РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ ИМЕНИ ПАТРИСА ЛУМУМБЫ

На правах рукописи

Иванова Дарья Вадимовна

МОДЕЛИ СОВМЕСТНОГО ОБСЛУЖИВАНИЯ ТРАФИКА С ПРИОРИТИЗАЦИЕЙ И РАЗДЕЛЕНИЕМ РЕСУРСОВ В ПРОМЫШЛЕННОМ РАЗВЕРТЫВАНИИ МОБИЛЬНЫХ СЕТЕЙ

Специальность 1.2.3. Теоретическая информатика, кибернетика

Диссертация
на соискание ученой степени кандидата
физико-математических наук

Научный руководитель
кандидат физико-математических наук
доцент
Маркова Екатерина Викторовна

Москва – 2024

Оглавление

Введение

Актуальность темы исследования. Развитие индустриального интернета вещей (англ. Industrial Internet of Things, IIoT) неразрывно связано с продолжающейся на данный момент четвертой промышленной революцией – Индустрией 4.0. Диапазон вариантов использования интернета вещей (англ. Internet of Things, IoT) включает в себя умные транспортные системы, умные сети, здравоохранение, умные измерения, наблюдение за общественной безопасностью, удаленное производство, а также промышленную автоматизацию.

Автоматизация всех стадий производства – это необходимое условие для развития современной промышленности. Основными целями промышленной автоматизации являются повышение производительности, скорости и качества производства за счет более эффективного использования имеющихся экономических ресурсов, таких как оборудование, рабочая сила, сырье, капиталовложения и энергия. Помимо этого, активное внедрение новых технологий в производственный процесс должно позволить значительно снизить себестоимость продукции, а также максимально увеличить срок и надежность эксплуатации оборудования и технических сооружений. Реализация данного сценария предполагает выполнение всех функций контроля и управления на предприятии с помощью автоматических систем и приборов, что, в свою очередь, должно привести к большей конкурентоспособности предприятий.

Промышленная автоматизация является одним из наиболее важных сценариев использования беспроводных сетей пятого поколения (англ. Fifth Generation, 5G) и характеризуется крайне высокими требованиями к качеству обслуживания на беспроводном участке доступа [58]. Например, для управления подвижными элементами производственного оборудования системы, генерирующие низкоскоростной трафик, требуют сверхнадежной передачи данных с ультрамалой задержкой (англ. Ultra-Reliable Low Latency Communication, URLLC). В то же время, для систем видеонаблюдения и

позиционирования необходима поддержка усовершенствованной мобильной широкополосной связи (англ. enhanced Mobile Broadband, eMBB) [59]. Обеспечить условия для функционирования новых приложений, таких как управление оборудованием на основе технологий телеприсутствия, совместное использование мобильных роботов, позиционирование, а также сервисы дополненной реальности [60], должна технология 5G NR (англ. New Radio).

Таким образом, базовые станции (БС) 5G NR должны одновременно поддерживать типы трафика, предъявляющие принципиально разные требования к качеству обслуживания, в частности нетолерантный к задержкам потоковый трафик, соответствующий URLLC услугам и характеризующийся гарантированной скоростью передачи данных и фиксированным временем обслуживания, и эластичный трафик, соответствующий eMBB услугам, скорость передачи которого может меняться в зависимости от загрузки системы.

Ожидается, что методы совместного обслуживания трафика с принципиально разными требованиями будут предложены в процессе стандартизации сетей 5G-Advanced в течение следующих пяти лет. В связи с этим, реализация одновременной поддержки URLLC и eMBB услуг, генерирующих кардинально отличающиеся типы трафика в беспроводных сетях 5G, является сложной и актуальной проблемой, требующей разработки новых подходов, алгоритмов и моделей обслуживания такого трафика. Следует отметить, что исследований в этом направлении проведено достаточно мало. В частности, остаётся неисследованным вопрос негативного влияния условий развертывания сети на процесс обслуживания трафика, нет четкого представления о методах разделения ресурсов на беспроводном участке доступа (приоритет, резервирование ресурсов), а также о том, какой из механизмов обслуживания (передача данных по протоколу неортогонального множественного доступа, прямая связь между устройствами, частичная адаптация скорости устройств к доступным

ресурсам системы) может позволить уменьшить нагрузку на систему и повысить эффективность обслуживания URLLC трафика в присутствии eMBB трафика. Кроме того, мало изучены системы массового обслуживания (СМО), позволяющие оценить показатели эффективности для различных подходов и на их основе предложить алгоритмы совместного обслуживания.

С учетом вышеизложенного, диссертационная работа посвящена исследованию трех основных актуальных направлений: 1) анализ моделей совместного обслуживания потокового и эластичного трафика, 2) анализ моделей схем одновременного предоставления услуг с реализацией абсолютного приоритета, 3) анализ моделей мультисервисной системы с резервированием индивидуальных зон и прерыванием обслуживания.

Степень разработанности темы. Исследованиям беспроводных сетей и построению моделей систем массового обслуживания посвящены работы ведущих российских и зарубежных ученых и специалистов: Башарин Г.П. [61, 62], Гайдамака Ю.В. [36, 63, 64, 65], Горшенин А.К. [66, 67, 68], Кучерявый А.Е. [69, 70, 71], Кучерявый Е.А. [10, 47, 51, 69, 72], Молчанов Д.А. [10, 35, 36, 39, 51, 56, 73], Мутханна А.А. [74, 75, 76], Наумов В.А. [50, 77, 78], Самуйлов К.Е. [63, 64, 77, 78, 79], Степанов С.Н. [46, 80, 81], Andrews J.G. [82, 83], Correia L.M. [65, 84, 85], Logothetis M.D. [86], Malanchini I. [87, 88], Pagano M. [89]. Кроме того, в настоящее время широко исследуются механизмы поддержки URLLC или eMBB услуг в отдельности на базовых станциях 5G NR [8, 9, 10, 11, 12, 13].

Для анализа беспроводных сетей применяются методы математического моделирования, теории массового обслуживания, математической теории телетрафика, а также теории вероятностей и теории случайных процессов. Значительный вклад в развитие данных областей внесли такие ученые, как Башарин Г.П. [49, 61, 90, 91, 92, 93], Бочаров П.П. [91, 94, 95], Вишневский В.М. [96, 97, 98, 99], Гайдамака Ю.В. [49, 100], Зейфман А.И. [101, 102], Дудин А.Н. [96, 97, 103], Моисеев А.Н. [104, 105, 106], Моисеева С.П. [107, 108], Назаров А.А. [48, 105, 107, 109], Наумов В.А.

[110, 111], Печинкин А.В. [94, 95], Пшеничников А.П. [112, 113], Рыков В.В. [114, 115, 116], Самуйлов К.Е. [49, 92, 100, 111, 117], Степанов С.Н. [118, 119, 120, 121], Цитович И.И. [121, 122], Шоргин С.Я. [64, 79, 123], Dohler M. [124, 125], Iversen V.B. [126, 127], Kelly F.P. [128], Roberts J.W. [129], Ross K.W. [130] и др.

Целью диссертационной работы является разработка марковских моделей совместного обслуживания различных видов трафика в промышленных развертываниях мобильных сетей, а также численный анализ показателей эффективности исследуемых систем.

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

1. Разработка и анализ марковских моделей схем одновременного предоставления услуг с прерыванием обслуживания в условиях промышленного развертывания беспроводных сетей связи пятого поколения.
2. Разработка модели мультисервисной системы массового обслуживания с реализацией механизмов прерывания обслуживания менее приоритетного трафика, реализация алгоритма выбора запросов, обслуживание которых должно быть прервано при поступлении более приоритетного запроса.
3. Разработка и анализ марковских моделей схем доступа с резервированием индивидуальных зон как без прерывания обслуживания, так и с прерыванием обслуживания менее приоритетного трафика.

Научная новизна диссертационной работы:

1. Построенная модель схемы одновременного предоставления услуг с реализацией явного приоритета, в отличие от исследуемых ранее моделей, учитывает особенности совместного обслуживания различных типов трафика в промышленных развертываниях

беспроводных сетей, а также динамическую блокировку в процессе передачи данных между устройствами.

2. Для модели схемы доступа с резервированием индивидуальных зон, помимо стандартного способа получения стационарного распределения вероятностей путем численного решения системы уравнений равновесия, предложено решение в мультипликативном виде.
3. Построенная модель мультисервисной системы массового обслуживания позволяет провести сравнение различных вариантов стратегий, основанных на приоритетах и резервировании, с точки зрения производительности, ориентированной на пользователя и оператора.
4. Для модели мультисервисной системы с прерыванием обслуживания реализован алгоритм выбора запросов, обслуживание которых должно быть прервано при поступлении более приоритетного запроса. В рамках сравнительного анализа моделей проведена численная оптимизация параметров для обеспечения гарантий качества обслуживания.

Теоретическая и практическая значимость работы. Полученные в ходе диссертационного исследования результаты могут использоваться операторами для развертывания и эксплуатации сетей 5G NR, а также обеспечения гарантированного уровня качества обслуживания трафика.

Разработанные математические модели и программные комплексы могут быть применены для управления ресурсами беспроводных сетей, расчета показателей эффективности и оценки производительности развертывания сетей 5G NR.

Методы исследования. В диссертационной работе применяются методы теории вероятностей, теории массового обслуживания, математической теории телетрафика и статистического моделирования.

Положения, выносимые на защиту:

1. Модель одновременной передачи потокового и эластичного трафика с приоритетами и снижением скорости обслуживания позволяет рассчитать вероятностные характеристики и оценить эффективность применяемой стратегии по качеству обслуживания и использованию ресурсов.
2. Модель промышленного развертывания беспроводных сетей связи пятого поколения применима для выбора оптимальной стратегии передачи двух типов трафика, позволяющей избежать создания дополнительной интерференции и демонстрирующей лучшие характеристики по сравнению с другими стратегиями.
3. Модель совместного обслуживания произвольного числа типов трафика с алгоритмом численной оптимизации параметров, определяющих минимальное число единиц ресурса для удовлетворения требований к качеству обслуживания, позволяет анализировать различные схемы разделения ресурсов с точки зрения вероятности блокировки, вероятности прерывания обслуживания, а также коэффициента использования ресурсов.

Степень достоверности и апробация результатов. Основные результаты, полученные в ходе диссертационного исследования по данной теме, докладывались на научных конференциях:

- международная конференция «Distributed Computer and Communication Networks: Control, Computation, Communications» (г. Москва, 2020);
- международная конференция «12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops» (г. Брно, 2020).

Основные результаты опубликованы в ведущих научных журналах: Mathematics, Future Internet, IEEE Access, Lecture Notes in Computer Science, Информатика и ее применения, Вестник Томского государственного университета, а также в трудах международных конференций, индексируемых в Web of Science и Scopus.

Реализация результатов работы. Результаты диссертационной работы включены в исследования в рамках гранта РФФИ № 20-37-70079 «Исследование и разработка моделей и интеллектуальных алгоритмов совместного обслуживания трафика с малыми задержками и широкополосного доступа в беспроводных сетях пятого поколения», а также гранта 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+».

Публикации. Основные результаты, изложенные в диссертационной работе, опубликованы в 7 печатных изданиях [24, 25, 40, 131, 132, 133, 134], входящих в базу данных Scopus/Web of Science, в 2 свидетельствах о государственной регистрации программ для ЭВМ [135, 136].

Соответствие паспорту специальности. Диссертационное исследование соответствует следующим разделам паспорта специальности 1.2.3. «Теоретическая информатика, кибернетика»: **п. 9** «Математическая теория исследования операций» в части исследования и разработки моделей систем массового обслуживания; **п. 11** «Распределенные многопользовательские системы» в части моделирования разделения ресурсов сети между различными типами трафика; **п. 12** «Модели информационных процессов и структур» в части моделирования схемы управления доступом к радиоресурсам сети.

Личный вклад. Все результаты диссертационного исследования, выносимые на защиту, получены автором лично, построение представленных в работе моделей и их численный анализ проведены автором самостоятельно. Программные средства, используемые для численного анализа, разработаны при непосредственном участии автора.

Объем и структура работы. Структура диссертационной работы включает в себя введение, три главы, заключение и список литературы из 136 источников. Диссертационная работа изложена на 119 страницах текста, содержит 40 рисунков и 6 таблиц.

Краткое изложение диссертации. Диссертационная работа состоит из трех глав. **В первой главе** исследуются сценарии использования беспроводных сетей пятого поколения в промышленной автоматизации, а также модели совместного обслуживания различных типов трафика. Раздел 1.1 посвящен особенностям развертывания беспроводных сетей пятого поколения в промышленной среде. В разделе 1.2 рассмотрена модель схемы одновременного предоставления услуг на основе приоритетов со снижением скорости обслуживания. В разделе 1.3 проведен расчет вероятностных характеристик модели и их численный анализ. В разделе 1.4 ставится задача исследований диссертационной работы, представляющая собой разработку моделей совместного обслуживания различных типов трафика в промышленных развертываниях беспроводных сетей, а также численный анализ показателей эффективности этих моделей.

Во второй главе разрабатывается и анализируется модель промышленного развертывания беспроводных сетей с приоритетным обслуживанием на базовой станции с прямой передачей между устройствами. В разделе 2.1 представлена системная модель, а также рассмотрены три стратегии одновременного предоставления услуг двух типов с использованием явного приоритета. В разделе 2.2 построена математическая модель, получено распределение вероятностей системы, выведены формулы для расчета вероятностных характеристик. В разделе 2.3 выполнен расчет характеристик модели в зависимости от плотности развертывания базовых станций, плотности размещения оборудования и других параметров системы, проведен сравнительный анализ рассмотренных стратегий.

Третья глава посвящена разработке и анализу моделей мультисервисной СМО с резервированием индивидуальных зон и приоритизацией. В разделе 3.1 изложена специфика разделения ресурсов в беспроводных сетях 5G, а также построена модель схемы доступа с прерыванием обслуживания для произвольного числа типов услуг, выполнен

расчет показателей эффективности для примера модели совместного обслуживания трех типов услуг. В разделе 3.2 рассмотрен частный случай модели с резервированием без прерывания обслуживания пользователей услуги с меньшим приоритетом. В разделе 3.3 рассмотрен частный случай модели с резервированием и прерыванием обслуживания пользователей менее приоритетной услуги. В разделе 3.4 представлены численные результаты сравнительного анализа различных стратегий разделения ресурсов для рассмотренных моделей, а также проведена численная оптимизация параметров для обеспечения гарантий производительности системы.

В **заключении** представлены основные результаты диссертационного исследования.

Глава 1. Построение и анализ модели обслуживания трафика в промышленных развертываниях сетей пятого поколения

1.1. Особенности развертывания мобильных сетей в условиях промышленной среды

Продолжающаяся четвертая промышленная революция или Индустрия 4.0 способствует стремительному развитию сферы интернета вещей (англ. Internet of Things, IoT). Согласно данным международных аналитических компаний, число устройств машинного типа, обслуживаемых в беспроводных сетях, уже в 2021 году значительно превысило количество классических абонентов сетей связи [1]. Варианты использования интернета вещей достаточно разнообразны и включают в себя умные транспортные системы, умные здравоохранение, промышленную автоматизацию, а также удаленное производство [2].

Развертывание беспроводных сетей пятого поколения (англ. Fifth Generation, 5G) в промышленной автоматизации является важным условием для улучшения качества, повышения производительности и эффективности производства путем оптимизации использования имеющихся экономических ресурсов [3]. Автоматизация должна позволить существенно снизить себестоимость продукции и увеличить срок эксплуатации оборудования. При этом функции контроля и управления на предприятии должны стать автоматизированными и осуществляться с помощью автоматических систем и приборов, что, в свою очередь, должно повысить конкурентоспособность предприятия.

В сетях пятого поколения определены три основных типа услуг (рис. 1.1) [4]. Первый тип услуг требует использования усовершенствованной мобильной широкополосной связи (англ. enhanced Mobile Broadband, eMBB), которая увеличивает пропускную способность сети. Второй тип услуг требует сверхнадежной связи с низкими задержками (англ. Ultra-Reliable Low

Latency Communication, URLLC). Третий тип услуг использует массовую межмашинную связь (англ. massive machine-type communications, mMTC), которая позволяет сетям пятого поколения поддерживать большое количество одновременно подключенных устройств. В то время как услуги третьего типа уже доступны на рынке благодаря стандарту сотовой связи для устройств телеметрии, который был утвержден 3GPP-консорциумом (англ. Third Generation Partnership Project) в 2016 году и основан на LTE (англ. Long-Term Evolution), ожидается, что услуги первого и второго типа будут поддерживаться с помощью активно развивающейся технологии 5G NR (англ. New Radio) [5].

Рис. 1.1. Услуги 5G NR.

Данная технология позволяет применять в области промышленной автоматизации совместное управление роботизированными устройствами, приложения на основе технологий телеприсутствия и дополненной реальности (англ. augmented reality, AR), а также другие новые сервисы. Беспроводная сеть должна заменить устаревшие проводные соединения [6]. Это особенно важно для снижения затрат на ранней стадии перехода современных производств к Индустрии 4.0. Устаревшие производственные

машины должны быть интегрированы в развивающиеся беспроводные сети там, где это возможно. Согласно стандарту 3GPP [7], основными нишами для беспроводной сети на производстве являются: управление движением и связь между системами управления; мобильные роботизированные платформы; мониторинг ресурсов и процессов; человеко-машинный интерфейс. При этом системы управления движением отвечают за движущиеся части машин и, как правило, генерируют низкоскоростной трафик, но при этом требуют сверхнадежной передачи данных с малой задержкой, что соответствует трафику URLLC. Системы мониторинга ресурсов и операций зависят от большого количества сенсорных устройств, установленных на предприятии, и предоставляют информацию о текущих процессах. Помимо выполнения измерений и дистанционного наблюдения, датчики могут также использоваться системами управления средой, например, для теплового видеонаблюдения, что влечет за собой использование усовершенствованной подвижной широкополосной связи eMBB.

Внедрение новых технологий на производстве требует обеспечения одновременной поддержки потокового URLLC трафика и эластичного eMBB трафика. Механизмы отдельной поддержки eMBB [8, 9, 10] или URLLC [11, 12, 13] услуг на базовых станциях (БС) NR в миллиметровом диапазоне (англ. millimeter waves, mmWave) широко исследуются в настоящее время. Однако, принципиальные различия в требованиях к качеству обслуживания, предъявляемых услугами URLLC и eMBB, например, ограничения к задержкам передачи данных для URLLC услуг в 1 мс и вероятность потери информации 10^{-5}, а также высокая скорость для eMBB услуг, на данный момент усложняют их одновременное предоставление в сетях 5G NR. Ожидается, что способы совместного обслуживания трафика с кардинально различающимися требованиями будут разработаны в ходе стандартизации сетей 5G-Advanced. Для успешного развертывания автоматизации в промышленности необходимо создание новых алгоритмов и моделей

одновременного обслуживания таких типов трафика на беспроводном интерфейсе 5G+ NR.

В текущих исследованиях, посвященных одновременной поддержке URLLC и eMBB трафика, был рассмотрен ряд подходов. Авторы [14] предложили решение проблемы мультиплексирования URLLC и eMBB трафика в восходящем канале связи с помощью схемы, основанной на неортогональном множественном доступе (англ. non-orthogonal multiple access, NOMA). В соответствии со схемой был разработан алгоритм планирования для сервисов eMBB, учитывающий ограничения, налагаемые URLLC трафиком. Оценка производительности предложенной схемы показала возможность увеличения пропускной способности при соблюдении строгих требований к качеству обслуживания (англ. quality of service, QoS) для URLLC трафика. Подобные методы также рассматривались в [15, 16, 17, 18], однако в данном случае авторы предлагают использовать сетевой слайсинг для удовлетворения требований всех типов трафика, обеспечения гарантий производительности и изоляции. В частности, в [15] авторы предложили использовать NOMA для увеличения количества устройств, генерирующих URLLC трафик, обслуживаемый одной базовой станцией, как для ортогонального, так и для неортогонального разделения сети с устройствами, генерирующими eMBB трафик. В работе [16] авторы рассмотрели потенциальные преимущества неортогонального совместного использования ресурсов сети радиодоступа (англ. Radio Access Network, RAN) сервисами URLLC, eMBB и mMTC. Явным преимуществом подхода NOMA является то, что передача критичных к задержке данных может быть запланирована немедленно, в том же временном интервале. Однако такой подход требует разработки сложного механизма упреждающей коррекции ошибок в канале передачи и выбора соответствующих ресурсных блоков.

Помимо NOMA в исследованиях был предложен метод резервирования ресурсов для одновременной поддержки URLLC и eMBB трафика через интерфейс NR [19]. В частности, в [20] авторы предложили резервировать

ресурсы случайного доступа для URLLC трафика во время первоначальной передачи запросов случайного доступа. С этой целью был разработан усовершенствованный механизм случайного доступа для различных политик резервирования, применение которого позволяет добиться ограничения задержки в 10 мс с вероятностью 95%. Однако данный подход может привести к неэффективному использованию ресурсов, поскольку интенсивность поступающего трафика может быть неизвестна заранее. Чтобы повысить эффективность механизма резервирования ресурсов, необходимо разработать легкий и точный алгоритм прогнозирования трафика, который бы динамически изменял количество ресурсов, выделяемых типам трафика, на временном интервале планирования, который составляет 1 мс для NR [21]. В [22] была предложена схема упреждающего резервирования ресурсов. Чтобы снизить негативное влияние на производительность eMBB трафика, авторы применили алгоритм прогнозирования траектории движения транспортного средства и продемонстрировали, что ограничение резервирования меньшим количеством ресурсов позволяет достичь требуемого уровня качества обслуживания URLLC трафика с меньшим воздействием на пропускную способность сети. В [23] авторы подробно изучили схемы планирования, в частности, планирование на основе резервирования, а именно полустатическое и динамическое резервирование, чтобы соответствовать ключевым требованиям URLLC трафика. Их результаты показали, что динамическое резервирование превосходит полустатическое резервирование с точки зрения задержки из-за быстрой адаптации к распределению ресурсов.

В качестве альтернативы для динамического распределения радиоресурсов между типами трафика можно также использовать планирование на основе приоритетов. В случае перегрузки при увеличении интенсивности поступления трафика один или несколько запросов на передачу трафика менее приоритетного типа могут быть прерваны. В предыдущих исследованиях [24], [25] разработана модель для оценки

эффективности этой стратегии и продемонстрировано, что четкая расстановка приоритетов действительно может привести к почти идеальному использованию ресурсов, обеспечивая при этом требуемые гарантии производительности URLLC трафика. Этот подход также был исследован в [26], где моделирование на системном и канальном уровнях продемонстрировало преимущества данного механизма с точки зрения производительности. Рассмотренные авторами методы эффективно снижают задержку, но могут не адаптироваться к динамически изменяющейся интенсивности трафика. В работе [27] авторы рассмотрели сценарий последовательного планирования для вытеснения запросов на передачу eMBB трафика в следующем временном интервале. Проведенное имитационное моделирование показало эффективность метода динамического планирования.

Планирование ресурсов для передачи URLLC трафика рассмотрено в [28]. Основной целью работы была формализация и решение задачи оптимизации, направленной на максимизацию скорости передачи eMBB трафика, с учетом требований к надежности, предъявляемых URLLC трафиком. Авторы рассмотрели два временных интервала: временной интервал, в котором выполняется распределение ресурсов для передачи eMBB трафика, и интервал, в котором планируется передача URLLC трафика. Решая в дальнейшем проблему оптимизации в [29], авторы предложили программный комплекс, основанный на глубоком обучении с подкреплением (англ. deep reinforcement learning, DRL) и включающий в себя фазу распределения ресурсов для eMBB трафика и фазу планирования для URLLC трафика. Авторы предложили приближенное решение для распределения ресурсов и подтвердили свои результаты с помощью компьютерного моделирования. В то же время, в работе [30] рассматривалась проблема минимизации риска, связанного с задержкой и надежностью передачи URLLC трафика. Авторы предложили чувствительный к риску

метод распределения ресурсов для сервисов, генерирующих URLLC трафик, и рассмотрели условное значение риска в качестве интересующей метрики.

Стоит отметить, что, несмотря на значительное количество проведенных к настоящему времени исследований, особенности одновременной поддержки URLLC и eMBB трафика в сценариях промышленной автоматизации рассмотрены не были. Таким образом, актуальными являются вопросы влияния условий развертывания сети (на заводах, фабриках) на процесс обслуживания трафика, в частности, на интенсивность прерывания соединений в результате динамической блокировки путей распространения сигнала, микромобильности оконечных устройств, вызываемой резкими смещениями подвижных частей машин, на которых могут быть установлены датчики, а также возникновения внешних помех, вызываемых электро-механическими приборами, работающими на производственных конвейерах. Для достижения границы в 1 мс следует определить методы использования радиоинтерфейса, которые позволят получить максимальный выигрыш по задержке, в частности технологии прямой связи между устройствами (англ. device-to-device, D2D) или же передачи согласно методам неортогонального множественного доступа. Недостаточно исследованными также остаются вопросы о типе изоляции трафика в пределах концепции нарезки радиоресурсов [31, 32, 33, 34].

Учитывая вышесказанное, можно заключить, что разработка моделей и методов одновременной поддержки URLLC и eMBB услуг, генерирующих принципиально разные типы трафика в беспроводных сетях, является сложной и актуальной проблемой, требующей разработки новых методов обслуживания такого трафика как на канальном, так и системных уровнях.

1.2. Модель обслуживания широкополосного трафика и трафика с малыми задержками

Рассмотрим базовую станцию NR, обслуживающую трафик двух типов: потоковый, соответствующий услугам URLLC, и эластичный,

соответствующий услугам eMBB. Примером такого сценария может служить одновременная поддержка прямой связи между производственным оборудованием и системами удаленного мониторинга в промышленной среде (рис. 1.2). Предполагается, что трафик URLLC характеризуется чрезвычайно малой продолжительностью обслуживания, в то время как трафик eMBB, характеризующийся большей продолжительностью обслуживания, является более гибким, адаптирующимся к изменениям параметров сети путем регулирования скорости передачи.

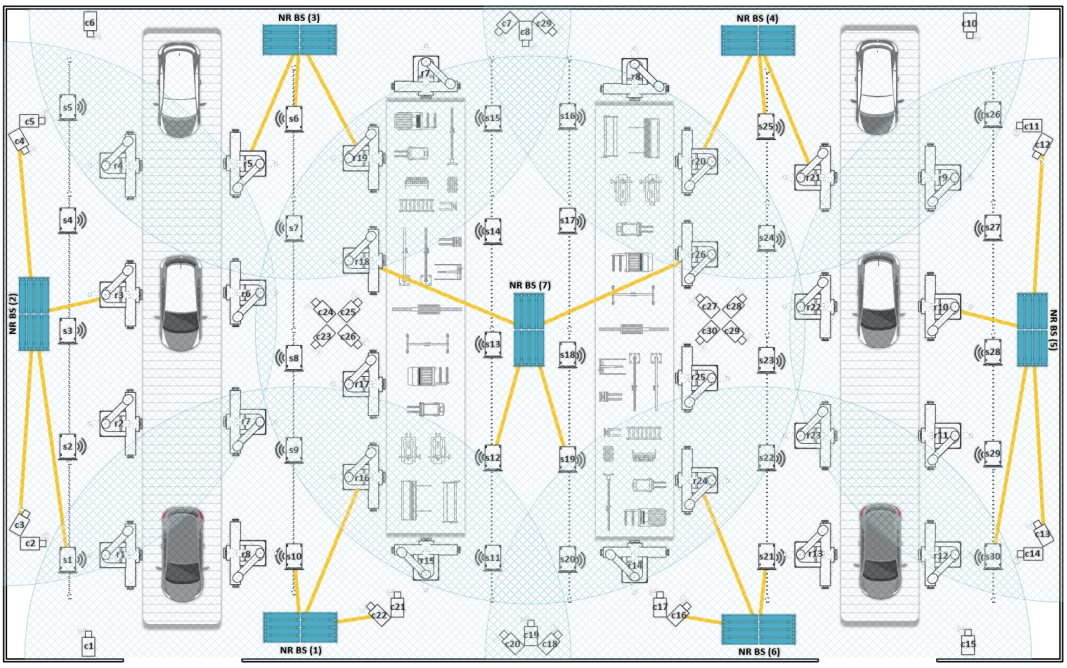

Рис. 1.2. Развертывание сетей 5G в промышленности.

Опишем модель с помощью системы массового обслуживания (СМО). Рассмотрим систему связи емкостью C каналов, в которой пользователям предоставляются услуги двух типов: услуга, генерирующая потоковый трафик, и услуга, генерирующая эластичный трафик. Пусть запросы на передачу потокового трафика соответствуют запросам первого типа, а запросы на передачу эластичного трафика – запросам второго типа. Запросы на предоставление услуг представляют собой пуассоновский поток с интенсивностями λ_1 и λ_2. Среднее время обслуживания запросов на

предоставление услуг, генерирующих потоковый и эластичный трафик, – μ_1^{-1} и μ_2^{-1} соответственно.

Требование к числу базовых цифровых каналов (БЦК), необходимых для предоставления услуги, генерирующей потоковый трафик, равно b_1, $b_1 \geq 1$. Рассмотрим две стратегии обслуживания eMBB трафика:

1. Фиксированная стратегия, при которой минимальное число БЦК, требуемых для обслуживания запроса на передачу эластичного трафика, равно $b_2^{\min}$, $b_2^{\min} = b_2^{\min_1} = b_2^{\min_2} \geq 1$. В случае недостаточности ресурсов для передачи потокового трафика обслуживание одного или нескольких запросов на передачу эластичного трафика, может быть прекращено.

2. Гибкая стратегия, при которой минимальное число БЦК, требуемых для обслуживания запроса на передачу эластичного трафика, равно $b_2^{\min_1}$, $b_2^{\min_1} \geq 1$, а управление доступом к ресурсам сети построено таким образом, что при недостаточности ресурсов для передачи потокового трафика число БЦК, требуемых для обслуживания запроса на передачу эластичного трафика, может быть снижено до порогового значения $b_2^{\min_2}$, $b_2^{\min_2} < b_2^{\min_1}$. При падении количества доступных ресурсов ниже порогового значения обслуживание одного или нескольких запросов может быть прекращено (рис. 1.3).

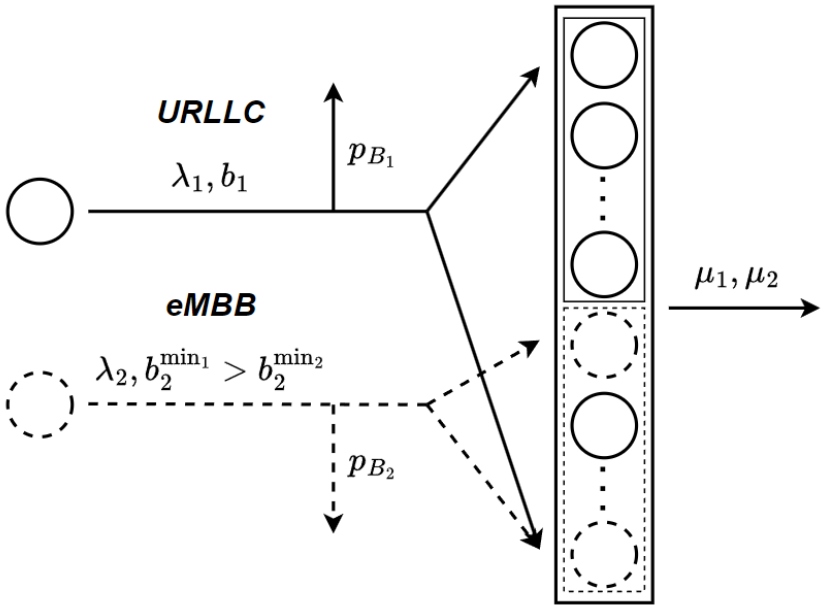

Рис. 1.3. Схема модели с приоритетным обслуживанием потокового трафика.

Обозначим максимальное число запросов первого типа, обслуживаемых в системе, $N_1 = \left\lfloor \frac{C}{b_1} \right\rfloor$, запросов второго типа – $N_2 = \left\lfloor \frac{C}{b_2^{\min_1}} \right\rfloor$. Функционирование рассматриваемой системы описывает двумерный марковский случайный процесс (СП) $\left\{ \left(N_1(t), N_2(t) \right), t \geq 0 \right\}$, где $N_1(t)$ – число обслуживаемых системой запросов первого типа, а $N_2(t)$ – число обслуживаемых системой запросов второго типа в момент времени t.

Состояние системы описывает двумерный вектор (n_1, n_2), где $n_1 = \{0, ..., N_1\}$ – число обслуживаемых системой запросов на предоставление услуги, генерирующей потоковый трафик, – запросов первого типа, $n_2 = \{0, ..., N_2\}$ – число обслуживаемых системой запросов на предоставление услуги, генерирующей эластичный трафик, – запросов второго типа. Тогда пространство состояний системы имеет вид

$$\mathbf{X} = \left\{ (n_1, n_2) : n_1 \geq 0,\ n_2 \geq 0,\ n_2 \leq N_2,\ n_1 b_1 + n_2 b_2^{\min_2} \leq C \right\}. \quad (1.1)$$

Число БЦК $b_2(n_1,n_2)$, выделяемое при обслуживании запросов второго типа, может меняться в зависимости от состояния системы и определяется как

$$b_2(n_1,n_2)=\left\lfloor \frac{C-n_1b_1}{n_2} \right\rfloor \geq b_2^{\min_2}. \quad (1.2)$$

Обозначим максимальное число запросов на предоставление услуги второго типа при условии, что в системе уже обслуживается n_1 запрос на предоставление услуги первого типа, как

$$k(n_1)=\left\lfloor \frac{C-n_1b_1}{b_2^{\min_1}} \right\rfloor. \quad (1.3)$$

Максимальное число запросов второго типа, которые могут быть обслужены со сниженным требованием к числу БЦК $b_2^{\min_2}$, при условии, что в системе уже обслуживается n_1 запрос первого типа, обозначим как

$$l(n_1)=\min\left(N_2,\left\lfloor \frac{C-n_1b_1}{b_2^{\min_2}} \right\rfloor \right). \quad (1.4)$$

Сформулируем правила приема и обслуживания запросов на предоставление услуги, генерирующей потоковый трафик:

- если в системе число свободных каналов больше или равно b_1, то запрос первого типа будет принят на обслуживание;
- если в системе число свободных каналов меньше b_1, $n_1<N_1$, $n_2>0$ и $b_2(n_1+1,n_2)\geq b_2^{\min_2}$, то запрос первого типа будет принят на обслуживание, а требование к числу БЦК для обслуживания запросов второго типа снизится до $\left\lfloor \frac{C-(n_1+1)b_1}{n_2} \right\rfloor$;
- если в системе число свободных каналов меньше b_1, $n_1<N_1$, $n_2>0$ и $b_2(n_1+1,n_2)<b_2^{\min_2}$, то запрос первого типа будет принят на

обслуживание за счет прерывания обслуживания $\left\lfloor \frac{b_1}{b_2(n_1, n_2)} \right\rfloor$ запросов второго типа;

- в иных случаях запрос первого типа будет заблокирован.

Сформулируем правила приема и обслуживания запросов на предоставление услуги, генерирующей эластичный трафик:

- если в системе число свободных каналов больше или равно $b_2^{\min_1}$, то запрос второго типа будет принят на обслуживание;
- в иных случаях запрос второго типа будет заблокирован.

С учетом изложенных правил приема и обслуживания запросов составлена диаграмма интенсивностей переходов, проиллюстрированная на рис. 1.4 в общем виде и на рис. 1.5 для центрального состояния.

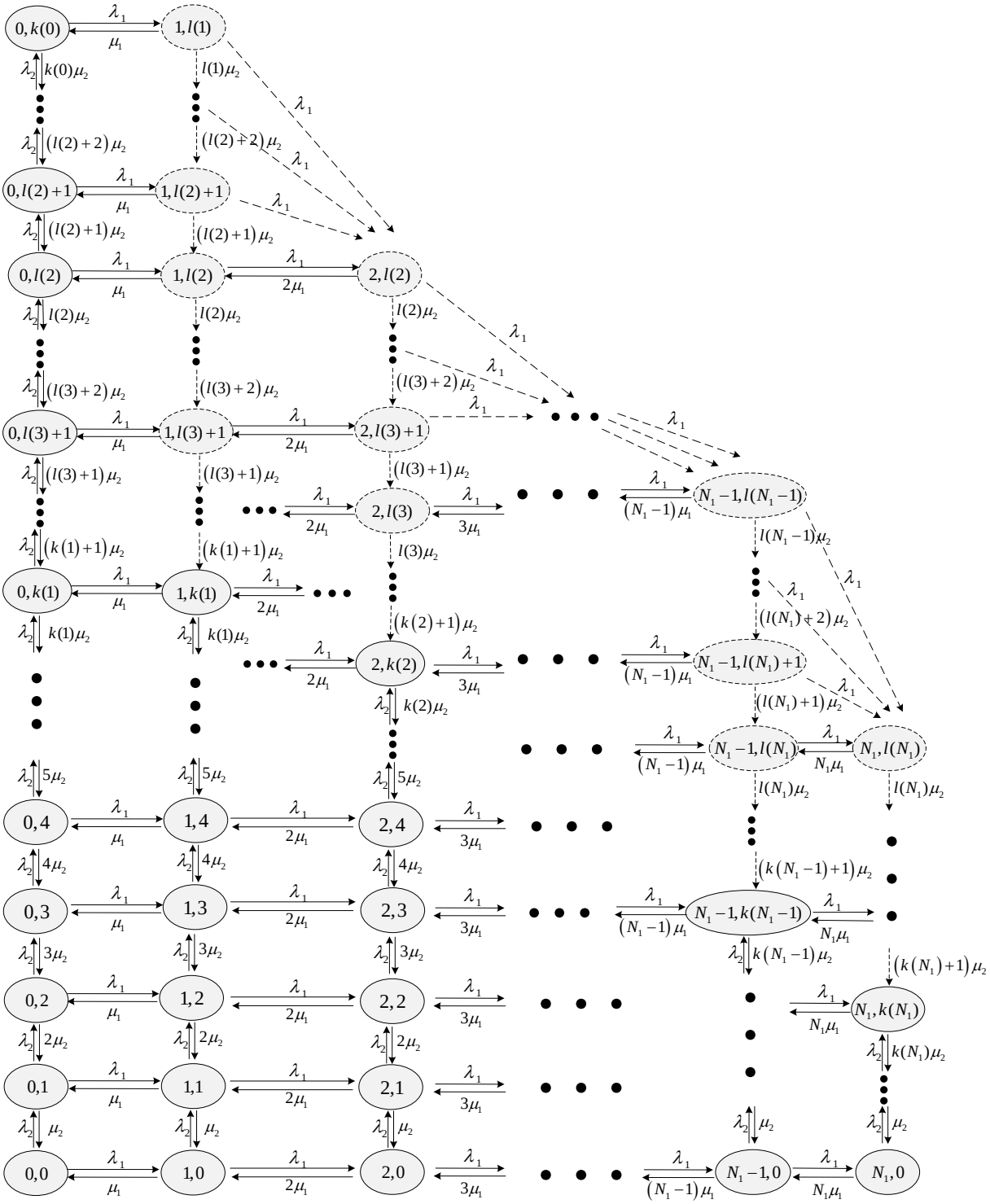

Рис. 1.4. Диаграмма интенсивностей переходов модели с приоритетным обслуживанием потокового трафика и гибкой стратегией обслуживания эластичного трафика.

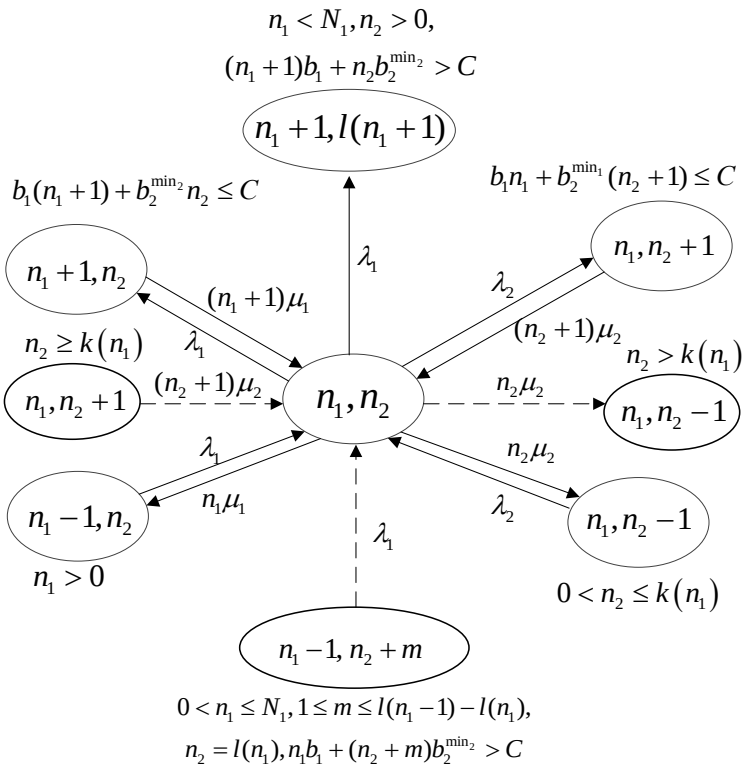

Рис. 1.5. Диаграмма интенсивностей переходов для центрального состояния модели с приоритетным обслуживанием потокового трафика и гибкой стратегией обслуживания эластичного трафика.

Основываясь на диаграмме интенсивностей переходов (рис. 1.5), рассматриваемый СП может быть описан системой уравнений глобального баланса (СУГБ) в общем виде:

$$\Big[\lambda_1 \cdot I\left\{n_1 < N_1, b_1(n_1+1) + b_2^{\min 2} n_2 \le C\right\} + \lambda_1 \cdot I\left\{n_1 < N_1, n_2 > 0, b_1(n_1+1) + b_2^{\min 2} n_2 > C\right\} +$$

$$+\lambda_2 \cdot I\left\{n_2 < N_2, b_1 n_1 + b_2^{\min 1}(n_2+1) \le C\right\} + n_1\mu_1 + n_2\mu_2\Big] \cdot p(n_1, n_2) = \lambda_1 \cdot p(n_1-1, n_2) \cdot$$

$$\cdot I\left\{n_1 > 0, b_1 n_1 + b_2^{\min 2} n_2 \le C\right\} + \lambda_2 \cdot I\left\{n_2 > 0, b_1 n_1 + b_2^{\min 1} n_2 \le C\right\} \cdot p(n_1, n_2-1) + (n_1+1)\mu_1 \cdot$$

$$\cdot I\left\{n_1 < N_1, b_1(n_1+1) + b_2^{\min 2} n_2 \le C\right\} \cdot p(n_1+1, n_2) + (n_2+1)\mu_2 \cdot p(n_1, n_2+1) \cdot$$

$$\cdot I\left\{n_2 < N_2, b_1 n_1 + b_2^{\min 2}(n_2+1) \le C\right\} + \lambda_1 \cdot p(n_1-1, n_2+m) \cdot I\left\{n_2 = l(n_1), 0 < n_1 \le N_1,\right.$$

$$\left. n_2 + m \le N_2, b_1(n_1-1) + b_2^{\min 2}(n_2+m) \le C, b_1 n_1 + b_2^{\min 2}(n_2+m) > C\right\}, m = \overline{1, l(n_1-1) - l(n_1)},$$

$$n_1 = \overline{0, N_1}, n_2 = \overline{0, N_2},$$

где $p(n_1, n_2)$, $(n_1, n_2) \in \mathbf{X}$ – стационарное распределение вероятностей состояний системы, $I\{x\}$ – функция-индикатор.

В связи с реализацией механизма прерывания обслуживания СП, описывающий поведение системы, не является обратимым, поэтому стационарное распределение $p(n_1, n_2)$, $(n_1, n_2) \in \mathbf{X}$, не представимо в мультипликативном виде.

Утверждение 1.1. Стационарное распределение $p(n_1, n_2)$, $(n_1, n_2) \in \mathbf{X}$, вероятностей состояний случайного процесса $\{(N_1(t), N_2(t)), t \geq 0\}$ определяется путем численного решения системы уравнений равновесия (СУР): $\mathbf{p}^T \cdot \mathbf{A} = \mathbf{0}^T$, $\mathbf{p}^T \cdot \mathbf{1} = 1$, где $\mathbf{A}$ – матрица интенсивностей переходов, элементы которой определены следующим образом:

$$a\left((n_1, n_2), (n_1', n_2')\right) =$$

$$= \begin{cases} \lambda_1, & \text{если } n_1' = n_1 + 1, n_2' = n_2, n_1 < N_1, b_1(n_1+1) + b_2^{\min 2} n_2 \leq C, \\ & \text{или } n_1' = n_1 + 1, n_2' = l(n_1+1), n_1 < N_1, n_2 > 0, b_1(n_1+1) + b_2^{\min 2} n_2 > C; \\ \lambda_2, & \text{если } n_1' = n_1, n_2' = n_2 + 1, n_2 < N_2, b_1 n_1 + b_2^{\min 1}(n_2+1) \leq C; \\ n_1\mu_1, & \text{если } n_1' = n_1 - 1, n_2' = n_2, n_1 > 0; \\ n_2\mu_2, & \text{если } n_1' = n_1, n_2' = n_2 - 1, n_2 > 0; \\ \varphi, & \text{если } n_1' = n_1, n_2' = n_2; \\ 0 & \text{в ином случае,} \end{cases} \tag{1.5}$$

$$\begin{aligned} \varphi = -\Big(& \lambda_1 \cdot I\left\{n_1 < N_1, b_1(n_1+1) + b_2^{\min 2} n_2 \leq C\right\} + \\ & + \lambda_1 \cdot I\left\{n_1 < N_1, n_2 > 0, b_1(n_1+1) + b_2^{\min 2} n_2 > C\right\} + \\ & + \lambda_2 \cdot I\left\{n_2 < N_2, b_1 n_1 + b_2^{\min 1}(n_2+1) \leq C\right\} + n_1\mu_1 + n_2\mu_2 \Big). \end{aligned} \tag{1.6}$$

Зная распределение вероятностей $p(n_1, n_2)$, $(n_1, n_2) \in \mathbf{X}$, можно вычислить следующие вероятностные характеристики системы:

– вероятность блокировки запросов первого типа

$$B_1 = \sum_{i=0}^{l(N_1)} p(N_1, i); \tag{1.7}$$

– вероятность блокировки запросов второго типа

$$B_2 = \sum_{i=0}^{N_1} \sum_{j=k(i)}^{l(i)} p(i,j); \tag{1.8}$$

– вероятность прерывания обслуживания запросов второго типа

$$p_{pre} = \sum_{i=0}^{N_1-1} \sum_{\substack{j=l(i+1)+1 \\ l(i)\neq l(i+1)}}^{l(i)} \frac{\lambda_1}{\lambda_1 + \lambda_2 \cdot I\{j < k(i)\} + i\mu_1 + j\mu_2} p(i,j); \tag{1.9}$$

– коэффициент использования ресурсов сети

$$U = \sum_{i=0}^{N_1} \sum_{j=1}^{l(i)} C \cdot p(i,j) + \sum_{i=1}^{N_1} i \cdot b_1 \cdot p(i,0); \tag{1.10}$$

– доля времени, в течение которого запросы второго типа обслуживаются со сниженным требованием к числу БЦК

$$\omega = \sum_{i=1}^{N_1} \sum_{\substack{j=k(i)+1 \\ k(i)<l(i)}}^{l(i)} j \cdot p(i,j) \Bigg/ \sum_{i=0}^{N_1} \sum_{j=0}^{l(i)} j \cdot p(i,j). \tag{1.11}$$

1.3. Численный анализ вероятностно-временных характеристик

Перейдем к численному анализу представленной модели, исходные данные для которого приведены в таблице 1.1. Отметим, что требования к ресурсам b_1 и $b_2^{\min_1}$ соответствуют скоростям передачи данных c_1 и c_2.

Покрытие базовой станции NR определяется с помощью методов, описанных в [10, 35], в то время как средние требования к ресурсам для запросов URLLC и eMBB рассчитываются, как описано в [36]. Для аппроксимации задержки передачи URLLC трафика соответствующее время обслуживания запроса принимается равным длительности кадра NR – 1 мс. Чтобы обеспечить надежную передачу данных, используется повторное кодирование в пределах одного кадра.

Таблица 1.1. Параметры системы для численного анализа

Обозначение		**Параметр системной модели**	**Значение**
Мат. модель	**Системная модель**		
C	C	Пропускная способность БС NR	20 МГц
λ_1	λ_1	Интенсивность поступления запросов на передачу URLLC трафика	$1000 \rightarrow 1100$ запросов/с
λ_2	λ_2	Интенсивность поступления запросов на передачу eMBB трафика	2 запроса/с
μ_1	μ_1	Среднее время обслуживания запроса на передачу URLLC трафика	0,1 с
μ_2	μ_2	Среднее время обслуживания запроса на передачу eMBB трафика	120 с
b_1	c_1	Скорость обслуживания запроса на передачу URLLC трафика	0,8 Мбит/с
$b_2^{\min_1}$	c_2	Минимальная скорость обслуживания запроса на передачу eMBB трафика	10 Мбит/с

Рассмотрим зависимость основных вероятностных характеристик – вероятности блокировки и вероятности прерывания – от интенсивности поступления запросов на передачу URLLC трафика (рис. 1.6 и 1.7 соответственно). Как можно заметить, вероятность блокировки запросов на передачу URLLC трафика не зависит от выбора стратегии со снижением скорости или без снижения. В то же время применение рассмотренных стратегий приводит к кардинально разным значениям вероятностей прерывания обслуживания запросов на передачу eMBB трафика. Вероятности блокировки и прерывания обслуживания для запросов на

передачу eMBB трафика при разных стратегиях совпадают только при увеличении интенсивности поступления запросов на передачу URLLC трафика. Стоит отметить, что гибкая стратегия приводит к более высокой вероятности блокировки запросов на передачу eMBB трафика, поскольку компенсируется более низкой вероятностью прерывания обслуживания запросов на передачу eMBB трафика.

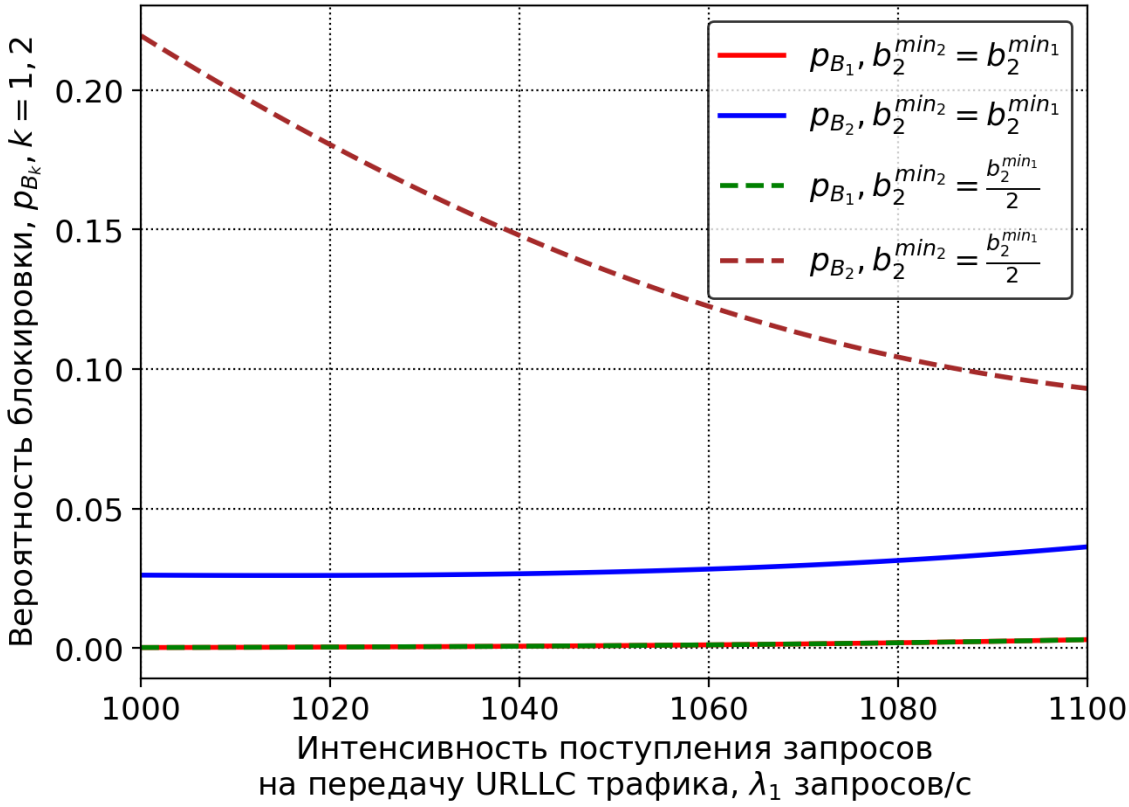

Рис. 1.6. Зависимость вероятности блокировки запросов на передачу URLLC/eMBB трафика от интенсивности поступления запросов на передачу URLLC трафика.

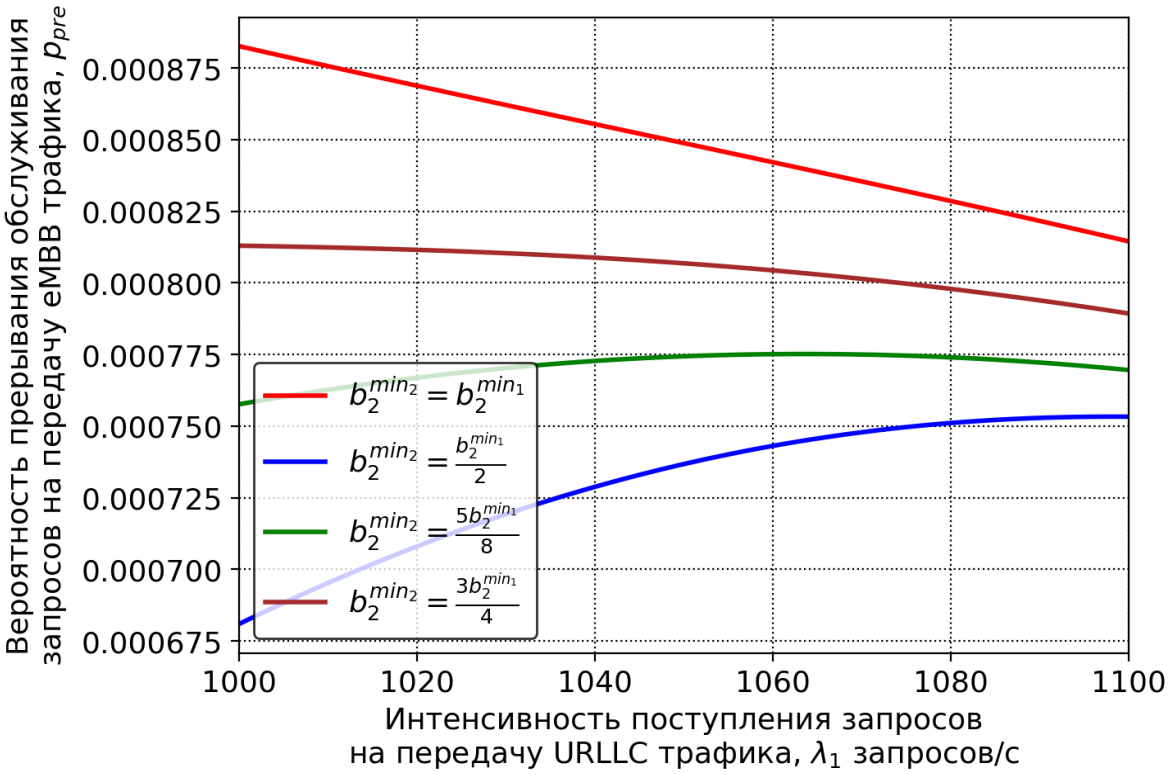

Рис. 1.7. Зависимость вероятности прерывания обслуживания запросов на передачу eMBB трафика от интенсивности поступления запросов на передачу URLLC трафика.

В дополнение к рис. 1.7 на рис. 1.8 продемонстрирована зависимость вероятности прерывания обслуживания запросов на передачу eMBB трафика от их интенсивности поступления для гибкой стратегии при различных значениях $b_2^{\min_2}$. Можно заметить, что, в отличии от влияния интенсивности поступления запросов на передачу URLLC трафика, выбор порогового значения для снижения скорости оказывает незначительный эффект на вероятность прерывания обслуживания.

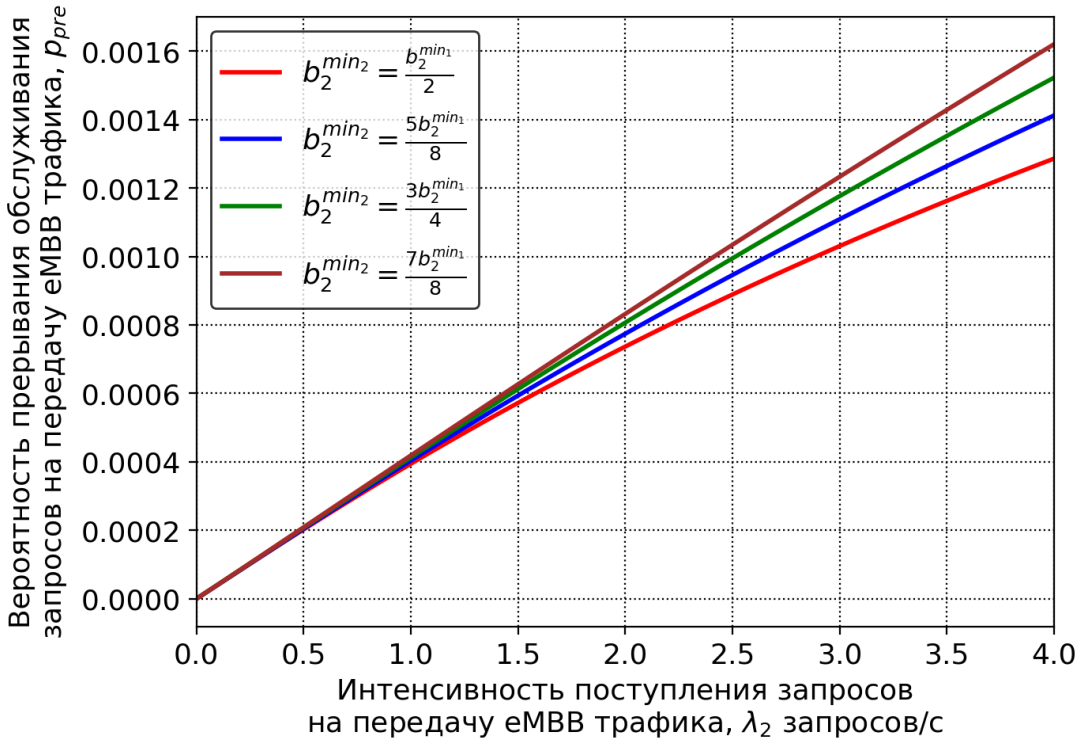

Рис. 1.8. Зависимость вероятности прерывания обслуживания запросов на передачу eMBB трафика от интенсивности поступления запросов на передачу eMBB трафика.

Оценка влияния рассмотренных стратегий на использование ресурсов (рис. 1.9) показывает, что гибкая стратегия, при которой запросы на передачу eMBB трафика некоторое время обслуживаются со сниженной скоростью, приводит к более высокому коэффициенту использования ресурсов на всем интервале значений интенсивностей поступления запросов на передачу URLLC трафика. В то же время, применение гибкой стратегии имеет также негативную сторону, которая продемонстрирована на рис. 1.10, иллюстрирующем долю времени, в течение которого скорость обслуживания запросов на передачу eMBB трафика снижена до порогового значения.

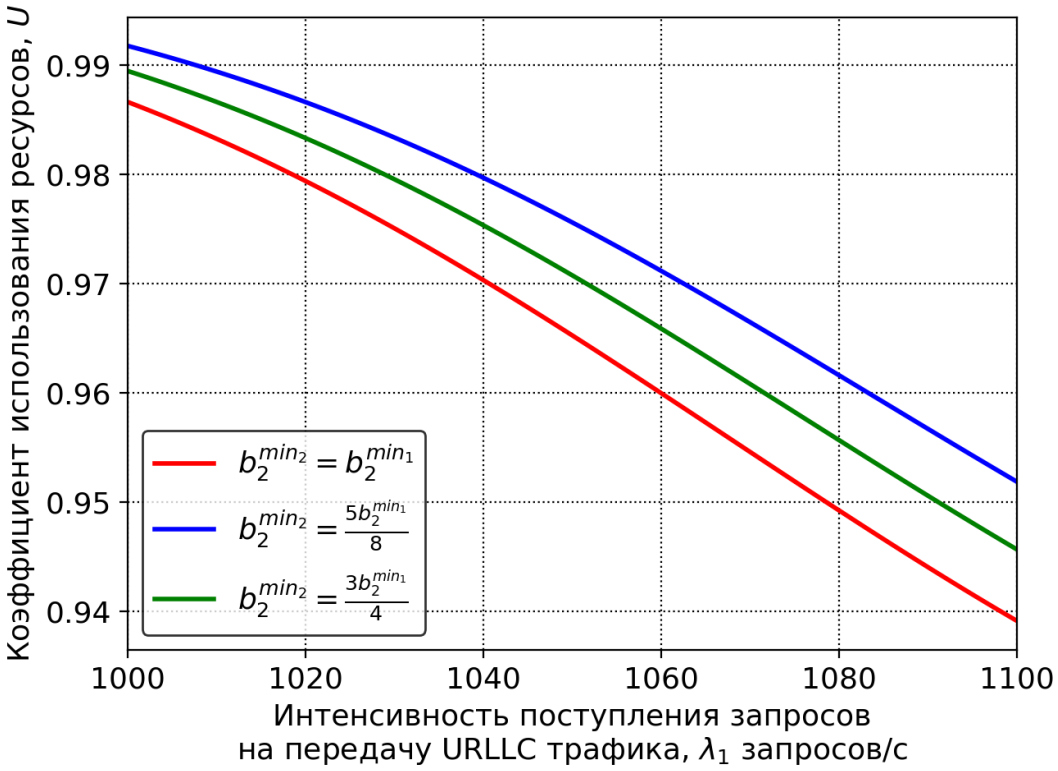

Рис. 1.9. Зависимость коэффициента использования ресурсов сети от интенсивности поступления запросов на передачу URLLC трафика.

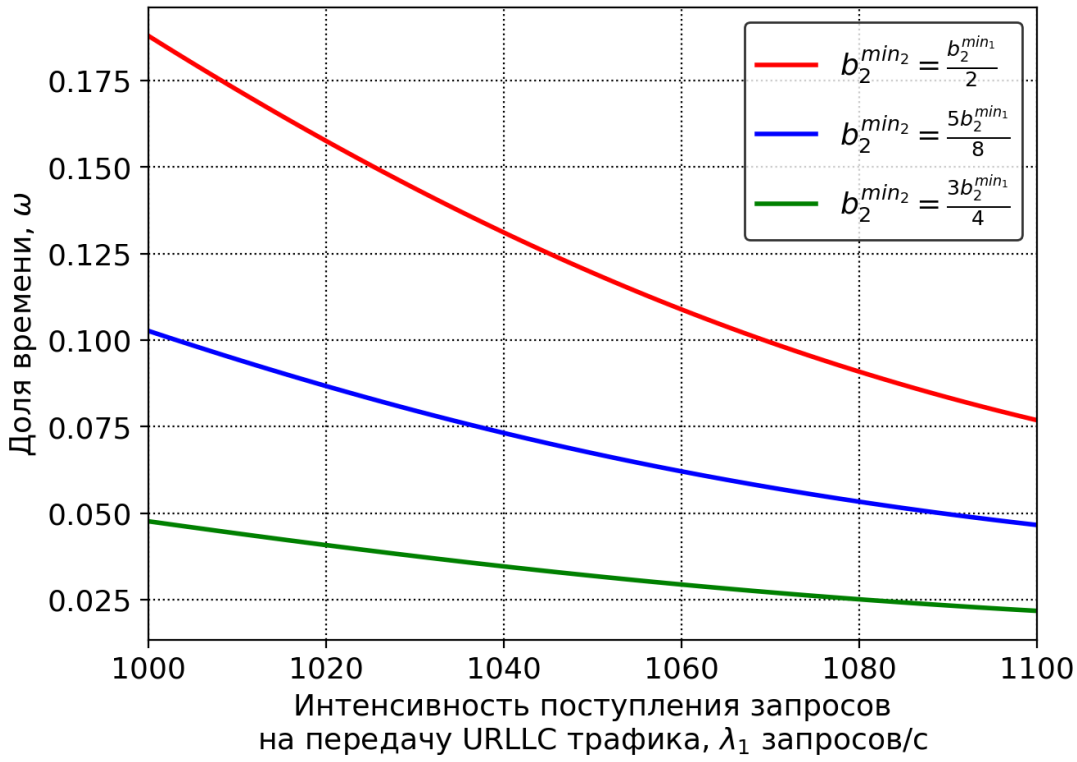

Рис. 1.10. Зависимость доли времени, в течение которого запросы на передачу eMBB трафика обслуживаются со сниженной скоростью, от интенсивности поступления запросов на передачу URLLC трафика.

Таким образом, использование пороговых значений для снижения скорости обслуживания запросов на передачу eMBB трафика предоставляет операторам сетей беспроводной связи простой и эффективный способ сбалансировать качество обслуживания и использование ресурсов, сохраняя при этом приоритет для обслуживания URLLC трафика. По сравнению со статическим механизмом резервирования ресурсов, обслуживание на основе приоритетов позволяет достичь использования 80-90% ресурсов сети, сохраняя при этом требования к задержке URLLC трафика. Кроме того, по сравнению с динамическим резервированием ресурсов, рассмотренный подход не требует процессов перераспределения ресурсов в режиме реального времени. Предлагаемая стратегия может быть использована в тех сферах, где предполагается обслуживание URLLC и eMBB трафика с динамически меняющейся нагрузкой на базовых станциях NR, например, на автоматизированных предприятиях.

1.4. Постановка задачи исследования

Проведенные в разделах 1.1–1.3 диссертационной работы исследования показали, что в области анализа обслуживания принципиально разных типов трафика в беспроводных сетях пятого поколения – URLLC и eMBB – ранее в основном рассматривались механизмы их отдельной поддержки. В то же время, существующие исследования моделей совместного обслуживания различных типов трафика не содержат подробного сравнения производительности систем при различных вариантах стратегий передачи данных, например, основанных на резервировании и приоритетах, с точки зрения пользователя и оператора. Рассматриваемые ранее традиционные методы совместного обслуживания разных типов трафика, основанные на полном резервировании или не предусматривающие разделения ресурсов [37], приводят либо к неэффективному использованию радиоресурсов, либо к невозможности обеспечить требуемый уровень качества обслуживания. В связи с этим, в диссертационной работе проводится построение моделей

совместного обслуживания потокового – URLLC и эластичного – eMBB трафика, основанное на введении приоритетов, а также сравнительный анализ различных стратегий обслуживания и разделения ресурсов.

Кроме того, существует необходимость проведения анализа влияния условий развертывания сети на процесс обслуживания трафика, например, на интенсивность прерывания обслуживания установленных соединений в результате динамической блокировки путей распространения сигнала.

Таким образом, цель диссертационной работы состоит в разработке набора марковских моделей совместного обслуживания трафика с приоритизацией и разделением ресурсов, описанных в виде СМО, применимых для анализа систем связи, требующих одновременной передачи сверхнадежного трафика с низкими задержками и мобильного трафика широкополосного доступа в условиях промышленного развертывания беспроводных сетей, а также анализе показателей эффективности разработанных моделей. Для достижения поставленной цели необходимо решить следующие задачи:

1. Разработка марковских моделей схем доступа к радиоресурсам сетей для одновременного предоставления услуг, предъявляющих различные требования к QoS, на основе приоритетного обслуживания в условиях промышленной среды.

2. Разработка алгоритмов расчета показателей эффективности моделей, а также проведение их сравнительного анализа при различных стратегиях обслуживания или разделения ресурсов.

С учетом особенностей, обозначенных в табл. 1.2, для каждого случая в работе представлено описание системной модели, проведено построение математической модели, рассчитано стационарное распределение вероятностей состояний системы, выведены формулы для расчета вероятностных характеристик модели, таких как вероятность блокировки, вероятность прерывания обслуживания, а также коэффициент использования ресурсов системы, на основе которых проведен численный эксперимент.

Сравнительный анализ результатов численного эксперимента позволяет оценить предложенные стратегии обслуживания и разделения ресурсов для различных системных параметров.

Таблица 1.2. Комплекс моделей, разработанных в рамках диссертационного исследования.

№	Особенности модели	Цель исследования
1	Модель мобильной сети с двумя типами услуг, генерирующими потоковый и эластичный трафик, **и снижением скорости обслуживания.**	Сравнение двух стратегий обслуживания эластичного трафика – без снижения скорости передачи данных и со снижением.
2	Модель мобильной сети с двумя типами услуг, генерирующими потоковый и эластичный трафик, **и реализацией абсолютного приоритета.**	Сравнение трех сценариев передачи двух типов трафика: I) передача трафика через БС; II) D2D-передача с полной координацией через БС; III) D2D-передача без координации через БС, с учетом динамической блокировки в промышленных развертываниях беспроводных сетей.
3	Модель мобильной сети с произвольным числом услуг, каждая из которых может генерировать как потоковый, так и эластичный трафик, **резервированием**	Получение стационарного распределения вероятностей состояний системы двумя способами: I) численное решение системы уравнений равновесия и II) аналитическое решение. Сравнение пяти стратегий разделения ресурсов: I) отсутствие резервирования и приоритетов; II) приоритетное обслуживание

№	Особенности модели	Цель исследования
	индивидуальных зон и приоритетным обслуживанием.	с прерыванием, но без резервирования; III) полное резервирование ресурсов; IV) частичное резервирование без прерывания; V) частичное резервирование с прерыванием. Численная оптимизация параметров для обеспечения гарантий производительности трафика.

Глава 2. Модель совместного обслуживания трафика с приоритизацией

Перейдем к анализу характеристик модели с двумя типами услуг, генерирующими потоковый и эластичный трафик, и реализацией абсолютного приоритета (№2 в табл. 1.2). В данной модели, в отличие от модели, исследованной в главе 1, не используется механизм снижения скорости передачи, однако учитывается влияние условий развертывания сети в промышленной среде. Основной целью данной главы является сравнительный анализ показателей эффективности модели для трех стратегий передачи данных: I) передача трафика через БС; II) D2D-передача с полной координацией через БС; III) D2D-передача без координации через БС.

2.1. Системная модель схемы одновременного предоставления услуг с реализацией абсолютного приоритета

Рассмотрим сценарий внедрения технологии 5G NR в промышленности, например на заводе с автоматизированными производственными линиями (рис. 2.1). Для автоматизации в промышленной среде необходимо обеспечение регулярного мониторинга процессов с помощью датчиков и камер видеонаблюдения. Предположим, базовые станции (англ. base station, BS) NR установлены на высоте h_{BS} метров с плотностью χ БС/м^2, образуя пуассоновский точечный процесс. Пользовательские устройства (англ. user equipment, UE), включающие в себя датчики и камеры, расположены на высоте h_{UE} метров в узлах сетки с шагом l метров. Ширина полосы пропускания каждой базовой станции составляет W Гц, что эквивалентно емкости соты сети связи C каналов.

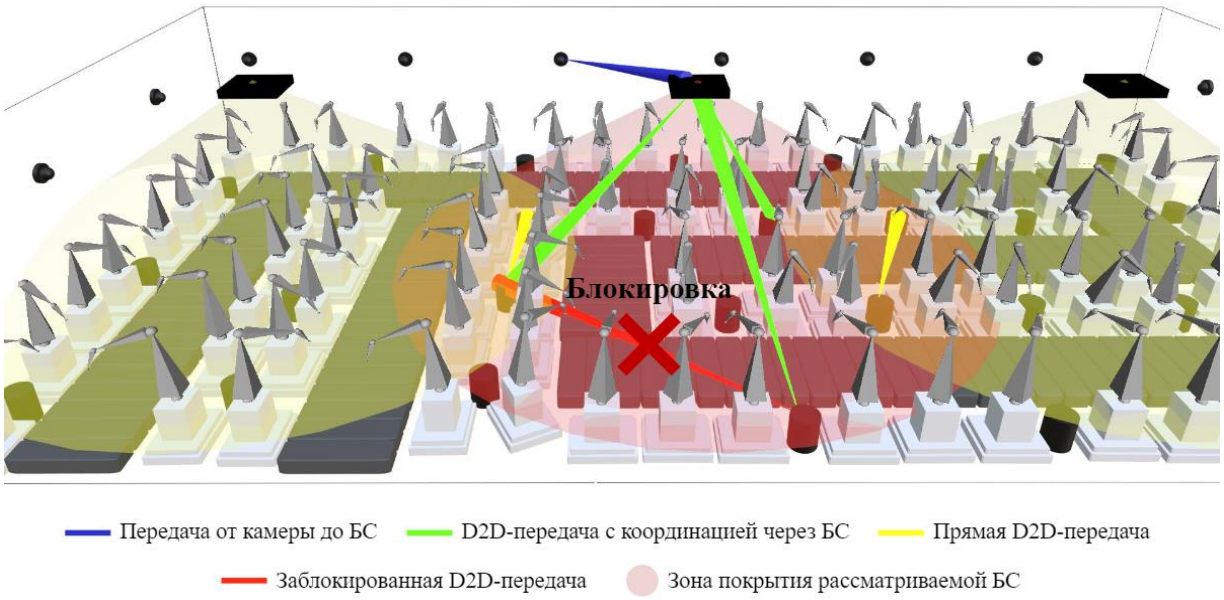

Рис. 2.1. Сценарий развертывания 5G NR.

Подключенные к производственному оборудованию датчики генерируют потоковый трафик, соответствующий URLLC услугам и имеющий гарантированную скорость передачи данных c_1, $c_1 \geq 1$. Эластичный трафик, соответствующий eMBB услугам, генерируется камерами для удаленного мониторинга и имеет минимальную скорость передачи данных $c_2^{\min}$, $c_2^{\min} \geq 1$. Средняя интенсивность поступления запросов на передачу данных от датчиков и камер предполагается равной λ_k, средняя длительность сессий по передаче данных – μ_k^{-1}, $k = 1,2$.

Рассмотрим три стратегии одновременного предоставления услуг двух типов с использованием абсолютного приоритета [38]:

1. базовая стратегия (англ. baseline) – передача трафика через БС NR;
2. стратегия согласованной прямой передачи (англ. D2D-aware) – использование прямой передачи между устройствами с полной координацией через базовую станцию, позволяющее избежать дополнительной интерференции;
3. стратегия несогласованной прямой передачи (англ. D2D-unaware) – использование D2D-передачи между устройствами без координации

через базовую станцию, позволяющее снизить задержку, но вызывающее дополнительную интерференцию.

Под абсолютным приоритетом будем понимать дисциплину обслуживания, в рамках которой запросы на передачу URLLC трафика имеют приоритет над запросами на передачу eMBB трафика и могут прерывать их обслуживание в условиях недостаточности ресурсов.

Требования к числу БЦК для передачи URLLC трафика через базовую станцию $b_{1,B}$ и при D2D-передаче $b_{1,D}$, а также требования к числу БЦК для передачи eMBB трафика $b_2^{\min}$ зависят от плотности размещения базовых станций и могут быть рассчитаны по формулам

$$b_2^{\min} = \frac{c_2^{\min}}{E[S_{e,B}]}, \tag{2.1}$$

$$b_{1,B} = \frac{c_1}{E[S_{e,B}]}, \tag{2.2}$$

$$b_{1,D} = \frac{c_1}{E[S_{e,D}]}, \tag{2.3}$$

где $E[S_{e,B}]$ – средняя спектральная эффективность при передаче трафика через базовую станцию, а $E[S_{e,D}]$ – при D2D-передаче.

Перейдем к расчету спектральной эффективности для двух видов передачи данных. Плотности вероятностей расстояния между двумя случайно выбранными пользовательскими устройствами, D, и расстояния от случайно выбранного устройства до базовой станции, B, определяются как [39]

$$f_B(x) = \frac{2x}{r}, \tag{2.4}$$

$$f_D(x) = \frac{2x}{r^2}\left[\frac{2}{\pi}\cos^{-1}\left(\frac{x}{2r}\right) - \frac{x}{r\pi}\sqrt{1-\frac{x^2}{4r^2}}\right]. \tag{2.5}$$

Чтобы получить радиус зоны покрытия базовой станции r, воспользуемся моделью распространения сигнала, рассмотренной в [40], согласно которой

$$r = \min\left(r_S, r_V\right), \tag{2.6}$$

где r_S – максимально возможное расстояние между пользовательским устройством и базовой станцией, а r_V – половина расстояния между двумя БС.

Отношение сигнала к шуму (англ. signal-to-noise ratio, SNR) на устройстве, находящемся на расстоянии x от базовой станции:

$$S(x) = \frac{P_U G_A G_U}{N_0 W + M_I} x^{-\zeta}, \tag{2.7}$$

где G_A и G_U – коэффициенты усиления антенны на базовой станции и на пользовательском устройстве соответственно, P_U – мощность пользовательского устройства, N_0 – спектральная плотность мощности шума, ζ – коэффициент распространения сигнала, M_I – мощность помех. Тогда

$$E[S_{e,B}] = \int_0^r f_B(x)\log_2[1+S(x)]dx, \tag{2.8}$$

$$E[S_{e,D}] = \int_0^{2r} f_D(x)\log_2[1+S(x)]dx. \tag{2.9}$$

Чтобы рассчитать вероятность блокировки D2D-передачи, в первую очередь, нужно найти вероятность перекрытия препятствием пути прямой видимости длиной x между двумя устройствами, $p_{B,1}(x)$. Используя методы интегральной геометрии [41], получим вероятность перекрытия

$$p_{B,1}(x) = \frac{2w\left(\pi w + 4x\right)\left(1-\kappa\right)}{\pi\left(2\pi r^2 - 4r^2 \sin^{-1}\left(x/(2r)\right) - x\sqrt{4r^2 - x^2}\right)}, \tag{2.10}$$

где w – ширина пользовательского устройства, κ – прозрачность пользовательского устройства.

Вероятность блокировки пути прямой видимости может быть рассчитана по формуле

$$p_B(x) = \sum_{j=1}^{N_R} \binom{N_R}{j} \nu^j (1-\nu)^{N_R - j} \left[1 - \left(1 - p_{B,1}(x)\right)\right]^j, \quad (2.11)$$

где N_R – максимальное количество устройств, расположенных в зоне покрытия базовой станции NR, которое может быть получено путем аппроксимации задачи о круге Гаусса, ν – вероятность нахождения устройства в точке сетки.

Тогда искомая вероятность блокировки будет рассчитываться как

$$p_B = \int_0^{2r} f_D(x)\, p_B(x) dx. \quad (2.12)$$

Полученные формулы будут использоваться для расчета требований к числу БЦК для передачи URLLC и eMBB трафика, а также для расчета вероятностных характеристик модели с учетом динамической блокировки. Далее перейдем к построению математической модели схемы совместного обслуживания запросов первого типа – запросов на передачу потокового трафика, генерируемого URLLC услугой, и второго типа – запросов на передачу эластичного трафика, генерируемого eMBB услугой, в виде СМО с приоритетным обслуживанием.

2.2. Построение математической модели

Функционирование рассматриваемой системы может быть описано двумерным марковским СП $\left\{\left(N_1(t), N_2(t)\right), t \geq 0\right\}$, где $N_1(t)$ и $N_2(t)$ – случайное количество обслуживаемых системой запросов первого и второго типа соответственно в момент времени t. Обозначим максимальное число запросов первого и второго типа, которое может находиться в системе, $N_1 = \lfloor C/b_1 \rfloor$ и $N_2 = \left\lfloor C/b_2^{\min} \right\rfloor$ соответственно, тогда $n_k = 0,...,N_k$ – число обслуживаемых системой запросов k-го типа, $k = 1,2$.

Состояние системы может быть описано двумерным вектором $\mathbf{n}=(n_1,n_2)$ над пространством состояний

$$\mathbf{X}=\left\{(n_1,n_2): n_1\geq 0, n_2\geq 0, n_1b_1+n_2b_2^{\min}\leq C\right\}. \quad (2.13)$$

Обозначим $k(n_1)=\left\lfloor (C-n_1b_1)/b_2^{\min}\right\rfloor$ максимальное число запросов второго типа, которые могут быть приняты в систему при обслуживании n_1 запросов первого типа. При этом число БЦК, выделяемое для обслуживания запроса на передачу eMBB трафика, может изменяться в зависимости от количества запросов в системе:

$$b_2(n_1,n_2)=\left\lfloor (C-n_1b_1)/n_2\right\rfloor\geq b_2^{\min}. \quad (2.14)$$

Сформулируем правила приема и обслуживания запросов:

- если число запросов k-го типа, которые обслуживаются в системе, меньше максимально возможного числа таких запросов N_k, а число свободных каналов, доступных для запросов этого типа, составляет не менее b_1 и $b_2^{\min}$ для первого и второго типа соответственно, то поступающий запрос принимается на обслуживание;
- если число запросов первого типа, которые обслуживаются в системе, меньше максимально возможного числа запросов N_1, число свободных каналов, доступных для запросов, меньше b_1, а число запросов второго типа, которые обслуживаются в системе, не меньше 1, то поступающий запрос первого типа принимается на обслуживание за счет прерывания обслуживания $q(n_1,n_2)=\left\lceil \left(b_1-C+\left(n_1b_1+n_2b_2^{\min}\right)\right)/b_2^{\min}\right\rceil$ запросов второго типа;
- в противном случае поступающий в систему запрос блокируется.

На рис. 2.2 и рис. 2.3 показана составленная на основе сформулированных выше правил диаграмма интенсивностей переходов в общем виде и для центрального состояния соответственно.

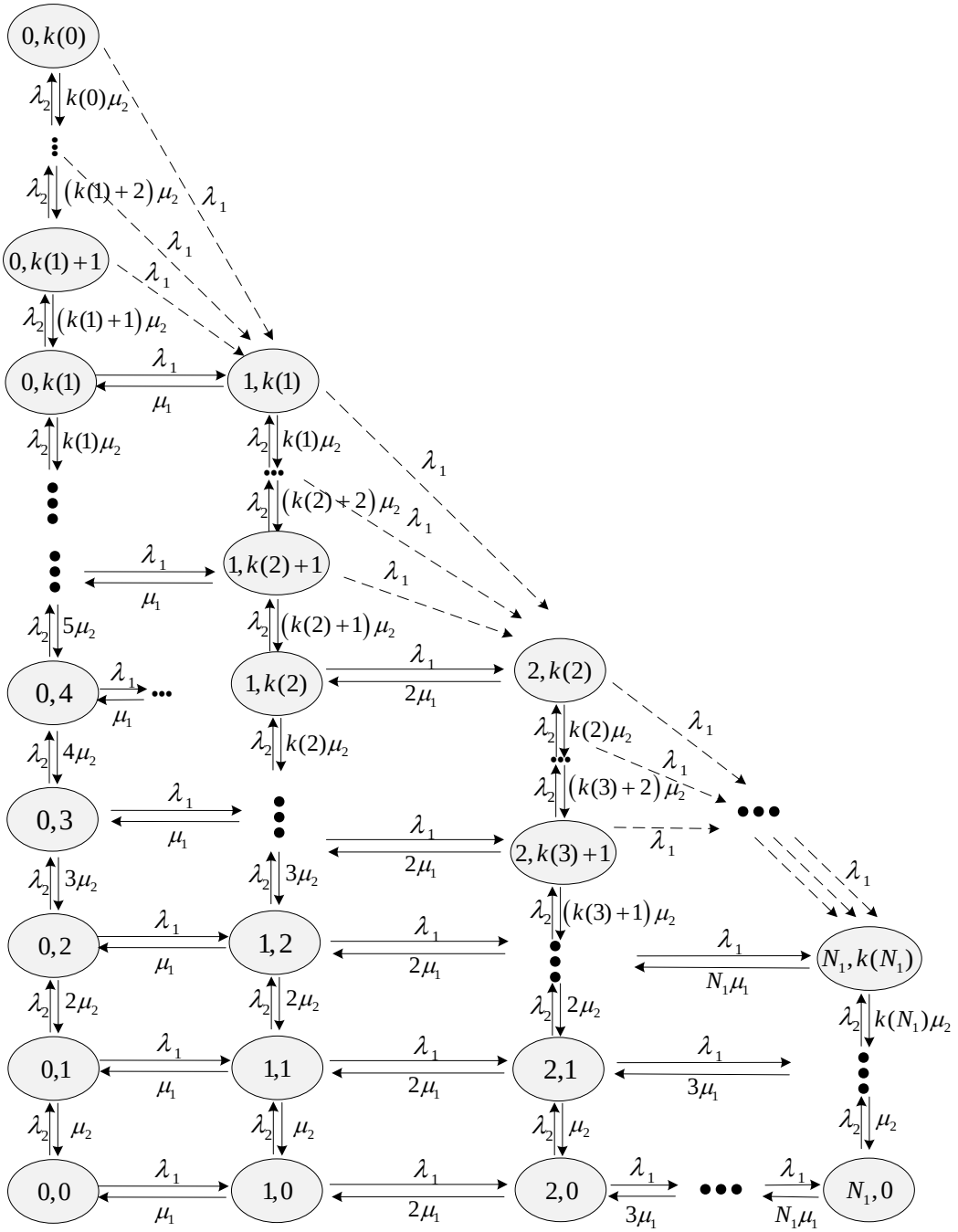

Рис. 2.2. Диаграмма интенсивностей переходов модели совместного обслуживания трафика с реализацией абсолютного приоритета.

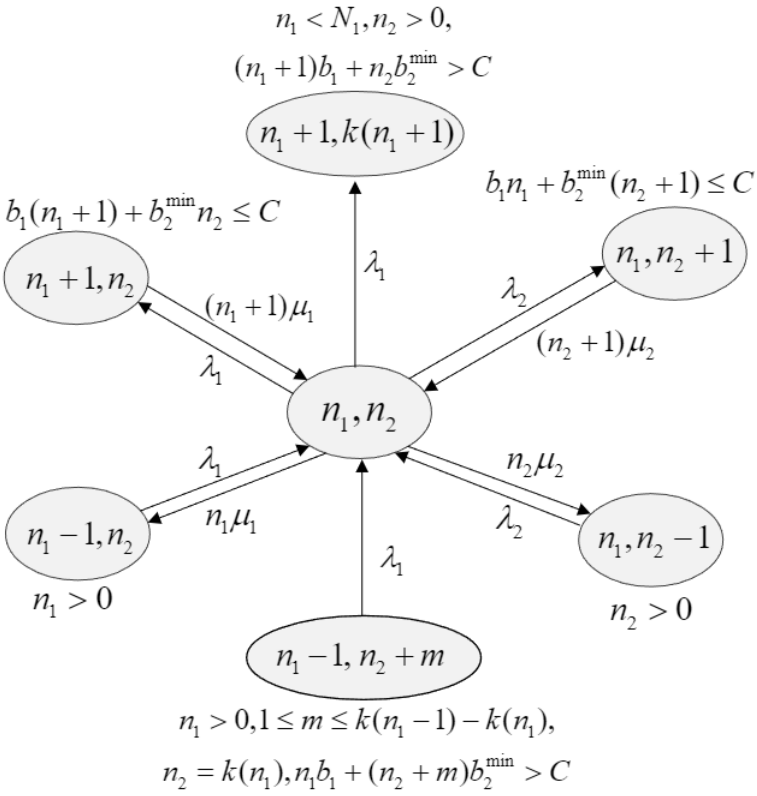

Рис. 2.3. Диаграмма интенсивностей переходов для центрального состояния модели совместного обслуживания трафика с реализацией абсолютного приоритета.

Основываясь на диаграмме интенсивностей переходов (рис. 2.3), рассматриваемый СП может быть описан СУГБ в общем виде:

$$\Big[\lambda_1 \cdot I\left\{n_1 < N_1,\, b_1(n_1+1) + b_2^{\min} n_2 \le C\right\} + \lambda_1 \cdot I\left\{n_1 < N_1,\, n_2 > 0,\, b_1(n_1+1) + b_2^{\min} n_2 > C\right\} +$$

$$+\lambda_2 \cdot I\left\{n_2 < N_2,\, b_1 n_1 + b_2^{\min}(n_2+1) \le C\right\} + n_1\mu_1 + n_2\mu_2\Big] \cdot p(n_1, n_2) = \lambda_1 \cdot p(n_1-1, n_2) \cdot$$

$$\cdot I\left\{n_1 > 0,\, b_1 n_1 + b_2^{\min} n_2 \le C\right\} + \lambda_2 \cdot I\left\{n_2 > 0,\, b_1 n_1 + b_2^{\min} n_2 \le C\right\} \cdot p(n_1, n_2-1) + (n_1+1)\mu_1 \cdot$$

$$\cdot I\left\{n_1 < N_1,\, b_1(n_1+1) + b_2^{\min} n_2 \le C\right\} \cdot p(n_1+1, n_2) + (n_2+1)\mu_2 \cdot p(n_1, n_2+1) \cdot$$

$$\cdot I\left\{n_2 < N_2,\, b_1 n_1 + b_2^{\min}(n_2+1) \le C\right\} + \lambda_1 \cdot p(n_1-1, n_2+1) \cdot$$

$$\cdot I\left\{n_1 > 0,\, n_2+1 \le N_2,\, b_1(n_1-1) + b_2^{\min}(n_2+1) \le C,\, b_1 n_1 + b_2^{\min}(n_2+1) > C\right\} + \ldots +$$

$$+\lambda_1 \cdot I\left\{n_1 > 0,\, b_1(n_1-1) + b_2^{\min} k(n_1-1) \le C,\, b_1 n_1 + b_2^{\min} k(n_1-1) > C\right\} \cdot p(n_1-1, k(n_1-1)),$$

$$n_1 = 0, \ldots, N_1,\; n_2 = 0, \ldots, N_2,$$

где $p(n_1, n_2)$, $(n_1, n_2) \in \mathbf{X}$ – стационарное распределение вероятностей состояний системы.

В связи с реализацией механизма прерывания обслуживания СП, описывающий поведение системы, не является обратимым, поэтому стационарное распределение $p(\mathbf{n})$, $\mathbf{n} \in \mathbf{X}$, не представимо в мультипликативном виде.

Утверждение 2.1. Стационарное распределение $p(\mathbf{n})$, $\mathbf{n} \in \mathbf{X}$, вероятностей состояний случайного процесса $\{(N_1(t), N_2(t)), t \geq 0\}$ определяется путем решения системы уравнений равновесия: $\mathbf{p}^T \cdot \mathbf{A} = \mathbf{0}^T$, $\mathbf{p}^T \cdot \mathbf{1} = 1$, где $\mathbf{A}$ – инфинитезимальная матрица, элементы которой определены в (2.15).

$$a(\mathbf{n}, \mathbf{n}') = \begin{cases} \lambda_1, & \text{если } \mathbf{n}' = \mathbf{n} + \mathbf{e}_1, n_1 < N_1, b_1(n_1 + 1) + b_2^{\min} n_2 \leq C, \\ & \text{или } n_1 < N_1, n_2 > 0, b_1(n_1 + 1) + b_2^{\min} n_2 > C, \\ & n_1' = n_1 + 1, n_2' = n_2 - q(n_1, n_2); \\ \lambda_2, & \text{если } \mathbf{n}' = \mathbf{n} + \mathbf{e}_2, n_2 < N_2, b_1 n_1 + b_2^{\min}(n_2 + 1) \leq C; \\ n_1 \mu_1, & \text{если } \mathbf{n}' = \mathbf{n} - \mathbf{e}_1, n_1 > 0; \\ n_2 \mu_2, & \text{если } \mathbf{n}' = \mathbf{n} - \mathbf{e}_2, n_2 > 0; \\ \varphi, & \text{если } \mathbf{n}' = \mathbf{n}; \\ 0 & \text{в ином случае,} \end{cases} \tag{2.15}$$

$$\begin{aligned} \varphi = -\Big[& \lambda_1 \cdot I\{n_1 < N_1, b_1(n_1 + 1) + b_2^{\min} n_2 \leq C\} + \\ & + \lambda_1 \cdot I\{n_1 < N_1, n_2 > 0, b_1(n_1 + 1) + b_2^{\min} n_2 > C\} + \\ & + \lambda_2 \cdot I\{n_2 < N_2, b_1 n_1 + b_2^{\min}(n_2 + 1) \leq C\} + n_1 \mu_1 + n_2 \mu_2 \Big]. \end{aligned} \tag{2.16}$$

Получив распределение вероятностей $p(\mathbf{n})$, $\mathbf{n} \in \mathbf{X}$, можно вычислить основные вероятностные характеристики системы: вероятность блокировки запросов и вероятность прерывания обслуживания запросов на передачу эластичного трафика.

Утверждение 2.2. Вероятность блокировки запросов на передачу URLLC трафика рассчитывается по формуле

$$B_1 = \sum_{n_1=0}^{k(N_1)} p\left(N_1, n_1\right). \tag{2.17}$$

Утверждение 2.3. Вероятность блокировки запросов на передачу eMBB трафика рассчитывается по формуле

$$B_2 = \sum_{n_1=0}^{N_1} p\left(n_1, k(n_1)\right). \tag{2.18}$$

Утверждение 2.4. Вероятность прерывания обслуживания запроса на передачу eMBB трафика рассчитывается по формуле

$$\Pi = \sum_{n_1=0}^{N_1-1} \sum_{\substack{n_2=k(n_1+1)+1 \\ k(n_1)\neq k(n_1+1)}}^{k(n_1)} \frac{\lambda_1 p\left(n_1, n_2\right)}{\lambda_1 + \lambda_2 \cdot I\left\{n_2 < k(n_1)\right\} + n_1\mu_1 + n_2\mu_2}. \tag{2.19}$$

2.3. Численный анализ показателей эффективности модели при разных стратегиях передачи трафика

В качестве численного эксперимента рассмотрим влияние плотности размещения пользовательских устройств (вероятности наличия устройства в точке сетки) ν и минимальной скорости передачи данных $c_2^{\min}$ на вероятность блокировки запроса на передачу данных при параметрах, представленных в таблице 2.1.

Таблица 2.1. Параметры системы для численных расчетов.

Параметр	**Значение**
Интенсивность поступления запросов на передачу URLLC трафика	5 000 запросов/с
Интенсивность поступления запросов на передачу eMBB трафика	100 запросов/с
Среднее время обслуживания запроса на передачу URLLC трафика	1 мс
Среднее время обслуживания запроса на передачу eMBB трафика	120 с

Параметр	**Значение**
Скорость передачи URLLC трафика	2 Мбит/с
Скорость передачи eMBB трафика	1 Мбит/с
Коэффициент усиления антенны на базовой станции	16×4
Коэффициент усиления антенны на пользовательском устройстве	4×4
Коэффициент распространения сигнала	5×10^{-4}

Для случая, изображенного на рис. 2.4, можно заметить рост вероятности блокировки в условиях увеличения плотности размещения пользовательских устройств ν. Как мы видим, стратегия, при которой все запросы проходят через базовую станцию, имеет постоянную вероятность блокировки запросов двух типов. В то же время использование стратегии D2D-aware значительно снижает вероятность блокировки запросов на передачу URLLC трафика. Стоит отметить, что преимущества прямой передачи данных в стратегии D2D-unaware незначительны по причине негативного влияния неконтролируемой интерференции на передачу данных, которое в условиях высокой нагрузки может оказаться критичным.

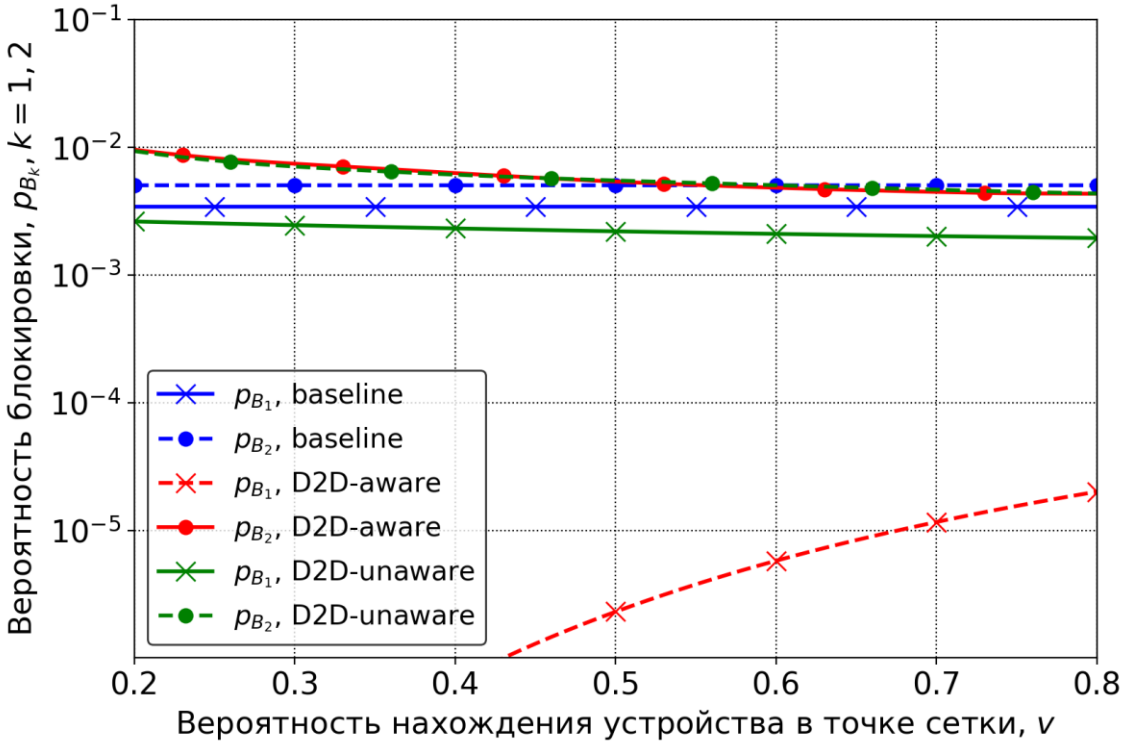

Рис. 2.4. Вероятность блокировки запросов на передачу URLLC/eMBB трафика в зависимости от плотности размещения пользовательских устройств.

Можно отметить, что передача eMBB трафика создает значительную нагрузку на систему, оказывая таким образом влияние на обслуживание URLLC трафика. На рис. 2.5 продемонстрирована зависимость вероятности блокировки запросов на передачу URLLC/eMBB трафика от требований к скорости обслуживания запросов на передачу eMBB трафика при $\nu = 0{,}5$.

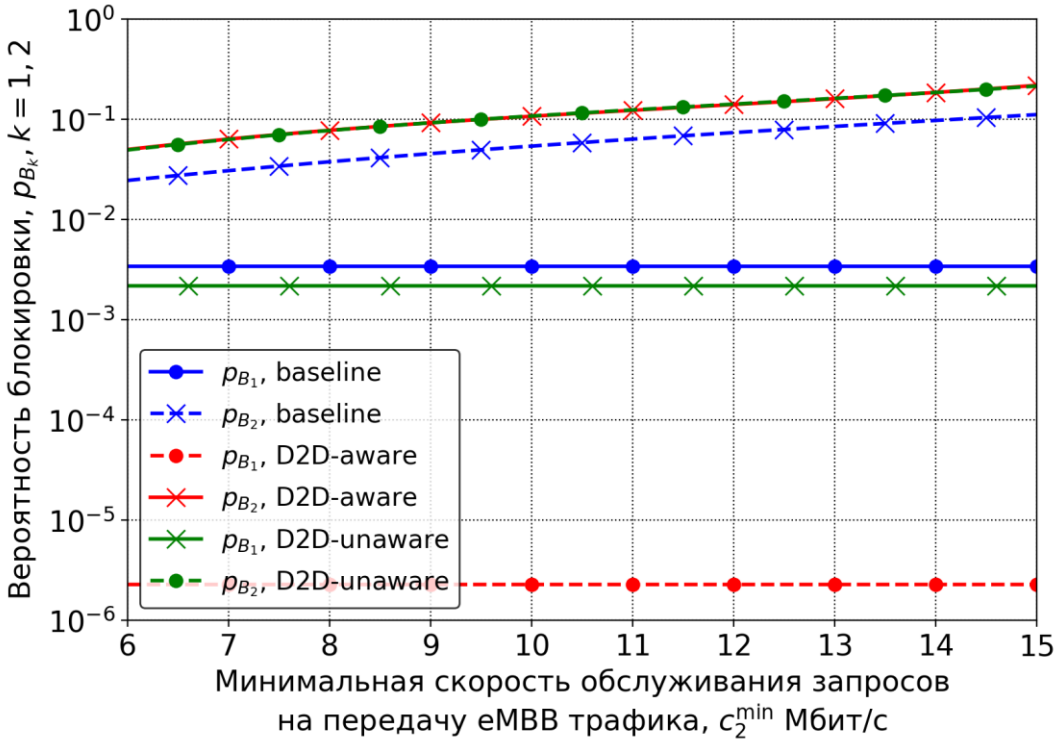

Рис. 2.5. Вероятность блокировки запросов на передачу URLLC/eMBB трафика в зависимости от минимальной скорости передачи eMBB трафика.

Как можно заметить, обслуживание на основе приоритетов эффективно при увеличении требований к запросам на передачу eMBB трафика, поскольку при этом вероятность блокировки запроса на передачу URLLC трафика остается неизменной. Использование приоритетов гарантирует защиту URLLC трафика от потенциально изменяющейся нагрузки, создаваемой eMBB трафиком. Таким образом, оба графика демонстрируют преимущество стратегии D2D-aware перед другими стратегиями на всем промежутке значений скорости обслуживания запросов $c_2^{\min}$.

Использование направленных антенн может существенно сказаться на эффективности рассматриваемых стратегий. На рис. 2.6 и 2.7 показано влияние количества антенных элементов, используемых в антенных решетках базовой станции и пользовательского оборудования, на вероятности блокировки запросов на передачу URLLC и eMBB трафика соответственно при $c_1 = 2$ Мбит/с, $c_2^{\min} = 1$ Мбит/с, $\chi = 5 \times 10^{-4}$, $\mu_1 = 10^3$, $\mu_2 = 1/120$. Стоит отметить, что количество антенных элементов не влияет на

выбор наилучшей стратегии, однако эффективность применения стратегии D2D-aware резко возрастает при увеличении количества антенных элементов базовой станции.

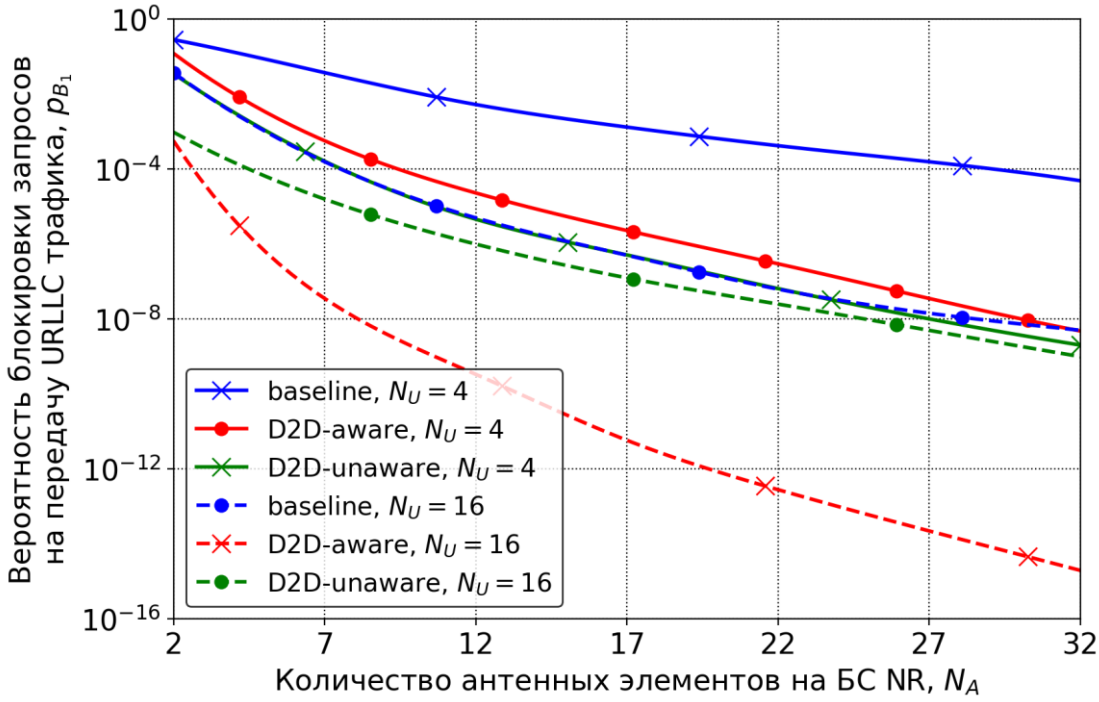

Рис. 2.6. Вероятность блокировки запросов на передачу URLLC трафика в зависимости от количества антенных элементов.

Для всех рассмотренных стратегий количество антенных элементов как пользовательского оборудования, так и базовой станции влияет на эффективность, а, следовательно, и на количество ресурсов, необходимых для обслуживания трафика. В результате вероятность блокировки запросов на передачу URLLC и eMBB трафика уменьшается по мере увеличения N_A или N_U. Кроме того, увеличение количества антенных элементов также снижает вероятность возникновения помех для стратегии, не использующей D2D-передачу, что положительно сказывается на вероятности блокировки запросов на передачу URLLC трафика, позволяя уменьшить ее в несколько раз.

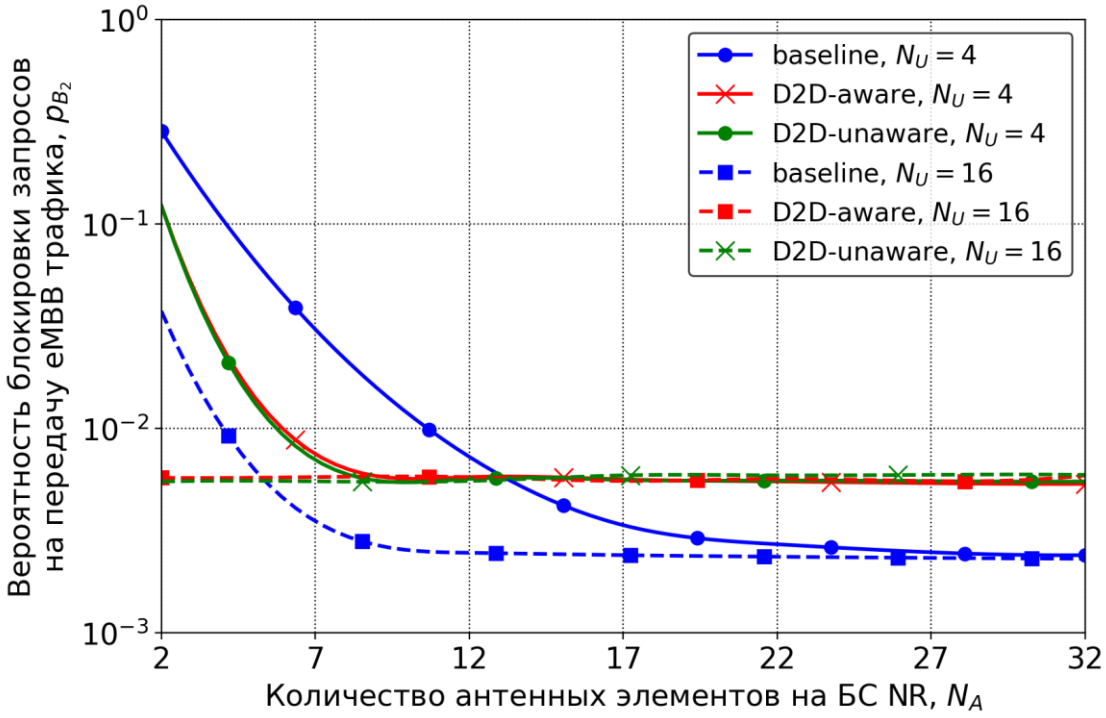

Рис. 2.7. Вероятность блокировки запросов на передачу eMBB трафика в зависимости от количества антенных элементов.

Рис. 2.8 иллюстрирует зависимость вероятности блокировки запросов на передачу URLLC и eMBB трафика от плотности развертывания базовых станций χ и антенных решеток БС при $c_1 = 2$ Мбит/с, $c_2^{\min} = 1$ Мбит/с, $\mu_1 = 10^3$, $\mu_2 = 1/120$. Как можно заметить, увеличение плотности расположения БС минимизирует вероятность блокировки запросов на передачу URLLC трафика, в то время как вероятность блокировки запросов на передачу eMBB трафика изменяется незначительно.

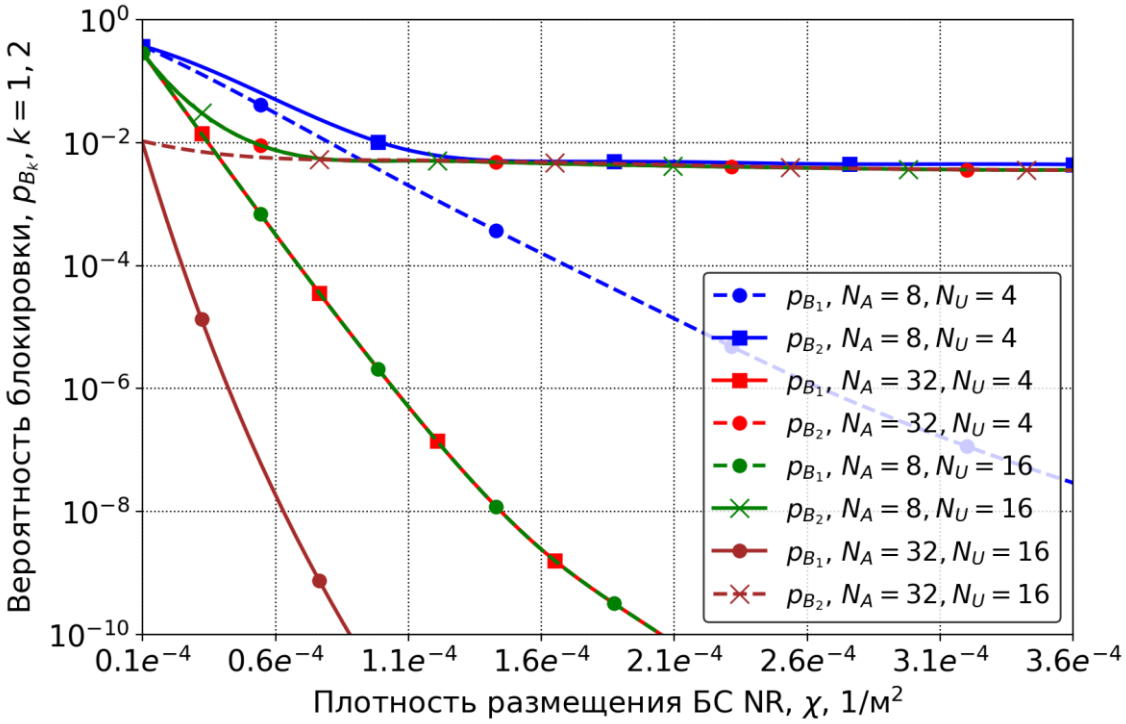

Рис. 2.8. Вероятность блокировки запросов на передачу URLLC/eMBB трафика в зависимости от плотности размещения базовых станций NR.

Полученные численные результаты показывают, что дисциплина обслуживания с приоритетами обеспечивает достаточную изоляцию приоритетного URLLC трафика, критичного к задержкам, от eMBB трафика с более низким приоритетом. Стратегия с поддержкой D2D, при которой прямая передача координируется через базовую станцию, превосходит стратегию, при которой координация через БС не используется, а также базовую стратегию, при которой весь трафик направляется через БС NR. Кроме того, число антенных элементов значительно влияет на показатели производительности: на количество необходимых для передачи данных ресурсов и на возникновение помех в рамках стратегии, не предусматривающей D2D-передачи. Предложенные методы могут быть использованы для оценки требуемой плотности размещения базовых станций NR в промышленных развертываниях таким образом, чтобы были обеспечены гарантии производительности для URLLC и eMBB трафика.

Глава 3. Модели совместного обслуживания трафика с приоритизацией и разделением ресурсов

Рассмотрим далее модели с произвольным числом услуг, генерирующих различные типы трафика, разделением ресурсов и приоритетным обслуживанием (№3 в табл. 1.2). В данной главе получены модели, использующие механизм резервирования индивидуальных зон для запросов каждого типа помимо механизма прерывания передачи менее приоритетного трафика, рассмотренного в главе 2. Основной целью данной главы является получение стационарного распределения вероятностей состояний системы двумя способами: путем численного и аналитического решения СУР для модели, не предусматривающей прерывания обслуживания запросов, а также сравнительный анализ пяти стратегий разделения ресурсов: I) отсутствие резервирования и приоритетов; II) приоритетное обслуживание с прерыванием, но без резервирования; III) полное резервирование ресурсов; IV) частичное резервирование без прерывания; V) частичное резервирование с прерыванием.

3.1. Модель схемы доступа к ресурсам мультисервисной сети

Для разработки моделей совместного обслуживания различных типов трафика с учетом свойств методов резервирования полосы пропускания и приоритетного обслуживания на участке доступа были исследованы способы реализации концепции сетевого слайсинга, являющегося важной функциональной особенностью систем 5G NR [42]. Слайсинг позволяет использовать одну и ту же физическую инфраструктуру для разделения ресурсов между операторами мобильных виртуальных сетей (англ. Mobile Virtual Network Operators, MVNO) или различными типами трафика [43]. Каждый оператор или тип трафика может предъявлять разные требования к качеству обслуживания, включая требования к пропускной способности, задержке и вероятности блокировки.

Ключевую роль в обеспечении требуемого уровня QoS для различных типов трафика играет сеть радиодоступа, условия радиоканала в которой могут динамически изменяться. Кроме того, согласно единому мнению стандартизирующих организаций (3GPP, ITU-R и GSMA) [44, 45], метод, используемый для сетевого слайсинга должен удовлетворять двум противоречивым требованиям: (а) строгая изоляция трафика и (б) эффективное использование радиоресурсов. Первое требование [22, 24] можно реализовать с помощью статических или динамических политик резервирования ресурсов [46, 47], тогда как второе – с помощью схем приоритетного обслуживания [14, 38]. Реализация схем, учитывающих оба типа требований, достаточно сложна, поэтому требуется разработка математических моделей, показывающих оптимальные варианты совместного использования различных схем доступа к радиоресурсам сети.

Рассмотрим развертывание базовой станции с зоной покрытия радиуса r (рис. 3.1). Предположим, что функция сетевого слайсинга используется для предоставления K типов услуг с различными требованиями к уровню QoS. Предположим, что пользовательские устройства, генерирующие запросы на предоставление услуг, равномерно распределены в зоне покрытия соты. Высота базовой станции и пользовательского оборудования равна h_{BS} метров и h_{UE} метров соответственно.

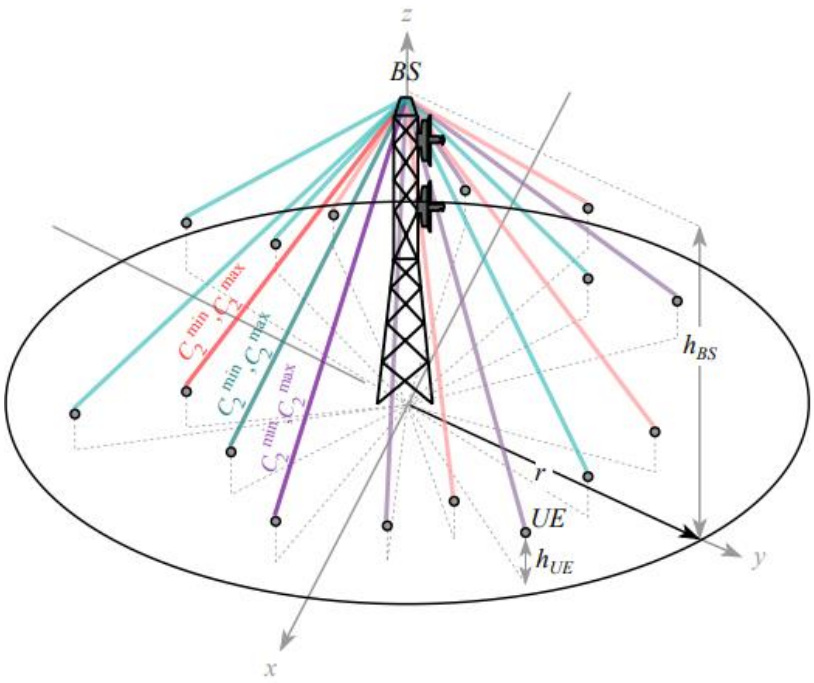

Рис. 3.1. Рассматриваемая системная модель.

На основе методов теории массового обслуживания (ТМО) [48, 49] сформулируем задачу сетевого слайсинга для K типов трафика. В качестве схемы доступа к радиоресурсам будем рассматривать совместное использование механизма резервирования ресурсов и приоритетного обслуживания. Модель будет описана в виде СМО [50], в которой каждому слайсу выделяется определенная индивидуальная зона емкостью g_k каналов, $k=1,...,K$ такая, что $\sum_{k=1}^{K} g_k < C$, где C каналов – емкость всей системы (рис. 3.2). Для повышения степени статистического мультиплексирования использование каналов из общего пула $c = C - \sum_{k=1}^{K} g_k$ регулируется процедурой приоритетного обслуживания. Отметим, что требования к ресурсам системы являются эластичными, т.е. могут меняться в некоторых пределах $\left(b_k^{\min}, b_k^{\max}\right)$, $k=1,...,K$, выражены в БЦК, соответствующих скоростям передачи данных $c_k^{\min}$ и $c_k^{\max}$, и могут быть определены, исходя из свойств радиоканала [51]. Введем вектора $\mathbf{b}_{\min} = \left(b_1^{\min}, b_2^{\min}, ..., b_K^{\min}\right)^T$ и

$\mathbf{b}_{\max} = \left(b_1^{\max}, b_2^{\max}, ..., b_K^{\max}\right)^T$. Запросы k-го типа поступают с интенсивностями λ_k, образуя пуассоновские потоки. Время обслуживания запросов распределено по экспоненциальному закону со средним μ_k^{-1}.

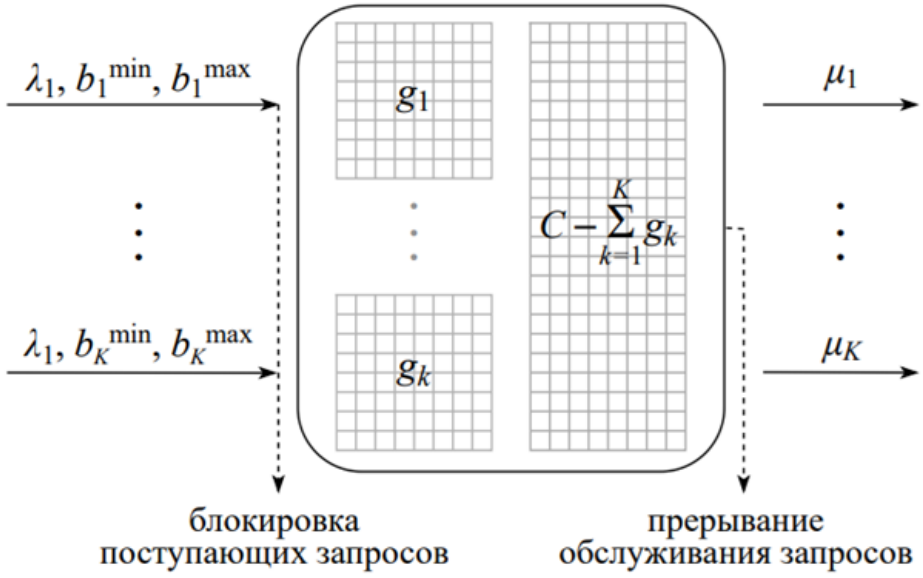

Рис. 3.2. Рассматриваемая СМО.

Функционирование рассматриваемой системы описывает K-мерный марковский СП $\left\{\left(N_1(t),...,N_K(t)\right), t \geq 0\right\}$, где $N_k(t)$, $k = 1,...,K$ – число обслуживаемых системой запросов k-го типа в момент времени t. Обозначим максимальное число запросов k-го типа, которое может находиться в системе, $N_k = \left\lfloor \left(C - \sum_{i \neq k} g_i\right) / b_k^{\min} \right\rfloor$, тогда число обслуживаемых системой запросов k-го типа $n_k = 0,...,N_K$, $i,k = 1,...,K$. Определим также максимальное гарантированное число запросов k-го типа, которое может быть обслужено в системе, как $N_k^g = \left\lfloor g_k / b_k^{\min} \right\rfloor$, $k = 1,...,K$.

Состояние системы описывает K-мерный вектор $\mathbf{n} = \left(n_1,...,n_K\right)$ над пространством состояний

$$\mathbf{X} = \left\{ \mathbf{n} : 0 \leq n_k \leq N_k, k = 1,...,K, \sum_{i=1}^{K} \max\left\{n_i b_i^{\min}, g_i\right\} < C \right\}. \quad (3.1)$$

Отметим, что число БЦК $b_k(\mathbf{n})$, $b_k^{\min} \le b_k(\mathbf{n}) \le b_k^{\max}$, выделяемых при обслуживании запросов k-го типа [52], может изменяться в зависимости от состояния системы и определяется следующим образом:

$$b_k(\mathbf{n}) = \\ = \min\left\{ \frac{C - \max\left\{ \sum\limits_{i \ne k} g_i, \sum\limits_{i=1}^{k-1} \max\left\{ n_i b_i(\mathbf{n}), g_i \right\} + \sum\limits_{i=k+1}^{K} \max\left\{ n_i b_i^{\min}, g_i \right\} \right\}}{n_k}, b_k^{\max} \right\}. \quad (3.2)$$

Пусть запросы разных типов имеют разные приоритеты в обслуживании, упорядоченные следующим образом: запросы первого типа имеют высший приоритет, запросы K-го типа – низший. Если при поступлении в систему запроса k-го типа БЦК для его обслуживания с минимальным требованием $b_k^{\min}$ недостаточно, обслуживание одного или нескольких менее приоритетных запросов может быть прекращено [53]. Предположим, что K-мерный вектор определяет число запросов каждого типа, обслуживание которых необходимо прервать при поступлении запроса типа j, $\mathbf{m} = (m_1, ..., m_K)$.

Пространство состояний системы $\mathbf{X}$ для каждого типа запросов является объединением множеств приема запросов $\mathbf{S}_k^{pre}$ и блокировки запросов $\mathbf{B}_k$. Множество приема запросов $\mathbf{S}_k^{pre}$ – множество состояний, в которых поступающие в систему запросы k-го типа принимаются на обслуживание:

$$\mathbf{S}_k^{pre} = \left\{ \mathbf{n} \in \mathbf{X} : n_k < N_k, \left(\sum_{i=1}^{k-1} \max\left\{ n_i b_i(\mathbf{n} + \mathbf{e}_k), g_i \right\} + \sum_{i=k+1}^{K} \max\left\{ n_i b_i^{\min}, g_i \right\} + \right. \right. \\ + (n_k + 1) b_k^{\min} \le C \vee \left(k < K, \sum_{i=1}^{k-1} \max\left\{ n_i b_i(\mathbf{n} + \mathbf{e}_k), g_i \right\} + (n_k + 1) b_k^{\min} + \right. \\ \left. \left. \left. + \sum_{i=k+1}^{K} \max\left\{ n_i b_i^{\min}, g_i \right\} - \sum_{i=k+1}^{K} b_i^{\min} (n_i - N_i^g) \cdot I(n_i > N_i^g) \le C \right) \right) \right\}, k = 1, ..., K. \quad (3.3)$$

Из множества приема запросов $\mathbf{S}_k^{pre}$ можно выделить подмножество $\mathbf{S}_k$ – множество состояний, в которых поступающие запросы k-го типа принимаются в систему, не прерывая обслуживание запросов менее приоритетных типов:

$$\mathbf{S}_k = \left\{ \mathbf{n} \in \mathbf{X} : n_k < N_k, \sum_{i=1}^{k-1} \max\left\{ n_i b_i \left(\mathbf{n} + \mathbf{e}_k \right), g_i \right\} + \right.$$
$$\left. + \sum_{i=k+1}^{K} \max\left\{ n_i b_i^{\min}, g_i \right\} + \left(n_k + 1 \right) b_k^{\min} \le C \right\}, k = 1,...,K. \tag{3.4}$$

Множество блокировки запросов $\mathbf{B}_k$ – множество состояний, в которых поступающие в систему запросы k-го типа блокируются из-за отсутствия свободных ресурсов:

$$\mathbf{B}_k = \left\{ \mathbf{n} \in \mathbf{X} : n_k = N_k \vee \sum_{i=1}^{k-1} \max\left\{ n_i b_i \left(\mathbf{n} + \mathbf{e}_k \right), g_i \right\} + \sum_{i=k+1}^{K} \max\left\{ n_i b_i^{\min}, g_i \right\} + \right.$$
$$\left. + \left(n_k + 1 \right) b_k^{\min} - \sum_{i=k+1}^{K} b_i^{\min} \left(n_i - N_i^g \right) \cdot I\left(n_i > N_i^g \right) > C \right\}, k = 1,...,K. \tag{3.5}$$

Множество прерывания запросов $\mathbf{\Pi}_k$ – множество состояний, в которых обслуживание одного или нескольких запросов, имеющих меньший приоритет, прерывается при поступлении в систему более приоритетного запроса k-го типа:

$$\mathbf{\Pi}_k = \left\{ \mathbf{n} \in \mathbf{X} : n_k < N_k, \left(\sum_{i=1}^{k-1} \max\left\{ n_i b_i \left(\mathbf{n} + \mathbf{e}_k \right), g_i \right\} + \sum_{i=k+1}^{K} \max\left\{ n_i b_i^{\min}, g_i \right\} + \right.$$
$$+ \left(n_k + 1 \right) b_k^{\min} > C, \sum_{i=1}^{k-1} \max\left\{ n_i b_i \left(\mathbf{n} + \mathbf{e}_k \right), g_i \right\} + \sum_{i=k+1}^{K} \max\left\{ n_i b_i^{\min}, g_i \right\} + \tag{3.6}$$
$$\left. \left. + \left(n_k + 1 \right) b_k^{\min} - \sum_{i=k+1}^{K} b_i^{\min} \left(n_i - N_i^g \right) \cdot I\left(n_i > N_i^g \right) \le C \right) \right\}, k = 1,...,K-1.$$

Кроме того, из множества $\mathbf{S}_k^{pre}$ можно выделить подмножество $\mathbf{S}_k^{\max}$ – множество состояний, в которых поступающие в систему запросы k-го типа принимаются на обслуживание с использованием максимального числа БЦК:

$$\mathbf{S}_k^{\max} = \left\{ \mathbf{n} \in \mathbf{X} : n_k < \left\lfloor \frac{C - \sum\limits_{i \neq k} g_i}{b_k^{\max}} \right\rfloor, \sum_{i=1}^{k-1} \max\left\{ n_i b_i \left(\mathbf{n} + \mathbf{e}_k \right), g_i \right\} + \right.$$
$$\left. + \left(n_k + 1 \right) b_k^{\max} + \sum_{i=k+1}^{K} \max\left\{ n_i b_i^{\min}, g_i \right\} \leq C \right\}, k, i = 1, \ldots, K. \qquad (3.7)$$

На рис. 3.3 представлен алгоритм выбора запросов, обслуживание которых должно быть прервано при поступлении более приоритетного запроса.

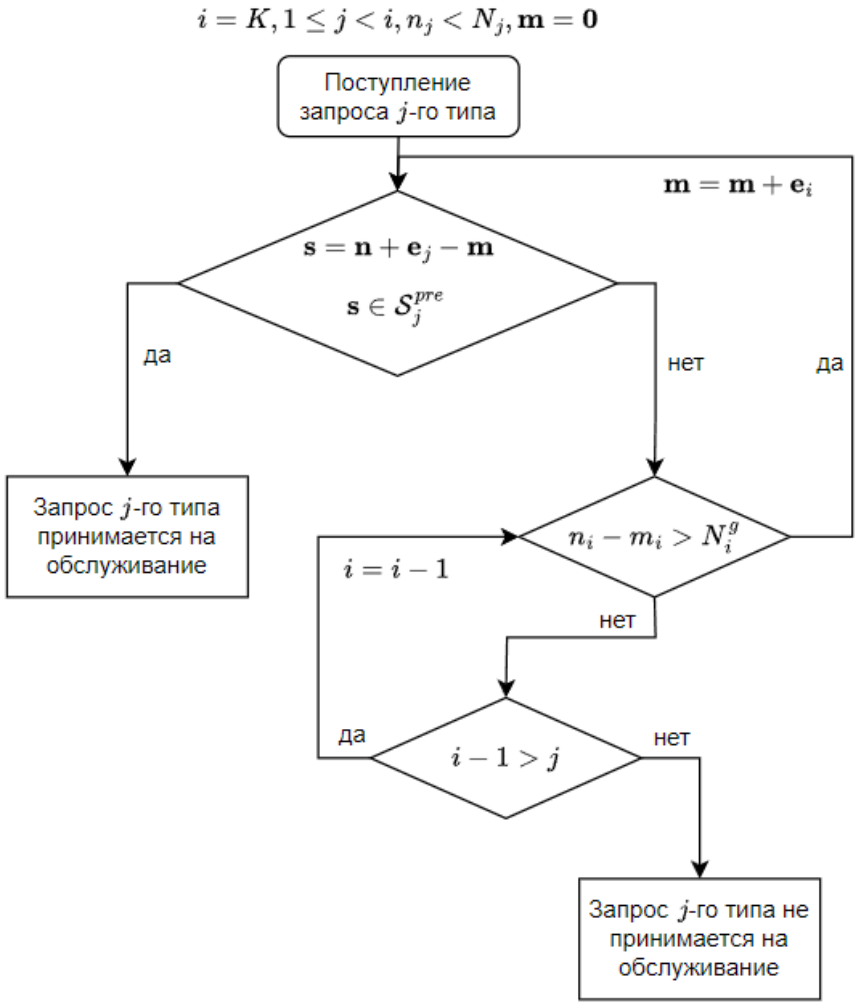

Рис. 3.3. Алгоритм выбора запросов, обслуживание которых должно быть прервано.

На основе описанных выше множеств можно сформулировать правила приема и обслуживания запросов k-го типа, $k = 1, \ldots, K$:

- если число запросов k-го типа, обслуживаемых в системе, меньше максимально возможного числа запросов данного типа N_k, и число

свободных каналов, доступных для запросов этого типа, составляет не менее $b_k^{\min}$ БЦК, то поступающий в систему запрос k-го типа принимается на обслуживание;

- если число запросов k-го типа, обслуживаемых в системе, меньше максимально возможного числа запросов k-го типа N_k, число свободных каналов, доступных для запросов данного типа, меньше $b_k^{\min}$ БЦК, а число каналов из общего пула c, занятых обслуживанием менее приоритетных запросов, составляет не менее $b_k^{\min}$ БЦК, то поступающий запрос k-го типа принимается на обслуживание за счет прерывания обслуживания $\sum_{i=2}^{K} m_i$ запросов менее приоритетных типов из общего пула, $k \neq K$;
- в противном случае поступающий в систему запрос k-го типа блокируется.

Составим диаграмму интенсивностей переходов (рис. 3.4).

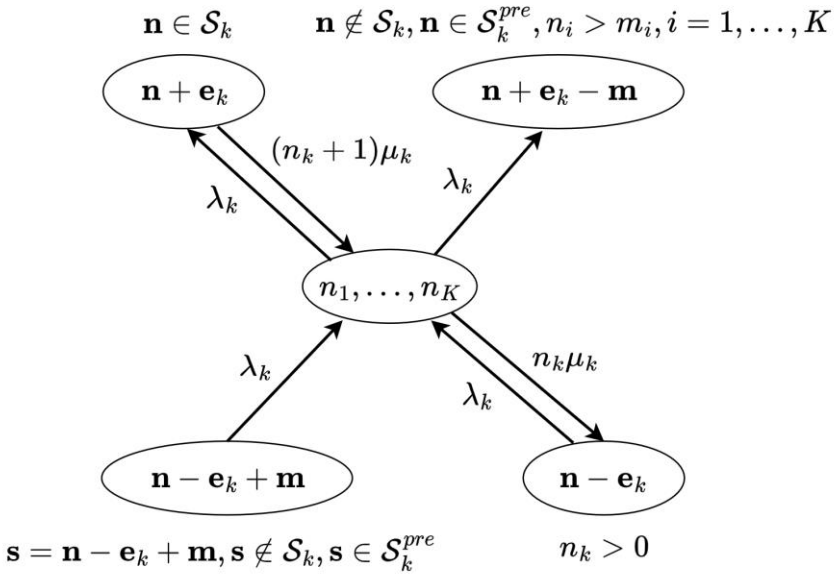

Рис. 3.4. Диаграмма интенсивностей переходов для центрального состояния модели с произвольным числом услуг, разделением ресурсов и прерыванием обслуживания неприоритетного трафика.

В связи с реализацией механизма прерывания обслуживания случайный процесс, описывающий поведение системы, не является обратимым, поэтому стационарное распределение $p(\mathbf{n})$, $\mathbf{n} \in \mathbf{X}$, не имеет аналитического решения.

Утверждение 3.1. Стационарное распределение $p(\mathbf{n})$, $\mathbf{n} \in \mathbf{X}$, вероятностей состояний СП $\{(N_1(t),...,N_K(t)), t \geq 0\}$ определяется путем численного решения СУР: $\mathbf{p}^T \mathbf{A} = \mathbf{0}^T$, $\mathbf{p}^T \mathbf{1} = 1$, где элементы инфинитезимальной матрицы $\mathbf{A}$ определяются следующим образом:

$$a(\mathbf{n},\mathbf{n}') = \begin{cases} \lambda_k, & \text{если } \mathbf{n}' = \mathbf{n} + \mathbf{e}_k, \mathbf{n} \in \mathbf{S}_k, k = 1,...,K, \\ & \text{или } \mathbf{n}' = \mathbf{n} - \mathbf{m} + \mathbf{e}_k, \mathbf{n} \notin \mathbf{S}_k, \mathbf{n} \in \mathbf{S}_k^{pre}, n_i > m_i, k,i = 1,...,K; \\ n_k \mu_k, & \text{если } \mathbf{n}' = \mathbf{n} - \mathbf{e}_k, n_k > 0, k = 1,...,K; \\ \varphi, & \text{если } \mathbf{n}' = \mathbf{n}, k = 1,...,K; \\ 0 & \text{в ином случае,} \end{cases} \quad (3.8)$$

где $\varphi = -\left[\sum_{k=1}^{K} \lambda_k \cdot I\{\mathbf{n} \in \mathbf{S}_k\} + \lambda_k \cdot I\left\{\mathbf{n} \notin \mathbf{S}_k, \mathbf{n} \in \mathbf{S}_k^{pre}, n_i > m_i, k,i = 1,...,K\right\} + n_k \mu_k \right]$.

Получив распределение вероятностей $p(\mathbf{n})$, $\mathbf{n} \in \mathbf{X}$, можем вычислить основные характеристики модели: вероятность блокировки запросов, вероятность прерывания обслуживания, среднее число каналов, занятых запросами, а также коэффициент использования ресурсов.

Утверждение 3.2. Вероятность блокировки запросов k-го типа рассчитывается по формуле

$$p_{B_k} = \sum_{n_1=0}^{N_1} ... \sum_{n_K=0}^{N_K} p(\mathbf{n}) \cdot I\{\mathbf{n} \in \mathbf{B}_k\}, k = 1,...,K. \quad (3.9)$$

Утверждение 3.3. Вероятность прерывания обслуживания запросов при поступлении запроса k-го типа рассчитывается по формуле

$$p_{pre_k} = \sum_{n_1=0}^{N_1} ... \sum_{n_K=0}^{N_K} p(\mathbf{n}) \cdot I\{\mathbf{n} \in \mathbf{\Pi}_k\}, k = 1,...,K-1. \quad (3.10)$$

Утверждение 3.4. Среднее число каналов, занятых запросами i-го типа, вычисляется по формуле

$$\bar{k}_i = \sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} n_i b_i(\mathbf{n}) p(\mathbf{n}) \cdot I\{\mathbf{n} \in \mathbf{X}\}, k = 1, \dots, K. \quad (3.11)$$

Утверждение 3.5. Коэффициент использования ресурсов вычисляется по формуле

$$U = \sum_{i=1}^{K} \sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} n_i b_i(\mathbf{n}) p(\mathbf{n}) \cdot I\{\mathbf{n} \in \mathbf{X}\}. \quad (3.12)$$

В качестве примера функционирования системы рассмотрим стратегию совместного обслуживания с использованием сетевого слайсинга в сетях 5G NR [54] трех типов трафика – URLLC, eMBB и mMTC [10, 38, 46], поддержка которых должна быть обеспечена радиоинтерфейсом 5G NR.

Для URLLC трафика должны быть обеспечены требования 5G International Mobile Telecommunication-2020 и рекомендации по слайсингу, определенные в 3GPP TS 23.501 [31], а именно наивысший приоритет и вероятность блокировки не выше 10^{-5}. При этом mMTC трафик имеет менее жесткие требования, в частности задержку не более 10 секунд, в то время как eMBB трафик характеризуется сбалансированными требованиями. Для расстановки приоритетов в рамках рассматриваемой стратегии данные условия будут учтены.

Для рассматриваемой СМО интерес представляют два типа метрик: (i) ориентированные на пользователя и (ii) ориентированные на оператора. Первый тип метрик включает в себя такие показатели как вероятность блокировки и вероятность прерывания обслуживания запроса. С точки зрения системы интерес представляет коэффициент использования ресурсов.

Сравним две схемы разделения ресурсов: I) полное резервирование, при котором общая емкость ресурсов равномерно распределена между слайсами и II) частичное резервирование с приоритетами, при котором часть ресурсов выделяется в общий пул. Параметры системы представлены в таблице 3.1.

Таблица 3.1. Параметры системы для численных расчетов.

Параметр	**Значение**	
	Схема I	**Схема II**
Интенсивность поступления запросов на передачу URLLC трафика	10 000 запросов/с	
Интенсивность поступления запросов на передачу eMBB трафика	1 запрос/с	
Интенсивность поступления запросов на передачу mMTC трафика	10 000 запросов/с	
Среднее время обслуживания запроса на передачу URLLC трафика	1 мс	
Среднее время обслуживания запроса на передачу eMBB трафика	10 с	
Среднее время обслуживания запроса на передачу mMTC трафика	1 мс	
Общая емкость ресурсов	39 БЦК	
Емкость ресурса для предоставления услуг URLLC	13 БЦК	10 БЦК
Емкость ресурса для предоставления услуг eMBB	13 БЦК	10 БЦК
Емкость ресурса для предоставления услуг mMTC	13 БЦК	10 БЦК
Требование к ресурсам для предоставления услуги URLLC	1 БЦК	
Минимальное требование к ресурсам для предоставления услуги eMBB	1 БЦК	
Максимальное требование к ресурсам для предоставления услуги eMBB	3 БЦК	
Требование к ресурсам для предоставления услуги mMTC	1 БЦК	

На рис. 3.5 представлено сравнение показателя вероятности блокировки запросов для eMBB и mMTC трафика p_{B_2} и p_{B_3} в зависимости от предложенной нагрузки, создаваемой запросами на передачу URLLC трафика, $\rho_1 = \lambda_1 b_1^{\min} / \mu_1$. Для схемы полного резервирования $g_k = 13$, для схемы частичного резервирования $g_k = 10$. В случае частичного резервирования выделен общий пул из 9 БЦК, которые используются всеми типами трафика. Как видно, механизм, основанный на частичном резервировании и приоритетах, превосходит механизм, основанный на полном резервировании с точки зрения значения вероятности блокировки запросов на передачу mMTC трафика для небольших значений ρ_1. Для eMBB и mMTC трафика система характеризуется двумя режимами: до $\rho_1 = 15$ лучше показывает себя частичное резервирование, однако при больших значениях ρ_1 схема полного резервирования превосходит ее. Объясняется это тем, что до $\rho_1 = 15$ eMBB и mMTC трафик эффективно используют общий пул доступных ресурсов. Однако при увеличении предложенной нагрузки, создаваемой запросами на передачу URLLC трафика с наивысшим приоритетом, соответствующие запросы занимают общий пул ресурсов, что приводит к потерям eMBB и mMTC трафика.

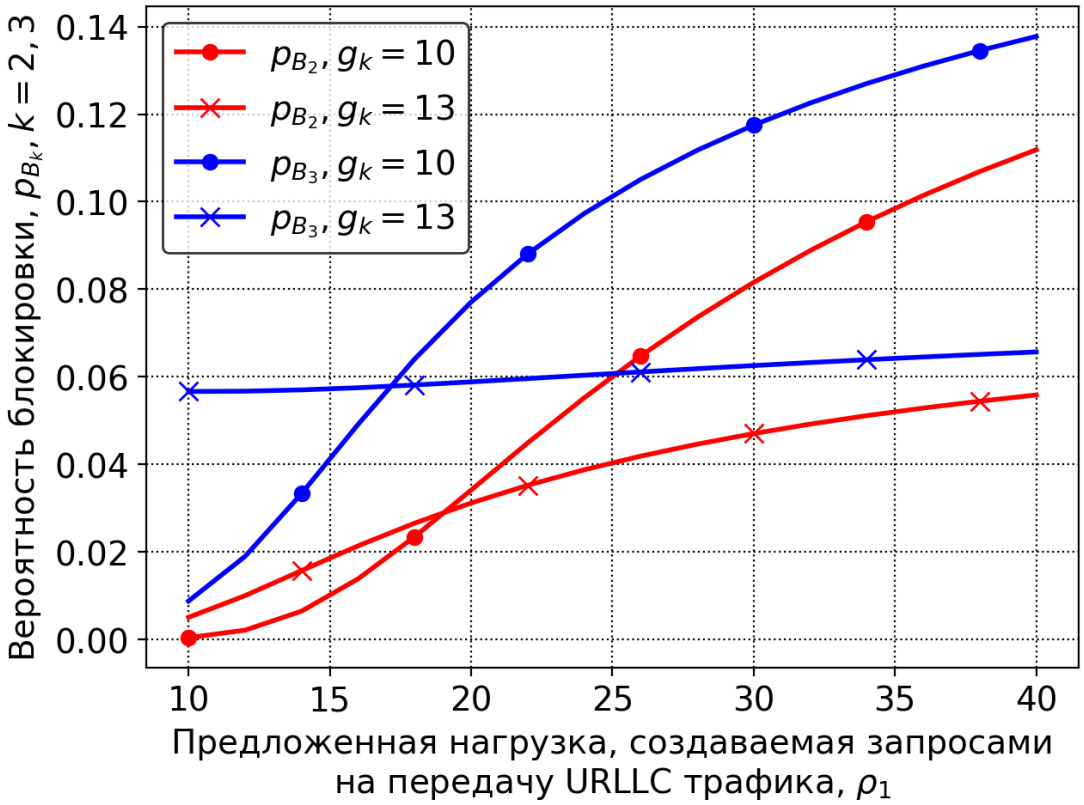

Рис. 3.5. Зависимость вероятности блокировки запросов на передачу eMBB/mMTC трафика от предложенной нагрузки, создаваемой запросами на передачу URLLC трафика.

Однако снижение вероятностей блокировки в схеме частичного резервирования приводит к более высоким вероятностям прерывания обслуживания запросов, как показано на рис. 3.6. Отметим, что вероятность прерывания обслуживания запросов менее приоритетных типов – eMBB и mMTC – ниже для схемы полного резервирования практически во всем диапазоне значений ρ_1. При этом больше всего заметно увеличение вероятностей прерывания обслуживания запросов на передачу eMBB трафика – разница между значениями для схем достигает трех раз при $\rho_1 = 20$.

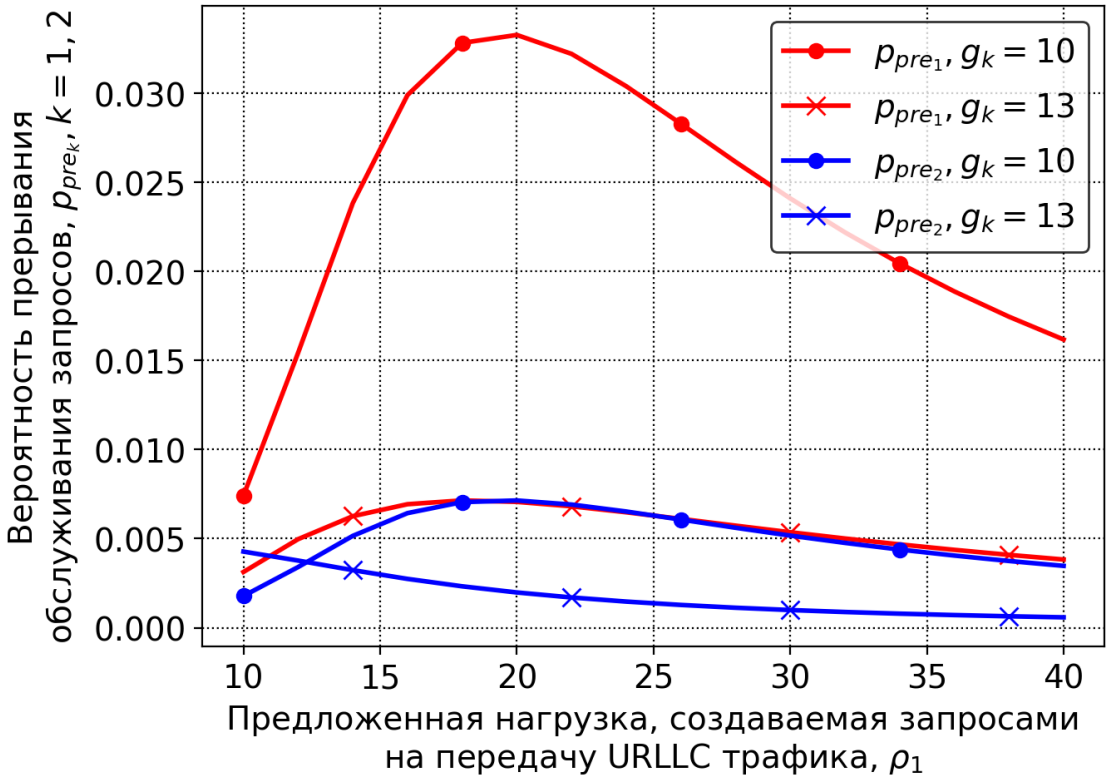

Рис. 3.6. Зависимость вероятности прерывания обслуживания при поступлении запросов на передачу URLLC/eMBB трафика от предложенной нагрузки, создаваемой запросами на передачу URLLC трафика.

Стоит отметить, что кратковременный легкий mMTC трафик лучше подходит для приоритетного обслуживания с точки зрения количества запросов, обслуживание которых необходимо прервать для поступления запроса с более высоким приоритетом. В то же время вытеснение eMBB трафика, требующего длительной передачи, оказывает более негативное влияние по сравнению с вытеснением mMTC трафика, что делает схему полного резервирования более подходящей для трафика, генерируемого услугами межмашинной связи – mMTC.

На рис. 3.7 можно увидеть, что во всем диапазоне значений ρ_1 схема частичного резервирования показывает на 10% лучшие результаты с точки зрения коэффициента использования ресурсов по сравнению с альтернативной схемой. Отметим, что в условиях высокой нагрузки на систему использование ресурсов снижается всего на 5%.

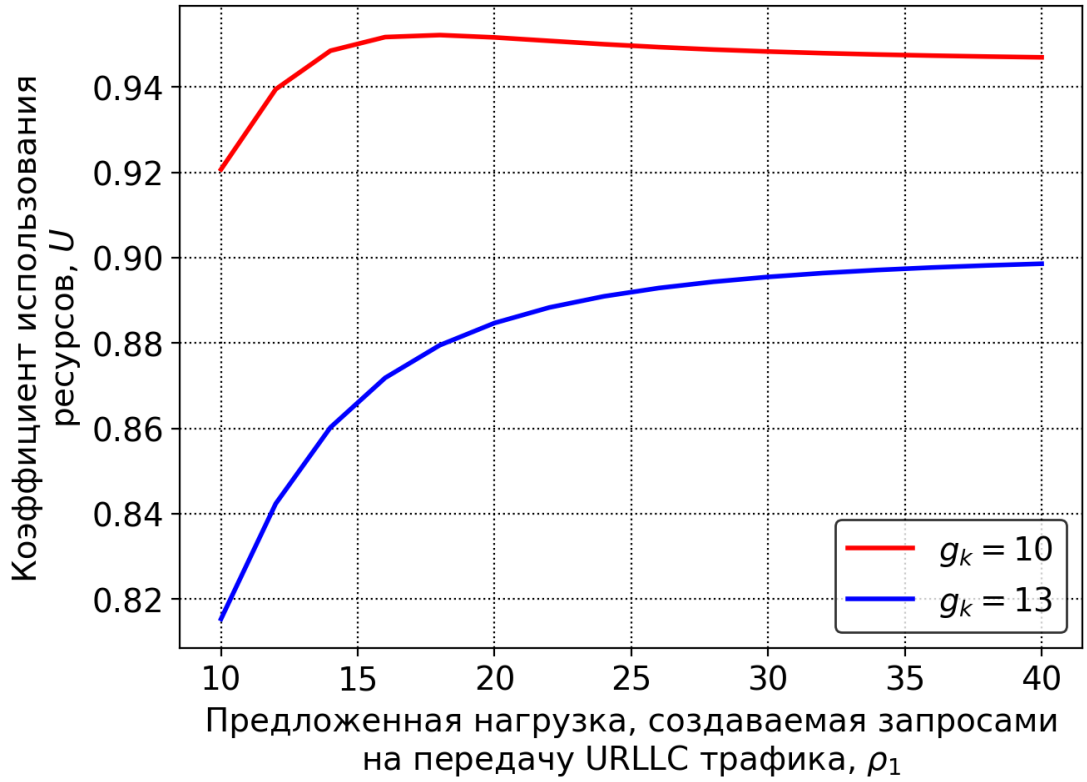

Рис. 3.7. Зависимость коэффициента использования ресурсов от предложенной нагрузки, создаваемой запросами на передачу URLLC трафика.

Аналогичные результаты наблюдаются на рис. 3.8, где показан коэффициент использования ресурсов в зависимости от максимального требования к ресурсам $b_2^{\max}$, предъявляемого при передаче eMBB трафика. При увеличении предложенной нагрузки схема частичного резервирования позволяет достичь 95% использования ресурсов, в то время как для схемы полного резервирования этот показатель достигает 80%.

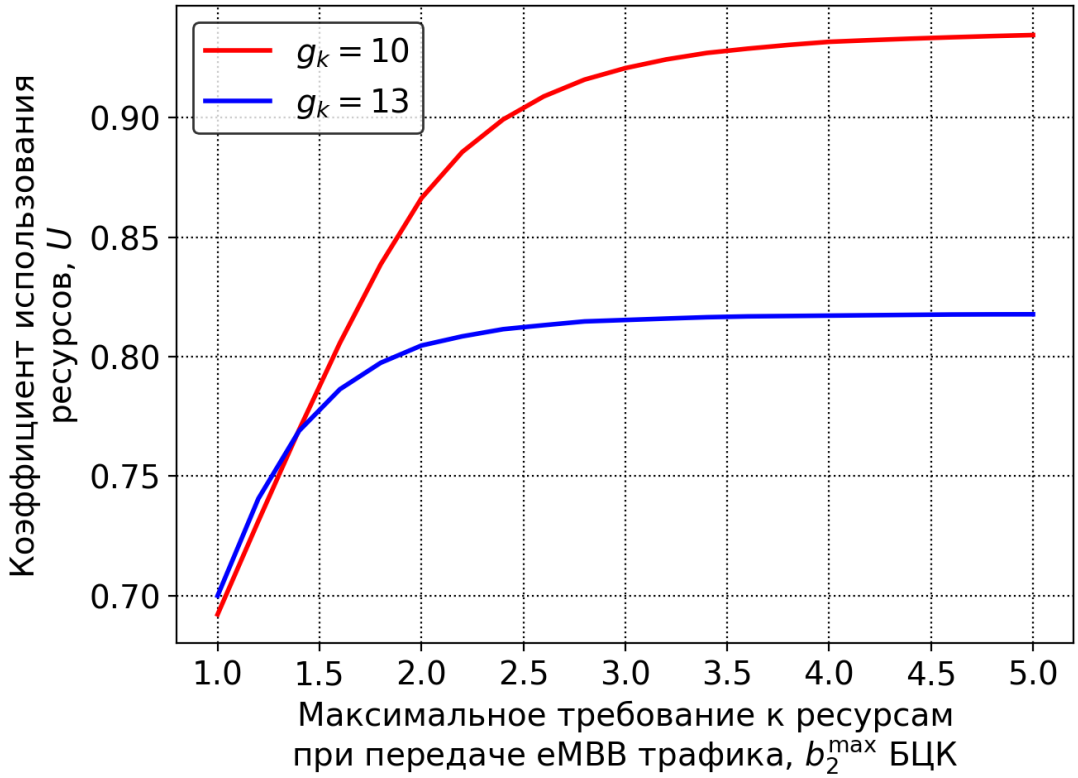

Рис. 3.8. Зависимость коэффициента использования ресурсов от максимального требования к ресурсам при передаче eMBB трафика.

Полученные численные результаты показывают, что предлагаемая стратегия, основанная на резервировании и приоритетах, позволяет повысить эффективность использования ресурсов до 95%, обеспечивая при этом изоляцию между типами трафика в условиях высокой предложенной нагрузки. Данная схема также имеет преимущество с точки зрения вероятности блокировки приоритетных запросов при низкой и умеренной предложенной нагрузке. В то же время, эластичный eMBB трафик имеет лучшие показатели в случае использования стратегии полного резервирования как с точки зрения вероятности блокировки, так и с точки зрения вероятности прерывания обслуживания.

3.2. Частный случай модели с резервированием индивидуальных зон без прерывания обслуживания неприоритетного трафика

В качестве частного случая рассмотрим систему емкостью C каналов, которая предоставляет услуги двух типов. Первый тип услуг генерирует

запросы на передачу потокового URLLC трафика, второй тип услуг – запросы на передачу эластичного eMBB трафика. Для краткости далее будем называть их запросами первого и второго типа соответственно. Запросы поступают в систему с интенсивностями λ_k, образуя пуассоновские потоки, а время обслуживания запросов распределено по экспоненциальному закону со средним μ_k^{-1}, $k=1,2$. Обозначим $\rho_k=\lambda_k/\mu_k$ интенсивность предложенной нагрузки, создаваемой запросами k-го типа, $k=1,2$. Для каждого типа услуг задано минимальное и максимальное число БЦК, требуемое для обслуживания, $1\le b_k^{\min}\le b_k^{\max}$, $k=1,2$. Обозначим $\mathbf{b}_{\min}=\left(b_1^{\min},b_2^{\min}\right)^T$, $\mathbf{b}_{\max}=\left(b_1^{\max},b_2^{\max}\right)^T$.

Предположим, что для запросов k-го типа выделены индивидуальные зоны емкостью g_k каналов, $g_1+g_2<C$ (рис. 3.9). Тогда $C-g_i$ – максимальное число каналов, доступных для услуги k-го типа, $i,k=1,2$, $i\ne k$, а $c=C-g_1-g_2$ – емкость общего пула для запросов любого типа.

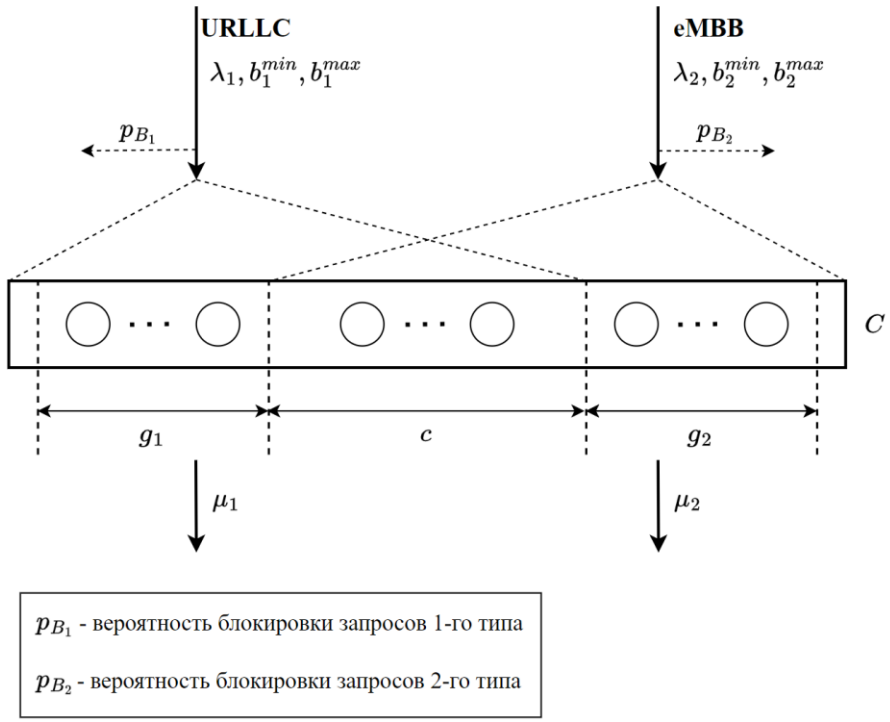

Рис. 3.9. Схема модели с резервированием для трафика URLLC и eMBB.

Функционирование системы описано двумерным марковским СП $\{(N_1(t), N_2(t)), t \geq 0\}$, где $N_1(t)$ и $N_2(t)$ – число обслуживаемых системой запросов на предоставление услуги первого и второго типа в момент времени t. Предположим, что максимальное число запросов на предоставление услуги k-го типа, которое может находиться в системе, $N_k = \left\lfloor \frac{C - \sum_{i \neq k} g_i}{b_k^{\min}} \right\rfloor$, $i,k = 1,2$, тогда $n_k = 0,...,N_k$ – число обслуживаемых системой запросов k-го типа. Состояние системы может быть описано двумерным вектором $\mathbf{n} = (n_1, n_2)$ над пространством состояний $\mathbf{X}$, имеющим вид

$$\mathbf{X} = \left\{ \mathbf{n} : 0 \leq n_k \leq N_k, \sum_{i=1}^{2} \max\left\{ n_i b_i^{\min}, g_i \right\} \leq C, k = 1,2 \right\}. \quad (3.13)$$

Отметим, что число БЦК $b_2(n_1, n_2)$, $b_2^{\min} \leq b_2(n_1, n_2) \leq b_2^{\max}$, необходимых для обслуживания запросов на предоставление услуг, генерирующих эластичный трафик, меняется в зависимости от количества запросов в системе и определяется следующим образом:

$$b_2(n_1, n_2) = \min\left\{ \frac{C - \max\left\{ n_1 b_1^{\min}, g_1 \right\}}{n_2}, b_2^{\max} \right\}, (n_1, n_2) \in \mathbf{X}. \quad (3.14)$$

Предположим, что $b_1^{\min} = b_1^{\max}$, $b_1^{\min} \leq b_2^{\min}$.

Обозначим максимальное число запросов второго типа при условии, что в системе уже обслуживается n_1 запросов первого типа, как

$$k(n_1) = \left\lfloor \frac{C - \max\left\{ n_1 b_1^{\min}, g_1 \right\}}{b_2^{\min}} \right\rfloor \leq N_2. \quad (3.15)$$

Тогда максимальное число запросов первого типа при условии, что в системе уже обслуживается n_2 запросов второго типа, можно обозначить как

$$l(n_2) = \left\lfloor \frac{C - \max\left\{ n_2 b_2^{\min}, g_2 \right\}}{b_1^{\min}} \right\rfloor \leq N_1. \quad (3.16)$$

Рассмотрим основные множества системы. Множество приема запросов $\mathbf{S}_k$ – это множество состояний системы, в которых поступающие запросы k-го типа принимаются на обслуживание, $k=1,2$:

$$\mathbf{S}_1=\left\{\mathbf{n}\in\mathbf{X}: n_1<N_1,\left(n_1+1\right)b_1^{\min}+\max\left\{n_2b_2^{\min},g_2\right\}\le C\right\}; \tag{3.17}$$

$$\mathbf{S}_2=\left\{\mathbf{n}\in\mathbf{X}: n_2<N_2,\max\left\{n_1b_1^{\min},g_1\right\}+\left(n_2+1\right)b_2^{\min}\le C\right\}. \tag{3.18}$$

Множество блокировки запросов $\mathbf{B}_k$ – это множество состояний, в которых поступающие в систему запросы k-го типа блокируются, $k=1,2$:

$$\mathbf{B}_1=\left\{\mathbf{n}\in\mathbf{X}: n_1=N_1\vee\left(n_1+1\right)b_1^{\min}+\max\left\{n_2b_2^{\min},g_2\right\}>C\right\}; \tag{3.19}$$

$$\mathbf{B}_2=\left\{\mathbf{n}\in\mathbf{X}: n_2=N_2\vee\max\left\{n_1b_1^{\min},g_1\right\}+\left(n_2+1\right)b_2^{\min}>C\right\}. \tag{3.20}$$

Из множества $\mathbf{S}_k$ можно выделить подмножество $\mathbf{S}_k^{\max}$ – множество состояний, в которых поступающие запросы k-го типа принимаются на обслуживание с использованием максимального числа БЦК, $k=1,2$:

$$\mathbf{S}_1^{\max}=\left\{\mathbf{n}\in\mathbf{X}: n_1<\left\lfloor\frac{C-g_2}{b_1^{\max}}\right\rfloor,\left(n_1+1\right)b_1^{\max}+\max\left\{n_2b_2^{\min},g_2\right\}\le C\right\}; \tag{3.21}$$

$$\mathbf{S}_2^{\max}=\left\{\mathbf{n}\in\mathbf{X}: n_2<\left\lfloor\frac{C-g_1}{b_2^{\max}}\right\rfloor,\max\left\{n_1b_1^{\max},g_1\right\}+\left(n_2+1\right)b_2^{\max}\le C\right\}. \tag{3.22}$$

Утверждение 3.6. Пространство состояний системы $\mathbf{X}$ для каждого типа запросов является объединением множеств приема запросов $\mathbf{S}_k$ и блокировки запросов $\mathbf{B}_k$, $\mathbf{S}_k\cup\mathbf{B}_k=\mathbf{X}$, $k=1,2$.

Доказательство. Из определения множеств приема запросов $\mathbf{S}_k$ (3.17, 3.18) и блокировки запросов $\mathbf{B}_k$ (3.19, 3.20) получим

$$\mathbf{S}_k\cup\mathbf{B}_k=\left\{\mathbf{n}\in\mathbf{X}: n_k<N_k,\left(n_k+1\right)b_k^{\min}+\max\left\{n_ib_i^{\min},g_i\right\}\le C\right\}\cup$$

$$\cup\left\{\mathbf{n}\in\mathbf{X}: n_k=N_k\vee\left(n_k+1\right)b_k^{\min}+\max\left\{n_ib_i^{\min},g_i\right\}>C\right\}, i\ne k, i,k=1,2;$$

$$\mathbf{S}_k \cup \mathbf{B}_k = \left\{\mathbf{n} \in \mathbf{X} : n_k < N_k, \left(n_k + 1\right) b_k^{\min} + \max\left\{n_i b_i^{\min}, g_i\right\} \le C \vee \right.$$

$$\left. \vee n_k = N_k \vee \left(n_k + 1\right) b_k^{\min} + \max\left\{n_i b_i^{\min}, g_i\right\} > C \right\}, i \ne k, i, k = 1, 2.$$

Пусть $a : n_k < N_k$, $b : \left(n_k + 1\right) b_k^{\min} + \max\left\{n_i b_i^{\min}, g_i\right\} \le C$, тогда $\bar{a} : n_k = N_k$, $\bar{b} : \left(n_k + 1\right) b_k^{\min} + \max\left\{n_i b_i^{\min}, g_i\right\} > C$, $i \ne k$, $i, k = 1, 2$. С учетом введенных обозначений объединение множеств $\mathbf{S}_k$ и $\mathbf{B}_k$ можно записать как $\left\{\mathbf{n} \in \mathbf{X} : a \wedge b \vee \bar{a} \vee \bar{b}\right\}$. По закону поглощения

$$a \wedge b \vee \bar{a} \vee \bar{b} = \left(a \wedge b \vee \bar{a}\right) \vee \bar{b} = \left(\bar{a} \vee b\right) \vee \bar{b} = \bar{a} \vee \left(b \vee \bar{b}\right) = 1.$$

Таким образом, $\mathbf{S}_k \cup \mathbf{B}_k = \mathbf{X}$, $k = 1, 2$.

Утверждение доказано. □

На основе описанных выше множеств можно сформулировать правила приема и обслуживания запросов k-го типа:

- если число запросов на предоставление услуги k-го типа, которые обслуживаются в системе, меньше максимально возможного числа запросов данного типа N_k, и число свободных каналов, доступных для запросов этого типа, составляет не менее $b_k^{\min}$, то запрос k-го типа принимается на обслуживание, $k = 1, 2$;
- в противном случае запрос на предоставление услуги k-го типа блокируется, $k = 1, 2$.

Составим диаграмму интенсивностей переходов в общем виде (рис. 3.10) и для центрального состояния (рис. 3.11).

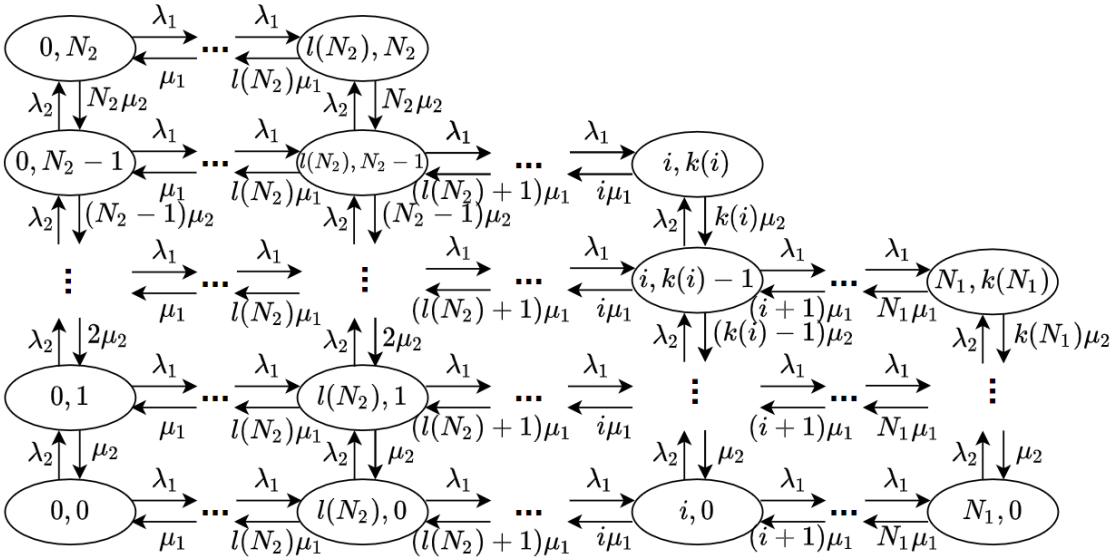

Рис. 3.10. Диаграмма интенсивностей переходов модели с разделением ресурсов без прерывания обслуживания неприоритетного трафика.

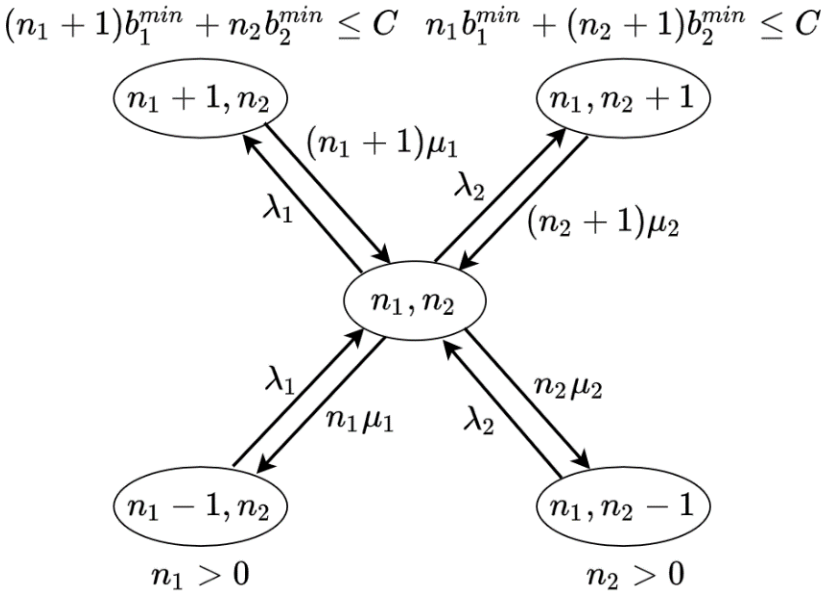

Рис. 3.11. Диаграмма интенсивностей переходов для центрального состояния модели с разделением ресурсов без прерывания обслуживания неприоритетного трафика.

Рассматриваемый случайный процесс является обратимым марковским СП и может быть описан системой уравнений глобального баланса в соответствии с диаграммой интенсивностей переходов (рис. 3.11):

$$\begin{aligned}&\left(\lambda_1 \cdot I\left\{n_1 < N_1, (n_1+1)b_1^{\min} + n_2 b_2^{\min} \le C\right\} + \lambda_2 \cdot I\left\{n_2 < N_2, n_1 b_1^{\min} + \right.\right.\\ &\left.\left. + (n_2+1)b_2^{\min} \le C\right\} + n_1\mu_1 \cdot I\{n_1 > 0\} + n_2\mu_2 \cdot I\{n_2 > 0\}\right) p(n_1, n_2) = \\ &= (n_1+1)\mu_1 \cdot I\left\{n_1 < N_1, (n_1+1)b_1^{\min} + n_2 b_2^{\min} \le C\right\} p(n_1+1, n_2) + \\ &+ (n_2+1)\mu_2 \cdot I\left\{n_2 < N_2, n_1 b_1^{\min} + (n_2+1)b_2^{\min} \le C\right\} p(n_1, n_2+1) + \\ &+ \lambda_1 \cdot I\{n_1 > 0\}\, p(n_1-1, n_2) + \lambda_2 \cdot I\{n_2 > 0\}\, p(n_1, n_2-1),\end{aligned} \tag{3.23}$$

где $p(n_1,n_2),(n_1,n_2) \in \mathbf{X}$ – стационарное распределение вероятностей состояний системы.

Утверждение 3.7. Стационарное распределение вероятностей $p(n_1,n_2),(n_1,n_2) \in \mathbf{X}$ представимо в мультипликативном виде:

$$p(n_1,n_2) = \frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}\, p(0,0); \tag{3.24}$$

$$p(0,0) = \left(\sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2} \frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}\right)^{-1}. \tag{3.25}$$

Доказательство. Используя диаграмму интенсивностей переходов (рис. 3.10), запишем систему уравнений локального баланса

$$\begin{cases}\lambda_1 p(n_1-1, n_2) = n_1\mu_1 p(n_1,n_2),\ (n_1,n_2) \in \mathbf{X},\ n_1 > 0,\\ \lambda_2 p(n_1, n_2-1) = n_2\mu_2 p(n_1,n_2),\ (n_1,n_2) \in \mathbf{X},\ n_2 > 0.\end{cases}$$

С учетом условия нормировки $\sum_{(n_1,n_2)\in\mathbf{X}} p(n_1,n_2) = 1$. Тогда

$$p(1,0) = \frac{\lambda_1}{\mu_1}\, p(0,0);\ p(0,1) = \frac{\lambda_2}{\mu_2}\, p(0,0);\ p(1,1) = \frac{\lambda_1}{\mu_1}\, p(0,1) = \frac{\lambda_1}{\mu_1}\frac{\lambda_2}{\mu_2}\, p(0,0);$$

$$p(2,0) = \frac{\lambda_1}{2\mu_1}\, p(1,0) = \frac{\lambda_1^2}{2\mu_1^2}\, p(0,0); \ldots; p(n_1,n_2) = \frac{\lambda_1^{n_1}}{n_1!\mu_1}\frac{\lambda_2^{n_2}}{n_2!\mu_2}\, p(0,0) = \frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}\, p(0,0).$$

Вычислим $p(0,0)$:

$$\sum_{(n_1,n_2)\in\mathbf{X}} \frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}\, p(0,0) = 1,\ p(0,0) \sum_{(n_1,n_2)\in\mathbf{X}} \frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!} = 1,\ p(0,0) = \left(\sum_{(n_1,n_2)\in\mathbf{X}} \frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}\right)^{-1}.$$

Утверждение доказано. □

Утверждение 3.8. Помимо мультипликативного решения, стационарное распределение вероятностей $\mathbf{p}=\{p(n_1,n_2),(n_1,n_2)\in\mathbf{X}\}$ может быть получено путем численного решения СУР: $\mathbf{p}^T\mathbf{A}=\mathbf{0}^T$, $\mathbf{p}^T\mathbf{1}=1$, где элементы $a\big((n_1,n_2),(n_1',n_2')\big)$ инфинитезимальной матрицы $\mathbf{A}$ для решения СУР определяются следующим образом:

$$a(\mathbf{n},\mathbf{n}')=$$

$$=\begin{cases}\lambda_1, & \text{если } n_1'=n_1+1,n_2'=n_2,n_1<N_1,(n_1+1)b_1^{\min}+n_2b_2^{\min}\le C;\\ \lambda_2, & \text{если } n_1'=n_1,n_2'=n_2+1,n_2<N_2,n_1b_1^{\min}+(n_2+1)b_2^{\min}\le C;\\ n_1\mu_1, & \text{если } n_1'=n_1-1,n_2'=n_2,n_1>0;\\ n_2\mu_2, & \text{если } n_1'=n_1,n_2'=n_2-1,n_2>0;\\ \varphi, & \text{если } n_1'=n_1,n_2'=n_2;\\ 0 & \text{в ином случае,}\end{cases} \tag{3.26}$$

$$\begin{aligned}\varphi=-\Big[&\lambda_1\cdot I\{n_1<N_1,(n_1+1)b_1^{\min}+n_2b_2^{\min}\le C\}+\\ &+\lambda_2\cdot I\{n_2<N_2,n_1b_1^{\min}+(n_2+1)b_2^{\min}\le C\}+n_1\mu_1+n_2\mu_2\Big].\end{aligned} \tag{3.27}$$

Рассчитав распределение вероятностей $\mathbf{p}$, можно вычислить такие вероятностные характеристики, как вероятность блокировки запросов, среднее число каналов, занятых запросами второго типа, а также коэффициент использования ресурсов.

Утверждение 3.9. Вероятность блокировки запросов первого типа рассчитывается по формуле

$$p_{B_1}=\sum_{j=0}^{N_2}\sum_{i=1}^{N_1}p(i,j)\cdot I\{i=N_1\vee(i+1)\cdot b_1^{\min}+j\cdot b_2^{\min}>C\}. \tag{3.28}$$

Утверждение 3.10. Вероятность блокировки запросов второго типа рассчитывается по формуле

$$p_{B_2}=\sum_{i=0}^{N_1}p\big(i,k(i)\big). \tag{3.29}$$

Доказательство. Используя множество блокировки запросов (3.20), определим состояния системы, в которых поступающие запросы второго типа будут заблокированы:

$$\mathbf{n} \in \mathbf{X} : n_2 = N_2 \vee \max\left\{n_1 b_1^{\min}, g_1\right\} + \left(n_2 + 1\right) b_2^{\min} > C;$$

$$\mathbf{n} \in \mathbf{X} : n_2 = N_2 \vee \left(n_2 + 1\right) b_2^{\min} > C - \max\left\{n_1 b_1^{\min}, g_1\right\};$$

$$\mathbf{n} \in \mathbf{X} : n_2 = N_2 \vee n_2 + 1 > \frac{C - \max\left\{n_1 b_1^{\min}, g_1\right\}}{b_2^{\min}}.$$

Учитывая максимальное число запросов второго типа (3.15), можем записать

$$\mathbf{n} \in \mathbf{X} : n_2 = \left\lfloor \frac{C - g_1}{b_2^{\min}} \right\rfloor \vee n_2 = \left\lfloor \frac{C - \max\left\{n_1 b_1^{\min}, g_1\right\}}{b_2^{\min}} \right\rfloor.$$

Следовательно, $\mathbf{n} \in \mathbf{X} : n_2 = k(n_1)$. Таким образом, просуммировав вероятности $p\left(n_1, k(n_1)\right)$ по n_1, получим вероятность блокировки (3.29).

Утверждение доказано. □

Утверждение 3.11. Среднее число каналов, занятых запросами второго типа, рассчитывается по формуле

$$\overline{k}_2 = \sum_{i=0}^{N_1} \sum_{j=1}^{k(i)} j \cdot b_2(i, j) \cdot p(i, j). \tag{3.30}$$

Утверждение 3.12. Коэффициент использования ресурсов вычисляется по формуле

$$\mathrm{UTIL} = \sum_{i=0}^{N_1} \sum_{j=0}^{k(i)} \left(i \cdot b_1^{\min} + j \cdot b_2(i, j)\right) \cdot p(i, j). \tag{3.31}$$

3.3. Частный случай модели с резервированием индивидуальных зон и прерыванием обслуживания неприоритетного трафика

Рассмотрим теперь аналогичную систему, в которой услуги имеют различные приоритеты в обслуживании, упорядоченные следующим образом – высший приоритет имеют услуги первого типа, низший – услуги второго

типа. Приоритетное обслуживание реализовано таким образом, что в случае недостаточности в системе БЦК для предоставления услуги 1-го типа с минимальным требованием $b_1^{\min}$ обслуживание одного или нескольких запросов на предоставление услуг 2-го типа может быть прекращено. Пусть двумерный вектор $\mathbf{m} = (m_1, m_2)$ определяет число запросов второго типа, обслуживание которых нужно прекратить для приема на обслуживание запроса первого типа. Определим максимальное гарантированное число запросов k-го типа, которое может быть обслужено в системе, как $N_k^g = \left\lfloor \frac{g_k}{b_k^{\min}} \right\rfloor$, $k = 1,2$.

Алгоритм 3.1. Выбор запросов второго типа, обслуживание которых должно быть прервано при поступлении запроса первого типа, реализован следующим образом:

$\mathbf{m} = (m_1, m_2) = (0,0)$

1: **repeat**

2: **if** $(n_1 + 1, n_2 - m_2) \in S_1^{pre}$ **then** запрос принимается на обслуживание

3: **else if** $n_2 - m_2 > N_2^g$ **then** $m_2 \leftarrow m_2 + 1$

4: **else** запрос не принимается на обслуживание

5: **until** $(n_1 + 1, n_2 - m_2) \notin S_1^{pre}$ **and** $n_2 > m_2 + N_2^g$

Пространство состояний системы определяется аналогично предыдущему разделу по формуле (3.13). Рассмотрим основные множества системы. Множество приема запросов $\mathbf{S}_k^{pre}$ – это множество состояний системы, в которых поступающие в систему запросы k-го типа принимаются на обслуживание, $k = 1,2$:

$$\mathbf{S}_1^{pre} = \Big\{\mathbf{n}\in\mathbf{X}: n_1 < N_1, \Big((n_1+1)b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} \le C \vee$$
$$\vee\Big((n_1+1)b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C, \qquad (3.32)$$
$$b_2^{\min}\left(n_2 - N_2^g\right)\cdot I\left\{n_2 > N_2^g\right\} \ge b_1^{\min}\Big)\Big)\Big\};$$

$$\mathbf{S}_2^{pre} = \Big\{\mathbf{n}\in\mathbf{X}: n_2 < N_2, \max\left\{n_1 b_1^{\min}, g_1\right\} + (n_2+1)b_2^{\min} \le C\Big\}. \qquad (3.33)$$

Множество блокировки запросов $\mathbf{B}_k$ – это множество состояний системы, в которых поступающие в систему запросы k-го типа блокируются, $k=1,2$:

$$\mathbf{B}_1^{pre} = \Big\{\mathbf{n}\in\mathbf{X}: n_1 = N_1 \vee \Big((n_1+1)b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C,$$
$$b_2^{\min}\left(n_2 - N_2^g\right)\cdot I\left\{n_2 > N_2^g\right\} < b_1^{\min}\Big)\Big\}; \qquad (3.34)$$

$$\mathbf{B}_2^{pre} = \Big\{\mathbf{n}\in\mathbf{X}: n_2 = N_2 \vee \Big(\max\left\{n_1 b_1^{\min}, g_1\right\} + (n_2+1)b_2^{\min} > C\Big)\Big\}. \qquad (3.35)$$

Множество прерывания запросов $\mathbf{\Pi}_1$ – это множество состояний системы, в которых обслуживание одного или нескольких запросов, имеющих меньший приоритет, прерывается при поступлении запросов первого типа:

$$\mathbf{\Pi}_1 = \Big\{\mathbf{n}\in\mathbf{X}: n_1 < N_1, (n_1+1)b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C,$$
$$b_2^{\min}\left(n_2 - N_2^g\right)\cdot I\left\{n_2 > N_2^g\right\} \ge b_1^{\min}\Big\}. \qquad (3.36)$$

Множество $\mathbf{S}_k^{\max}$ – множество состояний, в которых поступающие запросы k-го типа принимаются на обслуживание с использованием максимального числа БЦК, $k=1,2$, определено в формулах (3.21), (3.22).

Утверждение 3.13. Пространство состояний системы $\mathbf{X}$ для каждого типа запросов является объединением множеств приема запросов $\mathbf{S}_k^{pre}$ и блокировки запросов $\mathbf{B}_k^{pre}$, $\mathbf{S}_k^{pre} \cup \mathbf{B}_k^{pre} = \mathbf{X}$, $k=1,2$.

Доказательство. Из определения множеств приема запросов $\mathbf{S}_k^{pre}$ (3.32, 3.33) и блокировки запросов $\mathbf{B}_k^{pre}$ (3.34, 3.35) получим

$$\mathbf{S}_1^{pre} \cup \mathbf{B}_1^{pre} = \left\{\mathbf{n} \in \mathbf{X} : n_1 < N_1, \left(\left(n_1 + 1\right) b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} \le C \vee \right.\right.$$

$$\left.\left. \vee \left(\left(n_1 + 1\right) b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C, b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} \ge b_1^{\min} \right) \right) \right\} \cup$$

$$\cup \left\{\mathbf{n} \in \mathbf{X} : n_1 = N_1 \vee \left(\left(n_1 + 1\right) b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C, \right.\right.$$

$$\left.\left. b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} < b_1^{\min} \right) \right\};$$

$$\mathbf{S}_1^{pre} \cup \mathbf{B}_1^{pre} = \left\{\mathbf{n} \in \mathbf{X} : n_1 < N_1, \left(\left(n_1 + 1\right) b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} \le C \vee \right.\right.$$

$$\left.\left. \vee \left(\left(n_1 + 1\right) b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C, b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} \ge b_1^{\min} \right) \right) \vee \right.$$

$$\left. \vee n_1 = N_1 \vee \left(\left(n_1 + 1\right) b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C, b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} < b_1^{\min} \right) \right\}.$$

Пусть $a : n_1 < N_1$, $b : \left(n_1 + 1\right) b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} \le C$,

$c : b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} \ge b_1^{\min}$, тогда $\bar{a} : n_1 = N_1$,

$\bar{b} : \left(n_1 + 1\right) b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C$, $\bar{c} : b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} < b_1^{\min}$. С учетом обозначений объединение множеств $\mathbf{S}_1^{pre}$ и $\mathbf{B}_1^{pre}$ можно записать как $\left\{\mathbf{n} \in \mathbf{X} : a \wedge \left(b \vee \bar{b} \wedge c\right) \vee \bar{a} \vee \bar{b} \wedge \bar{c}\right\}$. По закону поглощения

$$a \wedge \left(b \vee \bar{b} \wedge c\right) \vee \bar{a} \vee \bar{b} \wedge \bar{c} = a \wedge b \vee a \wedge \bar{b} \wedge c \vee \bar{a} \vee \bar{b} \wedge \bar{c} =$$

$$= \left(\bar{a} \vee a \wedge b\right) \vee a \wedge \bar{b} \wedge c \vee \bar{b} \wedge \bar{c} = \bar{a} \vee b \vee a \wedge \bar{b} \wedge c \vee \bar{b} \wedge \bar{c} =$$

$$= \bar{a} \vee \left(b \vee \bar{b} \wedge \bar{c}\right) \vee a \wedge \bar{b} \wedge c = \bar{a} \vee b \vee \bar{c} \vee a \wedge \bar{b} \wedge c = b \vee \bar{c} \vee \left(\bar{a} \vee a \wedge \bar{b} \wedge c\right) =$$

$$= b \vee \bar{c} \vee \bar{a} \vee \bar{b} \wedge c = \bar{a} \vee \bar{c} \vee \left(b \vee \bar{b} \wedge c\right) = \bar{a} \vee \bar{c} \vee b \vee c = 1$$

Таким образом, $\mathbf{S}_1^{pre} \cup \mathbf{B}_1^{pre} = \mathbf{X}$.

Аналогично,

$$\mathbf{S}_2^{pre} \cup \mathbf{B}_2^{pre} = \left\{\mathbf{n} \in \mathbf{X} : n_2 < N_2, \max\left\{n_1 b_1^{\min}, g_1\right\} + \left(n_2 + 1\right) b_2^{\min} \le C\right\} \cup$$

$$\cup \left\{\mathbf{n} \in \mathbf{X} : n_2 = N_2 \vee \max\left\{n_1 b_1^{\min}, g_1\right\} + \left(n_2 + 1\right) b_2^{\min} > C\right\};$$

$$\mathbf{S}_2^{pre} \cup \mathbf{B}_2^{pre} = \left\{\mathbf{n} \in \mathbf{X} : n_2 < N_2, \max\left\{n_1 b_1^{\min}, g_1\right\} + \left(n_2 + 1\right) b_2^{\min} \le C \vee \right.$$

$$\left. \vee n_2 = N_2 \vee \max\left\{n_1 b_1^{\min}, g_1\right\} + \left(n_2 + 1\right) b_2^{\min} > C\right\}.$$

Пусть $a : n_2 < N_2$, $b : \max\left\{n_1 b_1^{\min}, g_1\right\} + \left(n_2 + 1\right) b_2^{\min} \le C$, тогда $\bar{a} : n_2 = N_2$, $\bar{b} : \max\left\{n_1 b_1^{\min}, g_1\right\} + \left(n_2 + 1\right) b_2^{\min} > C$. С учетом обозначений объединение множеств $\mathbf{S}_1^{pre}$ и $\mathbf{B}_1^{pre}$ можно записать как $\left\{\mathbf{n} \in \mathbf{X} : a \wedge b \vee \bar{a} \vee \bar{b}\right\}$. По закону поглощения

$$a \wedge b \vee \bar{a} \vee \bar{b} = \left(a \wedge b \vee \bar{a}\right) \vee \bar{b} = \left(\bar{a} \vee b\right) \vee \bar{b} = \bar{a} \vee \left(b \vee \bar{b}\right) = 1.$$

Таким образом, $\mathbf{S}_2^{pre} \cup \mathbf{B}_2^{pre} = \mathbf{X}$.

Утверждение доказано. □

С учетом выведенных выше множеств сформулируем правила приема и обслуживания запросов на предоставление услуги k-го типа:

- если число запросов на предоставление услуги k-го типа, которые обслуживаются в системе, меньше максимально возможного числа таких запросов N_k, и число свободных каналов, доступных для запросов этого типа, составляет не менее $b_k^{\min}$, то запрос на предоставление услуги данного типа принимается на обслуживание, $k = 1, 2$;
- если число запросов на предоставление услуги первого типа, которые обслуживаются в системе, меньше максимально возможного числа таких запросов N_1, число свободных каналов, доступных для запросов этого типа, меньше $b_1^{\min}$, а число каналов общего пула, занятых обслуживанием запросов второго типа, составляет не менее $b_2^{\min}$, то запрос первого типа принимается на обслуживание за счет прерывания обслуживания запросов второго типа из общего пула;
- в противном случае запрос на предоставление услуги k-го типа блокируется, $k = 1, 2$.

Составим диаграмму интенсивностей переходов в общем виде (рис. 3.12) и для центрального состояния (рис. 3.13).

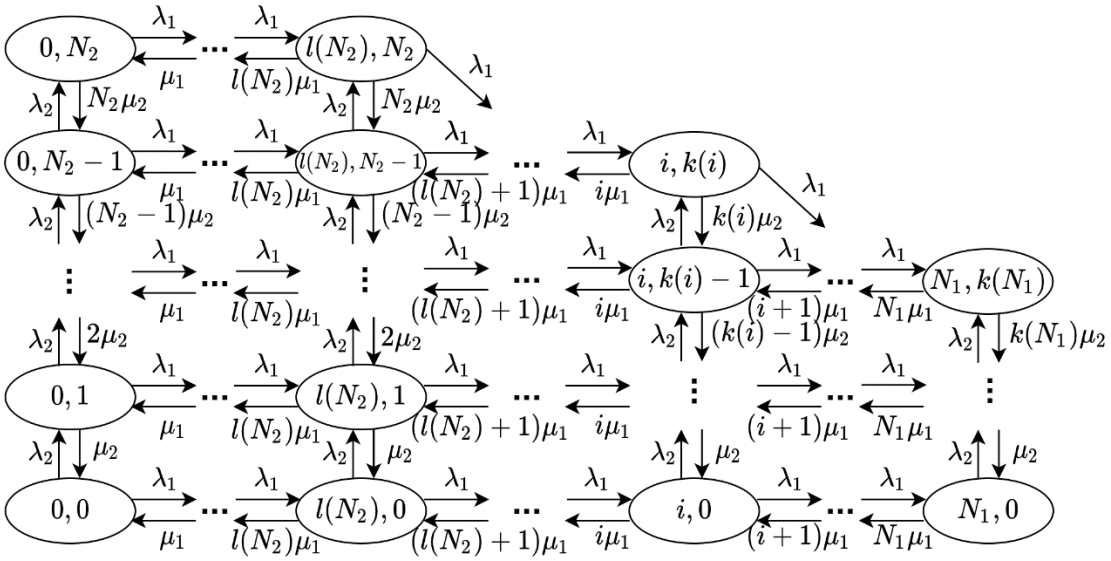

Рис. 3.12. Диаграмма интенсивностей переходов модели с разделением ресурсов и прерыванием обслуживания неприоритетного трафика.

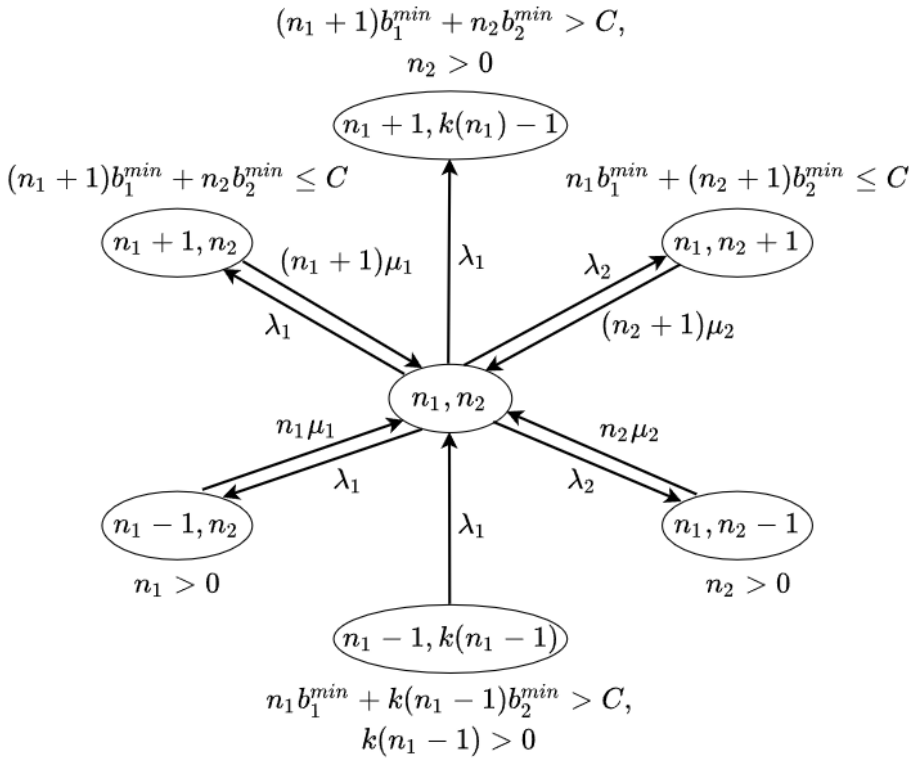

Рис. 3.13. Диаграмма интенсивностей переходов для центрального состояния модели с разделением ресурсов и прерыванием обслуживания неприоритетного трафика.

Согласно диаграмме интенсивностей переходов (рис. 3.13), рассматриваемый СП $X(t)=\{(N_1(t),N_2(t)),t\ge 0\}$ описывается следующей СУГБ:

$$\begin{aligned}
&\big(\lambda_1\cdot I\{n_1<N_1,(n_1+1)b_1^{\min}+n_2b_2^{\min}\le C\}+\lambda_1\cdot I\{n_1<N_1,n_2>0,\\
&(n_1+1)b_1^{\min}+n_2b_2^{\min}>C\}+\lambda_2\cdot I\{n_2<N_2,n_1b_1^{\min}+(n_2+1)b_2^{\min}\le C\}+\\
&+n_1\mu_1\cdot I\{n_1>0\}+n_2\mu_2\cdot I\{n_2>0\}\big)\cdot p(n_1,n_2)=\lambda_1\cdot I\{n_1>0\}\cdot\\
&\cdot p(n_1-1,n_2)+\lambda_1\cdot I\{n_1b_1^{\min}+k(n_1-1)b_2^{\min}>C,k(n_1-1)>0\}\cdot\\
&\cdot p(n_1-1,k(n_1-1))+\lambda_2\cdot I\{n_2>0\}\cdot p(n_1,n_2-1)+(n_1+1)\mu_1\cdot\\
&\cdot I\{n_1<N,(n_1+1)b_1^{\min}+n_2b_2^{\min}\le C\}\cdot p(n_1+1,n_2)+(n_2+1)\mu_2\cdot\\
&\cdot I\{n_2<N_2,n_1b_1^{\min}+(n_2+1)b_2^{\min}\le C\}\cdot p(n_1,n_2+1),
\end{aligned}\qquad (3.37)$$

где $p(n_1,n_2),(n_1,n_2)\in\mathbf{X}$ – стационарное распределение вероятностей состояний системы.

Утверждение 3.14. Стационарное распределение вероятностей $\mathbf{p}=\{p(n_1,n_2),(n_1,n_2)\in\mathbf{X}\}$ можно получить путем численного решения системы уравнений равновесия $\mathbf{p}^T\mathbf{A}=\mathbf{0}^T$, $\mathbf{p}^T\mathbf{1}=1$. Элементы $a((n_1,n_2),(n_1',n_2'))$ инфинитезимальной матрицы $\mathbf{A}$ определяются следующим образом:

$$a(\mathbf{n},\mathbf{n}')=\begin{cases}
\lambda_1, & \text{если } n_1'=n_1+1,n_2'=n_2,n_1<N_1,(n_1+1)b_1^{\min}+n_2b_2^{\min}\le C,\\
& \text{или } n_1'=n_1+1,n_2'=n_2-1,n_1<N_1,n_2>0,\\
& (n_1+1)b_1^{\min}+n_2b_2^{\min}>C;\\
\lambda_2, & \text{если } n_1'=n_1,n_2'=n_2+1,n_2<N_2,n_1b_1^{\min}+(n_2+1)b_2^{\min}\le C;\\
n_1\mu_1, & \text{если } n_1'=n_1-1,n_2'=n_2,n_1>0;\\
n_2\mu_2, & \text{если } n_1'=n_1,n_2'=n_2-1,n_2>0;\\
\varphi, & \text{если } n_1'=n_1,n_2'=n_2;\\
0 & \text{в ином случае,}
\end{cases}\qquad (3.38)$$

$$\varphi = -\Big[\lambda_1 \cdot I\left\{n_1 < N_1, n_2 > 0, (n_1+1)b_1^{\min} + n_2 b_2^{\min} > C\right\} +$$
$$+\lambda_1 \cdot I\left\{n_1 < N_1, (n_1+1)b_1^{\min} + n_2 b_2^{\min} \le C\right\} + \qquad (3.39)$$
$$+\lambda_2 \cdot I\left\{n_2 < N_2, n_1 b_1^{\min} + (n_2+1)b_2^{\min} \le C\right\} + n_1\mu_1 + n_2\mu_2\Big].$$

Зная распределение вероятностей $\mathbf{p}$, можно вычислить такие вероятностные характеристики системы, как вероятность блокировки запросов, вероятность прерывания обслуживания запросов, среднее число каналов, занятых запросами, а также коэффициент использования ресурсов.

Утверждение 3.15. Вероятность блокировки запросов первого типа рассчитывается по формуле

$$p_{B_1} = \sum_{i=0}^{k(N_1)} p(N_1, i). \qquad (3.40)$$

Доказательство. Используя множество блокировки запросов (3.34), определим состояния системы, в которых поступающие запросы первого типа будут заблокированы:

$$\mathbf{n} \in \mathbf{X} : n_1 = N_1 \vee \Big((n_1+1)b_1^{\min} + \max\left\{n_2 b_2^{\min}, g_2\right\} > C,$$
$$b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} < b_1^{\min}\Big);$$

$$\mathbf{n} \in \mathbf{X} : n_1 = N_1 \vee \Big((n_1+1)b_1^{\min} > C - \max\left\{n_2 b_2^{\min}, g_2\right\},$$
$$b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} < b_1^{\min}\Big);$$

$$\mathbf{n} \in \mathbf{X} : n_1 = N_1 \vee \Big(n_1 + 1 > \Big(C - \max\left\{n_2 b_2^{\min}, g_2\right\}\Big) \Big/ b_1^{\min},$$
$$b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} < b_1^{\min}\Big).$$

В случае предположения, что $b_1^{\min} = b_1^{\max}$, $b_1^{\min} \le b_2^{\min}$, неравенство $b_2^{\min}\left(n_2 - N_2^g\right) \cdot I\left\{n_2 > N_2^g\right\} < b_1^{\min}$ не может быть выполнено. Следовательно, $\mathbf{n} \in \mathbf{X} : n_1 = N_1$. Таким образом, просуммировав вероятности $p(N_1, n_2)$ по n_2, получим вероятность блокировки (3.40).

Утверждение доказано. □

Утверждение 3.16. Вероятность блокировки запросов второго типа рассчитывается по формуле

$$p_{B_2} = \sum_{i=0}^{N_1} p\big(i,k(i)\big). \tag{3.41}$$

Утверждение 3.17. Вероятность прерывания обслуживания запросов второго типа рассчитывается по формуле

$$p_{pre} = \sum_{i=0}^{N_1-1} \frac{\lambda_1}{\lambda_1 + \lambda_2 + i \cdot \mu_1 + k(i) \cdot \mu_2} p\big(i,k(i)\big) \cdot I\big(k(i) \neq k(i+1)\big). \tag{3.42}$$

Утверждение 3.18. Среднее число каналов, занятых запросами второго типа, рассчитывается по формуле

$$\bar{k}_2 = \sum_{i=0}^{N_1} \sum_{j=1}^{k(i)} j \cdot b_2(i,j) \cdot p(i,j). \tag{3.43}$$

Утверждение 3.19. Коэффициент использования ресурсов вычисляется по формуле

$$U = \sum_{i=0}^{N_1} \sum_{j=0}^{k(i)} \Big(i \cdot b_1^{\min} + j \cdot b_2(i,j)\Big) \cdot p(i,j). \tag{3.44}$$

3.4. Сравнительный анализ стратегий распределения ресурсов

В качестве примера рассмотрим процесс обслуживания базовой станцией емкости $C = 100$ каналов запросов пользователей на предоставление двух типов услуг: услуг, генерирующих потоковый трафик (URLLC), и услуг, генерирующих эластичный трафик (eMBB).

На основе предложенной модели СМО выполним сравнение пяти стратегий разделения ресурсов [55, 56]:

- отсутствие резервирования и приоритетов (англ. dynamic, DYN);
- приоритетное обслуживание запросов 1-го типа с прерыванием, но без резервирования (англ. dynamic with preemption, DYN+PRE);
- полное резервирование ресурсов, $g_1 + g_2 = C$ (англ. reservation, RES);
- частичное резервирование, $g_1 + g_2 < C$, без прерываний (англ. dynamic with reservation, DYN+RES);

- частичное резервирование с прерываниями (англ. dynamic with preemption and reservation, DYN+RES+PRE).

Чтобы гарантировать производительность URLLC и eMBB трафика с точки зрения вероятности блокировки, для схем RES, DYN+RES и DYN+RES+PRE необходимо выполнить численную оптимизацию параметров g_1 и g_2. Проведем анализ минимального числа каналов, необходимого для удовлетворения требований к качеству обслуживания трафика URLLC и eMBB. В качестве ограничений установим вероятности блокировки URLLC и eMBB равными 10^{-5} и 10^{-1} соответственно [57]. Параметры системы представлены в таблице 3.2.

Таблица 3.2. Параметры системы.

Параметр	**Значение**
Интенсивность поступления запросов на передачу URLLC трафика	5000 запросов/с
Интенсивность поступления запросов на передачу eMBB трафика	1 запрос/с
Среднее время обслуживания запросов на передачу URLLC трафика	1 мс
Среднее время обслуживания запросов на передачу eMBB трафика	10 с

В таблице 3.3 приведены параметры числа зарезервированных БЦК g_1 и g_2 для трех схем, а также выбранные значения коэффициентов σ_1 и σ_2. Согласно полученным данным для схемы DYN+RES+PRE, конкурирующей по производительности со схемой DYN+PRE, нет необходимости резервировать ресурсы для трафика URLLC, поскольку значения g_1 никогда не превышают 1 БЦК.

Таблица 3.3. Минимальное число каналов для рассматриваемых схем.

Коэффициент	**Схема**	g_1	g_2	C
$\sigma_1 = 0,4$	RES	23	26	49
	DYN+RES	20	1	47
	DYN+RES+PRE	1	1	24
$\sigma_1 = 0,8$	RES	36	26	62
	DYN+RES	32	1	59
	DYN+RES+PRE	1	1	37
$\sigma_1 = 1,2$	RES	48	26	74
	DYN+RES	43	1	70
	DYN+RES+PRE	1	1	49
$\sigma_1 = 1,6$	RES	59	26	85
	DYN+RES	54	1	81
	DYN+RES+PRE	1	1	60
$\sigma_1 = 2,0$	RES	69	26	95
	DYN+RES	65	1	92
	DYN+RES+PRE	1	1	70
$\sigma_2 = 0,4$	RES	42	14	56
	DYN+RES	38	1	53
	DYN+RES+PRE	1	1	43
$\sigma_2 = 0,8$	RES	42	22	64
	DYN+RES	38	1	61
	DYN+RES+PRE	1	1	43
$\sigma_2 = 1,2$	RES	42	30	72
	DYN+RES	38	1	69
	DYN+RES+PRE	1	1	43
$\sigma_2 = 1,6$	RES	42	38	80
	DYN+RES	38	1	77
	DYN+RES+PRE	1	1	43

Коэффициент	**Схема**	g_1	g_2	C
$\sigma_2 = 2,0$	RES	42	46	88
	DYN+RES	38	1	85
	DYN+RES+PRE	1	1	43

На рис. 3.14 минимальные требования к числу каналов представлены в виде функции аргумента $\sigma_2 = \rho_2/\rho_1$, где $\rho_1 = \lambda_1 b_1^{\min}/\mu_1$ и $\rho_2 = \lambda_2 b_2^{\min}/\mu_2$ – интенсивности предложенной нагрузки, создаваемой запросами на передачу URLLC и eMBB трафика соответственно, при этом интенсивность поступления запросов λ_2 на передачу eMBB трафика увеличивается, в то время как интенсивность поступления запросов λ_1 на передачу URLLC трафика остается постоянной. Согласно полученным результатам больше всего каналов требует схема DYN, где для обоих типов трафика не предусмотрены специальные дисциплины обслуживания. Второй по числу используемых каналов является схема полного резервирования RES. Прерывание без резервирования, DYN+PRE, является одной из двух лучших схем с точки зрения занятых каналов и сравнима с наиболее сложной схемой, сочетающей в себе как резервирование, так и прерывание, DYN+RES+PRE. Напомним, что схема DYN+PRE обеспечивает абсолютный приоритет в обслуживании URLLC трафика, оставляя при этом eMBB трафик уязвимым для прерывания.

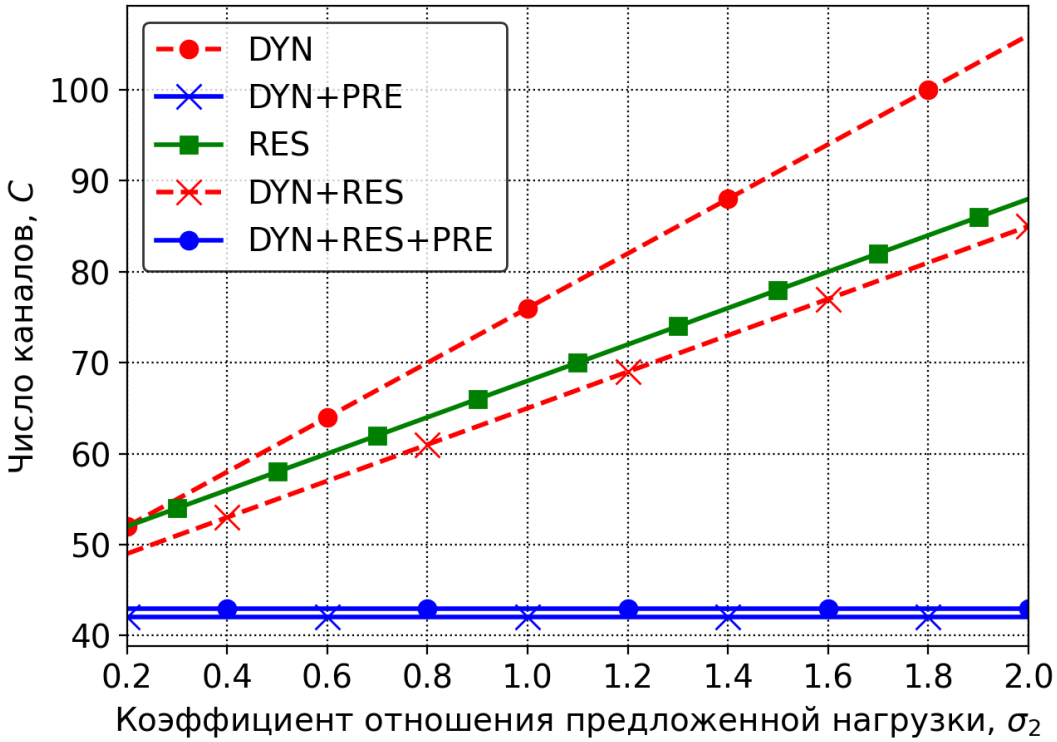

Рис. 3.14. Зависимость минимального числа каналов C от коэффициента отношения предложенной нагрузки σ_2 при увеличении интенсивности поступления запросов на передачу eMBB трафика.

Полученные на рис. 3.14 результаты дополнительно подтверждаются рис. 3.15, где варьируется интенсивность поступления запросов λ_1 на передачу URLLC трафика, влияющая на аргумент функции $\sigma_1 = \rho_1/\rho_2$. Обе линии, соответствующие DYN+RES+PRE и DYN+PRE, остаются близкими друг к другу и лежат значительно ниже по сравнению с другими рассмотренными схемами. Численно преимущество при изменении λ_2 остается постоянным и достигает 50%, а при изменении λ_1 сильно зависит от относительных нагрузок трафика и уменьшается при увеличении λ_1.

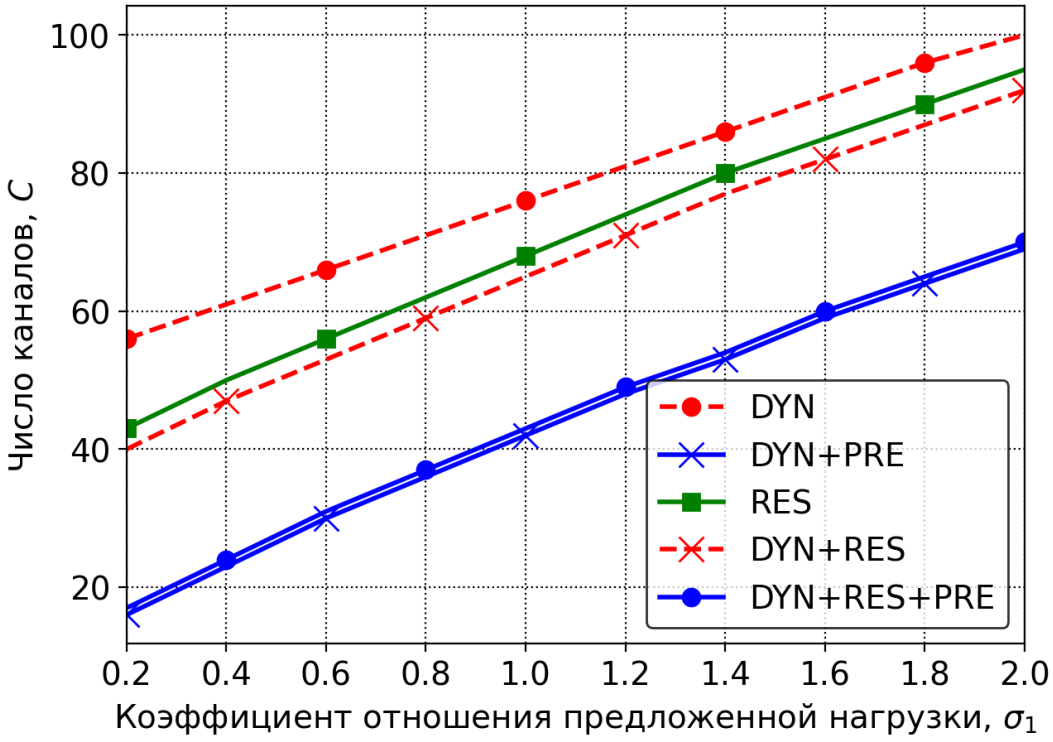

Рис. 3.15. Зависимость минимального числа каналов C от коэффициента отношения предложенной нагрузки σ_1 при увеличении интенсивности поступления запросов на передачу URLLC трафика.

Рассмотрим поведение модели при следующих параметрах: $b_1 = 1$, $b_2^{\min} = 2$, $b_2^{\max} = 4$, $\lambda_2 = 1$, $\mu_1 = 10^3$, $\mu_2 = 0{,}1$. Предположим, что для стратегии полного резервирования ресурсов $g_1 = g_2 = 25$, в то время как для стратегий частичного резервирования $g_1 = g_2 = 10$. Проанализировав данные, представленные на рис. 3.16, можно сделать вывод о том, что худшую производительность с точки зрения вероятности блокировки запросов URLLC во всем диапазоне значений λ_1 имеют схемы DYN и DYN+RES, что объясняется отсутствием механизма прерывания обслуживания менее приоритетных запросов и значительной нагрузкой, создаваемой eMBB трафиком. Более высокую производительность показывают схемы с прерыванием обслуживания, особенно DYN+PRE и DYN+RES+PRE, в рамках которых выполняется требование к вероятности блокировки запросов URLLC (менее 10^{-5}).

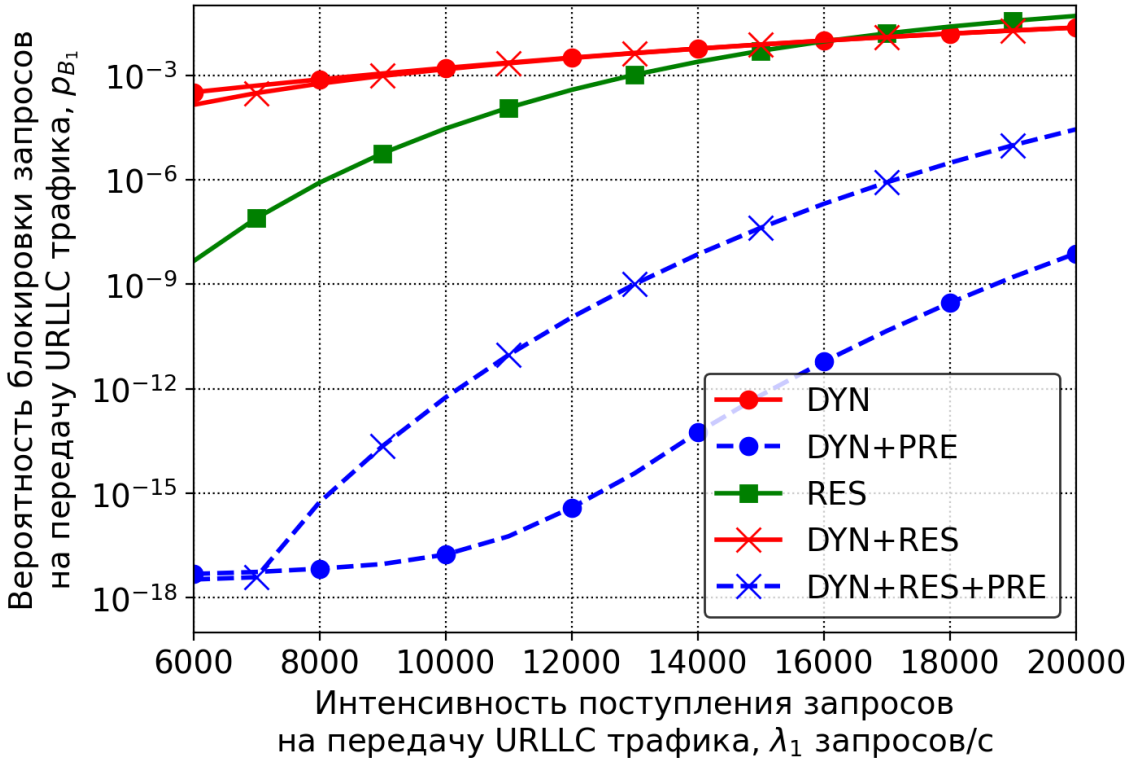

Рис. 3.16. Зависимость вероятности блокировки запросов на передачу URLLC трафика от интенсивности поступления запросов на передачу URLLC трафика.

Противоположное поведение наблюдается на рис. 3.17, где стратегия полного резервирования характеризуется наибольшей вероятностью блокировки. В то же время показатели для схем DYN и DYN+RES удовлетворяют требованиям для eMBB трафика (вероятность блокировки менее 10^{-1}) во всем диапазоне значений λ_1. Стоит отметить, что использование механизмов прерывания обслуживания крайне важно для одновременного предоставления услуг URLLC и eMBB, поскольку запросы на передачу eMBB трафика характеризуются более длительным временем обслуживания, что приводит к блокировке запросов на передачу URLLC трафика. Этот вывод подтверждается поведением вероятности прерывания обслуживания запросов на передачу eMBB трафика для схем DYN+PRE и DYN+RES+PRE, использующих данный механизм, на рис. 3.18. Можно заметить, что дополнительные потери eMBB трафика с точки зрения

вероятности прерывания незначительны, так как соответствующие значения не превышают 10^{-5} для всего рассматриваемого диапазона λ_1.

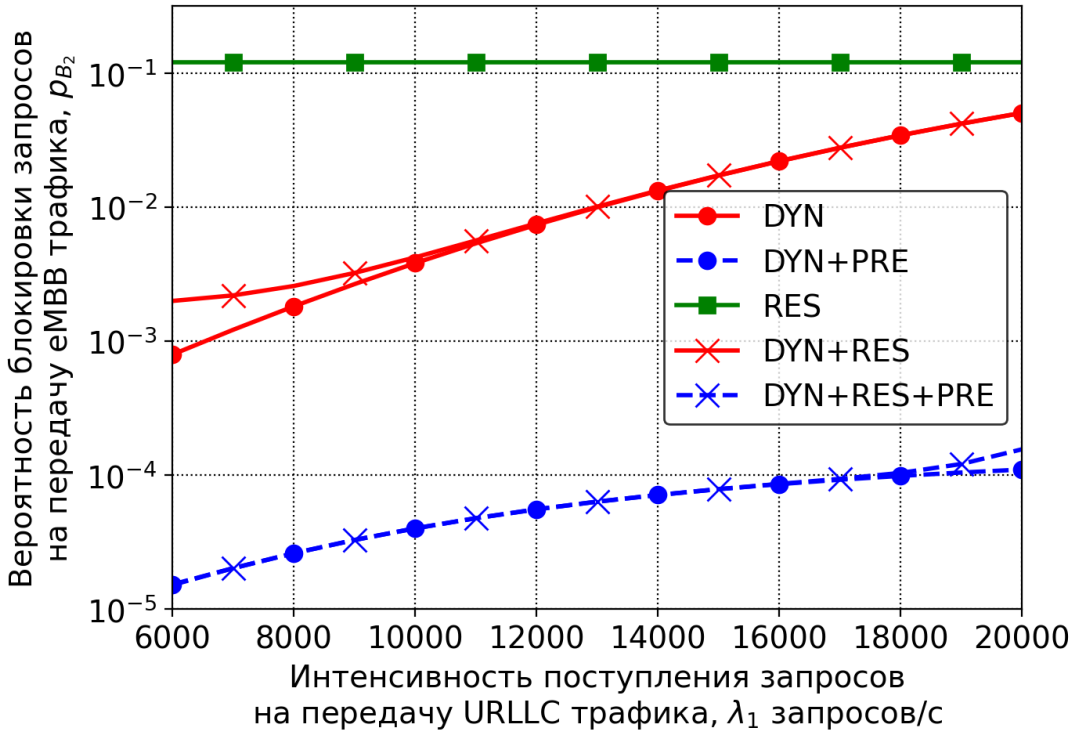

Рис. 3.17. Зависимость вероятности блокировки запросов на передачу eMBB трафика от интенсивности поступления запросов на передачу URLLC трафика.

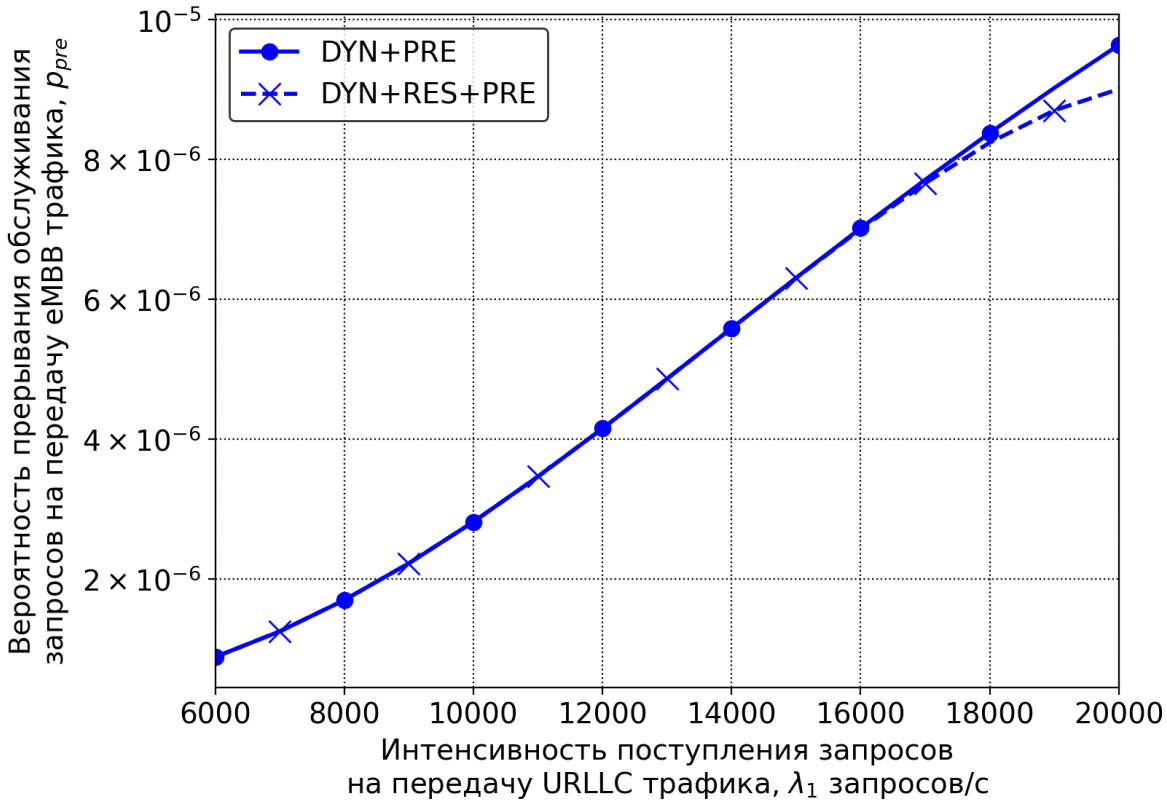

Рис. 3.18. Зависимость вероятности прерывания обслуживания запросов на передачу eMBB трафика от интенсивности поступления запросов на передачу URLLC трафика.

На рис. 3.19 показано использование ресурсов системы для всех рассмотренных схем. Отметим, что наименьший коэффициент использования ресурсов характерен для системы с полным разделением ресурсов. Это объясняется тем, что параметры g_1 и g_2 необходимо адаптировать к изменяющимся условиям трафика, чтобы данная стратегия была эффективной. В то же время наибольший коэффициент наблюдается для полностью динамической схемы, которая не использует прерывание обслуживания, DYN. При более высоких значениях λ_1 использование ресурсов также очень велико для схемы DYN+RES с частичным резервированием и без прерывания. Однако данные схемы имеют худшую производительность для URLLC трафика, как показано на рис. 3.16.

Стратегии, которые способны обеспечить наилучшую производительность одновременно для двух типов трафика, DYN+PRE и DYN+RES+PRE, характеризуются нелинейным использованием ресурсов,

которое увеличивается с ростом λ_1 до некоторого значения, а затем начинает снижаться. Это объясняется тем, что запросы на передачу URLLC трафика, характеризующиеся коротким временем обслуживания, начинают вытеснять запросы на передачу eMBB трафика, характеризующиеся длительным временем обслуживания, что приводит к снижению нагрузки системы. Однако для схемы DYN+RES+PRE могут быть настроены параметры g_1 и g_2 для улучшения показателя использования ресурсов.

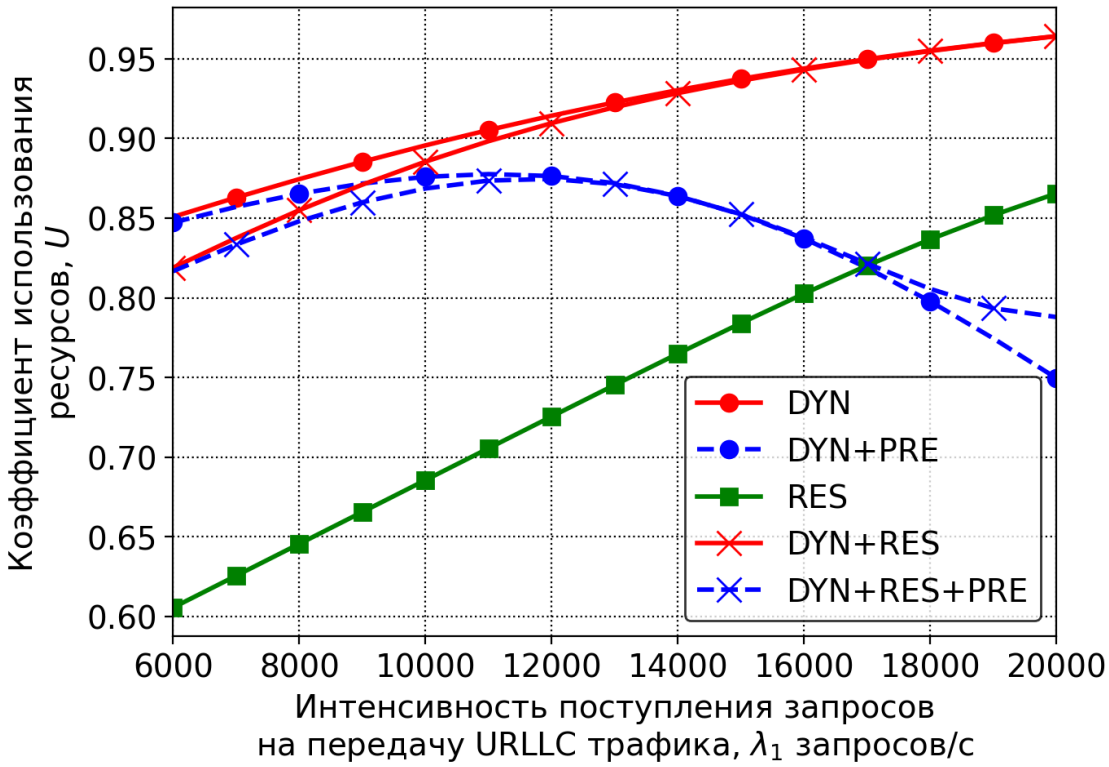

Рис. 3.19. Зависимость коэффициента использования ресурсов от интенсивности поступления запросов на передачу URLLC трафика.

Далее рассмотрим зависимость показателей эффективности модели от параметров радиоканала. Наиболее интересным является влияние антенной решетки, поскольку оно распространяется не только на зону покрытия БС, но и на качество канала и, следовательно, на количество ресурсов, необходимых для достижения определенной скорости передачи данных. На рис. 3.20 и 3.21 показаны вероятности блокировки запросов на передачу URLLC и eMBB трафика, а на рис. 3.22 – использование системных ресурсов для $C=70$, $c_1^{\min}=5$, $c_2^{\min}=10$, $c_2^{\max}=20$, $\lambda_1=5000$, $\lambda_2=1$, $\mu_1=10^3$, $\mu_2=0{,}1$. Для схемы

RES $g_1 = g_2 = 35$, а для стратегий с использованием частичного резервирования, DYN+RES и DYN+RES+PRE, полагаем $g_1 = g_2 = 15$.

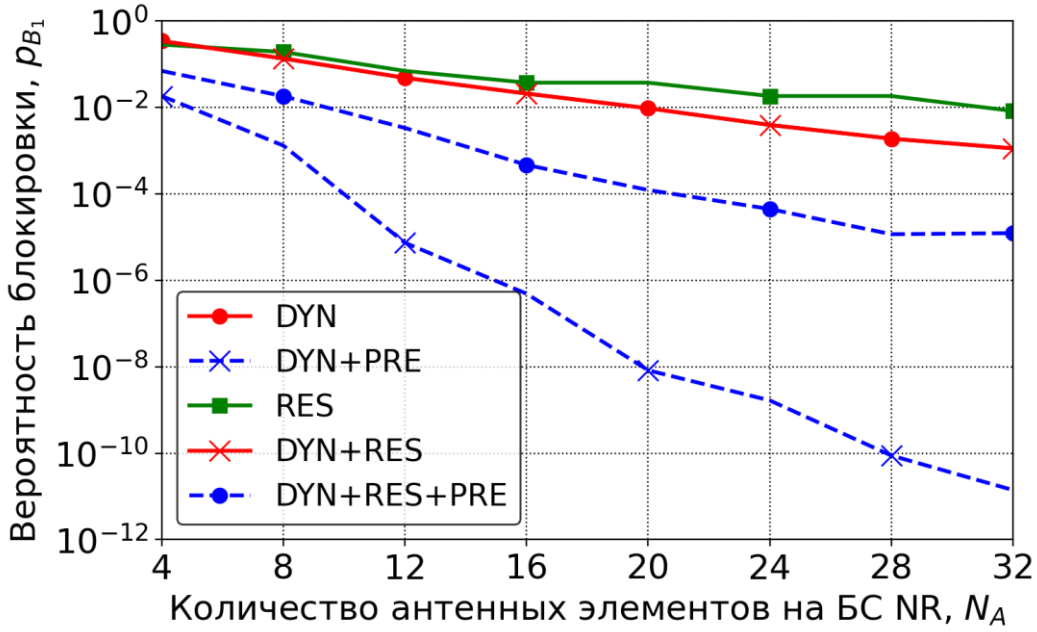

Рис. 3.20. Зависимость вероятности блокировки запросов на передачу URLLC трафика от количества антенных элементов.

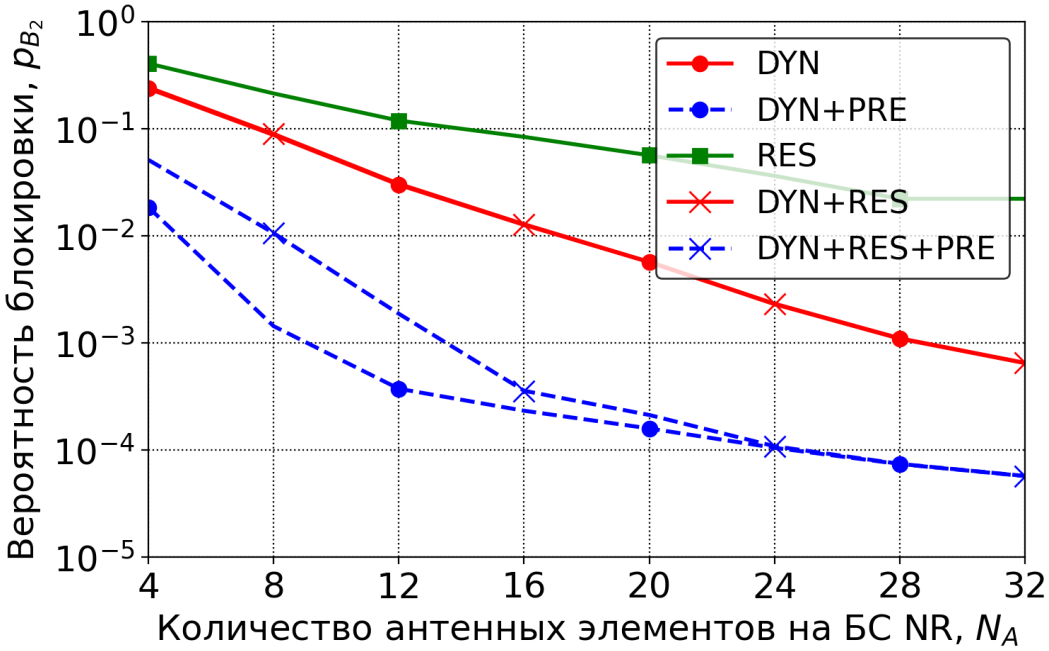

Рис. 3.21. Зависимость вероятности блокировки запросов на передачу eMBB трафика от количества антенных элементов.

Анализируя представленные данные, можно заметить, что увеличение количества антенных элементов положительно влияет на процесс обслуживания обоих видов трафика для всех рассмотренных схем, при этом

использование системных ресурсов, показанное на рис. 3.22, уменьшается с увеличением размера массива.

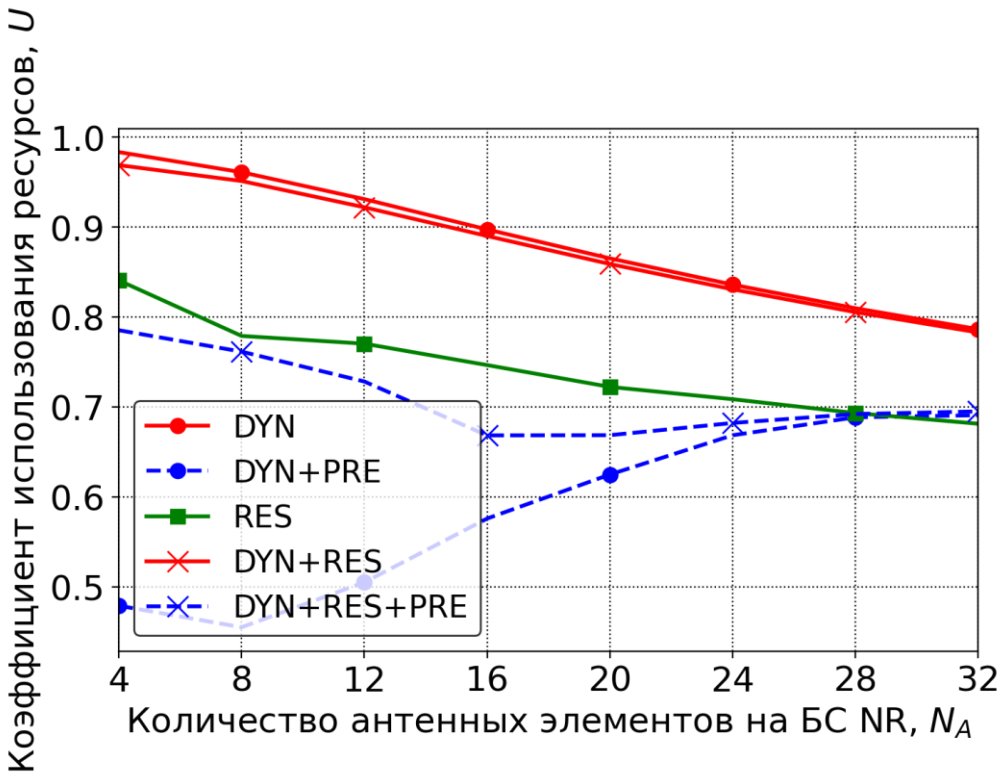

Рис. 3.22. Зависимость коэффициента использования ресурсов от количества антенных элементов.

Таким образом, численные результаты показывают, что использование механизма прерывания обслуживания имеет важное значение для одновременной поддержки URLLC и eMBB трафика, поскольку последний характеризуется более высокими требованиями к передаче и более длительным временем обслуживания, блокируя передачу URLLC трафика. Кроме того, прерывание обслуживания имеет решающее значение для минимизации количества ресурсов, необходимых для обеспечения гарантированной производительности как для eMBB, так и для URLLC трафика. При незначительной разнице с точки зрения нагрузки, создаваемой каждым из двух типов трафика, резервирование ресурсов не является необходимым. Наиболее оптимальная схема с приоритетами и отсутствием резервирования ресурсов позволяет достичь преимущества в количестве минимальных требуемых для передачи ресурсов по сравнению с системами

без приоритетов или с полным резервированием. Изменение параметров радиосвязи, включая увеличение количества антенных элементов, имеет положительное влияние на вероятность блокировки, позволяя снизить ее в несколько раз, и использование системных ресурсов.

Описанная стратегия одновременной поддержки двух типов трафика с реализацией механизма прерывания рекомендуется для использования в радиоинтерфейсе 5G NR, обслуживающем типы трафика с принципиально различными требованиями. Эта схема также может использоваться в сочетании с другими механизмами для улучшения показателей трафика, критичного к задержке, а именно вероятности блокировки и коэффициента использования системных ресурсов.

Заключение

В заключение сформулируем основные результаты и выводы диссертационной работы.

1. Исследованы особенности моделей беспроводных сетей при промышленном развертывании. Предложена модель схемы доступа к ресурсам мобильной сети со снижением скорости передачи эластичного трафика. Получено численное решение СУР. Проведен анализ вероятности блокировки, вероятности прерывания обслуживания запросов, коэффициента использования ресурсов и доли времени, в течение которого запросы обслуживаются со сниженной скоростью.
2. Разработана модель одновременного предоставления услуг с реализацией явного приоритета, учитывающая особенности совместного обслуживания разных типов трафика в условиях промышленного развертывания мобильных сетей. Проведен сравнительный анализ показателей эффективности модели с потоковым, критичным к задержкам, и эластичным трафиком для трех стратегий передачи данных.
3. Разработана модель совместного обслуживания K-типов трафика с резервированием индивидуальных зон и приоритетами. Разработан алгоритм выбора запросов, обслуживание которых должно быть прервано. Получены элементы инфинитезимальных матриц. Предложено численное решение СУР для расчета основных показателей эффективности модели.
4. Разработаны модели совместного обслуживания потокового и эластичного трафика с резервированием индивидуальных зон, с прерыванием или без прерывания передачи эластичного трафика. Получены элементы инфинитезимальных матриц. Предложено аналитическое решение СУР. Получены формулы для расчета вероятности блокировки, прерывания запросов, коэффициента

использования ресурсов. Выполнена численная оптимизация параметров для обеспечения гарантий производительности трафика в моделях совместного обслуживания. Проведен сравнительный анализ стратегий распределения ресурсов: I) отсутствие резервирования и приоритетов; II) приоритетное обслуживание с прерыванием, но без резервирования; III) полное резервирование ресурсов; IV) частичное резервирование без прерывания; V) частичное резервирование с прерыванием.

СПИСОК ОСНОВНЫХ СОКРАЩЕНИЙ

БС	–	Базовая станция
БЦК	–	Базовый цифровой канал
СМО	–	Система массового обслуживания
СП	–	Случайный процесс
СУГБ	–	Система уравнений глобального баланса
СУР	–	Система уравнений равновесия
ТМО	–	Теория массового обслуживания
3GPP	–	Third Generation Partnership Project
5G	–	Fifth Generation
AR	–	Augmented reality
BS	–	Base stations
D2D	–	Device-to-device
eMBB	–	Enhanced Mobile Broadband
IIoT	–	Industrial Internet of Things
IMT	–	International Mobile Telecommunication
IoT	–	Internet of Things
LTE	–	Long-Term Evolution
mMTC	–	Massive Machine-Type Communications
mmWave	–	Millimeter wave
MVNO	–	Mobile Virtual Network Operators
NR	–	New Radio
PRB	–	Primary resource blocks
QoS	–	Quality of Service
RAN	–	Radio Access Network
SNR	–	Signal-to-noise ratio

UE – User equipment

URLLC – Ultra-Reliable Low Latency Communication

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

h_{BS} – Высота расположения базовых станций

h_{UE} – Высота расположения пользовательских устройств

χ – Плотность размещения базовых станций

l – Шаг сетки размещения пользовательских устройств

W – Ширина полосы пропускания базовой станции

C – Количество каналов

λ_k – Интенсивность поступления запросов k-го типа

μ_k^{-1} – Среднее время обслуживания запроса k-го типа

c_k – Скорость обслуживания запроса k-го типа

$c_k^{\min}$ – Минимальная скорость обслуживания запроса k-го типа

$b_{k,B}$ – Число БЦК, требуемых для обслуживания запроса k-го типа при передаче через базовую станцию

$b_{k,D}$ – Число БЦК, требуемых для обслуживания запроса k-го типа при передаче в режиме D2D

$b_k^{\min}$ – Минимальное число БЦК, требуемых для обслуживания запроса k-го типа

$E[S_{e,B}]$ – Средняя спектральная эффективность при передаче трафика через базовую станцию

$E[S_{e,D}]$ – Средняя спектральная эффективность при D2D-передаче трафика

D – Расстояние между пользовательскими устройствами

B – Расстояние между пользовательским устройством и базовой станцией

r – Радиус зоны покрытия базовой станции

r_S – Максимальный радиус зоны покрытия базовой станции

r_V – Расстояние между базовыми станциями

P_U – Мощность приемника

G_A – Коэффициент усиления антенны на базовой станции

G_U – Коэффициент усиления антенны на пользовательском устройстве

N_0 – Спектральная плотность мощности шума

M_I – Мощность помех

ζ – Коэффициент распространения сигнала

$p_{B,1}(x)$ – Вероятность перекрытия пути прямой видимости длиной x устройством

κ – Прозрачность пользовательского устройства

w – Ширина пользовательского устройства

N_R – Число устройств в зоне покрытия базовой станции

ν – Вероятность расположения устройства в точке сетки

$N_k(t)$ – Случайное число обслуживаемых запросов k-го типа в момент времени t

N_k – Максимальное число обслуживаемых запросов k-го типа

n_k – Число обслуживаемых запросов k-го типа

$\mathbf{X}$ – Пространство состояний системы

$\mathbf{A}$ – Матрица интенсивностей переходов

g_k – Число каналов, зарезервированных для услуг k-го типа

$\mathbf{S}_k$ – Множество приема запросов k-го типа

$\mathbf{B}_k$ – Множество блокировки запросов k-го типа

$\mathbf{П}_k$ – Множество прерывания менее приоритетных запросов при поступлении запросов k-го типа

$\mathbf{S}_k^{\max}$ – Множество приема запросов k-го типа с использованием максимального числа БЦК

p_{B_k} – Вероятность блокировки запроса k-го типа

ρ_k – Интенсивность предложенной нагрузки, создаваемой запросами k-го типа

СПИСОК ЛИТЕРАТУРЫ

1. Cisco Global Networking Trends Report. – Cisco Systems. – 2020. – 95 p.
2. Ericsson White Paper: Cellular Networks for Massive IoT. – Ericsson. – 2020. – 16 p.
3. Gundall M. et al. 5G as enabler for Industrie 4.0 use cases: challenges and concepts // Proc. of 2018 IEEE 23rd international conference on emerging technologies and factory automation (ETFA). – 2018. – Vol. 1. – P. 1401–1408.
4. 3GPP TR 38.913: Study on scenarios and requirements for next generation access technologies: Release 16. – ETSI 3GPP. – 2020.
5. 3GPP TS 38.133: NR; Requirements for support of radio resource management: Release 16. – 3GPP. – 2020.
6. Sachs J., Wallstedt K., Alriksson F., and Eneroth G. Boosting smart manufacturing with 5G wireless connectivity // Ericsson Technology Review. – 2019. – P. 1–12.
7. 3GPP TS 22.104: Service requirements for cyber-physical control applications in vertical domains: Release 16. – 3GPP. – 2019.
8. Firyaguna F., Kibilda J., Galiotto C., and Marchettiet N. Performance analysis of indoor mmWave networks with ceiling-mounted access points // IEEE Transactions on Mobile Computing. – 2020. – Vol. 20, No. 5. – P. 1940–1950.
9. Humadi K., Trigui I., Zhu W.-P., and Ajibet W. Dynamic base station clustering in user-centric mmWave networks: Performance analysis and optimization // IEEE Transactions on Communications. – 2021. – Vol. 69, No. 7. – P. 4847–4861.
10. Begishev V., Moltchanov D., Sopin E., Samuylov A., Andreev S., Koucheryavy Y., and Samouylov K. Quantifying the impact of guard capacity on session continuity in 3GPP New Radio systems // IEEE Transactions on Vehicular Technology. – 2019. – Vol. 68, No. 12. – P. 12345–12359.

11. Mahmood N.H., Lopez M., Laselva D., Pedersen K., and Berardinelli G. Reliability oriented dual connectivity for URLLC services in 5G New Radio // Proc. of 2018 15th International Symposium on Wireless Communication Systems (ISWCS). – 2018.
12. Rao J. and Vrzic S. Packet duplication for URLLC in 5G: Architectural enhancements and performance analysis // IEEE Network. – 2018. – Vol. 32, No. 2. – P. 32–40.
13. Mahmood N.H., Karimi A., Berardinelli G., Pedersen K.I., and Laselva D. On the resource utilization of multi-connectivity transmission for URLLC services in 5G New Radio // Proc. of 2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW). – 2019.
14. Gerasin I., Krasilov A., and Khorov E. Flexible Multiplexing of Grant-Free URLLC and eMBB in Uplink // Proc. of 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. – 2020.
15. Tominaga E.N., Alves H., Souza R.D., Rebelatto J.L., and Latva-aho M. Non-Orthogonal Multiple Access and Network Slicing: Scalable Coexistence of eMBB and URLLC // Proc. of 2021 IEEE 93rd Vehicular Technology Conference (VTC 2021). – 2021.
16. Popovski P., Trillingsgaard K.F., Simeone O., and Durisi G. 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View // IEEE Access. – 2018. – Vol. 6. – P. 55765–55779.
17. Dos Santos E.J., Souza R.D., Rebelatto J.L., and Alves H. Network Slicing for URLLC and eMBB With Max-Matching Diversity Channel Allocation. // IEEE Communications Letters. – 2020. – Vol. 24, No. 3. – P. 658–661.
18. Tebe P.I., Ntiamoah-Sarpong K., Tian W., Li J., Huang Y., and Wen G. Using 5G Network Slicing and Non-Orthogonal Multiple Access to Transmit Medical Data in a Mobile Hospital System // IEEE Access. – 2020. – Vol. 8. – P. 189163–189178.

19. Banchs A., de Veciana G., Sciancalepore V., and Costa-Perez X. Resource allocation for network slicing in mobile networks // IEEE Access. – 2020. – Vol. 8. – P. 214696–214706.
20. Chen Y.-J., Cheng L.-Y., and Wang L.-C. Prioritized resource reservation for reducing random access delay in 5G URLLC // Proc. of 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). – 2017.
21. 3GPP TS 38.211: NR; Physical Channels and Modulation: Release 15. – 3GPP. – 2017.
22. Naddeh N., Jemaa S.B., Eddine Elayoubi S., and Chahed T. Proactive RAN resource reservation for URLLC vehicular slice // Proc. of 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). – 2021.
23. Ji H., Park S., Yeo J., Kim Y., Lee J., and Shim B. Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects // IEEE Wireless Communications. – 2018. – Vol. 25, No. 3. – P. 124–130.
24. Markova E., Moltchanov D., Pirmagomedov R., Ivanova D., Koucheryavy Y., and Samouylov K. Prioritized Service of URLLC Traffic in Industrial Deployments of 5G NR Systems // Lecture Notes in Computer Science. – 2020. – Vol. 12563. – P. 497–509.
25. Markova E., Moltchanov D., Pirmagomedov R., Ivanova D., Koucheryavy Y., and Samouylov K. Priority-based Coexistence of eMBB and URLLC Traffic in Industrial 5G NR Deployments // Proc. of 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). – 2020.
26. Yang W., Li C.-P., Fakoorian A., Hosseini K., and Chen W. Dynamic URLLC and eMBB multiplexing design in 5G New Radio // Proc. of 2020 IEEE 17th Annual Consumer Communications Networking Conference (CCNC). – 2020.

27. Pandey S.R., Alsenwi M., Tun Y.K., and Hong C.S. A downlink resource scheduling strategy for URLLC traffic // Proc. of 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). – 2019.
28. Bairagi A.K. et al. Coexistence Mechanism Between eMBB and uRLLC in 5G Wireless Networks // IEEE Transactions on Communications. – 2021. – Vol. 69, No. 3. – P. 1736–1749.
29. Alsenwi M., Tran N.H., Bennis M., Pandey S.R., Bairagi A.K., and Hong C.S. Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach // IEEE Transactions on Wireless Communications. – 2021. – Vol. 20, No. 7. – P. 4585–4600.
30. Alsenwi M., Tran N.H., Bennis M., Kumar Bairagi A., and Hong C.S. eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach // IEEE Communications Letters. – 2019. – Vol. 23, No. 4. – P. 740–743.
31. 3GPP TS 23.501: System architecture for the 5G system: Release 15. – ETSI 3GPP. – 2018.
32. Guan W., Wen X., Wang L., Lu Z., and Shen Y. A service-oriented deployment policy of end-to-end network slicing based on complex network theory // IEEE Access. – 2018. – Vol. 6. – P. 19691–19701.
33. Sun G., Xiong K., Boateng G.O., Liu G., and Jiang W. Resource slicing and customization in RAN with dueling deep Q-network» // Journal of Network and Computer Applications. – 2020. – Vol. 157.
34. Caballero P., Banchs A., de Veciana G., and Costa-Perez X. Network slicing games: Enabling customization in multi-tenant mobile networks // IEEE/ACM Transactions on Networking. – 2019. – Vol. 27, No. 2. – P. 662–675.
35. Samuylov A., Moltchanov D., Kovalchukov R., Pirmagomedov R., Gaidamaka Y., Andreev S., Koucheryavy Y., and Samouylov K. Characterizing resource allocation trade-offs in 5G NR serving multicast and

unicast traffic // IEEE Transactions on Wireless Communications. – 2020. – Vol. 19, No. 5. – P. 3421–3434.

36. Kovalchukov R., Moltchanov D., Gaidamaka Y., and Bobrikova E. An Accurate Approximation of Resource Request Distributions in Millimeter Wave 3GPP New Radio Systems // Lecture notes in computer science. – 2019. – Vol. 11660. – P. 572–585.
37. Koucheryavy Y., Lisovskaya E., Moltchanov D., Kovalchukov R., and Samuylov A. Quantifying the millimeter wave new radio base stations density for network slicing with prescribed SLAs // Computer Communications. – 2021. – Vol. 174 – P. 13–27.
38. Кочеткова И.А., Власкина А.С., Ву Н.Н., Шоргин В.С. Система массового обслуживания с управляемым по сигналам перераспределением приборов для анализа нарезки ресурсов сети 5G. Информатика и её применения – 2021. – Т. 15, № 3. – С. 91–97.
39. Moltchanov D. Distance distributions in random networks // Ad Hoc Networks. – 2012. – Vol.10, No. 6. – P. 1146–1166.
40. Ivanova D., Markova E., Moltchanov D., Pirmagomedov R., Koucheryavy Y., and Samouylov K. Performance of Priority-Based Traffic Coexistence Strategies in 5G mmWave Industrial Deployments // IEEE Access. – 2022. – Vol. 10. – P. 9241–9256.
41. Santalo L.A. Integral geometry and geometric probability. – Cambridge: Cambridge University Press. – 2004.
42. ITU-T: Requirements of the IMT-2020 Network. – ITU-T. – 2018. – 26 p.
43. Li Y., Zheng J., Li Z., Liu Y., Qian F., Bai S., Liu Y., and Xin X. Understanding the ecosystem and addressing the fundamental concerns of commercial MVNO // IEEE/ACM Transactions on Networking. – 2020. – Vol. 28. – P. 1364–1377.
44. GSM Association: Generic Network Slice Template. – GSM Association. – 2020. – 66 p.

45. ITU-T: Framework for the Support of Network Slicing in the IMT-2020 Network. – ITU-T. – 2018. – 20 p.
46. 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.
47. Begishev V., Sopin E., Moltchanov D., Kovalchukov R., Samuylov A., Andreev S., Koucheryavy Y., and Samouylov K. Joint Use of Guard Capacity and Multiconnectivity for Improved Session Continuity in Millimeter-Wave 5G NR Systems // IEEE Transactions on Vehicular Technology. – 2021. – Vol. 70, No. 3. – P. 2657–2672.
48. Горцев А.М., Назаров А.А., Терпугов А.Ф. Управление и адаптация в системах массового обслуживания. – Томск: Изд-во Томского университета, 1978. – 208 с.
49. Башарин Г.П., Гайдамака Ю.В., Самуйлов К.Е. Математическая теория телетрафика и ее приложения к анализу мультисервисных сетей связи следующих поколений // Автоматика и вычислительная техника. – 2013. – № 2. – С. 11–21.
50. Горбунова А.В., Наумов В.А., Гайдамака Ю.В., Самуйлов К.Е. Ресурсные системы массового обслуживания как модели беспроводных систем связи // Информатика и её применения – 2018. – Т. 12, № 3. – С. 48–55.
51. Moltchanov D., Sopin E., Begishev V., Samuylov A., Koucheryavy Y., and Samouylov K. A tutorial on mathematical modeling of 5G/6G millimeter wave and terahertz cellular systems // IEEE Communications Surveys & Tutorials. – 2022. – Vol. 24, No. 2. – P. 1072–1116.
52. Yashkov S. and Yashkova A. Processor sharing: A survey of the mathematical theory // Automation and Remote Control. – 2007. – Vol. 68. – P. 1662–1731.

53. Макеева Е.Д., Поляков Н.А., Харин П.А., Гудкова И.А. Вероятностная модель для анализа характеристик совместной передачи трафика URLLC и eMBB в беспроводных сетях // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. – 2020. – № 52. – С. 33–42.

54. Haque M.E., Tariq F., Khandaker M.R.A., Wong K.-K., and Zhang Y. A Survey of Scheduling in 5G URLLC and Outlook for Emerging 6G Systems // IEEE Access. – 2023. – Vol. 11. – P. 34372–34396.

55. Бегишев В.О., Самуйлов А.К., Молчанов Д.А., Самуйлов К.Е. Стратегии распределения радиоресурсов в гетерогенных сетях с трафиком Narrow-Band IoT // Системы и средства информатики. – 2017. – Т. 27, № 4. – С. 64–79.

56. Begishev V., Petrov V., Samuylov A., Moltchanov D., Andreev S., Koucheryavy Y., and Samouylov K. Resource allocation and sharing for heterogeneous data collection over conventional 3GPP LTE and emerging NB-IoT technologies // Computer Communications. – 2018. – Vol. 120. – P. 93–101.

57. ITU-R: Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s). – ITU-R. – 2017. – 11 p.

58. Ghosh A., Ratasuk R., and Rao A.M. Industrial IoT Networks Powered by 5G New Radio // Microwave Journal. – 2019. – Vol. 62, No. 12.

59. Navarro-Ortiz J., Romero-Diaz P., Sendra S., Ameigeiras P., Ramos-Munoz J.J., and Lopez-Soler J.M. A survey on 5G usage scenarios and traffic models. // IEEE Communications Surveys & Tutorials. – 2020. – Vol. 22, No. 2. – P. 905–929.

60. Gangakhedkar S., Cao H., Ali A.R., Ganesan K., Gharba M., and Eichinger J. Use cases, requirements and challenges of 5G communication for industrial automation // Proc. of 2018 IEEE International Conference on Communications Workshops (ICC Workshops). – 2018.

61. Башарин Г.П., Штатное С.В. Мультисервисная модель обслуживания эластичного трафика с конечным числом источников // T-Comm. – 2010. – № 7. – С. 4–7.

62. Basharin G.P. and Aterekova T.V. Analytical model of streaming and elastic traffic with dynamic channel allocation scheme // Proc. of 2010 International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT 2010). – 2010. – P. 1086–1090.

63. Gaidamaka Y.V. and Samouylov K.E. Analytical model of multicast network and single link performance analysis // Proc. of the 6-th International Conference on Telecommunications (ConTEL-2001). – 2001. – P. 169–175.

64. Samouylov K.E., Gaidamaka Y.V., Gudkova I.A., Zaripova E.R., and Shorgin S.Y. Baseline Analytical Model for Machine-type Communications over 3GPP RACH in LTE-advanced Networks // Computer and Information Sciences. – 2016. – Vol. 659. – P. 203–213.

65. Yarkina N., Gaidamaka Y., Correia L.M., and Samouylov K. An analytical model for 5G network resource sharing with flexible SLA-oriented slice isolation // Mathematics. – 2020. – Vol. 8, No. 7.

66. Gorshenin A., Kozlovskaya A., Gorbunov S., and Kochetkova I. Mobile network traffic analysis based on probability-informed machine learning approach // Computer Networks. – 2024. – Vol. 247.

67. Zeifman A., Satin Y., Morozov E., Nekrasova R., and Gorshenin A. On the ergodicity bounds for a constant retrial rate queueing model // Proc. of 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). – 2016. – P. 269–272.

68. Kochetkova I., Makeeva E., Ageeva A., and Gorshenin A. Model for Analyzing Impact of Path Loss on eMBB Bit Rate Degradation Under Priority URLLC Transmission in 5G Network // Lecture Notes in Computer Science. – 2022. – Vol. 13766. – P. 176–189.

69. Кучерявый Е.А. Кучерявый А.Е., Футахи А. LTE и беспроводные сенсорные сети // Мобильные телекоммуникации. – 2012. – №9. – С. 38–41.

70. Кучерявый А.Е., Парамонов А.И., Кучерявый Е.А. Сети связи общего пользования. Тенденции развития и методы расчета. – М.: ФГУП ЦНИИС, 2008. – 296 с.

71. Гольдштейн Б.С., Кучерявый А.Е. Сети связи пост-NGN. – СПб: БХВ-Петербург, 2013. – 160 с.

72. Кучерявый Е.А. Управление трафиком и качество обслуживания в сети интернет. – М.: Наука и техника, 2004. – 336 с.

73. Moltchanov D., Samuylov A., Petrov V., Gapeyenko M., Himayat N., Andreev S., and Koucheryavy Y. Improving session continuity with bandwidth reservation in mmWave communications // IEEE Wireless Communications Letters. – 2018. – Vol. 8, No. 1. – P. 105–108.

74. Mahmood O.A., Khakimov A., Muthanna A., and Paramonov A. Effect of Heterogeneous Traffic on Quality of Service in 5G Network // Lecture Notes in Computer Science. – 2019. – Vol. 11965. – P. 469–478.

75. Ateya A.A., Alhussan A.A., Abdallah H.A., Al duailij M.A., Khakimov A., and Muthanna A. Edge Computing Platform with Efficient Migration Scheme for 5G/6G Networks // Computer Systems Science and Engineering. – 2023. – Vol. 45, No. 2. – P. 1775–1787.

76. Ateya A.A., Muthanna A., Koucheryavy A., Maleh Y., and El-Latif A.A. Energy efficient offloading scheme for MEC-based augmented reality system // Cluster Computing. – 2023. – Vol. 26. – P. 789–806.

77. Naumov V.A. and Samouylov K.E. On the modeling of queueing systems with multiple resources // RUDN Journal of Mathematics, Information Sciences and Physics. – 2014. – Vol. 3. – P. 60–64.

78. Naumov V., Samouylov K., Yarkina N., Sopin E., Andreev S., and Samuylov A. LTE performance analysis using queuing systems with finite resources and random requirements // Proc. of the 7th International Congress

on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). – 2015. – P. 100–103.

79. Самуйлов К.Е., Сопин Э.С., Шоргин С.Я. Система массового обслуживания с ограниченными ресурсами и сигналами для анализа показателей эффективности беспроводных сетей // Информатика и ее применения. – 2017. – Т. 11, № 3. – С. 99–105.
80. Степанов С.Н. Модель совместного обслуживания трафика сервисов реального времени и трафика данных. I // Автоматика и телемеханика. – 2011. – № 4. – С. 121–132.
81. 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.
82. Haenggi M., Andrews J.G., Baccelli F., Dousse O., and Franceschetti M. Stochastic geometry and random graphs for the analysis and design of wireless networks // IEEE Journal on Selected Areas in Communications. – 2009. – Vol. 27, No. 7. – P. 1029–1046.
83. Andrews J.G., Bai T., Kulkarni M.N., Alkhateeb A., Gupta A.K., and Heath R.W. Modeling and Analyzing Millimeter Wave Cellular Systems // IEEE Transactions on Communications. – 2017. – Vol. 65, No. 1. – P. 403–430.
84. Yarkina N., Correia L.M., Moltchanov D., Gaidamaka Y., and Samouylov K. Multi-tenant resource sharing with equitable-priority-based performance isolation of slices for 5G cellular systems // Computer Communications. – 2022. – Vol. 188. – P. 39–51.
85. Ageev K., Garibyan A., Golskaya A., Gaidamaka Yu., Sopin E., Samouylov K., and Correia L.M. Modelling of Virtual Radio Resources Slicing in 5G Networks // Communications in Computer and Information Science. – 2019. – Vol. 1109. – P. 150–161.
86. Vassilakis V.G., Moscholios I.D., and Logothetis M.D. Call-Level Performance Modelling of Elastic and Adaptive Service-Classes with Finite

Population // IEICE Transactions on Communications. – 2008. – Vol. 91, No. 1. – P. 151–163.

87. Malanchini I., Valentin S., and Aydin O. Generalized resource sharing for multiple operators in cellular wireless networks // Proc. of 2014 International Wireless Communications and Mobile Computing Conference (IWCMC). – 2014. – P. 803–808.
88. Malanchini I., Valentin S., and Aydin O. Wireless resource sharing for multiple operators: Generalization, fairness, and the value of prediction // Computer Networks. – 2016. – Vol. 100. – P. 110–123.
89. Lisovskaya E., Moiseeva S., and Pagano M. The Total Capacity of Customers in the Infinite-Server Queue with MMPP Arrivals // Communications in Computer and Information Science. – 2015. – Vol. 678. – P. 110–120.
90. Башарин Г.П. Лекции по математической теории телетрафика. – М.: РУДН, 2009. – 342 с.
91. Башарин Г.П., Бочаров П.П., Коган Я.А. Анализ очередей в вычислительных сетях. Теория и методы расчета. – М.: Наука, 1989. – 336 с.
92. Башарин Г.П., Самуйлов К.Е., Яркина Н.В., Гудкова И.А. Новый этап развития математической теории телетрафика // Автоматика и телемеханика. – 2009. – № 12. – С. 16–28.
93. Башарин Г.П. Введение в теорию вероятностей: Учебное пособие. – М.: Изд-во РУДН, 1990. – 228 с.
94. Бочаров П.П., Печинкин А.В. Теория массового обслуживания: Учебник. – М.: Изд-во РУДН, 1995. – 529 с.
95. Bocharov P.P., D'Apice C., Pechinkin A.V., and Salerno S. Queueing Theory. – Brill Academic Publishers, 2004. – 457 p.
96. Dudin A.N., Klimenok V.I., and Vishnevsky V.M. The theory of queuing systems with correlated flows // Cham: Springer International Publishing, 2019. – 410 p.

97. Вишневский В.М., Дудин А.Н., Клименок В.И. Стохастические системы с корреляционными потоками. Теория и применение в телекоммуникационных сетях. – М.: Техносфера, 2018. – 564 с.
98. Вишневский В.М. Теоретические основы проектирования компьютерных сетей. – М.: Техносфера, 2003. – 512 с.
99. Вишневский В.М., Портной С.Л., Шахнович И.В. Энциклопедия WiMAX. Путь к 4G. – М.: Техносфера, 2009. – 472 с.
100. Borodakiy V., Samouylov K., Gaidamaka Yu., Abaev P., Buturlin I., and Etezov Sh. Modelling a Random Access Channel with Collisions for M2M Traffic in LTE Networks // Lecture Notes in Computer Science. – 2014. – Vol. 8638. – P. 301–310.
101. Разумчик Р.В., Зейфман А.И., Коротышева А.В., Сатин Я.А. Анализ энергоэффективности вычислительного комплекса, моделируемого с помощью системы обслуживания с пороговым управлением и интенсивностями, зависящими от времени // Системы и средства информатики. – 2015. – Т. 25, №4. – С. 19–30.
102. Зейфман А.И., Бенинг В.Е., Соколов И.А. Марковские цепи и модели с непрерывным временем. – М.: Элекс-КМ, 2008. – 167 с.
103. Семенова О.В., Дудин А.Н. Система массового обслуживания M|M|N с управляемым режимом обслуживания и катастрофическими сбоями // Автоматика и вычислительная техника. – 2007. – № 6. – С. 72–80.
104. Moiseev A., Shklennik M., and Polin E. Infinite-server queueing tandem with Markovian arrival process and service depending on its state // Annals of Operations Research. – 2023. – Vol. 326. – P. 261–279.
105. Moiseev A. and Nazarov A. Asymptotic Analysis of the Infinite-Srever Queueing System with High-Rate Semi-Arrivals // Proc. of the IEEE International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT 2014). – 2014. – P. 607–613.
106. Сонькин М.А., Моисеев А.Н., Сонькин Д.М., Буртовая Д.А. Объектная модель приложения для имитационного моделирования циклических

систем массового обслуживания // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2017. – № 40. – С. 71–80.

107. Назаров А.А., Моисеева С.П. Метод асимптотического анализа в теории массового обслуживания: монография. – Томск: Изд-во Научно-технической литературы, 2006. – 109 с.
108. Моисеева С.П., Панкратова Е.В., Убонова Е.Г. Исследование бесконечнолинейной системы массового обслуживания с разнотипным обслуживанием и входящим потоком марковского восстановления // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2016. – Т. 2, Вып. 35. – С. 46–53.
109. Назаров А.А., Терпугов А.Ф. Теория вероятностей и случайных процессов: Учебное пособие. – Томск: Изд-во Научно-технической литературы, 2006. – 204 с.
110. Наумов В.А. Численные методы анализа марковских систем. – М.: Изд-во УДН, 1985. – 37 с.
111. Наумов В.А., Самуйлов К.Е., Яркина Н.В. Теория телетрафика мультисервисных сетей: Монография. – М.: РУДН, 2007. – 191 с.
112. Пшеничников А.П., Даудов И.М. Концептуальные основы будущих сетей // REDS: Телекоммуникационные устройства и системы. – 2022. – Т. 12, № 2. – С. 24–28.
113. Корнышев Ю.Н., Пшеничников А.П., Харкевич А.Д. Теория телетрафика: Учебник для вузов. – М.: Радио и связь, 1996. – 272 с.
114. Рыков В.В., Самуйлов К.Е. К анализу вероятностей блокировок ресурсов сети с динамическими многоадресными соединениями // Электросвязь. – 2000. – № 10. – С. 27–30.
115. Рыков В.В., Ефросинин Д.В. К анализу характеристик производительности СМО с неоднородными приборами // Автоматика и телемеханика. – 2008. – №1. – С. 64–82.

116. Рыков В.В. Управляемые системы массового обслуживания // Теория вероятностей. Математическая статистика. Теоретическая кибернетика. – 1975. – Т. 12. – С. 43–153.
117. Самуйлов К.Е. Метод расчета вероятностных характеристик модели сети с многоадресными соединениями // Вестник РУДН. Прикладная и компьютерная математика. – 2003. – Т. 2, № 1. – С. 45–51.
118. Лагутин В.С., Степанов С.Н. Телетрафик мультисервисных сетей связи. – М.: Радио и связь, 2000. – 320 с.
119. Степанов С.Н. Основы телетрафика мультисервисных сетей. – М.: Изд-во «Эко-Трендз», 2010. – 392 с.
120. Andrabi U.M., Stepanov S.N., Stepanov M.S., Kanishcheva M.G., and Habinshuti F.X. The Model of Conjoint Servicing of Real Time and Elastic Traffic Streams Through Processor Sharing (PS) Discipline with Access Control // Proc. of International Conference Engineering and Telecommunication (En&T). – 2021.
121. Степанов С. Н., Цитович И. И. Эквивалентные определения вероятностных характеристик моделей с повторными вызовами и их применение // Проблемы передачи информации. – 1989. – Т. 25, Вып. 2. – С. 79–90.
122. Сегайер А., Цитович И.И. Построение моделей мультисервисных сетей // Электросвязь. – 2009. – № 9. – С. 54–57.
123. Shorgin S., Samouylov K., Gudkova I., Galinina O., and Andreev S. On the benefits of 5G wireless technology for future mobile cloud computing // Proc. of 2014 International Science and Technology Conference (Modern Networking Technologies) (MoNeTeC). – 2014. – P. 151–154.
124. Dohler M., Watteyne T., Alonso-Zrate J. Machine-to-machine: an emerging communication paradigm // Tutorial in the Second International ICST Conference on Mobile Networks And Management (MONAMI 2010). – 2010.

125. Dohler M., Li Y. Wireless Relay Channel in Cooperative Communications: Hardware, Channel & Phy. – John Wiley & Sons, 2010. – 464 p.
126. Iversen V.B. Teletraffic Engineering Handbook. – Technical University of Denmark, 2002. – 324 p.
127. Iversen V.B. Teletraffic engineering and network planning. – Technical University of Denmark, 2011. – 583 p.
128. Kelly F.P. Reversibility and stochastic networks. – Cambridge University Press, 2011. – 238 p.
129. Benameur N., Fredj S.B., Oueslati-Boulahia S., and Roberts J.W. Quality of service and flow level admission control in the Internet // Computer Networks. – 2002. – Vol. 40, No. 1. – P. 57–71.
130. Ross K.W. Multiservice loss models for broadband telecommunication networks. – London: Springer-Verlag, 1995. – 343 p.
131. Kondratyeva A., Ivanova D., Begishev V., Markova E., Mokrov E., Gaidamaka Y., and Samouylov K. Characterization of Dynamic Blockage Probability in Industrial Millimeter Wave 5G Deployments // Future Internet. – 2022. – Vol. 14, No. 7.
132. Иванова Д.В., Жбанкова Е.А., Маркова Е.В., Гайдамака Ю.В. Модели совместного обслуживания трафика eMBB и URLLC на основе приоритетов в промышленных развертываниях 5G NR // Информатика и её применения. – 2023. – Т. 17, № 4. – С. 64–70.
133. Ivanova D., Adou Y., Markova E., Gaidamaka Y., and Samouylov K. Mathematical Framework for Mixed Reservation- and Priority-Based Traffic Coexistence in 5G NR Systems // Mathematics. – 2023. – Vol. 11, No. 4.
134. Иванова Д.В., Жбанкова Е.А., Маркова Е.В., Гайдамака Ю.В. СМО с прерыванием обслуживания для моделирования нарезки радиоресурсов в беспроводных сетях 5G // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2023. – №65. – С. 36–46.

135. Иванова Д.В., Маркова Е.В. Расчет характеристик прерывания передачи eMBB трафика в сетях 5G при реализации абсолютного приоритета в обслуживании URLLC // Свидетельство о государственной регистрации программы для ЭВМ, № RU2021661768, 15.07.2021 г.
136. Иванова Д.В., Маркова Е.В., Молчанов Д.А. Расчет характеристик прерывания передачи eMBB трафика в сетях 5G при реализации относительного приоритета в обслуживании URLLC трафика // Свидетельство о государственной регистрации программы для ЭВМ, № RU2021661641, 14.07.2021 г.