ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ
НАУКИ ИНСТИТУТ ПРОБЛЕМ УПРАВЛЕНИЯ
ИМ. В. А. ТРАПЕЗНИКОВА РОССИЙСКОЙ АКАДЕМИИ НАУК

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

ДИССЕРТАЦИЯ

на соискание ученой степени

доктора физико-математических наук

Научный консультант:

д.ф.-м.н.

Лебедев Алексей Викторович

Москва — 2025

Содержание

Введение **5**

1 Методы машинного обучения и интеллектуального анализа для сложных систем массового обслуживания **23**

1.1 Искусственные нейронные сети и ТМО 25

1.2 Применение методов машинного обучения для оценки характеристик различных систем/сетей массового обслуживания 31

1.3 Применение методов машинного обучения для оценки характеристик моделей систем массового обслуживания 38

1.4 Примеры 49

1.4.1 Применение методов машинного обучения для оценки характеристик замкнутых экспоненциальных сетей 50

1.4.2 Применение методов машинного обучения для оценки характеристик открытых неэкспоненциальных сетей 65

1.5 Выводы к главе 1 80

2 Получение основных стационарных характеристик систем с разделением и параллельным обслуживанием в случае подсистем с пуассоновским входящим потоком и экспоненциальным распределением времени обслуживания **81**

2.1 Математическая модель системы с разделением и параллельным обслуживанием 82

2.2 Оценка основных характеристик системы с помощью ИНС 86

2.3 Оценка основных характеристик системы с помощью комплексного метода 89

2.4 Коэффициенты корреляции в системе с разделением и параллельным обслуживанием 99

2.5 Мета-гауссовская модель для времени отклика системы с разделением и параллельным обслуживанием 120
2.6 Копулы и квантили в fork-join системах массового обслуживания с подсистемами типа $M|M|1$ 125
2.7 Об особенностях имитационного моделирования системы 148
2.8 Выводы к главе 2 156

3 Получение основных стационарных характеристик систем с разделением и параллельным обслуживанием в случае распределения Парето времени обслуживания 159

3.1 Математическая модель системы с разделением и параллельным обслуживанием 159
3.2 Аналитические оценки времени отклика в fork-join системе 161
3.3 Оценка основных характеристик системы с помощью ИНС 168
3.4 Оценка основных характеристик системы с помощью комплексного метода 177
3.5 Оценка коэффициентов корреляции 188
3.6 Квантили распределения времени отклика в fork-join системах с распределением Парето времени обслуживания 196
3.7 О еще одном методе оценки квантилей и копулы распределения времени отклика 216
3.8 Выводы к главе 3 231

4 Об особенностях управления интенсивностью обслуживания в системах с разделением и параллельным обслуживанием 234

4.1 Математическая модель определения стоимости функционирования fork-join системы с экспоненциальным распределением времени обслуживания 235
4.2 Анализ fork-join СМО с двумя подсистемами типа $M|M|1$. Формула Нельсона–Тантави 237

4.3 Анализ fork-join СМО с $K > 2$ подсистемами типа $M|M|1$. Формула Нельсона–Тантави 246
4.4 Анализ fork-join СМО с $K > 2$ подсистемами типа $M|M|1$. Обобщение формулы Нельсона–Тантави 249
4.5 Асимптотика поведения оптимального решения для системы с разделением и параллельным обслуживанием с подсистемами типа $M|M|1$ 252
4.6 Математическая модель определения стоимости функционирования fork-join системы с распределением Парето времени обслуживания 258
4.7 Анализ fork-join СМО с $K \geqslant 2$ подсистемами типа $M|Pa|1$ 260
4.8 Численный эксперимент для fork-join СМО с распределением Парето времени обслуживания 263
4.9 Асимптотика поведения оптимального решения для системы с разделением и параллельным обслуживание с подсистемами $M|Pa|1$ 266
4.10 Выводы к главе 4 271

5 Остаточное время обслуживания в системе с разделением и параллельным обслуживанием с пуассоновским входящим потоком и произвольным распределением для времени обслуживания 273

5.1 Математическая модель исследуемой системы с разделением и параллельным обслуживанием 274
5.2 Случай интенсивности входящего потока, не зависящей от времени 277
5.3 Случай интенсивности, заданной функцией от времени 305
5.4 Случай интенсивности, заданной случайным процессом 307
5.5 Выводы к главе 5 311

Введение

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. К наиболее ранним публикациям в области данной проблематики, где были получены первые точные результаты для среднего времени отклика, относится статья R. Nelson и A.N. Tantawi 1988 года [163]. 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 системы 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.

Описанный механизм функционирования подходит для моделирования многих реальных процессов. Так, в области здравоохранения используют такую СМО для прогнозирования времени пребывания пациентов в медицинском учреждении: как правило, при поступлении перед осмотром пациента у врача должны быть готовы результаты разного рода медицинских тестов/анализов, которые могут проводиться в удаленных друг от друга лабораторных подразделениях. Аналогичная ситуация складывается и при выписке пациентов, кроме того, могут возникать дополнительные задержки, связанные с транспортировкой сложных пациентов или необходимых лекарств. Среди публикаций, посвященных анализу подобных задержек в сфере здравоохранения с использованием СМО с разделением и параллельным обслуживанием, можно перечислить работы [62, 63, 127].

В банковском секторе структура с разделением и параллельным обслуживанием также может использоваться для определения времени, затрачиваемого сотрудниками разных отделов, на обработку запросов, поступающих от клиентов. Например, одобрение ипотеки может потребовать согласования у сотрудников нескольких подразделений банка одновременно, в частности, кредитного отдела и отдела страхования жизни [133]. Множество статей посвящено мо-

делированию производственных процессов с использованием структур с разделением и параллельным обслуживанием в качестве ключевого или одного из составляющих его элементов [160]. Еще в [67] параллельная структура рассматривалась как адекватная модель для производственной линии по изготовлению многокомпонентных изделий. Каждый станок изготавливает определенную деталь, и изделие не будет готово до тех пор, пока не будут произведены все составляющие его элементы. Так, в статье [98] анализируется процесс производства электробытовой техники, большая часть которого приходится на закупку множества необходимых комплектующих у различных поставщиков. Заказ клиента инициирует процесс сборки заказа на складе, проводится анализ стратегий на основе знаний о распределении времени поставки компонентов. Более поздние работы учитывают неизбежные изменения в технологических процессах, связанные с прогрессом, однако система с разделением и параллельным обслуживанием как основа производственных моделей продолжает оставаться актуальной [175]. В [175] анализируется масштабные производственные процессы, которые возникают в таких крупнейших компаниях, как Airbus и ASML.

Отдельного упоминания заслуживают работы, посвященные моделированию сетей связи, вычислительных систем, анализу больших данных и их составных частей на основе параллельных структур (MapReduce, многопутевая маршрутизация, приложения с интенсивным использованием данных), например, [164, 174, 196, 207]. В области телекоммуникаций, системы с параллельным обслуживанием хорошо описывают процесс передачи дейтаграмм, которые являются частью одного сообщения. Параллельные и распределенные вычисления значительно увеличивают скорость работы приложений с интенсивным использованием данных и не только, а процессы, описывающие функционирование подобных систем также моделируются СМО с разделением и параллельным обслуживанием.

Что касается теоретических результатов исследования систем то, как уже упоминалось выше, в статье [163] было получено точное выражение для средне-

го времени отклика СМО с двумя подсистемами типа $M|M|1$, а также представлена оценка среднего времени отклика системы для произвольного количества подсистем $K > 2$. В статье [194] представлена еще одна приближенная формула, которая, как и в предыдущем случае, была получена благодаря сочетанию аналитического подхода с эмпирическим, т. е. была проведена коррекция аналитических выражений на основе данных имитационного моделирования. С некоторыми аспектами обработки больших потоков экспериментальных данных можно ознакомится, например, в [44]. В работе [195] приведена оценка для среднего времени отклика, полученная методом интерполяции системы в условиях предельно высокой и слабой входных нагрузок. Также для анализа СМО с разделением и параллельным обслуживанием применяется матрично-геометрический метод, причем не только в случае с экспоненциальным распределением, но и с распределением Кокса времени обслуживания, были получены верхние и нижние границы для моментов времени отклика [71]. Для анализа СМО с произвольным распределением длительностей обслуживания довольно часто используются элементы теории порядковых статистик [88], т. к. время отклика является максимумом из K случайных величин времен пребывания подзаявок в соответствующих подсистемах и, соответственно, представляет собой K-ую порядковую статистику [95, 146]. Однако при этом делается упрощающее допущение о независимости времен пребывания в подсистемах, хотя в реальности эти величины являются коррелированными. Подробный обзор методов исследования систем с разделением и параллельным обслуживанием до 2014 года можно найти в [191].

В последние годы активность исследований СМО с разделением и параллельным обслуживанием несколько снизилась, однако новые результаты все же появляются в силу актуальности параллельных структур в математических моделях многих реально функционирующих систем. Так, среди более свежих публикаций стоит отметить следующие. В [168] исследуется СМО с подсистемами типа $MAP|PH|1$, на основе аналитических результатов для случая с двумя

подсистемами строится оценка для среднего времени отклика и для процентелей времени отклика высокого уровня вплоть до 99-го процентиля, при этом погрешность аппроксимации полученных оценок остается относительно низкой даже для большого числа подсистем. В [165] исследуется система с разделением и параллельным обслуживанием более общего вида, в которой подсистемы могут представлять собой СМО типа $M|G|1$ или $G|G|1$. Поскольку модель с параллельным обслуживанием заявок используется для моделирования технических систем, в работе предлагается два подхода к их анализу, а именно как к модели “белого ящика”, когда известны распределения для входящего и обслуживающих потоков, и как к модели “черного ящика”, когда данные распределения неизвестны, однако технически возможно оценить средние характеристики подсистем. В результате получены оценки для среднего времени отклика и для процентилей высокого уровня, которые показывают неплохое качество приближения при высоких значениях загрузки системы. Кроме того, с помощью представленных в работе методов рассматриваются системы с разделением и параллельным обслуживанием с отличным от классического механизмом функционирования, например, когда заявка разделяется на подзаявки, количество которых вариативно и может быть меньше числа приборов в системе. В [94] разработан метод для оценки среднего значения и дисперсии времени отклика СМО с разделением и параллельным обслуживанием, который подходит для оценки параметров производительности системы для многих случаев распределения времени обслуживания на приборах, в том числе и для распределений с тяжелыми хвостами.

Каждый из известных подходов к приближенному анализу систем с разделением и параллельным обслуживанием имеет определенные достоинства и недостатки, обычно связанные с качеством аппроксимации и применимостью подхода для определенного набора распределений, входных параметров. Так, например, несмотря на приведенное разнообразие примеров распределений, для которых возможен метод анализа, предложенный в [165], он имеет ограничения

связанные с тем, что справедлив только в области высоких значений нагрузки и его погрешность может колебаться в пределах до 20%.

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 [12, 27, 30, 34, 41, 42, 45–47, 51, 197]. В серии статей, включающей [27, 30, 41, 42], проведен анализ системы с бесконечнолинейными подсистемами с входящим пуассоновским или MAP-потоком, постановка задач для подобных систем и методы их анализа имеют свою специфику. В [45] проанализирована одна из модификаций системы, когда при поступлении заявка разделяется не на фиксированное, а на переменное число подзаявок, которое определяется состоянием системы. В [12, 34, 197] получен аналитический результат в форме выражений для верхней и нижней границы (или алгоритма вычисления границ) среднего времени отклика для частного случая системы с двумя подсистемами, входящим пуассоновским потоком (или MAP-потоком), фазовым распределением времен обслуживания и конечным размером очереди одной из подсистем. В [51] предложен подход на основе инвариантов отношения для приближения среднего времени отклика системы с пуассоновским входящим потоком. В [46, 47] система с разделением и параллельным обслуживанием является составной частью сети с пуассоновским входящим потоком и экспоненциальным распределением времен обслуживания на приборах и некоторыми индивидуальными особенностями архитектуры: в [47] предложенная сеть используется для моделирования и исследования характеристик производительности сервисных платформ транзакционных услуг, а в [46] — для моделирования распределенной системы обработки и передачи данных.

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 растёт, 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, обработки данных, а также управления для стохастических систем с разделением и параллельным обслуживанием, улучшающих качество известных результатов, а также позволяющих оценить ранее не изученные характеристики, например, остаточное время обслуживания, является актуальной задачей в рамках системного анализа.

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 анализ 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 математическое 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 [9, 16–18, 21, 25, 26, 100–111, 198], 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 337 страницах, содержит 33 таблицы и 110 иллюстраций. Библиография включает 210 наименований.

Содержание работы. Глава 1. В главе 1 приводится обзор известных работ, связанных в той или иной степени с применением методов машинного обучения при исследовании моделей систем и сетей массового обслуживания различной природы. Обоснована идея применимости различных методов машинного обучения и искусственных нейронных сетей, в частности, к анализу

сложных стохастических систем. Сформулированы основные положения нового метода, а также сам алгоритм анализа стохастических систем. Приводятся численные примеры.

Глава 2. Во второй главе исследуется система с разделением и параллельным обслуживанием заявок с пуассоновским входящим потоком и экспоненциальным распределением времени обслуживания на приборах в ее подсистемах. Проводится оценка основных вероятностно-временных показателей производительности системы (математического ожидания и дисперсии времени отклика) с помощью искусственных нейронных сетей, а также с помощью комплексного подхода с использованием нескольких методов интеллектуального анализа данных, получены аналитические приближения характеристик. Получено точное выражения для коэффициентов корреляции Пирсона и Спирмена с помощью метода производящих функций, а также преобразования Лапласа–Стилтьеса. С помощью комплексного подхода получена приближенная формула для оценки коэффициента корреляции Кендалла. Предложена мета-гауссовская модель для определения среднего времени отклика, использующая полученную формулу для расчета коэффициента корреляции Спирмена. На основе элементов теории копул строится оценка квантилей распределения времени отклика системы с разделением и параллельным обслуживанием с двумя подсистемами. Представлен алгоритм расчета доверительных интервалов для оценки математического ожидания времени отклика системы, полученной с помощью имитационного моделирования системы.

Глава 3. В третьей главе проводится анализ характеристик системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком и распределением Парето времени обслуживания на приборах в ее подсистемах. Получены аналитические оценки для верхней границы среднего времени отклика на основе элементов теории порядковых статистик и преобразования Лапласа–Стилтьеса. Проводится оценка математического ожидания и дисперсии времени отклика системы с помощью искусственных нейронных сетей. Так-

же с помощью комплексного подхода с использованием нескольких методов интеллектуального анализа данных, получены аналитические выражения для среднего и дисперсии времени отклика, а также коэффициентов корреляции между временами пребывания подзаявок в подсистемах системы с разделением и параллельным обслуживанием. Предложен метод оценки квантилей распределения времени отклика посредством аппроксимации его распределения распределением Фреше, а также метод оценки квантилей с помощью элементов теории копул.

Глава 4. В четвертой главе строится общая модель управления интенсивностью обслуживания в системах с разделением и параллельным обслуживанием, в основе которой лежит оптимизация ее финансовых показателей и, соответственно, баланс между временем отклика и необходимыми ресурсами для поддержания оптимального уровня загрузки системы при изменении интенсивности входящего потока заявок. Для системы с пуассоновским входящим потоком и экспоненциальным распределением времени обслуживания на приборах в частном случае двух подсистем получено соотношение, определяющее оптимальное значение интенсивности обслуживания в аналитическом виде. Когда число подсистем больше двух, и для случая экспоненциального распределения времени обслуживания, и для случая распределения Парето времени обслуживания получены уравнения, численное решение которых определяет оптимальное значение интенсивности обслуживания. Кроме того, проведен асимптотический анализ полученных решений, его результаты могут использоваться для управления режимом функционирования системы на границах допустимых значений параметров, от которых зависит ее поведение.

Глава 5. 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. Под остаточным временем обслуживания понимается время, необходимое для корректного завершения работы системы.

Всевозможные сбои в системах должны эффективно обрабатываться надлежащими механизмами отказоустойчивости, поэтому время, необходимое для корректного завершения работы системы после отключения входящего потока заявок является одним из важнейших показателей, используемых при разработке эффективных методов повышения работоспособности распределенных систем. Получены аналитические выражения для совместной функции распределения максимальных остаточных времен обслуживания на некоторый момент времени. Для отдельных распределений найдены копула-функции и коэффициенты Бломквиста.

1 Методы машинного обучения и интеллектуального анализа для сложных систем массового обслуживания

В диссертации предлагается новый метод анализа систем с разделением и параллельным обслуживанием заявок, который основывается в том числе на применении интеллектуального анализа данных. Концепция метода подразумевает, что он может быть использован не только для анализа систем массового обслуживания типа fork-join, но и для других видов систем или сетей. Поэтому в рамках данной главы будет приведен обзор результатов применения машинного обучения и, в частности, искусственных нейронных сетей для исследования различных моделей массового обслуживания, сформулирована концепция, важные положения и этапы нового подхода.

Рассмотренные публикации делятся на несколько категорий — статьи, в которых машинные алгоритмы служат для прогнозирования интересующих параметров реальных систем массового обслуживания 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 авторами представленных в обзоре статей для анализа интересующих систем — их архитектура, входные и выходные параметры, а также алгоритмы обучения.

Хотя идея нового метода возникла относительно недавно, в настоящей главе рассматриваются и ранние публикации, в которых данная тематика была затронута лишь косвенно. Тем не менее, указанные публикации заслуживают внимания, поскольку позволяют проследить развитие новой методики от ее исходной точки до современного состояния. При этом четкая формулировка метода применительно к плоскости теории массового обслуживания (ТМО) и его

апробация на различных примерах моделей массового обслуживания отражены преимущественно в работах автора данной диссертации, например [109, 198].

Глава 1 написана на основе успешного применения нового метода к анализу различных структур систем с разделением и параллельным обслуживанием, а также сетей массового обслуживания и, соответственно, полученных результатов и сделанных на их базе выводов, в работах [9, 10, 21, 108, 109, 111, 198]. С деталями проведенного анализа для случая замкнутых экспоненциальных сетей и случая открытых неэкспоненциальных сетей можно подробно ознакомиться в разделе 1.4 данной главы.

Итак, в настоящее время известны три основные группы методов решения задач теории массового обслуживания: аналитические методы, численные методы и имитационное моделирование.

Нахождение основных показателей производительности систем массового обслуживания (СМО) не всегда является возможным аналитически. Это объясняется ограничениями, накладываемыми распределениями, характеризующими входящий поток и время обслуживания заявок, например такими, которые нельзя отнести к распределениям фазового типа, либо в случаях, когда речь идет о многофазных системах с непуассоновским входящим потоком. Численные же методы анализа математических моделей массового обслуживания [120], к примеру, итерационные или матрично-геометрические, могут оказаться довольно ресурсоемкими и энергозатратными, даже несмотря на существование упрощающих методов для некоторых типов СМО, целью разработки которых и являлось снижение трудоемкости матричных алгоритмов.

Еще одним методом, применяемым с целью определения различных показателей функционирования систем массового обслуживания, является имитационное моделирование. Оно позволяет решать достаточно широкий спектр реальных задач, в том числе и на этапе проектирования физических систем с целью их оптимизации, что может позволить при этом еще и значительно сэкономить, поскольку в этом случае не требуется создание прототипов.

Теперь более подробно остановимся еще на одной методике решения задач ТМО, а именно методах машинного обучения и в частности, нейронных сетях. 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 СМО.

1.1 Искусственные нейронные сети и ТМО

Идея применения искусственных нейронных сетей (ИНС) для исследований СМО в сущности обуславливается задачами, которые могут быть решены с их помощью. Вообще говоря, 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. В последующем нас будет интересовать применение искусственных нейронных сетей и других методов машинного обучения именно

в контексте решения данной проблемы, поскольку, например, нейросети считаются одним из лучших инструментов аппроксимации функций [15, 178].

Прежде чем непосредственно переходить к заявленной проблематике, немного остановимся на архитектуре искусственных нейронных сетей. Итак, в настоящий момент известно несколько типов структур ИНС, однако в большинстве исследований по тематике ТМО и не только используется персептрон, характеризуемый одним или несколькими скрытыми слоями и прямым распространением сигнала (рис. 1).

На вход персептрона поступает вектор $\mathbf{x} = (x_1, \ldots, x_n)$, где n — это число нейронов входного слоя. Выходным значением нейросети будет являться вектор $\mathbf{y} = f(\mathbf{x}) = (y_1, \ldots y_m)$, где m — это число нейронов выходного слоя. Если персептрон содержит только один скрытый слой, то элементы выходного слоя будут определяться выражением

$$y_i = \phi_i^{(1)}\left(\sum_{j=0}^{l} w_{ji}^{(2)}\phi_j^{(1)}\Big(\sum_{k=0}^{n} w_{kj}^{(1)}x_k\Big)\right), \quad i = \overline{1, m}, \tag{1.1}$$

где $x_0 = \phi_0^{(1)} = 1$ выступают в роли коэффициентов смещения для скрытого слоя, l — это число нейронов в скрытом слое, $w_{kj}^{(1)}$ и $w_{ji}^{(2)}$ — весовые коэффициенты, а $\phi_j^{(1)}(\cdot)$ и $\phi_i^{(2)}(\cdot)$ — функции активации.

Как видно из формулы, на каждом нейроне происходит суммирование произведений входных данных и соответствующих весовых коэффициентов, которые после прохождения через активационную функцию в качестве аргумента становятся потом входными данными для нейронов следующего слоя.

Наиболее распространенными активационными функциями являются пороговая

$$\phi(x) = \begin{cases} 0 & \text{при } x \geqslant 1, \\ 1 & \text{при } x < 1, \end{cases} \tag{1.2}$$

логистическая

$$\phi(x) = \frac{1}{1 + e^{-x}}, \tag{1.3}$$

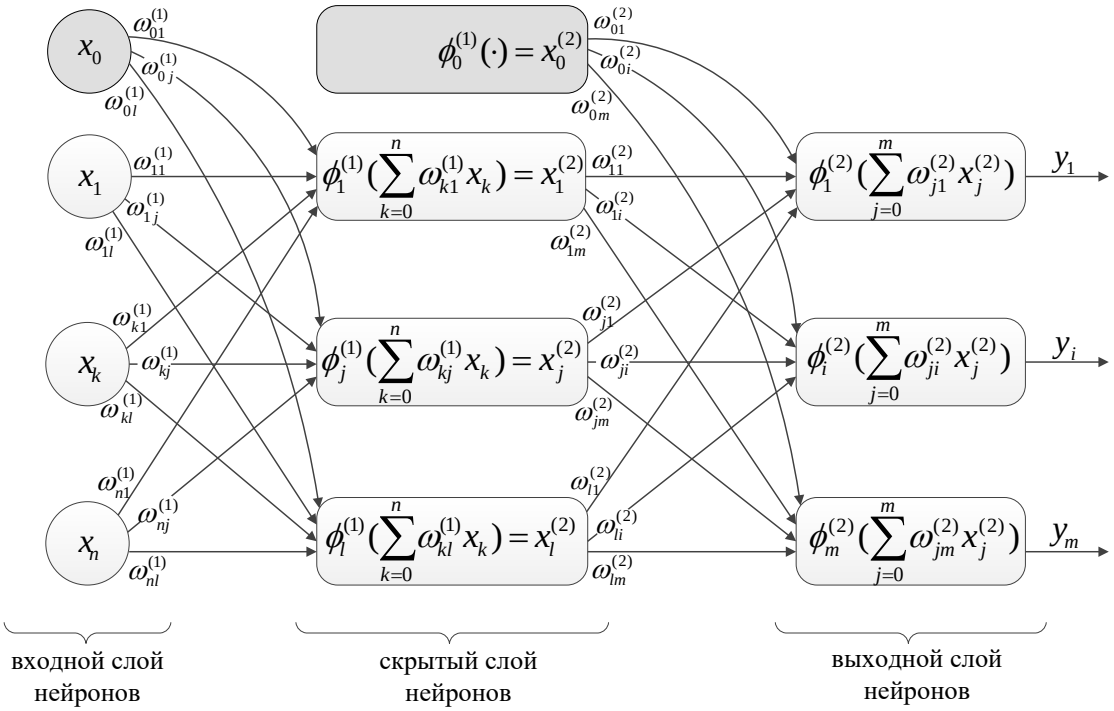


Рис. 1: Двухслойный персептрон.

и гиперболический тангенс

$$\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{1.4}$$

Как правило, в качестве основной активационной функции выбирают именно сигмоидальные функции, к семейству которых относятся две последние из перечисленных. Класс этих функций характеризуется тем, что

$$\lim_{x\to-\infty} \phi(x) = 0, \quad \lim_{x\to+\infty} \phi(x) = 1 \tag{1.5}$$

либо

$$\lim_{x\to-\infty} \phi(x) = -1, \quad \lim_{x\to+\infty} \phi(x) = 1, \tag{1.6}$$

т. е. график по своей форме похож на букву S в случае, если для функции выполняется условие монотонности, что не всегда обязательно [178].

Теперь скажем несколько слов об аппроксимационных свойствах нейронных сетей. Дело в том, что согласно теоремам об аппроксимации функций, приведенных в [15,187], а также в обзоре [178], любую непрерывную функцию можно

представить в виде комбинации линейных операций и единственного нелинейного элемента, что позволяет их применить к многослойному персептрону, где в качестве нелинейного элемента выступает функция активации.

Впервые возможности аппроксимации непрерывной функции произвольной линейности со значениями на единичном гиперкубе с помощью персептрона были продемонстрированы в работе [85], причем было строго доказано, что для этого достаточно только одного скрытого слоя. При этом вопрос об оптимальном количестве нейронов в этом скрытом слое остается открытым и в большинстве случаев решается экспериментальным путем, хотя на этот счет и существуют некоторые рекомендации [176].

Кроме того, несмотря на достаточность единственного скрытого слоя для решения задачи аппроксимации, в некоторых сложных случаях продуктивнее будет использовать архитектуру нейросети с двумя или даже тремя скрытыми слоями с целью увеличения рычагов управления процессом обучения [81, 190].

Заметим, что проблема с выбором структуры нейросети (количество скрытых слоев, количество нейронов в них, а также число связей для каждого нейрона) или, другими словами, ее оптимизация является, пожалуй, одной из самых серьезных, поскольку тесно связана с эффектом переобучения или недообучения сети, что, очевидно, является одним из ключевых критических факторов при работе с ИНС.

Также стоит отметить, что существует множество подходов к обучению нейронных сетей, которые, в свою очередь, оказывают влияние на процесс аппроксимации. Под обучением понимают процесс нахождения оптимальных весовых коэффициентов, которые бы в полной мере отражали зависимость между входными значениями и выходом нейросети. Одним из самых известных и распространенных методов обучения является метод обратного распространения ошибки (англ. backpropagation, BP). Он имеет свои достоинства и недостатки. К основным проблемам этого алгоритма можно отнести сложности, связанные с его сходимостью, а также с переобучением и даже параличом сети, при ре-

шении непростых и трудоемких задач. Улучшенной версией этого алгоритма является метод упругого распространения ошибки (англ. resilient propagation, RProp), преимущество которого заключается в большей надежности работы и в необходимости меньшего количества эпох обучения, что значительно увеличивает скорость обучения.

Среди других широко применяемых методов обучения ИНС, в том числе и для решения задач ТМО, можно назвать алгоритм обучения Левенберга-Маркарта (англ. Levenberg–Marquardt algorithm, LMA или LM)) [117], метод шкалированных связанных градиентов (англ. scaled conjugate gradient algorithm) [157] и усовершенствованная версия метода Левенберга-Маркарта — метод байесовской регуляризации [97]. Вопрос выбора определенного алгоритма из всего множества существующих, которые, разумеется приведены здесь не все, можно разрешить только экспериментальным путем при реализации конкретной задачи. Руководствоваться при этом можно, например, оптимальным соотношением таких характеристик как точность и трудоемкость выбранных для сравнения алгоритмов [52] либо специальными метриками для машинных алгоритмов, как, например, в [159].

Как уже упоминалось выше, первые публикации, косвенно затрагивающие применение нейросетей к анализу моделей массового обслуживания, появились относительно давно, тем не менее, были достаточно разрозненными и не обратили на себя должного внимания. Так, к одному из первых упоминаний применения методов машинного обучения, в частности алгоритма построения дерева решений ID3 (англ. iterative dichotomiser 3), к решению задач ТМО можно отнести работу [135]. Здесь говорится в общем о потенциале методов машинного обучения, а также обсуждаются новые, благодаря этим методам, возможности моделирования поведения сложных систем и преимущества их использования при реализации так называемых экспертных систем. Предложенный подход иллюстрируется на примере симуляции одной системы массового обслуживания.

В работе [156] был проведен сравнительный анализ оценки времени откли-

ка телекоммуникационной сети, полученной с помощью классических методов теории массового обслуживания, а точнее анализа модели $M|M|2$, и оценки, полученной с помощью обучения нейронной сети методом обратного распространения ошибки на наборе реальных данных, собранных при мониторинге работы исследуемой сети. Естественно, что нейронная сеть дала лучшую оценку в контексте меньшего значения среднеквадратической ошибки модели. Это было связано с простотой предложенной аналитической модели, чего авторы не скрывают, но подчеркивают, что построение более сложной математической модели системы массового обслуживания, учитывающей большее количество аспектов функционирования реальной сети, привело бы к значительному увеличению сложности ее анализа и спровоцировало бы рост финансовых затрат на ее исследование по сравнению с затратами на проектирование и обучение нейронной сети.

По большому счету, существующие работы можно разделить на несколько категорий в зависимости от целей применения методов машинного обучения. С одной стороны, машинное обучение может напрямую использоваться для моделирования работы реальных систем массового обслуживания. С другой стороны, их применение возможно для анализа сложных математических моделей массового обслуживания, для которых вычисление оценок интересующих характеристик не всегда возможно аналитически, а численные методы не всегда продуктивны.

Проанализируем имеющиеся тематические публикации более подробно, чтобы отразить одно из преимуществ применения машинного обучение, а именно разнообразие задач, которые могут быть решены с помощью его методов.

1.2 Применение методов машинного обучения для оценки характеристик различных систем/сетей массового обслуживания

Анализ характеристик реальных систем массового обслуживания различного происхождения методами машинного обучения опосредованно свидетельствует о нарастающей популярности подобной практики и может служить своего рода промежуточным этапом перед переходом к непосредственному анализу математических моделей массового обслуживания этими же методами. Поэтому остановимся и на описании публикаций подобного содержания.

Сам же вопрос построения строгой с точки зрения ТМО математической модели является непростым и решается для каждой системы индивидуально. Хотя стоит отметить, что данная научная дисциплина изначально была создана как прикладная в контексте ее применения к анализу именно систем массового обслуживания, и обладает мощным математическим инструментарием для эффективного анализа широкого круга реальных процессов. Так, например, для оценки показателей эффективности передачи данных по линиям связи различной физической природы исторически широко применяются модели и методы теории массового обслуживания [91]. Поэтому использование математического аппарата теории очередей может позволить избежать множества ошибок и неверных интерпретаций при проведении подобных исследований.

Применение ИНС для анализа моделей сетей связи. Как уже говорилось ранее, от выбора способа проектирования системы зависят ее будущие реальные характеристики. Поэтому построение адекватной имитационной модели для оценки основных вероятностно-временных показателей интересующей системы в качестве альтернативы поиску параметров производительности (качества обслуживания) в аналитическом виде является довольно актуальной задачей. В частности, при описании трафика мультисервисных телекоммуни-

кационных систем хорошо зарекомендовали себя самоподобные процессы (потоки), поэтому одним из направлений исследований в этой области является изучение моделей самоподобных процессов в контексте их использования для имитационного моделирования, и, кроме того, определения оценок параметров агрегированного потока, полученного в результате объединения трафика от индивидуальных пользователей [54, 147, 177].

Трудности, вызываемые моделированием потока, поступающего от индивидуального источника, с помощью On/Off-модели порождают новые проблемы, которые не всегда могут быть решены с помощью математических инструментов, широко используемых в классической теории массового обслуживания. Таким образом, одним из возможных решений данной проблемы стали нейронные сети. В [64,116] они использовались для прогнозирования кратковременного поведения трафика, в [80] — для поиска параметров модели, наилучшим образом описывающей характеристики реального импульсного источника, в [121, 167] — для оценки среднего числа потерянных пакетов данных. В [65] нейронные сети используются уже не для моделирования отдельных параметров сети, а самой системы в целом. В результате чего НС обучается выводить характеристику избыточной нагрузки в очереди. В [66] применяется эта же техника для моделирования небольшой сети массового обслуживания, состоящей из нескольких СМО с входящим траффиком, моделируемым тремя On/Off источниками. Представленные здесь результаты показывают, что применение нейронных сетей обеспечивает довольно надежную оценку среднего количества ожидающих пакетов, коэффициента потери пакетов и коэффициента вариации времени между отправлениями пакетов.

Снижение потерь в системах массового обслуживания в результате использования эффективной схемы управления буфером является одной из наиболее важных проблем при разработке алгоритмов управления трафиком. Уменьшение потери пакетов эквивалентно повышению эффективности, которая обычно считается инструментом оценки производительности. Так, в [209] успешно ис-

следуются возможности применения нейронных сетей с целью определения оптимальной схемы динамического управления буфером для адаптации в режиме реального времени к изменяющимся во времени условиям трафика

В более свежих работах [148, 181] решается проблема оптимального распределения ресурсов в беспроводных сетях связи уже с помощью глубокого обучения, которое, в свою очередь, является одной из форм машинного обучения. В [148] анализируются автономные сверхплотные сети (self-powered ultra-dense networks) благодаря применению глубокого Q-обучения (deep Q-learning), а в [181], если конкретизировать, то с помощью глубинных нейросетей (deep neural network) аппроксимируются два сложных итерационных алгоритма из области беспроводных сетей, предназначенных для распределения мощности сигнала, что снижает временные затраты по сравнению с использованием этих численных алгоритмов напрямую, и как следствие позволяет проводить оптимизацию распределения ресурсов сети в реальном времени.

Реальные коммуникационные сети предполагают большие объемы информации, которые должны анализироваться и интерпретироваться специальными приложениями для упреждающего и корректирующего управления, чтобы избежать перегрузок, сбоев и других аномалий в сети, которые могут в конечном итоге привести к ухудшению качества предлагаемых услуг. В [122] применяются нейронные сети для краткосрочного и среднесрочного прогнозирования трафика гетерогенной сети с целью оптимизации ее ресурсов и, тем самым, повышения качества обслуживания пользователей. В статье предлагается два различных приложения на основе нейронных и нейро-нечетких систем для управления качеством обслуживания в сетях следующего поколения для оказания услуг передачи голосовых сообщений и видеоданных (V2oIP, Voice and Video over IP). В качестве архитектуры нейронной сети выбран многослойный персептрон, а для нейро-нечеткой системы — эволюционирующая нейро-фаззи система. Кроме того, для визуализации текущего состояний сети с точки зрения качества оказываемых услуг, были использованы самоорганизующиеся карты Кохоне-

на. Тесты показали многообещающую перспективность применения нейронных сетей для прогнозирования нагрузки сети, и, кроме того, инструменты визуализации самоорганизующихся карт Кохонена обеспечили наглядность данного представления.

В [166] предлагается новый метод прогнозирования трафика беспроводной ячеистой сети (Wireless Mesh Network), основанный на использовании сети глубокого убеждения (deep belief network) в комбинации с гауссовской моделью. Результаты численного эксперимента подтверждают превосходство этого метода над тремя наиболее известными.

В [57] с помощью нейросетей подбираются основные параметры для настройки оптимальной конфигурации беспроводной сенсорной сети (англ. Wireless Sensor Network, WSN). Оптимизация выполняется на основании прогноза ИНС для таких характеристик, как срок службы сети, уровень мощности передачи, а также расстояние между узлами (network lifetime, transmission power level, internode distance). Этот способ ожидаемо значительно выгоднее с экономической точки зрения по сравнению с традиционными точными методами математического программирования, как, например, смешанное целочисленное программирование (СЦП) (англ. Mixed Integer Programming, MIP). Здесь в качестве структуры нейросети выбран многослойный персептрон, а в качестве метода обучения — метод обратного распространения ошибки. Результат численного эксперимента показывают относительную ошибку прогноза указанных параметров, не превышающую 7%.

С подробным обзором применения ИНС к исследованию беспроводных сетей связи можно ознакомиться в [56]. В [128] освещаются серьезные перспективы применения различных методов машинного обучения в целом, как мощного инструмента искусственного интеллекта, к исследованию сетей связи следующего поколения на примерах описания общей концепции решения различного рода технических проблем для гетерогенных сетей (heterogeneous networks), крупномасштабных MIMOs (massive MIMOs), умных сетей электроснабжения (smart

grid) и т. п.

Обзор [150] посвящен применению глубокого обучения, основой которого также являются ИНС, к анализу беспроводных сетей связи большой размерности и с очень сложной топологией. В [149] представлен обзор особенностей применения еще одного способа машинного обучения — глубокого обучения с подкреплением (англ. deep reinforcement learning) — к решению различных проблем в области коммуникаций и сетей связи, включая безопасность (network security), маршрутизацию трафика (traffic routing), управление скоростью передачи данных (data rate control) и многое другое, с достаточно подробным описанием самой концепции и методов обучения.

В [155] с помощью рекуррентной нейросети исследуется новый алгоритм повышения энергоэффективности беспроводной сети 5G, а в [193] ИНС предлагаются в качестве инструмента прогнозирования соотношения сигнал/шум (Signal-to-Interference-and-Noise-Ratio, SINR) в том числе и с целью снижения энергопотребления.

Среди последних работ в области применения методов машинного обучения при исследовании различных аспектов функционирования беспроводных сетей связи нового поколения можно перечислить [74,129,144,158]. Также стоит отметить, что в этой области существует множество нерешенных проблем, которые уже решаются или могут быть решены с помощью методов машинного обучения, что подтверждается множеством публикаций по данной тематике [149,150].

Применение НС для прогнозирования времени пребывания в реальных очередях. Физическая постановка в очередь является реальностью во многих сферах человеческой жизни, в частности, это касается отрасли предоставления услуг или продажи товаров. Чрезмерное ожидание в очереди может быть напряженным и утомительным, что в свою очередь может привести к снижению удовлетворенности клиентов конкретной компанией, которая совсем в этом не заинтересована. Поэтому исследования в этой области, нацеленные на

сокращение времени ожидания, а, соответственно, повышение эффективности обслуживания, остаются актуальными и сейчас.

Одним из основных инструментов оценки времени ожидания клиента в очереди остаются методы теории массового обслуживания. Однако в последнее время появилось множество исследований, посвященных прогнозированию времени пребывания в очередях с помощью методов машинного обучения [84, 119, 143, 159].

Так, в [143] прогнозируется время ожидания клиента в очереди в банк. В качестве входных значений выступают: время прихода клиента в банк (время поступления заявки в систему) — день недели, часы и минуты, место, которое клиент (заявка) занимает в очереди. Также предполагалось в качестве входного параметра использовать число обслуживающих приборов, однако в используемой авторами банковской статистике этих данных не оказалось. В качестве выходного значения выступает время ожидания в очереди до начала обслуживания. Для обучения нейросети доступные данные были разделены на тренировочный и тестовый наборы в соотношении 80% и 20%. В результате была обучена нейронная сеть с двумя скрытыми слоями из 12 и 8 нейронов, соответственно. В качестве метода обучения был выбран алгоритм оптимизации Адама для итеративного обновления весов сети на основе обучающих данных. Таким образом, здесь машинное обучение оказывается жизнеспособной альтернативой теории очередей для прогнозирования времени ожидания.

В работе [119] аналогично с помощью искусственной нейронной сети прогнозируется время пребывание клиента банка в очереди, однако акцент делается скорее на определении эффективного набора входных переменных нейросети с точки зрения наибольшего влияния на прогнозируемый параметр — время ожидания в очереди, при этом система массового обслуживания предполагается уже многолинейной. В качестве метода обучения используется алгоритм упругого обратного распространения ошибки. Архитектура нейросети представляет собой персептрон с одним скрытым слоем с различными вариациями числа

нейронов в нем. Для обучения используются реальные данные одного из крупнейших банков Индонезии, 65% из которых используется для обучения сети, а 35%, соответственно, для тестирования.

В [159] также исследуется время ожидания в очереди на обслуживание в кассе банка. При этом сравнивается точность прогнозирования трех методов машинного обучения: глубокое обучение (англ. deep learning, DL) случайный лес (англ. random forest, RF), градиентный бустинг (англ. gradient boosting machine, GBM), и СМО $M|M|1$. Все перечисленные интеллектуальные методы дают хорошее приближение по сравнению с СМО, что было вполне ожидаемо, поскольку выбранная модель массового обслуживания не лучшим образом соответствует реальной. Наиболее точный результат с точки зрения качества прогноза и значений F-меры (меры Ван Ризбергена), характеризующей соотношение точности и полноты алгоритма, дает метод градиентного бустинга.

В работе [84] был проведен анализ десяти различных методов машинного обучения для прогнозирования времени ожидания в очереди в медицинском учреждении, в число которых входят: нейронная сеть, случайный лес, метод опорных векторов (англ. support vector machine, SVM), модель эластичной сети (англ. elastic net, EN) [210], которую можно считать обобщением регрессии с регуляризацией, многомерные адаптивные регрессионные сплайны (англ. multivariate adaptive regression splines, MARS), метод k ближайших соседей (англ. k-th nearest neighbor, KNN), градиентный бустинг, бутстрэп-агрегирование или бэггинг (англ. bagging = bootstrap aggregating), дерево классификации и регрессии (англ. classification and regression tree, CART), а также линейная регрессия. Для каждой их перечисленных моделей в качестве обучающей выборки было использовано 70% имеющихся данных, а остальные 30% — для тестирования алгоритмов. Для оценки точности прогнозирования после применения обученных моделей к тестовому набору данных была вычислена среднеквадратическая ошибка, на основании чего была выбрана лучшая, в том числе и с точки зрения производительности, для оценки времени ожидания в очереди

модель, которой оказалась эластичная сеть.

1.3 Применение методов машинного обучения для оценки характеристик моделей систем массового обслуживания

Несмотря на наличие вышеописанных статей, в которых применяются методы машинного обучения при решении различного рода задач, лежащих в той или иной степени в плоскости ТМО, ни в одной из них полноценно не была сформулирована концепция новой методики. Все публикации на эту тему носят достаточно разрозненный характер. Выявить из них какую-то общую идею и уж тем более выделить ее в отдельное новое направление при решении сложных задач ТМО наравне с классическими методами довольно сложно.

Фактически основные положения применения нового метода в том виде, который позволяет его использовать при анализе моделей массового обслуживания любой сложности, были сформулированы именно в работах автора диссертации. Об этих публикациях, в том числе, речь пойдет далее. В начале для внесения ясности изложим ключевые принципы нового метода.

Основная идея и ее новизна заключаются в комбинации имитационного моделирования с различными методами интеллектуального анализа данных и применением ИНС, в частности. Имитационное моделирование является одним из способом получения высокоточных оценок показателей производительности моделей массового обслуживания, таких как среднее время отклика (среднее время пребывания заявки в системе), средняя длина очереди, среднее время ожидания заявки в очереди и других. Так, на вход имитационной модели подаются значения параметров, от которых зависят эти показатели, например, среднее время между поступлениями заявок или среднее время обслуживания на приборах, а на ее выходе мы получаем искомую величину. Однако время, затрачиваемое на получение одного значения интересующей характеристики

модели, может колебаться в пределах от нескольких секунд до нескольких минут. Это зависит от сложности моделируемой СМО, программной среды для симуляции, аппаратного обеспечения вычислительной системы (hardware).

В случае зависимости основных показателей производительности Y_i от параметров X_j, принимающих значения на заданных числовых интервалах $X_j \in [A_j, B_j]$, время симуляции для получения необходимого количества оценок в целях составления полноценного представления об их поведении может превысить все разумные пределы. Поэтому если с помощью симуляции получить набор значений интересующих характеристик для отдельных значений входных параметров в рамках заданного числового промежутка, то далее можно на полученных данных обучить нейросеть, которая с необходимой степенью точности будет давать оценку уже для любых промежуточных значений входных параметров из этих же интервалов без каких-либо ограничений на их количество.

В результате, необходимо будет затратить время на имитационное моделирование не для всех требуемых значений входных параметров, а уже для их ограниченного количества, а также на непосредственное обучение нейросети или какой-либо другой интеллектуальной модели. Сам же процесс прогнозирования фактически не требует временных затрат.

Конечно, в случае, если разработан эффективный вычислительный алгоритм для оценки вероятностно-временных характеристик определенной сети или системы массового обслуживания, но он требует слишком много временных и вычислительных затрат, то можно аналогично с помощью этого алгоритма сначала получить оценки характеристик для ограниченного набора входных параметров, а после обучить нейросеть и решать задачу прогнозирования, как это было сделано, например в [52] для немарковской СМО с «разогревом».

Что касается построения имитационной модели, то здесь существует несколько вариантов. Можно воспользоваться специализированными программными приложениями, разработанными для этих целей. Одними из наи-

более популярных приложений являются GPSS World, AnyLogic и Arena [90]. В качестве альтернативы уже готовым вариантам можно разработать собственную имитационную модель, например, в программной среде Python с довольно широким набором возможностей и множеством готовых библиотек в том числе и для обучения искусственных нейронных сетей.

Итак, резюмируя, выделим следующие основные этапы метода с использованием машинного обучения:

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.

Ввиду того, что одной из составных частей новой методики является имитационное моделирование, просто необходимо сделать несколько замечаний на этот счет. Сам по себе процесс имитационного моделирования является довольно прозрачным, особенно если построена четкая математическая модель симулируемой системы. Со спецификой построения имитационных моделей систем массового обслуживания можно ознакомиться, например, в [44, 75, 82]. Однако довольно сложным остается вопрос длины прогона модели, т. е. в данном случае подразумевается количество заявок, которое необходимо пропустить через систему для получения одного выходного значения (или одного набора выходных

значений).

Поскольку в большинстве случаев производится оценка стационарных характеристик, то на практике, как правило, поступают следующим образом. Задают некоторое малое положительное число ε и выбирают начальное значение для количества заявок L_0, которые будут проходить через систему, получают оценку искомой характеристики. Далее для тех же значений входных параметров увеличивают число заявок и сравнивают полученное новое значение с предыдущим. Если модуль их разности превышает заданное ε, то продолжают увеличивать длину прогона.

Так происходит до тех пор, пока разность текущего значения оценки Y_{L_i} и оценки $Y_{L_{i-1}}$, полученной на предыдущем шаге, не станет меньше ε

$$|Y_{L_i} - Y_{L_{i-1}}| < \varepsilon, \quad i = 1, 2, 3, ...$$

Чем меньше ε, тем выше точность полученного значения исследуемой характеристики.

При этом стоит отметить, что данные, получаемые при реализации одного прогона для оценки среднего значения практически любой исследуемой случайной величины, являются коррелированными. Причем, например, как в случае с временами пребывания заявок в системе, эта зависимость является положительной, что приводит к необходимости значительного увеличения числа реализаций по сравнению со случаем независимых случайных величин. К тому же длина прогона не является фиксированной величиной, а определяется индивидуально для каждого набора входных параметров, что дополнительно увеличивает время симуляции.

Подчеркнем, что имитационное моделирование иногда является единственно возможным способом определения вероятностно-временных характеристики сложной СМО. Однако при этом оно является довольно ресурсоемким. Поэтому применение различных методов машинного обучения позволяет существенно снизить временные затраты, сузив до минимума множество значений входных

параметров, для которых проведение симуляции необходимо. Но при этом без каких-либо ограничений на количество значений этих же входных параметров в рамках заданного числового интервала, которые впоследствии могут быть задействованы при получении других оценок показателей производительности анализируемой системы.

Применение ИНС для анализа классических моделей массового обслуживания. Работы [182, 183] посвящены построению модели классической системы массового обслуживания $M|M|1$ с помощью искусственной нейронной сети и анализу адекватности этого моделирования. В [182] нейросеть была разработана с использованием алгоритма обратного распространения ошибки с одним скрытым слоем, содержащим 28 нейронов. Эксперименты показали, что значения, моделируемые с использованием нейросети, совпадают с расчетными значениями, получаемыми благодаря использованию классического математического подхода.

В [183] была также разработана модель искусственной нейросети для моделирования СМО $M|M|1$. Входной слой состоял из четырех нейронов, соответствующих значениям интенсивности входящего потока λ, интенсивности обслуживания μ, максимального числа заявок N, одновременно находящихся в системе, и общего числа заявок в системе n, а выходной слой — из восьми нейронов, которые соответствовали значениям загрузки системы, вероятности простоя системы, среднего числа заявок в очереди и в системе, среднего времени ожидания начала обслуживания и пребывания в системе, а также вероятности того, что в системе находится n заявок. Количество скрытых слоев начиналось с одного и впоследствии увеличивалось до трех для получения большей точности решения. Моделирование сети выполнялось с использованием алгоритма обратного распространения ошибки. При этом 70% данных, полученных из аналитической модели СМО, использовалось для обучения нейросетевой модели, 20% данных — для тестирования, а оставшиеся 10% — для финальной провер-

ки (оценки). Полученная модель была протестирована, и проверка показала, что нейронная сеть в высокой степени соответствует аналитической модели и способна прогнозировать целевые параметры для заданных входных данных с минимальной незначительной ошибкой. В [184] аналогичным образом нейросеть, моделирующая классическую СМО, используется для планирования и оптимизации очереди на взлетно-посадочной полосе аэропорта.

Применение ИНС для анализа сложных моделей массового обслуживания. Собственно говоря, применение нейросетей для анализа немарковских систем массового обслуживания представляется одним из наиболее перспективных направлений исследований в области теории массового обслуживания в настоящий момент.

На сегодняшний день публикаций, посвященных данной тематике, не так много. Тем не менее, данный пробел начинает восполняться, поскольку именно посредством немарковских систем моделируется большинство реальных физических систем и процессов в них, а вот их исследование, как уже упоминалось ранее, с помощью классических методов не всегда дает удовлетворительный результат.

Одной из первых работ (если не самой первой), затронувшей применение аппарата нейронных сетей к анализу немарковских моделей СМО, является статья [52]. Здесь исследуется так называемая немарковская СМО с «разогревом» [13], которая может использоваться для моделирования процесса активации пустой системы в случае поступления в нее первой после перерыва в ее работе заявки. Данная система успешно «марковизируется» с помощью ее приближения посредством СМО с распределением входящего потока или времени обслуживания фазового типа, которое, как известно, может применяться в качестве аппроксимирующей функции произвольного закона распределения. Однако уже в случае СМО с «разогревом» вида $H_2|M|M|3$ и $M|H_2|M|3$ численные алгоритмы расчета стационарных вероятностей состояний оказываются

очень трудоемкими и ресурсозатратными. Применение нейронных сетей в этом случае позволило заметно снизить трудоемкость без потери в точности вычислений.

В качестве структуры нейросети был выбран двухслойный персептрон. В качестве функции активации на скрытом слое использовался гиперболический тангенс, а на выходах сети — линейная функция. На вход нейронной сети подавались интенсивности входящего и обслуживающего потоков, а также «разогрева» и коэффициент вариации. Выходными параметрами в случае СМО $H_2|M|M|3$ было стационарное распределения числа заявок, а для $M|H_2|M|3$ — среднее время ожидания обслуживания и среднее время пребывания в системе.

Для обучения ИНС использовались три алгоритма: алгоритм обучения Левенберга-Маркарта, метод шкалированных связанных градиентов и метод Байесовской регуляризации. Последний оказался самым точным в смысле минимальной среднеквадратической ошибки приближения в случае с числом нейронов в скрытом слое равным 40.

Новая методика была апробирована и на нескольких вариантах топологий систем из широкого класса систем поллинга. Под системами поллинга подразумеваются системы, состоящие из нескольких очередей и одного обслуживающего прибора [202], который в соответствии с некоторой стратегией перемещается между очередями и обслуживает находящиеся там заявки. Системы поллинга применяются для моделирования и анализа производительности различного рода реальных систем и протекающих в них процессов, например, речь может идти о производственных, транспортных или телекоммуникационных системах, системах управления запасами и т. д. [76]. Однако несмотря на востребованность подобных систем, в этой области остается множество нерешенных задач.

В работе [203] с помощью нового подхода исследуются различные характеристики для трех типов систем поллинга. Первая из них представляет собой систему, состоящую из N очередей типа $M|M|1$, т. е. в каждую i-ю очередь

$(i = 1, ..., N)$ поступает пуассоновский поток заявок с интенсивностью λ_i, обслуживание заявок из каждой очереди единым для всех очередей прибором происходит циклически в соответствии со шлюзовой дисциплиной и показательным законом распределения с интенсивностью μ_i. Время переключения прибора между соседними очередями также имеет показательное распределение с математическим ожиданием s_i. Для описанной архитектуры системы поллинга известно решение в аналитическом виде, в частности среднее время пребывания в каждой из очередей определяется выражением

$$v_i = \frac{q_i(i,i)(1+\rho_i)}{2\lambda_i^2 C} + \frac{1}{\mu_i}, \quad i = 1, ..., N, \tag{1.7}$$

где $\rho_i = \lambda_i/\mu_i$, C — время цикла, т. е. время, затрачиваемое прибором на один обход всех очередей, включая время переключения между очередями, а $q_i(j,k)$, $i,j,k = 1, ..., N$ — решение некоторой системы линейных уравнений. Обучение нейросети происходило на данных, рассчитанных с помощью известных аналитических формул. В результате нейросеть показала хорошее качество приближения для среднего взвешенного времени пребывания заявок в системе

$$v = \sum_{i=1}^{N} \rho_i v_i.$$

Кроме того, в статье [203] рассматриваются еще два вида структур систем поллинга, а именно — система с коррелированным входным потоком типа $MAP|M|1$, а также типа $M|M|1$ с адаптивным циклическим опросом. В отличие от предыдущего случая для таких систем известных решений в аналитическом виде не существует, поэтому для обучения нейросети использовались данные имитационного моделирования. Обученная нейросеть аналогично продемонстрировала хорошее качество приближения для среднего времени пребывания в очередях для обоих вариантов систем.

В статье [29] исследуется система вида k-из-n с точки зрения надежности. Математические модели k-из-n имеют широкое практическое приложение в различных отраслях, например, телекоммуникационной отрасли и робототехнике,

криптографии, системах мониторинга подводных трубопроводов, добыче полезных ископаемых и т. д. Для анализа различных типов k-из-n систем, как правило, применяют аналитические методы, базирующиеся на многомерных марковских процессах.

Так, в статье рассматривается система k-из-n, $k < n$, которая представляет собой ремонтируемую СМО с единственной ремонтной единицей. Отказ данной системы происходит в случае выхода из строя k элементов, каждый из которых начинает немедленно ремонтироваться после прекращения функционирования, а после завершения восстановления начинает заново работать. Предполагается, что время жизни компонент системы имеет экспоненциальное распределение с параметром α, а время ремонта имеет произвольное распределение со средним значением b.

Описанную замкнутую СМО можно обозначить в терминах классификации Кендалла, как $\langle M_{k<n}|G|1\rangle$. Множество состояний данной системы описывается двумерным марковским процессом $Z(t)_{t\geq 0} = \{J(t), X(t)\}_{t\geq 0}$, где $J(t)$ — это состояние системы за время t (характеризует количество элементов системы, вышедших из строя, и, соответственно, работоспособность всей системы в целом), а $X(t)$ — время, прошедшее с момента ремонта отказавшего элемента или всей системы. Заметим, что заданный таким образом процесс функционирования системы, становится марковским, благодаря введению второй компоненты $X(t)$.

В результате, после составления системы дифференциальных уравнений Колмогорова и последующего ее решения были получены выражения для стационарных вероятностей состояний СМО в терминах преобразования Лапласа.

Далее для проверки работоспособности нового метода в рамках численного эксперимента проводится сравнительный анализ результатов обучения нейросети и аналитики. Поскольку здесь было получено точное решение для стационарных вероятностей системы, которое позволяет оценить наиболее важную характеристику надежности ее работы, называемую коэффициентом готовно-

сти

$$K_{av} = 1 - \pi_k,$$

где π_k — это вероятность того, что k компонент вышли из строя, т. е. вся система вышла из строя и ремонтируется, то обучение нейросети проходило не на имитационных данных, а на данных, рассчитанных с помощью аналитических выражений.

В качестве распределений для времени ремонта рассматриваются экспоненциальное распределение, распределение Гнеденко-Вейбулла, а также гамма-распределение. При этом во всех случаях основные компоненты архитектуры системы $\langle M_{k<n}|G|1\rangle$ принимали одинаковые значения: $k = 3$ и $n = 6$. В качестве структуры нейросети выбран двухслойный персептрон с двумя входными нейронами, соответствующими среднему времени жизни и среднему времени ремонта, на выходе нейросети — один нейрон, выдающий значения коэффициента готовности, в скрытом слое содержится 16 нейронов, а в качестве активационных функций выступают гиперболический тангенс и его производная, алгоритм обучения нейросети — метод Адама. Результаты работы нейросети на тестовом наборе данных говорят о хорошем качестве приближения коэффициента готовности, что открывает новые перспективы при изучении систем надежности более общего вида $\langle G_{k<n}|G|1\rangle$.

В работе [92] нейронные сети используются для поиска оптимальной политики распределения заявок в многолинейной СМО с накопителем неограниченной емкости, неоднородными приборами и эксплуатационными расходами. В систему поступает пуассоновский поток заявок с интенсивностью λ, время обслуживания заявки на j-м приборе имеет экспоненциальное распределение с параметром μ_j, причем приборы упорядочены в порядке возрастания среднего времени обслуживания

$$\mu_1 \geqslant \mu_2 \geqslant ... \geqslant \mu_K.$$

Кроме того, в системе предполагаются расходы, а именно стоимость пребыва-

ния заявки в очереди (англ. the holding cost of waiting in the queue) равна $c_0 > 0$, а эксплуатационные расходы на обслуживание заявки (англ. operating cost) на j-м приборе составляют c_j в единицу времени, т. е. в среднем на одну заявку при обслуживании на j-м приборе расходуется $c_j\mu_j^{-1}$ и

$$c_1\mu_1^{-1} \leqslant c_2\mu_2^{-1} \leqslant ... \leqslant c_K\mu_K^{-1}.$$

В системе имеется контроллер, который на основе информации о состоянии системы принимает решение о распределении заявок между приборами в соответствии с некоторой политикой f. Контроллер в момент поступления новой заявки в систему или в момент окончания обслуживания заявки на приборе может либо отправить первую в очереди заявку на обслуживание, либо оставить ее в очереди.

Оптимальная политика распределения заявок между приборами с точки зрения минимизации долгосрочных средних расходов для такой системы имеет пороговый вид

$$1 \leqslant q_1 \leqslant q_2 \leqslant ... \leqslant q_K < \infty,$$

где q_j, $j = \overline{1, K}$ — пороговые уровни. Если число заявок в очереди больше или равно q_k, но не превышает порог q_{k+1}, то должны быть задействованы k первых наиболее быстрых приборов.

Для того, чтобы определить оптимальные пороги используется универсальный итерационный алгоритм (англ. policy-iteration algorithm) [123,189], поскольку минимизация функции средних расходов (анг. average cost function) напрямую может оказаться слишком трудоемкой. Однако и у этого алгоритма есть некоторые недостатки, в частности, возникают сложности с его сходимостью в условиях высокой загрузки системы, а также имеются ограничения на пространство состояний исследуемого процесса. Поэтому в статье предлагается два пути решения этой непростой задачи. С одной стороны, сформулировано эври-

стическое решение для нахождения пороговых уровней следующего вида

$$q_k \approx \min\left\{1, \left\lfloor \frac{\sum_{j=1}^{k-1}\mu_j - \lambda}{c_0}\left[\frac{c_k}{\mu_k} - \frac{\sum_{j=1}^{k-1} c_j}{\sum_{j=1}^{k-1}\mu_j}\right]\right\rfloor\right\}.$$

С другой стороны, представлено решение на основе нейронных сетей. На вход нейросети поступают значения λ, μ_j, c_0, c_j, $j = \overline{1,K}$, а на выходе нейросети — искомые оптимальные пороги. Обучение нейросети (шестислойный персептрон) проходило на 70% данных, полученных с помощью итерационного алгоритма, методом Адама в программной среде Mathematica (Wolfram Research). Проверка на оставшихся 30% данных показала низкую погрешность аппроксимации, что указывает на высокий потенциал решения задач оптимизации в области теории массового обслуживания.

И в завершении стоит отметить монографию [8], посвященную применению методов машинного обучения к теории очередей. Она включает в себя часть из описанных в данной главе статей и не только, а также более поздние работы по сравнению с представленными в диссертации, связанные с применением методов машинного обучения к анализу fork-join систем, включая [12, 197], где исследуется система с входным MAP-потоком, фазовым распределением времени обслуживания и конечными очередями, причем для частного случая двух подсистем получен аналитический результат в виде алгоритма для оценки нижней и верхней границ математического ожидания времени пребывания заявки в системе, а для отдельных случаев большего числа подсистем $MAP|PH|1|R$ применяется симуляция в комбинации с градиентным бустингом, нейросетью и деревьями решений.

1.4 Примеры

Сети массового обслуживания (СеМО) традиционно используются для моделирования а, соответственно, оценки производительности и оптимизации многих сложных систем, чем и объясняется интерес к изучению как открытых, так

и замкнутых сетей. Например, речь может идти об анализе компьютерных систем совместно с их подсистемами, сетях связи и о многих других различного рода технических, экономических, производственных, транспортных, медицинских, военных и т. п. системах [11].

Одним из наиболее востребованных направлений проведения исследований, а также часто встречающихся примеров является применение СеМО в различных вариантах их конфигурации в качестве моделей компьютерных сетей, которые, в свою очередь, стали уже неотъемлемой частью современной жизни. Так, компьютерные сети предоставляют своим пользователям довольно широкий спектр услуг, к которым относится и электронная почта, и всевозможные новостные сервисы, и работа с удаленными базами данных и многое другое. Кроме того, на базе компьютерных сетей реализуется дистанционное обучение, телемедицина, телеконференции, которые в свете известных событий 2020 года наглядно продемонстрировали свою актуальность, хотя запрос на улучшение сетевых технологий существовал и будет существовать всегда.

Таким образом, довольно быстрый и непрекращающийся рост числа компьютерных сетей и их пользователей приводит к необходимости развития теоретических основ проектирования компьютерных сетей и, как следствие, разработке новых подходов к анализу основных характеристик производительности построенных аналитических моделей сетей [7].

1.4.1 Применение методов машинного обучения для оценки характеристик замкнутых экспоненциальных сетей

Обзор известных методов анализа замкнутых СеМО. Более подробно остановимся на описании существующих методов исследования замкнутых СеМО, основных достоинствах и недостатках этих методов. Необходимо заметить, что каких-то серьезных сдвигов в направлении разработки новых подходов изучения замкнутых сетей так и не произошло, а самих методов анализа известно не так уж и много, причем все они зависят от принадлежности сети к опреде-

ленному типу.

Все подходы к анализу замкнутых сетей можно разделить на точные и приближенные. Сперва обратимся к точным методам исследования. В этой связи необходимо упомянуть об аналоге теоремы Джексона в случае открытых СеМО, но уже для замкнутых сетей. Это так называемая теорема Гордона–Ньюэлла, которая распространяет свое действие на однородные замкнутые сети с экспоненциальными временами обслуживания и накопителями неограниченной емкости [113]. Из этой теоремы следует, что стационарное распределение числа заявок в узлах замкнутой сети имеет мультипликативный вид.

Позднее теорема была обобщена теоремой BCMP, названной в честь авторов статьи, в которой впервые была описана сеть одноименного типа (Баскетта, Чанди, Мунца и Паласиоса), распространив мультипликативный вид стационарных вероятностей на более широкий круг сетей, но, тем не менее, с ограничением на вид функции распределения времени обслуживания, которая должна быть либо экспоненциальной, либо просто иметь рациональное представление Лапласа–Стилтьеса, причем в последнем случае количество приборов в узлах сети должно равняться одному или быть равным общему числу заявок, циркулирующих в сети [72].

Зная стационарные вероятности, можно определить основные вероятностно-временные характеристики всей сети. Однако ключевая проблема, возникающая при вычислении вероятностей состояний заключается в расчете нормирующей константы, присутствующей в определяющем их выражении. Прямое ее вычисление требует значительных ресурсов и больших временных затрат даже при относительно небольшой размерности сети. Данное обстоятельство привело к появлению специальных рекуррентных методов для вычисления нормирующей константы. Один из первых методов, который впоследствии стал основой для многих других позже разработанных алгоритмов — это метод Бузена или, как еще его называют, алгоритм свертки [78]. Хотя и здесь есть свои сложности, а именно велика вероятность при увеличении, например, числа узлов или ко-

личества заявок, одновременно находящихся в сети, получить машинный ноль, т. е. получить обнуление результатов, или их переполнение в памяти вычислительной машины.

Также существует еще один рекуррентный метод, сравнительно более эффективный в вычислительном отношении, поскольку не требует знания стационарных вероятностей и соответственно, предварительного вычисления нормализующей константы. Он относится к так называемому анализу средних значений (англ. Mean Value Analysis, MVA) и базируется на равенстве интенсивностей входящего в узел и выходящего из этого же узла потоков заявок, а также на формуле Литтла [172]. В результате применения этого итерационного метода можно вычислить такие важные характеристики производительности для любой сети, как средние времена пребывания заявок в узлах сети и среднее число заявок в каждом из узлов, кроме того, можно вычислить и другие показатели, например, маргинальное распределение длины очереди в отдельном узле. Однако, несмотря на отсутствие проблемы с обнулением и/или переполнением, с ростом числа узлов и общего числа заявок растет и вычислительная сложность алгоритма, а вместе с ней и необходимый объем вычислительных ресурсов и времени затрачиваемого на выполнение расчетов. Причем не исключена ситуация, когда требуемые затраты могут превысить возможности даже современных вычислительных машин, тогда единственным решением становится применение приближенных методов анализа.

Если говорить о замкнутых СеМО, не удовлетворяющих теореме BCMP, то точные методы, как правило, разработаны только для отдельных видов сетей, состоящих не более, чем из двух узлов, причем за редким исключением с более, чем одним прибором в каждой из подсистем [59, 70, 77, 86], а также ограниченным набором типов распределений для обслуживания на самих приборов, например, экспоненциальным или фазовым распределением, распределением Кокса [79]. При этом, стоит отметить, встречаются отдельные работы, в которых очереди к узлам сети могут быть и конечными, что, разумеется, приводит

к появлению блокировок до или после обслуживания. Так, в [69, 83] предлагается точное решение для сетей нелинейной топологии с конечными накопителями в однолинейных узлах и, соответственно, различными вариантами блокировок, но с экспоненциальными временами обслуживания на приборах.

Теперь перейдем к обсуждению приближенных методов. Несмотря на то, что в случае сетей с экспоненциальными узлами и накопителями неограниченной емкости известны точные методы решения, разработке приближенных процедур вычислений с целью анализа их производительности было посвящено не малое число работ. Большинство аппроксимационных алгоритмов основывается на методе анализа средних, а их преимущество заключается в большей эффективности и, соответственно, скорости работы по сравнению с точными методами. Приближенные методы были разработаны также и для экспоненциальных сетей аналогичной архитектуры, но уже и с накопителями ограниченной емкости. Большинство процедур также базируется на методе анализа средних либо на декомпозиционном подходе, применение которого не ограничивается экспоненциальными временами обслуживания, поэтому далее рассмотрим этот метод уже в контексте произвольных распределений времен обслуживания.

Основная идея метода декомпозиции заключается в разбиении сети на подсистемы, их изолированном анализе и дальнейшем объединении результатов для получения характеристик сети в целом. В роли подсистем могут выступать как слабо связанные подсети, так и непосредственно сами узлы сети. В последнем случае предполагается уже применение различных подходов к анализу систем массового обслуживания (СМО). При этом дополнительная сложность будет заключаться в определении основных параметров функционирования СМО, а именно — параметров входящего и выходящего потоков для каждого из узлов. Поэтому довольно часто декомпозиционный подход комбинируется с диффузионной аппроксимацией, которая, в свою очередь, применяется для анализа более общих моделей узлов типа $GI|G|1$. Причем этот метод оказался эффективен с некоторыми ограничениями и в случае открытых сетей массового

обслуживания с произвольной функцией распределения длительности обслуживания [73, 141, 205]. Однако его точность сильно зависит от уровня загрузки фрагментов сети, т. е. чем уровень загрузки выше, тем приближение лучше, а также от значения коэффициента вариации и конкретного типа распределения времени обслуживания, причем погрешность метода иногда может носить неприемлемый характер.

Под декомпозицией иногда понимают еще и переход к рассмотрению эквивалентных сетей. Так, довольно эффективная итеративная процедура была предложена R.A. Marie, которую в зарубежной литературе так и называют по имени автора [151, 152]. Она основана на теореме Нортона и относится к методам, которые позволяют анализировать сети с узлами типа $\cdot|G|m$. Для того, чтобы оценить производительность отдельного узла, оставшаяся часть сети преобразуется в единый узел с экспоненциальной длительностью времени обслуживания, зависящей от числа находящихся там заявок. Затем вычисляются интенсивности обслуживания в композиционном узле, что не совсем просто. После чего составляется система уравнений для изолированного узла с произвольной функцией распределения времени обслуживания и интенсивностью входящего потока, зависящей от нагрузки, которая имеет решение в замкнутом виде. В результате, удается получить маргинальные вероятности распределения числа заявок в анализируемом узле. Аналогичным образом исследуется каждый из узлов сети. Однако при этом все-таки предполагается, что произвольная функция распределения времени обслуживания в узловых фрагментах сети имеет рациональное представление Лапласа, т. е. должна аппроксимироваться обобщенным распределением Кокса с различными параметрами в зависимости от значения коэффициента вариации исходной функции.

Если говорить о случае сетей с конечной буферной емкостью и произвольными законами распределения времен обслуживания в узлах, то методов их приближенного анализа известно не так много. Так, например, в [58] анализ такой сети фактически сводится к ее аппроксимации эквивалентной сетью с

неограниченной емкостью накопителя в узлах и с последующим применением метода Marie. Также переходят от анализа замкнутой сети к исследованию открытой сети с аналогичными характеристиками, возможны и другие варианты декомпозиции, но, как правило, в комбинации с известными методами для анализа сетей без блокировок.

Как уже упоминалось выше, каких-то серьезных сдвигов в направлении появления новых методов исследования СеМО не наблюдается, тем не менее, среди последних работ можно отметить, например, [208], в которой автор адаптировал метод анализа средних значений для экспоненциальных сетей с блокировками и однолинейными узлами, в [180] проводится оценка параметров экспоненциальной замкнутой сети с блокировками для многолинейных узлов с помощью комбинации метода декомпозиции, заключающегося фактически в разбиении каждого узла сети на две СМО, и анализа средних значений, в [89] был предложен метод декомпозиции применительно к сетям с блокировками и произвольным распределением времени обслуживания.

Резюмируя вышесказанное, можно заключить, что для сетей не из категории BCMP известные методы при определенных условиях могут приводить либо к недопустимым погрешностям оцениваемых характеристик, либо сами вычисления могут занимать значительное время сопоставимое со временем, затрачиваемым на имитационное моделирование, причем большие временные затраты и серьезные вычислительные трудности могут возникать и для сетей, обладающих свойством мультипликативной формы при условии, например, большой размерности сети и большого числа заявок, циркулирующих в ней, что в общем-то характерно для реальных компьютерных сетей. Поэтому применим новый подход к получению характеристик замкнутых СеМО.

Поскольку основная задача состоит в демонстрации нового подхода, то в качестве наглядного примера и иллюстрации применения нового подхода рассмотрим простую экспоненциальную замкнутую сеть, которая, как известно, поддается аналитическому анализу с некоторыми ограничениями.

Математическая модель замкнутой СеМО и численный эксперимент. Итак, рассмотрим замкнутую СеМО, состоящую из M узлов, каждый из которых представляет собой однолинейную СМО с накопителем неограниченной емкости. Порядок выбора заявок на обслуживание в каждом узле определяется дисциплиной «первым пришел – первым обслужился» (англ. First Come First Served, FCFS). Функция распределения времени обслуживания на приборе i-го узла является экспоненциальной с параметром μ_i. В замкнутой сети заявки не поступают извне и не покидают сеть, поэтому их число постоянно. Обозначим их количество через K. Заявки перемещаются из одного узла сети в другой в соответствии с маршрутной матрицей $\Theta = (\theta_{ij})$, где θ_{ij} — это вероятность мгновенного перехода в j-й узел после обслуживания на приборе i-го узла. Поскольку заявки не могут покинуть сеть, то для маршрутной матрицы должно выполняться условие

$$\sum_{j=1}^{M} \theta_{ij} = 1, \quad i = \overline{1, M}.$$

Через λ_i обозначим интенсивность выходящего потока из узла i. Схема описанной сети представлена на рисунке 2, причем $\theta_j = \sum_{i=1}^{M} \theta_{ij}$ — это суммарная вероятность попадания в j-ю подсистему массового обслуживания. Характери-

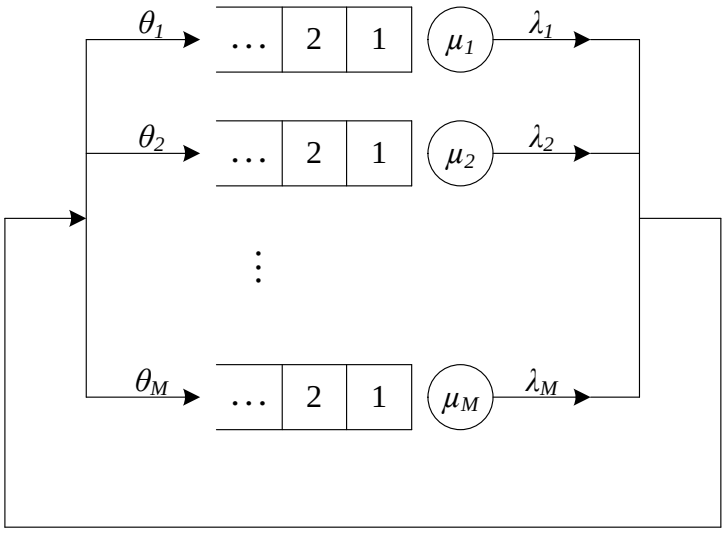


Рис. 2: Схема замкнутой сети массового обслуживания.

стики представленной сети могут быть получены в аналитическом виде, однако в силу того, что мы хотим проанализировать результативность нового подхода, то будем следовать именно ему.

На первом этапе необходимо разработать имитационную модель сети. С этой целью был выбран язык для программирования Python, а не готовое программное приложение. Теперь зададим конкретные значения параметров сети. Пусть матрица вероятностей переходов по узлам сети равна

$$\Theta = \begin{pmatrix} 0.1 & 0.3 & 0.4 & 0.1 & 0.1 \\ 0.2 & 0.1 & 0.2 & 0.3 & 0.2 \\ 0.3 & 0.1 & 0.1 & 0.3 & 0.2 \\ 0.2 & 0.2 & 0.1 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0.2 & 0.3 & 0.3 \end{pmatrix},$$

число узлов $M = 5$, а количество заявок $K = 100$. В качестве оцениваемых показателей производительности сети выберем среднее число заявок и среднее время пребывания заявок в каждом из узлов, которые обозначим через N_i и v_i, соответственно, $i = \overline{1,5}$. Входные данные, которые будут меняться — это интенсивности обслуживания на каждом из приборов μ_i. Все параметры μ_i будут варьироваться на интервале $[3.0, 4.0]$ с шагом 0.25 (табл. 1). В результате с

Таблица 1: Входные данные для обучения ИНС

№ п/п	μ_1	μ_2	μ_3	μ_4	μ_5
1	3.00	3.00	3.00	3.00	3.00
2	3.00	3.00	3.00	3.00	3.25
3	3.00	3.00	3.00	3.25	3.00
4	3.00	3.00	3.00	3.25	3.25
5	3.00	3.00	3.25	3.00	3.00
...	...	...	...	...	...
3125	4.00	4.00	4.00	4.00	4.00

помощью имитационной модели необходимо будет получить в общей сложности 3125 наборов данных.

Далее будем проводить обучение нейросети также в программной среде Python. Для этого полученную выборку разобьем на обучающую и тестовую в классическом соотношении 80% и 20%, соответственно. На обучающей выборке мы будем фактически тренировать ИНС, а для оценки проведенного обучения будем использовать оставшиеся 20%, которые не участвовали в обучении, но позволят оценить, насколько хорошо обученная нейросеть будет решать задачу прогнозирования в реальности для новых входных данных.

Чтобы улучшить качество прогноза с помощью ИНС данные как обучающей, так и тестовой выборок были подвержены стандартизации и нормализации. Под стандартизацией будем понимать приведение набора входных данных для каждого из оказывающих на систему влияние параметров к такому виду, что его математическое ожидание станет равным нулю, а стандартное отклонение — единице. Под нормализацией данных здесь подразумевается масштабирование значений входных параметров в пределах от 0 до 1.

В качестве структуры нейросети выберем многослойный персептрон. Несмотря на то, что одного скрытого слоя уже было бы достаточно, мы все-таки остановимся на двух слоях для возможности большей управляемости процессом обучения. Каждый скрытый слой будет состоять из 10 нейронов с логистической функцией активации $\varphi(x) = 1/(1 + e^{-x})$ на каждом из них. Обучение будет происходить методом обратного распространения ошибки. На выходе будет всего один нейрон, соответствующий одному из оцениваемых параметров, т. е. фактически мы будем строить 10 нейросетей (рис. 3). Таким образом, это положительно должно сказаться на качестве приближения оцениваемых показателей функционирования замкнутой сети.

Для оценки качества прогноза, выдаваемого нейросетью, будем использовать среднеквадратическую ошибку (MSE), среднюю абсолютную ошибку (MAE) и среднюю абсолютную процентную ошибку ($MAPE$), которые опре-

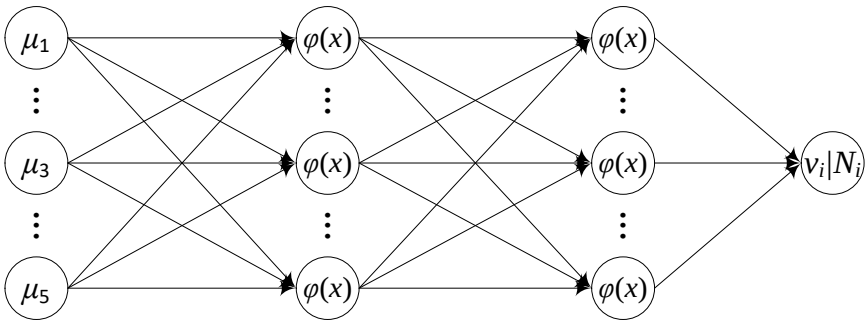


Рис. 3: Схема трехслойного персептрона, используемого для оценки характеристик производительности замкнутой СеМО.

деляются следующими выражениями

$$MSE = \frac{1}{N}\sum_{j=1}^{n}(y_j - \widehat{y}_j)^2, \tag{1.8}$$

$$MAE = \frac{1}{N}\sum_{j=1}^{n}|y_j - \widehat{y}_j|, \tag{1.9}$$

$$MAPE = \frac{1}{N}\sum_{j=1}^{n}\left|\frac{y_j - \widehat{y}_j}{y_j}\right| \cdot 100\%, \tag{1.10}$$

где $\widehat{y}_j$ — это оценка исследуемой характеристики, полученная с помощью ИНС на тестовой или какой-либо другой выборке для j-го входного набора данных. В нашем случае речь идет о среднем числе заявок в одном из узлов СеМО N_i либо о среднем времени, проведенным заявками в конкретном узле сети v_i, $i = \overline{1,5}$. Величина y_j — реальное значение оцениваемой характеристики, полученное в результате имитационного моделирования замкнутой сети, $j = \overline{1,N}$, а N — количество наборов данных в выборке.

Разумеется, на процесс построения оптимальной архитектуры нейросети (количество скрытых слоев и нейронов в них), на выбор функций активации на нейронах, а также конкретного метода обучения, которых не так уж и мало, и прочего можно было бы потратить гораздо больше времени, однако все перечисленные манипуляции позволяют получить уже достаточно хороший результат, как видно из таблицы 3.

Таблица 2: Входные данные для оценки характеристик производительности замкнутой сети массового обслуживания с помощью ИНС

№ п/п	μ_1	μ_2	μ_3	μ_4	μ_5
1	3.10	3.10	3.10	3.10	3.10
2	3.10	3.10	3.10	3.10	3.20
3	3.10	3.10	3.10	3.20	3.10
4	3.10	3.10	3.10	3.20	3.20
5	3.10	3.10	3.20	3.10	3.10
...	...	...	...	...	...
59049	3.90	3.90	3.90	3.90	3.90

Теперь, чтобы окончательно удостовериться в работоспособности построенной нейросети и в качестве оценок, получаемых с ее помощью, сделаем прогноз на каждую из вышеуказанных характеристик производительности замкнутой СеМО для значений интенсивностей времен обслуживания, меняющихся в том же диапазоне, что и раньше $[3.0, 4.0]$, но уже с шагом 0.1, например (табл. 2). Таким образом, количество различных наборов входных данных будет составлять уже 59049 единиц. Это примерно в 19 раз больше, чем число наборов данных, которое потребовалось для проведения имитационного моделирования с целью обучения нейросети в данном случае. Напомним, что в отличие от имитационного моделирования прогноз по ИНС делается практически мгновенно, поэтому выигрыш от их применения очевиден.

Возможность осуществить проверку работы ИНС на таком значительном объеме данных у нас имеется благодаря выбору именно экспоненциальной замкнутой СеМО в качестве объекта исследования, поскольку характеристики производительности данной сети можно вычислить аналитически и, соответственно, провести необходимый сравнительный анализ. Для вычисления нормирующих констант, присутствующих в выражениях для стационарных веро-

Таблица 3: Погрешности приближений оценок характеристик производительности замкнутой сети массового обслуживания, полученных с помощью ИНС, на тестовом наборе данных

Оцениваемая характеристика	Типы ошибок		
	MSE	MAE	$MAPE$, %
v_1	0.000012	0.002821	0.306338
v_2	0.000006	0.001542	0.185111
v_3	0.000042	0.003707	0.301359
v_4	0.001378	0.027710	0.128102
v_5	0.000449	0.013940	0.893533
N_1	0.000019	0.003138	0.115259
N_2	0.000043	0.004066	0.227245
N_3	0.000691	0.018952	0.646385
N_4	0.011222	0.076211	0.101165
N_5	0.002921	0.032369	0.845084

ятностей, в Python был запрограммирован алгоритм Бузена, с которым более подробно можно ознакомиться в первоисточнике [78] или, например, в [7]. После чего в той же программной среде были рассчитаны средние характеристики узлов замкнутой сети.

Кратко опишем последовательность действий при проведении расчетов по аналитическим формулам. Известно, что для замкнутых однородных экспоненциальных СеМО существует аналог теоремы Джексона, который носит название теоремы Гордона–Ньюелла. Из этой теоремы следует, что стационарное распределение числа заявок в узлах замкнутой сети имеет мультипликативный вид [7]

$$p(\mathbf{k}) = G^{-1}(M, K) \prod_{i=1}^{M} d_i^{k_i}, \quad \mathbf{k} \in S(M, K), \tag{1.11}$$

$$S(M,K) = \left(\mathbf{k} = (k_1, ...k_M), k_i \geqslant 0, i = 1, ..., M, \sum_{i=1}^{K} k_i = K\right),$$

где $S(M,K)$ — это пространство состояний сети. Нормирующая константа определяется выражением

$$G(M,K) = \sum_{\mathbf{k} \in S(M,K)} \prod_{i=1}^{M} d_i^{k_i}, \tag{1.12}$$

где d_i рассчитывается в соответствии с формулой

$$d_i = \frac{h_i}{\mu_i},$$

причем $(h_1, ..., h_M)$ — это одно из нетривиальных (ненормированных) решений следующей системы уравнений равновесия

$$h_j = \sum_{i=1}^{M} h_i \theta_{ij}, \quad j = 1, ..., M, \tag{1.13}$$

а h_i имеет смысл среднего числа посещений заявкой узла j (между двумя последовательными посещениями узла с порядковым номером 1, $h_1 = 1$).

Формулы (1.11) и (1.12) имеют довольно простой вид, однако прежде чем ими воспользоваться необходимо решить систему уравнений равновесия (1.13). Кроме того, вычисление нормирующей константы по формуле (1.12) может представлять серьезную сложность в случае большого числа циркулирующих заявок и узлов, поскольку количество слагаемых в этой сумме соответствует мощности пространства состояний сети, которая комбинаторно возрастает при увеличении числа узлов и заявок в нем. Таким образом, расчет по прямым формулам становится если не невозможным, то очень затруднительным. Поэтому были разработаны специальные методы для вычисления нормирующей константы. Один из первых методов, который впоследствии стал основой для многих других позже разработанных алгоритмов —- это метод Бузена. Далее представлены формулы для вычисления нормирующей константы посредством

алгоритма Бузена [78]. Вводится новая функция

$$g(m,k) = \sum_{\mathbf{k}\in S(m,k)} \prod_{i=1}^{m} d_i^{k_i}, \quad m = 1,...,M, \quad k-1,...,K$$

которая фактически является нормирующей константой в случае замкнутой СеМО с m узлами и k заявками. Очевидно, что для этой функции будет выполняться условие

$$g(M,K) = G(M,K).$$

Рекуррентное же соотношение для вычисления нормирующей константы имеет вид

$$g(m,k) = g(m-1,k) + d_m g(m,k-1), \quad m = 1,...,M, \quad k = 1,...,K.$$

Кроме того, понадобятся граничные условия

$$g(m,0) = 0, \quad m = 1,...,M,$$

$$g(0,k) = 0, \quad k = 1,...,K.$$

В результате, для вычисления нормирующей константы потребуется в общей сложности $2KM(K+1)$ арифметических операций.

Для вычисления стационарных характеристик производительности узлов сети необходимо вычислить маргинальное распределение числа заявок в этих узлах по формуле

$$p_i(k) = d_i^k G^{-1}(M,K)\big[g(M,K-k) - u(K-k)d_i g(M,K-k-1)\big].$$

Среднее число заявок в узле i определяется формулой

$$N_i = G^{-1}(M,K)\sum_{k=1}^{K} d_i^k G(M,K-k), \quad i = 1,..,M,$$

а интенсивность выходящего из i-го узла потока — выражением

$$\lambda_i = h_i \frac{g(M,K-1)}{G(M,K)}.$$

Тогда среднее время пребывания заявки в i-м узле согласно формуле Литтла вычисляется по формуле

$$v_i = \frac{N_i}{\lambda_i}.$$

Таблица 4: Погрешности приближений оценок характеристик производительности замкнутой сети массового обслуживания, полученных с помощью ИНС, для входного набора данных из табл. 2

Оцениваемая характеристика	Типы ошибок		
	MSE	MAE	$MAPE$, %
v_1	0.000014	0.002876	0.284568
v_2	0.000003	0.001275	0.161997
v_3	0.000120	0.003230	0.235173
v_4	0.004063	0.055876	0.227230
v_5	0.000239	0.009265	0.639649
N_1	0.000904	0.010727	0.287468
N_2	0.000024	0.003369	0.203218
N_3	0.001016	0.018032	0.551451
N_4	0.015708	0.087965	0.104054
N_5	0.010458	0.042261	0.765629

Для нашего примера средние значения ошибок на тестовой выборке, содержащей 20% случайным образом выбранных данных из общего набора, представлены в таблице 3, а средние значения тех же типов ошибок, но уже для входных параметров из таблицы 2 представлены в таблице 4. Как видно, средние значения ошибок различаются, но это неудивительно, поскольку количество наборов данных в тестовой выборке во много раз меньше, хотя при этом средняя относительная погрешность приближений для всех оцениваемых характеристик не превышает 1%, что, разумеется, является хорошим результатом.

Для того, чтобы понять, что скрывается за усредненными результатами раз-

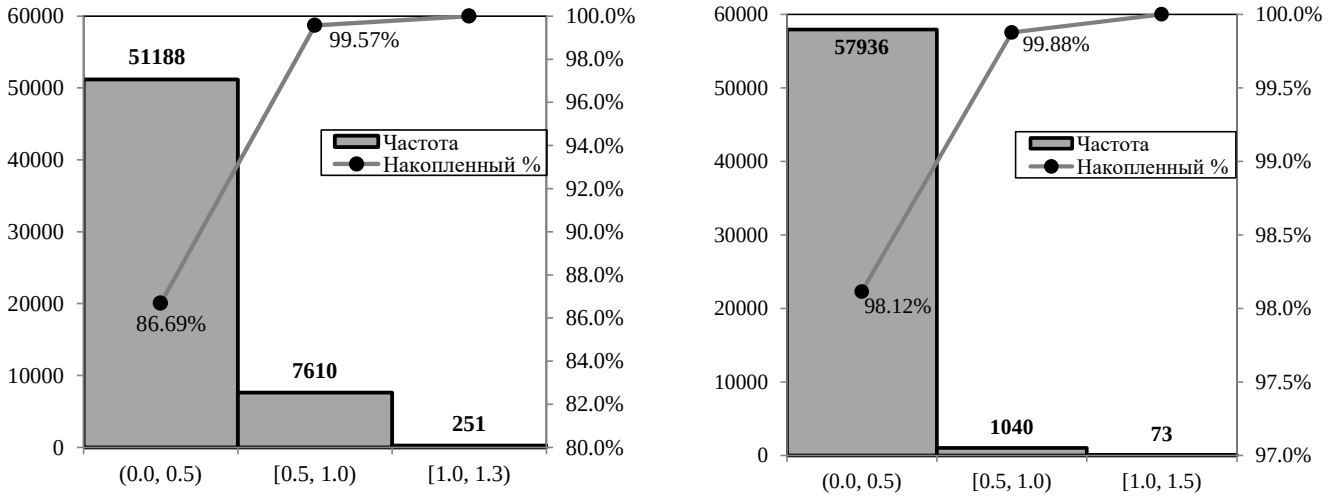
берем более подробно структуру относительных ошибок, входящих в состав $MAPE$. В силу того, что сделать какие-либо выводы на основании графика, содержащего практически 60 тысяч точек, очевидно, будет затруднительно, то построим гистограммы частот попадания в отдельные интервалы относительных ошибок приближения с отражением накопительного эффекта в процентном выражении для наглядности (рис. 4–8).

Как видно из графиков, погрешность приближения в целом не превышает 5%. Причем в самом лучшем случае максимальное значение относительной погрешности приближения не больше 1.55%, что справедливо для величин v_1, v_2, v_4 и N_2 (рис. 4, 5б, 7а), для показателей N_1 и N_4 максимум в первом случае не превышает, а во втором случае равен 3% (рис. 6б, 8а), для v_5 — не превышает 3.7% (рис. 6а). В условно худшем случае относительные ошибки аппроксимации попадают в интервал от 4% до 5% для величин v_3, N_3 и N_5 (рис. 5а, 7б, 8б), однако количество таких ошибок в общей сложности не превышает полпроцента от их общего числа, т. е. 59049, что говорит о хорошем качестве полученных с помощью ИНС оценок основных показателей производительности замкнутой сети.

1.4.2 Применение методов машинного обучения для оценки характеристик открытых неэкспоненциальных сетей

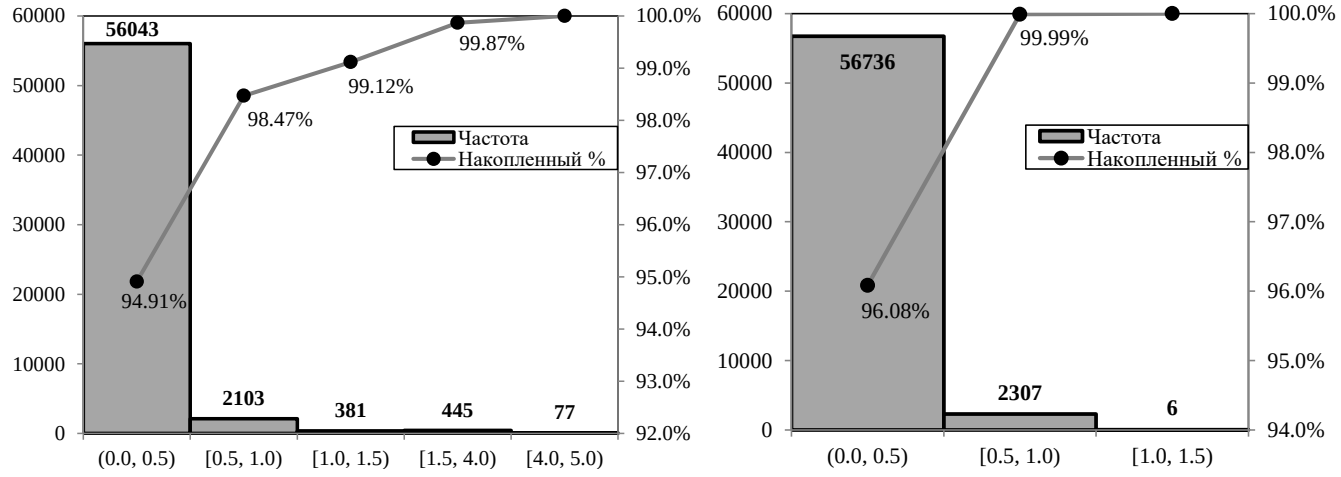
В данном разделе будут рассмотрены так называемые многофазные или тандемные системы массового обслуживания, структура которых более детально будет описана далее. Что же касается реальных физических систем, которые адекватно моделируется подобными сетями, то одним из наиболее ярких примеров является функционирование широкополосных беспроводных сетей связи с линейной топологией [199–201].

Причем в данном контексте речь будет идти именно об открытых системах, в которые заявки поступают из внешней среды и после последовательного обслуживания на всех ее фазах или узлах покидают сеть. Также будем считать,



а) б)

Рис. 4: Гистограмма частоты значений модуля относительных погрешностей приближений (%) для оценок: *а)* — среднего времени v_1, *б)* — среднего времени v_2, полученных с помощью нейросети.

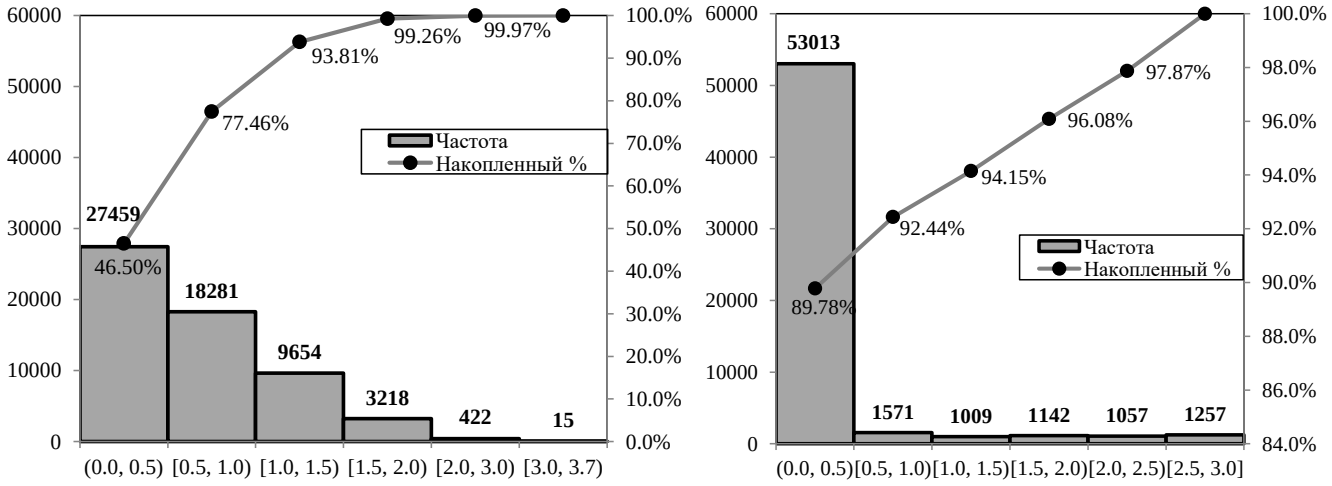


а) б)

Рис. 5: Гистограмма частоты значений модуля относительных погрешностей приближений (%) для оценок: *а)* — среднего времени v_3, *б)* — среднего времени v_4, полученных с помощью нейросети.

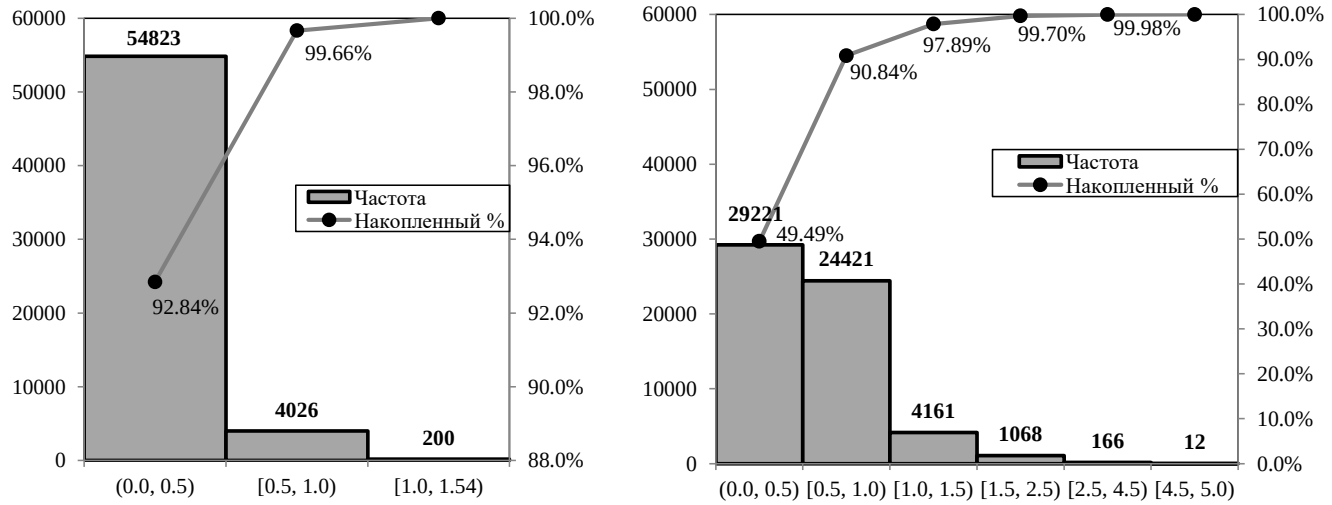
что каждый узел сети имеет накопитель неограниченной емкости.

Обзор известных методов анализа открытых СеМО. Точные аналитические результаты были получены совсем для небольшого класса сетей мас-



а) б)

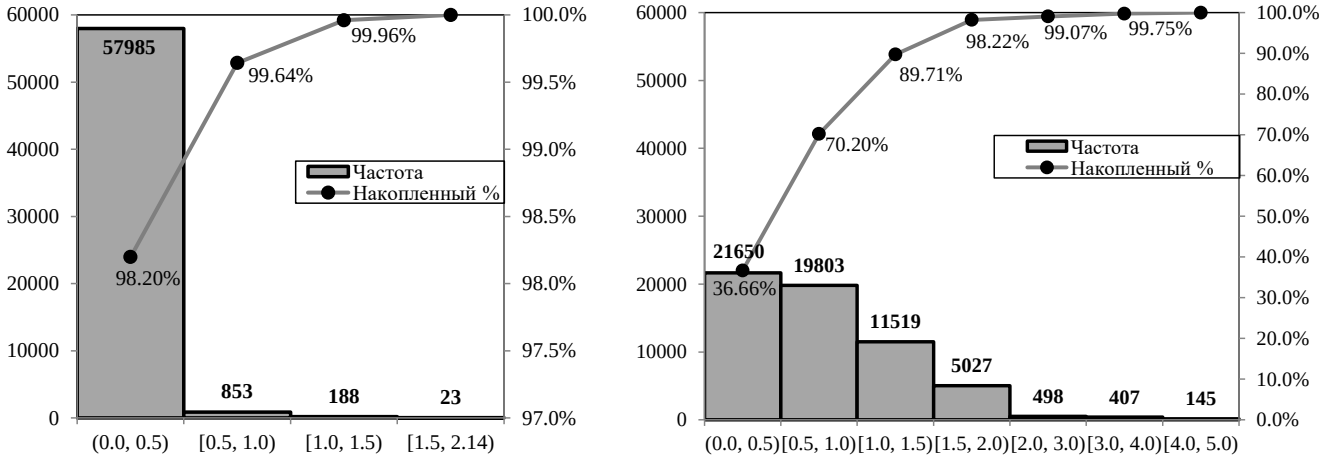
Рис. 6: Гистограмма частоты значений модуля относительных погрешностей приближений (%) для оценок: *а)* — среднего времени v_5, *б)* — среднего числа заявок N_1, полученных с помощью нейросети.



а) б)

Рис. 7: Гистограмма частоты значений модуля относительных погрешностей приближений (%) для оценок: *а)* — среднего числа заявок N_2, *б)* — среднего числа заявок N_3, полученных с помощью нейросети.

сового обслуживания [124–126, 130, 131]. В частности, речь идет об открытых сетях Джексона с пуассоновским входящим потоком и экспоненциальными временами обслуживания. В этом случае узлы сети представляют собой системы



а) б)

Рис. 8: Гистограмма частоты значений модуля относительных погрешностей приближений (%) для оценок: *а)* — среднего числа заявок N_4, *б)* — среднего числа заявок N_5, полученных с помощью нейросети.

массового обслуживания типа $M|M|1$, и совместное стационарное распределение числа заявок в узлах сети имеет довольно простой мультипликативный вид. Поэтому нахождение всевозможных показателей качества функционирования таких систем не составляет особого труда. Также отметим, что вид решения в форме произведения справедлив и для некоторых других случаев, указанных в теореме BCMP [7, 72].

Для большинства же других видов сетей точных решений в замкнутой форме не существует, поэтому для их анализа применяются различные методы аппроксимации и, соответственно, их комбинации. Особенно это касается сетей массового обслуживания, количество узлов в которых превышает число два. В результате, приближенный анализ сетей общего вида фактически становится единственно возможным решением. Хотя, конечно, нельзя не отметить, что в некоторых отдельных случаях точные аналитические методы исследований все же возможны, но из-за серьезной размерности пространства состояний изучаемых сетей, даже в условиях дальнейшего применения существенных возможностей современных вычислительных систем, являются нецелесообразны-

ми [11, 68, 137].

Таким образом, одним из наиболее распространенных методов анализа СеМО общего вида становится метод декомпозиции, который подразумевает переход к анализу каждого узла сети в отдельности с последующим изучением их взаимодействия между собой и дальнейшей компоновкой полученных результатов для оценки основных показателей производительности сети в целом [170]. При этом возникает потребность в получении оценок первых и вторых моментов выходных потоков каждого из узлов сети. Поэтому такой подход еще называют приближенной теорией второго порядка для СеМО [5].

В контексте указанного подхода один из способов исследования СеМО общего вида называется методом диффузионного приближения, который фактически используется в рамках анализа изолированных фрагментов сети типа $GI|G|1$. Пожалуй, одним из существенных недостатков данного вида аппроксимации является сравнительно низкая точность в условиях малой и средней нагрузок на узлы сети, а также ее сильная зависимость от выбора конкретного распределения, в случае анализа одномерного диффузионного процесса. Что, впрочем, может быть отчасти нивелировано за счет, например, введения аналогичного, но уже двумерного процесса, приближенно описывающего процесс формирования очереди [3,4]. При этом относительная погрешность аппроксимации показателей изолированного узла, как показали численные эксперименты, может быть значительно снижена.

Кроме того, в [188] для тандемных сетей был описан так называемый эффект «узкого места», которым является последняя фаза обслуживания. И здесь даже усовершенствованные процедуры диффузионной аппроксимации, одна из которых описана в [205], не всегда показывают удовлетворительные результаты, хотя позднее в [87, 206] были предложены алгоритмы, улучшающие точность анализа. Тем не менее, несмотря на некоторые, иногда весьма существенные в отношении точности приближений недостатки, описанный способ позволяет определить необходимые параметры выходных потоков из узловых подсистем,

что является существенным при анализе именно сетей, а не систем массового обслуживания.

И, наконец, третий подход к исследованию СеМО заключается в имитационном моделировании. Вероятно, это один из наиболее реалистичных подходов к анализу физически существующих СеМО. Однако он не всегда подразумевает своего рода универсальность, присущую теории массового обслуживания, поскольку моделирование может производиться для отдельно взятой реальной физической системы и, соответственно, возможно исключение этапа построения математической модели как таковой.

Отдельно стоит сказать о том, что существует множество работ, посвященных исследованию двухузловых тандемных систем с марковским входным потоком (англ. Markovian Arrival Process, MAP) и его обобщением — групповым марковским входным потоком (англ. Batch Markovian Arrival Process, BMAP), которые, вообще говоря, не являются рекуррентными, но тем не менее позволяют учесть сложный и, в том числе коррелированный, характер потоков в современных телекоммуникационных сетях, например, таких как широкополосные сети 4G или сети нового поколения 5G. Однако, несмотря на всю сложность и общность таких моделей, они не могут использоваться для анализа сетей с немарковскими потоками, которые описываются, например, такими законами распределения как гамма-распределение, распределения Вейбулла, Парето, логнормальное, равномерное, детерминированное и т. д. [33].

Математическая модель многофазной СеМО. Средняя межконцевая задержка. Рассмотрим открытую сеть массового обслуживания с линейной топологией, которая состоит из K узлов, $K \geqslant 2$ (рис. 9). Причем первый узел сети представляет собой систему массового обслуживания типа $GI|G|1$, а остальные — $\cdot|G|1$. Подобные сети фактически относятся к так называемым многофазным или тандемным системам массового обслуживания, в которых заявка после обслуживания на одной фазе или узле последовательно переходит

на обслуживание на следующей фазе, и так происходит до тех пор, пока она не покинет систему после окончания обслуживания на последнем K-м узле.

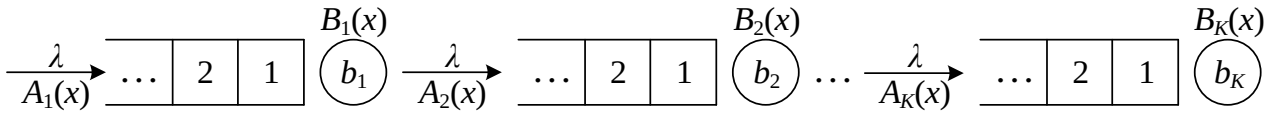


Рис. 9: Схема функционирования тандемной СеМО.

Одной из наиболее важных характеристик производительности любой системы или сети массового обслуживания является ее время отклика. Когда речь идет о сетях, данную характеристику чаще называют межконцевой задержкой. Фактически эта величина позволяет оценить время, проведенное пользователем сети в ожидании ответа на свой запрос.

Оценка характеристик СеМО на основе метода декомпозиции. Как известно, точное выражение для среднего значения межконцевой задержки, когда речь идет о более, чем двух последовательных подсистемах, было получено только в единичных случаях, в частности, когда узлы сети представляют собой системы массового обслуживания типа $M|M|1$ и для некоторых расширений данной СеМО [7, 130]. В остальных же ситуациях, когда входящий поток не является пуассоновским были предложены различные аппроксимации математического ожидания времени отклика.

Поскольку мы имеем дело с тандемной системой, то средняя межконцевая задержка представляет собой сумму средних времен пребывания заявок на каждом узле (рис. 10)

$$v = v_1 + v_2 + ... + v_K.$$

Таким образом, для того чтобы получить приближение для искомой величины необходимо отдельно рассмотреть каждый i-й узел сети, который характеризуется входящим потоком с функций распределения времени между поступлениями заявок $A_i(x)$ и функцией распределения времени обслуживания заявок

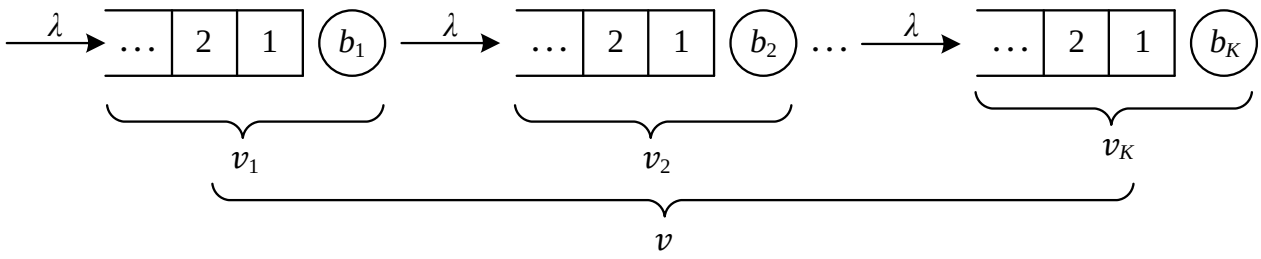


Рис. 10: Средняя межконцевая задержка v в тандемной СеМО.

$B_i(x)$, $i = \overline{1, K}$. Причем среди всех функций распределения, описывающих входящие в узлы потоки, предполагается известной функция распределения только лишь для входящего в сеть потока, т. е. $A_1(x)$.

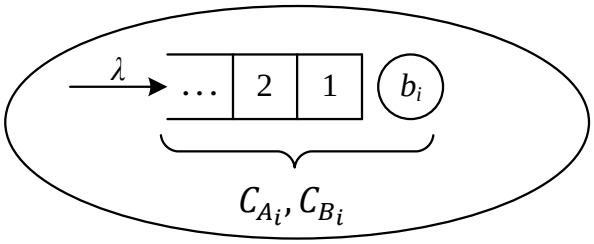


Рис. 11: Узел многофазной СеМО типа $G|G|1$.

Основная проблема заключается в необходимости оценки параметров входящего потока для второй и последующих узлов сети, помимо, разумеется, того, что до настоящего времени не существует методов нахождения точного выражения для среднего времени пребывания заявок в СМО вида $GI|G|1$, а только его приближения.

Тем не менее, в силу неограниченности накопителей узлов рассматриваемой сети, интенсивность входящего потока заявок в узел i совпадает с интенсивностью выходящего из него потока, т. е. этот параметр фактически инвариантен относительно вида функций $A_i(x)$ и $B_i(x)$. Следовательно, если интенсивность входящего потока на первую фазу мы обозначим через λ, то в силу линейности СеМО, интенсивность входящих потоков на все остальные узлы будет также равна λ.

Основная формула, используемая для приближения среднего времени от-

клика i-й фазы, была предложена в [139] и имеет вид

$$\widehat{v}_i = \frac{b_i \rho_i}{2(1-\rho_i)}(C_{A_i}^2 + C_{B_i}^2) g(\rho_i, C_{A_i}, C_{B_i}) + b_i, \tag{1.14}$$

где

$$g(\rho, C_A, C_B) = \begin{cases} \exp\left\{ -\frac{2(1-\rho)}{3\rho} \cdot \frac{(1-C_A^2)^2}{C_A^2 + C_B^2} \right\}, & \text{если} \quad 0 \leqslant C_A \leqslant 1, \\ \exp\left\{ -(1-\rho) \cdot \frac{C_A^2 - 1}{C_A^2 + 4C_B^2} \right\}, & \text{если} \quad C_A > 1, \end{cases} \tag{1.15}$$

$\rho_i = \lambda b_i$ — загрузка узла i, C_{A_i} — коэффициент вариации (отношение среднеквадратического отклонения к математическому ожиданию) для входящего потока в узел i, т. е. для распределения $A_i(x)$, C_{B_i} — коэффициент вариации для случайной величины времени обслуживания на приборе в i-м узле, т. е. для функции распределения времени обслуживания $B_i(x)$, а b_i — это математическое ожидание времени обслуживания на приборе в этом же узле сети, $i = \overline{1, K}$.

Данное выражение было получено эмпирическим путем при исследовании СМО $GI|G|1$, причем при $g(\rho, C_A, C_B) = 1$ оно полностью совпадает с формулой для первого момента времени пребывания заявки в СМО $M|G|1$.

В итоге, для того, чтобы вычислить среднее время пребывания заявки, к примеру, в узле $(i+1)$ необходимо оценить еще и коэффициент вариации входящего в эту фазу потока. Для этой цели существуют следующие формулы:

$$C_{A_{i+1}} = C_{B_i}, \tag{1.16}$$

$$C_{A_{i+1}} = \rho_i(1-\rho_i) + \rho_i^2 C_{B_i}^2 + (1-\rho_i) C_{A_i}^2, \tag{1.17}$$

$$C_{A_{i+1}} = C_{A_i}^2 + 2\rho_i C_{B_i}^2 - \rho_i (C_{A_i}^2 + C_{B_i}^2) g(\rho_i, C_{A_i}, C_{B_i}), \tag{1.18}$$

$$C_{A_{i+1}} = \rho_i^2 C_{B_i}^2 + (1-\rho_i^2) C_{A_i}^2, \tag{1.19}$$

где $g(\rho_i, C_{A_i}, C_{B_i})$ определяется в (1.15). Первая из них была предложена в работе [171] и основывается на предположении о том, что при большой нагрузке сети вероятность того, что на фазе i не производится обслуживание, близка к нулю, из чего следует, что выходящий из этого узла поток стремится к рекуррентному

и его распределение совпадает с распределением времени обслуживания здесь же.

Формула (1.17) была получена в [99]. А выражение под номером (1.18) представил П. Кюн в своей работе [140], оно является следствием формулы для коэффициента вариации выходящего потока из [153] для СМО $GI|G|1$:

$$C_{A_{i+1}} = C_{A_i}^2 + 2\rho_i C_{B_i}^2 - \frac{2\rho_i(1-\rho_i)(\widehat{v}_i - b_i)}{b_i},$$

И, наконец, соотношение (1.19) является упрощением (1.18) в предположении, что $g(\rho_i, C_{A_i}, C_{B_i}) = 1$ [140].

Несмотря на то, что все перечисленные результаты были получены относительно давно, до настоящего времени существенных сдвигов при проведении исследований в данном направлении, к сожалению, не произошло. Более точные приближения были получены лишь для систем, ограничивающихся, как правило, множеством условий. Одно из наиболее распространенных заключается в дуальности тандемных систем, т. е. в наличии только двух последовательных фаз обслуживания [11].

Оценка характеристик СеМО с помощью ИНС. Численный эксперимент. В качестве альтернативы описанному выше методу изолированного анализа узлов сети воспользуемся подходом, основанным на комбинации имитационного моделирования с методами интеллектуального анализа данных и с обучением искусственной нейросети, в частности.

Чтобы продемонстрировать эффективность описанного подхода далее в рамках численного эксперимента сравним результаты оценки средней межконцевой задержки, полученные посредством метода декомпозиции и с помощью ИНС.

Для оценки времени отклика рассмотрим тандемную СеМО с различным числом фаз и несколькими типами вероятностных распределений для входящего потока и времен обслуживания. Метод декомпозиции в контексте формул

для приближенного расчета математического ожидания времени отклика дает совершенно разную точность в зависимости от величины загрузки узлов и видов распределения для интервалов времени между поступлениями заявок и времен их обслуживания. Относительная погрешность вычислений может быть относительно не велика и принимать значение в пределах доли процента для отдельного узла, и наряду с этим, в некоторых случаях ее значение может доходить до нескольких десятков процентов.

Для сравнительного анализа рассмотрим два варианта распределений, которые показывают в случае изолированного анализа узлов сети линейной топологии неплохие результаты, а именно — равномерного распределения и распределения Парето при определенных ограничениях на параметр формы (см. далее). Т. к. в ситуации с большой относительной погрешностью, превышающей десятки процентов, применение приближенной теории второго порядка будет, по большому счету, не столь рациональным.

Для оценки погрешности приближений как и раньше будем использовать среднеквадратическую ошибку (MSE), среднюю абсолютную ошибку (MAE) и среднюю абсолютную процентную ошибку ($MAPE$), которые определяются следующими выражениями

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(v_j - \widehat{v}_j)^2, \quad MAE = \frac{1}{N}\sum_{j=1}^{N}|v_j - \widehat{v}_j|,$$

$$MAPE = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{v_j - \widehat{v}_j}{v_j}\right| \cdot 100\%,$$

где $\widehat{v}_j$ — это оценка средней межконцевой задержки, полученная либо с помощью ИНС на тестовой выборке, либо посредством одной из формул (1.16)–(1.19) на том же наборе входных параметров, v_j — реальное значение, выдаваемое системой, т. е. в данном случае результат имитационного моделирования сети, N — количество оцениваемых элементов (число элементов в тестовой выборке), $j = \overline{1, N}$. Для проведения численного эксперимента была разработана имита-

ционная модель в программной среде Phyton, в ней же и происходило обучение ИНС методом обратного распространения ошибки.

Итак, в первом случае рассмотрим тандемную СеМО с равномерным распределением входящего в нее потока на отрезке $[1; d]$, а также с равномерной функцией распределения времени обслуживания на приборе с параметрами 1 и 29, т. е. $R[1; 29]$, одинаковой для каждого из узлов сети, так же как и нагрузка $\rho_i = 30/(1 + d)$, $i = \overline{1, K}$. Значения параметра d подбираются таким образом, что коэффициент загрузки меняется от 0.1 до 0.9 с шагом 0.1. При этом количество фаз обслуживания K будет изменяться в пределах от 2 до 200 включительно. В качестве структуры обучаемой нейросети выберем трехслойный персептрон с двумя входными параметрами — числом фаз обслуживания K и коэффициентом загрузки ρ, а также единственным выходным параметром — средней межконцевой задержкой, и логистической функцией активации на каждом нейроне $\varphi(x) = 1/(1 + e^{-x})$ (рис. 12).

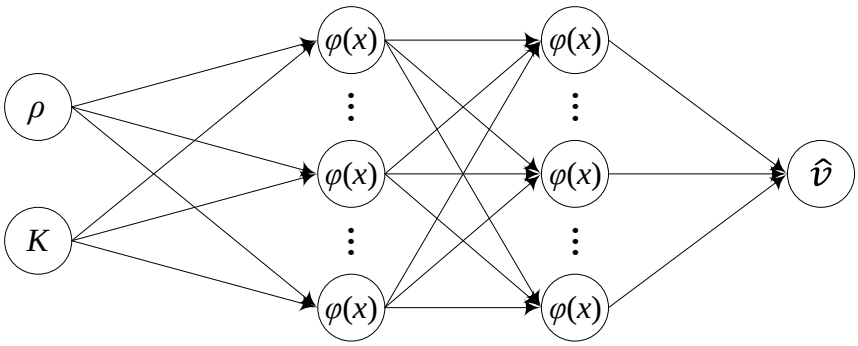


Рис. 12: Схема трехслойного персептрона, используемого при обучении нейросети.

С целью улучшения качества оценки с помощью ИНС, также как и ранее, данные обучающей и тестовой выборок были подвержены предварительной обработке, а именно стандартизации и нормализации.

Погрешность оценки приближения вычислим для конкретного значения параметра $d = 99$ и, соответственно, $\rho = 0.3$. Как видно из таблицы 5, в случае равномерного распределения формулы показывают относительно невысокую

Таблица 5: Погрешности приближений при вычислении средней межконцевой задержки с помощью формул (1.16)–(1.19) и ИНС в случае равномерного распределения для $\rho = 0.3$

Тип ошибки	MSE	MAE	$MAPE$, %
Формула (1.16)	15435.175	108.865	5.984
Формула (1.17)	12386.602	94.106	4.753
Формула (1.18)	12330.974	94.128	4.755
Формула (1.19)	12393.388	94.140	4.756
Нейросеть	4.703	1.829	0.137

погрешность аппроксимации — не более 6%. Тем не менее в этом случае удается обучить нейронную сеть таким образом, что результат, прогнозируемый с ее помощью оказывается значительно лучше.

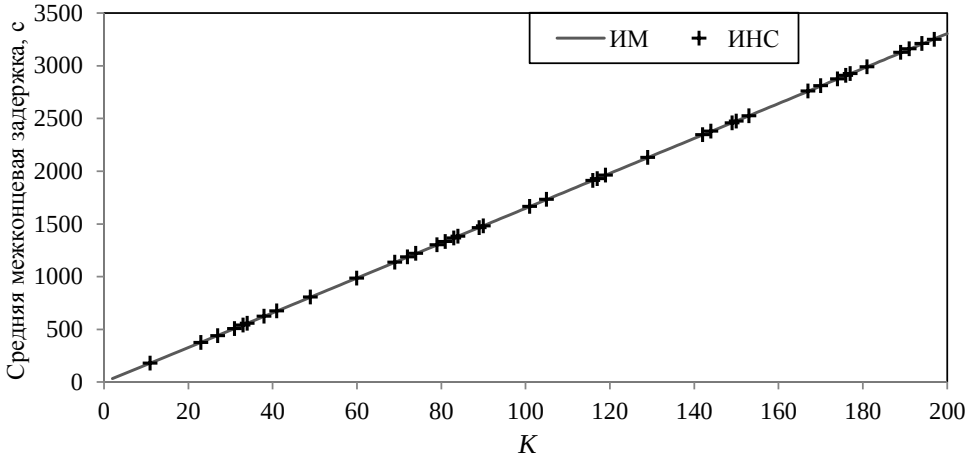


Рис. 13: Результат обучения ИНС в случае равномерного распределения для входящего в сеть потока и времен обслуживания в узлах сети, $\rho = 0.3$.

Что касается распределения Парето, то стоит напомнить, что оно является двухпараметрическим. Первый параметр — коэффициент формы — обозначим через α. Для второго параметра — коэффициента масштаба m, который по сути определяет минимальное значение, принимаемое случайной величиной с указанным распределением, выберем значение единица ($\alpha, m > 0$).

При этом следует отметить, что на практике интересен случай, когда параметр формы для случайной величины с распределением Парето меняется в пределах от 1 до 2 включительно, тогда ее среднее значение конечно, а дисперсия бесконечна. Таким образом может моделироваться, например, фрактальный трафик, характеризующийся высоким уровнем пульсаций, в рамках некоторой физической системы. Однако в силу того, что мы сравниваем два подхода, то вариант с бесконечной дисперсией не имеет смысла, поэтому ограничимся случаем с $\alpha > 2$.

Так, для времени обслуживания выберем значение параметра формы 25, тогда значения аналогичного параметра для входящего потока будут находиться в пределах от ≈ 2.119 до 15.625. Следовательно, это позволит нам получить интервал $[0.55, 0.975]$ для значений коэффициента загрузки с шагом 0.025 с учетом указанного выше ограничения на параметр α.

Таблица 6: Погрешности приближений при вычислении средней межконцевой задержки с помощью формул (1.16)–(1.19) и ИНС в случае распределения Парето

Тип ошибки	MSE	MAE	$MAPE$, %
Формула (1.16)	1.080	0.655	2.321
Формула (1.17)	1.648	1.106	6.387
Формула (1.18)	1.595	1.088	6.224
Формула (1.19)	1.604	1.091	6.209
Нейросеть	0.335	0.319	0.927

В качестве структуры ИНС будем использовать также трехслойный персептрон с двумя входными параметрами, а именно числом узлов сети K и коэффициентом загрузки ρ (рис. 12). Результаты обучения с помощью ИНС представлены в таблице 6. Погрешность аппроксимации здесь рассчитана на основе тестовой выборки, содержащей 248 элементов, при этом число элемен-

тов в обучающей выборке составляет 994. Формулы (1.17)–(1.19) в отличие от (1.16) в данной ситуации показывают более низкую точность, что было ожидаемо, поскольку вывод формулы (1.16) как раз основывается на предположении о высоком уровне загрузки системы. Однако результат обучения ИНС оказывается все равно лучше.

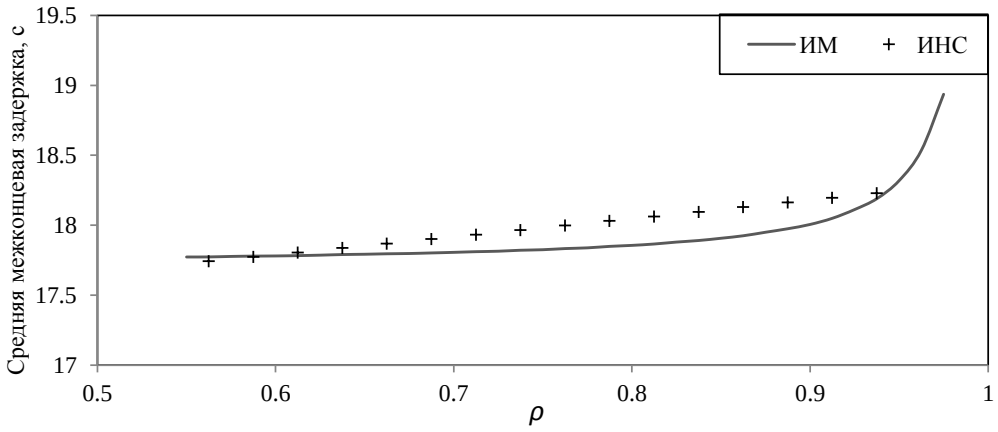


Рис. 14: Результат обучения ИНС в случае распределения Парето для входящего в сеть потока и времен обслуживания в узлах сети для $K = 17$.

Заметим, что при обучении ИНС данные разбивались на обучающую и тестовую выборки в соотношении 80% и 20% случайным образом в обоих примерах. Так, на рисунке 13 представлена зависимость межконцевой задержки от количества узлов сети K при фиксированной величине загрузки $\rho = 0.3$. Сплошной линией изображен результат имитационного моделирования (ИМ), а маркерами отмечены значения, полученные посредством обучения нейросети.

На рисунке 14 аналогично сплошной линией показана зависимость межконцевой задержки от загрузки узлов для $K = 17$. Маркерами отмечены значения, полученные с помощью нейросети для промежуточных значений загрузки узлов, которые не использовались в качестве входных параметров при обучении сети. Величина $MAPE$, рассчитанная для этих данных, составляет примерно 0.7%, когда величина аналогичной ошибки для формул (1.16)–(1.19) меняется в пределах от 1.8% до 6%. При этом максимальное значение относительной ошибки для нейросети составляет всего 1.4%, в то время, как для аналитиче-

ской формулы (1.16), показывающей наилучшее приближение, максимум уже равен 6.3%.

1.5 Выводы к главе 1

В настоящей главе представлен обзор работ, посвященных применению методов машинного обучения к анализу стохастических систем в рамках теории массового обслуживания. Новизна метода заключается в применении ИНС для прогнозирования основных характеристик производительности систем в комбинации с предварительным имитационным моделированием на ограниченном наборе данных.

В большей части упомянутых статей исследуется применение ИНС или других методов машинного обучения к исследованию реальных систем массового обслуживания в области беспроводных сетей связи или при анализе реальных очередей в банках и медицинских учреждениях. Также представлены работы, посвященные применению новой методики применительно к анализу математических моделей массового обслуживания. Сформулированы основная идея и принципы проведения исследований с применением методов машинного обучения, также обозначены важные особенности, в том числе связанные с проведением имитационного моделирования.

В качестве иллюстрации применения нового метода, а также его эффективности был проведен анализ параметров производительности открытой и замкнутой сетей массового обслуживания. 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.

2 Получение основных стационарных характеристик систем с разделением и параллельным обслуживанием в случае подсистем с пуассоновским входящим потоком и экспоненциальным распределением времени обслуживания

Исследуется классическая система массового обслуживания с разделением и параллельным (англ. fork-join) обслуживанием заявок с пуассоновским входным потоком и экспоненциальными временами обслуживания на приборах. Fork-join система является математической моделью для множества реально существующих систем, в которых происходит распараллеливание задач, но несмотря на традиционные для ТМО предположения о распределениях для входящего и обслуживающего потоков, анализ такой системы довольно сложен. Точные результаты известны только для частного случая такой системы (число подсистем равно двум), к примеру, известно аналитическое выражение для среднего времени отклика такой системы. Для более общего случая СМО, когда число подсистем больше двух, известны только разнообразные аппроксимации различной степени точности для среднего времени отклика и его дисперсии. Таким образом, исследования такой архитектуры fork-join СМО остается актуальным до настоящего момента.

В данной главе система с разделением и параллельным обслуживанием заявок сначала будет исследована с помощью нового метода с использованием нейронных сетей, особенности которого были описаны в предыдущей главе, а далее для анализа системы будет применен комплексный метод, результатом которого является уже не обученная нейросеть, а аналитические выражения хорошего качества приближения, причем не только для моментов времени отклика системы, но и для квантилей высоких уровней распределения времени

отклика. Также в главе представлены точные аналитические выражения для коэффициентов корреляции между временами пребывания подзаявок в подсистемах; построена мета-гауссовская модель, позволяющая с использованием коэффициентов корреляции, получить оценки не только среднего времени отклика, но и дисперсии, квантилей и других характеристик; изучены приближения совместного распределения времен пребывания подзаявок с помощью теории копул; описываются и важные аспекты проведения имитационного моделирования системы с разделением и параллельным обслуживанием, а также оценки его результатов. Результаты Главы 2 отражены в работах [16, 26, 102–104, 107]

2.1 Математическая модель системы с разделением и параллельным обслуживанием

Рассматривается fork-join система массового обслуживания, состоящая из K очередей и K приборов, $K \geqslant 2$. В систему поступает пуассоновский поток заявок с параметром λ, $\lambda > 0$, случайное время обслуживания на каждом k-ом приборе имеет экспоненциальное распределение с одним и тем же параметром μ, $\mu > 0$, $\eta_k \sim Exp(\mu)$, $k = 1, ..., K$.

Теперь опишем более подробно структуру системы и основные принципы ее функционирования. При поступлении в систему заявка разделяется (fork point) на число подзаявок, равное числу подсистем и, соответственно, общему числу приборов K. Все подсистемы фактически представляют собой самостоятельные системы массового обслуживания с бесконечной очередью и единственным прибором. Каждая из подзаявок поступает в одну из подсистем, обслуживается там, после чего попадает в условный буфер синхронизации (join point), где ожидает обслуживания оставшихся частей заявки. После окончания обслуживания всех подзаявок происходит мгновенная сборка целой заявки, и только после этого она может покинуть систему. Особенности функционирования рассматриваемой СМО с разделением заявок заключаются в следующем (рис. 15):

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.

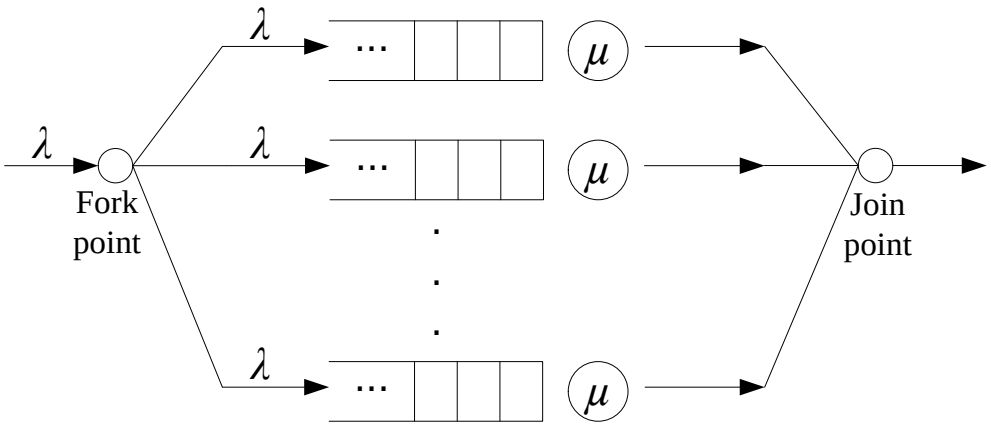


Рис. 15: Модель fork-join системы массового обслуживания с подсистемами типа $M_\lambda|M_\mu|1$.

Заметим, что известны и другие схемы обслуживания подзаявок, когда, например, приборы блокируются до окончания обслуживания всех составляющих заявку частей, либо подзявки являются неразличимыми в том смысле, что для сборки заявки необходимы любые K подзаявок и т. д. Более подробный обзор можно найти, например, в работах [20, 23, 191].

Одним из основных показателей производительности любой системы массового обслуживания является ее время отклика. Fork-join система в этом смысле не является исключением, поэтому конкретизируем данное понятие в контексте системы с разделением и параллельным обслуживанием. Поскольку заявка считается обслуженной только в момент окончания обслуживания ее последней подзаявки, то для вычисления времени пребывания заявки в системе R_K 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 $K = 2$ [163], а для $K > 2$ были получены лишь приближения различной степени точности [163,194,195]. 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.

Ниже представлены формулы, оценивающие среднее время отклика в системе с разделением заявок в случае одинаковых интенсивностей обслуживания на всех приборах $\mu_k = \mu$, $k = \overline{1, K}$ и интенсивностью входящего потока равной λ [163, 194, 195].

$$E[R_K] \approx \left[\frac{H_K}{H_2} + \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right)\rho\right]\frac{12 - \rho}{8}\frac{1}{\mu - \lambda}, \tag{2.1}$$

$$E[R_K] \approx \left[H_K + \left(\sum_{i=1}^{K} \binom{K}{i} (-1)^{i-1} \sum_{m=1}^{i} \binom{i}{m} \frac{(m-1)!}{i^{m+1}} - H_K\right) \frac{\lambda}{\mu}\right] \frac{1}{\mu - \lambda}, \quad (2.2)$$

$$E[R_K] \approx \frac{1}{\mu}\left[H_K + \frac{\rho}{2(1-\rho)}\left(\sum_{k=1}^{K} \frac{1}{k-\rho} + (1-2\rho)\sum_{k=1}^{K} \frac{1}{k(k-\rho)}\right)\right], \quad (2.3)$$

$$E[R_K] \approx \frac{1}{\mu - \lambda} H_K, \quad (2.4)$$

$$E[R_K] \approx \frac{1}{\mu - \lambda}\left(1 + \frac{K-1}{\sqrt{(2K-1)}}\right), \quad (2.5)$$

где $H_K = \sum_{i=1}^{K} 1/i$ — это частичная сумма гармонического ряда, а $\rho = \lambda/\mu < 1$ — коэффициент загрузки. Первые две из приведенных формул получены благодаря комбинации аналитического подхода с эмпирическим, заключающейся в наблюдении за поведением исследуемой случайной величины посредством имитационного моделирования и последующем использовании итогов этого наблюдения при составлении соответствующих оценок. Третья формула выведена методом интерполяции высокой и слабой входных нагрузок. Аппроксимация (2.4) является результатом анализа среднего времени отклика для K независимых параллельно функционирующих СМО $M|M|1$, что является не только естественным, но и объективно оправданным допущением, поскольку выражения для маргинальных вероятностей числа подзаявок k-го типа, $k = \overline{1, K}$, в исходной fork-join системе совпадают с выражениями для стационарных вероятностей в системе $M|M|1$ [22]. И, наконец, последнее выражение (2.5) из перечисленных представляет собой верхнюю границу среднего значения максимума порядковых статистик для независимых случайных величин с данными средним и дисперсией.

Помимо оценки величины среднего времени отклика должного внимания заслуживают приближения для дисперсии и для более высоких моментов исследуемой величины, в силу того, что эти числовые характеристики позволяют составить полноценное представление о поведении времени отклика. Однако здесь уже нет такого разнообразия аналитических выражений, а имеющиеся оценки

желательно было бы улучшить. Так, в работе [112] представлено выражение для оценки дисперсии среднего времени отклика как для случая одинаковых параметров экспоненциальных времен обслуживания μ, так и для различных μ_k, $k = \overline{1,K}$ (выведенное в предположении независимости случайных величин)

$$Var[R_K] \approx \sum_{l=1}^{K} \frac{2}{(\mu_l - \lambda)^2} - \sum_{1 \leqslant l < m \leqslant K} \frac{2}{(\mu_l + \mu_m - 2\lambda)^2} + \\ + \sum_{1 \leqslant l < m < k \leqslant K} \frac{2}{(\mu_l + \mu_m + \mu_k - 3\lambda)^2} + \ldots + (-1)^{K-1} \frac{2}{(\mu_1 + \mu_2 + \ldots + \mu_K - K\lambda)^2} - \\ - \left(\sum_{l=1}^{K} \frac{1}{\mu - \lambda} - \sum_{1 \leqslant l < m \leqslant K} \frac{1}{\mu_l + \mu_m - 2\lambda} + \sum_{1 \leqslant l < m < k \leqslant K} \frac{1}{\mu_l + \mu_m + \mu_k - 3\lambda} + \right. \\ \left. + \ldots + (-1)^{K-1} \frac{1}{\mu_1 + \mu_2 + \ldots + \mu_k - K\lambda} \right)^2, \tag{2.6}$$

$$Var[R_K] \approx \frac{2}{(\mu - \lambda)^2} \sum_{i=1}^{K} \binom{K}{i} (-1)^{i-1} \frac{1}{i^2} - \left(\frac{1}{\mu - \lambda} \sum_{i=1}^{K} \frac{1}{i} \right)^2. \tag{2.7}$$

2.2 Оценка основных характеристик системы с помощью ИНС

Для оценки основных характеристик fork-join системы воспользуемся подходом, подробно описанным в Главе 1. Будем оценивать математическое ожидание и среднеквадратическое отклонение времени отклика, т. е. $E[R_K]$ и $\sqrt{Var[R_K]}$, для которых известны приближенные выражения, представленные в предыдущем разделе, чтобы можно было сравнить качество оценок, рассчитываемых с помощью нейросети и известной аналитики.

Для аппроксимации среднего значения и дисперсии времени отклика будем использовать многослойный персептрон. Причем для повышения точности приближения будем обучать не один персептрон с двумя выходными нейронами, соответствующими каждой из указанных характеристик, а два персептрона с одним выходным нейроном в каждом для оценки математического ожидания и среднеквадратического отклонения времени отклика.

Несмотря на достаточность одного скрытого слоя согласно упомянутым ранее теоремам об аппроксимации функций, остановимся на ИНС с двумя скрытыми слоями — опять же с точки зрения повышения точности оценки — с логистической функцией активации на каждом нейроне $\varphi(x) = 1/(1+e^{-x})$ (рис. 16). В качестве входных данных будем использовать коэффициент загрузки ρ и число подзаявок K, $K = 3, ..., 20$, на которое расщепляется поступающая в систему заявка при фиксированном значении $\lambda = 2$ (табл. 7).

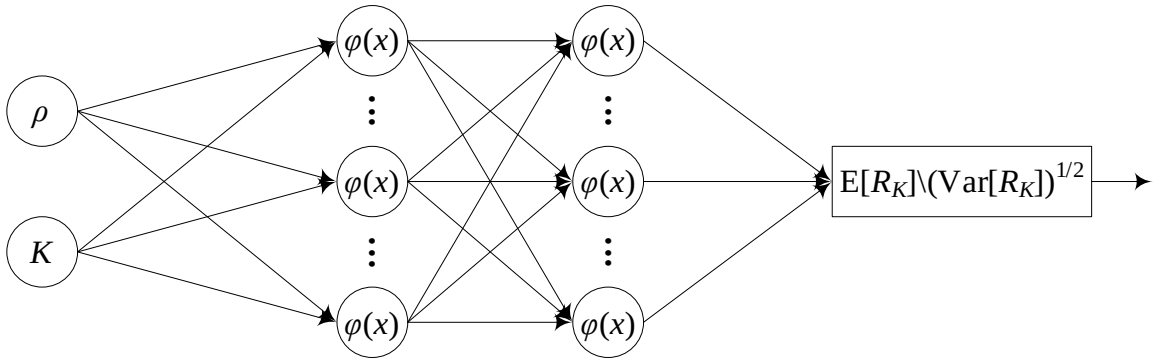


Рис. 16: Схема двухслойного персептрона для определения характеристик системы с разделением и параллельным обслуживанием заявок с K подсистемами типа $M|M|1$.

Таблица 7: Входные данные для расчета характеристик системы с подсистемами типа $M|M|1$

ρ	0.1	0.2	...	0.9	0.1	0.2	...	0.9	0.1	...	0.9
K	3	3	...	3	4	4	...	4	5	...	20

Векторы значений обучающей и тестовой выборок были получены с помощью имитационного моделирования в специализированной программной среде GPSS (англ. General Purpose Simulation System), а построение самой модели нейросети и ее обучение проводилось на языке программирования Python. В процессе обучения ИНС посредством метода обратного распространения ошибки была выбрана наилучшая конфигурация нейросети на основе значений сред-

неквадратической ошибки, средней абсолютной ошибки, а также средней абсолютной процентной ошибки, которые рассчитываются по формулам (1.8),(1.9) и (1.10). Выбор последней из представленных здесь мер, обусловлен еще и тем, что именно она использовалась многими авторами для оценки качества полученных ими приближенных аналитических выражений.

Поскольку наилучшее приближение для среднего времени отклика даёт формула (2.1), в частности, погрешность этой формулы для $K \leqslant 32$ не превышает 5%, то результаты работы ИНС сравниваются именно с результатами применения данной формулы. Так, в таблице 8 для сравнения представлены значения приведенных выше типов ошибок, рассчитанных уже на тестовой выборке в случае вычисления с помощью имитационного моделирования, применения нейросети и формулы (2.1). Как видно, наилучший результат показывает именно обученный персептрон даже несмотря на сравнительно небольшое количество векторов в обучающей выборке, измеряемое сотнями, что, вероятно, в общем случае не способствует улучшению качества аппроксимации, но, тем не менее, полученный результат говорит сам за себя и лишь подтверждает перспективность используемого подхода.

Таблица 8: Сравнение значений ошибок при вычислении среднего времени отклика с помощью формулы (2.1) и ИНС для системы с разделением заявок с подсистемами типа $M|M|1$

Тип ошибки	MSE	MAE	$MAPE$, %
Формула (2.1)	0.022164	0.081634	1.592066
Нейросеть	0.000649	0.015875	0.739039

В случае оценки среднеквадратического отклонения времени отклика известные формулы показывают большую погрешность приближения в отличие от оценки среднего значения этой величины, поэтому ИНС ожидаемо дают лучший результат (табл. 9).

Таблица 9: Сравнение значений ошибок при вычислении среднеквадратического отклонения времени отклика с помощью формулы (2.7) и ИНС для системы с разделением заявок с подсистемами типа $M|M|1$

Тип ошибки	MSE	MAE	$MAPE$, %
Формула (2.7)	0.080494	1.696545	6.896426
Нейросеть	0.000642	0.010495	0.363507

Благодаря использованию нейросети удалось улучшить $MAPE$ для оценки среднего времени отклика примерно в 2.2 раза, а для оценки среднеквадратического отклонения — практически в 19 раз.

2.3 Оценка основных характеристик системы с помощью комплексного метода

В данном разделе приводятся новые аналитические оценки для математического ожидания и дисперсии времени отклика fork-join СМО с K подсистемами $M|M|1$. 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 частности, с помощью имитационной модели, написанной в программной среде Python, были рассчитаны средние значения и дисперсии времени отклика при постоянной интенсивности входящего потока $\lambda = 1$ и меняющемся значении интенсивности обслуживания μ. А именно, коэффициент загрузки $\rho = \lambda/\mu$ принимал значения от 0.1 до 0.9 включительно с шагом 0.05, а число подсистем K менялось от 3 до 20 включительно (поскольку при $K = 2$ известно точное значение среднего времени отклика).

Для оценки качества приближения помимо меры $MAPE$ из (1.10), будем использовать максимальное и минимальное значения модуля относительной по-

грешности приближения

$$MaxAPE = \max_{1 \leqslant j \leqslant N} \left| \frac{\widehat{y}_j - y_j}{y_j} \right| \cdot 100\%, \quad MinAPE = \min_{1 \leqslant j \leqslant N} \left| \frac{\widehat{y}_j - y_j}{y_j} \right| \cdot 100\%,$$

где как и раньше под $\widehat{y}_j$ подразумевается оценка исследуемой характеристики (математического ожидания или дисперсии времени отклика), рассчитанная для j-го набора входных данных, а величина y_j — реальное значение оцениваемой характеристики, полученное в результате имитационного моделирования, N — количество наборов данных в выборке, $j = 1, ..., N$.

Количество испытаний (реализаций с. в.) в рамках одного прогона имитационной модели (для j-го набора входных данных) менялось от 5 млн. при условии низкой загрузки системы до 10 млн. при условии высокой загрузки системы. Далее опишем процесс вывода новых оценок на основе результатов симуляции и к чему он приводит.

При этом сразу стоит отметить, что в случае комплексного метода линейка используемых методов, в том числе интеллектуального анализа данных, значительно расширяется, поскольку, например, наряду с имитационным моделированием используется графический анализ данных, нелинейная регрессия и оптимизация.

Среднее время отклика. Как и ранее, будем обозначать через $E[R_K]$ и $\sqrt{Var[R_k]}$ — математическое ожидание и среднеквадратическое отклонение времени отклика в fork-join СМО с K ветвями типа $M_\lambda|M_\mu|1$. Напомним для удобства, что согласно приближенной формуле Нельсона–Тантави для оценки среднего времени отклика [163], которая дает наименьшую относительную погрешность по сравнению с другими известными формулами [107], не превышающую 5% для $K \leqslant 32$, справедливо

$$E[R_K] \approx E[R_K]_{NT} = \left[\frac{H_K}{H_2} + \frac{4}{11} \left(1 - \frac{H_K}{H_2} \right) \rho \right] \frac{12 - \rho}{8} \frac{1}{\mu - \lambda}, \tag{2.8}$$

где $H_K = \sum_{i=1}^{K} 1/i$ — частичная сумма гармонического ряда.

Поскольку в формуле (2.8) наблюдается зависимость среднего времени отклика от параметров ρ, $(\mu - \lambda)$ и $(H_K/H_2 - 1)$, то естественно предложить, что поправка будет зависеть от тех же параметров. В частности, допустим, что улучшенная оценка для среднего времени отклика имеет следующий вид

$$E[R_K] \approx \frac{\left(\frac{H_K}{H_2} - 1\right)\rho}{\mu - \lambda} \cdot \tilde{\mu} + E[R_K]_{NT}, \tag{2.9}$$

где исходим из того, что оценка Нельсона–Тантави точна при $K = 2$ и асимптотически точна при $\rho \to 0$, а масштабный множитель $1/(\mu - \lambda)$ сохраняется. Далее, чтобы детализировать $\tilde{\mu}$, построим графики для модифицированного выражения среднего времени отклика в зависимости от ρ и $(H_K/H_2 - 1)$. При этом под модифицированным выражением будем понимать отношение

$$\tilde{\mu} = \frac{(E[R_K] - E[R_K]_{NT})(\mu - \lambda)}{(H_K/H_2 - 1)\rho}, \tag{2.10}$$

расчет правой части которого возможен благодаря результатам имитационного моделирования для $E[R_K]$ и проведению вычислений по формуле (2.8) для $E[R_K]_{NT}$.

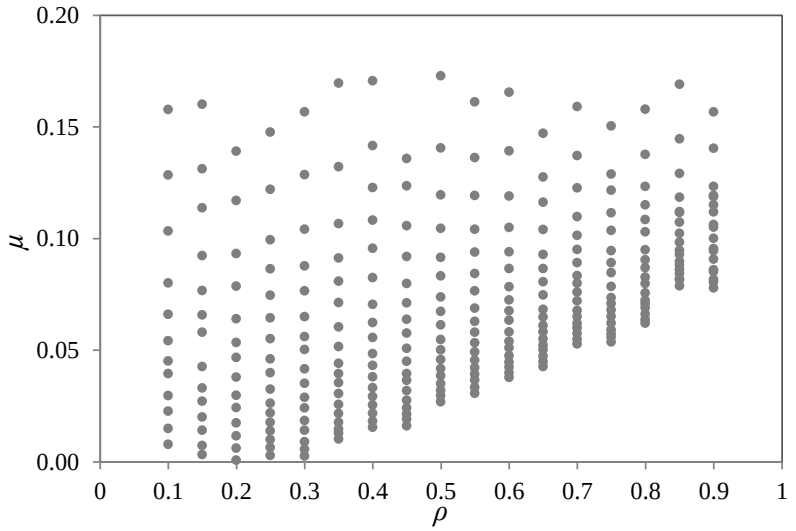


Рис. 17: Значения $\tilde{\mu}$ из (2.10) в зависимости от ρ.

На рисунках 17 и 18 можно уловить линейную зависимость $\tilde{\mu}$ от ρ и $H_K/H_2 - 1$ (хотя и не очень строгую). Таким образом, анализ графиков позво-

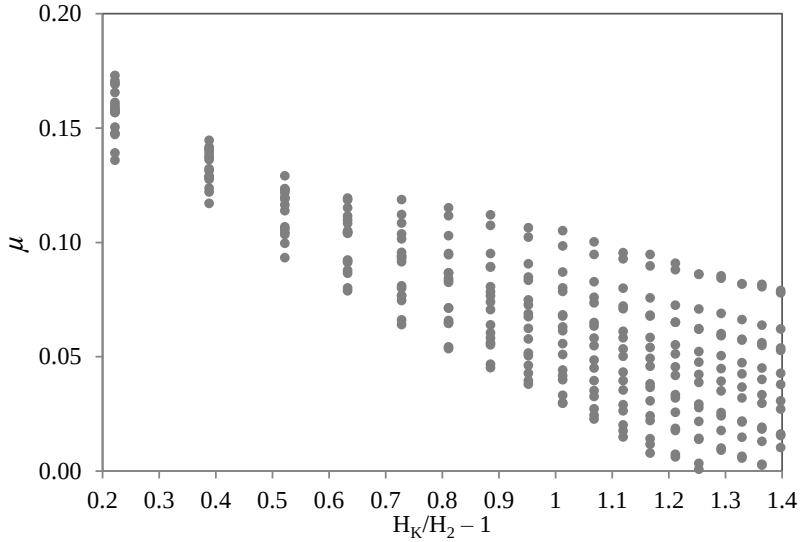


Рис. 18: Значения $\tilde{\mu}$ из (2.10) в зависимости от $H_K/H_2 - 1$.

ляет предложить выражение следующего вида

$$\tilde{\mu} = \tilde{\mu}(\rho, H_K) \approx C_1 - C_2\left(\frac{H_K}{H_2} - 1\right) + C_3\rho, \tag{2.11}$$

т. е.

$$E[R_K] \approx \frac{\left(\frac{H_K}{H_2} - 1\right)\rho}{\mu - \lambda} \cdot \left(C_1 - C_2\left(\frac{H_K}{H_2} - 1\right) + C_3\rho\right) + E[R_K]_{NT}. \tag{2.12}$$

Для определения значений коэффициентов C_i, $i = 1, 2, 3$, воспользуемся методом оптимизации Нелдера–Мида [138, 161]. Идея применения данного метода заключается в минимизации максимального значения модуля относительной ошибки приближения среднего времени отклика выражением (2.12) к его "истинным" значениям, рассчитанным с помощью имитационного моделирования:

$$\max\left|\frac{\hat{E}[R_K] - E[R_K]}{E[R_K]} \cdot 100\%\right| \to \min. \tag{2.13}$$

В результате минимизации ошибки (2.13) на всём множестве данных, полученных посредством симуляции, определяются оптимальные значения коэффициентов C_1, C_2 и C_3, которые наилучшим образом отображают зависимость (2.12) в рамках выбранного метода оптимизации.

В итоге после применения метода оптимизации Нелдера–Мида к (2.13) как

к функции от нескольких переменных C_i в программной среде Python получаем

$$C_1 \approx 0.087197, \quad C_2 \approx 0.070236, \quad C_3 \approx 0.09638. \tag{2.14}$$

В этом случае для указанных выше значений $\lambda = 1$, $\rho \in [0.1, 0.9]$ с шагом 0.05 и $K = 3, ..., 20$ выполняется

$$MaxAPE \approx 0.343207\%, \quad MinAPE \approx 0.001150\%, \quad MAPE \approx 0.142565\%,$$

в то время как для формулы Нельсона–Тантави (2.8) при тех же значениях λ, ρ и K верно

$$MaxAPE \approx 3.944700\%, \quad MinAPE \approx 0.000628\%, \quad MAPE \approx 1.322934\%.$$

Таким образом, введенная поправка улучшает $MaxAPE$ в 11.5 раз, а $MAPE$ в 9.3 раза.

Результаты аппроксимации среднего времени отклика с помощью формулы (2.12) в сравнении с результатами имитационного моделирования представлены на рисунке 19.

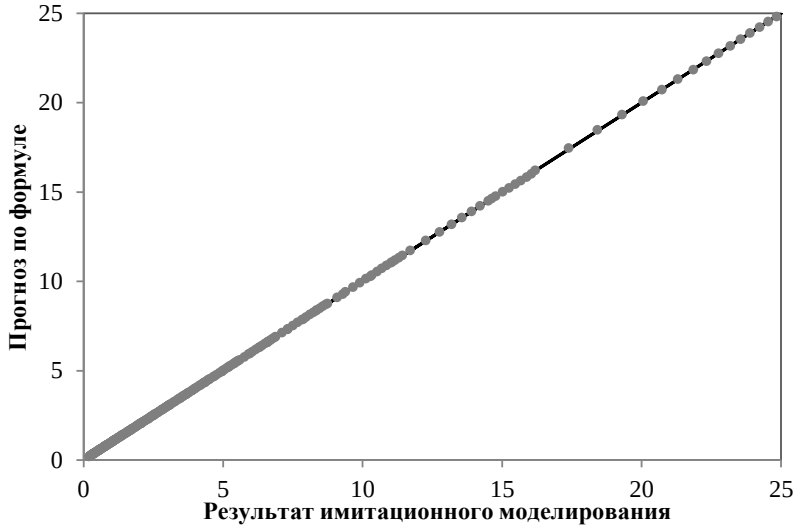


Рис. 19: Среднее время отклика системы с разделением и параллельным обслуживанием с подсистемами $M|M|1$, рассчитанное с помощью (2.12), в сравнении с результатами имитационного моделирования.

Важно отметить, что эффективность выражения (2.12) не ограничивается максимальным значением $K = 20$. Эта формула ожидаемо будет хороша и для большего количества подсистем в силу выстроенной логики предложенного подхода к аппроксимации среднего времени отклика. В частности, для $K = 100$ и значений $\rho \in [0.1, 0.9]$ с шагом 0.05 имеем следующие ошибки приближения:

$$MaxAPE \approx 2.666589\%, \quad MinAPE \approx 0.024029\%, \quad MAPE \approx 0.642155\%.$$

Это лучше формулы Нельсона-Тантави, которая дает

$$MaxAPE \approx 2.741596\%, \quad MinAPE \approx 0.100878\%, \quad MAPE \approx 1.084311\%.$$

Для значения $K = 1000$ в силу экономии вычислительных и временных ресурсов имитационного моделирования, ограничимся проверкой качества аппроксимации в условиях низкой нагрузки $\rho = \{0.1, 0.2, 0.3, 0.4, 0.5\}$, поскольку с увеличением значений не только K, но и ρ необходимо существенно увеличивать количество испытаний (реализаций с. в.) в рамках одного прогона имитационной модели. В результате ошибки приближения формулы (2.12) составили

$$MaxAPE \approx 1.485234\%, \quad MinAPE \approx 0.010079\%, \quad MAPE \approx 0.491061\%,$$

а по формуле Нельсона–Тантави имеем

$$MaxAPE \approx 3.389362\%, \quad MinAPE \approx 1.090899\%, \quad MAPE \approx 2.530636\%.$$

Заметим, что если добавить в формулу (2.11) еще одно слагаемое, т. е.

$$\tilde{\mu} \approx C_1 - C_2\left(\frac{H_K}{H_2} - 1\right) + C_3\rho + C_4\rho\left(\frac{H_K}{H_2} - 1\right), \tag{2.15}$$

и пересчитать оптимальные коэффициенты после подстановки (2.15) в (2.9), которые оказываются равны

$$C_1 \approx 0.150293, \quad C_2 \approx 0.152088, \quad C_3 \approx 0.012684, \quad C_4 \approx 0.105121, \tag{2.16}$$

то получим дальнейшее существенное улучшение оценок при K от 3 до 20, а именно,

$$MaxAPE \approx 0.252453\%, \quad MinAPE \approx 0.000495\%, \quad MAPE \approx 0.105624\%.$$

При $K = 100$ получаем

$$MaxAPE \approx 1.548824\%, \quad MinAPE \approx 0.467368\%, \quad MAPE \approx 1.185327\%.$$

Увеличение $MAPE$ может означать влияние небольшой систематической ошибки, которая становится заметной при больших K. Возможно, корректировка коэффициентов (при сохранении общих формул) с учетом моделирования при $K > 20$ может еще улучшить приближение.

Среднеквадратическое отклонение времени отклика. Для оценки дисперсии времени отклика fork-join СМО с K подсистемами $M_\lambda|M_\mu|1$ в общем случае различных интенсивностей обслуживания μ_i, $1 \leq i \leq K$, в [112] была предложена формула, которая в случае единой интенсивности обслуживания μ принимает вид (2.7). Отметим, что приведенная формула по своему построению исходит из независимости времен пребывания в подсистемах, т. е. речь идет о дисперсии максимума независимых показательных случайных величин. Данное выражение содержит не только суммы, но еще и комбинаторные элементы, что не очень удобно для построения новых оценок, однако оказалось, что его можно упростить.

Итак, формула для оценки дисперсии времени отклика (2.7) эквивалентна выражению

$$Var[R_K] \approx \frac{Q_K}{(\mu - \lambda)^2},$$

где

$$Q_K = \sum_{i=1}^{K} \frac{1}{i^2}.$$

Предположим, что наша оценка будет зависеть, также как и само выражение (2.7), от дисперсии среднего времени отклика в СМО $M|M|1$, т. е. от $1/(\mu-\lambda)^2$,

и от частичной суммы ряда обратных квадратов. Положим

$$Var[R_K] \approx \frac{Q_K - 1}{(\mu - \lambda)^2} \cdot \tilde{v} + \frac{1}{(\mu - \lambda)^2}, \tag{2.17}$$

где будем исходить из того, что оценка точна при $K = 1$. Далее, чтобы конкретизировать выражение для $\tilde{v}$, построим и проанализируем графики зависимости модифицированной дисперсии от ρ, $(Q_K - 1)$ и $(H_K - 1)$. Под модифицированным выражением дисперсии будем понимать следующее:

$$\tilde{v} = \frac{Var[R_K](\mu - \lambda)^2 - 1}{Q_K - 1}, \tag{2.18}$$

где значения $Var[R_K]$ были получены с помощью имитационного моделирования.

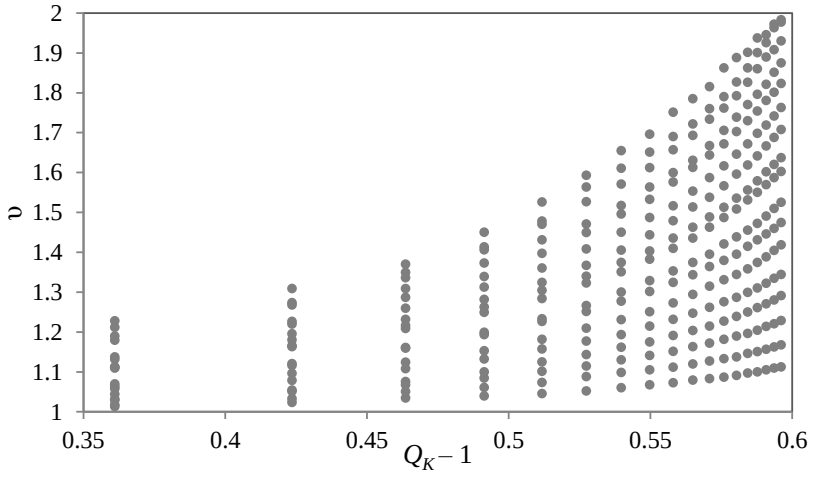


Рис. 20: Значения $\tilde{v}$ из (2.18) в зависимости от $Q_K - 1$.

На рисунках 20 и 21 при фиксированных значениях ρ наблюдается квадратичная зависимость $\tilde{v}$ от $(Q_K - 1)$ и от $(H_K - 1)$, соответственно. Однако анализ экспериментальных данных при больших K показал, что ближе к истине зависимость от $(H_K - 1)$, т. к. фактические значения дисперсии неограниченно растут с ростом K. В противном случае значения дисперсии быстро стабилизировались бы, поскольку, как известно, Q_K имеет конечный предел при $K \to \infty$, равный $\pi^2/6$. Далее, на рисунке 22 наблюдается явная линейная зависимость от ρ с угловым коэффициентом, зависящим от K, причем прямые проходят через

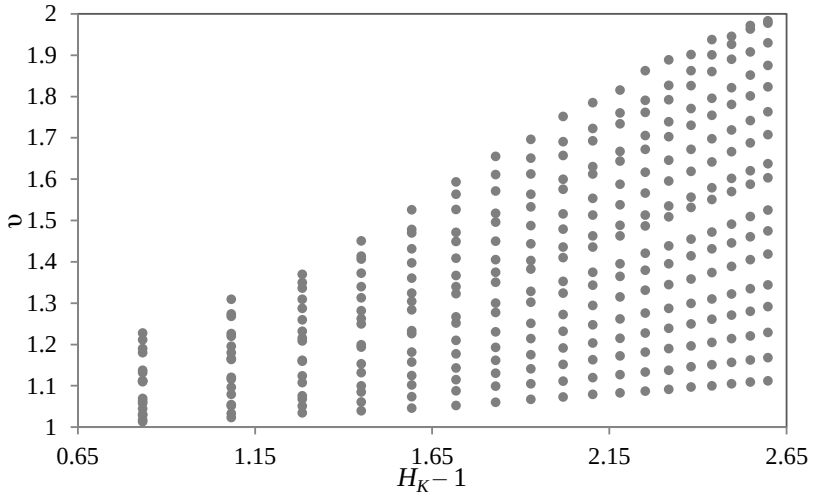


Рис. 21: Значения $\tilde{v}$ из (2.18) в зависимости от $H_K - 1$.

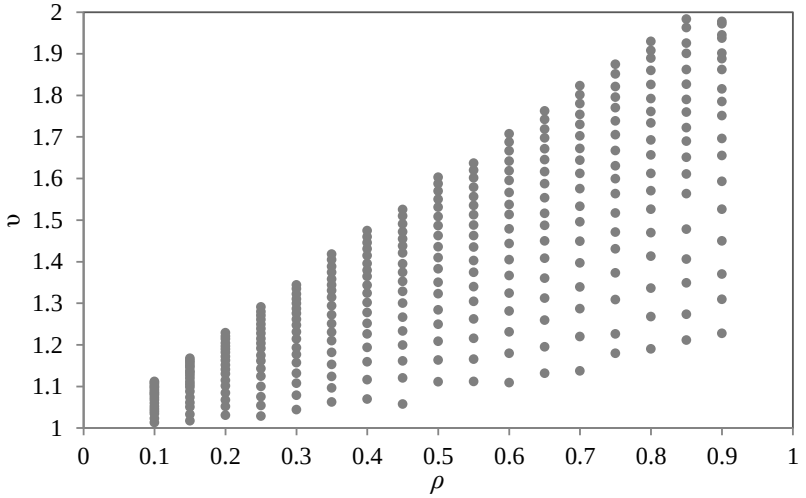


Рис. 22: Значения $\tilde{v}$ из (2.18) в зависимости от ρ.

точку $(0, 1)$. Поэтому предположим, что

$$\tilde{v} = \tilde{v}(\rho, H_K) \approx 1 + \rho\big(C_1 + C_2(H_K - 1) + C_3(H_K - 1)^2\big) \tag{2.19}$$

и, соответственно,

$$Var[R_K] \approx \frac{Q_K - 1}{(\mu - \lambda)^2} \cdot \big(1 + \rho\big(C_1 + C_2(H_K - 1) + C_3(H_K - 1)^2\big)\big) + \frac{1}{(\mu - \lambda)^2}, \tag{2.20}$$

Затем аналогично случаю с математическим ожиданием с помощью метода Нелдера–Мида решаем оптимизационную задачу по минимизации максимальной ошибки аппроксимации формулы (2.20)

$$\max \left| \frac{\hat{V}ar[R_K] - Var[R_K]}{Var[R_K]} \cdot 100\% \right| \to \min . \tag{2.21}$$

В результате получаем следующие значения коэффициентов C_i

$$C_1 \approx -0.113658, \quad C_2 \approx 0.339780, \quad C_3 \approx 0.053745. \tag{2.22}$$

На рисунке 23 для наглядности качества полученного приближения изобра-

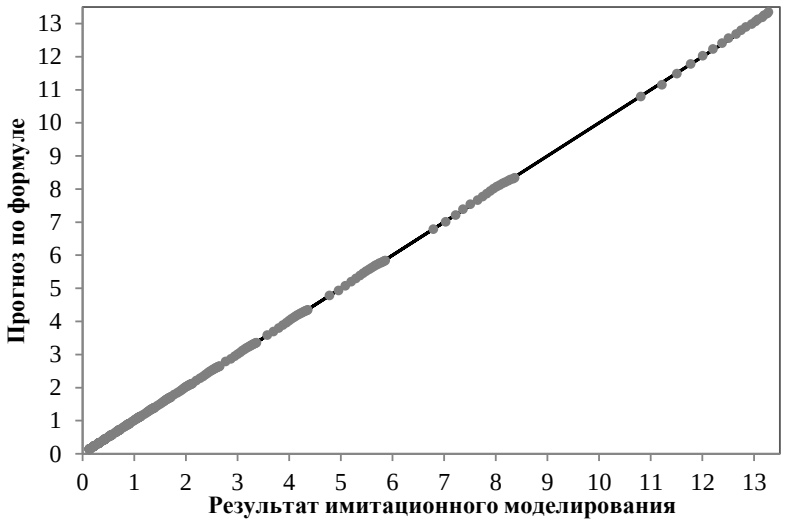


Рис. 23: Среднеквадратическое отклонение времени отклика системы с разделением и параллельным обслуживанием с подсистемами $M|M|1$, рассчитанное с помощью (2.20), в сравнении с результатами имитационного моделирования.

жены результаты, полученные с помощью формулы (2.20) и с помощью имитационного моделирования. Для среднеквадратического времени отклика при значениях $\lambda = 1$, $\rho \in \{0.1, 0.15, 0.20, ..., 0.90\}$ и $K = 3, ..., 20$ справедливо

$$MaxAPE \approx 0.564247\%, \quad MinAPE \approx 0.000755\%, \quad MAPE \approx 0.188812\%,$$

в то время как для формулы (2.7) для тех же значений λ, ρ и K верно

$$MaxAPE \approx 14.475134\%, \quad MinAPE \approx 0.173239\%, \quad MAPE \approx 6.055298\%.$$

Таким образом, новая оценка улучшает $MaxAPE$ в 25.7 раз, а $MAPE$ в 32.1 раза.

Также как и в случае с математическим ожиданием времени отклика, формула для среднеквадратического отклонения времени отклика оказывается хороша не только для максимального значения $K = 20$, но и гораздо большего

числа подсистем. Так, для $K = 100$ и значений $\rho \in [0.1, 0.9]$ с шагом 0.05 имеем следующие ошибки приближения:

$$MaxAPE \approx 1.211417\%, \quad MinAPE \approx 0.094622\%, \quad MAPE \approx 0.724898\%.$$

Для значения $K = 1000$ по аналогичным причинам, что и в случае с формулой для математического ожидания, приведём результаты погрешности аппроксимации только в условиях низкой загрузки системы $\rho = \{0.1, 0.2, 0.3, 0.4, 0.5\}$:

$$MaxAPE \approx 1.029874\%, \quad MinAPE \approx 0.203567\%, \quad MAPE \approx 0.592974\%.$$

В результате, имеет место очень хорошее приближение.

2.4 Коэффициенты корреляции в системе с разделением и параллельным обслуживанием

Основная причина сложности проведения анализа fork-join СМО заключается в имеющейся зависимости между временами пребывания подзаявок в подсистемах. Эта зависимость возникает из-за общности момента появления подзаявок в самой системе, поскольку они являются составными элементами одной заявки, которая разделяется на части в момент поступления в систему. Зависимость между временами пребывания подзаявок 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. Тем не менее, исследования в этом направлении практически отсутствуют и найти публикации, в которых проводился бы подобный анализ даже в случае подсистем типа $M|M|1$, не удалось.

Однако, как оказалось, можно получить даже не оценки коэффициентов корреляции Пирсона и Спирмена, а точные формулы для них, также было получено приближение для коэффициента корреляции Кендалла.

В разделе приводится подробное описание подхода к выводу выражений для коэффициентов корреляции между временами пребывания пары подзаявок в соответствующих подсистемах. Сам по себе подход для вывода корреляций Пирсона и Спирмена является классическим и основывается на теории производящих функций и преобразованиях Лапласа–Стилтьеса, при этом рассматривается система, для которой число подсистем равно двум ($K = 2$). Для оценки корреляции Кендалла используется комбинированный подход, который включает в себя метод оптимизации Нелдера–Мида и графический анализ.

Сразу стоит отметить, что на значение коэффициентов корреляции между любой парой подзаявок не влияет количество подсистем, т. е. значение коэффициента корреляции не зависит от K и будет одинаковым для любой парной комбинации подзаявок из двух разных подсистем при фиксированных параметрах входящего потока и времени обслуживания, а следовательно, верно и для всех $K > 2$.

Обнаружен интересный феномен: при низкой загрузке оказывается больше коэффициент корреляции Пирсона, а при высокой — Спирмена. При этом коэффициент корреляции Кендалла всегда меньше их обоих.

Производящая функция. Итак, перейдем к рассмотрению случая, когда $K = 2$, т. к. на попарную зависимость времен пребываний в подсистемах общее количество подсистем не оказывает никакого влияния. Следовательно, случайное время пребывания заявки в СМО (или время отклика) системы является максимумом из двух случайных времен пребывания подзаявок ξ_i, $i = 1, 2$, в соответствующих подсистемах

$$R_2 = R = \max\{\xi_1, \xi_2\}.$$

Функционирование системы описывается марковским процессом

$$S(t) = \{S_1(t), S_2(t)\},$$

где $S_i(t)$ — число подзаявок в i-й подсистеме, $i = 1, 2$. Тогда множество состояний процесса можем записать следующим образом: $S = \{(i, j), i \geqslant 0, j \geqslant 0\}$. Далее пусть p_{ij} — стационарная вероятность того, что в первой подсистеме находится i заявок, а во второй подсистеме — j заявок. Условие эргодичности системы стандартное и общее для обеих подсистем массового обслуживания, т. е. $\rho = \lambda/\mu < 1$.

Теперь введем производящую функцию для числа подзаявок в системе

$$P(z, w) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} z^i w^j p_{ij}. \tag{2.23}$$

В соответствии с [96] она примет вид

$$P(z, w) = \frac{N(z, w)}{Q(z, w)}, \tag{2.24}$$

где при $\lambda = 1$ (что без ограничения общности будем полагать везде далее)

$$N(z, w) = \mu z(w - 1)P(z, 0) + \mu w(z - 1)P(0, w),$$

$$Q(z, w) = (1 + 2\mu)zw - \mu w - \mu z - z^2 w^2,$$

$$P(z, 0) = \frac{(\mu - 1)^{3/2}}{\mu(\mu - z)^{1/2}} = \frac{(1 - \rho)^{3/2}}{(1 - \rho z)^{1/2}}, \quad P(0, w) = \frac{(1 - \rho)^{3/2}}{(1 - \rho w)^{1/2}},$$

тогда

$$P(z, w) = \frac{z(w - 1)P(z, 0) + w(z - 1)P(0, w)}{(2 + \rho)zw - w - z - \rho z^2 w^2}. \tag{2.25}$$

В следующих разделах перейдем к описанию последовательности действий, необходимых для вычисления коэффициентов корреляции нескольких видов.

Коэффициент корреляции Пирсона. Зависимость между временами пребывания подзаявок в подсистемах fork-join системы с параллельным обслуживанием заявок возникает в силу общих для них моментов поступления в эти подсистемы. Флуктуации входного потока в большую или меньшую сторону (по числу поступлений за какое-то время) приводят к увеличению или уменьшению

длины очередей в подсистемах и, соответственно, увеличению или уменьшению времен пребывания подзаявок от одной заявки в подсистемах.

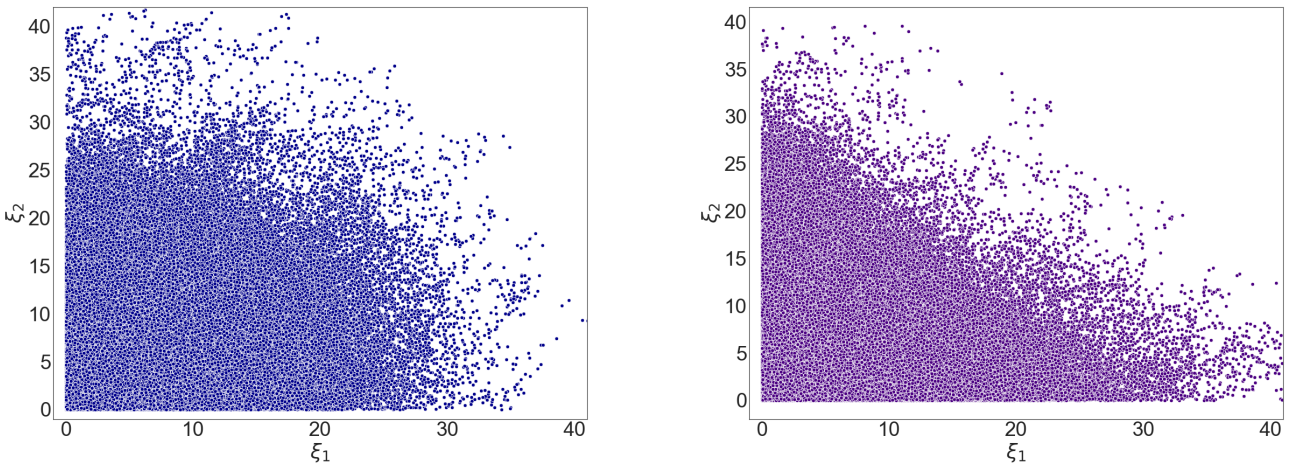
Визуализация данных позволяет продемонстрировать зависимость времен пребывания в подсистемах. Так, на рис. 24 представлены значения ξ_1 и ξ_2, полученные с помощью имитационного моделирования в случае fork-join СМО и для случая двух параллельно функционирующих СМО $M|M|1$ с идентичными значениями параметров λ и μ. Количество пар точек (ξ_1, ξ_2) в обоих случаях одинаково и составляет один миллион. Разумеется, для получения хорошей оценки коэффициента корреляции такого количества точек недостаточно, но для иллюстрации наличия разницы в поведении случайных величин для двух вариантов функционирования систем вполне приемлемо.

Из теории понятно и наглядно видно, что для независимых времен пребывания (рис. 24б) линии уровня совместной плотности имеют вид $x_1 + x_2 = const$ (поскольку частные распределения показательные) , в то время как для зависимых времен пребывания (рис. 24а) видно, что линии имеют выпуклость от начала координат, которая оказывается тем больше, чем больше загрузка. Это отражает большую вероятность совместно больших значений, чем при независимости.

Заметим, что визуально характер зависимости не похож на классический в статистике случай, когда имеется функциональная зависимость величин (линейная или монотонная), на которую накладывается случайный шум. Поэтому возникает вопрос, как известные коэффициенты корреляции сумеют уловить (отразить) эту зависимость.

Прежде чем формулировать теорему, касающуюся аналитического выражения для коэффициента корреляции Пирсона, докажем несколько вспомогательных утверждений.

Для вычисления коэффициента корреляции Пирсона будем пользоваться классическими для теории массового обслуживания инструментами. Рассмотрим преобразование Лапласа–Стилтьеса для времен пребывания подзаявок в



а) б)

Рис. 24: Иллюстрация наличия/отсутствия зависимости между случайными величинами ξ_1 и ξ_2 при $\rho = 0.8$ в случае а) fork-join СМО с двумя подсистемами $M|M|1$; б) двух параллельно функционирующих СМО $M|M|1$.

подсистемах.

Лемма 2.1. *Преобразование Лапласа–Стилтьеса (ПЛС) времен пребывания подзаявок в подсистемах системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ и экспоненциальными временами обслуживания с параметром μ можно представить в виде*

$$\varphi(s,t) = P\left(\frac{1}{1+\rho s}, \frac{1}{1+\rho t}\right) \cdot \frac{1}{1+\rho s}\frac{1}{1+\rho t}. \tag{2.26}$$

где $P(\cdot,\cdot)$ — это производящая функция из (2.23)*, $\rho = \lambda/\mu < 1$.*

Доказательство. ПЛС для времен пребывания подзаявок в подсистемах определяется выражением

$$\varphi(s,t) = \int_0^\infty \int_0^\infty e^{-sx} e^{-ty} v(x,y)\,dx\,dy, \tag{2.27}$$

где

$$v(x,y) = \frac{\partial^2 V(x,y)}{\partial x \partial y}$$

это двумерная плотность распределения времен пребывания подзаявок в подсистемах, $V(x,y) = P(\xi_1 < x, \xi_2 < y)$.

Далее проводим стандартные рассуждения. Если при поступлении в систему заявка застает i подзаявок в первой СМО и j подзаявок во второй СМО, то время ожидания в очереди каждого из двух составляющих ее родственных элементов в соответствующих подсистемах будет состоять из суммы случайных времен обслуживания всех стоящих перед ними в очереди i или j подзаявок, соответственно, а также времени обслуживания самого вновь поступившего элемента. С учетом того, что время обслуживания одной подзаявки имеет экспоненциальное распределение с параметром μ, то сумма, состоящая из $(i+1)$ и $(j+1)$ таких случайных величин будет иметь распределение Эрланга с функцией распределения $E_{i+1}(x)$ и $E_{j+1}(x)$. Тогда двумерная функция распределения времени пребывания подзаявок в подсистемах примет вид

$$V(x,y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} E_{i+1}(x)E_{j+1}(y)p_{ij}. \tag{2.28}$$

Пользуясь свойствами ПЛС, с учетом выражения (2.28) получаем, что

$$\varphi(s,t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \varepsilon_{i+1}(s)\varepsilon_{j+1}(t)p_{ij} = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \left(\frac{\mu}{\mu+s}\right)^{i+1}\left(\frac{\mu}{\mu+t}\right)^{j+1} p_{ij},$$

где $\varepsilon_{i+1}(s)$ и $\varepsilon_{j+1}(t)$ — ПЛС для $E_{i+1}(x)$. Таким образом, ПЛС времен пребывания подзаявок в подсистемах (2.27) можно представить в виде

$$\varphi(s,t) = P\left(\frac{\mu}{\mu+s}, \frac{\mu}{\mu+t}\right) \cdot \frac{\mu}{\mu+s} \cdot \frac{\mu}{\mu+t}, \tag{2.29}$$

где $P(\cdot,\cdot)$ — это производящая функция из (2.23).

С учетом $\lambda = 1$ имеем

$$\varphi(s,t) = P\left(\frac{1}{1+\rho s}, \frac{1}{1+\rho t}\right) \cdot \frac{1}{1+\rho s}\frac{1}{1+\rho t}.$$

□

Лемма 2.2. *Математическое ожидание произведения случайных величин* ξ_1 *и* ξ_2 *времен пребывания подзаявок в подсистемах системы с разделением и параллельным обслуживанием с подсистемами типа* $M_\lambda|M_\mu|1$ *определяется выражением*

$$E[\xi_1 \cdot \xi_2] = \frac{\rho^2(4\rho - \rho^2 + 8)}{8(1-\rho)^2}, \tag{2.30}$$

Доказательство. Вычислить $E[\xi_1 \cdot \xi_2]$ можно с помощью ПЛС $\varphi(s,t)$, учитывая лемму 2.1 имеем

$$E[\xi_1 \cdot \xi_2] = \int_0^\infty \int_0^\infty xy \cdot v(x,y)dxdy = \frac{\partial^2 \varphi(s,t)}{\partial s \partial t}\bigg|_{s=0,t=0} =$$

$$= \frac{\partial^2 \left[P\left(\frac{1}{1+\rho s}, \frac{1}{1+\rho t}\right) \cdot \frac{1}{1+\rho s}\frac{1}{1+\rho t}\right]}{\partial s \partial t}\bigg|_{s=0,t=0},$$

при этом

$$P\left(\frac{1}{1+\rho s}, \frac{1}{1+\rho t}\right) \cdot \frac{1}{1+\rho s}\frac{1}{1+\rho t} =$$

$$= \frac{(1-\rho)^{3/2}\left(t\frac{\sqrt{\rho s+1}}{\sqrt{\rho s-\rho+1}} + s\frac{\sqrt{\rho t+1}}{\sqrt{\rho t-\rho+1}}\right)}{\rho^2 s^2 t + \rho^2 s t^2 - \rho^2 s t + \rho s^2 + 2\rho s t - \rho s + \rho t^2 - \rho t + s + t}.$$

Если ввести следующие обозначения

$$F(s) = \frac{(1-\rho)^{3/2}\sqrt{1+\rho s}}{\sqrt{1+\rho s - \rho}},$$

$$G(s,t) = \rho^2 s^2 t + \rho^2 s t^2 - \rho^2 s t + \rho s^2 + 2\rho s t - \rho s + \rho t^2 - \rho t + s + t,$$

то формулу можно записать в более компактном виде

$$P\left(\frac{1}{1+\rho s}, \frac{1}{1+\rho t}\right) \cdot \frac{1}{1+\rho s}\frac{1}{1+\rho t} = \frac{tF(s) + sF(t)}{G(s,t)}.$$

Взятие производных приведет к следующему выражению

$$\frac{\partial^2}{\partial s \partial t}\left[\frac{tF(s) + sF(t)}{G(s,t)}\right] = \frac{f(s,t)}{g(s,t)},$$

где

$$f(s,t) = G^2(s,t)[F'(s) + F'(t)] +$$

$$+G'_s(s,t)G'_t(s,t)[2tF(s)+2sF(t)]-G'_s(s,t)G(s,t)[F(s)+sF'(t)]-$$

$$-G'_t(s,t)G(s,t)[F(t)+tF'(s)]-G''_{st}(s,t)G(s,t)[tF(s)+sF(t)],$$

$$g(s,t)=G^3(s,t).$$

Далее необходимо вычислить

$$\left.\frac{\partial^2\varphi(s,t)}{\partial s\partial t}\right|_{s=0,t=0}=\lim_{\substack{s\to0\\t\to0}}\frac{\partial^2\varphi(s,t)}{\partial s\partial t}=\lim_{\substack{s\to0\\t\to0}}\frac{f(s,t)}{g(s,t)}.$$

Причем заметим, что функции $f(s,t)$ и $g(s,t)$ являются бесконечно малыми третьего порядка при $(s,t)\to(0,0)$, т. е.

$$\lim_{\substack{s\to0\\t\to0}}f(s,t)=\lim_{\substack{s\to0\\t\to0}}g(s,t)=0,$$

$$f'_s(0,0)=f'_t(0,0)=g'_s(0,0)=g'_t(0,0)=0,$$

$$f''_{ss}(0,0)=f''_{tt}(0,0)=f''_{st}(0,0)=g''_{ss}(0,0)=g''_{tt}(0,0)=g''_{st}(0,0)=0,$$

$$f^{(3)}_{s^{n_1}t^{n_2}}(0,0)\neq0,\quad g^{(3)}_{s^{n_1}t^{n_2}}(0,0)\neq0,\quad n_1+n_2=3.$$

Поэтому для существования двойного предела в точке $(0,0)$ необходимо и достаточно выполнения равенства [31, 32]

$$\frac{f^{(3)}_{s^{n_1}t^{n_2}}(0,0)}{g^{(3)}_{s^{n_1}t^{n_2}}(0,0)}=m,\quad m\neq0,m\neq\pm\infty,\quad n_1+n_2=3. \tag{2.31}$$

причем сам двойной предел будет равен m из (2.31)

$$\lim_{\substack{s\to0\\t\to0}}\frac{f(s,t)}{g(s,t)}=m.$$

После соответствующих вычислений получаем, что

$$f^{(3)}_{s^{n_1}t^{n_2}}(0,0)=\frac{3(1-\rho)\rho^2(4\rho-\rho^2+8)}{4},\quad g^{(3)}_{s^{n_1}t^{n_2}}(0,0)=6(1-\rho)^3,\quad n_1+n_2=3.$$

Таким образом, имеем

$$\left.\frac{\partial^2\varphi(s,t)}{\partial s\partial t}\right|_{s=0,t=0}=\lim_{\substack{s\to0\\t\to0}}\frac{\partial^2\varphi(s,t)}{\partial s\partial t}=\frac{f^{(3)}_{s^{n_1}t^{n_2}}(0,0)}{g^{(3)}_{s^{n_1}t^{n_2}}(0,0)}=\frac{\rho^2(4\rho-\rho^2+8)}{8(1-\rho)^2},$$

□

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*

$$r_p = \frac{\rho(4-\rho)}{8}. \tag{2.32}$$

Доказательство. Коэффициент корреляции (Пирсона) между временами пребывания подзаявок в подсистемах определяется выражением

$$r_p = \frac{E[\xi_1 \cdot \xi_2] - E[\xi_1]E[\xi_2]}{\sqrt{Var[\xi_1]Var[\xi_2]}}. \tag{2.33}$$

С учетом того, что, как известно, время пребывания заявки в системе типа $M_\lambda|M_\mu|1$ имеет экспоненциальное распределение с параметром $(\mu-\lambda)$, т. е.

$$E[\xi_1] = E[\xi_2] = \frac{1}{\mu-\lambda}, \quad Var[\xi_1] = Var[\xi_2] = \frac{1}{(\mu-\lambda)^2},$$

выражение (2.33) при $\lambda = 1$ преобразуется к виду

$$r_p = (\mu-1)^2 E[\xi_1 \cdot \xi_2] - 1 = \mu^2(1-\rho)^2 E[\xi_1 \cdot \xi_2] - 1. \tag{2.34}$$

Поэтому, чтобы определить коэффициент корреляции, необходимо вычислить $E[\xi_1 \cdot \xi_2]$. В соответствии с леммой 2.2 имеем, что

$$E[\xi_1 \cdot \xi_2] = \frac{\rho^2(4\rho - \rho^2 + 8)}{8(1-\rho)^2},$$

тогда

$$r_p = \mu^2(1-\rho)^2 \frac{\rho^2(4\rho-\rho^2+8)}{8(1-\rho)^2} - 1 = \frac{\rho(4-\rho)}{8}.$$

□

Справедливость формулы (2.32) для коэффициента корреляции Пирсона между временами пребывания двух любых подзаявок в соответствующих под-

системах подтверждается численным экспериментом. С помощью имитационного моделирования для значений $\lambda = 1$, $\rho \in [0.1, 0.9]$ с шагом 0.05 были рассчитаны значения r_p. Результаты сравнения с аналитическим выражением (2.32) представлены на рисунке 25. Небольшие отклонения в области значений ρ, близ-

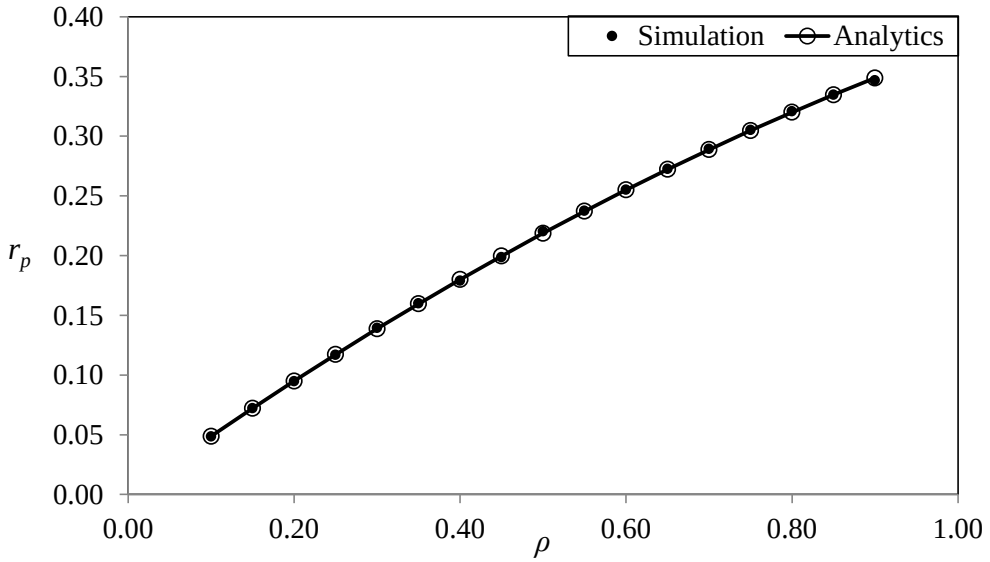


Рис. 25: Коэффициент корреляции Пирсона r_p.

ких к единице, объясняются следующим образом. Увеличение коэффициента загрузки системы требует значительного увеличения длительности прогона в рамках одного запуска имитационной модели, поскольку рост коэффициента загрузки приводит к росту корреляции данных, используемых для построения точечной оценки исследуемой величины.[1] Увеличение длительности прогона, в свою очередь, приводит к значительному увеличению времени, затрачиваемому на проведение численного эксперимента, что не является в данном случае рациональным, поскольку данные совпадают вплоть до третьей цифры после запятой, а модуль максимальной относительной погрешности не превышает 0.62%.

Исходя из полученных результатов можно сделать вывод, что зависимость между временами пребывания в подсистемах ξ_1 и ξ_2 хорошо отражается коэффициентом корреляции Пирсона, причем он возрастает с увеличением загрузки

[1] С особенностями проведения имитационного моделирования fork-join СМО и построении доверительных интервалов полученных оценок можно ознакомиться в последнем разделе данной главы.

ρ квадратичным образом (с замедлением).

Коэффициент корреляции Спирмена. Для составления полноценного представления о зависимости между случайными временами пребывания подзаявок в подсистемах ξ_1 и ξ_2 проанализируем коэффициент корреляции Спирмена.

В статистике коэффициент корреляции Спирмена вычисляется как коэффициент корреляции Пирсона применительно к рангам наблюдений (по возрастанию) в качестве характеристики силы монотонной зависимости (возрастающей или убывающей) между случайными величинами, обычно для проверки гипотез об их независимости.

С точки зрения теории вероятностей, для непрерывных случайных величин коэффициент корреляции Спирмена удобно определить следующим образом. Пусть случайные величины X_1 и X_2 имеют функции распределения F_1 и F_2, положим $U_1 = F_1(X_1)$ и $U_2 = F_2(U_2)$, тогда коэффициент корреляции Спирмена X_1 и X_2 можно выразить как коэффициент корреляции Пирсона случайных величин U_1 и U_2 [162, с. 170]: $r_s(X_1, X_2) = r_p(U_1, U_2)$.

Коэффициент r_s, также как и r_p, принимает значения на отрезке $[-1, 1]$. Для определения коэффициента r_s проведем некоторые подготовительные выкладки.

Лемма 2.3. *Преобразование Лапласа–Стилтьеса $\varphi(s,t)$ времен пребывания подзаявок в любых двух подсистемах системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ и экспоненциальными временами обслуживания с параметром μ после умножении каждого из аргументов на величину, обратную среднему времени пребывания подзаявки в подсистеме, т. е. $(\mu - \lambda)$, можно представить в виде*

$$\varphi((\mu-\lambda)s, (\mu-\lambda)t) = \\ = \frac{t(1+(1-\rho)s)^{1/2}(s+1)^{-1/2} + s(1+(1-\rho)t)^{1/2}(t+1)^{-1/2}}{(s+t) - \rho st + (s+t)^2 + (1-\rho)(s+t)st}. \qquad (2.35)$$

Доказательство. Рассмотрим $\varphi(s,t)$ из (2.29), где s и t умножим на $(\mu-\lambda)$, тогда с учётом того, что $\rho=\lambda/\mu$, получим

$$\varphi((\mu-\lambda)s,(\mu-\lambda)t)=$$

$$=P(\frac{\mu}{\mu+(\mu-\lambda)s},\frac{\mu}{\mu+(\mu-\lambda)t})\cdot\frac{\mu}{\mu+(\mu-\lambda)s}\cdot\frac{\mu}{\mu+(\mu-\lambda)t}=$$

$$=P(\frac{1}{1+(1-\rho)s},\frac{1}{1+(1-\rho)t})\cdot\frac{1}{1+(1-\rho)s}\cdot\frac{1}{1+(1-\rho)t}.$$

Теперь пусть $\lambda=1$, тогда подставим соответствующее выражение для производящей функции. Получим

$$\varphi((\mu-\lambda)s,(\mu-\lambda)t)=$$

$$=\frac{\frac{1}{1+(1-\rho)s}\left(\frac{1}{1+(1-\rho)t}-1\right)\cdot\frac{(1-\rho)^{3/2}}{(1-\frac{\rho}{1+(1-\rho)s})^{1/2}}+\frac{1}{1+(1-\rho)t}\left(\frac{1}{1+(1-\rho)s}-1\right)\cdot\frac{(1-\rho)^{3/2}}{(1-\frac{\rho}{1+(1-\rho)t})^{1/2}}}{(2+\rho)\frac{1}{1+(1-\rho)s}\cdot\frac{1}{1+(1-\rho)t}-\frac{1}{1+(1-\rho)s}-\frac{1}{1+(1-\rho)t}-\frac{\rho}{(1+(1-\rho)s)^2(1+(1-\rho)t)^2}}\cdot$$

$$\cdot\frac{1}{1+(1-\rho)s}\cdot\frac{1}{1+(1-\rho)t}=\frac{A(\rho)}{B(\rho)}\cdot\frac{1}{1+(1-\rho)s}\cdot\frac{1}{1+(1-\rho)t}.$$

Упростим числитель $A(\rho)$

$$A(\rho)=\frac{1}{1+(1-\rho)s}\cdot\frac{1-1-(1-\rho)t}{1+(1-\rho)t}\cdot\frac{(1-\rho)^{3/2}}{\left(\frac{1+(1-\rho)s-\rho}{1+(1-\rho)s}\right)^{1/2}}+$$

$$+\frac{1}{1+(1-\rho)t}\cdot\frac{1-1-(1-\rho)s}{1+(1-\rho)s}\cdot\frac{(1-\rho)^{3/2}}{\left(\frac{1+(1-\rho)t-\rho}{1+(1-\rho)t}\right)^{1/2}}=$$

$$=\frac{-(1-\rho)^2}{(1+(1-\rho)s)(1+(1-\rho)t)}\cdot\frac{t(1+(1-\rho)s)^{1/2}}{(s+1)^{1/2}}+$$

$$+\frac{-(1-\rho)^2}{(1+(1-\rho)s)(1+(1-\rho)t)}\cdot\frac{s(1+(1-\rho)t)^{1/2}}{(t+1)^{1/2}}=$$

$$=\frac{-(1-\rho)^2\cdot C(\rho)}{(1+(1-\rho)s)(1+(1-\rho)t)},$$

где

$$C(\rho)=\frac{t(1+(1-\rho)s)^{1/2}}{(s+1)^{1/2}}+\frac{s(1+(1-\rho)t)^{1/2}}{(t+1)^{1/2}}. \tag{2.36}$$

Упростим знаменатель $B(\rho)$ с учетом дополнительного множителя

$$B(\rho)\cdot(1+(1-\rho)s)(1+(1-\rho)t)=$$

$$= 2 + \rho - (1 + (1-\rho)s) - (1 + (1-\rho)t) - \frac{\rho}{(1 + (1-\rho)s)(1 + (1-\rho)t)} =$$

$$= \rho - (1-\rho)(s+t) - \frac{\rho}{(1 + (1-\rho)s)(1 + (1-\rho)t)} =$$

$$\frac{\rho(1-\rho)(s+t) + \rho(1-\rho)^2 st - (1-\rho)(s+t) - (1-\rho)^2(s+t)^2 - (1-\rho)^3(s+t)st}{(1 + (1-\rho)s)(1 + (1-\rho)t)} =$$

$$= \frac{-(1-\rho)^2 \cdot \big((s+t) - \rho st + (s+t)^2 + (1-\rho)(s+t)st\big)}{(1 + (1-\rho)s)(1 + (1-\rho)t)} =$$

$$= \frac{-(1-\rho)^2 \cdot D(\rho)}{(1 + (1-\rho)s)(1 + (1-\rho)t)},$$

где

$$D(\rho) = (s+t) - \rho st + (s+t)^2 + (1-\rho)(s+t)st. \tag{2.37}$$

Тогда

$$\varphi((\mu-\lambda)s, (\mu-\lambda)t) = \frac{C(\rho)}{D(\rho)},$$

где C и D определены в (2.36) и (2.37), т. е.

$$\varphi((\mu-\lambda)s, (\mu-\lambda)t) = \frac{t(1 + (1-\rho)s)^{1/2}(s+1)^{-1/2} + s(1 + (1-\rho)t)^{1/2}(t+1)^{-1/2}}{(s+t) - \rho st + (s+t)^2 + (1-\rho)(s+t)st}.$$

□

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*

$$r_s = \frac{12\sqrt{2}\sqrt{2-\rho}}{8-3\rho} - 3. \tag{2.38}$$

Доказательство. Напомним, что времена пребывания подзаявок в подсистемах fork-join СМО имеют экспоненциальное распределение с параметром $\mu - \lambda$, т. е. $\xi_i \sim Exp(\mu - \lambda)$. Рассмотрим случайные величины $U_i = F(\xi_i)$, где $F_{\xi_i}(x) = 1 - e^{-(\mu-\lambda)x}$, $x > 0$.

Согласно методу обратного преобразования (преобразования Н. В. Смирнова) известно, что если функция распределения F_ξ некоторой случайной величины ξ имеет обратную функцию (строго возрастает на всей области определения), то $F^{-1}(U) = \xi$, где U — это равномерно распределенная случайная величина на отрезке $[0, 1]$.

Таким образом, случайные величины $U_i = F(\xi_i)$ будут иметь равномерное распределение на отрезке $[0, 1]$, т. е. $U_i \sim R[0, 1]$, $i = 1, 2$. Тогда r_s — коэффициент корреляции Спирмена для случайных величин ξ_1 и ξ_2 будет являться коэффициентом корреляции Пирсона для случайных величин U_1 и U_2 [162, с. 170], т. е.

$$r_s = \frac{E[U_1 \cdot U_2] - E[U_1]E[U_2]}{\sqrt{Var[U_1] \cdot Var[U_2]}}.$$

Вычислим $E[U_1 \cdot U_2]$:

$$E[U_1 \cdot U_2] = E[(1 - e^{-(\mu-\lambda)\xi_1})(1 - e^{-(\mu-\lambda)\xi_2})] =$$

$$= 1 - E[e^{-(\mu-\lambda)\xi_1}] - E[e^{-(\mu-\lambda)\xi_2}] + E[e^{-(\mu-\lambda)\xi_1} e^{-(\mu-\lambda)\xi_2}]$$

В силу того, что $\eta_i = (\mu - \lambda)\xi_i \sim Exp(1)$, т. е. $p_{\eta_i}(x) = e^{-x}$, $x > 0$, имеем

$$E[e^{-(\mu-\lambda)\xi_i}] = E[e^{-\eta_i}] = \int\limits_0^\infty e^{-x} e^{-x} dx = \frac{1}{2}.$$

Далее получаем

$$E[U_1 U_2] = 1 - \frac{1}{2} - \frac{1}{2} + E[e^{-(\mu-\lambda)\xi_1} e^{-(\mu-\lambda)\xi_2}] = E[e^{-(\mu-\lambda)\xi_1} e^{-(\mu-\lambda)\xi_2}] =$$

$$= \int\limits_0^\infty \int\limits_0^\infty e^{-(\mu-\lambda)x} e^{-(\mu-\lambda)y} p_{\xi_1\xi_2}(x, y) dx dy =$$

$$= \int\limits_0^\infty \int\limits_0^\infty e^{-(\mu-\lambda)sx} e^{-(\mu-\lambda)sy} p_{\xi_1\xi_2}(x, y)\, dx dy \Bigg|_{s=1,t=1} =$$

$$= \varphi((\mu - \lambda)s, (\mu - \lambda)t)\Big|_{s=1,t=1}.$$

Теперь, подставив $s = 1$ и $t = 1$ в формулу (2.35), имеем

$$E[U_1 \cdot U_2] = \frac{\sqrt{2}\sqrt{2-\rho}}{8-3\rho}.$$

И учитывая, что $E[U_i] = 1/2$, а $Var[U_i] = 1/12$, $i = 1, 2$, можем вычислить коэффициент корреляции Спирмена. Окончательно получаем

$$r_s = \frac{\frac{\sqrt{2}\sqrt{2-\rho}}{8-3\rho} - \frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{12}} = \frac{12\sqrt{2}\sqrt{2-\rho}}{8-3\rho} - 3.$$

□

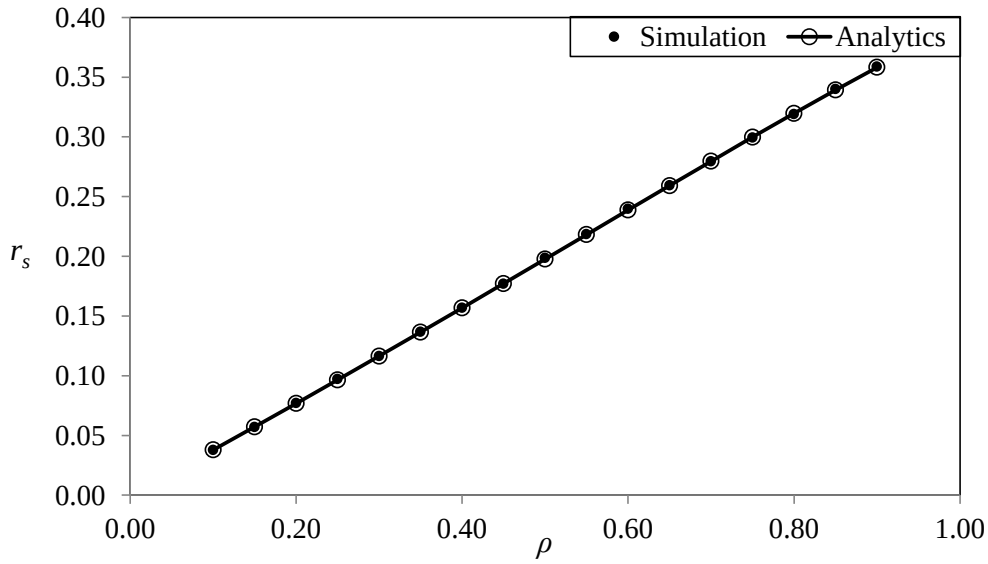


Рис. 26: Коэффициент корреляции Спирмена r_s.

Имитационное моделирование подтверждает корректность формулы (2.38) для коэффициента Спирмена (рис. 26).

Отметим, что коэффициент корреляции Спирмена также возрастает с увеличением загрузки нелинейным образом, который однако визуально ближе к линейному, чем для коэффициента корреляции Пирсона.

Коэффициент корреляции Кендалла. В данном разделе построим приближение для коэффициента корреляции Кендалла, поскольку вывести точную формулу для него не удается. Коэффициент корреляции Кендалла, также как и

коэффициент корреляции Спирмена, является ранговым коэффициентом корреляции. С его помощью также оценивается характер монотонной зависимости между случайными величинами и тесноты этой связи [55, 134].

Статистически, в выборке оценивается доля пар наблюдений случайных векторов (из двух случайных величин в качестве компонент), у которых компоненты имеют одинаковый порядок (характер монотонности), т. е., например, одна компонента возрастает с ростом другой, или наоборот, убывает с ростом другой. Коэффициент корреляции Кендалла оценивается как разность между долями пар векторов, для которых порядки совпадают и для которых различаются. Формулу для него можно записать следующим образом:

$$\hat{r}_k = 1 - \frac{4}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{1}\{[\xi_{1i} < \xi_{1j}] \neq [\xi_{2i} < \xi_{2j}]\}, \tag{2.39}$$

где $\mathbf{1}\{\cdot\}$ — функция-индикатор события $\{\cdot\}$, а (ξ_{1i}, ξ_{2i}), $1 \leq i \leq N$, — случайная выборка N векторов из случайных величин ξ_1 и ξ_2.

С точки зрения теории вероятностей, коэффициент корреляции Кендалла случайных величин X и Y определяется как

$$r_k = \mathbf{E}\ \mathrm{sign}(X_1 - X_2)(Y_1 - Y_2),$$

где (X_1, Y_1), (X_2, Y_2) — независимые случайные вектора, распределенные как (X, Y).

Для аппроксимации коэффициента корреляции Кендалла воспользуемся также как и раньше комбинацией нескольких методов. Сначала проведем графический анализ данных, полученных с помощью имитационного моделирования. После построения графика зависимости r_k от загрузки системы ρ, можно заметить, что как и в случае с r_s, эта зависимость по внешнему виду имеет характер близкий к линейному (сравните рис. 26 и рис. 27), хотя на самом деле, как следует из формулы (2.38), это не так. Предположим для простоты квадратичную зависимость r_k от ρ, т. е.

$$r_k \approx \rho(C_1 + C_2\rho).$$

Теперь для нахождения неизвестных коэффициентов воспользуемся методом оптимизации Нелдера–Мида. В процессе оптимизации относительно C будем минимизировать модуль относительной погрешности приближения

$$APE = \left| \frac{r_k - \widehat{r}_k}{r_k} \right| \cdot 100\% \to \min$$

между данными, полученными с помощью имитационного моделирования (r_k — “истинные” значения коэффициента корреляции Кендалла), и прогнозами, рассчитанными по предложенной аналитической формуле ($\widehat{r}_k$ — оценка коэффициента корреляции Кендалла), с начальными значениями коэффициентов, оцененными по графику.

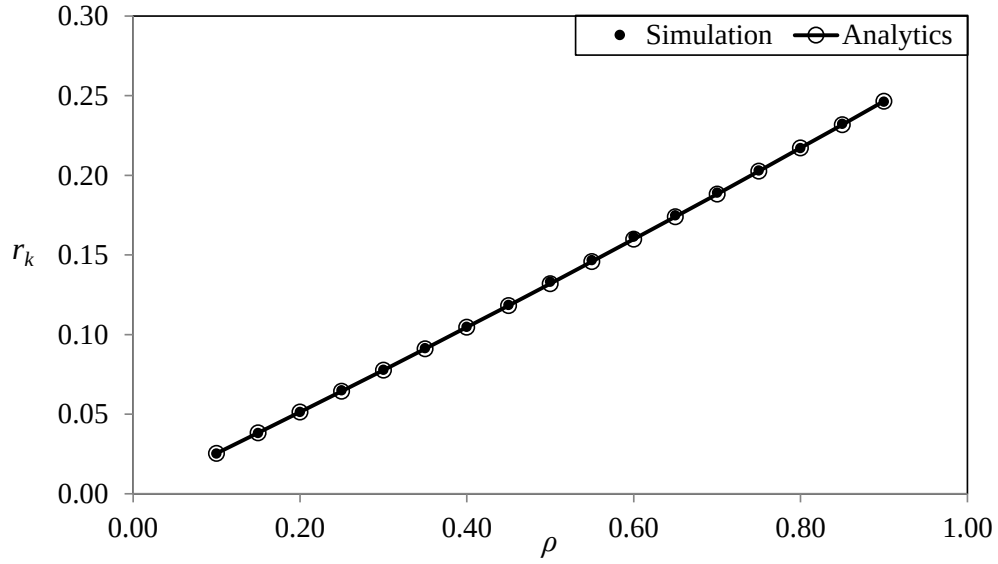


Рис. 27: Коэффициент корреляции Кендалла r_k.

Таблица 10: Погрешности приближений коэффициента корреляции Кендалла r_k формулой (2.40) для значений $\rho \in \{0.10, 0.15, ..., 0.90\}$

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
r_k	0.92828	0.09158	0.43138

В результате проведения оптимизации в программной среде Python получа-

ем значения коэффициентов

$$C_1 \approx 0.25134, \quad C_2 \approx 0.02517.$$

Следовательно, окончательное выражение для r_k будет иметь вид

$$r_k \approx \rho(0.25134 + 0.02517\rho). \tag{2.40}$$

На рисунке 27 и в таблице 10 можно сравнить результаты имитационного моделирования коэффициента Кендалла с результатами вычислений по аналитической формуле (2.6). Как видно, погрешность аппроксимации совсем не велика и не превышает 1%.

Что касается самих значений, то, например, для $\rho = 0.8$ верно $r_k \approx 0.2$, и это означает, что примерно для 60% пар наблюдений векторов (ξ_1, ξ_2) из всей совокупности их порядок (характер монотонности) совпадает, а для примерно 40% пар (ξ_1, ξ_2) — нет. При этом каждый вектор соответствует заявке и состоит из времен пребывания ее подзаявок.

Таким образом, характер зависимости коэффициентов корреляции Спирмена и Кендалла от загрузки в обоих случаях близок к линейному, но коэффициент Кендалла принимает меньшие значения.

Предельное двумерное распределение и соотношения коэффициентов корреляции. Из полученных результатов видно, 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.

Теорема 2.3. *Предел преобразования Лапласа–Стилтьеса совместного времени пребывания подзаявок в любой паре подсистем системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ и экспоненциальными временами обслуживания с параметром μ,*

где каждый из аргументов ПЛС умножен на величину обратную среднему времени пребывания подзаявки в подсистеме, т. е. $(\mu - \lambda)$

$$\lim_{\rho \to 1} \varphi((\mu-\lambda)s, (\mu-\lambda)t) = \frac{t(s+1)^{-1/2} + s(t+1)^{-1/2}}{s+t-st+(s+t)^2}, \tag{2.41}$$

является ПЛС некоторого двумерного распределения, а именно, это предельное распределение нормированных времен пребывания

$$\eta_1 = (\mu-\lambda)\xi_1, \quad \eta_2 = (\mu-\lambda)\xi_2,$$

чье совместное распределение описывается ПЛС $\varphi((\mu-\lambda)s, (\mu-\lambda)t)$*, при этом случайные величины* η_1 *и* η_2 *по отдельности всегда распределены одинаково (являются стандартными показательными), а зависимость между ними определяется загрузкой* ρ*.*

Доказательство. Чтобы получить решение предела (2.41) достаточно подставить $\rho = 1$ в выражение (2.35).

Далее согласно определению ПЛС

$$\varphi(s,t) = \varphi_{\xi_1,\xi_2}(s,t) = \int_0^\infty \int_0^\infty e^{-sx} e^{-ty} v(x,y)dxdy = E[e^{-\xi_1 s} \cdot e^{-\xi_2 t}],$$

где, как упоминалось ранее, $v(x,y)$ — это двумерная плотность распределения времен пребывания подзаявок в подсистемах. Для ПЛС случайных величин η_1 и η_2 справедливо следующее

$$\varphi_{\eta_1.\eta_2}(s,t) = E[e^{-\eta_1 s} \cdot e^{-\eta_2 t}] = E[e^{-\xi_1(\mu-\lambda)s} \cdot e^{-\xi_2(\mu-\lambda)t}] =$$

$$= \varphi_{\xi_1,\xi_2}((\mu-\lambda)s, (\mu-\lambda)t) = \varphi((\mu-\lambda)s, (\mu-\lambda)t).$$

Поскольку ξ_1 и ξ_2 являются экспоненциальными случайными величинами с параметром $(\mu - \lambda)$, то случайные величины η_1 и η_2 имеют стандартное показательное распределение, т. е. $\eta_i \sim Exp(1)$, $i = 1, 2$.

Коэффициент корреляции между случайными величинами η_1 и η_2 по свойствам равен коэффициенту корреляции между случайными величинами ξ_1 и ξ_2, т. е. $r_p = \rho(4-\rho)/8$ и, соответственно, определяется уровнем загрузки ρ. □

Все предельные значения коэффициентов корреляции, таким образом, являются значениями коэффициентов корреляции для предельного распределения (2.41).

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

$$r_k \approx 0.276.$$

Доказательство. Если подставить в формулы (2.32) и (2.38) значение коэффициента загрузки $\rho = 1$, то получатся соответствующие точные значения

$$r_p = \frac{\rho(4-\rho)}{8} = \frac{3}{8}, \quad r_s = \frac{12\sqrt{2}\sqrt{2-\rho}}{8-3\rho} - 3 = \frac{12\sqrt{2}}{5} - 3,$$

а если — в выражение (2.40) для оценки коэффициента корреляции Кендалла, то получим примерное значение

$$r_k \approx \rho(0.25134 + 0.02517\rho) \approx 0.276.$$

□

На рисунке 28 приведем также совместный график всех трех коэффициентов (с добавлением предельных значений). Как видно из рисунка, линии коэффициентов корреляции Пирсона и Спирмена пересекаются в одной точке (кроме нуля).

Следствие 2.2. *Уровень загрузки, при котором значения коэффициентов корреляции Пирсона и Спирмена времен пребывания подзаявок в подсистемах*

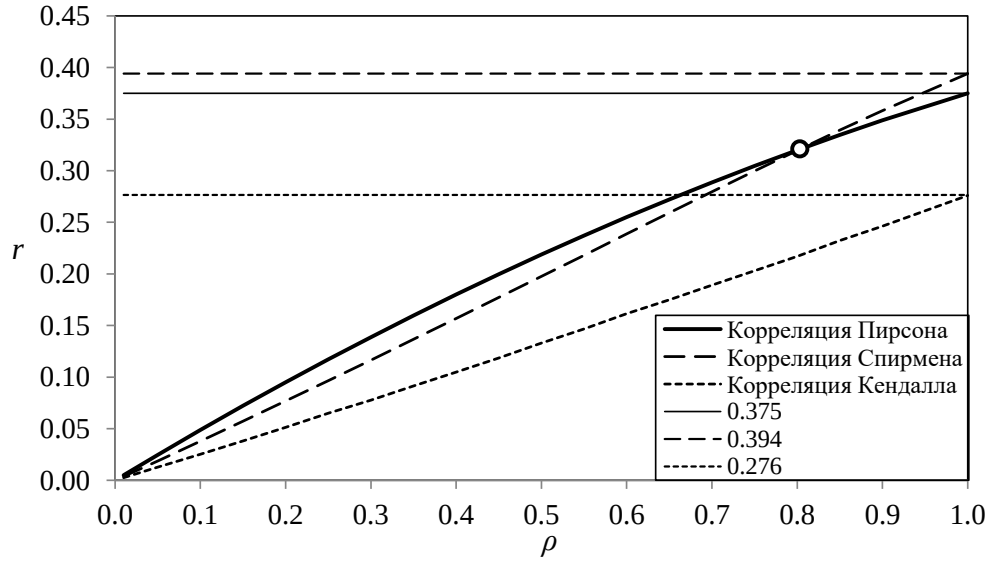


Рис. 28: Коэффициенты корреляции и их предельные значения.

системы с разделением и параллельным обслуживанием совпадают, определяется решением уравнения

$$9\rho^6 - 120\rho^5 + 160\rho^4 + 2752\rho^3 - 6080\rho^2 + 3072\rho = 0. \qquad (2.42)$$

Доказательство. Координаты точки пересечения можно найти, приравняв выражения (2.32) и (2.38)

$$\frac{\rho(4-\rho)}{8} = \frac{12\sqrt{2}\sqrt{2-\rho}}{8-3\rho} - 3,$$

$$(4\rho - \rho^2 + 24)(8-3\rho) = 96\sqrt{4-2\rho},$$

возводим в квадрат

$$9\rho^6 - 120\rho^5 + 160\rho^4 + 2752\rho^3 - 6080\rho^2 - 15360\rho + 36864 = 9216(4-2\rho)$$

и после упрощения получаем уравнение

$$9\rho^6 - 120\rho^5 + 160\rho^4 + 2752\rho^3 - 6080\rho^2 + 3072\rho = 0.$$

□

Численное решение уравнения (2.42) позволяет найти единственный корень, принадлежащий интервалу $\rho \in (0,1)$, а именно $\rho \approx 0.803146$. Таким образом, точка пересечения имеет следующие примерные координаты:

$(0.803146, 0.320943)$. Слева от нее больше коэффициент корреляции Пирсона, а справа — Спирмена.

2.5 Мета-гауссовская модель для времени отклика системы с разделением и параллельным обслуживанием

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.

Далее мы рассмотрим мета-гауссовскую модель, основанную на сведении произвольного распределения к многомерному нормальному. Такие модели применяются в финансовой математике [154, гл. 5, 9], гидрологии [132] и др. При этом не утверждается, что эта модель в данном случае является оптимальной. Тем не менее, она позволяет получить оценки производительности системы с разделением и параллельным обслуживанием более простым методом по сравнению, например, с организацией имитационного моделирования напрямую с той же целью.

Лемма 2.4. *Пусть для системы с разделением и параллельным обслуживанием с $K \geq 2$ подсистемами типа $M_\lambda|M_\mu|1$, в которой случайные величины ξ_i, $1 \leq i \leq K$ являются временами пребывания подзаявок от одной заявки, выполняется*

$$\zeta_i = \Phi^{-1}\left(1 - e^{-(\mu-\lambda)\xi_i}\right), \quad 1 \leq i \leq K, \tag{2.43}$$

где Φ^{-1} — обратная функция стандартного нормального распределения. Тогда случайные величины ζ_i, $1 \leq i \leq K$, имеют стандартное нормальное распределение и

$$\xi_i = -\frac{1}{\mu - \lambda} \ln\left(1 - \Phi(\zeta_i)\right), \quad 1 \leq i \leq K.$$

Доказательство. Справедливость данного утверждения следует из свойств обратного преобразования на интервале $(0,1)$ для случайной величины с непрерывной и строго монотонной функцией распределения. Поскольку времена пребывания подзаявок имеют показательное распределение с параметром $(\mu-\lambda)$, то, очевидно, справедливо

$$1-e^{-(\mu-\lambda)\xi_i}=\Phi[\Phi^{-1}(1-e^{-(\mu-\lambda)\xi_i})].$$

Обозначим

$$\Phi^{-1}(1-e^{-(\mu-\lambda)\xi_i})=\zeta_i\sim N(0,1),$$

тогда

$$1-e^{-(\mu-\lambda)\xi_i}=\Phi(\zeta_i)$$

и, соответственно,

$$\zeta_i=\Phi^{-1}\left(1-e^{-(\mu-\lambda)\xi_i}\right).$$

Что касается случайных величин ξ_i, то они легко выражаются из соотношения (2.43), т е.

$$1-e^{-(\mu-\lambda)\xi_i}=\Phi(\zeta_i),$$

$$-(\mu-\lambda)\xi_i=\ln(1-\Phi(\zeta_i))$$

и, следовательно,

$$\xi_i=-\frac{1}{\mu-\lambda}\ln\left(1-\Phi(\zeta_i)\right).$$

□

Поскольку случайные величины ζ_i, $1\le i\le K$, из (2.43) имеют стандартное нормальное распределение, то естественно приблизить их совместное распределение также многомерным нормальным. В связи с этим сформулируем следующее утверждение.

Лемма 2.5. *Если для системы с разделением и параллельным обслуживанием с $K\ge 2$ подсистемами типа $M_\lambda|M_\mu|1$, в которой случайные величины ξ_i, $1\le i\le K$, являются временами пребывания подзаявок от одной заявки,*

справедливо, что ζ_i, $1 \le i \le K$ *из* (2.43), *имеют приближенно совместное многомерное нормальное распределение, тогда коэффициент корреляции Пирсона* r *между любой парой случайных величин* ζ_i *и* ζ_j, $i \ne j$, $1 \le i, j \le K$, *оценивается выражением*

$$r = 2 \sin \frac{\pi r_s}{6}, \tag{2.44}$$

где коэффициент корреляции Спирмена r_s *вычисляется по формуле* (2.38).

Доказательство. В силу симметрии системы, любая пара величин ζ_i и ζ_j, $i \ne j$, имеет один и тот же коэффициент корреляции Пирсона, который обозначен через r. Кроме того, эта пара имеет тот же коэффициент корреляции Спирмена r_s, что и исходная пара ξ_i и ξ_j, поскольку коэффициент корреляции Спирмена сохраняется при непрерывных монотонно возрастающих преобразованиях случайных величин.

Для многомерного нормального распределения верно [154, теорема 5.36, с. 215]:

$$r_s = \frac{6}{\pi} \arcsin \frac{r}{2},$$

откуда можем найти

$$r = 2 \sin \frac{\pi r_s}{6},$$

при известном r_s, вычисляемом по формуле (2.38). □

Теорема 2.4. *Для системы с разделением и параллельным обслуживанием с* $K \ge 2$ *подсистемами типа* $M_\lambda|M_\mu|1$ *в мета-гауссовской модели справедливо приближение для времени отклика вида*

$$\widehat{R}_K = -\frac{1}{\mu - \lambda} \ln \left(1 - \Phi(\sqrt{r}\varepsilon_0 + \sqrt{1-r} \max\{\varepsilon_1, \ldots, \varepsilon_K\})\right), \tag{2.45}$$

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

Доказательство. Известно, что время отклика в системе с разделением и параллельным обслуживанием определяется как

$$R_K = \max\{\xi_1, \ldots, \xi_K\}.$$

Соответственно, согласно лемме 2.4 имеем

$$\max\{\xi_1, \ldots, \xi_K\} =$$

$$= \max\Big\{ -\frac{1}{\mu - \lambda} \ln(1 - \Phi(\zeta_1)), ..., -\frac{1}{\mu - \lambda} \ln(1 - \Phi(\zeta_K)) \Big\}.$$

Этот максимум в силу того, что функция $f(x) = -1/(\mu - \lambda) \cdot \ln(1 - \Phi(x))$ является монотонно возрастающей благодаря свойствам функции распределения нормальной случайной величины, свойствам логарифмической функции и условия $\mu > \lambda$ (рис. 29), равен

$$R_K = -\frac{1}{\mu - \lambda} \ln(1 - \Phi(\max\{\zeta_1, ..., \zeta_K\})).$$

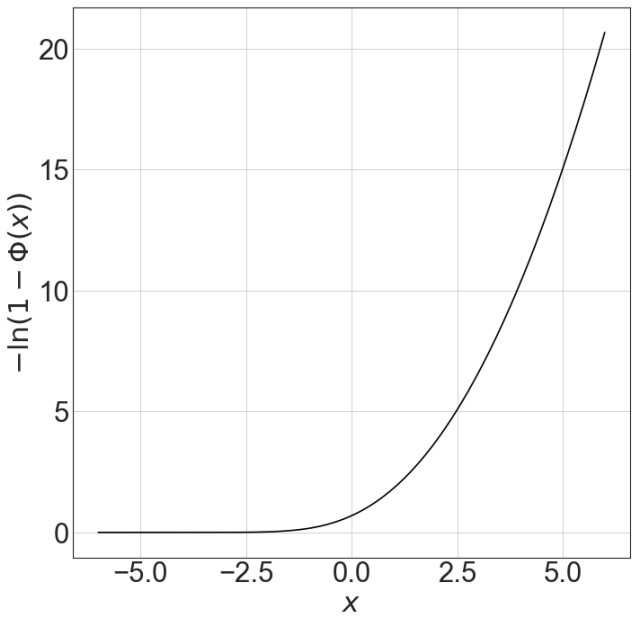


Рис. 29: График функции $-1/(\mu - \lambda) \cdot \ln(1 - \Phi(x))$ при $(\mu - \lambda) = 1$.

Набор случайных величин ζ_i с нужными распределениями и корреляциями можно получить по формулам [154, с. 445]

$$\widehat{\zeta}_i = \sqrt{r} \cdot \varepsilon_0 + \sqrt{1-r} \cdot \varepsilon_i, \quad 1 \leq i \leq K,$$

где ε_i, $0 \leq i \leq K$, — независимые стандартные нормальные случайные величины, а коэффициент корреляции Пирсона r в соответствии с утверждением леммы 2.5 определяется выражениями (2.44) и (2.38).

Таким образом, максимум случайных величин $\widehat{\zeta}_1, ..., \widehat{\zeta}_K$ равен

$$\max\{\widehat{\zeta}_1, ..., \widehat{\zeta}_K\} =$$

$$= \max\{\sqrt{r} \cdot \varepsilon_0 + \sqrt{1-r} \cdot \varepsilon_1, ..., \sqrt{r} \cdot \varepsilon_0 + \sqrt{1-r} \cdot \varepsilon_K\} =$$

$$= \sqrt{r} \cdot \varepsilon_0 + \sqrt{1-r} \cdot \max\{\varepsilon_1, ..., \varepsilon_K\}.$$

В результате, получаем оценку

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

□

С помощью формул (2.38), (2.44) и (2.45) была проведена симуляция для значений $\lambda = 1$, $\rho \in [0.1, 0.9]$ с шагом 0.05 и K от 3 до 20, с целью оценить среднее время отклика в сравнении с результатами симуляции fork-join системы массового обслуживания, проведенной ранее в [102, 107].

Погрешности полученных приближений представлены в таблице 11.

Таблица 11: Погрешности приближений среднего времени отклика с помощью формулы (2.45) (мета-гауссовская модель) для значений $K = 3, ..., 20$ и $\rho \in \{0.10, 0.15, ..., 0.90\}$

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
$E[R_K]$	4.18222	0.29829	2.46825

Отметим, что $MaxAPE$ оказывается близка к погрешности классической формулы Нельсона–Тантави [163] на данном множестве значений параметров (около 4%).

Модель легко использовать для оценки не только среднего времени отклика, но и дисперсии, квантилей и др.

К недостаткам мета-гауссовской модели можно отнести то, что она требует симуляции, а не дает явную формулу, однако эта симуляция гораздо проще и быстрее, чем симуляция исходной fork-join системы массового обслуживания.

Разумеется, возможны и более совершенные модели, параметризуемые коэффициентами корреляции. Если же модель использует параметризацию из других соображений, то представляется желательным, чтобы ее прогнозы значений коэффициентов корреляции не слишком расходились с фактическими.

2.6 Копулы и квантили в fork-join системах массового обслуживания с подсистемами типа $M|M|1$

Помимо первых или вторых моментов случайной величины времени отклика интерес представляют и квантили ее распределения. Установление квантиля заданного уровня вероятности означает, что система гарантирует обслуживание заявки не более чем за установленное время с данной вероятностью. Такой подход к оценке работы системы уместен, если важна срочность обслуживания (например, в медицине) или если долгое обслуживание вызывает недовольство клиентов и это не компенсируется быстрым обслуживанием других. Таким образом, оценка квантилями является более тонкой, чем оценка средним, в смысле качества обслуживания в современном мире.

В основе подхода к построению оценки квантилей времени отклика 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 многомер-

ное распределение раскладывается на частные распределения и копулу. Таким образом, копула исчерпывающе описывает зависимость случайных величин в чистом виде.

Будем рассматривать частный случай двух подсистем ($K = 2$), однако напомним, что количество подсистем никак не влияет на зависимость в любой паре времен пребывания подзаявок одной заявки (рис. 15). В систему как и ранее поступает пуассоновский поток заявок с интенсивностью $\lambda > 0$. В момент поступления в систему заявка мгновенно разделяется на 2 подзаявки, каждая из которых попадает в соответствующую подсистему, имеющую накопитель неограниченной емкости и один прибор. Все приборы являются однородными, время обслуживания имеет экспоненциальное распределение с параметром $\mu > 0$. Таким образом, подсистемы представляют собой две идентичных СМО типа $M|M|1$. Поскольку заявка считается обслуженной только после окончания обслуживания обеих составляющих ее подзаявок, то случайное время пребывания заявки в СМО (время отклика) R является максимумом из двух случайных времен пребывания подзаявок ξ_i, $i = 1, 2$, в каждой из двух подсистем:

$$R = \max\{\xi_1, \xi_2\}.$$

Случайные величины ξ_1 и ξ_2 являются коррелированными в силу того, что все подзаявки (части одной заявки) поступают в подсистемы в одно и то же время. Так как данный раздел является логическим продолжением исследований, начатых в разделе 2.4 (в контексте изучения зависимостей между временами пребывания подзаявок), то для удобства и понимания целостности всей картины приведем полученные там точные аналитические формулы для коэффициентов корреляции Пирсона и Спирмена

$$r_p = \frac{\rho(4-\rho)}{8}, \quad r_s = \frac{12\sqrt{2}\sqrt{2-\rho}}{8-3\rho} - 3,$$

а также приближенное выражение для коэффициента корреляции Кендалла

$$r_k \approx \rho(0.25134 + 0.02517\rho),$$

где $\rho = \lambda/\mu$ — это коэффициент загрузки системы.

Далее мы для простоты также полагаем $\lambda = 1$, $\mu = 1/\rho$. Такие параметры использовались при моделировании системы ранее.

Отметим, что различные коэффициенты корреляции, в общем случае, хотя и отражают зависимость, но лишь частично. В полной мере отражают зависимость только копулы.

Элементы теории копул. *Копулой* C называется функция многомерного распределения на $[0,1]^d$, $d \geq 2$, если все частные распределения являются равномерными на $[0,1]$ [162]. Согласно знаменитой теореме Скляра [179], любая функция многомерного распределения в $\mathbb{R}^d$ представима в виде

$$F(x_1, \dots x_d) = C(F_1(x_1), \dots F_d(x_d)),$$

где F_i, $1 \leq i \leq d$, — функции частных распределений. Таким образом, всякому многомерному распределению можно поставить в соответствие его копулу. Если частные распределения непрерывны, то такое представление единственно.

В качестве классического источника основ теории копул можно указать [162].

Далее ограничимся случаем двумерных копул ($d = 2$).

Диагональным сечением (двумерной) копулы называется функция $\delta(u) = C(u,u)$, $u \in [0,1]$. Она обладает следующими (необходимыми и достаточными) свойствами:

$$\begin{gathered} \max\{2u-1, 0\} \leq \delta(u) \leq u; \quad 0 \leq \delta(u_2) - \delta(u_1) \leq 2(u_2 - u_1), \\ 0 \leq u_1 \leq u_2 \leq 1. \end{gathered} \tag{2.46}$$

Смысл изучения диагональных сечений, например, 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

$$F_{\max}(x) = P(X_1 < x, X_2 < x) = C(F(x), F(x)) = \delta(F(x)), \tag{2.47}$$

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

Легко заметить, что условиям (2.46) удовлетворяет степенная функция (рис. 30)

$$\delta(u) = u^{\alpha}, \quad 1 \leq \alpha \leq 2,$$

тогда случаю $\alpha = 1$

$$F_{\max}(x) = F(x)$$

соответствует совершенная положительная зависимость (комонотонность), а случаю $\alpha = 2$

$$F_{\max}(x) = (F(x))^2 = P(X_1 < x) \cdot P(X_2 < x)$$

— независимость случайных величин.

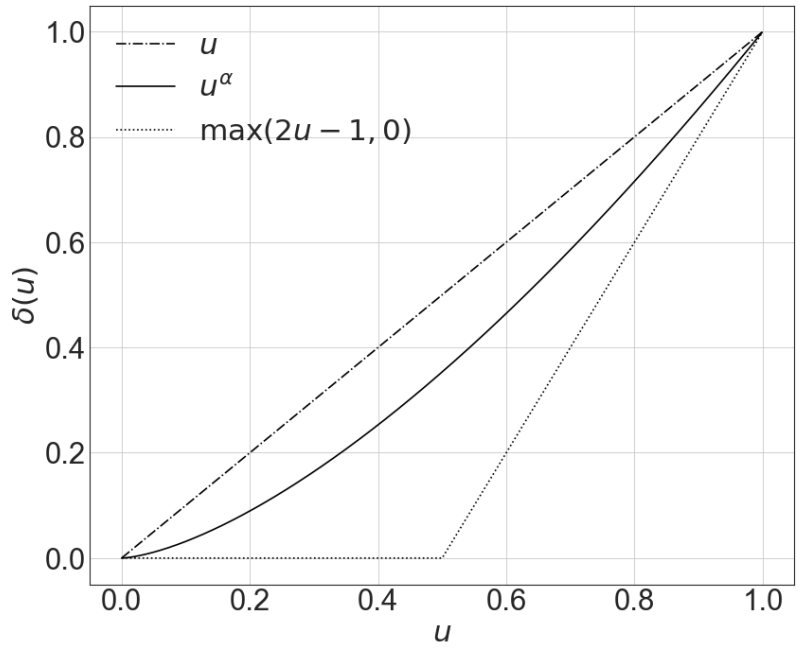


Рис. 30: Иллюстрация свойств (2.46) степенного диагонального сечения копулы $\delta(u) = u^{\alpha}$, $\alpha = 1.5$.

Классическим примером абсолютно непрерывной (имеющей плотность) ко-

пулы со степенным диагональным сечением является копула Гумбеля

$$C(u_1, u_2) = \exp\{-((-\ln u_1)^\theta + (-\ln u_2)^\theta)^{1/\theta}\}, \quad \theta \geq 1, \quad u_1, u_2 \in [0, 1],$$

тогда

$$\delta(u) = u^{2^{1/\theta}}.$$

Более точно, копула Гумбеля относится к классу копул экстремальных значений, которые всегда имеют степенные диагональные сечения. О таких копулах можно прочитать в [162, § 3.3.4], [114].

Приближения диагонального сечения и квантилей времени отклика. Для приближения квантилей распределения случайной величины времени отклика $R = \max\{\xi_1, \xi_2\}$ воспользуемся элементами теории копул. Будем рассматривать двумерную копулу $C(u_1, u_2)$ случайных векторов времен пребывания в подсистемах (ξ_1, ξ_2).

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*

$$x_p = F_R^{-1}(p) = F^{-1}(\delta^{-1}(p)), \tag{2.48}$$

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

Доказательство. Каждая компонента случайного вектора имеет экспоненциальное распределение с функцией распределения $F(x) = 1 - e^{-(\mu-\lambda)x}$, $x \geqslant 0$. Тогда в соответствии с теоремой Скляра представление с помощью копулы совместного распределения (ξ_1, ξ_2) существует и единственно

$$F_{\xi_1,\xi_2}(x_1, x_2) = P(\xi_1 < x_1, \xi_2 < x_2) = C(F(x_1), F(x_2)).$$

В силу (2.47) получаем

$$F_R(x) = C(F(x), F(x)) = \delta(F(x)), \tag{2.49}$$

где $\delta(u) = C(u, u)$ — диагональное сечение копулы, что дает нам уравнение для квантили уровня p распределения времени отклика

$$F_R(x_p) = \delta(F(x_p)) = p,$$

поэтому

$$x_p = F_R^{-1}(p) = F^{-1}(\delta^{-1}(p)).$$

□

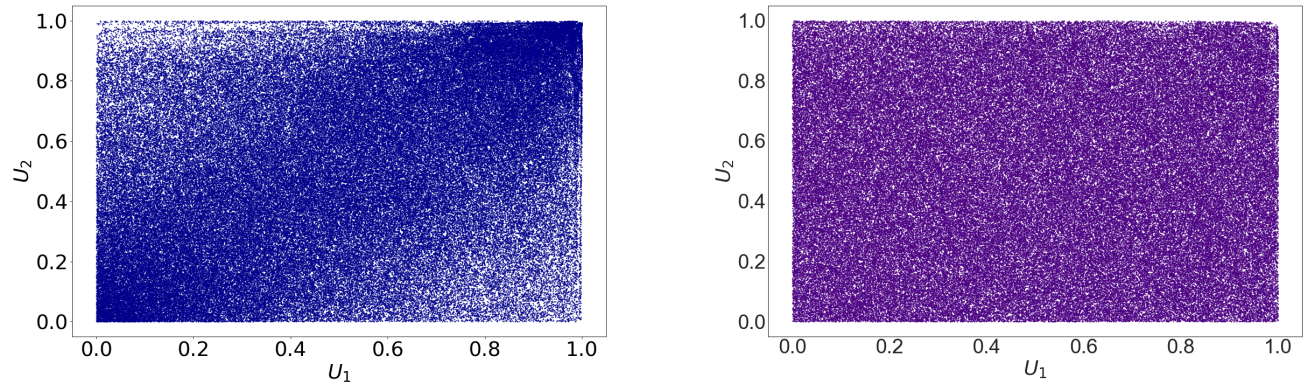
Стоит напомнить, что случайные величины ξ_i, $i = 1, 2$, имеют экспоненциальное распределение с параметром $(\mu - \lambda)$; далее рассмотрим

$$U_i = 1 - e^{-(\mu-\lambda)\xi_i}, \quad i = 1, 2.$$

Эти случайные величины будут иметь равномерное распределение на отрезке $[0, 1]$, т. е. $U_i \sim R[0, 1]$. Тогда

$$V = \max\{U_1, U_2\} = 1 - e^{-(\mu-\lambda)\cdot\max\{\xi_1,\xi_2\}} = 1 - e^{-(\mu-\lambda)R}. \tag{2.50}$$

На рисунке 31 изображено множество точек (U_1, U_2) при $\rho = 0.9$ для fork-join системы с двумя подсистемами $M|M|1$ и для случая двух независимых параллельно функционирующих СМО $M|M|1$ с одинаковыми параметрами. Количество пар точек равно 200 тысячам, увеличение их числа перегружает иллюстрацию. Как видно, на рисунке 31 б) точки распределены равномерно внутри единичного квадрата, что характерно для случая независимых случайных величин. На 31 а) подобная равномерность в распределении точек уже не наблюдается, даже несмотря на то, что значение коэффициента корреляции Пирсона между ξ_1 и ξ_2 невелико ($r_p = 0.34875$), и визуальный анализ несколько затруднен, тем не менее, зависимость между $U_1 = F(\xi_1)$ и $U_2 = F(\xi_2)$ в данном случае, очевидно, прослеживается.



а) б)

Рис. 31: Иллюстрация наличия/отсутствия зависимости между U_1 и U_2 при $\rho = 0.9$ в случае а) fork-join СМО с двумя подсистемами $M|M|1$; б) двух параллельно функционирующих СМО $M|M|1$.

Диагональное сечение копулы можно оценить следующим образом. Имеем

$$\delta(u) = C(u,u) = P(U_1 < u, U_2 < u) = P(\max(U_1, U_2) < u) = P(V < u) = p,$$

т. е.

$$\delta(u_p) = P(V < u_p) = p,$$

где u_p — это квантиль распределения с. в. V. С помощью реализаций V_i случайной величины V, полученных посредством имитационного моделирования значений случайных времен пребывания в fork-join СМО R_i и дальнейшей подстановкой их в формулу (2.50), строим оценку диагонального сечения $\delta(u)$, а фактически вероятностей p. Иными словами, строим эмпирическую оценку диагонального сечения с помощью квантилей распределения V. Для этого упорядочиваем полученные посредством симуляции величины V: $V_{(1)}$, $V_{(2)}$,..., $V_{(N)}$, где $V_{(k)}$ — это k-я порядковая статистика, $k = 1, ..., N$, и по точкам $(V_{(k)}, k/(N+1))$ определяем оценки (u_p, p) для значений вероятности из интересующего нас интервала $p \in \{0.2, 0.25, 0.30, ..., 0.90\}$, при конкретном фиксированном значении коэффициента загрузки $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$. Выбор значений p обусловлен тем, что, как правило, больший интерес представляют квантили более

высокого уровня, поэтому значения p начинаем рассматривать с 0.2. Далее на основе имеющихся данных будем строить прогноз вероятностей p в зависимости от квантилей u_p и коэффициента загрузки ρ.

$$p \approx \widehat{p} = \widehat{\delta}(u_p, \rho).$$

Теперь для определения вида функциональной зависимости проведем графи-

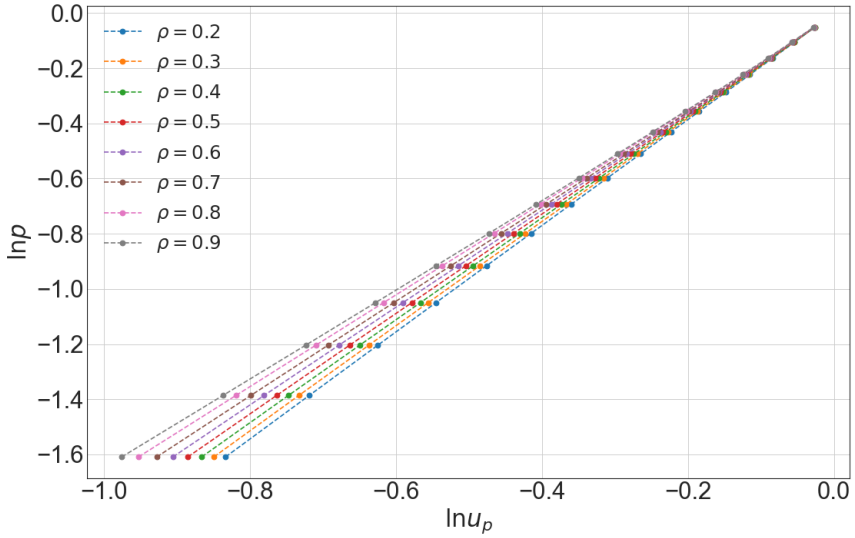


Рис. 32: Зависимость $\ln p$ от $\ln u_p$.

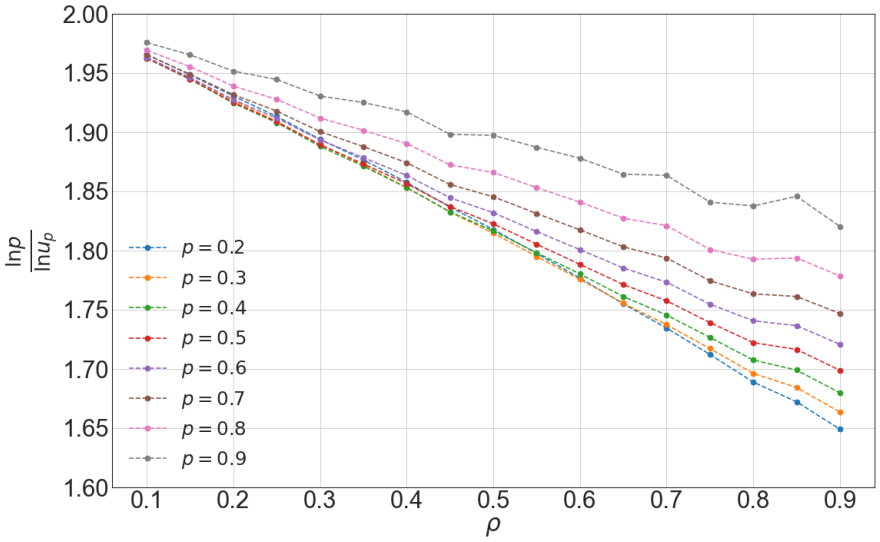


Рис. 33: Зависимость $(\ln p / \ln u_p)$ от ρ.

ческий анализ полученных данных. Прежде всего, было замечено, что зависимость p от u_p хорошо описывается степенной функцией, что соответствует линейной зависимости для логарифмов (см. рис. 32). Зависимость показателя степени α от ρ также оказалась близка к линейной (см. рис. 33). Отметим, что при $\rho \to 0$ времена пребывания подзаявок асимптотически независимы, откуда $\alpha \to 2$. Как видно из рисунка 33, график зависимости напоминает пучок близких прямых, проходящих через точку $(0, 2)$, поэтому естественно предположить (в качестве первого приближения), что

$$\frac{\ln p}{\ln u_p} \approx 2 - C \cdot \rho,$$

а следовательно

$$p = \delta(u_p, \rho) \approx u_p^{2-C\cdot\rho}. \tag{2.51}$$

Остается только вычислить значение коэффициента C. Аналогично ситуации с коэффициентом корреляции Кендалла [103], будем минимизировать методом Нелдера–Мида модуль относительной погрешности аппроксимации относительно данных имитационного моделирования, в результате чего получим значение

$$C \approx 0.370608. \tag{2.52}$$

Таким образом, имеем

$$p = \delta(u_p, \rho) \approx u_p^{2-0.370608\cdot\rho}. \tag{2.53}$$

На рисунке 34 представлены результаты имитационного моделирования вероятностей или уровней p квантилей u_p случайной величины $V = F(R)$ в сравнении с результатами вычислений по аналитической формуле (2.53) в диапазоне $[0.20, 0.95]$ с шагом 0.05. Каждая точка, изображенная на графике, фактически представляет собой множество из 17 точек по числу значений коэффициента загрузки $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$, которые накладываются друг на друга и практически сливаются, что и должно быть при хорошем уровне приближения вероятностей p. Небольшое расслоение (отклонение в пределах 2%) наблюдается

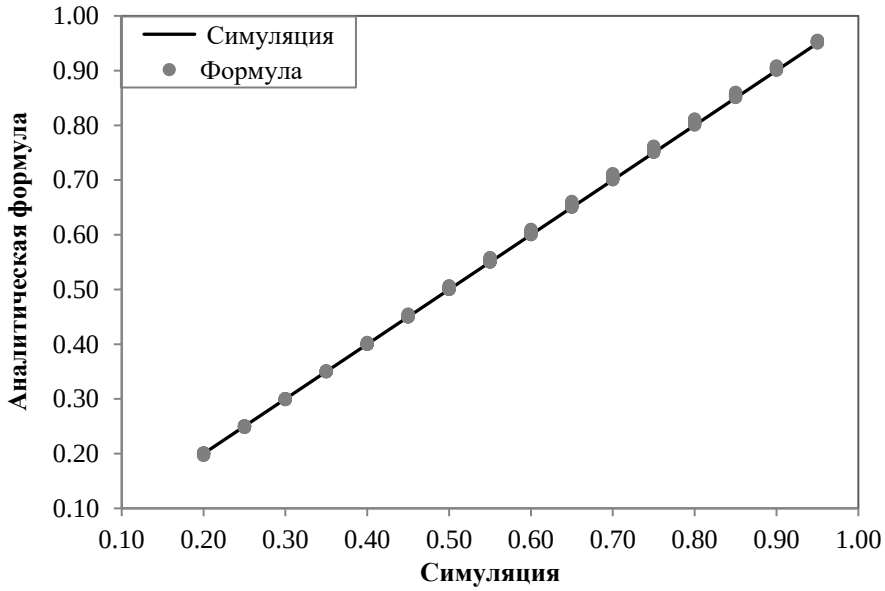


Рис. 34: Сравнение аналитических результатов формулы (2.53) с имитационным моделированием значений p квантилей u_p случайной величины $V = F(R)$ для значений $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$.

с ростом значений p. Для ясности в таблице 12 приведены абсолютные значения относительных погрешностей приближений для 272 рассчитанных значений p.

Таблица 12: Погрешности приближений значений вероятностей p, рассчитанных с помощью аналитической формулы (2.53) по сравнению с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Вероятность p из формулы (2.53)	1.679144	0.002731	0.438597

Теперь, учитывая (2.48), можем записать

$$\delta^{-1}(p) = F(x_p), \tag{2.54}$$

при этом из (2.51) следует, что $\delta^{-1}(p) \approx p^{\frac{1}{2-C\cdot\rho}}$. Подставляем оценку $\delta^{-1}(p)$ в (2.54) и получаем соотношение

$$p^{\frac{1}{2-C\cdot\rho}} = 1 - e^{-(\mu-\lambda)x_p},$$

откуда следует, что квантиль уровня p распределения случайной величины времени отклика fork-join СМО R определяется выражением

$$x_p \approx -\frac{\ln(1-p^{\frac{1}{2-C\cdot\rho}})}{\mu-\lambda}. \tag{2.55}$$

Далее оценим качество аппроксимации полученного выражения на следующем наборе данных: $\rho \in \{0.10, 0.15, ..., 0.90\}$, $p \in \{0.20, 0.25, ..., 0.90\}$, т. е. всего получим 272 значения, для которых и будем проводить оценку погрешности приближений формулой (2.55). В таблице 13 в первой строке для значения $C \approx 0.370608$ представлены относительные ошибки приближения, причем модуль максимальной погрешности приближения ($MaxAPE$) составляет около 3%, а среднее значение модуля относительной ошибки не превышает 1%. Однако если снова воспользоваться методом оптимизации Нелдера–Мида, но уже для минимизации погрешности квантили x_p из формулы (2.55), то в результате получим значение $C \approx 0.348284$. В этом случае результат несколько улучшится, однако, как показывают дальнейшие исследования, значение коэффициента $C \approx 0.37$ оказывается более обоснованным в других аспектах.

Таблица 13: Погрешности приближений квантилей времени отклика для двух вариантов значений коэффициента C в формуле (2.55)

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль x_p, $C \approx 0.370608$	3.123971	0.002328	0.734956
Квантиль x_p, $C \approx 0.348284$	2.819299	0.007276	0.699130

На рисунке 35 наглядно продемонстрировано качество аппроксимации квантилей времени отклика. Как следует из графиков, большие погрешности возникают для больших значений коэффициента загрузки системы ρ, но не превышают при этом и 3%, что является приемлемым результатом.

Ради уточнения аппроксимации квантилей, возвращаясь к рис. 33, можно отметить зависимость наклона прямых от p. Это наводит на мысль вместо кон-

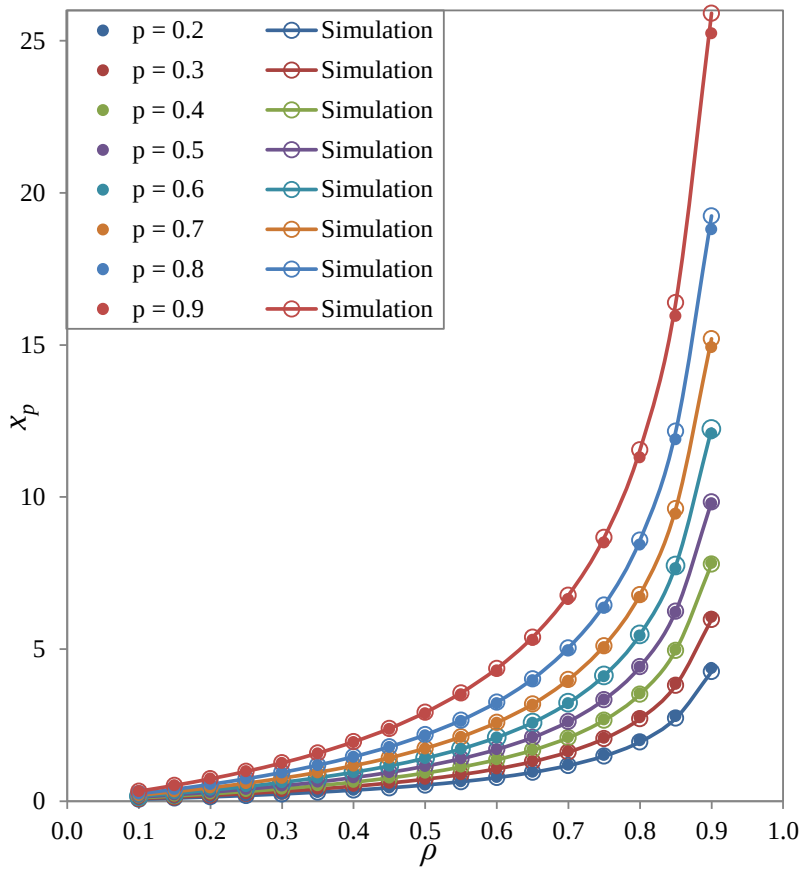


Рис. 35: Сравнение аналитических результатов формулы (2.55) с имитационным моделированием квантилей x_p случайной величины времени отклика R.

станты C в (2.55) использовать выражения вида $C_1 - C_2 p$ или $C_1 - C_2 p^2$. Подбор констант методом Нелдера–Мида и сравнительный анализ точности показывают, что лучше второй вариант, а именно, приближение

$$x_p \approx -\frac{\ln(1 - p^{\frac{1}{2-(C_1 - C_2 p^2)\rho}})}{\mu - \lambda}, \tag{2.56}$$

где

$$C_1 \approx 0.390327, \quad C_2 \approx 0.237842,$$

при этом погрешность составляет всего 0.62%, что меньше прежнего в 4.6 раза. Более подробно полученные результаты представлены в таблице 14.

Аналогичным образом строится оценка квантилей распределения времени отклика в случае разделения подзаявок на большее число подзаявок, в частности для значений $K = 2, ..., 20$, вероятностей $p \in \{0.20, 0.25, ..., 0.90\}$ и загрузки

Таблица 14: Погрешности приближений квантилей времени отклика, рассчитанные с помощью аналитической формулы (2.56) по сравнению с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль x_p	0.617316	0.004912	0.304632

$\rho \in \{0.10, 0.15, ..., 0.90\}$ получаем выражение

$$x_{p,K} \approx -\frac{\ln(1 - p^{\frac{1}{K^{1-(C_1 - C_2 p^2)\rho}}})}{\mu - \lambda},$$

где

$$C_1 \approx 0.390797, \quad C_2 \approx 0.221811.$$

В этом случае $MaxAPE \approx 5.498891\%$, $MinAPE \approx 0.000306\%$ и $MAPE \approx 1.116448\%$. Стоит отметить, что основной задачей в данном случае было получение равномерной (по относительной точности) оценки квантилей по всем значениям K от 2 до 20. При конкретных K аналогичным образом можно получить гораздо более точные оценки, как это было сделано для $K = 2$.

К сожалению, этот подход не позволяет получить явное выражение или удобное приближение для диагонального сечения, поэтому вернемся к формуле (2.51) и перейдем от нее к копулам.

Приближение копулы времен пребывания подзаявок копулой Гумбеля. В предыдущем разделе была получена оценка диагонального сечения копулы $\delta(u)$. В данном разделе будет представлено аналитическое выражение, оценивающее саму копулу $C(u_1, u_2)$. Для этого потребуются эмпирические данные, проанализировав которые, можно будет сделать вывод о близости исследуемой копулы к одному из известных семейств.

Алгоритм построения эмпирической копулы будет следующим:

1. имитационное моделирование множества пар (ξ_1^k, ξ_2^k) случайных величин времен пребывания в подсистемах $M|M|1$ fork-join СМО, где k — это порядковый номер смоделированной пары значений, $k = 1, ..., N$, N — объем выборки (общее число пар случайных величин);

2. преобразование случайных величин с экспоненциальным распределением $\xi_i \sim Exp(\mu - \lambda)$ с помощью функции распределения в случайные величины с равномерным распределением на отрезке $[0, 1]$, $U_i \sim R[0, 1]$, $i = 1, 2$,

$$(U_1^k, U_2^k) = (F(\xi_1^k), F(\xi_2^k)) = (1 - e^{-(\mu-\lambda)\xi_1^k}, 1 - e^{-(\mu-\lambda)\xi_2^k});$$

3. разбиение единичного квадрата на более мелкие квадраты (сетку) со сторонами длиной $h = 1/m$, где, например, $m = 20$ и определение числа точек (U_1^k, U_2^k), попадающих в каждый из квадратов, вершинами которого являются точки $(0, 0)$, $(ih, 0)$, $(0, jh)$, (ih, jh), $i, j = 1, ..., m$, и нормирование полученного значения, т. е.

$$C_{ij} = C(ih, jh) \approx \widehat{C}_{ij} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}\{U_1^k < ih, U_2^k < jh\},$$

где $\mathbf{1}\{\cdot\}$ — функция-индикатор события $\{\cdot\}$.

На рисунке 36 представлен график эмпирической копулы или, что то же самое, совместной функции распределения случайного вектора (U_1, U_2), построенной в соответствии с представленным выше алгоритмом.

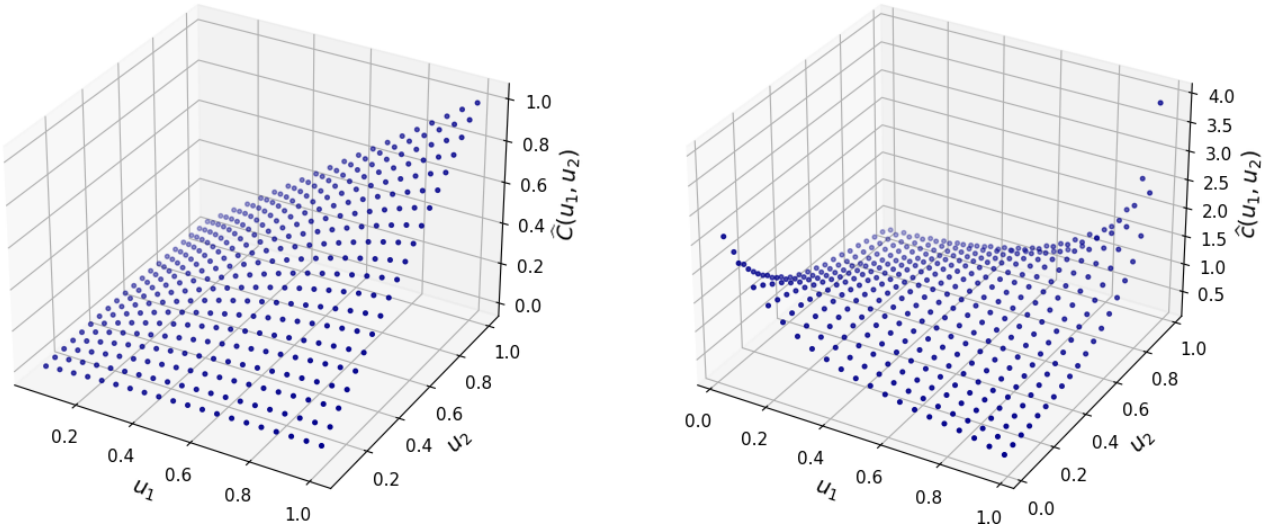
Также построим плотность копулы по следующем **алгоритму**:

1. используя результаты шагов 1 и 2 из предыдущего алгоритма, получим множество пар случайных величин (U_1^k, U_2^k), $k = 1, ..., N$;

2. разбиваем единичный квадрат на более мелкие квадраты (сетку) со сторонами длиной $h = 1/m$, где, например, $m = 20$ и определяем число точек (U_1^k, U_2^k), попадающих в каждый из квадратов, вершинами которого являются точки $((i-1)h, (j-1)h)$, $(ih, (j-1)h)$, $((i-1)h, jh)$, (ih, jh),

$i, j = 1, ..., m$, и нормируем полученные значения, т. е.

$$c_{ij} = c(ih, jh) \approx \widehat{c}_{ij} = \frac{1}{Nh^2} \sum_{k=1}^{N} \mathbf{1}\{(i-1)h < U_1^k < ih, (j-1)h < U_2^k < jh\}.$$



а) б)

Рис. 36: а) Эмпирическая копула $\widehat{C}(u_1, u_2)$, $\rho = 0.9$; б) эмпирическая плотность копулы $\widehat{c}(u_1, u_2)$, $\rho = 0.9$.

Исходя из внешнего вида полученных эмпирических функций на рисунках 36 а) и б), а также учитывая, что диагональное сечение рассматриваемой копулы было приближено в предыдущем разделе выражением вида

$$\delta(u) \approx u^{\alpha}, \quad \alpha = 2 - C\rho, \tag{2.57}$$

будем приближать искомую копулу $C(u_1, u_2)$ копулой Гумбеля, которая имеет вид

$$C_g(u_1, u_2) = \exp\{-[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}]^{\frac{1}{\theta}}\}, \tag{2.58}$$

где $\theta \in [1, +\infty)$ — параметр копулы, который предстоит оценить. Плотность копулы Гумбеля определяется взятием смешанной частной производной второго

порядка от функции копулы (2.58)

$$c_g(u_1,u_2)=\frac{\partial^2 C_g(u_1,u_2)}{\partial u_1\partial u_2}= \\ =\frac{C_g(u_1,u_2)(-\ln u_1)^{\theta}(-\ln u_2)^{\theta}[\theta+[(-\ln u_1)^{\theta}+(-\ln u_2)^{\theta}]^{\frac{1}{\theta}}-1]}{u_1u_2\ln u_1\ln u_2[(-\ln u_1)^{\theta}+(-\ln u_2)^{\theta}]^{\frac{2\theta-1}{\theta}}}. \tag{2.59}$$

Поскольку для копулы Гумбеля диагональное сечение имеет следующий вид

$$\delta_g(u)=C_g(u,u)=u^{2^{1/\theta}},$$

то с учетом (2.57) получаем, что

$$\theta\approx\frac{\ln 2}{\ln\alpha}=\frac{\ln 2}{\ln(2-C\rho)}. \tag{2.60}$$

Далее снова воспользуемся методом оптимизации Нелдера–Мида для минимизации модуля относительной ошибки приближения функции копулы Гумбеля (2.58) с учетом того, что параметр θ определяется выражением (2.60), при сравнении с “истинными” значениями функции копулы Гумбеля, полученными с помощью имитационного моделирования для различных коэффициентов загрузки $\rho\in\{0.10,0.15,0.20,...,0.90\}$. Как и раньше, не будем рассматривать квантили низкого уровня, т. е. пусть $u_1,u_2\in\{0.20,0.25,...,0.90\}$. В результате получаем следующее значение искомого коэффициента

$$C\approx 0.369250, \tag{2.61}$$

поэтому

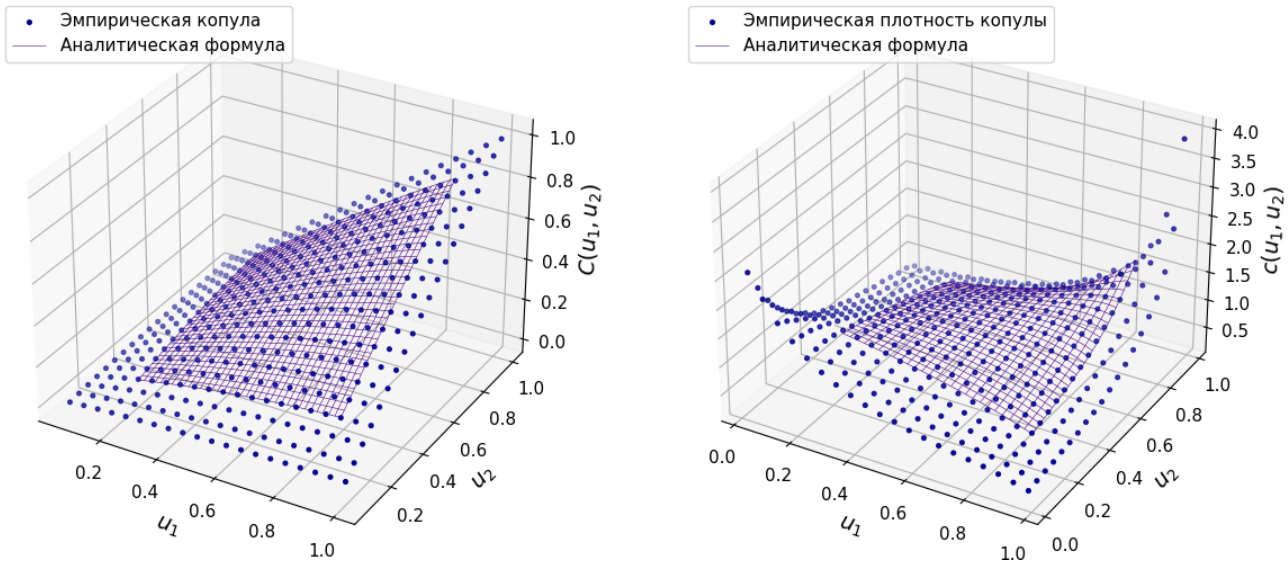
$$C(u_1,u_2)\approx\exp\{-((-\ln u_1)^{\frac{\ln 2}{\ln(2-0.36925\rho)}}+(-\ln u_2)^{\frac{\ln 2}{\ln(2-0.36925\rho)}})^{\frac{\ln(2-0.36925\rho)}{\ln 2}}\}. \tag{2.62}$$

Как видно, из (2.52) и (2.61) значения коэффициентов очень близки и фактически согласуются между собой, что подтверждает качество полученных приближений. Что касается погрешности аппроксимации формулы (2.62), то в таблице 15 представлены значения максимальной ($MaxAPE$), минимальной ($MinAPE$) и средней относительной ошибки аппроксимации ($MAPE$), первая из которых не превышает 5%, на наборе данных из 4913 троек (ρ,u_1,u_2).

На рисунке 37 а) также представлены графики эмпирической функции копулы и копулы, определяемой выражением (2.62) на заданном диапазоне значений $0.2 \leqslant u_1, u_2 \leqslant 0.9$.

Таблица 15: Погрешности приближений функции копулы Гумбеля $C(u_1, u_2)$ формулой (2.62) и плотности копулы Гумбеля формулой (2.59)

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
$C(u_1, u_2)$	4.966424	0.000000	0.411003
$c(u_1, u_2)$	11.164042	0.000805	1.634484



а) б)

Рис. 37: Сравнение эмпирических и аналитических значений при $\rho = 0.9$ для: а) функции копулы $C(u_1, u_2)$ (формула (2.62)); б) плотности распределения копулы $c(u_1, u_2)$ (формула (2.59)).

Заметим также, что если проводить оценку параметра θ классическим методом максимального правдоподобия (соответствующей функцией Python для копулы Гумбеля), то полученные значения, количество которых в данном случае будет соответствовать количеству значений коэффициента корреляции

$\rho \in [0.1, 0.9]$ с шагом 0.05, т. е. их будет всего 17, на тех же 4913 тройках значений (ρ, u_1, u_2) приближение копулой Гумбеля показывает большие погрешности. В этом случае $MaxAPE \approx 12.385819\%$, $MinAPE \approx 0.000000\%$ и $MAPE \approx 1.037800\%$.

Что касается плотности копулы (2.59), то здесь результат сравнения с эмпирическими данными (данными имитационного моделирования) несколько хуже, однако остается приемлемым. В таблице 15 представлены значения относительных погрешностей приближения для $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$, $u_1, u_2 \in \{0.225, 0.250, ..., 0.875\}$, т. е. общее количество троек (ρ, u_1, u_2), для которых проводился расчет, составляет 3332. Некоторая суженность диапазона значений u_1, u_2 объясняется особенностями проведенного расчета эмпирической плотности копулы в данном конкретном случае. При этом несмотря на то, что максимальная относительная погрешность составляет около 11%, в общей совокупности рассмотренных значений число относительных ошибок, превышающих порог в 10%, всего 2. Количество погрешностей приближений, превышающих 5%, составляет всего примерно 1.95% от общего количества данных, что подтверждает показатель $MAPE$, который примерно равен 1.63%. Отметим, что погрешность приближения плотности копулы увеличивается с ростом значений u_1 и u_2, однако полученные оценки будут являться оценками сверху, кроме того, такое явление может объясняться недостаточностью количества испытаний в области верхних квантилей и высоких значений коэффициента загрузки. Как уже упоминалось выше, увеличение точности оценок имитационного моделирования в области значений ρ, а также квантилей, близких к единице, требует значительного увеличения длительности симуляции [102].

Сравнение приближения копулой Гумбеля с ранее известными результатами. Далее, проверим, насколько полученный результат согласуется с точной формулой для математического ожидания времени отклика, получен-

ной в [163], поскольку должно выполняться

$$E[R] = \int_0^{+\infty} [1 - F_R(x)]dx = \int_0^{+\infty} [1 - \delta(F(x))]dx = \int_0^{+\infty} [1 - \delta(1 - e^{-(\mu-\lambda)x})]dx, \quad (2.63)$$

где оценка диагонального сечения (исходя из копулы, подобранной в предыдущем разделе) имеет вид

$$\delta(u) \approx u^{2-C\rho}, \quad C \approx 0.369250. \quad (2.64)$$

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* (2.64) *копулы совместного распределения случайных времен пребывания подзаявок в подсистемах, то среднее время отклика системы аппроксимируется выражением*

$$E[R] \approx [\psi(3 - C\rho) - \psi(1)] \frac{\rho}{1 - \rho}, \quad (2.65)$$

где $\psi(\cdot)$ — дигамма-функция.

Доказательство. Из [115] известно, что случайная величина X, имеющая стандартное обобщенное экспоненциальное распределение с функцией и плотностью распределения вида

$$F_\eta(x) = (1 - e^{-x})^\alpha, \quad f_\eta(x) = \alpha(1 - e^{-x})^{\alpha-1}e^{-x}, \quad x \geqslant 0, \quad \alpha > 0,$$

имеет математическое ожидание

$$E[X] = \psi(\alpha + 1) - \psi(1) \quad (2.66)$$

и дисперсию

$$Var[X] = \psi'(1) - \psi'(\alpha + 1), \quad (2.67)$$

где $\psi(\cdot)$ — дигамма-функция, которая определяется как логарифмическая производная гамма-функции [204].

Далее в силу предположения (2.64) для интеграла из (2.63) имеем

$$E[R] = \int\limits_0^{+\infty} \left[1 - \delta\left(1 - e^{-(\frac{1}{\rho}-1)x}\right)\right] dx \approx \int\limits_0^{+\infty} \left[1 - \left(1 - e^{-(\frac{1}{\rho}-1)x}\right)^{2-C\rho}\right] dx =$$

$$= \int\limits_0^{+\infty} \left[1 - \left(1 - e^{-\frac{1-\rho}{\rho}x}\right)^{2-C\rho}\right] dx.$$

Соответственно, фактически, исходя из степенного диагонального сечения, получаем для распределения времени отклика приближение обобщенным показательным распределением общего вида [2]

$$F_R(x) \approx (1 - e^{-x/\beta})^{\alpha}, \quad x \geqslant 0, \quad \alpha, \beta > 0,$$

где $\alpha = 2 - C\rho$, $\beta = \rho/(1-\rho)$.

Поэтому с учетом (2.66) интеграл (2.63) окончательно преобразуется следующим образом

$$E[R] \approx \frac{\rho}{1-\rho} \int\limits_0^{+\infty} \left[1 - \left(1 - e^{-y}\right)^{2-C\rho}\right] dy =$$

$$= \left[\psi(3 - C\rho) - \psi(1)\right] \frac{\rho}{1-\rho}.$$

□

Согласно [163], среднее время отклика fork-join СМО с двумя подсистемами $M_\lambda|M_\mu|1$ равно

$$E[R] = \frac{12-\rho}{8} \cdot \frac{1}{\mu - \lambda}.$$

С учетом того, что в рассматриваемом случае для простоты полагается $\lambda = 1$ и, соответственно, $\mu = 1/\rho$, можно переписать данное выражение следующим образом

$$E[R] = \frac{(12-\rho)\rho}{8(1-\rho)}. \tag{2.68}$$

[2] Такое приближение ранее постулировалось в [165], однако в данном случае оно было выведено естественным образом, с эмпирическим и теоретическим обоснованием

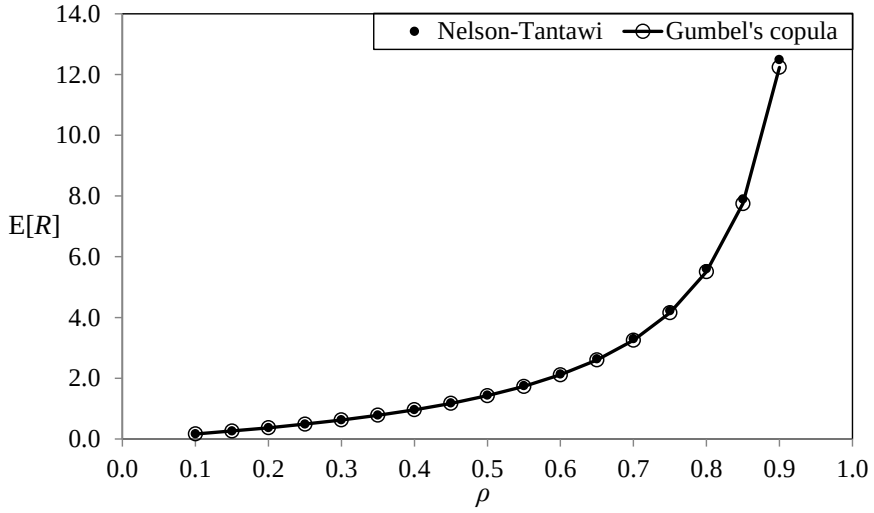


Рис. 38: Сравнение результатов вычислений среднего времени отклика fork-join СМО $E[R]$ с помощью точной формулы (2.68) и с помощью аппроксимации копулы Гумбеля (2.62) в соответствии с равенством (2.65).

На рисунке 38 можно сравнить приближения интеграла из (2.65) с истинными значениями среднего времени отклика (2.68) согласно [163] для значений $\rho \in \{0.10, 0.15, ..., 0.90\}$. Кроме того, в таблице 16 приведены результаты вычислений математического ожидания времени отклика, из которых следует, что модуль максимальной относительной погрешности приближения копулой Гумбеля не превышает для значений ρ из указанного диапазона 2.03%, откуда следует хорошая согласованность полученных аппроксимаций. Отметим, что если вместо $C \approx 0.369250$ взять $C \approx 0.370608$ из (2.52), то результат отличается крайне мало.

Поскольку среднее время отклика при $K = 2$ известно точно, разобранный пример имеет иллюстративный характер, однако данный подход может оказаться полезным при $K > 2$.

Применим теперь метод копул к оценке коэффициента корреляции Кендалла, который в отличие от коэффициентов Пирсона и Спирмена не вычисляется точно, и ранее был оценен формулой (2.40). Согласно [162, с. 164], для копулы

Таблица 16: Погрешность аппроксимации среднего времени отклика копулой Гумбеля в соответствии с формулой (2.65)

№	ρ	$E[R]_{NT}$	$E[R]_G$	Error, %
1	0.10	0.16527778	0.16503456	0.14716
2	0.15	0.26139706	0.26080338	0.22712
3	0.20	0.36875000	0.36760144	0.31148
4	0.25	0.48958333	0.48762334	0.40034
5	0.30	0.62678571	0.62369050	0.49382
6	0.35	0.78413462	0.77949216	0.59205
7	0.40	0.96666667	0.95994706	0.69513
8	0.45	1.18125000	1.17176220	0.80320
9	0.50	1.43750000	1.42432705	0.91638
10	0.55	1.74930556	1.73120372	1.03480
11	0.60	2.13750000	2.11273490	1.15860
12	0.65	2.63482143	2.60088709	1.28792
13	0.70	3.29583333	3.24893708	1.42290
14	0.75	4.21875000	4.15278226	1.56368
15	0.80	5.60000000	5.50421629	1.71042
16	0.85	7.89791667	7.75075620	1.86328
17	0.90	12.48750000	12.23495071	2.02242

Гумбеля (2.58) коэффициент корреляции Кендалла равен

$$r_k = 1 - \frac{1}{\theta},$$

откуда с учетом формулы (2.60), получаем приближение

$$r_k \approx 1 - \frac{\ln(2 - C\rho)}{\ln 2}. \tag{2.69}$$

Проведем анализ качества приближения с учетом результатов моделирования из [103].

К сожалению, при $C = 0.37$ (что соответствует значениям, полученным при оценке диагонального сечения и копулы) формула дает завышенные значения с ошибкой от 5.19% до 7.46%. Однако, если провести оптимизацию (2.69) по C методом Нелдера–Мида на основе результатов моделирования (как мы поступали и ранее), то при оптимальном значении $C \approx 0.349237$ (что близко к значению $C \approx 0.348284$, полученному при оценке квантилей времени отклика) получаем ошибку всего в 1.04%, что по качеству близко к эмпирическому приближению квадратичной функцией (2.6), полученному ранее. Последний вариант поэтому представляет собой альтернативное приближение коэффициента корреляции Кендалла.

Таким образом, приближения методом копул не всегда следует понимать буквально. Они могут подсказать удобные аналитические выражения для некоторых характеристик, в то время как параметры этих выражений могут потребовать уточнения с помощью дополнительной оптимизации на основе фактических данных.

Отметим еще интересный факт, что формула (2.69) позволяет заново оценить предельное значение коэффициента корреляции Кендалла при $\rho \to 1$. Ранее в Следствии 2.1, исходя из приближения (2.40), была найдена оценка коэффициента корреляции $r_k \approx 0.276$, теперь получаем значение, совпадающее с предыдущим с точностью до трех знаков, что говорит о хорошем соответствии.

Как выяснилось при дальнейших исследованиях, полученное приближение (копулы и диагонального сечения), к сожалению, не работает для оценки дисперсии (а значит, и среднего квадратического отклонения). С учетом (2.67), дисперсия нормированного времени отклика должна убывать с ростом загрузки, как и математическое ожидание, фактически же она возрастает, согласно результатам моделирования, и никакое $C > 0$ здесь не подходит. Это лишний раз показывает, что одно и то же приближение может быть хорошим для одних целей и плохим для других, поэтому в приложениях с этим следует быть осторожным.

2.7 Об особенностях имитационного моделирования системы

Система массового обслуживания с разделением и параллельным обслуживанием относится к трудно исследуемым системам. Точные аналитические результаты известны только для малого числа частных случаев. В большинстве работ этой тематики для анализа характеристик функционирования fork-join СМО было разработано множество приближенных методов, включая численные алгоритмы.

Эффективность того или иного метода может определяться по различным критериям. Наиболее значимыми среди них являются экономичность и высокая точность. Под экономичностью понимается необходимый объем ресурсов: чем меньше требуется памяти вычислительной машины или чем меньше время работы самого алгоритма, тем лучше. Что касается точности метода, то здесь требуются данные для сравнения. Единственно возможным инструментом их получения в данном контексте может выступить только имитационное моделирование. Таким образом, оценка точности различных методов исследования fork-join СМО напрямую зависит от точности полученных значений в результате имитационного моделирования.

Имитационное моделирование основывается на методе Монте-Карло, который заключается в многократном воспроизведении случайного процесса функционирования системы и дальнейшей статистической обработке полученных результатов. В результате имитационного моделирования fork-join СМО можно получить набор данных или вектор, представляющий собой последовательность реализаций случайной величины времени пребывания заявки в системе. Далее по этим данным необходимо построить точечную оценку среднего времени отклика и соответствующий доверительный интервал. Однако основная сложность построения доверительного интервала заключается в том, что полученные данные являются коррелированными. В частности, ясно, что время

пребывания n-й заявки в системе будет зависеть от времени пребывания предыдущей, т. е. $(n-1)$-й заявки, время пребывания которой, в свою очередь, зависит от времени пребывания $(n-2)$-й заявки, и т. д. Поэтому вопрос достаточного количества реализаций исследуемой величины в течении длительности одного прогона имитационной модели является довольно актуальным.

Конечно, на практике возможно экспериментальным путем определиться с необходимым размером последовательности, постепенно увеличивая количество ее элементов при каждом прогоне модели и наблюдая при этом, как меняется значение выборочного среднего времени отклика. Если эти изменения становятся незначительными, происходит остановка эксперимента и выбор конечного результата. Описанный подход к определению длительности прогона имитационной модели является наиболее распространенным.

В данном же разделе предлагается подход к построению доверительных интервалов для среднего времени отклика fork-join СМО в условиях коррелированности данных имитационного моделирования.

Оценка корреляции между реализациями случайной величины времени отклика fork-join системы R_i в течение длительности одного прогона, находящихся на расстоянии n, $n = 1, 2, ...$ определяется выражением

$$\hat{r}(n) = \frac{\hat{E}[R_i R_{i+n}] - \hat{E}[R]^2}{\hat{V}ar[R]}, \tag{2.70}$$

где

$$\hat{E}[R_i R_{i+n}] = \frac{1}{N-n} \sum_{i=1}^{N-n} R_i \cdot R_{i+n}, \quad \hat{\sigma}^2 = \hat{V}ar[R] = \hat{E}[R^2] - \hat{E}[R]^2,$$

$$\hat{R} = \hat{E}[R] = \frac{1}{N} \sum_{i=1}^{N} R_i, \quad \hat{E}[R^2] = \frac{1}{N} \sum_{i=1}^{N} R_i^2.$$

Далее рассчитаем дисперсию оценки времени отклика, получаемого с помощью имитационного моделирования:

$$Var[\hat{R}] = Var\left[\frac{1}{N} \sum_{i=1}^{N} R_i\right] =$$

$$= \frac{1}{N^2} Var\left[\sum_{i=1}^{N} R_i\right] = \frac{1}{N^2}\left(\sum_{i=1}^{N} Var[R_i] + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Cov(R_i, R_j)\right) =$$

$$= \frac{1}{N^2}\left(\sum_{i=1}^{N} Var[R_i] + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Corr(R_i, R_j)\sqrt{Var[R_i]Var[R_j]}\right) =$$

$$= \frac{1}{N^2}\left(N\sigma^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Corr(R_i, R_j)\sigma^2\right) =$$

$$= \frac{\sigma^2}{N}\left(1 + \frac{2}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Corr(R_i, R_j)\right) = \frac{\sigma^2}{N}\left(1 + 2\Delta_N\right),$$

где

$$\Delta_N = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Corr(R_i, R_j) = \frac{1}{N}\sum_{n=1}^{N-1}(N-n)r(n),$$

$r(n)$ — коэффициент корреляции между парой элементов (времен отклика), отстоящих друг от друга (по номерам обслуженных заявок) на расстояние n.

Тогда

$$\Delta_N = \frac{1}{N}\sum_{n=1}^{N-1}(N-n)r(n) = \sum_{n=1}^{N-1}\left(1 - \frac{n}{N}\right)r(n) \xrightarrow[N\to\infty]{} \sum_{n=1}^{\infty} r(n) = \Delta.$$

Возникает вопрос, как оценить сумму бесконечного ряда Δ, в то время как на практике можно оценить лишь конечное число автокорреляций $r(n)$. В работе [28] было предложено использовать для этого экспоненциальную асимптотику

$$r(n) \sim ae^{-bn}, \quad n \to \infty, \tag{2.71}$$

наблюдавшуюся для рассматриваемых там типов СМО. Расчеты в случае системы с разделением и параллельным обслуживанием также показывают хорошее соответствие экспоненциальному закону убывания автокорреляций (при больших ρ). Для наглядности на рисунке 39 приведен график зависимости выборочных автокорреляций, рассчитанных по формуле (2.70), от длины шага n при $\rho = 0.7$.

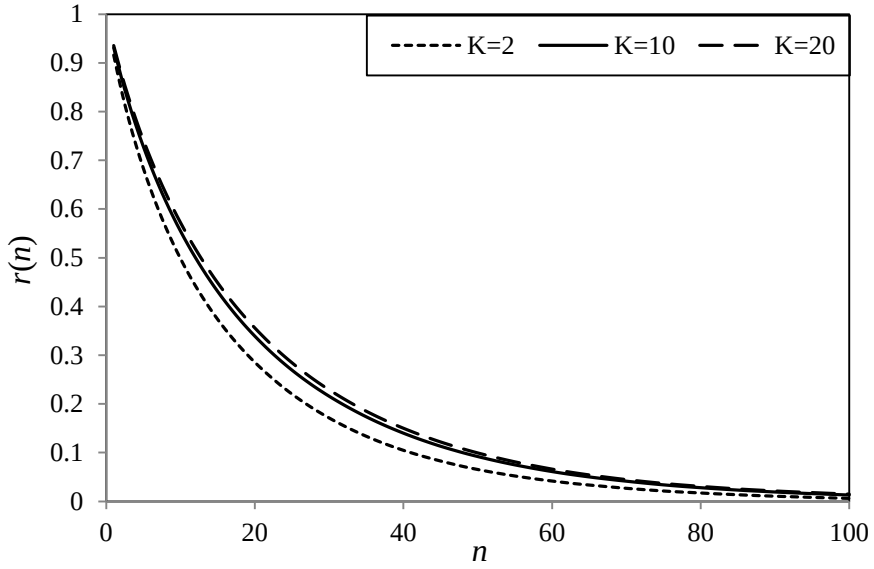


Рис. 39: Коэффициент корреляции между временами пребывания i-й и $(i+n)$-й заявками в системе, $\rho = 0.7$, $n = 1, 2, ...100$.

Теперь можем оценить Δ, рассчитав значения $\hat{r}(n)$ с помощью имитационного моделирования для конечного числа шагов, т. е. для $n = 1, 2, ..., m$,

$$\Delta = \sum_{n=1}^{+\infty} r(n) = \sum_{n=1}^{m} r(n) + \sum_{n=m+1}^{+\infty} r(n), \tag{2.72}$$

с учетом того, что в силу (2.71) верно

$$\sum_{n=m+1}^{+\infty} r(n) \sim a\frac{e^{-b(m+1)}}{1-e^{-b}}, \quad m \to \infty. \tag{2.73}$$

Возможны две ситуации. Суммируем $\hat{r}(n)$ до тех пор, пока справедливо, что $\hat{r}(n) \geqslant 2/\sqrt{N}$, в противном случае считаем значения $\hat{r}(n)$ статистически не значимыми. В первом случае для суммирования может оказаться достаточным число полученных элементов $\hat{r}(n)$, $n = 1, 2, ..., m$, т. е. $\hat{r}(m) < 2/\sqrt{N}$.

В противном случае, когда $\hat{r}(m) \geqslant 2/\sqrt{N}$, необходимо будет оценить остатки суммирования для n от $(m+1)$ до $+\infty$ из (2.73). Сделать это можно с помощью построения регрессионной модели. Полученную оценку для хвоста суммы прибавляем к сумме по n от 1 до m. В результате описанных действий мы вычислим некоторую оценку $\hat{\Delta}$ для Δ из (2.72).

Таким образом, для конкретных значений параметров рассматриваемой системы, например, для определенного значения коэффициента загрузки ρ и числа подсистем K имеем следующий **алгоритм**:

1. выбирается некоторое значение m;

2. на основе результатов имитационного моделирования для большого числа реализаций N вычисляется набор значений $\hat{r}(n)$ с помощью формулы (2.70), $n = 1, 2, ...m$;

3. суммируются элементы $\hat{r}(n)$ до тех пор, пока $\hat{r}(n) < 2/\sqrt{N}$, $n = 1, 2, ...m$;

4. если $\hat{r}(m) \geqslant 2/\sqrt{N}$

   1) строится регрессионная модель вида (2.71), которая приводится к линейной посредством логарифмирования обеих частей выражения: $\ln r(n) = \ln a - bn + \varepsilon$ (ε — случайная ошибка модели);

   2) оцениваются параметры a и b построенной регрессионной модели с помощью метода наименьших квадратов;

   3) вычисляется оценка хвоста суммирования по формуле (2.73) с помощью подстановки в неё значения полученных оценок параметров регрессии a и b, а также m;

   4) вычисляется оценка Δ согласно формуле (2.72) с помощью подстановки в неё соответствующие значения

$$\hat{\Delta} = \sum_{n=1}^{m} \hat{r}(n) + \hat{a}\frac{e^{-\hat{b}(m+1)}}{1 - e^{-\hat{b}}};$$

5. если $\hat{r}(n^* + 1) < 2/\sqrt{N}$, причём $n^* \leqslant m$, получаем оценку

$$\hat{\Delta} = \sum_{n=1}^{n^*} \hat{r}(n).$$

В силу асимптотической нормальности оценки среднего, можем строить асимптотические доверительные интервалы произвольных уровней надежности, например, применить правило “трёх сигм”. Тогда с вероятностью $p = 0.997$ будет справедливо

$$E[R] = \hat{E}[R] \pm 3\sigma_{\hat{E}[R]} \approx \hat{E}[R] \pm \frac{3\sigma}{\sqrt{N}}\sqrt{1+2\Delta_N} \approx \hat{E}[R] \pm \frac{3\hat{\sigma}}{\sqrt{N}}\sqrt{1+2\hat{\Delta}}. \quad (2.74)$$

Как видно, по сравнению со случаем независимых случайных величин, количество испытаний N при одном прогоне модели потребуется больше в $(1+2\Delta_N)$ раз из-за имеющейся корреляции между данными.

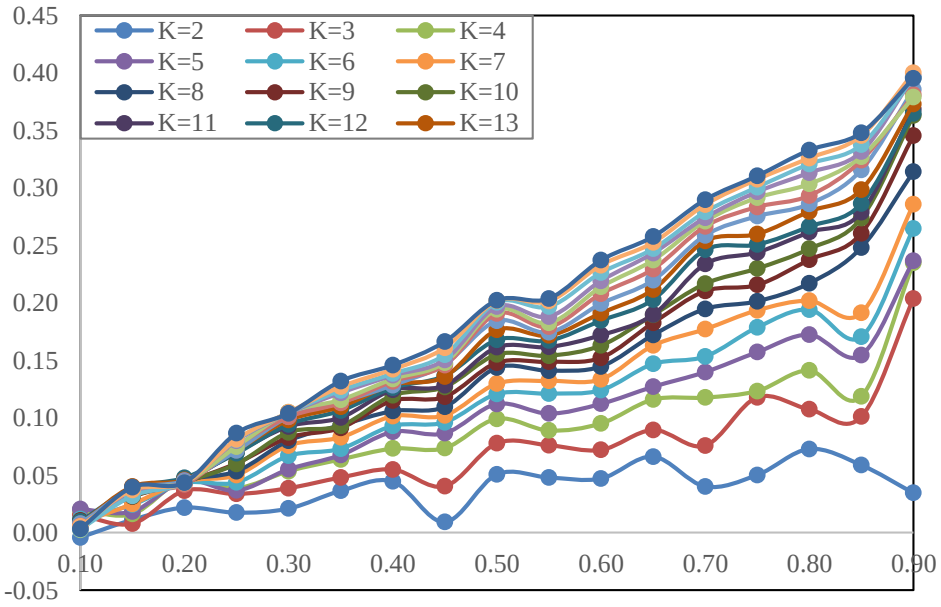


Рис. 40: Значения $\hat{\Delta}/(2\rho/(1-\rho^2)) - 1$ в зависимости от ρ.

Далее построим оценку величины Δ для fork-join СМО с K подсистемами типа $M|M|1$ в виде некоторого аналитического выражения в зависимости от величин ρ и K. Для этого с помощью имитационного моделирования для различных комбинаций пар значений $K = 2, ..., 20$ и $\rho = 0.1, 0.2, ...0.9$ по описанному выше алгоритму рассчитаем значения $\hat{\Delta} = \hat{\Delta}(K, \rho)$. При этом значения m выбирались в зависимости от ρ, а именно $m = 10, 20, 30, 40, 50, 50, 100, 250, 1000$.

Отметим, что здесь начинаем моделирование с $K = 2$, поскольку для данной характеристики даже в этом случае до сих пор нет точных результатов, как не

было и приближенных.

Теперь определим конкретный вид функциональной зависимости с помощью анализа результатов построения различных графиков. В частности, график $\hat{\Delta}/(2\rho/(1-\rho^2)) - 1$ в зависимости от ρ напоминает пучок прямых с различным углом наклона, проходящих через точку $(0, 0)$, с учетом колебаний, вызванных случайными ошибками (рис. 40). Поэтому остается подобрать формулу, описывающую эти прямые.

Учитывая полученные в ранних публикациях приближения для среднего времени отклика fork-join системы, в которых присутствует частичная сумма гармонического ряда, естественно предположить, что Δ будет аналогично зависеть от K не напрямую, а через H_K. Относительно H_K также наблюдалась зависимость, близкая к линейной. Поэтому допустим, что прямые можно описать формулой $C \cdot \rho(H_K - 1)$. В конечном итоге получаем выражение для оценки Δ следующего вида:

$$\Delta(K, \rho) \approx \frac{2\rho}{(1-\rho)^2}\big(1 + C \cdot \rho(H_K - 1)\big), \tag{2.75}$$

где C — это некоторый коэффициент, значение которого необходимо найти. Поиск коэффициента C осуществляем с помощью метода оптимизации Нелдера–Мида. Таким образом, получаем

$$C \approx 0.186722.$$

Чтобы удостовериться в качестве полученной аппроксимации, построим соответствующий график (рис. 41) и проанализируем ошибки приближения. Пусть

$$APE = \frac{\hat{\Delta} - \Delta}{\Delta} \cdot 100\%,$$

тогда получаем

$$MaxAPE \approx 5.27449\%, \quad MinAPE \approx 0.013941\%,$$

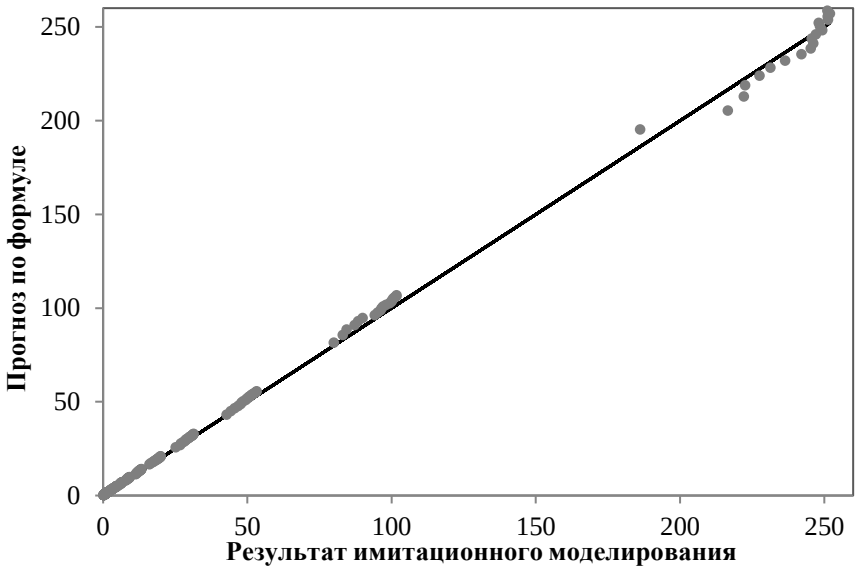


Рис. 41: Коэффициент Δ: сравнение результатов имитационного моделирования с формулой (2.75).

причем среднее значение модулей относительных ошибок приближения

$$MAPE \approx 0.835525\%,$$

что свидетельствует о хорошем качестве полученного выражения. При этом можно предположить, что формула (2.75) может успешно использоваться для построения доверительных интервалов не только в области $2 \leqslant K \leqslant 20$, но и для больших значений K.

Отметим также, что значения Δ, наблюдаемые в изучаемой области параметров, достигают примерно 252, что означает в таком случае увеличение необходимого числа испытаний N более чем в 500 раз по сравнению со случаем независимых испытаний (для достижения той же точности оценки среднего времени отклика), и подобным эффектом пренебрегать нельзя.

Представленный выше алгоритм можно использовать для построения доверительных интервалов вида (2.74) для оценки среднего времени отклика не только в случае подсистем типа $M|M|1$, но и в более сложных случаях, например, для подсистем типа $G|G|1$ и др.

Что касается метода построения функциональной зависимости для коэффи-

циента Δ, то здесь возможна более трудоемкая работа по подбору вида зависимости в каждом конкретном случае, т. е. для конкретных распределений входящего и обслуживающих потоков. В частности, зависимость от H_K естественно возникает для показательного распределения, а например, в случае распределений с тяжелыми (степенными) хвостами (например, как в работе [105]) будет естественна зависимость от некоторой степени K. Тем не менее, общий подход к построению оценки для Δ, описанный на примере вывода выражения (2.75), может быть применим и для других типов подсистем fork-join СМО.

2.8 Выводы к главе 2

В Главе 2 представлены результаты применения нового метода с использованием машинного обучения (нейронных сетей) к анализу характеристик системы с разделением и параллельным обслуживанием. Обученная нейронная сеть демонстрирует хорошее качество приближения для математического ожидания и дисперсии времени отклика системы. Также в рамках численного эксперимента проведен сравнительный анализ с полученными ранее аппроксимациями, обученная нейросеть на заданном наборе параметров показала лучшее приближение.

Предложен новый комплексный метод, который может базироваться, в том числе и на известных приближениях для математического ожидания и дисперсии времени пребывания заявки в системе. С помощью визуализации данных, а также применения метода оптимизации Нелдера–Мида были преобразованы известные ранее приближения характеристик системы с разделением и параллельным обслуживанием с K подсистемами $M|M|1$ и значительно (в разы) улучшено их качество аппроксимации. Результаты получены в виде аналитических формул.

Кроме того, в главе представлены точные аналитические выражения для коэффициентов корреляции Пирсона и Спирмена между временами пребыва-

ния подзаявок в подсистемах fork-join СМО. Данные формулы были получены с помощью классического метода производящих функций и преобразований Лапласа–Стилтьеса.

Для коэффициента корреляции Кендалла было получено приближенное выражение, точность аппроксимации которого достаточно высока. Для вывода оценки корреляции Кендалла использовался графический анализ данных и метод оптимизации Нелдера–Мида. Полученные результаты в рамках численного эксперимента сравнивались с данными имитационного моделирования, которое подтвердило их корректность.

Показано, что все коэффициенты растут с увеличением загрузки, причем коэффициент корреляции Пирсона квадратичным образом (с замедлением), а Спирмена и Кендалла более сложным нелинейным образом (но близким к линейному). При высокой загрузке они приближаются к некоторым предельным значениям, обусловленным свойствами предельного двумерного распределения нормированных времен пребывания подзаявок.

Также построена мета-гауссовская модель, которая позволяет с использованием коэффициентов корреляции, получить оценки не только среднего времени отклика, но и дисперсии, квантилей и других характеристик. Мета-гауссовская модель, с одной стороны, требует проведения симуляции, а не дает явную формулу для оценки характеристик, однако, с другой стороны, эта симуляция гораздо проще и быстрее, чем симуляция исходной fork-join системы массового обслуживания.

Кроме того, изучены приближения совместного распределения времен пребывания подзаявок с помощью теории копул. Получено хорошее соответствие с данными для степенных диагональных сечений и копулы Гумбеля. На основе оценок диагональных сечений выведены оценки квантилей времени отклика, в широком диапазоне уровней и загрузок. Получена также новая оценка коэффициента корреляции Кендалла.

В Главе 2 описываются и важные аспекты проведения имитационного мо-

делирования системы с разделением и параллельным обслуживанием, а также оценки его результатов. Приводится алгоритм построения доверительных интервалов для выборочных оценок среднего времени отклика, полученных в результате симуляции функционирования fork-join системы, а также даются некоторые рекомендации.

3 Получение основных стационарных характеристик систем с разделением и параллельным обслуживанием в случае распределения Парето времени обслуживания

В Главе 3 исследуется система с разделением и параллельным обслуживанием в случае распределения Парето времени обслуживания. Результаты Главы 3 отражены в публикациях [18, 25, 101, 105, 110].

3.1 Математическая модель системы с разделением и параллельным обслуживанием

Рассматривается система с разделением и параллельным обслуживанием заявок более общего вида. Считаем, что случайное время между соседними поступлениями заявок ζ имеет функцию распределения $F_\zeta(x)$, причем $E[\zeta] = a$. Случайное время обслуживания на приборах η имеет функцию распределения $F_\eta(x)$ с математическим ожиданием $E[\eta] = b$, а коэффициент загрузки системы $\rho = b/a < 1$.

В дальнейшем будет рассмотрено несколько вариантов распределений для входящего потока системы с разделением и параллельным обслуживанием, о чем более детально будет сказано в соответствующем разделе (раздел 3.6). Однако большая часть исследований посвящена именно случаю пуассоновского входящего потока, т. е. $F_\zeta(x) = 1 - e^{-\lambda x}$, $x \geq 0$, тогда

$$E[\zeta] = a = \frac{1}{\lambda}, \quad \rho = \lambda b.$$

Поэтому здесь и далее, пока не будет указано обратное, считаем, что время между соседними поступлениями заявок является показательным, а подсистемы fork-join СМО имеют вид $M|G|1$.

Случайная величина времени обслуживания подзаявок η на каждом из приборов имеет распределение Парето со следующей функцией распределения

$$F_\eta(x) = 1 - \left(\frac{\alpha-1}{\alpha}\cdot\frac{1}{x}\right)^\alpha, \quad x \geq \frac{\alpha-1}{\alpha}, \quad \alpha > 3, \tag{3.1}$$

тогда первые три момента времени обслуживания равны

$$E[\eta] = b = b_{Pa} = 1, \tag{3.2}$$

$$E[\eta^2] = b^{(2)} = b_{Pa}^{(2)} = \frac{(\alpha-1)^2}{\alpha(\alpha-2)}, \tag{3.3}$$

$$E[\eta^3] = b^{(3)} = b_{Pa}^{(3)} = \frac{(\alpha-1)^3}{\alpha^2(\alpha-3)}. \tag{3.4}$$

Таким образом, среднее время обслуживания на приборе будет равно одной условной временной единице, что в общем-то удобно и в контексте проведения численного анализа. Предполагается, что все приборы являются однородными.

Особенности функционирования fork-join системы в данном случае ничем не отличаются от случая подсистем типа $M|M|1$ (рис. 42):

1) в момент поступления заявки в систему она мгновенно разделяется на K ($K \geqslant 2$) подзаявок, каждая из которых становится в свою очередь на обслуживание к прибору (если он занят) или мгновенно начинает обслуживаться, если соответствующий прибор свободен;

2) после окончания обслуживания подзаявка остается в системе до тех пор, пока все родственные ей подзаявки, т. е. подзаявки, изначально составляющие одну заявку, не закончат свое обслуживание, далее происходит мгновенная сборка целой заявки и только после этого заявка считается обслуженной и может покинуть систему.

Поскольку заявка остается в системе до момента окончания обслуживания последней ее составляющей подзаявки, то время отклика системы R_K как и ранее будет являться максимумом из K случайных величин ξ_1, ξ_2, ..., ξ_K времен

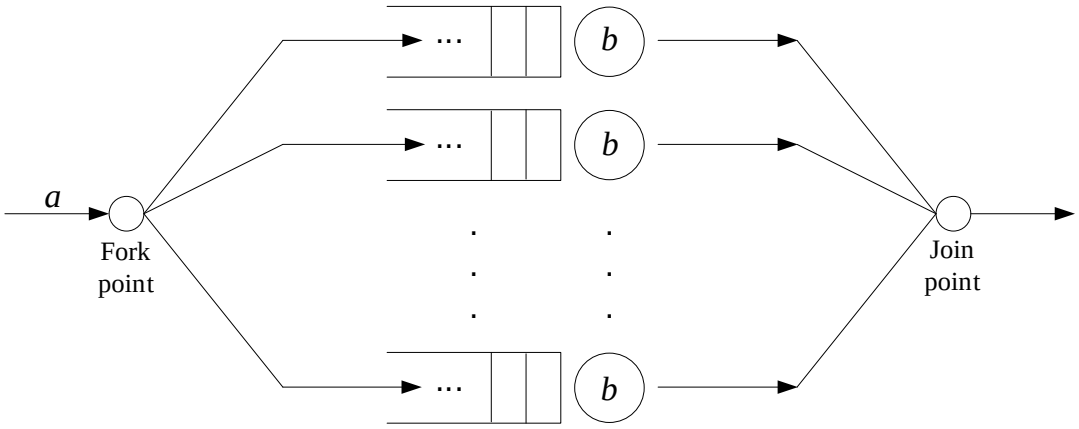


Рис. 42: Модель fork-join СМО с K подсистемами типа $G|G|1$.

пребывания подзаявок в подсистемах

$$R_K = \xi_{(K)} = \max(\xi_1, ..., \xi_K).$$

Данная величина согласно теории математической статистики называется максимальной (K-ой) порядковой статистикой [88].

3.2 Аналитические оценки времени отклика в fork-join системе

В случае с произвольным распределением времен обслуживания, т. е. в случае с K подсистемами типа $M|G|1$ в системе с разделением заявок, известно не так много формул для аппроксимации среднего времени отклика по сравнению со случаем экспоненциального обслуживания. Причем большинство из этих приближений основывается на теории порядковых статистик, поскольку случайная величина времени отклика по своей сути является K-й порядковой статистикой. При этом для упрощения анализа экстремальных значений, как правило, делается допущение о независимости указанных случайных величин (хотя на самом деле они положительно зависимы).

Так, например, для одинаково распределенных и независимых времен пребывания заявок в подсистемах с математическим ожиданием $E[\xi_k] = \mu$ и вто-

рым моментом $E[\xi_k^2] = \mu^{(2)}$ в [118] представлена формула

$$E[R_K] \approx \mu + \frac{\mu^{(2)}}{2\mu}(H_K - 1), \quad H_K = \sum_{i=1}^{K} \frac{1}{i}. \tag{3.5}$$

В [191] можно найти выражение для определения $E[\xi_{(K)}]$ в условиях различных распределений, и, соответственно, различных первых и вторых моментов случайных величин ξ_k. Кроме того, для некоторых конкретных распределений времен обслуживания, например, таких как распределение Эрланга, получены явные выражения для математического ожидания экстремального значения $\xi_{(K)}$ [191, 192].

Также теория порядковых статистик позволяет определить верхнюю границу для математического ожидания величины $\xi_{(K)}$, которую в свою очередь, можно использовать для аппроксимации среднего времени отклика [88]:

$$E[R_K] \approx \mu + \sigma \frac{K-1}{\sqrt{2K-1}}, \tag{3.6}$$

где σ — это среднеквадратическое отклонение.

Для подсистем типа $M_\lambda|G|1$ не всегда представляется возможным получить конкретное распределение времени пребывания в явном виде. Тем не менее, в этой ситуации, чтобы воспользоваться формулой (3.5) или (3.6), достаточно установить его первые два момента. Формула математического ожидания времени пребывания в $M_\lambda|G|1$ широко известна и имеет вид

$$\mu = b + \frac{\lambda b^{(2)}}{2(1-\rho)}, \tag{3.7}$$

где b — среднее время обслуживания на приборе, а $b^{(2)}$ — второй момент этого времени, $\rho = \lambda b$ — загрузка системы. Для того, чтобы получить второй момент времени пребывания в СМО $M_\lambda|G|1$ достаточно дважды продифференцировать преобразование Лапласа-Стилтьеса (ПЛС) соответствующей функции распределения $\psi(s)$ и вычислить ее предельное значение в нуле, т. е.

$$\mu^{(2)} = \lim_{s\to 0} \frac{d^2\psi(s)}{ds^2}, \quad \psi(s) = \frac{s(1-\rho)\beta(s)}{s - \lambda + \lambda\beta(s)},$$

где $\beta(s)$ — ПЛС функции распределения времени обслуживания.

Сформулируем утверждение

Утверждение 3.1. *Второй момент времени пребывания в подсистеме типа* $M|G|1$ *системы с разделением и параллельным обслуживанием с пуассоновским входным потоком с параметром* λ *определяется выражением*

$$\mu^{(2)} = b^{(2)} + \frac{\lambda b^{(3)} + 3\rho b^{(2)}}{3(1-\rho)} + \frac{\lambda^2 (b^{(2)})^2}{2(1-\rho)^2}, \tag{3.8}$$

где b, $b^{(2)}$ *и* $b^{(3)}$ *— первый, второй и третий моменты времени обслуживания, а* $\rho = \lambda b$ *— загрузка системы.*

Доказательство. Через $\varphi(s)$ обозначим ПЛС времени ожидания начала обслуживания, выражение для которого имеет вид [6]

$$\varphi(s) = (1-\rho) \cdot \frac{s}{s - \lambda + \lambda\beta(s)},$$

тогда по свойства ПЛС $\psi(s) = \varphi(s) \cdot \beta(s)$. Поэтому

$$\psi''(s) = \varphi''(s)\beta(s) + \varphi(s)\beta''(s) + 2\varphi'(s)\beta'(s).$$

Далее определим значение каждой компоненты данной формулы в нуле. По свойствам ПЛС известно, что

$$\varphi(0) = \beta(0) = 1, \quad \beta'(0) = -b, \quad \beta''(0) = b^{(2)}, \quad \beta'''(0) = -b^{(3)},$$

следовательно

$$\psi''(0) = \varphi''(0) + \beta''(0) + 2\varphi'(0)\beta'(0). \tag{3.9}$$

Для ПЛС времени ожидания начала обслуживания имеем

$$\varphi'(s) = (1-\rho) \cdot \frac{s - \lambda + \lambda\beta(s) - s(1 + \lambda\beta'(s))}{(s - \lambda + \lambda\beta(s))^2} =$$

$$= (1-\rho) \cdot \frac{-\lambda + \lambda\beta(s) - \lambda s\beta'(s)}{(s - \lambda + \lambda\beta(s))^2}.$$

Числитель и знаменатель полученной дроби являются бесконечно малыми второго порядка при $s \to 0$, поэтому, чтобы вычислить значение $\varphi'(0)$ дважды воспользуемся правилом Лопиталя

$$\varphi'(0) = (1-\rho) \cdot \lim_{s\to 0} \frac{-\lambda + \lambda\beta(s) - \lambda s\beta'(s)}{(s - \lambda + \lambda\beta(s))^2} =$$

$$= (1-\rho) \cdot \lim_{s\to 0} \frac{-\lambda s\beta''(s)}{2(s - \lambda + \lambda\beta(s))(1 + \lambda\beta'(s))} =$$

$$= (1-\rho) \cdot \lim_{s\to 0} \frac{-\lambda\beta''(s) - \lambda s\beta'''(s)}{2\left((1 + \lambda\beta'(s))^2 + (s - \lambda + \lambda\beta(s)\lambda\beta'(s))\right)} =$$

$$= (1-\rho) \cdot \frac{-\lambda\beta''(0)}{2(1 + \lambda\beta'(0))^2} = (1-\rho) \cdot \frac{-\lambda b^{(2)}}{2(1-\lambda b)^2} =$$

$$= (1-\rho) \cdot \frac{-\lambda b^{(2)}}{2(1-\rho)^2} = \frac{-\lambda b^{(2)}}{2(1-\rho)}.$$

Далее после некоторого упрощения получаем, что

$$\varphi''(s) = (1-\rho) \cdot \frac{A(s)}{B(s)},$$

где

$$A(s) = -\lambda s\beta''(s)(s - \lambda + \lambda\beta(s)) - 2(-\lambda + \lambda\beta(s) - \lambda s\beta'(s))(1 + \lambda\beta'(s)),$$

$$B(s) = (s - \lambda + \lambda\beta(s))^3.$$

Функции $A(s)$ и $B(s)$ являются бесконечно малыми третьего порядка при $s \to 0$, поэтому, чтобы вычислить необходимый предел воспользуемся правилом Лопиталя трижды

$$\varphi''(0) = (1-\rho) \cdot \lim_{s\to 0} \frac{A(s)}{B(s)} = (1-\rho) \cdot \frac{A'''(0)}{B'''(0)}. \tag{3.10}$$

Итак,

$$B'(s) = 3(s - \lambda + \lambda\beta(s))^2(1 + \lambda\beta'(s)),$$

$$B''(s) = 6(s - \lambda + \lambda\beta(s))(1 + \lambda\beta'(s))^2 + 3(s - \lambda + \lambda\beta(s))^2\lambda\beta''(s).$$

Понятно, что большая часть слагаемых в выражении для третьей производной функции $B(s)$ будет содержать множитель $(s - \lambda + \lambda\beta(s))$, который при подстановке $s = 0$, обратится в ноль, поэтому сразу можем записать единственный ненулевой элемент и, соответственно, искомое значение

$$B'''(0) = 6(1 + \lambda\beta'(0))^3 = 6(1 - \lambda b)^3 = 6(1 - \rho)^3.$$

Теперь для $A(s)$

$$A'(s) = (-\lambda\beta''(s) - \lambda s\beta'''(s))(s - \lambda + \lambda\beta(s)) +$$

$$+\lambda s\beta''(s)(1 + \lambda\beta'(s)) - 2\lambda\beta''(s)(-\lambda + \lambda\beta(s) - \lambda s\beta'(s)),$$

$$A''(s) = (-2\lambda\beta'''(s) - \lambda s\beta^{(4)}(s))(s - \lambda + \lambda\beta(s)) +$$

$$+3\lambda^2 s(\beta''(s))^2 - 2\lambda\beta'''(s)(-\lambda + \lambda\beta(s) - \lambda s\beta'(s)).$$

Далее опять же с учетом того, что часть слагаемых, содержащих множители $(s - \lambda + \lambda\beta(s))$, $(-\lambda + \lambda\beta(s) - \lambda s\beta'(s))$ и s, при $s = 0$ обратится в ноль, можем записать

$$A'''(0) = -2\lambda\beta'''(0)(1 + \lambda\beta(0)) + 3\lambda^2(\beta''(0))^2 =$$

$$= 2\lambda b^{(3)}(1 - \lambda b) + 3\lambda^2(b^{(2)})^2 = 2\lambda b^{(3)}(1 - \rho) + 3\lambda^2(b^{(2)})^2.$$

Теперь, подставляя полученные значения для $A'''(0)$ и $B'''(0)$ в (3.10), окончательно получаем

$$\varphi''(0) = (1 - \rho) \cdot \frac{2\lambda b^{(3)}(1 - \rho) + 3\lambda^2(b^{(2)})^2}{6(1 - \rho)^3} =$$

$$= \frac{2\lambda b^{(3)}(1 - \rho) + 3\lambda^2(b^{(2)})^2}{6(1 - \rho)^2} = \frac{\lambda b^{(3)}}{3(1 - \rho)} + \frac{\lambda^2(b^{(2)})^2}{2(1 - \rho)^2}.$$

Наконец, можем подставить все рассчитанные компоненты в (3.9), не забывая, что $\rho = \lambda b$,

$$\psi''(0) = \frac{\lambda b^{(3)}}{3(1 - \rho)} + \frac{\lambda^2(b^{(2)})^2}{2(1 - \rho)^2} + b^{(2)} + \frac{\rho b^{(2)}}{1 - \rho} =$$

$$= b^{(2)} + \frac{\lambda b^{(3)} + 3\rho b^{(2)}}{3(1 - \rho)} + \frac{\lambda^2(b^{(2)})^2}{2(1 - \rho)^2}.$$

□

Отметим, что рассматриваются только аналитические оценки, не использующие результатов имитационного моделирования и основанные на моментах времени обслуживания. С одной стороны, это позволяет их применить к произвольным распределениям, чьи моменты существуют и известны. С другой стороны, оценки не используют более подробную информацию о распределениях, и для различных распределений с одними и теми же значениями моментов могут работать по-разному, лучше или хуже.

Учитывая, что рассматривается модель fork-join СМО с распределением Парето для времени обслуживания, можно сформулировать следующую теорему, касающуюся верхней границы среднего времени отклика и которая может быть использована для оценки этой характеристики.

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

$$E[R_K] \leqslant \mu_{Pa} + \sigma_{Pa}\frac{K-1}{\sqrt{2K-1}}, \tag{3.11}$$

где

$$\mu_{Pa} = 1 + \frac{\rho(\alpha-1)^2}{2\alpha(\alpha-2)(1-\rho)}, \tag{3.12}$$

$$\sigma_{Pa} = \sqrt{\mu_{Pa}^{(2)} - \mu_{Pa}^2}, \tag{3.13}$$

а

$$\begin{gathered}\mu_{Pa}^{(2)} = \frac{(\alpha-1)^2}{\alpha(\alpha-2)} + \frac{\rho(\alpha-1)^3}{3\alpha^2(\alpha-3)(1-\rho)} + \\ + \frac{\rho(\alpha-1)^2}{\alpha(\alpha-2)(1-\rho)} + \frac{\rho^2(\alpha-1)^4}{2\alpha^2(\alpha-2)^2(1-\rho)^2}.\end{gathered} \tag{3.14}$$

Доказательство. Выражение (3.11) справедливо в соответствии с элементами теории порядковых статистик. В частности, известно, что верхняя граница для K-ой порядковой статистики, т. е. максимума среди K независимых одинаково распределенных случайных величин с математическим ожиданием μ и

дисперсией σ^2 имеет вид (3.6) [88]. Однако времена пребывания подзаявок в системе не являются независимыми, а наоборот — они положительно ассоциированные (коррелированные) случайные величины, а в силу этого их максимум стохастически не больше максимума независимых случайных величин с тем же распределением [163]. Поэтому верхняя граница (3.11) сохранится и для них.

Таким образом, остается только получить выражения для математического ожидания и среднеквадратического отклонения времени пребывания подзаявки в СМО типа $M_\lambda|Pa|1$.

В соответствии с формулами (3.7) и (3.8) после подстановки в них выражений (3.2), (3.3), (3.4) первый и второй моменты времени пребывания подзаявки с учетом того, что $\rho = \lambda$, будут определяться после преобразований следующим образом

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

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

Среднеквадратическое отклонение времени пребывания подзаявки в соответствии с определением равно

$$\sigma_{Pa} = \sqrt{\mu_{Pa}^{(2)} - \mu_{Pa}^2}.$$

□

Более общий случай, когда времена обслуживания распределены по Парето с функцией распределения

$$\widetilde{F}(x) = 1 - \left(\frac{\beta}{x}\right)^\alpha, \quad x \geqslant \beta > 0, \quad \alpha > 3,$$

(двухпараметрическое семейство распределений) сводится к (3.1) заменой времени

$$t \longmapsto \frac{\alpha-1}{\alpha\beta}t,$$

откуда получаем

$$E[\widetilde{R}_K] = \frac{\alpha\beta}{\alpha-1}E[R_K], \quad Var[\widetilde{R}_K] = \left(\frac{\alpha\beta}{\alpha-1}\right)^2 Var[R_K],$$

при

$$\lambda = \frac{\alpha\beta}{\alpha - 1}\widetilde{\lambda},$$

где с помощью "$\sim$" обозначены параметры и характеристики новой (более общей) системы. При этом загрузка ρ не меняется.

Стоит заметить, что характеристики времени пребывания подзаявки, описываемые формулами (3.12),(3.13) и (3.14) могут служить оценками соответствующих характеристик времени отклика и сами по себе (если при каких-то параметрах системы обслуживание подзаявок происходит достаточно синхронно).

3.3 Оценка основных характеристик системы с помощью ИНС

Концепция применения искусственных нейронных сетей (или каких-либо других методов, входящих в довольно обширное понятие интеллектуального анализа данных) к анализу систем или сетей массового обслуживания подробно описана в Главе 1. Заключается она, если кратко, в комбинации имитационного моделирования на ограниченном наборе входных параметров на заранее определенных числовых интервалах с последующим обучением на полученных данных нейросети, которая в дальнейшим позволит получить оценки интересующих характеристик системы уже для других (например, промежуточных) значений входных параметров.

Итак, в соответствии с описанным методом, сперва требуется разработать имитационную модель исследуемой системы, чтобы с ее помощью получить набор данных как для обучения нейросети, так и для ее проверки потом. Напомним, что мы рассматриваем fork-join систему с K ветвями типа $M|G|1$ с распределением Парето времени обслуживания из (3.1). Для этого необходимо определиться с перечнем входных параметров и с интервалом принимаемых ими значений, а также выходными характеристиками, которые будут рассчи-

тываться на этапе симуляции модели.

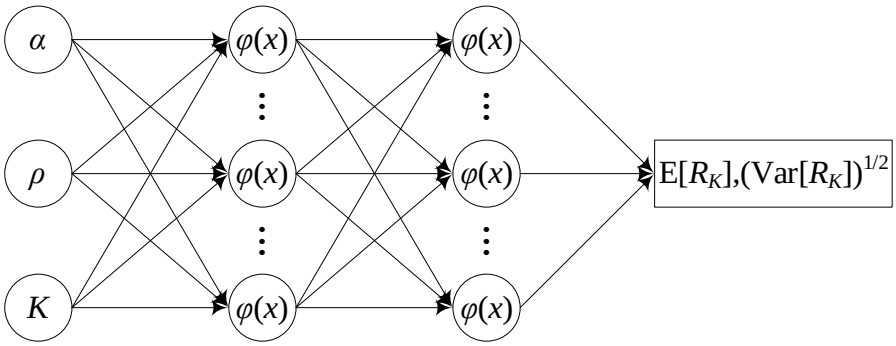


Рис. 43: Схема персептрона для оценки характеристик производительности fork-join системы с подсистемами типа $M|G|1$.

Допустим, что параметр α будет принимать только целочисленные значения на отрезке $[4.0, 10.0]$, количество подсистем K будет меняться от 2 до 20 включительно, а коэффициент загрузки ρ — от 0.1 до 0.9 с шагом 0.1 (табл. 17). В качестве характеристик производительности системы, что естественно, оста-

Таблица 17: Входные данные для обучения ИНС, а также для построения аналитических оценок среднего времени отклика и его среднеквадратического отклонения в разделе 3.4

№ п/п	1	2	...	19	20	21	...	1197
α	4.0	4.0	...	4.0	4.0	4.0	...	10.0
ρ	0.1	0.1	...	0.1	0.2	0.2	...	0.9
K	2	3	...	20	2	3	...	20

новим свой выбор на среднем времени пребывания в СМО $E[R_K]$, а также на среднеквадратическом отклонении этой случайной величины $\sqrt{Var[R_K]}$. Таким образом, необходимо будет получить 1197 наборов данных.

Имитационная модель описанной fork-join системы была разработана в программной среде Python, в ней же проходило обучение искусственной нейросети. Хотя, как упоминалось ранее, для моделирования СМО можно воспользоваться и готовыми коммерческими программными продуктами, такими, как например,

GPSS World, AnyLogic, Arena. В качестве структуры нейросети был выбран трехслойный персептрон с 10 нейронами в каждом из двух скрытых слоев с логистической функцией активации $\varphi(x) = 1/(1 + e^{-x})$ (рис. 43). Выходной слой состоит всего из одного нейрона, отвечающего за какую-то одну из оцениваемых характеристик, т. е. фактически были построены две нейросети, за счет чего ожидаемо должна была повыситься точность прогноза. Кроме того, с той же целью входные данные были подвергнуты предварительной стандартизации и нормализации.

Далее выборка из табл. 17 случайным образом была разбита на обучающую и тестовую в соотношении 80% и 20%, соответственно. В результате, тренировка данных происходила на 958 условных единицах из обучающего набора данных методом Адама, являющимся по сути расширением классического алгоритма градиентного спуска [136]. В процессе обучения из тренировочного набора выделялась валидационная выборка, по которой оценивалось качество обучения после прохождения каждой эпохи обучения.

Критерием оценки погрешности прогноза были выбраны среднеквадратическая ошибка (MSE), средняя абсолютная ошибка (MAE) и средняя абсолютная ошибка в процентах ($MAPE$) из (1.8), (1.9), (1.10), где под $\widehat{y}_j$ понимается оценка исследуемой характеристики (математического ожидания или среднеквадратического отклонения времени отклика), полученная либо с помощью нейросети либо с помощью аналитических формул, а под y_j — реальное значение одной из оцениваемых характеристик, полученное в результате имитационного моделирования системы с разделением и параллельным обслуживанием заявок, а N — количество наборов данных в выборке, предназначенной для оценки погрешности аппроксимации.

Значения ошибок (1.8)–(1.10), полученные на тестовой выборке, не участвующей в обучении нейросети, представлены в таблице 18, что уже свидетельствует об удовлетворительном качестве прогноза, выдаваемом нейросетью на абсолютно новых, незнакомых ей входных данных. Однако, чтобы окончательно

Таблица 18: Погрешности приближений оценок характеристик производительности fork-join системы с подсистемами $M|G|1$, полученных с помощью ИНС на тестовом наборе данных

Оцениваемая характеристика	Типы ошибок		
	MSE	MAE	$MAPE$, %
$E[R_K]$	0.000668	0.013629	0.490643
$\sqrt{Var[R_K]}$	0.000308	0.012031	0.884236

удостовериться в своего рода стабильности качественного прогноза, вычислим аналогичные ошибки, но уже для промежуточных входных данных, которые представлены в таблице 19.

Таблица 19: Промежуточные входные данные для оценки характеристик производительности fork-join системы с подсистемами $M|G|1$, полученных с помощью ИНС и для анализа аналитических оценок среднего времени отклика и его среднеквадратического отклонения в разделе 3.4

№ п/п	1	2	...	19	20	21	...	915
α	4.50	4.50	...	4.50	4.50	4.50	...	9.50
ρ	0.15	0.15	...	0.15	0.25	0.25	...	0.85
K	2	3	...	20	2	3	...	20

Чтобы провести сравнительный анализ, на этом же наборе данных вычислим ошибки аппроксимации для формул (3.5) и (3.6) с учетом выражений (3.12)–(3.14), а также самой формулы (3.12). Аналогичным образом сравним погрешность аппроксимации аналитической формулы (3.13) с прогнозом нейросети. Результаты вычислений представлены в таблице 20. Очевидно, что аналитические формулы допускают неприемлемую относительную погрешность.

Для того, чтобы более детально ознакомиться со структурой полученных

Таблица 20: Погрешности приближений оценок характеристик производительности fork-join системы, полученных с помощью ИНС и аналитических формул на наборе данных из таблицы 19.

Оцениваемая характеристика	Типы ошибок		
	MSE	MAE	$MAPE$, %
$E[R_K]$, ИНС	0.001939	0.021409	0.707818
$E[R_K]$, формула (3.5)	5.892661	1.878592	67.785755
$E[R_K]$, формула (3.6)	6.722487	1.817356	59.983739
$E[R_K]$, формула (3.12)	0.621682	0.642618	24.425016
$\sqrt{Var[R_K]}$, ИНС	0.002679	0.038941	3.355246
$\sqrt{Var[R_K]}$, формула (3.13)	0.107970	0.254335	18.701897

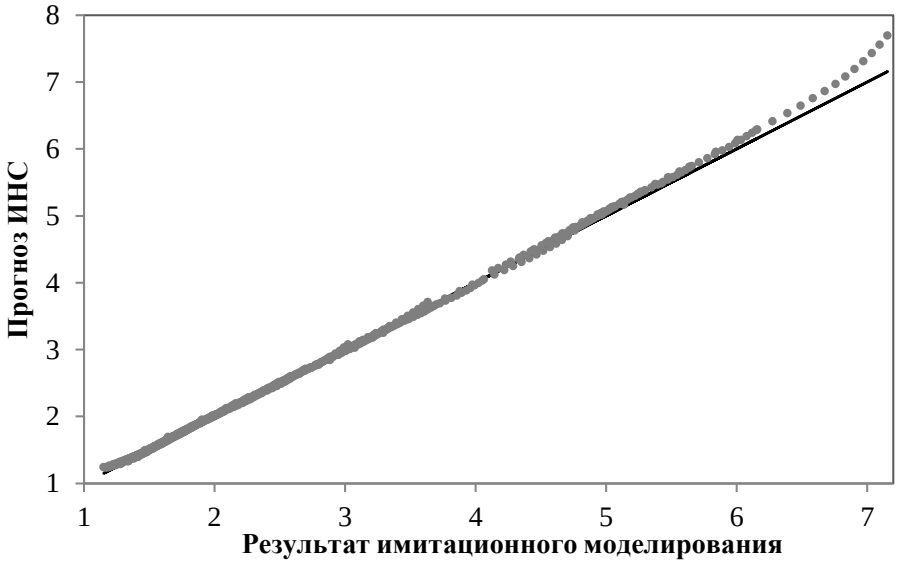


Рис. 44: Среднее время отклика fork-join СМО с подсистемами $M|G|1$, полученное в результате имитационного моделирования и в результате применения ИНС.

оценок среднего значения и среднеквадратического отклонения времени отклика, построим графики, на которых наглядно отражено отклонение этих оценок от истинных значений показателей. На оси абсцисс откладываются значения среднего времени отклика, полученные на наборе данных из таблицы 19 с по-

мощью имитационной модели, а по оси ординат — значения, полученные либо с помощью нейросети, либо с помощью формул (3.5), (3.6), (3.12), как на рисунках 44 и 45, 46, 47, соответственно, причем для полноты картины в случае аналитических формул область построения графиков расширяется за счет входных данных из табл. 17. Аналогичным образом строятся графики для среднеквадратического отклонения (рис. 48, 49).

Таблица 21: Входные данные, для которых относительная погрешность приближения аналитических формул меньше 10%

Формула (3.12)			Формула (3.13)		
K	α	ρ	K	α	ρ
2	4.5	–	2	4.5	0.65-0.85
	5.5	0.15-0.45		5.5	0.35-0.85
	6.5-9.5	0.15-0.85		6.5-9.5	0.15-0.85
3	6.5	–	3	6.5	0.55-0.85
	7.5	–		7.5	0.35-0.85
	8.5	0.15-0.55		8.5	0.25-0.85
	9.5	0.15-0.85		9.5	0.15-0.85
4	–	–	4	7.5	0.65-0.85
	–	–		8.5	0.45-0.85
	–	–		9.5	0.25-0.85
5	–	–	5	8.5	0.55-0.85
	–	–		9.5	0.35-0.85
6	–	–	6	8.5	0.75-0.85
	–	–		9.5	0.45-0.85
7	–	–	7	8.5	0.85
	–	–		9.5	0.65-0.85
8	–	–	8	9.5	0.75-0.85

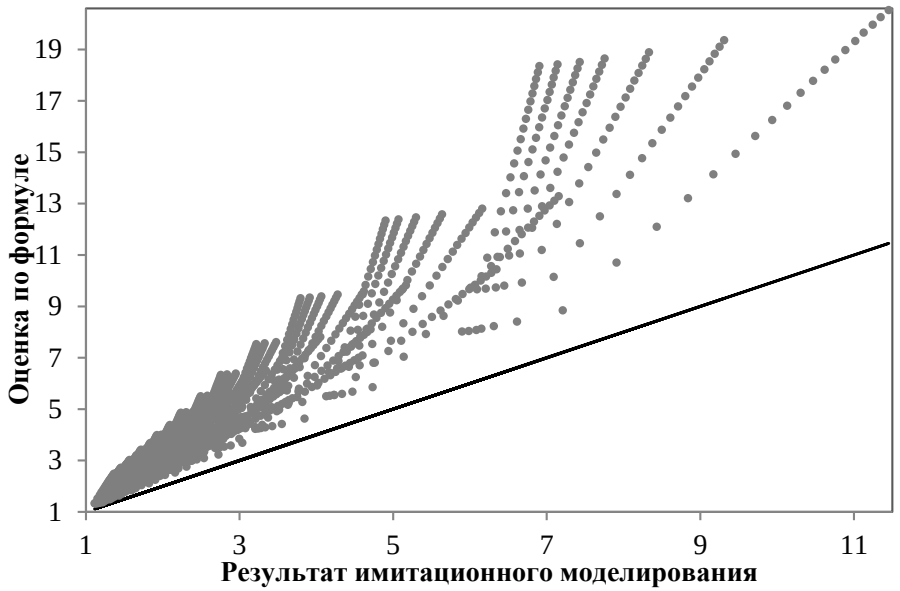


Рис. 45: Среднее время отклика fork-join СМО с подсистемами $M|G|1$, полученное в результате имитационного моделирования и в результате применения формулы (3.5).

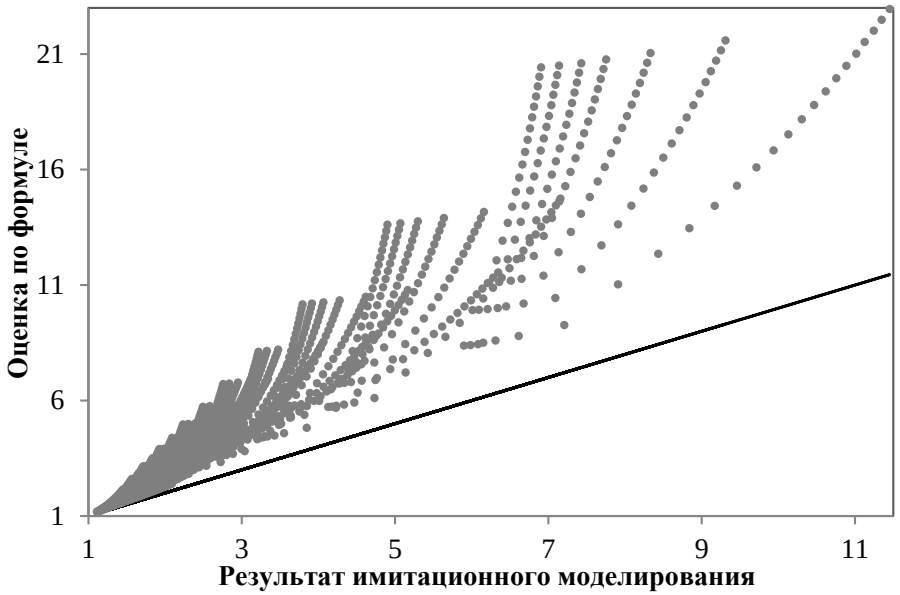


Рис. 46: Среднее время отклика fork-join СМО с подсистемами $M|G|1$, полученное в результате имитационного моделирования и в результате применения формулы (3.6).

Как видно из графиков (рис. 45, 46), для подавляющего большинства значений входных параметров аналитические формулы (3.5) и (3.6) выдают излишне завышенный результат, причем в случае выражения (3.5) все значения

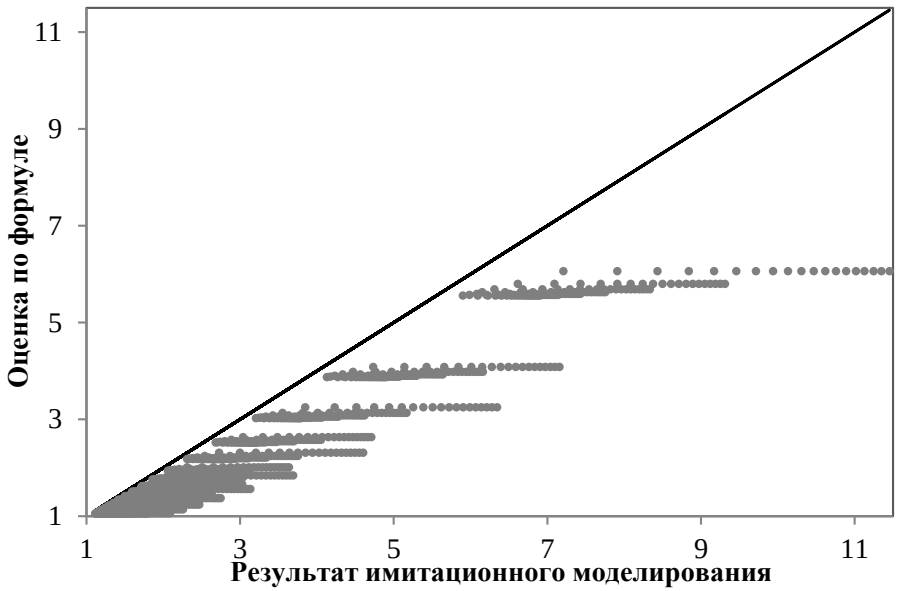


Рис. 47: Среднее время отклика fork-join СМО с подсистемами $M|G|1$, полученное в результате имитационного моделирования и в результате применения формулы (3.12).

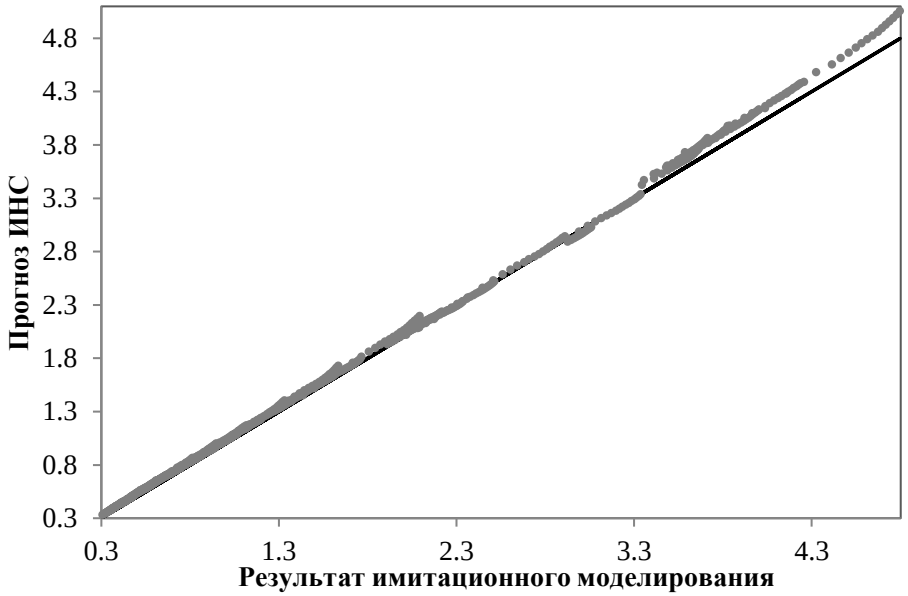


Рис. 48: Среднеквадратическое отклонения времени отклика fork-join СМО с подсистемами $M|G|1$, полученное в результате имитационного моделирования и в результате применения ИНС.

относительных ошибок приближения превышают 14%, а в случае с (3.6) — количество точечных оценок, относительная погрешность которых не превышала бы порога в 10%, равно 6. Заметим при этом, что оценка с помощью формулы

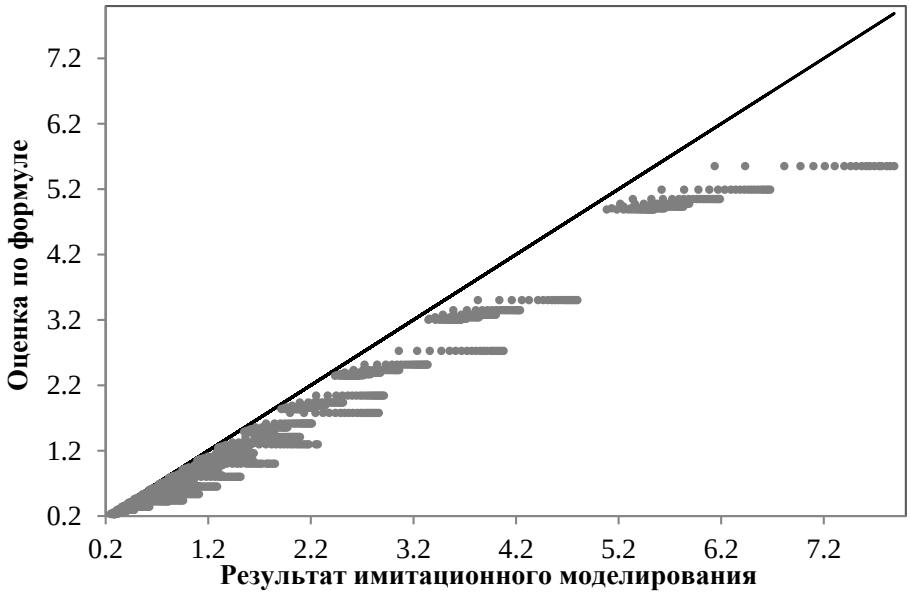


Рис. 49: Среднеквадратическое отклонения времени отклика fork-join СМО с подсистемами $M|G|1$, полученное в результате имитационного моделирования и в результате применения формулы (3.13).

(3.12) для среднего времени отклика в СМО $M|G|1$ дает более реалистичную оценку, в том смысле, что относительная ошибка приближения менее 10% справедлива уже для 49 значений из 912 (табл. 21). Лучшая ситуация в случае со среднеквадратическим отклонением (рис. 49), только здесь уже оценка данного показателя с помощью формулы (3.13) фактически получается снизу и порог в 10% действует примерно для 11.4% значений от всего количества оценок и распространяется на область больших значений параметра α и больших значений загрузки ρ (табл. 21).

Оценки, полученные с помощью обученных нейросетей (рис. 44, 48), демонстрируют приемлемое качество приближения. Так, в случае среднего времени отклика относительная погрешность аппроксимации не превышает 5% для 99.23% от общего количества полученных оценок, а для оставшихся 7 оценок, значения которых выходят за указанную границу, максимальная погрешность составляет всего 7.57%, что вполне допустимо. В случае среднеквадратического отклонения результат обучения получился чуть хуже, поскольку количество

оценок, выходящих за рамки 5% составляет уже 19.7% от их общего количества, тем не менее, порядок максимальной относительной погрешности тот же и составляет всего 7.61%. В обоих случаях видно, что наибольшая относительная погрешность наблюдается в области больших значений соответствующей характеристики.

Заметим, что цели наилучшим образом обучить нейросеть не стояло, поэтому, затратив большее количество времени, вполне вероятно, можно подобрать ИНС с другой архитектурой, функциями активации и т. д., которая могла бы давать более точный прогноз. Однако и без того ясно, что в силу отсутствия аналитических оценок, которые давали бы приемлемую погрешность результатов, предложенный метод представляется более перспективным, в том числе и применительно к анализу более сложной модели с подсистемами типа $G|G|1$.

3.4 Оценка основных характеристик системы с помощью комплексного метода

В данном разделе приводятся результаты дальнейшего развития нового метода, основанного на применении методов машинного обучения, что позволяет значительно повысить точность аналитического приближения для среднего времени отклика и его среднеквадратического отклонения, а также оценить другие характеристики системы (коэффициенты корреляции) в аналитическом виде. Вместо искусственной нейросети используется множественная регрессия (линейная и нелинейная). Имитационное моделирование fork-join СМО происходит в программной среде Python. При этом высокая точность каждой полученной числовой оценки достигается благодаря моделированию порядка 5–10 миллионов соответствующих величин для каждой комбинации значений параметров модели.

Комплексный метод уже использовалась в Главе 2 для нахождения математического ожидания и среднеквадратического отклонения времени отклика

fork-join СМО с показательным распределением времени обслуживания. Однако в данной Главе с помощью комплексного метода оцениваются и другие важные характеристики fork-join СМО. Поэтому более детально остановимся на описании общего алгоритма нахождения оценок различных показателей fork-join системы.

Метод включает в себя несколько этапов.

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

Эта часть предварительного анализа для специалистов из соответствующей области не должна представлять особых сложностей, поскольку интуитивно понятна. Очевидно, что в данном случае и не только, основные параметры, от которых зависит поведение fork-join модели и ее характеристик — это нагрузка на систему/подсистемы (интенсивность входящего и обслуживающего потоков), параметры соответствующих распределений, а также число подсистем. При этом для оценивания характеристик сложных систем удобно опираться на известные характеристики простых систем (подсистем), как на базовые (например, центрируя и нормируя, как в (3.15)). Также необходимо отслеживать, чтобы при переходе от сложной системы к простой формулы из приближенных становились точными (в нашем случае, при $K = 1$).

2. Далее проводится имитационное моделирование с целью получения точных (максимально близких к точным) оценок исследуемой характеристики в зависимости от различных значений рассматриваемых параметров на некотором ограниченном (но достаточно обширном для приложений) интервале.

Имитационное моделирование также имеет свои особенности, которым довольно часто не уделяется должного внимания, но от этого зависит точность получаемых результатов. Несмотря на то, что традиционно длительность прогона

имитационной модели для получения одного значения определяется экспериментально, в Главе 2 даются некоторые рекомендации относительно организации вычислительного эксперимента для классической fork-join СМО.

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

На этом этапе строятся графики искомой характеристики, полученные посредством симуляции, в зависимости от различных параметров модели. На основании визуального анализа корректируется предполагаемый вид функциональной зависимости и строится модель множественной (линейной или нелинейной) регрессии, в которой остаются неизвестными некоторые постоянные коэффициенты.

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

Этот этап позволяет значительно улучшить точность полученных аналитических оценок. На выходе мы получаем аналитическое выражение, аппроксимирующее интересующую характеристику с довольно высокой степенью точности. Метод Нелдера–Мида используется потому, что он не требует гладкости функции и прост в обращении. Возможны, конечно, и другие варианты.

Описанные этапы сочетают в себе комбинацию классических методов из нескольких математических областей. Например, имитационное моделирование, это практически единственный способ определения неизвестных параметров производительности модели массового обслуживания в случае, если вывод характеристик аналитическими методами затруднителен. Однако он требует значительных временных и вычислительных затрат и некоторой квалификации. Поэтому решение, полученное в форме аналитического выражения, имеет явное преимущество.

Что касается графического анализа и определения конкретного вида функциональной зависимости, то это, пожалуй, наиболее сложный и трудоемкий этап в предложенном подходе. Естественно, что при этом предполагается индивидуальность подбора типа зависимости между параметрами для конкретной архитектуры модели, и универсальных решений в этом смысле здесь нет. При этом трудоемкость работы и некоторые потери в общности компенсируются хорошим качеством аппроксимации, ведь, как правило, на практике рассматриваются конкретные модели физических процессов, а не их обобщения. В то же время, этот способ понятен для широкого круга исследователей вне зависимости от специализации, кроме того, визуальный анализ данных традиционно является частью data mining. И именно благодаря наличию данного этапа становится возможным находить не только начальные моменты показателей производительности системы, но и определять другие ее характеристики. Например, как в данном случае, речь может идти об определении зависимостей между исследуемыми случайными величинами (коэффициенты корреляции). До настоящего момента проведение подобного анализа представлялось крайне затруднительным, о чём свидетельствует отсутствие исследований в данном направлении.

Математическое ожидание времени отклика. Для среднего времени отклика естественно за основу взять формулу (3.6), т. е. по-прежнему считаем, что среднее время отклика fork-join системы зависит от математического ожидания и от среднеквадратического отклонения времени пребывания подзаявки в соответствующей подсистеме. Положим

$$\mu_K^* = \frac{E[R_K] - \mu_{Pa}}{\sigma_{Pa}}, \quad \sigma_K^* = \frac{\sqrt{Var[R_K]}}{\sigma_{Pa}} - 1. \tag{3.15}$$

Для распределения Парето известно, что максимум N независимых одинаково распределенных случайных величин растет как $N^{1/\alpha}$ при $N \to \infty$, в отличие от случая показательного распределения, когда максимум растет как $\log N$.

Исходя из этих соображений, рассмотрим зависимость μ_K^* и σ_K^* от $K^{1/\alpha}$.

График 50 зависимости μ_K^* от $K^{1/\alpha}$ и график 51 зависимости σ_K^* от $K^{1/\alpha}$ выглядят как пучок прямых, проходящих через точку с координатами $(1.0, 0.0)$ Это приводит к рассмотрению переменных

$$\mu_K^{**} = \frac{\mu_K^*}{K^{\frac{1}{\alpha}} - 1}, \quad \sigma_K^{**} = \frac{\sigma_K^*}{K^{\frac{1}{\alpha}} - 1},$$

слабо зависящих от K.

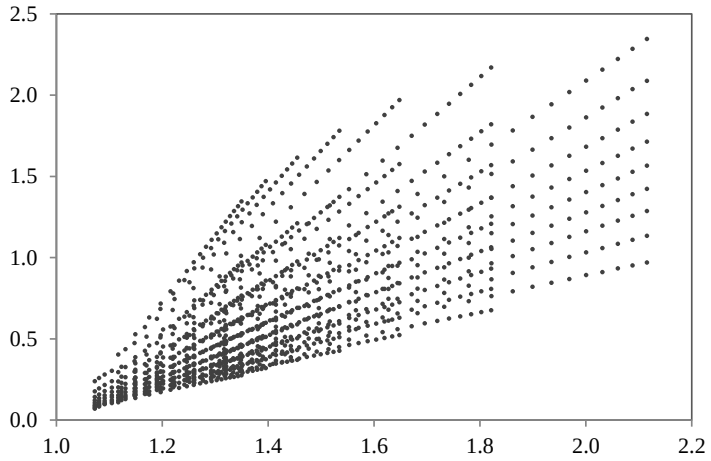


Рис. 50: μ_K^* из (3.15) в зависимости от $K^{1/\alpha}$.

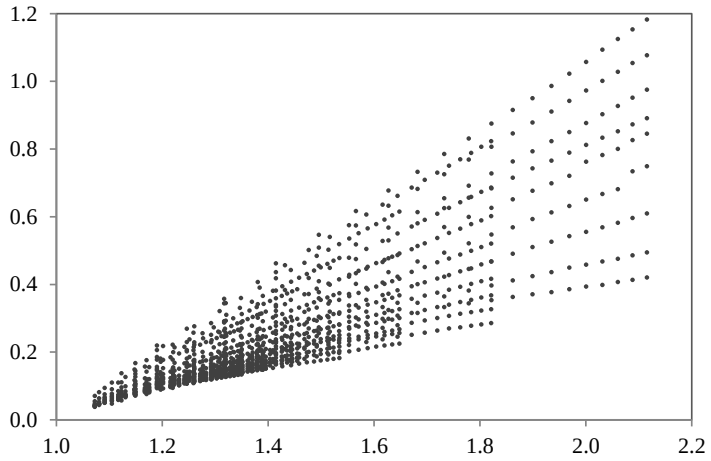


Рис. 51: σ_K^* из (3.15) в зависимости от $K^{1/\alpha}$.

Таким образом, приходим к следующей формуле

$$E[R_K] = \mu_{Pa} + \sigma_{Pa}(K^{\frac{1}{\alpha}} - 1)\tilde{\mu}_K, \tag{3.16}$$

где $\tilde{\mu}_K = \tilde{\mu}_K(\alpha, \rho)$ является функцией двух переменных α и ρ. Конкретное выражение, определяющее искомую функцию, будем находить с помощью сочетания двух методов — метода наименьших квадратов и метода оптимизации Нелдера–Мида.

Для проведения дальнейших манипуляций необходимо задать числовые значения для входных параметров. Остановимся на тех же входных данных, что рассматривались в предыдущем разделе в таблицах 17 и 19. В таблице 17 параметр α принимает только целочисленные значения на отрезке $[4.0, 10.0]$, коэффициент загрузки ρ меняется от 0.1 до 0.9 включительно с шагом 0.1, а число подсистем K варьируется от 2 до 20 включительно. Ранее данные из таблицы 17 использовались для обучения и тестирования нейронной сети; точность прогноза была позже проверена на данных таблицы 19. В данном случае будут использованы комбинированные данные из обеих таблиц. Для всех этих данных с помощью симуляции предварительно были получены значения $E[R_K]$. Задача состоит в получении формулы для аппроксимации среднего времени отклика, работающей для любых промежуточных значений указанных входных параметров с высокой степенью точности.

Истинные значения функции $\tilde{\mu}_K(\alpha, \rho)$, соответствующие определенным выше числовым значениям α и ρ, получим, выражая, $\tilde{\mu}_K(\alpha, \rho)$ из формулы (3.16), подставляя при этом в $E[R_K]$ результаты имитационного моделирования. Далее, построив график зависимости полученных истинных значений $\tilde{\mu}_K(\alpha, \rho)$ от α и ρ, можем сделать предположение о виде искомой функции

$$\tilde{\mu}_K \approx C_{\mu 1} + C_{\mu 2}\alpha + C_{\mu 3}\rho + C_{\mu 4}\alpha\rho + C_{\mu 5}\rho^2 + C_{\mu 6}\alpha^2. \tag{3.17}$$

Для поиска первого приближения коэффициентов $C_{\mu i}$, $i = 1, ..., 6$, воспользуемся методом наименьших квадратов. Далее, подставив функцию (3.17) в (3.16), воспользуемся оптимизационным методом Нелдера–Мида [138,161]. Этот метод применяется для минимизации максимальной относительной ошибки на всём наборе данных. В качестве минимизируемой функции выступает следую-

щее выражение

$$\left|\frac{\mu_{Pa}+\sigma_{Pa}(K^{\frac{1}{\alpha}}-1)\tilde{\mu}_K(\alpha,\rho)-E[R_K]}{E[R_K]}\right|\to\min. \tag{3.18}$$

Скажем несколько слов о самом методе Нелдера–Мида, который также известен как метод деформируемого многогранника. Процесс поиска оптимальной точки с его помощью заключается в вычислении значений исследуемой функции в вершинах некоторого симплекса и последующей его деформации, точнее говоря, изменения размера и формы многомерного тетраэдра с помощью различных преобразований, в направлении точки экстремума. Данный метод позволяет получить результат для функции нескольких переменных без необходимости вычисления ее градиента, что делает его применимым к негладким функциям.

В силу вероятного наличия большого числа локальных точек экстремума у анализируемой функции конечный результат может сильно зависеть от выбранного начального приближения, в нашем случае — конкретных значений $C_{\mu i}$. Однако первое приближение было уже получено с помощью метода наименьших квадратов, поэтому после применения метода Нелдера–Мида имеем (рис. 52)

$$\begin{gathered}\tilde{\mu}_K\approx 1.25918+0.36996\alpha-1.97400\rho-\\ -0.28495\alpha\rho+1.40841\rho^2-0.0112\alpha^2.\end{gathered} \tag{3.19}$$

Теперь для выведенной оценки среднего времени отклика (3.16) и (3.19) рассчитаем величину средней абсолютной ошибки в процентах ($MAPE$) на наборе данных из таблицы 17, а также промежуточных данных из таблицы 19. $MAPE$ определяется выражением (1.10), где $\widehat{y}_j$ — это оценка исследуемой характеристики (математического ожидания или среднеквадратического отклонения времени отклика), полученная с помощью с помощью аналитических формул, а y_j — реальное значение одной из оцениваемых характеристик, полученное в результате имитационного моделирования fork-join СМО, N — количество наборов данных в выборке, предназначенной для оценки погрешности аппроксимации.

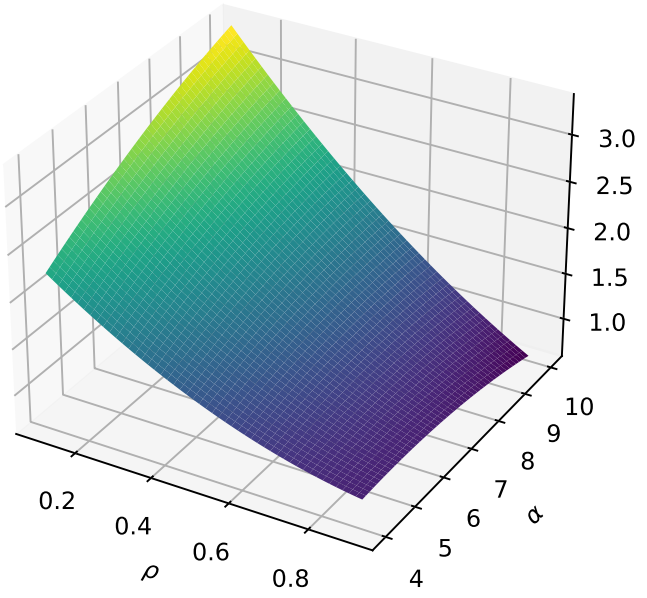


Рис. 52: Модифицированное среднее время отклика $\tilde{\mu}_K$ из (3.19) в зависимости от α и ρ.

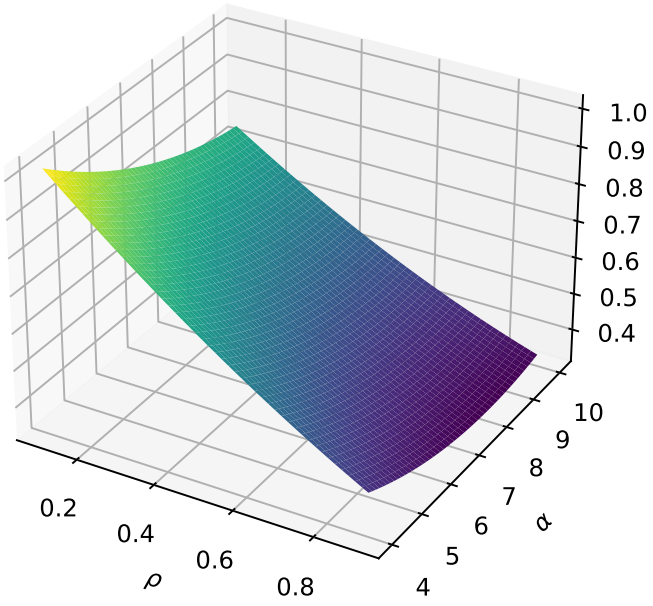


Рис. 53: Модифицированное среднеквадратическое отклонение времени отклика $\tilde{\sigma}_K(\alpha, \rho)$ из (3.23) в зависимости от α и ρ.

В результате имеем, что для выведенной оценки среднего времени отклика fork-join СМО ошибка вида $MAPE$ не превышает 1.6% (см. табл. 22). Также рассчитаны максимальная ($MaxAPE$) и минимальная ($MinAPE$) абсолютные процентные ошибки.

Таблица 22: Погрешности приближений оценок среднего времени отклика fork-join системы, полученные с помощью формул (3.16) и (3.19) на наборе данных из таблиц 17 и 19

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
$E[R_K]$	3.93942	0.00099	1.59993

Более детально ознакомиться со структурой приближений можно на рисунке 54. Здесь на оси абсцисс отложены значения среднего времени отклика, полученные с помощью имитационного моделирования на входных наборах данных из таблиц 17 и 19, а на оси ординат — оценки среднего времени отклика, рассчитанные с помощью аналитических формул на тех же наборах входных данных.

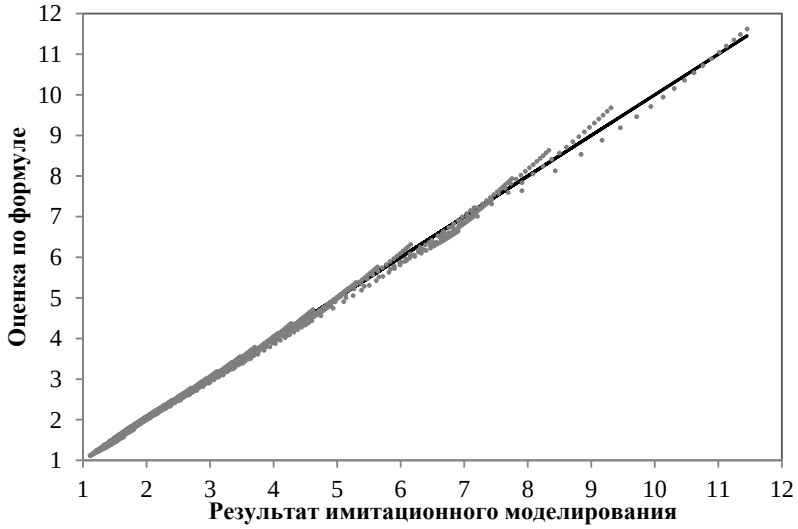


Рис. 54: Оценка качества аппроксимации среднего времени отклика $E[R_K]$ формулами (3.16) и (3.19).

Среднеквадратическое отклонение времени отклика. Аналогичным образом найдем оценку для среднеквадратического отклонения времени отклика:

$$\sqrt{Var[R_K]} = \sigma_{Pa}(1 + (K^{\frac{1}{\alpha}} - 1)\tilde{\sigma}_K), \tag{3.20}$$

где $\tilde{\sigma}_K = \tilde{\sigma}_K(\alpha, \rho)$. После визуального анализа графика для функции $\tilde{\sigma}_K(\alpha, \rho)$, который мы можем построить, выразив $\tilde{\sigma}_K$ из (3.20) и подставив предварительно смоделированные значения $\sqrt{Var[R_K]}$, предположим следующую зависимость

$$\tilde{\sigma}_K(\alpha, \rho) \approx C_{\sigma 1} + C_{\sigma 2}\alpha + C_{\sigma 3}\rho + C_{\sigma 4}\alpha\rho + C_{\sigma 5}\rho^2 + C_{\sigma 6}\alpha^2. \tag{3.21}$$

Далее после нахождения первого приближения для коэффициентов $C_{\sigma i}$, $i = 1, ..., 6$, методом Нелдера–Мида минимизируем функцию

$$\left| \frac{\sigma_{Pa}[1 + (K^{\frac{1}{\alpha}} - 1)\tilde{\sigma}_K(\alpha, \rho)] - \sqrt{\mathrm{Var}[R_k]}}{\sqrt{\mathrm{Var}[R_k]}} \right| \to \min \tag{3.22}$$

В результате получаем оптимальные значения коэффициентов $C_{\sigma i}$ и приходим к формуле (рис. 53).

$$\begin{gathered} \tilde{\sigma}_K(\alpha, \rho) \approx 1.54316 - 0.13036\alpha - 1.10517\rho + \\ +0.03936\alpha\rho + 0.20804\rho^2 + 0.00575\alpha^2. \end{gathered} \tag{3.23}$$

В результате, для среднеквадратического отклонения времени отклика значение $MAPE$ на наборах исходных и промежуточных данных из таблиц 17 и 19 не превышает 1.5% (табл. 23).

Таблица 23: Погрешности приближений среднеквадратического отклонения fork-join системы, полученные с помощью формул (3.20) и (3.23) на наборе данных из таблиц 17 и 19

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
$\sqrt{\mathrm{Var}[R_K]}$	4.06028	0.00111	1.46589

На рисунке 55 отражено, насколько оценка среднеквадратического отклонения времени отклика на основе аналитических формул (3.20) и (3.23) отличается от их истинных значений (результата симуляции).

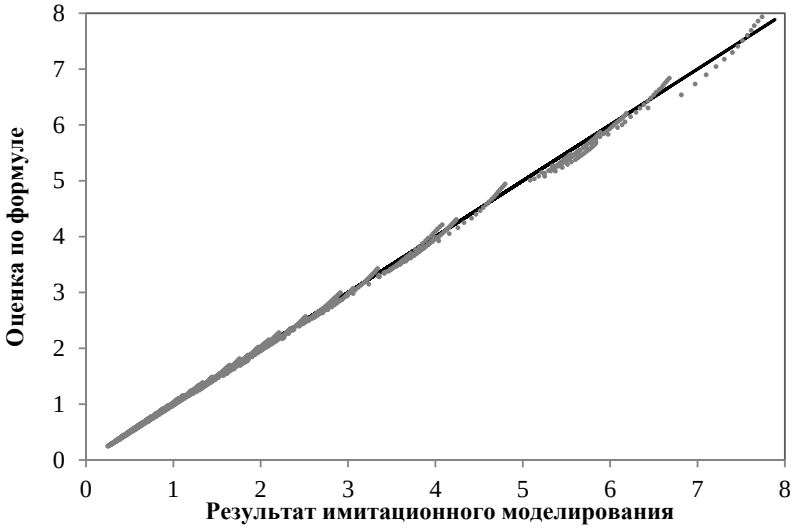


Рис. 55: Оценка качества аппроксимации среднеквадратического отклонения времени отклика $\sqrt{Var[R_K]}$ формулами (3.20) и (3.23).

Было бы интересно сравнить поведение модифицированных характеристик (см. рис. 52 и 53). Видно, что обе они возрастают с ростом ρ, но первая характеристика увеличивается с ростом α, а вторая убывает с ростом α.

Графики полученных оценок этих характеристик показаны на рис. 56.

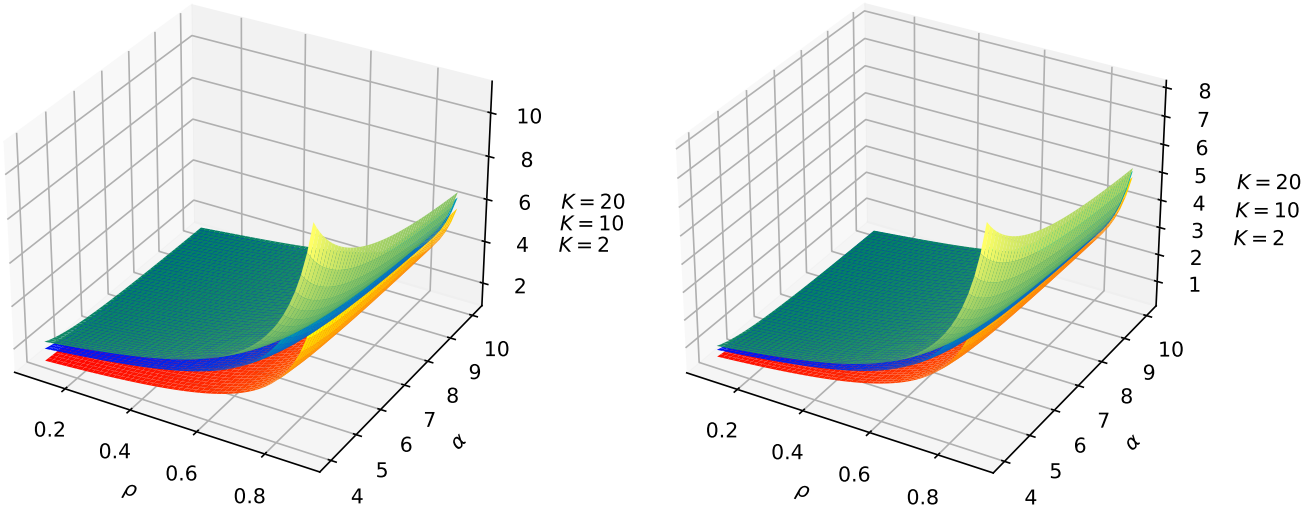


Рис. 56: Зависимость от α и ρ а) среднего времени отклика $E[R_K]$ из (3.16), б) среднеквадратического отклонения времени отклика $\sqrt{Var[R_K]}$ из (3.20).

Для полноты картины построим также диаграмму зависимости среднего времени отклика и стандартного отклонения (см. рис. 57), из которой ясно,

что каждая из этих характеристик не может быть адекватно выражена через другую, поэтому требуется их отдельная оценка.

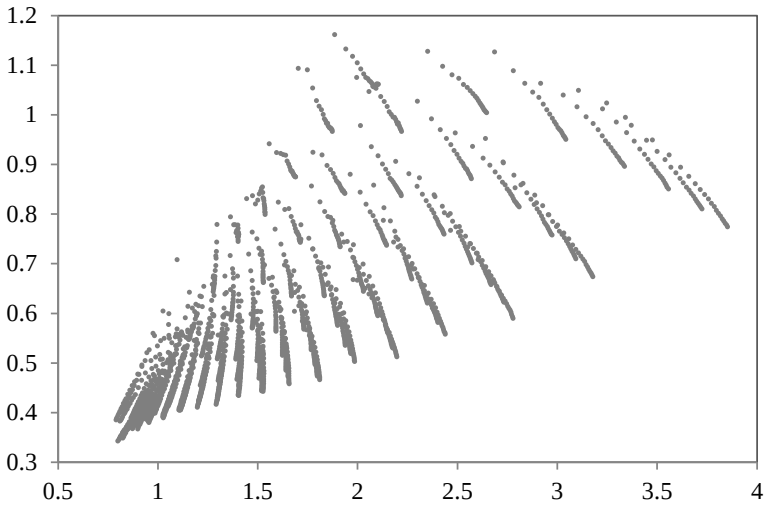


Рис. 57: Среднеквадратическое отклонение времени отклика $\sqrt{Var[R_K]}$ в зависимости от среднего времени отклика системы $E[R_K]$.

3.5 Оценка коэффициентов корреляции

Для оценки коэффициентов корреляции используем тот же метод, что и в предыдущем разделе, с параметрами из таблиц 17 и 19. Поэтому немного сократим описание использования промежуточных элементов анализа и более подробно остановимся только на этапе выбора типа функциональной зависимости.

Коэффициент корреляции Пирсона. Для начала предположим зависимость коэффициента корреляции Пирсона от двух параметров fork-join СМО — ρ и α. Далее для оценки коэффициента корреляции Пирсона удобно ввести следующую величину[3]

$$r^*_{\text{Pearson}} = r^*_{\text{Pearson}}(\alpha, \rho) = \frac{r_{\text{Pearson}}}{1 - r_{\text{Pearson}}}. \tag{3.24}$$

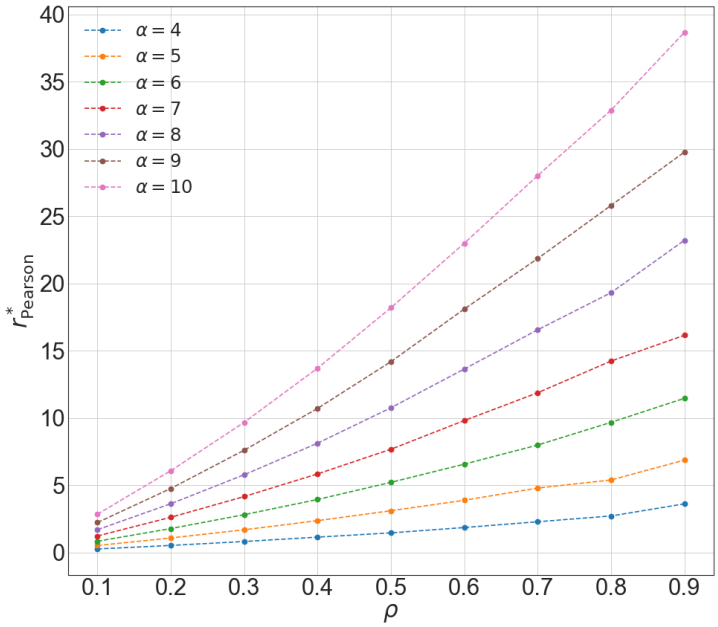


Рис. 58: Зависимость r^*_{Pearson} из (3.24) от ρ.

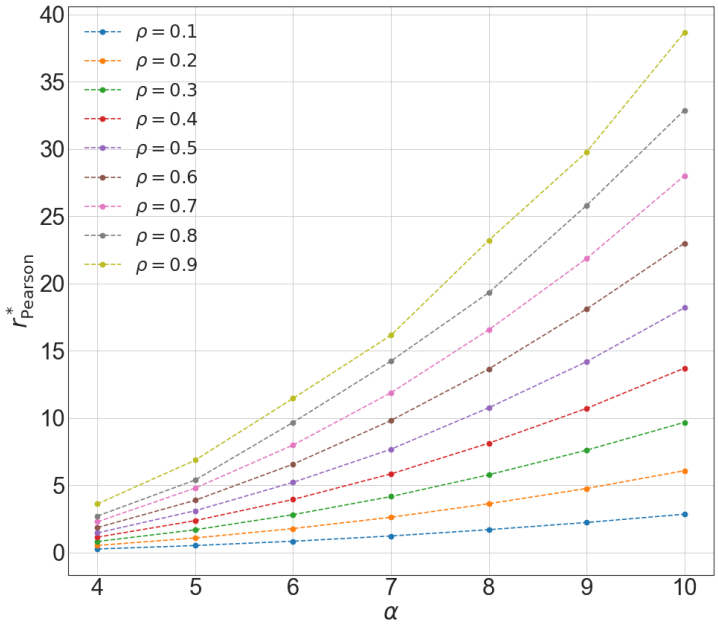


Рис. 59: Зависимость r^*_{Pearson} из (3.24) от α.

Теперь для определения конкретного типа функциональной зависимости построим графики для r^*_{Pearson} в зависимости от α и ρ (рис. 58 и 59). На рисунке 58

[3] Сначала была сделана попытка приближения без такого преобразования, но получалось хуже.

мы можем наблюдать пучок парабол, отличающихся масштабным коэффициентом. При этом ясно, что параболы проходят через точку $(0.0, 0.0)$, поскольку $r^*_{\text{Pearson}} = 0$ при $\rho = 0$. Таким образом, можем предположить, что

$$r^*_{\text{Pearson}} \approx C^* \cdot \rho(C_1 + C_2\rho),$$

где $C^* = C^*(\alpha)$ — это тот самый масштабный коэффициент, который меняет свое значение с изменением значения параметра α. Поэтому далее аналогичным образом, исходя из вида графика, представленного на рисунке 59, можем сделать вывод о квадратичном типе зависимости r^*_{Pearson} от α вида[4]

$$r^*_{\text{Pearson}} \approx C^{**} \cdot (C_3\alpha^2 + \alpha - C_4),$$

где $C^{**} = C^{**}(\rho) = \rho(C_1 + C_2\rho)$ — также масштабный коэффициент, который меняет свое значение в зависимости от значения параметра ρ. Следовательно, получаем, что

$$r^*_{\text{Pearson}}(\alpha, \rho) \approx C^*(\alpha) \cdot C^{**}(\rho) = \rho(C_1 + C_2\rho)(C_3\alpha^2 + \alpha - C_4).$$

Далее возвращаемся к исходной величине $r_{\text{Pearson}}(\alpha, \rho)$ и методом Нелдера–Мида определяем оптимальные значения коэффициентов.

В результате получаем, что коэффициент корреляции Пирсона между временами пребывания двух подзаявок одной заявки аппроксимируется следующим выражением (см. рис. 60)

$$r_{\text{Pearson}}(\alpha, \rho) \approx 1 - \frac{1}{1 + \rho(C_1 + C_2\rho)(C_3\alpha^2 + \alpha - C_4)}, \tag{3.25}$$

где

$$C_1 \approx 0.25037, \quad C_2 \approx 0.12668,$$

$$C_3 \approx 1.00000, \quad C_4 \approx 10.06767,$$

[4]Для однозначного оценивания набора коэффициентов, проводимого далее, один из коэффициентов должен быть зафиксирован. Здесь мы берем единичный коэффициент при α. При другом выборе получатся другие значения коэффициентов, но та же самая оценка после раскрытия скобок.

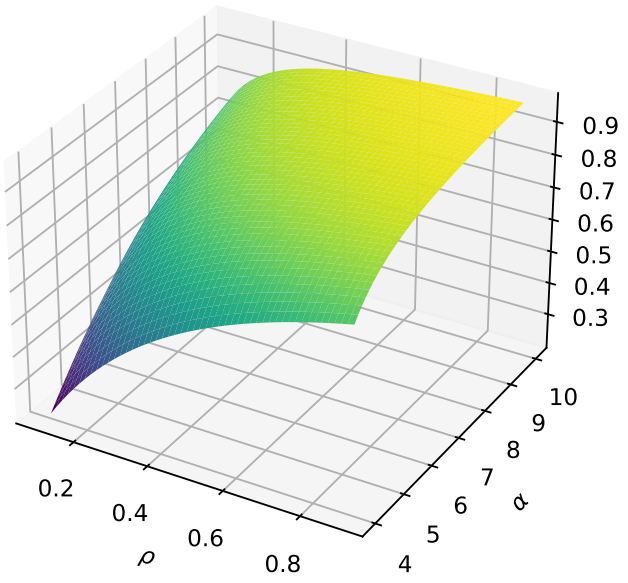


Рис. 60: Коэффициент корреляции Пирсона из (3.25) в зависимости от α и ρ.

$$MAPE \approx 0.85752\%, \quad MaxAPE \approx 2.37988\%, \quad MinAPE \approx 0.05274\%,$$

Качество полученного аппроксимационного выражения наглядно демонстрируется на рисунке 61.

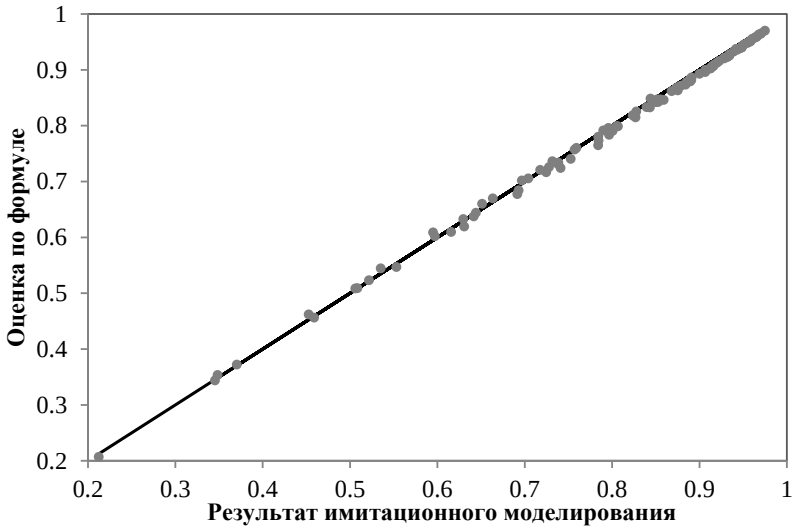


Рис. 61: Коэффициент корреляции Пирсона (оценка качества аппроксимации формулы (3.25)).

Коэффициент корреляции Кендалла. Теперь с помощью графического анализа определим вид функциональной зависимости для коэффициента корреляции Кендалла. На рисунке 62, где изображена зависимость r_{Kendall} от ρ можем

наблюдать пучок парабол с ветвями, направленными вниз, отличающихся масштабным коэффициентом. А на рисунке 63, где изображена зависимость r_{Kendall} от $1/\alpha$, наблюдаем пучок прямых с отрицательным наклоном. Следовательно,

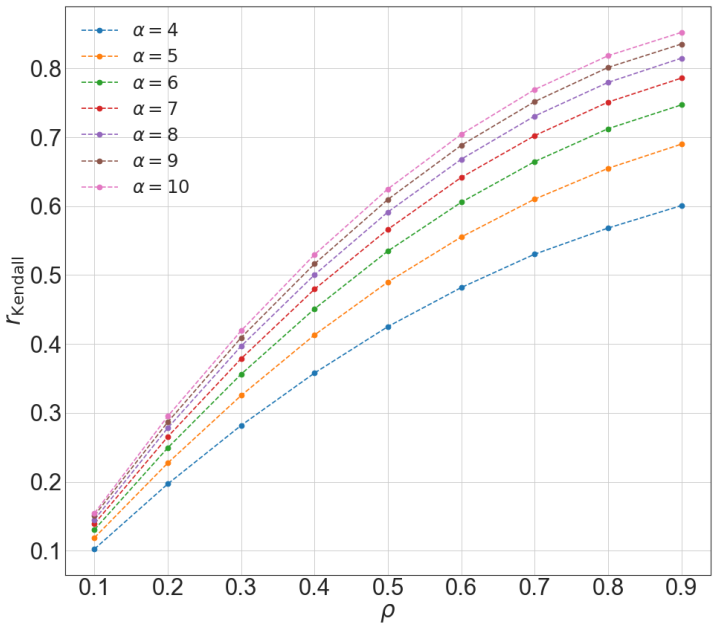


Рис. 62: Зависимость коэффициента корреляции Кендалла r_{Kendall} от ρ.

можем предположить аналитическое выражение

$$r_{\text{Kendall}}(\alpha, \rho) \approx \rho(C_1 - C_2\rho)(C_3 - \frac{C_4}{\alpha}). \tag{3.26}$$

Затем, аналогично повторяя описанные выше шаги, приходим к тому, что

$$C_1 \approx 3.93896, \quad C_2 \approx 1.73733,$$

$$C_3 \approx 0.49037, \quad C_4 \approx 0.85491,$$

и

$$MAPE \approx 0.533877\%, \quad MaxAPE \approx 1.568803\%, \quad MinAPE \approx 0.008721\%.$$

Соответствующий график представлен на рисунке 64, а точность аппроксимации иллюстрируется рисунком 65.

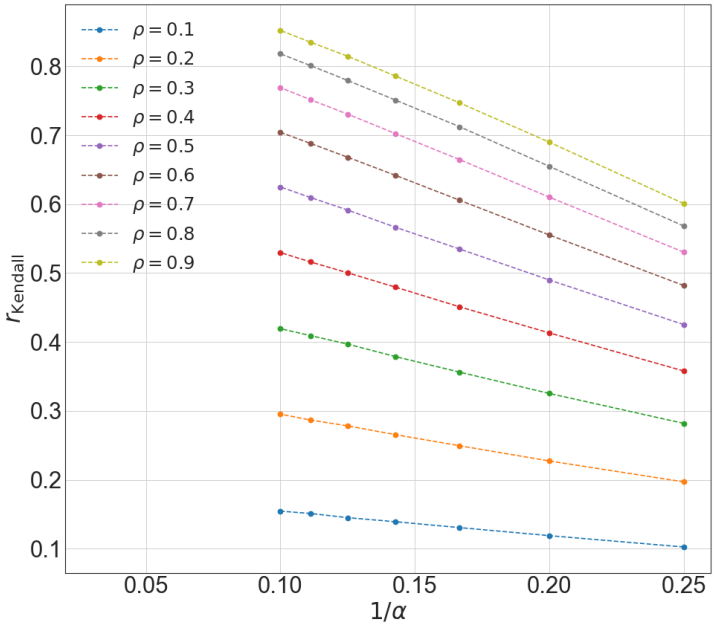


Рис. 63: Зависимость коэффициента корреляции Кендалла r_{Kendall} от $1/\alpha$.

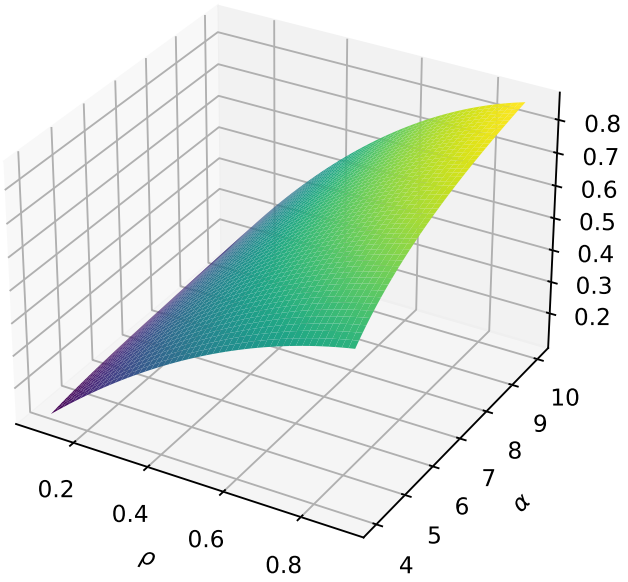


Рис. 64: Коэффициент корреляции Кендалла из (3.26) в зависимости от α и ρ.

Коэффициент корреляции Спирмена. Оказывается возможным найти относительно точное выражение для коэффициента корреляции Спирмена через коэффициент корреляции Кендалла.

Рисунок 66 наводит на мысль о кубической функции, связывающей коэффициент r_{Spearman} с r_{Kendall}. Для уточнения этого, построим график зависимости

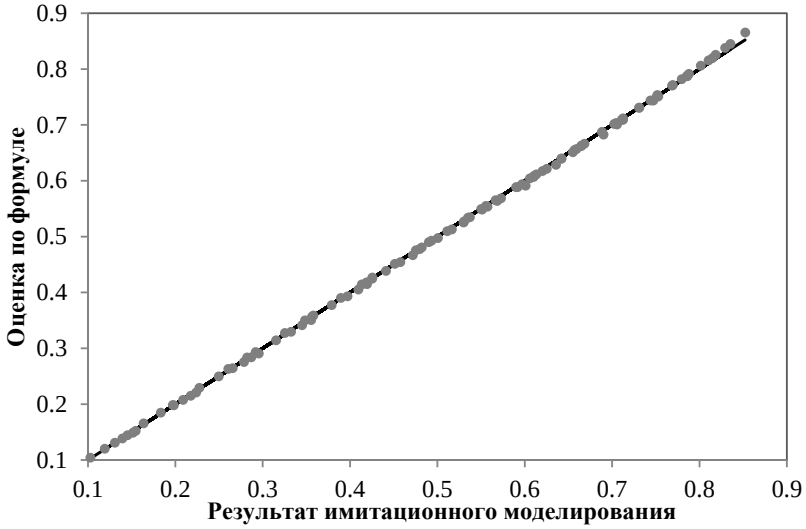


Рис. 65: Коэффициент корреляции Кендалла (оценка качества аппроксимации формулы (3.26)).

$r_{\text{Spearman}}/r_{\text{Kendall}}$ от r_{Kendall}. Анализируя рисунок 67, можем предположить квадратичную зависимость, причем ветви параболы направлены вниз, а максимум достигается в нуле. Таким образом, имеем

$$r_{\text{Spearman}}(r_{\text{Kendall}}) \approx r_{\text{Kendall}}(C_1 - C_2 r_{\text{Kendall}}^2); \tag{3.27}$$

здесь (см. рис. 68)

$$C_1 \approx 1.48030, \quad C_2 \approx 0.47393,$$

$$MAPE \approx 0.48771\%, \quad MaxAPE \approx 1.03572\%, \quad MinAPE \approx 0.02020\%.$$

Подставив $r_{\text{Kendall}}(\rho, \alpha)$ из (3.26) в последнее выражение (3.27), получим для результирующей оценки (см. рис 69),

$$MAPE \approx 0.510878\%, \quad MaxAPE \approx 1.375997\%, \qquad MinAPE \approx 0.005760\%.$$

Более того, коэффициенты корреляции Спирмена и Кендалла оказываются связанными с коэффициентом корреляции Пирсона, но лишь в достаточно слабой степени.

На рисунках 70 и 71 также показана зависимость модифицированных характеристик от коэффициентов корреляции. Зависимость, безусловно, прояв-

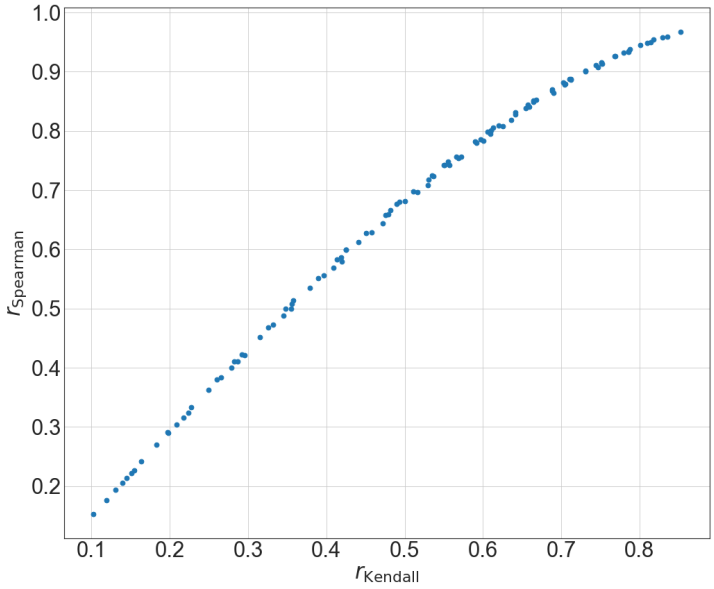


Рис. 66: График зависимости коэффициента корреляции Спирмена r_{Spearman} от коэффициента корреляции Кендалла r_{Kendall}.

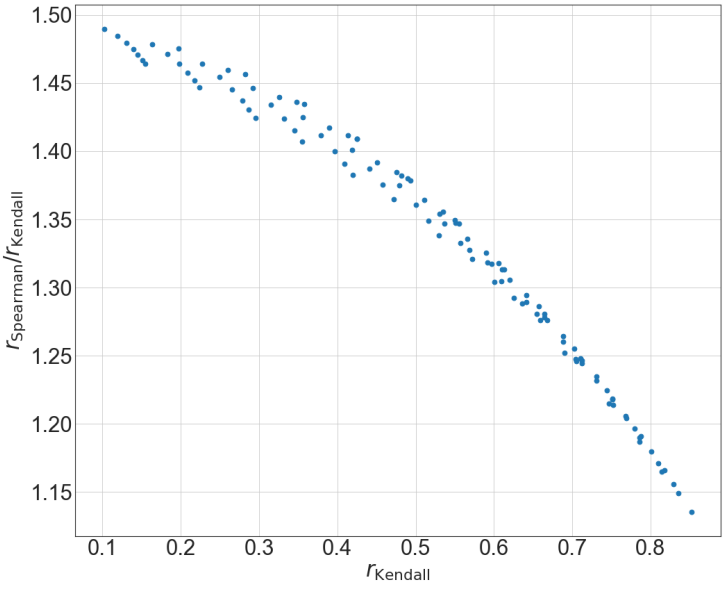


Рис. 67: График зависимости отношения $r_{\text{Spearman}}/r_{\text{Kendall}}$ от коэффициента корреляции Кендалла r_{Kendall}.

ляется, но ее нельзя использовать в качестве основы для достаточно точного прогноза.

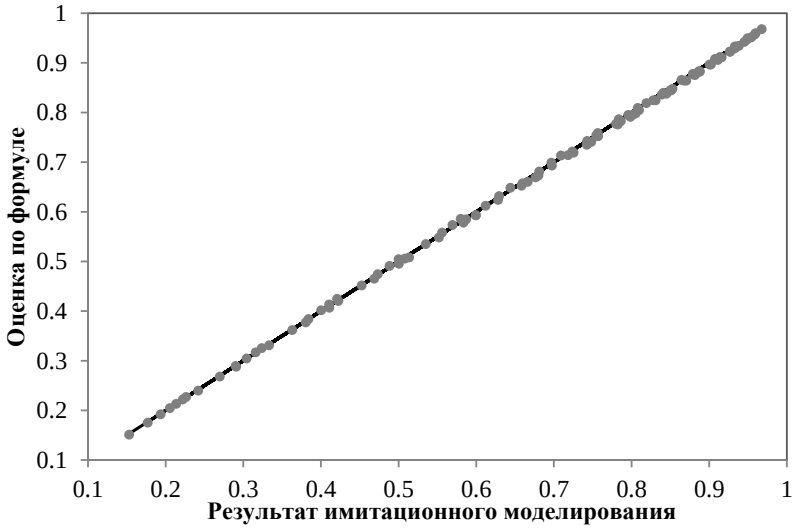


Рис. 68: Коэффициент корреляции Спирмена, рассчитанный с помощью коэффициента корреляции Кендалла по формуле (3.27).

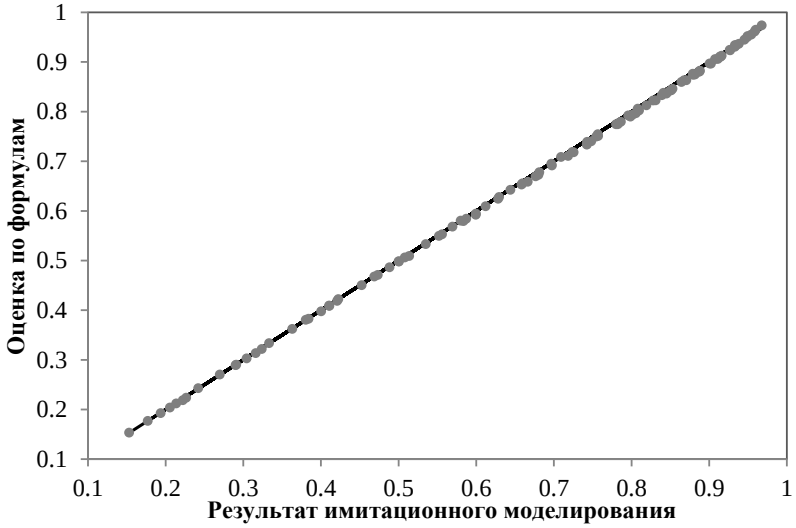
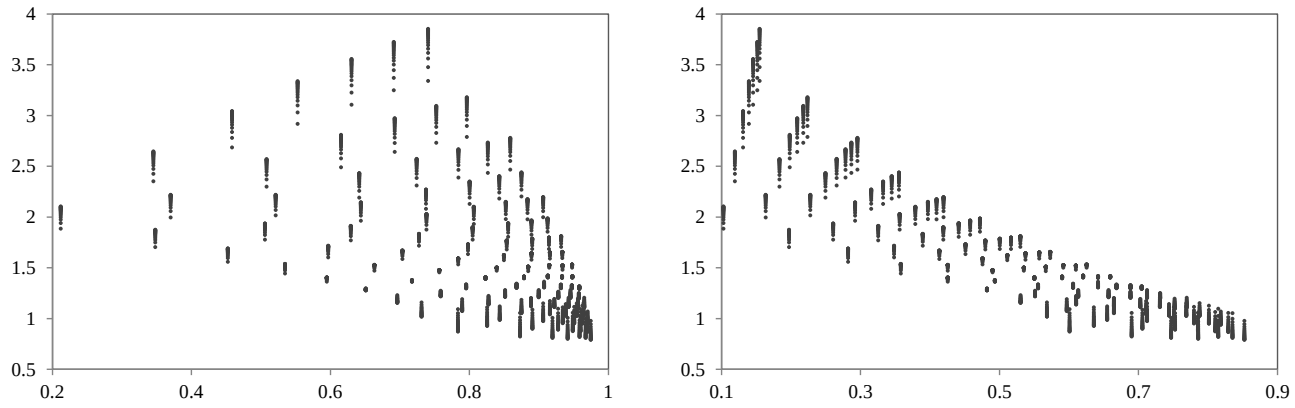


Рис. 69: Коэффициент корреляции Спирмена (аппроксимация с помощью формул (3.26) и (3.27), оценка качества).

3.6 Квантили распределения времени отклика в fork-join системах с распределением Парето времени обслуживания

Большой проблемой в изучении систем массового обслуживания с тяжелыми хвостами времен обслуживания является отсутствие явных формул распределений времен пребывания. Для систем типа fork-join эта проблема усугубляется



а) б)

Рис. 70: Модифицированное среднее время отклика (левая часть формулы (3.17)), рассчитанное с помощью имитационного моделирования, в зависимости от а) корреляции Пирсона, б) корреляции Кендалла.

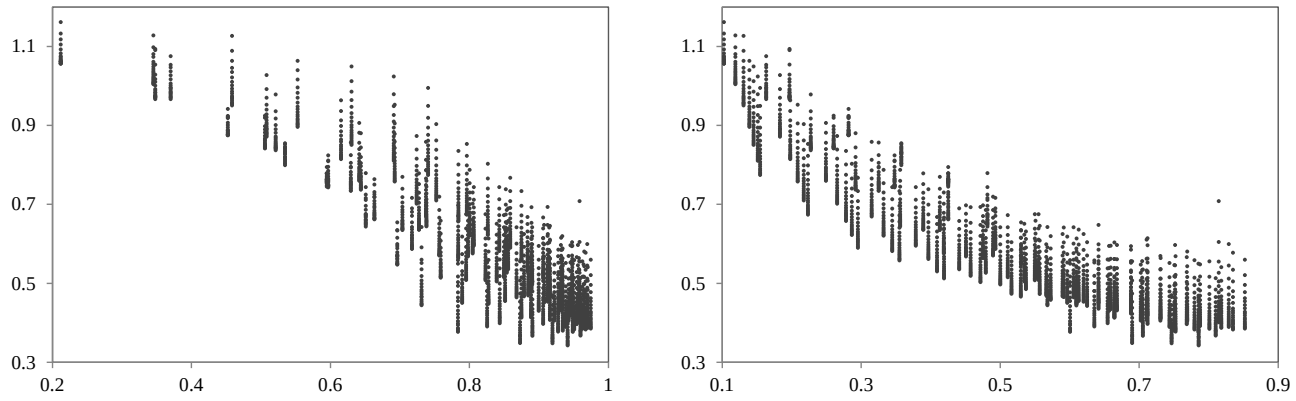
зависимостью времен пребывания в подсистемах, что приводит к отсутствию явных формул для распределений времени отклика даже в показательном случае [163, 191]. Однако эти распределения и их характеристики всегда можно приблизить какими-то формулами с той или иной точностью, вопрос только в том, как их подобрать.

Метод оценки квантилей времени отклика. Для системы с временами обслуживания, распределенными по Парето, предлагается использовать распределение Фреше, как известное распределение с тяжелым (степенным) хвостом, с целью приближения квантилей распределения времени отклика.

Функция распределения Фреше имеет вид

$$\Phi_{a,b,\gamma}(x) = \begin{cases} 0, & x \le a, \\ \exp\left\{-\left(\dfrac{x-a}{b}\right)^{-\gamma}\right\}, & x > a, \end{cases} \quad b, \gamma > 0. \tag{3.28}$$

Покажем, что хвост распределения имеет степенную асимптотику, и выведем выражение для квантилей данного распределения.



а) б)

Рис. 71: Модифицированное среднеквадратическое отклонение времени отклика (левая часть формулы (3.21)), рассчитанное с помощью имитационного моделирования, в зависимости от а) корреляции Пирсона, б) корреляции Кендалла.

Лемма 3.1. *Для случайной величины ξ с функцией распределения вида* (3.28) *справедливо, что хвост ее распределения имеет степенную асимптотику*

$$\bar{\Phi}_{a,b,\gamma}(x) \sim (x/b)^{-\gamma}, \quad x \to \infty, \tag{3.29}$$

а квантили распределения определяются следующим выражением

$$x_p = a + b(-\ln p)^{-1/\gamma}, \quad 0 < p < 1.$$

Доказательство. Воспользуемся разложением в ряд Тейлора, а точнее его частным случаем — рядом Маклорена — для хвоста распределения (3.28), для удобства обозначим через t отношение $(x-a)/b$, тогда

$$\bar{\Phi}_{a,b}(x) = P(\xi > x) = 1 - P(\xi \leqslant x) = 1 - e^{-t^{-\gamma}} =$$

$$= 1 - \left(1 + \frac{-t^{-\gamma}}{1!} + \frac{(-t^{-\gamma})^2}{2!} + \frac{(-t^{-\gamma})^3}{3!} + \ldots\right) =$$

$$= \left(\left(\frac{x-a}{b}\right)^{-\gamma} - \frac{1}{2!}\left(\frac{x-a}{b}\right)^{-2\gamma} + \frac{1}{3!}\left(\frac{x-a}{b}\right)^{-3\gamma} - \ldots\right) \sim \left(\frac{x}{b}\right)^{-\gamma}, \quad x \to \infty.$$

Теперь выразим квантиль x_p

$$p = P(\xi < x_p) = \exp\left\{-\left(\frac{x_p - a}{b}\right)^{-\gamma}\right\},$$

$$\ln p = -\left(\frac{x_p - a}{b}\right)^{-\gamma},$$

$$\frac{x_p - a}{b} = (-\ln p)^{-\frac{1}{\gamma}},$$

окончательно получаем

$$x_p = a + b(-\ln p)^{-1/\gamma}, \quad 0 < p < 1.$$

□

Распределение Фреше относится к так называемым распределениям экстремальных значений или максимум-устойчивым распределениям [93, 145] и обладает следующими интересными свойствами.

1. Пусть случайные величины $\xi_1, \ldots, \xi_n$, $n \geq 1$, независимы и имеют распределение $\Phi_{a,b,\gamma}$, тогда их максимум

$$M_n = \max(\xi_1, \xi_2, ..., \xi_n)$$

имеет распределение того же типа, а именно

$$P(M_n \leq x) = P(\xi_1 < x, ..., \xi_n < x) = P(\xi_1 < x) \cdot ... \cdot P(\xi_n < x) = \left(\Phi_{a,b,\gamma}(x)\right)^n$$

и

$$P(M_n \leq x) = \left(\Phi_{a,b,\gamma}(x)\right)^n = \Phi_{a,bn^{1/\gamma},\gamma}(x). \tag{3.30}$$

2. В более общем случае, если случайные величины $\xi_1, \ldots, \xi_n$, $n \geq 1$, независимы и имеют одинаковое распределение F с хвостом

$$\bar{F}(x) \sim (x/b)^{-\gamma}, \quad x \to \infty,$$

то

$$P\left(\frac{M_n}{n^{1/\gamma}} \leq x\right) \to \Phi_{0,b,\gamma}(x), \quad n \to \infty.$$

Исходя из этого, логично использовать распределение Фреше в тех случаях, когда речь идет о максимумах одинаково распределенных случайных величин с тяжелыми хвостами. В нашем случае времена отклика заявок представляют собой максимумы K времен пребывания подзаявок в подсистемах.

Заметим, что тип распределения Фреше может сохраняться и когда случайные величины зависимы определенным образом, как в случае fork-join системы. В этой связи обратимся к теории копул [162].

Напомним, что копулой C называется функция многомерного распределения на $[0,1]^d$, $d \geq 2$, если все частные распределения являются равномерными на $[0,1]$. Согласно знаменитой теореме Скляра, любая функция многомерного распределения в $\mathbb{R}^d$ представима в виде

$$F(x_1, \dots x_d) = C(F_1(x_1), \dots F_d(x_d)),$$

где F_i, $1 \leq i \leq d$, — функции частных распределений. Таким образом, всякому многомерному распределению можно поставить в соответствие его копулу. Если частные распределения непрерывны, то такое представление единственно.

Диагональным сечением копулы называется функция

$$\delta(u) = C(u, \dots, u).$$

Теорема 3.2. *Пусть случайные величины $\xi_1, \dots, \xi_n$, $n \geq 1$, 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.*

Доказательство.

$$P(M_n \leq x) = P(\max(\xi_1, \xi_2, ..., \xi_n) < x) =$$

$$= P(\xi_1 < x, \xi_2 < x, ...\xi_n < x) = F(x, ..., x) = C(F(x), F(x)..., F(x)) =$$

$$= C(\Phi_{a,b,\gamma}(x), \Phi_{a,b,\gamma}(x), ..., \Phi_{a,b,\gamma}(x)) = \delta(\Phi_{a,b,\gamma}(x)).$$

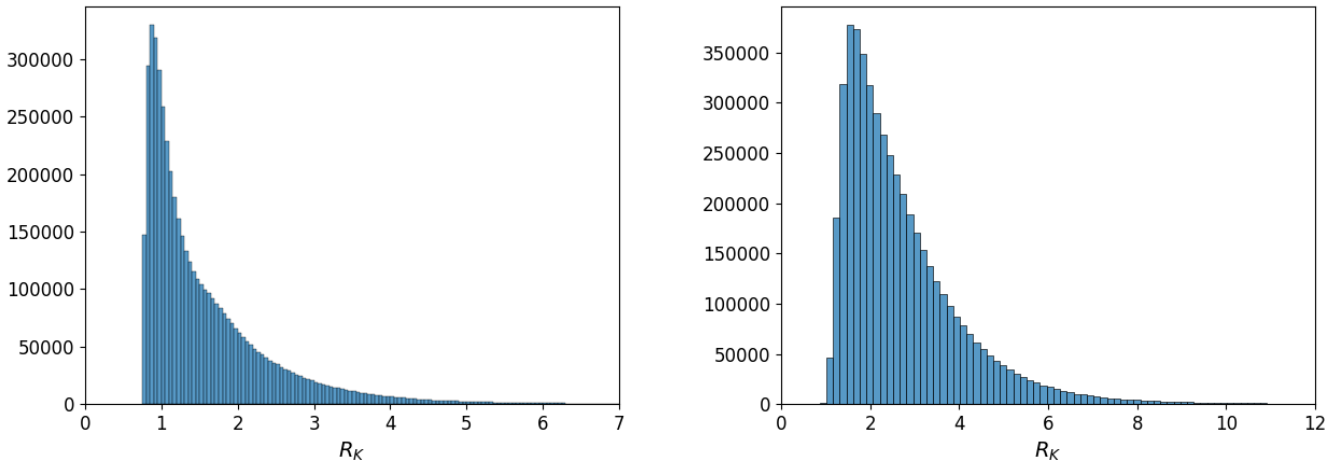
Далее с учетом того, что распределение $F = \Phi_{a,b,\gamma}$ имеет диагональное сечение, т. е. $\delta(u) = u^\nu$, $\nu > 0$, а также свойства распределения Фреше $\Phi_{a,b,\gamma}$ (3.30) получаем

$$\delta(\Phi_{a,b,\gamma}(x)) = \big(\Phi_{a,b,\gamma}(x)\big)^\nu = \Phi_{a,b\nu^{1/\gamma},\gamma}(x).$$

Итак, максимум M_n также имеет распределение Фреше вида $\Phi_{a,b\nu^{1/\gamma},\gamma}(x)$. $\square$

Таким образом, использование распределения Фреше не исключается и с учетом зависимости времен пребывания в подсистемах. А степенные диагональные сечения наблюдаются, например, у копул экстремальных значений [114, 162].

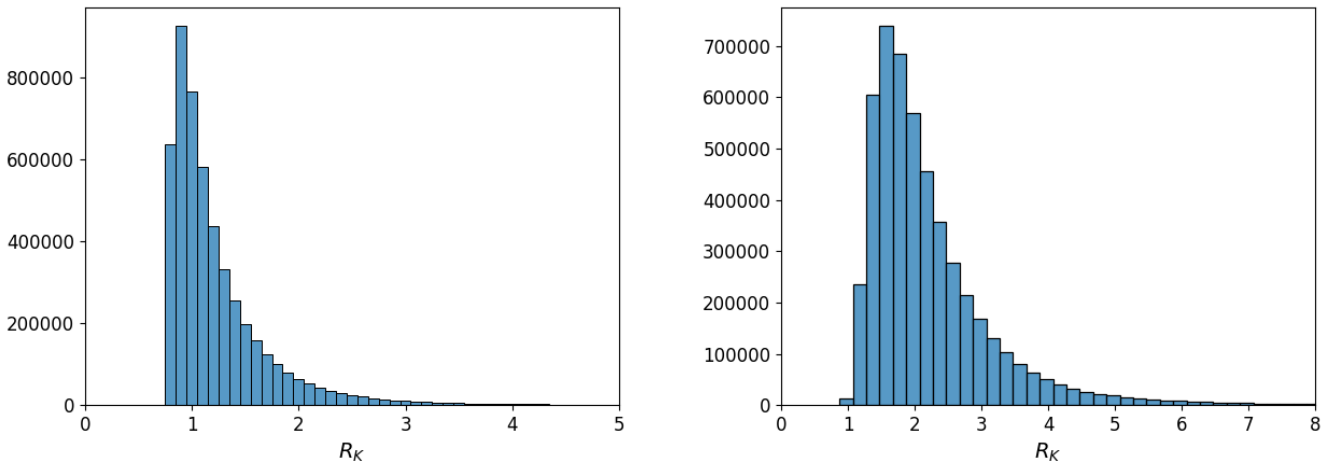
Предложение о допустимости использования распределения Фреше для случайной величины времени отклика R_K подтверждается и экспериментальными данными. Так, на рисунках 72 и 73 представлены эмпирические (выборочные) плотности распределения времени отклика fork-join системы массового обслуживания в случае с пуассоновским входным потоком и распределением Эрланга (с параметром формы $m = 3$) времени между соседними поступлениями заявок для $K = 2$ и $K = 20$. Как следует из вида гистограмм, распределение времени отклика fork-join системы схоже с видом аналитической плотности распределения для случайной величины, имеющей распределение Фреше.



а) б)

Рис. 72: Эмпирическая плотность распределения в случае пуассоновского входящего потока и распределения Парето времени обслуживания при $\alpha = 4$, $\rho = 0.4$: а) $K = 2$, б) $K = 20$.

Теорема 3.3. *Если случайная величина времени отклика системы с разделением и параллельным обслуживанием R_K приближается распределением*



а) б)

Рис. 73: Эмпирическая плотность распределения в случае распределения Эрланга (параметр формы $m = 3$) для входящего потока и распределения Парето времени обслуживания при $\alpha = 4$, $\rho = 0.4$: а) $K = 2$, б) $K = 20$.

Фреше с функцией распределения $\Phi_{a,b,\gamma}(x)$ *вида* (3.28), *то оценки квантилей времени отклика уровня* p *представимы в виде*

$$\hat{x}_p = \hat{a} + \hat{b}(-\ln p)^{-1/\gamma}, \quad 0 < p < 1, \tag{3.31}$$

где

$$\hat{a} = E[R_K] - \hat{b}E[\xi_0], \quad \hat{b} = \sqrt{\frac{Var[R_K]}{Var[\xi_0]}}, \tag{3.32}$$

а случайная величина ξ_0 *имеет стандартное распределение Фреше.*

Доказательство. Обратимся к характеристикам распределения Фреше. Известно, что случайная величина ξ_0 со стандартным распределением Фреше имеет моменты

$$E[\xi_0] = \Gamma\left(1 - \frac{1}{\gamma}\right), \quad \gamma > 1; \quad Var[\xi_0] = \Gamma\left(1 - \frac{2}{\gamma}\right) - \Gamma^2\left(1 - \frac{1}{\gamma}\right), \quad \gamma > 2.$$

Следовательно, случайная величина времени отклика fork-join системы с распределением $\Phi_{a,b,\gamma}$ будет распределена также, как величина $a + b\xi_0$, т. е.

$$R_K = a + b\xi_0. \tag{3.33}$$

Соответственно, распределение времени отклика fork-join СМО преобразуется в стандартное распределение Фреше при значениях параметров $a = 0$ и $b = 1$.

При известных γ, $E[R_K]$ и $Var[R_K]$ можно оценить параметры распределения Фреше a и b методом моментов, а затем и квантили. Приравнивая моменты с учетом выражения (3.33), получаем следующую систему уравнений

$$\begin{cases} a + bE[\xi_0] & = E[R_K], \\ b^2 Var[\xi_0] & = Var[R_K], \end{cases} \tag{3.34}$$

откуда следуют оценки параметров

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

и оценки квантилей (лемма 3.1)

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

□

Отметим, что под известными $E[R_K]$ и $Var[R_K]$ могут подразумеваться как их значения, оцененные статистически или экспериментально при каких-то параметрах СМО, так и значения, полученные по приближенным формулам (3.16) и (3.20) или с помощью нейронных сетей (при этом итоговая точность, конечно, снижается). Здесь же будут использоваться значения $E[R_K]$ и $Var[R_K]$, полученные с помощью симуляции.

Возникает вопрос о параметре γ. Наиболее естественным представляется выбор $\gamma = \alpha$ (из распределения Парето) или $\gamma = \alpha - 1$ (теоретическая асимптотика хвоста времени пребывания [169]). Числовая проверка показала, что лучшее соответствие дает $\gamma = \alpha$ (в рассматриваемой области значений параметров). Выяснилось также, что метод плохо пригоден для оценивания низких квантилей, в связи с чем было решено остановиться на высоких квантилях уровней 70–95% с шагом 5%.

Алгоритм построения оценок квантилей времени отклика для различных типов входящего потока. С помощью описанного в предыдущем разделе подхода опишем алгоритм построения оценок квантилей распределения времени отклика для различных вариантов распределений входящего потока.

Итак, последовательность действий алгоритма следующая.

1. С помощью имитационного моделирования (приближенных аналитических формул или нейросетей) определяем значения математического ожидания $E[R_K]$ и дисперсии $Var[R_K]$ времени отклика, а также квантилей его распределения x_p на сетке параметров, от которых зависят эти значения (при этом под сеткой подразумевается совокупность значений параметров на заранее выбранном интервале для каждого из них с определенным шагом).

Заметим, что параметрами, от которых зависит поведение рассматриваемой системы, традиционно в соответствии с теорий массового обслуживания являются средняя интенсивность входящего потока $1/E[\zeta]$, средняя интенсивность обслуживания на приборах $1/E[\eta]$ или обратные им величины, которые представляют собой математические ожидания для времени между соседними поступлениями заявок и времени обслуживания подзаявок (в данном случае), соответственно. Еще одним важным показателем является их отношение, представляющее собой коэффициент загрузки системы $\rho = E[\eta]/E[\zeta] < 1$. Поскольку для времени обслуживания было выбрано распределение Парето с функцией распределения (3.1), то среднее время обслуживания всегда равно одной условной временной единице $E[\eta] = 1$, что представляется довольно удобным, поэтому $\rho = 1/E[\zeta]$. Также, очевидно, имеется зависимость от числа подсистем K, параметра α распределения Парето времени обслуживания, а также от параметров распределения входящего потока, которых может быть и несколько в зависимости от конкретного вида распределения. И, конечно, поскольку данный алгоритм оценивает квантили распределения времени отклика системы x_p,

то появляется зависимость и от значения вероятности p.

2. По полученным данным имитационного моделирования для различных наборов значений математического ожидания и дисперсии времени отклика, количества подсистем K, параметра α и других параметров распределения входящего потока вычисляются оценки $\hat{a}$ и $\hat{b}$ по формулам (3.32).

Напомним, что в каждом случае в рамках имитационного моделирования методом Монте-Карло проводилось от 5 миллионов (в случае более слабой загрузки системы) до 10 миллионов (в случае более высокой загрузки системы) испытаний, благодаря чему оценки как моментов $E[R_K]$ и $Var[R_K]$, так и квантилей x_p представляются достаточно точными, поэтому рассматриваем их как «истинные» значения.

3. Подставляем полученные значения $\hat{a}$ и $\hat{b}$ в формулу (3.31), которая позволяет получить оценку квантилей распределения времени отклика $\hat{x}_p$.

Численный анализ показал, что несмотря на относительно небольшое значение средней погрешности приближения, максимальная ошибка $MaxAPE$ оказалась довольно высока, хотя и осталась в рамках инженерной погрешности. Поэтому было принято решение ввести поправочный коэффициент. Соответственно, алгоритм определения оценки квантилей дополнился следующими пунктами.

4. Построение множественной линейной или нелинейной регрессии вида

$$x_p^* = \hat{x}_p f(\alpha, p, \rho, \mathbf{C}), \tag{3.35}$$

где $f(\alpha, p, \rho, \mathbf{C})$ — это функция поправочного коэффициента, зависящая от параметров α, ρ, p и их комбинаций, а вектор $\mathbf{C}$ — это вектор числовых коэффициентов перед соответствующими переменными.

Выбор именно такой функциональной зависимости был сделан после изучения отношения $x_p/\hat{x}_p$. Анализ показал, что данное отношение практически не зависит от K и слабо зависит от α, ρ и p (по отдельности), но при этом всегда наблюдается большой разброс.

5. Определение оптимальных значений числовых коэффициентов $\mathbf{C}$ в выражении (3.35) с помощью метода оптимизации Нелдера–Мида, при этом под оптимизацией понимается минимизация модуля максимального значения относительной погрешности приближения

$$MaxAPE = \left| \frac{x_p - \hat{x}_p f(\alpha, p, \rho, \mathbf{C})}{x_p} \right| \cdot 100\% \to \min . \tag{3.36}$$

Описанный алгоритм представляется относительно трудоемким, т. к. включает в себя проведение длительного имитационного моделирования, построение оценок, подбор функции поправочного коэффициента, а также проведение оптимизации. Тем не менее, полученные результаты того стоят, поскольку конечным итогом предложенного подхода является оценка очень хорошего качества в виде аналитической формулы. Полученное выражение позволяет оценить квантили распределения времени отклика уже для любых промежуточных значений входных параметров из заранее определенных интервалов, при этом вычисление представляет собой выполнение несложных арифметических операций, что несоизмеримо сокращает временные затраты, которые потребовались на получение аналогичных значений, но уже с помощью имитационного моделирования. Кроме того, аналитическое выражение представляется более выигрышным результатом в сравнении с применением других методов интеллектуального анализа для получения интересующих оценок, например, искусственных нейронных сетей.

В следующем разделе приведены результаты применения данного подхода для нескольких вариантов распределений входящего потока. Стоит отметить, что несмотря на серьезность временных затрат на проведение имитационного моделирования, была проверена результативность предложенного алгоритма

на большом количестве промежуточных значений входных параметров, чтобы подтвердить его работоспособность.

Численный эксперимент. С помощью алгоритма построим оценки квантилей распределения времени отклика для следующих типов распределений входящего потока: распределение Эрланга, показательное распределение и гиперэкспоненциальное распределение.

Выбор был остановлен на данных типах распределений, потому что, с одной стороны, эти распределения были использованы в работах других авторов для моделирования процесса параллельной обработки данных в области информационных технологий [164, 165, 168], что, соответственно, позволяет говорить о возможности применении данной модели и подхода на практике. С другой стороны, коэффициент вариации каждого из этих распределений находится по разные стороны от единицы, т. е. для распределения Эрланга коэффициент вариации всегда меньше одного, для пуассоновского входящего потока — равен единице, а для гиперэкспоненциального распределения — больше одного. Подобная вариативность типов распределений свидетельствует о широте потенциала применения предложенного метода.

В нашем случае, речь идет об оценках для ρ от 0.1 до 0.9 с шагом 0.1, целочисленных значений α от 4 до 10 и количестве подсистем K от 2 до 20, а также вероятности $p \in \{0.70, 0.95\}$ с шагом 0.05, поскольку численный анализ показал, что предложенный метод подходит для оценивания квантилей высокого уровня. В дальнейшем будем называть данное множество параметров *исходным*.

Несмотря на то, что процесс имитационного моделирования, как уже упоминалось выше, является довольно длительным и трудоемким процессом, чтобы подтвердить результативность предложенного подхода и проверить качество полученных аналитических приближений, были смоделированы оцениваемые характеристики (квантили) для следующих промежуточных значений пара-

метров: $\rho \in \{0.15, 0.85\}$ с шагом 0.1, значений $\alpha \in \{4.5, 9.5\}$ с шагом 1, для того же количества подсистем K от 2 до 20 и тех же значений вероятностей $p \in \{0.70, 0.95\}$ с шагом 0.05. Соответственно, перечисленное множество значений параметров будем называть *промежуточным*.

Распределение Эрланга. Проанализируем случай распределения Эрланга времени обслуживания со следующей плотностью распределения

$$p_\zeta(x) = \frac{\lambda(\lambda x)^{m-1}}{(m-1)!}\exp\{-\lambda x\}, \quad x \geqslant 0, \quad \lambda > 0, \quad m \in \mathbb{N}, \tag{3.37}$$

моментами и коэффициентом вариации

$$E[\zeta] = m/\lambda, \quad Var[\zeta] = m/\lambda^2, \quad CV[\zeta] = \sqrt{Var[\zeta]}/E[\zeta] = 1/\sqrt{m}.$$

Рассмотрим значения параметра формы $m = 2$ и $m = 3$, тогда в общей сложности с помощью имитационного моделирования необходимо будет получить 14364 различных наборов данных, включающих значения вероятностей p, соответствующих им квантилей, а также значений $E[R_K]$, $Var[R_K]$, параметров распределений m и α, количества подсистем K, коэффициента загрузки ρ, параметров $\hat{a}$ и $\hat{b}$.

Погрешности приближения квантилей $\hat{x}_p$ представлены в таблице 24. Срав-

Таблица 24: Погрешности приближений значений квантилей $\hat{x}_p$, рассчитанных с помощью аналитической формулы (3.31) для случая распределения Эрланга, по сравнению с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль $\hat{x}_p$ из формулы (3.31)	11.316355	0.000157	2.167589

нительный график истинных квантилей x_p и приближенных $\hat{x}_p$ на рисунке 74 а). Видно, что хотя в среднем ошибка невелика, ее максимальное значение выглядит относительно большим.

После введения поправочного множителя в соответствии с представленным выше алгоритмом и изучения отношения $x_p/\hat{x}_p$, была получена формула

$$x_p^* = \hat{x}_p(C_1 + C_2\alpha + C_3\rho + C_4 p + C_5\rho^2 + C_6 p^2 + C_7\alpha\rho + C_8\alpha p + C_9\rho p). \quad (3.38)$$

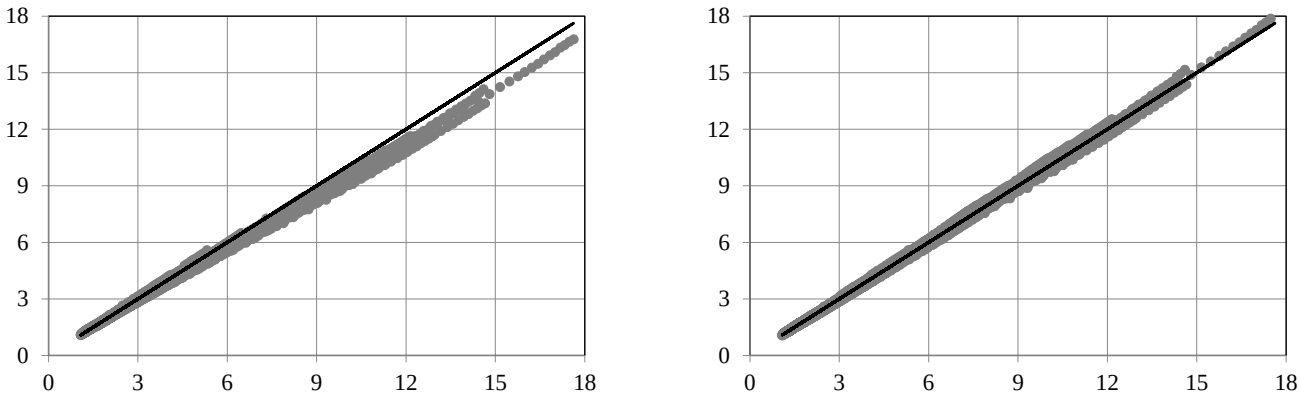
Оценка коэффициентов поправочного множителя методом оптимизации Нелдера–Мида приводит к следующим значениям

$$C_1 \approx 1.426570, \quad C_2 \approx -0.044745, \quad C_3 \approx -0.252554,$$

$$C_4 \approx -0.621259, \quad C_5 \approx 0.283248, \quad C_6 \approx 0.148713,$$

$$C_7 \approx -0.008820, \quad C_8 \approx 0.056996, \quad C_9 \approx 0.110999.$$

Погрешности приближения квантилей x_p^* представлены в первой строке таблицы (25), а сравнительный график истинных квантилей x_p и приближенных x_p^* на рис. 74 б). Заметно существенное улучшение, погрешность снижается при-



а) б)

Рис. 74: Квантили распределения времени отклика (отклонение от истинных значений, отложенных на оси абсцисс) в случае распределения Эрланга, рассчитанные а) с помощью формулы (3.31), б) с помощью формулы (3.38).

мерно в 2.3 раза и становится меньше 5%. К сожалению, далее снизить погрешность не получилось.

В соответствии с результатами, представленными во второй строке таблицы 25 для погрешностей приближений на промежуточных значениях параметров (10944 наборов данных), модуль максимальной относительной погрешности

Таблица 25: Погрешности приближений значений квантилей $\hat{x}_p$, рассчитанных с помощью аналитической формулы (3.38) для случая распределения Эрланга, по сравнению с результатами имитационного моделирования

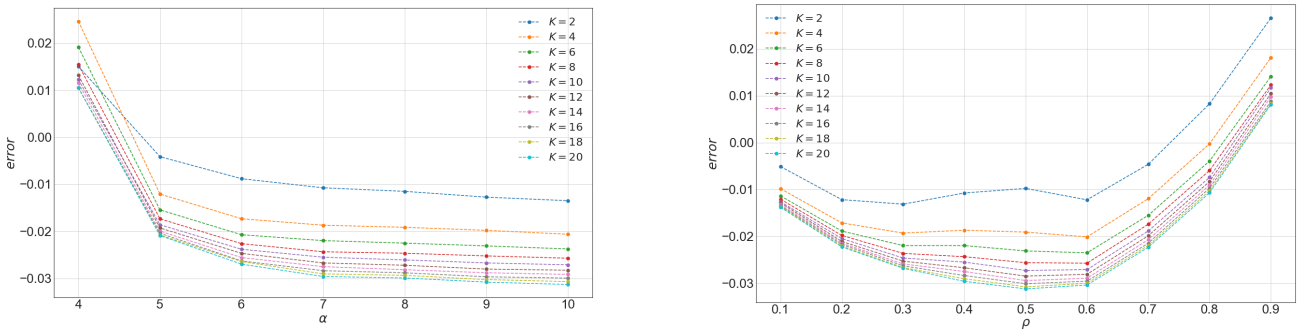
Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль x_p^* из формулы (3.38) на множестве исходных значений параметров	4.993868	0.000236	1.544850
Квантиль x_p^* из формулы (3.38) на множестве промежуточных значений параметров	5.116439	0.000366	1.457678

($MaxAPE$, %) практически не отличается от соответствующего значения в случае исходных значений параметров. Аналогичная ситуация и для двух других типов ошибок — модуля минимальной относительной погрешности ($MinAPE$, %), а также его среднего значения ($MAPE$, %).

Подобный эффект был ожидаем, поскольку изначально наблюдалась плавность изменения погрешности приближения на исходной сетке параметров. Этот факт обеспечивает небольшие отклонения уже на промежуточном множестве значений параметров. В подтверждение приведем для наглядности графики относительной погрешности приближения в зависимости от α (рис. 75 а) и от ρ (рис. 75 б)

$$error = \frac{\hat{x}_p - x_p}{x_p}$$

при фиксированных m и p для различного количества подсистем K. Как следует из графиков, разброс значений совсем не велик. Поскольку и для других типов распределений поведение погрешности (error) в контексте ее плавности повторяется, то приводить их уже не будем.



а) б)

Рис. 75: Относительная погрешность приближения (error) в случае распределения Эрланга для входящего потока и распределения Парето времени обслуживания для $m = 3$ и $p = 0.8$ в зависимости: а) от α при фиксированном значении $\rho = 0.4$, б) от ρ при фиксированном значении $\alpha = 7$.

Экспоненциальное распределение. Далее мы применим тот же метод для новых данных, смоделированных для случая пуассоновского входного потока, т. е. с плотностью распределения вида

$$p_\zeta(x) = \lambda \exp\{-\lambda x\}, \quad x \geqslant 0, \quad \lambda > 0, \tag{3.39}$$

моментами и коэффициентом вариации

$$E[\zeta] = 1/\lambda, \quad Var[\zeta] = 1/\lambda^2, \quad CV[\zeta] = 1,$$

В этом случае имеем, что $\rho = \lambda$.

Для вывода приближения было смоделировано 7182 набора данных. Погрешности аппроксимации формулы (3.31) представлены в таблице 26. Несмотря на небольшое значение средней погрешности аппроксимации, максимальное ее значение также желательно снизить. После анализа соотношения $x_p/\hat{x}_p$, введения поправочного множителя, а также оптимизации значений коэффициентов полученного выражения, имеем

$$x_p^* = \hat{x}_p(C_1 + C_2\alpha + C_3\rho + C_4 p + C_5\rho^2 + C_6 p^2), \tag{3.40}$$

Таблица 26: Погрешности приближений значений квантилей $\hat{x}_p$, рассчитанных с помощью аналитической формулы (3.31) для случая экспоненциального распределения, по сравнению с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль $\hat{x}_p$ из формулы (3.31)	12.090656	0.000060	3.536269

где

$$C_1 \approx 0.985358, \quad C_2 \approx -0.008832, \quad C_3 \approx -0.0119934,$$

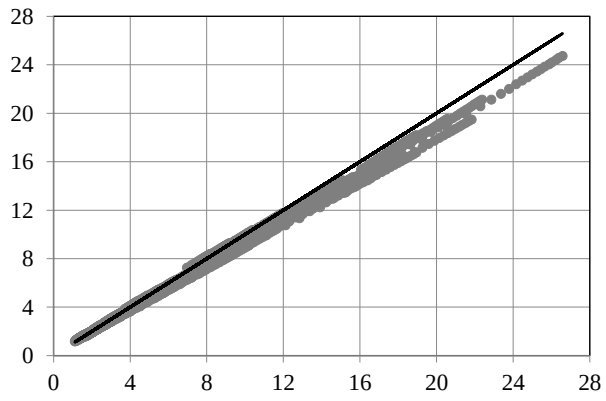
$$C_4 \approx -0.066961, \quad C_5 \approx 0.090650, \quad C_6 \approx 0.188999.$$

Как видно, число элементов в формуле (3.40) меньше, чем в предыдущем случае, т. к. получается удовлетворительное качество приближения без учета различных пар произведений α, ρ и p. Погрешности приближений x_p^* для исходного набора данных представлены в первой строке таблицы 27, а для множества промежуточных значений (5472 набора данных) — во второй строке этой же таблицы. Сравнить качество приближения до и после введения поправочного

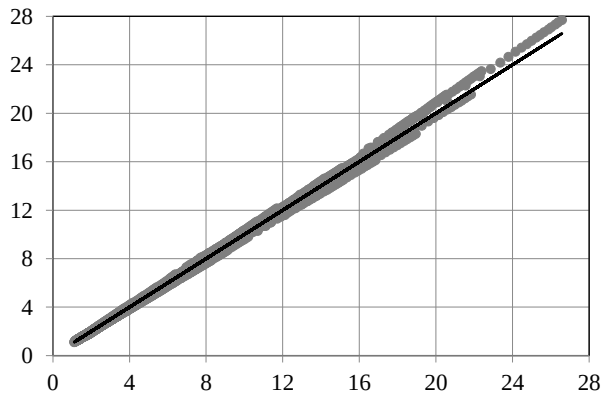
Таблица 27: Погрешности приближений значений квантилей $\hat{x}_p$, рассчитанных с помощью аналитической формулы (3.40) для случая экспоненциального распределения, по сравнению с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль x_p^* из формулы (3.40) на множестве исходных значений параметров	4.861944	0.000118	1.662532
Квантиль x_p^* из формулы (3.40) на множестве промежуточных значений параметров	5.614176	0.000740	1.531700

коэффициента можно на рисунках 76 а) и 76 б). Погрешность приближения снижается примерно в 2.5 раза.



а)



б)

Рис. 76: Квантили распределения времени отклика (отклонение от истинных значений, отложенных на оси абсцисс) в случае экспоненциального распределения, рассчитанные а) с помощью формулы (3.31), б) с помощью формулы (3.40).

Гиперэкспоненциальное распределение. Теперь проанализируем случай гиперэкспоненциального распределения для входного потока. Рассмотрим плотность распределения вида

$$p_\zeta(x) = p_1\lambda_1 \exp\{-\lambda_1 x\} + p_2\lambda_2 \exp\{-\lambda_2 x\}, \quad x \geqslant 0,$$
$$\lambda_1, \lambda_2 > 0, \quad p_1, p_2 \geqslant 0, \quad p_1 + p_2 = 1.$$

Пусть

$$\lambda_1 = \frac{\lambda}{1+\epsilon}, \quad \lambda_2 = \frac{\lambda}{1-\epsilon}, \quad 0 < \epsilon < 1.$$

Тогда при $p_1 = p_2 = 1/2$,

$$E[\zeta] = \frac{p_1}{\lambda_1} + \frac{p_2}{\lambda_2} = \frac{1}{2}\left(\frac{1+\epsilon}{\lambda} + \frac{1-\epsilon}{\lambda}\right) = \frac{1}{\lambda},$$

$$Var[\zeta] = 2\left(\frac{p_1}{\lambda_1^2} + \frac{p_2}{\lambda_2^2}\right) - E[\xi]^2 = \frac{(1+\epsilon)^2}{\lambda^2} + \frac{(1-\epsilon)^2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{2\epsilon^2+1}{\lambda^2},$$

а коэффициент вариации

$$CV[\zeta] = \sqrt{Var[\zeta]}/E[\zeta] = \sqrt{1 + 2\epsilon^2}$$

с учетом предположения $0 < \epsilon < 1$ принимает значения на интервале $(1, \sqrt{3})$. Допустим, что $\epsilon = 0.5$, следовательно $CV[\zeta] = \sqrt{1.5} \approx 1.225$. В этом случае также имеем, что $\rho = \lambda$.

Далее, действуя в рамках предложенного алгоритма, получаем следующие значения для погрешности аппроксимации на исходном множестве параметров, состоящем из 7182 наборов данных, до введения поправочного коэффициента (см. таб. 28). Затем после введения поправочного коэффициента и оптимизации

Таблица 28: Погрешности приближений значений квантилей $\hat{x}_p$, рассчитанных с помощью аналитической формулы (3.31) для случая гиперэкспоненциального распределения, по сравнению с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль $\hat{x}_p$ из формулы (3.31)	12.599848	0.003113	3.757279

получаем формулу

$$x_p^* = \hat{x}_p(C_1 + C_2\alpha + C_3\rho + C_4 p + C_5\rho^2 + C_6 p^2 + C_7\alpha\rho + C_8\alpha p + C_9\rho p). \quad (3.41)$$

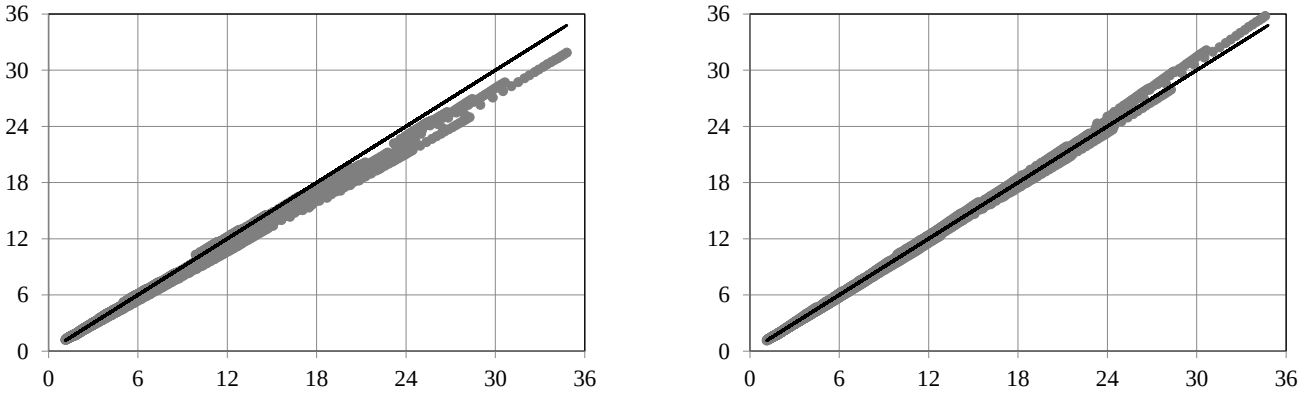
где

$$C_1 \approx 0.966123, \quad C_2 \approx -0.017217, \quad C_3 \approx 0.024434,$$

$$C_4 \approx 0.011531, \quad C_5 \approx 0.134103, \quad C_6 \approx 0.122878,$$

$$C_7 \approx -0.004496, \quad C_8 \approx 0.011486, \quad C_9 \approx -0.055572.$$

С погрешностями аппроксимации для исходного и промежуточного множества параметров, состоящего из 5472 наборов элементов, можно ознакомится в таблице 29 и на рисунках 77 а) и 77 б). Максимальное значения ошибки приближения снизилось примерно в 2.6 раз. Обратим внимание на то, что в данном



а) б)

Рис. 77: Квантили распределения времени отклика (отклонение от истинных значений, отложенных на оси абсцисс) в случае гиперэкспоненциального распределения, рассчитанные а) с помощью формулы (3.31), б) с помощью формулы (3.41).

Таблица 29: Погрешности приближений значений квантилей $\hat{x}_p$, рассчитанных с помощью аналитической формулы (3.40) для случая гиперэкспоненциального распределения, по сравнению с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль x_p^* из формулы (3.41) на множестве исходных значений параметров	4.938953	0.000107	1.882116
Квантиль x_p^* из формулы (3.41) на множестве промежуточных значений параметров	4.588924	0.001066	1.792223

случае модуль максимального значения погрешности аппроксимации для множества промежуточных значений параметров также не превышает 5%.

3.7 О еще одном методе оценки квантилей и копулы распределения времени отклика

В предыдущем разделе был предложен подход для оценки квантилей распределения времени отклика fork-join системы с распределением Парето времени обслуживания и различными вариантами распределений для входящего потока, а именно распределением Эрланга, гиперэкспоненциальным распределением и пуассоновским входящим потоком. В рамках предложенного подхода была обоснована допустимость использования распределения Фреше для аппроксимации распределения времени отклика R_K fork-join системы с распределением Парето времени обслуживания и числом $K \geq 2$ подсистем.

Далее применялся метод моментов для нахождения оценок параметров $\hat{a}$ и $\hat{b}$, участвующих в выражении для приближения квантилей времени отклика $\hat{x}_p$ уровня p, в результате было получено следующее аналитические выражение

$$\hat{x}_{p,R_K} = \hat{a}_K + \hat{b}_K(-\ln p)^{-1/\alpha}, \quad 0 < p < 1, \quad 2 \leqslant K \leqslant 20. \tag{3.42}$$

где $\hat{a}_K$ и $\hat{b}_K$ определялись с помощью оценок математического ожидания $E[R_K]$ и дисперсии времени отклика $Var[R_K]$ fork-join СМО.

Поскольку рассматривались различные варианты распределений для времен между соседними поступлениями заявок, а формул для оценок математического ожидания и дисперсии времени отклика fork-join систем известно не так много, то в качестве оценок $E[R_K]$ и $Var[R_K]$ использовались значения, полученные с помощью имитационного моделирования.

В данном разделе предлагается еще один подход к оцениванию 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). Подход тесно связан с элементами теории копул и аналогичен подходу, предложенному в Главе 2 для fork-join системы с двумя подсисте-

мами 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. Этим фактором и обуславливается использование аппроксимации вида (3.42).

Выражение (3.42) применимо не только для диапазона значений $2 \leqslant K \leqslant 20$, который рассматривался в более ранних работах при анализе fork-join СМО с распределением Парето времени обслуживания вида (3.1), но в том числе и для базового случая, а именно, когда $K = 1$. Поскольку подход основывается на оценке двумерных копул, то предлагается использовать результаты оценки для базового случая $K = 1$ при оценке квантилей в случае $K = 2$. При этом в выражении (3.42) вероятность p необходимо будет заменить на его оценку, которая будет получена далее при анализе случая $K = 2$ с помощью диагонального сечения копул.

Итак, в дальнейшем будет использоваться следующее выражение для оценки квантилей распределения времени отклика при $K = 1$

$$\hat{x}_{p,R_1} = \hat{a}_1 + \hat{b}_1(-\ln p)^{-1/\alpha}, \quad 0 < p < 1, \tag{3.43}$$

где для параметров $\hat{a}_1$ и $\hat{b}_1$ предположим следующие оценки:

$$\hat{a}_1 = \frac{A_0 + A_1\rho + A_2\rho^2}{1-\rho}, \quad \hat{b}_1 = \frac{B_0 + B_1\rho + B_2\rho^2}{1-\rho}. \tag{3.44}$$

Здесь исходим из того, что времена отклика растут асимптотически пропорционально $1/(1-\rho)$ при $\rho \to 1$, а значит, также ведут себя коэффициенты a и b, так что если их умножить на $1-\rho$, то получим какие-то нелинейные ограниченные функции от ρ на отрезке $[0, 1]$, которые попробуем приблизить квадратичными.

Для определения коэффициентов A_i и B_i, $i = 0, 1, 2$, из (3.44) воспользуемся методом оптимизации Нелдера–Мида [161]. А именно с помощью симуляции системы $M|Pa|1$ получим множество реализаций случайной величины времени

пребывания заявки в СМО, после чего статистически на основе эмпирических данных определим для нее квантили распределения x_{p,R_1}. Затем минимизируем с помощью метода Нелдера–Мида модуль относительной погрешности приближения оценки квантилей, рассчитанных по формулам (3.43) и (3.44), относительно данных, полученных с помощью имитационного моделирования, что позволит определить искомые коэффициенты.

$$\left| \frac{x_{p,R_1} - \left(\hat{a}_1 + \hat{b}_1 (-\ln p)^{-1/\alpha} \right)}{x_{p,R_1}} \right| \xrightarrow[A_0,A_1,A_2, \quad B_0,B_1,B_2]{} \min$$

Для коэффициента загрузки системы рассматривается диапазон значений $\rho = \{0.1, 0.2, ..., 0.9\}$. А для вероятностей p, т. е. уровней квантилей, выбраны значения $\{0.30, 0.35, ..., 0.85, 0.90\}$, поскольку, как правило, на практике в большей степени интерес представляют квантили именно более высокого порядка.

В результате получим

$$\begin{aligned} &A_0 \approx 0.129674, \quad A_1 \approx -1.650335, \quad A_2 \approx 0.316858, \\ &B_0 \approx 0.702442, \quad B_1 \approx 0.917551, \quad B_2 \approx -0.149744, \end{aligned} \tag{3.45}$$

Оценки погрешностей приближений для формул (3.43)–(3.45) представлены в таблице 30 и для наглядности изображены на рисунке 81 а). В таблице приведены абсолютные значения относительных погрешностей приближений для 117 рассчитанных значений $\hat{x}_{p,R_1}$: максимальная погрешность приближения ($MaxAPE$, %), минимальная погрешность приближения ($MinAPE$, %) и средняя погрешность ($MAPE$, %).

Теорема 3.4. *Для системы с разделением и параллельным обслуживанием с двумя подсистемами ($K = 2$) вида $M|Pa|1$, где ξ_1, ξ_2 — случайные величины времен пребывания подзаявок в этих подсистемах со строго возрастающей функцией распределения $G(x)$, выражение для определения вероятностей p квантилей распределения случайной величины времени отклика системы R_2 при предположении о степенном виде диагонального сечения его копулы име-*

Таблица 30: Погрешности приближений значений квантилей распределения времени отклика системы x_{p,R_1} $(K = 1)$, рассчитанных с помощью аналитических формул (3.43),(3.44) и (3.45) в сравнении с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль времени отклика x_{p,R_1}	5.731694	0.004086	2.240293

ет вид

$$p = u_p^{\beta}, \quad 1 \leqslant \beta \leqslant 2,$$

где u_p *— это квантиль уровня* p *распределения случайной величины* $V = \max(G(\xi_1), G(\xi_2))$.

Доказательство. В силу допущения о том, что частная функция распределения случайных величин ξ_1 и ξ_2 времен пребывания подзаявок в подсистемах $M|Pa|1$ имеет вид $G(x)$ и является строго возрастающей, их совместная функция распределения согласно теореме Скляра представима в виде

$$G_{\xi_1,\xi_2}(x_1, x_2) = P(\xi_1 < x_1, \xi_2 < x_2) = C(G(x_1), G(x_2)),$$

где $C(u_1, u_2)$ — это копула–функция совместного распределения случайных величин ξ_1 и ξ_2.

Тогда для случайной величины их максимума $R_2 = \max(\xi_1, \xi_2)$ функция распределения примет вид

$$G_{R_2}(x) = P(\max(\xi_1, \xi_2) < x) = P(\xi_1 < x, \xi_2 < x) = C(G(x), G(x)) = C(u, u).$$

При этом величина

$$\delta(u) = C(u, u), \quad 0 \leqslant u \leqslant 1$$

называется диагональным сечением копула–функции. Соответственно, имеем, что

$$G_{R_2}(x) = C(G(x), G(x)) = \delta(G(x)).$$

Поэтому уравнение для определения квантили времени отклика принимает вид

$$G_{R_2}(x_{p,R_2}) = \delta(G(x_{p,R_2})) = p,$$

откуда

$$x_{p,R_2} = x_p = G^{-1}\left(\delta^{-1}(p)\right) = x_{\delta^{-1}(p),R_1}. \tag{3.46}$$

Стоит напомнить, что диагональное сечение характеризуется (необходимыми и достаточными) свойствами

$$\begin{gathered} \max\{2u-1,0\} \le \delta(u) \le u; \\ 0 \le \delta(u_2) - \delta(u_1) \le 2(u_2 - u_1), \quad 0 \le u_1 \le u_2 \le 1. \end{gathered} \tag{3.47}$$

которыми обладает, в частности, степенная функция

$$\delta(u) = u^{\beta}, \quad 1 \le \beta \le 2.$$

При этом значение параметра $\beta = 1$ определяет абсолютную положительную зависимость случайных величин, а значение $\beta = 2$ — их полную независимость.

Введем случайные величины $U_1 = G(\xi_1)$ и $U_2 = G(\xi_2)$, которые будут иметь равномерное распределение на отрезке $[0,1]$, т. е.

$$U_i = G(\xi_i) \sim R[0,1], \quad i = 1,2.$$

Затем рассмотрим случайную величину $V = \max(U_1, U_2)$, для которой в силу предположения о том, что функция распределения $G(x)$ является строго возрастающей, будет справедливо следующее

$$\begin{gathered} V = \max(U_1, U_2) = \max(G(\xi_1), G(\xi_2)) = \\ = G(\max(U_1, U_2)) = P(V < u) = p, \end{gathered} \tag{3.48}$$

где p — это вероятность, $0 \le p \le 1$.

С другой стороны, для диагонального сечения справедливо

$$\delta(u) = C(u,u) = P(\xi_1 < x, \xi_2 < x) = P(G(\xi_1) < G(x), G(\xi_2) < G(x)) =$$

$$= P(U_1 < u, U_2 < u) = P(\max(U_1, U_2) < u) = P(V < u) = p,$$

т. е.

$$\delta(u_p) = P(V < u_p) = p,$$

где u_p — это квантиль уровня p распределения случайной величины V. В условиях предположения о степенном виде диагонального сечения получаем выражение для определения вероятности p, которое впоследствии будет использоваться для оценки квантилей времени отклика R_2

$$p = \delta(u_p) = u_p^{\beta}. \tag{3.49}$$

□

При дальнейших исследования обратимся к копуле Гумбеля

$$C(u_1, u_2) = \exp\{-((-\ln u_1)^{\theta} + (-\ln u_2)^{\theta})^{1/\theta}\}, \quad \theta \geq 1, \quad u_1, u_2 \in [0, 1], \tag{3.50}$$

т. к. она характеризуется степенным диагональным сечением следующего вида

$$\delta(u) = u^{2^{1/\theta}}. \tag{3.51}$$

Далее с помощью имитационного моделирования системы с разделением и параллельным обслуживанием заявок определяется множество пар значений величин (ξ_1, ξ_2). Речь идет о порядка 5–10 миллионов таких пар, т. е. получаем $N = 5$ миллионов (для низких уровней загрузки системы) или $N = 10$ миллионов (для высоких уровней загрузки) реализаций случайных величин (ξ_1^i, ξ_2^i), где $i = 1, ..., N$.

Затем для каждой пары (ξ_1^i, ξ_2^i) необходимо определить величину V_i, которая согласно (3.48) определяется выражением $G(\max[G(\xi_1^i), G(\xi_2^i)]) = G(\max(U_1^i, U_2^i))$. Однако, как уже упоминалось ранее, в силу того, что вид функции распределения $G(x)$ неизвестен, напрямую определить V_i по значениям (ξ_1^i, ξ_2^i) не получится. Поэтому воспользуемся универсальным методом нормированных рангов (описанным, например, в [154, §5.5.2]), согласно которому можно

использовать асимптотическую оценку

$$\hat{V}_i = \max\left(\hat{G}(\xi_1^i), \hat{G}(\xi_2^i)\right) = \max\left(\hat{U}_1^i, \hat{U}_2^i\right) = \max\left(\frac{\mathrm{rang}(\xi_1^i)}{N+1}, \frac{\mathrm{rang}(\xi_2^i)}{N+1}\right),$$

где $\mathrm{rang}(\cdot)$ — это ранг, т. е. порядковый номер реализации аргумента после сортировки в порядке возрастания, $i = 1, ..., N$.

Далее для полученных значений реализаций случайной величины $\hat{V}$ оцениваем квантили ее распределения, т. е. находим статистические значения (u_p, p), упорядочивая значения $\hat{V}_i$ по возрастанию, в результате получаем

$$(u_p, p) = \left(\hat{V}_{(k)}, \frac{k}{N+1}\right),$$

где $\hat{V}_{(k)}$ — это k-я порядковая статистика, $k = 1, ..., N$. Пары (u_p, p) оцениваются, как и ранее, для различных значений коэффициента загрузки системы $\rho = \{0.10, 0.15, ..., 0.90\}$ и вероятностей $p = \{0.30, 0.25, ..., 0.85, 0.90\}$.

Далее на имеющихся данных (u_p, p) для каждого выбранного значения коэффициента загрузки ρ проводится графический анализ для определения конкретного вида функциональной зависимости p от u_p из (3.49)

$$p \approx \widehat{p} = \widehat{\delta}(u_p, \rho) = u_p^{\widehat{\beta}}.$$

На рисунке 78 изображена зависимость $\ln p$ от значений $\ln u_p$. Как следует из вида графика, между $\ln p$ и $\ln u_p$ можно допустить линейную зависимость, причем пучок прямых проходит через начало координат, соответственно, свободный коэффициент отсутствует, т. е. получаем соотношение

$$\ln p \approx \widehat{\beta}(\rho) \cdot \ln u_p,$$

что соответствует предположению о степенной зависимости p от u_p. Далее подберем функциональную зависимость для $\widehat{\beta}(\rho)$. Для этого построим на имеющихся данных график зависимости отношения $\ln p / \ln u_p$. Из вида рисунка 79 можно предложить квадратичную зависимость. При $\rho \to 0$ времена пребывания подзаявок асимптотически независимы, что в соответствии с теорией копул

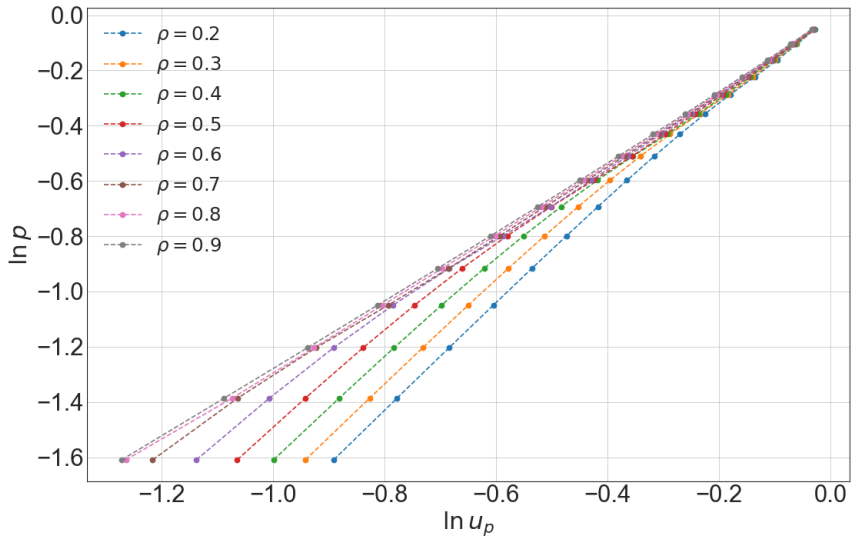


Рис. 78: Зависимость $\ln p$ от $\ln u_p$.

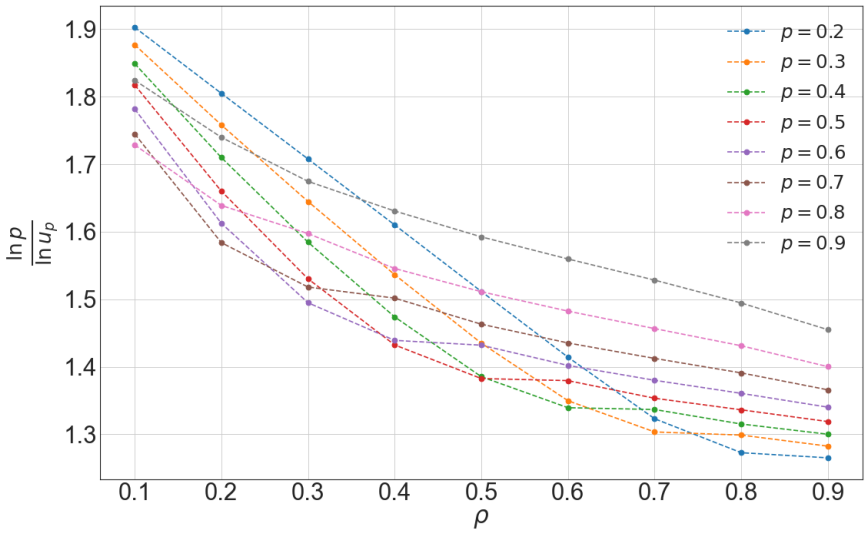


Рис. 79: Зависимость $(\ln p / \ln u_p)$ от ρ.

дает $\beta \to 2$. Поэтому логично допустить, что кривые проходят через точку с координатами $(0, 2)$, при этом вершина параболы сдвинута вправо и обращена вниз, что говорит об отрицательном значении числового коэффициента перед ρ и положительном при ρ. Таким образом, имеем

$$\frac{\ln p}{\ln u_p} = \beta(\rho) \approx 2 - C_1\rho + C_2\rho^2.$$

В результате получаем следующее выражение

$$p = \delta(u_p, \rho) \approx u_p^{2-C_1\rho+C_2\rho^2}. \tag{3.52}$$

Теперь необходимо найти значения коэффициентов C_1 и C_2. Для этого воспользуемся методом оптимизации Нелдера–Мида и минимизируем модуль относительной погрешности полученного аналитического (степенного) приближения $\widehat{p}$ по сравнению со значениями p, полученными с помощью имитационного моделирования

$$\left| \frac{p - u_p^{2-C_1\rho+C_2\rho^2}}{p} \right| \xrightarrow[C_1,C_2]{} \min.$$

Таким образом, значения коэффициентов равны

$$C_1 \approx 1.334476, \quad C_2 \approx 0.550919, \tag{3.53}$$

соответственно, искомая оценка имеет вид

$$p = \delta(u_p, \rho) \approx u_p^{2-1.334476\rho+0.550919\rho^2}. \tag{3.54}$$

На рисунке 80 представлены результаты имитационного моделирования вероятностей или уровней p квантилей u_p случайной величины $V = G(\max(U_1, U_2))$ в сравнении с результатами вычислений по аналитической формуле (3.54) в диапазоне значений $[0.30, 0.90]$ с шагом 0.05. Каждая точка, изображенная на графике, представляет собой множество из 9 точек по числу значений коэффициента загрузки $\rho \in \{0.10, 0.20, ..., 0.90\}$, которые накладываются друг на друга. Для ясности в таблице 31 приведены абсолютные значения относительных погрешностей приближений для 117 рассчитанных значений p Сравнительная анализ результатов имитационного моделирования уровня p квантилей u_p с.в. V с результатами аналитической формулы (3.54) показал, что средняя погрешность приближения составляет около 2.2%.

Далее поскольку вид функции распределения $G(x)$ случайных величин ξ_1 и ξ_2 времен пребывания подзаявок в подсистемах $M|Pa|1$ неизвестен, то вос-

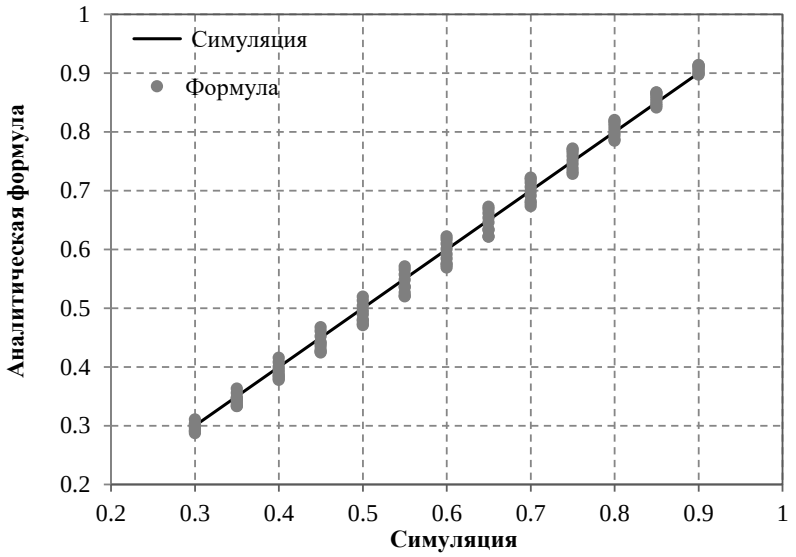


Рис. 80: Сравнение аналитических результатов формулы (3.54) с имитационным моделированием значений p квантилей u_p случайной величины $V = G(\max(U_1, U_2))$ для значений $\rho \in \{0.10, 0.20, ..., 0.90\}$.

Таблица 31: Погрешности приближений значений вероятностей p, рассчитанных с помощью аналитической формулы (3.54) в сравнении с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Вероятность p из формулы (3.54)	5.731694	0.061445	2.311958

пользуемся соотношением (3.43) для квантилей. А именно при $K = 1$ имеем

$$G(x_{p,R_1}) = p, \quad x_{p,R_1} = G^{-1}(p) = \hat{a}_1 + \hat{b}_1(-\ln p)^{-1/\alpha}.$$

Теперь с учетом уравнения (3.46) для определения квантилей времени отклика R_2 в случае $K = 2$ выразим $\delta^{-1}(p)$

$$\delta(u_p) = u_p^{2-C_1\rho+C_2\rho^2} = p,$$

$$(2 - C_1\rho + C_2\rho^2) \ln u_p = \ln p,$$

$$u_p = \delta^{-1}(p) = p^{\frac{1}{2-C_1\rho+C_2\rho^2}}.$$

Поэтому можем записать следующее

$$x_{p,R_2} = x_p = G^{-1}(\delta^{-1}(p)) = G^{-1}\left(p^{\frac{1}{2-C_1\rho+C_2\rho^2}}\right) = \hat{x}_p = \hat{a}_1 + \hat{b}_1\left(-\ln p^{\frac{1}{2-C_1\rho+C_2\rho^2}}\right)^{-1/\alpha}.$$

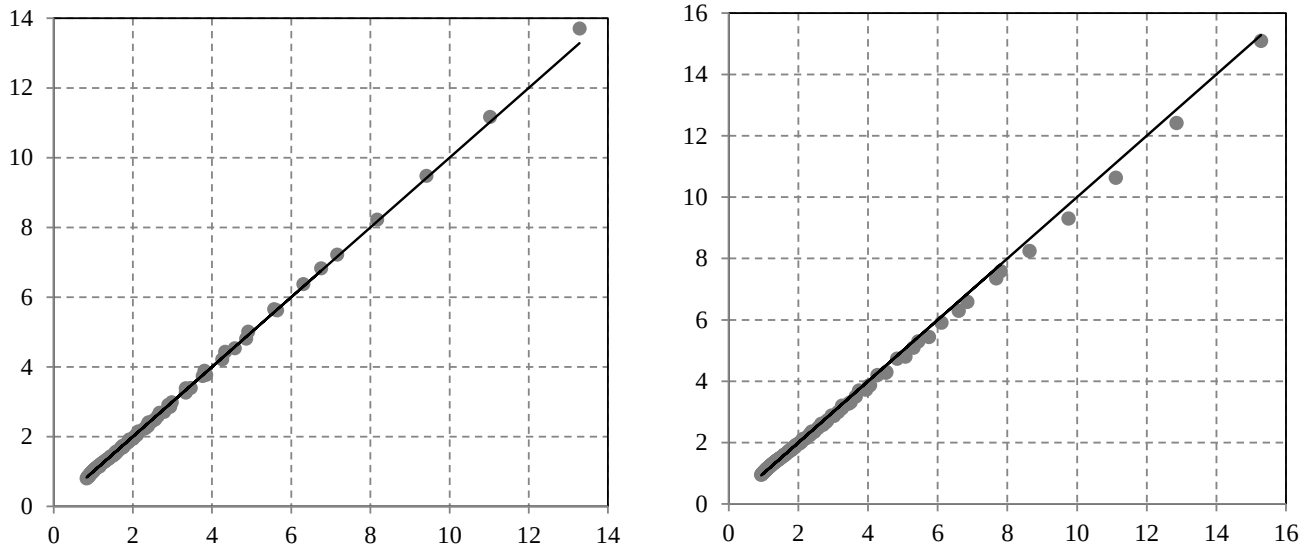
В результате, окончательно получаем следующее выражение для оценки квантилей времени отклика R_2 уровня p, $0.30 \leqslant p \leqslant 0.90$,

$$\hat{x}_p = \hat{a}_1 + \hat{b}_1\left(-\frac{\ln p}{2 - 1.334476\rho + 0.550919\rho^2}\right)^{-1/\alpha}, \quad 0 < p < 1, \tag{3.55}$$

где

$$\begin{aligned} \hat{a}_1 &\approx \frac{0.129674 - 1.650335\rho + 0.316858\rho^2}{1-\rho}, \\ \hat{b}_1 &\approx \frac{0.702442 + 0.917551\rho - 0.149744\rho^2}{1-\rho}. \end{aligned} \tag{3.56}$$

Далее аналогично оценим качество аппроксимации полученного выражения



а) б)

Рис. 81: Сравнение эмпирических и аналитических квантилей распределения случайной величины времени отклика R_K системы с разделением и параллельным обслуживанием: а) рассчитанных по формулам (3.43)–(3.45), $K = 1$; б) рассчитанных по формулам (3.55)–(3.56), $K = 2$.

(3.55) для 117 рассчитанных значений квантилей при $\rho \in \{0.10, 0.20, ..., 0.90\}$ и $p \in \{0.20, 0.25, ..., 0.90\}$. Результаты представлены в таблице 32. Получаем,

Таблица 32: Погрешности приближений значений квантилей распределения времени отклика системы x_p $(K = 2)$, рассчитанных с помощью аналитических формул (3.55) и (3.56) в сравнении с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
Квантиль времени отклика x_{p,R_2}	5.476451	0.001325	2.949008

что максимум модуля относительной ошибки составляет около 5.7%, а среднее значение этого модуля равно примерно 2.2%. На рисунке 81 б) также можно наглядно ознакомится результатами.

Приближение копулы времен пребывания подзаявок в подсистемах копулой Гумбеля. В предыдущем разделе была получена оценка диагонального сечения копулы $\delta(u)$. В данном разделе будет представлено аналитическое выражение, оценивающее саму копулу $C(u_1, u_2)$. Для этого потребуются эмпирические данные, проанализировав которые, можно будет сделать вывод о близости исследуемой копулы к одному из известных семейств, в частности, копулам Гумбеля.

Алгоритм построения эмпирической копулы будет следующим:

1. имитационное моделирование множества пар (ξ_1^k, ξ_2^k) случайных величин времен пребывания в подсистемах $M|Pa|1$ fork-join СМО, где k — это порядковый номер смоделированной пары значений, $k = 1, ..., N$, N — объем выборки (общее число пар случайных величин);

2. преобразование случайных величин (ξ_1^k, ξ_2^k) методом нормированных рангов (см. [154, §5.5.2]) в случайные величины с асимптотически равномер-

ным распределением на отрезке $[0,1]$, $U_i \sim R[0,1]$, $i = 1,2$,

$$(U_1^k, U_2^k) = \left(\frac{\text{rang}(\xi_1^k)}{N+1}, \frac{\text{rang}(\xi_2^k)}{N+1} \right);$$

3. разбиение единичного квадрата на более мелкие квадраты (сетку) со сторонами длиной $h = 1/m$, где, например, $m = 20$ и определение числа точек (U_1^k, U_2^k), попадающих в каждый из квадратов, вершинами которого являются точки $(0,0)$, $(ih,0)$, $(0,jh)$, (ih,jh), $i,j = 1,...,m$, и нормирование полученного значения, т. е.

$$C_{ij} = C(ih,jh) \approx \widehat{C}_{ij} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}\{U_1^k < ih, U_2^k < jh\},$$

где $\mathbf{1}\{\cdot\}$ — функция-индикатор события $\{\cdot\}$.

На рисунке 82 представлен график эмпирической копулы или, что то же самое, совместной функции распределения случайного вектора (U_1, U_2), построенной в соответствии с представленным выше алгоритмом.

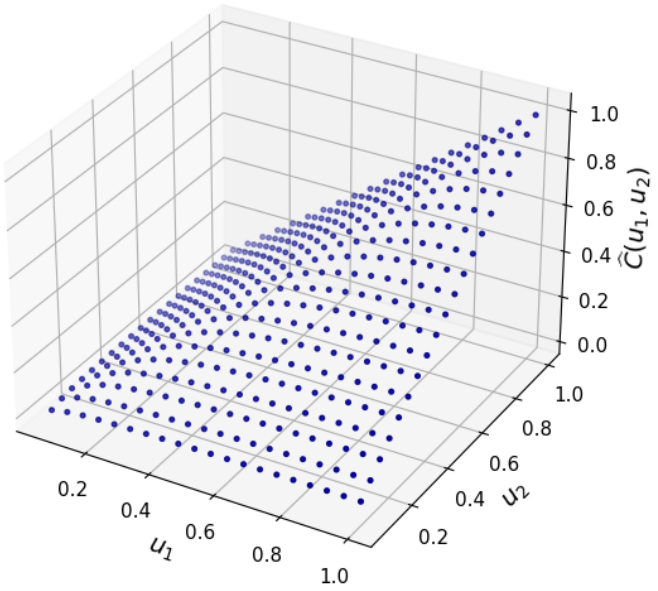


Рис. 82: Эмпирическая копула $\widehat{C}(u_1, u_2)$.

Исходя из внешнего вида полученной эмпирической функции на рисунке 82, а также учитывая, что диагональное сечение рассматриваемой копулы было

приближено в предыдущем разделе выражением вида

$$\delta(u) \approx u^{\beta}, \quad \beta = 2 - C_1\rho + C_2\rho^2, \tag{3.57}$$

будем приближать искомую копулу $C(u_1, u_2)$ копулой Гумбеля, которая имеет вид

$$C_g(u_1, u_2) = \exp\{-[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}]^{\frac{1}{\theta}}\}, \tag{3.58}$$

где $\theta \in [1, +\infty)$ — параметр копулы, который предстоит оценить.

Поскольку для копулы Гумбеля диагональное сечение имеет следующий вид

$$\delta_g(u) = C_g(u, u) = u^{2^{1/\theta}},$$

то с учетом (3.57) получаем, что

$$\theta \approx \frac{\ln 2}{\ln \beta} = \frac{\ln 2}{\ln(2 - C_1\rho + C_2\rho^2)}. \tag{3.59}$$

Далее снова воспользуемся методом оптимизации Нелдера–Мида для минимизации модуля относительной ошибки приближения функции копулы Гумбеля (3.58) с учетом того, что параметр θ определяется выражением (3.59), при сравнении с “истинными” значениями функции копулы Гумбеля, полученными с помощью имитационного моделирования для различных коэффициентов загрузки $\rho \in \{0.10, 0.20, ..., 0.90\}$. Как и раньше, не будем рассматривать квантили низкого уровня, т. е. пусть $u_1, u_2 \in \{0.30, 0.35, ..., 0.90\}$. В результате получаем следующие значения искомых коэффициентов

$$C_1 \approx 1.068768, \quad C_2 \approx 0.202125 \tag{3.60}$$

поэтому

$$C(u_1, u_2) \approx$$
$$\approx \exp\left\{-\left((-\ln u_1)^{\frac{\ln 2}{\ln(2-1.069\rho+0.202\rho^2)}} + (-\ln u_2)^{\frac{\ln 2}{\ln(2-1.069\rho+0.202\rho^2)}}\right)^{\frac{\ln(2-1.069\rho+0.202\rho^2)}{\ln 2}}\right\}. \tag{3.61}$$

Как видно из (3.53) и (3.60), полученные оценки коэффициентов различны, однако соответствующие им оценки функции $\beta(\rho)$ мало различаются между собой на рассматриваемом промежутке загрузки ρ, что позволяет говорить о значительной согласованности между ними.

Что касается погрешности аппроксимации формулы (3.61), то в таблице 33 представлены значения максимальной ($MaxAPE$), минимальной ($MinAPE$) и средней относительной ошибки аппроксимации ($MAPE$), первая из которых не превышает 10%, на наборе данных из 1521 троек (ρ, u_1, u_2). В той же таблице приведены результаты в случае оценки по коэффициентам из (3.53). На рисунке 83 также представлены графики эмпирической функции копулы и копулы, определяемой выражением (3.61) на заданном диапазоне значений $0.3 \leqslant u_1, u_2 \leqslant 0.9$.

Таблица 33: Погрешности приближений функции копулы Гумбеля $C(u_1, u_2)$ формулой (3.61)

Оцениваемая характеристика	Типы ошибок		
	$MaxAPE$, %	$MinAPE$, %	$MAPE$, %
$C(u_1, u_2)$, значения C_1, C_2 из (3.60)	8.248810	0.000686	2.735456
$C(u_1, u_2)$, значения C_1, C_2 из (3.53)	14.733185	0.000385	1.999350

Заметим также, что если проводить оценку параметра θ классическим методом максимального правдоподобия (соответствующей функцией Python для копулы Гумбеля), то полученные значения, количество которых в данном случае будет соответствовать количеству значений коэффициента корреляции $\rho \in [0.1, 0.9]$ с шагом 0.10, т. е. их будет всего 9, на тех же 1521 тройках значений (ρ, u_1, u_2) приближение копулой Гумбеля показывает большие погрешности. В этом случае $MaxAPE \approx 15.7\%$, $MinAPE \approx 0.0009\%$ и $MAPE \approx 2.34\%$.

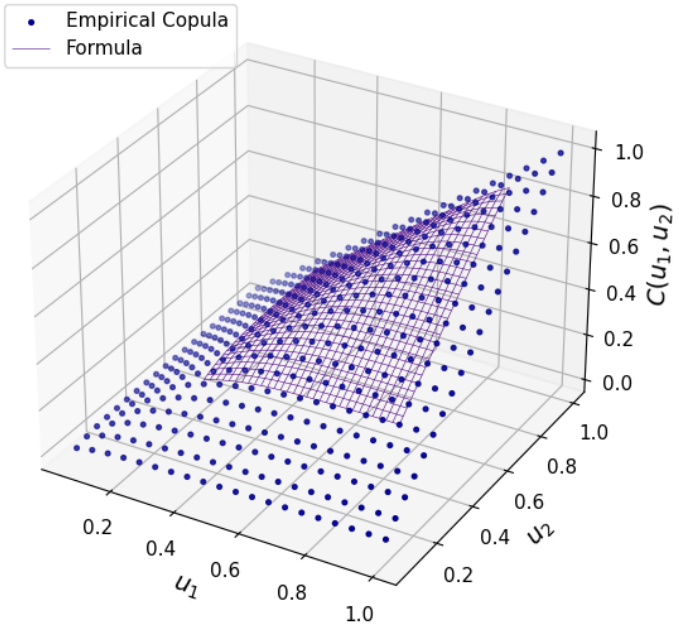


Рис. 83: Сравнение эмпирической копулы $\widehat{C}(u_1, u_2)$ и аналитической формулы (3.61).

3.8 Выводы к главе 3

В Главе 3 с помощью ИНС получены оценки математического ожидания и среднеквадратического отклонения времени отклика системы с разделением и параллельным обслуживанием с пуассоновским входным потоком и с распределением Парето времени обслуживания на приборах, а также представлено аналитическое выражение для верхней границы среднего времени отклика, полученное с помощью элементов теории массового обслуживания и теории порядковых статистик. Точность аппроксимации ИНС довольно высока в том числе и по сравнению с известными ранее аналитическими результатами.

Применение нейросетей к анализу показателей качества обслуживания в СМО типа fork-join позволяет расширить класс рассматриваемых моделей в силу отсутствия ограничений на предположения о типе входящего потока или распределения времен обслуживания, которые, как правило, сильно затрудняют аналитический или численный процесс решения поставленной задачи, при этом без потери в качестве оценки искомых параметров.

На примере fork-join системы того же типа представлен новый метод постро-

ению оценок ее различных характеристик уже в аналитическом виде. Метод основан на комбинации имитационного моделирования, визуального анализа данных с линейной (и нелинейной) регрессией, а также методом нелинейной оптимизации Нелдера–Мида. Благодаря использованию этого метода, получены аналитические нелинейные выражения для оценок среднего времени отклика, среднеквадратического отклонения времени отклика и коэффициентов корреляции Пирсона, Спирмена и Кендалла. Максимальная относительная погрешность аппроксимации в первом случае оказалась не более 3.939% и 4.061%, соответственно, а для коэффициентов корреляции — 2.380%, 1.569% и 1.376%. Стоит отметить, что погрешность до 5% считается хорошей в литературе по этой теме [163, 165, 194, 195]. Таким образом, полученные формулы хорошо работают в рассмотренных областях значений параметров. При необходимости можно получить аналогичные формулы и в других областях, используя описанные методы построения оценок.

Также в Главе 3 представлен новый метод оценки квантилей распределения времени отклика fork-join системы массового обслуживания с распределением Парето времени обслуживания на приборах и различными вариантами распределений времен между соседними поступлениям заявок.

Метод основан на приближении распределения времени отклика распределением Фреше, допустимость которого подтверждается как аналитическими, так и экспериментальными данными. Кроме того в рамках предложенного метода используется имитационное моделирование и метод оптимизации.

Полученные аналитические выражения демонстрируют хорошее качество аппроксимации квантилей распределения времени отклика высоких уровней, как на исходном наборе значений параметров (по которому строились оценки), так и на наборе промежуточных значений этих параметров. Описанный метод можно применить для анализа характеристик fork-join СМО и других конфигураций в том случае, когда речь идет о распределениях с тяжелыми хвостами. Тем не менее, аналитические выражения должны подбираться индивидуально

для каждого конкретного случая.

Также изучены приближения совместного распределения времен пребывания подзаявок с помощью теории копул. Получено хорошее соответствие с данными для степенных диагональных сечений. На основе оценок диагональных сечений выведены оценки квантилей времени отклика в широком диапазоне значений вероятностей или уровней квантилей, а также значений коэффициента загрузки системы. Несмотря на то, что рассматривалось определенное значение параметра распределения Парето времени обслуживания, представленный метод, основанный на элементах теории копул, можно применить аналогичным образом и для других значений параметра, а также обобщить на системы с большим количеством подсистем.

4 Об особенностях управления интенсивностью обслуживания в системах с разделением и параллельным обслуживанием

В настоящей Главе исследуется другой аспект производительности системы с разделением и параллельным обслуживанием, а именно строится модель стоимости функционирования системы, позволяющая управлять её режимом с целью оптимизации финансовых показателей. 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 его стоимости. Кроме того, с увеличением скорости обслуживания уменьшается время отклика системы. В результате, 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 мы хотим минимизировать. Таким образом, ставится задача стоимостной оптимизации управления.

Подобные постановки задач можно найти в монографии [186], посвященной

оптимальному дизайну СМО, в том числе оптимальному выбору скоростей поступления и обслуживания заявок (в предположении, что эти параметры управляемы) для различных систем и сетей массового обслуживания. Однако для fork-join систем такие задачи ранее не рассматривались.

Оптимальное значение интенсивности обслуживания определяется для классической системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком и двумя вариантами распределений времен обслуживания: экспоненциальным и распределением Парето. При поступлении в данную систему заявка разделяется на K одинаковых частей (подзаявок), количество которых соответствует числу подсистем. Каждая подсистема представляет собой систему с объемом накопителя бесконечной емкости и единственным прибором. Предполагается, что интенсивности обслуживания на всех имеющихся приборах идентичные. Заявка считается обслуженной после обслуживания всех составляющих её частей. Соответственно, время отклика системы (время пребывания заявки в системе) определяется максимумом из K случайных времен пребывания подзаявок в подсистемах. Результаты Главы 4 отражены в публикациях [17, 19, 100].

4.1 Математическая модель определения стоимости функционирования fork-join системы с экспоненциальным распределением времени обслуживания

Анализируется fork-join система с пуассоновским входящим потоком с интенсивностью $\lambda > 0$ и экспоненциальным распределением времен обслуживания на $K \geqslant 2$ однородных приборах с интенсивностью $\mu > 0$. Загрузка системы $\rho = \lambda/\mu < 1$.

Изложим единый математический подход, который будет далее применяться в работе.

Обозначим стоимость функционирования системы через S и введем функ-

цию $f(\rho)$, 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:

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

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

$$S = \frac{1}{\lambda}\left(f(\rho) + \frac{c_1}{\rho}\right). \tag{4.1}$$

Далее для нахождения оптимального значения уровня загрузки системы определим точку экстремума функции стоимости $S(\rho)$, а именно точку минимума. Для этого найдем производную последнего выражения и приравниваем ее к нулю

$$S'_\rho = \frac{1}{\lambda}\left(f'(\rho) - \frac{c_1}{\rho^2}\right) = 0,$$

откуда получим уравнение

$$f'(\rho)\rho^2 = c_1, \tag{4.2}$$

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

Рассмотрим базовый случай, когда $K = 1$, т. е. фактически систему $M|M|1$. Среднее время отклика в такой системе равно

$$E[R] = \frac{1}{\mu - \lambda} = \frac{1}{\lambda} \cdot \frac{1}{\mu/\lambda - 1}.$$

Тогда функция $f(\rho)$ с учетом того, что $\rho = \lambda/\mu$, определяется как

$$f(\rho) = \frac{1}{1/\rho - 1} = \frac{\rho}{1-\rho} = \frac{1}{1-\rho} - 1,$$

а производная этой функции имеет вид

$$f'(\rho) = \frac{1}{(1-\rho)^2}.$$

Теперь подставляем полученные выражения в уравнение (4.2)

$$\left(\frac{\rho}{1-\rho}\right)^2 = c_1,$$

$$\frac{\rho}{1-\rho} = \sqrt{c_1},$$

$$\rho = \sqrt{c_1} - \sqrt{c_1}\rho,$$

откуда получаем следующее значение для искомой оптимальной загрузки системы при $K = 1$

$$\rho_0 = \frac{\sqrt{c_1}}{1+\sqrt{c_1}}. \tag{4.3}$$

На следующем шаге можем вычислить оптимальную интенсивность обслуживания, а именно

$$\mu_0 = \frac{\lambda}{\rho_0} = \lambda\left(1 + \frac{1}{\sqrt{c_1}}\right) = \lambda + \frac{1}{\sqrt{c}}.$$

Аналогичная задача была решена в [186, § 1.1] для случая, когда в стоимости учитывается не среднее время отклика, а среднее время ожидания.

Подчеркнем, что поскольку мы получили уравнение (4.2) относительно загрузки ρ, то в дальнейшем будем искать только оптимальную загрузку, понимая, что ее можно при необходимости пересчитать в оптимальную интенсивность обслуживания.

4.2 Анализ fork-join СМО с двумя подсистемами типа $M|M|1$. Формула Нельсона–Тантави

Для начала разберем частный случай и определим оптимальное значение загрузки системы, когда число подсистем равно двум.

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 (4.1) 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[5]

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

Доказательство. Для среднего времени отклика при $K = 2$ известна точная формула, полученная в [163], а именно

$$E[R_2] = \frac{12 - \rho}{8}\frac{1}{\mu - \lambda} = \frac{1}{\lambda} \cdot \frac{12 - \rho}{8}\frac{\rho}{1 - \rho} = \frac{1}{\lambda} \cdot \frac{\rho(12 - \rho)}{8(1 - \rho)}.$$

Таким образом, имеем

$$f(\rho) = \frac{\rho(12 - \rho)}{8(1 - \rho)}.$$

Далее находим производную по ρ

$$f'(\rho) = \frac{\rho^2 - 2\rho + 12}{8(1 - \rho)^2}$$

[5] выражение для t_1 при $Q < 0$ определяется формулами (4.12) и (4.10)

и, подставляя выражение для $f'(\rho)$ в (4.2) можем записать следующее уравнение

$$\frac{\rho^2(\rho^2-2\rho+12)}{8(1-\rho)^2}=c_1,$$
$$\frac{\rho^2(\rho^2-2\rho+12)}{(1-\rho)^2}=8c_1.$$

Для удобства сделаем замену

$$c_2=8c_1$$

и после упрощения получаем уравнение четвертой степени

$$\rho^4-2\rho^3+(12-c_2)\rho^2+2c_2\rho-c_2=0, \tag{4.4}$$

которое и будем решать, поскольку, как известно, для уравнения четвертой степени существует аналитическое решение в радикалах. Для этого воспользуемся методом Феррари.

Введем следующие обозначения

$$A=-2,\quad B=12-c_2,\quad C=c_2,\quad D=-c_2,$$

тогда получим

$$\rho^4+A\rho^3+B\rho^2+C\rho+D=0.$$

С помощью замены $\rho=y-A/4$ сведем уравнение четвертой степени (4.4) к каноническому виду

$$y^4+A_1y^2+B_1y+C_1=0, \tag{4.5}$$

где

$$A_1=B-\frac{3A^2}{8}=-c_2+\frac{21}{2},$$
$$B_1=\frac{A^3}{8}-\frac{AB}{2}+C=c_2+11,$$
$$C_1=-\frac{3A^4}{256}+\frac{A^2B}{16}-\frac{AC}{4}+D=-\frac{1}{4}c_2+\frac{45}{16}.$$

Далее согласно методу Феррари [35, 36, 185] необходимо найти одно частное действительное решение следующего кубического уравнения

$$A_2 t^3 + B_2 t^2 + C_2 t + D_2 = 0, \tag{4.6}$$

где

$$A_2 = 2, \quad B_2 = -A_1 = c_2 - \frac{21}{2},$$
$$C_2 = -2C_1 = \frac{1}{2}c_2 - \frac{45}{8}, \quad D_2 = A_1 C_1 - \frac{B_1^2}{4} = -\frac{175}{16}c_2 - \frac{23}{32}.$$

С помощью замены $t = z - B_2/(3A_2)$ приводим уравнение (4.6) к каноническому виду уравнения третьей степени

$$z^3 + A_3 z + B_3 = 0, \tag{4.7}$$

где

$$A_3 = \frac{3A_2C_2 - B_2^2}{3A_2^2} = -\frac{1}{12}c_2^2 + 2c_2 - 12,$$
$$B_3 = \frac{2B_2^3 - 9A_2B_2C_2 + 27A_2^2D_2}{27A_2^3} = \frac{1}{108}c_2^3 - \frac{1}{3}c_2^2 - \frac{3}{2}c_2 - 16.$$

Вещественный корень уравнения (4.7) согласно методу Кардано [35, 36, 185] определяется следующим образом

$$z_1 = \left(-\frac{B_3}{2} + \sqrt{Q}\right)^{\frac{1}{3}} + \left(-\frac{B_3}{2} - \sqrt{Q}\right)^{\frac{1}{3}}, \quad Q > 0, \tag{4.8}$$

где

$$Q = \left(\frac{A_3}{3}\right)^3 + \left(\frac{B_2}{2}\right)^2 = -\frac{11}{432}c_2^4 + \frac{11}{12}c_2^3 - \frac{55}{16}c_2^2 + 44c_2. \tag{4.9}$$

При этом не забываем, что должно выполняться условие

$$\left(-\frac{B_3}{2} + \sqrt{Q}\right)^{\frac{1}{3}} \cdot \left(-\frac{B_3}{2} - \sqrt{Q}\right)^{\frac{1}{3}} = -\frac{A_3}{3}.$$

Отметим, что для случая $Q < 0$ для z_1 после преобразований в конечном счете будет справедлива следующая формула

$$z_1 = 2\sqrt{-\frac{A_3}{3}}\cos\frac{w}{3}, \tag{4.10}$$

где

$$w = \begin{cases} \operatorname{arctg}\left(\frac{-2\sqrt{-Q}}{B_3}\right), & \text{если } B_3 < 0 \\ \operatorname{arctg}\left(\frac{-2\sqrt{-Q}}{B_3}\right) + \pi, & \text{если } B_3 > 0 \\ \frac{\pi}{2}, & \text{если } B_3 = 0, \end{cases}$$

а для случая $Q = 0$ имеем

$$z_1 = 2\left(-\frac{B_3}{2}\right)^{\frac{1}{3}}. \tag{4.11}$$

Выражение для Q из (4.9) преобразуется к виду

$$-\frac{11}{432}c_2\left(c_2^3 - 36c_2^2 + 135c_2 - 1728\right)$$

и меняет свой знак в зависимости от значения c_2 (хотя c_2 должно быть положительным исходя из своего физического смысла), поэтому явно укажем области знакопостоянства выражения для Q, нули которого можно опять же получить с помощью метода Кардано, а именно

$$\begin{aligned} &Q > 0, \quad \text{если } c_2 \in (0; \tilde{c}_2)\ , \\ &Q < 0, \quad \text{если } c_2 \in (-\infty; 0) \cup (\tilde{c}_2; +\infty)\ , \end{aligned}$$

где $\tilde{c}_2 = (3\cdot 11^{\frac{1}{3}} + 3\cdot 11^{\frac{2}{3}} + 12) \approx 33.51$. График для $Q(c_2)$ представлен на рисунке 84.

Соответственно, получаем

$$t_1 = z_1 - \frac{B_2}{3A_2}, \tag{4.12}$$

где z_1 в зависимости от значения Q определяется выражениями (4.8), (4.10) или (4.11). Частное решение t_1 позволяет представить каноническое уравнение четвертой степени (4.5) в виде произведения двух квадратных трехчленов

$$\left(y^2 - y\sqrt{2t_1 - A_1} + \frac{B_1}{2\sqrt{2t_1 - A_1}}\right) \cdot \left(y^2 + y\sqrt{2t_1 - A_1} - \frac{B_1}{2\sqrt{2t_1 - A_1}}\right) = 0.$$

Решение одного из квадратных уравнений (с учетом обратной замены) будет являться искомым решением исходного уравнения четвертой степени (4.4).

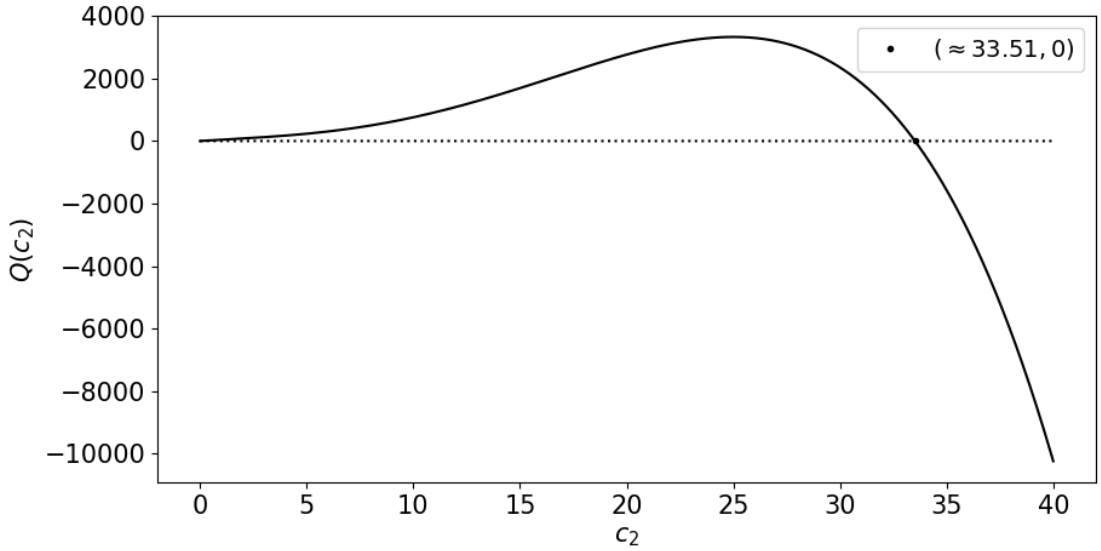


Рис. 84: График зависимости $Q(c_2)$.

Дискриминант первого уравнения будет определяться выражением

$$Dis_1 = -2t_1 - A_1 - \frac{B_1}{2\sqrt{2t_1 - A_1}},$$

а дискриминант второго уравнения равен

$$Dis_2 = -2t_1 - A_1 + \frac{B_1}{2\sqrt{2t_1 - A_1}}.$$

На рисунках 85 и 86 представлены графики зависимости значения выражений для дискриминантов от значения c_2. Как видно из графиков, дискриминант первого уравнения принимает отрицательные значения до $c_2 = \tilde{c}_2 \approx 33.51$, соответственно, его корни

$$y_1 = \frac{-\sqrt{2t_1 - A_1} + \sqrt{Dis_1}}{2}, \quad y_2 = \frac{-\sqrt{2t_1 - A_1} - \sqrt{Dis_1}}{2},$$

имеют смысл только для значений $c_2 > \tilde{c}_2 \approx 33.51$, причем сразу очевидно исходя из вида аналитического соотношения, что $y_2 < 0$ на всей области определения. На рисунках 87 и 88 представлены графики зависимостей корней y_1 и y_2 от c_2, которые наглядно подтверждают, что y_1 возрастает $(y_2' > 0)$ и $y_1 \to -0.5$ при $c_2 \to +\infty$, а $y_2 < -0.5$ для рассматриваемых значений c_2, поскольку убывает $(y_2' < 0)$ и $y_{2\,\min} = y_2(\tilde{c}_2) \approx -2.72$.

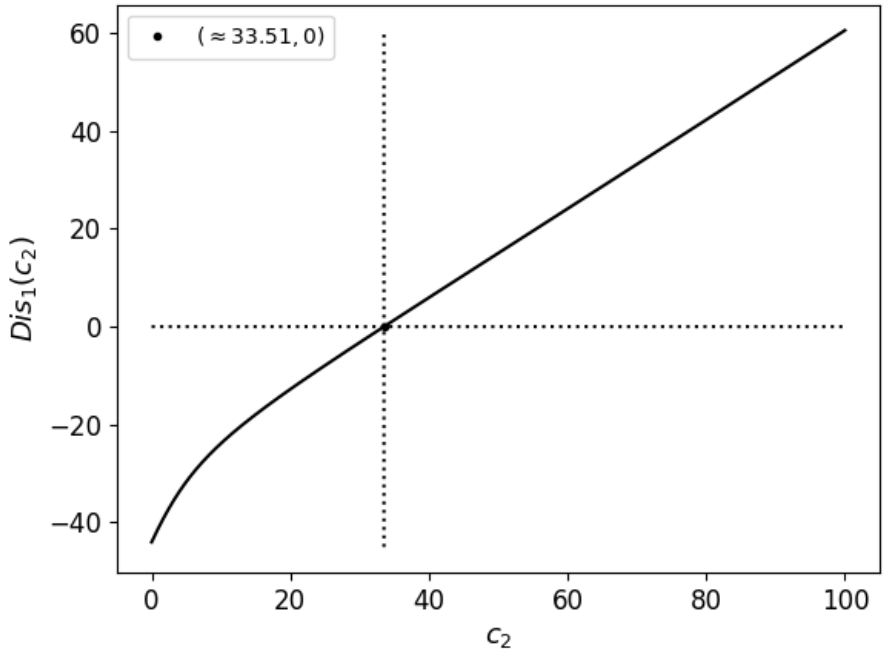


Рис. 85: График зависимости значения дискриминанта Dis_1 от значения c_2.

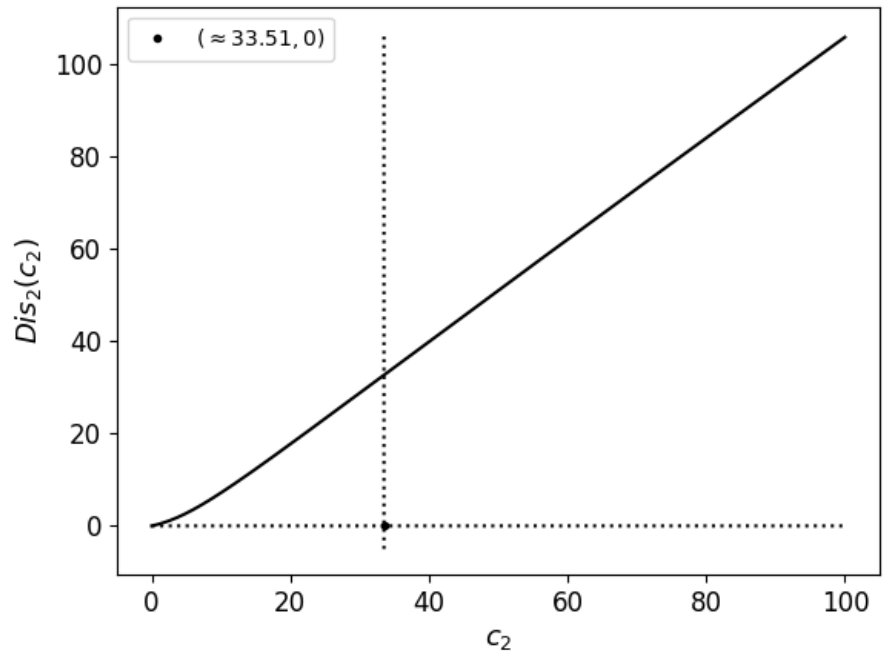


Рис. 86: График зависимости значения дискриминанта Dis_2 от значения c_2.

Корни второго уравнения имеют вид

$$y_3 = \frac{-\sqrt{2t_1 - A_1} + \sqrt{Dis_2}}{2}, \quad y_4 = \frac{-\sqrt{2t_1 - A_1} - \sqrt{Dis_2}}{2},$$

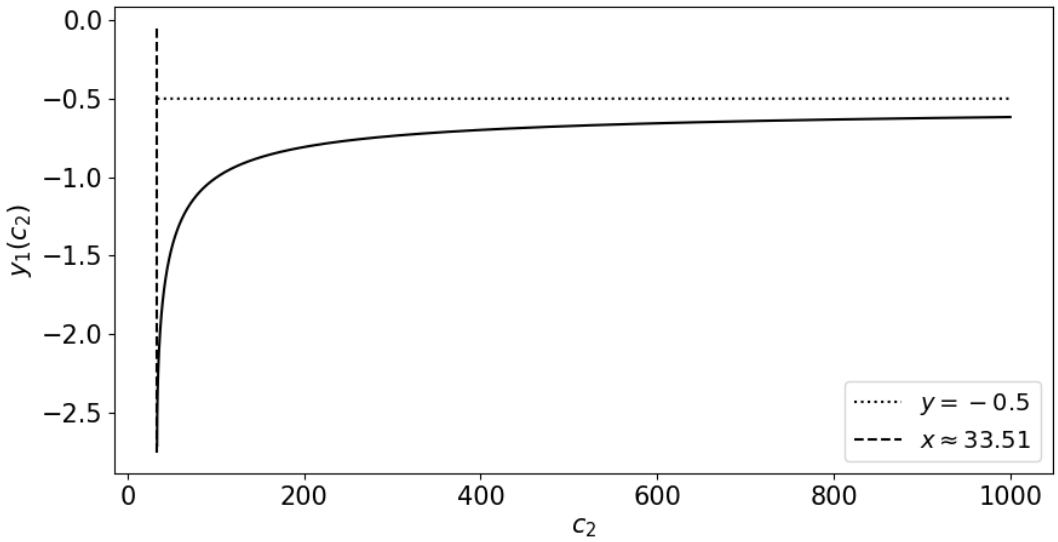


Рис. 87: График зависимости значения выражения для y_1 от значения c_2.

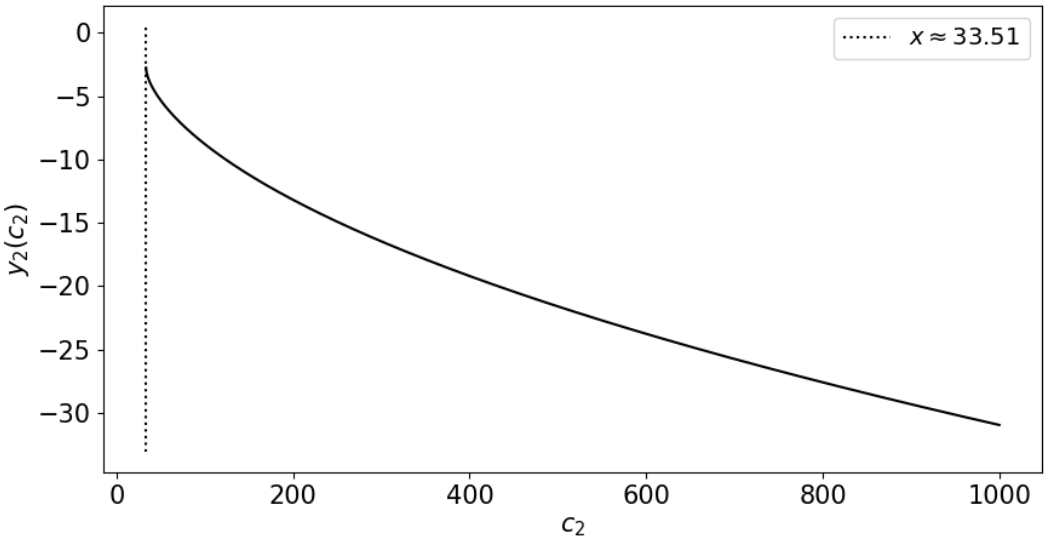


Рис. 88: График зависимости значения выражения для y_2 от значения c_2.

На рисунках 89 и 90 представлены графики зависимостей корней y_3 и y_4 от $c_2 \geqslant 0$. Значения y_3 и y_4 в нуле совпадают и равны -0.5, при этом $y_3 \to 0.5$, $c_2 \to +\infty$, а y_4 убывает и все его значения, очевидно, отрицательные, исходя из аналитического вида выражения для y_4, и при этом меньше $y_{4\min} = y_4(0) = -0.5$, $c_2 \in [0, +\infty)$. Поэтому можем сделать вывод и записать итоговое решение, как

$$y_0 = y_3 = \frac{-\sqrt{2t_1 - A_1} + \sqrt{Dis_2}}{2},$$

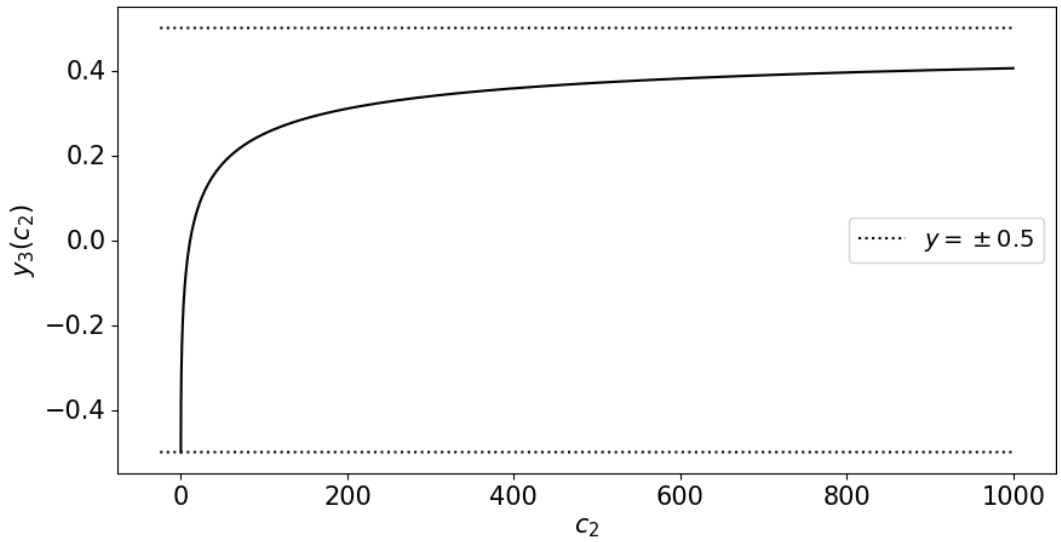


Рис. 89: График зависимости значения выражения для y_3 от значения c_2.

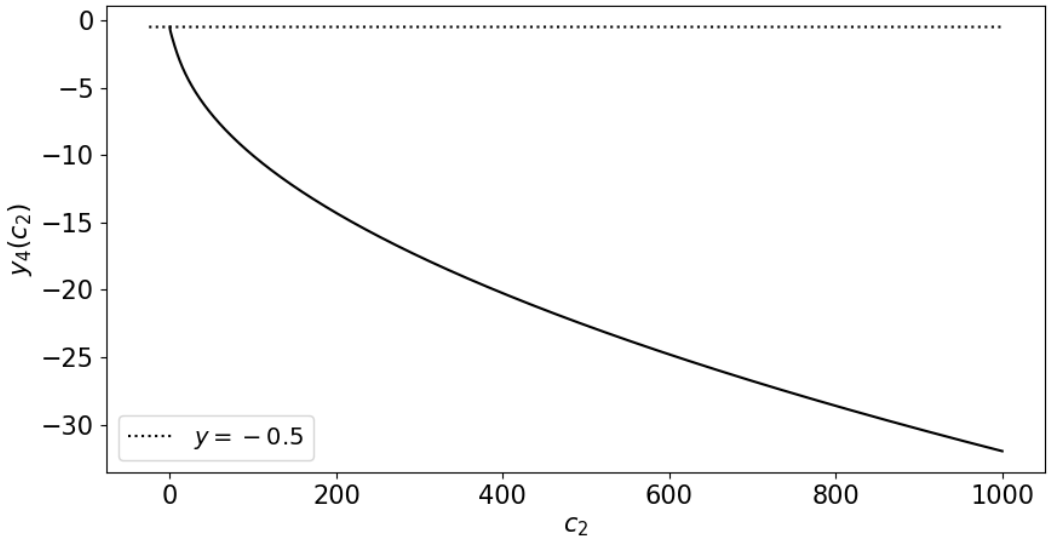


Рис. 90: График зависимости значения выражения для y_4 от значения c_2.

и, соответственно,

$$\rho_0 = y_0 - \frac{A}{4} = y_0 + \frac{1}{2}.$$

Далее при подстановке в полученное выражение для y_0 соответствующих параметров и некоторого упрощения получаем необходимые выражения из формулировки теоремы. □

4.3 Анализ fork-join СМО с $K > 2$ подсистемами типа $M|M|1$. Формула Нельсона–Тантави

Выведем уравнение для определения оптимального значения загрузки системы в общем случае, когда $K > 2$.

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 (4.1) и предположения о том, что формула Нельсона–Тантави (4.14) для определения среднего времени отклика системы является точной, оптимальное значение загрузки системы ρ_0 определяется решением уравнения*

$$\begin{gathered} 8(H-1)\rho^5 + (60-71H)\rho^4 + (118H-96)\rho^3 + \\ +11(c_2-12H)\rho^2 - 22c_2\rho + 11c_2 = 0, \end{gathered} \tag{4.13}$$

где $c_2 = 8c_1$, $H = H_K/H_2$.

Доказательство. Для математического ожидания времени отклика fork-join системы приближенная формула Нельсона–Тантави, которая считается одной из наиболее точных среди известных, имеет вид [163]:

$$E[R_K] \approx \left[\frac{H_K}{H_2} + \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right)\rho\right]\frac{12-\rho}{8}\frac{1}{\mu-\lambda}, \tag{4.14}$$

где $H_K = \sum_{i=0}^{K} 1/i$ — это частичная сумма гармонического ряда.

Решаем задачу для приближения Нельсона–Тантави, в предположении, что формула (4.14) является точной.

Введем следующие обозначения:

$$H = \frac{H_K}{H_2}, \quad M = \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right) = \frac{4}{11}(1-H).$$

Тогда

$$E[R_K] = \frac{1}{\lambda} \cdot (H + M\rho)\frac{12-\rho}{8}\frac{\rho}{1-\rho},$$

следовательно,

$$f(\rho) = (H + M\rho)\frac{12-\rho}{8}\frac{\rho}{1-\rho},$$

а

$$f'(\rho) = M\frac{12\rho - \rho^2}{8(1-\rho)} + (H + M\rho)\frac{\rho^2 - 2\rho + 12}{8(1-\rho)^2}.$$

После преобразований для $f'(\rho)$ получаем

$$f'(\rho) = \frac{1}{8(1-\rho)^2}(2M\rho^3 + (H - 15M)\rho^2 + (24M - 2H)\rho + 12H). \qquad (4.15)$$

Затем подставляем полученное выражение в (4.2)

$$\frac{\rho^2}{8(1-\rho)^2}(2M\rho^3 + (H - 15M)\rho^2 + (24M - 2H)\rho + 12H) = c_1.$$

Также для удобства вводим обозначение $c_2 = 8c_1$ и в результате получаем уравнение пятой степени

$$2M\rho^5 + (H - 15M)\rho^4 + (24M - 2H)\rho^3 + (12H - c_2)\rho^2 + 2c_2\rho - c_2 = 0.$$

И если учесть, что $M = 4(1 - H)/11$, то оно сводится к виду

$$-\frac{1}{11}\big(8(H-1)\rho^5 + (60 - 71H)\rho^4 +$$

$$+(118H - 96)\rho^3 + 11(c_2 - 12H)\rho^2 - 22c_2\rho + 11c_2\big) = 0$$

и, соответственно, имеем

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

Полученное уравнение можно решить численно и определить оптимальное значение ρ_0 и, соответственно искомое значение μ_0. □

На рисунке 91 представлены графики зависимости поведения оптимального значения загрузки $\rho = \rho_0$, являющегося решением уравнения (4.13), в зависимости от значения параметра c_1 для различного числа подсистем K fork-join СМО. На начальном этапе наблюдается довольно стремительный рост требуемого уровня загрузки с увеличением цены единицы ресурса и, соответственно, производительности системы в целом. Причем даже для числа подсистем

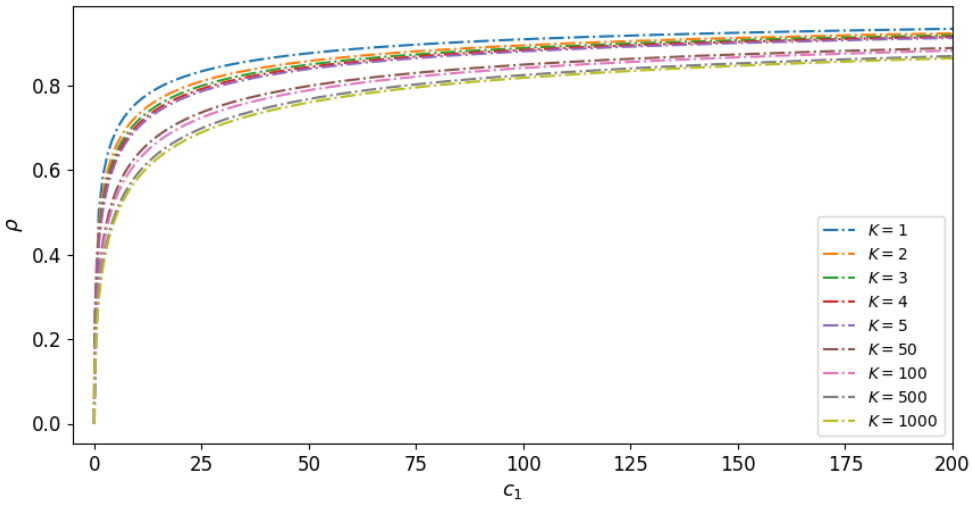


Рис. 91: График зависимости значения оптимального уровня загрузки системы ρ (решение уравнения (4.13)) от параметра c_1 для различного числа подсистем K.

$K = 1000$, т. е. иными словами, при условии разделения задачи и обработки ее подзадач на довольно большом количестве устройств, уже при $c_1 > 5.5$ необходимый уровень загрузки $\rho > 0.5$. Одновременно с этим при дальнейшем росте параметра c_1 наблюдается довольно медленный рост требуемой загрузки, например, при $c_1 \approx 158$, что почти в 29 раз больше $c_1 = 5.5$, значение $\rho_0 = 0.85$, т. е. все еще не превышает 90%. Кроме того, исходя из вида графиков, наблюдается эффект того, что с ростом K оптимальный уровень загрузки на систему снижается, что было ожидаемо и вполне естественно.

На рисунке 92 представлен график зависимости поведения оптимального значения загрузки системы от оптимального значения ρ при $K = 1$, согласно формуле (4.3). График позволяет сравнить уровень оптимальной загрузки для различных значений $K \geqslant 2$ с уровнем оптимальной загрузки в случае $K = 1$. Как видно, с увеличением числа подсистем уровень требуемой загрузки для оптимальной работы системы падает, т. е. линия все сильнее прогибается под прямой $\rho = \rho_{K=1}$ и становится всё более выпуклой (ниже любой хорды, соединяющей любые две точки на графике, выбранные на рассматриваемом

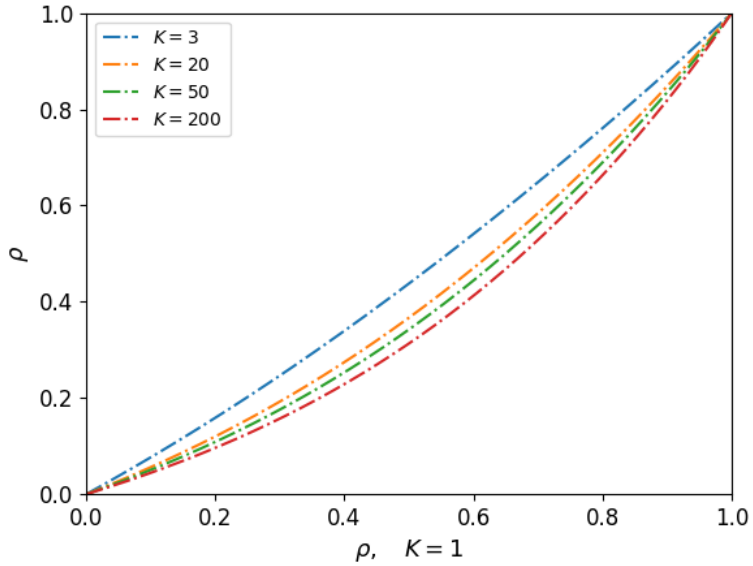


Рис. 92: График зависимости оптимального значения загрузки системы ρ в зависимости от оптимального значения загрузки ρ при $K = 1$.

интервале).

4.4 Анализ fork-join СМО с $K > 2$ подсистемами типа $M|M|1$. Обобщение формулы Нельсона–Тантави

Рассмотрим обобщение формулы Нельсона–Тантави из Главы 2, которое дает лучшее приближение для среднего времени отклика. Улучшение достигается за счет поправки к выражению из (4.14), которое теперь обозначим через $E[R_K]_{NT}$.

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 (4.1) и предположения о том, что формула (4.16) для определения среднего времени отклика системы является точной, оптимальное значение загрузки системы ρ_0 опре-*

деляется 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

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

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

Доказательство. Аппроксимация для среднего времени отклика системы с разделением и параллельным обслуживанием заявок имеет вид

$$E[R_K]=\frac{\rho}{\mu-\lambda}\left(\frac{H_K}{H_2}-1\right)\cdot\left(Q_1-Q_2\left(\frac{H_K}{H_2}-1\right)+Q_3\rho\right)+E[R_K]_{NT},\quad (4.16)$$

где

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

Определим выражение для $f(\rho)$ в данном случае, для этого вынесем $1/\lambda$

$$E[R_K]=\frac{1}{\lambda}\left[-\frac{\rho^2}{1-\rho}\left(1-\frac{H_K}{H_2}\right)\cdot\left(Q_1+Q_2\left(1-\frac{H_K}{H_2}\right)+Q_3\rho\right)+\lambda E[R_K]_{NT}\right],$$

в итоге получим

$$f(\rho)=-\frac{\rho^2}{1-\rho}\left(1-\frac{H_K}{H_2}\right)\cdot\left(Q_1+Q_2\left(1-\frac{H_K}{H_2}\right)+Q_3\rho\right)+\lambda E[R_K]_{NT},$$

где $\lambda E[R_K]_{NT}$ было рассчитано ранее и имеет вид

$$\lambda E[R_K]_{NT}=f_2(\rho)=(H+M\rho)\frac{12-\rho}{8}\frac{\rho}{1-\rho}.$$

Таким образом, с учетом предложенной ранее замены $H=H_K/H_2$ имеем

$$f(\rho)=-\frac{\rho^2}{1-\rho}\cdot(1-H)(Q_1+(1-H)Q_2+Q_3\rho)+f_2(\rho)=f_1(\rho)+f_2(\rho).$$

Далее берем производную от полученного выражения

$$f'(\rho)=f_1'(\rho)+f_2'(\rho),$$

где

$$f_1'(\rho) = \frac{\rho(1-H)}{(1-\rho)^2}\Big(2Q_3\rho^2 - \rho(3Q_3 - (1-H)Q_2 - Q_1) - (2Q_1 + 2(1-H)Q_2)\Big),$$

$$f_2'(\rho) = \frac{1}{8(1-\rho)^2}\Big(2M\rho^3 + (H-15M)\rho^2 + (24M-2H)\rho + 12H\Big).$$

Соответственно, после подстановки полученных выражений в (4.2) получаем, что

$$c_1 = \rho^2 f'(\rho) = \rho^2(f_1'(\rho) + f_2'(\rho)). \tag{4.17}$$

Уравнение (4.17), как и в случае с формулой Нельсона–Тантави из предыдущего раздела, является уравнением пятой степени, которое может быть решено численно, после чего будет найдено оптимальное значение μ_0. □

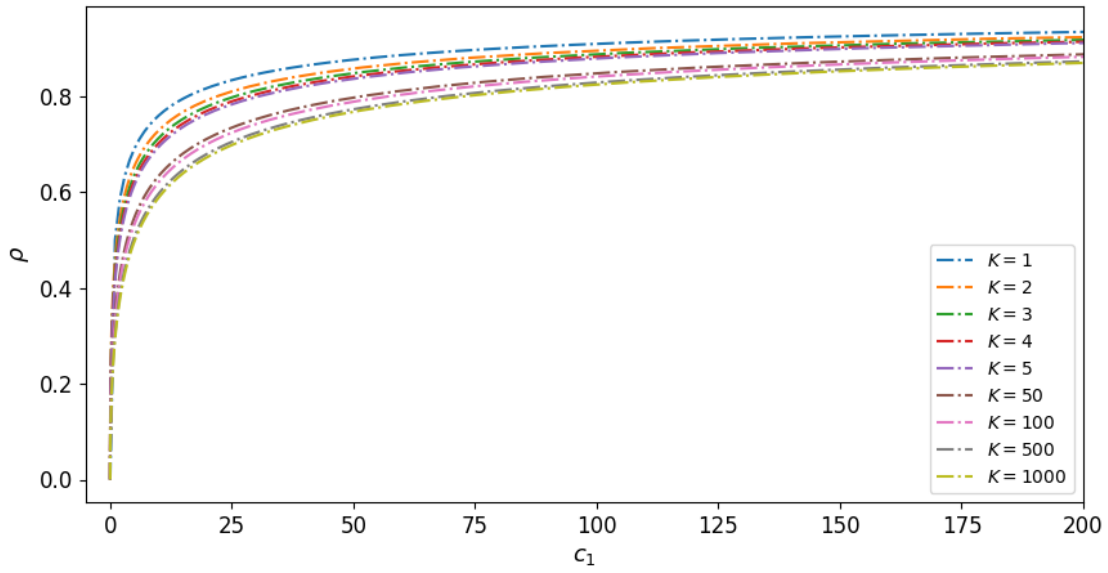


Рис. 93: График зависимости значения оптимального уровня загрузки системы ρ (решение уравнения (4.17)) от параметра c_1 для различного числа подсистем K.

На рисунке 93 представлены графики оптимального значения загрузки $\rho = \rho_0$, являющегося решением уравнения (4.17). Поведение графиков в целом аналогично случаю формулы Нельсона–Тантави (4.14). Чтобы более детально

проанализировать разницу между полученными результатами, проведем сравнение поведения оптимального решения для частных случаев числа подсистем $K = 20$ и $K = 200$.

На рисунках 94 представлены графики зависимости значения оптимального уровня загрузки системы $\rho = \rho_0$ от параметра c_1 для случаев формулы Нельсона–Тантави для среднего времени отклика системы $E[R_K]_{NT}$ из (4.14) и уравнения (4.13), а также обобщения формулы Нельсона–Тантави для среднего времени отклика системы $E[R_K]$ из (4.16) и уравнения (4.17) при $K = 20$ (рис. 94а и 94б) и $K = 200$ (рис. 94в и 94г), в том числе и в масштабе (рис. 94б и 94г).

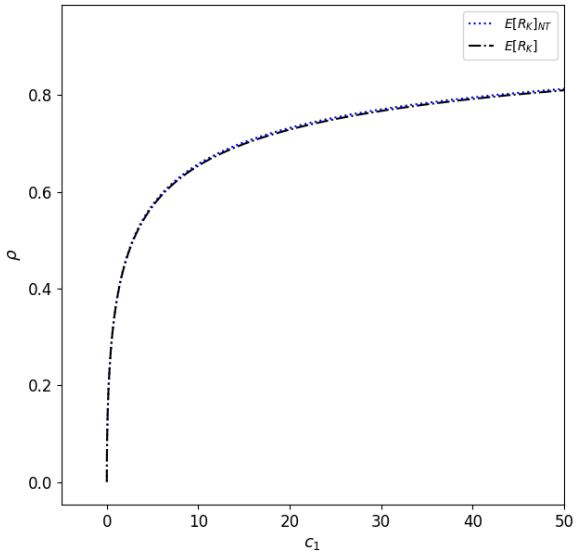
На рисунках 94а, 94б для $K = 20$ видно, что для случая формулы Нельсона–Тантави (4.14) для оценки среднего времени отклика системы, оптимальное значение загрузки ρ превышает оптимальное значение, рассчитанное по обобщенной формуле (4.16). Однако для графиков 94в, 94г при $K = 200$, можно наблюдать обратную ситуацию.

Для того, чтобы обстоятельно разобраться в поведении оптимальных решений в обоих случаях, далее проанализируем их асимптотику.

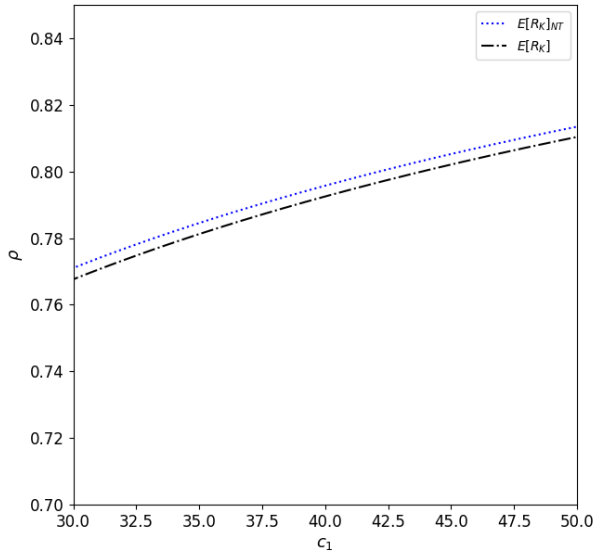
4.5 Асимптотика поведения оптимального решения для системы с разделением и параллельным обслуживанием с подсистемами типа $M|M|1$

Рассмотрим уравнение (4.2) и определим поведение его решения при стремлении $c_1 \to 0$ ($\rho \to 0$) и $c_1 \to +\infty$ ($\rho \to 1$) в общем виде.

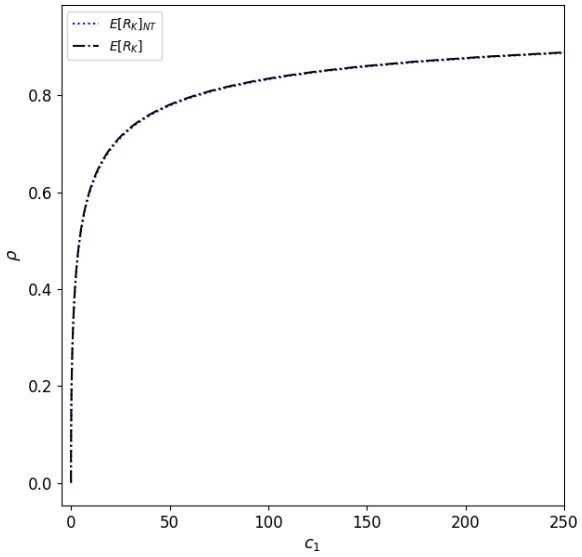
Теорема 4.4. *Для системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром $\lambda > 0$ и $K > 2$ подсистемами типа $M|M|1$ с показательным распределением времени обслуживания с параметром $\mu > 0$ ($\lambda < \mu$) в условиях стоимостной модели* (4.1) *асимпто-*



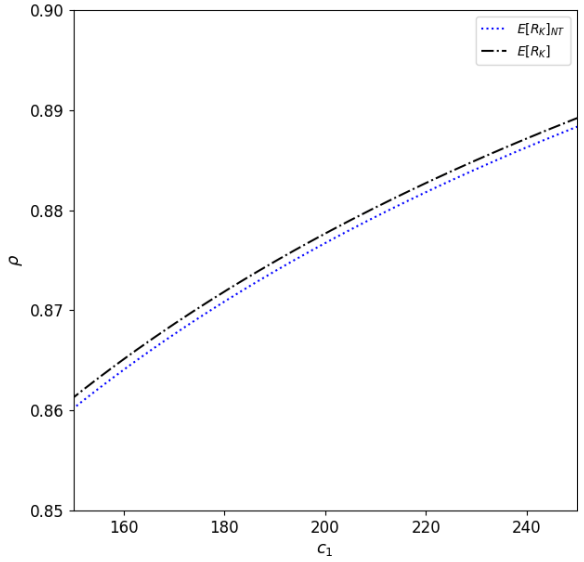
а)



б)



в)



г)

Рис. 94: Графики оптимального значения загрузки $\rho = \rho_0$, являющегося решением уравнения (4.17) при $K = 20$ (а и б) и $K = 200$ (в и г).

тическое поведение оптимального значения загрузки системы ρ имеет вид

$$\rho \sim \sqrt{\frac{c_1}{f'(0)}}, \quad c_1 \to 0,$$

$$\rho = 1 - \sqrt{\frac{L}{c_1}} \cdot (1 + o(1)), \quad c_1 \to \infty,$$

где

$$L = \lim_{\rho \to 1} (1-\rho)^2 f'(\rho).$$

Доказательство. Для начала проанализируем случай $\rho \to 0$, соответственно, $c_1 \to 0$. Тогда имеем

$$\rho^2 f'(\rho) \sim \rho^2 f'(0),$$

далее подставляем полученное в (4.2), т. е.

$$\rho^2 f'(\rho) = c_1,$$

$$\rho^2 f'(0) \sim c_1, \quad \rho \to 0, c_1 \to 0,$$

откуда следует, что

$$\rho \sim \sqrt{\frac{c_1}{f'(0)}}, \quad c_1 \to 0. \tag{4.18}$$

Теперь проанализируем случай $\rho \to 1$, соответственно, $c_1 \to +\infty$. В общем случае имеем

$$\rho^2 f'(\rho) \sim f'(\rho).$$

Таким образом, если существует такое число $L \in (0, +\infty)$, что

$$L = \lim_{\rho \to 1} (1-\rho)^2 f'(\rho),$$

то

$$f'(\rho) \sim \frac{L}{(1-\rho)^2},$$

поэтому при подстановке полученного выражения для $f'(\rho)$ с L в (4.2) и учетом асимптотики имеем следующее

$$\rho^2 f'(\rho) = c_1,$$

$$f'(\rho) \sim c_1, \quad \rho \to 1, c_1 \to +\infty,$$
$$\frac{L}{(1-\rho)^2} \sim c_1, \quad \rho \to 1, c_1 \to +\infty,$$
$$(1-\rho)^2 \sim \frac{L}{c_1}, \quad \rho \to 1, c_1 \to +\infty,$$

поэтому

$$\rho = 1 - \sqrt{\frac{L}{c_1}} \cdot (1 + o(1)), \quad c_1 \to \infty. \tag{4.19}$$

□

Далее предметно рассмотрим имеющиеся модели, а именно случаи формулы Нельсона–Тантави и его обобщение, и определим конкретные выражения для полученных эквивалентностей из общего случая.

Теорема 4.5. *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 (4.1) и предположения о том, что формула Нельсона–Тантави (4.14) для определения среднего времени отклика системы является точной, асимптотическое поведение оптимального значения загрузки системы ρ имеет вид*

$$\rho \sim \sqrt{\frac{2}{3} \cdot \frac{H_2}{H_K} \cdot c_1}, \quad c_1 \to 0,$$

$$\rho = 1 - \sqrt{\frac{1}{c_1}\left(\frac{7}{8} \cdot \frac{H_K}{H_2} + \frac{1}{2}\right)} \cdot (1 + o(1)), \quad c_1 \to +\infty.$$

Доказательство. Итак, для формулы Нельсона–Тантави при $K \geqslant 2$ с учетом выражения (4.15) справедливо следующее

$$f'(0) = \frac{3}{2} \cdot \frac{H_K}{H_2},$$

поэтому при подстановке в (4.18) для $c_1 \to 0$ ($\rho \to 0$) получим

$$\rho \sim \sqrt{\frac{2}{3} \cdot \frac{H_2}{H_K} \cdot c_1}, \quad c_1 \to 0. \tag{4.20}$$

Теперь определим асимптотику решения при $c_1 \to +\infty$ ($\rho \to 1$). Для этого

найдем значение L

$$\begin{gathered} L = \lim_{\rho\to 1}(1-\rho)^2 f'(\rho) = \\ = \lim_{\rho\to 1} \frac{2M\rho^3 + (H-15M)\rho^2 + (24M-2H)\rho + 12H}{8} = \\ = \frac{2M + H - 15M + 24M - 2H + 12H}{8} = \\ = \frac{11H+11M}{8} = \frac{7H+4}{8} = \frac{7}{8}\cdot\frac{H_K}{H_2} + \frac{1}{2}. \end{gathered} \tag{4.21}$$

Таким образом, окончательно получаем

$$\rho = 1 - \sqrt{\frac{1}{c_1}\left(\frac{7}{8}\cdot\frac{H_K}{H_2} + \frac{1}{2}\right)}\cdot(1+o(1)).$$

□

Теперь проанализируем обобщение формулы Нельсона–Тантави.

Теорема 4.6. *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 (4.1) и предположения о том, что формула (4.16) для определения среднего времени отклика системы является точной, асимптотическое поведение оптимального значения загрузки системы ρ имеет вид*

$$\rho \sim \sqrt{\frac{2}{3}\cdot\frac{H_2}{H_K}\cdot c_1}, \quad c_1 \to 0,$$

$$\rho = 1 - \sqrt{\frac{L}{c_1}}\cdot(1+o(1)), \quad c_1 \to +\infty,$$

где

$$L = \left(\frac{H_K}{H_2} - 1\right)\cdot\left[Q_1 - \left(\frac{H_K}{H_2} - 1\right)Q_2 + Q_3\right] + \frac{7}{8}\cdot\frac{H_K}{H_2} + \frac{1}{2},$$
$$Q_1 \approx 0.087197, \quad Q_2 \approx 0.070236, \quad Q_3 \approx 0.09638.$$

Доказательство. Рассмотрим уравнение (4.17) и определим поведение его решения при стремлении $c_1 \to 0$ ($\rho \to 0$) и $c_1 \to +\infty$ ($\rho \to 0$). Проанализируем случай $\rho \to 0$, соответственно, $c_1 \to 0$.

Для формулы (4.16) при $K \geqslant 2$ справедливо

$$f'(0) = f_1'(0) + f_2'(0) = 0 + \frac{3}{2} \cdot \frac{H_K}{H_2} = \frac{3}{2} \cdot \frac{H_K}{H_2},$$

поэтому, как и ранее

$$\rho \sim \sqrt{\frac{2}{3} \cdot \frac{H_2}{H_K} \cdot c_1}, \quad c_1 \to 0.$$

Теперь проанализируем случай $\rho \to 1$ ($c_1 \to \infty$). Для формулы (4.16) при $K \geqslant 2$ получим

$$L = \lim_{\rho \to 1}(1-\rho)^2 f'(\rho) = \lim_{\rho \to 1}(1-\rho)^2 (f_1'(\rho) + f_2'(\rho)) =$$

$$= \lim_{\rho \to 1}(1-\rho)^2 f_1'(\rho) + \lim_{\rho \to 1}(1-\rho)^2 f_2'(\rho) = L_1 + L_2.$$

Фактически значение L_2 было вычислено ранее и соответствует значению L из (4.21)

$$L_2 = \frac{7}{8} \cdot \frac{H_K}{H_2} + \frac{1}{2}.$$

Теперь рассчитаем L_1

$$L_1 = \lim_{\rho \to 1}(1-\rho)^2 f_1'(\rho) =$$

$$= \lim_{\rho \to 1} \rho(1-H)\Big(2Q_3\rho^2 - \rho(3Q_3 - (1-H)Q_2 - Q_1) - (2Q_1 + 2(1-H)Q_2)\Big) =$$

$$= (H-1)(Q_1 - (H-1)Q_2 + Q_3) = \left(\frac{H_K}{H_2} - 1\right) \cdot \left[Q_1 - \left(\frac{H_K}{H_2} - 1\right) Q_2 + Q_3\right] \approx$$

$$\approx \left(\frac{H_K}{H_2} - 1\right) \cdot \left[0.183577 - 0.070236 \left(\frac{H_K}{H_2} - 1\right)\right].$$

таким образом, получаем, что

$$\rho = 1 - \sqrt{\frac{L}{c_1}} \cdot (1 + o(1)),$$

где

$$L = \left(\frac{H_K}{H_2} - 1\right) \cdot \left[Q_1 - \left(\frac{H_K}{H_2} - 1\right) Q_2 + Q_3\right] + \frac{7}{8} \cdot \frac{H_K}{H_2} + \frac{1}{2}. \tag{4.22}$$

□

На рисунке 95 представлен график зависимости значения L от числа подсистем для случая формулы Нельсона–Тантави (4.21) и случая обобщающей формулы (4.22). После значения $K = 126$ происходит пересечение графиков, и если сперва значение L_{NT} превышало значение L для обобщающей формулы, то для значений $K \geqslant 127$ ситуация поменялась с точностью до наоборот. Что в том числе и подтверждается графиками, представленными ранее, для

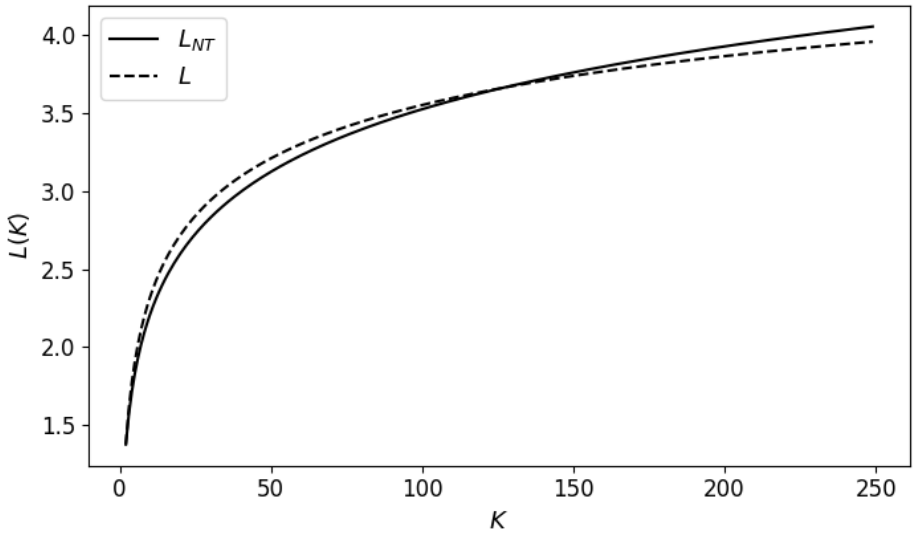


Рис. 95: График зависимости значения L от K для случая формулы Нельсона–Тантави (4.14) и случая обобщающей формулы (4.16).

оптимального значения загрузки в зависимости от цены c_1 на рисунках 94 для $K = 20$ и $K = 200$.

4.6 Математическая модель определения стоимости функционирования fork-join системы с распределением Парето времени обслуживания

Анализируется fork-join система с пуассоновским входящим потоком с интенсивностью $\lambda > 0$ и распределением Парето времен обслуживания на $K \geqslant 2$ однородных приборах со скоростью $\mu > 0$. А именно, полагаем, что при $\mu = 1$

время обслуживания имеет базовое распределение Парето F_{Pa} случайной величины η с $E[\eta] = 1$, а при произвольном $\mu > 0$ распределение случайной величины η/μ, т.е. время обслуживания сокращается в μ раз по сравнению с базовым случаем. Загрузка системы $\rho = \lambda/\mu < 1$.

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

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

где c — это цена единицы скорости обслуживания (единицы ресурса). Соответственно, с учетом введенной функции $g(\rho)$ можем переписать данное выражение следующим образом

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

Пусть $c_1 = c\lambda^2/c_0$, тогда

$$S = \frac{1}{\lambda}\left(\rho g(\rho) + \frac{c_1}{\rho}\right). \tag{4.23}$$

Далее для нахождения оптимального значения уровня загрузки системы определим точку экстремума функции стоимости $S(\rho)$, а именно точку минимума. Для этого найдем производную последнего выражения и приравниваем ее к нулю

$$S'_\rho = \frac{1}{\lambda}\left(\rho g'(\rho) + g(\rho) - \frac{c_1}{\rho^2}\right) = 0,$$

откуда получим уравнение

$$(\rho g'(\rho) + g(\rho))\rho^2 = c_1, \tag{4.24}$$

решив которое, найдем оптимальное значение ρ_0 и, соответственно, оптимальное значение скорости обслуживания $\mu_0 = \lambda/\rho_0$. Подчеркнем, что поскольку мы получили уравнение (4.24) относительно загрузки ρ, то в дальнейшем будем искать только оптимальную загрузку, понимая, что ее можно при необходимости пересчитать в оптимальную скорость обслуживания.

4.7 Анализ fork-join СМО с $K \geqslant 2$ подсистемами типа $M|Pa|1$

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 Главе 3 была получена приближенная формула (в области $4 \leq \alpha \leq 10$, $2 \leq K \leq 20$), которая и будет использоваться в дальнейших вычислениях.

При этом стоит напомнить предположение о виде функции распределения времени обслуживания подзаявки

$$F_{Pa}(x) = 1 - \left(\frac{\alpha - 1}{\alpha} \cdot \frac{1}{x}\right)^{\alpha}, \quad x \geq \frac{\alpha - 1}{\alpha}, \quad \alpha > 3. \tag{4.25}$$

Среднее время обслуживания в этом случае равно единице, а само распределение далее используется как базовое.

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* (4.25) *в условиях стоимостной модели* (4.23) *и предположения о том, что формула* (4.27) *для определения среднего времени отклика системы является точной, оптимальное значение загрузки системы ρ_0 определяется решением уравнения*

$$\rho^3 g'(\rho) + \rho^2 g(\rho) = c_1, \tag{4.26}$$

где $g(\rho)$ *и* $g'(\rho)$ *определяются выражениями* (4.31) *и* (4.32).

Доказательство. В оценочной формуле для среднего времени отклика используются выражения для математического ожидания и среднеквадратического отклонения времени пребывания заявки (подзаявки) в системе (подсистеме) $M|G|1$ с распределением Парето времени обслуживания (4.25), а также множитель, связанный с асимптотикой роста максимума случайных величин с распределением Парето. Кроме того, для улучшения качества аппроксимации был введен поправочный множитель $\widetilde{\mu}(\alpha, \rho)$ квадратичной формы. Таким образом, оценка имеет вид

$$E[R_K] \approx \mu_{Pa} + \sigma_{Pa}(K^{\frac{1}{\alpha}} - 1) \cdot \widetilde{\mu}(\alpha, \rho), \tag{4.27}$$

где μ_{Pa} и σ_{Pa} — математическое ожидание и среднеквадратическое отклонение времени пребывания в подсистемах типа $M|Pa|1$, которые определяются следующими выражениями

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

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

а второй момент времени пребывания подзаявки в подсистеме типа $M|Pa|1$ равен

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

Соотношение для поправочного коэффициента представляет собой следующее выражение:

$$\widetilde{\mu}(\alpha, \rho) = Q_1 + Q_2\alpha + Q_3\rho + Q_4\alpha\rho + Q_5\rho^2 + Q_6\alpha^2, \tag{4.28}$$

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

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

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

Таким образом, в общем виде получаем следующее:

$$\begin{gathered} E[R_K] \approx 1 + \frac{\rho(\alpha-1)^2}{2\alpha(\alpha-2)(1-\rho)} + (K^{\frac{1}{\alpha}} - 1)\cdot \\ \cdot \left(Q_1 + Q_2\alpha + Q_3\rho + Q_4\alpha\rho + Q_5\rho^2 + Q_6\alpha^2\right) \cdot \\ \cdot \sqrt{\mu_{Pa}^{(2)} - \left(1 + \frac{\rho(\alpha-1)^2}{2\alpha(\alpha-2)(1-\rho)}\right)^2}. \end{gathered} \tag{4.29}$$

Решается задача для данного приближения в предположении, как если бы формула была точной.

Тогда функция $g(\rho)$ из модели (4.23) будет определяться (4.29), т. е. $g(\rho) = E[R_K]$. Чтобы в дальнейшем использовать полученное соотношения в решении уравнения (4.24), его необходимо упростить. Упростим подкоренное выражение, в результате получим

$$\begin{gathered} \sigma_{Pa}^2 = \mu_{Pa}^{(2)} - \left(1 + \frac{\rho(\alpha-1)^2}{2\alpha(\alpha-2)(1-\rho)}\right)^2 = \\ = \frac{1}{\alpha(\alpha-2)} + \frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{\rho}{1-\rho} + \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}. \end{gathered} \tag{4.30}$$

Таким образом, окончательно получаем

$$\begin{gathered} g(\rho) = 1 + \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{\rho}{1-\rho} + (K^{\frac{1}{\alpha}} - 1)\cdot \\ \cdot \left((Q_1 + Q_2\alpha + Q_6\alpha^2) + (Q_3 + Q_4\alpha)\rho + Q_5\rho^2\right) \cdot \\ \cdot \left(\frac{1}{\alpha(\alpha-2)} + \frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{\rho}{1-\rho} + \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}\right)^{\frac{1}{2}}. \end{gathered} \tag{4.31}$$

Теперь определим производную полученной функции, она будет иметь вид

$$g'(\rho) = \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{1}{(1-\rho)^2} + (K^{\frac{1}{\alpha}}-1)\cdot\Bigg[(Q_3+\alpha Q_4+2Q_5\rho)\cdot$$
$$\cdot\left(\frac{1}{\alpha(\alpha-2)} + \frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{\rho}{1-\rho} + \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}\right)^{\frac{1}{2}} +$$
$$+\frac{1}{2}\left((Q_1+Q_2\alpha+Q_6\alpha^2)+(Q_3+Q_4\alpha)\rho+Q_5\rho^2\right)\cdot \tag{4.32}$$
$$\cdot\left(\frac{1}{\alpha(\alpha-2)} + \frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{\rho}{1-\rho} + \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}\right)^{-\frac{1}{2}}\cdot$$
$$\cdot\left(\frac{(\alpha-1)^3}{3\alpha^2(\alpha-3)}\frac{1}{(1-\rho)^2} + \frac{(\alpha-1)^4}{2\alpha^2(\alpha-2)^2}\frac{\rho}{(1-\rho)^3}\right)\Bigg].$$

В результате имеем уравнение

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

где $g(\rho)$ и $g'(\rho)$ определяются выражениями (4.31) и (4.32). □

Полученное уравнение (4.26) можно решить численно и определить оптимальное значение загрузки ρ_0 и, соответственно искомое значение оптимальной интенсивности обслуживания μ_0.

4.8 Численный эксперимент для fork-join СМО с распределением Парето времени обслуживания

В данном разделе анализируется поведение численного решения уравнения (4.26) при различном количестве подсистем системы с разделением и параллельным обслуживанием заявок, а также при различных значениях параметра $\alpha > 3$ распределения Парето времени обслуживания. Выражение для аппроксимации среднего времени отклика использовалось для ограниченного набора параметров $\alpha \in [4, 10]$ и $K = \{2, 3, ..., 20\}$, при этом не исключена справедливость оценки и на большем диапазоне. Тем не менее, в рамках численного эксперимента ограничимся указанным набором данных.

Стоит подчеркнуть, что функция распределения Парето вида (4.25), рассматриваемая для случайной величины времени обслуживания и, соответственно, для оценки среднего времени отклика (4.29), была подобрана таким образом, что среднее время обслуживания составляло одну условную единицу. Однако в построенной модели для нахождения оптимального значения загрузки системы ρ_0 она обобщается за счет множителя $b = 1/\mu$ на случай произвольного значения скорости обслуживания.

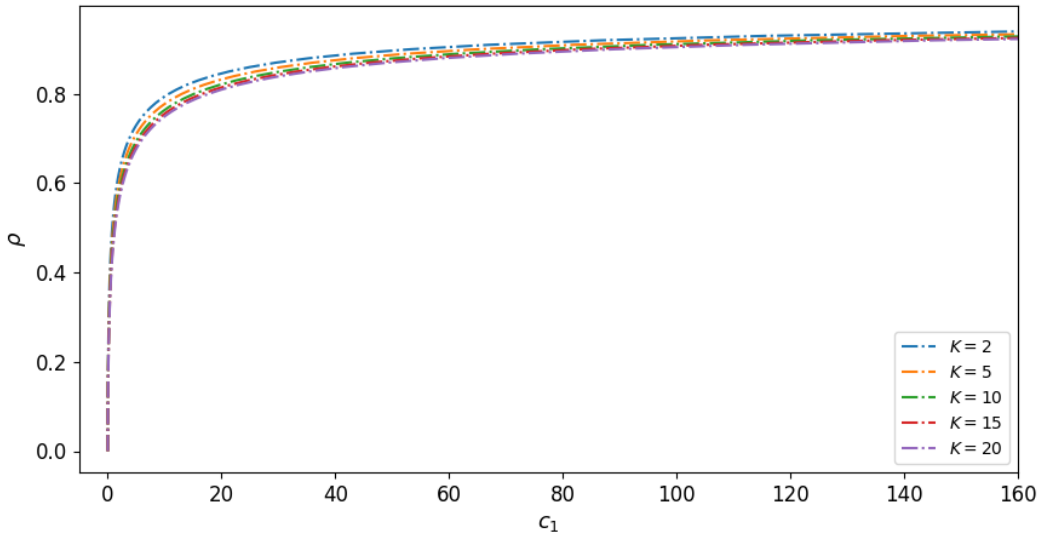


Рис. 96: График зависимости значения оптимального уровня загрузки системы ρ (решение уравнения (4.26)) от параметра c_1 для различного числа подсистем K при $\alpha = 4$.

Для начала рассмотрим поведение оптимального решения при различном числе подсистем $K \geqslant 2$. Для этого зафиксируем значение параметра $\alpha = 4$. На рисунке 96 представлены графики зависимости поведения оптимального значения загрузки $\rho = \rho_0$, являющегося решением уравнения (4.26), в зависимости от значения параметра c_1 для различного числа подсистем K fork-join СМО. Параметр c_1 пропорционален значению цены единицы ресурса c и поэтому, изучая зависимость оптимального значения загрузки от параметра c_1, мы тем самым изучаем зависимость от цены единицы ресурса c.

На начальном этапе, как и в случае подсистем типа $M|M|1$ системы с разделением и параллельным обслуживанием, наблюдается довольно стремительный рост оптимальной загрузки с увеличением значения c_1. Так, для числа подсистем $K = 20$ для сохранения оптимальной производительности системы при $c_1 > 1.3$ необходим уровень загрузки не ниже среднего ($\rho > 0.5$). При дальнейшем увеличении значения параметра c_1 рост оптимального значения загрузки заметно замедляется. Так, например, при $c_1 \approx 35.3$, что почти в 27 раз больше $c_1 = 1.3$, значение $\rho_0 = 0.85$, т. е. не превышает 90%.

Также стоит отметить, что в целом с увеличением количества подсистем K или количества подзадач, на которое разделяется исходная задача, оптимальный уровень загрузки системы снижается, что было ожидаемо.

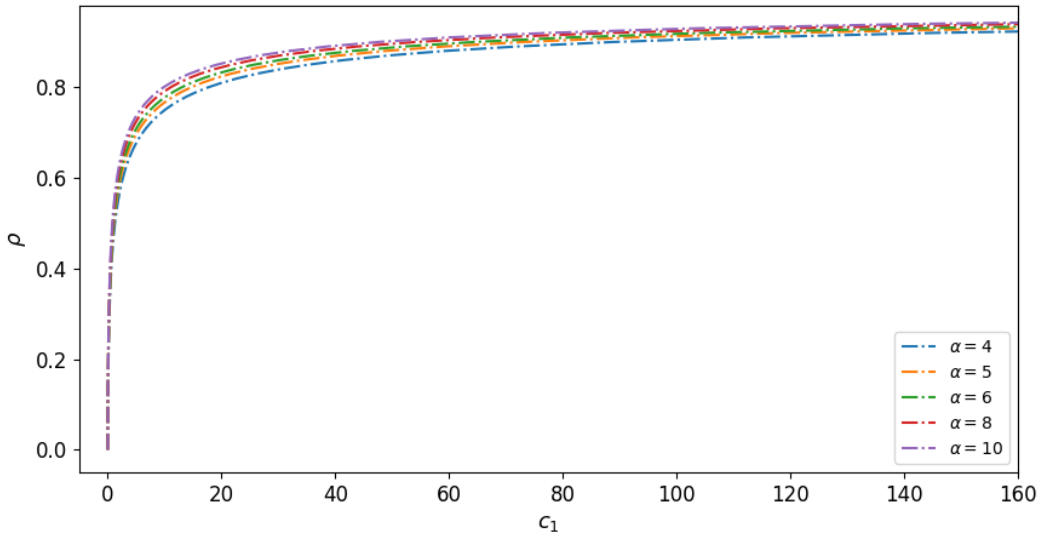


Рис. 97: График зависимости значения оптимального уровня загрузки системы ρ (решение уравнения (4.26)) от стоимости c_1для различных значений параметра α при $K = 20$.

На рисунке 97 представлены графики оптимального значения загрузки $\rho = \rho_0$, являющегося решением уравнения (4.26), в зависимости от значения параметра c_1 для фиксированного числа подсистем $K = 20$ fork-join системы при различных значениях параметра α, а именно $\alpha = 4, 5, 6, 8, 10$. Аналогично графикам на рисунке 96 наблюдается довольно быстрый рост с увеличением

значения c_1, однако затем рост заметно замедляется. Характер зависимости оптимальной загрузки от параметра α отличается от характера зависимости от параметра K, а именно с ростом значения α требуемый уровень загрузки системы повышается, в то время как с ростом K наоборот — требуемый уровень оптимальной загрузки снижается.

4.9 Асимптотика поведения оптимального решения для системы с разделением и параллельным обслуживание с подсистемами $M|Pa|1$

Рассмотрим уравнение (4.24) и определим поведение его решения при стремлении $c_1 \to 0$ ($\rho \to 0$) и $c_1 \to +\infty$ ($\rho \to 1$) в общем виде.

Теорема 4.8. *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* (4.25) *асимптотическое поведение оптимального значения загрузки системы ρ имеет вид*

$$\rho \sim \sqrt{\frac{c_1}{g(0)}}, \quad c_1 \to 0. \tag{4.33}$$

$$\rho = 1 - \sqrt{\frac{L}{c_1}} \cdot (1 + o(1)), \quad c_1 \to +\infty, \tag{4.34}$$

где

$$L = \lim_{\rho \to 1} (1 - \rho)^2 g'(\rho).$$

Доказательство. Для начала проанализируем случай $\rho \to 0$, соответственно, $c_1 \to 0$. Если $g'(0) < \infty$ и $g(0) > 0$, то $\rho g'(\rho) = o(g(\rho))$ Тогда из соотношения (4.24) имеем

$$\rho^2 g(0) \sim c_1, \quad \rho \to 0, c_1 \to 0,$$

откуда следует, что

$$\rho \sim \sqrt{\frac{c_1}{g(0)}}, \quad c_1 \to 0.$$

Теперь проанализируем случай $\rho \to 1$, соответственно, $c_1 \to +\infty$. В общем случае имеем

$$\rho g'(\rho) \sim g'(\rho), \quad \rho \to 1.$$

Таким образом, если существует такое число $L \in (0, +\infty)$, что

$$L = \lim_{\rho \to 1} (1-\rho)^2 g'(\rho),$$

то

$$g'(\rho) \sim \frac{L}{(1-\rho)^2}, \quad \rho \to 1,$$

а, соответственно,

$$g(\rho) \sim \frac{L}{(1-\rho)}, \quad \rho \to 1.$$

Отсюда следует

$$g(\rho) = o(\rho g'(\rho))$$

и тогда

$$g(\rho) = o(g'(\rho)), \quad \rho \to 1,$$

поэтому при подстановке полученного выражения для $g'(\rho)$ с L в (4.24) и учетом асимптотики имеем следующее

$$\rho^2 g'(\rho) \sim c_1, \quad \rho \to 1, c_1 \to +\infty,$$

$$g'(\rho) \sim c_1, \quad \rho \to 1, c_1 \to +\infty,$$

$$\frac{L}{(1-\rho)^2} \sim c_1, \quad \rho \to 1, c_1 \to +\infty,$$

$$(1-\rho)^2 \sim \frac{L}{c_1}, \quad \rho \to 1, c_1 \to +\infty,$$

поэтому

$$\rho = 1 - \sqrt{\frac{L}{c_1}} \cdot (1 + o(1)), \quad c_1 \to \infty.$$

□

Далее проанализируем поведение оптимальной загрузки и определим конкретные выражения для полученных в общем виде эквивалентностей в случае fork-join системы с распределением Парето времени обслуживания.

Теорема 4.9. *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* (4.25) *в условиях стоимостной модели* (4.23) *и предположения о том, что формула* (4.27) *для определения среднего времени отклика системы является точной, асимптотическое поведение оптимального значения загрузки системы ρ имеет вид*

$$\rho \sim \sqrt{\frac{c_1}{1 + (K^{\frac{1}{\alpha}} - 1)\left(Q_1 + Q_2\alpha + Q_6\alpha^2\right)\left(\alpha(\alpha-2)\right)^{-\frac{1}{2}}}}, \quad c_1 \to 0,$$

$$\rho = 1 - \sqrt{\frac{1}{c_1}\left(\frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\left[1 + (K^{\frac{1}{\alpha}} - 1)\left(\widetilde{Q}_1 + \widetilde{Q}_2\alpha + Q_6\alpha^2\right)\right]\right)} \cdot (1 + o(1)),$$

$$c_1 \to +\infty,$$

где

$$Q_1 \approx 1.25918, \quad Q_2 \approx 0.36996, \quad Q_6 \approx -0.01122,$$

$$\widetilde{Q}_1 \approx 0.69360, \quad \widetilde{Q}_2 \approx 0.08502.$$

Доказательство. Сперва рассмотрим вариант $c_1 \to 0$ ($\rho \to 0$). Рассчитаем значение функции $g(\rho)$ в нуле, подставив соответствующее значение в формулу (4.31):

$$g(0) = 1 + (K^{\frac{1}{\alpha}} - 1)\left(Q_1 + Q_2\alpha + Q_6\alpha^2\right).\frac{1}{\sqrt{\alpha(\alpha-2)}}. \tag{4.35}$$

Далее при подстановке в (4.33) получим

$$\rho \sim \sqrt{\frac{c_1}{1 + (K^{\frac{1}{\alpha}} - 1)\left(Q_1 + Q_2\alpha + Q_6\alpha^2\right)\left(\alpha(\alpha-2)\right)^{-\frac{1}{2}}}}, \quad c_1 \to 0. \tag{4.36}$$

Теперь определим асимптотику решения при $c_1 \to +\infty$ ($\rho \to 1$). Для этого найдем значение L. Для начала перепишем формулы для $g(\rho)$ из (4.31) и $g'(\rho)$ из (4.32) в более компактном виде

$$g(\rho) = \mu_{Pa} + (K^{\frac{1}{\alpha}} - 1)\widetilde{\mu}(\alpha, \rho)\sigma_{Pa},$$

тогда

$$g'(\rho) = \mu'_{Pa} + (K^{\frac{1}{\alpha}} - 1)\left(\widetilde{\mu}(\alpha,\rho)\sigma'_{Pa} + \widetilde{\mu}'(\alpha,\rho)\sigma_{Pa}\right),$$

где

$$\mu'_{Pa} = \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{1}{(1-\rho)^2}, \quad \widetilde{\mu}'(\alpha,\rho) = Q_3 + Q_4\alpha + 2Q_5\rho,$$

σ^2_{Pa} определяется выражением (4.30), а $\widetilde{\mu}(\alpha,\rho)$ — выражением (4.28).

Тогда из вида (4.30) следует

$$\sigma^2_{Pa} \sim \frac{(\alpha-1)^4}{4\alpha^2(\alpha-2)^2}\frac{\rho^2}{(1-\rho)^2}, \quad \rho \to 1,$$

следовательно,

$$\sigma_{Pa} \sim \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{\rho}{1-\rho}, \quad \rho \to 1,$$

и тогда

$$\sigma'_{Pa} \sim \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{1}{(1-\rho)^2}, \quad \rho \to 1.$$

Кроме того,

$$\widetilde{\mu}(\alpha,1) = Q_1 + Q_2\alpha + Q_3 + Q_4\alpha + Q_5 + Q_6\alpha^2 =$$

$$= Q_1 + Q_3 + Q_5 + (Q_2 + Q_4)\alpha + Q_6\alpha^2 = \widetilde{Q}_1 + \widetilde{Q}_2\alpha + Q_6\alpha^2,$$

где

$$\widetilde{Q}_1 = Q_1 + Q_3 + Q_5 = 0.69360, \quad \widetilde{Q}_2 = Q_2 + Q_4 = 0.08502.$$

Теперь можем записать

$$g'(\rho) \sim \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\frac{1}{(1-\rho)^2}\left[1 + (K^{\frac{1}{\alpha}} - 1)\left(\widetilde{Q}_1 + \widetilde{Q}_2\alpha + Q_6\alpha^2\right)\right]$$

$$L = \lim_{\rho\to 1}(1-\rho)^2 g'(\rho) = \frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\left[1 + (K^{\frac{1}{\alpha}} - 1)\left(\widetilde{Q}_1 + \widetilde{Q}_2\alpha + Q_6\alpha^2\right)\right]. \quad (4.37)$$

Таким образом, окончательно получаем

$$\rho = 1 - \sqrt{\frac{1}{c_1}\left(\frac{(\alpha-1)^2}{2\alpha(\alpha-2)}\left[1 + (K^{\frac{1}{\alpha}} - 1)\left(\widetilde{Q}_1 + \widetilde{Q}_2\alpha + Q_6\alpha^2\right)\right]\right)}\cdot(1+o(1)). \quad (4.38)$$

□

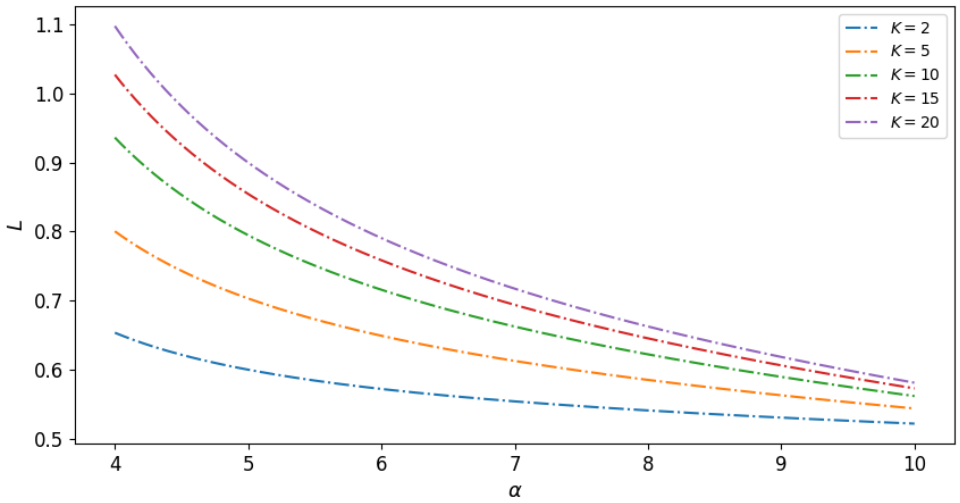


Рис. 98: График зависимости значения показателя L из (4.37) от α для различного числа подсистем K.

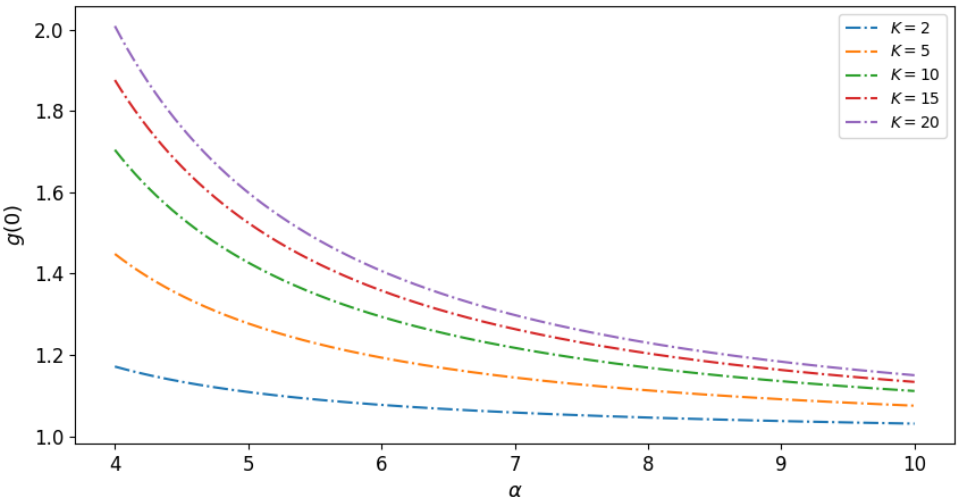


Рис. 99: График зависимости значения показателя $g(0)$ из (4.35) от α для различного числа подсистем K.

На рисунках 98 и 99 представлены зависимости значения L из (4.37) и значения функции $g(0)$ из (4.31) при $\rho = 0$ от значения параметра $\alpha \in [4, 10]$ распределения Парето времени обслуживания для различного числа подсистем $K = 2, 5, 10, 15, 20$. Как видно, в обоих случаях графики демонстрируют убывание по α и возрастание по K. Такое поведение ожидаемо, поскольку при увеличении α среднее время пребывания подзаявки сокращается, а при увеличении K

время отклика представляет собой максимум, соответственно, большего числа случайных величин (времен пребывания подзаявок).

4.10 Выводы к главе 4

В Главе исследуется система с разделением и параллельным обслуживанием с точки зрения управления режимом ее функционирования в зависимости от цены единицы ресурса и интенсивности входящего потока, влияющего на производительность системы и, соответственно, время ее отклика. Построена математическая модель, учитывающая оптимальное соотношение стоимости и эффективности работы системы как для случая показательного распределения времени обслуживания, так и для случая распределения Парето времени обслуживания. Проведен анализ на базе полученных ранее (в предыдущих главах) аппроксимаций для среднего времени отклика. Для частного случая системы с экспоненциальным распределением времени обслуживания, т. е. когда число подсистем равно двум, удается вывести в явном виде выражения для оптимальной загрузки системы.

Для большего числа подсистем представлены уравнения, численное решение которых позволяет определить искомые значения загрузки на систему и, соответственно, интенсивности обслуживания, которая фактически характеризует “мощность“ необходимых ресурсов. Также проанализировано асимптотическое поведение системы.

Поскольку используемые приближения для среднего времени отклика являются довольно точными, с их помощью также хорошо описывается суммарная стоимость затрат, поэтому при расчетных значениях оптимальной загрузки (и соответственно, оптимальной скорости обслуживания) стоимость будет близка к оптимальной. Соответственно, расчетные значения загрузки можно рекомендовать для использования в качестве оценок их фактических значений для fork-join систем. Поскольку в общем случае задача решается только численно,

то в случаях высоких и низких цен на единицу скорости обслуживания можно рекомендовать использовать асимптотические формулы, которые дают более простые, но при этом более грубые оценки оптимальной загрузки.

5 Остаточное время обслуживания в системе с разделением и параллельным обслуживанием с пуассоновским входящим потоком и произвольным распределением для времени обслуживания

Изучение систем с разделением и параллельным обслуживанием заявок, как уже неоднократно упоминалось, является актуальной задачей в настоящее время, поскольку организация вычислений подобным образом приносит существенный результат с точки зрения увеличения скорости и, соответственно, сокращения времени обработки информации, улучшая таким образом качество обслуживания пользователей таких систем. Это в значительной степени объясняет то, что, подход, основанный на параллельности процессов обработки данных, при построении современных компьютерных и информационно-вычислительных сетей получил достаточно широкое распространение [49, 61].

Одним из наиболее ярких примеров описанных систем с точки зрения высокой востребованности в силу наличия целого ряда преимуществ являются облачные вычисления. Однако распределенная природа данной технологии приводит к проблемам, связанным с надежностью и доступностью предоставляемых ею услуг [60, 142]. В частности, речь идет о нескольких типах сбоев, которые могут приводить к снижению производительности, поломке или даже отключению системы, кроме того в облачной среде существует так называемое понятие частичных отказов. Более того, если говорить в целом, то нередки ситуации преднамеренного выключения любой вычислительной системы в ответ на инцидент информационной безопасности, что, вероятно, является наиболее радикальным решением, но иногда самым лучшим при развитии определенных сценариев.

В стохастических 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 данной главе будут рассматриваться в качестве подсистем системы с разделением и параллельным обслуживанием системы с бесконечным числом приборов, которые на практике приближают системы с очень большим объемом ресурсов. Результаты Главы 5 отражены в статье [106].

5.1 Математическая модель исследуемой системы с разделением и параллельным обслуживанием

Рассмотрим систему с разделением и параллельным обслуживанием, для которой введем обозначение вида $M^{(2)}|G_2|\infty$. Это означает, что в систему с интенсивностью λ поступает пуассоновский поток заявок. В момент поступления заявка мгновенно разделяется на две идентичные подзаявки, каждая из которых поступает на обслуживание в первую и вторую подсистемы исходной системы, соответственно.

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 при-

борах второй подсистемы (рис. 100).

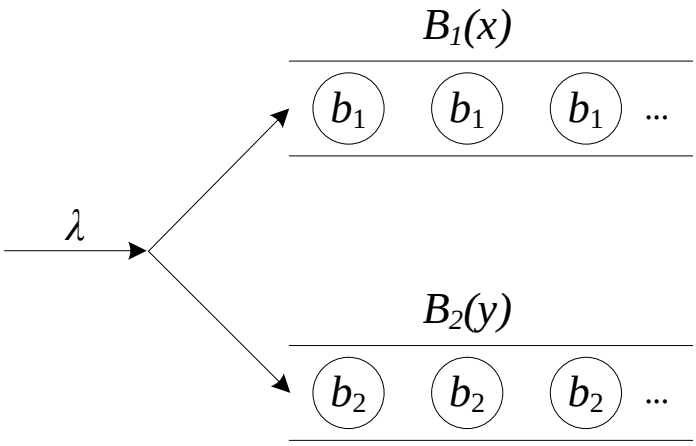


Рис. 100: Схема системы с разделением и параллельным обслуживанием с подсистемами типа $M|G|\infty$.

Вообще говоря, бесконечнолинейные системы массового обслуживания $M|G|\infty$ изучаются еще со времен работы [173]. Основные результаты для них изложены в учебниках [1, 6]. Дальнейшие обобщения были в основном связаны с рассмотрением более сложного входящего потока [37–40, 53]. Так, например, в [37–39] изучались системы с групповым поступлением $M^{[X]}|G|\infty$ с точки зрения максимального числа заявок. В работах [23, 24, 27, 30, 41, 48] исследовались системы с параллельным обслуживанием кратных заявок, т. е. подобные той, что рассматривается здесь. В [27, 30, 41] были получены аналитические выражения для производящих функций распределения вероятностей состояний цепи Маркова, характеризующей число заявок в каждом блоке.

Что касается такой характеристики, как максимальное остаточное время обслуживания, то оно изучалось в сравнительно меньшей степени. Так, в работах [14, 43] были рассмотрены бесконечнолинейные системы с дважды стохастическим пуассоновским входящим потоком, предполагалось, что время обслуживания распределено показательно, а интенсивность входящего потока переключается между значениями λ_1 и λ_2 через показательно распределенные промежутки времени с параметрами α_1 и α_2. Были выведены рекуррентные формулы,

позволяющие оценивать плотность распределения численно. В [40] максимальное остаточное время обслуживания изучалось в бесконечнолинейных системах типа $M|G|\infty$ на данный момент и в стационарном режиме в условиях различных вариантов интенсивностей входящего потока.

Для интенсивности входящего потока проанализируем три варианта, а именно — случай интенсивности λ, не зависящей от времени, случай интенсивности $\lambda(t)$, заданной функцией от времени и случай интенсивности $\lambda(t)$, заданной случайным процессом.

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. Под коэффициентом Бломквиста двух случайных величин ξ и η с функциями распределения $G_1(x)$ и $G_2(y)$ понимается [163]

$$\begin{gathered}\beta_C = P[(\xi - x_{1/2})(\eta - y_{1/2}) > 0] - P[(\xi - x_{1/2})(\eta - y_{1/2}) < 0] = \\ = \mathbf{E}\,\mathrm{sign}(\xi - x_{1/2})(\eta - y_{1/2}),\end{gathered} \tag{5.1}$$

где $x_{1/2}$ и $y_{1/2}$ — это квантили уровня 0.5, т. е. медианы случайных величин ξ и η, соответственно. Поэтому коэффициент Бломквиста еще иногда называют медиальным коэффициентом корреляции.

Если известна двумерная функция распределения $G(x, y)$ для ξ и η, то (5.1)

сводится к

$$\beta_C = 4G(x_{1/2}, y_{1/2}) - 1, \tag{5.2}$$

а поскольку по определению и согласно теореме Скляра

$$G(x, y) = C(G_1(x), G_2(y)),$$

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

$$\beta_C = 4C\left(\frac{1}{2}, \frac{1}{2}\right) - 1. \tag{5.3}$$

Последним выражением и будем пользоваться для расчета коэффициента Бломквиста.

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 [2, 50, 163].

Сразу отметим, что в рамках исследуемой модели имеется следующий нюанс: рассматриваемые частные функции распределения $G_{1,T}(x)$ и $G_{2,T}(y)$ не будут являться непрерывными, они имеют скачки в нуле, от нуля до положительных значений (имеющих смысл вероятностей того, что первая и вторая подсистемы свободны). Поэтому копулу $C_T(u, v)$ имеет смысл определять только при $u \geqslant G_{1,T}(0)$ и $v \geqslant G_{2,T}(0)$. Кроме того, коэффициент Бломквиста имеет смысл, только когда медианы максимальных остаточных времен обслуживания положительны, т. е. при условиях $G_{1,T}(0) < 1/2$ и $G_{2,T}(0) < 1/2$.

5.2 Случай интенсивности входящего потока, не зависящей от времени

Рассмотрим сначала 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*

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

Доказательство. Будем использовать свойство стационарного пуассоновского потока: при известном числе k заявок, поступивших на отрезке $[0, T]$, моменты их поступления (без учета порядка) независимы и равномерно распределены на этом отрезке, т. е. имеют плотность $1/T$ при $t \in [0, T]$.

При условии, что одна заявка поступила в момент $(T - t)$, вероятность ее обслуживания до момента $(T+x)$ имеет распределение $B(t+x)$. Обслуживание заявок происходит независимо. Тем не менее, в нашем случае в систему факти-

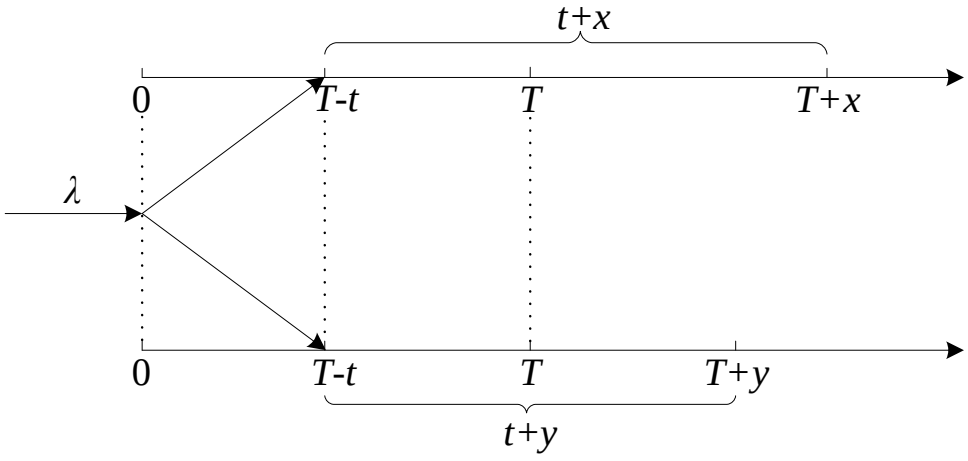


Рис. 101: Временная схема.

чески поступают две заявки одновременно, т. е. по одной заявке в каждую из подсистем.

Итак, вероятность того, что пара таких заявок на момент T будет обслужена до нужных моментов, т. е. первая — до момента $(T+x)$, а вторая — до момента $(T+y)$, будет равна $B_1(t+x)B_2(t+y)$ (рис. 101). А вероятность того, что за время T в систему поступит k заявок, равна $(\lambda T)^k e^{-\lambda T}/k!$ в силу пуассоновости входящего потока.

Следовательно, совместное распределением максимумов остаточных времен обслуживания в подсистемах можно представить следующим образом:

$$G_T(x,y) = \sum_{k=0}^{\infty} \frac{(\lambda T)^k}{k!} e^{-\lambda T} \left(\frac{1}{T} \int_0^T B_1(t+x)B_2(t+y)dt \right)^k =$$

$$= \sum_{k=0}^{\infty} \frac{\left(\lambda \int\limits_0^T B_1(t+x)B_2(t+y)dt \right)^k}{k!} e^{-\lambda T},$$

а с учетом разложения в ряд Маклорена можем переписать выражения для $G_T(x,y)$ следующим образом

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

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

□

Выражение для $G_T(x,y)$ можно представить следующим образом

$$G_T(x,y) = \exp\left\{ -\lambda \int_0^T \bar{B}_1(t+x)dt - \lambda \int_0^T \bar{B}_2(t+y)dt + \right.$$
$$\left. + \lambda \int_0^T \bar{B}_1(t+x)\bar{B}_2(t+y)dt \right\} = G_{1,T}(x)G_{2,T}(y)D_T(x,y), \quad (5.4)$$

где

$$G_{1,T}(x) = \exp\left\{ -\lambda \int_0^T \bar{B}_1(t+x)dt \right\},$$
$$G_{2,T}(y) = \exp\left\{ -\lambda \int_0^T \bar{B}_2(t+y)dt \right\} \quad (5.5)$$

— частные функции распределения максимальных остаточных времен обслуживания в первой и второй подсистемах,

$$D_T(x,y) = \exp\left\{\lambda \int_0^T \bar{B}_1(t+x)\bar{B}_2(t+y)dt\right\}, \tag{5.6}$$

— множитель, описывающий зависимость (в его отсутствие они были бы независимы). С помощью верхней черты обозначаются хвосты распределений

$$\bar{B}(\cdot) = 1 - B(\cdot).$$

Через $G_\infty(x,y)$ будем обозначать предельное распределение ($T \to +\infty$), если оно существует.

Далее сформулируем и докажем несколько следствий из теоремы 5.1.

Следствие 5.1. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ имеют показательное распределение, т. е.*

$$B_1(x) = 1 - e^{-\beta_1 x}, \quad B_2(y) = 1 - e^{-\beta_2 y}, \quad x, y \geqslant 0, \quad \beta_1, \beta_2 > 0, \tag{5.7}$$

то тогда функция распределения максимумов остаточных времен обслуживания и их копула на момент времени T определяются как

$$G_T(x,y) = \exp\left\{ -\lambda\left(\frac{e^{-\beta_1 x}(1-e^{-\beta_1 T})}{\beta_1} + \frac{e^{-\beta_2 y}(1-e^{-\beta_2 T})}{\beta_2} - \right.\right.$$
$$\left.\left. - \frac{e^{-(\beta_1 x+\beta_2 y)}(1-e^{-(\beta_1+\beta_2)T})}{\beta_1+\beta_2}\right)\right\}, \quad x \geqslant 0, \quad y \geqslant 0, \quad 0 < T \leqslant \infty,$$

$$C_T(u,v) = u \cdot v \cdot \exp\left\{\frac{\beta_1\beta_2}{\lambda(\beta_1+\beta_2)}\frac{1-e^{-(\beta_1+\beta_2)T}}{(1-e^{-\beta_1 T})(1-e^{-\beta_1 T})}\ln u \ln v\right\},$$

а их предельные значения и коэффициент Бломквиста имеют вид

$$G_\infty(x,y) = \exp\left\{-\lambda\left(\frac{1}{\beta_1}e^{-\beta_1 x} + \frac{1}{\beta_2}e^{-\beta_2 y} - \frac{1}{\beta_1+\beta_2}e^{-(\beta_1 x+\beta_2 y)}\right)\right\}, \quad x \geqslant 0, \quad y \geqslant 0,$$

$$C_\infty(u,v) = u \cdot v \cdot \exp\left\{\frac{\beta_1\beta_2}{\lambda(\beta_1+\beta_2)}\ln u \ln v\right\},$$

$$\beta_C = 2^{\frac{\beta_1\beta_2}{\lambda(\beta_1+\beta_2)}\ln 2} - 1.$$

Доказательство. Сперва найдем выражение для $G_T(x,y)$. Для этого определим все составляющие его компоненты в (5.4). Итак, подставив конкретные выражения для $B_1(t+x)$ и $B_2(t+y)$ в (5.5), получаем для частной функции распределения $G_{1,T}(x)$ следующее

$$G_{1,T}(x) = \exp\left\{ -\lambda \int\limits_0^T e^{-\beta_1(t+x)}dt \right\} = \exp\left\{ -\lambda e^{-\beta_1 x} \int\limits_0^T e^{-\beta_1 t}dt \right\} =$$

$$= \exp\left\{ \frac{\lambda}{\beta_1} e^{-\beta_1 x} e^{-\beta_1 t}\Big|_0^T \right\} = \exp\left\{ -\frac{\lambda}{\beta_1} e^{-\beta_1 x}(1 - e^{-\beta_1 T}) \right\},$$

а для частной функции распределения $G_{2,T}(y)$ аналогичным образом —

$$G_{2,T}(y) = \exp\left\{ -\frac{\lambda}{\beta_2} e^{-\beta_2 y}(1 - e^{-\beta_2 T}) \right\}.$$

Теперь найдем множитель из (5.6), определяющий зависимость

$$D_T(x,y) = \exp\left\{ \lambda \int\limits_0^T e^{-\beta_1(t+x)} e^{-\beta_2(t+y)}dt \right\} = \exp\left\{ \lambda e^{-(\beta_1 x + \beta_2 y)} \int\limits_0^T e^{-(\beta_1+\beta_2)t}dt \right\} =$$

$$= \exp\left\{ -\frac{\lambda e^{-(\beta_1 x + \beta_2 y)}}{\beta_1 + \beta_2} e^{-(\beta_1+\beta_2)t}\Big|_0^T \right\} = \exp\left\{ \frac{\lambda e^{-(\beta_1 x + \beta_2 y)}}{\beta_1 + \beta_2}(1 - e^{-(\beta_1 + \beta_2 T)}) \right\}.$$

В результате, получаем требуемое выражение для $G_T(x,y)$. Далее, устремляя T к бесконечности для $G_\infty(x,y)$ имеем

$$G_\infty(x,y) = \lim_{T\to+\infty} G_{1,T}(x) G_{2,T}(y) D_T(x,y) =$$

$$= \exp\left\{ -\lambda\left(\frac{1}{\beta_1} e^{-\beta_1 x} + \frac{1}{\beta_2} e^{-\beta_2 y} - \frac{1}{\beta_1 + \beta_2} e^{-(\beta_1 x + \beta_2 y)} \right) \right\}, \quad x \geqslant 0, \quad y \geqslant 0.$$

Теперь перейдем к нахождению копулы. Поскольку копула в соответствии со следствием из теоремы Скляра равна

$$C_T(u,v) = G_T(G_{1,T}^{-1}(u), G_{2,T}^{-1}(v)),$$

то найдем функции, обратные частным функциям распределения. Для $G_{1,T}^{-1}(u)$ справедливо

$$\exp\left\{ -\frac{\lambda}{\beta_1} e^{-\beta_1 x}(1 - e^{-\beta_1 T}) \right\} = u,$$

$$-\frac{\lambda}{\beta_1}e^{-\beta_1 x}(1-e^{-\beta_1 T})=\ln u,$$

$$e^{-\beta_1 x}=-\frac{\beta_1\ln u}{\lambda(1-e^{-\beta_1 T})},$$

следовательно,

$$G^{-1}_{1,T}(u)=-\frac{1}{\beta_1}\ln\Big(-\frac{\beta_1\ln u}{\lambda(1-e^{-\beta_1 T})}\Big).$$

Аналогичное выражение получаем для $G^{-1}_{2,T}(v)$

$$G^{-1}_{2,T}(v)=-\frac{1}{\beta_2}\ln\Big(-\frac{\beta_2\ln v}{\lambda(1-e^{-\beta_2 T})}\Big).$$

Далее

$$C_T(u,v)=G_T(G_1^{-1}(u),G_2^{-1}(v))=$$

$$=\exp\Bigg\{-\lambda\Bigg(\frac{e^{-\beta_1\frac{-1}{\beta_1}\ln\left(-\frac{\beta_1\ln u}{\lambda(1-e^{-\beta_1 T})}\right)}}{\beta_1}(1-e^{-\beta_1 T})+\frac{e^{-\beta_2\frac{-1}{\beta_2}\ln\left(-\frac{\beta_2\ln v}{\lambda(1-e^{-\beta_2 T})}\right)}}{\beta_2}(1-e^{-\beta_2 T})-$$

$$-\frac{\beta_1\beta_2\ln u\ln v}{\lambda^2(\beta_1+\beta_2)(1-e^{-\beta_1 T})(1-e^{-\beta_2 T})}(1-e^{-(\beta_1+\beta_2)T})\Bigg)\Bigg\}=$$

$$=\exp\Bigg\{\ln u+\ln v+\frac{\beta_1\beta_2\ln u\ln v}{\lambda(\beta_1+\beta_2)(1-e^{-\beta_1 T})(1-e^{-\beta_2 T})}(1-e^{-(\beta_1+\beta_2)T})\Bigg\}.$$

В итоге получаем, что копула определяется следующим выражением

$$C_T(u,v)=u\cdot v\cdot\exp\Bigg\{\frac{\beta_1\beta_2}{\lambda(\beta_1+\beta_2)}\frac{1-e^{-(\beta_1+\beta_2)T}}{(1-e^{-\beta_1 T})(1-e^{-\beta_1 T})}\ln u\ln v\Bigg\},$$

а при $T\to+\infty$

$$C_\infty(u,v)=\lim_{T\to+\infty}C_T(u,v)=u\cdot v\cdot\exp\Bigg\{\frac{\beta_1\beta_2}{\lambda(\beta_1+\beta_2)}\ln u\ln v\Bigg\}.$$

Что касается коэффициента Бломквиста, то для него имеем

$$\beta_C=4C_\infty\Big(\frac{1}{2},\frac{1}{2}\Big)-1=\exp\Bigg\{\frac{\beta_1\beta_2}{\lambda(\beta_1+\beta_2)}\ln^2\frac{1}{2}\Bigg\}-1=$$

$$=\exp\Bigg\{\frac{\beta_1\beta_2}{\lambda(\beta_1+\beta_2)}\ln 2\cdot\ln 2\Bigg\}-1=2^{\frac{\beta_1\beta_2}{\lambda(\beta_1+\beta_2)}\ln 2}-1.$$

□

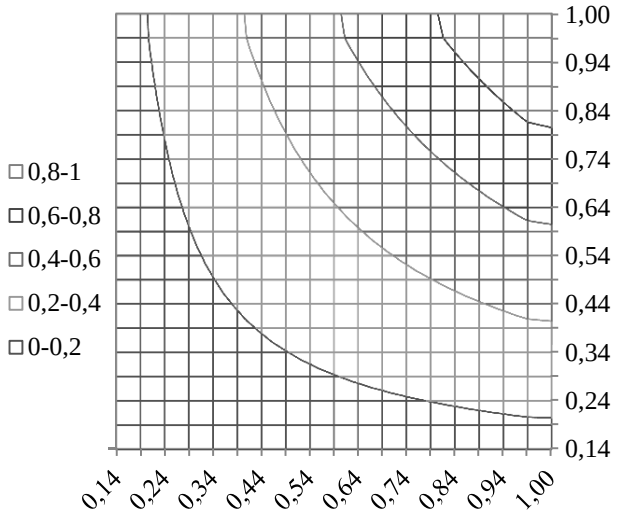


Рис. 102: Копула для случая показательного распределения, $\lambda = 2$: контурный график.

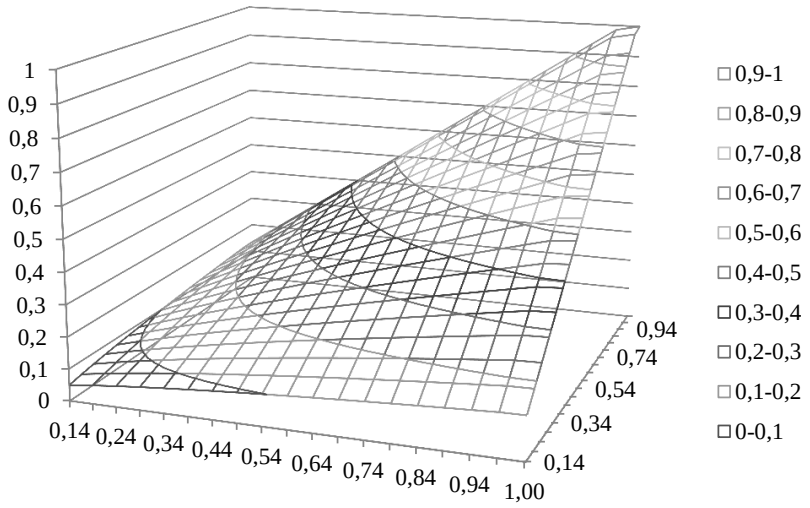


Рис. 103: Копула для случая показательного распределения, $\lambda = 2$: трехмерный график.

На рисунках 102 и 103 представлены графики копулы в случае показательного распределения времени обслуживания на приборах ($\beta_1 = \beta_2 = 1$) системы с разделением и параллельным обслуживанием.

Следствие 5.2. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с парамет-*

ром λ имеют гиперэкспоненциальное распределение, т. е.

$$B_1(x) = 1 - \sum_{i=1}^{n} c_{1i} e^{-\beta_{1i} x}, \quad x \geqslant 0, \quad \beta_{1i} > 0, \quad \sum_{i=1}^{n} c_{1i} = 1, \quad c_{1i} > 0,$$

$$B_2(y) = 1 - \sum_{i=1}^{n} c_{2i} e^{-\beta_{2i} y}, \quad y \geqslant 0, \quad \beta_{2i} > 0, \quad \sum_{i=1}^{n} c_{2i} = 1, \quad c_{2i} > 0,$$

то тогда функция распределения максимумов остаточных времен обслуживания на момент времени T и ее предельное значение определяются как

$$G_T(x,y) = \exp\Bigg\{ -\lambda\Bigg(\sum_{i=1}^{n} \frac{c_{1i}}{\beta_{1i}} e^{-\beta_{1i} x}(1 - e^{-\beta_{1i} T}) + \sum_{i=1}^{n} \frac{c_{2i}}{\beta_{2i}} e^{-\beta_{2i} y}(1 - e^{-\beta_{2i} T}) -$$
$$- \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{c_{1i} c_{2i}}{\beta_{1i} + \beta_{2i}} e^{-\beta_{1i} x - \beta_{2i} y}\big(1 - e^{-(\beta_{1i} + \beta_{2i}) T}\big)\Bigg)\Bigg\}, \quad x \geqslant 0, \quad y \geqslant 0, \quad 0 < T \leqslant \infty,$$

$$G_\infty(x,y) = \exp\Bigg\{ -\lambda\Bigg(\sum_{i=1}^{n} \frac{c_{1i}}{\beta_{1i}} e^{-\beta_{1i} x} + \sum_{i=1}^{n} \frac{c_{2i}}{\beta_{2i}} e^{-\beta_{2i} y} -$$
$$- \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{c_{1i} c_{2i}}{\beta_{1i} + \beta_{2i}} e^{-\beta_{1i} x - \beta_{2i} y}\Bigg)\Bigg\}, \quad x \geqslant 0, \quad y \geqslant 0.$$

Доказательство. Найдем выражение для $G_T(x,y)$. Для этого определим все составляющие его компоненты в (5.4). Подставляя конкретные выражения для $B_1(t+x)$ и $B_2(t+y)$ в (5.5), получаем для частной функции распределения $G_{1,T}(x)$ следующее

$$G_{1,T}(x) = \exp\Bigg\{ -\lambda \int_0^T \sum_{i=1}^{n} c_{1i} e^{-\beta_{1i}(t+x)} dt \Bigg\} = \exp\Bigg\{ -\lambda \sum_{i=1}^{n} c_{1i} e^{-\beta_{1i} x} \int_0^T e^{-\beta_{1i} t} dt \Bigg\} =$$

$$= \exp\Bigg\{ \lambda \sum_{i=1}^{n} \frac{c_{1i} e^{-\beta_{1i} x}}{\beta_{1i}} e^{-\beta_{1i} t} \Bigg|_0^T \Bigg\} = \exp\Bigg\{ -\lambda \sum_{i=1}^{n} \frac{c_{1i} e^{-\beta_{1i} x}}{\beta_{1i}} (1 - e^{-\beta_{1i} T}) \Bigg\},$$

а для частной функции распределения $G_{2,T}(y)$ аналогичным образом —

$$G_{2,T}(y) = \exp\Bigg\{ -\lambda \sum_{i=1}^{n} \frac{c_{2i} e^{-\beta_{2i} y}}{\beta_{2i}} (1 - e^{-\beta_{2i} T}) \Bigg\}.$$

Теперь найдем множитель из (5.6), определяющий зависимость

$$D_T(x,y) = \exp\left\{\lambda \int\limits_0^T \sum_{i=1}^n c_{1i} e^{-\beta_{1i}(t+x)} \sum_{j=1}^n c_{2j} e^{-\beta_{2j}(t+y)} dt\right\} =$$

$$= \exp\left\{\lambda \sum_{i=1}^n \sum_{j=1}^n c_{1i} c_{2j} e^{-\beta_{1i}x - \beta_{2j}y} \int\limits_0^T e^{-(\beta_{1i}+\beta_{2j})t} dt\right\} =$$

$$= \exp\left\{\lambda \sum_{i=1}^n \sum_{j=1}^n c_{1i} c_{2j} e^{-\beta_{1i}x - \beta_{2j}y} \frac{e^{-(\beta_{1i}+\beta_{2j})t}}{-(\beta_{1i}+\beta_{2j})}\Big|_0^T\right\} =$$

$$= \exp\left\{\lambda \sum_{i=1}^n \sum_{j=1}^n \frac{c_{1i} c_{2j}}{\beta_{1i}+\beta_{2j}} e^{-\beta_{1i}x - \beta_{2j}y}\left(1 - e^{-(\beta_{1i}+\beta_{2i})T}\right)\right\}.$$

В результате, получаем требуемое выражение для $G_T(x,y)$. Далее, устремляя T к бесконечности для $G_\infty(x,y)$ имеем

$$G_\infty(x,y) = \lim_{T\to+\infty} G_{1,T}(x) G_{2,T}(y) D_T(x,y) =$$

$$= \exp\left\{ -\lambda\left(\sum_{i=1}^n \frac{c_{1i}}{\beta_{1i}} e^{-\beta_{1i}x} + \sum_{i=1}^n \frac{c_{2i}}{\beta_{2i}} e^{-\beta_{2i}y} - \right.\right.$$

$$\left.\left. - \sum_{i=1}^n \sum_{j=1}^n \frac{c_{1i} c_{2i}}{\beta_{1i}+\beta_{2i}} e^{-\beta_{1i}x - \beta_{2i}y}\right)\right\}, \quad x \geqslant 0, \quad y \geqslant 0.$$

□

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

вания 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

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

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

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

Доказательство. Сперва найдем выражение для $G_T(x,y)$. Для этого определим все составляющие его компоненты в (5.4). Подставляя конкретные выражения для $B_1(t+x)$ и $B_2(t+y)$ в (5.5), получаем для частной функции распределения $G_{1,T}(x)$ следующее

$$G_{1,T}(x)=\exp\left\{-\lambda\int\limits_0^T\frac{1}{(1+x+t)^2}dt\right\}=\exp\left\{\lambda\cdot\frac{1}{1+x+t}\Big|_0^T\right\}=$$

$$=\exp\left\{\lambda\left(\frac{1}{1+x+T}-\frac{1}{1+x}\right)\right\}=\exp\left\{-\frac{\lambda T}{(x+T+1)(x+1)},\right\}$$

а для частной функции распределения $G_{2,T}(y)$ аналогичным образом —

$$G_{2,T}(y)=\exp\left\{-\frac{\lambda T}{(y+T+1)(y+1)}\right\}.$$

Теперь найдем множитель из (5.6), определяющий зависимость

$$D_T(x,y)=\exp\left\{\lambda\int\limits_0^T\frac{1}{(1+x+t)^2}\cdot\frac{1}{(1+y+t)^2}dt\right\}.$$

Для этого разложим подынтегральное выражение на простые дроби относительно переменной t

$$\frac{1}{(1+x+t)^2(1+y+t)^2} = \frac{2}{(x-y)^3(1+x+t)} - \frac{2}{(x-y)^3(1+y+t)} + \\ + \frac{1}{(x-y)^2(1+x+t)^2} + \frac{1}{(x-y)^2(1+y+t)^2}$$

и вернемся к вычислению интеграла

$$D_T(x,y) = \exp\left\{\lambda \int_0^T \left[\frac{2}{(x-y)^3}\left(\frac{1}{1+x+t} - \frac{1}{1+y+t}\right) + \right.\right. \\ \left.\left. + \frac{1}{(x-y)^2}\left(\frac{1}{(1+x+t)^2} - \frac{1}{1+y+t)^2}\right)\right] dt\right\} = \\ = \exp\left\{\lambda\left[\frac{2}{(x-y)^3}\ln\frac{1+x+t}{1+y+t} - \frac{1}{(x-y)^2}\left(\frac{1}{1+x+t} + \frac{1}{1+y+t}\right)\right]\Bigg|_0^T\right\} = \\ = \exp\left\{\lambda\left[\frac{2}{(x-y)^3}\left(\ln\frac{1+x+T}{1+y+T} - \ln\frac{1+x}{1+y}\right) - \right.\right. \\ \left.\left. - \frac{1}{(x-y)^2}\left(\frac{1}{1+x+T} + \frac{1}{1+y+T} - \frac{1}{1+x} - \frac{1}{1+y}\right)\right]\right\} = \\ = \exp\left\{\lambda\left[\frac{2}{(x-y)^3}\ln\frac{(1+y)(1+x+T)}{(1+x)(1+y+T)} + \right.\right. \\ \left.\left. + \frac{T}{(x-y)^2}\left(\frac{1}{(1+x)(1+x+T)} + \frac{1}{(1+y)(1+y+T)}\right)\right]\right\}.$$

В результате, после подстановки и некоторого упрощения получаем требуемое выражение для $G_T(x,y)$

$$G_T(x,y) = \exp\left\{ -\lambda\left[\frac{T}{(1+x)(1+x+T)}\left(1 - \frac{1}{(x-y)^2}\right) + \right.\right. \\ \left.\left. + \frac{T}{(1+y)(1+y+T)}\left(1 - \frac{1}{(x-y)^2}\right) - \frac{2}{(x-y)^3}\ln\frac{(1+y)(1+x+T)}{(1+x)(1+y+T)}\right]\right\} = \\ = \exp\left\{ -\lambda\left(\frac{(x-y+1)(x-y-1)}{(x-y)^2}\left(\frac{T}{(x+T+1)(x+1)} + \right.\right.\right. \\ \left.\left.\left. + \frac{T}{(y+T+1)(y+1)}\right) - \frac{2}{(x-y)^3}\ln\frac{(x+T+1)(y+1)}{(y+T+1)(x+1)}\right)\right\}.$$

Далее, устремляя T к бесконечности для частных функций распределения и $D_\infty(x,y)$ имеем

$$G_{1,\infty}(x) = \exp\left\{ -\frac{\lambda}{x+1}\right\},$$

$$G_{2,\infty}(y) = \exp\left\{ -\frac{\lambda}{y+1}\right\}.$$

$$D_\infty(x,y) = \exp\left\{\lambda\left[\frac{2}{(x-y)^3}\ln\frac{y+1}{x+1} + \frac{1}{(x-y)^2}\left(\frac{1}{x+1}+\frac{1}{y+1}\right)\right]\right\}.$$

Соответственно, получаем

$$G_\infty(x,y) = \lim_{T\to+\infty} G_{1,T}(x)G_{2,T}(y)D_T(x,y) =$$

$$= \exp\left\{ -\lambda\left(\frac{(x-y+1)(x-y-1)}{(x-y)^2}\left(\frac{1}{x+1}+\frac{1}{y+1}\right)-\right.\right.$$

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

Теперь перейдем к нахождению копулы. Поскольку копула в соответствии со следствием из теоремы Скляра равна

$$C_\infty(u,v) = G_\infty(G^{-1}_{1,\infty}(u), G^{-1}_{2,\infty}(v)),$$

то найдем функции, обратные частным функциям распределения

$$G^{-1}_{1,\infty}(u) = -\frac{\lambda}{\ln u} - 1, \quad G^{-2}_{2,\infty}(v) = -\frac{\lambda}{\ln v} - 1,$$

Далее после подстановки соответствующих выражений и некоторых преобразований копула примет вид

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

При $x = y$ получим

$$D_T(x,x) = \exp\left\{\lambda\int\limits_0^T \frac{1}{(1+x+t)^4}dt\right\} = \exp\left\{-\lambda\left(\frac{1}{3(x+T+1)^3} - \frac{1}{3(x+1)^3}\right)\right\}$$

и, соответственно,

$$G_T(x,x) = \exp\left\{ -\lambda\left(\frac{2T}{(x+T+1)(x+1)} + \frac{1}{3(x+T+1)^3} - \frac{1}{3(x+1)^3}\right)\right\} =$$

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

Тогда при $T \to \infty$ можем записать

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

что позволяет получить следующее выражение для копулы

$$C_\infty(u,u) = G_\infty(G^{-1}_{1,\infty}(u), G^{-1}_{2,\infty}(u)) = \exp\left\{ -\lambda\left(\frac{2\ln u}{-\lambda} - \frac{\ln^3 u}{-3\lambda^3}\right)\right\} =$$

$$\exp\left\{\left(\ln u^2 - \frac{\ln^3 u}{3\lambda^2}\right)\right\} = u^2 \exp\left\{ -\frac{\ln^3 u}{3\lambda^2}\right\}.$$

Следовательно, коэффициент Бломквиста равен

$$\beta_C = 4C_\infty\left(\frac{1}{2}, \frac{1}{2}\right) - 1 = \exp\left\{ -\frac{\ln^3 \frac{1}{2}}{3\lambda^2}\right\} - 1 = 2^{\frac{\ln^2 2}{3\lambda^2}} - 1.$$

□

На рисунках 104 и 105 представлены графики копулы в случае распределения Парето времени обслуживания на приборах системы с разделением и параллельным обслуживанием.

Следствие 5.4. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ имеют распределение Парето с параметром $\alpha = 3$, т. е.*

$$B_1(x) = 1 - (x+1)^{-3}, \quad B_2(y) = 1 - (y+1)^{-3}, \quad x \geqslant 0, \quad y \geqslant 0,$$

то тогда функция распределения максимумов остаточных времен обслужи-

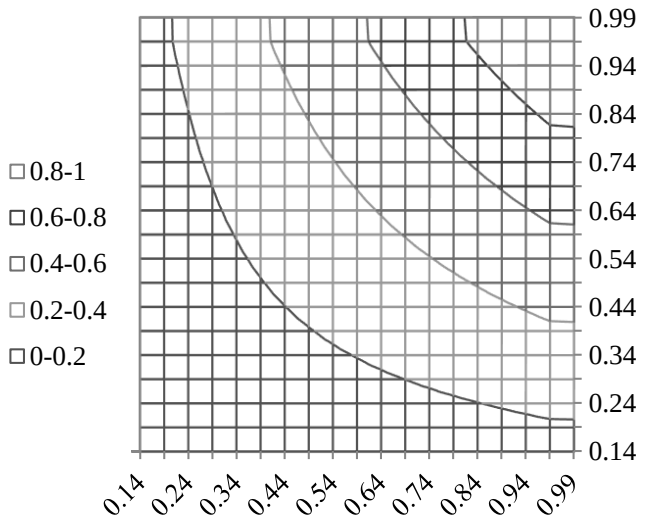


Рис. 104: Копула для случая распределения Парето, $\alpha = 2, \lambda = 2$: контурный график.

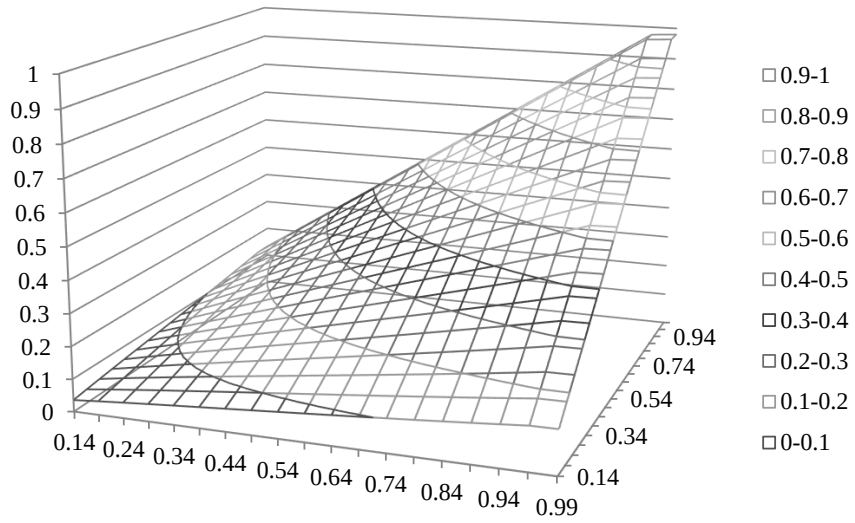


Рис. 105: Копула для случая распределения Парето, $\alpha = 2, \lambda = 2$: трехмерный график.

вания на момент времени T определяется как

$$G_T(x,y) = \exp\left\{-\lambda\left(\frac{1}{2(x+1)^2}+\frac{1}{2(y+1)^2}+\frac{1}{(x-y)^3}\left(\frac{1}{2(x+1)^2}-\frac{1}{2(y+1)^2}\right)+\right.\right.$$
$$+\frac{3}{(x-y)^4}\left(\frac{1}{x+1}+\frac{1}{y+1}\right)-\frac{1}{2(x+T+1)^2}-\frac{1}{2(y+T+1)^2}-$$
$$-\frac{1}{(x-y)^3}\left(\frac{1}{2(x+T+1)^2}-\frac{1}{2(y+T+1)^2}\right)-\frac{3}{(x-y)^4}\left(\frac{1}{x+T+1}+\frac{1}{y+T+1}\right)+$$
$$\left.\left.+\frac{6}{(x-y)^5}\ln\frac{(x+T+1)(y+1)}{(y+T+1)(x+1)}\right)\right\}, \quad x \geqslant 0, \quad y \geqslant 0, \quad 0 < T \leqslant \infty, \quad x \neq y,$$

а ее предельное значение ($T \to \infty$) и коэффициент Бломквиста имеет вид

$$G_\infty(x,y) = \exp\left\{-\lambda\left(\frac{1}{2(x+1)^2}+\frac{1}{2(y+1)^2}+\frac{1}{(x-y)^3}\left(\frac{1}{2(x+1)^2}-\frac{1}{2(y+1)^2}\right)+\right.\right.$$
$$\left.\left.+\frac{3}{(x-y)^4}\left(\frac{1}{x+1}+\frac{1}{y+1}\right)-\frac{6}{(x-y)^5}\ln\frac{x+1}{y+1},\right)\right\} \quad x \geqslant 0, \quad y \geqslant 0, \quad x \neq y,$$

$$\beta_C = 2^{\frac{2}{5}\left(\frac{\lambda}{2\ln 2}\right)^{-\frac{3}{2}}} - 1.$$

Доказательство. Сперва найдем выражение для $G_T(x,y)$. Для этого определим все составляющие его компоненты в (5.4). Подставляя конкретные выражения для $B_1(t+x)$ и $B_2(t+y)$ в (5.5), получаем для частной функции распределения $G_{1,T}(x)$ следующее

$$G_{1,T}(x) = \exp\left\{-\lambda\int\limits_0^T \frac{1}{(1+x+t)^3}dt\right\} = \exp\left\{\lambda\cdot\frac{1}{2(1+x+t)^2}\Big|_0^T\right\} =$$

$$= \exp\left\{-\lambda\left(\frac{1}{2(x+1)^2}-\frac{1}{2(x+T+1)^2}\right)\right\}, \quad x \geqslant 0,$$

а для частной функции распределения $G_{2,T}(y)$ аналогичным образом —

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

Теперь найдем множитель из (5.6), определяющий зависимость

$$D_T(x,y) = \exp\left\{\lambda\int\limits_0^T \frac{1}{(1+x+t)^3}\cdot\frac{1}{(1+y+t)^3}dt\right\}.$$

Для этого разложим подынтегральное выражение на простые дроби относительно переменной t

$$\frac{1}{(1+x+t)^3(1+y+t)^3} = -\frac{6}{(x-y)^5(1+x+t)}+\frac{6}{(x-y)^5(1+y+t)}-$$
$$-\frac{3}{(x-y)^4(1+x+t)^2}-\frac{3}{(x-y)^4(1+y+t)^2}-$$
$$-\frac{1}{(x-y)^3(1+x+t)^3}+\frac{1}{(x-y)^4(1+y+t)^3}$$

и вернемся к вычислению интеграла

$$D_T(x,y)=\exp\left\{\lambda\int\limits_0^T\left[\frac{-6}{(x-y)^5}\left(\frac{1}{1+x+t}-\frac{1}{1+y+t}\right)-\right.\right.$$

$$-\frac{3}{(x-y)^4}\left(\frac{1}{(1+x+t)^2}+\frac{1}{(1+y+t)^2}\right)-$$

$$\left.\left.-\frac{1}{(x-y)^3}\left(\frac{1}{(1+x+t)^3}-\frac{1}{(1+y+t)^3}\right)\right]dt\right\}=$$

$$=\exp\left\{\lambda\left[-\frac{6}{(x-y)^5}\ln\frac{1+x+t}{1+y+t}+\frac{3}{(x-y)^4}\left(\frac{1}{1+x+t}+\frac{1}{1+y+t}\right)+\right.\right.$$

$$\left.\left.+\frac{1}{(x-y)^3}\left(\frac{1}{2(1+x+t)^2}-\frac{1}{2(1+y+t)^2}\right)\right]\Bigg|_0^T\right\}=$$

$$=\exp\left\{\lambda\left[-\frac{6}{(x-y)^3}\left(\ln\frac{1+x+T}{1+y+T}-\ln\frac{1+x}{1+y}\right)+\right.\right.$$

$$+\frac{3}{(x-y)^4}\left(\frac{1}{1+x+T}+\frac{1}{1+y+T}-\frac{1}{1+x}-\frac{1}{1+y}\right)+$$

$$\left.\left.+\frac{1}{(x-y)^3}\left(\frac{1}{2(1+x+T)^2}-\frac{1}{2(1+y+T)^2}+\frac{1}{2(1+x)}-\frac{1}{2(1+y)}\right)\right]\right\}=$$

$$=\exp\left\{-\lambda\left[\frac{6}{(x-y)^5}\ln\frac{(1+y)(1+x+T)}{(1+x)(1+y+T)}-\right.\right.$$

$$-\frac{3}{(x-y)^4}\left(\frac{1}{1+x+T}+\frac{1}{1+y+T}-\frac{1}{1+x}-\frac{1}{1+y}\right)-$$

$$\left.\left.-\frac{1}{(x-y)^3}\left(\frac{1}{2(1+x+T)^2}-\frac{1}{2(1+y+T)^2}-\frac{1}{2(1+x)}+\frac{1}{2(1+y)}\right)\right]\right\}.$$

В результате, после подстановки и некоторого упрощения получаем требуемое выражение для $G_T(x,y)$

$$G_T(x,y)=\exp\left\{-\lambda\left(\frac{1}{2(x+1)^2}+\frac{1}{2(y+1)^2}+\frac{1}{(x-y)^3}\left(\frac{1}{2(x+1)^2}-\frac{1}{2(y+1)^2}\right)+\right.\right.$$

$$+\frac{3}{(x-y)^4}\left(\frac{1}{x+1}+\frac{1}{y+1}\right)-\frac{1}{2(x+T+1)^2}-\frac{1}{2(y+T+1)^2}-$$

$$-\frac{1}{(x-y)^3}\left(\frac{1}{2(x+T+1)^2}-\frac{1}{2(y+T+1)^2}\right)-\frac{3}{(x-y)^4}\left(\frac{1}{x+T+1}+\frac{1}{y+T+1}\right)+$$

$$\left.\left.+\frac{6}{(x-y)^5}\ln\frac{(x+T+1)(y+1)}{(y+T+1)(x+1)}\right)\right\},\quad x\geqslant 0,\quad y\geqslant 0,\quad 0<T\leqslant\infty,\quad x\neq y,$$

При $x = y$ получим

$$D_T(x,x) = \exp\left\{\lambda \int\limits_0^T \frac{1}{(1+x+t)^6}dt\right\} = \exp\left\{-\lambda\left(\frac{1}{5(x+T+1)^5} - \frac{1}{5(x+1)^5}\right)\right\}$$

и, соответственно,

$$G_T(x,x) = \exp\left\{-\lambda\left(\frac{1}{5(x+T+1)^5} - \frac{1}{(x+T+1)^2} - \frac{1}{5(x+1)^5} + \frac{1}{(x+1)^2}\right)\right\},$$
$$x \geqslant 0, \quad 0 < T \leqslant \infty,$$

Тогда при $T \to \infty$ можем записать

$$G_\infty(x,x) = \exp\left\{-\lambda\left(\frac{1}{(x+1)^2} - \frac{1}{5(x+1)^5}\right)\right\}, \quad x \geqslant 0,$$

что позволяет получить следующее выражение для копулы

$$C_\infty(u,u) = G_\infty(G_{1,\infty}^{-1}(u), G_{2,\infty}^{-1}(u)) = u^2 \exp\left\{\frac{4\ln^2 u}{5\lambda}\sqrt{-\frac{2\ln u}{\lambda}}\right\}.$$

Тогда коэффициент Бломквиста равен

$$\beta_C = 4C_\infty\left(\frac{1}{2},\frac{1}{2}\right) - 1 = \exp\left\{\frac{2\ln 2}{5}\left(\frac{\lambda}{2\ln 2}\right)^{-\frac{3}{2}}\right\} - 1 = 2^{\frac{2}{5}\left(\frac{\lambda}{2\ln 2}\right)^{-\frac{3}{2}}} - 1.$$

□

Следствие 5.5. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ имеют равномерное распределение на отрезке $[0,1]$, т. е.*

$$B_1(x) = x, \quad B_2(y) = y, \quad x \in [0,1], \quad y \in [0,1],$$

то предельное распределения максимумов остаточных времен обслуживания на момент времени $T \to \infty$ и соответствующая копула определяется как

$$G_\infty(x,y) = \exp\left\{-\lambda\left(\frac{(1-x)^2}{2} + \frac{(1-y^2)}{2} - \frac{(1-\max\{x,y\})^2}{6}\big[3|x-y| + \right.\right.$$
$$\left.\left. + 2(1-\max\{x,y\})\big]\right)\right\}, \quad x \in [0,1], \quad y \in [0,1],$$

$$C_\infty(u,v) = uv \cdot (\max\{u,v\})^{\frac{1}{\lambda}\left(\frac{1}{3}\sqrt{-2\ln(\max(u,v))}-\sqrt{-2\ln(\min(u,v))}\right)},$$

а коэффициент Бломквиста имеет вид

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

Доказательство. Сперва найдем выражение для $G_\infty(x,y)$. Для этого определим все составляющие его компоненты в (5.4) при $T \to \infty$. Подставляя конкретное выражения для $B_1(t+x)$ в (5.5), получаем для частной функции распределения $G_{1,\infty}(x)$ следующее (рис. 106)

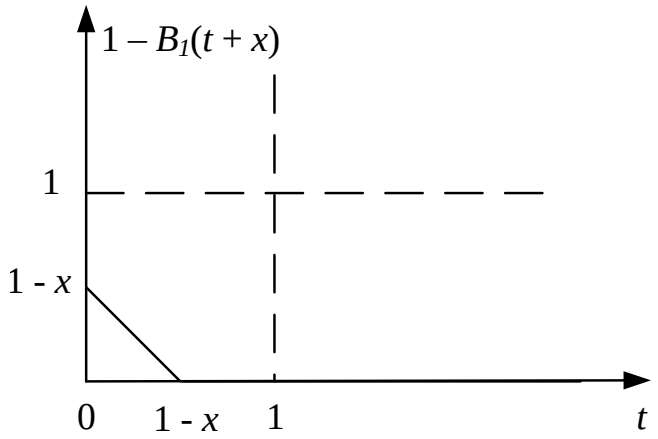


Рис. 106: График функции $1 - B_1(t+x)$ для случае равномерного распределения.

$$G_{1,\infty}(x) = \exp\left\{-\lambda\int\limits_0^\infty (1-(x+t))dt\right\} = exp\left\{-\lambda\int\limits_0^{1-x}(1-(x+t))dt\right\} =$$

$$= \exp\left\{-\lambda\left((1-x)t - \frac{t^2}{2}\right)\Big|_0^{1-x}\right\} = \exp\left\{-\lambda\frac{(1-x)^2}{2}\right\}, \quad x,y \in [0,1].$$

Аналогичное выражение получаем для $G_{2,\infty}(x)$

$$G_{2,T}(y) = \exp\left\{-\lambda\frac{(1-y)^2}{2}\right\}, \quad x,y \in [0,1].$$

Теперь найдем множитель из (5.6) при $T \to \infty$, определяющий зависимость

$$D_\infty(x,y) = \exp\left\{\lambda\int\limits_0^\infty (1-(t+x))(1-(t+y))dt\right\} =$$

$$= \exp\left\{\lambda \int_0^\infty (1-(x+t))dt + \int_0^\infty (1-(x+t))dt - \int_0^\infty (1-(x+t)(y+t))dt\right\}.$$

Для первых двух слагаемых результат уже получен, поэтому рассмотрим последнюю составляющую выражения.

Возможны два случая. В первом случае при $T > 1$ имеем, если $x \leqslant y$ (рис. 107)

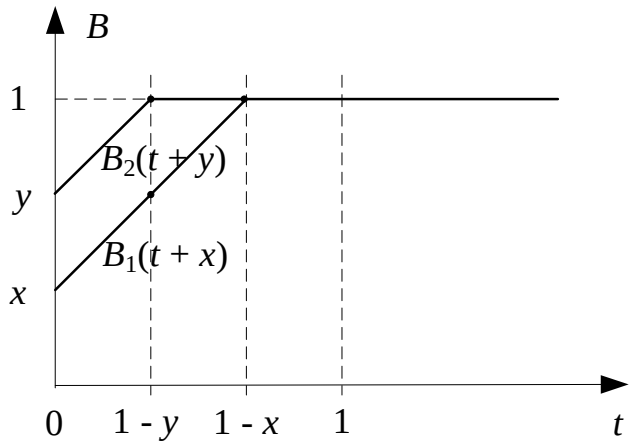


Рис. 107: Графики функций $B_1(t+x)$ и $B_2(t+y)$ для случае равномерного распределения, $x \leqslant y$.

$$\int_0^T \big(1-B_1(t+x)B_2(t+y)\big)dt = \int_0^{1-y}(1-(t+x)(t+y))dt + \int_{1-y}^{1-x}(1-(t+x))dt =$$

$$= \left((1-xy)t - (x+y)\frac{t^2}{2} - \frac{t^3}{3}\right)\Big|_0^{1-y} + \left((1-x)t - \frac{t^2}{2}\right)\Big|_{1-y}^{1-x} =$$

$$= (1-xy)(1-y) - \frac{(x+y)(1-y)^2}{2} - \frac{(1-y)^3}{3} + \frac{(1-x)^2}{2} - (1-x)(1-y) + \frac{(1-y)^2}{2} =$$

$$= \frac{(1-y)^2}{2}(1+x-y) + \frac{(1-x)^2}{2} - \frac{(1-y)^3}{3},$$

а во втором — если $x > y$ аналогично получаем

$$\int_0^T \big(1-B_1(t+x)B_2(t+y)\big)dt = \int_0^{1-x}(1-(t+x)(t+y))dt + \int_{1-x}^{1-y}(1-(t+y))dt =$$

$$= \frac{(1-x)^2}{2}(1+y-x) + \frac{(1-y)^2}{2} - \frac{(1-x)^3}{3}.$$

Таким образом, при $x \leqslant y$

$$D_\infty(x,y) = \exp\left\{\lambda\frac{(1-y)^2}{2}(y-x) + \frac{(1-y)^3}{3}\right\},$$

а при $x > y$

$$D_\infty(x,y) = \exp\left\{\lambda\frac{(1-x)^2}{2}(x-y) + \frac{(1-x)^3}{3}\right\},$$

следовательно, окончательно получаем

$$G_\infty(x,y) = \exp\left\{-\lambda\left(\frac{(1-x)^2}{2} + \frac{(1-y^2)}{2} - \frac{(1-\max\{x,y\})^2}{6}\big[3|x-y| + \right.\right.$$

$$\left.\left. +2(1-\max\{x,y\})\big]\right)\right\}, \quad x \in [0,1], \quad y \in [0,1],$$

Теперь перейдем к нахождению копулы. Поскольку копула в соответствии со следствием из теоремы Скляра равна

$$C_\infty(u,v) = G_\infty(G^{-1}_{1,\infty}(u), G^{-1}_{2,\infty}(v)),$$

то найдем функции, обратные частным функциям распределения:

$$G_1^{-1}(u) = 1 - \sqrt{-\frac{2\ln u}{\lambda}}, \quad G_2^{-1}(v) = 1 - \sqrt{-\frac{2\ln v}{\lambda}}.$$

Далее после подстановки соответствующих выражений и некоторых преобразований копула примет вид

$$C_\infty(u,v) = uv \cdot (\max\{u,v\})^{\frac{1}{\lambda}\left(\frac{1}{3}\sqrt{-2\ln(\max(u,v))} - \sqrt{-2\ln(\min(u,v))}\right)}.$$

При $u = v$ справедливо

$$C_\infty(u,u) = u^2 u^{-\frac{2}{3}\sqrt{-\frac{2\ln u}{\lambda}}},$$

тогда коэффициент Бломквиста –

$$\beta_C = 4C_\infty\left(\frac{1}{2},\frac{1}{2}\right) - 1 = 2^{\frac{2}{3}\sqrt{\frac{2\ln 2}{\lambda}}} - 1.$$

□

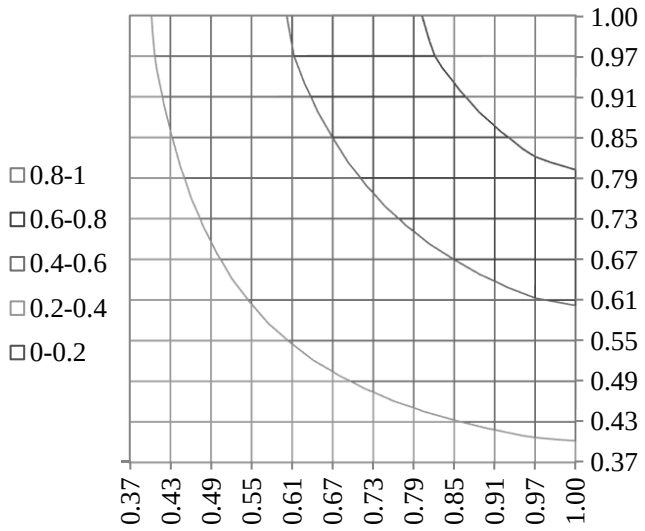


Рис. 108: Копула для случая равномерного распределения, $\lambda = 2$: контурный график.

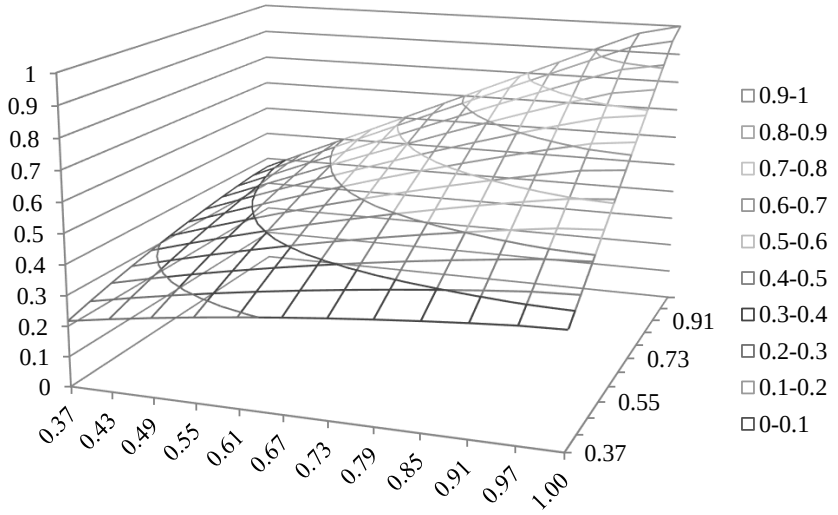


Рис. 109: Копула для случая равномерного распределения, $\lambda = 2$: трехмерный график.

Для случаев показательного распределения, распределения Парето ($\alpha = 2$, $\beta_1 = \beta_2 = 1$) и равномерного распределения при $\lambda = 2$ выше были построены графики копула-функций (рис. 102 – 105 и 108 – 109). Теперь проиллюстрируем поведение коэффициента Бломквиста (рис. 110) в зависимости от загрузки ρ для тех же распределений, чтобы учесть различия в средних временах обслуживания.

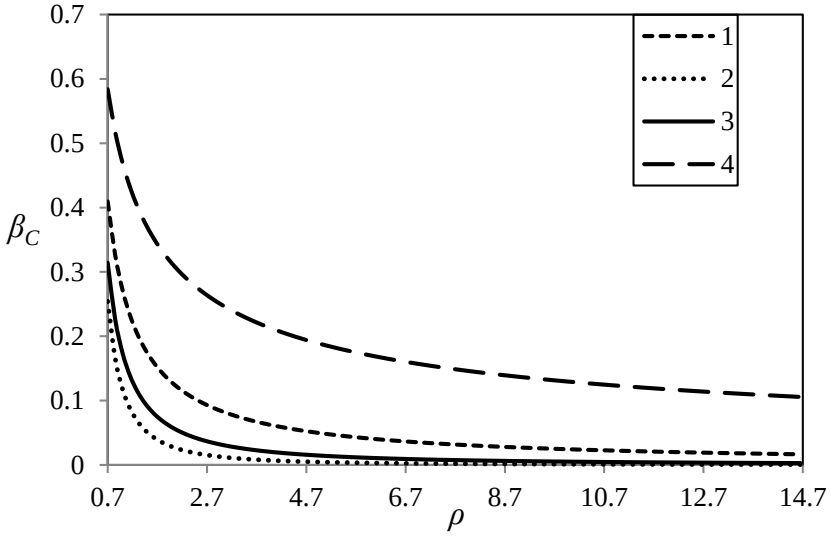


Рис. 110: Коэффициент Бломквиста: 1 — показательное распределение; 2 — распределение Парето, $\alpha = 2$; 3 — распределение Парето, $\alpha = 3$; 4 — равномерное распределение.

Формулы зависимости коэффициента Бломквиста от нагрузки для показательного распределения:

$$\beta_C = 2^{\ln 2/2\rho} - 1,$$

для распределения Парето, $\alpha = 2$,

$$\beta_C = 2^{\ln^2 2/3\rho^2} - 1,$$

для распределения Парето, $\alpha = 3$,

$$\beta_C = 2^{(\ln^2 2/5\rho)\sqrt{\ln 2/\rho}} - 1,$$

для равномерного распределения

$$\beta_C = 2^{(2/3)\sqrt{\ln 2/\rho}} - 1.$$

Следствие 5.6. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ имеют распределения Парето с параметром $\alpha > 1$, т. е.*

$$B(x) = 1 - (x+1)^{-\alpha}, \quad x \geqslant 0,$$

то тогда коэффициент Бломквиста максимумов остаточных времен обслуживания определяется следующим выражением

$$\beta_C = 2^{\frac{\alpha-1}{2\alpha-1}\left(\frac{\lambda}{(\alpha-1)\ln 2}\right)^{-\frac{\alpha}{\alpha-1}}} - 1.$$

Доказательство. Сперва найдем выражение для $G_\infty(x,x)$. Для этого определим все составляющие его компоненты в (5.4) при $T \to \infty$, $x = y$. Подставляя конкретное выражения для $B(t+x)$ в (5.5), получаем для частной функции распределения $G_{1,\infty}(x)$ следующее

$$G_{1,\infty}(x) = \exp\left\{ -\lambda \int_0^\infty (x+t+1)^{-\alpha} dt \right\} = exp\left\{ -\lambda \frac{(x+t+1)^{-\alpha+1}}{-\alpha+1}\Big|_0^\infty \right\} =$$

$$= \exp\left\{ -\frac{\lambda}{(\alpha-1)(x+1)^{\alpha-1}} \right\}.$$

Далее рассчитываем множитель, описывающий зависимость

$$D_\infty(x,x) = \exp\left\{ -\lambda \int_0^\infty (x+t+1)^{-2\alpha} dt \right\} = \exp\left\{ -\frac{\lambda}{(2\alpha-1)(x+1)^{2\alpha-1}}. \right\}.$$

В результате получаем

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

Для функции, обратной частной функции распределения, имеем

$$-\frac{\lambda}{(\alpha-1)(x+1)^{\alpha-1}} = \ln u,$$

$$x+1 = \left(-\frac{\lambda}{(\alpha-1)\ln u} \right)^{\frac{1}{\alpha-1}},$$

соответственно,

$$G^{-1}_{1,\infty}(u) = \left(-\frac{\lambda}{(\alpha-1)\ln u} \right)^{\frac{1}{\alpha-1}} - 1.$$

Следовательно, копула при $u = v$ будет определяться выражением

$$C_\infty(u,u) = G_\infty(G^{-1}_{1,\infty}(u), G^{-1}_{1,\infty}(u)) =$$

$$= \exp\left\{ -\lambda\left[\frac{2\ln u}{-\lambda} - \frac{1}{(2\alpha-1)\left(-\frac{\lambda}{(\alpha-1)\ln u}\right)^{\frac{2\alpha-1}{\alpha-1}}}\right]\right\} =$$

$$= u^2 \exp\left\{\frac{\lambda}{2\alpha-1}\left(-\frac{(\alpha-1)\ln u}{\lambda}\right)^{\frac{2\alpha-1}{\alpha-1}}\right\} = u^2 u^{\frac{(\alpha-1)^2\ln u}{\lambda(2\alpha-1)}\left(-\frac{(\alpha-1)\ln u}{\lambda}\right)^{\frac{1}{\alpha-1}}},$$

Тогда для коэффициента Бломквиста справедливо

$$\beta_C = 4C_\infty\left(\frac{1}{2},\frac{1}{2}\right) - 1 = \exp\left\{\frac{\lambda}{2\alpha-1}\left(-\frac{(\alpha-1)\ln 2}{\lambda}\right)^{\frac{2\alpha-1}{\alpha-1}}\right) - 1 =$$

$$= \exp\left\{\frac{(\alpha-1)\ln 2}{(2\alpha-1)}\left(\frac{\lambda}{(\alpha-1)\ln 2}\right)^{-\frac{\alpha}{\alpha-1}}\right\} - 1 = 2^{\frac{\alpha-1}{2\alpha-1}\left(\frac{\lambda}{(\alpha-1)\ln 2}\right)^{-\frac{\alpha}{\alpha-1}}} - 1.$$

□

Следствие 5.7. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ имеют распределения следующего вида*

$$B_1(x) = 1-(1-x)^\alpha, \quad B_2(y) = 1-(1-y)^\alpha, \quad x\in[0,1], \quad y\in[0,1], \alpha \geqslant 1,$$

то тогда коэффициент Бломквиста максимумов остаточных времен обслуживания определяется следующим выражением

$$\beta_C = 2^{\frac{\alpha+1}{2\alpha+1}\left(\frac{\lambda}{(\alpha+1)\ln 2}\right)^{-\frac{\alpha}{\alpha+1}}} - 1.$$

Доказательство. Сперва найдем выражение для $G_\infty(x,x)$. Для этого определим все составляющие его компоненты в (5.4) при $T\to\infty$, $x=y$. Подставляя конкретное выражения для $B(t+x)$ в (5.5), получаем для частной функции распределения $G_{1,\infty}(x)$ следующее

$$G_{1,\infty}(x) = \exp\left\{-\lambda\int_0^\infty (1-(x+t))^\alpha dt\right\} =$$

$$= \exp\left\{-\lambda\int_0^{1-x}(1-(x+t))^\alpha dt\right\} = exp\left\{-\lambda\frac{(1-(x+t))^{\alpha+1}}{\alpha+1}\Big|_0^{1-x}\right\} =$$

$$= \exp\left\{ -\lambda \cdot \frac{(1-x)^{\alpha+1}}{\alpha+1} \right\}.$$

Далее рассчитываем множитель, описывающий зависимость

$$D_\infty(x,x) = \exp\left\{ -\lambda \int\limits_0^{1-x} (1-(x+t))^{2\alpha} dt \right\} = \exp\left\{ -\lambda \cdot \frac{(1-x)^{2\alpha+1}}{2\alpha+1} \right\}.$$

В результате получаем

$$G_\infty(x,x) = \exp\left\{ -\lambda \left(\frac{2(1-x)^{\alpha+1}}{\alpha+1} - \frac{(1-x)^{2\alpha+1}}{(2\alpha+1)} \right) \right\}.$$

Для функции, обратной частной функции распределения, имеем

$$-\frac{\lambda}{\alpha+1}(1-x)^{\alpha+1} = \ln u,$$

$$1-x = \left(-\frac{(\alpha+1)\ln u}{\lambda} \right)^{\frac{1}{\alpha+1}},$$

соответственно,

$$G^{-1}_{1,\infty}(u) = 1 - \left(-\frac{(\alpha+1)\ln u}{\lambda} \right)^{\frac{1}{\alpha+1}}.$$

Следовательно, копула при $u = v$ будет определяться выражением

$$C_\infty(u,u) = G_\infty(G^{-1}_{1,\infty}(u), G^{-1}_{1,\infty}(u)) =$$

$$= \exp\left\{ -\lambda \left[-\frac{2\ln u}{\lambda} - \frac{1}{(2\alpha+1)} \left(-\frac{(\alpha+1)\ln u}{\lambda} \right)^{\frac{2\alpha+1}{\alpha+1}} \right] \right\} =$$

$$= u^2 \exp\left\{ \frac{\lambda}{2\alpha+1} \left(-\frac{(\alpha+1)\ln u}{\lambda} \right)^{\frac{2\alpha+1}{\alpha+1}} \right\} =$$

$$= u^2 \exp\left\{ \left[-\left(\frac{\lambda}{2\alpha+1} \right)^{\frac{\alpha+1}{2\alpha+1}} \frac{\alpha+1}{\lambda} \ln u \right]^{\frac{2\alpha+1}{\alpha+1}} \right\} =$$

$$= u^2 \exp\left\{ \left[\ln u^{-\frac{\alpha+1}{\lambda}\left(\frac{\lambda}{2\alpha+1}\right)^{\frac{\alpha+1}{2\alpha+1}}} \right]^{\frac{2\alpha+1}{\alpha+1}} \right\} =$$

$$= u^2 u^{\frac{(\alpha+1)^2 \ln u}{\lambda(2\alpha+1)}\left(-\frac{\lambda}{(\alpha+1)\ln u}\right)^{\frac{1}{\alpha+1}}}.$$

Тогда для коэффициента Бломквиста справедливо

$$\beta_C = 4C_\infty\left(\frac{1}{2},\frac{1}{2}\right) - 1 = \exp\left\{ \frac{\lambda}{2\alpha+1} \left(-\frac{(\alpha+1)\ln\frac{1}{2}}{\lambda} \right)^{\frac{2\alpha+1}{\alpha+1}} \right) - 1 =$$

$$= \exp\left\{\frac{\lambda}{2\alpha+1}\left(\frac{(\alpha+1)\ln 2}{\lambda}\right)^{1+\frac{\alpha}{\alpha+1}}\right\} - 1 =$$

$$= \exp\left\{\ln 2 \cdot \frac{\alpha+1}{2\alpha+1}\left(\frac{\lambda}{(\alpha+1)\ln 2}\right)^{-\frac{\alpha}{\alpha+1}}\right\} - 1 =$$

$$= 2^{\frac{\alpha+1}{2\alpha+1}\left(\frac{\lambda}{(\alpha+1)\ln 2}\right)^{-\frac{\alpha}{\alpha+1}}} - 1.$$

□

Теперь введем следующие функции распределения

$$\begin{aligned} F_{1,T}(x) &= 1 - \frac{1}{\mu_{1,T}}\int_0^T \bar{B}_1(t+x)dt, \\ F_{2,T}(y) &= 1 - \frac{1}{\mu_{2,T}}\int_0^T \bar{B}_2(t+y)dt, \end{aligned} \tag{5.8}$$

где

$$\mu_{1,T} = \int_0^T \bar{B}_1(t)dt, \quad \mu_{2,T} = \int_0^T \bar{B}_2(t)dt.$$

Тогда можем записать

$$G_1(x) = \exp\left\{-\lambda\mu_{1,T}(1 - F_{1,T}(x))\right\},$$

$$G_2(y) = \exp\left\{-\lambda\mu_{2,T}(1 - F_{2,T}(y))\right\},$$

и, соответственно,

$$G_T(x,y) = D_T(x,y)\exp\big\{-\lambda\mu_{1,T}\bar{F}_{1,T}(x)\big\}\exp\big\{-\lambda\mu_{2,T}\bar{F}_{2,T}(x)\big\}. \tag{5.9}$$

Далее сформулируем предельную теорему общего характера в условиях большой загрузки.

Теорема 5.2. *Пусть функции $F_{i,T}$ из (5.8) для системы с разделением и параллельным обслуживанием, состоящей из двух подсистем с бесконечным числом приборов с пуассоновским входящим потоком с параметром λ, принадлежат области притяжения максимум-устойчивых распределений H_i, т.е. существуют нормирующие константы $a_i(s) > 0$ и $b_i(s)$, $s > 0$, $i = 1, 2$ такие что*

$$F_{1,T}^s(a_1(s)x + b_1(s)) \to H_1(x), \quad F_{2,T}^s(a_2(s)x + b_2(s)) \to H_2(y), \quad s \to \infty,$$

то тогда

$$G_T(a_1(\lambda)x+b_1(\lambda),a_2(\lambda)y+b_2(\lambda))\to H_1^{\mu_{1,T}}(x)H_2^{\mu_{2,T}}(y),\quad \lambda\to\infty.$$

Доказательство. Из теоремы 2 в [40] следует, что

$$G_{1,T}(a_1(\lambda)x+b_1(\lambda))\to H_1^{\mu_{1,T}}(x),\quad G_{2,T}(a_2(\lambda)y+b_2(\lambda))\to H_2^{\mu_{2,T}}(y),\quad \lambda\to\infty,$$

в силу того, что

$$\bar F_{1,T}(a_1(\lambda)x+b_1(\lambda))\sim-\frac{\ln H_1(x)}{\lambda},\quad \bar F_{2,T}(a_2(\lambda)y+b_2(\lambda))\sim-\frac{\ln H_2(y)}{\lambda},\quad \lambda\to\infty.$$

Теперь рассмотрим множитель $D_T(x,y)$. Поскольку при $x,y\to+\infty$

$$\int_0^T\bar B_1(t+x)\bar B_2(t+y)dt=o\bigg(\int_0^T\bar B_1(t+x)dt\bigg)$$

или

$$\int_0^T\bar B_1(t+x)\bar B_2(t+y)dt=o\bigg(\int_0^T\bar B_2(t+y)dt\bigg)$$

в силу того, что

$$\int_0^T\bar B_1(t+x)\bar B_2(t+y)dt\leqslant\bar B_2(y)\int_0^T\bar B_1(t+x)dt$$

и

$$\int_0^T\bar B_1(t+x)\bar B_2(t+y)dt\leqslant\bar B_1(x)\int_0^T\bar B_1(t+y)dt.$$

Следовательно, имеем

$$D_T(a_1(\lambda)x+b_1(\lambda),a_2(\lambda)y+b_2(\lambda))\to\exp\bigg\{\lambda o\bigg(\int_0^T\bar B_1(t+a_1(\lambda)x+b_1(\lambda))dt\bigg)\bigg\}=$$

$$=\exp\bigg\{\lambda o\bigg(\frac{\ln G_{1,T}^{-1}(a_1(\lambda)x+b_1(\lambda))}{\lambda}\bigg)\bigg\}\to1,\quad\lambda\to\infty.$$

□

Далее рассмотрим случай бесконечного среднего времени обслуживания (при степенных хвостах).

Теорема 5.3. *Если для функций распределения $B_1(x)$ и $B_2(y)$ времен обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с пуассоновским входящим потоком с параметром λ выполняется*

$$\bar{B}_1(x) \sim c_1 x^{-\alpha_1}, \quad \bar{B}_2(y) \sim c_2 y^{-\alpha_2}, \qquad x, y \to \infty, \quad c_1, c_2 > 0, \quad \alpha_1, \alpha_2 \in (0, 1],$$

тогда при $0 < \alpha_1, \alpha_2 < 1$ справедливо для функции распределения максимумов остаточных времен обслуживания на момент времени T

$$G_T(T^{1/\alpha_1} x, T^{1/\alpha_2} y) \to \exp\{-c_1 \lambda x^{-\alpha_1}\} \exp\{-c_2 \lambda y^{-\alpha_2}\}, \quad x, y > 0, \quad T \to \infty$$

а при $\alpha_1 = \alpha_2 = 1$ выполняется

$$G_T(Tx, Ty) \to \left(\frac{x}{x+1}\right)^{c_1 \lambda} \left(\frac{y}{y+1}\right)^{c_2 \lambda} \quad x, y > 0, \quad T \to \infty.$$

Доказательство. Сперва рассмотрим случай, когда $0 < \alpha_1, \alpha_2 < 1$. Из теоремы 3 в [40] следует, что

$$G_{1,T}(T^{1/\alpha_1} x) \to \exp\{-c_1 \lambda x^{-\alpha_1}\}, \quad x > 0, T \to \infty,$$

$$G_{2,T}(T^{1/\alpha_2} y) \to \exp\{-c_2 \lambda y^{-\alpha_2}\}, \quad y > 0, T \to \infty.$$

Поскольку $G_T(x, y) = D_T(x, y) G_{1,T}(x) G_{2,T}(y)$, остается рассмотреть только выражение

$$\begin{aligned}
\int_0^T \bar{B}_1(t + T^{1/\alpha_1} x) \bar{B}_2(t + T^{1/\alpha_2} y) dt &\sim c_1 c_2 \int_0^T (t + T^{1/\alpha_1} x)^{-\alpha_1} (t + T^{1/\alpha_2} y)^{-\alpha_2} dt = \\
&= \frac{c_1 c_2}{x^{\alpha_1} y^{\alpha_2} T^2} \int_0^T \left(1 + \frac{t}{T^{1/\alpha_1} x}\right)^{-\alpha_1} \left(1 + \frac{t}{T^{1/\alpha_2} y}\right)^{-\alpha_2} dt \sim \\
&\sim \frac{c_1 c_2}{x^{\alpha_1} y^{\alpha_2} T^2} \int_0^T \left(1 - \frac{\alpha_1 t}{T^{1/\alpha_1} x}\right) \left(1 - \frac{\alpha_2 t}{T^{1/\alpha_1} y}\right) dt = \\
&= \frac{c_1 c_2}{x^{\alpha_1} y^{\alpha_2} T^2} \left(T - \frac{\alpha_1 T}{2 T^{1/\alpha_1} x} - \frac{\alpha_2 T}{2 T^{1/\alpha_2} y} + \frac{\alpha_1 \alpha_2 T^3}{3 T^{1/\alpha_1} T^{1/\alpha_2} xy}\right) = \\
&= \frac{c_1 c_2}{x^{\alpha_1} y^{\alpha_2}} \left(\frac{1}{T} - \frac{\alpha_1}{2 T^{1/\alpha_1} x} - \frac{\alpha_2}{2 T^{1/\alpha_2} y} + \frac{\alpha_1 \alpha_2}{3 T^{1/\alpha_1 + 1/\alpha_2 - 1} xy}\right) \to 0, \quad T \to \infty.
\end{aligned}$$

Таким образом, получаем, что множитель $D_T(T^{1/\alpha_1}x, T^{1/\alpha_2}y) \to 1$, $T \to \infty$.

Аналогично, при $\alpha_1 = \alpha_2 = 1$ согласно теореме 3 из [40] следует, что

$$G_{1,T}(Tx) \to \left(\frac{x}{x+1}\right)^{c_1\lambda}, \quad x > 0, \quad T \to \infty,$$

$$G_{2,T}(Ty) \to \left(\frac{y}{y+1}\right)^{c_2\lambda}, \quad y > 0, \quad T \to \infty.$$

Как и раньше рассмотрим выражение

$$\int_0^T \bar{B}_1(t+Tx)\bar{B}_2(t+Ty)dt \sim c_1c_2 \int_0^T (t+Tx)^{-1}(t+Ty)^{-1}dt =$$

$$= \frac{c_1c_2}{T(x-y)} \int_0^T \left(\frac{1}{t+Ty} - \frac{1}{t+Tx}\right)dt =$$

$$= \frac{c_1c_2}{T(x-y)} \Big(\ln(t+Ty) - \ln(t+Tx) \Big) \Big|_0^T =$$

$$= \frac{c_1c_2}{T(x-y)} \ln \frac{(1+y)x}{(1+x)y} \to 0, \quad T \to \infty.$$

Следовательно, $D_T(T^{1/\alpha_1}x, T^{1/\alpha_2}y) \to 1$, $T \to \infty$. Таким образом, получаем, что и требовалось доказать. □

5.3 Случай интенсивности, заданной функцией от времени

Теперь рассмотрим случай, когда интенсивность входящего потока задана ограниченной неотрицательной функцией $\lambda(t)$, $t \geqslant 0$, т. е. имеет место нестационарный пуассоновский поток. Тогда будет справедлива следующая теорема.

Теорема 5.4. *Для системы с разделением и параллельным обслуживанием с нестационарным 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 — $B_2(y)$, выражение для совместной функции распределения максимальных остаточных времен обслуживания на момент времени T, $0 < T < \infty$, имеет вид

$$G_T(x,y) = \exp\Big\{ -\int_0^T \lambda(t)\big(1 - B_1(T-t+x)B_2(T-t+y)\big)dt\Big\}. \qquad (5.10)$$

Доказательство. Будем использовать свойство нестационарного пуассоновского потока, которое заключается в том, что при известном числе k заявок, поступивших на отрезке времени $[0,T]$, моменты их поступления (без учета порядка) независимы и имеют плотность распределения $\lambda(t)/\Lambda(T)$, $t \in [0,T]$, где

$$\Lambda(T) = \int_0^T \lambda(t)dt.$$

Если заявка поступила в момент времени t, то вероятность обслуживания ее первой подзаявки до момента $(T+x)$, а второй – до момента $(T+y)$ имеет распределение $B_1(T-t+x)B_2(T-t+y)$. Обслуживание заявок происходит независимо, следовательно, имеем

$$G_T(x,y) = \sum_{k=0}^{\infty} \frac{(\Lambda(T))^k}{k!} e^{-\Lambda(T)} \left(\frac{1}{\Lambda(T)} \int_0^T \lambda(t) B_1(T-t+x) B_2(T-t+y)dt \right)^k =$$

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

□

Следствие 5.8. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с нестационарным пуассоновским входящим потоком с интенсивностью $\lambda(t)$ имеют показательное распределение, т. е.*

$$B_1(x) = 1 - e^{-\beta_1 x}, \quad B_2(y) = 1 - e^{-\beta_2 y}, \quad x, y \geqslant 0, \quad \beta_1, \beta_2 > 0, \qquad (5.11)$$

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

$$G_T(x,y) = \exp\Big\{ -A_1(T)e^{-\beta_1 x} - A_2(T)e^{-\beta_2 y} + A_3(T)e^{-\beta_1 x - \beta_2 y}\Big\}, \quad x, y \geqslant 0,$$

где

$$A_1(T) = \int_0^T \lambda(T-t)e^{-\beta_1 t}dt, \quad A_2(T) = \int_0^T \lambda(T-t)e^{-\beta_2 t}dt,$$

$$A_3(T) = \int_0^T \lambda(T-t)e^{-(\beta_1+\beta_2)t}dt.$$

Доказательство. Для начала немного преобразуем формулу 5.10, сделав замену переменной

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

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

Теперь, подставляя конкретные выражения для $B_1(t+x)$ и $B_2(t+y)$ получаем требуемое

$$G_T(x,y) = \exp\Big\{ -\int_0^T \lambda(T-t)\big(e^{-\beta_1 x}e^{-\beta_1 t} + e^{-\beta_2 y}e^{-\beta_2 t} - e^{-\beta_1 x-\beta_2 y}e^{-(\beta_1+\beta_2)t}\big)dt\Big\} =$$

$$= \exp\Big\{ -e^{-\beta_1 x}\int_0^T \lambda(T-t)e^{-\beta_1 t} - e^{-\beta_2 y}\int_0^T \lambda(T-t)e^{-\beta_2 t} +$$

$$+e^{-\beta_1 x-\beta_2 y}\int_0^T \lambda(T-t)e^{-(\beta_1+\beta_2)t}dt\Big\}.$$

□

5.4 Случай интенсивности, заданной случайным процессом

Далее проанализируем ситуацию, когда интенсивность входящего потока задана случайным процессом $\lambda(t)$, $t \geqslant 0$, т. е. имеет место дважды стохастический пуассоновский поток. Тогда будет верна нижеследующая теорема.

Теорема 5.5. *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 — $B_2(y)$, выражение для совместной функции распределения максимальных остаточных времен обслуживания на момент времени T, $0 < T < \infty$, имеет вид

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

Доказательство. Доказательство данной теоремы следует из теоремы 5.4, если провести усреднение по всем возможным траекториям процесса $\lambda(t)$. □

Сформулируем следующую теорему.

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*

$$\lambda(t) = \max\{\gamma(t), 0\}, \quad \gamma(t) = \lambda_0 + \sigma\xi(t), \quad \lambda_0, \sigma > 0, \tag{5.12}$$

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

$$G_T(x,y) \leqslant \exp\Big\{-\lambda_0\int_0^T \big(1 - B_1(t+x)B_2(t+y)\big)dt + \\ + \frac{\sigma^2}{2}\int_0^T\int_0^T R(u-v)\big(1 - B_1(u+x)B_2(u+y)\big)\big(1 - B_1(v+x)B_2(v+y)\big)dudv\Big\}. \tag{5.13}$$

Доказательство. Введем случайную величину

$$\begin{aligned}\zeta &= -\int_0^T \gamma(t)\big(1 - B_1(T-t+x)B_2(T-t+y)\big)dt = \\ &= -\int_0^T \gamma(T-t)\big(1 - B_1(t+x)B_2(t+y)\big)dt.\end{aligned} \tag{5.14}$$

Эта случайная величина будет иметь нормальное распределение с параметрами

$$\mathbf{E}\zeta = -\lambda_0 \int_0^T \big(1 - B_1(t+x)B_2(t+y)\big)dt,$$

$$\mathbf{D}\zeta = \sigma^2 \int_0^T \int_0^T R(u-v)\big(1 - B_1(u+x)B_2(u+y)\big)\big(1 - B_1(v+x)B_2(v+y)\big)dudv.$$

Далее, воспользовавшись формулой для математического ожидания экспоненты от нормальной случайной величины, получаем, что

$$\mathbf{E}e^{\zeta} = \exp\big\{\mathbf{E}\zeta + \frac{1}{2}\mathbf{D}\zeta\big\}.$$

Поскольку интенсивность входящего потока не может принимать отрицательных значений, рассмотрим случайную величину ζ_1, которая получается из выражения (5.14) заменой $\gamma(t)$ на $\lambda(t) = \max\{\gamma(t), 0\}$. А из того, что $\zeta_1 \leqslant \zeta$, следует $G_T(x,y) = \mathbf{E}e^{\zeta_1} \leqslant \mathbf{E}e^{\zeta}$. □

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 (с оценкой погрешности в одномерном случае можно ознакомиться в [40].

Следствие 5.9. *Если времена обслуживания заявок на приборах каждой из двух подсистем с бесконечным числом приборов системы с разделением и параллельным обслуживанием с нестационарным пуассоновским входящим потоком с интенсивностью* $\lambda(t)$ *из* (5.12) *имеют показательное распределение, т. е.*

$$B_1(x) = 1 - e^{-\beta_1 x}, \quad B_2(y) = 1 - e^{-\beta_2 y}, \quad x, y \geqslant 0, \quad \beta_1, \beta_2 > 0, \tag{5.15}$$

то тогда для функция распределения максимумов остаточных времен обслу-

живания на момент времени T справедливо следующее неравенство

$$G_T(x,y) \leqslant \exp\Big\{ - A_1(T)e^{-\beta_1 x} - A_2(T)e^{-\beta_2 y} + A_3(T)e^{-\beta_1 x-\beta_2 y} + \\ + D_1(T)e^{-2\beta_1 x} + D_2(T)e^{-2\beta_2 y} + D_3(T)e^{-\beta_1 x-\beta_2 y} - \\ - D_4(T)e^{-2\beta_1 x-\beta_2 y} - D_5(T)e^{-\beta_1 x-2\beta_2 y} + D_6(T)e^{-2\beta_1 x-2\beta_2 y}\Big\}, \quad x,y \geqslant 0,$$

где

$$A_1(T) = \frac{\lambda_0}{\beta_1}(1-e^{-\beta_1 T}), \quad A_2(T) = \frac{\lambda_0}{\beta_2}(1-e^{-\beta_2 T}),$$

$$A_3(T) = \frac{\lambda_0}{\beta_1+\beta_2}(1-e^{-(\beta_1+\beta_2)T}),$$

$$D_1(T) = \frac{\sigma^2}{2}\int_0^T\int_0^T R(u-v)e^{-\beta_1(u+v)}dudv,$$

$$D_2(T) = \frac{\sigma^2}{2}\int_0^T\int_0^T R(u-v)e^{-\beta_2(u+v)}dudv,$$

$$D_3(T) = \frac{\sigma^2}{2}\int_0^T\int_0^T R(u-v)\big(e^{-(\beta_1 u+\beta_2 v)} + e^{-(\beta_2 u+\beta_1 v)}\big)dudv,$$

$$D_4(T) = \frac{\sigma^2}{2}\int_0^T\int_0^T R(u-v)\big(e^{-(\beta_1 u+(\beta_1+\beta_2)v)} + e^{-((\beta_1+\beta_2)u+\beta_1 v)}\big)dudv,$$

$$D_5(T) = \frac{\sigma^2}{2}\int_0^T\int_0^T R(u-v)\big(e^{-(\beta_2 u+(\beta_1+\beta_2)v)} + e^{-((\beta_1+\beta_2)u+\beta_2 v)}\big)dudv,$$

$$D_6(T) = \frac{\sigma^2}{2}\int_0^T\int_0^T R(u-v)\big(e^{-(\beta_1+\beta_2)(u+v)}\big)dudv.$$

Доказательство. Подставив конкретные выражения для $B_1(t+x)$ и $B_2(t+y)$ в (5.13), получим требуемое. □

Далее если допустить, что, например, $R(t) = e^{-\delta|t|}$, $\delta > 0$, тогда получим следующее

$$D_1(\infty) = \frac{\sigma^2}{2\beta_1(\delta+\beta_1)}, \quad D_2(\infty) = \frac{\sigma^2}{2\beta_2(\delta+\beta_2)},$$

$$D_3(\infty) = \sigma^2\frac{\beta_1+\beta_2+2\delta}{(\delta+\beta_1)(\delta+\beta_2)(\beta_1+\beta_2)},$$

$$D_4(\infty) = \sigma^2\frac{2\beta_1+\beta_2+2\delta}{(\delta+\beta_1)(\delta+\beta_1+\beta_2)(2\beta_1+\beta_2)},$$

$$D_5(\infty) = \sigma^2 \frac{\beta_1 + 2\beta_2 + 2\delta}{(\delta + \beta_2)(\delta + \beta_1 + \beta_2)(\beta_1 + 2\beta_2)},$$

$$D_6(\infty) = \frac{\sigma^2}{2(\delta + \beta_1 + \beta_2)(\beta_1 + \beta_2)}.$$

5.5 Выводы к главе 5

В данной Главе изучено максимальное остаточное время обслуживания в бесконечнолинейных системах типа $M^{(2)}|G_2|\infty$, понимаемое как максимум остаточных времен обслуживания по всем занятым приборам на текущий момент времени или в стационарном режиме. С практической точки зрения это время, необходимое для корректного завершения работы системы после отключения входящего потока заявок. Рассмотрены различные случаи интенсивности входящего потока: 1) не зависящей от времени, 2) заданной функцией от времени, 3) заданной случайным процессом. В последнем случае конкретные результаты получены для стационарного гауссовского процесса.

В качестве распределений времен обслуживания рассмотрены показательное, гиперэкспоненциальное, Парето и равномерное. В случае постоянной интенсивности изучены эффекты, возникающие при степенных хвостах распределений. Для отдельных распределений найдены копула-функции и коэффициенты Бломквиста.

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 (на-

пример, 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.

Список литературы

[1] *Афанасьева Л.Г., Булинская Е.В.* Случайные процессы в теории массового обслуживания и управления запасами. М.: Изд-во МГУ, 1980. 110 с.

[2] *Благовещенский Ю.Н.* Основные элементы теории копул // Прикладная эконометрика. 2012. № 2. С. 113–130.

[3] *Бахарева Н.Ф.* Обобщенная двумерная диффузионная модель массового обслуживания типа $GI|G|1$ // Телекоммуникации. 2009. № 7. С. 2–8.

[4] *Бахарева Н.Ф., Тарасов В.Н., Коннов А.Л.* Декомпозиция сетей массового обслуживания без ограничений на длину очереди // Научно-технические ведомости СПбГПУ. 2008. № 2. С. 31–35.

[5] *Башарин Г.П., Бочаров П.П., Коган Я.А.* Анализ очередей в вычислительных сетях. Теория и методы расчета // Москва: Наука, 1989. 336 с.

[6] *Бочаров П.П., Печинкин А.В.* Теория массового обслуживания. М.: Изд-во РУДН, 1995. 529 с.

[7] *Вишневский В.М.* Теоретические основы проектирования компьютерных сетей. М: Техносфера, 2003. 512 с.

[8] *Вишневский В.М., Ефросинин Д.В.* Теория очередей и машинное обучение. М.: ИНФРА-М, 2024. 370 с.

[9] *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.

[10] *Вишневский В.М., Горбунова А.В.* Применение методов машинного обучения к решению задачи теории массового обслуживания // Труды 20-й Международной конференции им. А. Ф. Терпугова “Информационные технологии и математическое моделирование”. 2022. С. 167–172.

[11] *Вишневский А.В., Дудин А.Н., Клименок В.И.* Стохастические системы с коррелированными потоками. Теория и применение в телекоммуникационных сетях. М: Техносфера, 2018. 564 с.

[12] *Вишневский В.М., Клименок В.И., Соколов А.М., Ларионов А.А.* Исследование fork-join системы с марковским входным потоком и распределением времени обслуживания фазового типа // Информационные технологии и вычислительные системы. 2023. № 4. С. 29–56.

[13] *Гиндин С.И., Хомоненко А.Д., Ададуров С.Е.* Численный расчет многоканальной системы массового обслуживания с рекуррентным входящим потоком и «Разогревом» // Известия Петербургского университета путей сообщения. 2013. Т. 37, № 4. С. 92–101.

[14] *Глухова Е.В., Орлов А.Б.* Средняя длительность периода занятости бесконечно линейных систем массового обслуживания с дважды стохастическим входящим потоком // Изв. вузов. Физика. 2003. № 3. С. 62–68.

[15] *Горбань А.Н.* Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей // Сибирский журнал вычислительной математики. 1998. Т. 1, № 1. С. 11–24.

[16] *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.

[17] *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.

[18] *Горбунова А.В.* Оценки копулы и квантилей распределения времени отклика системы с разделением и параллельным обслуживанием заявок и распределением Парето времени обслуживания // Управление большими системами: сборник трудов. 2024. № 112. С. 7–29.

[19] *Горбунова А.В.* О подходе к управлению скоростью обслуживания в системах с разделением и параллельным обслуживанием заявок // Материалы X Всероссийской научно-практической конференции “Современные проблемы физико-математических наук” (СПФМН-2024, Орел). Орел: ОГУ имени И.С. Тургенева, 2024. С. 173–179.

[20] *Горбунова А.В.* Анализ моделей массового обслуживания для оценки времени отклика в системе облачных вычислений: дис. ... канд. физ.-мат. наук.- М., 2017.

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

[22] *Горбунова А.В., Зарядов И.С., Матюшенко С.И., Самуйлов К.Е.* Аппроксимация времени отклика системы облачных вычислений // Информатика и её применения. 2015. Т. 9, Вып. 3. С. 32–38.

[23] *Горбунова А.В., Зарядов И.С., Самуйлов К.Е., Сопин Э.С.* Обзор систем параллельной обработки заявок // Вестник Российского университета дружбы народов. Серия: Математика, информатика, физика. 2017. Т. 25, № 4. С. 350–362.

[24] *Горбунова А.В., Зарядов И.С., Самуйлов К.Е.* Обзор систем параллельной обработки заявок. Часть II // Вестник Российского университета дружбы народов. Серия: Математика, информатика, физика. 2018. Т. 26, № 1. С. 13–27.

[25] *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.

[26] *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.

[27] *Жидкова Л.А., Моисеева С.П.* Исследование системы параллельного обслуживания кратных заявок простейшего потока // Вест. Том. политехн. ун-та. Управление, вычислительная техника и информатика. 2011. Т. 17, № 4. С. 49–54.

[28] *Задорожный В.Н., Захаренкова Т.Р.* Методы планирования имитационных экспериментов при моделировании фрактальных очередей // Омский научный вестник. Сер. Приборы, машины и технологии. №. 3. 2016. С. 87–92.

[29] *Иванова Н.М., Вишневский В.М.* Оценка надежности привязных высотных беспилотных платформ с использованием моделей систем k-из-n и методов машинного обучения // Проблемы информатики. 2021. Т. 53, № 4. С.16–39.

[30] *Ивановская И.А., Моисеева С.П.* Исследование математической модели параллельного обслуживания заявок смешанного типа // Изв. Том. политехн. ун-та. Управление, вычислительная техника и информатика. 2010. Т. 317, № 5. С. 32–34.

[31] *Ивлев В.В.* Неопределенности функций многих переменных (часть I) // Матем. обр. 2002. Вып. 4. С. 90–100.

[32] *Ивлев В.В.* Неопределенности функций многих переменных (часть II) // Матем. обр. 2003. Вып. 3. С. 77–85.

[33] *Клименок В.И., Тарамин О.С.* Двухфазная система $GI|PH|1 \to |PH|1|0$ с потерями // Автоматика и телемеханика. 2011. Вып. 5. С. 113–126.

[34] *Клименок В.И.* Характеристики производительности системы массового обслуживания с расщеплением запросов // Информатика. 2023. Т. 20. С. 50–60.

[35] *Курош А.Г.* Алгебраические уравнения произвольных степеней (Популярные лекции по математике; вып. 7). Изд. 2-е. М.: Наука, 1975. 32 с.

[36] *Курош А.Г.* Курс высшей алгебры: Учебник. Изд. 17-е. СПб.: Лань. 2008. 432 с.

[37] *Лебедев А.В.* Асимптотика максимумов в бесконечнолинейной системе с ограниченным размером групп // Фундамент. и прикл. матем. 1996. Т. 2, № 4. С. 1107–1115.

[38] *Лебедев А.В.* Экстремумы некоторых процессов массового обслуживания: Дис. ... канд. физ.-мат. наук: 01.01.05. М., 1997.

[39] *Лебедев А.В.* Максимумы в системе $M^{[X]}|G|\infty$ с "тяжелыми хвостами" размеров групп // Автомат. и телемех. 2000. № 12. С. 115–121.

[40] *Лебедев А.В.* Максимальное остаточное время обслуживания в бесконечнолинейных системах // Проблемы передачи информации. 2018. Т. 54, Вып. 2. С. 86–102.

[41] *Моисеева С.П., Захорольная И.А.* Математическая модель параллельного обслуживания кратных заявок с повторными обращениями // Автометрия. 2011. Т. 47. № 6. С. 51–58.

[42] *Моисеева С.П., Панкратова Е.В., Убонова Е.Г.* Исследование бесконечнолинейной системы массового обслуживания с разнотипным обслуживанием и входящим потоком марковского восстановления // Вест. Том. политехн. ун-та. Управление, вычислительная техника и информатика. 2016. Т. 35. № 2. С. 46–53.

[43] *Орлов А.Б.* Плотность вероятностей максимального остаточного времени обслуживания на занятых приборах // Вычисл. технол. Спец. вып. 5: Избр. докл. VI Междунар. научн.-практ. конф. “Информационные технологии и математическое моделирование”. 2008. Т 13. С. 93–98.

[44] *Орлов Ю.Н., Соловьев В.О.* Об оценке точности обработки больших потоков экспериментальных данных // Препринты ИПМ им. М.В. Келдыша. 2022. 083. 24 с.

[45] *Осипов О. А.* Система обслуживания с делением и слиянием требований, в которой требование занимает все свободные обслуживающие приборы // Вест. Росс. ун-та дружбы народов. Серия: Математика. Информатика. Физика. 2018. Т. 26. № 1. С. 28–38.

[46] *Постнова О.С., Тананко И.Е., Рогачко Е.С.* Приближенный анализ длительности пребывания требований в сети массового обслуживания с делением и слиянием требований // Управление большими системами. 2025. Т. 115. С. 33–51.

[47] *Редругина Н.М.* Метод вычисления временных характеристик обслуживания в сервисных платформах инфокоммуникационных транзакционных услуг с параллельной обработкой запросов // Тр. уч. заведений связи. 2023. Т. 9. № 3. С. 82–90.

[48] *Синякова И.А.* Математические модели и методы исследования систем

параллельного обслуживания сдвоенных заявок случайных потоков: Дис. ... канд. физ.-мат. наук: 05.13.18. Т., 2013.

[49] *Топорков В.В.* Модели распределенных вычислений. М.: ФИЗМАТЛИТ. 2004. 320 с.

[50] *Фантаццини Д.* Моделирование многомерных распределений с использованием копула-функций. I // Прикладная эконометрика. № 2 (22) 2011. 98-134.

[51] *Хабаров Р.С., Лохвицкий В.А., Дудкин А.С.* Аппроксимация времени пребывания для системы массового обслуживания fork-join на основе инвариантов отношения // Интеллектуальные технологии на транспорте. 2020. Т. 22. № 2. С. 46–50.

[52] *Хомоненко А.Д., Яковлев, Е.Л.* Нейросетевая аппроксимация характеристик многоканальных немарковских систем массового обслуживания // Труды СПИИРАН. 2015. Вып. 41, № 4. С. 81–93.

[53] *Чернавская Е.А.* Предельные теоремы для бесконечноканальных систем с тяжелыми хвостами распределений времен обслуживания: Дис. ... канд. физ.-мат. наук: 01.01.05. М., 2017.

[54] *Ageyev D., Mohsin A., Radivilova T., Kirichenko L.* Infocommunication networks design with self-similar traffic // Proc. IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM). 2019. P. 24–27.

[55] *Agresti A.* Analysis of Ordinal Categorical Data. New York: John Wiley & Sons. 2010. 419 p.

[56] *Ahad N., Qadir Ju., Ahsan N.* Neural networks in wireless networks:

Techniques, applications and guidelines // Journal of Network and Computer Applications. 2016. Vol. 68. P. 1–27.

[57] *Akbas A., Yildiz H., Ozbayoglu A., Tavli B.* Neural network based instant parameter prediction for wireless sensor network optimization models // Wireless Network. 2019. Vol. 25. P. 3405–3418.

[58] *Akyildiz I.F.* General closed queueing networks with blocking // Proceedings of the 12th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation. 1988. P. 283–303.

[59] *Akyildiz I.F., Brand H.* Exact solutions for networks of queues with blocking-after-service // Theoretical Computer Science. 1994. Vol. 125, No. 1. P. 111–130.

[60] *Alkasem A., Liu H.* A Survey of Fault-tolerance in Cloud Computing: Concepts and Practice // Research Journal of Applied Sciences, Engineering and Technology. 2015. Vol. 11, No. 12. P. 1365–1377.

[61] *Andrews G.R.* Foundations of Multithreaded, Parallel, and Distributed Programming. B.: Addison Wesley, 1999. 688 p.

[62] *Armony M., Israelit S., Mandelbaum A., Marmor Y.N., Tseytlin Y., Yom-Tov G.B.* Patient flow in hospitals: a data-based queueing-science perspective // Stochastic Systems. 2015. Vol. 5, No. 1. P. 146–194.

[63] *Atar R., Mandelbaum A., Zviran A.* Control of fork-join networks in heavy traffic // 50th Annual Allerton Conference on Communication, Control, and Computing, Allerton. 2012. P. 823–830.

[64] *Aussem A., Murtagh F.* A neuro-wavelet strategy for Web traffic forecasting // Research in Oficial Statistics. 1998. Vol. 1, No. 1. P. 65–87.

[65] *Aussem A., Rouxel S., Marie R.* Neural-based queueing system modeling for service quality estimation in B-ISDN networks // Proceedings of the ninth International Conference on Artificial Neural Networks (ICANN'99). 1999. Vol. 2, No 470. P. 970–975.

[66] *Aussem A., Rouxel S., Marie R.* Neural-based queueing system modeling for service quality estimation in B-ISDN networks // Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN. Neural Computing: New Challenges and Perspectives for the New Millennium. 2000. Vol. 5. P. 392–397.

[67] *Baccelli F., Makowski A.M.* Queueing models for systems with synchronization constraints // Proceedings of the IEEE. 1989. Vol. 77, No. 1. P. 138-161.

[68] *Baldwin R., Davis I.V.N., Midkiff S., Kobza J.* Queueing network analysis: concepts, terminology, and methods // Journal of Systems and Software. 2003. Vol. 66, No 2. P. 99–117.

[69] *Balsamo S., Clò M.* A convolution algorithm for product-form queueing networks with blocking // Annals of Operations Research. 1998. Vol. 79. P. 97–117.

[70] *Balsamo S., Donatiello L.* On the cycle time distribution in a two-stage cyclic network with blocking // IEEE Transactions on Software Engineering. 1989. Vol. 15, No. 11. P. 1206–1216.

[71] *Balsamo S., Donatiello L., Van Dijk N.M.*, Bound Performance Models of Heterogeneous Parallel Processing Systems // IEEE Transactions on Parallel and Distributed Systems. 1998. Vol. 9, No. 10. P. 1041–1056.

[72] *Baskett F., Chandy K.M., Muntz R.R., Palacios F.G.* Open, closed and mixed

networks of queues with different classes of customers // Journal of the ACM. 1975. Vol. 22. P. 248–260.

[73] *Baynat B., Dallery Y.* A unified view of product-form approximation techniques for general closed queueing networks // Performance Evaluation. 1993. Vol. 18, No. 3. P. 205–224.

[74] *Bhatti M.A., Riaz M.A., Rizvi S.S., Shokat S., Riaz F., Kwon S.J.* Outlier detection in indoor localization and Internet of Things (IoT) using machine learning // Journal of Communications and Networks. 2020. Vol. 22, No. 3. P. 236–243.

[75] *Bolch G., Greiner S., Meer H., Trivedi K.S.* Queueing networks and Markov chains: Modeling and performance evaluation with computer science applications. John Wiley & Sons, 2006. 896 p.

[76] *Boon M.A.A., van der Mei M.A.A., Winands E.M.M.* Applications of polling systems // Surveys in Operations Research and Management Science. 2011. Vol. 16, No. 2. P. 67–82.

[77] *Boxma O.J., Donk P.* On response time and cycle time distribution in a two-stage cyclic queue // Performance Evaluation. 1982. Vol. 2, No. 3. P. 181–194.

[78] *Buzen J.P.* Computational algorithms for closed queueing networks with exponential servers // Communications of the ACM. 1973. Vol. 16, No. 9. P. 527–531.

[79] *Carbini S., Donatiello L., Iazeolla G.* An efficient algorithm for the cycle time distribution in two-stage cyclic queues with a non-exponential server // Proceedings of the International Seminar in Teletraffic Analysis and Computer Performance Evaluation, Amsterdam. 1986. P. 99–115.

[80] *Casilari E., Alfaro A., Reyes A., Diaz-Estrella A., Sandoval F.* Neural

modelling of Ethernet traffic over ATM networks // Proceedings of the fourth International Conference on Engineering Applications of Neural Networks (EANN'98). 1998. P. 304–307.

[81] *Chester D.L.* Why two hidden layers are better than one // Proc. International Joint Conference on Neural Networks. 1990. P. 265–268.

[82] *Choi B.K., Kang D.* Modeling and simulation of discrete-event systems, John Wiley & Sons Inc, Hoboken, New Jersey, 2013. 432 p.

[83] *Clò M. C.* MVA for product-form cyclic queueing networks with blocking // Annals of Operations Research. 1998. Vol. 79. P. 83–96.

[84] *Curtis C., Liu Ch, Bollerman Th.J., Pianykh O.S.* Machine learning for predicting patient wait times and appointment delays // Journal of the American College of Radiology. 2018. Vol. 15, No. 9. P. 1310–1316.

[85] *Cybenko G.* Aproximation by superpositions of a sigmoidal function // Math. Control Signal Systems. 1989. Vol. 24. P. 303–314.

[86] *Daduna H.* Two-stage cyclic queues with nonexponential servers: Steady-state and cyclic time // Operations Research. 1986. Vol. 34, No. 3. P. 455–459.

[87] *Dai J., Nguyen V., Reiman M.* Sequential bottleneck decomposition: an approximation method for generalised jackson networks // Operations Research. 1995. Vol. 42, No 1. P. 119–136.

[88] *David H.A., Nagaraja H.N.* Order Statistics. John Wiley & Sons, 2004. 488 p.

[89] *Dayar T., Meri A.* Kronecker representation and decompositional analysis of closed queueing networks with phase-type service distributions and arbitrary buffer sizes // Annals of Operations Research. 2008. Vol. 164, No 1. P. 193–210.

[90] *Dias L.M.S., Vieira A.A.C., Pereira G.A.B., Oliveira J.A.* Discrete Simulation Software Ranking — a Top list of the Worldwide most Popular

and Used Tools // Proceedings of the 2016 Winter Simulation Conference (WSC). 2016. P. 1060–1071.

[91] *Dudin A.N., Klimenok V.I., Vishnevsky V.M.* The theory of queuing systems with correlated flows. Springer, Heidelberg, Germany, 2020.

[92] *Efrosinin D., Stepanova N.* Estimation of the Optimal Threshold Policy in a Queue with Heterogeneous Servers Using a Heuristic Solution and Artificial Neural Networks // Mathematics. 2021. Vol. 9. Article Number: 1267.

[93] *Embrechts P., Klüppelberg C., Mikosch T.* Modelling Extremal Events for Insurance and Finance. Springer-Verlag Berlin Heidelberg, 1997. 644 p.

[94] *Enganti P., Rosenkrantz T., Sun L., Wang Z., Che H., Jiang H.* ForkMV: Mean-and-Variance Estimation of Fork-Join Queuing Networks for Datacenter Applications // Proc. IEEE International Conference on Networking, Architecture and Storage (NAS). 2022. P. 1–8

[95] *Fiorini P.M.* Analytic approximations of fork-join queues // Proceedings of IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). 2015. Vol. 2. P. 966–971.

[96] *Flatto L., Hahn S.* Two Parallel Queues Created by Arrivals with Two Demands I // SIAM Journal on Applied Mathematics. 1984. Vol. 44, No 5. Pp. 1041–1053.

[97] *Foresee F.D., Hagan M.T.* Gauss-Newton approximation to Bayesian learning // Proceedings of International Conference on Neural Networks (ICNN'97). 1997. Vol. 3. P. 1930–1935.

[98] *Gallien J., Wein L.M.* A Simple and Effective Component Procurement Policy

for Stochastic Assembly Systems // Queueing Systems. 2001. Vol. 38. P. 221–248.

[99] *Gelenbe E., Pujolle G.* The behaviour of a single queue in a general queueing network // Acta Informatica. 1976. Vol. 7, No 2. P. 123–136.

[100] *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.
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.

[101] *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.

[102] *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.

[103] *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.

[104] *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.

[105] *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.

[106] *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.
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.

[107] *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.

[108] *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.

[109] *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.

[110] *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.

[111] *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.

[112] *Gorbunova A.V., Zaryadov I.S., Matushenko S.I., Sopin E.S.* The Estimation of Probability Characteristics of Cloud Computing Systems with Splitting of

Requests // Communications in Computer and Information Science. 2016. Vol. 678. P. 418–429.

[113] *Gordon W.J., Newell G.F.* Closed Queuing Systems with Exponential Servers // Operations Research. 1967. Vol. 15. P. 254–265.

[114] *Gudendorf G., Segers J.* Extreme-Value Copulas // Copula theory and Its Application. 2010. Vol. 198. P. 127–145

[115] *Gupta R.D., Kundu D.* Generalized exponential distributions // Australian New Zealand J. Statist. 1999. Vol. 41, No. 2, P. 173–188.

[116] *Habib I.W.* Applications of neurocomputing in traffic management of ATM networks // Proceedings of the IEEE. 1996. Vol. 84, No. 10. P. 1430–1441.

[117] *Hagan M.T., Menhaj M.* Training feedforward networks with the Marquardt algorithm, IEEE Transactions on Neural Networks. 1994. Vol. 5, No. 6. P. 989–993.

[118] *Harrison P., Zertal S.* Queueing Models with Maxima of Service Times // Computer Performance Evaluation. Modelling Techniques and Tools. 2003. P. 152–168.

[119] *Hermanto R.P.S., Suharjito S., Nugroho A.* Waiting-time estimation in bank customer queues using RPROP neural networks // Procedia Computer Science. 2018. Vol. 135. P. 35–42.

[120] *Heyman D.* Ch. 11. Numerical methods in queueing theory, Handbook of Statistics, Elsevier, 2003. VoL. 21. P. 407–429.

[121] *Hiramatsu A.* Integration of ATM call admission control and link capacity control by distributed neural networks // IEEE Journal on Selected Areas in Communications. 1991. Vol. 9, No. 7. P. 1131–1138.

[122] *Hoyo-Alonso R., Fernández-de-Alarcón P., Navamuel-Castillo J.-J., Medrano-Marqués N.J., Martin-del-Brio B., Fernández-Navajas J., Abadía-Gallego D.* Neural networks for QoS network management // Lecture Notes in Computer Science. 2007. Vol. 4507. P. 887–894.

[123] *Howard R.A.* Dynamic Programming and Markov Processes. Technology Press of Massachusetts Institute of Technology and John Wiley & Sons, 1960. 136 p.

[124] *Jackson R.R.P.* Queueing systems with phase type service // Journal of the Operational Research Society. 1954. Vol. 5, No 4. P. 109–120.

[125] *Jackson J.R.* Networks of waiting lines // Operations Research. 1957. Vol. 5, No 4. P. 518–521.

[126] *Jackson J.R.* Job shop-like queueing systems // Management Science. 1963. Vol. 10, No 1. P. 131–142.

[127] *Jiang L., Giachetti R.E.* A queueing network model to analyze the impact of parallelization of care on patient cycle time // Health Care Management Science. 2008. Vol. 11. P. 248–261.

[128] *Jiang C., Zhang H., Ren Y., Han Z., Chen K.-C., Hanzo L.* Machine learning paradigms for next-generation wireless networks // IEEE Wireless Communications. 2017. Vol. 24, No. 2. P. 98–105.

[129] *Kaur R., Kaur Sandhu J., Sapra L.* Machine learning technique for wireless sensor networks // Proc. Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). 2020. P. 332–335.

[130] *Kelly F.P.* Networks of queues with customers of different types // Journal of Applied Probability. 1975. Vol. 12, No 3. P. 542–554.

[131] *Kelly F.P.* Networks of queues // Advances in Applied Probability. 1976. Vol. 8, No 2. P. 416–432.

[132] *Kelly K.S., Krzysztofowicz R. A* A bivariate meta-Gaussian density for use in hydrology // Stochastic Hydrology and Hydraulics. 1997. Vol. 11. P. 17–31.

[133] *Kemper B., Mandjes M.* Mean sojourn times in two-queue fork-join systems: bounds and approximations // OR Spectrum. 2012. Vol. 34. P. 723–742.

[134] *Kendall M.* A New Measure of Rank Correlation // Biometrika. 1938. Vol. 30, Issue 1-2. P. 81–93.

[135] *Khoshnevis B., Parisay S.* Machine learning and simulation: application in queuing systems // Simulation. 1993. Vol. 61, No. 5. P. 294–302.

[136] *Kingma D.P., Ba J.* Adam: A method for stochastic gradient descent // Proc. ICLR, San Diego, CA, USA. 2015. P. 1–15.

[137] *Klimenok V., Dudin A.,Vishnevsky V.* On the Stationary Distribution of Tandem Queue Consisting of a Finite Number of Stations // Communications in Computer and Information Science. 2012. Vol. 291. P. 383–392.

[138] *Kochenderfer M.J., Wheeler T.A.* Algorithms for Optimization. MIT Press, 2019. 500 p.

[139] *Kraemer W., Langenbach-Belz M.* Approximate formulae for the delay in the queueing system $GI|G|1$ // Proceedings of the 8th International Teletraffic Congress. 1976. Vol. 235. P. 1–8.

[140] *Kühn P.* Analysis of complex queuing networks by decomposition // Proceedings of the 8.th. International Teletraffic Congress. 1976.

[141] *Kühn P.J.* Approximate analysis of general queuing networks by decomposition // IEEE Trans. Comm.. 1979. Vol. 27, No 1. P. 113–126.

[142] *Kumari P., Kaur P.* A survey of fault tolerance in cloud computing // Journal of King Saud University – Computer and Information Sciences. 2018. Vol. 33. P.1159–1176.

[143] *Kyritsis A.I., Deriaz M.* A machine learning approach to waiting time prediction in queueing scenarios // Second IEEE International Conference on Artificial Intelligence for Industries. 2019. P. 17-21.

[144] *Laha S., Chowdhury N., Karmakar N.* How can machine learning impact on wireless network and IoT? – A survey // Proc. 1th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2020. P. 1–7.

[145] *Leadbetter M.R., Lindgren G., Rootzen H.* Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag New York Inc. 1983. 336 p.

[146] *Lebrecht A.S., Knottenbelt W.J.* Response Time Approximations in Fork-Join Queues // Proceedings of the 23rd Annual UK Performance Engineering Workshop (UKPEW'07), 2007.

[147] *Leland W.E., Taqqu M.S., Willinger W., Wilson D.V.* On the self-similar nature of Ethernet traffic (extended version) // IEEE/ACM Transactions on Networking. 1994. Vol. 2, No. 1. P. 1–15. https://doi.org/10.1109/90.282603

[148] *Li H., Gao H., Lv T., Lu Y.* Deep q-learning based dynamic resource allocation for self-powered ultra-dense networks // IEEE international conference on communications workshops (ICC Workshops). 2018. P. 1–6.

[149] *Luong N.C., Hoang N.C., Gong S., Niyato D., Wang D., Liang D.* Applications of deep reinforcement learning in communications and networking: A Survey // IEEE Communications Surveys & Tutorials. 2019. Vol. 21, No. 4. P. 3133–31741.

[150] *Mao Q., Hu F., Hao F.* Deep learning for intelligent wireless networks: A Comprehensive survey // IEEE Communications Surveys & Tutorials. 2018. Vol. 20, No. 4. P. 2595–2621.

[151] *Marie R.A.* An approximate analytical method for general queueing networks // Transactions on Software Engineering. 1979. Vol. 5, No 5. P. 530–538.

[152] *Marie R.A.* Calculating equilibrium probabilities for $\lambda(n)/c_k/1/n$ queues // Proceedings of the 1980 International Symposium on Computer Performance Modelling, Measurement and Evaluation. 1980. P. 117–125.

[153] *Marshall K.T.* Some Inequalities in Queuing // Operations Research. 1968. Vol. 16, No 3. P. 651–668.

[154] *McNeil A.J., Frey R., Embrechts P.* Quantitative risk management. Princeton University Press. 2015. 720 p.

[155] *Memon M.L., Maheshwari M.K., Saxena N., Roy A., Shin D.R.* Artificial intelligence-based discontinuous reception for energy saving in 5G networks // Electronics. 2019. Vol. 8, No. 7. Article Number: 778.

[156] *Merlo G., Britos P., Rossi B., García Martínez R.* Neural networks applied to automatic estimation of networks performance // Proceedings of the International Conference on Intelligent Systems and Control. 2004. P. 167–171.

[157] *Moller M.F.* A scaled conjugate gradient algorithm for fast supervised learning // Neural Networks. 1993. Vol. 6. P. 525–533.

[158] *Morshedi M., Noll J.* Estimating PQoS of video streaming on Wi-Fi networks using machine learning // Sensors. 2021. Vol. 21, No. 2. Article Number: 621.

[159] *Mourõo R.N., Carvalho R.S., Carvalho R.N., Ramos G.N.* Predicting waiting time overflow on bank teller queues // Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA). 2018. P 842–847.

[160] *Narahari Y., Sundarrajan P.* Performability Analysis of Fork-Join Queueing Systems // The Journal of the Operational Research Society. 1995. Vol. 46, No. 10. P. 1237–1249.

[161] *Nelder J.A., Mead R.* A Simplex Method for Function Minimization, Computer Journal. 1965. Vol. 7. P. 308–313.

[162] *Nelsen R.* An introduction to copulas. Berlin, Germany: Springer, 2006. 272 p.

[163] *Nelson R., Tantawi A.N.* Approximate analysis of fork/join synchronization in parallel queues // IEEE Transactions on Computers. 1988. Vol. 37, No. 6. P. 739–743.

[164] *Nguyen M., Alesawi S., Li N., Che H., Jiang H.* ForkTail: A black-box fork-join tail latency prediction model for user-facing datacenter workloads // HPDC '18: Proceedings of the 27th International Symposium on High-Performance Parallel and Distributed Computing. 2018. P. 206–217.

[165] *Nguyen M., Alesawi S., Li N., Che H., Jiang H.* A Black-Box Fork-Join Latency Prediction Model for Data-Intensive Applications // IEEE Transactions on Parallel and Distributed Systems. 2020. Vol. 31, No. 9, P. 1983–2000

[166] *Nie L., Jiang L., Yu S., Song H.* Network traffic prediction based on deep belief network in wireless mesh backbone networks // IEEE Wireless Communications and Networking Conference (WCNC). 2017. P. 1–5.

[167] *Nordstrom E., Carlstrom J., Gallmo 0., Asplund L.* Neural networks for adaptive traffic control in ATM networks // IEEE Communications Magazine. 1995. Vol. 33, No. 10. P. 43–49.

[168] *Qiu Zh., Perez J.F., Harrison P.G.* Beyond the mean in fork-join queues:

Efficient approximation for response-time tails // Performance Evaluation. 2015. Vol. 91. P. 99–116.

[169] *Raaijmakers Y., Borst S., Boxma O.* Fork–join and redundancy systems with heavy-tailed job sizes // Queueing Systems. 2023. Vol. 103. P. 131–159.

[170] *Rabta B.* A review of decomposition methods for open queueing networks // Rapid Modelling for Increasing Competitiveness. Springer, London. 2009. P. 25–42.

[171] *Reiser M., Kobayashi H.* Accuracy of the diffusion approximation for some queuing systems // IBM Journal of Research and Development. 1974. Vol. 18, No 2. P. 110–124.

[172] *Reiser M., Lavenberg S.S.* Mean-value analysis of closed multichain queueing networks // Journal of the Association for Computing Machinery. 1980. Vol. 27, No. 2. P. 313–322.

[173] *Riordan J.* Telephone Traffic Time Averages // Bell Syst. Tech. 2018. Vol. 30, No. 4. P. 1129–1144.

[174] *Rizk A., Poloczek F., Ciucu F.* Stochastic bounds in Fork–Join queueing systems under full and partial mapping // Queueing Systems. 2016. Vol. 83. P. 261–291.

[175] *Schol D., Vlasiou M., Zwart B.* Large Fork-Join Queues with Nearly Deterministic Arrival and Service Times // Mathematics of Operations Research. 2022. Vol. 47, No. 2. P. 1335-1364.

[176] *Serpen G., Palabiyik Y., Serpen U.* An artificial neural network model for Na/K geothermometer // Proceedings 34th Workshop on Geothermal Reservoir Engineering. 2009. P. 1–10.

[177] *Sheng W.F.* Network Coding Method for Self-Similar Streaming Media Flow in Wireless Mesh Network // Proc. International Conference on Engineering Simulation and Intelligent Control (ESAIC). 2018. P. 334–339.

[178] *Shvedov A.S.* Functions approximating by neural networks and fuzzy systems // Control Sciences. 2018. Vol. 1. P. 21–29.

[179] *Sklar A.* Random Variables, Joint Distribution Functions, and Copulas // Kybernetika. 1973. Vol. 9. P. 449–460.

[180] *Smith J.M., Barnes R.* Optimal server allocation in closed finite queueing networks // Flexible Services and Manufacturing Journal. 2015. Vol. 27. P. 58–85.

[181] *Sun H., Chen X., Shi Q., Hong M., Fu X., Sidiropoulos N.D.* Learning to optimize: Training deep neural networks for wireless resource management // IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2017. P. 1–6.

[182] *Sundaria M.S., Palaniammalb S.* Simulation of $M|M|1$ queuing system using ANN // Malaya Journal of Matematik: Special Issue. 2015. Vol. 1. P. 279–294.

[183] *Sundaria M.S., Palaniammalb S.* An ANN simulation of single server with infinite capacity queuing system // International Journal of Innovative Technology and Exploring Engineering. 2019. Vol. 8, No. 12. P. 4067–4071.

[184] *Sundari M.S., Yamini S., Kalicharan R., Kumar S.S., Palaniammalb S.* Artificial neural network simulation for Markovian queuing models in a busy airport // Proc. International Conference on Computer Science, Engineering and Applications (ICCSEA). 2020. P. 1–6.

[185] *Spiegel M.R., Lipschutz, S. Liu J.* Mathematical Handbook of Formulas and

Tables, McGraw Hill Professional, 3rd (Third) edition. Schaum's Outline Series. 2008. 312 p.

[186] *Stidham S.* Optimal design of queueing systems. Boca Raton: CRC Press/Taylor & Francis, 2009. 384 p.

[187] *Stone M.N.* The generalized Weierstrass approximation theorem // Mathematics Magazine. 1948. Vol. 21, No. 4. P. 167–184.

[188] *Suresh S, Whitt W.* The heavy-traffic bottleneck phenomenon in open queueing networks // Operations Research Letters. 1990. Vol. 9, No 6. P. 355–362.

[189] *Tijms H.C.* Stochastic Models. An Algorithmic Approach. John Wiley & Sons, 1994. 375 p.

[190] *Thomas A.J., Petridis M., Walters S.D., Gheytassi S.M., Morgan R.E.* Two hidden layers are usually better than one // Communications in Computer and Information Science. 2017. Vol. 744. P. 279–290.

[191] *Thomasian A.* Analysis of Fork/Join and Related Queueing Systems // ACM Computing Surveys (CSUR). 2014. Vol. 47, No. 2. P. 17:1–17:71.

[192] *Thomasian A., Menon J.* RAID5 performance with distributed sparing // IEEE Transactions on Parallel and Distributed Systems. 1997. Vol. 8, No. 6. P. 640–657.

[193] *Ullah R., Marwat S.N.K., Ahmad A.M., Ahmed S., Hafeez A., Kamal T., Tufail M.* A machine learning approach for 5G SINR prediction // Electronics. 2020. Vol. 9, No. 10. Article Number: 1660.

[194] *Varki E., Merchant A., Chen H.* The $M|M|1$ Fork-Join Queue with Variable Subtasks. [Online] http://www.cs.unh.edu/ varki/publication/2002-nov-open.pdf

[195] *Varma S., Makowski A.M.* Interpolation approximations for symmetric fork-join queues // Performance Evaluation. 1994. Vol. 20. P. 245–265.

[196] *Vianna E., Comarela G., Pontes T., Almeida J., Almeida V., Wilkinson K., Kuno H., Dayal U.* Analytical performance models for MapReduce workloads // International Journal of Parallel Programming. 2013. Vol. 41, No. 4. P. 495–525.

[197] *Vishnevsky V.M., Klimenok V.I., Sokolov A.M., Larionov A.A.* Investigation of the Fork–Join System with Markovian Arrival Process Arrivals and Phase-Type Service Time Distribution Using Machine Learning Methods // Mathematics. 2024. Vol. 12, No. 5. Article Number: 659.

[198] *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.

[199] *Vishnevsky V.M., Dudin A.N., Kozyrev D.V., Larionov A.A.* Methods of Performance Evaluation of Broadband Wireless Networks Along the Long Transport Routes // Communications in Computer and Information Science. 2016. Vol. 601. P. 72–85.

[200] *Vishnevsky V.M., Krishnamoorthy A., Kozyrev D.V., Larionov A.A.* Review of methodology and design of broadband wireless networks with linear topology // Indian Journal of Pure and Applied Mathematics. 2016. Vol. 47. P. 329–342.

[201] *Vishnevsky V.M., Larionov A.A., Ivanov R.E.* An Open Queueing Network with a Correlated Input Arrival Process for Broadband Wireless Network Performance Evaluation // Communications in Computer and Information Science. 2016. Vol. 638. P. 354–365.

[202] *Vishnevsky V., Semenova O.* Polling systems and their application to telecommunication networks // Mathematics. 2021. Vol. 9, No. 2. Article Number: 117.

[203] *Vishnevsky V.M., Semenova O.V., Bui D.T.* Using a machine learning for analysis of polling systems with correlated arrivals // Lecture Notes in Computer Science. 2021. Vol. 13144. P. 336-345.

[204] *Weisstein E.W.* Digamma Function, MathWorld – A Wolfram Web Resource // https://mathworld.wolfram.com/DigammaFunction.html

[205] *Whitt W.* The Queueing Network Analyzer // The Bell System technical journal. 1983. Vol. 62, No 9. P. 2779–2815.

[206] *Whitt W.* Variability functions for parametric-decomposition approximations of queueing networks // Management Science. 1995. Vol. 41, No 10. P. 1704–1715.

[207] *Xia Y., Tse D.* On the large deviation of resequencing queue size: 2-M/M/1 case // IEEE Transactions on Information Theory. 2008. Vol. 54, No. 9. P. 4107–4118.

[208] *Yuzukirmizi M.* Closed finite queueing networks with multiple servers and multiple customer types. Ph.D. Dissertation. 2005.

[209] *Yousefi'zadeh H., Jonckheere H.* Dynamic neural-based buffer management for queuing systems with self-similar characteristics // IEEE Transactions on Neural Networks. 2005. Vol. 16. P. 1163–1173.

[210] *Zou H., Hastie T.* Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society. Series B (Statistical Methodology). 2005. Vol. 67, No. 2. P. 301–320.