающих на r-ю станцию из $(r-1)$-й станции. По известным эргодическим соображениям, отношение этих двух интенсивностей определяет искомую вероятность $P_{loss/1}^{(r)}$. Аналогичные рассуждения приводят к формуле (7.83) для вероятности $P_{loss/2}^{(r)}$. □

Используя теорему 7.27, можно получить альтернативное выражение для вероятности суммарных потерь $P_{loss}^{(r)}$ на r-й станции, ранее определенной формулой (7.91) вследствии 4. Альтернативная формула (7.84), а также формулы для некоторых совместных вероятностей, ассоциированных с тандемом, приведены в следующем утверждении.

Следствие 7.21. *Вероятность того, что произвольный запрос из суммарного потока на r-ю станцию принадлежит выходящему потоку из $(r-1)$-й станции и будет потерян из-за отсутствия свободных приборов на r-й станции, вычисляется как*

$$P_{loss,1}^{(r)} = P_{loss/1}^{(r)} \frac{\lambda_{r-1}}{\tilde{\lambda}_r},\ r = 2, \ldots, R.$$

Вероятность того, что произвольный запрос из суммарного потока на r-ю станцию принадлежит дополнительному потоку и будет потерян из-за отсутствия свободных приборов на станции, вычисляется следующим образом:

$$P_{loss,2}^{(r)} = P_{loss/2}^{(r)} \frac{h_r}{\tilde{\lambda}_r},\ r = 2, \ldots, R.$$

Вероятность $P_{loss}^{(r)}$ потери произвольного запроса на r-й станции может быть вычислена как

$$P_{loss}^{(r)} = P_{loss,1}^{(r)} + P_{loss,2}^{(r)}. \tag{7.84}$$

Следствие 7.22. *В случае $R = 2$ вероятность того, что запрос из входящего потока на первую станцию не будет потерян в системе, определяется как*

$$P_{succ} = (1 - P_{loss/1}^{(1)})(1 - P_{loss/1}^{(2)}).$$

7.7.4 Исследование системы на основе построения цепи Маркова с использованием подхода Рамасвами – Лукантони

Как видно из формул (7.78)-(7.79), теоремы 7.25 и следствия 7.18, размерность систем линейных уравнений, определяющих стационарные распределения фрагментов тандема и всего тандема в целом, существенно зависит от числа N_r обслуживающих приборов на станциях и от размерности M_r пространства состояний управляющего процесса PH-обслуживания. Используя те же доводы, что и в подразделе 7.7.6, приходим к выводу, что размерность фазового пространства процесса, описывающего PH-процесс

обслуживания в системе, а значит, и размерность фазового пространства цепей, описывающих фрагменты тандема и тандем в целом, может быть существенно снижена, если при построении этих цепей использовать подход, предложенный в [20], [167] американскими авторами В. Рамасвами и Д. Лукантони.

Используя подход Рамасвами – Лукантони, мы описываем процесс изменения состояний системы в терминах неприводимой многомерной цепи Маркова с непрерывным временем

$$\tilde{\xi}_t = \{n_t^{(R)}, \vartheta_{1,t}^{(R)}, \ldots, \vartheta_{M_R,t}^{(R)}, \nu_t^{(R)};\ n_t^{(R-1)}, \vartheta_{1,t}^{(R-1)}, \ldots, \vartheta_{M_{R-1},t}^{(R-1)}, \nu_t^{(R-1)}; \ldots;$$

$$n_t^{(1)}, \vartheta_{1,t}^{(1)}, \ldots, \vartheta_{M_1,t}^{(1)};\ \nu_t\},\ t \geq 0,$$

где

- $n_t^{(r)}$ число занятых приборов на r-й станции, $n_t^{(r)} \in \{0, \ldots, N^{(r)}\},\ r \in \{1, \ldots, R\}$;
- $\vartheta_{l,t}^{(r)}$ – число приборов r-й станции, находящихся на l-й фазе обслуживания, $\vartheta_{l,t}^{(r)} = \overline{1, n^{(r)}}$, $r = \overline{1, R}$, $l = \overline{1, M_r}$;
- $\nu_t^{(r)}$, $\nu_t^{(r)} \in \{0, \ldots, W_r\}$ – состояние управляющего процесса дополнительного MAP_r-потока
  в момент времени t.

Фазовое пространство состояний такой цепи имеет размерность

$$\prod_{r=1}^{R}(W_r + 1)\sum_{n=0}^{N_r} C_{n+M_r-1}^{M_r-1}$$

и при больших значениях N_r, M_r значительно меньше размерности фазового пространства цепи ξ_t, построенной ранее на основе классического подхода.

Теорема 7.28. Выходящий поток из r-й станции $r \in \{1, 2, \ldots, R\}$, принадлежит классу MAP-потоков. Этот MAP-поток задается матрицами $D_0^{(r)}$ и $D_1^{(r)}$, которые вычисляются по следующим рекуррентным формулам:

$$D_0^{(r)} = diag\{A_n(N_r, S^{(r)}),\ n = \overline{0, N_r}\} \otimes I_{K_r} + \qquad (7.85)$$

$$+\begin{pmatrix} \tilde{D}_0^{(r)} & \tilde{D}_1^{(r)} \otimes P_0(\beta^{(r)}) & 0 & \dots & 0 & 0 \\ 0 & I_{C_{M_r}^{M_r-1}} \otimes \tilde{D}_0^{(r)} & I_{C_{M_r}^{M_r-1}} \otimes \tilde{D}_1^{(r)} \otimes P_1(\beta^{(r)}) & \dots & 0 & 0 \\ 0 & 0 & I_{C_{M_r+1}^{M_r-1}} \otimes \tilde{D}_0^{(r)} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & I_{C_{N_r+M_r-2}^{M_r-1}} \otimes \tilde{D}_0^{(r)} & I_{C_{N_r+M_r-2}^{M_r-1}} \otimes \tilde{D}_1^{(r)} \otimes P_{N_r-1}(\beta^{(r)}) \\ 0 & 0 & 0 & \dots & 0 & I_{C_{N_r+M_r-1}^{M_r-1}} \otimes (\tilde{D}_0^{(r)} + \tilde{D}_1^{(r)}) \end{pmatrix},$$

$$D_1^{(r)} = \tag{7.86}$$

$$= \begin{pmatrix} 0_1 & 0 & \dots & 0 & 0 & 0 \\ L_{N_r-1}(N_r, S^{(r)}) & 0_{C_{M_r}^{M_r-1}} & \dots & 0 & 0 & 0 \\ 0 & L_{N_r-2}(N_r, S^{(r)}) & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & L_1(N_r, S^{(r)}) & 0_{C_{N_r+M_r-2}^{M_r-1}} & 0 \\ 0 & 0 & \dots & 0 & L_0(N_r, S^{(r)}) & 0_{C_{N_r+M_r-1}^{M_r-1}} \end{pmatrix} \otimes I_{K_r},$$

$$r = 1, 2, \dots, R-1,$$

где

$$\tilde{D}_0^{(r)} = D_0^{(r-1)} \oplus H_0^{(r)}, \;\; \tilde{D}_1^{(r)} = D_1^{(r-1)} \oplus H_1^{(r)}, r = \overline{2, R},$$
$$\tilde{D}_0^{(1)} = D_0, \;\; \tilde{D}_1^{(1)} = D_1,$$

величина K_r вычисляется как

$$K_r = \prod_{l=1}^{r-1}(W_l + 1) \sum_{n=0}^{N_l} C_{n+M_l-1}^{M_l-1}, \; r = 1, 2, \dots, R,$$

а матрицы $P_{n^{(r)}}(\boldsymbol{\beta}^{(r)})$, $A_{n^{(r)}}(N_r, S^{(r)})$, $L_{N_r-n^{(r)}}(N_r, S^{(r)}$ описаны в подразделе 7.7.6.

Используя выражения (7.85)-(7.86) для матриц $D_0^{(r)}, D_1^{(r)}$, можем вычислить стационарное распределение тандема и его фрагментов, а также некоторые вероятности потерь в тандеме по формулам, приведенным выше, в теоремах 7.24-7.26 и следствиях 7.17-7.20. Формулы для условных вероятностей потерь, приведенные в теореме 7.27, модифицируются вследствие того, что размерности фазовых пространств при использовании подхода Рамасвами – Лукантони отличаются от таковых при использовании классического подхода к построению цепей Маркова. Модифицированные формулы приведены ниже, в утверждении теоремы 7.29.

Теорема 7.29. Вероятность потери произвольного запроса на 1-й станции тандема вычисляется как

$$P_{loss/1}^{(1)} = \frac{\boldsymbol{\theta} diag\{O_{\sum_{l=0}^{N_1-1} C_{l+M_l-1}^{M_l-1}(W_1+1)}, \; I_{C_{N_1+M_1-1}^{M_1-1}} \otimes D_1\}\mathbf{e}}{\lambda},$$

Вероятность потери на r-й станции тандема произвольного запроса, поступившего с $(r-1)$-й станции, вычисляется как

$$P_{loss/1}^{(r)} = \frac{\boldsymbol{\theta}^{(r)} diag\{O_{\sum_{l=0}^{N_r-1} C_{l+M_l-1}^{M_l-1} K_r},\ I_{C_{N_r+M_r-1}^{M_r-1}} \otimes D_1^{(r-1)} \otimes I_{W_r+1}\}\mathbf{e}}{\lambda_{r-1}},\ r = \overline{2,R}.$$

Вероятность потери на r-й станции тандема произвольного запроса из дополнительного потока вычисляется по формуле

$$P_{loss/2}^{(r)} = \frac{\boldsymbol{\theta}^{(r)} diag\{O_{\sum_{l=0}^{N_r-1} C_{l+M_l-1}^{M_l-1} K_r},\ I_{C_{N_r+M_r-1}^{M_r-1}} \otimes I_{\frac{K_r}{W_r+1}} \otimes H_1^{(r)}\}\mathbf{e}}{h_r},\ r = \overline{2,R}.$$

7.8 Многофазные тандемы однолинейных СМО с конечными буферами и кросс-трафиком

В данном разделе рассматривается тандемная система, состоящая из произвольного конечного числа фаз (станций), представленных однолинейными системами с конечными буферами, на первую фазу которой поступает MAP-поток запросов, каждый из которых должен получить последовательное обслуживание на всех станциях тандема. Кроме того, на каждую из станций поступает дополнительный MAP-поток запросов, которые должны обслужиться на этой станции и на всех последующих станциях тандема. В силу ограниченности мест для ожидания на станциях, запросы из любого потока могут быть потеряны. Для обнаружения и предотвращения так называемых узких мест в тандемной сети важно определить вероятности потерь запросов, поступающих на каждую из станций тандема, а также стационарное распределение вероятностей числа запросов и времени пребывания запроса на каждой фазе и в тандеме в целом.

7.8.1 Описание системы

Рассматривается тандемная система массового обслуживания, состоящая из $R,\ R > 1$, станций. В терминах обозначений Кендалла эта система может быть описана как

$$MAP_1/PH/1/N_1 \to \cdot, MAP_2/PH/1/N_2 \to \ldots \to \cdot, MAP_R/PH/1/N_R.$$

Станция номер r, $r = 1, \ldots, R$, представлена однолинейной системой массового обслуживания с буфером емкости $N_r - 1$.

На вход первой станции поступает марковский поток запросов, который мы обозначим как MAP_1. Кроме потока запросов, поступающих на r-ю, $r > 1$, станцию из $(r-1)$-й станции, на нее поступает дополнительный MAP-поток запросов, который мы обозначим как MAP_r.

Обозначим матрицы, задающие MAP_1, как D_0, D_1, а аналогичные матрицы, задающие MAP_r, как $H_0^{(r)}$, $H_1^{(r)}$. Интенсивность поступления запросов в MAP_1 обозначим как $\tilde{\lambda}_1$, интенсивность поступления запросов в MAP_r, $r > 1$, – как h_r.

Времена обслуживания запросов на r-й станции тандема имеет PH-распределение, заданное неприводимым представлением $(\boldsymbol{\beta}^{(r)}, S^{(r)})$, где $\boldsymbol{\beta}^{(r)}$ и $S^{(r)}$ – вектор и матрица порядка M_r.

Если запрос, поступающий на r-ю станцию тандема, застает прибор занятым, то он помещается в буфер, если в нем есть свободное место. В противном случае запрос покидает тандем навсегда (теряется). Дисциплина выбора запросов на обслуживание из буфера: первым пришел – первым обслужен.

7.8.2 Выходящие потоки из станций и входящие потоки на станции тандема

Процесс изменения состояний тандемной системы описывается в терминах неприводимой многомерной цепи Маркова с непрерывным временем

$$\xi_t = \{n_t^{(R)}, \eta_t^{(R)}, \nu_t^{(R)}, n_t^{(R-1)}, \eta_t^{(R-1)}, \nu_t^{(R-1)}, \ldots, n_t^{(1)}, \eta_t^{(1)}, \nu_t^{(1)}\},\ t \geq 0,$$

где

- $n_t^{(r)}$ – число запросов на r-й станции, $n_t^{(r)} \in \{0, \ldots, N_r\}$,
- $\eta_t^{(r)}$ – состояние процесса обслуживания в приборе r-й станции, $\eta_t^{(r)} \in \{1, \ldots, M_r\}$,
- $\nu_t^{(r)}$ – состояние управляющего процесса MAP_r – потока, $\nu_t^{(r)} \in \{0, \ldots, W_r\}$,

  $r \in \{1, \ldots, R\}$, в момент времени t.

Заметим, что во избежание необходимости отдельного рассмотрения состояний, когда некоторые буфера являются пустыми и прибор простаивает, будем предполагать, что в момент окончания обслуживания на станции с номером r тандема состояние управляющего процесса обслуживания

в этом приборе для следующего запроса устанавливается в соответствии с вероятностями, заданными компонентами вектора $\boldsymbol{\beta}^{(r)}$, независимо от того, имеются ли запросы в буфере этой станции. Если буфер пуст, то состояния управляющего процесса обслуживания начнут изменяться только после начала обслуживания поступаюшего на этот прибор следующего запроса.

Пространство состояний цепи Маркова $\xi_t, t \geq 0$, имеет размерность

$$\prod_{r=1}^{R}(N_r+1)M_r(W_r+1).$$

В данном исследовании мы развиваем простой, точный и удобный метод вычисления маргинальных стационарных распределений вероятностей фрагментов тандема, а также всего тандема и соответствующих вероятностей потерь, предложенный в [151], а также находим распределение времени пребывания запроса в тандеме.

Как и для описанных в двух предыдущих разделах многофазных тандемах многолинейных СМО, исследование рассматриваемого в даннном раделе тандема начинаем с анализа выходящих и входящих потоков на станциях тандема, в результате которого доказано, что такие потоки принадлежат классу MAP-потоков. Соответствующие результаты сформулированы в следующей теореме и ее следствии.

Теорема 7.30. *Выходящий поток из r-й станции тандема, $r \in \{1,2,\ldots,R\}$, принадлежит классу MAP-потоков. Этот MAP-поток задается матрицами $D_0^{(r)}$ и $D_1^{(r)}$, которые вычисляются по следующим рекуррентным формулам:*

$$D_0^{(r)} = diag\{O, S^{(r)}, \ldots, S^{(r)}\} \oplus (D_0^{(r-1)} \oplus H_0^{(r)}) + \tilde{E}_r^{+} \otimes (D_1^{(r-1)} \oplus H_1^{(r)}), \quad (7.87)$$

$$D_1^{(r)} = E_r^{-} \otimes \mathbf{S}_0^{(r)} \boldsymbol{\beta}^{(r)} \otimes I_{K_r},\ r = 1,2,\ldots,R, \quad (7.88)$$

с начальными условиями

$$D_0^{(0)} = D_0,\ D_1^{(0)} = D_1,\ H_0^{(1)} = H_1^{(1)} = O.$$

Здесь

$\tilde{E}_r^+ = E_r^+ \otimes I_{M_r}$, *где* E_r^+ *и* E_r^- *– квадратные матрицы размера* $N_r + 1$, *задаваемые формулами*

$$E_r^+ = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \; E_r^- = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix},$$

а величины K_r *вычисляются по формуле*

$$K_r = \prod_{r'=1}^{r-1}(N_{r'}+1)M_{r'}\prod_{r'=1}^{r}(W_{r'}+1).$$

Доказательство проводится аналогично доказательству теоремы 7.23.

Следствие 7.23. *Входящий поток на* r*-ю станцию,* $r \in \{2, \dots, R\}$, *тандема принадлежит классу* MAP*-потоков. Этот* MAP*-поток задается матрицами*

$$\tilde{D}_0^{(r)} = D_0^{(r-1)} \oplus H_0^{(r)}, \quad \tilde{D}_1^{(r)} = D_1^{(r-1)} \oplus H_1^{(r)}, \quad r \in \{2, \dots, R\}, \qquad (7.89)$$

где матрицы $D_0^{(r)}$, $D_1^{(r)}$ *вычисляются по рекуррентным формулам (7.87), (7.88).*

Замечание 7.4. Положив в формуле (7.89) $r = 1$, получим естественные соотношения

$$\tilde{D}_k^{(1)} = D_k, \; k = 0, 1.$$

Замечание 7.5. В дальнейшем мы будем обозначать суммарный входящий MAP– поток на r-ю станцию как $MAP^{(r)}$, $r = 1, 2, \dots, R$. Заметим, что обозначение $MAP^{(1)}$ означает то же, что и ранее введенное обозначение MAP_1. Оба этих обозначения используются для входящего потока на первую станцию.

Используя результаты теоремы 7.30, можно рассчитать маргинальное стационарное распределение r-й станции тандема как стационарное распределение системы массового обслуживания $MAP^{(r)}/PH/1/N_r$, $r = 1, 2, \dots, R$.

Для удобства читателя в следующем разделе приведем алгоритм вычисления стационарного распределения такого типа систем. Для краткости опустим индекс r в обозначении матриц, описывающих MAP-поток, и процесс обслуживания, а также в обозначениях числа мест в системе.

7.8.3 Стационарное распределение вероятностей состояний и времени пребывания запроса в системе $MAP/PH/1/N$

Функционирование системы $MAP/PH/1/N$ может быть описано трехмерной цепью Маркова $\zeta_t = \{n_t, \eta_t, \nu_t\}$, где n_t – число запросов в системе, $n_t \in \{0, \ldots, N\}$, η_t – текущее состояние процесса, управляющего обслуживанием запросов, $\eta_t \in \{1, \ldots, M\}$, а ν_t, $\nu_t \in \{0, \ldots, W\}$ – состояние управляющего процесса MAP в момент времени t.

Перенумеруем состояния цепи в лексикографическом порядке. Тогда инфинитезимальный генератор этой цепи имеет вид

$$A = \begin{pmatrix} \hat{D}_0 & \hat{D}_1 & O & \ldots & O & O \\ \hat{\mathbf{S}} & S \oplus D_0 & \hat{D}_1 & \ldots & O & O \\ O & \hat{\mathbf{S}} & S \oplus D_0 & \ldots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \ldots & S \oplus D_0 & \hat{D}_1 \\ O & O & O & \ldots & \hat{\mathbf{S}} & S \oplus (D_0 + D_1) \end{pmatrix},$$

где $\hat{D}_k = I_M \otimes D_k,\ k = 0, 1$, $\hat{\mathbf{S}} = \mathbf{S}_0 \boldsymbol{\beta} \otimes I_{W+1}$.

Пусть $\mathbf{q}$ является вектором-строкой стационарного распределения вероятностей состояний цепи. Этот вектор определяется как единственное решение системы линейных алгебраических уравнений

$$\mathbf{q}A = \mathbf{0},\ \mathbf{q}\mathbf{e} = 1.$$

В случае большой размерности данной системы для ее решения целесообразно использовать специальный алгоритм, учитывающий вероятностный смысл элементов матрицы A (см. [152]). Этот алгоритм существенно учитывает разреженную структуру матрицы A и описывается следующим образом.

Представим вектор $\mathbf{q}$ как $\mathbf{q} = (\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_N)$, где векторы $\mathbf{q}_i$, $i = 0, \ldots, N$, имеют размер $M(W + 1)$.

Алгоритм 7.2. Векторы стационарного распределения $\mathbf{q}_i$, $i = 0, \ldots, N$, вычисляются как

$$\mathbf{q}_l = \mathbf{q}_0 F_l,\ l = 1, \ldots, N,$$

где матрицы F_l вычисляются по следующим рекуррентным формулам:

$$F_0 = I,\ \ F_i = -F_{i-1}\hat{D}_1(S \oplus D_0 + (S \oplus D_1)\delta_{i,N} + \hat{D}_1 G_i)^{-1},\ \ i = 1, \ldots, N,$$

а матрицы $G_i, i = \overline{0, N-1}$, вычисляются с помощью обратной рекурсии

$$G_i = -(S \oplus D_0 + \hat{D}_1 G_{i+1})^{-1}\hat{\mathbf{S}},\ \ i = N-2, N-3, \ldots, 0,$$

при начальном условии

$$G_{N-1} = -(S \oplus (D_0 + D_1))^{-1}\hat{\mathbf{S}},$$

вектор $\mathbf{q}_0$ является единственным решением системы линейных алгебраических уравнений:

$$\mathbf{q}_0(\hat{D}_0 + \hat{D}_1 G_0) = \mathbf{0},\quad \mathbf{q}_0 \sum_{l=0}^{N} F_l \mathbf{e} = 1.$$

Утверждение 7.1. *Преобразование Лапласа – Стилтьеса $v(s)$ распределения времени пребывания запроса в системе $MAP/PH/1/N$ задается следующим образом:*

$$v(s) = \lambda^{-1}\left[\sum_{i=0}^{N-1} \mathbf{q}_i (I_M \otimes D_1\mathbf{e})(sI - S)^{-1}\mathbf{S}_0(\beta(s))^i + \mathbf{q}_N(\mathbf{e}_M \otimes D_1\mathbf{e})\right].$$

Доказательство легко проводится с использованием формулы полной вероятности и вероятностного смысла преобразования Лапласа – Стилтьеса в терминах понятия катастроф. Отметим, что компоненты вектор-столбца $(sI - S)^{-1}\mathbf{S}_0$ задают вероятность ненаступления катастрофы за оставшуюся длительность времени, имеющего распределение фазового типа, при фиксированном значении фазы обслуживания в настоящий момент времени.

Утверждение 7.2. *Среднее время v_1 пребывания запроса в системе $MAP/PH/1/N$ задается следующим образом:*

$$v_1 = \lambda^{-1} \sum_{i=0}^{N-1} \mathbf{q}_i (I_M \otimes D_1\mathbf{e})((-S)^{-1} + ib_1 I)\mathbf{e}.$$

7.8.4 Стационарное распределение тандема и его фрагментов

Теорема 7.31. *Стационарное распределение фрагмента* $\langle r, r+1, \ldots, r' \rangle$ *рассматриваемого тандема может быть рассчитано как стационарное распределение тандема* $MAP^{(r)}/PH/1/N_r \to \cdot/PH/1/N_{r+1} \to \ldots \to \cdot/PH/1/N_{r'}$, *где* $MAP^{(r)}$ *определяется по формулам (7.87)-(7.89).*

Следствие 7.24. *Вектор* $\mathbf{p}^{(r)}$ *маргинального стационарного распределения* r*-й станции тандема вычисляется как стационарное распределение системы массового обслуживания* $MAP^{(r)}/PH/1/N_r$, $r = 1, 2, \ldots, R$, *и может быть рассчитано с помощью алгоритма, описанного в разделе 3.*

Теорема 7.32. *Совместное стационарное распределение* $\mathbf{p}^{(1,\ldots,r)}$ *вероятностей состояний первых* r *станций тандема может быть вычислено как стационарное распределение управляющего процесса выходящего из* r*-й станции* $MAP^{(r)}$*-потока, т.е. имеет место формула*

$$\mathbf{p}^{(1,\ldots,r)} = \boldsymbol{\theta}^{(r)},$$

где вектор $\boldsymbol{\theta}^{(r)}$ *– единственное решение системы линейных алгебраических уравнений*

$$\boldsymbol{\theta}^{(r)}(D_0^{(r)} + D_1^{(r)}) = \mathbf{0}, \;\; \boldsymbol{\theta}^{(r)}\mathbf{e} = 1, \; r = 1, 2, \ldots, R,$$

а матрицы $D_0^{(r)}$, $D_1^{(r)}$ *вычисляются по рекуррентным формулам (7.87)-(7.88).*

В случае большой размерности эта система может быть успешно решена с помощью устойчивого алгоритма 7.2, описанного выше в подразделе 7.6.3. Очевидно, что вектор $\mathbf{p}$ стационарного распределения всего тандема совпадает с вектором $\mathbf{p}^{(1,\ldots,R)}$.

Следствие 7.25. *Вектор стационарного распределения тандема вычисляется как единственное решение системы линейных алгебраических уравнений*

$$\mathbf{p}(D_0^{(R)} + D_1^{(R)}) = \mathbf{0}, \;\; \mathbf{pe} = 1. \tag{7.90}$$

Как следует из системы (7.90), матрица $D_0^{(R)} + D_1^{(R)}$ совпадает с инфинитезимальным генератором Q цепи Маркова $\xi_t, t \geq 0$, описывающей функционирование тандема.

Таким образом, используя результаты анализа выходящих потоков со станций тандема, мы построили матрицу Q, избежав трудоемкой работы, требующейся при прямом подходе к построению этой матрицы.

7.8.5 Вероятности потерь

Теорема 7.33. *Вероятность потери $P_{loss}^{(r)}$ произвольного запроса на r-й станции тандема вычисляется как*

$$P_{loss}^{(r)} = \frac{\tilde{\lambda}_r - \lambda_r}{\tilde{\lambda}_r},\ r = 1, 2, \ldots, R, \tag{7.91}$$

Здесь $\tilde{\lambda}_r$ – интенсивность суммарного $MAP^{(r)}$-потока, поступающего на r-ю станцию,– вычисляется по формуле

$$\tilde{\lambda}_r = \tilde{\boldsymbol{\theta}}^{(r)} \tilde{D}_1^{(r)} \mathbf{e},$$

где вектор $\boldsymbol{\theta}^{(r)}$ единственное решение системы

$$\tilde{\boldsymbol{\theta}}^{(r)} (\tilde{D}_0^{(r)} + \tilde{D}_1^{(r)}) = \mathbf{0},\ \ \tilde{\boldsymbol{\theta}}^{(r)} \mathbf{e} = 1,$$

λ_r – интенсивность выходящего потока с r-й станции – вычисляется по формуле

$$\lambda_r = \boldsymbol{\theta}^{(r)} D_1^{(r)} \mathbf{e},$$

где вектор $\boldsymbol{\theta}^{(r)}$ единственное решение системы

$$\boldsymbol{\theta}^{(r)} (D_0^{(r)} + D_1^{(r)}) = \mathbf{0},\ \ \boldsymbol{\theta}^{(r)} \mathbf{e} = 1.$$

Теорема 7.34. *Вероятность потери на r-й станции тандема произвольного запроса, поступившего с $(r-1)$-й станции, вычисляется как*

$$P_{loss/1}^{(r)} = \frac{\tilde{\boldsymbol{\theta}}^{(r)} (D_1^{(r-1)} \otimes I_{W_r+1}) \mathbf{e}}{\lambda_{r-1}},\ r = 1, 2, \ldots, R. \tag{7.92}$$

Здесь $D_1^{(0)} = D_1$, λ_0 – интенсивность входящего потока в тандем.

Вероятность потери на r-й станции тандема произвольного запроса из дополнительного потока вычисляется по формуле

$$P_{loss/2}^{(r)} = \frac{\tilde{\boldsymbol{\theta}}^{(r)}(I_{K_r(N_r+1)} \otimes H_1^{(r)})\mathbf{e}}{h_r},\ r = 2, \ldots, R. \tag{7.93}$$

Доказательство. Числитель правой части (7.92) есть интенсивность потока запросов, поступающего на r-ю станцию из $(r-1)$-й станции и застающих все приборы r-й станции занятыми, а знаменатель – интенсивность потока всех запросов, поступающих на r-ю станцию из $(r-1)$-й станции. По известным эргодическим соображениям, отношение этих двух интенсивностей определяет искомую вероятность $P_{loss/1}^{(r)}$. Аналогичные рассуждения приводят к формуле (7.93) для вероятности $P_{loss/2}^{(r)}$. □

Используя теорему 7.34, можно получить альтернативное выражение для вероятности суммарных потерь $P_{loss}^{(r)}$ на r-й станции, ранее определенной формулой (7.91) в теореме 7.33 . Альтернативная формула (7.96), а также формулы для некоторых совместных вероятностей, ассоциированных с тандемом, приведены в следующем утверждении.

Следствие 7.26. *Вероятность того, что произвольный запрос из суммарного потока на r-ю станцию принадлежит выходящему потоку из $(r-1)$-й станции, и будет потерян из-за отсутствия свободных приборов на r-й станции, вычисляется как*

$$P_{loss,1}^{(r)} = P_{loss/1}^{(r)} \frac{\lambda_{r-1}}{\tilde{\lambda}_r},\ r = 1, 2, \ldots, R. \tag{7.94}$$

Вероятность того, что произвольный запрос из суммарного потока на r-ю станцию принадлежит дополнительному потоку и будет потерян из-за отсутствия свободных приборов на станции, вычисляется следующим образом:

$$P_{loss,2}^{(r)} = P_{loss/2}^{(r)} \frac{h_r}{\tilde{\lambda}_r},\ r = 2, \ldots, R. \tag{7.95}$$

Вероятность $P_{loss}^{(r)}$ потери произвольного запроса на r-й станции может быть вычислена как

$$P_{loss}^{(r)} = P_{loss,1}^{(r)} + P_{loss,1}^{(r)}. \tag{7.96}$$

Доказательство. Формулы (7.94)– (7.96) выводятся из очевидных вероятностных соображений. □

Следствие 7.27. *В случае $R = 2$ вероятность того, что запрос из входящего потока на первую станцию не будет потерян в системе, определяется как*

$$P_{succ} = (1 - P^{(1)}_{loss/1})(1 - P^{(2)}_{loss/1}).$$

7.8.6 Стационарное распределение времени пребывания запроса на станциях тандема и в тандеме в целом

Для точного расчета стационарного распределения времени пребывания в тандеме произвольного запроса необходимо проследить процесс прохождения этого запроса по станциям тандема. Время пребывания произвольного запроса в рассматриваемом тандеме зависит от состояния тандема в момент поступления запроса на первую станцию тандема и процессов, управляющих поступлением кросс-трафика на станциях тандема, до которых еще не дошел помеченный запрос. Отметим плату, которую пришлось заплатить за получение в более-менее простом рекуррентном виде генератора марковского процесса, описывающего поведение системы. Эта плата заключается в том, что этот процесс описан как

$$\xi_t = \{n_t^{(R)}, \eta_t^{(R)}, \nu_t^{(R)}, n_t^{(R-1)}, \eta_t^{(R-1)}, \nu_t^{(R-1)}, \ldots, n_t^{(1)}, \eta_t^{(1)}, \nu_t^{(1)}\},\ t \geq 0,$$

т.е. состояния станций тандема в нем описываются не в порядке возрастания номеров станций, а в порядке их убывания. В принципе, это не является существенным недостатком, поскольку выше описан путь нахождения маргинального распределения вероятностей состояний станций и даны выражения для различных вероятностей потерь в рассматриваемом тандеме. В конце концов, если является желательным знать стационарные вероятности процесса

$$\tilde{\xi}_t = \{n_t^{(1)}, \eta_t^{(1)}, \nu_t^{(1)}, n_t^{(2)}, \eta_t^{(2)}, \nu_t^{(2)}, \ldots, n_t^{(R)}, \eta_t^{(R)}, \nu_t^{(R)}\},\ t \geq 0,$$

то векторы стационарных вероятностей процесса $\tilde{\xi}_t$ могут быть получены путем соответствующих перестановок компонент векторов стационарных вероятностей процесса ξ_t. В принципе так же можно было бы поступить и для нахождения генератора процесса $\tilde{\xi}_t$ путем соответствующих перестановок компонент полученного выше генератора процесса ξ_t. Однако такие

перестановки неминуемо влекут утрату специфики блочной структуры генератора. Если генератор процесса ξ_t имеет блочную трехдиагональную структуру, то для вычисления стационарного распределения можно эффективно использовать приведенный в разделе 3 алгоритм. В то же время полученный путем перестановок генератор процесса $\tilde{\xi}_t$ не имеет выраженной блочной структуры, что резко снижает возможности расчета характеристик тандема на компьютере даже при относительно небольшом числе станций и размере буферов.

В силу сказанного, представляется логичным предложить точные или приближенные формулы для расчета вероятности потери в процессе прохождения обслуживания на станциях произвольного запроса, поступившего на первую станцию тандема, преобразования Лапласа-Стилтьеса распределения времени пребывания такого запроса на станциях тандема и в тандеме в целом и соответствующих средних времен (задержек на станциях тандема и в тандеме в целом), не основанные на строгом анализе процесса прохождения помеченного запроса по тандему.

Формула (7.94) задает вероятность $P_{loss,1}^{(r)}$ того, что произвольный запрос из суммарного потока на r-ю станцию принадлежит выходящему потоку из $(r-1)$-й станции и будет потерян из-за отсутствия свободных приборов на r-й станции. Хотя наряду с запросами, поступившими на первую станцию тандема, выходящий поток из $(r-1)$-й станции включает и другие запросы (поступившие в кросс-трафике на предыдущие станции), можно предположить, что $P_{loss,1}^{(r)}$ есть вероятность того, что запрос, поступивший на первую станцию тандема, будет потерян на r-й станции тандема. При таком предположении справедливы следующие утверждения.

Теорема 7.35. *Вероятность P_{loss} того, что запрос, поступивший на первую станцию тандема, будет потерян при прохождении по станциям тандема, задается выражением*

$$P_{loss} = \sum_{r=0}^{R-1} \prod_{k=1}^{r} (1 - P_{loss,1}^{(k)}) P_{loss,1}^{(r+1)}.$$

Теорема 7.36. *Преобразование Лапласа – Стилтьеса $v^{(r)}(s)$ распределения времени пребывания на r-й станции запроса, поступившего из $(r-1)$-й станции, $r \in \{1, \ldots, R\}$, задается следующим образом:*

$$v^{(r)}(s) = P_{loss,1}^{(r)} + \lambda_{r-1}^{-1} \sum_{i=0}^{N_r-1} \mathbf{p}_i^{(r)} (I_{M_r} \otimes (D_1^{(r-1)} \otimes I_{W_r+1}) \mathbf{e}) (sI - S^{(r)})^{-1} \mathbf{S}_0^{(r)} (\beta^{(r)}(s))^i,$$

где

$$\beta^{(r)}(s) = \boldsymbol{\beta}^{(r)}(sI - S^{(r)})^{-1}\mathbf{S}_0^{(r)}, \mathrm{Re}\, s > 0,$$

$\lambda_0 = \lambda$.

Доказательство следует из Утверждения 7.1.

Следствие 7.28. *Преобразование Лапласа – Стилтьеса $v_a^{(r)}(s)$ распределения времени пребывания на r-й станции запроса, поступившего из $(r-1)$-й станции и не потерянного на r-й станции, $r \in \{1, \ldots, R\}$, задается следующим образом:*

$$v_a^{(r)}(s) =$$

$$= (1-P_{loss,1}^{(r)})^{-1}\lambda_{r-1}^{-1}\sum_{i=0}^{N_r-1}\mathbf{p}_i^{(r)}(I_{M_r}\otimes(D_1^{(r-1)}\otimes I_{W_r+1})\mathbf{e})(sI-S^{(r)})^{-1}\mathbf{S}_0^{(r)}(\beta^{(r)}(s))^i.$$

Следствие 7.29. *Среднее время $v_1^{(r)}$ пребывания на r-й станции запроса, поступившего из $(r-1)$-й станции, $r \in \{1, \ldots, R\}$, задается следующим образом:*

$$v_1^{(r)} = \lambda_{r-1}^{-1}\sum_{i=0}^{N_r-1}\mathbf{p}_i^{(r)}(I_{M_r}\otimes(D_1^{(r-1)}\otimes I_{W_r+1})\mathbf{e})((-S^{(r)})^{-1} + ib_1^{(r)}I)\mathbf{e}.$$

Доказательство следует из Утверждения 7.2.

Теорема 7.37. *Преобразование Лапласа – Стилтьеса $v(s)$ распределения времени пребывания в системе запроса, поступившего на первую станцию, задается следующим образом:*

$$v(s) = P_{loss}+\prod_{r=1}^{R}\left[\lambda_{r-1}^{-1}\sum_{i=0}^{N_r-1}\mathbf{p}_i^{(r)}(I_{M_r}\otimes(D_1^{(r-1)}\otimes I_{W_r+1})\mathbf{e})(sI-S^{(r)})^{-1}\mathbf{S}_0^{(r)}(\beta^{(r)}(s))^i\right].$$

Следствие 7.30. *Преобразование Лапласа – Стилтьеса $v_a(s)$ распределения времени пребывания в системе запроса, поступившего на первую станцию и не потерянного в системе, задается следующим образом:*

$$v_a(s) =$$

$$= (1-P_{loss})^{-1}\prod_{r=1}^{R}\left[\lambda_{r-1}^{-1}\sum_{i=0}^{N_r-1}\mathbf{p}_i^{(r)}(I_{M_r}\otimes(D_1^{(r-1)}\otimes I_{W_r+1})\mathbf{e})(sI-S^{(r)})^{-1}\mathbf{S}_0^{(r)}(\beta^{(r)}(s))^i\right].$$

Следствие 7.31. *Среднее время v_1 пребывания в системе запроса, поступившего на первую станцию, задается следующим образом:*

$$v_1 = \sum_{r=1}^{R} \left[\lambda_{r-1}^{-1} \sum_{i=0}^{N_r-1} \mathbf{p}_i^{(r)} (I_{M_r} \otimes (D_1^{(r-1)} \otimes I_{W_r+1}))((-S^{(r)})^{-1} + ib_1^{(r)} I)\mathbf{e} \right].$$

Следствие 7.32. *Среднее время V пребывания в системе запроса, поступившего на первую станцию и не потерянного в системе, задается следующим образом:*

$$V = (1 - P_{loss})^{-1} v_1.$$

ГЛАВА 8
СИСТЕМЫ ДИНАМИЧЕСКОГО ПОЛЛИНГА

8.1 Введение

Интерес к исследованию систем стохастического поллинга со стороны математиков и прикладников не ослабевает уже в течение длительного времени, начиная с работ С. Мака и др. [199, 200], опубликованных в 1957 г. и до настоящего времени (см., например, [201–208]. В значительной мере это связано с тем, что модели стохастического поллинга эффективно используются для оценки производительности, проектирования и оптимизации структуры телекоммуникационных систем и сетей, транспортных систем и систем управления дорожным движением, производственных систем и систем управления запасами, и т.д. (см., например, [71, 209, 210]).

Системы поллинга, или системы циклического опроса, являются разновидностью систем массового обслуживания с несколькими очередями. В каждую очередь поступает свой поток заявок. Обслуживающий прибор (сервер) по определенному правилу посещает очереди и обслуживает находящиеся в них заявки. Разновидностью систем поллинга являются также приоритетные системы массового обслуживания, в которых заявки с более высоким приоритетом должны быть обслужены раньше заявок с низким приоритетом. В общем случае в системах поллинга сервер назначает приоритет очередям по определенному правилу (в соответствии с таблицей поллинга).

Правило, следуя которому сервер выбирает очередь для обслуживания, называют порядком обслуживания. Примерами такого правила могут служить циклический опрос очередей, когда сервер посещает очереди от первой до последней и вновь возвращается к первой очереди, или случайный порядок, при котором очередь на обслуживание выбирается случайно, и т.д.

Очереди системы поллинга обслуживаются согласно заданной дисциплине обслуживания. Она характеризуется числом заявок, которое может обслужить сервер за одно посещение очереди. Наиболее распространенными дисциплинами обслуживания являются: исчерпывающая дисциплина, при которой сервер обслуживает заявки до тех пор, пока очередь не опустеет; шлюзовая дисциплина, при которой сервер обслуживает лишь те

заявки, которые находились в очереди в момент подключения к ней сервера; ограниченная дисциплина, при которой число заявок, которое может обслужить сервер, ограничено.

Режим циклического опроса широко используется в телекоммуникационных системах с централизованным управлением. Соответственно, модели стохастического поллинга, адекватно описывающие функционирование таких систем, применяются для оптимизации и оценки их производительности.

Современные и перспективные протоколы широкополосных беспроводных сетей предусматривают методы и алгоритмы по решению "проблемы скрытых станций". Указанная проблема возникает, когда абонентская станция (называемая скрытой) по причине удаленности от базовой станции или нахождения в ее радиотени лишается возможности прослушивать передачу от базовой станции другой абонентской станции и может инициировать собственную передачу, которая приведет к коллизии. Иная проблема возникает в беспроводных сетях миллиметрового диапазона радиоволн и называется проблемой "глухоты". В этом диапазоне сигналы, излучаемые станциями, являются узко направленными, так что абонентская станция может "не услышатьчужую передачу и начать интерференцию. Эффективным способом борьбы с обеими проблемами является применение механизма циклического опроса базовой станцией очередей пакетов абонентских станций. Одним из первых примеров его реализации является протокол централизованного управления (или координации) с методом допуска Point Coordination Function – PCF, появившийся во время первой редакции стандарта IEEE 802.11. Его дальнейшее развитие привело к появлению метода HCCA – Hybrid Control Channel Access в дополнении IEEE 802.11e и его дальнейшее развитие в сетях сантиметрового и миллиметрового диапазонов радиоволн – IEEE 802.11ac и 802.11ad соответственно. При использовании этих методов на базовую станцию возлагается задача управления коллективным доступом всех остальных станций (узлов) сети к среде передачи данных на основе заданного алгоритма опроса или приоритетов узлов сети. Таким образом, базовая станция опрашивает все узлы, внесенные в список опроса, и на основании этого опроса организует передачу данных между всеми станциями сети. Список опроса может динамически меняться самой базовой станцией, что позволяет обслуживать новые станции. В отличии от механизмов DCF или EDCA, в которых каждый узел инициирует передачу данных без согласования с другими участ-

никами сети, в случае функционирования HCCA, разрешение на передачу узлам может предоставить только базовая станция. Важно отметить, что такой подход полностью исключает конкурентный доступ к среде (как в случае DCF и EDCA) и делает невозможным возникновение коллизий, вызванных "скрытоййли "глухойстанцией. При этом для чувствительных ко времени передачи приложений гарантируется приоритет доступа к сети.

Несмотря на то, что оборудование, поддерживающее централизированое управление, сложнее в разработке и производстве, а значит и дороже, оно гораздо эффективнее использует временные ресурсы беспроводной сети, что приводит к росту ее пропускной способности и уменьшению задержек передачи данных. Централизированное управление позволяет полностью устранить как проблему "скрытыхстанций, так и проблему "глухоты"; позволяет строго организовать порядок доступа станций к среде передачи данных, гибко управлять всей работой сети и менять ее параметры в зависимости от конкретной обстановки, настраивая только базовую станцию и не затрагивая абонентских. В настоящей главе дано описание новых методов исследования систем поллинга и оптимизации их характеристик.

Систематизация и обобщение теоретических результатов, полученных в области исследования систем поллинга до 1985 г., проведены в монографии Х. Такаги [214]. Дальнейшее развитие теоретических результатов в этом направлении, опубликованных до 1995 г., нашло отражение в монографии С. Борста [212], а работы, опубликованные в 1996-2009 годах, отражены в обзорах [179, 211]. Обобщению и систематизации моделей и методов исследования стохастических систем с циклическим опросом и их применению для проектирования широкополосных беспроводных сетей посвящены монографии [71, 208, 210]. В них рассмотрены модели поллинга. адекватно описывающие функционирование широкополосных беспроводных сетей под управлением протоколов Wi-Fi и Wi-Max с централизованными механизмом управления.

Типичной чертой большинства реальных систем поллинга является немгновенность переключения прибора с обслуживания одного буфера на другой. Поэтому важной задачей является оптимизация расписания переключения. Классические поллинговые дисциплины не предполагают зависимости времени между последовательными подключениями прибора к обслуживанию той или иной очереди. Вместе с тем, учет такой зависимости может помочь уменьшить затраты времени на переключения. Например, с точки зрения применения к широкополосным беспроводным сетям

связи, в [71, 213] была предложена дисциплина адаптивного поллинга. Эта дисциплина нацелена на избежание затраты времени на подключение к абонентской станции, которая в настоящее время не имеет никакой информации для передачи. Знать, имеет ли абонентская станция информацию для передачи в текущем цикле опроса, заранее невозможно. Однако высока вероятность того, что если длина цикла опроса станций невелика, то станция не имеет информации для передачи и в текущем цикле опроса, если она не имела такой информации в предыдущем цикле.

Поллинговые системы являются довольно сложными для математического анализа, поскольку необходимо одновременно учитывать состояние очередей во всех буферах системы и стохастический процесс, описывающий поведение системы, является многомерным. Поэтому для их исследования целесообразно использовать различные методы декомпозиции. Например, анализ поллинговой системы можно проводить через совместное рассмотрение совокупности так называемых систем массового обслуживания с отдыхами, поскольку с точки зрения отдельного конкретного пользователя время, затрачиваемое сервером на обслуживание других пользователей, можно интерпретировать как отдых сервера. Для ссылок на литературу по таким системам см., например, [176].

Методика приближенного анализа поллинговой системы, основанная на использовании результатов анализа $M/G/1$ системы с простейшим видом адаптивного отдыха и шлюзовым обслуживанием описана в [180]. Этот простейший вид предполагает наличие двух видов отдыха, отличающихся длительностью. Если в некоторый момент окончания отдыха буфер не пуст, то после завершения обслуживания запросов из него прибор берет первый вид отдыха (более короткий). Если же в момент окончания отдыха буфер пуст, то прибор немедленно берет второй вид отдыха (более длинный).

В работе [196] анализ такой модели с адаптивными отдыхами был распространен на систему типа $BMAP/G/1$. В данной главе результаты из [196] обобщены на случай более сложной адаптивной дисциплины отдыхов. А именно, эта дисциплина предполагает, что если буфер системы пуст в момент окончания отдыха, и был пуст в предыдущие $r-1$ моменты окончания отдыхов подряд, $r \geq 1$, то прибор берет отдых, названный отдыхом типа-r. Продолжительность этого отдыха характеризуется функцией распределения $H_r(t)$. В [196] предполагалось, что целое число r может принимать только значения 1 или 2. Здесь мы предполагаем, что $r \in \{2, \ldots, R\}$,

где R принимает конечное ($R \geq 2$) или бесконечное значение.

Стоит отметить, что исследуемая математическая модель может быть полезна не только для анализа производительности и настройки параметров протоколов в широкополосных беспроводных сетях связи. Она может быть полезна в применении к многочисленным реальным системам и сетям связи, где емкость батареи пользователя (или пользовательской станции) очень ограничена и частые опросы базовой станцией (прерывание режима ожидания), даже если пользовательская станция не имеет информации для передачи, приводят к быстрому ухудшению обслуживания пользователей. Базовая станция может регулировать время интервалов между моментами опроса помеченной станции, увеличивая эти интервалы, когда число предыдущих моментов опроса подряд, в которые пользовательская станция не имела информации для передачи, увеличивается. Кроме того, используя статистику предыдущей работы помеченного пользователя, базовая станция может наоборот увеличивать частоту опросов данного пользователя, когда велика вероятность того, что, выйдя из режима ожидания, пользователь будет иметь актуальную информацию для передачи.

8.2 Анализ системы $BMAP/G/1$ со шлюзовым обслуживанием и адаптивной продолжительностью отдыхов

8.2.1 Математическая модель

Рассматривается однолинейная система массового обслуживания с двумя бесконечными буферами, соединенными шлюзом.

Запросы поступают в систему в соответствии с $BMAP$-потоком. В момент поступления группа запросов размещается в первом буфере (Буфер 1). Обслуживание запросов из этой группы будет возможно только после того, как переключатель между Буфером 1 и Буфером 2 будет включен, и все запросы (если таковые имеются) мгновенно переместятся из Буфера 1 в Буфер 2.

Возможны следующие ситуации. Первая ситуация, которую назовем ситуацией-0, возникает когда Буфер 1 не пуст. В этом случае прибор начинает обслуживание запросов из Буфера 2 в порядке их поступления (для запросов, принадлежащих одной группе некоторого размера, скажем k, их порядок в группе выбирается произвольно в интервале $[1, k]$ с ве-

роятностью $\frac{1}{k}$). Время обслуживания запроса характеризуется функцией распределения $B(t)$, с преобразованием Лапласа – Стилтьеса (*ПЛС*) $\beta(s) = \int\limits_0^\infty e^{-st} dB(t),\ \ Re\ s > 0$, и конечными начальными моментами $b_k = \int\limits_0^\infty t^k dB(t),\ k \geq 1$. Обслуживание запросов продолжается до тех пор, пока не будет завершена обработка всех запросов из Буфера 2. После этого прибор берет отдых типа-0. Продолжительность этого отдыха характеризуется распределением $H_0(t)$, с *ПЛС* $h_0(s) = \int\limits_0^\infty e^{-st} dH_0(t),\ \ Re\ s > 0$, и конечными начальными моментами $h_k^{(0)} = \int\limits_0^\infty t^k dH_0(t),\ k \geq 1$. После завершения отдыха переключатель между Буфером 1 и Буфером 2 снова включается.

Ситуацией-r, $r \geq 1$, назовем ситуацию, когда Буфер 1 пуст в данный момент окончания отдыха и был пуст во все $r-1$ предыдущие моменты окончания отдыхов подряд.

В такой ситуации прибор незамедлительно берет другой отдых. Продолжительность этого отдыха характеризуется распределением $H_r(t)$, с *ПЛС* $h_r(s) = \int\limits_0^\infty e^{-st} dH_r(t),\ Re\ s > 0$, и конечными начальными моментами $h_k^{(r)} = \int\limits_0^\infty t^k dH_r(t),\ k \geq 1$.

Предполагается, что параметр r принимает значения в диапазоне $r \in \{1, \ldots, R\}$. Можно считать, что число R конечно или бесконечно. В частности, можно предположить, что существует такое число $\hat{R}$, что все распределения $H_r(t)$ совпадают для $r \geq \hat{R}$. Случай с $\hat{R} = 1$ рассматривался в [196]. При выводе формул ниже предполагается, что R бесконечно. Изменения в формулах в случае конечного R очевидны.

Проанализируем поведение описанной модели системы массового обслуживания.

Рассмотрим процесс

$$\zeta(t) = \{i_1(t), i_2(t), r(t), \nu_t\},\ \ t \geq 0,$$

где $i_k(t)$ число запросов в Буфере k, $k = 1, 2$, и процесс $r(t)$ описывает состояние прибора в момент t, $t \geq 0$:

$$r(t) = \begin{cases} *, & \text{если прибор осуществляет обслуживание запросов,} \\ r, & \text{если прибор находится на отдыхе типа-}r,\ r \geq 0. \end{cases}$$

Очевидно, что $i_2(t) = 0$, если $r(t) \neq *$. Стохастический процесс $\zeta(t),\ \ t \geq 0$, не является марковским, и для его исследования применим метод вложенных цепей Маркова.

8.2.2 Стационарное распределение вложенной цепи Маркова

Пусть t_n – это n-ый момент окончания отдыха, $n \geq 1$, j_n – число запросов в Буфере 1 в момент $t_n - 0$ (число запросов в Буфере 2 в момент $t_n + 0$). Пусть r_n равно 0, если Буфер 1 не был пустым в момент $t_n - 0$, и равно r, если Буфер 1 был пуст во все моменты $t_{n-r},\ t_{n-r+1},\ \ldots\ ,\ t_{n-1}$, $r \geq 1$, и пусть ν_n определяет ν_{t_n}.

Лемма 8.1. *Легко видеть, что процесс* $\xi_n = \{j_n, r_n, \nu_n\}$, $n \geq 1$, *является цепью Маркова с дискретным временем, и ненулевые матрицы* $\mathbf{P}_{(i,r),(j,r')}$, *образованные ее одношаговыми вероятностями переходов*

$$P\{j_{n+1} = j,\ r_{n+1} = r',\ \nu_{n+1} = \nu^{'} | j_n = i,\ r_n = r,\ \nu_n = \nu\},\ \nu, \nu^{'} = \overline{0, W},$$

$i,\ j \geq 0,\ r,\ r' \geq 0$, *имеют следующий вид:*

$$\mathbf{P}_{(i,r),(j,0)} = \int_0^\infty \mathbf{P}(j,t) dB^{*(i)} * H_0(t),\ i \geq 1,\ j \geq 0,\ r \geq 0, \tag{8.1}$$

$$\mathbf{P}_{(0,r),(j,r+1)} = \int_0^\infty \mathbf{P}(j,t) dH_{r+1}(t),\ j \geq 0,\ r \geq 0, \tag{8.2}$$

где матрицы $\mathbf{P}(l,t)$ *определяются как коэффициенты матричного разложения в ряд*

$$e^{\mathbf{D}(z)t} = \sum_{l=0}^{\infty} \mathbf{P}(l,t) z^l,$$

$B^{*(i)}(t)$ *– свертка* i*-го порядка функции распределения* $B(t)$, *и* $A * H(t)$ *– свертка функций распределения* $A(t)$ *и* $H(t)$.

По аналогии с доказательством Теоремы 1 в [196] можно показать, что стационарные вероятности

$$q(j,r,\nu) = \lim_{n \to \infty} P\{j_n = j,\ r_n = r,\ \nu_n = \nu\},\ j \geq 0,\ r \geq 0,\ \nu = \overline{0, W},$$

цепи Маркова $\xi_n,\ \ n \geq 1$, существуют, если выполняется неравенство

$$\rho = \lambda b_1 < 1.$$

Далее предполагаем, что эти неравенство выполняется.

Сгруппируем это вероятности в векторы-строки

$$\mathbf{q}(j,r) = (q(j,r,0),\ \ldots,\ q(j,r,W)),\ j \geq 0,\ r \geq 0.$$

Эти векторы удовлетворяют следующей системе уравнений Чепмена – Колмогорова:

$$\mathbf{q}(j,0) = \sum_{r=0}^{\infty}\sum_{i=1}^{\infty}\mathbf{q}(i,r)\int_0^{\infty}\mathbf{P}(j,t)dB^{*(i)} * H_0(t),\ j \geq 0, \tag{8.3}$$

$$\mathbf{q}(j,r+1) = \mathbf{q}(0,r)\int_0^{\infty}\mathbf{P}(j,t)dH_{r+1}(t),\ j \geq 0, r \geq 0. \tag{8.4}$$

Найдем решение бесконечной системы уравнений (8.3)-(8.4). Для этого введем в рассмотрение векторные производящие функции

$$\mathbf{Q}_r(z) = \sum_{j=0}^{\infty}\mathbf{q}(j,r)z^j,\ \ |z| < 1,\ r \geq 0.$$

Умножая уравнения системы (8.3)-(8.4) на соответствующие степени z и затем суммируя их, можно доказать следующее утверждение.

Лемма 8.2. *Векторные производящие функции* $\mathbf{Q}_r(z),\ \ |z| < 1,\ r \geq 0,$ *удовлетворяют следующей системе функциональных уравнений:*

$$\mathbf{Q}_0(z) = \sum_{r=0}^{\infty}\left(\mathbf{Q}_r(\boldsymbol{\beta}(-\mathbf{D}(z))) - \mathbf{Q}_r(0)\right)\mathbf{h}_0(-\mathbf{D}(z)), \tag{8.5}$$

$$\mathbf{Q}_{r+1}(z) = \mathbf{Q}_r(0)\mathbf{h}_{r+1}(-\mathbf{D}(z)), r \geq 0, \tag{8.6}$$

где

$$\beta(-\mathbf{D}(z)) = \int_0^{\infty}e^{\mathbf{D}(z)t}dB(t),\ \ \mathbf{h}_r(-\mathbf{D}(z)) = \int_0^{\infty}e^{\mathbf{D}(z)t}dH_r(t),\ r \geq 0,$$

матрицы, полученные из ПЛС $\beta(s),\ h_r(s),\ r \geq 0,$ *заменой скалярного аргумента* s *матричной производящей функцией* $-\mathbf{D}(z)$. *Эта замена справедлива (интегралы сходятся), поскольку все собственные значения матрицы* $\mathbf{D}(z)$ *имеют отрицательные действительные части.*

После ряда преобразований система (8.5), (8.6) может быть приведена к более простому виду:

$$\mathbf{Q}_0(z) = \mathbf{Q}_0(\boldsymbol{\beta}(-\mathbf{D}(z)))\mathbf{h}_0(-\mathbf{D}(z)) + \mathbf{Q}_0(0)\mathbf{A}(z), \tag{8.7}$$

$$\mathbf{Q}_{r+1}(z) = \mathbf{Q}_0(0)\mathbf{H}_r\mathbf{h}_{r+1}(-\mathbf{D}(z)),\ r \geq 0, \tag{8.8}$$

где

$$\mathbf{A}(z) = \mathbf{h}_0(-\mathbf{D}(z)) \left[\sum_{r=0}^{\infty} \mathbf{H}_r(\mathbf{h}_{r+1}(-\mathbf{D}(\boldsymbol{\beta}(-\mathbf{D}(z)))) - \mathbf{h}_{r+1}(-\mathbf{D}(0))) - I\right],$$

$$\mathbf{H}_r = \prod_{j=1}^{r} \mathbf{h}_j(-\mathbf{D}(0)).$$

Уравнение (8.7) является векторным функциональным уравнением с неизвестной векторной производящей функцией $\mathbf{Q}_0(z)$. Если эта функция будет вычислена, остальные векторные производящие функции $\mathbf{Q}_r(z),\ r \geq 1$, будут легко подсчитаны из (8.8).

Используя идею [196], будем решать векторное функциональное уравнение (8.7) следующим образом. Введем последовательность рекуррентно определенных матричных функций $\mathbf{N}_m(z),\ m \geq 0$:

$$\mathbf{N}_0(z) = zI,\ \ \mathbf{N}_{m+1}(z) = \boldsymbol{\beta}(-\mathbf{D}(\mathbf{N}_m(z))),\ m \geq 0. \tag{8.9}$$

Из [16] известно, что для любого $z,\ |z| < 1$, последовательность матриц $\mathbf{N}_m(z),\ m \geq 0$, сходится к матрице $\mathbf{G}$, которая определяется как решение нелинейного матричного уравнения

$$\mathbf{G} = \boldsymbol{\beta}(-\mathbf{D}(\mathbf{G})) = \int_0^{\infty} e^{\mathbf{D}(\mathbf{G})t} dB(t).$$

Оказывается, что матрица $\mathbf{G}$ играет ключевую роль не только при анализе цепей Маркова типа $M/G/1$, см. раздел 3.4, но и при исследовании цепи Маркова $\xi_n,\ n \geq 1$, у которой структура матрицы одношаговых вероятностей переходов, определенно, очень далека от структуры матрицы вероятностей переходов цепей Маркова типа $M/G/1$.

Последовательно подставляя матрицы $\mathbf{N}_m(z),\ m \geq 0$, в правой части функционального уравнения (8.7) вместо скалярного аргумента z, получаем следующее соотношение:

$$\mathbf{Q}_0(\mathbf{N}_0(z)) = \mathbf{Q}_0(\mathbf{N}_n(z)) \prod_{m=0}^{n-1} \mathbf{h}_0(-\mathbf{D}(\mathbf{N}_m(z)) +$$

$$+\mathbf{Q}_0(0)\sum_{m=0}^{n-1}\mathbf{A}(\mathbf{N}_m(z))\prod_{k=0}^{n-2}\mathbf{h}_0(-\mathbf{D}(\mathbf{N}_k(z))),\ n\geq 1. \tag{8.10}$$

Устремляя в (8.10) параметр n к бесконечности, получаем соотношение

$$\mathbf{Q}_0(z)=\mathbf{Q}_0(\mathbf{G})\prod_{l=0}^{\infty}\mathbf{h}_0(-\mathbf{D}(\mathbf{N}_l(z)))+$$

$$+\mathbf{Q}_0(0)\sum_{r=0}^{\infty}\mathbf{A}(\mathbf{N}_r(z))\prod_{m=0}^{r-1}\mathbf{h}_0(-\mathbf{D}(\mathbf{N}_m(z))). \tag{8.11}$$

Это функциональное уравнение проще по сравнению с уравнением (8.7), которое содержит неизвестную функцию $\mathbf{Q}_0(z)$ в точках $z,\ \boldsymbol{\beta}(-\mathbf{D}(z)),\ 0$, две из которых зависят от z, в то время как уравнение (8.11) содержит неизвестную векторную функцию $\mathbf{Q}_0(z)$ в точках $z,\ \mathbf{G},\ 0$, среди которых только одна зависит от переменной z.

Для вычисления наизвестных постоянных векторов $\mathbf{Q}_0(0)$ и $\mathbf{Q}_0(\mathbf{G})$, осуществим замену $z=\mathbf{G}$ в уравнении (8.7) и $z=0$ в уравнении (8.11). В результате получим следующее уравнение:

$$(\mathbf{Q}_0(0),\mathbf{Q}_0(\mathbf{G}))= \tag{8.12}$$

$$=(\mathbf{Q}_0(0),\mathbf{Q}_0(\mathbf{G}))\begin{pmatrix}\sum\limits_{r=0}^{\infty}\mathbf{A}(\mathbf{N}_r(0))\prod\limits_{m=0}^{r-1}\mathbf{h}_0(-\mathbf{D}(\mathbf{N}_m(0))) & \mathbf{A}(\mathbf{G})\\ \prod\limits_{l=0}^{\infty}\mathbf{h}_0(-\mathbf{D}(\mathbf{N}_l(0))) & \mathbf{h}_0(-\mathbf{D}(\mathbf{G}))\end{pmatrix}.$$

Из (8.8) следует

$$\mathbf{Q}_r(z)=\mathbf{Q}_0(0)\prod_{l=1}^{r}\mathbf{H}_{l-1}\mathbf{h}_l(-\mathbf{D}(z)),\ r\geq 1. \tag{8.13}$$

Из (8.11),(8.13) и условия нормировки $\sum\limits_{r=0}^{\infty}\mathbf{Q}_r(1)\mathbf{e}=1$, получаем еще одно уравнение для неизвестного вектора $(\mathbf{Q}_0(0),\mathbf{Q}_0(\mathbf{G}))$:

$$1=(\mathbf{Q}_0(0),\mathbf{Q}_0(\mathbf{G}))\times \tag{8.14}$$

$$\times\begin{pmatrix}\sum\limits_{r=0}^{\infty}\mathbf{A}(\mathbf{N}_r(1))\prod\limits_{m=0}^{r-1}\mathbf{h}_0(-\mathbf{D}(\mathbf{N}_m(1)))+\sum\limits_{r=0}^{\infty}\mathbf{H}_r\mathbf{h}_{r+1}(-\mathbf{D}(1)) & 0\\ 0 & \prod\limits_{l=0}^{\infty}\mathbf{h}_0(-\mathbf{D}(N_l(1)))\end{pmatrix}\mathbf{e}.$$

Таким образом, доказано следующее утверждение.

Теорема 8.1. *Векторная производящая функция* $\mathbf{Q}_0(z)$ *задается формулой (8.11), где неотрицательные векторы* $\mathbf{Q}_0(0)$ *и* $\mathbf{Q}_0(\mathbf{G})$ *определяются как решение системы линейных алгебраических уравнений (8.12), (8.14).*

Векторные производящие функции $\mathbf{Q}_r(z),\ r \geq 1$, *вычисляются по формуле (8.13).*

Следствие 8.1. *Векторная производящая функция* $\mathbf{Q}(z)$ *стационарного распределения длины очереди в моменты окончания отдыхов вычисляется по формуле*

$$\mathbf{Q}(z) = \sum_{r=0}^{\infty} \mathbf{Q}_r(z) = \mathbf{Q}_0(z) \sum_{r=0}^{\infty} \prod_{l=1}^{r} \mathbf{H}_{l-1}\mathbf{h}_l(-\mathbf{D}(z)).$$

По умолчанию предполагается, что произведение равно I, *если нижнее значение индекса произведения больше, чем верхнее.*

Доказательство следует из формулы (8.13).

Следствие 8.2. *Среднее число* L *запросов в системе в моменты окончания отдыхов вычисляется как* $L = \mathbf{Q}'(1)\mathbf{e}$.

Среднее число моментов окончания отдыхов, в которые система была пуста, определяемое в данный момент окончания отдыха, вычисляется как $\sum\limits_{r=0}^{\infty} r\mathbf{Q}_r(1)\mathbf{e}$.

8.2.3 Стационарное распределение состояний системы в произвольный момент времени

Введем стационарные вероятности состояний системы в произвольный момент времени следующим образом:

$$p(i_1, i_2, *, \nu) = \lim_{t\to\infty} P\{i_1(t) = i_1,\ i_2(t) = i_2,\ r(t) = *,\ \nu_t = \nu\},\ i_1 \geq 0,\ \ i_2 \geq 1,$$

$$p(i_1, r, \nu) = \lim_{t\to\infty} P\{i_1(t) = i_1,\ r(t) = r,\ \nu_t = \nu\},\ i_1 \geq 0,\ r \geq 0,\ \nu = \overline{0, W}.$$

Эти вероятности существуют при условии $\rho = \lambda b_1 < 1$, которое выше было предположено выполненным.

Введем векторы-строки

$$\mathbf{p}(i_1, i_2, *) = (p(i_1, i_2, *, 0), \ldots, p(i_1, i_2, *, W)),$$

$$\mathbf{p}(i_1, r) = (p(i_1, r, 0), \ldots, p(i_1, r, W)),\ r \geq 0.$$

Справедливо следующее утверждение.

Теорема 8.2. *Векторы* $\mathbf{p}(i_1, i_2, *)$, $\mathbf{p}(i_1, r)$, $r \geq 0$, *выражаются через векторы* $\mathbf{q}(i, r)$, $i \geq 0$, $r \geq 0$, *следующим образом:*

$$\mathbf{p}(i_1, i_2, *) = \tau^{-1}\Big[\sum_{j=i_2+1}^{\infty} \sum_{r=0}^{\infty} \mathbf{q}(j, r) \sum_{l=0}^{i_1} \int_0^{\infty} \mathbf{P}(l, t) dB^{*(j-i_2)}(t) \times$$

$$\int_0^{\infty} \mathbf{P}(i_1 - l, t)(1 - B(t))dt + \sum_{r=0}^{\infty} \mathbf{q}(i_2, r) \int_0^{\infty} \mathbf{P}(i_1, t)(1 - B(t))dt\Big], \quad (8.15)$$

$$\mathbf{p}(i_1, 0) = \tau^{-1} \sum_{j=1}^{\infty} \sum_{r=0}^{\infty} \mathbf{q}(j, r) \sum_{l=0}^{i_1} \int_0^{\infty} \mathbf{P}(l, t) dB^{*(j)}(t) \int_0^{\infty} \mathbf{P}(i_1 - l, t)(1 - H_0(t))dt, \quad (8.16)$$

$$\mathbf{p}(i_1, r) = \tau^{-1} \mathbf{q}(0, r-1) \int_0^{\infty} \mathbf{P}(i_1, t)(1 - H_r(t))dt, \ r \geq 1, \quad (8.17)$$

где среднее время τ *между последовательными моментами окончания отдыхов вычисляется как*

$$\tau = b_1 \mathbf{Q}'(1)\mathbf{e} + h_1^{(0)}(1 - \mathbf{Q}(0)\mathbf{e}) + \mathbf{Q}_0(0) \sum_{r=0}^{\infty} \mathbf{H}_r h_1^{(r+1)} \mathbf{e}. \quad (8.18)$$

Доказательство теоремы основывается на известных результатах для процессов марковского восстановления (см., например, [112]) и вероятностном смысле введенных матриц.

Введем следующие векторные производящие функции:

$$\mathbf{\Pi}(z, y, *) = \sum_{i_1=0}^{\infty} \sum_{i_2=1}^{\infty} \mathbf{p}(i_1, i_2, *) z^{i_1} y^{i_2}, \ |z| \leq 1, \ |y| \leq 1,$$

$$\mathbf{\Pi}(z, r) = \sum_{i_1=0}^{\infty} \mathbf{p}(i_1, r) z^{i_1}, \ |z| \leq 1, \ \ r \geq 0.$$

Следствие 8.3. *Векторные производящие функции* $\mathbf{\Pi}(z, y, *)$, $\mathbf{\Pi}(z, r)$, $r \geq 0$, *вычисляются по формулам*

$$\mathbf{\Pi}(z, y, *) = \tau^{-1}(\mathbf{Q}(\boldsymbol{\beta}(-\mathbf{D}(z))) - \mathbf{Q}(y))y(\boldsymbol{\beta}(-\mathbf{D}(z)) - yI)^{-1}(-\mathbf{D}(z))^{-1} \times$$

$$(I - \boldsymbol{\beta}(-\mathbf{D}(z))), \quad (8.19)$$

$$\mathbf{\Pi}(z, 0) = \tau^{-1}(\mathbf{Q}(\boldsymbol{\beta}(-\mathbf{D}(z))) - \mathbf{Q}(0))(-\mathbf{D}(z))^{-1}(I - \mathbf{h}_0(-\mathbf{D}(z))), \quad (8.20)$$

$$\mathbf{\Pi}(z, r) = \tau^{-1} \mathbf{Q}_{r-1}(0)(-\mathbf{D}(z))^{-1}(I - \mathbf{h}_r(-\mathbf{D}(z))), \ r \geq 1. \quad (8.21)$$

Доказательство. Формула (8.19) получается умножением уравнения (8.15) на соответствующие степени z и y и суммированием их. Аналогично, формулы (8.20)-(8.21) получаются умножением уравнений (8.16) и (8.17) на соответствующие степени z и суммированием их. □

Следствие 8.4. *Вероятность того, что прибор занят (осуществляет обслуживание запросов) в произвольный момент времени, определяется как*

$$\mathbf{\Pi}(1,1,*)\mathbf{e} = \tau^{-1}\mathbf{Q}'(1)\mathbf{e}b_1.$$

Вероятность того, что прибор находится в состоянии отдыха типа-0 в произвольный момент времени, определяется как

$$\mathbf{\Pi}(1,0)\mathbf{e} = \tau^{-1}(1-\mathbf{Q}(0)\mathbf{e})h_1^{(0)}.$$

Вероятность того, что прибор находится в состоянии отдыха типа-r в произвольный момент времени, определяется как

$$\mathbf{\Pi}(1,r)\mathbf{e} = \tau^{-1}\mathbf{Q}_{r-1}(0)\mathbf{e}h_1^{(r)},\ r \geq 1.$$

Доказательство очевидно.

Следствие 8.5. *Векторная производящая функция $\mathbf{R}(z)$ распределения числа запросов в Буфере 1 в произвольный момент времени, определяется как*

$$\mathbf{R}(z) = \tau^{-1}(\mathbf{Q}(1)-\mathbf{Q}(z))(-\mathbf{D}(z))^{-1}.$$

Доказательство. Очевидно, что

$$\mathbf{R}(z) = \mathbf{\Pi}(z,1,*) + \sum_{r=0}^{\infty}\mathbf{\Pi}(z,r).$$

Подставляя в эту формулу выражения (8.19)-(8.21) и принимая во внимание функциональные уравнения (8.5)-(8.6), легко можно убедиться в справедливости доказываемого утверждения. □

Следствие 8.6. *Векторная производящая функция $\mathbf{P}(z)$ распределения числа запросов в системе (в Буфере 1 и Буфере 2) в произвольный момент времени определяется как*

$$\mathbf{P}(z) = \tau^{-1}(\mathbf{Q}(\boldsymbol{\beta}(-\mathbf{D}(z)))-\mathbf{Q}(z))\boldsymbol{\beta}(-\mathbf{D}(z))(1-z)(\boldsymbol{\beta}(-\mathbf{D}(z))-zI)^{-1}(-\mathbf{D}(z))^{-1}.$$

Доказательство. Очевидно, что

$$\mathbf{P}(z) = \mathbf{\Pi}(z, z, *) + \sum_{r=0}^{\infty} \mathbf{\Pi}(z, r).$$

В остальном доказательство аналогично доказательству предыдущего следствия. □

Замечание 8.1. Среднее число запросов в Буфере 1 в произвольный момент времени вычисляется как $\mathbf{R}'(1)\mathbf{e}$. Среднее число запросов в системе вычисляется как $\mathbf{P}'(1)\mathbf{e}$. Эти выражения выглядят достаточно просто. Однако на практике вычисления на их основе не являются тривиальными, поскольку выражения для $\mathbf{R}(z)$ и $\mathbf{P}(z)$ содержат неопределенность вида $\frac{0}{0}$ в точке $z = 1$, и требуют применения правила Лопиталя. Полезные для вычислений результаты можно найти в Приложении Б в [196].

8.2.4 Распределение времени ожидания произвольного запроса в системе

Пусть $W(x)$ есть функция распределения времени ожидания произвольного запроса в системе и $w(s)$ – ее *ПЛС*:

$$w(s) = \int_0^{\infty} e^{-sx} dW(x),\ Re\ s > 0.$$

Как уже было сказано выше, предполагается, что запросы обслуживаются в порядке их поступления в систему (дисциплина обслуживания $FIFO$). Если запрос поступает в группе размера k, он будет обслужен r-м в группе с вероятностью $\frac{1}{k},\ r = 1, \ldots, k$.

Теорема 8.3. *ПЛС $w(s)$ времени ожидания произвольного запроса определяется формулой*

$$w(s) = \tag{8.22}$$

$$= (\lambda\tau)^{-1}\Big(\mathbf{Q}(\beta(s))(h_0(s)-1)+\sum_{r=0}^{\infty}\mathbf{Q}_r(0)h_{r+1}(s)-\mathbf{Q}(0)h_0(s)\Big)(1-\beta(s))^{-1}\mathcal{B}(s),$$

где

$$\mathcal{B}(s) = (sI + \mathbf{D}(\beta(s)))^{-1}\mathbf{D}(\beta(s))\mathbf{e}. \tag{8.23}$$

Доказательство. Получим выражение (8.22) для *ПЛС* $w(s)$, используя метод коллективных меток (метод катастроф), см., например, [143, 178]. Будем рассматривать переменную s как интенсивность некоторого виртуального стационарного пуассоновского потока событий, называемых катастрофами. Тогда легко видеть, что *ПЛС* $w(s)$ равно вероятности того, что в течение времени ожидания не наступит катастрофа. Соответственно, *ПЛС* $\beta(s),\ h_r(s)$ равны вероятностям того, что в течение времени обслуживания и времени отдыха типа-r, $r \geq 0$, не наступит катастрофа.

Введем также функции $w_r(s),\ r = *, 0, 1, 2, \ldots$, которые равны вероятностям того, что произвольный запрос поступит тогда, когда состояние прибора – r (описание состояний см.выше) и в течение времени ожидания этого запроса не наступит катастрофа. Очевидно, что

$$w(s) = w_*(s) + \sum_{r=0}^{\infty} w_r(s). \tag{8.24}$$

Время ожидания произвольного (помеченного) запроса, который поступает, когда прибор занят, определяется следующим образом. Пусть запрос поступает в момент времени $t,\ t \geq 0$, после окончания отдыха, к которому $j,\ j \geq 1$, запросов перешли из Буфера 1 в Буфер 2. И пусть в течение интервала времени $(0, t)$, i запросов поступило в Буфер 1, $i \geq 0$. Кроме того, помеченный запрос поступает в произвольный момент времени t в составе группы, состоящей из $k,\ k \geq 1$, запросов, и занимает r-ю позицию в этой группе, $r = \overline{1, k}$. Вероятность этих событий определяется формулой

$$\mathbf{q}_j \mathbf{P}(i, t)\lambda^{-1} k \mathbf{D}_k dt \mathbf{e} \frac{1}{k}.$$

Пусть в некоторый момент времени $x,\ x \leq t$, равно $m,\ m = 0, \ldots, j-1$, запросов из Буфера 2 закончили обслуживание и $(m+1)$-й запрос продолжает обслуживание в момент времени t. Пусть также его оставшееся время обслуживания равно $u,\ u > 0$ (то есть общее время обслуживания запроса, находящегося на обслуживании в момент поступления помеченного запроса, равно $t - x + u$), и катастрофа не наступит в течение этого оставшегося времени обслуживания. Вероятность этих событий определяется формулой

$$dB^{(*(m))}(x) e^{-su} dB(t - x + u).$$

После момента $t + u$, время ожидания помеченного запроса состоит из времени обслуживания: (а) $j - m - 1$ запросов, которые находились в Буфере 2 в момент поступления помеченного запроса, (б) i запросов, которые

поступили в Буфер 1 до момента поступления помеченного запроса, и (в) $r-1$ запросов, которые поступили в Буфер 1 в той же группе, что и помеченный запрос, но располагаются перед ним, и времени отдыха типа-0. Вероятность того, что в течение этого времени не наступит катастрофа, равна $(\beta(s))^{j-m-1+i+r-1}h_0(s)$.

Принимая во внимание приведенные выше пояснения и используя формулу полной вероятности, вероятностную интерпретацию *ПЛС* и теорию полурегенерирующих процессов, получим следующее выражение для функции $w_*(s)$:

$$w_*(s) = \tau^{-1}\sum_{j=1}^{\infty}\sum_{r=0}^{\infty}\mathbf{q}(j,r)\sum_{m=0}^{j-1}\int_0^{\infty}\int_0^{t}dB^{*(m)}(x)\sum_{i=0}^{\infty}\mathbf{P}(i,t)\lambda^{-1}\sum_{k=1}^{\infty}k\mathbf{D}_k dt\mathbf{e}\times$$

$$\times\int_0^{\infty}e^{-su}dB(t-x+u)\sum_{l=1}^{k}\frac{1}{k}(\beta(s))^{j-m-1+i+l-1}h_0(s). \qquad (8.25)$$

После ряда преобразований, аналогичных приведенным в [19], выражение (8.25) может быть переписано в более простом виде:

$$w_*(s) = (\lambda\tau)^{-1}(\mathbf{Q}(\beta(s)) - \mathbf{Q}(\beta(-\mathbf{D}(\beta(s)))))h_0(s)(1-\beta(s))^{-1}\mathcal{B}(s). \quad (8.26)$$

Используя аналогичные рассуждения, получаем следующие выражения для функций $w_r(s),\ r \geq 0$:

$$w_0(s) = \tau^{-1}\sum_{j=1}^{\infty}\sum_{r=0}^{\infty}\mathbf{q}(j,r)\int_0^{\infty}\int_0^{t}dB^{*(j)}(x)\sum_{i=0}^{\infty}\mathbf{P}(i,t)\lambda^{-1}\sum_{k=1}^{\infty}k\mathbf{D}_k dt\mathbf{e}\times$$

$$\times\int_0^{\infty}e^{-su}dH_0(t-x+u)\sum_{l=1}^{k}\frac{1}{k}(\beta(s))^{i+l-1},$$

$$w_r(s) = \tau^{-1}\mathbf{q}(0,r-1)\int_0^{\infty}\sum_{i=0}^{\infty}\mathbf{P}(i,y)\lambda^{-1}\sum_{k=1}^{\infty}k\mathbf{D}_k dy\mathbf{e}\times$$

$$\int_0^{\infty}e^{-su}dH_r(y+u)\sum_{l=1}^{k}\frac{1}{k}(\beta(s))^{i+l-1},\ r\geq 1,$$

которые могут быть переписаны в более простом виде

$$w_0(s) =$$

$$= -(\lambda\tau)^{-1}(\mathbf{Q}(\boldsymbol{\beta}(-\mathbf{D}(\beta(s))))-\mathbf{Q}(0))(\mathbf{h}_0(-\mathbf{D}(\beta(s)))-h_0(s)I)(1-\beta(s))^{-1}\mathcal{B}(s)),$$

$$w_r(s) = -(\lambda\tau)^{-1}\mathbf{q}(0,r-1)(\mathbf{h}_r(-\mathbf{D}(\beta(s)))-h_r(s)I)(1-\beta(s))^{-1}\mathcal{B}(s)), \quad (8.27)$$

$$r \geq 1.$$

Подставляя (8.26) и (8.27) в формулу (8.24) и принимая во внимание функциональное уравнение (8.5), получим формулу (8.22). □

Следствие 8.7. *Среднее время ожидания W_1 произвольного запроса вычисляется как*

$$W_1 = \frac{1}{\lambda\tau}\Bigg[\Bigg((\mathbf{Q}(1)-\mathbf{Q}(0))\frac{(h_1^{(0)}b_2 - h_2^{(0)}b_1)}{2b_1^2} - \mathbf{Q}^{'}(1)h_1^{(0)} +$$

$$+\sum_{r=0}^{\infty}\mathbf{Q}_r(0)\frac{(h_1^{(r+1)}b_2 - h_2^{(r+1)}b_1)}{2b_1^2}\Bigg)\mathcal{B}(0) +$$

$$+b_1^{-1}\Bigg((\mathbf{Q}(1)-\mathbf{Q}(0))h_1^{(0)} + \sum_{r=0}^{\infty}\mathbf{Q}_r(0)h_1^{(r+1)}\Bigg)\mathcal{B}^{'}(0)\Bigg], \quad (8.28)$$

где

$$\mathcal{B}(0) = -\frac{\rho}{1-\rho}\mathbf{e}, \quad (8.29)$$

$$\mathcal{B}^{'}(0) = \frac{1}{1-\rho}\Bigg(\tilde{I}\mathbf{D}(1) + \hat{\mathbf{e}}\boldsymbol{\theta}(I - b_1\mathbf{D}^{'}(1))\Bigg)^{-1}\times$$

$$\Bigg(\tilde{I}(\rho\mathbf{e} - b_1\mathbf{D}^{'}(1)\mathbf{e}) + \frac{1}{2}\hat{\mathbf{e}}\boldsymbol{\theta}(b_1^2\mathbf{D}^{''}(1) + b_2\mathbf{D}^{'}(1))\mathbf{e}\Bigg). \quad (8.30)$$

Здесь $\tilde{\mathbf{I}}$ *диагональная матрица с диагональными элементами* $\{0,1,\ldots,1\}$, *и вектор-столбец* $\hat{\mathbf{e}}$ *с элементами* $\{1,0,\ldots,0\}$.

Доказательство. В принципе, доказательство достаточно простое. Оно основано на очевидной формуле $W_1 = -w'(0)$ и формулах (8.22) и (8.23). Однако на практике оно оказывается довольно утомительным, поскольку выражения (8.22) и (8.23) и их производные в точке $s = 0$ содержат неопределенность вида $\frac{0}{0}$, и требуют применения правила Лопиталя. Соответствующую информацию можно найти в доказательстве Следствия 5 в [196]. □

8.2.5 Численные результаты

Целью этого подраздела является демонстрация работоспособности разработанных алгоритмов для вычислений, и формирование некоторого представления о количественном поведении изучаемой системы. Мы представляем результаты трех экспериментов.

Эксперимент 8.1. Данный эксперимент иллюстрирует необходимость учитывать корреляцию во входном потоке.

Будем рассматривать следующие пять различных входных потоков с одинаковой интенсивностью λ=2.5, но разными корреляцией и вариацией интервалов между моментами поступления групп и разными распределениями числа запросов в группе.

- M, стационарный пуассоновский поток, который определяется матрицами
$$\mathbf{D}_0 = \begin{pmatrix} -2.5 \end{pmatrix}, \mathbf{D}_1 = \begin{pmatrix} 2.5 \end{pmatrix}.$$
- IPP, прерывающийся пуассоновский поток, который определяется матрицами
$$\mathbf{D}_0 = \begin{pmatrix} -3.725 & 0.6 \\ 2.4 & -2.4 \end{pmatrix}, \mathbf{D}_1 = \begin{pmatrix} 3.125 & 0 \\ 0 & 0 \end{pmatrix}.$$
Процессы M и IPP являются некоррелированными.
- $BMAP_{0.035}$, $BMAP$ – групповой марковский входной поток с коэффициентом корреляции c_{cor}=0.035, определяется матрицами
$$\mathbf{D}_0 = \begin{pmatrix} -1.45 & 0.45 \\ 0.6 & -2.6 \end{pmatrix}, \mathbf{D}_1 = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.5 \end{pmatrix}, \mathbf{D}_2 = \begin{pmatrix} 0.75 & 0 \\ 0 & 1.5 \end{pmatrix}.$$
- $BMAP_{0.16}$, $BMAP$ – групповой марковский входной поток с коэффициентом корреляции c_{cor}=0.16, определяется матрицами
$$\mathbf{D}_0 = \begin{pmatrix} -7.5 & 0.9375 \\ 0.1875 & -1.5 \end{pmatrix}, \mathbf{D}_1 = \begin{pmatrix} 5.625 & 0 \\ 0 & 1.125 \end{pmatrix}, \mathbf{D}_2 = \begin{pmatrix} 0.9375 & 0 \\ 0 & 0.1875 \end{pmatrix}.$$
- $BMAP_{0.3}$, $BMAP$ – групповой марковский входной поток с коэффициентом корреляции c_{cor}=0.3, определяется матрицами
$$\mathbf{D}_0 = \begin{pmatrix} -3.785 & 0.035 \\ 0.013 & -0.628 \end{pmatrix}, \mathbf{D}_1 = \begin{pmatrix} 1.5 & 0 \\ 0 & 0.03 \end{pmatrix}, \mathbf{D}_2 = \begin{pmatrix} 2.25 & 0 \\ 0 & 0.585 \end{pmatrix}.$$

Предполагаем, что время обслуживания имеет экспоненциальное распределение с параметром 7. Пусть $\hat{R} = 2$, что означает, что отдыхи могут быть трех различных типов. Длительность отдыха типа-r, $r = 0, 1, 2$, имеет экспоненциальное распределение с параметром 0.3, 0.13 и 0.1 соответственно.

На следующих рисунках мы изменяем значение основной интенсивности λ с 0.8875 до 4.6375, умножая матрицы $\mathbf{D}_k$, $k \geq 0$, на скаляр γ, который принимает значения в интервале $(0.355, 1.855)$. Таким образом, коэффициент загрузки системы ρ изменяется от 0.13 до 0.66. Зависимость средней длины интервала τ и среднего времени ожидания W_1 от интенсивности λ представлена на Рисунке 8.1.

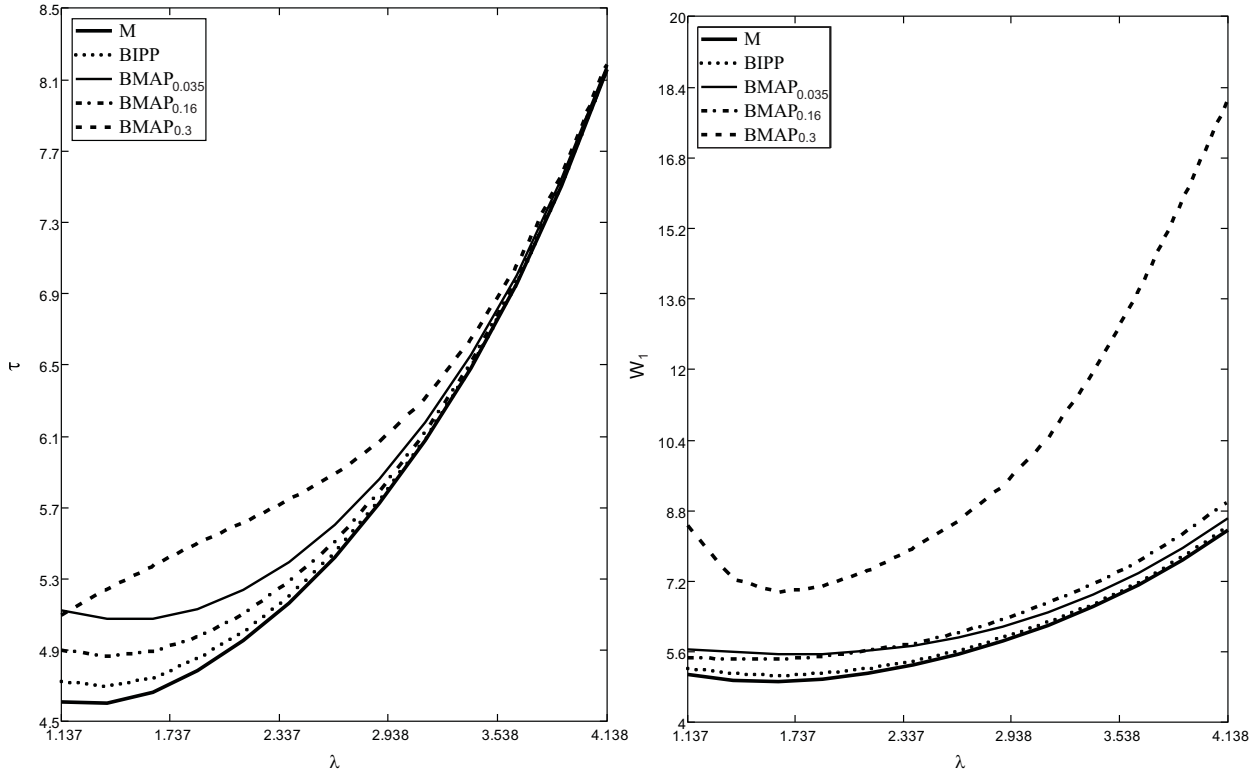

Рисунок 8.1. Зависимость τ и W_1 от λ

Как видно из Рисунка 8.1, корреляция имеет большое значение. Когда корреляция высокая (0.3), среднее время ожидания больше, по сравнению со случаями с низкой корреляцией. Разница существенно увеличивается, когда система становится более загруженной. Таким образом, пренебрежение корреляцией и предположение, что любой входной поток хорошо аппроксимируется стационарным пуассоновским потоком, может привести к огромным ошибкам в оценке эффективности для рассматриваемой модели с отдыхами.

Стоит также отметить, что среднее время ожидания W_1 увеличивается не только с ростом коэффициента загрузки системы ρ (что вполне естественно), но оно также увеличивается и когда коэффициент загрузки ρ становится маленьким. Этот факт, очевидно, объясняется следующим образом. Когда загрузка ρ существенно снижается, система чаще оказывается пустой в моменты окончания отдыхов и, таким образом, отдыхи типа-1 и типа-2 используются намного чаще. Поскольку в данном примере длительность этих отдыхов длиннее, чем длительность отдыха типа-0, это влечет увеличение среднего времени ожидания.

Эксперимент 8.2. Этот эксперимент частично дает ответ на вопрос: важно ли учитывать распределение времени отдыха или достаточно принимать во внимание только среднее значение времени отдыха. Другими словами, есть ли необходимость рассматривать произвольное рапределение времени отдыха, или достаточно ограничиться более простым и популярным экспоненциальным распределением.

Рассмотрим два вида распределения длительности отдыха. Одним из них является экспоненциальное распределение со средним временем отдыха 3.33, 7.69 и 10 для отдыхов типа 0,1 и 2 соответственно. Параметры соответствующей функции распределения равны 0.3, 0.13 и 0.1. В качестве второго вида распределения возьмем гиперэкспоненциальное распределение:

$$H_0(t) = 1 - (0.4e^{-0.132t} + 0.6e^{-2t}),$$
$$H_1(t) = 1 - (0.4e^{-0.053t} + 0.6e^{-3t}),$$
$$H_2(t) = 1 - (0.4e^{-0.04t} + 0.6e^{-4t}).$$

Средние длительности отдыхов снова равны 3.33, 7.69 и 10 для отдыхов типа 0, 1 и 2 соответственно.

Предполагаем, что входной поток совпадает с $BMAP_{0.16}$.

Зависимость τ и W_1 от интенсивности λ для рассматриваемых распределений времени отдыха изображена на Рисунке 8.2.

Можно видеть, что среднее время ожидания значительно дольше для гиперэкспоненциального распределения и имеет более высокое изменение длительности отдыха по сравнению с экспоненциальным распределением.

На Рисунке 8.3 изображена зависимость τ и W_1 от средней длительности $h_1^{(0)}$ отдыха типа-0, которая изменяется в интервале (0.1,2). Средние длительности отдыхов типа-1 и типа-2 зафиксированы и равны 2 и 4 соответственно.

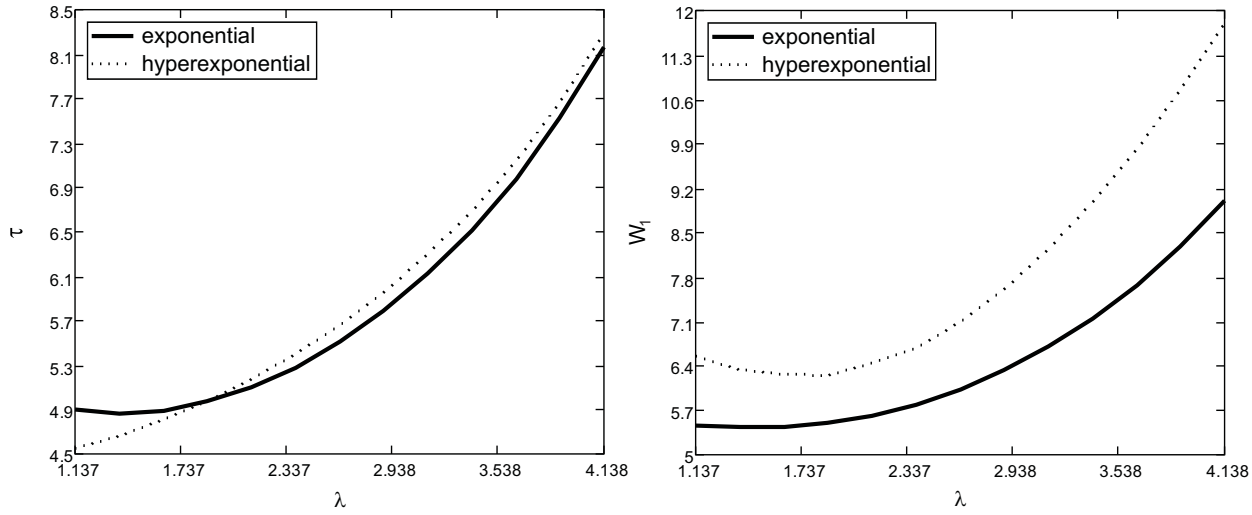

Рисунок 8.2. Зависимость τ и W_1 от λ

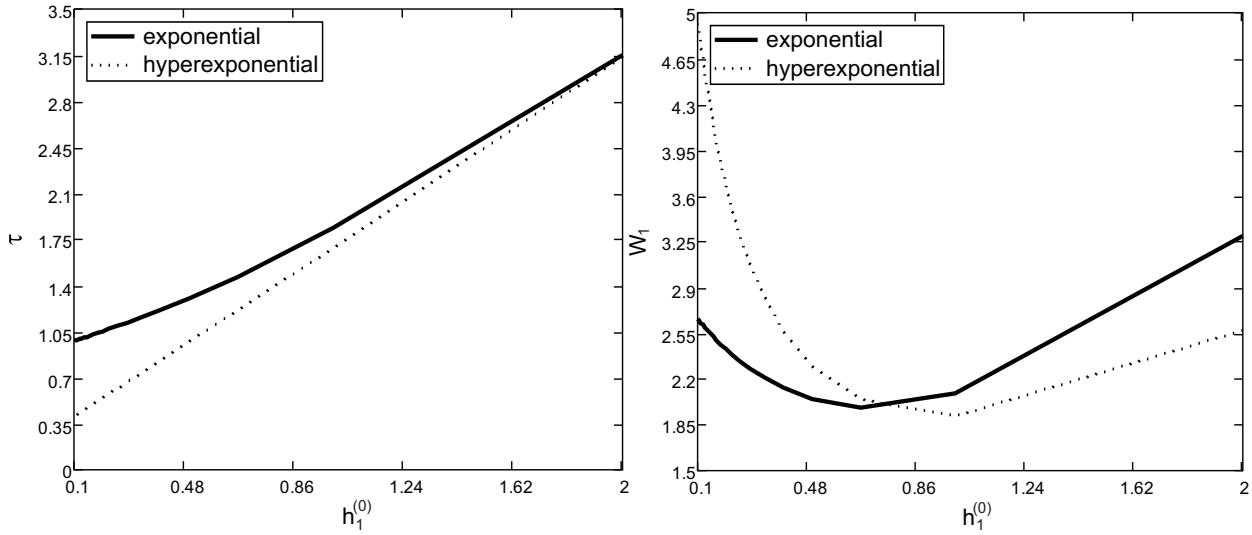

Рисунок 8.3. Зависимость τ и W_1 от от средней длительности отдыха типа-0

Вполне ожидаемым является то, что среднее время ожидания увеличивается, когда значение $h_1^{(0)}$ возрастает от 0.1 до 1. Чем дольше отдых, тем дольше время ожидания. Менее ожидаемым является тот факт, что W_1 увеличивается, когда среднее время отдыха типа-0 уменьшается от 1 до 0.1. Объяснение этому факту следующее. Когда $h_1^{(0)}$ становится маленьким, отдых после периода обслуживания короткий, и с высокой вероятностью запросы не поступают в этот период. Таким образом, отдыхи типа-1 и типа-2 происходят намного чаще. Их средняя продолжительность дольше, что приводит к более долгому времени ожидания поступающих запросов.

Как видно из результатов данного эксперимента, среднее время ожидания зависит от средней продолжительности отдыха типа-0. Такой же

вывод можно сделать, основываясь на других численных результатах, относительно существенной зависимости W_1 от средней продолжительности отдыхов разных типов. Таким образом, хороший выбор продолжительности отдыхов разных типов может свести к минимуму среднее время ожидания.

Таким образом, проблема выбора оптимальной продолжительности отдыхов имеет важное практическое значение.

В **эксперименте 8.3** еще раз иллюстрируется влияние корреляции входного потока, и выгода от оптимального выбора продолжительности отдыха типа-0.

Рисунок 8.4 отображает зависимость W_1 от средней продолжительности отдыха типа-0 $h_1^{(0)}$ для входных потоков M, IPP и $BMAP_{0.16}$, и $h_1^{(1)}$=2 и $h_1^{(2)}$=4.

Для кривой, соответствующей входному потоку M, среднее время ожидания равно 2.337, если $h_1^{(0)}$=0.1 и равно 5.097, если $h_1^{(0)}$=2. Оптимальное значение $h_1^{(0)}$ равно 0.4 и при этом среднее значение времени ожидания равно 1.804, что на 23 % меньше, чем 2.337 и на 65 % меньше, чем 5.097. Для IPP и $BMAP_{0.16}$ потоков, оптимальное значение $h_1^{(0)}$ снова 0.4, а оптимальное значение среднего времени ожидания равно 1.942 и 2.49, соответственно.

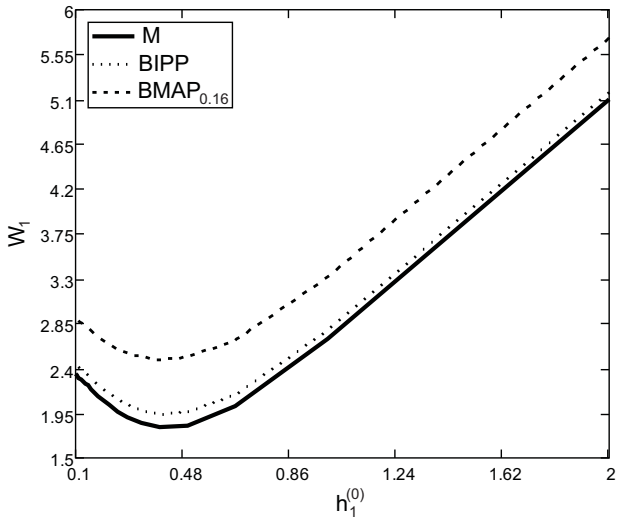

Рисунок 8.4. Зависимость W_1 от от средней длительности отдыха типа-0

8.3 Пример исследования модели циклического обслуживания с адаптивной дисциплиной просмотра буферов

В данном разделе мы проиллюстрируем путь применения результатов исследования системы массового обслуживания с адаптивными отдыхами для исследования поллинговой системы с динамической стратегией повторов. Для упрощения изложения будем предполагать, что потоки являются стационарными пуассоновскими и $R = 1$, т.е. имеется всего два типа отдыхов.

8.3.1 Постановка задачи

Имеется система массового обслуживания, состоящая из одного прибора и N буферов, имеющих бесконечную емкость. Входной поток в j-ю очередь (буфер, систему) является стационарным пуассоновским потоком с параметром $\lambda_j, \quad j = \overline{1,N}$. Время обслуживания запросов имеет функцию распределения $B_j(t)$ с преобразованием Лапласа – Стилтьеса $\beta_j(s) = \int\limits_0^\infty e^{-st} dB_j(t), \quad Re\ s > 0$, и конечными начальными моментами $b_j^{(k)} = \int\limits_0^\infty t^k dB_j(t), \ k \geq 1, \quad j = \overline{1,N}$. Время, затрачиваемое прибором на подключение к j-й системе, имеет функцию распределения $G_j(t)$ с преобразованием Лапласа – Стилтьеса (*ПЛС*) $g_j(s) = \int\limits_0^\infty e^{-st} dG_j(t), \ Re\ s > 0$, и конечными начальными моментами $g_j^{(k)} = \int\limits_0^\infty t^k dG_j(t), \ k \geq 1, \ j = \overline{1,N}$.

Очереди опрашиваются прибором в циклическом порядке. После подключения прибора к очереди происходит обслуживание всех запросов, находящихся в очереди на момент подключения. После окончания их обслуживания прибор переходит к следующей очереди. Запросы, пришедшие в систему за время их обслуживания, будут обслужены только при следующем подключении. Если в момент подключения к некоторой очереди она оказалась пустой, в следующем цикле данная очередь не опрашивается (пропускается). Если в текущем цикле все очереди подлежат пропуску, прибор уходит на отдых на время, имеющее распределение $F(t)$ с *ПЛС* $\varphi(s) = \int\limits_0^\infty e^{-st} dF(t), \ Re\ s > 0$, и конечными начальными моментами

$\varphi^{(k)} = \int\limits_0^\infty t^k dF(t),\ k \geq 1$. После отдыха все очереди подлежат опросу. Такая стратегия опроса названа адаптивной в [179].

Требуется найти среднее значение и второй начальный момент времени ожидания произвольного запроса в каждой очереди рассматриваемой системы.

8.3.2 Обсуждение возможных путей решения задачи

Близкая постановка задачи, но без пропуска очередей, оказавшихся пустыми в предыдущем цикле опроса, была исследована в [70], стр. 162-165.

Попытка использовать подход из [70] приводит к следующему. Для упрощения изложения будем считать, что прибор обязательно опрашивает все приборы по циклу. Для соответствия адаптивной стратегии опроса будем считать, что время на подключение к очереди, которая пропускается в данном цикле, нулевое и прибор немедленно переходит к следующей очереди. Для формализации адаптивной стратегии каждый буфер будем снабжать меткой j, которая принимает значение 1, если при следующем подключении к данному буферу он будет опрашиваться, и значение 0, если при следующем подключении к данному буферу он опрашиваться не будет. Метка буфера может изменяться в моменты после подключения к нему прибора.

Пусть t_k есть k-й момент переключения прибора, $k \geq 1$. Рассмотрим случайный процесс

$$\zeta_k = \{i_{t_k}, \eta_{t_k}^{(1)}, \ldots, \eta_{t_k}^{(N)}; j_{t_k}^{(1)}, \ldots, j_{t_k}^{(N)}\},\ k \geq 1,$$

где i_{t_k} – номер буфера, к которому происходит подключение в момент t_k, $\eta_{t_k}^{(i)}$ – число запросов в i-м буфере в этот момент, $j_{t_k}^{(i)}$ – метка буфера с номером i перед моментом t_k, $i = \overline{1, N}$.

Известно, что при выполнении условий

$$g_i^{(1)} < \infty,\ i = \overline{1, N}, \quad \sum_{i=1}^{N} \lambda_i b_i^{(1)} < 1, \tag{8.31}$$

рассматриваемая системы эргодична и существуют пределы

$$\pi(i, m_1, \ldots, m_N; j_1, \ldots, j_N) = \lim_{k\to\infty} P\{i_{t_k} = i, \eta_{t_k}^{(l)} = m_l,\ j_{t_k}^{(l)} = j_l,\ l = \overline{1, N}\},$$

$$i = \overline{1, N},\ m_l \geq 0,\ j_l = 0 \ \vee 1,\ l = \overline{1, N}.$$

Введем обозначения

$$\mathbf{j}=(j_1,\ldots,j_N),\ j_l=0\ \vee 1,\ ;\ \mathbf{z}=(z_1,\ldots,z_N),\ |z_l|\le 1,\ l=\overline{1,N},$$

$$\boldsymbol{\pi}(i,\mathbf{j},\mathbf{z})=\sum_{m_1=0}^{\infty}\cdots\sum_{m_N=0}^{\infty}\pi(i,m_1,\ldots,m_N;j_1,\ldots,j_N)z^{m_1}\ldots z^{m_N},$$

$$\mathbf{z}^{(i)}=(z_1,\ldots,z_{i-1},0,z_{i+1},\ldots,z_N),\ \hat{\beta}_i(\mathbf{z})=\beta_i(\sum_{m=1}^{N}\lambda_m(1-z_m)),$$

$$\hat{g}_i(\mathbf{z})=g_i(\sum_{m=1}^{N}\lambda_m(1-z_m)),\ i=\overline{1,N},\ \hat{\varphi}(\mathbf{z})=\varphi(\sum_{m=1}^{N}\lambda_m(1-z_m)).$$

Утверждение 8.1. *Производящие функции* $\boldsymbol{\pi}(i,\mathbf{j},\mathbf{z})$ *удовлетворяют рекуррентным соотношениям:*

$$\boldsymbol{\pi}(i+1,\mathbf{j},\mathbf{z})=\boldsymbol{\pi}(i,\mathbf{j}+\mathbf{e}_i,\mathbf{z}^{(i)})\hat{g}_i(\mathbf{z}), \tag{8.32}$$

если $j_i=0,\ i=\overline{1,N}$,

$$\boldsymbol{\pi}(i+1,\mathbf{j},\mathbf{z})=\boldsymbol{\pi}(i,\mathbf{j}-\mathbf{e}_i,\mathbf{z})+(\boldsymbol{\pi}(i,\mathbf{j},\mathbf{z}^{(i)}+\mathbf{e}_i\hat{\beta}_i(\mathbf{z}))-\boldsymbol{\pi}(i,\mathbf{j},\mathbf{z}^{(i)}))\hat{g}_i(\mathbf{z}), \tag{8.33}$$

если $j_i=1,\ i=\overline{1,N},\ \mathbf{j}\ne\mathbf{e}$,

$$\boldsymbol{\pi}(i+1,\mathbf{j},\mathbf{z})=\boldsymbol{\pi}(i,\mathbf{j}-\mathbf{e}_i,\mathbf{z})+(\boldsymbol{\pi}(i,\mathbf{j},\mathbf{z}^{(i)}+\mathbf{e}_i\hat{\beta}_i(\mathbf{z}))-\boldsymbol{\pi}(i,\mathbf{j},\mathbf{z}^{(i)}))\hat{g}_i(\mathbf{z})+ \tag{8.34}$$

$$+\boldsymbol{\pi}(i,\mathbf{e}_i,\mathbf{z}^{(i)})\hat{\varphi}(\mathbf{z}),$$

если $\mathbf{j}=\mathbf{e}$.

Похожая система уравнений для производящих функций $\boldsymbol{\pi}(i,\mathbf{z}),\ i=\overline{1,N}$, формально получающихся из рассматриваемых нами производящих функций $\boldsymbol{\pi}(i,\mathbf{j},\mathbf{z})$ путем удаления из записи вектора $\mathbf{j}$ меток буферов была получена в [70] для классической схемы циклического опроса с шлюзовым механизмом обслуживания. Эта система уравнений имела вид

$$\boldsymbol{\pi}(i+1,\mathbf{z})=\boldsymbol{\pi}(i,\mathbf{z}^{(i)}+\mathbf{e}_i\hat{\beta}_i(\mathbf{z}))\hat{g}_i(\mathbf{z}),\ i=\overline{1,N}. \tag{8.35}$$

Система (8.35) в [70] не решалась, а была использована для нахождения моментов распределения числа запросов в буферах в моменты подключения к ним прибора. Система уравнений (8.32)-(8.34) на порядок сложнее, чем система уравнений (8.35). Сложность состоит в том, что:

- Система (8.35), рассмотренная в [70], есть система относительно N производящих функций $\boldsymbol{\pi}(i, \mathbf{z}),\ i = \overline{1, N}$. А система (8.32)-(8.34) есть система относительно $N2^N$ производящих функций $\boldsymbol{\pi}(i, \mathbf{j}, \mathbf{z})$.
- Уравнения системы (8.35) представляют собой систему функциональных уравнений. Эти уравнения содержат неизвестные производящие функции $\boldsymbol{\pi}(i, \mathbf{z}),\ i = \overline{1, N}$, и эти же функции, в которых i-й аргумент z_i заменен на аргумент $\hat{\beta}_i(\mathbf{z})$. При подстановке в систему уравнений (8.35) в качестве вектора $\mathbf{z}$ вектора $\mathbf{e}$, получаем, что $z_i = \hat{\beta}_i(\mathbf{z}) = 1$ и функциональность уравнения в точке $\mathbf{z} = \mathbf{e}$ исчезает. Система (8.32)-(8.34) также представляет собой систему функциональных уравнений. Кроме большей размерности системы, отмеченной выше, эти уравнения сложнее, чем уравнения в (8.35). Наряду с неизвестными производящими функции $\boldsymbol{\pi}(i, \mathbf{j}), \mathbf{z}),\ i = \overline{1, N}$, в них входят не только эти же функции, в которых i-й аргумент z_i заменен на аргумент $\hat{\beta}_i(\mathbf{z})$, как и в (8.35), но и функции, в которых i-й аргумент z_i заменен на аргумент 0. Появление таких функций связано с различием поведения системы после моментов подключения, в которых буфер оказался непустым или пустым, обусловленным адаптивным механизмом опроса.

Хорошо известно, что функциональные уравнения нескольких переменных, в которые входит аргумент $\mathbf{z}$ и более чем один из аргументов типа $\mathbf{z}^{(i)}$, удается решить крайне редко. Поскольку мы имеем целую систему таких уравнений, отягощенную также присутствием функциональности по аргументу $\mathbf{j}$, решить эту систему не представляется возможным.

В такой ситуации представляется целесообразным использование упомянутого выше подхода к исследованию систем с циклическим опросом буферов, заключающегося в декомпозиции системы на ряд отдельных моделей систем массового обслуживания (СМО) с отдыхами приборов и последующего агрегирования характеристик систем, используя некоторую разумную эвристическую "сшивку" параметров и характеристик отдельных моделей СМО с отдыхами приборов.

Заметим, что при данном адаптивном механизме опроса, в отличие от классического механизма, достаточно сложными явились обе задачи: (а) исследование СМО с отдыхами прибора при шлюзовом доступе и зависимостью продолжительности отдыха от занятости или пустого состояния буфера в предыдущий момент подключения прибора и (б) подбор распределений времени отдыха в такой СМО, адекватно учитывающих тот факт,

что отдых данного прибора на самом деле является периодом, когда прибор опрашивает и обслуживает другие буферы системы.

8.3.3 Модель системы обслуживания с отдыхами прибора при шлюзовом доступе и зависимостью продолжительности отдыха от занятости буфера в предыдущий момент окончания отдыха

Рассмотрим следующую модель СМО с отдыхами прибора, описывающую процесс обслуживания запросов, поступающих в произвольный буфер системы с опросом. Предполагаем, что емкость буфера бесконечна. Будем использовать обозначения для параметров СМО, введенные выше. Для упрощения записи будем временно (в рамках данного подраздела) опускать индекс, указывающий номер системы, то есть вместо интенсивности входного потока λ_j, распределения времени обслуживания $B_j(t)$, его *ПЛС* $\beta_j(s)$ и начальных моментов $b_j^{(k)},\ j = \overline{1,N}$, будем использовать обозначения λ, $B(t)$, $\beta(s)$, $b^{(k)}$.

Предполагаем, что дисциплина обслуживания шлюзовая – прибор обслуживает только те запросы, которые находились в системе в момент окончания отдыха. Остальные запросы обслуживаются после следующего отдыха. Длительность отдыха предполагается различной в зависимости от того, не был ли буфер пуст в предыдущий момент окончания отдыха. Если буфер не был пуст, длительность следующего отдыха характеризуется функцией распределения $H(t)$ с *ПЛС* $h(s) = \int\limits_0^\infty e^{-st}H(t),\ \ Re\ s > 0$, и конечными начальными моментами $h^{(k)} = \int\limits_0^\infty t^k dH(t),\ k \geq 1$. Если же буфер был пуст, длительность следующего отдыха характеризуется функцией распределения $\tilde{H}(t)$ с *ПЛС* $\tilde{h}(s) = \int\limits_0^\infty e^{-st}\tilde{H}(t),\ \ Re\ s > 0$, и конечными начальными моментами $\tilde{h}^{(k)} = \int\limits_0^\infty t^k d\tilde{H}(t),\ k \geq 1$.

Исследуем данную СМО. Пусть t_k есть k-й момент окончания отдыха, $k \geq 1$, и i_{t_k} есть число запросов в буфере в момент t_k. Несложно видеть, что процесс $i_{t_k},\ k \geq 1$, является цепью Маркова с дискретным временем и ее одношаговые вероятности переходов

$$p_{i,j} = P\{i_{t_{k+1}} = j | i_{t_k} = i\},\ i,j \geq 0,$$

имеют вид:

$$p_{i,j} = \sum_{l=0}^{j} a_l^{(i)} y_{j-l},\ i > 0,\ j \geq 0,$$

$$p_{0,j} = \tilde{y}_j,\ j \geq 0,$$

где

$$a_l^{(i)} = \int_0^\infty \frac{(\lambda t)^l}{l!} e^{-\lambda t} dB^{(*i)}(t),$$

$$y_l = \int_0^\infty \frac{(\lambda t)^l}{l!} e^{-\lambda t} dH(t),\ \tilde{y}_l = \int_0^\infty \frac{(\lambda t)^l}{l!} e^{-\lambda t} d\tilde{H}(t),\ l \geq 0,$$

где $B^{(*i)}(t)$ есть свертка i-го порядка распределения $B(t)$.

Можно показать, что при выполнении условия $\rho = \lambda b^{(1)} < 1$ существуют стационарные вероятности состояний системы

$$q_j = \lim P\{i_{t_k} = j\},\ j \geq 0.$$

Эти вероятности удовлетворяют системе линейных алгебраических уравнений

$$q_j = q_0 \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} d\tilde{H}(t) +$$

$$+ \sum_{i=1}^{\infty} \sum_{l=0}^{j} q_i \int_0^\infty \frac{(\lambda t)^l}{l!} e^{-\lambda t} dB^{(*i)}(t) \int_0^\infty \frac{(\lambda t)^{j-l}}{(j-l)!} e^{-\lambda t} dH(t),\ j \geq 0. \qquad (8.36)$$

Умножая уравнения системы (8.36) на соответствующие степени z и суммируя, легко убедиться, что производящая функция

$$Q(z) = \sum_{j=0}^{\infty} q_j z^j,\ |z| \leq 1,$$

этих вероятностей удовлетворяет функциональному уравнению

$$Q(z) = (Q(\beta(\lambda - \lambda z)) - q_0) h(\lambda - \lambda z) + q_0 \tilde{h}(\lambda - \lambda z). \qquad (8.37)$$

Уравнение (8.37) включает функцию $Q(x)$ при аргументе x, равном z, $\beta(\lambda - \lambda z)$ и 0, соответственно.

Задача решения функционального уравнения является весьма сложной. Но это уравнение уже было решено в статье С. Сумиты [173] и соответствующий результат включен в первый том монографии Х. Такаги [214] (стр. 223-225). При решении С. Сумита использовал технику решения функциональных уравнений типа (8.37), описанную в книге М. Кучмы [156].

Для этого вводится последовательность функций $\eta_j(z),\ j \geq 0$, задаваемых рекуррентным образом:

$$\eta_0(z) = z,\ \eta_{j+1}(z) = \beta(\lambda - \lambda\eta_j(z)),\ j \geq 0. \tag{8.38}$$

С. Сумита доказал, что при выполнении условия $\lambda b_1 < 1$ последовательность функций $\eta_j(z),\ j \geq 0$, равномерно сходится к 1 при всех $z,\ 0 \leq z \leq 1$.

Подставляя в уравнение (8.37) в качестве аргумента z величину $\eta_j(z)$, получаем уравнение

$$Q(\eta_j(z)) = Q(\eta_{j+1}(z))h(\lambda - \lambda\eta_j(z)) + q_0(\tilde{h}(\lambda - \lambda\eta_j(z)) - h(\lambda - \lambda\eta_j(z))). \tag{8.39}$$

Повторяя эту операцию при $j = 0, 1, \ldots, n-1$, получаем соотношение

$$Q(z) = Q(\eta_n(z)) \prod_{j=0}^{n-1} h(\lambda - \lambda\eta_j(z)) +$$

$$+q_0 \sum_{j=0}^{n-1} (\tilde{h}(\lambda - \lambda\eta_j(z)) - h(\lambda - \lambda\eta_j(z))) \prod_{k=0}^{j-1} h(\lambda - \lambda\eta_k(z)). \tag{8.40}$$

Здесь понимается, что $\prod\limits_{k=a}^{b} c_k = 1$, если $b < a$.

Устремляя в соотношении (8.40) величину n к бесконечности и учитывая условие нормировки $Q(1) = 1$ и равномерную сходимость последовательности функций $\eta_j(z),\ j \geq 0$, к 1 при всех $z,\ 0 \leq z \leq 1$, получаем уравнение

$$Q(z) = \prod_{j=0}^{\infty} h(\lambda - \lambda\eta_j(z)) + q_0 \sum_{j=0}^{\infty} (\tilde{h}(\lambda - \lambda\eta_j(z)) - h(\lambda - \lambda\eta_j(z))) \prod_{k=0}^{j-1} h(\lambda - \lambda\eta_k(z)). \tag{8.41}$$

Подставляя в это соотношение $z = 0$, получаем следующее выражение для вероятности q_0:

$$q_0 = \frac{\prod\limits_{j=0}^{\infty} h(\lambda - \lambda z_j)}{1 - \sum\limits_{j=0}^{\infty} (\tilde{h}(\lambda - \lambda z_j) - h(\lambda - \lambda z_j)) \prod\limits_{k=0}^{j-1} h(\lambda - \lambda z_k)}, \tag{8.42}$$

где числа $z_j = \eta_j(0),\ j \geq 0$, задаются рекурсией

$$z_0 = 0,\ z_{j+1} = \beta(\lambda - \lambda z_j),\ j \geq 0. \tag{8.43}$$

Соотношения (8.38), (8.41), (8.42) и (8.43) полностью определяют (в принципе) распределение вероятностей $q_j,\ j \geq 0$.

При решении задачи нахождения характеристик системы с адаптивным пуллингом нам понадобятся формулы для значений первых трех производных $Q^{'}(1),\ \ Q^{(''')}(1),\ Q^{(''')}(1)$ производящей функции $Q(z)$ в точке $z = 1$, которые можно получить, используя формулу (8.41).

В формулу (8.41) входят *ПЛС* $(\tilde{h}(\lambda - \lambda\eta_j(z)),\ h(\lambda - \lambda\eta_j(z))$, где функции $\eta_j(z),\ j \geq 0$, заданы рекурсией (8.38). Поэтому для вывода формул для значений первых трех производных $Q^{'}(1),\ Q^{(''')}(1),\ Q^{(''')}(1)$ производящей функции $Q(z)$ необходимо предварительно вывести формулы для первых трех производных функций $\eta_j(z),\ j \geq 0$, в точке $z = 1$:

$$\eta_j(1) = 1,\ j \geq 0,$$

$$\eta_j{}'(1) = \rho^j,\ j \geq 0,$$

$$\eta_j{}''(1) = \lambda^2 b^{(2)} \rho^{j-1} \frac{1 - \rho^j}{1 - \rho},\ j \geq 0,$$

$$\eta_j{}'''(1) = \lambda^3 b^{(3)} \rho^{j-1} \frac{1 - \rho^{2j}}{1 - \rho^2} + 3\lambda^4 (b^{(2)})^2 \rho^{j-1} \frac{(1 - \rho^{j-1})(1 - \rho^j)}{(1 - \rho)(1 - \rho^2)},\ j \geq 0.$$

Используя эти формулы, можно получить следующие выражения для первых трех производных функций $h(\lambda - \lambda\eta_j(z)),\ j \geq 0$, в точке $z = 1$:

$$h(\lambda - \lambda\eta_j(z))|_{z=1} = 1,\ j \geq 0,$$

$$(h(\lambda - \lambda\eta_j(z)))'|_{z=1} = \lambda h^{(1)} \rho^j,\ j \geq 0,$$

$$(h(\lambda - \lambda\eta_j(z)))''_{z=1} = \lambda^2 h^{(2)} \rho^{2j} + \lambda^3 h^{(1)} b^{(2)} \rho^{j-1} \frac{1 - \rho^j}{1 - \rho},\ j \geq 0,$$

$$(h(\lambda-\lambda\eta_j(z)))^{'''}_{z=1}=\lambda^3h^{(3)}\rho^{3j}+3\lambda^4h^{(2)}b^{(2)}\rho^{j-1}\frac{1-\rho^j}{1-\rho}+$$

$$+\lambda h^{(1)}\rho^{j-1}\left[\lambda^3b^{(3)}\frac{1-\rho^{2j}}{1-\rho^2}+3\lambda^4(b^{(2)})^2\frac{(1-\rho^{j-1})(1-\rho^j)}{(1-\rho)(1-\rho^2)}\right],\ j\geq 0.$$

Используя эти вспомогательные формулы, после длительных и громоздких вычислений можно получить следующие формулы для значений первых трех производных $Q^{'}(1),\ Q^{(''')}(1),\ Q^{(''')}(1)$ производящей функции $Q(z)$ в точке $z=1$:

$$Q^{'}(1)=\frac{\lambda((1-q_0)h^{(1)}+q_0\tilde{h}^{(1)})}{1-\rho}=\frac{\lambda h^{(1)}}{1-\rho}+q_0\frac{\lambda(\tilde{h}^{(1)}-h^{(1)})}{1-\rho},\qquad(8.44)$$

$$Q^{''}(1)=\frac{\lambda^2((1-q_0)h^{(2)}+q_0\tilde{h}^{(2)})}{1-\rho^2}+\frac{\lambda^2((1-q_0)h^{(1)}+q_0\tilde{h}^{(1)})(\lambda b^{(2)}+2\rho h^{(1)})}{(1-\rho)(1-\rho^2)}=$$

$$=\frac{\lambda^2h^{(2)}}{1-\rho^2}+\frac{\lambda^3h^{(1)}b^{(2)}+(\lambda h^{(1)})^22\rho}{(1-\rho)(1-\rho^2)}+\qquad(8.45)$$

$$q_0\left[\frac{\lambda^2(\tilde{h}^{(2)}-h^{(2)})}{1-\rho^2}+(\tilde{h}^{(1)}-h^{(1)})\frac{\lambda^3b^{(2)}+2\rho\lambda^2h^{(1)})}{(1-\rho)(1-\rho^2)}\right],$$

$$Q^{'''}(1)=\frac{\lambda^3h^{(3)}}{1-\rho^3}+\frac{\lambda^4h^{(1)}b^{(3)}}{(1-\rho)(1-\rho^3)}+\frac{3\rho\lambda^4h^{(2)}b^{(2)}+3\rho(1+2\rho)\lambda^3h^{(1)}h^{(2)}}{(1-\rho^2)(1-\rho^3)}+\qquad(8.46)$$

$$+\frac{3\rho\lambda^5h^{(1)}(b^{(2)})^2+3\lambda^4(1+2\rho^2)(h^{(1)})^2b^{(2)}}{(1-\rho)(1-\rho^2)(1-\rho^3)}+\frac{6\rho^3(\lambda h^{(1)})^3}{(1-\rho)(1-\rho^2)(1-\rho^3)}+$$

$$+q_0\left[\frac{\lambda^3(\tilde{h}^{(3)}-h^{(3)})}{1-\rho^3}+\frac{\lambda^4(\tilde{h}^{(1)}-h^{(1)})b^{(3)}}{(1-\rho)(1-\rho^3)}+\frac{3\rho\lambda^4(\tilde{h}^{(2)}-h^{(2)})b^{(2)}}{(1-\rho^2)(1-\rho^3)}+\right.$$

$$+\frac{3\rho^2\lambda^3(\tilde{h}^{(2)}-h^{(2)})h^{(1)}}{(1-\rho^2)(1-\rho^3)}+\frac{3\rho(1+\rho)\lambda^3(\tilde{h}^{(1)}-h^{(1)})h^{(2)}}{(1-\rho^2)(1-\rho^3)}+$$

$$+\frac{3\rho\lambda^5(\tilde{h}^{(1)}-h^{(1)})(b^{(2)})^2+3\lambda^4(1+2\rho^2)((\tilde{h}^{(1)}-h^{(1)})h^{(1)}b^{(2)}}{(1-\rho)(1-\rho^2)(1-\rho^3)}+$$

$$\left.+\frac{6\rho^3\lambda^3(h^{(1)})^2(\tilde{h}^{(1)}-h^{(1)})}{(1-\rho)(1-\rho^2)(1-\rho^3)}\right],$$

Отметим трудоемкость получения этих формул и следующий факт. В книге [215] приведены формулы для значений первых двух производных $Q^{'}(1),\ Q^{(''')}(1)$ производящей функции $Q(z)$ в точке $z=1$. Первая из них

(формула (5.77a) на стр. 224) есть формула (8.44). Формула (5.77b) на стр. 225 для $Q^{''}(1)$ неверна. Она дана в виде

$$Q^{''}(1) = \frac{\lambda^2((1-q_0)h^{(2)} + q_0\tilde{h}^{(2)})}{(1-\rho)(1-\rho^2)} + \frac{\lambda^2((1-q_0)h^{(1)} + q_0\tilde{h}^{(1)})(\lambda b^{(2)} + 2\rho h^{(1)})}{1-\rho^2},$$

в то время как правильное выражение имеет вид (8.45). Если распределения длин отдыхов $H(t)$ и $\tilde{H}(t)$ одинаковые, то для контроля вычислений величины $Q^{'''}(1)$ можно было бы попытаться использовать формулу (5.22b) на стр. 208 книги [215]. Выражение до слагаемого с множителем q_0 в формуле (8.46) должно совпадать с (5.22b). Однако эта формула (5.22b) содержит ошибку. В ней пропущено слагаемое $\frac{6\rho^3(\lambda h^{(1)})^3}{(1-\rho)(1-\rho^2)(1-\rho^3)}$, которое имеется в формуле (8.46).

Отметим, что если сомножители типа $\tilde{h}^{(r)} - h^{(r)}$ в скобке при величине q_0 в (8.45) заменить на $h^{(r)}$, $r = 1, 2, 3$, то выражение в скобке при q_0 становится равным выражению до скобки с сомножителем q_0.

Величина $Q^{'}(1)$ задает среднее число L_1 запросов в системе в моменты окончания отдыхов

$$L_1 = Q^{'}(1), \tag{8.47}$$

второй начальный момент L_2 распределения числа запросов в системе в моменты окончания отдыхов задается формулой

$$L_2 = Q^{''}(1) + Q^{'}(1), \tag{8.48}$$

третий начальный момент L_3 распределения числа запросов в системе в моменты окончания отдыхов задается формулой

$$L_3 = Q^{'''}(1) + 3L_2 - 2L_1. \tag{8.49}$$

Используя полученные выражения, можно получить начальные моменты $\hat{\psi}^{(r)}$, $r = 1, 2, 3$, распределения длительности ζ обслуживания прибором запросов данной системы между последовательными отдыхами прибора. Число запросов данной системы, обслуженных между последовательными отдыхами прибора, является случайной величиной τ с начальными моментами L_1, L_2, L_3, а время обслуживания k-го запроса между последовательными отдыхами прибора, является случайной величиной ξ_k с начальными моментами $b^{(r)}$, $r = 1, 2, 3$.

Очевидно, что

$$\zeta = \sum_{k=1}^{\tau} \xi_k, \tag{8.50}$$

причем одинаково распределенные случайные величины ξ_k независимы между собой и не зависят от случайной величины τ. Таким образом, случайная величина ζ равна сумме случайного числа случайных величин. Используя аппарат условных математических ожиданий, можно получить следующие формулы для моментов $\hat{\psi}^{(r)},\ r = 1, 2, 3$, распределения случайной величины ζ:

$$\hat{\psi}^{(1)} = b^{(1)} L_1, \tag{8.51}$$

$$\hat{\psi}^{(2)} = b^{(2)} L_1 + (b^{(1)})^2 (L_2 - L_1), \tag{8.52}$$

$$\hat{\psi}^{(3)} = L_1 b^{(3)} + 3 b^{(1)} b^{(2)} (L_2 - L_1) + (b^{(1)})^3 (L_3 - 3L_2 + 2L_1). \tag{8.53}$$

Формула (8.51) широко известна в вероятностной литературе под названием тождества Вальда.

Обозначим через $\psi^{(r)},\ r = 1, 2, 3$, условные начальные моменты распределения длительности ζ обслуживания прибором запросов данной системы между последовательными отдыхами прибора при условии, что система была непуста. Очевидно, что условные начальные моменты $\psi^{(r)},\ r = 1, 2, 3$, выражаются через начальные моменты $\hat{\psi}^{(r)},\ r = 1, 2, 3$, следующим образом:

$$\psi^{(r)} = \frac{\hat{\psi}^{(r)}}{1 - q_0},\ r = 1, 2, 3.$$

Итак, формулы (8.41)-(8.43) задают стационарное распределение $q_i,\ i \geq 0$, числа запросов в системе в моменты подключения к ней прибора, а формулы (8.44)-(8.49) задают первые три момента этого распределения.

Далее получим стационарное распределение вероятностей состояний системы в произвольный момент времени, распределение времени ожидания и моменты распределения вероятностей времени ожидания запросов в буферах рассматриваемой системы. В частности, выведем следующие формулы, которые понадобятся нам в следующем разделе при исследовании системы с адаптивным пуллингом:

Формула для среднего времени W_1 ожидания требования в буфере следующая:

$$W_1 = \frac{(1 - q_0) h^{(2)} + q_0 \tilde{h}^{(2)}}{2((1 - q_0) h^{(1)} + q_0 \tilde{h}^{(1)})} + \frac{\lambda b^{(2)} + 2\rho h^{(1)}}{2(1 - \rho)}. \tag{8.54}$$

(она приведена в статье [173] и книге [215]).

Формула для второго начального момента W_2 времени ожидания требования в буфере имеет следующий вид:

$$W_2 = \frac{\rho(\lambda b^{(2)} + 2\rho h^{(1)})(\lambda b^{(2)} + \rho h^{(1)})}{(1-\rho)(1-\rho^2)} + \frac{\rho((1-q_0)h^{(2)} + q_0\tilde{h}^{(2)})(\lambda b^{(2)} + \rho h^{(1)})}{(1-\rho^2)((1-q_0)h^{(1)} + q_0\tilde{h}^{(1)})} +$$
$$+ \frac{\lambda b^{(3)} + 3\lambda b^{(2)} h^{(1)} + 3\rho h^{(2)}}{3(1-\rho)} + \frac{(1-q_0)h^{(3)} + q_0\tilde{h}^{(3)}}{3((1-q_0)h^{(1)} + q_0\tilde{h}^{(1)})} +$$
$$+ \frac{\lambda b^{(2)}(\lambda b^{(2)} + 2\rho h^{(1)})}{2(1-\rho^2)} + \frac{\lambda b^{(2)}((1-q_0)h^{(2)} + q_0\tilde{h}^{(2)})}{2(1+\rho)((1-q_0)h^{(1)} + q_0\tilde{h}^{(1)})}. \qquad (8.55)$$

Обозначим через $\pi^{(0)}(i,l)$ вероятность того, что в произвольный момент времени прибор занят обслуживанием запросов, в системе находится l запросов и на входе в систему ожидают i запросов, $i, l \geq 0$, через $\pi^{(1)}(i,0)$ вероятность того, что в произвольный момент времени прибор находится на отдыхе, длительность которого имеет функцию распределения $H(t)$ (обычном отдыхе) и на входе в систему ожидают i запросов, через $\pi^{(2)}(i,0)$ вероятность того, что в произвольный момент времени прибор находится на отдыхе, длительность которого имеет функцию распределения $\tilde{H}(t)$ (особом отдыхе) и на входе в систему ожидают i запросов, $i \geq 0$.

Утверждение 8.2. *Вероятности* $\pi^{(0)}(i,l)$, $\pi^{(k)}(i,0)$, $k = 1,2$, $i \geq 0$, *высчитываются следующим образом:*

$$\pi^{(0)}(i,l) = \tau^{-1}\left[q_l\hat{f}_i + \sum_{j=l+1}^{\infty} q_j \sum_{k=0}^{i} f_k^{(*(j-l))}\hat{f}_{i-k}\right],$$

$$\pi^{(1)}(i,0) = \tau^{-1}\left[\sum_{l=1}^{\infty} q_l \sum_{k=0}^{i} f_k^{(*(l))}\hat{h}_{i-k}\right],$$

$$\pi^{(2)}(i,0) = \tau^{-1}q_0\hat{\tilde{h}}_i,\ i \geq 0,$$

где величина τ *среднего времени между моментами окончания отдыхов считается по формуле*

$$\tau = \frac{(1-q_0)h_1 + q_0\tilde{h}_1}{1-\rho},$$

а остальные величины определяются следующим образом:

$$\hat{f}_i = \int_0^{\infty} \frac{(\lambda t)^i}{i!} e^{-\lambda t}(1 - B(t))dt,\ \ \hat{h}_i = \int_0^{\infty} \frac{(\lambda t)^i}{i!} e^{-\lambda t}(1 - H(t))dt,$$

$$\hat{\tilde{h}}_i = \int_0^\infty \frac{(\lambda t)^i}{i!} e^{-\lambda t}(1-\tilde{H}(t))dt, \quad f_i^{(*(l))} = \int_0^\infty \frac{(\lambda t)^i}{i!} e^{-\lambda t} dB^{(*l)}(t),$$

где $B^{(*l)}(t)$ *есть свертка* l*-го порядка распределения* $B(t)$, $l \geq 1$.

Введем в рассмотрение производящие функции

$$\Pi^{(0)}(z,y) = \sum_{i=0}^{\infty}\sum_{l=1}^{\infty} \pi^{(0)}(i,l) z^i y^l,$$

$$\Pi^{(k)}(z) = \sum_{i=0}^{\infty} \pi^{(k)}(i,0) z^i, \ k = 1,2, \ |z| \leq 1.$$

Утверждение 8.3. *Производящие функции* $\Pi^{(0)}(z,y)$, $\Pi^{(k)}(z)$, $k = 1,2$, *определяются следующим образом:*

$$\Pi^{(0)}(z,y) = \tau^{-1}\frac{1-\beta(\lambda(1-z))}{\lambda(1-z)}\frac{y}{y-\beta(\lambda(1-z))}(Q(y) - Q(\beta(\lambda(1-z)))),$$

$$\Pi^{(1)}(z) = \tau^{-1}\frac{1-h(\lambda(1-z))}{\lambda(1-z)}(Q(\beta(\lambda(1-z))) - q_0),$$

$$\Pi^{(2)}(z) = \tau^{-1}\frac{1-\tilde{h}(\lambda(1-z))}{\lambda(1-z)} q_0.$$

Следствие 8.8. *Среднее число* $L_1^{(0)}$ *запросов на входе в систему, среднее число* $L_2^{(0)}$ *запросов на обслуживании, среднее число* $L^{(0)}$ *запросов в системе в целом в периоды времени, когда прибор занят обслуживанием, определяются следующим образом:*

$$L_1^{(0)} = \frac{1}{2\lambda\tau}(Q''(1)\rho^2 + Q'(1)\lambda^2 b_2),$$

$$L_2^{(0)} = \tau^{-1} b_1 (Q'(1) + Q''(1)),$$

$$L^{(0)} = \tau^{-1}(Q'(1)(b_1 + \frac{\lambda^2 b_2}{2})) + \frac{1}{2}Q''(1)b_1(1+\rho)).$$

Следствие 8.9. *Производящая функция* $P(z)$ *числа запросов на входе в систему в произвольный момент времени выражается через производящую функцию* $Q(z)$ *числа запросов на входе в систему в момент окончания отдыха следующим образом:*

$$P(z) = \Pi^{(0)}(z,1) + \Pi^{(1)}(z) + \Pi^{(2)}(z) = \frac{1-Q(z)}{\lambda\tau(1-z)}.$$

Пусть $W(x),\ x \geq 0$, есть функция распределения времени ожидания произвольного запроса в системе, $w(s) = \int\limits_0^\infty e^{-sx} dW(w)$ – ее *ПЛС*.

Очевидно, что

$$w(s) = w^{(0)}(s) + w^{(1)}(s) + w^{(2)}(s),$$

где $w^{(0)}(s)$ есть *ПЛС* распределения времени ожидания произвольного запроса, поступившего в систему, когда прибор был занят обслуживанием запросов, $w^{(1)}(s)$ есть *ПЛС* распределения времени ожидания произвольного запроса, поступившего в систему, когда прибор находился на обычном отдыхе, $w^{(2)}(s)$ есть *ПЛС* распределения времени ожидания произвольного запроса, поступившего в систему, когда прибор находился на особом отдыхе.

Используя вероятностную интерпретацию *ПЛС*, можно убедиться в справедливости следующих формул:

$$w^{(0)}(s) = \tau^{-1} h(s) \frac{Q(\beta(\lambda(1-\beta(s)))) - Q(\beta(s))}{s - \lambda(1-\beta(s))},$$

$$w^{(1)}(s) = \tau^{-1} (Q(\beta(\lambda(1-\beta(s)))) - q_0) \frac{h(\lambda(1-\beta(s))) - h(s)}{s - \lambda(1-\beta(s))},$$

$$w^{(2)}(s) = \tau^{-1} q_0 \frac{\tilde{h}(\lambda(1-\beta(s))) - \tilde{h}(s)}{s - \lambda(1-\beta(s))}.$$

Из этих формул следует справедливость следующего результата.

Утверждение 8.4. *ПЛС $w(s)$ распределения времени ожидания произвольного запроса вычисляется следующим образом:*

$$w(s) = \frac{\tau^{-1}}{s - \lambda(1-\beta(s))} (Q(\beta(s))(1-h(s)) + q_0(h(s) - \tilde{h}(s))).$$

Дифференцируя эту функцию в точке $s = 0$, получаем формулы (8.54), (8.55).

Отметим, что если мы введем обозначение

$$v_k = (1-q_0) h^{(k)} + q_0 \tilde{h}^{(k)},\ k = 1, 2, 3,$$

то формулы (8.54) и (8.55) можно записать несколько короче:

$$W_1 = \frac{v_2}{2v_1} + \frac{\lambda b^{(2)} + 2\rho h^{(1)}}{2(1-\rho)},$$

$$W_2 = \frac{v_3}{3v_1} + \frac{\lambda b^{(3)} + 3\lambda b^{(2)} h^{(1)} + 3\rho h^{(2)}}{3(1-\rho)} + W_1 \left(\frac{\lambda b^{(2)}}{1-\rho} + \frac{2\rho^2 h^{(1)}}{1-\rho^2} \right).$$

8.3.4 Стыковка параметров моделей буферов

Как следует из формул (8.42)-(8.55), ключевую роль в вычислении характеристик рассмотренной системы обслуживания имеет вид *ПЛС* $h(s)$, $\tilde{h}(s)$ распределения длины отдыха после подключения прибора к непустому и пустому буферу соответственно. Для того чтобы использовать данную СМО для исследования системы с адаптивным опросом, необходимо оценить эти *ПЛС*. Очевидно, что они, в свою очередь, определяются временем, затрачиваемым прибором на обслуживание других буферов, и неизвестны априори. Требуется построить процедуру последовательных приближений для расчета первых двух начальных моментов ожидания требования в каждом из буферов на основе рассмотренной в предыдущем разделе СМО, на каждом из шагов которой производится уточнение вида неизвестных *ПЛС*.

В данном разделе будет предложена такая эвристическая процедура. В ней мы будем использовать величины $z_j^{(i)}$, $q_0^{(i)}$, $\psi_i^{(r)}$, $r = 1, 2, 3$, $W_r^{(i)}$, $r = 1, 2$, $i = \overline{1, N}$, заданные формулами (8.42)-(8.55), в которых все *ПЛС* и соответствующие моменты снабжены верхним индексом i – номером соответствующего буфера.

Поскольку после подключения прибора к пустому буферу в следующем цикле просмотра данный буфер просматриваться не будет и подключение к нему произойдет через еще один цикл, логично считать, что время отсутствия подключения прибора к данному буферу практически равно удвоенному времени отсутствия подключения прибора к данному буферу после подключения, в момент которого буфер не был пуст, за вычетом одного времени подключения прибора к данному буферу. При этом, однако, надо учесть, что если в данном цикле опроса все буферы оказались пустыми, то в следующем цикле, следующем за принудительным простоем прибора, все буферы будут опрашиваться.

Отсюда следует, что разумным является считать, что

$$\tilde{h}_i(s) = \tilde{\chi}_i(s) g_i(s), \tag{8.56}$$

где

$$\tilde{\chi}_i(s) = (\chi_i(s))^2 + Q_i(\varphi(s + \lambda_i(1 - \beta_i(s)))r_i(s) - \chi_i(s)),$$

где функции $\chi_i(s)$ заданы формулой (8.62), приведенной ниже,

$$Q_i = \prod_{j=1, j\neq i}^{N} q_j,$$

$$r_i(s) = \prod_{j=1, j\neq i}^{N} (q_j + (1-q_j)\psi_j(s))g_j(s).$$

Отсюда следует, что начальные моменты распределения $\tilde{H}(t)$ и $H(t)$ длин отдыха в модели СМО, рассмотренной в предыдущем разделе, высчитываются следующим образом:

$$\tilde{h}_i^{(1)} = \tilde{\chi}_i^{(1)} + g_i^{(1)},$$
$$\tilde{h}_i^{(2)} = \tilde{\chi}_i^{(2)} + g_i^{(2)} + 2\tilde{\chi}_i^{(1)} g_i^{(1)},$$
$$\tilde{h}_i^{(3)} = \tilde{\chi}_i^{(3)} + g_i^{(3)} + 3\tilde{\chi}_i^{(2)} g_i^{(1)} + 3\tilde{\chi}_i^{(1)} g_i^{(2)},$$

где моменты $\tilde{\chi}_i^{(l)}$ вычисляем по следующим формулам:

$$\tilde{\chi}_i^{(1)} = 2\chi_i^{(1)} + Q_i(\hat{\varphi}^{(1)} + r_i^{(1)} - \chi_i^{(1)}),$$

$$\tilde{\chi}_i^{(2)} = 2\chi_i^{(2)} + 2(\chi_i^{(1)})^2 + Q_i(\hat{\varphi}^{(2)} + 2\hat{\varphi}^{(1)} r_i^{(1)} + r_i^{(2)} - \chi_i^{(2)}),$$
$$\tilde{\chi}_i^{(3)} = 2\chi_i^{(3)} + 6\chi_i^{(1)}\chi_i^{(2)} + Q_i(\hat{\varphi}^{(3)} + 3\hat{\varphi}^{(2)} r_i^{(1)} + 3\hat{\varphi}^{(1)} r_i^{(2)} + r_i^{(3)} - \chi_i^{(3)}),$$

где

$$\hat{\varphi}^{(1)} = \varphi^{(1)}(1+\rho_i), \;\; \rho_i = \lambda_i b_i^{(1)},$$
$$\hat{\varphi}^{(2)} = \varphi^{(2)}(1+\rho_i)^2 + \varphi^{(1)}\lambda_i b_i^{(2)},$$
$$\hat{\varphi}^{(3)} = \varphi^{(3)}(1+\rho_i)^3 + 3\varphi^{(2)}(1+\rho_i)\lambda_i b_i^{(2)} + \varphi^{(1)}\lambda_i b_i^{(3)},$$

а моменты $r_i^{(l)}$ вычисляем по следующим формулам:

$$r_i^{(1)} = \sum_{j=1,\, j\neq i}^{N} (q_j g_j^{(1)} + (1-q_j) a_j^{(1)}),$$

$$r_i^{(2)} = \sum_{j=1,\, j\neq i}^{N} (q_j g_j^{(2)} + (1-q_j) a_j^{(2)} + (q_j g_j^{(1)} + (1-q_j) a_j^{(1)}) \sum_{k=1,\, k\neq i, k\neq j}^{N} (q_k g_k^{(1)} + (1-q_k) a_k^{(1)})),$$

$$r_i^{(3)} = \sum_{j=1,\, j\neq i}^{N} (q_j g_j^{(3)} + (1-q_j) a_j^{(3)} + 2(q_j g_j^{(2)} + (1-q_j) a_j^{(2)}) \sum_{k=1,\, k\neq i, k\neq j}^{N} (q_k g_k^{(1)} + (1-q_k) a_k^{(1)}) +$$

$$+(q_j g_j^{(1)} + (1-q_j)a_j^{(1)}) \sum_{k=1,\ k\neq i, k\neq j}^{N} (q_k g_k^{(2)} + (1-q_k)a_k^{(2)} + (q_k g_k^{(1)} + (1-q_k)a_k^{(1)}) \times$$

$$\times \sum_{m=1,\ m\neq i, m\neq j, m\neq k}^{N} (q_m g_m^{(1)} + (1-q_m)a_m^{(1)}))).$$

Поэтому далее мы будем обсуждать проблему нахождения *ПЛС* $h(s)$.

На начальном шаге итерационной процедуры будем считать, что в каждом цикле происходит подключение прибора к каждому буферу и прибор обслуживает ровно по одному запросу из каждого буфера.

Из этого предположения следует, что на начальном шаге в качестве *ПЛС* $h_i(s)$ времени отдыха в модели i-го буфера следует взять

$$h_i(s) = \frac{\sigma_i}{s+\sigma_i},$$

где

$$(\sigma_i)^{-1} = \sum_{j=1,\ j\neq i}^{N} (b_j^{(1)} + g_j^{(1)}),\ i = \overline{1,N}.$$

Соответственно, начальные моменты распределения времени отдыха в модели i-го буфера следует взять в виде

$$h_i^{(1)} = \sigma_i^{-1},\ h_i^{(2)} = 2\sigma_i^{-2},\ h_i^{(3)} = 6\sigma_i^{-3}.$$

Используя такие *ПЛС* и начальные моменты распределения времени отдыха, проводим расчет величин $z_j^{(i)},\ q_0^{(i)},\ \psi_i^{(r)},\ r = 1,2,3,\ W_r^{(i)},\ r = 1,2,\ i = \overline{1,N}$, по формулам (8.42)-(8.55) в модели каждого i-го буфера, $i = \overline{1,N}$.

Отметим, что для контроля вычисления всех начальных моментов в процессе реализации данного алгоритма на ЭВМ можно использовать неравенство Ляпунова. Например, для моментов $\psi_i^{(r)},\ r = 1,2,3$, должны выполняться неравенства:

$$\psi_i^{(2)} \geq (\psi_i^{(1)})^2,\ \psi_i^{(3)} \geq (\psi_i^{(1)})^3,\ \psi_i^{(3)} \geq (\psi_i^{(2)})^{\frac{3}{2}}.$$

Анализируя вид начальных моментов $\psi_i^{(r)},\ r = 1,2,3$, непрерывного времени обслуживания запросов в i-м буфере, оцениваем *ПЛС* $\psi_i(s)$ распределения этого времени. При этом если оказалось, что величина $c_\psi = \frac{\psi_i^{(2)}}{(\psi_i^{(1)})^2}$, примерно равна 1, то считаем, что

$$\psi_i(s) = \frac{1}{1+s\psi_i^{(1)}}, \tag{8.58}$$

то есть распределение отдыха имеет показательное распределение.

Если величина c_ψ примерно равна $\frac{1}{k}$, где k – некоторое натуральное число, то считаем, что

$$\psi_i(s) = \left(\frac{1}{1 + s\psi_i^{(1)}}\right)^k, \tag{8.59}$$

то есть распределение отдыха имеет распределение Эрланга порядка k.

Если величина c_ψ больше 1, используем *ПЛС* $\psi_i(s)$ вида:

$$\psi_i(s) = p_i \frac{\mu_i^{(1)}}{\mu_i^{(1)} + s} + (1 - p_i)\frac{\mu_i^{(2)}}{\mu_i^{(2)} + s}. \tag{8.60}$$

Подбор параметров $p_i, \mu_i^{(1)}, \mu_i^{(2)}$ гиперэкспоненциального распределения с *ПЛС* (8.60) по трем начальным моментам $\psi_i^{(r)}$, $r = 1, 2, 3$, можно осуществить с помощью компьютерных программ.

В случае когда $c_\psi < 1$, но эта величина не равна примерно $\frac{1}{k}$, распределение может быть аппроксимировано распределением фазового типа. Если величина c_ψ очень близка к нулю, то распределение может быть аппроксимировано вырожденным распределением:

$$\psi_i(s) = e^{-\psi_i^{(1)} s}.$$

Итак, считаем, что *ПЛС* $\psi_i(s)$, $i = \overline{1, N}$, известны. Заметим, что не исключено, что для различных буферов *ПЛС* $\psi_i(s)$ будут иметь различные виды из (8.58)-(8.60).

Далее делаем упрощающее предположение, что вероятность того i-й буфер пуст в произвольный момент подключения, не зависит от текущего состояния других буферов. Это – критическое предположение в проводимом анализе. Оно, вообще говоря, не соответствует реальности (если какие-либо другие буферы пусты, данный буфер получает больше времени обслуживания прибором и, соответственно, вероятность того, что данный буфер пуст, выше, чем если бы все другие буферы не были пусты). Но отказ от этого предположения исключает возможность анализа рассматриваемой системы аналитическими методами.

При сделанном предположении можно перерасчитать *ПЛС* $h_i(s)$ времени отдыха в модели i-го буфера следующим образом:

$$h_i(s) = \chi_i(s) g_i(s), \tag{8.61}$$

где

$$\chi_i(s) = \prod_{j=1,\, j\neq i}^{N} (q_j + (1-q_j)\psi_j(s)g_j(s)). \qquad (8.62)$$

Вывод формул (8.61)-(8.62) основан на вероятностной интерпретации *ПЛС* как вероятности ненаступления за интересующее исследователя время катастрофы из стационарного пуассоновского потока с параметром s. Чтобы не наступило катастрофы за время отдыха прибора в модели i-го буфера, надо, чтобы ее не наступило за времена подключения к данному буферу (вероятность этого есть $g_i(s)$) и за времена подключения ко всем остальным буферам во время текущего цикла и времена обслуживания запросов из этих буферов (вероятность этого есть $\chi_i(s)$). Первое слагаемое в (8.62) учитывает тот факт, что катастрофа не наступает за время подключения к j-му буферу и обслуживания запросов из этого буфера с вероятностью 1, если подключения к буферу не производилось, и с вероятностью $\psi_j(s)g_j(s)$, если подключение производилось. Последнее слагаемое в (8.62) учитывает тот факт, что в ситуации, когда в начинающемся цикле не требуется подключения ни к одному из буферов, прибор уходит на отдых, длительность которого имеет *ПЛС* $\varphi(s)$.

Из формул (8.61), (8.62) очевидным образом следуют формулы для первых трех начальных моментов времени отдыха в модели i-го буфера:

$$h_i^{(1)} = \chi_i^{(1)} + g_i^{(1)},$$

$$h_i^{(2)} = \chi_i^{(2)} + g_i^{(2)} + 2\chi_i^{(1)} g_i^{(1)},$$

$$h_i^{(3)} = \chi_i^{(3)} + g_i^{(3)} + 3\chi_i^{(2)} g_i^{(1)} + 3\chi_i^{(1)} g_i^{(2)},$$

где

$$\chi_i^{(1)} = \sum_{j=1,\, j\neq i}^{N} (1-q_j) a_j^{(1)},$$

$$\chi_i^{(2)} = \sum_{j=1,\, j\neq i}^{N} (1-q_j) a_j^{(2)} + \sum_{j=1,\, j\neq i}^{N} (1-q_j) a_j^{(1)} \sum_{k=1,\, k\neq i, k\neq j}^{N} (1-q_k) a_k^{(1)},$$

$$\chi_i^{(3)} = \sum_{j=1,\, j\neq i}^{N} (1-q_j) a_j^{(3)} + 2 \sum_{j=1,\, j\neq i}^{N} (1-q_j) a_j^{(2)} \sum_{k=1,\, k\neq i, k\neq j}^{N} (1-q_k) a_k^{(1)} +$$

$$+ \sum_{j=1,\, j\neq i}^{N} (1-q_j) a_j^{(1)} \sum_{k=1,\, k\neq i, k\neq j}^{N} (1-q_k) a_k^{(1)} \sum_{m=1,\, m\neq i, m\neq j, m\neq k}^{N} (1-q_m) a_m^{(1)}, \qquad (8.63)$$

где для $m = \overline{1, M}$

$$a_m^{(1)} = g_m^{(1)} + \psi_m^{(1)},$$
$$a_m^{(2)} = g_m^{(2)} + \psi_m^{(2)} + 2g_m^{(1)}\psi_m^{(1)},$$
$$a_m^{(3)} = g_m^{(3)} + \psi_m^{(3)} + 3g_m^{(2)}\psi_m^{(1)} + 3g_m^{(1)}\psi_m^{(2)}.$$

Используя формулы (8.61)-(8.63) для *ПЛС* и начальных моментов распределения времени отдыха, проводим снова расчет величин $q_0^{(i)}$, $\psi_i^{(r)}$, $r = 1, 2, 3$, $W_r^{(i)}$, $r = 1, 2$, $i = \overline{1, N}$, по формулам (8.42)-(8.55) в модели каждого i-го буфера, $i = \overline{1, N}$.

Описанную процедуру последовательных приближений повторяем до тех пор, пока на последовательных итерациях не получим совпадения величин $q_0^{(i)}$, $\psi_i^{(r)}$, $r = 1, 2, 3$, $W_r^{(i)}$, $r = 1, 2$, $i = \overline{1, N}$, с желаемой точностью.

Полученные после остановки процедуры величины $W_r^{(i)}$, $r = 1, 2$, $i = \overline{1, N}$, дают решение поставленной задачи.

8.3.5 Результаты численных экспериментов

В данном подразделе приводятся численные эксперименты, иллюстрирующие работу алгоритма для систем с циклическим опросом, различным числом очередей и интенсивностью трафика в сравнении с результатами имитационного моделирования, полученными с помощью пакета GPSS World [168]. Объектом моделирования являлась региональная широковещательная беспроводная сеть, состоящая из нескольких приборов и одной базовой станции. Интенсивность поступления пакетов на приборы и интенсивность их обслуживания различны. Приборы опрашиваются циклически. Дисциплина обслуживания пакетов шлюзовая. Это значит, что обслуживаются только те пакеты, которые были в очереди в момент открытия шлюза. Входной поток предполагается пуассоновским, а времена обслуживания пакетов и инициализация опроса имеют экспоненциальное распределение.

Для имитационного моделирования предположим, что система функционирует в стационарном режиме, если при удваивании числа пакетов, проходящих через систему ни один из сравниваемых параметров не изменяется более чем на 0.5 %. В эксперименте через систему проходит более миллиона пакетов.

Рассмотрим три случая с различным числом очередей N.

Таблица 8.1. Система с двумя очередями

Параметры	A	S	Δ, %
$\lambda_1 = \lambda_2 = 0.321$, $\rho = 0.2$	0.289	0.268	7.8
$\lambda_1 = \lambda_2 = 0.5$, $\rho = 0.311$	0.392	0.358	9.5
$\lambda_1 = \lambda_2 = 0.803$, $\rho = 0.5$	0.659	0.601	9.7
$\lambda_1 = \lambda_2 = 1.28$, $\rho = 0.8$	1.73	1.93	10.4
$\varphi^{(1)} = 0.05$	0.392	0.358	9.5
$\varphi^{(1)} = 0.1$	0.417	0.384	8.6

Случай N = 2. Для начала рассмотрим симметричную систему с двумя очередями и средним временем обслуживания $b_1^{(1)} = b_2^{(1)} = 0.311$, среднее время переключения $s_1^{(1)} = s_2^{(1)} = 0.091$ и среднее время отдыха прибора $\varphi^{(1)} = 0.005$. Среднее время ожидания вычисляется с помощью алгоритма, разработанного в предыдущем подразделе (столбец "A"). Результаты моделирования (столбец "S") и относительная погрешность сравнения (столбец "Δ") приведены в таблице 8.1. Первый столбец содержит интенсивности потока запросов и соответствующие загрузки системы. Две последние строки таблицы содержат результаты, полученные для различных средних времен отдыха прибора $\varphi^{(1)}$ при условии $\lambda_1 = \lambda_2 = 0.5$.

Случай N = 3. Рассмотрим случай системы с тремя очередями и симметричным обслуживанием $b_i^{(1)} = 0.044$, $s_i^{(1)} = 0.1$, $i = \overline{1,3}$, $\varphi^{(1)} = 0.1$. Среднее время ожидания $W_i^{(1)}$, $i = \overline{1,3}$, вычисленное с помощью алгоритма и моделирования для различных интенсивностей входного потока запросов, представлено в таблице 8.2. Последние две строки содержат результаты для полностью симметричной системы (все λ_i, $i = \overline{1,3}$, одинаковые).

Случай N = 5. Наконец, рассмотрим случай пяти очередей с $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 0.5$, $\lambda_4 = 6$, $\lambda_5 = 0.5$, $b_i^{(1)} = 0.05$, $s_i^{(1)} = 0.05$, $i = \overline{1,5}$, $\varphi^{(1)} = 0.05$. Мы изменяем входную интенсивность с помощью умножения всех λ_i на α принимающие значения 0.285, 0.714, 1 и 1.143. Таким образом, загрузка системы ρ, изменяется от 0.2 до 0.8. Результаты приведены в таблице 8.3. Последние четыре строки представляют результаты для полностью симметричной системы с $\lambda_i = 2$ умноженной на одни и те же значения α.

Результаты для несимметричной системы ($b_1^{(1)} = 0.07$, $b_2^{(1)} = 0.015$, $b_3^{(1)} = 0.1$, $b_4^{(1)} = 0.025$, $b_5^{(1)} = 0.4$) приведены в таблице 8.4.

Ниже мы обсудим полученные результаты. Для симметричных систем

Таблица 8.2. Система с тремя очередями

Параметры	$W^{(1)}$	A	S	Δ, %
$\lambda_1 = 2.5$, $\lambda_2 = 6$,	$W_1^{(1)}$	0.342	0.365	6.3
$\lambda_3 = 0.5$,	$W_2^{(1)}$	0.335	0.361	7.2
$\rho = 0.4$	$W_3^{(1)}$	0.410	0.440	6.8
$\lambda_1 = 4.375$, $\lambda_2 = 10.5$,	$W_1^{(1)}$	0.658	0.698	5.7
$\lambda_3 = 0.875$,	$W_2^{(1)}$	0.781	0.834	3.0
$\rho = 0.7$	$W_3^{(1)}$	0.778	0.805	3.4
Симметричная система				
$\lambda_i = 3$, $i = \overline{1,3}$, $\rho = 0.4$	$W_i^{(1)}$	0.387	0.382	1.3
$\lambda_i = 5.25$, $i = \overline{1,3}$, $\rho = 0.7$	$W_i^{(1)}$	0.702	0.771	8.9

Таблица 8.3. Система с пятью очередями

Параметры	$W^{(1)}$	A	S	Δ, %	Параметры	$W^{(1)}$	A	S	Δ, %
	$W_1^{(1)}$	0.203	0.220	7.7		$W_1^{(1)}$	0.714	0.672	6.2
$\alpha = 0.285$	$W_2^{(1)}$	0.199	0.215	7.4	$\alpha = 1$	$W_2^{(1)}$	0.661	0.618	7.0
$\rho = 0.2$	$W_3^{(1)}$	0.205	0.222	7.7	$\rho = 0.7$	$W_3^{(1)}$	0.776	0.738	5.1
	$W_4^{(1)}$	0.197	0.203	3.0		$W_4^{(1)}$	0.705	0.679	3.8
	$W_5^{(1)}$	0.205	0.224	8.5		$W_5^{(1)}$	0.776	0.745	4.2
	$W_1^{(1)}$	1.036	0.974	6.4		$W_1^{(1)}$	0.374	0.398	6.0
$\alpha = 0.714$	$W_2^{(1)}$	0.353	0.370	4.6	$\alpha = 1.143$	$W_2^{(1)}$	0.967	0.934	3.5
$\rho = 0.5$	$W_3^{(1)}$	0.393	0.419	6.2	$\rho = 0.8$	$W_3^{(1)}$	1.153	1.080	6.8
	$W_4^{(1)}$	0.340	0.355	4.2		$W_4^{(1)}$	1.152	1.120	2.9
	$W_5^{(1)}$	0.393	0.422	6.9		$W_5^{(1)}$	1.153	1.090	5.8
Симметричная система									
$\rho = 0.2$	$W_i^{(1)}$	0.207	0.216	4.2	$\rho = 0.7$	$W_i^{(1)}$	0.759	0.686	10.6
$\rho = 0.5$	$W_i^{(1)}$	0.455	0.391	16.4	$\rho = 0.8$	$W_i^{(1)}$	1.003	1.040	3.6

Таблица 8.4. Система с пятью очередями и несимметричным обслуживанием

Параметры	$W^{(1)}$	A	S	Δ, %	Параметры	$W^{(1)}$	A	S	Δ, %
$\alpha = 0.4$, $\rho = 0.2$	$W_1^{(1)}$	0.251	0.250	0.4	$\alpha = 1$, $\rho = 0.5$	$W_1^{(1)}$	0.570	0.506	12.7
	$W_2^{(1)}$	0.248	0.244	1.6		$W_2^{(1)}$	0.535	0.475	12.7
	$W_3^{(1)}$	0.251	0.254	1.2		$W_3^{(1)}$	0.592	0.548	7.9
	$W_4^{(1)}$	0.244	0.228	7.0		$W_4^{(1)}$	0.516	0.455	13.4
	$W_5^{(1)}$	0.223	0.254	12.2		$W_5^{(1)}$	0.538	0.559	3.8
$\alpha = 0.6$, $\rho = 0.3$	$W_1^{(1)}$	0.318	0.314	1.3	$\alpha = 1.4$, $\rho = 0.7$	$W_1^{(1)}$	1.016	0.902	12.6
	$W_2^{(1)}$	0.311	0.302	3.0		$W_2^{(1)}$	0.901	0.831	8.4
	$W_3^{(1)}$	0.322	0.325	0.9		$W_3^{(1)}$	1.095	0.994	10.2
	$W_4^{(1)}$	0.305	0.281	8.5		$W_4^{(1)}$	0.938	0.896	4.7
	$W_5^{(1)}$	0.281	0.325	13.5		$W_5^{(1)}$	1.082	1.080	0.2

с $N = 2$ и $N = 3$ относительная погрешность сравнения растет с ростом загрузки системы ρ. Ситуация меняется в случае $N = 5$: результаты аппроксимации и результаты моделирования совпадают с точностью до 5 % для малых и больших значений загрузки, но при $\rho = 0.5$ и $\rho = 0.7$, погрешность становится неприемлемой (более, чем 10 %). Для системы с пятью очередями зависимость относительной погрешности от общей загрузки системы не совсем понятна. В случае несимметричной системы с пятью очередями, см. таблицы 8.3 и 8.4, результаты сложно объяснить. При несимметричном поступлении запросов, как представлено в таблице 8.3, совпадение увеличивается при росте ρ, но когда мы делаем обслуживание несимметричным, см. таблицу 1.4, совпадение можно наблюдать для очередей с относительно большими загрузками, а именно для очередей 4 и 5, в то время как результаты для остальных очередей иные (относительная погрешность увеличивается с ростом ρ_i, $i = \overline{1,3}$). Отметим, что в таблице 8.3, результаты для систем с количеством очередей 3 и 5 должны быть одинаковыми, поскольку очереди идентичны, но результаты моделирования отличаются примерно на 1 %, что объясняется погрешностями имитации.

Для полного численного анализа сравним две схемы опроса в системах, работающих в режиме сбора данных.

1. Базовая станция всегда циклически опрашивает станции абонентов, даже если некоторые из них были пустыми в предыдущем цикле (циклический опрос);

2. Базовая станция пропускает (не опрашивает) станции абонентов, которые были пустыми в предыдущем цикле (адаптивный циклический опрос).

Сравнение основывается на модели радиоячейки с параметрами $N = 4$, входными интенсивностями $\lambda_1 = \lambda_2 = 7000$, $\lambda_3 = \lambda_4 = \lambda$ изменяющимися от 1 до 500. Стоит отметить, что среднее время обслуживания пакетов, время переключения и т.д., используемые в данном примере, взяты из реальной IEEE 802.11 широкополосной беспроводной сети в режиме PCF, со скоростью передачи данных 54 Mbps, с реалистичными размерами пакетов и уровнями загрузки. Значения параметров взяты с учетом следующего: PCF режим часто используется для передачи чувствительного к задержке трафика, в этом случае размер пакетов довольно мал и мы взяли его равным 100 байтам. Принимая во внимание заголовок пакета, преамбулу, хвостовик и модуляции, мы можем получить, что такой пакет передается на скорости 44 μs. В соответствии с процедурой опознавания, которая используется в стандарте, пакет следует по SIFS интервалу и ACK пакету, чьи размеры 16 и 24 μs соответственно. Таким образом, полное время передачи пакета равно 84 μs, следовательно, среднее время обслуживания $b_i^{(1)} = 8.4 \times 10^{-5}$, $i = \overline{1,4}$.

В стандарте время переключения состоит из передачи пакета CF-POLL, следующим за интервалом PIFS простоя передающей среды. PIFS равны 25 μs, и CF-POLL занимают в среднем 44 μs, потому что он должен быть передан на самой низкой базовой скорости передачи, равной 6 Mbps для IEEE 802.11 стандарта. Тем не менее, если у опрашиваемой станции нечего передавать, она должна ответить на это сообщение NULL пакетом, чтобы проинформировать базовую станцию, что она должна переходить к следующей очереди. Это занимает 16 μs для SIFS интервала и 24 μs для самого NULL пакета, так как осуществляется ACK процедура. Это значит, что процедура переключения может занять 69 или 109 μs, в зависимости от состояния очереди. Мы не принимали это во внимание в аналитической модели, но можем сказать, что среднее время переключения есть среднее арифметическое этих двух значений. Это утверждение не является абсолютно верным, но такой аппроксимации вполне достаточно для нашей модели. Следовательно, среднее время $s_i^{(1)} = 8.9 \times 10^{-5}$, $i = \overline{1,4}$.

Среднее время отдыха прибора стандартом не определено и выбрано как $\varphi^{(1)} = 5 \times 10^{-5}$.

Сравним средневзвешенное время ожидания в системе

$$W = \sum_{i=1}^{N} \frac{\rho_i}{\rho} W_i^{(1)}$$

для циклического опроса и адаптивного циклического опроса, см. рисунок 8.5. Значение W для системы с циклическим опросом получается с помощью закона псевдосохранения, см. [102].

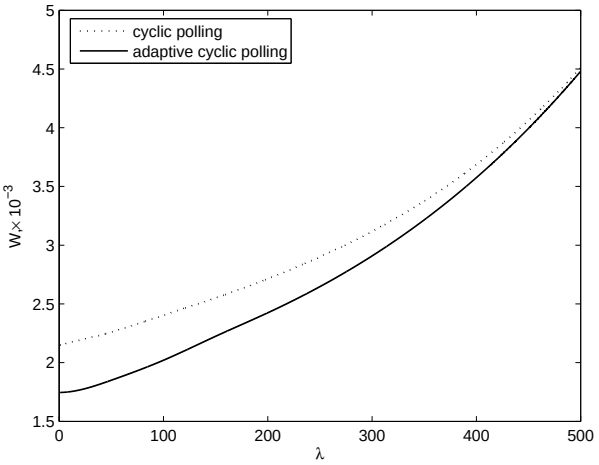

Рисунок 8.5. Зависимость средневзвешенного времени ожидания от λ для циклического и адаптивного опросов

Таким образом, используя стандартные значения, мы получили результаты, которые соответствуют таким сетям, и приведенные результаты показывают, что среднее время ожидания удовлетворяет требованиям реальной системы для чувствительного к задержке трафика.

В приведенном примере адаптивный динамический опрос дает выгоду до 23 % (в случае $\lambda = 1$). Но если λ возрастает и ρ стремится к 1, выгода существенно уменьшается, так как в случае тяжелого трафика система с адаптивным опросом ведет себя как система с циклическим опросом, поскольку буферы редко бывают пустыми в моменты их опроса.

ГЛАВА 9
МАТЕМАТИЧЕСКАЯ МОДЕЛЬ СОТЫ СЕТИ МОБИЛЬНОЙ СВЯЗИ С ПРОЦЕДУРОЙ HANDOVER

9.1 Обзор литературы

Известно, что концепция сотовой сети связи предполагает наличие большого числа маленьких областей (сот), что позволяет эффективнее использовать имеющиеся радиочастоты. Сотовые сети связи также предполагают мобильность абонентов (пользователей). Поэтому одним из базовых требований является автоматическая передача обслуживания пользователя при перемещении его из соты в соту без прерывания соединения. Процесс такой передачи называется "handover"в Европе и "handoff"в Америке. Этот процесс очень важен, поскольку при невозможности его надежного проведения возрастает доля прерванных соединений, которые являются одним из основных факторов неудовлетворенности пользователей качеством связи.

Для обеспечения приоритетного предоставления ресурса пользователям, прибывающим в данную соту при установленном соединении (далее будем называть их хэндовер-пользователями), по сравнению с пользователями, намеревающимися начать соединение в данной соте (далее будем называть их новыми пользователями), используются различные стратегии. Одной из наиболее широко известных стратегий является так называемая стратегия с защищенными каналами (Guard Channel Policy), см., например, [130]. Эта стратегия предусматривает резервирование определенного числа каналов для обслуживания хэндовер-пользователей. Таким образом устанавливается приоритет для мобильных пользователей, который реализуется путем резервирования некоторого числа каналов исключительно для обслуживания этих пользователей. При этом возникает задача выбора оптимального числа зарезервированных каналов.

Пусть общее число каналов в соте равно N. При стратегии Guard Channel Policy, фиксируется число R, $R < N$. Новый пользователь принимается для обслуживания, только если количество занятых приборов в момент попытки генерации им соединения меньше, чем R. Очевидно, что

задача оптимального выбора числа R нетривиальна. Если это число выбрано слишком большим, хэндовер-пользователи не получают достаточно приоритета перед новыми пользователями. Если же это число выбрано слишком малым, ухудшается качество обслуживания новых пользователей и растут простои приборов. Частный случай стратегии с защищенными каналами при $R = N - 1$ анализировался в [177]. В этой статье также содержится ряд полезных ссылок на ранние исследования по вопросу о резервировании каналов. В [177], так же как и в [160], отмечается, что резервирование даже только одного канала ($R = N - 1$) позволяет значительно улучшить качество обслуживания хэндовер-пользователей. В общем случае возникает проблема оптимального выбора числа R. В качестве критерия для оценки качества работы системы могут быть рассмотрены различные стоимостные функции. Эти стоимостные функции должны учитывать как вероятность потери хэндовер пользователей (ее иногда называют вероятностью сброса – dropping probability), вероятность потери новых пользователей (ее иногда называют вероятностью блокировки – blocking probability), а также коэффициент использования пропускной способности базовой станции соты. В результате при фиксированных величинах штрафов за потерю запросов различных типов и за недостаточное использование пропускной способности требуется найти оптимальное число R^* резервируемых каналов.

Краткий обзор литературы по данной проблеме содержится в статье [147], в которой, в частности, упоминаются работы [93,107,108,114,157,198]. Преимуществом работы [157] перед всеми предыдущими работами является точный математический анализ модели, в которой времена обслуживания хэндовер-пользователей и новых пользователей могут различаться друг от друга. Предположение об одинаковости распределения времени обслуживания хэндовер-пользователей и новых пользователей кардинально упрощает математический анализ модели, поскольку не требуется отдельно отслеживать текущее число хэндовер-пользователей и новых пользователей и анализ системы сводится к более-менее тривиальной задаче анализа одномерного процесса гибели и размножения. Однако в целом ряде работ обсуждались реальные статистические данные, свидетельствующие о существенном отличии распределений времен обслуживания хэндовер-пользователей и новых пользователей. Недостатками модели, рассмотренной в работе [157], являются следующие: (i) блокированный новый пользователь покидает систему, а не делает повторные попытки сгенерировать

соединение, и (ii) потоки хэндовер-пользователей и новых пользователей образуют стационарный пуассоновский поток.

Первое из этих предположений означает игнорирование явления повторных вызовов, типичного для сотовых сетей связи. В силу этой типичности далее мы будем упоминать только работы, учитывающие эффект повторных вызовов. Второе из этих предположений противоречит известному факту, что траффик в современных сетях характеризуется корреляцией и большой дисперсией времен между моментами поступления.

Эффект корреляции и большой дисперсией времен между моментами поступления запросов различных типов может быть эффективно учтен, если предположить, что прибытие запросов определяется маркированным марковским входным потоком (Marked Markovian arrival process – $MMAP$), см. [131], который является обобщением MAP-потока на случай потока разнородных запросов.

В работе [107] рассмотрена многолинейная СМО с независимыми MAP-потоками хэндовер-пользователей и новых пользователей. Если все приборы заняты в момент поступления хэндовер-пользователя, он регистрируется в буфере бесконечной емкости. В такой же ситуации новый пользователь начинает генерировать повторные запросы, зарегистрировавшись в так называемой орбите. Емкость орбиты в [107] предполагается конечной. Резервирования каналов не предусматривается.

В работе [93] рассмотрена довольно общая модель. Времена обслуживания и между моментами прихода запросов имеют распределение фазового типа, более общее, чем показательное распределение, предполагающееся в большинстве других работ. Недостатком анализа модели, рассмотренной в [93], является применение усечения размера орбиты. В силу мобильности пользователей вряд ли возможно провести усечение числа одновременно пребывающих в соте пользователей числом, при котором остается работоспособным вычислительный алгоритм, применяемый в [93].

Модель СМО с повторными вызовами, рассмотренная в [198], похожа на модель, изученную в [93]. Но для упрощения математического анализа авторы [198] предположили, что суммарная интенсивность повторов с орбиты не зависит от текущего числа запросов на орбите, что, конечно, не является справедливым на практике. В работе [108] рассматривается модель СМО, похожая на рассмотренную позднее в работе [147], но имеющая следующие два существенных ограничения: (i) потоки пользователей обоих типов являются стационарными пуассоновскими; (ii) времена обслужи-

вания хэндовер-пользователей и новых пользователей имеют одинаковое распределение. Как уже отмечалось выше, эти ограничения не выполняются на практике. В работе [114] анализируется модель, близкая к модели из [108]. Недостатком модели, рассмотренной в [114], по сравнению с моделью из [108], является то, что в ней, так же, как и в [198], предположено, что суммарная интенсивность повторов с орбиты не зависит от текущего числа запросов на орбите.

Модель, рассмотренная в работе [147], свободна от почти всех перечисленных недостатков предыдущих работ. Входной поток предполагается $MMAP$-потоком. Распределения времен обслуживания запросов обоих типов – экспоненциальные, но с интенсивностью, зависящей от типа запроса. В [147] получено условие эргодичности многомерной ЦМ, описывающей динамику СМО. Это условие имеет довольно прозрачную вероятностную трактовку, хотя на первый взгляд кажется неочевидным. Далее в [147] использованы результаты работы [150] для разработки алгоритмов нахождения стационарного распределения вероятностей состояний системы и значения основных характеристик производительности системы. Приведены численные результаты, иллюстрирующие эффективность стратегии с защищенными каналами при оптимальном выборе числа R резервируемых каналов.

Материал данного раздела книги основан на результатах исследований, проведенных в работе [146]. Модель, рассмотренная в [146], улучшает модель, изученную в [147], в следующих трех аспектах:

- В [147] рассмотрена чистая стратегия защищенных каналов. Это означает, что произвольный хэндовер-пользователь теряется, если в момент его поступления все каналы заняты. В реальных сотовых сетях соты частично перекрываются, вследствие чего существует некоторый интервал времен, в течение которого запрос, прибывающий из другой соты, может ждать предоставления канала в новой соте без разрыва соединения. Чтобы учесть этот факт, в [146] предполагается, что если в момент поступления хэндовер-пользователя все каналы заняты, то он не теряется мгновенно, а регистрируется в буфере ограниченной длины, если он не заполнен. Если буфер заполнен, пользователь получает отказ. Таким образом, стратегия доступа хэндовер-пользователей в [146] является гибридом стратегии защищенных каналов и стратегии с буферизацией хэндовер-пользователей. Время пребывания хэндовер-пользователя в буфере предполагается ограни-

ченным экспоненциально распределенной величиной. Пользователь, не получивший канал до истечения этого времени, теряется. Нетерпеливость пользователя, получившего место в буфере, можно трактовать, как потерю им (из-за слабости радиосигнала) соединения с базовой станцией соты, из которой он перемещается, или получение канала в еще одной, третьей, соте с которой перекрывается сота, из которой перемещается данный пользователь, а данная сота, не предоставила канал хэндовер-пользователю за время его ожидания, или завершение необходимого пользователю времени разговора.

- В [147] предполагается, что хэндовер-пользователь, получивший отказ в обслуживании из-за занятости всех каналов, покидает соту навсегда. В реальной жизни такой пользователь может начать инициацию повторных попыток попасть на обслуживание как обычный новый пользователь. Такая возможность допускается в [146].
- В [147] было предположено, что времена обслуживания пользователей обоих типов имеют экспоненциальное распределение, в то время как имеется статистика, см., например, [164], что времена занятия каналов в мобильных сотовых телефонных сетях имеют более сложное гиперэкспоненциальное распределение. Из соображений математической общности, в [146] предположено, что времена занятия каналов в рассматриваемой системе имеют еще более общее распределение фазового типа. Это предположение приводит в существенному усложнению анализа модели, поскольку существенно повышается многомерность ЦМ, описывающей поведение системы. Это приводит как к чисто техническим трудностям в выписывании генератора ЦМ, так и к значительным трудностям компьютерной реализации алгоритмов расчета характеристик системы ввиду ограниченного размера RAM современных персональных компьютеров. В значительной мере преодолеть упомянутые трудности позволяет использование так называемого обобщенного распределения фазового типа, см. [144], в комбинации с результатами [20, 167], иллюстрация использования которых была выше приведена в разделе 5.2.7.

Отметим также, что появление в модели буфера для ожидания хэндовер-пользователей (по сравнению с моделью в [147]) приводит к необходимости анализа также распределения времени ожидания.

Заметим, что эффективное управление процедурой хэндовер через предоставление приоритетов и резервирование каналов является важной

проблемой и в различных существующих стандартах и типах мобильных сотовых сетей (GSM/GPRS, CDMA, WCDMA, UMTS, HSDPA, etc.), и в сетях LTE-advanced (см. технические спецификации 3G PP TS - 3GA - 36.413 (Rel.13) V.13.1.0 Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP), 2015). Поэтому при соответствующем выборе параметров системы приведенные в данном разделе результаты можно использовать для оптимизации различных видов процедуры хэндовер, например системы с hard, softer, soft horizontal и vertical handover. Результаты можно также использовать для оптимизации процедуры хэндовер в когнитивных радиосетях, см., например, [94], [197].

В данном разделе приведены результаты исследования СМО, моделирующей работу соты с учетом эффекта хэндовер, основанные на статье [146].

9.2 Описание системы

Рассмотрим систему массового обслуживания с повторными вызовами и двумя типами запросов. Система имеет N идентичных приборов и конечное пространство для ожидания (буфер) емкости $K - N$. Структура исследуемой системы представлена на рисунке 9.1.

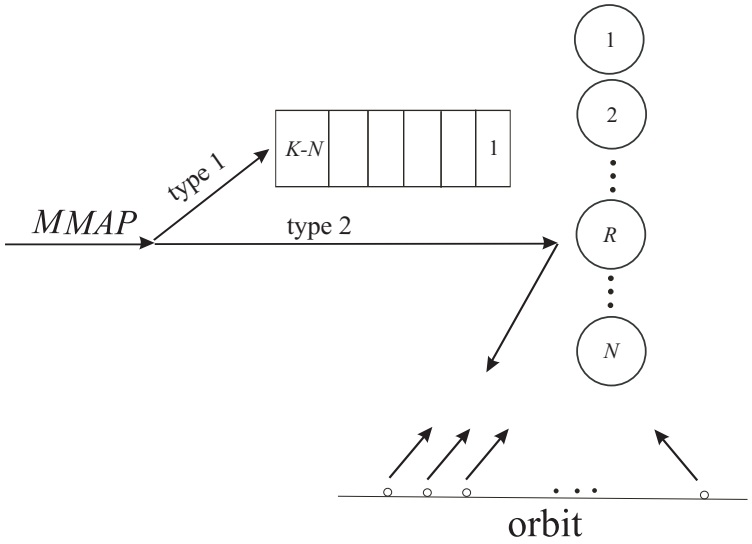

Рисунок 9.1. Структура исследуемой системы

В систему поступает марковский маркированный входной поток запросов ($MMAP$), заданный неприводимой цепью Маркова ν_t, $t \geq 0$, с непрерывным временем и конечным пространством состояний $\{0, 1, ..., W\}$. Время пребывания цепи в состоянии ν экспоненциально распределено с поло-

жительным параметром λ_ν. Когда время пребывания в состоянии ν истекло, с вероятностью $p^{(0)}_{\nu,\nu'}$ процесс ν_t переходит в состояние ν' без генерации запросов, $\nu \neq \nu'$, а с вероятностью $p^{(r)}_{\nu,\nu'}$ процесс ν_t переходит в состояние ν' с генерацией запроса r-го типа, $r = 1,\ 2$, $\nu, \nu' = \overline{0, W}$.

Поведение $MMAP$ полностью характеризуется матрицами $D_0,\ D_1^{(r)},\ r = 1, 2$, которые определяются следующим образом:

$$(D_1^{(r)})_{\nu,\nu'} = \lambda_\nu p^{(r)}_{\nu,\nu'},\ \nu, \nu' = \overline{0, W},\ r = 1, 2,$$

и

$$(D_0)_{\nu,\nu} = -\lambda_\nu, \nu = \overline{0, W},\ (D_0)_{\nu,\nu'} = \lambda_\nu p^{(0)}_{\nu,\nu'},\ \nu, \nu' = \overline{0, W}, \nu \neq \nu'.$$

Матрица $D(1) = D_0 + D_1^{(1)} + D_1^{(2)}$ представляет собой инфинитезимальный генератор цепи $\nu_t, t \geq 0$.

Средняя интенсивность поступления запросов λ имеет вид

$$\lambda = \boldsymbol{\theta}(D_1^{(1)} + D_1^{(2)})\mathbf{e},$$

где $\boldsymbol{\theta}$ – вектор стационарного распределения цепи Маркова $\nu_t, t \geq 0$. Вектор $\boldsymbol{\theta}$ является единственным решением системы линейных алгебраических уравнений

$$\boldsymbol{\theta} D(1) = \mathbf{0},\ \boldsymbol{\theta}\mathbf{e} = 1.$$

Средняя интенсивность λ_r поступления запросов r-го типа определяется формулой

$$\lambda_r = \boldsymbol{\theta}\, D_1^{(r)}\mathbf{e},\ r = 1,\ 2.$$

Коэффициент вариации c_{var} длин интервалов между моментами поступления запросов задается формулой

$$c_{var} = 2\lambda\boldsymbol{\theta}(-D_0)^{-1}\mathbf{e} - 1.$$

Коэффициент вариации $c^{(r)}_{var}$ длин интервалов между моментами поступления запросов r-го типа имеет вид

$$c^{(r)}_{var} = 2\lambda_r\boldsymbol{\theta}(-D_0 - D_1^{(\bar{r})})^{-1}\mathbf{e} - 1,\ \ \bar{r} \neq r, \bar{r}, r = 1, 2.$$

Коэффициент корреляции c_{cor} длин двух соседних интервалов между поступлением запросов вычисляется следующим образом:

$$c_{cor} = (\lambda\boldsymbol{\theta}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\mathbf{e} - 1)/c_{var}.$$

Коэффициент корреляции $c_{cor}^{(r)}$ длин двух соседних интервалов между поступлением запросов r-го типа вычисляется как

$$c_{cor}^{(r)} = (\lambda_r \boldsymbol{\theta}(D_0 + D_1^{(\bar{r})})^{-1} D_1^{(r)} (D_0 + D_1^{(\bar{r})})^{-1} \mathbf{e} - 1)/c_{var}^{(r)}, \bar{r} \neq r.$$

Следует отметить, что проведенный ниже анализ проводится в предположении, что параметры входного потока известны. Полученные результаты могут быть использованы для изучения практической системы только после того, как эти параметры для этой конкретной системы будут вычислены. Методы построения $MMAP$ или MAP потоков по реальным данным хорошо известны в литературе. Существующие результаты позволяют оценить значение параметра W и матрицы $D_0,\ D_1^{(k)},\ k = 1, 2$.

Время обслуживания запроса типа r каждым прибором имеет распределение фазового типа PH_r с неприводимым представлением $(\boldsymbol{\beta}_r, S_r)$, $r = 1, 2$. Время обслуживания, имеющее распределение фазового типа, можно интерпретировать как время, в течение которого управляющий марковский процесс $m_t^{(r)}, t \geq 0$, с конечным пространством состояний $\{1, \ldots, M_r, M_r + 1\}$ достигнет единственного поглощающего состояния $M_r + 1$ при условии, что начальное состояние этого процесса выбирается во множестве $\{1, \ldots, M_r\}$ согласно стохастическому вектору-строке $\boldsymbol{\beta}_r$, $r = 1, 2$. Интенсивности переходов процесса $m_t^{(r)}$ во множестве состояний $\{1, \ldots, M_r\}$ определяются субгенератором S_r, а интенсивности переходов в поглощающее состояние определяются элементами вектор-столбца $\boldsymbol{S}_0^{(r)} = -S_r \mathbf{e}$, $r = 1, 2$. Отметим, что представление $(\boldsymbol{\beta}_r, S_r)$ неприводимое, если матрица $S_r + \boldsymbol{S}_0^{(r)} \boldsymbol{\beta}_r$ неприводима, $r = 1, 2$.

Функция распределения времени обслуживания имеет вид

$$A_r(x) = 1 - \boldsymbol{\beta}_r e^{S_r x} \mathbf{e},$$

преобразование Лапласа – Стилтьеса (ПЛС) $\int\limits_0^\infty e^{-sx} dA_r(x)$ этого распределения имеет вид

$$\boldsymbol{\beta}_r (sI - S_r)^{-1} \boldsymbol{S}_0^{(r)},\ \mathrm{Re}\, s > 0,\ r = 1, 2.$$

Среднее время обслуживания запроса типа r вычисляется по формуле

$$b_1^{(r)} = \boldsymbol{\beta}_r (-S_r)^{-1} \mathbf{e},\ r = 1, 2.$$

Коэффициент вариации времени обслуживания запроса типа r имеет вид

$$c_{var}^{(r)} = b_2^{(r)} / (b_1^{(r)})^2 - 1,$$

где $b_2^{(r)} = 2\boldsymbol{\beta}_r(-S_r)^{-2}\mathbf{e}$.

Для получения более подробной информации о распределении фазового типа, см. [163].

Класс PH-распределений всюду плотен (в смысле слабой сходимости) во множестве всех вероятностных распределений неотрицательных случайных величин. Таким образом, PH-распределение может быть использовано для аппроксимации произвольного распределения, см. [91]. Предположение о том, что время обслуживания имеет распределение фазового типа, а не экспоненциальное распределение, существенно усложняет дальнейший анализ системы. Однако, как следует из [164], время обслуживания в мобильных сетях имеет гиперэкспоненциальное, а не экспоненциальное распределение. Поэтому необходимо рассматривать систему с более общим распределением, чем экспоненциальное. Гиперэкспоненциальное распределение, рекомендуемое для исследования в [164], является частным случаем PH-распределения. Рассмотрение конкретно гиперэкспоненциального распределения времени обслуживания ничем не упростит исследование системы по сравнению с PH-распределением. Поэтому для математической общности предположим, что времена обслуживания для обоих типов запросов имеет PH-распределение. Это гарантирует, что мы рассмотрим наиболее общий сценарий.

Поскольку мы применяем исследуемую систему для моделирования соты сети мобильной связи, запросы первого типа соответствуют хэндовер-запросам, в то время как запросы второго типа соответствуют новым вызовам, сгенерированным внутри соты.

Если в момент поступления произвольного запроса первого типа есть свободный прибор, запрос занимает прибор и начинает обслуживание. Если во время поступления произвольного запроса первого типа все приборы заняты, но буфер не полон, запрос становится в буфер. В противном случае, прибывающий запрос первого типа идет на орбиту с вероятностью p, а с дополнительной вероятностью покидает систему навсегда.

Запросы, находящиеся в буфере, могут быть нетерпеливыми и уходить из системы после случайного интервала времени, имеющего экспоненциальное распределение с параметром φ, $\varphi > 0$. В случае, когда запрос покидает систему из-за нетерпеливости, нетерпеливый запрос также уходит на орбиту с вероятностью p, а с дополнительной вероятностью покидает систему навсегда. Мы предполагаем, что запрос первого типа, который уходит на орбиту, изменяет тип и становится запросом второго типа.

Прием запросов второго типа в систему осуществляется в соответствии с пороговым механизмом. Пусть задан порог R, $0 < R \leq N$. Входящий запрос второго типа принимается на обслуживание в систему только в том случае, тогда число занятых приборов в момент поступления запроса меньше R. Это эквивалентно резервированию $N - R$ приборов исключительно для обслуживания запросов первого типа. Если запрос второго типа принимается в систему, запрос занимает произвольный свободный прибор и начинает обслуживание. Если запрос второго типа не получает разрешения начать обслуживание, с вероятностью q, $0 \leq q \leq 1$, этот запрос повторяет попытку получить обслуживание позже. С дополнительной вероятностью, запрос второго типа покидает систему навсегда (теряется).

Запрос с орбиты повторяет попытки попасть на обслуживание, независимо от других запросов на орбите, после интервала времени, имеющего экспоненциальное распределение с параметром $\alpha, \alpha > 0$. Попытка будет успешной, если число занятых приборов в момент совершения этой попытки меньше, чем R. Если попытка была успешной, запрос сразу же занимает свободный прибор и начинает обслуживание. Если попытка не удалась, запрос возвращается на орбиту с вероятностью q. С дополнительной вероятностью запрос второго типа покидает систему навсегда.

Запросы, находящиеся на орбите, нетерпеливы и могут уйти из системы после случайного интервала времени, имеющего экспоненциальное распределением с параметром γ, $\gamma > 0$.

9.3 Процесс изменения состояний системы

Для того чтобы значительно упростить исследование рассматриваемой системы, мы используем два очень полезных приема. Один прием состоит в описании множества состояний управляющего процесса обслуживания в каждом занятом приборе, основываясь не на состоянии управляющего процесса обслуживания в каждом занятом приборе, а на основе числа приборов, предоставляющих обслуживание на каждой существующей фазе обслуживания. Этот прием очень эффективен, когда число фаз в процессе обслуживания намного меньше, чем количество приборов. Этот прием довольно прост. Однако его конкретная реализация, впервые осуществленная в [20,167], далека от тривиальной. Этот прием был уже применен нами в разделе 5.3. Второй прием, изначально представленный в работе [144], состоит в следующем. Вместо отдельного рассмотрения времени обслужи-

вания запросов первого и второго типов будем рассмотривать обобщенное время обслуживания с распределением, которое мы называем обобщенным распределение фазового типа с неприводимым представлением $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, S)$, где

$$S = \begin{pmatrix} S_1 & O \\ O & S_2 \end{pmatrix}.$$

Этот прием значительно упрощает аналитическую работу, связанную с построением генератора многомерном цепи Маркова, которая описывает динамику рассматриваемой системы. Это упрощение происходит в виду того, что прием позволяет учитывать только общее число занятых приборов, без отдельного учета количества приборов, предоставляющих обслуживание для запросов первого и второго типа. Важно отметить, что в сочетании с концепцией, представленной в [20, 167], использование обобщенного распределения фазового типа не приводит к увеличению пространства состояний процесса, что крайне важно при компьютерной реализации представленных ниже алгоритмов.

Время обслуживания, имеющее обобщенное распределение фазового типа, можно интерпретировать как время, в течение которого управляющий марковский процесс $\eta_t, t \geq 0$, с конечным пространством состояний $\{1, \ldots, M, M+1\}$, где $M = M_1 + M_2$, достигнет единственного поглощающего состояния $M+1$. Начальное состояние этого процесса выбирается во множестве состояний $\{1, \ldots, M\}$ в зависимости от типа запроса, выбираемого для обслуживания. Если для обслуживания выбирается произвольный запрос первого типа, начальное состояние этого процесса выбирается в соответствии с вероятностным вектором-строкой $\boldsymbol{\beta}^{(1)} = (\boldsymbol{\beta}_1, \mathbf{0}_{M_2})$, а если для обслуживания выбирается запрос второго типа, начальное состояние выбирается в соответствии с вероятностным вектором-строкой $\boldsymbol{\beta}^{(2)} = (\mathbf{0}_{M_1}, \boldsymbol{\beta}_2)$. Интенсивности переходов процесса η_t во множестве $\{1, \ldots, M\}$ определяются суб-генератором

$$S = \begin{pmatrix} S_1 & O \\ O & S_2 \end{pmatrix},$$

а интенсивности переходов в поглощающее состояние (которые приводят к окончанию обслуживания) задаются элементами вектора столбца $\boldsymbol{S}_0 = -S\mathbf{e}$.

Пусть i_t, $i_t \geq 0$, – количество запросов на орбите, k_t, $k_t = \overline{0, K}$, – количество запросов в системе, ν_t, $\nu_t = \overline{0, W}$, – состояние управляюще-

го процесса $MMAP$, $\eta_t^{(m)}$ – количество приборов на фазе m обобщенного обслуживания, $m = \overline{1,M}$, $\eta_t^{(m)} = \overline{0, \min\{k_t, N\}}$, $\sum\limits_{m=1}^{M} \eta_t^{(m)} = \min\{k_t, N\}$, в момент времени t, $t \geq 0$.

Легко заметить, что многомерный процесс

$$\xi_t = \{i_t, k_t, \nu_t, \eta_t^{(1)}, \ldots, \eta_t^{(M)}\}, \;\; t \geq 0,$$

является неприводимой цепью Маркова с непрерывным временем.

Перенумеруем состояния цепи ξ_t в обратном лексикографический порядке компонент $(\eta_t^{(1)}, \ldots, \eta_t^{(M)})$ и в прямом порядке компонент (i_t, k_t, ν_t). Множество состояний со значениями (i, k) двух первых компонент называется макро-состоянием (i, k).

Пусть Q – генератор цепи Маркова ξ_t, $t \geq 0$, состоящий из блоков $Q_{i,j}$, которые, в свою очередь, состоят из матриц $(Q_{i,j})_{k,k'}$ интенсивностей переходов этой цепи из макро-состояния (i, k) в макро-состояние (j, k'), k, $k' = \overline{0, K}$. Диагональные элементы $Q_{i,i}$ отрицательны, а модули диагональных элементов матриц определяют полную интенсивность выхода из соответствующего состояния цепи Маркова ξ_t, $t \geq 0$.

Для ясности и упрощения объяснения формы блоков генератора Q, использованные ранее, и некоторые новые обозначения приведены в таблице 9.1.

Таблица 9.1. Обозначения и сокращения

K	емкость системы
N	число приборов
R	порог управления допуском запросов второго типа
D_0, D_1, D_2	квадратные матрицы размера $W+1$, которые характеризуют $MMAP$
λ_r, $r=1,2$	средняя интенсивность поступления запросов r-го типа
$(\boldsymbol{\beta}_r, S_r)$, $r=1,2$	неприводимое представление распределения времени обслуживания запросов r-го типа
φ	интенсивность нетерпеливости запросов первого типа
α	индивидуальная интенсивность повторных попыток
γ	интенсивность нетерпеливости запросов, находящихся на орбите
p	вероятность того, что запрос первого типа идет на орбиту в случае, когда нет свободного места в системе в момент его прихода, или когда запрос покидает буфер из-за нетерпеливости
q	вероятность того, что запрос второго типа идет (возвращается) на орбиту в момент поступления (совершения повторной попытки), когда число занятых приборов больше или равно R
T_n	$\binom{n+M-1}{M-1} = \frac{(n+M-1)!}{n!(M-1)!}$
$(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, S)$	неприводимое представление распределения обобщенного времени обслуживания фазового типа
M	число фаз обобщенного распределения фазового типа
ξ_t	цепь Маркова, описывающая поведение системы
Q	генератор цепи Маркова ξ_t
$\tilde{S}$	$\begin{pmatrix} 0 & \mathbf{0} \\ \boldsymbol{S}_0 & S \end{pmatrix}$
$P_k(\boldsymbol{\beta}^{(r)})$	матрица, которая определяет вероятности переходов процесса $\{\eta_t^{(1)}, \ldots, \eta_t^{(M)}\}$ в момент начала обслуживания запроса r-го типа при условии, что k приборов заняты в этот момент

Продолжение таблицы 2.1

$L_{N-k}(N,\tilde{S})$	матрица, которая определяет интенсивности переходов процесса $\{\eta_t^{(1)},\dots,\eta_t^{(M)}\}$ в момент завершения обслуживания при условии, что k приборов заняты в данный момент
$A_k(N,S)$	матрица, которая определяет интенсивности переходов процесса $\{\eta_t^{(1)},\dots,\eta_t^{(M)}\}$, которые не приводят к окончанию обслуживания при условии, что k приборов заняты
$\Delta^{(k)}$	диагональная матрица с диагональными элементами, которые определяют суммарную интенсивность выхода из соответствующего состояния процесса $\{\eta_t^{(1)},\dots,\eta_t^{(M)}\}$, при условии, что k приборов заняты

Лемма 9.1. *Генератор Q цепи Маркова ξ_t имеет следующую блочно-трехдиагональную структуру:*

$$Q=\begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & \dots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{9.1}$$

Ненулевые блоки $Q_{i,j}$, $i,j\geq 0$, вычисляются как:

$$Q_{i,i}=\mathrm{diag}\{C_i^{(0)},\dots,C_i^{(K)}\}+\mathrm{diag}^-\{\bar{C}^{(1)},\dots,\bar{C}^{(K)}\}+\mathrm{diag}^+\{\tilde{C}^{(0)},\dots,\tilde{C}^{(K-1)}\}, \tag{9.2}$$

где

$$C_i^{(k)}=D_0\oplus[A_{\min\{k,N\}}(N,S)+\Delta^{(\min\{k,N\})}]-i(\alpha+\gamma)I_{\bar{W}T_k}+ \tag{9.3}$$

$$+\begin{cases} O, & 0\leq k<R,\\ iq\alpha I_{\bar{W}T_k}+(1-q)D_2\otimes I_{T_k}, & R\leq k\leq N,\\ (iq\alpha-(k-N)\varphi)I_{\bar{W}T_N}+(1-q)D_2\otimes I_{T_N}, & N<k<K,\\ (iq\alpha-(k-N)\varphi)I_{\bar{W}T_N}+((1-p)D_1+(1-q)D_2)\otimes I_{T_N}, & k=K, \end{cases}$$

$$\Delta^{(0)}=0,\ \Delta^{(k)}=-\mathrm{diag}\{A_k(N,S)\mathbf{e}+L_{N-k}(N,\tilde{S})\mathbf{e}\},\ k=\overline{1,N}, \tag{9.4}$$

$$\bar{C}^{(k)}=\begin{cases} I_{\bar{W}}\otimes L_{N-k}(N,\tilde{S}), & 1\leq k\leq N,\\ I_{\bar{W}}\otimes L_0(N,\tilde{S})P_{N-1}(\boldsymbol{\beta}^{(1)})+(1-p)(k-N)\varphi I_{\bar{W}T_N}, & N<k\leq K, \end{cases} \tag{9.5}$$

$$\tilde{C}^{(k)} = \begin{cases} D_1 \otimes P_k(\boldsymbol{\beta}^{(1)}) + D_2 \otimes P_k(\boldsymbol{\beta}^{(2)}), & 0 \le k < R, \\ D_1 \otimes P_k(\boldsymbol{\beta}^{(1)}), & R \le k < N, \\ D_1 \otimes I_{T_N}, & N \le k < K, \end{cases} \tag{9.6}$$

$$Q_{i,i+1} = Q^+ = \mathrm{diag}\{H^{(0)}, \ldots, H^{(K)}\} + \mathrm{diag}^-\{\bar{H}^{(1)}, \ldots, \bar{H}^{(K)}\},\ i \ge 0, \tag{9.7}$$

где

$$H^{(k)} = \begin{cases} O_{\bar{W}T_k \times \bar{W}T_k}, & 0 \le k < R, \\ qD_2 \otimes I_{T_{\min\{k,N\}}}, & R \le k < K, \\ (pD_1 + qD_2) \otimes I_{T_N}, & k = K, \end{cases} \tag{9.8}$$

$$\bar{H}^{(k)} = \begin{cases} O_{\bar{W}T_{k-1} \times \bar{W}T_k}, & 1 \le k \le N, \\ p(k-N)\varphi I_{\bar{W}T_N}, & N < k \le K, \end{cases} \tag{9.9}$$

$$Q_{i,i-1} = \mathrm{diag}\{B_i^{(0)}, \ldots, B_i^{(K)}\} + \mathrm{diag}^+\{\tilde{B}_i^{(0)}, \ldots, \tilde{B}_i^{(K-1)}\},\ i \ge 1, \tag{9.10}$$

где

$$B_i^{(k)} = \begin{cases} i\gamma I_{\bar{W}T_k}, & 0 \le k < R, \\ i(\gamma + (1-q)\alpha) I_{\bar{W}T_{\min\{k,N\}}}, & R \le k \le K, \end{cases} \quad i \ge 1, \tag{9.11}$$

$$\tilde{B}_i^{(k)} = \begin{cases} i\alpha I_{\bar{W}} \otimes P_k(\boldsymbol{\beta}^{(2)}), & 0 \le k < R, \\ O_{\bar{W}T_{\min\{k,N\}} \times \bar{W}T_{\min\{k+1,N\}}}, & R \le k < K, \end{cases} \quad i \ge 1. \tag{9.12}$$

Подробное описание матриц $P_k(\boldsymbol{\beta}^{(1)})$, $P_k(\boldsymbol{\beta}^{(2)})$, $k = \overline{0, N-1}$, $A_k(N, S)$ *и* $L_k(N, \tilde{S})$, $k = \overline{0, N}$, *и алгоритмы, используемые для их вычисления, можно найти в [145].*

Доказательство леммы осуществляется путем исчерпывающего анализа интенсивностей переходов цепи Маркова ξ_t за бесконечно малый промежуток времени. Блочно-трехдиагональная структура (9.1) генератора Q объясняется тем, что вероятность того, что количество запросов на орбите увеличится с i до $i + l$ или уменьшится с i до $i - l$ за бесконечно малый интервал времени, ничтожно мала для $l \ge 2$. Таким образом, только блоки $Q_{i,i-1}$, $Q_{i,i}$, $Q_{i,i+1}$ в i-й блочной строке являются ненулевыми. Блок $(Q_{i,j})_{k,k'}$ состоит из интенсивностей переходов компонент $\{i_t, k_t\}$ цепи Маркова ξ_t из состояния (i, k) в состояние (j, k'). Здесь компонента k_t является числом запросов в системе. Точно так же, как было упомянуто выше, блоки $(Q_{i,j})_{k,k'}$ могут быть отличны от нуля, только если $k' = k - 1$, $k' = k$, $k' = k + 1$. Эти рассуждения объясняют блочно-трехдиагональную форму (9.2) матрицы $Q_{i,i}$ и блочно-двухдиагональную форму (9.7) и (9.10) матриц $Q_{i,i-1}$ и $Q_{i,i+1}$ соответственно.

Диагональными блоками матрицы $Q_{i,i}$ являются матрицы $C_i^{(k)}$, $k = \overline{0,K}$. Форма (9.3) этих матриц объясняется следующим образом. Недиагональные элементы матрицы $C_i^{(k)}$ представляют интенсивности перехода компонент $\{\nu_t, \eta_t^{(1)}, \ldots, \eta_t^{(M)}\}$ цепи Маркова ξ_t в другое состояние без изменений компонент $\{i_t, k_t\}$ этой цепи. Такие переходы могут происходить, когда наступает одно (и только одно) из следующих событий:

1) Управляющий процесс поступления запросов делает переход без генерации запроса (интенсивности такого перехода определяются недиагональными элементами матрицы D_0);

2) Генерируется запрос второго типа, но этот запрос не принимается в систему, так как число занятых приборов не меньше R, и запрос принимает решение покинуть систему (интенсивности такого перехода определяются матрицей $(1-q)D_2$);

3) Запрос первого типа поступает, но не может быть принят в систему, так как все приборы заняты и буфер заполнен; запрос принимает решение покинуть систему (интенсивности такого перехода определяются матрицей $(1-p)D_1$);

4) Происходит изменение фазы обслуживания в одном из занятых приборов, не повлекших завершение обслуживания (интенсивности такого перехода определяются матрицей $A_{\min\{k,N\}}(N,S)$).

Диагональные элементы матрицы $C_i^{(k)}$ отрицательны. Их модули представляют интенсивности выхода цепи Маркова ξ_t из соответствующих состояний. Эти элементы равны сумме соответствующих диагональных элементов матриц, приведенных выше, минус суммарная интенсивность запросов, уходящих с орбиты (делают успешную попытку или уходят из-за нетерпеливости). В результате мы получаем формулы (9.3) и (9.4).

Матрица $Q_{i,i}$ также имеет поддиагональные блоки $(Q_{i,i})_{k,k-1}$, определенные как $\bar{C}^{(k)}$, $k = \overline{1,K}$, и наддиагональные блоки $(Q_{i,i})_{k,k+1}$, определенные как $(Q_{i,i})_{k,k+1}$. Форма (9.5) поддиагональных блоков вытекает из того факта, что переходы компоненты k_t цепи Маркова ξ_t из состояния k в состояние $k-1$ происходят, когда количество запросов в системе уменьшается на 1. Такая ситуация возникает, когда завершается обслуживание в одном из занятых приборов или запрос покидает буфер и систему из-за нетерпеливости. Интенсивности завершения обслуживания задаются матрицами $L_{N-k}(N,\tilde{S})$, если $k = \overline{1,N}$, и матрицами $L_0(N,\tilde{S})$, если $k = \overline{N+1,K}$. В последнем случае новое обслуживание начинается сразу после завершения предыдущего. Фаза управляющего процесса обслуживания запроса перво-

го типа задается в соответствии с вероятностной матрицей $P_{N-1}(\boldsymbol{\beta}^{(1)})$. Интенсивность ухода из системы из-за нетерпеливости равна $(1-p)(k-N)\varphi$. Эти выкладки приводят к формуле (9.5). Форма (9.6) наддиагональных блоков матрицы $Q_{i,i}$ вытекает из того факта, что переходы компоненты k_t цепи Маркова ξ_t из состояния k в состояние $k+1$ происходят, когда число запросов в системе увеличивается на 1. Эта ситуация возникает, когда новый запрос принимается в систему. Интенсивности поступления запросов r-го типа определяются матрицей D_r, $r = 1, 2$. Если имеются доступные приборы, инициируется обслуживание приходящего запроса. Фаза управляющего процесса обслуживания запроса r-го типа инициируется в соответствии с заданными вероятностными матрицами $P_k(\boldsymbol{\beta}^{(r)})$, $r = 1, 2$. В результате, мы получаем формулу (9.6).

Матрица $Q_{i,i+1}$ имеет диагональные блоки $(Q_{i,i+1})_{k,k}$, определенные как $H^{(k)}$, $k = \overline{0, K}$, и поддиагональные блоки $(Q_{i,i-1})_{k,k-1}$, определенные как $\bar{H}^{(k)}$, $k = \overline{1, K}$. Переходы, интенсивность которых задается матрицей $H^{(k)}$, происходят, когда поступающий запрос не может быть допущен в систему и он решает присоединиться к орбите. Переходы, интенсивности которых задаются матрицей $\bar{H}^{(k)}$, возникают, когда запрос покидает буфер из-за нетерпеливости и решает присоединиться к орбите. Посредством этого краткого анализа мы легко получим формулы (9.7)-(9.9).

Матрица $Q_{i,i-1}$ имеет диагональные блоки $(Q_{i,i-1})_{k,k}$, определенные как $B_i^{(k)}$, $k = \overline{0, K}$, и наддиагональные блоки $(Q_{i,i-1})_{k,k+1}$, определенные как $\tilde{B}_i^{(k)}$, $k = \overline{0, K-1}$. Переходы, интенсивность которых задается матрицей $B_i^{(k)}$, возникают, когда запрос покидает орбиту из-за нетерпеливости или после неудачной попытки. Эти интенсивности определяются формулой (9.11). Переходы, интенсивности которых задаются матрицей $\tilde{B}_i^{(k)}$, происходят, когда запрос покидает орбиту после удачной попытки. Обслуживание этих запросов начинается сразу в момент попытки. Фаза управляющего процесса обслуживания запроса r-го типа инициируется в соответствии с заданной вероятностной матрицей $P_k(\boldsymbol{\beta}^{(2)})$. Эти интенсивности определяются формулой (9.12).

Лемма доказана.

Замечание 9.1. Можно проверить, что существуют следующие пределы:

$$Y_0 = \lim_{i\to\infty} R_i^{-1} Q_{i,i-1},\ Y_1 = \lim_{i\to\infty} R_i^{-1} Q_{i,i} + I,\ Y_2 = \lim_{i\to\infty} R_i^{-1} Q_{i,i+1},$$

где матрица R_i – диагональная матрица с диагональными элементами, определенными как модули соответствующих диагональных элементов

матрицы $Q_{i,i},\ i \geq 0$. Таким образом, в соответствии с определением цепь Маркова ξ_t принадлежит к классу асимптотически квазитеплицевых цепей Маркова (АКТЦМ).

Можно проверить, что матрицы Y_0, Y_1 и Y_2 определяются следующими выражениями:

$$Y_0 =$$

$$= \begin{pmatrix} \frac{\gamma}{\gamma+\alpha} I_{\bar{W}} & \frac{\alpha}{\gamma+\alpha} I_{\bar{W}} \otimes P_0(\boldsymbol{\beta}^{(2)}) & \dots & O & O & O & \dots & O \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & \dots & \frac{\gamma}{\gamma+\alpha} I_{\bar{W}T_{R-1}} & \frac{\alpha}{\gamma+\alpha} \alpha I_{\bar{W}} \otimes P_{M-1}(\boldsymbol{\beta}^{(2)}) & O & \dots & O \\ O & O & \dots & O & I_{\bar{W}T_R} & O & \dots & O \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & \dots & O & O & O & \dots & O \\ O & O & \dots & O & O & O & \dots & I_{\bar{W}T_N} \end{pmatrix},$$

$$Y_1 = O,\ Y_2 = O.$$

Как следует из [150], достаточным условием для эргодичности АКТЦМ $\xi_t,\ t \geq 0$, является выполнение неравенства

$$\mathbf{y} Y_0 \mathbf{e} > \mathbf{y} Y_2 \mathbf{e}, \tag{9.13}$$

где вектор-строка $\mathbf{y}$ является единственным решением системы линейных алгебраических уравнений

$$\mathbf{y}(Y_0 + Y_1 + Y_2) = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1. \tag{9.14}$$

Очевидно, что для рассматриваемой цепи Маркова ξ_t неравенство (9.13) и систему (9.14) можно переписать в виде

$$\mathbf{y} Y_0 \mathbf{e} > 0,$$

$$\mathbf{y} Y_0 = \mathbf{y},\ \mathbf{y}\mathbf{e} = 1.$$

Таким образом, неравенство (9.13) эквивалентно неравенству

$$\mathbf{y} Y_0 \mathbf{e} = \mathbf{y}\mathbf{e} = 1 > 0,$$

которое справедливо для всех значений параметров исследуемой системы.

Таким образом, существуют следующие пределы (стационарные вероятности):

$$\pi(i, k, \nu, \eta^{(1)}, \dots, \eta^{(M)}) = \lim_{t \to \infty} P\{i_t = i,\ k_t = k,\ \nu_t = \nu,\ \eta_t^{(m)} = \eta^{(m)}, m = \overline{1, M}\},$$

$$i \geq 0,\ k = \overline{0, K},\ \nu = \overline{0, W}, \eta^{(m)} = \overline{0, \min\{k, N\}},\ m = \overline{1, M}.$$

Составим векторы-строки $\boldsymbol{\pi}(i, k, \nu)$ из этих вероятностей, перенумерованных в обратном лексикографическом порядке компонент $\eta^{(1)}, \ldots, \eta^{(M)}$. Затем сформируем векторы-строки стационарных вероятностей $\boldsymbol{\pi}_i$ следующим образом:

$$\boldsymbol{\pi}(i, k) = (\boldsymbol{\pi}(i, k, 0), \boldsymbol{\pi}(i, k, 1), \ldots, \boldsymbol{\pi}(i, k, W)),\ \ k = \overline{0, K},$$

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i, 0), \boldsymbol{\pi}(i, 1), \ldots, \boldsymbol{\pi}(i, K)),\ i \geq 0.$$

Хорошо известно, что вероятностные векторы $\boldsymbol{\pi}_i$, $i \geq 0$, удовлетворяют следующей системе линейных алгебраических уравнений:

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)Q = \mathbf{0},\ \ (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots)\mathbf{e} = 1, \tag{9.15}$$

где Q является генератором цепи Маркова $\xi_t,\ t \geq 0$. Эта система имеет бесконечный размер. Чтобы решить эту систему, мы использовали численно устойчивый алгоритм для вычисления вероятностных векторов $\boldsymbol{\pi}_i$, $i \geq 0$, разработанный в [150]. Алгоритм, разработанный в [150], предполагает, что генератор АКТЦМ имеет верхне-хессенбергову структуру. Этот алгоритм был адаптирован для случая более простой блочно-трехдиагональной структуры в работе [119].

9.4 Характеристики производительности системы

Вычислив векторы стационарных вероятностей $\boldsymbol{\pi}_i,\ i \geq 0$, можно вычислить различные характеристики производительности системы.

Распределение числа запросов на орбите вычисляется как

$$\lim_{t \to \infty} P\{i_t = i\} = \boldsymbol{\pi}_i \mathbf{e},\ i \geq 0.$$

Среднее число запросов в системе вычисляется как

$$L = \sum_{i=0}^{\infty} \sum_{k=0}^{K} k \boldsymbol{\pi}(i, k) \mathbf{e}.$$

Замечание 9.2. Здесь и в следующем подразделе формулы для основных характеристик производительности содержат бесконечные суммы. Эти

суммы не создают критические трудности в вычислениях. Хорошо известно, что если цепь Маркова эргодична, векторы стационарных вероятностей $\boldsymbol{\pi}_i$ сходятся по норме к нулевому вектору, если i стремится к бесконечности. Таким образом, вычисление бесконечной суммы может быть прекращено, если норма слагаемого становится меньше наперед заданного значения ϵ (например, $\epsilon = 10^{-10}$).

Среднее число занятых приборов вычисляется как

$$N_{server} = \sum_{i=0}^{\infty} \sum_{k=1}^{K} \min\{k, N\} \boldsymbol{\pi}(i,k)\mathbf{e}.$$

Среднее количество запросов первого типа в буфере вычисляется как

$$N_{buffer} = \sum_{i=0}^{\infty} \sum_{k=N+1}^{K} (k-N) \boldsymbol{\pi}(i,k)\mathbf{e}.$$

Среднее количество запросов на орбите вычисляется как

$$L_{orbit} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}.$$

Интенсивность выходного потока запросов r-го типа

$$\lambda_{out}^{(r)} = \sum_{i=0}^{\infty} \sum_{k=1}^{K} \boldsymbol{\pi}(i,k)(I_{\bar{W}} \otimes L_{N-\min\{k,N\}}(N, \tilde{S}^{(r)}))\mathbf{e},\ r = 1, 2,$$

где

$$\tilde{S}^{(1)} = \begin{pmatrix} 0 & \mathbf{0} \\ \begin{pmatrix} -(S_1\mathbf{e})^T, \mathbf{0}_{M_2} \end{pmatrix}^T & S \end{pmatrix},\ \tilde{S}^{(2)} = \begin{pmatrix} 0 & \mathbf{0} \\ \begin{pmatrix} \mathbf{0}_{M_1}, -(S_2\mathbf{e})^T \end{pmatrix}^T & S \end{pmatrix}.$$

Интенсивность выходного потока запросов из системы вычисляется как

$$\lambda_{out} = \lambda_{out}^{(1)} + \lambda_{out}^{(2)}.$$

Вероятность потери запроса первого типа по прибытии из-за переполнения буфера вычисляется как

$$P_1^{(arr-loss)} = (1-p)\lambda_1^{-1} \sum_{i=0}^{\infty} \boldsymbol{\pi}(i,K)(D_1 \otimes I_{T_N})\mathbf{e}.$$

Вероятность того, что произвольный запрос первого типа пойдет на орбиту по прибытии из-за переполнения буфера, вычисляется как

$$P_1^{(arr-to-orbit)} = p\lambda_1^{-1} \sum_{i=0}^{\infty} \boldsymbol{\pi}(i, K)(D_1 \otimes I_{T_N})\mathbf{e}.$$

Вероятность потери запроса второго типа на входе в систему вычисляется как

$$P_2^{(arr-loss)} = (1-q)\lambda_2^{-1} \sum_{i=0}^{\infty} \sum_{k=R}^{K} \boldsymbol{\pi}(i, k)(D_2 \otimes I_{T_{\min\{k,N\}}})\mathbf{e}.$$

Вероятность того, что запрос второго типа пойдет на орбиту по прибытии в систему, вычисляется как

$$P_2^{(arr-to-orbit)} = q\lambda_2^{-1} \sum_{i=0}^{\infty} \sum_{k=R}^{K} \boldsymbol{\pi}(i, k)(D_2 \otimes I_{T_{\min\{k,N\}}})\mathbf{e}.$$

Вероятность потери запроса первого типа вычисляется как

$$P_1^{(loss)} = 1 - \frac{\lambda_{out}^{(1)}}{\lambda_1}.$$

Вероятность того, что произвольный запрос первого типа покинет буфер из-за нетерпеливости и потеряется, вычисляется как

$$P_1^{(imp-loss)} = (1-p)(P_1^{(loss)} - P_1^{(arr-to-orbit)} - P_1^{(arr-loss)}).$$

Вероятность того, что произвольный запрос первого типа покинет буфер из-за нетерпеливости и пойдет на орбиту, вычисляется как

$$P_1^{(imp-orb)} = p(P_1^{(loss)} - P_1^{(arr-to-orbit)} - P_1^{(arr-loss)}).$$

Вероятность потери запроса второго типа вычисляется как

$$P_2^{(loss)} = 1 - \frac{\lambda_{out}^{(2)}}{\lambda_1(P_1^{(arr-to-orbit)} + P_1^{(imp-orb)}) + \lambda_2}.$$

Вероятность потери запроса второго типа с орбиты вычисляется как

$$P_2^{(loss-from-orbit)} = P_2^{(loss)} - P_2^{(arr-loss)}.$$

Вероятность того, что произвольный запрос первого типа занимает прибор по прибытии, вычисляется как

$$P_1^{(imm)} = \lambda_1^{-1} \sum_{i=0}^{\infty} \sum_{k=0}^{N-1} \boldsymbol{\pi}(i,k)(D_1 \otimes I_{T_k})\mathbf{e}.$$

Вероятность того, что произвольный запрос второго типа занимает прибор по прибытии, вычисляется как

$$P_2^{(imm)} = \lambda_2^{-1} \sum_{i=0}^{\infty} \sum_{k=0}^{R-1} \boldsymbol{\pi}(i,k)(D_2 \otimes I_{T_k})\mathbf{e}.$$

9.5 Распределение времени ожидания произвольного запроса первого типа

Пусть $V(x)$ – функция распределения времени ожидания произвольного запроса первого типа в системе, и пусть $v(s) = \int\limits_0^\infty e^{-sx} dV(x)$, $\mathrm{Re}\ s > 0$, – ее преобразование Лапласа – Стилтьеса (ПЛС).

Пометим произвольный запрос первого типа и проследим за его нахождением в буфере. Получим выражение для ПЛС $v(s)$, используя метод коллективных отметок (смотри, например, [143], [178]). С этой целью будем интерпретировать переменную s как интенсивность виртуального стационарного пуассоновского потока катастроф. Таким образом, $v(s)$ представляет собой вероятность того, что катастрофа не наступит в течение времени пребывания помеченного запроса.

Пусть $w(s, l, \eta^{(1)}, \ldots, \eta^{(M)})$ – это вероятность того, что катастрофа не наступит в течение времени ожидания в системе помеченного запроса первого типа при условии, что в данный момент позиция помеченного запроса в буфере есть l, $l = \overline{1, K-N}$, а состояния процессов $\eta_t^{(1)}, \ldots, \eta_t^{(M)}$, $t \geq 0$, есть $\eta^{(1)}, \ldots, \eta^{(M)}$, соответственно.

Перенумеруем вероятности $w(s, l, \eta^{(1)}, \ldots, \eta^{(M)})$ в лексикографическом порядке компонент, как было указано выше, и образуем из этих вероятностей векторы-столбцы $\mathbf{w}(s, l)$.

Теорема 9.1. *ПЛС $v(s)$ распределения времени ожидания произвольного запроса первого типа в системе вычисляется как*

$$v(s) = P_1^{(arr-to-orbit)} + P_1^{(arr-loss)} + P_1^{(imm)} +$$

$$+\lambda_1^{-1}\sum_{i=0}^{\infty}\sum_{k=N}^{K-1}\boldsymbol{\pi}(i,k)(D_1\mathbf{e}\otimes I_{T_N})\mathbf{w}(s,k-N+1).$$

Доказательство. Возможны следующие ситуации во время прибытия помеченного запроса первого типа:

1) Буфер полон, и запрос покидает систему навсегда или идет на орбиту. Вероятность этого события $P_1^{(arr-to-orbit)}+P_1^{(arr-loss)}$, а вероятность того, что катастрофа не произойдет за время ожидания, равна 1.

2) Есть свободный прибор, и помеченный запрос сразу начинает получать обслуживание. Вероятность этого события – $P_1^{(imm)}$. В этом случае вероятность того, что катастрофа не произойдет за время ожидания, также равна 1.

3) Все приборы заняты, буфер не заполнен и помеченный запрос присоединяется к буферу. Вероятность этого события – $\lambda_1^{-1}\sum_{i=0}^{\infty}\sum_{k=N}^{K-1}\boldsymbol{\pi}(i,k)(D_1\mathbf{e}\otimes I_{T_N})\mathbf{e}$ и вероятность того, что катастрофа не наступит в течение времени ожидания помеченного запроса при условии, что число запросов в системе равно k, $k=\overline{N,K-1}$, равна $\mathbf{w}(s,k-N+1)\mathbf{e}$.

Используя формулу полной вероятности, можно легко убедиться в справедливости теоремы.

Для того чтобы разработать алгоритм для вычисления вектор-столбцов $\mathbf{w}(s,l)$, мы используем вероятностную интерпретацию ПЛС, вероятностный смысл матриц $P_{N-1}(\boldsymbol{\beta}^{(1)})$, $A_N(N,S)$ и $L_0(N,\tilde{S})$, приведенный выше, и формулу полной вероятности. В результате получаем следующее утверждение.

Лемма 9.2. *Векторы* $\mathbf{w}(s,l)$ *могут быть рекурсивно вычислены следующим образом:*

$$\mathbf{w}(s,1)=\Big[(s+\varphi)I-(A_N(N,S)+\Delta^{(N)})\Big]^{-1}(L_0(N,\tilde{S})\mathbf{e}+\varphi\mathbf{e}),$$

$$\mathbf{w}(s,l)=\Big[(s+l\varphi)I-(A_N(N,S)+\Delta^{(N)})\Big]^{-1}\times$$

$$\times\Big(\varphi\mathbf{e}+(L_0(N,\tilde{S})P_{N-1}(\boldsymbol{\beta}^{(1)})+(l-1)\varphi I_{T_N})\mathbf{w}(s,l-1)\Big),$$

$$l=\overline{2,K-N}.$$

Следствие 9.1. *Среднее время ожидания V^{wait} произвольного запроса первого типа рассчитывается как*

$$V^{wait} = -\lambda_1^{-1} \sum_{i=0}^{\infty} \sum_{k=N}^{K-1} \boldsymbol{\pi}(i,k)(D_1\mathbf{e} \otimes I_{T_N})\mathbf{w}'(s, k-N+1)|_{s=0},$$

где векторы-столбцы $\mathbf{w}'(s,l)|_{s=0}$, $l = \overline{1, K-N}$, могут быть вычислены с помощью рекурсии

$$\mathbf{w}'(s,1)|_{s=0} = \left[-\varphi I + (A_N(N,S) + \Delta^{(N)})\right]^{-1}\mathbf{e},$$

$$\mathbf{w}'(s,l)|_{s=0} = \left[-l\varphi I + (A_N(N,S) + \Delta^{(N)})\right]^{-1} \times$$

$$\times[\mathbf{e} - (L_0(N,\tilde{S})P_{N-1}(\boldsymbol{\beta}^{(1)}) + (l-1)\varphi I_{T_N})\mathbf{w}'(s,l-1)|_{s=0}],$$

$$l = \overline{2, K-N}.$$

9.6 Численный эксперимент

Цели представленного в этом подразделе численного примера следующие: продемонстрировать целесообразность использования предложенных алгоритмов для расчета ключевых показателей эффективности системы, проиллюстрировать поведение рассматриваемой системы. Мы предполагаем, что входной поток определяется матрицами:

$$D_0 = \begin{pmatrix} -0.8109843388145143 & 0 \\ 0 & -0.026322138550243148 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 0.20139810624083473 & 0.0013479784627937906 \\ 0.003665226557387846 & 0.0029153080801729413 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 0.6041943187225042 & 0.004043935388381372 \\ 0.010995679672163538 & 0.008745924240518824 \end{pmatrix}.$$

Средняя скорость прибытия запросов составляет $\lambda = 0.6$. Коэффициент корреляции времен между последовательными поступлениями запросов равен $c_{cor} = 0.2$, а коэффициент вариации времени между поступлениями составляет $c_{var} = 12.34$. Средняя скорость поступления запросов первого типа λ_1 составляет 0.15, а средняя скорость прибытия запросов второго типа λ_2 составляет 0.45.

Важность учета корреляции входного потока в похожей на нашу модель была продемонстрирована в [147]. Поэтому мы не будем рассматривать вопросы, связанные с влиянием корреляции на параметры системы в данном примере. Одним из существенных обобщений модели [147], которая рассмотрена в данном разделе, является рассмотрение более сложного процесса обслуживания. Для изучения влияния дисперсии времени обслуживания запросов рассмотрим два набора распределений времени обслуживания с одинаковым средним временем обслуживания запросов первого типа $b_1^{(1)} = 2$ и запросов второго типа $b_1^{(2)} = 3\frac{1}{3}$.

В первом случае, обозначаемом как $M + M$, мы предполагаем, что процесс обслуживания запросов первого типа определяется вектором $\boldsymbol{\beta}_1 = (1)$ и матрицей $S_1 = (-0.5)$, а процесс обслуживания запросов второго типа определяется вектором $\boldsymbol{\beta}_2 = (1)$ и матрицей $S_2 = (-0.3)$. Обратим внимание, что в этом случае времена обслуживания запросов первого и второго типов имеют показательное распределение с параметрами 0.5 и 0.3 соответственно. Коэффициент вариации времени обслуживания запросов r-го типа составляет $c_{var}^{(r)} = 1$, $r = 1, 2$.

Во втором случае, обозначаемом как $PH + PH$, мы предполагаем, что время обслуживания запросов первого типа определяется вектором $\boldsymbol{\beta}_1 = (0.05, 0.95)$ и матрицей $S_1 = \begin{pmatrix} -0.03104 & 0 \\ 0 & -2.441 \end{pmatrix}$. Коэффициент вариации такого времени обслуживания равен $(c_{var}^{(1)})^2 = 25.027$. Процесс обслуживания запросов второго типа определяется вектором $\boldsymbol{\beta}_2 = (0.1, 0.9)$ и матрицей $S_2 = \begin{pmatrix} -0.03359 & 0 \\ 0 & -2.52625 \end{pmatrix}$. Коэффициент вариации равен $(c_{var}^{(2)})^2 = 14.979$.

Остальные параметры системы предполагаются следующими:

$$N = 7,\ K = 9,\ \gamma = 0.5,\ \varphi = 0.3,\ \alpha = 2,\ q = 0.7,\ p = 0.85.$$

Будем изменять порог R в интервале $[1, N]$. Таблицы 9.2, 9.3 показывают зависимость некоторых характеристик производительности системы от порога R для случаев $M + M$ (Таблица 9.2) и $PH + PH$ (Таблица 9.3).

Таблица 9.2. Зависимость характеристик от порога R для случая $M + M$

R	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	$R = 6$	$R = 7$
$P_1^{(loss)}$	7.79E-8	3.82E-7	2.01E-6	1.06E-5	5.41E-5	2.602E-4	0.00118
$P_2^{(loss)}$	0.70952	0.44918	0.2516	0.12314	0.0523	0.01936	0.00649

N_{server}	0.7357	1.1262	1.4226	1.6153	1.7215	1.771	1.7904
N_{buffer}	3.77E-8	1.85E-7	9.75E-7	5.14E-6	2.61E-5	1.25E-4	5.68E-4
$P_1^{(arr-loss)}$	3.5E-10	1.7E-9	9.8E-9	5.4E-8	2.88E-7	1.45E-6	6.89E-6
$P_2^{(arr-loss)}$	0.23138	0.15809	0.09464	0.04905	0.02189	0.00843	0.00285
$P_1^{(imp-loss)}$	1.1E-8	5.5E-8	2.9E-7	1.54E-6	7.82E-6	3.75E-5	1.703E-4
$P_1^{(imm)}$	0.999999	0.999995	0.999979	0.999897	0.999507	0.997770	0.990501
$P_2^{(imm)}$	0.22873	0.47304	0.68455	0.83650	0.92706	0.97190	0.990501
V^{wait}	2.5E-7	1.2E-6	6.5E-6	3.42E-5	1.739E-4	8.352E-4	0.003784

Таблица 9.3. Зависимость характеристик от порога R для случая $PH + PH$

R	$R = 1$	$R = 2$	$R = 3$	$R = 4$	$R = 5$	$R = 6$	$R = 7$
$P_1^{(loss)}$	8.6E-8	3.3E-7	1.6E-6	9.6E-6	6E-5	3.78E-4	0.00224
$P_2^{(loss)}$	0.70831	0.44153	0.24301	0.1169	0.04885	0.01784	0.00622
N_{server}	0.7375	1.1377	1.4355	1.6246	1.7267	1.7733	1.7909
N_{buffer}	3.99E-8	1.53E-7	7.68E-7	4.46E-6	2.78E-5	1.74E-4	0.00102
$P_1^{(arr-loss)}$	9.8E-10	3.7E-9	1.8E-8	1.08E-7	6.83E-7	4.44E-6	2.84E-5
$P_2^{(arr-loss)}$	0.22320	0.14609	0.08418	0.04226	0.01837	0.00692	0.00230
$P_1^{(imp-loss)}$	1.2E-8	4.6E-8	2.3E-7	1.3E-6	8.34E-6	5.23E-5	3.07E-4
$P_1^{(imm)}$	0.999999	0.999998	0.999991	0.999949	0.999699	0.998349	0.992335
$P_2^{(imm)}$	0.25600	0.51304	0.71941	0.85914	0.93878	0.97693	0.992335
V^{wait}	2.6E-7	1.0E-6	5.1E-6	2.97E-5	1.85E-4	0.001163	0.006835

На рисунке 9.2 графически показана зависимость вероятности потери произвольного запроса первого типа P_1^{loss} от параметра R.

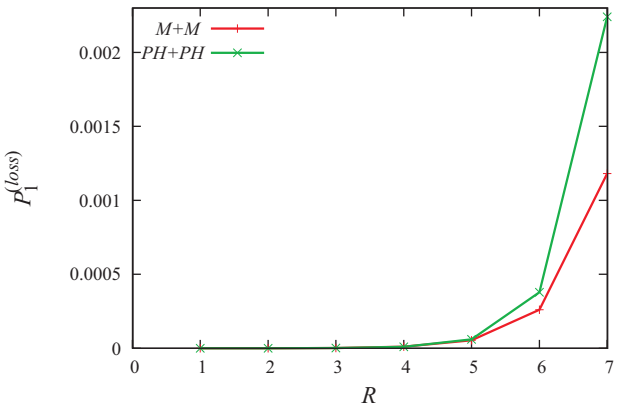

Рисунок 9.2.– Зависимость $P_1^{(loss)}$ от порога R

Таблицы 9.2 и 9.3 показывают, что некоторые показатели эффективности системы чувствительны к дисперсии времени обслуживания. Например, вероятность $P_1^{(arr-loss)}$ в случае $PH + PH$ для некоторых значений R

в два-три раза выше, чем для случая $M+M$. Результаты, представленные в таблицах 9.2 и 9.3, обеспечивают количественное подтверждение интуитивно понятному факту, что увеличение порога R приводит к улучшению качества обслуживания запросов второго типа и ухудшению качества обслуживания запросов первого типа. Таким образом, на основе упомянутых выше аналитических и алгоритмических результатов можно формулировать и решать различные задачи оптимизации, например проблему выбора значения порога R таким образом, что одна из заданных характеристик системы является минимальной или максимальной при условии, что еще одна характеристика превышает (или не превышает) наперед заданную величину.

Другим существенным обобщением модели [147], которая рассматривается в данном разделе, является предположение о том, что хэндовер-запросы (запросы первого типа) могут быть помещены в буфер, когда все приборы заняты. Для анализа влияния буферизации зафиксируем порог $R=6$ и будем изменять емкость буфера $K-N$ в интервале $[0,6]$. В таблице 9.4 представлена зависимость некоторых показателей производительности системы от емкости буфера $K-N$ в случае $M+M$.

Таблица 9.4. Зависимость характеристик от емкости буфера $K-N$ в случае $M+M$

буфер	0	1	2	3	4	5	6
$P_1^{(loss)}$	0.002085	3.714E-4	2.602E-4	2.536E-4	2.532E-4	2.532E-4	2.532E-4
$P_2^{(loss)}$	0.019618	0.019379	0.019361	0.019360	0.019359	0.019359	0.019359
N_{server}	1.770815	1.770975	1.770988	1.770989	1.770989	1.770989	1.770989
N_{buffer}	0.0	1.111E-4	1.253E-4	1.265E-4	1.266E-4	1.266E-4	1.266E-4
$P_1^{(arr-loss)}$	3.13E-4	2.24E-5	1.45E-6	8.58E-8	4.67E-9	2.36E-10	1.11E-11
$P_2^{(arr-loss)}$	0.008405	0.008427	0.008429	0.008429	0.008429	0.008429	0.008429
$P_1^{(imp-loss)}$	0	3.33E-5	3.75E-5	3.79E-5	3.79E-5	3.79E-5	3.79E-5
$P_1^{(imm)}$	0.99791	0.99778	0.99777	0.99777	0.99777	0.99777	0.99777
$P_2^{(imm)}$	0.97198	0.97191	0.97190	0.97190	0.97190	0.97190	0.97190
V^{wait}	0	7.41E-4	8.35E-4	8.43E-4	8.44E-4	8.44E-4	8.44E-4

В таблице 9.4, в частности, показано, что когда предоставляется одно место в буфере, вероятность $P_1^{(arr-loss)}$ потери запроса первого типа на входе в систему становится в 10-20 раз меньше. Таким образом, буферизация оказывает важное положительное влияние. Тем не менее, не имеет смысла делать емкость буфера больше двух, поскольку увеличение буферной

емкости влечет за собой увеличение вероятности $P_1^{(imp-loss)}$ потери запроса первого типа из-за нетерпеливости, а интегральная вероятность $P_1^{(loss)}$ потери запроса первого типа снижается несущественно.

ПРИЛОЖЕНИЕ А

НЕКОТОРЫЕ СВЕДЕНИЯ ИЗ ТЕОРИИ МАТРИЦ И ФУНКЦИЙ ОТ МАТРИЦ

1 Стохастические и субстохастические матрицы. Генераторы и субгенераторы

Пусть $A = (a_{ij})_{i,j=\overline{1,n}}$ – некоторая квадратная матрица размерности $n \times n$.

Характеристическим многочленом матрицы A называют многочлен вида $\det(\lambda I - A)$. Характеристический многочлен имеет степень n. Уравнение вида

$$\det(\lambda I - A) = 0 \qquad (A1.1)$$

называют характеристическим уравнением.

Корни характеристического уравнения называют характеристическими числами матрицы A.

Квадратная матрица $P = (p_{ij})_{i,j=\overline{1,n}}$ называется стохастической, если для всех $i,\ i = \overline{1,n}$, выполняются неравенства $p_{ij} \geq 0,\ j = \overline{1,n}$, и равенство $\sum\limits_{j=1}^{n} p_{ij} = 1$.

Квадратная матрица $\tilde{P} = (\tilde{p}_{ij})_{i,j=\overline{1,n}}$ называется субстохастической, если для всех $i,\ i = \overline{1,n}$, выполняются неравенства $\tilde{p}_{ij} \geq 0,\ j = \overline{1,n}$, и неравенство $\sum\limits_{j=1}^{n} \tilde{p}_{ij} \leq 1$.

Квадратная матрица $Q = (q_{ij})_{i,j=\overline{1,n}}$ называется генератором, если для всех $i,\ i = \overline{1,n}$, выполняются условия $q_{ii} < 0$, $q_{ij} \geq 0,\ j = \overline{1,n},\ j \neq i$, и равенство $\sum\limits_{j=1}^{n} q_{ij} = 0$.

Квадратная матрица $Q = (\tilde{q}_{ij})_{i,j=\overline{1,n}}$ называется субгенератором, если для всех $i,\ i = \overline{1,n}$, выполняются условия $\tilde{q}_{ii} < 0$, $\tilde{q}_{ij} \geq 0,\ j = \overline{1,n},\ j \neq i$, и неравенство $\sum\limits_{j=1}^{n} \tilde{q}_{ij} \leq 0$.

Нетрудно видеть, что если матрица P является стохастической, то матрица $P - I$ является генератором. Если матрица $\tilde{P}$ является субстохастической, то матрица $\tilde{P} - I$ является субгенератором. Поэтому свойства стохастических матриц и генераторов, субстохастических матриц и субгенераторов связаны.

Лемма. Неотрицательная матрица является стохастической тогда и только тогда, когда она имеет собственный вектор $(1, 1, \ldots, 1)$, соответствующий характеристическому числу 1. Характеристическое число 1 является максимальным для стохастической матрицы.

Из этой леммы следует, что если матрица A стохастическая, то матрица $I - A$ вырожденная.

Лемма. Генератор имеет собственный вектор $(1, 1, \ldots, 1)$, соответствующий характеристическому числу 0.

Теорема (Теорема Адамара)**.** Если элементы a_{ij} матрицы A, таковы, что для всех i, $i = \overline{1, n}$, выполняются неравенства $|a_{ii}| > \sum\limits_{j=1, j\neq i}^{n} |a_{ij}|$, то есть диагональные элементы доминируют, то матрица A – невырожденная.

Теорема (Теорема О. Таусски)**.** Если матрица A – неразложимая (неприводимая), для всех i, $i = \overline{1, n}$, выполняются неравенства $|a_{ii}| \geq \sum\limits_{j=1, j\neq i}^{n} |a_{ij}|$ и хотя бы для одного i неравенство выполнено строго, то матрица A – невырожденная.

Лемма. Пусть $\tilde{P}$ – субстохастическая матрица. Если существует ненулевая неотрицательная матрица $\bar{P}$ такая, что матрица $P = \tilde{P} + \bar{P}$ является стохастической и неприводимой, то матрица $\tilde{P}$ является невырожденной.

Лемма. Пусть $\tilde{Q}$ – субгенератор. Если существует ненулевая неотрицательная матрица $\bar{Q}$ такая, что матрица $Q = \tilde{Q} + \bar{Q}$ является неприводимым генератором, то матрица $\tilde{Q}$ является невырожденной. Все ее характеристические числа имеют отрицательную действительную часть.

Леммы А3 и А4 доказываются аналогично. Докажем лемму А4.

Если субгенератор $\tilde{Q}$ таков, что неравенства $\sum\limits_{j=1}^{n} \tilde{q}_{ij} \leq 0$ выполняются строго для всех i, $i = \overline{1, n}$, то невырожденность субгенератора следует из теоремы Адамара. То, что характеристические числа матрицы $\tilde{Q}$ имеют отрицательную действительную часть, доказывается от противного. Пусть существует характеристическое число λ, такое, что $Re\ \lambda \geq 0$. Тогда должно выполняться равенство $\det(\lambda I - \tilde{Q}) = 0$. Но это невозможно по теореме Адамара, так как все диагональные элементы матрицы $\lambda I - \tilde{Q}$ заведомо доминируют, поскольку это имеет место уже для матрицы $\tilde{Q}$.

Пусть не все неравенства $\sum_{j=1}^{n} \tilde{q}_{ij} \leq 0$ выполняются строго. Наличие хотя бы одного $i,\ i = \overline{1,n}$, такого, что неравенство $\sum_{j=1}^{n} \tilde{q}_{ij} \leq 0$ выполняется строго, следует из предположений леммы о том, что матрица $\bar{Q}$ неотрицательная и имеет положительные элементы, а матрица $Q = \tilde{Q} + \bar{Q}$ является генератором и, следовательно, имеет нулевые суммы элементов во всех строках.

Если матрица $\tilde{Q}$ является неприводимой, утверждение леммы следует непосредственно из теоремы О. Таусски.

Пусть теперь матрица $\tilde{Q}$ является приводимой. Тогда перестановкой строк и столбцов она может быть приведена к канонической нормальной форме

$$\tilde{Q} = \begin{pmatrix} Q^{(1)} & O & \dots & O & O \\ O & Q^{(2)} & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & Q^{(m)} & O \\ Q^{(m+1,1)} & Q^{(m+1,2)} & \dots & Q^{(m+1,m)} & Q^{(m+1)} \end{pmatrix}, \qquad (A1.2)$$

где матрицы $Q^{(r)},\ r = \overline{1,m+1}$, являются неразложимыми, а среди матриц $Q^{(m+1,r)},\ r = \overline{1,m}$, есть хотя бы одна ненулевая.

Известно, что

$$\det\tilde{Q} = \prod_{r=1}^{m+1} \det Q^{(r)}.$$

Матрица $Q^{(m+1)}$ является невырожденной в силу теоремы О. Таусски, поскольку она неприводимая, в ней имеется нестрогое доминирование диагональных элементов во всех строках и строгое доминирование хотя бы в одной строке, что следует из того, что среди матриц $Q^{(m+1,r)},\ r = \overline{1,m}$, есть хотя бы одна ненулевая.

Матрицы $Q^{(r)},\ r = \overline{1,m}$, являются также невырожденными в силу теоремы О. Таусски. Наличие строгого доминирования хотя бы в одной строке следует из того, что в противном случае матрица $Q^{(r)}$ является генератором, из чего следует, что каноническая нормальная форма матрицы Q имеет структуру, аналогичную (А1.2), что противоречит предположению о ее неприводимости. Таким образом, $\det\tilde{Q}$ является произведением ненулевых определителей и сам является ненулевым. То есть матрица $\tilde{Q}$ является невырожденной.

Утверждение об отрицательности действительной части характеристических чисел проводится совершенно аналогично рассмотренному выше случаю, когда строгое доминирование имеется во всех строках матрицы $\tilde{Q}$. $\square$

Следствие. Матрица D_0 в определении $BMAP$-потока является невырожденной. Все ее характеристические числа имеют отрицательную действительную часть.

Матрица A называется устойчивой, если все ее характеристические числа имеют отрицательную действительную часть. Матрица B является полуустойчивой, если действительные части всех ее характеристических чисел отрицательные или равны нулю.

Лемма. Если матрица A устойчивая, то она невырожденная.

Лемма. Если матрица A такова, что $A^n \to O$ при $n \to \infty$, то матрица $I - A$ имеет обратную, причем

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

Лемма. Пусть матрица P является матрицей переходных вероятностей неприводимой ЦМ, а стохастическая матрица A – ее предельная матрица, то есть $P^n \to A$ при $n \to \infty$. Тогда матрица A имеет вид:

$$A = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha} \\ \vdots \\ \boldsymbol{\alpha} \end{pmatrix},$$

где $\boldsymbol{\alpha}$ – левый стохастический собственный вектор матрицы, P соответствующий собственному значению 1, то есть этот вектор является единственным решением системы $\boldsymbol{\alpha}P = \boldsymbol{\alpha}$, $\boldsymbol{\alpha}\mathbf{e} = 1$. Вектор $\boldsymbol{\alpha}$ называется предельным вектором неприводимой ЦМ.

Справедливы соотношения $PA = AP = A$.

Матрица Z, заданная соотношением $Z = (I - (P - A))^{-1}$, называется фундаментальной матрицей неприводимой ЦМ с матрицей переходных вероятностей P, с предельной матрицей A и предельным вектором $\boldsymbol{\alpha}$.

С учетом того, что справедливо соотношение $A^k = A,\ k \geq 1$, матрица Z также определяется соотношениями:

$$Z = \sum_{n=0}^{\infty}(P - A)^n = I + \sum_{n=1}^{\infty}(P^n - A).$$

Лемма. Справедливы формулы:

$$PZ = ZP,\ Z\mathbf{e} = \mathbf{e},\ \boldsymbol{\alpha}Z = \boldsymbol{\alpha}, I - Z = A - PZ.$$

Лемма. Пусть матрица P неприводимая стохастическая, а $\boldsymbol{\alpha}$ – левый стохастический собственный вектор матрицы P.

Тогда матрица $I - P + \mathbf{e}\boldsymbol{\alpha}$ является невырожденной.

Доказательство. Утверждение леммы будет непосредственно следовать из леммы 6, если мы покажем, что выполняется соотношение

$$(P - \mathbf{e}\boldsymbol{\alpha})^n \to O$$

при $n \to \infty$.

Согласно введенному обозначению, выполняются соотношения

$$\mathbf{e}\boldsymbol{\alpha} = A,\ PA = A,\ P^n \to A \text{ при } n \to \infty.$$

С учетом этих соотношений по формуле бинома Ньютона имеем:

$$(P - A)^n = \sum_{k=0}^{n} C_n^k P^{n-k}(-1)^k A^k = P^n + \sum_{k=1}^{n} C_n^k P^{n-k}(-1)^k A^k =$$

$$= P^n + \sum_{k=1}^{n} C_n^k (-1)^k A = P^n - A \to O.$$

□

Лемма. Пусть $A_i,\ i \geq 0$, – субстохастические матрицы порядка $K \times K$ и их матричная ПФ $A(z) = \sum\limits_{i=0}^{\infty} A_i z^i$ такая, что $A(1)$ — неразложимая стохастическая матрица и $A'(z)|_{z=1} < \infty$. Тогда существует вектор $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_K)^T$, знак каждой из компонент которого совпадает со знаком определителя $(\det(zI - A(z)))'|_{z=1}$ (или равный нулю, если

$(\det(zI - A(z)))'|_{z=1} = 0)$ такой, что система линейных алгебраических уравнений для компонент вектора $\mathbf{u} = (u_1, \ldots, u_K)^T$

$$(I - A(1))\mathbf{u} = (A(z) - zI)'|_{z=1}\mathbf{e} + \mathbf{\Delta} \qquad (A1.3)$$

имеет бесконечное множество решений.

Доказательство. Так как матрица $A(1)$ стохастическая и неразложимая, то ранг матрицы $I - A(1)$ равен $K - 1$. Пусть $D_m(\mathbf{\Delta})$ — определитель, полученный путем замены m-го столбца $\det(I - A(1))$ на столбец свободных членов системы (A1.3). Система (A1.3) имеет решение тогда и только тогда, когда

$$D_m(\mathbf{\Delta}) = 0,\ m = \overline{1, K}.$$

Разлагая определитель $D_m(\mathbf{\Delta})$ по m-му столбцу, получим систему линейных уравнений для элементов вектора $\mathbf{\Delta}$:

$$\mathbf{\nabla}_m\mathbf{\Delta} = -\mathbf{\nabla}_m(A(z) - zI)'|_{z=1}\mathbf{e},\ m = \overline{1, K}. \qquad (A1.4)$$

Здесь $\mathbf{\nabla}_m$ — вектор-строка алгебраических дополнений m-го столбца определителя $\det(I - A(1))$. Легко видеть, что матрица коэффициентов при неизвестных в (A1.4) является матрицей $\mathrm{Adj}(I - A(1))$ алгебраических дополнений определителя $\det(I - A(1))$. Известно, что размерность пространства правых нулевых векторов стохастической неразложимой матрицы равна единице. Из соотношений

$$(I - A(1))\mathrm{Adj}(I - A(1)) = \det(I - A(1))I = 0$$

следует, что каждый из столбцов матрицы $\mathrm{Adj}(I - A(1))$ является правым нулевым вектором матрицы $I - A(1)$. Очевидно, вектор $\mathbf{e}$ — также правый нулевой вектор этой матрицы. Тогда все столбцы $\mathrm{Adj}(I - A(1))$ с точностью до постоянных множителей c_k совпадают с вектором $\mathbf{e}$, то есть

$$\mathrm{Adj}(I - A(1))_{(k)} = c_k\mathbf{e},\ k = \overline{1, K},$$

а сама матрица $\mathrm{Adj}(I - A(1))$ имеет вид:

$$\mathrm{Adj}(I - A(1)) = \begin{pmatrix} c_1 & c_2 & \cdots & c_n \\ c_1 & c_2 & \cdots & c_n \\ \cdot & \cdot & \cdots & \cdot \\ c_1 & c_2 & \cdots & c_n \end{pmatrix}.$$

Известно, что в случае неразложимой стохастической матрицы $A(1)$ все алгебраические дополнения определителя $\det(I - A(1))$ положительны, то есть

$$c_k > 0,\ k = \overline{1, K}. \qquad (A1.5)$$

Нетрудно показать, что правая часть (A1.4) равна $(\det(zI - A(z)))'|_{z=1}$. Очевидно также, что

$$\boldsymbol{\nabla}_m = (c_1, c_2, \ldots, c_n),\ m = \overline{1, K}.$$

Тогда система (A1.4) эквивалентна одному уравнению

$$\sum_{k=1}^{n} c_k \Delta_k = (\det(zI - A(z)))'|_{z=1}. \qquad (A1.6)$$

Из (A1.5), (A1.6) следует, что существует вектор $\boldsymbol{\Delta}$, знак которого совпадает со знаком правой части (A1.6) (или равный нулю, если правая часть (A1.6) равна 0), удовлетворяющий уравнению (A1.6). Этот вектор является и решением системы (A1.4). Подставляя это решение в (A1.3), получим систему линейных алгебраических уравнений для компонент вектора $\mathbf{u}$, которая имеет бесконечное множество решений. □

2 Функции от матриц

Функции от матриц могут быть заданы разными способами. Один из них следующий. Пусть $f(\lambda)$ – некоторая функция. Если функция $f(\lambda)$ является полиномом, то есть

$$f(\lambda) = a_0 + a_1\lambda + \ldots + a_K\lambda^K,$$

то для любой матрицы A функция $f(A)$ от этой матрицы определяется соотношением

$$f(A) = a_0 I + a_1 A + \ldots + a_K A^K.$$

Пусть $\tilde{\lambda}_k, k = \overline{1, n_1}$ – характеристические числа $n \times n$ матрицы A, где n_1 – количество различных характеристических чисел, $\tilde{r}_k, k = \overline{1, n_1}$ – их кратности, $\sum\limits_{k=1}^{n_1} \tilde{r}_k = n$, $\tilde{r}_k \geq 1$.

Множество чисел $\tilde{\lambda}_k$ образуют спектр матрицы A. Максимум из модулей характеристических чисел матрицы A называют спектральным радиусом этой матрицы и обозначают $\varrho(A)$. Если матрица A – неотрицательная,

то характеристическое число, на котором достигается максимум, является действительным и ему соответствует неотрицательный собственный вектор. Спектральный радиус стохастической матрицы равен 1, а соответствующий собственный вектор есть $\mathbf{e}$. Если вычислить дополнительные миноры $(n-1)$-го порядка для всех элементов матрицы $\lambda I - A$ и их наибольший общий делитель обозначить через $D_{n-1}(\lambda)$, то многочлен

$$\psi(\lambda) = \frac{\det(\lambda I - A)}{D_{n-1}(\lambda)}$$

называют минимальным характеристическим многочленом матрицы A с корнями $\lambda_k, k = \overline{1, n_1}$, кратности $r_k,\ r_k \le \tilde{r}_k$.

Тогда для любой функции $f(\lambda)$ такой, что для всех $\lambda_k,\ k = \overline{1, n_1}$, существуют $f(\lambda_k), ..., f^{(r_k-1)}(\lambda_k), k = \overline{1, n_1}$, функция $f(A)$ существует и определяется следующим образом:

$$f(A) = \sum_{k=1}^{n_1} \sum_{j=0}^{r_k-1} f^{(j)}(\lambda_k) Z_{kj}, \qquad (A1.7)$$

где Z_{kj} – некоторые матрицы, не зависящие от выбора функции f, называемые составляющими матрицами данной матрицы A, причем

$$Z_{kj} = (j!)^{-1}(A - \lambda_k I)^j Z_{k0},\ j \ge 0.$$

Для иллюстрации приведем пример вычисления матриц Z_{kj}. Пусть матрица A имеет вид

$$A = \begin{pmatrix} 2 & -1 & 1 \\ 0 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix}.$$

Характеристическое уравнение запишется как

$$\det(\lambda I - A) = (\lambda - 1)^2(\lambda - 2) = 0.$$

Это означает, что матрица A имеет характеристические числа $\lambda_1 = 1$ кратности $r_1 = 2$ и $\lambda_2 = 2$ кратности $r_2 = 1$.

Наибольший общий делитель $D_2(\lambda) = 1$. Поэтому характеристический и минимальный многочлены равны. Уравнение (А1.7) принимает вид:

$$f(A) = f(1)Z_{10} + f'(1)Z_{11} + f(2)Z_{20}. \qquad (A1.8)$$

Будем формировать систему уравнений для нахождения матриц Z_{ij} путем выбора различных функций $f(\lambda)$ в уравнении (A1.8). Взяв функцию $f(\lambda) = 1$, получаем уравнение

$$I = Z_{10} + Z_{20}, \qquad (A1.9)$$

взяв функцию $f(\lambda) = \lambda - 1$, получаем уравнение

$$A - I = Z_{11} + Z_{20}, \qquad (A1.10)$$

взяв функцию $f(\lambda) = (\lambda - 1)^2$, получаем уравнение

$$(A - I)^2 = Z_{20}. \qquad (A1.11)$$

Решая систему уравнений (A1.9)-(A1.11) и подставляя полученные выражения для матриц Z_{ij} в уравнение (A1.8), получаем формулу

$$f(A) = f(1)\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix} + f'(1)\begin{pmatrix} 1 & -1 & 1 \\ 1 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix} + f(2)\begin{pmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 1 & 0 \end{pmatrix}.$$

Эта формула позволяет легко вычислять различные функции от матрицы .Например,

$$e^A = e^1\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix} + e^1\begin{pmatrix} 1 & -1 & 1 \\ 1 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix} + e^2\begin{pmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 1 & 0 \end{pmatrix}.$$

Формула (A1.7) иногда называется формулой разложения матрицы на спектре. Она бывает полезной при проверке некоторых неочевидных свойств функций от матриц. Проиллюстрируем это на примере доказательства следующего утверждения.

Следствие. Для матрицы D_0 из определения $BMAP$-потока интеграл $\int\limits_0^\infty e^{D_0 t}dt$ существует и равен $(-D_0)^{-1}$.

Доказательство. Пусть $\lambda_k, k = \overline{1, n_1}$, корни кратности r_k минимального характеристического многочлена матрицы D_0. Тогда, согласно (A1.7), для любой функции $f(\lambda)$ такой, что она и соответствующее число ее производных существует в точках $\lambda_k, k = \overline{1, n_1}$, справедлива формула

$$f(D_0) = \sum_{k=1}^{n_1}\sum_{j=0}^{r_k-1} f^{(j)}(\lambda_k)Z_{kj}, \qquad (A1.12)$$

где Z_{kj} – составляющие матрицы для матрицы D_0.

Возьмем в качестве функции $f(\lambda)$ функцию $f(\lambda) = \int\limits_0^\infty e^{\lambda t} dt$. Из следствия 1 следует, что эта функция и ее производные существуют во всех точках $\lambda_k, k = \overline{1, n_1}$. Поэтому

$$\int\limits_0^\infty e^{D_0 t} dt = \sum_{k=1}^{n_1} \sum_{j=0}^{r_k - 1} \int\limits_0^\infty (e^{\lambda_k t})^{(j)} dt Z_{kj} = \sum_{k=1}^{n_1} \sum_{j=0}^{r_k - 1} \int\limits_0^\infty t^j e^{\lambda_k t} dt Z_{kj} =$$

$$= \sum_{k=1}^{n_1} \sum_{j=0}^{r_k - 1} \frac{j!}{(-\lambda_k)^{j+1}} Z_{kj} = \sum_{k=1}^{n_1} \sum_{j=0}^{r_k - 1} (\frac{1}{-\lambda_k})^{(j)} Z_{kj} = (-D_0)^{-1}. \qquad (A1.13)$$

Последний переход в (A1.13) сделан с использованием формулы (A1.12) для функции $f(\lambda) = \frac{1}{-\lambda}$. □

Отметим следующее полезное свойство функций от матриц. Если сложная функция $h(\lambda) = g(\varphi(\lambda))$ существует на спектре матрицы A (то есть она и соответствующее число ее производных существует во всех точках, являющихся характеристическими числами этой матрицы), то справедлива формула $h(A) = g(\varphi(A))$.

3 Нормы матриц

Для оценивания близости матриц часто используют понятие нормы от матрицы. Пусть A – квадратная матрица размера $n \times n$ с элементами $a_{i,j},\ i, j = \overline{1, n}$. Функцию $||A||$ матрицы A, принимающую значения в множестве действительных чисел, называют нормой, если она удовлетворяет следующим аксиомам:

(1) $||A|| \geq 0$ (неотрицательность);

(1a) $||A|| = 0$ тогда и только тогда, когда $A = 0$ (положительность);

(2) $||cA|| = |c|||A||$ для всех комплексных чисел c (абсолютная однородность);

(3) $||A + B|| \leq ||A|| + ||B||$ (неравенство треугольника);

(4) $||AB|| \leq ||A||||B||$ (кольцевое свойство).

Наиболее популярными нормами матриц являются следующие:

- Норма l_1, определенная как $||A||_{l_1} = \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} |a_{i,j}|$;

- Норма l_2 (евклидова норма, норма Фробениуса, норма Шура, норма Гильберта – Шмидта), определенная как $||A||_{l_2} = (\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}|a_{i,j}|^2)^{\frac{1}{2}}$;
- Норма l_∞, определенная как $||A||_{l_\infty} = \max\limits_{i,j=\overline{1,n}}|a_{i,j}|$;
- Максимальная столбцовая норма $||A||_1$, определенная как

$$||A||_1 = \max_{j=\overline{1,n}}\sum_{i=1}^{n}|a_{i,j}|;$$

- Максимальная строчная норма $||A||_\infty$, определенная как

$$||A||_\infty = \max_{i=\overline{1,n}}\sum_{j=1}^{n}|a_{i,j}|.$$

Справедливо следующее утверждение. Для любой матричной нормы $||A||$ любой матрицы A выполнено неравенство $\varrho(A) \leq ||A||$, где $\varrho(A)$ – спектральный радиус матрицы A.

4 Кронекеровы произведение и сумма матриц

Пусть имеются матрицы $A = (a_{ij})_{i=\overline{1,n},j=\overline{1,m}}$ порядка $n \times m$ и $B = (b_{ij})_{i=\overline{1,r},j=\overline{1,s}}$ порядка $r \times s$.

Кронекеровым произведением этих матриц будем называть матрицу

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}$$

порядка $nr \times ms$.

Свойства кронекерова произведения:

1. $(A + B) \otimes C = A \otimes C + B \otimes C$.
2. $A \otimes (B + C) = A \otimes B + A \otimes C$.
3. $(\alpha A) \otimes B = A \otimes \alpha B = \alpha A \otimes B, \alpha = const$.
4. $A \otimes (B \otimes C) = (A \otimes B) \otimes C$.
5. $(A \otimes B)^T = A^T \otimes B^T$.
6. Если матрицы A и B квадратные и невырожденные, то $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

7. Вообще говоря, матрицы $(A \otimes B)$ и $(B \otimes A)$ не равны. Но существуют матрицы перестановок P и Q, такие, что

$$(A \otimes B) = P(B \otimes A)Q.$$

Если при этом матрицы A и B квадратные, то $P = Q^T$.

8. Если матрицы A, B, C, D таковы, что произведения матриц AB и CD имеют смысл, то справедлива формула

$$(AB) \otimes (CD) = (A \otimes C)(B \otimes D),$$

называемая правилом смешанного произведения (mixed product rule). Это правило легко распространяется и на случай произведения многих матриц.

9. Если матрицы A и B квадратные размера n и m, соответственно, то

$$detA \otimes B = (detA)^m (detB)^n,$$

$$trA \otimes B = (trA)(trB),$$

где trA означает след матрицы A, который равен сумме элементов главной диагонали матрицы A,

$$rankA \otimes B = (rankA)(rankB).$$

10. Если матрицы A и B квадратные, $\mathbf{x}$ и $\mathbf{y}$ – их собственные векторы, соответствующие собственным числам λ и μ соответственно, то вектор $\mathbf{x} \otimes \mathbf{y}$ является собственным вектором матрицы $A \otimes B$, соответствующим собственному числу $\lambda\mu$.

Пусть A и B – квадратные матрицы размерности $n \times n$ и $m \times m$ соответственно. Кронекеровой суммой этих матриц называется матрица

$$A \oplus B = A \otimes I_m + I_n \otimes B.$$

Полезным свойством кронекеровой суммы матриц является следующее:

$$e^A \otimes e^B = e^{A \oplus B}.$$

Если матрица A устойчива, а матрица G – полуустойчивая, то интеграл $\int\limits_0^\infty e^{At} \otimes e^{Gt} dt$ существует и равен $-(A \oplus G)^{-1}$.

Литература

1. Kleinrock L. Queueing System. Volume II: Computer Applications.- John Wiley & Sons. New York. 1976.

2. Schwartz M. Computer Communication Network. Design and Analysis.- Prentice-Hall. New Jersey. 1977.

3. Башарин Г.П., Бочаров П.П., Коган Я.А. Анализ очередей в вычислительных сетях. Теория и методы расчета.- Москва. Наука. 1989.

4. Вишневский В.М., Жожикашвили В.А. Сети массового обслуживания. Теория и применение в сетях ЭВМ.- Москва. Радио и связь. 1988.

5. Вишневский В.М. Теоретические основы проектирования компьютерных сетей.- Москва. Техносфера. 2003.

6. Вишневский В.М., Портной С.Л., Шахнович И.В. Энциклопедия WiMAX. Путь к 4G.- Москва. Техносфера. 2010.

7. M. A. Lema; E. Pardo; O. Galinina; S. Andreev; M. Dohler. Flexible Dual-Connectivity Spectrum Aggregation for Decoupled Uplink and Downlink Access in 5G Heterogeneous System // IEEE Journal on Selected Areas in Communications.- 2016.- №.99.- P.1-2.

8. S. Niknam, A. A. Nasir, H. Mehrpouyan and B. Natarajan. A Miltiband OFDMA Heterogeneous Network for Millimater Wave 5G Wireless Applications // IEEE Access.- 2016.- V. 4.- P. 640-648.

9. Vishnevsky V., Larionov A., Frolov S. Design and Scheduling in 5G Stationary and Mobile Communication Systems Based in Wireless Millimeter-Wave Mesh Network. Communication in Computer and Information Science.- 2014.- V. 279.- P. 11-27.

10. Vishnevsky V., Larionov A.A., Ivanov R.E. Applying Graph-Theoretic Approach for Time-Frequency Resource Allocation in 5G MmWave Backhaul Network // IEEE Xplore digital library.- 2016.- P. 71-78

11. Leland W.E., Taqqu Murad S., Willinger W., Wilson D.V. On the Self-Similar Nature of Ethernet traffic // Journal IEEE/ ACM Transactions on Networking.- 1994.- V. 2.- №. 1. P. 1-15.

12. Цыбаков Б.С. Модель телетрафика на основе самоподобного случайного процесса // Радиотехника. - 1999. - № 5. - С. 24-31.

13. Neuts M.F. Versatile Markovian point process // J. Appl. Probab.- 1979. - V. 16, № 4.- P. 764-779.

14. Ramaswami V. The $N/G/1$ queue and its detailed analysis // Advances Appl. Probab.- 1980.- V. 12.- No. 1.- P. 222-261.

15. Дудин А.Н., Листопад Н.И., Царенков Г.В. Улучшенный алгоритм оптимизации работы узла сети интернет / Пробл. проектирова-ния информ.-телекомм. систем. - Минск: Изд-во БГУ, 2001. - С. 28-43.

16. Neuts M.F. Structured stochastic matrices of $M/G/1$ type and their applications.- N. Y.: Marcel Dekker, 1989.

17. Lucantoni D.M. New results on the single server queue with a batch Markovian arrival process // Commun. Stat.- Stoch. Models.- 1991.- V. 7, №1.- P. 1-46.

18. Lucantoni D.M. The $BMAP/G/1$ queue: A Tutorial // Models and Techniques for Performance Evaluation of Computer and Communications Systems, L. Donatiello and R. Nelson Eds, London, Springer Verlag.- 1993.- P. 330-358.

19. Lucantoni D.M., Neuts M.F. Some steady-state distributions for the $MAP/SM/1$ queue // Commun. Statist. Stochast. Models.- 1994.- V. 10.- P. 575-598.

20. Ramaswami V., Lucantoni D. Algorithm for the multi-server queue with phase-type service // Commun. Statist.-Stochast. Models. - 1985. - V. 1. - P. 393–417.

21. Chakravarthy S.R. The batch Markovian arrival process: a review and fu-ture work // Advances Probab. Theory Stochast. Proc. - New Jersey, Notable Publicat, 2001. - P. 21-49.

22. Chakravarthy S., Dudin A.N. A multi-server retrial queue with $BMAP$ arrivals and group services // Queueing Syst.- 2002.- V. 42.- P. 5-31.

23. Chakravarthy S., Dudin A.N. Multi-threshold control for the $BMAP/SM/1/K$ queue with group services // J. Appl. Math. Stochast. Anal. - 2003. - V. 16, № 4. - P. 327-348.

24. Chakravarthy S., Dudin A.N. Analysis of a retrial queueing model with MAP arrivals and two types of customers // Math. Comput. Modell. - 2003. - V.37. - P. 343-364.

25. Kim Jiseung, Dudin A., Dudin S., Kim Chesoong Analysis of a Semi-Open Queueing Network with Markovian Arrival Process // Performance Evaluation. - 2018. - V. 120. - P. 1-19.

26. Vishnevsky V.M., Dudin A.N. Queueing systems with correlated arrival flows and their applications to modeling telecommunication // Automation and Remote Control. - 2017. - Vol. 78, Issue 8. - P. 1361–1403.

27. Dudin A., Nazarov A. On a tandem queue with retrials and losses and state dependent arrival, service and retrial rates // International Journal of Operational Research.- 2017. - V. 29, № 2. - P. 170-182.

28. Brugno A., Dudin A.N., Manzo R. Retrial Queue with Discipline of Adaptive Permanent Pooling // Applied Mathematical Modelling. - 2017. - V. 50. - P.1-16.

29. Chakravarthy S.R., Dudin A.N. A Queueing Model for Crowdsourcing // Journal of the Operational Research Society. - 2017. - V. 68, № 3. - P. 221–236.

30. Brugno A., Dudin A.N., Marzo R. Analisys of a strategy of adaptive group admission of customers to single server retrial system // Journal of Ambient Intelligence & Humanized Computing. - 2018. - V. 9. - P. 123–135.

31. Dudin A., Deepak T.G., Varghese C.J., Krishnamoorty A., Vishnevsky V. On a BMAP/G/1 Retrial System with Two Types of Search of Customers from the Orbit // Communications in Computer and Information Science. Volume. - 2017. - P. 1-12.

32. Dudin A., Lee M., Dudina O., Lee S. Analisys of priority retrial queue with many types of costumers and servers reservation as a model of cognitive radio system // IEEE Transactions on Communications. - 2017. - V. 65, № 1. - p.186-199.

33. Brugno A., D'Apice C., Dudin A.N., Manzo R. Analysis of a MAP/PH/1 Queue with Flexible Group Service // Applied Mathematics and Computer Science. - 2017. - V. 27, № 1. - P. 119-131.

34. Klimenok V., Dudin A., Samouylov K. Analysis of the $BMAP/PH/N$ queueing system with backup servers // Applied Mathematical Modelling.- 2018. - V. 57. - P. 64-84.

35. Klimenok V., Dudina O. Retrial Tandem queue with controlled strategy of repeated attempts / V. Klimenok, // Quality Technology and Quantitative Management. - 2017. - V. 14, № 1. - P. 74-93.

36. V.Klimenok, O.Dudina, V.Vishnevsky, K.Samuylov. Retrial Tandem Queue with BMAP-Input and Semi-Markovian Service Process // Communications in Computer and Information Science. Springer International Publishing. - 2017. - Vol. 700. - P. 159-173.

37. Kim C.S., Klimenok V., Dudin A. Analysis of unreliable $MAP/PH/N$ type queue with Markovian flow of breakdowns // Applied Mathematics and Computation. - 2017. - V. 314. - P. 154-172.

38. Klimenok V.I. Two-Server Queueing System with Unreliable Servers and Markovian Arrival Process // Communications in Computer and Information Science. - 2017. - V. 800. - P. 42-55.

39. Klimenok V., Dudin A., Samouylov K. Recursive Formulas for the Moments of Queue Length Distribution in the $BMAP/SM/1$ Queue // Operations Research Letters. - 2017. - V. 45, № 5. -P. 467-470.

40. Dudin A., Klimenok V., Samouylov K. Stationary Distribution of Waiting Time in $MAP/G/1/N$ Queueing System with LIFO Service Discipline // Lecture Notes in Computer Science. - 2017. - V. 10372. - P. 50–61.

41. Larionov A., Vishnevsky V., Ivanov R., Semenova O. Estimation of IEEE 802.11 DCF access Performance in Wireless Networks with Topology using PH Service Time Approximations and MAP Input. Proceedings 11th IEEE International Conference Application of Information and Communication Technologies. - 2017. - Vol. 2. - P. 85-89.

42. Вишневский В.М., Минниханов Р.Н., Дудин А.Н. и др. Новое поколение систем безопасности на автодорогах и их применение в интеллектуальных транспортных системах // Информ. технологии и вычисл. системы. - 2013. - № 4. - С.17-26.

43. Dudin A.N., Vishnevsky V.M., Sinjugina Ju.V. Analysis of the $BMAP/G/1$ queue with gated service and adaptive vacations duration // Telecommun. Syst. J.- 2016. - V.61, № 3. - P. 403-415.

44. Vishnevsky V.M., Dudin A.N., Semenova O.V., Klimenok V.I. Performance Analysis of the $BMAP/G/1$ Queue with Gated Servicing and Adaptive Vacations // Performance Evaluat. - 2011. - V. 68, № 5. - P. 446-462.

45. Larionov A., Ivanov R., Vishnevsky V. UHF RFID in Automatic Vehicle Identification: Analysis and Simulation // IEEE Journal of Radio Frequency Identification. - 2017. - Volume1, Issue: 1. - P. 3-12.

46. V.M.Vishnevsky, A.N.Andronov. Estimating the Throughput of Wireless Hybrid Systems Operating in a Semi-Markov Stochastic Environment // Automation and Remote Control. - 2017. - Vol. 78, № 12. - P. 2154-2165.

47. Dudin A., Klimenok V., Samouylov K. Stationary Distribution of Waiting Time in $MAP/G/1/N$ Queueing System with LIFO Service Discipline // Lecture Notes in Computer Science. - 2017. - V. 10372. - P. 50–61.

48. Dudin A., Dudin S., Dudina O., Kim C.S. Analysis of a wireless sensor node with varying rates of energy harvesting and consumption // Lecture Notes in Computer Science. - 2017. - V. 10684. - P. 172-182.

49. Klimenok V.I., Dudina O.S., Vishnevsky V.M., Samouylov K.E. Retrial tandem queue with BMAP-input and semi-Markovian service process // Communications in Computer and Information Science. - 2017.- V. 700.- P. 159-173.

50. V.Vishnevsky, K.Samouylov, V.Naumov, A.Krishnamoorty, N.Yarkina. Multiservice Queueing System with MAP Arrivals for Modelling LTE Cell with H2H and M2M Communications and M2M Aggregation // Communications in Computer and Information Science. Springer International Publishing. - 2017.- Vol. 700.- P. 63-74.

51. Kim C.S., Klimenok V.I., Dudin A.N. Optimization of Guard Channel Policy in Cellular Mobile Networks with Account of Retrials // Comput. Oper. Res. - 2014. - V. 43. - P. 181-190.

52. Lee M.H., Dudin A., Klimenok V. Mathematical analysis of the multi-server queueing model for dynamic channel reservation in wireless networks // IEEE Commun. Lett. - 2006. - V. 10. - P. 855-857.

53. Do T.V. Solution for a retrial queueing problem in cellular networks with the Fractional Guard Channel policy // Math. Comput. Modell. - 2011. - V. 53. - P. 2059-2066.

54. Zhou Z., Zhu Y. Optimization of the (MAP1,MAP2)/(PH1, PH2)/N retrial queue model of wireless cellular networks with channel allocation // Comput. Electric. Engineer. - 2013. - V. 39. - P. 1637-1649.

55. Kim C.S., Dudin A., Dudin S., Dudina O. Analysis and optimization of Guard Channel Policy with buffering in cellular mobile networks // Comput. Networks. - 2016. - V. 107. - P. 258-269.

56. Vishnevsky V., Larionov A., Semenova O., Ivanov R. State Reduction in Analysis of a Tandem Queueing System with Correlated Arrivals // Communications in Computer and Information Science. Springer International Publishing. - 2017.- Vol. 800. - P. 215-230.

57. Dudin A., Dudin S., Dudina O., Kim C.S. Analysis of a wireless sensor node with varying rates of energy harvesting and consumption // Lecture Notes in Computer Science. - 2017. - V. 10684. - P. 172-182.

58. Vishnevsky V.M., Larionov A.A., Ivanov R.E. An Open Queueing Network with a Correlated Input Arrival Process for Broadband Wireless Network Performance Evaluation // Inform. Technol. Math. Modell. - Queueing Theory Appli. - 2016. - CCIS. V. 638. - P. 354-365.

59. Vishnevsky V.M., Dudin A.N., Kozyrev D.V., Larionov A.A. Methods of Performance Evaluation of Broadband Wireless Networks Along the Long Transport Routes // Commun. Comput. Inform. Sci. - 2016. - V. 601. - P. 72-85.

60. Vishnevsky V.M., Krishnamoorthy A., Kozyrev D.V., Larionov A.A. Review of methodology and design of broadband wireless networks with linear topology // Indian J. Pure Appl. Math. - 2016. - V. 47, № 2. - P. 329-342.

61. Vishnevsky V.M., Larionov A.A., Smolnikov R.V. Optimization of Topological Structure of Broadband Wireless Networks Aiong the Long Traffic Routes // Distributed Comput. Commun. Networks. - 2016. - V. 601. - p. 30-39.

62. Lakatos, Laszlo, Szeidl, Laszlo, Telek, Miklos // Introduction to Queueing Systems with Telecommunication Applications. Springer Science+Business Media, LCC, 2013.

63. Boucherie, Richard J., van Dijk Nico M. Queueing Networks. // Springer Science+Business Media, LLC, 2011.

64. Artalejo, Jesus, Gomez-Corral, Antonio Retrial Queueing Systems - A Computational Approach // Springer-Verlag Berlin Heidelberg, 2008.

65. Lagershausen, Svenja Performance Analysis of Closed Queueing Networks // Springer-Verlag Berlin Heidelberg, 2013.

66. Alfa, Attahiru Sule Queueing Theory for Telecommunications // Springer Science+Business Media, LLC, 2010.

67. Степанов С.Н. Теория телетрафика. Концепции, модели, приложения. -М.: Горячая линия - Телеком. 2015. - С. 860.

68. Беллман Р. Введение в теорию матриц. Наука, Москва, 1969.

69. Бочаров П.П., Печинкин А.В. Теория массового обслуживания. - М.: Изд-во Рос. ун-та дружбы народов, 1995.

70. Бронштейн О.И., Духовный И.М. Модели приоритетного обслуживания в информационно - вычислительных сетях. - М.: Наука, 1976.

71. Вишневский В.М., Семенова О.В. Системы поллинга: теория и применение в широкополосных беспроводных сетях. - М.: Техносфера, 2007.

72. Гантмахер Ф.Р. Теория матриц. Наука, Москва, 1966.

73. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания. - М.: Наука, 1987.

74. Гнеденко Б.В. и др. Приоритетные системы обслуживания. - М.:

75. Джейсуол Н. Очереди с приоритетами. - М.: Мир, 1973.

76. Дудин А.Н., Клименок В.И. Системы массового обслуживания с коррелированными потоками. - Мн.: Изд-во Белорус. ун-та, 2000.

77. Дудин А.Н., Медведев Г.А., Меленец Ю.В. Практикум на ЭВМ по теории массового обслуживания. Минск: Университетское. 2000.

78. Дудин А.Н., Клименок В. И., Царенков Г.В. Расчет характеристик однолинейной системы массового обслуживания с групповым марковским потоком, полумарковским обслуживанием и конечным буфером // Автоматика и телемеханика. - 2002. - № 8. - С. 87-101.

79. Дудин С., Дудина О. Двухфазная система $MAP/PH/N/K-N \to \bullet/PH/R/\infty$ как модель функционирования центра информационной и технической поддержки // Проблемы передачи информации. - 2013. - V. 49. - № 1. - С. 58-72.

80. Кемени Д.Дж., Снелл Дж.Л. Конечные цепи Маркова. Москва. Наука. 1970.

81. Клейнрок Л. Теория массового обслуживания. М.: Машиностроение, 1979. -432 с.

82. Клейнрок Л. Вычислительные системы с очередями. М.: Мир, 1979. -598 с.

83. Клименок В.И., Тарамин О.С. Двухфазные системы обслуживания с коррелированными потоками.- Минск.: РИВШ, 2011.- 142 с.

84. Клименок В.И., Тарамин О.С. Двухфазная система обслуживания с групповым марковским потоком и повторными вызовами // Автоматика и телемеханика. - 2010. - Т. 71, № 1. - С. 3-17.

85. Клименок В.И., Брейер Л., Царенков Г.В., Дудин А.Н. Двухфазная система $BMAP/G/1/N \to PH/1/M-1$ с блокировкой // Автоматика и телемеханика. - 2004. - № 1. - С. 104-115.

86. Климов Г. П. Стохастические системы обслуживания. - М.: Наука, 1966. - 244 с.

87. Матвеев В. Ф, Ушаков В. Г. Системы массового обслуживания. - М.: МГУ, 1984. - 240 с.

88. Феллер В. Введение в теорию вероятностей и ее приложения. Т.1.- М.: Мир, 1984.

89. Хинчин А.Я. Работы по математической теории массового обслуживания. - М.: Физматгиз, 1963.

90. Яшков С.Ф. Анализ очередей в ЭВМ. - М.: Радио и связь, 1989.

91. Asmussen S. Applied probability and queues. – Springer Science & Business Media, 2008. - T. 51.

92. Alfa A.S. On approximating higher order MAPs with MAPs of order two / A.S. Alfa, J.E. Diamond // Queueing Systems. – 2000. – V. 34. – P. 269-288.

93. Alfa A.S., Li W. PCS networks with correlated arrival process and retrial phenomenon // IEEE Transactions on Wireless Communications. - 2002. - V. 1. - P. 630-637.

94. Aman M., Mahfooz S., Rehman W.U. A Handoff using Guard Channels Scheme (HGCS) for Cognitive Radio Networks // Global Journal of Computer Science and Technology. - 2011. - V. 11, № 20.

95. Arnon S., Barry J., Karagiannidis G., Schober R., Uysal M. (Eds): Advanced Optical Wireless Communication Systems. Cambridge University Press, 2012. - 404 p.

96. Artalejo J.R. Accessible bibliography on retrial queues // Mathematical and Computer Modelling. - 1999. - V. 30. - P. 1-6.

97. Artalejo J.R. A classified bibliography of research on retrial queues: Progress in 1990 - 1999 // Top. - 1999. - V. 7. - P. 187-211.

98. Artalejo J.R. G - networks: a versatile approach for work removal in queueing networks // European Journal of Operational Research. - 2000. - V. 126. - P. 233-249.

99. Balsamo S. A review on queueing network models with finite capacity queues for software architectures performance prediction / Balsamo S., Persone V.D.N., Inverardi P. // Performance Evaluation. - 2003. – V. 51. – P. 269-288.

100. Bocharov P. P., Manzo R., Pechinkin A. V. Analysis of a two-phase queueing system with a Markov arrival process and losses // Journal of Mathematical Sciences. - 2005. - V. 131.- № 3. - P. 5606-5613.

101. Bocharov P. P., Manzo R., Pechinkin A. V. Analysis of a two-phase queueing system with a Markov arrival process and blocking // Journal of Mathematical Sciences. - 2006. - V. 132. № 5. - P. 578-589.

102. Boxma O.J., Groenendijk W.P. Pseudo conservation laws in cyclic-service systems // Journal of Applied Probability. – 1987. - V. 24. – P. 949-964.

103. Bromberg M.A. Multi-phase systems with losses with exponential servicing / Bromberg M.A. // Automation and Remote Control. – 1979. – V. 40. – P. 27-31.

104. Bromberg M.A. Service by a cascaded network of instruments / Bromberg M.A., Kokotushkin V.A., Naumov V.A. // Automation and Remote Control. – 1977. – V. 38. – P. 60-64.

105. Burke P.J. The output of a queueing system // Operations Research.- 1956. - V. 4. - P. 699-704.

106. Chakravarthy S. The batch markovian arrival process: a review and future work // Advances in probability theory and stochastic processes. - 2001. - P. 21-39.

107. Choi B.D., Chang Y. $MAP_1, MAP_2/M/c$ retrial queue with the retrial group of finite capacity and geometric loss // Mathematical and Computer Modelling. - 1999. - V. 30. - P. 99-114.

108. Choi B.D., Melikov A., Velibekov A. A simple numerical approximation of joint probabilities of calls in service and calls in the retrial group in a picocell // Applied Computational Mathematics. - 2008. - V. 7. P. 21-30.

109. D. Heyman, D. Lucantoni. Modelling multiple IP traffic streams with rate limits // IEEE ACM Transactions on Networking. - 2003. - V.11. - P. 948-958.

110. A. Klemm, C.Lindermann, M. Lohmann. Modelling IP traffic using the batch Marcovian arrival process // Performance Evaluation. - 2008. - V.54. - P.149-173.

111. Vishnevsky V., Larionov A., Ivanov R., Semenova O. Estimation of IEEE 802.11 DCF access Performance in Wireless Networks with Topology using PH Service Time Approximations and MAP Input. Proceedings 11th IEEE International Conference Application of Information and Communication Technologies - AIST2017, 2017. - Vol. 2 . - P.85-89.

112. Cinlar E. Introduction to stochastic processes. - New Jersej: Prentice-Hall, 1975.

113. Daigle J.N. Queueing theory for telecommunications. - Addison-Wesley Publishing Company, Inc., 1992. Изд-во МГУ, 1973.

114. Do T.V. Solution for a retrial queueing problem in cellular networks with the Fractional Guard Channel policy // Mathematical and Computer Modelling. - 2011. - V. 53. - P. 2059-2066.

115. Dudin, AN, VI Klimenok, IA Klimenok, et al. (2000). Software "SIRIUS+" for evaluation and optimization of queues with the $BMAP$-input. In *Advances in Matrix Analytic Methods for Stochastic Models*, G Latouche and P Taylor (eds.), 115–133. New Jersey: Notable Publications, Inc. - P. 115–133.

116. Dudin A.N., Shaban A.A., Klimenok V.I. Analysis of a queue in the $BMAP/G/1/N$ system // International Journal of Simulation: Systems, Science and Technology. - 2005. - V. 6. № 1-2. - P. 13-23.

117. Dudin A.N., Shaban A.A. Analysis of the $BMAP/SM/1/N$ type system with randomized choice of customers admission discipline // Communications in Computer and Information Science. - 2016. - V. CCIS 638.- P. 44-56.

118. Dudin A., Dudin S., Dudina O. Tandem queueing system $MAP/M/N/K - N \rightarrow \bullet M/R/\infty$ with impatient customers as a model of remote technical support // Proceedings of the 2nd Baltic Congress on Future Internet Communications. - 2012. - № BCFIC 2012. 1569570915. - P. 1-6.

119. O. Dudina, C.S. Kim, S. Dudin, Retrial Queueing System with Markovian Arrival Flow and Phase Type Service Time Distribution, Computers and Industrial Engineering, 66 (2013) 360-373.

120. Falin G.I., Templeton J.G.C. Retrial queues. - London: Chapman and Hall, 1997.

121. Ferng H.W., Chang J.F. Departure processes of $BMAP/G/1$ queues // Queueing Systems. - 2001. - V. 39. - P. 109-135.

122. Ferng H.W., Chang J.F. Connection-wise end-to-end performance analysis of queueing networks with $MMPP$ inputs // Performance Evaluation. - 2001. - V.43. - P. 39-62.

123. Gail H.R., Hantler S.L., Taylor B.A. Spectral analysis of $M|G|1$ and $G|M|1$ type Markov chains // Advances in Applied Probability. - 1996. - V. 28. - P. 114-165.

124. Gelenbe E. Product form networks with negative and positive customers // Journal of Applied Probability. - 1991. - V. 28. - P. 655-663.

125. Gnedenko B.V., Koenig D. Handbuch der bedienungstheorie. - Berlin: Akademie - Verlag, 1983.

126. Gomez-Corral A. A tandem queue with blocking and Markovian arrival process // Queueing Systems. - 2002. - V. 41. - P. 343-370.

127. Gomez-Corral A. On a tandem G-network with blocking // Advances in Applied Probability. - 2002. - V.34, № 3. - P. 626-661.

128. Graham A. Kronecker products and matrix calculus with applications. Cichester: Ellis Horwood, 1981.

129. Grassmann W.K., Stanford D.A. Matrix analytic methods // Computational Probability. - Boston: Kluwer Academic, 2000. - P. 153-203.

130. Guerin R. Queueing-blocking system with two arrival streams and guard channels // IEEE Transactions on Communications. - 1998. - V. 36. - P.153-163.

131. He Q.M. Queues with marked calls. Advances in Applied Probability. - 1996. - V. 28. - P. 567-587.

132. Heindl A. Decomposition of general tandem networks with $MMPP$ input // Performance Evaluation. - 2001. - V. 44. - P. 5-23.

133. Heindl A. Output models of $MAP/PH/1(/K)$ queues for an efficient network decomposition / Heindl A., Telek M. // Performance Evaluation. – 2002. – V. 49. – P. 321-339.

134. Heindl A. Decomposition of general queue networks with $MMPP$ inputs and customer losses // Performance Evaluation. - 2003. - V. 51. - P. 117-136.

135. Heindl A. Correlation bounds for second order MAPs with application to queueing network decomposition / Heindl A., Mitchell K., van de Liefvoort A. // Performance Evaluation. – 2006. – V. 63. – P. 553-577.

136. Heyman D.P. Modelling multiple IP traffic streams with rate limits / Heyman D.P., Lucantoni D. // IEEE/ACM Transactions on Networking.. – 2003. – V. 11. – P. 948-958.

137. Kemeni J. G., Snell J. L., Knapp A. W. Denumerable Markov Chains. Van Nostrand, New York, 1966.

138. Kim C.S., Park S.H., Dudin A., Klimenok V., Tsarenkov G. Investigaton of the $BMAP/G/1 \to \bullet/PH/1/M$ tandem queue with retrials and losses // Applied Mathematical Modelling. - 2010. - V.34, № 10. - P. 2926-2940.

139. Kim C.S., Klimenok V.I., Taramin O.S., Dudin A. A tandem $BMAP/G/1 \to \bullet/M/N/0$ queue with heterogeneous customers // Mathematical Problems in Engineering. - 2012. - V. 2012, Article ID 324604. - 26. P.

140. Kim C.S., Klimenok V., Taramin O. A tandem retrial queueing system with two Markovian flows and reservation of channels // Computers and Operations Research. - 2010. - V. 37, № 7. - P. 1238-1246.

141. Kim Ch., Dudin A., Klimenok V. Tandem retrial queueing system with correlated arrival flow and operation of the second station described by a Markov chain // Communications in Computer and Information Science. - 2012. - V. 291. - P. 370-382.

142. Kim C.S., Dudin A., Dudin S., Dudina O. Tandem queueing system with impatient customers as a model of call center with Interactive Voice Response // Performance Evaluation. - 2013. - V. 70. - P. 440-453.

143. Kesten H., J.Th. Runnenburg J.Th. Priority in waiting line problems, CWI, Amsterdam, 1956. - 25 p.

144. Kim C.S., Dudin A., Dudina O., Dudin S. Tandem queueing system with infinite and finite intermediate buffers and generalized phase-type service time distribution // European Journal of Operational Research. - 2014. - V. 235. - P. 170-179.

145. Kim C.S., Dudin S., Taramin O., Baek J. Queueing system $MMAP/PH/N/N+R$ with impatient heterogeneous customers as a model of call center // Applied Mathematical Modelling. - 2013. - V. 37, № 3. - P. 958-976.

146. Kim C.S., Dudin A., Dudin S., Dudina O. Analysis and optimization of Guard Channel Policy with buffering in cellular mobile networks // Computer Networks. - 2016. - V. 107. - P. 258-269.

147. Kim C.S., Klimenok V., Dudin A. Optimization of Guard Channel Policy in Cellular Mobile Networks with Account of Retrials // Computers and Operation Research. - 2014. - V. 43. - P. 181-190.

148. Kim C.S., Klimenok V., Dudin A. Priority tandem queueing system with retrials and reservation of channels as a model of call center // Computers and Industrial Engineering. - 2016. - V. 96. - P. 61-71.

149. Klemm A. Modelling IP traffic using the batch Markovian arrival process / Klemm A., Lindermann C., Lohmann M. // Performance Evaluation. – 2003. – V. 54. – P. 149-173.

150. Klimenok, V.I., Dudin A.N.: Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queueing Systems, 54. - 2006. - P. 245–259.

151. Klimenok V., Dudin A., Vishnevsky V. On the stationary distribution of tandem queue consisting of a finite number of stations // Communications in Computer and Information Science. - 2012, - V.291. - P. 383-392.

152. Klimenok V.I., Kim C.S., Orlovsky D.S., Dudin A.N. Lack of invariant property of Erlang $BMAP/PH/N/0$ model // Queueing Systems. - 2005. - V.49. - P. 187-213.

153. Klimenok V.I., Breuer L., Tsarenkov G.V., Dudin A.N. The $BMAP/G/1/N \to PH/1/M-1$ tandem queue with losses // Performance Evaluation. - 2005. - V. 61.- P. 17-60.

154. Klimenok V., Kim C.S., Tsarenkov G.V., Breuer L., Dudin A.N. The $BMAP/G/1 \rightarrow \cdot/PH/1/M$ tandem queue with feedback and losses // Performance Evaluation. - 2007.- V. 64. - P. 802-818.

155. Kulkarni V.G., Liang H.M. Retrial queues revisited // Frontiers in Queueing. - Boca Raton: CRC Press, 1997. - P. 19-34.

156. Kuczma M. Functional equations in a single variable // PWN - Polish Scientific Publishers. Warsaw. – 1968.

157. Lee M.H., Dudin A., Klimenok V. Mathematical analysis of the multi-server queueing model for dynamic channel reservation in wireless networks // IEEE Communications Letters. - 2006. - V. 10. - P. 855-857.

158. Lian Z., Liu L. A tandem network with MAP inputs // Operations Research Letters. - 2008. - V. 36. - P. 189 – 195.

159. Lucantoni D.M. New results on the single server queue with a batch markovian arrival process // Communications in Statistics - Stochastic Models. - 1991. - V. 7. - P. 1-46.

160. Mandjes M., Tutschku K. Efficient call handling procedures in cellular mobile networks. Forschungsbericht // Preprint-Reihe Nr 14X, Universitaet Wuerzburg, Institut fuer Informatik, 1996.

161. Mitrani I. The spectral expansion solution method for Markov processes on lattice strips. In "Advances in Queueing: Theory, Methods and Open Problems CRC Press. Boca Raton. - 1995. - Chapter 13. - P. 337-352.

162. Moustafa M.D. Input – output Markov process, Proc. Koninkl. Net. Akad. Wetensch, 1957. - Vol. A60. - p. 112–118.

163. Neuts M.F. Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach. Johns Hopkins University Press. 1981.

164. Pattavina A., Parini A. Modelling voice call inter-arrival and holding time distributions in mobile networks, in: Performance Challenges for Efficient Next Generation Networks - Proc.of 19th International Teletraffic Congress, Aug.-Sept. - 2005. - P. 729-738.

165. Perros H.G. A bibliography of papers on queueing networks with finite capacity queues / Perros H.G. // Performance Evaluation. – 1989. – V. 10. – P. 255-260.

166. Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. Numerical recipes. - Cambridge: Cambridge University Press, 1986.

167. Ramaswami V. Independent Markov process in parallel // Communications in Statistics-Stochastic Models. - 1985. - V. 1. - P. 419-432.

168. Schriber T.J. Simulation using GPSS // John Wiley & Sons. – 1974

169. Skorokhod, A. Probability theory and random processes / Skorokhod A. // Kiev: High School, 1980.

170. Sennot L.I., Humblet P.A., Tweedie R.L. Mean drifts and non-ergodicity of Markov chains, Operations Research. - 1983. - Vol. 31. - P. 783–789.

171. Sharov S.Yu., Semenova O.V. Simulation model of wireless channel based on FSO and RF technologies. Distributed Computer and Communication Networks. Theory and Applications (DCCN-2010). - 2010. - P. 368-376.

172. Shioda S. Departure process of the $MAP/SM/1$ queue // Queueing Systems. - 2003. - V. 44. - P. 31-50.

173. Sumita S. Performance analysis of interprocessor communications in an electronic switching system with distributed control // Performance Evaluation. – 1988. – V. 9. – P. 83-91.

174. Syski R. A personal view of queueing theory. - В кн.: Frontiers in Queueing.- Boca Raton - New York - London - Tokyo: CRC, 1997. - P. 3-18.

175. Tijms H. Stochastic Models: An Algorithmic Approach. Chichester: John Wiley and Sons, 1994.

176. Tian N., Zhang Z.G. Vacation Queueing Models - Theory and Applications. Springer: Heidelberg, 2006. - 385 p.

177. Tran-Gia, P., Mandjes M. Modeling of customer retrial phenomenon in cellular mobile networks // IEEE Journal on Selected Areas in Communications. - 1997. - V.15. - P. 1406 – 1414.

178. van Dantzig D. Chaines de Markof dans les ensembles abstraits et applications aux processus avec regions absorbantes et au probleme des boucles, Ann. de l'Inst. H. Pioncare, 14. - 1955. - P. 145-199.

179. Vishnevskii V.M., , Semenova O.V. Mathematical methods to study the polling systems // Automation and Remote Control. – 2006. – V. 67. – P. 173-220.

180. Vishnevsky V., Dudin A., Klimenok V., Semenova O., Shpilev S. Approximate Method to Study $M/G/1$-Type Polling System with Adaptive Polling Mechanism // Quality Technology and Quantitative Management. - 2012. - V. 9. № 2. - P. 211-228.

181. A. Douik, H. Dahrouj, T. Y. Al-Naffouri, и M. S. Alouini. Hybrid Radio/Free-Space Optical Design for Next Generation Backhaul Systems // IEEE Transactions on Communications. : IEEE. - 2016. - P. 2563–2577.

182. M. A. Esmail, H. Fathallah, и M. S. Alouini. Outdoor FSO Communications under Fog: Attenuation Modeling and Performance Evaluation // IEEE Photonics J. - 2016. - Т. 8. № 4. - P. 1–22.

183. Y. Wu, Q. Yang, D. Park, и K. S. Kwak. Dynamic Link Selection and Power Allocation with Reliability Guarantees for Hybrid FSO/RF Systems // IEEE Access. - 2017. - Т. 5. - P. 13654–13664.

184. K. Zhou, C. Gong, N. Wu, и Z. Xu. Distributed Channel Allocation and Rate Control for Hybrid FSO/RF Vehicular Ad Hoc Networks // Journal of Optical Communications and Networking. - 2017. - Т. 9. № 8. - P. 669–681.

185. B. Makki, T. Svensson, T. Eriksson, и M. S. Alouini. On the Performance of RF-FSO Links with and Without Hybrid ARQ // IEEE Transactions on Wireless Communications. : IEEE. - 2016. - P. 4928–4943.

186. Y. Zhang, Y.-J. Chu, и T. Nguyen. Coverage Algorithms for WiFO: A Hybrid FSO-WiFi Femtocell Communication System // 2017 26th International Conference on Computer Communication and Networks (ICCCN). : IEEE. - 2017. - P. 1–6.

187. M. Najafi, V. Jamali, D. W. K. Ng, и R. Schober. C-RAN with Hybrid RF/FSO Fronthaul Links: Joint Optimization of RF Time Allocation and Fronthaul Compression // GLOBECOM 2017 - 2017 IEEE Global Communications Conference. : IEEE/ - 2017. - С. 1–7.

188. M. Atakora и H. Chenji. Optimal multicasting in hybrid RF/FSO DTNs // 2016 IEEE Global Communications Conference, GLOBECOM 2016 - Proceedings. : IEEE. - 2016. - С. 1–6.

189. Advanced Optical Wireless Communication Systems. Cambridge University Press, 2012. - 404 P.

190. Vishnevsky, V.M., Semenova, O.V., Sharov, S.Yu. Modeling and Analysis of a Hybrid Communication Channel Based on Free-space Optical and Radio-frequency Technologies. Automation and Remote Control. - 2013. - V. 74. - P. 521-528.

191. Vishnevsky V.M., Kozyrev D.V., Semenova O.V. Redundant queuing system with unreliable servers // Proceedings of the 6th International Congress on Ultra Modern. Telecommunications and Control Systems. IEEE Xplore Digital Library, 2014. - P. 383-386.

192. Vishnevsky V.M., Klimenok V.I. Unreliable Queueing System with Cold Redundancy // Communication in Computer and Information Science. - 2015. - V. 522. - P. 336-347.

193. Dudin, A., Klimenok, V., Vishnevsky V. Analysis of Unreliable Single Server Queuing System with Hot Back-up Server // Communication in Computer and Information Science. - 2015. - V. 499. - P. 149-161.

194. Vishnevsky V., Kozyrev D., Semenova O.V.: Redundant queueing system with unreliable servers. Proceedings of the 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). Moscow, 2016.- P. 383-386.

195. Vishnevsky V.M., Semenova O.V., Sharov S.Yu.: Modeling and analysis of a hybrid communication channel based on free-space optical and radio-frequency Technologies // Automation and Remote Control 72. - 2013. - P. 345-352.

196. Vishnevsky V.M., Dudin A.N., Semenova O.V., Klimenok V.I. Performance Analysis of the $BMAP/G/1$ Queue with Gated Servicing and Adaptive Vacations // Performance Evaluation. - 2011. - V. 68. № 5. - P. 446-462.

197. Yang S.F., Wu J.S., Huang J.W. Spectrum Handover with Queues and Guard Channels in Cognitive Radio Networks // 2012 Spring World Congress on Engineering and Technology. 1-4 DOI: 10.1109/SCET.2012.6342146.

198. Zhou Z., Zhu Y. Optimization of the $(MAP_1, MAP_2)/(PH_1, PH_2)/N$ retrial queue model of wireless cellular networks with channel allocation // Computers and Electrical Engineering. - 2013. - V. 39. - P. 1637-1649.

199. Mack C. The efficiency of N machines uni-directionally patrolled by one operative when walking time is constant and repair times are variable // Journal of the Royal Statistical Society Series B. - 1957. - Vol. 19, №. 1. - P. 173-178.

200. Mack C., Murphy T., Webb N.L. The efficiency of N machines uni-directionally patrolled by one operative when walking time and repair times are constants // Journal of the Royal Statistical Society Series B. - 1957. - Vol. 19, №. 1. - P. 166-172.

201. Abidini M.A., Boxma O., Resing J. Analysis and optimization of vacation and polling models with retrials // Performance Evaluation. - 2016. - Vol. 98. - P. 52-69.

202. Adan I.J.B.F., Boxma O.J., Kapodistria V.G. The shorter queue polling model // Annals of operations Research. - 2016. - Vol. 241, №. 1. - P.167-200.

203. Avrachenkov K., Perel E., Yechiali U. Finite-buffer polling system with threshold-based switching policy // TOP. - 2016. - Vol. 24, №. 3. - P. 541-571.

204. Boon M.A.A., van der Mei R.D., Winands E.M.M. Heavy traffic analysis of roving server networks // Stochastic Models. - 2016. - P. 1-39.

205. Jiang T., Liu L., Zhu Y. Analysis of a batch service polling system in a multi-phase random environment // Methodology and Computing in Applied Probability. - 2017. - P. 1-20.

206. Liu Z., Chu Y., Wu J. On the three-queue priority polling system with threshold service policy // Journal of Applied Mathematics and Computing. - 2017. - Vol. 53, №. 1. - P. 445-470.

207. Perel E., Yechiali U. Two-queue polling systems with switching police based on the queue that is not being served // Stochastic Models. May, 2017.

208. Вишневский В.М., Семенова О.В. Системы адаптивного динамического поллинга с коррелированными входными потоками // ИПУ РАН. - 2017. - 89 с.

209. Boon M.A.A., van der Mei R.D., Winands E.M.M. Applications of polling system // Surveys in Operations Research and Management Science. - 2011. - Vol. 16, №. 2. - P. 67-81.

210. Vishnevsky V., Semenova O. Polling System: Theory and applications for Broadband Wireless Network // Academic Piblishing. - 2012. - P. 317.

211. Вишневский В.М., Мишкой Г.К., Семенова О.В. Новые модели и методы исследования систем поллинга // Proceedings of the International Conference Distributed Computer and Communication Network. Theory and applications (DCCN-2009, Moscow). - 2009. - P. 79-85.

212. Borst S.C. Polling system. Amsterdam: Stichting Mathematisch Centrum, 1996.

213. Vishnevsky V.M., Semenova O.V. Adaptive dynamical polling in wireless network // Cybernetics and Information Technologies. - 2008. - V. 8, №. 1. - P. 3-11.

214. Tagaki H. Analysis of polling systems. MIT Press, 1986.

215. Takagi H. Queueing Analysis. Amsterdam: North Holland- V. 1.- 1991. V. 2.- 1993. V. 3.- 1993.- P. 1503.