Монография посвящена разработке и применению метода потенциалов для определения характеристик одноканальных систем обслуживания с пороговыми и гистерезисными стратегиями функционирования. Рассматриваются системы с групповым поступлением заявок, пуассоновским стационарным входящим потоком, общими распределениями времени обслуживания, ограниченной и неограниченной очередью. Различные режимы функционирования могут отличаться распределением времени обслуживания и интенсивностью входящего потока. Переключение режимов осуществляется в моменты начала обслуживания заявок. Сущность метода потенциалов состоит в использовании стольких базовых случайных блужданий, сколько применяется различных режимов функционирования. Каждое блуждание определяется функцией распределения времени обслуживания и параметрами входящего потока соответствующего режима. Полученные результаты проверены с помощью имитационных моделей, построенных с использованием инструментальных средств GPSS World. Книга предназначена для научных сотрудников, аспирантов и студентов, занимающихся исследованием систем массового обслуживания.

Жерновый Юрий, кандидат физ.-мат. наук, доцент Львовского национального университета имени Ивана Франко. Автор более 85 научных работ, 2-х монографий и 3-х учебных пособий. Жерновый Константин, кандидат физ.-мат. наук, научный сотрудник Львовского национального университета имени Ивана Франко, Львов, Украина. Автор 25 научных работ.

978-3-659-33412-2

Метод потенциалов

Жерновый, Жерновый

Юрий Жерновый
Константин Жерновый

Метод потенциалов для пороговых стратегий обслуживания

Определение характеристик одноканальных систем

LAP LAMBERT Academic Publishing

Юрий Жерновый
Константин Жерновый

Метод потенциалов для пороговых стратегий обслуживания

Юрий Жерновый
Константин Жерновый

Метод потенциалов для пороговых стратегий обслуживания

Определение характеристик одноканальных систем

LAP LAMBERT Academic Publishing

Impressum / Выходные данные
Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
Alle in diesem Buch genannten Marken und Produktnamen unterliegen warenzeichen-, marken- oder patentrechtlichem Schutz bzw. sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Inhaber. Die Wiedergabe von Marken, Produktnamen, Gebrauchsnamen, Handelsnamen, Warenbezeichnungen u.s.w. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Библиографическая информация, изданная Немецкой Национальной Библиотекой. Немецкая Национальная Библиотека включает данную публикацию в Немецкий Книжный Каталог; с подробными библиографическими данными можно ознакомиться в Интернете по адресу http://dnb.d-nb.de.
Любые названия марок и брендов, упомянутые в этой книге, принадлежат торговой марке, бренду или запатентованы и являются брендами соответствующих правообладателей. Использование названий брендов, названий товаров, торговых марок, описаний товаров, общих имён, и т.д. даже без точного упоминания в этой работе не является основанием того, что данные названия можно считать незарегистрированными под каким-либо брендом и не защищены законом о брендах и их можно использовать всем без ограничений.

Coverbild / Изображение на обложке предоставлено: www.ingimage.com

Verlag / Издатель:
LAP LAMBERT Academic Publishing
ist ein Imprint der / является торговой маркой
OmniScriptum GmbH & Co. KG
Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Германия
Email / электронная почта: info@lap-publishing.com

Herstellung: siehe letzte Seite /
Напечатано: см. последнюю страницу
ISBN: 978-3-659-33412-2

Содержание

Список основных обозначений

$\mathrm{M}^{\theta}/\mathrm{G}/1/\mathrm{m}$ – система типа $\mathrm{M}/\mathrm{G}/1/\mathrm{m}$ с групповым поступлением заявок.

$\mathrm{M}^{\theta}/\mathrm{G}/1$ – система типа $\mathrm{M}^{\theta}/\mathrm{G}/1/\mathrm{m}$ в случае $m=\infty$.

m – максимальная длина очереди.

$\xi(t)$ – число заявок в системе в момент времени t.

$\eta(x,\lambda)$ – число заявок, прибывших в систему на промежутке времени $[0; x)$, при условии, что интенсивность простейшего потока групп заявок равна λ.

$\mathbf{E}$ ($\mathbf{P}$) – условное математическое ожидание (условная вероятность) при условии, что система начинает работать в момент прибытия первой группы заявок.

$\mathbf{P}_n(A)=\mathbf{P}\{A\,|\,\xi(0)=n\}$ – условная вероятность события A при условии, что в начальный момент времени в системе находится $n\geq 0$ заявок.

$a_k=\mathbf{P}\{\theta_n=k\}$, где θ_n – число заявок в n-ой группе.

$F(x)$ – функция распределения вероятностей времени обслуживания в основном режиме работы системы.

$f(s)=\int\limits_0^{\infty} e^{-sx}dF(x);$

$\tilde{f}(s)=\int\limits_0^{\infty} e^{-sx}d\tilde{F}(x);$

$M=\int\limits_0^{\infty} x\,dF(x);$

$\tilde{M}=\int\limits_0^{\infty} x\,d\tilde{F}(x);$

$\overline{F}(x)=1-F(x);$

$\overline{a}_n=\sum\limits_{k=n}^{\infty} a_k;$

$b=\sum\limits_{k=1}^{\infty} ka_k;$

$\rho=\lambda Mb;$

$\tilde{\rho}=\lambda\tilde{M}b.$

$p_k(m) = \lim_{t \to \infty} \mathbf{P}\{\xi(t) = k\}$ – стационарное распределение числа заявок в системе.

$\mathbf{P}_{\mathrm{sv}}(m)$ – стационарное значение вероятности обслуживания.

$\mathbf{E}\,\tau(m)$ – средняя продолжительность периода занятости.

$\mathbf{E}Q(m)$ – стационарное значение средней длины очереди.

$\mathbf{E}w(m)$ – стационарное значение среднего времени ожидания в очереди.

ВВЕДЕНИЕ

В последние десятилетия происходит бурное развитие информационных технологий. Появляются новые, разнообразные вычислительные средства и информационно-вычислительные сети. Общей особенностью информационных процессов, протекающих в этих сетях, является их стохастическая природа. Поэтому, как на этапе проектирования и создания сетевых систем, так и в процессе их функционирования возникает необходимость решения ряда математических проблем, связанных с оценкой производительности сетей и их компонентов. Для решения поставленных теоретических задач исследователи часто используют методы теории массового обслуживания.

Системы обслуживания, которые могут работать в нескольких режимах, отличающихся друг от друга интенсивностями входящего потока и обслуживания, являются адекватной математической моделью процессов передачи информации в информационно-вычислительных сетях. Среди них, например, процесс передачи информации в каналах цифровых сетей интегрального обслуживания (ISDN) в режиме гибридной коммутации, процесс передачи информации с использованием протоколов типа скользящего окна, leaky bucket [14], процессы передачи данных в сетях ATM с использованием технологий мультиплексирования [64]. Пороговое управление входящим потоком и временем обслуживания используется с целью предотвращения перегрузок в таких системах.

В простейшем случае суть пороговой стратегии заключается в следующем: если длина очереди превышает пороговое значение, то интенсивность входящего потока может быть снижена (или сведена к нулю в случае блокировки), а интенсивность обслуживания должна быть увеличена. С помощью систем обслуживания с пороговыми стратегиями функционирования, в частности, проанализирован ряд схем управления процессами передачи информации в сетях ATM [42–44, 49, 56, 63–66].

Целью применения блокировки входящего потока и режимов с разными интенсивностями обслуживания является уменьшение числа потерянных заявок и времени ожидания заявки на обслуживание. Система такого типа может играть роль основного канала в двухканальной системе обслуживания со вторым резервным каналом. В случае чрезмерной загрузки основного канала посылается соответствующий сигнал и новые заявки направляются в резервный канал, что позволяет предотвратить потери заявок. Такая схема поступления заявок позволяет уменьшить общую загрузку системы, что, в свою очередь, сокращает время ожидания обслуживания.

С целью предотвращения перегрузок в узлах сетей с коммутацией пакетов (ATM, TCP/IP и др.) используются алгоритмы активного управления очередью (англ. *Active Queue Management* – AQM). В соответствующей системе обслуживания, моделирующей работу узла сети, каждый поступающий пакет может быть отброшен с определённой вероятностью, зависящей от длины очереди, даже если буфер ещё полностью не заполнен. Зависимость вероятности отбрасывания пакетов от длины очереди называют функцией отбрасывания [45].

Механизм функции отбрасывания является мощным средством для регулирования параметров системы обслуживания. С его помощью можно регулировать не только длину очереди, вероятность потери заявок, время ожидания, дисперсию длины очереди, но и несколько из этих параметров одновременно. Модели с функцией отбрасывания имеют также свой глубокий универсальный смысл. Во всех случаях, когда нет возможности регулировать параметры системы обслуживания через изменения входящего потока либо процесса обслуживания, применение функции отбрасывания является простым и эффективным способом для обеспечения требуемых параметров системы обслуживания.

Механизм функции отбрасывания заявок можно реализовать, рассматривая модель, предусматривающую функционирование системы в нескольких режимах, отличающихся друг от друга интенсивностями входящего потока.

Системы с переключениями режимов функционирования и групповым поступлением заявок являются более сложными для исследования по сравнению с классическими, поэтому для них еще острее стоит проблема получения рациональных вычислительных алгоритмов для стационарных характеристик, которые можно было бы применить как для систем с ограниченной, так и с неограниченной очередью. Под "рациональными" мы понимаем алгоритмы, быстродействие которых слабо зависит от значений ограничителя длины очереди и порогов переключения.

В настоящей работе мы строим именно такие алгоритмы для одноканальных систем обслуживания типа $M^{\theta}/G/1/m$ с пороговыми стратегиями функционирования, опираясь на метод потенциала, разработанный В. С. Королюком [35] для изучения непрерывных снизу случайных блужданий. Главное внимание мы сосредоточим на определении средней продолжительности периода занятости, стационарного распределения числа заявок в системе, стационарных значений вероятности потери заявок, средней длины очереди и среднего времени ожидания. Вместе с тем, наш подход, в отличие от большинства методов, используемых для изучения полумарковских моделей обслуживания, позволяет провести исследование не только стационарного, но и переходного режима работы системы, в частности, найти преобразования Лапласа для распределения числа заявок в течение периода занятости и для функции распределения периода занятости.

Остановимся на основных предположениях и обозначениях, используемых в настоящей работе.

Для обозначения систем с групповым поступлением заявок будем использовать верхний индекс θ. Например, через $M^\theta/G/1/m$ будем обозначать систему типа $M/G/1/m$ с групповым поступлением заявок. В системе $M^\theta/G/1/m$ входящий поток представляет собой пуассоновский стационарный (простейший) поток групп заявок, а длительности обслуживания заявок предполагаются независимыми одинаково распределёнными случайными величинами с конечным средним значением и произвольной функцией распределения вероятностей $F(x)$. Пусть $\theta_n -$ число заявок в n-ой группе, тогда $a_k = \mathbf{P}\{\theta_n = k\}$, $\sum_{k=1}^{\infty} a_k = 1$, $\sum_{k=1}^{\infty} k a_k = b < \infty$.

В первой главе настоящей монографии, состоящей из трёх глав, мы обсуждаем результаты и методы исследования для систем, близких к изучаемым в нашей работе. Здесь же приведены вспомогательные результаты, которые образуют математический аппарат исследования, базирующийся в основном на методе потенциала В. С. Королюка. Мы используем метод потенциала в модернизированном виде, приспособив к нашим потребностям. Для исследования систем с несколькими режимами функционирования возникает необходимость использования стольких базовых случайных блужданий и их потенциалов, сколько применяется различных режимов функционирования. Поэтому используемый нами подход мы называем *методом потенциалов*.

Во второй главе изучаются характеристики систем с пороговой блокировкой входящего потока, в которых в зависимости от длины очереди могут применяться различные режимы обслуживания. Рассматриваются также системы с двумя независимыми входящими потоками, в которых применяются пороговые и гистерезисные стратегии управления временем обслуживания и интенсивностью потока заявок.

Остановимся на обозначениях и описании рассмотренных здесь систем.

$M^{\theta}(h)/G,G_{h+1},\ldots,G_{m+1}/1/m$ – система типа $M^{\theta}/G/1/m$ с блокировкой входящего потока в режиме перегрузки. Режим перегрузки включается в момент t начала обслуживания заявки при выполнении условия $\xi(t) > h,$ где $\xi(t)$ – число заявок в системе в момент времени t. В режиме перегрузки функция распределения времени обслуживания заявки $G_n(x)$ зависит от числа n заявок в системе.

$M^{\theta}(h)/G,G_{h+1},\ldots,G_{\tilde{h}}/1$ – система типа $M^{\theta}(h)/G,G_{h+1},\ldots,G_{m+1}/1/m$ при $m = \infty$ и $G_n(x) = G_{\tilde{h}}(x)$ для $n \geq \tilde{h}.$

$M^{\theta}(h)/G,\tilde{G}/1$ – система типа $M^{\theta}(h)/G,G_{h+1},\ldots,G_{\tilde{h}}/1$ при $G_n(x) = G_{h+1}(x)$ для $n \geq h+1.$

$M^{\theta}(h_1,h_2)/G,\tilde{G}/1/m$ – система типа $M^{\theta}/G/1/m$ с блокировкой входящего потока в режиме перегрузки. Режим перегрузки включается в момент t начала обслуживания заявки при выполнении условия $\xi(t) > h_2.$ Процесс поступления заявок возобновляется в момент $\tilde{t}$ начала обслуживания той заявки, для которой $\xi(\tilde{t}) = h_1 \leq h_2.$ От момента начала обслуживания первой заявки во время блокировки до момента завершения блокировки входящего потока время обслуживания каждой заявки распределено по закону $\tilde{F}(x).$

$M_2^{\theta}(h)/G,\tilde{G}/1/m$ – система типа $M^{\theta}/G/1/m,$ в которую поступают два независимых потока групп заявок и используются три режима управления интенсивностью входящего потока: основной, режим частичной блокировки и режим полной блокировки. Промежутки времени между моментами поступления групп заявок i-го потока – независимые случайные величины, распределённые по показательному закону с параметром λ_i ($i = 1, 2$). В основном режиме в систему допускаются заявки первого и второго потоков, а функция распределения времени обслуживания каждой заявки равна $F(x).$ В режиме частичной блокировки принимаются лишь заявки первого потока, а распределение времени обслуживания не изменяется. В режиме полной блокировки прекращается приём всех заявок, а время обслуживания заявки распределено по закону $\tilde{F}(x).$ Пусть t_b – момент начала обслуживания заявки. Если $\xi(t_b) \leq h,$ то во время обслуживания данной заявки применяется основной режим работы системы. Переключение на режим частичной блокировки осуществляется в момент t_b начала обслуживания той первой заявки, для которой выполняются неравенства $h+1 \le \xi(t_b) \le m.$ Режим полной блокировки включается с момента достижения длины очереди значения m и длится до момента t_b начала обслуживания той заявки, для которой выполнится равенство $\xi(t_b) = h.$ Переключение на режим обслуживания с функцией распределения времени обслуживания $\tilde{F}(x)$ осуществляется не в момент начала функционирования режима полной блокировки, а в момент начала обслуживания первой заявки во время действия этого режима.

$\mathrm{M}_2^{\theta}(\mathrm{h}_1,\mathrm{h}_2)/\mathrm{G},\tilde{\mathrm{G}}/1/\mathrm{m}$ – система типа $\mathrm{M}^{\theta}/\mathrm{G}/1/\mathrm{m},$ в которую поступают два независимых потока групп заявок и используются три режима управления интенсивностью входящего потока: основной, режим частичной блокировки и режим полной блокировки. Промежутки времени между моментами поступления групп заявок i-го потока – независимые случайные величины, распределённые по показательному закону с параметром λ_i $(i=1, 2)$. В основном режиме в систему допускаются заявки первого и второго потоков, а функция распределения времени обслуживания каждой заявки равна $F(x).$ В режиме частичной блокировки принимаются лишь заявки первого потока. В режиме полной блокировки прекращается приём всех заявок. В режимах частичной и полной блокировки время обслуживания заявки распределено по закону $\tilde{F}(x).$ Пусть t_b – момент начала обслуживания заявки. С момента поступления в систему первой заявки и во время обслуживания всех заявок, для которых выполнено условие $\xi(t_b) \le h_2,$ применяется основной режим работы системы. Переключение из основного режима на режим частичной блокировки осуществляется в момент t_b начала обслуживания той первой заявки, для которой выполняются неравенства $h_2+1 \le \xi(t_b) \le m.$ Режим полной блокировки включается с момента достижения длины очереди числа m и длится до момента t_b начала обслуживания той заявки, для которой выполнится равенство $\xi(t_b) = h_2.$ В этот момент происходит переключение на режим частичной блокировки. Из режима частичной блокировки возможно переключение или на основной режим, или на режим полной блокировки. Переключение на основной режим осуществляется в момент t_b начала обслуживания той заявки, для которой выполняется равенство $\xi(t_b) = h_1 \le h_2.$

Третья глава монографии посвящена исследованию систем с параметрами функционирования, зависящими от длины очереди. Время обслуживания каждой заявки определяется по правилу: если в момент начала обслуживания данной заявки в системе находится n заявок, то её времени обслуживания соответствует функция распределения $F_n(x)$. Параметры входящего потока для каждого режима также могут быть разными и определяются в моменты окончания обслуживания заявок. Здесь же рассмотрены системы с двухпороговой гистерезисной стратегией функционирования.

Приведём описание систем, изученных в третьей главе.

$\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G}_1,\ldots,\mathrm{G}_{\mathrm{m+1}}/1/\mathrm{m}$ – система типа $\mathrm{M}^\theta/\mathrm{G}/1/\mathrm{m}$, в которой различные режимы функционирования системы отличаются друг от друга параметрами входящего потока λ_n, a_{nj} $(n \geq 0)$ и функциями распределения $F_n(x)$ $(n \geq 1)$ времени обслуживания одной заявки, а переключения режимов может осуществляться в моменты окончания обслуживания заявок.

$\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G}_1,\ldots,\mathrm{G}_\mathrm{h}/1/\mathrm{m}$ – частный случай системы типа $\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G}_1,\ldots,\mathrm{G}_{\mathrm{m+1}}/1/\mathrm{m}$ при $F_n(x)=\tilde{F}(x)$, $\lambda_n=\tilde{\lambda}$ для $n \in \{h, h+1, \ldots, m+1\}$.

$\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G}_1,\ldots,\mathrm{G}_\mathrm{h}/1$ – система типа $\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G}_1,\ldots,\mathrm{G}_\mathrm{h}/1/\mathrm{m}$ при $m=\infty$.

$\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G},\tilde{\mathrm{G}}/1/\mathrm{m}$ – система типа $\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G}_1,\ldots,\mathrm{G}_{\mathrm{m+1}}/1/\mathrm{m}$, в которой $\lambda_n=\lambda$, $a_{ni}=a_i$ $(i \geq 1)$, $F_n(x)=F(x)$ для $n \in \{1, 2, \ldots, h-1\}$ и $\lambda_n=\tilde{\lambda}$, $a_{ni}=\tilde{a}_i$ $(i \geq 1)$, $F_n(x)=\tilde{F}(x)$ для $n \in \{h, h+1, \ldots, m+1\}$.

$\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G},\tilde{\mathrm{G}}/1$ – система типа $\mathrm{M}_\mathrm{n}^{\theta_\mathrm{n}}/\mathrm{G},\tilde{\mathrm{G}}/1/\mathrm{m}$ при $m=\infty$.

$\mathrm{M}^\theta(\mathrm{h}_1,\mathrm{h}_2)/\mathrm{G}(\mathrm{h}_1,\mathrm{h}_2),\tilde{\mathrm{G}}(\mathrm{h}_1,\mathrm{h}_2)/1/\mathrm{m}$ – система типа $\mathrm{M}^\theta/\mathrm{G}/1/\mathrm{m}$, в которой применяются два режима функционирования (основной режим и режим перегрузки) с функциями распределения времени обслуживания $F(x)$, $\tilde{F}(x)$ и параметрами λ, $\tilde{\lambda}$ простейших потоков групп заявок соответственно. Если t – момент начала обслуживания очередной заявки и $\xi(t) > h_2$, то во время обслуживания данной заявки применяется режим перегрузки. Возвращение к основному режиму осуществляется в момент начала обслуживания очередной заявки, для которой $\xi(t)=h_1 \leq h_2$. Итак, если в момент начала обслуживания заявки $\xi(t) \leq h_1$, то применяется основной режим, если $\xi(t) > h_2$ – режим перегрузки, а если $h_1 < \xi(t) \leq h_2$, то для очередной заявки сохраняется режим функционирования, который применялся для предыдущей заявки.

$\mathrm{M}^\theta/\mathrm{G}(\mathrm{h}_1,\mathrm{h}_2),\tilde{\mathrm{G}}(\mathrm{h}_1,\mathrm{h}_2)/1/\mathrm{m}$ – частный случай системы типа $\mathrm{M}^\theta(\mathrm{h}_1,\mathrm{h}_2)/\mathrm{G}(\mathrm{h}_1,\mathrm{h}_2),\tilde{\mathrm{G}}(\mathrm{h}_1,\mathrm{h}_2)/1/\mathrm{m}$ при $\tilde{\lambda}=\lambda$.

$M^{\theta}(h_1,h_2)/G(h_1,h_2),\tilde{G}(h_1,h_2)/1$ – система типа $M^{\theta}(h_1,h_2)/G(h_1,h_2),\tilde{G}(h_1,h_2)/1/m$ при $m=\infty$.

$M^{\theta}/G(h_1,h_2),\tilde{G}(h_1,h_2)/1$ – частный случай системы типа $M^{\theta}(h_1,h_2)/G(h_1,h_2),\tilde{G}(h_1,h_2)/1$ при $\tilde{\lambda}=\lambda$.

В монографии приведены примеры вычисления стационарных характеристик рассматриваемых систем. Результаты, полученные в каждом из этих примеров, проверены с помощью имитационных моделей соответствующих систем обслуживания, построенных с использованием инструментальных средств GPSS World [2, 29]. Тексты соответствующих программ на языке GPSS приведены в Приложениях 1–8.

Результаты исследований, представленные в настоящей монографии, основаны на работах авторов [18–28, 31].

1 ОБЗОР ЛИТЕРАТУРЫ. БАЗОВОЕ СЛУЧАЙНОЕ БЛУЖДАНИЕ

В данной главе мы приводим математический аппарат, основанный на модификации метода потенциала для случайных блужданий, который будет использован в последующих главах. Начинаем главу с обзора известных результатов, связанных с темой нашего исследования.

1.1 Обзор литературы

Остановимся на результатах и методах исследования для систем, близких к изучаемым в настоящей работе. Главное внимание сконцентрируем на различиях нашего подхода от известных.

Интенсивные исследования систем массового обслуживания начались в начале 60-х годов прошлого века и продолжаются до сих пор. Все известные идеи и методы этих исследований описанных во многих учебниках и монографиях. Укажем, в частности, на следующие: Анисимов В. В., Закусило О. К., Донченко В. С. [1], Боровков А. А. [3], Бочаров П. П., Печинкин А. В. [4], Вишневский В. М. [12], Гнеденко Б. В., Коваленко И. Н. [13], Дудин А. Н. [15], Ивченко Г. И., Каштанов В. А., Коваленко И. Н. [33], Cohen J. W. [46], Cooper R. B. [47], Kleinrock L. [54, 55], Medhi J. [58], Ng Chee-Hock [60], Takagi H. [67], Tijms H. C. [68], Wolff R. W. [70].

Большинство результатов теории массового обслуживания получены для систем с входящими потоками марковского типа. Системы с более общими потоками практически не изучены. Для них, в лучшем случае, получены лишь отдельные результаты о существовании стационарных режимов функционирования и сходимости к таким режимам [4, с. 77].

Исследование систем обслуживания с входящими потоками марковского типа можно условно разделить на два направления: 1) рассмотрение моделей с более сложными входящими потоками по сравнению с пуассоновским стационарным; 2) изучение систем с пуассоновским входящим потоком (систем типа M/G), в том числе систем с групповым поступлением заявок.

Первое направление исследований вызвано потребностями адекватного учёта особенностей потоков в современных телекоммуникационных сетях, что привело к моделированию входящих потоков в виде *марковского входящего потока* (Markov arrival process – MAP, поступление заявок описывает однородный марковский процесс с матрицей интенсивностей переходов блочного вида), его частного случая – *пуассоновского потока, управляемого цепью Маркова* (Markov modulated Poisson process – MMPP) [4, с. 84] и *группового марковского входящего потока* (batch Markov arrival process – BMAP), впервые рассмотренного Lucantoni D. [57] в 1991 г. Исследование систем с такими входящими потоками чрезвычайно трудоемкое, требует применения сложного и громоздкого аналитического аппарата и не всегда позволяет получить необходимую информацию о характеристиках системы обслуживания.

Нашу работу можно отнести к тем исследованиям второго направления, которые посвящены изучению систем с различными стратегиями функционирования. Среди систем типа M/G лучше всего изучены системы с неограниченной очередью $M/G/1$ [4, 12, 13, 15, 33, 47, 54, 58, 60, 67, 68]. Понятно, что традиционные методы, разработанные для таких систем (метод вложенных цепей Маркова, метод введения дополнительной переменной, метод введения дополнительного события, методы теории восстановления), исследователи менее или более успешно пытаются применять и к более сложным системам, в частности, и к тем, которые является объектом нашего исследования, – системам с групповым поступлением заявок, с возможностью применения различной интенсивности обслуживания, обусловленного изменением длины очереди, и с пороговой блокировкой входящего потока. Понятно также, что не все результаты, полученные этими методами для системы $M/G/1$ могут быть получены для более сложных систем. Поэтому мы сначала кратко охарактеризуем традиционные методы и известные результаты для системы $M/G/1$.

Метод вложенных цепей Маркова и метод введения дополнительной переменной основываются на построении некоторого марковского процесса $X(t)$, распределение и характеристики которого, если их удаётся найти, позволяют найти распределение и характеристики того немарковского процесса $Y(t)$, который нас интересует. Для построения процесса $X(t)$ применяют внешнюю или внутреннюю марковизацию процесса $Y(t)$.

Внешняя марковизация заключается в расширении пространства состояний системы за счёт введения дополнительных переменных с целью получения марковского процесса, для которого можно вывести системы дифференциальных уравнений в частных производных или интегро-дифференциальных уравнений для распределения вероятностей процесса. Для системы $M/G/1$ источником немарковости процесса является зависимость поведения этого процесса не только от его собственного значения в момент t, но и от времени, прошедшего до момента t с момента начала обслуживания заявки, обслуживаемой в момент t. Поэтому естественно марковский случайный процесс $X(t)$ взять в виде двумерного процесса $\{Y(t), \nu(t)\}$, где $\nu(t)$ – время, прошедшее с момента начала обслуживания заявки, обслуживаемой в момент t, или остаток времени до завершения обслуживания этой заявки. Недостатком внешней марковизации является сложность исследования расширенного марковского процесса $X(t)$, а преимуществом – простота получения распределения $Y(t)$ с помощью распределения процесса $X(t)$.

В меньшей степени этим недостатком (но и этим преимуществом) обладает подход, связанный с внутренней марковизацией, который разработан Кендаллом и называется методом вложенных цепей Маркова. Его суть заключается в том, что для немарковского случайного процесса $Y(t)$ ищут последовательность моментов времени τ_n $(n \geq 1)$, в которые этот процесс является марковским, и тогда процесс $Y(\tau_n)$ является дискретной цепью Маркова. Методами цепей Маркова ищут распределение вероятностей состояний цепи $Y(\tau_n)$, а затем по этому распределению восстанавливают распределение вероятностей состояний процесса $Y(t)$.

Кратко опишем суть метода введения дополнительного события. Пусть требуется найти некоторое распределение, которое характеризует функционирование системы обслуживания. Производящей функции этого распределения (если распределение дискретное) или его преобразованию Лапласа-Стилтьеса предоставляется вероятностный смысл за счет "раскрашивания" заявок или рассмотрения потока "катастроф". Затем рассматривают некоторое (дополнительное) случайное событие, вероятность которого вычисляют двумя различными способами, используя вероятностную интерпретацию преобразования Лапласа-Стилтьеса и производящей функции. Полученные выражения приравнивают и получают уравнение для искомой характеристики системы.

Использование теории восстановления основывается на наличии моментов регенерации, во время которых процесс, описывающий функционирование системы, полностью забывает своё "прошлое". Это позволяет свести анализ характеристик системы на заданном временном интервале $[0, t]$ к изучению этих характеристик на отдельном периоде регенерации, то есть периоде между восстановлениями, и, как результат, записать (преимущественно в терминах преобразований) совместное распределение некоторых характеристик системы.

Основные результаты для системы M/G/1, которые удалось получить традиционными методами, относятся к стационарному режиму. Переход к нестационарному случаю, то есть к изучению характеристик обслуживания как функции от времени, обычно связан с большими аналитическими трудностями. Итак, для стационарного режима функционирования системы M/G/1 получена формула Поллячека-Хинчина для производящей функции распределения числа заявок (см., например, [4, с. 259; 15, с. 44; 33, с. 101]); формула Поллячека-Хинчина для преобразования Лапласа-Стилтьеса функции распределения времени ожидания заявки в системе (см., например, [4, с. 276; 15, с. 45]); средние значения времени ожидания и времени пребывания заявки в системе [4, с. 276]; выведено функциональное уравнение для производящей функции распределения числа заявок, обслуженных за период занятости [15, с. 55], и функциональное уравнение для преобразования Лапласа-Стилтьеса функции распределения периода занятости [12, с. 58; 15, с. 54]. Решение первого из этих уравнений можно искать только приближённо, а решение второго известно лишь для случаев, когда распределение времени обслуживания является показательным или вырожденным [12, с. 59], но с помощью этого уравнения всё-таки удалось найти среднюю продолжительность периода занятости [15, с. 55]. Метод введения дополнительной переменной [4, с. 286] и использование теории восстановления [4, с. 310] дали одинаковый результат в терминах двойного преобразования для совместного стационарного распределения числа заявок в системе и времени ожидания начала обслуживания.

Для функции распределения виртуального времени ожидания системы M/G/1 в нестационарном случае выведено интегро-дифференциальное уравнение Такача [4, с. 275; 33, с. 112], решение которого найдено в терминах двойных преобразований [4, с. 278; 33, с. 117]. Полученный результат достаточно трудно использовать на практике, ведь для этого нужно обратить двойное преобразование.

Хотя ограничения на объём буферных накопителей является важным фактором, который необходимо учитывать при аналитическом моделировании очередей в реальных информационно-вычислительных системах и сетях, но системам с ограниченной очередью M/G/1/m посвящено значительно меньше исследований, чем системам M/G/1. Остановимся на некоторых результатах для системы M/G/1/m. Это – формулы для стационарного распределения длины очереди [53], точные изображения для преобразования Лапласа времени ожидания и его асимптотические свойства при $m \to \infty$ [69], преобразование Лапласа для периода занятости и формула для средней длительности этого периода [59].

Впервые систему $\mathrm{M}^{\theta}/\mathrm{G}/1$ с групповым поступлением заявок исследовал Gaver D. P. [48] с помощью метода вложенных цепей Маркова и теории восстановления. Методом вложенных цепей Маркова для этой системы получено среднее значение стационарного времени ожидания [47, с. 241] и производящая функция стационарного распределения числа заявок в моменты завершения обслуживания [41, с. 130].

Для систем $\mathrm{M}^{\theta}/\mathrm{G}/1$ с временем обслуживания, зависящим от длины очереди, традиционными методами не удаётся получить все результаты, которые получены для обычной системы $\mathrm{M}^{\theta}/\mathrm{G}/1$. Так, например, в [41, с. 127; 50] рассмотрена система типа $\mathrm{M}^{\theta}/\mathrm{G}/1$, в которой время обслуживания имеет функцию распределения $F_n(x)$, если в момент начала обслуживания заявки в системе находится n заявок. Применение для её исследования метода вложенных цепей Маркова позволяет записать систему линейных алгебраических уравнений для стационарного распределения числа заявок в моменты завершения обслуживания, но не удается найти производящую функцию этого распределения. К аналогичному результату приводит применение к этой системе марковской теории восстановления [43]. Используя метод введения дополнительной переменной, Ivnitskiy V. A. [52] для системы $\mathrm{M}^{\theta}/\mathrm{G}/1$, в которой интенсивности входящего потока, вероятности размера группы заявок и интенсивность обслуживания зависят от длины очереди, получил рекуррентные соотношения для преобразований Лапласа от переходных вероятностей распределения длины очереди. Эти соотношения очень сложны, поэтому не пригодны для числовых расчётов [43].

Опишем некоторые результаты, полученные традиционными методами теории массового обслуживания для систем типа $\mathrm{M}/\mathrm{G}/1/\mathrm{m}$ и $\mathrm{M}^{\theta}/\mathrm{G}/1/\mathrm{m}$ с пороговыми стратегиями обслуживания (с ограниченной и неограниченной очередью). Простейшими из них являются системы с одним порогом переключения режимов обслуживания (в наших обозначениях – это системы $\mathrm{M}/\mathrm{G},\widetilde{\mathrm{G}}/1/\mathrm{m}$ и $\mathrm{M}^{\theta}/\mathrm{G},\widetilde{\mathrm{G}}/1/\mathrm{m}$). Одним из первых систему $\mathrm{M}/\mathrm{G},\widetilde{\mathrm{G}}/1/\mathrm{m}$ исследовал Рыжиков Ю. И. [38] с помощью законов сохранения для линейчатых процессов [37] и предложил метод рекуррентного расчёта стационарного распределения числа заявок в системе. Методом дополнительных переменных в статье [44] для систем $M/G,\widetilde{G}/1/m$ с ограниченной и неограниченной очередью найдено значение стационарных вероятностей состояний в виде контурных интегралов (в том числе двойных), которые удалось вычислить аналитически только для случая показательных распределений времени обслуживания.

Двухпороговая гистерезисная стратегия переключения интенсивности обслуживания (пороги $h_1 < h_2$, две интенсивности обслуживания, переключение режимов происходит, когда количество заявок превышает h_2 и меньше h_1) впервые предложена в статье [61], где для системы $M/G/1$ найдено выражение для производящей функции стационарного распределения длины очереди, содержащее два неизвестных параметра. Для отыскания этих параметров применяется рекурсивный алгоритм. Такая же стратегия рассмотрена в [16] для системы $M^{\theta}/G/1$, в которой два режима работы отличаются друг от друга не только интенсивностями обслуживания, но и интенсивностями входящего потока. С помощью метода вложенных цепей Маркова найдено выражение для производящей функции стационарного распределения числа заявок в моменты завершения обслуживания и явная зависимость критерия качества функционирования системы от параметров h_1 и h_2. В статье [62] для системы $M^{\theta}/G/1$ применена такая же гистерезисная стратегия, как в [61], и с помощью марковской теории принятия решений определён класс правил переключения режимов обслуживания, позволяющий минимизировать стационарное среднее число заявок в системе.

В случае поступления заявок по одной входящий поток для систем типа $M/G/1$ является обычным пуассоновским процессом, а в случае группового поступления (система $M^{\theta}/G/1/m$) этот процесс становится обобщенным пуассоновским, что значительно усложняет аналитический аппарат исследования. В 1974 г. В. С. Королюк [34] предложил новый подход к изучению функционалов, связанных с флуктуациями полунепрерывного обобщённого процесса Пуассона, который в его монографии [35] был перенесён и на случай непрерывных снизу случайных блужданий. В трудах его учеников (Братийчука М. С. [10], Хусанова М. [40], Пирджанова Б. [36]) этот подход был успешно применён к анализу систем типа $M/G/1$ (как с неограниченной, так и с ограниченной очередью), системы $M^{\theta}/G/1/m$ и системы $G/M/1$ с двухпороговой гистерезисной стратегией переключения интенсивности обслуживания [11]. В 2007–2008 гг. появились работы Братийчука А. М. [5–9], в которых найдены точные оценки скорости сходимости распределения длины очереди к стационарному для системы $M^{\theta}/G/1/m$, разработан рациональный алгоритм вычисления этого стационарного распределения и исследована система $M^{\theta}/G/1/m$ с блокировкой входящего потока, которая является частным случаем изученной нами системы $M^{\theta}(h_1,h_2)/G,\tilde{G}/1/m$).

Системы с блокировкой входящего потока изучены недостаточно. Впервые они рассматривались в работах [66, 67], где были исследованы некоторые стационарные характеристики (длина очереди, время ожидания) таких систем. Аналогичные задачи рассматривались традиционными методами в [51, 71]. В нашей работе впервые исследованы системы типа $M^{\theta}/G/1/m$ с ограниченной и неограниченной очередью, в которых вместе с блокировкой входящего потока применяется пороговое переключение времени обслуживания.

Метод потенциала к системам типа $M^{\theta}/G/1/m$ с переключениями интенсивности обслуживания до сих пор не применялся. Как уже отмечалось во Введении, мы применяем подход, который базируется именно на методе потенциала В. С. Королюка, поэтому остановимся на его преимуществах перед другими методами теории массового обслуживания. Этот подход, в некотором смысле, является общим, поскольку его можно применять к различным функционалам от процесса обслуживания и одинаково успешно к системам с групповым и индивидуальным поступлением заявок.

Применение метода потенциала позволяет получить некоторые новые результаты, которые не были ранее получены традиционными методами. Как отмечалось выше, методом введения дополнительного события для системы $M/G/1$ было выведено функциональное уравнение для преобразования Лапласа-Стилтьеса функции распределения периода занятости [12, с. 58; 15, с. 54], решение которого известно лишь для случаев, когда распределение времени обслуживания является показательным или вырожденным. Аналогичного результата для системы $M/G/1/m$ вообще нет. Наш подход позволяет найти в явном виде преобразования Лапласа функции распределения периода занятости для систем с пороговыми стратегиями функционирования с ограниченной очередью, которые являются более сложными по сравнению с системой $M/G/1/m$. В настоящей работе впервые для этих систем найдено также преобразование Лапласа для распределения числа заявок в течение периода занятости.

Ещё одно преимущество нашего подхода заключается в том, что он даёт удобные алгоритмы для вычисления стационарного распределения числа заявок и других стационарных характеристик. Для систем с ограниченной очередью M/G/1/m и M^{θ}/G/1/m традиционными методами обычно получают рекуррентные соотношения для стационарных вероятностей, числовая реализация которых позволяет выразить все стационарные вероятности через вероятность отсутствия заявок в системе $p_0(m)$, которую ещё необходимо определять из условия нормировки (см, например, [38; 47, с. 235–236]). Такие алгоритмы требуют большого объёма вычислений и, если, имея распределение для конкретного m (m – максимальная длина очереди), необходимо определить его для другого значения (такая необходимость часто возникает в задачах оптимизации систем [15–17, 30, 32, 39]), то каждый раз нужно заново находить $p_0(m)$ из условия нормировки. Кроме этого, не все традиционные методы позволяют сразу определять стационарное распределение, соответствующее произвольным моментам времени. Например, стационарные вероятности, полученные методом вложенных цепей Маркова, соответствуют моментам завершения обслуживаний и для системы с ограниченной очередью не совпадают со стационарными вероятностями для произвольных моментов времени. Для соответствующих систем с неограниченной очередью некоторые авторы вообще не приводят формул для стационарных вероятностей, ограничиваясь выражением для производящей функции стационарного распределения (см, например, [15, с. 59; 33, с. 101]). Если стационарное распределение определять с помощью этой функции, то это можно сделать, выполняя последовательные трудоёмкие операции её дифференцирования.

Преимущество нашего подхода в том, что мы получаем явные формулы для всех сразу стационарных вероятностей систем с ограниченной очередью, все они выражаются через среднюю продолжительность периода занятости системы. Чтобы найти стационарные характеристики для других значений порога h или параметра m, можно использовать уже найденные ранее значения характеристик базового блуждания R_i, π_i и q_i, которые не зависят от h и m и являются коэффициентами в формулах для $\mathbf{E}\tau(m)$ и $p_k(m)$.

Имея формулы для средней продолжительности периода занятости и стационарного распределения числа заявок для системы с ограниченной очередью, мы можем осуществить предельный переход $m \to \infty$ и сразу получить соответствующие формулы для системы с неограниченной очередью. Пользуясь полученными явными выражениями для стационарных характеристик, мы можем исследовать характер зависимостей этих характеристик от параметров h и m.

Информация о характере этих зависимостей может быть использована для решения задач оптимального синтеза систем с заданными характеристиками.

Купить книгу

Чтобы скопировать адрес, нажмите «просмотр текста»

https://www.ljubljuknigi.ru/store/gb/book/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4-%D0%BF%D0%BE%D1%82%D0%B5%D0%BD%D1%86%D0%B8%D0%B0%D0%BB%D0%BE%D0%B2-%D0%B4%D0%BB%D1%8F-%D0%BF%D0%BE%D1%80%D0%BE%D0%B3%D0%BE%D0%B2%D1%8B%D1%85-%D1%81%D1%82%D1%80%D0%B0%D1%82%D0%B5%D0%B3%D0%B8%D0%B9-%D0%BE%D0%B1%D1%81%D0%BB%D1%83%D0%B6%D0%B8%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F/isbn/978-3-659-33412-2

Buy the book

To copy the address, click "Text Viewer"

https://www.morebooks.de/store/gb/book/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4-%D0%BF%D0%BE%D1%82%D0%B5%D0%BD%D1%86%D0%B8%D0%B0%D0%BB%D0%BE%D0%B2-%D0%B4%D0%BB%D1%8F-%D0%BF%D0%BE%D1%80%D0%BE%D0%B3%D0%BE%D0%B2%D1%8B%D1%85-%D1%81%D1%82%D1%80%D0%B0%D1%82%D0%B5%D0%B3%D0%B8%D0%B9-%D0%BE%D0%B1%D1%81%D0%BB%D1%83%D0%B6%D0%B8%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F/isbn/978-3-659-33412-2