На правах рукописи

Кущазли Анна Ивановна

МОДЕЛИ МАССОВОГО ОБСЛУЖИВАНИЯ ДЛЯ АНАЛИЗА ЭФФЕКТИВНОСТИ МИГРАЦИИ СЕРВИСОВ В ГРАНИЧНЫХ ОБЛАЧНЫХ ВЫЧИСЛЕНИЯХ

1.2.3 – Теоретическая информатика, кибернетика

Автореферат

диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2026

Работа выполнена на кафедре теории вероятностей и кибербезопасности Федерального государственного автономного образовательного учреждения высшего образования «Российский университет дружбы народов имени Патриса Лумумбы»

Научный руководитель: доктор физико-математических наук, доцент, доцент кафедры теории вероятностей и кибербезопасности факультета физико-математических и естественных наук Российского университета дружбы народов имени Патриса Лумумбы (РУДН)
Кочеткова Ирина Андреевна

Официальные оппоненты: доктор технических наук, профессор, заведующий кафедрой интеллектуальных сетевых и облачных технологий факультета сетевой инженерии Московского технического университета связи и информатики (МТУСИ)
Степанов Сергей Николаевич,

доктор физико-математических наук, ведущий научный сотрудник, заведующий лабораторией телекоммуникационных систем Института проблем управления им. В.А. Трапезникова Российской академии наук (ИПУ РАН)
Семенова Ольга Валерьевна,

кандидат физико-математических наук, доцент кафедры прикладной информатики института прикладной математики и компьютерных наук Национального исследовательского Томского государственного университета (ТГУ)
Лапатин Иван Леонидович

Защита диссертации состоится «19» июня 2026 г. в 13 час. 00 мин. на заседании диссертационного совета ПДС 0200.006 при Российском университете дружбы народов имени Патриса Лумумбы по адресу: г. Москва, ул. Орджоникидзе, д. 3, ауд. 208.
С диссертацией можно ознакомиться в Научной библиотеке Российского университета дружбы народов имени Патриса Лумумбы по адресу: 117198, Москва, ул. Миклухо-Маклая, дом 6.
Автореферат разослан «___» _____ 2026 г.

Ученый секретарь диссертационного совета
ПДС 0200.006, к.ф.-м.н., доцент М.Н. Геворкян

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования.

Развитие сетей пятого и шестого поколений (5G/6G) сопровождается стремительным ростом трафика иммерсивных приложений – дополненной и виртуальной реальности, облачного гейминга, телеприсутствия – для которых ключевым требованием является обеспечение малой межконцевой задержки (end-to-end delay, E2E-delay) и гарантированного качества обслуживания (quality of service, QoS). Централизованные облачные вычисления не всегда способны выполнить эти требования ввиду значительного расстояния между пользователем и удаленным дата-центром. Это обусловило широкое внедрение архитектуры периферийных, граничных вычислений с множественным доступом (multi-access edge computing, MEC), предполагающей размещение вычислительных узлов вблизи конечных пользователей на границе (edge) сети радиодоступа, что позволяет существенно снизить E2E-задержку по сравнению с облачным сервером.

В облачной инфраструктуре единицей выделения вычислительных ресурсов является виртуальная машина (ВМ) – программная среда, изолированно выполняющая задачи пользователей на физическом сервере. В гранично-облачной архитектуре аналогичным объектом является сервис – прикладная функциональность (например, облачный гейминг или потоковое видео), обслуживающая запросы пользователей беспроводной сети на ресурсах MEC-узла и/или облачного сервера. Таким образом, виртуальная машина и сервис являются двумя самостоятельными объектами миграции: в облачной инфраструктуре миграция – это перенос ВМ вместе с выполняемыми задачами между физическими серверами; в гранично-облачной архитектуре – перераспределение пользователей сервиса между MEC-узлом и облачным сервером.

Политика миграции определяется тремя компонентами: целевой функцией (критерием оптимальности – например, минимизацией занятой пропускной способности серверов или суммарной E2E-задержки), алгоритмом принятия решения (правилом перемещения ВМ или пользователей) и моментом принятия решения (событием, инициирующим проверку необходимости миграции, – поступлением новой заявки, завершением обслуживания или сменой фазы входного потока). Выбор допустимых моментов принятия решения существенно влияет на эффективность системы. В существующих моделях миграции ограничения на такие моменты не формализованы в рамках теории массового обслуживания: данная проблема, как правило, решается введением двойных порогов загрузки, задержками между миграциями или предсказанием нагрузки.

Модели миграции сервисов в MEC-системах формулировались, как правило, для одного сервиса с пороговой политикой по числу пользователей либо как задача марковского принятия решений и решались численно методом итерации значений без получения аналитического решения. Случай нескольких конкурирующих сервисов, одновременно разделяющих ресурсы MEC-узла, с аналитически вычислимым стационарным распределением в мультипликативном виде в литературе не рассматривался. Помимо этого, входной поток заявок в большинстве моделей описывался простейшим

пуассоновским процессом. Вместе с тем реальный сетевой трафик носит коррелированный характер. Адекватной моделью такого трафика является марковски-модулированный пуассоновский процесс (Markov-modulated Poisson process, MMPP) – интерпретируемый частный случай марковского потока (MAP), удобный для описания трафика с переключаемой интенсивностью. Однако адаптация политики миграции к фазе MMPP-потока, а также оценивание параметров потока по реальным данным в контексте задач миграции в существующих исследованиях не рассматривались.

Изложенное определяет научную проблему настоящего исследования: разработку математических моделей граничных облачных вычислений, включающих механизмы миграции виртуальных машин и сервисов с формализацией допустимых моментов принятия решения, получением политик миграции и расчетом показателей качества обслуживания.

Степень разработанности темы исследования. Значительный вклад в развитие данной тематики внесли следующие российские и зарубежные ученые и исследователи. В области методов анализа показателей эффективности сетей связи, включая облачные вычисления и граничных облачных вычислений: Андреев С.Д., Барабанова Е.А., Бегишев В.О., Волков А.Н., Вытовтов К.А., Гольдштейн Б.С., Киричек Р.В., Крук Е.А., Кучерявый А.Е., Кучерявый Е.А., Кулябов Д.С., Ляхов А.И., Маколкина М.А., Молчанов Д.А., Мутханна А.С.А., Нетес В.А., Орлов Ю.Н., Парамонов А.И., Пшеничников А.П., Росляков А.В., Смелянский Р.Л., Хакимов А.А., Хоров Е.М., Яновский Г.Г., Ateya A.A., Buyya R., Correia L.M., Dustdar S., Shi W., Taleb T., Wang S. и др. Отметим ученых, которые внесли значительный вклад в развитие методов математической теории телетрафика и теории массового обслуживания: Башарин Г.П., Бочаров П.П., Гайдамака Ю.В., Горцев А.М., Ефросинин Д.В., Зейфман А.И., Зорин А.В., Ибрагимов Б.Г., Ивницкий В.А., Карташевский В.Г., Лапатин И.Л., Меликов А.З., Пауль С.В., Печинкин А.В., Разумчик Р.В., Рыков В.В., Самуйлов К.Е., Сатин Я.А., Соколов Н.А., Сопин Э.С., Степанов С.Н., Терпугов А.Ф., Тюрликов А.М., Фархадов М.П., Федоткин М.А., Цитович И.И., Цициашвили Г.Ш., Шнепс М.А. и др., включая по моделям с коррелированными входными потоками: Вишневский В.М., Дудин А.Н., Клименок В.И., Меликов А.З., Моисеев А.Н., Моисеева С.П., Морозов Е.В., Назаров А.А., Наумов В.А., Нежельская Л.А., Пауль С.В., Румянцев А.С., Семенова О.В., Chakravarthy S.R., Fischer W., Latouche G., Lucantoni D.M., Meier-Hellstern K.S., Neuts M.F., Ramaswami V. и др.

Цель исследования состоит в разработке моделей миграции виртуальных машин в облачной инфраструктуре и сервисов в гранично-облачной архитектуре для анализа и расчета вероятностно-временных показателей качества обслуживания пользователей и эффективности миграции.

Достижение сформулированной цели достигается путем решения следующих **задач исследования**:

1. Разработка моделей миграции виртуальных машин в облачной инфраструктуре и сервисов между граничным и облачными серверами в виде систем массового обслуживания с политиками миграции на основе

критериев занятой пропускной способности и суммарной межконцевой задержке и принятием решения в моменты изменения состояний системы.

2. Разработка алгоритмов для анализа и расчета показателей эффективности миграции, в том числе вероятности миграции, средней высвобождаемой пропускной способности сервера, средней суммарной межконцевой задержки, с учетом коррелированного характера входного потока заявок.

Научная новизна результатов исследования состоит в следующем:

1. Модель миграции виртуальных машин в облачной инфраструктуре реализует перемещение всех заявок класса между группами приборов по алгоритмам, минимизирующим занятую пропускную способность серверов, с оценкой занятости приборов до и после размещения поступившей заявки. Решение о миграции принимается только в момент поступления новой заявки соответствующего класса. Ранее основным критерием являлась загрузка процессора и энергопотребление, без формализации ограничений на момент принятия решения в модели массового обслуживания – двойными порогами загрузки, задержками между миграциями, предсказанием нагрузки.
2. Модель миграции сервисов в гранично-облачной архитектуре реализует перемещение заявок между общей группой приборов с эксклюзивным обслуживанием одного класса и индивидуальными приборами по политике, минимизирующей суммарную межконцевую задержку. Оптимальное распределение заявок получено аналитически в виде функции от числа заявок в системе, а стационарное распределение вероятностей состояний имеет мультипликативный вид. Ранее задача миграции сервисов в MEC-системах формулировалась для одного сервиса с пороговой политикой по числу пользователей или как марковский процесс принятия решений и решалась численно методом итерации без получения аналитического решения в мультипликативном виде.
3. Модель миграции сервиса в гранично-облачной архитектуре учитывает коррелированный характер входного потока заявок пользователей в виде MMPP и реализует адаптивную политику миграции, зависящую от фазы потока: в фазе высокой интенсивности решение пересматривается при каждом событии, в фазе низкой интенсивности – только при поступлении заявки или смене фазы. Ранее модели миграции сервисов в MEC-системах исследовались с пуассоновским входным потоком и без адаптации политики миграции к фазе потока.

Теоретическая значимость работы определяется следующим. Доказаны утверждения и теоремы, позволяющие аналитически вычислять стационарные распределения вероятностей состояний моделей миграции: для модели миграции сервисов в гранично-облачной архитектуре с пуассоновским входным потоком – в мультипликативном виде, для модели с коррелированным потоком MMPP – на основе матричного рекуррентного алгоритма, использующего блочно-трехдиагональную структуру инфинитезимальной матрицы. Для каждой модели определены оптимальные политики миграции и получены формулы расчета показателей эффективности обслуживания. В рамках решаемых задач продуктивно применен комплекс методов математической теории телетрафика,

теории массового обслуживания, марковских случайных процессов, матричных аналитических методов и статистического оценивания параметров входного потока. Систематизированы объекты миграции (виртуальная машина, сервис), критерии принятия решения о миграции (по занятой пропускной способности и по суммарной межконцевой задержке), а также допустимые моменты принятия решения, включая адаптивную политику, зависящую от фазы MMPP-потока. Исследовано влияние коррелированного характера входного потока на показатели эффективности миграции сервисов в гранично-облачной архитектуре.

Практическая значимость работы состоит в следующем. Разработанные модели, алгоритмы и расчетные формулы могут быть применены сотовыми операторами и провайдерами облачных и граничных сервисов при проектировании и эксплуатации беспроводных сетей 5G/6G с поддержкой иммерсивных приложений. Для облачной инфраструктуры получены алгоритмы выбора сервера-назначения при миграции виртуальной машины и формулы расчета вероятности миграции и средней высвобождаемой пропускной способности, позволяющие оценить эффективность использования ресурсов серверов. Для гранично-облачной архитектуры разработана политика распределения пользователей между MEC-узлом и облачным сервером, минимизирующая суммарную E2E-задержку; получены аналитические выражения для ее расчета в зависимости от числа активных сервисов и загрузки MEC-узла. Матричный алгоритм расчета стационарного распределения для модели с MMPP-потоком позволяет учитывать коррелированный характер реального сетевого трафика: численный анализ выполнен с параметрами, оцененными по данным реального мобильного оператора для пяти типов сервисов, что подтверждает применимость результатов к практическим сценариям развертывания MEC.

Методология и методы исследования. Для решения поставленных задач использовались методы математической теории телетрафика, теории массового обслуживания, марковских случайных процессов, теории вероятностей и математической статистики.

Положения, выносимые на защиту.

1. Модель миграции виртуальных машин в облачной инфраструктуре по алгоритмам, минимизирующим занятую пропускную способность серверов, позволяет рассчитать показатели эффективности миграции, такие как вероятность миграции виртуальной машины и среднюю высвобождаемую пропускную способность сервера.
2. Модель миграции сервисов в гранично-облачной архитектуре по политике, минимизирующей суммарную межконцевую задержку по всем пользователям, и стационарное распределение в мультипликативном виде применимы для расчета показателей эффективности миграции, которые зависят от распределения пользователей между граничным и облачным серверами и размещенного на граничном узле сервиса.
3. Модель миграции сервиса в гранично-облачной архитектуре по политике, зависящей от фазы входного MMPP-потока, и матричный алгоритм

расчета стационарного распределения позволяют рассчитать среднюю суммарную межконцевую задержку по всем пользователям при коррелированном входном потоке заявок.

Степень достоверности результатов работы обеспечивается: использованием при доказательстве утверждений и теорем общепринятых методов теории массового обслуживания, марковских случайных процессов и матричных аналитических методов; опорой на анализ и обобщение актуального отечественного и зарубежного опыта в области миграции виртуальных машин в облачной инфраструктуре и миграции сервисов в гранично-облачной архитектуре; установленным качественным совпадением частных случаев разработанных моделей с известными моделями теории массового обслуживания — в частности, при отсутствии миграции модели сводятся к классическим многоканальным системам с потерями; согласованностью результатов численного анализа с физически ожидаемым поведением системы.

Апробация результатов работы.

Основные положения работы были апробированы на следующих конференциях: International Conference on Next Generation Wired/Wireless Networks and Systems, NEW2AN (2025: Абу-Даби, ОАЭ), International Conference on Information Technologies and Mathematical Modelling, ITMM (2024: Карши, Узбекистан), International Conference on Information, Control, and Communication Technologies, ICCT (2024: Владикавказ), International Conference on Computer-Aided Technologies in Applied Mathematics, ICAM (2024: Катунь), International Conference on Distributed Computer and Communication Networks, DCCN (2024: Москва, РУДН), международный молодежный научный форум «Ломоносов» (2025: Москва, МГУ), всероссийская конференция с международным участием «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем» ИТТММ (2022-2025: Москва, РУДН). Положения работы обсуждались также на научном межвузовском семинаре «Современные телекоммуникации и математическая теория телетрафика», проводимом РУДН, МТУСИ, ТГУ, ИПМ РАН (24.04.2026).

Автор является победителем конкурсного отбора на назначение стипендии Президента РФ для аспирантов и адъюнктов, обучающихся по очной форме обучения в российских организациях, осуществляющих образовательную деятельность, и проводящих научные исследования в рамках реализации приоритетов НТР РФ, определенных в СНТР РФ (2024-2025), а также победителем конкурсного отбора программы РУДН «Аспирантура полного дня» (2022-2025). В 2024 г. как стипендиат Президента РФ автор была приглашена и участвовала в Конгрессе молодых ученых на федеральной территории «Сириус».

Реализация результатов работы. Автор является исполнителем грантов системы грантовой поддержки научных проектов РУДН «Разработка моделей и алгоритмов нарезки радиоресурсов и приоритетного доступа в беспроводной сети 6G» (2023-2024), «Модели математической теории телетрафика для анализа приоритетного обслуживания потокового и эластичного трафика в сетях новых поколений» (2025-2026).

Публикации. Результаты исследования представлены в 12 публикациях, в том числе 3 опубликованы в рецензируемых научных изданиях Перечня ВАК РФ (К-1, К-2) / Перечня РУДН / МБЦ WoS, Scopus, получены 3 свидетельства о государственной регистрации программы для ЭВМ (К-3).

Соответствие паспорту специальности. Работа соответствует следующим пунктам паспорта научной специальности 1.2.3 «Теоретическая информатика, кибернетика»:

- п. 9 «Математическая теория исследования операций» в части моделей массового обслуживания для исследования политик управления миграцией виртуальных машин между облачными серверами и миграцией пользователей сервисов между граничным и облачными серверами;
- п. 11 «Распределенные многопользовательские системы» в части моделирования многосерверной облачной системы выполнения задач пользователей на виртуальных машинах и системы предоставления сервисов пользователям в распределенной беспроводной гранично-облачной архитектуре;
- п. 12 «Модели информационных процессов и структур» в части моделирования процесса выполнения задач пользователей в облачной инфраструктуре и процесса передачи коррелированного трафика в виде MMPP-потока в гранично-облачной архитектуре.

Личный вклад автора. Разработанные модели, программные средства и анализ моделей выполнены автором самостоятельно. Все результаты, вынесенные на защиту, получены лично автором..

Объем и структура работы. Текст диссертации включает в себя введение, основную часть из трех глав и заключение. Диссертация включает в себя список литературы из 145 библиографических ссылок. Работа изложена на 101 странице текста, содержит 37 рисунков и 13 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Введение содержит обоснование актуальности темы, анализ степени разработанности предметной области, цель и задачи работы. Отражены научная новизна, теоретическая и практическая значимость результатов, использованные методы и положения, выносимые на защиту. Приведены сведения о достоверности и апробации результатов, соответствии паспорту специальности, личном вкладе автора и основных публикациях.

В **главе 1** проведен аналитический обзор вопросов миграции сервисов в граничных облачных вычислениях и описаны методы представления сетевого трафика, используемые в главе 3. В разделе 1.1 рассмотрены архитектурные сценарии взаимодействия облачной инфраструктуры и гранично-облачной архитектуры с использованием технологии MEC: определены роли облачных серверов и MEC-узлов при предоставлении иммерсивных сервисов – дополненной и виртуальной реальности, облачного гейминга, телеприсутствия – в сетях 5G/6G. Систематизированы объекты, критерии и инициирующие события процедуры миграции. Сформулирована постановка задачи исследования.

В *разделе 1.2* решена задача классификации сетевого трафика по типам сервисов на основе номеров портов методами машинного обучения. Анализ выполнен на наборе данных реального мобильного оператора Vodafone, включающем 44 млн записей о 16 типах сервисов. Из рассмотренных классификаторов (kNN, Random Forest, XGBoost, LightGBM, CatBoost, LR, SVM, NB) наилучший результат по метрике macro-averaged F1 показал XGBoost (F1 = 0,97, Accuracy = 0,97). Классификация позволяет выделять трафик конкретных приложений и формировать агрегированные профили нагрузки для групп сервисов, используемые в последующих разделах.

В *разделе 1.3* профили трафика описаны с использованием моделей временных рядов. Для прогнозирования нисходящего трафика рассмотрены модели SARIMA и Holt-Winters; оценка качества проводилась по критерию MAPE. Модель Holt-Winters обеспечила MAPE ≈ 11,2%, модель SARIMA – MAPE ≈ 15%, что свидетельствует о предпочтительности первой для краткосрочного прогнозирования трафика с выраженными сезонными компонентами.

В *разделе 1.4* предложено моделирование сетевого трафика в виде марковского потока (MAP). Параметры MMPP- и MAP-потоков оцениваются по реальным данным трафика Vodafone методом максимального правдоподобия на основе EM-алгоритма с использованием пакета mapfit для языка R. Оцененные матрицы $\boldsymbol{D_0}$ и $\boldsymbol{D_1}$ для MMPP-потока служат входными данными для модели в разделах 3.4-3.6.

В главе 2 исследована модель миграции виртуальных машин в облачной инфраструктуре в виде системы массового обслуживания с перемещением заявок между группами приборов, соответствующих физическим серверам. Отличительная особенность модели состоит в том, что миграция реализуется как перенос всех заявок класса между группами приборов, а проверка необходимости миграции выполняется только в момент поступления новой заявки на соответствующую виртуальную машину. Сформулированы и сравниваются два алгоритма выбора сервера-назначения, различающихся моментом оценки занятой пропускной способности: до и после размещения поступившей заявки.

В разделе 2.1 рассматривается облачная вычислительная система, предназначенная для обслуживания потоков задач, поступающих от пользователей и выполняемых на виртуальных машинах. Пусть $\mathcal{S} = \{1, \dots, S\}$ множество серверов, $\mathcal{V} = \{1, \dots, V\}$ множество виртуальных машин. Каждый сервер $s \in \mathcal{S}$ обладает фиксированной пропускной способностью (максимальной загрузкой, ресурсом) C_s бит/с и трактуется как группа приборов. Каждая виртуальная машина $v \in \mathcal{V}$ обслуживает задачи одного типа: задачи поступают пуассоновским потоком с интенсивностью λ_v задача/с, время обслуживания одной задачи экспоненциально распределено с параметром μ_v 1/с, каждая задача требует ресурс b_v бит/с. Схема модели показана на рисунке 1.

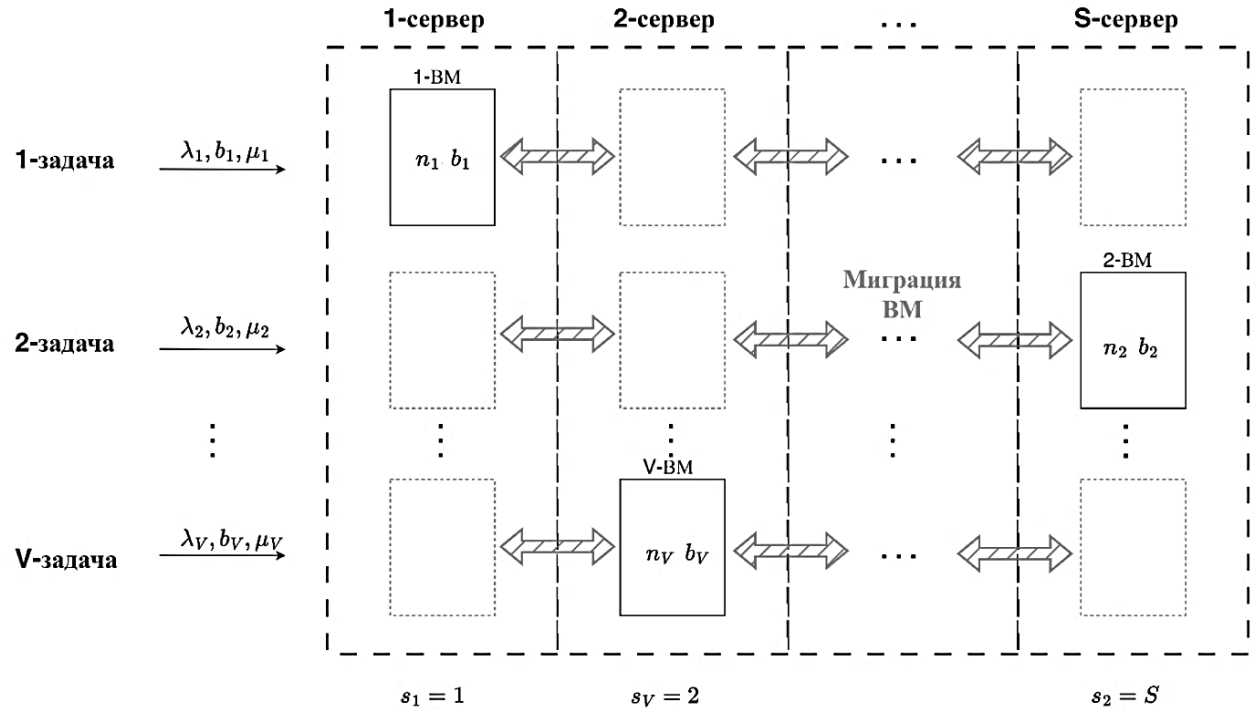

Рисунок 1 – Схема модели с перемещением заявок между группами приборов

Система описывается случайным процессом $\boldsymbol{X}(t)$, $t \geq 0$, где состояние $\boldsymbol{x} = (\boldsymbol{n}, \boldsymbol{s})$ определяется числом задач на виртуальных машинах $\boldsymbol{n} = (n_1, \dots, n_V)$ и номерами серверов $\boldsymbol{s} = (s_1, \dots, s_V)$, на которых размещены ВМ. Занятая пропускная способность сервера s в состоянии $\boldsymbol{x}$ равна

$$c_s(\boldsymbol{x}) = \sum_{v \in \mathcal{V}} n_v \; b_v \; \mathbf{1}\{s_v = s\}, s \in \mathcal{S}. \tag{1}$$

Пространство состояний системы:

$$\mathcal{X} = \begin{Bmatrix} \boldsymbol{x} = (\boldsymbol{n}, \boldsymbol{s}) \colon (n_v > 0, s_v \in \mathcal{S}) \vee \; (n_v = 0, s_v = 0), v \in \mathcal{V}; \\ c_s(\boldsymbol{x}) \leq C_s, s \in \mathcal{S} \end{Bmatrix}. \tag{2}$$

Для каждой ВМ $v \in \mathcal{V}$ введено подмножество состояний, в которых поступление новой задачи на ВМ v привело бы к превышению допустимой пропускной способности текущего сервера s_v

$$\mathcal{X}_v = \left\{\boldsymbol{x} \in \mathcal{X} \colon c_{s_v}(\boldsymbol{x}) + b_v > C_{s_v}\right\}, v \in \mathcal{V}. \tag{3}$$

Для каждого состояния $\boldsymbol{x} \in \mathcal{X}_v$, в котором миграция возможна, определяется целевой сервер $s_v^*(\boldsymbol{x}) \in \mathcal{S}_v(\boldsymbol{x})$ в соответствии с заданной политикой миграции, где $\mathcal{S}_v(\boldsymbol{x})$ множество допустимых серверов-назначений для ВМ v в состоянии $\boldsymbol{x}$.

В разделе 2.2 описана задача оптимизации, которая состоит в выборе сервера-назначения $s_v^*(\boldsymbol{x})$ при поступлении новой задачи на ВМ v в состоянии $\boldsymbol{x} \in \mathcal{X}_v$, так, чтобы минимизировать суммарную занятую пропускную способность серверов. Сформулированы два алгоритма миграции, различающихся моментом оценки занятости приборов. В алгоритме 1 целевой сервер выбирается по минимуму занятой пропускной способности после размещения поступившей задачи на ВМ v на сервере

$$s_v^*(\boldsymbol{x}) = \arg \min_{s' \in \mathcal{S}_v(\boldsymbol{x})} \left(c_{s'}(\boldsymbol{x}) + b_v \mathbf{1}\{s_v \neq s'\}\right), v \in \mathcal{V}, \boldsymbol{x} \in \mathcal{X}. \tag{4}$$

В алгоритме 2 целевой сервер выбирается по минимуму текущей занятой пропускной способности до учета поступившей задачи. В обоих алгоритмах

проверка необходимости миграции выполняется исключительно в момент поступления новой задачи на соответствующую ВМ.

В разделе 2.3 рассмотрен частный случай политики миграции, при которой решение о переносе ВМ v принимается в зависимости от текущего числа задач на серверах. Переход ВМ v инициируется, если число задач на текущем сервере s_v превышает заданный порог, а на сервере-назначении число задач ниже соответствующего порога.

В разделе 2.4 на основе стационарного распределения $\pi(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}$ получены формулы для следующих показателей эффективности: вероятность блокировки задач, среднее использование полосы пропускания серверов и ВМ, среднее число работающих серверов, вероятность миграции ВМ, средняя высвобождаемая пропускная способность сервера. Для вероятности миграции доказано следующее утверждение.

Утверждение 1. Для модели с перемещением заявок между группами приборов вероятность миграции произвольной ВМ равна

$$P^{mg} = \sum_{\boldsymbol{x} \in \mathcal{X}} P^{mg}(\boldsymbol{x}) \cdot \pi(\boldsymbol{x}), \tag{5}$$

где условная вероятность миграции любой ВМ в состоянии $\boldsymbol{x}$:

$$P^{mg}(\boldsymbol{x}) = \sum_{v \in \mathcal{V}} P_v^{mg}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}, \tag{6}$$

а вероятность миграции ВМ v в состоянии $\boldsymbol{x} = (\boldsymbol{n}, \boldsymbol{s})$

$$P_v^{mg}(\boldsymbol{x}) = \frac{\lambda_v}{\sum_{v' \in \mathcal{V}} (\lambda_{v'} + n_{v'}\mu_{v'})} \cdot \mathbf{1}\{\boldsymbol{x} \in \mathcal{M}_v\}, v \in \mathcal{V}, \boldsymbol{x} \in \mathcal{X}, \tag{7}$$

где $\mathcal{M}_v$ множество состояний, инициирующих миграцию ВМ v после поступления новой задачи

$$\mathcal{M}_v = \{\boldsymbol{x} \in \mathcal{X}_s : \mathcal{S}_v(\boldsymbol{x}) \neq \emptyset\}, v \in \mathcal{V}. \tag{8}$$

Средняя высвобождаемая пропускная способность сервера источника при миграции ВМ v в состоянии $\boldsymbol{x}$ определяется как разность занятой пропускной способности сервера s_v до и после миграции и вычисляется аналогично через стационарное распределение $\pi(\boldsymbol{x})$.

В разделе 2.5 проведен численный анализ модели для сценария обслуживания иммерсивных сервисов (AR/VR, облачный гейминг). Рассмотрены наборы параметров, отражающие различную интенсивность поступления задач и характеристики серверов. Алгоритм 1 обеспечивает более равномерное распределение нагрузки между серверами при высокой интенсивности входного потока.

В главе 3 исследованы модели миграции сервисов в гранично-облачной архитектуре. В отличие от главы 2, объектом миграции здесь является не виртуальная машина, а сервис: пользователи перераспределяются между MEC-узлом с эксклюзивным обслуживанием одного класса и облачным сервером с индивидуальными приборами. Глава содержит два самостоятельных результата, различающихся моделью входного потока: пуассоновский поток (разделы 3.1–3.3) и коррелированный MMPP-поток (разделы 3.4–3.6).

В разделах 3.1–3.3 рассматривается модель миграции сервисов с пуассоновским входным потоком. Оптимальная политика миграции получена аналитически в виде функции от числа заявок в системе, а стационарное

распределение при данной политике имеет мультипликативный вид, что позволяет вычислять все показатели эффективности непосредственно по аналитическим формулам.

В разделе 3.1 рассматривается K конкурирующих сервисов ($\mathcal{K} = \{1, \dots, K\}$). Для каждого сервиса $k \in \mathcal{K}$ входной поток пуассоновский с интенсивностью λ_k, время обслуживания экспоненциально с параметром μ_k, каждая заявка требует полосу пропускания b_k бит/с. Параметры ресурсных ограничений $M_s = \left\lfloor \frac{C_0}{b_s} \right\rfloor, N_k = \left\lfloor \frac{C_k}{b_k} \right\rfloor$, где C_0 пропускная способность канала к MEC-узлу; C_k – к облачному серверу для сервиса k; M_s максимальное число пользователей сервиса s на MEC-узле, N_k максимальное число пользователей сервиса k в облаке. Задержки следующие: d_0 на MEC-узле, d_k в облаке для сервиса k; предполагается $d_0 < d_k$ для всех $k \in \mathcal{K}$. Предполагается, что MEC-узел в каждый момент времени может обслуживать заявки не более одного сервиса класса s (эксклюзивное использование ресурса).

Случайный процесс $\boldsymbol{X}(t), t \geq 0$ имеет состояние $\boldsymbol{x} = (\boldsymbol{n}, m, s)$, где $\boldsymbol{n} = (n_1, \dots, n_K)$ числа заявок по сервисам, m число заявок на MEC-узле, s индекс сервиса, закрепленного на MEC-узле ($s = 0$ означает пустой MEC-узел). Пространство состояний $\mathcal{X} \subseteq \widetilde{\mathcal{X}}$:

$$\begin{gathered}\widetilde{\mathcal{X}} = \{(\mathbf{0}, 0,0)\} \cup \\ \cup \{(\boldsymbol{n}, m, s)\colon n_s = m,\ 0 < m < M_s,\ 0 \leq n_k \leq N_k, k \in \mathcal{K} \setminus \{s\},\ s \in \mathcal{K}\} \cup \\ \cup \{(\boldsymbol{n}, M_s, s)\colon M_s \leq n_s \leq N_s,\ 0 \leq n_k \leq N_k, k \in \mathcal{K} \setminus \{s\},\ s \in \mathcal{K}\}.\end{gathered} \quad (9)$$

Суммарная E2E-задержка в состоянии $(\boldsymbol{n}, m, s)$ равна

$$d(\boldsymbol{n}, m, s) = \sum_{k \in \mathcal{K}} n_k\, d_k \ + \ m\,(d_0 - d_s). \quad (10)$$

Далее в *разделе 3.2* обсуждается задача оптимизации, которая состоит в выборе в каждом состоянии $\boldsymbol{x}$ и при каждом событии A_{ik} допустимого действия (m', s'), минимизирующего суммарную E2E-задержку:

$$(m^*(\boldsymbol{n}, m, s, i, k), s^*(\boldsymbol{n}, m, s, i, k)) = \arg \min_{(m', s') \in \mathcal{A}_{ik}(\boldsymbol{x})} d(\boldsymbol{n}', m', s'), \quad (11)$$

где $\mathcal{A}_{ik}(\boldsymbol{x})$ множество допустимых действий в состоянии $\boldsymbol{x}$ при наступлении события A_{ik} ($i = 1$ поступление заявки; $i = 2$ завершение обслуживания, $k \in \mathcal{K}$). Структура множества $\mathcal{A}_{ik}(\boldsymbol{x})$ формально описана в Лемме 1.

Лемма 1. Для модели с перемещением заявок между общей и индивидуальными группами приборов, множество допустимых действий при наступлении события A_{ik} в состоянии $\boldsymbol{x}$ имеет вид

$$\mathcal{A}_{ik}(\boldsymbol{x}) = \begin{cases} \{(1, k)\}, s = 0, k \in \mathcal{K}, i = 1, \\ \{(\min(n_s + 1, M_s), s), (\min(n_j, M_j), j)\}, j \in \mathcal{K}\backslash\{s\}, s \neq 0, k = s, i = 1, \\ \{(\min(n_k + 1, M_k), k), (\min(n_s, M_s), s)\}, k \in \mathcal{K}\backslash\{s\}, s \neq 0, i = 1, \\ \{(\min(n_s - 1, M_s), s), (\min(n_j, M_j), j)\}, j \in \mathcal{K}\backslash\{s\}, s \neq 0, k = s, m > 0, i = 2, \\ \{(\min(n_s, M_s), s)\}, s \neq 0, k = s, n_s > 0, i = 2. \end{cases} \quad (12)$$

Теорема 1. Для модели с перемещением заявок между общей и индивидуальными группами приборов, оптимальная политика минимизации суммарной E2E-задержки определяется явно для каждого события A_{ik} в состоянии $\boldsymbol{x} = (\boldsymbol{n}, m, s)$ следующим образом

$$\big(m^*(\boldsymbol{n},m,s,i,k),s^*(\boldsymbol{n},m,s,i,k)\big)=$$
$$=\begin{cases}(1,k), & i=1,s=0,\\ (m+1,k), & i=1,s\neq 0,k=s,n_k<N_k,m<M_k,m(d_k-d_0)\geq \max\limits_{l\in\mathcal{K}\setminus\{s\}}\min(n_l,M_l)(d_l-d_0),\\ (M_k,k), & i=1,s\neq 0,k=s,n_k<N_k,m=M_k,M_k(d_k-d_0)\geq \max\limits_{l\in\mathcal{K}\setminus\{s\}}\min(n_l,M_l)(d_l-d_0),\\ (n_l,l), & i=1,s\neq 0,k=s,n_k<N_k,0<n_l\leq M_l,min(n_l,M_l)\,(d_l-d_0)>m(d_k-d_0),\\ (M_l,l), & i=1,s\neq 0,k=s,n_k<N_k,n_l>M_l,min(n_l,M_l)\,(d_l-d_0)>m(d_k-d_0),\\ (m,s), & i=1,s\neq 0,k\neq s,n_k<N_k,m(d_s-d_0)\geq min(n_k+1,M_k)\,(d_k-d_0),\\ (n_k+1,k), & i=1,s\neq 0,k\neq s,n_k<M_k,m(d_s-d_0)<(n_k+1)(d_k-d_0),\\ (M_k,k), & i=1,s\neq 0,k\neq s,M_k\leq n_k<N_k,m(d_s-d_0)<M_k(d_k-d_0),\\ (m-1,s), & i=2,s\neq 0,k=s,m>0,n_s=m,(m-1)(d_s-d_0)\geq \max\limits_{l\in\mathcal{K}\setminus\{s\}}\min(n_l,M_l)(d_l-d_0),\\ (M_s,s), & i=2,s\neq 0,k=s,m=M_s,n_s>M_s,M_s(d_s-d_0)\geq \max\limits_{l\in\mathcal{K}\setminus\{s\}}\min(n_l,M_l)(d_l-d_0),\\ (n_l,l), & i=2,s\neq 0,k=s,m>0,0<n_l\leq M_l,min(n_l,M_l)\,(d_l-d_0)>(m-1)(d_s-d_0),\\ (M_l,l), & i=2,s\neq 0,k=s,m>0,n_l>M_l,min(n_l,M_l)\,(d_l-d_0)>(m-1)(d_s-d_0),\\ (m,s), & i=2,s\neq 0,k\neq s,n_k>0.\end{cases}\tag{13}$$

В разделе 3.3 отмечено, что поскольку политика Теоремы 1 однозначно определяет (m,s) как функцию от $\boldsymbol{n}$, система сводится к случайному процессу $\boldsymbol{N}(t)=(N_1(t),\dots,N_K(t))$ с пространством состояний

$$\mathcal{N}=\{\boldsymbol{n}=(n_1,\dots,n_K)\colon 0\leq n_k\leq N_k,\ k\in\mathcal{K}\}.\tag{14}$$

Оптимальные m и s восстанавливаются из $\boldsymbol{n}$:

$$s(\boldsymbol{n})=\arg\max_{k\in\mathcal{K}}(n_k,M_k)(d_k-d_0),$$
$$m(\boldsymbol{n})=\min\big(n_{s(\boldsymbol{n})},M_{s(\boldsymbol{n})}\big).\tag{15}$$

При оптимальной политике миграции стационарное распределение вероятностей состояний системы имеет мультипликативный вид.

Результаты численного анализа представлены на рисунке 2. С ростом интенсивности поступления заявок средняя суммарная E2E-задержка возрастает в обоих сценариях, однако использование MEC-узла обеспечивает ее снижение в 5–6 раз по сравнению со сценарием без граничных вычислений (Cloud Only).

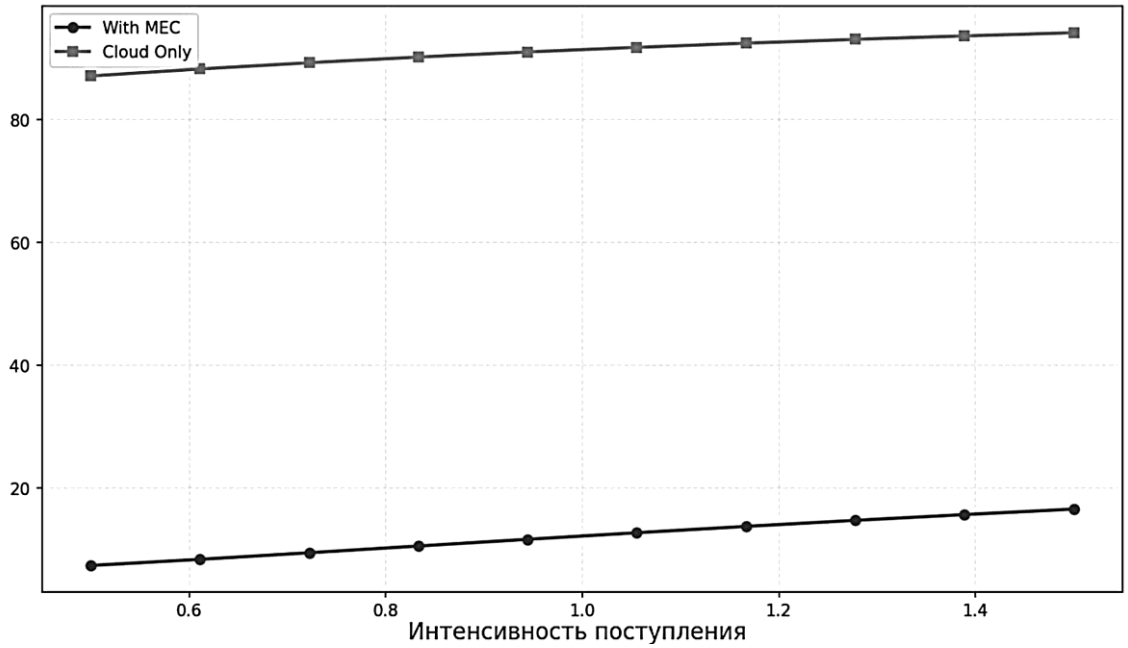

Рисунок 2 – Средняя суммарная E2E-задержка в зависимости от интенсивности поступления заявок при наличии MEC-узла (With MEC) и без него (Cloud Only)

В разделах 3.4–3.6 рассмотрена модель миграции одного сервиса с коррелированным входным потоком. Входной трафик описывается MMPP с двумя фазами нагрузки $l \in \{0,1\}$: фазовое состояние задается непрерывным марковским случайным процессом с интенсивностями переходов α (из фазы 0 в 1) и β (из фазы 1 в 0), интенсивности поступления заявок $\lambda_1 > \lambda_0$.

В разделе 3.4 описана архитектура системы: MEC-узел емкостью M, облачный сервер с общим числом задач N. Единственный класс сервиса с интенсивностью обслуживания μ. Адаптивная политика миграции работает следующим образом: в фазе $l = 1$ (высокая нагрузка) решение о миграции пересматривается при каждом событии; в фазе $l = 0$ (низкая нагрузка) – только при поступлении заявки или смене фазы. Выбор моментов принятия решений о миграции обусловлен характером нагрузки: в фазе высокой интенсивности ($l = 1$) каждое событие может существенно изменить соотношение задержек на MEC и в облаке, поэтому пересмотр политики при каждом событии необходим. В фазе низкой интенсивности ($l = 0$) промежуточные пересмотры нецелесообразны ввиду малой вероятности переполнения MEC-узла.

Система описывается случайным процессом $\boldsymbol{X}(t), t \geq 0$ с состояниями (n, m, l), где n общее число заявок, m число на MEC-узле, l фаза. Пространство состояний имеет вид

$$\begin{gathered} \mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1, \\ \mathcal{X}_0 = \{(n,m,0): 0 \leq n \leq N,\ 0 \leq m \leq \min(n,M)\}, \\ \mathcal{X}_1 = \{(n,m,1): 0 \leq n \leq N,\ m = \min(n,M)\}. \end{gathered} \tag{16}$$

В фазе $l = 1$ состояние m однозначно определяется n ($m = \min(n,M)$), поскольку MEC всегда заполнен до предела. В фазе $l = 0$ m может принимать любое значение от 0 до $\min(n,M)$.

В разделе 3.5, при лексикографическом упорядочивании состояний по числу задач n матрица интенсивностей переходов $\boldsymbol{Q}$ принимает блочно-трехдиагональный вид.

Лемма 2. Для модели с перемещением заявок и коррелированным потоком, при лексикографическом порядке на пространстве состояний $\mathcal{X}$

$$\begin{gathered} \boldsymbol{x} = (n,m,l) < (n',m',l') = \boldsymbol{x}' \Leftrightarrow \\ (n < n') \vee \big(n = n', (l < l') \vee (l = l', m < m')\big) \end{gathered} \tag{17}$$

матрица интенсивностей переходов случайного процесса $\boldsymbol{X}(t)$ представима в блочно-трехдиагональном виде

$\boldsymbol{Q}$	0	1	2	⋯	$M-1$	M	$M+1$	⋯	$N-1$	N
0	$\boldsymbol{B}_0$	$\boldsymbol{A}_0$	0	0	0	0	0	0	0	0
1	$\boldsymbol{C}_1$	$\boldsymbol{B}_1$	$\boldsymbol{A}_1$	0	0	0	0	0	0	0
2	0	$\boldsymbol{C}_2$	$\boldsymbol{B}_2$	⋱	0	0	0	0	0	0
⋮	0	0	⋱	⋱	⋱	0	0	0	0	0
$M-1$	0	0	0	⋱	$\boldsymbol{B}_{M-1}$	$\boldsymbol{A}_{M-1}$	0	0	0	0
M	0	0	0	0	$\boldsymbol{C}_M$	$\boldsymbol{B}_M$	$\boldsymbol{A}_M$	0	0	0
$M+1$	0	0	0	0	0	$\boldsymbol{C}_{M+1}$	$\boldsymbol{B}_M$	$\boldsymbol{A}_M$	0	0
⋮	0	0	0	0	0	0	$\boldsymbol{C}_{M+1}$	⋱	⋱	0
$N-1$	0	0	0	0	0	0	0	⋱	$\boldsymbol{B}_M$	$\boldsymbol{A}_M$
N	0	0	0	0	0	0	0	0	$\boldsymbol{C}_{M+1}$	$\boldsymbol{B}_N$

(18)

с ненулевыми блоками

$$\boldsymbol{A}_n = (\boldsymbol{A}_{n0}, \boldsymbol{a}_{n1}^T),$$
$$\boldsymbol{A}_{n0} = diag^+(\lambda_0, \dots, \lambda_0), \boldsymbol{a}_{n1}^T = \lambda_1 \cdot \boldsymbol{e}_{n+2}, n = 0, \dots, M-1, \quad (19)$$

$$\boldsymbol{A}_M = \begin{pmatrix} \boldsymbol{A}_{M2} & \boldsymbol{0^T} \\ \boldsymbol{0} & \lambda_1 \end{pmatrix},$$
$$\boldsymbol{A}_{M2} = diag^+(\lambda_0, \dots, \lambda_0) + diag\,(0, \dots, 0, \lambda_1), n = M, \dots, N, \quad (20)$$

$$\boldsymbol{C}_n = \begin{pmatrix} \boldsymbol{C}_{n0} \\ \boldsymbol{c}_{n1} \end{pmatrix},$$
$$\boldsymbol{C}_{n0} = diag\,(n\mu, (n-1)\mu, \dots, \mu, 0) + diag^-(\mu, 2\mu, \dots, n\mu), \quad (21)$$
$$\boldsymbol{c}_{n1} = n\mu\, \boldsymbol{e}_{n+1}, n = 0, \dots, M,$$

$$\boldsymbol{C}_{M+1} = \begin{pmatrix} \boldsymbol{C}_{M2} & \boldsymbol{0^T} \\ \boldsymbol{0} & n\mu \end{pmatrix},$$
$$\boldsymbol{C}_{M2} = diag\,(n\mu, (n-1)\mu, \dots, \mu, 0) + diag^-(\mu, 2\mu, \dots, n\mu), \quad (22)$$

$$\boldsymbol{B}_0 = \begin{pmatrix} -(\lambda\_0 + \alpha) & \alpha \\ \beta & -(\lambda_1 + \beta) \end{pmatrix}, \quad (23)$$

$$\boldsymbol{B}_n = \begin{pmatrix} \boldsymbol{B}_{n00} & \boldsymbol{b}_{n01}^{\boldsymbol{T}} \\ \boldsymbol{b}_{n10} & -(\lambda_1 + n\mu + \beta) \end{pmatrix}, \quad (24)$$

$$\boldsymbol{B}_{n00} = (\lambda_0 + n\mu + \alpha) \cdot \boldsymbol{I}_{M+1}, \boldsymbol{b}_{n01} = (\alpha, \dots . \alpha), \quad \boldsymbol{b}_{n10} = (0, \dots . 0, \beta), \quad (25)$$
$$n = 1, \dots, N-1,$$

$$\boldsymbol{B}_N = \begin{pmatrix} \boldsymbol{B}_{N00} & \boldsymbol{b}_{N01}^{\boldsymbol{T}} \\ \boldsymbol{b}_{N10} & -(N\mu + \beta) \end{pmatrix}, \quad (26)$$
$$\boldsymbol{B}_{N00} = -[N\mu + \alpha] \cdot \boldsymbol{I}_N, \boldsymbol{b}_{N01}^{\boldsymbol{T}} = \boldsymbol{b}_{n01}, \boldsymbol{b}_{N10} = \boldsymbol{b}_{n10}.$$

Теорема 2. Для модели с перемещением заявок и коррелированным потоком, стационарное распределение случайного процесса $\boldsymbol{X}(t)$ рассчитывается в матричном виде по формуле

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 \prod_{i=0}^{n-1} \boldsymbol{R}_i, \qquad n = \overline{1, N}\,, \quad (27)$$

где $\boldsymbol{\pi}_0$ является единственным решением системы уравнений

$$\begin{cases} \boldsymbol{\pi}_0(\boldsymbol{B}_0 + \boldsymbol{R}_0\boldsymbol{C}_1) = 0, \\ \sum_{n=0}^{N} \boldsymbol{\pi}_0 \prod_{i=0}^{n-1} \boldsymbol{R}_i\, \boldsymbol{1}^{\boldsymbol{T}} = 1, \end{cases} \quad (28)$$

а матрицы $\boldsymbol{R}_i$ вычисляются по рекуррентным соотношениям

$$\boldsymbol{R}_n = -\boldsymbol{A}_n\,(\boldsymbol{B}_{n+1} + \boldsymbol{R}_{n+1}\boldsymbol{C}_{n+2}\,)^{-1}, \;\; n = 0, \dots, M-1,$$
$$\boldsymbol{R}_n = -\boldsymbol{A}_M\,(\boldsymbol{B}_M + \boldsymbol{R}_{n+1}\boldsymbol{C}_{M+1}\,)^{-1}, \;\; n = M, \dots, N-2, \quad (29)$$
$$\boldsymbol{R}_{N-1} = -\boldsymbol{A}_M\,\boldsymbol{B}_N^{-1}.$$

В разделе 3.6 параметры MMPP-потока оценены по реальному трафику мобильного оператора Vodafone (EM-алгоритм, пакет mapfit/R). Выбраны 5 классов сервисов: потоковое видео, веб-приложения, игровые приложения, IP-телефония, обмен мгновенными сообщениями. Например, матрицы для сервиса «Игровые приложения» выглядят следующим образом

$$\boldsymbol{D}_0^4 = \begin{bmatrix} -5{,}2699 & 1{,}2505 \\ 7{,}0706 & -11{,}0899 \end{bmatrix}, \boldsymbol{D}_1^4 = \begin{bmatrix} 4{,}0194 & 0{,}00 \\ 0{,}00 & 4{,}0193 \end{bmatrix}. \quad (30)$$

На рисунке 3 представлены параметры MMPP-потока. Наибольшие интенсивности поступления заявок – как в низкой, так и в высокой фазе –

характерны для веб-приложений, наименьшие – для потокового видео, что отражает существенно коррелированный характер потокового видео.

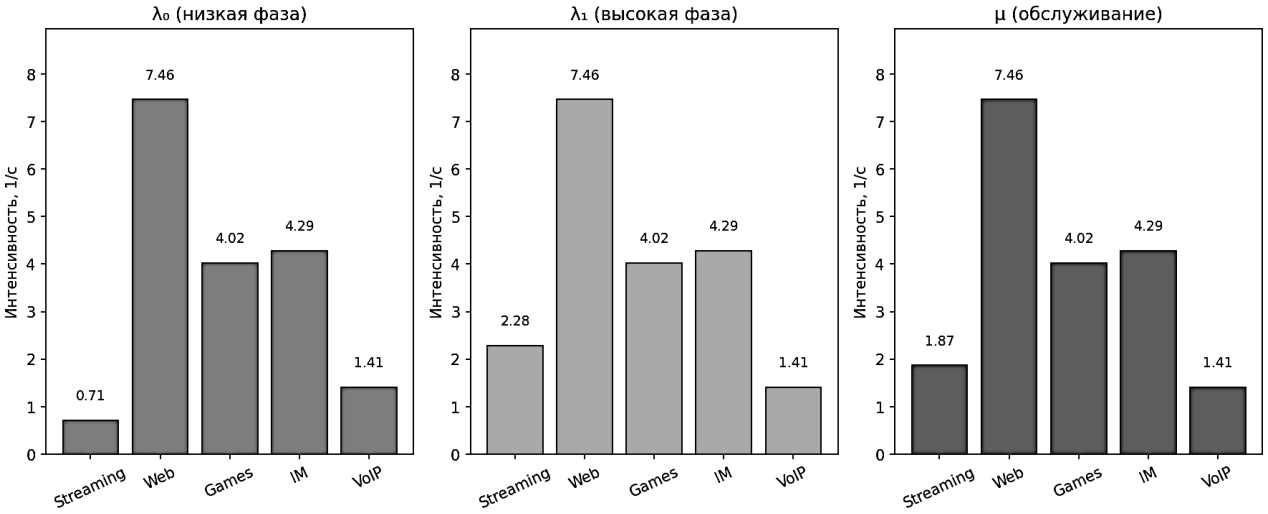

Рисунок 3 – Параметры MMPP-потока (λ_0, λ_1, μ) 1/с для пяти классов сервисов по данным мобильного оператора Vodafone

Матрицы $\boldsymbol{D}_0, \boldsymbol{D}_1$ использованы в качестве входных данных при численном анализе модели. Рассматривались следующие параметры сценария $N = 100$, $M = 5$. Результаты численного анализа представлены в Таблице 1. Высокая доля обслуживания на MEC-узле (свыше 84%) для всех классов сервисов подтверждает эффективность предложенной политики миграции при ограниченной емкости граничного узла. Вероятность блокировки для веб-приложений на 6 порядков ниже, чем для потокового видео. Это объясняется существенно большим коэффициентом вариации MMPP-потока для потокового видео ($\lambda_1/\lambda_0 \approx 3{,}2$) по сравнению с веб-приложениями ($\lambda_1/\lambda_0 \approx 1{,}0$), что приводит к более выраженным всплескам нагрузки и увеличению вероятности переполнения системы.

Таблица 1 – Сравнение характеристик модели по классам сервисов

Сервис	Загрузка MEC	Доля на MEC	Доля в облаке	Вероятность блокировки
Потоковое видео	0,1883	0,8452	0,1547	7,25E-06
Веб-приложения	0,1998	0,9992	0,0007	1,67E-12
Игровые приложения	0,1997	0,9989	0,0010	3,89E-11
Обмен сообщениями	0,1996	0,9976	0,0023	1,36E-08
IP-телефония	0,1994	0,9971	0,0028	1,92E-09

Заключение содержит основные научные результаты работы.

ЗАКЛЮЧЕНИЕ

Основные научные результаты работы состоят в следующем.

1. Разработана модель миграции виртуальных машин с обслуживаемыми задачами между облачными серверами в виде системы массового обслуживания с перемещением всех заявок класса между группами приборов. Решение о миграции принимается только в момент поступления новой заявки соответствующего класса. Формализованы два алгоритма миграции, направленные на минимизацию занятой пропускной способности серверов в облачной инфраструктуре и различающиеся моментом оценки занятости приборов: до и после принятия поступившей заявки. Получены формулы для расчета показателей эффективности миграции виртуальных машин, в том числе вероятности миграции и средней высвобождаемой пропускной способности сервера.
2. Разработана модель миграции пользователей сервисов между граничным и облачными серверами в виде системы массового обслуживания с перемещением заявок между общей группой приборов с эксклюзивным обслуживанием одного класса и индивидуальными приборами. Политика миграции формализована как задача минимизации суммарной межконцевой задержки по всем пользователям при каждом изменении состояния системы. Получено оптимальное распределение заявок между группами приборов в виде функции от числа заявок в системе. Показано, что стационарное распределение вероятностей состояний системы при оптимальной политике имеет мультипликативный вид.
3. Модель миграции с коррелированным потоком заявок пользователей в беспроводной гранично-облачной архитектуре построена в виде системы массового обслуживания с входным потоком, моделируемым марковски-модулированным пуассоновским процессом MMPP. Политика миграции зависит от текущей фазы MMPP-потока: в фазе высокой интенсивности решение о миграции пересматривается при каждом событии, в фазе низкой интенсивности – только при поступлении новой заявки или смене фазы. Матрица интенсивностей переходов представлена в блочно-трехдиагональном виде, на основе которого разработан матричный алгоритм расчета стационарного распределения. Численный анализ модели выполнен с параметрами MMPP-потока, оцененными по реальному сетевому трафику методом максимального правдоподобия на основе EM-алгоритма для различных типов сервисов.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

1. **Kushchazli A.,** Leonteva K., Gaidamaka E., Kochetkova I. A delay-aware queuing model for performance analysis of service migration in MEC-Cloud environments // Lecture Notes in Computer Science. – 2026. – Vol. 16461. – P. 218–233.
2. **Kushchazli A.,** Leonteva K., Kochetkova I., Khakimov A. Evaluating QoS in dynamic virtual machine migration: A multi-class queuing model for edge-cloud systems // Journal of Sensor and Actuator Networks. – 2025. – Vol. 14, No. 3. – Art. No. 47.
3. **Kushchazli A.,** Safargalieva A., Kochetkova I., Gorshenin A. Queuing model with customer class movement across server groups for analyzing virtual machine migration in cloud computing // Mathematics. – 2024. – Vol. 12, No. 3. – Art. No. 468.
4. Kochetkova I., **Kushchazli A.,** Burtseva S., Gorshenin A. Short-term mobile network traffic forecasting using seasonal ARIMA and Holt-Winters models // Future Internet. – 2023. – Vol. 15, No. 9. – Art. No. 290.
5. **Кущазли А.И.,** Леонтьева К.А., Кочеткова И.А. Расчет показателей эффективности модели миграции виртуальных машин между серверами облачной инфраструктуры с выбором наименее загруженного сервера // Свидетельство о государственной регистрации программы для ЭВМ № 2025681795 РФ : заявл. 04.07.2025 : опубл. 18.08.2025.
6. Ермолаев А.М., **Кущазли А.И.,** Кочеткова И.А. Классификация пакетов передачи данных по типам услуг методом двухэтапного анализа метаданных для инструментов захвата сетевого трафика // Свидетельство о государственной регистрации программы для ЭВМ № 2025685024 РФ : заявл. 18.08.2025 : опубл. 19.09.2025.
7. Сафаргалиева А.И., **Кущазли А.И.,** Нохуров М., Кочеткова И.А. Расчет характеристик модели выполнения задач в среде облачных вычислений с миграцией виртуальных машин // Свидетельство о государственной регистрации программы для ЭВМ № 2023663762 РФ : № 2023662447 : заявл. 15.06.2023 : опубл. 28.06.2023.
8. **Кущазли А.И.** Система массового обслуживания с перемещением классов заявок по группам приборов для анализа миграции сервисов в облачной инфраструктуре // Ломоносов-2025 : материалы Международного молодежного научного форума, Москва, 11–25 апреля 2025 г. – М.: МАКС Пресс, 2025. – С. 216–217.
9. Ермолаев А.М., **Кущазли А.И.,** Хакимов А.А., Кочеткова И.А. Разработка модели трафика в задаче миграции сервисов между граничными МЕС-узлами и облаком // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем : материалы Всероссийской конференции с международным участием, Москва, 07–11 апреля 2025 г. – М.: РУДН, 2025. – С. 144–148.
10. **Кущазли А.И.,** Власкина А.С., Кочеткова И.А. Система массового обслуживания с перемещением классов заявок между группами приборов для анализа миграции виртуальных машин в облачных вычислениях // Информационные технологии и технические средства управления (ICCT-2024): материалы VIII Международной научной конференции, Владикавказ, 01–05 октября 2024 г. – М. : ИПУ РАН, 2024. – С. 404–405.
11. **Кущазли А.И.,** Сафаргалиева А.И., Кочеткова И.А. Система массового обслуживания с перемещением классов заявок между группами приборов для анализа миграции виртуальных машин // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: материалы Всероссийской конференции с международным участием, Москва, 08–12 апреля 2024 г. – М.: РУДН, 2024. – С. 69–72.
12. Силкина М.А., **Кущазли А.И.,** Кочеткова И.А., Горшенин А.К. К статистическому анализу профиля трафика мобильного оператора // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: материалы Всероссийской конференции с международным участием, Москва, 18–22 апреля 2022 г. – М.: РУДН, 2022. – С. 152–155.

Кущазли Анна Ивановна (Российская Федерация)

Модели массового обслуживания для анализа эффективности миграции сервисов в граничных облачных вычислениях

В диссертации разработаны модели миграции виртуальных машин и сервисов пользователей в виде систем массового обслуживания, позволившие выявить закономерности процессов миграции в облачной инфраструктуре и гранично-облачной архитектуре для сценариев предоставления сервисов дополненной и виртуальной реальности, облачного гейминга в сетях шестого поколения 6G. Предложены алгоритмы для анализа и расчета показателей эффективности миграции, а также определения оптимальных политик миграции виртуальных машин между облачными серверами по критерию минимизации занятой пропускной способности и сервисов пользователей между граничным и облачным серверами по критерию минимизации суммарной межконцевой задержки с учетом коррелированного характера входного потока заявок в виде MMPP. Доказаны утверждения, позволяющие вычислять стационарные распределения вероятностей состояний моделей миграции, в том числе в мультипликативном виде и на основе матричного алгоритма с блочно-трехдиагональной структурой матрицы интенсивностей переходов.

Kushchazli Anna Ivanovna (Russia)

Queuing models for performance analysis of service migration in cloud and multi access edge computing

This dissertation develops queueing models for virtual machine and user service migration that reveal patterns governing migration processes in cloud infrastructure and multi-access edge computing (MEC) architecture for scenarios involving augmented and virtual reality services and cloud gaming in sixth-generation (6G) networks. The study proposes algorithms for analyzing and computing migration performance metrics, as well as for determining optimal migration policies: virtual machine migration between cloud servers based on the criterion of minimizing occupied server bandwidth, and user service migration between edge and cloud servers based on the criterion of minimizing the total end-to-end delay, accounting for correlated arrival processes modeled as a Markov-modulated Poisson process (MMPP). The work establishes theoretical propositions for computing stationary probability distributions of migration model, including a closed-form product-form solution and a matrix algorithm exploiting the block-tridiagonal structure of the transition rate matrix.