На правах рукописи

Горбунова Анастасия Владимировна

МЕТОДЫ И АЛГОРИТМЫ АНАЛИЗА И УПРАВЛЕНИЯ ДЛЯ СТОХАСТИЧЕСКИХ СИСТЕМ С РАЗДЕЛЕНИЕМ И ПАРАЛЛЕЛЬНЫМ ОБСЛУЖИВАНИЕМ

Специальность 2.3.1 — Системный анализ, управление и обработка информации, статистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

доктора физико-математических наук

Москва — 2026

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институт проблем управления им. В.А. Трапезникова Российской академии наук (ИПУ РАН).

Научный консультант:	**Лебедев Алексей Викторович**, доктор физико-математических наук, доцент, МГУ им. М.В. Ломоносова.
Официальные оппоненты:	**Зейфман Александр Израилевич**, доктор физико-математических наук, профессор, заведующий кафедрой прикладной математики Федерального государственного бюджетного образовательного учреждения высшего образования «Вологодский государственный университет». **Моисеева Светлана Петровна**, доктор физико-математических наук, профессор, заведующий кафедрой теории вероятностей и математической статистики Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский Томский государственный университет». **Орлов Юрий Николаевич**, доктор физико-математических наук, доцент, главный научный сотрудник Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук»
Ведущая организация:	**Федеральное государственное бюджетное учреждение науки Институт прикладной математики Дальневосточного отделения Российской академии**

Защита состоится 20 октября 2026 г. в 11:00 на заседании диссертационного совета 24.1.224.01 на базе ФИЦ ИУ РАН по адресу: 117312, г. Москва, пр.-т 60-летия Октября, д. 9.

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН по адресу: 119333, г. Москва, ул. Вавилова, 44, корп. 2 и на официальном сайте http://www.frccsc.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: 117312, г. Москва, пр.-т 60-летия Октября, д. 9, ученому секретарю диссертационного совета 24.1.224.01.

Автореферат разослан “__”_____ 2026 года.

Ученый секретарь диссертационного совета 24.1.224.01, д.т.н. / И.В. Смирнов

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Стохастическая система с разделением и параллельным обслуживанием представляет собой сложную систему, в которой объемная задача разделяется на более мелкие составляющие, процесс обработки которых происходит в параллельном режиме. Параллельная структура лежит в основе множества реальных процессов, начиная с производственных систем и заканчивая техническими приложениями с использованием параллельных или распределенных вычислений, поэтому построение и исследование их моделей в рамках системного анализа имеют высокую прикладную востребованность.

Системы массового обслуживания (СМО) с параллельной архитектурой как вероятностные модели стохастических систем с разделением и параллельным обслуживанием различной природы являются предметом исследования множества авторов на протяжении уже многих лет. При этом исследования продолжаются и ведутся вплоть до настоящего времени. Это объясняется, с одной стороны, актуальностью использования данной СМО как математической модели для современных рабочих или производственных процессов, а также для различных вычислительных систем. С другой стороны, одной из причин столь растянувшихся во времени исследований является сложность анализа данной системы массового обслуживания. Даже в наиболее простом ее варианте с пуассоновским входящим потоком и экспоненциальными временами обслуживания при числе подзаявок больше двух точных решений для среднего времени отклика до сих пор не получено. Многообразие индивидуальных характеристик составных элементов таких систем приводит к многообразию различного рода методов и подходов к определению их основных характеристик производительности.

Стоит отметить, что известны несколько типов СМО с различными механизмами распараллеливания заявок. Начало исследований таких систем приходится примерно на конец 1970-х годов. Однако в диссертации пойдет речь о классической системе с разделением и параллельным обслуживанием. Основная особенность функционирования этой систе-

мы заключается в том, что при поступлении в нее заявка разделяется на фиксированное количество подзаявок. Затем каждая из подзаявок встает в очередь на обслуживание к своему прибору. Заявка считается полностью обслуженной только после того, как обслужится последняя из ее подзаявок. Таким образом, время пребывания заявки в системе характеризуется наиболее длительным временем пребывания одной из составляющих ее частей.

Возвращаясь к наиболее востребованной области применения математической модели системы с разделением и параллельным обслуживанием стоит отметить, что объем информации, подвергающейся обработке в различных целях, заметно растет, поэтому применение параллельных вычислений актуально для большинства центров обработки данных. Возможности параллельных вычислений используют и социальные сети и поисковые системы. Кроме того, программные приложения в производственных, медицинских, научных и других отраслях обращаются к услугам вычислительных центров, в том числе и облачным, для анализа больших данных. С целью поддержания высокого качества обслуживания пользователей либо его улучшения в условиях конкурентной борьбы, поставщики услуг, очевидно, заинтересованы в более точных прогнозах показателей качества обслуживания при различных уровнях загрузки системы и, соответственно, в разработке методов и алгоритмов их получения в том числе с целью управления такими системами, т. к. от этого напрямую зависит количество выделяемых ресурсов (необходимого оборудования), а соответственно, и материальных затрат на его покупку и эксплуатацию.

Таким образом, разработка новых методов и алгоритмов анализа, а также управления для стохастических систем с разделением и параллельным обслуживанием, улучшающих качество известных результатов, а также позволяющих оценить ранее не изученные характеристики систем, например, остаточное время обслуживания, является актуальной задачей в рамках системного анализа.

Степень научной разработанности темы

R. Nelson и A.N. Tantawi в своей работе представили точное выражение для среднего времени отклика системы с разделением и параллельным

обслуживанием для случая двух подсистем с пуассоновским входящим потоком и экспоненциальными временами обслуживания на приборах. Для большего количества подсистем известны лишь приближения для среднего времени отклика различной степени точности, которые были получены, например, E. Varki, E. Merchant, H. Chen, S. Varma, A.M. Makowski. R. Nelson и A.N. Tantawi также вывели свою оценку, причем она считается одной из лучших по качеству приближения на определенной области параметров модели. Дисперсия времени отклика изучена гораздо в меньшей степени, тем не менее, здесь тоже есть некоторые приближенные оценки.

Для оценки моментов времени отклика систем различной архитектуры или их верхних и нижних границ авторы S. Balsamo, L. Donatiello, N.M. Van Dijk, A.S. Lebrecht, W.J. Knottenbelt применяли и матрично-геометрические методы, и элементы теории порядковых статистик.

В последние годы наряду с исследованиями моментов времени отклика появились работы M. Nguyen, S. Alesawi, N. Li, H. Che, Zh. Qiu, J.F. Perez, P.G. Harrison по анализу квантилей распределения времени отклика высокого уровня, в том числе и для случая распределений с тяжелыми хвостами, которые показывают неплохое качество приближения при высоких значениях загрузки системы.

Каждый из известных методов приближенного анализа систем с разделением и параллельным обслуживанием имеет определенные достоинства и недостатки, обычно связанные с качеством аппроксимации и применимостью метода для определенного набора распределений, входных параметров.

Несмотря на то, что исследования рассматриваемой системы велись в большей степени иностранными авторами, стоит отметить работы отечественных авторов, связанные с исследованиями различных типов и структур систем с разделением и параллельным обслуживанием, в частности, работы С.П. Моисеевой, В.И. Клименок, В.М. Вишневского, И.Е. Тананко, И.А. Синяковой (Ивановской), И.А. Захорольной, Л.А. Жидковой, О.А. Осипова, Р.С. Хабарова, В.А. Лохвицкого, А.С. Дудкина, Н.М. Редругиной и др.

Объект и предмет исследования

Объектом исследования диссертационной работы являются стохастические системы с разделением и параллельным обслуживанием. **Предметом** исследования являются вероятностные модели, методы и алгоритмы анализа и управления для систем с разделением и параллельным обслуживанием.

Цели и задачи

Диссертация посвящена решению фундаментальной научной проблемы — разработке вероятностных моделей, методов и алгоритмов анализа и управления для стохастических систем с разделением и параллельным обслуживанием. Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Разработать комплекс методов и алгоритмов оценки основных характеристик времени отклика систем с разделением и параллельным обслуживанием: моментов и квантилей его распределения, а также коэффициентов корреляции между временами пребывания в подсистемах.

2. Разработать метод определения оптимальной интенсивности обслуживания в системах с разделением и параллельным обслуживанием в зависимости от интенсивности входящего потока.

3. Разработать метод определения характеристик остаточного времени обслуживания, т. е. времени, необходимого для корректного завершения работы системы после отключения входящего потока, в системах с разделением и параллельным обслуживанием.

4. Продемонстрировать применимость предложенных методов на примерах конкретных типов распределений для входящего и обслуживающего потоков.

Научная новизна

Научная новизна диссертации заключается в следующем:

1. Разработан новый метод анализа систем с разделением и параллельным обслуживанием на основе машинного обучения, результатом применения которого является обученная интеллектуальная модель, позволяющая оценить интересующие характеристики

для любых промежуточных значений входных параметров из заданного интервала.

2. Разработан новый комплексный метод анализа систем с разделением и параллельным обслуживанием, включающий в себя множественную регрессию, визуальный анализ данных, имитационное моделирование и метод оптимизации, результатом применения которого являются аналитические выражения для оценки различных характеристик системы с разделением и параллельным обслуживанием.

3. Впервые получены точные формулы для коэффициентов корреляции Пирсона и Спирмена, а также аналитическое приближение коэффициента корреляции Кендалла между временами пребывания в подсистемах системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком и экспоненциальным распределением времен обслуживания.

4. Впервые разработана мета-гауссовская модель для оценки характеристик времени отклика системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком и экспоненциальным распределением времен обслуживания.

5. Разработан новый метод оценки квантилей распределения времени отклика систем с разделением и параллельным обслуживанием с пуассоновским входящим потоком на основе элементов теории копул и их диагональных сечений.

6. Разработан новый метод оценки квантилей распределения времени отклика систем с разделением и параллельным обслуживанием с различными вариантами распределений для входящего потока и распределением со степенным хвостом времен обслуживания на примере распределения Парето на основе аппроксимации распределения времени отклика распределением Фреше и метода моментов.

7. Разработаны новая модель и на ее основе метод определения оптимальной интенсивности обслуживания для систем с разделением и параллельным обслуживанием на примере систем с пуассоновским входящим потоком и экспоненциальным распределением или распределением Парето времен обслуживания.

8. Разработан новый метод определения характеристик остаточного времени обслуживания систем с разделением и параллельным обслуживанием с двумя бесконечнолинейными подсистемами с различными вариантами распределений для времени обслуживания.

Теоретическая значимость. Разработан комплекс алгоритмов и методов системного анализа, а также управления системой с разделением и параллельным обслуживанием. Для анализа впервые предложены метод на основе машинного обучения, комплексный метод с использованием множественной регрессии, метода оптимизации и визуального анализа данных. Впервые получены точные выражения для коэффициентов корреляции между временами пребывания подзаявок в подсистемах системы с пуассоновским входящим потоком и экспоненциальным распределением времен обслуживания на приборах, а также приближенные формулы для оценки коэффициентов корреляции между временами пребывания подзаявок в подсистемах системы с пуассоновским входящим потоком и распределением Парето времен обслуживания на приборах. Оценивание коэффициентов корреляции важно потому, что расширяет довольно ограниченную линейку известных методов для анализа времени отклика системы. Получены оценки для копул времен пребывания подзаявок в подсистемах системы с разделением и параллельным обслуживанием. Копула исчерпывающе описывает зависимость случайных величин в чистом виде. Современный математический аппарат теории копул активно развивается и применяется в последние десятилетия, однако в теории массового обслуживания он пока представлен мало. Таким образом, указанные характеристики позволяют провести полноценный системный анализ, поскольку ранее имеющаяся зависимость между временами пребывания подзаявок не исследовалась. Предложена модель управления для систем с

разделением и параллельным обслуживанием, которая позволяет определить оптимальное значение интенсивности обслуживания в зависимости от интенсивности входящего потока. Также получены формулы для функции распределения остаточного времени обслуживания системы с разделением и параллельным обслуживанием с двумя бесконечнолинейными подсистемами с различными вариантами распределений для времени обслуживания на приборах.

Практическая значимость. Системы массового обслуживания с разделением и параллельным обслуживанием широко используются для моделирования различного рода процессов, в рамках которых происходит разделение или распараллеливание задачи, в частности, в области информационных технологий при моделировании процесса функционирования высокопроизводительных вычислительных сред, использующих для повышения производительности различные методы распараллеливания. Определение различных характеристик физических систем с разделением и параллельным обслуживанием, например таких как математическое ожидание, дисперсия или квантили распределения времени отклика системы, а также оптимальное значение интенсивности обслуживания с точки зрения оптимизации финансовых показателей системы, очевидно, позволяет провести полноценный системный анализ и, как следствие, корректное проектирование системы, а также адекватное прогнозирование ее поведения в различных условиях.

Основные положения, выносимые на защиту

1. Метод анализа основных характеристик систем с разделением и параллельным обслуживанием на основе машинного обучения, результатом применения которого является обученная интеллектуальная модель, позволяющая оценить интересующие характеристики для любых промежуточных значений входных параметров из заданного интервала.

2. Комплексный метод получения аналитических оценок характеристик систем с разделением и параллельным обслуживанием, включающий в себя множественную регрессию, визуальный ана-

лиз данных, имитационное моделирование и метод оптимизации.

3. Мета-гауссовская модель для оценки характеристик времени отклика системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком и экспоненциальным распределением времен обслуживания.

4. Метод оценки квантилей распределения времени отклика систем с разделением и параллельным обслуживанием на основе элементов теории копул и их диагональных сечений, а также на основе аппроксимации распределения времени отклика системы распределением Фреше и метода моментов для случая распределения со степенным хвостом времен обслуживания на примере распределения Парето.

5. Метод определения оптимальной интенсивности обслуживания для систем с разделением и параллельным обслуживанием на примере систем с пуассоновским входящим потоком, экспоненциальным распределением или распределением Парето времен обслуживания.

6. Метод определения характеристик остаточного времени обслуживания для систем с разделением и параллельным обслуживанием с двумя бесконечнолинейными подсистемами с различными вариантами распределений для времени обслуживания.

Соответствие паспорту специальности

Данная диссертационная работа соответствует специальности 2.3.1 “Системный анализ, управление и обработка информации, статистика” по следующим пунктам:

— п.1. Теоретические основы и методы системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта.

— п.2. Формализация и постановка задач системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта.

— п.4. Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта.

— п.11. Методы и алгоритмы прогнозирования и оценки эффективности, качества, надежности функционирования сложных систем управления и их элементов.

Методы исследования

В диссертационной работе применяются методы теории массового обслуживания, теории вероятностей и случайных процессов, методы статистического анализа данных и теории порядковых статистик, элементы теории копул, методы математического моделирования (метод Монте–Карло), численные методы решения уравнений, методы машинного обучения (искусственные нейронные сети), множественная регрессия, методы оптимизации. При реализации имитационного моделирования систем массового обслуживания, численных методов и методов оптимизации используется программная среда Python.

Степень обоснованности и достоверности полученных результатов

Достоверность полученных в диссертационной работе результатов обосновывается строгими математическими доказательствами теорем, лемм и утверждений, корректностью разработанных методов исследования с использованием классических методов исследования, а также подтверждается согласованностью теоретических результатов с результатами вычислительных экспериментов, проведенных с помощью компьютерного моделирования.

Апробация результатов

По тематике диссертационной работы были сделаны доклады на следующих российских и международных конференциях: Аналитические и вычислительные методы в теории вероятностей и её приложениях (АВМТВ–2017, Москва); IX Московская международная конференция по Исследованию Операций (ORM-2018 Germeyer-100, Москва); IX Всероссийская конференция с международным участием «Информационно-телекоммуникационные технологии и математи-

ческое моделирование высокотехнологичных систем» (2019, Москва); 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT-2019, Dublin); XX Всероссийский Симпозиум по прикладной и промышленной математике (осенняя открытая сессия) (2019, Сочи); VI Всероссийская научно-практическая конференция с международным участием «Современные проблемы физико-математических наук» (2020, Орел); XX Международная конференция имени А. Ф. Терпугова «Информационные технологии и математическое моделирование» (ИТММ-2021, Томск); 24-я Международная научная конференция «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь» (DCCN-2021, Москва); XVII Всероссийская школа-конференция молодых ученых «Управление большими системами» (Москва, 2021); 25-я Международная научная конференция «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь» (DCCN-2022, Москва); XXIII Международная конференция им. А. Ф. Терпугова «Информационные технологии и математическое моделирование» (ИТММ-2024, Томск); X Всероссийская научно-практическая конференция «Современные проблемы физико-математических наук» (СПФМН-2024, Орел); XXIV International conference Mathematical Optimization Theory and Operations Research (MOTOR 2025, Новосибирск); 1-ая Международная научная конференция «Школа теории массового обслуживания» (ШТМО-2025, Томск).

Также основные положения диссертации докладывались и обсуждались в рамках докладов на следующих научных семинарах: расширенный семинар лаборатории №27 в ИПУ РАН (2024 г.); общемосковский постоянный научный семинар «Теория автоматического управления и оптимизации» в ИПУ РАН (2024 г.); семинар «Свертки, теория информации, массового обслуживания, надежности» кафедры теории вероятностей механико–математического факультета МГУ им. М.В. Ломоносова (2024 г.); общемосковский постоянный научный семинар «Теория автоматического управления и оптимизации» в ИПУ РАН (2025 г.).

Публикации

По теме диссертационной работы опубликовано 20 работ, из которых

15 — в изданиях из списка ВАК категории К1 и приравненных к ним, а именно: 6 статей в рецензируемых научных изданиях из Перечня ВАК (категория К1), 9 публикаций в журналах, индексируемых в Scopus (1 — Q1, 1 — Q2, 7 — Q3), и 5 статей в других сборниках, индексируемых в Scopus.

Личный вклад соискателя.
Все результаты исследований, изложенные в диссертационной работе и вынесенные на защиту, выполнены лично соискателем. Направления исследований, формулировки проблем и постановки задач обсуждались с научным консультантом, доктором физико-математических наук А. В. Лебедевым, что отражено в совместных публикациях, в которых основные результаты и их доказательства принадлежат автору диссертационной работы. В работах, опубликованных в соавторстве с доктором технических наук В. М. Вишневским, все статьи подготовлены автором диссертации самостоятельно (включая текстовое оформление, математические выкладки, анализ результатов и формулировку выводов). В. М. Вишневским осуществлялась постановка задач, определившая общее направление исследований.

Структура и объем диссертации
Диссертационная работа состоит из введения, пяти глав, заключения и списка литературы. Работа изложена на 338 страницах, содержит 33 таблицы и 110 иллюстраций. Библиография включает 210 наименований.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы диссертационного исследования, а также ее научная ценность и значимость; сформулированы цели и задачи работы, представлен обзор литературы, посвященной анализу стохастических систем с разделением и параллельным обслуживанием, приведены основные положения, выносимые на защиту, а также описаны структура и краткое содержание диссертации.

Первая глава диссертации посвящена обзору результатов применения интеллектуального анализа и методов машинного обучения, в

частности, искусственных нейронных сетей для исследования различных стохастических систем, сформулирована концепция, важные положения и этапы нового метода.

Рассмотренные публикации делятся на несколько категорий — статьи, в которых машинные алгоритмы служат для прогнозирования интересующих параметров реальных систем массового обслуживания (СМО) с использованием накопленной реальной статистики иногда без разработки их строгих математических моделей как таковых и работы, в которых данные методы применяются для анализа моделей массового обслуживания. Приводятся основные параметры интеллектуальных моделей, построенных для анализа интересующих систем — их архитектура, входные и выходные параметры, а также алгоритмы обучения. Поскольку идея нового подхода возникла относительно недавно, в настоящей главе рассматриваются и довольно ранние публикации, в которых данная тематика была затронута лишь косвенно. Тем не менее, подобные публикации заслуживают внимания, поскольку позволяют проследить развитие новой методики от ее исходной точки до современного состояния.

Идея применения искусственных нейронных сетей (ИНС) для исследований СМО в сущности обуславливается задачами, которые могут быть решены с их помощью. Как правило, выделяют три основных типа задач, которые могут быть решены посредством применения различных алгоритмов машинного обучения:

1) классификация (отнесение новых объектов к заранее определенным классам);

2) кластеризация (разбиение множества объектов на классы в ситуации, когда ни количество классов, ни их характерные свойства заранее не известны);

3) прогнозирование.

Последняя задача заключается в предсказании поведения системы по ее предыдущим реакциям, что фактически сводится к задаче аппроксимации функции нескольких переменных. Таким образом, основной

интерес представляет применение искусственных нейронных сетей и других методов машинного обучения (или интеллектуального анализа) именно в контексте решения данной проблемы, поскольку, например, нейросети считаются одним из лучших инструментов аппроксимации функций.

Среди основных классических методов решения задач в области теории массового обслуживания (ТМО) можно выделить следующие три группы: аналитические и численные методы, имитационное моделирование. При этом нахождение основных показателей производительности систем массового обслуживания не всегда является возможным аналитически, что объясняется ограничениями, накладываемыми распределениями, характеризующими входящий поток и время обслуживания заявок. Численные же методы анализа математических моделей массового обслуживания, к примеру, итерационные или матрично-геометрические, могут оказаться довольно ресурсоемкими и энергозатратными даже несмотря на возможности современных вычислительных систем. Что касается имитационного моделирования, то оно является иногда единственно возможным методом анализа СМО.

Новый метод является комбинацией машинного обучения с имитационным моделированием систем. С одной стороны, он устраняет один из существенных недостатков имитационного моделирования, который заключается иногда в превышающих все разумные пределы временных затратах на его проведение, особенно если речь идет о сложных системах и, с другой стороны, обладает, пожалуй, одним из главных его преимуществ — универсальностью, т. е. подходит для исследования практически любых систем. Итак, перечислим следующие основные этапы метода с использованием машинного обучения и интеллектуального анализа:

1. получение посредством имитационного моделирования значений интересующих характеристик анализируемой системы для конечного набора значений из заданных числовых промежутков для входных параметров, от которых зависит производительность системы;

2. обучение интеллектуальной модели на полученных с помощью симуляции данных одним из методов машинного обучения с целью решения задачи прогнозирования;

3. практически мгновенная оценка искомых характеристик производительности для любых других промежуточных значений входных параметров на тех же числовых промежутках с помощью обученной интеллектуальной модели.

В качестве наглядных примеров и иллюстрации применения нового метода рассмотрены два варианта систем — случай замкнутых экспоненциальных сетей и случай открытых неэкспоненциальных сетей. Результаты применения нового метода позволяют говорить о возможности и перспективах его применения к оценке основных вероятностно-временных характеристик других сложных систем массового обслуживания и, в частности, систем с разделением и параллельным обслуживанием заявок, об анализе которых речь пойдет далее.

Во **второй главе** исследуется система с разделением и параллельным обслуживанием с пуассоновским входящим потоком и экспоненциальным распределением времени обслуживания на приборах. В силу того, что первые исследования, посвященные данной СМО появились за рубежом, и в русскоязычной литературе пока нет устоявшегося перевода для обозначения данной системы, используется в том числе и англоязычная версия названия системы, а имеено fork-join СМО. Особенности функционирования рассматриваемой СМО с разделением заявок описаны в первом разделе и заключаются в следующем (рис. 1):

1) в момент поступления заявки в систему происходит ее мгновенное разделение на K ($K \geqslant 2$) более мелких составляющих, т. е. подзаявок, каждая из которых, становится в свою очередь на обслуживание к прибору (если он занят) или мгновенно начинает обслуживаться, если соответствующий прибор свободен, причем считаем, что каждая подзаявка имеет свой тип, который должен соответствовать номеру прибора, на котором она будет обслуживаться;

2) после окончания обслуживания подзаявка попадает в так называемый буфер синхронизации и остается там до тех пор, пока все родственные ей подзаявки, т. е. подзаявки, изначально принадлежащие одной заявке, не закончат свое обслуживание, далее происходит мгновенная сборка целой заявки и только после этого заявка считается обслуженной и может покинуть систему.

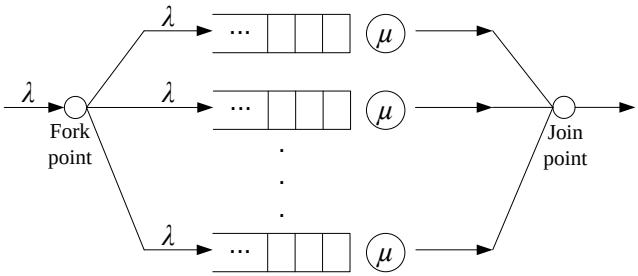

Рис. 1: Модель fork-join системы массового обслуживания с подсистемами типа $M_\lambda|M_\mu|1$.

Поскольку заявка считается обслуженной только в момент окончания обслуживания ее последней подзаявки, то для вычисления времени пребывания заявки в системе с учетом того, что моменты появления всех ее подзаявок в системе совпадают, достаточно определить максимум из всех времен пребывания ее подзаявок

$$R_K = \max(\xi_1, ..., \xi_K),$$

где ξ_i — это время пребывания подзаявки в i-ой подсистеме, $i = 1, ..., K$.

Однако задача вычисления точного значения среднего времени отклика даже несмотря на пуассоновский характер входящего потока и экспоненциальное время обслуживания на всех приборах в случае разделения на более чем две подзаявки остается до сих пор не решенной. Точное выражение в аналитическом виде известно только для $K = 2$, а для $K > 2$ были получены лишь приближения различной степени точности. Это объясняется сложностью анализа времен пребывания частей заявки в системе из-за существующей между ними зависимости

в силу общих моментов поступления. Времена пребывания подзаявок в системе являются положительно ассоциированными случайными величинами, и в силу этого их максимум стохастически не больше максимума независимых случайных величин с тем же распределением.

Для оценки основных характеристик fork-join системы во втором разделе используется подход с применением ИНС, подробно описанный в первой главе. Оцениваются математическое ожидание и среднеквадратическое отклонение времени отклика, т. е. $E[R_K]$ и $\sqrt{Var[R_K]}$, для которых известны приближенные выражения, чтобы можно было сравнить качество оценок, рассчитываемых с помощью нейросети и известной аналитики. Наилучший результат показывает именно обученный персептрон.

В третьем разделе приводятся новые аналитические оценки для математического ожидания и дисперсии времени отклика fork-join СМО. Предлагаемые оценки основываются на модификации известных приближений, полученных ранее. Для корректировки оценок также как и ранее необходимы экспериментальные данные, которые были получены путем имитационного моделирования. В итоге для среднего времени отклика имеем

$$E[R_K] \approx \frac{\left(\frac{H_K}{H_2} - 1\right)\rho}{\mu - \lambda} \cdot \left(C_1 - C_2\left(\frac{H_K}{H_2} - 1\right) + C_3\rho\right) + E[R_K]_{NT}. \quad (1)$$

Далее путем минимизации модуля максимального значения относительной погрешности приближения на всем рассматриваемом множестве данных, полученных посредством симуляции, определяются оптимальные значения коэффициентов C_1, C_2 и C_3, которые наилучшим образом отображают зависимость (1) в рамках выбранного метода оптимизации. В итоге после применения метода оптимизации Нелдера–Мида как к функции от нескольких переменных C_i в программной среде Python получаем

$$C_1 \approx 0.087197, \quad C_2 \approx 0.070236, \quad C_3 \approx 0.09638. \quad (2)$$

Введенная поправка улучшает модуль максимального значения отно-

сительной погрешности приближения $MaxAPE$ в 11.5 раз. Для оценки дисперсии времени отклика аналогичным образом получаем выражение

$$Var[R_K] \approx \frac{Q_K - 1}{(\mu - \lambda)^2} \cdot \big(1 + \rho\big(C_1 + C_2(H_K - 1) + \\ + C_3(H_K - 1)^2\big)\big) + \frac{1}{(\mu - \lambda)^2}, \tag{3}$$

где значения коэффициентов C_i

$$C_1 \approx -0.113658, \quad C_2 \approx 0.339780, \quad C_3 \approx 0.053745. \tag{4}$$

Новая оценка улучшает $MaxAPE$ для ранее известной в 25.7 раз.

В следующем разделе исследуется зависимость между временами пребывания подзаявок, которая является отличительной чертой fork-join СМО с K подсистемами $M|M|1$ (анализируемый тип СМО в данном конкретном случае) от K параллельно функционирующих СМО $M|M|1$. Поэтому отдельный интерес представляет оценка коэффициентов корреляции между временами пребывания подзаявок. Оказалось, можно получить даже не оценки коэффициентов корреляции Пирсона и Спирмена, а точные формулы для них традиционными методами для ТМО (производящие функции, преобразование Лапласа–Стилтьеса), также было получено приближение для коэффициента корреляции Кендалла.

Теорема 2.1. *Для системы с разделением и параллельным обслуживанием заявок с пуассоновским входящим потоком с параметром λ и экспоненциальным распределением времен обслуживания на однородных приборах с параметром μ, т. е. с K подсистемами типа $M_\lambda|M_\mu|1$, коэффициент корреляции Пирсона r_p между временами пребывания подзаявок для любой пары подсистем определяется следующим выражением*

$$r_p = \frac{\rho(4 - \rho)}{8}. \tag{5}$$

Теорема 2.2. *Для системы с разделением и параллельным обслуживанием заявок с пуассоновским входящим потоком с параметром λ*

и экспоненциальным распределением времен обслуживания на однородных приборах с параметром μ, т. е. с K подсистемами типа $M_\lambda|M_\mu|1$, коэффициент корреляции Спирмена r_s между временами пребывания подзаявок для любой пары подсистем определяется следующим выражением

$$r_s = \frac{12\sqrt{2}\sqrt{2-\rho}}{8-3\rho} - 3. \tag{6}$$

Из полученных результатов ясно, что все коэффициенты корреляции имеют некоторые пределы при высокой загрузке ($\rho \to 1$), причем эти пределы отличны как от 0, так и от 1, что наводит на мысль предположить в них содержательный смысл.

Следствие 2.1. *Предельные значения коэффициентов корреляции Пирсона и Спирмена времен пребывания подзаявок в любых двух подсистемах системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком и экспоненциальными временами обслуживания равны*

$$r_p = \frac{3}{8} = 0.375, \quad r_s = \frac{12\sqrt{2}}{5} - 3 \approx 0.394,$$

а оценка предельного значения коэффициента корреляции Кендалла имеет вид

$$r_k \approx 0.276.$$

Вычисление коэффициентов корреляции времен пребывания имеет не только академический интерес, но и может быть полезно для оценки характеристик системы. Действительно, когда речь идет о подборе приближенной модели зависимости времен пребывания, такая модель может быть параметризована одним из коэффициентов корреляции. В пятом разделе рассматривается мета-гауссовскую модель, основанная на сведении произвольного распределения к многомерному нормальному.

Теорема 2.4. *Для системы с разделением и параллельным обслуживанием с $K \geq 2$ подсистемами типа $M_\lambda|M_\mu|1$ в мета-гауссовской мо-*

дели справедливо приближение для времени отклика вида

$$\widehat{R}_K = -\frac{1}{\mu - \lambda} \ln\left(1 - \Phi(\sqrt{r}\varepsilon_0 + \sqrt{1-r}\max\{\varepsilon_1, \ldots, \varepsilon_K\})\right), \qquad (7)$$

где ε_i, $0 \le i \le K$, *— независимые стандартные нормальные случайные величины, а* Φ *— функция стандартного нормального распределения.*

Помимо первых или вторых моментов случайной величины времени отклика интерес представляют и квантили ее распределения. В основе подхода к построению оценки квантилей времени отклика в шестом разделе диссертации находится работа с копулами и их диагональными сечениями. Копулы представляют собой функции многомерного распределения на единичном кубе с равномерными частными распределениями. Согласно теореме Скляра, любое многомерное распределение раскладывается на частные распределения и копулу. Таким образом, копула исчерпывающе описывает зависимость случайных величин в чистом виде. Рассматривается частный случай двух подсистем ($K = 2$), однако напомним, что количество подсистем никак не влияет на зависимость в любой паре времен пребывания подзаявок одной заявки. Смысл изучения диагональных сечений копул в следующем. Если заданы случайные величины X_1 и X_2 с одинаковыми частными распределениями $F_1 = F_2 = F$ и копулой совместного распределения C, то их максимум $X_{\max} = \max\{X_1, X_2\}$ имеет функцию распределения

$$F_{\max}(x) = P(X_1 < x, X_2 < x) = C(F(x), F(x)) = \delta(F(x)), \qquad (8)$$

так что для ее вычисления достаточно знать только диагональное сечение, а не всю копулу.

Утверждение 2.1. *Для системы с разделением и параллельным обслуживанием с двумя (*$K = 2$*) подсистемами типа* $M_\lambda|M_\mu|1$*, в которой случайные величины* ξ_i, $i = 1, 2$, *являются временами пребывания подзаявок от одной заявки, квантили распределения времени отклика уровня* p *определяются выражением*

$$x_p = F_R^{-1}(p) = F^{-1}(\delta^{-1}(p)), \qquad (9)$$

где $F_{\xi_i}(x) = F(x) = 1 - e^{-(\mu-\lambda)x}$, $x \geqslant 0$.

В результате преобразований и применения методов интеллектуального анализа имеем, что квантили времени отклика аппроксимируются выражением

$$x_p \approx -\frac{\ln(1 - p^{\frac{1}{2-(C_1-C_2p^2)\rho}})}{\mu - \lambda}, \tag{10}$$

где $C_1 \approx 0.390327$, $C_2 \approx 0.237842$.

Также в рамках данного раздела представлены аналитические выражения, оценивающие саму копулу $C(u_1, u_2)$ и ее плотность. Оценка диагонального сечения (исходя из полученного выражения для копулы) имеет вид

$$\delta(u) \approx u^{2-C\rho}, \quad C \approx 0.369250, \tag{11}$$

поэтому можем сформулировать следующее утверждение.

Утверждение 2.2. *Если для системы с разделением и параллельным обслуживанием с двумя ($K = 2$) подсистемами типа $M_\lambda|M_\mu|1$ справедлива оценка диагонального сечения* (11) *копулы совместного распределения случайных времен пребывания подзаявок в подсистемах, то среднее время отклика системы аппроксимируется выражением*

$$E[R] \approx [\psi(3 - C\rho) - \psi(1)] \frac{\rho}{1 - \rho}, \tag{12}$$

где $\psi(\cdot)$ — *дигамма-функция.*

В седьмом разделе главы предлагается метод построения доверительных интервалов для среднего времени отклика fork-join СМО в условиях коррелированности данных имитационного моделирования.

Третья глава посвящена анализу системы с разделением и параллельным обслуживанием с распределением Парето времени обслуживания с функцией распределения вида

$$F(x) = 1 - \left(\frac{\alpha - 1}{\alpha} \cdot \frac{1}{x}\right)^{\alpha}, \quad x \geq \frac{\alpha - 1}{\alpha}, \quad \alpha > 3, \tag{13}$$

В первом разделе более детально описана математическая модель рассматриваемой системы. Во втором разделе получена верхняя граница

для среднего времени отклика системы на основе известных результатов для СМО $M|G|1$, а также элементов теории порядковых статистик.

Теорема 3.1. *Для системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ и распределением Парето времени обслуживания на приборах* (13) *верхняя граница для среднего времени отклика имеет вид*

$$E[R_K] \leqslant \mu_{Pa} + \sigma_{Pa}\frac{K-1}{\sqrt{2K-1}}, \tag{14}$$

где

$$\mu_{Pa} = 1 + \frac{\rho(\alpha-1)^2}{2\alpha(\alpha-2)(1-\rho)}, \tag{15}$$

$$\sigma_{Pa} = \sqrt{\mu_{Pa}^{(2)} - \mu_{Pa}^2}, \tag{16}$$

а

$$\begin{gathered}\mu_{Pa}^{(2)} = \frac{(\alpha-1)^2}{\alpha(\alpha-2)} + \frac{\rho(\alpha-1)^3}{3\alpha^2(\alpha-3)(1-\rho)} + \\ + \frac{\rho(\alpha-1)^2}{\alpha(\alpha-2)(1-\rho)} + \frac{\rho^2(\alpha-1)^4}{2\alpha^2(\alpha-2)^2(1-\rho)^2}.\end{gathered} \tag{17}$$

В третьем разделе с помощью применения нового метода с использованием машинного обучения обучается ИНС для оценки среднего времени отклика и его среднеквадратического отклонения, проводится сравнение с аналитикой. Обученная нейросеть показывает гораздо лучшие результаты.

В четвертом разделе проводится оценка основных характеристик системы с помощью комплексного метода на основе интеллектуального анализа данных. Такой же метод использовался при анализе характеристик системы в третьем разделе первой главы. Более детально остановимся на описании общего алгоритма нахождения оценок различных показателей fork-join системы.

1. Для начала анализируется искомая характеристика, т. е. делаются некоторые предположения о её зависимости от известных параметров модели.

2. Далее проводится имитационное моделирование с целью получе-

ния точных (максимально близких к точным) оценок исследуемой характеристики в зависимости от различных значений рассматриваемых параметров на некотором ограниченном (но достаточно обширном для приложений) интервале.

3. Затем проводится визуальный анализ предполагаемой зависимости на основе имеющихся данных.

4. На заключительном этапе с помощью метода оптимизации определяются оптимальные значения постоянных коэффициентов, минимизирующие максимальную относительную погрешность приближения при сравнении с данными, полученными с помощью симуляции.

С помощью представленного метода получены аналитические оценки для математического ожидания и среднеквадратического отклонения времени отклика системы, а также в пятом разделе — для коэффициентов корреляции Пирсона, Кендалла и Спирмена.

В шестом разделе для системы с Парето-распределенными временами обслуживания (и различными типами распределений для времени между соседними поступлениями заявок) предлагается использовать распределение Фреше (18)

$$\Phi_{a,b,\gamma}(x) = \begin{cases} 0, & x \le a, \\ \exp\left\{-\left(\dfrac{x-a}{b}\right)^{-\gamma}\right\}, & x > a, \end{cases} \quad , \quad b, \gamma > 0, \tag{18}$$

как известное распределение с тяжелым (степенным) хвостом, с целью приближения квантилей распределения времени отклика.

Теорема 3.2. *Пусть случайные величины $\xi_1, \ldots, \xi_n$, $n \ge 1$, имеют одинаковое частное распределение F и копулу совместного распределения C. Тогда при степенном диагональном сечении $\delta(u) = u^\nu$, $\nu > 0$, и $F = \Phi_{a,b,\gamma}$ получаем для их максимума M_n распределение того же типа.*

Тогда будет справедлива теорема, с помощью которой можно оценить квантили распределения времени отклика системы с разделением

и параллельным обслуживанием.

Теорема 3.3. *Если случайная величина времени отклика системы с разделением и параллельным обслуживанием R_K приближается распределением Фреше с функцией распределения $\Phi_{a,b,\gamma}(x)$ вида* (18)*, то оценки квантилей времени отклика уровня p представимы в виде*

$$\hat{x}_p = \hat{a} + \hat{b}(-\ln p)^{-1/\gamma}, \quad 0 < p < 1, \tag{19}$$

где

$$\hat{a} = E[R_K] - \hat{b}E[\xi_0], \quad \hat{b} = \sqrt{\frac{Var[R_K]}{Var[\xi_0]}}, \tag{20}$$

а случайная величина ξ_0 имеет стандартное распределение Фреше.

В седьмом разделе предлагается еще один метод оценивания квантилей времени отклика fork-join СМО для частного случая системы с разделением и параллельным обслуживанием, когда число подсистем $K = 2$ на примере, когда параметр $\alpha = 4$ (при этом предложенный метод может распространяться и на другие значения α). Метод тесно связан с элементами теории копул и аналогичен методу, предложенному во второй главе для fork-join системы с двумя подсистемами типа $M|M|1$. Однако основное отличие случая подсистем типа $M|Pa|1$ от случая экспоненциального обслуживания заключается в том, что вид функции распределения времени пребывания подзаявки в данной подсистеме неизвестен. В то время как для подсистемы типа $M|M|1$ время пребывания подзаявки имеет также экспоненциальное распределение.

В **четвёртой главе** решается задача управления режимом функционирования систем с разделением и параллельным обслуживанием: на основе построенной модели стоимости определяется оптимальная интенсивность обслуживания, обеспечивающая наилучшие финансовые показатели.

Базируется модель на естественных предположениях о необходимости минимизации среднего времени отклика системы для сохранения ее конкурентоспособности при разумных затратах на требуемые для этого ресурсы. В частности, под ресурсами может пониматься мощность

необходимого оборудования, которое позволяет быстрее обрабатывать клиентский запрос, если речь идет об информационно-вычислительных или производственных системах, например. Понятно, что чем мощнее оборудование, тем больше затрат требуется на его покупку, техническое обслуживание и содержание в целом. Таким образом, скорость работы оборудования (или в терминах СМО интенсивности обслуживания) пропорциональна росту его стоимости. Кроме того, с увеличением скорости обслуживания уменьшается время отклика системы. Таким образом, стоимость функционирования системы складывается из оптимального баланса между временем отклика и скоростью работы обслуживающих приборов.

Более подробно, предполагается, что установлены: 1) цена (штраф) за единицу среднего времени отклика и 2) цена за единицу интенсивности обслуживания. При этом первая цена для простоты полагается равной единице. Далее вычисляются стоимости времени отклика и обслуживания и складываются в общую стоимость затрат, которую хотим минимизировать. Таким образом, ставится задача стоимостной оптимизации управления.

В первом разделе главы более детально описывается математическая модель определения оптимальной стоимости функционирования fork-join системы с экспоненциальным распределением времени обслуживания. Обозначим стоимость функционирования системы через S и введем функцию $f(\rho)$, которая определяет выражение для среднего времени отклика системы в случае $\lambda = 1$, т. е. $f(\rho) = E[R_K]$ при $\lambda = 1$. Тогда в общем случае будет справедливо следующее выражение:

$$E[R_K] = \frac{1}{\lambda} f(\rho).$$

Далее, поскольку стоимость функционирования системы S зависит от среднего времени отклика системы (цену за единицу времени принимаем за единицу, $c_0 = 1$) и стоимости затрат на обслуживание, то можем для нее записать:

$$S = c_0 E[R_K] + c \cdot \mu,$$

где c — это стоимость единицы интенсивности обслуживания. Соответственно, с учетом введенной функции $f(\rho)$ можем переписать данное выражение следующим образом

$$S = \frac{c_0}{\lambda} f(\rho) + c\frac{\lambda}{\rho} = \frac{c_0}{\lambda}\left(f(\rho) + \frac{c\lambda^2}{c_0\rho}\right).$$

Пусть $c_1 = c\lambda^2/c_0$, тогда

$$S = \frac{1}{\lambda}\left(f(\rho) + \frac{c_1}{\rho}\right), \tag{21}$$

откуда получим уравнение

$$f'(\rho)\rho^2 = c_1, \tag{22}$$

решив которое, найдем оптимальное значение ρ_0 и, соответственно, оптимальное значение интенсивности обслуживания $\mu_0 = \lambda/\rho_0$.

Во втором разделе изучается частный случай системы с разделением, когда число подсистем равно двум и, соответственно, известна точная формула для среднего времени отклика.

Теорема 4.1. *Для системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром $\lambda > 0$ и двумя подсистемами типа $M|M|1$ с показательным распределением времени обслуживания с параметром $\mu > 0$ $(\lambda < \mu)$ в условиях стоимостной модели* (21) *оптимальное значение загрузки системы определяется выражением*

$$\rho_0 = y_0 + \frac{1}{2},$$

где

$$y_0 = \frac{-\sqrt{2t_1 + 8c_1 - 10.5} + \sqrt{Dis_2}}{2},$$

$$Dis_2 = -2t_1 + 8c_1 - 10.5 + \frac{16c_1 + 22}{\sqrt{2t_1 + 8c_1 - 10.5}},$$

причем выражение для t_1 при условии $Q = -\frac{44}{27}c_1\big(64c_1^3 - 288c_2^2 +$

$+135c_1 - 216\big) \geqslant 0$ *имеет вид*[1]

$$t_1 = \left(\frac{32}{3}c_1^2 - \frac{64}{27}c_1^3 + 6c_1 + 8 + \sqrt{\frac{1408}{3}c_1^3 - \frac{2816}{27}c_1^4 - 220c_1^2 + 352c_1}\right)^{\frac{1}{3}} +$$
$$+ \left(\frac{32}{3}c_1^2 - \frac{64}{27}c_1^3 + 6c_1 + 8 - \sqrt{\frac{1408}{3}c_1^3 - \frac{2816}{27}c_1^4 - 220c_1^2 + 352c_1}\right)^{\frac{1}{3}} -$$
$$- \frac{8c_1 - 10.5}{6}.$$

В третьем разделе рассматривается более общий случай, когда число подсистем системы с разделением и параллельным обслуживанием больше двух, а для среднего времени отклика используется оценка, полученная Нельсоном и Тантави

$$E[R_K] \approx E[R_K]_{NT} = \left[\frac{H_K}{H_2} + \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right)\rho\right]\frac{12-\rho}{8}\frac{1}{\mu - \lambda}. \qquad (23)$$

Теорема 4.2. *Для системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром $\lambda > 0$ и $K > 2$ подсистемами типа $M|M|1$ с показательным распределением времени обслуживания с параметром $\mu > 0$ ($\lambda < \mu$) в условиях стоимостной модели* (21) *и предположения о том, что формула Нельсона–Тантави* (23) *для определения среднего времени отклика системы является точной, оптимальное значение загрузки системы ρ_0 определяется решением уравнения*

$$8(H-1)\rho^5 + (60 - 71H)\rho^4 + (118H - 96)\rho^3 + \\ +11(c_2 - 12H)\rho^2 - 22c_2\rho + 11c_2 = 0, \qquad (24)$$

где $c_2 = 8c_1$, $H = H_K/H_2$.

Полученное уравнение (24) можно решить численно и определить оптимальное значение ρ_0 и, соответственно, искомое значение μ_0.

В четвертом разделе рассматривается обобщение формулы

[1] выражение для t_1 при $Q < 0$ определяется другими формулами, представленными в основном тексте диссертации

Нельсона–Тантави из первой главы, которое дает лучшее приближение для среднего времени отклика. Улучшение достигается за счет поправки к выражению из (23)

$$E[R_K] = \frac{\rho}{\mu - \lambda}\left(\frac{H_K}{H_2} - 1\right) \cdot \\ \cdot \left(Q_1 - Q_2\left(\frac{H_K}{H_2} - 1\right) + Q_3\rho\right) + E[R_K]_{NT}. \tag{25}$$

Теорема 4.3. *Для системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром $\lambda > 0$ и $K > 2$ подсистемами типа $M|M|1$ с показательным распределением времени обслуживания с параметром $\mu > 0$ ($\lambda < \mu$) в условиях стоимостной модели* (21) *и предположения о том, что формула* (25) *для определения среднего времени отклика системы является точной, оптимальное значение загрузки системы ρ_0 определяется решением уравнения*

$$8\rho^3(1-H)\Big(2Q_3\rho^2 - \rho(3Q_3 - (1-H)Q_2 - Q_1) - \\ -(2Q_1 + 2(1-H)Q_2)\Big) + \rho^2\Big(2M\rho^3 + (H - 15M)\rho^2 + \\ +(24M - 2H)\rho + 12H\Big) = 8c_1(1-\rho)^2,$$

где

$$Q_1 \approx 0.087197, \quad Q_2 \approx 0.070236, \quad Q_3 \approx 0.09638,$$

$$H = H_K/H_2, \quad M = 4(1-H)/11.$$

Для того, чтобы обстоятельно разобраться в поведении оптимальных решений в обоих случаях в пятом разделе проанализирована их асимптотика.

В следующем шестом разделе представлена математическая модель определения оптимальной стоимости функционирования fork-join системы с распределением Парето времени обслуживания аналогичная описанной во втором разделе. Обозначим стоимость функционирования системы с разделением и параллельным обслуживанием заявок

через S и введем функцию $g(\rho)$, которая определяет выражение для среднего времени отклика системы в случае $\mu = 1$, т. е. $g(\rho) = E[R_K]$ при $\mu = 1$. Тогда в общем случае будет справедливо следующее

$$E[R_K] = \frac{1}{\mu} g(\rho).$$

Далее аналогично, поскольку стоимость функционирования системы S зависит от среднего времени отклика системы и стоимости затрат на обслуживание, то можем для нее записать:

$$S = c_0 E[R_K] + c \cdot \mu,$$

$$S = \frac{c_0}{\mu} g(\rho) + c\frac{\lambda}{\rho} = \frac{c_0}{\lambda}\left(\rho g(\rho) + \frac{c\lambda^2}{c_0 \rho}\right).$$

Пусть $c_1 = c\lambda^2/c_0$ ($c_0 = 1$), тогда

$$S = \frac{1}{\lambda}\left(\rho g(\rho) + \frac{c_1}{\rho}\right), \tag{26}$$

откуда получим уравнение

$$(\rho g'(\rho) + g(\rho))\rho^2 = c_1, \tag{27}$$

решив которое, найдем оптимальное значение ρ_0 и, соответственно, оптимальное значение скорости обслуживания $\mu_0 = \lambda/\rho_0$.

Далее в седьмом разделе выведем уравнение для определения оптимального значения загрузки системы в общем случае, когда $K \geqslant 2$.

Для среднего времени отклика fork-join системы $E[R_K]$ с распределением Парето времени обслуживания в третьей главе была получена приближенная формула (в области $4 \leq \alpha \leq 10$, $2 \leq K \leq 20$), которая после некоторого упрощения в рамках доказательства следующей

теоремы будет использоваться в дальнейших вычислениях

$$E[R_K] \approx 1 + \frac{\rho(\alpha-1)^2}{2\alpha(\alpha-2)(1-\rho)} + (K^{\frac{1}{\alpha}} - 1)\cdot$$
$$\cdot \left(Q_1 + Q_2\alpha + Q_3\rho + Q_4\alpha\rho + Q_5\rho^2 + Q_6\alpha^2\right)\cdot \qquad (28)$$
$$\cdot\sqrt{\mu_{Pa}^{(2)} - \left(1 + \frac{\rho(\alpha-1)^2}{2\alpha(\alpha-2)(1-\rho)}\right)^2},$$

где коэффициенты Q_i, $i = 1, ..., 6$, имеют следующие значения:

$$Q_1 \approx 1.25918, \quad Q_2 \approx 0.36996, \quad Q_3 \approx -1.97400,$$

$$Q_4 \approx -0.28495, \quad Q_5 \approx 1.40841, \quad Q_6 \approx -0.01122.$$

Теорема 4.7. *Для системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром $\lambda > 0$ и $K \geqslant 2$ подсистемами типа $M|Pa|1$ с распределением Парето времени обслуживания вида* (13) *в условиях стоимостной модели* (26) *и предположения о том, что формула* (28) *для определения среднего времени отклика системы является точной, оптимальное значение загрузки системы ρ_0 определяется решением уравнения*

$$\rho^3 g'(\rho) + \rho^2 g(\rho) = c_1, \qquad (29)$$

где $g(\rho)$ и $g'(\rho)$ определяются выражениями

$$g(\rho) = 1 + \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{\rho}{1-\rho} + (K^{\frac{1}{\alpha}} - 1)\cdot$$
$$\cdot \left((Q_1 + Q_2\alpha + Q_6\alpha^2) + (Q_3 + Q_4\alpha)\rho + Q_5\rho^2\right)\cdot \qquad (30)$$
$$\cdot \left(\frac{1}{\alpha(\alpha-2)} + \frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{\rho}{1-\rho} + \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}\right)^{\frac{1}{2}},$$

$$g'(\rho) = \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{1}{(1-\rho)^2} + (K^{\frac{1}{\alpha}}-1)\cdot\Bigg[(Q_3+\alpha Q_4+2Q_5\rho)\cdot$$
$$\cdot\left(\frac{1}{\alpha(\alpha-2)} + \frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{\rho}{1-\rho} + \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}\right)^{\frac{1}{2}} +$$
$$+\frac{1}{2}\left((Q_1+Q_2\alpha+Q_6\alpha^2)+(Q_3+Q_4\alpha)\rho+Q_5\rho^2\right)\cdot$$
$$\cdot\left(\frac{1}{\alpha(\alpha-2)} + \frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{\rho}{1-\rho} + \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}\right)^{-\frac{1}{2}}\cdot$$
$$\cdot\left(\frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{1}{(1-\rho)^2} + \frac{(\alpha-1)^4}{2\alpha^2(\alpha-2)^2}\frac{\rho}{(1-\rho)^3}\right)\Bigg]. \qquad (31)$$

В следующем восьмом разделе анализируется поведение численного решения уравнения (29) при различном количестве подсистем системы с разделением и параллельным обслуживанием заявок, а также при различных значениях параметра $\alpha > 3$ распределения Парето времени обслуживания.

В последнем разделе главы изучается асимптотика поведения оптимального решения для системы с разделением и параллельным обслуживание с подсистемами $M|Pa|1$.

В **пятой главе** диссертации проводится анализ остаточного времени обслуживания для системы с разделением и параллельным обслуживанием с двумя бесконечнолинейными подсистемами с различными вариантами распределений для времени обслуживания. В системах с разделением и параллельным обслуживанием, как и в любых других системах, могут происходить различного рода сбои, следствием которых может стать отключение и ремонт. Отключение системы может иметь и плановый (профилактический) характер. При этом желательно, если это возможно, чтобы процесс отключение системы происходил корректно. В первую очередь, должен прекратиться прием новых задач, а все задачи, уже имеющиеся в системе на момент запуска процесса отключения, должны быть обязательно обработаны (дообслужены), и только после этого система может полностью завершить свою работу. Таким образом, возникает потребность в изучении остаточного времени

обслуживания в системах с разделением и параллельным обслуживанием, т. е. времени, необходимого для корректного завершения работы системы.

В первом разделе главы подробно описывается математическая модель исследуемой системы с разделением и параллельным обслуживанием. Каждая подсистема обслуживания содержит бесконечное число приборов. Обслуживание в каждой подсистеме имеет произвольное распределение с функцией распределения $B_1(x)$ — на приборах первой подсистемы и $B_2(y)$ — на приборах второй подсистемы (рис. 2).

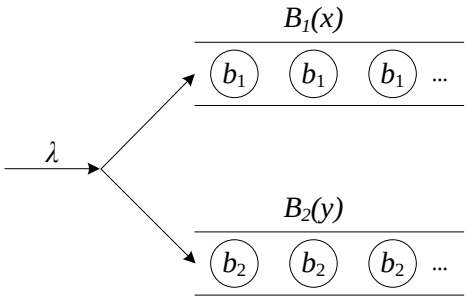

Рис. 2: Схема системы с разделением и параллельным обслуживанием с подсистемами типа $M|G|\infty$.

Сконцентрируемся на изучении максимального остаточного времени обслуживания, под которым будем понимать максимум из остаточных времен обслуживания по всем занятым приборам на момент времени T. Для этого введем двумерную функцию распределения максимумов остаточных времен обслуживания подзаявок, находящихся в двух подсистемах системы с разделением и параллельным обслуживанием, в момент времени T, которую обозначим через $G_T(x, y)$. Основная задача заключается в определении вида данной функции, поскольку этого вполне достаточно для получения основных характеристик (моментов) исследуемой случайной величины.

Кроме того, в силу двумерности рассматриваемых распределений отдельное внимание уделяется вычислению копула-функций и коэффициентов Бломквиста. Коэффициент Бломквиста еще иногда назы-

вают медиальным коэффициентом корреляции. Для его расчета будем пользоваться следующим выражением

$$\beta_C = 4C\left(\frac{1}{2}, \frac{1}{2}\right) - 1, \tag{32}$$

где через $C(\cdot, \cdot)$ обозначена копула.

Коэффициент Бломквиста может обеспечить довольно точное приближение коэффициентов Спирмена и Кендалла, и в целом изучение копул совместно с медиальным коэффициентом корреляции помогает составить некоторое представление о стохастической зависимости, хотя и не всегда может разрешить все вопросы зависимостей.

Во втором разделе рассматривается случай интенсивности входящего потока, не зависящей от времени. Иными словами, имеет место стационарный пуассоновский поток. Оценим двумерную случайную величину максимумов остаточных времен обслуживания в такой системе, а точнее ее функцию распределения $G_T(x, y)$. В результате будет справедлива следующая теорема.

Теорема 5.1. *Для системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром $\lambda = const$ и двумя подсистемами типа $M|G|\infty$ с функцией распределения времени обслуживания на приборах первой подсистемы $B_1(x)$ и на приборах второй подсистемы — $B_2(y)$, выражение для совместной функции распределения максимальных остаточных времен обслуживания на момент времени T, $0 < T < \infty$, имеет вид*

$$G_T(x, y) = \exp\left\{ -\lambda \int_0^T \big(1 - B_1(t + x)B_2(t + y)\big)dt \right\}.$$

Из данной теоремы вытекает множество следствий для различных типов распределения времени обслуживания на приборах: экспоненциального, гиперэкспоненциального, распределения Парето, равномерного и т. д. Так, например.

Следствие 5.3. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с раз-*

делением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ имеют распределение Парето с параметром $\alpha = 2$, т. е.

$$B_1(x) = 1 - (x+1)^{-2}, \quad B_2(y) = 1 - (y+1)^{-2}, \quad x \geqslant 0, \quad y \geqslant 0,$$

то тогда функция распределения максимумов остаточных времен обслуживания на момент времени T определяется как

$$G_T(x,y) = \exp\left\{ -\lambda\left(\frac{(x-y+1)(x-y-1)}{(x-y)^2}\left(\frac{T}{(x+T+1)(x+1)} + \right.\right.\right.$$
$$\left.\left.\left. + \frac{T}{(y+T+1)(y+1)}\right) - \frac{2}{(x-y)^3}\ln\frac{(x+T+1)(y+1)}{(y+T+1)(x+1)}\right)\right\},$$
$$x \geqslant 0, \quad y \geqslant 0, \quad 0 < T \leqslant \infty, \quad x \neq y,$$

а предельные значения функции распределения и копулы, а также коэффициент Бломквиста имеют вид

$$G_\infty(x,y) = \exp\left\{ -\lambda\left(\frac{(x-y+1)(x-y-1)}{(x-y)^2}\left(\frac{1}{x+1} + \frac{1}{y+1}\right) - \right.\right.$$
$$\left.\left. - \frac{2}{(x-y)^3}\ln\frac{y+1}{x+1}\right)\right\}, \quad x \geqslant 0, \quad y \geqslant 0, \quad x \neq y,$$

$$C_\infty(u,v) = uv\exp\left\{\frac{\ln^2 u \ln^2 v\left(2\ln u \ln v \ln(\frac{\ln u}{\ln v}) - \ln(uv)\ln(\frac{u}{v})\right)}{\lambda^2(\ln u - \ln v)^3}\right\}, \quad u \neq v,$$

$$\beta_C = 2^{\frac{\ln^2 2}{3\lambda^2}} - 1.$$

Также сформулированы предельная теорема общего характера в условиях большой загрузки и теорема для случая бесконечного среднего времени обслуживания (при степенных хвостах).

В третьем разделе исследуется случай интенсивности, заданной функцией от времени, когда интенсивность входящего потока задана ограниченной неотрицательной функцией $\lambda(t)$, $t \geqslant 0$, т. е. имеет место нестационарный пуассоновский поток. Тогда будет справедлива теорема, аналогичная первой теореме из второго раздела, касающая-

ся общего вида выражения для функции распределения максимумов остаточных времен обслуживания на момент времени T. Представлено следствие из теоремы для случая показательного распределения времени обслуживания на приборах.

В четвертом разделе проанализирована ситуация, когда интенсивность входящего потока задана случайным процессом $\lambda(t)$, $t \geqslant 0$, т. е. имеет место дважды стохастический пуассоновский поток. В этом случае также сформулирована теорема, касающаяся общего вида выражения для функции распределения максимумов остаточных времен обслуживания на момент времени T, а также следствие для случая показательного распределения времени обслуживания на приборах. Кроме того, имеет место следующая теорема.

Теорема 5.6. *Если для системы с разделением и параллельным обслуживанием с дважды стохастическим пуассоновским входящим потоком с интенсивностью $\lambda = \lambda(t)$ и двумя подсистемами с бесконечным числом приборов и с функцией распределения времени обслуживания на приборах первой подсистемы $B_1(x)$, а на приборах второй подсистемы — $B_2(y)$, выполняется, что интенсивность имеет вид*

$$\lambda(t) = \max\{\gamma(t), 0\}, \quad \gamma(t) = \lambda_0 + \sigma\xi(t), \quad \lambda_0, \sigma > 0, \tag{33}$$

где $\xi(t), t \geqslant 0$, — стационарный гауссовский процесс с нулевым математическим ожиданием и ковариационной функцией $R(t), t \geqslant 0; R(0) = 1$, тогда для совместной функции распределения максимальных остаточных времен обслуживания на момент времени T, $0 < T < \infty$ будет справедливо неравенство

$$G_T(x,y) \leqslant \exp\Bigg\{ -\lambda_0 \int_0^T \big(1 - B_1(t+x)B_2(t+y)\big)dt + \\ + \frac{\sigma^2}{2} \int_0^T \int_0^T R(u-v)\big(1 - B_1(u+x)B_2(u+y)\big)\cdot \\ \cdot \big(1 - B_1(v+x)B_2(v+y)\big)dudv \Bigg\}. \tag{34}$$

Стоит отметить, что точное описание интенсивности гауссовским

процессом невозможно, поскольку тогда интенсивность должна принимать иногда отрицательные значения. Однако вероятность таких значений и их вклад в результат при $\sigma \ll \lambda_0$ будут очень малы, а значит, G_T должна быть близка к указанной верхней границе

Среди полученных результатов стоит выделить явление зависимости между распределением времени обслуживания и скоростью убывания коэффициента Бломквиста с ростом интенсивности входного потока (или загрузки с учетом среднего времени обслуживания).

Зависимость между максимальными остаточными временами в подсистемах порождается синхронным поступлением туда подзаявок. Так, при постоянном времени обслуживания возникает совершенная зависимость (идентичность). Напротив, различия в состояниях подсистем порождаются разнообразием во временах обслуживания подзаявок. Чем это разнообразие больше (например, в смысле тяжести хвоста), тем меньше должна быть зависимость. При бесконечном среднем времени обслуживания для этого даже высокая загрузка не нужна, максимумы оказываются асимптотически независимыми.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИОННОЙ РАБОТЫ

В диссертации в рамках решения фундаментальной научной проблемы по исследованию стохастических систем с разделением и параллельным обслуживанием разработаны вероятностные модели, методы и алгоритмы анализа, обработки данных и управления для систем с разделением и параллельным обслуживанием.

1. Разработан комплекс методов и алгоритмов оценки основных характеристик времени отклика систем с разделением и параллельным обслуживанием: моментов и квантилей его распределения, а также коэффициентов корреляции между временами пребывания в подсистемах.

2. Разработан метод определения оптимальной интенсивности обслуживания в системах с разделением и параллельным обслужи-

ванием в зависимости от интенсивности входящего потока.

3. Разработан метод определения характеристик остаточного времени обслуживания, т. е. времени, необходимого для корректного завершения работы системы после отключения входящего потока, в системах с разделением и параллельным обслуживанием.

4. Продемонстрирована применимость предложенных методов на примерах конкретных типов распределений для входящего и обслуживающего потоков.

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

В изданиях из списка ВАК РФ и приравненных к ним

1. *Горбунова А.В.* Нелинейная аппроксимация квантилей распределения времени отклика fork-join системы массового обслуживания с подсистемами M|M|1 // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2025. № 72. С. 16–27. **ВАК (К1)**.

2. *Горбунова А.В.* Об особенностях управления скоростью обслуживания в fork-join системах с распределением Парето времени обслуживания // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2024. № 4. С. 53–62. **ВАК (К1)**.

3. *Горбунова А.В., Лебедев А.В.* Квантили распределения времени отклика в fork-join системах с распределением Парето времени обслуживания // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2024. № 3. С. 5–16. **ВАК (К1)**.

4. *Горбунова А.В.* Оценки копулы и квантилей распределения времени отклика системы с разделением и параллельным обслужи-

ванием заявок и распределением Парето времени обслуживания // Управление большими системами: сборник трудов. 2024. № 112. С. 7–29. **ВАК (К1)**.

5. *Горбунова А.В., Лебедев А.В.* О новом подходе к оценке квантилей времени отклика системы с разделением и параллельным обслуживанием заявок // Управление большими системами: сборник трудов. 2024. № 108. С. 6–21. **ВАК (К1)**.

6. *Вишневский В.М., Горбунова А.В.* Применение методов машинного обучения к решению задач теории массового обслуживания // Информационные технологии и вычислительные системы. 2021. № 4. С. 70–82. **ВАК (К1)**.

7. *Gorbunova A.V., Lebedev A.V.* On the Features of Service Rate Control in Fork-Join Queueing System // Automation and Remote Control. 2024. Vol. 85, Issue 12. P. 1184–1198. **Scopus (Q3)**.

   Горбунова А.В., Лебедев А.В. Об особенностях управления скоростью обслуживания в системах с разделением и параллельным обслуживанием заявок // Автоматика и телемеханика. 2024. № 12. С. 70–88.

8. *Gorbunova A.V., Lebedev A.V.* Correlations of the Sojourn Times of Subtasks in Fork-Join Queueing Systems with $M|M|1$-type Subsystems // Advances in Systems Science and Applications. 2024. Vol. 24. No. 2. P. 1–18. **Scopus (Q3)**.

9. *Gorbunova A.V., Lebedev A.V.* Copulas and Quantiles in Fork-Join Queueing Systems // Advances in Systems Science and Applications. 2024. Vol. 24, No. 1. P. 1–19. **Scopus (Q3)**.

10. *Gorbunova A.V., Lebedev A.V.* Nonlinear Approximation of Characteristics of a Fork-Join Queueing System with Pareto Service as a Model of Parallel Structure of Data Processing // Mathematics and Computers in Simulation. 2023. Vol. 214. P. 409–428. **Scopus (Q1)**.

11. *Gorbunova A.V., Lebedev A.V.* On Estimating the Characteristics of a Fork-Join Queueing System with Poisson Input and Exponential Service Times // Advances in Systems Science and Applications. 2023. Vol. 23, No. 2. P. 99–114. **Scopus (Q3)**.

12. *Горбунова А.В., Вишневский В.М.* Оценка времени отклика среды для вычислений с интенсивным использованием данных // Информационно-управляющие системы. 2022. № 4. С. 12–19. **Scopus (Q3)**.

13. *Gorbunova A.V., Vishnevsky V.M.* The Analysis of Big Data Centers Performance // Advances in Systems Science and Applications. 2022. Vol. 22, No. 3. P. 70–83. **Scopus (Q3)**.

14. *Gorbunova A.V., Lebedev A.V.* Bivariate Distributions of Maximum Remaining Service Times in Fork-Join Infinite-Server Queues // Problems of Information Transmission. 2020. Vol. 56, No. 1. P. 73–90. **Scopus (Q2)**.

    Горбунова А.В., Лебедев А.В. Двумерные распределения максимальных остаточных времен обслуживания в бесконечнолинейных системах с разделением заявок // Проблемы передачи информации. 2020. Т. 56, Вып. 1. С. 80–98.

15. *Gorbunova A.V., Vishnevsky V.M.* Estimating the Response Time of a Cloud Computing System with the Help of Neural Networks // Advances in Systems Science and Applications. 2020. Vol. 20, No. 3, P. 105–112. **Scopus (Q3)**.

Публикации в сборниках, индексируемых в Scopus

16. *Gorbunova A.V., Vishnevsky V.M.* On Estimating the Average Response Time of High-Performance Computing Environments // Lecture Notes in Computer Science. 2023. Vol. 13766. P. 371–384.

17. *Gorbunova A.V., Lebedev A.V.* Response Time Estimate for a Fork-Join System with Pareto Distributed Service Time as a

Model of a Cloud Computing System Using Neural Networks // Communications in Computer and Information Science. 2022. Vol. 1552. P. 318–332.

18. *Vishnevsky V.M., Gorbunova A.V.* Application of Machine Learning Methods to Solving Problems of Queuing Theory // Communications in Computer and Information Science. 2022. Vol. 1605. P. 304–316.

19. *Gorbunova A.V., Vishnevsky V.M.* Evaluation of the Performance Parameters of a Closed Queuing Network Using Artificial Neural Networks // Lecture Notes in Computer Science. 2021. Vol. 13144. P. 265-278.

20. *Gorbunova A.V., Vishnevsky V.M., Larionov A.A.* Evaluation of the End-to-End Delay of a Multiphase Queuing System Using Artificial Neural Networks // Lecture Notes in Computer Science. 2020. Vol. 12563. P. 631–642.

Научное издание

Горбунова Анастасия Владимировна

МЕТОДЫ И АЛГОРИТМЫ АНАЛИЗА И УПРАВЛЕНИЯ ДЛЯ СТОХАСТИЧЕСКИХ СИСТЕМ С РАЗДЕЛЕНИЕМ И ПАРАЛЛЕЛЬНЫМ ОБСЛУЖИВАНИЕМ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора физико-математических наук

Подписано в печать ___.___.2026. Формат $60 \times 90/16$.
Тираж ___ экз. Заказ № ___.

Федеральное государственное бюджетное учреждение науки
Институт проблем управления им. В.А. Трапезникова
Российской академии наук
117342, г. Москва, вн. тер. г. муниципальный округ Коньково,
ул. Профсоюзная, д. 65, стр. 2
www.ipu.ru