ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ»

На правах рукописи

Рыкова Татьяна Владимировна

АНАЛИЗ ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ РАСПРЕДЕЛЕНИЯ РЕСУРСОВ В МОБИЛЬНЫХ СЕТЯХ С ПОМОЩЬЮ ДВУХФАЗНЫХ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ

Специальность 05.13.17 – теоретические основы информатики

Диссертация
на соискание ученой степени кандидата
физико-математических наук

Научный руководитель
доктор технических наук,
профессор Самуйлов Константин Евгеньевич

Москва – 2021

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. Рост количества пользователей систем мобильных сетей и их повышенные потребности в ресурсах беспроводной сети приводит к стремительному развитию телекоммуникационных технологий. 5G — поколение мобильных сетей, работающих в соответствии со стандартами телекоммуникаций, следующих за технологией LTE (Long-Term Evolution), предлагающих еще более высокие скорости передачи за счет использования более широкого спектра частот и повышения спектральной эффективности, а также снижение задержки передачи данных. Одним из методов улучшения спектральной эффективности является переход к гетерогенной сети, суть которой заключается в наличии в сети наряду с традиционными базовыми станциями (БС) ретрансляционных станций (РС). Использование РС, ретранслирующих передаваемую от БС к пользователю информацию, позволяет снизить расходы на развертывание сети, исключая необходимость обеспечения проводного доступа и ускоряя процесс построения сети [20,71,73,84]. При этом, учитывая ограниченность радио ресурсов в сети, возникновение ошибок в результате интерференции, высокие требования пользователей к предоставляемым услугам, задача эффективного распределения ресурсов между БС и РС относится к первостепенным и широко исследуется [82,107,112,113,123,157]. В литературе, однако, отсутствуют модели, рассматривающие одновременно функционирование БС и РС в соте в условиях совместно используемых ими частотно-временных ресурсов.

К повышению пропускной способности в сетях с динамически изменяющимся каналом также относится применение принципов межуровневой оптимизации (CLA, Cross-Layer Adaptation), которые позволяют за счет протокольного взаимодействия обеспечить оптимизированную передачу данных [10,76,106,114,125,127,135,136,148,149,151,152,153,159,160,161,166]. Наиболее известное применение принципов межуровневой оптимизации

связано с появлением технологии адаптивной модуляции и кодирования (AMC, Adaptive Modulation and Coding), которая позволила значительно повысить пропускную способность в сетях с динамически изменяющимся каналом [117]. В частности, вопрос адаптации видео контента в зависимости от состояния качества канала и других факторов является одним из важных направлений исследований для CLA. В [135] был предложен адаптивный алгоритм на клиентской стороне, нацеленный на минимизацию повторной буферизации пакетов при передаче видео в сети LTE. В [153] данные о качестве и скорости видео кодирования добавлены в пересылаемую пользователям сигнальную информацию для улучшения восприятия видео пользователями. В [127] проанализирован сценарий видео передачи на основе протокола передачи гипертекста (HTTP, HyperText Transfer Protocol), при котором скорость кодирования пересылаемого видео выбирается на основе полученных оценок пропускных способностей всех пользователей в сети. Данная задача приводит к необходимости исследования системы массового обслуживания (СМО) с поступающим потоком и длительностью обслуживания, которые могут изменяться во времени в зависимости от состояния канала в соте. В настоящее время известно ограниченное число работ, рассматривающих аспекты реализации подобной задачи [160, 161,166], а математические модели практически отсутствуют, что делает актуальной разработку модели функционирования одновременно БС и пользовательской станции (UE, User Equipment), в условиях изменяющихся во времени нагрузочных параметров, зависящих от состояния канала передачи от БС к UE.

Несмотря на высокий потенциал сети мобильной связи пятого поколения, согласно последнему релизу стандарта 5G, до сих пор существует ряд нерешенных задач по обеспечению требуемой низкой задержки, что подчеркивает релевантность поиска новых протокольных рещений на пути к услугам сверхнадежных коммуникаций с низкой задержкой (URLLC, Ultra Reliable and Low Latency Communications).

Существует несколько стратегий по уменьшению задержки для данных услуг, например, за счет усовершенствования механизма обратной связи для гибридного автоматического запроса на повторение (HARQ, Hybrid Automatic Repeat reQuest) [69,70,99,162,163]. Главным его недостатком является ограничение RTT (Round Trip Time) – временного интервала между моментами отправки пакета и получением обратной связи на передатчике. На данный момент предложено множество схем, направленных на снижение временного интервала RTT [125,128]. В [125] уменьшение данного временного интервала достигается за счет сокращения длины передачи, равной одному символу ортогонального частотного мультиплексирования (OFDM, Orthogonal Frequency Division Multiplexing), что приводит к более высоким требованиям к полосе пропускания, мгновенной обработки приемником и к ограничению мощности передатчика. Другой подход, обсуждаемый в 3GPP (3rd Generation Partnership Project), включает в себя автоматическую передачу избыточных версий данных для достижения необходимого уровня надежности до тех, пор пока передатчик не получит первое подтверждение об успехе (ACK, ACKnowledgement) [126]. Значительные исследовательские усилия были направлены на создание схем предсказания мгновенного результата декодирования, также называемых схемами раннего предсказания на базе HARQ (early HARQ, e-HARQ), с применением искусственного интеллекта. Следует отметить, что большинство известных предикторов используют заданное пороговое значение в алгоритмах как механизм классификации между сообщениями ACK и NACK (NACK, Non-ACKnowledgement). Следовательно, выбор порогового значения является критически важным для повышения эффективности схем предсказания. При этом в литературе отсутствуют аналитические модели, учитывающие механизм ранней адаптации канала в сети 5G с помощью параллельного моделирования механизма предсказания и традиционного механизма обратной передачи без

предсказания с целью повышения эффективности за счет поиска оптимального порогового значения.

Цифровая, дискретная природа современных пакетных технологий и частотно-временных ресурсов передачи в мобильной сети приводит к необходимости исследования рассмотренных выше задач в дискретном времени. Развитие методов анализа СМО ограниченной емкости и в дискретном времени, которые позволяли бы учитывать как дискретный характер передаваемых данных, так и существенно дискретный характер функционирования реальных сетей, является актуальным. Изучению СМО в дискретном времени посвящено значительное число работ (Башарин Г.П., Бочаров П.П., Ефимушкин В.А., Разумчик Р.В., Bruneel H., Kobayashi H., Takagi H., Wu D., и др., см. [4,5,8,11,24,122,160] и литературу в них). Следует отметить при этом, что работы, посвященные исследованию СМО сложной структуры в дискретном времени, позволяющих провести анализ различных алгоритмов распределения частотно-временных ресурсов в рамках конкретных решений для мобильных сетей, практически отсутствуют.

Анализ источников, рекомендаций и стандартов международных организаций, таких как 3GPP, IEEE (Institute of Electrical and Electronics Engineers), ETSI (European Telecommunications Standards Institute), позволил установить, что необходимы комплексные модели, которые адекватно описывали бы особенности алгоритмов управления доступом и распределением ресурсов сетей 5G.

В диссертации исследуются показатели эффективности распределения ресурсов. Традиционно под эффективностью использования ресурса понимается либо доля используемого ресурса, например, доля постоянно занятых каналов от общего числа каналов в системе передачи, либо доля времени, в течение которого ресурс успешно используется для обслуживания запросов на этот ресурс. Например, среда передачи в локальной сети при применении протоколов случайного множественного доступа может находиться в состоянии простоя,

успешной передачи от одного источника, конфликтной передачи от нескольких источников и периода разрешения конфликта. Таким образом, доля времени, в течение которого среда занята успешной передачей, есть эффективность ее использования.

Выше и далее по тексту под показателями эффективности распределения ресурсов подразумевается набор показателей, характеризующих исследуемый алгоритм распределения ресурсов и позволяющих провести сравнение с другими алгоритмами. Этот набор может состоять из одного показателя, например, наиболее часто используемого – вероятности потерь поступающих в соту мобильной сети пакетов или вызовов, либо нескольких, предполагающих оценку эффективности алгоритма на качественном уровне или рассмотрение их в качестве критериев в оптимизационных задачах при выборе того или иного алгоритма распределения ресурсов (см. [13,15,16,31,98,104,154,161,162] и литературу в них).

Особенностью настоящей работы является создание нового метода анализа показателей эффективности фрагментов мобильных сетей с помощью двухфазных моделей массового обслуживания в дискретном времени, позволяющего дать рекомендации по улучшению известных протоколов. Ввиду вышеупомянутого тема работы является актуальной.

Степень разработанности темы. Различным аспектам решения данной задачи посвящены работы российских и зарубежных исследователей. Исследование вопросов эффективного использования ресурсов, ширины полосы пропускания в телекоммуникационных системах и сетях нашло отражение в работах российских ученых (Башарин Г.П., Вишневский В.М., Гайдамака Ю.В., Гнеденко Б.В., Гольдштейн Б.С., Гудкова И.А., Ефимушкин В.А., Зейфман А.И., Кучерявый А.Е., Кучерявый Е.А., Назаров А.А., Наумов В.А., Нейман В.И., Орлов Ю.Н., Печинкин А.В., Пшеничников А.П., Ромашкова О.Н., Самуйлов К.Е., Севастьянов Б.А., Семенова О.В., Соколов Н.А., Степанов С.Н., Харкевич А.Д., Цитович И.И., Шнепс-

Шнеппе М.А., Шоргин С.Я., Яновский Г.Г. и др.) [1,2,3,7,9,12-14,17-25,40,44-49,51-54,59-62,64,65,155,156,164,165] и зарубежных авторов (Дудин В.Н., Bohge M., Capozzi F., Iversen B., Kelly F.P., Kleinrock L., Kobayashi H., Rappaport S., Shariat M., Wang L., Wu D. и др.), разрабатывавших математические модели и методы анализа, широко применяющиеся при расчетах и планировании сетей [39-42,76,80,85,101,105,110,115,116,121,148,157,160].

Обсуждению и анализу вопросов распределения ресурсов в мобильных сетях последующего поколения (NGMN, Next Generation Mobile Network) посвящено значительное число публикаций [1,12,13,17,76-82,97,112-114,122-124,129,136,145-147,157-161,166].

Новые услуги в современных сетях делают задачу управления доступом к сетевым ресурсам для обеспечения характеристик качества функционирования сети и предоставления услуг одной из наиболее актуальных. В ряде работ российских ученых изложен подход к анализу показателей качества обслуживания с помощью мультисервисных моделей теории телетрафика [3,23,48,52,54,62,63,72].

Актуальность проблемы распределения ресурсов возросла с переходом к сетям 4G, и затем к 5G, ориентированным на услуги со сложными моделями нагрузки [5,24,45,50,51,55,95,133,134]. В связи с этим возникают задачи исследования сетей со специальными механизмами управления потоком, изменяющимся в процессе функционирования системы. Исследованию математических моделей систем с такими потоками посвятили свои работы российские и зарубежные авторы: Башарин Г.П., Бочаров П.П., Лагутин В.С., Наумов В.А., Нейман В.И., Самуйлов К.Е., Степанов С.Н., Шоргин С.Я., Iversen V.B., Wang L., Wong C.Y., Wu D. и др. [1,9,45,48,53,110,120,157-159,160].

Цель и задачи исследований. В связи с изложенным, целью диссертационной работы является построение вероятностных моделей в виде двухфазных систем массового обслуживания сложной структуры в

дискретном времени для анализа показателей эффективности распределения ресурсов беспроводных сетей.

Для достижения поставленной цели в диссертационной работе решаются следующие задачи:

1. Построение и анализ модели двухфазной СМО в дискретном времени для решения задачи повышения пропускной способности соты беспроводной гетерогенной сети, позволяющей учитывать различные алгоритмы распределения ресурсов между фазами. Разработка пропорционального алгоритма распределения ресурсов с ограничениями.
2. Построение и анализ модели двухфазной СМО в дискретном времени для оценки показателей эффективности распределения ресурсов при решении задачи межуровневой оптимизации – повышения пропускной способности сети и качества восприятия видео потока на пользовательской станции.
3. Разработка модели двухфазной СМО в дискретном времени, позволяющей произвести оценку среднего времени пребывания заявки в системе и других показателей эффективности путем итеративного моделирования механизма предсказания повторной передачи и механизма обратной связи для решения задачи снижения задержки передачи пакета на пользовательскую станцию.

Структура и объем работы. Перейдем к общей характеристике результатов диссертации с продолжением обзора литературы. Диссертационная работа состоит из введения, трех глав, заключения, библиографии из 166 наименований на русском и английском языках. Результаты диссертационной работы изложены на 119 страницах. Текст работы иллюстрируется 35 рисунками и 4 таблицами.

Краткое изложение диссертации. В главе 1 рассматриваются вопросы распределения ресурсов передачи информации в сетях NGMN на примере сети LTE, ставятся задачи моделирования и оценки показателей

эффективности алгоритмов распределения ресурсов. Показывается актуальность разработки моделей функционирования соты NGMN в виде двухфазных СМО в дискретном времени. В разделе 1.2 предложена модель функционирования соты гетерогенной сети NGMN, состоящей из одной БС и нескольких РС, совместно разделяющих частотно-временные ресурсы соты, в виде двухфазной СМО в дискретном времени сложной структуры с буферным накопителем (БН) первой фазы (БС) конечной емкости, несколькими БН второй фазы (РС) также конечной емкости, разделяемым на каждом такте между БС и РС ограниченным множеством приборов и неординарным потоком заявок на первую фазу. Поскольку в гетерогенных сетях нерациональное распределение ресурсных блоков между БС и РС может приводить к значительным потерям, становится необходимым исследовать гибкие алгоритмы распределения ресурсов соты такой сети. Для СМО выведены система уравнений равновесия (СУР), выражения для основных вероятностно-временных характеристик (ВВХ), разработан программный комплекс аналитического моделирования, проведен численный анализ для четырех алгоритмов распределения ресурсов между БС и РС. Разработанная модель позволяет проводить сравнительный анализ показателей эффективности различных алгоритмов распределения ресурсов внутри соты гетерогенной сети NGMN. Предложен алгоритм пропорционального распределения приборов с ограничениями, показавший свою эффективность.

Основные результаты первой главы опубликованы в работах автора [26,32,33,38,93,94].

В главе 2 предлагается и исследуется двухфазная СМО в дискретном времени с управляемыми цепью Маркова ординарным геометрическим поступающим потоком и ординарным обслуживанием на фазе 1, и обслуживанием по геометрическому закону с опустошением на фазе 2.

Данная СМО может служить аналитической моделью процесса передачи видео потока по нисходящему каналу соты сети NGMN с учетом CLA, решающей задачу улучшения характеристик передачи информации

за счет межпротокольного взаимодействия. В СМО первая фаза моделирует процесс передачи видео с учетом распределения ресурсов в соте NGMN, и вторая фаза – процесс декодирования видео потока терминалом пользователя. Межуровневая оптимизация учитывается зависимостью от состояния индикатора качества канала, параметров поступающего потока заявок и вероятности обслуживания заявок на первой фазе СМО, что соответствует зависимости разделения частотно-временных ресурсов в соте от качества канала. В главе осуществляется постановка задачи и описание СМО, проводится декомпозиция системы, выводятся СУР для первой и второй фаз.

Показано, что решение для стационарного распределения вероятностей данной СМО вычисляется мультипликативно на основе полученных распределений для первой и второй фаз. Найдены ВВХ функционирования СМО и проводится их численный анализ. Предложенная модель может применяться для получения быстрой оценки эффективности распределения ресурсов для видео передачи в сети NGMN.

Основные результаты второй главы опубликованы в работах автора [28,31,34,88,89,90].

В главе 3 диссертационной работы исследуется механизм ранней адаптации канала на базе предсказания повторной передачи e-HARQ для нисходящего канала мобильной сети, решающей задачу улучшения характеристик передачи информации. Модель представляет собой двухфазную СМО, в которой на первой фазе моделируется процесс предсказания с возможностью ретрансляции в случае NACK, а на второй фазе рассматривается процесс обработки сообщения терминалом пользователя на базе HARQ. Первая и вторая фаза данной СМО в дискретном времени характеризуются геометрическим поступающим потоком и геометрическим процессом обслуживания. В разделах 3.1 и 3.2 осуществляется постановка задачи и описание модели, в то время как в разделе 3.3 выводится СУР и ВВХ функционирования СМО. В разделе 3.4

проводится численный анализ полученных ВВХ. На первом этапе были найдены вероятности переходов за счет моделирования канального уровня сети 5G. Получение реалистичных значений для вероятностей переходов позволяет использовать данную аналитическую модель для анализа и оптимизации существующих схем предсказания e-HARQ. Далее был разработан имитационный комплекс для проверки корректности аналитической модели, и решена оптимизационная задача поиска оптимальных параметров: длин сообщений, качества предсказания, а также пороговых значений, при которых наблюдается наименьшая длительность успешного обслуживания заявки при условии, что ошибка предсказания не превышает допустимых значений.

Основные результаты третьей главы опубликованы в работах автора [36-38,57,58, 103, 138,139].

В заключении сформулированы основные результаты диссертации.

Для проведения численных экспериментов и анализа функционирования предложенных моделей в диссертации был разработан комплекс программных средств на языке C++, MATLAB, PYTHON.

Положения, выносимые на защиту.

1. Для решения задачи повышения пропускной способности беспроводной гетерогенной сети с ретрансляторами данных применима предложенная в диссертации модель двухфазной СМО в дискретном времени с распределением ресурсов приборов между фазами. Для расчета показателей эффективности СМО предложен алгоритм пропорционального распределения фиксированного числа приборов с ограничениями между буферными накопителями первой и второй фазы.
2. Для анализа среднего времени пребывания заявки в системе и вероятности потери заявки при решении задачи межуровневой оптимизации – повышения пропускной способности сети и качества восприятия видео потока – применима предложенная в диссертации

модель двухфазной СМО в дискретном времени, и получены формулы для расчета характеристик и стационарного распределения.

3. Для решения задачи снижения задержки передачи пакета на пользовательскую станцию мобильной сети применима предложенная модель двухфазной СМО в дискретном времени, моделирующая механизм предсказания повторной передачи и механизм обратной связи, формализована задача оптимизации, результаты решения которой могут быть использованы как исходные данные в алгоритме предсказания повторной передачи.

Научная новизна диссертационной работы.

1. Для решения задачи повышения пропускной способности соты беспроводной гетерогенной сети предложена модель многопотоковой двухфазной СМО в дискретном времени с групповым поступающим потоком и второй фазой сложной структуры, состоящей из параллельных СМО конечной емкости. Отличительной особенностью модели является распределение множества приборов между системами первой и второй фаз.

2. Для решения задачи межуровневой оптимизации – повышения пропускной способности сети и качества восприятия видео потока на пользовательской станции – предложена модель двухфазной СМО в дискретном времени, в которой в отличие от известных входящий поток и обслуживание на фазе 1 управляются цепью Маркова, на фазе 2 применяется обслуживание «с опустошением».

3. Для решения задачи снижения задержки передачи пакета на пользовательскую станцию предложена модель двухфазной СМО в дискретном времени, которая в отличие от известных моделей учитывает итеративное моделирование механизма предсказания повторной передачи и механизма обратной связи.

Методы исследования. В диссертации применяются методы теории массового обслуживания, теории вероятностей, теории марковских

случайных процессов, математической теории телетрафика, теории матриц и имитационного моделирования.

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

Разработанные математические модели могут быть использованы профильными подразделениями университетов и институтов высшего образования в учебной деятельности, научно-исследовательскими и проектными институтами в практических разработках при расчете и планировании мобильных сетей для получения быстрой оценки показателей эффективности алгоритмов распределения ресурсов в беспроводных гетерогенных сетях без больших затрат времени на имитационное моделирование сети и без применения дорогих экспериментальных установок.

Реализация результатов диссертации. Полученные при подготовке диссертации результаты использовались при выполнении работ по гранту 19-07-00933 А – «Стохастические модели и задачи оптимизации для разработки информационных технологий виртуализации и управления ресурсами в беспроводных мультисервисных сетях» и внедрены в учебный процесс – в научно-образовательные курсы «Модели для анализа качества сетей подвижной связи» и «Анализ производительности сетей сотовой подвижной связи» для студентов бакалавриата направлений подготовки 02.03.02 «Фундаментальная информатика и информационные технологии в РУДН».

Обоснованность и достоверность результатов. Обоснованность результатов подтверждается адекватностью выбранных методов цели и задачам исследования, актуальностью и репрезентативностью источников, используемых в работе. О достоверности результатов диссертации свидетельствует сравнительный анализ расчетов для

построенных моделей технических систем с соответствующими вычислительными экспериментами, проведенными на базе близких к реальным исходных данных.

Апробация результатов диссертации. Результаты работы докладывались и обсуждались на следующих научных конференциях и семинарах: VII и VIII международные конференции «Finnish-Russian University Cooperation in Telecommunications (FRUCT)» (Санкт-Петербург, 2010 г., Лаппеенранта, 2010 г.); международная конференция «Consumer Communications and Networking Conference» IEEE CCNC (Лас-Вегас, США, 2011 г.); V всероссийская конференция (с международным участием) «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем» ИТММ (Москва, 2011 г.); XIV и XVII международные конференции «Distributed Computer and Communication Networks (DCCN)» (Москва, 2011 г., 2013 г.); XII Всероссийское совещание по проблемам управления (ВСПУ), (Москва, 2014 г.); научный межвузовский семинар «Современные телекоммуникации и математическая теория телетрафика» (Москва, 2014 г.); XXXII International Seminar on Stability Problems for Stochastic Models (Трондхейм, Норвегия, 2014 г.); XX International Conference on Next Generation Wired/Wireless Advanced Networks and Systems (NEW2AN, Санкт-Петербург, 2020 г.); семинар кафедры прикладной информатики и теории вероятностей Российского университета дружбы народов (Москва, 2021 г.).

Результаты главы 2 диссертационной работы, представленные автором в заявке «Комплекс моделей, методик и программных средств оптимизации ресурсов в сетях LTE» на Конкурс инноваций и инновационных проектов 2013/2014, проводившийся Международной академией связи при поддержке Московского технического университета связи и информатики и Общественного совета при Федеральном агентстве связи получил диплом лауреата первой степени в номинации «Конкурс концептуальных идей, методик, рекомендаций».

Соответствие паспорту специальности. Диссертационное исследование выполнено в соответствии с паспортом специальности 05.13.17 «Теоретические основы информатики» и включает оригинальные результаты в области исследования информационных процессов и требований их пользователей к показателям эффективности, в области разработки моделей информационных процессов в мобильных сетях, разработки общих принципов организации телекоммуникационных систем и оценки их эффективности. Таким образом, диссертационное исследование соответствует следующим разделам паспорта специальности 05.13.17 «Теоретические основы информатики»:

1. Исследование, в том числе с помощью средств вычислительной техники, информационных процессов, информационных потребностей коллективных и индивидуальных пользователей.
2. Исследования и разработка требований к программно-техническим средствам современных телекоммуникационных систем на базе вычислительной техники.
3. Общие принципы организации телекоммуникационных систем и оценки их эффективности.

Личный вклад. Предлагаемые в диссертации модели, системы массового обслуживания разработаны и исследованы автором самостоятельно. Выносимые на защиту результаты в виде формул, математических процедур, алгоритмов, программных средств получены автором лично.

Публикации. По теме диссертации опубликовано 37 работ, из них 12 [10,16,26,29,30,32,33,34,36,90,137,142] – в рецензируемых научных журналах, рекомендованных ВАК Минобрнауки России, 21 [15,28,31,35, 37,38,57,58,87-89,91-96,103,138,139,143] – в рецензируемых трудах международных конференций, 1 [27] – в сетевом издании, 1 российский и 2 зарубежных патента [56,140,141].

ГЛАВА 1

МОДЕЛЬ ДВУХФАЗНОЙ СМО В ДИСКРЕТНОМ ВРЕМЕНИ С РАСПРЕДЕЛЯЕМЫМИ МЕЖДУ ФАЗАМИ ПРИБОРАМИ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ПОВЫШЕНИЯ ПРОПУСКНОЙ СПОСОБНОСТИ БЕСПРОВОДНОЙ ГЕТЕРОГЕННОЙ СЕТИ

1.1. Двухфазные модели для анализа показателей эффективности алгоритмов распределения ресурсов в беспроводных сетях

Стандарты беспроводных сетей последующих поколений NGMN [132], к которым относятся сети 4G и 5G характеризуются развитыми алгоритмами распределения частотно-временных ресурсов для передачи информации в соте между БС и UE. В данных сетях частотно-временные ресурсы формируются с использованием мультиплексирования с ортогональным частотным разделением каналов [109]. Вследствие многоэтапности процессов передачи в соте NGMN, наиболее приемлемыми для исследования сетей NGMN и алгоритмов распределения ресурсов в них являются многофазные СМО. Однофазные СМО, даже при использовании распределения длительности обслуживания фазового типа, позволяющего придать необходимый физический смысл этапам обслуживания, как это делается, например, при моделировании локальных сетей [6], в случае NGMN не дают возможности учесть сложную структуру сети с промежуточным хранением передаваемой информации.

Как уже указывалось анализу многофазных СМО посвящено большое число публикаций, рассматривающих различные варианты структурных параметров: емкости БН фаз, числа приборов на фазах, ординарного или неординарного входящего потока, блокировки обслуживания на фазе или потери заявки при полной занятости БН следующей фазы, возможности повторного обслуживания на фазе или в системе, распределений входящего потока и обслуживания заявок приборами фаз. В публикациях

число фаз обычно ограничивается двумя, и рассматриваются они в основном в непрерывном времени.

Следует отметить, что, насколько известно автору, в российских и зарубежных публикациях не рассматривались аналитические модели в виде многофазных СМО с распределяемым между фазами ограниченным множеством приборов [137]; данные модели позволяют исследовать распределение частотно-временных ресурсов в сети NGMN между БС и РС.

Исследованию многофазных СМО (двухфазных) в дискретном времени посвящены лишь несколько работ, см. например, [5,11,75,86,118], которые, однако, не позволяют учесть сложную структуру фаз при моделировании процессов передачи в соте и, в связи с этим, не полностью соответствуют решению задачи исследования эффективности распределения ресурсов в соте сети NGMN. Следует отметить, что большое число зарубежных публикаций, содержащих в названии термины «discrete» и «tandem queue», по сути, посвящены исследованию систем циклического обслуживания в дискретном времени, но не многофазных систем.

Количество фаз в многофазной СМО в рассматриваемой в качестве аналитической модели для оценки показателей эффективности распределения ресурсов в соте NGMN в большинстве случаев, по мнению автора, следует принимать равным двум, поскольку каждая фаза чаще всего сама по себе является структурно-сложной СМО с непростыми правилами функционирования, и дальнейшее увеличение числа фаз резко усложняет формализацию всей системы, приводит к многомерным процессам, описывающим ее поведение, затрудненному практическому использованию; анализ при этом становится крайне громоздким с большими рисками получения неточных результатов. Декомпозиция такой системы с анализом отдельных фаз или групп фаз чаще всего не применима вследствие существенного взаимовлияния фаз, в отличие от почти полностью разложимых систем [2,83], и может приводить в

большинстве случаев к существенным ошибкам моделирования. Случаи независимости функционирования фазы от предыдущей фазы и, соответственно, допустимой декомпозиции редки и возникают, при выполнении условий [52], например, при использовании экспоненциальных распределений и БН неограниченной емкости [111], либо в предположениях о специфических условиях функционирования фаз [28,90].

В связи с вышесказанным, возникающими новыми задачами исследования алгоритмов распределения ресурсов в NGMN в диссертации далее предлагаются и исследуются двухфазные системы конечной емкости в дискретном времени специальной структуры для анализа показателей эффективности распределения ресурсов в беспроводных сетях.

1.2 Построение модели беспроводной гетерогенной сети для решения задачи повышения пропускной способности

Рассмотрим функционирование соты гетерогенной сети NGMN, содержащей одну БС и K РС, $K < \infty$, при централизованном распределении ресурсов, рис.1.1.

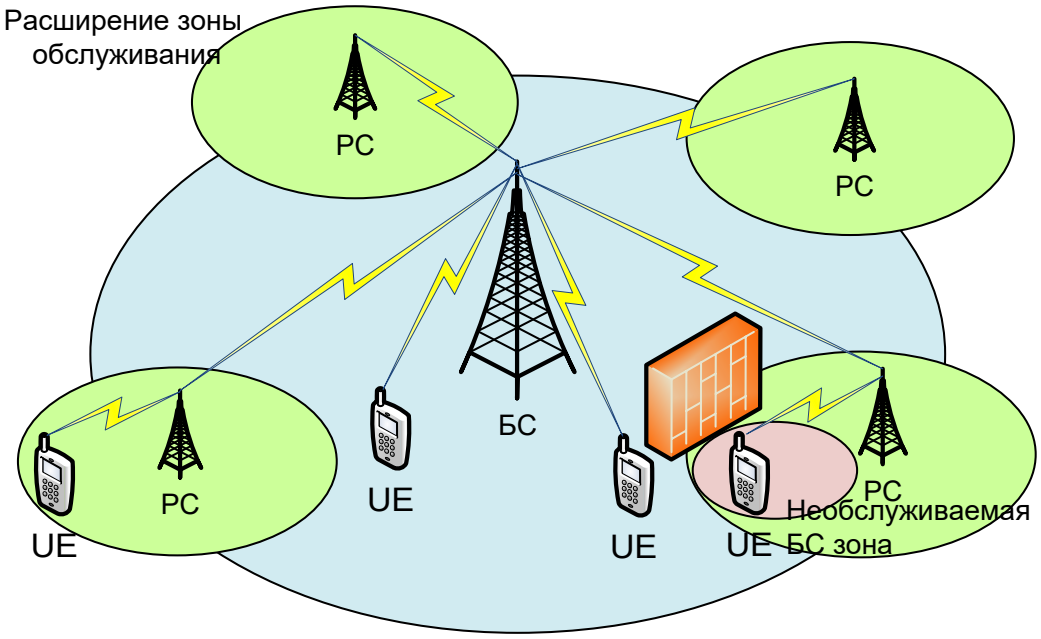


Рис.1.1. Гетерогенная сеть NGMN с одной и БС и несколькими РС

Кадр нисходящего канала, в течение которого происходят возможные передачи пакетов в соте, разбит на S ресурсных блоков (РБ). Все S РБ каждого кадра распределяются между БС и K РС для передачи пакетов затем в направлении UE. В предлагаемой далее в качестве модели СМО будем под заявкой рассматривать пакет, а в качестве прибора – один РБ. Таким образом, ресурсы одного кадра нисходящего канала соответствует S приборам в СМО.

Будем считать, что на систему поступают заявки $K+1$ типов. Заявка k-го типа (k-заявка) при $k=0$ должна быть передана на UE, находящуюся в зоне обслуживания БС, и при $k=1,2,\dots,K$ – на UE в зоне обслуживания k-й РС (далее – РС$_k$). Поступающие на фазу 1 (БС) новые заявки и на K СМО фазы 2 (РС) со стороны фазы 1 буферизуются в буферном накопителе (БН) фазы 1 или БН этих СМО фазы 2, соответственно. Будем полагать емкость БН БС (далее – БН$_0$) равной $r_0, r_0 < \infty$, и емкость БН РС$_k$ (далее – БН$_k$) равной $r_k, r_k < \infty,\ k=1,2,\dots,K$. Будем считать, что поступившие заявки, которым не хватило мест для буферизации, теряются, не возобновляются и не оказывают влияния на дальнейшее функционирование системы. Наконец, заметим, что обслуживаемая заявка занимает одно место в БН.

Подсистему рассматриваемой СМО, образованную из приборов и буферного накопителя БН$_0$, относящихся к БС, будем обозначать СМО$_0$. При этом, группу приборов в СМО$_0$ в количестве s_0, выделенную для обслуживания k - заявок будем обозначать П$_k$, $k=0,\dots,K$. Аналогично, приборы в количестве s_k и БН$_k$, относящиеся к РС$_k$, обозначим СМО$_k$, а приборы в СМО$_k$ будем обозначать П$_{K+k}$, $k=1,2,\dots,K$. Структура предлагаемой СМО приведена на рис.1.2.

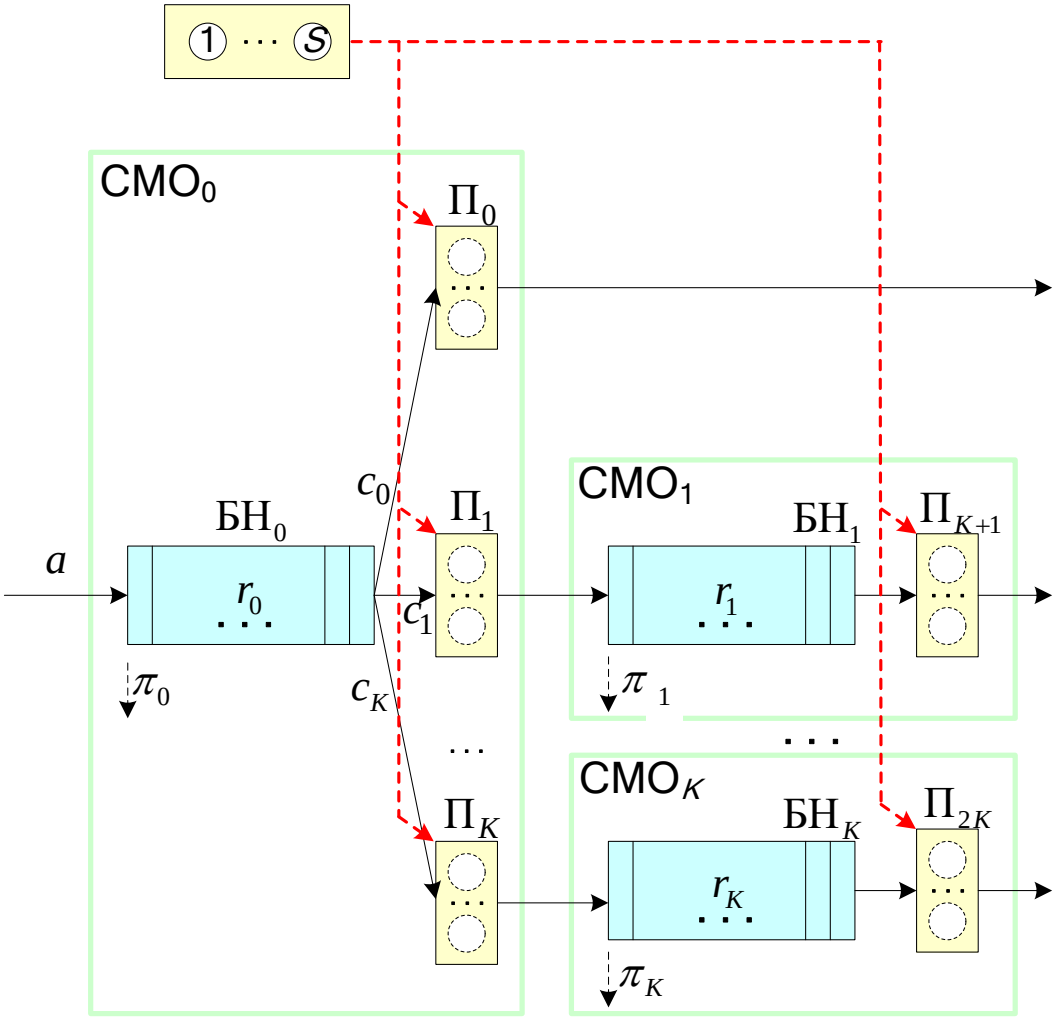


Рис.1.2. Двухфазная система с K СМО на фазе 2, S приборами, распределяемыми среди СМО$_k$, $k = 0, \dots, K$, () и K+1 типами заявок

Будем рассматривать функционирование СМО в дискретном времени с тактом постоянной длины h и следующей последовательностью событий в момент nh:

- окончание обслуживания заявок на приборах П$_{K+k}$ фазы 2 и освобождение мест, занимаемых этими заявками, в БН$_k$, $k = 1,2, \dots, K$;
- окончание обслуживания заявок на приборах П$_k$, $k = 0,1, \dots, K$, фазы 1;
- поступление заявок, ориентированных на СМО$_k$, с приборов СМО$_0$ на БН СМО$_k$, $k = 1,2, \dots, K$, и освобождение мест в БН СМО$_0$, занимавшихся обслуженными за такт заявками;
- поступление новых заявок на СМО$_0$;

- перераспределение S приборов между СМО$_k$, $k = 0,1, \dots, K$;
- фиксация состояния.

Для лучшего понимания функционирования СМО на рис.1.3 приведена последовательность указанных выше событий в привязке к событиям в соте гетерогенной сети.

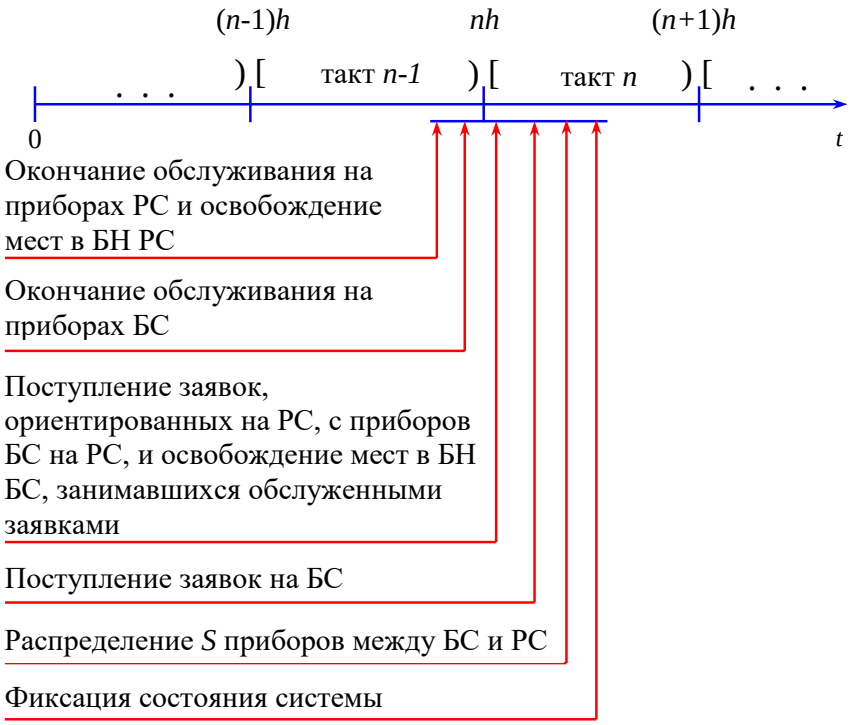


Рис.1.3. Диаграмма последовательности событий в двухфазной СМО в дискретном времени в привязке к событиям в соте гетерогенной NGMN

В данном случае перераспределение приборов между СМО фазы 1 и фазы 2 является завершающим активным событием, позволяющем учесть при обслуживании в следующем такте сформировавшиеся в БН фаз объемы заявок разных типов.

Пусть $\eta_n, \eta_n \in \{0,1\}$, число поступлений групп заявок за n-й такт, причем все $\eta_n, n \geq 0$, – независимые одинаково распределенные случайные величины с производящей функцией (ПФ)

$A(z) = Mz^{\eta_n} = 1 - a + az, |z| \leq 1, a = P\{\eta_n = 1\}, 0 < a < 1, n \geq 0.$

Число заявок χ_n в поступившей группе является независящей от n случайной величиной с ПФ

$G(z) = Mz^{\chi_n} = \sum_{i\geq 1} g_i z^i, |z| \leq 1, G(1) = 1, g_i = P\{\chi_n = i\}, n \geq 0.$

Таким образом, поступающий поток является геометрическим групповым Geom^G, поскольку период времени между поступлениями групп имеет геометрическое распределение со средним $\frac{1}{a}$, и характеризуется ПФ

$$A\big(G(z)\big) = 1 - a + aG(z) = \sum_{i\geq 1} a_i z^i, |z| \leq 1, a_0 = \bar{a} = 1 - a,$$

$a_i = ag_i, i \geq 1.$

В соответствии со сказанным выше, пусть каждая заявка из поступившей группы принадлежит типу k (k -заявка) с вероятностью $c_k, k = 0, \dots, K, c_\cdot = 1$. Точка вместо индекса означает полную сумму переменной по этому индексу. С физической точки зрения будем считать, что 0-заявка соответствует пакету, предназначенному для передачи напрямую UE данной БС, а k -заявка – для передачи в направлении РС_k, $k = 1,2, \dots, K$. Таким образом, поступающий на СМО поток является $(K+1)$-мерным групповым геометрическим.

Функционирование соты предполагает, что поступающие пакеты могут быть различной длины и требовать для передачи нескольких РБ. Соответственно, в СМО в общем случае обслуживание заявки должно быть описано случайным законом распределения, например, геометрическим. Однако в условиях существенно сложной структуры при неординарном неоднородном входящем потоке далее в целях упрощения рассматривается детерминированный закон обслуживания заявок с длительностью, равной одному такту, что эквивалентно времени передачи пакета, не превышающем длительности одного РБ.

Исходя из сделанных предположений о функционировании СМО, каждая заявка обслуживается в течение одного такта, после чего освобождает прибор и занимаемое во время обслуживания место в БН.

Для введенной СМО можно использовать, таким образом, мнемоническое обозначение, содержащей K параллельных СМО: $Geom_{K+1}^{G}|D=1|s_0|r_1<\infty \rightarrow (D=1|s_k|r_2<\infty)^K$.

В ряде работ (см. ссылки на источники, например, в [1,7]) показано, что в неоднородных СМО рассматриваемого типа можно не различать заявки в очереди (в нашем случае в $БН_0$) до начала обслуживания; розыгрыш же типа осуществляется по полиномиальной схеме с вероятностями $c_0, \dots, c_K$ в момент выбора заявки из очереди на обслуживание. Отметим, что данное положение действует для анализируемой соты NGMN в случаях детерминированных алгоритмов распределения РБ и для схем распределения, учитывающих лишь общее число пакетов в $БН_0$ и не принимающих во внимание число пакетов в $БН_0$ каждого типа.

Поведение СМО описывается однородной цепью Маркова (ЦМ) ξ_n по моментам $nh, n \geq 0$ над пространством состояний:

$X = \{\mathbf{x} = (x_0, x_1, \dots, x_K): x_K = 0, \dots, r_k, k = 0, \dots, K\}, |X| = \prod_{k=0}^{K}(r_k + 1)$,

где x_k – число k -заявок, находящихся в $БН_k$.

1.3. Алгоритмы распределения приборов

Пусть далее $\lfloor y \rfloor$ - означает округление y в сторону наименьшего целого.

Для учета распределения приборов между БС и РС в сети на шаге n введем вектор $\boldsymbol{s}^n := (s_0^n, s_1^n, \dots, s_{2K}^n)^T$, определяющий распределение приборов, значения которого зависят от состояния **x** системы на шаге n. Далее в выражении неравенства для векторов, будем, как обычно, полагать его поэлементное выполнение. Вектор $\boldsymbol{s}^n$ задается алгоритмом распределения приборов; предполагается $\boldsymbol{s}^n(\mathbf{x}) = \mathbf{s}(\mathbf{x}), n \geq 0$.

Далее в главе будем исследовать следующие четыре алгоритма, формализованные в терминах предложенной модели:

A1. Детерминированный алгоритм (не зависит от **x**):

$s_k = \left\lfloor \frac{S}{2K+1} \right\rfloor, k = 1, \dots, 2K, s_0 = S - \sum_{k=1}^{2K} s_k$.

A2. Детерминированный алгоритм (не зависит от **x**), исследуется в [17]:

$$s_{K+k} = \left\lfloor \frac{S - \left\lfloor \frac{S}{2} \right\rfloor}{K+1} \right\rfloor, k = 1, \dots, K, \tilde{S} = S - \sum_{k=1}^{K} s_{K+k}, s_k = \left\lfloor \frac{\tilde{S}}{K+1} \right\rfloor,$$

$k = 1, \dots, K, s_0 = \tilde{S} - \sum_{k=1}^{K} s_k$.

Отметим, что введенные выше алгоритмы **A1** и **A2** не учитывают изменения объемов трафика в гетерогенной сети и приводят к снижению ее пропускной способности. В **A1** в отличие от **A2** меньшее число РБ по сравнению с **A2** предоставляется БС, и соответственно РС$_k$ имеют больше РБ, чем при алгоритме **A2**.

A3. Пропорциональный алгоритм с приоритетом для СМО фазы 2 (зависит от **x**):

$s_{K+k} = \left\lfloor \frac{x_k S}{x_.} \right\rfloor, \tilde{S} = S - \sum_{k=1}^{K} s_{K+k}, s_k = \left\lfloor \frac{\lfloor x_0 c_i \rfloor \tilde{S}}{x_0} \right\rfloor, k = 1, \dots, K,.$

$s_0 = \tilde{S} - \sum_{k=1}^{K} s_k$.

A4. Пропорциональный алгоритм с ограничениями (зависит от **x**):

Алгоритм **A4**, предлагаемый далее, ориентирован в первую очередь на обслуживание заявок в СМО фазы 2, уже прошедших обслуживание на фазе 1 и, в некотором смысле, более ценных. В случае, когда суммарное число заявок в БН СМО фазы 2 превосходит либо равно общему числу приборов S, все имеющиеся приборы назначаются для обслуживания СМО$_k$, $k = 1, \dots, K$. При этом в процессе действия **A4** возможны две ситуации:

1) Суммарное число заявок фазы 2 превосходит S: $\sum_{i=1}^{K} x_i > S$. В этом случае выполняется пропорциональное распределение всех приборов между СМО фазы 2:

$$s_{K+k} = \left\lfloor \frac{x_k S}{\sum_{i=1}^{K} x_i} \right\rfloor \; k = 1, \dots, K.$$

2) Суммарное число заявок в СМО фазы 2 не превосходит S: $\sum_{i=1}^{K} x_i \leq S$. В данном случае необходимое количество приборов предоставляется для обслуживания всех заявок, находящихся в СМО фазы

2, и $s_{K+k} = x_k,\ \ k = 1, \dots, K$. Технически обеспечение этого может осуществляться следующим пересчетом: $s_{K+k} = min(x_k, \tilde{S}_k)$, $\tilde{S}_k = S - \sum_{i=0}^{k-1} s_{K+i}, s_K = 0, k = 1, \dots, K$.

Оставшиеся $\tilde{\tilde{S}} = S - \sum_{i=1}^{K} s_{K+i}$, приборов после распределения между СМО$_k$ предоставляются в СМО$_0$ для обслуживания заявок в БН$_0$. Итерационная процедура предоставления приборов продолжается до тех пор, пока все они не будут распределены между k -заявками, $k = 1, \dots, K$.

<u>Алгоритм.</u> *Итерационный алгоритм пропорционального распределения ресурсов с ограничениями между буферами первой и второй фазы для расчета показателей эффективности двухфазной СМО в дискретном времени состоит из следующих шагов:*

Шаг 1*. Входные параметры:*

$n = 0: \tilde{\tilde{S}}_{(n)} = \tilde{\tilde{S}}, c_k^{(n)} = c_k, s_k^{(n)} = 0, k = 0,1, \dots, K.$

Шаг 2*. Если $\tilde{\tilde{S}}_{(n)} > 0$ переходим к шагу 3. В противном случае, происходит выход из алгоритма.*

Шаг 3*.* $n = n + 1$:

$$s_k^{(n)} = min\left(\sum_{m=0}^{n-1} \left\lfloor \frac{\lfloor x_0 c_k^{(m)} \rfloor \tilde{\tilde{S}}_{(m)}}{x_0} \right\rfloor, r_k - (x_k - s_{K+k})\right) k = 1,2, \dots, K,$$

$$s_0^{(n)} = \sum_{m=0}^{n-1} \left\lfloor \frac{\lfloor x_0 c_0^{(m)} \rfloor \tilde{\tilde{S}}_{(m)}}{x_0} \right\rfloor,$$

$$\tilde{\tilde{S}}_{(n)} = \tilde{\tilde{S}}_{(n-1)} - \sum_{k=0}^{K} (s_k^{(n)} - s_k^{(n-1)}),$$

$$c_k^{(n)} = \frac{c_k^{(n-1)}(1 - \delta(s_k^{(n)}, r_k - (x_k - s_{K+k})))}{c_\cdot^{(n)}},$$

где

$$c_\cdot^{(n)} = \sum_{k=1}^{K} \left(c_k^{(n-1)} \left(1 - \delta\left(s_k^{(n)}, r_k - (x_k - s_{K+k})\right)\right)\right) + c_0^{(n-1)}.$$

Шаг 4*. Чтобы избежать зацикливания алгоритма необходимо проверить условие $\tilde{\tilde{S}}_{(n)} \neq \tilde{\tilde{S}}_{(n-1)}$; если оно выполняется, переходим к шагу 2. В противном случае, все оставшиеся приборы $\tilde{\tilde{S}}_{(n)}$ назначаются на*

обслуживание заявок 0-типа: $s_0^{(n)} = s_0^{(n)} + \tilde{\tilde{S}}_{(n)}$, *и происходит завершение алгоритма.*

Пропорциональные алгоритмы распределения приборов (**A3, A4**), зависящие от состояния системы, нацелены на повышение общей пропускной способности соты за счет адаптированного к нагрузке распределения приборов между БС и РС, поэтому они могут быть отнесены к классу планировщиков Proportional Fair [46]. Данный класс планировщиков характеризуется выделением РБ пользователям с наилучшим качеством канала в целях повышения пропускной способности сети, учитывая при этом среднее число переданных бит всех активных UE в предыдущих тактах, что позволяет также обеспечить минимальным числом РБ пользователей, находящихся в плохих канальных условиях. Особенностью алгоритма **A4** является выделение БС на обслуживание k-заявок группы РБ объема, не приводящего к потерям на буферных накопителях РС.

Таким образом, преимуществом алгоритма **A4** является возможность ограничения предоставления приборов фазе 1 для обслуживания k-заявок, для которых не хватает места в $БН_k$ в $СМО_k$ фазы 2. Отметим, что данный алгоритм может применяться в сети NGMN с централизованной архитектурой, при которой БС обладает информацией о состоянии БН каждой РС. При этом в ситуации 2, когда суммарное число заявок в СМО фазы 2 не превосходит S, распределение $\tilde{\tilde{S}}$ приборов в $СМО_0$ осуществляется по итерационному алгоритму, пропорционально вероятностям c_k принадлежности заявок типу k. Алгоритм **A4** предоставляет в $СМО_0$ для k -заявок только такие по объему группы приборов, которые не приводят далее к потерям на $БН_k$ в $СМО_k$, $k = 1, \dots, K$.

1.4. Система уравнений равновесия

При $0 < a < 1$ ЦМ ξ_n, $n \geq 0$ – неразложима и апериодична, поэтому стационарное распределение вероятностей $[\mathbf{x}], \mathbf{x} \in X$ существует.

Вследствие непростой структуры исследуемой СМО СУР имеет крайне сложное представление; для получения СУР в компактной записи введем обозначения для некоторых выражений с пояснением смысла обозначения, используется обозначение $x \circ y := \min(x, y)$.

Так, число обслуженных заявок в $СМО_k$, очевидно, есть $s_k^{\min} := x_k \circ s_{K+k}, k = 1, \dots, K$; суммарное число обслуженных заявок на фазе 2 за такт равно $s_\cdot^{\min} := \sum_{k=1}^{K} s_k^{\min}$; $s_0^{\min} = s_0 \circ x_0 - s_\cdot^{\min}$;

Число $\tilde{r}_0$ оставшихся заявок в $БН_0$ с учетом числа обслуженных 0-заявок и суммарного числа обслуженных заявок приборами $П_1,\dots,П_K$ в $СМО_0$, и с учетом того, что заявки, обслуженные приборами $П_{K+1},\dots,П_{K+k}$ в $СМО_1,\dots,СМО_K$ восполнятся в том же объеме в $БН_1,..,БН_k$, поступив из $СМО_0$ равно

$\tilde{r}_0 := x_0 - s_\cdot^{\min} - s_0^{\min}$.

Число n_k заявок, поступивших на $П_k$, которым не хватило мест для буферизации в $БН_k$ равно

$n_k := s_k - s_k^{\min} \circ \tilde{r}_0 - \sum_{i=1}^{k-1} \delta(x_i, r_i) n_i, k = 1, \dots, K$.

Данные заявки не входят в число k -заявок, восполняющих $БН_k$ после обслуживания приборами $П_{K+k}$ в $СМО_k$, и рассматриваются только в случае, когда число k -заявок в $БН_k$ совпадает с емкостью накопителя.

Число $s_{0,q}^{\min}$ обслуженных 0-заявок в $СМО_0$ за такт с учетом изменения x_k числа заявок в $БН_k$ на q_k единиц, где q_k принимает как положительные, так и отрицательные значения есть

$s_{0,q}^{\min} := s_0 \circ x_0 - \sum_{k=1}^{K} (s_k^{\min} - q_k)$.

Число $\tilde{r}_{0,q}$ оставшихся заявок в $БН_0$ с учетом обслуженных 0-заявок и суммарного числа обслуженных заявок приборами $П_1,\dots,П_K$ в $СМО_0$ при изменении x_k числа заявок в $БН_k$ на q_k единиц равно

$\tilde{r}_{0,q} := x_0 - \sum_{k=1}^{K} (s_k^{\min} - q_k) - s_{0,q}^{\min}$.

Число $n_{k,q}$ заявок, поступивших на $П_k$, которым не хватило мест для буферизации в $БН_k$ с учетом изменения x_k числа заявок в $БН_k$ на q_k единиц равно

$n_{k,q} \coloneqq s_k - s_k^{\min} + q_k \circ \tilde{r}_{0,q} - \sum_{i=1}^{k-1} \delta(x_i, r_i) n_{i,q}\,, k = 1, \dots, K.$

Данные заявки не входят в число k -заявок, восполняющих БН$_k$ до x_k после обслуживания приборами П$_{K+k}$ в СМО$_k$, и рассматриваются только в случае когда, x_k в БН$_k$ совпадает с r_k.

Введем следующие пространства состояний:

Ω_0 – пространство состояний, удовлетворяющих условию: число заявок в БН$_k$ не превышает числа выделенных для обслуживания приборов П$_{K+k}$ в данной СМО$_k$, при этом рассматриваются все СМО, кроме СМО$_0$:

$\Omega_0 \coloneqq \{\mathbf{x}: x_k \le s_{K+k}, k = 1, \dots, K\}$ для $\forall \mathbf{x} \in X \backslash \mathbf{0}$.

Ω_1 – пространство состояний, при которых число приборов П$_k$ в СМО$_0$ превышает число обслуженных заявок в СМО$_k$ за такт при условии восполнения в БН$_k$ числа обслуженных заявок СМО$_k$ есть

$\Omega_1 \coloneqq \left\{\mathbf{x}: x_0 \ge s_{\cdot}^{\min}, s_k \ge s_k^{\min}, k = 1, \dots, K\right\}$ для $\forall \mathbf{x} \in X \backslash \mathbf{0}$.

Ω_2 – пространство состояний, при которых число приборов П$_k$ в СМО$_0$ превышает число обслуженных заявок в СМО$_k$ за такт с учетом изменения x_k числа заявок в БН$_k$ на q_k единиц есть

$\Omega_2 \coloneqq \left\{\mathbf{x}: x_0 \ge \sum_{k=1}^{K} \left(s_k^{\min} - q_k\right) \ge 0\right\}, q_k = -x_k, \dots, r_k - x_k, k = 1, \dots, K$ для $\forall \mathbf{x} \in X \backslash \mathbf{0}$.

Данное условие гарантирует число x_k заявок в БН$_k$ в СМО$_k$ после фиксации состояния в конце такта.

Введем также два обозначения для выражений, касающихся поступления заявок. Вероятность $\tilde{a}$ поступления числа заявок для сохранения статус-кво за такт, когда система не пуста равна

$$\tilde{a} \coloneqq \begin{cases} \sum_{i=1}^{\infty} a_{s_0^{\min} + s_{\cdot}^{\min} + \sum_{k=1}^{K} \delta(x_k, r_k) n_k + i}\,, \text{если } x_0 = r_0, \\ a_{s_0^{\min} + s_{\cdot}^{\min} + \sum_{k=1}^{K} \delta(x_k, r_k) n_k}, \text{в противном случае.} \end{cases}$$

Следует отметить случай, когда x_0 в БН$_0$ совпадает с r_0, при котором в СМО$_0$ поступают заявки, которые теряются, не оказывая действия на функционирование системы.

Вероятность $\tilde{a}_q$ поступления группы заявок, восполняющих покинувшие СМО заявки для всех q_k до состояния $\mathbf{x}$ есть

$$\tilde{a}_q := \begin{cases} \sum_{i=1}^{\infty} a_{s_{0,q}^{\min}+\sum_{k=1}^{K}\left(s_k^{\min}-q_k\right)-q_0+\sum_{k=1}^{K}\delta(x_k,r_k)n_{k,q}+i}, \text{если } x_0 = r_0, \\ a_{s_{0,q}^{\min}+\sum_{k=1}^{K}\left(s_k^{\min}-q_k\right)-q_0+\sum_{k=1}^{K}\delta(x_k,r_k)n_{k,q}}, \text{в противном случае.} \end{cases}$$

Для случая, когда x_0 в БН$_0$ совпадает с r_0, можно сделать аналогичное приведенному выше замечание.

Перейдем к получению СУР для рассматриваемой двухфазной СМО, используя введенные обозначения.

Утверждение 1.1. *Система уравнений равновесия для цепи Маркова* $\xi_n, n \geq 0$ *имеет вид:*

$$a[\mathbf{0}] = \bar{a} \sum_{\Omega_0} c_0^{x_0} [\mathbf{x}], \tag{2.1}$$

$$1 - \sum_{\Omega_1} c_0^{s_0^{\min}} \prod_{k=1}^{K} c_k^{s_k^{\min}+\delta(x_k,r_k)(1+\cdots+n_k)} \tilde{a}[\mathbf{x}] =$$

$$\sum_{\Omega_2} c_0^{s_{0,q}^{\min}} \prod_{k=1}^{K} c_k^{s_k^{\min}-q_k+\delta(x_k,r_k)(1+\cdots+n_{k,q})} \tilde{a}_q \quad [\mathbf{x} + \sum_{k=0}^{K} q_k \mathbf{e}_k]. \tag{2.2}$$

Стационарное распределение вероятностей $[\mathbf{x}], \mathbf{x} \in X$, *находится из (2.1), (2.2) и нормировочного условия* $\sum_{\mathbf{x}\in X}[\mathbf{x}] = 1$.

Доказательство. Рассмотрим уравнение (2.1). Из состояния $\mathbf{0}$ можно выйти с вероятностью a лишь за счет поступления одной или более заявок. В состояние $\mathbf{0}$ можно войти в случае, когда произойдет освобождение всех БН$_k$ в СМО$_k$ за счет обслуживания k-заявок приборами П$_{K+k}$, $k = 1, \dots, K$. При этом в СМО$_0$ происходит обслуживание 0-заявок, адресованных UE, что соответствует слагаемому $c_0^{x_0}$, и с вероятностью $\bar{a}$ не поступают заявки в БН$_0$.

Рассмотрим уравнение (2.2). Его левая часть представляет собой интенсивность выхода за такт из состояния $\mathbf{x}$, где второе слагаемое соответствует вероятности сохранения статус-кво за такт, когда система не пуста. Действительно, выполнение неравенства $s_k \geq s_k^{\min}, k = 1, \dots, K$, предполагает, что k-заявки, обслуженные приборами П$_{K+k}$ в СМО$_k$,

восполнятся в БН$_k$, поступив с приборов П$_k$ из СМО$_0$, обслуживание на которых указывается с помощью $\prod_{k=1}^{K} c_k^{s_k^{\min}}$. Принимая во внимание число оставшихся заявок в БН$_0$, обслуживание 0-заявок в СМО$_0$ определено с помощью слагаемого $c_0^{s_0 \circ x_0 - s_\cdot^{\min}}$. Перед фиксацией состояния необходимо восполнить все покинувшие СМО заявки за счет поступления с вероятностью a с индексом, соответствующим их суммарному числу.

Правая часть уравнения (2.2) описывает все возможные состояния входа в состояние **x** за счет изменения $q_k = -x_k, \dots, r_k - x_k$, далее $k = 1, \dots, K$. При этом выполнение неравенства $s_k \geq s_k^{\min} - q_k$ предполагает наличие необходимого числа приборов П$_k$ в СМО$_0$, обслуживание числа k -заявок на которых, указанное в виде $\prod_{k=1}^{K} c_k^{s_k^{\min} - q_k}$, позволяет восполнить недостающие до **x** заявки в БН$_k$. Аналогично левой части уравнения, число обслуженных 0-заявок на П$_0$ определяется в виде $c_0^{s_0^{\min}}$, а восполнение числа покинувших СМО заявок для всех q_k до состояния **x** происходит за счет поступления группы заявок с вероятностью a с индексом, соответствующим их суммарному числу. Следует отметить случай, когда число заявок x_k соответствует размеру БН$_k$. При этом все поступившие заявки теряются и не оказывают влияния на функционирование СМО. Для учета подобных состояний используется символ Кронекера.

Следствие 1.1. *Стационарное распределение вероятностей позволяет получить формулы для основных ВВХ.*

Вероятность π_k потерь k-заявки в СМО$_k$, $k = 0, \dots, K$, есть

$\pi_0 = \sum_{g=r_0-\tilde{r}_0+1}^{\infty} a_g \sum_X [\mathbf{x}]$,

$\pi_k = \sum_{\tilde{s}_k = r_k - x_k + s_k^{\min}}^{n_k} c_k^{\tilde{s}_k} \sum_{\vec{s}: s_k > r_k - x_k + s_k^{\min}} [\mathbf{x}], k = 1, \dots, K$,

а вероятность π потерь в системе –

$\pi = 1 - \prod_{k=0}^{K} \pi_k$.

Среднее число потерянных заявок на СМО$_k$, $k = 0, \dots, K$, за такт вычисляется по формулам

$$P_0 = \sum_{g=r_0-\tilde{r}_0+1}^{\infty} g a_g \sum_X [\mathbf{x}],$$

$$P_k = \sum_{\tilde{s}_k=r_k-x_k+s_k^{\min}}^{n_k} \tilde{s}_k \, c_k^{\tilde{s}_k} \sum_{s:s_k>r_k-x_k+s_k^{\min}} [\mathbf{x}], k = 1, \dots, K.$$

Среднее число потерянных заявок в системе за такт есть

$P = P_{\cdot}$.

Среднее число N_k заявок в СМО$_k$, $k = 0, \dots, K$ –

$N_k = \sum_X x_k[\mathbf{x}]$.

Среднее число N заявок в СМО есть

$N = N_{\cdot}$.

Среднее число S_k приборов в СМО$_k$ (например, для пропорционального варианта распределения приборов) есть

$$S_k = \sum_X \left\lfloor \frac{x_k S}{x_{\cdot}} \right\rfloor [\mathbf{x}], k = 1, \dots, K.$$

$$S_0 = \sum_X \left(\left(\tilde{S} - \sum_{k=1}^{K} s_k \right) + \sum_{k=1}^{K} \left\lfloor \frac{x_k \tilde{S}}{x_{\cdot}} \right\rfloor \right) [\mathbf{x}], \tilde{S} = S - \sum_{k=1}^{K} s_{K+k}.$$

Среднее число U_k обслуженных заявок в СМО$_k$ за такт и общее число U обслуженных заявок в СМО есть, соответственно:

$U_k = \min(N_k, S_k)$,

$U_0 = \min(N_0, S_0)$,

$U = U_{\cdot}$.

Среднее время пребывания T_0 в СМО$_0$ есть

$$T_0 = \frac{N_0}{a_0^{cp}(1-\pi_0)}, a_0^{cp} = a \sum_{i=1}^{\infty} i g_i.$$

1.5. Сравнительный численный анализ показателей эффективности для различных алгоритмов распределения ресурсов

Представленные выше ВВХ можно рассматривать в качестве основных показателей эффективности функционирования соты гетерогенной сети NGMN. Для получения численных результатов и анализа алгоритмов распределения РБ в соте сети NGMN, был разработан программный комплекс на языке программирования C++, реализующий итерационный метод для расчета $[\mathbf{x}], \mathbf{x} \in X$. Применение итерационного метода в данном случае рационально вследствие сильной разреженности

матрицы переходных вероятностей ЦМ $\xi_n, n \geq 0$ и сложности получения матрично-рекуррентного решения из-за ее многомерности. Моделирование проводится для модели с числом приборов, равным 30. При этом предполагается, что емкости буферных накопителей принимают значения, заданные вектором $\mathbf{r}^T = (45,7,7,7)$, что соответствует пропорциям структурных данных реальной соты LTE [67]. Объем поступающей группы заявок g имеет распределение Пуассона со средним равным 18. Распределение вероятностей принадлежности заявки типу k=0,1,2,3 задается вектором $\mathbf{c}^T = (\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$. На рис.1.4-1.13 представлены графики для некоторых показателей исследуемых алгоритмов распределения ресурсов при изменении вероятности поступления a за такт.

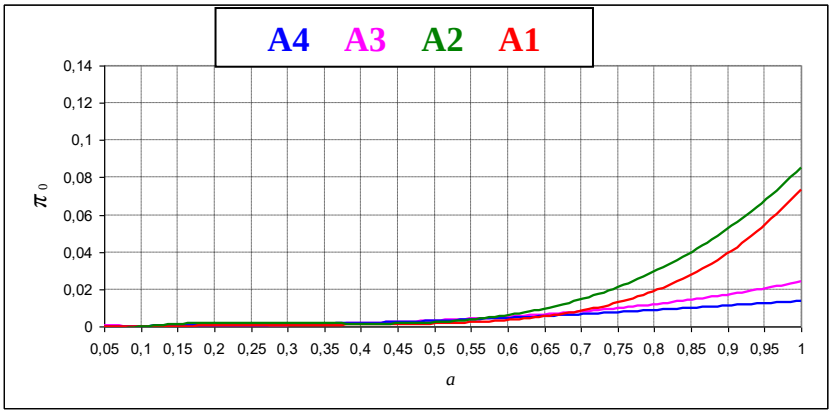


Рис.1.4. Графики зависимости вероятности потерь в СМО$_0$ от вероятности поступления заявки на СМО

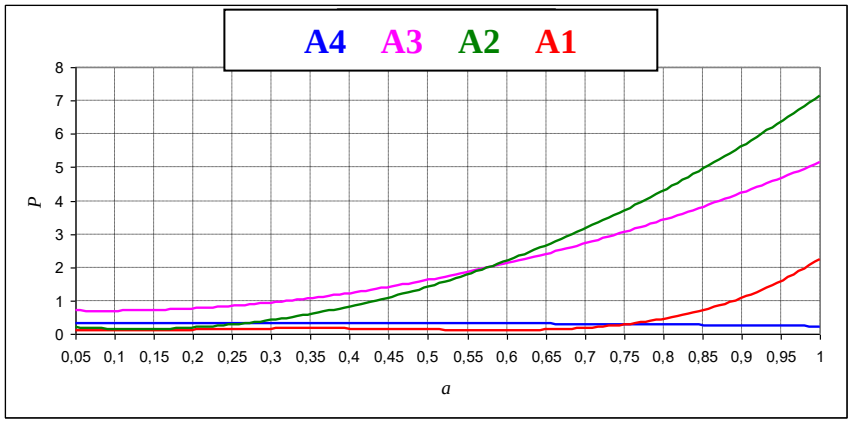


Рис.1.5. Графики зависимости среднего числа потерянных заявок в СМО за такт от вероятности поступления заявки на СМО

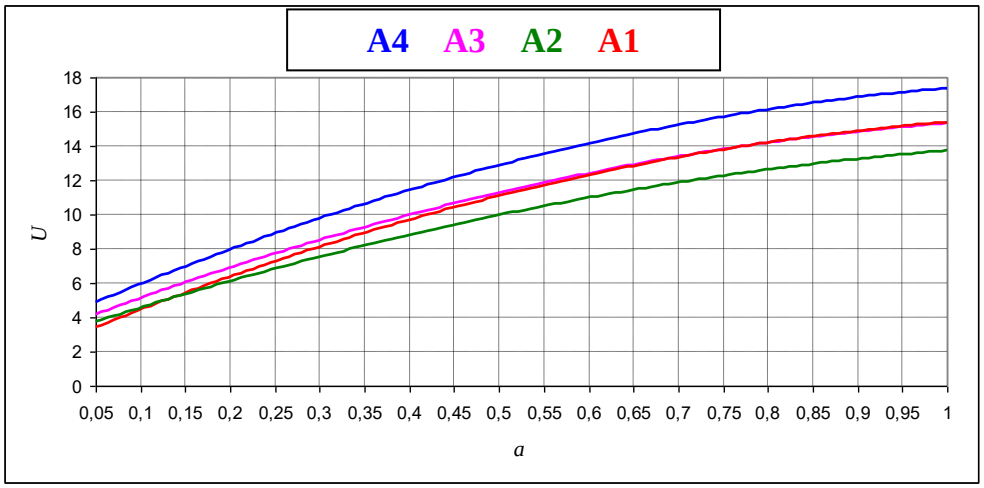


Рис.1.6. Графики зависимости среднего числа обслуженных заявок в СМО за такт от вероятности поступления заявки на СМО

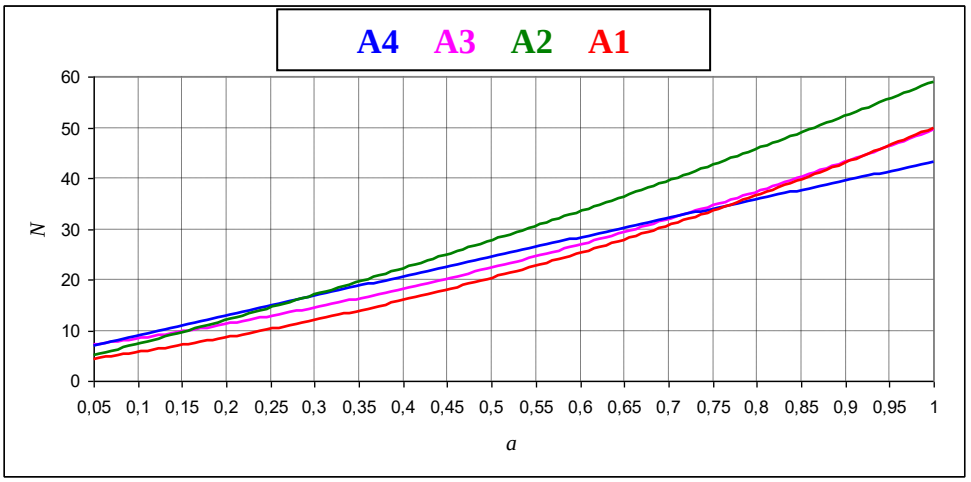


Рис.1.7. Графики зависимости среднего числа заявок в СМО от вероятности поступления заявки на СМО

A1. Как показано на рис.1.4 данный алгоритм характеризуется предоставлением фиксированного наименьшего числа приборов в $СМО_0$, и соответственно более высоким предоставлением приборов в $СМО_k$, рис.1.13. Следует отметить довольно высокую вероятность потери заявок в $СМО_0$, рис.1.4, но относительно низкое суммарное число потерянных заявок в СМО, рис.1.5, что говорит о низких потерях в $СМО_k$, рис.1.11. Однако, как видно из рис.1.6 число обслуженных заявок уступает пропорциональным алгоритмам **A3** и **A4**.

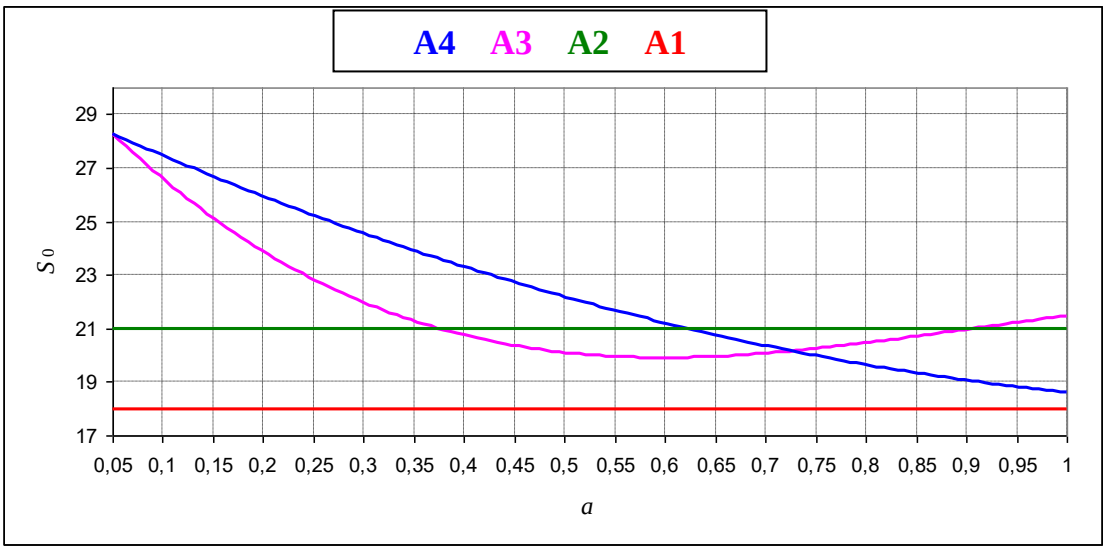


Рис.1.8. Графики зависимости среднего числа приборов в $СМО_0$ от вероятности поступления заявки на СМО

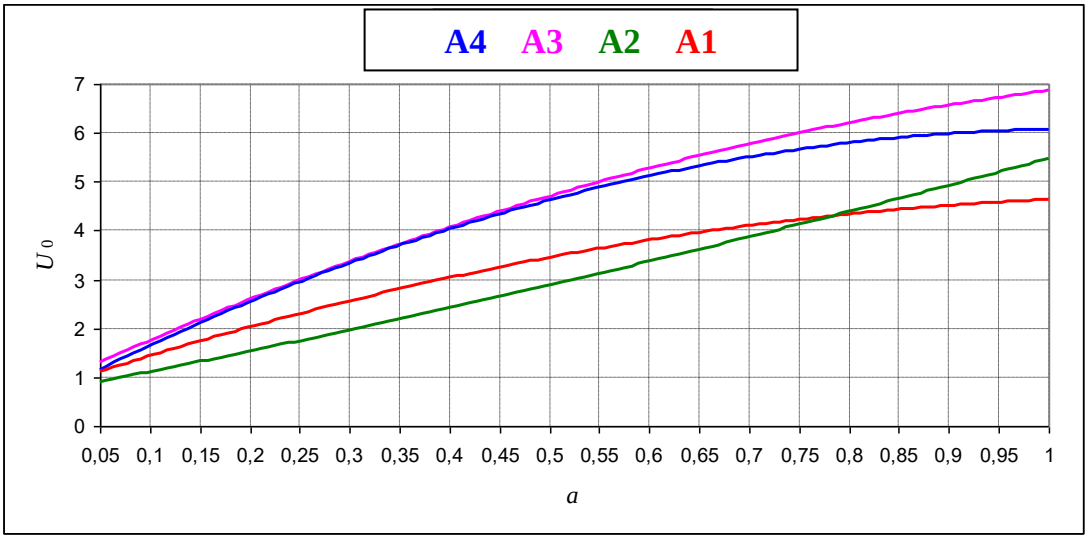


Рис.1.9. Графики зависимости среднего числа обслуженных 0-заявок в $СМО_0$ за такт от вероятности поступления заявки на СМО

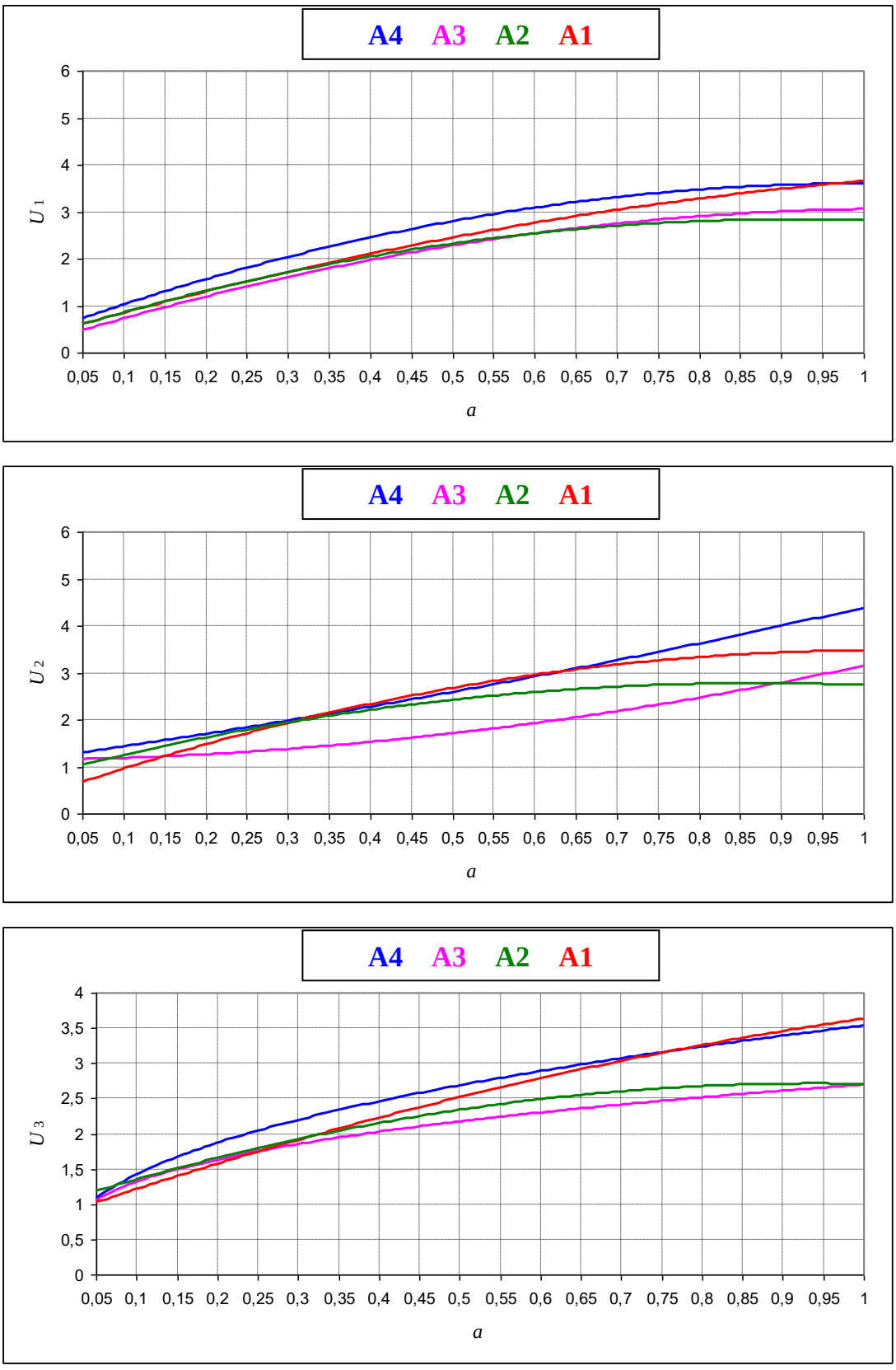


Рис.1.10. Графики зависимости среднего числа обслуженных заявок в СМО$_k$, k=1,2,3, за такт от вероятности поступления заявки на СМО

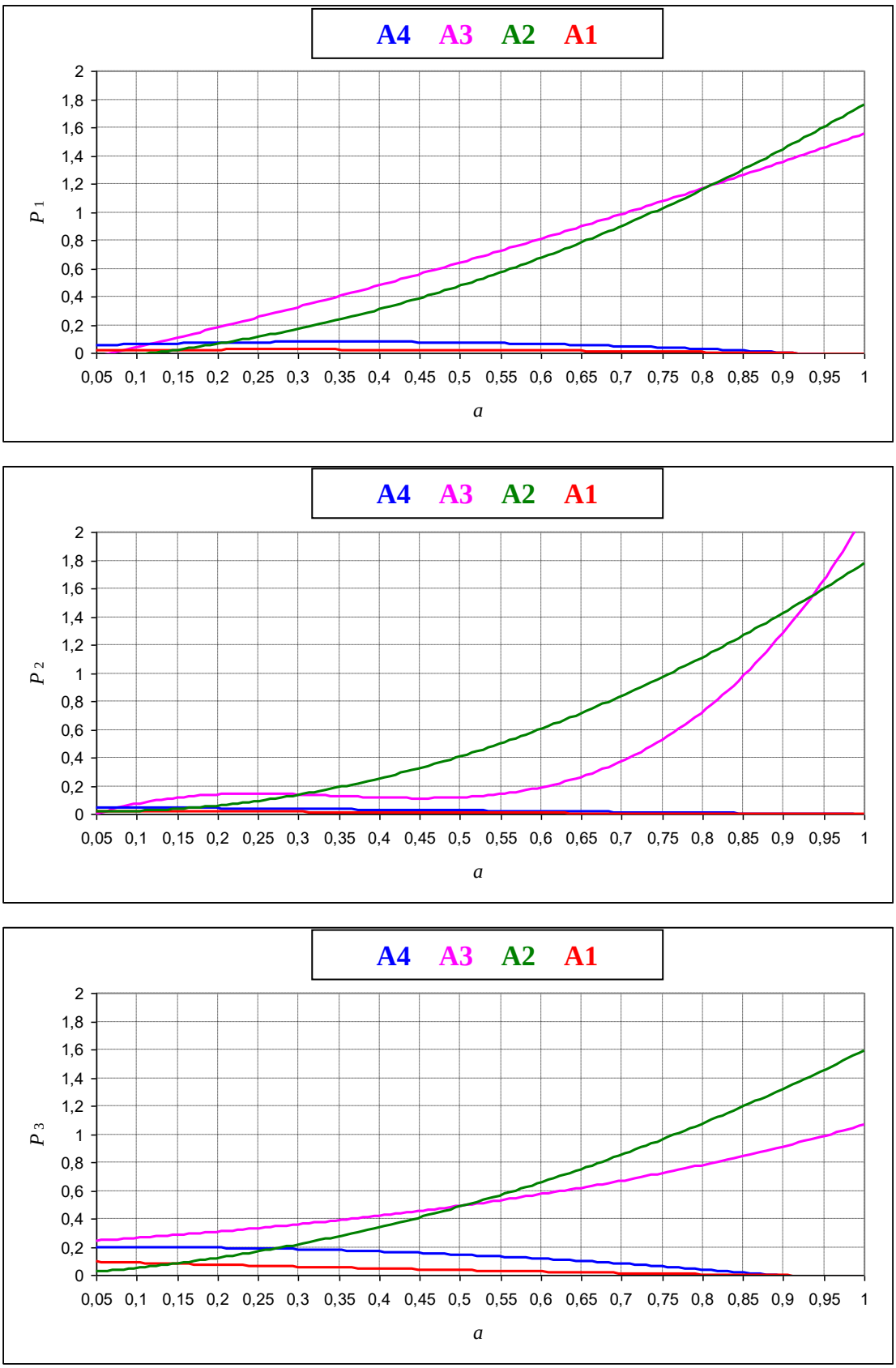


Рис.1.11. Графики зависимости среднего числа потерянных заявок в СМО$_k$, k=1,2,3, за такт от вероятности поступления заявки на СМО

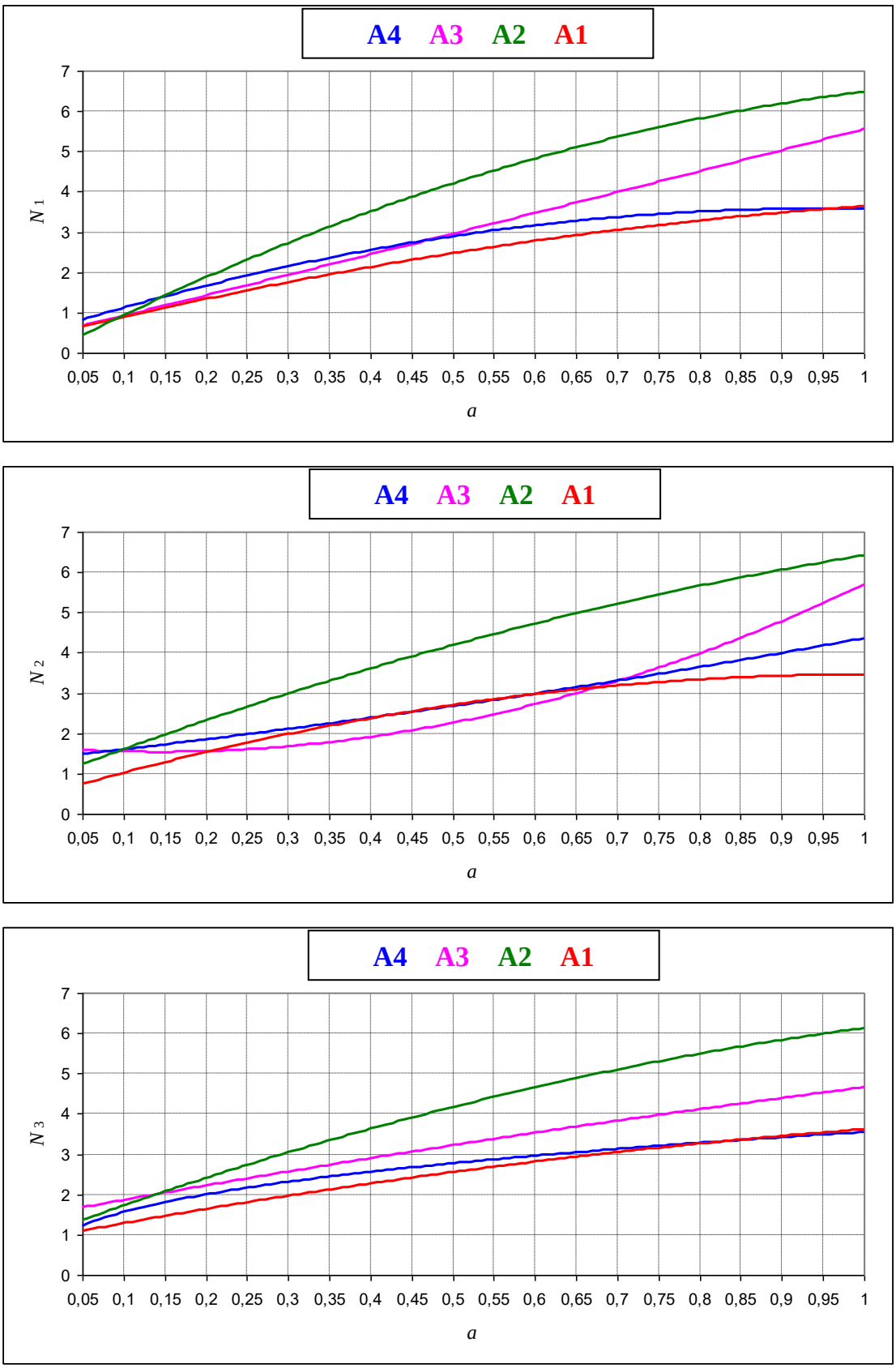
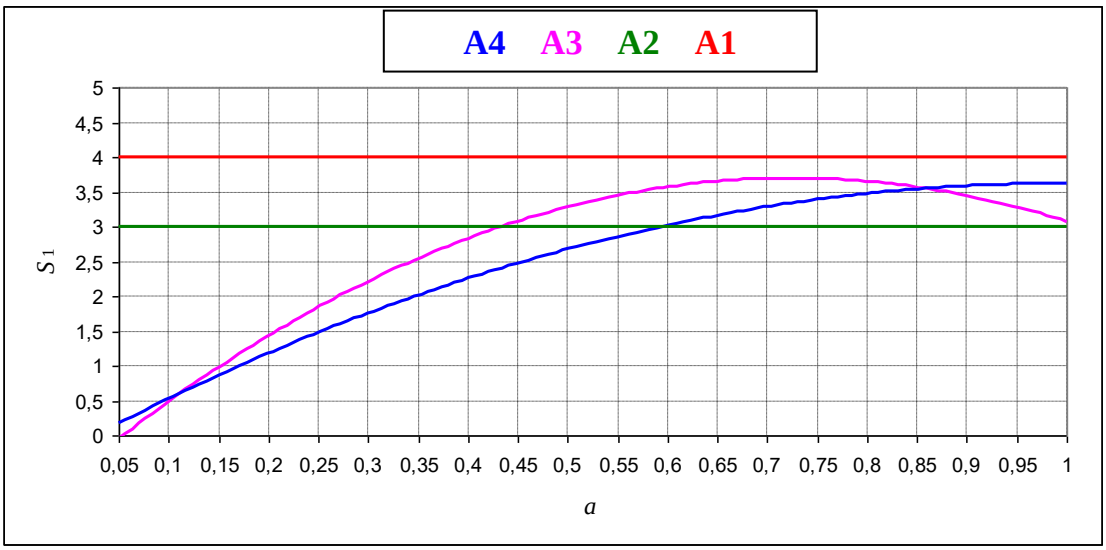
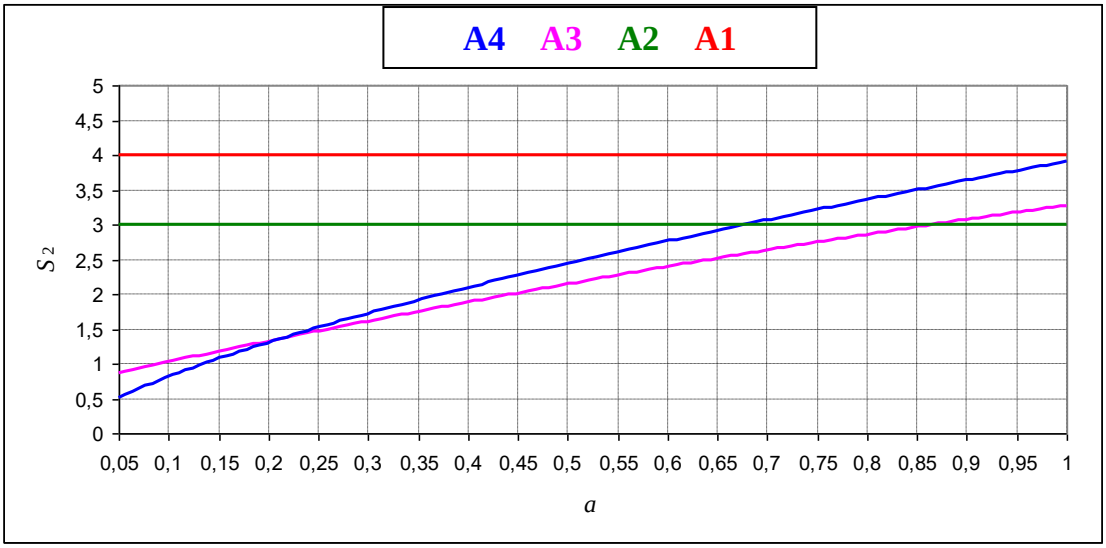


Рис.1.12. Графики зависимости среднего числа заявок в СМО$_k$, k=1,2,3, от вероятности поступления заявки на СМО





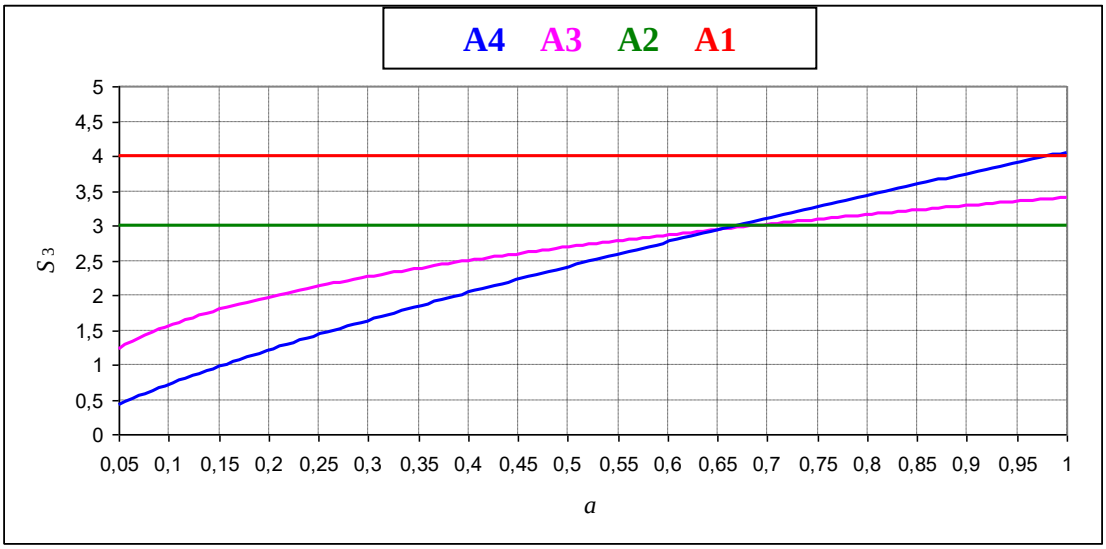


Рис.1.13. Графики зависимости среднего числа приборов в СМО$_k$, k=1,2,3, от вероятности поступления заявки на СМО

A2. Данный детерминированный алгоритм назначает большее число приборов в $СМО_0$ по сравнению с **A1**, что продемонстрировано на рис.1.8, предоставляя при этом меньшее число приборов в $СМО_k$, k=1,2,3, рис.1.13. Отметим, что для гетерогенной сети с одним уровнем ретрансляторов данный алгоритм распределения является наименее рациональным, что подтверждается наибольшей вероятностью потерь в $СМО_0$, рис.1.4, наибольшим суммарным средним числом потерянных заявок в СМО, рис.1.5, и наибольшим средним числом заявок в СМО, рис.1.7.

A3. Пропорциональный алгоритм **A3** показывает рациональное поведение при невысоких нагрузках, однако при вероятности поступления группы заявок в СМО выше 0.7, начинает предоставлять высокое число приборов в $СМО_0$, рис.1.8, что сказывается на снижении среднего числа приборов в $СМО_k$, рис.1.13, увеличении среднего числа потерянных заявок в $СМО_k$, рис.1.11, и, соответственно, совокупном высоком среднем числе потерянных заявок, рис.1.5. При этом, имея необходимое число приборов в $СМО_0$, вероятность потери заявок в $СМО_0$ меньше, чем при детерминированных алгоритмах, рис.1.4. Однако, совокупное среднее число обслуженных заявок в СМО уступает алгоритму **A4**, рис.1.6.

A4. Вероятность потери для алгоритма **A4** является наименьшей в $СМО_0$, рис.1.4, а также **A4** характеризуется наименьшим совокупным средним числом потерянных заявок в СМО, рис.1.5. Более того, среднее число обслуженных заявок в СМО является наибольшим, рис.1.6, а среднее число заявок в СМО – наименьшим, рис.1.7, что говорит о высокой пропускной способности при использовании алгоритма **A4**. Графики на рис.1.8 демонстрируют постоянное уменьшение приборов в $СМО_0$ с увеличением нагрузки, и смещение их в сторону $СМО_k$, рис.1.13. Это позволяет снизить вероятность потери в $СМО_k$, как показано на рис.1.11. Учитывая полученные результаты, отметим, что данный усовершенствованный алгоритм с ограничениями позволяет наиболее адекватно отвечать требованиям гетерогенной сети, характеризующейся изменением объемов трафика.

ГЛАВА 2

МОДЕЛЬ ДВУХФАЗНОЙ СМО В ДИСКРЕТНОМ ВРЕМЕНИ С УПРАВЛЯЕМЫМИ ЦЕПЬЮ МАРКОВА ПОСТУПАЮЩИМ ПОТОКОМ И ОБСЛУЖИВАНИЕМ ДЛЯ РЕШЕНИЯ ЗАДАЧИ МЕЖУРОВНЕВОЙ ОПТИМИЗАЦИИ

2.1. Формулирование задачи распределения ресурсов на основе межуровневого подхода при передаче видео

Как уже было упомянуто в первой главе, к преимуществам NGMN можно отнести высокую скорость передачи данных, повышение эффективности функционирования, снижение задержек обработки и передачи, расширение предоставляемых услуг, что позволяет поддерживать большие объемы высокоскоростных видео приложений. Несмотря на повышение эффективности NGMN, существует теоретический верхний предел количества передаваемых бит за единицу времени в сети с ограниченной шириной полосы пропускания и ненулевым фоновым шумом [147].

Учитывая необходимость разделения частотного спектра среди пользователей в сети, характеризующейся динамическим изменением состояния каналов, можно сделать вывод о том, что изменяющаяся во времени пропускная способность останется серьезной проблемой в будущем, несмотря на все более совершенные технологии. Поэтому для повышения качества восприятия (QoE, Quality of Experience) на смену традиционным методам передачи видео с фиксированным качеством приходят технологии адаптации видео потока при передаче.

Под данным понятием понимают адаптацию видео контента на основе нескольких факторов: состояния канала, занятости буфера пользователя, типа устройства и т.п., и, соответственно, относят его к принципам межуровневой оптимизации. В общем случае, межуровневая оптимизация позволяет за счет протокольного взаимодействия обеспечить

оптимизированную передачу данных [89]. Одним из способов адаптации скорости передачи данных является использование масштабируемых форматов медиа кодирования, включая технологии MDC (Multiple Description Coding), SVC (Scalable Video Coding), SP/SI развитие видео кодека H.264/AVC и другие. Однако эти технологии не имеют широкого распространения, в то время как недавно появившаяся технология адаптивной передачи потокового видео по протоколу HTTP (AHS, Adaptive HTTP Streaming) становится все более популярной за счет повсеместного использования протокола HTTP и традиционных видео кодеков, например H.264/AVC.

Метод AHS был стандартизирован 3GPP и стал стартовой точкой создания рабочей группой MPEG (Moving Pictures Experts Group) Объединенного технического комитета № 1 ISO/МЭК технологии MPEG-DASH (Dynamic Adaptive Streaming over HTTP) [107]. Результатом дальнейшей работы партнерства 3GPP стало появление стандарта 3GP-DASH, который имеет общую структуру с MPEG-DASH, и далее для упрощения будет именоваться DASH.

На рис.2.1 представлена последовательность основных событий, возникающих при передаче видео сегментов на базе технологии DASH для нисходящего канала сети LTE, а на рис.2.2 представлена схема передачи видео информации [66], где приведены основные компоненты структуры сети, а также перечислены функции модулей, участвующих в организации передачи видео по протоколу HTTP. Следует отметить, что стандарт DASH определяет только способ описания информации и порядок ее доставки от сервера до пользователя, в то время как применяемые видео кодеки, форматы и процедура выбора видео сегментов на клиентской части не имеют ограничений [149].

На подготовительном этапе организации сеанса видео передачи генерируются видео сегменты, содержащие различные варианты кодирования медиа контента, различающиеся, например, размерами или скоростью видео кадров. Данные видео сегменты хранятся на одном или

нескольких серверах наряду с системой данных описания представления медиа (MPD, Media Presentation Description), включающей описание структуры, технические характеристики и адреса сегментов. Структурированная информация MPD пересылается по сети пользователю, который использует ее для запросов видео сегментов [153].

Наряду с механизмом DASH на рис.2.2 также приведен другой принцип межуровневой адаптации: выбор схемы модуляции и кодирования (MCS, Modulation and Coding Scheme) базовой станцией eNB на основе полученной от пользователя UE информации о состоянии канала.

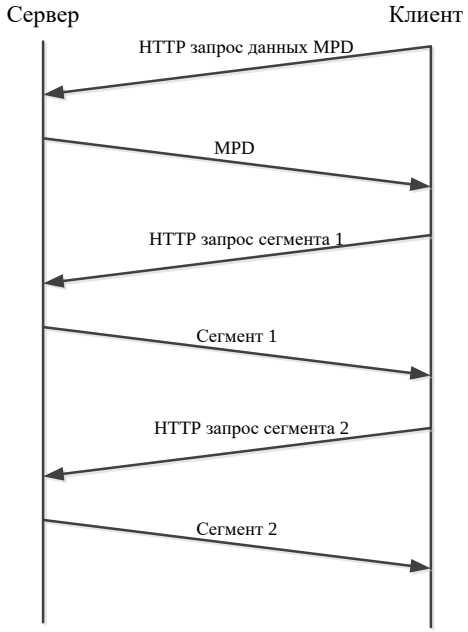


Рис.2.1. Последовательность основных событий, возникающих при передаче видео сегментов на базе технологии DASH

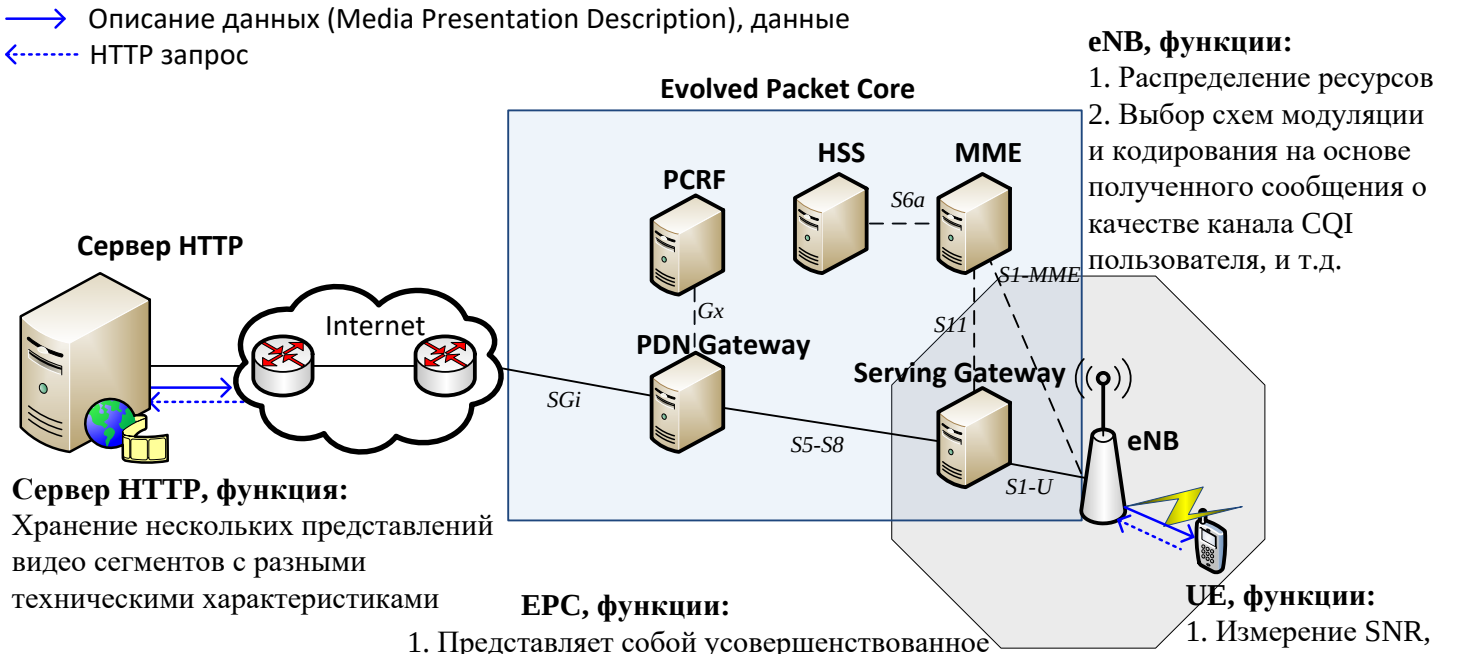


Рис.2.2. Схема передачи видео на базе технологии DASH для нисходящего канала сети LTE

Технологией LTE предусмотрена передача индикатора качества канала (CQI, Channel Quality Indicator) от UE на eNB, содержащего квантованную информацию измерения пользователем отношения SNR. Более детальная информация о различных типах сообщения CQI представлена в [117].

Величина, характеризующая состояние канала, используется на eNB для выбора соответствующей схемы MCS, т.е. модуляции (QPSK, 16-QAM, 64-QAM) и кодирования (1/2, 3/4, 9/16), и распределения ресурсных блоков, что позволяет учитывать изменения канала и адаптивно менять пропускную способность, выделяемую пользователю.

Передача данных по нисходящему каналу сети 5G в настоящее время является популярной исследовательской темой ([72,77,160], см. также библиографию в [89]). Большинство новых схем предполагают существование контроллера на базовой станции, который решает задачи межуровневой оптимизации: динамически изменяет параметры, соответствующие нескольким уровням модели взаимодействия открытых систем, используя при этом исходные данные, включая индикаторы качества, особенности видео контента и выбранную схему компенсации потерянных видео пакетов.

Адаптация обычно происходит за счет решения многомерной задачи оптимизации, например максимизации оценки качества видео контента на кодере. Однако исследования в области адаптивной видео передачи в сети LTE на базе механизма DASH также не стоят на месте [125,127,135,144]. В [135] был предложен адаптивный алгоритм на клиентской стороне, нацеленный на минимизацию повторной буферизации пакетов при передаче видео в сети LTE. В [153] данные о качестве и скорости видео кодирования добавлены в пересылаемую пользователям информацию MPD для улучшения восприятия видео пользователями. В [125] анализируется сценарий видео передачи на основе протокола HTTP, при котором скорость кодирования пересылаемого видео выбирается на основе полученных оценок пропускных способностей всех пользователей в сети LTE. Главным ограничением предложенных ранее алгоритмов

является невозможность одновременного анализа видео передачи по нисходящему каналу на базовой и пользовательской станциях.

2.2. Построение двухфазной СМО в дискретном времени для повышения пропускной способности сети и качества восприятия видео потока на пользовательской станции

Рассмотрим функционирование соты сети NGMN, в которой происходит передача видео по протоколу HTTP от БС к одному UE. В модели используется понятие заявки, имеющей физическое значение пакета с видео контентом. Поступающие новые заявки на БС со стороны сервера HTTP и на UE со стороны eNB буферизуются в БН БС и UE, соответственно.

Будем полагать, что емкость БН БС равна $r_1, r_1 < \infty$, и емкость БН UE – $r_2, r_2 < \infty$. При этом будем считать, что поступившие заявки на БН, которым не хватило мест для буферизации, теряются, не возобновляются и не оказывают влияния на дальнейшее функционирование системы. Наконец, заметим, что обслуживаемая заявка занимает одно место в БН.

На рис.2.3 показана структура рассматриваемой двухфазной СМО. Первая фаза моделирует процесс видео передачи на базе технологии DASH от БС к одному UE, а на фазе 2 рассматривается процесс видео декодирования на терминале.

Представленные на рис.2.3 параметры будут определены далее в разделе. Будем рассматривать функционирование системы в дискретном времени с тактом h постоянной длины, равным длительности одного субкадра в сети LTE, что позволяет учесть дискретный характер физических процессов видео передачи в сети LTE. Следует отметить, что распределение ресурсов в нисходящем канале на БС происходит в каждый временной интервал (TTI, Time Transmission Interval), равный 1 мс, что соответствует длительности субкадра [68]. Разделим ось времени на такты h и примем, что все изменения в системе происходят лишь в моменты $nh, n = 1,2, \ldots$. Для определенности будем считать, как и ранее в диссертации, что такт n есть полуинтервал $[nh, (n+1)h)$.

В каждом временном интервале TTI пользователь производит измерение качества канала, и передает квантованное значение SNR в виде сообщения CQI на БС. В свою очередь, согласно технологии DASH UE контролирует адаптацию видео контента за счет выбора следующего сегмента на основе измеренного состояния канала и передачи DASH запроса на сервер по протоколу HTTP. Учитывая, что оба сообщения, DASH запрос и CQI, содержат коррелированную информацию о состоянии канала, предположим существование между ними строгого соответствия. В модели будем рассматривать передачу только CQI, подразумевая при этом, что значение DASH запроса может быть легко найдено из полученного индикатора. Таким образом, в каждом временном такте значение CQI $s, s = \overline{1,S}$, где S – общее число его возможных значений, доступно на БС, а значение запроса DASH доступно на сервере HTTP, что показано пунктирной линией на рис.2.3. Отметим, что изменение переменной s в общем случае моделируется с помощью графа вероятностей переходов, где s_{ij} – есть вероятность перехода s из состояния i в j. При этом общий случай данного графа с возможными переходами из состояния в любое другое состояние используется для описания городской зоны с высокой плотностью застройки и частыми многолучевыми замираниями.

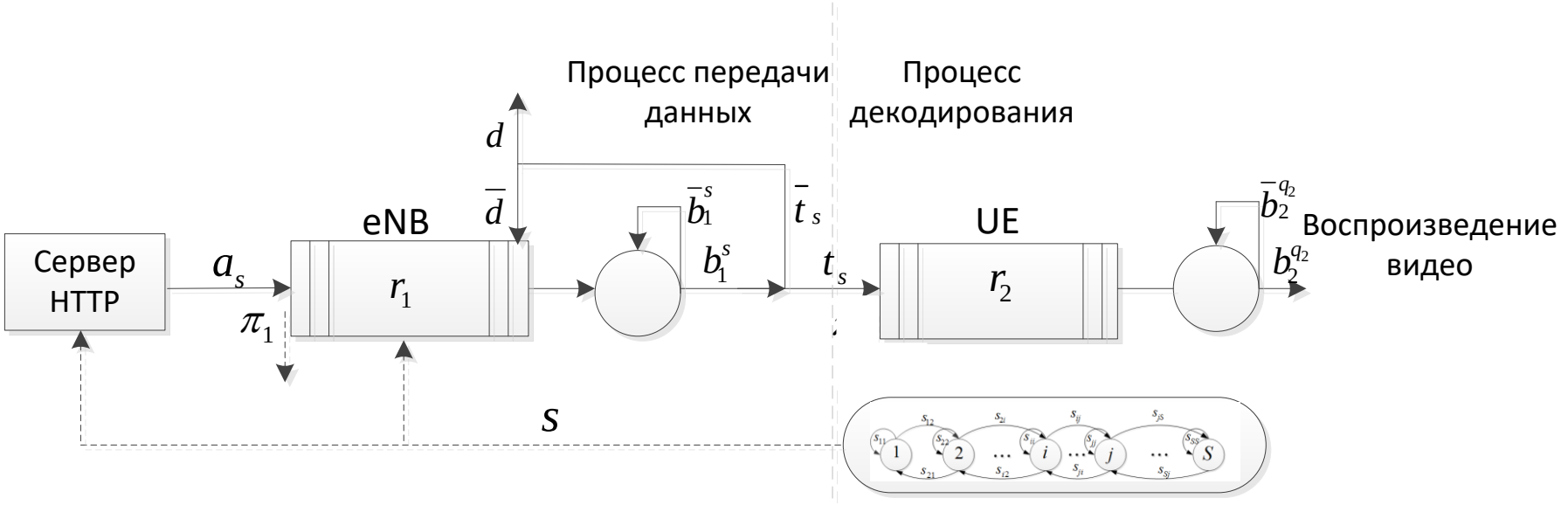


Рис.2.3. Структура двухфазной модели, описывающей процесс передачи видео от БС (eNB) к UE (фаза 1) и процесс видео декодирования на терминале UE (фаза 2)

В качестве частного случая в модели рассматривается граф с переходами только в соседние состояния, что соответствует видео передаче в местности с плавным изменением CQI, например, в сельской. В целях формализации функционирования СМО будем предполагать следующую, изображенную на рис.2.4, последовательность возможных событий, происходящих в такте n (в момент nh) и, в целом, соответствующих функционированию реальной системы.

Адаптация видео передачи в модели учитывается изменением поступающего потока заявок в зависимости от состояния запроса DASH. Предполагая, что видео пакеты имеют одинаковую длину, запрос сегмента требуемого качества (соответствующего видео кодирования) и, соответственно, не одинаковой длины моделируется в двухфазной модели с помощью изменения интенсивности поступающего потока a_s: более длинному сегменту соответствует большая интенсивность.

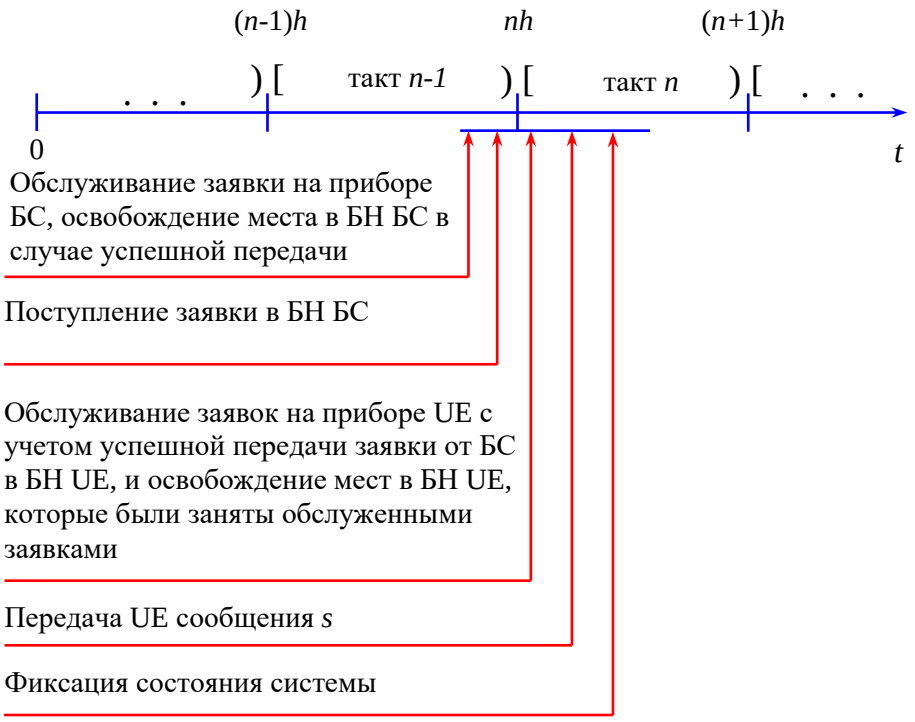


Рис.2.4. Временная диаграмма последовательности событий в рассматриваемой СМО в дискретном времени

Далее параметры с индексом s будут определять аналогичную с a_s функциональную зависимость от s. Таким образом, поступающий поток описывается геометрическим законом распределения, т.е. в течение такта заявка поступает на БН БС с вероятностью a_s, с дополнительной вероятностью $\overline{a_s} = 1 - a_s$ поступления заявки нет.

Процесс обслуживания на первой фазе описывается геометрическим законом распределения с параметром b_1^s, соответствующим возникновению ситуаций недостаточности выделяемых БС ресурсов для передачи пакета в одном или нескольких субкадрах (тактах).

При этом сценарий с несколькими пользователями будем учитывать следующим образом: число пользователей в сети считается неизменным, что позволяет принимать различные значения b_1^s, отражающие влияние других пользователей в сети на вероятность обслуживания рассматриваемого пользователя.

Межуровневая адаптация обеспечивается за счет изменения вероятности обслуживания b_1^s в зависимости от состояния индикатора CQI s в данном такте. Например, в условиях отличного состояния канала БС использует модуляцию в выделенных пользователю ресурсных блоках с более высокой скоростью канального кодирования, что позволяет повысить пропускную способность.

Потеря заявки, возникающая в канале в связи с интерференцией и затуханием сигнала, приводит к повторной передаче с вероятностью $\overline{t_s}\bar{d}$ (рис.2.3), в то время как $\overline{t_s}d$ – есть вероятность истечения времени действия заявки, что соответствует ситуации непригодности пакета для видео воспроизведения по таймауту.

На второй фазе процесс обслуживания описывается геометрическим законом с опустошением $Geom^E$, при котором все заявки в БН второй фазы одновременно обслуживаются с вероятностью $b_2^{q_2}$, зависящей от числа q_2 заявок в БН, с учетом поступления заявки с прибора БС на БН UE. Выбор данного распределения имеет следующее физическое объяснение: для того, чтобы декодировать видео сегмент, UE ожидает

передачи всех пакетов данного сегмента. Данный подход позволяет избежать ошибок при декодировании [136,161]. Вероятность обслуживания $b_2^{q_2}$ возрастает с ростом числа q_2 заявок в БН UE. В модели предполагается, что в случае возникновения полной занятости БН второй фазы при поступлении заявки с первой фазы, вероятность обслуживания на фазе 2 становится равной 1, все заявки второй фазы обслуживаются, что исключает при этом потери заявок на ней.

Таким образом, для предложенной модели в дискретном времени можно использовать следующее мнемоническое обозначение: $Geom_s|Geom_s|1|r_1 < \infty \to Geom^E(q_2)|1|r_2 < \infty$, где $Geom_s$ определяет зависимость геометрического распределения от значения параметра s, а $Geom^E(q_2)$ – зависимость геометрического распределения с опустошением от числа q_2 заявок на второй фазе.

2.3. Система уравнений равновесия и ее решение

Поведение СМО описывается однородной ЦМ $\zeta_n = (\xi_n, \psi_n, \eta_n)$ по моментам $nh + 0, n \geq 0$, над пространством состояний:

$X = (\mathbf{x}^T = (q_1, q_2, s): q_1 = 0,1, \dots, r_1; q_2 = 0,1, \dots, r_2 - 1; s = 1,2, \dots, S)$,

где q_1 и q_2 – число заявок в БН первой и второй фазы, соответственно, а s – значение CQI в текущем состоянии системы.

Для ЦМ $\zeta_n, n \geq 0$, несложно выписать СУР и получить распределение вероятностей в матричном виде, но учитывая, что функционирование первой фазы и изменение состояния CQI не зависят от функционирования второй фазы, произведем декомпозицию системы, и проведем анализ ее фаз отдельно, при этом при анализе второй фазы используются результаты, полученные при анализе первой фазы. Этот подход позволяет резко снизить вычислительную сложность задачи. Таким образом, будем рассматривать следующие этапы анализа:

- анализ первой фазы с учетом изменения индикатора CQI s, результатом которого станет нахождения стационарного распределения $\mathbf{p}$ состояния числа заявок на первой фазе и состояния CQI;

- анализ второй фазы с учетом найденного стационарного распределения на первой фазе, и соответственно, входящего потока на вторую фазу, в результате чего будет найдено стационарное распределение **g** состояния числа заявок на второй фазе.

Анализ первой фазы

Функционирование системы $Geom_s|Geom_s|1|r_1 < \infty$ описывается однородной ЦМ $\zeta_n^1 = (\xi_n, \psi_n)$ по моментам $nh + 0, n \geq 0$, над пространством состояний

$X^1 = \cup_{q_1=0}^{r_1} X_{q_1}^1, X_{q_1}^1 = \{(q_1, s), s = 1,2, \dots, S\}$.

С учетом сделанных предположений ЦМ ζ_n^1 – неразложима и апериодична, поэтому существует стационарное распределение вероятностей $\mathbf{p}^T = (\mathbf{p}_0^T, \mathbf{p}_1^T, \dots, \mathbf{p}_{r_1}^T)$, где $\mathbf{p}_{q_1}^T = (p_{q_1 1}, p_{q_1 2}, \dots, p_{q_1 S})$ для $q_1 = 0,1, \dots, r_1$.

Введем матрицу **S** порядка S, описывающую вероятности перехода переменной s:

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1S} \\ s_{21} & s_{22} & \cdots & s_{2S} \\ \vdots & \vdots & \ddots & \vdots \\ s_{S1} & s_{S2} & \cdots & s_{SS} \end{pmatrix}.$$

Стационарное распределение **p** находится из СУР:

$$\mathbf{p}^T(\widetilde{\mathbf{A}} - \mathbf{I}) = \mathbf{0}^T, \quad (2.1)$$

с нормировочным условием

$$\mathbf{p}^T \mathbf{1} = 1, \quad (2.2)$$

где $\mathbf{0}^T = (0,0,\dots,0)$, **I** – единичная матрица, $\mathbf{1}^T = (1,1,\dots,1)$.

Здесь клеточная порядка $r_i + 1$ трехдиагональная матрица $\widetilde{\mathbf{A}}$ определена следующим образом:

$$\widetilde{\mathbf{A}} = \begin{pmatrix} \widetilde{\mathbf{A}}_{00} & \widetilde{\mathbf{A}}_{01} & \mathbf{0} & \cdots & \mathbf{0} \\ \widetilde{\mathbf{A}}_{10} & \widetilde{\mathbf{A}}_{11} & \widetilde{\mathbf{A}}_{12} & \ddots & \vdots \\ \mathbf{0} & \widetilde{\mathbf{A}}_{21} & \widetilde{\mathbf{A}}_{22} & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \widetilde{\mathbf{A}}_{r_{1-1} r_1} \\ \mathbf{0} & \cdots & \mathbf{0} & \widetilde{\mathbf{A}}_{r_1 r_{1-1}} & \widetilde{\mathbf{A}}_{r_1 r_1} \end{pmatrix},$$

где $\mathbf{0}$ – нулевая квадратная матрица соответствующего порядка, а ненулевые подматрицы представлены в виде $\widetilde{\mathbf{A}}_{ij} = \mathbf{A}_{ij} * \mathbf{S}, i, j = 0,1, \dots, r_1$. Используя обозначение $\|d_s\|_{s=1,2,\dots,S}^{D} = diag\{d_1, d_2, \dots, d_S\}$ определим подматрицы $\mathbf{A}_{ij}, i, j = 0,1, \dots, r_1$:

$\mathbf{A}_{00} = \|\bar{a}_s\|_{s=1,2,\dots,S}^{D}$,

$\mathbf{A}_{01} = \|a_s\|_{s=1,2,\dots,S}^{D}$,

для подматриц нижней диагонали:

$\mathbf{A}_{q_1 q_1 - 1} = \|b_s^1 \bar{t}_s d\bar{a}_s + b_s^1 t_s \bar{a}_s\|_{s=1,2,\dots,S}^{D}, q_1 = 1, \dots, r_1$,

для диагональных подматриц:

$\mathbf{A}_{q_1 q_1} = \left\| b_s^1 \bar{t}_s d a_s + \left(\bar{b}_s^1 + b_s^1 \bar{t}_s \bar{d}\right) \bar{a}_s^{1-\delta(r_1, q_1)} + b_s^1 t_s a_s \right\|_{s=1,2,\dots,S}^{D}, q_1 = 1, \dots, r_1$,

для подматриц верхней диагонали:

$\mathbf{A}_{q_1 q_1 + 1} = \left\| \bar{b}_s^1 a_s + b_s^1 \bar{t}_s \bar{d} a_s \right\|_{s=1,2,\dots,S}^{D}, q_1 = 1, \dots, r_1 - 1.$

Принимая во внимание, что $\widetilde{\widetilde{\mathbf{A}}}_{q_1 q_1} = \widetilde{\mathbf{A}}_{q_1 q_1} - \mathbf{I}, q_1 = 0, \dots, r_1$, перейдем к записи решения СУР (2.1) с нормировочным условием (2.2).

<u>Утверждение 2.1.</u> *Для двухфазной СМО в дискретном времени, моделирующей процессы передачи видео и его декодирования на терминале, стационарное распределение вероятностей* $\boldsymbol{p}^T = (\boldsymbol{p}_0^T, \boldsymbol{p}_1^T, \dots, \boldsymbol{p}_{r_1}^T)$, $\boldsymbol{p}_{q_1}{}^T = (p_{q_1 1}, p_{q_1 2}, \dots, p_{q_1 S})$ *для первой фазы имеет следующее рекуррентное представление:*

$\boldsymbol{p}_m^T = \boldsymbol{p}_{m-1}^T \boldsymbol{W}_{m-1}, m = 1, \dots, r_1$,

где

$\boldsymbol{W}_{m-1} = -\widetilde{\boldsymbol{A}}_{m-1m} \left(\widetilde{\widetilde{\boldsymbol{A}}}_{mm} + \boldsymbol{W}_m \widetilde{\boldsymbol{A}}_{m+1m}\right)^{-1}, m = 1, \dots, r_1 - 1,$

$\boldsymbol{W}_{r_1 - 1} = -\widetilde{\boldsymbol{A}}_{r_1 - 1 r_1} \left(\widetilde{\widetilde{\boldsymbol{A}}}_{r_1 r_1}\right)^{-1},$

и вектор $\boldsymbol{p}_0$ *определяется из системы уравнений*

$\boldsymbol{p}_0^T \widetilde{\boldsymbol{W}} = \boldsymbol{e}_S^T$,

где $\widetilde{\boldsymbol{W}}$ *– матрица* $\boldsymbol{W} = \widetilde{\widetilde{\boldsymbol{A}}}_{00} + \boldsymbol{W_0} \widetilde{\boldsymbol{A}}_{10}$, *у которой последний столбец заменен вектором*

$\sum_{q_1=-1}^{r_1 - 1} \left(\prod_{m=0}^{q_1} \boldsymbol{W}_m\right) \mathbf{1}$, *причем* $\prod_{m=0}^{-1} \boldsymbol{W}_m = \mathbf{I}$, *а* $\boldsymbol{e}_S^T = (0, \dots, 0, 1)$.

Анализ второй фазы

Функционирование данной системы описывается однородной ЦМ $\zeta_n^2 = (\eta_n)$ по моментам $nh + 0, n \geq 0$ над пространством состояний: $X^2 = \{(q_2), q_2 = 0,1, \dots, r_2 - 1\}$.

С учетом сделанных предположений ЦМ ζ_n^2 – неразложима и апериодична, стационарное распределение вероятностей $\mathbf{g}^T = (g_0, g_1, \dots, g_{r_2-1})$ существует и находится из СУР:

$$\mathbf{g}^T(\mathbf{B} - \mathbf{I}) = \mathbf{0}^T, \quad (2.3)$$

с нормировочным условием

$$\mathbf{g}^T\mathbf{1} = 1, \quad (2.4)$$

где матрица $\mathbf{B}$ порядка r_2 имеет следующий вид:

$$\mathbf{B} = \begin{pmatrix} -c\bar{b}_2^1 & c\bar{b}_2^1 & & \cdots & 0 \\ cb_2^2 & -c & c\bar{b}_2^2 & \ddots & \vdots \\ cb_2^3 & & -c & \ddots & \\ \vdots & \ddots & \ddots & \ddots & c\bar{b}_2^{r_2-1} \\ c & 0 & & & -c \end{pmatrix}.$$

Используя найденное стационарное распределение $\mathbf{p}$ состояний первой фазы, выражение для вероятности поступления заявки в БН UE имеет вид $c = \sum_{q_1=1}^{r_1} \sum_{s=1}^{S} p_{q_1 s} b_1^s t_s$, и для вероятности, что заявка не поступит на вторую фазу – $\bar{c}$.

Утверждение 2.2. *Стационарное распределение вероятностей* $\boldsymbol{g}^T = (g_0, g_1, \dots, g_{r_2-1})$ *для второй фазы имеет следующий вид:*

$$g_m = \prod_{i=1}^{m} \bar{b}_2^i \, g_0, m = 1, \dots, r_2 - 1, \quad (2.5)$$

где

$$g_0 = \left[1 + \sum_{i=1}^{r_2-1} \prod_{j=1}^{i} \bar{b}_2^j\right]^{-1}. \quad (2.6)$$

Как видно из (2.5), (2.6) полученное распределение $\mathbf{g}$ не зависит от вероятности c поступления заявки с первой фазы в БН UE, и полностью определяется вероятностью обслуживания $b_2^{q_2}$. Отметим, что полученный результат соответствует заданному при описании модели условию: все заявки q_2, находящиеся в БН второй фазы, обслуживаются одновременно

с вероятностью $b_2^{q_2}$ только в случае поступления новой заявки с прибора eNB на БН UE.

Стационарное распределение двухфазной СМО

Получение стационарного распределения вероятностей $\widehat{\mathbf{p}}^T = \left(\hat{p}_{q_1 q_2 s}\right), q_1 = 0, \dots, r_1, q_2 = 0, \dots, r_2 - 1, s = 1, \dots, S$, как указывалось выше, осуществляется с использованием матричных методов, но является громоздким, и при больших значениях параметров системы может приводить к накоплению вычислительной ошибки. Следует отметить, что найденное распределение $\mathbf{g}$ удовлетворяет условию 2 теоремы 1 о независимой работе фаз сложной системы [52]:

«***Теорема 2.1.*** *Стационарное распределение двумерного процесса* $\boldsymbol{f}(x, v) > 0, (x, v) \in X$*, описывающего функционирование двух подсистем СМО с пространством состояний* X*, удовлетворяющее условиям:*

1) $\sum_{(x,v)\in X} \boldsymbol{f}(x, v)\{a_0(x) + a_1(x) + b(v)\} < \infty,$

где $a_1(x)$ *- интенсивность выхода требований из первой подсистемы на вторую,* $a_0(x)$ *- интенсивность переходов первой подсистемы не связанных с выходами таких требований, когда подсистема находится в состоянии* $x \in X_1$*, а* $b(v) = \sum_{\substack{u\in X_2 \\ u\neq v}} b(v, u)$ *- матрица плотностей переходов второй подсистемы;*

2) $f(x, v) = \sum_{u\in X_2} f(x, u) \sum_{y\in X_1} f(y, v),$

существует тогда и только тогда, когда выполнено одно из условий:

Условие 2: $f_2(v) = \sum_{u\in X_2} f_2(u) q(u, v),$

где $q(u, v)$ *- есть вероятность перехода в состояние* v *из состояния* u*. Заметим, что* $u, v \in X_2$*. При этом для любых* $(x, v) \in X$ $f(x, v) = f_1(x) f_2(v)$*».*

Как видно из условия 2 данной теоремы стационарное распределение второй подсистемы или фазы не зависит от входящего потока на данную систему с первой фазы, что полностью соответствует полученному распределению $\mathbf{g}$ (2.5), (2.6). Из вышеизложенного следует следующее утверждение:

Утверждение 2.3. *Для модели двухфазной СМО в дискретном времени стационарное распределение* $\hat{\boldsymbol{p}}$ *имеет мультипликативный вид:*

$\hat{p}_{q_1 q_{2s}} = p_{q_1 s} g_{q_2}, q_1 = 0, \dots, r_1, q_2 = 0, \dots, r_2 - 1, s = 1, \dots, S.$

и вычисляется последовательно на основе полученных распределений $\boldsymbol{p}$ *и* $\boldsymbol{g}$ *для первой и второй фазы, соответственно.*

2.4. Вероятностно-временные характеристики и их численный анализ

Следствие 2.1. *Стационарные распределения вероятностей* $\boldsymbol{p}$, $\boldsymbol{g}$ *и* $\hat{\boldsymbol{p}}$ *позволяют получить основные ВВХ, перечисленные далее отдельно для каждой из фаз и для системы в целом вместе с их краткими описаниями.*

Одной из важнейших характеристик СМО ограниченной емкости является вероятность π_s потери заявки на первой фазе в состоянии s:

$\pi_s = p_{r_1 s}\left(\bar{b}_1^s + b_1^s \bar{t}_s \bar{d}\right),$

а вероятность π потери заявки на фазе 1 в СМО есть

$\pi = \pi_{\cdot} .$

Среднее число N_s^1 заявок в состоянии s на фазе 1 в СМО вычисляется по формуле

$N_s^1 = \sum_{q_1=0}^{r_1} q_1 p_{q_1 s},$

а среднее число N^1 заявок на фазе 1 в СМО равно

$N^1 = N_{\cdot}^1.$

Среднее время пребывания T_s^1 заявок в состоянии s CQI на фазе 1 в СМО находится по формуле

$T_s^1 = \frac{N_s^1}{a_s \bar{\pi}_s}.$

Среднее время пребывания T_s^1 заявок на первой фазе в состоянии s CQI найдено по закону Литтла, однако оно может быть также получено в результате решения системы уравнений равновесия для матрицы $\mathbf{S}$ аналогично уравнениям (2.1), (2.2). Маргинальная вероятность $p_{\cdot s}$ пребывания фазы 1 в состоянии s CQI на фазе 1 в СМО равна

$p_{\cdot s} = \sum_{q_1=0}^{r_1} p_{q_1 s}, s = 1,2, \dots, S,$

а маргинальная вероятность $p_{q_1.}$ нахождения q_1 заявки на фазе 1 в СМО есть

$p_{q_1.} = \sum_{s=1}^{S} p_{q_1 s}, q_1 = 0,1, \dots, r_1.$

Среднее число заявок L_s, покинувших фазу 1 в СМО в состоянии s CQI из-за истечения срока их действия, имеет следующий вид

$L_s = \sum_{q_1=1}^{r_1} q_1 b_1^s \bar{t}_s d p_{q_1 s},$

а вероятность c поступления заявки на фазу 2 в СМО есть

$c = \sum_{q_1=1}^{r_1} \sum_{s=1}^{S} b_1^s t_s p_{q_1 s}.$

Среднее число N^2 заявок на фазе 2 в СМО (пакетов в буфере UE) равно

$N^2 = \sum_{q_2=0}^{r_2-1} q_2 g_{q_2}.$

Среднее время пребывания T^2 заявки на фазе 2 в СМО вычисляется по формуле

$T^2 = \frac{N^2}{c}.$

Вероятность G заполнения фазы 2 в СМО есть

$G = g_{r_2-1}.$

Среднее время пребывания заявки в системе с учетом состояния s с момента поступления на eNB до момента воспроизведения сегмента –

$T_s = T_s^1 + T^2, s = 1,2, \dots, S.$

Вероятность нахождения системы в состоянии простоя равна

$p_{0.} g_0,$

и среднее число заявок в системе находится по формуле

$N = N^1 + N^2.$

Для проведения численных экспериментов было разработано программное обеспечение на языке программирования MATLAB. Согласно [117] существует 16 значений индикатора качества канала CQI, т.е. можно принять $S = 16$, причем увеличение значения соответствует улучшению состояния канала.

Как было упомянуто выше, выбор вероятностей обслуживания на первой фазе b_1^s зависит от пересылаемого пользователем сообщения CQI s. Передатчик eNB на основе данного значения s выбирает схему

модуляции и кодирования, в то время как планировщик eNB в соответствии с используемой схемой распределения ресурсов выделяет их пользователям [80].

Пусть, без ограничения общности, вектор вероятностей обслуживания определяется следующей функциональной зависимостью от параметра γ: $\mathbf{b}_1 = \mathbf{x}^\gamma$, где $\mathbf{x}^T = (x_1, x_2, \dots, x_S)$ – вектор равномерно распределенных значений $0 < x_i < 1, x_i < x_{i+1}, i = 1, \dots, S-1$, и γ – параметр, управляющий формой кривой $\mathbf{b}_1$. Выбор данной функции определяется возможностями получения значений вектора вероятностей обслуживания, вполне соответствующих реальным параметрам соты сети LTE. Могут рассматриваться и другие варианты задания $\mathbf{b}_1$. Очевидно, вероятность обслуживания b_1^s увеличивается с ростом значения индикатора s; однако, вектор вероятностей обслуживания, включающий значения для всех состояний s $\mathbf{b}_1^T = (b_1^1, b_1^2, \dots, b_1^S)$ может быть представлен множеством вариантов возрастающего поведения как показано на рис.2.5, что обусловлено изменением пары параметров: схемы MCS и числа выделенных пользователю ресурсных блоков.

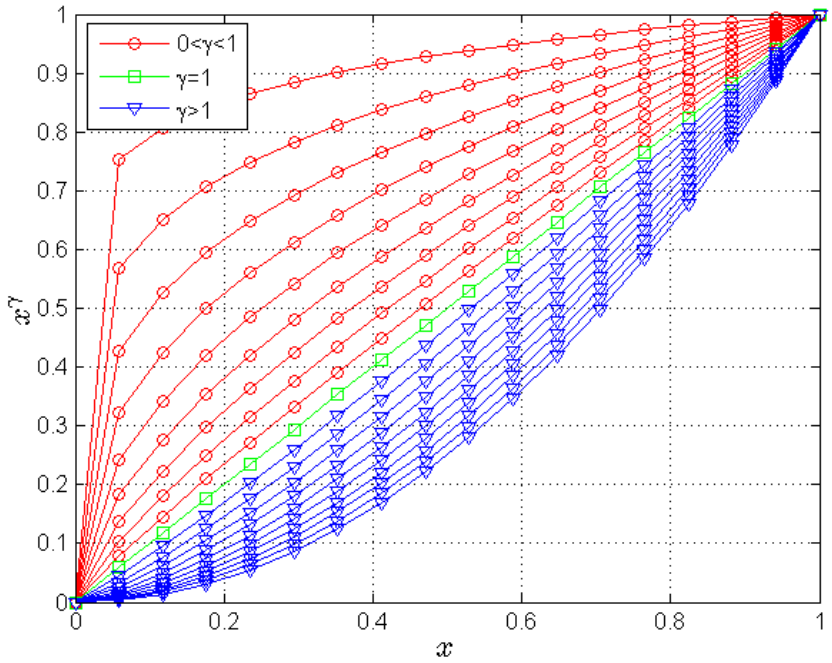


Рис.2.5. Семейство значений вектора $\mathbf{b}_1$ вероятностей обслуживания на первой фазе

В таблице 2.1 представлены значения основных параметров, используемых в численном эксперименте. В соответствии с технологией DASH рассматриваются 4 видео потока, закодированные с разной скоростью передачи. Как было упомянуто выше, существует строгое соответствие между сообщением CQI и DASH запросом следующего сегмента, и интенсивность входящего потока задается на основе состояния индикатора s следующим способом, указанным в таблице 2.1. Следует отметить, что, проводя численный анализ, уделяется внимание скорее качественным, чем количественным характеристикам.

Таблица 2.1. Значения параметров, используемых в численном эксперименте

Параметр	**Значение**
Число значений CQI, S	16
Городская местность, **S**	Задается случайным образом, $s_{i.} = 1, i = 1,2, \dots, S$
Сельская местность, **S**	$\mathbf{S} = \begin{pmatrix} 0.9 & 0.1 & & & \\ 0.1 & 0.8 & 0.1 & & \\ & \ddots & \ddots & \ddots & \\ & & 0.1 & 0.8 & 0.1 \\ & & & 0.1 & 0.9 \end{pmatrix}$
Емкость БН, r_1, r_2	30,10
Вероятность поступления за такт, a_s	$a_s = \begin{cases} 0.25, s = 1,2, \dots,7 \\ 0.5, s = 8,9,10 \\ 0.75, s = 11,12,13,14 \\ 0.95, s = 15,16 \end{cases}$
Вероятность успешной передачи видео потока на фазу 2, t_s	$0.9 < t_s < 1$
Вероятность ухода заявки с фазы 1 из-за истечения срока действия заявки, d	0.2
Вероятность обслуживания на фазе 2, $b_2^{q_2}\, b_2^{q_2}$	Равномерное распределение

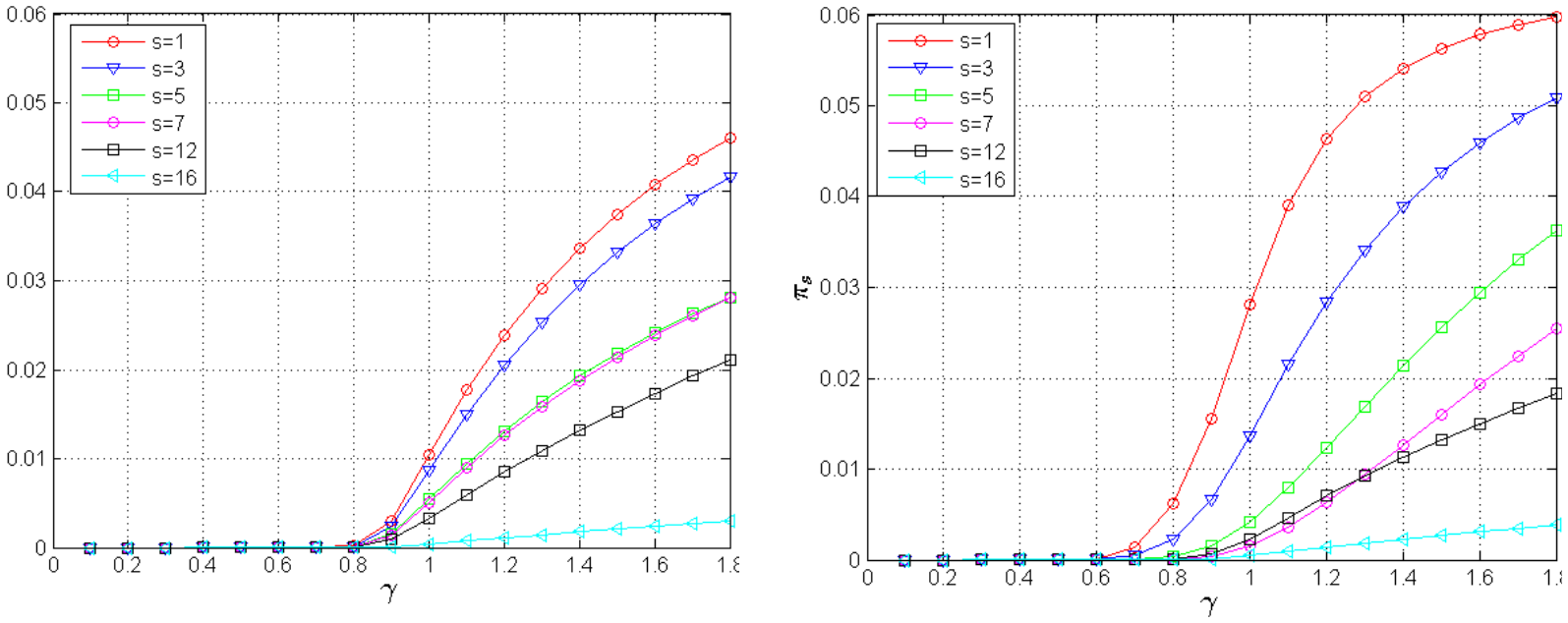
В ходе анализа рассматриваются два сценария передачи видео в сети LTE в городской местности (ГМ) и сельской местности (СМ). На рис.2.6-2.9 изображены графики, представляющие основные ВВХ СМО для городской и сельской местностей в зависимости от изменения вектора вероятностей обслуживания $\mathbf{b}_1$, определяемого параметром γ.

На рис.2.6 представлена вероятность потери заявок на первой фазе в состоянии s (приведено 6 состояний канала) сообщения CQI. Как видно из рис.2.6. в случае, когда вектор вероятностей обслуживания $\mathbf{b}_1$ принимает значения (рис.2.5) над прямой ($\gamma \leq 1$), вероятность потери для сценариев с ГМ и СМ не превышает 0.01, что считается приемлемым в сети LTE. Отметим, что при $\gamma > 1$ векторы вероятностей обслуживания заявок характеризуются очень медленным ростом, что приводит к увеличению вероятности потерь заявок.

Это более характерно для случая СМ и плохих состояний канала s<7, и обусловлено более длительным пребыванием в данных состояниях (см. таблицу 2.1, матрица **S** с переходами только в соседние состояния).

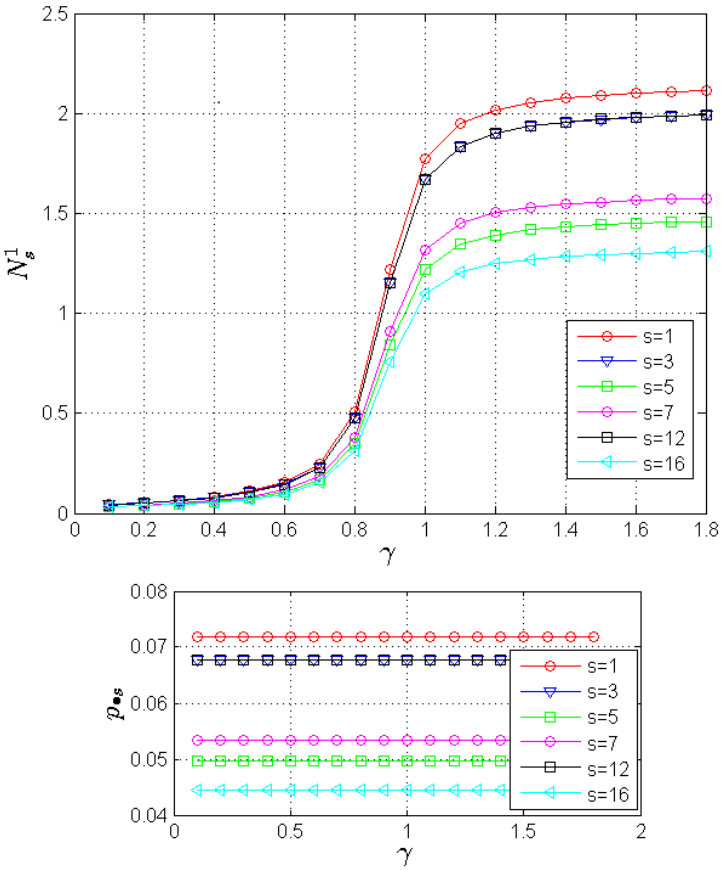
Рис.2.7 демонстрирует графики для среднего числа заявок N_s^1 на первой фазе в состоянии s сообщения CQI, и маргинальную вероятность $p_{.s}$ пребывания в состоянии s, полученную для городской местности. Некоторое непоследовательное поведение данной ВВХ в сценарии ГМ объясняется поведением маргинальной вероятности $p_{.s}$, приведенной для шести состояний s.

На рис.2.8 изображены два графика, иллюстрирующие среднее число заявок, покинувших фазу 1 из-за истечения срока их действия. Оба сценария с ГМ и СМ имеют максимум за счет изменения вектора вероятностей обслуживания $\mathbf{b}_1$. Очевидно, с уменьшением вероятности обслуживания ($\gamma \leq 1$), величина L возрастает, однако, при еще большем увеличении ($\gamma > 1$) у заявки нет возможности покинуть прибор, и соответственно величина L начинает уменьшаться.

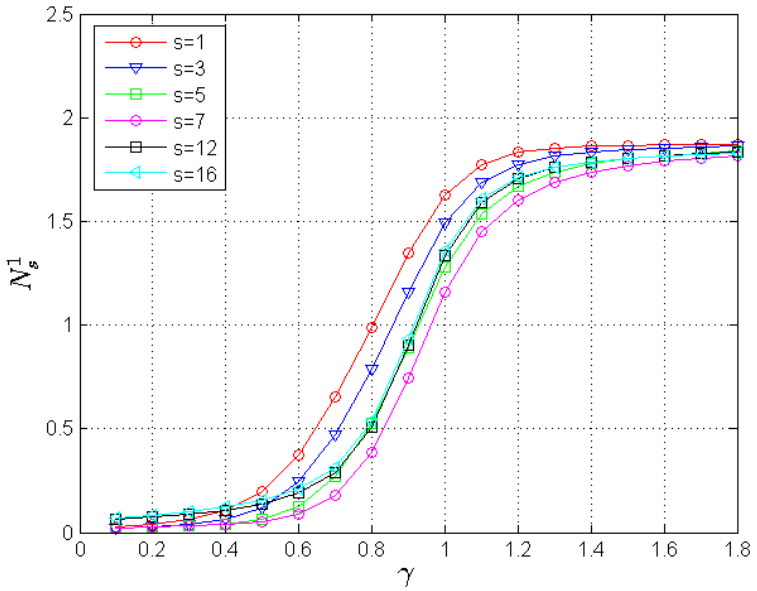


ГМ СМ

Рис.2.6. Графики зависимости вероятности π_s потерь заявок в состоянии s сообщения CQI от изменения вектора вероятностей обслуживания $\mathbf{b}_1$

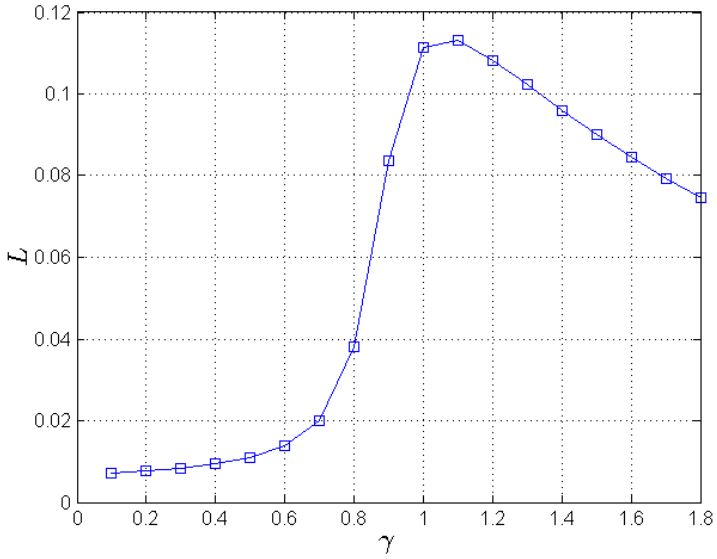


ГМ

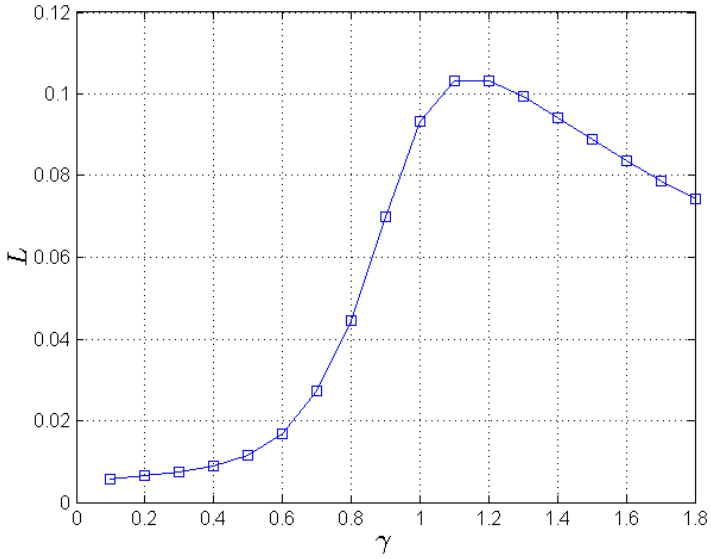


СМ

Рис.2.7. Графики зависимости среднего число N_s^1 заявок на первой фазе в состоянии s и маргинальной вероятности $p_{.s}$ для случая городской местности от изменения вектора вероятностей обслуживания $\mathbf{b}_1$

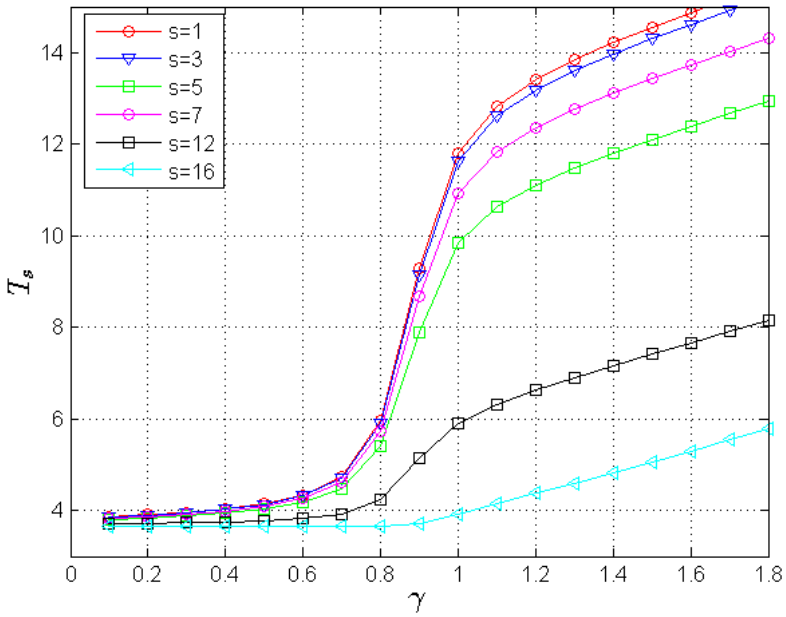


ГМ

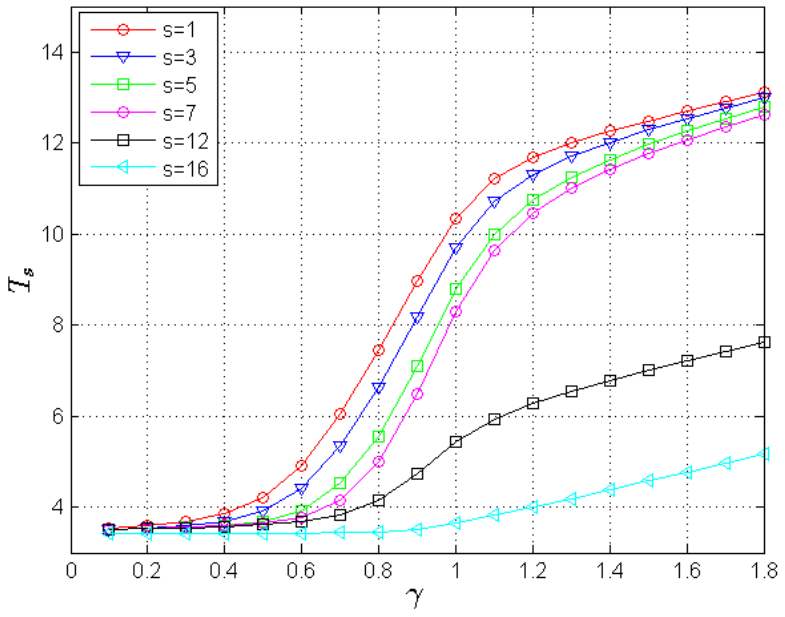


СМ

Рис.2.8. Графики зависимости среднего число заявок L, покинувших систему из-за истечения срока их действия, от изменения вектора вероятностей обслуживания $\mathbf{b}_1$



ГМ



СМ

Рис.2.9. Графики зависимости среднего время пребывания T_s заявки в СМО в состоянии s сообщения CQI от изменения вектора вероятностей обслуживания

ГЛАВА 3

МОДЕЛЬ ДВУХФАЗНОЙ СМО В ДИСКРЕТНОМ ВРЕМЕНИ С УЧЕТОМ МЕХАНИЗМА ПРЕДСКАЗАНИЯ ПОВТОРНОЙ ПЕРЕДАЧИ И МЕХАНИЗМА ОБРАТНОЙ СВЯЗИ ДЛЯ РЕШЕНИЯ ЗАДАЧИ СНИЖЕНИЯ ЗАДЕРЖКИ ПЕРЕДАЧИ ПАКЕТА

3.1. Формулирование задачи ранней адаптации канала на основе механизма предсказания повторной передачи e-HARQ

Мобильные сети 5G обладают огромным потенциалом для инноваций в во всех отраслях экономики и социальной сферы, что обусловлено высокой скоростью передачи данных (1-2 Гбит/с), сниженной задержкой и меньшим расходом энергии батарей пользовательского оборудования UE. К основным услугам, для которых требуется создание сетей нового поколения мобильной связи, относится сверхнадежная межмашинная связь с низкими задержками URLLC, которая характеризуется задержкой до 0.5 мс для обеспечения передачи критически важной информации от датчиков класса critical IoT [154].

Несмотря на высокий потенциал сети мобильной связи пятого поколения, согласно последнему релизу стандарта 5G [69], до сих пор существует ряд нерешенных задач по обеспечению требуемой низкой задержки, что подчеркивает необходимость поиска новых протокольных решений дляуслуг URLLC.

Существует несколько стратегий по уменьшению задержки, например, за счет усовершенствования механизма обратной связи для гибридного автоматического запроса на повторение HARQ [69]. Следует отметить, что HARQ зарекомендовал себя как надежный механизм поиска компромисса между задержкой и спектральной эффективностью. HARQ – протокол физического уровня, предоставляющий передатчику сообщение подтверждения приема ACK в случае успешного декодирования пакета, и NACK – в случае ошибки. Главным его недостатком является ограничение

на RTT – временной интервал между процессами первоначальной и повторной передачи. На данный момент предложено множество схем, направленных на снижение временного интервала RTT [126,128]. В [128] уменьшение RTT достигается за счет сокращения фактической длины передачи, равной одному символу ортогонального частотного мультиплексирования, что приводит к более высоким требованиям, вплоть до мгновенной обработки приемником, и к ограничению мощности передатчика. Другой подход, обсуждаемый в 3GPP, включает в себя автоматическую передачу избыточных версий данных для достижения необходимого уровня надежности до тех пор, пока передатчик не получит первое подтверждение ACK [126]. Тем не менее, это может снизить спектральную эффективность из-за ненужных повторных передач вследствие задержек обратной связи.

Значительные исследовательские усилия были направлены на создание схем предсказания результата декодирования, или предикторов, также называемых схемами раннего предсказания на базе HARQ (early HARQ, e-HARQ), с применением искусственного интеллекта. Большинство подходов предсказания e-HARQ проводят оценку коэффициента битовой ошибки (BER, Bit Error Rate) на основе данных, принятых в виде логарифмов отношения правдоподобия (LLR, Log-Likelihood Ratios), и далее используют пороговые значения для принятия решения о выборе ACK/NACK сообщения обратной связи [74,102,126].

В [102] авторы представили алгоритм прогнозирования e-HARQ, который использует субструктуры кода с малой плотностью проверок на четность (LDPC, Low-Density Parity Check), чтобы начать принятие решения обратной связи уже по частично полученным кодовым словам, и, следовательно, значительно сократить RTT. Более того, методы машинного обучения, которые позволяют более точно предсказать результат декодирования на базе частично полученных кодов детально изучены в [99]. В [139] авторы используют данные результаты для моделирования алгоритма прогнозирования e-HARQ с учетом решений планировщика.

Следует отметить, что большинство известных предикторов e-HARQ используют заданное пороговое значение в алгоритмах как механизм выбора между сообщениями ACK и NACK. Следовательно, выбор порогового значения является критически важным для повышения эффективности схем предсказания e-HARQ.

На рис.3.1,3.2 представлены примеры сравнительной временной диаграммы передачи сообщения от БС к UE между схемами предсказания e-HARQ на базе полученных кодов LDPC разной длины и HARQ [102]. Здесь длительность n_3 сообщения принимается равной 6 временным слотам, причем один временной слот соответствует длине одного OFDM символа. Предположим, что n_0 – это количество временных слотов, необходимых для распространения сигнала. Алгоритм предсказания e-HARQ использует n_1 OFDM символов для последующих их обработки и принятия ACK/NACK решения, и выдает результат на БС через $2n_0 + n_1 + n_2$ временных слотов, где n_2 – время, затрачиваемое на обработку принятых символов и формирование предсказания. В свою очередь, обратная связь на базе механизма HARQ может быть получена через $2n_0 + n_3 + n_4$ временных слотов, где n_4 – время, затрачиваемое на обработку всего сообщения механизмом HARQ.

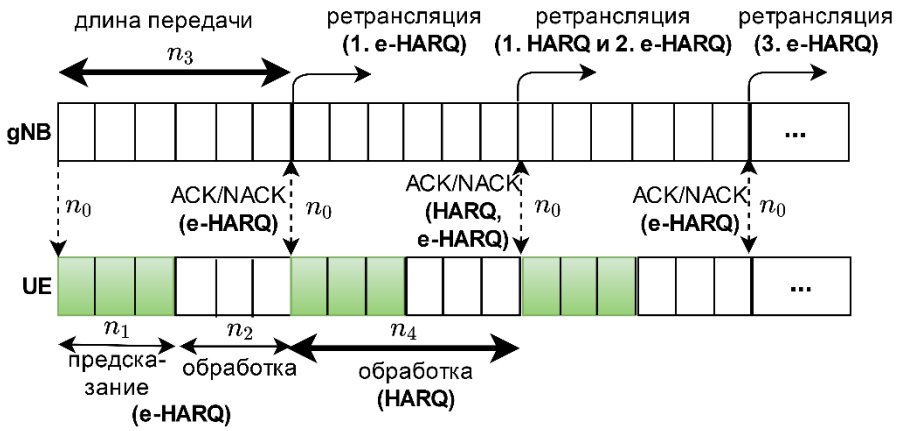


Рис.3.1. Сравнительная временная диаграмма между схемами предсказания e-HARQ на базе 3 OFDM символов

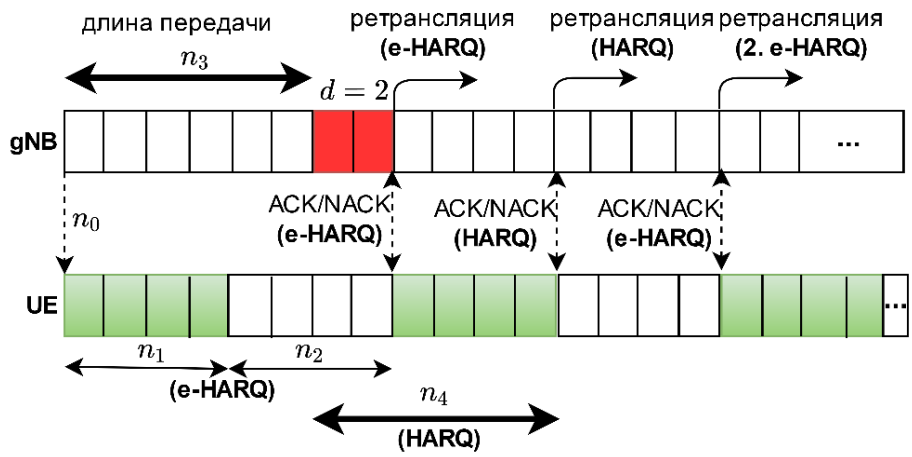


Рис.3.2. Сравнительная временная диаграмма между схемами предсказания e-HARQ на базе 4 OFDM символов

Масштабируемость времени обработки учитывается за счет предположения равенств $n_1 = n_2$ и $n_3 = n_4$, в то время как несущественное время распространения сигнала для сервисов URLLC – за счет предположения $n_0 = 0$, что является адекватным согласно [100]. Из рис.3.1,3.2 видно, что схема предсказания e-HARQ позволяет гораздо быстрее отправить сообщение обратной связи и инициировать повторную передачу по сравнению с традиционным механизмом HARQ. Отличием двух примеров (рис.3.1,3.2) является разная длина полученных LDPC кодов, равная 3 и 4 OFDM символам, соответственно, при неизменной длине сообщения. При этом, в первом случае $n_1 + n_2 = n_3$, и следовательно, в момент времени n_3 на БС уже имеется полученное от предиктора ACK/NACK сообщение, и может быть начат процесс ретрансляции.

На рис.3.2 предсказание осуществляется на базе кода длиной, равной 4 OFDM символам, что соответствует $n_1 + n_2 = n_3 + d$, где $0 \leq d < N$ – временной сдвиг относительно окончания передачи сообщения n_3, измеряемый в OFDM символах, а N – фиксированное число OFDM символов, необходимых для передачи пакета. С увеличением значения d очевидно, растет точность прогнозирования механизма e-HARQ, что приводит к сокращению числа возможных ретрансляций. Следует отметить, что предсказание e- HARQ может быть сделано также на базе полученных LDPC кодов длины $n_1 + n_2 < n_3$, однако данный случай не будет

рассматриваться из-за высокой ошибки предиктора и сложности аналитического моделирования.

Далее будет предложена аналитическая модель с учетом механизма ранней адаптации канала в сети 5G на основе предсказания повторной передачи с целью поиска оптимального порогового значения, получены решение для аналитической модели и ее основные характеристики. В разделе 3.2 вводится двухфазная СМО в дискретном времени с управляемым цепью Маркова поступающим потоком и обслуживанием, имеющие геометрическое распределение. В разделе 3.3 представлены система уравнений равновесия для двухфазной модели и ее решение, в то время как в разделе 3.4 приведен анализ основных ВВХ, полученных на базе вероятностного распределения.

3.2. Построение модели двухфазной СМО в дискретном времени с учетом итеративного моделирования механизма предсказания повторной передачи и механизма обратной связи

Рассмотрим функционирование соты сети 5G, в которой происходит передача пакета от БС к одному UE. При этом процесс прогнозирования на базе e-HARQ происходит на первой фазе, в то время как обработка пакета на базе традиционного механизма обратной связи HARQ – на второй. В модели используется понятие заявки, имеющей физическое значение пакета. Будем полагать, что пакеты имеют одинаковую длину, равную длительности N OFDM символов, или микрослотов. С учетом всех ретрансляций, необходимых для успешной доставки пакета, заявка должна быть успешно получена на UE за T_N микрослотов, иначе она считается потерянной.

Будем рассматривать функционирование системы в дискретном времени с тактом h постоянной длины, равным длительности одного микрослота, или одного OFDM символа, что позволяет учесть дискретный характер физических процессов передачи сообщения в сети 5G и добиться правомерного сравнения численных результатов для случаев с разными значениями N, d. Разделим ось времени на такты h и примем, что все

изменения в системе происходят лишь в моменты $nh, n = 1,2, \ldots$. Для определенности будем считать, как и ранее в диссертации, что такт n есть полуинтервал $[nh, (n+1)h)$.

Заявка поступает в систему с вероятностью $a, 0 < a \leq 1$, и, пока она находится в СМО, поступление новой заявки не происходит. Если предиктор принял положительное решение ACK на первой фазе, что соответствует событию $\{\text{eHARQ} = \text{ACK}\}$, то заявка переходит на вторую фазу с вероятностью $b_1(l)$, где l – текущее число попыток передачи данного пакета на UE. В противном случае, если решение предиктора, отправленное на БС – NACK, т.е. событие $\{\text{eHARQ} = N\text{ACK}\}$, заявка будет ретранслирована с дополнительной вероятностью $\bar{b}_1(l)$. Однако, обработка предыдущей заявки, которой предиктор вынес отрицательное заключение, продолжается на второй фазе с детерминированной вероятностью, равной 1. Таким образом, моделируются процедуры параллельной работы механизма предсказания и механизма обратной связи без предсказания, отдавая при этом приоритет решениям традиционного протокола HARQ во избежание лишних ретрансляций, сгенерированных в результате ошибок предиктора.

В свою очередь, решение об успешности передачи, принятое механизмом без предсказания, позволяет освободить систему с вероятностью $b_i(l)$, или продолжить процесс ретрансляции в случае отрицательного решения с дополнительной вероятностью $\bar{b}_i(l), i = 2,3$.

Вероятности $b_2(l)$ и $b_3(l)$ успешного обслуживания на второй фазе могут быть представлены в следующем виде:

$b_2(l) \coloneqq P(\{\text{HARQ} = \text{ACK}\}|\{\text{eHARQ} = \text{ACK}\})$,

$b_3(l) \coloneqq P(\{\text{HARQ} = \text{ACK}\}|\{\text{eHARQ} = \text{NACK}\})$,

где $\{\text{HARQ} = \text{ACK}\}$ – событие, при котором механизм обратной связи на второй фазе принимает положительное решение ACK.

Устанавливая зависимость между ACK/NACK решениями предиктора на первой фазе и механизма обратной связи на второй фазе за счет введения различных вероятностей перехода $b_2(l)$ и $b_3(l)$, мы получаем системную модель, которую можно использовать для анализа существующих схем

предсказания e-HARQ. Следует учесть, что число передач l данного пакета имеет непосредственное влияние на вероятность его обслуживания $b_i(l), i = 2,3$, и растет с увеличением l. Если же отправка пакета на UE превышает максимальное число микрослотов T_N или максимальное число возможных передач $T_c = \left\lfloor \frac{T_N}{N} \right\rfloor$, пакет считается потерянным и покидает систему.

На рис.3.3 показана структура рассматриваемой двухфазной СМО с прогнозированием на первой фазе и механизмом обратной связи без предсказания на второй фазе, соответственно.

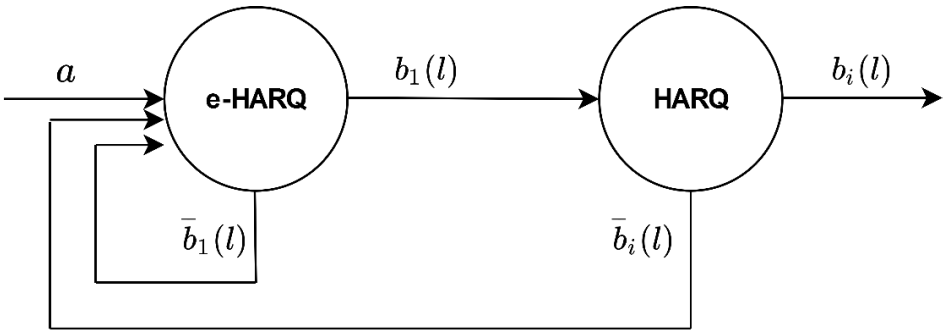


Рис.3.3. Структура двухфазной модели, описывающей процесс прогнозирования (e-HARQ) на фазе 1 и процесс обработки пакета на терминале UE на базе HARQ на фазе 2, $i = 2,3$

3.3. Система уравнений равновесия и основные вероятностно-временные характеристики

Поведение СМО описывается однородной ЦМ ξ_n по моментам $nh + 0, n \geq 0,$ над пространством состояний $X = X_m^* \cup X_t$, где X_m^* определяет множество основных состояний системы, в которых происходят главные активные события, например, выход из системы, переход на другую фазу, и другие:

$$X_m^* = \left\{(0,0,0,0)^*, (1,1,1,t)^*, (s,l,v,t)^*: s \in \{1,2,3\}, l = \left\lceil \frac{v}{2} \right\rceil + u(s,v), \dots, v-1, \atop v = 2, \dots, T_c, t \le T_N \right\}, \quad (3.1)$$

где v – номер временного слота, и $s = 0,1,2,3$ –параметр, отвечающий за процедуру обработки на фазах,

$$u(s,v) = \begin{cases} 1, \text{если } \{s=1\} \cap \{v = 2n, n = 2,3, \dots\}, \\ 0, \text{в противном случае,} \end{cases}$$

$$t(l,v) = \begin{cases} vN + 2d(l - \left\lceil \frac{v}{2} \right\rceil), \text{если } v = 2n, n = 1,2, \dots, \\ vN + d(2\left(l - \left\lceil \frac{v}{2} \right\rceil\right) + 1), \text{в противном случае.} \end{cases}$$

Однако, так как функционирование модели рассматривается в дискретном времени с тактом h постоянной длины, равным длительности одного микрослота, или одного OFDM символа, необходимо определить множество переходных состояний X_t, позволяющих попасть в основные состояния системы:

$$X_t = \{\{(s-1, l-1, v-1, t-j): s = 1\}, \{\gamma(l,v)(s, l, v-1, t-i): s \in \{2,3\}\}, \{\delta(l,v)(s+2, l-1, v-1, t-j): s \in \{2,3\}\}, l, v, t \in X_m^*, i = 1, \dots, N - d - 1, j = 1, \dots, N + d - 1\}. \quad (3.2)$$

Здесь,

$$\gamma(l,v) = \begin{cases} 1, \text{если } \{l < v-1\} \cup \{v = 2\}, \\ 0, \text{в противном случае,} \end{cases}$$

$$\delta(l,v) = \begin{cases} 1, \text{если } \left\{l > \left\lceil \frac{v}{2} \right\rceil, v = 2n\right\} \cup \{v = 2n+1\}, n = 1,2, \dots \\ 0, \text{в противном случае.} \end{cases}$$

Как видно из (3.1), (3.2) мы используем два временных параметра: v – номер временного слота, равного длительности передачи пакета, или N OFDM символам, и t – номер микрослота, равного одному OFDM символу. Использование двух данных параметров позволяет упростить логику математических выражений и дифференциацию между состояниями системы.

В таблице 3.1 определен параметр s, отвечающий за процедуру обработки на фазах, причем положение единицы в бинарной записи соответствует функционированию данной фазы, и соответственно e-HARQ

и/или HARQ механизмов. Следует отметить, что описание функционирования фаз включает только основные состояния системы X_m^*, в которых происходят активные события. Однако, путь переходов дает понять, какие и сколько переходных состояний X_t должна пройти заявка, чтобы достичь состояния X_m^*: например, чтобы попасть в состояние фазы $s = 1$, в котором происходит прогнозирование сообщения ACK/NACK, заявка должна пройти $N + d - 1$ переходных состояний, в которых $s = 0$.

Таблица 3.1. Значения параметра s для основных состояний системы и путь перехода к ним

s	Описание	Количество состояний из X_t
0 (0,0)	Система в состоянии простоя	
1 (1,0)	Пакет частично получен на UE и вынесено ACK/NACK решение с помощью механизма e-HARQ на первой фазе	$s = 0, (N + d - 1)$
2 (0,1)	Пакет обработан на второй фазе механизмом HARQ вследствие ACK решения по данному пакету, вынесенному предиктором на фазе 1	$s = 2, (N - d - 1)$ $s = 4, (N + d - 1)$
3 (1,1)	Пакет ретранслируется из-за NACK решения на первой фазе, и предыдущая передача обрабатывается механизмом HARQ на второй фазе	$s = 3, (N - d - 1)$ $s = 5, (N + d - 1)$

Граф вероятностей переходов между основными состояниями системы представлен на рис.3.4. Следует учесть, что случай $b_1(l) = 1$ исключает работу предиктора e-HARQ, а значит позволяет промоделировать функционирование традиционного механизма обратной связи HARQ.

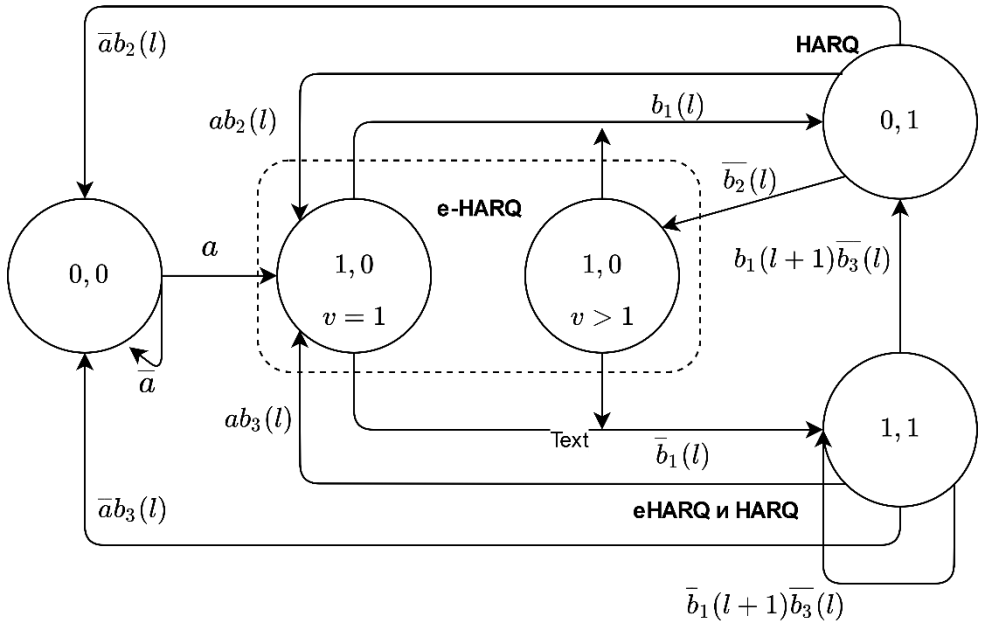


Рис.3.4 Граф вероятностей переходов между основными состояниями системы, X_m^*

При $0 < a \leq 1$ ЦМ ξ_n, $n \geq 0$, на пространстве состояния X – неразложима и апериодична, поэтому стационарное распределение вероятностей $[s, l, v, t]$, $(s, l, v, t) \in X$ существует и находится из СУР.

Утверждение 3.1. *Система уравнений равновесия для цепи Маркова* $\xi_n, n \geq 0$ *имеет следующий вид:*

$$a[0,0,0,0]^* = \bar{a}\left(\sum_{s=2}^{3} \sum_{(s,l,v,t)\in X_m^* \setminus X_{\text{end}}} b_s(l)[s,l,v,t]^* + \sum_{(s,l,v,t)\in X_{\text{end}}} [s,l,v,t]^*\right),$$

$$[0,0,0,1]^* = a\Bigg([0,0,0,0] + \sum_{s=2}^{3} \sum_{(s,l,v,t)\in X_m^* \setminus X_{\text{end}}} b_s(l)[s,l,v,t]^* + \sum_{(s,l,v,t)\in X_{\text{end}}} [s,l,v,t]^*\Bigg),$$

$$[1,1,1,t]^* = [0,0,0,t-1] = \cdots = [0,0,0,1],$$

$$[1,l,v,t]^* = [0,l-1,v-1,t-1] = [0,l-1,v-1,t-2] = \cdots = [0,l-1,v-1,t-(N+d-1)] = \bar{b}_2(l-1)[2,l-1,v-1,t-(N+d)]^*, v = 3,4,\ldots,T_c,$$

$$[s,l,v,t]^* = \gamma(l,v)[s,l,v-1,t-1] + \delta(l,v)[s+2,l-1,v-1,t-1],$$

где $s = 2,3$, *и*

$$[2,l,v-1,t-1] = [2,l,v-1,t-2] = \cdots = [2,l,v-1,t-(N-d-1)] = b_1(l)[1,l,v-1,t-(N-d)]^*,$$

$$[3,l,v-1,t-1] = [3,l,v-1,t-2] = \cdots = [3,l,v-1,t-(N-d-1)] = \bar{b}_1(l)[1,l,v-1,t-(N-d)]^*,$$

$$[4,l-1,v-1,t-1] = [4,l-1,v-1,t-2] = \cdots = [4,l-1,v-1,t-(N+d-1)] = b_1(l)\bar{b}_3(l-1)\delta(l,v)[3,l-1,v-1,t-(N+d)]^*,$$

$$[5,l-1,v-1,t-1] = [5,l-1,v-1,t-2] = \cdots = [5,l-1,v-1,t-(N+d-1)] = \bar{b}_1(l)\bar{b}_3(l-1)\delta(l,v)[3,l-1,v-1,t-(N+d)]^*,$$

где $(s,l,v,t) \in X$.

Далее введем множество состояний выхода из системы X_{end}, включающее в себя состояния, которые достигли граничного значения T_N:

$$X_{\text{end}} = X_{\text{end}}(1) \cup X_{\text{end}}(2) \cup X_{\text{end}}(3) \subset X_m^*,$$

$$X_{\text{end}}(1) = \{\mu(t(l,v+1) - T_N) * (1,l,v,t(l,v))^*\},$$

$$X_{\text{end}}(2) = \{\mu(t(l+1,v+1) - T_N) * (2,l,v,t(l,v))^*\},$$

$$X_{\text{end}}(3) = \{\mu(t(l+1,v+1) - T_N) * (3,l,v,t(l,v))^*\},$$

$$\mu(x) = \begin{cases} 1, \text{если x} > 0, \\ 0, \text{в противном случае.} \end{cases}$$

Нормировочное условие имеет вид:

$$\sum_{(s,l,v,t)\in X} [s,l,v,t] = 1.$$

СМО, моделирующие подобные процессы в дискретном времени, достаточны сложны в определении пространства состояний. Мощность такого пространства состояний напрямую влияет на вычислительную сложность нахождения стационарного распределения вероятностей.

Определим мощность пространства состояний данной модели как $|X| = |X_m^*| + |X_t|$,

где $|X_m^*|$- общее число главных состояний системы:

$$|X_m^*| = |(0,0,0,0)| + |(1,1,1,t(1,1))| + \sum_{v=3}^{\left\lfloor \frac{T_N}{N} \right\rfloor} \sum_{l=\left\lceil \frac{v}{2} \right\rceil + u(1,v)}^{v-1} (1 - \mu(t(l, v+1) - T_N)) \, |(1, l, v, t(l, v))| + \sum_{s=2}^{3} \sum_{v=2}^{\left\lfloor \frac{T_N}{N} \right\rfloor} \sum_{l=\left\lceil \frac{v}{2} \right\rceil}^{v-1} (1 - \mu(t(l+1, v+1) - T_N)) \, |(s, l, v, t(l, v))| ,$$

и $|X_t|$ - общее число переходных состояний системы:

$$|X_t| = \sum_{(1,l,v,t) \in X_m^*} \sum_{j=1}^{N+d-1} |(0, l-1, v-1, t-j)| + \sum_{s=2}^{3} \sum_{(s,l,v,t) \in X_m^*} \left(\sum_{i=1}^{N-d-1} \gamma(l, v) \, |(s, l, v-1, t-i)| + \sum_{j=1}^{N+d-1} \delta(l, v) \, |(s+2, l-1, v-1, t-j)| \right).$$

Например, для случая $N = 4, d = 1$ и $T_N = 28$ мощность пространства состояний составляет 122 состояния, из которых 26 – главных и 96 переходных.

Стационарное распределение вероятностей с нормировочным условием позволяет получить формулы для основных ВВХ. Для этого введем X_{quit} – множество состояний системы, из которых заявка может покинуть систему как в случае успеха, так и в случае, когда превышено граничное значение T_N:

$$X_{\text{quit}} := \left\{ (s, l, v, t)^* : s \in \{2,3\}, l = \overline{\left\lceil \frac{v}{2} \right\rceil, v-1}, v = \overline{2, T_c}, t \le T_N \right\} \cup \{(1, l, T_c, t) : l = \overline{\left\lceil \frac{T_c}{2} \right\rceil + u(1, T_c), T_c - 1}, t \le T_N\}.$$

<u>Утверждение 3.2.</u> *Вероятность π неуспешного обслуживания заявки вычисляется по формуле:*

$$\pi = \sum_{x \in X_{quit}} P_{quit}(x)(1 - B(x)),$$

где $B(x), x \in X_{quit}$ находится следующим образом:

$$B(x) = \begin{cases} 0, \text{если } s_x = 1 \\ b_2(l_x), \text{если } s_x = 2 \text{ и } x \in X_{end}(2) \\ b_3(l_x), \text{если } s_x = 3 \text{ и } x \in X_{end}(3) \\ 1, \text{в противном случае} \end{cases},$$

и $P_{quit}: X_{quit} \to [0,1]$ *условная вероятность выхода заявки из системы в состоянии* $x \in X_{quit}$ *в случае успеха или потери вычисляется по формуле:*

$$P_{quit}(x) = \frac{b_{quit}(x)[x]}{\sum_{x \in X_{quit}} b_{quit}(x)[x]},$$

где

$$b_{quit}(x) = \begin{cases} b_2(l_x), \text{если } s_x = 2 \text{ и } x \notin X_{end} \\ b_3(l_x), \text{если } s_x = 3 \text{ и } x \notin X_{end} \\ 1, \text{в противном случае} \end{cases}.$$

Среднее число передач заявки можно посчитать аналогичным образом:

$$M = \sum_{x \in X_{\text{quit}}} P_{\text{quit}}(x) L(x),$$

где $L(x)$ – число передач в состоянии системы x вычисляется по формуле:

$$L(x) := \begin{cases} l_x - 1, \text{если } s_x = 1 \\ l_x, \text{если } s_x = 2 \text{ или } s_x = 3 \end{cases}.$$

Среднее время пребывания T заявки в системе до момента успешной передачи находится по формуле:

$$T = \sum_{x \in X_{\text{quit}}} P_{\text{success}}(x) v_x,$$

где $P_{\text{success}}(x): X_{\text{quit}} \to [0,1]$ определяется аналогичным способом, как и $P_{\text{quit}}(x)$, при условии, что $b_{\text{quit}}(x)$ заменяется на $b_{\text{success}}(x)$:

$$b_{\text{success}}(x) = \begin{cases} b_2(l_x), \text{если } s_x = 2 \\ b_3(l_x), \text{если } s_x = 3 \\ 0, \text{в противном случае} \end{cases}.$$

Вероятность нахождения системы в состоянии простоя равна:

$$P_0 = [0,0,0,0].$$

Ошибки предсказания e-HARQ классифицируются на FN (False Negative) в случае, когда решение предиктора e-HARQ – NACK, в то время как механизмом HARQ вынесено решение ACK о подтверждении данного пакета; и FP,(False Positive) в случае, когда предиктор e-HARQ предсказал ACK, а механизм HARQ вынес NACK решение об ошибке. Ошибочно-отрицательные предсказания предиктора приводят к ненужным

ретрансляциям, ухудшающим спектральную эффективность канала, но не оказывают влияния на задержку и отношение числа ошибочных блоков к общему числу блоков (BLER, Block Error Rate), передаваемых в сети. Таким образом, ошибочно-отрицательные решения предиктора могут быть допустимыми до определенного предела. Ошибочно-положительные предсказания соответствуют непосредственным ошибкам предиктора. Большинство известных схем предсказания e-HARQ используют фиксированное пороговое значение θ в своих алгоритмах в качестве способа классификации между принятием ACK/NACK решения, поэтому выбор данного значения θ существенным образом влияет на эффективность схемы предсказания. Данная аналитическая модель позволять получить формулы в виде BBX на базе стационарного распределения для FN и FP вероятностей, соответственно:

$$P_{\text{FN}} = \sum_{v=2}^{T_c} \sum_{l=\frac{\lceil v \rceil}{2}}^{v-1} b_3(l)[3, l, v, t],$$

$$P_{\text{FP}} = \sum_{v=2}^{T_c} \sum_{l=\frac{\lceil v \rceil}{2}}^{v-1} \bar{b}_2(l)[2, l, v, t].$$

3.4. Численный анализ

В данном разделе представлен численный анализ полученных BBX. На первом этапе были найдены вероятности переходов $b_1(l), b_2(l)$ и $b_3(l)$ за счет канального моделирования сети 5G. Получение реалистичных значений для вероятностей переходов позволяет использовать данную аналитическую модель для анализа и оптимизации существующих схем предсказания e- HARQ. Далее был разработан имитационный комплекс для проверки корректности аналитической модели, и решена оптимизационная задача поиска оптимальных параметров: длины сообщения, качества предсказания, а также порогового значения, при которых наблюдается наименьшая длительность успешного обслуживания заявки при условии, что ошибка предсказания не превышает допустимых значений.

Имитационное моделирование сети 5G

Для проведения численного анализа предложенной аналитической модели, были найдены вероятности переходов $b_1(l), b_2(l)$ и $b_3(l)$ за счет канального моделирования сети 5G. Основные исходные данные данной симуляции приведены в таблице 3.2, в соответствии с которыми пакет размером 500 бит был закодирован посредством схемы LDPC, широко применяемой в технологии 5G, и назначен на нисходящую OFDM-передачу в канале с пропускной способностью, равной 1.08 MHz, и модуляцией 64-QAM. Данный закодированный сигнал был передан по каналу с замираниями TDL-C, и обработан на приемной станции с помощью процедуры (MMSE, Minimum Mean Square Error) в частотной области, после чего декодирован посредством канального декодера LDPC (Min-Sum).

Более того, в симуляции рассматривался предиктор e-HARQ на базе логических регрессий (LR, Logistic Regressions) [150], принимающий решение о декодируемости пакета с учетом 5 итераций LDPC декодера.

Таблица 3.2. Исходные данные для имитационного моделирования сети 5G

Размер транспортного блока (число бит)	500
Пропускная способность	1.08 MHz (6 РБ)
Схема канального кодирования	1/5 LDPC [130]
Алгоритм и порядок модуляции	64 QAM, Approx. LLR
Распределение мощности	Постоянное значение $\frac{E_b}{N_0}$
Форма волны (waveform)	3GPP OFDM с нормальным циклическим префиксом и разнесением поднесущих 15kHz
Тип канала	1Tx1Rx, TDL-C 100 нс, 7 ГГц, 3 км/ч (скорость пользователя)

Эквалайзер	MMSE в частотной области
SNR	5.0 dB – 12.0 dB
Тип декодера	Min-Sum (50 итераций)
Тип предиктора (e-HARQ)	Предиктор на базе логических регрессий (LR) [150] с 5 итерациями

Следует отметить, что показатели эффективности предиктора LR напрямую зависят от значения переходной вероятности $b_1(l)$. При малых значениях вероятности перехода на первой фазе $b_1(l)$, предиктор показывает более консервативное поведение с точки зрения принятых ACK решений, в то время как при высоких значениях $b_1(l)$, предиктор имеет тенденцию слишком частого предсказания в пользу ACK. В связи с этим, на первом этапе предиктор был настроен таким образом, чтобы сперва было достигнуто некоторое значение $b_1(l)$, а далее найдены вероятности переходов $b_2(l)$ и $b_3(l)$. Для этого были проведены симуляции передачи пакета на канальном уровне с учетом вышеизложенных исходных данных на базе метода Монте-Карло с 1.7 млн итераций. Отношение сигнал-шум (SNR, Signal-to-Noise Ratio) варьировался между 5dB и 12dB, и были рассмотрены различные длины LDPC кодов.

Чтобы довести время моделирования и процессов обработки данных до трактабельных, вероятности перехода были аппроксимированы следующим образом:

$$b_2(l) = P(\{\mathrm{HARQ}_l = \mathrm{ACK}\}|\{\mathrm{eHARQ}_l = \mathrm{ACK}, \mathrm{HARQ}_{l-1} = \mathrm{NACK}\}) \approx 1 - \frac{P(\{\mathrm{HARQ}_l = N\mathrm{ACK}\}|\{\mathrm{eHARQ}_l = \mathrm{ACK}\})}{P(\{\mathrm{HARQ}_{l-1} = N\mathrm{ACK}\}|\{\mathrm{eHARQ}_{l-1} = \mathrm{ACK}\})},$$

$$b_3(l) = P(\{\mathrm{HARQ}_l = \mathrm{ACK}\}|\{\mathrm{eHARQ}_l = N\mathrm{ACK}, \mathrm{HARQ}_{l-1} = \mathrm{NACK}\}) \approx 1 - \frac{P(\{\mathrm{HARQ}_l = N\mathrm{ACK}\}|\{\mathrm{eHARQ}_l = N\mathrm{ACK}\})}{P(\{\mathrm{HARQ}_{l-1} = N\mathrm{ACK}\}|\{\mathrm{eHARQ}_{l-1} = N\mathrm{ACK}\})}.$$

Значения вероятностей перехода $b_1(l)$ для первого порогового значения $\theta = 0$ были получены по формуле:

$$b_1(l) = P(\{\mathrm{HARQ}_l = \mathrm{ACK}\}|\{\mathrm{HARQ}_{l-1} = \mathrm{NACK}\}),$$

и далее увеличивались на 0.1 для каждой последующей трансляции. Всего рассмотрено 10 пороговых значений, последнее из которых ($\theta = 9$) соответствует передаче пакета на базе традиционного механизма HARQ с отсутствием e-HARQ прогнозирования на первой фазе, $b_1(l) = 1$.

На рис. 3.5 приведен пример вероятностей перехода, полученных в результате имитационного моделирования сети 5G на уровне канала для случая SNR=10 dB и длительности пакета, равной 14 OFDM символам. Данные значения переходных вероятностей будут использованы в следующем подразделе для проверки корректности модели. Графики на рис.3.5 демонстрируют вероятности переходов только для первой передачи ($l = 1$), однако, показывают поведение для всех длин кодов с соответствующим смещением $d = 0,2, \dots ,12$, используемых для e-HARQ прогнозирования.

С увеличением порогового значения θ, мы видим монотонный рост вероятности $b_1(l)$ до 1, начиная от нуля для случая небольших длин кодов, и 0.5 для длинного кода ($d = 12$), что объясняется его высокой точностью предсказания и заданным уровнем SNR. Вероятности обслуживания на второй фазе $b_2(l)$ и $b_3(l)$ очевидным образом имеют высокие значения при частых ретрансляциях в случае низких пороговых значений, и уменьшаются с увеличением θ, приближаясь к результатам традиционной HARQ схемы.

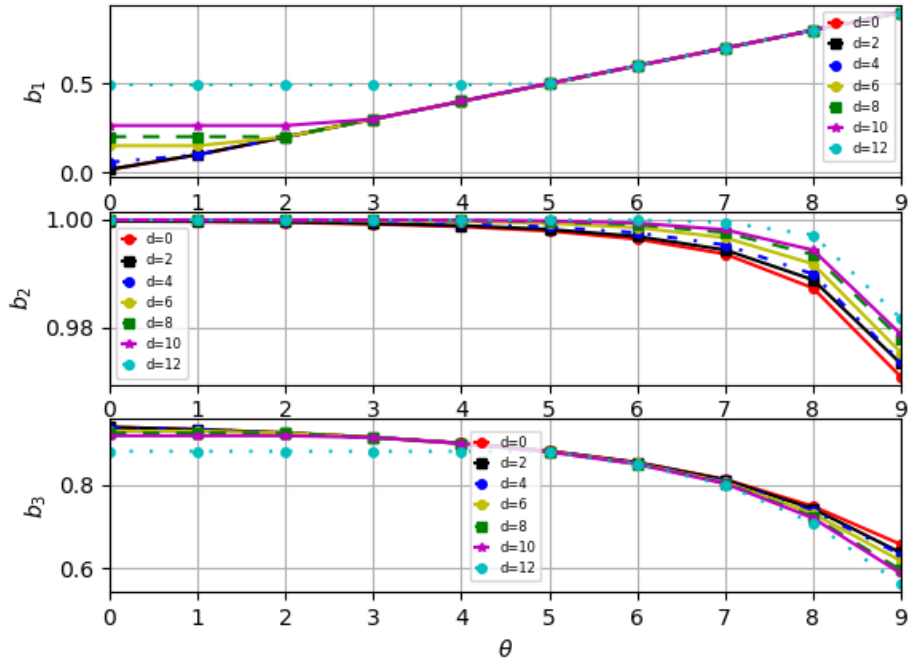


Рис.3.5. Вероятности переходов $b_1(l), b_2(l)$ и $b_3(l)$, найденные в результате имитационного моделирования сети 5G

Проверка корректности модели

Для проверки корректности аналитической модели был проведен имитационный анализ, в ходе которого был смоделирован процесс перехода пакета по состояниям модели до его успешного обслуживания или потери с учетом заданного времени функционирования. Данный процесс продолжался до момента сходимости распределения вероятностей. Имитационное моделирование позволило собрать основные статистические данные о количестве обслуженных/потерянных пакетов, времени пребывания и др., на базе которых были найдены характеристики, соответствующие основным ВВХ аналитической модели.

В данном эксперименте были использованы вероятности переходов $b_1(l), b_2(l)$ и $b_3(l)$, представленные на рис.3.5, не меняющие значений с увеличением числа ретрансляций. Вероятность поступления в систему пакета длительностью 14 OFDM символов в течение микрослота принята равной 0.2, что соответствует типам услуг с постоянным потоком данных.

Программное обеспечение для двух моделей было разработано на языке программирования PYTHON.

Рис.3.6 демонстрирует графики для среднего времени пребывания заявки в системе до успешной передачи, измеряемого в микрослотах, для различных длин кодов с соответствующими смещениями относительно окончания передачи пакета $d = 0,2, \dots ,12$ для аналитической и имитационной моделей. Следует отметить, что процесс имитации данной модели, функционирующей в микрослотах, очень ресурснозатратный, и требует гораздо больше вычислительных усилий в сравнении с аналитической моделью, основные ВВХ которой вычисляются мгновенно даже для высоких значений временных ограничений T_N и длительностей пакетов N. Поэтому, были получены результаты только для нескольких пороговых значений θ, что, однако, является достаточным, чтобы сделать вывод о совпадении результатов двух исследуемых подходов.

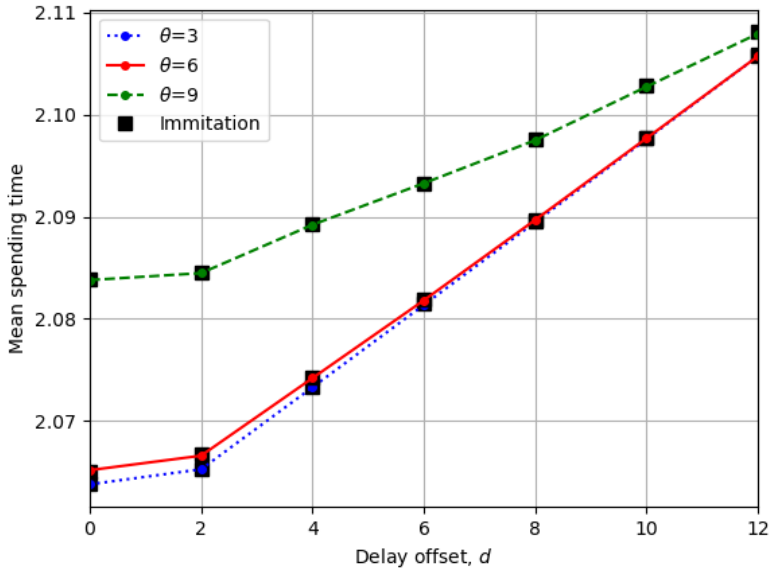


Рис.3.6. Графики сравнения среднего времени пребывания для аналитической и имитационной моделей

Таким образом, в дальнейшем мы будем проводить численный анализ только для аналитической модели, предполагая ее математическую корректность.

Анализ основных показателей эффективности модели

Был проведен обширный сравнительный анализ полученных характеристик аналитической модели с различными значениями SNR (5.0 dB–12.0 dB), длинами передачи пакета $N = 4,5,\dots,12$ OFDM символам, длинами кодов с соответствующими сдвигами $d = 0,2,\dots,12$ относительно окончания первой передачи и пороговыми значениями $\theta = 0,\dots,9$. Следует отметить, что тенденции поведения основных ВВХ сохраняются для исследованных условий качества канала, поэтому будут приведены в качестве примера только для случая SNR=10.0 dB. Ограничение времени T_N, выделенного на успешную доставку пакета, выбрано равным 84 микрослотам, что соответствует различным вариантам максимального числа трансляций данного пакета $T_c = \left\lfloor \frac{T_N}{N} \right\rfloor$ для разных длин сообщений. Рис.3.7 демонстрирует семейство графиков зависимости вероятности простоя системы и среднего времени пребывания заявки в системе от изменения порогового значения θ, соответственно. На графиках используются следующие обозначения: N, d, где длина сообщения N отмечена цветом, а временной сдвиг d, соответствующий различным длинам LDPC кодов показан маркером. Таким образом, мы можем одновременно сравнить поведение всех возможных длин LDPC кодов, используемых для прогнозирования обратной связи для различных длин передачи пакета.

Вероятность простоя системы является одной из важнейших характеристик, которая косвенным образом демонстрирует эффективность обслуживания пакета в системе. Чем быстрее пакет обслуживается, тем скорее система переходит в состояние $[0,0,0,0]$ и пребывает там до прихода нового пакета. Таким образом, чем выше эта вероятность, тем эффективнее система справляется с передачей пакета на пользовательскую станцию.

На рис.3.7 мы видим, что система с наименьшей длиной передачи $N = 4$ и с самой быстрой возможностью ретрансляции $d = 0$ показывает наилучшее поведение вероятности простоя с небольшим ухудшением характеристики при увеличении длины кода d, и соответственно RTT. Этот

результат может быть объяснен возможностью для систем с короткими передачами осуществить больше ретрансляций, чем в системах с длинными передачами за время T_N. При увеличении порогового значения θ вероятность простоя системы уменьшается, приближаясь к поведению системы с традиционным HARQ механизмом без прогнозирования.

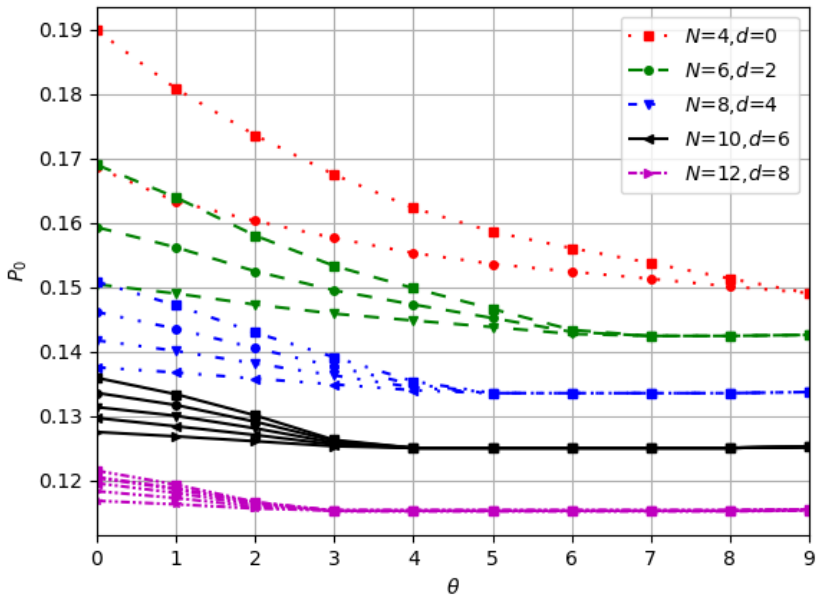


Рис.3.7. Семейство графиков зависимости вероятности простоя системы с параметрами N, d от изменения порогового значения θ

На рис.3.8 показано семейство графиков зависимости среднего времени пребывания заявки в системе, измеряемого в микрослотах, с параметрами N, d от изменения порогового значения θ. Тенденции поведения данных кривых не противоречат результатам вероятности простоя, и демонстрируют наименьшее время пребывания в системе до успешной передачи пакета в случае $N = 4, d = 0$ коротких передач с частыми ретрансляциями.

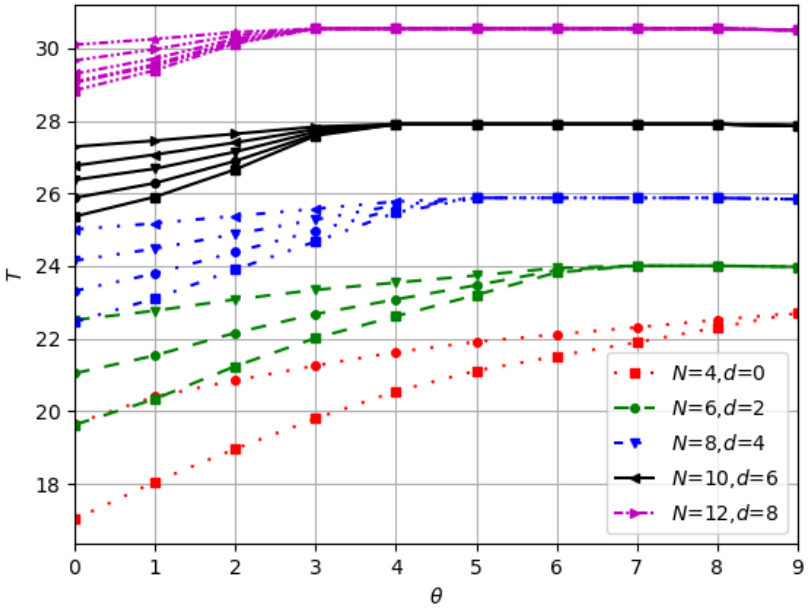


Рис.3.8. Семейство графиков зависимости среднего времени пребывания заявки в системе с параметрами N, d от изменения порогового значения θ

Однако, короткие передачи подвержены риску частых ошибок прогнозирования, поэтому наряду с тенденциями поведения основных ВВХ функционирования системы, важно также проанализировать вероятности FN и FP ошибок предиктора e-HARQ, представленные на рис.3.9 для случая $d = 0$ и различных длин передач N. Длинная передача N характеризуется более высокой точностью предсказания, а значит гораздо больше подвержена ошибочно-отрицательным решениям предиктора при низких пороговых значениях θ, и менее – ошибочно-положительным решениям предиктора при любых θ. В свою очередь, чем короче длина сообщения, тем выше вероятность FP – вероятность, которая непосредственным образом указывает на эффективность работы используемого предиктора e-HARQ. Таким образом, данные графики показывают важные тенденции, необходимые при проведении анализа, поиска оптимальных параметров сети и прогнозирования e-HARQ.

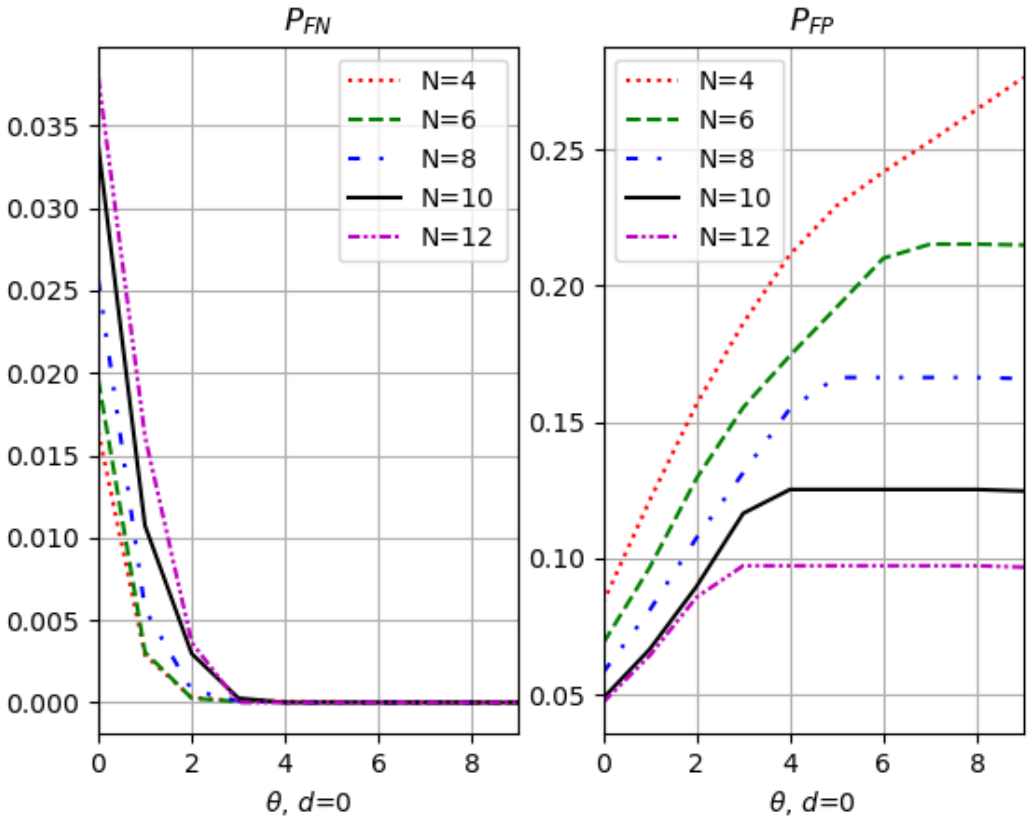


Рис.3.9. Графики зависимости вероятностей FN и FP для случая $d = 0$ и различных длин сообщений N от изменения порогового значения θ

Формализация задачи оптимизации и ее решение

Основной задачей данной аналитической модели является получение быстрой оценки основных ВВХ для поиска оптимальных параметров, например порогового значения θ, повышающих эффективность схемы предсказания e-HARQ. Принимая во внимание строгие требования к задержке и надежности передачи (BLER), характерные для услуг URLLC, и учитывая тенденции поведения основных ВВХ, изображенные на рис. 3.7-3.9, сформулируем задачу оптимизации следующим образом.

Утверждение 3.3. *В модели двухфазной СМО в дискретном времени, моделирующей механизм предсказания повторной передачи и механизм обратной связи, значение порогового параметра θ предсказания повторной передачи, может быть найдено путем решения задачи минимизации целевой функции затрат:*

$min(w_1 P_{FN}(\theta) + w_2 T^*(\theta))$,

с ограничениями $w_1 + w_2 = 1$ *и* $P_{FN}(\theta) < FN_{max}$, *где* w_1 *и* w_2 – *коэффициенты веса целевых функций, и* $T^* = \frac{T}{T_N}$ – *среднее время пребывания в системе до успешной передачи, нормированное относительно числа микрослотов* T_N. Задача решена методом прямого поиска.

Вместо T^* в качестве целевой функции также можно рассматривать вероятность π неуспешного обслуживания пакета, однако во время численного анализа было замечено, что при исследуемых условиях качества канала и максимального времени на передачу T_N, данная характеристика была близка к нулю на всей области значений. Рис.3.10 демонстрирует функции затрат $Cost(\theta)$ для различных значений N, d и качества канала SNR, соответственно. Рис.3.11 показывает зависимость изменения локального минимума функции затрат и соответствующего ему порогового значения от выбранного веса w_1 слева направо, соответственно. В общем случае, более низкому значению веса w_1 соответствует более высокое значение минимума функции затрат, что объясняется изначально более высокими значениями целевой функции $T^*(\theta)$.

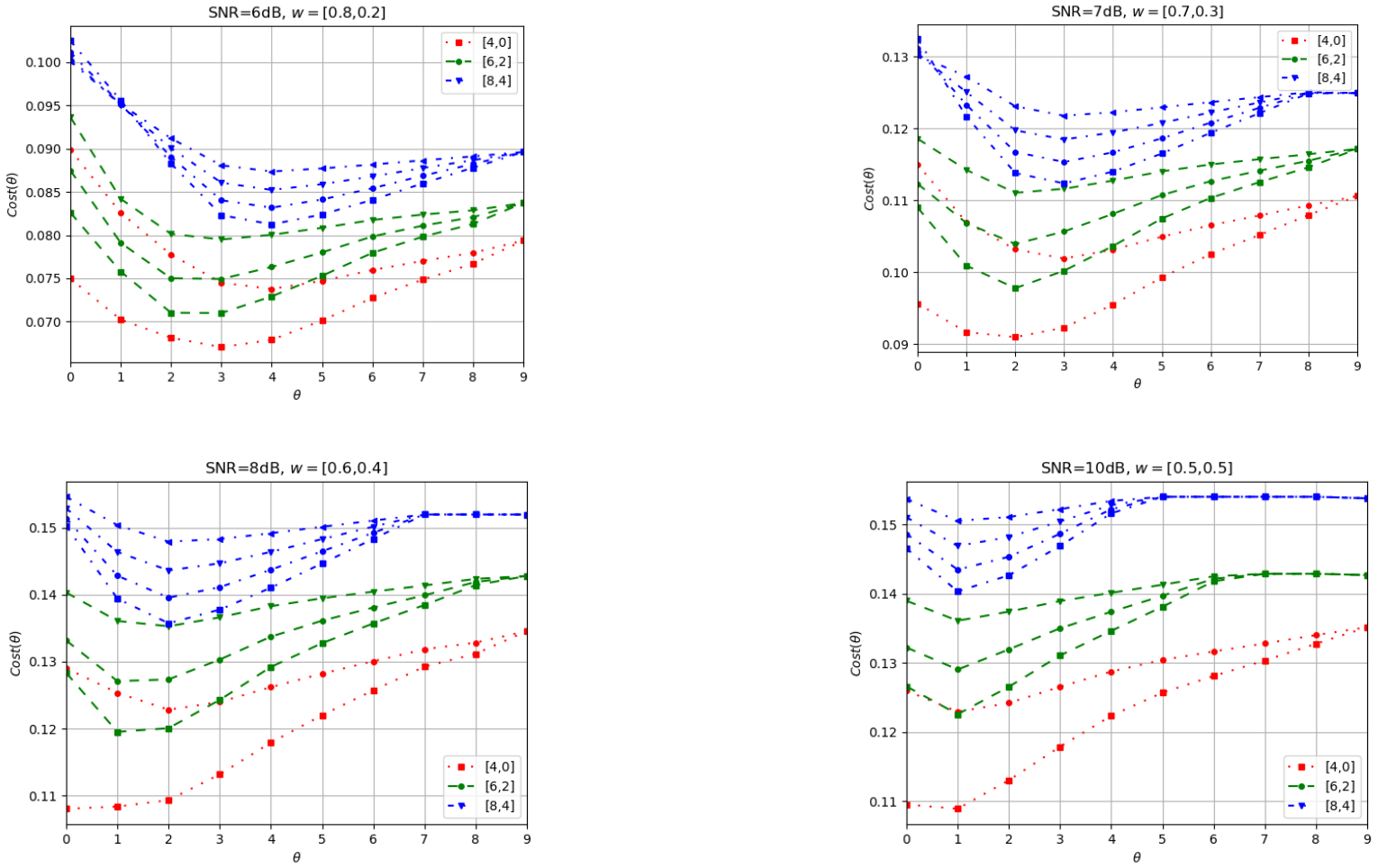


Рис..3.10. Функция затрат для различных уровней качества канала SNR и параметров N, d

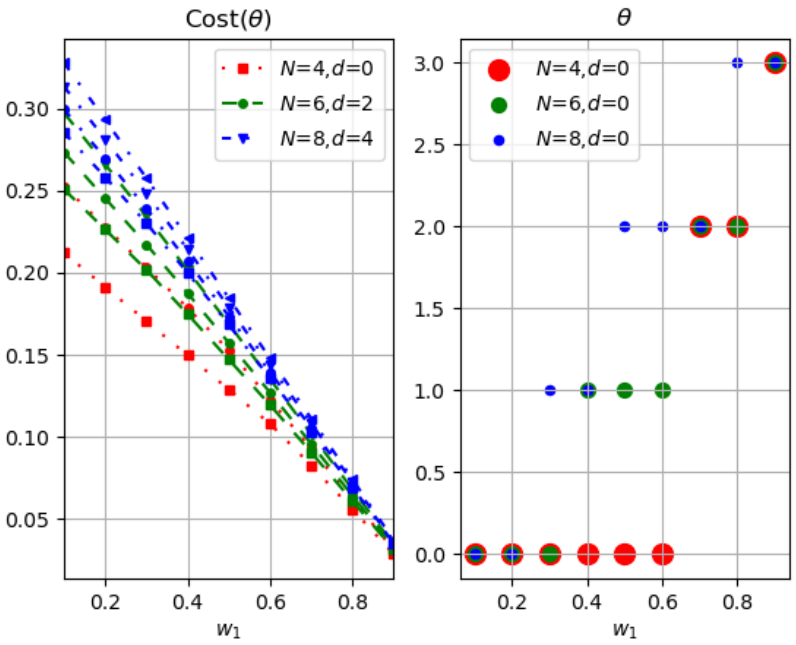


Рис.3.11. Зависимость изменения локального минимума функции затрат и соответствующего ему порогового значения от выбранного веса w_1 слева направо, соответственно.

Поэтому наблюдается монотонное снижение функции затрат $Cost(\theta)$ для семейства кривых с различными значениями N, d с увеличением веса w_1. При этом пороговое значение θ увеличивается, как показано на рис.3.11 (справа), что приводит к ухудшению основных ВВХ системы. Таким образом, данные значения параметра w_1 могут быть настроены в соответствии с заданными характеристиками сети. Как видно из рис.3.10 в плохих канальных условиях SNR=6dB имеет смысл принять $w_1 = 0.8$ для того, чтобы снизить ошибочно-отрицательную вероятность предиктора. При этом наблюдается смещение локального минимума влево по оси абсцисс с уменьшением веса w_1 при улучшении качества канала. При SNR=10dB функция затрат имеет свой минимум в $\theta = 1$ для коротких передач и $\theta = 2$ для более длинных передач $N = 8$ из-за довольно быстрого снижения вероятности FN к нулю. Важно отметить, что с увеличением длины кодов с соответствующим смещением относительно конца первой передачи d функция затрат растет.

Таким образом, при наличии основных параметров сети и схемы прогнозирования e-HARQ, на базе предложенной аналитической модели можно получить быструю и аккуратную оценку основных ВВХ и найти оптимальное пороговое значение θ для повышения эффективности прогнозирования e-HARQ.

Далее проведем оптимизационный анализ и найдем максимальный выигрыш, получаемый при использовании схемы с предсказанием в сравнении с традиционным механизмом HARQ при заданных ограничениях на FN, FP вероятности ошибок предиктора e-HARQ.

Утверждение 3.4. *Значения параметров длины пакета N, длины кода с соответствующим сдвигом d, порогового значения предсказания повторной передачи θ для обеспечения эффективной передачи пакета могут быть найдены путем решения задачи максимизации выигрыша, получаемого схемой с предсказанием в сравнении со схемой без предсказания:*

$max\ G(T(N,d,\theta),T_{HARQ}(N,d,\theta))$,

с ограничениями $P_{FN}(N,d,\theta) < FN_{max},\ \ P_{FP}(N,d,\theta) < FP_{max}$.

Здесь $G(x,y) = \frac{|x-y|}{y} * 100\%, FN_{max} = max(P(\{HARQ = ACK\}|\{eHARQ = NACK\}))$, $FP_{max} = max(P(\{HARQ = NACK\}|\{eHARQ = ACK\}))$. Задача решена методом прямого поиска.

В первую очередь, установим ограничение относительно вероятности FN, и найдем минимальное пороговое значение θ^*, для которого $P_{FN}(\theta^*) < FN_{\max}$. Поведение вероятности FP при этом для различных пар значений N, d показано на рис.3.12. Принимая во внимание тенденции основных ВВХ функционирования системы, было бы разумно выбрать наименьшие пары параметров N, d, соблюдая, однако при этом ограничение по вероятности ошибки предиктора FP. На рис.3.13 показан выигрыш (в %) среднего времени пребывания в системе схемы с предсказанием относительно системы без предсказания. Как видно из графика, ограничивая вероятность FP до 0.1, можно получить выигрыш в

20% при $FN_{\max} = 0.1$ и выигрыш около 7.5% при $FN_{\max} = 0.001$. В Таблице 3.3 приведены результаты выигрыша (в %) вероятности простоя в системе схемы с предсказанием относительно системы без предсказания.

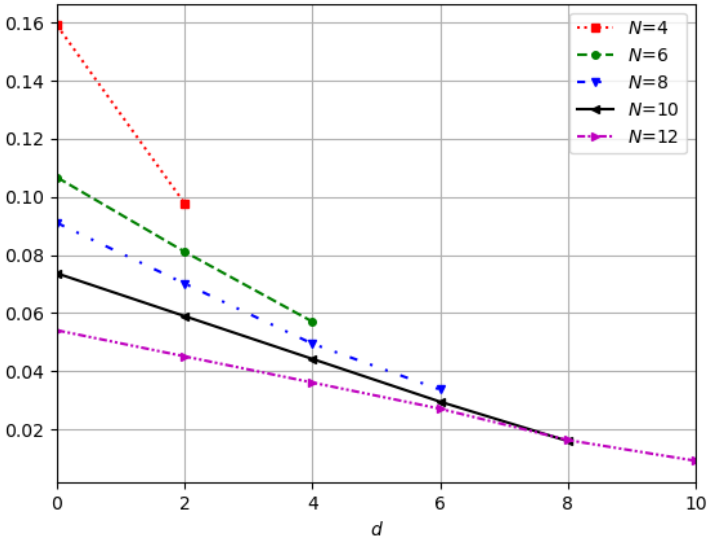


Рис.3.12. Вероятность FP с найденными оптимальными θ для SNR=10 и $FN_{\max} = 0.0001$

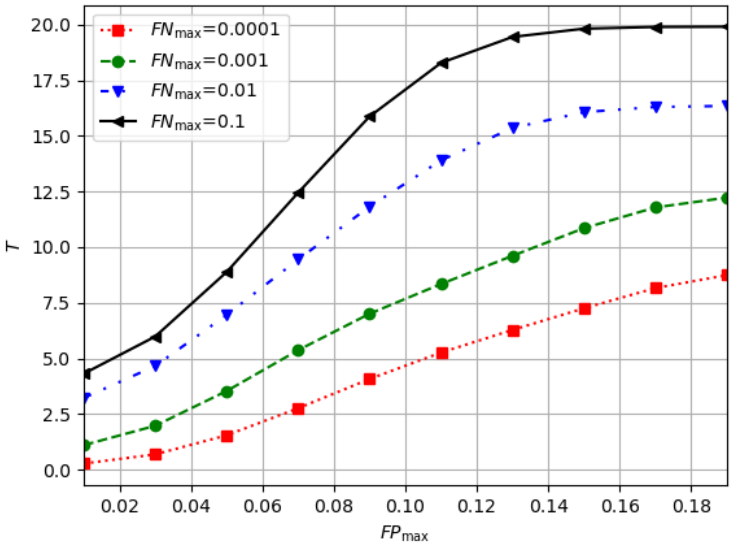


Рис.3.13. Выигрыш (в %) среднего времени пребывания в системе с предсказанием относительно системы без предсказания

Таблица 3.3. Выигрыш (в %) вероятности простоя в системе с предсказанием относительно системы без предсказания

	FP_{max}=0.5				FP_{max}=0.1			
	FN_{max}				FN_{max}			
SNR[dB]	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
5	1.32	4.41	5.23	6.17	1.60	5.62	8.17	40.8
7	1.27	3.12	5.72	6.76	1.98	4.65	15.04	39.09
9	0.55	1.67	3.38	6.02	2.87	8.78	18.07	32.33
10	0.3	1.47	2.87	2.87	3.3	7.51	21.3	27.37
11	0.71	1.42	2.08	9.85	4.73	8.10	18.14	22.78
12	0.68	2.29	7.62	7.62	5.32	7.15	15.51	19.53

ЗАКЛЮЧЕНИЕ

В заключение сформулируем основные результаты диссертации:

1. Разработана модель двухфазной СМО в дискретном времени для решения задачи повышения пропускной способности соты беспроводной гетерогенной сети, позволяющая учитывать различные алгоритмы распределения ресурсов между фазами. Разработан пропорциональный алгоритм распределения ресурсов с ограничениями. Получен метод вычисления основных вероятностных характеристик. Численный анализ показал снижение среднего числа потерянных заявок вследствие адаптации предложенного алгоритма распределения ресурсов к изменениям нагрузки трафика.

2. Разработана модель двухфазной СМО в дискретном времени для оценки показателей эффективности распределения ресурсов при решении задачи межуровневой оптимизации – повышения пропускной способности сети и качества восприятия видео потока на пользовательской станции. Получено матрично-рекуррентное решение для стационарного распределения цепи Маркова, описывающего функционирование модели.

3. Для решения задачи снижения задержки передачи пакета на пользовательскую станцию разработана модель двухфазной СМО в дискретном времени, позволяющая произвести оценку среднего времени пребывания заявки в системе и других показателей эффективности путем итеративного моделирования механизма предсказания повторной передачи и механизма обратной связи. Разработана событийная имитационная модель двухфазной СМО для оценки точности аналитической модели.

4. Формализована и решена задача оптимизации длин пакетов, показателей качества предсказания, а также пороговых значений,

при которых наблюдается наименьшая длительность обслуживания заявки при условии, что ошибка предсказания не превышает допустимых значений. Задача решена методом прямого поиска. Результаты решения задачи оптимизации использованы как исходные данные в алгоритме предсказания повторной передачи, что дает возможность снизить задержку при передаче пакета пользователю.

СПИСОК ОСНОВНЫХ СОКРАЩЕНИЙ

3(4,5)G	–	Мобильная сеть поколения 3(4,5)
БН	–	Буферный накопитель
БС	–	Базовая станция
ВВХ	–	Вероятностно-временные характеристики
ГМ	–	Городская местность
ПФ	–	Производящая функция
РБ	–	Ресурсный блок
РС	–	Ретрансляционная станция
СМО	–	Система массового обслуживания
СУР	–	Система уравнений равновесия
ЦМ	–	Цепь Маркова
3GPP	–	3rd Generation Partnership Project, консорциум, разрабатывающий спецификации для мобильной телефонии
AHS	–	Adaptive HTTP Streaming Адаптивная потоковая передача по протоколу HTTP
BER	–	Bit Error Rate, коэффициент битовой ошибки
BLER		Block Error Rate, отношение числа ошибочных блоков к общему числу блоков
CLA	–	Cross Layer Adaptation, межуровневая адаптация
CQI	–	Channel Quality Indicator, индикатор качества канала
DASH	–	Dynamic Adaptive Streaming over HTTP, динамическое адаптивное потоковое HTTP-вещание
gNB	–	Next Generation Node B, базовая станция
HSPA	–	High Speed Packet Access, высокоскоростной пакетный доступ
HTTP	–	HyperText Transfer Protocol, протокол передачи гипертекста
IoT	–	Internet of Things, интернет вещей

LLR	–	Log-Likelihood Ratios, логарифм отношения правдоподобия
LTE	–	Long Term Evolution, эволюция в долгосрочной перспективе
MCS	–	Modulation and Coding Scheme, схема модуляции и кодирования
MDC	–	Multiple Description Coding, кодирование с множественным описанием
MPD	–	Media Presentation Description, описание представления медиа
NGMN	–	Next Generation Mobile Network, мобильная сеть последующего поколения
OFDM	–	Orthogonal Frequency Division Multiplexing, мультиплексирование с ортогональным частотным разделением
QoE	–	Quality of Experience, качество восприятия
RTT	–	Round Trip Time, временной интервал между процессами первоначальной и повторной передачи
SNR	–	Signal to Noise Ratio, отношение сигнал-шум
TTI	–	Transmission Time Interval, временной интервал передачи
UE	–	User Equipment, пользовательская станция
URLLC	–	Ultra Reliable and Low Latency Communications, сверхнадежные коммуникации с низкой задержкой

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

$:=$	–	знак введения обозначения (со стороны двоеточия)
$u(x) = \begin{cases} 1, x \geq 0; \\ 0, x < 0, \end{cases}$	–	функция Хевисайда
$\delta(a,b) = \begin{cases} 0, a \neq b; \\ 1, a = b, \end{cases}$	–	символ Кронекера
$Geom(\mathbf{q})$	–	ординарный геометрический поток второго рода
$Geom^G(\mathbf{q})$	–	неординарный геометрический поток второго рода
$Geom_s$	–	геометрическое распределение, зависящее от параметра s
$Geom^E(q_2)$	–	геометрическое распределение с опустошением, зависящее от числа q_2 заявок на второй фазе
h	–	длина такта
n	–	номер такта $[nh, (n+1)h)$
$\bar{x} = 1 - x$	–	дополнение до 1 вероятности x
$x_\cdot$	–	полная сумма переменной x_i по индексу i
$\binom{i}{j}$	–	биномиальный коэффициент
$\mathbf{0}$	–	нулевой вектор, нулевая матрица (определяется контекстом)
$\mathbf{I}$	–	единичная матрица
$\mathbf{1}$	–	единичный вектор
$\|X\|$	–	мощность множества X
$\lfloor y \rfloor$	–	округление y в сторону наименьшего целого

СПИСОК ЛИТЕРАТУРЫ

1. Башарин Г.П. Лекции по математической теории телетрафика // М.: РУДН, 2009. – 342 с.
2. Башарин Г.П., Бочаров П.П., Коган Я.А. Анализ очередей в вычислительных сетях. Теория и методы расчета. // М.: Наука, 1989. – 336 с.
3. Башарин Г.П., Гайдамака Ю.В., Самуйлов К.Е. Математическая теория телетрафика и ее приложения к анализу мультисервисных сетей связи следующих поколений // Автоматика и вычислительная техника. – 2013. - № 2. - С. 11-21.
4. Башарин Г.П., Ефимушкин В.А. Исследование однолинейной системы с заявками нескольких типов в дискретном времени // Проблемы передачи информации. – 1984. – № 1. – С. 95-104.
5. Башарин Г.П., Ефимушкин В.А. Алгоритмический анализ структурно сложных систем конечной емкости с двумерным пространством состояний // В кн.: Модели теории телетрафика в системах связи и вычислительной технике. М.: Наука, 1985. – С. 28-41.
6. Башарин Г.П., Ефимушкин В.А. Графо-матричные модели локальных вычислительных сетей // М.: Изд-во УДН, 1986. – 40 с.
7. Башарин Г.П., Харкевич А.Д., Шнепс М.А. Массовое обслуживание в телефонии // М.: Наука, 1968. – 247 с.
8. Бочаров П.П., Громов А.И. О пуассоновской двухфазной система ограниченной емкости // В кн.: Методы теории телетрафика в системах распределения информации. М.: Наука, 1975. – С. 15-28.
9. Бочаров П.П., Печинкин В.А. Теория массового обслуживания // М.: Изд-во РУДН, 1995. – 529 с.

10. Бутурлин И.А., Гайдамака Ю.В., Ефимушкина Т.В., Самуйлов А.К., Самуйлов К.Е. Задачи оптимального планирования межуровневого интерфейса в беспроводных сетях // Информатика и ее применения. – 2012. – № 3. – С. 74-80.

11. Вискова Е.В. Двухфазная система массового обслуживания с марковскими потоком и обслуживанием в дискретном времени // Информационные процессы. – 2005. – Том 5. – № 3. – С. 247-257.

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

13. Вишневский В.М., Ляхов А.И., Портной С.Л., Шахнович И.В. Широкополосные беспроводные сети передачи информации // М.: Техносфера, 2005. – 592 с.

14. Вишневский В.М., Портной С.Л., Шахнович И.В. Энциклопедия WiMAX. Путь к 4G // М.: Техносфера, 2009. – 472 с.

15. Гайдамака Ю., Ефимушкина Т., Самуйлов А., Самуйлов К. Обзор задач оптимального планирования межуровневого интерфейса на базе ортогонального частотного мультиплексирования в беспроводных сетях // В кн.: Труды 14-й Международной конференции «Распределенные компьютерные и телекоммуникационные сети: теория и приложения» DCCN-2011, 26-28 октября 2010 г. – Москва. – М.: НПФ ИНСЕТ, 2011. – С. 180-187.

16. Гайдамака Ю.В., Ефимушкина Т.В., Самуйлов А.К., Самуйлов К.Е. Задачи оптимального планирования межуровневого интерфейса в беспроводных сетях // Информатика и ее применения. – 2012. – Том 6. – Вып. 3. – С.74-80.

17. Гайдамака Ю.В., Зарипова Э.Р., Самуйлов К.Е. Модели обслуживания вызовов в сети сотовой подвижной связи: Учебно-метод. пособие // М.: РУДН, 2008. – 72 с.

18. Гарайшина И.Р., Моисеева С.П., Назаров А.А. Методы исследования коррелированных потоков и специальных систем массового обслуживания // Томск: Изд-во научно-технической литературы, 2010. – 202 с.
19. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания // М.: Наука, ГРФМЛ, 1987. – 336 с.
20. Гольдштейн Б.С., Кучерявый А.Е. Сети связи пост-NGN // СПб.: БХВ-Петербург, 2013. – 160 с.
21. Гольдштейн Б.С., Соколов Н.А., Яновский Г.Г. Сети связи: Учебник для ВУЗов // СПб.: БХВ-Петербург, 2010. – 400 с.
22. Горелов Г.В., Ромашкова О.Н., Чан Туан Ань. Качество управления речевым трафиком в телекоммуникационных сетях // М.: Радио и связь, 2001. – 112 с.
23. Деарт В.Ю. Мультисервисные сети связи. Транспортные сети и сети доступа // М.: Инсвязьиздат, 2007. – 166 с.
24. Ефимушкин В.А. Анализ системы конечной емкости с обслуживанием общего вида и неоднородными заявками в дискретном времени // В кн.: Модели информационных сетей. М.: Наука, 1984. – С. 76-83.
25. Ефимушкин В.А. Классификация дисциплин циклического обслуживания // В кн.: Численные методы и информатика // М.: Изд-во УДН, 1988. – С. 60-69.
26. Ефимушкина Т.В. Исследование вероятностно-временных характеристик для усовершенствованной схемы распределения ресурсов в гетерогенной сети LTE // T-Comm – Телекоммуникации и Транспорт. – 2013. – № 7. – С. 58-65.
27. Ефимушкина Т.В. Модель распределения ресурсов в мобильной гетерогенной сети в виде двухфазной СМО с общими для фаз приборами // Научно-просветительский портал «Академия современных инфокоммуникационных технологий», ЭЛ № ФС 77-50669. [Электронный ресурс] – Режим доступа:

http://www.acikt.ru/index.php/obrazovanie/tekhnicheskoe-napravlenie/seti-podvizhnoj-svyazi . – 41 с. (дата обращения – 15.01.2021).

28. Ефимушкина Т.В., Габуж М., Самуйлов К.Е. Исследование процесса межуровневой адаптации при передаче видео потока в сетях LTE // В кн.: Сб. трудов XII Всероссийского совещания по проблемам управления ВСПУ-2014 16-19 июня 2014 г. [Электронный ресурс] М.: ИПУ РАН, 2014. – С.8544-8553. – Электрон. опт. диск. Номер гос. регистрации 0321401153.
29. Ефимушкина Т.В., Молчанов Д.А., Кучерявый Е.А. Исследование вероятностно-временных характеристик функционирования соты WiMAX с несколькими режимами модуляции и эластичным трафиком данных // T-Comm – Телекоммуникации и Транспорт. – 2010. – № 7. – С. 203-204.
30. Ефимушкина Т.В., Молчанов Д.А., Кучерявый Е.А. Исследование модели управления доступом к каналам сети WiMAX // T-Comm – Телекоммуникации и Транспорт. – 2011. – № 7. – С. 68-71.
31. Ефимушкина Т.В., Самуйлов К.Е. Обзор задач оптимизации ресурсов в беспроводных сетях LTE // В кн.: Труды Всероссийской конференции (с международным участием) «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем». Москва: Изд-во РУДН, 2011. – С. 81-84.
32. Ефимушкина Т.В., Самуйлов К.Е. Исследование методов распределения нагрузки в сетях LTE с ретрансляторами // T-Comm – Телекоммуникации и Транспорт. – 2012. – № 7. – С. 101-106.
33. Ефимушкина Т.В., Самуйлов К.Е. Аналитическая модель схем распределения нагрузки в сетях LTE с разнородными узлами

связи // Discrete and Continuous Models and Applied Computational Science. – 2013. – Вып.1. – С.37-44.

34. Ефимушкина Т.В., Самуйлов К.Е. Двухфазная модель процесса передачи видео с учетом межуровневой адаптации в сети LTE // T-Comm – Телекоммуникации и Транспорт. – 2014. – № 5. – С. 16-21.
35. Ефимушкина Т.В. Исследование двухфазной системы конечной емкости в дискретном времени с распределяемым между фазами множеством приборов // Тез. докл. IX Международной отраслевой научной конференции «Технологии информационного общества». – 24 марта 2015 г. – М.: ИД Медиа Паблишер, 2015. – С.42-43.
36. Ефимушкина Т.В. Многофазная СМО в дискретном времени с распределяемым между фазами множеством приборов // T-Comm – Телекоммуникации и Транспорт. – 2015. – № 7. – С. 60-68.
37. Ефимушкина Т.В. Анализ многофазной СМО с единственным прибором // Тез. докл. IX Международной отраслевой научной конференции «Технологии информационного общества». – 24 марта 2015 г. – М.: ИД Медиа Паблишер, 2015. – С.43-44.
38. Ефимушкина Т.В. Вероятностно-временные характеристики функционирования многофазной СМО с распределяемым множеством приборов для анализа гетерогенной сети подвижной связи // Тез. докл. X Международной отраслевой научной конференции «Технологии информационного общества». – 16-17 марта 2016 г. – М.: ИД Медиа Паблишер, 2016. – С.374-375.
39. Кениг Д., Штойян Д. Методы теории массового обслуживания // М.: Радио и связь, 1981. – 128 с.
40. Кислицын А.А., Орлов Ю.Н. Моделирование эволюции выборочных распределений случайных величин с помощью уравнения Лиувилля// Математическое моделирование, 2020. – Т. 32. – № 1. – С.111-128.

41. Клейнрок Л. Теория массового обслуживания // М.: Машиностроение, 1979. – 432 с.
42. Клейнрок Л. Вычислительные системы с очередями // М.: Мир, 1979. – 600 с.
43. Климов Г.П. Стохастические системы обслуживания // М.: Наука, 1966. – 244 с.
44. Корнышев Ю.Н., Пшеничников А.П., Харкевич А.Д. Теория телетрафика. Учебник для ВУЗов // М.: Радио и связь, 1996. – 270 с.
45. Королев В.Ю, Шоргин С.Я. Математические методы анализа стохастической структуры информационных потоков // М.: ИПИ РАН, 2011. – 130 с.
46. Кучерявый Е.А. Управление трафиком и качество обслуживания в сети Интернет // СПб.: Наука и техника, 2004. – 336 с.
47. Кучерявый А.Е., Цуприков А.Л. Сети связи следующего поколения // М.: ЦНИИС, 2006. – 278 с.
48. Лагутин В.С., Степанов С.Н. Телетрафик мультисервисных сетей связи // М.: Радио и связь, 2000. – 320 с.
49. Лившиц Б.С., Пшеничников А.П., Харкевич А.Д. Теория телетрафика // М.: Связь, 1979. – 224 с.
50. Мардер Н.С. Современные телекоммуникации // М.: ИРИАС, 2006. – 384 с.
51. Назаров А.Н., Сычев К.И. Модели и методы расчета показателей качества функционирования узлового оборудования и структурно-сетевых параметров сетей связи следующего поколения // Красноярск: Изд-во ООО «Поликом», 2010. – 389 с.
52. Наумов В.А. О независимой работе подсистем сложной системы. // В кн. «Труды 3-й Всесоюзной школы-совещания по теории массового обслуживания». М.: Изд-во МГУ, 1976. – Т. 2. – С. 169-177.

53. Наумов В.А., Самуйлов К.Е., Яркина Н.В. Теория телетрафика мультисервисных сетей // М.: РУДН, 2007. – 191 с.
54. Нейман В.И. Телетрафик и теория массового обслуживания // Автоматика и телемеханика. – 2009. – № 12. – С. 29-38.
55. Ромашкова О.Н. Обработка пакетной нагрузки информационных сетей // М.: МИИТ, 2001. – 191 с.
56. Рыкова Т., Хелльге К., Санчес Я., Ширль Т., Хауштайн Т., Тиле Л., Вирт Т., Куррас М., Рашковский Л. Передача сигнала данных в системе беспроводной связи с уменьшенной сквозной задержкой // Заявка №2019124612/07(047981). Патентообладатель: ФРАУНХОФЕР-ГЕЗЕЛЛЬШАФТ ЦУР ФЕРДЕРУНГ ДЕР АНГЕВАНДТЕН ФОРШУНГ Е.Ф. – Дата подачи заявки: 02.08.2019 (Российская Федерация).
57. Рыкова Т. Алгоритм расчета стационарного распределения для многофазной системы конечной емкости в дискретном времени с распределяемым между фазами множеством приборов // Труды XI Международной отраслевой научно-технической конференции «Технологии информационного общества». Москва, 15-16.03.2017 г. – М.: Медиа Паблишер, 2017. – С. 416-418.
58. Рыкова Т. Снижение задержки обработки пакетов в сети 5G с применением искусственного интеллекта // Цифровая инфраструктура для трансформации экономики: задачи и возможности. – М.: МФЮА, 2020. – С.112-114.
59. Саати Т.Л. Элементы теории массового обслуживания и ее приложения // М.: Советское Радио, 1971. – 520 с.
60. Севастьянов Б.А. Курс теории вероятностей и математической статистики // М.: Изд-во ИКИ, 2004. – 272 с.
61. Соколов Н.А. Задачи планирования сетей электросвязи // СПб.: Техника связи, 2012.— 428 с.

62. Степанов С.Н. Основы телетрафика мультисервисных сетей // М.: Эко-Трендз, 2010. – 392 с.
63. Сычев К.И. Многокритериальное проектирование мультисервисных сетей связи // СПб.: Изд-во Политехн. ун-та, 2008. – 272 с.
64. Тихвинский В.О., Терентьев С.В., Юрчук А.Б. Сети мобильной связи LTE: технология и архитектура // М.: Эко-Трендз, 2010. – 284 с.
65. Шнепс-Шнеппе М.А. Системы распределения информации. Методы расчета // М.: Связь, 1979. – 342 с.
66. 3GPP R1-072578. Summary of Downlink Performance Evaluation. May 2007.
67. 3GPP TR 36.913 v8.0.0: Requirements for Further Advancements for E-UTRA (LTE-Advanced). Release 8, 2008.
68. 3GPP TS 36.201 v1.0.0: LTE Physical Layer General Description, 2012.
69. 3GPP Release 16. Technical Report // https://www.3gpp.org/release-16.
70. 3GPP. Study on physical layer enhancements for NR ultra reliable and low latency case (URLLC) .– 3GPP. – Tech. – Rep. TS38.824. – Mar. 2019.
71. Acharya J., Gao L., Gaur S. Heterogeneous Networks in LTE-Advanced // Chichester: John Wiley & Sons. Ltd, 2014. – 271 p.
72. Basharin G.P., Gaidamaka Yu.V., Samouylov K.E. Mathematical Theory of Teletraffic and Its Application to the Analysis of Multiservice Communication of Next Generation Networks // Automatic Control and Computer Sciences Journal. – 2013. – V. 47. – No. 2. – Pp. 62-69.
73. Berezdivin R., Breinig R., Topp R. Next-Generation Wireless Communications Concepts and Technologies // IEEE Communications Magazine. – 2002. – No. 3. – Pp. 108-116.

74. Berardinelli G., Khosravirad S. R., Pedersen K. I., Frederiksen F., Mogensen P. On the benefits of early HARQ feedback with non-ideal prediction in 5G networks // In: Proc. 2016 International Symposium on Wireless Communication Systems (ISWCS), Poznan. – 2016. – Pp. 11-15.
75. Bocharov P.P., Pechinkin A.V., Sanchez S. Stationary state probabilities of a two-phase queueing system with a Markov arrival process and internal losses // In: Proc. of the Fourth Int. Workshop on Queueing Networks with Finite Capacity. Ilkley: 2000. – Pp. 06/1-10.
76. Bohge M., Gross J., Wolisz A., Meyer M. Dynamic Resource Allocation in OFDM Systems: an Overview of Cross-Layer Optimization Principles and Techniques // IEEE Networks. – 2007. – V. 21. – No. 1. – Pp. 53-59.
77. Borodakiy V.Y., Buturlin I.A., Gudkova I.A., Samouylov K.E. Modelling and Analysing a Dynamic Resource Allocation Scheme for M2M Traffic in LTE Networks // In: Lecture Notes in Computer Science. V. 8121. – Berlin, Heidelberg: Springer, 2013. – Pp. 420–426.
78. Bruneel H. Performance of discrete-time queueing systems // Computers and Operations Research. – 1993. – V. 20. – No.3. – Pp. 303-320.
79. Can B., Yanikomeroglu H., Onat F. A., Carvalho E. D., Yomo H. Efficient Cooperative Diversity Schemes and Radio Resource Allocation for IEEE 802.16j // In: Proc. of Wireless Communications and Networking Conference WCNC IEEE. – 2008. – Pp. 36-41.
80. Capozzi F., Piro G., Grieco L.A., Boggia G., Camarda P. Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey // IEEE Communications Surveys & Tutorials. – 2013. – V. 15. – No. 2. – Pp. 678-700.
81. Chen W.Y., Wu C., Lu L.L. Performance Comparisons of Dynamic Resource Allocation With/Without Channel De-Allocation in

GSM/GPRS Networks // IEEE Communications Letters. – 2003. – Vol. 7. – No. 1. – Pp. 10-12.

82. Choand J., Haas Z. On the Throughput Enhancement of the Downstream Channel in Cellular Radio Networks through Multihop Relaying // IEEE Journal on Selected Areas of Communications. – 2004. – V. 22. – No. 9. – Pp. 1206-1219.
83. Courtois P.J. Decomposability. Queueing and Computer System Application // New York: Academic Press, 1977. – 201 p.
84. Damnjanovic A., Montojo J., Yongbin W., Tingfang J. et. al. A Survey on 3GPP Heterogeneous Networks // IEEE Wireless Communications.– 2011. – V. 18. – No. 3. – Pp. 10-21.
85. Deniz D.Z., Mohamed N.O. Performance of CAC Strategies for Multimedia Traffic in Wireless Networks // IEEE Journal of Selected Areas in Communications. – 2003. – V .21.– No. 10. – Pp. 1557-1565.
86. Doshi B.T. Analysis of a Two Phase Queueing System with General Service Times // Operations Research Lett. – 1991. – V. 10. – No.5. – Pp. 265-272.
87. Efimushkina T. Performance Evaluation of a Tandem Queue with Common for Phases Servers // In: Proc. of the 18-th International Conference on Distributed Computer and Communication Networks (DCCN-2015): 19-22 October 2015. – Moscow, Russia. – Moscow: Technosphere, 2015. – Pp.44-51.
88. Efimushkina T., Gabbouj M. Survey on Cross-Layer Adaptation for Video Downlink Communications over LTE // In: Proc. of the 17-th International Conference on on Distributed computer and communication networks: control, computation, communications (DCCN-2013), Moscow, Russia, 7-10 October 2013. – Moscow: Technosfera, 2013. – Pp. 102-109.
89. Efimushkina T., Gabbouj M. Cross-Layer Adaptation-Based Video Downlink Transmission over LTE: Survey // In: Communications in

Computer and Information Science. – No. 279. – Berlin, Heidelberg: Springer, 2014. – Pp. 101–113.

90. Efimushkina T., Gabbouj M., Samuylov K. Analytical model in discrete time for cross-layer video communication over LTE // Automatic Control and Computer Sciences . – 2014. – V. 48. – No. 6. – Pp. 345-357.
91. Efimushkina T., Moltchanov D., Koucheryavy Y. Analysis of WiMAX Cell with two AMC modes and Elastic Data Traffic // In: Proc. of 7th Finnish-Russian University Cooperation in Telecommunications (FRUCT) Conference, Saint-Petersburg, April 26-30, 2010. – Pp. 26-29.
92. Efimushkina T., Moltchanov D., Koucheryavy Y. Analytical Model of a WiMAX Cell in AMC Environment // In: Proc. of the 8th FRUCT Conference, Lappeenranta, Finland, November 9-12, 2010. – Pp. 46-48.
93. Efimushkina T., Samuylov K. Resource Allocation in LTE Heterogeneous Networks // In: Proc. of the 17-th International Conference on on Distributed computer and communication networks: control, computation, communications (DCCN-2013), Moscow, Russia, 7-10 October 2013. – Moscow: Technosfera, 2013. – Pp. 36-43.
94. Efimushkina T., Samuylov K. Analysis of the Resource Distribution Schemes in LTE-Advanced Relay-Enhanced Networks // In: Communications in Computer and Information Science. – No. 279. – Berlin, Heidelberg: Springer, 2014. – Pp. 43-57.
95. Efimushkina T., Samuylov K., Borodakiy V. Queuing Model of Resource Allocation in LTE Uplink Channel // In: Proc. of XXXII International Seminar on Stability Problems for Stochastic Models, Trondheim, Norway, 16-21 June 2014. – Pp. 117-119.
96. Efimushkina T., Vassileva N., Moltchanov D., Koucheryavy Y. Analytical Performance Evaluation of a WiMAX Cell with

VoIP/Elastic Data Traffic // In: Proc. of the IEEE CCNC 2011, Las Vegas, USA, January 9-12, 2011. – Pp. 509-514.

97. Efrosinin D., Gudkova I., Samouylov K., Stepanova N. // Algorithmic Analysis of a Two-Class Multi-server Heterogeneous Queueing System with a Controllable Cross-connectivity. – In Gribaudo M., Sopin E., Kochetkova I. (eds) Analytical and Stochastic Modelling Techniques and Applications. – ASMTA 2019. – Lecture Notes in Computer Science. – V. 12023. – Springer.
98. Elnashar A., El-saidny M., Sherif M. Design, Deployment and Performance of 4G-LTE Networks: A Practical Approach // Chichester: John Wiley & Sons. Ltd, 2014. – 580 p.
99. Enhanced Industrial Internet of Things (IoT) and URLLC support 3GPP. – N.S.B. Nokia. Technical Report RP-193233. – Dec. 2019.
100. Ericsson. Way forward on processing timing reduction for sTTI // Technical Report R1-165854. – 3GPP. – 2016.
101. Fu J., Karasawa Y. Fundamental Analysis on Throughput Characteristics of Orthogonal Frequency Division Multiple Access OFDMA in Multipath Propagation Environments // IEICE Trans. – 2002. – V. J85-B. – No. 11. – Pp. 1884-1894.
102. Göektepe B., Faehse S., Thiele L., Schierl T., Hellge C. Subcode-Based Early HARQ for 5G // In Proc. 2018 IEEE International Conference on Communications Workshops (ICC Workshops), Kansas City, MO. – 2018. – Pp. 1-6.
103. Göktepe B., Rykova T., Fehrenbach T., Schierl T., Hellge C. Feedback Prediction for Proactive HARQ in the Context of Industrial Internet of Things // Proc. of the IEEE Globecom, Taipei, Taiwan, December 2020. arXiv: 2009.06301v1 [eess.SP] 14 Sep 2020. – 7 p.
104. Gopalam S., Hanly S. V., Whiting P. Distributed User Association and Resource Allocation Algorithms for Three Tier HetNets // IEEE Transactions on Wireless Communications. –2020. – V.19. – No.12. – Pp.7913-7926.

105. Hong D., Rappaport S. Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Non-prioritized Handoff Procedures // IEEE Transaction on Vehicular Technology. – 1986. – VT-35. – Pp.77-92.
106. Iannone L. and Fdida S. Evaluating a Cross-Layer Approach for Routing in Wireless Mesh Networks // Telecommunication Systems Journal. Special Issue: Next Generation Networks. – Architectures, Protocols, Performance. – 2006. – V. 31. – No.2-3. – Pp. 173-193.
107. IEEE S802.16j-08/050. Maximum Number of Hops for Centralized Scheduling Mode. Jan. 2008.
108. ISO/IEC 23009-1: Dynamic Adaptive Streaming over HTTP (DASH)-Part 1: Media Presentation Description and Segment Formats. Draft International Standard. 2011.
109. ITU-R M.2134. Requirements Related to Technical Performance for IMT-Advanced Radio Interface(s), October 2008.
110. Iversen V.B. Teletraffic Engineering Handbook // ITU-D, SG 2 Q 16/2. – May 2008.
111. Jackson J.R. Networks of Waiting Lines // Operations Research. – 1957. – V. 5. – No. 4. – Pp. 518–521.
112. Jia S., Li.W., Zhang X., Liu Y., Gu X. Advanced Load Balancing Based on Network Flow Approach in LYE-A Heterogeneous Network // International Journal of Antennas and Propagation, 2014. – Article ID 934101 – 10 p.
113. Kaneko M., Popovski P. Adaptive Resource Allocation in Cellular OFDMA Systems with Multiple Relay Stations // In: Proc. of Vehicular Technology Conference VTC-Spring IEEE, 2007. – Pp. 3026-3030.
114. Kawadia V., Kumar P.R. A Cautionary Perspective on Cross-layer Design // IEEE Wireless Communications. – 2005. – V. 12. – No. 1. – Pp. 3-11.

115. Kelly F.P. Blocking Probabilities in Large Circuit Switched Networks // Advances in Applied Probability. – 1986. – V. 18. – No. 2. – Pp. 473-505.
116. Kelly F.P. Reversibility and Stochastic Networks // Chichester: John Wiley & Sons, 1979. – 630 p.
117. Khan F. LTE for 4G Mobile Broadband. Air Interface Technologies and Performance // Cambridge: Cambridge University Press, 2009. – 506 p.
118. Kim T.-S., Chang S.H., Chae K.C. Performance Analysis of a Discrete-Time Two-Phase Queueing System // ETRI Journal. – 2003. – V. 25. – No. 4. – Pp. 238-246.
119. Kirina-Lilinskaya E. P., Zenyuk D. A., Bobrikova E. V., Orlov Y. N., Gaidamaka Y. V. and Samouylov K. E. Simulating interference in D2D link using fractal random walk model for elasticity analysis // International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. – 2019. – V. 2019. – 6 p.
120. Kivanc D., Li G., Liu H. Computationally Efficient Bandwidth Allocation and Power Control for OFDMA // IEEE Transactions on Wireless Communications. – 2003. – V. 2. – No. 6. – Pp. 1150-1158.
121. Klimenok V., Dudin A. Dual tandem queueing system with multi-server stations and retrials // In: Proc. Int. Conf. on Distributed computer and communication networks: control, computation, communications (DCCN-2013), Moscow, Russia, 7-10 October 2013. – Moscow: Technosfera, 2013. – Pp. 394-401.
122. Kobayashi H., Konheim A. Queueing Models for Computer Communications System Analysis // IEEE Trans. Communications. – 1977. – V. 25. – No. 1. – Pp. 2-29.
123. Kwak R., Cioffi J. Resource Allocation for OFDMA Multi-Hop Relaying Downlink Systems // In: Proc. IEEE Global Telecommunications Conference Globecom, 2007. – Pp. 3225-3229.

124. Kwon E., Lee J., Jung K., Ryu S. A Performance Model for Admission Control in IEEE 802.16 // In: Lecture Notes in Computer Science, 2005. – V. 3610. – Pp. 159-168.
125. Liu C., Bouazizi I., Hannuksela M., Gabbouj M. Rate Adaptation for Dynamic Adaptive Streaming over HTTP in Content Distribution Network // Signal Processing: Image Communications Journal. – 2012. – V. 27. – No. 4. – Pp. 288-311.
126. Liu Y., Deng Y., Elkashlan M., Nallanathan A., Karagiannidis G. Analyzing Grant-Free Access for URLLC Service // IEEE Journal on Selected Areas in Communications. – 2021. – Vol.39. –No.3 – Pp.741-755.
127. Ma K., Bartos R., Bhatia S., Nair R. Mobile Video Delivery with HTTP // IEEE Communications Magazine. – 2011. – V. 49. – No. 4. – Pp. 166–175.
128. Mahmood N. H., Abreu R., Böhnke R., Schubert M., Berardinelli G. and Jacobsen T. H. Uplink Grant-Free Access Solutions for URLLC services in 5G New Radio // In: Proc. of the 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, Finland. – 2019. – Pp. 607-612.
129. Markoval E., Moltchanov D., Pirmagomedov R., Ivanova D., Koucheryavy Y. and Samouylov K. // Priority-based Coexistence of eMBB and URLLC Traffic in Industrial 5G NR Deployments. – 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic. – 2020. – Pp. 1-6.
130. MCC Support. 3GPP TS 38.212 v16.0.0 // Technical Report. – 3GPP. – 2020. – Pp.19–30.
131. Medvedeva E., Gorbunova A., Gaidmaka Y., Samouylov K. A Discrete Queueing Model for Performance Analysis of Scheduling Schemes in Multi-User MIMO Systems // In: Proc. 11th International

Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). – Dublin, Ireland. – 2019 .– Pp.1-5.

132. Next Generation Mobile Network // The NGMN Alliance. [Электронный ресурс] – Режим доступа: http://www.ngmn.org/uploads/media/NGMN_at_a_Glance_-_January_2014.pdf. (дата обращения – 30.01.2015).
133. Osman A., Mohammed A. Performance Evaluation of a Low-Complexity OFDM UMTS-LTE System // In: Proc. Of IEEE Vehicular Technology Conf. (VTC'08), Singapore, May 2008. – Pp. 2142-2146.
134. Otani Y., Ohno S., Teo K., Teo D., Hinamoto T. Subcarrier Allocation for Multi-User OFDM System // In: Proc. Asia-Pacific Communication Conference, Oct. 2005. – Pp. 1073-1077.
135. Oyman O., Singh S. Quality of Experience for HTTP Adaptive Streaming Services // IEEE Comm. Magazine. – 2012. – V. 50. – No. 4. – Pp. 20–27.
136. Psannis K., Ishibashi Y. Efficient Error Resilient Algorithm for H.264/AVC: Mobility Management in Wireless Video Streaming // Springer Telecommunication Systems Journal. – 2009. – V. 41. – No. 2. – Pp. 65-76.
137. T. V. Rykova. Towards the analysis of the performance measures of heterogeneous networks by means of two-phase queueing systems // Discrete and Continuous Models and Applied Computational Science 29 (3) (2021) 242–250. DOI: 10.22363/2658-4670-2021-29-3-242-250.
138. Rykova T., Göktepe B., Schierl T., Hellge C. Analytical Model of Early HARQ Feedback Prediction // In: Lecture Notes in Computer Science. – Springer International Publishing, 2020. – Pp. 222–239.
139. Rykova T., Göktepe B., Schierl T., Hellge C. Analytical Model and Feedback Predictor Optimization for Combined Early-HARQ and HARQ // MDPI Mathematics 2021, *9*, 2104.

140. Rykova T., Hellge C., Sanchez Y., Schierl T., Haustein T., Thiele L., Wirth T., Kurras M., Raschkowski L. Data Signal Transmission in a Wireless Communication System with Reduced End-To-End Latency // Applicant: Fraunhofer-Gesellschaft zur Förderung Erangewandten Forschung e.V. Patent Application Number: 2018- 532115. Drafting Date: October 23, 2020. Representative: Daisuke Noguchi (Japan).
141. Rykova T., Hellge C., Sanchez Y., Schierl T., Haustein T., Thiele L., Wirth T., Kurras M., Raschkowski L. Data Signal Transmission in a Wireless Communication System with Reduced End-To-End Latency // Applicant: Fraunhofer-Gesellschaft zur Förderung Erangewandten Forschung e.V. Patent Application Number: US 2021/0218536 A1 Pub. Date: July 15, 2021. (United States).
142. Rykova T., Algorithmic calculation of stationary distribution for multiphase queue with common for phases servers in discrete time // T-Comm – Телекоммуникации и Транспорт, 2017. – №12. – С.71-76.
143. Rykova T. Marginal loss probabilities at phases for multiphase queue with common for phases servers in discrete time // In: Proc. of 2018 Systems of Signals Generating and Processing in the Field of on Board Communications, 2018. – Pp.1-10.
144. Samouylov K.E., Gudkova I.A. Analysis of an Admission Model in a Fourth Generation Mobile Network with Triple Play Traffic // Automatic Control and Computer Sciences Journal. – 2013. – V. 47. – No. 4. – Pp. 202–210.
145. Samuylov A. et al. // Characterizing Resource Allocation Trade-Offs in 5G NR Serving Multicast and Unicast Traffic. – in IEEE Transactions on Wireless Communications. – V. 19. – No. 5. – Pp. 3421-3434. – May 2020.
146. Sesia S., Toufik I., Baker M. The UMTS Long Term Evolution: From Theory to Practice // Chichester: John Wiley & Sons, 2009. – 648 p.

147. Shannon C. E. A Mathematical Theory of Communication // Bell Syst. Tech. Journal. – 1948. – V. 27. – Pp. 379-423, 623-656.
148. Shariat M., Quddus A.U., Ghorashi S.A., Tafazolli R. Scheduling as an Important Cross-Layer Operation for Emerging Broadband Wireless Systems // IEEE Communication Surveys & Tutorials. – 2009. – V. 11. – No. 2. – Pp. 74-86.
149. Stockhammer T. Dynamic Adaptive Streaming over HTTP – Standards and Design Principles // In: Proc. MMSys'2011. California, USA, New York, ACM, 2011. – Pp. 133-144.
150. Strodthoff N., Göktepe B., Schierl T., Samek W., Hellge C. Machine Learning for Early HARQ Feedback Prediction in 5G // 2018 IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates. – 2018. – Pp. 1-6.
151. Sudame P., Badrinath B.R. On Providing Support for Protocol Adaptation in Mobile Networks // Mobile Networks and Applications, 2001. – V. 6. – No. 1. – Pp. 43-55.
152. Takagi H. Queueing Analysis, Vol. III: Discrete-Time Systems. // Amsterdam: North-Holland Publishing, 1993. – 470 p.
153. Thang T., Ho Q., Kang J. and Pham A. Adaptive Streaming of Audiovisual Content Using MPEG DASH // IEEE Trans. on Consumer Electronics Journal. – 2012. – V. 58. – No. 1. – Pp. 78–85.
154. Tullberg H., Popovski P., Li Z., Uusitalo M.A., Hoglund A., Bulakci O., Fallgren M., and Monserrat J.F. The METIS 5G System Concept: Meeting the 5G Requirements. // IEEE Communications Magazine. – Vol.54. – No.12. – 2016. – Pp.132–139.
155. Vishnevsky V., Larionov A., Semenova O. and Ivanov R. State reduction in analysis of a tandem queueing system with correlated arrivals // 16th International Conference on Information Technologies and Mathematical Modelling. – Vol.800. – 2017. – Pp.215–230.

156. Vishnevsky V. and Semenova O. Polling systems and their application to telecommunication networks // Mathematics. – Vol.9. – No.2. – 2021. – Pp.1–30.
157. Wang L. Resource Allocation in OFDMA Relay-Enhanced Cellular Networks // SOKENDAI Publ. – May 2010. – 117 p.
158. Wong C.Y., Cheng R.S., Letaief K. B. Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation // IEEE Journal on Selected Areas in Communications. – 1999. – V. 17. – No. 10. – Pp. 1747-1757.
159. Wong C., Shen Z., Evans L., Andrews J.G. A Low Complexity Algorithm for Proportional Resource Allocation in OFDMA Systems // In: Proc. IEEE Workshop on Signal Processing Systems, Texas, USA, Oct. 2004. – Pp. 1-6.
160. Wu D., Ci S., Zhang W., Zhang J. Cross-Layer Rate Adaptation for Video Communications over LTE Networks // In: Proc. IEEE Global Communications Conference. Anaheim, CA, 2012. – Pp. 5056-5061.
161. Xue Z., Loo K., Cosmas J., Tun M., et.al. Error-Resilient Scheme for Wavelet Video Codec Using Automatic ROI Detection and Wyner-Ziv Coding Over Packet Erasure Channel // IEEE Transactions on Broadcasting Journal. – 2010. – V. 56. – No. 4. – Pp. 481-493.
162. Yang H., Zhang K., Zheng K., Qian Y. Joint Frame Design and Resource Allocation for Ultra-Reliable and Low-Latency Vehicular Networks // IEEE Transactions on Wireless Communications. – 2020. – V.19. – No.5. – Pp.3607-3622.
163. Yin H., Zhang L., Roy S. Multiplexing URLLC Traffic Within eMBB Services in 5G NR: Fair Scheduling // IEEE Transactions on Communications. – 2021.– V.69. – No.2. – Pp.1080-1093.
164. Zeifman, A., Satin, Ya., Razumchik, R., Kryukova, A., Shilova, G., Bounding the Rate of Convergence for One Class of Finite Capacity Time Varying Markov Queues.// In: Gribaudo M., Iacono M., Phung-Duc T., Razumchik R. (eds) Computer Performance Engineering.

EPEW 2019. – Lecture Notes in Computer Science. – V. 12039. Springer, 2020. – Pp. 148–159.

165. Zeifman, A., Satin, Ya., Kryukova, A., Razumchik, R., Kiseleva, K., Shilova, G., On the Three Methods for Bounding the Rate of Convergence for some Continuous-time Markov Chains.– 2020. – Int. J. Appl. Math. Comput. Sci., 2020. – V. 30. – No. 2. – Pp. 251 – 266.
166. Zhang Y.J., Letaief K.B. Multiuser Adaptive Subcarrier and Bit Allocation with Adaptive Cell Selection for OFDM Systems // IEEE Transactions in Wireless Communications. – 2004. – V. 3. – No. 5. – Pp. 1566-1575.