РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ
ИМЕНИ ПАТРИСА ЛУМУМБЫ

На правах рукописи

Платонова Анна Алексеевна

ПОСТРОЕНИЕ И АНАЛИЗ МОДЕЛИ ДЛЯ РАСЧЕТА ВЕРОЯТНОСТНО-ВРЕМЕННЫХ ХАРАКТЕРИСТИК СЕТИ ИНТЕГРИРОВАННОГО ДОСТУПА И ТРАНЗИТА С РАЗДЕЛЕНИЕМ РЕСУРСОВ

Специальность 1.2.3 – Теоретическая информатика, кибернетика

Диссертация
на соискание ученой степени кандидата
физико-математических наук

Научный руководитель
доктор физико-математических наук
профессор
Гайдамака Юлия Васильевна

Москва – 2026

Оглавление

Введение

Актуальность темы исследования.

Современные и перспективные беспроводные сети, такие как 5G и будущие 6G, сталкиваются с беспрецедентными вызовами, связанными с экспоненциальным ростом числа подключенных устройств, разнообразием сервисов и ужесточением требований к ключевым показателям качества обслуживания (Quality of Service, QoS). В частности, для критически важных приложений, таких как промышленный Интернет вещей, телемедицина и автономный транспорт, необходимы сверхнизкие задержки, высокая надежность и актуальность передаваемой информации. Для удовлетворения этих зачастую противоречивых требований требуются инновационные подходы к управлению сетевыми ресурсами.

В этом контексте технология нарезки сети (Network Slicing) в сетях радиодоступа (Radio Access Network, RAN) пятого поколения (5G) является фундаментальным механизмом, позволяющим эффективно использовать единую физическую инфраструктуру для предоставления специализированных логических сетей (слайсов) с гарантированными показателями производительности. Однако эффективное управление динамическим выделением ресурсов, обеспечение строгой изоляции слайсов и оптимизация производительности в условиях постоянно меняющейся абонентской нагрузки остаются сложными задачами, требующими глубокого теоретического анализа и разработки новых методов.

Параллельно с технологией нарезки сети, для развертывания 5G/6G в миллиметровом (mmWave) и суб-терагерцевом (sub-THz) диапазонах, характеризующихся высоким затуханием сигнала и критической чувствительностью к блокировке прямой видимости между приемо-передающими устройствами, активно внедряются многошаговые беспроводные сети интегрированного доступа и транзита (Integrated Access and Backhaul, IAB). Эти сети используют беспроводные ретрансляторы для создания гибкой транспортной инфраструктуры. Однако многошаговая передача и характерный для IAB-сетей полудуплексный режим передачи данных неизбежно приводят к увеличению

сквозной задержки пакетов и влияют на актуальность передаваемой информации. Таким образом, разработка моделей и методов для оптимизации распределения ресурсов в IAB сетях с учетом этих специфических ограничений, направленных на минимизацию сквозной задержки и возраста информации (Age of Information, AoI), является крайне актуальной научной и практической задачей.

Степень разработанности темы.

Исследования в области нарезки сети 5G активно развиваются, охватывая вопросы оптимального использования ресурсов, справедливого их распределения между слайсами и поддержания надежной изоляции. В источниках [1-4] представлены разнообразные подходы к моделированию задач распределения ресурсов и их оптимизации, основанные на аппарате оптимизации, теории игр и методах машинного обучения. Однако, большинство этих исследований сосредоточено на общих принципах нарезки и не учитывает специфику многошаговых беспроводных транспортных сетей, к которым относится IAB-сеть.

В области IAB-сетей значительный объем исследований посвящен вопросам пропускной способности, надежности, маршрутизации и координации помех, как показано в обзорах [5-12], [13-25]. Многие из этих работ ориентированы на полнодуплексные режимы передачи, что упрощает управление ресурсами, но не отражает реальные ограничения полудуплексных систем, особенно в mmWave и sub-THz диапазонах. Комплексный анализ и оптимизация сквозной задержки и, в особенности, AoI в условиях полудуплексных многошаговых IAB сетей остаются недостаточно изученными. Известные работы [26-29] затрагивают отдельные аспекты задержки и AoI, но не предлагают комплексных моделей и методов управления ресурсами, учитывающих специфические ограничения полудуплексной передачи и динамику трафика в многошаговых IAB-сетях. Таким образом, существует явный пробел в исследованиях, касающихся минимизации задержки и анализа AoI в таких системах.

Для исследования показателей эффективности многошаговых сетей интегрированного доступа и транзита построены математические модели, основанные на теории вероятностей, теории марковских случайных процессов (СП), теории

массового обслуживания, теории телетрафика и стохастической геометрии. К российским и зарубежным ученым, исследователям, внесшим большой вклад в эти области и практическое применение результатов к анализу сетей связи, относятся Башарин Г.П., Бочаров П.П., Горцев А.М., Дудин А.Н., Ефросинин Д.В., Зейфман А.И., Зорин А.В., Ибрагимов Б.Г., Ивницкий В.А., Карташевский В.Г., Ляхов А.И., Клименок В.И., Меликов А.З., Моисеев А.Н., Моисеева С.П., Молчанов Д.А., Морозов Е.В., Назаров А.А., Наумов В.А., Нежельская Л.А., Пауль С.В., Печинкин А.В., Разумчик Р.В., Румянцев А.С., Рыков В.В., Самуйлов К.Е., Сатин Я.А., Семенова О.В., Соколов Н.А., Сопин Э.С., Степанов М.С., Степанов С.Н., Терпугов А.Ф., Тюрликов А.М., Фархадов М.П., Федоткин М.А., Цитович И.И., Циципашвили Г.Ш., Шнепс М.А., L.M. Correia, E. Gelenbe, M. Haenggi, V.B. Iversen, F.P. Kelly, L. Kleinrock, O. Martikainen, K.W. Ross и другие ученые, включая исследователей, занимающихся вопросами анализа качества предоставления услуг в сетях связи – Барабанова Е.А., Вишневский В.М., Вытовтов К.А., Гольдштейн Б.С., Докучаев В.А., Ефимушкин В.А., Киричек Р.В., Крук Е.А., Кулябов Д.С., Кучерявый А.Е., Маколкина М.А., Нетес В.А., Орлов Ю.Н., Парамонов А.И., Пауль С.В., Пшеничников А.П., Росляков А.В., Смелянский Р.Л., Яновский Г.Г., и анализом сетей подвижной связи 5G – Андреев С.Д., Бегишев В.О., Волков А.Н., Кочеткова И.А., Кучерявый Е.А., Молчанов Д.А., Мутханна А.С.А., Хакимов А.А., Хоров Е.М., Яркина Н.В., а также зарубежных исследователей – J.G. Andrews, M. Dohler, J. M. Jornet, A. Goldsmith, H. V. Poor.

Целью диссертационной работы является анализ и расчет показателей качества предоставления услуг в сети интегрированного доступа и транзита с разделением ресурсов с использованием марковских моделей систем массового обслуживания.

Для достижения этой цели в диссертационной работе решаются следующие **задачи**.

– Разработка метода разделения ресурсов беспроводной сети при динамическом выделении ресурса на основе максминной справедливости в условиях приоритизации слайсов.

– Построение математической модели слайса в виде системы массового обслуживания с дисциплиной разделения процессора и эластичным трафиком с ограничением на максимальную скорость передачи, позволяющей провести сравнительный анализ влияния методов вызова процедуры нарезки по показателям эффективности – вероятности деградации обслуживания, коэффициенту использования ресурса и частоты вызова процедуры нарезки.
– Разработка алгоритма расчета оптимальных долей времени активности каналов сети интегрированного доступа и транзита с разделением ресурсов по времени, минимизирующих среднюю сквозную задержку. Вычисление правостороннего квантиля заданного уровня пикового возраста информации на маршруте, а также получение в явном виде функции распределения сквозной задержки с помощью распределения фазового типа.

Научная новизна диссертационной работы**.**

№ 1. Разработан метод разделения ресурсов между слайсами сети при динамическом выделении ресурса на основе максминной справедливости с учетом приоритизации слайсов, который в отличие от известных предусматривает избыточное резервирование ресурсов типа «овербукинг».

№ 2. Предложено понятие деградации обслуживания абонента в случае падения скорости передачи данных при предоставлении услуги абоненту ниже заданного порога, которое позволило получить аналитическое выражение для вероятности деградации обслуживания и сравнить показатели эффективности обслуживания абонентов для нескольких методов вызова процедуры нарезки ресурса.

№ 3. Формализована и решена проблема минимизации средней по сети интегрированного доступа и транзита сквозной задержки и среднего пикового возраста информации, при этом в отличие от известных результатов показано, что случайная величина задержки на маршруте сети имеет функцию распределения фазового типа.

Теоретическая и практическая значимость работы. Полученные в диссертационной работе результаты обладают существенной теоретической и практической значимостью. В теоретическом плане они развивают математический аппарат теории массового обслуживания и теории оптимизации, предлагая новые подходы к анализу управления ресурсами в современных и перспективных беспроводных сетях. Разработанные модели и выведенные аналитические выражения для таких показателей эффективности, как коэффициент использования ресурса, вероятность деградации обслуживания, сквозная задержка и возраст информации, расширяют теоретические основы для глубокого понимания поведения сложных сетевых систем с нарезкой ресурса и многошаговой передачей. Особый вклад вносится в исследование возраста информации за счет его анализа в контексте IAB сетей с полудуплексными ограничениями. Наряду с этим, работа имеет значительную практическую ценность, выражающуюся в разработке конкретных моделей и алгоритмов, применимых для проектирования, планирования и оптимизации сетей 5G/6G. Предложенные стратегии нарезки ресурса и методы оптимизации долей времени активности каналов в IAB сетях позволяют операторам связи эффективно управлять сетевыми ресурсами, гарантировать требуемое качество обслуживания для разнообразных услуг, минимизировать сквозные задержки и обеспечивать требования к свежести информации для критически важных приложений. Это способствует не только повышению производительности сети, но и снижению вычислительной сложности управления, поскольку алгоритмы разделения ресурсов Глав 1 и 3 могут быть интегрированы в систему управления радиоресурсами и в менеджер нарезки ресурса в реальных сетях.

Методы исследования. В диссертации применяются методы теории массового обслуживания, стохастического моделирования (пуассоновские процессы), теории оптимизации (линейное и нелинейное программирование, метод проекции градиента), имитационное моделирование с использованием дискретно-событийного симулятора OMNeT++ с расширенной библиотекой *queueinglib*, а также методы математического анализа и теории вероятностей.

Положения, выносимые на защиту.

№ 1. Разработанная методика разделения ресурсов между слайсами сети при динамическом выделении ресурса на основе максминной справедливости с учетом приоритизации слайсов позволяет получить оптимальные доли ресурса, выделяемые каждому слайсу, как решение проблемы оптимизации, учитывающей максимальное использование ресурса и ограничения на диапазон скоростей передачи данных при предоставлении услуги.

№ 2. Динамическая нарезка ресурса при обнаружении деградации обслуживания, т.е. падения скорости передачи данных при предоставлении услуги абоненту ниже заданного порога, обеспечивает более высокие показатели эффективности обслуживания абонентов по сравнению с обслуживанием абонентов с статической нарезкой и другими методами вызова процедуры нарезки ресурса – при поступлении нового запроса, по окончании получения услуги абонентом в сети, при срабатывании регулярного таймера нарезки.

№ 3. Разработанный алгоритм расчета оптимальных долей времени активности каналов сети интегрированного доступа и транзита с разделением ресурсов по времени позволяет минимизировать среднюю по сети сквозную задержку и средний по сети пиковый возраст информации, а также вычислять правосторонний квантиль заданного уровня для функций распределения указанных случайных величин на маршруте от донора к абонентскому устройству в виде распределения фазового типа.

Степень достоверности и апробация результатов обеспечивается использованием строгого математического аппарата теории массового обслуживания и теории оптимизации, корректностью применяемых допущений, а также верификацией аналитических моделей с помощью дискретно-событийного имитационного моделирования в среде OMNET++.

Основные результаты диссертации были апробированы на международных и всероссийских научных конференциях:

– международная конференция «International Conference on Next Generation Wired/Wireless Advanced Networks and Systems» (г. Абу-Даби, Объединенные Арабские Эмираты, 2025);

– всероссийская конференция с международным участием «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем» (г. Москва, РУДН, 2022, 2023, 2025);

– международная конференция «Distributed computer and communication networks: control, computation, communications» (г. Москва, РУДН, 2022).

Основные результаты опубликованы в рецензируемых научных журналах Future Internet (Scopus Q2), Информатика и ее применения (Scopus Q3), Системы и средства информатики (список ВАК), а также в трудах международных конференций, индексируемых в Scopus – Lecture Notes in Computer Science (Scopus Q2), Communications in Computer and Information (Scopus Q3).

Реализация результатов работы. Результаты диссертационной работы включены в исследования по грантам Российского научного фонда (РНФ) № 22-79-10053 «Разработка моделей и алгоритмов обслуживания критичного к задержке и надежности доставки трафика в сценариях промышленной автоматизации на основе беспроводных систем 5G», № 24-19-00804 «Исследование возраста информации в задачах обеспечения качества предоставления услуг URLLC и mMTC в беспроводных сетях 5G», № 23-79-10084 «Математические модели и практические алгоритмы повышения энергоэффективности в гетерогенных миллиметровых и терагерцовых сетях пятого и шестого поколения (5G/6G)».

Публикации. Основные результаты диссертации изложены в 11 работах [30-40], в том числе в 5 изданиях, входящих в базу данных Scopus [30-34], в 1 издании из списка журналов, рекомендованного Высшей аттестационной комиссией при Минобрнауки России [35], в 3 свидетельствах о государственной регистрации программ для электронно-вычислительной машины (ЭВМ) [36-38].

Соответствие паспорту специальности. Диссертационное исследование выполнено в соответствии с паспортом специальности 1.2.3. «Теоретическая

информатика, кибернетика» и соответствует **п. 9** Математическая теория исследования операций (физико-математические науки), а именно формализация исследуемой системы как объекта исследования операций в части описания сети IAB в виде совокупности взаимодействующих элементов (абонентские устройства, базовые станции, радиоканалы) и в части описания протекающих в сети IAB процессов передачи пакетов с учетом ограничений полудуплексного и полнодуплексного режимов передачи данных в радиоканале (раздел 3.2), в части описания процедуры нарезки ресурса сети радиодоступа в виде алгоритма с входными параметрами, определяющими объем ресурса системы, ограничения на объемы ресурса слайсов, приоритизацию слайсов, и критерий справедливости при распределении ресурса. Методика разделения ресурсов между слайсами основана на решении задачи оптимизации, учитывающей динамическое выделение ресурсов и максминную справедливость (раздел 1.2). Сравнительный анализ альтернатив, как ключевой элемент исследования операций, реализован при выборе метода вызова процедуры нарезки ресурса (раздел 2.2). Алгоритм расчета долей времени активности радиоканалов предназначен для минимизации средней сквозной задержки и среднего пикового возраста информации на всех маршрутах от IAB-донора до абонентского устройства (раздел 3.3). Критериями при анализе выступают показатели эффективности, которые определены в разделе 2.1.

Личный вклад. Автором лично разработана методика разделения ресурсов между слайсами, учитывающая динамическое выделение ресурсов; самостоятельно реализован алгоритм численного решения оптимизационной задачи распределения ресурсов; построена математическая модель слайса услуги «Best Effort» в виде СМО с дисциплиной Processor Sharing; самостоятельно выведены аналитические выражения для ключевых показателей эффективности предоставления услуг в сети IAB, проведен численный анализ влияния различных методов вызова процедуры нарезки ресурса на эти показатели; построена модель сети IAB в виде открытой экспоненциальной сети массового обслуживания, учитывающая разделение ресурса по времени в полудуплексном и полнодуплексном режимах передачи данных в радиоканале и древовидную

топологию; самостоятельно получены аналитические выражения для функций распределения сквозной задержки пакета и пикового возраста информации с помощью распределения фазового типа; решена задача оптимизации долей времени активности каналов для минимизации средней сквозной задержки и среднего пикового возраста информации.

Объем и структура работы. Структура диссертационной работы построена из введения, трех глав, заключения и списка литературы из 85 источников. Диссертационная работа изложена на 106 страницах текста, содержит 23 рисунка и 4 таблицы.

Краткое изложение диссертации. Результаты диссертационной работы изложены в трех главах. Во **введении** обоснована актуальность темы диссертационной работы, определены цели и задачи исследования, сформулированы научная новизна и практическая ценность работы.

В **первой главе** представлены предварительные исследования технологии нарезки ресурса в сетях радиодоступа 5G, для которой разработана методика разделения ресурса между слайсами для систем с нарезкой ресурсов без приоритизации слайсов и с приоритизацией слайсов, а также сформулированы задачи исследования. Раздел 1.1 посвящен общим принципам технологии нарезки ресурса, включая изоляцию слайсов и динамическое масштабирование ресурсов. В разделе 1.2 описана системная модель и архитектура нарезки сети радиодоступа 5G, а также математическая модель распределения ресурсов, учитывающая динамику абонентской нагрузки и ограничения на пропускную способность. Раздел 1.3 формулирует постановку задачи исследования, акцентируя внимание на вызовах, связанных с IAB-сетями и метриками сквозной задержки и возраста информации.

Основными результатами Главы 1 являются Алгоритм 1.1 численного решения задачи разделения ресурса между S слайсами, учитывающей динамическое выделение ресурсов и максминную справедливость, а также

Теоремы 1.1 и 1.2 с методикой разделения ресурса между S слайсами для системы без приоритизации слайсов и с приоритизацией слайсов.

Алгоритм 1.1. Численное решение $\mathbf{R}^* = \underset{\mathbf{R}}{\operatorname{argmax}}\, U(\mathbf{R})$ оптимизационной задачи разделения ресурса C между слайсами из множества $\mathcal{S}$, $S = |\mathcal{S}|$, в системе с нарезкой ресурса в ограничениях на нижнее $\mathbf{R}^{\min} = \left(R_s^{\min}\right)_{s\in\mathcal{S}}$ и верхнее $\mathbf{R}^{\max}$ пороговые значения скорости передачи данных в каждом слайсе, а также на минимальное число $\mathbf{N}^{\text{cont}}$ абонентов каждого слайса, которым услуга должна быть предоставлена на скорости не ниже $\mathbf{R}^{\min}$, для числа абонентов в каждом слайсе, заданного вектором $\mathbf{N}$,

$$U(\mathbf{R}) = \sum_{s\in\mathcal{S}} W_s\,(N_s)N_s * \ln(R_s) \;\rightarrow\; \max,$$

$$\mathbf{N}^T\mathbf{R} \;=\; C,$$

$$\mathbf{R} \in \mathbb{R}_+^S : R_s^{\min} \le R_s \le R_s^{\max}, s \in \mathcal{S},$$

может быть вычислено с помощью следующего алгоритма.

Входные параметры: $C, S, \mathbf{N}, \mathbf{R}^{\min}, \mathbf{R}^{\max}, \mathbf{N}^{\text{cont}}$

Выходной параметр: $\mathbf{R}$

1. инициализация
2. $\mathbf{W}^T \coloneqq [W_1(N_1),\ \ldots,\ W_S(N_S)]$ // вектор весовых функций
3. $\mathbf{X}^{\text{stat}} \coloneqq \mathbf{W}C(\mathbf{W}N)^{-1}$ // стационарная точка
4. **if** $R_i^{\min} \le X_i^{\text{stat}} \le R_i^{\max},\ i = 1,\ldots,S$ **then**
5. **return** $\mathbf{X}^{\text{stat}}$
6. $M_{[1\times S]} \coloneqq \mathbf{N}^T$
7. $P_{[S\times S]} \coloneqq \mathbf{I} - \mathbf{N}^T \cdot (\mathbf{N}\mathbf{N}^T)^{-1} \cdot \mathbf{N}$
8. $\mathbf{X}^0 \coloneqq \mathbf{R}^{\min} + \left(C - \mathbf{N}^T\mathbf{R}^{\min}\right)\left(\mathbf{N}\left(\mathbf{R}^{\max} - \mathbf{R}^{\min}\right)^T\right)^{-1}\left(\mathbf{R}^{\max} - \mathbf{R}^{\min}\right)$
9. $\tau \coloneqq \|\mathbf{X}^0 - \mathbf{X}^{\text{stat}}\|,\ \delta \coloneqq 1$
10. **while** $\delta > 0.0001$ **do**
11. $\quad \mathbf{X}^1 \coloneqq \mathbf{X}^0 + \tau\mathbf{P}\, div\left(\mathbf{N}^T\mathbf{W}, \mathbf{X}^0\right)$
12. $\quad t_{bound} \coloneqq 2,\ t_{coord} \coloneqq -1,\ \delta_+ \coloneqq 0$
13. $\quad$ **for** $i = \overline{1,S}$ **do**
14. $\quad\quad$ **if** $N_i > 0$ **then**
15. $\quad\quad\quad$ **if** $X_i^1 < R_i^{\min}$ **then**
16. $\quad\quad\quad\quad$ **if** $t_{bound} > \left(R_i^{\min} - X_i^0\right)\left(X_i^1 - X_i^0\right)^{-1}$ **then**

17. $t_{bound} \coloneqq \left(R_i^{\min} - X_i^0\right)\left(X_i^1 - X_i^0\right)^{-1},\ t_{coord} \coloneqq i$
18. **if** $X_i^1 > R_i^{\max}$ **then**
19. **if** $t_{bound} > \left(R_i^{\max} - X_i^0\right)\left(X_i^1 - X_i^0\right)^{-1}$ **then**
20. $t_{bound} \coloneqq \left(R_i^{\max} - X_i^0\right)\left(X_i^1 - X_i^0\right)^{-1},\ t_{coord} \coloneqq i$
21. **if** $i_{bound} < 2$ **then**
22. $\mathbf{X}^1 \coloneqq \mathbf{X}^0 + t_{bound}(\mathbf{X}^1 - \mathbf{X}^0)$
23. **if** *Число строк матрицы* $\mathbf{M} < S - 1$ **then**
24. Добавить пустую строку к $\mathbf{M}$
25. $\delta_+ \coloneqq 1$
26. Последняя строка матрицы $\mathbf{M} \coloneqq \mathbf{I}[t_{coord}]$
27. **if** $\left\|\mathbf{M}\mathbf{M}^T\right\| > 0.0000001$ **then**
28. $\mathbf{P} \coloneqq \mathbf{I} - \mathbf{M}^T\left(\mathbf{M}\mathbf{M}^T\right)^{-1}\mathbf{M}$
29. $\delta \coloneqq \delta_+ + \|\mathbf{X}^0 - \mathbf{X}^1\|\ ; \mathbf{X}^0 \coloneqq \mathbf{X}^1$
30. **return** $\mathbf{X}^0$

Теорема 1.1. Пусть система с нарезкой ресурса между слайсами из множества $\mathcal{S}$, $S = |\mathcal{S}|$, задана следующими параметрами:

C – суммарный ресурс системы (емкость БС);

$\mathbf{R}^{\min} = \left(R_s^{\min}\right)_{s\in\mathcal{S}}$ – вектор минимальных скоростей передачи данных для абонента слайса s, определенных в SLA между владельцем сети и арендатором слайса s;

$\mathbf{N}^{\min} = \left(N_s^{\min}\right)_{s\in\mathcal{S}}$ – вектор числа абонентов слайса s, которым услуга может быть предоставлена на максимальной скорости;

$\mathbf{N} = (N_s)_{s\in\mathcal{S}}$ – вектор числа абонентов слайса s в состоянии, когда ресурсов недостаточно для обеспечения всем абонентам минимальных скоростей передачи данных;

$\mathbf{C}(\mathbf{N}) = \left(C_s(\mathbf{N})\right)_{s\in\mathcal{S}}$ – вектор емкостей слайсов.

В системе с нарезкой радиоресурсов без приоритизации слайсов в состоянии $\mathbf{N}_{[1\times S]} = (N_1, N_2, \dots, N_S)$, когда ресурсов недостаточно для обеспечения всем абонентам минимальных скоростей передачи данных, емкость слайса $C_s(\mathbf{N})$ определяется следующим образом:

$$C_s(\mathbf{N}) = \begin{cases} \dfrac{N_s^{\min} R_s^{\min}}{\mathbf{N}^{\min}\mathbf{R}^{\min}} C, & \mathbf{N}^{\min}\mathbf{R}^{\min} \geq C; \\ N_s^{\min} R_s^{\min} + \dfrac{(N_s - N_s^{\min}) R_s^{\min}}{(\mathbf{N} - \mathbf{N}^{\min})\mathbf{R}^{\min}} \left(C - \mathbf{N}^{\min}\mathbf{R}^{\min}\right), & \mathbf{N}^{\min}\mathbf{R}^{\min} < C. \end{cases}$$

Теорема 1.2. Пусть система с нарезкой ресурса между слайсами задана параметрами Теоремы 1.1. В системе с нарезкой радиоресурсов, с приоритизацией слайсов в ситуации дефицита ресурсов в состоянии $\mathbf{N}_{[1 \times S]} = (N_1, N_2, \dots, N_S)$, когда суммарная минимальная требуемая емкость превышает доступную пропускную способность БС, емкость C распределяется в соответствии с установленными приоритетами следующим образом.

1. Если $c_1(\mathbf{N}) \geq C$ (недостаточно ресурсов даже для наивысшего приоритета), емкость распределяется только среди слайсов наивысшего приоритета $\mathcal{S}_1$:

$$C_s(\mathbf{N}) = \begin{cases} \dfrac{N_s^{\min} R_s^{\min}}{c_1(\mathbf{N})} C, & s \in \mathcal{S}_1, \\ 0, & s \in \mathcal{S} \setminus \mathcal{S}_1, \end{cases}$$

2. Если $c_{u^*}(\mathbf{N}) < C$ и $c_{u^*+1}(\mathbf{N}) \geq C$ (дефицит возникает на уровне $u^* + 1$):

$$C_s(\mathbf{N}) = \begin{cases} N_s^{\min} R_s^{\min}, & s \in \mathcal{S}_i : 1 \leq i \leq u^*; \\ \dfrac{N_s^{\min} R_s^{\min}}{c_{u^*+1}(\boldsymbol{N}) - c_{u^*}(\boldsymbol{N})} \left(C - c_{u^*}(\boldsymbol{N})\right), & s \in \mathcal{S}_{u^*+1}; \\ 0, & s \in \mathcal{S}_i : i > u^* + 1. \end{cases}$$

Вторая глава посвящена анализу показателей качества предоставления услуги категории BE с помощью построенной в разделе 2.1 математической модели слайса услуги BE в виде СМО с дисциплиной PS и эластичным трафиком с ограничением на максимальную скорость передачи. Исследован новый показатель качества обслуживания – деградация обслуживания абонента, которая наступает при пересечении доступной абоненту скоростью обслуживания некоторого порога, т.о. для абонентов, получающих услугу на скорости ниже пороговой, услуга предоставляется с ненадлежащим качеством. В разделе 2.2 приведены численные результаты моделирования и их анализ, демонстрирующий влияние различных методов вызова процедуры нарезки на вероятность деградации (снижения

качества) обслуживания, коэффициент использования ресурса и частоту вызова процедуры нарезки ресурса.

Основным результатом Главы 2 является Утверждение 2.6. для расчета вероятности деградации обслуживания абонента в слайсе.

Утверждение 2.6. Вероятность $P^{\deg}$ деградации обслуживания абонента в слайсе услуги BE с эластичным трафиком и ограничением на максимальную скорость передачи может быть вычислена по формуле

$$P^{\deg} = 1 - \sum_{N=0}^{N^{\text{cont}}} p_N,$$

где стационарное распределение $\{p_N, N \geq 0\}$ вероятностей состояний марковского процесса (МП) числа заявок в СМО $\begin{matrix} M \\ \lambda \end{matrix} \left| \begin{matrix} M \\ \theta \end{matrix} \right| C \Big| 0,\ R^{\max}$ с дисциплиной разделения процессора PS над пространством состояний $\mathcal{X} = \{0, 1, \dots, \infty\}$ имеет вид

$$p_N = \begin{cases} \dfrac{1}{N!}\left(\dfrac{\lambda\theta}{R^{\max}}\right)^N p_0, & 1 \leq N \leq M; \\ \left(\dfrac{\lambda\theta}{C}\right)^{n-M} p_M = \dfrac{1}{M!}\dfrac{(\lambda\theta)^N}{C^{N-M}(R^{\max})^M}\, p_0, & n \geq M, \end{cases}$$

$$p_0 = \left(1 + \sum_{N=1}^{M-1} \frac{1}{N!}\left(\frac{\lambda\theta}{R^{\max}}\right)^n + \frac{1}{M!}\left(\frac{\lambda\theta}{R^{\max}}\right)^M \sum_{N=M}^{\infty} \left(\frac{\lambda\theta}{C}\right)^{N-M}\right)^{-1},$$

а условия существования стационарного режима определяется неравенством

$$\lambda < \frac{C}{\theta}.$$

Третья глава посвящена построению модели передачи трафика в сети IAB с учетом особенностей организации многошаговой беспроводной сети, включая полудуплексный и полнодуплексный режимы передачи данных в радиоканале на основе технологии нарезки ресурса по времени. В разделе 3.1 выполнен литературно-аналитический обзор задач анализа качества обслуживания в сети IAB. В комплексную задачу оптимизации сети IAB включен новый показатель качества обслуживания – возраст информации AoI, который отражает актуальность информации от удаленной подсистемы на центральной системе управления и

мониторинга. Возраст информации равен интервалу времени, прошедшему с момента генерации на удаленной подсистеме последней информации, полученной от нее центральной системой. В разделе 3.2 с помощью построенной в виде CeMO модели решена задача разделения ресурсов сети IAB, которая сформулирована в виде проблемы минимизации средней по сети сквозной задержки или среднего пикового возраста информации. Решением проблемы минимизации стали доли времени активности каналов доступа / транзита, которые являются параметрами алгоритма разделения временного ресурса менеджером нарезки ресурса. Раздел 3.3 содержит анализ метрик производительности, ограничения и оптимизацию в сети IAB, формулировку задач минимизации средней сквозной задержки и среднего пикового возраста информации, а также результаты их решения.

Основным результатом Главы 3 являются Алгоритм 3.1 расчета оптимальных долей времени активности каналов сети IAB, минимизирующих среднюю по сети сквозную задержку, и вычисления правостороннего квантиля заданного уровня пикового возраста информации на p-пути, а также Теорема 3.1, определяющая в явном виде функцию распределения сквозной задержки на p-пути в сети IAB.

Алгоритм 3.1. Пусть древовидная сеть IAB задана следующими параметрами:

множество $\mathcal{B}$ базовых станций (донор и ретрансляторы), для каждой из которых задано число N_n секторов антенны для БС n, $n \in \mathcal{B}$;

множество $\mathcal{E}$ каналов доступа / транзита;

множество $\mathcal{P}$ маршрутов от донора к АУ в секторе (n, i), $n \in \mathcal{B}$, $i = 1,2, \dots, N_n$;

подмножества $\mathcal{E}_p$ составляющих маршрут p каналов, $\mathcal{E}_p \subseteq \mathcal{E}$, $p \in \mathcal{P}$;

матрица конфликтов $\mathbf{F} = (f_{ne})_{n\in\mathcal{B},e\in\mathcal{E}}$, где в строке, соответствующей БС n, элементы $f_{ne_1} = f_{ne_2} = \cdots = 1$, если каналы e_1, e_2, … не могут быть активны одновременно;

интенсивности Λ_p потоков пакетов для АУ в секторе (n, i) конечного узла n маршрута p, $n \in \mathcal{B}$, $i = 1,2, \dots, N_n$, $p \in \mathcal{P}$;

емкость C радиоканала;

емкости $C_e = C$ каналов доступа / транзита, $e \in \mathcal{E}$;

доли q_e времени активности каналов доступа / транзита, $e \in \mathcal{E}$.

Тогда вектор $\mathbf{q}^* = \left(q_1^*, q_1^*, \ \dots\ , \ q_{|\mathcal{E}|}^*\right)$ оптимальных долей времени активности каналов, минимизирующий значение средней по сети сквозной задержки в нисходящем направлении от донора к АУ, и правосторонний квантиль $\delta_{A_p}^+(\alpha)$ заданного уровня α пикового возраста информации для маршрута p в сети IAB, $p \in \mathcal{P}$, могут быть вычислены по следующему алгоритму.

Входные данные: $\mathcal{B}$; $N_n, n \in \mathcal{B}$; $\mathcal{E}$; $\mathcal{P}$; $\mathcal{E}_p \subseteq \mathcal{E}, p \in \mathcal{P}$; $\mathbf{F}$; $\boldsymbol{\Lambda_p} = \left(\Lambda_p\right)_{p\in\mathcal{P}}$; C.

Выходные параметры: $\mathbf{q}^* = (q_e^*)_{e\in\mathcal{E}}$; $\delta_{A_p}^+(\alpha)\ , p \in \mathcal{P}$.

ШАГ 1. Вычисление компонент вектора интенсивностей $\boldsymbol{\lambda_e} = (\lambda_e)_{e\in\mathcal{E}_p}$ потоков пакетов на e-канал по формуле

$$\lambda_e = \textstyle\sum_{p:e\in\mathcal{E}_p} \Lambda_p,\ e \in \mathcal{E}.$$

ШАГ 2. Вычисление компонент вектора $\mathbf{q}^* = \left(q_1^*, q_1^*, \ \dots\ , \ q_{|\mathcal{E}|}^*\right)$ оптимальных долей времени активности каналов как решение методом множителей Лагранжа проблемы оптимизации $U(\mathbf{q}) \to \min\limits_{\mathbf{q}}$ с целевой функцией

$$U(\mathbf{q}) = \sum_{p\in\mathcal{P}} \left(\frac{\Lambda_p}{\sum_{p\in\mathcal{P}} \Lambda_p} \cdot d_p\right) = \frac{1}{\sum_{p\in\mathcal{P}} \Lambda_p} \sum_{p\in\mathcal{P}} \left(\Lambda_p \cdot \sum_{e\in\mathcal{E}_p} \frac{1}{C_e q_e - \lambda_e}\right)$$

в ограничениях

$$R_1: 0 \le q_e \le 1, \forall e \in \mathcal{E};$$

$$R_2: \lambda_e \le C_e q_e, \forall e \in \mathcal{E};$$

$$R_3: \mathbf{F}\mathbf{q}^T \le \mathbf{1};$$

для существования стационарного режима функционирования сети IAB с матрицей конфликтов $\mathbf{F} = (f_{ne})_{n\in\mathcal{B},e\in\mathcal{E}}$, отражающей полудуплексный и полнодуплексный режимы передачи данных в каналах доступа / транзита.

ШАГ 3. Получение правостороннего квантиля $\delta_{A_p}^+(\alpha)$ заданного уровня α пикового возраста информации для маршрута p в сети IAB как $\delta_{A_p}^+(\alpha) = \delta_{1-\alpha}$, где $\delta_{1-\alpha}$ – решение уравнения

$$1 - \boldsymbol{\beta}^T e^{\mathbf{N}_p \delta_{1-\alpha}}\, \mathbf{1} = 1 - \alpha,$$

в котором $\boldsymbol{\beta}^T = (1,0,\dots,0)$ – вектор-строка размерности $|\mathcal{E}_p| + 1$, $\beta_1 = 1$ соответствует ближнему к донору каналу,

матрица $\mathbf{N}_p$ – диагональная матрица размерности $(|\mathcal{E}_p| + 1) \times (|\mathcal{E}_p| + 1)$

$$\mathbf{N}_p = \begin{pmatrix} -\Lambda_p & 0 & 0 & \dots & 0 \\ 0 & -(C_1 q_1 - \lambda_1) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\left(C_{|\mathcal{E}_p|} q_{|\mathcal{E}_p|} - \lambda_{|\mathcal{E}_p|}\right) \end{pmatrix}, p \in \mathcal{P}.$$

Теорема 3.1. Для древовидной сети IAB с множеством $\mathcal{E}$ каналов доступа / транзита; интенсивностями λ_e потоков пакетов на e-канал; емкостями $C_e q_e$ каналов доступа / транзита с учетом долей q_e времен активности каналов, $e \in \mathcal{E}$; емкостью C общего ресурса радиоканала, разделяемого между e-каналами на основе временного разделения ресурса, случайная величина D_p сквозной задержки на p-пути, начинающемся с e^*-канала, $e^* \in \mathcal{E}_p \subseteq \mathcal{E}$, имеет PH-распределение с функцией распределения вида

$$F_{D_p}(x) = 1 - \boldsymbol{\beta}^T e^{\mathbf{M}_p x} \mathbf{1},\ p \in \mathcal{P},$$

где $\boldsymbol{\beta}^T = (\beta_e)_{e \in \mathcal{E}_p}$ – вектор-строка размерности $|\mathcal{E}_p|$ с компонентами

$$\beta_e = \begin{cases} 1, & e = e^*; \\ 0, & e^* \in \mathcal{E}_p \backslash \{e^*\}, \end{cases}$$

матрица $\mathbf{M}_p$ – диагональная матрица размерности $|\mathcal{E}_p| \times |\mathcal{E}_p|$ с ненулевыми элементами вида $(\lambda_e - C_e q_e)$, $e \in \mathcal{E}_p$, $p \in \mathcal{P}$.

В **заключении** представлены основные результаты диссертационной работы.

ГЛАВА 1
ИССЛЕДОВАНИЕ МЕХАНИЗМОВ НАРЕЗКИ РЕСУРСОВ В СЕТЯХ РАДИОДОСТУПА 5G

1.1. Технология нарезки сети в 5G

Нарезка сети (Network slicing) представляет собой одну из фундаментальных технологий беспроводных сетей пятого поколения (5G), предназначенную для эффективного предоставления широкого спектра телекоммуникационных услуг с существенно различными требованиями к QoS на базе единой физической инфраструктуры. Помимо этого, данная технология способна поддерживать разнообразные бизнес-модели, включая операторов виртуальных мобильных сетей (Mobile Virtual Network Operator, MVNO), обеспечивая им необходимую степень автономии в управлении услугами.

Сетевой слайс определяется как логическая сеть, которая предоставляет специфический набор функциональных возможностей и характеристик производительности, адаптированных под конкретные сервисы или группы абонентов. Ключевыми принципами функционирования нарезки сети являются:

1. Изоляция слайсов: Обеспечение строгой изоляции между различными сетевыми слайсами для минимизации взаимного влияния и гарантирования предсказуемой производительности каждого слайса, т.о. колебания трафика в одном слайсе сети не должны отрицательно влиять на QoS других слайсов.

2. Динамическое масштабирование ресурсов: Возможность оперативного выделения и высвобождения сетевых ресурсов с минимальным воздействием на качество обслуживания. Этот принцип может включать механизмы прерывания или вытеснения ресурсов в пользу слайсов с более высоким приоритетом при возникновении конкуренции за ограниченные ресурсы.

Сетевые слайсы могут охватывать различные сегменты сетевой инфраструктуры, начиная от АУ и RAN до опорной сети. Ресурсы, выделяемые слайсам, могут быть как специализированными (выделенными исключительно для

данного слайса), так и совместно используемыми (например, вычислительная мощность, оперативная память, полоса пропускания).

Центральной проблемой, возникающей при реализации технологии нарезки сети, является эффективное управление и распределение ресурсов. Это включает в себя обеспечение оптимального использования доступных ресурсов, их справедливого и недискриминационного распределения между различными сетевыми слайсами и их абонентами, а также поддержание гибкой, но надежной изоляции слайсов.

Современные исследования в данной области охватывают концепцию сквозной (end-to-end) нарезки сети, зачастую базирующуюся на представлении сетевой топологии в виде графовых моделей, а также задачи распределения ресурсов в RAN [41-44]. При этом рассматриваются механизмы совместного управления гетерогенными ресурсами, включая пропускную способность каналов связи, объем кэш-памяти и вычислительные мощности в архитектуре Cloud-RAN. Исследование структурировано по сценариям обслуживания 5G, каждому из которых соответствует свой тип слайса: расширенная мобильная широкополосная связь (Enhanced mobile broadband, eMBB), сверхнадёжная связь с малой задержкой (Ultra Reliable Low Latency Communications, URLLC) и массовая межмашинная связь/ мобильный Интернет вещей (Massive Machine Type Communications/ Mobile Internet of Things, mMTC/MIoT) [45, 46]. Кроме того, исследуются обобщенные сценарии, характеризующиеся наличием эластичного или неэластичного трафика, а также жесткими требованиями соглашений об уровне обслуживания (Service Level Agreement, SLA).

Целевые установки стратегий нарезки варьируются в зависимости от постановки задачи и могут включать максимизацию суммарной скорости передачи данных или функций полезности, минимизацию удельных ресурсных затрат, а также обеспечение строгой изоляции слайсов с точки зрения показателей производительности и соблюдения детерминированных требований к QoS и величине задержек. Для нахождения оптимальных решений применяется широкий спектр математических методов, в частности: методы теории оптимизации

(решение линейных и нелинейных задач), аппарат теории игр, алгоритмы машинного обучения, а также стохастическое моделирование на основе марковских цепей с непрерывным временем.

В научно-технической литературе выделяют два фундаментальных подхода к управлению ресурсами:

1. Прямое распределение: выделение ресурсов непосредственно оконечным абонентам с последующей агрегацией показателей в рамках конкретного слайса.

2. Иерархическое распределение: двухуровневая схема, при которой на первом этапе системный ресурс распределяется между слайсами, а на втором – внутри каждого слайса между его абонентами.

В диссертационной работе применяется второй подход. Процедура нарезки ресурсов может осуществляться в различных временных масштабах: от оперативного управления на уровне интервалов передачи (Transmission Time Interval, TTI) до долгосрочного планирования на основе прогнозных моделей динамики трафика.

Рассмотренные технологические принципы нарезки сети определяют архитектурные границы реализации сервисов, однако для оценки их эффективности требуется переход от качественного описания к формализованным математическим моделям. В связи с этим, в разделе 1.2 будет представлена системная модель и архитектура нарезки RAN, позволяющая количественно описать процессы взаимодействия менеджера нарезки ресурса (Slicing Manager, SliM) и конечных абонентов.

1.2. Моделирование процесса нарезки ресурсов

Системная модель и архитектура нарезки RAN

Для целей настоящего исследования разработана системная модель, описывающая функционирование RAN пятого поколения (5G) в условиях применения технологий виртуализации ресурсов и нарезки сети [47]. В рамках данной модели рассматривается процесс передачи данных по нисходящей линии

связи БС. Выбор нисходящей линии связи для приоритетного анализа обусловлен тем, что она, как правило, характеризуется более высокими требованиями к пропускной способности по сравнению с восходящей, хотя предложенный аналитический аппарат, при соответствующей адаптации, может быть распространен и на восходящую линию связи.

Предлагаемая архитектура разделения ресурсов сети радиодоступа представлена на рис. 1.1.

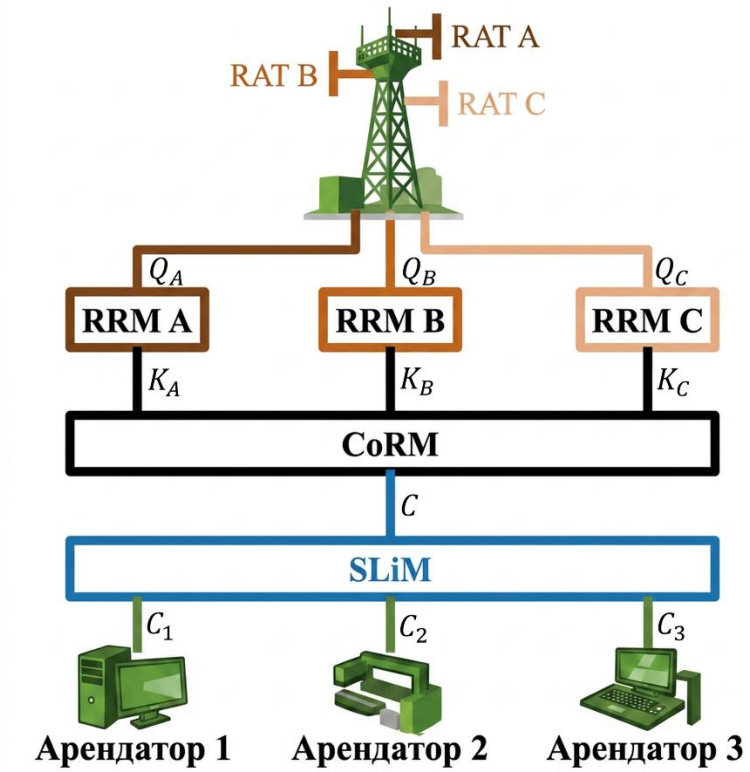


Рис. 1.1. Архитектура нарезки RAN

БС функционирует на основе нескольких технологий радиодоступа (Radio Access Technology, RAT), обозначаемых как RAT A, RAT B, RAT C и т. д. Данные технологии могут соответствовать различным частотным диапазонам стандарта 5G или представлять собой комбинацию стандартов 4G и 5G. Каждая RAT характеризуется набором физических каналов (блоков ресурсов), выделяемых абонентам, пропускная способность которых обозначается как $Q_{A_{1[\text{Гбит/с}]}}$, $Q_{A_{2[\text{Гбит/с}]}}$, …, $Q_{A_{M_A[\text{Гбит/с}]}}$; $Q_{B_{1[\text{Гбит/с}]}}$, $Q_{B_{2[\text{Гбит/с}]}}$, …, $Q_{B_{M_B[\text{Гбит/с}]}}$; …, где M_X представляет число каналов в каждой RAT. Фактическая скорость, доступная абоненту m на канале $Q_{X_{m[\text{Гбит/с}]}}$ зависит от мгновенных условий в радиоканале, таких как отношение

мощности сигнала к сумме мощностей помех и шума (Signal-to-Interference-plus-Noise Ratio, SINR). В рамках каждой RAT распределение и контроль радиоресурсов выполняются системой управления радиоресурсами (Radio Resource Management, RRM), при этом пиковая пропускная способность каждого блока определяется заданными параметрами $K_{A[\text{Гбит/с}]}, K_{B[\text{Гбит/с}]}, K_{C[\text{Гбит/с}]}, \ldots$ соответственно, т.е. $Q_{X_{m[\text{Гбит/с}]}} \leq K_{X[\text{Гбит/с}]},\ m = 1, \ldots, X_m$. Координация работы всех объектов RRM на глобальном уровне обеспечивается общим менеджером ресурсов (Common Resource Manager, CoRM).

Указанная ранее структура соответствует архитектуре гетерогенной сети единого оператора мобильной сети (Mobile Network Operator, MNO) – владельца инфраструктуры (Infrastructure Provider, InP). Поставщик инфраструктуры, часто называемый поставщиком инфраструктуры как услуги (Infrastructure as a Service, IaaS), предлагает базовые ИТ-ресурсы – такие как виртуальные серверы, сети, хранилища и центры обработки данных – в аренду или с оплатой по мере использования. Примером могу служить Amazon Web Services (AWS), Microsoft Azure и Google Cloud Platform (GCP), Alibaba Cloud (APAC), Oracle Cloud (OCI), IBM Cloud. Для реализации механизмов нарезки сети в данную модель на более высоких уровнях иерархии вводится менеджер нарезки ресурса SliM, который агрегирует максимальные емкости всех доступных RAT в суммарную емкость БС, обозначаемую как $C_{[\text{Гбит/с}]}$, и осуществляет ее распределение между арендаторами слайсов – MVNO.

В пределах каждой RAT пропускная способность определяется предоставленным спектральным ресурсом, а также зависит от используемых схем адаптивной модуляции и кодирования (Adaptive Modulation and Coding, AMC) и настроек многоантенных систем (Multiple Input Multiple Output, MIMO). Вместе с тем реальная скорость, доступная абонентам, является случайной и меняется в соответствии с мгновенными условиями радиоканала, при решении задач оптимизации распределения ресурсов на высоком уровне допустимо оперировать

значениями максимальной пропускной способности. При необходимости данные значения могут быть скорректированы в соответствии с заданными критериями.

Кроме того, учитывая, что в рассматриваемых сетях выделение ресурсов происходит в рамках временных интервалов (подкадров) длительностью 1 мс, предположим, что для исследуемых частотных диапазонов и скоростей перемещения абонентов условия радиоканала меняются реже, раз в подкадр, т.е. интервал когерентности радиоканала (интервал с постоянными условиями радиоканала) превышает 1 подкадр. Тогда состояние сети в пределах одного шага планирования распределения ресурсов менеджером нарезки можно считать стационарным. На основании вышеизложенного, в рамках данной аналитической модели вводится допущение о том, что суммарная емкость БС $C_{[\text{Гбит/с}]}$ является константой и представляет собой общий объем ресурсов, подлежащий распределению.

В рамках предложенной системной модели предполагается существование S сетевых слайсов в пределах рассматриваемой БС. Множество слайсов обозначено $\mathcal{S}$, при этом его мощность составляет $|\mathcal{S}| = S$. Емкость, выделяемая слайсу $s \in \mathcal{S}$, обозначена $C_{s[\text{Гбит/с}]} \geq 0$ Общая пропускная способность БС, обозначенная $C_{[\text{Гбит/с}]}$, ограничивает суммарную емкость, выделяемую слайсам, следующим неравенством:

$$\sum_{s \in \mathcal{S}} C_s \;\; \leq \; C. \tag{1.1}$$

Обозначим через N_s число абонентов в слайсе $s \in \mathcal{S}$. Вектор-строка $\mathbf{N}$, содержащая число абонентов во всех слайсах, представлен как $\mathbf{N}_{[1 \times S]} = (N_s)_{s \in \mathcal{S}}$. Для упрощения модели принимается допущение: один абонент поддерживает не более одного соединения в одном слайсе. Если абонент инициирует несколько параллельных соединений, каждое из них трактуется как самостоятельный логический абонент.

Предполагается, что каждый сетевой слайс ориентирован на один класс услуг (например, видеостриминг, видеоконференции, игровые приложения, передача

файлов, веб-доступ). В рамках этого допущения трафик внутри слайса рассматривается как однородный по характеристикам и требованиям к QoS. Скорость $R_{s[\text{Гбит/с}]}$, предоставляемая отдельному абоненту в слайсе s, определена принципом равного распределения: суммарная емкость слайса равномерно распределяется между всеми его активными абонентами:

$$R_s = \frac{C_s}{N_s}, \quad s \in \mathcal{S}. \tag{1.2}$$

Следует учитывать, что в реальной сети скорости отдельных абонентов внутри одного слайса не совпадают из-за стохастической природы радиоканала. При этом детальная модель внутри слайсового распределения ресурсов выходит за рамки настоящей работы.

Обозначим вектор-столбец средних скоростей по всем слайсам как $\mathbf{R}_{[S \times 1]} = (R_s)_{s \in \mathcal{S}}$.

Далее предполагается, что в соответствии с SLA между владельцем сети и арендатором слайса средняя скорость в слайсе s, определенная формулой (1.2), не опускается ниже порогового значения R_s^{min}., пока число абонентов не превышает согласованного лимита N_s^{cont}.

Формально выразим это в следующем виде:

$$0 < R_s^{\min} \leq R_s, \tag{1.3}$$

$$\text{при условии } N_s \leq N_s^{\text{cont}}. \tag{1.4}$$

Пороговое значение $R_s^{\min}$ соответствует минимальной скорости передачи данных, необходимой для удовлетворения требований к QoS услуги, предоставляемой в слайсе s. Следовательно, владелец сети предоставляет изоляцию производительности слайса до тех пор, пока число абонентов в нем не превышает число N_s^{cont}, согласованное с арендатором. Вариации параметра $R_s^{\min}$ существенны: для некоторых услуг, например, передачи файлов, $R_s^{\min}$ может быть относительно низким (приближающимся к нулю), однако достаточным для базового функционирования сервиса. Для других услуг, таких как видеоконференцсвязь, требуется строгое минимальное значение $R_s^{\min}$, при

несоблюдении которого услуга не может быть корректно предоставлена абонентам. В любом случае, $R_s^{\min}$ является параметром, согласованным между владельцем сети и арендатором слайса.

Для повышения эффективности использования сетевых ресурсов в SLA для каждого слайса устанавливается верхняя граница индивидуальной скорости:

$$R_s \leq R_s^{\max} \leq C. \tag{1.5}$$

Значение $R_s^{\max}$ определяется спецификой услуги и, как правило, соответствует уровню насыщения, выше которого дальнейшее увеличение скорости практически не улучшает показатели QoS или качества субъективного восприятия (Quality of Experience, QoE) абонента.

Очевидно, что если для ряда услуг (например, фоновая передача файлов) абонент может получить преимущество от сверхвысоких скоростей за счет сокращения времени загрузки, то для других сервисов (например, видеоконференцсвязи) превышение определенного порога скорости не повышает качество обслуживания, так как интенсивность потока ограничена параметрами используемого алгоритма кодирования. Параметр $R_s^{\max}$ является предметом договора между владельцем инфраструктуры и арендатором слайса.

Векторы-столбцы минимальных (минимально допустимых) и максимальных (максимально эффективных) скоростей передачи данных обозначаются

$$\mathbf{R}_{[S\times 1]}^{\min} = \left(R_s^{\min}\right)_{s\in\mathcal{S}} \text{ и } \mathbf{R}_{[S\times 1]}^{\max} = \left(R_s^{\max}\right)_{s\in\mathcal{S}} \tag{1.6}$$

соответственно. Поскольку емкость базовой станции (БС) конечна, обеспечить полную изоляцию по производительности при произвольно высокой нагрузке во всех слайсах одновременно невозможно. Поэтому, следуя подходу, изложенному в [1], предполагается, что изоляция слайса гарантируется до тех пор, пока число активных абонентов в нем не превышает SLA-порог N_s^{cont} (контрактный порог):

$$0 \leq N_s^{\text{cont}} \leq \left\lfloor \frac{C}{R_s^{\min}} \right\rfloor. \tag{1.7}$$

Данный порог также может быть указан в SLA (в контракте) через т.н. «контрактную емкость» слайса $C_{s[\text{Гбит/с}]}^{\text{cont}}$ или выражен через коэффициент q_s, представляющий собой долю от общего ресурса C (емкости БС):

$$q_s = \frac{C_s^{\text{cont}}}{C}. \tag{1.8}$$

В этом случае максимальное число абонентов N_s^{cont} определяется как

$$N_s^{\text{cont}} = \left\lfloor \frac{q_s C}{R_s^{\min}} \right\rfloor. \tag{1.9}$$

Предполагается, что $0 \leq q_s \leq 1$ для всех $s \in \mathcal{S}$ и выполняется условие:

$$0 \leq \sum_{s \in \mathcal{S}} q_s \leq S. \tag{1.10}$$

Последнее неравенство допускает возможность так называемого «овербукинга» (overbooking) – избыточного резервирования ресурсов [48], при котором суммарная емкость слайсов, согласованная в рамках SLA со всеми арендаторами, может превышать физическую пропускную способность БС C.

Наконец, каждому слайсу назначается приоритет $v_s \in \mathbb{N}$, где значение $v_s = 1$ соответствует наивысшему уровню приоритета. Вектор приоритетов имеет вид

$$\mathbf{v}_{[1 \times S]} = (v_s)_{s \in \mathcal{S}}. \tag{1.11}$$

Процедура приоритизации касается лишь контрактной емкости C_s^{cont} и срабатывает при наличии конфликта резервирования: когда несколько слайсов запрашивают частично совпадающие гарантированные ресурсы. Следовательно, для критически важных слайсов значение коэффициента q_s должно быть равно или близко к 1.

Предлагаемая схема нарезки обеспечивает гибкую изоляцию слайсов по показателям производительности. Эта изоляция достигается за счет использования параметров $R_s^{\min}, R_s^{\max}$ и $(q_s)_{s \in \mathcal{S}}$ (рис. 1.2). Схема предполагает динамическое выделение емкости слайсам, адаптированное к их текущим потребностям, которые определяются числом абонентов $\mathbf{N}$. Гибкость схемы заключается в том, что невостребованные договорные ресурсы не простаивают: они оперативно направляются на обслуживание более нагруженных слайсов. Это обеспечивает возможность временного увеличения выделенной емкости относительно значений, закрепленных в SLA.

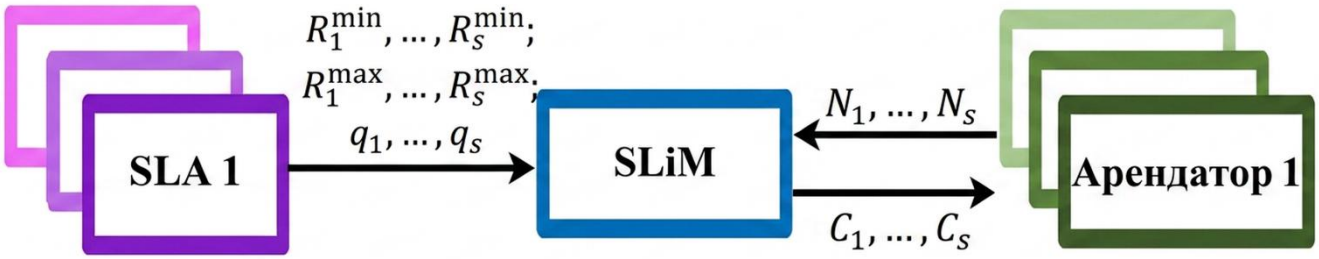


Рис. 1.2. Параметры схемы слайсинга

Иными словами, для обеспечения эффективного использования общих ресурсов, контрактная емкость не резервируется статически за каждым слайсом, как это делается при жестком резервировании ресурсов для обеспечения изоляции слайсов. Вместо этого слайс получает приоритет в доступе к ресурсам по сравнению с другими. Более того, этот приоритет является вытесняющим: при полной загрузке БС схема в первую очередь выделяет радиоресурсы слайсам, для которых выполняются условия, соответствующие требованию (1.4) (что означает, что число абонентов не превышает N_s^{cont}), отбирая ресурсы у слайсов-«нарушителей», в которых число абонентов превышает N_s^{cont}. В сценариях избыточного резервирования перекрывающаяся контрактная емкость распределяется пропорционально назначенным приоритетам слайсов, определенным вектором $\mathbf{v}_{[1 \times S]}$.

Представленная архитектура разделения ресурсов сети радиодоступа (рис. 1.1) формализует иерархию управления и взаимодействия компонентов. На основе данной архитектуры далее будет разработана математическая модель распределения ресурсов и сформулирована задача оптимизации, учитывающая динамику абонентской нагрузки и ограничения на пропускную способность.

Математическая модель распределения ресурсов и задача оптимизации

Как было отмечено выше, емкость, выделяемая слайсам, определяется динамически на каждом интервале когерентности радиоканала (интервале с постоянными условиями радиоканала) в зависимости от числа абонентов в них, которое задается вектором: $\mathbf{N} = (N_1, \dots, N_S)$. Каждый арендатор слайса может реализовывать собственную политику управления доступом к сети. В наиболее общем случае вектор состояния абонентов $\mathbf{N}$ принадлежит пространству $\Omega =$

$\{\mathbb{N} \cup \{0\}\}^S$: $\mathbf{N} \in \Omega$. Предлагаемая схема нарезки определяет для каждого состояния $\mathbf{N} \in \Omega$ соответствующие емкости слайсов $C_1(\mathbf{N}), \ldots, C_S(\mathbf{N}) \in \mathbb{R}_+$. При этом суммарная выделенная емкость ограничена общей пропускной способностью БС $C_{[\text{Гбит/с}]}$:

$$\sum_{s \in \mathcal{S}} C_s(\mathbf{N}) \leq C\,. \tag{1.12}$$

Множество всех возможных состояний Ω можно разделить на три непересекающихся подмножества, каждое из которых характеризуется специфическим режимом распределения ресурсов:

$$\Omega = \Omega^{\max} \cup \Omega^{\text{opt}} \cup \Omega^{\text{cong}}\,. \tag{1.13}$$

Подмножество состояний $\Omega^{\max}$ определяется следующим условием: $\Omega^{\max} = \{\mathbf{N} \in \Omega : \mathbf{N}\mathbf{R}^{\max} \leq C\}$. Данное подмножество соответствует ситуациям избытка ресурсов, когда совокупная максимальная требуемая емкость не превышает доступную пропускную способность БС C. Такой сценарий реализуется, если число абонентов в слайсах невелико и/или характеристики предоставляемых услуг не предъявляют высоких требований к ресурсам.

В этих условиях абонентам во всех слайсах могут быть предоставлены максимальные скорости передачи данных. Таким образом, для $\mathbf{N} \in \Omega^{\max}$: $\mathbf{R}(\mathbf{N}) = \mathbf{R}^{\max}$ и, соответственно, емкость каждого слайса s принимает значение: $C_s(\mathbf{N}) = N_s R_s^{\max}, s \in \mathcal{S}$.

Подмножество состояний Ω^{opt} определяется как: $\Omega^{\text{opt}} = \{\mathbf{N} \in \Omega: \mathbf{N}\mathbf{R}^{\min} \leq C < \mathbf{N}\mathbf{R}^{\max}$. В данном режиме доступных ресурсов недостаточно для обеспечения максимальной скорости передачи данных всем абонентам во всех слайсах, однако в каждом слайсе всем абонентам может быть обеспечена гарантированная минимальная скорость $R_s^{\min}$. В таких состояниях распределение ресурсов целесообразно осуществлять таким образом, чтобы:

– соблюдались установленные границы скорости для каждого абонента: $R_s^{\min} \leq R_s \leq R_s^{\max}$;
– обеспечивалось максимально полное использование доступной емкости БС C;

– реализовывался принцип максминной справедливости (max-min fairness) для абонентов, принимая во внимание соблюдение контрактных обязательств по числу абонентов $N_s \leq N_s^{\text{cont}}$.

Обозначим через $\mathbf{R}^*$ оптимальное распределение емкости. Оно получается из решения следующей оптимизационной постановки [2]:

$$U(\mathbf{R}) = \sum_{s \in \mathcal{S}} W_s(N_s) N_s * \ln(R_s) \rightarrow \max, \tag{1.14}$$

$$\mathbf{N}^T \mathbf{R} = C, \tag{1.15}$$

$$\mathbf{R} \in \mathbb{R}_+^S : R_s^{\min} \leq R_s \leq R_s^{\max}, s \in \mathcal{S}, \tag{1.16}$$

т.е. $\mathbf{R}^* = \underset{\mathbf{R}}{\operatorname{argmax}} U(\mathbf{R})$. Здесь под $\ln(R_s)$ понимается $\ln\left(R_{s[\text{Гбит/с}]}/1_{[\text{Гбит/с}]}\right)$, что обеспечивает безразмерность аргумента логарифмической функции.

Целевая функция $U(\mathbf{R})$ задана в логарифмической форме (1.14), что, по работе [3], поддерживает принцип «proportional fairness» посредством максимизации суммарных логарифмов скоростей. В этой работе критерий трактуется как максминно-ориентированный [4]. Соответственно, оптимизация стремится к максимизации максминной справедливости между абонентами с учетом положения текущей нагрузки относительно SLA-порога N_s^{cont}.

Весовые функции $W_s(N_s)$ в (1.14) определяются следующим образом:

$$W_s(N_s) = \begin{cases} 1, & N_s \leq N_s^{\text{cont}}, \\ N_s^{\text{cont}}/N_s, & N_s > N_s^{\text{cont}}. \end{cases} \tag{1.17}$$

Применяемые весовые функции обеспечивают максминно-справедливое распределение ресурсов между абонентами до тех пор, пока численность в соответствующих слайсах не превышает контрактный предел N_s^{cont}. Одновременно они реализуют механизм санкционирования для «нарушающих» слайсов: при $N_s > N_s^{\text{cont}}$ их веса уменьшаются, что понижает приоритет при аллокации. Ограничение (1.15) гарантирует не только невыхождение суммарного выделения за рамки доступной емкости C, но и полное использование доступного ресурса. Набор ограничений (1.16) фиксирует нижние и верхние границы скоростей: минимумы $R_s^{\min}$ обеспечивают соблюдение требований QoS, тогда как максимумы

$R_s^{\max}$ предотвращают неэффективное расходование ресурса, направляя его в те слайсы, где прирост скорости действительно повышает QoS/QoE.

Поскольку $U(\mathbf{R})$ является дифференцируемой и строго вогнутой, а допустимая область, заданная (1.15) – (1.16), выпукла и невырожденна, оптимизационная задача имеет единственное глобальное решение; его можно получить стандартными методами множителей Лагранжа. В рамках диссертационной работы для нахождения вектора распределения ресурса между слайсами, т.е. $\mathbf{R}^* = \underset{\mathbf{R}}{\operatorname{argmax}}\, U(\mathbf{R})$ путем численного решения задачи (1.14) – (1.16) использован метод проекции градиента, на основе которого сформулирован Алгоритм 1.1.

Алгоритм 1.1. Численное решение задачи (1.14) – (1.16)

Входные параметры: $C, S, \mathbf{N}, \mathbf{R}^{\min}, \mathbf{R}^{\max}, \mathbf{N}^{\text{cont}}$

Выходной параметр: $\mathbf{R}$

1. инициализация
2. $\mathbf{W}^T \coloneqq [W_1(N_1),\ \ldots,\ W_S(N_S)]$ // вектор весовых функций
3. $\mathbf{X}^{\text{stat}} \coloneqq \mathbf{W}C\left(\mathbf{W}\mathbf{N}^T\right)^{-1}$ // стационарная точка
4. **if** $R_i^{\min} \le X_i^{\text{stat}} \le R_i^{\max},\ i = 1, \ldots, S$ **then**
5. **return** $\mathbf{X}^{\text{stat}}$
6. $M_{[1\times S]} \coloneqq \mathbf{N}^T$
7. $P_{[S\times S]} \coloneqq \mathbf{I} - \mathbf{N}^T \cdot (\mathbf{N}\mathbf{N}^T)^{-1} \cdot \mathbf{N}$
8. $\mathbf{X}^0 \coloneqq \mathbf{R}^{\min} + \left(C - \mathbf{N}^T\mathbf{R}^{\min}\right)\left(\mathbf{N}\left(\mathbf{R}^{\max} - \mathbf{R}^{\min}\right)^T\right)^{-1}\left(\mathbf{R}^{\max} - \mathbf{R}^{\min}\right)$
9. $\tau \coloneqq \|\mathbf{X}^0 - \mathbf{X}^{\text{stat}}\|,\ \delta \coloneqq 1$
10. **while** $\delta > 0.0001$ **do**
11. $\mathbf{X}^1 \coloneqq \mathbf{X}^0 + \tau\mathbf{P}\, div\left(\mathbf{N}^T\mathbf{W}, \mathbf{X}^0\right)$
12. $t_{bound} \coloneqq 2,\ t_{coord} \coloneqq -1,\ \delta_+ \coloneqq 0$
13. **for** $i = \overline{1,S}$ **do**
14. **if** $N_i > 0$ **then**
15. **if** $X_i^1 < R_i^{\min}$ **then**

16. **if** $t_{bound} > \left(R_i^{\min} - X_i^0\right)\left(X_i^1 - X_i^0\right)^{-1}$ **then**
17. $t_{bound} := \left(R_i^{\min} - X_i^0\right)\left(X_i^1 - X_i^0\right)^{-1},\ t_{coord} := i$
18. **if** $X_i^1 > R_i^{\max}$ **then**
19. **if** $t_{bound} > \left(R_i^{\max} - X_i^0\right)\left(X_i^1 - X_i^0\right)^{-1}$ **then**
20. $t_{bound} := \left(R_i^{\max} - X_i^0\right)\left(X_i^1 - X_i^0\right)^{-1},\ t_{coord} := i$
21. **if** $i_{bound} < 2$ **then**
22. $\mathbf{X}^1 := \mathbf{X}^0 + t_{bound}(\mathbf{X}^1 - \mathbf{X}^0)$
23. **if** *Число строк матрицы* $\mathbf{M} < S - 1$ **then**
24. Добавить пустую строку к $\mathbf{M}$
25. $\delta_+ := 1$
26. Последняя строка матрицы $\mathbf{M} := \mathbf{I}[t_{coord}]$
27. **if** $\left\|\mathbf{M}\mathbf{M}^T\right\| > 0.0000001$ **then**
28. $\mathbf{P} := \mathbf{I} - \mathbf{M}^T\left(\mathbf{M}\mathbf{M}^T\right)^{-1}\mathbf{M}$
29. $\delta := \delta_+ + \|\mathbf{X}^0 - \mathbf{X}^1\|; \mathbf{X}^0 := \mathbf{X}^1$
30. **return** $\mathbf{X}^0$

Для решения задач оптимизации с линейными ограничениями используется проецированный градиентный метод – классический алгоритм, реализуемый через стандартную итерационную схему [49]:

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + \tau\mathbf{d}^{(k)}. \tag{1.18}$$

Здесь $\mathbf{X}^{(k)}$ обозначает точку, полученную на k-й итерации алгоритма; τ – шаг итерации; $\mathbf{d}^{(k)}$ – вектор приращения. Вектор $\mathbf{d}^{(k)}$ вычисляется как проекция градиента целевой функции $\nabla U\left(\mathbf{X}^{(k)}\right)$ на подпространство допустимых направлений, задаваемое линейными ограничениями

$$\mathbf{d}^{(k)} = \mathbf{P}\nabla U\left(\mathbf{X}^{(k)}\right), \tag{1.19}$$

где матрица проектирования $\mathbf{P}$ изначально задается как

$$\mathbf{P} = \mathbf{I} - \mathbf{N}^{\mathbf{T}}(\mathbf{N}\mathbf{N}^{\mathbf{T}})^{-1}\mathbf{N}. \tag{1.20}$$

Теперь необходимо расширить разработанную политику управления ресурсами на состояния $\mathbf{N} \in \Omega^{\text{cong}}$, то есть на ситуации, когда ресурсов

недостаточно для обеспечения всем абонентам даже минимальных скоростей передачи данных. Известен подход, состоящий в определении отображения состояний из Ω^{cong} в одно из менее нагруженных подмножеств $\Omega^{\max} \cup \Omega^{\mathrm{opt}}$. В диссертационной работе разработан второй подход, заключающийся в непосредственном определении $C_s(\mathbf{N})$ для состояний из подмножества $\Omega^{\mathrm{cong}} = \{\mathbf{N} \in \Omega : \mathbf{N}\mathbf{R}^{\min} > C\}$.

Для состояний $\mathbf{N} \in \Omega^{\mathrm{cong}}$ хотя бы в одном слайсе имеется недостаток ресурса для обеспечения требований QoS. В отличие от модели [2] с жестким разделением ресурса, где пространство состояний модели не содержало таких состояний ввиду строгого контроля доступа, поскольку прием заявок-«нарушителей» блокировался на уровне доступа в систему, концепция нарезки ресурсов предполагает эффективное использование ресурсов, которые не должны простаивать. В диссертационной работе предлагается методика определения емкости слайсов при наличии принятых в систему заявок-«нарушителей», сформулированная в виде теорем.

Вначале рассмотрим сценарий, при котором все слайсы обладают равным приоритетом. Для вектора $\mathbf{N} = (N_1, N_2, \dots, N_s)$ обозначим $N_s^{\min}(\mathbf{N}) = \min\{N_s, N_s^{\mathrm{cont}}\}$ эффективное число абонентов в слайсе s, ограниченное контрактным порогом N_s^{cont}. Вектор эффективного числа абонентов имеет вид $\mathbf{N}_{[1\times S]}^{\min} = \left(N_s^{\min}\right)_{s\in\mathcal{S}}$.

Теорема 1.1. В системе с нарезкой радиоресурсов без приоритизации слайсов в состоянии $\mathbf{N}_{[1\times S]} = (N_1, N_2, \dots, N_S)$, когда ресурсов недостаточно для обеспечения всем абонентам минимальных скоростей передачи данных, емкость слайса $C_s(\mathbf{N})$ определяется следующим образом:

$$C_s(\mathbf{N}) = \begin{cases} \dfrac{N_s^{\min} R_s^{\min}}{\mathbf{N}^{\min}\mathbf{R}^{\min}} C, & \mathbf{N}^{\min}\mathbf{R}^{\min} \geq C; \\ N_s^{\min} R_s^{\min} + \dfrac{\left(N_s - N_s^{\min}\right) R_s^{\min}}{\left(\mathbf{N} - \mathbf{N}^{\min}\right)\mathbf{R}^{\min}} \left(C - \mathbf{N}^{\min}\mathbf{R}^{\min}\right), & \mathbf{N}^{\min}\mathbf{R}^{\min} < C. \end{cases} \quad (1.21)$$

Доказательство.

Если суммарная минимально необходимая емкость для эффективных абонентов превышает общую емкость БС, т.е., $\mathbf{N}^{\min}\mathbf{R}^{\min} \geq C$, тогда емкость $C_s(\mathbf{N})$ распределяется пропорционально этому требованию:

$$C_s(\mathbf{N}) = \frac{N_s^{\min} R_s^{\min}}{\mathbf{N}^{\min}\mathbf{R}^{\min}} C. \tag{1.22}$$

В противном случае, условие $\mathbf{N}^{\min}\mathbf{R}^{\min} < C$ означает наличие некоторого избытка ресурсов даже после обеспечения минимальных скоростей эффективным абонентам. Тогда емкость слайса $C_s(\mathbf{N})$ определяется как

$$C_s(\mathbf{N}) = N_s^{\min} R_s^{\min} + \frac{\left(N_s - N_s^{\min}\right) R_s^{\min}}{\left(\mathbf{N} - \mathbf{N}^{\min}\right)\mathbf{R}^{\min}} \left(C - \mathbf{N}^{\min}\mathbf{R}^{\min}\right), \tag{1.23}$$

где $\mathbf{N}_{[1\times S]}^{\min} = \left(N_1^{\min}, N_2^{\min}, \dots, N_S^{\min}\right)$.

Теорема доказана.

Теперь предположим, что слайсы обладают различными приоритетами. Множество всех слайсов $\mathcal{S}$ разбивается на V непересекающихся подмножеств: $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_V$, где $\mathcal{S}_1$ содержит слайсы с наивысшим приоритетом, $\mathcal{S}_2$ – слайсы с приоритетом следующего уровня, и так далее, вплоть до $\mathcal{S}_V$, содержащего слайсы с самым низким приоритетом.

Для каждого уровня приоритета введем величину $c_u(\mathbf{N})$:

$$c_u(\mathbf{N}) = \sum_{k=1}^{u} \sum_{s \in \mathcal{S}_k} N_s^{\min} R_s^{\min}, \quad 1 \leq u \leq V.$$

Величина $c_u(\mathbf{N})$ представляет собой суммарную минимально требуемую емкость в состоянии $\mathbf{N}$, которую владелец сети обязан гарантировать для всех слайсов с приоритетом u и выше.

Теорема 1.2. В системе с нарезкой радиоресурсов с приоритизацией слайсов в ситуации дефицита ресурсов в состоянии $\mathbf{N}_{[1\times S]} = (N_1, N_2, \dots, N_S)$, когда суммарная минимальная требуемая емкость превышает доступную пропускную способность БС, емкость C распределяется в соответствии с установленными приоритетами следующим образом.

1. Если $c_1(\mathbf{N}) \geq C$ (недостаточно ресурсов даже для наивысшего приоритета), емкость распределяется только среди слайсов наивысшего приоритета $\mathcal{S}_1$:

$$C_s(\mathbf{N}) = \begin{cases} \dfrac{N_s^{\min} R_s^{\min}}{c_1(\mathbf{N})} C, & s \in \mathcal{S}_1, \\ 0, & s \in \mathcal{S} \setminus \mathcal{S}_1, \end{cases} \tag{1.24}$$

2. Если $c_{u^*}(\mathbf{N}) < C$ и $c_{u^*+1}(\mathbf{N}) \geq C$ (дефицит возникает на уровне $u^* + 1$):

$$C_s(\mathbf{N}) = \begin{cases} N_s^{\min} R_s^{\min}, & s \in \mathcal{S}_i : 1 \leq i \leq u^*; \\ \dfrac{N_s^{\min} R_s^{\min}}{c_{u^*+1}(\boldsymbol{N}) - c_{u^*}(\boldsymbol{N})} \left(C - c_{u^*}(\boldsymbol{N})\right), & s \in \mathcal{S}_{u^*+1}; \\ 0, & s \in \mathcal{S}_i : i > u^* + 1. \end{cases} \tag{1.25}$$

Доказательство.

Слайсы с приоритетами от 1 до u^* гарантированно получают свою минимальную емкость. Остаток доступной емкости $\left(C - c_{u^*}(\boldsymbol{N})\right)$ распределяется между слайсами уровня $u^* + 1$, пропорционально их минимальным требованиям. Слайсы с более низкими приоритетами ($i > u^* + 1$) в этом сценарии не получают ресурсов. При $\mathbf{N} \in \Omega^{\text{cong}}$, как и для других состояний, скорость передачи данных для каждого абонента в слайсе s определяется как $R_s(\boldsymbol{N}) = \frac{C_s(\boldsymbol{N})}{N_s}$, $s \in \mathcal{S}$.

Заметим, что при наличии гарантированной емкости $c_V(\mathbf{N}) = \mathbf{N}^{\min}\mathbf{R}^{\min} < C$, то есть для каждого слайса может быть предоставлена гарантированная емкость, емкости слайсов $C_s(\mathbf{N})$ рассчитываются по формуле (1.23).

Теорема доказана.

Алгоритм 1.1 и Теоремы 1.1 и 1.2 составляют разработанную в диссертационной работе методику определения емкости слайсов, обеспечивающую выполнение двух ключевых требований концепции сетевой нарезки радиоресурсов: обеспечения изоляции слайсов при максимальном использовании ресурса системы. Методика соответствует решению оптимизационной задачи (1.14) с ограничениями (1.15) – (1.16).

1.3. Постановка задачи исследования

Исследования, проведенные в разделах 1.1–1.2 диссертационной работы, показали, что технология нарезки ресурсов сети является одним из фундаментальных механизмов в сетях радиодоступа 5G, предназначенных для эффективного управления и распределения ресурсов между услугами с разнообразными требованиями к QoS.

При этом разработанная выше методика нарезки ресурсов предполагает, что ресурс для приоритетных услуг выделяется за счет снижения качества предоставления услуг категории BE, генерирующих эластичный трафик без ограничений на минимальную скорость передачи и задержку доставки данных. Таким образом, актуальной является задача исследования качества предоставления услуги BE, для решения которой в Главе 2 необходимо построить математическую модель слайса услуги BE в виде СМО с PS и эластичным трафиком для анализа основных показателей качества предоставления услуги, а именно, средней доли занятого ресурса слайса, среднего числа абонентов в слайсе, средней скорости, на которой абонент получает услугу, а также важного нового показателя – вероятности деградации обслуживания в слайсе. Также в Главе 2 необходимо с помощью имитационной модели исследовать качество обслуживания абонента слайса BE без применения механизма нарезки ресурса и с применением этого механизма, при этом событиями, вызывающими процедуру нарезки ресурса, служат поступление запроса на предоставление услуги от нового абонента в слайсе, окончание получения услуги абонентом в слайсе, обнаружение деградации обслуживания, срабатывание регулярного таймера.

Кроме указанных проблем с деградацией качества предоставления услуг BE с эластичным трафиком в случае применения механизма нарезки ресурса при переходе к развертыванию сетей пятого и шестого поколений (5G/6G) в mmWave и sub-THz диапазонах возникают новые задачи распределения ресурсов в условиях соблюдения требований к QoS, связанные с ограниченным радиусом распространения сигнала и высокой чувствительностью к блокировкам. Для

решения этих проблем активно внедряется технология IAB, которая использует беспроводные ретрансляторы (IAB-узлы) для создания многошаговой транспортной сети. Такая архитектура, несмотря на свои преимущества в расширении покрытия и снижении затрат на развертывание, неизбежно приводит к увеличению сквозной задержки пакетов из-за многошаговой передачи, буферизации на промежуточных узлах и, что особенно важно, ограничений разделения ресурса по времени при мультиплексировании с временным разделением (Time Division Multiplexing, TDM). В контексте критически важных приложений 5G/6G, таких как промышленный Интернет вещей, телемедицина и автономный транспорт, помимо низкой сквозной задержки, ключевое значение приобретает актуальность передаваемых данных, что требует применения важной метрики возраста информации AoI, которая количественно отражает свежесть информации от удаленного источника, имеющейся в системе управления и мониторинга удаленных подсистем. Таким образом, в Главе 3 диссертационной работы необходимо выполнить аналитический обзор научно-технической литературы в части исследований показателей качества сети IAB и решить задачу разделения ресурсов сети IAB, которая сформулирована в виде проблемы минимизации сквозной задержки и пикового возраста информации на всех маршрутах от удаленных подсистем. Решением проблемы минимизации должны стать доли пропускных способностей звеньев маршрута, выделенные для передачи трафика от соответствующей удаленной подсистемы, которые являются параметрами методики разделения менеджером нарезки SliM ресурса БС по времени при мультиплексировании с временным разделением, обеспечивающей минимизацию сквозной задержки и пикового возраста информации, а также оценку правостороннего квантиля уровня 0,999 пикового возраста информации на всех маршрутах от IAB-донора до АУ как в полудуплексном, так и дуплексном режимах передачи данных. Полудуплексный режим (Half-Duplex, HD) позволяет передавать данные в обоих направлениях, но только по очереди (не одновременно). Полнодуплексный режим (Full-Duplex, FD) обеспечивает одновременную передачу и прием данных обоими участниками соединения.

Таким образом, **цель** диссертационной работы состоит в анализе и расчете показателей качества предоставления услуг в сети интегрированного доступа и транзита с разделением ресурсов с использованием марковских моделей систем массового обслуживания.

Для достижения этой цели в диссертационной работе решаются следующие **задачи**.

– Разработка метода разделения ресурсов беспроводной сети при динамическом выделении ресурса на основе максминной справедливости в условиях приоритизации слайсов.
– Построение математической модели слайса в виде системы массового обслуживания с дисциплиной разделения процессора и эластичным трафиком с ограничением на максимальную скорость передачи, позволяющей провести сравнительный анализ влияния методов вызова процедуры нарезки по показателям эффективности – вероятности деградации обслуживания, коэффициенту использования ресурса и частоты вызова процедуры нарезки.
– Разработка алгоритма расчета оптимальных долей времени активности каналов сети интегрированного доступа и транзита с разделением ресурсов по времени, минимизирующих среднюю сквозную задержку. Вычисление правостороннего квантиля заданного уровня пикового возраста информации на маршруте, а также получение в явном виде функции распределения сквозной задержки с помощью распределения фазового типа.

Логика изложения материала в работе с учетом поставленных задач построена следующим образом (табл. 1.1). Для каждой из задач построены модели обслуживания эластичного трафика с учетом особенностей IAB сетей и ограничений полудуплексной передачи данных в радиоканале. Различные способы анализа и алгоритмы позволяют оценить предложенные принципы управления ресурсами, отраженные в моделях с различных аспектов (в показателях эффективности, в функциях полезности для оптимизационных задач, в показателях производительности сети).

Табл. 1.1. Структура диссертационной работы

	Особенности моделей	**Результат**
Результат 1	– приоритизация слайсов, – динамическое выделение ресурсов на основе максминной справедливости; – избыточное резервирование ресурсов (овербукинг). (Глава 1, разделы 1.1, 1.2)	Методика нарезки ресурсов в условиях недостатка ресурса. Алгоритм численного решения задач оптимизации распределения ресурсов (метод проекции градиента). (Глава 1, раздел 1.2)
Результат 2	– слайс с ограничением на максимальную скорость передачи. Методы вызова процедуры нарезки: – по поступлению заявки; – по уходу заявки; – по регулярному таймеру; – по событию деградации обслуживания; – по любому из перечисленных событий. (Глава 2, раздел 2.1)	Аналитические выражения для показателей эффективности (коэффициент использования ресурса, среднее число заявок, средняя скорость обслуживания абонента, вероятность деградации обслуживания). Сравнительный анализ методов вызова процедуры нарезки. (Глава 2, разделы 2.1, 2.2)
Результат 3	– полудуплексный / полнодуплексный режимы передачи данных (Глава 3, разделы 3.1, 3.2)	ФР фазового типа сквозной задержки и пикового возраста информации на маршруте. Алгоритм расчёта среднего значения и правостороннего квантиля заданного уровня сквозной задержки и пикового возраста информации на маршруте и средних значений по сети. (Глава 3, разделы 3.3, 3.4)

ГЛАВА 2
АНАЛИЗ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ НАРЕЗКИ И ВЕРОЯТНОСТНАЯ МОДЕЛЬ

2.1. Вероятностная модель и аналитические выражения показателей эффективности

Для анализа системы, представленной в Главе 1, используется аппарат теории массового обслуживания [50-56]: каждый сетевой слайс трактуется как отдельная СМО. Тип СМО подбирается так, чтобы адекватно отражать профиль услуги, которую обслуживает соответствующий слайс. Заявки в модели интерпретируются как абонентские сессии. Поскольку все слайсы функционируют в рамках единой сети с общей емкостью $C_{[\text{бит}\backslash\text{ед.вр.}]}$, набор S СМО разделяет общий ресурс C; часть ресурса, доступная СМО с индексом s, обозначается как C_s, $s \in \mathcal{S}$, при соблюдении условия (1.1).

В настоящем исследовании рассматривается один класс слайсов BE без контроля доступа, но с ограничением на максимальную скорость обслуживания абонента; далее обозначим его как $BE^{\max}$. Такой слайс моделируется СМО с PS и эластичным трафиком [54,57]. Индивидуальная скорость обслуживания $R_{s\,[\text{бит}\backslash\text{ед.вр.}]}$ одинакова для всех абонентов слайса и обратно пропорциональна их числу N_s, однако ограничена сверху значением $R_s^{\max}{}_{\,[\text{бит}\backslash\text{ед.вр.}]}$. Эта постановка хорошо аппроксимирует услуги передачи файлов и буферизованные трансляции.

Рассматриваемую модель иллюстрирует рис. 2.1: $A_s(x)$ – функция распределения интервалов между поступлениями заявок, $B_s(x)$ – функция распределения размеров заявок, где $s \in \mathcal{S}$.

Политики управления доступом и внутрислайсового распределения ресурсов задаются индивидуально для каждого типа слайса и выбираются в соответствии с характером предоставляемой услуги. Для трафика $BE^{\max}$ предполагается отсутствие механизма контроля доступа и единая для всех абонентов скорость обслуживания; число одновременно обслуживаемых заявок не ограничивается.

Поскольку в СМО с дисциплиной разделения процессора PS отсутствует механизм контроля доступа и не предусмотрена очередь для ожидания заявками освобождения ресурса (приборов), при высокой интенсивности поступления заявок ресурс, выделенный каждой, будет стремиться к нулю, что соответствует $R_s^{\min} = 0$ для требования к минимальной скорости обслуживания в слайсе s в обозначениях модели Главы 1. В этих условиях имеет смысл ввести понятие «деградации обслуживания» абонента в слайсе, которая наступает при пересечении доступной абоненту скоростью обслуживания R_s некоторого порога $R^{\deg}_{[\text{бит\ед.вр.}]}$, $0 < R^{\deg} \leq C$, и считать, что для абонентов, получающих услугу на скорости ниже пороговой, услуга предоставляется с ненадлежащим качеством. Событием, вызвавшим деградацию обслуживания, может стать поступление очередной заявки и/или результат выполнения процедуры нарезки ресурса в пользу слайсов с более высоким приоритетом. Количественной мерой деградации обслуживания может служить вероятность деградации, т.е. доля времени получения абонентами слайса услуги на скорости ниже R_s, в течение достаточно большого интервала наблюдения за слайсом.

Анализ нарезки ресурсов проведен на примере произвольного слайса $s \in \mathcal{S}$, относящегося к типу $BE^{\max}$ «наилучшее из возможного с ограничением $R_s^{\max}$ на максимальную скорость обслуживания» (на рис. 2.1 выделен пунктирной линией). Распределение ресурсов в рассматриваемом слайсе моделируется с помощью СМО с дисциплиной разделения процессора PS

$$\begin{matrix} G \\ A(x) \end{matrix} \Big| \begin{matrix} G \\ B(x) \end{matrix} \Big| C \Big| 0, \qquad R^{max}, \tag{2.1}$$

при этом для удобства до конца раздела 2.1 опустим индекс s у параметров этой единственной рассматриваемой в разделе СМО: $A(x), B(x), C, R^{\max}$ и т.д. Примеры систем с потоками разных видов можно найти в [54,58-61].

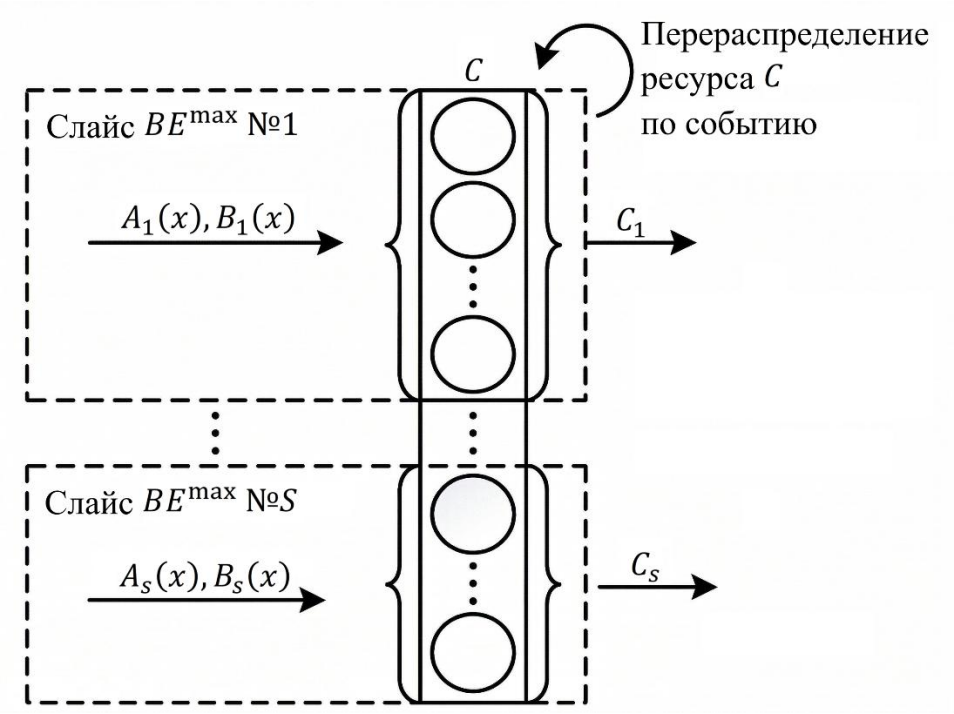


Рис. 2.1. Система из S слайсов типа $BE^{\max}$

Рассмотрим СМО (2.1) в предположении о пуассоновско-экспоненциальной нагрузке: законы распределения интервалов между поступлениями заявок и размера заявок имеют вид $A(x) = Exp(\lambda)$ и $B(x) = Exp(\theta)$, где $\lambda_{[\text{заявок\ед.вр.}]}$ – интенсивность поступления заявок, а $\theta_{[\text{бит}]}$ – средний размер заявки:

$$\begin{matrix} M \\ \lambda \end{matrix} \Big| \begin{matrix} M \\ \theta \end{matrix} \Big| \; C \; \Big| 0, \qquad R^{\max}. \tag{2.2}$$

Ресурс, выделяемый заявке в СМО с дисциплиной разделения процессора PS, зависит от числа заявок в системе, т.о. событиями, приводящими к перераспределению ресурса, являются поступление и при уходе заявки. Длительность обслуживания заявки эластичного трафика складывается из интервалов времени между соседними событиями от момента поступления этой заявки в СМО (момент приема на обслуживание) до момента ухода этой заявки из СМО (момент окончания обслуживания) и зависит от ресурса (скорости обслуживания), который выделяется для обслуживания этой заявки на каждом таком интервале, с ограничением максимального ресурса $R^{\max}$.

СП $X(t)$, соответствующий числу заявок в СМО в момент времени $t \geq 0$, вследствие предположения о пуассоновско-экспоненциальной нагрузке СП $X(t)$

является марковским (МП). Пространство состояний МП $X(t)$ имеет вид $\mathcal{X} = \{0, 1, \dots, \infty\}$.

Как и в формуле (1.13) для модели Главы 1, пространство состояний СМО (2.2) можно разделить на три подпространства:

$$\mathcal{X} = \mathcal{X}^{\max} \cup \mathcal{X}^{\text{opt}} \cup \mathcal{X}^{\text{cong}}, \tag{2.3}$$

$$\mathcal{X}^{\max} = \{ N \in \mathcal{X} : NR^{\max} \leq C\}, \tag{2.4}$$

$$\mathcal{X}^{\text{opt}} = \{ N \in \mathcal{X} : NR^{\max} > C\}, \tag{2.5}$$

$$\mathcal{X}^{\text{cong}} = \emptyset. \tag{2.6}$$

где $\mathcal{X}^{\max}$ – подмножество состояний, в которых поступающая заявка принимается в СМО и все заявки обслуживаются с максимальной скоростью $R^{\max}$, $\mathcal{X}^{\text{opt}}$ – подмножество состояний, в которых поступающая заявка принимается в СМО, после чего вызывается процедура нарезки ресурса для всех заявок в соответствии с дисциплиной разделения процессора PS, $\mathcal{X}^{\text{cong}}$– подмножество состояний, в которых поступающая заявка не принимается в СМО.

Введем $M = \left\lfloor \frac{C}{R^{\max}} \right\rfloor$ – максимальное число заявок в СМО, которые можно обслужить со скоростью $R^{\max}$, т.о. $|\mathcal{X}^{\max}| = M + 1$.

В предположении о существовании стационарного режима введем стационарные вероятности

$$p_N = \lim_{t\to\infty} \mathrm{P}\{X(t) = N\}, \quad N \in \mathcal{X}. \tag{2.7}$$

Теорема 2.1. Для МП $X(t)$ над пространством состояний $\mathcal{X} = \{0, 1, \dots, \infty\}$, описывающего число заявок в СМО (2.2), стационарное распределение вероятностей состояний $\{p_N, N \geq 0\}$ имеет вид

$$p_N = \begin{cases} \dfrac{1}{N!}\left(\dfrac{\lambda\theta}{R^{\max}}\right)^N p_0, & 1 \leq N \leq M; \qquad (2.8) \\ \left(\dfrac{\lambda\theta}{C}\right)^{n-M} p_M = \dfrac{1}{M!}\dfrac{(\lambda\theta)^N}{C^{N-M}(R^{\max})^M}\, p_0, & n \geq M, \qquad (2.9) \end{cases}$$

где p_0 имеет вид

$$p_0 = \left(1 + \sum_{N=1}^{M-1} \frac{1}{N!}\left(\frac{\lambda\theta}{R^{\max}}\right)^n + \frac{1}{M!}\left(\frac{\lambda\theta}{R^{\max}}\right)^M \sum_{N=M}^{\infty} \left(\frac{\lambda\theta}{C}\right)^{N-M}\right)^{-1}, \tag{2.10}$$

а условия существования стационарного режима определяется неравенством

$$\lambda < \frac{C}{\theta}. \tag{2.11}$$

Доказательство.

Граф интенсивностей переходов $X(t)$ представлен на рис. 2.2.

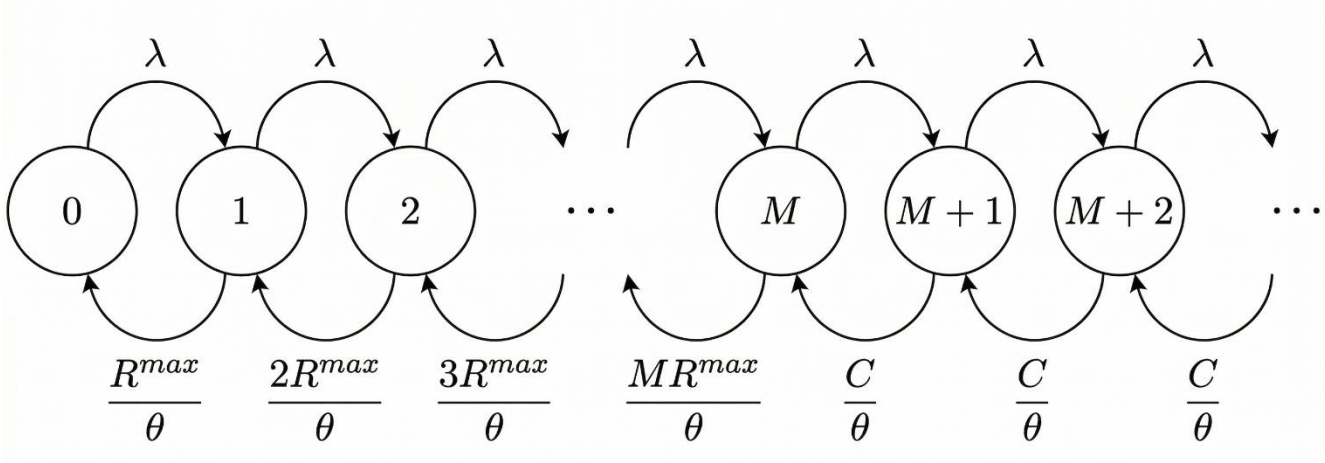


Рис. 2.2. Граф интенсивностей переходов МП $X(t)$

Элементы матрицы $\mathbf{A} = \left[a_{ij}\right]_{i,j\in\Omega}$ интенсивностей переходов МП $X(t)$ имеют вид

$$a_{ij} = \begin{cases} \lambda, & j = i+1; \\ \dfrac{jR^{\max}}{\theta}, & j = i-1, i, j \in \mathcal{X}^{\max}; \\ \dfrac{C}{\theta}, & j = i-1, j \in \mathcal{X}^{\text{opt}}; \\ -\left(\lambda + \dfrac{jR^{\max}}{\theta}\right), & j = i,\ \ i,j \in \mathcal{X}^{\max}; \\ -\left(\lambda + \dfrac{C}{\theta}\right), & j = i,\ \ i,j \in \mathcal{X}^{\text{opt}}; \\ 0, & \text{в ост. случаях.} \end{cases} \tag{2.12}$$

С момента, когда число N заявок в СМО превышает значение M, интенсивность обслуживания достигает своего верхнего предела $\frac{C}{\theta}$, а скорость $R(N)$ обслуживания каждого из N абонентов, $R(N) < R^{\max}$, определяется в соответствии с дисциплиной разделения процессора PS:

$$R(N) = \frac{C}{N}, \quad N > M, \qquad N \in \mathcal{X}^{\text{opt}}. \tag{2.13}$$

Таким образом, условие существования стационарного режима имеет вид (2.11)

$$\lambda < \frac{C}{\theta},$$

т.е. интенсивность поступления заявок не должна превышать максимальную интенсивность их обслуживания.

Стационарные вероятности $\{p_N, N \geq 0\}$, для которых по формуле полной вероятности выполняется условие нормировки

$$\sum_{N=0}^{\infty} p_N = 1, \tag{2.14}$$

удовлетворяют системе уравнений глобального баланса (СУГБ):

$$\begin{cases} \lambda\, p_0 = \dfrac{R^{\max}}{\theta} p_1, \quad \text{если } N = 0; \\ \left(\lambda + \dfrac{N\, R^{\max}}{\theta}\right) p_N = \lambda\, p_{N-1} + \dfrac{(N+1)R^{\max}}{\theta}\, p_{N+1}, \text{если } 1 \leq N < M; \\ \left(\lambda + \dfrac{M\, R^{\max}}{\theta}\right) p_M = \lambda\, p_{M-1} + \dfrac{C}{\theta} p_{M+1}, \quad \text{если } N = M; \\ \left(\lambda + \dfrac{C}{\theta}\right) p_N = \lambda p_{N-1} + \dfrac{C}{\theta} p_{(N+1)}, \text{если } N > M. \end{cases} \tag{2.15}$$

Кроме того, $\{p_N, N \geq 0\}$ удовлетворяют системе уравнений локального баланса (СУЛБ):

$$\begin{cases} \lambda p_{N-1} = \dfrac{N\, R^{\max}}{\theta} p_N, \quad \text{если } 1 \leq N \leq M; \\ \lambda\, p_{N-1} = \dfrac{C}{\theta} p_N, \qquad \text{если } N > M. \end{cases} \tag{2.16}$$

Решение $\{p_N, N \geq 0\}$ СУГБ и СУЛБ совпадают и имеют вид (2.8) – (2.9)

$$p_N = \begin{cases} \dfrac{1}{N!}\left(\dfrac{\lambda\theta}{R^{\max}}\right)^N p_0, & 1 \leq N \leq M; \\ \left(\dfrac{\lambda\theta}{C}\right)^{n-M} p_M = \dfrac{1}{M!}\dfrac{(\lambda\theta)^N}{C^{N-M}(R^{\max})^M}\, p_0, & n \geq M, \end{cases}$$

где p_0 определяется из условия нормировки (2.14) и имеет вид

$$p_0 = \left(1 + \sum_{N=1}^{M-1} \frac{1}{N!}\left(\frac{\lambda\theta}{R^{\max}}\right)^n + \frac{1}{M!}\left(\frac{\lambda\theta}{R^{\max}}\right)^M \sum_{N=M}^{\infty} \left(\frac{\lambda\theta}{C}\right)^{N-M}\right)^{-1}.$$

Распределение (2.8) – (2.9) можно также записать в виде

$$p_N = \frac{1}{\min(N,M)!}\frac{(\lambda\theta)^N}{C^{\max(0,N-M)}(R^{max})^{\min(N,M)}}\,p_0, \quad N \geq 0. \tag{2.17}$$

Поскольку $\frac{\lambda\theta}{C} < 1$, второе слагаемое в (2.10) представляет собой сумму членов бесконечно убывающей геометрической прогрессии

$$\sum_{n=M}^{\infty}\left(\frac{\lambda\theta}{C}\right)^{n-M} = \sum_{n=0}^{\infty}\left(\frac{\lambda\theta}{C}\right)^{n} \tag{2.18}$$

со знаменателем $q = \frac{\lambda\theta}{C}$ и первым членом $b_1 = 1$, сумма S_∞ которой определяется выражением

$$S_\infty = \lim_{n\to\infty}\sum_{n=0}^{\infty} b_1 q^n = \frac{b_1}{1-q}. \tag{2.19}$$

Следовательно, в условиях $\left|\frac{\lambda\theta}{C}\right| < 1$ эргодичности МП $X(t)$ находим

$$\sum_{n=0}^{\infty}\left(\frac{\lambda\theta}{C}\right)^{n} = \frac{1}{1-\frac{\lambda\theta}{C}} = \frac{C}{C-\lambda\theta}, \tag{2.20}$$

и из (2.10) p_0 имеет вид

$$p_0 = \left(1 + \sum_{n=1}^{M-1}\frac{1}{n!}\left(\frac{\lambda\theta}{R^{\max}}\right)^n + \frac{1}{M!}\left(\frac{\lambda\theta}{R^{\max}}\right)^M \frac{C}{C-\lambda\theta}\right)^{-1}. \tag{2.21}$$

Теорема доказана.

Зная стационарное распределение $\{p_N, N \geq 0\}$ числа заявок в СМО (2.2), докажем несколько утверждений, определяющих основные показатели эффективности СМО, а именно, среднюю долю *UTIL* занятого ресурса, среднее число N^{avg} абонентов в системе, среднюю скорость R^{avg}, на которой абонент получает услугу, а также вероятность P^{deg} деградации обслуживания.

Утверждение 2.1. Средняя доля *UTIL* занятого ресурса СМО (2.2) в стационарном режиме может быть найдена по формуле

$$UTIL = \frac{R^{\max}}{C}\sum_{N=0}^{M} N\,p_N + \frac{p_0}{M!}\left(\frac{\lambda\theta}{R^{\max}}\right)^M \frac{\lambda\theta}{C-\lambda\theta}. \tag{2.22}$$

Доказательство.

При $0 \le N \le M$ используется только часть ресурса C СМО, равная $NR^{\max}$. В остальных случаях, при $N > M$, используется весь ресурс. Отсюда получаем:

$$UTIL = \frac{1}{C}\left[\sum_{N=0}^{M} N\,R^{\max} p_N + \sum_{N=M+1}^{\infty} C\,p_N\right] = \frac{1}{C}\left[R^{\max}\sum_{N=0}^{M} N\,p_N + C\sum_{N=M+1}^{\infty} p_N\right] =$$

$$= \frac{R^{\max}}{C}\sum_{N=0}^{M} N\,p_N + \frac{p_0}{M!}\left(\frac{\lambda\theta}{R^{\max}}\right)^M \sum_{N=M+1}^{\infty}\left(\frac{\lambda\theta}{C}\right)^{N-M} =$$

$$= \frac{R^{\max}}{C}\sum_{N=0}^{M} N\,p_N + \frac{p_0}{M!}\left(\frac{\lambda\theta}{R^{\max}}\right)^M \sum_{N=1}^{\infty}\left(\frac{\lambda\theta}{C}\right)^{N} =$$

$$= \frac{R^{\max}}{C}\sum_{N=0}^{M} N\,p_N + \frac{p_0}{M!}\left(\frac{\lambda\theta}{R^{\max}}\right)^M \frac{\lambda\theta}{C-\lambda\theta}.$$

Получили выражение (2.22).

Утверждение доказано.

Можно получить альтернативную формулу записи показателя *UTIL.*

<u>Утверждение 2.2.</u> Средняя доля *UTIL* занятого ресурса СМО (2.2) в стационарном режиме может быть найдена по формуле

$$UTIL = 1 - \sum_{N=0}^{M}\left(1 - \frac{NR^{\max}}{C}\right)p_N. \qquad (2.23)$$

Доказательство.

Проведем рассуждения аналогично доказательству Утверждения 2.1, сделав замену $\sum_{N=M+1}^{\infty} p_N$:

$$UTIL = \left[\sum_{N=0}^{M} N\,R^{\max} p_N + \sum_{N=M+1}^{\infty} C\,p_N\right] = \frac{1}{C}\left[R^{\max}\sum_{N=0}^{M} N\,p_N + C\sum_{N=M+1}^{\infty} p_N\right] =$$

$$= \frac{R^{\max}}{C}\sum_{N=0}^{M} N\,p_N + \left(1 - \sum_{N=0}^{M} p_N\right) = 1 + \frac{R^{\max}}{C}\sum_{N=0}^{M} N\,p_N - \sum_{N=0}^{M} p_N =$$

$$= 1 - \sum_{N=0}^{M}\left(1 - \frac{NR^{\max}}{C}\right)p_N = 1 - \sum_{N=0}^{M}\left(1 - \frac{NR^{\max}}{C}\right)p_N.$$

Получили выражение (2.23).

Утверждение доказано.

Утверждение 2.3. Среднее число N^{avg} заявок в СМО (2.2) имеет вид

$$N^{\text{avg}} = \sum_{n=1}^{M-1} n\,p_n + Mp_M \sum_{n=0}^{\infty} \left(\frac{\lambda\theta}{C}\right)^n + p_M \sum_{n=0}^{\infty} n \left(\frac{\lambda\theta}{C}\right)^n. \tag{2.24}$$

Доказательство.

По определению математического ожидания дискретной случайной величины (с.в.)

$$N^{\text{avg}} = \sum_{n=1}^{\infty} n\,p_n = \sum_{n=1}^{M-1} n\,p_n + \sum_{n=M}^{\infty} n\,p_n = \sum_{n=1}^{M-1} n\,p_n + p_M \sum_{n=M}^{\infty} n \left(\frac{\lambda\theta}{C}\right)^{n-M} =$$

$$= \sum_{n=1}^{M-1} n\,p_n + p_M \sum_{n=0}^{\infty} (n+M) \left(\frac{\lambda\theta}{C}\right)^n = \sum_{n=1}^{M-1} n\,p_n + Mp_M \sum_{n=0}^{\infty} \left(\frac{\lambda\theta}{C}\right)^n + p_M \sum_{n=0}^{\infty} n \left(\frac{\lambda\theta}{C}\right)^n.$$

Получили выражение (2.24).

Утверждение доказано.

Приведем также альтернативную формулу записи показателя N^{avg}.

Утверждение 2.4. Среднее число N^{avg} заявок в СМО (2.2) имеет вид

$$N^{\text{avg}} = \sum_{n=1}^{M-1} n\,p_n + Cp_M \left(\frac{M}{C-\lambda\theta} + \frac{\lambda\theta}{(C-\lambda\theta)^2}\right). \tag{2.25}$$

Доказательство.

Выполним предварительные преобразования членов ряда:

$$\sum_{n=0}^{\infty} n \left(\frac{\lambda\theta}{C}\right)^n = \frac{\lambda\theta}{C} \sum_{n=0}^{\infty} n \left(\frac{\lambda\theta}{C}\right)^{n-1} = \frac{\lambda\theta}{C} \left(\sum_{n=0}^{\infty} \left(\frac{\lambda\theta}{C}\right)^n\right)'_{\frac{\lambda\theta}{C}} =$$

$$= \frac{\lambda\theta}{C} \left(\frac{1}{1-\frac{\lambda\theta}{C}}\right)'_{\frac{\lambda\theta}{C}} = \frac{\frac{\lambda\theta}{C}}{\left(1-\frac{\lambda\theta}{C}\right)^2}. \tag{2.26}$$

По определению математического ожидания дискретной с.в.

$$N^{\text{avg}} = \sum_{n=1}^{M-1} n\,p_n + Mp_M\left(\frac{C}{C-\lambda\theta}\right) + p_M\frac{\lambda\theta C}{(C-\lambda\theta)^2} =$$

$$= \sum_{n=1}^{M-1} n\,p_n + Mp_M\left(\frac{C}{C-\lambda\theta}\right) + p_M\frac{\lambda\theta C}{(C-\lambda\theta)^2} =$$

$$= \sum_{n=1}^{M-1} n\,p_n + Cp_M\left(\frac{M}{C-\lambda\theta} + \frac{\lambda\theta}{(C-\lambda\theta)^2}\right).$$

Получили выражение (2.25).

Утверждение доказано.

<u>Утверждение 2.5.</u> Средняя скорость R^{avg} обслуживания заявки имеет вид

$$R^{\text{avg}} = R^{\max}\sum_{n=1}^{M} p_n - Cp_M\left(\frac{C}{\lambda\theta}\right)^M\left[\ln\left(1-\frac{\lambda\theta}{C}\right) + \sum_{n=1}^{M}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^n\right]. \qquad (2.27)$$

Доказательство.

По определению математического ожидания дискретной с.в.

$$R^{\text{avg}} = R^{\max}\sum_{n=1}^{M} p_n + \sum_{n=M+1}^{\infty}\frac{C}{n}p_n = R^{\max}\sum_{n=1}^{M} p_n + C\sum_{n=M+1}^{\infty}\frac{p_n}{n} =$$

$$= R^{\max}\sum_{n=1}^{M} p_n + Cp_M\sum_{n=M+1}^{\infty}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^{n-M} =$$

$$= R^{\max}\sum_{n=1}^{M} p_n + Cp_M\left(\sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^n - \sum_{n=1}^{M}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^n\right) = \qquad (2.28)$$

$$= R^{\max}\sum_{n=1}^{M} p_n + Cp_M\left(\frac{C}{\lambda\theta}\right)^M\left(\sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^n - \sum_{n=1}^{M}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^n\right).$$

Выполним предварительные преобразования с членами ряда:

$$\sum_{n=1}^{\infty}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^n = \int\sum_{n=1}^{\infty}\left(\frac{\lambda\theta}{C}\right)^n d\left(\frac{\lambda\theta}{C}\right) =$$

$$= \int\frac{1}{1-\frac{\lambda\theta}{C}}d\left(\frac{\lambda\theta}{C}\right) = \left|\begin{matrix}u = 1-x\\ du = -dx\end{matrix}\right| = -\int\frac{1}{u}du = \qquad (2.29)$$

$$= -\ln u + const = -\ln\left(1 - \frac{\lambda\theta}{C}\right) + const = |const = 0| =$$

$$= -\ln\left(1 - \frac{\lambda\theta}{C}\right).$$

Подставим (2.29) в формулу для математического ожидания:

$$R^{\text{avg}} = R^{\max}\sum_{n=1}^{M} p_n + Cp_M\left(\frac{C}{\lambda\theta}\right)^M\left[-\ln\left(1 - \frac{\lambda\theta}{C}\right) - \sum_{n=1}^{M}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^n\right] =$$

$$= R^{\max}\sum_{n=1}^{M} p_n - Cp_M\left(\frac{C}{\lambda\theta}\right)^M\left[\ln\left(1 - \frac{\lambda\theta}{C}\right) + \sum_{n=1}^{M}\frac{1}{n}\left(\frac{\lambda\theta}{C}\right)^n\right].$$

Получили выражение (2.27).

Утверждение доказано.

Утверждение 2.6. Вероятность P^{deg} деградации обслуживания в СМО $\underset{\lambda}{M}\left|\underset{\theta}{M}\right|C\left|0,\ R^{max}\right.$ с дисциплиной разделения процессора PS может быть вычислена по формуле

$$P^{\text{deg}} = 1 - \sum_{N=0}^{N^{\text{cont}}} p_N, \tag{2.30}$$

где стационарное распределение $\{p_N, N \geq 0\}$ вероятностей состояний МП числа заявок в СМО над пространством состояний над пространством состояний $\mathcal{X} = \{0, 1, \ldots, \infty\}$ имеет вид (2.8) – (2.20) с условием существования стационарного режима (2.11).

Доказательство.

По предположению при построении модели абонент слайса получает услугу с надлежащим качеством, если его скорость обслуживания не ниже минимальной допустимой скорости обслуживания R^{deg}, $0 < R^{\text{deg}} \leq C$. Используем подход (1.7) для «контрактного порога» при определении максимального числа N^{cont} абонентов с надлежащим качеством обслуживания в слайсе:

$$N^{\text{cont}} = \left\lfloor \frac{C}{R^{\text{deg}}} \right\rfloor. \tag{2.31}$$

Поскольку ресурс слайса делится между абонентами в соответствии с дисциплиной разделения процессора PS, то в состояниях $N > N^{\mathrm{cont}}$ все абоненты одновременно испытывают деградацию обслуживания. Вероятность деградации обслуживания P^{deg} для СМО (2.2), согласно формуле полной вероятности, определяется как вероятность того, что число заявок превысит порог N^{cont}. В этом случае

$$P^{\mathrm{deg}} = 1 - \sum_{N=0}^{N^{\mathrm{cont}}} p_N.$$

Утверждение доказано.

В разделе 2.2 показатели эффективности СМО, включая вероятность деградации обслуживания (2.30), численно исследованы также с помощью имитационной модели для нескольких методов вызова процедуры нарезки ресурса.

2.2. Численные результаты моделирования и их анализ

Численный эксперимент проведен с использованием математической модели раздела 2.1 и имитационной модели [62], разработанной в среде OMNeT++ с использованием библиотеки *queueinglib* и расширенной для поддержки концепции сетевой нарезки. Модель построена по иерархическому принципу, где основными компонентами являются модули, соединения (каналы) и параметры. Для реализации алгоритмов оптимизации и работы с матричными операциями в среду моделирования интегрирована библиотека *Boost*. Взаимодействие компонентов основано на обмене сообщениями и системе сигналов (*@signal*), что позволяет собирать статистику (*@statistic*) без нарушения инкапсуляции модулей.

Для оценки эффективности предложенных методов вызова процедуры нарезки ресурса используются следующие метрики: коэффициент использования ресурса $UTIL$ по формуле (2.22), среднее число заявок в слайсе N_s^{avg} по формуле (2.24), средняя скорость R_s^{avg} по формуле (2.27), вероятность деградации P_s^{deg} по формуле (2.30), и статистика частоты вызовов алгоритма нарезки ресурса v^{alg}, собранная при имитационном моделировании.

Анализ проводился для системы из $S = 5$ слайсов (табл. 2.1), предоставляющих услуги с двумя типами трафика – буферизованное потоковое видео и скачивание файла.

Исследованы два сценария:

– сценарий I (С-I) – высокая нагрузка на потоковое видео;

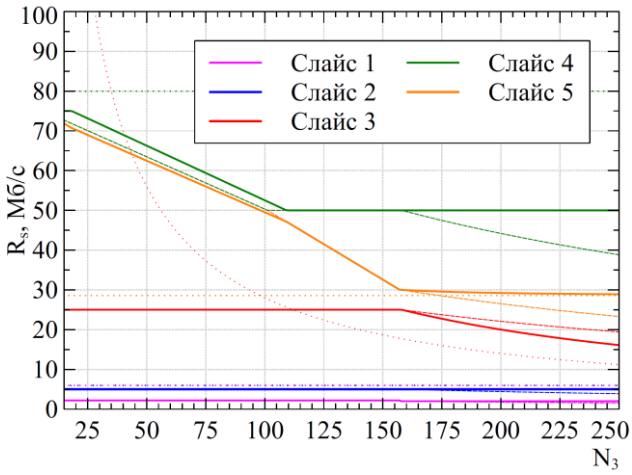
– сценарий II (С-II) – наличие приоритетного слайса и овербукинга.

Табл. 2.1. Исходные данные

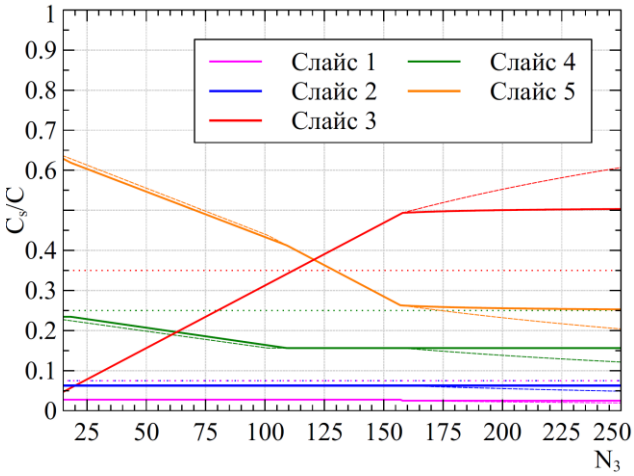
C, Гбит/с	8				
S	5				
Номер слайса	1	2	3	4	5
Тип трафика	Буферизованное потоковое видео				Скачивание файла
Описание трафика	SD	HD	UHD (4K)	VR (8K)	Обновление ПО
Параметры схемы нарезки					
$R_s^{\min}$, Мбит/с	2	5	25	50	30
$R_s^{\max}$, Мбит/с	2,2	8	30	75	C
q_s (С-I)	0,075	0,075	0,35	0,25	0,25
q_s (С-II)	**1**	0,075	0,35	0,25	0,25
v_s (С-I)	1	1	1	1	1
v_s (С-II)	1	**2**	**2**	**2**	**3**
Нагрузочные параметры для стохастического анализа					
Сред. время между вызовами, с	1,65	7,25	16	19	5
Средний размер файла, ГБ	0,3	1,2	2,5	5	1

На рис. 2.3 показана зависимость от состояния системы средней скорости в слайсе R_s^{avg} по (2.27); доли емкости слайса $\frac{C_s}{C}$ по (1.21), (1.23) для С-I без приоритизации и по (1.24), (1.25) для С-II с приоритизацией; коэффициента овербукинга слайса $\frac{C_s}{C_s^{\text{cont}}}$ по (1.21), (1.23), (1.8) для С-I без приоритизации и по (1.24), (1.25), (1.8) для С-II с приоритизацией. При превышении суммарного спроса над емкостью ($N_3 > 158$ для С-I) система переходит в область перегрузки Ω_{cong}. На рис. 2.3 приведены также результаты для сравнения гибкой схемы нарезки сети (сплошная линия) с классическими схемами совместного использования – жестким резервированием в схеме с полным разделением ресурсов Complete Partitioning (точечная линия) и мягкой полнодоступной схемой Complete Sharing (пунктирная

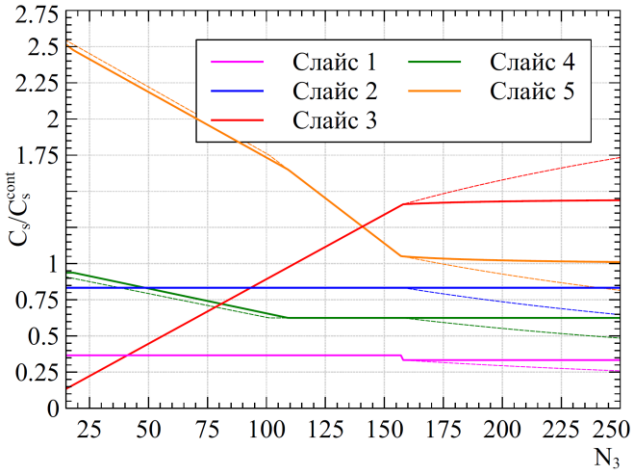
линия), которые позволяют продемонстрировать преимущества схемы нарезки сети. Наблюдается эффект изоляции производительности: несмотря на деградацию обслуживания в слайсах 3 и 5, показатели приоритетных слайсов (1, 2, 4) остаются стабильными, что подтверждает преимущество динамической нарезки перед полным совместным использованием ресурсов.



а) Скорость передачи данных



б) Доли емкости слайса



в) Перегруженность слайса

Линии: сплошная – нарезка сети, пунктирная – совместное использование, точечная – полное разделение

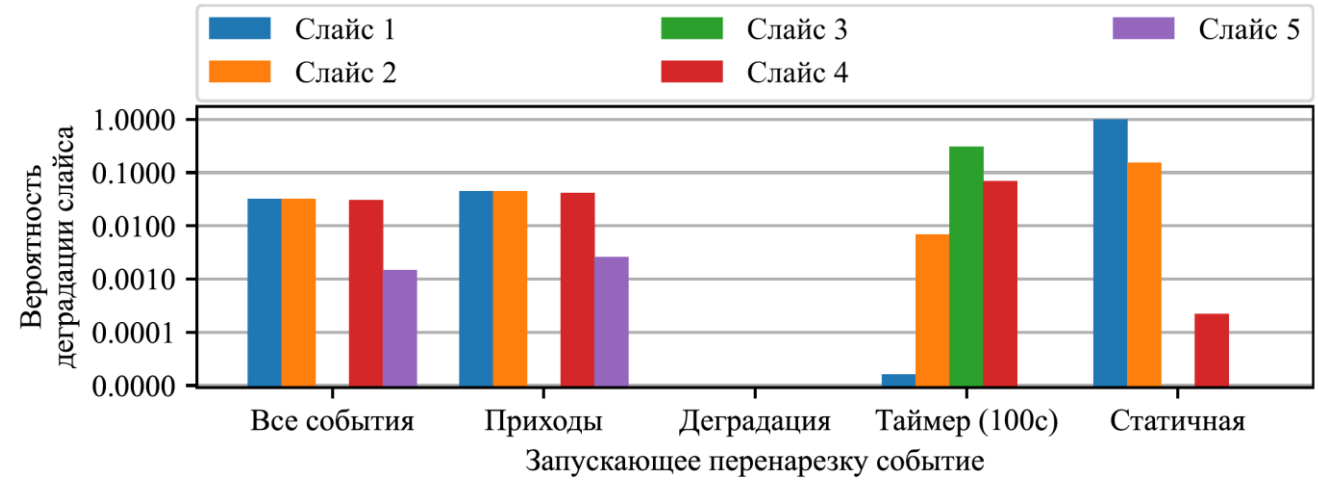
Рис. 2.3. Показатели слайсов в состояниях $\mathbf{N} = (100{,}100, N_3, 25{,}70)$ к числу N_3 абонентов в Слайсе 3 для С-I

На рис. 2.4 сопоставляется вероятность деградации обслуживания (2.30) для нескольких вариантов инициирования процедуры ресурсного слайсинга:

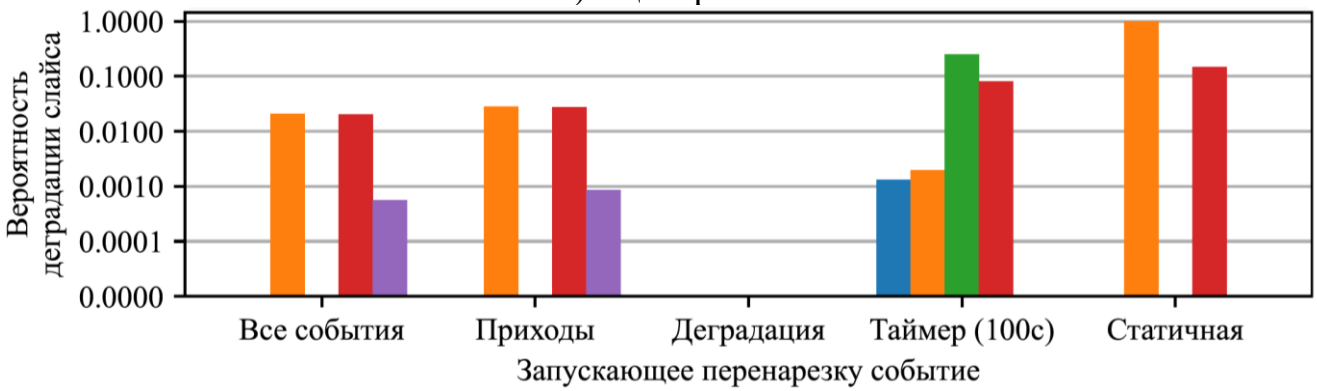
– по событиям (поступление заявки, окончание обслуживания заявки, деградация, таймер);

– по приходам (при поступлении новой заявки);
– при деградации обслуживания (при падении скорости получения услуги абонентами слайса ниже порога $R_s^{\deg}$);
– по таймеру;
– статичная (первоначальная нарезка ресурса не меняется).

Показано, что вызов процедуры нарезки ресурса по событию деградации значительно снижает вероятность деградации для всех слайсов, часто до нуля для критически важных.



а) Сценарий I



б) Сценарий II

Рис. 2.4. Вероятность деградации $P_s^{\deg}$ для разных методов вызова процедуры нарезки ресурса

Рис. 2.5 демонстрирует, что вызов процедуры нарезки ресурса по всем событиям или по приходам обеспечивает максимальное использование ресурсов. Однако вызов процедуры нарезки ресурса по деградации, при более низком

коэффициенте $UTIL$, достигает сравнимой эффективности в предотвращении деградации.

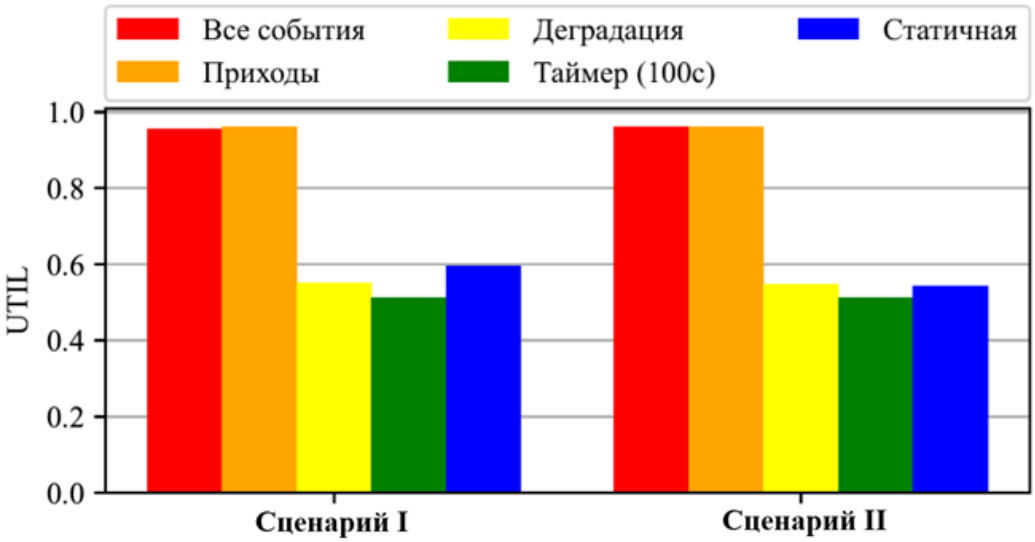


Рис. 2.5. Коэффициент использования ресурса $UTIL$ для различных методов вызова процедуры нарезки ресурса

Рис. 2.6 показывает, что вызов процедуры нарезки ресурса по деградации происходит значительно реже (на два порядка), чем по всем событиям или приходам, что указывает на ее вычислительную экономичность.

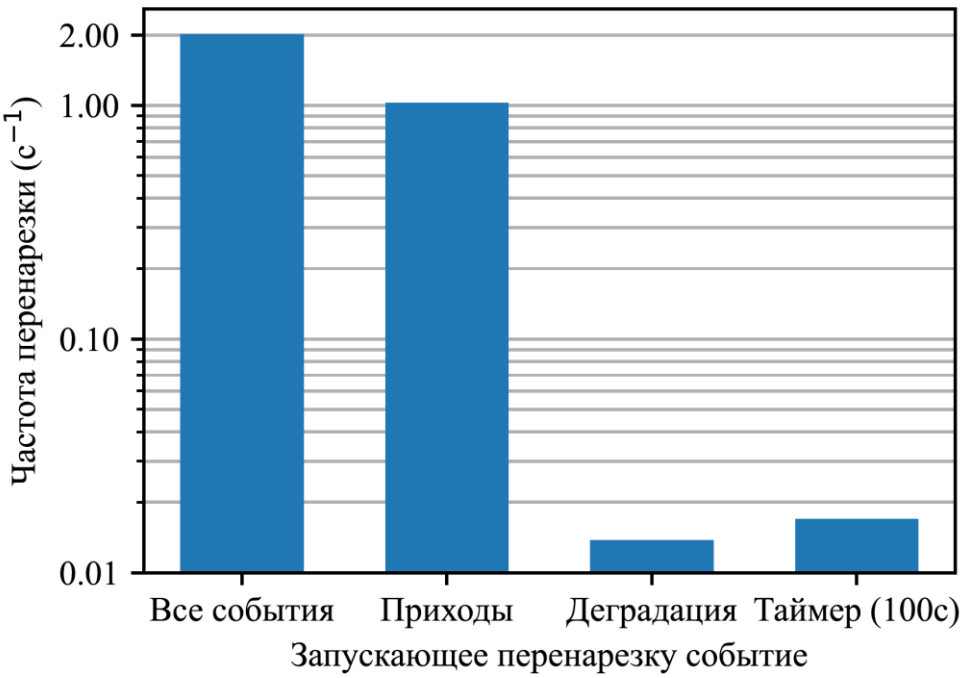
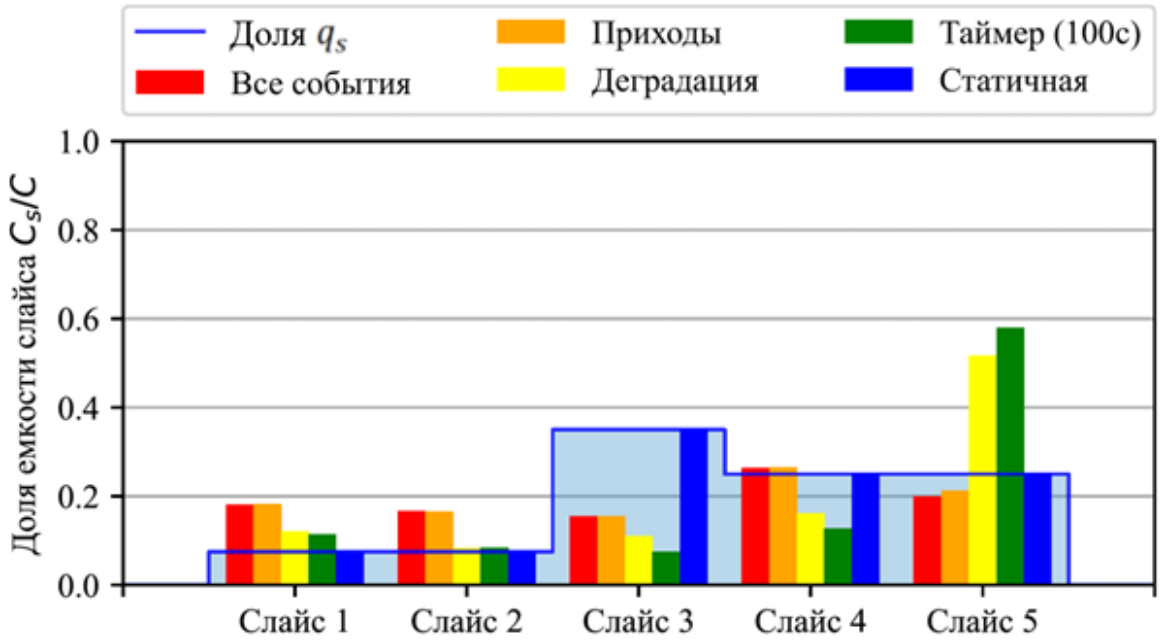
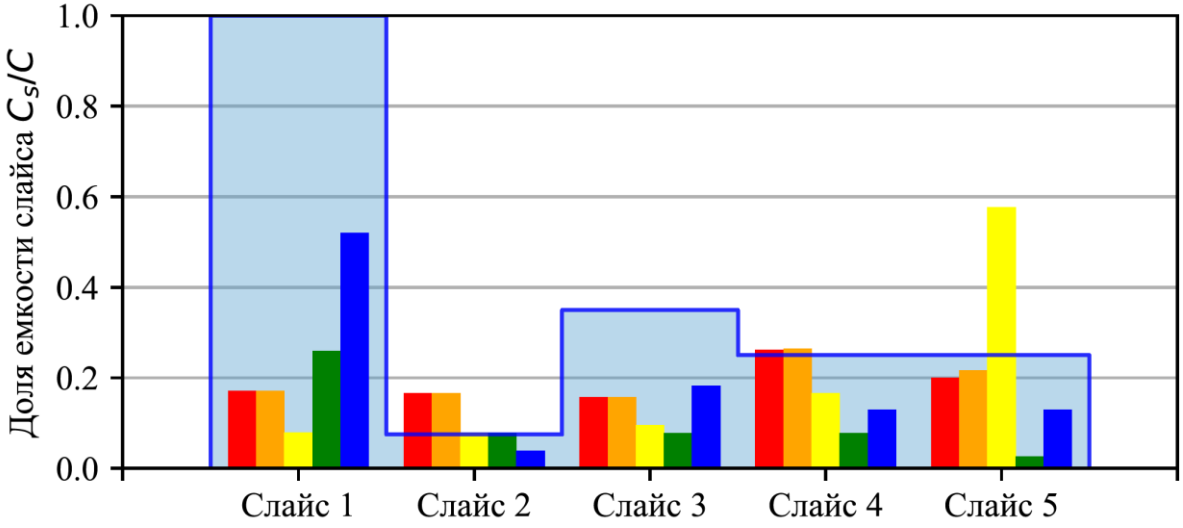


Рис. 2.6. Частота исполнения алгоритма нарезки ресурса для различных методов вызова процедуры нарезки ресурса

На рис. 2.7 сравнивается средняя доля выделенной емкости $\frac{C_s}{C}$ по (1.21), (1.23) для C-I без приоритизации и по (1.24), (1.25) для C-II с приоритизацией с гарантированными долями γ_s по (1.8). В условиях избыточного резервирования (овербукинга) частый вызов процедуры нарезки ресурса (по всем событиям/приходам) может выделять слайсам больше ресурсов, чем гарантировано, тогда как вызов процедуры нарезки ресурса по деградации более консервативен, но при этом более эффективен.



а) Сценарий I



б) Сценарий II

Рис. 2.7. Средняя доля емкости слайса $\frac{C_s}{C}$ по сравнению с q_s для различных методов вызова процедуры нарезки ресурса

Основные выводы по результатам моделирования.

1. Вероятность деградации (рис. 2.4): статичная нарезка не справляется с флуктуациями трафика. Метод вызова процедуры нарезки ресурса «по событиям» минимизирует деградацию, однако метод вызова процедуры нарезки ресурса «по деградации» показывает сопоставимые результаты при меньших вычислительных затратах.

2. Эффективность ресурсов (рис. 2.5, рис. 2.6): вызов алгоритма по всем событиям обеспечивает максимальное значение коэффициента использования ресурса $UTIL$, но требует высокой частоты перерасчета (до 2 раз в с). Вызов процедуры нарезки ресурса по деградации происходит на два порядка реже, сохраняя стабильность системы.

3. Динамика емкости (рис. 2.7): средние доли выделяемого ресурса в динамических сценариях соответствуют теоретическим весам q_s, но позволяют временно использовать излишки ресурсов других слайсов («овербукинг»).

Таким образом, показано, что мгновенный отклик на любое событие не является обязательным: вызов процедуры нарезки ресурса по факту снижения качества обслуживания (деградации) обеспечивает необходимый баланс между QoS и вычислительной сложностью управления сетью.

ГЛАВА 3
МОДЕЛЬ ДЛЯ АНАЛИЗА ЗАДЕРЖКИ В МНОГОШАГОВОЙ СЕТИ

3.1. Задача анализа показателей качества сети IAB

В сетях пятого поколения (5G) требования к пропускной способности и задержке значительно ужесточились по сравнению с действующими системами предыдущих поколений [63-65]. Одновременно диапазоны микроволновых частот, уже задействованные в мобильной связи, практически исчерпаны и не способны покрыть растущий пользовательский спрос [66,67]. Новые возможности обеспечения высокого уровня QoS открыла стандартизированная консорциумом «Проект партнерства третьего поколения» (3rd Generation Partnership Project, 3GPP) технология «Новое радио» (New Radio, NR) в mmWave на частотах 24–100 ГГц.

Вместе с тем использование mmWave осложняется требованием прямой видимости (Line-of-Sight, LoS) между передатчиком и приемником, а также повышенными потерями при распространении сигнала, включая атмосферное поглощение. Это приводит к необходимости более плотного размещения БС 5G [68] и, как следствие, к росту затрат на развертывание. Для смягчения этих ограничений 3GPP стандартизировала технологию IAB, предполагающую построение многошаговой беспроводной транспортной сети: лишь часть БС (IAB-доноры) имеет проводной доступ к опорной сети, тогда как остальные (IAB-ретрансляторы) подключаются к донорам по воздуху напрямую или через несколько промежуточных узлов (рис. 3.1).

Сеть интегрированного доступа и транзита IAB на базе mmWave 5G NR обязана состоять из одного IAB-донора (далее – «донор»), нескольких IAB-ретрансляторов (далее – «ретрансляторов»), при этом с каждым из вышеназванных узлов сети IAB, выполняющих функции БС (IAB-узлы – донор и ретрансляторы), связаны несколько АУ. Узлы сети IAB на расстоянии одного шага по нисходящей линии связи от некоторого узла (донора или ретранслятора) называются его «дочерними» узлами (их ноль или более), а одного шага от узла-ретранслятора по

восходящей линии связи – «родительскими» узлами. Два типа соединений (доступ и транзит) в сети IAB показаны на рис. 3.1: каналы доступа (access) соединяют АУ с ретранслятором или донором, каналы транзита (backhaul, BH) соединяют два ретранслятора или узел и донор. Кроме того, на рис. 3.1 показаны два направления передачи – нисходящее (downlink) и восходящее (uplink).

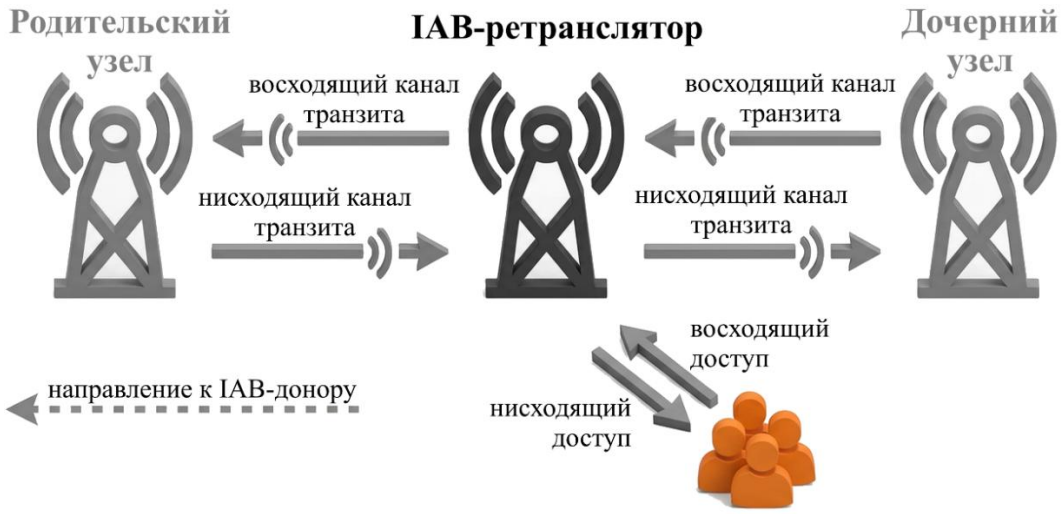


Рис. 3.1. Типы соединений IAB

В такой конфигурации сквозная задержка пакетов может значительно увеличиться из-за многошаговой передачи. Проблема усугубляется требованиями к полудуплексной передаче данных в радиоканале, когда узлы и доноры не могут одновременно передавать и принимать информацию по нескольким имеющимся у них антеннам.

Отличительные особенности технологии IAB

Технология IAB предназначена для повышения эффективности и рентабельности транзита в сетях стандарта долговременного развития (Long-Term Evolution, LTE) и 5G NR, в то время как микро- и фемтосоты, впервые появившиеся в LTE, используются для обеспечения улучшенного покрытия и пропускной способности в районах со слабыми сетевыми сигналами или вообще без них. Технология IAB объединяет функции доступа и транзита в единый сетевой узел, что сокращает число сетевых элементов и упрощает сетевую архитектуру. Это может привести к снижению затрат, повышению эффективности сети и более быстрому развертыванию по сравнению с традиционными сетями обратной передачи.

Модуль транзита может использовать протоколы маршрутизации, такие как протокол маршрутизации по состоянию канала (Open Shortest Path First, OSPF), протокол граничного шлюза (Border Gateway Protocol, BGP) и протокол информации маршрутизации (Routing Information Protocol, RIP), для маршрутизации пакетов между IAB-узлами и интернетом по проводной сети и беспроводной сети. Эти протоколы являются протоколами сетевого уровня в модели OSI, а значит не зависят от базового физического соединения, используемого для передачи данных. Независимо от того, передаются ли данные по проводному соединению, такому как Ethernet, или по беспроводному соединению, такому как Wi-Fi, протоколы маршрутизации работают одинаково, определяют наилучший путь для передачи пакетов на основе текущего состояния сети. Эти технологии и протоколы позволяют сети IAB эффективно управлять полудуплексным соединением, разделять трафик и обеспечивать беспроводной транзит, эффективно используя доступный спектр и уменьшая помехи за счет использования протоколов управления радиоресурсами (Radio Resource Control, RRC) и технологии с несколькими антеннами для одновременной передачи и приема данных (Multiple Input Multiple Output, MIMO), что особенно важно для услуг Интернета вещей [69].

Преимущества полудуплексного соединения

Основным преимуществом использования в технологии IAB полудуплексного режима передачи данных в радиоканале является масштабируемость. Ресурс – полоса частот одного радиоканала – доступен нескольким БС на основе мультиплексирования с временным разделением TDM, что означает, что каждой БС назначается определенный временной интервал для передачи данных по радиоканалу. Это позволяет нескольким БС совместно использовать один и тот же радиоканал как канал транзита или канал доступа, не создавая помех друг другу, что позволяет развернуть больше БС в пределах одной сети. По мере увеличения числа БС в сети полудуплексное соединение можно легко масштабировать, добавляя дополнительные временные интервалы для размещения большего числа БС. Напротив, для полнодуплексного соединения требуется

выделенная полоса пропускания для каждой линии связи, что означает, что требования к полосе пропускания сети увеличиваются с каждой дополнительной БС. Это затрудняет масштабирование сети по мере увеличения числа БС, поскольку доступная полоса пропускания становится ограничивающим фактором.

В дополнение к масштабируемости полудуплексное соединение проще и дешевле в реализации по сравнению с полнодуплексным. Это позволяет внедрять технологию IAB более эффективно и с меньшими затратами, что делает ее более привлекательным решением для многих операторов. Полудуплексное соединение также менее подвержено помехам и более эффективно с точки зрения использования полосы пропускания, поскольку не требует выделенной полосы пропускания для каждой линии связи. Это делает технологию IAB более надежной, стабильной и эффективной с точки зрения использования доступной полосы пропускания, снижая затраты на развертывание и эксплуатацию сети.

Особенности многошаговой беспроводной передачи

Сети IAB – первая архитектура коммерческих сотовых сетей, использующая многошаговую беспроводную передачу. Проблемы, которые необходимо исследовать для удовлетворения сети IAB требованиям к качеству обслуживания, установленным стандартами ITU-R и 3GPP, определены в 3GPP TR 38.874 [70]. В части управления топологией основное внимание уделяется разработке сетевых протоколов и архитектуры, а также описанию процедур управления трафиком. В части выбора маршрута для многошагового доступа классическая проблема маршрутизации трафика и распределения ресурсов усложняется полудуплексным ограничением и необходимостью учета помех из-за беспроводной транспортной сети. В дополнение к традиционным задачам топологии сети и выбора маршрута актуальным становится исследование влияния числа шагов маршрута на сквозную задержку, которая также является одним из показателей качества предоставления услуг. Под сквозной (end-to-end) задержкой понимается время прохождения информации от одного конца системы до другого через все промежуточные узлы и сети (включая задержки обработки, очереди, передачи и распространения). Вторая временнАя метрика, недавно привлекшая внимание исследователей – так

называемая «возраст информации» (AoI), которая предложена для оптимизации услуг мониторинга состояния удаленных систем и может рассматриваться как метрика QoS. Концептуально метрика AoI является явной функцией времени и представляет собой сумму интервала времени между соседними моментами генерации пакетов в источнике и задержки при передаче пакета по сети и предполагает, что только своевременно полученные обновления могут отражать текущее состояние удаленной системы. Пиковый возраст информации (Peak AoI, PAoI) представляет собой значение возраста информации в момент непосредственно перед получением системой мониторинга очередного пакета от удаленной системы [71]. Метрики возраста информации и пикового возраста информации востребованы и, например, при мониторинге промышленных процессов, который осуществляется путем обмена статусом рабочего оборудования, при синхронизации между роботами, выполняющими совместные действия на производственной линии, когда актуальность информации ограничена по времени, в энергосистемах, в том числе на атомных электростанциях, где мониторинг показателей функционирования критически важен для безопасности, или в системах «умного города», где есть потребность в своевременной доставке информации в центры управления и мониторинга.

Как видно из табл. 3.1, проблемы технологии IAB рассматривались в литературе; однако способ решения любой из них во многом зависит от того, предполагается ли дуплексная или полнодуплексная передача. Табл. 3.1 систематизирует результаты анализа публикаций, посвященных актуальным вопросам организации связи и управления ресурсами в современных и будущих сетях, таких как 5G/6G, IAB, БПЛА-сети и другие.

Табл. 3.1. Обзор проблем, связанных с IAB, в научной литературе

Проблема	**Подпроблема**	**Статья**	**Предпола-гаемый режим передачи**
1. Оптимизация планирования ресурсов и	Динамическое планирование каналов с учетом требований абонентов	[5]	Дуплекс
		[6]	Дуплекс / Полудуплекс

управления трафика для минимизации задержек	Выбор пути с гарантией задержки	[7]	Дуплекс
	Комплексное планирование ресурсов	[8]	Дуплекс
2. Координация помех и повышение спектральной эффективности	Подавление помех, управление диаграммами направленности	[9]	Дуплекс
		[10]	Дуплекс
	Подавление помех в полнодуплексной связи	[11]	Дуплекс
		[12]	Дуплекс
3. Позиционирование, развертывание и управление мобильными узлами (UAV/BAP)	Оптимизация развертывания БПЛА	[72]	Дуплекс
		[73]	Дуплекс
	Интеграция IAB в воздушных низковысотных платформах	[74]	Дуплекс
	Управление и планирование ресурсов для БПЛА-систем	[75]	Дуплекс
4. Эффективность и управление энергопотреблением	Энергоэффективная оптимизация воздушных IAB-сетей	[76]	Дуплекс
	Энергоэффективная оптимизация гетерогенных сетей IAB	[77]	Дуплекс
5. Интеграция с новыми технологиями и архитектурами (NTN, RIS, THz/FSO)	IAB-сети с отражающими поверхностями	[78]	Дуплекс
		[13]	Дуплекс (возможно, TDD/FDD)
	Многошаговые и меш-сети, интеграция наземных и спутниковых сетей	[14]	Дуплекс / Полудуплекс
		[15], [16]	Дуплекс
	Неназемные сети, включая спутники	[17]	Дуплекс
6. Обеспечение надежности и отказоустойчивости	Повышение надежности с использованием mmWave / THz	[13]	Дуплекс (возможно, TDD/FDD)
		[14]	Дуплекс / Полудуплекс
	Риск-ориентированное обучение для надежной самотранзитной связи	[18]	Дуплекс
7. Маршрутизация, ассоциация абонентов и выбор пути	Оптимальный выбор пути с улучшением QoS, интеллектуальный выбор транзитного канала	[19], [20], [21]	Дуплекс

	Агенты пересылки, взаимодействие абонентов	[22], [23]	Дуплекс
8. Теоретический анализ и моделирование производительности	Анализ производительности полнодуплексных многоячеистых IAB-сетей	[12]	Дуплекс
	Анализ схем дуплексирования, моделирование распространения в THz сетях	[24]	Дуплекс / Полудуплекс
		[25]	Дуплекс / Полудуплекс
9. Применение в специфических сценариях и сервисах	Механизм выгрузки задач в сетях с поддержкой БПЛА.	[79]	Дуплекс
	IAB-сети для экстренных коммуникаций.	[76]	Дуплекс
		[80]	Дуплекс

Анализ современной научно-технической литературы (табл. 3.1) показал, что большая часть исследований в области IAB сосредоточена на вопросах пропускной способности и надежности, в меньшей степени – на оптимизации сквозной задержки, при этом автору не известны работы, где ставилась и решалась задача оптимизации возраста и пикового возраста информации в сети IAB.

В диссертационной работе исследовано решение проблем, связанных с разделением ресурсов для минимизации задержки, которая критически важна для многих современных приложений и является составной частью возраста информации. Проведен теоретический анализ и моделирование процесса передачи пакетов в сети IAB в терминах CeMO, направленные на разработку подходов, учитывающих особенности как полнодуплексных, так и полудуплексных режимов, с целью достижения оптимальной производительности многошаговой сети.

Системная модель сети IAB

Рассмотрена сеть IAB, работающая в полудуплексном режиме передачи данных в радиоканале, при этом полоса пропускания $B_{[\text{Гц}]}$ используется как в каналах доступа, так и в каналах транзита. Согласно предположениям [81], каждая из физических антенн IAB-узла может одновременно либо передавать, либо принимать данные, однако одновременная передача и прием невозможны из-за интерференции (помех). Для системы IAB в миллиметровом диапазоне пропускная

способность каналов доступа и каналов транзита может быть вычислена с помощью закона Шеннона:

$$C_{[\text{бит/с}]} = B \log_2 \left(1 + \frac{P_T G_T G_R}{PL(d) N_0 B_{PRB} M_{SF} L}\right), \quad (3.1)$$

где P_T – излучаемая мощность, G_T и G_R – коэффициенты усиления антенн передатчика и приемника соответственно, N_0 – спектральная плотность мощности теплового шума, B_{PRB} – размер физического ресурсного блока (Physical Resource Block, PRB), $M_{SF_{[\text{дБ}]}} \sim Norm\,(0, \sigma_{SFLoS/SFnLoS})$ – медленное замирание в условиях наличия/отсутствия прямой видимости, а $L_{[dB]}$ – совокупные потери, учитывающие запас помехоустойчивости.

Узлы сети IAB перенумерованы: номер 1 для донора и номера от 2 до N для $(N-1)$ ретранслятора. Сетевой IAB-узел n имеет N_n секторов антенны, пронумерованных от 1 до N_n, $n = 1, \ldots, N$. Каждый сектор антенны обеспечивает канал доступа по восходящей и нисходящей линиям связи с АУ и может поддерживать один беспроводной канал транзита с другим сетевым узлом. Топология сети предполагается фиксированной и древовидной, другими словами, между донором и каждым ретранслятором существует один маршрут. Такая топология может быть представлена матрицей $\mathbf{M} = \left(M_{i,j}\right)_{i,j=1,\ldots,N}$, где элемент $M_{i,j} \in \{0, 1, \ldots, N_i\}$ – либо номер сектора антенны в узле i, через который узел i связывается с узлом $j \neq i$, либо 0, если между этими узлами нет связи. Заметим, что если $M_{i,j} = 0$, то $M_{j,i} = 0$, и наоборот, если $M_{i,j} > 0$, то $M_{j,i} > 0$. Положим $M_{i,i} = 0$ для всех $i = 1, \ldots, N$.

Пример сети с $N = 3$, $N_1 = 4$ и $N_i = 3$ для $i = 2, 3$ изображен на рис. 3.2.

Будем считать, что IAB-узел относится к ярусу (tier) k, если кратчайшее число шагов от узла 1 (донора) до этого узла равняется $k - 1$. Отсюда следует, что узел 1 (донор) – единственный элемент первого яруса. На рис. 3.2 узел 2 принадлежит ярусу 2: он связан с донором и с узлом 3 через свои секторы 1 и 2 соответственно. Транспортная топология той же сети задается матрицей

$$\mathbf{M} = \begin{vmatrix} 0 & 3 & 0 \\ 1 & 0 & 2 \\ 0 & 2 & 0 \end{vmatrix},$$

где $M(i,j)$ указывает номер сектора узла i, используемый для связи с узлом j (0 означает отсутствие прямого соединения). Каждый сектор любого узла может обслуживать прикрепленные к нему АУ. Передача пакетов осуществляется между АУ и опорной транспортной сетью, к которой подключен донор. Предполагается, что сама передача пакета по активному каналу происходит мгновенно (временные задержки передачи пренебрежимо малы относительно масштабов моделирования), однако из-за невозможности одновременной активации всех каналов требуется буферизация пакетов и ожидание доступности соответствующего канала.

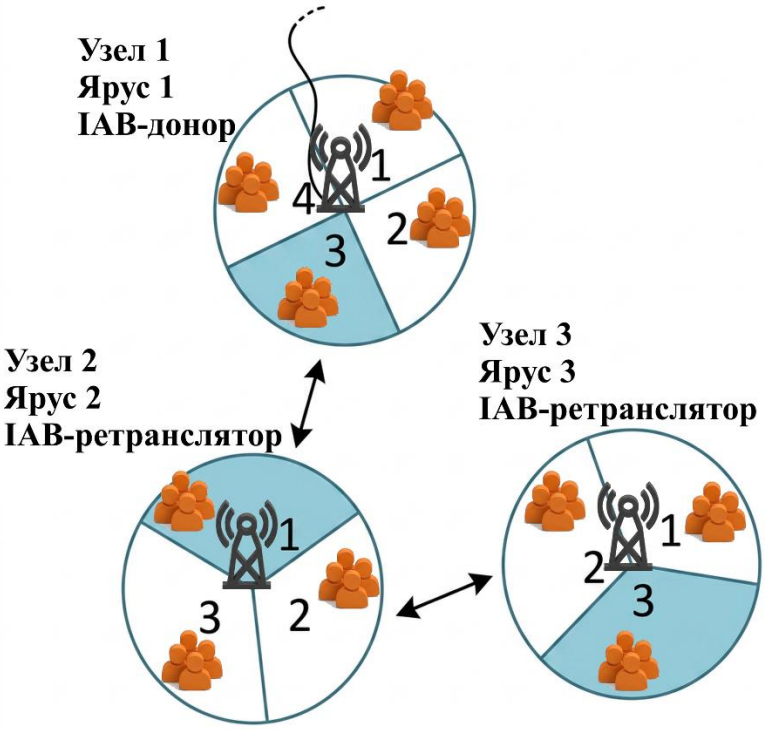


Рис. 3.2. Пример сети IAB

Как показано на рис. 3.3, каждый сектор каждого узла имеет один канал транзита с буфером ВН и один канал доступа с двумя буферами доступа – исходящим буфером доступа (для нисходящей линии связи к АУ, AC_{DL}) и входящим буфером доступа (для восходящей линии от АУ, AC_{UL}). Буфер транзита сектора узла содержит пакеты, ожидающие передачи по каналу транзита, поддерживаемому секторной антенной. Входящий буфер доступа сектора узла содержит полученные от родительского узла пакеты, предназначенные для

передачи по нисходящей линии к АУ, связанным с сектором. Исходящий буфер доступа сектора узла хранит пакеты от всех АУ, связанных с сектором, ожидающие передачи по восходящей линии к родительскому узлу. Буфер транзита сектора узла хранит пакеты для передачи по нисходящей или по восходящей линии к связанному узлу на шаг дальше от узла 1 или на шаг ближе к узлу 1.

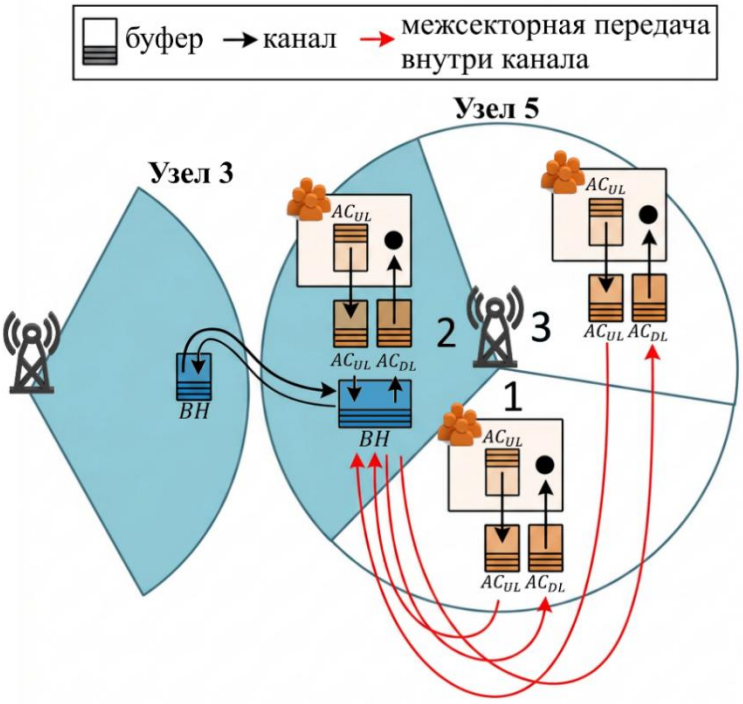


Рис. 3.3. Схема сети массового обслуживания узла сети IAB

Для восходящего направления маршрутизация пакета от АУ, связанного с сектором i узла n (обозначается как (n, i)) происходит следующим образом. Пусть узел n принадлежит ярусу k. При активации канала доступа для восходящей линии в секторе (n, i) пакет от АУ мгновенно достигает узла n. Если $n = 1$ (узел является донором), то пакет сразу покидает систему. Если $n > 1$, то пакет помещается в буфер транзита сектора (n, j), обращенного к ярусу $k - 1$, т. е. сектора, связывающего узел n с узлом, принадлежащим ярусу $k - 1$. Чтобы пакет переместился на шаг вверх в направлении донора, должен быть активирован канал транзита между узлом n и узлом m, для которого $M_{n,m} = j$. Как только это произойдет, пакет мгновенно достигает узла m. Затем снова, если $m = 1$, пакет сразу покидает систему, а если $m > 1$, то пакет присоединяется к буферу транзита в секторе, обращенном к ярусу $k - 2$. Передача продолжается до тех пор, пока

пакет не достигнет узла 1 (донора), после чего пакет уходит в опорную транспортную сеть по проводному подключению и покидает систему. Под сквозной задержкой пакета в восходящем направлении будем понимать интервал от момента его поступления во входящий буфер доступа сектора соответствующего АУ до момента выхода из транзитного буфера узла второго яруса в направлении узла 1. Поскольку предполагается мгновенная передача пакетов между узлами, при достижении узла 1 пакет немедленно передается по проводному каналу в опорную сеть.

Для нисходящего направления маршрутизация пакета от опорной проводной транспортной сети к АУ, связанному с сектором (n, i) узла n, принадлежащего ярусу k, происходит следующим образом. Если $n = 1$, то пакет входит в систему, присоединяясь к исходящему буферу доступа в секторе $(1, i)$, и уходит из системы, как только активируется соответствующий канал доступа в нисходящем направлении к АУ. Если $n > 1$, то пакет передается внутри узла 1 и сначала присоединяется к буферу транзита сектора $(1, j)$, связанного с соответствующим сектором узла n, если $k = 2$; или с узлом яруса 2, через который может быть достигнут узел n. Как только соответствующий канал транзита активирован, пакет достигает узла m, такого что $M_{1,m} = j$. Если $m = n$, то пакет присоединяется к исходящему буферу доступа в секторе (n, i) узла назначения и ожидает активации канала доступа для нисходящей линии до АУ, после чего покидает систему. Если $m \neq n$, пакет помещается в буфер транзита в секторе, связывающем узел m с узлом n, или в узле уровня 3, через который может быть достигнут узел n, и т. д. Процесс продолжается до тех пор, пока пакет не покинет исходящий буфер доступа в узле n по каналу доступа, активированному для нисходящей линии к АУ. Под сквозной задержкой пакета в нисходящем направлении будем понимать интервал от момента поступления пакета в исходящий буфер сектора доступа на узле 1 до момента его выхода из системы по каналу доступа к АУ.

Поскольку с целью минимизации собственной интерференции и интерференции соседних БС как в полудуплексном, так и в полнодуплексном

режимах радиоканал активен не постоянно, предусмотрен менеджер нарезки ресурса SliM на основе мультиплексирования с временным разделением– планировщик (scheduler) активации радиоканалов, представляющий собой программный компонент или инструмент, предназначенный для управления последовательностью активации и длительностью интервалов активности каналов транзита и каналов доступа, который активирует антенну сектора узла сети IAB для выполнения одного из соответствующих действий:

- «передачи транзита» (передача пакетов на связанный узел по каналу транзита);
- «прием транзита» (получение пакетов от связанного узла по каналу транзита);
- «передача доступа» (передача пакетов на связанные АУ по каналу доступа в нисходящем направлении);
- «прием доступа» (получение пакетов от связанных АУ по каналу доступа в восходящем направлении).

Также с целью ограничения интерференции сделано предположение, что в каждый момент времени все секторы одного и того же узла могут либо передавать, либо получать пакеты. Например, в момент времени $t = t_1$ в узле n Сектор 1 может получать пакеты по каналу транзита, а Секторы 2 и 3 получать пакеты по каналу доступа, в то время как в момент времени $t = t_2$ все секторы могут передавать пакеты по каналам доступа.

Для активации каналов доступа достаточно, чтобы соответствующий сектор перешел в состояния передачи доступа или приема доступа. Таким образом, например, как только сектор (n, i) переходит в состояние приема доступа, все содержимое входящего буфера доступа сектора достигает узла n, и, кроме того, любой пакет, присоединяющийся к этому буферу, когда сектор находится в состоянии приема доступа, также достигает узла n.

Для активации канала транзита требуются скоординированные действия двух связанных секторов на связанных узлах: один должен находиться в состоянии передачи транзита, а другой – приема транзита. Так, на рис. 3.2, канал транзита между узлами 1 и 2 активируется для передачи по нисходящей линии (от узла 1 к

узлу 2) всякий раз, когда Сектор (1, 3) переходит в состояние передачи транзита, а Сектор (2, 1) – приема транзита. В течение этого времени все пакеты из буфера транзита в Секторе (1, 3) перемещаются на узел 2. Точно так же канал транзита между узлами 2 и 3 активируется для передачи по восходящей линии, если Сектор (3, 2) переходит в состояние передачи транзита, а Сектор (2, 2) – приема транзита. Если в заданной топологии сектор не используется для транзита, его допустимыми состояниями являются передача доступа или прием доступа, поскольку нахождение в состояниях передачи транзита или приема транзита не приведет к активации канала.

Далее в разделе 3.1 рассматривается сценарий развертывания сети IAB на нескольких репрезентативных топологиях. Для наборов параметров, приближенных к реальным, проведена оценка сквозной задержки пакетов в сети IAB по двум направлениям:

– восходящее (Uplink, UL): интервал от момента попадания пакета во входящий буфер доступа в его секторе до достижения пакетом узла 1;
– нисходящее (Downlink, DL): интервал от момента помещения пакета в буфер на узле 1, предназначенный для передачи по каналу доступа к АУ, до выхода пакета из системы по нисходящей линии.

Сценарий и параметризация

Развертывание 5G IAB NR ориентировано прежде всего на плотную городскую застройку, где распространение радиосигнала осложняется крупными препятствиями. В таких условиях сигнал преимущественно «ведется» вдоль уличных каньонов и, как правило, не огибает углы кварталов из-за слабого проникновения через стены. В качестве иллюстративного примера для численных экспериментов используется «манхэттенская» планировка – ортогональная сетка «проспектов» и «улиц», типичная для центров мегаполисов. Естественные точки размещения БС – уличные перекрестки, а ключевым параметром сценария выступает плотность размещения БС.

В сценарии использованы кварталы размером 76 × 183 м, при средней ширине проспекта/улицы 30 м. Предполагалось, что радиус зоны покрытия антенн узлов (БС) составляет не более 210 м [85]. Размещение узлов соответствует вариантам топологии, изображенным на рис. 3.4. Пример развертывания Топологии № 2 показан на рис. 3.5, где $N_1 = 4$ и $N_n = 3$ для $n = 2, \ldots, N$.

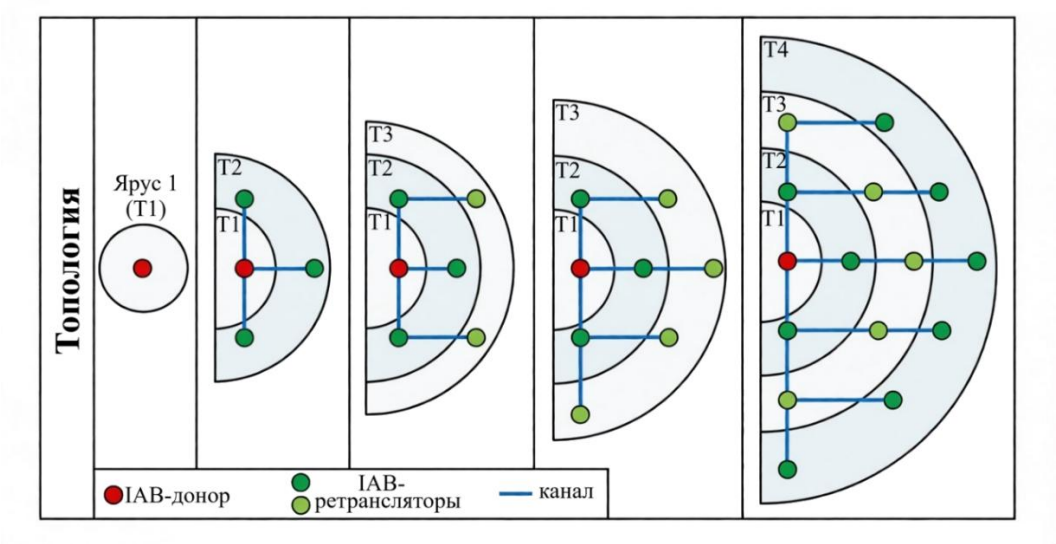


Рис. 3.4. Пять вариантов топологии сети IAB

Пусть АУ равномерно распределены в пространстве $\mathbb{R}^2$ в границах улиц. Поступление на узел пакетов в восходящем направлении от всех АУ, связанных с сектором (n, i), моделируется пуассоновским потоком с параметром $\lambda_{n,j} = L_{n,j}\lambda$, где $L_{n,j}$ – среднее число АУ в секторе (n, i), а λ – средняя интенсивность поступления пакетов от одного АУ (далее – «удельная нагрузка»). Поступление на узел предназначенных для АУ в секторе (n, i) пакетов от родительского узла в нисходящем направлении, представлено в виде пуассоновского процесса с параметром $\alpha_{DL/UL}\lambda_{n,i}$, где коэффициент $\alpha_{DL/UL}$ – отношение интенсивности инициированного АУ трафика в нисходящем и восходящем направлениях («к АУ» и «от АУ»).

Ближайший к донору шаг в многошаговой сети IAB должен обеспечивать пропускную способность, достаточную для агрегированного трафика многошаговых маршрутов от всех узлов IAB. Это «бутылочное горлышко» – одно из основных ограничений масштабирования сети IAB.

Модель сети IAB реализована в среде имитационного моделирования OMNeT++, представляющей собой универсальный дискретно-событийный симулятор с поддержкой базовых сетевых функций и гибкими возможностями расширения. Основными оцениваемыми метриками выступают сквозные задержки доставки пакетов по восходящей и нисходящей линиям, анализируемые в зависимости от параметров системы.

Для оценки интенсивности поступления пакетов 1000 АУ были размещены в области размером 1260 × 1260 м. Радиус зоны покрытия базовых станций – 105 м для IAB-ретрансляторов и 210 м для IAB-донора. Интервал передачи TTI $t_{TTI} = 1$ мс, цикл передачи при мультиплексировании с временным разделением TDM $T = 6t_{TTI}, \lambda = 200$ с$^{-1}$.

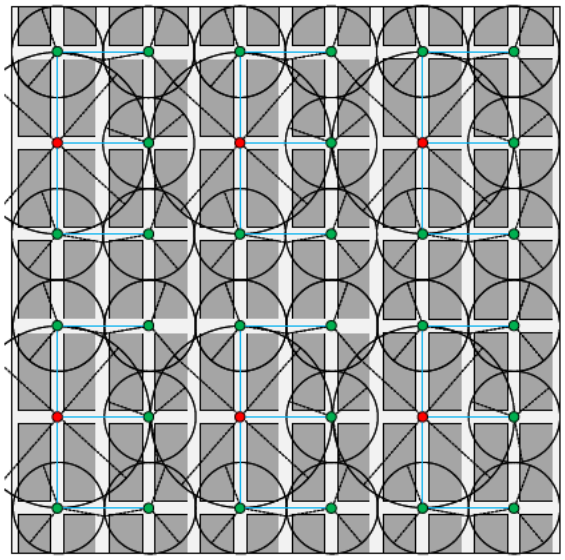


Рис. 3.5. Пример развертывания Топологии № 2:
красная точка – донор, зеленая точка – ретранслятор

Обеспечение покрытия изображенной на рис. 3.5 области требует применения одного из указанных вариантов топологии:

- 9 сетей Топологии № 0 (только донор),
- 7 сетей Топологии № 1 (донор + 3 ретранслятора = 4 IAB-узла),
- 6 сетей Топологии № 2 (донор + 5 ретрансляторов = 6 IAB-узлов),
- 4 сети Топологии № 3 (донор + 7 ретрансляторов = 8 IAB-узлов),
- 2 сети Топологии № 4 (донор + 14 ретрансляторов = 15 IAB-узлов).

На рис. 3.6 показана сквозная задержка в зависимости от числа узлов в сети.

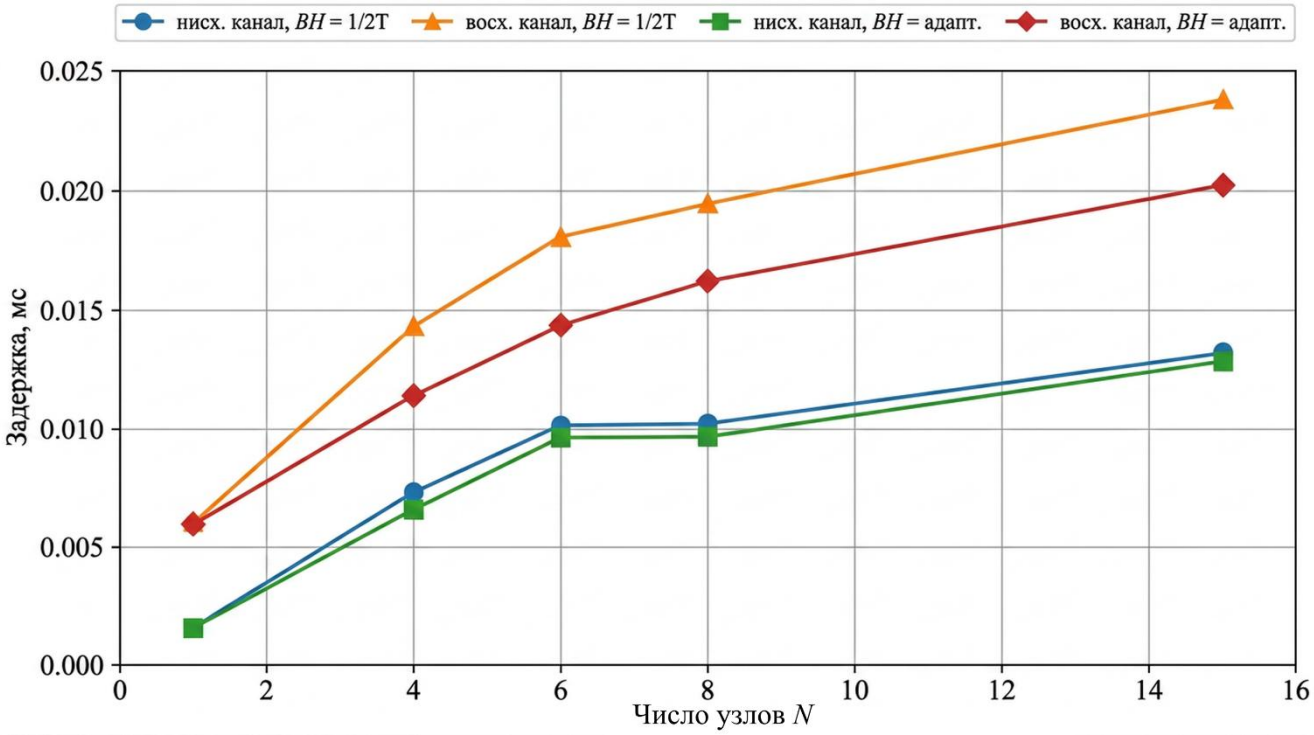


Рис. 3.6. Средняя задержка в зависимости от числа узлов сети

На рис. 3.7 показана средняя сквозная задержка в зависимости от числа шагов в многошаговом маршруте. Можно наблюдать, что тренд близок к линейному, и средняя сквозная задержка увеличивается примерно на 4–6 мс при добавлении к сети очередного яруса.

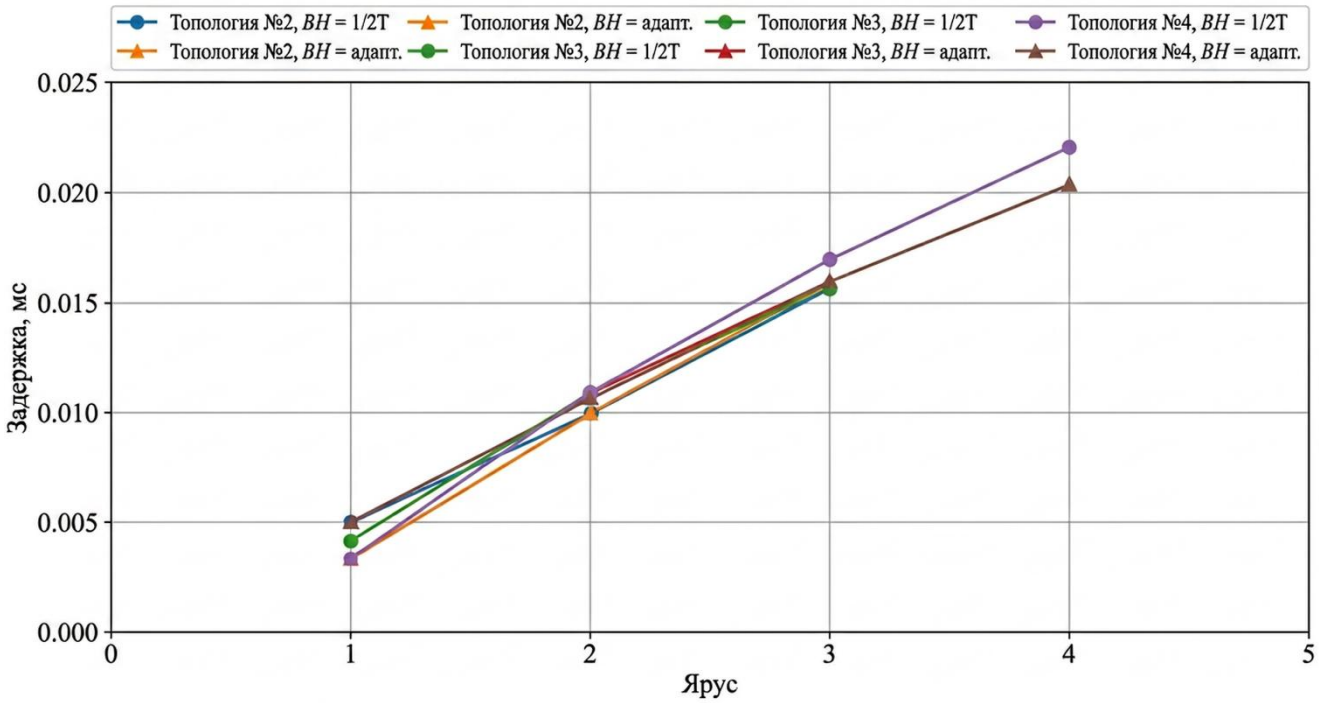


Рис. 3.7. Средняя задержка в зависимости от уровня сети

Несомненно, имитационное моделирование является надежным инструментом, с помощью которого можно учесть многие системные параметры, которыми приходится пренебрегать при построении аналитических моделей. Более детально с имитационным моделированием сети IAB можно познакомиться в работе [33] с участием автора диссертационной работы. Однако основной целью диссертационной работы является построение и анализ математической модели, допускающей аналитическое решение в замкнутом виде для функции распределения маршрутных сетевых задержек. Этому посвящены следующие два раздела Главы 3 диссертационной работы.

3.2. Математическая модель в виде экспоненциальной сети массового обслуживания

Сеть IAB имеет древовидную структуру, при этом БС соответствуют корню (IAB-донор) и вершинам ветвления (IAB-ретрансляторы), а секторы (n, i) – листьевым вершинам. Пример графовой модели сети IAB показан на рис. 3.8.

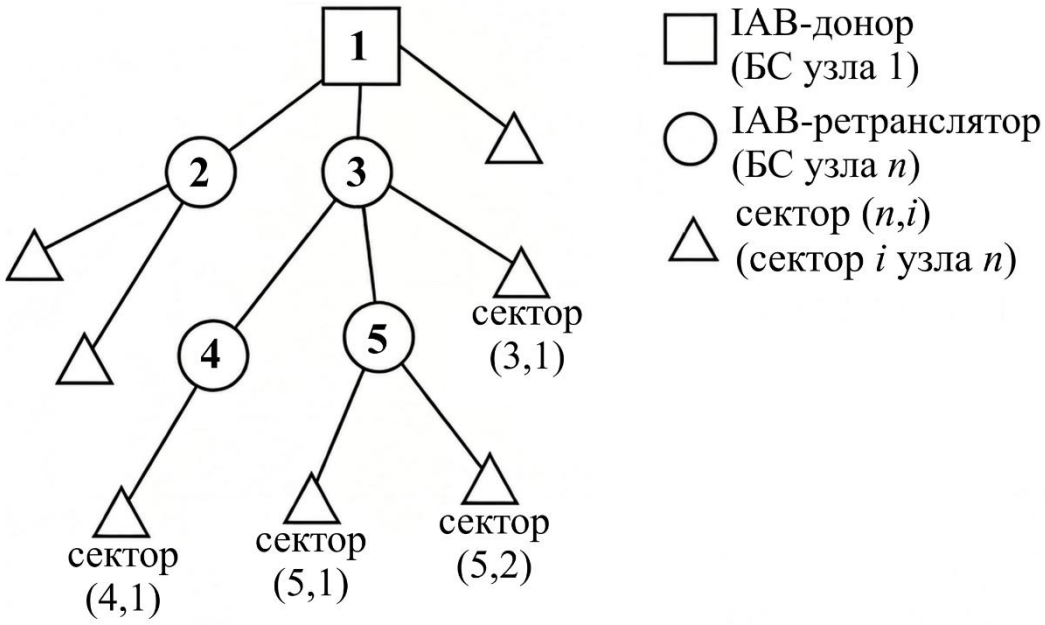


Рис. 3.8. Схема сети IAB в виде древовидного графа

В разделе 3.2, с учетом [27], показан подход к построению математической модели процесса передачи пакетов в сети IAB, заданной моделью древовидного графа, в виде модели открытой экспоненциальной сети массового обслуживания

для анализа полудуплексного (half-duplex, HD) и полнодуплексного (full-duplex, FD) режимов передачи с точки зрения сквозной задержки.

Без потери общности рассмотрена передача только в нисходящем направлении от донора к АУ, поскольку передача в восходящем направлении моделируется аналогично. Т.о. при активации планировщиком канала транзита секторы связанных этим каналом родительского и дочернего узлов переводятся в состояния «передача транзита» и «прием транзита» соответственно, а при активации планировщиком канала доступа сектор узла, связанного с АУ – в состояние «передача доступа».

Табл. 3.2 показывает соответствие между сетью IAB и ее математической моделью в виде открытой экспоненциальной СеМО.

Табл. 3.2. Обозначения математической модели.

Сеть IAB	**Графовая модель**	**Открытая СеМО**
Множество узлов сети IAB	Множество вершин древовидного графа	
Множество БС	Корень и множество $\mathcal{B}$ вершин ветвления графа	
Множество каналов сети IAB	Множество $\mathcal{E}$ ребер древовидного графа	Множество узлов СеМО без узла 0 «внешняя среда»
Радиоканал сети IAB	e-ребро древовидного графа, $e \in \mathcal{E}$	Узел СеМО = СМО типа M|M|1|∞
Множество многошаговых маршрутов в сети IAB от донора к каждому АУ	Множество $\mathcal{P}$ физических путей древовидного графа от корня до каждой из листьевых вершин	Множество всех маршрутов по СеМО
Множество каналов сети IAB, составляющих маршрут от донора к АУ p	Множество $\mathcal{E}_p$ ребер древовидного графа, составляющих p-путь, $p \in \mathcal{P}$	Множество узлов СеМО на маршруте p
Интенсивность поступления на донор из проводной опорной транспортной сети пакетов для АУ p	Интенсивность генерации в корне древовидного графа заявок на p-путь	Интенсивность поступления извне (из узла 0) на узел 1 заявки для маршрута p
Интенсивность поступления пакетов на e-канал сети IAB	Суммарная интенсивность $\lambda_e = \sum_{p:e\in\mathcal{E}_p} \Lambda_p$ поступления пакетов по p-путям	Интенсивность поступления заявок на e-узел СеМО

Сеть IAB	Графовая модель	Открытая СеМО
	древовидного графа, содержащим e-ребро, $e \in \mathcal{E}$	
Пропускная способность канала доступа/транзита сети IAB на интервалах активности	Вес C_e ребра древовидного графа, $C_e = C,\ e \in \mathcal{E}$	Интенсивность обслуживания в e-узле СеМО при работе без простоя
Вектор параметров для менеджера нарезки ресурса SliM	Вектор $\mathbf{q} = (q_e)_{e\in\mathcal{E}}, q_e \in [0,1]$ долей времени активности канала	Вектор, где компонент q_e – доля времени исправной работы прибора e-узла СеМО
Пропускная способность канала доступа/транзита сети IAB с учетом интервалов неактивности	Урезанный вес $\mu_e = C_e q_e$ ребра древовидного графа, $e \in \mathcal{E}$	Интенсивность обслуживания в e-узле СеМО с учетом простоя из-за неактивности приборов

Используя принятые в [26] обозначения и модифицируя их, введем нотацию для анализа модели сети IAB, иллюстрируя нотацию табл. 3.2 на примере сети на рис. 3.8:

- $\mathcal{B}$ – множество БС (донор и ретрансляторы) сети IAB, $\mathcal{B} = \{0,1,2,3,4\}$;
- $\mathcal{E}$ – множество каналов доступа и каналов транзита сети IAB, $\mathcal{E} = \{1,2,3,4,5,6,7,8,9,10,11\}$;
- C – ресурс сети IAB, пропускная способность радиоканала [бит/с];
- $\mathbf{c} = (C_e)_{e\in\mathcal{E}}$ – вектор пропускных способностей каналов доступа/транзита, [бит/с];
- $\mathcal{P}$ – множество маршрутов сети IAB от донора к соответствующему сектору родительского узла для АУ, далее – множество физических путей (p-путей), $\mathcal{P} = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$;
- $\mathcal{E}_p$ – подмножество радиоканалов, составляющих p-путь, $\mathcal{E}_{p_1} = \{1\}$, $\mathcal{E}_{p_2} = \{2,4\}$, $\mathcal{E}_{p_3} = \{2,5,9\}$, $\mathcal{E}_{p_4} = \{2,5,10\}$, $\mathcal{E}_{p_5} = \{2,6,11\}$, $\mathcal{E}_{p_6} = \{3,7\}$, $\mathcal{E}_{p_7} = \{3,8\}$

  которое может быть задано в виде матрицы маршрутов

$$\mathbf{R} = \left(r_{ep}\right)_{e\in\mathcal{E},\ p\in\mathcal{P}}, \text{ где } r_{ep} = \begin{cases} 1, & e \in \mathcal{E}_p, \\ 0, & e \notin \mathcal{E}_p. \end{cases}$$

Отметим, что матрица маршрутов для сети IAB на рис. 3.1 имеет вид

$$\mathbf{R} = \begin{array}{c|ccccccc} e/p & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 9 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 11 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}.$$

Обозначим $\Lambda_p \left[\frac{1}{\text{мс}}\right],\ p \in \mathcal{P}$ – интенсивность поступления пакетов на p-путь, $\lambda_e = \sum_{p:e\in\mathcal{E}_p} \Lambda_p\,,\ e \in\ \mathcal{E}$ – интенсивность потока пакетов между смежными узлами, соединенными e-каналом. Вектор $\boldsymbol{\lambda} = (\lambda_e)_{e\in\mathcal{E}}$ для сети на рис. 3.8 имеет вид

$$\boldsymbol{\lambda} = \left(\Lambda_{p_1}, \Lambda_{p_2} + \Lambda_{p_3} + \Lambda_{p_4} + \Lambda_{p_5}, \Lambda_{p_7} + \Lambda_{p_8}, \Lambda_{p_2} \cdot \Lambda_{p_3} + \Lambda_{p_4} \cdot \Lambda_{p_5} \cdot \Lambda_{p_6} \cdot \Lambda_{p_7}, \Lambda_{p_3}, \Lambda_{p_4}, \Lambda_{p_5}\right).$$

Поскольку и в HD, и в FD режиме доступа/транзита активен не постоянно, а часть времени простаивает, введем долю времени активности e-канала $q_e \in [0; 1]$, которая представляет отношение суммарного времени, в течение которого по e-каналу передаются пакеты, к длительности некоторого достаточно большого интервала наблюдения за сетью. Соответствующий вектор $\mathbf{q} = (q_e)_{e\in\mathcal{E}}$ долей времени активности каналов доступа / транзита является решением сформулированной в разделе 3.3 задачи минимизации средней по сети сквозной задержки или среднего по сети пикового возраста информации и может быть найден в соответствии с разработанным Алгоритмом 3.1 численного решения проблемы оптимизации сквозной задержки и возраста информации. Оптимальные доли времени активности каналов доступа/транзита используются менеджером нарезки ресурса SliM при планировании последовательности активации и длительности интервалов активности каналов.

Введем $\mu_e = C_e q_e$ – пропускную способность e-канала доступа / транзита, $e \in \mathcal{E}$, и соответствующий вектор $\boldsymbol{\mu} = (\mu_e)_{e\in\mathcal{E}}$. Заметим, что доля q_e времени активности e-канала – аналог доли q_s от общего ресурса системы C вида (1.8), а

значение μ_e – аналог выделенной слайсу $s \in \mathcal{S}$ емкости C_s, определенных в Главе 1 при формализации систем с нарезкой ресурса.

Введем $\mathbf{F} = (f_{ne})_{n\in\mathcal{B},e\in\mathcal{E}}$ матрицу конфликтов, где строка соответствует донору или ретранслятору, а столбец – каналу доступа/транзита сети. Единицы в матрице соответствуют конфликтующим каналам, т.е. каналам, которые не могут быть активны одновременно. Матрицы конфликтов $\mathbf{F}_{HD}$ для HD и $\mathbf{F}_{FD}$ для FD режимов в сети на рис. 3.8 имеют вид

$$\mathbf{F}_{HD} = \begin{matrix} n/e & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{matrix}, \tag{3.2}$$

$$\mathbf{F}_{FD} = \begin{matrix} n/e & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}. \tag{3.3}$$

Теперь, когда описана математическая модель сети IAB в виде открытой экспоненциальной СеМО, можно формализовать ограничения на параметры сети IAB, которые необходимо учитывать в задаче оптимизации сквозной задержки и возраста информации.

Для сети IAB должно выполняться условие своевременной доставки, т.е. доля пакетов, доставленных от донора на АУ с задержкой, не превышающей δ_α секунд, должна быть выше α. Также необходимо найти оптимальные значения интенсивностей поступления $\boldsymbol{\lambda}$ и обслуживания $\boldsymbol{\mu}$, которые максимизируют пропускную способность сети. Прокомментируем каждое из ограничений.

Пусть p-путь используется для доставки трафика некоторой услуги, для которой требованиями QoS определена минимальная ненулевая интенсивность поступления пакетов $\lambda_{\min} > 0$. Тогда для успешного предоставления услуги интенсивность поступления пакетов на p-путь должна быть не менее $\lambda_{\min}$:

$$\Lambda_p \geq \lambda_{\min},\ p \in \mathcal{P}.$$

Условия существования стационарного режима в открытой экспоненциальной СеМО имеют вид

$$\boldsymbol{\lambda} < \boldsymbol{\mu}, \tag{3.4}$$

т.е. для каждого канала сети IAB в нисходящем направлении интенсивность λ_e входящего потока не должна быть выше интенсивности μ_e обслуживания.

Ограничение длительностей активности каналов сети IAB задается с помощью матрицы конфликтов следующим образом:

$$\mathbf{F}\mathbf{q}^T \leq \mathbf{1}. \tag{3.5}$$

Обозначим D_p, $p \in \mathcal{P}$, с.в. сквозной задержки на p-пути. Интерес представляет ее математическое ожидание d_p и правосторонний квантиль $\delta_{D_p}^{+}(\alpha)$ заданного уровня α, т.е. значение с.в. D_p, которое превышают лишь $(1-\alpha)$ наблюдений. Правосторонний квантиль $\delta_{D_p}^{+}(\alpha)$ уровня α связан с квантилем δ_α следующим соотношением:

$$\delta_{D_p}^{+}(\alpha) = \delta_{1-\alpha}, \tag{3.6}$$

где квантиль δ_α имеет вид

$$\mathrm{P}\{D_p \leq \delta_\alpha\} = \alpha. \tag{3.7}$$

Например, для $\alpha = 0.999$ правосторонний квантиль $\delta_{D_p}^{+}(0.999)$ указывает границу, выше которой значения с.в. D_p встречаются редко – с вероятностью $1-\alpha = 0.001$.

Заметим, что при заданном ограничении τ на время доставки пакета формула (3.7) определяет долю своевременно доставленных пакетов $\mathrm{P}\{D_p \leq \tau\}$. Таким образом возникает задача анализа функции распределения с.в. D_p.

Как упоминалось, интерес представляет формулировка задачи максимизации пропускной способности сети IAB при минимизации сквозной задержки пакетов и пикового возраста информации в ограничениях на вероятность своевременной доставки пакетов. Решение задачи оптимизации зависит от топологии сети и различается для HD и FD режимов передачи данных в сети IAB.

3.3. Анализ сквозной задержки

Будем моделировать каждый e-узел СеМО как систему массового обслуживания типа $M/M/1/\infty$. Интенсивности поступления пакетов на каналы сети IAB соответствуют интенсивностям входящих потоков в узлы СеМО и заданы вектором $\boldsymbol{\lambda}$, пропускные способности каналов сети IAB соответствуют интенсивностям обслуживания в узлах СеМО заданы вектором $\boldsymbol{\mu}$.

Теорема 3.1. Для древовидной сети IAB с множеством $\mathcal{E}$ каналов доступа / транзита; интенсивностями λ_e потоков пакетов на e-канал; емкостями $C_e q_e$ каналов доступа / транзита с учетом долей q_e времен активности каналов, $e \in \mathcal{E}$; емкостью C общего ресурса радиоканала, разделяемого между e-каналами на основе временного разделения ресурса, случайная величина D_p сквозной задержки на p-пути, начинающемся с e^*-канала, $e^* \in \mathcal{E}_p \subseteq \mathcal{E}$, имеет PH-распределение с функцией распределения вида

$$F_{D_p}(x) = 1 - \boldsymbol{\beta}^T e^{\mathbf{M}_p x}\mathbf{1},\ p \in \mathcal{P}, \tag{3.8}$$

где $\boldsymbol{\beta}^T = (\beta_e)_{e\in\mathcal{E}_p}$ – вектор-строка размерности $|\mathcal{E}_p|$ с компонентами

$$\beta_e = \begin{cases} 1, & e = e^*; \\ 0, & e^* \in \mathcal{E}_p \backslash \{e^*\}, \end{cases} \tag{3.9}$$

матрица $\mathbf{M}_p$ – диагональная матрица размерности $|\mathcal{E}_p| \times |\mathcal{E}_p|$ с ненулевыми элементами вида $(\lambda_e - C_e q_e)$, $e \in \mathcal{E}_p$, $p \in \mathcal{P}$.

Доказательство.

Обозначим W_e с.в. времени пребывания в e-узле, $e \in \mathcal{E}$. Известно [82,83], что в сделанных предположениях с.в. W_e распределена экспоненциально с параметром

$$\gamma_e = \mu_e - \lambda_e. \tag{3.10}$$

Тогда с.в. D_p задержки на p-пути определяется формулой

$$D_p = \sum_{e\in\mathcal{E}_p} W_e\,,\ p \in \mathcal{P}, \tag{3.11}$$

и, следовательно, имеет распределение Эрланга, которое является частным случаем распределения фазового типа (PH-распределения) [84]. В этом случае функция

распределения с.в. D_p сквозной задержки на p-пути, начинающемся с e^*-канала, $e^* \in \mathcal{E}_p \subseteq \mathcal{E}$, имеет вид (3.8) с диагональной матрицей $\mathbf{M}_p$ размерности $|\mathcal{E}_p| \times |\mathcal{E}_p|$, ненулевыми элементами которой являются параметры γ_e, $e \in \mathcal{E}_p$, вида (3.10) и вектором $\boldsymbol{\beta}^T = (\beta_e)_{e \in \mathcal{E}_p}$, компоненты которого имеют вид (3.9).

Теорема доказана.

Следствие 3.1. Для сети IAB, описанной в Теореме 3.1, ФР с.в. A_p пикового возраста информации p-пакета имеет вид

$$F_{A_p}(x) = 1 - \boldsymbol{\beta}^T e^{\mathbf{N}_p x}\mathbf{1},\ p \in \mathcal{P}, x > 0, \tag{3.12}$$

где $\boldsymbol{\beta}^T = (1,0,\dots,0)$ – вектор-строка размерности $|\mathcal{E}_p| + 1$, $\beta_1 = 1$ соответствует ближнему к донору каналу,

матрица $\mathbf{N}_p$ – диагональная матрица размерности $(|\mathcal{E}_p| + 1) \times (|\mathcal{E}_p| + 1)$

$$\mathbf{N}_p = \begin{pmatrix} -\Lambda_p & 0 & 0 & \dots & 0 \\ 0 & -(\mu_1 - \lambda_1) & & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\left(\mu_{|\mathcal{E}_p|} - \lambda_{|\mathcal{E}_p|}\right) \end{pmatrix}. \tag{3.13}$$

Доказательство.

Для произвольного p-пути обозначим t_k момент поступления k-пакета на донор из опорной транспортной сети (или момент генерации k-пакета донором) и t_k' – момент получения k-пакета соответствующим АУ сети IAB. Согласно [29], возраст информации $\Delta(t)$ на АУ в момент времени t равен времени, прошедшему с момента генерации последнего принятого АУ пакета, т.е.

$$\Delta(t) = t - t_{N(t)},\ p \in \mathcal{P},$$

где $N(t)$ – номер последнего пакета, принятого АУ на момент времени t. Пусть $Y^{(k)}$ – интервал времени между моментами генерации $(k-1)$-пакета и k-пакета, $D^{(k)}$ – время доставки k-пакета от донора до АУ, т.е.

$$Y^{(k)} = t_k - t_{k-1},\quad D^{(k)} = t_k{}' - t_k.$$

С учетом введенных обозначений пиковый возраст $A^{(k)}$ информации, содержащейся в k-пакете, определяется формулой

$$A^{(k)} = t_k{}' - t_{k-1} = Y^{(k)} + D^{(k)}.$$

Обозначим с.в. Y_p время между моментами генерации (поступления на донор) пакетов для АУ по p-пути, и с.в. D_p время доставки пакета от донора до АУ по p-пути. В сделанных выше предположениях о мгновенной передаче пакета между узлами сети IAB, если соединяющий их канал доступа/транзита активен, с.в. D_p соответствует сквозной задержке на p-пути. Из построения математической модели можно считать, что Y_p распределена экспоненциально с параметром Λ_p. Тогда с.в. A_p пикового возраста информации [71] для заявок на p-пути

$$A_p = Y_p + D_p, p \in \mathcal{P},$$

имеет обобщенное распределение Эрланга с $\left|\mathcal{E}_p\right| + 1$ этапом, время пребывания на каждом из которых распределено экспоненциально с параметром Λ_p для первого этапа и параметрами $\gamma_e = \mu_e - \lambda_e$, $e \in \mathcal{E}_p$, для остальных этапов, $p \in \mathcal{P}$. Тогда для p-пути ФР с.в. A_p пикового возраста информации является частным случаем распределения фазового типа (3.12) с $\left|\mathcal{E}_p\right| + 1$ этапом, где $\boldsymbol{\beta}^T = (1,0,\dots,0)$, $\mathbf{N}_p$ имеет вид (3.13).

Следствие доказано.

Среднее по времени значение d_p сквозной задержки пакета по p-пути определяется формулой

$$d_p = MD_p = \sum_{e \in \mathcal{E}_p} \frac{1}{\mu_e - \lambda_e},\ p \in \mathcal{P},$$

а среднее по всем путям значение d сквозной задержки пакета в сети имеет вид

$$d = \sum_{p \in \mathcal{P}} \left(\frac{\Lambda_p}{\sum_{p \in \mathcal{P}} \Lambda_p} \cdot d_p \right) = \frac{1}{\sum_{p \in \mathcal{P}} \Lambda_p} \sum_{p \in \mathcal{P}} \left(\Lambda_p \cdot \sum_{e \in \mathcal{E}_p} \frac{1}{\mu_e - \lambda_e} \right),\ p \in \mathcal{P}. \qquad (3.14)$$

Из (3.14) среднее по времени значение a_p пикового возраста информации на p-пути определяется формулой

$$a_p = MA_p = MY_p + MD_p = \frac{1}{\Lambda_p} + \sum_{e \in \mathcal{E}_p} \frac{1}{\gamma_e},\ p \in \mathcal{P},$$

а среднее по путям значение a пикового возраста информации в сети имеет вид

$$a = \sum_{p\in\mathcal{P}} \left(\frac{\Lambda_p}{\sum_{p\in\mathcal{P}} \Lambda_p} \cdot a_p \right) = \frac{1}{\sum_{p\in\mathcal{P}} \Lambda_p} \sum_{p\in\mathcal{P}} \left(1 + \Lambda_p \cdot \sum_{e\in\mathcal{E}_p} \frac{1}{\gamma_e} \right). \quad (3.15)$$

Введя все обозначения и выписав основные формулы, сформулируем проблему оптимизации средней сквозной задержки и среднего пикового возраста информации в сети:

$$U(\mathbf{q}) \to \min_{\mathbf{q}} \quad (3.16)$$

$$R_1: 0 \le q_e \le 1, \forall e \in \mathcal{E};$$

$$R_2: \lambda_e \le C_e q_e, \forall e \in \mathcal{E}; \quad (3.17)$$

$$R_3: \mathbf{F}\mathbf{q}^T \le \mathbf{1}.$$

Вектор $\mathbf{q}^*$ оптимальных долей времени активности каналов предлагается находить как решение проблемы оптимизации (3.16) с целевой функцией $U(\mathbf{q})$ вида (3.14) для задачи минимизации среднего значения сквозной задержки и вида (3.15) для задачи минимизации среднего значения пикового возраста информации (3.15), в одинаковых для обеих задач ограничениях (3.17), задаваемых матрицей конфликтов **F** (вида (3.2) для HD-режима и вида (3.3) для FD-режима функционирования сети), учитывая условия существования стационарного режима (3.4).

Для количественной оценки показателей качества предоставления услуг в сети интегрированного доступа и транзита с разделением ресурсов разработан алгоритм расчета оптимальных долей времени активности каналов сети IAB, минимизирующих значение средней по сети сквозной задержки, а также вычисления правостороннего квантиля заданного уровня пикового возраста информации.

Алгоритм 3.1. Пусть древовидная сеть IAB задана следующими параметрами:

множество $\mathcal{B}$ базовых станций (донор и ретрансляторы), для каждой из которых задано число N_n секторов антенны для БС n, $n \in \mathcal{B}$;

множество $\mathcal{E}$ каналов доступа / транзита;

множество $\mathcal{P}$ маршрутов от донора к АУ в секторе (n,i), $n \in \mathcal{B}$, $i = 1,2,\dots,N_n$;

подмножества $\mathcal{E}_p$ составляющих маршрут p каналов, $\mathcal{E}_p \subseteq \mathcal{E}$, $p \in \mathcal{P}$;

матрица конфликтов $\mathbf{F} = (f_{ne})_{n\in\mathcal{B},e\in\mathcal{E}}$, где в строке, соответствующей БС n, элементы $f_{ne_1} = f_{ne_2} = \cdots = 1$, если каналы e_1, e_2, … не могут быть активны одновременно;

интенсивности Λ_p потоков пакетов для АУ в секторе (n,i) конечного узла n маршрута p, $n \in \mathcal{B}$, $i = 1,2,\dots,N_n$, $p \in \mathcal{P}$;

емкость C радиоканала;

емкости $C_e = C$ каналов доступа / транзита, $e \in \mathcal{E}$;

доли q_e времени активности каналов доступа / транзита, $e \in \mathcal{E}$.

Тогда вектор $\mathbf{q}^* = \left(q_1^*, q_1^*, \dots, q_{|\mathcal{E}|}^*\right)$ оптимальных долей времени активности каналов, минимизирующий значение средней по сети сквозной задержки в нисходящем направлении от донора к АУ, и правосторонний квантиль $\delta_{A_p}^+(\alpha)$ заданного уровня α пикового возраста информации для маршрута p в сети IAB, $p \in \mathcal{P}$, могут быть вычислены по следующему алгоритму.

Исходные данные: $\mathcal{B}$; $N_n, n \in \mathcal{B}$; $\mathcal{E}$; $\mathcal{P}$; $\mathcal{E}_p \subseteq \mathcal{E}$, $p \in \mathcal{P}$; $\mathbf{F}$; $\boldsymbol{\Lambda_p} = \left(\Lambda_p\right)_{p\in\mathcal{P}}$; C.

Результат: $\mathbf{q}^* = (q_e^*)_{e\in\mathcal{E}}$; $\delta_{A_p}^+(\alpha), p \in \mathcal{P}$.

ШАГ 1. Вычисление компонент вектора интенсивностей $\boldsymbol{\lambda_e} = (\lambda_e)_{e\in\mathcal{E}_p}$ потоков пакетов на e-канал по формуле

$$\lambda_e = \textstyle\sum_{p:e\in\mathcal{E}_p} \Lambda_p,\ e \in \mathcal{E}.$$

ШАГ 2. Вычисление компонент вектора $\mathbf{q}^* = \left(q_1^*, q_1^*, \dots, q_{|\mathcal{E}|}^*\right)$ оптимальных долей времени активности каналов как решение методом множителей Лагранжа проблемы оптимизации $U(\mathbf{q}) \to \min\limits_{\mathbf{q}}$ с целевой функцией

$$U(\mathbf{q}) = \sum_{p\in\mathcal{P}} \left(\frac{\Lambda_p}{\sum_{p\in\mathcal{P}} \Lambda_p} \cdot d_p \right) = \frac{1}{\sum_{p\in\mathcal{P}} \Lambda_p} \sum_{p\in\mathcal{P}} \left(\Lambda_p \cdot \sum_{e\in\mathcal{E}_p} \frac{1}{C_e q_e - \lambda_e} \right)$$

в ограничениях

$$R_1: 0 \le q_e \le 1, \forall e \in \mathcal{E};$$

$$R_2: \lambda_e \le C_e q_e, \forall e \in \mathcal{E};$$

$$R_3: \mathbf{F}\mathbf{q}^T \le \mathbf{1};$$

для существования стационарного режима функционирования сети IAB с матрицей конфликтов $\mathbf{F} = (f_{ne})_{n\in\mathcal{B}, e\in\mathcal{E}}$, отражающей полудуплексный и полнодуплексный режимы передачи данных в каналах доступа / транзита.

ШАГ 3. Получение правостороннего квантиля $\delta^+_{A_p}(\alpha)$ заданного уровня α пикового возраста информации для маршрута p в сети IAB как $\delta^+_{A_p}(\alpha) = \delta_{1-\alpha}$, где $\delta_{1-\alpha}$ – решение уравнения

$$1 - \boldsymbol{\beta}^T e^{\mathbf{N}_p \delta_{1-\alpha}} \mathbf{1} = 1 - \alpha,$$

в котором $\boldsymbol{\beta}^T = (1,0,\dots,0)$ – вектор-строка размерности $|\mathcal{E}_p| + 1$, $\beta_1 = 1$ соответствует ближнему к донору каналу,

матрица $\mathbf{N}_p$ – диагональная матрица размерности $(|\mathcal{E}_p| + 1) \times (|\mathcal{E}_p| + 1)$

$$\mathbf{N}_p = \begin{pmatrix} -\Lambda_p & 0 & 0 & \dots & 0 \\ 0 & -(C_1 q_1 - \lambda_1) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\left(C_{|\mathcal{E}_p|} q_{|\mathcal{E}_p|} - \lambda_{|\mathcal{E}_p|}\right) \end{pmatrix}, \; p \in \mathcal{P}.$$

3.4. Анализ задачи оптимизации задержки в сети

Численный анализ сквозной задержки и пикового возраста информации проведен для сценария, описанного в разделе 3.1 (рис. 3.4).

В принятых выше допущениях геометрия городской застройки задается кварталами размером 76 × 183 м и улицами/проспектами средней ширины около 30 м. Радиус зоны покрытия антенн БС – 105 м для ретрансляторов и 210 м для IAB-донора, т.е. площадь зоны покрытия IAB-донора в 4 раза больше площади зоны покрытия ретранслятора. Размещение узлов соответствует вариантам топологии, показанным на рис. 3.4. согласно работе [85]. Пример реализации варианта топологии, которая наилучшим образом покрывает квартал, детально показан на

рис. 3.9, первый узел оснащен 4 антеннами ($N_1 = 4$), остальные – тремя антеннами ($N_n = 3,\ n = 2 \ldots N$). В [33] показано, что необходимо 6 сетей рассматриваемой топологии, чтобы полностью покрыть выбранную область.

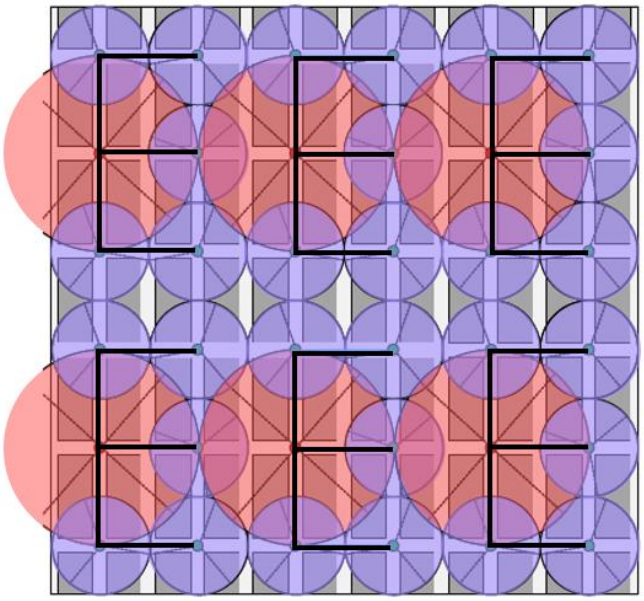


Рис. 3.9. Пример развертывания топологии:

красный круг – зона покрытия донора; фиолетовый круг – ретранслятора

На территории квартала в пространстве $\mathbb{R}^2$ в границах улиц равномерно распределены 10 000 АУ. Трафик ко всем АУ, связанным с сектором (*n, i*) (сектор *i* узла *n*) моделируется пуассоновским потоком с интенсивностью

$$\lambda_{(n,i)} = L_{(n,i)} \cdot \lambda. \qquad (3.18)$$

Здесь $L_{(n,i)}$ –среднее число АУ в секторе, а λ – удельная нагрузка одного АУ (средняя интенсивность нагрузки, которую генерирует одно АУ). В полудуплексном (HD) режиме каналы, связанные с одним узлом (канал доступа на АУ и канал транзита на другую БС), не могут быть активны одновременно, поэтому передача доступа и передача транзита моделируется с помощью двух разных каналов. На рис. 3.10 показана схема рассматриваемой сети, где обозначены БС-донор (квадрат), БС-ретрансляторы (круги) и секторы (треугольники), а также пронумерованы узлы сети IAB (в квадрате и кругах), многошаговые пути (в треугольниках) и каналы (на отрезках/ребрах).

Основной интересующей метрикой является с.в. сквозной задержки доставки пакета по многошаговому маршруту от донора до АУ в нисходящем направлении

как функция от параметров системы. Сквозная задержка в разделе 3.1 определена как случайный интервал времени с момента, когда пакет из опорной транспортной сети попадает в буфер транзита в узле 1, до момента, когда пакет покидает соответствующий сектор родительского узла для АУ по нисходящей линии к АУ.

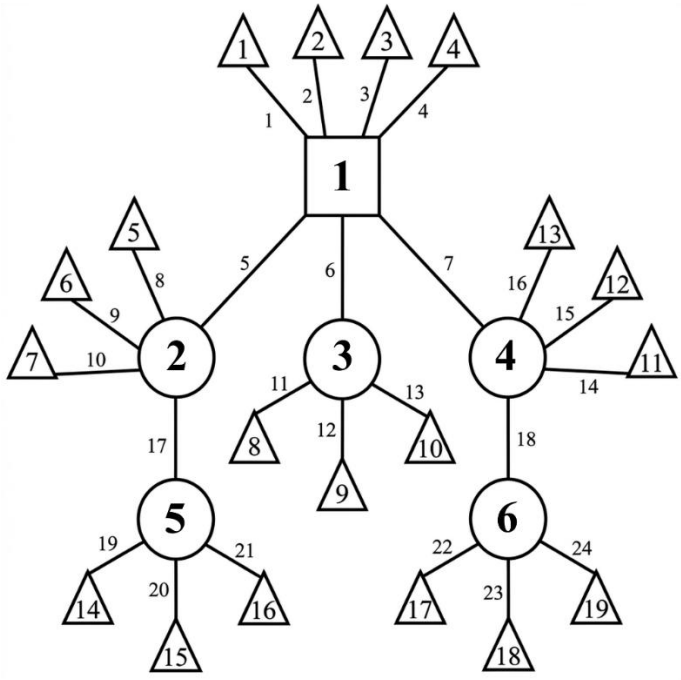


Рис. 3.10. Схема сети: квадрат – БС-донор, круг – БС-ретранслятор, треугольник – сектор узла сети IAB

Для сети IAB на рис. 3.10 в указанных условиях имеют место следующие значения параметров модели.

$|\mathcal{B}| = 6$ – число БС сети IAB (донор и ретрансляторы);

$|\mathcal{P}| = 19$ – суммарное число секторов в узлах сети IAB;

$|\mathcal{E}| = 24$ – суммарное число каналов доступа и каналов транзита сети IAB;

$L_{(1,i)} = 185$ – число абонентов в секторе i узла 1 (БС-донор IAB), $i = 1, \dots, 4$;

$L_{(n,i)} = 62$ – число абонентов в секторе i узла n (БС-узла IAB), $i = 1, \dots, 3$, $n = 2, \dots, 6$;

$\lambda = 0{,}2$ мс$^{-1}$ – удельная нагрузка (средняя интенсивность потока пакетов из опорной транспортной сети для каждого АУ).

Предполагается $\theta_{[\text{бит}]}$=1500 байт – длина пакета, т.е. для $C_{[\text{пакет/ед.вр.}]} = 2000$ мс$^{-1}$ $C_{[\text{бит/с}]} = 12$ кбит/с – формула (3.1).

По формуле (3.18) получаем:

$\lambda_{(1,i)} = 37$ мс$^{-1}$, $i = 1, \dots, 4$ – интенсивность потока пакетов из опорной транспортной сети для абонентов в секторе i узла 1 (БС-донор IAB), $i = 1, \dots, 4$;

$\lambda_{(n,i)} = 12.4$ мс$^{-1}$, где $n = 2, \dots, 6$ и $i = 1, \dots, 3$ – интенсивность потока пакетов из опорной транспортной сети для абонентов в секторе i узла n (БС-узла IAB), $i = 1, \dots, 3$, $n = 2, \dots, 6$.

Предположим, что из опорной транспортной сети на донор для каждого АУ пакеты поступают с одинаковой интенсивностью (симметричная нагрузка на секторы узлов). Тогда получаем следующие значения компонентов векторов $\mathbf{\Lambda}_p = \left(\Lambda_p\right)_{p\in\mathcal{P}}$, $\boldsymbol{\lambda}_e = (\lambda_e)_{e\in\mathcal{E}_p}$, $\mathbf{q} = (q_e)_{e\in\mathcal{E}}$ и матрицы конфликтов $\mathbf{F}_{\mathrm{HD}}$ размерности $|\mathcal{B}| \times |\mathcal{E}|$.

$\mathbf{\Lambda}_p = (37, 37, 37, 37, 12.4, 12.4, 12.4, 12.4, 12.4,$

$12.4, 12.4, 12.4, 12.4, 12.4, 12.4, 12.4, 12.4, 12.4, 12.4, 12.4),$

$\boldsymbol{\lambda}_e = (37, 37, 37, 37, 74, 37, 74, 12.4, 12.4, 12.4, 12.4,$

$12.4, 12.4, 12.4, 12.4, 12.4, 37, 37, 12.4, 12.4, 12.4, 12.4, 12.4, 12.4),$

$$\mathbf{F}_{\mathrm{HD}} = \begin{array}{c|cccccccccccccccccccccccc} n/e & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}.$$

Вектор $\mathbf{q}^* = \left(q_1^*, q_1^*, \dots, q_{|\mathcal{E}|}^*\right)$ оптимальных долей времени активности каналов доступа/транзита будем искать как решение задачи (3.16) минимизации средней задержки по долям $\mathbf{q}$ с ограничениями (3.17) для полудуплексного режима $\mathbf{F}_{\mathrm{HD}}$ вида (3.2) для топологии сети IAB на рис. 3.9, которая записывается следующим образом:

$$U(\mathbf{q}) = \sum_{p\in\mathcal{P}} \left(\sum_{e\in\mathcal{E}_p} \frac{1}{C_e q_e - \lambda_e} \right) \to \min_{\mathbf{q}}$$

с ограничениями

$$\begin{cases} q_1 + q_2 + q_3 + q_4 + q_5 + q_6 + q_7 = 1; \\ q_5 + q_8 + q_9 + q_{10} + q_{17} = 1; \\ q_6 + q_{11} + q_{12} + q_{13} = 1; \\ q_7 + q_{14} + q_{15} + q_{16} + q_{18} = 1; \\ q_{17} + q_{19} + q_{20} + q_{21} = 1; \\ q_{18} + q_{22} + q_{23} + q_{24} = 1. \end{cases}$$

В предположении симметричной нагрузки на АУ, т.е.

$$\begin{cases} q_1 = \cdots = q_4; \\ q_5 = q_7; \\ q_8 = \cdots = q_{10} = q_{14} = \cdots = q_{16}; \\ q_{17} = q_{18}; \\ q_{19} = \cdots = q_{24}, \end{cases}$$

число переменных сокращается, и для определения точек экстремума функции применим метод множителей Лагранжа. Получаем систему нелинейных уравнений, которую решаем численно методом Ньютона и находим решение, то есть вектор $\mathbf{q}^* = \left(q_1^*, q_1^*, \ \ldots\ , \ q_{|\mathcal{E}|}^*\right)$ оптимальных долей времени активности каналов доступа/транзита при симметричной нагрузке трафика на АУ:

$$\begin{cases} q_1^* = \cdots = q_4^* \approx 0.1 \\ q_5^* = q_7^* \approx 0.22 \\ q_6^* = 0.16 \\ q_8^* = \cdots = q_{16}^* \approx 0.17\,. \\ q_{11}^* = \cdots = q_{13}^* \approx 0.28 \\ q_{17}^* = q_{18}^* \approx 0.26 \\ q_{19}^* = \cdots = q_{24}^* \approx 0.25 \end{cases} \qquad (3.19)$$

Среднее по всем путям значение d сквозной задержки пакета в сети при найденных оптимальных долях $\mathbf{q}^*$ времени активности каналов по формуле (3.14) равна $d \approx 0.006$ мс.

Рис. 3.11 демонстрирует график зависимости средней задержки по сети d по формуле (3.14) от удельной (для одного АУ) нагрузки λ при разных значениях C емкости радиоканала для оптимальных значений долей активации каналов (3.19).

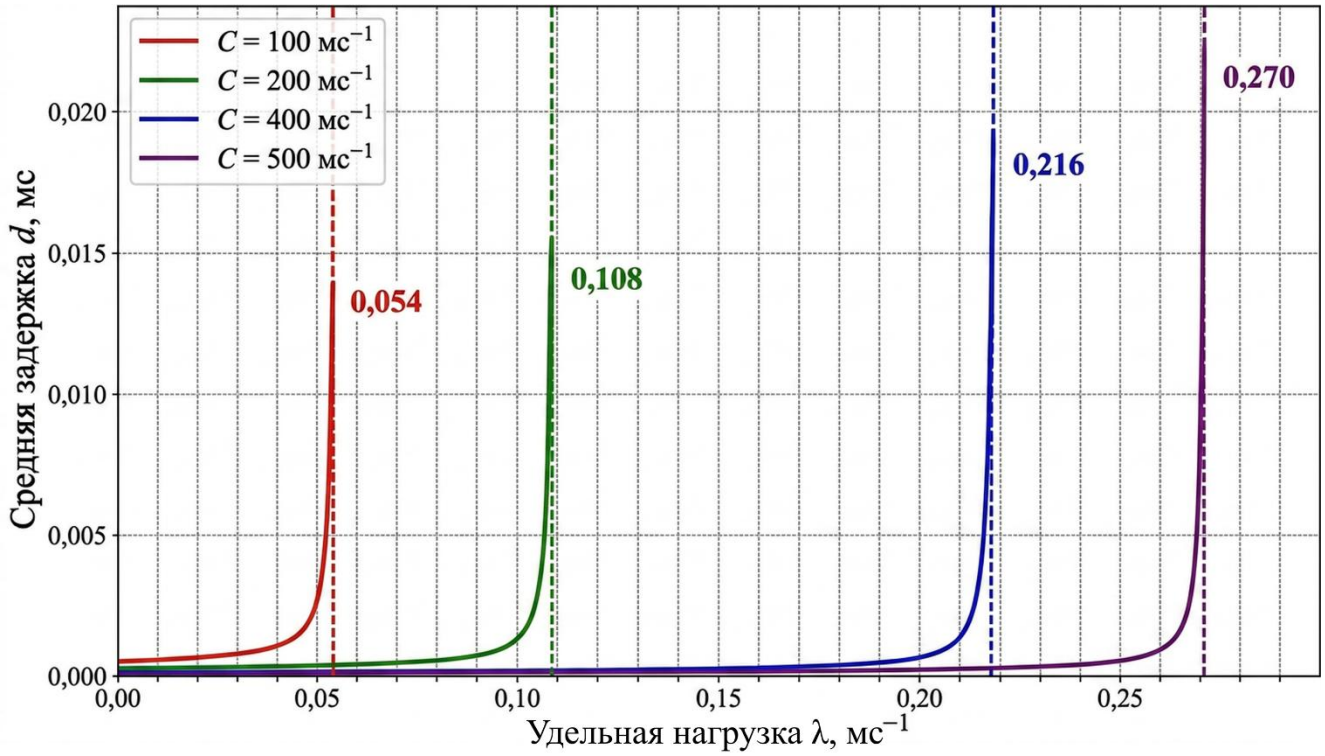


Рис. 3.11. Зависимость средней по сети задержки d от удельной нагрузки λ при различных значениях C емкости радиоканала для $\mathbf{q}^*$ по (3.19)

Далее в случае одного из самых длинных путей ($p = 14$) при найденных оптимальных долях $\mathbf{q}^*$ активности каналов, используя формулу (3.12) для функции распределения $F_{A_p}(x)$ пикового возраста информации A_p

$$F_{A_{14}}(x) = \mathrm{P}\{a_{14} < \delta_{0.999}\} = 1 - \boldsymbol{\beta}^T e^{\mathbf{N}_p \delta_{0.999}} \mathbf{1}$$

и определение квантиля $\delta_{0.999}$ уровня 0.999

$$F_{A_{14}}(x) = 0.999,$$

находим квантиль $\delta_{0.999}$, который, согласно формуле (3.7), соответствует правостороннему квантилю $\delta_{A_p}^+(\alpha)$ уровня $\alpha = 0{,}001$ возраста информации донора на АУ по

$$\delta_{A_p}^+(\alpha) = \delta_{A_{14}}^+(0.001) = \delta_{0.999}.$$

В результате получен правосторонний квантиль уровня 0.001 $\delta_{A_{14}}^+(0.001) \approx 0.026$, что означает, что с вероятностью 99,9% пиковый возраст информации на самом длинном с точки зрения числа шагов пути не превышает 0,026 мс.

График на рис. 3.12 показывает процесс поиска квантиля $\delta_{0.001}$ относительно вероятности превышения заданного значения порога. Можно отметить быструю сходимость численного метода к искомому квантилю.

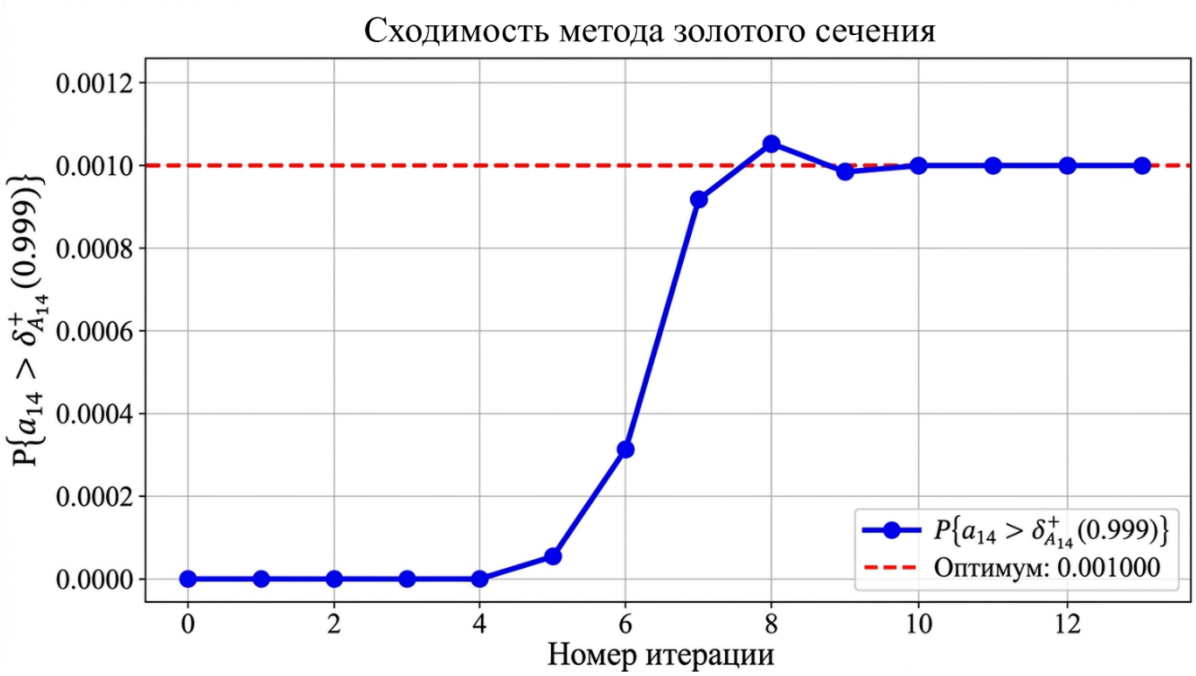


Рис. 3.12. Сходимость вероятности превышения

График на рис. 3.13 изменения параметра $\delta_{0.001}$ демонстрирует монотонное убывание от начального приближения до установившегося значения около 0,0261 мс. Стабилизация после 6–7 итераций подтверждает, что найденное значение является корректным решением и не зависит от начального приближения.

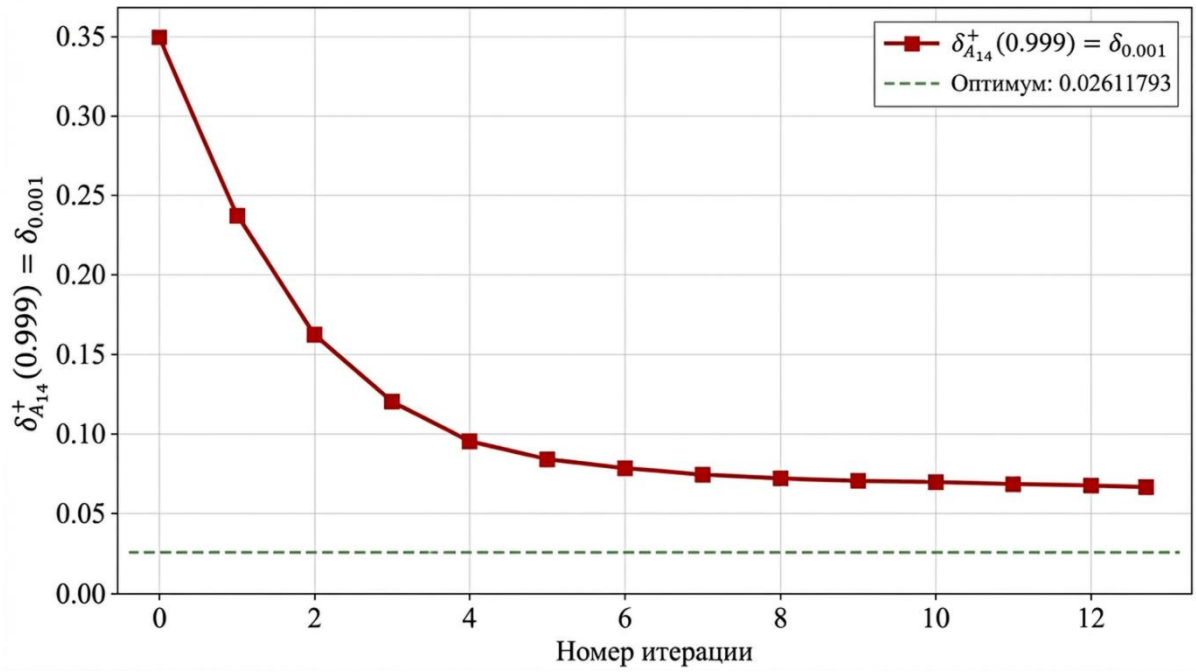


Рис. 3.13. Траектория поиска значения $\delta_{1-\alpha} = \delta_{0.001}$

На графике рис. 3.14 показана траектория, по которой значение $\delta_{0.001}$ движется по методу Ньютона от начальной точки к точке, соответствующей оптимальному квантилю. График наглядно показывает, как метод, начиная с грубой оценки, постепенно приближается к точке, где $\mathrm{P}\{d > \delta_{0.001}\} = \alpha$, корректируя $\delta_{0.001}$ на каждом шаге. Конечное значение $\delta_{0.001} \approx 0.0261$ мс означает, что для самого длинного пути сети с вероятностью 99.9 % пиковый возраст информации не превышает 26 микросекунд.

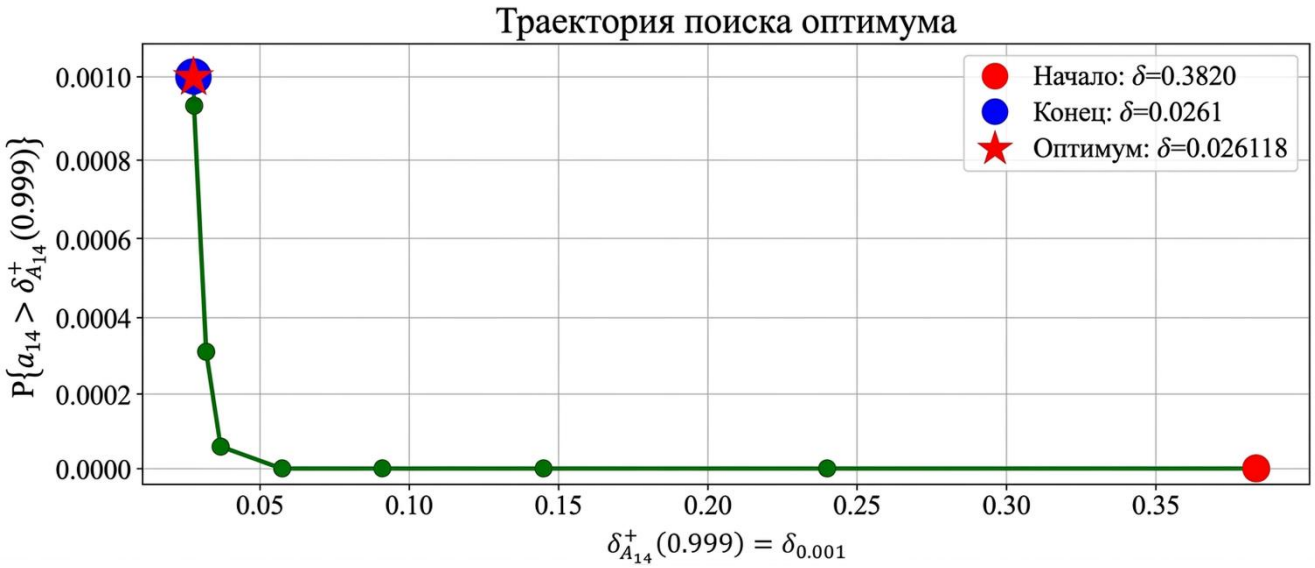


Рис. 3.14. Траектория поиска значения вероятности $1 - \alpha = 0.001$

Таким образом, разработанный алгоритм расчета оптимальных долей времени активности каналов сети IAB с разделением ресурсов по времени позволяет минимизировать среднюю по сети сквозную задержку и средний по сети пиковый возраст информации, а также вычислять правосторонний квантиль заданного уровня пикового возраста информации на маршруте от донора к абонентскому устройству для функции распределения сквозной задержки на маршруте сети IAB в виде распределения фазового типа.

Заключение

В заключение сформулируем основные результаты и выводы научно-квалификационной работы.

– Разработан метод разделения ресурсов между слайсами сети с нарезкой ресурса при динамическом выделении ресурса на основе максминной справедливости с учетом приоритизации слайсов, которая в отличие от известных методик предусматривает избыточное резервирование ресурсов типа «овербукинг».
– Построена математическая модель слайса услуги «Best Effort» в виде системы массового обслуживания с дисциплиной разделения процессора и эластичным трафиком с ограничением на максимальную скорость передачи. Предложено понятие деградации обслуживания в случае падения скорости передачи данных при предоставлении услуги абоненту ниже заданного порога, которое позволило получить аналитическое выражение для вероятности деградации обслуживания и сравнить показатели эффективности обслуживания абонентов для нескольких методов вызова процедуры нарезки ресурса – при поступлении нового запроса, по окончании получения услуги абонентом в сети, при обнаружении деградации обслуживания, при срабатывании регулярного таймера нарезки. Построенная модель позволяет провести сравнительный анализ влияния методов вызова процедуры нарезки ресурса на показатели эффективности обслуживания абонентов – вероятность деградации обслуживания, коэффициент использования ресурса и частоту вызова процедуры нарезки ресурса.
– Разработан алгоритм численного расчета оптимальных долей времени активности каналов сети IAB с разделением ресурса по времени, минимизирующих среднюю по сети сквозную задержку и средний пиковый возраст информации, а также алгоритм вычисления правостороннего квантиля заданного уровня пикового возраста информации на маршруте от донора к абонентскому устройству, при этом в отличие от известных результатов

случайная величина сквозной задержки на маршруте сети IAB имеет функцию распределения фазового типа.

Выражаю глубочайшую признательность нашему Учителю – заведующему кафедрой ТВиК РУДН, профессору К.Е. Самуйлову за мудрую поддержку и терпение, за совет включить в диссертацию «золотую» формулу правостороннего квантиля возраста информации в сети IAB. Я искренне благодарна к.ф.-м.н. Н.В. Яркиной за предложенную ей оригинальную концепцию формализации сетевого слайсинга, которая определила вектор моих исследований. Особая, теплая признательность – магистрантам кафедры ТВиК РУДН Антонине Паращенко и Никите Полякову, моим друзьям и единомышленникам, за их энтузиазм, живую дискуссию и бесценную помощь в проведении численных экспериментов.

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

$\mathcal{S}$ – Множество сетевых слайсов

$|\mathcal{S}|$ – Число сетевых слайсов

$C_{[\text{бит/с}]}$ – Общая емкость системы с нарезкой ресурсов

C_s – Емкость, выделяемая для слайса s, $s \in \mathcal{S}$

N_s – Число абонентов в слайсе s, $s \in \mathcal{S}$

$\mathbf{N}$ – Вектор числа абонентов во всех слайсах

R_s – Скорость передачи данных, предоставляемая абоненту в слайсе s, $s \in \mathcal{S}$

$R_s^{\min}$ – Минимальная допустимая скорость передачи данных для абонента слайса s, $s \in \mathcal{S}$

$R_s^{\max}$ – Максимальная эффективная скорость передачи данных для слайса s, $s \in \mathcal{S}$

N_s^{cont} – Предварительно согласованное («контрактное») число абонентов для слайса s, $s \in \mathcal{S}$

q_s – Отношение «контрактной» емкости слайса s к общей емкости системы с нарезкой ресурса, $s \in \mathcal{S}$

v_s – Приоритет слайса s, $s \in \mathcal{S}$

Ω – Пространство состояний системы с нарезкой ресурса

$\Omega^{\max}, \Omega^{\text{opt}}, \Omega^{\text{cong}}$ – Подмножества состояний, характеризующие режимы распределения ресурсов системы с нарезкой ресурса

$W_s(N_s)$ – Весовая функция для слайса s, $s \in \mathcal{S}$

$UTIL$ – Коэффициент использования ресурса в СМО

P^{deg} – Вероятность деградации обслуживания в СМО

N^{avg} – Среднее число заявок в СМО

R^{avg} – Средняя скорость обслуживания заявки в СМО

λ – Интенсивность поступления заявок в СМО

$\theta_{[\text{бит}]}$ – Средний размер заявки в СМО

M – Максимальное число заявок в СМО, которые можно обслужить с максимальной скоростью R^{max}

p_n – Стационарная вероятность состояния n (число заявок в СМО), $n \in \{0, 1, \dots, \infty\}$.

$\mathcal{E}$ – Множество каналов сети IAB

$\mathcal{P}$ – Множество многошаговых маршрутов (p-путей) в сети IAB от донора к каждому АУ

$\mathcal{E}_p$ – Множество каналов, составляющих p-путь, $p \in \mathcal{P}$

Λ_p – Интенсивность поступления пакетов на p-путь, $p \in \mathcal{P}$

$\boldsymbol{\lambda} = (\lambda_e)_{e \in \mathcal{E}}$ – Вектор интенсивностей поступления пакетов на e-канал

$C_{e\,[\text{пакет/ед.вр.}]}$ – Скорость передачи пакетов e-канале, $e \in \mathcal{E}$

$\mathbf{c} = (C_e)_{e \in \mathcal{E}}$ – Вектор емкостей каналов

$\mathbf{q} = (q_e)_{e \in \mathcal{E}}$ – Вектор долей времени активности канала

$\boldsymbol{\mu} = (\mu_e)_{e \in \mathcal{E}}$ – Интенсивность обслуживания пакета на e-канале

$\mathbf{F} = (f_{ne})_{n \in \mathcal{B}, e \in \mathcal{E}}$ – Матрица конфликтов

N – Число IAB-узлов в сети

N_n – Число секторов антенны для узла n, $n = 1, \dots, N$

$\mathbf{M}$ – Матрица топологии сети

t_{TTI} – Интервал передачи при мультиплексировании с временным разделением TDM

T – Цикл передачи при мультиплексировании с временным разделением TDM

D_p – Случайная величина сквозной задержки на p-пути, $p \in \mathcal{P}$

d_p – Среднее по времени значение сквозной задержки для p-пути, $p \in \mathcal{P}$

d – Среднее по путям значение сквозной задержки пакета в сети

$\delta_{D_p}^{+}(\alpha)$ – Правосторонний квантиль уровня α сквозной задержки на p-пути.

τ – Пороговое значение задержки для своевременной доставки

α – Минимальная доля пакетов, доставленных своевременно

$A(t)$ – Возраст информации донора на АУ в момент времени t

A_p – Случайная величина пикового возраста информации для p-пути

a_p – Среднее по времени значение пикового возраста информации на p-пути

a – Среднее по путям значение пикового возраста информации в сети

$\delta_{A_p}^{+}(\alpha)$ – Правосторонний квантиль уровня α пикового возраста информации для p-пути

Литература

1. Bega D. et al. A machine learning approach to 5G infrastructure market optimization //IEEE Transactions on Mobile Computing. – 2019. – Т. 19. – №. 3. – С. 498-512.

2. Yarkina N. et al. An analytical model for 5G network resource sharing with flexible SLA-oriented slice isolation //Mathematics. – 2020. – Т. 8. – №. 7. – С. 1177.

3. Kelly F. Charging and rate control for elastic traffic //European transactions on Telecommunications. – 1997. – Т. 8. – №. 1. – С. 33-37.

4. Bertsekas D., Gallager R. Data networks. – Athena Scientific, 2021.

5. Sadovaya Y. и др. Delay-aware link scheduling in IAB networks with dynamic user demands // IEEE Transactions on Vehicular Technology, 2024, 73(10), с. 15125-15139. DOI: 10.1109/TVT.2024.3397224.

6. Zhivtsova A. и др. A Survey of Delay-Oriented Dynamic Link Scheduling Policies for 5G/6G Integrated Access and Backhaul Systems // IEEE Access, 2024, 12, с. 118565–118586. DOI: 10.1109/ACCESS.2024.3446569.

7. Eftetahi P.S., Cai L., Ren X. Delay-Guaranteed Path Selection and Scheduling in IAB Networks // IEEE Transactions on Cognitive Communications and Networking, 2024.

8. Cho C.W., Pan M.S. Resource Scheduling in MU-MIMO and NOMA Enabled Integrated Access and Backhaul Networks // IEEE Open Journal of the Communications Society, 2025.

9. Yoon S.H., Lim B., Ko, Y.C. IRS-assisted interference mitigation for full-duplex IAB network // IEEE Wireless Communications Letters, 2024, 13(6), с. 1680-1684.

10. Linfu Z. и др. Joint beamforming and combining design for mmWave integrated access and backhaul networks // IEEE Open Journal of the Communications Society, 2024, 5, с. 503-513.

11. Hong Z.H. и др. Cancelling Adjacent Channel Interference for In-Band Full-Duplex Communications // 2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2024, с. 1-6.

12. Zhang J., Ratnarajah, T. Performance analysis of in-band-full-duplex multi-cell wideband IAB networks // IEEE Access, 2024, 12, с. 47024-47040.

13. Sambhwani S. и др. Extending mmWave deployment in the next-generation network: Coverage and reliability enhancements // IEEE Wireless Communications, 2025, 32(1), с. 83-89.

14. Singya P.K. и др. High-rate reliable communication using multi-hop and mesh THz/FSO networks // IEEE Open Journal of the Communications Society, 2024, 5, с. 3804-3823.

15. Kwon G. и др. Access-backhaul strategy via gNB cooperation for integrated terrestrial-satellite networks // IEEE Journal on Selected Areas in Communications, 2024, 42(5), с. 1403-1419.

16. Hu Z., Han C., Wang X. Deep reinforcement learning based cross-layer design in terahertz mesh backhaul networks // IEEE/ACM Transactions on Networking, 2023, 32(3), с. 2159-2173.

17. Belmekki B.E.Y., Alouini M.S. NOMA as the next-generation multiple access in nonterrestrial networks // Proceedings of the IEEE, 2024.

18. Gargari A.A. и др. Risk-averse learning for reliable mmwave self-backhauling // IEEE/ACM Transactions on Networking, 2024.

19. Morgado A.J. и др. Intelligent backhaul link selection for traffic offloading in B5G networks // IEEE Access, 2024.

20. Ullah, I. и др. Enhancing QoS in 6G networks through multi-IAB relaying strategies with Optimal Path Selection // IEEE Access, 2025.

21. Morgado A.J. и др. Intelligent backhaul link selection for traffic offloading in B5G networks // IEEE Access, 2024.

22. Pueyo J., Camps-Mur D., Catalan-Cid M. PHaul: A PPO-Based Forwarding Agent for Sub6 Enhanced Integrated Acess and Backhaul Networks // IEEE Transactions on Network and Service Management, 2024.

23. Yarkina N. и др. Coexistence of Multicast and Unicast Services in mmWave/sub-THz Self-Backhauled Systems: User Associations and Performance Gains // IEEE Transactions on Vehicular Technology, 2024.

24. Tafintsev N. и др. Analysis of duplexing patterns in multi-hop mmWave integrated access and backhaul systems // IEEE Open Journal of the Communications Society, 2024.

25. Mahmood A. и др. Analysis of terahertz (THz) frequency propagation and link design for federated learning in 6G wireless systems // IEEE Access, 2024, 12, с. 23782-23797.

26. Polese M., Giordani M., Zugno T., Roy A., Goyal S., Castor D., Zorzi M. Integrated Access and Backhaul in 5G mmWave Networks: Potential and Challenges // IEEE Communications Magazine, 2020, 58, с. 62–68. DOI: 10.1109/MCOM.001.1900570.

27. Gupta M., Roberts, I.P., Andrews, J.G. System-level analysis of full-duplex self-backhauled millimeter wave networks // IEEE Transactions on Wireless Communications, 2022, 22(2), с. 1130–1144. DOI: 10.1109/TWC.2022.3204874

28. Kosta A., Pappas N., Angelakis V. Age of Information: A New Concept, Metric, and Tool // Foundations and Trends® in Networking, 2017, 12(3), с. 162–259. DOI: 10.1561/1300000050.

29. Kaul S., Gruteser M. Rai V., Kenney J. Minimizing age of information in vehicular networks // 2011 8th Annual IEEE communications society conference on sensor, mesh and ad hoc communications and networks, 2011, с. 350-358.

30. Gaidamaka E.A., Platonova A.A., Parashchenko A.D., Zhivtsova A.A., Gaidamaka Yu.V. End-to-end Delay Analysis for an Integrated Access and Backhaul Network // Lecture Notes in Computer Science. – 2026. – Т. 16461. – С. 62-73.

31. Самуйлов А.К., Платонова А.А., Шоргин В.С., Гайдамака Ю.В. On modeling the effects of multicast traffic servicing in 5G NR networks [К моделированию эффектов обслуживания многоадресного трафика в сетях 5G NR] // Информатика и ее применения. – 2023. – Т. 17. – № 2. – С. 71-77.

32. Polyakov N., Platonova A. Characterizing the Effects of Base Station Variable Capacity on 5G Network Slicing Performance // Communications in Computer and Information Science. – 2023. – Т. 1748. – С. 135-146.

33. Polyakov N., Platonova A. Assessing latency of packet delivery in the 5G 3GPP integrated access and backhaul architecture with half-duplex constraints // Future Internet. – 2022. – Т. 14. – № 11. – С. 345.

34. Khayrov E. M., Prosvirov V. A., Platonova A.A. Traffic Arrival Model for Millimeter Wave 5G NR Systems // Lecture Notes in Computer Science. – 2022. – Т. 13766. – С. 161-175.

35. Бобрикова Е.В., Платонова А.А., Гайдамака Ю.В., Шоргин С.Я. Пример применения аппарата нейронных сетей при назначении модуляционно-кодовой схемы планировщиком базовой станции сети 5G // Системы и средства информатики. – 2021. – Т. 31. – № 3. – С. 135-143.

36. Гайдамака Е.А., Николаев Д.И., Платонова А.А., Самуйлов К.Е. Расчет задержки и пикового возраста информации пакетов в сети DECT с запланированным доступом // Свидетельство о государственной регистрации программы для ЭВМ № 2025669968 РФ. – Опубл. 01.08.2025.

37. Паращенко А.Д., Платонова А.А., Гайдамака Ю.В. Расчёт характеристик задержки в многошаговой сети // Свидетельство о государственной регистрации программы для ЭВМ № 2025680621 РФ. – Опубл. 07.08.2025.

38. Гайдамака Е.А., Николаев Д.И., Платонова А.А., Самуйлов К.Е. Расчет задержки и пикового возраста информации пакетов на листьевом узле сети IAB // Свидетельство о государственной регистрации программы для ЭВМ № 2025680762 РФ. – Опубл. 08.08.2025.

39. Паращенко А.Д., Платонова А.А., Гайдамака Ю.В. Анализ квантиля сквозной задержки и пикового возраста информации в сети интегрированного доступа и транзита // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем : материалы Всероссийской конференции с международным участием, Москва, 2025. – С. 57-62.

40. Гайдамака Е.А., Платонова А.А., Ким Р., Гайдамака Ю.В. К анализу возраста информации в одноуровневой сети с многоадресной доставкой информации // Новые информационные технологии в исследовании сложных структур : материалы Пятнадцатой международной конференции, Томск, 2024. – С. 70-72.

41. Stepanov M. S., Stepanov S. N., Andrabi U., Petrov D., Ndayikunda J. The Increasing of Resource Sharing Efficiency in Network Slicing Implementation // Communications in Computer and Information Science. – 2022. – Vol. 1552. – P. 18–35.

42. 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.

43. Stepanov M.S., Stepanov S.N., Andrabi U., Petrov D., and Ndayikunda J. The Increasing of Resource Sharing Efficiency in Network Slicing Implementation // Communications in Computer and Information Science. – 2022. – Vol. 1552. – P. 18–35.

44. Stepanov M. S., Stepanov S. N., Kanischeva M. G., Kroshin F. S. Analysis of Procedures to Ensure the Required QoS Indicators in Multiservice Access Nodes // Distributed computer and communication networks: control, computation, communications (DCCN-2023). – 2023. – P. 47–55.

45. Ateya A.A., Ahmed A. Abd El-Latif, A. Muthanna, A. Volkov, A. Koucheryavy. Enabling Metaverse and Telepresence Services in 6G Networks. River Publishers. 2025. ISBN: 9788770046732.

46. Abdellah A. R., Mahmood O. A. K., Paramonov A., Koucheryavy A. IoT traffic prediction using multi-step ahead prediction with neural network // 2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). – Dublin, Ireland, 2019. – Pp. 1–4. – DOI: 10.1109/ICUMT48472.2019.8970675.

47. Rouzbehani, B. A Service-Oriented Approach for Radio Resource Management in Virtual RANs [Текст] / B. Rouzbehani, L. M. Correia, L. Caeiro // Hindawi Wireless Communications and Mobile Computing. – 2018.

48. Salvat J. X. et al. Overbooking network slices through yield-driven end-to-end orchestration //Proceedings of the 14th international conference on emerging networking experiments and technologies. – 2018. – С. 353-365.

49. Gladkih, B. Metody optimizacii i issledovanie operacij dlya bakalavrov informatiki. CH. 2. Nelinejnoe i dinamicheskoe programmirovanie [Текст] / B. Gladkih. – Tomsk: Izd-vo NTL, 2011.

50. Башарин Г. П. Лекции по математической теории телетрафика. – Москва : РУДН, 2009. – 342 с.

51. Basharin G. P., Gaidamaka Y., Samouylov K. Mathematical theory of teletraffic and its application to the analysis of multiservice communication of next generation networks // Automatic Control and Computer Sciences. – 2013. – Vol. 47. – P. 62–69.

52. Назаров А. А., Моисеева С. П. Метод асимптотического анализа в теории массового обслуживания : монография. – Томск : Изд-во Научно-технической литературы, 2006. – 109 с.

53. Вишневский В. М. Теоретические основы проектирования компьютерных сетей. – Москва : Техносфера, 2003. – 512 с.

54. Степанов С. Н. Теория телетрафика концепции, модели, приложения. М.: Горячая линия - Телеком, 2015. – 867 с.

55. Назаров А. А., Терпугов А. Ф. Теория вероятностей и случайных процессов : учебное пособие. – Томск : Изд-во Научно-технической литературы, 2006. – 204 с.

56. Пшеничников А. П. Теория телетрафика : учебник для вузов. – Москва : Горячая линия-Телеком, 2020. – 212 с.

57. Яшков С.Ф. Математические вопросы теории систем обслуживания с разделением процессора. Итоги науки и техн. Сер. Теор. вероятн. Мат. стат. Теор. кибернет., 29, ВИНИТИ, М., 1990, 3–82.

58. Вишневский В. М., Дудин А. Н., Клименок В. И. Стохастические системы с корреляционными потоками. Теория и применение в телекоммуникационных сетях. – Москва : Техносфера, 2018. – 564 с.

59. Полин Е. П., Моисеева С. П., Моисеев А. Н. Применение отрицательного биномиального распределения для аппроксимации стационарного распределения числа заявок в СМО с входящим MAP-потоком, интенсивность которого зависит от состояния системы // Управление большими системами : сборник трудов. – 2024. – Вып. 108. – С. 40–56.

60. Moiseeva S. P., Turenova I. A., Imomov A. A. Asymptotic analysis of multi-arrival heterogenous resource queueing system MMPP/GI(2)/∞ in a random Markovian environment // Журнал Сибирского федерального университета. Серия: Математика и физика. – 2025. – Т. 18, № 6. – С. 770–781.

61. Моисеева С. П., Панкратова Е. В. Асимптотический анализ многопоточной гетерогенной СМО в условии предельно редких изменений состояний управляющей входящими потоками цепи Маркова // Управление большими системами. – 2024. – Вып. 112. – С. 30–44.

62. Wehrle K. Modeling and Tools for Network Simulation [Текст] / K. Wehrle,M. G…unes, J. Gross. – Springer, Berlin, Heidelberg, 2010.

63. ITU-R. Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface : Technical Report M.2410-0, 2017. URL: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf (дата обращения: 01.10.2022).

64. Jin H., Huang H., Su L., Nahrstedt K. Cost-Minimizing Mobile Access Point Deployment in Workflow-Based Mobile Sensor Networks // Proceedings of the 2014 IEEE 22nd International Conference on Network Protocols, Raleigh, NC, USA, 21–24 October 2014, 2014, с. 83–94. DOI: 10.1109/ICNP.2014.6970402.

65. Yang X., Lin H., Li Z., Qian F., Li X., He Z., Wu X., Wang X., Liu Y., Liao Z., [и др.]. Mobile access bandwidth in practice: Measurement, analysis, and implications // Proceedings of the ACM SIGCOMM 2022 Conference, Renton, WA, USA, 4–6 April 2022, 2022, с. 114–128.

66. Lopez A.V., Chervyakov A., Chance G., Verma S., Tang Y. Opportunities and Challenges of mmWave NR // IEEE Wireless Communications, 2019, 26, с. 4–6. DOI: 10.1109/MWC.001.1900060.

67. Скоробогатова С. А., Викулов А. С., Парамонов А. И. Анализ эволюционного развития сетей Wi-Fi за первую четверть 21 века // Труды учебных заведений связи. – 2025. – Т. 11, № 6. – С. 68–77. – DOI: 10.31854/1813-324X-2025-11-6-68-77.

68. Хоанг Ф. Н., Парамонов А. И. Метод балансировки задержки и потерь данных в гетерогенных сетях высокой плотности Интернета вещей // Вестник СПбГУТ. – 2025. – Т. 3, № 2. – С. 5.

69. Парамонов А. И., Бушеленков С. Н. Метод выбора маршрута в сети интернета вещей // Информационные технологии и телекоммуникации. – 2022. – Т. 10, № 1. – С. 34–44. – DOI: 10.31854/2307-1303-2022-10-1-34–44.

70. 3GPP TR 38.874 v16.0.0. NR; Study on Integrated Access and Backhaul : Technical Report. 2018. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3232 (дата обращения: 01.10.2022).

71. Akar N., Dogan O., Atay E. U., Finding the exact distribution of (Peak) age of information for queues of PH/PH/1/1 and M/PH/1/2 type, IEEE Trans. Commun., vol. 68, no. 9, pp. 5661–5672, Sep. 2020.

72. Zhang Y., Kishk M.A., Alouini M.S. Deployment optimization of tethered drone-assisted integrated access and backhaul networks // IEEE Transactions on Wireless Communications, 2023, 23(4), с. 2668-2680.

73. Lee Y. и др. D3QN-Based IAB Resource Allocation and Tethered UAV Positioning for IoT Networks // IEEE Transactions on Intelligent Transportation Systems, 2025.

74. Alavicheh R.G., Razavizadeh S.M., Yanikomeroglu H. Integrated Access and Backhaul (IAB) in Low Altitude Platforms // IEEE Open Journal of the Communications Society, 2024.

75. Zhang, H. и др. 5G network on Wings: A deep reinforcement learning approach to the UAV-based integrated access and backhaul // IEEE Transactions on Machine Learning in Communications and Networking, 2024.

76. Zhang Y., Kishk M.A., Alouini M.S. Energy-Efficient Optimization in Aerial IAB Networks for Emergency Communications // IEEE Transactions on Aerospace and Electronic Systems, 2024.

77. Shang, W., Friderikos V. Energy Efficient Optimization of In-Band Integrated Access and Backhaul Heterogeneous Networks // IEEE Transactions on Vehicular Technology, 2024.

78. Park C. и др. Aerial Reconfigurable Intelligent Surface-Assisted Integrated Access and Backhaul Networks // IEEE Wireless Communications Letters, 2025.

79. Tariq M.N. и др. Toward Optimal Resource Allocation: A Multi-Agent DRL Based Task Offloading Approach in Multi-UAV-Assisted MEC Networks // IEEE Access, 2024, 12, с. 81428-81440.

80. Lin H., Kishk M.A., Alouini M.S. Virtual Backhaul Connectivity for Enhanced Coverage in Fiber-Less Areas // IEEE Wireless Communications, 2024, 31(4), с. 324-330.

81. Sadovaya Y., Moltchanov D., Nikopour H., Yeh S.p., Mao W., Orhan O., Talwar S., Andreev S. Self-Interference Assessment and Mitigation in 3GPP IAB Deployments // Proceedings of the ICC 2021-IEEE International Conference on Communications, Montreal, QC, Canada, 14–23 June 2021, 2021, с. 1–6.

82. Pollaczek, F. Über eine Aufgabe der Wahrscheinlichkeitstheorie. I. Math Z 32, 64–100 (1930). https://doi.org/10.1007/BF01194620.

83. Хинчин А.Я. Математическая теория стационарной очереди. Матем. сб. 39, 1932, с.73–84.

84. Bocharov P.P., Naumov V.A. Matrix-geometric stationary distribution of PH|PH|1|R queue. RR-0304. INRIA. 1984

85. Nashiruddin M.I., Rahmawati P., Nugraha M.A., Akhmad A. Deployment of 5G NR at mmWave Frequency for Mobile Network in Indonesia’s Market // Proceedings of the 2021 2nd International Conference on ICT for Rural Development (IC-ICTRuDev), Virtual, 27–28 October 2021, 2021, с. 1–6. DOI: 10.1109/ICICTRUDEV53934.2021.9641427.