МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ,
СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Санкт-Петербургский государственный университет телекоммуникаций
им. проф. М.А. Бонч-Бруевича»

На правах рукописи

Редругина Наталия Михайловна

МОДЕЛИ И МЕТОДЫ ОЦЕНКИ ХАРАКТЕРИСТИК ТЕЛЕКОММУНИКАЦИОННЫХ КОМБИНИРОВАННЫХ СЛАБОСВЯЗАННЫХ УСЛУГ

Специальность 2.2.15. Системы, сети и устройства телекоммуникаций

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
Кандидат технических наук, доцент
Зарубин Антон Александрович

Санкт-Петербург – 2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 4
ГЛАВА 1 ТЕЛЕКОММУНИКАЦИОННЫЕ УСЛУГИ: ЭВОЛЮЦИЯ И МОДЕЛИРОВАНИЕ 11
1.1 Представление объекта исследований 11
1.2 Теория эволюции телекоммуникационных услуг 12
1.3 Моделирование телекоммуникационных систем 19
1.4 Классификация телекоммуникационных услуг 27
1.5 Микросервисная архитектура - идеальный симбиоз технологий. 31
1.6 Выводы по главе 1 38
ГЛАВА 2 МОДЕЛИ ОЦЕНКИ ВРЕМЕННЫХ ХАРАКТЕРИСТИК КОМБИНИРОВАННЫХ СЛАБОСВЯЗАННЫЕ УСЛУГ 40
2.1 Подход к разделению услуг по уровню взаимодействия с пользователем 40
2.2 Принципы реализации и оценки характеристик комбинированных телекоммуникационных услуг 42
2.3 Модели и методы вычисления задержек при предоставлении пользователям сеансовых услуг 46
2.4 Математический инструмент расчета временных характеристик сеансовых услуг 55
2.5 Выводы по главе 2 61
ГЛАВА 3 МОДЕЛЬ ОЦЕНКИ ХАРАКТЕРИСТИК ТРАНЗАКЦИОННЫХ УСЛУГ С ПАРАЛЛЕЛЬНОЙ ОБРАБОТКОЙ ЗАПРОСОВ 63
3.1. Транзакционные услуги – сценарии выполнения рабочих процессов. 63
3.2 Параллельная обработка запросов через механизм Fork-Join 73
3.3 Моделирование рабочего процесса математическим аппаратом сетей Джексона 79

3.4 Аппроксимация неэкспоненциальных распределений с высоким коэффициентом вариации 93
3.5 Оценка временных задержек в последовательной цепочке локальных транзакций 102
3.6 Выводы по главе 3 110
ГЛАВА 4 МЕТОДЫ РАСЧЕТА ХАРАКТЕРИСТИК ТЕЛЕКОММУНИКАЦИОННЫХ УСЛУГ 112
4.1 Методика проектирования телекоммуникационных услуг 112
4.2 Пример расчета временных характеристик комбинированной услуги 118
4.3 Оценка применимости имитационного моделирования 125
4.4 Выводы по главе 4 130
ЗАКЛЮЧЕНИЕ 131
СПИСОК ЛИТЕРАТУРЫ 133
ПРИЛОЖЕНИЕ А. АКТЫ О ВНЕДРЕНИИ РЕЗУЛЬТАТОВ ДИССЕРТАЦИИ. 149

ВВЕДЕНИЕ

Актуальность темы исследования

В современном информационном обществе, где растущее количество пользователей ожидает высокое качество и эффективность услуг, разработка и повышение эффективности телекоммуникационных систем становится особенно важной задачей. Технологические и социокультурные изменения, такие как быстрое развитие сетей связи, увеличение объемов данных, повышение требований к безопасности и удобству использования, вносят новые вызовы и возможности в область телекоммуникационных технологий. В рамках данной работы понятие телекоммуникационных услуг включается в себя

Распределенные системы, в том числе системы, основанные на микросервисной архитектуре [1], становятся все более популярными и предлагают новые подходы к проектированию и развертыванию сложных информационных решений. Однако, такие системы требуют комплексного исследования, особенно в отношении их временных характеристик и производительности, чтобы обеспечить эффективную работу и удовлетворение потребностей пользователей [2].

Отдельно стоит отметить тенденцию к разработке цифровых двойников, в основе концепции, которой лежит численное моделирование позволяющее спрогнозировать эксплуатационные характеристики проектируемого объекта, в том числе системы или услуги [3].

Эффективность является важным критерием, который должен быть оценен конечными пользователями и поставщиками услуг, прежде чем внедрять телекоммуникационные услуги для работы на сложных глобальных рынках [4]. Риск связан с возможностью несоответствия внедряемой конечной услуги соглашению об уровне обслуживания (SLA).

Описанное выше позволяет сделать вывод об актуальности исследований, связанных с оценкой качественных характеристик цифровых услуг на всех этапах проектирования. В связи с этим в данной работе исследуются методы и подходы к оценке временных характеристик систем и их зависимость от различных

архитектурных решений. Особое внимание уделяется микросервисной архитектуре и её влиянию на производительность системы. Имитационное моделирование используется для получения точных результатов различных сценариев работы системы и сравнения их эффективности.

Степень разработанности темы

В современном информационном обществе происходит активный рост количества и разнообразия телекоммуникационных услуг, разработчики которых наряду с научными деятелями активно публикуют научные работы, связанные с анализом аспектов качества обслуживания многопользовательских услуг [4],[5] архитектурных решений с поддержкой QoS [6], разработкой составных услуг телекоммуникаций [7] и использованием существующих [8]. Результаты данных работ могут быть полезны при внедрении и использовании услуг телекоммуникаций на различных предприятиях.

На данный момент множество ученых обратились к тематике моделирования услуг телекоммуникаций, среди которых можно отметить авторов Б.С. Гольдштейна, А.Е. Кучерявого, А.Б. Гольдштейна, А.А. Зарубина, К.Е. Самуйлова, М.А. Шнепс-Шнеппе, В.И. Вишневского, Р.В. Киричека, А.И. Парамонова, С.Н. Степанова, И.В. Карташевского.

Множество статей, научных публикаций и диссертаций посвящены исследованию производительности и повышению эффективности телекоммуникаций в различных областях их применения. Однако, в существующих исследованиях может наблюдаться недостаточное внимание к определенным аспектам, таким как анализ требований к архитектурным решениям, сравнение различных сценариев реализации процесса предоставления услуги, и последующее получение качественных характеристик методами математического и имитационного моделирования для принятия обоснованных решений на этапах проектирования телекоммуникационных услуг.

Теоретические и практические инженерные исследования, раскрывающие часть вопросов проблематики, которая является основой данной диссертационной работы. В работе [9] авторы предлагают подход для оценки сценариев

развертывания архитектуры слабосвязанных программных компонентов с помощью автоматизированного тестирования производительности. Данный подход основан на нагрузочном тестировании на основе домена.

Авторы [10] используя графовую модель, детально анализируют механизм взаимодействия между слабосвязанными услугами в контексте рабочих процессов. В работе [11] в свою очередь, раскрываются принципы масштабирования веб-сервисов, а также обсуждаются проблемы, которые успешно решаются при данном подходе. Эти вопросы включают обработку большого числа симультанных запросов, администрирование ограниченной пропускной способности, а также управление задержками при операциях чтения и записи данных. В [12] описывают разработанный метод оценки времени отклика для производительности веб-служб. Авторы рассчитывают применить данный метод при проектировании вычислительных систем.

Данная диссертационная работа реализует важное дополнение к существующей литературе в области анализа телекоммуникационных услуг. Она фокусируется на исследовании временных характеристик и производительности при различных архитектурных подходах, включая микросервисную архитектуру. Синтез аналитического и имитационного моделирования дает возможность получить точные результаты для различных сценариев работы системы. Полученные результаты могут служить основой для принятия обоснованных решений по управлению качественными и количественными показателями на всех этапах проектирования и эксплуатации телекоммуникационных услуг, что делает данную работу важным вкладом в развитие данной области.

Объектом исследования являются телекоммуникационные слабосвязанные комбинированные услуги обслуживания пользовательских сеансов и реализации транзакционных рабочих процессов разрабатываемые на микросервисной архитектуре.

Предметом исследования являются модели и методы оценки вероятностно-временные характеристики различных сценариев реализации рабочих процессов транзакционных услуг и пользовательских сеансов.

Цели и задачи диссертации

Цель диссертации в повышении эффективности проектирования и предоставления телекоммуникационных услуг путем применения новых моделей и методов расчета временных характеристик систем обслуживания пользовательских сеансов и рабочих процессов в микросервисных архитектурах.

Для достижения поставленной цели в работе были поставлены следующие *задачи*:

1) Разработка модели оценки временных характеристик услуг реализующих поддержку пользовательских сеансов.

2) Разработка модели для расчета вероятностно-временных характеристик рабочих процессов транзакционных услуг с параллельной обработкой запросов.

3) Разработка модели с применением методов аппроксимации для расчета вероятностно-временных характеристик сетей систем общего вида G/G/1.

4) Разработка методов расчета вероятностно-временных характеристик комбинированных слабосвязанных услуг.

5) Разработка комплекса имитационных моделей и библиотеки структурных элементов для апробации и расчета характеристик систем.

Научная задача: Разработка моделей и методов их применения для оценки характеристик слабосвязанных услуг с целью повышения качества обслуживания и возможности принятия обоснованного решения, по управлению количественными показателями систем

Научная новизна результатов исследования работы обусловлена:

— новым подходом к моделированию телекоммуникационных услуг с разделением их на услуги реализующие пользовательские сеансы и рабочие процессы выполнения транзакций, а также их интеграция в комбинированные телекоммуникационные услуги.

— разработкой моделей для расчета характеристик последовательных сетей систем реализующих рабочие процессы транзакционных слабосвязанных

услуг *с внедрением математического аппарата параллельной обработки запросов Fork-Join* в том числе систем произвольного вида систем G/G/1.

— внедрением оригинального метода аппроксимации для определения характеристик выходного потока в сетях системах произвольного вида G/G/1, для моделирования последовательных сетей систем массового обслуживания.

— отсутствием аналогов разработанного комплекса имитационных моделей, в том числе для оценки характеристик системы параллельной обработки запросов Fork-Join для коэффициента репликации более двух

Теоретическая значимость работы состоит в разработке метода применения моделей для оценки вероятностных характеристик сценариев рабочих процессов транзакционных услуг с применением аппарата параллельной обработки, в применении методов аппроксимации для определения верних и нихних границ временных характеристик рабочих процессов в последовательных сетях систем произвольного вида и синтеза с инструментами имитационного моделирования, для комплексного анализа. В разработке модели на базе математического аппарата сетей Джексона для оценки характеристик многопользовательских услуг поддерживающих пользовательские сеансы, а также для их синтеза в рамках комбинированных услуг.

Практическая значимость работы заключается в разработке методов применения моделей для расчета вероятностно-временных характеристик различных типов телекоммуникационных услуг, которые могут быть внедрены в основу подхода к планированию, проектированию и разработке требований к предоставлению телекоммуникационных услуг, основанного на оригинальных моделях, полученных в целях прогнозирования и повышения эффективности качественных характеристик.

Также результаты представленной диссертации были использованы на этапе проектирования систем технического учета «Аргус NRI» и управлении рабочей силой «Аргус WFM», в разработке программного-аппаратного комплекса «Профит», в разработке программного комплекса «Контактный центр SIGURD-

MIND для непальского оператора связи NepalTelecom» в учебном процессе СПбГУТ на кафедре инфокоммуникационных систем (ИКС), что подтверждено соответствующими актами о внедрении в Приложении А.

Методология и методы диссертационного исследования. Выполненные в работе исследования основаны на положениях теории массового обслуживания, теории вероятностей и математической статистики, методах гибридного (имитационно-аналитического) моделирования дискретных событийных систем. Разработка комплекса имитационных моделей, выполненная с использованием программного обеспечения Anylogic. Расчет сложных математических моделей, и результатов аппроксимации были реализованы на ПО Mathcad.

Положения, выносимые на защиту:

1. Эффективный метод применения разработанных моделей расчета временных характеристик при предоставлении телекоммуникационных комбинированных слабосвязанные услуг.

2. Модели для расчёта временных характеристик транзакционных услуг с параллельной обработкой запросов эффективны для сетей систем с процессами поступления и обслуживания, распределенными экспоненциально, с величиной расхождения не более 1% между имитационным моделированием и аналитическим расчётом. Разработанные модели так же эффективны для расчёта сетей систем произвольного вида G/G/1.

3. Комплекс имитационных моделей сеансовых услуг и рабочих процессов транзакционных услуг, обеспечивающий выявление искомых временных характеристик в условиях не экспоненциального характера протекающих процессов в моделируемых сетях систем

Степень достоверности полученных результатов подтверждена результатами аналитических расчётов и имитационного моделирования. Результаты, полученные в ходе выполнения исследования, не противоречат ранее полученным данным, опубликованным в открытых источниках.

Апробация результатов. Результаты работы докладывались и обсуждались на международной научной конференции «Актуальные проблемы

инфотелекоммуникаций в науке и образовании» (АПИНО). (Санкт-Петербург, 2021, 2022);

Публикации. По теме данной диссертационной работы опубликовано 8 научных работ, в том числе в 4 работы в рецензируемых научных изданиях, рекомендованных ВАК РФ (2020, 2023) и 4 в других изданиях и материалах конференций. Получены свидетельства о государственной регистрации программы для ЭВМ «Программный модуль для расчета характеристик систем параллельной обработки транзакций, разработанный в среде Anylogic» № 2023616789 от 07.04.2023, «Программный модуль для расчета качественных характеристик комбинированных инфокоммуникационных услуг поддерживающих обслуживание пользовательских сессий и обработку системных транзакций в среде Anylogic» № 2023662354 от 14.06.2023.

Соответствие диссертации паспорту научной специальности. Диссертационная работа соответствует пунктам 1, 2, 4 и 5 паспорта научной специальности 2.2.15 – Системы, сети и устройства телекоммуникаций. Проведенные исследования соответствуют формуле специальности.

ГЛАВА 1 ТЕЛЕКОММУНИКАЦИОННЫЕ УСЛУГИ: ЭВОЛЮЦИЯ И МОДЕЛИРОВАНИЕ

1.1 Представление объекта исследований

Объектом исследования данной работы являются телекоммуникационные услуги. Государственный стандарт [13] определяет услугу как «Способ предоставления ценности заказчику через содействие ему в получении конечных результатов, которых заказчик хочет достичь». Исследование [14] определяет услугу как программный компонент для реализации описанной функциональности в виде механизма запроса и ответа, с возможностью определения конкретных свойств, которые можно назвать качеством обслуживания. Услуги в рамках телекоммуникационных технологий представляют собой комплексные сервисы и функциональности в области информационных и коммуникационных технологий, предоставляемые пользователям для обмена информацией, связи, доступа к данным и других потребностей. Они охватывают широкий спектр услуг, таких как мобильная и фиксированная связь, интернет-провайдинг, облачные сервисы, видеоконференции, электронная почта и многое другое.

Исследование услуг позволяет оценить их производительность, эффективность, надежность и адаптацию к различным условиям и требованиям пользователей. Рассматриваемые услуги применяются во множестве сфер деятельности, включая:

– Бизнес и предпринимательство: обеспечение связи и коммуникации между сотрудниками, клиентами и партнерами, с предоставлением инструментов для электронной коммерции, управления клиентскими отношениями и проектами, ведения бухгалтерии.

– Образование: обеспечение доступа к знаниям и информации, электронному обучению и дистанционному образованию, создание виртуальных классов и совместная работа студентов и преподавателей.

– Здравоохранение: электронный медицинский документооборот, телемедицина, медицинский мониторинг и диагностика, управление медицинскими записями пациентов, улучшение связи между врачами и пациентами.

– Государственное управление и администрирование: ИКТ-услуги предоставляются для улучшения эффективности и прозрачности административных процессов, электронного правительства, электронных услуг для граждан, управления информационной безопасностью и т.д.

– Транспорт и логистика: управление и контроль транспортных сетей, отслеживание грузов и транспортных средств, оптимизация маршрутов, электронное билетирование и т.д.

– Медиа и развлечения: ИК услуги используются в медиаиндустрии для распространения контента через интернет и цифровые платформы, стримингового видео, музыки и игр, с внедрением технологий виртуальной реальности.

В работе [15] описана концепция цифровой экономики как конвергенции современных технологий, существующих платформ и инновационных решений для предоставления доступа пользователей и умных вещей к определенному набору услуг.

1.2 Теория эволюции телекоммуникационных услуг

Эволюция услуг телекоммуникаций прошла через несколько важных этапов, начиная с появления первых телефонных линий в конце XIX века, когда телефонные услуги стали основным средством связи. Затем, в середине века, стали популярны услуги радиовещания и телевидения. В 1960-х годах появился компьютер, и с ним возникла возможность электронной связи, включая электронную почту и интернет.

С появлением мобильных телефонов в 1980-х годах и развитием беспроводной связи в 1990-х годах произошла революция в мобильной связи. Появление смартфонов, объединивших в себе функции мобильного телефона,

компьютера и цифрового фотоаппарата, кардинально повлияло на изменение всего спектра услуг. Процессы и принципы перехода к инфокоммуникациям подробно описаны в работе Гольдштейна [16].

Сегодня телекоммуникационные услуги включают широкий спектр технологий и сервисов, таких как голосовые и видео вызовы, обмен сообщениями, мессенджеры, социальные сети, облачные вычисления, потоковое видео и музыка, онлайн-игры, электронная коммерция и многое другое.

Эволюция сетей имеет прямое влияние на функциональность систем, и это взаимодействие является важным аспектом развития информационных технологий и телекоммуникаций. Важно понимать, что функциональность системы зависит от нескольких аспектов эволюции сетей, таких как архитектура, пропускная способность, скорость передачи данных, надежность, безопасность и доступность.

1) Эволюция сетей ведет к увеличению функциональности системы. Когда сети становятся более быстрыми и мощными, системы могут обрабатывать больший объем данных, что приводит к улучшению их функциональности.

2) Существует положительная корреляция между пропускной способностью сетей и функциональностью систем.

3) Современные сети предоставляют лучшие методы шифрования и защиты данных, что влияет на увеличение безопасности систем.

4) С развитием беспроводных технологий и мобильных сетей системы становятся более доступными в различных местах и условиях, что способствует повышению их функциональности.

5) Развитие облачных сервисов так же способствует улучшению функциональности систем.

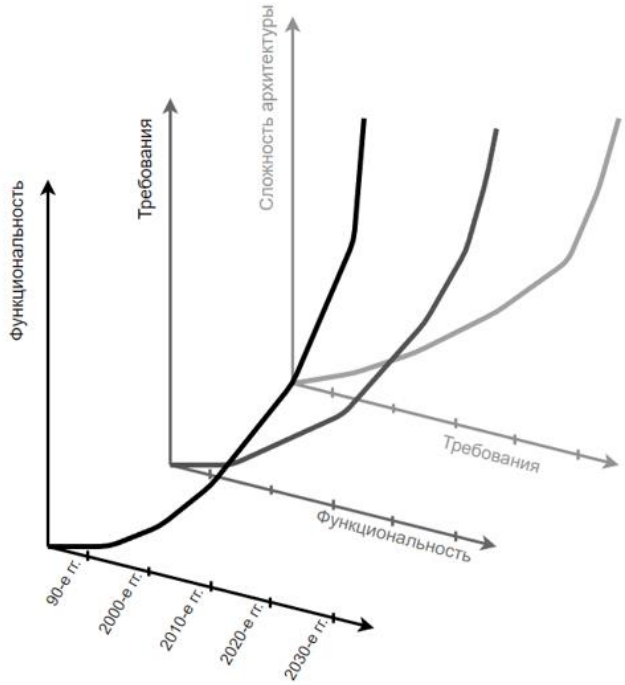


Рисунок 1.1 - Проблематика разработки телекоммуникационных услуг

Функциональность системы, контекст её использования и поставленные задачи влияют на предъявляемые к разрабатываемым системам требования. Концептуальные требования к системам описаны в исследованиях Сетей 2030 [17], сетей 6G [18-19]. Например, системы с высокой функциональностью могут требовать возможности масштабирования для поддержки роста количества пользователей или объема данных [20-21]. Также, если функциональность системы критична, вероятно, потребуется высокая надежность, чтобы минимизировать сбои и перерывы в работе. Поэтому проектирование и анализ требований [22-24] играют критическую роль в успешной реализации высокофункциональных систем. Следовательно, одной из ключевых задач, поставленных перед данным исследовательским трудом, заключается в разработке алгоритма, применяемого в фазе проектирования телекоммуникационных систем.

Зависимость сложности архитектуры от требований к системам является еще одной важной переменной в разработке программных продуктов. Сложность архитектуры возрастает вместе с уровнем функциональности, производительности, надежности, безопасности и другими характеристиками, необходимыми для

удовлетворения требований. Так, например если система должна поддерживать масштабируемость для обработки растущего объема данных или пользователей, это может потребовать разработки распределенной архитектуры, где различные компоненты системы работают независимо и могут масштабироваться горизонтально, а также поддерживать возможность балансировки входящей нагрузки от различных источников. В [25],[26] авторы рассматривают вопрос балансировки нагрузки от запросов приложений на центры обработки данных. Так же если система должна взаимодействовать с различными сторонними системами и сервисами, может потребоваться разработка гибкой архитектуры с хорошо определенными интерфейсами и протоколами обмена данными.

Для реализации требований, предъявленных к системам, важно отметить эволюцию и сетевой инфраструктуры. Требования к сетям будущего и методы обеспечения QoS на сетях нового поколения являются предметом постоянного внимания со стороны ITU, ETSI, IETF и иных организаций по стандартизации в области телекоммуникаций. Так же проблемы поиска методов и подходов по повышению качества обслуживания на сетях будущего рассматривали Б.С. Гольдштейн, А.С. Гольдштейн, А.Е. Кучерявый, А.И. Парамонов, М.А. Маколкина, А.С.А. Мутханна, Р.В. Киричек, О.А. Симонина, В.К. Сарьяна, С.Н. Степанов, В.О. Тихвинский, А.С. Бородин, Ю.В. Гайдамаки и другие авторы многих научных работ таких как [27-30].

Это открывает новые возможности для создания и использования приложений, которые ранее были недоступны или непрактичны в более медленных сетях. Интересно то, что для полной оценки преимуществ сетей нового поколения, необходимо также оценить качество приложений, которые используют эти сети.

Проблематика, поставленная во главу данной работы, заключается в следующем: важно не только обеспечить высокое качество базовых услуг сети, но также необходимо обеспечить высокое качество сервисов, которые используют сетевую инфраструктуру, это позволит полностью оценить преимущества новых технологий. С данным вопросом можно ознакомиться при анализе научно-технической литературы таких авторов как Б.С. Гольдштейн, А.А. Зарубин, Ю.В.

Малиновский, Е.Б. Фишман, Т.И. Алиев по построению эффективных систем, теории очередей [31-33], прогнозированию состояния систем [34] и многих других источников.

Для обеспечения высокого качества предоставляемых услуг, необходимо проводить анализ и моделирование системы, что позволит прогнозировать возможные проблемы и улучшать качество сервисов на ранних стадиях его разработки. Это одна из задач, поставленных во главу развития теории массового обслуживания, к которой впоследствии будет ссылаться данная диссертационная работа. Применение концепций теории массового обслуживания занимает многие десятилетия научных изысканий и связано с выдающимися именами ученых, в числе которых Л. Клейнрок [35], Ю.К.Беляев, И.Н.Коваленко [36], П.П. Бочаров [37-38], A. Ghosal [39], Б.А.Севастьянов [40] в теории вероятностей и В.А.Диткин [41] в математическом анализе. Их значимый вклад оказал существенное воздействие на развитие предметной области и на её приложение в анализе информационных систем и сетей связи, наряду с другими выдающимися исследователями.

Для дальнейшего анализа данного вопроса необходимо понимать, что рассматриваемые приложения могут являться как отдельной услугой, так и поддерживать взаимодействие нескольких компонентных услуг [42].

В дальнейшем анализ телекоммуникационных услуг будет проводится с использованием модели абстрактной социальной сети, изображенной на рисунке 1.2.

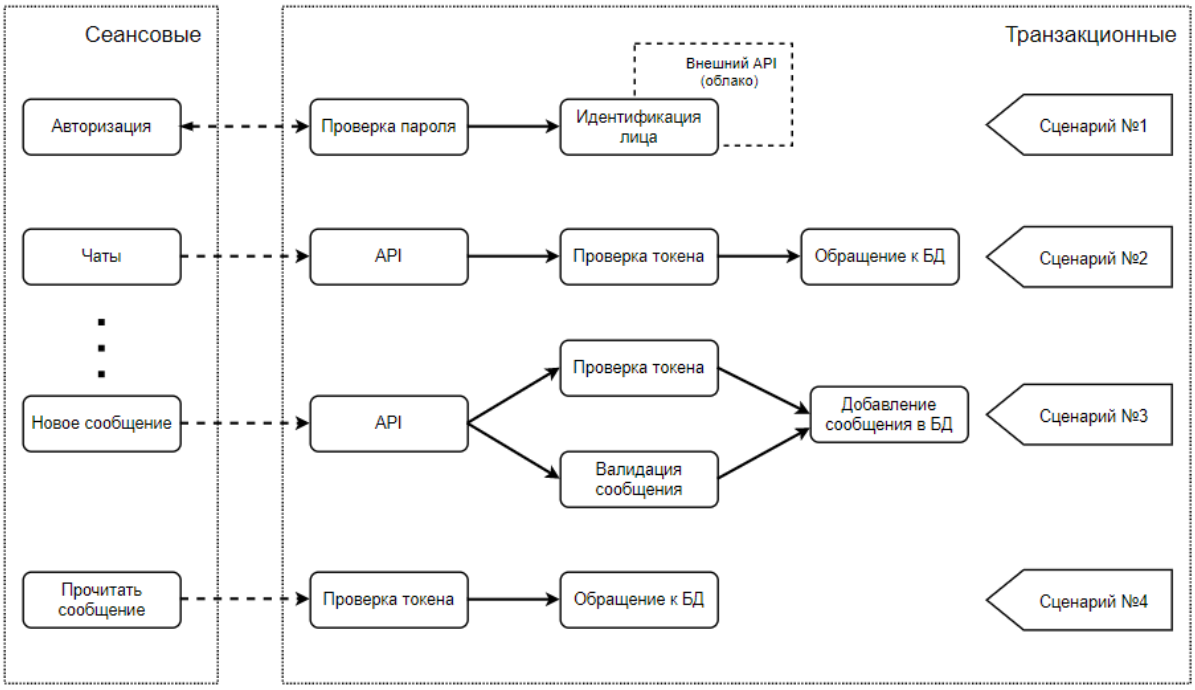


Рисунок 1.2 – Функциональная модель абстрактной социальной сети

Для разработки функциональной модели и дальнейшего анализа, предлагается ознакомиться с одним из подходов к моделированию систем - использование методов и моделей Domain-Driven Design (DDD) [43]. Он предполагает создание полной и точной модели предметной области, которая отражает бизнес-правила, процессы и концепции, помогает разбить систему на отдельные блоки и описать их функциональность и их взаимодействие. Это позволяет лучше понять систему в целом и повысить эффективность её работы для достижения нужного качества обслуживания.

В данном исследовании была поставлена задача разработки алгоритма оценки вероятностно-временных характеристик (ВВХ) комбинированных (сеансовых и транзакционных) услуг с ослабленной связью между компонентами архитектуры.

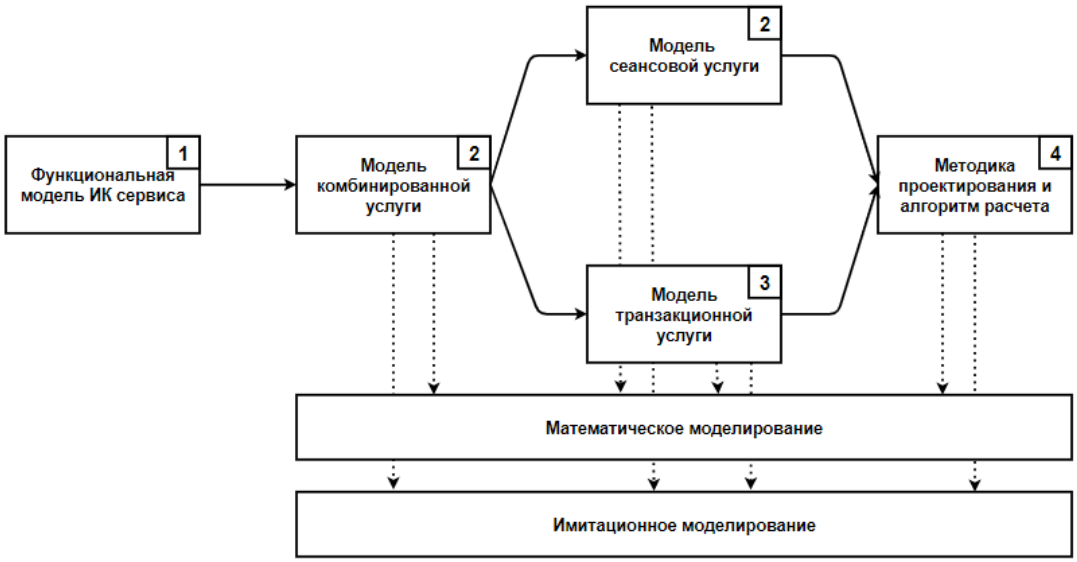


Рисунок 1.3 – Структура диссертационной работы

Для успешного выполнения этой задачи предполагается следующее:

1) Провести анализ требований к моделируемым услугам. Это позволит определить основные параметры и характеристики, которые должны быть учтены в реализуемых моделях.

2) Выбрать соответствующие методы моделирования, учитывая цель работы, доступные данные и ресурсы. Это может включать имитационное моделирование, математическое моделирование, аналитические методы и другие подходы.

3) Создать семейство моделей комбинированных услуг телекоммуникаций с учетом слабой связи между элементами архитектуры. Модели должны быть достаточно адекватными и точными, чтобы обеспечить надежный прогноз и анализ производительности системы.

4) Проверить и валидировать разработанные модели, используя реальные данные и сценарии работы системы. Это позволит убедиться в их правильности и соответствии реальным условиям.

5) Произвести прогнозирование и анализ производительности комбинированных услуг на основе созданных моделей. Исследовать различные сценарии работы системы и оценить их эффективность.

По представленной схеме и описанию реализации одной из основных задач можно заметить, что в данной работе возникает необходимость использования комбинации математических и имитационных методов моделирования на всех этапах исследования. Оба метода моделирования взаимодополняют друг друга и позволяют более полно и точно исследовать временные характеристики телекоммуникационных услуг с учетом различных архитектурных решений.

1.3 Моделирование телекоммуникационных систем

1.3.1 Методы моделирования ИКС

Моделирование систем на этапе проектирования и разработки для прогнозирования функциональных возможностей и характеристик системы, позволяет не только создавать более качественный продукт, но и сокращать время на его разработку и доработку. Формула для определения возможностей изменения характеристик системы, описанная в [44] может быть представлена как:

$$\Delta H = \sum_{i=1}^{n} X_i \times W_i \qquad (1.1)$$

где ΔH- относительное изменение выходной характеристики H (например, производительности, надежности, доступности) системы, X_i - относительное изменение i-й входной характеристики (например, количества ресурсов, скорости обработки, времени отклика), W_i - весовой коэффициент i-й переменной.

Для моделирования телекоммуникационных услуг используются различные методы, включая:

1. *Математическое моделирование* используется для описания поведения систем и услуг с помощью математических уравнений и моделей. Оно может быть использовано для прогнозирования трафика, определения пропускной способности сетей и расчета надежности систем.

2. *Имитационное моделирование* используется для симуляции работы систем, оценки их производительности и надежности. Оно позволяет создавать виртуальные модели сетей и систем, которые могут быть использованы для

тестирования и повышение их производительности. Пример реализации для корпоративной системы [45].

3. *Анализ данных* позволяет проводить обработку и анализ больших объемов данных, собранных из телекоммуникационных систем и услуг. Он может использоваться для определения трендов и паттернов использования услуг, прогнозирования спроса и определения областей, требующих улучшения.

4. *Методы машинного обучения* используются для создания прогностических моделей на основе анализа больших объемов данных. Они могут использоваться для прогнозирования спроса на услуги, определения рыночных тенденций и улучшения качества обслуживания пользователей, за счет прогнозирования нагрузки.

5. *Экспертные методы* позволяют производить анализ эффективных решений в управлении трафиком и прогнозировании результатов для проектирования сетей и систем используя экспортов в конкретной области.

В зависимости от конкретной задачи, требуемой точности прогнозирования и доступных ресурсов, могут быть использованы различные комбинации методов моделирования. Как можно было заметить в данной работе для оценки вероятностно-временных характеристик системы будет использована гибридная имитационно-аналитическая модель рисунок 1.4 [46].

Рисунок 1.4 – Гибридная имитационно-аналитическая модель

Если рассматривать наиболее практическое применение моделирования в рамках разработки информационных систем и услуг, стоит упомянуть подход Model-Driven Architecture (Модельно-ориентированная архитектура, MDA). Данный подход к разработке программного обеспечения, основан на создании

абстрактных моделей системы, которые затем используются для автоматической генерации кода и других компонентов системы.

В развитии информационного пространства стала выражена тенденция в создании готовых решений и шаблонов для дальнейшей разработки, что сокращает сроки и автоматизирует процесс [44],[47].

Данный принцип лег в основу разработки итогового метода оценки ВВХ систем, реализующий разработанный математический инструментарий и библиотеку имитационных моделей.

1.3.2 Математический инструментарий для моделирования ИКС

Комплексные телекоммуникационные услуги представляют из себя сети систем массового обслуживания (СеМО). Математическое моделирование которых использует знания теории массового обслуживания, теорию телетрафика, в частности математический аппарат сетей Джексона. На рисунке 1.5 представлена абстрактная архитектура СеМО открытого типа, с обратными связями.

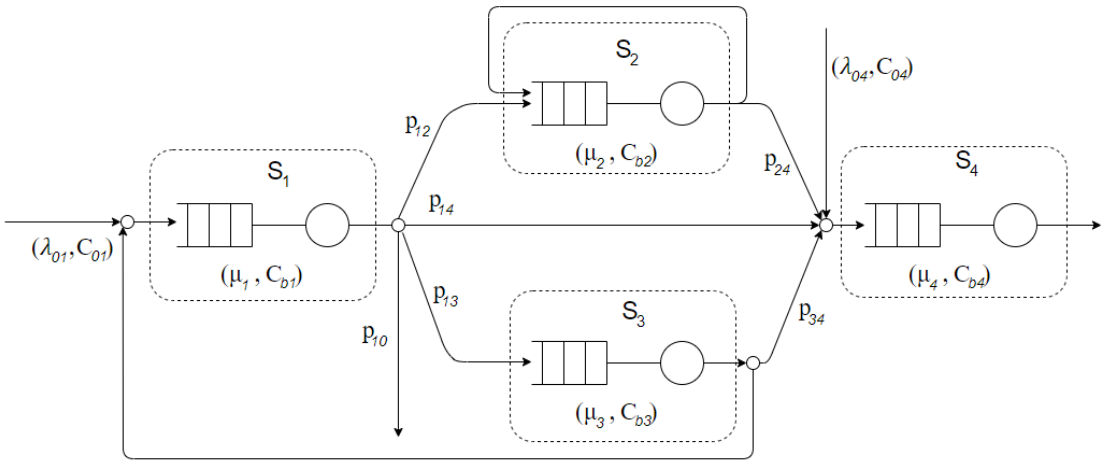


Рисунок 1.5 – Пример модели СеМО открытого типа.

Параметры, определяющие структуру СеМО с учетом её топологии и разветвления сценариев:

μ_i- интенсивность обслуживания в узле i;

λ_{ij}- интенсивность входного потока с узла i на узел j;

C_{ij}- коэффициент вариации процесса поступления из узла i в узле j;

C_{bi}- коэффициент вариации процессов обслуживания в узле i;

Y – общее количество систем массового обслуживания;

R = (p_{ij}) – матрица переходов, где p_{ij} – вероятность перехода для запросов, поступающих их узла i в узел j, i=1…Y, j=1…Y.

Предполагается, что время перехода запроса между элементами равно нулю, однако в случае, когда данная гипотеза не может быть принята, исследуемая модель усложняется. Например, для учета времени распространения сигнала между узлами коммутации необходимо ввести дополнительный элемент СМО G/G/∞. Данная модель предусматривает выделение обслуживающего устройства для каждого запроса, что гарантирует мгновенное поступление на обслуживание.

Расширенная модель последовательности действий может быть упрощена путем выделения каждой последовательности в новый узел с более сложным действием и определением плотности вероятности распределения (ПРВ) времени выполнения этого действия. Этот подход может быть полезен как при анализе системы сначала в целом, а затем разбиении ее на более мелкие подсистемы с сохранением характеристик элемента, так и при использовании обратного свойства преобразования, когда последовательный анализ цепочки узлов возможно упростить, если использовать свойство сцепления.

Результатом данных преобразований являются системы массового обслуживания как отдельные элементы с уникальными свойствами и законами функционирования, но при этом сохраняется их общая суть, по принципу черного ящика.

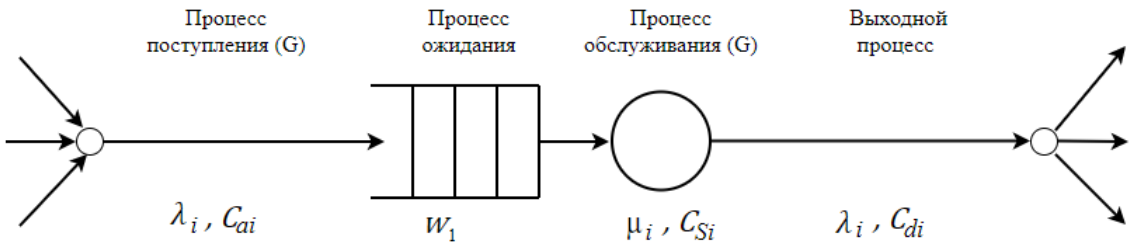


Рисунок 1.6 – Обобщенная система массового обслуживания G/G/1.

Время между прибытием запросов в рассматриваемую систему задается произвольным законом распределения (G), после поступления в случае, если обслуживающий прибор свободен запрос поступает на обслуживание с произвольным характером (G), в ином случае ожидает освобождения прибора обслуживания в буфере, данные процессы обладают следующими характеристиками:

λ_i – интенсивность поступления на узел i;

C_{Ai} – коэффициент вариации процессов поступления;

μ_i – интенсивность обслуживания на узле i, где $t_i = 1/\lambda_i$- среднее время обслуживания запросов;

C_{Si} – коэффициент вариации процессов обслуживания.

Для корректного функционирования сети все её элементы СМО должны находится в стационарном состоянии, условие для которого:

$$\rho_i = \lambda_i/\mu_i < 1 \qquad (1.2)$$

Наличие неточных решений для общего класса сетей и систем [48] побудило исследователей разработать приближенные методы оценки показателей производительности. В рамках данной проблематики можно выделить четыре категории, применяемых для общего случая.

1) Диффузионное приближение [49-50] — это аналитический подход, в рамках которого система описывается в терминах случайных колебаний или диффузии, что позволяет рассматривать ее стохастические характеристики и прогнозировать ее поведение во времени.

2) Анализ среднего значения [51-52] фокусируется на измерении и интерпретации среднего значения показателей производительности, таких как среднее время ожидания, среднее время обслуживания или среднее число запросов в системе. Помогает определить уровень производительности системы и выявить потенциальные проблемы или улучшения.

3) Оперативный анализ [53] включает в себя сбор данных о работе системы, измерение показателей производительности и их интерпретацию с целью принятия оперативных решений.

4) Методы декомпозиции

Методы декомпозиции представляют собой стратегии, направленные на обобщение концепции независимости и создание общих аппроксимаций для сетей типа Джексона.

Данный метод включает в себя следующие этапы:

1. Анализ взаимодействия между узлами путем рассмотрения сети как совокупности трех основных процессов:

Поток через очередь или станцию схема, изображенная на рисунке 1.6. В данном сценарии базовые характеристики сети систем массового обслуживания задаются следующими параметрами:

– интенсивность потока, поступающего на следующий узел:

$$\lambda_{i+1} = \lambda_i \qquad (1.3)$$

– коэффициент вариации потока поступления на следующий узел в цепочке последовательных систем [48]:

$$C_{ai+1}^2 \approx 1 + \rho_i^2\left(C_{si}^2 - 1\right) + (1 - \rho_i^2)(C_{ai}^2 - 1) \qquad (1.4)$$

где C_{ai}^2 - квадрат коэффициента вариации (scv) между поступающими требованиями в узел i, равен csv выходному потока из предыдущего узла C_{di-1}^2

ρ_i – загрузка узла i, стоит отметить, что при $\rho \to 0, C_d^2 \to C_{ai}^2$.

Данный параметр является дополнительной величиной описывающий события происходящие в системах массового обслуживания и определяется как отношение ее квадрата дисперсии к квадрату среднего значения.

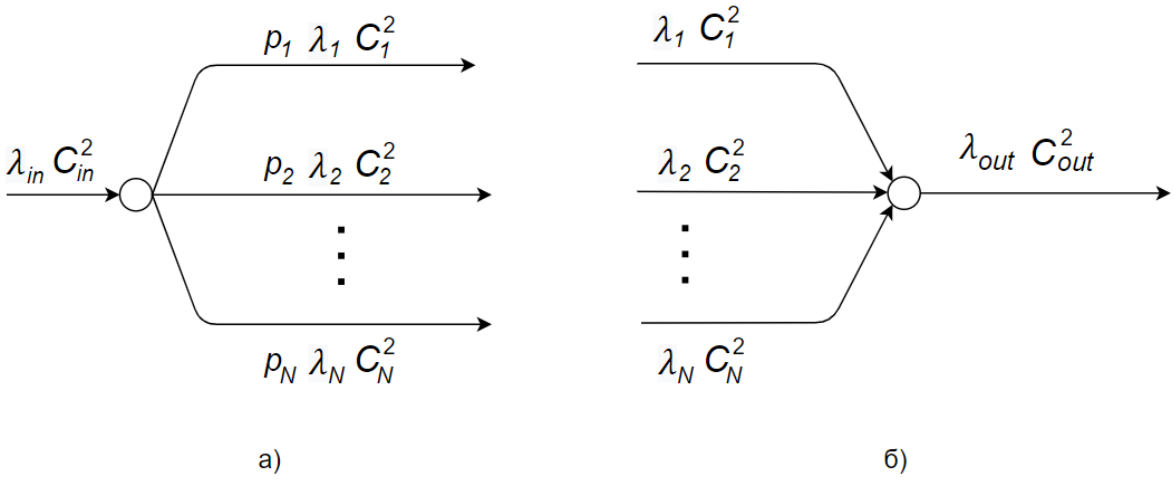


Рисунок 1.7 – Сценарии взаимодействия потоков запросов в СеМО

– Расщепление или декомпозиция, показанные на рисунке 1.7, относится к процессу разделения или разбиению общего потока запросов на более мелкие и управляемые подпотоки запросов. Этот процесс является полезным в ситуациях, когда необходимо осуществить более гибкое управление обработкой запросов, распределение нагрузки или применение различных политик обработки для разных категорий запросов.

Для данного процесса параметры выходного потока, следующие:

– выходная интенсивность:

$$\lambda_j = \lambda_i p_{ij} \tag{1.5}$$

– коэффициент вариации потока отправления:

$$C_j^2 \approx 1 + p_{ij}(C_i^2 - 1) \tag{1.6}$$

Наложение или слияние потоков запросов [54] (как показано на рисунке 5 (б)) реализуется, при объединении нескольких потоков запросов или транзакций в один общий поток для целей управления системами массового обслуживания. Параметры потока определены выражениями:

– интенсивность исходящего потока:

$$\lambda_{out} = \sum_{i=1}^{N} \lambda_i \tag{1.7}$$

– коэффициент вариации исходящего потока:

$$C_{out}^2 \approx 1 + \sum_{i=1}^{N} \left(\frac{\lambda_i}{\lambda_{out}}\right)^2 (C_i^2 - 1) \quad (1.8)$$

Процессы разделения и слияния моделируют маршрутизацию продукта в сети систем массового обслуживания. Такой подход позволяет эффективно управлять потоком запросов и распределять ресурсы для достижения высокой производительности и эффективности обслуживания. Декомпозиция сети на подсистемы отдельных узлов и их анализ, где каждый узел станция анализируется на основе частичной информации, полученной на предыдущем этапе. Анализ систем M/G/1 и GI/G/1 для оценки производительности рассматривался в работе [54]

Метод для моделирования системы путем декомпозиции использует следующие методы:

1) Декомпозицию сети массового обслуживания на подсистемы – системы массового обслуживания (СМО).
2) Анализ подсистем в отдельности, оценка взаимодействия с сетевым окружением подсистемы - процессов прибытия и отправления.
3) Аппроксимацию всех невозобновляемых процессов стационарными процессами обновления
4) Оценку временных характеристик системы – среднего значения и дисперсии.
5) Сведение общего анализа к нескольким элементарным операциям.

Так же основные параметры системы для времени между прибытиями T_A, времени обслуживания T_S, времени между отправлениями T_D характеризуются параметрами:

- функция распределения вероятностей времени будет равна вероятности того, что случайная величина T будет меньше или равна заданному t:

$$F(t)=P(T \leq t) \quad (1.9)$$

- k-й момент:

$$E[T^k] = \int_0^\infty t^k F`(t) dt \ , \quad k = 1,2\ldots \quad (1.10)$$

- коэффициент вариации:

$$c = \sqrt{\frac{E[T^k]}{E[T]^2} - 1} \quad (1.11)$$

2. Агрегировав полученные результаты, можно провести оценку эффективности сети с использованием таких метрик производительности, как среднее время ожидания обслуживания и время отклика.

$$\bar{T} = \bar{W} + \bar{t}_{\text{обсл.}} \quad (1.12)$$

Изначально мы предполагаем, что все системы являются системами G/G/v, однако в дальнейшем для обозначения типа процесса будут использоваться сокращения по классификации Кендалла, например M - марковских, D - детерминированный, H_k гиперэкспоненциальных процессов порядка k.

Для выбора правильной стратегии моделирования телекоммуникационных услуг с целью прогнозирования необходимо определить тип услуги для чего далее будет введена их классификация. Это позволит определить свойства, заложенные внутри субъекта такие как характеристики поступающих на обслуживание запросов, характер обслуживания (дисциплина очередей, количество обслуживающих устройств, характеристики очереди, функция вероятности времени обслуживания и т.д.), архитектурные решения, включающие описание взаимосвязи между элементами услуг.

1.4 Классификация телекоммуникационных услуг

Классификация услуг позволяет выбрать правильную стратегию моделирования для получения качественных характеристик в зависимости от назначения, типа информации которой оперируют услуги, типа используемых технологий, архитектурных решений и т.д. Дальнейшее разделение услуг:

- По качеству обслуживания (QoS) согласно рекомендации [55]:

— Критически важные услуги (Critical Services) требуют наивысшего уровня качества обслуживания с нулевой допустимой неработоспособностью.

Примеры: экстренные вызовы, системы управления трафиком и безопасностью, медицинские услуги в реальном времени;

— Временно-критические услуги (Real-time Interactive Services), требуют высокой доступности и надежности, с минимальными задержками и джиттером. Примеры: голосовая связь в режиме реального времени, видеоконференции, онлайн-игры;

— Интерактивные услуги (Interactive Services), требуют надежной передачи данных и достаточной скорости для обеспечения интерактивности. Примеры: мультимедийные услуги, стриминг видео, онлайн-чаты;

— Транзакционные услуги (Transactional Services), в которых производительность более важна, чем интерактивность. Примеры: онлайн-банкинг, электронная коммерция, электронные транзакции;

— Несрочные услуги (Bulk Transfer Services) имеют низкие требования к качеству обслуживания, которые могут быть выполнены в течение продолжительного времени. Примеры: загрузка больших файлов, пакетная передача данных.

- По функциональному назначению

– Коммуникационные сервисы - услуги, связанные с передачей голосовой, текстовой и видео информации, такие как мобильная связь, фиксированная связь, Интернет-телефония (VoIP), видеосвязь, текстовая связь (SMS, чаты), электронная почта и др.

– Услуги безопасности, связанные с защитой данных и информации, такие как антивирусная защита, защита персональных данных, защита от кибератак и др.

– Интернет-сервисы обеспечивают доступ к информации и услугам в Интернете, такие как поиск информации, онлайн-медиа (видео, музыка, игры), услуги электронной коммерции (онлайн-магазины, услуги платежных систем), услуги социальных сетей и др.

– Управление процессами: включают услуги, связанные с управлением бизнес-процессами и проектами.

– Аналитика и бизнес-интеллект - услуги, предоставляющие анализ и обработку данных.

- По типу используемых технологий:

– Кабельная связь — это технология, которая использует физический кабель для передачи данных;

– Беспроводная связь позволяет передавать данные без использования физического кабеля;

– Спутниковая связь использует искусственные спутники для передачи данных;

– Облачные технологии — это технологии, которые позволяют пользователю получить доступ к вычислительным ресурсам, хранению и программному обеспечению удаленно через Интернет;

– Интернет вещей — это технология, которая позволяет устройствам подключаться к Интернету и взаимодействовать друг с другом. Примеры включают умный дом, умные города, промышленный Интернет вещей и т.д.;

– Виртуальная и дополненная реальность — ряд технологий, которые позволяют создавать и использовать виртуальную или дополненную реальность. Примеры включают виртуальные игры, тренировки, конференции и т.д.;

– Блокчейн технология, которая обеспечивает распределенный реестр для хранения данных и транзакций;

– Искусственный интеллект - технология, которая позволяет компьютерам симулировать интеллектуальное поведение.

- Архитектурные решения для систем

Архитектура системы – это комплекс основополагающих принципов организации системы, которые воплощены в наборе её компонентов, связях компонентов друг с другом и с внешним окружением, а также принципов

проектирования и развития системы [56]. Про архитектурные принципы построения программного обеспечения так же писал Robert C. Martin [57].

– *Монолитная архитектура*: это классическая архитектура, в которой приложение представляет собой единое целое. В монолите все компоненты приложения взаимодействуют в рамках единого процесса;

– В *сервис-ориентированной архитектуре (SOA*) приложение состоит из набора независимых сервисов, выполняющих определенную функцию, которые взаимодействуют между собой через сеть;

– В *событийно-ориентированная архитектуре* сервисы взаимодействуют между собой через события, которые передаются через шину сообщений;

– Отдельно стоит упомянуть *бессерверную архитектуру* [58], которая относится к стилю облачных вычислений, в которой серверная инфраструктура, обычно предоставляемая облачным провайдером, управляет масштабированием и выполнением функций в ответ на события. Разработчик не заботится о настройке и управлении серверами и инфраструктурой, что позволяет создавать и развертывать сервисы более быстро и эффективно, не тратя время на управление инфраструктурой.

Согласно современным требованиям и растущей сложности архитектурных решений в разрабатываемых сервисах, предпочтительным выбором для многих компаний и организаций в разработки сервисов стала микросервисная архитектура (МСА). Так, например авторы [59] предлагают реализацию систем управления данными на основе микросервисов для своих веб-сервисов для обеспечения эффективного управления и низких затрат на разработку и производство. Авторы [60] утверждают, что МСА может стать возможностью привнести такую гибкость и динамичность в системы промышленной автоматизации. Поэтому в данной научной работе в целях анализа и моделирования высоконагруженных сервисов и оценки их производительности, будут учитываться особенности именно этой архитектуры.

1.5 Микросервисная архитектура - идеальный симбиоз технологий.

1.5.1 MCA – архитектура слабосвязанных услуг

Довольно большое количество исследований посвящено микросервисной архитектуре. В том числе данные исследования включают в себя сравнительный анализ MCA и иных архитектур [61-62], а также возможности миграция микросервисов из других архитектур [63-64].

Определения микросервисов даны в работах [65] «Микросервис — это целостный, независимый процесс, взаимодействующий посредством сообщений». А также определение, данное в [66] “слабо связанный сервис в ограниченном контексте” - понятие, которое будет в дальнейшем определять элементы структуры сервисов, построенных на MCA как *слабосвязанные сервисы*.

Микросервисная архитектура была ответом на проблематику при создании крупномасштабных распределенных веб-приложений, что объясняет её популярность. Разработчики ASR- платформы в работе [67] описали принцип выбора MCA для управления метаданными пользователя в системе. В [68] предложена реализация интернет-банковской системы на микросервисной архитектуре и сервисов для обработки больших данных (big data) в [69]. Применение инфраструктуры микросервисов для встраиваемых систем описано в работе [70], а для слияния перекрестных сервисов [71].

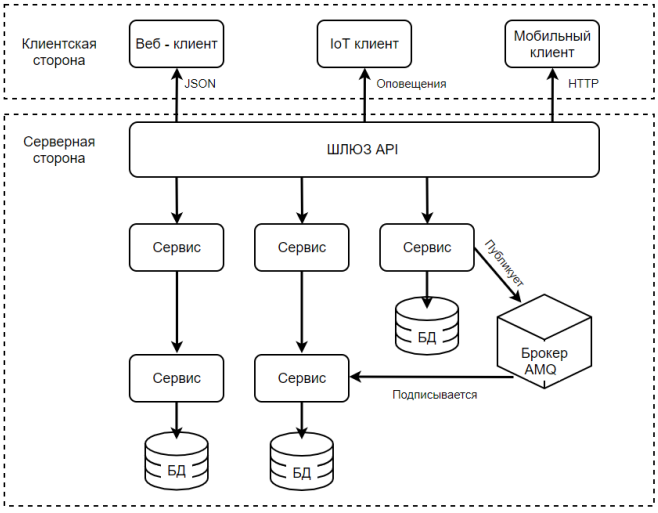


Рисунок 1.8 – Типовой пример микросервисной архитектуры

Ранее был обозначен вопрос использования паттернов построения цифровых услуг, однако необходимо понимать, что в микросервисной архитектуре как новой концепции построения систем используются иные принципы реализация шаблонов проектирования. В работах [72-74] рассматриваются облачные шаблоны проектирования как абстрактные решения повторяющихся проблем проектирования. Результатом данных исследований являются измерения, которыми выделено увеличение производительности и другие качественные характеристики системы при переходе на архитектуру микросервисов.

В рамках данной работы по разработке моделей и методов расчета качественных характеристик слабосвязанных услуг, будут предложены решения применимые к проектированию микросервисной архитектуры.

Стоит так же отметить, что в рамках данной работы слабосвязанные услуги будут рассматривать с точки зрения их взаимодействия в пользователя и делиться на сеансовые и транзакционные услуги, функциональность которых будет рассмотрена далее.

1.5.2 Свойства микросервисов

Микросервис представляет собой отдельно развертываемую и масштабируемую единицу, которая выполняет конкретную бизнес-функцию и общается с другими микросервисами посредством явно определенных интерфейсов. Авторы работы [75-76] - указывают на отсутствие правил относительно того насколько маленьким должен быть микросервис, однако он должен быть сосредоточен на единице работы и иметь следующие свойства:

1. Микросервисы разделяют функциональность приложения на небольшие, автономные и изолированные компоненты, каждый из которых выполняет свою конкретную задачу.

2. Каждый микросервис ограничивает свою функциональность определенным контекстом бизнес-задачи или доменной областью.

3. Микросервисы могут быть независимо развернутыми и масштабированными. Что позволяет проводить эффективное управление,

использование ресурсов и легко вносить изменения или обновления в отдельные компоненты без воздействия на всю систему.

4. Коммуникация посредством API, позволяет микросервисам общаться и передавать информацию друг другу, сохраняя слабую связанность между компонентами системы.

5. Каждый микросервис отвечает за свою собственную функциональность и имеет свою собственную базу кода, базу данных или другие ресурсы, не зависящие от других микросервисов.

1.5.3 Эквивалентные микросервисы

Эквивалентные микросервисы — это микросервисы, которые выполняют одну и ту же функцию, но реализованы с использованием различных технологий или подходов [77].

Преимущество использования эквивалентных микросервисов заключается в том, что они могут использоваться для обеспечения высокой доступности системы. Если один из микросервисов выходит из строя, его функции могут быть перехвачены другим эквивалентным микросервисом. Также использование эквивалентных микросервисов может увеличить масштабируемость системы, так как разные микросервисы могут быть запущены на разных платформах или в разных облачных системах.

Использование эквивалентных микросервисов может улучшить временные характеристики рабочих процессов за счет распределения нагрузки между микросервисами и возможности их параллельной обработки. Методы параллельной обработки запросов будут более детально рассмотрены в третьей главе данной работы. Выбор сервиса, основанный на качестве обслуживания из функционально эквивалентных сервисов для приложения более высокого качества описан в [78].

1.5.4 Рабочие процессы в микросервисной архитектуре

В микросервисной архитектуре автоматическая генерация рабочих процессов используется для координации и управления различными микросервисами в рамках одного приложения или системы. Каждый микросервис выполняет свою узкоспециализированную функцию и может использоваться в нескольких рабочих процессах.

На рисунке представлен пример рабочего процесса в микросервисной архитектуре [79]:

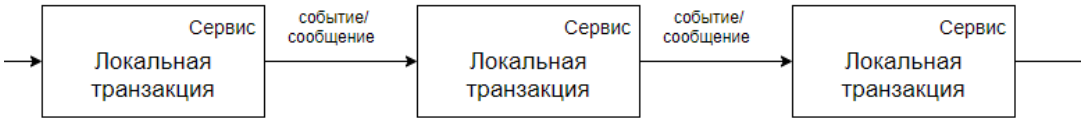


Рисунок 1.9 – Шаблон проектирования рабочего процесса в МСА

В контексте микросервисной архитектуры, существуют шаблоны, используемые для управления последовательностью операций (шагов), которые взаимодействуют с различными микросервисами. Каждый шаг транзакции выполняется в отдельном микросервисе, а координатор контролирует выполнение этих шагов.

В упомянутых паттернах можно использовать разные виды технологий [80] в зависимости от конкретной реализации.

– *Централизованная логика* управления выполнением шагов транзакции. Данный механизм определяет последовательность и координирует выполнение каждого шага транзакции, обращаясь к соответствующим микросервисам. Он может быть реализован как отдельный сервис или компонент внутри системы. На рисунке 1.10 изображен сценарий работы с отказом одного из сервисов.

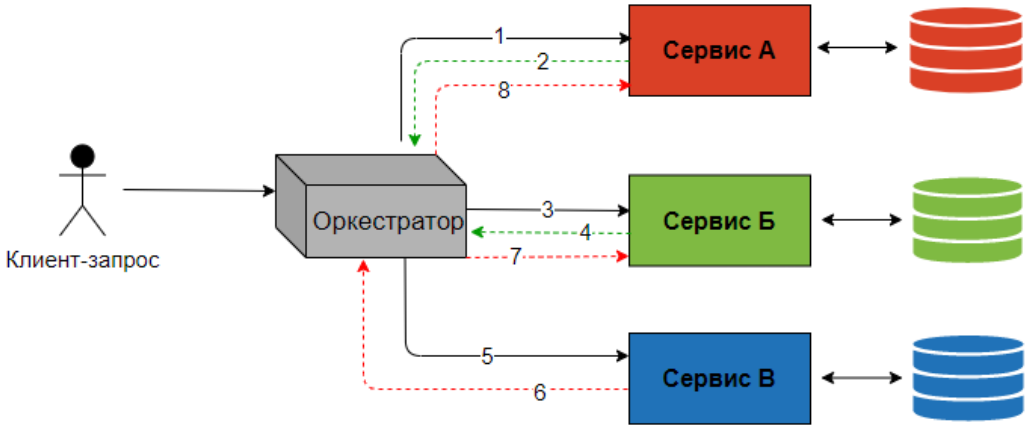


Рисунок 1.10 – МСА с централизованной логикой взаимодействия

где этапы 1,3,5 – запрос на обработку задачи, этапы 2,4 – положительный ответ о статусе выполнения задачи, этап 6 – отказ в обслуживании транзакции сервисом В, который приводит к необходимости отмены предыдущих изменений в сервисах Б (этап 7) и сервисах А (этап 8).

Данный принцип организации рабочего процесса позволяет масштабировать систему добавляя в неё новые сервисы. Однако управляющий элемент является дополнительной точкой отказа и в случае высокой нагрузки – «*узким местом*».

– *Событийно-ориентированная архитектура*, реализована таким образом что каждый микросервис знает, какую роль он играет в транзакции и как должен взаимодействовать с другими микросервисами, принятие решения происходит на основе событий. Элементы архитектуры взаимодействуют с использованием методов обмена сообщениями (командами) или других асинхронных механизмов коммуникации, для достижения целостности транзакции.

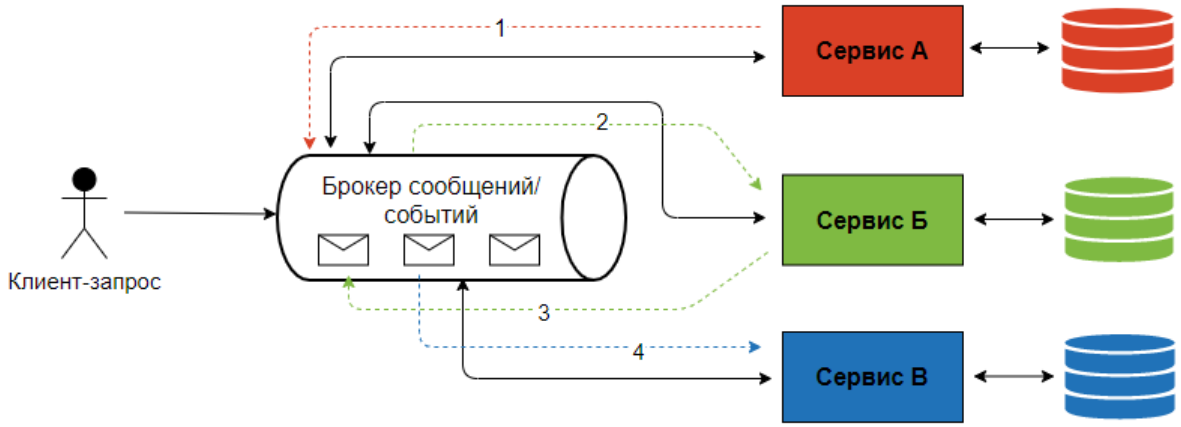


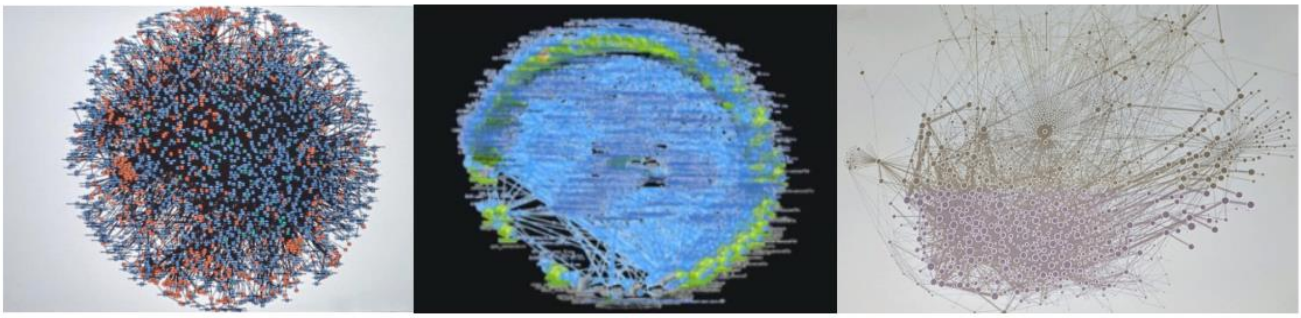
Рисунок 1.11 – МСА с распределенной коммуникацией через события

Данный способ исключает единый механизм управления как точку отказа, и применяет подход, где сервисы общаются между собой через брокера сообщений. Используя следующие этапы: этапы 1 и 3 – сервис А выполняет обработку и посылает сообщение брокеру, этапы 2 и 4 – сервисы прослушивают брокер сообщения.

Каждый подход имеет свои преимущества и соответствует определенным сценариям использования. Оркестратор обеспечивает более централизованное управление и контроль выполнения транзакции, в то время как распределенное управление предлагает подход, где каждый микросервис обладает большей автономией в принятии решений.

Для управления микросервисами и координации их работы крупные сервисы используют автоматическую генерацию рабочих процессов, чтобы управлять своими сервисами и повысить их эффективность. Они используют свои собственные фреймворки [81-84], которые позволяют разработчикам создавать, запускать и управлять масштабируемыми рабочими процессами, а также позволяет сервисам эффективно использовать свою инфраструктуру. Данные решения позволяют создавать процессы для обработки запросов на воспроизведение видео, управление потоками данных, обновление информации об аккаунтах и многое другое.

В автоматизации рабочих процессов нуждаются все компании со сложной архитектурой и огромным количеством микросервисов. Существует несколько примеров успешного использования автоматической генерации рабочих процессов для повышения эффективности своих бизнес-процессов и обеспечения качественного обслуживания своих клиентов в архитектуре микросервисов, примеры которых представлены на рисунке 1.12.



а) б) в)

Рисунок 1.12 – Карта микросервисной архитектуры популярных сервисов

Рабочие процессы могут быть реализованы в компактной микросервисной архитектуре для онлайн-магазина, которые охватывают несколько микросервисов, таких как «управление корзиной покупок», «обработка платежей», «управление складом и доставкой».

В работе [85] предложена методология и архитектура фреймворка для оценки масштабируемости каждой конфигурации разработки архитектуры. Авторы используют операционные профили с эмпирическими распределениями рабочих ситуаций и результаты производительности, которые были получены при нагрузочном тестировании, для оценки возникновения каждой рабочей ситуации в производственной среде. Это необходимо для оценки различных архитектурных конфигураций работы в условиях определенной рабочей нагрузки.

Авторы оценивают долю успешного выполнения всех услуг как:

$$\hat{s}(\lambda) = \sum_{j=1}^{n} \delta_j c_j \qquad (1.13)$$

где δ_j-доля сервисов, которые были оценены как успешные (по требованиям к масштабируемости) для набора из n сервисов, c_j = 1 в случае если $x(\lambda)_j < \Gamma_j(\lambda_0)$ – среднее время отклика по данным эксперимента не превышает базовый порог, $c_j = 0$ в ином случае.

Авторы работы так же указали что для общей применимости подхода было бы точнее разработать модель работы *сети очередей* для вычисления $\hat{s}(\lambda)$, что будет реализованно в данной диссертационной работе для сценариев рабочих процессов обслуживания задач в транзакционных системах.

1.6 Выводы по главе 1

Высокие требования к современным телекоммуникационным услугам появились в следствии высокого роста функциональных возможностей для разработки сервисов и веб-продуктов. Это привело к необходимости разработки услуг с высоким качеством (QoS), что подразумевало моделирование с целью прогнозирования для предоставления пользователям готового продукта.

В данной главе получены следующие результаты:

1) Проведен анализ методов и подходов к моделированию телекоммуникационных слабосвязанных услуг. Описан математический инструмент теории массового обслуживания необходимых для дальнейшего анализа.

2) Введена классификация телекоммуникационных услуг, для правильного выбора методов моделирования.

3) Определено архитектурное решение для построения современных высоконагруженных сервисов – микросервисная архитектура, особенности которой будут основой для будущего анализа и разработки моделей. Так же введено понятие рабочего процесса обработки запросов в *слабосвязанных* архитектурных элементах.

4) Вопросы оценки вероятностно-временных характеристик систем и использование предложенных методов моделирования являются основой данной диссертационной работы.

Последующая работа будет заключаться в решении задач по разработке моделей и методик их использования для оценки ВВХ слабосвязанных услуг.

ГЛАВА 2 МОДЕЛИ ОЦЕНКИ ВРЕМЕННЫХ ХАРАКТЕРИСТИК КОМБИНИРОВАННЫХ СЛАБОСВЯЗАННЫЕ УСЛУГ

2.1 Подход к разделению услуг по уровню взаимодействия с пользователем

Для последующего моделирования ИКУ предлагается подход к разделению услуг на сеансовые и транзакционные, в зависимости от типа взаимодействия между пользователем и системой.

Сеансовые услуги (SE) как правило связаны с продолжительным взаимодействием пользователя с системой, например, приложения для обмена сообщениями или онлайн-игры, где пользователь может проводить несколько часов или даже дней в системе. В таких приложениях часто используются авторизация и аутентификация пользователей, сохранение состояния сессии, и обработка запросов на основе текущего состояния. Сеансовые телекоммуникационные услуги являются услугами, которые предоставляются в рамках непрерывного сеанса связи между пользователями. Примерами сеансовых услуг могут служить видеоконференции, IP-телефония и многопользовательские онлайн игры.

Транзакционные услуги (TR), с другой стороны, связаны с краткосрочным взаимодействием пользователя с системой, они являются услугами, которые предоставляются для выполнения определенной операции или транзакции. Эти услуги часто связаны с передачей информации, обработкой данных и осуществлением финансовых операций. Примерами транзакционных услуг могут служить онлайн-банкинг, электронная коммерция и передача файлов.

Одним из главных отличий сеансовых и транзакционных услуг является подход к обеспечению требований по качеству обслуживания. Во-первых, время обслуживания T_{SE} напрямую зависит от поведения пользователя, иначе говоря количество времени необходимого пользователю на получение услуги. В данной работе не учитывается возможность влияния на поведение пользователя и его изменение, в том числе возможность временного ограничения, что может повлиять

на качество восприятия услуги (QoE). В транзакционных системах время обслуживания запросов зависит от характеристик системы, таких как производительность и надежность. Изменение этого аспекта системы зависит от производительности системы и возможности изменения характеристик программно-аппаратных средств.

Во-вторых, требования по возможности масштабирования и производительности системы, в разной степени затрагивают данные типы услуг. Так, для обеспечения высокого уровня качества обслуживания для пользовательских сессий необходимы грамотное распределение серверных ресурсов и возможность дальнейшего масштабирование как системы в целом, так и её составляющих. Что касается транзакционных услуг, для высоких показателей качества при обслуживании операций и пользовательских задач, необходима высокая скорость обработки и низкий уровень ошибочной обработки запросов.

В сравнении с временем пользовательской сессии T_{SE} время обслуживания запроса T_{TR} стремиться к нулю, то есть является незначительным. Следовательно, нецелесообразно включать время обслуживания в расчет времени нахождения пользователя в системе после инициирования сессии, на основе чего можно сделать вывод, в рамках моделирования сеансовых услуг рациональным является игнорирование занятия оборудования пользовательским запросом (TR) в указанных условиях.

Стоит учитывать, что единой модели для оценки данных типов услуг не существует, однако для расчета вероятностно-временных характеристик как сеансовых, так и транзакционных услуг, можно использовать математический аппарат теории массового обслуживания [86]. Как например в работе [87] был выполнен расчет сеансовой услуги с апробацией игрового сервиса с разделением ресурсов локаций.

В процессе разделения на микросервисы транзакционные сервисы могут быть как вынесены в отдельный микросервис, который будет отвечать за хранение и обработку данных, так и разбиты на отдельные микросервисы, выполняющие ограниченную по бизнес-функциям работу. Сеансовые сервисы в свою очередь

следует разделять на отдельные микросервисы, которые будут отвечать за конкретные функции приложения, например, аутентификацию, управление сессиями, работу с почтовыми сервисами, либо реализовывать функцию масштабирования для высоконагруженного сервиса. Этот подход к разделению на микросервисы реализует ряд преимуществ используя рабочие процессы взаимодействия между сервисами.

2.2 Принципы реализации и оценки характеристик комбинированных телекоммуникационных услуг

Сервис может предоставлять несколько услуг различного типа в рамках единой инфраструктуры. При этом, данные услуги могут взаимодействовать друг с другом и образовывать комбинированные услуги, которые могут являться более функциональными и удобными для конечного пользователя.

Выбор микросервисной архитектуры является рациональным для реализации комбинированных услуг, так как разбивает приложение на малые компоненты, каждый из которых может предоставлять несколько обособленных услуг. Это позволяет легче управлять каждой из услуг и их взаимодействием, а также более гибко настраивать и модифицировать функциональность в соответствии с потребностями пользователя.

Однако при разработке комбинированных услуг не стоит забывать, что сеансовые и транзакционные функциональности имеют разные требования к масштабируемости, надежности и производительности. Поэтому для каждой функциональности могут быть выделены отдельные ресурсы и инфраструктура, чтобы обеспечить их полноценную работу.

Примеры комбинированных услуг:

А. Интернет-коммерция

– Онлайн-магазин: пользователь может просматривать каталог товаров и добавлять их в корзину (сеансовая услуга), а затем совершать покупку и проводить транзакцию (транзакционная услуга).

– Система онлайн-бронирования отеля. Поиск доступных номеров и выбор их по критериям (сеансовая услуга), бронирование выбранного номера и оплата (транзакционная услуга).

B. Интернет-банкинг. Просмотр баланса и истории транзакций на счете (сеансовая услуга), осуществление переводов и платежей между своими счетами или на другие счета (транзакционная услуга).

C. Также можно использовать комбинированные микросервисы для управления игровым процессом, где сеансовые функции будут включать в себя сохранение игровых данных пользователей, а транзакционные функции – обработку платежей за покупку игровых ресурсов или дополнительных возможностей.

D. В социальных сетях можно использовать комбинированные микросервисы для управления личными данными пользователей, где сеансовые функции будут включать в себя отображение личных данных, а транзакционные – обработку запросов на их изменение.

Для обеспечения универсальности предлагаемых методов, предлагается рассмотреть моделирование абстрактного сервиса. Сценарий взаимодействия данных типов услуг может происходить по следующему сценарию:

1) Пользовательские запросы на обслуживание с интенсивность λ_{SE} поступают на узел обслуживания.

2) Далее запрос либо ожидает, либо начинает пользовательскую сессию в момент t_1 .

3) С этого момента t_1 он инициирует обращения к услугам обработки транзакций.

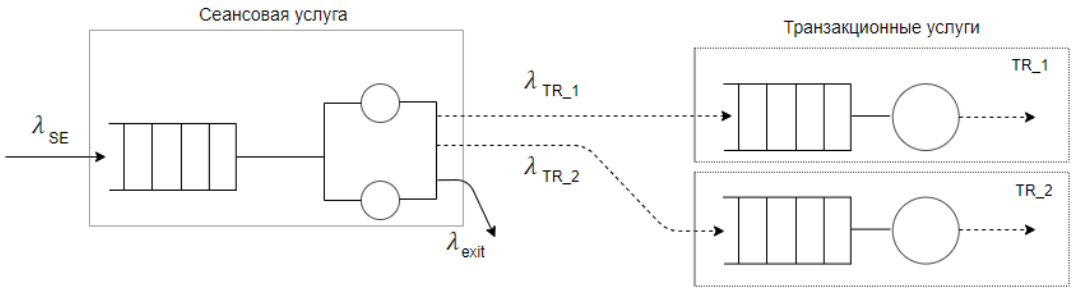


Рисунок 2.1 – Функциональная схема взаимодействие сеансовой и транзакционной услуги.

В данном сценарии взаимодействия, характер работы транзакционных услуг зависит напрямую от функционирования пользовательских сессий.

Из-за заметной разницы в функциональных требованиях и атрибутах качества, можно утверждать, что моделирование транзакционных и сеансовых услуг в рамках одной модели нецелесообразно. К такому выводу можно придти при анализе временных характеристик системы. На рисунке 2.2 изображена временная линия пользовательской сессии T_{se}в течении которой он инициирует два типа потоков на обработку задач T_{TR-1} и T_{TR-2} с интенсивностями $\lambda_{TR\_1} = \frac{1}{t_{RI\_1}}$ и $\lambda_{TR\_2} = \frac{1}{t_{RI\_2}}$ соответственно.

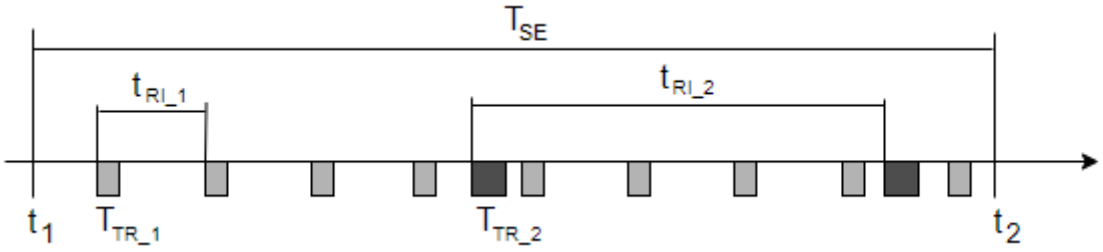


Рисунок 2.2 – Граничные значения распределения временного предоставления границ услуг Tr и Se типов.

Общая скорость поступления запросов от всех активных $\bar{K}$ сессий к TR рассчитывается следующим образом:

$$\lambda`_{TR\_i} = \lambda_{TR\_i} \times A_{SE} = \lambda_{TR\_i} \times \bar{t}_{SE} \times \lambda_{SE} \qquad (2.1)$$

где A_{SE} – нагрузка узла пользовательских сессий $A_{SE} = \lambda/\mu$, $\bar{t}_{SE}$ - средняя

продолжительность пользовательской сессии, λ_{SE}- интенсивность поступления пользователей на сервис, λ_{TR_i} – интенсивность потока запросов от S_i к TR.

Сценарий, при котором услуга пользовательских сессий обращается к транзакционной услуге, описывает обращение микросервиса к своей базе данных. Однако возможна иная реализация этого процесса, при котором несколько микросервисов инициируют запросы к одной базе данных как показано на рисунке 2.3.

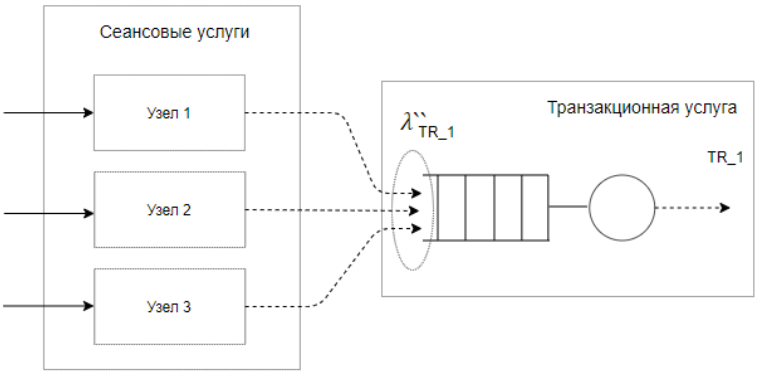


Рисунок 2.3 – Функциональная схема отправки обращений от SE-услуги к TR-услуге

В случае если разные микросервисы в результате обслуживания пользовательских сессий разной длительности $\bar{t}_{SE}\_j$ обращаются с разной интенсивностью λ_{TR_j} к услуге обслуживающей транзакции, интенсивность запросов на вход TR рассчитывается следующим образом:

$$\lambda``_{TR} = \sum_{j}^{Y_{SE}} \lambda_{j\_TR} \times A_j \qquad (2.2)$$

где Y_{SE} – количество сеансовых услуг, обращающихся к сервису обслуживания транзакций, A_j – нагрузка j-го узла пользовательской услуги.

В целом при рассмотрении телекоммуникационных услуг существуют входные параметры системы, которые влияют на качество её функционирования, и без определения которых невозможно провести полноценное моделирование:

– C = $\{c_1, c_2, \dots c_n\}$ – множество внешних воздействий на сервис: общая входная интенсивность, приоритеты поступления запросов или потока;

– B = $\{b_1, b_2, b_3 \dots b_n\}$ – множество внутренних характеристики систем: количество узлов Y, количество приборов v_i и интенсивность обслуживания μ_i в каждом узле, вероятность потерь $p_{\text{отк}\_i}$, i∈Y;

– H = $\{h_1, h_2, h_3 \dots h_n\}$– множество выходных характеристик системы. Стоит отметить, что для систем в рамках теории массового обслуживания все вероятностно-временные характеристики одинаковые, однако в контексте разделения на разные типы услуг, предложенного в данной работе, интерес к отдельным характеристикам может различаться, требуя использования различий в терминологии и методах их расчета, ВВХ описаны в таких стандартах как [55], [88]. Конкретные вероятностно-временные характеристики, идентифицирующие каждый тип услуг, будут описаны далее.

Решением поставленной задачи является разработка математической модели для расчета характеристик сеансовых услуг, определения их вероятностно-временных характеристик и степени воздействия на транзакционные. Так же разработка математической модели качественных характеристик транзакционных услуг, что позволит интегрировать математические модели в алгоритм расчета комбинированных слабосвязанных услуг.

2.3 Модели и методы вычисления задержек при предоставлении пользователям сеансовых услуг

2.3.1 Услуги поддерживающие пользовательские сеансы

Определение "сеансовые услуги" в IT-индустрии производится на основе общего понятия "сеанс" в компьютерных системах, которое относится к периоду работы программы или взаимодействия между пользователем и компьютерной системой.

В стандартах и рекомендациях таких как [55],[89],[90] определение сеансовых услуг представлено в контексте телекоммуникационных сетей и связи,

каждый из которых уточняет контекст и особенности использования данного термина в своей области.

"Сеансовая услуга (session service) — это услуга, которая обеспечивает установление, поддержание и завершение коммуникационной сессии между двумя или более пользователями, а также передачу данных между ними в течение этой сессии." - определена в стандарте ITU-T E.800 [55].

"Услуга, которая предоставляет конкретную функцию клиенту в течение определенного временного периода, включая обработку информации или использование системных ресурсов" – определена стандартом [91]. Так же отмечено, что сеансовая услуга может включать взаимодействие с другими услугами или системами.

В стандарте ITIL (Information Technology Infrastructure Library) не предоставляется явное определение сеансовых услуг, однако в рамках процесса управления уровнем обслуживания (Service Level Management) упоминаются сеансовые параметры, такие как время отклика и время доступности, которые могут использоваться для оценки качества сеансовых услуг.

В данной работе вероятностно-временные характеристики для сеансовых услуг определены как:

— Коэффициент загрузки узлов системы — это показатель, отражающий степень использования ресурсов конкретного узла в рамках распределенной системы;

— Среднее время ожидания обслуживания — это среднее время, начиная от момента поступления запроса в систему до момента получения им обслуживания;

— Вероятность ожидания начала пользовательской сессии — это вероятность того, что пользователь будет ожидать, прежде чем сможет начать использовать систему;

— Интенсивность поступления пользовательских запросов на конкретный узел, для системы, состоящей из нескольких узлов, параметр, определяющий долю запросов из общей интенсивности Λ поступающую на выбранный узел λ_i;

— Среднее время нахождение в системе – это время, рассчитанное от входа пользователя в систему, до его выхода из неё;

— Пропускная способность — это количество запросов, которые система может обработать за единицу времени.

Возможные реализации сеансовых услуг в контексте веб-сервисов:

1) Мессенджеры, такие как WhatsApp, Telegram, Viber и другие, которые обеспечивают обмен сообщениями в режиме реального времени между пользователями.

2) Платформы для видеоконференций, такие как Zoom, Microsoft Teams, Skype и другие, которые позволяют пользователям проводить видеоконференции и общаться друг с другом в режиме реального времени.

3) Интернет-телефония, которая обеспечивает передачу голосовой информации по сети Интернет. Например, такие сервисы, как Skype, Viber, WhatsApp, могут использоваться для проведения голосовых звонков.

4) Игры, как уже упоминалось выше, которые могут использовать различные технологии для обеспечения коммуникации между игроками в режиме реального времени.

5) Системы видеонаблюдения и безопасности, которые могут использоваться для мониторинга и управления безопасностью объектов в режиме реального времени.

6) Системы управления производством, которые обеспечивают мониторинг и управление производственными процессами в режиме реального времени.

7) Также существуют специализированные системы телемедицины, которые позволяют хирургам и другим медицинским специалистам проводить консультации и общаться между собой в режиме реального времени.

Использование таких технологий позволяет улучшить координацию между хирургами и другими медицинскими специалистами, повысить качество медицинской помощи и улучшить результаты лечения.

2.3.2 Модели и методы вычисления задержек при предоставлении пользователям сеансовых услуг

Для получения качественной оценки разрабатываемой телекоммуникационной многопользовательской услуги, необходимо определить её входные характеристики. Начнем с того, что для разработки модели расчета обслуживаемой нагрузки на сеансовую модель системы, описывающую функционал представленных выше многопользовательских услуг, следует провести логическое разделение обслуживающих серверных ресурсов. Далее необходимо определить входные характеристики, такие как:

— Характеристики времени между поступлением запросов пользователей в систему.

— Вероятности перехода пользователей между узлами системы.

— Характеристики времени пребывания запросов пользователей в каждом узле системы.

В работах [92],[93] авторы проводят моделирование web-сервисов и многопользовательских игровых сервисов и выдвигают предположении о том, что переходы между состояниями системы, в том числе перемещение между серверами игровой логики описывается целью Маркова. В [94] описан алгоритм подготовки к тестированию сервиса и его результаты опираясь на пользовательское поведение.

Проблематика оценки временных характеристик сеансовых услуг основана на отсутствии четкого определения пользовательского поведения, предсказать поведение и временные характеристики для анализа возможном следующими методами:

1) Ретроспектива (оценка предыдущих значений).
2) Анализ схожих по функционалу сервисов.

3) Аналитика данных функционирующего сервиса, для моделирования системы на этапе корректировки.

Для услуг, реализующих обслуживание пользовательских сеансов, параметры временных характеристик были определены в работах [95-99], описывающих возможные распределения для входного потока – как экспоненциальное, обратное Гауссовское, Вейбулла, в дополнение к данным распределениям определены процессы обслуживания [100-102],[93] характеризующиеся смешанным экспоненциальным, распределением Парето и др.

Методы для получения входных характеристик описаны в работах [103-104] где авторы предлагают два способа измерения параметров игровых сессий: измерение сетевого трафика и получение информации из трассировок, или внедрение дополнительного ПО в устройства игроков для измерения продолжительности сеанса.

Закон Парето [105] определен следующей функцией (2.3) и плотностью (2.4) распределения вероятности:

$$F(x) = P(X < x) = 1 - \left(\frac{\sigma}{x}\right)^k, \quad x \geq \sigma \tag{2.3}$$

$$f(x) = \begin{cases} \frac{k\sigma^k}{x^{k+1}}, & x \geq \sigma \\ 0, & x < \sigma \end{cases} \tag{2.4}$$

Моменты случайной величины и дисперсия, имеющей распределение Парето, задаются формулой:

– оценка n-го момента:

$$E[X^n] = \frac{k\sigma^n}{k-n} \tag{2.5}$$

– дисперсия:

$$D[X] = (\frac{\sigma}{k-1})^2 \times \frac{k}{k-2} \tag{2.6}$$

Помимо этого, длиннохвостые распределения для определения длительности игровой сессии упоминаются в работах [106][107], в последней предложен метод решения проблем с синтезированием распределения Вейбулла-Гнеденко, при аппроксимации его совместным распределением трех случайных величин. Данный метод, и метод аппроксимации распределения Вейбулла-Гнеденко гиперэкспоненциальным распределением [108] позволяет упрощать расчет «сложных» распределений для определения качественных характеристик системы.

Распределение Вейбулла-Гнеденко X~W(b,a) определено следующими выражениями:

– функция распределения вероятности:

$$F(x) = 1 - e^{-\left(\frac{x}{a}\right)^b}, \ x \geq 0 \tag{2.7}$$

– плотность распределения вероятности:

$$f(x) = \frac{b}{a} \times (\frac{x}{a})^{b-1} \times e^{-(\frac{x}{a})^b} \tag{2.8}$$

– оценка n-го момента случайной величины

$$E[X^n] = a^n \Gamma(1 + \frac{n}{b}) \tag{2.9}$$

где Γ- гамма-функция:$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$, b - коэффициент формы b>0, a коэффициент масштаба a>0.

Как видно из [99],[95] продолжительность пользовательского сеанса как на одном узле, так и в системе в целом согласуются с экспоненциальным распределением. Данное распределение мы возьмем за основу последующего моделирования. Поэтому следует указать причины:

– В случае проектирования системы данный вид моделирования достаточно прост в расчете что позволяет проводить глубокий анализ ВВХ и определять узкие места в структуре СеМО. Это является достаточным на этапе планирования систем, без учета реальных данных.

– Они обладают свойством отсутствия памяти, и означает, что вероятность того, что запрос будет обслужен в течение определенного времени, не зависит от того, сколько времени запрос уже находится в очереди.

– Экспоненциальное распределение часто используется для моделирования случайных событий, таких как случайное поступление запросов, или отказа оборудования.

Разрабатываемая модель включает в себя M модулей, обслуживающих пользовательские сессии и обозначены они будут $S_1, S_2 \dots S_M$. Мы предполагаем использование математического аппарата сетей Джексона для расчета качественных характеристик сервиса, следовательно весь сервис следует рассматривать как сеть систем массового обслуживания (СеМО).

Каждый модуль является многоканальной системой массового обслуживания (СМО) с бесконечной очередью и принципом обслуживания FIFO (First - In - First - Out - «первым пришел - первым обслужен»).

На рисунке 2.4 представлена разрабатываемая абстрактная модель общего вида для многопользовательского сервиса. Она состоит из блока генерации входящей пользовательской нагрузки S_0 на обслуживание модулями системы, а также блока, в который поступают запросы после окончания обслуживания. Мы указываем на то, что нулевой блок не является частью сети систем обслуживания запросов, поэтому он был введен как внешняя среда, взаимодействующая с системами.

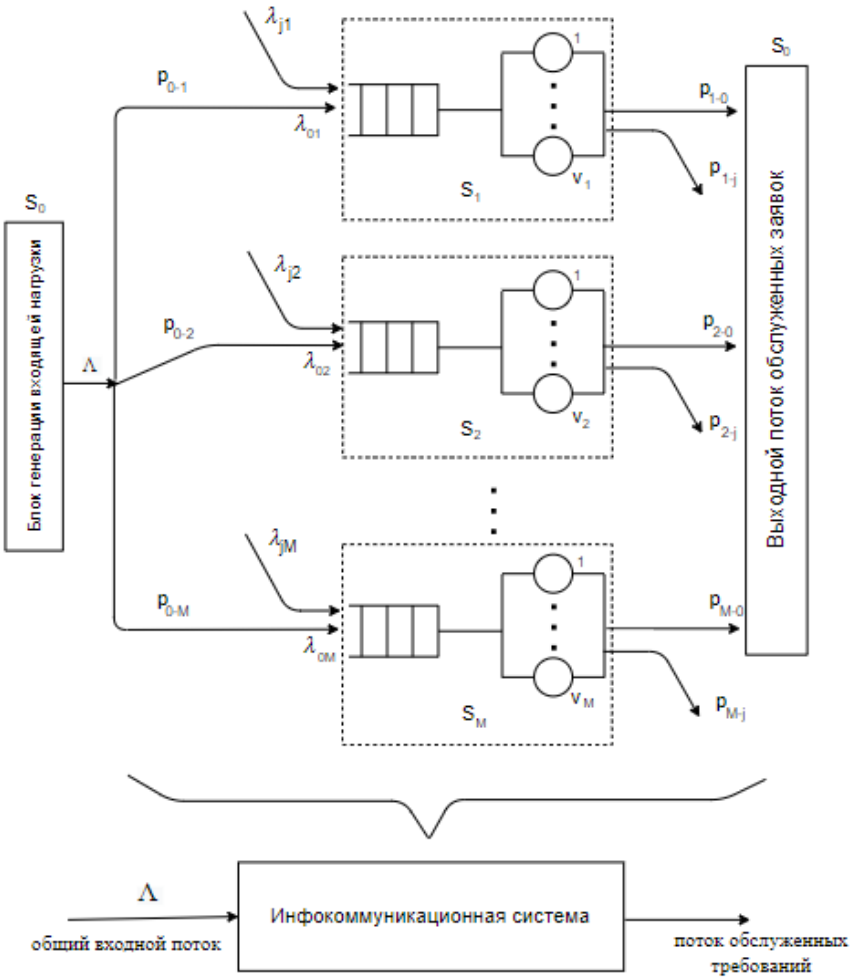


Рисунок 2.4 – Обобщенная модель многопользовательского сеансового сервиса

Однако в ранее указанной работе [109], авторы предложили иной способ моделирования - замкнутую сеть систем массового обслуживания, с подсистемой моделирующей активные сессии приложения как показано на рисунке 2.5. Сеансы моделируются с помощью системы массового обслуживания, которая генерирует запросы на систему. Предполагается, что каждый сеанс занимает один сервер в S_0, а время, проведенное в этом сервере, моделирует время размышления пользователя перед следующей активной сессией в сети обслуживания. Однако авторы так же признают, что для такой модели предполагается, что сеансы никогда не будут завершены, но на практике их количество будет меняться по мере завершения существующих и появления новых сеансов.

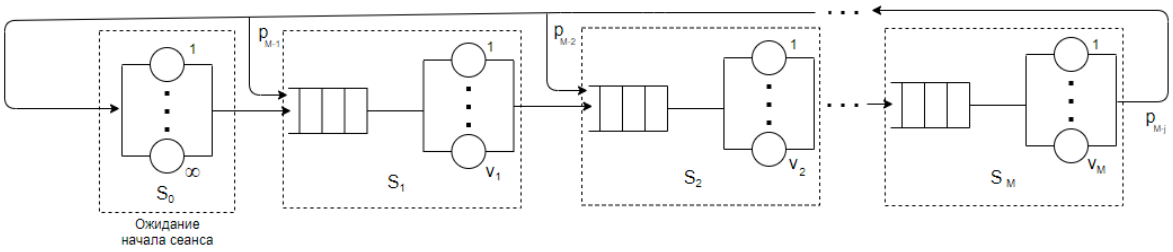


Рисунок 2.5 – Альтернативная модель обслуживания сервисов

Обозначенное на рис. 2.4 архитектурное решение является более удобным для анализа и дальнейшей разработки математической модели, поэтому обращаться далее будем именно к нему.

Из подсистемы S_0 на обслуживающие приборы поступает общая входящая нагрузка с интенсивностью Λ, и далее она с определенной вероятностью попадает на одну из M систем. После поступления пользовательского запроса на один из узлов СеМО, выделенный узел будет указан как S_i, происходит его обслуживание одним из свободных обслуживающих приборов из n_i возможных, с временем обслуживания $t_{\text{обсл}.i}$. После обслуживания пользовательского запроса в СМО, перейти в другой узел с вероятностью p_{i-j} для продолжения обслуживания своей пользовательской сессии. Вероятность переходов из узла S_i в узел S_j , как и вероятность перехода из нулевого узла на один из узлов системы обработки запросов, для удобства могут быть заданы матрицей $M \times M$ вероятностей переходов $R = \|p_{ij}\|$, в случае системы из четырех узлов:

$$R = \begin{Vmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{Vmatrix} \tag{2.10}$$

Она позволяет упростить расчет входной интенсивности, поступающий на каждый узел сети, рассчитываемый по теореме Джексона:

$$\lambda_i = \Lambda p_{0i} + \sum_{j=0}^{M} \lambda_j p_{ji} = \Lambda p_{0i} + \lambda_j R \tag{2.11}$$

Однако для рассмотрения открытой системы, не допускается бесконечная циркуляции пользовательских сессий в системе, из чего следует существование вероятности окончательного выхода из системы $p_{i-0} = 1 - \sum_{j=1}^{M} p_{i-j}$.

Мы рассматриваем непрерывные случайные переходы по узлам сети массового обслуживания с интенсивностями переходов $\lambda_{\alpha\beta}$ из состояния $\alpha = (\alpha^1, \ldots, \alpha^M)$ и $\beta = (\beta^1, \ldots, \beta^M)$,

$$\lambda_i = \begin{cases} \mu_{0i} \doteq \lambda_i & \text{если } \beta - \alpha = e_i \\ \mu_{i0} \doteq \mu_i p_{i0} & \text{если } \beta - \alpha = -e_i \\ \mu_{ij} \doteq \mu_i p_{ij} & \text{если } \beta - \alpha = -e_i + e_i \end{cases} \quad (2.12)$$

Это необходимо для дальнейшего расчета временных характеристик скомпонованных комплексных услуг.

2.4 Математический инструмент расчета временных характеристик сеансовых услуг

В [110] определено стационарное среднее время $T_{\text{сист.}}$ длительности пользовательской сессии в системе S, представленное в виде:

$$T_{\text{сист.}} = \sum_{i=1}^{Y} a_i u_i = \sum_{i=1}^{Y} a_i (w_i + t_{\text{обсл.}i}) \quad (2.13)$$

где a_i - среднее число посещений пользователем i-го узла за все время пребывания в системе, u_i- среднее время пребывания запроса, w_i – среднее время ожидания обслуживания пользователя, $t_{\text{обсл.}i}$ – среднее время получения пользователем услуги в i-м узле.

Определить значения коэффициентов a_i позволяет система уравнений баланса:

$$a_i = \sum_{i=1}^{Y} a_j \times p_{ji} \ , \quad i = \overline{1, Y} \ \ a_0 = 1 \quad (2.14)$$

Полученные данные позволяют определить интенсивности входных потоков узлов $\lambda_i = \Lambda \times a_i$, которые, в свою очередь, необходимы для оценки качества функционирования рассматриваемых телекоммуникационных систем, за счет

получения таких выходных характеристик как: коэффициенты загрузки узлов системы ρ_i, средняя длинна очереди L_i и время ожидания на узлах системы w.

Если интенсивность входного потока была нами определена, то интенсивность обслуживания задается следующим выражением:

$$\mu_n = \begin{cases} n\mu, & 0 \le n < v \\ v\mu, & n \ge v \end{cases} \quad (2.15)$$

где v – количество обслуживающих приборов, μ - интенсивность обслуживания.

В сеансовых услугах одной из качественных характеристик определяющей уровень обслуживания пользователя является время ожидания обслуживания, для его расчета используем формулу Литтла:

$$\overline{W} = \frac{\overline{L}}{\lambda} \quad (2.16)$$

где λ – интенсивность входного потока, $\overline{L}$ – средняя длина очереди.

Как уже говорилось ранее вычисление качественных характеристик возможно в стационарном состоянии системы, когда $A_i < v_i$, где $A_i = \lambda_i/\mu_i$ – поступающая нагрузка.

Средняя длина очереди определяется выражением:

$$\bar{L} = \sum_{k=v+1}^{\infty} (k - v) * [k] = \frac{A \times v}{(v - A)^2} \times P_v \quad (2.17)$$

где P_v - вероятность состояния задана выражением:

$$P_v = \frac{A^V}{v!} P_0 \quad (2.18)$$

где P_0 - вероятность отсутствия запроса в системе определяется выражением:

$${P_0}^{-1} = \sum_{x=0}^{v} \frac{A^x}{x!} + \frac{A^V}{v!} \times \frac{A}{v - A} \quad (2.19)$$

Коэффициент загрузки узла i при известной входной интенсивности на данный узел, рассчитывается выражением:

$$\rho_i = \frac{\lambda_i t_{\text{обсл}.i}}{v_i} \quad (2.20)$$

Условие стационарного состояния системы $\rho_i < 1$ - гарантирует отсутствие перегрузки в системе.

Из-за возможностей современный телекоммуникационных технологий социальные сети способны передавать текстовую, аудио- и видеоинформацию. Это позволяет утверждать, что время получения пользователем услуги может варьировать от малых значений (прочтение текстовой информации) до больших значений (просмотр вложенного видео).

В качестве примера применения математического инструмента сетей Джексона, предложен расчет сеансовой услуги для абстрактного сервиса (рис 1.2) схема которой изображена на рис.2.6. Данная система состоит из четырех взаимосвязанных узлов SE_i, реализующих свою функциональность.

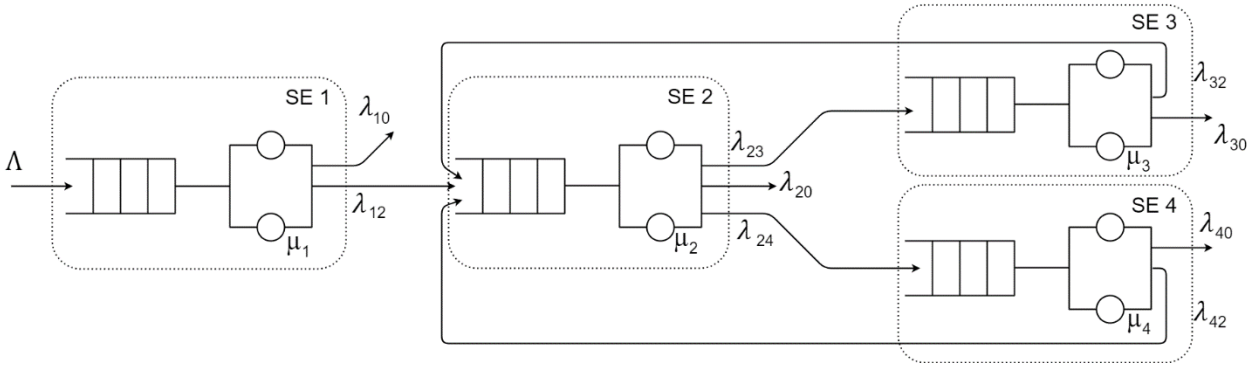


Рисунок 2.6 – Математическая модель сеансовой услуги

Для расчета интенсивности входного потока λ_1 на конкретную систему s_i необходимо использовать матрицу вероятности переходов, включающую узлы системы с учетом внешнего поступления (нулевого узла). Расчет интенсивности входного потока будет включать в себя решение системы уравнений для всех узлов:

$$\begin{cases} \lambda_1 = \Lambda * p_{01} \\ \lambda_2 = \lambda_1 * p_{12} + \lambda_3 * p_{32} + \lambda_4 * p_{42} \\ \lambda_3 = \lambda_2 * p_{23} \\ \lambda_4 = \lambda_2 * p_{24} \end{cases} \quad (2.21)$$

Получив результаты моделирования в данной модели узел «Se 2» был определен как высоконагруженный. На что влияет структура и взаимосвязи между элементами системы, в связи с этом данный узел получает потоки запросов на обслуживания со всех узлов. Зависимость средней задержки на первом и втором

узле от входной нагрузки представлены на рисунке 2.7 для разных вероятностей переходов на узел p_{j2}.

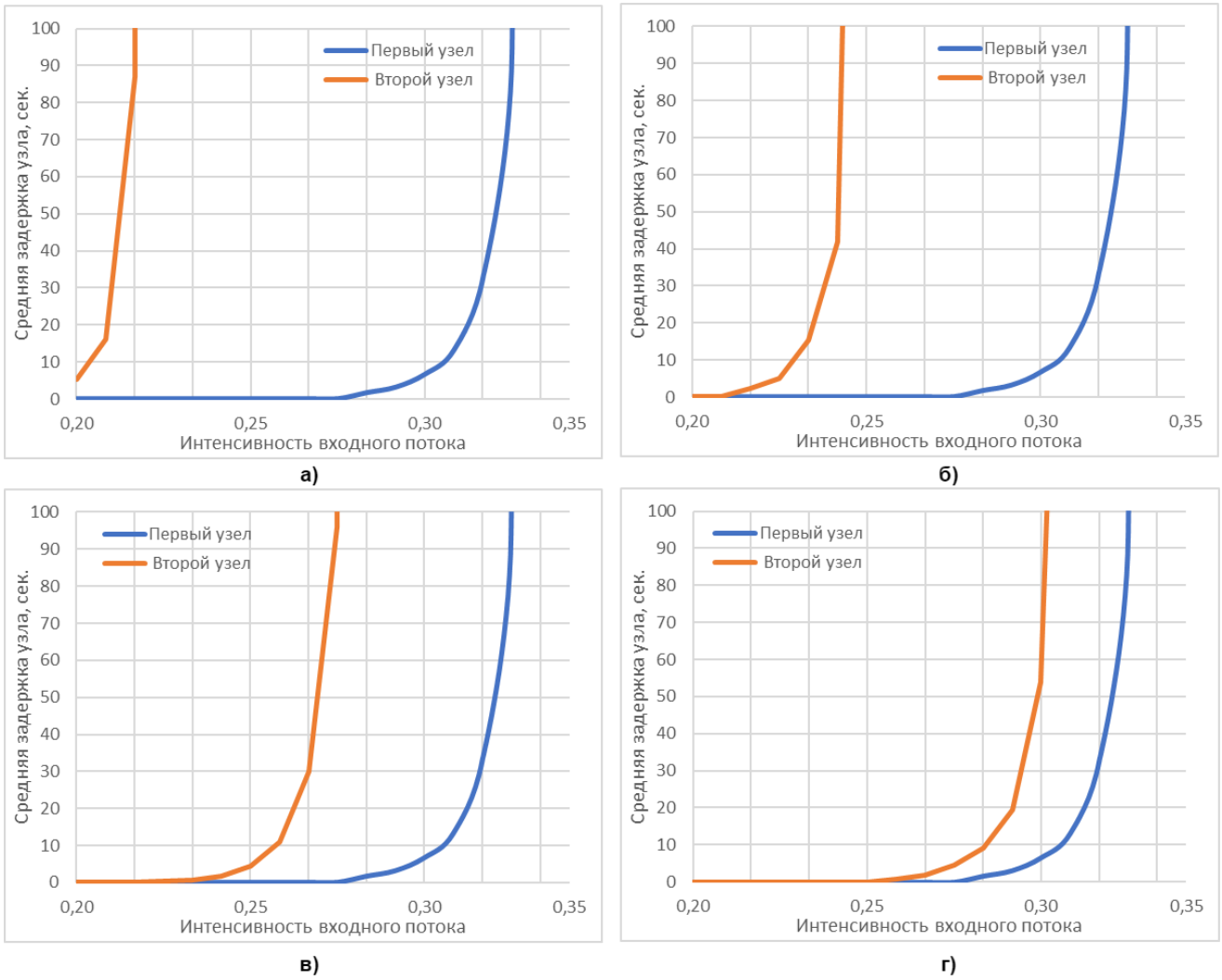


Рисунок 2.7 – Графики значений средних задержек для первого и второго узла в СеМО для разных вероятностей переходов p_{j2} а) 0,5 б) 0,4 в) 0,3 г) 0,2.

Как видно из представленных графиков, при уменьшении вероятности перехода потока на второй узел, а значит уменьшении входного потока λ_2 , средняя задержка на втором узле снижается и стремится к среднему значению для всех узлов СеМО.

Так же модель позволяет определить среднее время пользовательской сессии во всей системе при использовании общего выражения для получения общего времени в системе:

$$S_{\text{общ.}} = T_{\text{общ.}} + W_{\text{общ.}} \quad (2.22)$$

где среднее временя пребывания в СМО S определено выражением:

$$S = T_w + T_s = \frac{1}{\mu}\frac{1}{(1-\rho)} \quad (2.23)$$

Стоит отметить, что при изменении интенсивности входного потока, чистое время обслуживания в системе без распределения нагрузки $T_{\text{общ.}} \sim const$. Для предоставления высокой точности его можно получить, используя имитационное моделирование, модель которого представлена на рисунке 2.8.

Однако при изменении входного потока, возможно увеличение очереди в высоконагруженном узле, а вместе с тем и $W_{\text{общ.}}$. Чтобы не проводить в ИМ множество итераций при разных значениях Λ, рекомендуется использовать общую математическую модель с выражениями (2.22) - (2.23) и известное $T_{\text{общ.}}$

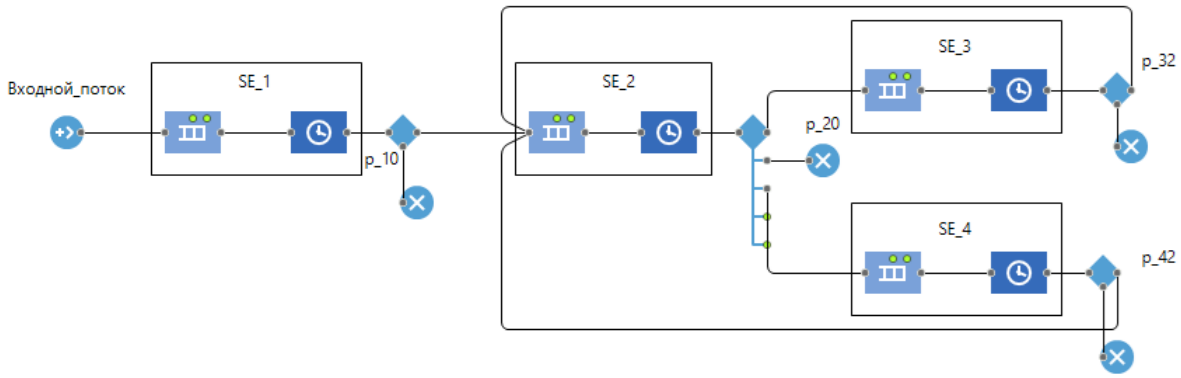


Рисунок 2.8 – Имитационная модель сеансовой услуги

В качестве результата анализа модели рис. 2.9 представлены графики зависимости среднего времени пользовательской сессии (S) и времени ожидания обслуживания (W) от общей входной интенсивности Λ. Как видно из рисунка 2.9 среднее время пользовательской сессии не имеет сильной тенденции к росту, однако изменяется, когда начинается ярко выраженное увеличение времени ожидания на узлах. Это связано с архитектурным решением, где все узлы сети идентичны, однако суммарная нагрузка на второй узел как видно из выражения (2.21) очевидно выше. Следовательно, система не может обслужить поток клиентских запросов на обслуживание который превышает 0,222 пользовательского запроса в секунду.

Рисунок 2.9 – Оценка временных характеристик CeMO

Очевидным решением является увеличение обслуживающих ресурсов на высоконагруженном узле. Что позволяет увеличить входящую нагрузку до 0,332 польз.запрос./сек. как видно на рис. 2.11.

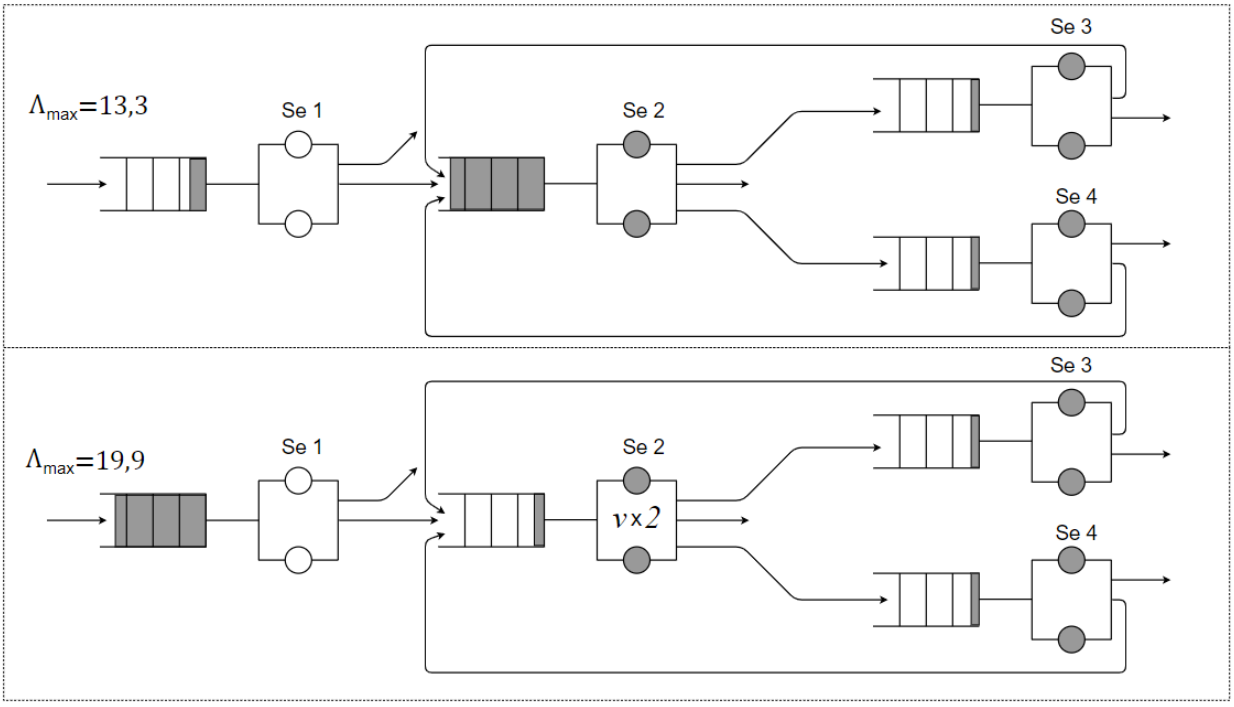


Рисунок 2.10 – Изменение входных параметров рассматриваемой CeMO.

Однако далее высоконагруженным узлом становится первое СМО, которое является единой точной входа нагрузки на CeMO, из чего характеристики системы изменяются как показано на графике (рис. 2.11).

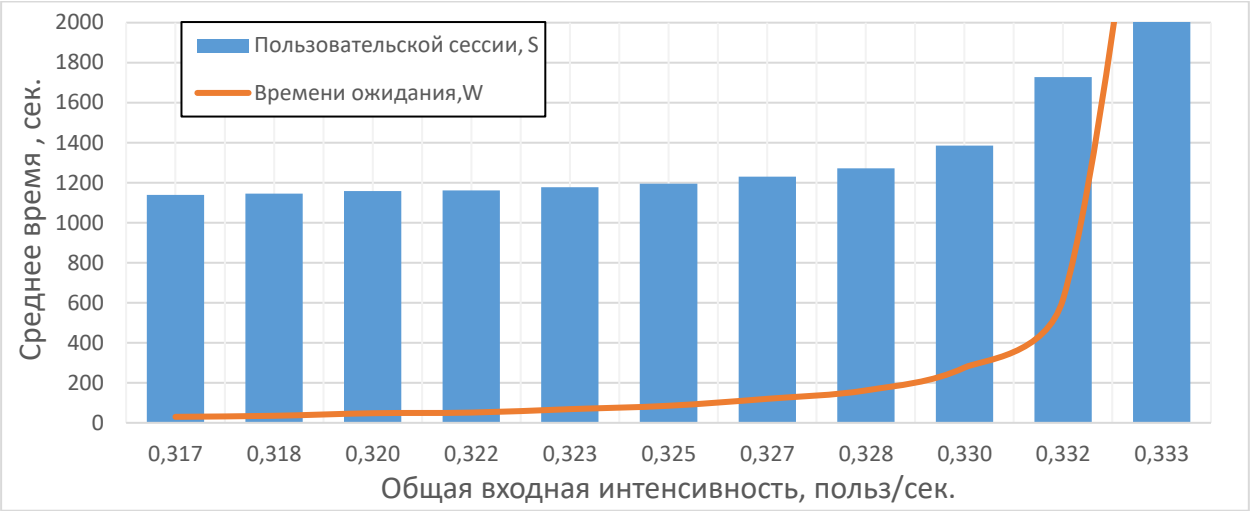


Рисунок 2.11 – Оценка временных характеристик СеМО

Из результатов данного анализа, можно сделать вывод что для обеспечения высокого качества функционирования сервиса необходимо проводить моделирование системы и получение требований к входным характеристикам его структурных элементов.

2.5 Выводы по главе 2

В результате проведенного моделирования можно утверждать, что в стабильной системе суммарная задержка заметно меньше общего времени обслуживания. Однако в сеансовых услугах как в реализации сетей массового обслуживания существует необходимость решения проблем, связанных с высоконагруженными узлами. Для этого необходимо внедрять методы математического и имитационного моделирования, комбинация которых позволит сократить трудозатраты как было описано в последнем разделе данной главы.

В данной главе диссертационной работы получены следующие результаты:

1) Разработан и формализован подход к разделению ИК услуг по степени взаимодействия с пользователем:

— Сеансовые – услуги реализующие коммуникационные пользовательские сессии.

— Транзакционные – услуги, предоставляющие возможность выполнения системных операций.

— Комбинированные – реализация сеансовых и транзакционных услуг в одном сервисе, для предоставления пользователю полной функциональности.

2) Исследованы особенности взаимодействия разных типов услуг в единых архитектурных решениях, а также подход к их моделированию.

3) Разработаны аналитические и имитационные модели для оценки временных характеристик сеансовых услуг веб-сервиса.

Проведен подробный анализ сеансового типа услуг, что подводит нас к дальнейшему решению поставленных задач, а именно к разработке математического инструмента для расчета ВВХ характеристик рабочих процессов транзакционных услуг

ГЛАВА 3 МОДЕЛЬ ОЦЕНКИ ХАРАКТЕРИСТИК ТРАНЗАКЦИОННЫХ УСЛУГ С ПАРАЛЛЕЛЬНОЙ ОБРАБОТКОЙ ЗАПРОСОВ

3.1. Транзакционные услуги – сценарии выполнения рабочих процессов.

3.1.1 Характеристики транзакционных услуг

Как было определено ранее, реализация транзакционных услуг необходима для предоставления возможностей выполнения системных операций. Данный тип услуг характеризуется высокой степенью взаимодействия с программно-аппаратной средой, высокой частотой запросов и быстрым выполнением операций. Поэтому системы, обслуживающие транзакционные услуги, должны быть производительными и надежными, чтобы обеспечить быстрое и точное выполнение операций, сократить время ответа на запросы и обеспечить надежную передачу данных.

Качество обслуживания TR-услуг рассматривается в работе [111]. В [112] предложены решения для уменьшения задержки в обработке сложных транзакций в ЦОДах. При анализе исследований было определено множество выходных характеристик H = $\{h_1, h_2, h_3 \ldots h_n\}$ для транзакционных услуг как следующий набор ВВХ

Время пребывания запроса в системе: это математическое ожидание и дисперсия времени, которое требуется системе для обработки запроса и отправки ответа. Для большинства транзакционных систем время отклика должно быть минимальным.

– Время обслуживания, которое требуется системе для выполнения операции. Для транзакционных систем время выполнения должно быть также минимальным.

– Время простоя - время, когда система не занята обработкой запросов. Чем меньше время простоя системы, тем выше её продуктивность.

– Надёжность системы: это вероятность того, что система сможет обработать запрос.

– Пропускная способность определяет количество запросов, которое система может обработать за единицу времени. Для транзакционных систем пропускная способность должна быть достаточно высокой для удовлетворения потребностей пользователей.

– Время ожидания - время, которое запрос тратит на ожидание, прежде чем будет выбран для обработки.

– Время процесса: время, затрачиваемое рабочим процессом на выполнение задачи в целом.

3.1.2 Транзакции – строительные блоки рабочих процессов

Транзакция — это логическая единица работы в рабочем процессе которая обычно включает в себя одну или несколько операций работы с данными или запросами на обслуживание. Транзакция должна обладать следующими свойствами:

– Атомарность. Транзакция должна быть выполнена целиком, или не выполнена совсем. Если какая-либо операция в транзакции завершится с ошибкой, все операции транзакции должны быть отменены, и база данных должна быть возвращена к состоянию, которое было до начала транзакции.

– Согласованность. Транзакция должна сохранять целостность данных. Если данные в транзакции нарушают ограничения целостности, транзакция должна быть отменена.

– Изолированность. Транзакция должна быть изолированной от других транзакций, которые выполняются параллельно. Транзакции должны выполняться одна за другой и не должны взаимодействовать друг с другом.

– Устойчивость. После завершения транзакции изменения, сделанные в базе данных, должны быть сохранены навсегда, даже в случае сбоя системы.

Отличительные свойства транзакций подтверждают рациональность реализации транзакционных услуг на базе микросервисной архитектуры. В таком случае каждая транзакция выполняется в рамках одного микросервиса.

3.1.3 Рабочие процессы в рамках транзакционных систем

В рамках исследования микросервисной архитектуры уже была затронута тематика рабочих процессов. Рабочий процесс как было описано ранее, является совокупностью установок для передачи запросов между элементами слабосвязанной структуры и обслуживания каждым из них.

Несколько исследователей определили рабочие процессы как вычислительную модель, которая позволяет стандартному методу создания приложений и процессов веб-сервисов подключаться и обмениваться информацией через Интернет [113-114]. Однако для реализации транзакционных услуг и сценариев их работы необходимо определить принципы реализации рабочих процессов.

В транзакционных системах, рабочий процесс (workflow) как правило описывает последовательность действий, необходимых для обработки транзакции (задачи). Стоит отметить, что рабочий процесс так же может являться набором шагов для выполнения локальной транзакции, таких как:

– Начало транзакции, когда система получает запрос на выполнение операции.

– Валидация - определение данных необходимых для выполнения операции, проверяются на правильность и целостность.

– Обработка –выполнение операции.

– Подтверждение – сохранение результата.

– Завершение транзакции.

В данной научной работе рабочий процесс описывает сценарий обслуживания задачи цепочкой локальных транзакций, т.е. последовательное выполнение действий для решения поставленной задачи. Сценарий и его

реализация зависят от поступающей задачи и набора сервисов, которые необходимы для её выполнения. В дальнейшей работе мы будем рассматривать именно этот сценарий реализации, так как нас интересуют вероятностно-временные характеристики, которые в свою очередь зависят от архитектуры системы и параметров её элементов.

В транзакционных системах существует несколько видов рабочих процессов (workflows), которые могут быть использованы для обработки транзакций. Некоторые из них включают в себя структуры, изображённые на рисунке 3.1:

– *Последовательный рабочий процесс* (а) (Sequential Workflow): в этом типе рабочего процесса каждый шаг выполняется последовательно, в том порядке, в котором они определены. Пока один шаг не будет выполнен, система не перейдет к следующему.

– *Параллельный рабочий процесс* (б) (Parallel Workflow): в этом типе рабочего процесса несколько шагов выполняются параллельно. Каждый шаг может завершиться независимо от других шагов, и система продолжит работу только после того, как все шаги будут завершены.

– *Правила* (в) (Rules-Based Workflow): в этом типе рабочего процесса, шаги определяются правилами, которые могут быть настроены для определенных условий. Если условие выполняется, то система выполняет определенный шаг, иначе переходит к следующему шагу.

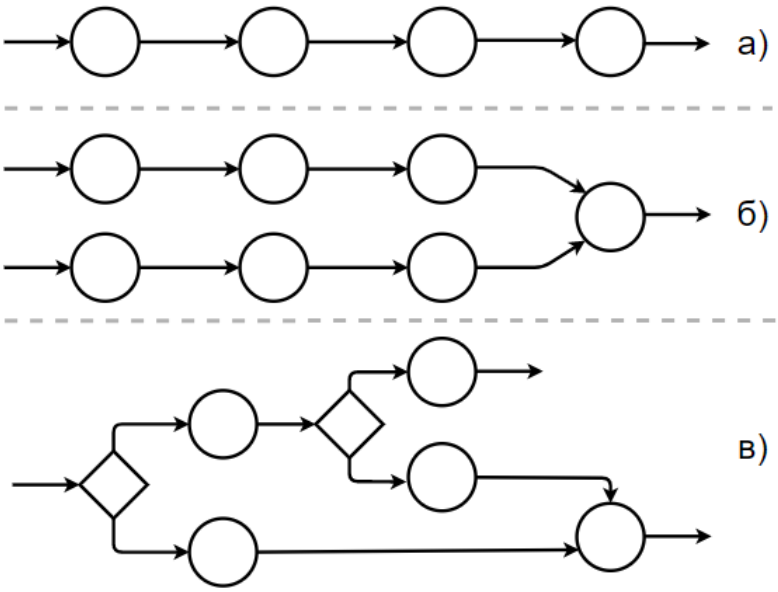


Рисунок 3.1 - Основные типы рабочих процессов.

Гибридный рабочий процесс (Hybrid Workflow): Этот тип рабочего процесса комбинирует два или более типов рабочих процессов, чтобы получить наилучший результат для конкретной задачи.

Выбор определенного типа рабочего процесса зависит от конкретной задачи, которую необходимо решить, а также от требований к производительности и надежности системы.

При разработке транзакционной услуги, основанной на рабочем процессе обслуживания задачи необходимо определить сценарий работы, графическое представление может быть реализовано с помощью UML диаграммы. К данному типу диаграмм достаточно часто обращаются при разработке сервисов, так в [115] предлагается метод, использующий модели действий UML для разработки композиции составных веб-сервисов, с возможностью повторного использования базовых сервисов при их разработке. Так анализ транзакционной услуги по авторизации пользователей (рисунок 1.2) позволяет изобразить ее в виде следующей диаграммы на рисунке 3.2.

Стоит уточнить что рассматриваемый сервис имеет следующий функционал: обрабатывает информацию о пользователе по фотографии как

облачный сервис, использующий методы машинного обучения и требующий высокие мощности для обработки транзакций. Анализ данной диаграммы позволяет определить взаимосвязи между сервисами.

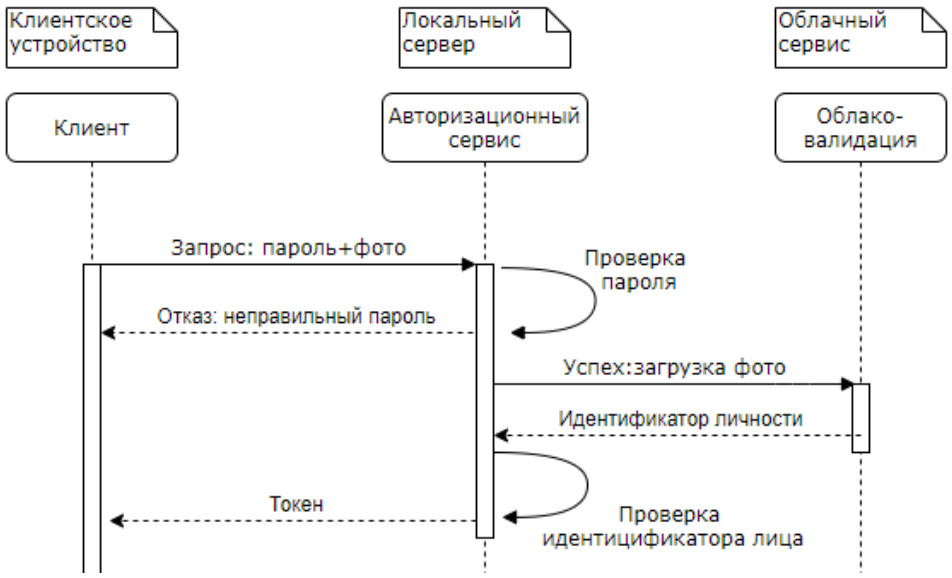


Рисунок 3.2 – UML диаграмма сценария авторизации.

Так с клиентского устройства поступает запрос (1) на авторизацию с использованием пароля и идентификатора лица. Мы рассматриваем сервис, который обрабатывает информацию о пользователе по фотографии как облачный сервис, использующий методы машинного обучения и требующий высокие мощности для обработки транзакции. Для сокращения затрат на обслуживания данного типа запросов, разработан сценарий работы, подразумевающий обработку, при которой сначала запрос поступает на сервер авторизации по паролю (2), чтобы в случае неудачной идентификации при обращении к локальному серверу не перенаправлять запрос в облако, так как мы рассчитываем, что при этом появляются видимые временные сетевые задержки. На данном этапе происходит разбиение сценария на успешную авторизацию по паролю (3.1) и отказ (3.2), что позволяет утверждать об изменении значения потока, приходящего на 3 сервис. В случае шага 3.1 производится обработка полученного изображения (4) и проверка

идентификатора личности (5), после чего передача токена обратно на клиентское устройство (6).

Данная диаграмма позволяет изучить структуру и характеристики системы, тем самым определяя функциональную модель. Математический анализ и имитационная модель будет представлены в четвертой главе данной работы. В проанализированной системе четко выражен рабочий процесс, что указывает на возможность его анализа в транзакционных системах.

Рабочие процессы обладают рядом следующих свойств:

– Автоматизация: Рабочие процессы в транзакционных системах должны быть автоматизированы, чтобы снизить процент ошибок, связанных с ручным вводом данных. Так же позволяет также снизить время выполнения процессов и повысить эффективность работы системы.

– Рабочие процессы должны быть *изолированы* друг от друга. Это обеспечивает более высокую степень надежности системы, так как возможные ошибки или сбои в одном процессе не будут влиять на другие процессы.

– Рабочие процессы должны быть *масштабируемыми*, чтобы поддерживать возрастающее количество транзакций и пользователей. Масштабируемость также позволяет расширять функциональность системы в соответствии с потребностями бизнеса.

– Рабочие процессы должны быть *гибкими,* чтобы адаптироваться к изменяющимся потребностям бизнеса. Гибкость позволяет быстро изменять и настраивать процессы, а также добавлять новые функции и интегрировать их в систему.

Вопрос математического моделирования рабочих процессов сопровождал развитие информационных систем, что видно в работах [116-117]. В работе [7] была разработана формулировка полумарковского процесса для составных веб-сервисов с точки зрения их производительности, надежности и узких мест. Данное исследование касается использования

нескольких веб-сервисов для создания новой услуги в рамках бизнес-процесса. Так же к расчету качества обслуживания для составных сервисов со сложной структурой обращался автор в работе [118].

3.1.4 Понятие эквивалентности в контексте рабочих процессов

В рабочих процессах описанных ранее эквивалентные микросервисы могут использоваться для повышения отказоустойчивости системы, а также для обеспечения более эффективной обработки задач. Если один микросервис выходит из строя, другой эквивалентный микросервис может продолжить работу вместо него. Использование эквивалентных микросервисов может увеличить масштабируемость системы, так как разные микросервисы могут быть запущены на разных платформах или в разных облаках.

На рисунке 3.3 представлена обобщённая схема реализации эквивалентности. На котором $S_{\text{старт}}$ – начальный узел, $S_{\text{конец}}$ - конечный узел и N – степень репликации задач на начальный узел.

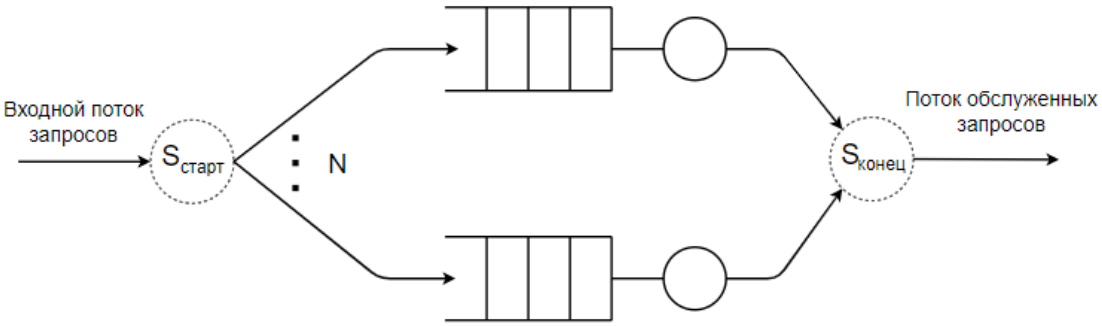


Рисунок 3.3 Обобщенная схема эквивалентности

Предложены следующие виды рабочих процессов, использующих эквивалентность:

Отказоустойчивость – система запускает одну ветвь процесса за раз, вплоть до момента, когда одна из N ветвей не получит успешное обслуживание. На рисунке показан сценарий отказоустойчивости, при котором если ветвь не получила успешного обслуживания в СМО S1, система

запускает процесс обслуживания дополнительной ветви в СМО S2, которая как показано на рисунке как закончиться успехом, так и нет. Если же при реализации всех N попыток обслуживания требование так и не было выполнено, оно считается не обслуженным. Математическая реализация данного процесса схожа с выражениями полученными в [119-120]. Однако в первом случае полученные при вычислении значения подразумевают что ответ от системы будет получен даже в случае отказа и равен будет времени полного успешного обслуживания. Во втором случае так же не учитывается время отказа узла, а выражение расчета времени обслуживания учитывает лишь два узла в рассматриваемой системе.

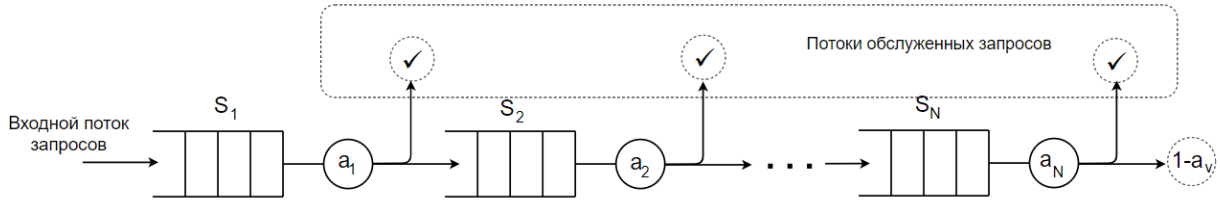


Рисунок 3.4 – Сценарий отказоустойчивой эквивалентности

В наше работе были приведены выражения для среднего времени успешного обслуживания в системе без учета времени отказа для N эквивалентных ветвей что снижает корректность расчетов, ведь время, затраченное на ошибочную обработку, может отличаться от времени успешной обработки.

Формула для расчета времени ответа в отказоустойчивой системе ,

$$T = \left(\sum_{i=1}^{N} a_i \times \sum_{k=1}^{i} t_k\right) \times \prod_{v=1}^{i-1}(1 - a_v) \Big/ 1 - \prod_{i=1}^{N}(1 - a_i) \qquad (3.1)$$

где N-количество активных узлов, a_i – надежность узла i (i=1…N), t_i – время обслуживания запроса в узле i.

Неудачное обслуживание может занимать как время меньшее времени обработки так как обслуживание не было оконченным, так и включать в себя время уведомления системы об ошибке в реализации транзакции.

Следовательно, учет времени ошибочной обработки транзакции необходим для более точного определения времени расчета сценария обработки требования более чем двух эквивалентных ветвях:

$$T = (\sum_{i=1}^{N} a_i \times \prod_{v=1}^{i-1}(1 - a_v) \times (t_i \times \sum_{k=1}^{i-1} t_{\text{отк}k}) \Big/ 1 - \prod_{i=1}^{N}(1 - a_i) \quad (3.2)$$

где $t_{\text{отк}k}$ – время, затраченное на уведомление об отказе в узле k (k=1…N)

Надежность все сети систем рассчитывается из выражения:

$$a_v = 1 - \prod_{i=1}^{N}(1 - a_i) \quad (3.3)$$

Спекулятивная параллельность – система включает одновременно в процесс все k ветвей, и результатом выполнения будет являться первый полученный результат успешного обслуживания. На рисунке изображен спекулятивно параллельный сценарий реализации обслуживания транзакции. Как только успешное выполнение процесса закончила одна из ветвей $S_{1.i}$ (где i номер узла в ветви процесса), остальные завершают выполнения обслуживания этой транзакции $S_{j.i}$ (где j количество оставшихся процессов и равно N-1).

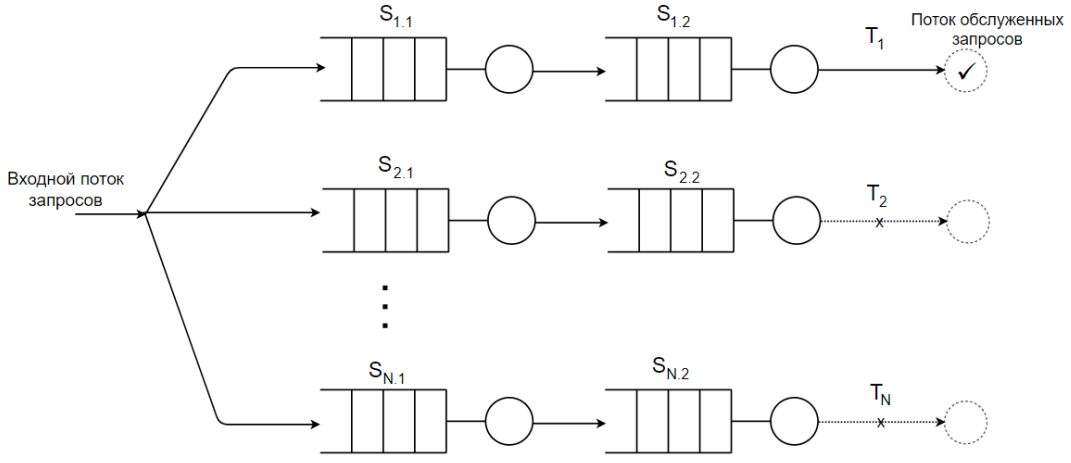


Рисунок 3.5 – Сценарий спекулятивного параллелизма в эквивалентности

Если хотя бы одна ветвь сценария успешно обслужила транзакцию, то временем обслуживания транзакции является время самого быстрого ответа системы. Если же реализация всех ветвей завершается неудачей, то временем ответа системы будет время ожидания ответа самой долгого рабочего процесса [92].

$$T = \begin{cases} min\{T_{\text{обс.}}\}: |T_{\text{обс.}}| > 0 \\ max\{T_{\text{неуд.}}\}: |T_{\text{обс.}}| = 0 \end{cases}; T = \{t_1, \dots, t_N\} \quad (3.4)$$

где $T_{\text{обс.}}$ – время успешного обслуживания транзакции в каждой запущенной ветви, $T_{\text{неуд.}}$ – время неудачного обслуживания транзакции.

Для расчета надежности системы, реализующей рассматриваемый сценарий необходимо использовать выражение 3.3.

Включение результата расчета наибольшего времени неудачного обслуживания необходимо для получения времени ответа системе о том, что все ветви сценария закончили свою работу, для принятия дальнейших мер.

Параллельность - системой запускается одновременно k ветвей исполнения сценария. Условием успешной реализации поставленной задачи является полное завершение всех k ветвей. Реализация данного процесса идентична работе параллельных ветвей Fork-Join из чего следует возможность применения данного математического аппарата для получения выходных характеристик системы.

3.2 Параллельная обработка запросов через механизм Fork-Join

Параллельность - включает в себя репликацию запроса на N узлов, что может быть использовано для двух целей. В первом случае, это может использоваться для моделирования процесса одновременного опроса двух или более слабосвязанных услуг с целью получения ответа от всех узлов. Если обслуживание хотя бы одной из услуг завершается провалом, то обслуживание оставшихся услуг также прекращается и запрос, поступивший на начальный узел, считается необслуженным.

Второй вариант использования параллельности заключается в необходимости разбиения задачи на более мелкие блоки данных и их одновременном обслуживании на параллельных узлах [121]. Следует отметить, что это требует изменения времени обслуживания транзакции:

$$t^{*}_{\text{обсл.}} = \frac{t_{\text{обсл.}}}{N} \tag{3.5}$$

Результатом реализации данного подхода будет зависимость (3.6) где для времени удачного обслуживания является максимальное время обработки реплики, а время, заложенное на ответ о неудачном обслуживании, является временем первой необслуженной реплики:

$$T = \begin{cases} max\{T_{\text{обс.}}\}: |T_{\text{обс.}}| > 0 \\ min\{T_{\text{неуд.}}\}: |T_{\text{обс.}}| = 0 \end{cases}; \; T = \{t_1, \ldots, t_N\} \tag{3.6}$$

где $T_{\text{обс.}}$ – время успешного обслуживания транзакции в каждой запущенной ветви, $T_{\text{неуд.}}$ – время неудачного обслуживания транзакции

Математический аппарат Fork-Join сохраняет выделенное свойство атомарности при обслуживании транзакций, что позволяет использовать его для расчета качественных характеристик элемента рабочего процесса в выделенной задачи расчёта. Применение Fork-Join в моделировании и проектировании телекоммуникационных систем имеет ряд преимуществ:

— Параллельное обслуживание позволяет эффективно использовать ресурсы системы, обеспечивая одновременное обслуживание нескольких запросов или задач.

— Увеличение пропускной способности позволяет распараллеливать обработку запросов, что может привести к увеличению пропускной способности системы и сокращению времени ответа.

— Использование Fork-Join структур дает возможность динамически адаптироваться к изменяющимся условиям и нагрузке, что повышает гибкость телекоммуникационной услуги.

— Моделирование сложных систем: Fork-Join структуры могут быть использованы для моделирования сложных систем, таких как распределенные системы хранения данных (РСХД), где параллельное обслуживание играет важную роль в обеспечении высокой доступности и масштабируемости.

Задача расчета характеристик процесса Fork-Join – является нетривиальной, однако имеет высокую актуальность и востребованность при моделировании систем с микросервисной архитектурой. В рамках данной работы необходимо провести анализ математического аппарата расчета систем FJ и его временных характеристик.

Моделирование Fork-Join структур является сложным процессом, так как требует учета параллельного обслуживания нескольких запросов и последующего их объединения. Для этого часто используются методы теории вероятностей, стохастического моделирования и численного анализа, чтобы оценить производительность и характеристики таких систем.

Ознакомиться с принципами внедрения данного процесса возможно в следующих работах [122-123]. Рассматриваемая структура, состоящая из *M* параллельных подсистем обслуживания, работает по следующему принципу: поступающее требование приходит на абстрактный блок Fork «*F*» в котором мгновенно делиться на N идентичных фрагментов, которые затем поступают на обслуживание в каждую из имеющихся подсистем. Далее реплики поступают по мере обслуживания в подсистемах в синхронизирующую очередь перед прибором объединения, и дожидаются окончания обслуживания всех реплик исходного запроса и объединяются в абстрактном блоке Join «*J*», после чего покидают СеМО в составе единого запроса.

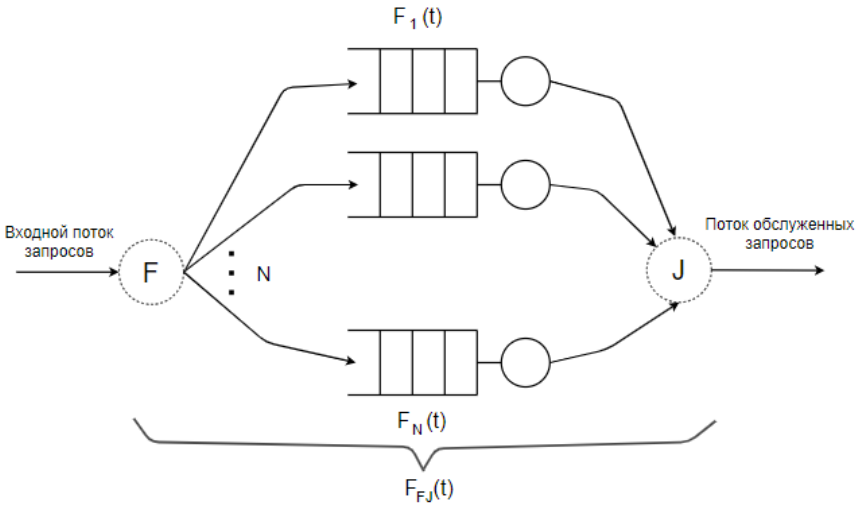


Рисунок 3.6 – Схема параллельной обработки запросов.

Математический аппарат Fork-Join сохраняет свойство атомарности при обслуживании транзакций: если все реплики поступили в блок слияния – обслуживание транзакции считается успешным, если же хоть одна реплика не была обслужена – реализация транзакции считается неудачной. Каждая подсистема $(1 \ldots N)$ рассматриваемой сети является обособленной системой массового обслуживания с функцией распределения времени нахождения в каждой системе $F_e(t)$.

Для произвольных параметров входящего потока и длительности обслуживания G/G/1 в работе [124] описан подход, основанный на анализе сети обслуживания в условиях слабой и сильной нагрузки. На основе предельных значений нагрузки выполняется интерполяция математического ожидания длительности пребывания требований в сети обслуживания.

Для случая экспоненциальной сети обслуживания G/M/1 выражение для TM имеет вид

$$T_M \approx [H_M + (\left(\sum_{i=1}^{M} \binom{M}{i} (-1)^{i-1} \sum_{m=1}^{i} \binom{i}{m} \frac{(m-1)!}{i^{m+1}}) - H_M\right) \frac{\lambda}{\mu}] \frac{1}{\mu - \lambda} \qquad (3.7)$$

Обратный же случай, описанный в работе [125] предполагает не экспоненциальное распределение длительности обслуживания в системе

M/G/1. Авторы работы [126] предлагают подход к моделированию и анализу систем с параллелизмом используя сети Петри.

Самый распространённый случай – сеть систем массового обслуживания вида M/M/1. Для рассматриваемой сети, состоящей из двух систем, было найдено точное значение для T_2:

$$T_2 = \frac{12 - \rho}{8} T_1 \tag{3.8}$$

где $T1 = 1/(\mu - \lambda)$, $\rho = \lambda/\mu$.

Предложенный метод [127] предполагает существование нижней и верхней границ для величины T_M. Верхняя граница получена в предположении о независимости длительностей пребывания фрагментов в системах обслуживания, тогда как нижняя граница получена при допущении отсутствия очередей фрагментов в системах обслуживания. Тогда в силу того, что длительность пребывания требования в сети обслуживания определяется как максимальное значение из длительностей пребывания его фрагментов в системах, получим

$$H_N \frac{1}{\mu} \leq T_N \leq H_N T_1 \tag{3.9}$$

где H_N обозначает N-ую частичную сумму гармонического ряда $H_N = \sum_{i=1}^{N} 1/i$. Учитывая одинаковую скорость роста $O(\ln N)$ этих границ, получено следующее приближённое выражение для T_N:

$$T_N \approx \left(\frac{H_N}{H_2} + \frac{4}{11}\left(1 - \frac{H_N}{H_2}\right)\rho\right) T_2, \qquad N \geq 2 \tag{3.10}$$

Приближенная функция распределения для всей сети с коэффициентом репликации равным m, рассчитывается выражением [127]:

$$F_{FJ}(t) \approx 1 - \sum_{i=1}^{N} c_i \left[1 - F_{Si}(t)\right] \prod_{j=1}^{i-1} F_{Si}(t) \approx \sum_{i=1}^{N} (c_i - c_{i+1}) \prod_{j=1}^{i} F_{Si}(t) \tag{3.11}$$

где c_i – коэффициент репликации, равный $c_{m+1} = 0, c_1 = 1, c_2 = 1 - p_1/4$ и

$$c_i \approx (1 - \frac{p_{i-1}}{8})\,(1 - \frac{p_{i-2}}{8}), \quad 3 \le i \le N \tag{3.12}$$

p_i – вероятность того, что требование будет ожидать обслуживания

С учетом того, что все узлы в блоке параллельного обслуживания упорядочены таким образом что $n_1\mu_1 \le \ldots \le n_N\mu_N$, в случае узлов с одним обслуживающим устройством не применяется иных ограничений и так же действует соглашение $\mu_1 \le \ldots \le \mu_N$.

Для учета блока параллельной обработки в расчете ВВХ прохождения всей сети локальных транзакций систем M/M/1, определённой в уравнении (3.16), были определены выражения для изображения ФРВ и ПРВ параллельной обработки двумя слабосвязанными услугами:

$$F_{FJ2}(s) = \frac{(\lambda - s - \mu)\left(-\lambda + 4\mu + \frac{4\mu(-2\lambda + s + 2\mu)}{s}\right)}{4\mu * (\lambda\ -\ s - \mu)(2\mu + s - 2\lambda)} + \frac{(-2\lambda + s + 2\mu)(8\mu - \lambda)}{4\mu * (\lambda\ -\ s - \mu)(2\mu + s - 2\lambda)} \tag{3.13}$$

$$f_{FJ2}(s) = \frac{8\mu^3 - 16\lambda\mu^2 + (\lambda s + 8\lambda^2)\mu - \lambda^2 s}{4\mu * (\mu + s - \lambda)(2\mu + s - 2\lambda)} \tag{3.14}$$

Fork-Join встречается в различных типах МСА, и не ограничивается только системами M/M/1. Такие структуры можно встретить в более общих моделях, таких как M/G/1, G/G/1 и других. Расчет моделей с коэффициентом репликации N>2 или с внедрением неэкпоненциальных распределений, рационально производить с использованием аппарата имитационного моделирования.

В рамках данной работы была разработана имитационная модель сценария Fork-Join с N-м количеством узлов и гибкой настройкой функционала [128].

3.3 Моделирование рабочего процесса математическим аппаратом сетей Джексона

3.3.1 Анализ математического аппарата сетей массового обслуживания

Основной задачей, поставленной в данной работе, является создание семейства моделей расчета ВВХ транзакционных услуг возможных рабочих процессов систем включая параллельную обработку запросов, описанную выше.

На рисунке 3.7 представлена функциональная модель рабочего процесса обслуживания цепочки последовательных локальных транзакций [129]. Сеть представляет совокупность эквивалентных обслуживающих узлов, для которых необходимо провести последовательный анализ перехода требований от узла к узлу для получения качественных временных характеристик.

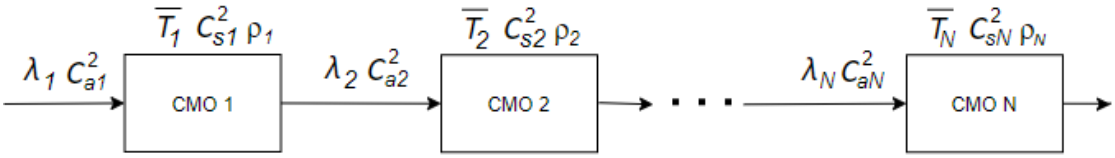


Рисунок 3.7 - Модель рабочего процесса обслуживания последовательных локальных транзакций

Общая средняя задержка T_s от начала до конца обслуживания в выделенной цепочке систем массового обслуживания получается суммированием отдельных задержек $\overline{T_i}$ на каждом из них:

$$\overline{T_{\text{СеМО}}} = \sum_{i=1}^{N} \overline{T_i} \qquad i = 1 \dots N \tag{3.15}$$

Так же последовательный анализ цепочки узлов возможно упростить, если использовать свойство сцепления как показано на рисунке 3.8. То есть преобразовать последовательное обслуживание требований цепочкой СМО в один эквивалентный узел с ФРВ времени прохождения $F_x(t)$.

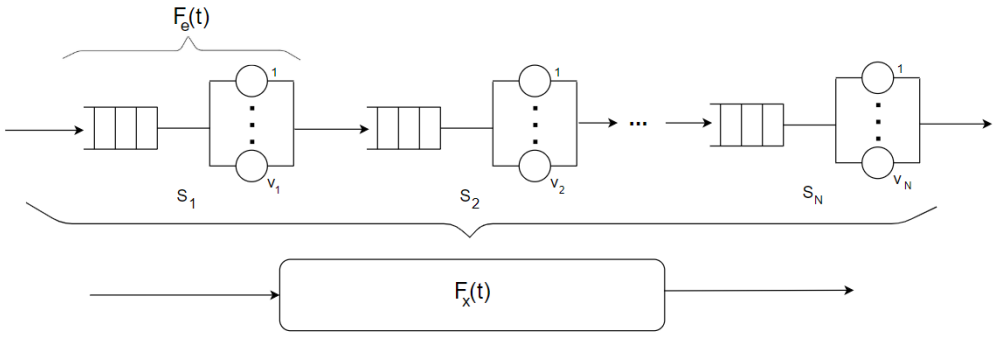


Рисунок 3.8 – Модель цепочки локальных транзакций в рабочем процессе системы.

Для нахождения математического ожидания и дисперсии временных характеристик для всей сети: нам необходимо провести расчет плотности распределения времени нахождения требования в рассматриваемой СеМО. Согласно [130] функция распределения времени прохождения всей цепочки микросервисов определяется сверткой ФРВ времени нахождения в каждой СМО:

$$F_x(t) = F_{e1}(t) * F_{e2}(t) * F_{e3}(t) * \cdots * F_{eN}(t) \qquad (3.16)$$

где $F_{ei}(t)$ - функция распределения временя окончания обслуживания в i-ой СМО, i=1…N

Для расчета ФРВ времени окончания обслуживания необходимо так же использовать процедуру свертки ФРВ времени ожидания и времени обслуживания:

$$F_e(t) = F_w(t) * F_s(t) \qquad (3.17)$$

где $F_w(t)$ - ФРВ времени ожидания обслуживания, $F_s(t)$- ФРВ времени обслуживания.

Для проведения процедуры свертки необходимо получить изображения для первоначальных функций и плотностей [131], что может быть реализовано с помощью преобразования Лапласа–Стилтьеса функции F(t), определяется следующим образом:

$$f(s) = \int_0^{\infty} e^{-st}\, dF(t) \qquad (3.18)$$

где s – комплексный параметр, s≥0

Так для систем M/M/1 определены ФРВ времени ожидания:

$$F_w(t) = 1 - \rho e^{-(\mu-\lambda)t} \qquad (3.19)$$

ФРВ времени нахождения в системе:

$$F_e(t) = 1 - e^{-(\mu-\lambda)t} \qquad (3.20)$$

В результате расчета функций распределения времени прохождения всей цепочки систем массового обслуживания (3.16), возможен расчет математического ожидания и дисперсии времени прохождения запроса СеМО или выделенного рабочего процесса следующими выражениями:

- первый начальный момент или математическое ожидание времени пребывания в выделенной сети:

$$\mathrm{M}[T] = \int_0^{+\infty} (1 - F_x(t))\, dt \qquad (3.21)$$

- дисперсия времени пребывания в выделенной сети:

$$D[T] = \mathrm{M}[(T)^2] - (\mathrm{M}[T])^2 \qquad (3.22)$$

Для анализа различных сценариев рабочих процессов необходимо определение основных как узловых характеристик, рассмотренных в первой главе, так и характеристик сетей систем массового обслуживания с различными архитектурными решениями.

Расчет сети систем массового обслуживания с ветвлением сценариев (рисунок 3.9) прохождения системы подразумевает использование принципа свертки ФРВ прохождения всех СМО в сценарии. Однако в данном случае на ФРВ прохождения системы $F_x(t)$, влияет разбиение потока запросов в сети при поступлении на параллельные узлы, изменяя интенсивность входного потока в каждый узел. Из [35] интенсивность входного потока λ_j в узел j из

предшествующего узла i определяется матрицей переходов $R = \|p_{ij}\|$ и равняется $\lambda_i \times R$.

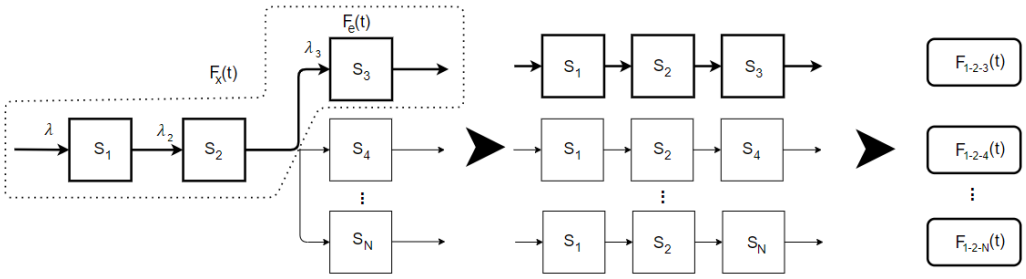


Рисунок 3.9 – Модель ветвления сценариев рабочих процессов системы.

Уравнение Лапласа-Стилтьеса для функции компонентного процесса j задается выражением:

$$f_{ij}(s) = \frac{p_{ij} f_i(s)}{1 - (1 - p_{ij}) f_i(s)} \tag{3.33}$$

где $f_i(s)$ - преобразование ЛС функции F(t).

Так же необходимо рассмотреть модель сценария использования одного элемента системы для обслуживания потоков запросов нескольких рабочих процессов.

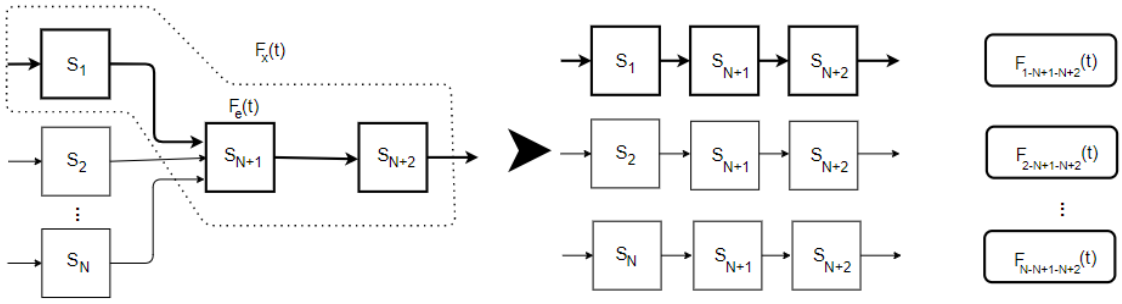


Рисунок 3.10 – Модель сценария наложения ресурсов рабочих процессов

Функция F(t) результирующего процесса двух потоков равна [132]:

$$F(t) = 1 - \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \times (F_1^C(t) \times \int_t^\infty F_2^C(u)du + F_2^C(t) \times \int_t^\infty F_1^C(u)du) \tag{3.34}$$

где $F_j^C(t) = 1\text{-}F_j(t)$ – дополнительная ФРВ, $F_j(t), j = 1,2$ – функция распределения интервалов времени между событиями в потоке j.

В случае N>2 переход от двухкомпонентных процессов к общему случаю (N+1) выполняется рекурсивно за N шагов в соответствии с рисунком 3.11.

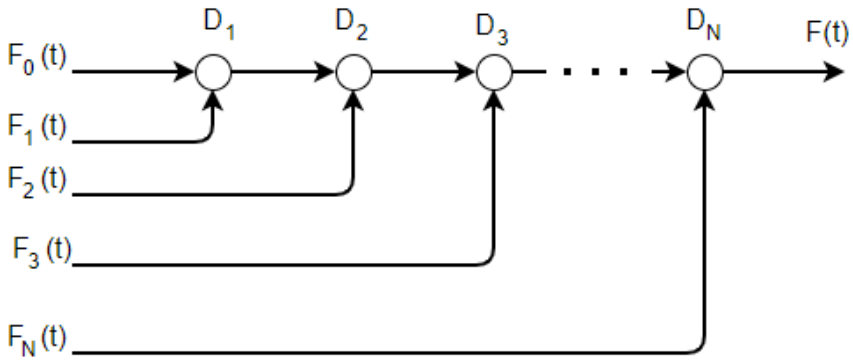


Рисунок 3.11 – Состав (N+1) компонентных процессов

Это позволяет получить характеристики основных процессов изменения потока в сетях систем массового обслуживания.

3.3.2 Параметры выходного потока для последовательной СеМО

Последовательное прохождение цепочки систем массового обслуживания подразумевает изменение выходного потока из каждой СМО в цепочке. Обозначим d(t) как плотность распределения вероятностей промежутков между последовательными требованиями на выходе узла *i*. D(s) – преобразование Лапласа для заданной плотности.

Расчет данного параметра проводится следующим выражением:

$$D(s) = (1-\rho)D(s)|_{\text{узел i свободен}} + \rho D(s)|_{\text{узел i занят}} \qquad (3.35)$$

- другой вид выражения данного выражения:

$$D(s) = (1-\rho)I`(s)B(s) + \rho B(s) \qquad (3.36)$$

где B(s) – преобразование ЛС ФРВ времени обслуживания, I`(s)- преобразование ЛС ФРВ времени до поступления следующего требования в систему.

Для систем M/M/1 в [35] предложено следующее решение:

Согласно теореме Бёрке исходящий поток стационарной СМО типа M/M/v с пуассоновским входящим потоком с параметром λ и показательным распределением времени обслуживания с параметром μ в каждом из v приборов является пуассоновским потоком с тем же самым параметром λ. Это еще раз доказывает предположение о том, что цепочка из нескольких узлов, состоящих из систем M/M/v сохраняет свойство разложения на отдельные узлы, которая позволяет рассматривать узлы отдельно друг от друга.

Из вышесказанного для систем M/M/v можно вывести следующее:

$$\mathrm{D(s)} = A(s) = \frac{\lambda}{s+\lambda} \qquad (3.37)$$

Следовательно, распределение промежутком времени между уходящими требованиями:

$$\mathrm{D}(t) = 1 - \mathrm{e}^{-\lambda t}, \qquad t \geq 0 \qquad (3.38)$$

Это упрощает расчет характеристик сетей систем с экспоненциальными распределениями, и зачастую является основой методов аппроксимации [132][54] для некоторых типов систем с произвольными распределениями G/G/1. Вопрос расчета сетей систем с не экспоненциальными параметрами включает определение ФРВ до поступления следующего требования в системах общего вида G/G/1 и будет рассмотрен позже.

3.3.3 Пример расчета цепочки систем M/M/1 с параллельным обслуживанием

В качестве примера проведем расчет цепочки локальных транзакций с внедрением параллельной обработки запроса Fork-Join изображенной на рисунке 3.12.

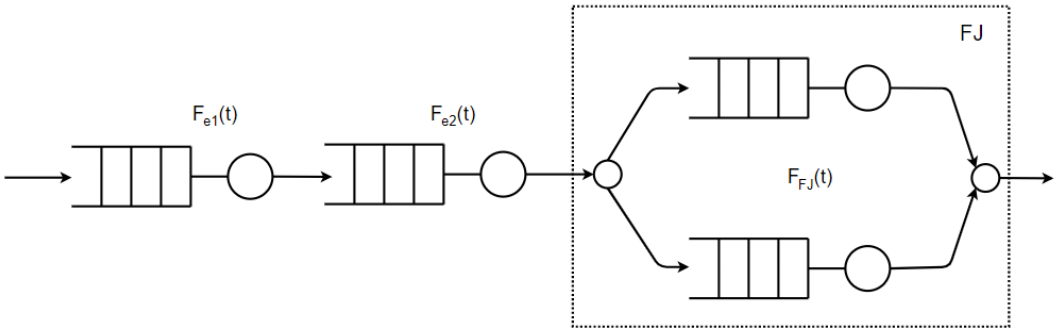


Рисунок 3.12 – Модель рабочего процесса с параллельной обработкой запроса

Для расчета функции распределения времени прохождения сети систем включая систему параллельной обработки запросов так же, как и в (3.16) используется уравнение свертки и выглядит следующим образом для примера, изображенного на рисунке:

$$F_x(t) = F_{e1}(t) * F_{e2}(t) * F_{FJ}(t) \tag{3.39}$$

На рисунке 3.13 изображен график ПРВ (а) и ФРВ (б) времени прохождения трех сценариев обслуживания транзакции: S (система Fork-Join), S1 (CMO_2 и система Fork-Join) и S2 (CMO_1, CMO_2 и система Fork-Join). Так же на рисунке 3.13 изображено математическое ожидания выделенных CeMO. Получение результатов моделирования необходимо для анализа влияния последовательных систем на выходные временные характеристики цепочки локальных транзакций. Плотность распределения для S2 задается выражением:

$$f_x(t) = \frac{1}{8\mu} \times (\mu - \lambda) * e^{t(\lambda-\mu)} \times (4\lambda^2 t + \lambda^3 * t^3 - 6\lambda^2 t^2 \mu - 15\lambda t^2 \mu^2 - 4\lambda e^{t(\lambda-\mu)} + 12\lambda t\mu - 16\mu e^{t(\lambda-\mu)} + 4\lambda + 8t^2\mu^3 - 16t\mu^2 + 16\mu \tag{3.40}$$

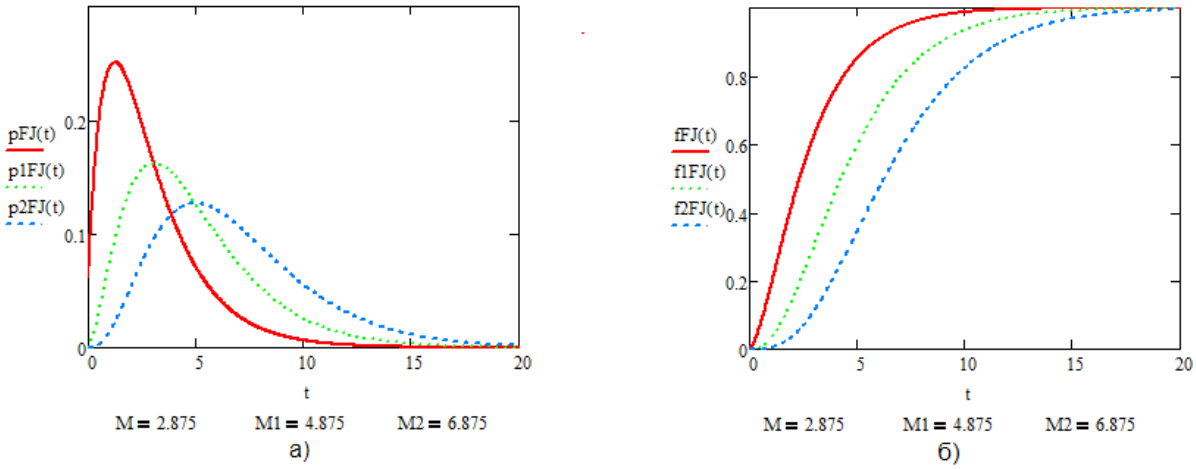


Рисунок 3.13 – Графики ПРВ и ФРВ времени обслуживания в трех выделенных сценариях

Так же были получены результаты иммитационного моделирования с аналогичными входными данными, с помощью программного обеспечения Anylogic.

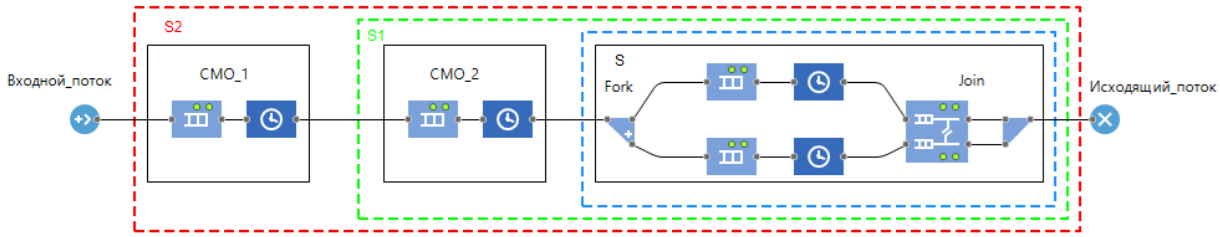


Рисунок 3.14 - Имитационная модель цепочки локальных транзакций с FJ

При сравнении данных результатов с результатами математического анализа, можно сделать вывод о рациональности использования предложенного математического аппарата.

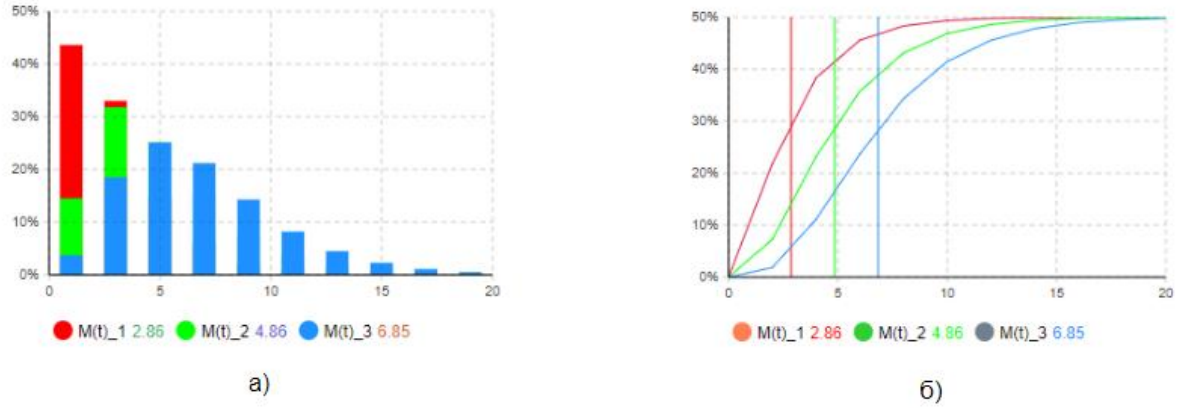


Рисунок 3.15 – Результаты иммитационного моделирования: плотность и функция распределения вероятностей времени обслуживания

Как видно из рисунка 3.16 при увеличении входной интенсивности на всю систему, увеличивается время обслуживания задачи, а точнее время ожидания обслуживания запроса в каждом элементе. Увеличение происходит равномерно до момента, пока коэффициент загрузки узлов $\rho_i \rightarrow 1$, так при входной интенсивности $\Lambda_{\text{крит.}}$=0,98 запрос/сек., в этот момент узлы стремятся к перегрузке и характеризуются недопустимым временем нахождения в системе.

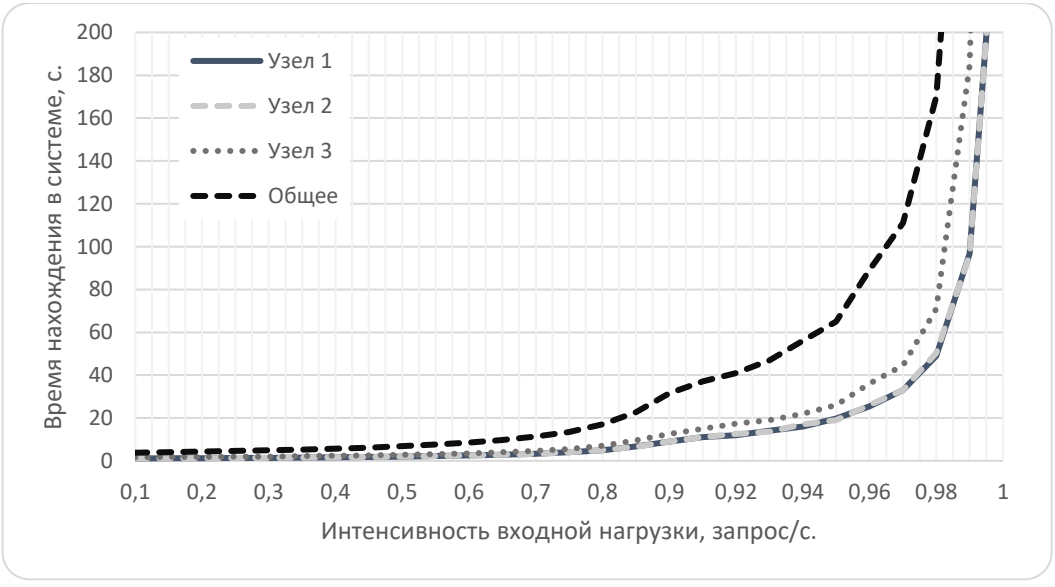


Рисунок 3.16 – Зависимость времени нахождения в системе от Λ

Дальнейшее моделирование предполагает при заданной критической входной нагрузке $\Lambda_{\text{крит.}}$, постепенно увеличивать интенсивность обслуживания на всех узлах систем, что как видно на рисунке 3.17 заметно снижает время нахождении запроса в узлах и СеМО в целом.

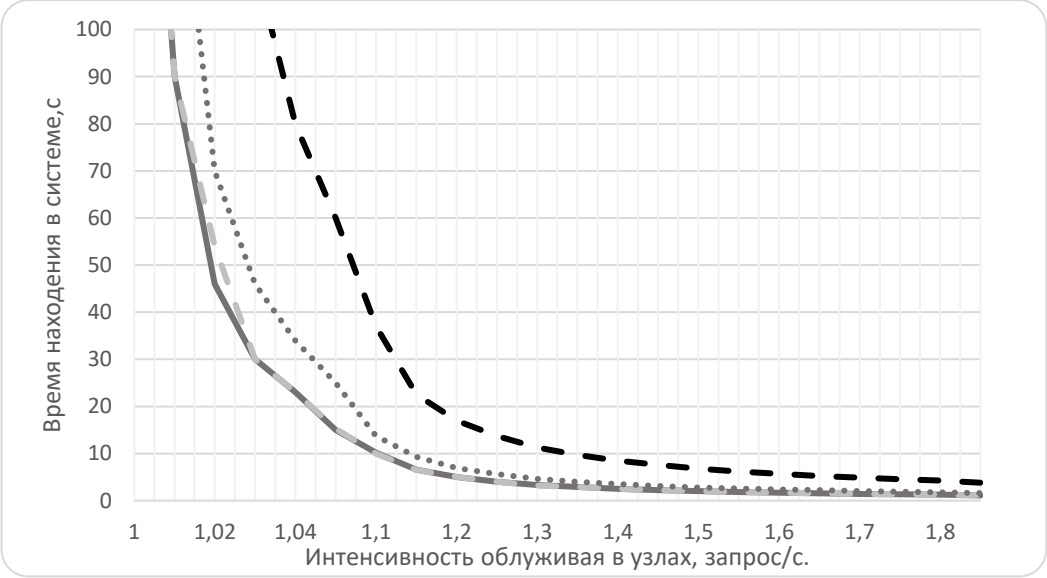


Рисунок 3.17 – Зависимость времени нахождения в системе от μ_i

Это позволяет предлагать вертикальное масштабирование (увеличение производительности) разрабатываемых транзакционных услуг как эффективный метод повышения качественных и количественных показателей услуг.

В целом по результатам проведенного анализа последовательной сети систем с элементов параллельного обслуживания Fork-Join, можно сделать вывод что математическое моделирование систем типа M/M/1 дает точные результаты, и является практичным решением для предварительных расчетов характеристик сетей и систем на этапе планирования.

3.3.4 Расчет параметров задержки систем произвольного вида

Ранее были представлены готовые решения для оценки времени ожидания обслуживания в системах M/M/1. Для систем G/G/1, которые представляют собой более общий тип обслуживания, был предложен метод расчета задержки транзакций в работе Клейнрока [35]. В этой работе было

получено интегральное уравнение Линди, которое применимо для всех значений аргумента в виде:

$$W(y) + W_{-}(y) = \begin{cases} \int_{-\infty}^{y} W(y-u)dC(u) & y \geq 0 \\ 0 & y < 0 \end{cases} \quad (3.41)$$

где через C(u) – обозначена функция распределения величины u.

Для решения данного уравнения было выбрано спектральное решение интегрального уравнения Линдли. Принимаем $\Phi_{-}(s)$ – преобразование Лапласа ФРВ $W_{-}(y)$, а $\Phi_{+}(s)$ – преобразование Лапласа ФРВ $W(y)$, которое связанно соотношением (3.42) с плотностью распределения времени ожидания W*(s).

$$s\Phi_{+}(s) = W^{*}(s) \quad (3.42)$$

Ссылаясь на выражения [35], с учетом того, что время ожидания является неотрицательной случной величиной, в правой части выражения должна находится свертка функции распределения времени ожидания и плотности случайной величины, следовательно

$$\Phi_{+}(s) + \Phi_{-}(s) = \Phi_{+}(s) * C^{*}(s) \quad (3.43)$$

где преобразование Лапласа-Стилтьеса функции C(u) находится выражением

$$C^{*}(s) = A^{*}(-s) \times B^{*}(s) \quad (3.44)$$

Так для аналитичной в области в области Re(s)<D функции $\Phi_{-}(s)$ получается равенство:

$$\Phi_{-}(s) = \Phi_{+}(s)[A * (-s)B * (s) - 1] \quad (3.45)$$

$$\Phi_{-}(s) = \Phi_{+}(s)\frac{\Psi_{+}(s)}{\Psi_{-}(s)} \quad (3.46)$$

Где функция $\Psi_{+}(s)$ аналитична по s при Re(s)>0 и не иметь нулей в правой полуплоскости, аналогично функция $\Psi_{-}(s)$ аналитична в плоскости Re(s)<D .

$$\Phi_+(s) = \frac{1}{\Psi_+(s)} \lim_{s\to 0} \frac{\Psi_+(s)}{s} = \frac{W(0^+)}{\Psi_+(s)} \qquad (3.47)$$

В работе [35] приведены примеры расчета систем M/M/1, G/M/1 и конкретно $E_2/M/1$. В качестве примера расчета в данной работе мы возьмем $H_2/H_2/1$, а в четвертой главе данной работы будет предложен алгоритм с методом расчета цепочки последовательных узлов с аппроксимирующей функцией распределения времени между поступлениями транзакций в систему.

Выражения для преобразования Лапласа-Стилтьеса ФРВ обслуживания и между поступлением запросов в систему с гиперэкспоненциальным распределением второго порядка (H_2):

$$\mathrm{A(s)} = \frac{\lambda_1 p}{\lambda_1 + s} + \frac{\lambda_2(1-p)}{\lambda_2 + s} \qquad (3.48)$$

где p – вероятность формирования случайной величины по первой экспоненте с интенсивностью λ_1, аналогично ПЛС для ФРВ времени обслуживания:

$$\mathrm{B(s)} = \frac{\mu_1 q}{\mu_1 + s} + \frac{\mu_2(1-q)}{\mu_2 + s} \qquad (3.49)$$

где q – вероятность формирования случайной величины по первой экспоненте с интенсивностью μ_1.

Используя выражения (3.42)-(3.49) на этапе разложения, получаем:

$$\frac{\Psi_+(s)}{\Psi_-(s)} = A*(-s)B*(s) - 1 = \left(\frac{\lambda_1 p}{\lambda_1 - s} + \frac{\lambda_2(1-p)}{\lambda_2 - s}\right)\left(\frac{\mu_1 q}{\mu_1 + s} + \frac{\mu_2(1-q)}{\mu_2 + s}\right) - 1 = \frac{s\,(s^3 - s^2 \times a - s \times b - c)}{(\lambda_2 - s)(\lambda_1 - s)(\mu_1 + s)(\mu_2 + s)} \qquad (3.50)$$

где коэффициенты уравнения *(3.50)*: $a = \lambda_2 + \lambda_1 - \mu_1 - \mu_2$, $b = -(p\lambda_1 + (1-p)\lambda_2) * (q\mu_1 + (1-p)\mu_2) - \lambda_1\lambda_2 - \mu_1\mu_2 + (\lambda_1 + \lambda_2)(\mu_1 + \mu_2)$, $c = \lambda_1\lambda_2 * (q\mu_1 + (1-p)\mu_2) - (p\lambda_1 + (1-p)\lambda_2)\mu_1\mu_2 - \lambda_1\lambda_2(\mu_1 + \mu_2) + \mu_1\mu_2(\lambda_1 + \lambda_2)$

Графическое представление результата из полученного выражения (3.50):

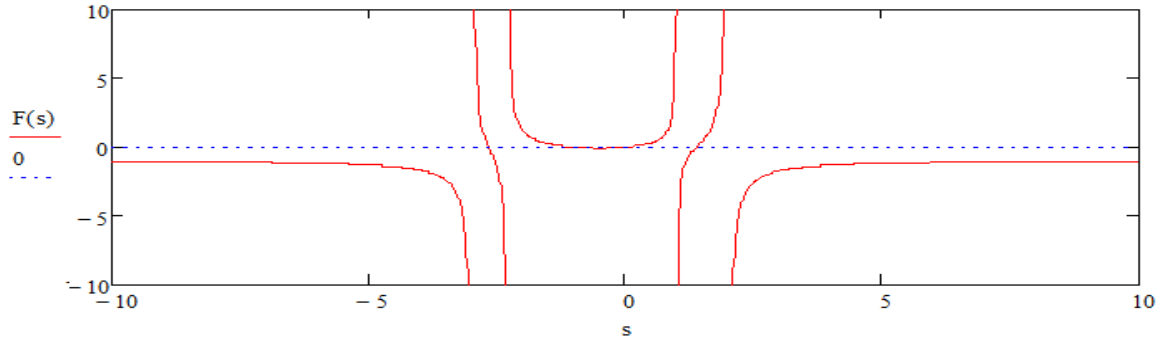


Рисунок 3.18 – График функции (3.50)

Исходя из выводов [35], спектральное разложение удобно решается при переносе его нулей и полюсов на комплексную плоскость как на рисунке 3.19.

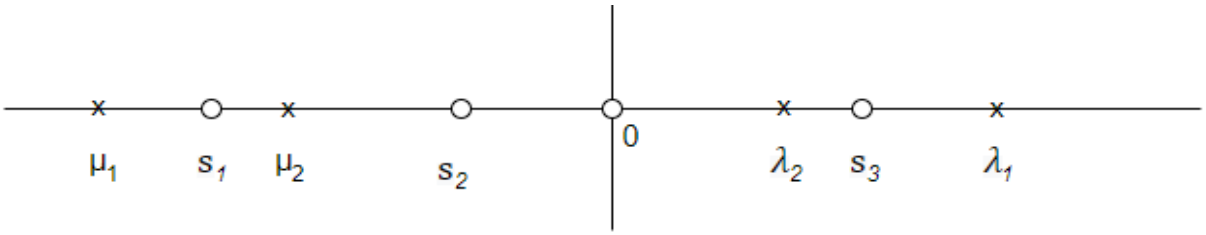


Рисунок 3.19 – Набор нулей и полюсов для системы $H_2/H_2/1$

где s_1, s_2, s_3 – корни уравнения если $\lambda_1 > \lambda_2$ и $\mu_1 > \mu_2$

Для решения уравнения (3.50) необходимо из рисунка 3.19 к фунции $\Psi_+(s)$ отнести нули и полюса лежащие в левой полулоскости, получив в результате выражение:

$$\Psi_+(s) = \frac{s(s+s_1)(s+s_2)}{(\mu_1+s)(\mu_2+s)} \qquad (3.51)$$

В случае систем типа G/M/1, дальнейшее решение данной задачи упрощенно реализовано в виде выражения:

$$W(t) = 1 - (1 - \frac{x}{\mu})\, e^{-xt}, \quad t \geq 0 \qquad (3.52)$$

где $x = \mu \times \lim_{s \to 0} \frac{\Psi_+(s)}{s}$

Приближение при большой нагрузке ρ⟶1 для G/G/1

$$A * (-s)B * (s) - 1 \approx s(s - s_0)\frac{\sigma_a^2 + \sigma_b^2}{2} \qquad (3.53)$$

где $s_0 \approx -\frac{2\bar{t}(1-\rho)}{\sigma_a^2+\sigma_b^2}$, σ_a^2 – дисперсия времени поступления запросов σ_b^2 - дисперсия времени обслуживания.

Используя выражение для s_0 и преобразование Лапласа, была получена аппроксимация распределения времени ожидания в области высоких значений времени ожидания при ρ ≈ 1:

$$W(t) \approx 1 - \exp\left(-\frac{2\bar{t}(1-\rho)}{\sigma_a^2 + \sigma_b^2}t\right) \qquad (3.54)$$

где σ_a^2 – дисперсия времени поступления запросов, σ_b^2 - дисперсия времени обслуживания, ρ – коэффициент загрузки.

Если взять во внимание что среднее время ожидания W определяется выражением $1/s_0$, таким образом:

$$\bar{W} = \frac{\sigma_a^2 + \sigma_b^2}{2\bar{t}(1-\rho)} \qquad (3.55)$$

В работе [133] представлена математическая модель для расчёта характеристик систем M/G/1, в которых информация о плотности вероятности времени обслуживания ограничена знанием нескольких моментов. Полученные результаты определяют теоретико-информационные приближения для G/G/1 в случае при котором известны только первые моменты для плотности вероятности времени обслуживания s(t) и плотности вероятности времени между поступлением запросов α(t), однако данные приближения являются точным результатом для M/M/1 что предполагает, что системы M/M/1 могут являться надлежащим теоретико-информационным приближением для систем G/G/1 при заданных первых моментах. Так плотность времени ожидания выражена следующим образом:

$$f_w(t) = (1-\rho)\delta(t) + \lambda(1-\rho)e^{-(1-\rho)\mu t} \qquad (3.56)$$

Для разработки общего алгоритма расчета ВВХ систем общего вида G/G/1, модели которых могут включаться в себя функции имеющие ограничения для процедур интегрирования, а также распределения, для которых функции вероятности времени не определены в явном виде (экспериментальные данные – ступенчатые функции), необходимо рассмотреть возможные подходы к аппроксимации.

3.4 Аппроксимация неэкспоненциальных распределений с высоким коэффициентом вариации

3.4.1 Необходимость аппроксимации и последствия ошибки

Любая рабочая нагрузка может быть представлена в виде комбинации функции плотности вероятности и определенных параметров для целей моделирования [134].

Важность выбора функций можно оценить, представив негативные последствия [135], с которыми можно столкнуться при неправильном выборе функции аппроксимации сложных распределений такие как:

– Выбранная аппроксимационная функция может смоделировать исходную с недостаточной точностью, что приведет к некорректным выводам и ошибочным прогнозам;

– Потеря важных параметров или свойств исходной функции распределения при выборе аппроксимационной функции, это может так же негативно сказаться на полноте информации и значимости аспектов модели.

– Высокая сложность функции аппроксимации, реализация которой потребует большого количества данных для оценки, может привести к недоступности аппроксимации моделей с ограниченным объемом исходных данных.

Чтобы избежать указанных последствий, необходимо тщательно анализировать исходную функцию распределения и выбирать аппроксимационные функции, которые наилучшим образом ее описывают. Кроме того, необходимо проверять результаты аппроксимации на соответствие реальным данным и корректировать модель при необходимости. Данные проблемы раскрыл автор работы [136], продемонстрировав негативные последствия на примере реальных данных, а в работе [137] описано негативное влияние на оценку максимального правдоподобия между двумя распределенными функциями. При этом авторы сходятся во мнении что проблема выбора точного распределения усложняется с уменьшением размера выборки первоначальных значений. Проблема оценки наблюдаемых процессов в аналитической форме для использования в сетевых вычислительных системах описана в [137].

Следующий рассматриваемый вопрос связан с оценкой аппроксимационных моделей. Для оценки данных моделей зачастую используются критерии точности аппроксимации и сложности модели. Хотя были предложены различные аналитическое модели [138] еще не многие из них могут обеспечить оба преимущества одновременно. Чтобы достичь высокого уровня аппроксимации часто требуется увеличить сложность модели. Это усложняет процесс оценки и модель становится трудной для реализации на практике. Далее будут представлены и описаны некоторые методы аппроксимации сложных распределений и приведен метод получения аппроксимирующей гиперэкспоненциальной функции для длиннохвостых распределений Вейбулла-Гнеденко.

Метод аппроксимации может быть использован для оценки H(t) когда замкнутое преобразование Лапласа ФРВ невозможно. Вместо аппроксимаций H(t) предлагается найти подходящую функцию $F_a(t)$, которая может аппроксимировать базовую функцию F(t). Требуется, чтобы $F_a(t)$ имела замкнутое преобразование Лапласа, и низкий коэффициент ошибки апроксимации.

Так же стоит отметить, что помимо аппроксимации уже существующих распределений, возможно внедрение понятия трансмутации некоторых существующих распределений для получения инструмента, более четко описывающего вероятностные события во всех сферах научной деятельности.

Так, например, в работе [139] приведена расширенная обобщенная форма логистического распределения, в [140] – введено понятие логарифмически-логистического распределения (TLLog).

В [141] предлагается решение для оценки нестационарных распределений с использованием среды имитационного моделирования Anylogic. Один из значимых результатов работы предполагает получение свойства потоков: если сумма средних значений интенсивности не превышает интенсивности обслуживая, то независимо от суммы максимальных значений интенсивностей система не перейдет в постоянно перегруженное состояние.

3.4.2 Эмпирическая функция распределения

Эмпирическая функция распределения (Empirical Distribution Function, EDF) - метод аппроксимации функции распределения путем построения ступенчатой функции на основе наблюдаемых данных.

Для построения эмпирической функции распределения, сначала упорядочиваются наблюдаемые значения по возрастанию. Затем для каждого значения вычисляется относительная частота, то есть доля наблюдений, которая меньше или равна данному значению. Наконец, строится ступенчатая функция (рисунок 3.20), где значение эмпирической функции распределения на каждом узле равно соответствующей относительной частоте. Между узлами функция остается постоянной, формируя ступеньки.

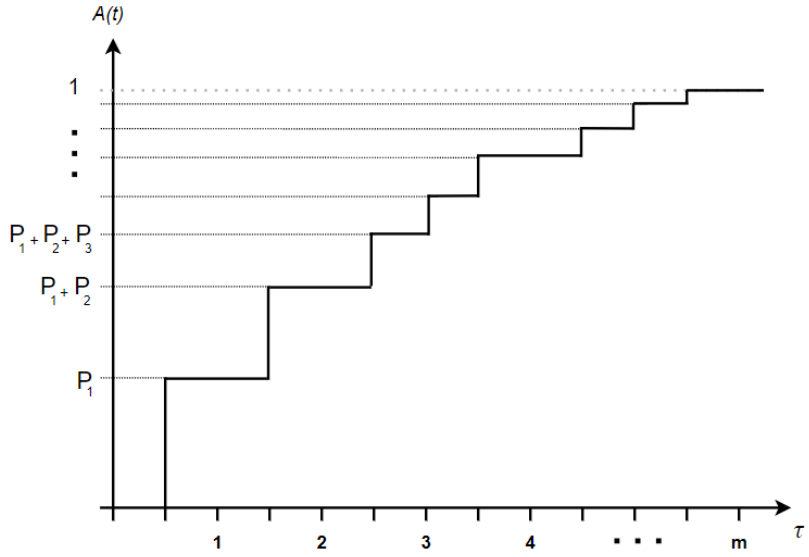


Рисунок 3.20 – Ступенчатая функция

Математическая модель для определения аппроксимационной функции, заключается в следующем: предположим, что информация о функции A(t) (или B(t)) основана на результатах измерений. Пусть с некоторым периодом квантования, равным τ, определяются соответствующие величины приращений исследуемой функции P_i. Если первое приращение функции A(t) зафиксировано в точке τ, то её преобразование Лапласа-Стилтьеса α(s) можно записать в таком виде:

$$\alpha(s) = \sum_{i=0}^{m} P_i e^{-i\tau s} \qquad (3.57)$$

Величина m определяет ту точку на оси «Время», в которой обнаружено последнее приращение функции A(t). Коэффициент вариации длительности интервалов между моментами поступлений соседних запросов C_A может меняться в широких пределах.

3.4.3 Аппроксимация длиннохвостых распределений

Длиннохвостые распределения могут оказывать значительное влияние на производительность, например, данный вид распределений времени обслуживания приводят к длиннохвостым распределениям времени ожидания

в очередях, и зачастую данный эффект сложно описать подробно, поскольку модели производительности с компонентными длиннохвостыми распределениям как правило трудно поддаются анализу.

Уже упомянутое классическое распределение Вейбулла-Гнеденко задается следующими выражениями для функции и (2.7) для плотности вероятности (2.8).

Существует множество подходов к аппроксимации одного распределения другим. Общепринятый способ заключается в приравнивании моментов. То есть чтобы аппроксимировать распределение Вейбулла нормальным распределением $N(\mu,\sigma^2)$, мы можем приравнять среднее значение (3.58) и дисперсию (3.59) нормального распределения к значениям распределения Вейбулла.

$$\mu = \Gamma(1/\beta + 1), \qquad (3.58)$$

$$\sigma^2 = \Gamma(2/\beta + 1) - (\Gamma(1/\beta + 1))^2 \qquad (3.59)$$

Ошибка аппроксимаций между распределениями Вейбулла и нормальным распределением с равными средними значениями и дисперсией рассчитывается как:

$$\varepsilon_{\text{ошиб.}} = F(x) - \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} exp\left\{-\frac{(u-\mu)^2}{2\sigma^2}\right\} du, \qquad x \geq 0 \qquad (3.60)$$

В работе [138] предложен метод использования двухпараметрического обобщенного экспоненциального распределения (GE) для ситуаций, когда требуется искаженное распределение для неотрицательной случайной величины.

– ФРВ двухпараметрического обобщенного экспоненциального распределения (GE):

$$F_{GE}(x;\alpha,\lambda) = (1 - e^{-\lambda x})^{\alpha}, \qquad \alpha,\lambda > 0 \qquad (3.61)$$

– ПРВ двухпараметрического обобщенного экспоненциального распределения (GE):

$$f_{GE}(x;\alpha,\lambda) = \alpha\lambda(1-e^{-\lambda x})^{\alpha-1}e^{-\lambda x}, \quad \alpha,\lambda > 0 \qquad (3.62)$$

Авторы уточняют что функции распределения (GE) и распределения Вейбулла являются обобщением экспоненциального распределения. И во многих ситуациях распределение (GE) обеспечивает лучшую подгонку чем распределение Вейбулла.

Для аппроксимации F(t) Вейбулла и Гамма распределений в [142] используется обобщенная экспоненциальная функция. Это связанно с тем что она математически точна для аппроксимации именно этих типов распределений и позволяет достаточно просто выполнять преобразование Лапласа. Для обобщенных экспоненциальных распределений

– ФРВ обобщенных экспоненциальных распределений

$$F_g(t) = (1-e^{-bt})^n \qquad (3.63)$$

– ПРВ обобщенных экспоненциальных распределений

$$f_g(t) = bne^{-bt}(1-e^{-bt})^{n-1} \qquad (3.64)$$

где b - параметр масштаба b>0, n – параметр формы n>0

Обобщённая экспоненциальная функция Fs(t) может быть использована для аппроксимации $F_w(t)$ следующим образом:

$$F_w(t) = (1-e^{-bt})^n + e(t) \qquad (3.65)$$

где e(t) – коэффициент, который включен в уравнение для представления ошибки аппроксимации. Учитывая конкретную функцию распределения Вейбулла с известными коэффициентами формы и масштаба, необходимо найти наилучшие значения b и n чтобы минимизировать значение e(t).

$$min\, w\{b,n\} = \frac{1}{m}\sum_{i=1}^{m}(F_w(t_i) - (1 - e^{-bt_i})^n)^2 \qquad (3.66)$$

где b>0 и n положительное целое число

Данное выражение позволяет минимизировать среднеквадратичную ошибку между $F_w(t)$ и $F_s(t)$. Где t – момент времени выбранный из временного горизонта для i=1…m. Реализация данного решения включает в себя только две переменные решения b и n. Для поиска наилучшего решения могут быть использованы итерационные алгоритмы или градиентные методы.

Стоит отметить, что в данном методу условие возможности аппроксимации для распределения Вейбулла является значение коэффициента формы b>1.

3.4.4 Подход к аппроксимации гиперэкспоненциальным распределением

Авторы работ [143-144] показывают результаты аппроксимации распределений суммой экспонент и доказывают возможность сколь угодно близкой аппроксимации распределений из большого класса длиннохвостых распределений включая распределение Вейбула-Гнеденко или Парето. Условием аппроксимации данным способом длиннохвостых распределений является их монотонность, условие которой для распределений Парето и Вейбула-Гнеденко является параметр формы b<1.

Характеристики гиперэкспоненциального распределения:

– плотность распределения, заданной выражением:

$$p(t) = \sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i t}, \qquad t \geq 0 \qquad (3.67)$$

– дополнительная кумулятивная функция распределения:

$$F^c(t) = \sum_{i=1}^{k} p_i e^{-\lambda_i t}, \qquad t \geq 0 \qquad (3.68)$$

– изображение плотности гиперэкспоненциального распределения:

$$p(s) = \int_0^\infty e^{st} p(t)dt = \int_0^\infty e^{st} dtF(t) = \sum_{i=1}^{k} \frac{p_i \lambda_i}{\lambda_i + s} \qquad (3.69)$$

Стоит отметить, что использование данного метода аппроксимации длиннохвостых распределений гиперэкспоненциальным распределением является рациональным решением, именно благодаря простоте переходов Лапласа для такого рода распределений.

Алгоритм подгонки гиперэкспоненциального распределения (H_k) к заданному длиннохвостому распределению состоит в следующем:

Принятие экспоненциальных параметров необходимых для выражений (3.67) - (3.69) такие чтобы $\lambda_1 < \lambda_2 < \ldots < \lambda_k$. При этом компоненты с более высоким индексом имеют хвосты, которые распадаются быстрее.

Рекурсивная подгонка компонент H_k начиная с пары (p_1, λ_1), переходя к (p_2, λ_2), и далее к более высоким индексам. Если λ_2 весомее чем λ_1 то $\sum_{i=2}^{k} e^{-\lambda_k * t}$ должно быть значительней чем $p_k e^{-\lambda_k * t}$ при времени достаточно большом в хвосте. Это позволяет не брать в расчет другие параметры H_k при выборе первой пары значений (p_1, λ_1).

В качестве примера в данной работе было выбрано распределение Вейбулла с параметрами a=0,5 и c=0,69 и полученное гиперэкспоненциальное распределение с параметрами, представленными в таблице

Таблица 3.1 – Значения параметров для аппроксимационной формулы

i	**1**	**2**	**3**	**4**	**5**	**6**
p_i	0,2	0,13	0,202	0,1559	0,104	0,208
λ_i	0,8	0,92	1,829	2,04	2,38	27,3418

На рисунке 3.21 показан результат для ФРВ (а) и ПРВ (б) при аппроксимации распределения Вейбула-Гнеденко, гипперэкспоненциальным распределением с шестью экспонентами.

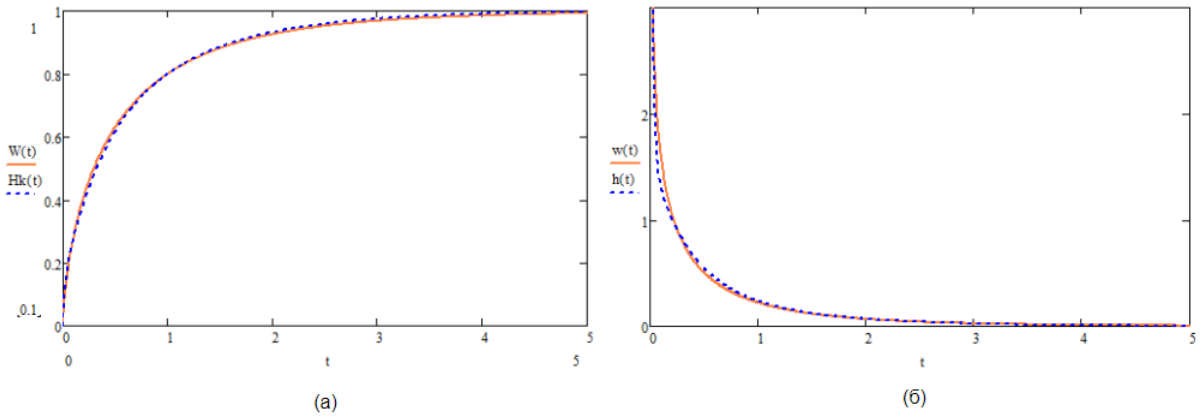


Рисунок 3.21 – ФРВ и ПРВ Вейбулла-Гнеденко и его аппроксимации ГЭ распределением 6 порядка

Можно утверждать, что кривые выглядят практически идентично, но требуются и объективные параметры оценки. В качестве количественных характеристик, определяющих точность апроксимации можно вычислить абсолютные (3.70) и относительные (3.71) погрешности, соответственно:

$$\Delta_{\text{абс.}} = |H^c(t) - F^c(t)| = |H(t) - F(t)| \tag{3.70}$$

$$\Delta_{\text{отн.}} = \frac{|H^c(t) - F^c(t)|}{min\{F^c(t), F(t)\}} * 100\% \tag{3.71}$$

В результате данного анализа мы получили апроксимационную формулу для распределения Вейбулла-Гнеденко, с относительно малой (<0,8 %) ошибкой аппроксимации как видно на рисунке 3.22. Это позволит производить преобразования Лапласа неэкмпоненциальных ФРВ для моделирования систем с целью получения временных характеристик систем далее в работе.

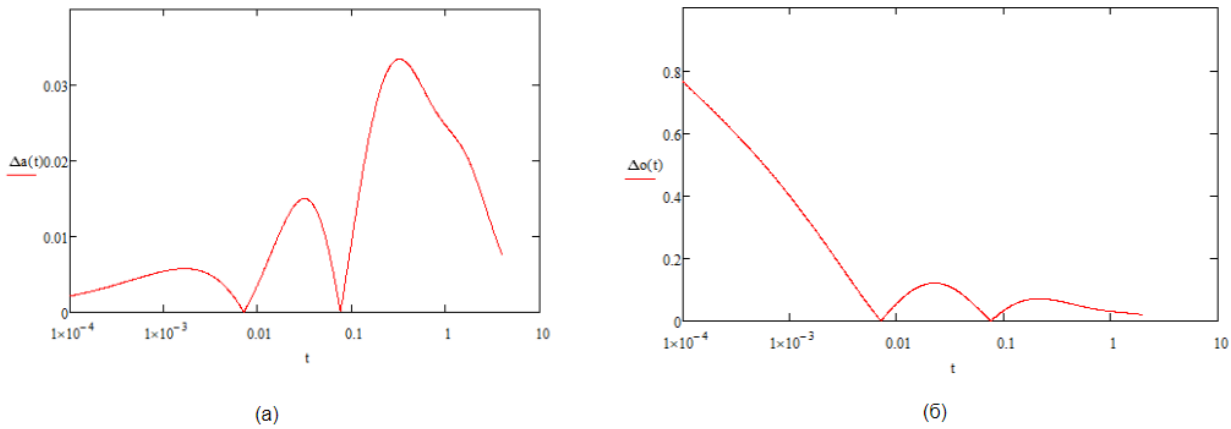


Рисунок 3.22 – Абсолютная и относительная ошибка аппроксимации

Аналогично данному методу в работе [145] автор предлагает аппроксимацию распределений с помощью метода максимального правдоподобия (MLE) (при которой оценка критериев согласия (GOF) осуществлялась с помощью статистического критерия Колмогорова-Смирнова), но уже по измерениям реального трафика.

3.5 Оценка временных задержек в последовательной цепочке локальных транзакций

3.5.1 Проблематика оценки выходного потока систем G/G/1

Как уже было описано в разделе 3.3.4 параметры задержки рассчитываются с использованием преобразования ЛС ФРВ времени между поступлением запросов и времени обслуживания запроса в системе. Данный подход применим для систем с произвольными распределениями для входных потоков и обслуживания G/G/1. Однако вопрос применения данного подхода в расчете цепочки локальных транзакций формирует проблему, связанную с определением ФРВ времени между поступлением запросов на последующие узлы j (j=2…Y) в цепочке.

На этапе определения выходного потока в цепочке последовательных узлов D(s) описанного в разделе 3.3.2, появилась необходимость определения параметра I`(s) – ЛС ФРВ времени до поступления следующего требования,

иначе говоря периода простоя системы. В общем случае для системы G/G/1 нет решения данной задачи, в связи с тем, что период простоя зависит от того, как проходил предыдущий период занятости.

Системы данного типа являются зависимыми и не обладают свойствами систем M/M/1 где поток запросов в системе независимый, следовательно, готового общего решения такие системы не имеют.

В таком случае точное решение имеют системы типа M/M/1 описанные в предыдущих главах и частные случаи, такие как системы типа M/G/1 и G/D/1.

3.5.2 Системы массового обслуживания M/G/1

Системы вида M/G/1 предусматривают произвольный характер функции B(t), анализ данных систем проведен в работе [108]. Данное семейство распределений могут включать такие распределения как M/D/1 или M/M/1.

Основные результаты исследования для систем с произвольным характером обслуживания определенные «формулой Полячека-Хинчина»: формулах, выражающих среднюю длину очереди (3.72), длительности ожидания обслуживания (3.73) определенные «формулой Полячека-Хинчина».

— средняя длина очереди

$$N = \rho + \frac{\rho^2(1 + C_S^2)}{2(1 - \rho)} \qquad (3.72)$$

— длительность ожидания обслуживания

$$W = \frac{\rho(1 + C_S^2)}{2(1 - \rho)}\bar{T} \qquad (3.73)$$

где C_s – коэффициент вариации длительности обслуживания запросов для (3.73)

$$C_s = \sqrt{1 + \frac{(1-2p)^2}{2p(1-p)}} \qquad (3.74)$$

где p - параметр формы.

Преобразование Лапласа-Стилтьеса для периода простоя в системе M/G/1 будет иметь показательное распределение и рассчитываться выражением (3.75). Что позволяет определить ЛС ФРВ выходного потока и точное решения для ФРВ времени ожидания и времени нахождения в каждой системе.

$$I`(s) = \frac{\mu}{\mu + s} \qquad (3.75)$$

Для реализации математического аппарата свертки необходимыми формулами систем M/G/1 являются разновидности «уравнения Полячека-Хинчина». Данные уравнения определяют преобразования Лапласа-Стилтьеса:

– длительности ожидания

$$w(s) = \frac{s(1-\rho)}{s - \lambda + \lambda\beta(s)} \qquad (3.76)$$

– задержки запросов в СМО

$$\varepsilon(s) = \frac{s(1-\rho)\beta(s)}{s - \lambda + \lambda\beta(s)} \qquad (3.77)$$

где $\beta(s)$ – преобразование Лапласа-Стилтьеса для ФРВ времени обслуживания, λ – интенсивность входного потока, ρ – коэффициент загрузки узла.

Узлы M/G/1 в цепочке локальных транзакций являются независимыми системами, что позволяет определить время прохождения цепочки методом свертки определённом в (3.16).

3.5.3 Применение метода аппроксимации для расчета последовательных сетей систем G/M/1

При расчете систем с произвольным законом поступления запросов в систему возникает сложность в определении функции распределения выходного потока запросов после обслуживания в системе с показательным распределением времени обслуживания. Предположение об аппроксимации выходного потока экспоненциальным распределением с интенсивностью, равной среднему времени между приходом запросов в систему, является ошибочным подходом. Данное предположение содержит неточности и не учитывает реальные условия функционирования системы. Системы G/M/1, в которых поступление запросов подчиняется произвольному закону, представляют сложные сценарии, для которых нет точных общих решений. В связи с этим, для определения преобразования ЛС функции распределения времени простоя, было предложено несколько методов аппроксимации.

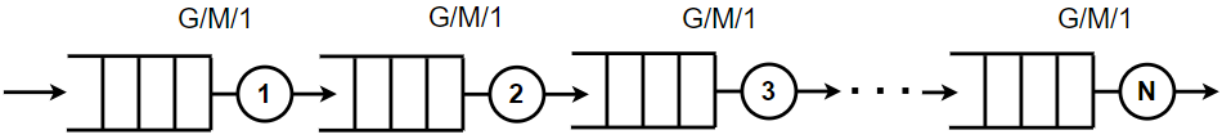


Рисунок 3.23 – Последовательная цепочка систем G/M/1

В качестве входных характеристик системы мы определили входной поток как гиперэкспоненциальное распределение шестого порядка H_6 с параметрами, определенными в Таблице 3.1 и экспоненциальным временем обслуживания. Оценка времени ожидания обслуживания получена с использованием метода, описанного в разделе 3.3.4.

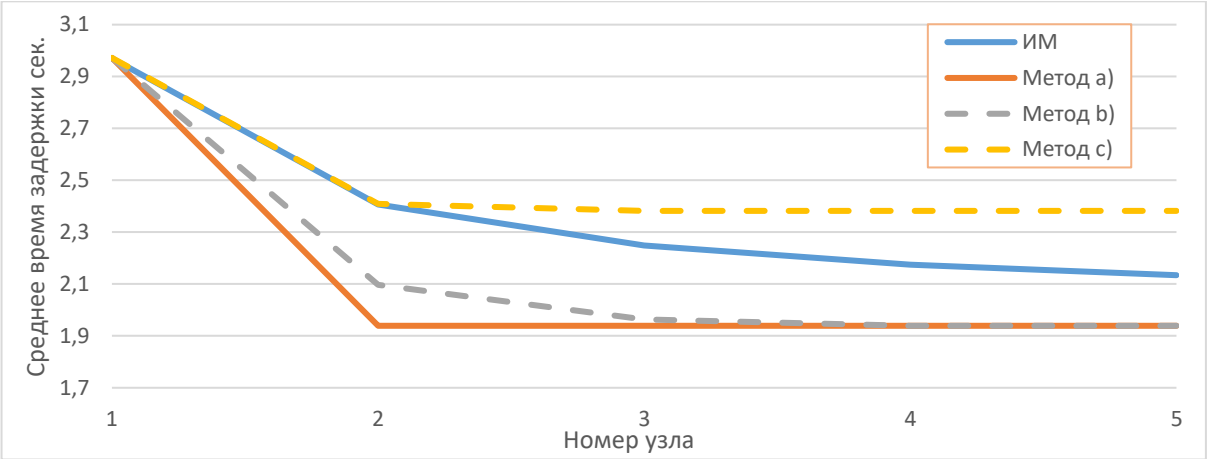


Рисунок 3.24 – Среднее время ожидания на каждом узле в зависимости от метода аппроксимации

a) Аппроксимация I`(s) экспоненциальным распределением. Как видно из рисунка данных подход не учитывает особенности распределений с высокими значениями дисперсии и имеет высокую ошибку аппроксимации (≈20 %) в расчёте времени ожидания обслуживания на втором узле.

b) Аппроксимация периода простоя временем между поступлением запросов в систему:

$$\mathrm{D(s)} = (1 - \rho)A(s)B(s) + \rho B(s) \qquad (3.78)$$

где A(s) – преобразование ЛС ФРВ времени между поступлением запросов.

В случае гипперэкспоненциального распределения вероятности времени между поступлением запросов A(s) будет определено выражением (3.69).

c) Аппроксимационное предположение, представленное в [35] где период простоя будет равен остаточному значению промежутка времени между поступающими требованиями. Основываясь на предположении [35, 191] «…длинные промежутки времени между моментами восстановления захватывают на оси времени большие участки, чем короткие промежутки, поэтому более вероятно, что случайно выбранная точка t попадет на длинный промежуток.» можно определить преобразование ЛС ФРВ периода простоя следующим образом:

$$\Gamma`(s) = \begin{cases} \dfrac{1 - I_a`(s)}{s\bar{x}} & \text{при } \rho \le 0{,}5 \\ \dfrac{1 - B(s)}{s\bar{m}} & \text{при } \rho > 0{,}5 \end{cases} \qquad (3.79)$$

где $\bar{m}$ – среднее время обслуживания

$B(s)$ - ПЛС ФРВ времени обслуживания

$I_a`(s)$ –аппроксимация ПЛС ФРВ остаточного времени $A`(s) = \frac{\varepsilon}{\varepsilon+s}$

$\varepsilon = 1/\bar{x} = 1/(\bar{t} - \bar{m})$, $\bar{x}$ – среднее значение остаточного времени, $\bar{t}$ – среднее время между приходом запросов.

Необходимые значения для расчета ФРВ времени задержки (3.52) на втором узле цепочки систем при реализации «b» и «c» методов аппроксимации возможно определить построив графики искомых функций на рисунке.

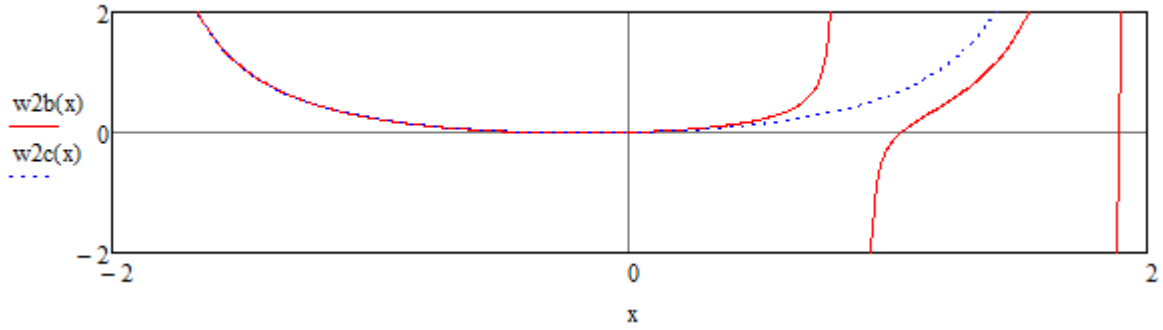


Рисунок 3.25 – Графики искомых функции

В данном случае стоит отметить высокую сходимость значений на втором узле в сети, однако дальше данный метод не дает возможности передать характер изменения выходного потока на последующих узлах (2..n). Из-за этого возрастает относительная ошибка аппроксимации, и становится сложно предсказать характер поведения очередей. Однако оценка задержки в цепочке СеМО при возрастающей нагрузке, определен для трех методов на рисунке 3.26.

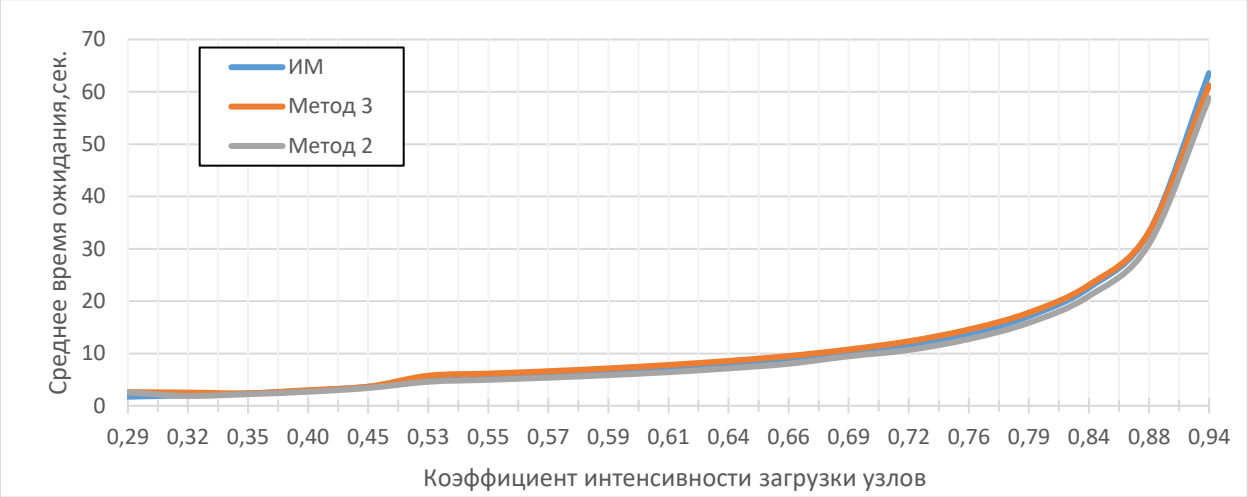


Рисунок 3.26 - Оценка средней задержки для всей СеМО для 3 методов моделирования

Точность данных методов можно оценить, рассматривая рисунок 3.27. Как видно метод «c» может принимать значения близкие к значениям, полученным способов имитационного моделирования, однако существуют такие уровни ρ, при которых ошибка аппроксимации может превышать 10 %. Однако при увеличении ρ→1 процент ошибки снижается что обусловлено выражением (3.35).

Обратный результат получается при использовании метода «b». При его реализации как видно из рисунка 3.27 ошибка аппроксимации постоянно варьируется в рамках 6-8 процентов, однако это метод не дает точных результатов при любой загрузке системы.

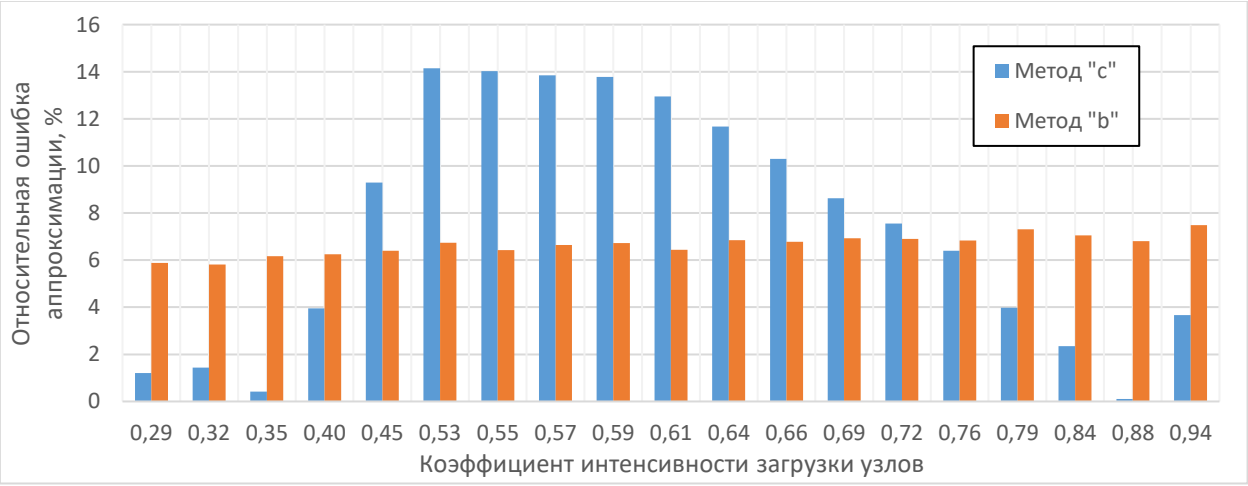


Рисунок 3.27 – Относительная оценка аппроксимации для 2 методов моделирования к результатам ИМ

Не один из предложенных методов не может гарантировать точные результаты, однако при проведении анализов полученных результатов была отмечена возможность использования двух методов (b) и (c) в качестве выражений для определения нижней и верхней границы значений искомой величины, как показано на рисунке 3.28.

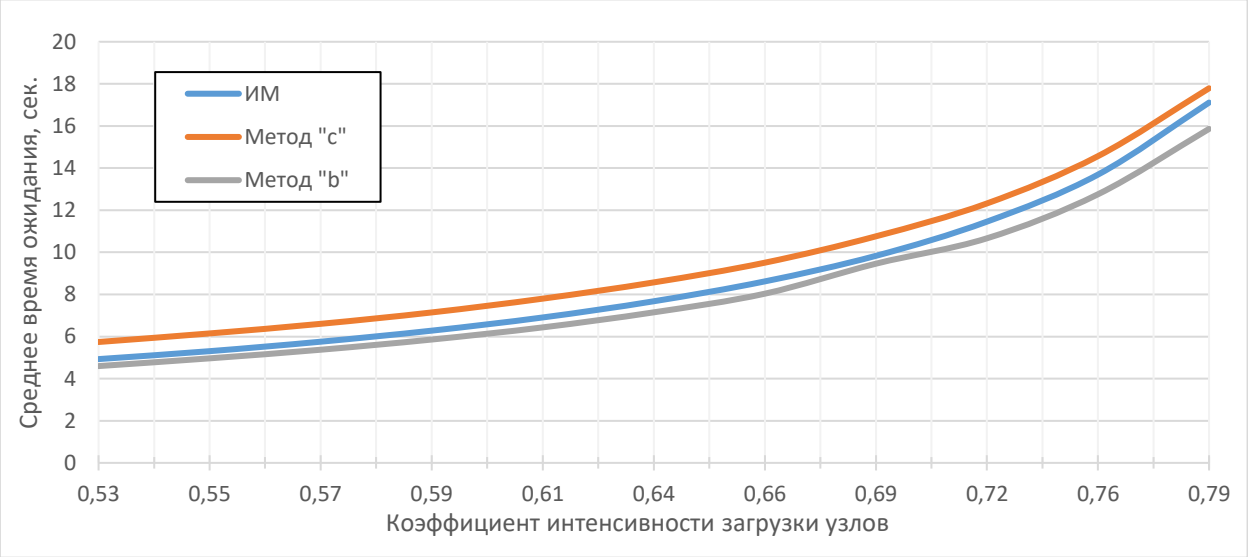


Рисунок 3.28 - Оценка задержки для средних значений нагрузки.

Интересные результаты наблюдаются при исследовании изменения характера потока в зависимости от количества узлов в сценарии рабочего процесса обслуживания запроса в транзакционной системе. Рисунок 3.29 позволяет оценить зависимость скорости сходимости гиперэкспоненциального входного потока к экспоненциальному распределению в зависимости от коэффициента загрузки узлов. Из графика видно, что чем больше нагрузка тем выше скорость приближения формы входного потока к форме закона распределения для времени обслуживания в системах G/M/1.

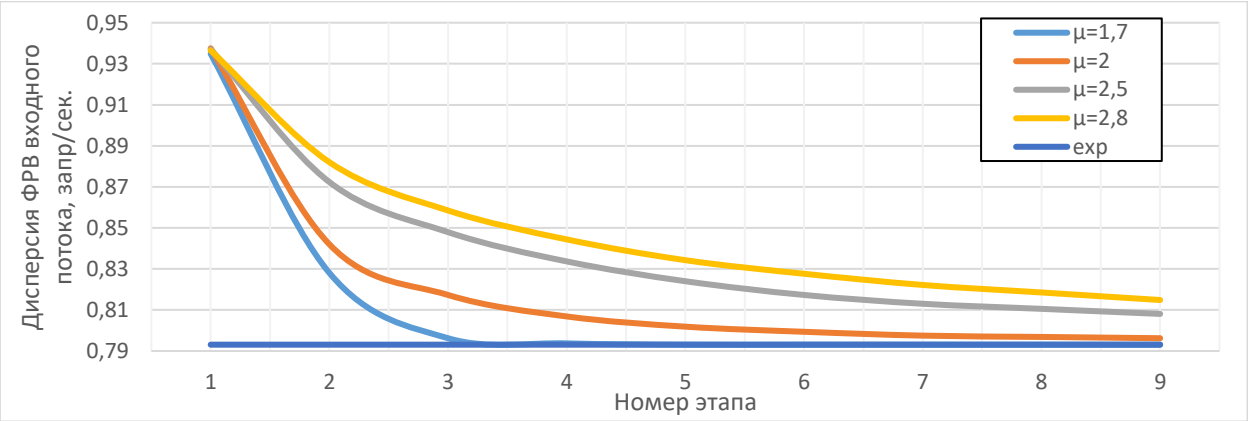


Рисунок 3.29 – Дисперсия ФРВ времени между поступлением запросов на каждый узел в сети для разных интенсивностей обслуживания

В основном на скорость сходимости влияет загруженность системы, что обусловлено выражением (3.35). Результаты измерения для таких распределений как Гамма, Парето и Логнормальное распределения, доказывают данный вывод на рисунке 3.30 при μ~const и $\bar{x}$~const.

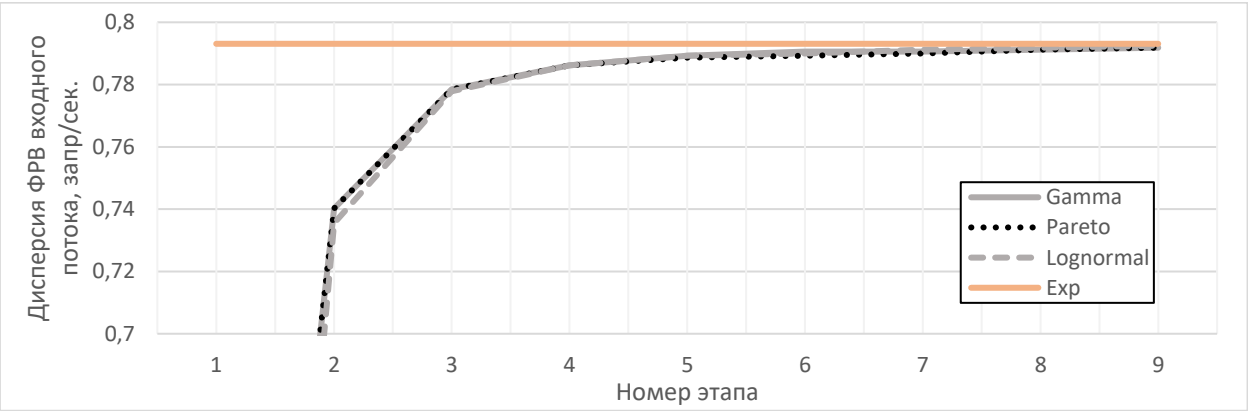


Рисунок 3.30 – Дисперсия ФРВ времени между поступлением запросов на каждый узел в сети для разных законов поступления

Как видно из рисунка 3.30 с каждым последующей обработкой потока запросов в узлах в СеМО характеристики входного потока в системы G/M/1 становятся близкими к экспоненциальному потоку.

3.6 Выводы по главе 3

В данном разделе был разработан метод для расчета временных характеристик рабочих процессов обслуживания запросов в транзакционных

слабосвязанных услугах с использованием инструмента свёртки. Данный метод включает математический инструмент, основанный не только на средних значениях, для расчета СеМО последовательных узлов включая ветвление, слияние рабочих процессов. Данный метод позволяет производить расчет систем с параллельной обработкой запросов в рамках рабочего процесса используя математический инструмент Fork-Join.

Методы для получения точных результатов расчета временных характеристик были предложены для экспоненциальных систем M/M/1, при этом для оценки характеристик цепочек систем произвольного типа G/G/1 необходимо внедрение методов аппроксимации, разработанных и формализованных в данной главе, а именно:

1) Методы аппроксимации *неэкспоненциальных* распределений с высокими коэффициентами вариации, позволяющие обойти сложности при аналитическом моделировании и при математических расчетах последовательных сетей систем массового обслуживания.

2) Методы аппроксимации оценки характеристик выходного потока, обслуженного в узлах рассматриваемых сетей систем произвольного вида G/G/1, для которых нет точного общего решения для оценки характеристик выходного потока при всех значениях нагрузки.

ГЛАВА 4 МЕТОДЫ РАСЧЕТА ХАРАКТЕРИСТИК ТЕЛЕКОММУНИКАЦИОННЫХ УСЛУГ

В данной работе была разработана методика, объединяющая результаты разработки аналитических математических моделей, имитационного моделирования для повышения эффективности процесса проектирования информационных систем.

Данная методика основана на использовании аппроксимационных формул, которые позволяют приближенно описывать характер поведения запросов в рабочих процессах. Эти формулы были разработаны на основе исследований, представленных в 3 главе данной диссертационной работы. Аппроксимационные формулы предоставляют возможность оценить различные параметры и характеристики запросов, что помогает в анализе и определении требований к рабочим процессам проектируемых и функционирующих систем.

4.1 Методика проектирования телекоммуникационных услуг

4.1.1 Особенности моделирования при проектировании ИК систем

На этапе планирования ИС возможно использование аналитического моделирования, представляющего из себя аппроксимационный расчет на основе стационарных, ординарных систем без последействия M/M/1. Это позволит определить возможные параметры системы в том числе этапы возникновения узких мест в выделенных рабочих процессах.

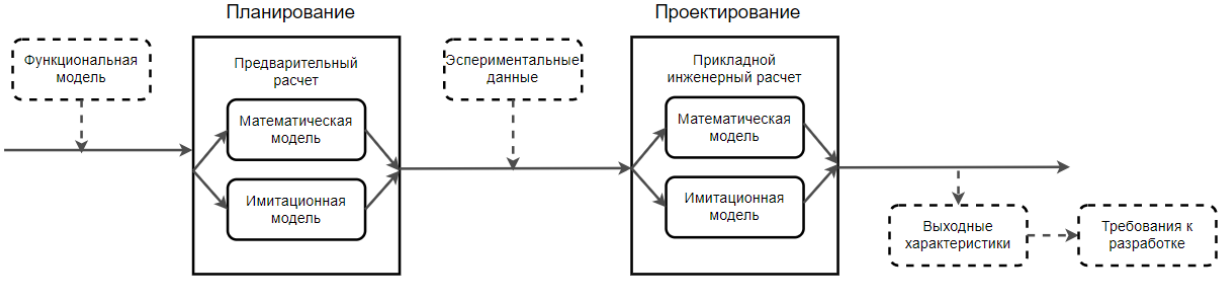


Рисунок 4.1 – Этапы проектирования телекоммуникационных услуг

Аппроксимационные расчеты с помощью математических моделей сетей систем M/M/1 могут помочь в идентификации узких мест в рабочих процессах ИКС. Узкие места — это компоненты или процессы, которые ограничивают общую производительность системы. Их идентификация важна для корректного функционирования систем и повышения их эффективности. Рекомендуется так же внедрение на данном этапе имитационного моделирования, с использованием данных, полученных путем построения функциональной и математической модели, для оценки корректности реализации математического аппарата.

Возможно дальнейшее проектирование систем, на основе данных реального рабочего процесса, полученных экспериментальным путем, которое может повлечь изменение входных характеристик системы и появление систем произвольного типа G/G/1. Для последующего моделирования систем возможно использование разработанной имитационной модели или методов моделирования систем, включающих неэкпоненциальные распределения с высоким коэффициентов вариации. Имитационное моделирование позволит получить точные характеристики системы, без применения методов аппроксимации, такие как моменты и дисперсию выходного потока. Так же имитационное моделирование позволяет производить последующую регулировку и наглядные испытания на основе реальных данных.

Однако ограничение использования данного метода моделирования заключается в невозможности получения единовременно выходных характеристик при значительном количестве итераций.

Моделирование ИС с помощью математических моделей позволяет нам анализировать и предсказывать поведение системы, оценивать производительность и эффективность ее компонентов, и управлять количественными и качественными параметрами системы до ее фактической реализации.

4.1.2 Алгоритм методики расчета ВВХ ИК систем

Объединив результаты полученные из 1-3 главы диссертации, предлагается следующий обобщенный алгоритм методики расчета вероятностно-временных характеристик телекоммуникационной системы:

Шаг 1. Получение основных входных характеристик системы, которые включают множество внешних воздействий на сервис C = $\{c_1, c_2, \dots c_n\}$ и множество внутренних характеристик системы B = $\{b_1, b_2, b_3 \dots b_n\}$.

Шаг 2. Определение матрицы взаимодействия сеансовых (SE) – услуг и транзакционных услуг (TR). Результатом должна стать матрица R_{ji}^{c} интенсивностей переходов из j-узла сеансовой услуги в i – рабочий процесс транзакционной услуги.

Шаг 3. Получение выходных характеристик сеансовых услуг

Шаг 3.1. Определяем архитектуру сеансовой услуги. Результатом должна быть матрица вероятностей переходов поступающих запросов на пользовательский сеанс между узлами системы - R_{ji} где j – узел отправления запроса, i – узел поступления запроса.

Шаг 3.2. Используя входную характеристику – интенсивность общего входного потока и матрицу переходов, по теореме Джексона (2.10) рассчитываем интенсивность входного потока λ_i на каждый узел Se_i.

Шаг 3.3. Имея в качестве входной характеристики интенсивность потока на каждый узел λ_i, рассчитываем качественные характеристики системы по следующим формулам (2.12) - (2.20).

Шаг 4. Получив в качестве входной характеристики нагрузку узла пользовательских сессий A_{SE}, а также используя матрицу R_{ij}^{c} проведем расчет входного потока с SE_j на транзакционные системы TR_i.

Шаг 4.1. В случае пуассоновского входного потока рекомендуется рассматривать генерацию запросов на транзакционную услугу как систему типа M/M/1/∞/N с конечным числом источников нагрузки, используя выражения (2.1) - (2.2).

Шаг 4.2. Если моделируется система с известными характеристиками, то входной поток возможно получить экспериментально. Так же в случае, если предполагается сложное распределение, то характеристику входного потока возможно получить с помощью имитационного моделирования и эмпирической ступенчатой функции.

Шаг 5. Получив в качестве входных характеристик интенсивность входного потока в транзакционную услугу λ_i^{TR}, проведем расчет качественных характеристик рабочих процессов транзакционных услуг.

Шаг 5.1. В качестве входных параметров для последующего анализа сценария транзакционной услуги рассматриваем ФРВ времени между поступлением запросов A(t) и ФРВ времени обслуживания запроса в системе B(t).

Шаг 5.2. Определяем сценарий рабочего процесса для транзакционной услуги.

Шаг 5.3. Проверка условия возможности интегрирования ФРВ, описывающих входные параметры. В случае если один из выбранных параметров не интегрируется или является ступенчатой функцией полученной экспериментально, необходимо произвести аппроксимацию распределения, методы которой описаны в разделе 3.2.

Шаг 5.4.1. Начиная с первого узла в цепочке локальных транзакций необходимо получить преобразование Лапласа-Стилтьеса (3.18) для ФРВ времени обслуживания B(s) и времени между поступлением запросов на узел A(s).

Шаг 5.4.2. Расчет параметров времени ожидания обслуживания в узле $W_i(s)$.

a) Для систем типа M/M/1 возможно использование выражения (3.19)

b) Параметры систем типа M/G/1 могут быть получены с использованием формул (3.73), (3.76).

c) Расчет систем типа G/G/1 предусматривает использование спектрального решения интегрального уравнения Линдли (3.41) - (3.56).

Шаг 5.4.3. В случае, если рабочий процесс обслуживания запроса в транзакционной системе включает количество узлов N>1 необходимо определить изображение плотности выходного потока D_i(s):

a) В случае Пуассоновского входного потока D(s)=A(s);
b) В случае произвольного входного потока, выходной поток определяется методами аппроксимации описанными в 3.5.3.
c) В случае систем G/G/1 следует использовать средства имитационного моделирования.

Шаг 5.4.4. Определение изображения плотности вероятности времени нахождения в системе $F_{ei}(t)$ используя свертку распределений времени ожидания $W_i(t)$, и обслуживания $F_i(t)$.

Шаг 5.4.5. Выполнение пунктов с 5.3. – 5.4.4, пока i≤A.

Шаг 5.5. Определение плотности вероятности времени обработки запроса в сценарии рабочего процесса $F_x(t)$, с использованием процедуры свертки ПРВ времени нахождения в системе (3.16) и обратное преобразование Лапласа результата.

Шаг 5.6 Проводим расчет первого момента и дисперсии используя выражения (3.21) -(3.22).

Алгоритм предложенной методики представлен в виде SDL-диаграмм на рисунках 4.2 - 4.3.

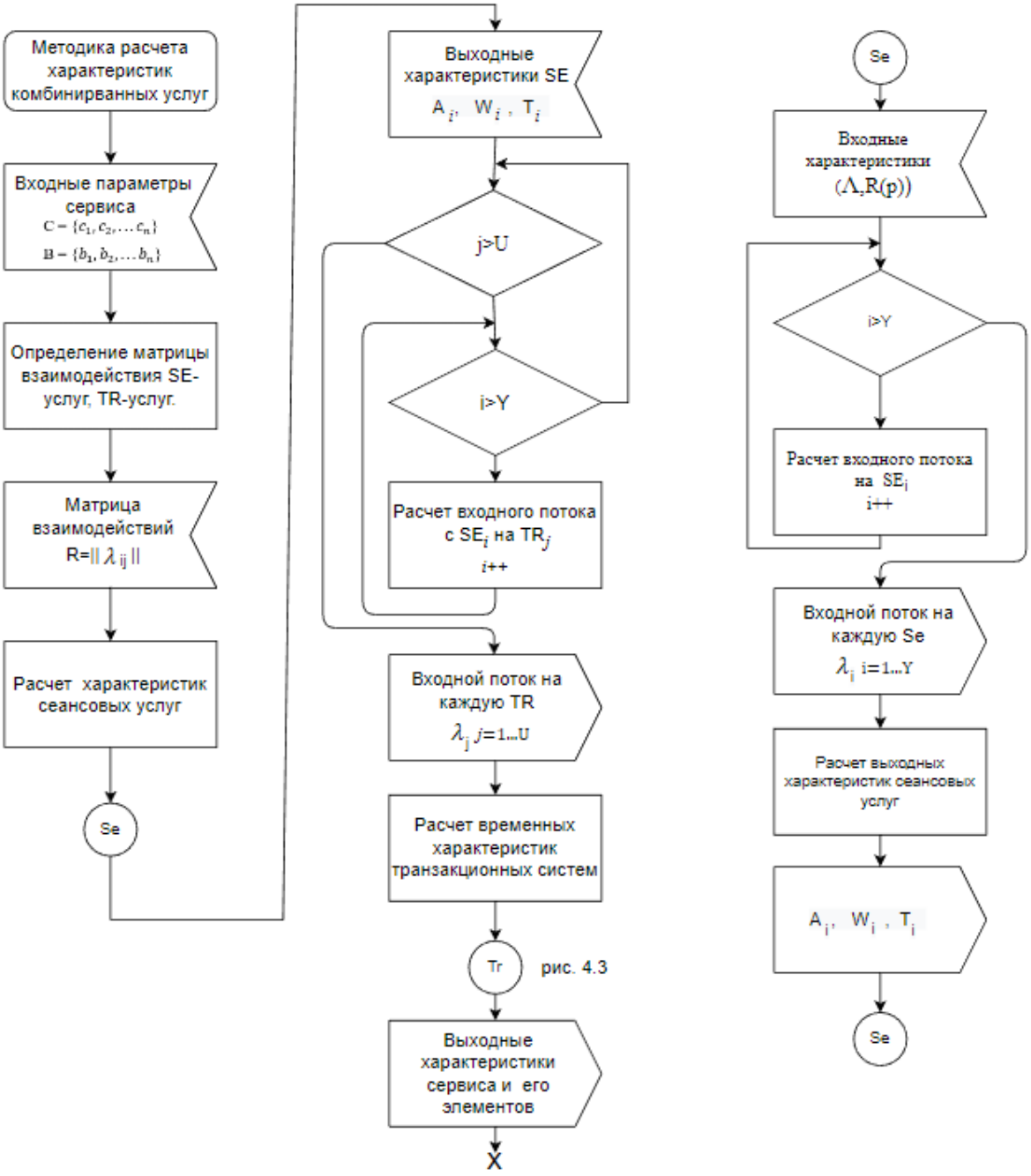


Рисунок 4.2 – Алгоритм расчета качественных характеристик комбинированных услуг

Отдельно представляется к рассмотрению алгоритм применения метода оценки временных характеристик рабочих процессов транзакционных услуг, которые могут иметь множество сценариев реализации, качественная оценка которых позволит принимать обоснованные решения на всех этапах

планирования и проектирования услуг без и с учетом реальных входных данных.

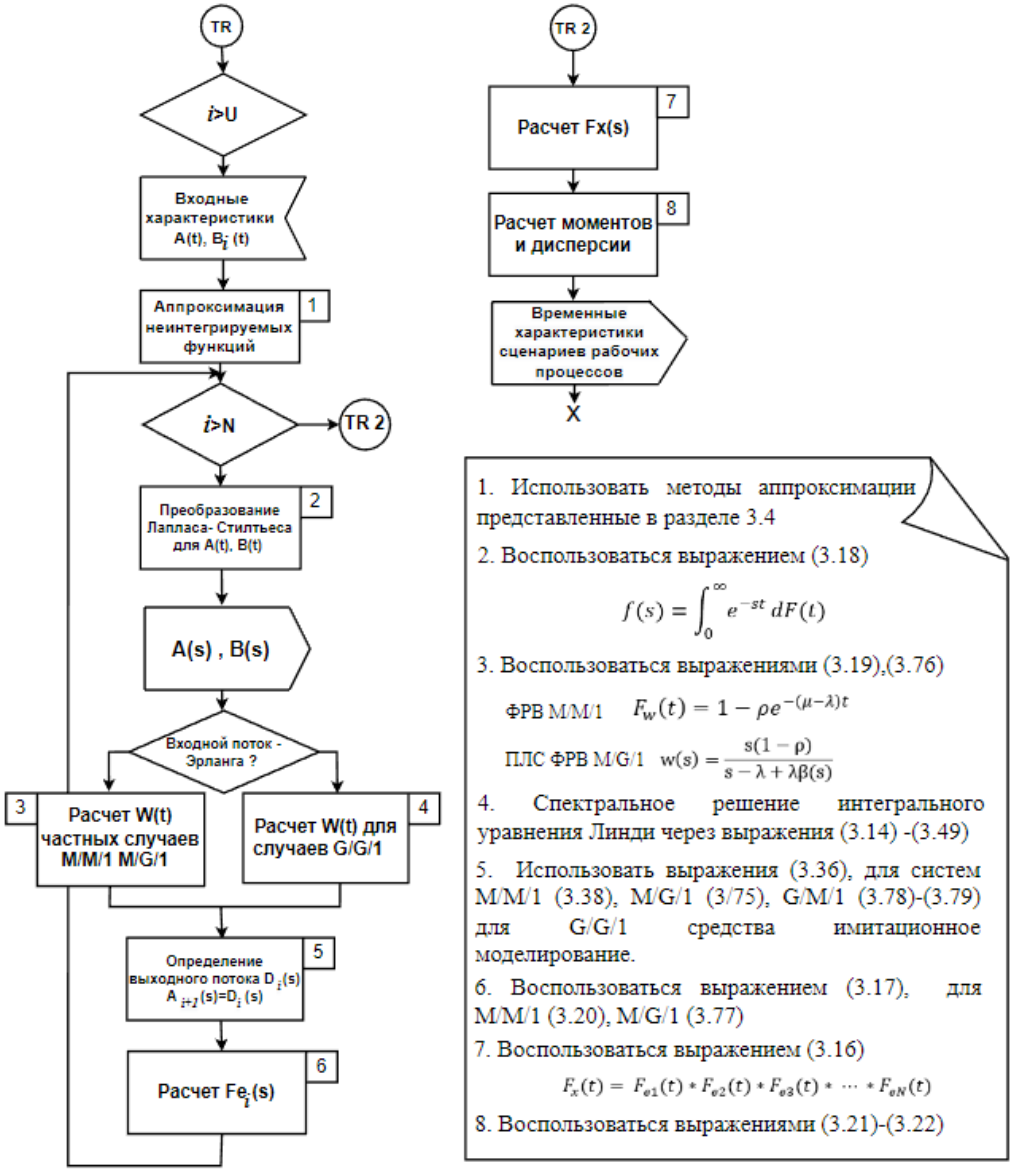


Рисунок 4.3 – Методика расчета транзакционных услуг

4.2 Пример расчета временных характеристик комбинированной услуги

4.2.1 Моделирование сценариев реализации услуги «Авторизации»

Отметим, что в представленном алгоритме возможны исключения. Например, представленный на рисунке 3.1 сценарий №1 авторизации в форме UML-диаграммы является транзакционной системой, предоставляющей

входные данные для сеансовой услуги, а не наоборот, как было описано ранее. Время прохождения цепочки транзакций считается временем обслуживания пользователя в первой сеансовой СМО, а интенсивность выходного потока из последнего узла в цепочке соответствует интенсивности выходного потока из первого узла сеансовой услуги. Тем не менее, данный пример явно демонстрирует множество возможных сценариев обслуживания в рабочем процессе, которые следует учитывать при проектировании и оценочном моделировании системы.

Согласно предусмотренному сценарию работы, возможно повторное отправление запроса на первый узел в цепочке в случае ошибочной авторизации, при этом в обслуживании будет отказано после трех неудачных попыток. Важно отметить, что весь этот процесс происходит в рамках одного пользовательского сеанса.

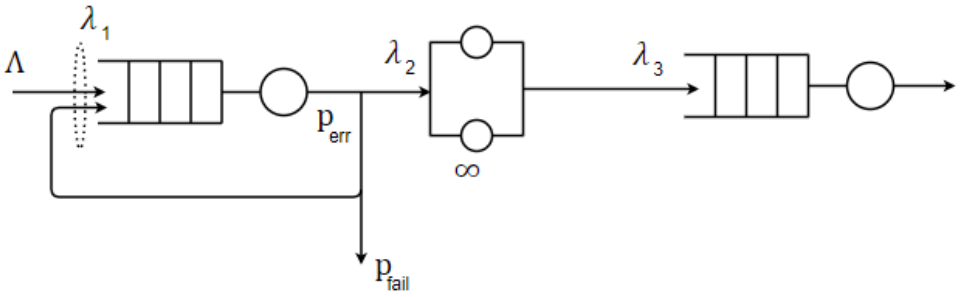


Рисунок 4.4 – Пример модели услуги «Авторизации»

Описанный сценарий предусматривает решение следующей системы уравнений для расчета интенсивности входного потока:

$$\begin{cases} \lambda_1 = \Lambda * (1 + p_{err} + p_{err}^2) \\ \lambda_2 = \lambda_1 * (1 - p_{fail}) \\ \lambda_3 = \lambda_2 \end{cases} \quad (4.1)$$

где p_{err} – вероятность ошибки в авторизации, $p_{fail} = p_{err}^3$ - вероятность отказа в авторизации после трех ошибок

С учетом (2.23) и (4.1) выражение для расчета среднего времени пользовательской сессии будет выглядеть следующим образом:

$$\bar{S} = \sum_{i}^{3} \bar{S_i} = \frac{(1 + p_{err} + p_{err}^2)(1 - p_{fail})}{\mu_1(1 - \frac{\Lambda * (1 + p_{err} + p_{err}^2)}{\mu_1})} + \frac{1}{\mu_2} + \frac{1}{\mu_3(1 - \frac{\Lambda * (1 + p_{err} + p_{err}^2)(1 - p_{fail})}{\mu_3})} \quad (4.2)$$

Второй вариант предусматривает в случае ошибочной авторизации отказ в обслуживании пользователя и его повторное обращение к сервису. Тем самым мы увеличивает входной поток к сеансовой услуге Λ на долю потерянных запросов в транзакционной p_{err} , тем самым изменив сценарий работы транзакционной системы. И тогда интенсивность входного потока на первых узел λ_1будет равна входной интенсивности Λ.

Комбинирование аналитических расчетов с имитационным моделированием, позволяет сократить трудозатраты и предоставить более точные данные для оценки качественных характеристик сложных комплексных систем. Таким образом, полученные результаты дают возможность принимать обоснованные решения при проектировании, внедрении систем и их последующем управлении, с учетом различных сценариев работы.

На рисунке получены временные характеристики работы системы при первом варианте (а) и втором варианте (б) реализации рабочего процесса имеющие незначительные различия.

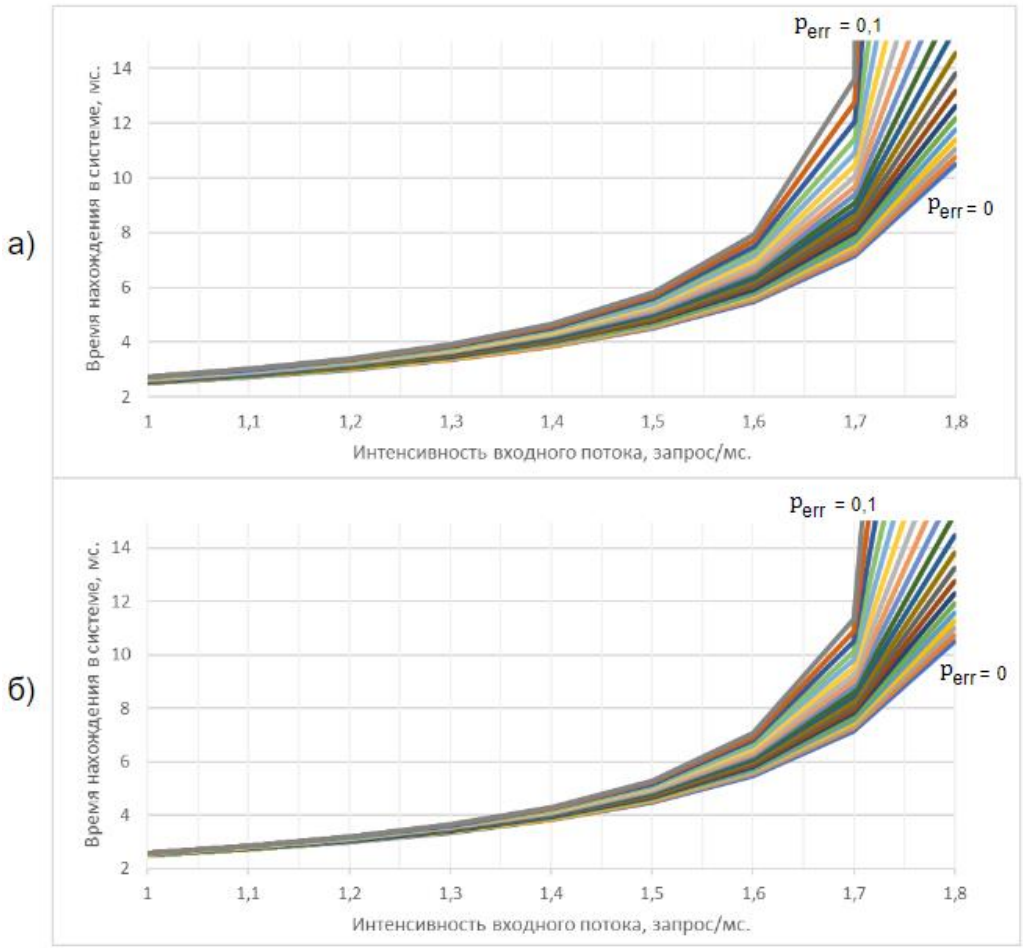


Рисунок 4.5 – Зависимость T_s от Λ при разных вероятностях ошибки.

На рис. 4.6 представлены ПРВ времени нахождения запроса в СеМО, представленном на схеме (рис. 4.4) при трех сценариях реализации услуги «Авторизации», полученные путем имитационного моделирования.

Сценарий №1 соответствует ранее описанному сценарию с повторным посещением пользователя сервиса при ошибочной попытке аутентификации. Сценарий №2 предполагает несколько повторных попыток аутентификации пользователя до отказа в обслуживании. Последний сценарий №3 предполагает полный отказ в обслуживании незарегистрированного пользователя.

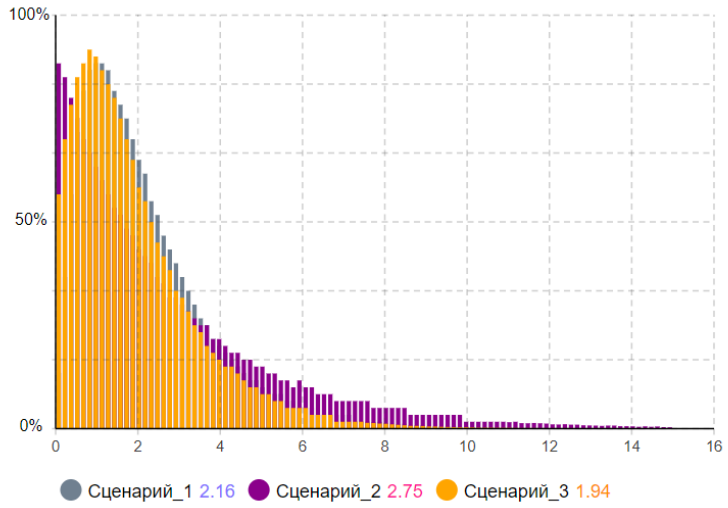


Рисунок 4.6 – ПРВ времени присутствия пользователя на сервере

Результаты моделирования позволяют предполагать возможность проектирования системы реализующий данный рабочего процесса, ориентируясь только на функциональные потребности, так как выбор между вариантами не значительно влияет на временные характеристики системы.

4.2.2 Оценка сервиса с разработкой имитационной модели

При разработке и эксплуатации готовой системы можно как продолжать использовать аналитический расчет, так и использовать разработанный инструмент имитационного моделирования.

На рисунке 4.7 изображена полная реализация имитационной модели, которая учитывает все характеристики и структуру сервиса схема которого представлена на рисунке 1.2, что позволяет исключить внедрение в моделирование сервиса математических расчетов.

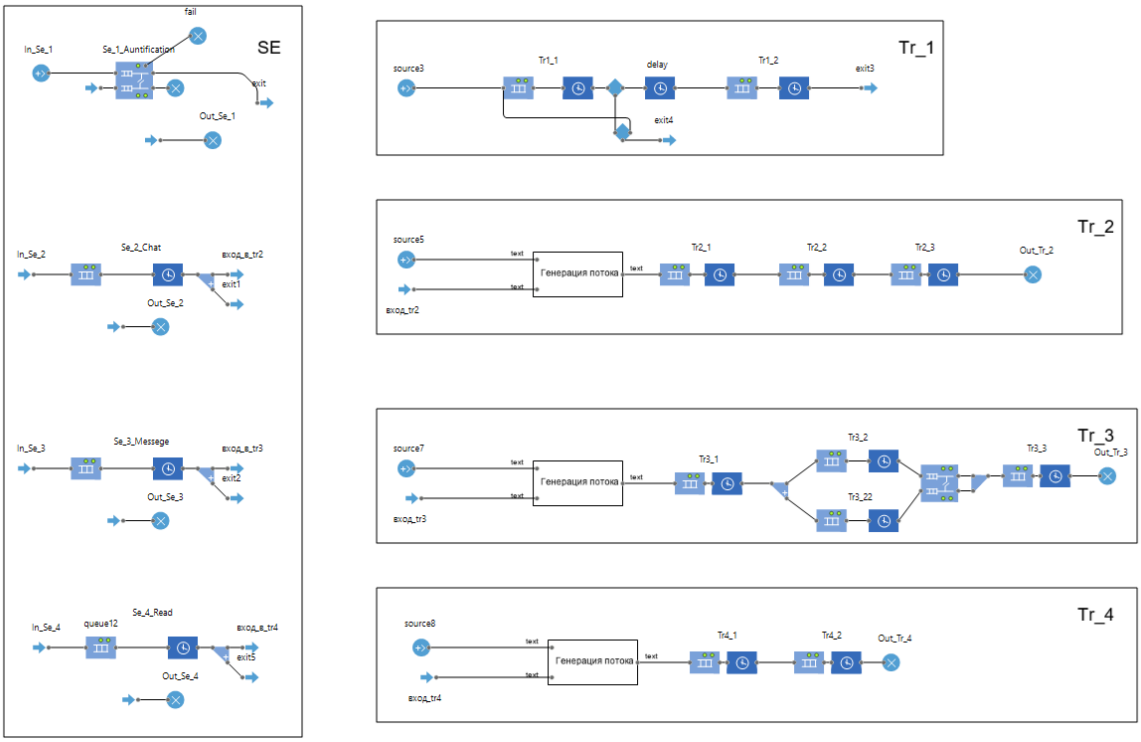


Рисунок 4.7 – Имитационная модель комбинированной услуги, реализующей функционал веб-сервиса.

Данный анализ позволяет определить временные характеристики реализации двух структурных решений. В первом решении каждый сценарий абстрактного сервиса (рис. 1.2) подразумевает обращение к своей базе данных через отдельный микросервис. Во втором решении предполагается, что последним этапом в каждом рабочем процессе является обращение к одной общей базе данных, с формированием необходимого количества ресурсов обслуживания.

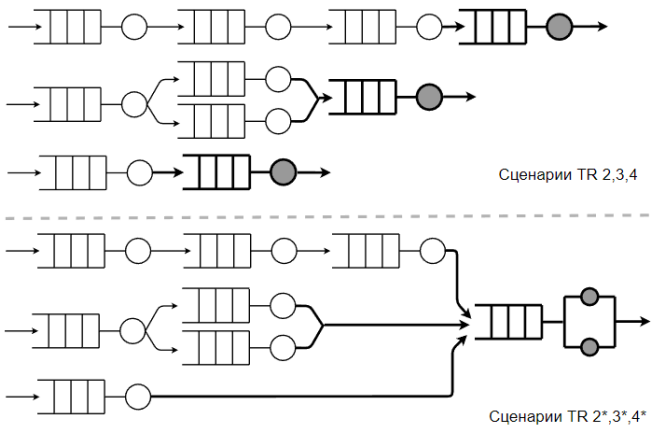


Рисунок 4.8 – Решения для реализации сценариев обслуживания запросов.

Из рисунка видно, что второй вариант сокращает время прохождения запросов рабочих процессов систем (Сценарий*). Это связано с возможностью сокращения времени простоя в обслуживающем узле при эффективном использовании, в том числе перераспределений ресурсов.

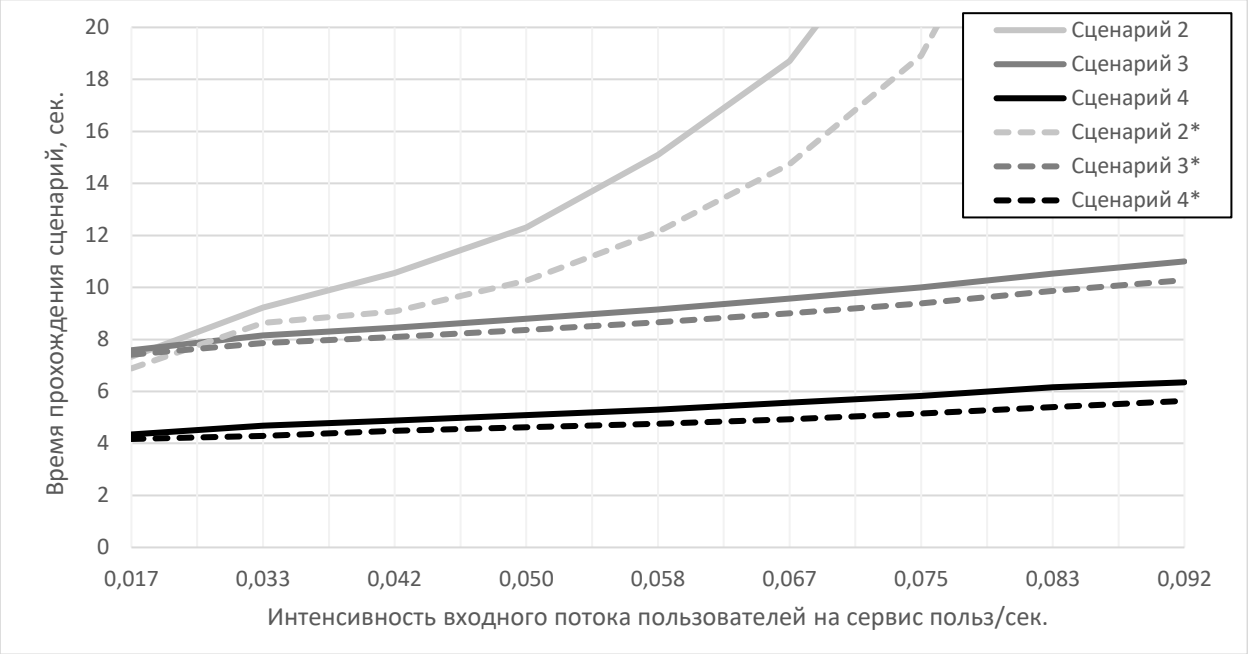


Рисунок 4.9 – Оценка времени прохождения рабочего процесса для нескольких сценариев.

Таким образом, второй вариант реализации рабочего процесса представляется более эффективным с точки зрения временных характеристик, что может сказаться на общей производительности системы. Выбор эффективного и обоснованного решения зависит от конкретных требований к системе, её нагрузки и ограничений. Однако проведенный анализ позволяет принять обоснованное решение при проектировании системы, учитывая временные аспекты работы рабочих процессов.

4.3 Оценка применимости имитационного моделирования

4.3.1 Рекомендации к применению методов моделирования при проектировании телекоммуникационных услуг

Применению имитационных моделей для оценки характеристик систем необходимо:

1) Когда существует сложность в расчете характеристик системы аналитическими методами. Применение ИМ является предпочтительным из вариантов, когда рассматриваемая система имеет сложную архитектуру, когда в расчете фигурирую входные параметры включающие сложные распределения, не поддающиеся прямым манипуляциям и предполагающие разработку или использование аппроксимирующих методов, описанных в разделах 3.4.3 и 3.5.3.

2) В случае если, нужно учесть случайные факторы: ИМ позволяет моделировать случайные события и стохастические процессы, что делает его полезным для систем с неопределенностью и случайными факторами.

3) Для проведения экспериментов без риска для реальной системы. Это позволяет изучать различные сценарии и повысить эффективность процессов. Данное применение ИМ актуально в вопросах поставленных в концепции цифровых двойников [3].

4) Когда данные о системе ограничены: если у вас недостаточно данных о системе или они неполны, ИМ может помочь вам получить представление о поведении системы на основе имитации реальных событий. Так в случае возможно использование имитационного моделирования процесса обслуживания системы G/M/1 для оценки параметров выходного потока как было предложено в разделе 3.5.3.

Далее будут приведены рекомендации, при которых целесообразно будет построение математических моделей для оценки качественных характеристик системы.

1) Простота структуры исходной системы. В случае, когда система имеет простую структуру и можно аналитически описать её поведение, математические модели могут быть более эффективным и точным подходом.

2) Возникает необходимость в проведении множественных опытов, для которых использование готового математического инструмента будет менее трудозатратным решением.

3) Математические модели позволяют анализировать систему в абстрактной форме и исследовать общие закономерности её поведения.

Однако часто рекомендуется использование комбинации данных подходов как показано на рисунке 4. Например, можно использовать имитационное моделирование для предварительного исследования системы и выявления ключевых особенностей, а затем создание математические модели для более точных аналитических результатов. Либо использование имитационной модели для оценки рациональности применения выбранного математического инструмента при оценке выходных характеристик системы.

4.3.2 Использование функционала имитационного моделирования

1) Удобная настройка

Моделирование СеМО которые включает в себя возможность выбора пути прохождения запросом узлов обслуживания в случае транзакционных услуг, так и выбор маршрутов в моделях сеансовых услуг задается матрицей переходов.

Моделирование данного процесса в среде Anylogic возможно несколькими способами, один из которых создание базы данных (а) и определение маршрутов (б) запросов как показано на рисунке:

	маршрут	вероятность_перехода
1	p01	1
2	p12	1
3	p23	0.3
4	p24	0.4
5	p25	0.3

а)

↓i→j	0	1	2	3	4	5
0	0	1	0	0	0	0
1	0	0	1	0	0	0
2	0	0	0	0,3	0,4	0,3

б)

Рисунок 4.10 – Маршрутная матрица переходов между узлами сервиса.

Следующим шагом необходимо в блоках выбора маршрута «SelectOutputOut» → «Вероятность», или «Переменная» → «Начальное значение» с ссылкой на БД прописать следующее условие:

SELECT вероятность_перехода **FROM** «Имя БД»

WHERE маршрут = "pij" // выбрать правильный переход, из i в узел j

Единый элемент для управления параметрами системы позволит снизить трудозатраты на моделирования в крупномасштабны моделях.

2) Возможность масштабирования и использования шаблонов из готовой библиотеки моделей.

В рамках данной работы были разработано семейство имитационных моделей, элементы которых могут быть применены для моделирования различных сценариев рабочих процессов, например:

а) Элемент «Генерации потока»

Элемент имитационной модели, изображенный на рисунке 4.11 был применен в модели рис. 4.7 для снижения сложности модели и возможности её последующего масштабирования.

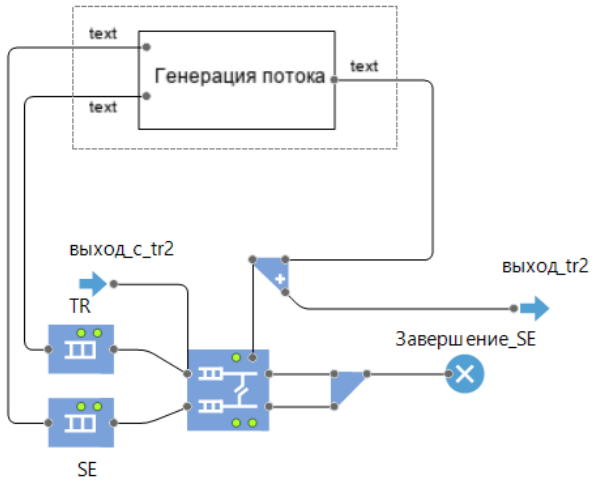


Рисунок 4.11 – Элемент ИМ «Генерация потока TR-услуг»

Данный блок процесса позволяет генерировать поток запросов в транзакционных услугах, в зависимости от состояния пользовательских

сессий в сеансовых услугах, с возможностью индивидуальной настройки параметров функционирования системы (например изменение типа распределения входящего потока и числовых параметров).

б) Элемент «Параллельная обработка запросов»

В работе был представлен математический аппарат параллельной обработки запросов Fork-Join, имитационная модель которого была представлена для параллельной обработки двумя узлами систем. Однако сложность модели резко возрастает при увеличении узлов (i=3…N), при сохранении свойств данного математического аппарата, что показано на рисунке 4.12.

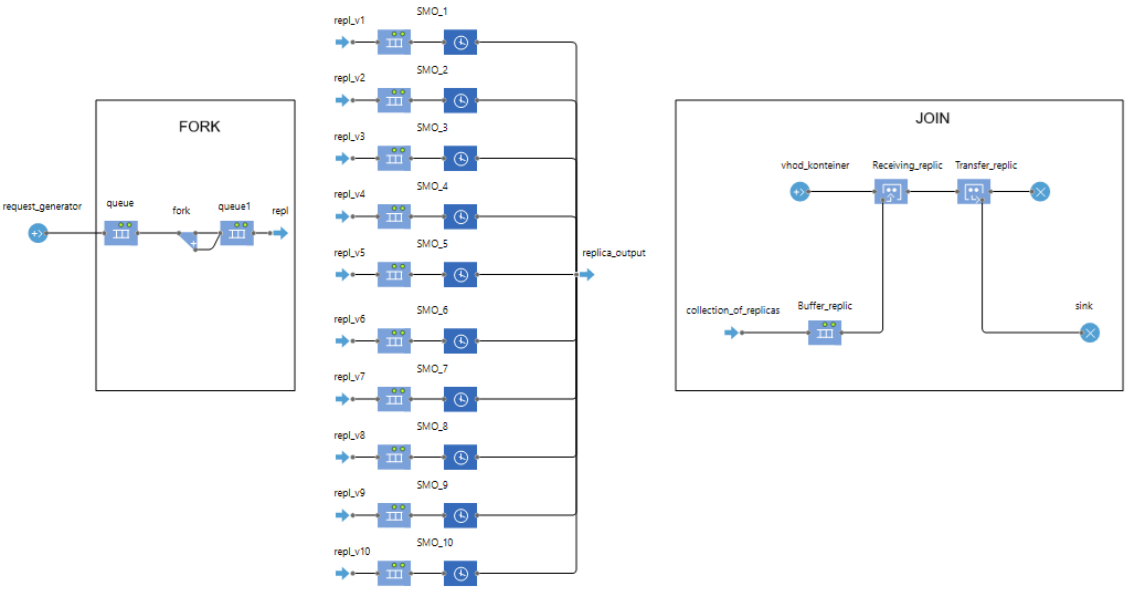


Рисунок 4.12 – Имитационная модель Fork-Join M узлов

Стоит отметить, что в данной имитационной модели автоматизирован процесс изменения количества реплик за счет введения функций, описанных ниже:

Таблица 4.1 - Описание блоков модели FJ

Блок	Назначение	Условие
Функция «number_of_nodes»	Изменяет количество активных узлов на значение, определенное в параметре «M_nodes»	**if** (M_nodes == 1) {Nodes_1();} **else if** (M_nodes == 2) {Nodes_2 ();} и т.д.
Функции	Определяет вероятности	Для коэффициента реплик M=2

«Nodes_i» где i=1…M	обработки реплик конкретными узлами	M_1=1/M_nodes; M_2=1/M_nodes; M_3=0; и т.д.

Queue «Buffer_replic»	Ожидает поступление всех M реплик, прежде чем объединить их в запрос и закончить обслуживание	ID=agent.number;
		replica_counter=0;
		for (i=0;i <Buffer_replic.size();i++)
		{
		if ((Buffer_replic.get(i)).number==ID){
		replica_counter++;
		}
		if (replica_counter==M_nodes){
		vhod_konteiner.inject();
		i=0;
		break;}}

Pickup «Receiving_replic»	Собирает пришедшие в блок «Buffer_replic» реплики в единый запрос	Условие: agent.number==ID;
Dropoff «Transfer_replic»	Передает запрос из «Receiving_replic» на выход системы	Условие: agent.v_replication==1

Данные примеры позволяют утверждать о возможности построения достаточно сложных моделей с использованием библиотеки готовых моделей и отдельных элементов. Модель, представленная на рисунке 4.7 может являться малой долей моделей реального крупномасштабного сервиса, сложность которого будет возрастать с увеличением функционала. Из чего возможно сделать вывод о необходимости совместной реализации математического и имитационного моделирования, что позволит производить расчет вероятностно-временных характеристик сервисов и услуг, для последующего прогнозирования их функциональных возможностей.

4.4 Выводы по главе 4

Разработанный алгоритм позволяет проводить расчет временных характеристик комбинированных телекоммуникационных услуг, с возможность использования методов аппроксимации для сетей систем произвольного вида G/G/1.

Так же был приведены примеры расчета транзакционной услуги с несколькими сценариями возможной реализации рабочих процессов обслуживания запроса, что наглядно показывает влияния моделирования на процесс разработки решений для веб-сервисов.

В дополнение к методам математического моделирования, были предложены методы интеграции имитационных моделей к оценке характеристик сложных сетей систем массового обслуживания. В рамках реализации данной работы была разработана библиотека имитационных моделей, для которой предложено решение для интеграции аппарата параллельной обработки запросов Fork-Join как обособленный элемент структуры, с последующей комбинацией в разнотипные масштабные модели.

ЗАКЛЮЧЕНИЕ

Разработка методов оценки качественных характеристик с целью прогнозирования возможностей на этапе проектирования и разработки современных телекоммуникационных систем в эпоху быстрого роста возможностей как самих систем, так и сетевой инфраструктуры на которых происходит их разработка, является актуальной и важной задачей, поставленной перед научным обществом.

В диссертационной работе получены следующие основные результаты, имеющие теоретическое и практическое значение:

1. Разработан подход к моделированию телекоммуникационных услуг с разделением их на услуги реализующие поддержку пользовательских сессий, рабочих процессов транзакционнцх услуг и интеграция их в комбинированные телекомуникационных услуги.

2. Разработаны математическая и имитационная модель сеансовой телекоммуникационной услуги, которые позволяют принимать обоснованные решения по изменению количетсвенных параметров системы, для улучшения качества предоставления услуг.

3. Разработаны модели и метод расчета временных характеристик рабгочих процессов транзакционных слабосвязанных услуг с внедрением математического аппарата, параллельной обработки запросов Fork-Join.

4. Предложенны методы аппроксимации для расчета временных характеристик систем произвольного вида G/G/1, и проведена оценка их применимости, для расчета последовательных сетей систем вида G/M/1. Основным результатом являются верхняя и нижняя границы значений временных характеристик для СеМО данного вида.

5. На основе разработанных моделей для сеансовых и транзакционных услуг, был синтезирован обобщенный алгоритм реализации метода расчета вероятностно-временных характеристик комбинированных телекоммуникационных систем.

6. Разработана комплекс имитационных моделей, структурные элементы которых возможно интегрировать для создания различных моделей существующих и разрабатываемых телекоммуникационных услуг

Результаты данной работы могут являться практическими рекомендациями и основой для принятия обоснованных решений при проектировании и повышении эффективности телекоммуникационных систем, что способствует повышению их производительности и улучшению качества обслуживания пользователей. Это имеет большое значение для различных областей, включая телекоммуникации, облачные сервисы, электронную коммерцию и многие другие, где качество и эффективность услуг являются ключевыми факторами успеха и конкурентоспособности.

СПИСОК ЛИТЕРАТУРЫ

1. *Shabani, I.* Design of modern distributed systems based on microservices architecture/ I.Shabani //International Journal of Advanced Computer Science and Applications. – 2021. – Т. 12. – №. 2

2. Zhu J. QoS-Aware Co-Scheduling for Distributed Long-Running Applications on Shared Clusters/ J. Zhu // IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 12, pp. 4818-4834

3. Меньшаева, А. А. Развитие цифровых двойников в российской промышленности / А. А. Меньшаева. // Молодой ученый. — 2021. — № 11 (353). — С. 25-27.

4. Medhi, Subhash & Bora, Abhijit & Bezboruah, Tulshi. (2017). Investigations on Evaluation of Some QoS Aspects of Service Oriented Computing System Based on Web Services. Sensors and Transducers. 209. 56-64.

5. Xiong, K. Web services performance modeling and analysis/ K. Xiong // International Symposium on High Capacity Optical Networks and Enabling Technologies, HONET - 2006. - pp. 1 – 6

6. Yu, T.The design of QoS broker algorithms for QoS-capable Web service/ T.Yu, K-J.Lin //4 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE’04) - 2004. - pp. 17- 24. 10.1109/EEE.2004.1287283.

7. Zheng, Z. Semi-Markov Models of Composite Web Services for their Performance, Reliability and Bottlenecks. / Z. Zheng , K. Trivedi , K. Qiu, R. Xia // IEEE Transactions on Services Computing. – 2016.- pp.1939-1374.

8. Cheung, L. A Study of Web Services Performance Prediction: A Client's Perspective. / L. Cheung, L. Golubchik, F. Sha // 9th Annual IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems.- Singapore, - 2011.- pp. 75-84.

9. Avritzer, A. A Quantitative Approach for the Assessment of Microservice Architecture Deployment Alternatives by Automated Performance

Testing. / A. Avritzer, V. Ferme, A. Janes, B. Russo, H. Schulz, A. van Hoorn. // Lecture Notes in Computer Science, - 2018 - pp. 159–174.

10. Guo, X. Graph-based trace analysis for microservice architecture understanding and problem diagnosis / X. Guo, X. Peng, H. Wang, W. Li, H. Jiang, D. Ding, T. Xie, L. Su // Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. – 2020 - pp.159-174

11. Макаров Д.А. Масштабирование веб-приложений/ Д.А. Макаров, А.Д. Шибанова // Теория и практика современной науки. - 2021- №1 (67) – 4 с.

12. Xiong, K. Web services performance modeling and analysis/ K. Xiong // International Symposium on High Capacity Optical Networks and Enabling Technologies, Charlotte. - 2006 - pp. 1-6,

13. ГОСТ Р 57392-2017/ISO/IEC TR 20000-10:2015. Информационные технологии. Управление услугами. Часть 10. Основные понятия и терминология. URL: https://docs.cntd.ru/document/1200144207 (дата обращения: 21.03.2023).

14. Dai W. Bridging Service-Oriented Architecture and IEC 61499 for Flexibility and Interoperability. / W. Dai, V. Vyatkin, S. Member, J. H. Christensen, and V. N. Dubinin // IEEE Transactions on Industrial Informatics - vol. 11 - no. 3 – 2015 - pp. 771–781

15. Воробьев А.И. Инфокоммуникация и архитектура цифровой экономики/ А.И. Воробьев, М.О. Колбанев // Наука России: Цели и задачи – 2017 - с. 13-17

16. Гольдштейн Б. С. Инфокоммуникационные сети и системы / Б. С. Гольдштейн - СПб.БХВ-Петербург – 2019 – vol.22(5) - 208 с.

17. Росляков А.В. СЕТЬ 2030: архитектура, технологии, услуги / А.В. Росляков – М.: ООО «ИКЦ «Колос-с», 2022 – 278 с.

18. Abdel Hakeem, S.A. Security Requirements and Challenges of 6G Technologies and Applications/ S.A. Abdel Hakeem, H.H. Hussein, H. Kim // Sensors – 2022 – 1969

19. Akhtar, M.W. The shift to 6G communications: vision and requirements. / M. W. Akhtar, S. A. Hassan, R. Ghaffar, H. Jung, S. Garg, M. S. Hossain // The shift to 6G communications: vision and requirements. Human-Centric Computing and Information Sciences - 2020 - vol.10:53 - 27 p.

20. Villegas D. An Analysis of Provisioning and Allocation Policies for Infrastructure-as-a-Service Clouds / A. Antoniou, S. Sadjadi, A. Iosup, // 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing – IEEE – 2012. - pp. 612–619.

21. Moens H. A Scalable Approach for Structuring Large-Scale Hierarchical Cloud Management Systems / H. Moens, F. De Turck // 9th International Conference on Network and Service Management (CNSM 2013) – IFIP – 2013 - pp. 1–8

22. Моргунов Е.П. Система поддержки принятия решений при исследовании эффективности сложных систем: принципы разработки, требования и архитектура/ Е.П. Моргунов // Сибирский аэрокосмический журнал – 2007 - №3. - с. 59-63.

23. Гарькина И.А., Единая концепция системного проектирвоания сложных систем / И.А. Гарькина, А.М.Данилов // Regional architecture and engineering – 2019 - №1 - с.119-123

24. Джарратано Д. Экспертные системы: принципы разработки и программирование 4-е излание. / Д. Джарратано, Г. Райли // Пер. с англ. - М. ООО "И.Д. Вильямс" - 2007 - 1152 с.

25. Breitgand D. On cost-aware monitoring for self-adaptive load sharing. / D. Breitgand, R. Cohen, A. Nahir, D. Raz // IEEE Journal on Selected Areas in Communications – 2010 - vol. 28 - no. 1 - pp. 70–83.

26. Lu Y. JoinIdle-Queue: A novel load balancing algorithm for dynamically scalable web services / Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg // Performance Evaluation – 2011- vol. 68, no. 11, pp. 1056–1071

27. Алзагир, А. А. Исследование качества обслуживания в сетях 5G и последующих поколений / А. А. Алзагир, А. И. Парамонов, А. Е. Кучерявый //

Электросвязь. – 2022. – № 6. – С. 2-7. – DOI 10.34832/ELSV.2022.31.6.001. – EDN XAWJRS

28. Кучерявый А.Е. Трехмерные многослойные гетерогенные сверхплотные сети / А. Е. Кучерявый, А. И. Парамонов, М. А. Маколкина [и др.] // Информационные технологии и телекоммуникации. – 2022. – Т. 10, - № 3. – С. 1-12.

29. Бородин А.С. Искусственный интеллект в сетях связи пятого и последующих поколений А.С. Бородин, А.Н. Волков, А.С.А Мутханна, А.Е. Кучерявый // Электросвязь. - 2021. - № 1.- С. 17-22.

30. Фук Х.Д. Анализ методов минимизации сетевой задержки в каналах связи при передаче трафика промышленного интернета вещей / Х. Д. Фук, А. А. Березкин, Д. Д. Парфенов [и др.] // Электросвязь. – 2023. – № 1. – С. 8-16.

31. Гольдштейн Б. С. Инфокоммуникационные сети и системы./ Б.С. Гольдштейн - Санкт-Петербург– 2019 – 208 с.

32. Фишман, Е. Б. Анализ алгоритмов обслуживания очередей в сетях с поддержкой "Качества обслуживания" (QoS) / Е. Б. Фишман // Качество. Инновации. Образование. – 2006. – № 6(22). – С. 63-71

33. Алиев Т.И. Основы моделирования дискретных систем. / Е.Б. Фишман – СПб: СПбГУ ИТМО, 2009. – 363 с

34. Щемелинин Д.А. Математические модели и методы мониторинга и прогнозирования состояния глобально распределенных вычислительных комплексов/ Д.А. Щемелинин // Труды учебных заведений связи. 2021 - 7(3): с. 73-78

35. Клейнрок, Л. Вычислительные системы с очередями /Л. Клейнрок / Под ред. Б.С. Цыбакова. - Москва: Мир, 1979. - 600 с

36. Беляев Ю. К. Основные направления исследований в теории массового обслуживания. / Ю. К. Беляев, Б. В. Гнёденко, И. Н. Коваленко // Тр. VI Всесоюзного совещания по теории вероятности и математической статистике, Вильнюс- 1962. - С. 341—357

37. Бочаров П.П., Б 72 Теория массового обслуживания./ П.П. Бочаров, А.В. Печинкин // Учебник. - М : Изд-во РУДН - 1995 - 529 с

38. Башарин Г.П. Анализ очередей в вычислительных сетях: Теория и методы расчета / Г.П. Башарин, П.П. Бочаров, Я.А. Коган - 1989. - 336 с

39. Ghosal A., Queues in Series /A. Ghosal // Royal Statistical Society: Series B (Methodological) - 2018. - pp. 491-496

40. Севастьянов Б. А., Предельные теоремы для ветвящихся случайных процессов специального вида. / Б. А. Севастьянов // Теория вероятностей и ее применения № 3, - 1957 - с. 339—348

41. Диткин В.А. Интегральные преобразования и операционное исчисление / В.А. Диткин, А.П., Прудников // М: ГИФМЛ - 1961. 524 с

42. ГОСТ Р 57392—2017 - Информационные технологии. Управление услугами. Часть 10. Основные понятия и терминология. Дата введения 01.01.2018. – М:Стандартинформ. 2018 - 25 с

43. Zimmermann O. Microservices tenets: Agile approach to service development and deployment," / O. Zimmermann //Comput. Sci. - Res. Dev., - 2017 - vol. 32, no. 3– 4, pp. 301–310

44. Птицына, Л. К. Анализ интеграции сервис-ориентированных средств в активных инфокоммуникационных средах / Л. К. Птицына, В. О. Веселов // Наукоемкие технологии в космических исследованиях Земли. – 2015. – Т. 7, № 2. – С. 42-47

45. Стоянченко, С. С. Simevent имитационная модель корпоративной информационной системы / С. С. Стоянченко // Технологические исследования: информационное обеспечение, алгоритм проведения, интерпретация результатов – Стерлитамак: 2020. – С. 79-83

46. Shanthikumar, J. G. A Unifying View of Hybrid Simulation/ J. G. Shanthikumar, R. G. Sargent //Analytic Models and Modeling. Operations Research – 1983.- 31(6), pp. 1030–1052.

47. Ardagna C. A. Scalability Patterns for Platform-as-a-Service / C. A. Ardagna, E. Damiani, F. Frati, D. Rebeccani and M. Ughetti, // 2012 IEEE Fifth

International Conference on Cloud Computing, Honolulu, HI, USA, - 2012 - pp. 718-725

48. Bitran, G. R. Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference. Management Science/ G. R. Bitran, D. Tirupati // Management Science - 1988 - no. 34(1), pp. 75–100

49. Iglehart D. L. Multiple channel queues in heavy traffic. II: Sequences, networks, and batches/ D. L. Iglehart, W. Whitt //Advances in Applied Probability. – 1970. – Т. 2. – №. 2. – pp. 355-369

50. Gaver D. P. Approximate models for processor utilization in multiprogrammed computer systems/ D. P. Gaver, G. S. Shedler //SIAM Journal on Computing. – 1973. – Т. 2. – №. 3. – pp. 183-192

51. Schweitzer P. J. The correction terms in approximate mean value analysis/ P. J.Schweitzer, A. Seidmann, S. Shalev-Oren //Operations Research Letters. – 1986. – Т. 4. – №. 5. – pp. 197-200

52. Schweitzer P.J. Approximate analysis of multiclass closed networks of queues/ P.J. Schweitzer // J. ACM. – 1981. – Т. 29. – №. 2.

53. Buzen J. P. Fundamental operational laws of computer system performance/ J. P. Buzen // Acta Informatica. – 1976. – Т. 7. – С. 167-182

54. Allen, A.O. Queueing Models of Computer Systems. / A.O. Allen // Computer, - 1980. – vol. 13. - pp.13-24

55. ITU-T E.800 Термины и определения, связанные с качеством обслуживания и производительностью сети, включая надежность. Утверждена 23 сентября 2008 года 2-й Исследовательской комиссией – 2008 – 32 р.

56. Астахова, Т. Н. Архитектура цифровой платформы / Т. Н. Астахова, М. О. Колбанев, А. А. Шамин // Перспективные направления развития отечественных информационных технологий : материалы IV межрегиональной научно-практической конференции, Севастополь, 18–22 сентября 2018 года - 2018. – С. 69-71

57. Martin R. C. Clean architecture. – 2017. – 320 p.

58. Бессерверная архитектура [Электронный ресурс] URL: https://learn.microsoft.com/ru-ru/dotnet/architecture/serverless/serverless-architecture (Дата обращения 16.04.2023)

59. Logesh S.K. Supply Chain Data Management for Web Services / S. K. Logesh, K. Kavin, S. A. Kumar, A. Manikandan, G. Sakthi, B. Adithya // 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India – 2022. - pp. 1584-1589

60. Homay A. A Survey: Microservices Architecture in Advanced Manufacturing Systems / A. Homay, A. Zoitl, M. D. Sousa, M. Wollschlaeger // 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), Helsinki, Finland – 2019 - pp. 1165-1168

61. Shadija D. Towards an understanding of microservices / D. Shadija, M. Rezai, R. Hill // 2017 23rd International Conference on Automation and Computing (ICAC), Huddersfield, UK - 2017 - pp. 1-6.

62. Raj V. Performance and complexity comparison of service oriented architecture and microservices architecture/ V. Raj, R. Sadam // International Journal of Communication Networks and Distributed Systems. – 2021. – Т. 27. – №. 1. – pp. 100-117.

63. Raj V. A service graph based extraction of microservices from monolith services of service-oriented architecture/ V. Raj, S. Ravichandra // Software: Practice and Experience. – 2022. – Т. 52. – №. 7. – pp. 1661-1678

64. Balalaie, A. Microservices migration patterns. / A. Balalaie, A. Heydarnoori, P. Jamshidi, D.A. Tamburri, T. Lynn // Software: Practice and Experience. – 2018.- 24 p.

65. Dragoni N. Microservices: yesterday, today, and tomorrow/ N. Dragoni, S. Giallorenzo, A. L. Lafuente, M. Mazzara, F. Montesi, R. Mustan, L. Sana // Present and Ulterior Software Engineering – 2017 – pp.17

66. State of the Art in Microservices [Электронный ресурс] URL: https://microxchg.io/2015/slides/01_08_AdrianCockcroftSlides.pdf (Дата обращения 13.04.2023).

67. Oyucu S. Web Service-Based Turkish Automatic Speech Recognition Platform / S. Oyucu, H. Polat, H. Sever // 2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey – 2020.- pp. 1-5.

68. Мелентьева, О. В. Моделирование интернет-банковской системы на основе микросервисной архитектуры в управлении маркетинговой деятельностью банков / О. В. Мелентьева // Инновационное развитие экономики. – 2020. – № 4-5(58-59). – С. 153-158

69. Miao K. A microservice-based big data analysis platform for online educational applications/ K. Miao, J. Li, W. Hong, M. Chen //Scientific Programming. – 2020. – Т. 2020. – С. 1-13

70. Wang S. Microservice Architecture for Embedded Systems / S. Wang, C. Du, J. Chen, Y. Zhang, M. Yang // 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Xi'an, China. – 2021. - pp. 544-549

71. Guo S. Crossover Service Fusion Approach Based on Microservice Architecture / S. Guo, C. Xu, S. Chen, X. Xue, Z. Feng and S. Chen // 2019 IEEE International Conference on Web Services (ICWS), Milan, Ital. – 2019.- pp. 237-241

72. Torkura, K. A. Leveraging Cloud Native Design Patterns for Security-as-a-Service Applications/ K. A. Torkura, M. I. H. Sukmana, F. Cheng, C. Meinel // 2017 IEEE International Conference on Smart Cloud (SmartCloud). New York, NY, USA. - 2017. - pp. 90-97

73. Khomh, F. Understanding the impact of cloud patterns on performance and energy consumption/ F. Khomh, S.A. Abtahizadeh // Journal of Systems and Software. – 2018. – v.141.- pp. 151–170

74. Hecht, G. An Empirical Study of the Impact of Cloud Patterns on Quality of Service (QoS). / G. Hecht, B. Jose-Scheidt, C. D. Figueiredo, N. Moha, F. Khomh // 2014 IEEE 6th International Conference on Cloud Computing Technology and Science. – 2014. – 6 p.

75. Daya S. Microservices from Theory to Practice Creating Applications in IBM Bluemix Using the Microservices Approach/ S. Daya, N.Van Duy, K. Eati, C. M Ferreira, D. Glozic, V. Gucer, M. Gupta, S. Joshi, V. Lampkin, M. Martins, S. Narain// International Business Machines Corporation -2015. - p. 170

76. Hassan, S. Microservice Ambients: An Architectural Meta-Modelling Approach for Microservice Granularity/ S. Hassan, N. Ali, R. Bahsoon // 2017 IEEE International Conference on Software Architecture (ICSA). – 2017 – pp. 1-10

77. Song, Z. Equivalence-Enhanced Microservice Workflow Orchestration to Efficiently Increase Reliability/ Z. Song, E. Tilevich // 2019 IEEE International Conference on Web Services (ICWS), Milan, Italy. - 2019. - pp. 426-433

78. Hiratsuka, N. Service Selection with Combinational Use of Functionally-Equivalent Services/ N. Hiratsuka, F. Ishikawa, S. Honiden // 2011 IEEE International Conference on Web Services. Washington, DC, USA. – 2011. - pp. 97-104

79. Фомин Д.С., Проблематика обработки транзакций при использовании микросервисной архитектуры/ Д.С. Фомин, А.В. Бальзамов // Известия вузов. Поволжский регион. Технические науки. - 2021 - №2 – 9 с

80. Лобанкова, Т. С. Проблемы согласованности данных в микросервисной архитектуре / Т. С. Лобанкова // Цифровая трансформация - шаг в будущее: Материалы II Международной научно-практической конференции молодых ученых, посвященной 100-летию Белорусского государственного университета – Минск: Белорусский государственный университет, 2021. – С. 167-170

81. Netflix-conductor [Электронный ресурс] URL: https://netflixtechblog.com/netflix-conductor-a-microservices-orchestrator-2e8d4771bf40 (Дата обращения 16.04.2023)

82. Mangu S. Robotic process automation approach/ S. Mangu // Technology. – 2020. – Т. 11. – №. 11. – С. 831-840.

83. Ruecker B. Practical Process Automation/ B. Ruecker – " O'Reilly Media, Inc." - 2021. – 265 p.

84. Tank A. Automate Your Busywork: Do Less, Achieve More, and Save Your Brain for the Big Stuff. – John Wiley & Sons,Wiley. - 2023 – 256 p.

85. Avritzer, A. Scalability Assessment of Microservice Architecture Deployment Configurations: A Domain-based Approach Leveraging Operational Profiles and Load Tests/ A. Avritzer, V. Ferme, A. Janes, B. Russo, A. Hoorn, H. Schulz, V. Rufino // Journal of Systems and Software. – 2020. – vol.165. – pp. 16

86. Клейнрок Л. Теория массового обслуживания / Л.Клейнрок - Пер. с англ. М.: Машиностроение. - 1979. - 432 с.

87. Зарубин А. А. Моделирование игровых многопользовательских сервисов / А. А. Зарубин, Н. М. Редругина // Вестник связи. - 2020. - № 8.- С. 11–16.

88. ETSI TR 103 559 Speech and multimedia Transmission Quality (STQ);

89. Best practices for robust network QoS benchmark testing and scoring (введен 08.2019) – 11 с.

90. 3GPP Technical Specification 23.107 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Quality of Service (QoS) concept and architecture (Дата введения 03.2022) – 42 с.

91. RFC 2119: Key words for use in RFCs to Indicate Requirement Levels/ S. Bradner// Harvard University (Дата введения 03.1997)

92. ГОСТ Р ИСО/МЭК 12207-2010 Информационная технология - Программное обеспечение для жизненного цикла процессов (Дата введения 01.03.2012) – М.: Стандартинформ, 2011. – 105 с.

93. Zheng Z. A Distributed Replication Strategy Evaluation and Selection Framework for Fault Tolerant Web Services. / Z. Zheng, M. Lyu // Proceedings of the IEEE International Conference on Web Services, ICWS. – 2008 - pp. 145-152

94. Suznjevic M. Modelling of Network Traffic for Multiplayer Role-Playing Games based on User Behaviour: PhD in Engineering / Suznjevic Mirko – M.2012 – с. 194

95. Medhi S. Investigations on evaluation of some QoS aspects of service oriented computing system based on web services/ S. Medhi, A. Bora, T. Bezboruah //Sensors & Transducers. – 2017. – Т. 209. – №. 2. – С. 56.

96. Zhuang X. Player Dynamics in Massively Multiplayer Online Games / X. Zhuang, A. Bharambe, J. Pang, S. Seshan // School of Computer Science, Carnegie Mellon University, Pittsburgh, Tech. Rep. CMU-CS-07-158. – 2007 – p.30

97. Lee Y.-T. World of Warcraft avatar history dataset / Y.-T Lee, K.-T. Chen, Y.-M. Cheng, C.-L. Lei // Proceedings of the second annual ACM conference on Multimedia systems. - 2011. - PP. 123–128

98. Casalicchio E. A simulation framework for cluster-based Web services / E. A Casalicchio //International Journal of Simulation. – 2007. – Т. 8. – №. 4. – С. 21-33

99. Tarng P-Y. An analysis of WoW players' game hours/ P-Y. Tarng, K-T.Chen, P. Huang // roceedings of the 7th ACM SIGCOMM Workshop on Network and System Support for Games, NETGAMES 2008, Worcester, Massachusetts, USA. – 2008. - pp. 47-52.

100. Chen K-T. Game traffic analysis: an MMORPG perspective / K-T. Chen, P. Huang, C-L. Lei // Computer Networks. - 2006. - No. 50 - pp. 3002-3023

101. Borella M. S. Source models of network game traffic/ M. S. Borella // Computer communications. - 2000. - Т. 23. - No. 4. - pp. 403–410

102. Qiu T. Modeling user activities in a large IPTV system / T.Qiu, Z.Ge, D. Lee, J. Wang, Q.Zhao //Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement. – 2009. – С. 430-441

103. Jiang Z. Online-offline activities and game-playing behaviors of avatars in a massive multiplayer online role-playing game/ Z. Jiang, W. Zhou, Q.Z. Tan // EPL (Europhysics Letters). - 2009 - p. 88

104. Svoboda P. Traffic analysis and modeling for World of Warcraft / P. Svoboda, W. Karner, M. Rupp // Communications IEEE International Conference - 2007. - pp. 1612–1617

105. Kihl M. Analysis of World of Warcraft Traffic Patterns and User behavior/ M. Kihl, A. Aurelius, C. Lagerstedt // International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. – 2010 - pp. 218–223

106. Арнольд, Б. К. Одномерные и многомерные модели Парето./ Б. К. Арнольд // Приложение J Stat Distrib. - 2014 - №11 – 16 с.

107. Feldmann A. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models / A. Feldmann, W.Ward // Proceedings of INFOCOM. – 1997. - vol.3. - pp. 1096-1104

108. Daley, D. J. The Correlation Structure of the Output Process of Some Single Server Queueing Systems / D. J. Daley // The Annals of Mathematical Statistics 39,- 1968. - no. 3 - pp.1007–1019.

109. Urgaonkar B. An analytical model for multi-tier internet services and its applications./ B. Urgaonkar, G. Pacifici, P. Shenoy, M. Spreitzer, A. Tantawi // Sigmetrics Performance Evaluation Review - SIGMETRICS, - 2005 - №33(1). - pp. 291-302.

110. Chen H. Overload control in QoS-aware web servers / H. Chen, P. Mohapatra // Computer Networks, - 2003. - pp. 119-133

111. Cardoso J. Quality of service for workflows and web service processes / J. Cardoso, A. Sheth, J. Miller, J. Arnold, K. Kochut //Journal of web semantics. – 2004. – Т. 1. – №. 3. – pp. 281-308

112. Johnsson, M. Meeting latency targets for complex SOA-based transactions in data centers / M. Johnsson, B. Jennings // NOMS 2016 IEEE/IFIP Network Operations and Management Symposium - 2016 – 4 p

113. Chen Q. Dynamic-agents, Workflow and XML for E-Commerce Automation / Q. Chen, U. Dayal, M. Hsu, M. Griss // EC-Web. - 2000.- pp. 314-323

114. Shegalov G. XML-enabled workflow management for e-services across heterogeneous platforms/ G. Shegalov, M. Shegalov, G. Weikum // VLDB Journal – 2001. - №10. – pp. 91-103

115. Skogan D. Web service composition in UML / D. Skogan, R. Grønmo, I. Solheim // Proceedings. Eighth IEEE International Enterprise Distributed Object Computing Conference. - 2004. – pp. 47-57

116. Kochut, K.J. Optimizing workflow. / K.J. Kochut, A. Sheth, J. Miller// Component Strategies. – 1999. - №1.- pp. 45-57

117. Cardoso J. Semantic e-workflow composition / J.Cardoso, A. Sheth //Journal of Intelligent Information Systems – 2003. – pp. 191-225

118. Huiyuan Z. QoS Analysis and Service Selection for Composite Services/ Z. Huiyuan, Y. Jian, Z. Weiliang // Proceedings of the IEEE International Conference of Services Computing. - 2010. - pp. 122-129

119. Nabil A. Modeling and Analysis of Workflows Using Petri Nets / A. Nabil, A. Vijayalakshmi, H. Wei-kuang //Journal of Intelligent Information Systems. – 1998. - №10. pp.131-158.

120. Hiratsuka N. Service Selection with Combinational Use of Functionally-Equivalent Services. / Hiratsuka, N., Ishikawa, F., Honiden, S. // Proceedings - 2011 IEEE 9th International Conference on Web Services, ICWS - 2011. - pp. 97-104

121. Gorbunova, 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. / A.V. Gorbunova, A.V Lebedev // Distributed Computer and Communication Networks. DCCN 2021. Communications in Computer and Information Science – 2022.- pp. 318-332

122. Li B. A user-oriented web service reliability model / B. Li, Z. Su, Y. Zhou, X. Gong // Systems, Man and Cybernetics, IEEE International Conference on – 2008.- pp. 3612–3617

123. Xia Y. Stochastic modeling and quality evaluation of workflow systems based on qwf-nets / Y. Xia, H. Wang, C. Xu, L. Li // Computational Science, 6th International Conference on. Springer – 2006. - pp. 988–995

124. Varma S. Interpolation approximations for symmetric Fork-Join queues/ S. Varma, A. M. Makowski // Performance Evaluation. – 1994. – Vol. 20, № 1–3. – pp. 245–265.

125. Thomasian A. Approximate Solutions for $M/G/1$ Fork/Join Synchronization / A. Thomasian, A. N. Tantawi // Proceedings of the 26th Conference on Winter Simulation. – Orlando : Society for Computer Simulation International. - 1994. – pp. 361–368

126. Murata T. Petri Nets: Properties, Analysis and Applications / T. Murata // Proceedings of the IEEE, - - Vol. 77- No. 4 - April 1989 - pp. 541-580

127. Горбунова А.В. Обзор систем параллельной обработки заявок. Часть 1/ И.С. Зарядов, К.Е. Самуйлов, Э.С. Сопин // Вестник Российского университета дружбы народов. Серия: Математика, информатика, физика. - 2017.- Т. 25. - No 4 - С. 350–362

128. Программный модуль для расчета характеристик систем параллельной обработки транзакций, разработанный в среде Anylogic Зарубин А.А., Редругина Н.М. Свидетельство о регистрации программы для ЭВМ 2023618304, 21.04.2023. Заявка № 2023616789 от 07.04.2023

129. Чернова А. А. Методы согласования изменений в распределённых системах / А. А. Чернова, К. А. Курицын // Системный анализ и логистика: журнал.: выпуск №1(31), ISSN 2007-5687. – СПб.: ГУАП., 2022 – с.38-43

130. Колмогоров А.Н. Элементы теории функций и функционального анализа / А.Н. Колмогоров, С.В. Фомин - М.: Наука, 1968 - 496 с.

131. Вишневский В.М. Теоретические основы проектирования компьютерных сетей/ В.М. Вишневский - Москва: Техносфера, 2003. – 512 с.

132. Daley, D. J. The Correlation Structure of the Output Process of Some Single Server Queueing Systems / D. J. Daley // The Annals of Mathematical Statistics 39,- 1968. - no. 3 - pp.1007–1019.

133. Shore, J. Information theoretic approximations for M/G/1 and G/G/1 queuing systems / J. Shore // Acta Informatica - 1982. - 17(1) – pp.43-61

134. Chhibber, A. Security analysis of cloud computing.” / A. Chhibber, Sunil Batra //International Journal of Advanced Research in Engineering and Applied Sciences 2.3 (2013): 2278-6252

135. Dey, A.K. Discriminating Among the Log-Normal, Weibull, and Generalized Exponential Distributions/ A.K Dey, D. Kundu// Reliability, IEEE Transactions on Reliability. – 2009 - №58. - pp. 416 - 424.

136. Cox, D.R. Tests of separate families of hypotheses / D.R. Cox // Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability, Berkeley, University of California Press. - 1961. - №32 – pp.323-344.

137. Wiens, B.L. (1999), When log-normal and gamma models give different results: acase study / B.L. Wiens //American Statistician- 1999. - vol. 53 - pp. 89-93

138. Gupta, R. D. Discriminating between Weibull and generalized exponential distributions/ R.D. Gupta, K. Debasis Kundu //Computational Statistics & Data Analysis - 2003. - № 43(2) - pp. 179-196

139. Aryal, G. R. Transmuted Log-Logistic Distribution / G.R. Aryal //Journal of Statistics Applications & Probability 2 - 2013.- pp. 11-20

140. Verhulst P. F. Notice sur la loi que la population suit dans son accroissement / P. F. Verhulst. // Corr. Math. Physics,- 1975 - №10 - pp. 113-121

141. Zhmylev S. Analytical methods of nonstationary processes modeling / S. Zhmylev, I. Martynchuk, V. Kireev, T. Aliev// Majorov International Conference on Software Engineering and Computer Systems - MICSECS. – 2018. – 11 p

142. Tongdan J. Weibull and Gamma Renewal Approximation Using Generalized Exponential Functions, Communications in Statistics / J. Tongdan, G. Lakshmana // Simulation and Computation. – 2008. - №38. - pp. 154-171

143. Блатов И.А. Аппроксимация произвольной плотности распределения суммами экспонент / И. А. Блатов, В. Г. Карташевский, Н. В. Киреева, Л. Р. Чупахина // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2013. – № 2. – С. 53-57

144. Feldmann A. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models/ A.Feldmann, W. Whitt //Performance evaluation. – 1998. – Т. 31. – №. 3-4. –pp. 245-279.

145. Киреева Н.В. Сравнение характеристик систем массового обслуживания M/M/1 И G/G/1 спектральным методом/ Н.В. Киреева, Л.Р. Чупахина // Современные наукоемкие технологии. – 2014. – № 12-2. – С. 160-164

ПРИЛОЖЕНИЕ А. АКТЫ О ВНЕДРЕНИИ РЕЗУЛЬТАТОВ ДИССЕРТАЦИИ.

SIGURD · IT
телекоммуникационная IT-компания

sigurd-it.ru
sales@sigurd-it.ru
+7(812)449-47-32
194044, г. Санкт-Петербург, ул.Гельсингфорсская, д.4, к.1, Литера В, пом.16-Н, оф.28, БЦ «Красная Заря»

ООО «СИГУРД-АйТи»
ИНН 7802695166, КПП 780201001

АКТ
о внедрении научных результатов,
полученных в диссертации Редругиной Наталии Михайловны

Комиссия в составе:
- Генеральный директор – Потишний Р.В.
- Технический директор – Сергеева М.Ю.
- Заместитель технического директора – Колесникова Т.Я.

составила настоящий акт о том, что научные результаты, полученные Редругиной Наталией Михайловной в диссертации «Модели и методы оценки характеристик телекоммуникационных комбинированных слабосвязанных услуг», а именно:

— Модели расчета временных характеристик многопользовательских сеансовых слабосвязанных услуг, для прогнозирования качества обслуживания, что позволяет определить эффективные границы использования телекоммуникационных систем.

— Библиотека имитационных моделей, структурные элементы которых возможно интегрировать для создания различных модели существующих и разрабатываемых телекоммуникационных услуг

— Методика применения разработанных моделей для расчета вероятностно-временных характеристик комбинированных слабосвязанных услуг.

использованы в разработке программного комплекса «Контактный центр SIGURD-MIND для Непальского оператора связи NepalTelecom».

Разработанные модели оценки качественных характеристик слабоструктурированных сервисов позволяют эффективно и обоснованно принимать решения для управления количественными и качественными показателями работы операторов, Разработанные имитационные модели, интегрированные в рамках проекта, позволяют проводить постоянный контроль, прогнозировать и оценивать качественные и количественные характеристики функционирующих систем, с низким уровнем влияния на качество восприятия услуги.

Члены комиссии:
- Генеральный директор – Потишний Р.В.
- Технический директор – Сергеева М.Ю.
- Заместитель технического директора – Колесникова Т.Я.

Генеральный директор
ООО «СИГУРД-АйТи» _____ Р.В. Потишний
М.П.

АРГУС
НТЦ

Общество с ограниченной ответственностью
«Научно-Технический Центр АРГУС»
Почтовый адрес: 197198, г. Санкт-Петербург,
ул. Красного Курсанта, д. 25, лит. Ж
Юридический адрес: 191028,
г. Санкт-Петербург, ул. Пестеля, д. 7
ИНН 7841364714
КПП 784101001
ОГРН 1077847466597
Тел.: +7-812-333-36-60
Факс: +7-812-333-36-59
E-mail: office@argustelecom.ru
URL: www.argustelecom.ru

Утверждаю
Генеральный директор
ООО «НТЦ АРГУС»
_____
А.Б. Гольдштейн

АКТ
о внедрении научных результатов
полученных в диссертации Редругиной Наталии Михайловны

Комиссия в составе:
- заместителя Генерального директора, старшего преподавателя СПбГУТ – Феноменова М.А.
- руководителя направления – Петровского Н.О.
- бизнес-аналитика, доцента СПбГУТ, к.т.н. – Кислякова С.В.

составила настоящий акт о том, что научные результаты, полученные Редругиной Наталией Михайловной в диссертации «Модели и методы оценки характеристик телекоммуникационных комбинированных слабосвязанных услуг», а именно:

— Модели расчета временных характеристик транзакционных инфокоммуникационных слабосвязанных услуг с внедрением математического аппарата параллельной обработки.

— Библиотека имитационных моделей, структурные элементы которых возможно интегрировать для создания различных моделей существующих и разрабатываемых инфокоммуникационных услуг.

— Методика расчета вероятностно-временных характеристик комбинированных инфокоммуникационных систем.

Разработанные модели расчета временных характеристик транзакционных инфокоммуникационных слабосвязанных услуг применены на этапе проектирования систем технического учета «Аргус NRI» и управления рабочей силой «Аргус WFM». Указанные системы спроектированы на основе компонентов открытой цифровой архитектуры (Open Digital Architecture). Перевод программных продуктов и решений НТЦ АРГУС на микро-сервисную архитектуру, поддерживаемую TM Forum Open Digital Architecture, привёл к необходимости решать задачи оптимального комбинирования микро-сервисов для случаев сквозной автоматизации бизнес-процессов. Для оптимизации решений для сквозной автоматизации услуг подключения клиента и восстановления повреждённой услуги применена библиотека имитационных моделей и методика расчета вероятностно-временных характеристик комбинированных инфокоммуникационных систем.

Члены комиссии:

Феноменов М.А. _____

Петровский Н.О _____

Кисляков С.В _____

Общество с ограниченной ответственностью
«Научно-технический центр СевенТест»
ИНН 7841364672 / КПП 784101001

АКТ
о внедрении научных результатов
полученных в диссертации Редругиной Н.М.
«Модели и методы оценки характеристик телекоммуникационных комбинированных слабосвязанных услуг»

Комиссия в составе:

Ехриель И.М. первый заместитель генерального директора, технический директор, к.т.н.

Рерле Р.Д. генеральный директор, к.т.н.

составила настоящий акт о том, что научные результаты, полученные Редругиной Наталией Михайловной в диссертации «Модели и методы оценки характеристик телекоммуникационных комбинированных слабосвязанных услуг», а именно:

— Комплекс имитационных и математических моделей оценки характеристик систем в слабосвязанной структуре.
— Методика применения разработанных моделей для расчета характеристик различных типов телекоммуникационных услуг для внедрения в основу подхода к проектированию и последующего прогнозирования функционирования систем.

использованы в разработке программного-аппаратного комплекса «Профит».

Разработанный комплекс моделей позволяет оценить характеристики системы на этапах принятия решений по реализации возможных конфигураций разработки архитектур. В свою очередь, предложенная методика позволила повысить эффективность разработки моделей систем на этапах планирования и проектирования, для прогнозирования функционирования систем с учетом реальных входных данных.

Первый заместитель генерального директора, технический директор, к.т.н. Ехриель И.М.

Генеральный директор, к.т.н. Рерле Р.Д.

Российская Федерация, 191028, Санкт-Петербург, ул. Пестеля, д.7
Почтовый адрес: Российская Федерация, 197198, Санкт-Петербург, ул. Красного Курсанта д.25 лит. Ж.
Телефон: +7 (812) 333-36-37, факс: +7 (812) 333-36-47

**МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ,
СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ТЕЛЕКОММУНИКАЦИЙ ИМ. ПРОФ. М.А. БОНЧ-БРУЕВИЧА»
(СПбГУТ)**

Санкт-Петербург

УТВЕРЖДАЮ
Первый проректор-проректор по
учебной работе
к.т.н., доцент _____ А.В. Абилов
«07» 09 2023 г.

Акт

о внедрении научных результатов,

полученных в диссертационной работе Редругиной Наталии Михайловны

«Модели и методы оценки характеристик телекоммуникационных

комбинированных слабосвязанных услуг».

Комиссия в составе декана факультета Инфокоммуникационных сетей и систем к.т.н., доцента Д.В. Окуневой, и.о. заведующего кафедрой инфокоммуникационных систем к.т.н., доцента В.С. Елагина и заведующей лабораторией кафедры инфокоммуникационных систем И.А.Лаюшка составила настоящий акт в том, что научные результаты, полученные Редругиной Наталией Михайловной в диссертации "Модели и методы оценки характеристик телекоммуникационных комбинированных слабосвязанных услуг», использованы:

1. При проведении практических занятий и лабораторных работ для бакалавров по дисциплине «Математические методы построения инфокоммуникационных сетей и систем» (Рабочая Программа регистрационный номер №_21.05/231-Д) раздел Программы:

— Качество обслуживания;

— Применение математических подходов к исследованию компьютерных сетей и систем;

— Задачи расчета производительности сети и систем;

— Примеры применения имитационного моделирования для расчета характеристик систем.

2. При проведении практических занятий и лабораторных работ для бакалавров по курсу «Теория телетрафика» (Рабочая Программа №_19.05/1783-Д), раздел Программы:

— Задачи прогнозирования при планировании сетей;

— Принципы выбора структуры сети;

— Задачи, связанные с качеством обслуживания.

3. При проведении практических занятий и лабораторных работ для бакалавров по дисциплине «Сети связи и системы коммутации» (Рабочая Программа регистрационный номер №_21.05/244-Д), раздел Программы:

— Переходные вероятности;

— Оценка надежности сетевых элементов;

— Обеспечение качества обслуживания в Интернет;

— Системы с ожиданием без потерь M/M/1, M/D/1;

— Понятие о качестве обслуживания.

В указанных дисциплинах используются следующие новые научные результаты, полученные Редругиной Наталией Михайловной в диссертационной работе:

- Комплекс имитационных и математических моделей оценки временных характеристик комбинированных слабосвязанных услуг.
- Методика проектирования с применением разработанных моделей инфокоммуникационных услуг

Декан факультета ИКСС,
к.т.н., доцент — Д.В. Окунева

И.о заведующего кафедрой ИКС,
к.т.н., доцент — В.С. Елагин

Заведующая лабораторией
кафедры ИКС — Н.А. Лаюшка