ОРДЕНА ТРУДОВОГО КРАСНОГО ЗНАМЕНИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ»

На правах рукописи

НДАЙИКУНДА ЖУВЕН

Разработка и анализ модели динамического распределения ресурса беспроводных узлов доступа при передаче неоднородного трафика IoT

Специальность 2.2.15— Системы, сети и устройства телекоммуникаций

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Степанов Сергей Николаевич

Москва, 2022

Оглавление

Введение

Актуальность темы исследования. Одной из основных тенденций развития телекоммуникаций является необходимость совместного обслуживания выделенным ресурсом потоков информационных сообщений, отличающихся существенным разнообразием в требованиях к ресурсу и качеству обслуживания. Часто источниками информационных сообщений являются устройства телеметрии, видеокамеры и т.п. Подобные устройства являются элементами сети виртуального оператора, предоставляющего услуги сбора и обработки данных разного рода наблюдений. Эта деятельность регулируется положениями концепции интернета Вещей.

Нередко виртуальные сети разворачиваются в местах, где ограничено или вообще не имеется возможности применения фиксированной проводной связи. Это вынуждает виртуального оператора использовать ресурс беспроводных сетей для обслуживания возникающих информационных потоков. В силу известных причин[1], этот ресурс ограничен и должен использоваться с максимальной эффективностью.

Чтобы добиться этого результата, нужно решить следующие две задачи. Во-первых, построить модель формирования и обслуживания сессий связи, которая более точно отражает реалии работы действующих узлов беспроводного доступа. Во-вторых, предложить и исследовать сценарии распределения ресурса между поступающими потоками информационных сообщений, которые бы позволили создать условия для их дифференцированного обслуживания. Иначе, как показали численные эксперименты, с ростом нагрузки на канал происходит неконтролируемое перераспределение ресурса в пользу потоков сессий с относительно малыми требованиями к скорости передачи. Этот результат может нарушить принятое соглашение об обслуживании. Решение сформулированных задач позволит находить предпочтительные соотношения между параметрами потоков запросов на информационное обслуживание и характеристиками пропускной способности мультисервисного узла доступа, обеспечивающие гарантированное качество обслуживания клиентов. Именно эти вопросы рассматривались в диссертационной работе, что говорит об актуальности выбранной тематики.

Степень разработанности темы. Поставленная задача решалась на базе моделей и методов теории телетрафика, а также возможностей, заложенных в механизмы управления процессом обслуживания сессий связи в современных беспроводных мультисервисных узлах доступа. Различным аспектам решения данной задачи посвящены работы российских и зарубежных авторов. В их числе: Г.П. Башарин, В.М. Вишневский, Ю.В. Гайдамака, В.Г. Карташевский, А.Е.

[1]В их число входят ограничения физического плана, а также действия регулятора, направленные на создание конкуренции.

Кучерявый, Е.А. Кучерявый, В.А. Наумов, А.П. Пшеничников, К.Е. Самуйлов, С.Н. Степанов, М.С. Степанов, И.И. Цитович, и др., а также – T.Bonald, F.P. Kelly, V.B. Iversen, K.W. Ross, J. Virtamo и др. Отдельные вопросы построения и исследования моделей распределения ресурса в беспроводных узлах доступа рассматривались в диссертационных работах: С.Д. Андреева, В.О. Бегишева, Е.А. Кучерявого, К.А. Агеева и др. Анализ публикаций и выполненных диссертационных исследований показал, что в большинстве теоретических работ либо изучалось действие какого-то одного фактора на процесс распределения ресурса узла доступа (например, зависимость требования к ресурсу от типа сервиса, ограничение доступа, резервирование ресурса и т.д.), либо процесс распределения ресурса рассматривался с избыточной детальностью, что в итоге затрудняло использование построенной математической модели. Задача построения модели, которая, с одной стороны, отражала основные реалии распределения ресурса, а с другой — могла бы использоваться в практических приложениях не рассматривалась, что и определило направление исследований, выполненных в диссертации.

Цели и задачи работы. Целью исследования является разработка и анализ процедуры динамического распределения ресурса беспроводного узла доступа, направленной на создание условий по дифференцированному обслуживанию неоднородного трафика и повышению эффективности использования ресурса передачи информации. Для достижения указанной цели необходимо решить следующие частные научные задачи: разработать модель динамического распределения ресурса беспроводного узла доступа при обслуживании неоднородного трафика при наличии ограничения по досупу; определить характеристики качества обслуживания поступающих сессий связи; построить алгоритмы оценки характеристик; сформулировать рекомендации по эффективному распределению ресурса между поступающими потоками разнородного трафика.

Научная новизна.

1. Построена и исследована обобщенная модель обслуживания неоднородного трафика в беспроводном узле доступа, которая в отличие от известных моделей позволила учесть совместное влияние основных значимых факторов, определяющих совместное обслуживание трафика реального времени и эластичных данных. Среди них: наличие приоритета у трафика реального времени; использование дисциплины Processor Sharing при передаче эластичного трафика; ограничение по доступу для всех видов трафика, зависящее от общего уровня занятости ресурса.

2. Получены выражения для оценки характеристик качества обслуживания заявок через значения входных параметров и стационарных вероятностей обобщенной модели беспроводного узла доступа. В отличие от более ранних исследований, полученные выражения позволяют анализировать действие разного рода процедур, направленных на повышение эффективности использования ресурса передачи узлов доступа и создание условий по дифференцированному обслуживанию потоков неоднородного трафика, основанных на ограничении доступа, зависящего от общего уровня занятости ресурса.
3. Построена система уравнений статистического равновесия, связывающая значения стационарных вероятностей модели и разработан алгоритм ее решения. В отличие от известных реализаций других стандартных методов разработанный алгоритм позволяет вести оценку характеристик для моделей с числом состояний в несколько миллионов, что достаточно для исследования условий по дифференцированному обслуживанию поступающих потоков неоднородного трафика для большинства практических приложений.

Теоретическая и практическая значимость работы. Теоретическая значимость работы заключается в построении и исследовании обобщенной модели обслуживания неоднородного трафика в беспроводном узле доступа, которая позволила учесть совместное влияние основных значимых факторов, определяющих совместное обслуживание трафика реального времени и эластичных данных, а также в разработке алгоритмов расчета характеристик подобных моделей. Получены программные реализации построенных в диссертации алгоритмов. Разработанный инструментарий рекомендуется использовать для создания условий по дифференцированному обслуживанию гетерогенного трафика в беспроводных узлах доступа и теоретическом обосновании действий администрации, направленных на повышение эффективности использования ресурса передачи. Результаты диссертации использованы в учебном процессе на кафедре «Сети связи и системы коммутации» МТУСИ. Реализация результатов работы подтверждена соответствующим актом.

Методы исследования. Для решения поставленной задачи применялись методы теории телетрафика, теории вероятностей и вычислительной математики.

Основные положения, выносимые на защиту:

1. Построенная обобщенная модель обслуживания неоднородного трафика в беспроводном узле доступа позволяет учесть совместное влияние основных значимых факторов, определяющих совместное обслуживание трафика реального времени и эластичных данных.

Среди них: наличие приоритета у трафика реального времени; использование дисциплины Processor Sharing при передачи эластичного трафика; ограничение по доступу для всех видов трафика, зависящее от общего уровня занятости ресурса.

2. Для оценки значений характеристик качества совместного обслуживания сессий передачи эластичного трафика и эластичного трафика данных заявок в рамках построенной модели беспроводного узла доступа рекомендуется использовать метод, основанный на решении системы уравнений равновесия итерационным алгоритмом Гаусса-Зейделя. Этот подход позволяет рассчитать характеристики для моделей с числом состояний в несколько миллионов, что достаточно для исследования условий по дифференцированному обслуживанию поступающих потоков неоднородного трафика для большинства практических приложений.
3. Разработанные модель и алгоритмы оценки ее характеристик позволяют анализировать действие разного рода процедур, направленных на повышение эффективности использования ресурса передачи узлов доступа и создание условий по дифференцированному обслуживанию потоков неоднородного трафика, основанных на ограничении доступа, зависящего от общего уровня занятости ресурса. Среди них динамичный слайсинг, когда распределение выделенного объема ресурса осуществляется на динамической основе и зависит от его загрузки. Для ограничения доступа сессий здесь предлагается использоваться процедуру резервирования, основанную на фильтрации поступающих сессий с использованием функции внутренней блокировки. Другой сценарий — статичный слайсинг. Для данного сценария имеющийся ресурс делится между поступающими потоками в определенной пропорции, зависящей от требований сессий связи к показателям качества обслуживания.
4. Выполненное численное исследование показало, что использование динамического слайсинга позволяет на 5–20% уменьшить требование к объему ресурса, обеспечивающего требуемый уровень потерь сессий, по сравнению с применением для этих же целей статичного слайсинга. Наибольший эффект применение предложенной версии динамического слайсинга приносит в ситуации обслуживания эластичного трафика данных с использованием дисциплины Processor Sharing.

Степень достоверности и апробация результатов. Полученные теоретические результаты обоснованы доказательствами с использованием математических методов теории телетрафика, подтверждены численными экспериментами. Достоверность положений и выводов диссертации

подтверждается апробацией работы, основные результаты которой обсуждались и докладывались на международной научно-технической конференции «Технологии информационного общества» (Москва, 2019 — 2021 гг.), на отраслевой научно-технической конференции «Телекоммуникационные и вычислительные системы» (Москва, 2019 гг.), на международной научной конференции «Systems of Signals Generating and Processing in the Field of on Board Communications» (Москва, 2020 — 2021 гг.), на международной научной конференции «Conference of Open Innovation Association, FRUCT» (Москва, 2019 гг.), на международной научной конференции «Distributed Computer and Communication Networks: Control, Computation, Communications» (Москва, 2020 — 2021 гг.). По материалам диссертации опубликованы 14 работ, в том числе 3 — в изданиях, включенных в список ВАК РФ и 4 в изданиях, входящих международную базу цитирования SCOPUS.

Основное содержание работы. Диссертация состоит из введения, четырех разделов, заключения, списка литературы и приложения. Основная часть (без приложения) изложена на 134 страницах машинописного текста, содержит 46 рисунков и 10 таблиц; список литературы состоит из 111 наименований. Приложение изложено на 1 странице машинописного текста.

Раздел 1

Анализ возможностей построения сетей IoT на основе существующей инфраструктуры сетей мобильной связи

1.1. Введение к разделу 1

Телекоммуникационные услуги передачи данных в комплексных сетях мобильной связи становятся более неоднородными, поскольку появляется необходимость поддерживать одновременно несколько категорий трафика, каждая из которых имеет свои собственные нагрузки, свое требование к качеству обслуживания и свою предпочтительную радиотехнологию. Решение перечисленных задач осуществляется путем внедрения новых технологий беспроводной связи и более совершенных алгоритмов распределения ресурсов, в том числе новых алгоритмов распределения ресурсов, предложенных в нашем исследовании. Технологии интернета вещей достигли значительного улучшения в сборе и в обработке больших данных, гетерогенности и производительности [76, 84]. В реализации концепции интернета вещей для беспроводной передачи данных важную роль играют такие качества, как отказоустойчивость, возможность самоорганизации, эффективность в условиях низких скоростей и адаптивность [61]. Главными направлениями эволюции беспроводных систем мобильной связи являются поддержка массовых соединения (см. рисунок 1.1), сверхнизкое энергопотребление, широкая зона покрытия и двунаправленный запуск между плоскостью сигнализации и плоскостью данных [58, 60], улучшение качества предоставления мультимедийных услуг, и т.д. Перечисленные направления развития мобильной связи достигаются благодаря внедрению технологий интернета вещей например технологии узкополосного интернета вещей NB-IoT (Narrow Band Internet of Things), которая отлично поддерживается в сетях мобильной связи [54] и использованию эффективных методов распределения ограниченного радиоресурсов сети мобильной связи. Технология NB-IoT является одной из самых перспективных технологий энергоэффективных сетей большого радиуса действия LPWAN (Low Power Wide Area Network).

Целью данного раздела является анализ возможности подключения массовых устройств к мобильным сетям LTE, а также реализация концепции интернета вещей в сетях мобильной связи пятого поколения. Проведен анализ способов разворачивания стандарта NB-IoT на инфраструктуре существующих сетей LTE. Технология LTE особенно удобна для реализации концепции интернета

вещей благодаря особым характеристикам радиоинтерфейса и применения методов распределения ограниченных ресурсов беспроводных сетей, учитывающих особенности устройств IoT.

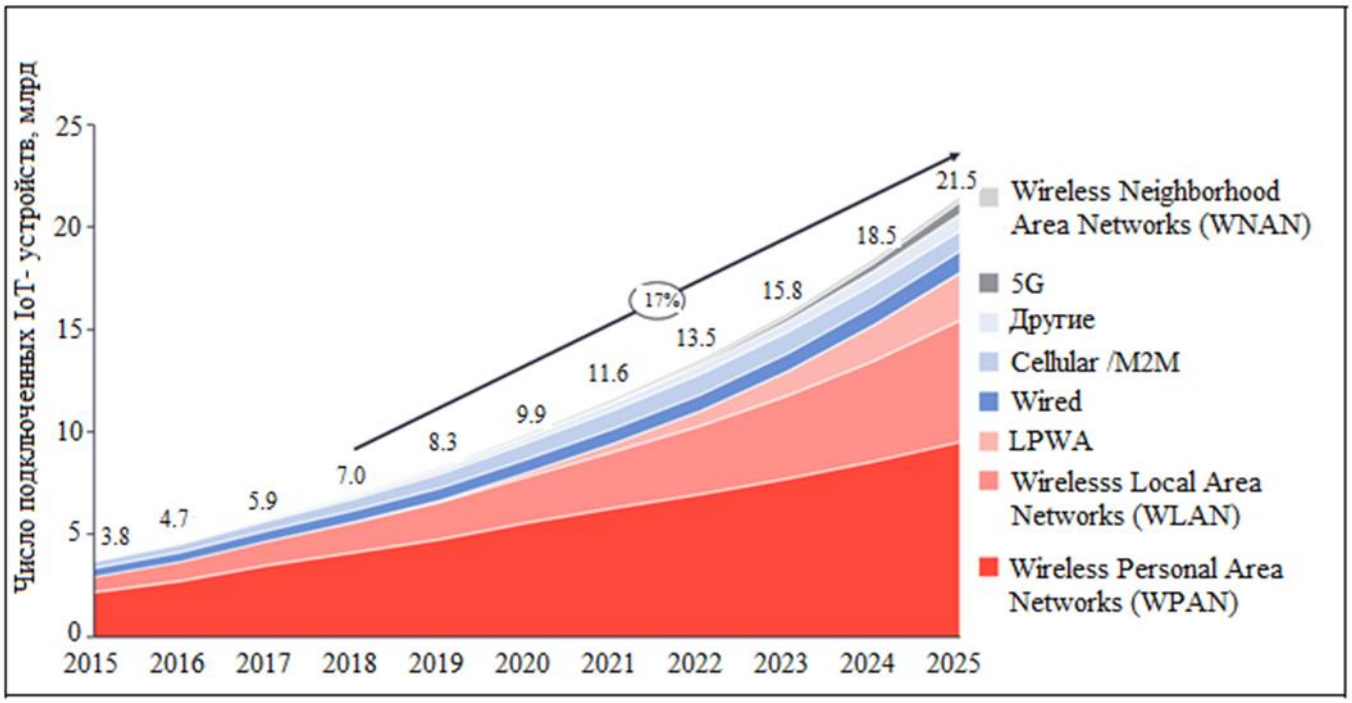


Рисунок 1.1 — Глобальное число подключенных IoT-устройств в мире [60]

1.2. Анализ особенностей построения беспроводных сетей LTE

В настоящее время технология LTE является самых используемых технологий, для сбора неоднородных информационных данных в беспроводных сетях доступа [6, 8]. В релизе 8 разработана архитектура сетей стандарта LTE для обеспечения более высокого уровня производительности. Архитектура 4G, известная как эволюция архитектуры системы SAE (System Architecture Evolution) [64], предлагает множество преимуществ по сравнению с архитектурами 2G и 3G, таких как новые методы маршрутизации, эффективные решения для совместного использования выделенной полосы частот, увеличение мобильности и пропускной способности [8, 15, 18, 48, 98]. Архитектура сети мобильной связи стандарта LTE с возможностью поддержки массовых подключений устройств интернета вещей показана на рисунке 1.2. Архитектура сети LTE состоит из пакетной сети EPC и сети радиодоступа E-UTRAN (Evolved Universal Terrestrial Radio Access Network) [9]. Пакетная сеть (англ. Evolved Packet Core — EPC) отвечает за общий контроль абонентских терминалов и настройку логических трактов передачи пакетов, называемых bearer. К основным элементам базовой сети EPC относятся [4, 9]:

- узел управления мобильностью MME (Mobility Management Entity);
- обслуживающий шлюз сети LTE S-GW (Serving Gateway);
- шлюз для взаимодействия с сетями других операторов P-GW (The Packet data network Gateway);
- сервер абонентских данных HSS (Home Subscriber Server);
- узел выставления счетов абонентам за оказанные услуги PCRF (Policy and Charging Resource Function).

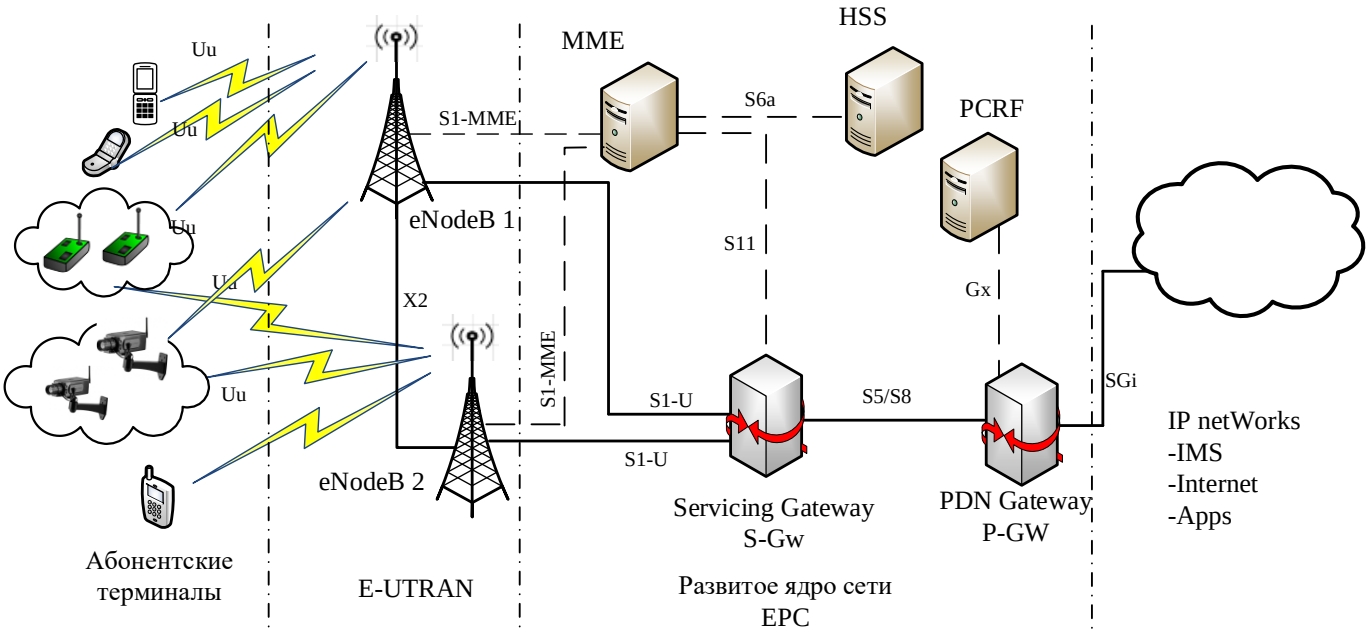


Рисунок 1.2 — Архитектура сети стандарта LTE

Шлюз для выхода на пакетные сети PGW осуществляет организацию точки доступа к внешним сетям. Например, он выделяет IP-адрес для новых абонентских терминалов, требующих подключение к сети. Он также отвечает за обеспечение качества обслуживания QoS (Quality of Service) и за тарификацию на основе используемых потоков данных. Через интерфейс SGi, каждый шлюз PDN обменивается данными с одним или несколькими внешними устройствами или сетями пакетной передачи данных, такими как серверы оператора IP-сети, интернет или мультимедийная подсистема IP. S-GW в LTE действует как маршрутизатор и отвечает за пересылку данных между eNodeB и PGW. S-GW служит локальной привязкой мобильности для хэндоверов между узлами eNodeB, а также привязкой мобильности для взаимодействия с другими технологиями 3GPP. MME выполняет операции сигнализации между абонентскими терминалами и базовой сетью CN. Он также отвечает за мобильность абонентских терминалов, хэндоверы,

механизмы отслеживания и пейджинга абонентского терминала при установлении соединения. Узел выставления счетов PCRF является управляющим сервером. Его основной задачей является централизация управления ресурсами сети, тарификация и учет предоставляемых услуг. Элементом HSS является большая база данных, хранящая данные абонентов. Он выполняет функции различных регистров VLR (Visitors Location Register), HLR (Home Location Register), AUC (Authentification Center), EIR (Equipment Identity Register), которые использовались в сетях 2G и 3G. E-UTRAN управляет радиосвязью между мобильными устройствами и развитым пакетным ядром, и имеет только два узла – базовая станция eNodeB и абонентский терминал. Каждый eNodeB является базовой станцией, которая управляет мобильными устройствами в одной или нескольких сотах. Мобильное устройство обменивается данными только с одной базовой станцией и одной сотой одновременно, что отличается от хэндовера в UMTS [47]. Базовая станция eNodeB имеет две основные функции. Во-первых, eNodeB посылает радиосигналы на все мобильные устройства, которые подключены к ней по нисходящей линии связи и принимает сигналы от них по восходящей линии связи, используя функции аналоговой и цифровой обработки сигналов радиоинтерфейса LTE. Во-вторых, eNodeB контролирует низкоуровневую работу всех своих мобильных устройств, посылая им сигнальные сообщения, такие как команды передачи, относящиеся к этим радиопередачам. При выполнении этих функций eNodeB объединяет более ранние функции узла NodeB и контроллера радиосети, чтобы уменьшить задержку, возникающую, когда мобильное устройство обменивается информацией с сетью. К основным функциям сети радиодоступа E-UTRAN относятся:

- Управление ресурсами. Функция управления ресурсами включает в себя такие функции, как контроль радиоканалов, контроль радиодоступа, контроль мобильности радиосвязи и планирование радиоресурсов для абонентских терминалов как в восходящей, так и в нисходящей линии связи.
- Функция сжатия заголовка. С помощью этой функции, RAN стремится эффективно уменьшить заголовки IP-пакетов.
- Безопасность. Функция безопасности позволяет шифровать все данные, передаваемые по радиоинтерфейсу.
- Позиционирование. Его роль заключается в предоставлении всей необходимой информации для определения местоположения абонентских терминалов.
- Возможность подключения к CN. Данная функция отвечает за сигнализацию между MME и SGW.

Технология LTE основана на базе IP-технологий в отличие от предыдущих беспроводных технологий мобильной связи. Радиоинтерфейс стандарта LTE разработан с целью повышения технических характеристик, в том числе максимальной пропускной способности, минимальной задержки пакетов (<5 мс) и высокой спектральной эффективности по сравнению со стандартами предыдущих поколений. Используемый метод радиопередачи и приема радиосигнала в LTE, известен как множественный доступ с ортогональным частотным разделением OFDMA (Orthogonal Frequency-Division Multiple Access). OFDMA выполняет те же функции, что и любой другой метод множественного доступа, позволяя базовой станции обмениваться данными с несколькими различными мобильными устройствами одновременно. Существует модифицированный метод радиопередачи, известный как множественный доступ с ортогональным частотным разделением с одной несущей SC-FDMA (Single-Carrier Frequency Division Multiple Access). В частности, SC-FDMA и OFDMA используются в направлениях восходящей линии связи и нисходящей линии связи соответственно. Однако, OFDMA отличается от SC-FDMA в том, что OFDMA использует преимущества поднесущих, распределенных внутри всего спектра, когда SC-FDMA использует только смежные поднесущие. Кроме того, OFDMA обеспечивает высокую масштабируемость и высокую устойчивость сигнала при замираниях [47]. Он также имеет высокую устойчивость к межсимвольной интерференции, возникающей при многолучевом распространении сигналов [47]. Данные технологии обеспечивают не только улучшение спектральной эффективности по сравнению с предыдущими мобильными сетями 3G, но и высокие скорости передачи данных и низкую задержку, даже для пользователей в сценариях высокой мобильности. Теоретически, пропускная способность соты рассчитывается как количество передаваемых символов в секунду. Далее преобразуется в битах в секунду в зависимости от того, сколько битов может нести символ. При ширине полосы частот 20МГц, доступные ресурсные блоки составляют 100 RB [87]. При этом, общее число ресурсных элементов составляет $100 \times 12 \times 7 = 8400$ REs, т.е. 8400 символов. В зависимости от степени модуляции каждый символ несёт 2, 3, 4 или 6 бит при использовании 4 QAM, 8 QAM, 16 QAM и 64 QAM соответственно. Из этого следует, что 100 RB содержатся не более $8400 \times 6 = 50400$ бит (64 QAM) без учета битов, отдаваемых для передачи служебной информации и без учета используемой скорости кодирования кодека (англ. Coding rate). Применение модуляции с определенной скоростью кодирования означает что часть битов отделяется для передачи полезной информации, и остальная часть содержит избыточные биты. Полученный объем данных передается в течении 0,5мс, следовательно, скорость передачи в радио-канале составляет $50400 \text{ бит} / 0,5 \text{мс} = 100800000 \text{ бит} / \text{мс} = 100,8 \text{ Мбит} / \text{с}$. Если используется

система MIMO 2×2 , то при ширине полосы 20МГц скорость передачи данных удваивается и становится $100,8\times2=201,6$ Мбит/с. С учетом скорости кодирования кодера, пиковая скорость уменьшается на коэффициент избыточности кодека. При использовании системы MIMO 4×4, скорость передача не умножается на 4 из за условия радиоканала [47, 48, 53]. На практике достигать такие скорости невозможно, реальные скорости передачи с учетом битов, отдаваемых для передачи служебной информации, используемой схемы кодирования и системы MIMO представлены в таблицах 1.1 и 1.2.

Таблица 1.1 — Скорость передачи информации вниз при частотном дуплексе с нормальном CP, Мбит/с [16]

Эффективный MCs	MIMO	1.4 МГц	3 МГц	5 МГц	10 МГц	15 МГц	20 МГц
QPSK$^{1/2}$	Single	0,85	2,21	3,71	7,46	11,21	14,96
16QAM$^{1/2}$	2×2	3,35	8,53	14,29	28,69	43,09	57,49
16QAM$^{3/4}$	2×2	5,02	12,79	21,42	43,03	64,63	86,23
16QAM1	2×2	6,69	17,06	28,58	57,40	86,18	114,98
64QAM$^{1/2}$	2×2	5,02	12,79	21,43	43,03	64,63	86,23
64QAM$^{3/4}$	2×2	7,53	19,19	32,15	64,55	96,95	129,35
64QAM$^{9/10}$	2×2	9,03	23,03	38,58	77,46	116,34	155,22
64QAM1	2×2	10,04	25,59	42,87	86,07	129,27	172,47
64QAM1	4×2	19,09	48,47	81,11	162,71	244,31	325,91

Таблица 1.2 — Скорость передачи информации вверх при частотном дуплексе с нормальном CP, Мбит/с [16]

Эффективный MCs	MIMO	1.4 МГц	3 МГц	5 МГц	10 МГц	15 МГц	20 МГц
QPSK$^{1/2}$	Single	0,72	2,02	3,46	7,06	10,66	14,26
16QAM$^{1/2}$	Single	1,14	4,03	6,91	14,11	21,31	28,51
16QAM$^{3/4}$	Single	2,16	6,05	10,37	21,17	31,97	42,77
16QAM $^{9/10}$	Single	2,60	7,26	12,44	25,40	38,36	51,32
16QAM1	Single	2,88	8,06	13,83	28,22	42,62	57,02
64QAM$^{1/2}$	Single	2,16	6,05	10,37	21,17	31,97	42,77
64QAM$^{3/4}$	Single	3,24	9,10	15,56	31,76	47,96	64,38
64QAM$^{9/10}$	Single	3,88	10,89	18,67	38,11	57,54	77,25
64QAM1	Single	4,32	12,10	20,74	42,34	63,94	85,84

1.3. Процесс планирования радиоресурсов

Планирование является частью важнейших функций сетей LTE и играет жизненно важную роль, поскольку планирование отвечает за эффективное распределение радиоресурсов. Для достижения требуемых целей, блок управления радиоресурсами LTE RRM (Radio Resource Management) использует набор функций MAC (Medium Access Control) и функций физического уровня, таких как совместное использование ресурсов, отчетность по индикатору качества канала CQI (Channel Quality Indicator), адаптация канала связи с помощью адаптивной модуляции и кодирования AMC (Adaptive Modulation and Coding) и гибридного автоматического запроса повторной передачи HARQ (Hybrid Automatic Retransmission Request) [55]. Чем эффективнее будут использоваться радиоресурсы, тем лучше будут достигнуты целевые показатели производительности системы и удовлетворяются потребности пользователей в соответствии с конкретными требованиями к качеству обслуживания QoS. Параметры качества обслуживания QoS состоят из идентификатора класса QoS QCI (QoS Class Identifier) и распределения и удерживания приоритета ARP (Allocation and Retention Priority). QCI — это скалярное стандартизированное значение, которое используется для доступа к параметрам, управляющих обработкой пересылкой пакетов в радиоканале. Таким образом, каждый QCI характеризуется уровнем приоритета, бюджетом задержки пакетов и приемлемым коэффициентом потери пакетов. Консорциум 3GPP определил в ходе разработки спецификации LTE несколько классов QoS-услуг через QCIs [2]. Стандартизированные QCIs и их характеристики приведены в таблице 1.3. В зависимости от требования к качеству обслуживания QoS, выделенные однонаправленные радиоканалы могут быть классифицированы как однонаправленные радиоканалы с гарантированной скоростью передачи GBR (Guaranteed Bit-Rate) или с негарантированной скоростью передачи данных N-GBR (Non-Guaranteed Bit-Rate) [1]. GBR радиоканалы имеют гарантированную скорость передачи в течении времени установленного сеанса связи. Радиоканалы GBR используются для передачи данных в режиме реального времени, таких как видеозвонки или потоковое видеокамеры. В отличие от однонаправленных радиоканалов с гарантированной скоростью передачи GBR, N-GBR радиоканалы не имеют гарантированную битовую скорость. Другими словами, никакие ресурсы (полоса пропускания, скорость передачи) постоянно не выделяются пользователям. N-GBR радиоканалы используются для приложений не чувствительных к задержкам, таких как просмотр веб-страниц или передача файлов.

Таблица 1.3 — Стандартизированные идентификаторы класса обслуживания QoS для LTE

QCI	Тип канала	Приоритет	Допустимые задержки, мс	Допустимые потери	Пример приложений
1	GBR	2	100	10^{-2}	Телефония в реальном времени
2		4	150	10^{-3}	Видеотелефония, видео в реальном времени
3		5	300	10^{-6}	Нетрадиционное видео с буферизацией
4		3	50	10^{-3}	Игры в реальном времени
5	Non-GBR	1	100	10^{-6}	IMS-сигнализация
6		7	100	10^{-3}	Аудио, видео в реальном времени, интерактивные игры
7		6	300	10^{-6}	Видео с буферизацией
8		8	300	10^{-6}	TCP/IP,чат, FTP, передача файлы P2P
9		9	300	10^{-6}	

Функция планирования выполняется в eNodeB, в котором решаются какие ресурсы будут выделены абонентским терминалам в течение одного TTI (Time Transmission Interval), как показано на рисунке 1.3. Ресурсы выражаются в ресурсных блоках RB или физических ресурсных блоках PRB (Physicals Resources Blocks). Ресурсные блоки RB охватывают один слот во временной области и один подканал в частотной области и соответствуют наименьшему единичному ресурсу, который может быть назначен одному пользователю [63]. Физические ресурсные блоки PRB охватывают один TTI во временной области и один подканал в частотной области. Основная задача планировщика радиоресурсов состоит в выделении части спектра, разделяемой между пользователями для обоих нисходящей и восходящей линий связи. Другими словами, конечная цель его функции, как правило, состоит в том, чтобы оправдать ожидание многих пользователей системы связи, принимая во внимание несколько параметров, таких как качество радиосвязи,

требования QoS, приоритеты обслуживания и так далее. Процесс планирования обычно основан на сравнении метрик для каждого RB. Это означает, что eNodeB принимает решения о планировании после сравнения метрик RB всех абонентских терминалов. Точнее, *n*-й RB системы может быть выделен пользователю *z* тогда и только тогда, когда метрика является самой высокой среди других метрик. Однако процесс планирования не стандартизирован, поскольку он в значительной степени является внутренним по отношению к eNodeB, что позволяет разрабатывать передовые алгоритмы и оптимизировать их для конкретных сценариев.

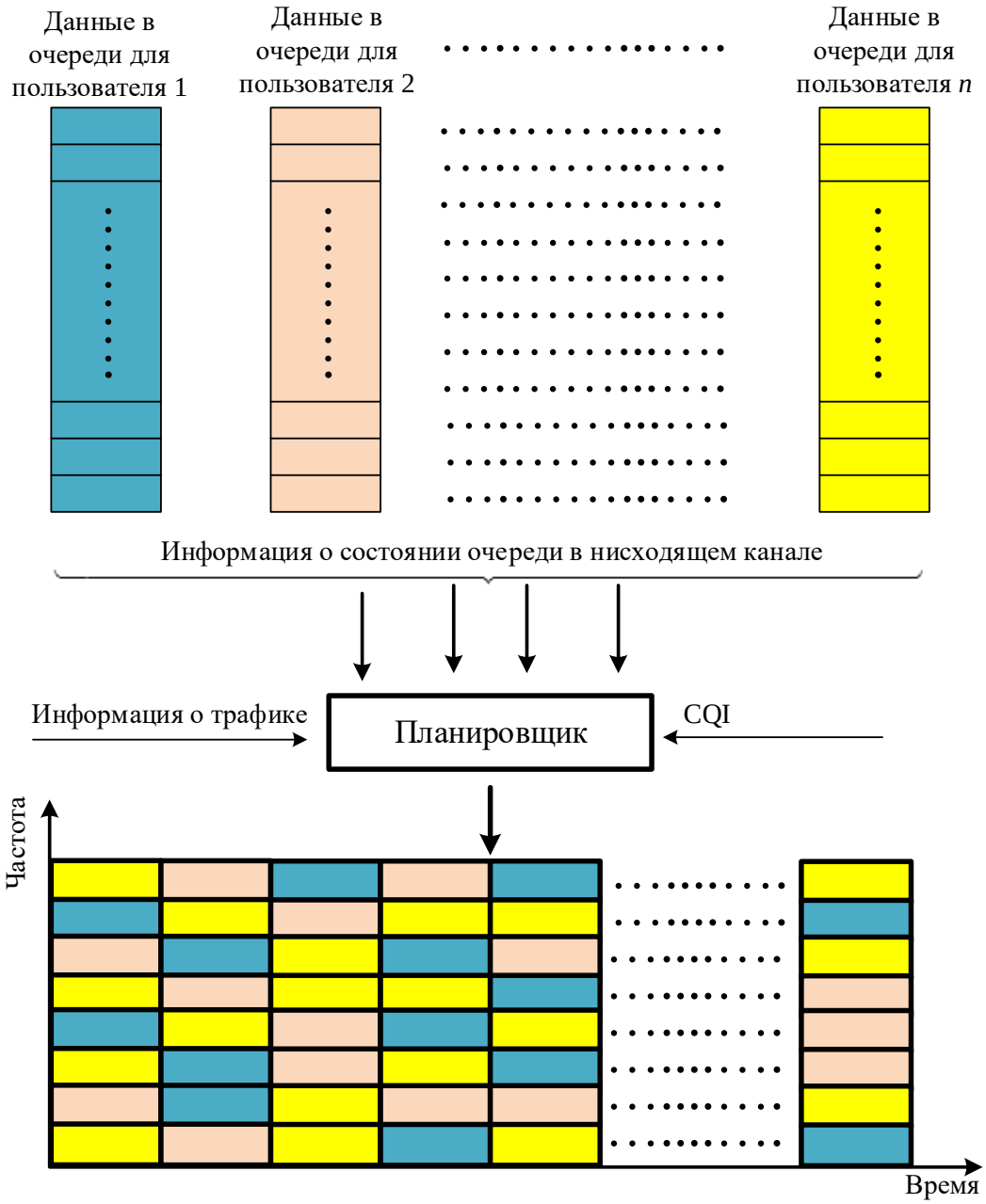


Рисунок. 1.3 — Общий вид планирования радиоресурсов в нисходящей линии связи

$$m_{z,n} = \max_{i} \left\{ m_{i,n} \right\} \quad (1.1)$$

где $m_{z,n}$ – метрика n-го ресурсного блока, выделяемого z-му пользователю.

Процесс планирования в нисходящей линии связи. В направлении нисходящей линии связи eNodeB распределяет RB непосредственно к потокам, а не к абонентским устройствам. Например, у абонентского терминала, запускающего разные приложения, будет приниматься решение о планировании для каждого из своих потоков вместо одного решения о планировании для всех потоков. Во время процесса планирования eNodeB выполняет список операций, которые повторяются и обновляются в целом каждый TTI [8]. Сначала eNodeB явно осведомлен о количестве данных, ожидающих в своем буфере, затем он подготавливает список потоков, которые могли бы быть запланированы и выделены RB в текущем TTI. Затем разные абонентские терминалы сообщают о своем CQI. Обратные связи CQI позволяют eNodeB, во-первых, оценить качество абонентского канала связи и иметь представление о скорости передачи данных, которая будет поддерживаться в радиоканале нисходящей линии связи. Во-вторых, обратные связи CQI будут использоваться в качестве входных параметров метрики планирования. На втором этапе, после обработки обратных связей CQI, eNodeB вычисляет для каждого RB, принадлежащего TTI, метрики планирования каждого потока в соответствии с определенной политикой планирования. Затем каждый RB выделяется потоку, имеющему наивысшую метрику на соответствующем RB, согласно уравнению (1.1). После этого eNodeB рассчитывает для каждого запланированного потока объем данных, которые будут переданы. Для этого используется модуль AMC, выбирающий соответствующую MCS и пытающийся максимизировать поддерживаемую пропускную способность с заданным коэффициентом блочных ошибок BLER (Block Error Rate). На третьем этапе, eNodeB передает решение о планировании абонентским терминалам через канал PDCCH (Physical Downlink Control Channel). Канал PDCCH несет DCIs (Downlink Control Information) информацию, содержащую все сведения, необходимые для того, чтобы абонентские терминалы могли идентифицировать свои ресурсы в канале PDSCH (Physical Downlink Shared Channel) и декодировать их. Наконец, каждый абонентский терминал считывает полезную нагрузку PDCCH и, если ресурсы были выделены, он обращается к части общего канала, содержащей его данные. Общий вид этого процесса показан на рисунке 1.4.

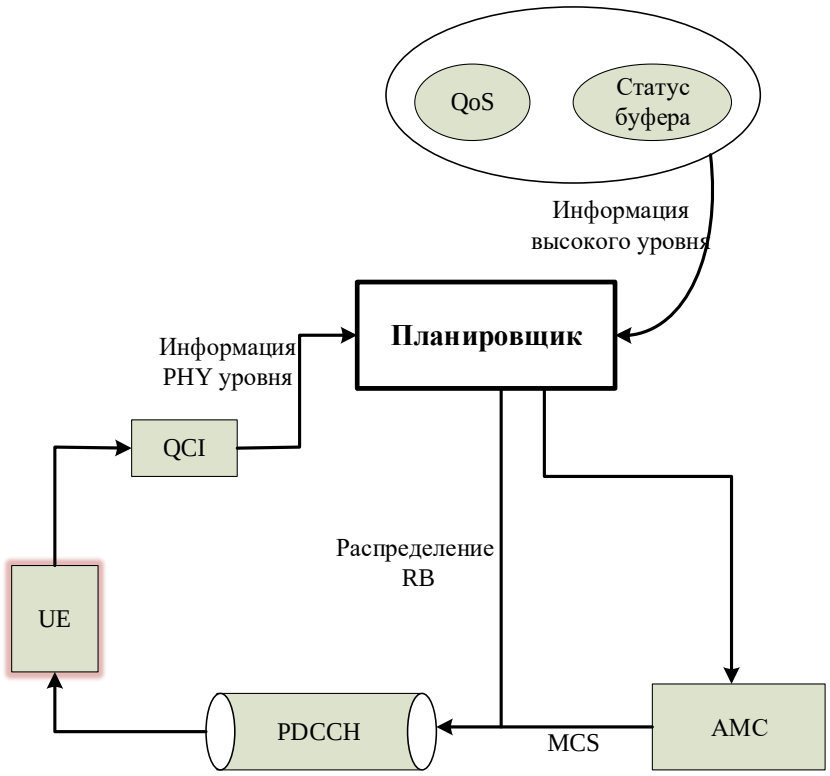


Рисунок.1 .4 — Общая модель планировщика пакетов нисходящей линии связи [53]

Процесс планирования в восходящей линии связи. Использование SC-FDMA в восходящей линии связи накладывает ряд ограничений, которые необходимо учитывать. В отличие от нисходящей линии связи, в восходящей линии связи распределение ресурсов выполняется для каждого абонентского терминала, даже если каждый абонентский терминал имеет несколько потоков [18]. Другими словами, абонентский терминал, запускающий различные приложения, получит единственное решение о планировании для всех потоков в рассматриваемом TTI. Кроме того, RB, выделенные абонентским терминалом должны быть смежными. eNodeB в начале процесса планирования не имеет информации об объеме абонентских данных. Абонентские терминалы обязаны информировать eNodeB о количестве буферизованных данных, ожидающих передачу, и их приоритете в начале процесса, поскольку eNodeB не имеет информации об объеме данных абонентских терминалов. Поэтому каждый абонентский терминал, имеющий данные в своем буфере, передает запрос планирования SR (Scheduling Request) через канал PUCCH (Physical Uplink Control Channel), позволяющий eNodeB узнать, что данные ожидают в буфере абонентского терминала. Однако SR является просто сигналом, посылаемым для предупреждения eNodeB, и не содержит никакой информации об объём данных, ожидающих передачу. Таким образом, eNodeB отправляет минимальное разрешение доступа обратно абонентским терминалам через канал

PDCCH. Как и в нисходящем канале, канал PDCCH несет DCIs информации, которые содержат всю информацию, необходимую для того, чтобы абонентские терминалы могли идентифицировать ресурсы в PUSCH (Physical Uplink Shared Channel) и использовать их для передачи. После получения минимального разрешения доступа, абонентский терминал осуществляет свою первую передачу потоков данных по каналу PUSCH, содержащих отчет о состоянии буфера (BSR– Buffer Status Report). BSR содержит информацию об объеме данных, ожидающих в буфере абонентского терминала. Процесс планирования затем выполняется на основе BSR. BSR поможет eNodeB принимать более точные решения о количестве радиоресурсов, которые будут предоставлены в следующих TTIs [1]. Общий вид этого процесса показан на рисунке 1.5.

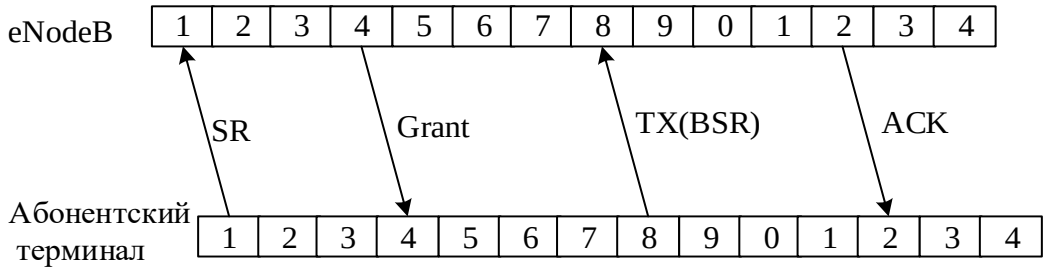
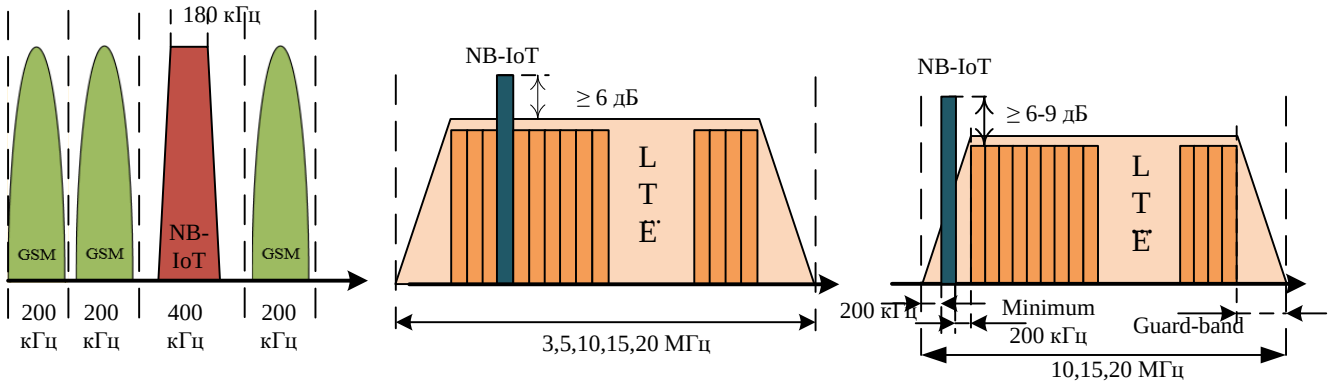


Рисунок 1.5 — Взаимодействие между eNodeB и абонентским терминалом в восходящей линии связи [53].

1.4. Анализ способов применений технологии узкополосной передачи данных NB-IoT в сетях LTE

Технология LTE предлагает множество категорий пользовательского оборудования с различными скоростями передачи данных, производительностью и стоимостью. Эта гибкость позволяет LTE работать со всем спектром приложений IoT, от приложений с высокой пропускной способностью и высокой скоростью передачи данных, таких как игры и мобильные вычисления, до приложений с низким энергопотреблением и низкой скоростью передачи данных, таких как интеллектуальный счётчик и датчики телеметрии. Большинство устройств IoT используются в приложениях с низкой скоростью передачи данных. IoT технологии могут развертываться в существующих сетях LTE. Технология узкополосного интернета вещей NB-IoT может быть развернута как внутри основной полосы за счет использования свободных ресурсных блоков LTE (внутриполосные —In band), так и вне ее, между соседними несущими LTE (в защитной полосе — Guard Band) или в свободном участке спектра GSM (Автономный — Stand Alone), как показан на рисунке 1.6.



А) Автономный Б) Внутриполосный С) На защитной полосе

Рисунок 1.6 — Варианты развертывания узкополосного интернета вещей

Общее требования к NB-IoT технологии:

- Низкое энергопотребление: NB-IoT использует режимы энергосбережения PSM (Power Saving Mode) и расширенный прерывистый прием eDRX (extended Discontinuous Reception), обеспечивающие максимальное время автономной работы батареи, теоретически до 10 лет [35].
- Низкая пропускная способность канала: Ширина полосы канала составляет 200 кГц (180 кГц), что делает её пригодным для реферминга канала GSM, поскольку позволяет заменить один канал GSM / GPRS на канал NB-IoT. NB-IoT работает на лицензированных диапазонах с высокой гибкостью [58].
- Низкая стоимость проектирования: возможность развертывания инфраструктуры сетей NB-IoT поверх уже существующей инфраструктуры сетей LTE/GSM или повторного использования существующих диапазонов GSM, позволяя операторам мобильной сети легко ее развертывать.
- Низкая стоимость для абонентского устройства: устройства NB-IoT просты в реализации. Для достижения этой цели, были реализованы некоторые функции, включающие в себя упрощенный процесс обработки в основной полосе, низкий объем памяти и уменьшение числа радиочастотных компонентов. В связи с этим ширина полосы устанавливается равной 180 кГц с понижением требования к временной синхронизации.
- Поддержка доставки IP пакетов и не-IP пакетов: NB-IoT поддерживает как доставку пакетов IP, так и доставку не-IP пакетов NIDD (Non-IP Data Delivery). NIDD был улучшен в нескольких аспектах, поскольку услуга SMS может также использоваться для доставки данных без использования протокола IP.

- Расширение покрытия: Если Максимальные потери соединения MCL (Maximum Coupling Loss) составляют 164 дБ, NB-IoT обеспечивает дополнительный бюджет канала в 20 дБ, что позволяет увеличить зону покрытия примерно в десять раз (на 20 дБ лучше по сравнению с GPRS) [81].
- Поддержка массового подключения узкополосных устройств: благодаря своей низкой полосе пропускания и расширенному покрытию, технология NB-IoT разработана с учетом требований к подключению массовых приложений и устройств MTC (Machine Type Communication).
- Улучшенная техника позиционирования: в релизе 14 консорциума 3GPP представлен усовершенствованный метод определения местоположения OTDOA (Observed Time Difference Of Arrival) для NB-IoT с целью улучшения измерения положения идентификатора соты CID (cell identity) конкретного абонентского терминала.
- Услуги Многоадресной Рассылки: Основной целью этого механизма является оптимизация ресурсов, а также уменьшение задержки передачи путем одновременной адресации данных группы абонентских терминалов, а не отправки их несколько раз на отдельные устройства.
- Поддержка малых сот: для дальнейшего улучшения как емкости, так и зоны покрытия в релизе 15, NB-IoT поддерживает развертывание небольших сот. Мощность нисходящей линии связи, которая будет повторно использоваться для небольших сот NB-IoT, указана в [5]. Абонентский терминал NB-IoT не может передавать мощность выше сконфигурированной максимальной мощности [27].
- Возможность поддержки дуплекса с временным разделением TDD (Time Division Duplex).

Устройства NB-IoT разработаны для работы при более низких уровнях сигнала, с низкой скоростью передачи. Технология NB-IoT позволяет передавать только короткие сообщения примерно 20-256 байт в сообщении, посылаемом несколько раз за день [92]. Стандарт NB-IoT принимает ту же структуру кадра, что и LTE [110]. Один суперкадр, состоит из 1024 кадров, которые в свою очередь содержат 10 субкадров каждый. В одном субкадре содержатся два временных интервала длительностью 0,5 мс каждый во временной области, следовательно, временный интервал передачи TTI составляет 1 мс. В частотной области, используется либо 12 поднесущих в каждом временном интервале длительностью 0,5 мс при частоте разноса 15 кГц в нисходящей и восходящей линиях связи, либо 48 поднесущих с длительностью интервала 2 мс при частоте разноса 3,75 кГц в восходящей линии связи. Стандарт NB-IoT использует тот же стек протоколов, что и LTE. Однако некоторые конструктивные изменения как на физическом уровне

PHY, так и на уровне MAC были введены для поддержки массовых соединений на большие расстояния с MCL до 20 дБ по сравнению с традиционными технологиями, такими как LTE, GSM и GPRS [79]. На физическом уровне NB-IoT применяются те же технологии, что и в LTE. Технологии OFDMA и SC-FDMA используются в стандарте NB-IoT для формирования сигналов при передаче информации в нисходящей линии связи и восходящей линии связи соответственно. Однако блок планирования ресурсов в NB-IoT является ресурсным элементом RE (Resource Element) или тоном вместо PRB, позволяющим устройствам NB-IoT передавать радиосигнал на одной поднесущей на частоте 15 кГц, что дает возможность обслуживать несколько устройств в полосе частот 180 кГц. Спектральная плотность сигнала увеличивается с использованием поднесущей на более низкой частоте 3,75 кГц и, следовательно, отношение сигнал/шум увеличивается, что очень важно для устройств, которые имеют гораздо менее мощные передатчики, чем базовая станция LTE. В сетях NB-IoT поддерживается многотональной передачи (англ. Multi-Tone Transmission)**.** Чтобы подключить массовые устройства на одну базовую станцию, в NB-IoT распределяются единицы ресурсов RU (Resource Units) между несколькими пользовательскими устройствами, в отличие от LTE, где весь ресурсный блок выделяется одному пользователю в восходящей линии связи [27]. RU (см. рисунок 1.7) — это очередной более крупный элемент, из которого образуются транспортные блоки TB (Transport Block), назначаемые пользователю. Передача информации по восходящей линии связи использует схему мультиплексирования SC-FDMA. Разнос частоты между поднесущими составляет 3,75 кГц или

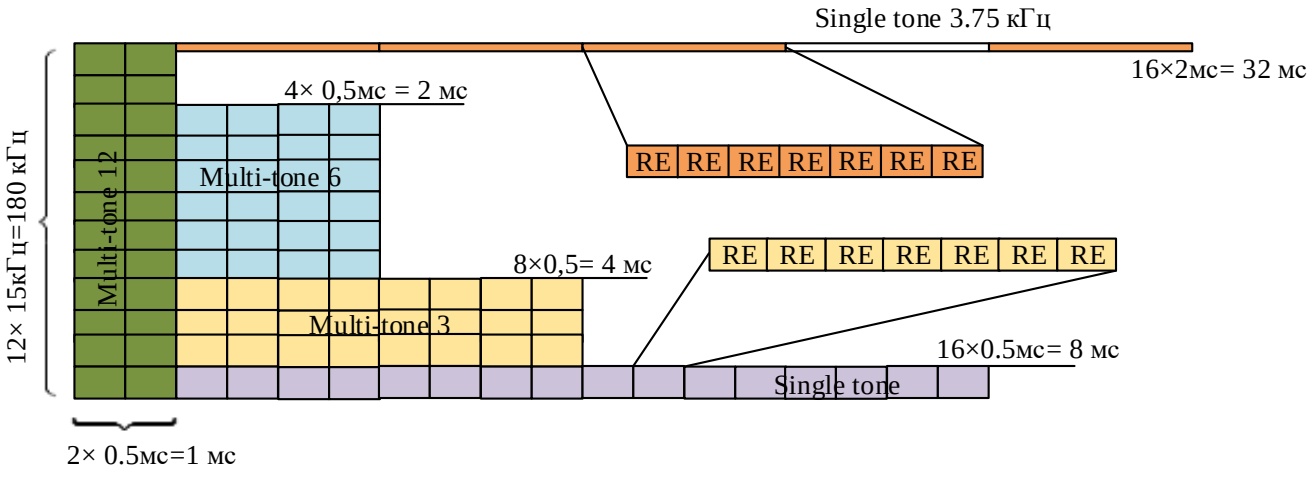


Рисунок 1.7 – Единичный ресурс RU [86]

15 кГц. По нисходящей линии связи, NB-IoT использует 15 кГц со схемой OFDMA, как и в LTE. С разносом 15 кГц, NB-IoT использует режим single-tone (8 мс) или режим multi-tone (3 тона, 6 тонов

и 12 тонов) с длительностью 4 мс, 2 мс и 1 мс, соответственно. С использованием 3,75 кГц поддерживается только режим single-tone с 48 подкадрами длительностью 32 мс.

1.5. Сервисы оператора систем видеонаблюдений

Устройства интернета вещей являются конечными точками системы, которые могут подключаться к сетевому домену либо напрямую с помощью встроенного модуля идентификации абонента SIM (Subscriber Identity Module), либо через шлюз, который выполняет функции агрегирования, мультиплексирования / демультиплексирования поток данных и отправляют их в сетевой домен LTE. Сервисы, предоставляемые конечными устройствами представлены в таблице.1.3. Видеокамеры и IP-телефония являются источниками трафика реального времени чувствительного к задержке. Интеллектуальные датчики интернета вещей являются источниками данных эластичного трафика. Такие устройства используют технологии интернета вещей для сбора данных. Технология NB-IoT является наиболее перспективным кандидатом для сбора больших данных в сфере IoT благодаря своим особым характеристикам, таким как дальность действия, которая может достигать 10 км, высокое энергоэффективное потребление и недорогая конструкция радиосвязи. Технология NB-IoT обладает также огромным потенциалом для работы в условиях с ограниченными ресурсами. При этом, технология NB-IoT позволяет продлить срок службы батареи до 10 лет при передаче пакета данных в среднем 200 байт в день. Максимальный размер полезной нагрузки для каждого NB-IoT сообщения составляет 2536 байт. В таблице 1.4 представлены характеристики некоторых технологий LPWAN. Более того, наряду с внедрением недорогих интеллектуальных счетчиков с низким трафиком, в системах интернета вещей наблюдается растущее влияние мультимедийных больших данных [74] особенно те, которые собираются массовыми системами видеонаблюдения, развернутыми в целях безопасности [37, 46]. Интеллектуальные счетчики и камеры видеонаблюдения могут быть развернуты на большой территории и размещены в местах, где проводное подключение невозможно по техническим или экономическим причинам. С целью повышения интеллекта и надежности в принятие решений и, следовательно, раскрытия потенциала интернета вещей IoT [46, 82], информации о обнаружении вторжения или пожара, передаваемые группой интеллектуальных счетчиков, могут быть немедленно проверены по данным видеомониторинга. Система видеонаблюдения началась в 1970-х годах с аналоговыми системами телевидения замкнутого контура CCTV (Closed-Circuit Television). Системы CCTV были построены из камер низкого качества изображения,

мультиплексоров, видеомагнитофонов и мониторов. Потребовалось большое количество коаксиальных кабелей для передачи и хранения видео на видеокассетах [49].

Таблица 1.4 — Характеристики некоторых технологий LPWAN

Технология	LTE cat 1	LTE-M		NB-IoT	
		LTE cat M1	LTE cat M2	LTE cat NB1	LTE cat NB1
3GPP release	Release 8	Release 13	Release 14	Release 13	Release 14
Пиковая скорость DL	10 Мбит/с	1 Мбит/с	~4 Мбит/с	26 кбит/с	127 кбит/с
Пиковая скорость UL	5 Мбит/с	1 Мбит/с	~7 Мбит/с	66 кбит/с (multi-tone) 16.9 кбит/с (Single-tone)	159 кбит/с
Задержка	50-100мс	10-15 мс		1,6-10 с	
Система MIMO	2	1	1	1	1
Мод Duplex	Full duplex	Full or Half Duplex	Full or Half Duplex	Half Duplex	Half Duplex
Полоса пропускания устройства	1,4-20 МГц	1,4 МГц	5 МГц	180 кГц	180 кГц
Максимальное расстояние/покрытие		156 дБ		164 дБ	

В новом тысячелетии современные видеокамеры с сетевыми интерфейсами (IP-камеры) стали широко доступными с высоким качеством изображения и с новыми совершенными функциональными возможностями. Объем мультимедийных данных, собираемых IP-видеокамерами во много раз превышает объем данных классических видеокамер. Передача большого объема мультимедийных данных в беспроводных сетях является серьезным препятствием, которое необходимо преодолеть, т.к. пропускная способность беспроводной связи ограничена действиями регулятора и физическими возможностями передачи радиосигналов. Использование методов уменьшения объема передаваемых видеоданных стало неизбежно для снижения требований к пропускной способности. Например, избыточные данные могут быть удалены путем применения встроенных алгоритмов локальной обработки LPA (Local Processing

Algorithms) информации или путем использование распределенного исходного кодирования DSC (Distributed Source Coding) [71].

Распределенное исходное кодирование. Размер видеопотока определяется размером сжатого видео, пропускной способностью и задержкой. К широко используемым форматам видеокодирования относятся: H.264/AVC [65], H.265/HEVC [103], MPEG-4 [75], AVS2 [66], VP6 [85], VP9 [77], VC-1 [68], и AV1 [67]. Являясь хорошим преемником H.264/AVC, H.265/HEVC позволяет повысить эффективность кодирования на 50% с сохранением качества видео. H.265 / HEVC удобен для кодирования в реальном времени и поддерживает видео HD, UHD и 8K UHD. Для адаптации к различным сетевым средам и различным требованиям пользователей в системах видеопередачи и хранения были предложены SVC и SHVC, являющиеся масштабируемыми расширениями H.264 / AVC и H.265 / HEVC [44, 94]. В средах передачи с потерями, скорость передачи данных и форматы видеокодирования могут быть скорректированы для адаптации к различным возможностям терминала или условиям сети. Кроме того, видеотрафик оказал огромное влияние на мобильные сети. Мобильные сети должны обеспечивать более высокую скорость передачи и меньшую задержку в сети. Технология адаптивного битрейта ABR (Adaptive Bitrate) [93] используется для решения выше перечисленных задач. Во-первых, в мобильной сети видео кодируется в различные версии потоковой передачи битовой скорости (версия потоковой передачи с битрейтом). Во-вторых, каждая потоковая версия разделена на несколько сегментов. Следовательно, в соответствии с возможностями терминала и условиями сети подходящая потоковая версия будет предоставляться пользователю динамически. Преимущество ABR заключается в том, что вероятность прерывистого (англ. probability of choppy) видео может быть уменьшена, а QoE пользователя может быть увеличено. Закодированные видеопотоки собираются в контейнер «битовый поток», и битовый поток передается с использованием транспортного протокола. К популярным протоколам видеопотоковой передачи включаются в себя MEPG-DASH [96], Apple HTTP Live Streaming (HLS) [32], Adobe HTTP Dynamic Streaming (HDS) [111] и Microsoft Smooth Streaming [28]. В таблице 1.5 представлены рекомендуемые конфигурации для более высоких и низких битрейтов видео при использовании видеокодек H264.

Передача HD-видеоданных требует высокую пропускную способность радиоканала чтобы поддерживать требуемое качество связи. Предполагается что при хорошем условии радиоканала, видеоданные передаются в реальном времени непосредственно с IP-камеры подключённой на прямую к базовой станции LTE (см. рисунок 1.8), а затем при ухудшении качества связи регулируется качество видео (разрешение видео, частота кадров, битрейт, и т.д.) [33, 93] уменьшая

при этом потребность к пропускной способности. В более худших условиях радиоканала, камера начнет записывать видеоданные на его внутреннюю память (на SD карту, установленную в камере

Таблица 1.5 — Битрейты кодека H264 для потоковой передачи видео [59].

Качество	Разрешение	Видео битрейт	Кадры/ с	Интервал опорного кадра, с
Низкое	480x270	400 кбит/с	25 / 30	1 с
Среднее	640x360	800 - 1200 кбит/с	25 / 30	1 с
Высокое	960x540 / 854x480	1200 - 1500 кбит/с	25 / 30	1 с
HD 720	1280x720	1500 - 4000 кбит/с	25 / 30	1 с
HD 1080	1920x1080	4000-8000 кбит/с	25 / 30	1 с
4K	3840x2160	8000-14000 кбит/с	25 / 30	1 с

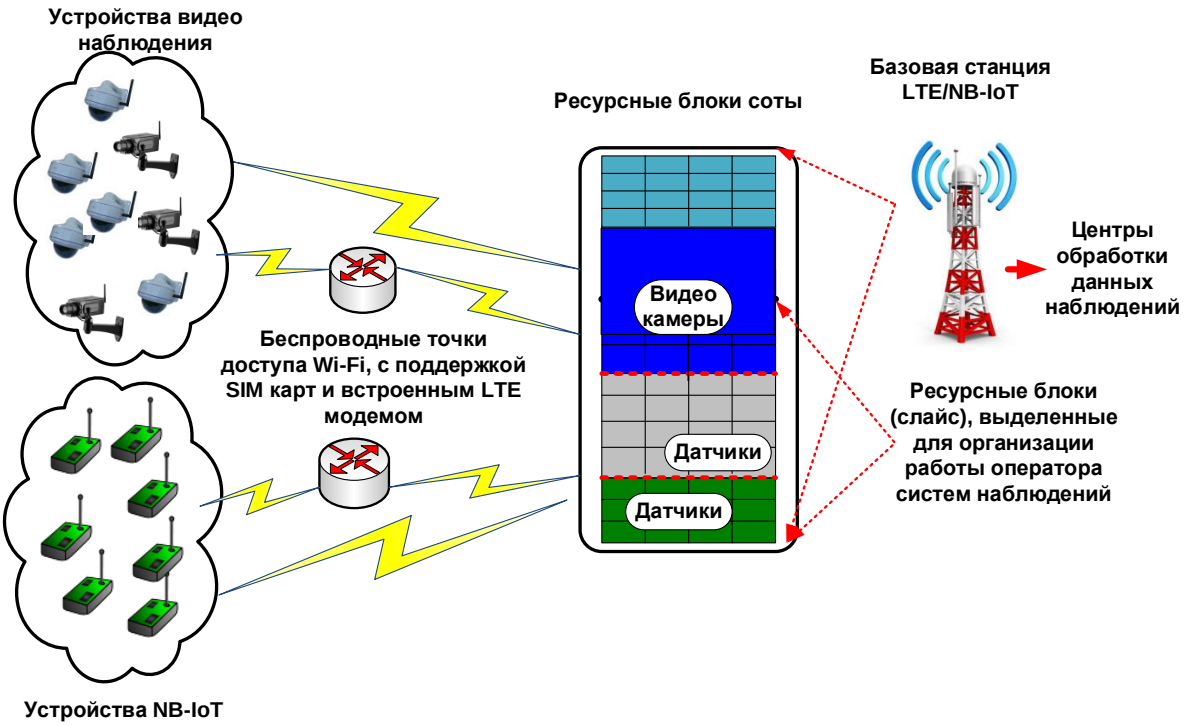


Рисунок 1.8 — Способы подключения IoT устройств к базовой станции LTE

наблюдения) и затем видеоданные передаются в качестве эластичного трафика. Камера сначала записывает видео на локальный видеорегистратор т.е. на карту памяти SD подключена к камере, а затем внешний видеорегистратор (сервер) может подключаться к карте памяти SD и создавать резервные копии видео. Это создает избыточную систему хранения видео, в которой есть две копии видео. Использование второго метода пригодно в случае нехватки пропускной способности канала связи для передачи видео в реальном времени. Видеоданные не теряются и сохраняются на SD карта памяти и транслируются при хорошей связи или по запросу что повышает отказоустойчивость.

Используемые видеокамеры в нашем исследовании имеют встроенные датчики и передают в сети кадры при обнаружении вторжения вместо потоковой передачи необработанных видеоданных и уменьшая существенно передаваемые видеоданные в сети LTE. Однако, когда допускается более локальная обработка, датчики камеры могут выполнять вычитание фона для обнаружения движущегося объекта и только если обнаруженный объект превышает определённое пороговое значение, то видеокамера передает в сети часть изображения, содержащего обнаруженный объект. Такой метод существенно уменьшает объем видеоданных передаваемых в сетях LTE. Таблица 1.6 содержит технические характеристики IP-видеокамер для разных производителей.

Таблица 1.6 — Технические характеристики IP камер [83, 88]

Характеристики	Устройства		
	Optimus Basic IP-P012.1 (4x) DWG	Reolink Argus Go	Millenium 433G PTZ
Максимальное разрешение видео	1920×1080	1920×1080	2560x1920
Частота кадров при максимальном разрешении	30 кадров в секунду	15 кадров в секунду	15 кадров в секунду
Количество мегапикселей	2	2	5
Интеллектуальные функции	Обнаружение движения, Обнаружение звука	датчик освещённости	Обнаружение движения, датчик освещённости
Беспроводная связь	3G, 4G LTE, GSM	3G, 4G LTE, GSM	3G, 4G LTE
Форматы сжатия	H.264/H.265/H.264+/H.265+/MJPE	H.264	H.265, H.264, MJPEG, AVI
Поддержка SD карты	до 256 Гб	до 128 Гб	до 128 Гб

Существуют несколько типов камер с встроенными датчиками в зависимости от функций датчика: IP-камеры с поддержкой детектора движения, IP-камеры с поддержкой зон маскирования, IP-камеры с детектором пересечения виртуальной линии, IP-камеры с детектором вторжения в зону, IP-камеры со встроенной аудио аналитикой, IP-камеры с возможностью подсчета людей, и т.д.

1.6. Параметры модели трафика интернета вещей

IoT-устройства с непериодической отчетностью. Устройства интернет вещей отслеживают физические явления и передают отчеты о событиях в аналитические центры для анализа и принятия решения. К таким событиям относятся уведомление дымовой сигнализации, уведомление об отключении питания, уведомление о несанкционированном вмешательстве, уведомление об обнаружении вторжений, уведомление о движении автомобилей и транспортных средств, уведомление об обнаружении деградации окружающей среды, и т.д. В целях анализа задержки предполагается что сообщения, передаваемые устройствами такого типа, содержат полезную нагрузку в 20 байт во восходящей линии связи и требуемая задержка составляет 10 секунд в режиме реального времени [56]. Предполагается что размер пакета в сообщении ACK в нисходящей линии связи равен 20 байт [56, 57].

IoT-устройства с периодической отчетностью. Отчет о наблюдаемых событиях в некоторых сферах жизни человека такие как умное сельское хозяйство, умная среда, освещение города, торговые автоматы, отслеживание используемого объема газа и воды, потребление электроэнергии, и т.д. осуществляется в определенных интервалах времени. В этом случае, принято использовать модель трафика с периодическими отчетами при моделировании систем связи для анализа пропускной способности. В таблице 1.7 приведены характеристики сессий связи IoT устройств в соответствии с параметрами, установленными 3GPP GERAN [3].

Таблица 1.7 — Характеристики модели трафика в восходящей линии связи IoT устройств

Приложение	Количество сообщений	Размер сообщения	Загрузка/ день	Модель трафика 3GPP
Отслеживание домашних животных VIP	2 сооб/час	50 байт	2400байт	Модель 2
Медицинское обслуживание	8 сооб/ день	100 байт	800 байт	Модель 1
Агростационарное отслеживание/мониторинг	4 сооб/ день	100 байт	400 байт	Модель 4

Продолжение таблицы 1.7

Учет воды / газа	8 сооб/ день	200 байт	1600 байт	Модель 1
Охрана зданий — сигнализация и актуатор	5 сооб/ день	50 байт	2500 байт	Модель 4
Детектор дыма для дома/предприятия	2 сооб/день	20 байт	40 байт	Модель 3

1.7. Анализ механизмов нарезки сети Network slicing с учетом гарантий обслуживания гетерогенного трафика

1.7.1. Основные понятия и термины

Технология интернета вещей нашла широкое применение во всем мире. Рост числа простых интеллектуальных мобильных устройств, предназначенных для передачи небольших пакетов с низкими скоростями приводит к увеличению мультимедийных данных, традиционно создаваемых массовыми системами видеонаблюдения, развернутые в целях обеспечения безопасности. Существующие решения межмашинного взаимодействия на основе стандарта 4G и 4G + не могут полностью поддерживать требование к масштабируемости и надежности массовых видеонаблюдения из-за недостаточной емкости для потоковой передачи огромных объемов данных реального времени и эластичных данных. Поэтому появляется необходимость разработать эффективные методы распределения ограниченных ресурсов сетей 4G и 4G +. Комбинация двух или более гетерогенных потоков больших данных является одной из ключевых особенностей мобильных сетей 5G, которая повышает интеллект и надежность принятия решений и, следовательно, раскрывает весь потенциал интернета вещей.

Как отмечалось выше, некоторые приложения, такие как видео сверхвысокой четкости UHD (ultra-high definition), дополненная реальность и виртуальная реальность требуют высокоскоростные каналы связи с высокой информационной ёмкостью, в то время как межмашинное взаимодействие требует сверхнизкие задержки и сверхнадежные услуги. Распределение ограниченного ресурса мобильной сети между устройствами с высокой скоростью передачи данных и устройствами с низкой скоростью передачи данных приводит к перераспределению ресурса в пользу потоков с низкой скоростью передачи данных [17, 21]. Устранение недостатков, вызванных этим явлением осуществляется либо путем резервирования ресурсов для потоков данных с большими потребностями к скорости передачи данных, либо путем

разделения низкоскоростных и высокоскоростных потоков данных по отдельным слайсам. Поставщики услуг связи ищут альтернативные механизмы, которые могут успешно обеспечить управление сетевыми ресурсами более динамичным и прогнозируемым. Виртуальные операторы в сетях мобильной связи арендует определенный частотный ресурс по механизму нарезки сети (англ. Network slicing) у оператора сети, чтобы гарантировать бесперебойную и предсказуемую работу службы эксплуатации независимо от фактической нагрузки на диапазоны спектра других операторов. Механизм нарезки сети Network Slicing позволяет совместно использовать общую физическую сеть [109]. Концептуальная архитектура Network Slicing в сети стандарта LTE иллюстрирована на рисунке 1.9.

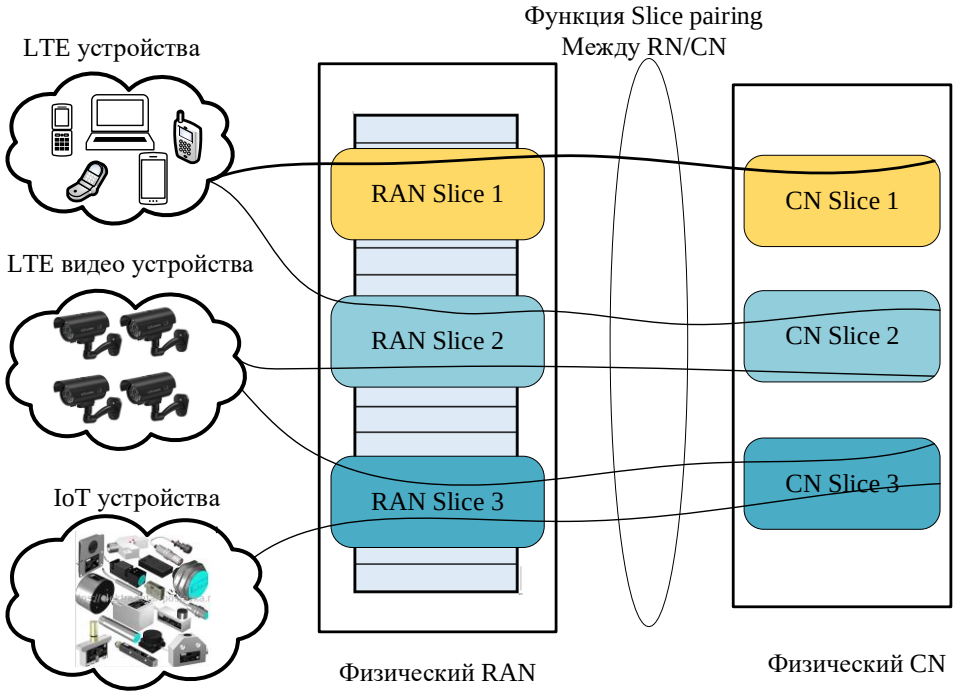


Рисунок 1.9 — Архитектура Network Slicing в сети стандарта LTE

На сегодняшний день было проведено множество исследований, направленных на улучшение моделей управления ресурсами в мобильных сетях. В некоторых из этих работ были предложены механизмы распределения ресурсов, основанные на назначении ряда физических ресурсных блоков PRB каждому запросу пользователя сотовой связи. Можно классифицировать механизм управления ресурсами на два уровня: модель управления ресурсом низкого уровня и модель управления ресурсом высокого уровня. Преимущество применения низкоуровневой модели заключается в том, что ее легко реализовать, поскольку любой запрос на выделение ресурса получает запрашиваемый ресурс в единицах ресурсов. Низкоуровневая модель обеспечивает

точность распределения ресурсов по каждому запросу в единицах ресурсов. Однако, органам управления высокого уровня (например, операторам и поставщикам услуг) трудно принять механизм управления низкого уровня, поскольку ресурсы в модели управления высокого уровня распределяются частично (например, 30% от общего объема доступных физических ресурсных блоков PRB). Будущие стандарты мобильной связи, в том числе и уже запущен стандарт беспроводной связи пятого поколения 5G будет охватывать исключительно сетевую виртуализацию), чтобы гарантировать бесперебойную и предсказуемую работу службы эксплуатации независимо от фактической нагрузки на диапазоны спектра других операторов. Кроме этого, основными факторами, которые привели к быстрому внедрению виртуализации сети являются экономичное совместное использование сетевых ресурсов и высокая степень использования ресурсов сети. С целью получения синергетических преимуществ сетевой виртуализации, наряду с разработкой эффективных сетевых архитектур, исследовательские усилия должны быть сосредоточены на разработке эффективных механизмов управления ресурсами в виртуальной сети. Виртуализированные сети нуждаются в новом механизме управления ресурсами, который обеспечивал бы точность распределения ресурсов и гарантированную изоляцию ресурсов (англ. resource isolation). Для достижения этих целей необходим новый механизм управления ресурсами, который учитывает модели распределения ресурсов как низкого, так и высокого уровня. Основная роль модели низкого уровня заключается в обеспечении распределения ресурсов на основе PRB по количеству единиц ресурсов, обеспечивая тем самым высокую точность распределения ресурсов. С другой стороны, модель высокого уровня должна обеспечивать изоляцию выделенных ресурсов. Для того чтобы обеспечить такое гибкое распределение ресурсов, главную роль играют программно-определяемая сеть SDN (Software Defined Network) и виртуализация сетевых функций NFV (Network Function Virtualization) [34, 72], позволяющие динамическую конфигурацию сети и экономическую эффективную работу сети.

Программно-определяемой сетью SDN является новая технология, в которой плоскость управления успешно отделяется от плоскости данных, что делает сеть программируемой и экономически эффективной. Программно-определяемая SDN предлагает ряд преимуществ по сравнению с обычными аппаратно-ориентированными сетями, включая политику переадресации трафика по требованию, снижение стоимости и улучшение качества обслуживания. Виртуализацией сетевых функций NFV является новая технология для будущих сетей, позволяющая одновременно распределить физическую сетевую инфраструктуру между несколькими сетями. NFV предлагает ряд преимуществ по сравнению с традиционными сетями,

таких как лучшее сетевое администрирование, программируемость и снижение затрат. SDN и NFV разделяют традиционные сети на виртуальные элементы, которые логически связаны между собой [80].

1.7.2. Основные преимущества использования механизма Network Slicing

Концептуальная архитектура нарезки сети Network slicing (см. рисунок 1.9), направленна на совместное использование общей физической инфраструктуры между несколькими виртуальными сетями с использованием тех же принципов, что и в SDN и NFV [29, 73]. В частности, существуют некоторые важные требования, которые должны быть соблюдены при применении сетевого сегментирования:

- Изоляция между слайсами (англ. Isolation among slices). Изоляция определяется как возможность ограничить влияние слайса на другие слайсы в той же сети, даже если они используют одну и ту же инфраструктуру. То есть, если в одном слайсе происходит какое-либо изменение состояния ресурсов (например, изменение нагрузки трафика), то такое изменение не должно влиять на выделенные ресурсы других слайсов.
- Настройка (англ. Customization). Управление ресурсами каждого слайса может осуществляться независимо. То есть политика контроля допуска каждого слайса к ресурсам может отличаться от других слайсов.
- Эффективное использование ресурсов. Максимальное использование ресурсов канала, позволяет увеличить пропускную способность базовой станции и эффективно использовать канала передачи.

Слайс можно рассматривать как набор потоков, принадлежащих разным конечным пользователям, использующим однотипные абонентские терминалы. В качестве примера, слайсы могут быть: все потоки, источником или получателем которых является устройства одного типа, например, потоки данных, создаваемых разными датчиками; все потоки услуг VoIP; все потоки голосового сообщения; или все потоки конечных пользователей определенного типа оператора. Слайс поддерживает потоки нескольких конечных пользователей, но в то же время конечный пользователь может участвовать в нескольких слайсах [89]. Кроме того, можно рассматривать слайс как подмножество сетевых ресурсов, выделенных провайдеру (виртуальному оператору или поставщику услуг), с полным контролем над этими ресурсами. Важным аспектом подхода к проектированию нарезки сети является предоставление клиенту инструментов настройки и

программирования. В зависимости от спецификации слайса конечный пользователь может участвовать в разных слайсах, но слайсы всегда независимы друг от друга. В контексте беспроводных сетей часто рассматриваются два сценария использования слайсов:

1. Качество обслуживания нарезки сети (англ. Quality of Service Slicing). Идея состоит в создание слайсов, чтобы предлагать различные услуги и обеспечивать некоторый тип QoS внутри слайса. Например, слайс может быть создан для обслуживания определенной группы устройств с одинаковыми требованиями (датчики, смартфоны, устройства видео наблюдения) или по типу приложения (слайс для мультимедийных услуг).
2. Слайсинг для совместного использования инфраструктуры (англ. Infrastructure Sharing Slicing). Это традиционная идея сетевой виртуализации, применяемая к беспроводному домену. Существующему провайдеру предоставляется слайс сети. У провайдера есть полный контроль над сетевой инфраструктурой и функциями в рамках слайса.

К преимуществам использования механизма нарезки сети беспроводной связи относятся гетерогенная дифференциация услуг (англ. Heterogeneous Service Differentiation), сетевое администрирование (англ. Network Management), гетерогенные технологии радиодоступа (англ. Heterogeneous Radio Access Technologies), совместное использование инфраструктуры (англ. Infrastructure Sharing), гибкость для новых услуг и бизнес-моделей (англ. Flexibility for New Services and Business Models).

Гетерогенная дифференциация услуг. В условиях, когда беспроводным сетям приходится иметь дело с широким спектром услуг и устройств, нарезка сети становится способом изолировать и одновременно выполнять разные требования. При совместном использовании ресурсов нарезка сети позволит создавать настраиваемые сервисы с точными функциями управления QoS [50]. Основная идея состоит в разделении сети на сегменты (слайсы), имеющие разные ресурсы и пропускные способности, чтобы предлагать дифференцированные услуги для разнородных вариантов использования. Более того, нарезка сети является одним из ключевых инструментов систем 5G и будущих сетей для управления неоднородными требованиями к ресурсам передачи телекоммуникационной информации. Можно определить так же слайс для конкретных приложений, которые могут потребовать настройку сетевых возможностей. Другой возможный подход заключается в том, чтобы настроить слайса для каждого типа устройств или для каждого типа требований пользователей. Сетевые нарезки обеспечивают эффективное использование

ресурсов, поскольку каждый слайс может быть настроен для конкретной услуги и в динамическом режиме по запросу.

Сетевое администрирование. Управление различными приложениями с противоречивыми требованиями к общей инфраструктуре может осуществляться через отдельные сегменты сети [70]. Нарезка сети позволяет индивидуально настраивать сети от конца до конца и определять конкретные функции для каждого случая, используя при этом одну и ту же инфраструктуру и избегая более высоких затрат. Например, нарезка сети позволяет выделить только необходимые функции и зарезервировать ресурсы на всем сеансе связи, позволяя настраивать сетевую конфигурацию для каждого случая. Нарезка сети также обеспечивает гибкость для динамического создания и уничтожения слайсов в зависимости от политики операторов с помощью NFV и SDN. Основная цель состоит в том, чтобы виртуализировать как можно больше функций, а те, которые не могут быть виртуализированы, должны быть программируемыми и настраиваемыми. Более того, в случае, когда слайсы определенны для каждого типа услуги или устройства и поскольку известно какую услугу обслуживает каждый слайс, сеть можно упростить, путем удаления ненужных функций. Например, если слайс предоставляет доступ к статическим датчикам, управление мобильностью может быть сведено к минимуму. Таким образом, сетевое управление упрощается, становится проще разработать автономное управление для каждого конкретного фрагмента сети.

Гетерогенные технологии радиодоступа. Нарезка также может помочь в управлении сетями с использованием гетерогенных технологий радиодоступа RATs (Radio Access Technologies). В настоящее время, разные технологии радиодоступа RAT работают чаще в одной сети, чтобы решить проблему нехватки спектра. Распределение ресурсов по различным технологиям может быть обработано с точки зрения сегментирования, где в зависимости от таких параметров, как пропускная способность, местоположение пользователя или затраты, каждому слайсу назначается наилучший RAT. Эффективность использования спектра также может быть улучшена с помощью нарезки сети, поскольку можно соответствовать различным требованиям к наилучшим доступным радиоресурсам [107]. Чтобы реализовать такую возможность, сеть должна включать в себя виртуализированные или программируемые беспроводные интерфейсы, а также различные беспроводные технологии, как ожидается, в будущих беспроводных сетях. Технологии будущего ведут к сосуществованию и конвергенции различных беспроводных технологий, составляющих сервис-ориентированную инфраструктуру. Нарезка сети представляется одним из

возможных решений, позволяющих обеспечить такое сосуществование за счет упрощения сетевого управления.

Совместное использование инфраструктуры. Еще одной важной мотивацией для нарезки сети является совместное использование инфраструктуры. Это похоже на концепцию дифференциации услуг, но в этом случае каждый слайс может использоваться другим оператором, предлагающим свои собственные услуги. Большинство из них предлагают аналогичные услуги голосовой связи, короткие текстовые сообщения SMS (Short Message Service) и передачи данных что и традиционные операторы. Нарезка сети облегчит управление инфраструктурой и обеспечит изоляцию между различными операторами. С другой точки зрения, идея совместного использования инфраструктуры даст операторам больше гибкости для изменения их логической сети и эффективного использования их ресурсов [106].

Гибкость для новых услуг и бизнес-моделей. С точки зрения бизнеса, сегментирование сети продвигает внедрению новых сценариев использования ресурсов без увеличения затрат благодаря возможности совместного использования инфраструктуры различными сегментами сети. Кроме того, в качестве стандартизированного программного интерфейса приложения API (Application Programming Interface) для программирования сети может быть предложено нарезке сети с использованием бизнес-модели «Все как услуга (анл. Everything as a Service XaaS)» и позволит третьим сторонам исследовать новые возможности.

1.7.3. Анализ модели распределения ресурсов сети LTE на основе концепции Network Slicing

Виртуализация проводной сети осуществляется на разных уровнях сети, таких как процессор, память, соединение портов и физический уровень связи. В отличие от виртуализации проводной сети, беспроводная сеть требует виртуализация ядра сети CN, так и радиодоступа RAN. Однако беспроводная виртуализация, по сравнению с виртуализацией проводной сети, включает в себя виртуализацию конкретного беспроводного оборудования и радиочастотного спектра. Концептуальная архитектура управления ресурсами нарезки сети (англ. Network Slicing Resource Management) для сети LTE изображена на рисунке 1.10. Данная архитектура в целом сегментирована на три уровня: уровень слайса (англ. Slice layer), уровень диспетчера контроллеров слайсов LTE LSCM (LTE Slice Controller Manager) и уровень слайсера (англ. Slicer layer). Кроме

того, данная архитектура позволяет нарезать виртуальной сети на несколько сегментов (слайсов), каждый из которых настраивается в зависимости от требований оператора.

Предположим, что в сети LTE имеются n слайсов, как показано на рисунке 1.10. Пусть каждый слайс принадлежит одному оператору и управляется его контроллером с помощью функции сопряжения срезов (англ. Slice pairing function). Контроллер отвечает за максимальное использование ресурсов слайса (всех виртуальных ресурсов). Как правило, один или несколько потоков может принадлежит одному пользователи. Также эти потоки могут принадлежать одному и тому же слайсу или разным слайсам [91]. В случае, когда потоки данных принадлежат одному и тому же слайсу необходимо контроллеру управлять ресурсами внутри слайса с целью выделения необходимых ресурсов для каждого потока. Кроме того, он должен обеспечить изоляцию потоков в слайсе. Изоляция между слайсами гарантирует предопределенный метод распределения радио ресурсов внутри каждого слайса и повышает коэффициент использования ресурсов сети и уменьшает влияние слайса на другие слайсы в той же сети. В нашем исследовании, предполагается что каждый уровень слайсера отвечает за изоляцию между слайсами (более подробно рассматривается во втором разделе). Как упоминалось в [91], изоляция ресурсов слайса может быть классифицирована на три общие категории в зависимости от группы пользователей с одним и тем же типом приложения; сквозной сети (различные сквозные потоки); ресурсов, распределенных между различными слайсами (количество выделенных ресурсов заранее определено в соответствии с политикой).

Уровень слайса. Как мы упоминали ранее, каждый слайс в этом уровне принадлежит конкретному владельцу слайса, а контроллер слайса отвечает за управление ресурсами слайса, как показано на рисунке 1.10. Контроллер координирует взаимодействие между элементами слайса и хранит всю информацию слайса в информационной базе данных пользователей UID (User Information Database), например, информация о пользователях и о требовании к ресурсам. Ниже приведены основные элементы уровень слайса [30]:

- Запросы пользователей UR (User Requests). Данный элемент содержит запросы пользователей. Когда пользователь хочет получить услугу из слайса, сначала возникает необходимость вызвать связанный элемент слайса. Затем пользователь отправляет запрос контроллеру слайса, в котором упоминается запрашиваемая услуга (например, услуги передачи видео для системы видеонаблюдения, услуги передачи эластичного трафика для IoT системы или услуги голосового сообщения, и т.д.), которую он требует. Затем контроллер слайса определяет количество требуемых ресурсов (например, необходимые числа PRB) для

удовлетворения требований пользователя в зависимости от запрашиваемой услуги. UR сохраняет информацию, получаемую от пользователя в UID. Контроллер слайса при необходимости извлекает требование пользователя из UID.

- Политика пользователя UP (User Policy). Данный элемент отвечает за политику каждого пользователя (т.е. каждый пользователь связан с определенной политикой). Политика определяется администратором политики. Контроллер слайса использует определенную политику для каждого пользователя при обработке любых пользовательских запросов.
- Вычисление ресурсов для каждого пользователя RCPU (Resource Computing Per User). RCPU определяет требуемые ресурсы, необходимые для удовлетворения потребности пользователя. RCPU извлекает информацию о пользователе из UR и UP перед вычислением требуемого количества ресурса удовлетворяющее потребности пользователя. Контроллер слайса в свою очередь использует RCPU для определения точного количества ресурсов слайса, необходимых для удовлетворения запроса пользователей данного слайса.

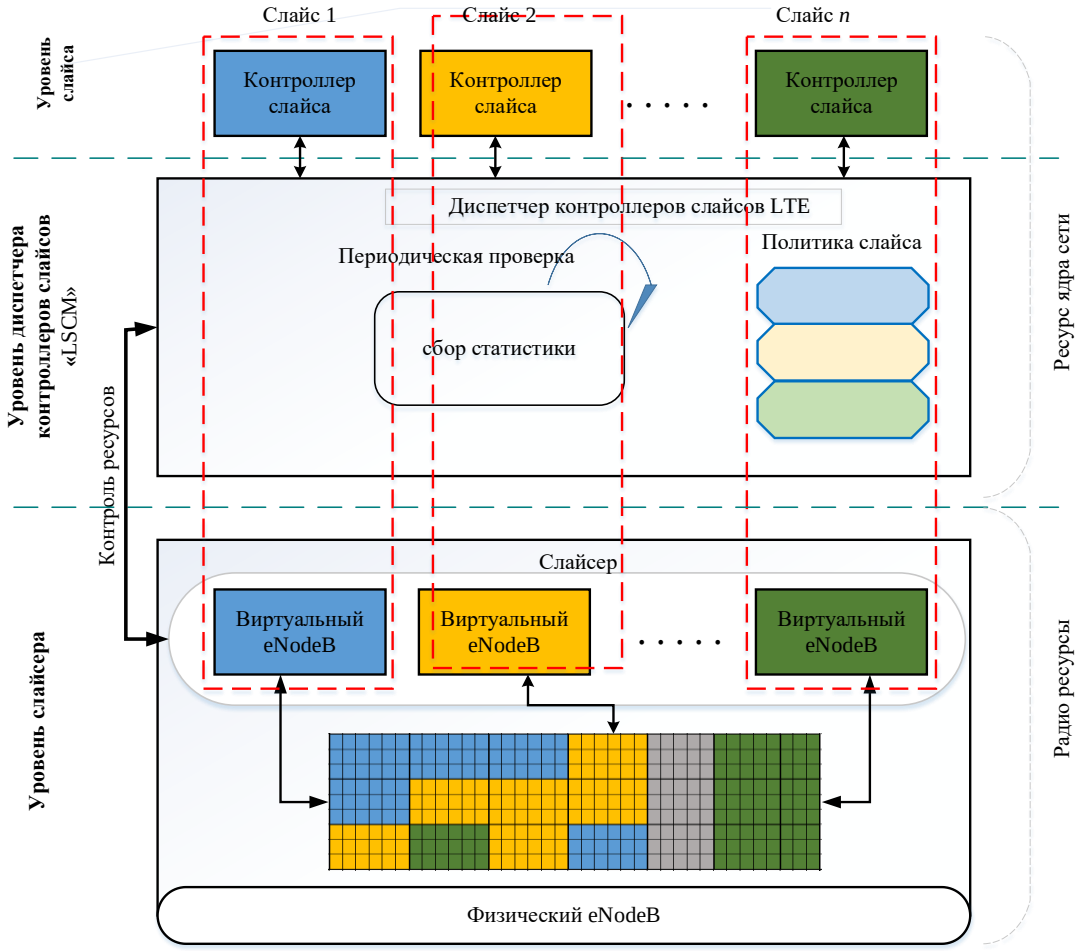


Рисунок 1.10 — Архитектура управления ресурсами сети по механизму Network slicing

- Статус пользователя (User Status). Пользователь может находиться в активном или неактивном режиме в данный момент времени [45]. Этот элемент периодически отслеживает статус пользователя, что помогает контроллеру освободить выделенные ресурсы пользователю если пользователь находится в режиме ожидания в данный момент времени. В нашем исследовании, предполагается что, когда пользователь с выделенными ресурсами переходит из активного режима в режим ожидания, он освобождает назначенные ему ресурсы. Затем контроллер слайса перераспределяет высвобожденные ресурсы между оставшимися пользователями, которые находятся в активном режиме внутри слайса. Такой подход позволит максимально использовать ресурсы слайса. В случае, когда пользователь возвращается из режима ожидания в активный режим, то контроллер слайса переназначает пользователю высвободившееся количество ресурсов. Такой подход возможно потому что весь процесс происходит в пределах одного и того же TTI. С другой стороны, после выхода из режима ожидания, если контроллер слайса не имеет необходимого количества ресурсов для удовлетворения потребности пользователя, он взывает слайсеру для назначения дополнительных ресурсов данному слайсу. Затем контроллер слайса обновляет выделение ресурсов слайса в следующем TTI.
- Отслеживание ресурсов слайса SRT (Slice Resource Tracker): данный элемент имеет глобальное представление о ресурсах слайса. Он периодически наблюдает за общим использованием ресурсов слайса и уведомляет об этом контроллеру слайса.
- Оценка ресурсов RE (Resource Estimation). Данный элемент отвечает за оценку будущего ожидаемого количества ресурсов для удовлетворения спроса пользователей в пределах слайса.

Уровень диспетчера контроллеров слайсов LTE «LSCM». Уровень LSCM управляет базовой сетью LTE (он облегчает связь между объектами ядра сети CN). Кроме того, LSCM имеет глобальное представление о требованиях к сетевым ресурсам. Он динамически отслеживает состояние сетевых ресурсов посредством статистики о требуемых ресурсах и политик назначения этих ресурсов. Уровень диспетчера контроллеров слайсов LTE использует два основных элемента для выполнения его задачу:

- Статистический сбор информации SGI (Statistics Gathering Information). Задача SGI состоит в том, чтобы получить статистику о ресурсах, необходимых для каждого слайса. Периодически, SGI собирает и сохраняет информацию о предполагаемом ресурсе для каждого слайса через элемент RE. Таким образом, он имеет историческую статистику о

ресурсе для каждого слайса. Чтобы реализовать точный расчет требуемого количества ресурса для каждого слайса, SGI измеряет среднее значение требуемого ресурса на основе полученных статистических данных о ресурсах, необходимых для удовлетворения требования каждого слайса.

- Политика распределения ресурсов RAP (Resource Allocation Policy). Данный элемент содержит все согласия (политики) между Поставщиками услуг SP (Service Providers) и Поставщиками инфраструктуры InP (Infrastructure Providers). Администратор политики сохраняет информацию о политике в RAP, что позволяет слайсеру получить информацию, связанную с политикой перед выделением ресурсов каждому слайсу (см. рисунок 1.11).

Уровень слайсера. Как показано на рисунке 1.11, в рассматриваемой модели нашего исследования вводится понятие виртуального уровня, называемого уровнем слайсера поверх физических ресурсов eNodeB. Слайсер отвечает за виртуализацию eNodeB путем создания несколько виртуальных eNodeB, где каждый из этих eNodeB представляет собой сетевая нарезка (сегмент). Он распределяет физические ресурсы eNodeB между слайсами. То есть слайсер выделяет ресурсные блоки PRB каждому слайсу с использованием алгоритма распределения полосы пропускания после учета предварительных определенных контрактов SLA (Service Level Agreement) между владельцем слайса SP и владельцем сетевой инфраструктуры InP. К основным элементам уровня слайсера относятся:

- Элементы виртуальных ресурсов VRs (Virtual Resources). Задача VRs состоит в том, чтобы создать логическую платформу и разделить эту платформу на различные логические экземпляры слайсов, где каждый логический экземпляр представляет собой слайс. Более того, VRS состоят из двух компонентов, выполняющих функциональность этой платформы (см. рисунок. 1.11): управление ресурсами по слайсу PSRM (Per Slice Resource Management). PSRM контролирует распределение ресурсов каждого слайса между пользователями слайса; вычисление ресурсов RC (Resource Computing). RC отвечает за вычисление предполагаемых ресурсов каждого слайса. RC использует модель экспоненциального сглаживания для расчета требуемых физических ресурсов PRB для каждого слайса за каждое время приема-передачи RTT (Round Trip Time). Более того, элементы SGI и RAP уровни LSCM предоставляют RC требуемые статистики и правила политики, чтобы дополнить процесс выделения ресурсов слайса.

- Мультиплексирование / демультиплексирование (англ. Multiplexing / DeMuliplexing). Он отвечает за управление несколькими потоками данных, поступающими из / в разных слайсах по радиоканалу базовой станции eNodeB.

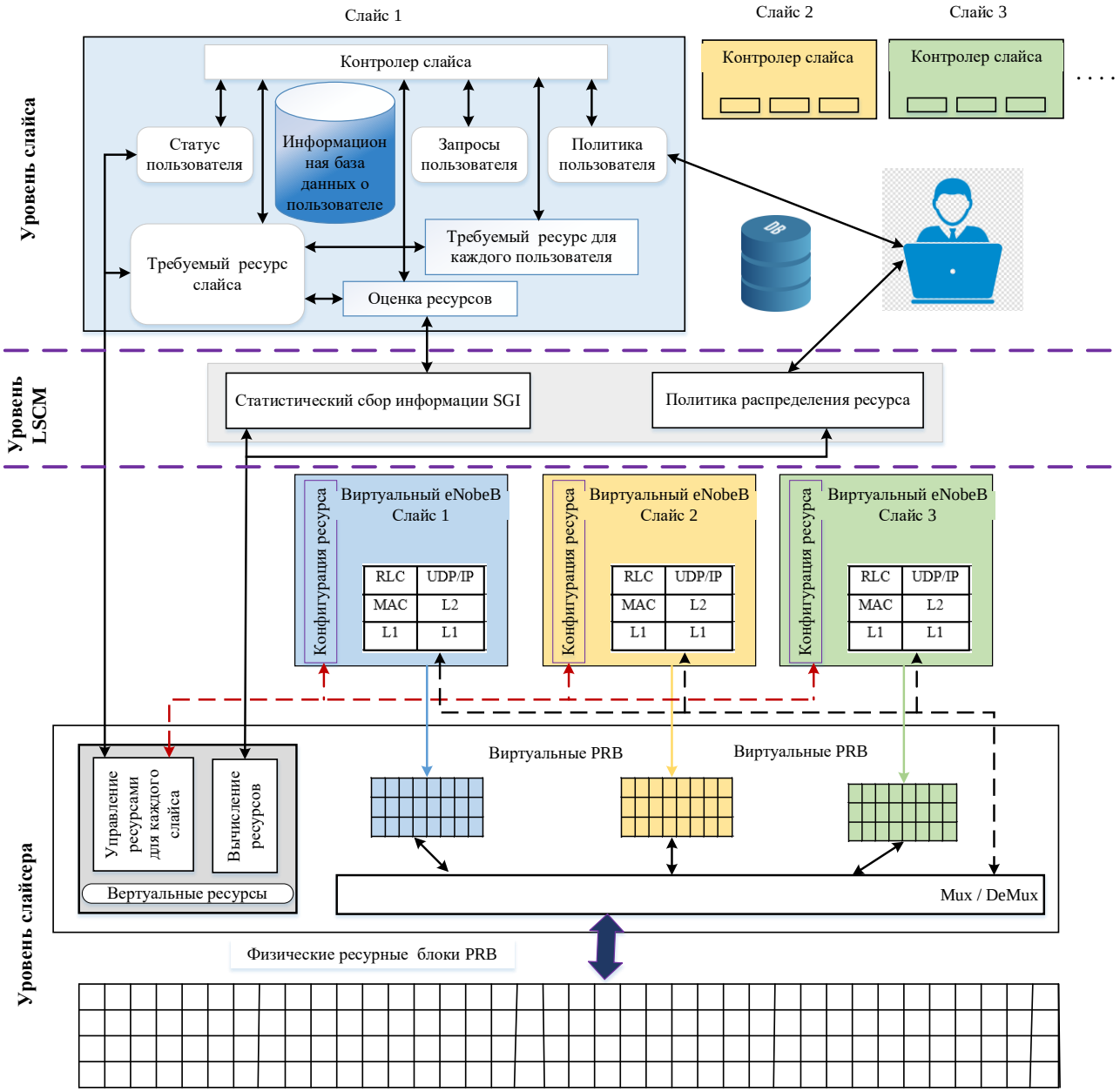


Рисунок 1.11 — Логическая взаимосвязь между элементами уровней при реализации стратегии Network slicing

Помимо вышеизложенных методов распределения ресурсов, в нашем исследовании рассматривается также способ распределения имеющегося ограниченного ресурса базовой станции eNodeB между поступающими запросами по дисциплине разделения процессора (англ. Processor

Sharing) внутри изолированного слайса. Использование такого подхода в распределении ресурсов существенно повышает коэффициент использования ресурсов системы мобильной связи.

1.8. Анализ выполненных исследований по тематике диссертационной работы

На сегодняшний день было проведено множество исследований, направленных на улучшение моделей распределения ресурсов в мобильных сетях. В некоторых из этих работ были предложены механизмы распределения ресурсов, основанные на назначении ряда физических ресурсных блоков PRB каждому запросу пользователя сотовой связи с учетом состояний канала и требований к качеству обслуживания QoS [6, 21, 22, 25, 64]. В работах [7, 36] проведены исследования, направленные на устранение недостатков, связанных с неэффективным использованием ресурса с использованием разных методов динамического его распределения. Неконтролируемое перераспределение ресурса в пользу потоков, требующих малую скорость передачи данных, появляется при совместном использовании ограниченного радиоресурса узлов доступа сетей беспроводной связи (см. подраздел 4.2). Для устранения отмеченных недостатков, в работах [13, 14, 22] предлагается использовать механизмы резервирования, основанные на перераспределении ресурса между обслуживаемыми заявками. Расчет характеристик качества обслуживания поступающих запросов с использованием функции блокировки для резервирования ресурсов в пользу выбранной группы потоков не вызывает трудности и осуществляется с использованием известных численных алгоритмов [12, 17, 19, 23, 26, 62, 105]. В общем случае приближенные алгоритмы [69, 90, 108] или имитационное моделирование [25] являются наиболее подходящими методами для оценки характеристик качества обслуживания потоков трафика в мультисервисных узлах доступа.

Анализ публикаций и выполненных диссертационных исследований показал, что в большинстве теоретических работ либо изучалось действие какого-то одного фактора на процесс распределения ресурса узла доступа (например, зависимость требования к ресурсу от типа сервиса, ограничение доступа, резервирование ресурса и т.д.), либо процесс распределения ресурса рассматривался с избыточной детальностью, что в итоге затрудняло использование построенной математической модели. Задача построения модели, которая, с одной стороны, отражала основные реалии распределения ресурса, а с другой — могла бы использоваться в практических приложениях не рассматривалась, что и определило направление исследований, выполненное в диссертации.

1.9. Постановка задачи диссертационного исследования

Достижение эффективного распределения радиоресурсов является сложной задачей в основном из-за изменчивости пропускной способности беспроводных каналов связи и из-за ограничения ресурсов. Целью данного исследования является разработка и анализ модели динамического распределения ресурса с резервированием при передаче неоднородного трафика IoT. Для достижения указанной цели необходимо решить следующие задачи:

- Разработать модель динамического распределения ресурса в беспроводном узле доступа, которая учитывает влияние основных факторов, которые определяют совместное обслуживание гетерогенного трафика реального времени и эластичных данных. Среди них: наличие приоритета у трафика реального времени; использование дисциплины Processor Sharing при передачи эластичного трафика; ограничение по доступу для всех видов трафика, зависящее от общего уровня занятости ресурса.
- Определить характеристики качества обслуживания поступающих запросов.
- Построить алгоритмы оценки характеристик.
- Сформулировать рекомендации по эффективному распределению ресурса между поступающими потоками неоднородного трафика.

Разворачивание стандартов IoT осуществляется на инфраструктуре существующих сетей LTE. Количество ресурсов передачи информации в телекоммуникационных сетях беспроводной связи ограничено действиями частотного регулирования и физическими возможностями передачи радиосигналов. По этой причине исследования направленные на определение эффективных сценариев распределения ресурса беспроводных узлов доступа при обслуживании неоднородного трафика современных коммуникационных приложений являются актуальными. Решение сформулированных задач будет рассмотрено в последующих разделах диссертации.

1.10. Выводы по результатам первого раздела

1. Проведен анализ основных параметров, используемых на практике для описания качества работы мобильных сетей на базе стандарте LTE. Показано, что процесс планирования радиоресурсов является важнейшей задачей сетей LTE, поскольку планирование отвечает за эффективное распределение радиоресурсов. Для достижения требуемых целей, блок управления радиоресурсами LTE RRM использует набор функций MAC и функций

физического уровня, таких как совместное использование ресурсов, отчетность по индикатору качества канала CQI, адаптация канала связи с помощью адаптивной модуляции и кодирования AMC и гибридного автоматического запроса повторной передачи HARQ.

2. Проведен анализ особенностей радиоинтерфейсов сетей стандарта LTE и NB-IoT и показано, что разворачивание стандарта NB-IoT осуществляется на инфраструктуре существующих сетей LTE путем обновления программного обеспечения оборудование LTE без покупки новых оборудований, т.е. с меньшими финансовыми затратами.
3. Показано, что применение механизма Network Slicing в сети мобильной связи LTE существенно повышает коэффициент занятия ресурсов сети путем использования преимуществ технологий NFV и SDN.
4. Устранение недостатков, вызванных неконтролируемым перераспределением радиоресурса в пользу потоков с низкой скоростью передачи данных, осуществляется путем резервирования ресурсов, либо путем разделения низкоскоростных и высокоскоростных потоков данных по отдельным слайсам. Для теоретического обоснования возможности применения перечисленных процедур необходимо построение математической модели узла доступа, учитывающей совместное влияние основных факторов, определяющих функционирование узла и разработка алгоритмов оценки характеристик обслуживания заявок.

Раздел 2

Модель совместного обслуживания трафика реального времени и эластичного трафика данных в узле доступа сети подвижной связи при наличии процедуры резервирования ресурса

2.1. Введение к разделу 2

Решение задач, сформулированных в конце предыдущего раздела и относящихся к достижению цели исследовательской работы, осуществляется путем построения и анализа математической модели узла доступа сети стандарта LTE, учитывающей особенности поступления и формирования гетерогенного трафика реального времени и трафика эластичных данных (см. подразделах 1.5 – 1.6). Информационные сообщения генерируются в виде сессий связи видеокамерами и разного рода датчиками, установленными оператором систем наблюдений на территории соты. Поступающие информационные потоки занимают выделенный оператору ресурс мультисервисного узла доступа сети беспроводной связи стандарта LTE. Разделение ресурса между принятыми заявками зависит о настройки комплекса RRM. В подразделе 2.2 в общем виде рассмотрено разделение ресурса по макроканалам, интегрирующим передаточные возможности выделенного ресурса соты.

В подразделе 2.3 исследуются особенности поступления и совместного обслуживания потоков информационных сообщений (сессий передачи трафика), поступающих от разнообразных технических устройств телеметрии, подключенных к сети. Характеристики устройств приведены в подразделах 1.5 и 1.6. В подразделе 2.4 построены математические модели формирования продолжительности интервалов времени между появлением заявок и времени их обслуживания. Введены компоненты марковского процесса, описывающего процесс изменения состояний модели, построена система уравнений равновнесия. Выражения для показателей качества обслуживания поступающих запросов получены в подразделе 2.5. Расчетные формулы для оценки характеристик представлены через значения стационарных вероятностей состояний модели. Зависимости между характеристиками, которые используются при их измерении или расчете получены в подразделе 2.6.

Модель мультисервисного узла доступа и алгоритмы, полученные на ее основе, используются:

- для анализа зависимости характеристик качества обслуживания рассматриваемых сессий от параметров возникающих информационных потоков и условий допуска сессий связи к занятию ресурса;
- для анализа разных способов увеличения эффективности использования ресурса узла доступа и создания условий по дифференцированному обслуживанию поступающих сессий связи.

Результаты, полученные при решении сформулированных задач рассмотрены в разделе 4. Одной из целей диссертации является анализ использования концепции Network Slicing (см. подраздел 1.7) при обслуживании гетерогенного трафика устройств телеметрии, и разработка алгоритмов повышения эффективности применения данной концепции. В процессе реализации концепции Network Slicing отдельные информационные потоки с примерно одинаковыми требованиями к условиям передачи обслуживаются отдельными слайсами, представляющими из себя часть выделенного ресурса соты, обособленного для реализации этих целей. Модели обслуживания отдельных информационных потоков в слайсах, которые являются частными случаями исследуемой модели узла, рассмотрены в разделе 3. В подразделе 2.7 сформулированы выводы по разделу 2.

2.2. Динамическое распределение ресурса передачи информации

2.2.1. Общие положения

В традиционных моделях теории телетрафика, относящихся к передаче данных, пересылка информационных потоков рассматривалась на уровне отдельных информационных сообщений. Это могут быть пакеты или некоторая группа пакетов. Ресурс, выделяемый для их обслуживания, не меняется в процессе перемещения сообщения по сети [11–14, 21, 22, 24, 25, 31, 36, 45, 78, 99–102]. Этот классический подход в настоящее время претерпевает существенные изменения. Они вызваны широким распространением услуг Интернета и сетей сотовой подвижной связи. В условиях дефицита ресурса, скорость передачи эластичных данных может уменьшиться до минимального значения при сохранении качества предоставляемого сервиса. Назначение ресурса, необходимого для передачи информационных данных осуществляется по определенным правилам,

зависящим от возможностей интегрирования свободного ресурса в один макроканал. Обычно, ресурс передачи информации делится в одинаковой пропорции между всеми запросами, находящимися в данный момент на обслуживании, но можно применять и разного рода преимущественные схемы его назначения.

Действие механизма перераспределения ресурса может происходит в разные моменты времени, но наиболее естественно предположить, что изменение скорости передачи информации происходит при изменении числа запросов, находящихся на обслуживании в рассматриваемый момент времени. Будем называть этот подход динамическим способом разделения канального ресурса. Достигаемое ускорение передачи данных в те моменты времени, когда ресурс узла частично или полностью освобождается от пересылки информационных данных, осуществляемых в реальном времени, повышает эффективность использования ресурса беспроводных узлов доступа (соответствующие численные примеры приведены в разделе 4). Этот эффект имеет особое значение для беспроводных сетей связи так как диапазон радиочастот, выделяемый для образования радиоканалов ограничен. Дефицит радиочастот связан с физическими ограничениями на передачу информации по радиоканалам и действиями регулятора, направленными на повышение конкуренции на рынке связи.

2.2.2. Особенности моделирования передачи эластичных данных

Передача данных, обладающих свойствами эластичности, анализируется на уровне отдельных информационных потоков, каждый из которых представляет из себя последовательность пакетов, формируемых используемой технологией передачи информации, и обрабатываемых коммутаторами по единым для всех пакетов процедурам. Отдельный поток в процессе моделирования будет представлять из себя информационное сообщение, связанное с пересылкой файлов, текстовых документов или их фрагментов. Информационные потоки данного типа обладают свойствами эластичности и могут передаваться по сети с переменной скоростью без снижения показателей качества предоставления сервиса. Скорость передачи пакетов в рассматриваемом потоке изменяется в соответствии с значением метрик, отражающих степень загрузки ресурса. Для TCP-соединений подобной метрикой является доля потерянных пакетов. Таким образом, скорость пересылки информации меняется в процессе обслуживания заявок и пользователь оценивает качество работы сети исходя из времени загрузки файла. Используемая шкала времени позволяет не принимать во внимание пакетную структуру информационного

потока и ограничиться исследованием процесса распределения ресурса в моменты начала и окончания обслуживания отдельных потоков.

Анализ структуры информационных потоков, относящихся к передаче эластичного трафика, формируемого разного рода датчиками, показывает, что потоки группируются в сессии. При этом можно утверждать, что внутри одной сессии существует зависимость между интервалами времени поступления отдельных потоков. По мнению экспертов[2], можно полагать, что отдельные сессии являются результатом независимой активности пользователей услуг связи в получении заказанного сервиса. Известно из теории вероятностей, что если число источников нагрузки велико и процесс генерации запросов можно считать независимым, то можно полагать, что поток заявок, инициирующих открытие сессий связи, будет пуассоновским. Более того, пуассоновским можно считать и поток информационных сообщений (файлов), составляющих сессию связи.

2.2.3. Распределение ресурса передачи информации при обслуживании эластичных данных

В зависимости от постановки задачи и анализируемой технологии передачи информации при распределении ресурса могут применяться разные подходы. Например, если требуется увеличить объем передаваемого трафика, то радиоканалы необходимо выделять абонентам, находящимся в лучших условиях приема радиосигналов. Существует много разных алгоритмов, реализующих распределение ресурса передачи информации в соответствии с так называемой функций полезности (англ. utility function), рассчитываемой для всех пользователей услуг связи, находящихся на обслуживании. К подобным решениям относится процедура пропорциональной равнодоступности ресурса (англ. proportional fairness) [25].

В моделировании процесса разделения ресурса передачи информации используются два подхода. В первом — ресурс задается битовой скоростью передачи и делится в соответствии с требованием к скорости от поступающих заявок. При этом скорость передачи эластичных данных при обслуживании одной заявки меняется в некотором диапазоне: от минимально возможной, до максимально возможной. Реально получаемая скорость получается делением незадействованной скорости передачи поровну между всеми заявками, принятыми на обслуживание. Этот подход основан на идеализированном предположении о возможности непрерывного разделения скорости передачи информации между активными пользователями. Теоретически заявка, находящаяся на обслуживании, может получить любую скорость пересылки информации, не превышающую заданных ограничений. При реализации второго подхода назначение ресурса происходит

[2]Bonald T. Internet and the Erlang formula / T. Bonald, J.Roberts // ACM SIGCOMM Computer Communication Review. — 2012. — V. 42. — N. 1. P. — 22–30.

порциями. Минимальная порция соответствует минимальной градации имеющегося ресурса передачи информации. Будем называть эту порцию единицей ресурса, канальной единицей, виртуальным каналом или просто каналом. В случае необходимости отдельные порции ресурса, могут быть сгруппированы в соответствии с каким-то алгоритмом и все вместе предоставлены пользователю услуг связи для передачи требуемой информации.

Покажем, как решается эта задача в ситуации, когда технология NB-IoT реализуется в полосе частот LTE, сценарий (англ. in band) [79] (см. подраздел 1.4). При этом блок планирования ресурсов в NB-IoT представлен ресурсным элементом RE (Resource Element) или тоном, позволяющим устройствам NB-IoT передавать радиосигнал на одной поднесущей на частоте 15 кГц, и предоставляющий возможность обслуживать несколько устройств в полосе частот 180 кГц. Кроме того, ресурсный блок NB-IoT делится на 12 или 48 поднесущих по 15 кГц или по 3,75 кГц соответственно. С разносом 15 кГц, NB-IoT использует режим single-tone (8 мс) или режим multi-tone (3 тона, 6 тонов и 12 тонов) с длительностью 4 мс, 2 мс и 1 мс, соответственно. С использованием 3,75 кГц поддерживается только режим single-tone с 48 подкадрами длительностью 32 мс. Данный режим организации пересылки информации устройствами NB-IoT позволяет реализовать принцип динамического распределения ресурса, о котором идет речь в данном подразделе.

Помимо эффективного использования имеющейся полосы пропускания этот принцип дает возможность уменьшить время активного режима работы устройства NB-IoT, что позволяет экономить ресурс батареи. Имея ввиду ожидаемые высокие темпы роста числа задействованных устройств NB-IoT, будем предполагать, что число ресурсных блоков LTE, задействованных для передачи трафика датчиков может быть при необходимости увеличено с единицы до нескольких штук. Данный подход объясняется следующими особенностями принятия стандартов передачи информации в телекоммуникационных системах.

Телекоммуникационный ресурс является важнейшей характеристикой объекта исследования. В широком смысле ресурс представляет из себя технические возможности, обеспечиваемые сетью связи или отдельными ее сегментами для поддержки запросов клиентов на информационное обслуживание. Если используемый ресурс представлен битовой скоростью, то для оценки его объема вводится понятие *виртуальной единицы ресурса*, которая определяется исходя из минимального требования к ресурсу от поступающих запросов. Далее общий объем ресурса оценивается через максимально возможное число создаваемых виртуальных единиц. При использовании технологии LTE и ее дальнейшего развития (5G, 6G) величина ресурса

определяется более сложными методами. Со стороны радиоинтерфейса ресурс представлен в виде ресурсных блоков. Часть из них выделяется для обеспечения скорости передачи информации, требуемой для обслуживания поступающего запроса. Необходимо отметить, что здесь нет линейной зависимости между числом выделяемых блоков и обеспечиваемой ими скоростью передачи. Соответствующий функционал имеет сложный характер и зависит от множества факторов. В их перечень входят: схема кодирования; алгоритм работы диспетчера пакетов; расстояние до базовой станции; число и типы запросов, находящихся в определенный момент на обслуживании, и соотношение между ними и т.д. В этой ситуации величина скорости определяется опытным путем, исходя из опыта эксплуатации подобных систем и натурных экспериментов, и последующей калибровки результатов расчетных алгоритмов. В качестве примерного значения C может быть выбрана величина максимальной пропускной способности соты LTE [см. таблицы 1.1 и 1.2].

В этом и последующем разделах будут рассмотрены теоретические и практические вопросы обслуживания запросов на передачу трафика эластичных данных в соответствии с изложенными выше принципами. Сформулируем предположения, необходимые для построения математической модели. Обозначим через v общее число единиц ресурса. Пусть $r-$ скорость пересылки информации, предоставляемая единицей ресурса в битах в секунду. На практике, в качестве значения одной единицы ресурса (канальное единицы) принимается минимальное требование к скорости передачи информационных данных от всех поступающих запросов. Понятно, что таковым в рассматриваемой постановке задачи будет минимальная величина скорости передачи информации, которая может быть предложена при пересылке эластичных данных. Пусть C — скорость передачи информации в битах в секунду, обеспечиваемая всем ресурсом рассматриваемого узла, $C = vr$. а $d-$ число запросов на передачу трафика эластичных данных. Предположим для простоты, что в модели рассматриваются только запросы на передачу эластичных данных. Имеющийся ресурс передачи информации используется потоком заявок на передачу данных со скоростью, которая меняется в соответствии с изменением числа запросов, находящихся на обслуживании. Предполагается, что единицы ресурса могут интегрироваться в одном макроканале для обслуживания запроса передачи данных.

Распределение ресурса на макроканалы удовлетворяет нескольким правилам. Обозначим через v_d скорость передачи эластичных данных одного макроканала, выраженную в канальных единицах. Множество значений, принимаемых v_d обозначим через ψ. Назначение скоростей

макроканалов при известных значениях d и v определяется диспетчером пакетов на основе используемой процедуры распределения канального ресурса между заявками, находящимися на обслуживании. Распределение ресурса при этом либо следует структуре макроканалов, заданной стандартом передачи информации, либо направлено на максимизацию общего числа занятых ресурсов. Необходимо, чтобы определение скоростей макроканалов при заданных значениях v происходило единственным образом.

Приведем примеры следования заданной структуре макроканалов, направленные на максимизацию числа используемых единиц ресурса. В первом случае можно рассмотреть двоичное назначение числа каналов в макроканале в соответствии с выражением $v_{d,i} = 2^i$, где i последовательно принимает значения от 0 до некоторого значения j, при котором скорость $v_{d,j} \le v$, а $v_{d,j+1} > v$ [25]. Другое распределение каналов в макроканалы используется при реализации процедуры multi-tone (см. выше). Там каналы интегрируются по схеме 3, 6 или 12.

Более простой с точки зрения моделирования вариант назначения ресурса направлен на использование всех имеющихся свободных единиц канального ресурса [25]. Предполагается, что скорость макроканала кратна одной канальной единице и может принимать значения от 1 к.е. до v к.е. Предположим, что $a = \left\lfloor \frac{v}{d} \right\rfloor$. В данном случае $v - ad$ макроканалов по $a+1$ к.е. и $d(a+1) - v$ макроканалов по a к.е. используются для обслуживания d запросов. Понятно, что все v к.е. заняты обслуживанием d запросов при использовании такого выбора скоростей макроканалов. Данный алгоритм выбора скоростей макроканалов будем называть дисциплиной разделения процессора (Processor Sharing — *PS*). Функциональная модель разделения ресурса между сессиями трафика реального времени и сессиями передачи эластичных данных показана на рисунке 2.1.

Ряд теоретических и практических исследований утверждает, что характеристики обслуживания эластичных данных слабо зависят от выбранного типа алгоритма назначения канальных ресурсов, если он основан на методах справедливого распределения ресурса между отдельными устройствами таких, как алгоритм максиминного назначения ресурса, алгоритм пропорционального назначения ресурса передачи информации и метод сбалансированного назначения ресурса передачи эластичных данных [38–41]. Следовательно, характеристики качества обслуживания трафика эластичных данных зависят в большей степени от объема передаваемых файлов, а не от того, какая между ними используется справедливая схема распределения ресурса. Данное положение позволяет упростить процесс моделирования

особенностей распределения ресурса по макроканалам, предоставляемым пользователю. Поэтому далее при распределении ресурса между запросами на передачу эластичных данными будем использовать введенную выше дисциплину *PS*.

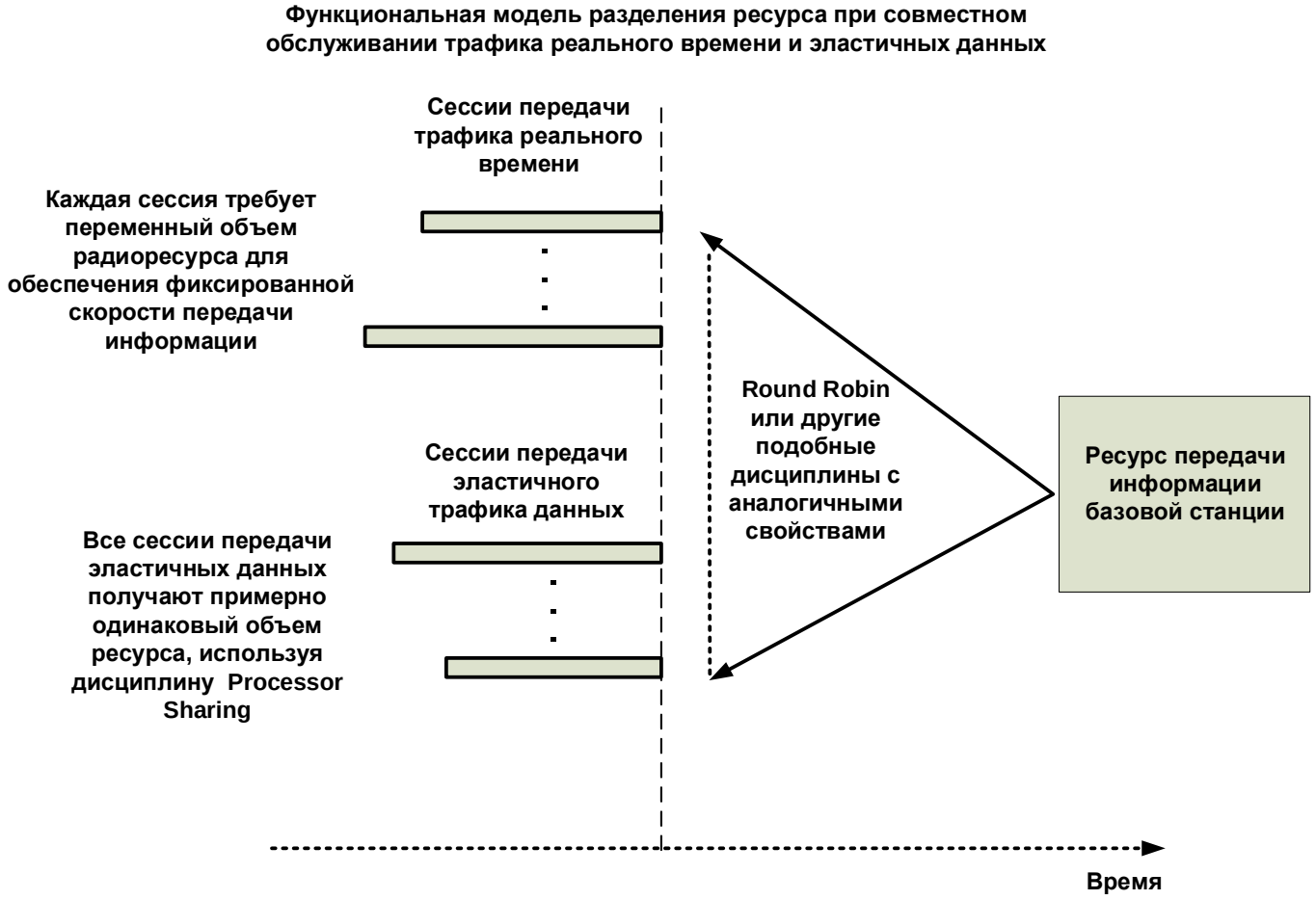


Рисунок 2.1 — Процедура разделения ресурса между сессиями трафика реального времени и сессиями передачи эластичных данных

2.3. Функциональная модель совместного обслуживания трафика реального времени и эластичного трафика данных

2.3.1 Формирование потоков запросов оператора систем наблюдения

Анализируемая система связи представляет собой изолированную соту сети стандарта LTE, часть ресурса которой (выделенный слайс) арендуется оператором систем наблюдения для передачи информационных потоков, инициируемых видеокамерами и разного рода датчиками. В

дальнейшем исследуется процесс разделения выделенного ресурса с целью повышения эффективности его использования и создания условий для дифференцированного обслуживания поступающих потоков. Понятно, что информационные потоки видеокамер (сессии связи) обладают свойствами трафика реального времени, т.е. требуют фиксированную скорость передачи на всё время соединения. Совместно с трафиком реального времени обслуживаются запросы на передачу эластичных данных. К ним относятся, например, видеокамеры с записью видеоконтента в буфер, а также датчики, генерирующие файлы измерений большого объема, например, фотографии высокого качества. Соответствующие сессии связи обладают свойством эластичности, т.е. могут менять скорость передачи в некоторых пределах без потери качества предоставления услуги. Предполагается, что каждая отдельная видеокамера функционирует как обычный пользователь услуг сети LTE, т.е. напрямую соединяется с базовой станцией.

Будем считать, что число видеокамер велико и они посылают запросы на информационное обслуживание независимо друг от друга. Известно из теории вероятностей, что если число источников нагрузки велико и процесс генерации запросов можно считать независимым, то можно полагать поток заявок, инициирующих начало сессий, пуассоновским с некоторой заданной интенсивностью [25]. Каждый из рассматриваемых потоков имеет свою интенсивность поступления запросов на информационное обслуживание и требование к величине скорости передачи информации, которая зависит от качества предоставления услуги передачи видеоконтента. По этой причине будем предполагать, что в модели анализируется процесс поступления n потоков запросов на пересылку сессий сервисов реального времени. Заявки сгруппированы в потоки в зависимости от требования к скорости передачи порождаемых ими информационных сообщений. Будем предполагать, что длительность сессии на передачу трафика сервисов реального времени имеет экспоненциальное распределение. Запросы подобного рода поступают от устройств телеметрии, подключенных к соте LTE для последующей передачи в аналитические центры результатов наблюдений. Рассматриваемая функциональная модель работы изолированной соты сети LTE, часть ресурса которой (выделенный слайс) арендуется оператором систем наблюдения для передачи информационных потоков, инициируемых видеокамерами и разного рода датчиками показана на рисунке 2.2.

Рисунок 2.2 — Функциональная модель работы изолированной соты сети LTE при совместном обслуживании трафика сервисов реального времени и эластичного трафика данных

2.3.2. Распределение ресурса между сессиями трафика реального времени и эластичных данных

Будем предполагать, что в соответствии со свойствами обслуживаемых информационных потоков, трафик реального времени получает преимущество в занятии ресурса передачи перед трафиком эластичных данных. В силу необходимости, используемый ресурс для передачи эластичных данных может уменьшен до одного канала, но не меньше. При появлении свободного канального ресурса скорость пересылки данных возрастает. Перераспределение ресурса происходит в момент изменения числа заявок, находящихся на обслуживании и осуществляется по динамической основе. Предполагается что назначение ресурса для передачи эластичных данных подчиняется правилам дисциплины *PS* (см. раздел 2.2.3).

Используемая процедура назначения канального ресурса соты позволяет существенно повысить его загрузку. Результаты расчетов показывают, что выигрыш может составить до

нескольких десятков процентов от его общего объёма. Численное исследование, иллюстрирующее сформулированное положение, будет проведено в разделе 4. Отмеченный эффект получен в результате действия процедур управления трафиком, которые переносят передачу данных на те моменты времени, когда выделенный ресурс узла беспроводной связи свободен от обслуживания тяжелого трафика, который передается в реальном времени. Подобные механизмы реализуются в настройках комплекса RRM (см. подраздел 1.3). Для создания условий по дифференцированному обслуживанию поступающих запросов будет использоваться процедура резервирования ресурса. Его действие будет ограничивать занятие ресурса каждым из поступающих потоков. Выбор ограничений зависит от общего уровня занятости выделенного ресурса всеми поступающими информационными потоками.

Для оценки эффективности динамического распределения канального ресурса соты необходимо построить математическую модель совместного обслуживания потоков трафика реального времени и потока трафика эластичных данных. В подразделе 2.4 будет построена и исследована математическая модель совместного обслуживания произвольного числа потоков трафика реального времени и потока трафика эластичных данных. Доступ запросов на обслуживание каждого вида информационных сообщений ограничен с использованием процедуры резервирования.

2.4. Математическая модель совместного обслуживания трафика реального времени и эластичного трафика данных в узле доступа LTE

2.4.1. Модель поступления запросов на информационное обслуживание

Напомним (см. подраздел 2.2.3) обозначения, используемые для характеристики ресурса узла доступа, выделенного для обслуживания информационных потоков оператора систем наблюдения, выраженного через значения создаваемой выделенным ресурсом битовой скорости пересылки информации и скорости требуемой каждым потоком для качественной передачи трафика. Через v обозначено общее число имеющихся виртуальных каналов или передаточных единиц ресурса соты, которые используются для совместного обслуживания поступающих запросов, абозначим через r скорость передачи информации, обеспечиваемая одним единичным ресурсам. Обычно в качестве таковой выбирается минимальное допустимое требование к скорости передачи

эластичных данных. Для сокращенного обозначения используемой виртуальной канальной единицы ресурса будем использовать аббревиатуру к.е. (канальная единица).

Пусть C — скорость передачи информации в битах в секунду, обеспечиваемая всем ресурсом модели $C = vr$. В модели имеются только один поток для передачи файлов и n потоков для передачи трафика сервисов реального времени. Все поступающие потоки на установеление сессий связи подчиняются закону Пуассона. Обозначим через λ_k параметр экспоненциального распределения интервала времени поступления запросов k-го потока. Пусть b_k – число канальных единиц, необходимых для обслуживания одного запроса k-го потока, а μ_k — параметр экспоненциального распределения длительности сессии k-го потока, где $k = 1, 2, \ldots, n$. Если в момент поступления запроса k-го потока нет достаточного числа свободных канальных единиц, но это количество может быть получено в результате уменьшения скорости передачи находящихся на обслуживании сессий эластичных данных, то соответствующее изменение скоростей выполняется и поступивший запрос принимается для организации сессий связи.

Поступивший запрос k-го потока на передачу трафика реального времени может получить отказ в обслуживании поскольку минимальное значение, до которого может снизиться скорость передачи одной сессии эластичных данных, равно скорости, обеспечиваемой одной канальной единицей. Сформулируем условия, необходимые для этого. Пусть i_r – число занятых канальных единиц обслуживанием тяжелого трафика в момент поступления запроса на передачу эластичных данных. Обозначим через d число обслуживаемых запросов на пересылку эластичных данных. Заявка k-го потока на пересылку трафика реального времени принимается для организации сессий, если выполняется условие $i_r + d + b_k \le v$. Выполнение данного условия означает, что все заявки на пересылку файлов уже получают для своего обслуживания один канал и дальнейшее уменьшение скорости передачи эластичных данных невозможно. Таким образом, в ситуации $i_r + d + b_k > v$ поступивший запрос получает отказ и не возобновляется.

Поступление запросов на передачу эластичных данных подчиняется пуассоновскому закону с интенсивностью λ_d. Поступивший запрос принимается к обслуживанию, если выполняется условие $i_r + d + 1 \le v$. Так же как при организации доступа сессий реального времени выполнение данного условия означает, что все заявки на пересылку файлов уже получают для своего обслуживания один канал и дальнейшее уменьшение скорости передачи эластичных данных невозможно. Таким образом, в ситуации $i_r + d + 1 > v$ поступивший запрос получает отказ и не

возобновляется. Распределение ресурса на обслуживание эластичных данных рассмотрено в разделе 2.2.3 и подчиняется правилам дисциплины *PS* (Processor Sharing). Средний объём пересылаемого файла, выраженный в битах и экспоненциально распределено, обозначим через F. Обозначим через $\mu_d = r/F$ параметр экспоненциального распределения времени передачи файла с использованием передаточных возможностей одной канальной единицы.

Ограничение доступа к информационным ресурсам для запросов на пересылку трафика реального времени осуществляется с помощью функции блокировки $\varphi_k(i_r + d)$. Функция $\varphi_k(i_r + d)$ зависит от требуемого b_k числа канальных единиц для обслуживания поступившей заявки и от значения $i_r + d$ канального ресурса, занятого всеми потоками. Причем обслуживание сессий эластичного трафика происходит с минимальной скоростью, т.е. с использованием одного канала, таким образом, дальнейшее уменьшение скорости передачи трафика эластичных данных невозможно. Заявка k-го потока на передачу трафика реального времени, поступившая в момент занятости i_r канальных единиц сессиями реального времени и наличии d сессий передачи эластичного трафика, принимается к обслуживанию с вероятностью $1-\varphi_k(i_r + d)$, а с дополнительной вероятностью $\varphi_k(i_r + d)$ запрос получает отказ и не приводит к появлению повторной попытки соединения. Величины $\varphi_k(i_r + d)$ удовлетворяют определенным ограничениям. Значение фильтрующей функции лежит в пределах от 0 до 1, т.е. $0 \le \varphi_k(i_r + d) \le 1,\ \ k = 1,2,\ldots,n,\ i_r + d = 0,1,\ldots,v$. Поступающий запрос k-го потока получает отказ на обслуживание из-за нехватки ресурса с вероятностью, равной единице т.е. $\varphi_k(i_r + d) = 1,$ $k = 1,2,\ldots,n$ если $i_r + d = v - b_k + 1,\ v - b_k + 2,\ldots,v$.

Аналогичным образом фильтруется допуск запросов на передачу эластичных данных. Обозначим через $\varphi_d(i_r + d)$ соответствующую функцию фильтрации. Запрос на передачу файлов, принимается к обслуживанию с вероятностью $1-\varphi_d(i_r + d)$, а с дополнительной вероятностью $\varphi_d(i_r + d)$ запрос получает отказ и не приводит к появлению повторной попытки соединения. Рассматриваемая математическая модель использования ресурса соты сети LTE, выделенного для обслуживания сессий связи оператора систем наблюдения, показана на рисунке 2.3.

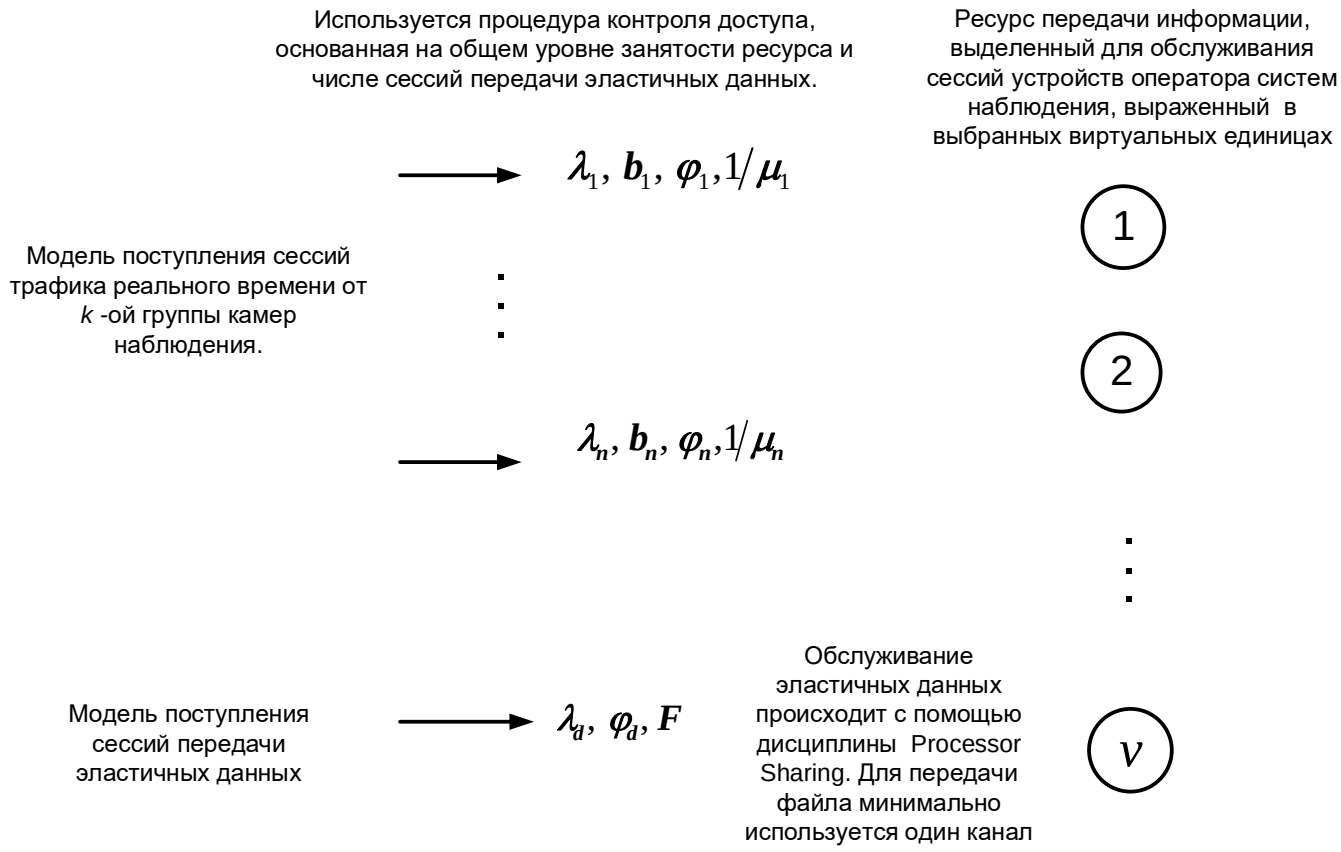


Рисунок 2.3 — Математическая модель использования ресурса соты сети LTE, выделенного для обслуживания сессий связи оператора систем наблюдения

Рассмотрим более подробно процесс выделения ресурса для поступившей заявки на передачу эластичных данных. Обозначим через v_d – скорость одного макроканала, принадлежащего множеству ψ, который используется для обслуживания одного запроса на передачу данных. Пусть f – число выделенных единиц ресурса, занятого для передачи эластичных данных, а d — число обслуживаемых сессий на передачу эластичных данных. Двумерный вектор с компонентами f , d обозначим через (f,d). Вектор (f,d) показывает состояние занятости ресурса передачи и число запросов, находящихся на обслуживании. По условиям построения модели при фиксированном f величина d не превосходит f. Назначение ресурса в виде макроканалов для обслуживания d запросов производится в соответствии с принципами, изложенными в подразделе 2.2.3 при известных значениях f и d . Перераспределение скоростей макроканалов осуществляется в соответствии с выбранной процедурой при изменении значения f и d в процессе поступления или окончания обслуживания сессий связи всех типов. Интервал времени обслуживания одной сессии передачи эластичных данных (файла) имеет

экспоненциальное распределение с параметром $\mu_d = r/F$. В этом случае обслуживается файл объема F одним каналом со скоростью передачи r бит в секунду. Если в какой-то момент происходит изменение числа сессий, находящихся на обслуживании, то по условиям построения модели меняется и величина используемого ресурса. В результате меняется и остаточное время обслуживания рассматриваемой сессии на передачу эластичных данных.

В построенной модели предполагается, что изменение скорости передачи выполняется только для сессий эластичных данных. В условиях перегрузки узла обслуживание принятых сессий на передачу файлов происходит на скорости, обеспечиваемой одной канальной единицей. Если в последнем случае на обслуживании находится один файл, то он передается на скорости, обеспечиваемой всеми свободными единицами ресурса, выделенными оператору систем наблюдения. Определим интенсивность обслуживания сессий на передачу эластичных данных в зависимости от состояния занятости выделенного ресурса. Предположим, что сота находится в состоянии (f,d). Напомним, что d — число сессий связи на передачу эластичных данных. Интервал времени до окончания обслуживания одного из d запросов на передачу эластичных данных имеет экспоненциальное распределение с параметром $\mu(f,d)$, зависящим от компонентов f,d и от принимаемого метода образования макроканалов. При использовании дисциплины *PS* (см. подраздел 2.2.3) все свободные от передачи трафика реального времени канальные единицы используются для передачи эластичных данных, принятых на обслуживание. Тогда $\mu(f,d) = f\mu_d = (v - i_r)\mu_d$. Здесь i_r — ресурс, занятый на обслуживание сессий на передачу трафика реального времени.

2.4.2. Марковский процесс и пространство состояний модели

Определим вид и перечислим состояния, которые входят в пространство состояний марковского процесса, который используется для описания процесса обслуживания анализируемых сессий связи. Пусть $i_k(t)$ - число обслуживаемых в момент t сессий k-го потока на передачу трафика реального времени, $k = 1,2,\ldots,n,$ а $d(t)$ – число обслуживаемых в момент t сессий пересылки эластичных данных. Изменение числа организованных сессий связи в зависимости от времени представлено многомерным случайным процессом

$$r(t) = (i_1(t),\ldots,i_n(t),d(t)),$$

определённым на конечном пространстве состояний, включающем вектора $(i_1,\ldots,i_n,d)$ с параметрами $i_1,\ldots,i_n,d$, принимающими значения [20]:

$$\begin{aligned} & i_1 = 0,1,\ldots,\left\lfloor \frac{v}{b_1} \right\rfloor;\ i_2 = 0,1,\ldots,\left\lfloor \frac{v-i_1b_1}{b_2} \right\rfloor; \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & i_n = 0,1,\ldots,\left\lfloor \frac{v-i_1b_1-\ldots-i_{n-1}b_{n-1}}{b_n} \right\rfloor; \\ & d = 0,1,\ldots,v+w-i_1b_1-\ldots-i_nb_n. \end{aligned} \qquad (2.1)$$

Все $(i_1,\ldots,i_n,d)$, представленные (2.1.) образуют множество теоретически возможных состояний Ω. Множество состояний, используемых в модели $S \subset \Omega$ и определяется выбором $\varphi_k(i)$.

При построении модели предполагалось, что все события происходят через независимые случайные промежутки, имеющие экспоненциальное распределение. По этой причине процесс $r(t)$ обладает марковскими свойствами. Пусть $p(i_1,\ldots,i_n,d)$ – стационарная вероятность произвольного состояния модели $(i_1,\ldots,i_n,d) \in S$. В силу марковских свойств модели эту вероятность можно трактовать как долю времени пребывания модели в $(i_1,\ldots,i_n,d)$ и использовать для вычисления показателей обслуживания сессий связи оператора систем наблюдения.

2.4.3. Система уравнений равновесия

Построенный марковский процесс дает возможность определить показатели качества обслуживания поступающих сессий через значения $p(i_1,\ldots,i_n,d)$. Для оценки $p(i_1,\ldots,i_n,d)$ достаточно составить и решить систему линейных уравнений стационарного равновесия, где неизвестными являются величины $p(i_1,\ldots,i_n,d)$ [25]. Делается это достаточно просто. Необходимо приравнять интенсивность выхода $r(t)$ из состояния $(i_1,\ldots,i_n,d)$ к интенсивности перехода $r(t)$ в $(i_1,\ldots,i_n,d)$ из других состояний модели, входящих в пространство состояний S. Поскольку S имеет достаточно много состояний и не обладает специальными свойствами, позволяющими использовать матричные блочные методы, то построенную далее систему уравнений равновесия приходится решать стандартными методами, развитыми в линейной алгебре. По мнению экспертов

[25] наиболее эффективным можно признать метод, суть которого составляет итерационный алгоритм Гаусса-Зейделя или Якоби (см. подраздел 3.2). Для удобства их применения на ЭВМ систему уравнений равновесия рекомендуется представить в виде одного равенства с рекурсивным вычислением коэффициентов при $p(i_1,\ldots,i_n,d)$, зависящих от компонент состояния $(i_1,\ldots,i_n,d)$. Изложим последовательность действий, которая позволит найти это соотношение.

Рассмотрим разные виды событий, которые меняют состояние модели в анализируемой модели:

- Появление запроса k-го потока на организацию сессии пересылки трафика сервисов реального времени, происходящее с интенсивностью λ_k.
- Появление запроса на организацию сессии пересылки эластичного трафика, происходящее с интенсивностью λ_d.
- Освобождение канального ресурса от обслуживания сессии трафика реального времени, относящегося к k-му потоку заявок. При наличии i_k заявок на обслуживании интенсивность события $i_k\mu_k$.
- Освобождение канального ресурса от обслуживания сессии передачи эластичного трафика данных. При наличии d заявок на обслуживании интенсивность события $(v-i_r)\mu_d$, где i_r — число канальных единиц занятых на организацию сессий трафика сервисов реального времени $i_r = i_1b_1+\ldots+i_nb_n$.

Построение системы уравнений статического равновесия для процесса $r(t)$ осуществляется путем исследования возможностей реализации перечисленных выше событий, которые меняют состояние модели. К ним относятся выход модели из состояния $(i_1,\ldots,i_n,d)$ или обратный переход в состояние $(i_1,\ldots,i_n,d)$. Сформулируем выражение для левой части СУР. Для этого найдем условия для реализации всех перечисленных выше событий и определим интенсивность выхода модели из состояния $(i_1,\ldots,i_n,d)$.

Поступление заявки k-го потока на организацию сессии пересылки трафика сервисов реального времени (интенсивность потока заявок λ_k) выводит модель из состояния $(i_1,\ldots,i_n,d)$ с вероятностью единица, если поступившая сессия принимается на обслуживание с учетом возможного изменения скорости передачи данных (т.е. для всех $(i_1,\ldots,i_n,d)$, удовлетворяющих

условию $i_r + d + b_k \le v$). Это ограничение ранее было учтено в определении вероятности допуска заявки к обслуживанию. Таким образом, сессия принимается к обслуживанию с вероятностью $\left(1-\varphi_k\left(i_r+d\right)\right)$, а с дополнительной вероятностью получает отказ и не возобновляется. В рассматриваемой ситуации с интенсивностью $P\left(i_1,\ldots,i_n,d\right)\lambda_k$ выполняется переход из $\left(i_1,\ldots,i_n,d\right)$ в $\left(i_1,\ldots,i_k+1,\ldots,i_n,d\right)$.

Поступление заявки на организацию сессии пересылки эластичного трафика данных (интенсивность потока заявок λ_d) выводит модель из состояния $\left(i_1,\ldots,i_n,d\right)$ с вероятностью единица, если поступившая заявка принимается на обслуживание с учетом возможного изменения скорости передачи данных (т.е. для всех $\left(i_1,\ldots,i_n,d\right)$, удовлетворяющих условию $i_r + d + 1 \le v$). Это ограничение ранее было учтено в определении вероятности допуска заявки к обслуживанию. Таким образом, сессия принимается к обслуживанию с вероятностью $\left(1-\varphi_d\left(i_r+d\right)\right)$, а с дополнительной вероятностью получает отказ и не возобновляется. В рассматриваемой ситуации с интенсивностью $P\left(i_1,\ldots,i_n,d\right)\lambda_d$ совершается переход из $\left(i_1,\ldots,i_n,d\right)$ в $\left(i_1,\ldots,i_n,d+1\right)$.

Завершение сессии k-го потока на пересылку трафика сервисов реального времени (интенсивность $i_k\mu_k$) изменит $\left(i_1,\ldots,i_n,d\right)$ с вероятностью, равной единице, если рассматриваемая сессия находится на обслуживании (т.е. для всех $\left(i_1,\ldots,i_n,d\right)$, для которых $i_k > 0$). В этой ситуации с интенсивностью $P\left(i_1,\ldots,i_n,d\right)i_k\mu_k$ выполняется переход из $\left(i_1,\ldots,i_n,d\right)$ в $\left(i_1,\ldots,i_k-1,\ldots,i_n,d\right)$.

Завершение сессии пересылки эластичного трафика (интенсивность $\left(v-i_r\right)\mu_d$) изменит $\left(i_1,\ldots,i_n,d\right)$ с вероятностью, равной единице, если рассматриваемая сессия находится на обслуживании, т.е. $d > 0$. В этой ситуации с интенсивностью $P\left(i_1,\ldots,i_n,d\right)\left(v-i_r\right)\mu_d$ совершается переход из $\left(i_1,\ldots,i_n,d\right)$ в $\left(i_1,\ldots,i_n,d-1\right)$.

Отдельные слагаемые, составляющие интенсивность выхода модели из состояния $\left(i_1,\ldots,i_n,d\right)$ получены. При составлении аналогичного выражения для правой части, следует таким же образом исследовать рассмотренные выше ситуации, которые приводят к переходу модели в $\left(i_1,\ldots,i_n,d\right)$.

Перемещение $r\left(t\right)$ в $\left(i_1,\ldots,i_n,d\right)$ происходит после появления заявки k-го потока на организацию сессии трафика сервисов реального времени (интенсивность λ_k). Для осуществления

рассматриваемого события надо, чтобы $(i_1,\ldots,i_k-1,\ldots,i_n,d)$ находилось в пространстве состояний S (таким образом, компоненты $(i_1,\ldots,i_k-1,\ldots,i_n,d)$ удовлетворяют условию $i_k>0$). Дополнительно необходимо учесть действие процедуры ограничения доступа. Сессия принимается к обслуживанию с вероятностью $\left(1-\varphi_k\left(i_r+d-b_k\right)\right)$, а с дополнительной вероятностью получает отказ и не возобновляется. В результате сформулированных условий с интенсивностью $P(i_1,\ldots,i_k-1,\ldots,i_n,d)\lambda_k$ выполняется перемещение модели из $(i_1,\ldots,i_k-1,\ldots,i_n,d)$ в $(i_1,\ldots,i_k,\ldots,i_n,d)$.

Перемещение $r(t)$ в $(i_1,\ldots,i_n,d)$ происходит после появления заявки на организацию сессии передачи эластичного трафика (интенсивность λ_d). Для осуществления события надо, чтобы $(i_1,\ldots,i_n,d-1)$ находилось в пространстве состояний S (таким образом, компоненты $(i_1,\ldots,i_n,d-1)$ удовлетворяют условию $d>0$). Дополнительно необходимо учесть действие процедуры ограничения доступа. Сессия принимается к обслуживанию с вероятностью $\left(1-\varphi_d\left(i_r+d-1\right)\right)$, а с дополнительной вероятностью получает отказ и не возобновляется. В результате сформулированных условий с интенсивностью $P(i_1,\ldots,i_n,d-1)\lambda_d$ выполняется перемещение модели из $(i_1,\ldots,i_n,d-1)$ в $(i_1,\ldots,i_n,d)$.

Перемещение $r(t)$ в $(i_1,\ldots,i_n,d)$ происходит в результате завершения сессии k-го потока на пересылку трафика сервисов реального времени (интенсивность $(i_k+1)\mu_k$). Для реализации события необходимо, чтобы состояние $(i_1,\ldots,i_k+1,\ldots,i_n,d)$ принадлежало пространству состояний модели S, т.е. выполнялось условие $i_r+b_k+d\le v$. В результате сформулированных условий с интенсивностью $P(i_1,\ldots,i_k+1,\ldots,i_n,d)(i_k+1)\mu_k$ выполняется перемещение модели из $(i_1,\ldots,i_k+1,\ldots,i_n,d)$ в $(i_1,\ldots,i_k,\ldots,i_n,d)$.

Перемещение $r(t)$ в $(i_1,\ldots,i_n,d)$ происходит в результате завершения сессии передачи эластичного трафика (интенсивность $(v-i_r)\mu_d$). Для реализации события необходимо, чтобы состояние $(i_1,\ldots,i_n,d+1)$ принадлежало пространству состояний модели S, т.е. выполнялось условие $i_r+d+1\le v$. В результате сформулированных условий с интенсивностью

$P(i_1,\ldots,i_n,d+1)(v-i_r)\mu_d$ выполняется перемещение модели из $(i_1,\ldots,i_n,d+1)$ в состояние $(i_1,\ldots,i_n,d)$.

Система уравнений статического равновесия, связывающая ненормированные вероятности и учитывающая события изменения состояний модели и условия их осуществления, имеет вид:

$$P(i_1,\ldots,i_n,d)\Big\{\ \sum_{k=1}^{n}\Big(\lambda_k\left(1-\varphi_k\left(i_r+d\right)\right)+i_k\mu_k I\left(i_k>0\right)\Big)+ \tag{2.2}$$

$$+\lambda_d\left(1-\varphi_d\left(i_r+d\right)\right)+(v-i_r)\mu_d I\left(d>0\right)\ \Big\}=$$

$$=\sum_{k=1}^{n}P(i_1,\ldots,i_k-1,\ldots,i_n,d)\lambda_k\left(1-\varphi_k\left(i_r+d-b_k\right)\right)I\left(i_k>0\right)+$$

$$+P(i_1,\ldots,i_n,d-1)\lambda_d\left(1-\varphi_d\left(i_r+d-1\right)\right)I\left(d>0\right)+$$

$$+\sum_{k=1}^{n}P(i_1,\ldots,i_k+1,\ldots,i_n,d)(i_k+1)\mu_k I\left(i_r+d+b_k\le v\right)+$$

$$+P(i_1,\ldots,i_n,d+1)(v-i_r)\mu_d I\left(i_r+d+1\le v\right).$$

В (2.2) $I(\cdot)$ — индикаторная функция, обозначающая результат осуществления события. Значение $I(\cdot)$ равно единице, если неравенство, указанное в скобках выполняется, и значение $I(\cdot)$ равно нулю, если это неравенство не выполняется.

Для значений $P(i_1,\ldots,i_n,d)$ выполнено условие нормировки

$$\sum_{(i_1,\ldots,i_n,d)\in S}P(i_1,\ldots,i_n,d)=1.$$

Процедура решения системы уравнений равновесия рассмотрен в разделе 3.

2.5. Характеристики качества обслуживания

Предположим, что известны величины $p(i_1,\ldots,i_n,d)$. Они могут быть найдены после решения системы линейных уравнений равновесия (2.2) алгоритмами линейной алгебры. Соответствующая процедура рассмотрена в разделе 3. Формальные определения характеристик через значения входных параметров записываются стандартным образом [25]. При этом используются свойства пуассоновского потока (PASTA) и интерпретация значений стационарных

вероятностей $p(i_1,\ldots,i_n,d)$, которые представляют собой долю времени пребывания модели в $(i_1,\ldots,i_n,d)$. Сформулируем определения характеристик обслуживания k-го потока заявок на организацию сессий трафика сервисов реального времени. Напомним, что через i_r обозначено число единиц ресурса соты, выделенных для обслуживания сессий связи оператора систем наблюдения и занятых в состоянии $(i_1,\ldots,i_n,d)$ на обслуживание сессий трафика сервисов реального времени.

Поскольку входной поток заявок обладает пуассоновскими свойствами, то в силу свойства PASTA, доля π_k запросов k-го потока на открытие сессий связи, потерянных из-за недостатка требуемых b_k единиц ресурса и действия процедуры ограничения допуска, определяется из выражения

$$\pi_k = \sum_{\{(i_1,\ldots,i_n,d)\in S\}} p(i_1,\ldots,i_n,d)\varphi_k(i_r+d).$$

Обычным образом формулируется выражение для вычисления m_k среднего числа канальных единиц, занятых передачей сессий k-го потока трафика сервисов реального времени,

$$m_k = \sum_{(i_1,\ldots,i_n,d)\in S} p(i_1,\ldots,i_n,d) i_k b_k$$

и для вычисления y_k среднего числа обслуживаемых сессий k-го потока

$$y_k = \sum_{(i_1,\ldots,i_n,d)\in S} p(i_1,\ldots,i_n,d) i_k .$$

Понятно, что выполняется соотношение $m_k = y_k b_k$. В силу указанного соотношения достаточно рассчитывать одну из двух характеристик m_k или y_k.

Сформулируем показатели обслуживания сессий передачи эластичных данных. Пусть π_d – доля потерянных запросов на организацию сессий передачи файлов. Соотношение для определения π_d имеет вид

$$\pi_d = \sum_{\{(i_1,\ldots,i_n,d)\in S\}} p(i_1,\ldots,i_n,d)\varphi_d(i_r+d).$$

Обозначим через y_d среднее число сессий передачи эластичных данных находящихся на обслуживании

$$y_d = \sum_{(i_1,\ldots,i_n,d)\in S} p(i_1,\ldots,i_n,d) d .$$

Обозначим через I_d интенсивность окончания обслуживания сессий передачи эластичных данных

$$I_d = \sum_{\{(i_1,\ldots,i_n,d)\in S | d>0\}} p(i_1,\ldots,i_n,d)(v-i_r)\mu_d.$$

Пусть k_d – среднее число канальных единиц (средний объем макроканала), который используется при передаче эластичных данных

$$k_d = \frac{I_d}{y_d \mu_d}.$$

Обозначим через T_d среднее время обслуживания сессий передачи эластичных данных

$$T_d = \frac{y_d}{\lambda_d (1-\pi_d)}.$$

2.6. Соотношения между характеристиками

Введенные характеристики обслуживания сессий связаны между собой соотношениями, которые можно применять для их вычисления или оценки погрешности численных методов решения системы уравнений равновесия. Они следуют из результатов преобразования системы уравнений статистического равновесия (2.2). Часть результатов может быть получена из формулы Литтла. Отмеченные соотношения имеют вид законов сохранения интенсивностей потоков заявок на передачу трафика сервисов реального времени и эластичных данных поступающих, заблокированных и обслуженных выделенным ресурсом соты.

Введенные определения дают возможность рассчитать значения характеристик в ситуации, когда известны значения стационарных вероятностей $p(i_1,\ldots,i_n,d)$. Значения характеристик можно рассчитать и приближенными методами. В этом случае используются либо частные случаи построенной модели, либо используются свойства процесса обслуживания заявок, характерные для предельных значений входных параметров, в частности значений интенсивности поступления заявок [25]. Для k-го потока запросов на передачу сессий трафика сервисов реального времени интенсивности поступающих, заблокированных и обслуженных выделенным ресурсом соты запросов связаны следующими соотношениями:

$$\lambda_k = \lambda_k \pi_k + y_k \mu_k, \qquad k = 1,2,\ldots,n. \tag{2.3}$$

Соотношения (2.3) доказываются стандартными методом [25], в результате умножения уравнения системы уравнений равновесия (2.2), содержащего вероятность $p(i_1,\ldots,i_n,d)$ в левой части, на i_k, $k=1,2,\ldots,n$ и суммирования полученных соотношений по всем состояниям $(i_1,\ldots,i_n,d)\in S$. Соотношения (2.3) означают, что интенсивность потока заявок на передачу трафика сервисов реального времени равна сумме интенсивностей соответствующих заявок, заблокированных из-за нехватки ресурса и обслуженных выделенным ресурсом соты.

Аналогичным способом можно установить подобное соотношение и для интенсивностей поступления и обслуживания сессий на передачу эластичного трафика. Оно имеет вид

$$\lambda_d = \lambda_d \pi_d + I_d. \qquad (2.4)$$

Как уже было сказано, приведенные соотношения позволяют установить альтернативные формулы, которые упрощают оценку введенных характеристик. В частности, из (2.4) и определения характеристик следует соотношение

$$T_d = \frac{1}{\mu_d} \cdot \frac{1}{k_d}.$$

Таким образом, среднее время передачи файла равно среднему времени передачи файла одним каналом, деленному на среднее число каналов, составляющих один макроканал.

2.7. Выводы по результатам второго раздела

1. Построена математическая модель процесса совместного обслуживания выделенным ресурсом изолированной соты сети стандарта LTE произвольного числа сессий связи на обслуживание трафика сессий реального времени и потока сессий на пересылку эластичного трафика данных. Предполагается, что источниками сессий являются видеокамеры и разнообразные устройства телеметрии, принадлежащие оператору систем наблюдения, использующему ресурс соты для организации своей работы. В модели трафик сессий реального времени имеет преимущество в использовании выделенного ресурса перед трафиком сессий пересылки эластичны данных. Это преимущество характеризуется в уменьшении скорости пересылки данных до некоторого заранее оговоренного минимального значения, которое соответствует нижней границе скорости передачи эластичных данных. При появлении свободного ресурса передачи информации скорость пересылки данных возрастает. Для создания условий по дифференцированному

обслуживанию поступающих гетерогенных информационных потоков в модели используется процедура ограничения доступа в занятии ресурса. Выбор ограничений зависит от типа сессий связи и от общего уровня занятости выделенного ресурса всеми поступающими информационными потоками.

2. Построенная модель может быть использована:
   - для анализа зависимости характеристик качества обслуживания поступающих запросов от параметров возникающих информационных потоков и условий допуска сессий связи к занятию ресурса;
   - для анализа действия разного рода процедур, конечной целью которых является повышение эффективности занятия ресурса передачи узлов доступа и создание условий по дифференцированному обслуживанию поступающих сессий связи.
3. Построенная модель позволяет определить характеристики качества обслуживания через величины входных параметров поступающих запросов и значения стационарных вероятностей состояний модели. Для сессий передачи трафика реального времени — это доля потерянных запросов, среднее число сессий, находящихся на обслуживании, и средний объем занятого ресурса. Для сессий передачи эластичного трафика данных — это доля потерянных запросов, среднее время доставки сообщения, среднее число сессий, находящихся на обслуживании, и среднее число единиц ресурса, используемых для передачи одного сообщения. Преобразование СУР дает возможность установить соотношения между характеристиками построенной модели, которые носят характер законов сохранения интенсивностей потоков сессий поступающих, получивших отказ и обслуженных на выделенном ресурсе. Полученные соотношения могут быть использованы для косвенной оценки характеристик при организации их измерения или вычисления.
4. Для точной оценки характеристик построенной модели используется алгоритм, основанный на составлении и решении СУР методами линейной алгебры. Для их реализации удобно представить СУР в виде одного соотношения. В процессе реализации цикла по изменению целочисленных компонент состояния модели получаются отдельные уравнения системы.

Раздел 3

Разработка и анализ алгоритмов оценки характеристик качества совместного обслуживания трафика реального времени и эластичного трафика данных с резервированием

3.1. Введение к разделу 3

Формулы для вычисления показателей качества совместного пересылки трафика реального времени и эластичного трафика данных выделенным ресурсом узла доступа сети подвижной связи основаны на использовании значений стационарных вероятностей состояний модели. Стандартный способ их оценки заключается в составлении и решении системы линейных уравнений равновесия. Обзор соответствующих процедур рассмотрен в подразделе 3.2. По мнению экспертов, лучше всего для этих целей подходит итерационный метод Гаусса-Зейделя. Основные его положения также обсуждаются в подразделе 3.2.

Реализация процедуры слайсинга предполагает интеграцию информационных потоков с примерно одинаковыми свойствами с целью их последующего обслуживания в отдельном слайсе. Действуя таким образом, можно определить ресурс отдельных слайсов, который достаточен для создания условий по дифференцированному обслуживанию только сессий трафика реального времени и только эластичных данных. Оценка характеристик обслуживания сессий трафика реального времени и эластичных данных в отдельных слайсах рассмотрена соответственно в подразделах 3.3 и 3.4.

Для практических приложений представляет интерес модель узла доступа сети подвижной связи, в которой исследуется процесс совместной пересылки двух потоков сессий трафика реального времени и одного потока сессий на передачу эластичных данных. В этом частном случае характеристики качества обслуживания могут быть найдены для любых значений входных параметров с использованием решения системы уравнений равновесия. Изложение алгоритма решения и описание модели приведены в подразделе 3.5.

В разделе 3.6 сформулированы выводы по результатам третьего раздела.

3.2. Решение системы уравнений равновесия

3.2.1. Общие положения

В общем виде систему уравнений статистического равновесия (2.2) можно представить в виде выражения

$$\begin{gathered} a_{1,1}P_1 + a_{1,2}P_2 + \ldots + a_{1,q}P_q = 0; \\ a_{2,1}P_1 + a_{2,2}P_2 + \ldots + a_{2,q}P_q = 0; \\ a_{l,1}P_1 + a_{l,2}P_2 + \ldots + a_{l,q}P_q = 0; \quad l = 3,4,\ldots,q-1; \\ a_{q,1}P_1 + a_{q,2}P_2 + \ldots + a_{q,q}P_q = 0. \end{gathered} \tag{3.1}$$

В матричном виде система уравнений равновесия (3.1) может быть представлена в виде соотношения

$$AP = 0. \tag{3.2}$$

В (3.2) $A = \left\| a_{l,j} \right\|_{l,j=1}^{q}$ — матрица системы уравнений равновесия после перемещения неизвестных в левую часть. Структура матрицы A записывается в следующем виде:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,q} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q,1} & a_{q,2} & \ldots & a_{q,q} \end{pmatrix}.$$

В приведенном выражении для матрицы символ q означает число неизвестных в СУР (2.2), которое равно числу элементов в используемом пространстве состояний S (2.1). Вектор $P = \left(P_1, P_2, \ldots, P_q\right)$ — представляет из себя вектор неизвестных вероятностей $p\left(i_1, \ldots, i_n, d\right)$, где $\left(i_1, \ldots, i_n, d\right) \in S$. Значение $p\left(i_1, \ldots, i_n, d\right)$ представляет из себя долю времени пребывания модели в $\left(i_1, \ldots, i_n, d\right)$ и может применяться для вычисления показателей совместного обслуживания сессий связи.

Будем предполагать, что состояния $\left(i_1, \ldots, i_n, d\right) \in S$ занумерованы в лексикографическом порядке. Соответствующая последовательность определяется порядком перебора состояний в (2.1) и определяется следующим циклом

$$i_1 = 0,1,\ldots,\left\lfloor \frac{v}{b_1} \right\rfloor;$$

$$i_2 = 0,1,\ldots,\left\lfloor \frac{v - i_1 b_1}{b_2} \right\rfloor;$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$i_n = 0,1,\ldots,\left\lfloor \frac{v - i_1 b_1 - \ldots - i_{n-1} b_{n-1}}{b_n} \right\rfloor;$$

$$d = 0,1,\ldots,v + w - i_1 b_1 - \ldots - i_n b_n.$$

Символ P_j определяет вероятность состояния с номером j. Ранг системы линейных уравнений (3.2) на единицу меньше, чем число уравнений, поэтому значения вероятностей состояний находятся из решения (3.2) после применения процедуры нормировки.

В исследуемом случае матрица A имеет большую размерность, достигающую нескольких миллионов строк и столбцов. Из-за этого свойства затрудняется использование стандартных алгоритмов решения (2.2), встроенных в пакеты программ типа MatLab или MathCad. Также матрица (2.2) в общем виде не обладает специальными свойствами, которые позволяют преобразовать (2.2) к удобному для расчетов рекурсивному алгоритму. Сделать это можно только в частных случаях, когда рассматривается модель обслуживания в узле доступа сети подвижной связи только сессий реального времени и только сессий эластичных данных. Структура моделей, формулы для оценки характеристик качества обслуживания заявок и изложение алгоритмов их оценки рассмотрено соответственно в подразделе 3.3 и в подразделе 3.4.

Учитывая тот факт, что в системе уравнений (2.2) матрица A имеет большое число нулей, а ненулевые элементы матрицы (2.2) связаны с компонентами состояния $(i_1,\ldots,i_n,d)$ простыми алгебраическими зависимостями, то представляется удобным для решения (2.2) использовать итерационные методы [25]. Наиболее простыми в реализации являются итерационные процедуры Якоби и Гаусса-Зейделя. Особенности их применения для решения (2.2) рассмотрены в следующем подразделе.

3.2.2. Итерационные методы решения систем уравнений равновесия

Рассмотрим в общем виде решение (2.2) итерационными методами Якоби и Гаусса-Зейделя. Обозначим вектор s-го приближения, полученного итерационным алгоритмом, к вектору $P=\left(P_1,P_2,\ldots,P_q\right)$ неизвестных вероятностей состояний модели узла сети через $P^{(s)}=\left(P_1^{(s)},P_2^{(s)},\ldots,P_q^{(s)}\right)$. В процессе применения итерационного алгоритма Якоби вектор $(s+1)$-го приближения получается из векторов s-го приближения с использованием следующих рекурсий:

$$P_1^{(s+1)}=-\frac{1}{a_{1,1}}\left(a_{1,2}P_2^{(s)}+a_{1,3}P_3^{(s)}+\ldots+a_{1,q}P_q^{(s)}\right); \tag{3.3}$$

$$P_2^{(s+1)}=-\frac{1}{a_{2,2}}\left(a_{2,1}P_1^{(s)}+a_{2,3}P_3^{(s)}+\ldots+a_{2,q}P_q^{(s)}\right);$$

$$P_l^{(s+1)}=-\frac{1}{a_{l,l}}\left(\sum_{j=1}^{l-1}a_{l,j}P_j^{(s)}+\sum_{j=l+1}^{q}a_{l,j}P_j^{(s)}\right),\quad l=3,4,\ldots,q-1;$$

$$P_q^{(s+1)}=-\frac{1}{a_{q,q}}\left(a_{q,1}P_1^{(s)}+a_{q,2}P_2^{(s)}+\ldots+a_{q,q-1}P_{q-1}^{(s)}\right).$$

В результате использования итерационного алгоритма Гаусса-Зейделя вектор $(s+1)$-го приближения получается из векторов s-го и $(s+1)$-го приближений с использованием следующих рекурсий:

$$P_1^{(s+1)}=-\frac{1}{a_{1,1}}\left(a_{1,2}P_2^{(s)}+a_{1,3}P_3^{(s)}+\ldots+a_{1,q}P_q^{(s)}\right); \tag{3.4}$$

$$P_2^{(s+1)}=-\frac{1}{a_{2,2}}\left(a_{2,1}P_1^{(s+1)}+a_{2,3}P_3^{(s)}+\ldots+a_{2,q}P_q^{(s)}\right);$$

$$P_l^{(s+1)}=-\frac{1}{a_{l,l}}\left(\sum_{j=1}^{l-1}a_{l,j}P_j^{(s+1)}+\sum_{j=l+1}^{q}a_{l,j}P_j^{(s)}\right),\quad l=3,4,\ldots,q-1;$$

$$P_q^{(s+1)}=-\frac{1}{a_{q,q}}\left(a_{q,1}P_1^{(s+1)}+a_{q,2}P_2^{(s+1)}+\ldots+a_{q,q-1}P_{q-1}^{(s+1)}\right).$$

В отличие от алгоритма Якоби при реализации алгоритма Гаусса-Зейделя при расчете $P_l^{(s+1)}$, $l=1,2,\ldots.q$, применяются уже найденные элементы $(s+1)$-го приближения

$P_j^{(s+1)}$, $j=1,2,\ldots,l-1$, и имеющиеся компоненты s-го приближения $P_j^{(s)}$, $j=l+1,\ l+2,\ldots,q$. При реализации итерационных методов необходимо определить начальное приближение, критерий окончания итерационного цикла и выяснить вопросы наличия или отсутствия сходимости.

В качестве начального 0-го приближения к вектору неизвестных вероятностей можно взять любое приближение с положительными компонентами $P^{(0)}=\left(P_1^{(0)},P_2^{(0)},\ldots,P_q^{(0)}\right)$. Для простоты обычно используют вектор $P^{(0)}=(1,1,\ldots,1)$.

Применение итерационных алгоритмов к решению (2.2) может приводить к сходящейся или к не сходящейся последовательности приближений. Сходимость алгоритма определяется из величины нормированной разности между двумя последовательными приближениями к значениям неизвестных вероятностей. Выбирается значение контрольного параметра ε порядка $10^{-8}-10^{-10}$ и проверяется справедливость соотношения

$$\frac{\left|P_1^{(s+1)}-P_1^{(s)}\right|+\ldots+\left|P_q^{(s+1)}-P_q^{(s)}\right|}{P_1^{(s+1)}+\ldots+P_q^{(s+1)}}\le\varepsilon. \qquad (3.5)$$

Если относительная разность, взятая по модулю, лежит в интервале $10^{-8}\ldots10^{-10}$, то итерационный цикл заканчивается. После этого применяются косвенные средства проверки сходимости. К ним относится контроль за выполнением (2.3), (2.4), связывающих интенсивности потоков сессий связи, поступивших в выделенный ресурс узла доступа, получивших отказ и попавших на обслуживание.

3.2.3. Сходимость итерационной процедуры

Вопрос сходимости является важным аспектом применения итерационных методов решения (2.2). Можно привести примеры, когда итерационные процедуры Якоби и Гаусса-Зейделя не сходятся, если их применять к решению (2.2). Для метода Гаусса-Зейделя новые приближения применяются сразу же по мере вычисления. В алгоритме Якоби они не используются до следующей итерации. В результате, в большинстве случаев алгоритм Гаусса-Зейделя сходится быстрее нежели алгоритм Якоби. По этой причине в дальнейшем для решения (2.2) будем использовать алгоритм Гаусса-Зейделя. Используя результаты [25], покажем, что можно внести некоторые изменения в структуру матрицы (3.1), которые приведут к сходящейся процедуре.

Выполним следующие действия. Положим одну из неизвестных в (3.1), например P_1, равной единице и уберем из рассмотрения первое уравнение. В результате (3.1) преобразуется к виду:

$$a_{2,2}P_2 + a_{2,3}P_3 + \ldots + a_{2,q}P_q = -a_{2,1}; \tag{3.6}$$

$$a_{3,2}P_2 + a_{3,3}P_3 + \ldots + a_{3,q}P_q = -a_{3,1};$$

$$a_{l,2}P_2 + a_{l,3}P_3 + \ldots + a_{l,q}P_q = -a_{l,1}; \quad l = 3,4,\ldots,q-1;$$

$$a_{q,2}P_2 + a_{q,3}P_3 + \ldots + a_{q,q}P_q = -a_{q,1}.$$

В результате проделанных преобразований сумма элементов по столбцам матрицы хотя для одного столбца будет строго больше нуля. Это свойство означает наличие слабого диагонального преобладания, которое обеспечивает сходимость соответствующей итерационной процедуры [25]. Для системы уравнений (3.6) рекурсия Гаусса-Зейделя приобретает вид:

$$P_2^{(s+1)} = -\frac{1}{a_{2,2}}\left(a_{2,3}P_3^{(s)} + a_{2,4}P_4^{(s)} + \ldots + a_{2,q}P_q^{(s)} + a_{2,1}\right); \tag{3.7}$$

$$P_3^{(s+1)} = -\frac{1}{a_{3,3}}\left(a_{3,2}P_2^{(s+1)} + a_{4,3}P_3^{(s)} + \ldots + a_{3,q}P_q^{(s)} + a_{3,1}\right);$$

$$P_l^{(s+1)} = -\frac{1}{a_{l,l}}\left(\sum_{j=1}^{l-1} a_{l,j}P_j^{(s+1)} + \sum_{j=l+1}^{q} a_{l,j}P_j^{(s)} + a_{l,1}\right), \quad l = 3,4,\ldots,q-1;$$

$$P_q^{(s+1)} = -\frac{1}{a_{q,q}}\left(a_{q,2}P_2^{(s+1)} + a_{q,3}P_3^{(s+1)} + \ldots + a_{q,q-1}P_{q-1}^{(s+1)} + a_{q,1}\right).$$

Как уже было сказано, проделанное изменение алгоритма решения СУР приводит к сходящемуся алгоритму. Однако, как показали результаты численных расчетов, скорость сходимости уменьшается по сравнению с реализацией (3.4) на несколько порядков. По этой причине применение (3.7) ограничено крайне редкими случаями отсутствия сходимости в (3.4).

3.2.4. Формулировка итерационной процедуры

Приведем итерационные соотношения для решения (2.2). Для простоты рассмотрим итерационный метод в реализации (3.4). Аналогичным образом можно записать рекурсию и для реализации (3.7). Введем обозначение $L(i_1,\ldots,i_n,d)$, для коэффициента при левой части (2.2) у вероятности $p(i_1,\ldots,i_n,d)$, где $(i_1,\ldots,i_n,d) \in S$. Величина $L(i_1,\ldots,i_n,d)$ определяется из равенства

$$L(i_1,\ldots,i_n,d)=\Big\{\ \sum_{k=1}^{n}\Big(\lambda_k\left(1-\varphi_k\left(i_r+d\right)\right)+i_k\mu_k I\left(i_k>0\right)\Big)+ \tag{3.8}$$

$$+\lambda_d\left(1-\varphi_d\left(i_r+d\right)\right)+\left(v-i_r\right)\mu_d I\left(d>o\right)\ \Big\}.$$

Рекурсивное соотношение имеет вид:

$$P^{(s+1)}\left(i_1,\ldots,i_n,d\right)=\frac{1}{L\left(i_1,\ldots,i_n,d\right)}\times \tag{3.9}$$

$$\times\Big\{\ \sum_{k=1}^{n}P^{(s,s+1)}\left(i_1,\ldots,i_k-1,\ldots,i_n,d\right)\lambda_k\left(1-\varphi_k\left(i_r+d-b_k\right)\right)I\left(i_k>0\right)+$$

$$+P^{(s,s+1)}\left(i_1,\ldots,i_n,d-1\right)\lambda_d\left(1-\varphi_d\left(i_r+d-1\right)\right)I\left(d>0\right)+$$

$$+\sum_{k=1}^{n}P^{(s,s+1)}\left(i_1,\ldots,i_k+1,\ldots,i_n,d\right)\left(i_k+1\right)\mu_k I\left(i_r+d+b_k\le v\right)+$$

$$+P^{(s,s+1)}\left(i_1,\ldots,i_n,d+1\right)\left(v-i_r\right)\mu_d I\left(i_r+d+1\le v\right)\ \Big\}.$$

Запись $P^{(s,s+1)}\left(i_1,\ldots,i_n,d\right)$ означает использование новых значений приближений сразу же по мере их получения. После того как сходимость установлена значения $P^{(s+1)}\left(i_1,\ldots,i_n,d\right)$ нормируются. Для этого используется соотношение:

$$P\left(i_1,\ldots,i_n,d\right)=\frac{P^{(s+1)}\left(i_1,\ldots,i_n,d\right)}{\sum\limits_{(i_1,\ldots,i_n,d)\in S}P^{(s+1)}\left(i_1,\ldots,i_n,d\right)}.$$

Здесь $(s+1)$— номер последней итерации в (3.9). Численный анализ сходимости итерационной схемы Гаусса-Зейделя рассмотрен в подразделе 3.5.3. Разработанная вычислительная схема использовалась для точной оценки значений характеристик совместного обслуживания трафика реального времени и эластичных данных и сравнения различных сценариев распределения ресурса. Эта задача рассмотрена в заключительном разделе работы.

Предположим, что ресурс, выделенный для обслуживания сессий связи оператора систем наблюдения, разделен на два слайса в соответствии с типом трафика. Один слайс обслуживает сессии трафика реального времени, другой — сессии эластичного трафика данных. Получим алгоритмы оценки характеристик в каждом слайсе, рассмотренном отдельно. Вначале построим процедуру вычисления характеристик обслуживания сессий трафика реального времени.

3.3. Оценка характеристик обслуживания сессий трафика реального времени в слайсе

3.3.1. Модель входного потока

Обозначим через v общее число имеющихся единиц ресурса передачи информации, выделенного в виде изолированного слайса для обслуживания сессий трафика реального времени. Обозначим через i — число единиц ресурса занятых обслуживанием запросов в момент поступления запроса k-го потока в слайсе. Модель поступления и обслуживания сессий связи k-го потока, $k = 1,\ldots,n$ выглядит следующим образом:

- поступление сессий происходит через случайное время, имеющее экспоненциальное распределение с параметром λ_k;
- длительность интервала времени обслуживания сессии имеет экспоненциальное распределение со средним $1/\mu_k$;
- для организации сессии одновременно требуется b_k единиц ресурса;
- сессия связи организуется с вероятностью $1-\varphi_k(i)$, а с дополнительной вероятностью $\varphi_k(i)$ этого не происходит.

Построенная математическая модель использования ресурса соты сети LTE, выделенного для обслуживания сессий трафика реального времени, показана на рисунке 3.1 [23].

Построим случайный процесс, который будет описывать процесс поступления и обслуживания рассматриваемых запросов. Пусть $i_k(t)$ — число сессий k-го потока, находящихся на обслуживании в моменте времени t. Изменение числа организованных сессий связи в зависимости от времени представлено многомерным случайным процессом, $r(t) = (i_1(t),\ldots,i_n(t))$, определённым на пространстве состояний Ω. Используемое пространство состояний $S \subset \Omega$ определяется выбором функции блокировки $\varphi_k(i)$. Процесс $r(t)$ будет марковским и можно показать, что время пребывания $r(t)$ в любом $(i_1,\ldots,i_n) \in S$ распределено экспоненциально с известным параметром и далее с известными вероятностями происходит переход $r(t)$ в другие состояния из S.

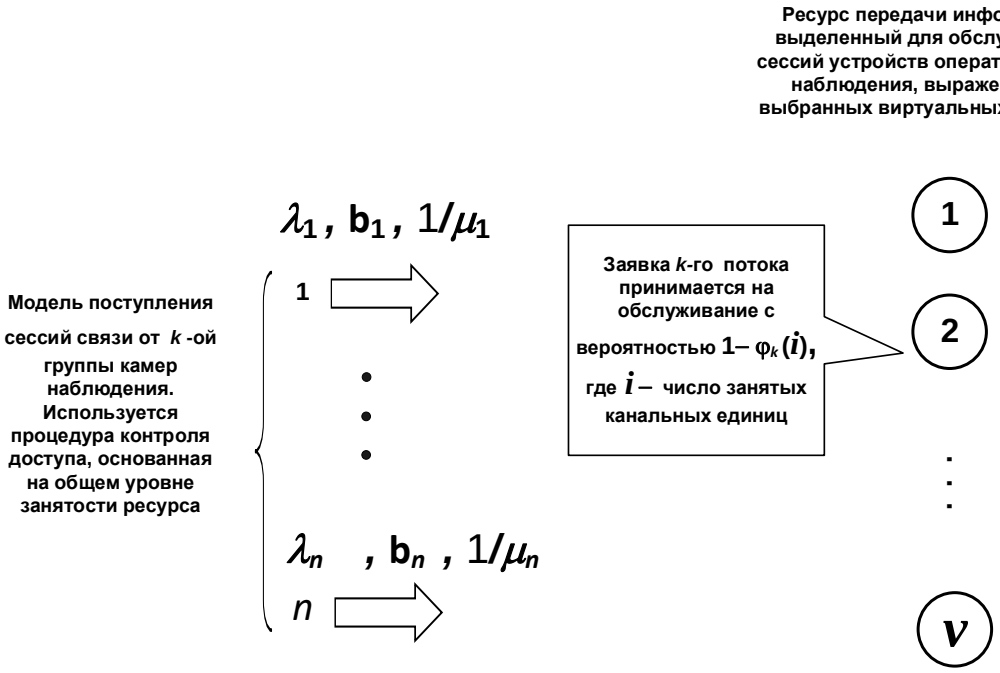


Рисунок 3.1 — Математическая модель использования ресурса соты сети LTE, выделенного для обслуживания сессий трафика реального времени

3.3.2. Характеристики обслуживания сессий

Пусть $p(i_1,\ldots,i_n)$ - стационарная вероятность состояния $(i_1,\ldots,i_n)$. Качество обслуживания сессий k-го потока определим долей π_k потерянных сессий. Значение π_k находится из отношения интенсивности потерянных сессий к интенсивности поступивших сессий

$$\pi_k = \sum_{(i_1,\ldots,i_n)\in S} p(i_1,\ldots,i_n)\varphi_k(i).$$

Загрузку ресурса обслуживанием сессий k-го потока оценим средним числом m_k занятых виртуальных каналов. Для расчета m_k используется соотношение

$$m_k = \sum_{(i_1,\ldots,i_n)\in S} p(i_1,\ldots,i_n) i_k b_k .$$

Среднее число обслуживаемых сессий k-го потока находится из соотношения $y_k = m_k / b_k$. Для оценки введенных характеристик в соответствии по введенным определениям надо найти $p(i_1,\ldots,i_n)$.

Стандартным образом (см. процедуру составления системы уравнений равновесия (2.2) для совместной модели обслуживания сессий оператора систем наблюдения) для построенной модели слайса можно выписать систему уравнений равновесия. Для всех $(i_1,\ldots,i_n)\in\Omega$

$$P(i_1,\ldots,i_n)\sum_{k=1}^{n}\left(\lambda_k\left(1-\varphi_k(i)\right)+i_k\mu_k\right)= \tag{3.10}$$

$$=\sum_{k=1}^{n}P(i_1,\ldots,i_k-1,\ldots,i_n)\lambda_k\left(1-\varphi_k(i-b_k)\right)I(i_k>0)+$$

$$+\sum_{k=1}^{n}P(i_1,\ldots,i_k+1,\ldots,i_n)(i_k+1)\mu_k I(i+b_k\le v).$$

В приведенной записи: $i=i_1b_1+\ldots+i_nb_n$ — ресурс, занятый обслуживанием всех сессий трафика реального времени, а $I(\cdot)$ — индикаторная функция события. Найденные величины $P(i_1,\ldots,i_n)$ необходимо нормировать. Значения $P(i_1,\ldots,i_n)$ не обладают свойством мультипликативности, поэтому определяются численными методами из решения СУР. Для этих целей лучше всего подходит итерационный алгоритм Гаусса-Зейделя (см. подраздел 3.2.2). Приведем запись соответствующей рекурсии.

3.3.3. Оценка характеристик

Чтобы иметь возможность вести расчет характеристик для всех практически интересных значений числа единиц ресурса придется ограничиться рассмотрением случая $n=3$, т.е. наличия только трех потоков сессий реального времени. В этом случае система уравнений равновесия (3.10) приобретает вид

$$P(i_1,i_2,i_3)\left(\lambda_1\left(1-\varphi_1(i)\right)+\lambda_2\left(1-\varphi_2(i)\right)+\lambda_3\left(1-\varphi_3(i)\right)+i_1\mu_1+i_2\mu_2+i_3\mu_3\right)= \tag{3.11}$$

$$=P(i_1-1,i_2,i_3)\lambda_1\left(1-\varphi_1(i-b_1)\right)I(i_1>0)+P(i_1,i_2-1,i_3)\lambda_2\left(1-\varphi_2(i-b_2)\right)I(i_2>0)+$$

$$+P(i_1,i_2,i_3-1)\lambda_3\left(1-\varphi_3(i-b_3)\right)I(i_3>0)+P(i_1+1,i_2,i_3)(i_1+1)\mu_1 I(i+b_1\le v)+$$

$$+P(i_1,i_2+1,i_3)(i_2+1)\mu_2 I(i+b_2\le v)+P(i_1,i_2,i_3+1)(i_3+1)\mu_3 I(i+b_3\le v).$$

Здесь $i=i_1b_1+i_2b_2+i_3b_3$ и значения $P(i_1,i_2,i_3)$ удовлетворяют условию нормировки

$$\sum_{(i_1,i_2,i_3)\in S}P(i_1,i_2,i_3)=1.$$

Чтобы получить все уравнения СУР достаточно сформировать цикл по i_1, i_2 и i_3 в следующем виде $i_1 = 0,1,\ldots,\left\lfloor \frac{v}{b_1} \right\rfloor$; $i_2 = 0,1,\ldots,\left\lfloor \frac{v - i_1 b_1}{b_2} \right\rfloor$; $i_3 = 0,1,\ldots,\left\lfloor \frac{v - i_1 b_1 - i_2 b_2}{b_3} \right\rfloor$ и подставить значения i_1, i_2 и i_3 в приведенное соотношение.

Рекурсивные соотношения алгоритма Гаусса-Зейделя (см. подраздел 3.2.2) найденные используя систему уравнений статического равновесия (3.11) выглядит следующем образом

$$P^{(s+1)}(i_1,i_2,i_3) = \frac{1}{\left(\lambda_1\left(1-\varphi_1(i)\right)+\lambda_2\left(1-\varphi_2(i)\right)+\lambda_3\left(1-\varphi_3(i)\right)+i_1\mu_1+i_2\mu_2+i_3\mu_3\right)}\times \tag{3.12}$$

$$\times\Big(P^{(s,s+1)}(i_1-1,i_2,i_3)\lambda_1\left(1-\varphi_1(i-b_1)\right)I(i_1>0)+P^{(s,s+1)}(i_1,i_2-1,i_3)\lambda_2\left(1-\varphi_2(i-b_2)\right)I(i_2>0)+$$

$$+P^{(s,s+1)}(i_1,i_2,i_3-1)\lambda_3\left(1-\varphi_3(i-b_3)\right)I(i_3>0)+P^{(s,s+1)}(i_1+1,i_2,i_3)(i_1+1)\mu_1 I(i+b_1\le v)+$$

$$+P^{(s,s+1)}(i_1,i_2+1,i_3)(i_2+1)\mu_2 I(i+b_2\le v)+P^{(s,s+1)}(i_1,i_2,i_3+1)(i_3+1)\mu_3 I(i+b_3\le v)\Big).$$

Полученное выражение легко представить на любом алгоритмическом языке программирования. Индикаторная функция записывается в виде условного оператора. Это значительно упрощает реализацию алгоритма на вычислительных машинах.

3.3.4. Приближенная оценка характеристик

Условия практической реализации модели во многом определяются временем, которое тратится на вычисление характеристик. В этой связи область применения точных методов, основанных на решении СУР, ограничена исследованием погрешности инженерных методик, которые несут основную нагрузку при решении задач планирования пропускной способности сети. Инженерные методики строятся на основе приближённых алгоритмов анализа модели. Рассмотрим один из способов приближенной оценки показателей совместного обслуживания запроса на передачу сессий тяжелого трафика для модели мультисервисного узла с резервированием. Идея метода основана на предполагаемой возможности выполнения для исследуемой модели соотношений детального баланса [25].

Сформулируем новые расчетные выражения для введенных характеристик. Обозначим через $p(i)$ долю времени пребывания узла в состояниях, когда занято ровно i единиц ресурса

$$p(i) = \sum_{i_1 b_1 + \ldots + i_n b_n = i} p(i_1, \ldots, i_n), \quad i = 0,1,\ldots,v.$$

Соотношение для расчёта π_k выглядит так

$$\pi_k = \sum_{(i_1,\ldots,i_n)\in S} p(i_1,\ldots,i_n)\varphi_k(i) = \sum_{i=0}^{v} \sum_{i_1 b_1 + \ldots + i_n b_n = i} p(i_1,\ldots,i_n)\varphi_k(i) = \sum_{i=0}^{v} p(i)\varphi_k(i).$$

Оценка m_k, $k = 1,\ldots,n$ с помощью $p(i)$ следует из формулы Литтла: $m_k = a_k b_k (1-\pi_k)$, где $a_k = \lambda_k / \mu_k$. Таким образом, чтобы рассчитать введенные характеристики достаточно построить способ оценки $p(i)$, $i = 0,1,\ldots,v$.

Положим в основу приближенного алгоритма следующее предположение. Будем считать, что для построенной модели слайса с резервированием выполняются соотношения детального баланса. Приходим к такому результату (знак ˆ, означает приближенный вид равенства)

$$\hat{p}(i_1,\ldots,i_k - 1,\ldots,i_n)\lambda_k\left(1-\varphi_k(i-b_k)\right) = \hat{p}(i_1,\ldots,i_k,\ldots,i_n) i_k \mu_k.$$

Введем обозначение $a_k = \lambda_k / \mu_k$. После суммирования приведенного выше соотношения по всем $(i_1,\ldots,i_n) \in S$ таким, что $i_1 b_1 + \ldots + i_n b_n = i$, Находим рекурсивное выражение, связывающее значения $\hat{P}(i)$, $i = 0,1,\ldots,v$

$$\hat{P}(i) = \frac{1}{i} \times \sum_{k=1}^{n} a_k b_k \hat{P}(i-b_k)\left(1-\varphi_k(i-b_k)\right).$$

Величины $\hat{P}(i)$ находятся с использованием рекурсивного алгоритма, аналогичного полученному для мультисервисной модели Эрланга [25].

1. Положим значение $\hat{P}(0) = 1$.
2. Выразим значения $\hat{P}(i)$ через $\hat{P}(0)$, используя соотношение

$$\hat{P}(i) = \frac{1}{i} \times \sum_{k=1}^{n} a_k b_k \hat{P}(i-b_k)\left(1-\varphi_k(i-b_k)\right) \qquad (3.13)$$

и последовательно увеличивая i от 1 до v. При фиксированном i значения $\hat{P}(i-b_k)$, $k = 1,\ldots,n$ либо уже представлены через $\hat{P}(0)$ (для $i - b_k \geq 0$), либо равны 0 (для $i - b_k < 0$). Рекурсия реализуется для всех i.

3. Находим величину нормировочной константы $N = \sum_{i=0}^{v} \hat{P}(i)$.

4. Определяем величины $\hat{p}(i) = \dfrac{\hat{P}(i)}{N}$, $i = 0,1,\ldots,v$.

5. Приводим формулы для вычисления характеристик анализируемых потоков

$$\hat{\pi}_k = \sum_{i=0}^{v} \hat{p}(i)\varphi_k(i),\ m_k = a_k b_k (1-\pi_k),\ k = 1,\ldots,n. \qquad (3.14)$$

6. Рекурсивный алгоритм отличается легкостью реализации, однако требует анализа погрешности оценки характеристик исследуемой модели.

3.3.5. Анализ погрешности приближенной оценки характеристик

Проведем исследование точности метода. Рассмотрим модель слайса с параметрами: $v = 200$ к.е.; $n = 3$; $b_1 = 1$ к.е.; $b_2 = 5$ к.е.; $b_3 = 20$ к.е.; $\lambda_k = v\rho / nb_k$; $\mu_k = 1$; $k = 1,2,3$. Предложенная нагрузка в эрлангах $a_k = v\rho / nb_k$ Эрл, $k = 1,2,3$. Исследуем зависимость погрешности оценки характеристик модели от изменения ρ интенсивности предложенного трафика на канал. Величина $\rho = \dfrac{a_1 b_1 + a_2 b_2 + a_3 b_3}{v}$ меняется от 0,5 до 1,5. Воспользуемся резервированием и сделаем равными потери заявок всех трех потоков. Для этого достаточно определить функции $\varphi_k(i)$ из следующих соотношений $\varphi_k(i) = 1$, $k = 1,2,3$, $i = v - b_3 + 1, v - b_3 + 2, \ldots, v$ и $\varphi_k(i) = 0$, $k = 1,2,3$, $i = 0,1,\ldots,v - b_3$. На рисунках 3.2 и 3.3 показаны результаты точного вычисления выравненных значений доли потерянных заявок $\pi_{c,k} = \pi$, $k = 1,2,3$ (рисунок 3.2) и среднего использования единицы ресурса $\delta = \dfrac{1}{v}(m_1 + m_2 + m_3)$ (рисунок 3.3), полученные в результате решения системы уравнений (3.11) итерационным методом Гаусса-Зейделя (3.12) и приближенные значения этих же характеристик, полученные после реализации рекурсии (3.13). Кривые, соответствующие точным значениям имеют номер 1, кривые, соответствующие приближенным значениям имеют номер 2.

Приведенные результаты показывают, что приближенный метод отличается высокой точностью. Погрешность оценки незначительно увеличивается в области перегрузки, когда $\rho > 1$.

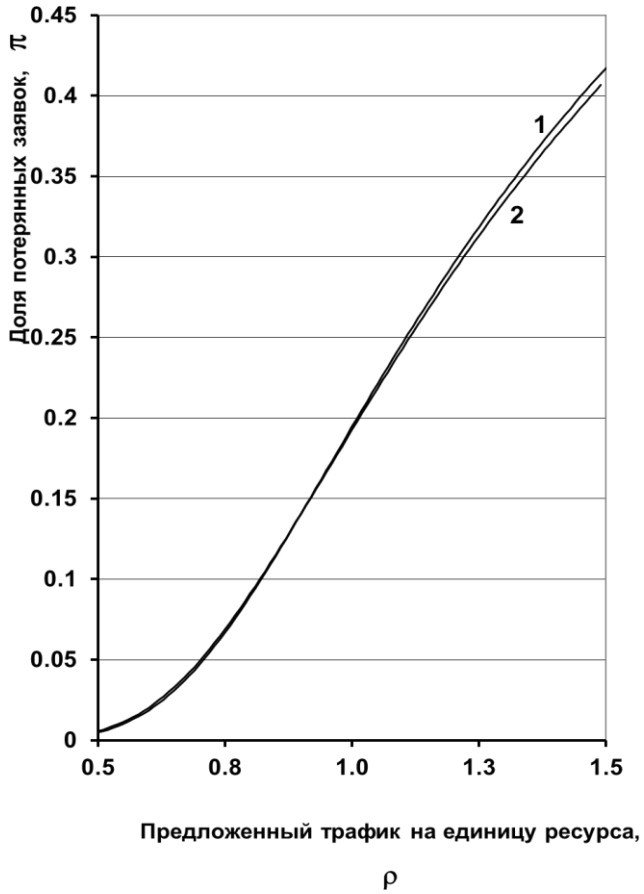


Рисунок 3.2 — Точный и приближенный расчет доли потерянных сессий на передачу трафика реального времени в зависимости от изменения потенциальной загрузки единицы ресурса

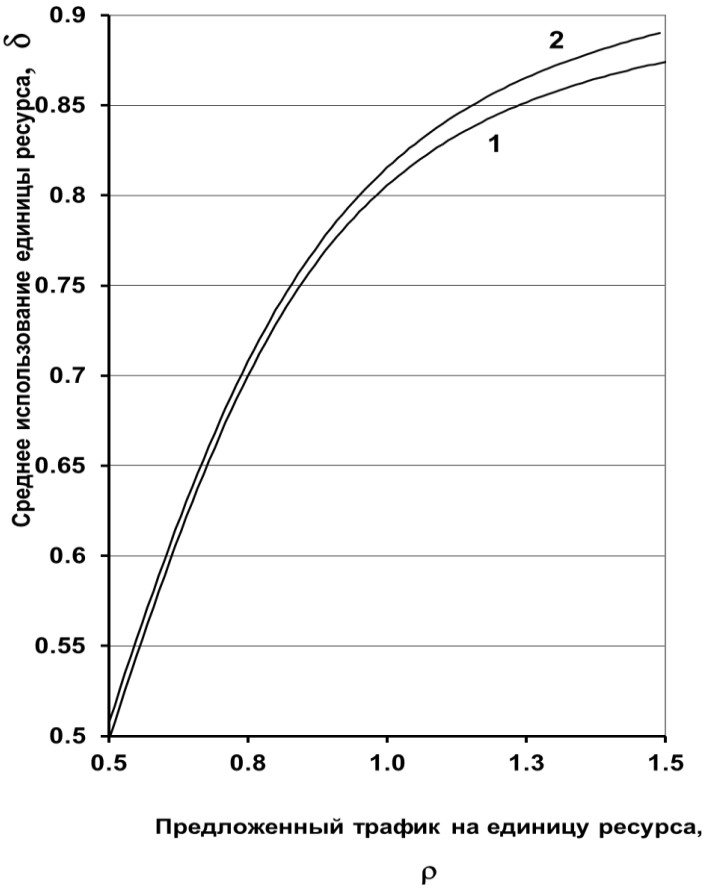


Рисунок 3.3 — Точный и приближенный расчет среднего использования единицы ресурса в зависимости от изменения ρ потенциальной загрузки единицы ресурса

Для дополнительной иллюстрации погрешности вычисления характеристик модели с резервированием сравним численные значения точного и приближенного вычисления характеристик. Соответствующие результаты приведены в таблице 3.1 для тех же значений входных параметров, что были использованы при построении диаграмм, показанных на рисунках 3.2 и 3.3. Полученные результаты подтверждают ранее сделанный вывод о высокой точности рассмотренного приближенного метода. Погрешность лежит в пределах нескольких процентов.

Таблица 3.1 — Точное и приближенное значения π и δ в зависимости от изменения ρ потенциальной загрузки единицы ресурса

ρ	π		δ	
	точное	прибл.	точное	прибл.
0,50	0,004929	0,004844	0,497536	0,497578
0,60	0,018648	0,018176	0,588811	0,589094
0,70	0,047077	0,045559	0,667046	0,668108
0,80	0,089580	0,086343	0,728336	0,730925
0,90	0,140694	0,135476	0,773375	0,778072
1,00	0,194456	0,187434	0,805544	0,812566
1,10	0,246813	0,238380	0,828506	0,837782
1,20	0,295706	0,286277	0,845152	0,856468
1,30	0,340391	0,330314	0,857492	0,870592
1,40	0,380814	0,370358	0,866860	0,881499
1,50	0,417242	0,406606	0,874138	0,890091
1,60	0,450056	0,439388	0,879911	0,896979

3.4. Оценка характеристик обслуживания сессий эластичного трафика

3.4.1. Модель поступления и обслуживания сессий

Предположим, что ресурс, выделенный для организации сессий передачи эластичных данных, обеспечивает пропускную способность C бит в секунду. Выделенный ресурс обслуживает пуассоновский поток заявок на передачу файлов интенсивности λ. Минимальный объем ресурса, который может быть выделен для обслуживания одной сессии, равен r бит в секунду. Пусть C делится на r без остатка. Тогда v число единиц ресурса определяется из равенства $v = \frac{C}{r}$. Величина передаваемого файла распределена экспоненциально и имеет среднее значением F бит. Отсюда получаем, что время пересылки файла единицей ресурса распределено экспоненциально и имеет среднее $1/\mu_d = F/r$ секунд. Данную модель можно использовать для оценки характеристик обслуживания сессий устройств NB-IoT выделенным ресурсом соты LTE, работающем в режиме multi-tone (см. подраздел 2.2.3) с агрегацией передаточных возможностей нескольких поднесущих.

Рассматриваемая математическая модель использования ресурса соты сети LTE, выделенного для обслуживания сессий эластичного трафика, показана на рисунке 3.4 [23].

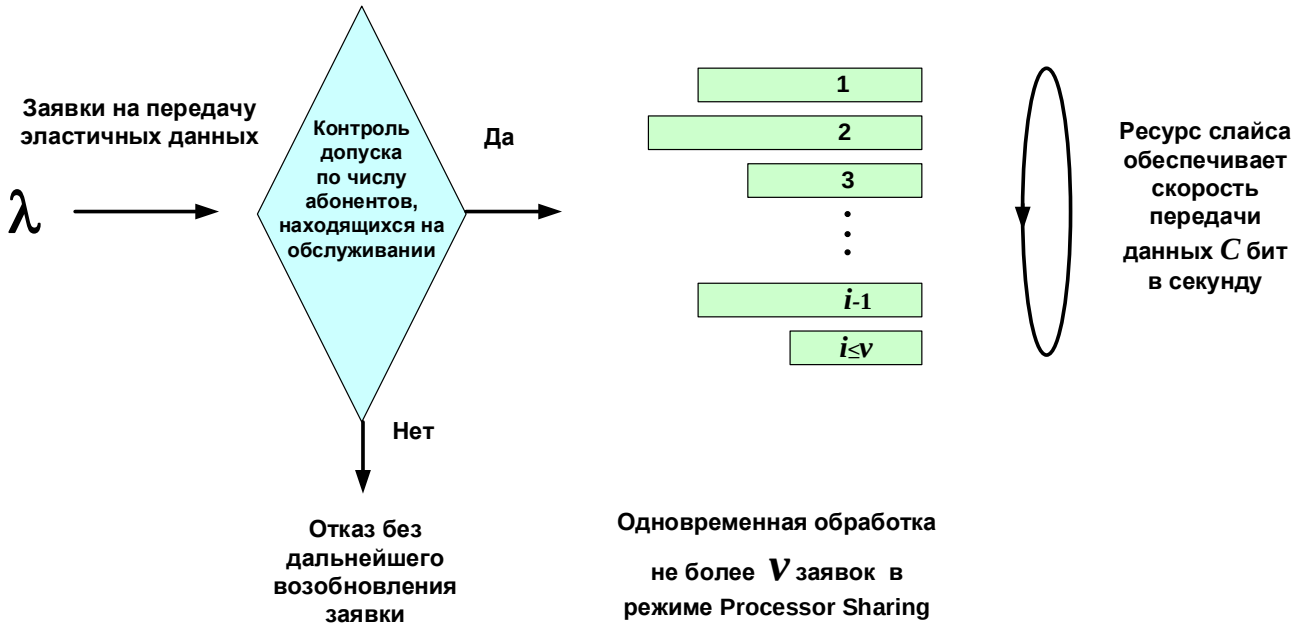


Рисунок 3.4 — Математическая модель использования ресурса соты сети LTE, выделенного для обслуживания эластичных данных

Пусть $i(t)$ — число обслуживаемых сессий в момент t. Изменение обслуживаемых сессий описывается марковским процессом $r(t)=(i(t))$, заданным на пространстве состояний $S=\{(i),\ i=0,1,\ldots,v\}$. Здесь i — число обслуживаемых сессий на пересылку эластичного трафика. Диаграмма переходов $r(t)$ такая же как диаграмма переходов модели $M/M/1/1+w$, при $1+w=v$. Понятно, что модель $M/M/1-PS$ и модель $M/M/1/1+w$, где w — число мест ожидания [25], эквивалентны. Следовательно, у них совпадают вероятности пребывания в состоянии (i). Воспользуемся этим свойством для оценки характеристик исследуемой модели. Диаграмма переходов $r(t)$ показана рисунке 3.5.

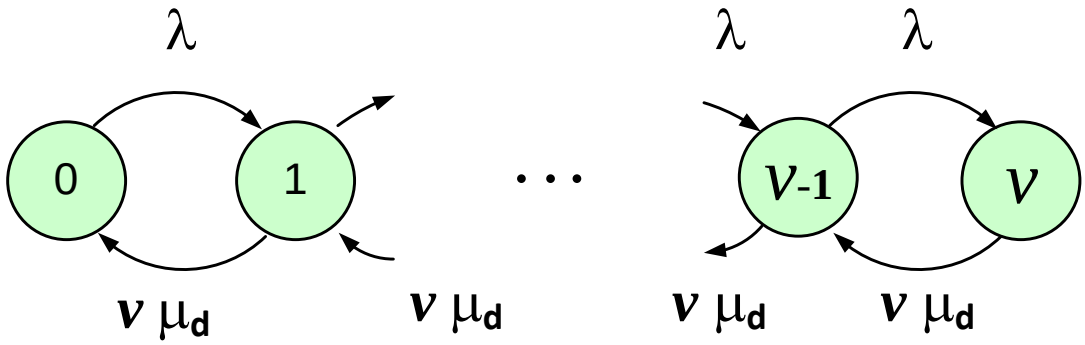


Рисунок 3.5 — Диаграмма переходов марковского процесса, описывающего изменение состояния числа сессий передачи эластичного трафика

3.4.2. Характеристики обслуживания сессий

Пусть интенсивность предложенного информационного потока находится из выражения $\beta = \lambda F$ и выражается в битах в секунду. Пусть $\rho = \frac{\beta}{C} = \frac{\lambda}{\nu\mu_d}$ представляет из себя коэффициент потенциальной загрузки обслуживающего ресурса. Из рисунка 3.5, а также из выражений детального баланса для системы $M/M/1/1+w$, находим рекурсии для стационарных вероятностей

$$p(i)\rho = p(i+1), \quad i = 0,1,\ldots,\nu-1. \tag{3.15}$$

Используя (3.15) и условие нормировки, находим формулы для оценки $p(i),\ i = 0,1,\ldots,\nu$ в стационарном режиме, т.е. для всех $\rho > 0$:

$$p(i) = \frac{(1-\rho)\rho^i}{1-\rho^{\nu+1}},\ \rho \neq 1; \qquad p(i) = \frac{1}{\nu+1},\ \rho = 1.$$

Вероятности стационарных состояний модели и значения характеристик качества обслуживания сессий связи, полученные с их использованием, сохраняют свои значения при изменении функции распределения объема передаваемого файла, если при этом не меняется его средний объем.

Приведем определения и расчетные выражения для характеристик обслуживания сессий передачи данных $(\rho \neq 1)$

$$\pi_d = p(v) = \frac{(1-\rho)\rho^v}{1-\rho^{v+1}}; \tag{3.16}$$

$$y_d = \sum_{i=1}^{v} p(i)i = p(0)\rho\left(1+2\rho+3\rho^2+\ldots+v\rho^{v-1}\right) = \frac{\rho}{1-\rho} - \frac{(v+1)\rho^{v+1}}{1-\rho^{v+1}} = \frac{\rho}{1-\rho}\frac{1-(v+1)\rho^v+v\rho^{v+1}}{1-\rho^{v+1}};$$

$$T_d = \frac{y_d}{\lambda(1-\pi_d)} = \frac{Fy_d}{\rho C(1-\pi_d)} = \frac{\rho}{\lambda(1-\rho)}\frac{1-(v+1)\rho^v+v\rho^{v+1}}{1-\rho^v};$$

$$I_d = \sum_{i=1}^{v} p(i)v\mu_d = v\mu_d\left(1-p(0)\right) = \frac{\lambda\left(1-\rho^v\right)}{1-\rho^{v+1}};$$

$$k_d = \frac{I_d}{y_d\mu_d} = \frac{v\left(1-p(0)\right)}{y_d} = \frac{\lambda\left(1-\rho^v\right)}{y_d\mu_d\left(1-\rho^{v+1}\right)}.$$

При $\rho = 1$ получаем такие формулы

$$\pi_d = \frac{1}{v+1};\ \ y_d = \frac{v}{2};\ \ T_d = \frac{F(v+1)}{2C} = \frac{\rho(v+1)}{2\lambda}; \tag{3.17}$$

$$I_d = \frac{v^2\mu_d}{v+1};\ \ k_d = \frac{2v}{v+1}.$$

Для вычисления характеристик обслуживания сессий передачи эластичных данных в слайсе можно использовать рекурсию (3.15) и определения характеристик либо явные выражения характеристик через значения входных параметров (3.16) и (3.17). Отметим, что в вычислительном плане первый из отмеченных подходов приводит к более эффективным расчетным процедурам.

3.4.3. Анализ эффективности дисциплины *PS* при обслуживании эластичного трафика

Проведем численное исследование эффективности применения процедуры *PS* при обслуживании сессий передачи эластичного трафика. Будем предполагать, что базовая станция имеет лицензию на 20 МГц. Из стандарта LTE следует, что базовая станция может обслужить одновременно максимум 80 абонента, предоставив каждому пользователю ее услуг скорость не ниже 1 Мбит/с. Из перечисленных данных следуют величины параметров модели передачи

эластичного трафика, используемые при проведении вычислений. Предположим, что: $v = 80$, минимальная скорость одного виртуального канала передачи информации $r = 1$ Мбит/с, средний объем передаваемого файла F примем равным 10 Мбайт или 80 Мбит. Величина μ_d определяется из соотношения $\frac{1}{\mu_d} = \frac{80}{1} = 80$ сек. Это максимальное среднее время передачи файла с использованием передаточных возможностей одного канала.

Исследуем зависимость веденных характеристик модели (3.16) и (3.17) от изменения $\rho = \beta / C = \lambda / v\mu_d$ потенциальной нагрузки на один канал и сравним найденные значения характеристик с их аналогами, рассчитанными при условии передачи файлов по правилам обслуживания трафика реального времени, т.е. при использовании одного канала для передачи одного файла. В этой ситуации значения характеристик обслуживания сессий рассчитываются с использованием мультисервисной модели Эрланга [25]. Эта модель является частным случаем модели обслуживания характеристик обслуживания сессий трафика реального времени, рассмотренной в разделе 3.3. Для того, чтобы ее получить нужно выбрать величины функции блокировки исходя из следующих равенств $\varphi_k(i) = 0,\ \ k = 1,\ldots,n,\ \ i = 0,1,\ldots,v - b_k$ и $\varphi_k(i) = 1,\ \ k = 1,\ldots,n,\ \ \ i = v - b_k + 1,$ $v - b_k + 2,\ldots,v.$ При данном выборе значений функции блокировки использование рекурсии (3.13) позволяет найти точные значения характеристик обслуживания сессий трафика реального времени без использования механизма резервирования. Для этого используются выражения (3.14). На рисунках 3.6 и рисунках 3.7 показана зависимость потерь сессий для двух диапазонов изменения ρ: $\rho \le 2$ (рисунок 3.6) и $\rho \le 1.1$ (рисунок 3.7).

Из приведенных данных можно сделать вывод о существенном снижении потерь сессий, которое достигается в результате использования дисциплины *PS*. Причем больше всего величины потерь уменьшаются для значений потерь, не превышающих 0,1. Понятно, что это происходит из-за большого объема свободного ресурса передачи информации, который в ситуации применения *PS* используется для ускорения обслуживания сессий передачи эластичных данных. В ситуации перегрузки, когда $\rho > 1$ характеристики потерь сессий выравниваются, поскольку при использовании обеих дисциплин *RT* и *PS* для передачи файла используется чуть больше (*PS*) или ровно (*RT*) один канал.

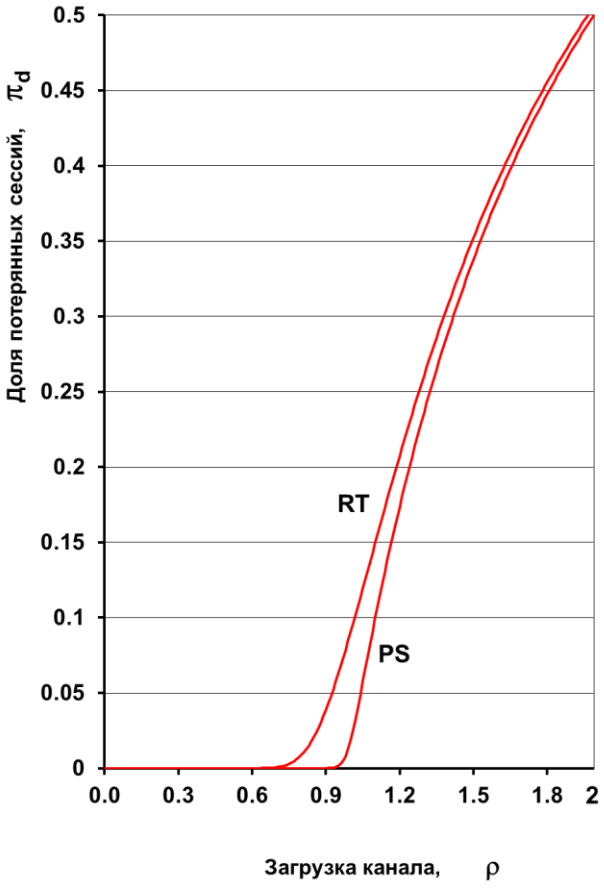


Рисунок 3.6 — Зависимость потерь сессий от $\rho \le 2$

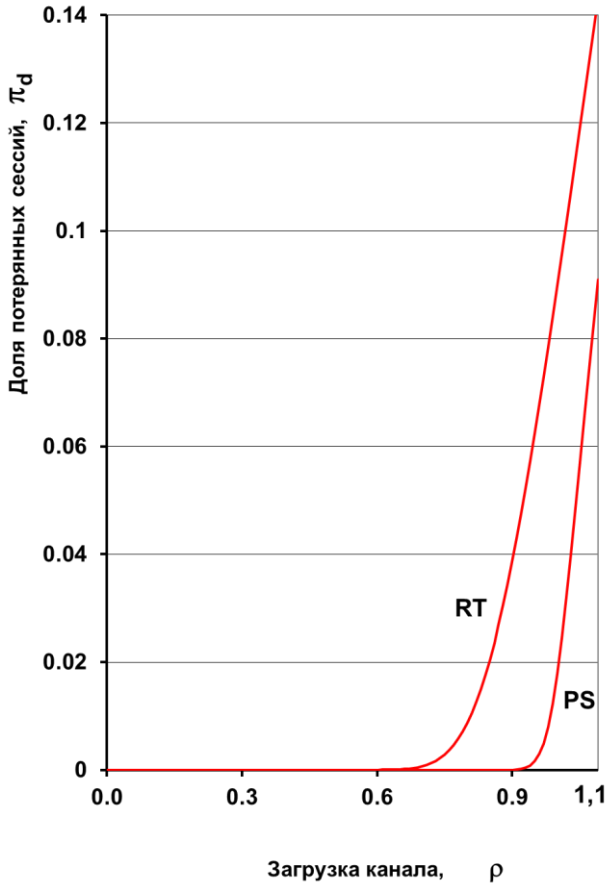


Рисунок 3.7 — Зависимость потерь сессий от $\rho \le 1.1$

На рисунках 3.8 и 3.9 приведены соответственно значения T_d среднего времени обслуживания одной сессии и значения δ среднего использования одного канала. Для дисциплины PS величина T_d рассчитывается из выражений (3.16) и (3.17), а для дисциплины RT величина $T_d = \frac{1}{\mu_d}$, т.е. средний интервал времени пересылки файла постоянен и в рассматриваемом случае просто равен 80 сек., поскольку передача файла идет на одном канале. Для дисциплины PS величина δ рассчитывается из выражения

$$\delta = \sum_{i=1}^{v} p(i) = 1 - p(0) = \frac{I_d}{v\mu_d}, \qquad (3.18)$$

а для дисциплины RT используется формула $\delta = \frac{m}{v}$, где m — среднее число занятых каналов. Использование канала определяется работой, которую надо совершить по передаче файла. В ситуации малых потерь эта работа, отнесенная на один канал, одинакова как для дисциплины PS ,

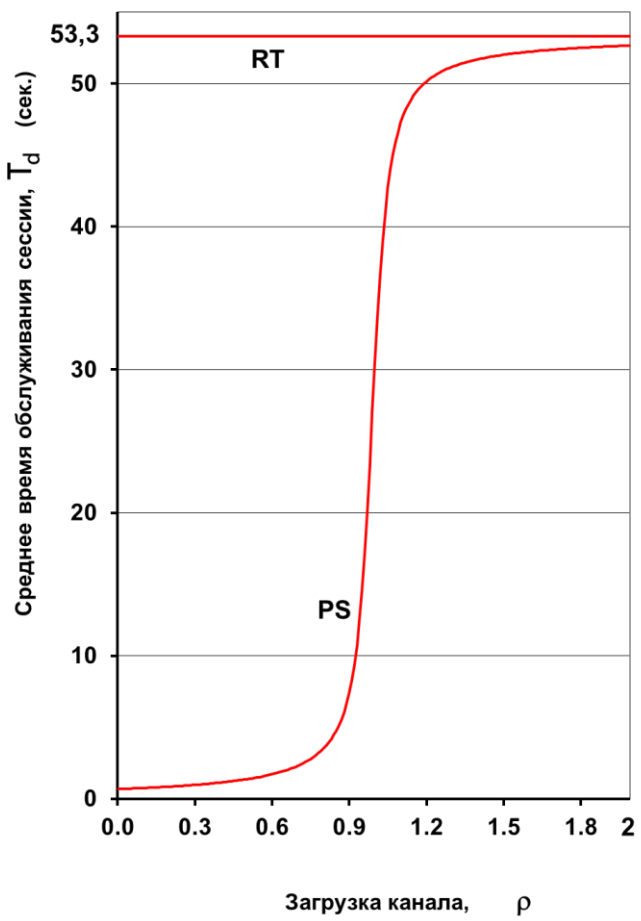


Рисунок 3.8 — Среднее время обслуживания сессии

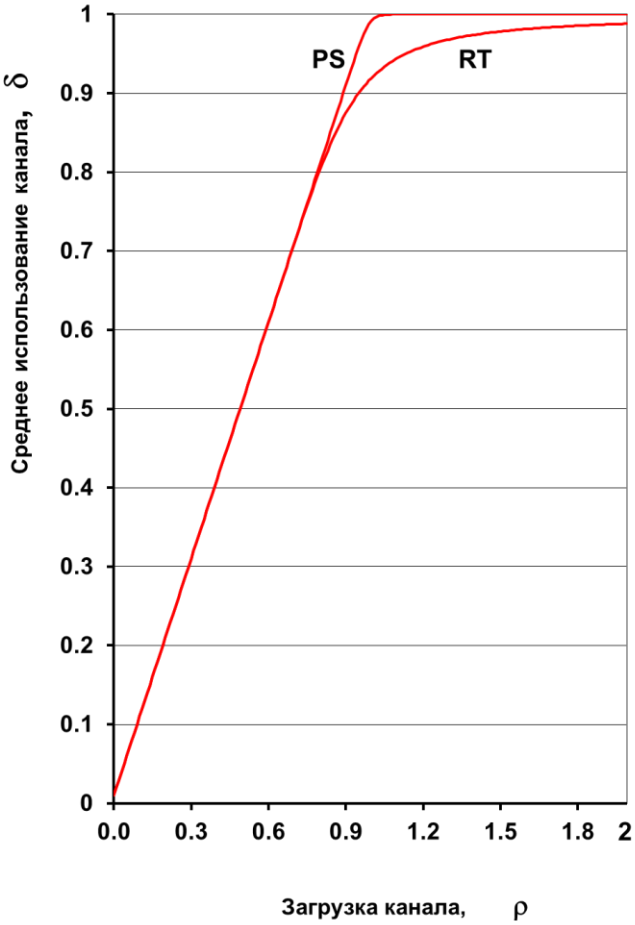


Рисунок 3.9 — Среднее использование канала

так и для дисциплины RT , поэтому соответствующие кривые близки друг к другу. С ростом ρ при использовании дисциплины PS уменьшаются потери, поэтому возрастает использование канала по сравнению с дисциплиной RT .

На рисунках 3.10 и рисунках 3.11 приведены соответственно значения k_d среднего числа каналов, используемых для обслуживания одной сессии и значения m_d среднего числа обслуживаемых сессий связи. Для дисциплины PS величина k_d и m_d рассчитывается из выражений (3.16) и (3.17), а для дисциплины RT значение $k_d = 1$, а m_d рассчитывается из выражения (3.14). Приведенные данные показывают, что эффективность дисциплины PS связана

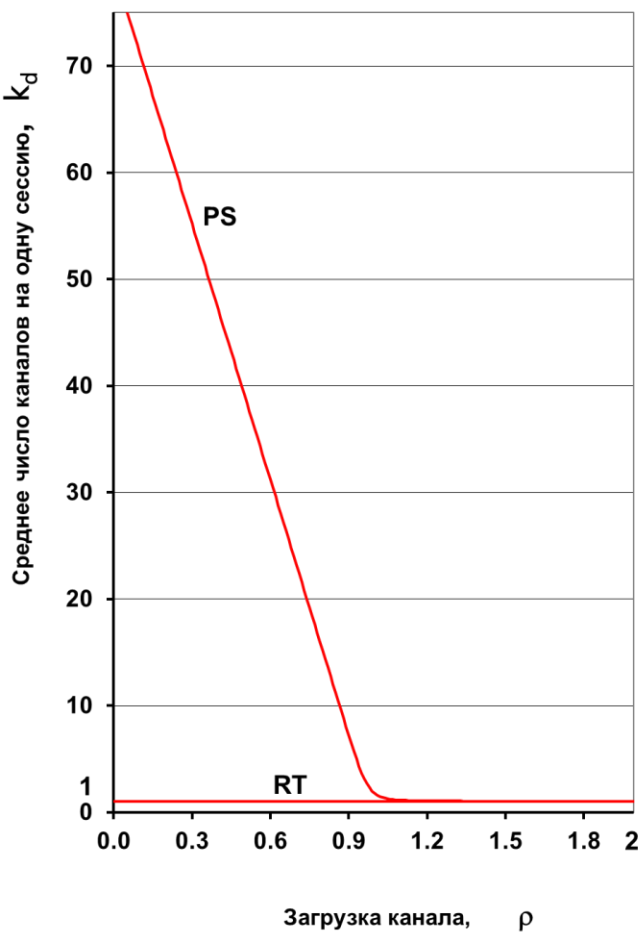


Рисунок 3.10 — Среднее число каналов на одну сессию

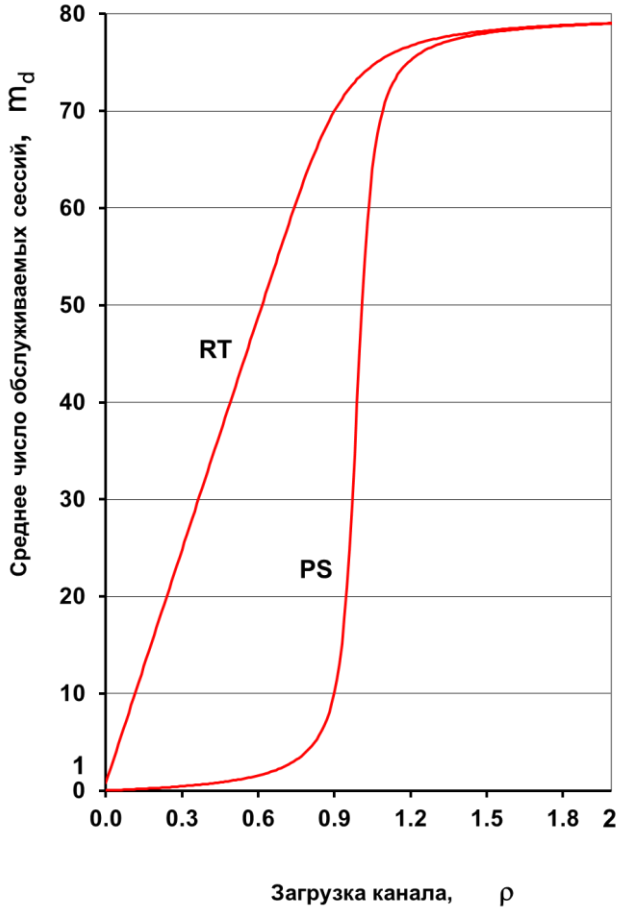


Рисунок 3.11 — Среднее число обслуживаемых сессий

с возможность использования ресурса свободных каналов для передачи эластичных данных. В пределе для одного пользователя может быть выделен весь имеющийся канальный ресурс (рисунок 3.10). Ускоренное обслуживание эластичных данных способствует уменьшению потерь трафика реального времени при их совместном обслуживании на выделенном ресурсе (см. раздел 4).

3.5. Анализ трехпотоковой модели

3.5.1. Описание модели

Оценка характеристик совместного обслуживания запросов на информационное обслуживание ресурсом соты, выделенным оператору систем наблюдения, проводится с помощью

решения СУР. Если ставить задачу оценки характеристик для реальных величин входных параметров, а это значит для значений v примерно равных 100, то число входных потоков следует взять равным три. Дальнейшее увеличение числа потоков требует либо уменьшения числа каналов v, либо применения более мощных вычислительных средств нежели персональные компьютеры. Имея ввиду важность трехпотоковой модели для проведения численных расчетов и практических приложений, рассмотрим ее исследование более подробно.

Будем предполагать, что в модели рассматривается процесс совместной организации сессий связи для двух потоков сервисов реального времени и одного потока эластичного трафика данных. Пусть v — общее число выделенных единиц ресурса соты; r — скорость пересылки информации, задаваемая одной единицей ресурса; C — скорость пересылки информации, предоставляемая всем ресурсом в битах в секунду $(C = vr)$. Поступление заявок на организацию сессий для первого и второго потока трафика реального времени происходит через случайное время имеющее экспоненциальное распределение с параметрами λ_1 и λ_2 соответственно. Длительность интервала времени обслуживания экспоненциально распределено с параметрами μ_1 и μ_2 соответственно. Обозначим через b_1 и b_2 – число канальных единиц, которые используются для обслуживания 1-го и 2-го потока сессий трафика реального времени соответственно.

Если в момент поступления сессии нет свободных единиц ресурса, тогда делается попытка уменьшить скорость передачи данных до минимального значения, которое принято за единицу ресурса и, если это удается, то рассматриваемая сессия принимается на обслуживание. Поступление сессий на передачу эластичных данных подчиняется пуассоновскому закону с интенсивностью λ_d . Величина передаваемого файла распределена экспоненциально и имеет среднее значением F бит. Отсюда получаем, что время пересылки файла единицей ресурса распределено экспоненциально и имеет среднее $1/\mu_d = F/r$ секунд. Для ограничения доступа поступающих сессий используются функции $\varphi_k(i)$, $k = 1, 2$, $i = 0, 1, \ldots, v$ для сессий трафика реального времени и $\varphi_d(i)$, $i = 0, 1, \ldots, v$ для сессий передачи эластичных данных. Здесь i — суммарное число каналов занятых на обслуживание сессий трафика реального времени и количество сессий передачи данных, находящихся на обслуживании. Более подробно описание рассматриваемой модели приведено в подразделе 2.4. Рассматриваемая математическая модель использования ресурса соты сети LTE, выделенного для обслуживания сессий связи оператора систем наблюдения, показана на рисунке 3.12.

Рисунок 3.12 — Математическая модель использования ресурса соты сети LTE, выделенного для обслуживания сессий связи оператора систем наблюдения

3.5.2 Марковский процесс и характеристики модели

Пусть $i_1(t)$ и $i_2(t)$ число обслуживаемых в момент t сессий 1-го и 2-го потоков на передачу трафика реального времени, $k=1,2,\ldots,n,$ а $d(t)-$ число обслуживаемых в момент t сессий пересылки эластичных данных. Изменение числа организованных сессий связи в зависимости от времени представлено двумерным случайным процессом $r(t)=(i_1(t),i_2(t),d(t))$, заданным на пространстве S. В него входят вектора (i_1,i_2,d) с компонентами, принимающими значения:

$$i_1=0,1,\ldots,\left\lfloor\frac{v}{b_1}\right\rfloor,\quad i_2=0,1,\ldots,\left\lfloor\frac{v-i_1b_1}{b_2}\right\rfloor; \tag{3.19}$$

$$d=0,1,\ldots,v-i_1b_1-i_2b_2.$$

Пространство реальных состояний определяется выбором $\varphi_k(i),\ k=1,2$.

По условиям построения модели все случайные величины имеют имеют экспоненциальное распределение и независимы. Следовательно, процесс $r(t)$ обладает марковскими свойствами. Пусть $p(i_1,i_2,d)$ стационарная вероятность состояния $(i_1,i_2,d)\in S$. Стандартным образом (см. подраздел 2.5) определяются характеристики совместного обслуживания сессий связи. Напомним, что через i_r обозначено число единиц ресурса соты, выделенных для обслуживания сессий связи оператора систем наблюдения и используемых в состоянии (i_1,i_2,d) на обслуживание сессий трафика сервисов реального времени. Обозначим суммирование по используемому пространству состояний в следующем упрощенном виде

$$\sum_{i_1=0}^{\left\lfloor \frac{v}{b_1} \right\rfloor} \sum_{i_2=0}^{\left\lfloor \frac{v-i_1b_1}{b_2} \right\rfloor} \sum_{d=0}^{v-i_1b_1-\ldots-i_2b_2} = \sum_{(i_1,i_2,d)\in S} . \qquad (3.20)$$

Используя (3.20), и определения характеристик из подраздела 2.5, получаем следующие расчетные формулы для вычисления их значений:

$$\pi_1 = \sum_{(i_1,i_2,d)\in S} p(i_1,i_2,d)\varphi_1(i_r+d); \quad \pi_2 = \sum_{(i_1,i_2,d)\in S} p(i_1,i_2,d)\varphi_2(i_r+d);$$

$$m_1 = \sum_{(i_1,i_2,d)\in S} p(i_1,i_2,d)i_1b_1; \quad m_2 = \sum_{(i_1,i_2,d)\in S} p(i_1,i_2,d)i_2b_2;$$

$$y_1 = \sum_{(i_1,i_2,d)\in S} p(i_1,i_2,d)i_1; \quad y_2 = \sum_{(i_1,i_2,d)\in S} p(i_1,i_2,d)i_2;$$

$$\pi_d = \sum_{(i_1,i_2,d)\in S} p(i_1,i_2,d)\varphi_d(i_r+d); \quad y_d = \sum_{(i_1,i_2,d)\in S} p(i_1,i_2,d)d;$$

$$I_d = \sum_{(i_1,i_2,d)\in S|d>0} p(i_1,i_2,d)(v-i_r)\mu_d; \quad k_d = \frac{I_d}{y_d\mu_d};$$

$$T_d = \frac{y_d}{\lambda_d(1-\pi_d)}.$$

Для оценки характеристик требуется составить и решить СУР. В общем виде эта система представлена в подразделе 2.4.3, см. соотношение (2.2). Приведем вид системы уравнений статистического равновесия в рассматриваемом частном случае.

$$P(i_1,i_2,d)\Big\{\ \lambda_1\left(1-\varphi_1\left(i_r+d\right)\right)+\lambda_2\left(1-\varphi_2\left(i_r+d\right)\right)+ \tag{3.21}$$

$$+i_1\mu_1+i_2\mu_2+\lambda_d\left(1-\varphi_d\left(i_r+d\right)\right)+\left(v-i_r\right)\mu_d I\left(d>0\right)\ \Big\}=$$

$$=P\left(i_1-1,i_2,d\right)\lambda_1\left(1-\varphi_1\left(i_r+d-b_1\right)\right)I\left(i_1>0\right)+$$

$$+P\left(i_1,i_2-1,d\right)\lambda_2\left(1-\varphi_2\left(i_r+d-b_2\right)\right)I\left(i_2>0\right)+$$

$$+P\left(i_1,i_2,d-1\right)\lambda_d\left(1-\varphi_d\left(i_r+d-1\right)\right)I\left(d>0\right)+$$

$$+P\left(i_1+1,i_2,d\right)\left(i_1+1\right)\mu_1 I\left(i+d+b_1\le v\right)+$$

$$+P\left(i_1,i_2+1,d\right)\left(i_2+1\right)\mu_2 I\left(i+d+b_2\le v\right)+$$

$$+P\left(i_1,i_2,d+1\right)\left(v-i_r\right)\mu_d I\left(i+d+1\le v\right).$$

В (2.2) $I\left(\cdot\right)$ — индикаторная функция, обозначающая результат осуществления события. Значение $I\left(\cdot\right)$ равно единице, если неравенство, указанное в скобках выполняется, и значение $I\left(\cdot\right)$ равно нулю, если это неравенство не выполняется. Для значений $P\left(i_1,i_2,d\right)$ выполнено условие нормировки

$$\sum_{(i_1,i_2,d)\in S}P\left(i_1,i_2,d\right)=1.$$

Система уравнений равновесия решается с использованием алгоритма Гаусса-Зейделя. Соответствующая рекурсивная последовательность является частным случаем $\left(3.8\right)$ и $\left(3.9\right)$. Зависимость между $\left(s+1\right)$-ым и s-ым приближениями определяется следующим выражением:

$$P^{(s+1)}\left(i_1,i_2,d\right)=\frac{1}{L\left(i_1,i_2,d\right)}\times \tag{3.22}$$

$$\times\Big\{\ P^{(s,s+1)}\left(i_1-1,i_2,d\right)\lambda_1\left(1-\varphi_1\left(i_r+d-b_1\right)\right)I\left(i_1>0\right)+$$

$$+P^{(s,s+1)}\left(i_1,i_2-1,d\right)\lambda_2\left(1-\varphi_2\left(i_r+d-b_2\right)\right)I\left(i_2>0\right)+$$

$$+P^{(s,s+1)}\left(i_1,i_2,d-1\right)\lambda_d\left(1-\varphi_d\left(i_r+d-1\right)\right)I\left(d>0\right)+$$

$$+P^{(s,s+1)}\left(i_1+1,i_2,d\right)\left(i_1+1\right)\mu_1 I\left(i+d+b_1\le v\right)+$$

$$+P^{(s,s+1)}\left(i_1,i_2+1,d\right)\left(i_2+1\right)\mu_2 I\left(i+d+b_2\le v\right)+$$

$$+P^{(s,s+1)}\left(i_1,i_2,d+1\right)\left(v-i_r\right)\mu_d I\left(i+d+1\le v\right).$$

где

$$L\left(i_1,i_2,d\right)=\Big\{\ \lambda_1\left(1-\varphi_1\left(i_r+d\right)\right)+\lambda_2\left(1-\varphi_2\left(i_r+d\right)\right)+$$

$$+i_1\mu_1+i_2\mu_2+\lambda_d\left(1-\varphi_d\left(i_r+d\right)\right)+\left(v-i_r\right)\mu_d I\left(d>0\right)\ \Big\}.$$

После того как сходимость итерационной процедуры (3.22) достигнута значения $P^{(s+1)}\left(i_1,i_2,d\right)$ нормируются. Для этого используется соотношение:

$$P\left(i_1,i_2,d\right)=\frac{P^{(s+1)}\left(i_1,i_2,d\right)}{\sum\limits_{(i_1,i_2,d)\in S}P^{(s+1)}\left(i_1,i_2,d\right)}.$$

Результаты вычисления характеристик в соответствии с разработанным алгоритмом рассмотрены в заключительном разделе работы.

3.5.3. Численный анализ сходимости итерационной процедуры

Приведем численный пример, иллюстрирующий скорость сходимости в зависимости от выбора ε в (3.5). Рассмотрим модель слайса с параметрами: $v=200$ к.е.; $n=3$; $b_1=10$ к.е.; $b_2=20$ к.е.; $b_d=1$ к.е.; $\lambda_k=v\rho/3b_k$; $\mu_k=1$; $k=1,2$; $\lambda_d=v\rho/3$; $\mu_d=1$. Величина ε принимает значения: 10^{-7}; 10^{-8}; 10^{-9}; 10^{-10}; 10^{-15}. Величина ρ потенциальной загрузки единицы ресурса принимает значения: $0,5$; $0,6$; $0,7$; $0,8$; $0,9$; $1,0$; $1,1$; $1,2$; $1,3$; $1,4$; $1,5$. В рассматриваемой модели слайса реализована процедура выравнивания потерь. Величины функции блокировки выбраны так, чтобы доли потерянных сессий связи у всех потоков были одинаковы. Этот результат достигается следующим выбором функций блокировки:

$$\varphi_1\left(i\right)=0,\ i=0,1,\ldots,v-b_2;\quad \varphi_1\left(i\right)=1,\ i=v-b_2+1,\ v-b_2+2,\ldots,v;$$
$$\varphi_2\left(i\right)=0,\ i=0,1,\ldots,v-b_2;\quad \varphi_2\left(i\right)=1,\ i=v-b_2+1,\ v-b_2+2,\ldots,v;$$

$$\varphi_d(i)=0,\ i=0,1,\ldots,v-b_2;\quad \varphi_d(i)=1,\ i=v-b_2+1,\ v-b_2+2,\ldots,v.$$

В таблице 3.2 приведены величины $\pi=\pi_1=\pi_2=\pi_d$ в зависимости от ε и ρ. В таблице 3.3 приведены величины N_i, где N_i — число шагов итерационного цикла в (3.9) до достижения относительной ошибки ε при вычислении π зависимости от изменения тех же параметров. В последнем столбце, рассчитанным при $\varepsilon=10^{-15}$, приведены десять точных значений значащих цифр величины потерь поступающих сессий связи.

Из приведенных данных можно сделать следующие выводы:

- Число итераций исчисляется несколькими сотнями для достижения приемлемой для практических целей точности в оценке характеристик. При увеличении относительной погрешности вычисления нормировочной константы на порядок число итераций возрастает примерно на 100.
- Требуемое число итераций для оценки характеристик с заданной погрешностью обычно растет с ростом нагрузки на канал, но не всегда этот рост носит монотонный характер.
- Для использования результатов вычислений для графического отображения зависимости характеристик от изменения входных параметров достаточно задать относительную погрешность вычисления характеристик на уровне $\varepsilon=10^{-6}-10^{-7}$.

Таблица 3.2 — Величины $\pi=\pi_1=\pi_2=\pi_d$ в зависимости от ε и ρ

ρ \ ε	π				
	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-15}
0,50	0,0012847040	0,0012848996	0,0012849187	0,0012849206	0,0012849208
0,60	0,0048799905	0,0048806876	0,0048807560	0,0048807629	0,0048807637
0,70	0,0135190693	0,0135207180	0,0135208797	0,0135208956	0,0135208973
0,80	0,0296567849	0,0296567849	0,0298783794	0,0298784013	0,0298784036
0,90	0,0556881104	0,0558024624	0,0558026187	0,0558026316	0,0558026328
1,00	0,0916132058	0,0916175792	0,0916424003	0,0916424003	0,0916423998
1,10	0,1360557283	0,1360584816	0,1360589023	0,1360589559	0,1360598165
1,20	0,1861551155	0,1861558592	0,1861559458	0,1861559562	0,1861559577
1,30	0,2380400218	0,2380400350	0,2380400376	0,2380400380	0,2380400380
1,40	0,2880684446	0,2880680928	0,2880680581	0,2880680547	0,2880680543
1,50	0,3340091258	0,3340086439	0,3340085961	0,3340085913	0,3340085908

Таблица 3.3 — Величины N_i, где N_i — число шагов итерационного цикла в (3.9) до достижения относительной ошибки ε при вычислении π зависимости от изменения ε и ρ

ε / ρ	π				
	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-15}
0,50	452	540	627	714	1151
0,60	507	608	708	809	1314
0,70	554	664	774	885	1445
0,80	384	384	828	944	1560
0,90	463	770	885	1001	1675
1,00	572	580	962	1076	1786
1,10	717	775	794	797	1891
1,20	844	981	1113	1239	1535
1,30	920	1090	1260	1430	2273
1,40	950	1137	1324	1510	2443
1,50	951	1144	1337	1530	2493

3.6. Выводы по результатам третьего раздела

1. Анализ численных методов расчета характеристик совместного обслуживания сессий связи в мультисервисных узлах доступа показал, что наиболее эффективным способом их вычисления являются процедуры, основанные на формировании и последующем решении СУР с помощью алгоритма Гаусса-Зейделя. Выбор алгоритма основан на использовании свойств матрицы СУР, среди которых важнейшими являются большое число неизвестных и большое количество нулевых элементов, а также наличие рекуррентных зависимостей для оценки коэффициентов. Приведена формулировка основных шагов алгоритма для мультисервисного узла доступа. Среди них: выбор начальных значений итерационного цикла; рекурсивная формула, связывающая величины последовательных приближений; анализ условий сходимости и критерий останова итерационного процесса.
2. В целях повышения эффективности реализации итерационного алгоритма Гаусса-Зейделя получено выражение для формирования системы уравнений статистического равновесия в виде цикла по целочисленным компонентам состояния. Полученное соотношение дает возможность определять коэффициента матрицы системы уравнений непосредственно в

процессе применения итерационного алгоритма, а не хранить их в памяти компьютера. Этот результат существенно упрощает реализацию итерационного метода и позволяет увеличить число состояний в исследуемой модели мультисервисного узла доступа до нескольких миллионов.

3. Проведено численное исследование особенностей реализации итерационного метода. Показано, что в практически интересных случаях число итераций исчисляется несколькими сотнями для достижения приемлемой точности в оценке характеристик. При увеличении относительной погрешности вычисления нормировочной константы на порядок число итераций возрастает примерно на 100. Требуемое число итераций для оценки характеристик с заданной погрешностью обычно растет с ростом нагрузки на канал, но не всегда этот рост носит монотонный характер. Для использования результатов вычислений для графического отображения зависимости характеристик от изменения входных параметров достаточно задать относительную погрешность вычисления характеристик на $\varepsilon = 10^{-6} - 10^{-7}$.
4. Построена вычислительная процедура оценки характеристик обслуживания трафика сервисов реального времени на отдельном слайсе с использованием возможностей ограниченного доступа. Сформулированы определения характеристик обслуживания заявок и рассмотрены алгоритмы их точной и приближенной оценки. Численно исследована точность приближенного алгоритма. Показано, что относительная погрешность оценки характеристик лежит в пределах нескольких процентов для большинства практически интересных случаев.
5. Построена вычислительная процедура оценки характеристик обслуживания эластичного трафика данных на отдельном слайсе. Построенную модель можно применять для оценки характеристик обслуживания сессий устройств NB IoT выделенным ресурсом соты LTE, работающем в режиме multi-tone с агрегацией передаточных возможностей нескольких поднесущих. Сформулированы определения характеристик обслуживания сессий передачи эластичного трафика. В рассматриваемом случае для оценки показателей обслуживания сессий можно использовать явные формулы и рекурсивные алгоритмы. Приведены соответствующие расчетные выражения через значения входных параметров модели. Численно исследована эффективность агрегации свободных каналов при обслуживании эластичного трафика. В области малых потерь (до 10 %) доля потерь сессий и среднее время передачи файла могут уменьшиться более чем в десять раз.

Раздел 4

Использование разработанной модели для решения задач эффективного распределения при совместном обслуживании трафика реального времени и эластичного трафика данных

4.1. Введение к разделу 4

Обслуживание неоднородного трафика общим ресурсом приводит к его неконтролируемому перераспределению в пользу «легкого» трафика. Эта проблема обсуждается в подразделе 4.2 на конкретных численных примерах. Отметим, что отрицательные последствия этого явления особенно ярко проявляют себя при обслуживании «легкого» трафика по правилам передачи трафика реального времени, т.е. при использовании передаточных возможностей одного канала. Если трафик данных обладает свойством эластичности, то острота проблем снижается, однако они также требуют своего решения. Сценарии эффективного обслуживания гетерогенного трафика рассматриваются в подразделе 4.3. В их числе статичный слайсинг и динамичный слайсинг. Первый отличается простотой реализации и имеет достаточно простые алгоритмы оценки показателей качества обслуживания [19, 20]. Однако этот сценарий требует большего объема ресурса для своей реализации.

Численный анализ использования каждого из отмеченных сценариев при создании условий по дифференцированному обслуживанию гетерогенного трафика рассмотрен в подразделах 4.4 и 4.5 для двух моделей формирования и обслуживания трафика данных. В первой модели трафик данных представляет из себя сессии передачи видеоконтента с низким качеством, требующим относительно невысокую скорость передачи. Он обслуживается по правилам трафика реального времени. Каждый файл передается с использованием возможностей одной канальной единицы. Численный анализ модели выполнен в подразделе 4.4. Для оценки характеристик совместного обслуживания сессий используются модели и алгоритмы, рассмотренные в подразделе 3.3. Во второй модели трафик данных обладает эластичными свойствами, например, представляя из себя файлы, получающиеся после записи видеоконтента в буфер. Он обслуживается по правилам эластичного трафика. Минимальный объем используемого ресурса составляет одну канальную единицу. Для оценки характеристик совместного обслуживания сессий используются модели и алгоритмы, рассмотренные в подразделах 2.4, 2.5, 3.3 — 3.5. Численный анализ второй модели

формирования трафика выполнен в подразделе 4.5. В заключительном разделе сформулированы выводы по 4-му разделу.

4.2. Численный анализ совместного обслуживания трафика реального времени и эластичных данных

4.2.1. Проблемы совместного обслуживания гетерогенного трафика

Использование ограниченного ресурса мультисервисного узла доступа при совместном обслуживании информационных потоков с ярко выраженной неоднородностью требований к ресурсу приводит к перераспределению ресурса в пользу так называемого «легкого» трафика. Проведем численное исследование этого явления. Предполагается, что базовая станция имеет лицензию на 20 МГц. Из стандарта LTE следует, что базовая станция может обслужить одновременно максимум 80 абонента, предоставив каждому пользователю ее услуг скорость не ниже 1 Мбит/с. Определим величины параметров модели передачи эластичного трафика, используемые при проведении вычислений. Предположим, что: $v = 80$ к.е., скорость одной к.е. определяется из соотношения $r = 1$ Мбит/с, средний объем передаваемого файла F примем равным 10 Мбайт или 80 Мбит. Величина μ_d — параметр экспоненциального распределения времени передачи файла с использованием одного канала определяется из соотношения $\frac{1}{\mu_d} = \frac{80}{1} = 80$ сек. Это максимальное среднее время передачи файла. Совместно с эластичным трафиком передается трафик реального времени. Предположим, что для обслуживания одной сессии требуется $b_1 = 20$ к.е. Выбор входных параметров указывает, на совместное обслуживание информационных потоков с ярко выраженной неоднородностью требований к ресурсу.

Исследуем зависимость веденных характеристик модели (см. раздел 2.5) от изменения ρ потенциальной нагрузки на один канал. Величина ρ рассчитывается из выражения

$$\rho = \frac{\lambda_1 b_1}{\mu_1} \cdot \frac{1}{v} + \frac{\lambda_d F}{C} = \frac{\lambda_1 b_1}{\mu_1} \cdot \frac{1}{v} + \frac{\lambda_d F}{rv} = \left(\frac{\lambda_1}{\mu_1} b_1 + \frac{\lambda_d}{\mu_d} \right) \frac{1}{v}. \quad (4.1)$$

Будем предполагать, что среднее время передачи сессии трафика реального времени и сессии передачи эластичных данных приняты за единицу $\mu_1 = \mu_d = 1$. Таким образом, все временные

характеристики обслуживания этих сессий (время доставки, частота поступления сессий) выражены в указанных единицах. Также предположим, что потенциальная нагрузка, создаваемая каждым видом трафика, одинакова. Это предположение позволяет получить выражения для интенсивностей поступления сессий в эрлангах через значение ρ

$$\lambda_1 = \frac{\nu\rho}{2b_1}; \quad \lambda_d = \frac{\nu\rho}{2}. \tag{4.2}$$

Выполним анализ совместного обслуживания гетерогенного трафика для двух сценариев обслуживания эластичного данных. В первом, — файл передается по правилам реального времени с использованием одного канала, во втором, — по правилам дисциплины *PS*. В последнем случае все заявки на передачу файлов получают одинаковую порцию ресурса, получающуюся после деления числа всех свободных каналов на общее число передаваемых файлов. На рисунке 4.1 рассмотрено поведение доли потерянных сессий π_1 и π_d с ростом ρ для 1-го сценария распределения ресурса, а на рисунке 4.2 — для 2-го. Цифры у кривой указывают на номер потока

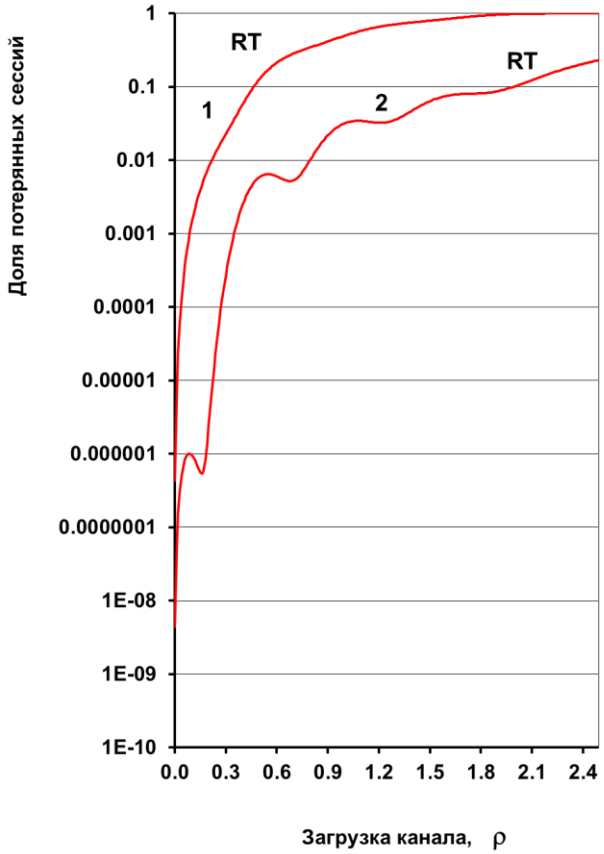


Рисунок 4.1— Зависимость потерь сессий от ρ для 1-го сценария распределения ресурса

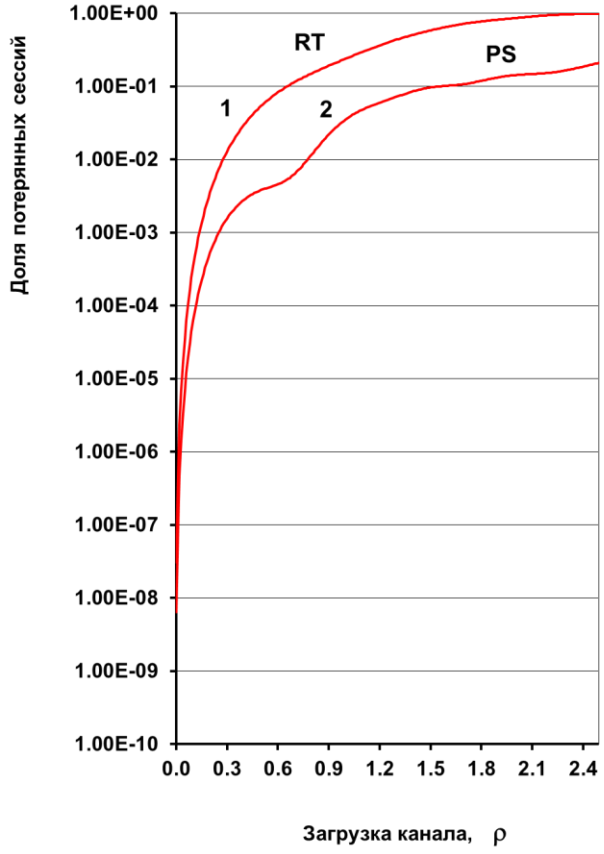


Рисунок 4.2 — Зависимость потерь сессий от ρ для 2-го сценария распределения ресурса

(1-ый поток — это сессии реального времени, 2-ой поток — сессии передачи эластичного трафика), а обозначения *RT* или *PS* показывают используемую дисциплину разделения ресурса. Для проведения вычислений использовались результаты подразделов 3.3 — 3.5.

На рисунке 4.3 анализируется среднее использование канала на пересылку трафика реального времени δ_r и эластичных данных δ_d с ростом ρ для 1-го сценария распределения ресурса, а на рисунке 4.4 — для 2-го. Характеристики δ_r и δ_d определялись из соотношений

$$\delta_r = \frac{m_1}{v}; \quad \delta_d = \frac{y_d k_d}{v}. \tag{4.3}$$

Из полученных данных видно, что с ростом нагрузки на канал «легкий» трафик получает преимущество в занятии ресурса передачи перед «тяжелым» трафиком, вытесняя последний из процесса обслуживания. Количество локальных минимумов у величины потерь «легкого» трафика равно $\left\lfloor \frac{v}{b_1} \right\rfloor = 4$. Первый минимум у π_1 следует из того, что с ростом нагрузки стало практически невозможным одновременное обслуживание четырех ресурсоемких заявок. Заявки с малыми требованиями к ресурсу получают дополнительные возможности к его использованию. Каждый следующий минимум у π_1 означает практическую невозможность одновременного обслуживания 3, 2 и, наконец, уже одной ресурсоемкой заявки.

Данные, приведенные на рисунке 4.3, показывают изменение доли занятости канала «легким» и «тяжелым» трафиком. Когда потери сессий малы, среднее число занятых единиц ресурса для обоих потоков примерно равны поскольку равны величины потенциального трафика каждого потока. С ростом потерь сессии «легкого» трафика уже доминируют в использовании ресурса узла доступа. Как видно из представленных данных, заявки с малым использованием ресурса могут практически полностью вытеснить из обслуживания ресурсоемкие заявки. Необходимо отметить, что острота обозначенных проблем снижается, если вести передачу трафика с эластичными свойствами с использованием дисциплины разделения ресурса *PS*. Это свойство показано на рисунке 4.2 и рисунке 4.4. Задачи, относящиеся к созданию условий по дифференцированному обслуживанию гетерогенного трафика можно решать, используя разделение ресурса на слайсы, а также применяя резервирование ресурса. Эти вопросы будут рассмотрены в подразделах 4.3 – 4.5.

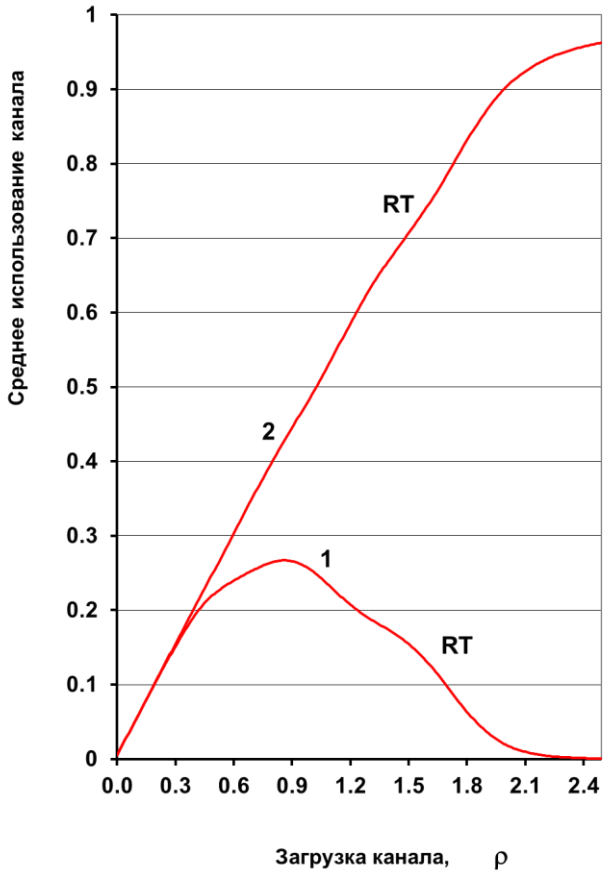


Рисунок 4.3 — Зависимость среднего использования канала от ρ для 1-го сценария распределения ресурса

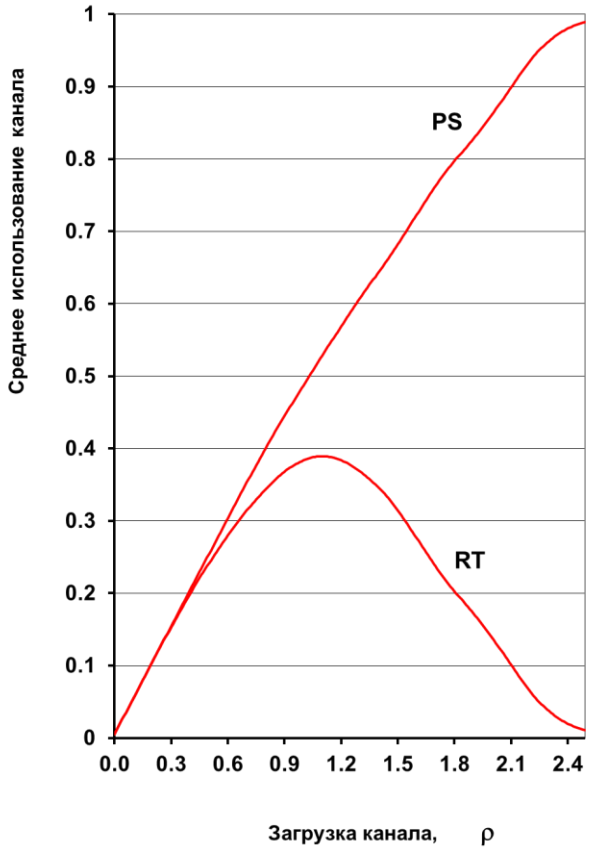


Рисунок 4.4 — Зависимость среднего использования канала от ρ для 2-го сценария распределения ресурса

4.2.2. Анализ эффективности совместного обслуживания гетерогенного трафика при использовании дисциплины *PS*

В разделе 3.4 было продемонстрировано, что использование дисциплины *PS* улучшает характеристики обслуживания эластичного трафика: уменьшаются вероятность потерь сессий и среднее время доставки файла и увеличивается коэффициент использования единицы ресурса. Покажем, что применение дисциплины *PS* также улучшает характеристики совместного обслуживания сессий трафика реального времени и данных по сравнению со сценарием, когда трафик данных обслуживается по сценарию обслуживания трафика реального времени, т.е. с использованием одной канальной единицы. В ситуации использования дисциплины *PS* графические зависимости характеристик будут обозначаться символом *PS*, в случае применения для всех потоков правил обслуживания трафика реального времени — символом *RT*.

Вернемся к исследованию особенностей использования выделенного ресурса трафиком реального времени и эластичным трафиком, рассмотренному в подразделе 4.2.1, и изменим требование к ресурсу у сессий трафика реального времени положив $b_1 = 10$ к.е. Величина ρ см. (4.1) меняется в пределах $0,5 \le \rho \le 1,5$. На рисунках 4.5 и 4.6 рассмотрены, соответственно, поведение доли потерянных сессий трафика реального времени и эластичных данных для обоих

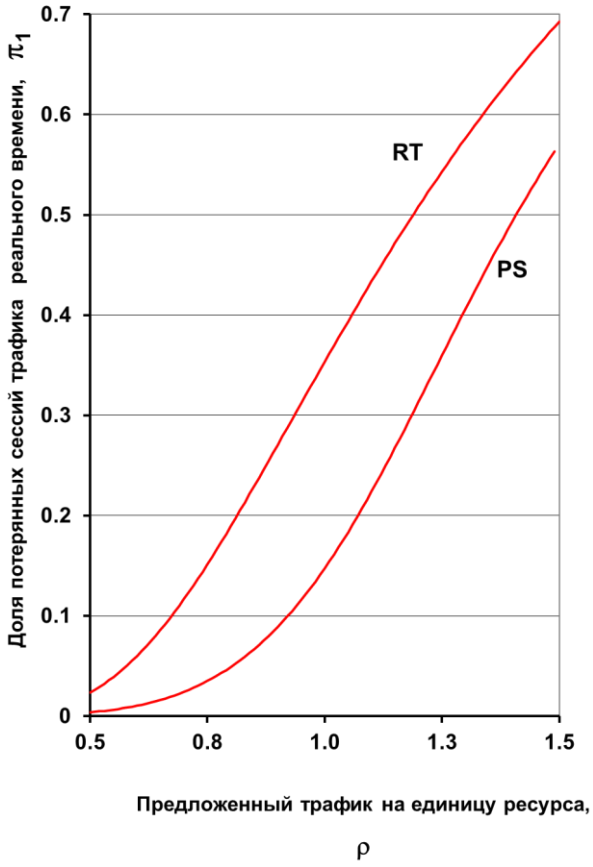


Рисунок 4.5 — Зависимость потерь сессий реального времени от ρ для дисциплин *PS* и *RT*

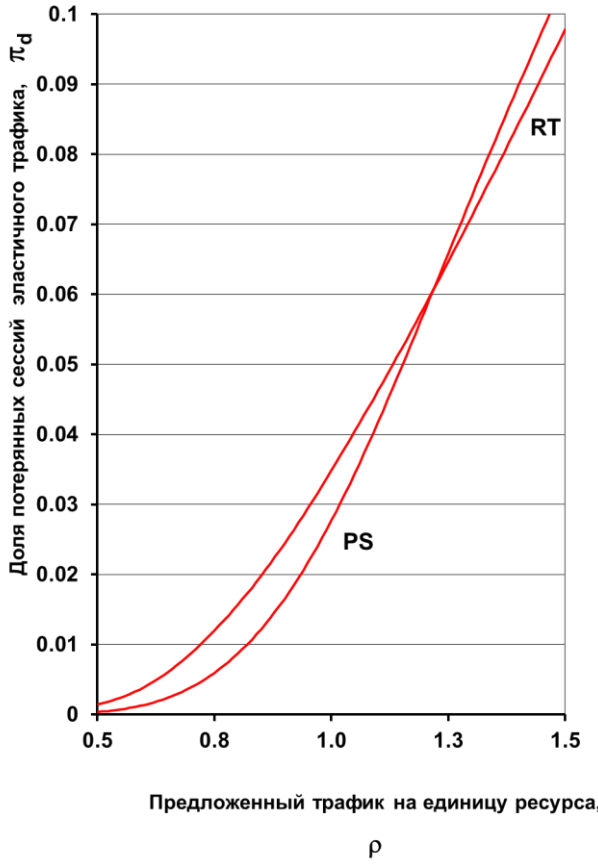


Рисунок 4.6 — Зависимость потерь сессий передачи эластичных данных от ρ для дисциплин *PS* и *RT*

сценариев распределения выделенного ресурса. Из представленных данных следует, что использование дисциплины *PS* существенно уменьшает значение потерь сессий трафика реального времени по сравнению с дисциплиной *RT* во всем диапазоне изменения ρ. Доля потерь сессий эластичного трафика вначале уменьшается, затем начиная со значения $\rho > 1$ увеличивается и становится больше аналогичной характеристики, найденной в условиях использования дисциплины *RT*. Этот результат объясняется тем, что в этой области изменения ρ для дисциплины

RT большая часть сессий трафика реального времени теряется, освобождая тем самым ресурс для обслуживания эластичного трафика.

На рисунках 4.7 и 4.8 рассмотрены, соответственно, зависимости от ρ среднего использования канала на передачу сессий трафика реального времени и эластичных данных. С ростом ρ среднее использование канала на пересылку трафика реального времени увеличивается и его значение существенно больше для сценария *PS* нежели для сценария *RT*.

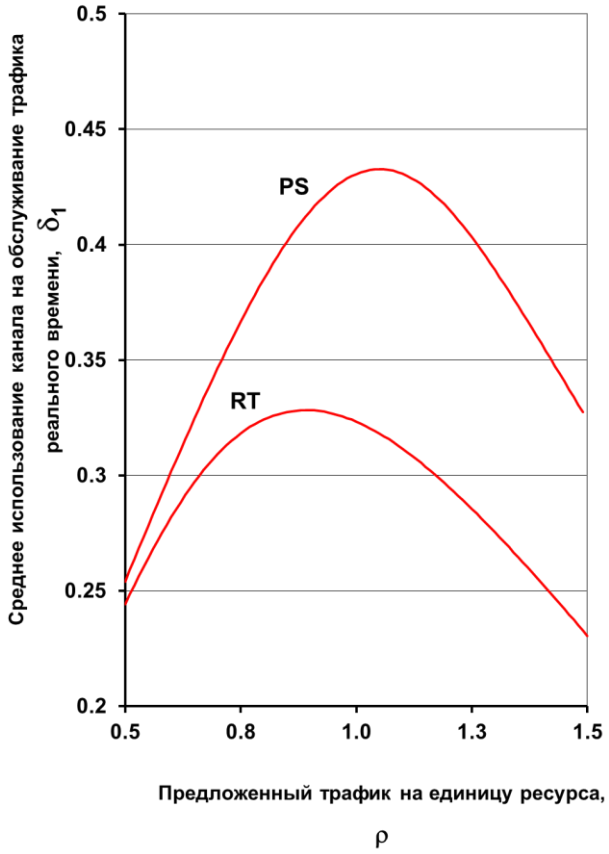


Рисунок 4.7 — Зависимость среднего использования канала от ρ для сессий трафика реального времени

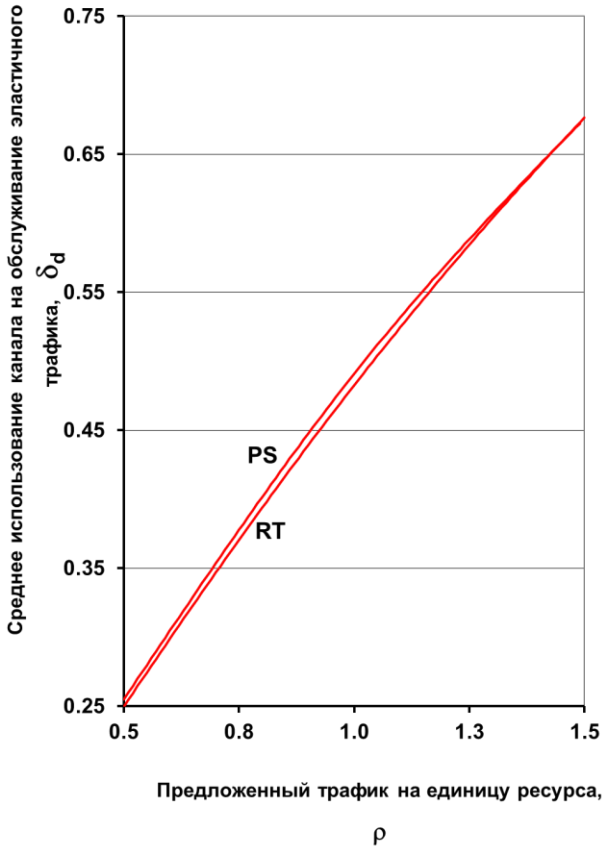


Рисунок 4.8 — Зависимость среднего использования канала от ρ для сессий передачи эластичных данных

В ситуации перегрузки происходит вытеснение трафика реального времени и среднее использование канала уменьшается для обоих сценариев распределения ресурса. Отметим, что и в этой ситуации остается преимущество *PS* над *RT*. Среднее использование канала на передачу эластичного трафика слабо зависит вида сценария распределения ресурса, монотонно увеличивается с ростом ρ с небольшим преимуществом сценария *PS* над сценарием *RT*. Этот факт объясняется тем, что в обоих случаях потери сессий эластичного трафика малы. В ситуации малых

потерь эта работа, отнесенная на один канал, одинакова как для дисциплины *PS*, так и для дисциплины *RT*, поэтому соответствующие кривые близки друг к другу (см. рисунок 3.9).

На рисунке 4.9 показано, что использование динамического распределения ресурса при обслуживании трафика данных позволяет существенно увеличить загрузку канала по сравнению с использованием при передаче данных сценария, основанного на принципах обслуживания трафика реального времени. Выигрыш достигает $10-20\%$. Этот факт необходимо учитывать при разработке сценариев распределения ресурса в мультисервисных беспроводных узлах доступа. Преимущество сценария *PS* связано с использованием для передачи файлов всех имеющихся свободных каналов. Среднее число каналов, занятых на передачу одного файла, в зависимости от ρ показано на рисунке 4.10. Понятно, что в области малых потерь эта величина достигает максимального значения и затем с увеличением ρ значение k_d стремится к единице.

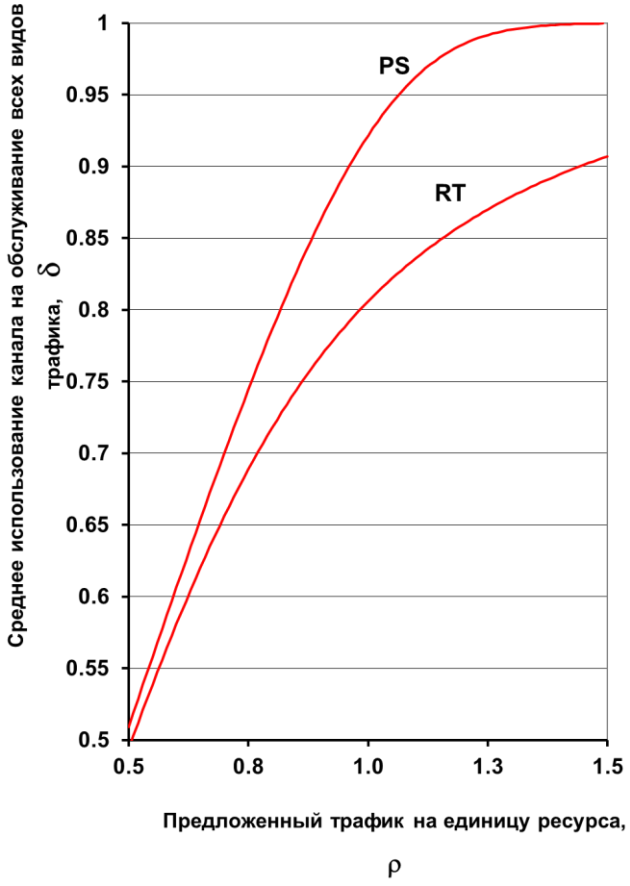


Рисунок 4.9 — Среднее использование канала в зависимости от ρ и сценария распределения ресурса при обслуживании трафика данных

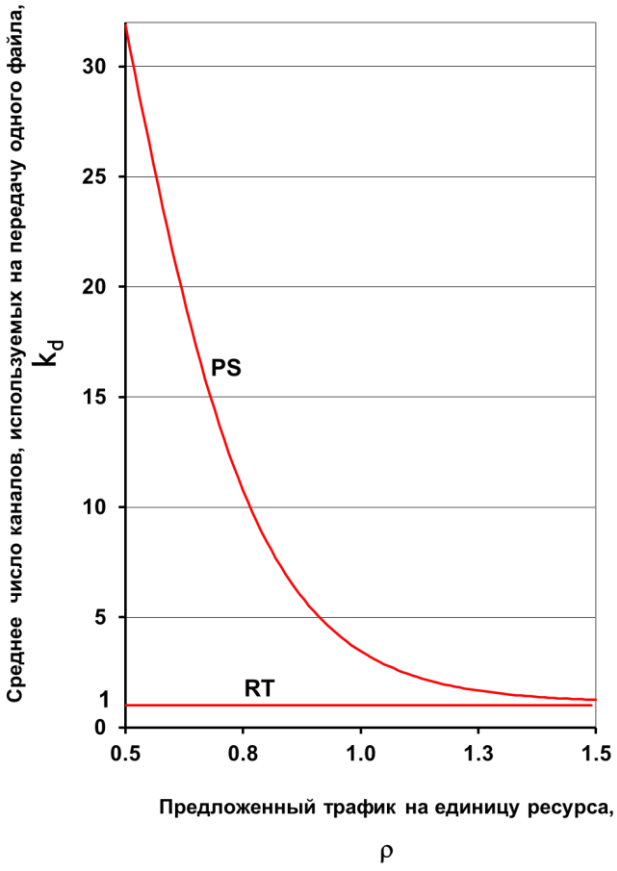


Рисунок 4.10— Среднее число каналов, используемых на передачу одного файла, в зависимости от ρ и сценария распределения ресурса при обслуживании трафика данных

4.3. Сценарии эффективного обслуживания гетерогенного трафика

Как было показано в предыдущем подразделе (см. рисунки 4.1 – 4.4) совместное обслуживание информационных потоков с ярко выраженной неоднородностью требований заявок к ресурсу передачи может привести к неконтролируемому перераспределению ресурса в пользу потоков сессий с относительно малыми требованиями к скорости передачи. Этот результат может нарушить принятое соглашение об обслуживании. Избавиться от перечисленных трудностей можно создав условия по дифференцированному обслуживанию входящих информационных потоков. Сделать это можно используя положения концепции Network Slicing. Рассмотрим две ее реализации:

- Статичный слайсинг (англ. Static Slicing — SS). Для данного сценария имеющийся ресурс делится между поступающими потоками в определенной пропорции, зависящей от требований сессий связи к показателям качества обслуживания. Выделение определенного объема ресурса, который носит название «слайс», выполняется для группы информационных потоков с примерно одинаковыми требованиями к скорости передачи, либо для одного потока, если отмеченное объединение потоков невозможно.
- Динамичный слайсинг (англ. Dynamic Slicing — DS). В рассматриваемом сценарии распределение выделенного объема ресурса осуществляется на динамической основе и зависит от его загрузки. До определенного уровня занятости ресурса он используется всеми поступающими потоками сессий связи. С увеличением загрузки ресурса часть поступающих потоков заявок получает отказ, создавая тем самым приоритет в использовании ресурса у выделенной группы информационных потоков. Ограничение доступа можно проводить разными способами. В данной работе будет использоваться процедура резервирования, введенная в подразделе 2.4.1.

Исследуем применение сформулированных сценариев для создания условий по дифференцированному обслуживанию гетерогенного трафика. Рассмотрим решение следующих трех задач:

1. Для заданного объема ресурса v, выраженного в к.е., найти разделение ресурса на слайсы с тем, чтобы поступающие потоки сессий обслуживались с одинаковыми характеристиками, выраженными в значениях доли потерянных сессий для трафика реального времени и данных.

2. Найти минимальный объем ресурса v, выраженный в к.е., и разделение ресурса на слайсы с тем, чтобы поступающие потоки сессий обслуживались с требуемыми π одинаковыми значениями характеристик, выраженными в значениях доли потерянных сессий для трафика реального времени и данных.
3. Найти минимальный объем ресурса v, выраженный в к.е., и разделение ресурса на слайсы с тем, чтобы поступающие потоки сессий обслуживались с требуемыми π_e и π_d значениями характеристик, выраженными в значениях доли потерянных сессий, соответственно, для трафика реального времени и данных.

Рассмотрим решение перечисленных задач для двух моделей обслуживания гетерогенного трафика, представляющего из себя смесь «тяжелого» трафика видеоконтента и «легкого» трафика данных. В первой модели трафик данных представляет из себя сессии передачи видеоконтента с низким качеством, требующим относительно невысокую скорость передачи. Он обслуживается по правилам трафика реального времени. Каждый файл передается с использованием возможностей одной канальной единицы. Для оценки характеристик совместного обслуживания сессий используются модели и алгоритмы, рассмотренные в подразделе 3.3. Во второй модели трафик данных обладает эластичными свойствами, например, представляя из себя файлы, получающиеся после записи видеоконтента в буфер. Он обслуживается по правилам эластичного трафика. Минимальный объем используемого ресурса составляет одну канальную единицу. Для оценки характеристик совместного обслуживания сессий используются модели и алгоритмы, рассмотренные в подразделах 2.4, 2.5, 3.3 — 3.5.

Решение второй и третьей из перечисленных задач близки по последовательности действий и по полученным результатам, поэтому ограничимся изложением решения только для второй задачи.

4.4. Дифференцированное обслуживание неоднородного трафика реального времени с использованием резервирования

4.4.1 Параметры модели

Рассмотрим модель совместного обслуживания гетерогенного трафика с параметрами, введенными в подразделе 4.2.1. Напомним их значения. Предполагается, что базовая станция LTE может обслужить одновременно максимум 80 абонентов, предоставив каждому пользователю ее

услуг скорость не ниже 1 Мбит/с. Предположим, что: $v = 80$ к.е., скорость одного виртуального канала передачи информации $r = 1$ Мбит/с, средний объем передаваемого файла F равен 10 Мбайт или 80 Мбит. Величина μ_d определяется из соотношения $\frac{1}{\mu_d} = \frac{80}{1} = 80$ сек. Это максимальное среднее время передачи файла с использованием передаточных возможностей одного канала. Совместно с эластичным трафиком обслуживаются два потока сессий трафика реального времени. Требуемая скорость передачи в к.е.: $b_1 = 5$ к.е., $b_2 = 10$ к.е.. Выбор входных параметров указывает, на совместное обслуживание информационных потоков с выраженной неоднородностью требований к ресурсу.

В дальнейшем для удобства при проведении вычислений примем, что среднее время обслуживания сессий всех видов трафика равно единице, т.е. $\mu_1 = \mu_d = \mu_d = 1$, а интенсивности поступления сессий пересчитаны в эрланги. При передаче эластичного трафика предполагается, что за единицу времени принято среднее время передачи файла одним каналом. Предполагается, что потенциальная нагрузка, создаваемая каждым потоком одинакова и удовлетворяет соотношению

$$\lambda_1 = \frac{v\rho}{3b_1}; \quad \lambda_2 = \frac{v\rho}{3b_2}; \quad \lambda_d = \frac{v\rho}{3}, \tag{4.4}$$

где ρ — потенциальная нагрузка на один канал (см. (4.1), (4.2))

$$\rho = \frac{\lambda_1 b_1 + \lambda_2 b_2 + \lambda_d}{v}. \tag{4.5}$$

При проведении вычислений будем полагать, что $\rho = 1$.

4.4.2. Статичный слайсинг

Рассмотрим решение 1-й и 2-й из сформулированных задач с использованием возможностей статичного слайсинга для 1-й модели формирования трафика данных (см. подраздел 4.3), где трафик данных представляет из себя сессии передачи видеоконтента с низким качеством, требующим относительно невысокую скорость передачи. Он обслуживается по правилам трафика реального времени. Каждый файл передается с использованием возможностей одной канальной единицы. Для используемого выбора значений входных параметров получаем: $\pi_1 = 0{,}180821$; $\pi_2 = 0{,}360495$; $\pi_d = 0{,}035750$. При одинаковой потенциальной загрузке

ресурса «легкий» трафик получает преимущество в его занятии. На рисунке 4.11 показан результат выравнивания значений потерь всех видов трафика на имеющемся объеме ресурса с использованием статичного слайсинга. Потери сессий «тяжелого» трафика выравнены с помощью процедуры резервирования ресурса следующим выбором функций блокировки:

$$\varphi_1(i)=0, \quad i=0,1,\ldots,v_1-b_2; \tag{4.6}$$

$$\varphi_1(i)=1, \quad i=v_1-b_2+1,\ v_1-b_2+2,\ldots,v_1;$$

$$\varphi_2(i)=0, \quad i=0,1,\ldots,v_1-b_2;$$

$$\varphi_2(i)=1, \quad i=v_1-b_2+1,\ v_1-b_2+2,\ldots,v_1.$$

Здесь v_1 — объем слайса, используемого для обслуживания «тяжелого» трафика. Объем слайса, используемого для передачи «легкого» трафика $v_2=v-v_1$. Для проведения вычислений использовалась модель и алгоритмы ее расчета, рассмотренные в подразделе 3.3.

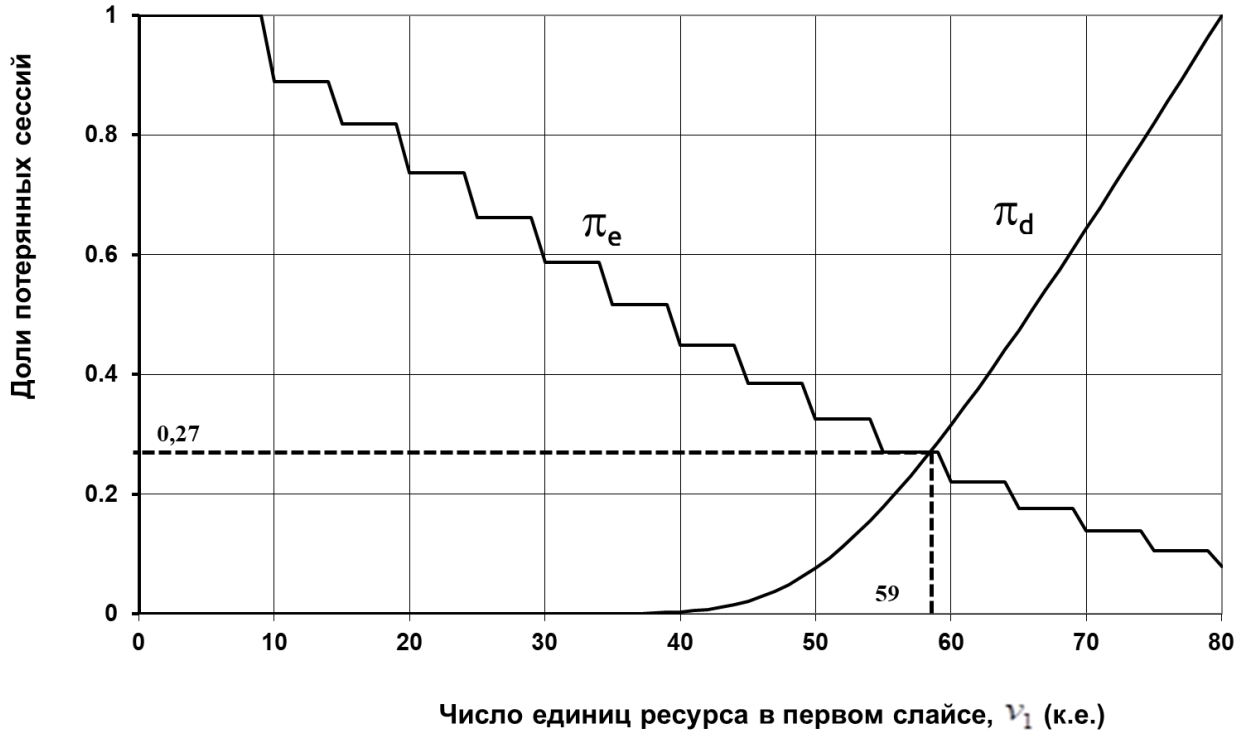


Рисунок 4.11— Использование статичного слайсинга для выравнивания потерь всех типов трафика. Первая модель формирования трафика.

Из результатов вычислений получаем, что объем 1-го слайса для обслуживания «тяжелого» трафика выбирается из соотношения $v_1 = 59$ к.е., объем 2-го слайса для обслуживания трафика данных $v_2 = 21$ к.е. Общий уровень потерь $\pi \approx 0,27$. Далее перейдем к решению 2-ой задачи. Найдем объем ресурса и размеры слайсов, чтобы достичь потерь сессий всех видов трафика на уровне 0,03.

Из результатов вычислений, представленных соответственно на рисунках 4.12 и 4.13, получаем, что объем 1-го слайса для обслуживания «тяжелого» трафика выбирается из соотношения $v_1 = 95$ к.е., объем 2-го слайса для обслуживания трафика данных $v_2 = 34$ к.е. Общий объем $v = 129$ к.е. Проделанные вычисления показывают, что решение обеих задач не вызывает затруднений и сводится к использованию модели совместного обслуживания трафика и алгоритмов ее расчета, рассмотренных в подразделе 3.3.

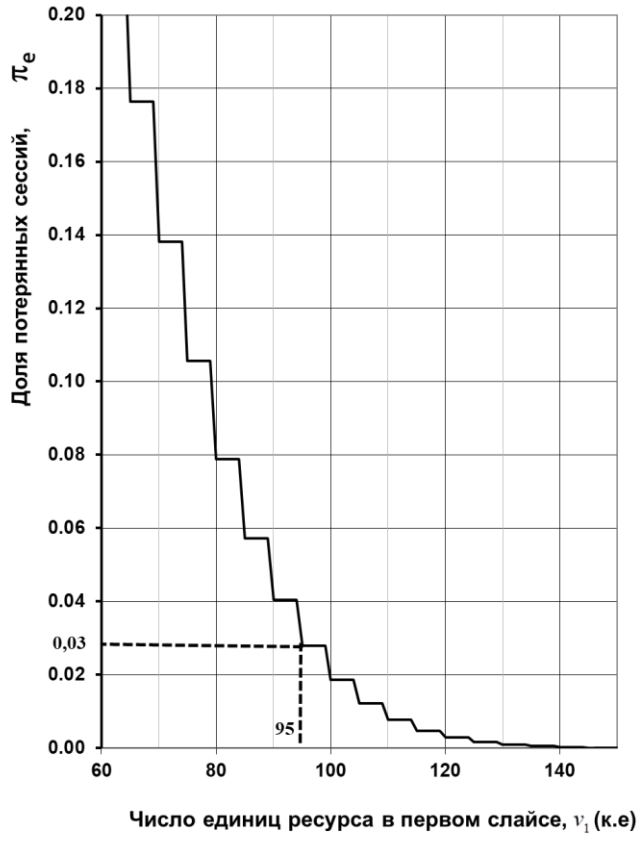


Рисунок 4.12 — Оценка требуемого объема 1-го слайса для обслуживания «тяжелого» трафика на уровне 0,03. Первая модель формирования трафика

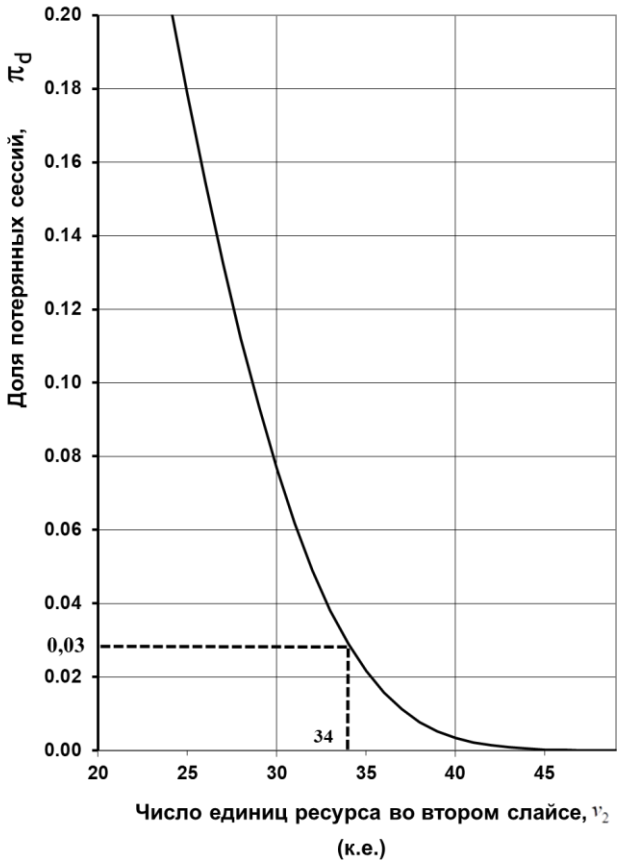


Рисунок 4.13 — Оценка требуемого объема 2-го слайса для обслуживания трафика данных на уровне 0,03. Первая модель формирования трафика

4.4.3. Динамичный слайсинг

Теперь рассмотрим решение 1-й и 2-й из сформулированных задач с использованием возможностей динамичного слайсинга. Начнем с 1-й задачи. Все три потока сессий обслуживаются общим ресурсом из 80 каналов. Потери сессий «тяжелого» трафика выравнены с помощью процедуры резервирования ресурса заданного соотношениями (4.6). Используется процедура контроля доступа данных. Она реализуется выбором функции блокировки

$$\varphi_d(i) = 0, \quad i = 0,1,\ldots, v_d - 1 - c_r; \tag{4.7}$$

$$\varphi_1(i) = 1, \quad i = v_d - c_r,\ v_d - c_r + 1,\ \ldots,\ v_d,$$

где: v_d — объем ресурса, который можно занять для передачи данных («легкий» трафик), c_r — число единиц ресурса, резервируемого в пользу «тяжелого» трафика. Для проведения вычислений использовались модель и алгоритмы ее расчета, рассмотренные в подразделе 3.3. Расчетные материалы представлены на рисунке 4.14.

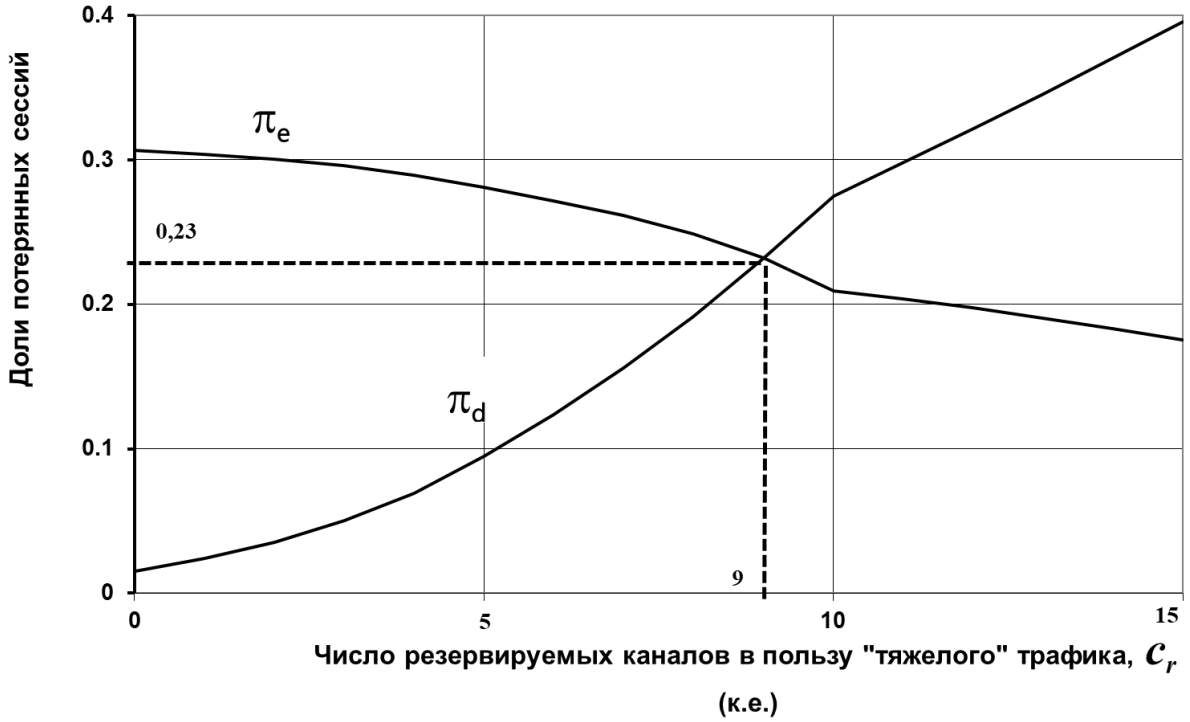


Рисунок 4.14 — Использование динамичного слайсинга для выравнивания потерь всех типов трафика. Первая модель формирования трафика

Результаты вычислений показывают, что динамичный слайсинг по сравнению с использованием статичного слайсинга (см. рисунок 4.11), обеспечивает на одном и том же объеме ресурса меньший уровень потерь. Число резервируемых каналов $c_r = 9$ к.е. Перейдем к решению 2-ой задачи. Значения характеристик показаны на рисунке 4.15.

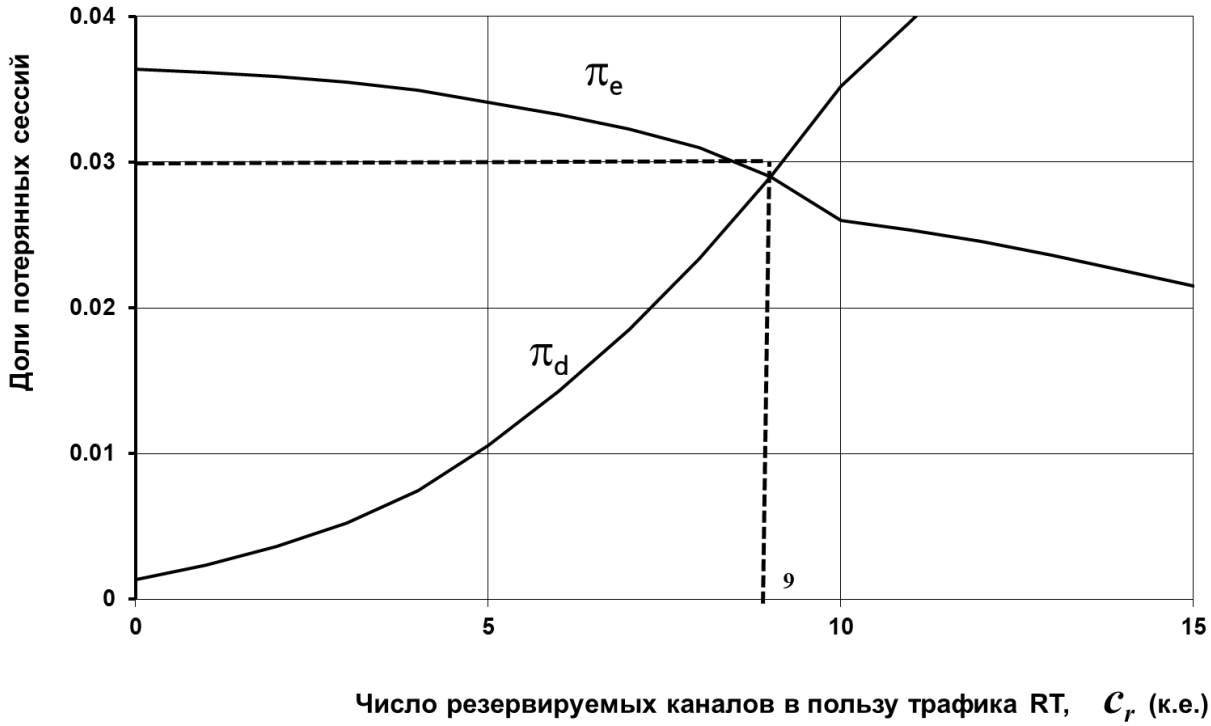


Рисунок 4.15 — Использование динамичного слайсинга для выравнивания потерь всех типов трафика на уровне 0,03. Первая модель формирования трафика

Потери сессий «тяжелого» трафика выравнены с помощью процедуры резервирования ресурса заданного соотношениями (4.6). Используется процедура контроля доступа данных. Она реализуется выбором функции блокировки из соотношений (4.7). Минимальное значение v, на котором достигается требуемое условие $\max(\pi_1, \pi_2, \pi_d) \le 0.03$ определяется из соотношения $v = 121$ к.е. Число резервируемых каналов $C_r = 9$ к.е. Выигрыш по сравнению с использованием статичного сценария составляет $\frac{129-121}{129} \approx 6\%$.

4.5. Дифференцированное обслуживание неоднородного трафика реального времени и эластичных данных с использованием резервирования

4.5.1. Статичный слайсинг

Сравним результаты использования статичного и динамичного слайсинга при решении 1-й и 2-й задач (см. подраздел 4.3) в условиях применения 2-ой модели формирования и обслуживания трафика данных. В этой ситуации трафик данных обладает свойством эластичности и обслуживается с использованием дисциплины *PS*. Выберем значения входных параметров модели из предпосылок, использованных в подразделе 4.4.1.

В анализируемой модели трафик данных представляет из себя последовательность поступления файлов, обладающих свойством эластичности. Это может быть, например, передача видеоконтента с промежуточной буферизацией. Он обслуживается с использованием дисциплины Processor Sharing, т.е. для передачи файла задействован весь свободный ресурс, оставшийся от обслуживания «тяжелого» трафика реального времени. В ситуации перегрузки используемый ресурс может уменьшен до одного канала, но не меньше. Для оценки характеристик совместной организации сессий используются модели и алгоритмы, рассмотренные в подразделах 2.4, 2.5, 3.3—3.5. Для используемого выбора значений входных параметров получаем: $\pi_1 = 0,094151;\ \pi_2 = 0,179877;\ \pi_d = 0,028384.$ В рассматриваемых условиях трафик данных получает преимущество в занятии ресурса. На рисунке 4.16 показан результат выравнивания значений потерь всех видов трафика на имеющемся объеме ресурса в результате использования статичного слайсинга. Потери сессий «тяжелого» трафика выравнены с помощью процедуры резервирования ресурса заданного соотношениями (4.6).

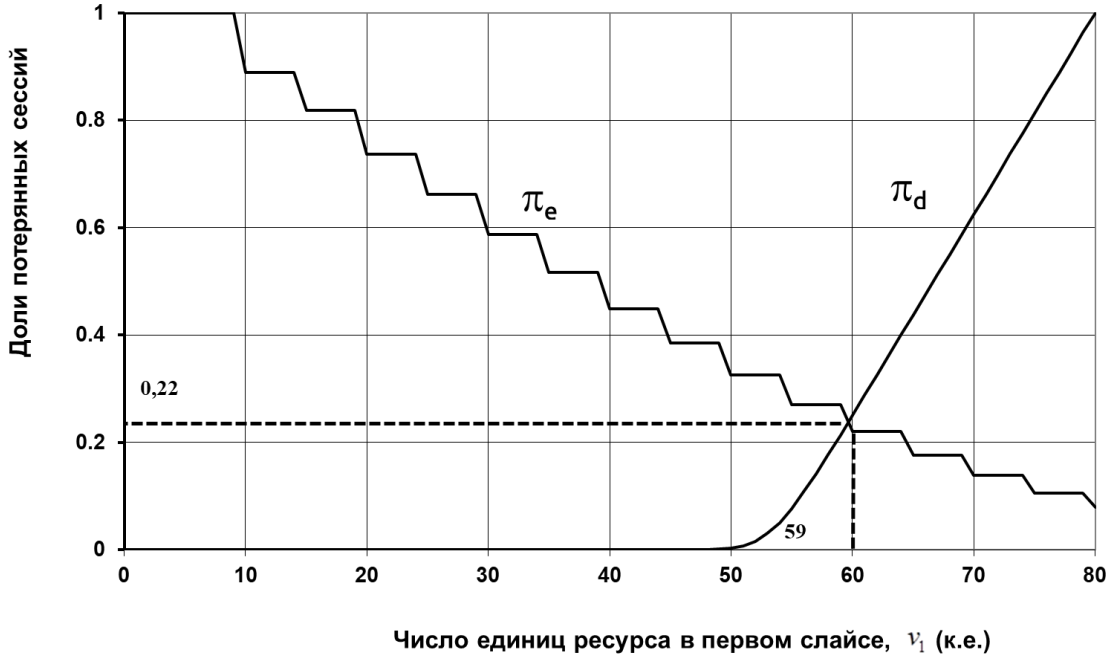


Рисунок 4.16 — Использование статичного слайсинга для выравнивания потерь всех типов трафика. Вторая модель формирования трафика.

Из результатов вычислений получаем, что объем 1-го слайса для обслуживания «тяжелого» трафика выбирается из соотношения $v_1 = 60$ к.е., объем 2-го слайса для обслуживания трафика данных $v_2 = 20$ к.е. Общий уровень потерь $\pi \approx 0,22$. Использование дисциплины *PS* уменьшает уровень выравнивания значений потерь по сравнению с обслуживанием трафика по правилам сервисов реального времени (см. рисунок 4.11). Далее перейдем к решению 2-ой задачи. Найдем объем ресурса и размеры слайсов, чтобы достичь потерь сессий всех видов трафика на уровне 0,03.

Из результатов вычислений, представленных соответственно на рисунках 4.17 и 4.18, получаем, что объем 1-го слайса для обслуживания «тяжелого» трафика выбирается из соотношения $v_1 = 95$ к.е., как и раньше, поскольку процесс обслуживания трафика реального времени в слайсе не изменился. Объем 2-го слайса для обслуживания трафика данных теперь стал равным $v_2 = 28$ к.е. Общий объем $v = 123$ к.е. Проделанные вычисления показывают, что решение обеих задач не вызывает затруднений и сводится к использованию модели совместного обслуживания трафика и алгоритмов ее расчета, рассмотренных в подразделах 3.3—3.5. Отметим, что общий объем использованного ресурса уменьшился из-за преимуществ в обслуживании трафика данных, создаваемых дисциплиной *PS*.

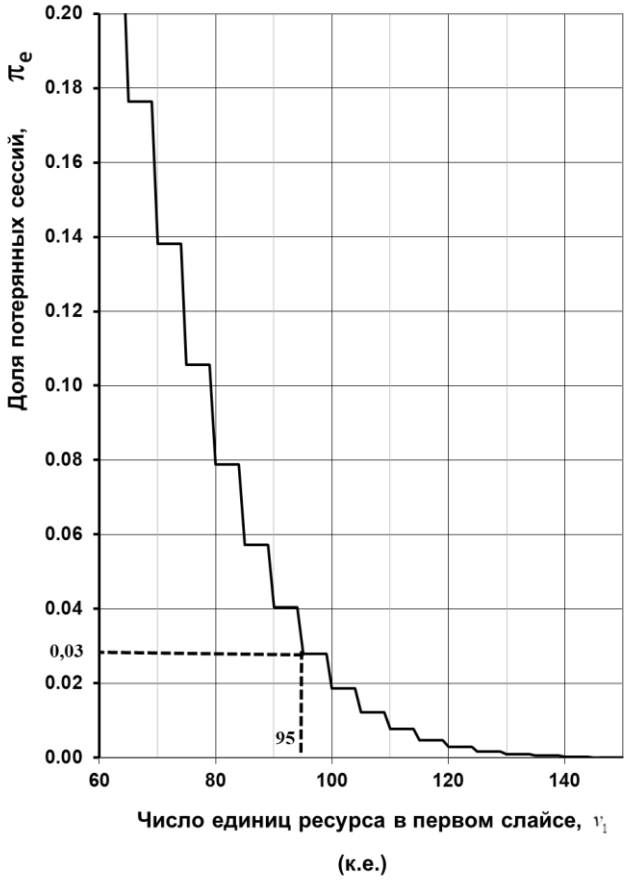


Рисунок 4.17 — Оценка требуемого объема 1-го слайса для обслуживания «тяжелого» трафика на уровне 0,03. Вторая модель формирования трафика.

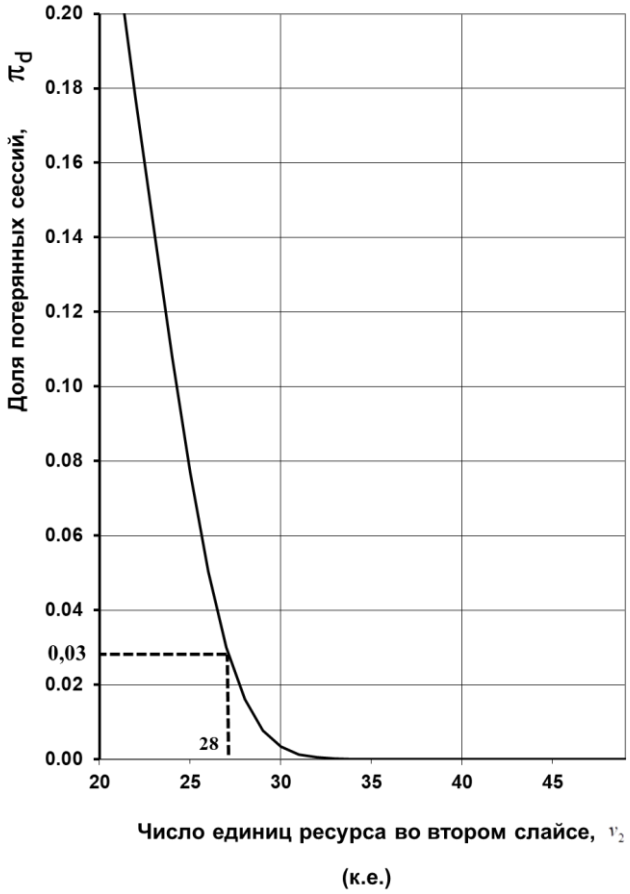


Рисунок 4.18 — Оценка требуемого объема 2-го слайса для обслуживания трафика данных на уровне 0,03. Вторая модель формирования трафика

4.5.2. Динамичный слайсинг

Теперь рассмотрим решение 1-й и 2-й из сформулированных задач с использованием возможностей динамичного слайсинга. Начнем с 1-й задачи. Все три потока сессий обслуживаются общим ресурсом из 80 каналов. Потери сессий «тяжелого» трафика выравняны с помощью процедуры резервирования ресурса заданного соотношениями (4.6). Используется процедура контроля доступа данных. Она реализуется выбором функции блокировки из соотношений (4.7). Для проведения вычислений использовалась модель и алгоритмы ее расчета, рассмотренные в подразделах 2.4, 2.5, 3.3—3.5. Расчетные материалы представлены на рисунке 4.19.

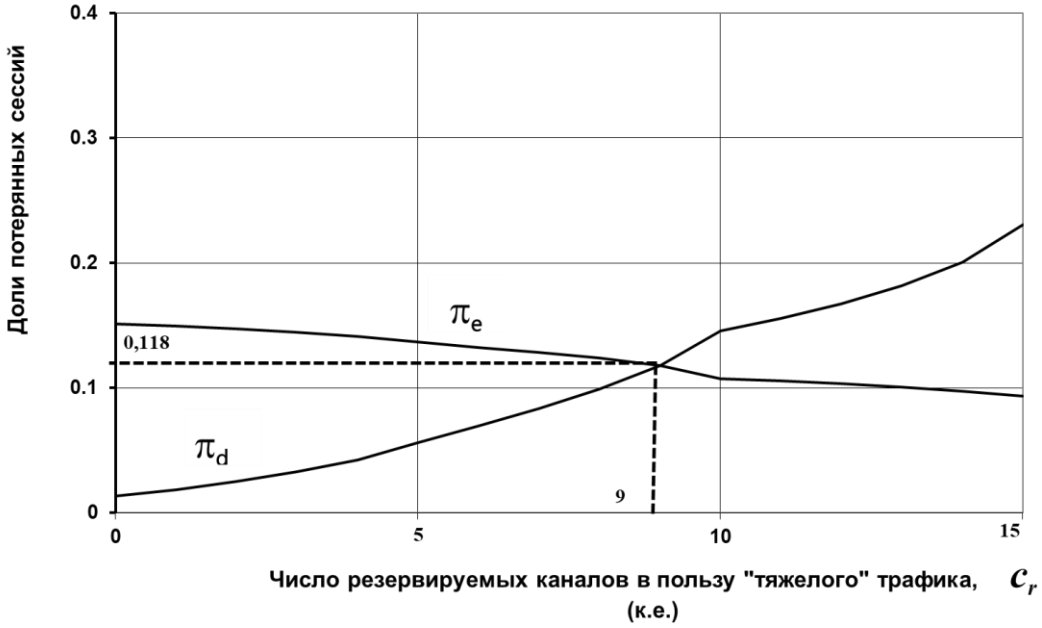


Рисунок 4.19 — Использование динамичного слайсинга для выравнивания потерь всех типов трафика. Вторая модель формирования трафика.

Результаты вычислений показывают, что динамичный слайсинг по сравнению с использованием статичного слайсинга (см. рисунок 4.11), обеспечивает на одном и том же объеме ресурса меньший уровень потерь. Число резервируемых каналов $c_r = 9$ к.е. Перейдем к решению 2-ой задачи. Значения характеристик показаны на рисунке 4.20.

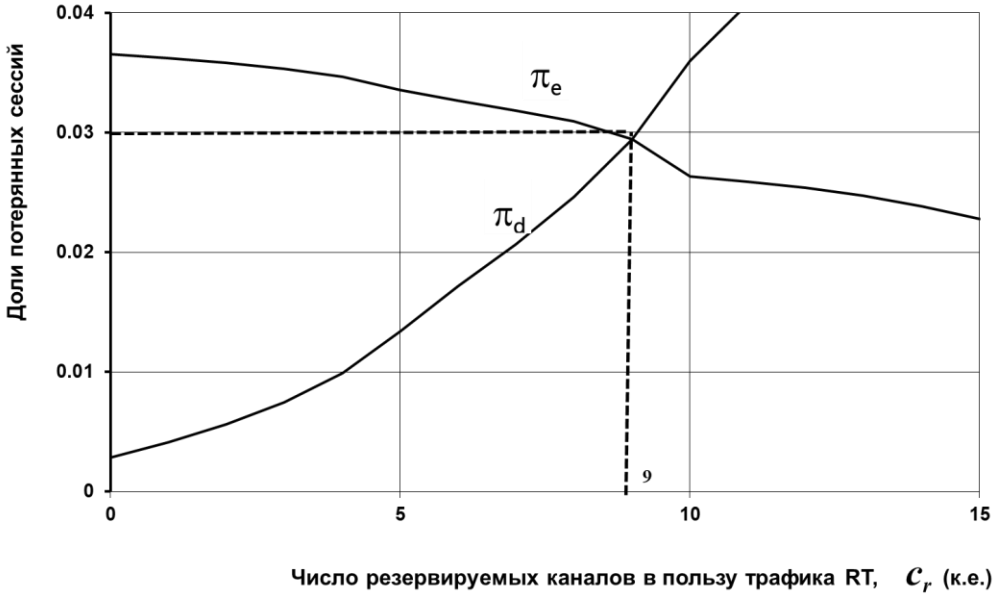


Рисунок 4.20 — Использование динамичного слайсинга для выравнивания потерь всех типов трафика на уровне 0,03. Вторая модель формирования трафика.

Потери сессий «тяжелого» трафика выравнены с помощью процедуры резервирования ресурса заданного соотношениями (4.6). Используется процедура контроля доступа данных. Она реализуется выбором функции блокировки из соотношений (4.7). Минимальное значение v, на котором достигается требуемое условие $\max(\pi_1, \pi_2, \pi_d) \le 0,03$ определяется из соотношения $v = 101$ к.е. Число резервируемых каналов $c_r = 9$ к.е. Выигрыш по сравнению с использованием статичного сценария составляет $\frac{123-101}{123} \approx 18\%$.

4.6. Выводы по результатам четвертого раздела

1. Выполнен анализ особенностей совместного обслуживания информационных потоков с ярко выраженной неоднородностью требований заявок к ресурсу передачи. Проведенное исследование показало, что с ростом нагрузки на канал происходит неконтролируемое перераспределение ресурса в пользу потоков сессий с относительно малыми требованиями к скорости передачи. Этот результат может нарушить принятое соглашение об обслуживании. Избавиться от перечисленных трудностей можно создав условия по дифференцированному обслуживанию входящих информационных потоков.
2. Для создания условий по дифференцированному обслуживанию гетерогенного трафика предлагается использовать два сценария:
    - Статичный слайсинг (Static Slicing— SS). Для данного сценария имеющийся ресурс делится между поступающими потоками в определенной пропорции, зависящей от требований сессий связи к показателям качества обслуживания. Выделение определенного объема ресурса, который носит название «слайс», выполняется для группы информационных потоков с примерно одинаковыми требованиями к скорости передачи, либо для одного потока, если отмеченное объединение потоков невозможно.
    - Динамичный слайсинг (Dynamic Slicing— DS). В рассматриваемом сценарии распределение выделенного объема ресурса осуществляется на динамической основе и зависит от его загрузки. До определенного уровня занятости ресурса он используется всеми поступающими потоками сессий связи. С увеличением загрузки ресурса часть поступающих потоков заявок получает отказ, создавая тем самым приоритет в использовании ресурса у выделенной группы информационных потоков.

Для ограничения доступа сессий предлагается использоваться процедуру резервирования, основанную на фильтрации поступающих сессий с использованием функции внутренней блокировки.

3. Показано, что эффективность реализации каждого из предложенных сценариев можно исследовать с помощью комплекса моделей совместного обслуживания неоднородного трафика и алгоритмов оценки их вероятностных характеристик, введенных и исследованных во втором и третьем разделах диссертации. Разработанные алгоритмы отличаются высокой эффективностью реализации и могут применяться для всех практически интересных значений входных параметров.

4. С использованием разработанных алгоритмов проведено сравнение предложенных сценариев создания условий по дифференцированному обслуживанию неоднородного трафика. Сравнение проводилось по результатам решения следующих двух задач:

   - Для заданного объема ресурса v , выраженного в к.е., найти разделение ресурса на слайсы с тем, чтобы поступающие потоки сессий обслуживались с одинаковыми характеристиками, выраженными в значениях доли потерянных сессий для трафика реального времени и данных.

   - Найти минимальный объем ресурса v , выраженный в к.е., и разделение ресурса на слайсы с тем, чтобы поступающие потоки сессий обслуживались с требуемыми π одинаковыми значениями характеристик, выраженными в значениях доли потерянных сессий для трафика реального времени и данных.

5. Перечисленные задачи решались для двух моделей генерации и обслуживания гетерогенного трафика, встречающихся в практических приложениях, представляющего из себя смесь «тяжелого» трафика видеоконтента и «легкого» трафика данных. В первой модели трафик данных представляет из себя сессии передачи видеоконтента с низким качеством, требующим относительно невысокую скорость передачи. Он обслуживается по правилам трафика реального времени. Каждый файл передается с использованием возможностей одной канальной единицы. Во второй модели трафик данных обладает эластичными свойствами, например, представляя из себя файлы, получающиеся после записи видеоконтента в буфер. Он обслуживается по правилам эластичного трафика. Минимальный объем используемого ресурса составляет одну канальную единицу.

6. Выполненное численное исследование показало, что использование динамического слайсинга позволяет на $5-50\%$ уменьшить потери при дифференцированном обслуживании, направленном на выравнивание потерь сессий на фиксированном объеме ресурса, и на $5-20\%$ уменьшить требование к объему ресурса, обеспечивающего требуемый уровень потерь сессий, по сравнению с применением для этих же целей статичного слайсинга. Наибольший эффект применение предложенной версии динамического слайсинга приносит в ситуации обслуживания эластичного трафика данных с использованием дисциплины Processor Sharing.

Заключение

Основные результаты работы состоят в следующем.

1. Построена и исследована обобщенная модель обслуживания неоднородного трафика в беспроводном узле доступа, которая в отличие от известных моделей позволила учесть совместное влияние основных значимых факторов, определяющих совместное обслуживание трафика реального времени и эластичных данных. Среди них: наличие приоритета у трафика реального времени; использование дисциплины Processor Sharing при передаче эластичного трафика; ограничение по доступу для всех видов трафика, зависящее от общего уровня занятости ресурса.
2. С использованием модели получены выражения для оценки характеристик качества обслуживания заявок через значения входных параметров и стационарных вероятностей обобщенной модели беспроводного узла доступа. Среди них для каждого типа трафика: доли потерянных сессий, средний объем занятого ресурса, среднее время доставки сообщения, средний объем ресурса, используемый каждой сессией и т.д. Полученные выражения позволяют анализировать действие разного рода процедур, направленных на повышение эффективности использования ресурса передачи узлов доступа и создание условий по дифференцированному обслуживанию потоков неоднородного трафика, основанных на ограничении доступа, зависящего от общего уровня занятости ресурса.
3. Получено алгебраическое представление системы уравнений равновесия исследуемой модели беспроводного узла доступа в виде, удобном для последующей реализации метода Гаусса-Зейделя. Найденное выражение дает возможность записать все уравнения системы в виде одного соотношения с коэффициентами, вычисляемыми с помощью рекуррентных формул, зависящих от компонент состояния модели. Это значительно упрощает реализацию метода и дает возможность увеличить число состояний в исследуемой модели до нескольких миллионов.
4. Получены соотношения между характеристиками обслуживания сессий, которые имеют характер законов сохранения интенсивностей поступающих и обслуженных системой потоков заявок. Найденные соотношения можно использовать для вычисления значений характеристик и косвенной оценки сходимости итерационного метода решения системы уравнений равновесия.

5. Разработанные модель и алгоритмы оценки ее характеристик позволяют анализировать действие разного рода процедур, направленных на повышение эффективности использования ресурса передачи узлов доступа и создание условий по дифференцированному обслуживанию потоков неоднородного трафика, основанных на ограничении доступа, зависящего от общего уровня занятости ресурса. Среди них динамичный слайсинг, когда распределение выделенного объема ресурса осуществляется на динамической основе и зависит от его загрузки. Для ограничения доступа сессий здесь предлагается использоваться процедуру резервирования, основанную на фильтрации поступающих сессий с использованием функции внутренней блокировки. Другой сценарий — статичный слайсинг. Для данного сценария имеющийся ресурс делится между поступающими потоками в определенной пропорции, зависящей от требований сессий связи к показателям качества обслуживания.
6. Выполненное численное исследование показало, что использование динамического слайсинга позволяет на 5 – 20% уменьшить требование к объему ресурса, обеспечивающего требуемый уровень потерь сессий, по сравнению с применением для этих же целей статичного слайсинга. Наибольший эффект применение предложенной версии динамического слайсинга приносит в ситуации обслуживания эластичного трафика данных с использованием дисциплины Processor Sharing.

Таким образом, в результате проведенных в диссертационной работе исследований построена и проанализирована процедура динамического распределения ресурса беспроводного узла доступа, позволяющая создать условия по дифференцированному обслуживанию неоднородного трафика современных коммуникационных приложений и повысить эффективность использования ресурса передачи информации. Тем самым, цель диссертационного исследования достигнута.

Список литературы

1. 3GPP TR 25.912. Feasibility Study for Evolved Universal Terrestrial Radio Access and Universal Terrestrial Radio Access Network, Release 10, section 13.5. April 2011.
2. 3GPP TR 36.814. Further Advancements for E-UTRA Physical Layer Aspects, 3-rd Generation Partnership Project, Release 9, section 10. March 2010.
3. 3GPP TR 45.820 V0.3.0 Release 13. —2015-03.
4. 3GPP TS 25.306. UE Radio Access Capabilities, Release 10, section 5. October 2011.
5. 3GPP TS 36.213. Evolved Universal Terrestrial Radio Access (E-UTRA), Physical layer procedures, version 15.2.0 Release 15. 2018.
6. Антонова В.М. Оценка ресурса передачи информации при обслуживании разнородного трафика в сетях LTE / В.М. Антонова, Д.О. Волков, М.С. Степанов // Естественные и технические науки. — 2016. — № 11. — С. 183-189.
7. Бегишев В.О. Стратегия распределения радиоресурсов в гетерогенных сетях с трафиком Narrow-Band IoT / В.О. Бегишев, А.К. Самуйлов, Д.А. Молчанов, К.Е Самуйлов // Системы и средства информатики. — 2017. — No 4. — Т. 27. — С. 64-79.
8. Вишневский В.М. Энциклопедия WiMAX. Путь к 4G / В.М. Вишневский, С.Л. Портной, И.В. Шахнович. — М.: Техносфера. — 2009. — 472 с.
9. Гельгор А.Л. Технология LTE мобильной передачи данных: учеб. пособие / А.Л. Гельгор, Е.А Попов. — СПб.: Изд-во Политехн. ун-та. — 2011. — 204 с.
10. Карякин В.Л. Методы ТВ вещания в стандарте DVB-T2 со вставкой регионального контента / В.Л. Карякин, Д.В. Карякин, Л.А. Морозова // T-Comm: Телекоммуникации и транспорт. — 2016. — Том 10. — №4. — С.41-46.
11. Клейнрок Л. Теория массового обслуживания: Пер. с англ. под ред. В. И. Неймана / Л. Клейнрок. — М.: Машиностроение. — 1979. — 452 с.
12. Корнышев Ю. Н. Теория телетрафика. Учебник для вузов / Ю. Н. Корнышев, А. П. Пшеничников, А. Д. Харкевич. —М.: Радио и связь. — 1996. — 272 с.
13. Ндайикунда Ж. Оценка качества обслуживания в сетях LTE с ограниченным числом пользователей / Ж. Ндайикунда // Технологии информационного общества. Сборник трудов

XIV Международной отраслевой научно-технической конференции «Технологии информационного общества». (18-19 марта 2020 г. Москва, МТУСИ). — М.: ИД Медиа Паблишер. — 2020. — С. 98-100.

14. Ндайикунда Ж. Построение модели совместного обслуживания разнородных устройств в гетерогенных сетях LTE / Ж. Ндайикунда // Технологии информационного общества. Сборник трудов XV Международной отраслевой научно-технической конференции «Технологии информационного общества». (3-4 марта 2021 г. Москва, МТУСИ). — М.: МТУСИ. — 2021. — С. 67-69.

15. Росляков А.В. ОКС №7: архитектура, протоколы, применение / А.В. Росляков. —М.: Эко-Трендз. —2008. — 320 с.

16. Рыжков А.Е. Системы и сети радиодоступа 4G: LTE, WiMax. / А.Е. Рыжков [и др.] // СПб.: — Линк. — 2012.— 226 с.

17. Саламех Немер. Анализ и разработка метода оценки скорости звеньев мультисервисной сети при совместном обслуживании неоднородного трафика реального времени: Дис. … канд. техн. наук: 05.12.13 / Немер Саламех: МТУСИ. — 2016. — 164 с.

18. Скрынников В.Г. Радиоподсистемы UMTS/LTE. Теория и практика / В.Г. Скрынников. — М.: Культура и спорт -2000. — 2012. — 864 с.

19. Степанов М.С. Разработка и анализ обобщённой модели обслуживания вызовов в перспективных контакт-центрах: Дис. … канд. техн. наук: 05.12.13 / М.С. Степанов: МТУСИ. — 2016. — 153 с.

20. Степанов С. Н. Планирование ресурса передачи при совместном обслуживании мультисервисного трафика реального времени и эластичного трафика данных / С. Н. Степанов, С. Степанов // Автомат и телемех. — 2017. — № 11. — С. 79-93.

21. Степанов С. Н. Эффективный алгоритм оценки требуемого объема ресурса беспроводных систем связи при совместном обслуживании гетерогенного трафика устройств Интернета Вещей / С. Н. Степанов, М. С. Степанов // Автомат и телемех. — 2019. — No. 11. — С. 108-126.

22. Степанов С.Н. Основы телетрафика мультисервисных сетей С.Н. Степанов. —М.: Эко-Трендз. — 2010. — 392 с.: ил.

23. Степанов С.Н. Построение и анализ двухпотоковой модели звена с конечным числом абонентов и возможностью внутренних блокировок / С.Н.Степанов, Немер Саламех // T-Comm: Телекоммуникации и транспорт. — 2016. — Том 10. — №9. — С. 30-37.

24. Степанов С.Н. Построение модели динамического распределения радиоресурсов LTE в гетерогенных сетях с трафиком NB-IoT / С.Н. Степанов, Ж. Ндайикунда // Технологии информационного общества. Сборник трудов XIII Международной отраслевой научно-технической конференции «Технологии информационного общества». (20-21 марта 2019 г. Москва, МТУСИ). В 2-х томах. —М.: ИД Медиа Паблишер. — 2019. Том.1. — С.136-138.

25. Степанов С.Н. Теория телетрафика: концепции, модели, приложения / С.Н. Степанов. — М.: горячая линия — Телеком. —2015. — 868 с.: ил. — (Серия «Теория и практика инфокоммуникаций»).

26. Шнепс-Шнеппе М. А. Системы распределения информации. Методы расчёта / М. А. Шнепс-Шнеппе // Справочное пособие. —М.: Связь. — 1979. — 344 с.

27. Abbasi M. NB-IoT Small Cell. A 3GPP Perspective / M. Abbasi. — 6 p.

28. Alex Z. IIS smooth streaming technical overview / Z. Alex // Microsoft Corporation. — Mar. 2009.

29. Alexandros Kaloxylos, [et al.]. View on 5G Architecture. — 2016. Available at: https://www.researchgate.net/publication/306107214 (accessed January 2022).

30. ALFOUDI Ali. An efficient resource management mechanism for network slicing in LTE network / Ali ALFOUDI, Shah NEWAZ, Abayomi OTEBOLAKU, Gyu Myoung LEE, Rubem PEREIRA // IEEE Access. — 2019. — Vol. 7. — P. 89441-89457.

31. Andrabi U.M. Cellular network resource distribution methods for the joint servicing of real-time multiservice traffic and grouped IoT traffic / U.M. Andrabi, S.N. Stepanov, J. Ndayikunda, M.G . Kanishcheva // T-Comm. — 2020. — Vol. 14. — No.10. — P. 61-69.

32. Andrew F.-L. A review of HTTP live streaming. — Jan. 2010.

33. Arun Raj L. Adaptive video streaming over HTTP through 4G wireless networks based on buffer analysis / L. Arun Raj, Dhananjay Kumar, H. Iswarya, S. Aparna and A. Srinivasan // EURASIP Journal on Image and Video Processing . — 2017. — Vol. 41. — P. 1-13.

34. Badach A. SDN Software Defined Networking. — 2020. Available at: https://www.researchgate.net/publication/341574902 (accessed February 2022).

35. Bardyn J. P. IoT: The era of LPWAN is starting now / J. P. Bardyn, T. Melly, O. Seller, N. Sornin // ESSCIRC Conference 2016: 42nd European SolidState Circuits Conference. — Sept. 2016. — P. 25-30.

36. Begishev V. Resource Allocation and Sharing for Heterogeneous Data Collection over conventional 3GPP LTE and Emerging NB-IoT Technologies / V. Begishev, V. Petrov, A. Samuylov, D. Moltchanov, S. Andreev, Y. Koucheryavy // Comput. Communicat. — 2018. — Vol. 120. — No 2. — P. 93-101.

37. Bjorklund F. Video Surveillance and Social Control in a Comparative Perspective / F. Bjorklund, O. Svenonius // Routledge. — 2013.

38. Bonald T. Calculating the flow level performance of balanced fairness in tree networks / T. Bonald, J. Virtamo // Performance Evaluation. — 2004. — Vol.58. — P.1-14.

39. Bonald T. A queueing analysis of max-min fairness, proportional fairness and balanced fairness / T. Bonald, L. Massoulie, A. Proutiere, J. Virtamo // Queueing Syst. Theory Appl. — 2006. — Vol.53. — Issue.1-2. — P.65-84.

40. Bonald T. Congestion in large balanced multirate links / T. Bonald, J. P. Haddad, R. R. Mazumdar // Proceedings of the 23rd International Teletraffic Congress. — 2011. — P.182-189.

41. Bonald T. Insensitive bandwidth sharing in data networks / T. Bonald A. Proutìere // Queueing Syst. Theory Appl. 2003. — Vol.44, issue.1. — P.69-100.

42. Bonald T. Insensitive traffic models for communication networks / T. Bonald // Discrete Event Dynamic Systems. — 2007. — Vol. 17. — No. 3. — P. 405-421.

43. Bormann C. CoAP over tcp, tls, and websockets. RFC 8323 / C. Bormann, S. Lemay, H. Tschofenig, K. Hartke, B. Silverajan, B. Raymor // RFC Editor. — Feb. 2018.

44. Boyce J. M. Overview of SHVC: Scalable extensions of the high efficiency video coding standard / J. M. Boyce, Y. Ye, J. Chen, and A. K. Ramasubramonian // IEEE Transactions on Circuits and Systems for Video Technolog. — Jan. 2016. — Vol. 26. — No. 1. — P. 20-34.

45. Brown J. A predictive resource allocation algorithm in the LTE uplink for event based M2M applications / J. Brown and J. Y. Khan // IEEE Trans. Mobile Comput. — Dec. 2015. — Vol. 14. — No. 12. — P. 2433-2446.

46. Che D. From big data to big data mining: challenges, issues, and opportunities / D. Che, M. Safran, Z. Peng // Proc. of the 18th International Conference on Database Systems for Advanced Applications. — 2013. — P. 1-15.

47. Christopher Cox. An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications / Cox Christopher // John Wiley & Sons, Ltd. — 2012. — 309 p.

48. Christopher Cox. An Introduction to LTE: LTE, LTE-Advanced, SAE, VoLTE and 4G Mobile Communications. 2nd Edition / Cox Christopher // John Wiley & Sons, Ltd. — 2014. — 488 p. ISBN: 978-1-118-81804-6.

49. Cieszynski J. Closed circuit television. 3rd edition, 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA, Elsevier. — 2007. — P. 2-10.

50. Derakhshani M. Virtualization of multi-cell 802.11 networks: Association and airtime control / M. Derakhshani, X. Wang, T. Le-Ngoc, A. Leon-Garcia // arXiv preprint arXiv: 1508.03554. — 2015.

51. Dizdarevic J., [et al.]. Survey of Communication Protocols for Internet of Things and Related Challenges of Fog and Cloud Computing Integration / J. Dizdarevic, [et al.] // ACM Computing Surveys. — 2019. — Vol. 51. — № 6. — P. 1-29.

52. Edited by Andrew Banks and Rahul Gupta. MQTT version 3.1.1. Oasis standard. — 29 October 2014.

53. Francesco C. Downlink packet scheduling in LTE cellular networks: Key design issues and a survey / C. Francesco, P. Giuseppe, A. Luigi , B. Gennaro, C. Pietro // Communications Surveys & Tutorials, IEEE, 15. Vol. 2. — 2013. — P. 678-700.

54. Ge X. 5G software defined vehicular networks / X. Ge, S. Tu, G. Mao, C.-X. Wang, T. Han // 5G Ultra-Dense Cellular Netw. — 2016. — Vol. 23. — No. 1. — P. 72-79.

55. Giuseppe P. A two level scheduling algorithm for QoS support in the downlink of LTE cellular networks / P. Giuseppe, A. G. Luigi, B. Gennaro, C. Pietro // European Wireless Conference IEEE. — 2010. — P. 246-253.

56. GSMA: 3GPP Low Power Wide Area Technologies.

57. GSMA. NB-IoT Deployment Guide to Basic Feature set Requirements. Available at: https://www.gsma.com/iot/wp-content/uploads/2019/07/201906-GSMA-NB-IoT-Deployment-Guide-v3.pdf (accessed January 2022).

58. Hoymann C. LTE release 14 outlook / C. Hoymann, D. Astely, M. Stattin, G. Wikstrom, J. F. Cheng, A. Hoglund, M. Frenne, R. Blasco, J. Huschke, and F. Gunnarsson // IEEE Communications Magazine. — June 2016. — Vol. 54. — No. 6. — P. 44-49.

59. IBM.COM. Internet connection and recommended encoding settings. Available at: https://support.video.ibm.com/hc/en-us/articles/207852117-Internet-connection-and-recommended-encoding-settings (accessed January 2022).

60. IoT ANALYTICS. State of IoT 2021: Number of connected IoT devices growing 9% to 12.3 billion globally, cellular IoT now surpassing 2 billion. — 2021.

61. iTech. Технологии связи: Введение в IoT (Интернет Вещей). Available at: https://itechinfo.ru/content/ (accessed January 2022).

62. Iversen V. B. Teletraffic Engineering and Network Planning / V. B. Iversen // Technical University of Denmark. — May 2010. — 370 p.

63. Jean Thierry Stephen Avocanh. An enhanced two level scheduler to increase multimedia services performance in LTE networks / Jean Thierry Stephen Avocanh, Marwen Abdennebi, Jalel Ben-Othman // IEEE International Conference on Communications (ICC). —2014. — P. 2351-2356.

64. Jean-Thierry Stephen Avocanh. Resources allocation in high mobility scenarios of LTE networks / Jean-Thierry Stephen Avocanh // Networking and Internet Architecture [cs.NI]. Université Sorbonne Paris Cité. — 2015. — 164 p.

65. Jiang X. Fast coding unit size decision based on probabilistic graphical model in high efficiency video coding inter prediction / X. Jiang, T. Song, W. Shi, T. Katayama, T. Shimamoto, L. Wang // IEICE Transactions on Information and Systems. — Nov. 2016. — Vol. 99. — No. 11. —P. 2836-2839.

66. Jiang X. Low-complexity and hardware-friendly H. 265/HEVC encoder for vehicular ad-hoc networks / X. Jiang, J. Feng, T. Song, T. Katayama // Sensors. — Apr. 2019. — Vol. 19. — No. 8. — 1927 p.

67. Jiang X. Quality oriented perceptual HEVC based on the spatiotemporal saliency detection model / X. Jiang, T. Song, D. Zhu, T. Katayama, L. Wang // Entropy. — Feb. 2019. — Vol. 21. — No. 2. — 165 p.

68. Kalva H. The VC-1 video coding standard / H. Kalva, J. Lee // IEEE MultiMedia. — Oct. 2007. — Vol. 14. —No. 4. —P. 88-91.

69. Kelly F.P. Reversibility and stochastic networks – New York: Willy. — 1979. — 238 p.

70. Kusume K., [et al.]. Deliverable D1.5. Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations / K. Kusume, [et al.] // ICT-317669 METIS Project, Public Deliverable ICT-317669-METIS/D1.5. — April. 2015. — 57 p.

71. Li S. A Survey of Energy-Efficient Communication Protocols with QoS Guarantees in Wireless Multimedia Sensor Networks / S. Li, J.G. Kim, D.H. Han, K.S/ Lee // Sensors. — 2019. — Vol. 19. — 199 p.

72. Li Y. Software-defined network function virtualization: A survey / Y. Li, M. Chen // IEEE Access. — 2015. — Vol. 3. — P. 2542-2553.

73. Liu J. Concert: A cloud based architecture for next-generation cellular systems / J. Liu, T. Zhao, S. Zhou, Y. Cheng, Z. Niu // IEEE Wireless Commun. — Dec. 2014. — Vol. 21. — No. 6. — P. 14-22.

74. Lyon D. Surveillance, snowden, and big data: capacities, consequences, critique / D. Lyon // Big Data & Society. — 2014. — Vol. 1. — No 2. — P. 1-13.

75. M. Grube. Applications of MPEG-4: Digital multimedia broadcasting / M. Grube, P. Siepen, C. Mittendorf, M. Boltz, M. Srinivasan // IEEE Transactions on Consumer Electronics. — 2001. —Vol. 47. —№.3. — P. 474-484.

76. Mocnej J. Network Traffic Characteristics of the IoT Application Use Cases / J. Mocnej, A. Pekar, W. K.G. Seah, I. Zolotova // School of Engineering and Computer Science, Victoria University of Wellington. — 2018. — 20 p.

77. Mukherjee D. A technical overview of VP9 -the latest opensource video codec / D. Mukherjee, J. Han, J. Bankoski, R. Bultje, A. Grange, J. Koleszar, P. Wilkins, Y. Xu // SMPTE Motion Imaging Journal. — Jan. 2015. — Vol. 124. — No. 1. — P. 44-54.

78. Muzata. A. R. The Modeling of Elastic Traffic Transmisson by the Mobile Network with NB-IoT Functionality. / A. R. Muzata, V. A. Pershina, M. S. Stepanov, F. Ndimumahoro, J. Ndayikunda. // 2021 Systems of Signals Generating and Processing in the Field of on Board Communications.— 2021. — P. 1-7.

79. Mwakwata C.B., [et al.]. Narrowband Internet of Things: From Physical and Media Access Control Layers Perspectives / C.B. Mwakwata [et al.] // Sensors. — 2019. — Vol. 19. — № 11. — 34 p.

80. Nguyen V.G. SDN/NFV based mobile packet core network architectures: A survey / V.G. Nguyen, A. Brunstrom, K.J. Grinnemo, J. Taheri // IEEE Commun. Surveys Tuts.— 2017. — Vol. 19. — No. 3. — P. 1567-1602.

81. Nokia. LTE evolution for IoT connectivity, Tech. Rep. White paper. Available at: https://halberdbastion.com/sites/default/files/201706/Nokia_LTE_Evolution_for_IoT_Connectivity_White_Paper.pdf (accessed January 2022).

82. Olshannikova E. Visualizing Big Data with augmented and virtual reality: challenges and research agenda / E. Olshannikova, A. Ometov, Y. Koucheryavy, T. Olsson // Journal of Big Data. — 2015. —Vol. 2. — No.1. — P. 1-27.

83. Optimus-cctv.ru: Системы безопасности и видеонаблюдения [Электронный ресурс]. Режим доступа: https://optimus-cctv.ru / (Дата обращения январь 2022).

84. Qiu M. Phase-change memory optimization for green cloud with genetic algorithm / M. Qiu, Z. Ming, J. Li, K. Gai, and Z. Zong // IEEE Trans. Comput. — 2015. — Vol. 64. — No. 12. — P. 3528-3540.

85. Rao K. R. VP6 Video Coding Standard / K. R. Rao, D. N. Kim, and J. J. Hwang // Dordrecht: Springer Netherlands. — Oct. 2014. — P. 159-197.

86. Ratasuk R. Overview of narrowband IoT in LTE Rel-13 / R. Ratasuk, N. Mangalvedhe, Y. Zhang, M. Robert, J. Koskinen // 2016 IEEE Conference on Standards for Communications and Networking (CSCN). — 2016. — P. 1-7.

87. Rathi Sonia. Throughput for TDD and FDD 4 G LTE Systems / Sonia Rathi, Nisha Malik, Nidhi Chahal, Sukhvinder Malik // International Journal of Innovative Technology and Exploring Engineering (IJITEE). — May 2014. — Vol. 3. — P. 73-77.

88. Reolink-russia.ru: IP Камеры Reolink [Электронный ресурс]. Режим доступа: https://reolink-russia.ru/ (Дата обращения январь 2022).

89. Richart M. Resource Slicing in Virtual Wireless Networks: A Survey / M. Richart, J. Baliosian, J. Serrat and J. Gorricho // in IEEE Transactions on Network and Service Management. — Sept. 2016. — Vol. 13. — No. 3. — P. 462-476.

90. Ross K.W. Multiservice loss models for broadband telecommunication networks / K.W. Ross // London, Berlin, New York: Springer– Verlag. — 1995. — 343 p.

91. Rost P., [et al.]. Network slicing to enable scalability and flexibility in 5G mobile networks / P. Rost, [et al.] // IEEE Commun. Mag. — May 2017. — Vol. 55. — No. 5. — P. 72-79.

92. Salman L. Energy efficient IoT-based smart home / L. Salman, S. Salman, S. Jahangirian, M. Abraham, F. German, C. Blair, P .Krenz // In Proceedings of the IEEE 3rd World Forum on Internet of Things (WF-IoT), Reston, VA, USA. — 2016. — Vol. 1. — P. 526-529.

93. Sani Y. Adaptive bitrate selection: A survey / Y. Sani, A. Mauthe, C. Edwards // IEEE Communications Surveys Tutorials. — 2017. — Vol. 19. — No. 4. — P. 2985-3014.

94. Schwarz H. Overview of the scalable video coding extension of the H.264/AVC standard / H. Schwarz, D. Marpe, and T. Wiegand // IEEE Transactions on Circuits and Systems for Video Technology. — Sep. 2007. — Vol. 17. — No. 9. — P. 1103-1120.

95. Shelby Z. The constrained application protocol (CoAP). RFC 7252/ Z. Shelby, K. Hartke, and C. Bormann // RFC Editor. — June 2014.

96. Sodagar I. The MPEG-DASH standard for multimedia streaming over the internet / I. Sodagar // IEEE MultiMedia. — Apr. 2011. — Vol. 18. — No. 4. — P. 62-67.

97. Stanford-Clark A. MQTT for Sensor Networks (MQTT-SN) Protocol Specification Version 1.2 / A. Stanford-Clark, H. Linh Troung // Mqtt.Org. — 2013.

98. Stasiak M. Modeling and Dimensioning of mobile networks from GSM to LTE / M. Stasiak, M. Głabowski, A. Wisniewski, P. Zwierzykowski // John Wiley & Sons Ltd. — 2011. — 136 p.

99. Stepanov S. Resource Allocation and Sharing for Transmission of Batched NB-IoT Traffic over 3GPP LTE / S. tepanov, M. Stepanov, A. Tsogbadrakh, J. Ndayikunda, U. Andrabi // Conference of Open Innovation Association, FRUCT. — 2019. — P. 422-429.

100. Stepanov S. N. Reservation Based Joint Servicing of Real Time and Batched Traffic in Inter Satellite Link / S. N. Stepanov, U. M. Andrabi, M. S. Stepanov, J. Ndayikunda // 2020 Systems of Signals Generating and Processing in the Field of on Board Communications, Moscow, Russia. — 2020. — P. 1-5.

101. Stepanov S.N. The Analysis of Resource Sharing for Heterogenous Traffic Streams over 3GPP LTE with NB-IoT Functionality / S.N. Stepanov, M.S. Stepanov, U. Andrabi, J. Ndayikunda. // Distributed Computer and Communication Networks. DCCN 2020. Lecture Notes in Computer Science. — 2020. —Vol 12563. — P. 422-435.

102. Stepanov S.N. The construction and analysis of generalized model of resource sharing for LTE technology with functionality of NB-IoT / S. N. Stepanov, M.S. Stepanov, E.E. Malikova, A. Tsogbadrakh, Ju. Ndayikunda // T-Comm. — 2018. — Vol. 12. — No.12. — P. 71-77.

103. Sullivan G. J. Overview of the high efficiency video coding (HEVC) standard / G. J. Sullivan, J. Ohm, W. Han, T. Wiegand // IEEE Transactions on Circuits and Systems for Video Technology. — Dec. 2012. — Vol. 22. — No. 12. — P. 1649-1668.

104. Sultana T. Choice of Application Layer Protocols for Next Generation Video Surveillance Using Internet of Video Things / T. Sultana, K.A. Wahid // IEEE Access. — 2019. — Vol. 7. — P. 41607-41624.

105. Surenda M. Gupta Queueing Model with State Dependent Balking and Reneging: Its Complementary and Equivalence / M. Surenda // ACM SIGMETRICS Performance Evaluation Review. — 1995. — Vol. 22. — No. 2-4. — P. 63-72.

106. TM Forum showcases telco collaboration with key verticals | Industry Trends | IBC. Available at: https://www.ibc.org/trends/tm-forum-showcases-telco-collaboration-with-key-erticals/3861.article (accessed February 2022).

107. Tong W. 5G: A thechnology vision / W. Tong, Z. Peiying // Huawei Technologies Co., Tech. Rep., 2013.

108. Walrand J., Varaiya P. High Performance Communications Networks (2nd ed) / J. Walrand, P. Varaiya // Morgan Kaufmann. — 2000.

109. Xin L., [et al.]. Network Slicing for 5G: Challenges and Opportunities / L. Xin, [et al.] // IEEE Internet Computing. — 2017. — T. 21. — P. 20-27.

110. Xu T. Non-Orthogonal Narrowband Internet of Things: A Design for Saving Bandwidth and Doubling the Number of Connected Devices / T. Xu, I .Darwazeh // IEEE Internet Things. — 2018. — Vol.5. — P. 2120-2129.

111. Yin X. Toward a principled framework to design dynamic adaptive streaming algorithms over HTTP / X. Yin, V. Sekar, B. Sinopoli // In Proc. the 13th ACM Workshop on Hot Topics in Networks, Los Angeles, USA. — Oct. 2014. — P. 1-7.

Приложение

Акт об использовании результатов диссертационной работы в учебном процессе МТУСИ

МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ, СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ РОССИЙСКОЙ ФЕДЕРАЦИИ

Ордена Трудового Красного Знамени федеральное государственное бюджетное образовательное учреждение высшего образования

«МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ» (МТУСИ)

MINISTRY OF DIGITAL DEVELOPMENT, COMMUNICATIONS AND MASS MEDIA OF THE RUSSIAN FEDERATION

MOSCOW TECHNICAL UNIVERSITY OF COMMUNICATIONS AND INFORMATICS (MTUCI)

ул. Авиамоторная, д. 8а, Москва, 111024,
www.mtuci.ru; мтуси.рф; e-mail: kanc@mtuci.ru
Телефон (495) 957-77-31; факс (495) 957-77-36
ОГРН 1027700117191; ИНН/КПП 7722000820/772201001; ОКПО 01179952;
ОКВЭД 85.22, 46.19, 58.19, 61.10, 68.32, 72.19, 85.21, 85.23, 85.42.9, 71.20, 33.13, 26.60 ; ОКТМО 45388000

_____ 20___ № _____

На №_____ от _____

УТВЕРЖДАЮ
Проректор МТУСИ по учебной работе
_____ Титов Е.В.
11.01.2022.

АКТ
об использовании результатов диссертационной работы
Ндайикунда Жувена
«Разработка и анализ модели динамического распределения ресурса беспроводных узлов доступа при передаче неоднородного трафика IoT»,
представленной на соискание ученой степени кандидата технических наук, в учебном процессе по кафедре «Сети связи и системы коммутации» МТУСИ

Комиссия в составе: председатель – зам. зав. кафедрой СС и СК, доцент Маликова Е.Е. члены доцент Степанова И.В. и доцент Данилов А.Н. составили настоящий акт в том, что результаты диссертационной работы:

- математическая модель обслуживания неоднородного трафика в беспроводном узле доступа, учитывающая наличие приоритета у трафика реального времени; использование дисциплины Processor Sharing при передачи эластичного трафика; ограничение по доступу, зависящее от общего уровня занятости ресурса;
- методика анализа эффективности процедур дифференцированного обслуживания неоднородного трафика, основанная на ограничении доступа, зависящего от общего уровня занятости ресурса;
- программы расчета характеристик пропускной способности беспроводного узла доступа

использованы в учебно-исследовательской работе бакалавров и магистров кафедры СС и СК, при курсовом проектировании по дисциплине «Основы Интернета Вещей».

Материалы диссертационной работы использованы при подготовке учебного пособия для выполнения практических работ по дисциплине «Основы Интернета Вещей».

Председатель комиссии: Маликова Е.Е.

Члены комиссии: Степанова И.В.
Данилов А.Н.