РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

На правах рукописи

Дараселия Анастасия Валерьевна

МОДЕЛИ И АНАЛИЗ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ МЕХАНИЗМОВ ВЫГРУЗКИ ТРАФИКА В ГЕТОРОГЕННЫХ БЕСПРОВОДНЫХ СЕТЯХ

Специальность 1.2.3. Теоретическая информатика, кибернетика

Диссертация
на соискание ученой степени кандидата
физико-математических наук

Научный руководитель
кандидат физико-математических наук,
Сопин Эдуард Сергеевич

Москва – 2022

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. Концепция выгрузки в инфокоммуникационных сетях становится все более актуальной. Выделяют два основных направления реализации этой концепции. Первым из них является выгрузка мобильных вычислений в распределенную систему туманно-облачных вычислений. Туманные вычисления предполагают размещение вычислительных узлов на границе сети, что является хорошим решением для критичных ко времени отклика ресурсоемких мобильных приложений. Система туманно-облачных вычислений позволяет мобильным устройствам выгрузить некоторые задачи и, следовательно, увеличить время автономной работы. Кроме того, в отличие от классических облачных решений, туманные вычисления имеют меньшее время отклика благодаря своей близости к конечному пользователю. Однако узлы туманных вычислений не обладают ресурсами облачных дата-центров, поэтому легко перегружаются. В связи с этим задача разработки моделей выгрузки мобильных вычислений в туманно-облачную инфраструктуру для анализа времени отклика является актуальной.

Вторым крупным направлением является выгрузка трафика из лицензированного спектра частот, традиционно используемого мобильными операторами, в нелицензированный для увеличения емкости сети. Впервые технология выгрузки была стандартизирована в рамках сетей связи четвертого поколения и недавно была включена в стандарты пятого поколения. Таким образом, задача разработки моделей для анализа дополнительной емкости сети, возникающей вследствие использования нелицензированного диапазона частот, является актуальной.

Степень разработанности темы. Анализ вероятностных характеристик выгрузки трафика в гетерогенных беспроводных сетях проведен с помощью аппарата теории вероятностей, теории массового обслуживания, теории случайных процессов, теории телетрафика. К российским ученым, исследователям, внесшим большой вклад в эти области, относятся Г.П. Башарин [79 - 83, 132], П.П. Бочаров [80, 84], В.М. Вишневский [85 - 88], Ю.В. Гайдамака [110, 111, 132, 133], А.Н.

Дудин [86], А.И. Зейфман [115-119], Гольдштейн Б.С [89], А.Е. Кучерявый [89, 90], Е.А. Кучерявый [69, 90, 95], А.Н. Моисеев [92, 93], С.П. Моисеева [91], Д.А. Молчанов [43, 69 - 72, 94, 95], А.А. Назаров [96], В.А. Наумов [97 - 100, 110, 111, 133], А.П. Пшеничников [101], В.В. Рыков [68], К.Е. Самуйлов [43, 82, 98 - 100, 110, 111, 121, 132, 133], С.Н. Степанов [106 - 108], М.С. Степанов [108], И.И. Цитович [122, 123], С.Я. Шоргин [113, 114, 120, 121] и др., а к зарубежным – M. Dohler [42, 43, 130], J.G. Andrews [126-129], F.P. Kelly [103, 104], V.B. Iversen [109], L. Kleinrock [57], E. Gelenbe [124, 125], Luis M Correia [65], K.W. Ross [105] и др.

Обзор конкретных работ содержится в главах диссертационной работы по мере изложения решений поставленных задач.

Целью диссертационной работы является разработка и анализ моделей для расчета показателей эффективности механизмов выгрузки задач мобильных вычислений и выгрузки трафика в гетерогенных беспроводных сетях.

Для достижения цели в диссертационной работе решаются следующие **задачи**:

1. Построение в виде системы массового обслуживания модели выгрузки в систему туманно-облачных вычислений задач мобильных вычислений, разработка метода расчета функции распределения времени отклика с учетом неоднородности задач по объему вычислений и данных.
2. Построение и анализ двухпараметрической модели в виде ресурсной системы массового обслуживания, описывающей обслуживание трафика в лицензированном диапазоне частот и дискретной цепи Маркова, описывающей процедуру случайного доступа при выгрузке в нелицензированный диапазон частот. Разработка метода расчета распределения скорости передачи в нелицензированном диапазоне.

Объем и структура работы. Структура диссертации построена из введения, трех глав, заключения и списка литературы из 103 источников. Научная работа изложена на 100 страницах текста, содержит 26 рисунков и 2 таблицы.

Краткое изложение диссертации. Диссертация состоит из трех глав. В **первой главе** работы рассмотрены основные принципы выгрузки задач мобильных

вычислений и трафика в беспроводных сетях, проведен общий обзор технологий и методов анализа и описываются базовые модели выгрузки. В разделе 1.1 изложены основные особенности выгрузки мобильных вычислений и трафика в беспроводных сетях, приведен краткий обзор литературы в направлении исследования. В разделе 1.2 рассмотрена базовая модель времени отклика с однопараметрическим критерием выгрузки мобильных вычислений в систему туманно-облачных вычислений. В разделе 1.3 рассмотрена базовая модель для анализа эффективности выгрузки трафика в нелицензированный диапазон. При написании разделов 1.1 – 1.3 использовались публикации [65, 67, 68, 69, 70, 76, 77] с участием автора.

Вторая глава посвящена построению и анализу модели выгрузки задач мобильных вычислений в туманно-облачную систему. В разделе 2.1 представлена расширенная модель с двухпараметрическим критерием выгрузки. В разделе 2.2 проведен анализ распределения времени отклика в условиях выгрузки. В разделе 2.3 приведены результаты численного анализа оптимальных порогов критерия выгрузки. При написании разделов 2.1 – 2.3 использовались публикации [66, 73] с участием автора.

В третьей главе построена и проанализирована модель ресурсных СМО для анализа показателей эффективности выгрузки трафика в нелицензированный диапазон. В разделе 3.1 представлена модель выгрузки трафика с использованием ресурсных СМО. Предложены три механизма выгрузки трафика на нелицензированный диапазон. В разделе 3.2 проведен сравнительный анализ показателей эффективности выгрузки для предложенных трех стратегий. При написании разделов 3.1-3.4 использовались публикации [71, 72, 74, 75] с участием автора.

В заключительном разделе представлены основные результаты диссертационной работы.

Положения, выносимые на защиту.

1. Разработанная модель механизма выгрузки задач с мобильных устройств в туманно-облачную инфраструктуру по пороговому значению объема

вычислений позволяет рассчитывать функцию распределения (ФР) времени отклика. Модель учитывает неоднородность поступающих задач по объему необходимых вычислений.

2. Модель двухпараметрического механизма выгрузки мобильных вычислений в туманно-облачную инфраструктуру учитывает неоднородность задач по объему вычислений и по объему передаваемых данных, а также задержку передачи на беспроводном участке сети. Модель позволяет рассчитывать дискретное распределение времени отклика и математическое ожидание потребления энергии мобильными устройствами.
3. Анализ показателей эффективности выгрузки трафика в нелицензированный диапазон частот беспроводной сети осуществляется при помощи модели, состоящей из двух компонентов. Модель в виде ресурсной системы массового обслуживания (РеСМО) описывает обслуживание сессий в лицензированном диапазоне и позволяет вычислять долю выгружаемого трафика, а также распределение требований к ресурсу выгружаемых сессий. Модель в виде дискретной цепи Маркова описывает процедуру случайного доступа в нелицензированном диапазоне и позволяет вычислять распределение скорости передачи.

Научная новизна диссертационной работы:

1. В однопараметрической модели выгрузки задач с мобильных устройств, в отличие от известных, была учтена неоднородность задач по объему необходимых вычислений. Объем вычислений, от которого зависит время обслуживания, полагается случайной величиной с заданным распределением.
2. В модели выгрузки задач мобильных вычислений с двухпараметрическим критерием выгрузки, помимо неоднородности объема вычислений, были учтены неоднородность объема данных и случайная задержка при выгрузке в узел туманных вычислений.
3. В модели ресурсной СМО, описывающей передачу трафика в лицензированном диапазоне частот, получено распределение требований

сброшенных заявок к ресурсу системы. Это позволило вычислить распределение скорости передачи данных в нелицензированных частотах.

Методы исследования. В диссертации применяются методы теории массового обслуживания, теории вероятностей, теории случайных процессов и математической теории телетрафика.

Теоретическая и практическая значимость работы. Полученные в диссертационной работе результаты могут быть использованы проектными телекоммуникационными компаниями, операторами сетей связи при планировании сетей радиодоступа для предоставления требуемого качества услуг.

Разработанные математические модели позволяют провести анализ показателей эффективности механизмов выгрузки задач и мобильных вычислений и выгрузки трафика в гетерогенных беспроводных сетях.

Результаты работы включены в исследования по грантам РФФИ №19-07-00933 “Стохастические модели и задачи оптимизации для разработки информационных технологий виртуализации и управления ресурсами в беспроводных мультисервисных сетях”, № 19-07-00919 “Исследование задач оптимизации топологии инфокоммуникационных сетей и разработка семейства точных и приближенных алгоритмов решения на основе декомпозиции их математических моделей”, № 20-07-01052 “Вероятностный анализ эффективности механизмов выгрузки задач интернета вещей и мобильных вычислений в туманно-облачную систему беспроводных сетей 5G”.

Реализация результатов работы. Основные научные достижения, полученные в диссертации, использованы в совместных исследовательских мероприятиях в рамках сотрудничества РУДН, в исследованиях по грантам РФФИ, в проекте «5-100» повышения конкурентоспособности ведущих российских университетов среди ведущих мировых научно-образовательных центров.

Степень достоверности и апробация результатов. Основные результаты, изложенные в диссертации, докладывались на научных конференциях и семинарах: международная конференция «International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops (ICUMT)» (Москва, ноябрь

2018); международная конференция имени А.Ф. Терпугова «Информационные технологии и математическое моделирование, ИТММ» (Томск, сентябрь 2020 г., декабрь 2021 г.); международная конференция «IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)» (Лондон, сентябрь 2020 г.); международная конференция по проводным и беспроводным сетям и системам следующего поколения (NEW2AN) (Санкт-Петербург, август 2018 г., август 2019 г.); всероссийская конференция с международным участием «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем (ИТТММ)» (Москва, апрель 2019 г., апрель 2020 г., апрель 2021 г.); международная конференция «12th International Workshop on Applied Problems in Theory of Probabilities and Mathematical Statistics (APTP+MS 2018, Summer Session)» (Португалия, Лиссабон, октябрь 2018г.).

Основные результаты опубликованы в ведущих научных журналах – Lecture Notes in Computer Science, IEEE Transactions on Vehicular Technology, Информатика и ее применение, а также в трудах международных конференций, индексируемых WoS (Web of Science) и Scopus.

Соответствие паспорту специальности. Диссертационное исследование соответствует следующим разделам паспорта специальности 1.2.3. Теоретическая информатика, кибернетика, а именно **п. 11** «Распределенные многопользовательские системы»; **п. 12** «Модели информационных процессов и структур»; **п. 23** «Новые интернет - технологии, включая средства поиска, анализа и фильтрации информации».

Личный вклад. Программные средства, используемые для численного анализа, и представленные в диссертации модели и результаты их анализа получены с участием автора.

Публикации. Основные результаты по теме диссертационного исследования изложены в 12 печатных изданиях [65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77], из которых издание [66, 72] из списка ВАК/РУДН, а издания [65, 67, 68, 69, 70, 71] входят в базы данных Scopus/WoS.

ГЛАВА 1. МОДЕЛИ ВЫГРУЗКИ МОБИЛЬНЫХ ВЫЧИСЛЕНИЙ И ТРАФИКА В БЕСПРОВОДНЫХ СЕТЯХ

1.1. Особенности выгрузки вычислительных задач и трафика

В задачах выгрузки большинство исследователей выделяют две крупные задачи. Первая из них – это выгрузка мобильных вычислений в распределенную инфраструктуру, например, в систему туманно-облачных вычислений. Помимо этого, существует подход выгрузки трафика из классического лицензированного спектра, который стандартно используется в мобильных сетях, в нелицензированный спектр, который в мобильных сетях обычно не используется.

Начнем с первой из них, выгрузки в систему туманно-облачных вычислений. Термин «туманные вычисления» (*англ.* Fog computing) был предложен в 2012 году исследователями из Cisco Systems, представляющие собой парадигму, которая расширяет облачные сервисы до границ сети [2]. Обработка логики приложений и данных на периферии - не новая концепция. Концепция периферийных, или граничных вычислений (*англ.* Edge computing), появилась примерно в 2000-х годах [3], [4], другая похожая концепция, облачность, была введена в 2009 году [5]. И «облачко» (*англ.* Cloudlet, небольшой облачный центр обработки данных), и туманные вычисления являются развитием аналогичной концепции, которая основана на обработке на периферийном уровне. В то время как облачные вычисления применяются в мобильной сети, туманные вычисления применяются на уровне подключенных вещей, в соответствии с концепцией интернета вещей (*англ.* internet of things, IoT) [6].

На рисунке 1.1 показана обобщенная схема туманно-облачных вычислений. Устройства туманных вычислений, серверы туманных вычислений и шлюзы являются основными вычислительными компонентами в среде туманных вычислений. Любое устройство, имеющее вычислительные, сетевые и запоминающие возможности, может действовать как устройство туманных вычислений. Эти устройства включают телевизионные приставки, коммутаторы, маршрутизаторы, базовые станции, прокси-серверы или любое другое

вычислительное устройство. Серверы туманных вычислений, которые управляют несколькими устройствами туманных вычислений и шлюзами туманных вычислений, отвечают за услуги перевода между разнородными устройствами в среде туманных вычислений. Шлюзы туманных вычислений также предоставляют услуги перевода между уровнями интернета вещей, туманных и облачных вычислений.

Рис. 1.1. Схема туманно-облачных вычислений.

Для принятия и развертывания на рынке туманные вычисления должны иметь стандартную архитектуру. На сегодняшний день нет доступной стандартной архитектуры, однако во многих исследовательских работах представлены варианты архитектуры туманных вычислений [1, 102]. Описание одной из таких архитектур, высокоуровневой, приведем ниже.

В этой архитектуре структура туманных вычислений имеет три разных уровня, основной из которых - уровень туманных вычислений. Этот уровень состоит из всех промежуточных вычислительных устройств. На нем можно использовать традиционные технологии виртуализации аналогично облаку. Этот уровень накапливает данные, генерируемые датчиками, с уровня интернета вещей и отправляет запросы, связанные со срабатыванием, после обработки. Хотя кажется, что проблема больших данных решается обработкой сгенерированных данных на уровне периферийных вычислений, однако слишком большое количество устройств создает проблему с их обработкой. Фактически, на этом

уровне можно использовать малые и средние объемы обработки больших данных. Было проведено множество исследовательских работ по обработке больших данных на уровне туманных вычислений [7] - [13].

Самый нижний уровень - это уровень интернета вещей, который состоит из всех подключенных устройств. Устройства на этом уровне выполняют процесс считывания и обработки. Для чувствительных ко времени приложений обработка должна выполняться исключительно в плоскости туманных вычислений, в то время как облако может выполнять другие задачи, не критичные ко времени. Однако уровень туманных вычислений будет управлять тем, что нужно отправлять в облако, а что нет. По запросу пользователи могут получать услуги как из туманных вычислений, так и из облачных. Облачный уровень отвечает за управление сложной обработкой и хранением.

Теперь поговорим о связных парадигмах и технологиях. Туманные вычисления используют вычислительные ресурсы рядом с сетями, расположенными между традиционными облачными и пограничными устройствами, для обеспечения более качественной и быстрой обработки приложений и услуг [13]. Помимо туманных вычислений существует несколько аналогичных парадигм вычислений, таких как мобильные облачные вычисления (*англ.* Mobile Cloud Computing, MCC), мобильные периферийные вычисления (англ. Mobile-Edge Computing, MEC), переферийные вычисления (*англ.* Edge Computing), росистые вычисления (*англ*. Dew Computing) и туманно-росистые вычисления (*англ.* Fog-dew computing). В облачных вычислениях все устройства IoT напрямую подключены к облаку, и вычисления полностью зависят от облака. Однако все вышеперечисленные аналогичные технологии не зависят исключительно от облака, а зависят от некоторых промежуточных устройств для вычислений; некоторые из них даже не требуют подключения к облаку. На рис. 1.2 показана высокоуровневая архитектура этих технологий.

Далее коротко разберем отдельно каждую из упомянутых выше схем.

А) Мобильные облачные вычисления, MCC. Удаленное выполнение выгруженных мобильных услуг осуществляется с использованием MCC рядом с

оконечными пользователями [32], [33]. Технология МСС преодолевает ограничения вычислительных ресурсов, ресурсов хранения данных мобильных устройств и их энергопотребления. Для этого, как правило, на границе сети размещается небольшой облачный сервер [34].

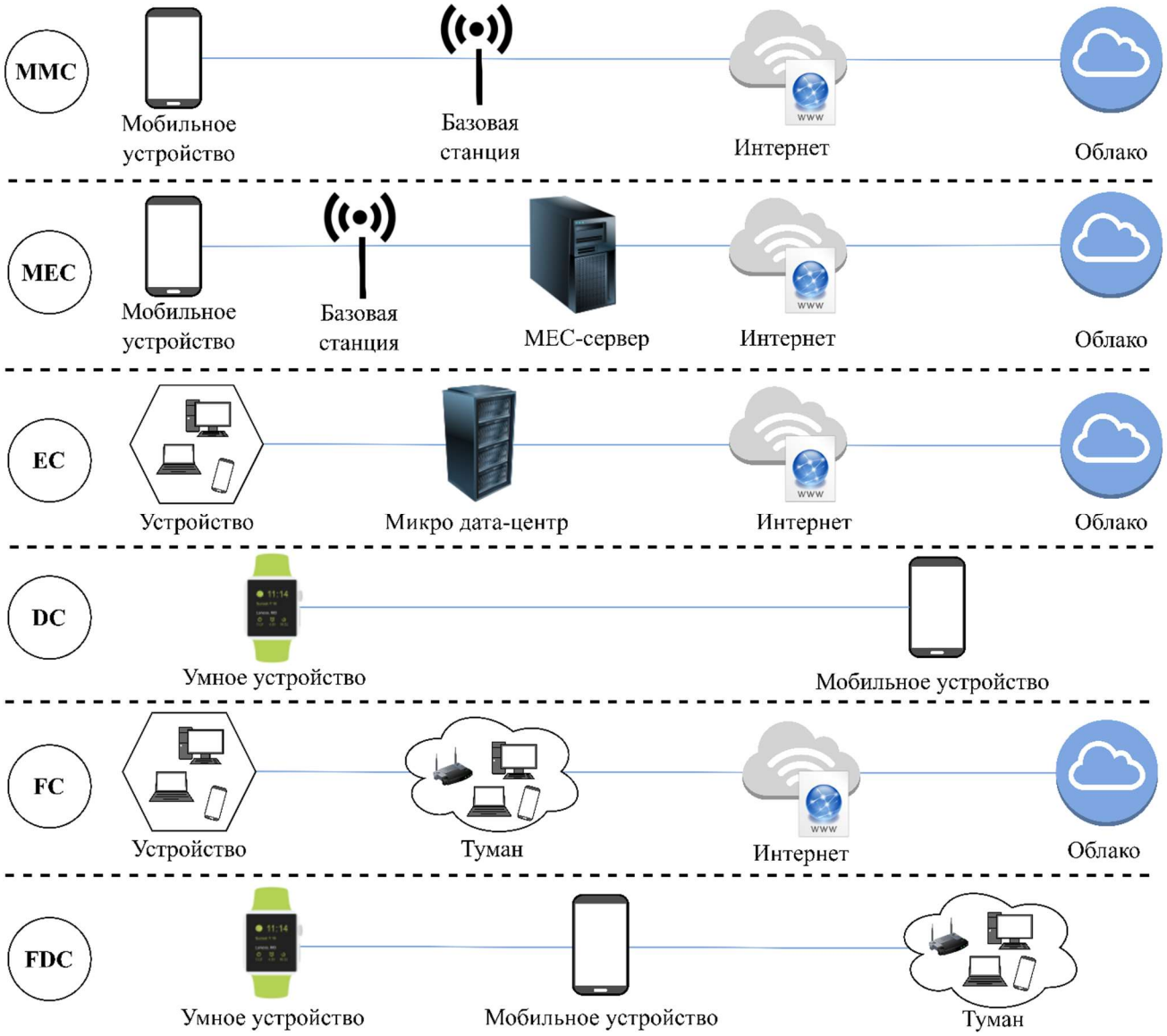

Рис. 1.2. Схема парадигм вычислений.

Данная технология обеспечивает доступ к данным, приложениям и облаку через Интернет для мобильных пользователей. Ожидается, что в будущем эта технология будет применяться в образовании, развитии городов и сельских районов, здравоохранении и социальных сетях [32]. Кроме того, в настоящее время становятся популярны приложения с интенсивными вычислениями, такие как дополненная реальность (*англ.* augmented reality, AR), распознавание речи,

машинное обучение (*англ.* machine learning), планирование и принятие решений, а также приложения для обработки естественного языка, где также планируется использовать ММС [33].

Б) Мобильные периферийные вычисления, MEC. MEC предлагает совместное размещение вычислительных ресурсов и ресурсов хранения данных на базовых станциях сотовых сетей [35]. Согласно [36], MEC является эволюцией мобильных базовых станций, представляющей собой совместное развертывание телекоммуникационных и ИТ-сетей.

MEC можно подключить к облачным центрам обработки данных в удаленном месте. Следовательно, MEC поддерживает двух- или трехуровневое иерархическое развертывание приложений вместе с оконечными мобильными устройствами [36]. В системе MEC новое устройство, называемое сервером MEC, необходимо развернуть рядом с базовыми станциями, чтобы обеспечить возможность обработки и хранения данных на периферии вычислений.

В) Периферийные, или граничные вычисления, EC. В исследовании [23] “периферия” определяется как любая сеть или вычислительный ресурс вблизи пути передачи данных между облачными центрами обработки данных и источниками данных. Любое смарт-устройство или датчик могут иметь источники данных, но “периферии” у них разные. Например, облачное хранилище и микроцентр обработки данных — это “периферия” между мобильным приложением и облаком, тогда как шлюз IoT — это “периферия” между датчиками IoT и облаком. Точно так же, если облачное приложение работает на смартфоне, то смартфон является “периферией” приложения и облака [37]. Основная цель периферийных вычислений заключается в том, что вычисления должны выполняться ближе к источникам данных.

Г) Росистые вычисления, DC. В рассматриваемой иерархии вычислений DC [38] находится на нижнем уровне среды облачных и туманных вычислений [39]. DC выходит за рамки концепции сервиса, хранилища и сети и превращается в подплатформу, основанную на концепции микросервиса, для которой его вычислительная иерархия распределяется по вертикали [39]. Подход DC облегчает

использование таких ресурсов, как датчики, планшеты и смартфоны, которые легко подключаются к сети. Из-за этого DC охватывает широкий спектр сетевых технологий на основе связи типа ad-hoc [39].

Д) Туманно-росистые вычисления, FDC. В архитектуре туманно-росистых вычислений устройствам IoT не требуется активное подключение к Интернету при подключении к серверу FDC. Сервер FDC будет взаимодействовать с облаком и отвечать за предоставление услуг устройствам IoT [41]. Облачные вычисления всегда требуют подключения к Интернету, что является основным недостатком облака. В то время как облако не может обслуживать пользователей без подключения к Интернету, туманно-росистые вычисления облегчают работу в автономном режиме без подключения к Интернету.

Перейдем теперь к наиболее распространенным сценариям использования механизмов выгрузки задач мобильных вычислений в систему туманно-облачных вычислений, которые выделяют на данный момент.

Некоторым приложениям требуется инфраструктура туманных вычислений для обеспечения бесперебойного обслуживания. К ним относятся интеллектуальные транспортные системы, дополненная (*англ.* augmented reality, AR) и виртуальная реальность (*англ.* virtual reality, VR), здравоохранение, потоковое видео, умные дома (*англ.* smart house) и умные города (*англ.* smart city). Требования к платформе и приложениям также необходимы для предоставления услуг. Ниже приведен обзор исследовательских работ, посвященных применению туманных вычислений.

А. Умная транспортная система. Было проведено несколько исследовательских работ по интеллектуальным транспортным системам, использующим туманные вычисления. В [14] автор предложил архитектуру автомобильных самоорганизующихся сетей (*англ.* Vehicular ad-hoc network, VANET), которая объединяет программно-конфигурируемые сети (*англ.* Software Defined Networking, SDN) и туманные вычисления. Поскольку SDN обладает функциями программирования, гибкости, глобальных знаний и масштабируемости, а преимуществами туманных вычислений являются привязка к

местоположению и быстрый отклик, сочетание этих двух факторов поможет решить ключевые проблемы VANET. Предлагаемая система способна улучшить связь между транспортными средствами, инфраструктурой и базовыми станциями посредством централизованного управления, уменьшить задержки и оптимизировать использование ресурсов. Исследование сетей VANET в туманных вычислениях также проводилось в [15]. В работе исследовалось моделирование приложений, но не рассматривались другие аспекты приложений, инфраструктура или платформа.

Б. Транспортные средства в инфраструктуре туманных вычислений. В [16] авторы предложили идею автомобильных туманных вычислений (*англ.* vehicular fog computing, VFC), которые будут использовать транспортное средство в качестве инфраструктуры для вычислений и связи. Архитектура VFC использует вычислительные ресурсы транспортных средств, предоставляя услуги периферийным устройствам, расположенным рядом с ними. Они объединят ресурсы каждого движущегося автомобиля, что позволит повысить качество обслуживания.

В. Дополненная и виртуальная реальность. Приложения дополненной реальности очень чувствительны ко времени; небольшая задержка может привести к серьезным ошибкам при использовании. Таким образом, решения на основе туманных вычислений будут иметь большой потенциал в этой области [21]. Эти утверждения также применимы к играм виртуальной реальности, или VR-играм. В [17] авторы предложили игру с расширенным взаимодействием мозга и компьютера, в которой использовались инфраструктура туманных и облачных вычислений. Туманные вычисления выполняли анализ в реальном времени, такой как обработка сигналов, которая должна классифицировать состояние мозга, и другие анализы, обновленные из облака.

Г. Здравоохранение. Подход туманных вычислений также позволяет предоставлять медицинские услуги в реальном времени. В [18] авторы предложили системную архитектуру с использованием туманных вычислений для системы здравоохранения. Система будет обрабатывать данные, полученные от датчиков,

оценивать их и сообщать о необходимости какой-либо неотложной медицинской помощи. Другая архитектура туманных вычислений для задач здравоохранения была предложена в [20]. Данная работа в основном сосредоточена на задержке сети, энергопотреблении и оптимизации связи в медицинских услугах на основе туманных вычислений.

Д. Умные города. Приложениям, связанным с умным городом, необходимо обрабатывать данные датчиков в режиме реального времени, где туманные вычисления могут играть важную роль. В [19] авторы предложили архитектуру, которая поддерживает различные приложения в умном городе. Предложенная структура оценивает три приложения умного города, включая картографирование шумового загрязнения, городские дренажные сети и умную улицу.

Вышеуказанные работы являются типичными примерами применения туманных вычислений в том смысле, что они выполняют критический по времени анализ на границе сети и при этом устойчивы к задержкам во времени при отправке в облачные вычисления. В данном контексте, туманные вычисления можно назвать расширением облачных вычислений.

Теперь рассмотрим некоторые существующие исследования по туманным вычислениям, анализирующие распределение ресурсов в туманно-облачной инфраструктуре. Они включают распределение ресурсов и планирование, обработку сбоев, инструменты моделирования и микро-сервисы.

В [22] авторы предложили методы распределения ресурсов для совместной платформы, состоящей из систем туманных и облачных вычислений. В работе был предложен алгоритм для балансировки рабочей нагрузки с учетом емкости виртуальных машин, времени завершения обработки запроса и объема вычислений для балансировки рабочей нагрузки. В [23] была представлена модель распределения рабочей нагрузки в среде туманных и облачных вычислений для нахождения компромисса между энергопотреблением и задержкой передачи. Однако сложный характер нагрузки и ресурсов в этой работе не изучался.

В [24] авторы разработали инструмент, удовлетворяющий требованиям к оборудованию, программному обеспечению и качеству обслуживания перед развертыванием приложения в инфраструктуре туманных вычислений, и учитывающий только качество обслуживания канала связи. Однако доступность и время ожидания более важны в среде туманных вычислений, нежели рассмотренные потребление ресурсов и каналы связи.

Для обеспечения эффективного использования ресурсов и сетевой инфраструктуры в среде туманных и облачных вычислений авторы статьи [25] предложили алгоритм для эффективного развертывания приложений интернета вещей в инфраструктуре туманно-облачных вычислений. В работе учитывались оперативная память и пропускная способность для распределения ресурсов в системе. Однако рассматривая только с этой точки зрения, использование ресурсов облачных вычислений всегда будет наиболее эффективным решением, поэтому необходимо учитывать и другие параметры, такие как время отклика и доступность указанных ресурсов.

В [26] был изучен сценарий потоковой передачи больших данных с помощью туманных вычислений, которые отвечают за предварительную обработку данных для приложений, размещенных на удаленном облаке и чувствительных ко времени отклика. Данный сценарий позволяет минимизировать вычислительную нагрузку за счет динамического распределения устройств туманных вычислений, облака и SDN с использованием обмена сообщениями, и определяет объем данных, который будет направлен на устройство туманных вычислений. В работе учитывались потеря ценности информации из-за предварительной обработки и стоимость эксплуатации туманных и облачных вычислений. В [27] авторы предложил алгоритм динамической оценки ресурсов путем интеграции записи истории потребителя облачных услуг в среду туманных вычислений, благодаря которому можно свести к минимуму недоиспользование ресурсов.

Подведем итоги. Говоря об актуальности исследований, туманные вычисления являются хорошим решением для критичных ко времени отклика ресурсоемких приложений. Туман позволяет мобильным устройствам выгрузить

некоторые задачи с мобильного устройства и, как следствие, это приводит к увеличению времени автономной работы мобильного устройства.

Отметим, что главное отличие туманных вычислений от облачных заключается в приближенной структуре к конечному потребителю, что позволяет существенно ускорить время отклика.

Кроме того, в настоящее время постоянно увеличивающееся число одновременно работающих приложений на мобильных устройствах увеличивает интенсивность вычислений, что также приводит к увеличению потребления энергии. Однако эти устройства имеют ограничения по производительности вычислений и источникам питания, особенно по сравнению с серверными машинами. Таким образом выгрузка в распределенную вычислительную среду туманно-облачных вычислений позволяет сэкономить заряд аккумулятора беспроводных устройств.

Перейдем ко второму типу задач выгрузки – выгрузка трафика из лицензированного спектра частот, традиционно используемого мобильными операторами, в нелицензированный для увеличения емкости сети.

С развитием сетей связи и информационных технологий спрос пользователей на услуги беспроводных сетей начал быстро расти, в результате чего также возросли требования к скорости и объему передачи информации. Поэтому научные и промышленные сообщества ищут новые методы и технологии для решения этого вопроса, одним из которых стало использование нелицензированного спектра для выгрузки трафика данных из лицензированных полос. На предыдущих этапах развития беспроводных сетей связи было предложено использовать различные планировщики распределения трафика в лицензированных и нелицензированных полосах для базовых станций, использующих технологию связи четвертого поколения LTE-U (*англ.* LTE-Unlicensed) и WiFi.

В 2016 году 3GPP представила концепцию технологий 5G в нелицензированном спектре (5G-U), которая получила название Новое радио в нелицензированном спектре (*англ.* New Radio Unlicensed, NR-U) в спецификациях TR 38.889 [37] и TR 38.716 [38]. По мере продвижения работы над спецификациями

технологии NR (*англ.* New Radio, NR), стандарты NR-U также развивались. Как и в технологии LAA (*англ.* License Assisted Access, LAA), основными “строительными блоками” для NR-U стали возможность агрегации несущих и динамический выбор частоты. Чтобы обеспечить справедливое сосуществование с технологиями, использующими нелицензированные полосы частот, например, с технологией WiGig (*англ.* Wireless Gigabit, стандарты IEEE 802.11ad/ay), процедуры случайного доступа к каналам были выделены в TR 38.889 [37] как критически важные функции.

Интеграция лицензированных и нелицензированных полос частот миллиметрового диапазона впервые была предложена в работах [42], [43]. В [42] авторы изучают сосуществование обеих систем с точки зрения скорости передачи данных по нисходящей линии связи, сравнивая три различных сценария: только WiGig, только 5G-U и сосуществование 5G-U с WiGig. Результаты показали, что рекомендуется использовать лицензированный сигнал в нелицензированном диапазоне частот, 5G-U может сосуществовать с WiGig, и это хорошее “соседство” с существующими сетями. Поскольку работа в одной и той же нелицензированной полосе частот может создавать помехи, в исследованиях [45], [46] был описан механизм сосуществования 5G-U и WiGig, динамический выбор частоты (*англ.* dynamic frequency selection, DFS) и т. д. В работе [46] были описаны некоторые аспекты проектирования и архитектурные детали NR для поддержания совместного использования спектра в нелицензированной полосе.

Литературу по анализу производительности технологии NR-U можно разделить на два направления. В первом исследуются механизмы сосуществования между NR-U и нелицензированными технологиями в одном частотном спектре. На данный момент предложено два подхода для реализации механизма: (i) подходы, основанные на рабочем цикле (англ. Duty cycle), и (ii) механизмы, основанные на процедуре прослушивания перед разговором (*англ.* Listen Before Talk, LBT).

Сосуществование технологий Wi-Fi и LTE-U и применение метода рабочего цикла были исследованы в нескольких работах [48], [49]. В [50] авторы исследуют производительность Wi-Fi при наличии передачи LTE-U на основе рабочего цикла

по тому же каналу. Авторы описывают математическую модель системы с учетом рабочего цикла с использованием цепи Маркова. В [49] авторы предложили несколько схем сосуществования для LTE и точки доступа Wi-Fi в качестве возможные улучшения в будущих 5G сетях. В [131] авторы представляют аналитическую модель, которая учитывает специфику системы LAA (*англ.* Licensed-Assisted Access) на основе рабочего цикла с контролем допуска соединения (*англ.* connection admission control, CAC), механизмом перераспределения ресурсов, и учетом эластичности трафика.

Согласно TR 38.889 [37], существует 4 возможных схемы доступа к каналу в нелицензированном спектре. Первая схема подразумевает немедленную передачу после короткого периода переключения, которая используется для того, чтобы передатчик передал сразу после коммутационного промежутка внутри времени занятости канала (*англ.* channel occupancy time, COT); Вторая схема подразумевает механизм LBT без случайного таймера обратного отсчета, который использует постоянный интервал времени в режиме ожидания перед передачей. Третья схема также использует LBT, но со случайным таймером обратного отсчета и фиксированным размером конкурентного окна (*англ.* contention window, CW). Перед тем, как размер CW выбирается случайным образом, этот размер интерпретируется как период времени для ожидания перед передачей. В диссертационной работе используется четвертый способ, LBT со случайным таймером обратного отсчета и переменным размером CW. Как и в предыдущем подходе, размер CW выбирается случайным образом, но этот параметр может быть изменен передающим объектом во время его цикла передачи.

В [44] исследуется механизм LBT, первоначально предложенный в 15 релизе спецификации TR 38.889 [37] консорциума 3GPP, и показывающий, что он может не полностью соответствовать требованиям сосуществования технологий NR-U и WiGig (стандарта IEEE 802.11ad) при реалистичной реализации NR. В [45] авторы предлагают метод прослушивания до получения (*англ.* Listen Before Receive, LBR) для совместного доступа к спектру и анализируют его возможности для обеспечения справедливого сосуществования нескольких технологий

радиодоступа (англ. Radio Access Technology, RAT) в нелицензированных миллиметровых диапазонах.

В работе [52] авторы разрабатывают теории эффективной емкости для измерения LAA в соответствии со статистическими требованиями качества обслуживания, они предлагают полумарковскую модель с четырьмя состояниями, разработанную для описания возникновения коллизии во время передачи и блокировки пути прямого распространения сигнала (*англ.* Line of Sight, LoS). Модель с четырьмя состояниями далее абстрагируется до упрощенного варианта модели с двумя состояниями. В диссертационной работе будет использоваться концепция успешной и неуспешной передачи данных с учетом вероятностей возникновения коллизии передачи и блокировки пути прямого распространения сигнала, а также методы расчета эффективной емкости системы.

В [54] авторы продолжили исследования, описанные в [52], рассмотрев взаимодействие NR-U и Wi-Fi, и разработали новый протокол LBT, относящийся к кооперативному LBT, в котором применяется форсирование нуля и предварительное кодирование для подавления многопользовательских помех.

Дальнейшее расширение схемы LBT для технологии NR-U связано с исследованием в [47], в котором была предложена гибридная схема, в которой обычная процедура LBT из TR 38.889 [37] используется в условиях перегруженного канала, в то время как специальный метод доступа не используется, когда общий нелицензированный канал относительно свободный. Проведенное моделирование показало, что этот подход потенциально обеспечивает гораздо меньшую задержку и пропускную способность. Оптимизация производительности порога обнаружения несущей для обычного LBT из TR 38.889 выполнена в [48]. Наконец, обзор и сравнение различных предложений доступа к каналу NR-U представлены в [49].

Второе направление исследований в контексте NR-U, которому на данный момент уделялось гораздо меньше внимания, связано с вопросами распределения ресурсов и производительностью пользователей в среде NR-U. Среди прочего, авторы в [50] рассмотрели применение технологии NR-U к приложениям

расширенной виртуальной реальности на спортивных мероприятиях. Их результаты показывают, что при наличии справедливого сосуществующего механизма могут быть достигнуты улучшения производительности на системном уровне в показателях, ориентированных на пользователя и оператора. Аналитическая модель оценки производительности системы из [50] была разработана и представлена в [51]. Среди прочих результатов авторы обнаружили, что пропорциональное разделение трафика между лицензированным и нелицензированным миллиметровыми диапазонами приводит к лучшей производительности. Однако в своем исследовании авторы рассмотрели только одну БС NR-U, а также не включили в свою модель схему произвольного доступа в нелицензированном диапазоне. Несмотря на то, что основное внимание уделяется технологии LAA, в [52] и [53] авторы разработали и проанализировали простую схему на основе рабочего цикла для справедливого разделения ресурсов между LTE-U и Wi-Fi, работающими в нелицензированном диапазоне.

Агрегация несущих впервые появилась в 8-м релизе 3GPP. Это позволило операторам более эффективно использовать свои частоты. Самый простой способ организовать агрегацию — это использовать смежные компонентные несущие в одной и той же рабочей полосе частот, как это было определено для LTE, так называемые внутриполосные смежные несущие. Это не всегда возможно из-за сценариев распределения частот оператором. Несмежное распределение может быть либо внутриполосным, т. е. компонентные несущие принадлежат одной и той же рабочей полосе частот, но имеют промежутки или промежутки между ними, либо междиапазонным, в этом случае компонентные несущие относятся к разным рабочим полосам частот.

Когда используется агрегация несущих, имеется несколько обслуживающих сот, по одной на каждую компонентную несущую (*англ.* Component Carriers, CC). Покрытие обслуживающих сот может различаться, например, из-за того, что компонентная несущая в разных диапазонах частот будет иметь разные потери в тракте, см. рисунок 1.3. Агрегация в LAA осуществляется с использованием первичных и вторичных сот. Первичные соты, работающие в лицензированном

диапазоне частот, используются для обеспечения передачи критической информации и гарантирования качества обслуживания (*англ.* Quality of Service, QoS). Вторичные соты, которые работают в нелицензированном диапазоне частот, увеливают скорость передачи данных за счет агрегации с первичными ячейками.

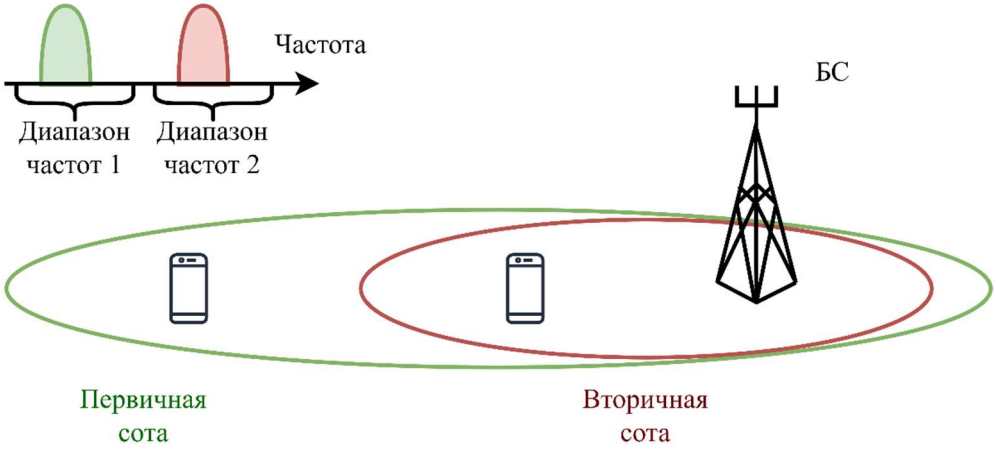

Рис. 1.3. Агрегация несущих. Первичные и вторичные соты

Рис. 1.4. Агрегация несущих. Лицензированный и нелицензированный спектр.

Таким образом при построении математической модели системы стоит учесть следующие особенности.

- Возможность использовать агрегацию несущих для использования разных частот в процессе передачи данных (рис. 1.4).
- Возможность использовать нелицензированный спектр для создания соединения с БС в нелицензированном спектре (аналогично рекомендациям LTE-U / LAA) во время передачи и снизить нагрузку на лицензированный спектр (рис.1.4).
- В NR-U лицензированные операторы объединяются с нелицензированными, чтобы увеличить пропускную способность нисходящего канала для пользователей, предлагая бесшовную поддержку мобильности.
- Улучшения должны включать протоколы связи, основанные на механизме прослушивания перед разговором (LBT).

– Возможность подключения к БС устройств, которые не могут использовать нелицензированный спектр (рис. 1.5).

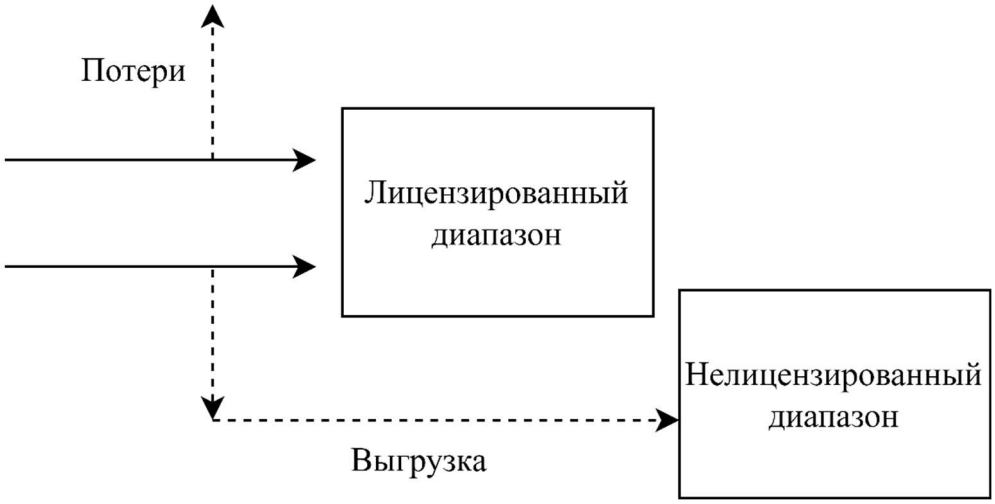

Рис. 1.5. Принцип выгрузки сессий в NR-U.

1.2. Модель времени отклика в системе туманных вычислений

Рассматривается схема выгрузки мобильных вычислений, которая состоит из узла туманных вычислений и удаленной системы облачных вычислений. Мобильные устройства запускают приложения (приложения для финансовой торговли в реальном времени, игровые приложения, приложения виртуальной реальности и т.д. [45]), которые потребляют большое количество вычислительных ресурсов и мощности. Согласно внутренней политике мобильного устройства, часть задач выгружается в узел туманных вычислений. Узел туманных вычислений имеет ограниченную емкость, поэтому может произойти перегрузка. В этом случае выгруженная задача отправляется на удаленный узел облачных вычислений.

Цель состоит в том, чтобы минимизировать энергопотребление мобильных устройств, сохраняя жесткие ограничения на время отклика. На рисунке 1.6 показан пример выгрузки вычислений в сети, состоящей из M мобильных устройств с небольшой вычислительной мощностью и ресурсом аккумуляторной батареи, узла M+1 туманных вычислений и узла M+2 облачных вычислений. Случайная величина (с.в.) времени отклика $\beta_i, 1 \le i \le M$ при решении задачи на узле i на верхней части рисунка соответствует времени обработки не выгруженных задач. В этом случае мобильное устройство потребляет много энергии, но имеет

наименьшее время отклика, которое состоит только из времени обработки, $\beta_i = t_i$. Время отклика β_i на средней части рисунка соответствует выгрузке задачи в узел туманных вычислений. В этом случае мобильное устройство тратит энергию на отправку задачи на узел туманных вычислений по беспроводному каналу, но экономит энергию на вычислениях. Полученное время отклика является суммой задержки передачи Δ_1 и времени обработки t_{M+1} в узле туманных вычислений: $\beta_i = 2\Delta_1 + t_{M+1}$. Время отклика β_i на нижней части рисунка соответствует случаю, когда выгруженная задача не может быть обработана в узле туманных вычислений и отправляется на удаленное облако. В этом случае потребление энергии мобильным устройством такое же, как в предыдущем случае, но время отклика намного больше, поскольку добавляется задержка передачи Δ_2 на узел облачных вычислений: $\beta_i = 2\Delta_1 + 2\Delta_2 + t_{M+2}$.

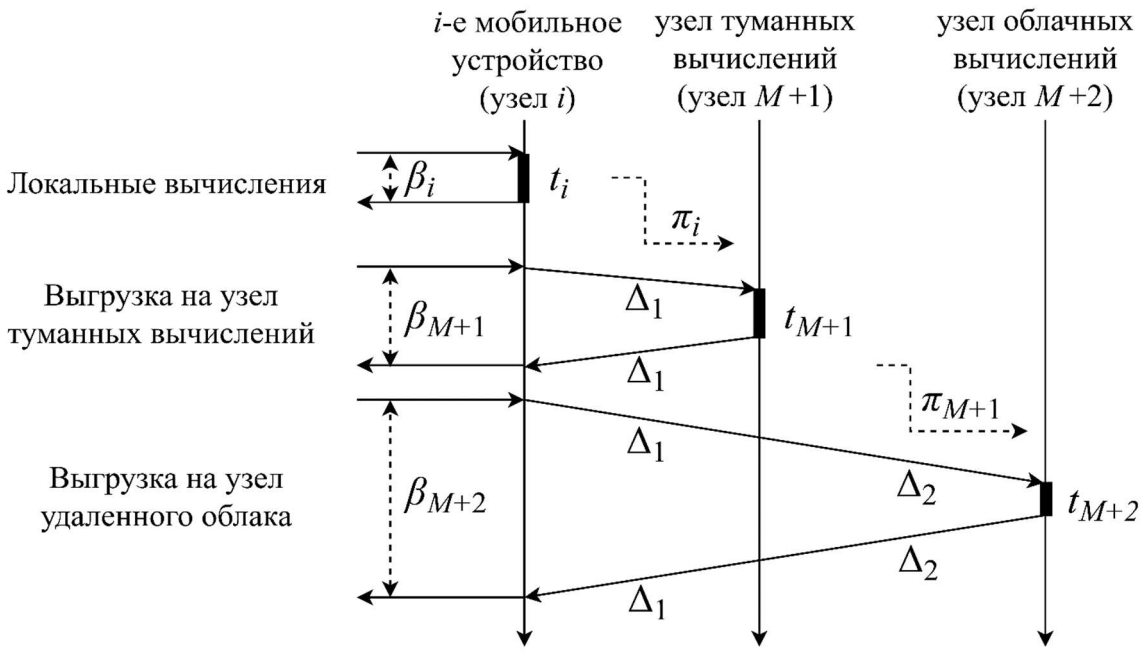

Рис. 1.6. Пример выгрузки мобильных вычислений.

Предлагаемая вычислительная схема выгрузки может быть описана в терминах сетей массового обслуживания. Пусть M - это число мобильных устройств в зоне действия узла туманных вычислений. Мобильное устройство, обозначенное как i-узел, $1 \le i \le M$, генерирует поток заявок, который считается пуассоновским потоком с интенсивностью $\lambda_i, 1 \le i \le M$ (заявок в секунду). Каждая

задача требует определенного объема вычислений (измеряется в операциях или операциях с плавающей запятой). Пусть $w_i, 1 \le i \le M$, с.в. объема вычислений, который должен быть выполнен для задачи на *i*-м мобильном устройстве с ФР $W_i(x) = P\{w_i \le x\}$. Пусть μ_i постоянная скорость обработки задач (инструкций в секунду) на *i*-м мобильном устройстве. Задача выгружается на узел туманных вычислений, если $w_i > w^*$, где w^* – порог объема вычислений, и обрабатывается локально на мобильном устройстве в противном случае. Вероятность выгрузки *i*-го мобильного устройства определяется как $\pi_i = 1 - W_i(w^*)$.

С учетом обозначений, описанных выше, математическую модель системы можно представить, как показано на рисунке 1.7.

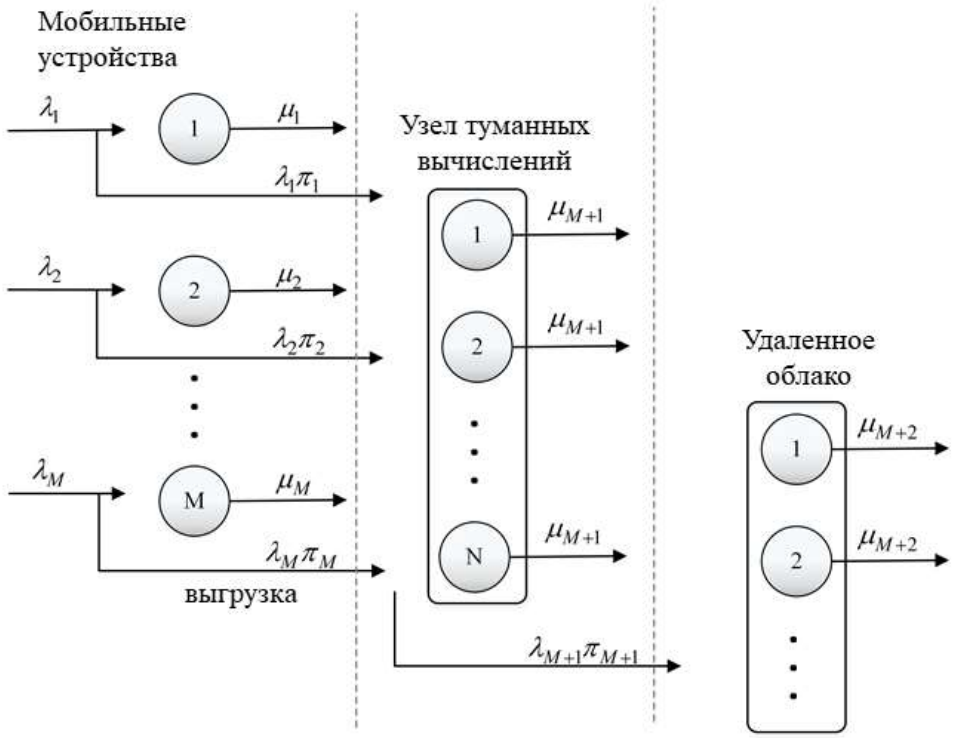

Рис. 1.7. Схема модели выгрузки мобильных вычислений.

Далее проводится вычисление среднего времени отклика и обработки задачи на мобильном устройстве, в узле туманных вычислений и узле облачных вычислений, соответственно.

A. Мобильные устройства

Пусть $p_{i,j}$ - вероятность того, что для обработки задачи с *i*-го узла мобильных вычислений потребуются *j* операций, $\sum_{j=0}^{+\infty} p_{i,j} = 1$. В этом случае ФР $W_i(x)$ равна

$$W_i(x) = \sum_{0 \le j < x} p_{i,j}. \tag{1.1}$$

ФР $V_i(x)$ объема вычислений для задачи, которая обрабатывается локально на *i*-м мобильном устройстве, определяется как

$$V_i(x) = \begin{cases} \dfrac{W_i(x)}{1-\pi_i}, x \le w^*; \\ 1, x > w^*. \end{cases}$$

После подстановки формулы (1.1) для ФР $W_i(x)$ формула будет иметь вид

$$V_i(x) = \begin{cases} \dfrac{1}{1-\pi_i} \sum\limits_{0 \le j < x} p_{i,j}, x \le w^*; \\ 1, x > w^*. \end{cases} \tag{1.2}$$

Время обработки на *i*-м мобильном устройстве определяется как отношение объема вычислений и скорости обработки задач λ_i, а его ФР $T_i(x)$ определяется как

$$T_i(x) = \begin{cases} V_i(\mu_i x), x \le \dfrac{w^*}{\mu_i}; \\ 1, x > \dfrac{w^*}{\mu_i}; \end{cases}$$

После подстановки формулы (1.2) для ФР $V_i(x)$ объема вычислений конечная формула ФР $T_i(x)$ имеет вид

$$T_i(x) = \begin{cases} \dfrac{1}{1-\pi_i} \sum\limits_{0 \le j < x} p_{i,j}, x \le \dfrac{w^*}{\mu_i}; \\ 1, x > \dfrac{w^*}{\mu_i}, \end{cases} \tag{1.3}$$

где $\frac{1}{1-\pi_i} p_{i,j}$ для $j \le \frac{w^*}{\mu_i}$ - вероятность того, что время обработки задачи на *i*-м мобильном устройстве равно $\frac{j}{\mu_i}$. Пусть t_i - с.в. времени обработки локально обрабатываемой задачи на *i*-м мобильном устройстве. Таким образом, среднее время отклика *i*-го мобильного устройства равно

$$t_i^{(1)} = \frac{1}{\mu_i} \sum_{j=0}^{w^*} j \frac{1}{1-\pi_i} p_{i,j} = \frac{1}{1-\pi_i} \frac{1}{\mu_i} \sum_{j=0}^{w^*} j p_{i,j}. \quad (1.4)$$

Б. Узел туманных вычислений

Все задачи со всех мобильных устройств с $w_i > w^*$ отправляются в узел туманных вычислений. Задачи поступают на узел туманных вычислений в соответствии с пуассоновским процессом с интенсивностью $\lambda_{M+1} = \sum_{i=0}^{M} \lambda_i \pi_i$, распределение объема вычислений заявок на *i*-м мобильном устройстве определяется ФР $V_{M+1,i}(x)$, где

$$V_{M+1,i}(x) = P\{w_i \le x \mid w_i > w^*\}. \quad (1.5)$$

Из определения очевидно, что $V_{M+1,i}(x) = 0$ для всех $j \le w^*$. С другой стороны, если $j > w^*$, то $V_{M+1,i}(x) = \frac{1}{1-\pi_i} \sum_{0 \le j < x} p_{i,j}, j \ge w^* + 1$. Тогда ФР имеет вид

$$V_{M+1,i}(x) = \begin{cases} 0, x < w^* + 1; \\ \sum_{w^*+1 \le j \le x} p_{M+1,i,j}, x \ge w^* + 1. \end{cases} \quad (1.6)$$

Пусть μ_{M+1} - постоянная скорость обработки задачи на виртуальной машине (ВМ) на узле туманных вычислений, *N* - число виртуальных машин. Тогда распределение времени обработки t_{M+1} задачи на узле туманных вычислений определяется ФР $T_{M+1}(x)$, которую можно найти аналогично (1.3), и которая имеет вид

$$T_{M+1,i}(x)=\begin{cases}0, x<\dfrac{w^*+1}{\mu_{M+1}};\\ \dfrac{1}{\pi_i}\displaystyle\sum_{w^*\le j\le\mu_{M+1}x} p_{i,j}, x\ge\dfrac{w^*+1}{\mu_{M+1}};\end{cases} \tag{1.7}$$

$$T_{M+1}(x)=\begin{cases}0, x<\dfrac{w^*+1}{\mu_{M+1}};\\ \displaystyle\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_{M+1}}\sum_{w^*\le j\le\mu_{M+1}x} p_{i,j}, x\ge\dfrac{w^*+1}{\mu_{M+1}};\end{cases} \tag{1.8}$$

т. е. вероятность того, что задача обрабатывается $\dfrac{j}{\mu_{M+1}}$ секунд, равна $\displaystyle\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_{M+1}}p_{i,j}$.

Тогда среднее время обработки задачи в узле туманных вычислений с любого мобильного устройства равно

$$t_{M+1}^{(1)}=\frac{1}{\mu_{M+1}}\sum_{j=w^*+1}^{\infty} j\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_{M+1}}p_{i,j}. \tag{1.9}$$

Вероятность того, что задача будет перенаправлена в удаленный узел облачных вычислений из-за перегрузки узла туманных вычислений, вычисляется как вероятность блокировки в системе типа M/G/1/N и равна

$$\pi_{M+1}=\left(\sum_{k=0}^{N}\frac{\left(\lambda_{M+1}t_{M+1}^{(1)}\right)^k}{k!}\right)^{-1}\frac{\left(\lambda_{M+1}t_{M+1}^{(1)}\right)^N}{N!}. \tag{1.10}$$

B. Узел облачных вычислений

Интенсивность поступления заявок в удаленный узел облачных вычислений составляет $\lambda_{M+1}\pi_{M+1}$, поскольку все задачи, заблокированные на узле туманных вычислений, перенаправляются в узел облачных вычислений. Пусть распределение рабочей нагрузки задачи в узле облачных вычислений такое же, как и в узле туманных вычислений, тогда распределение времени обработки t_{M+2} на удаленном узле облачных вычислений определяется ФР $T_{M+2}(x)$, которая имеет вид

$$T_{M+2,i}(x)=\begin{cases}0, x<\dfrac{w^*+1}{\mu_{M+2}};\\ \dfrac{1}{\pi_i}\sum\limits_{w^*\le j\le\mu_{M+2}x} p_{i,j}, x\ge\dfrac{w^*+1}{\mu_{M+2}};\end{cases} \tag{1.11}$$

$$T_{M+2}(x)=\begin{cases}0, x<\dfrac{w^*+1}{\mu_{M+1}};\\ \sum\limits_{i=1}^{M}\dfrac{\lambda_i}{\lambda_{M+1}}\sum\limits_{w^*\le j\le\mu_{M+2}x} p_{i,j}, x\ge\dfrac{w^*+1}{\mu_{M+2}};\end{cases} \tag{1.12}$$

то есть вероятность того, что задача обрабатывается в течение $\dfrac{j}{\mu_{M+2}}$ секунд, составляет $\sum\limits_{i=1}^{M}\dfrac{\lambda_i}{\lambda_{M+1}}p_{i,j}$. Тогда среднее время обработки задачи с любого мобильного устройства в узле облачных вычислений равно

$$t_{M+2}^{(1)}=\frac{1}{\mu_{M+2}}\sum_{j=w^*+1}^{\infty} j\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_{M+1}}p_{i,j}. \tag{1.13}$$

Теперь сформулируем задачу оптимизации и приведём численные результаты анализа полученных ранее характеристик.

Одна из основных целей концепции туманных вычислений – это снизить нагрузку и энергопотребление мобильных устройств. В терминах модели это означает максимизацию вероятности π_i выгрузки с *i*-го мобильного устройства и полной вероятности выгрузки π, которая равна

$$\pi=\sum_{i=1}^{M}\frac{\lambda_i}{\lambda_\bullet}\pi_i \text{, где } \lambda_\bullet=\sum_{i=1}^{M}\lambda_i. \tag{1.14}$$

Однако с ростом вероятности выгрузки π нагрузка узла туманных вычислений также увеличивается, и это может привести к быстрому росту вероятности перегрузки π_{M+1} узла туманных вычислений и, как следствие, к резкому увеличению времени отклика.

Пусть t_{M+2} - с.в. времени обработки в узле облачных вычислений, β_i - с.в. времени отклика на задачу с *i*-го мобильного устройства и β - общее время

отклика. Считая, что задержки передачи постоянны, распределение вероятностей β_i будет таким, как показано в таблице 1.1.

Таблица 1.1. Распределение вероятностей для с.в. времени отклика β_i

β_i	t_i	$2\Delta_1 + t_{M+1}$	$2\Delta_1 + 2\Delta_2 + t_{M+2}$
P	$1-\pi_i$	$\pi_i\,(1-\pi_{M+1})$	$\pi_i\,\pi_{M+1}$

Таким образом, можно определить среднее время отклика $\beta_i^{(1)}$ задачи с *i*-го мобильного устройства и время отклика $\beta^{(1)}$ произвольной задачи.

Утверждение 1.1. ФР времени отклика *i*-го узла имеет вид

$$B_i(x) = (1-\pi_i)T_i(x) + \pi_i(1-\pi_{M+1})T_{M+1,i}(x-2\Delta_1) + \pi_i\pi_{M+1}T_{M+2,i}(x-2\Delta_1-2\Delta_2). \quad (1.15)$$

Следствие 1.1. Среднее время отклика *i*-го узла имеет вид

$$\beta_i^{(1)} = t_i^{(1)}(1-\pi_i) + (2\Delta_1 + t_{i,M+1})\pi_i\,(1-\pi_{M+1}) + (2\Delta_1 + 2\Delta_2 + t_{M+2})\pi_i\,\pi_{M+1}. \quad (1.16)$$

Время отклика $\beta^{(1)}$ произвольной задачи равно

$$\beta^{(1)} = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_\bullet}\beta_i^{(1)}. \quad (1.17)$$

Помимо средних значений, для заявок, критичных ко времени отклика, еще одним важным показателем производительности является вероятность $B(t^*)$ того, что время отклика превысит заранее определенный порог t^*, то есть $B(t^*) = P\{\beta > t^*\}$. Вероятность выгрузки π задается формулами (2.13) и (2.1), среднее время отклика $\beta^{(1)}$ формулой (2.15), а $B(t^*)$ рассчитывается как

$$P\{\beta > t^*\} = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_\bullet}P\{\beta_i > t^*_1\}, \quad (1.18)$$

где вероятность превышения времени отклика на задачу с *i*-го мобильного устройства порогового значения определяется как

$$\begin{aligned} P\{\beta_i > t^*\} &= (1-\pi_i)(1-T_i(t^*)) + \pi_i\,(1-\pi_{M+1})(1-T_{M+1,i}(t^*-2\Delta_1)) + \\ &+\pi_i\,\pi_{M+1}(1-B_{M+2}(t^*-2\Delta_1+2\Delta_2)). \end{aligned} \quad (1.19)$$

Перейдем к численному анализу полученных характеристик. Предположим, что объем вычислений имеет пуассоновское распределение с параметром $\alpha = 100$.

Вероятность того, что для обработки задачи потребуются j операций будет одинаковым для каждого мобильного устройства и задается гамма-распределением с параметрами k_i и $\delta_i, i = 1,2$.

Предполагается, что в радиусе действия одного узла туманных вычислений находятся $M = 20$ мобильных устройств. Каждое мобильное устройство генерирует задачи с одинаковой интенсивностью $\lambda_i = 20, 1 \le i \le M$, и их скорости обработки также равны друг другу, $\mu_i = 1000$ операций в секунду, $1 \le i \le M$. Узел туманных вычислений обрабатывает задачи с помощью виртуальных машин, в этом узле $N = 10$ виртуальных машин, скорость обработки каждой составляет $\mu_{M+1} = 5000$ операций в секунду, в то время как скорости обработки в удаленном узле облачных вычислений равняется $\mu_{M+2} = 10000$. Наконец, предполагается, что задержка передачи между мобильным устройством и туманным узлом Δ_1 составляет 10 мс, а задержка передачи между туманным узлом и удаленным облаком Δ_2 составляет 100 мс.

На рисунке 1.8 показано, что вероятность выгрузки π убывает в зависимости от порога выгрузки w^*, что удовлетворяет логике построенной модели.

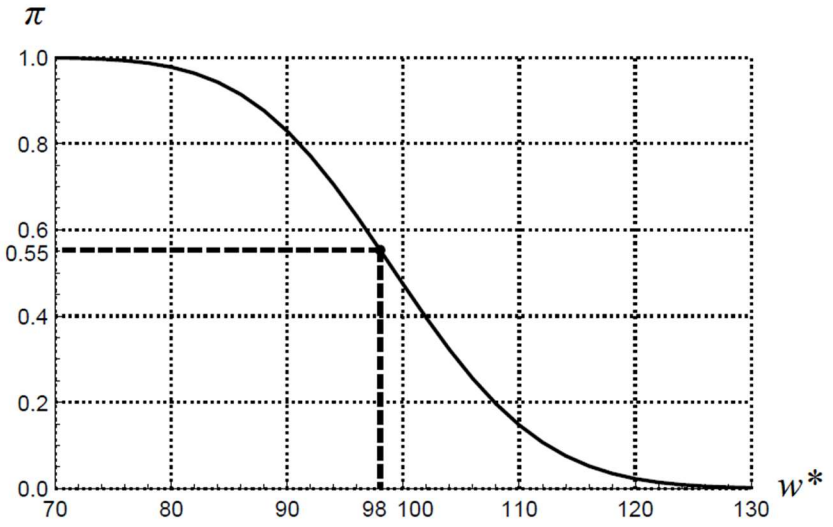

Рис. 1.8. Зависимость вероятности выгрузки π от порога выгрузки w^*.

На рисунке 1.9 показан график зависимости среднего времени отклика $\beta^{(1)}$ от порога выгрузки w^*. Для $\lambda_i = 20$, минимальное среднее время отклика

достигается при $w^* = 91$, в то время как в случае $\lambda_i = 30$, минимум достигается при $w^* = 98$. Когда w^* мало, почти все задачи выгружаются в узел туманных вычислений и он становится перегруженным. Следовательно, многие задачи перенаправляются в удаленный узел облачных вычислений со сравнительно большой задержкой передачи. С увеличением порога выгрузки w^* вероятность перегрузки узла туманных вычислений π_{M+1} становится меньше, что приводит к уменьшению среднего времени отклика. Однако в некоторый момент, после прохождения точки локального минимума, среднее время отклика начинает увеличиваться, потому что вероятность выгрузки π становится слишком малой, и большая часть заявок обрабатывается локально на мобильных устройствах со сравнительно небольшой емкостью. Другими словами, в точках локального минимума нагрузка оптимально сбалансирована между мобильными устройствами и узлом туманных вычислений в отношении среднего времени отклика.

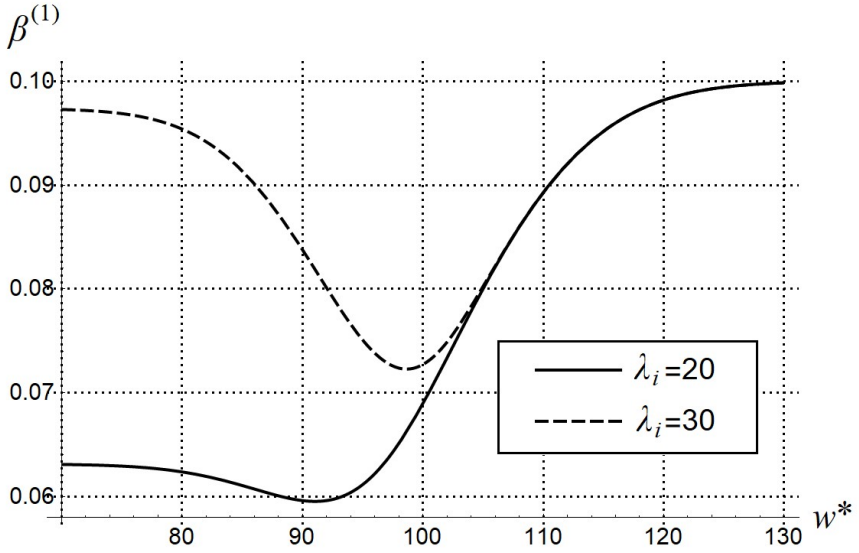

Рис. 1.9. Зависимость среднего времени отклика $\beta^{(1)}$ от порога вычислений w^* для выгрузки в узел туманных вычислений.

Теперь, чтобы найти оптимальный порог выгрузки при ограничениях на среднее время отклика $\beta^{(1)}$ и вероятность $B(t^*) = B(t^*, w^*)$ того, что время отклика превысит заданный предел, сформулируем задачу оптимизации:

$$\begin{cases} \pi(w^*) \to \max, \\ \text{R1}: \beta^{(1)}(w^*) \le \tau^*, \\ \text{R2}: B(w^*, t^*) \le B^*, \end{cases} \tag{1.20}$$

где τ^* и B^* - заранее определенные пороговые значения времени отклика и вероятности того, что время отклика превысит заранее определенный порог t^*, соответственно.

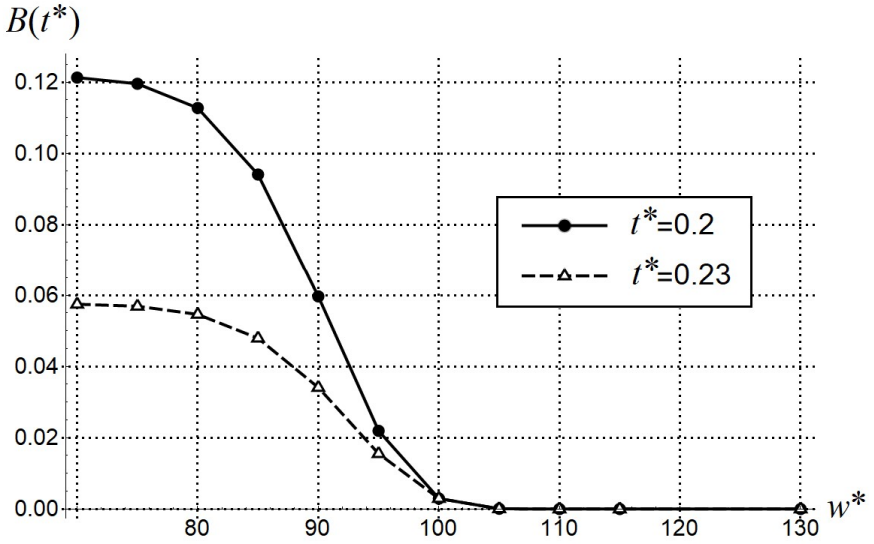

Рис. 1.10. Зависимость $B(t^*)$ от порога вычислений w^* для выгрузки в узел туманных вычислений.

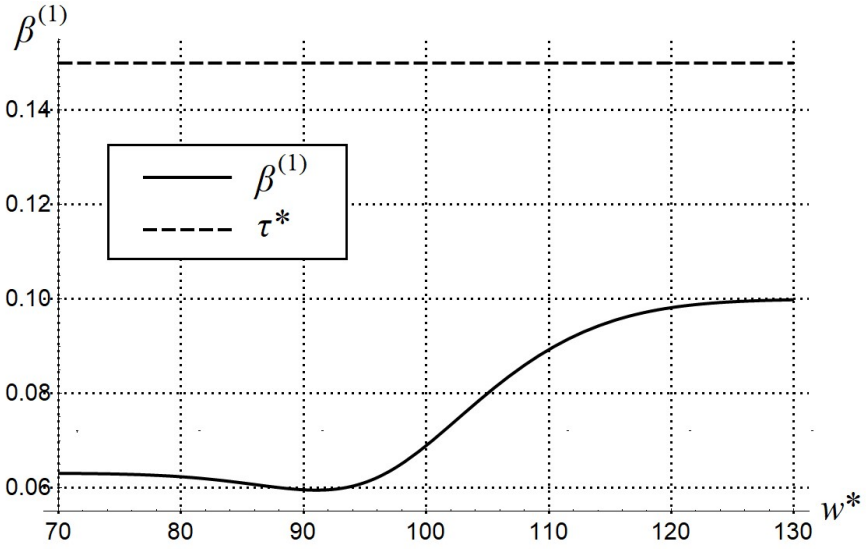

Рис. 1.11. Численный пример решения задачи оптимизации. Зависимость среднего времени отклика от порога выгрузки.

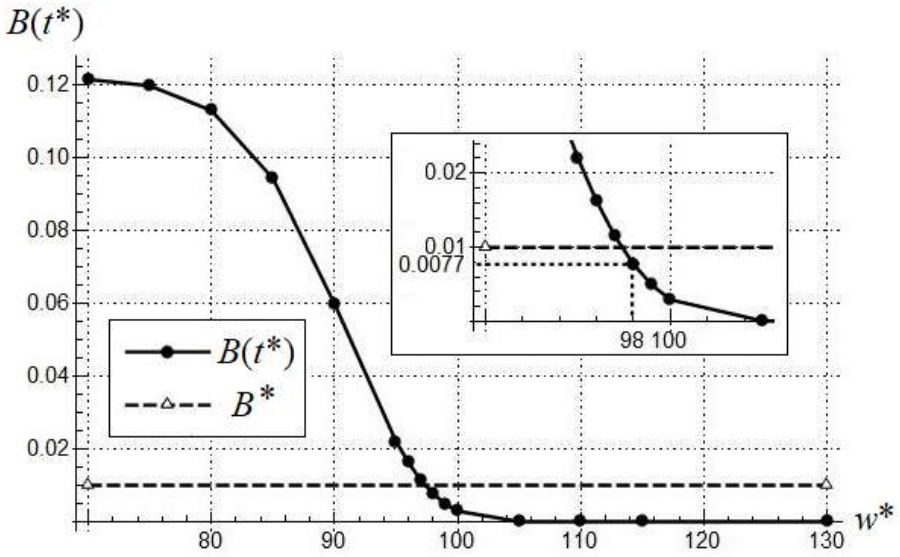

Рис. 1.12. Численный пример решения задачи оптимизации. Зависимость вероятности $B(t^*)$ от порога выгрузки

На рис. 1.11 и 1.12 представлен численный пример решения задачи оптимизации для пороговых значений $\tau^* = 0.15$ и B^*=0.1, на рис. 1.11 представлено ограничение на среднее время отклика и на рис. 1.12 представлено ограничение на вероятность того, что время отклика превысит заранее определенный порог t^*.

Оптимальное значение вероятности выгрузки π для таких пороговых значений равно 0.55316, которое достигается при $w^* = 98$, как показано на рис.1.3.

Подведем итоги, в данном разделе была разработана модель для анализа времени отклика, которая учитывает изменение задач с точки зрения объема вычислений. Отметим, что одной из целей выгрузки ресурсоемких задач является повышение срока автономной работы мобильных устройств. В базовой модели, описанной выше, этот вопрос не рассматривался, поэтому было решено рассмотреть модель, учитывающую энергопотребление и во второй главе диссертации энергопотреблению будет посвящен один подраздел. Также в модели необходимо рассмотреть дополнительный параметр - объем данных, которые должны быть переданы в случае выгрузки на узел туманных вычислений.

1.3. Анализ эффективности выгрузки трафика в нелицензированный диапазон частот

Рассматривается система с технологиями NR и WiGig, физически размещенными на одной и той же базовой станции (БС). Технология NR использует лицензированный диапазон 28 ГГц, используя канал $B_1 = 200$ МГц [55]. Предполагается, что технология WiGig работает в нелицензированном диапазоне 60 ГГц с использованием одного канала с полосой пропускания $B_2 = 160$ МГц [56]. Предполагается, что БС развернуты согласно точечному пуассоновскому процессу (*англ.* Poisson point process, PPP) на плоскости с плотностью $\lambda_{\bullet}$ БС на квадратный метр, см. рис. 1.13. Высота БС устанавливается равной $h_{\bullet}$.

В рассматриваемом развертывании есть два типа устройств: пользовательские устройства NR-U и пользовательские устройства WiGig. Устройства первого типа способны работать как в диапазонах NR, так и в диапазонах WiGig, а устройства WiGig используют только технологию WiGig. Расположение пользовательских устройств NR-U и WiGig определяется также в соответствии с точечным процессом Пуассона. Высота всех пользовательских устройств составляет h.

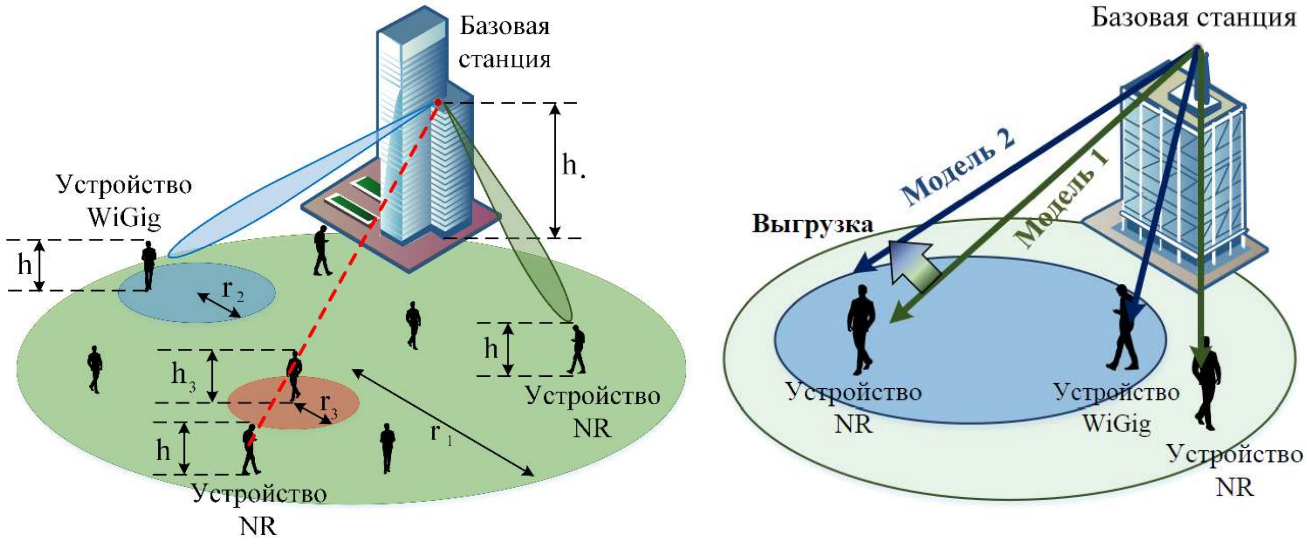

Рис. 1.13. Иллюстрация рассматриваемого сценария развертывания.

Все пользовательские устройства, работающие в нелицензированном диапазоне, используют подход прослушивания перед разговором (*англ.* Listen

Before Talk, LBT). Чтобы повысить скорость сессий NR, предполагается, что технология NR-U использует механизм LBT на основе наблюдения за каналом (*англ.* Channel Observation-Based Listen-Before Talk, CoLBT), основанный на концепциях конкурентного окна (*англ.* contention window, CW) и таймера обратного отсчета процедуры LBT на основе наблюдения за каналом (рис. 1.14), который аналогичен рекомендованному 3GPP для технологии четвертого поколения LAA [58]. Изначально размер CW установлен на 32 слота. *T* обозначает максимальное количество неудачных повторных передач при удвоении CW.

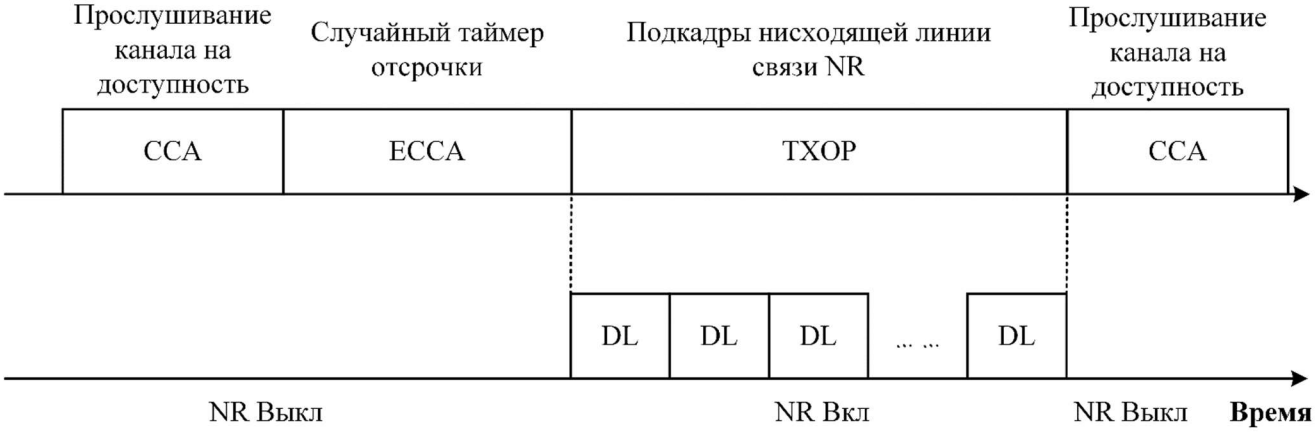

Рис. 1.14. Иллюстрация рассматриваемого механизма случайного доступа.

Пользовательские устройства, у которых есть пакет, готовый к передаче, выбирают таймер обратного отсчета в пределах (0, CW). Значение таймера обратного отсчета уменьшается на единицу на каждом такте времени, где определяется, что канал свободен. Если канал занят, пользовательское устройство приостанавливает таймер обратного отсчета и продолжает слушать канал. Когда значение таймера обратного отсчета достигает единицы, появляется возможность передачи (*англ.* transmit opportunity, TxOp), и пользовательское устройство передает свой пакет. Возможны три следующих исхода: (а) успешная передача, (б) неудачная передача из-за коллизии с другой передачей устройства NR-U или WiGig и (в) неудачная передача из-за блокировки пути прямого распространения сигнала (*англ.* Line of Sight, LoS).

На процедуру разрешения конфликтов NR и WiGig влияет направленность используемых антенных решеток. То есть коллизия может произойти только тогда,

когда пользовательские устройства, расположенные в одном и том же секторе покрытия WiGig, пытаются передавать в одно и то же время. Фактически, с точки зрения моделирования, это означает, что эффективное количество пользовательских устройств ограничено направленностью антенной решетки WiGig.

Основным интересующим показателем является вероятность возможной потери сессии пользовательского устройства NR-U, т. е. вероятность того, что сессия пользовательского устройства NR-U, выгруженная в нелицензированную часть БС, не будет поддерживаться на минимальной скорости передачи $R_{\min}$ и, таким образом, будет потеряна.

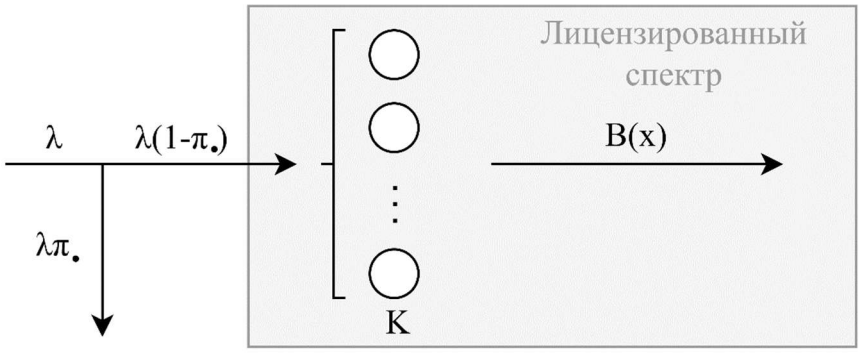

Рис. 1.15. Модель выгрузки трафика в терминах систем массового обслуживания.

Процесс обслуживания сессий пользовательских устройств NR-U на лицензированном спектре БС моделируется с использованием системы массового обслуживания M/G/K/0, см. рисунок выше (рис. 1.15). Поступление сессий с устройств NR-U моделируется как входящий поток заявок, распределенный экспоненциально с интенсивностью λ. Обратим внимание, что для заданной плотности развертывания БС, определенной интенсивности поступления сессий устройств NR-U и средних минимальных требований к ресурсам, можно определить долю нагрузки, которую устройства не могут обработать в лицензированном спектре, используя NR технологию на БС. Фактически интенсивность переполнения на нелицензированной части БС может быть рассчитана как $\lambda\pi_{\bullet}$, где $\pi_{\bullet}$ — вероятность потери сессии в M/G/K/0.

Обозначим через K максимальное количество пользовательских устройств NR-U в системе, т. е. количество заявок, которые могут одновременно

обрабатываться на лицензированном спектре БС. Для систем массового обслуживания типа M/G/K/0 стационарные вероятности наличия в системе i активных заявок пользовательских устройств NR-U равны,

$$p_i = \frac{\rho^i}{i!} p_0, \quad p_0 = \left(\sum_{i=0}^{K} \frac{\rho^i}{i!} p_i \right)^{-1}, i = 1,...,n.$$

Известно, что вероятность потери в системе M/G/K равна

$$\pi_\bullet = \frac{\rho^K}{K!} \Big/ \sum_{i=0}^{K} \frac{\rho^i}{i!},$$

где предлагаемая нагрузка $\rho = \lambda r_1^2 \pi / \mu$, и μ — средняя интенсивность обработки заявок пользовательских устройств NR-U.

Обратим внимание, что существует неотъемлемое технологическое несоответствие между NR и WiGig. В частности, более простые антенные решетки, используемые в системах WiGig, а также более высокая несущая частота часто приводят к тому, что радиус покрытия технологией WiGig меньше радиуса покрытия технологией NR, т.е. $r_2 < r_1$. Таким образом, только часть пользовательских устройств NR, $\Lambda = r_1^2 / r_2^2$, потерянная в лицензированном диапазоне частот БС, может конкурировать за ресурсы в БС на нелицензированных диапазонах частот, где $\pi_\bullet \lambda r_1^2 \pi$ — интенсивность переполнения на лицензированном диапазоне частот БС. Обратим внимание, что часть заявок с устройств NR-U теряется, $1 - \Lambda$, и эти потери можно минимизировать только за счет увеличения плотности БС.

Теперь опишем процесс обслуживания на нелицензированном спектре.

Заявки устройств NR-U, выгруженные в нелицензированный диапазон БС, конкурируют за ресурсы с устройствами WiGig. Перейдем к получению вероятности успешной передачи обоих типов пользовательских устройств, которая используется далее для определения скорости, полученной пользовательским устройством NR в нелицензированном диапазоне БС.

Пусть $\varsigma(n,m)$ - вероятность коллизии при наличии n устройств NR и m устройств WiGig, а ψ - вероятность того, что путь прямого распространения

сигнала заблокирован. Тогда вероятность успеха попытки передачи может быть выражена как

$$\theta = \theta(n,m) = \big(1-\varsigma(n,m)\big)\big(1-\psi\big). \tag{1.21}$$

Поведение системы можно описать цепью Маркова $\{X_k, k \geq 0\}$, где $X_k = i$, $i = 0,1,...,T$, обозначает номер попытки передачи, когда значения таймера обратного отсчета находится в интервале $[0, 2^i W - 1]$, и W - минимальное значение CW.

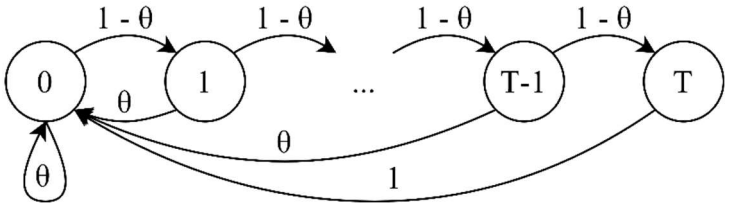

Рис. 1.16. Граф переходов между состояниями цепи Маркова, показывающих число неудачных попыток передачи.

Используя диаграмму вероятностей переходов между состояниями, показанную на рис. 1.16, получим систему уравнений для расчета стационарных вероятностей

$$\begin{cases} q_0 = q_0\theta + q_1\theta + ... + q_T, \\ q_i = q_{i-1}\big(1-\theta\big), i = 1,...,T-1, \\ q_T = q_{T+1}\big(1-\theta\big), \end{cases} \tag{1.22}$$

где q_i — стационарная вероятность вида

$$q_i = \lim_{n\to\infty} P\{X_n = i\}, i = 0,1,...,T. \tag{1.23}$$

Решая систему уравнений (1.22), получаем формулу для вычисления стационарных вероятностей

$$q_i = \frac{\theta}{1-\big(1-\theta\big)^{T+1}}\big(1-\theta\big)^i, i = 0,1,...,T. \tag{1.24}$$

Пусть теперь π_1^* и π_2^*— вероятности того, что пользовательские устройства NR-U и WiGig передают в произвольно выбранных временных тактах. Тогда, если есть *n* NR и *m* WiGig конкурирующих сессий, вероятность коллизии равна

$$\varsigma(n,m) = 1-\big(1-\pi_1^*(n,m)\big)^n\big(1-\pi_2^*(n,m)\big)^m. \tag{1.25}$$

Поскольку пользовательские устройства пытаются передавать только в одном временном такте каждого состояния цепи Маркова X_k, вероятность совершения устройством NR попытки передачи π_1^* может быть вычислена как обратное значение средней длительности пребывания (в тактах) в одном состояния цепи Маркова. Таким образом, чтобы найти вероятность того, что устройство выполнит попытку передачи, необходимо просуммировать среднее число тактов b_j, которые устройство проводит в состоянии *j*, умноженное на вероятность q_j того, что устройство находится в состоянии *j*, т.е.

$$\pi_1^* = \left[\sum_{j=0}^{T_1} q_i b_j \right]^{-1}, \tag{1.26}$$

где T_1 - число повторных передач устройством NR, среднее число тактов b_j в состоянии j имеет вид

$$b_j = \sum_{i=1}^{2^j W} \frac{1}{2^j W} i = \frac{2^j W + 1}{2}, \; j = 0,1,...,T_1. \tag{1.27}$$

Подставив (1.24), (1.26) в (1.27), вероятность передачи π_1^* можно записать в следующем виде:

$$\pi_1^*(n,m) = \left[\frac{\theta(n,m) W \left(1 - 2^{T_1+1}\left(1-\theta(n,m)\right)^{T_1+1}\right)}{2\left(1-\left(1-\theta(n,m)\right)^{T_1+1}\right)\left(2\theta(n,m)-1\right)} + \frac{1}{2} \right]^{-1} \tag{1.28}$$

Также заметим, что вероятности $\theta(n,m)$, $\varsigma(n,m)$ $\pi_1^*(n,m)$ и $\pi_2^*(n,m)$ фактически являются функциями от числа заявок *n* и *m*, конкурирующих за возможность передачи. Таким образом, решая нелинейную систему (1.21), (1.25) и (1.28) для каждой пары значений *n* и *m*, полученные значения можно использовать для вычисления успешной передачи Π_1^* следующим образом:

$$\Pi_1^* = \sum_{i=1}^{\infty} \frac{\left(\rho_1^*\right)^n}{n!} e^{-\rho_1^*} \sum_{j=0}^{\infty} \frac{\left(\rho_2^*\right)^m}{m!} e^{-\rho_2^*} \pi_1^*(n,m)\theta(n,m), \tag{1.29}$$

где предложенная нагрузка $\rho_1^* = \frac{\lambda_2 \pi_2}{\mu}$. Отметим, что на практике суммирование в Π_1^* идет до достижения некоторого заданного уровня точности ε. Для проведения численных расчетов были взяты $\varepsilon = 10^{-6}$ и $n \le 100$, $m \le 100$.

Вероятность успешной передачи для устройства WiGig рассчитывается аналогичным образом. Скорость, достигаемая устройством NR нелицензированном диапазоне по технологии WiGig, определяется с использованием формулы Шеннона:

$$\mathrm{M}\,\nu = \Pi_1^* B_2 \mathrm{M}\left[\log_2\left(1 + S(x)\right)\right], \tag{1.30}$$

где $S(x)$ — функция соотношения сигнала к шуму на расстоянии x. Выражение в правой части функции (1.30) получается следующим образом

$$\mathrm{M}\left[\log_2\left(1 + S(x)\right)\right] = \int_0^{r_2} \log_2\left(1 + S(x)\right) f(x) dx, \tag{1.31}$$

где r_2 — радиус покрытия WiGig, $f(x)$ —плотность распределения с.в. расстояния от БС до пользовательского устройства NR, рассчитываемая по формуле

$$f(x) = 2x / r_2^2,\ 0 < x < r_2. \tag{1.32}$$

Скорость, достигаемая устройством WiGig, получается аналогичным образом. Определим Q как вероятность снижения скорости устройства NR ниже требуемого порога, т. е. вероятность того, что сессия пользовательского устройства NR, потерянная в лицензированном спектре БС, также будет в конечном итоге потеряна в нелицензированном спектре БС из-за недостаточной скорости. Используя $\mathrm{M}\,\nu$, получаем

$$Q = \pi_{\bullet} I\left(\mathrm{M}\,\nu\right), \quad I(x) = \begin{cases} 1, & x < R_{\min}, \\ 1 - \dfrac{r_2^2}{r_1^2}, & x \ge R_{\min}, \end{cases} \tag{1.33}$$

где $R_{\min}$ - минимальное значение скорости передачи для поддержания сессии устройства NR-U.

Перейдем к оценке точности модели и анализу интересующих показателей. Вероятность успешной передачи — самая сложная часть разрабатываемой модели,

которая может повлиять на ее общую точность. Таким образом, сначала оценим точность этой части, сравнивая вероятность успешной передачи, полученную с помощью разработанной модели, и рассчитанную с помощью компьютерного моделирования процедуры произвольного доступа, показанные на рис. 1.17. Как можно заметить, результаты модели очень близки к результатам, полученным с помощью компьютерного моделирования. Таким образом, в дальнейшем можно опираться на разработанную модель для оценки полученных характеристик.

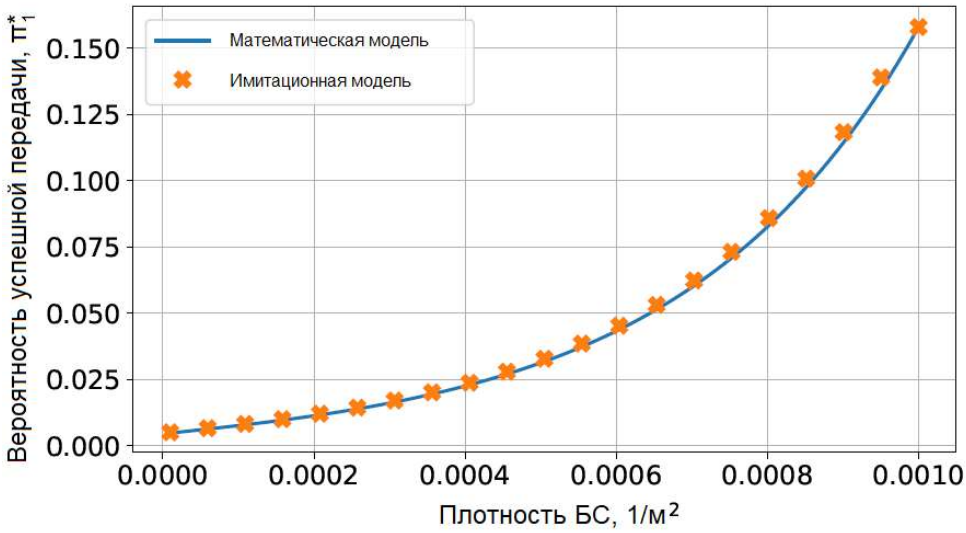

Рис. 1.17. Сравнение математической модели с имитационной для вероятности успешной попытки передачи.

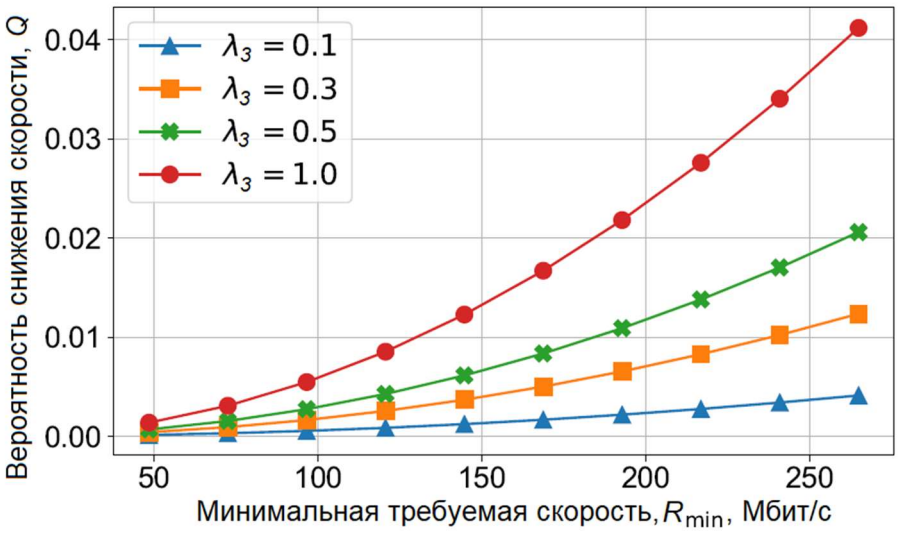

Рис. 1.18. График зависимости вероятности снижения скорости устройства NR-U ниже требуемого порога.

Теперь приступим к оценке реакции системы на входные параметры, фокусируясь на вероятности снижения скорости устройства NR ниже требуемого порога, Q.

На рис. 1.18 показано влияние минимальной скорости сессии устройства NR-U, $R_{\min}$, на вероятность снижения скорости устройства NR ниже требуемого порога, Q, для нескольких значений плотности блокаторов. Общая плотность блокаторов представляет собой плотность, полученную путем суммирования плотности пользовательских устройств WiGig, NR-U и внешних блокаторов. Данный график подтверждает, что минимальная требуемая скорость отрицательно влияет на вероятность снижения скорости устройства NR ниже требуемого порога, так как для заданной плотности БС, плотности устройств WiGig и NR достигнутая скорость уменьшается с увеличением $R_{\min}$. Кроме того, обратим внимание, что увеличение плотности блокаторов отрицательно влияет на вероятность снижения скорости устройства NR ниже требуемого порога в конечном итоге.

В модели, представленной в разделе 1.3, разработана модель для описания работы технологии NR-U, описывающая произвольный доступ в нелицензированном диапазоне и характеризующая вероятность потери сессии NR-U. Однако для описания особенности распределения ресурсов в лицензируемом диапазоне была использована очень простая модель M/M/K/0. В связи с этим необходимо разработать более точную модель процесса обслуживания в лицензированном диапазоне.

На основании вышеизложенного формулируются следующие задачи исследования.

1. Построение в виде системы массового обслуживания двухпараметрической модели выгрузки в систему туманно-облачных вычислений задач мобильных вычислений, разработка метода расчета функции распределения времени отклика с учетом неоднородности задач по объему вычислений и данных.
2. Построение и анализ модели в виде ресурсной системы массового обслуживания, описывающей обслуживание трафика в лицензированном

диапазоне частот, и дискретной цепи Маркова, описывающей случайный доступ при выгрузке в нелицензированный диапазон частот. Разработка метода расчета распределения скорости передачи в нелицензированном диапазоне.

ГЛАВА 2. АНАЛИЗ МОДЕЛИ ВЫГРУЗКИ ЗАДАЧ МОБИЛЬНЫХ ВЫЧИСЛЕНИЙ В ТУМАННО-ОБЛАЧНОЙ СИСТЕМЕ

В предыдущей главе разработана аналитическая структура для исследования времени отклика, которая учитывает изменение задач с точки зрения объема вычислений. В этой главе описывается двухпараметрический механизм выгрузки, который учитывает, как вычислительную сложность через объем вычислений, так и объем данных, которые должны быть переданы в случае выгрузки на узел туманных вычислений. После этого приводится функция распределения времени отклика с использованием преобразования Лапласа-Стилтьеса. В завершение будет сформулирована задача оптимизации, которая минимизирует энерго-эффективность, с учетом ограничения на среднее время отклика и с учетом вероятности того, что это время превысит заданный порог.

В диссертации, основываясь на теории массового обслуживания, построена математическая модель для систем выгрузки задач в узлы туманных и облачных вычислений. Опишем системную модель.

2.1. Модель с двухпараметрическим критерием выгрузки

Рассматривается распределенная вычислительная система, состоящая из мобильных устройств, беспроводной сети, узла туманных вычислений и удаленного узла облачных вычислений. Мобильные устройства запускают приложения реального времени, которые требуют значительного количества вычислительных ресурсов. Для каждой задачи мобильное устройство принимает решение, будет ли она выгружена в узел туманных вычислений или обслужена локально. Емкость узла туманных вычислений ограничена. Рассматриваемую систему можно представить схематически так, как показано на рис. 3.1.

Предположим, что существует M мобильных устройств, каждое из которых генерирует пуассоновский поток задач с интенсивностью $\lambda_i, 1 \le i \le M$. Каждая задача характеризуется требуемым объемом вычислений и объемом данных, которые необходимо передать в случае выгрузки. Предположим, что объем

вычислений (измеряемый в миллионах инструкций) и объем данных (измеренный в Мбайтах) являются независимыми случайными величинами (с.в.), имеющими гамма-распределение с параметрами k_l и δ_l, $l=1,2$, с плотностями распределения $w_i(x)$ и $s_i(x)$ и с ФР $W_i(x)$ и $S_i(x)$, соответственно. Мобильные устройства локально обрабатывают задачи в порядке поступления (*англ.* First Come First Serve, FCFS) с постоянной скоростью обработки $\mu_i, 1 \le i \le M$ (измеряется в миллионах инструкций в секунду).

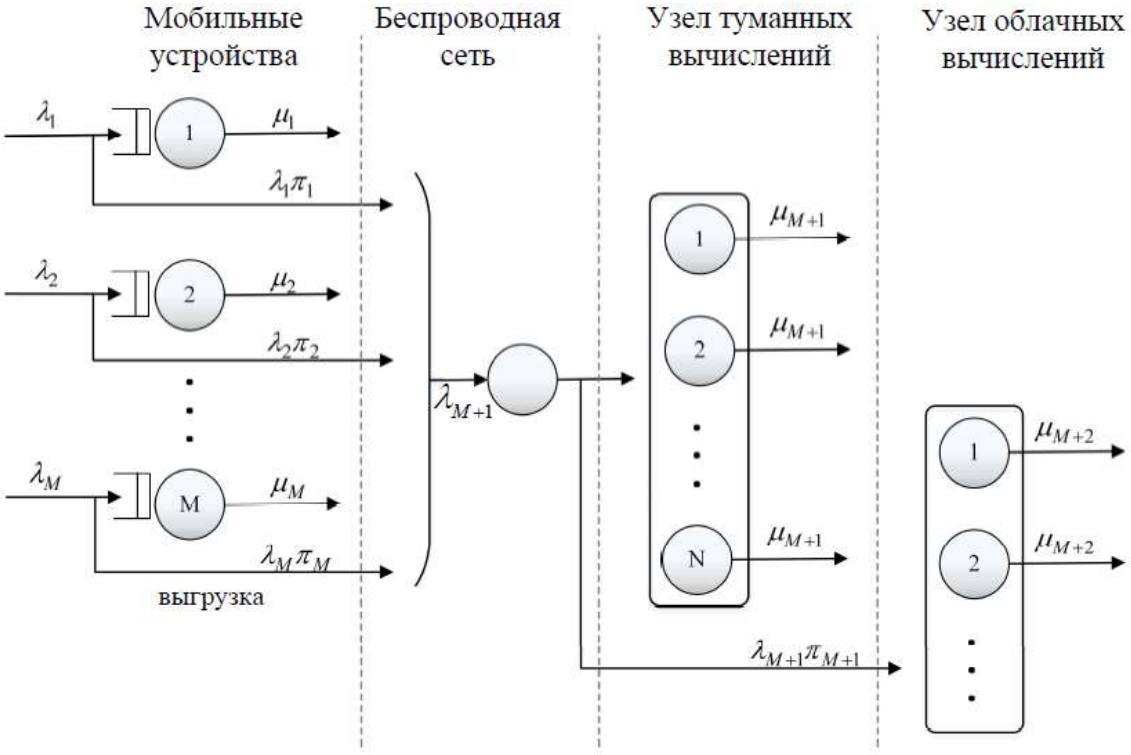

Рис. 2.1. Схема модели выгрузки мобильных вычислений.

Предлагаемый механизм выгрузки подразумевает выгрузку задач, которые являются «тяжелыми» с точки зрения объема вычислений и «легкими» с точки зрения объема данных. Разделение на «тяжелые» и «легкие» задачи выполняется нижним пороговым значением w^* для объема вычислений и верхним пороговым значением s^* для объема данных. Следовательно, вероятность выгрузки π_i на i-м мобильном устройстве имеет вид

$$\pi_i = \int_{w^*}^{\infty} w_i(x)dx \int_0^{s^*} s_i(y)dy = (1 - W_i(w^*))S_i(s^*). \qquad (2.1)$$

Если задача обрабатывается локально, то время отклика состоит только из времени обработки на мобильном устройстве. Если задача выгружается на узел туманных вычислений, то общее время отклика является суммой времени передачи задачи в узел туманных вычислений через беспроводную сеть, времени обработки в узле туманных вычислений и времени передачи обратно на мобильное устройство. Если узел туманных вычислений перегружен и выгруженная задача отправляется в удаленный узел облачных вычислений, то время обработки в узле туманных вычислений заменяется временем передачи между узлами туманных и облачных вычислений, временем обработки в узле облачных вычислений и временем передачи назад к узлу туманных вычислений.

На узел беспроводной сети поступает поток выгружаемых из мобильных устройств задач, являющийся по сути просеянным пуассоновским потоком с интенсивностью λ_{M+1}. Будем считать, что беспроводная сеть обеспечивает общую скорость передачи *R*, которая распределяется поровну между всеми одновременно передающими мобильными устройствами. Следовательно, процесс передачи по беспроводной сети моделируется в терминах СМО M/G/1/0. Время передачи по беспроводной сети равно отношению объема данных задачи к скорости передачи *R*. С другой стороны, время передачи между узлами туманных и облачных вычислений считаем постоянным.

После прохода через беспроводную сеть интервалы между поступлениями заявок уже не будут иметь экспоненциального распределения, и поток не является пуассоновским. Однако в данном исследовании будем считать в качестве упрощающего предположения, что на узел туманных вычислений поступает пуассоновский поток задач с интенсивностью λ_{M+1}. Узел туманных вычислений предоставляет вычислительные ресурсы в виде виртуальных машин (ВМ), каждая из которых имеет постоянную скорость обработки μ_{M+1}. Общее количество виртуальных машин в узле туманных вычислений составляет *N*. Постоянная скорость обработки μ_{M+2} ВМ в узле облачных вычислений больше, чем μ_{M+1}, и предполагается, что количество вычислительных ресурсов, т.е. виртуальных

машин, в узле облачных вычислений достаточно велико, чтобы его нельзя было перегрузить.

2.2. Анализ распределения времени отклика в условиях выгрузки

2.2.1 Анализ времени отклика. ФР компонентов времени отклика

В этом подразделе приведены ФР для всех компонентов времени отклика, которые будут использоваться для анализа распределения общего времени отклика в условиях выгрузки.

Процесс обслуживания в *i*-м мобильном устройстве моделируется в терминах СМО типа M/G/1 с интенсивностью поступления входящих заявок $\lambda_i, \sum_{i=1}^{M} \lambda_i = \lambda$.

Распределение объема вычислений локально обслуживаемых заявок имеет следующую ФР:

$$V_i(x) = \begin{cases} \dfrac{W_i(x)}{1-\pi_i}, x \le w^*, \\ \dfrac{W_i(w^*) + (W_i(x) - W_i(w^*))(1 - S_i(s^*))}{1-\pi_i}, x > w^*. \end{cases} \quad (2.2)$$

Получив функцию распределения объема вычислений $V_i(x)$, можно найти время обработки заявки на мобильном устройстве. В силу того, что скорость обработки на *i*-м мобильном устройстве постоянна и равна μ_i, его ФР легко получается как

$$T_i(x) = V_i(\mu_i x). \quad (2.3)$$

Среднее время обработки в *i*-м мобильном устройстве можно найти путем интегрирования с использованием ФР $T_i(x)$.

Если задача выгружена в распределенную вычислительную инфраструктуру, она сначала передается через беспроводную сеть в узел туманных вычислений. Задержки в беспроводных сетях получаются аналогично, используя СМО типа M/G/1. Интенсивность поступления заявок λ_{M+1} представляет собой сумму интенсивностей выгрузки от всех мобильных устройств:

$$\lambda_{M+1} = \sum_{i=1}^{M} \lambda_i \pi_i \,. \tag{2.4}$$

ФР размера передаваемого файла выгруженной задачи $G_i(x)$ имеет вид

$$G_i(x) = \begin{cases} \dfrac{1}{\pi_i}(1 - W_i(w^*))S_i(x),\ x \le s^*, \\ 1,\ x > s^*. \end{cases} \tag{2.5}$$

ФР времени обслуживания в беспроводной сети задается по формуле

$$D_i(x) = G_i(Rx). \tag{2.6}$$

В узле туманных вычислений находится N виртуальных машин для обработки выгруженных задач, поэтому процесс обслуживания может быть смоделирован в виде системы типа M/G/N/0, где заблокированные задачи перенаправляются на следующий уровень – узел облачных вычислений. Интенсивность поступления задач такая же, как для беспроводной сети - λ_{M+1}. ФР объема вычислений для задач, поступающих с *i*-го мобильного устройства на узел туманных вычислений определяется как

$$V_{M+1,i}(x) = \begin{cases} 0,\ x \le w^*, \\ \dfrac{1}{\pi_i}(W_i(x) - W_i(w^*))S_i(s^*),\ x > w^*. \end{cases} \tag{2.7}$$

Время обработки равна объему вычислений, деленному на интенсивность обслуживания μ_{M+1}, поэтому ФР времени обработки задач в узле туманных вычислений имеет вид

$$T_{M+1,i}(x) = V_{M+1,i}(\mu_{M+1}x), \tag{2.8}$$

$$T_{M+1}(x) = \frac{\lambda_i \pi_i}{\lambda_{M+1}} T_{M+1,i}(x). \tag{2.9}$$

Обозначим β_{M+1} за среднее время обработки задачи в узле туманных вычислений, которое можно оценить из ФР $T_{M+1}(x)$. Вероятность перегрузки, т.е. вероятность того, что задача будет перенаправлена в узел облачных вычислений π_{M+1}, вычисляется по формуле Эрланга:

$$\pi_{M+1} = \frac{(\lambda_{M+1}\beta_{M+1})^N}{N!}\left(\sum_{k=0}^{N}\frac{(\lambda_{M+1}\beta_{M+1})^k}{k!}\right)^{-1}. \quad (2.10)$$

Вероятность перегрузки узла туманных вычислений π_{M+1} не зависит от объема обработки заявок, поэтому ФР времени обработки $T_{M+2}(x)$ в узле облачных вычислений равна

$$T_{M+2,i}(x) = W_{M+1,i}(\mu_{M+2}x). \quad (2.11)$$

Затем после обработки в узле облачных вычислений задача возвращается на мобильное устройство с постоянной задержкой передачи между узлами туманных и облачных вычислений и беспроводной сетью. Задача считается обработанной.

2.2.2 Анализ общего времени отклика

В этом подразделе приводятся преобразования Лапласа-Стилтьеса (ПЛС) для всех компонентов задержки передачи и получается ПЛС общего времени отклика.

ПЛС функции существенно зависит от ее типа. Для простоты предполагаем, что и объем вычислений, и объем данных имеют гамма-распределение с параметрами k_1, δ_1 и k_2, δ_2, соответственно:

$$w_i(x) = x^{k_1-1}\frac{e^{-\frac{x}{\delta_1}}}{\delta_1^{k_1}\Gamma(k_1)}, \qquad s_i(x) = x^{k_2-1}\frac{e^{-\frac{x}{\delta_2}}}{\delta_2^{k_2}\Gamma(k_2)}.$$

Утверждение 2.1. ПЛС времени обработки задачи в *i*-м мобильном устройстве имеет вид

$$\tilde{T}_i(s) = \frac{\mu_i}{1-\pi_i}\left[\frac{1}{(\delta_1 s+\mu_i)^{k_1}} - e^{-\left(\frac{sw^*}{\mu_i}+\frac{w^*}{\delta_1}\right)}\sum_{n=0}^{k_1-1}\frac{\left(w^*/\mu_i\right)^n}{n!\delta_1^n(\delta_1 s+\mu_i)^{k_1-n}}\left(1-e^{-\frac{s^*}{\delta_2}}\sum_{n=0}^{k_2-1}\frac{\left(s^*/\delta_2\right)^n}{n!}\right)\right]. \quad (2.12)$$

Доказательство: С учетом формул (2.2) и (2.3) для вычисления $T_i(x)$, ПЛС $\tilde{T}_i$ времени обработки задачи в *i*-м мобильном устройстве имеет вид

$$\tilde{T}_i(s) = \int_0^{+\infty} e^{-sx}d(T_i(x)) = \frac{1}{1-\pi_i}\left[\int_0^{\frac{w^*}{\mu_i}} e^{-sx}d\left(W_i(\mu_i x)\right) + \right.$$

$$+\int_{\frac{w^*}{\mu_i}}^{\infty} e^{-sx} d\left(W_i(\mu_i x)\left(1-S_i\left(s^*\right)\right)+W_i(w^*)S_i(s^*)\right)\Bigg]. \qquad (2.13)$$

После подстановки $w_i(x)$ функция имеет вид

$$\tilde{T}_i(s)=\int_0^{\frac{w^*}{\mu_i}} e^{-sx}\frac{e^{-\mu_i\frac{x}{\delta_1}}x^{k_1-1}}{\delta_1^{\,k}(k_1-1)!}dx+\int_{\frac{w^*}{\mu_i}}^{\infty}\left(1-S_i\left(s^*\right)\right)e^{-sx}\frac{e^{-\mu_i\frac{x}{\delta_1}}x^{k_1-1}}{\delta_1^{\,k}(k_1-1)!}dx. \qquad (2.14)$$

Интегрируя первое слагаемое (2.14) по частям, получается

$$I_1=\int_0^{\frac{w^*}{\mu_i}}\frac{x^{k_1-1}}{(k_1-1)!}\frac{e^{-\left(s+\frac{\mu_i}{\delta_1}\right)x}}{\delta_1^{\,k}}dx=-\frac{x^{k_1-1}}{(k_1-1)!\delta_1^{\,k_1-1}\left(s\delta_i+\mu_i\right)}e^{-\left(\frac{s\delta_i+\mu_i}{\delta_1}\right)x}\Bigg|_0^{\frac{w^*}{\mu_i}}+$$

$$+\int_0^{\frac{w^*}{\mu_i}}\frac{x^{k_1-2}}{(k_1-2)!\delta_1^{\,k_1-1}\left(s\delta_i+\mu_i\right)}e^{-\left(\frac{s\delta_1+\mu_i}{\delta_1}\right)x}dx=\ldots=$$

$$=-\sum_{n=1}^{k_1-1}\frac{\left(w^*/\mu_i\right)^n}{n!\delta_1^{\,n}\left(s\delta_i+\mu_i\right)^{k_1-n}}e^{-\left(\frac{sw^*}{\mu_i}+\frac{w^*}{\delta_i}\right)}+\int_0^{\frac{w^*}{\mu_i}}\frac{1}{\delta_1\left(s\delta_1+\mu_i\right)^{k_1-1}}e^{-\left(\frac{s\delta_1+\mu_i}{\delta_1}\right)x}dx=$$

$$=-\sum_{n=1}^{k_1-1}\frac{\left(w^*/\mu_i\right)^n}{n!\delta_1^{\,n}\left(s\delta_1+\mu_i\right)^{k_1-n}}e^{-\left(\frac{sw^*}{\mu_i}+\frac{w^*}{\delta_1}\right)}-\frac{1}{\left(s\delta_1+\mu_i\right)^{k_1}}e^{-\left(\frac{sw^*}{\mu_i}+\frac{w^*}{\delta_1}\right)}+$$

$$+\frac{1}{\left(s\delta_i+\mu_i\right)^{k_1}}=\frac{1}{\left(s\delta_1+\mu_i\right)^{k_1}}-\sum_{n=0}^{k_1-1}\frac{\left(w^*/\mu_i\right)^n}{n!\delta_1^{\,n}\left(s\delta_1+\mu_i\right)^{k_1-n}}e^{-\left(\frac{sw^*}{\mu_i}+\frac{w^*}{\delta_1}\right)}. \qquad (2.15)$$

Интегрируя второе слагаемое (2.14) по частям, получается

$$I_2=\int_{\frac{w^*}{\mu_i}}^{\infty}\left(1-S_i\left(s^*\right)\right)e^{-sx}\frac{e^{-\mu_i\frac{x}{\delta_1}}x^{k_1-1}}{\delta_1^{\,k}(k_1-1)!}dx=$$

$$=\left(1-S_i\left(s^*\right)\right)\left[-\sum_{n=0}^{k_1-1}\frac{x^n}{n!\delta_1^{\,n}\left(s\delta_i+\mu_i\right)^{k_1-n}}e^{-\left(\frac{s\delta_i+\mu_i}{\delta_i}\right)x}\Bigg|_{\frac{w^*}{\mu_i}}^{+\infty}\right]=$$

$$= \left(1 - S_i\left(s^*\right)\right) \sum_{n=0}^{k_1-1} \frac{\left(w^* / \mu_i\right)^n}{n! \delta_1^{\,n} \left(s\delta_i + \mu_i\right)^{k_1-n}} e^{-\left(\frac{sw^*}{\mu_i} + \frac{w^*}{\delta_i}\right)x}. \qquad (2.16)$$

Аналогичным образом интегрируется ФР объема данных:

$$S_i(s^*) = \int_0^{s^*} \frac{e^{-\frac{x}{\delta_2}} x^{k_2-1}}{\delta_2^{\,k} (k_2-1)!} dx = -\frac{x^{k_2-1} e^{-\frac{x}{\delta_2}}}{(k_2-1)! \delta_2^{\,k_2-1}}\Bigg|_0^{s^*} + \int_0^{s^*} \frac{x^{k_2-2}}{(k_2-2)!} \frac{e^{-\frac{x}{\delta_2}}}{\delta_2^{\,k}} dx =$$

$$-\frac{\left(s^*\right)^{k_2-1} e^{-\frac{s^*}{\delta_2}}}{(k_2-1)! \delta_2^{\,k}} - \frac{x^{k_2-2}}{(k_2-2)!} \frac{e^{-\frac{x}{\delta_2}}}{\delta_2^{\,k-2}}\Bigg|_0^{s^*} - \int_0^{s^*} \frac{x^{k_2-3}}{(k_2-3)!} \frac{e^{-\frac{x}{\delta_2}}}{\delta_2^{\,k-2}} dx = \ldots =$$

$$= -\sum_{n=1}^{k_2-1} \frac{\left(s^*\right)^n e^{-\frac{s^*}{\delta_2}}}{n! \delta_2^{\,n}} - \int_0^{s^*} \frac{e^{-\frac{x}{\delta_2}}}{\delta_2} dx = 1 - \sum_{n=0}^{k_2-1} \frac{\left(s^* / \delta_2\right)^n e^{-\frac{s^*}{\delta_2}}}{n!}. \qquad (2.17)$$

После подстановки формул (2.15-2.17) в (2.14) получается выражение (2.12).

■

Затем, ПЛС ω_i времени ожидания и ПЛС ϕ_i времени пребывания на *i*-м мобильном устройстве, задаются формулами (2.18) и (2.19) соответственно.

$$\omega_i(s) = \frac{s(1-\rho_i)}{s - \lambda_i + \lambda_i \tilde{T}_i(s)}, \qquad (2.18)$$

$$\phi_i(s) = \tilde{T}_i(s) \omega_i(s), \qquad (2.19)$$

где ρ_i - предлагаемая нагрузка на *i*-м мобильном устройстве.

Те же шаги проделываются для вычисления времени передачи через беспроводную сеть.

Утверждение 2.2. ПЛС распределения времени передачи имеет вид

$$\tilde{D}_i(s) = \frac{R}{\pi_i} \left[\frac{1}{(\delta_2 s + R)^{k_2}} - e^{-\left(\frac{ss^*}{R} + \frac{s^*}{\delta_2}\right)} \sum_{n=0}^{k_2-1} \frac{\left(s^*/R\right)^n}{n! \delta_2^n (\delta_2 s + R)^{k_2-n}} \right] \sum_{m=0}^{k_1-1} \frac{\left(w^*/\delta_1\right)^m}{m!} e^{-\left(\frac{w^*}{\delta_1}\right)}, \qquad (2.20)$$

где *R* – скорость передачи в беспроводной сети.

Доказательство: ПЛС $\tilde{D}_i(s)$ распределения времени передачи получается непосредственно из $D_i(x)$ (2.6). ПЛС распределения времени передачи имеет вид

$$\tilde{D}_i(s) = \int_0^{+\infty} e^{-sx} d(D_i(x)) = \frac{R}{\pi_i} \int_0^{\frac{s^*}{R}} e^{-sx} \left(1 - W_i(w^*)\right) \frac{e^{-R\frac{x}{\delta_2}} x^{k_2-1}}{\delta_2{}^k (k_2-1)!} dx. \qquad (2.21)$$

По аналогии с (2.15) из доказательства утверждения 2.1, ПЛС интегрируется по частям следующим образом:

$$\tilde{D}_i(s) = R\frac{1 - W_i(w^*)}{\pi_i}\left[\frac{1}{(\delta_2 s + R)^{k_2}} - \sum_{n=0}^{k_2-1} \frac{\left(s^*/R\right)^n}{n!\delta_2^n (\delta_2 s + R)^{k_2-n}} e^{-\left(\frac{ss^*}{R} + \frac{s^*}{\delta_2}\right)}\right], \qquad (2.22)$$

где аналогично (2.17) ФР имеет вид

$$W_i(w^*) = \int_0^{w^*} \frac{e^{-\frac{x}{\delta_1}} x^{k_1-1}}{\delta_1{}^k (k_1-1)!} dx = 1 - \sum_{n=0}^{k_1-1} \frac{\left(w^*/\delta_1\right)^n e^{-\frac{w^*}{\delta_1}}}{n!}. \qquad (2.23)$$

■

Поскольку беспроводная сеть также моделируется в терминах СМО типа M/G/1/0, ПЛС $\psi_i(s)$ времени ожидания и ПЛС $\varphi_i(s)$ времени пребывания в беспроводной сети равны:

$$\psi_i(s) = \frac{s(1-\rho_\bullet)}{s - \lambda_{M+1} + \lambda_{M+1}\tilde{D}_i(s)}, \qquad (2.24)$$

$$\varphi_i(s) = \tilde{D}_i(s)\psi_i(s). \qquad (2.25)$$

Здесь $\rho_\bullet$ - предлагаемая нагрузка в беспроводной сети, равная произведению λ_{M+1} на средний размер передаваемых данных, который получается из ФР (2.5).

Утверждение 2.3. ПЛС распределения времени обслуживания на узле туманных вычислений имеет вид

$$\tilde{T}_{M+1,i}(s) = \frac{\mu_{M+1}}{\pi_i}\left[e^{-\left(\frac{sw^*}{\mu_{M+1}} + \frac{w^*}{\delta_1}\right)} \sum_{n=0}^{k_1-1} \frac{\left(w^*/\mu_{M+1}\right)^n}{n!\delta_1^n (\delta_1 s + \mu_{M+1})^{k_1-n}} \left(1 - e^{-\frac{s^*}{\delta_2}} \sum_{n=0}^{k_2-1} \frac{\left(s^*/\delta_2\right)^n}{n!}\right)\right]. \qquad (2.26)$$

Доказательство: ПЛС $\tilde{T}_{M+1,i}(s)$ распределения времени обслуживания в узле туманных вычислений получается из ФР $T_{M+1,i}(x)$ (2.8) и ФР $V_{M+1,i}(x)$ (2.7) и имеет вид

$$\tilde{T}_{M+1,i}(s)=\int_0^{+\infty} e^{-sx}d(T_{M+1,i}(x))=\frac{\mu_{M+1}}{\pi_i}\int_{\frac{w^*}{\mu_{M+1}}}^{+\infty}\left(S_i\left(s^*\right)\right)e^{-sx}\frac{e^{-\mu_{M+1}\frac{x}{\delta_1}}x^{k_1-1}}{\delta_1{}^k(k_1-1)!}dx. \qquad (2.27)$$

По аналогии с формулой (2.16) из доказательства утверждения 2.1, ПЛС интегрируется по частям, в результате чего получается следующее выражение

$$\tilde{T}_{M+1,i}(s)=\mu_{M+1}\frac{S_i(s^*)}{\pi_i}\left[\sum_{n=0}^{k_1-1}\frac{\left(w^*/\mu_{M+1}\right)^n}{n!\delta_1^n(\delta_1 s+\mu_{M+1})^{k_1-n}}e^{-\left(\frac{sw^*}{\mu_{M+1}}+\frac{w^*}{\delta_1}\right)}\right], \qquad (2.28)$$

где $S_i(s^*)$ вычисляется по ранее полученной формуле (2.17).

■

Обратим внимание, что как в узле туманных вычислений, так и узле облачных вычислений нет очереди ожидания.

Утверждение 2.4. ПЛС распределения времени обслуживания на узле удаленного облака имеет вид

$$\tilde{T}_{M+2,i}(s)=\frac{\mu_{M+2}}{\pi_i}\left[e^{-\left(\frac{sw^*}{\mu_{M+2}}+\frac{w^*}{\delta_1}\right)}\sum_{n=0}^{k_1-1}\frac{\left(w^*/\mu_{M+2}\right)^n}{n!\delta_1^n(\delta_1 s+\mu_{M+2})^{k_1-n}}\left(1-e^{-\frac{s^*}{\delta_2}}\sum_{n=0}^{k_2-1}\frac{\left(s^*/\delta_2\right)^n}{n!}\right)\right]. \qquad (2.29)$$

Доказательство: Аналогично тому, как была получена формула (2.28) из утверждения 2.3, получается формула (2.29). ПЛС $\tilde{T}_{M+2,i}(s)$ распределения времени обслуживания в узле облачных вычислений получается из ФР $T_{M+2,i}(x)$ (2.11) и имеет вид

$$\tilde{T}_{M+2,i}(s)=\int_0^{+\infty} e^{-sx}d(T_{M+2,i}(x))=\frac{\mu_{M+2}}{\pi_i}\int_{\frac{w^*}{\mu_{M+2}}}^{+\infty}\left(S_i\left(s^*\right)\right)e^{-sx}\frac{e^{-\mu_{M+2}\frac{x}{\delta_1}}x^{k_1-1}}{\delta_1{}^k(k_1-1)!}dx. \qquad (2.31)$$

По аналогии с формулой (2.28) из доказательства утверждения 2.3, ПЛС $\tilde{T}_{M+2,i}(s)$ интегрируется по частям, в результате чего получается следующее выражение

$$\tilde{T}_{M+2,i}(s) = \mu_{M+1}\frac{S_i(s^*)}{\pi_i}\left[\sum_{n=0}^{k_1-1}\frac{\left(w^*/\mu_{M+2}\right)^n}{n!\delta_1^n(\delta_1 s+\mu_{M+2})^{k_1-n}}e^{-\left(\frac{sw^*}{\mu_{M+2}}+\frac{w^*}{\delta_1}\right)}\right], \quad (2.32)$$

где $S_i(s^*)$ так же вычисляется по ранее полученной формуле (2.17).

■

Наконец, ПЛС времени передачи между узлом туманных вычислений и узлом облачных вычислений задается выражением $e^{-\Delta_2 s}$, где Δ_2 - постоянное время передачи между узлом туманных и облачных вычислений.

Теперь можно перейти к вычислению ПЛС общего времени отклика. Общее время отклика – это условная сумма задержек обработки и передачи. Задача из *i*-го мобильного устройства обрабатывается локально с вероятностью $1-\pi_i$, на узле туманных вычислений с вероятностью $\pi_i(1-\pi_{M+1})$ и в узле облачных вычислений с вероятность $\pi_i\pi_{M+1}$. Зная ПЛС для всех компонентов задержки и применяя свойства ПЛС от свертки, получаем ПЛС $\tilde{\beta}(s)$ распределения времени отклика задачи из *i*-го мобильного устройства.

Следствие 2.1 ПЛС времени отклика имеет вид

$$\tilde{\beta}_i(s) = (1-\pi_i)\phi_i(s) + \pi_i(1-\pi_{M+1})\tilde{T}_{M+1,i}(s)\varphi_i^2(s) + \pi_i\pi_{M+1}\tilde{T}_{M+2,i}(s)\varphi_i^2(s)e^{-\Delta_2 s}. \quad (2.32)$$

где Δ_2 - время передачи между узлами туманных и облачных вычислений.

Для вывода ФР $\beta_i(s)$ времени отклика было использовано обратное ПЛС. Среднее время отклика может быть представлено в виде

$$\beta = \left(\sum\nolimits_{j=1}^{M}\lambda_j\right)^{-1}\sum_{i=1}^{M}\lambda_i\beta_i. \quad (2.33)$$

Полученные выражения позволяют получить вероятность того, что время отклика меньше порога t^* как $B(t^*) = \beta(t^*)$. Тогда вероятность превышения заданного порога t^* представляет собой $1-B(t^*)$.

2.2.3 Анализ энергопотребления

В этом подразделе представлены формулы для среднего энергопотребления мобильных устройств.

Потребление энергии для задач, обрабатываемых на мобильном устройстве, пропорционально объемам обработки задач, поэтому среднее потребление энергии $E_{1,i}$ (измеряемое в Джоулях) при локальной обработке задачи на *i*-м мобильном устройстве составляет

$$E_{1,i} = P_{1,i} t_i, \tag{2.34}$$

где $P_{1,i}$ – потребляемая мощность (Вт) во время обработки задачи *i*-м мобильным устройством, которая принимается постоянной для простоты.

Средний размер файла, передаваемого с *i*-го мобильного устройства, можно вычислить путем интегрирования с использованием ФР $G_i(x)$ из предыдущего подраздела.

Среднее потребление энергии *i*-го мобильного устройства во время передачи также пропорционально времени передачи и составляет

$$E_{2,i} = P_{2,i} \frac{\theta_i}{R}, \tag{2.35}$$

где $P_{2,i}$ – мощность передачи (Вт) во время обработки задачи с *i*-го мобильного устройства, θ_i – средний размер файла, передаваемого с *i*-го узла. Тогда среднее потребление энергии для любого *i*-го мобильного устройства представляет собой взвешенную сумму потреблений энергии на обработку и передачу:

$$E_i = (1 - \pi_i) E_{1,i} + \pi_i E_{2,i}. \tag{2.36}$$

Наконец, можно оценить среднее потребление энергии для задачи с произвольного мобильного устройства следующим образом

$$E = \left(\sum\nolimits_{j=1}^{M} \lambda_j \right)^{-1} \sum_{i=1}^{M} \lambda_i E_i. \tag{2.37}$$

2.3. Численный анализ оптимальных порогов критерия выгрузки

В этом разделе представлены численные результаты исследования. Рассматривается система с $M = 20$ однородными мобильными устройствами, которые запускают одни и те же приложения, поэтому распределения объема вычислений и размера данных задач также одинаковы. Узлы туманных вычислений

могут запускать максимум N = 8 виртуальных машин. Все значения параметров, используемых в разделе, собраны в Таблице 3.1.

Таблица 2.1. Исходные данные для численного анализа.

Параметр	Значение
M	20
N	8
R	150 Мбит / с
λ_i	2 задачи / с
μ_i	4 млн инструкций /с
μ_{M+1}	6 млн инструкций /с
μ_{M+2}	10 млн инструкций /с
δ_2	0.25
δ_1	0.75
Δ_2	0.5 с
P_1	16 Вт
P_2	0.2 Вт

На рисунке 2.2 показана вероятность $B(t^*)$, что время отклика меньше порога t^* для разных пороговых значений объема вычислений w^*. Как видно из графика, при небольших значениях порога объёма данных, s^* = 0.2, ничего не выгружается и почти все обрабатывается локально, из-за чего мобильные устройства оказываются перегруженными и с увеличением w^* нагрузка на устройства продолжает расти. При небольшом увеличении порога объёма данных, s*=0.3, часть задач выгружается на узел туманных вычислений, из-за чего нагрузка на устройства ниже. С увеличением порога объема данных до s*=0.4, при небольших пороговых значениях объема вычислений w* туманный узел оказывается перегружен, с увеличением порога w* происходит балансировка нагрузки, и в точке w*=1.4 достигается оптимальное перераспределение нагрузки с точки зрения $B(t^*)$.

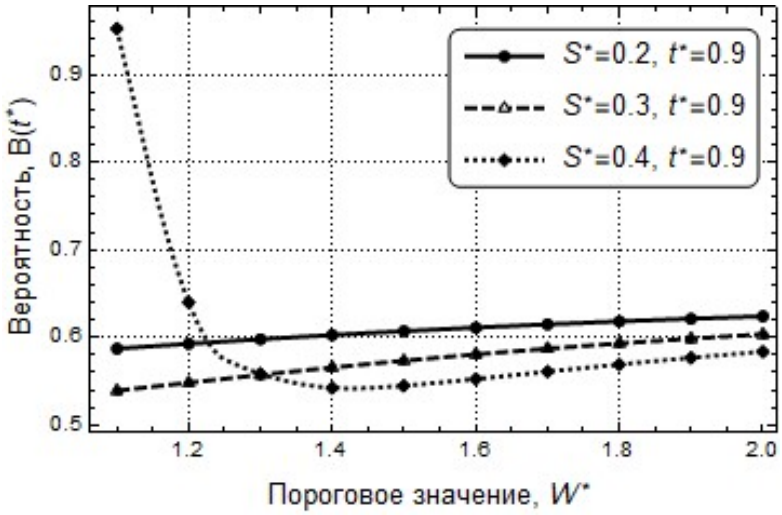

Рис. 2.2. Вероятность того, что время отклика меньше порога t^* от порогового значения w^*.

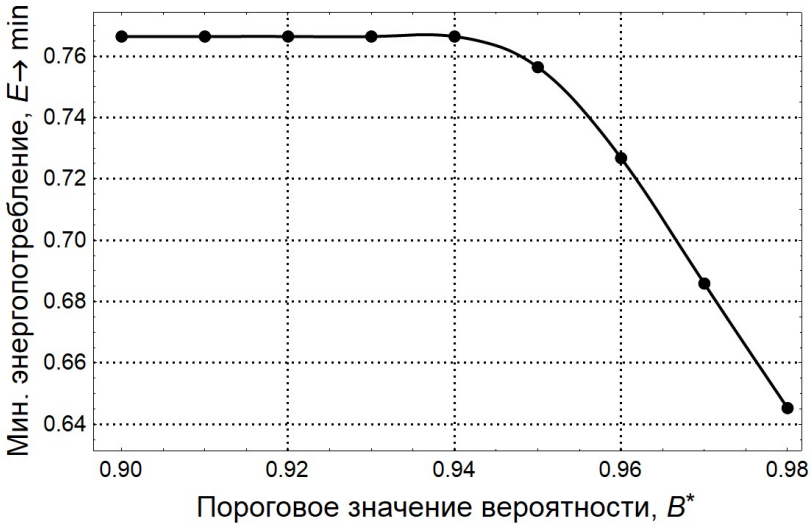

Рис. 2.3. Зависимость минимальной потребляемой энергии $E \to \min$ от порогового значения B^*, $t^* = 0{,}5$

Кроме того, основной интересующей метрикой является минимальное потребление энергии E для задачи с произвольного мобильного устройства при некоторых ограничениях пороговыми значениями.

Чтобы найти минимальное потребление энергии E, заданное формулой (2.37), при ограничениях на среднее время отклика β, заданное формулой (2.33), и вероятность $B(t^*) = B(t^*, w^*, s^*)$ того, что время отклика ниже порогового значения t*, была сформулирована задача оптимизации следующим образом:

$$\begin{cases} E\left(w^*, s^*\right) \to \min, \\ \mathrm{R1}: \beta\left(w^*, s^*\right) \le t^*, \\ \mathrm{R2}: B(t^*, w^*, s^*) \le B^*. \end{cases} \tag{2.38}$$

Задача (2.38) решалась численно перебором. На рисунке 2.3 показана зависимость вероятности $B(t^*)$ от порогового значения объема обработки с пороговым значением времени отклика $t^* = 0{,}5$.

График энергопотребления начинает убывать только при значениях $B^* = 0.94$ и выше. Это показывает, что только при очень высоком пороговом значении вероятности $B(t^*)$ того, что время отклика ниже порогового значения t^*, будет выигрыш в виде энергозатрат.

ГЛАВА 3. АНАЛИЗ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ ВЫГРУЗКИ ТРАФИКА В НЕЛИЦЕНЗИРОВАННЫЙ ДИАПАЗОН ЧАСТОТ

В главе 1 была введена модель, описывающая поведение произвольного доступа в нелицензированном диапазоне технологии NR-U и характеризующая вероятность потери сессии. Чтобы отразить особенности распределения ресурсов в лицензируемом диапазоне, была использована базовая модель СМО типа M/M/K/0. В этой главе приводится более детальная модель процесса обслуживания в лицензированном диапазоне с использованием ресурсных систем массового обслуживания, а также предлагаются 3 стратегии выгрузки трафика на нелицензированный диапазон частот.

Будем считать, что для пользовательских устройств NR-U скорость передачи, требуемая от сети, составляет $R_{\min}$. Введение ограничения на минимальную скорость соответствует требовательным к скорости приложениям, таким как потоковое видео, дополненная (*англ.* Augmented Reality, AR) и виртуальная реальность (*англ.* Virtual Reality, VR), дистанционное теле-присутствие. В зависимости от удаленности устройства от БС поддержание минимальной требуемой скорости $R_{\min}$ может потребовать разного количества физических ресурсов. Интенсивность поступления запросов на установление сессии (далее для краткости - сессий) NR-U составляет λ сессий в секунду. Предполагается, что трафик WiGig полностью эластичен, то есть скорость может динамически адаптироваться к изменяющимся условиям.

Запросы на установление сессии поступают на БС, поддерживающую две технологии - NR-U и WiGig. Устройства NR-U и WiGig устанавливают соединение с БС на основе средней мощности приема опорного сигнала (*англ.* reference signal receive power, RSRP). На практике это означает, что выбирается физически ближайшая точка доступа.

Рассматриваются следующие схемы выгрузки:

– *Базовая выгрузка* (рис. 3.1). Устройства пытаются подключиться к ближайшей базовой станции (БС) и использовать лицензированный

диапазон. Если у БС недостаточно ресурсов для обслуживания сессии, она перенаправляется на нелицензированный диапазон.

- *«Тяжелая» выгрузка.* (рис.3.2). Если объем требуемых ресурсов больше порогового значения R_1, то сессия сразу направляется в нелицензированный диапазон. В противном случае сессия сначала пытается зарезервировать ресурсы в лицензированном диапазоне БС и только при их недостатке выгружается в нелицензированный диапазон.
- *«Легкая» выгрузка*. Если объем требуемых ресурсов меньше порогового значения R_2, то сессия сразу направляется в нелицензированный диапазон. В противном случае сессия сначала отправляется в лицензированный диапазон БС и только в случае нехватки ресурсов, выгружается в нелицензированный диапазон.

Обоснование последних двух схем заключается в том, что достигнутая скорость в нелицензированном диапазоне является нелинейной функцией от числа конкурирующих пользовательских устройств. Обратим внимание, что только те пользователи, которые находятся ближе к БС чем на расстоянии r_2, где r_2 — радиус покрытия нелицензированной технологии, могут быть выгружены на нелицензированный диапазон.

Основной показатель, представляющий интерес в этом исследовании — вероятность потери сессии NR-U, т.е. вероятность прерывания сессии из-за невозможности поддерживать минимальную скорость $R_{\min}$ на БС, использующую как лицензированные, так и нелицензированные диапазоны. Обратим внимание, что этот показатель рассчитывается по-разному для рассматриваемых стратегий выгрузки. Используя эту метрику, можно вычислить плотность БС, необходимую для поддержания заданной вероятности потери сессии NR-U в присутствии конкурирующего трафика WiGig.

3.1. Ресурсная модель выгрузки трафика

Следуя описанной методологии, можно проанализировать все три определенные выше стратегии выгрузки, используя одну и ту же структуру, состоящую из двух основных компонентов: (а) система массового обслуживания со случайными требованиями к ресурсам и (б) цепь Маркова для анализа произвольного доступа к нелицензированной общей среде. Для «базовой» стратегии, см. рис. 3.1, вероятность потери из-за нехватки ресурсов на БС фактически является вероятностью выгрузки в нелицензированную полосу, а вероятность потери на нелицензированном спектре – вероятностью того, что запрошенная скорость там не удовлетворяет минимальным требованиям. Для последних двух схем, схем «тяжелой» и «легкой» выгрузки, см. рис. 3.2, решение о изначальном направлении сессии на лицензированный или нелицензированный диапазон, принимается по прибытии сессии устройства на БС. Таким образом, для этих двух схем вероятность возможных потерь определяется как вероятность того, что минимальная скорость $R_{\min}$ не будет обеспечена для сессии в нелицензированном диапазоне.

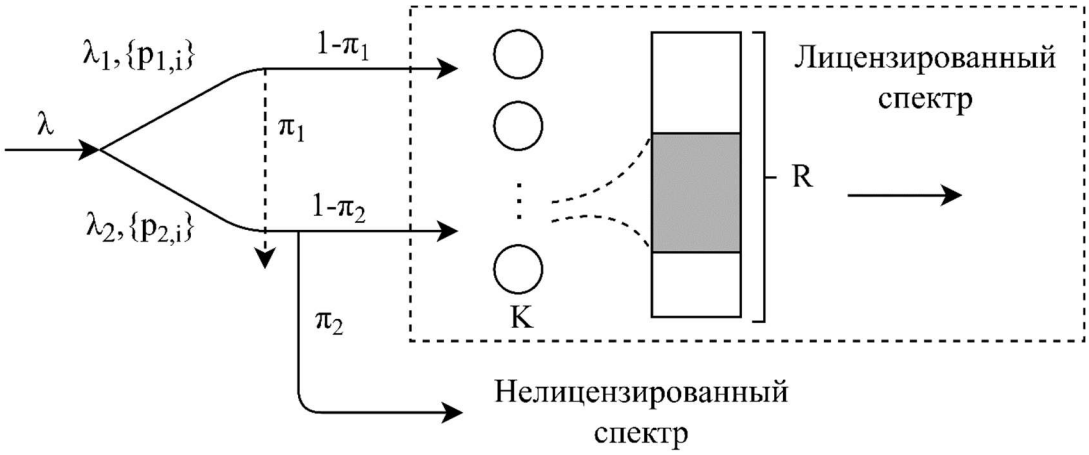

Рис. 3.1. Модель ресурсной системы массового обслуживания для базовой стратегии.

Принимая во внимание разницу между радиусами покрытия лицензированных и нелицензированных диапазонов, r_1 и r_2 (т. е. $r_1 > r_2$), вводятся два типа сессий. Сессии первого типа могут обслуживаться только в лицензированном диапазоне. Когда ресурсов для обработки этой сессии

недостаточно, сессия теряется. Местоположения пользователей, инициирующих такие сессии, находятся в зоне покрытия, представляющей собой кольцо $r_2 < x < r_1$. Сессии второго типа могут быть выгружены на нелицензированный диапазон, так как поступают из круга радиуса r_2 вокруг БС.

Пусть $\lambda = \chi r_1^2 \pi \lambda_\bullet$ — начальная интенсивность поступления заявок, где χ — плотность устройств NR на квадратный метр, $r_1^2 \pi$ — площадь зоны покрытия БС для лицензированной технологии, $\lambda_\bullet$ — интенсивность поступления сессий с устройства. Поскольку $r_2 \le r_1$, λ представляется как $\lambda = \lambda_1 + \lambda_2$, где λ_1 и λ_2 — интенсивности поступления заявок с устройств из окружности $0 < x < r_2$ и кольца $r_2 < x < r_1$ соответственно. Назовем их заявками первого и второго типа.

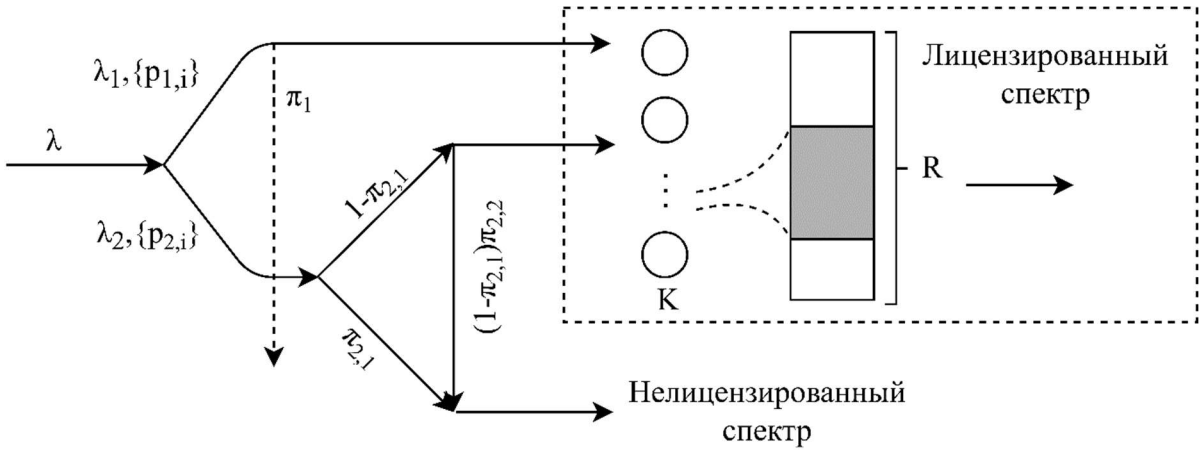

Рис. 3.2. Модель ресурсной системы массового обслуживания для стратегии «тяжелой» выгрузки.

Базовая стратегия. Для базовой стратегии интенсивность поступления на лицензированную полосу частот представляет собой сумму интенсивностей заявок обоих типов, т. е. $\lambda_1 + \lambda_2$.

Доля нагрузки, которую не может выдержать лицензированный диапазон БС, определяет интенсивность поступления заявок второго типа на нелицензированный диапазон БС. Этот параметр рассчитывается как $\lambda_2 \pi_2$, где π_2 — вероятность того, что заявка второго типа будет перенаправлена на нелицензированный диапазон. Обозначим также π_1 вероятность того, что заявка первого типа, которая не может быть выгружена в нелицензированный диапазон,

будет потеряна при поступлении на лицензированный спектр из-за нехватки ресурсов.

Стратегия «тяжелой» выгрузки. Для стратегий «тяжелой» и «легкой» выгрузки (рис. 3.2), распределение требований к ресурсам в БС зависит от порогов, R_1 и R_2, соответственно. В частности, для «тяжелой» стратегии поток поступающих заявок второго типа делится по «весу» заявки, т.е. требуемому ею объему ресурсов. "Более тяжелые" заявки изначально направляются на нелицензированный диапазон с вероятностью $\pi_{2,1}$, а с обратной вероятностью $(1-\pi_{2,1})$ "более легкие" сессии направляются на лицензированный диапазон. Таким образом, общая интенсивность поступления обоих типов заявок на лицензированном диапазоне составляет $\lambda_1 + \lambda_2\left(1-\pi_{2,1}\right)$.

Заметим, что заявки второго типа поступают на нелицензированный диапазон в двух случаях: (а) когда «вес» заявки больше определенного порога R_1, (б) когда нет достаточного количества доступных ресурсов для заявки, которая изначально была направлена на лицензированный диапазон. Вероятность π_2 того, что заявка второго типа будет направлена на нелицензированный спектр, равна сумме вероятности $\pi_{2,1}$ того, что заявка была "тяжелой", и вероятности $\pi_{2,2}$ что "легкая" заявка не может быть обработана на лицензированном диапазоне и поэтому выгружается на нелицензированный, т.е.

$$\pi_2 = \pi_{2,1} + \left(1-\pi_{2,1}\right)\pi_{2,2} \tag{3.1}$$

Обратим внимание, что интенсивность поступления на нелицензированный диапазон БС ранее была записана как $\lambda_2\pi_2$.

Стратегия «легкой» выгрузки. Аналогично стратегии «тяжелой» выгрузки, для стратегии «легкой» выгрузки интенсивности поступления на лицензированный диапазон БС можно записать как $\lambda_1 + \lambda_2\left(1-\pi_{2,1}\right)$, где вероятность $\pi_{2,1}$ представляет сбой вероятность того, что заявка - "легкая", то есть заявка, требующая меньше R_2 ресурсов и изначально направленная на нелицензированный диапазон.

В отличие от «тяжелой» стратегии заявки попадают в нелицензированный диапазон в двух случаях: (а) когда «вес» заявки меньше порогового значения R_2, или (б) когда заявка изначально была направлена на лицензированный диапазон, но объем доступных ресурсов оказался недостаточным для ее обслуживания. Принимая во внимание вышенаписанное, вероятность того, что заявка второго типа будет направлена на нелицензированный диапазон, определяется выражением, аналогичным формуле (3.1), где $\pi_{2,2}$ — вероятность того, что «тяжелая» заявка не может быть обслужена в лицензированном диапазоне и должна быть выгружена в нелицензированный диапазон. Общую интенсивность поступления на нелицензированный диапазон можно рассчитать как $\lambda_2\pi_2$.

Для моделирования процесса обслуживания заявки в лицензированном диапазоне была использована ресурсная СМО [59, 60, 61] с $K<\infty$ приборами и $R<\infty$ единицами ресурса. В систему поступают заявки двух типов, поступление заявок описывается пуассоновским процессом с интенсивностью поступления λ_1 для первого типа и λ_2 для второго. Таким образом, суммарный поступающий поток является пуассоновским с параметром $\lambda=\lambda_1+\lambda_2$. Распределение времени обработки заявок экспоненциальное со интенсивностью μ.

Для процесса обслуживания каждой заявки требуется прибор и произвольное число единиц ресурса, $0\le r\le R$. Распределения требований к ресурсам для рассматриваемых типов заявок задаются как $\left\{p_{l,j}\right\}_{j\ge 0}$, l=1,2, где $p_{l,j}$ - вероятность того, что для заявки типа l потребуется j единиц ресурсов. Согласно [59], ресурсная СМО с двумя потоками может быть проанализирована как система с одним агрегированным потоком следующим образом:

$$\tilde{p}_{1,j}=\frac{\rho_1}{\rho}p_{1,j}+\frac{\rho_2}{\rho}p_{2,j}, \qquad (3.2)$$

где предлагаемая нагрузка трафика составляет $\rho^*=\rho_1^*+\rho_2^*$, $\rho_i=\lambda_i/\mu, i=1,2$.

Система работает следующим образом. Поступающая заявка принимается системой, если на момент поступления имеется достаточное количество

доступного ресурса. В противном случае, поступающая заявка первого типа теряется, а заявка второго типа перенаправляется на нелицензированный диапазон. Когда время обслуживания заявки заканчивается, она покидает систему, освобождая все занятые ресурсы. Поведение системы можно описать случайным процессом $X(t)=\left(\xi(t),\gamma(t)\right)$, где $\xi(t)$ — число заявок в системе, а $\gamma(t)=\left(\gamma_1(t),\gamma_2(t),...,\gamma_{\xi(t)}(t)\right)$, $\gamma_i(t)$ — вектор числа единиц ресурса, выделенных для *i*-й заявки в момент времени *t*.

Обозначим через $P_k(j)$ стационарные вероятности того, что в системе есть *k* заявок, суммарно занимающих *j* ресурсов, т. е.

$$P_k(j)=\lim_{t\to\infty}P\left\{\xi(t)=k,\sum_{i=1}^{\xi(t)}\gamma_i(t)=j\right\},0\le j\le R. \tag{3.3}$$

Согласно [62], стационарное распределение определяется выражением

$$P_k(r)=P_0\frac{\rho^k}{k!}\tilde{p}_{1,r}^{(k)},\ k=1,2,...,K, \tag{3.4}$$

$$P_0=\left(1+\sum_{k=1}^{K}\frac{\rho^k}{k!}\sum_{r=0}^{R}\tilde{p}_{1,r}^{(k)}\right)^{-1}, \tag{3.5}$$

где $\left\{\tilde{p}_{1,r}^{(k)}\right\}_{r\ge0}$ - *k*-кратная свертка распределения $\tilde{p}_{1,r}$ вероятностей, которая вычисляется рекуррентно по формуле

$$\tilde{p}_{1,r}^{(k)}=\sum_{j=0}^{r}\tilde{p}_{1,r-j}^{(k-1)}\tilde{p}_{1,j},\ k\ge2, \tag{3.6}$$

где $\tilde{p}_{1,r}^{(1)}=\tilde{p}_{1,r}$, $r\ge0$.

Вероятность π_2 выгрузки заявок второго типа, т. е. вероятность того, что заявка будет перенаправлена в нелицензированный спектр БС, и вероятность π_1 потери заявки первого типа , т.е. вероятность того, что в лицензированной полосе не будет достаточного количества ресурса для обслуживания заявки первого типа, определяются выражением

$$\pi_i=1-P_0\sum_{k=0}^{K-1}\frac{\rho^k}{k!}\sum_{r=0}^{R}\sum_{j=0}^{r}p_{i,j}\tilde{p}_{1,r-j}^{(k+1)}. \tag{3.7}$$

Для больших значений K и R согласно (3.7) расчеты требуют больших вычислительных ресурсов. В этом случае можно использовать рекуррентный вычислительный алгоритм, разработанный в [60]. Обозначим

$$G(k,r)=\sum_{i=0}^{k}\frac{\tilde{\rho}^{i}}{i!}\sum_{j=0}^{r}\tilde{p}_{1,j}^{(i)},\quad P_0=G^{-1}(K,R). \tag{3.8}$$

Согласно алгоритму, функция $G(k,r)$ вычисляется рекуррентно следующим образом:

$$G(k,r)=G(k-1,r)+\frac{\tilde{\rho}}{k}\sum_{j=0}^{r}\tilde{p}_{1,j},\ 2\le k\le N,0\le r\le R, \tag{3.9}$$

с начальными условиями

$$G(0,r)=1,r\ge 0, \tag{3.10}$$

$$G(1,r)=1+\tilde{\rho}\sum_{0\le j\le r}\tilde{p}_{1,j},\ 0\le r\le R. \tag{3.11}$$

Следствие 3.1. Вероятность π_1 потери заявки первого типа и вероятность π_2 выгрузки заявки второго типа имеет вид

$$\pi_i=1-G^{-1}(K,R)\sum_{j=0}^{R}p_{i,j}G(K-1,R-j),\quad i=1,2 \tag{3.12}$$

Описанная методика расчета вероятности потери заявки в лицензированном диапазоне остается неизменной для всех рассматриваемых стратегий. Однако выбор стратегии влияет на входные параметры, включая интенсивность поступающего потока и распределение количества запрошенного ресурса.

Для стратегии «тяжелой» выгрузки вероятность того, что для заявки второго типа потребуется j ресурсов в лицензированной полосе, имеет вид

$$p_{2,j}^{\bullet}=\left(\sum_{i=0}^{R_1}p_{2,i}\right)^{-1}p_{2,j},\ 0\le j\le R_1. \tag{3.13}$$

По аналогии с предыдущей стратегией, для стратегии «легкой» выгрузки вероятность того, что для заявки второго типа требуется j ресурсов в лицензированной полосе частот, определяется как

$$p^{\bullet}_{2,j} = \left(1 - \sum_{i=0}^{R_2} p_{2,i}\right)^{-1} p_{2,j}, 0 \le j \le R_2 . \qquad (3.14)$$

Тогда, аналогично (3.2), распределение запросов ресурсов для заявок в агрегированном потоке при стратегиях «тяжелой» и «легкой» выгрузки имеет вид

$$\tilde{p}_{1,j} = \frac{\rho_1}{\rho^*} p_{1,j} + \frac{\rho_2^*}{\rho^*} p^{\bullet}_{2,j}, \qquad (3.15)$$

где предлагаемая нагрузка трафика $\rho^* = \rho_1 + \rho_2^*$, и $\rho_2^* = \lambda_1\left(1 - \pi_{2,1}\right)\big/\mu$.

По аналогии с базовой стратегией вероятность того, что заявка первого типа не может быть обслужена в лицензированном диапазоне при стратегии «тяжелой» или «легкой» выгрузки, рассчитывается по формуле (3.12).

Далее для каждой рассматриваемой стратегии мы охарактеризуем распределение вероятностей требуемых ресурсов и интенсивности поступления заявок на нелицензированный диапазон.

Для базовой стратегии вероятность того, что выгруженная заявка, изначально отправленная в лицензированный диапазон БС, потребует *j* ресурсов на нелицензированном диапазоне, определяется выражением вида

$$\tilde{p}_{2,j} = \frac{p_{2,j}}{\pi_2}\left(\sum_{r=0}^{R} P_K(r) + \sum_{k=0}^{K-1}\sum_{r=R-j+1}^{R} P_k(r)\right), 0 \le j \le R. \qquad (3.16)$$

Эта формула может быть записана с использованием описанного выше рекуррентного алгоритма. Используя функцию *G*(*k,r*), получаем

$$G(K,R) - G(K-1,R) = \sum_{i=0}^{K} \frac{\rho^i}{i!} \sum_{j=0}^{R} \tilde{p}_{1,j}^{(i)} - \sum_{i=0}^{K-1} \frac{\rho^i}{i!} \sum_{j=0}^{R} \tilde{p}_{1,j}^{(i)} = \frac{\rho^K}{K!} \sum_{j=0}^{R} \tilde{p}_{1,j}^{(K)}. \qquad (3.17)$$

Умножая полученное выражение (3.17) на P_0, получаем $\sum_{r=0}^{R} P_K(r)$. Тогда первая сумма в (3.16) может быть записана как

$$\sum_{r=0}^{R} P_K(r) = \frac{G(K,R) - G(K-1,R)}{G(K,R)}, \qquad (3.18)$$

а вторая сумма в формуле (3.16) может быть представлена как

$$\sum_{k=0}^{K-1}\sum_{r=R-j+1}^{R} P_k(r) = \sum_{k=0}^{K-1}\left(\sum_{r=0}^{R} P_k(r) - \sum_{r=0}^{R-j} P_k(r)\right). \quad (3.19)$$

Заметим, что

$$\sum_{k=0}^{K-1}\sum_{r=R-j+1}^{R} P_k(r) = \frac{G(K-1,R) - G(K-1,R-j)}{G(K,R)}. \quad (3.20)$$

Подставив получившиеся выражения (3.18) и (3.20), можно преобразовать исходную формулу (3.16) для вычисления требований к ресурсу через рекуррентный алгоритм.

Утверждение 3.1. Распределение требований к ресурсам выгружаемых заявок для базовой стратегии имеет вид

$$\tilde{p}_{2,j} = \frac{1}{\pi_2} p_{2,j} \frac{G(K,R) - G(K-1,R-j)}{G(K,R)}, 0 \le j \le R. \quad (3.21)$$

Отметим, что фактически распределение требований к ресурсам отражает распределение с.в. спектральной эффективности η, связанной с заявкой сессии, $\tilde{p}_{2,j} = P\{\eta = s_j\}$.

Для стратегии «тяжелой» выгрузки вероятность $\pi_{2,1}$, что заявка является «тяжелой», т. е. заявка требует ресурсов, превышающих R_1, и, таким образом, изначально направлена на нелицензированный спектр, может быть записана как

$$\pi_{2,1} = 1 - \sum_{j=0}^{R_1} p_{2,j} . \quad (3.22)$$

По аналогии с вероятностью потерь в «базовой» стратегии (3.7) вероятность $\pi_{2,2}$, что «легкая» заявка не может быть обработана в лицензированном диапазоне и, таким образом, выгружается на нелицензированную полосу, может быть рассчитана по формуле

$$\pi_{2,2} = 1 - P_0 \sum_{k=0}^{K-1} \frac{\rho^k}{k!} \sum_{r=0}^{R} \sum_{j=0}^{r} p^{\bullet}_{2,j} \tilde{p}^{(k+1)}_{1,r-j}. \quad (3.23)$$

После использования рекуррентного алгоритма формула (3.23) может быть записана аналогично формуле (3.16):

$$\pi_{2,2} = 1 - G^{-1}(K,R)\sum_{j=0}^{R} p_{2,j}^{\bullet} G(K-1, R-j). \tag{3.24}$$

Вероятность того, что для заявки требуется j ресурсов в нелицензированном диапазоне, необходимо рассчитывать отдельно для двух случаев: (а) когда заявка является «тяжелой» и, таким образом, изначально направляется в нелицензированный диапазон, и (б) когда заявка сначала направляется на лицензированный диапазон, но ресурсов для ее обслуживания оказалось недостаточно. Данное поведение можно описать следующим образом:

$$\tilde{p}_{2,j} = \begin{cases} \dfrac{1-\pi_{2.1}}{\pi_2} p_{2,j} \left(\sum_{r=0}^{R} P_K(r) + \sum_{k=0}^{K-1} \sum_{r=R-j+1}^{R} P_k(r) \right), j \le R_1, \\ \dfrac{1}{\pi_2} p_{2,j}, \; j > R_1. \end{cases} \tag{3.25}$$

Используя полученные выше формулы (3.16) и (3.21) и подставляя их в (3.25) получаем формулу вероятности того, что заявка требует j ресурсов в нелицензированном диапазоне.

Утверждение 3.2. Вероятность того, что заявка, выгруженная в нелицензированный диапазон, потребует j ресурсов для стратегии «тяжелой» выгрузки, вычисляется согласно формуле

$$\tilde{p}_{2,j} = \begin{cases} \dfrac{1-\pi_{2.1}}{\pi_2} p_{2,j} \dfrac{G(K,R) - G(K-1, R-j)}{G(K,R)}, j \le R_1, \\ \dfrac{1}{\pi_2} p_{2,j}, \; j > R_1. \end{cases} \tag{3.26}$$

Для стратегии «легкой» выгрузки вероятность $\pi_{2,1}$ того, что заявка является «легкой», т. е. заявка требует менее R_2 ресурсов и, таким образом, изначально сразу после поступления выгружается на нелицензированный диапазон, определяется выражением

$$\pi_{2,1} = \sum_{j=0}^{R_2} p_{2,j} \,. \tag{3.27}$$

Вероятность $\pi_{2,2}$ того, что "тяжелая" заявка не может быть обслужена в лицензированном диапазоне и, таким образом, переведена в нелицензированный,

вычисляется по аналогии с формулой (3.23) стратегии «тяжелой» выгрузки для интервала $j \in [0, R_2]$.

Аналогично (3.25), вероятность $\tilde{p}_{2,j}$ того, что потребителю потребуются j ресурсов в нелицензированном диапазоне, имеет вид

$$\tilde{p}_{2,j} = \begin{cases} \dfrac{1-\pi_{2.1}}{\pi_2} p_{2,j} \left(\sum\limits_{r=0}^{R} P_K(r) + \sum\limits_{k=0}^{K-1} \sum\limits_{r=R-j+1}^{R} P_k(r) \right), j \geq R_2, \\ \dfrac{1}{\pi_2} p_{2,j}, j < R_2. \end{cases} \tag{3.27}$$

Утверждение 3.3. Вероятность того, что заявка, выгруженная в нелицензированный диапазон, потребует j ресурсов для стратегии «легкой» выгрузки, вычисляется согласно формуле

$$\tilde{p}_{2,j} = \begin{cases} \dfrac{1-\pi_{2.1}}{\pi_2} p_{2,j} \dfrac{G(K,R) - G(K-1, R-j)}{G(K,R)}, j \geq R_2, \\ \dfrac{1}{\pi_2} p_{2,j}, j < R_2. \end{cases} \tag{3.28}$$

Теперь перейдем к оценке эффективности выгрузки на нелицензированном спектре. Напомним, что для всех рассмотренных стратегий сессия NR-U может быть выгружена на нелицензированный диапазон, где она конкурирует с сессиями WiGig за ресурсы. Чтобы получить вероятность того, что для выгруженной сессии NR-U выполняются требования минимальной скорости $R_{\min}$, необходимо сначала определить вероятность успешной передачи пакетов для пользовательских устройств NR-U и WiGig в нелицензированном диапазоне.

Пусть $\varsigma(n,m)$ — вероятность коллизии при n устройств NR и m устройств WiGig, а ψ — вероятность того, что путь прямого распространения сигнала заблокирован. Тогда вероятность успеха попытки передачи может быть выражена как $\theta = \theta(n,m) = (1-\varsigma(n,m))(1-\psi)$.

Поведение системы описывается цепью Маркова $\{X_k, k \geq 0\}$, аналогичной заданной в упрощенной модели из раздела 1.3. Здесь $X_k = i$, $i = 0,1,...,T$, обозначает порядковый номер попытки передачи, когда значение таймера обратного отсчета

находится в интервале $[0, 2^i W - 1]$, и W — минимальное значение конкурентного окна.

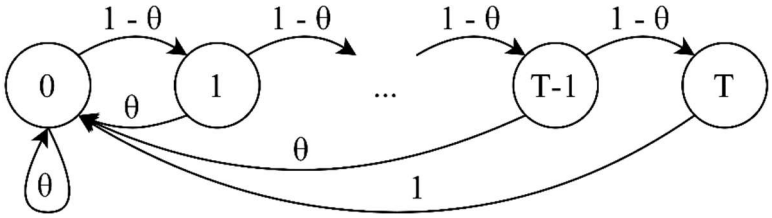

Рис. 3.3. Граф переходов между состояниями цепи Маркова, показывающих число неудачных попыток передачи.

Стационарные вероятности q_i для такой цепи Маркова, среднее число тактов b_i, которое проведет заявка в состоянии *j,* вычисляются аналогично формулам (1.24), (1.25) и (1.26), полученным в разделе 1.3.

Утверждение 3.4. Вероятность $\pi_i^*(n,m), i = 1,2$ того, что устройство совершит попытку передачи, и вероятность коллизии $\varsigma(n,m)$ являются решением системы уравнений:

$$\begin{cases} \pi_i^*(n,m) = \left[\dfrac{\theta(n,m) W \left(1 - 2^{T_i+1} \left(1 - \theta(n,m)\right)^{T_i+1}\right)}{2\left(1 - \left(1 - \theta(n,m)\right)^{T_i+1}\right)\left(2\theta(n,m) - 1\right)} + \dfrac{1}{2} \right]^{-1}, \quad i = 1,2, \\ \varsigma(n,m) = (1-\psi)\left(1 - \pi_1^*(n,m)\right)^n \left(1 - \pi_2^*(n,m)\right)^m. \end{cases} \tag{3.29}$$

Утверждение 3.5. Вероятность успешной передачи устройств NR имеет вид

$$\Pi_1^* = \sum_{i=1}^{\infty} \frac{\left(\rho_1^*\right)^n}{n!} e^{-\rho_1^*} \sum_{j=0}^{\infty} \frac{\left(\rho_2^*\right)^m}{m!} e^{-\rho_2^*} \pi_1^*(n,m)\theta(n,m), \tag{3.30}$$

где $\rho_1^* = \lambda_2 \pi_2 / \mu$ - нагрузка, вызванная устройствами NR, $\rho_2^* = \vartheta / \omega$ - нагрузка, вызванная устройствами WiGig, ϑ и ω - интенсивности поступления и обслуживания для устройств WiGig, соответственно.

Пусть ν — с.в. скорости передачи данных на нелицензированном диапазоне частот. С.в. ν скорости передачи в нелицензированном диапазоне частот является линейной функцией случайной величины спектральной эффективности,

однозначно определяемой для каждого значения j числа единиц ресурса, с распределением $\tilde{p}_{2,j}$.

Следствие 3.4. С.в. скорости ν на нелицензированном диапазоне частот является линейной функцией от с.в. спектральной эффективности η с распределением $\left\{\tilde{p}_{2,j}\right\}_{j\geq 0}$:

$$\nu = \Pi_1^* B\eta \text{, где } P\{\eta = s_j\} = \tilde{p}_{2,j}. \tag{3.31}$$

Здесь B - ширина полосы пропускания на нелицензированном спектре.

Теперь можно оценить ожидаемое значение скорости передачи данных, достигаемой пользовательским устройством в нелицензированной полосе частот.

Следствие 3.5. Средняя скорость передачи, достигаемая устройством NR в нелицензированном диапазоне, составляет

$$\mathrm{M}\,\nu = \sum_{j=0}^{R} \tilde{p}_{2,j} \Pi_1^* B s_j. \tag{3.32}$$

Получив скорость, достижимую на пользовательском устройстве NR-U в нелицензированном диапазоне, можно определить вероятность потери сессии NR-U.

Следствие 3.6. Вероятность снижения скорости устройства NR ниже требуемого порога имеет вид

$$Q = \pi_2 P\left\{\nu < R_{\min}\right\} = \pi_2 \sum_{j:\, \Pi_1^* B s_j < R_{\min}} \tilde{p}_{j,2}. \tag{3.33}$$

3.2. Численный анализ показателей эффективности стратегий выгрузки

В этом разделе представлены численные результаты. Имея дело со сложными стратегиями обслуживания, сначала будет проведена оценка производительности процедуры произвольного доступа как функции системных параметров. Затем проведен анализ показателей системы, включая возможную вероятность потери сессии и сравнивая предлагаемые стратегии выгрузки.

Перед тем, как представить численные результаты анализа полученных показателей, опишем параметризацию системы.

В качестве модели блокировки рассматривалась модель (см. рис. 3.4), в которой учитывается перекрытие пешеходами путей распространения сигнала. Предполагается, что пешеходы движутся по плоскости в случайном направлении [63], со скоростью v м/с и экспоненциально распределенной длиной пробега со средним значением τ метров. Пешеходы моделируются в виде цилиндров высотой h_3 и радиусом r_3. На практике h_3 — средний рост человека, 1,7 м.

Обозначенная в формуле (1.21) через $\psi(x)$ вероятность блокировки пути прямого распространения сигнала представляет собой вероятность того, что пользовательское устройство, расположенное на расстоянии x от БС, заблокировано. Обобщая результаты из [63,64], вероятность $\psi(x)$ вычисляется по формуле

$$\psi(x) = 1 - e^{-2\lambda_3 r_3\left(x\frac{h_3-h}{h_*-h}+r_3\right)}, \tag{3.34}$$

где λ_3 - плотность пешеходов.

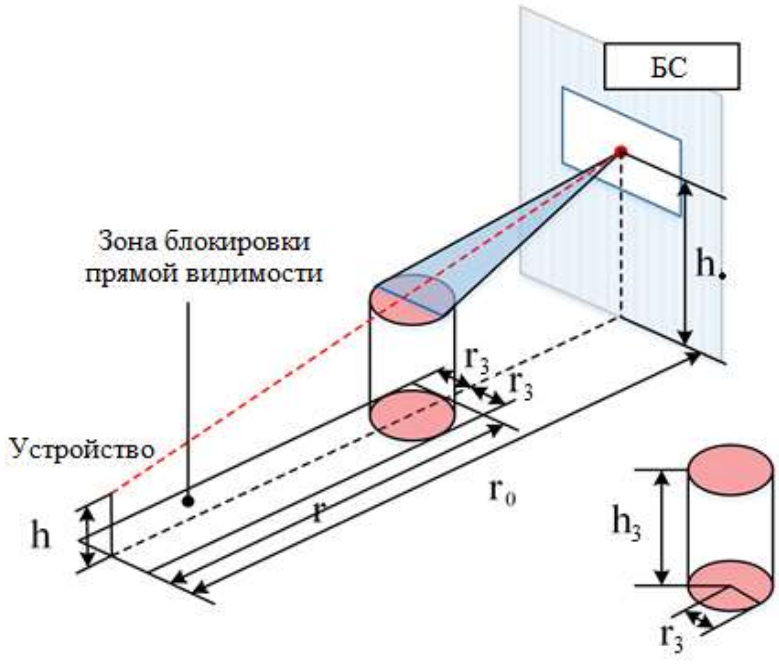

Рис. 3.4. Модель блокировки пути прямого распространения сигнала.

Учитывая, что устройства NR-U распределены по точечному пуассоновскому процессу на плоскости для лицензированного спектра БС, а также принимая во внимание значения спектральной эффективности модуляционно-кодовых схем, среднюю спектральную эффективность можно получить по формуле

$$\mathrm{M}\,\eta = \int_0^{r_1} \frac{2x}{r_1} \log_2\left(1 + S(x)\right) dx, \tag{3.35}$$

где *S(x)* – отношение сигнал–шум плюс помехи (*англ.* Signal-to-Interference & Noise Ratio, SINR) на приемнике, расположенном на расстоянии *x*, которое вычисляется по следующей формуле:

$$S(x) = Cx^{-2.1}\left[1 - \psi(x)\right] + Cx^{-3.19}\left[\psi(x)\right], \quad C = \frac{PAU}{\left(N_0 B + M\right)10^{2\log_{10} f + 3.24}}, \tag{3.36}$$

где P — мощность передачи пользовательского устройства; A и U — коэффициенты усиления антенной решетки на БС и пользовательском устройстве, соответственно; N_0 — спектральная плотность мощности шума; B — рабочая полоса пропускания; f — рабочая частота, ГГц; M — параметр интерференции.

Кроме того, для определения искомых параметров требуется определить эффективные радиусы покрытия NR и WiGig частей БС, r_1 и r_2. Поскольку оба радиуса получаются аналогично, ниже рассматривается только r_1. Эффективный радиус r_1 покрытия БС для лицензированного спектра равен минимуму от расстояния между БС $r_{1,1}$ и радиуса максимального покрытия БС на лицензированном спектре $r_{1,2}$, т. е. $r_1 = \min\left(r_{1,1}, r_{1,2}\right)$. Ниже приведены эти компоненты.

Радиус $r_{1,1}$ определяется как максимальное расстояние между пользовательским устройством и БС, так что пользовательское устройство в условиях блокировки пути прямого распространения сигнала не находится в условиях простоя.

$$r_{1,1} = \sqrt{\left(S^* M^* / C\right)^{0.627} - \left(h_3 - h\right)^2}, \tag{3.37}$$

где *S** является значением отношения сигнал–шум, соответствующим самой низкой возможной модуляционно-кодовой схеме (*англ.* modulation and coding scheme, MCS), *M** — граница теневого замирания.

Радиус $r_{1,2}$, равный половине расстояния между соседними БС , определяется аппроксимацией кругом ячейки Вороного, образованной местоположением БС на

плоскости. Поскольку фактическая площадь ячейки Вороного неизвестна [78], используется компьютерное моделирование с входным параметром плотности БС, чтобы получить $r_{1,2}$. Радиус покрытия БС по технологии WiGig r_2 получается аналогично.

Теперь, получив радиусы покрытия можно определить среднее значение вероятности блокировки пути прямого распространения сигнала

$$\psi = \int_0^{r_1} \psi(x)\left((2x)/r_1^2\right)dx .$$

Таблица 3.1. Параметры системы по умолчанию.

Параметр	**Значение**
Рабочие частоты NR/WiGig, f	28/60 ГГц
Ширина полосы пропускания NR/WiGig, B_1, B_2	400МГц / 2,16 ГГц
Высота БС, $h_{\bullet}$	10 м
Радиус блокатора, r_3	0,2 м
Высота блокатора, h_3	1,7 м
Высота пользовательского устройства, h	1,5 м
Мощность передачи NR/WiGig, P	33/23 дБ ▪ м
Тепловой шум, N_0	−174 дБ м/Гц
Параметр интерференции, M	3 дБ
Порог мощности приема сигнала, S*	−9 дБ
Интенсивность блокаторов, λ_3	0,3

Остальные входные параметры системы рассматривались как постоянные, значения которых, использованные для расчета, приведены в таблице 3.2.

Таблица 3.2. Параметры математической модели по умолчанию.

Параметр	**Значение**
Начальный размер конкурентного окна, W	16
Максимальное число заявок, K	10
Число повторных попыток передач, T	10

Параметр	**Значение**
Интенсивность обслуживания, μ	0.2
Минимальная требуемая скорость сессии, $R_{\min}$	50 Мбит/с
Вероятность блокировки пути прямого распространения сигнала	0.166

Начнем с анализа производительности механизмов произвольного доступа в зависимости от параметров системы. С этой целью на рисунках ниже (рис. 3.5-3.7) показана вероятность коллизии как функции от количества активных пользовательских устройств NR-U для различных значений плотности блокаторов, λ_3, начального размера конкурентного окна, *W*, и число попыток повторной передачи, *T*.

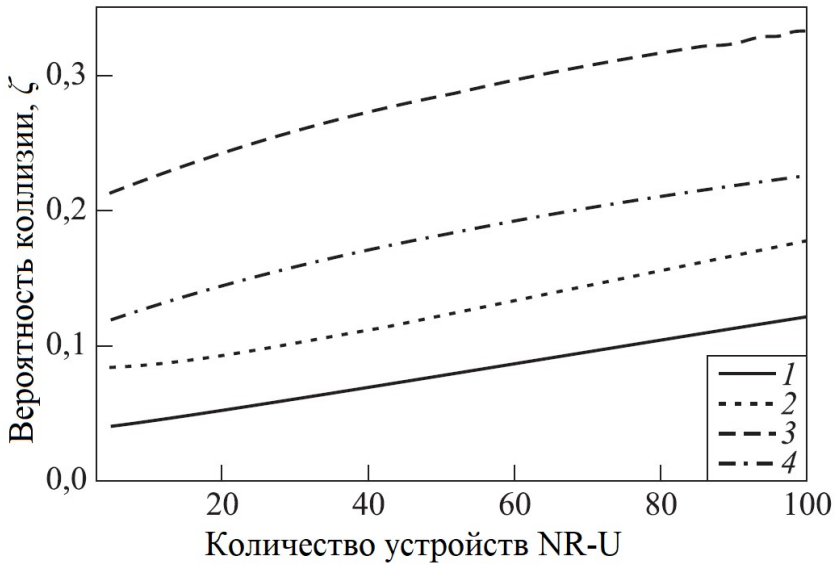

Рис. 3.5. Вероятность коллизии $\varsigma(n,m)$ от числа пользовательских устройств NR-U для разных плотностей блокаторов:

1 — λ_3 = 0,1; 2 — λ_3 =0,3; 3 — λ_3 = 0,5; 4 — λ_3 =0.7.

Анализируя данные, представленные на рис. 3.5, можно заметить, что, вероятность коллизий увеличивается с ростом числа устройств NR-U. Однако зависимость от плотности блокаторов более сложная. В частности, с увеличением значения λ_3 сначала увеличивается вероятность коллизии. Это объясняется тем фактом, что большее количество устройств сталкивается с блокировкой, как только

они выигрывают конкуренцию за среду передачи. Однако дальнейшее увеличение плотности блокаторов, например при λ_3=0.7, наоборот, приводит к уменьшению вероятности коллизии. Этот эффект вызван увеличением значения текущего размера конкурентного окна в результате неудачных попыток передачи. Рассматриваемая метрика также сильно зависит от других значений системных параметров, включая начальное конкурентное окно и число попыток повторной передачи. Такое поведение отрицательно влияет на производительность задержки системы.

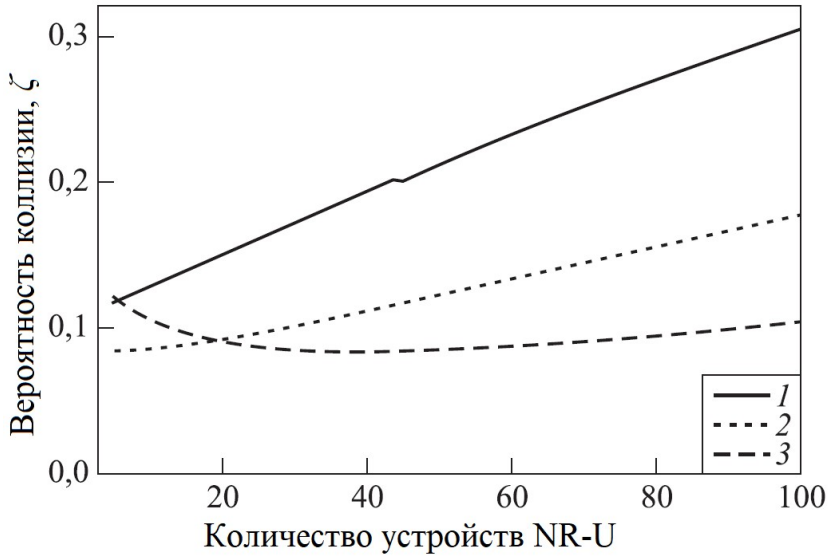

Рис. 3.6. Вероятность коллизии от числа пользовательских устройств NR-U для нескольких размеров конкурентного окна *W: 1 —W = 8; 2 — 16; 3 —W = 32*.

Влияние начального конкурентного окна на вероятность коллизий показано на рис. 3.6. Здесь мы снова наблюдаем поведение, нехарактерное для низкочастотных систем, особенно при больших значениях начального конкурентного окна. При малых значениях конкурентного окна (например, *W*=8 и *W*=16) и рассматриваемой вероятности блокировки λ_3=0,3 наблюдается линейный рост вероятности коллизии. Однако при *W*=32 поведение рассматриваемой метрики более сложное с явным минимумом, достигаемым примерно при 40 устройствах. Это объясняется эффектом блокировки, которая положительно влияет на вероятность коллизии для небольшого числа активных устройств. Однако, при

дальнейшем росте числа устройств, вероятность коллизии снова начинает увеличиваться.

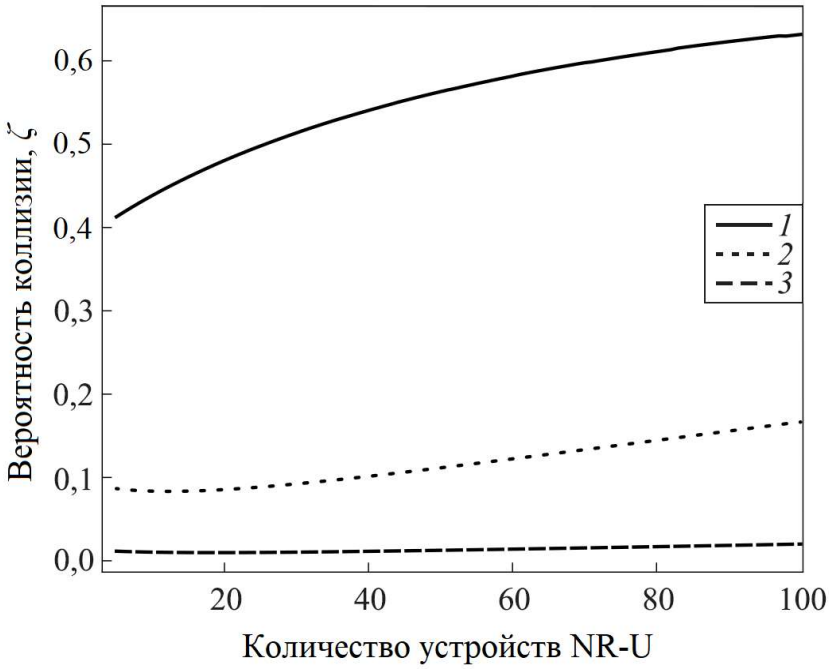

Рис. 3.7. Вероятность коллизии от числа пользовательских устройств NR-U для разного числа повторных передач *T*: *1 —T = 5; 2 — T = 10; 3 —T = 15*.

Результат попыток повторной передачи показан на рис. 3.7. Мы видим ожидаемое уменьшение вероятности коллизии при увеличении количества повторных передач. Тем не менее, для небольшого числа активных устройств мы замечаем эффект блокаторов, который снижает вероятность коллизий. Однако, начиная уже с 10 устройств, обе кривые демонстрируют последовательное линейное увеличение.

Теперь мы можем перейти к сравнению рассмотренных стратегий выгрузки. С этой целью ниже мы рассмотрим три варианта стратегий: (а) базовая, когда клиент освобождается от нагрузки, когда в лицензированном диапазоне нет доступных ресурсов, (б) стратегия «тяжелой» выгрузки, в которой «более тяжелые» клиенты изначально направляются на нелицензированный диапазон, и (в) стратегия «легкой» выгрузки при которой «более легкие» клиенты изначально направляются в нелицензированный диапазон.

На рис. 3.8 показана вероятность успешной передачи для всех рассмотренных стратегий. Обратите внимание, что базовая стратегия и стратегия «тяжелой»

выгрузки дают примерно одинаковую вероятность успешной передачи данных. При небольшой суммарной предлагаемой нагрузке все три стратегии дают примерно одинаковый результат, при увеличении нагрузки выигрыш сохраняется за базовой стратегией и стратегией «тяжелой» выгрузки.

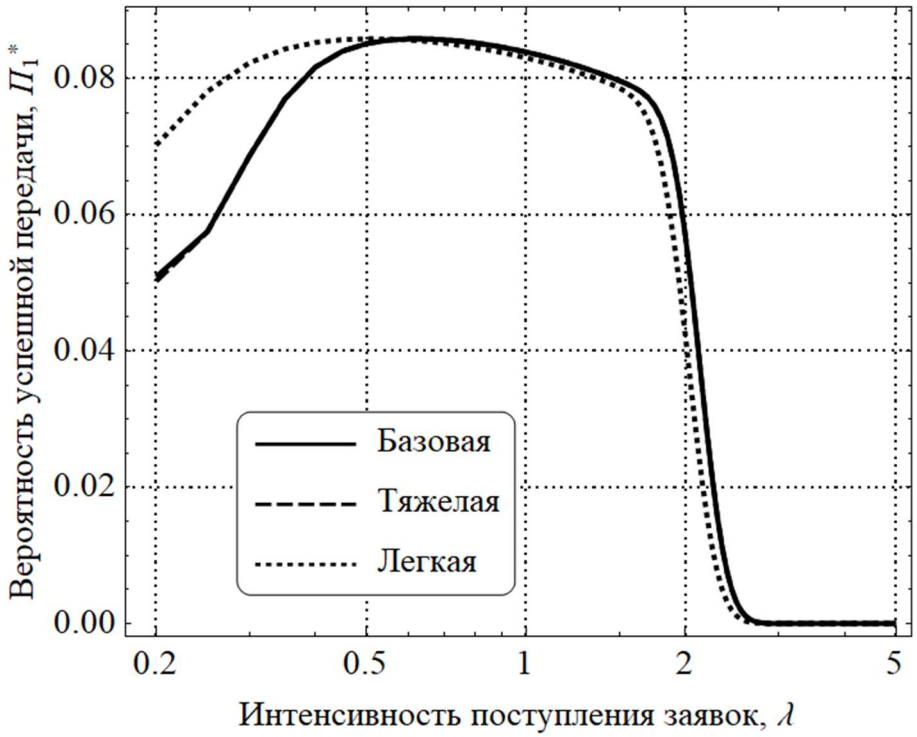

Рис. 3.8. Вероятность перенаправления заявок для разных стратегий.

На следующем рис. 3.9 показана скорость передачи данных, достигнутая устройством в нелицензированном диапазоне для всех рассматриваемых стратегий. Наибольший прирост скорости достигается при стратегии «легкой» выгрузки в основном для легких нагрузок. При увеличении нагрузки происходит рост вероятности передачи (Рис. 3.8), а вместе с ним рост скорости передачи (Рис. 3.9), так как появляется больше сессий на нелицензированной полосе WiGig за счет переполнения NR, и нелицензированная полоса еще не перегружена. После этого наступает насыщение и вероятность успешной передачи и скорость передачи начинают падать. Падение сначала медленное, так как коллизии разрешаются более-менее равномерно, и наконец случайный доступ на полосе WiGig не выдерживает нагрузки, и параметры начинают резко падать, наличие такого режима – особенность случайного доступа.

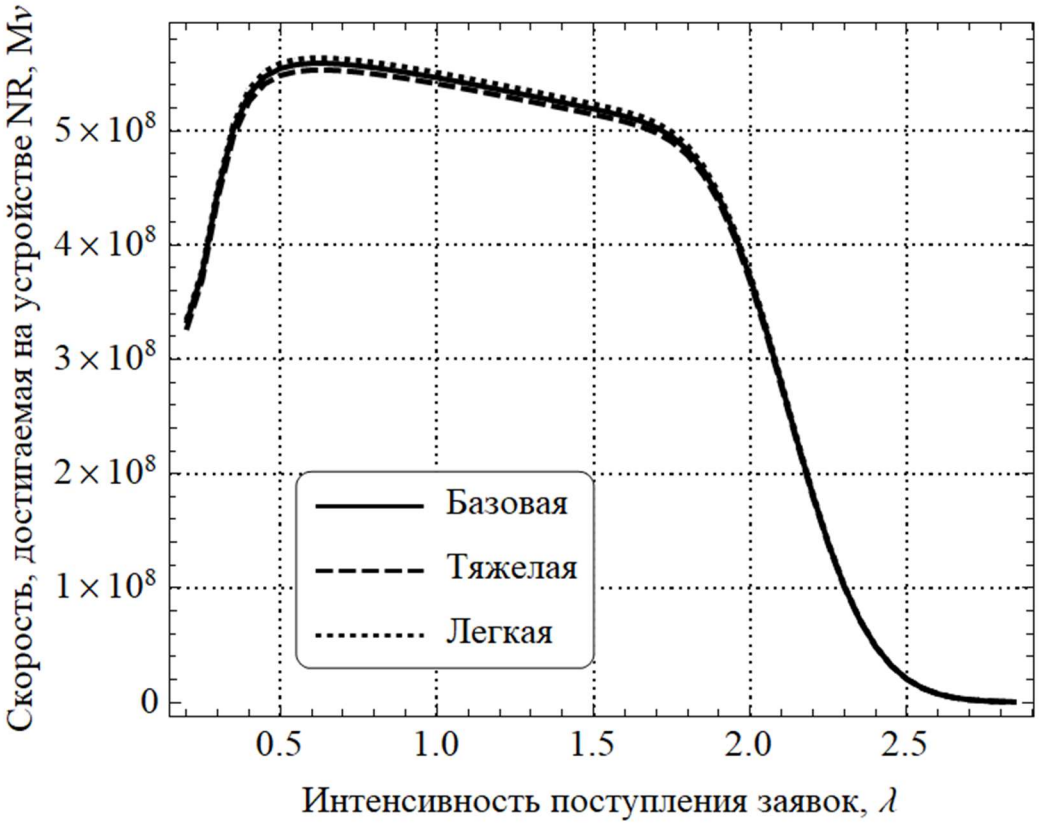

Рис. 3.10. Скорость, достигаемая устройством NR на нелицензированном диапазоне (бит/с), для разных стратегий.

Рис. 3.10 иллюстрирует итоговую вероятность снижения скорости ниже заданного порога для всех рассмотренных стратегий и требованиям к ресурсам сессии: $R_{\min} = 50$ Мбит/с. Анализируя представленные результаты, можно заметить, что базовая стратегия, при которой сессия выгружается на нелицензированную полосу, когда ресурсы для ее обслуживания отсутствуют в лицензированной, связана с минимальными значениями вероятностей окончательного сброса сессии. При этом выгрузка тяжелых сессий на нелицензированный диапазон приводит к наибольшему выигрышу стабильно на всем рассматриваемом предложенном интервале нагрузки, тогда как стратегия «тяжелой» выгрузки, при которой «тяжелые» сессии выгружаются на нелицензированный диапазон, характеризуется худшей производительностью.

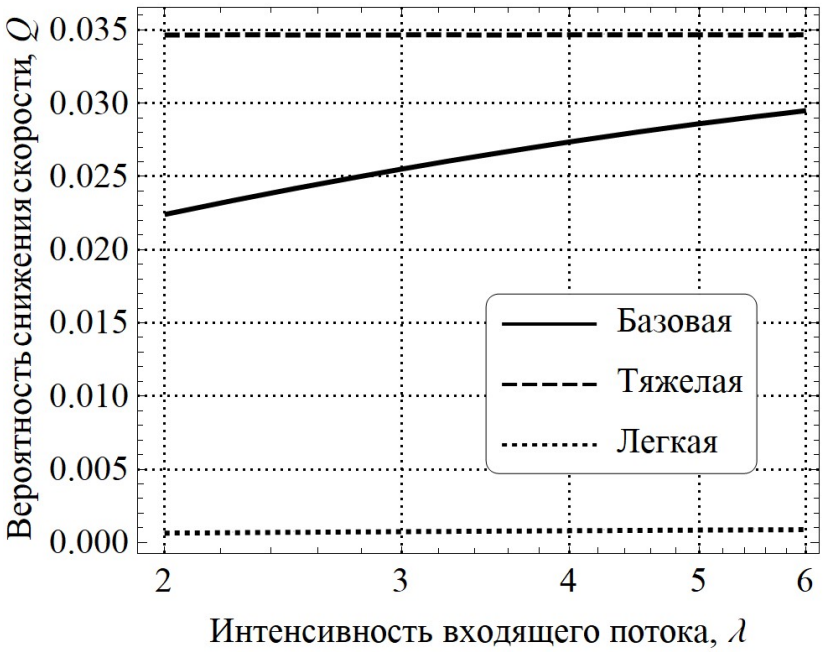

Рис. 3.11. Вероятность снижения скорости ниже требуемого порога $R_{\min}$ устройства NR от интенсивности поступления заявок.

Таким образом, в данном разделе с использованием инструментов теории массового обслуживания, описана ресурсная модель с стратегиями выгрузки заявок на нелицензированный диапазон, основанными на «весе» заявки, и было получено распределение требований к ресурсам выгруженных заявок. Было проведено сравнение 3 стратегий выгрузки: базовая, при которой сессии выгружается, когда в лицензированном диапазоне нет доступных ресурсов, стратегия «тяжелой» выгрузки, при которой «тяжелые» заявки изначально направляются в нелицензированный диапазон, и стратегия «легкой» выгрузки, при которой «более легкие» заявки, изначально направляются на нелицензированную часть диапазона. Для параметров, представленных в таблице 3.1, можно отметить, что наиболее выигрышной с точки зрения производительности стратегией выгрузки для этого случая является базовая стратегия. Также можно отметить, что значения характеристик для стратегий базовой и «тяжелой» выгрузки оказались очень близкими, поскольку в базовой мы перенаправляли на нелицензированную полосу все то, что не удовлетворяло требованиям к ресурсам, что по логике близко к выгрузке тяжелых заявок.

ЗАКЛЮЧЕНИЕ

В заключении диссертационной работы сформулируем основные результаты и выводы диссертации.

1. Разработана модель выгрузки задач с мобильных устройств по пороговому значению объема вычислений в туманно-облачную инфраструктуру. Модель в виде сети массового обслуживания содержит три типа узлов, моделирующие обработку задач на мобильных устройствах, узлах туманных и облачных вычислений. Получена формула для функции распределения времени отклика системы.
2. Разработана модель двухпараметрического механизма выгрузки задач мобильных вычислений в туманно-облачную инфраструктуру. Модель учитывает неоднородность задач по объему вычислений и размеру данных для выгрузки. В предположении, что объем вычислений и объем данных имеют гамма-распределение, получено преобразование Лапласа-Стилтьеса функции распределения времени отклика системы. Модель позволяет вычислять энергопотребление мобильных устройств в условиях выгрузки задач.
3. Разработана модель выгрузки трафика из лицензированного диапазона в нелицензированный спектр частот мобильной сети. Модель состоит из двух компонентов – ресурсная система массового обслуживания для анализа передачи трафика в лицензированном диапазоне и цепь Маркова, моделирующая механизм случайного доступа к среде передачи в нелицензированном диапазоне. В первой модели были получены дискретное распределение с.в. требований выгружаемых сессий к ресурсам СМО и вероятность выгрузки заявок в нелицензированный диапазон, во второй модели - дискретное распределение скорости передачи, достигаемой в нелицензированных частотах, с учетом вероятности успешной передачи и дискретного распределения первой модели в терминах ресурсных СМО.

СПИСОК ОСНОВНЫХ СОКРАЩЕНИЙ

3GPP	-	3rd Generation Partnership Project, партнерский проект развития связи 3-го поколения, консорциум, разрабатывающий спецификации для мобильной телефонии
БС	-	Базовая станция
ПЛС	-	Преобразование Лапласа-Стилтьеса
СВ	-	Случайная величина
СМО	-	Система массового обслуживания
СУР	-	Система уравнений равновесия
ФР	-	Функция распределения
AR	-	Augmented Reality, дополненная реальность
CCA	-	Clear Channel Assesment, оценка доступности канала
DFS	-	Dynamic Frequency Selection, динамический выбор частоты
DL	-	Downlink, нисходящая линия связи
ECCA	-	Extended Clear Channel Assesment, расширенная оценка доступности канала
EWS	-	Early Warning Score, оценка раннего предупреждения
FCFS	-	First Come First Serve, первый пришел - первый обслужился
FSDN	-	Fog Software Defined Networking, программно-конфигурируемые сети туманных вычислений
GBD	-	Generalized Benders decomposition, обобщенное разложение Бендера
H-ADMM	-	Hybrid Alternating Direction Method of Multipliers, гибридный метод множителя с переменным направлением
IoT	-	Internet of Things, интернет-вещей

LAA	-	License Assisted Access, сетевая технология 4-го поколения
LBR	-	Listen Before Receive, прослушивание до получения
LBT	-	Listen Before Talk, прослушивание до разговора
LoS	-	Line of Sight, прямое распространение сигнала
LTE	-	Longterm Evolution, сетевая технология 4-го поколения
NR	-	New Radio, Новое радио, сетевая технология 5-го поколения
PPP	-	Poisson point process, точечному процессу Пуассона
QoS	-	Quality of Service, качество предоставления услуг
RAT	-	Radio Access Technology, технология радиодоступа
SDN	-	Software Defined Networking, программно-конфигурируемые сети
SLA	-	Service Level Agreement, Соглашение о качестве услуг
TxOp	-	Transmit opportunity, возможность передачи
VANET	-	Vehicular ad-hoc network, автомобильные самоорганизующиеся сети
VR	-	Virtual Reality, виртуальная реальность

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

Глава 1 Раздел 1.2

M – количество узлов мобильных вычислений

N количество виртуальных машин

w_i – случайная величина, объем вычислений для *i*-го узла

$W_i(x)$ – ФР объема вычислений задач с *i*-го узла мобильных вычислений

$V_i(x)$ – объема вычислений для задачи, которая обрабатывается на *i*-м узле, $i = 1,...,M + 2$;

$p_{i,j}$ – вероятность того, что для обработки задачи с *i*-го узла мобильных вычислений потребуются *j*

$\{p_{i,j}\}_{j\geq 1}$ – дискретное распределение с.в. w_i

β_i – случайная величина, время отклика *i*-го узла

$B_i(x)$ – ФР времени отклика *i*-го узла

t_i – случайная величина, время обработки задачи на *i*-м узле, $i = 1,...,M + 2$;

$T_i(x)$ – ФР времени обработки задачи на *i*-м узле

w^* – порог объёма вычислений для выгрузки с узла мобильных вычислений в узел туманно-облачных вычислений;

π_i – вероятность выгрузки с узла *i* на узел туманных вычислений *M*+1 из-за превышения порога

π_{M+1} – вероятность выгрузки с узла *M+1* на узел *M+2* облачных вычислений из-за перегрузки узла

λ_i – интенсивность поступления задач на i-й узел

μ_i – скорость обработки задач на *i*-м узле, $i = 1,...,M + 2$;

Δ_1 – Задержка передачи между узлом мобильных вычислений и узлом туманных вычислений

Δ_2 – Задержка передачи между узлом туманных вычислений и узлом облачных вычислений

$t_i^{(1)}$ Среднее время обработки задачи на *i*-м узле

$\beta_i^{(1)}$ – среднее время отклика задачи с *i*-го узла мобильных вычислений

$\beta^{(1)}$ – среднее время отклика задачи с произвольного узла мобильных вычислений

$B(t^*)$ – вероятность того, что время отклика превысит порог t^*

Глава 1 Раздел 1.3

λ – интенсивность поступления заявок на лицензированный спектр

μ – интенсивность обслуживания

π – вероятность выгрузки заявки

p_i – стационарные вероятности наличия в системе *i* активных заявок пользовательских устройств

r_1 – радиус покрытия технологией NR

r_2 – радиус покрытия технологией WiGig

$\varsigma(i,j)$ – коллизии при *i* устройств NR и *j* устройств WiGig

ψ – вероятность блокировки пути прямого распространения сигнала

$\theta(i,j)$ – вероятность успешной передачи

T – максимальное количество повторных передач с удвоением таймера обратного отсчета

W – размер начального окна конкурентного доступа

$\pi_1^*(i,j)$ – вероятность совершения попытки передачи во временном слоте устройства NR

$\pi_2^*(i,j)$ – вероятность совершения попытки передачи во временном слоте устройства WiGig

Π_1^* – вероятность успешной передачи устройств NR

Π_2^* – вероятность успешной передачи устройств WiGig

ν – случайная величина, скорость передачи данных на нелицензированном диапазоне частот

$R_{\min}$ – минимальная требуемая скорость передачи

Q – вероятность снижения скорости устройства NR ниже требуемого порога

Глава 2

s_i – случайная величина, объем данных для выгрузки для *i*-го узла

$S_i(x)$ – ФР объема данных для выгрузки задач с *i*-го узла мобильных вычислений

w^* – порог объёма данных для выгрузки с узла мобильных вычислений в узел туманно-облачных вычислений

$G_i(x)$ – ФР размера передаваемого файла выгруженной

$D_i(x)$ – ФР времени передачи в беспроводной сети

$\tilde{T}_i$ – ПЛС времени обработки задачи в *i*-м узле

ω_i – ПЛС распределения времени ожидания

$\phi_i(s)$ – ПЛС распределения времени пребывания на *i*-м узле мобильных вычислений

$\tilde{D}_i(s)$ – ПЛС распределения времени передачи в беспроводной сети

R – скорость передачи в беспроводной сети

$\psi_i(s)$ – ПЛС распределения времени ожидания в беспроводной сети

$\varphi_i(s)$ – ПЛС времени пребывания в беспроводной сети

$\tilde{\beta}_i(s)$ – ПЛС распределения времени отклика задачи из i-го узла мобильных вычислений

$\tilde{\beta}(s)$ – ПЛС распределения времени отклика задачи из произвольного узла мобильных вычислений

$E_{1,i}$ – среднее потребление энергии $E_{1,i}$ при локальной обработке задачи на i-м мобильном устройстве

$E_{2,i}$ – среднее потребление энергии i-го мобильного устройства во время передачи

$P_{1,i}$ – потребляемая мощность во время обработки задачи i-м мобильным устройством

$P_{2,i}$ – мощность передачи во время обработки задачи с i-го узла мобильных вычислений

E – среднее потребление энергии для задачи с произвольного мобильного устройства

Глава 3

λ_i – Интенсивность поступления заявок i-го типа

R – Единицы ресурсов

$p_{l,j}$ – вероятность того, что для заявки типа l потребуется j ресурсов

$\{p_{l,j}\}_{j\geq 0}$ – Распределение требований к ресурсам заявок i-го типа

ρ – предлагаемая нагрузка трафика

$P_k(j)$ – стационарные вероятности того, что в системе есть k заявок, суммарно занимающих r ресурсов

π_1 – вероятность потери заявки первого типа

π_2 – вероятность выгрузки заявки второго типа

$\{\tilde{p}_{1,j}\}$ – Распределение требований к ресурсам заявок лицензированного спектра

$\{\tilde{p}_{2,j}\}$ – Распределение требований к ресурсам заявок лицензированного спектра

СПИСОК ЛИТЕРАТУРЫ

1. *R. K. Naha et al.,* Fog Computing: Survey of Trends, Architectures, Requirements, and Research Directions // IEEE Access. – 2018. – Vol. 6. – Pp. 47980–48009.
2. *F. Bonomi, R. Milito, J. Zhu, and S. Addepalli*, Fog computing and its role in the Internet of Things // Proc. 1st Ed. MCC Workshop Mobile Cloud Comput. – 2012. – Pp. 13–16.
3. *X. Xie, H.-J. Zeng, and W.-Y. Ma,* Enabling personalization services on the edge // Proc. 10th ACM Int. Conf. Multimedia. – 2002. – Pp. 263–266.
4. *P. P. Gelsinger,* Microprocessors for the new millennium: Challenges, opportunities, and new frontiers // Proc. IEEE Int. Solid-State Circuits Conf., Feb. – 2001. – Pp. 22–25.
5. *S. Ibrahim, H. Jin, B. Cheng, H. Cao, S. Wu, and L. Qi,* CLOUDLET: Towards mapreduce implementation on virtual machines // 18th ACM Int. Symp. High Perform. Distrib. Comput. – 2009. – Pp. 65–66.
6. *N. M. Gonzalez et al.,* Fog computing: Data analytics and cloud distributed processing on the network edges // Proc. 35th Int. Conf. Chilean Comput. Sci. Soc. (SCCC). – 2016. – Pp. 1–9.
7. *H. Dubey, J. Yang, N. Constant, A. M. Amiri, Q. Yang, and K. Makodiya,* Fog data: Enhancing telehealth big data through fog computing // Proc. ASE BigData SocialInform. – 2015. – Pp. 14.
8. *M. Ahmad, M. Bilal, S. Hussain, B. Ho, T. Cheong, and S. Lee,* Health fog: A novel framework for health and wellness applications // J. Supercomput. – 2016. – Vol. 72, no. 10. – Pp. 3677–3695.
9. *B. Tang et al.,* Incorporating intelligence in fog computing for big data analysis in smart cities // IEEE Trans. Ind. Informat. –2017. – Vol. 13, no. 5. – Pp. 2140–2150.
10. *B. Tang, Z. Chen, G. Hefferman, T. Wei, H. He, and Q. Yang,* A hierarchical distributed fog computing architecture for big data analysis in smart cities // Proc. ASE BigData SocialInform. – 2015 – Pp. 28.

11. *W. Zhang, Z. Zhang, and H.-C. Chao,* Cooperative fog computing for dealing with big data in the Internet of vehicles: Architecture and hierarchical resource management // IEEE Commun. Mag. – 2017. – Vol. 55, no. 12. –Pp. 60–67.
12. *B. Yin, W. Shen, Y. Cheng, L. X. Cai, and Q. Li,* Distributed resource sharing in fog-assisted big data streaming // Proc. IEEE Int. Conf. Commun. –2017. – Pp. 1–6.
13. *R. Pecori,* A virtual learning architecture enhanced by fog computing and big data streams, Future Internet. – 2018. – Vol. 10, no. 1. – Pp. 1–30.
14. *N. B. Truong, G. M. Lee, and Y. Ghamri-Doudane,* Software defined networking-based vehicular adhoc network with fog computing // Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. – 2015. –Pp. 1202–1207.
15. *N. K. Giang, V. C. M. Leung, and R. Lea,* On developing smart transportation applications in fog computing paradigm // Proc. 6th ACM Symp. Develop. Anal. Intell. Veh. Netw. – 2016. – Pp. 91–98.
16. *X. Hou, Y. Li, M. Chen, D. Wu, D. Jin, and S. Chen,* Vehicular fog computing: A viewpoint of vehicles as the infrastructures // IEEE Trans. Veh. Technol. – 2016. – Vol. 65, no. 6. – Pp. 3860–3873.
17. *J. K. Zao et al.,* Augmented brain computer interaction based on fog computing and linked data // Proc. Int. Conf. Intell. Environ. (IE). – 2014. – Pp. 374–377.
18. *A. M. Rahmani et al.,* Exploiting smart e-health gateways at the edge of healthcare Internet-of-Things: A fog computing approach // Future Gener. Comput. Syst. – 2018. – Vol. 78. – Pp. 641–658.
19. *A. Giordano, G. Spezzano, and A. Vinci,* Smart agents and fog computing for smart city applications // Proc. Int. Conf. Smart Cities. London, U.K.: Springer. – 2016. – Pp. 137–146.
20. *R. Mahmud, F. L. Koch, and R. Buyya,* Cloud-fog interoperability in IoT-enabled healthcare solutions // Proc. 19th Int. Conf. Distrib. Comput. Netw. (ICDCN). New York, NY, USA: ACM. – 2018. –Pp. 32:1–32:10.

21. *A. V. Dastjerdi, H. Gupta, R. N. Calheiros, S. K. Ghosh, and R. Buyya*, Fog computing: Principles, architectures, and applications // Internet of Things: Principle & Paradigms. San Mateo, CA, USA: Morgan Kaufmann. – 2016.
22. *A. A. Alsaffar, H. P. Pham, C.-S. Hong, E.-N. Huh, and M. Aazam,* An architecture of IoT service delegation and resource allocation based on collaboration between fog and cloud computing // Mobile Inf. Syst. – 2016. –Vol. 2016, Art. no. 6123234.
23. *R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang,* Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption // IEEE Internet Things. – 2016. – Vol. 3, no. 6. –Pp. 1171–1181.
24. *A. Brogi, S. Forti, and A. Ibrahim*, How to best deploy your fog applications, probably // Proc. Int. Conf. Edge Fog Comput. – 2017. – Pp. 105–114.
25. *M. Taneja and A. Davy,* Resource aware placement of IoT application modules in fog-cloud computing paradigm // Proc. IFIP/IEEE Symp. Integr. Netw. Service Manage. (IM). – 2017. – Pp. 1222–1228.
26. *B. Yin, W. Shen, Y. Cheng, L. X. Cai, and Q. Li,* Distributed resource sharing in fog-assisted big data streaming // Proc. IEEE Int. Conf. Commun. – 2017. – Pp. 1–6.
27. *M. Aazam, M. St-Hilaire, C.-H. Lung, I. Lambadaris, and E.-N. Huh,* IoT resource estimation challenges and modeling in fog // Fog Computing in the Internet of Things. Springer. – 2018. – Pp. 17–31.
28. *F. Bonomi, R. Milito, J. Zhu, and S. Addepalli,* Fog computing and its role in the Internet of Things // Proc. 1st Ed. MCC Workshop Mobile Cloud Comput. – 2012. – Pp. 13–16.
29. *Z. Sanaei, S. Abolfazli, A. Gani, and R. Buyya,* Heterogeneity in mobile cloud computing: Taxonomy and open challenges // IEEE Commun. Surveys Tuts. – 2014. – Vol. 16, no. 1. – Pp. 369–392.
30. *P. Bahl, R. Y. Han, L. E. Li, and M. Satyanarayanan,* Advancing the state of mobile cloud computing // Proc. 3rd ACM Workshop Mobile Cloud Comput. Services. – 2012. – Pp. 21–28.

31. *M. Satyanarayanan, G. Lewis, E. Morris, S. Simanta, J. Boleng, and K. Ha,* The role of cloudlets in hostile environments // IEEE Pervas. Comput. – 2013. – Vol. 12, no. 4. – Pp. 40–49.
32. *M. T. Beck, M. Werner, S. Feld, and T. Schimper,* Mobile edge computing: A taxonomy // Proc. 6th Int. Conf. Adv. Future Internet. – 2014. – Pp. 48–54.
33. *G. I. Klas.* Fog Computing and Mobile Edge Cloud Gain Momentum Open Fog Consortium ETSI MEC and Cloudlets. – 2015. http://yucianga.info/?p=938
34. *W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu,* 'Edge computing: Vision and challenges // IEEE Internet Things J. –2016. – Vol. 3, no. 5. – Pp. 637–646.
35. *Y. Wang,* 'Cloud-dew architecture // Int. J. Cloud Comput. – 2015. – Vol. 4, no. 3. – Pp. 199–210.
36. *T. Mane.* (2017). Fog-Dew Architecture for Better Consistency. [Online]. Available: https://eye3i.wordpress.com/2016/07/02/adressinginconsistency-in-dew-computing-using-fog-computing/
37. 3GPP TR 38.889: Study on NR-based access to unlicensed spectrum (Release 16).
38. 3GPP TR 38.716: NR inter-band Carrier Aggregation (CA) / Dual Connectivity (DC) Rel-16 for 3 bands Down Link (DL) / 1 bands Up Link (UL).
39. 3GPP RP-181339: Revision of Study on NR-based Access to Unlicensed Spectrum.
40. 3GPP TR 37.890 Technical Specification Group Radio Access Network; Feasibility Study on 6 GHz for LTE and NR in Licensed and Unlicensed Operations.
41. Recommendation ITU-R M.1652-1, Dynamic frequency selection in wireless access systems including radio local area networks for the purpose of protecting the radio determination service in the 5GHz band, 05/2011.
42. *X. Lu, M. Lema, T. Mahmoodi, and M. Dohler*, Downlink data rate analysis of 5G-U (5G on Unlicensed Band): Coexistence for 3GPP 5Gand IEEE 802. 11ad WiGig // 2017 European Wireless (EW). – 2017. – Pp. 1–6.
43. *X. Lu, E. Sopin, V. Petrov, O. Galinina, D. Moltchanov, K. Ageev, S. Andreev, K. Yevgeni, K. Samouylov, and M. Dohler*, Integrated use of licensed- and unlicensed-

band mmWave radio technology in 5G and beyond // IEEE Access. – 2019. – Vol. 7. – Pp. 24376–24391.

44. *A. I. Sulyman, A. Alwarafy, G. R. MacCartney, T. S. Rappaport, and A. Alsanie*, Directional radio propagation path loss models for millimeter-wave wireless networks in the 28-, 60-, and 73-GHz bands // IEEE Transactions on Wireless Communications. – 2016. – Vol. 15, no. 10. – Pp. 6939–6947.
45. *S. Lagen and L. Giupponi*, Listen before receive for coexis-tence in unlicensed mmWave bands // Wireless Communicationsand Networking Conference (WCNC). IEEE. – 2018. – Pp. 1–6. doi:10.1109/WCNC.2018.8377293.
46. *E. Semaan, J. Ansari, G. Li, E. Tejedor, and H. Wiemann,* An outlookon the unlicensed operation aspects of NR // Wireless Communicationsand Networking Conference (WCNC). IEEE. – 2017. – Pp. 1–6.
47. *M. Rebato, J. Park, P. Popovski, E. De Carvalho, M. Zorzi*, Stochastic Geometric Coverage Analysis in mmWave Cellular Net-works with Realistic Channel and Antenna Radiation Models // IEEE Transactions on Communications. – 2019. – Vol. 67, no. 5. – Pp. 3736–3752. doi:10.1109/TCOMM.2019.2895850
48. *Yi. Pang, A. Babaei, Jennifer Andreoli-Fang, Belal Hamzeh,* Wi-Fi Coexistence with Duty Cycled LTE-U // Wireless Communications and Mobile Computing. – 2017. – Vol. 2017. – Article ID 6486380. – Pp.1–10 doi:10.1155/2017/6486380.
49. *Shaoyi Xu,Yan Li, Yuan Gao, Yang Liu, Haris Gaanin*, Opportunistic Coexistence of LTE and WiFi for Future 5G System: Experimental Performance Evaluation and Analysis // IEEE Access. – 2017. – Vol. 6. – Pp. 8725–8741 doi:10.1109/ACCESS.2017.2787783.
50. *K. Venugopal, M. C. Valenti, and R. W. Heath,* Analysisof Millimeter Wave Networked Wearables in Crowded Environments // Asilomar Conference on Signals, Systems, and Computers. – Nov. 2015. doi:10.1109/ACSSC.2015.7421261.
51. *V. Begishev, D. Moltchanov, E. Sopin, A. Samuylov, S. Andreev, Ye. Koucheryavy, and K. Samouylov,* Quantifying the Impact of Guard Capacity on Session Continuity in 3GPP New Radio Systems // IEEE Transactions on Vehicular

Technology. – 2019. – Vol. 68, no. 12. – Pp. 12345–12359. doi: 10.1109/TVT.2019.2948702.

52. *Qimei Cui, Yu Gu, Wei Ni, and Ren Ping Liu,* Effective Capacity of Licensed-Assisted Access in Unlicensed Spectrum for 5G: From Theory to Application // IEEE Journal on Selected Areas in Communications. – 2017. – Vol. 35, Issue: 8, – Pp. 1754 -1767. doi:10.1109/JSAC.2017.2710023.
53. *N. Bitar, M.O. Al Kalaa, S.J. Seidman, H.H. Refai*, On the Co-existence of LTE-LAA in the Unlicensed Band: Modeling and Performance Analysis // IEEE Access. – 2018. – Vol. 6. – Pp. 52668–52681.
54. *Hengguo Song, Qimei Cui, Yu Gu, Gordon L. Stber, Yong Li, ZesongFei, Chongtao Guo,* Cooperative LBT Design and Effective Capacity Analysis for 5G NR Ultra Dense Networks in Unlicensed Spectrum // IEEE Access. – Vol. 7. – Pp. 50265–50279. DOI: 10.1109/AC-CESS.2019.2910582.
55. *3GPP*, NR; User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone, http://www.3gpp.org/, 3GPP TR 38.101, v.16.2.0, Jan. 2020.
56. *IEEE*, IEEE Standard for Information technology–Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band, http://www.ieee.org/, IEEESTD.2012.6392842, June. 2013.
57. *L. Kleinrock,* Queueing systems, volume 1: Theory. Wiley New York. – 1976. – Vol. 66.
58. *R. Ali, N. Shahin, A. Musaddiq, B.-S. Kim, and S. W. Kim,* Fair and efficient channel observation-based listen-before talk (CoLBT) for LAA-WiFi coexistence in unlicensed LTE // 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN). IEEE. – 2018. – Pp. 154–158.
59. *K. Samouylov, E. Sopin, and O. Vikhrova,* Analyzing blocking probability in LTE wireless network via queuing system with finite amount of resources //

International Conference on Information Technologies and Mathematical Modelling. Springer. – 2015. –Pp. 393–403.

60. *E. Sopin, K. Ageev, E. Markova et al.*, Performance analysis of M2M traffic in LTE network using queuing systems with random resource requirements // Automatic Control and Computer Sciences. – 2018. – Vol. 52. – Pp. 345–353.
61. *V. Begishev, D. Moltchanov, E. Sopin, A. Samuylov, S. Andreev, Y. Koucheryavy, and K. Samouylov,* Quantifying the impact of guard capacity on session continuity in 3gpp new radio systems // IEEE Transactions on Vehicular Technology. – 2019. – Vol. 68, no. 12. – Pp. 12 345– 12 359.
62. *V. Naumov, K. Samuilov, and A. Samuilov,* On the total amount of resources occupied by serviced customers // Automation and Remote Control. – 2016. – Vol. 77. – Pp. 1419–1427.
63. *P. Nain et al.,* Properties of random direction models // Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE. – 2005. – Vol. 3. – Pp. 1897–1907.
64. *M. Gapeyenko et al.,* Analysis of human-body blockage in urban millimeter-wave cellular communications // 2016 IEEE International Conference on Communications (ICC). IEEE. – 2016. – Pp. 1–7.
65. *Sopin, E., Daraseliya, A., Correia*, L.M., Performance Analysis of the Offloading Scheme in a Fog Computing System // 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). IEEE. Moscow. – 2018. – Pp. 1–5.
66. *A.V. Daraseliya, E.S. Sopin*, Optimization of mobile device energy consumption in a fog-based mobile computing offloading mechanism // Discrete and Continuous Models and Applied Computational Science. – 2021. – Vol. 29, № 1 – Pp. 53-62. DOI:10.223 63/2658-4670-2021- 29-1-53-62.
67. *Дараселия А.В., Сопин Э.С.*, Comparative Analysis of the Mechanisms for Energy Efficiency Improving in Cloud Computing Systems // Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer Nature

Switzerland AG 2018 O. Galinina et al. (Eds.): NEW2AN 2018/ruSMART 2018. LNCS. – Vol. 11118. – Pp. 1–9.

68. *Дараселия А.В., Сопин Э.С., Рыков В.В*, On optimization of energy consumption in cloud computing system // Proceedings of the Selected Papers of the 12th International Workshop on Applied Problems in Theory of Probabilities and Mathematical Statistics (Summer Session) in the framework of the Conference on Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems (APTP+MS'2018). – 2018. – Pp. 23–31. http://ceur-ws.org/Vol-2332/paper-03-005.pdf.
69. *A. Daraseliya, M. Korshykov, E. Sopin, D. Moltchanov, Y. Koucheryavy and K. Samouylov*, Handling Overflow Traffic in Millimeter Wave 5G NR Deployments using NR-U Technology // 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. – 2020. – Pp. 1-7. DOI: 10.1109/PIMRC48278.2020.9217313.
70. *Maksym V. Korshykov, Anastasia V. Daraseliya, Eduard S. Sopin*, Development of Analytical Framework for Evaluation of LTE-LAA Probabilistic Metrics // 19th International Conference, NEW2AN 2019, and 12th Conference, ruSMART 2019, Lecture Notes in Computer Science. – Vol. 11660. – Pp. 318–328. DOI: 10.1007/978-3-030-30859-9_27.
71. *A. Daraseliya, M. Korshykov, E. Sopin, D. Moltchanov, S. Andreev and K. Samouylov*, Coexistence Analysis of 5G NR Unlicensed and WiGig in Millimeter-Wave Spectrum // IEEE Transactions on Vehicular Technology. – 2021. – Vol. 70, no. 11. – Pp. 11721–11735. doi: 10.1109/TVT.2021.3113617.
72. *A. V. Daraseliya E. S. Sopin D. A. Moltchanov K. E. Samouylov*, Анализ стратегии разгрузки базовых станций 5G NR с помощью технологии NR-U, Информатика и ее применения. – Т. 15. – Вып. 3. – С. 98-111. DOI: 10.14357/19922264210313.
73. *Daraseliya A.V., Sopin E.S.*, Optimization of task offloading thresholds in the fog computing system // Information technologies and mathematical modelling

(ITMM-2020). Материалы XIX Международной конференции имени А.Ф. Терпугова. Томск. – 2021. – С. 31-36.

74. *A. Daraseliya, M. Korshykovy, E. Sopin,* Оценка показателей эффективности разгрузки базовых станции 5G NR с помощью технологии NR-U // Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems 2021 (ITTMM 2021). –2021. – С. 84-88.
75. *Дараселия А.В., Сопин Э.С.*, Вычисления добавочной скорости передачи данных в нелицензируемом спектре в системе 5G-U // Материалы Всероссийской конференции с международным участием «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем (ИТТММ-2019)». Москва. – 2019. – С. 137–139.
76. *М. В. Коршиков, А.В. Дараселия, Э. С. Сопин*, К обзору спецификаций технологий для сетей связи пятого поколения // Материалы Всероссийской конференции с международным участием «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем (ИТТММ-2019). Москва. – 2019. – С. 152-155.
77. *Дараселия А.В.*, Анализ механизмов повышения энергоэффективности облачных систем // Материалы Всероссийской конференции с международным участием «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем (ИТТММ-2018). Москва: РУДН. – 2018. – С. 118-120.
78. *Tanemura, M.* Statistical distributions of poisson voronoi cells in two and three dimensions. Forma-Tokyo. – 2003. – 18(4). – Pp. 221–247.
79. *Башарин Г.П.* Лекции по математической теории телетрафика // Учеб. пособие. Изд. 3-е, испр. и доп. – М.: Изд-во РУДН. – 2009. – С. 342.
80. *Башарин Г.П., Бочаров П.П., Коган Я.А.* Анализ очередей в вычислительных сетях. Теория и методы расчета. – М.: Главная редакция физико-математической литературы изд-ва «Наука». – 1989. – 336 с.

81. *Башарин Г.П., Толмачев А.Л.* Теория сетей массового обслуживания и ее приложения к анализу информационно-вычислительных систем // Итоги науки и техники. Серия «Теория вероятностей. Математическая статистика. Теоретическая кибернетика». – 1983. – Т. 21. – С. 3–119.
82. *Башарин Г.П., Харкевич А.Д., Шпенс-Шнеппе М.А.* Массовое обслуживание в телефонии. – М.: Наука. – 1968. – 247 с.
83. *Basharin G. P., Samouylov K. E., Yarkina N. V., Gudkova I. A*. A new stage in mathematical teletraffic theory // Automat. Rem. Contr. – 2009. – Vol. 70. No. 12. – P. 1954– 1964.
84. *Bocharov P.P., D'Apice C., Pechinkin A.V., and Salerno S*. Queueing Theory. – Brill Academic Publishers. – 2004. – 457 p.
85. *Вишневский В.М.* Теоретические основы проектирования компьютерных сетей. – М.: Техносфера. – 2003. – 512 с.
86. *Вишневский В.М., Дудин А.Н., Клименок В.И.* Стохастические системы с корреляционными потоками. Теория и применение в телекоммуникационных сетях. М.: Техносфера. – 2018.
87. *Вишневский В.М., Портной С.Л., Шахнович И.В.* Энциклопедия WiMAX. Путь к 4G // М.: Техносфера. – 2009. – С. 472.
88. *Вишневский В.М., Семенова О.В.* Системы поллинга. Теория и применение в широкополосных беспроводных сетях. – М.: Техносфера, – 2007. – 312 с.
89. *Гольдштейн Б.С., Кучерявый А.Е.*, Сети связи пост-NGN // СПб: БХВ-Петербург. – 2013. – С. 160.
90. *Кучерявый А.Е., Парамонов А.И., Кучерявый Е.А.* Сети связи общего пользования. Тенденции развития и методы расчета // М.:ФГУП ЦНИИС. – 2008. – С. 296.
91. *Моисеева С.П., Панкратова Е.В., Убонова Е.Г.* Исследование бесконечнолинейной системы массового обслуживания с разнотипным обслуживанием и входящим потоком марковского восстановления // Вестник Томского государственного университета. Серия «Управление,

вычислительная техника и информатика». – 2016. – № 2. – Вып. 35. – С. 46-53.

92. *Сонькин М.А., Моисеев А.Н., Сонькин Д.М., Буртовая Д.А.* Объектная модель приложения для имитационного моделирования циклических систем массового обслуживания // Вестн. Том. гос. ун-та. УВТиИ. – 2017. – № 40. – С. 71–80.
93. *Moiseev A., Nazarov A. Asymptotic* Analysis of the Infinite-Srever Queueing System with High-Rate Semi-Arrivals // Proc. of the IEEE International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT 2014). – St. Petersburg: IEEE. – 2014. – Pp. 607–613.
94. *Moltchanov D.,* Survey paper: Distance distributions in random networks // Ad Hoc Netw. – 2021. – Vol. 10, no. 6. – Pp. 1146–1166.
95. *Moltchanov D., Samuylov A., Petrov V., Gapeyenko M., Himayat N., Andreev S., and Koucheryavy Y*, Improving session continuity with bandwidth reservation in mmwave communications, // IEEE Wireless Communications Letters. – 2019. – Vol. 8, no. 1. – Pp. 105–108.
96. *Назаров А.А., Терпугов А.Ф.* Теория вероятностей и случайных процессов // Учебное пособие. – Томск: Изд-во НТЛ. – 2006. – С. 204.
97. *Наумов В.А.* Численные методы анализа марковских систем // М.: Изд-во УДН. – 1985. – С. 37.
98. *Naumov V., Samouylov K., Yarkina N., Sopin E., Andreev S., and Samuylov A.* LTE performance analysis using queuing systems with finite resources and random requirements // Proc. of the 7th International Congress on Ultra Modern Telecommunications and Control Systems ICUMT-2015 (October 6–8, 2015, Brno, Czech Republic). – USA, New Jersey, Piscataway, IEEE. – 2015. – P. 100–103.
99. *Naumov V.A., Samouylov K.E.,* On the modeling of queuing systems with multiple resources // PFUR Bulletin, Series Informatics. Mathematics. Physics. – 2014. – Vol. 3. – Pp. 58–62.

100. *Naumov, V.A., Samuilov, K.E., Samuilov, A.K.,* On the total amount of resources occupied by serviced customers // Autom Remote Control. – 2016. – Vol. 77, issue 8. – Pp. 1419–1427.
101. *Корнышев Ю.Н., Пшеничников А.П., Харкевич А.Д.* Теория телетрафика: Учебник для вузов. – М.: Радио и связь. – 1996. – 272 с.
102. *OpenFog Consortium Architecture Working Group,* OpenFog architecture overview, OpenFog Consortium, Tokyo, Japan, White Paper OPFWP001.0216. – Feb. 2016.
103. *Kelly F. P*. Loss networks // Ann. Appl. Probab., – 1991. – No. 1. – P. 319–378.
104. *Kelly F.P*. Reversibility and Stochastic Networks. – New York: J. Wiley & Sons. – 1979. – Pp. 630.
105. *Ross K.W.* Multiservice loss models for broadband telecommunication networks // Springer-Verlag. – 1995. – Pp. 343.
106. *Степанов С.Н.* Основы телетрафика мультисервисных сетей // М.: Изд-во «Эко-Трендз». – 2010. – С. 392.
107. *Степанов С.Н*. Теория телетрафика: концепции, модели, приложения // М.: Горячая линия – Телеком. – 2015. – 868 с
108. *Степанов С.Н., Степанов М.С.* Планирование ресурса передачи при совместном обслуживании мультисервисного трафика реального времени и эластичного трафика данных // Автоматика и Телемеханика. – 2017. – №11. – С. 79–93.
109. *Iversen V.B.* Teletraffic engineering and network planning // ITU-D. – May 2011. – Pp. 567.
110. *А. В. Горбунова, В. А. Наумов, Ю. В. Гайдамака, К. Е. Самуйлов,* Ресурсные системы массового обслуживания как модели беспроводных систем связи // Информ. и её примен. – 2018. – 12:3. – С. 48–55
111. *А. В. Горбунова, В. А. Наумов, Ю. В. Гайдамака, К. Е. Самуйлов,* Ресурсные системы массового обслуживания с произвольным обслуживанием // Информ. и её примен. – 2019. – 13:1. – С. 99–107

112. *Степунин А.Н., Николаев А.Д.,* Мобильная связь на пути к 6G // Москва, Вологда, Инфра-Инженерия. – 2021.

113. *E. Sopin, K. Samouylov, S. Shorgin,* The analysis of the computation offloading scheme with two-parameter offloading criterion in fog computing // Lecture Notes in Computer Science. – 2019. – Pp. 11–20. doi:10.1007/978-3-030-34914-1_2.

114. *E. Sopin, N. Zolotous, K. Ageev, S. Shorgin,* Analysis of the response time characteristics of the fog computing enabled real-time mobile applications // 20th International Conference NEW2AN 2020, Lecture Notes in Computer Science 12525. – 2020. – Pp. 764–779. doi:10.1007/978-3-030-65726-0_9

115. *Р. В. Разумчик, А. И. Зейфман, А. В. Коротышева, Я. А. Сатин*, Анализ энергоэффективности вычислительного комплекса, моделируемого с помощью системы обслуживания с пороговым управлением и интенсивностями, зависящими от времени // Системы и средства информ. – 2015. – 25:4. –С. 19–30

116. *А. И. Зейфман, В. Е. Бенинг, И. А. Соколов,* Марковские цепи и модели с непрерывным временем // М.: Элекс-КМ. – 2008.

117. *Zeifman A.I., Korolev V.Yu, Sipin A.S.(Eds)*, Stability Problems for Stochastic Models. Theory and Applications // MDPI, Basel, Switzerland, ISBN 978-3-0365-0452-0. – 2021. https://doi.org/10.3390/books978-3-0365-0453-7

118. *Kochetkova I, Satin Y, Kovalev I, Makeeva E, Chursin A, Zeifman A.* Convergence Bounds for Limited Processor Sharing Queue with Impatience for Analyzing Non-Stationary File Transfer // Wireless Network. Mathematics. – 2022. – 10(1):30. https://doi.org/10.3390/math10010030.

119. *Y.A. Satin, R.V. Razumchik, A.I. Zeifman, I.A. Kovalev,* Upper bound on the rate of convergence and truncation bound for non-homogeneous birth and death processes on Z // Applied Mathematics and Computation. – 2022. – Vol. 423. – 127009. https://doi.org/10.1016/j.amc.2022.127009.

120. *Gorbunova, A., Vishnevsky, V.,* Estimating the Response Time of a Cloud Computing System with the Help of Neural Networks // Systems Science and Applications. – 2022. – 20(3). – Pp. 105-112.

121. *A. V. Gorbunova, I. S. Zaryadov, S. I. Matyushenko, K. E. Samouylov, S. Ya. Shorgin*, The approximation of response time of a cloud computing system // Inform. Primen. – 2015. – 9:3. – Pp. 32–38.

122. *Tsitovich, I., Titov, I.*, Analysis of loss probability for multimedia resource's traffic // Information Technology and Systems Conference, Moscow. –2012. – Pp. 484–489 (in Russian)

123. *Titov, I., Tsitovich, I., Poryazov, S.,* Use of Time-Scale for Analysis of Data Source Traffic // Modern Probabilistic Methods for Analysis of Telecommunication Networks. BWWQT 2013. Communications in Computer and Information Science. – Springer, Berlin, Heidelberg. – 2013. – Vol. 356. https://doi.org/10.1007/978-3-642-35980-4_21.

124. *L. Wang and E. Gelenbe*, Adaptive Dispatching of Tasks in the Cloud // IEEE Transactions on Cloud Computing. – 2018. – Vol. 6, no. 1. – Pp. 33-45. doi: 10.1109/TCC.2015.2474406.

125. *E. Gelenbe, R. Lent and M. Douratsos,* Choosing a Local or Remote Cloud // 2012 Second Symposium on Network Cloud Computing and Applications. – 2012. – Pp. 25-30. doi: 10.1109/NCCA.2012.16.

126. *J. G. Andrews et al.,* What Will 5G Be? // IEEE Journal on Selected Areas in Communications. – 2014. – Vol. 32, no. 6. – Pp. 1065-1082. doi: 10.1109/JSAC.2014.2328098.

127. *M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse and M. Franceschetti,* Stochastic geometry and random graphs for the analysis and design of wireless networks // IEEE Journal on Selected Areas in Communications. – 2009. – Vol. 27, no. 7. – Pp. 1029-1046. doi: 10.1109/JSAC.2009.090902.

128. *J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta and R. W. Heath*, Modeling and Analyzing Millimeter Wave Cellular Systems // IEEE Transactions on Communications. –2017. – Vol. 65, no. 1. – Pp. 403-430. doi: 10.1109/TCOMM.2016.2618794.

129. *S. Weber, J. G. Andrews and N. Jindal,* An Overview of the Transmission Capacity of Wireless Networks // IEEE Transactions on Communications. – 2010. – Vol. 58, no. 12. – Pp. 3593–3604. doi: 10.1109/TCOMM.2010.093010.090478.
130. *Afif Osseiran, Jose F. Monserrat , Patrick Marsch, Mischa Dohler, Takehiro Nakamura*, 5G Mobile and Wireless Communications Technology // Cambridge University Press; 1st edition. – October 3, 2016.
131. *E. Markova, D. Moltchanov, I. Gudkova, K. Samouylov and Y. Koucharyavy,* Performance Assessment of QoS-Aware LTE Sessions Offloading Onto LAA/WiFi Systems // IEEE Access. – 2019. – Vol. 7. – Pp. 36300–36311. doi: 10.1109/ACCESS.2019.2905035.
132. *Башарин Г.П., Гайдамака Ю.В., Самуйлов К.Е., Яркина Н.В.*, Управление качеством и вероятностные модели функционирования сетей связи следующего поколения: учебное пособие // М. РУДН. –2008. – 157 с.
133. *Naumov V.A., Gaidamaka Y.V., Yarkina N.V., Samouylov K.E.,* Matrix and Analytical Methods for Performance Analysis of Telecommunication Systems // Springer Nature Switzerland AG. – 2021. – 308 c.
134. *A. Daraseliya, E. Sopin, D. Moltchanov, Y. Koucheryavy and K. Samouylov*, "Performance of Offloading Strategies in Collocated Deployments of Millimeter Wave NR-U Technology // IEEE Transactions on Vehicular Technology. – 2022. doi: 10.1109/TVT.2022.3213927.