РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ
ИМЕНИ ПАТРИСА ЛУМУМБЫ

На правах рукописи

Власкина Анастасия Сергеевна

МОДЕЛИ С ЭЛАСТИЧНЫМ ТРАФИКОМ И СИГНАЛАМИ ДЛЯ АНАЛИЗА И РАСЧЁТА ПОКАЗАТЕЛЕЙ ЭФФЕКТИВНОСТИ НАРЕЗКИ СЕТЕВЫХ РЕСУРСОВ

Специальность 1.2.3 – Теоретическая информатика, кибернетика

Диссертация
на соискание ученой степени кандидата
физико-математических наук

Научный руководитель
кандидат физико-математических наук
доцент
Кочеткова Ирина Андреевна

Москва – 2023

Оглавление

Введение

Актуальность темы исследования. В сетях пятого (англ. 5th generation, 5G) и последующих поколений (англ. Next Generation Networks, NGN) одной из важнейших концепций является технология нарезки сети (англ. Network Slicing), которая позволяет разделять вычислительные, сетевые и радиоресурсы базового оператора между сегментами сети. При этом между базовым и виртуальными операторами заключаются соглашения о качестве обслуживания (англ. Service Level Agreement, SLA), в соответствии с которыми базовый оператор планирует распределение ресурса между сегментами сети, например, виртуальными операторами, в условиях нормальной загрузки. Такое распределение будем называть начальным, справедливым или тем, которое соответствует соглашению о качестве обслуживания. Однако с увеличением нагрузки на сеть могут возникать ситуации простоя одного сегмента при наличии ожидающих запросов другого сегмента, в таком случае можно осуществить динамическое перераспределение ресурса. Этим перераспределением управляет система мониторинга или контроллер, который направляет сигналы с определенной периодичностью по которым осуществляется проверка необходимости перераспределения ресурса. С одной стороны, частые сигналы позволяют гибко настроить систему, с другой – увеличивается сигнальная нагрузка. Отсюда возникает проблема настройки частоты поступления сигналов контроллера таким образом, чтобы учесть соответствие начальному распределению, максимально использовать ресурсы и повысить число «успешных» сигналов. Второй момент связан с вопросом, по какому правилу и на сколько изменять объем ресурсов сегмента. Таким образом, диссертационная работа посвящена разработке и анализу таких моделей динамической нарезки радиоресурсов, что и обуславливает ее актуальность.

Степень разработанности темы. В настоящее время вопросами исследования и анализа беспроводных сетей пятого и шестого поколений активно занимаются ведущие российские и зарубежные ученые. Применение новой

технологии нарезки сети вносит свои особенности в построение моделей беспроводных сетей. Среди ученых, внесших вклад в развитие этой области, можно отметить Вишневского В.М. [36], Гайдамака Ю.В. [68, 89, 107, 111, 112, 116–119], Кучерявого А.Е. [14,71], Кучерявого Е.А. [56, 115], Молчанова Д.А. [56, 112], Мутханна А.А. [6, 7, 12], Парамонов А.И. [9], Пшеничникова А.П. [8], Самуйлова К.Е. [4, 89, 101, 107, 108, 112, 113, 115–118, 120, 133, 138], Сопина Э.С. [107, 113, 120], Степанова С.Н. [31, 53], Correia L. [54, 107, 112, 118], Taleb T. [28]. Классическими методами анализа беспроводных сетей является теория массового обслуживания, математическая теория телетрафика и теория случайных процессов, которые также применяются в диссертационной работе. Значительный вклад в развитие этой области внесли ученые Башарин Г.П. [94, 96], Гайдамака Ю.В. [90, 96], Горшенин А.К. [39, 46, 47], Дудин А.Н. [36, 131, 134], Моисеев А.Н. [43], Моисеева С.П. [37], Назаров А.А. [37, 48], Парамонов А.И. [44, 45], Самуйлов К.Е. [90, 95, 96], Степанов С.Н. [33], Цитович И.И. [38].

В диссертационной работе для построения моделей перераспределения ресурса применяются разные классы систем массового обслуживания. Исследованию моделей с эластичным трафиком и дисциплиной разделения процессора посвящены работы ученых Башарина Г.П. [93], Яшкова С.Ф. [114], E. Altman [74], O.J. Boxma [75], R.J. Boucherie [76], M.D. Logothetis [72], J.W. Roberts [73], M. Telek [77]. Значительный вклад в исследования систем массового обслуживания с сигналами внесли ученые Наумов В.А. [42], Сопин Э.С. [40, 41]. Вопросами построения моделей управляемых систем массового обслуживания занимались Горцев А.М.[48], Ефросинин Д.В. [128, 130, 140], Рыков В.В. [125–127], Семенова О.В. [131, 132, 134], Howard R.A. [130].

Целью диссертационной работы является разработка моделей с нетерпеливым эластичным трафиком и минимальной скоростью передачи для анализа и расчета показателей эффективности динамической нарезки радиоресурсов по сигналам в беспроводной сети.

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

- Разработка моделей нарезки сети с нетерпеливым эластичным трафиком и минимальной скоростью передачи и двумя стратегиями перераспределения ресурса по сигналам – фиксированной и с управлением выбором объема ресурса.
- Анализ и разработка алгоритмов расчета показателей эффективности нарезки сети, отражающих занятость ресурса, соответствие распределения ресурса соглашению о качестве обслуживания, вероятность перераспределения ресурса по сигналу, а также влияние на показатели частоты поступления сигналов.

Научная новизна диссертационной работы**:**

№ 1. Модель динамической нарезки радиоресурсов в виде системы массового обслуживания с эластичным трафиком включает контроллер, который отправляет поток сигналов на проверку необходимости перераспределения ресурса. Ранее в системах массового обслуживания, применявшихся для моделирования нарезки ресурсов, перераспределение могло произойти в любой момент времени при изменении состояния системы.

№ 2. Построенная управляемая система массового обслуживания моделирует выбор объема перераспределения ресурса для динамической нарезки радиоресурсов. Ранее рассматривались системы с фиксированной стратегией либо с использованием методов машинного обучения для перераспределения ресурса.

№ 3. В формулировку задачи выбора частоты поступления сигналов и объема перераспределения ресурса заложены занятость ресурса, соответствие распределения ресурса соглашению о качестве обслуживания, вероятность перераспределения ресурса по сигналу. Ранее в системах массового обслуживания, применявшихся для моделирования нарезки

ресурсов, исследовались показатели обслуживания пользователей виртуального операторов.

Теоретическая и практическая значимость работы. Полученные в диссертационной работе результаты по основным соотношениям между параметрами исследуемых систем и качеством обслуживания пользователей могут использоваться для успешного развертывания и эксплуатации сетей операторами связи и обеспечения гарантий соглашений об уровне обслуживания в случае нехватки радиоресурсов.

Разработанные математические модели могут быть применены при управлении ресурсами мобильных беспроводных сетей для оптимизации разделения пропускной способности сети. Учет коэффициентов соответствия соглашению об уровне обслуживания, доли полезных вызовов нарезки и среднего значения по коэффициентам занятости ресурсов позволяет динамически перераспределять ресурсы между сегментами сети. Построенная имитационная модель позволяет определять оптимальные интервалы времени между нарезками сети с учетом загруженности сети для достижения высокой производительности, избегания простоя ресурсов и выполнения минимальных гарантированных требований по передаче данных.

Методы исследования. В диссертации применяются методы теории массового обслуживания, математической теории телетрафика и статистического моделирования.

Положения, выносимые на защиту.

№ 1. Система массового обслуживания с нетерпеливым эластичным трафиком, минимальной скоростью передачи и перераспределением ресурса по сигналам, матричный алгоритм расчета стационарного распределения позволяют рассчитать показатели эффективности нарезки ресурсов как со стороны базового оператора – вероятность перераспределения ресурса по сигналу, так и со стороны виртуальных операторов – вероятность блокировки запросов на передачу эластичного трафика.

№ 2. Управляемая система массового обслуживания с эластичным трафиком и стратегией выбора объема перераспределения ресурса по сигналам, итерационный алгоритм вычисления оптимальной стратегии применимы для настройки параметров динамической нарезки сети с учетом простоя ресурса, отклонения распределения ресурса от значений в соглашении о качестве обслуживания, вероятности перераспределения ресурса по сигналу.

№ 3. Дискретно-событийная модель для произвольного числа сегментов сети с алгоритмом перераспределения ресурса позволяет настроить частоту поступления сигналов для максимизации взвешенных коэффициентов использования ресурса и соответствия распределения ресурса соглашению о качестве обслуживания, вероятности перераспределения ресурса по сигналу.

Степень достоверности и апробация результатов обеспечивается корректным использованием строгих математических доказательств, а также численными экспериментами с применением имитационного моделирования и численного анализа. Основные результаты диссертационной работы представлены на всероссийских и международных конференциях и семинарах:

- международная молодежная научная конференция «Математическое и программное обеспечение информационных, технических и экономических систем» (г. Томск, ИПМКН ТГУ, 2019, 2020);
- конференция «XIII Всероссийское совещание по проблемам управления» (г. Москва, ИПУ РАН, 2019);
- всероссийская конференция с международным участием «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем» (г. Москва, РУДН, 2019-2023);
- международная конференция «International Conference on Next Generation Wired/Wireless Advanced Networks and Systems» (г. Санкт-Петербург, 2019, 2020).

Основные результаты опубликованы в ведущих научных журналах: Discrete and Continuous Models and Applied Computational Science, Lecture Notes in Computer Science, Информатика и ее применение, Известия Саратовского университета, а также в трудах международных конференций, индексируемых в Web of Science и Scopus.

Реализация результатов работы. Результаты диссертационной работы включены в исследования по грантам Министерства науки и высшего образования РФ (грант Президента РФ) № 075-15-2019-1124 «Вероятностные модели сегментации радиоресурсов беспроводных сетей и методы расчета характеристик обслуживания пользователей», РФФИ № 20-37-70079 «Исследование и разработка моделей и интеллектуальных алгоритмов совместного обслуживания трафика с малыми задержками и широкополосного доступа в беспроводных сетях пятого поколения» и научному проекту РУДН «Разработка моделей и алгоритмов нарезки радиоресурсов и приоритетного доступа в беспроводной сети 6G».

Публикации. Основные результаты диссертации изложены в 10 работах [102, 104, 106, 140, 141, 144, 146, 149] в том числе в 5 изданиях, входящих в базу данных Scopus/Web of Science [102, 106, 141, 146, 149], в 1 издании, рекомендованном ВАК РФ [140], в 2 свидетельствах о государственной регистрации программ для ЭВМ [104, 144].

Соответствие паспорту специальности. Диссертационное исследование соответствует следующим разделам паспорта специальности 1.2.3 «Теоретическая информатика, кибернетика», а именно **п. 12** «Модели информационных процессов и структур» в части моделирования процесса передачи данных пользователей сети с нарезкой ресурсов; **п. 30** «Теория управляющих систем» в части моделирования системы управления перераспределением ресурса; **п. 11** «Распределенные многопользовательские системы» в части моделирования системы доступа виртуального оператора к распределенному ресурсу базового оператора.

Личный вклад. Построенные в диссертационной работе модели и результаты их последующего анализа получены автором самостоятельно.

Программные средства, используемые для численного анализа, разработаны с участием автора.

Объем и структура работы. Структура диссертационной работы построена из введения, трех глав, заключения и списка литературы из 145 источников. Диссертационная работа изложена на 130 страницах текста, содержит 50 рисунков и 9 таблиц.

Краткое изложение диссертации. Диссертационная работа состоит из трех глав. Во **введении** обоснована актуальность темы диссертационной работы, определены цели и задачи исследования, сформулированы научная новизна и практическая ценность работы.

В **первой главе** представлены предварительные исследования технологии управления ресурсами, анализ одного сегмента сети, а также аппарат управляемых систем массового обслуживания. Раздел 1.1 посвящен общим принципам управления радиоресурсами в беспроводных сетях пятого поколения при использовании технологии нарезки сети. В разделе 1.2 записан явный вид распределения вероятностей для модели обслуживания пользователей одного сегмента сети с эластичным трафиком и минимально-гарантированной скоростью передачи данных, построена имитационная модель для анализа вероятностно-временных характеристик. В разделе 1.3 получено распределение вероятностей в мультипликативном виде для модели с двумя сегментами без управления радиоресурсами. Раздел 1.4 содержит исследование управления занятия ресурсов в модели облачных вычислений из двух групп виртуальных машин, записана функция получаемого среднего вознаграждения, построена имитационная модель для анализа показателей эффективности.

Вторая глава посвящена модели с фиксированной политикой перераспределения ресурсов между двумя сегментами сети. При этом, раздел 2.1 содержит фиксированный алгоритм управления ресурсами модели при воздействии внешних сигналов контроллера о перераспределении. В разделе 2.2 записана матрица интенсивностей переходов в блочном трехдиагональном виде для фиксированной стратегии выбора объема перераспределения ресурса,

ориентированной на максимальное его использование, и в разделе 2.3 получен матричный рекуррентный алгоритм расчета стационарного распределения вероятностей. Раздел 2.4 содержит показатели эффективности нарезки сети с точки зрения простоя ресурса, отклонения распределения ресурса от значений в соглашении о качестве обслуживания, вероятности перераспределения ресурса по сигналу. При выборе частоты поступления сигналов учитываются также ограничения на вероятности блокировки запросов на передачу эластичного трафика виртуальных операторов, что отражено в разделе 2.5.

В **третьей главе** приводится еще более гибкая модель, позволяющая осуществлять выбор нового объема ресурса при динамической нарезке сети. В разделе 3.1 строится управляемая система массового обслуживания для двух сегментов сети, записывается множество допустимых стратегий выбора нового объема ресурса. Далее в разделе 3.2 строится марковский процесс принятия решения в непрерывном времени и определяется вид функции вознаграждения, компоненты которой отражают три принципа эффективной нарезки ресурсов. Применение итерационного метода решения системы уравнений относительно функций среднего вознаграждения показано в разделе 3.3. Здесь также получен вид целевой функции для улучшения стратегии управления перераспределением. В разделе 3.4 представлено дискретно-событийное моделирование системы с произвольным числом сегментов и формализована задача максимизации показателей эффективности нарезки ресурсов со стороны базового оператора. Анализ показателей эффективности нарезки ресурсов для нескольких сегментов содержит раздел 3.5.

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

ГЛАВА 1
ПОСТРОЕНИЕ И АНАЛИЗ МОДЕЛЕЙ ЗАНЯТИЯ РЕСУРСОВ СЕТИ ПЯТОГО ПОКОЛЕНИЯ

1.1. Динамическая нарезка радиоресурсов

В настоящее время наблюдается рост числа пользовательских устройств различных услуг мобильной сети. Стремясь удовлетворить потребности пользователей с точки зрения качества обслуживания, операторы мобильной связи разрабатывают новое программное и аппаратное обеспечение, внедряя современные принципы построения сетевой архитектуры на основе разрабатываемых стандартов. Новые технологии сетей пятого поколения имеют фундаментальные преимущества, позволяющие поддерживать высокие скорости передачи данных, большее количество пользователей, предоставлять широкий спектр услуг [1, 2]. Существуют три основных сценария использования Международной мобильной связи (англ. International Mobile Telecommunications, IMT-2020) – это расширенная мобильная широкополосная связь (англ. Enhanced Mobile Broadband, eMBB), массовая связь машинного типа (англ. Massive Machine Type Communications, mMTC), сверхнадежная связь с малой задержкой (англ. Ultra Reliability Low Latency Communication, URLLC) [3, 4].

Однако сеть 5G не будет отвечать всем требованиям будущего в 2030 году и далее ожидается, что сети беспроводной связи шестого поколения (6G) обеспечат глобальное покрытие, повышенную спектральную, энергетическую и экономическую эффективность, более высокий уровень безопасности и т.д. [5–7]. В соответствии с этими требованиями сети 6G будут опираться на новые передовые технологии такие, как методы работы с большими данными и Интернета вещей (англ. Internet of Things, IoT) [8, 9], а также новую сетевую архитектуру, включающую множественный доступ, схемы кодирования каналов,

многоантенную технологию, нарезку сети, бессотовую архитектуру и облачные/туманные/граничные вычисления [10].

Нарезка сети – это сквозная концепция, охватывающая сетевые и облачные сегменты сети (сеть радиодоступа, транспортная сеть, сеть граничных вычислений). Она обеспечивает одновременное развертывание нескольких логических, автономных и независимых разделенных сетевых ресурсов, а также группы сетевых и сервисных функций на единой инфраструктурной платформе [11–14]. При применении данного механизма управления поставщик ресурсов может выделять («нарезать») пользователям логически изолированные сегменты сети, каждый из которых спроектирован и оптимизирован для конкретных требований (например, один для сотовой связи, другой для Интернета вещей). Согласно текущим стандартам, сетевой сегмент представляет собой управляемую группу подмножеств ресурсов, сетевых функций/сетевых виртуальных функций на уровне данных, контроля, управления и обслуживания в любой момент времени [11, 15, 16]. Стандарты также указывают на необходимость минимального изменения в качестве обслуживания (англ. Quality of Service, QoS) [17] при динамическом изменении числа ресурсов в сегменте [16, 18] и изолированности сегментов друг от друга, чтобы их влияние друг на друга было минимальным [19].

Такая нарезка исходной сетевой архитектуры на несколько логических и независимых сетей осуществляется с точки зрения, например, вычислительной мощности или пропускной способности [20]. В табл. 1.1 приведены некоторые примеры разделения сети, которые учитывают авторы в своих исследовательских работах. Для большинства случаев сегменты адаптируются под три стандартных сценария eMBB, mMTC и URLLC (рис. 1.1). В качестве примера выступает использование автомобиля с автономным управлением, когда пользователь сети одновременно обращается к разным сегментам сети. Автомобиль с автономным управлением подключается к сети через службу совместной работы автомобиля и дороги (англ. Vehicle-to-Everything, V2X). Пользователь, сидящий в автомобиле, инициирует услугу потоковой передачи видео с высоким разрешением (англ. High-Definition Video, HD) через информационно-развлекательную систему,

имеющуюся в автомобиле. В этом случае служба связи V2X требует малой задержки, но не обязательно высокой скорости передачи данных, в то время как служба потоковой передачи HD-видео требует высокой скорости передачи данных, но терпима к задержкам. Таким образом, служба связи V2X и служба потоковой передачи HD-видео подключаются к разным сегментам сети, например, сегменту URLLC и сегменту eMBB.

Табл. 1.1. Механизмы динамической нарезки радиоресурсов

Нарезка ресурсов с точки зрения	**Основные работы**
стандартных сценариев использования сети IMT-2020 (eMBB, mMTC и URLLC), рис. 1.1	[16, 21–25]
виртуальных операторов, рис. 1.2	[26–29]
трафика (услуг, ресурса): эластичный, потоковый	[30–35]
Динамическое изменение границ сегментов в зависимости от	**Основные работы**
весов услуг (требований трафика к скорости передачи данных или к задержке)	[35, 50 –53]
соглашения об уровне обслуживания (SLA)	[54–56]
приоритета сегментов	[57–59]
стоимости	[60, 61]
модели аукциона	[62–64]
функций полезности	[65]
обеспечения изоляции (гарантированный минимум и/или максимум)	[29, 66, 67]
деградации качества	[68]
обеспечения справедливости с помощью трех критериев: a) the product of user powers (PPC) – отношение средней пропускной способности к средней задержке конкретного пользователя или сегмента; b) the modified throughput/delay criterion (MTD) – критерий пропускной способности/задержки, предназначенный для максимизации средней пропускной способности сети с учетом ограничений на среднюю задержку; c) критерий пороговой функции.	[69]
обеспечения справедливости с помощью трех критериев: a) коэффициент совместного использования; b) максимальное отклонение от пороговых значений; c) временной интервал, в пределах которого отклонение возможно.	[70]

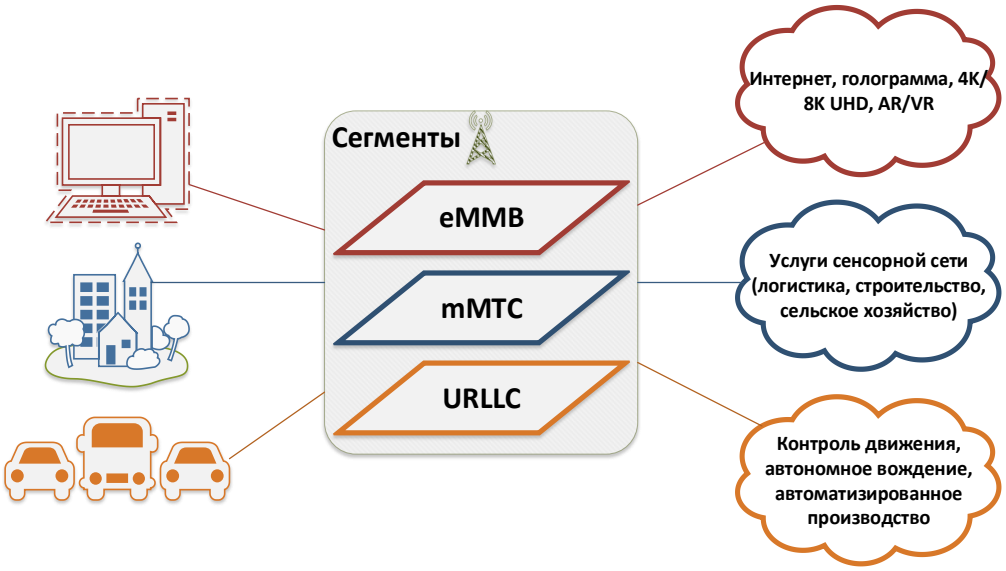


Рис. 1.1. Нарезка ресурсов по сценариям использования сетей 5G [19]

Другим сценарием управления распределением ресурсов из табл. 1.1 является нарезка сети с точки зрения виртуальных операторов. Примером является архитектура сети (рис. 1.2), в которой поставщик телекоммуникационной инфраструктуры (далее – базовый оператор) сдает в аренду свои радиоресурсы виртуальным операторам (далее – оператор), не владеющим своей собственной физической инфраструктурой беспроводной сети, для совместного использования базовой сети. Виртуальный и базовый операторы заключают соглашение об уровне обслуживания, в котором устанавливаются границы сегментов в зависимости от скорости передачи данных. Именно этот способ нарезки ресурсов исследуется в диссертационной работе.

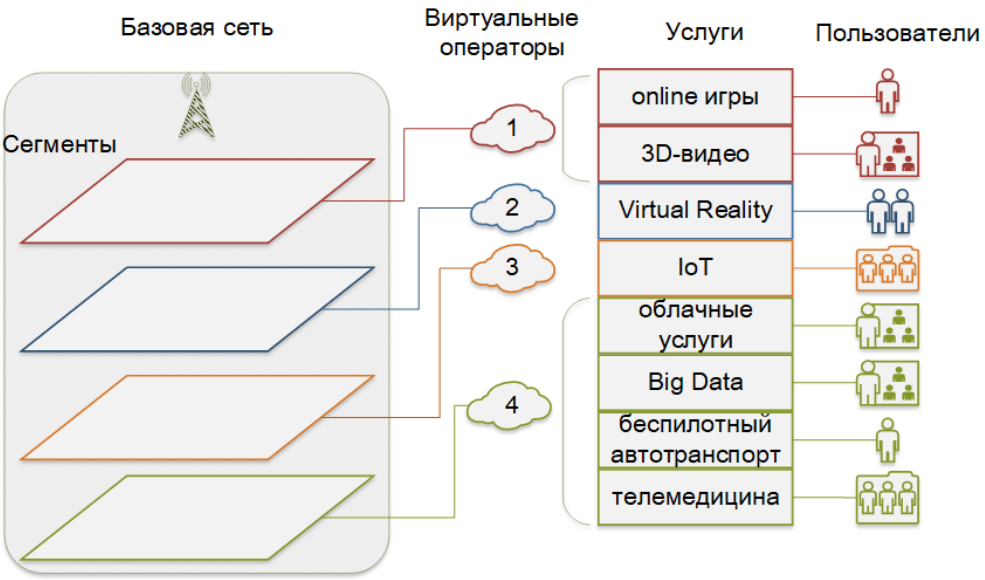


Рис. 1.2. Нарезка ресурсов по виртуальным операторам

Нарезка сети также может осуществляться с точки зрения типа предоставляемых услуг (табл. 1.1). Данная классификация постоянно обновляется, в текущей версии услуги сгруппированы по типу ресурса (табл. 1.2): GBR – трафик, для которого задано значение минимальной скорости передачи данных (данный тип услуг исследуется в диссертационной работе); и Non-GBR – трафик, для которого нет гарантий по скорости передачи данных. Услуги также могут быть разделены по типу передаваемого трафика: услуги потокового трафика (услуга характеризуется продолжительностью) и услуги эластичного трафика (услуга характеризуется объемом данных) [71–77]. В диссертационной работе рассматриваются услуги, генерирующие эластичный трафик с объемом передаваемой информации (блок данных).

Табл. 1.2. Классы услуг [78]

Приоритет услуги	**Примеры услуг**
GBR (с гарантированной скоростью передачи)	
2	Передача голоса
4	Передача видео (прямая трансляция)
3	Игры в реальном времени
2,5	Сообщения V2X
Non-GBR (с негарантированной скоростью передачи)	
1	Сигнализация IMS
6	Передача видео на основе TCP (www, электронная почта)
7	Интерактивные игры
0,5	Критическая сигнализация, чувствительная к задержке (сигнализация MC-PTT, сигнализация MC Video)
6,8	Приложения eMBB, дополненная реальность

В зависимости от нагрузки на сеть задача распределения ресурсов может решаться как в статичной, так и в динамической постановках. Под статичным распределением ресурсов понимается начальное распределение, когда распределение ресурсов производится единовременно без учета возможности изменения границ сегментов во времени. В случае нарезки сети по виртуальным операторам и их услугам ресурс базового оператора V делится между ними в условиях существующих возможностей и ограничений. Виртуальный оператор на основе имеющихся у него ресурсов может предоставлять абонентам всего

$\mathcal{M}=\{1,2,...,M\}$ услуг, где M – число всех возможных услуг. В зависимости от типа ресурса (табл. 1.2) услугам назначается минимальная $b_m^{\min}$ и максимальная $b_m^{\max}$ скорости передачи данных, где m – номер услуги из множества всех услуг, $m\in\mathcal{M}$. Деление услуг по типу трафика (табл. 1.1) отразим следующим образом: услуги потокового трафика $\mathcal{M}_s\subseteq\mathcal{M}$, характеризующегося продолжительностью s_m, $m\in\mathcal{M}_s$; и услуги эластичного трафика $\mathcal{M}_e\subseteq\mathcal{M}$, характеризующегося объемом μ_m^{-1}, $m\in\mathcal{M}_e$, $\mathcal{M}_s\cup\mathcal{M}_e=\mathcal{M}$. Кроме того, каждый n-оператор, $n=1...N$, предоставляет пользователям свой собственный набор услуг, $\mathcal{M}_n\subseteq\mathcal{M}$, а (n,m) – номер услуги для n-оператора.

Объем выделяемого ресурса под каждый сегмент (услугу n-оператора) обозначим V_{nm}, $\sum_{n=1}^{N}\sum_{m\in\mathcal{M}_n}V_{nm}\le V$, $m\in\mathcal{M}_n$: $0\le V_{nm}^{\min}\le V_{nm}\le V_{nm}^{\max}\le V$, $V_n=\sum_{m\in\mathcal{M}_n}V_{nm}$, где $V_{nm}^{\min}$ и $V_{nm}^{\max}$ – являются минимальным и максимальным объемами выделяемого ресурса. Из табл. 1.1 следует, что услуга характеризуется приоритетом α_{nm}, $0\le\alpha_{nm}\le 1$, $\sum_{n=1}^{N}\sum_{m\in\mathcal{M}_n}\alpha_{nm}=1$. Число пользователей, имеющих доступ к m-услуге n-оператора обозначим как K_{nm}, тогда $K_n=\sum_{m\in\mathcal{M}_n}K_{nm}$ и $K=\sum_{n=1}^{N}K_n=\sum_{n=1}^{N}\sum_{m\in\mathcal{M}_n}K_{nm}$. Целевая функция записывается как $f(V)=\sum_{n=1}^{N}\sum_{m\in\mathcal{M}_n}\alpha_{nm}V_{nm}$ при одновременном учете приоритета услуг и других ограничений. Следовательно, процедура распределения ресурсов может быть сформулирована как задача оптимизации следующим образом [78]

$$\max_{V} f(V)=\sum_{n=1}^{N}\sum_{m\in\mathcal{M}_n}\alpha_{nm}V_{nm},$$

$$s.t: \begin{cases} \sum_{n=1}^{N} \sum_{m \in \mathcal{M}_n} V_{nm} \le V, \\ 0 \le V_{nm}^{\min} \le V_{nm} \le V_{nm}^{\max} \le V. \end{cases} \quad (1.1)$$

Динамическая постановка предполагает гибкое управление распределением ресурса во времени, т.е. использование статичного планирования ресурсов, которое затем будет изменяться с учетом нагрузки на сеть для обеспечения лучшего качества обслуживания (рис. 1.3). Для достижения этой цели в диссертационной работе предложена модель с системой мониторинга (далее – контроллером), которая управляет перераспределением ресурса посредством отправки сигнала о проверке необходимости перераспределения ресурсов. Результаты показали, что не при каждом поступлении сигнала будет инициировано перераспределение из-за состояния системы, те сигналы, которые привели к перераспределению ресурса, будем называть успешными (далее – успешные сигналы).

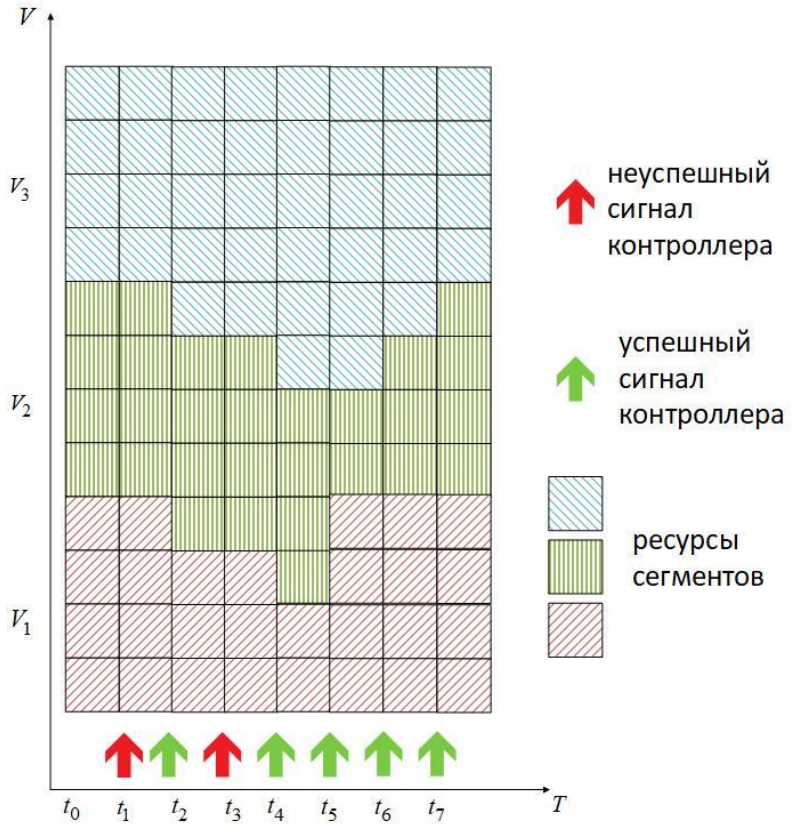


Рис. 1.3. Динамическая нарезка – перераспределение ресурсов по сигналам контроллера

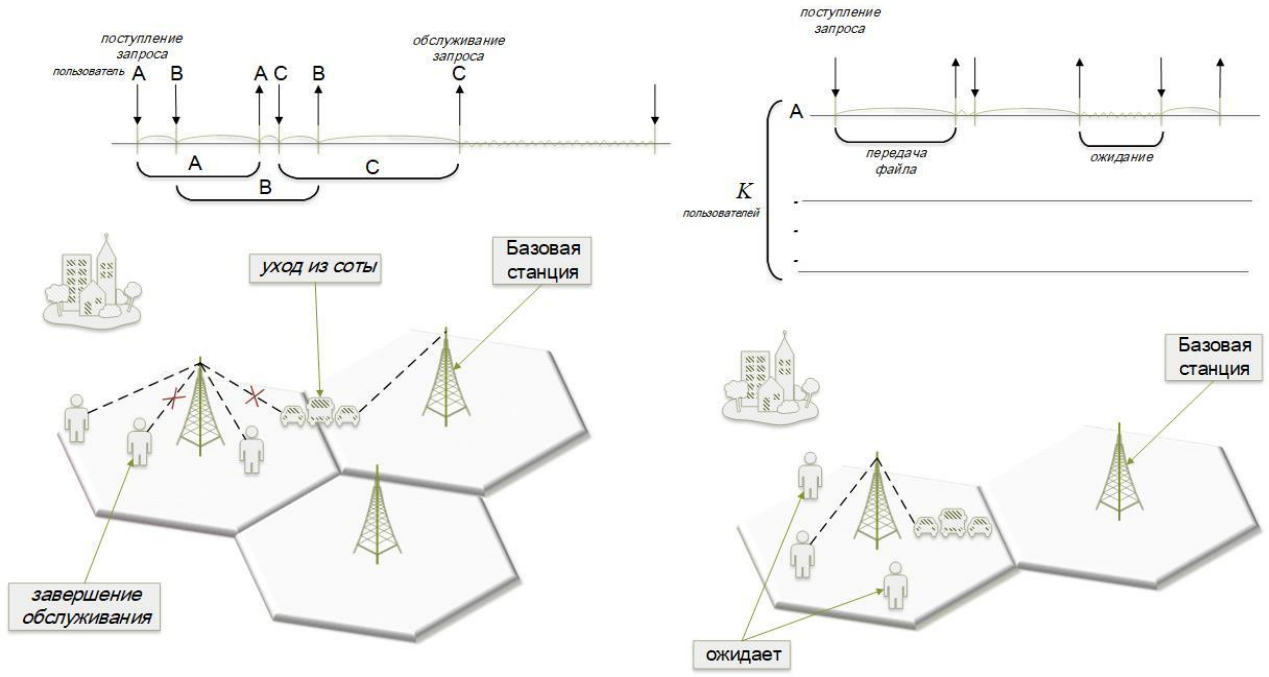
Эффективность использования ресурсов напрямую зависит от нахождения адекватной политики распределения ресурсов между сегментами сети. Анализ способов эффективного управления сетевыми ресурсами по процедуре нарезки сети изложен во многих научных изданиях, что свидетельствует об актуальности выбранной тематики. Для решения задач оптимизации наиболее часто используются математические методы: аппарат теории массового обслуживания [36–45]; теории случайных процессов [46, 47]; теории оптимизации [35, 50–52, 54, 57, 58, 80]; теории игр [62–64, 81]; кооперативная теория игр (равновесие Нэша); машинного обучения [80, 82]; марковских процессов принятия решений (англ. Markov Decision Process, MDP); цепь Маркова с непрерывным временем (англ. Continuous-Time Markov Chain, CTMC) [83, 84].

Совместно с О.В. Семёновой автор проводила исследования [85–88] в части осуществления доступа к ресурсам беспроводных сетей по радиоканалу случайного доступа (англ. Random Access Channel, RACH) путем пересылки четырех сигнальных сообщений между базовой станцией (англ. eNodeB, eNB) и конечным устройством (англ. User Equipment, UE), целью которых была разработка улучшенной, по сравнению с работами [89, 90], процедуры установления связи так, чтобы повысить вероятность успешного соединения и уменьшить среднюю задержку доступа. Полученные результаты позволяют оценить указанные показатели качества при условии возможности повторной передачи последнего сигнального сообщения (результирующий ответ в процедуре установления соединения). Автор также проводила исследования в части совместного использования технологий мультивещания и прямого взаимодействия устройств [91, 92], целью которых была разработка модели для минимизации задержки передачи данных в сети.

1.2. Модель с нетерпеливым эластичным трафиком и минимальной скоростью

Поскольку в сегментах сети функционируют классы услуг с различными характеристиками, при управлении ресурсами необходимо учитывать различные аспекты такие, как требуемые объемы ресурсов, QoS для каждого класса услуг. Ранее в разделе 1.1 упоминалось деление трафика на две крупные категории – потоковый и эластичный [93, 94]. Первый требует выполнения гарантий обслуживания из-за чувствительности к задержке (такие услуги, как передача голоса, видеоконференции). Второй относится к услугам, которые могут регулировать свои скорости в соответствии с доступной полосой пропускания канала (англ. Bandwidth), т.к. данный класс трафика допускает задержки (передача данных, веб-просмотр, электронная почта). Различные модели эластичного [94–98] и потокового [96, 99, 101] трафика опубликованы в исследовательских работах.

Рассмотрим сеть с одним виртуальным оператором [102], который обеспечивает пользователям высокоскоростной доступ к услуге передачи данных (загрузка файла). Такой тип услуги требует выполнения минимально гарантированных скоростей передачи данных b единиц на ресурсе V, т.е. число занимаемых ресурсов блоком данных всегда $\geq b$. Предполагается, что входящий поток запросов на передачу блоков эластичных данных является пуассоновским потоком первого рода с интенсивностью $0<\lambda<\infty$, объем блока распределен по экспоненциальному закону со средним $0<\mu^{-1}<\infty$. Обозначим r длину очереди для ожидающих начала обслуживания запросов, при этом запросы покидают систему по причине превышения времени ожидания начала обслуживания с интенсивностью $0<\varepsilon<\infty$. Исходя из функционирования системы рассматриваются два сценария поведения пользователей (обозначим их сценариями I и II) и соответствующие им модели трафика (рис. 1.4).



I II

Рис. 1.4. Сценарии поведения пользователей – переменное (I) и фиксированное (II) число пользователей

Сценарий I с переменным числом пользователей (рис. 1.5): каждый пользователь отправляет запрос на доступ к услуге (загрузка файла), занимает ресурс для получения услуги (загружает файл) и исчезает из системы. Исчезновение может быть связано как с уходом пользователя из границ рассматриваемой соты, сменой услуги, так и с завершением обслуживания. Функционирование системы описывается одномерным случайным процессом $\boldsymbol{N}(t) \in \left\{1,...,N = \left\lfloor \frac{V}{b} \right\rfloor\right\}$ – число запросов, находящихся в системе в момент времени $t \geq 0$ над пространством состояний $\mathcal{X} = \left\{\boldsymbol{n} : 0,\ldots,N,\ldots,(N+r)\right\}$. Диаграмма интенсивностей переходов изображена на рис. 1.6 и представляет собой процесс рождения и гибели. Распределение вероятностей для модели с переменным числом пользователей приводится в утверждении 1.1.

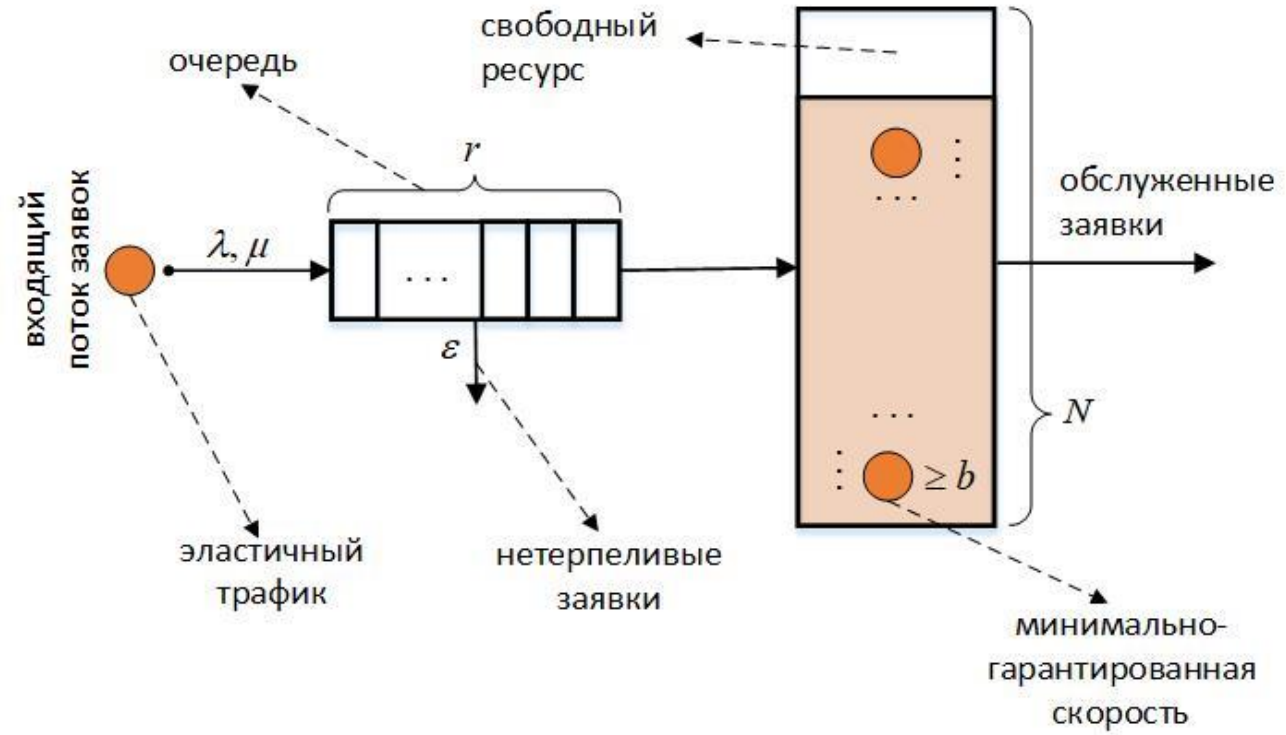


Рис. 1.5. Схема СМО для сценария с переменным числом пользователей (I)

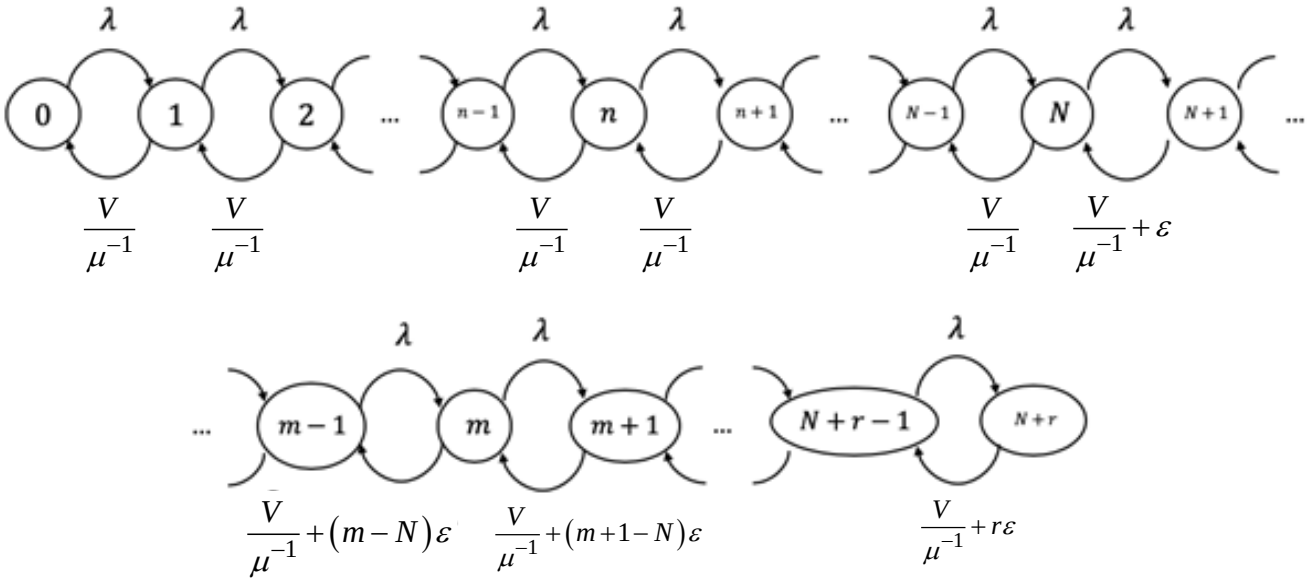


Рис. 1.6. Диаграмма интенсивностей переходов для СМО с переменным числом пользователей (I)

Утверждение 1.1. Распределение вероятностей для модели с переменным числом пользователей вычисляется по формуле

$$p_n = \begin{cases} \left(\frac{\lambda}{V\mu}\right)^n p_0, \ \ n = \overline{1; N}, \\ \left(\frac{1}{V\mu}\right)^N \frac{\lambda^n}{\prod_{i=1}^{n-N}(V\mu + i\varepsilon)} p_0, \ \ n = \overline{N+1; N+r}, \end{cases} \tag{1.2}$$

$$p_0 = \left(\sum_{n=0}^{N} \left(\frac{\lambda}{V\mu}\right)^n + \left(\frac{1}{V\mu}\right)^N \sum_{n=N+1}^{N+r} \frac{\lambda^n}{\prod_{i=1}^{n-N}(V\mu + i\varepsilon)} \right)^{-1}.$$

Доказательство. Используя диаграмму интенсивностей переходов (рис. 1.6), запишем систему уравнений глобального баланса (СУГБ) для модели с переменным числом пользователей

$$\begin{cases} \lambda p_0 = V\mu p_1, \\ (\lambda + V\mu) V\mu p_n = \lambda p_{n-1} + V\mu p_{n+1}, \ n = \overline{1; N-1}, \\ \left(\lambda + V\mu + (m-N)\varepsilon\right) p_m = \lambda p_{m-1} + \left(V\mu + (m+1-N)\varepsilon\right) p_{m+1}, m = \overline{N; N+r-1}, \\ \left(V\mu + r\varepsilon\right) p_{N+r} = \lambda p_{N+r-1}, \end{cases}$$

с условием нормировки $\sum_{i=0}^{N+r} p_i = 1$. Тогда если $n = \overline{1, N}$:

$$p_0 = \frac{V\mu}{\lambda} p_1, p_1 = \frac{\lambda}{V\mu} p_0; \ p_1 = \frac{V\mu}{\lambda} p_2, p_2 = \frac{\lambda}{V\mu} p_1 = \left(\frac{\lambda}{V\mu}\right)^2 p_0; \ldots;$$

$$p_n = \left(\frac{\lambda}{V\mu}\right)^n p_0.$$

Если $n = \overline{N+1, N+r}$: $p_N = \frac{V\mu + \varepsilon}{\lambda} p_{N+1}, p_{N+1} = \frac{\lambda}{V\mu + \varepsilon} p_N$;

$$p_{N+1} = \frac{V\mu + 2\varepsilon}{\lambda} p_{N+2}, p_{N+2} = \frac{\lambda}{V\mu + 2\varepsilon} p_{N+1} = \frac{\lambda}{V\mu + 2\varepsilon} \frac{\lambda}{V\mu + \varepsilon} p_N.$$

Отсюда

$$p_n = \begin{cases} \left(\frac{\lambda}{V\mu}\right)^n p_0, \ \ n = \overline{1;N}, \\ \left(\frac{1}{V\mu}\right)^N \frac{\lambda^n}{\prod_{i=1}^{n-N}(V\mu + i\varepsilon)} p_0, \ \ n = \overline{N+1;N+r}. \end{cases}$$

Вычислим p_0,

$$\frac{1}{p_0} = \sum_{n=0}^{N}\left(\frac{\lambda}{V\mu}\right)^n + \sum_{n=N+1}^{N+r}\left(\frac{1}{V\mu}\right)^N \frac{\lambda^n}{\prod_{i=1}^{n-N}(V\mu + i\varepsilon)},$$

$$p_0 = \left(\sum_{n=0}^{N}\left(\frac{\lambda}{V\mu}\right)^n + \left(\frac{1}{V\mu}\right)^N \sum_{n=N+1}^{N+r} \frac{\lambda^n}{\prod_{i=1}^{n-N}(V\mu + i\varepsilon)}\right)^{-1}.$$

Утверждение доказано. □

Сценарий II с фиксированным числом пользователей (рис. 1.7): в сети функционирует фиксированное число пользователей, каждый из которых отправляет запрос на доступ к услуге (загрузка файла), занимает ресурс для получения услуги (загружает файл), потом ожидает, снова отправляет запрос на доступ к услуге и т.д. Поэтому вместо одного пуассоновского потока запросов рассматривается k, $k = \overline{1;K}$, $0 < K \le N$ малоинтенсивных независимых источников запросов, каждый из которых не может подать новый запрос, пока не будет обработан предыдущий, отправленный им. Каждый из k источников в свободном состоянии может с интенсивностью λ' сгенерировать один запрос, который мгновенно займет один из свободных ресурсов, если он имеется, либо займет место в очереди. Входящая нагрузка будет являться пуассоновским потоком второго рода, а пространство состояний примет вид $\mathcal{X} = \{n : 0,\ldots,N,\ldots,\min(K, N+r)\}$, где число состояний зависит от числа источников запросов K и общего числа мест в системе $(N+r)$.

Из-за такой зависимости числа состояний от числа источников запросов и общего числа мест в системе модель распадается на три случая (соотношения) – $0 < K \le N$, $N < K \le N + r$ и $K > N + r$. Дальнейшие рассуждения будут строиться

относительно последнего соотношения, как наиболее обобщенного случая, предполагая, что в некоторый момент времени может происходить только одно событие – поступление или уход из системы запроса по причине обслуживания или нетерпеливости. Аналогично сценарию I для последующего построения СУГБ для стационарных вероятностей построена диаграмма интенсивностей переходов, рис. 1.8. Распределение вероятностей для модели с фиксированным числом пользователей приводится без доказательства (аналогично утверждению 1.1) в утверждении 1.2.

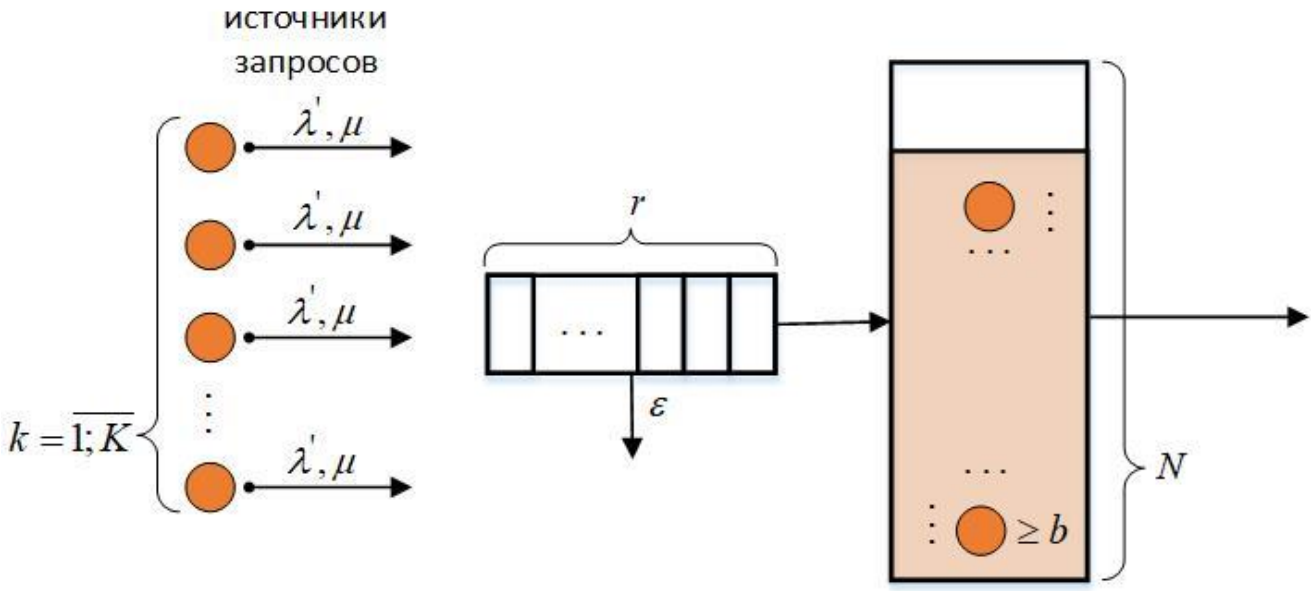


Рис. 1.7. Схема СМО для сценария с фиксированным числом пользователей (II)

Утверждение 1.2. [103] Распределение вероятностей для модели с фиксированным числом пользователей вычисляется по формуле

$$p_n = \begin{cases} \left(\dfrac{\lambda'}{V\mu}\right)^n A_k^n \, p_0, \;\; n = \overline{1, \min(N,k)}, \\ \left(\dfrac{1}{V\mu}\right)^N \dfrac{\left(\lambda'\right)^n}{\prod_{i=1}^{n-N}(V\mu + i\varepsilon)} A_k^n \, p_0, \;\; n = \overline{N+1, \min(N+r,k)}, \end{cases} \tag{1.3}$$

$$p_0 = \left(\sum_{n=0}^{\min(N,\,k)} \left(\frac{\lambda'}{V\mu}\right)^n A_k^n + \left(\frac{1}{V\mu}\right)^N \sum_{n=1}^{\min(r,k-N)} \frac{\left(\lambda'\right)^n}{\prod_{i=1}^{n-N}(V\mu + i\varepsilon)} A_k^n \right)^{-1},$$

где $A_k^n = \frac{n!}{(n-k)!}$.

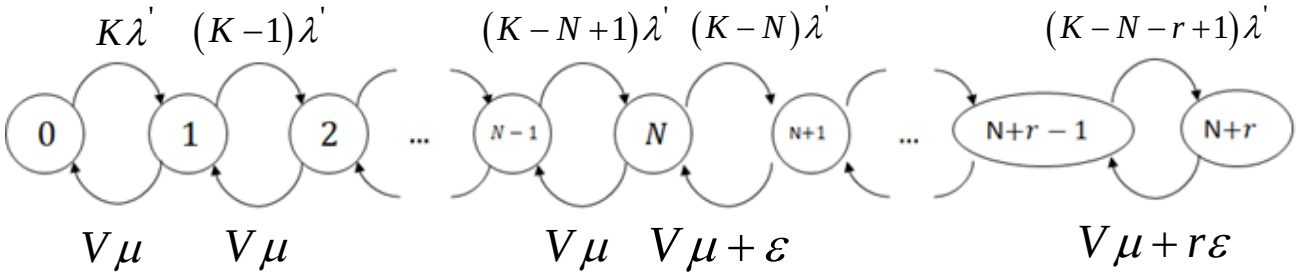


Рис. 1.8. Диаграмма интенсивностей переходов для СМО с фиксированным числом пользователей (II) при $K > N + r$

Пример 1.1. Проиллюстрируем поведение показателей эффективности в зависимости от модели поведения пользователей с помощью построенной имитационной модели [104]. В качестве исходных данных выбраны значения: число пользователей системы варьируется в рамках $K = [1;350]$; минимальная гарантированная скорость передачи данных $b = 0,384$ [Мб/с]; ресурс объема $V = 16,76$ [Мб/с]; интенсивность входящего потока для сценария II (экспоненциальное распределение) $\lambda' = 0,0056$; интенсивность входящего потока для сценария I (экспоненциальное распределение) с параметром $\lambda = \lambda' K = 0,28$; средний размер файла (экспоненциальное распределение и усеченное логнормальное распределение с параметрами $\mu = 14,45, \sigma = 0,35$) 2 [МБ]; интенсивность ухода из системы по причине нетерпеливости $\varepsilon = 0,000001$; длина очереди $r = 20$.

Результаты сравнения сценариев поведения пользователей и законов распределения длин файла показали, что экспоненциальный закон является оценкой сверху для усеченного логнормального (рис. 1.9). Можно найти момент, в который модель с пуассоновским потоком второго рода может быть заменена моделью с пуассоновским потоком первого рода, т.к. с точки зрения моделирования и вычислительной сложности проще работать с пуассоновским потоком первого рода. При заданных исходных данных это наблюдается, начиная примерно с 350 пользователей (рис. 1.10).

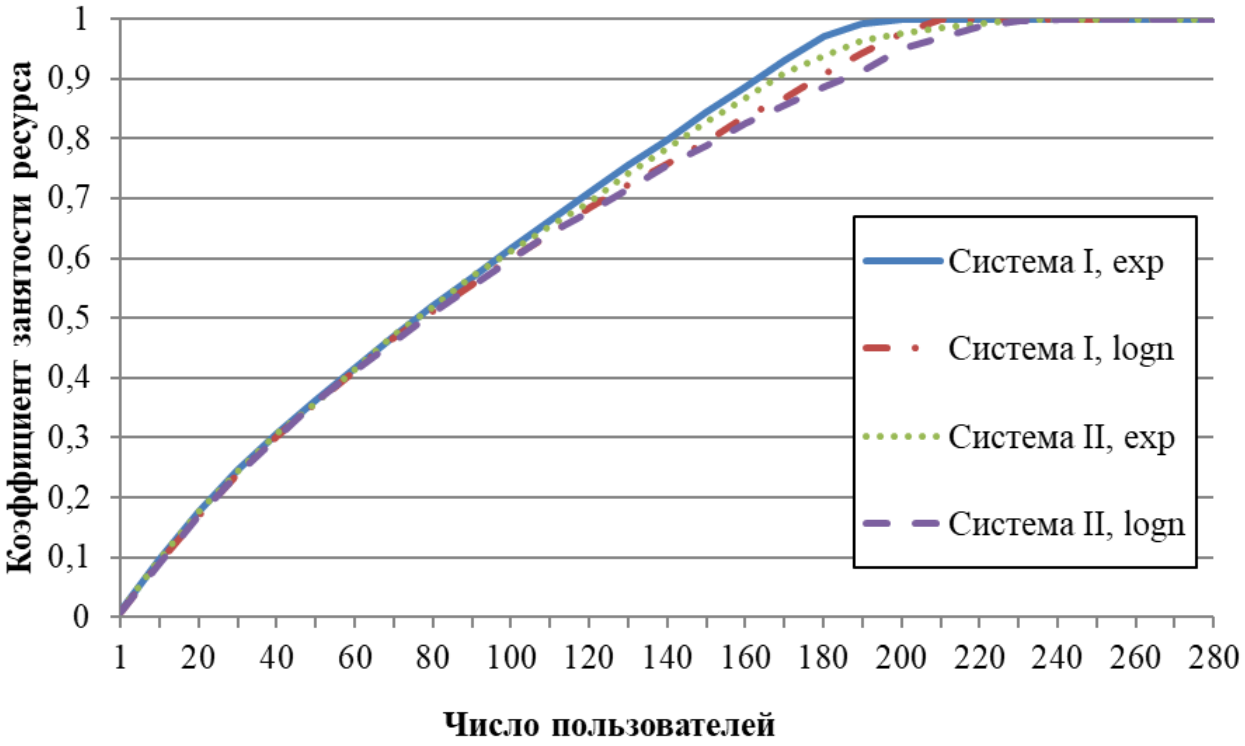


Рис. 1.9. Коэффициент использования ресурса для двух моделей поведения пользователей (переменное I, фиксированное II) и разных законов распределения размера файла

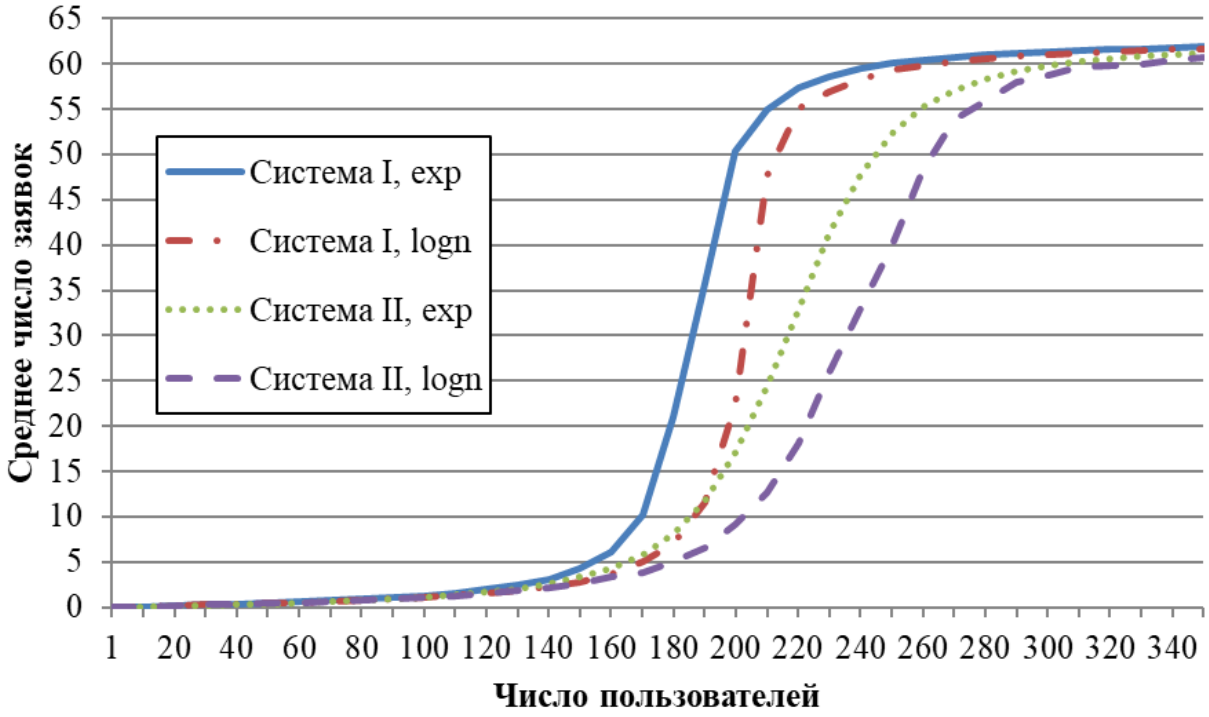


Рис. 1.10. Среднее число пользователей в системе для двух моделей поведения пользователей (переменное I, фиксированное II) и разных законов распределения размера файла

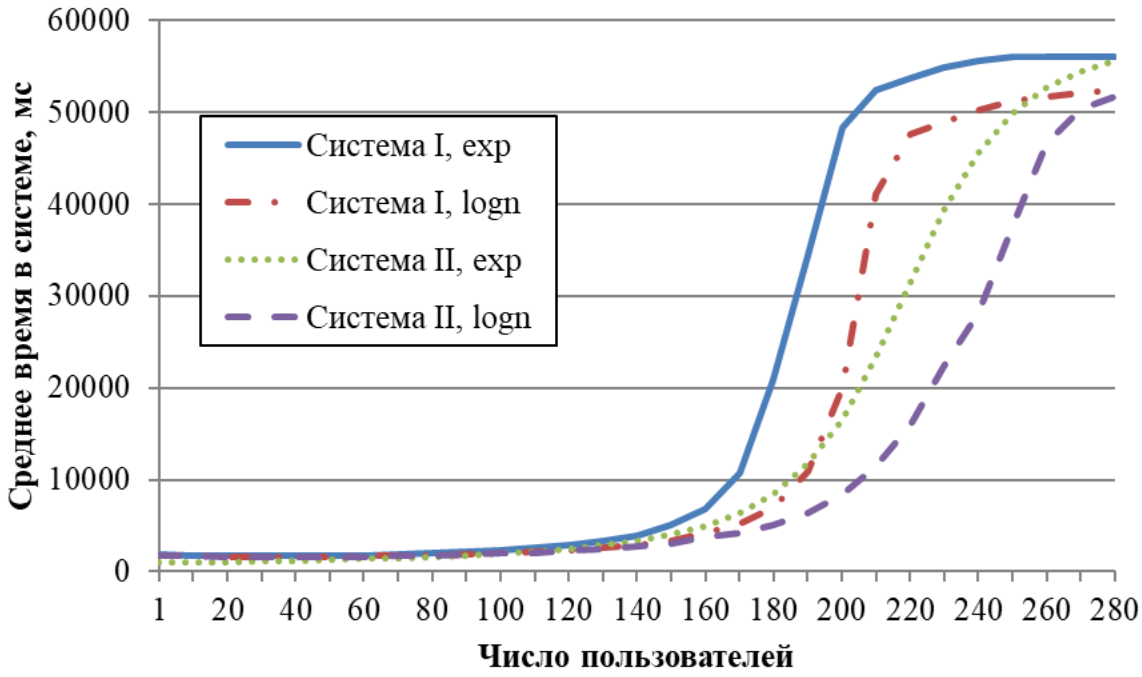


Рис. 1.11. Среднее время пребывания пользователя в системе для двух моделей поведения пользователей (переменное I, фиксированное II) и разных законов распределения размера файла

Пример 1.2. Проиллюстрируем поведение вероятностно-временных характеристик для системы с переменным числом пользователей (формулы для расчета представлены в [102]). В качестве исходных данных выбраны значения: число пользователей системы $K = 50$; минимальная гарантированная скорость передачи данных $b = 0,384$ [Мб/с]; ресурс объема $V = 16,76$ [Мб/с] и $V = 3500$ [Мб/с]; интенсивность входящего потока (экспоненциальное распределение) $\lambda' = [0.01; 20]$; средний размер файла (экспоненциальное распределение и усеченное логнормальное распределение с параметрами $\mu = 14,45, \sigma = 0,35$) 2 [МБ]; интенсивность ухода из системы по причине нетерпеливости $\varepsilon = 0,000001$; длина очереди $r = 20$.

Рис. 1.12 (а), (б) [105] отражает зависимость вероятностно-временных характеристик от интенсивности поступления заявок в систему при пропускной способности всего ресурса в $V = 16,76$ [Мб/с]. Заметим, что в этом случае общее время для служб отправки превышает 40 с, а размер файла составляет 2 Мб. При незначительном увеличении интенсивности поступления запросов система быстро

заполняется и формируется очередь. Чтобы приблизить задержку к реальным значениям были проведены расчеты для $V = 3500$ [Мб/с], рис. 1.12 (в), (г). Для таких исходных данных значение среднего времени пребывания заявки в системе колеблется от 8,27 мс до 156,9 мс и почти напрямую зависит от интенсивности поступления запросов. Очередь с выбранной интенсивностью не формируется.

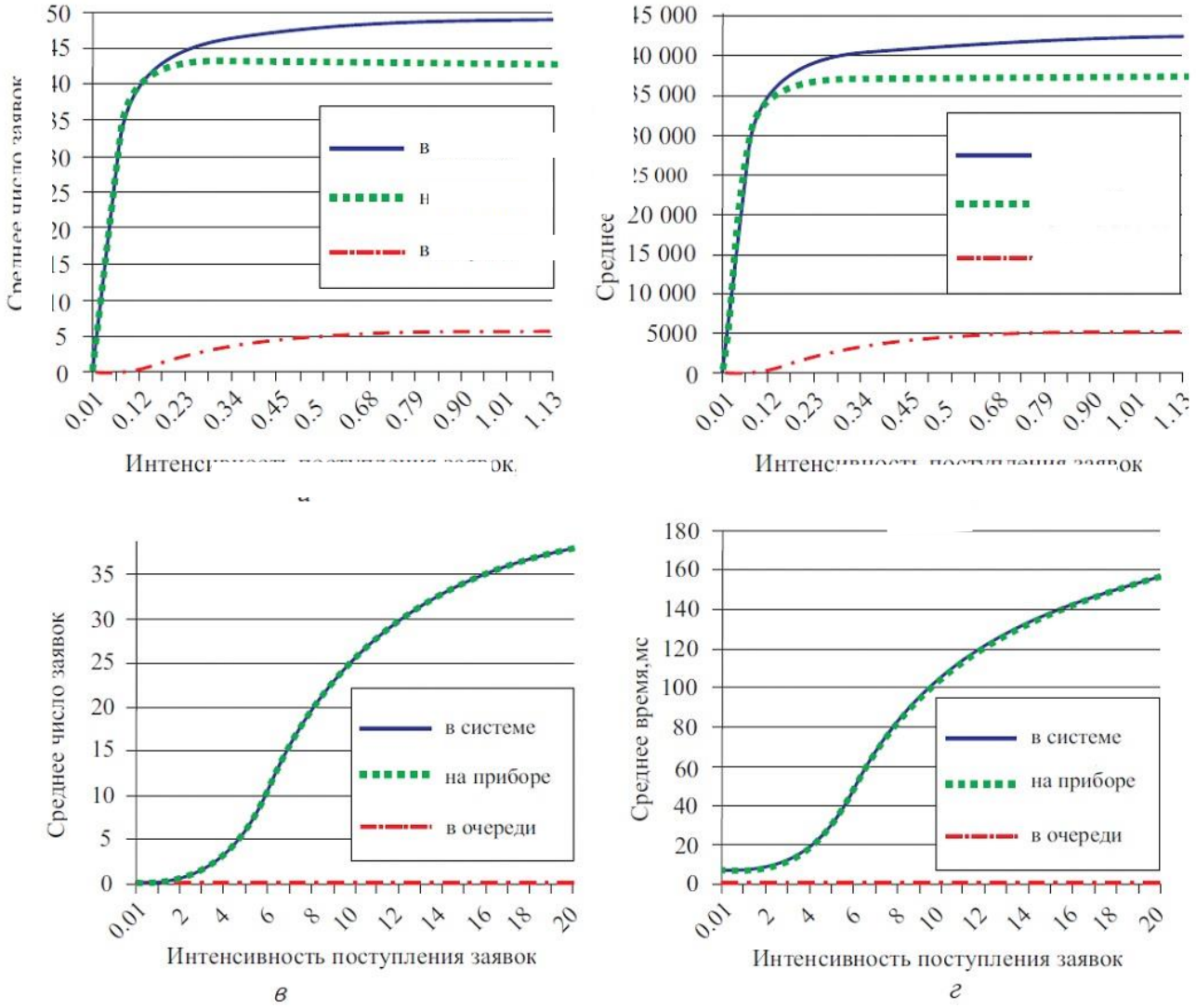


Рис. 1.12. Среднее число и среднее время пребывания пользователя при (а, б) V = 16,76 Мбит/с; (в, г) V = 3500 Мбит/с [106]

1.3. Алгоритм перераспределения ресурсов между сегментами сети

Рассматриваемые в представленной области проблемы исследовались на кафедре прикладной информатики и теории вероятностей РУДН под руководством К.Е. Самуйлова. Начальная постановка задачи была сформулирована на основе научных результатов, представленных в диссертации PhD студента профессора Университета Лиссабона (Португалия) Correia L. [26]. Целью являлось создание системы управления (имитационного симулятора) виртуальными радиоресурсами операторов, которая бы оптимизировала использование сети с учетом SLA [107].

Одновременно с этим начинаются исследования технологии нарезки сети с точки зрения математического моделирования, которые группируются по двум направлениям по схемам распределения ресурсов – исследования по моделированию системы с повторными вызовами и бесконечной орбитой [108, 109], которые в последствии также стали учитывать схемы приоритизации обслуживания [110], резервирования ресурсов (англ. Resource Reservation, RR) [111] и полного разделения (англ. Processor Sharing, PS) [112]. И исследования по моделированию нетерпеливого эластичного трафика, которые представлены в данной диссертационной работе. Отдельно следует отметить исследования, посвященные анализу управления распределением ресурсов при нарезке сети с использованием моделей ресурсных систем массового обслуживания [113]. Особенностью моделей является поступление сигнала, при котором заявка выталкивается с обслуживания и поступает снова в систему с уже новыми требованиями к объему занимаемых ресурсов.

Задача разработки гибких и легко настраиваемых моделей нарезки ресурсов привела к формализации критериев эффективности использования ресурсов таких, например, как справедливость (т.е. равное деление ресурсов каждого сегмента между его пользователями [114]), доступность (процент от времени безотказной работы) [115] и изоляция [116, 117]. Понятие изоляции недостаточно формализовано в спецификациях, однако понимается как отсутствие влияния

быстрого роста трафика в одном сегменте на производительность в других, или их влияние сведено к минимуму, и обеспечивается до тех пор, пока число пользователей сегмента не превысит заданный порог [118, 119]. В случае нехватки свободных ресурсов заявки сегмента сети, нарушившего изоляцию, могут быть вытеснены заявками других сегментов [120].

Изоляция тесно связана с деградацией сегмента, которая может быть определена как состояния системы, в которых один или несколько пользователей приняты на обслуживание, но не обеспечены необходимым количеством ресурсов для выполнения гарантий по скоростям передачи данных. Понятие деградации может применяется как к потоковому (например, временное ухудшение качества видео во время видеоконференции), так и к эластичному трафику (нарушение требований задержки передачи данных) [115]. В моделях с орбитами это состояния системы, когда общий объем занятого ресурса пользователями сегмента меньше порогового значения, а число пользователей в соответствующей орбите больше 0 [68].

Рассмотрим сеть с двумя виртуальными операторами, которые предоставляют доступ к загрузке файлов разного объема $0 < \mu_k^{-1} < \infty, k = 1,2$ и имеют разные гарантии обслуживания b_k, $k = 1,2$. Предполагается, что входящий поток запросов на передачу блоков эластичных данных является пуассоновским потоком с интенсивностями $0 < \lambda_k < \infty, k = 1,2$. Т.к. рассматриваются услуги равного приоритета, ресурс будет разделен на два сегмента в равных долях $V = V_1 + V_2, V_k = V/2$. Запросы попадают в очереди R_1 и R_2 и могут покидать систему по причине нетерпеливости с интенсивностями ε_1 и ε_2. Возможны две схемы нарезки радиоресурсов – статичное и динамическое распределение (в разделе 1.1 изложены отличия). На рис. 1.13 представлена схема модели для статичного случая. Схема для динамического случая в диссертационной работе не приводится, отличием является возможность изменения границы между двумя сегментами, которое обозначается пунктирной линией на рисунке (следовательно,

запрос, находящийся на ожидании во второй очереди, поступит на обслуживание после перераспределения ресурсов).

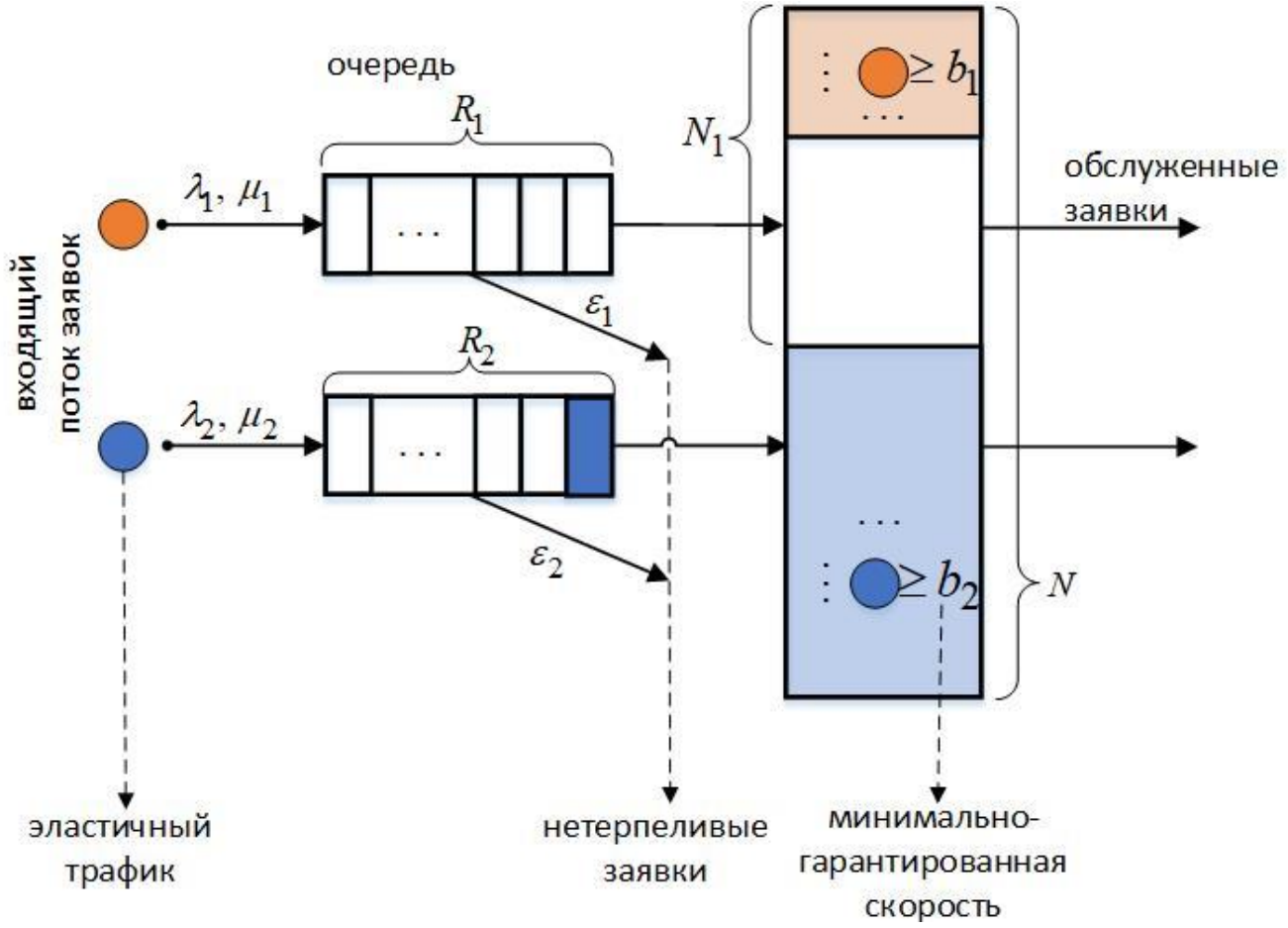


Рис. 1.13. Схема СМО для фиксированного распределения ресурса

Сначала рассматривается статичное распределение ресурсов. Функционирование системы описывается случайным процессом $\mathbf{X}(t)$ с состояниями вида $\boldsymbol{x} = (n_1, n_2)$, где n_1 число запросов 1-оператора, n_2 число запросов 2-оператора в системе. Тогда пространство состояний $\mathbf{X}(t)$ будет иметь вид $\mathcal{X} = \{n_k \in \{0,\ldots,N_k,\ldots,N_k + R_k\} : n_1 + n_2 \leq N,\ k = \{1,2\}\}$, где $N_k = \left\lfloor \frac{V_k}{b_k} \right\rfloor$, $N = N_1 + N_2$ – максимальное число одновременно обсуживающихся запросов k-типа. На основе диаграммы интенсивностей переходов (строится аналогично рис. 1.5) может быть получено стационарное распределение вероятностей (утверждение 1.3).

Утверждение 1.3. [122] Распределение вероятностей в мультипликативном виде для модели с двумя сегментами без управления радиоресурсами вычисляется по формуле

$$p_{n_1 n_2} = \begin{cases} \left(\frac{\lambda_1}{V_1\mu_1}\right)^{n_1}\left(\frac{\lambda_2}{V_2\mu_2}\right)^{n_2} p_{00},\ n_1 = \overline{1, N_1},\ n_2 = \overline{1, N_2}, \\ \frac{\left(\frac{1}{V_1\mu_1}\right)^{N_1}\lambda_1{}^{n_1}}{\prod_{i=1}^{n_1-N_1}\left(V_1\mu_1 + i\varepsilon_1\right)}\frac{\left(\frac{1}{V_2\mu_2}\right)^{N_2}\lambda_2{}^{n_2}}{\prod_{i=1}^{n_2-N_2}\left(V_2\mu_2 + i\varepsilon_2\right)}, \begin{array}{l} n_1 = \overline{N_1+1, N_1+R_1}, \\ n_2 = \overline{N_2+1, N_2+R_2}. \end{array} \end{cases} \tag{1.4}$$

$$p_{00} = \left(\sum_{n_1=0}^{N_1}\sum_{n_2=0}^{N_2}\left(\frac{\lambda_1}{V_1\mu_1}\right)^{n_1}\left(\frac{\lambda_2}{V_2\mu_2}\right)^{n_2} + \left(\frac{1}{V_1\mu_1}\right)^{N_1}\left(\frac{1}{V_2\mu_2}\right)^{N_2} \cdot \right.$$

$$\left. \cdot \sum_{n_1=N_1+1}^{N_1+R_1}\sum_{n_2=N_2+1}^{N_2+R_2}\frac{\lambda_1{}^{n_1}}{\prod_{i=1}^{n_1-N_1}\left(V_1\mu_1 + i\varepsilon_1\right)}\frac{\lambda_2{}^{n_2}}{\prod_{i=1}^{n_2-N_2}\left(V_2\mu_2 + i\varepsilon_2\right)} \right)^{-1}.$$

Доказательство. Используя диаграмму интенсивностей переходов запишем систему уравнений частичного баланса (СУЧБ):

$$\begin{cases} p_{n_1 n_2} V_1\mu_1 = p_{n_1-1, n_2-1}\lambda_1, & 0 < n_1 \le N_1, \\ p_{n_1 n_2}\left(V_1\mu_1 + \left(n_1 - N_1\right)\varepsilon_1\right) = p_{n_1-1, n_2-1}\lambda_1, & n_1 > N_1, \\ p_{n_1 n_2} V_2\mu_2 = p_{n_1-1, n_2-1}\lambda_2, & 0 < n_2 \le N_2, \\ p_{n_1 n_2}\left(V_2\mu_2 + \left(n_2 - N_2\right)\varepsilon_2\right) = p_{n_1-1, n_2-1}\cdot\lambda_2, & n_2 > N_2. \end{cases} \tag{1.5}$$

Далее аналогично утверждению 1.3 выводится распределение вероятностей, выражается $p_{n_1 n_2}$ через p_{n_1-1, n_2-1} и далее рекуррентно приводим к виду (1.4).

Утверждение доказано. □

Далее представлено динамическое распределение ресурсов [123]. Функционирование системы описывается трехмерным Марковским процессом $\mathbf{X}(t) = \left\{N_1(t), N_2(t), R_1(t), R_2(t)\right\}_{t\ge 0}$, где $N_k(t)$ общее число пользователей k-оператора в момент времени t, $R_k(t)$ общее число пользователей, находящихся на ожидании k-оператора в очереди в момент времени t, с состояниями вида

$\boldsymbol{x} = (n_1, n_2, r_1, r_2)$, где n_1 число запросов 1-оператора в системе, n_2 число запросов 2-оператора в системе, r_1 число запросов в очереди 1-оператора, r_2 число запросов в очереди 2-оператора. Следовательно, пространство состояний для этой системы представимо в виде

$$\mathcal{X} = \{(n_1, n_2, m_1, m_2) : n_1 \geq 0, n_2 \geq 0, m_1 \geq 0, m_2 \geq 0;$$
$$(n_1, n_2, 0, 0) : n_1 + n_2 \leq N;$$
$$(n_1, n_2, m_1, m_2) : n_1 + n_2 = N, m_1 \leq M_1, m_2 \leq M_2\}.$$

Запишем систему уравнений равновесия в матричном виде $\vec{p}^T \mathbf{A} = \vec{0}^T$, где $\mathbf{A}$ матрица интенсивностей переходов, $\vec{p}^T$ вектор-строка стационарных вероятностных состояний. Уравнение решается совместно с нормирующим условием. Элементы матрицы интенсивнотей переходов представлены в табл. 1.3.

Табл. 1.3. Интенсивности переходов для СМО с перераспределением ресурса по всем событиям системы [121]

Интенсивность	**Состояние**	**Условия перехода и случайные события**
λ_1	$(n_1+1,\ n_2,\ 0,\ 0)$	Очередь отсутствует. Запрос 1-го типа поступает на 1-ый прибор
λ_2	$(n_1,\ n_2+1,\ 0,\ 0)$	Очередь отсутствует. Запрос 2-го типа поступает на 2-ой прибор
$V\mu_1$	$(n_1-1,\ n_2,\ 0,\ 0)$	Очередь отсутствует. Запрос 1-го типа обслуживается
$V\mu_2$	$(n_1,\ n_2-1,\ 0,\ 0)$	Очередь отсутствует. Запрос 2-го типа обслуживается
λ_1	$(n_1,\ n_2,\ r_1+1,\ r_2)$	Ресурс занят. Запрос 1-го типа пришел в очередь
λ_2	$(n_1,\ n_2,\ r_1,\ r_2+1)$	Ресурс занят. Запрос 2-го типа пришел в очередь
$\frac{V\mu_1}{2}+r_1\varepsilon_1$	$(n_1,\ n_2,\ r_1-1,\ r_2)$	Ресурс занят. Обе очереди заняты. 1) Запрос обслужился. На его место встает запрос из 1-ой очереди. 2) Нетерпеливый запрос
$\frac{V\mu_2}{2}+r_2\varepsilon_2$	$(n_1,\ n_2,\ r_1,\ r_2-1)$	Ресурс занят. Обе очереди заняты. 1) Запрос обслужился. На его место встает запрос из 2-ой очереди. 2) Нетерпеливый запрос
$V\mu_1+r_1\varepsilon_1$	$(n_1,\ n_2,\ r_1-1,\ 0)$	Ресурс занят. Очередь 2-го типа пуста. 1) Обслужится запрос из очереди 1-го типа. 2) Нетерпеливый запрос

$V\mu_2 + r_2\varepsilon_2$	$(n_1,\ n_2, 0,\ r_2 - 1)$	Ресурс занят. Очередь 1-го типа пуста. 1) Обслужится запрос из очереди 2-го типа. 2) Нетерпеливый запрос
$V\mu_1$	$(n_1 - 1,\ n_2 + 1, 0,\ r_2 - 1)$	Ресурс занят. Очередь 1-го пуста. Обслужился запрос 1-го типа, на прибор поступил запрос 2-го типа, так как его очередь не пуста
$V\mu_2$	$(n_1 + 1,\ n_2 - 1,\ r_1 - 1,\ 0)$	Ресурс занят. Очередь 2-го пуста. Обслужился запрос 2-го типа, на прибор поступил запрос 1-го типа, так как его очередь не пуста
$\frac{V\mu_1}{2}$	$(n_1 - 1,\ n_2 + 1,\ r_1,\ r_2 - 1)$	Ресурс занят. Обе очереди заняты. Запрос 1-го типа обслуживается и на его место приходит запрос 2-го типа из очереди, так как вероятность попадания из каждой очереди – $\frac{1}{2}$
$\frac{V\mu_2}{2}$	$(n_1 + 1,\ n_2 - 1,\ r_1 - 1,\ r_2)$	Ресурс занят. Обе очереди заняты. Запрос 2-го типа обслуживается и на его место приходит запрос 1-го типа из очереди, так как вероятность попадания из каждой очереди – $\frac{1}{2}$

Пример 1.3. Далее представлены результаты численного анализа для следующих исходных данных: $V = 10$ Мбит/с, $b = 1$ кбит/с, λ_1 от 1 до 10 1/мин, $\lambda_2 = 5$ 1/мин, $\varepsilon_1 = \varepsilon_2 = 0{,}1$ 1/мин, $\mu_1^{-1} = \mu_2^{-1} = 1$ бит. Рис. 1.14 для модели с динамическим разделением ресурсов отражает, что среднее число пользователей для каждого сегмента будет варьироваться в зависимости от потребностей системы. Для модели со статичным разделением ресурсов повторного распределения ресурсов не произойдет, и будет зафиксировано максимальное число пользователей для каждого сегмента. Из рис. 1.15 следует, что число пользователей в очередях в модели с динамической нарезкой ресурсов будет стремиться к балансу благодаря использованию нарезки. В модели со статичным распределением ресурсов число пользователей будет зависеть от максимальной емкости очереди.

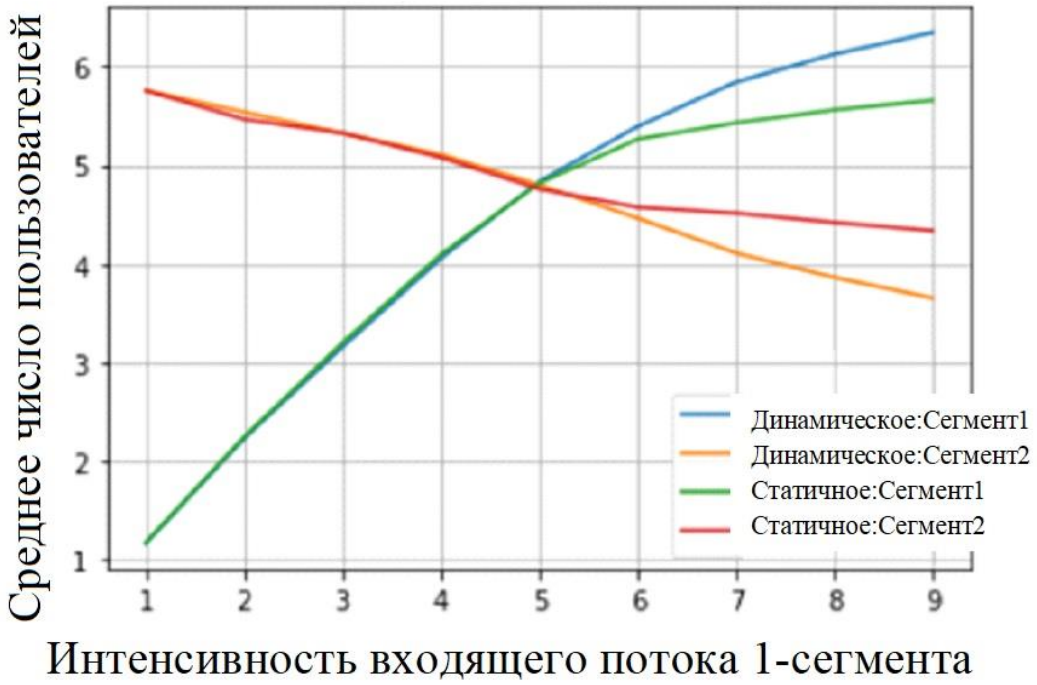


Рис. 1.14. Среднее число пользователей в системе для модели с перераспределением ресурса по всем событиям системы

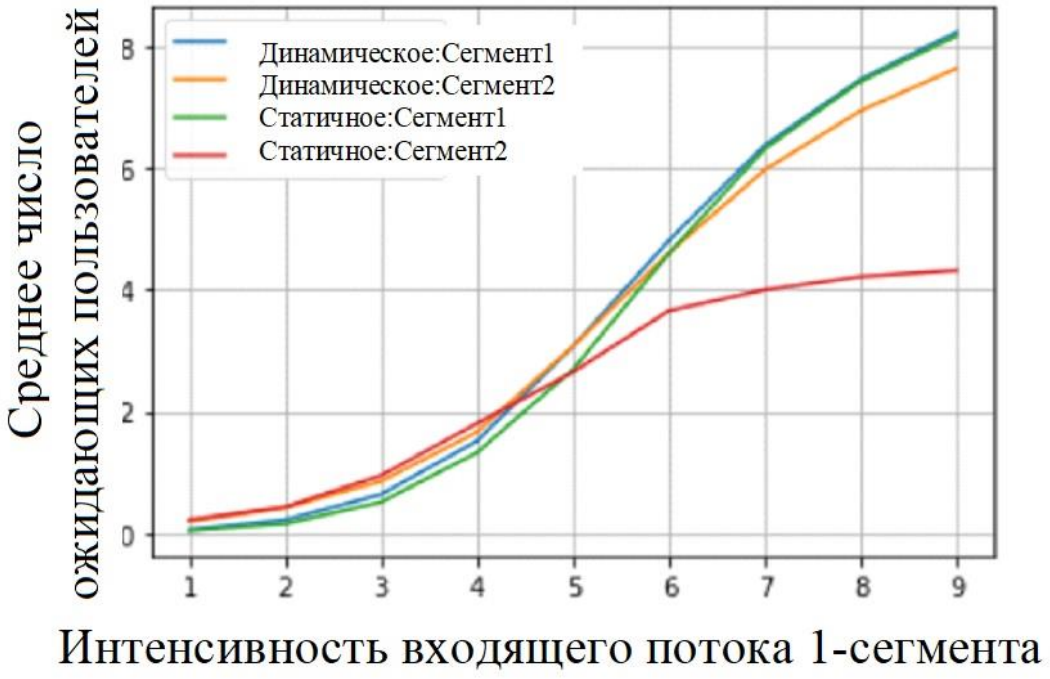


Рис. 1.15. Среднее число пользователей в очереди для модели с перераспределением ресурса по всем событиям системы

1.4. Управляемая система массового обслуживания для доступа к ресурсам

При управлении нарезкой сети необходимо учитывать доступность ресурсов нескольким виртуальным операторам, при этом должен быть гарантирован адекватный уровень производительности. В настоящее время большинство работ в области нарезки ресурсов отсутствует единая стратегия управления ресурсами, способная интегрировать выбор, настройку и управление ресурсами как часть единого объекта [68]. В диссертационной работы предложен динамический подход, при котором создается гибкая модель обслуживания пользователей, способная повысить пропускную способность сети. При этом важно построить модель, которая позволяет динамически перераспределять ресурсы (менять границы сегментов) с помощью различных алгоритмов и целей оптимизации. Следовательно, задачей является построение политики выбора наилучшего с точки зрения заданного критерия использования ресурсов, т.е. количества ресурсов, выделяемых каждому из сегментов. Причем такая политика должна динамически реагировать на изменения в состояниях системы, а не задаваться фиксированным алгоритмом на стадии проектирования системы.

В диссертационной работе для решения этой цели выбран математический аппарат управляемых систем массового обслуживания (УпрСМО). Существует большое количество публикаций теоретического [125–128,131,132] и прикладного характера [130], касающихся вопросов УпрСМО [134–137]. Так аппарат УпрСМО используется для решения различных прикладных задач, например, при исследовании транспортных систем, в задачах организации работы систем обработки информации, а также распределения ресурсов между различными приложениями в области облачных вычислений. Выбранный аппарат УпрСМО применим для решения задачи обслуживания пользователей двух классов веб-приложений, которые развертываются в облаке [138] и распараллеливаются на двух разных серверах.

Схема системы представлена на рис. 1.16, подробное описание функционирования представлено в работе [139], а пример с двумя серверами (два класса пользователей), на каждом из которых развернуто по одной виртуальной машине, показан на рис. 1.17. Задачей является распределение пользователей между двумя серверами таким образом, чтобы максимально сократить потери при ожидании обслуживания в случаях, когда в системе имеются оба класса ожидающих начала обслуживания пользователей. Каждый раз при изменении состояния системы в зависимости от количества запросов, находящихся на ожидании, и от того виртуальная машина какого сервера освободилась, принимается решение, из какой очереди взять на обслуживание пользователя.

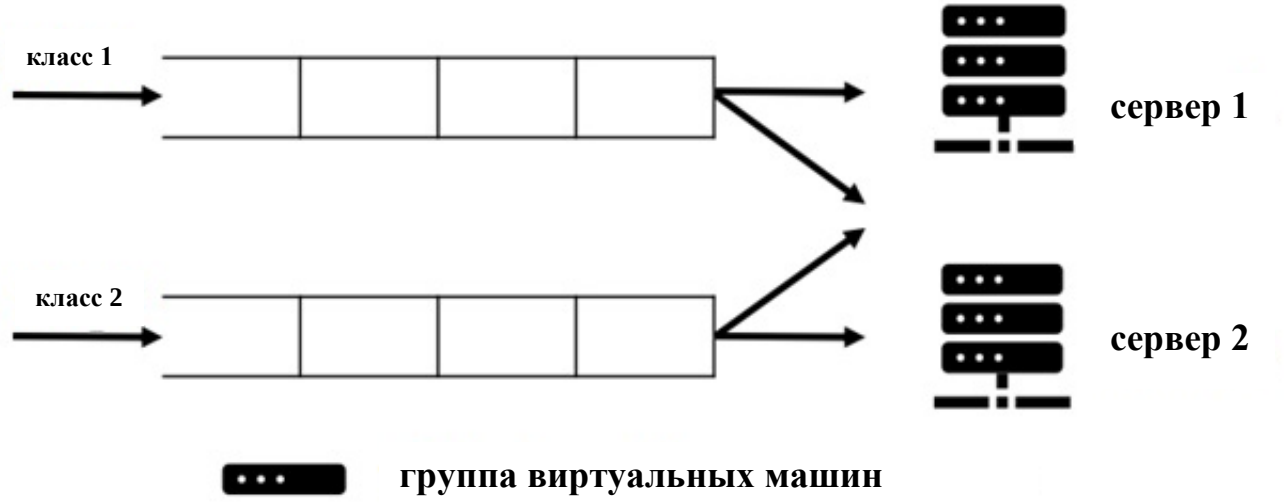


Рис. 1.16. Схема модели управляемого занятия ресурсов в среде облачных вычислений

Т.е. используя аппарат УпрСМО необходимо найти оптимальную политику распределения пользователей при минимизации стоимости обслуживания и ожидания начала обслуживания. Обозначим: k – сервер, $k=\{1;2\}$; N_k – число приборов (виртуальных машин) на k-сервере; μ_k – интенсивность обслуживания k-сервера (экспоненциальное распределение); j – класс запроса/ пользователя, $j=\{1;2\}$; λ_j – интенсивность входящего j-потока (пуассоновский поток); q_1, q_2 – очередь пользователей; c_{k0} – стоимость ожидания пользователя в k-очереди; c_{k1} – стоимость обслуживания k-пользователя на своем приборе; c_{k2} – стоимость

обслуживания k -пользователя на альтернативном приборе; $Q_k(t)$ – число запросов в k -очереди в момент времени t и $D_{kj}(t)$ – число j -запросов на k -сервере в момент времени t. Отсюда, в произвольный момент времени d_{11} – число 1- типа пользователей на приборах 1-сервера; d_{12} – число 2-типа пользователей на приборах 1-сервера; d_{21} – число первого типа пользователей на приборах 2-сервера; d_{22} – число второго типа пользователей на приборах 2-сервера. Изображение модели в виде системы массового обслуживания с двумя очередями и перекрестным обслуживанием с дополнительными штрафами представлено на рис. 1.18.

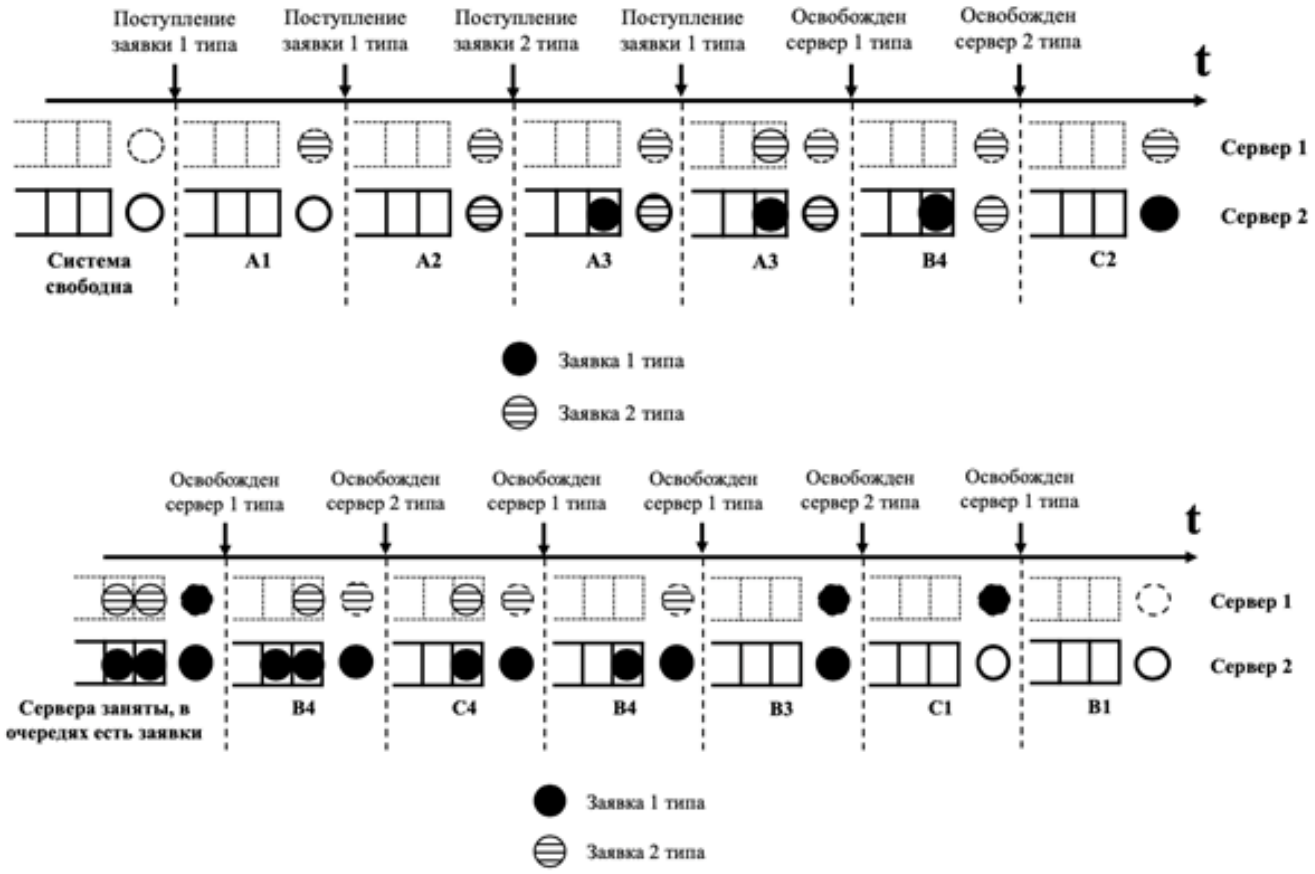


Рис. 1.17. Изменение состояний модели управляемого занятия ресурсов во времени

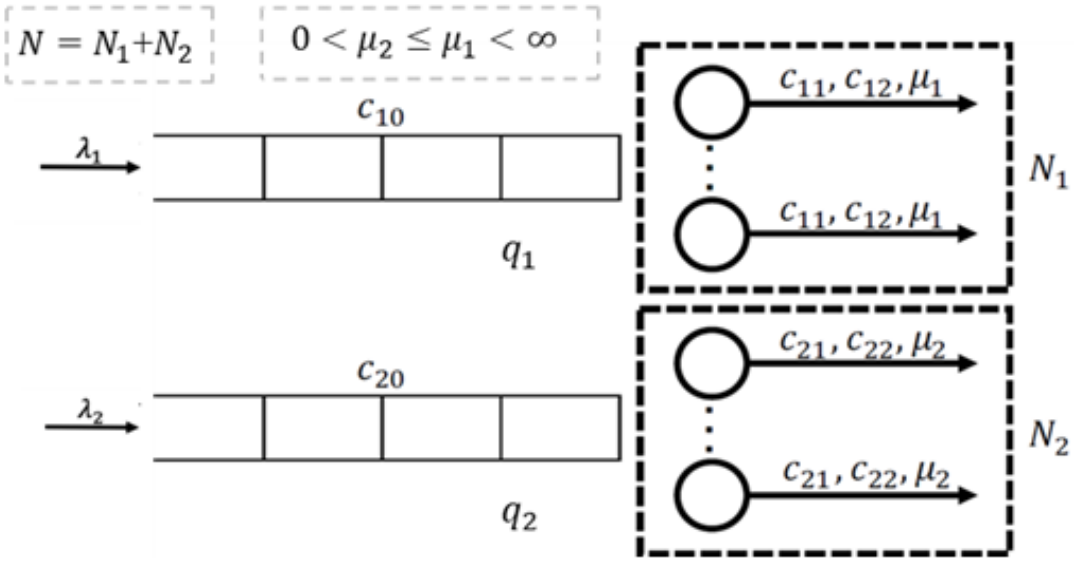


Рис. 1.18. Схема управляемой СМО для управляемого занятия ресурсов

Функционирование системы описывается многомерным Марковским случаным процессом $\boldsymbol{X}(t) = \{Q_1(t), Q_2(t), D_{11}(t), D_{12}(t), D_{21}(t), D_{22}(t)\}$ – число запросов, находящихся в системе в момент времени $t \ge 0$ над пространством состояний

$$\mathcal{X} = \{\boldsymbol{x} = (q_1, q_2, d_{11}, d_{12}, d_{21}, d_{22}) : d_{kj} \ge 0,\ q_k \ge 0,\ k,\ j = 1,2;$$
$$(0,0,d_{11},d_{12},d_{21},d_{22}) : d_{k1} + d_{k2} \le N_k;$$
$$q_1 + q_2 > 0 : d_{k1} + d_{k2} = N_k, k = 1,2\}.$$

Матрица интенсивностей переходов составляется в соответствии с представленными выше правилами при $q_1 + q_2 = 0$. В случаях, когда обе очереди заняты, $d_{kj} > 0,\ q_1 + q_2 \ge 1$, в соответствии с функцией выбора очереди в фиксированном состоянии $\boldsymbol{x}$ при обслуживании j-типа пользователей на приборах k-сервера

$$f_{kj}(\boldsymbol{x}) \in \{1,2\}, \boldsymbol{x} \in \mathcal{X} : q_1 + q_2 > 0. \qquad (1.6)$$

Обозначим $\boldsymbol{f}(\boldsymbol{x}) = (f_{11}(\boldsymbol{x}), f_{12}(\boldsymbol{x}), f_{21}(\boldsymbol{x}), f_{22}(\boldsymbol{x}))$ – вектор политики выбора очередей при различных значениях k и j. Назовем политикой маршрутизации вектор

$$f = \boldsymbol{f} = (\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X} : q_1 + q_2 > 0) \qquad (1.7)$$

из четырех компонент таких моделей управления выбором очередей. Таким образом, если определить фиксированную политику $\boldsymbol{f}$, можно составить соответствующую систему уравнений равновесия и найти распределение вероятностей $\pi^{\boldsymbol{f}}(\boldsymbol{x}) = \mathrm{P}\left[\boldsymbol{X}^{\boldsymbol{f}}(t) = \boldsymbol{x}\right]$.

Утверждение 1.4. [140] Функция среднего вознаграждения в состоянии $\boldsymbol{x} \in \mathcal{X}$ вычисляется по формуле

$$c(\boldsymbol{x}) = \sum_{k=1}^{2} \left(c_{k0} q_k + c_{k1} d_{k1} + c_{k2} d_{k2}\right). \tag{1.8}$$

Доказательство. Исходя из выбранного критерия оптимизации сокращения среднего числа запросов, рассчитывается стоимость обслуживания запросов на своих и альтернативных приборах с учетом дополнительных штрафов $c_{k1} d_{k1}$ и $c_{k2} d_{k2}$, а также стоимости ожидания начала обслуживания в своих очередях $c_{k0} q_k$. Просуммировав данные значения по каждому типу сервера, получим функцию среднего вознаграждения (1.8) в конкретном состоянии $\boldsymbol{x} \in \mathcal{X}$.

Утверждение доказано. □

Далее приводятся основные показатели эффективности модели: среднее число запросов каждого типа в очереди $\bar{Q}_k = \sum_{\boldsymbol{x} \in \mathrm{X}} q_k \pi^{\boldsymbol{f}}(\boldsymbol{x})$; среднее число приборов, обслуживающих 1 и 2-классы пользователей $\bar{C}_j = \sum_{\boldsymbol{x} \in \mathrm{X}} \left(d_{11} + d_{12} + d_{21} + d_{22}\right) \pi^{\boldsymbol{f}}(\boldsymbol{x})$; и среднее число пользователей в системе

$$\bar{N} = \sum_{j=1}^{2} \left(\bar{Q}_j + \bar{C}_j\right).$$

Пример 1.4. Для анализа модели строится имитационная модель в среде AnyLogic (рис. 1.19), основанной на использовании объектно-ориентированного языка Java. Описание элементов симулятора: source1 – входящий поток 1-типа; source2 – входящий поток 2-типа; TS1, TS2, TS, TS3 – элементы для разметки запросов временными метками; queue1, queue2 – 1- и 2-тип очереди; TE1, TE2 – элементы

для чтения меток; selectOutput1, selectOutput2 – элементы, которые распределяются по группам инструментов согласно SQL-запросам; dalay1, delay2 – группы приборов; delA – состояние занятости 1-типа группы приборов; delS – состояние занятости 2-типа группы приборов; qu1 – текущее состояние 1-типа очереди; qu2 – текущее состояние 2-типа очереди. Пусть в системе развернуто $N_1 = N_2 = 5$ виртуальных машин на каждом сервере. Предположим, что входящие потоки 1- и 2-типов равны $\lambda_1 = \lambda_2 = 30$. Поскольку первый сервер работает быстрее, интенсивность обслуживания на приборах 1-сервера $\mu_1 = 20$, а на 2-сервере $\mu_2 = 5$. Также имеются две очереди бесконечной емкости.

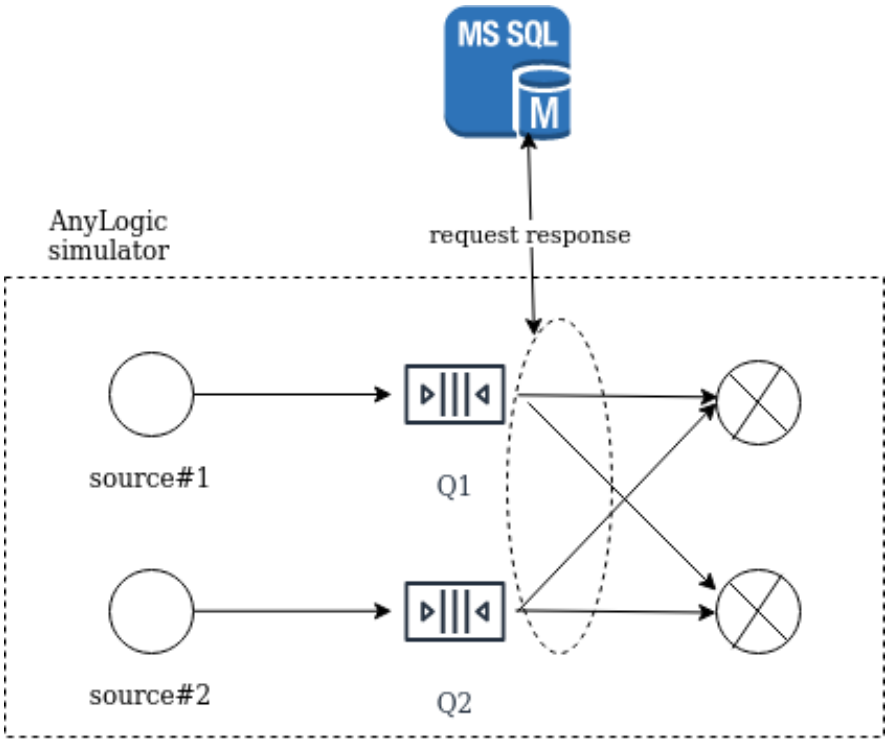


Рис. 1.19. Схема структуры имитационной модели управляемого занятия ресурсов

Схема модели с двумя очередями и перекрестным обслуживанием с дополнительными штрафами представлены на рис. 1.20. Поскольку входные данные для симулятора представляют собой фиксированную политику маршрутизации, в примере исследуется фиксированный вариант распределения пользователей между двумя группами приборов. На основании полученных результатов в экпоненциальном и нормальном предположениях (табл. 1.4) можно сделать вывод, что в системе практически отсутствуют запросы, ожидающие

начала обслуживания, следовательно используемую политику управления занятия ресурсов можно считать оптимальной с данной точки зрения. Поток событий является простейшим пуассоновским потоком, тем не менее, для моделирования и анализа такой системы он приемлем.

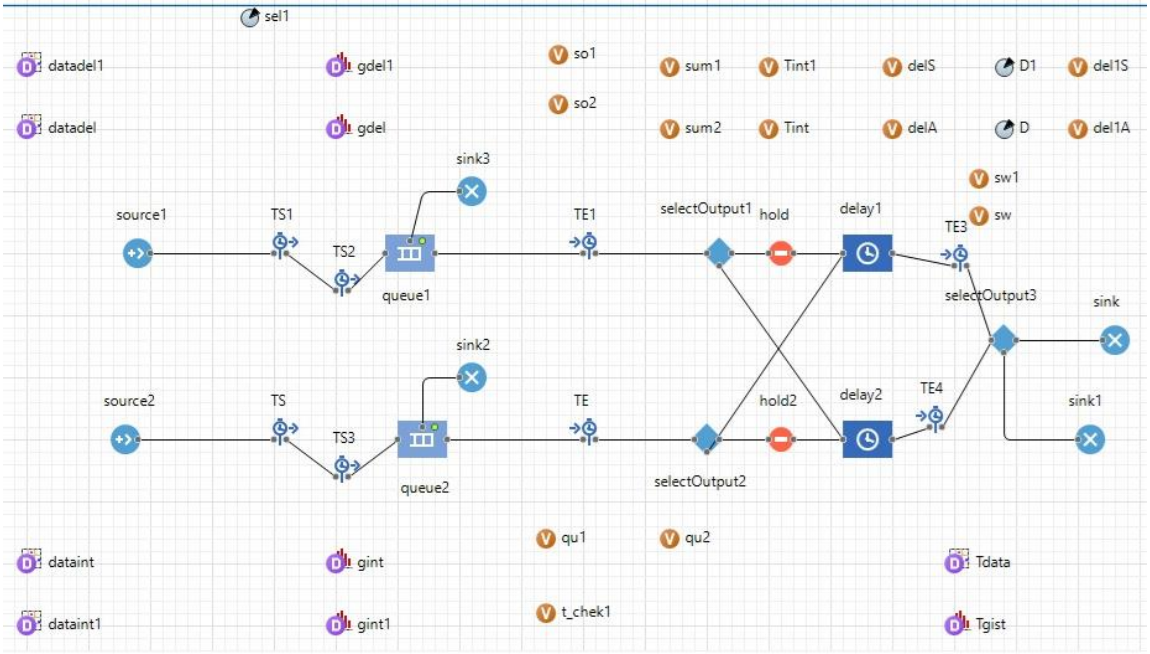


Рис. 1.20. Схема имитационной модели управляемого занятия ресурсов в Anylogic

Табл. 1.4. Характеристики модели с управляемым занятием ресурсов [140]

Показатель	**Значение**	
	Exp (30)	**Norm (30; 0,001)**
средняя длина 1-очереди	0,206	0,289
средняя длина 2-очереди	0,153	0,168
среднее число пользователей, обслуживаемых на 1-группе виртуальных машин	3,85	3,87
среднее число пользователей, обслуживаемых на 2-группе виртуальных машин	3,91	3,94
среднее число 1-типа пользователей в системе	4,11	4,18
среднее число 2-типа пользователей в системе	4,08	4,10
среднее время ожидания 1-типа пользователей	0,007	0,008
среднее время ожидания 2-типа пользователей	0,005	0,004
среднее время нахождения в системе пользователей, обслуживаемых на 1-группе виртуальных машин	0,0503	0,0503
среднее время нахождения в системе пользователей, обслуживаемых на 2-группе виртуальных машин	0,199	0,199
среднее время нахождения в системе 1-типа пользователей	0,068	0,070
среднее время нахождения в системе 2-типа пользователей	0,133	0,135

1.5. Постановка задачи исследования

Исследования, проведенные в разделах 1.1–1.4 диссертационной работы, показали, что в области анализа моделей нарезки ресурсов ранее в основном рассматривались системы с фиксированной политикой управления ресурсами, либо с использованием методов машинного обучения для перераспределения ресурсов. При этом перераспределение могло произойти в любой момент времени при изменении состояния системы. Схему распределения ресурсов базового оператора между двумя виртуальными операторами/сегментами можно представить в виде рис. 1.21. Как видно из схемы, нерешенной задачей является планирование ресурсов при изменении нагрузки на сеть, когда возникают ситуации простоя одного сегмента, при наличии ожидающих запросов другого сегмента. Проверка необходимости перераспределения ресурсов при каждом изменении состояния системы может, конечно, позволить гибко настроить систему, однако увеличит сигнальную нагрузку на сеть. Поэтому в диссертационной работе рассматривается система мониторинга, контроллер, который направляет сигналы о проверке системы с некоторой периодичностью.

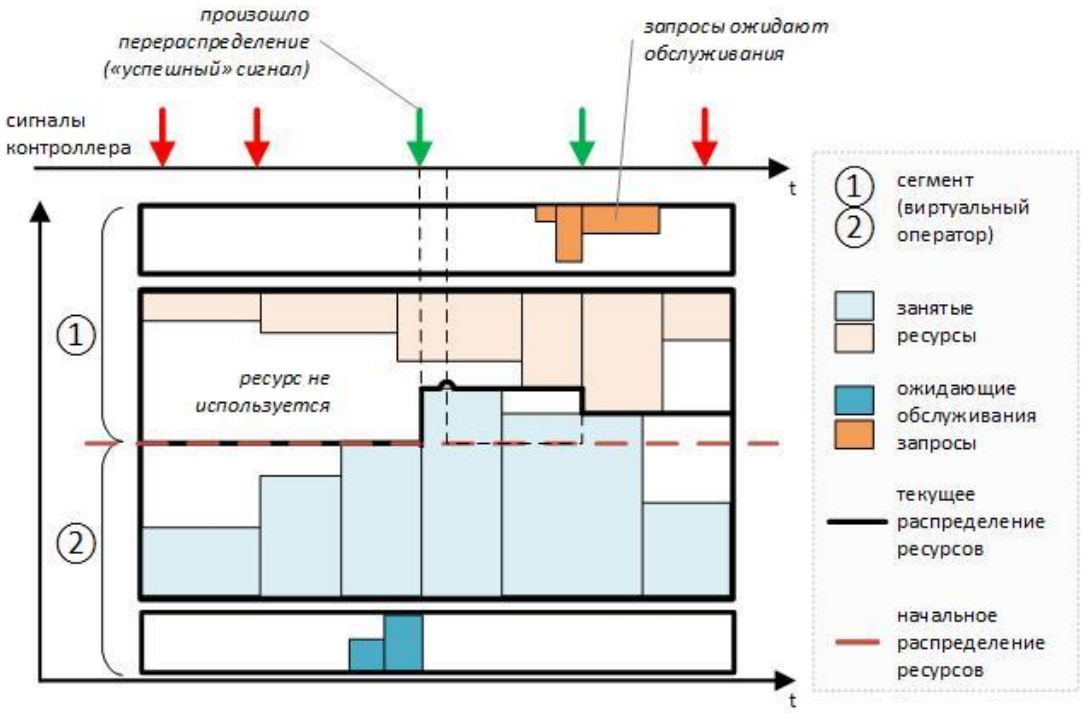


Рис. 1.21. Динамическое перераспределение ресурса по сигналу

При внедрении динамического перераспределения ресурсов по сигналу контроллера возникает задача настройки частоты поступления этих сигналов таким образом, чтобы повысить пропускную способность сети. Ранее в системах массового обслуживания, применявшихся для моделирования нарезки ресурсов, исследовались показатели обслуживания пользователей виртуального операторов. Однако, в диссертационной работе предлагается учитывать показатели эффективности самой нарезки сети с точки зрения занятости ресурса, соответствия распределения ресурса соглашению о качестве обслуживания, вероятности перераспределения ресурса по сигналу. При этом инициация перераспределения ресурсов только по сигналу контроллера накладывает на себя некоторые ограничения, связанные с вопросом выбора нового объема ресурсов, предоставляемого виртуальным операторам (по какому правилу и на сколько двигать границу между сегментами сети).

В диссертационной работе модель строится в виде системы массового обслуживания с нетерпеливым эластичным трафиком с минимально-гарантированной скоростью обслуживания и потоком сигналов, которые моделируют сообщения от контроллера, схематическое изображение которой представлено на рис. 1.22. Все обозначения для описания функционирования системы введены выше и также отражены на схеме. В качестве услуг, которые предоставляются виртуальными операторами, в диссертационной работе рассматриваются услуги передачи данных, для которых требуется соблюдать минимальную скорость и имеются ограничения на время ожидания начала обслуживания.

Таким образом, цель диссертационной работы состоит в разработке моделей с нетерпеливым эластичным трафиком и минимальной скоростью передачи для анализа и расчета показателей эффективности динамической нарезки радиоресурсов по сигналам в беспроводной сети. Для достижения цели необходимо решить две задачи:

1. Разработка моделей нарезки сети с нетерпеливым эластичным трафиком и минимальной скоростью передачи и двумя стратегиями перераспределения

ресурса по сигналам – фиксированной и с управлением выбором объема ресурса.

2. Анализ и разработка алгоритмов расчета показателей эффективности нарезки сети, отражающих занятость ресурса, соответствие распределения ресурса соглашению о качестве обслуживания, вероятность перераспределения ресурса по сигналу, а также влияние на показатели частоты поступления сигналов.

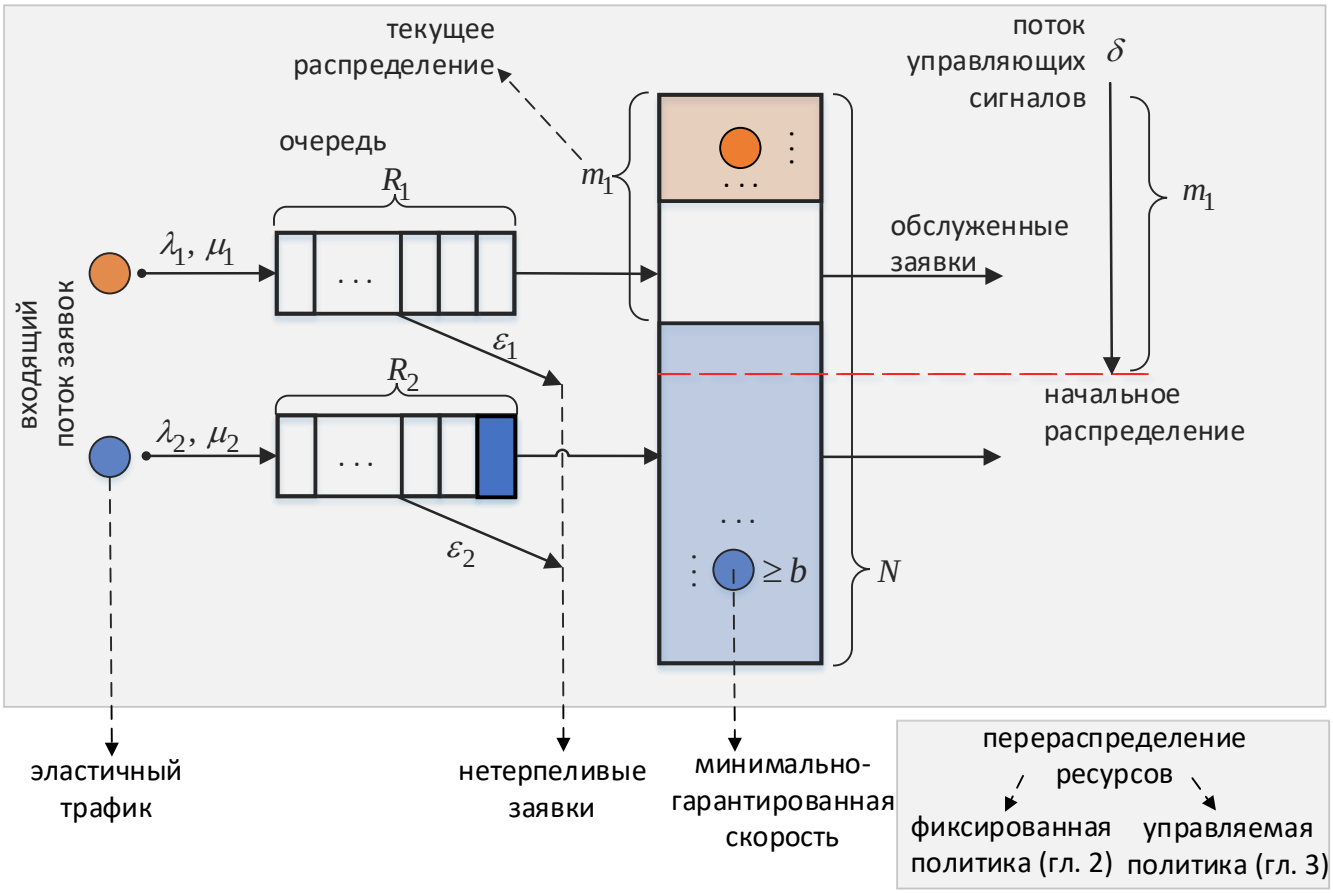


Рис. 1.22. Схема СМО для динамического перераспределения ресурса по сигналу

Логика изложения материала в работе с учетом поставленных задач будет строиться следующим образом (табл. 1.5). Для каждой из задач строятся модели обслуживания эластичного трафика с сигналами контроллера и нетерпеливыми запросами, отличающиеся числом сегментов и политикой распределения ресурсов – фиксированной и управляемой. Различные способы анализа и алгоритмы позволяют оценить предложенные принципы нарезки ресурсов, отраженные в моделях с различных аспектов (в показателях эффективности, в функции

вознаграждения для управляемой СМО, и в показателях эффективности со стороны базового оператора).

Табл. 1.5. Структура диссертационной работы

	Особенности моделей (задача 1)	**Способы анализа (задача 2)**
Результат 1	Модель СМО для **двух** классов **эластичного** трафика с **сигналами** и нетерпеливыми запросами с **фиксированным** выбором объема ресурса (глава 2).	**Матричный рекуррентный алгоритм** расчета стационарного распределения вероятностей (глава 2 – раздел 2.3); **принципы нарезки** ресурсов отражены в рассчитываемых показателях эффективности (глава 2 – раздел 2.4).
Результат 2	Модель СМО для **двух** классов **эластичного** трафика с **сигналами** и нетерпеливыми запросами с **управлением** выбором объема ресурса (глава 3 – разделы 3.1-3.3).	**Итерационный алгоритм** вычисления оптимальной стратегии управлением выбором объема ресурса (глава 3 – раздел 3.3); **принципы нарезки** ресурсов отражены в функции вознаграждения для УпрСМО (глава 3 – раздел 3.2).
Результат 3	Модель для **произвольного** числа сегментов с **фиксированным** алгоритмом перераспределения ресурса (глава 3 – разделы 3.4-3.5)	Дискретно-событийная **имитационная** модель (глава 3 – раздел 3.4); **принципы нарезки** ресурсов отражены в рассчитываемых показателях эффективности нарезки ресурсов базового оператора (глава 3 – раздел 3.5).

ГЛАВА 2
МОДЕЛЬ С ФИКСИРОВАННОЙ ПОЛИТИКОЙ ПЕРЕРАСПРЕДЕЛЕНИЯ РЕСУРСА

2.1. Построение модели с эластичным трафиком и сигналами

В главе 2 получен результат №1, сформулированный в разделе 1.5, а также заложены основы для получения результатов №2 и №3. Основной целью данной главы является построение модели управления ресурсами для двух сегментов с фиксированным выбором объема ресурса. В отличие от модели, исследованной в разделе 1.3, текущая модель включает контроллер, который отправляет поток сигналов на проверку необходимости перераспределения ресурса. Таким образом, перераспределение ресурсов инициируется не в любой момент времени при изменении системы.

Перейдем к построению математической модели для двух классов эластичного трафика, между которыми динамически перераспределяется ресурс V, в виде системы массового обслуживания с сигналами (рис. 2.1). Обозначим $N/2$ начальное (равное) распределение ресурса для 1- и 2-класса трафика, а m_1 (и $N - m_1$) распределение ресурса в некоторый момент времени t. Предположим, что входящие потоки запросов на передачу трафика являются пуассоновскими с интенсивностями λ_1 и λ_2, объем трафика распределен по экспоненциальному закону с параметрами μ_1 и μ_2. С учетом порога b скорости для передачи трафика максимальное число обслуживаемых сессий составляет $N = \lfloor V/b \rfloor$, а число мест в очереди для ожидающих начала обслуживания сессий R_1 и R_2. Пусть пороги на время ожидания начала обслуживания сессий распределены по экспоненциальному закону с параметрами ε_1 и ε_2, а поток сигналов, управляющий перераспределением ресурса, является пуассоновским с интенсивностью δ. Как уже было указано ранее, этим перераспределением управляет контроллер, который

направляет сигналы, по которым осуществляется проверка необходимости перераспределения ресурса. Далее остановимся подробнее на механизме управления доступом эластичного трафика к ресурсу.

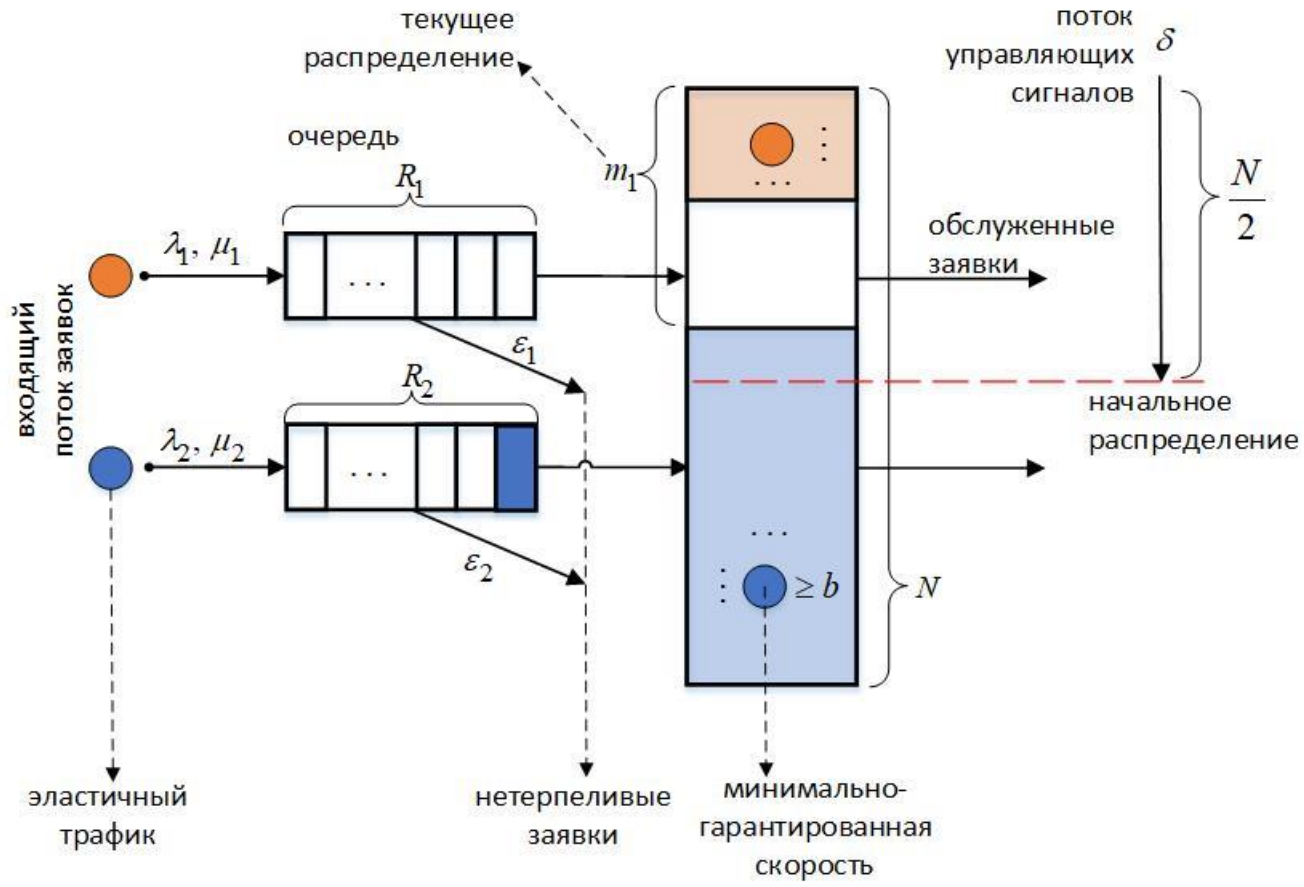


Рис. 2.1. Схема СМО с фиксированной политикой перераспределения ресурса по сигналу

Управление доступом (англ. Admission Control). В случае поступления сессии,

- если ожидающих обслуживания сессий нет и ресурсы, соответствующие данному классу сессии, свободны, то можно инициировать обслуживание сессии;
- если ресурсы, соответствующие данному классу сессии, заняты, то сессия ожидает обслуживания в очереди.

В случае обслуживания сессии,

- если ожидающих обслуживания сессий, соответствующих классу обслужившейся сессии, нет, то освободившийся ресурс простаивает;

- если есть ожидающие обслуживания сессии, соответствующие классу обслужившейся сессии, то на освободившийся ресурс поступает на обслуживание сессия из очереди.

В случае превышения времени ожидания сессии могут покинуть очередь.

Управление перераспределением ресурсов (англ. Reallocation Control). При поступлении сигнала контроллера распределение ресурсов не будет изменено, если

- система пуста;
- не все ресурсы заняты обслуживанием (есть свободные ресурсы для 1- и 2-классов трафика);
- все ресурсы заняты обслуживанием (нет свободных ресурсов для 1- и 2-классов трафика), нет ожидающих обслуживания сессий;
- все ресурсы заняты обслуживанием (нет свободных ресурсов для 1- и 2-классов трафика), есть ожидающие обслуживания сессии одного из классов трафика;
- все ресурсы заняты обслуживанием (нет свободных ресурсов для 1- и 2-классов трафика), есть ожидающие обслуживания сессии 1- и 2-классов.

При поступлении сигнала контроллера распределение радиоресурсов будет изменено, если

- ресурсы для 1-класса трафика заняты обслуживанием, есть ожидающие обслуживания сессии 1-класса, ресурсы для 2-класса трафика свободны (свободные ресурсы для 2-класса трафика перераспределяются для 1-класса);
- ресурсы для 1-класса трафика заняты обслуживанием, есть ожидающие обслуживания сессии 1-класса, не все ресурсы для 2-класса трафика заняты обслуживанием (свободные ресурсы для 2-класса трафика перераспределяются для 1-класса);
- ресурсы для 2-класса трафика заняты обслуживанием, есть ожидающие обслуживания сессии 2-класса, ресурсы для 1-класса трафика свободны (свободные ресурсы для 1-класса трафика перераспределяются для 2-класса);

– ресурсы для 2-класса трафика заняты обслуживанием, есть ожидающие обслуживания сессии 2-класса, не все ресурсы для 1-класса трафика заняты обслуживанием (свободные ресурсы для 1-класса трафика перераспределяются для 2-класса).

Функционирование системы описывает случайный процесс $\mathbf{X}(t)$ с состояниями вида $\boldsymbol{x} = (m_1, m_2, n_1, n_2, r_1, r_2)$, где m_k порог на максимальное число обслуживаемых сессий k-класса, n_k число обслуживаемых сессий k-класса, r_k число ожидающих начала обслуживания сессий k-класса, $k = 1, 2$, над пространством состояний

$$\begin{aligned}
&\mathcal{X} = \{(m_1, N - m_1, n_1, n_2, 0, 0) : 0 \le m_1 \le N,\ 0 \le n_1 \le m_1,\ 0 \le n_2 \le m_2\} \cup \\
&\cup \{(N - m_2, m_2, N - m_2, n_2, r_1, 0) : 0 \le m_2 \le N,\ 0 \le n_2 \le m_2,\ 0 < r_1 \le R_1\} \cup \\
&\cup \{(m_1, N - m_1, n_1, N - m_1, 0, r_2) : 0 \le m_1 \le N,\ 0 \le n_1 \le m_1,\ 0 < r_2 \le R_2\} \cup \\
&\cup \{(m_1, N - m_1, m_1, N - m_1, r_1, r_2) : 0 \le m_1 \le N,\ 0 < r_1 \le R_1,\ 0 < r_2 \le R_2\}.
\end{aligned} \tag{2.9}$$

Алгоритм 2.1. Управление перераспределением ресурсов между двумя виртуальными операторами по фиксированной стратегии задается как

$\delta\downarrow,\ \boldsymbol{x} = (m_1, m_2, n_1, n_2, r_1, r_2)$

▷ *ресурсы второго сегмента простаивают*

1: **if** $(r_1 > 0, n_2 < m_2)$ **then**

2: **if** $(r_1 \le m_2 - n_2)$ **then** $m_1 \leftarrow m_1 + r_1, m_2 \leftarrow m_2 - r_1$

3: **else** $m_1 \leftarrow m_1 + m_2 - n_2, m_2 \leftarrow n_2$

▷ *ресурсы первого сегмента простаивают*

4: **else if** $(r_1 = 0, n_1 < m_1)$ **then**

5: **if** $(r_2 \le m_1 - n_1)$ **then** $m_1 \leftarrow m_1 - r_2, m_2 \leftarrow m_2 + r_2$

6: **else** $m_2 \leftarrow m_2 + m_1 - n_1, m_1 \leftarrow n_1$

▷ *ресурсов достаточно или оба сегмента перегружены*

7: **else** $m_1 \leftarrow m_1,\ m_2 \leftarrow m_1$

В табл. 2.1 перечислены интенсивности переходов между состоянием $\boldsymbol{x}$ и другими состояниями системы $\boldsymbol{x}'$, управление перераспределением ресурсов отражено в строках 4.

Табл. 2.1. Интенсивности переходов для модели с фиксированной политикой перераспределения ресурса по сигналу [142]

№ п/п	Интенсивность события	Условие на $\boldsymbol{x}$	Состояние $\boldsymbol{x}'$
1а-1	λ_1	$n_1+1\le m_1$	$(m_1, N-m_1, n_1+1, n_2, 0, r_2)$
1б-1	λ_1	$n_1+1>m_1,\ r_1+1\le R_1$	$(m_1, N-m_1, m_1, n_2, r_1+1, r_2)$
1а-2	λ_2	$n_2+1\le m_2$	$(N-m_2, m_2, n_1, n_2+1, r_1, 0)$
1б-2	λ_2	$n_2+1>m_2,\ r_2+1\le R_2$	$(N-m_2, m_2, n_1, m_2, r_1, r_2+1)$
2а-1	$\frac{m_1}{N}V\mu_1$	$r_1>0$	$(m_1, N-m_1, m_1, n_2, r_1-1, r_2)$
2б-1	$\frac{m_1}{N}V\mu_1$	$r_1=0,\ n_1>0$	$(m_1, N-m_1, n_1-1, n_2, 0, r_2)$
2а-2	$\frac{m_2}{N}V\mu_2$	$r_2>0$	$(N-m_2, m_2, n_1, m_2, r_1, r_2-1)$
2б-2	$\frac{m_2}{N}V\mu_2$	$r_2=0,\ n_2>0$	$(N-m_2, m_2, n_1, n_2-1, r_1, 0)$
3а-1	$r_1\varepsilon_1$	$r_1>0$	$(m_1, N-m_1, m_1, n_2, r_1-1, r_2)$
3а-2	$r_2\varepsilon_2$	$r_2>0$	$(N-m_2, m_2, n_1, m_2, r_1, r_2-1)$
4а-1	δ	$n_1=m_1,\ r_1>0,$ $n_2<m_2,\ r_2=0,$ $r_1\le m_2-n_2$	$(m_1+r_1, m_2-r_1, m_1+r_1, n_2, 0, 0)$
4б-1	δ	$n_1=m_1,\ r_1>0,$ $n_2<m_2,\ r_2=0,$ $r_1>m_2-n_2$	$\begin{pmatrix} m_1+m_2-n_2, n_2, \\ m_1+m_2-n_2, n_2, \\ r_1-m_2+n_2, 0 \end{pmatrix}$
4а-2	δ	$n_2=m_2,\ r_2>0,$ $n_1<m_1,\ r_1=0,$ $r_2\le m_1-n_1$	$(m_1-r_2, m_2+r_2, n_1, m_2+r_2, 0, 0)$
4б-2	δ	$n_2=m_2,\ r_2>0,$ $n_1<m_1,\ r_1=0,$ $r_2>m_1-n_1$	$\begin{pmatrix} n_1, m_2+m_1-n_1, \\ n_1, m_2+m_1-n_1, \\ 0, r_2-m_1+n_1 \end{pmatrix}$

Пример 2.1. Для заданных пороговых значений $N=2, R_1=1, R_2=1$ построим диаграмму интенсивностей переходов (рис. 2.2) и матрицу интенсивностей переходов $\mathbf{\Lambda}$ (рис. 2.3).

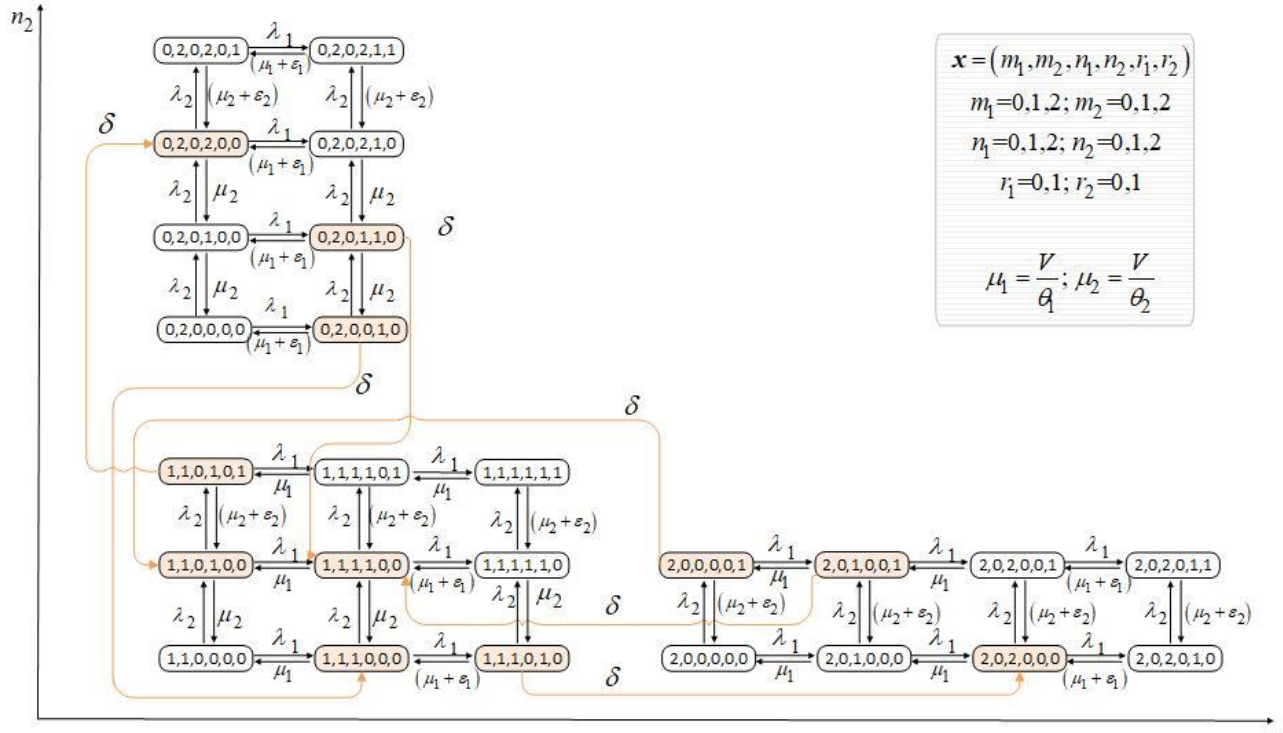


Рис. 2.2. Пример диаграммы интенсивностей переходов для СМО с фиксированной политикой перераспределения ресурса по сигналу при $N=2, R_1=1, R_2=1$

Рис. 2.3. Фрагмент матрицы интенсивностей переходов для СМО с фиксированной политикой перераспределения ресурса по сигналу при $N=2, R_1=1, R_2=1$

2.2. Блочная трехдиагональная матрица интенсивностей переходов

Далее сформулировано и доказано утверждение, позволяющее найти стационарное распределение вероятностей $\boldsymbol{\pi\Lambda}=\mathbf{0},\ \boldsymbol{\pi}\mathbf{e}^T=1$ (раздел 2.3), где $\boldsymbol{\Lambda}$ – матрица интенсивностей переходов.

Утверждение 2.1. В случае если на множестве $\mathbf{X}(t)$ введен следующий лексикографический порядок

$$\begin{aligned}
&\boldsymbol{x}'=(m_1',m_2',n_1',n_2',r_1',r_2')\succ(m_1,m_2,n_1,n_2,r_1,r_2)=\boldsymbol{x}\Leftrightarrow\\
&n_1'+n_2'+r_1'+r_2'>n_1+n_2+r_1+r_2,\\
&n_1'+n_2'+r_1'+r_2'=n_1+n_2+r_1+r_2,\quad m_1'>m_1,\\
&n_1'+n_2'+r_1'+r_2'=n_1+n_2+r_1+r_2,\quad m_1'=m_1,\quad n_1'+r_1'>n_1+r_1,
\end{aligned}\tag{2.10}$$

то матрица интенсивностей переходов СП $\mathbf{X}(t)$ представима в блочном трехдиагональном виде

$$\boldsymbol{\Lambda}=\begin{pmatrix}
\boldsymbol{\Lambda}_{00} & \boldsymbol{\Lambda}_{11} & 0 & 0 & 0 & . & . & 0\\
\boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{01} & \boldsymbol{\Lambda}_{12} & 0 & 0 & . & . & 0\\
0 & \boldsymbol{\Lambda}_{22} & \boldsymbol{\Lambda}_{02} & \boldsymbol{\Lambda}_{13} & 0 & . & . & 0\\
0 & 0 & \boldsymbol{\Lambda}_{23} & \boldsymbol{\Lambda}_{03} & \boldsymbol{\Lambda}_{14} & . & . & 0\\
0 & 0 & 0 & \boldsymbol{\Lambda}_{24} & . & . & . & .\\
. & . & . & . & . & . & . & .\\
. & . & . & . & . & . & . & \boldsymbol{\Lambda}_{1,N+R_1+R_2}\\
0 & 0 & 0 & 0 & 0 & 0 & \boldsymbol{\Lambda}_{2,N+R_1+R_2} & \boldsymbol{\Lambda}_{0,N+R_1+R_2}
\end{pmatrix},\tag{2.11}$$

$$\begin{aligned}
&\boldsymbol{\Lambda}=\operatorname{diag}\left(\boldsymbol{\Lambda}_{00},\boldsymbol{\Lambda}_{01},\ldots,\boldsymbol{\Lambda}_{0,N+R_1+R_2}\right)+\operatorname{diag}^{+}\left(\boldsymbol{\Lambda}_{11},\boldsymbol{\Lambda}_{12},\ldots,\boldsymbol{\Lambda}_{1,N+R_1+R_2}\right)+\\
&+\operatorname{diag}^{-}\left(\boldsymbol{\Lambda}_{21},\boldsymbol{\Lambda}_{22},\ldots,\boldsymbol{\Lambda}_{2,N+R_1+R_2}\right),
\end{aligned}\tag{2.12}$$

где блоки матрицы при $a=n_1+n_2+r_1+r_2$ и $b=\min\left(m_1+R_1,m_2+R_2\right)$ имеют размерность

$$\dim\left(\boldsymbol{\Lambda}_{1a}\right)=\sum_{m_1=0}^{N}\left(\min\left(a-1,b\right)-\max\left(0,a-1+b-N-R_1-R_2\right)+1\right)\times$$

$$\times \sum_{m_1=0}^{N} \left(\min(a,b) - \max\left(0, a+b-N-R_1-R_2\right)+1\right), a = \overline{1, N+R_1+R_2}.$$

$$\dim\left(\mathbf{\Lambda}_{2a}\right) = \sum_{m_1=0}^{N} \left(\min(a,b) - \max\left(0, a+b-N-R_1-R_2\right)+1\right)\times$$

$$\sum_{m_1=0}^{N} \left(\min(a-1,b) - \max\left(0, a-1+b-N-R_1-R_2\right)+1\right), a = \overline{1, N+R_1+R_2}.$$

$$\dim\left(\mathbf{\Lambda}_{0a}\right) = \sum_{m_1=0}^{N} \left(\min(a,b) - \max\left(0, a+b-N-R_1-R_2\right)+1\right)\times \tag{2.13}$$

$$\sum_{m_1=0}^{N} \left(\min(a,b) - \max\left(0, a+b-N-R_1-R_2\right)+1\right), a = \overline{0, N+R_1+R_2}.$$

Отсюда, размерность матрицы $\mathbf{\Lambda}$:

$$\dim\left(\mathbf{\Lambda}\right) = \sum_{a=0}^{N+R_1+R_2} \sum_{m_1=0}^{N} \left(\min(a,b) - \max\left(0, a+b-N-R_1-R_2\right)+1\right)\times \tag{2.14}$$

$$\sum_{a=0}^{N+R_1+R_2} \sum_{m_1=0}^{N} \left(\min(a,b) - \max\left(0, a+b-N-R_1-R_2\right)+1\right).$$

с ненулевыми положительными элементами, вычисляемыми по формулам

$$\mathbf{\Lambda}_{0a}\left(\boldsymbol{x}, \boldsymbol{x}'\right) =$$

$$= \begin{cases} \delta, \quad m_1' = m_1 + r_1, m_2' = m_2 - r_1, n_1' = m_1 + r_1, n_2' = n_2, r_1' = r_1 = 0, r_2' = r_2 = 0, \\ \quad n_1 = m_1, r_1 \le m_2 - n_2, \\ \quad \text{или} \\ \quad m_1' = m_1 + m_2 - n_2, m_2' = n_2, n_1' = m_1 + m_2 - n_2, n_2' = n_2, r_1' = r_1 - m_2 + n_2, \\ \quad r_2' = r_2 = 0, n_1 = m_1, r_1 > m_2 - n_2, \\ \quad \text{или} \\ \quad m_1' = m_1 - r, m_2' = m_2 + r_2, n_1' = n_1, n_2' = m_2 + r_2, r_1' = r_1 = 0, r_2' = r_2 = 0, \\ \quad n_2 = m_2, \ r_2 \le m_1 - n_1, \\ \quad \text{или} \\ \quad m_1' = n_1, m_2' = m_2 + m_1 - n_1, n_1' = n_1, n_2' = m_2 + m_1 - n_1, r_1' = r_1 = 0, \\ \quad r_2' = r_2 - m_1 + n_1, n_2 = m_2, r_2 > m_1 - n_1. \end{cases} \tag{2.15}$$

$$\boldsymbol{\Lambda}_{1a}\left(\boldsymbol{x},\boldsymbol{x}'\right)=$$

$$\begin{cases}\lambda_1, & m_1'=m_1,m_2'=m_2,n_1'=n_1+1,n_2'=n_2,r_1'=r_1=0,r_2'=r_2,n_1+1\le m_1,\\ & \text{или}\\ & m_1'=m_1,m_2'=m_2,n_1'=m_1,n_2'=n_2,r_1'=r_1+1,r_2'=r_2,n_1+1>m_1,\\ \lambda_2, & m_1'=m_1,m_2'=m_2,n_1'=n_1,n_2'=n_2+1,r_1'=r_1,r_2'=r_2=0,n_2+1\le m_2,\\ & \text{или}\\ & m_1'=m_1,m_2'=m_2,n_1'=n_1,n_2'=m_2,r_1'=r_1,r_2'=r_2+1,n_2+1>m_2.\end{cases} \tag{2.16}$$

$$\boldsymbol{\Lambda}_{2a}\left(\boldsymbol{x},\boldsymbol{x}'\right)=$$

$$\begin{cases}\dfrac{m_1}{N}V\mu_1, & m_1'=m_1,m_2'=m_2,n_1'=m_1,n_2'=n_2,r_1'=r_1-1,r_2'=r_2,\\ & \text{или}\\ & m_1'=m_1,m_2'=m_2,n_1'=n_1-1,n_2'=n_2,r_1'=r_1=0,r_2'=r_2,\\ \dfrac{m_1}{N}V\mu_2, & m_1'=m_1,m_2'=m_2,n_1'=n_1,n_2'=m_2,r_1'=r_1,r_2'=r_2-1,\\ & \text{или}\\ & m_1'=m_1,m_2'=m_2,n_1'=n_1,n_2'=n_2-1,r_1'=r_1,r_2'=r_2=0,\\ r_1\varepsilon_1, & m_1'=m_1,m_2'=m_2,n_1'=m_1,n_2'=n_2,r_1'=r_1-1,r_2'=r_2,\\ r_2\varepsilon_2, & m_1'=m_1,m_2'=m_2,n_1'=n_1,n_2'=m_2,r_1'=r_1,r_2'=r_2-1.\end{cases} \tag{2.17}$$

Доказательство. Ввиду введенного на множестве $\mathscr{X}$ лексикографического порядка (2.10), можно сделать вывод, что сортировка состояний на множестве осуществляется в три этапа. Первый этап – сортировка состояний по возрастанию числа запросов двух классов эластичного трафика $n_1+n_2+r_1+r_2$, т.е. элемент $\boldsymbol{x}'=\left(m_1',m_2',n_1',n_2',r_1',r_2'\right)\succ\left(m_1,m_2,n_1,n_2,r_1,r_2\right)=\boldsymbol{x}$, если справедливо соотношение $n_1'+n_2'+r_1'+r_2'>n_1+n_2+r_1+r_2$. Таким образом, пространство состояний $\mathscr{X}$ разбивается на подмножества $\mathscr{X}(a)=\{\boldsymbol{x}\in\mathscr{X}:n_1+n_2+r_1+r_2=a\}$, $a=\overline{0,N+R_1+R_2}$, такие что $\mathscr{X}=\bigcup\limits_{a=0}^{N+R_1+R_2}\mathscr{X}(a)$:

$$\mathscr{X}(0)=\left\{(0,N,0,0,0,0);(1,N-1,0,0,0,0);(2,N-2,0,0,0,0);\ldots;(N,0,0,0,0,0)\right\},$$

$$\mathscr{X}(1)=\left\{\begin{matrix}(0,N,0,1,0,0);(0,N,0,0,1,0);(1,N-1,0,1,0,0);(1,N-1,1,0,0,0);\ldots;\\ (N,0,0,0,0,1);(N,0,1,0,0,0)\end{matrix}\right\},$$

$$\mathcal{X}(2)=\left\{\begin{matrix}(0,N,0,2,0,0);(0,N,0,1,1,0);(0,N,0,0,2,0);\\(1,N-1,0,2,0,0);(1,N-1,1,1,0,0);(1,N-1,1,0,1,0);...;\\(N,0,0,0,0,2);(N,0,1,0,0,1);(N,0,2,0,0,0)\end{matrix}\right\},$$

$$\vdots$$

$$\mathcal{X}(k)=\left\{\begin{matrix}(0,N,0,k,0,0);(0,N,0,k-1,1,0);(0,N,0,k-2,2,0);\\(1,N-1,0,k,0,0);(1,N-1,1,k-1,0,0);...;\\(N,0,0,0,0,k);(N,0,1,0,0,k-1);...;(N,0,k,0,0,0)\end{matrix}\right\},$$

$$k=\overline{1,N+R_1+R_2-1},$$

$$\vdots$$

$$\mathcal{X}\left(N+R_1+R_2\right)=\left\{\begin{matrix}(0,N,0,N,R_1,R_2);(1,N-1,1,N-1,R_1,R_2);\\(2,N-2,2,N-2,R_1,R_2);...;(N,0,N,0,R_1,R_2)\end{matrix}\right\}. \qquad (2.18)$$

Вторым этапом сортировки является сортировка состояний внутри каждого подмножества $\mathcal{X}(a),\ a=\overline{0,N+R_1+R_2}$ следующим образом: $\boldsymbol{x}'\succ\boldsymbol{x}<=>(n_1'+n_2'+r_1'+r_2'=n_1+n_2+r_1+r$ и $m_1'>m_1)$, т.е. по возрастанию порога на максимальное число обслуживаемых сессий 1-класса m_1.

Третьим этапом сортировки является сортировка состояний внутри каждого подмножества $\mathcal{X}(a),\ a=\overline{0,N+R_1+R_2}$ следующим образом: $\forall\boldsymbol{x}',\boldsymbol{x}\in\mathcal{X}(a)$, $\boldsymbol{x}'\succ\boldsymbol{x}<=>(n_1'+n_2'+r_1'+r_2'=n_1+n_2+r_1+r,\ m_1'=m_1$ и $n_1'+r_1'>n_1+r_1)$, т.е. по возрастанию числа сессий 1-класса n_1+r_1.

- Блоки верхней диагонали $\boldsymbol{\Lambda}_{1a}$ представляют переходы СП $\mathbf{X}(t)$ из состояний множества $\mathcal{X}(a-1)$ в состояния множества $\mathcal{X}(a),\ a=\overline{1,N+R_1+R_2}$.
- Блоки нижней диагонали $\boldsymbol{\Lambda}_{2a}$ представляют переходы СП $\mathbf{X}(t)$ из состояний множества $\mathcal{X}(a)$ в состояние множества $\mathcal{X}(a-1),a=\overline{1,N+R_1+R_2}$.
- Блоки центральной диагонали $\boldsymbol{\Lambda}_{0a}$ представляют переходы СП $\mathbf{X}(t)$ внутри множества $\mathcal{X}(a),\ a=\overline{0,N+R_1+R_2}$.

Соответственно, ненулевые элементы блоков будут определяться согласно формулами (2.15)–(2.17) по построению. Ненулевой диагональный элемент

представляет собой сумму всех выходящих интенсивностей из рассматриваемого состояния.

Далее докажем, что размерность блоков определяется соотношением (2.14). Зафиксируем число запросов двух классов эластичного трафика a и значение порога на максимальное число обслуживаемых сессий 1-го класса m_1 (из соотношения $N = m_1 + m_2$, при известном m_1 известно и m_2). Отсюда, задача поиска размерности блоков сводится к комбинаторной задаче разложения a неразличимых шаров по k различным ящикам. В качестве «ящиков» будет выступать максимально возможное число сессий 1-го и 2-го классов в системе: объемы данных ящиков $b_1 = m_1 + R_1$ и $b_2 = m_2 + R_2$ для 1-го и 2-го классов, соответственно. В качестве «шаров» будет выступать число сессий 1- и 2-типа, находящихся в системе: $a = n_1 + n_2 + r_1 + r_2$. Таким образом, будем раскладывать a шаров в два ящика размерностью b_1 и b_2. Предположим, если в первый ящик положим a_1 шаров, то во второй ящик положим оставшиеся $a - a_1$ шаров, при условии $a_1 < b_1$ и $a - a_1 < b_2$. Отсюда, задача разложения шаров по ящикам может быть сведена к задаче поиска числа возможных вариантов наполнения a шарами одного ящика.

Определим ящик с наименьшим объемом $\min(b_1, b_2)$. Пусть первый ящик меньше $b_1 < b_2$. Далее необходимо определить, сколько шаров можно положить в первый ящик с учетом его объема. Минимальное число шаров, которые поместятся в ящике, будет определяться как $\max(0, a - b_2)$:

1) все a шаров поместятся во втором ящике, $a \le b_2 \to 0$;

2) все a шаров не поместятся во втором ящике, $a > b_2 \to a - b_2$.

Теперь определим максимальное число шаров, которые поместятся в ящике, $\min(a, b_1)$:

1) все a шаров поместятся в ящике, $a \le b_1 \to a$;

2) все a шаров не поместятся в ящике, $a > b_1 \to b_1$.

Отсюда, число шаров в первом ящике $\{\max(0,a-b_2),...,\min(a,b_1)\}$, т.е. $\min(a,b_1)-\max(0,a-b_2)+1$. Аналогично для случая, когда второй ящик меньше: пусть

$$b_1 > b_2 : \{\max(0,a-b_1),...,\min(a,b_2)\} \Rightarrow \min(a,b_2)-\max(0,a-b_1)+1.$$

Выразим b_2 через b_1:

$$b_2 = m_2 + R_2 = \{N = m_1 + m_2\} = N - m_1 + R_2 = \{m_1 = m_1 + R_1 - R_1\} =$$
$$= N - (m_1 + R_1 - R_1) + R_2 = \{b_1 = m_1 + R_1\} = N - b_1 - R_1 + R_2.$$

Тогда, в общем случае

$$\begin{aligned} &b = \min(m_1 + R_1, m_2 + R_2), \\ &a = n_1 + n_2 + r_1 + r_2, \\ &\min(a,b) - \max(0, a + b - N - R_1 - R_2) + 1. \end{aligned} \qquad (2.19)$$

Отсюда следует, что размерность подмножества $\mathcal{X}(a)$ равна

$$\sum_{m_1=0}^{N} \left(\min(a,b) - \max(0,a+b-N-R_1-R_2)+1\right)\times$$

$$\times \sum_{m_1=0}^{N} \left(\min(a,b) - \max(0,a+b-N-R_1-R_2)+1\right), a = \overline{0, N+R_1+R_2}.$$

Исходя из указанного выше, что блок верхней диагонали $\mathbf{\Lambda}_{1a}$ представляют переходы СП $\mathbf{X}(t)$ из состояний множества $\mathcal{X}(a-1)$ в состояния множества $\mathcal{X}(a)$, $a = \overline{1, N+R_1+R_2}$, делаем вывод, что размерность блока

$$\mathbf{\Lambda}_{1a} = \sum_{m_1=0}^{N} \left(\min(a-1,b) - \max(0,a-1+b-N-R_1-R_2)+1\right)\times$$

$$\sum_{m_1=0}^{N} \left(\min(a,b) - \max(0,a+b-N-R_1-R_2)+1\right), a = \overline{1, N+R_1+R_2}.$$

Аналогично, для блоков $\mathbf{\Lambda}_{2a}$ и $\mathbf{\Lambda}_{0a}$ мы получим все соотношения (2.13) соответственно.

Утверждение доказано. □

2.3. Матричный алгоритм расчета стационарного распределения

Для расчета стационарного распределения введем макро-вектор строку стационарных вероятностей состояний согласно лексикографическому порядку (2.10)

$$\boldsymbol{\pi} = \Big(\pi_{(0,N,0,0,0,0)}, \pi_{(1,N-1,0,0,0,0)}, \pi_{(2,N-2,0,0,0,0)}, \ldots, \pi_{(m_1,m_2,n_1,n_2,r_1,r_2)}, \ldots,$$

$$\pi_{(N,0,N,0,R_1,R_2)}\Big),$$

где $(m_1,m_2,n_1,n_2,r_1,r_2) \in \mathcal{X}$, $\dim(\boldsymbol{\pi}) = \big(\min(a,b) - \max(0, a+b-N-R_1-R_2)+1\big)$,

компоненты которого $\boldsymbol{\pi}_k = \Big\{\pi_{(0,N,0,k,0,0)}, \pi_{(0,N,0,k-1,1,0)}, \ldots, \pi_{(N,0,1,0,0,k-1)}, \ldots,$

$\pi_{(N,0,k,0,0,0)}\Big\}, k = n_1+n_2+r_1+r_2 = \overline{0, N+R_1+R_2}$. Макро-вектор стационарных вероятностей удовлетворяет системе

$$\boldsymbol{\pi}\boldsymbol{\Lambda} = \mathbf{0},\ \boldsymbol{\pi}\mathbf{e}^T = 1. \tag{2.20}$$

Лемма 2.1. Стационарные вероятности состояний модели с управляемым по сигналам перераспределением приборов вычисляются в матричном виде по формуле

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_{N+R_1+R_2} \prod_{i=1}^{N+R_1+R_2-k} \mathbf{M}_{N+R_1+R_2-i}, \text{ где } k = \overline{0, N+R_1+R_2-1}, \tag{2.21}$$

где вектор $\boldsymbol{\pi}_{N+R_1+R_2}$ является единственным решением системы уравнений

$$\begin{cases} \boldsymbol{\pi}_{N+R_1+R_2}\left(\sum\limits_{k=0}^{N+R_1+R_2-1}\ \prod\limits_{i=1}^{N+R_1+R_2-k} \mathbf{M}_{N+R_1+R_2-i}\right)\mathbf{e}^T = 1, \\ \boldsymbol{\pi}_{N+R_1+R_2}\left(\mathbf{M}_{N+R_1+R_2-1} \cdot \boldsymbol{\Lambda}_{1,N+R_1+R_2} + \boldsymbol{\Lambda}_{0,N+R_1+R_2}\right) = \mathbf{0}, \end{cases} \tag{2.22}$$

а матрицы $\mathbf{M}_k$ вычисляются по рекуррентным соотношениям

$$\mathbf{M}_0 = -\mathbf{\Lambda}_{21} \cdot \mathbf{\Lambda}_{00}{}^{-1},$$
$$\mathbf{M}_k = -\mathbf{\Lambda}_{2,k+1} \cdot \left(\mathbf{M}_{k-1} \cdot \mathbf{\Lambda}_{1k} + \mathbf{\Lambda}_{0k}\right)^{-1}, k = \overline{1, N + R_1 + R_2 - 1}, \qquad (2.23)$$
$$\mathbf{M}_{N+R_1+R_2} = \left(\mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{\Lambda}_{1,N+R_1+R_2} + \mathbf{\Lambda}_{0,N+R_1+R_2}\right)^{-1}.$$

Доказательство. Первое уравнение СУР (2.20) имеет вид

$$\boldsymbol{\pi}_0 \cdot \mathbf{\Lambda}_{00} = -\boldsymbol{\pi}_1 \cdot \mathbf{\Lambda}_{21},$$
$$\boldsymbol{\pi}_0 = -\boldsymbol{\pi}_1 \cdot \mathbf{\Lambda}_{21} \cdot \mathbf{\Lambda}_{00}{}^{-1} = \left\{\mathbf{M}_0 = -\mathbf{\Lambda}_{21} \cdot \mathbf{\Lambda}_{00}{}^{-1}\right\} = \boldsymbol{\pi}_1 \cdot \mathbf{M}_0.$$

Второе уравнение СУР (2.20) имеет вид

$$\boldsymbol{\pi}_0 \cdot \mathbf{\Lambda}_{11} + \boldsymbol{\pi}_1 \cdot \mathbf{\Lambda}_{01} + \boldsymbol{\pi}_2 \cdot \mathbf{\Lambda}_{22} = \mathbf{0},$$
т.к. $\left\{\boldsymbol{\pi}_0 = \boldsymbol{\pi}_1 \cdot \mathbf{M}_0\right\}$, то
$$\boldsymbol{\pi}_1 \cdot \mathbf{M}_0 \cdot \mathbf{\Lambda}_{11} + \boldsymbol{\pi}_1 \cdot \mathbf{\Lambda}_{01} + \boldsymbol{\pi}_2 \cdot \mathbf{\Lambda}_{22} = \mathbf{0},$$
$$\boldsymbol{\pi}_1 \cdot \left(\mathbf{M}_0 \cdot \mathbf{\Lambda}_{11} + \mathbf{\Lambda}_{01}\right) + \boldsymbol{\pi}_2 \cdot \mathbf{\Lambda}_{22} = \mathbf{0},$$
$$\boldsymbol{\pi}_1 \cdot \left(\mathbf{M}_0 \cdot \mathbf{\Lambda}_{11} + \mathbf{\Lambda}_{01}\right) = -\boldsymbol{\pi}_2 \cdot \mathbf{\Lambda}_{22}.$$

$$\boldsymbol{\pi}_1 = -\boldsymbol{\pi}_2 \cdot \mathbf{\Lambda}_{22} \cdot \left(\mathbf{M}_0 \cdot \mathbf{\Lambda}_{11} + \mathbf{\Lambda}_{01}\right)^{-1} = \left\{\mathbf{M}_1 = -\mathbf{\Lambda}_{22}\left(\mathbf{M}_0 \cdot \mathbf{\Lambda}_{11} + \mathbf{\Lambda}_{01}\right)^{-1}\right\} = \boldsymbol{\pi}_2 \cdot \mathbf{M}_1.$$

Третье уравнение СУР (2.20) имеет вид

$$\boldsymbol{\pi}_1 \cdot \mathbf{\Lambda}_{12} + \boldsymbol{\pi}_2 \cdot \mathbf{\Lambda}_{02} + \boldsymbol{\pi}_3 \cdot \mathbf{\Lambda}_{23} = \mathbf{0},$$
т.к. $\left\{\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 \cdot \mathbf{M}_1\right\}$, то
$$\boldsymbol{\pi}_2 \cdot \mathbf{M}_1 \cdot \mathbf{\Lambda}_{12} + \boldsymbol{\pi}_2 \cdot \mathbf{\Lambda}_{02} + \boldsymbol{\pi}_3 \cdot \mathbf{\Lambda}_{23} = \mathbf{0},$$
$$\boldsymbol{\pi}_2 \cdot \left(\mathbf{M}_1 \cdot \mathbf{\Lambda}_{12} + \mathbf{\Lambda}_{02}\right) + \boldsymbol{\pi}_3 \cdot \mathbf{\Lambda}_{23} = \mathbf{0},$$
$$\boldsymbol{\pi}_2 = -\boldsymbol{\pi}_3 \cdot \mathbf{\Lambda}_{23} \cdot \left(\mathbf{M}_1 \cdot \mathbf{\Lambda}_{12} + \mathbf{\Lambda}_{02}\right)^{-1} = \left\{\mathbf{M}_2 = -\mathbf{\Lambda}_{23} \cdot \left(\mathbf{M}_1 \cdot \mathbf{\Lambda}_{12} + \mathbf{\Lambda}_{02}\right)^{-1}\right\} = \boldsymbol{\pi}_3 \cdot \mathbf{M}_2.$$

k -уравнение СУР (2.20) имеет вид

$$\boldsymbol{\pi}_{k-2} \cdot \mathbf{\Lambda}_{1,k-1} + \boldsymbol{\pi}_{k-1} \cdot \mathbf{\Lambda}_{0,k-1} + \boldsymbol{\pi}_k \cdot \mathbf{\Lambda}_{2k} = \mathbf{0},$$
т.к. $\left\{\boldsymbol{\pi}_{k-2} = \boldsymbol{\pi}_{k-1} \cdot \mathbf{M}_{k-2}\right\}$, то
$$\boldsymbol{\pi}_{k-1} \cdot \mathbf{M}_{k-2} \cdot \mathbf{\Lambda}_{1,k-1} + \boldsymbol{\pi}_{k-1} \cdot \mathbf{\Lambda}_{0,k-1} + \boldsymbol{\pi}_k \cdot \mathbf{\Lambda}_{2k} = \mathbf{0},$$
$$\boldsymbol{\pi}_{k-1} \cdot \left(\mathbf{M}_{k-2} \cdot \mathbf{\Lambda}_{1,k-1} + \mathbf{\Lambda}_{0,k-1}\right) + \boldsymbol{\pi}_k \cdot \mathbf{\Lambda}_{2k} = \mathbf{0},$$
$$\boldsymbol{\pi}_{k-1} = -\boldsymbol{\pi}_k \cdot \mathbf{\Lambda}_{2k} \cdot \left(\mathbf{M}_{k-2} \cdot \mathbf{\Lambda}_{1,k-1} + \mathbf{\Lambda}_{0,k-1}\right)^{-1} =$$

$$= \left\{ \mathbf{M}_{k-1} = -\mathbf{\Lambda}_{2k} \cdot \left(\mathbf{M}_{k-2} \cdot \mathbf{\Lambda}_{1,k-1} + \mathbf{\Lambda}_{0k-1} \right)^{-1} \right\} = \boldsymbol{\pi}_k \cdot \mathbf{M}_{k-1}.$$

$(k+1)$-уравнение СУР (2.20) имеет вид

$$\boldsymbol{\pi}_{k-1} \cdot \mathbf{\Lambda}_{1k} + \boldsymbol{\pi}_k \cdot \mathbf{\Lambda}_{0k} + \boldsymbol{\pi}_{k+1} \cdot \mathbf{\Lambda}_{2,k+1} = \mathbf{0},$$

т.к. $\left\{ \boldsymbol{\pi}_{k-1} = \boldsymbol{\pi}_k \cdot \mathbf{M}_{k-1} \right\}$, то

$$\boldsymbol{\pi}_k \cdot \mathbf{M}_{k-1} \cdot \mathbf{\Lambda}_{1k} + \boldsymbol{\pi}_k \cdot \mathbf{\Lambda}_{0k} + \boldsymbol{\pi}_{k+1} \cdot \mathbf{\Lambda}_{2,k+1} = \mathbf{0},$$

$$\boldsymbol{\pi}_k \cdot \left(\mathbf{M}_{k-1} \cdot \mathbf{\Lambda}_{1k} + \mathbf{\Lambda}_{0k} \right) + \boldsymbol{\pi}_{k+1} \cdot \mathbf{\Lambda}_{2,k+1} = \mathbf{0},$$

$$\boldsymbol{\pi}_k = -\boldsymbol{\pi}_{k+1} \cdot \mathbf{\Lambda}_{2,k+1} \cdot \left(\mathbf{M}_{k-1} \cdot \mathbf{\Lambda}_{1k} + \mathbf{\Lambda}_{0k} \right)^{-1} = \left\{ \mathbf{M}_k = -\mathbf{\Lambda}_{2,k+1} \cdot \left(\mathbf{M}_{k-1} \cdot \mathbf{\Lambda}_{1k} + \mathbf{\Lambda}_{0k} \right)^{-1} \right\} =$$

$$= \boldsymbol{\pi}_{k+1} \cdot \mathbf{M}_k,\ k = \overline{1, N + R_1 + R_2 - 1}.$$

$N + R_1 + R_2 + 1$-уравнение СУР (2.20) имеет вид

$$\boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{\Lambda}_{1,N+R_1+R_2} + \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{\Lambda}_{0,N+R_1+R_2} = \mathbf{0},$$

$$\boldsymbol{\pi}_{N+R_1+R_2} \left(\mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{\Lambda}_{1,N+R_1+R_2} + \mathbf{\Lambda}_{0,N+R_1+R_2} \right) = \mathbf{0},$$

$$\boldsymbol{\pi}_{N+R_1+R_2} = \left(\mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{\Lambda}_{1,N+R_1+R_2} + \mathbf{\Lambda}_{0,N+R_1+R_2} \right)^{-1} =$$

$$\boldsymbol{\pi}_{N+R_1+R_2-1} \cdot \mathbf{\Lambda}_{1,N+R_1+R_2} + \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{\Lambda}_{0,N+R_1+R_2} = \mathbf{0},$$

т.к. $\left\{ \boldsymbol{\pi}_{N+R_1+R_2-1} = \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \right\} =$

$$= \left\{ \mathbf{M}_{N+R_1+R_2} = \left(\mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{\Lambda}_{1,N+R_1+R_2} + \mathbf{\Lambda}_{0,N+R_1+R_2} \right)^{-1} \right\} = \mathbf{M}_{N+R_1+R_2}.$$

Отсюда, СУР для СП $\mathbf{X}(t)$ с матрицей $\mathbf{\Lambda}$ имеет вид

$$\begin{cases} \boldsymbol{\pi}_0 = \boldsymbol{\pi}_1 \mathbf{M}_0, \text{ где } \mathbf{M}_0 = -\mathbf{\Lambda}_{21} \mathbf{\Lambda}_{00}^{-1}, \\ \boldsymbol{\pi}_k = \boldsymbol{\pi}_{k+1} \mathbf{M}_k, \text{где } \mathbf{M}_k = -\mathbf{\Lambda}_{2,k+1} \left(\mathbf{M}_{k-1} \mathbf{\Lambda}_{1k} + \mathbf{\Lambda}_{0k} \right)^{-1}, k = \overline{1, N + R_1 + R_2 - 1}, \\ \boldsymbol{\pi}_{N+R_1+R_2} = \mathbf{M}_{N+R_1+R_2}, \text{ где } \mathbf{M}_{N+R_1+R_2} = \left(\mathbf{M}_{N+R_1+R_2-1} \mathbf{\Lambda}_{1,N+R_1+R_2} + \right. \\ \left. + \mathbf{\Lambda}_{0,N+R_1+R_2} \right)^{-1}, \end{cases} \tag{2.24}$$

с нормировочным условием (2.20) $\sum_{k=0}^{N+R_1+R_2} \boldsymbol{\pi}_k \mathbf{e}^T = 1$, где $\mathbf{e}^T$ – единичный вектор, $\dim(\mathbf{e}) = \left(\min(a,b) - \max(0, a+b-N-R_1-R_2)+1\right)$. Далее выразим все $\boldsymbol{\pi}_k$, $k = \overline{0, N+R_1+R_2-1}$ через $\boldsymbol{\pi}_{N+R_1+R_2}$:

$$\boldsymbol{\pi}_{N+R_1+R_2-1} = \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1},$$

$$\boldsymbol{\pi}_{N+R_1+R_2-2} = \boldsymbol{\pi}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} = \left\{\boldsymbol{\pi}_{N+R_1+R_2-1} = \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1}\right\} =$$

$$= \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2},$$

$$\boldsymbol{\pi}_{N+R_1+R_2-3} = \boldsymbol{\pi}_{N+R_1+R_2-2} \cdot \mathbf{M}_{N+R_1+R_2-3} =$$

$$= \left\{\boldsymbol{\pi}_{N+R_1+R_2-2} = \boldsymbol{\pi}_{N+R_1+R_2-3} = \boldsymbol{\pi}_{N+R_1+R_2-2} \cdot \mathbf{M}_{N+R_1+R_2-3}\right\} =$$

$$= \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \mathbf{M}_{N+R_1+R_2-3},$$

$$\vdots$$

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_{k+2} \cdot \mathbf{M}_{k+1} = \left\{\boldsymbol{\pi}_{k+2} = \boldsymbol{\pi}_{k+3} \cdot \mathbf{M}_{k+2} =\right.$$

$$\left. \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_{k+2}\right\} =$$

$$= \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_{k+2} \cdot \mathbf{M}_{k+1}.$$

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_{k+1} \cdot \mathbf{M}_k = \left\{\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_{k+2} \cdot \mathbf{M}_{k+1} =\right.$$

$$\left. = \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_{k+1}\right\} =$$

$$= \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_{k+1} \cdot \mathbf{M}_k,$$

$$\boldsymbol{\pi}_{k-1} = \boldsymbol{\pi}_k \cdot \mathbf{M}_{k-1} = \left\{k = \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \boldsymbol{\pi} \cdot \mathbf{M}_{k+1} \cdot \mathbf{M}_k\right\} =$$

$$= \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_{k+1} \cdot \mathbf{M}_k \cdot \mathbf{M}_{k-1},$$

$$\vdots$$

$$\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 \cdot \mathbf{M}_1 = \left\{\boldsymbol{\pi}_2 = \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_2\right\} =$$

$$= \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_2 \cdot \mathbf{M}_1,$$

$$\boldsymbol{\pi}_0 = \boldsymbol{\pi}_1 \cdot \mathbf{M}_0 = \left\{\boldsymbol{\pi}_1 = \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_1\right\} =$$

$$= \boldsymbol{\pi}_{N+R_1+R_2} \cdot \mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{M}_{N+R_1+R_2-2} \cdot \ldots \cdot \mathbf{M}_1 \cdot \mathbf{M}_0.$$

Следовательно, каждая компонента $\boldsymbol{\pi}_k$, $k = \overline{0, N+R_1+R_2-1}$ имеет вид

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_{N+R_1+R_2} \prod_{i=1}^{N+R_1+R_2-k} \mathbf{M}_{N+R_1+R_2-i}, \text{ где } k = \overline{0, N+R_1+R_2-1}, \qquad (2.25)$$

где матрицы вычисляются рекуррентно

$$\mathbf{M}_0 = -\mathbf{\Lambda}_{21} \cdot \mathbf{\Lambda}_{00}^{-1},$$
$$\mathbf{M}_k = -\mathbf{\Lambda}_{2,k+1} \cdot \left(\mathbf{M}_{k-1} \cdot \mathbf{\Lambda}_{1k} + \mathbf{\Lambda}_{0k}\right)^{-1}, k = \overline{1, N + R_1 + R_2 - 1}, \quad (2.26)$$
$$\mathbf{M}_{N+R_1+R_2} = \left(\mathbf{M}_{N+R_1+R_2-1} \cdot \mathbf{\Lambda}_{1,N+R_1+R_2} + \mathbf{\Lambda}_{0,N+R_1+R_2}\right)^{-1}.$$

Условие нормировки можно представить в виде

$$\boldsymbol{\pi}_{N+R_1+R_2} \left(\sum_{k=0}^{N+R_1+R_2-1} \prod_{i=1}^{N+R_1+R_2-k} \mathbf{M}_{N+R_1+R_2-i} \right) \mathbf{e}^T = 1. \quad (2.27)$$

Вектор $\boldsymbol{\pi}_{N+R_1+R_2}$ является единственным решением системы уравнений (2.22).

Лемма доказана. □

2.4. Анализ показателей эффективности нарезки ресурсов

Перейдем к оценке эффективности управления ресурсами на основании трех предложенных ниже принципов эффективности нарезки ресурсов.

1) Насколько сильно происходит отклонение от начального распределения ресурса, т.е. число сессий, которые могли бы обслуживаться, но находятся на ожидании из-за несправедливого деления ресурсов. Представим, что нагрузка от второго класса трафика увеличивается, и при этом простаивают ресурсы первого. Предположим, что для фиксированного алгоритма управления при поступлении сигнала от контроллера необходимо будет все свободные ресурсы передать второму сегменту. Благодаря чему система справится с этой перегрузкой сессий второго класса. Пусть с течением времени увеличивается поток сессий от первого класса. В текущем распределении ресурсов возникнет ситуация, когда сессии 1-класса будут ущемлены из-за предыдущего перераспределения ресурсов.
2) Насколько часты случаи, когда при поступлении сигнала перераспределения ресурса не происходит.

3) Насколько много простаивает свободного ресурса при ожидающих в это время сессиях трафика. Если в первом принципе задержка обслуживания связана с несправедливым занятием всех ресурсов, то здесь наоборот с тем, что система выделила недостаточное количество ресурсов для обслуживания трафика.

Как упоминалось ранее, в сети одновременно функционируют базовый и виртуальные операторы, каждый из которых стремиться удовлетворить свои потребности. С точки зрения базового оператора, необходимо поддерживать высокие скорости передачи данных, эффективно и успешно распределять и перераспределять ресурсы между сегментами сети. А с точки зрения виртуального оператора, необходимо обеспечивать обслуживание большего количества пользователей и не допускать высокие значения блокировок запросов пользователей.

Показатели эффективности базового оператора:

– коэффициент успеха перераспределения ресурса $\beta = \sum_{\boldsymbol{x}\in\mathscr{B}} \pi\left(m_1, m_2, n_1, n_2, r_1, r_2\right)$ над пространством состояний $\mathscr{B} = \left\{ \begin{array}{r} \boldsymbol{x} \in \mathscr{X} : n_1 = m_1, n_2 < m_2, r_1 > 0, r_2 = 0; \\ n_2 = m_2, n_1 < m_1, r_2 > 0, r_1 = 0 \end{array} \right\}$;

– коэффициент использования ресурса $\gamma = \frac{\gamma_1 + \gamma_2}{2}$, где

$$\gamma_k = \sum_{\boldsymbol{x}\in\mathscr{X}: m_k > 0} \frac{n_k}{m_k} \pi\left(m_1, m_2, n_1, n_2, r_1, r_2\right).$$

Показатели эффективности виртуального оператора: вероятность блокировки запроса по заявкам $B_k = \sum_{\boldsymbol{x}\in\mathscr{B}_k} \pi\left(m_1, m_2, n_1, n_2, r_1, r_2\right), k = 1,2$ над пространствами состояний $\mathscr{B}_1 = \left\{\left(m_1, m_2, n_1, n_2, r_1, r_2\right) \in \mathscr{X} : n_1 = m_1, r_1 = R_1\right\}$ и

$\mathscr{B}_2 = \{(m_1, m_2, n_1, n_2, r_1, r_2) \in \mathscr{X} : n_2 = m_2, r_2 = R_2\}$; среднее число обслуживаемых пользователей $N_k = \sum_{\boldsymbol{x} \in \mathscr{X}} n_k \pi(m_1, m_2, n_1, n_2, r_1, r_2)$; среднее число пользователей в очереди $N_k^r = \sum_{\boldsymbol{x} \in \mathscr{X}} r_k \pi(m_1, m_2, n_1, n_2, r_1, r_2)$.

Пример 2.2. [143] Далее представлены некоторые числовые результаты, полученные по вышеописанным формулам. На графиках (рис. 2.4, рис. 2.5) показана зависимость коэффициентов эффективности перераспределения ресурса γ и β от среднего значения времени δ^{-1} между поступлениями сигналов. Чем чаще поступают сигналы, тем более полно используются предоставленные для операторов ресурсы, на что указывает коэффициент γ. При этом менее полезным для работы системы становится каждый ее вызов, что и демонстрирует поведение показателя β.

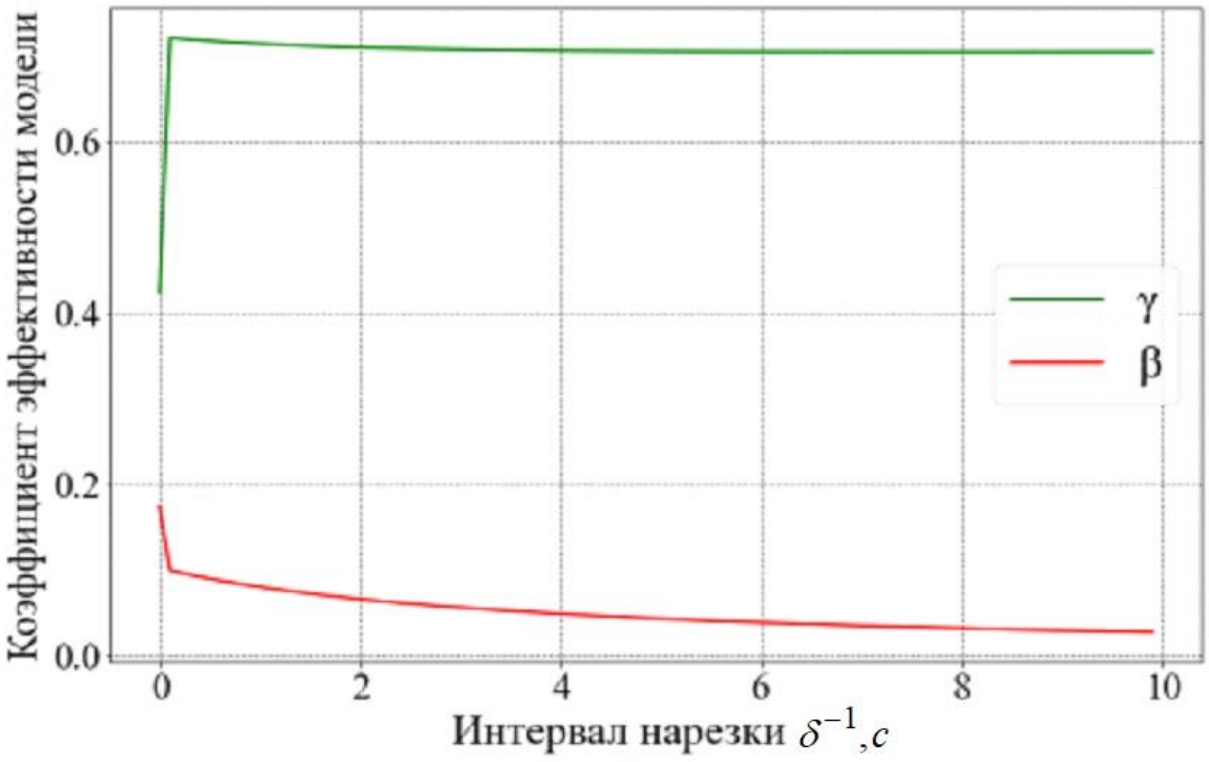


Рис. 2.4. Показатели эффективности для СМО с фиксированной политикой перераспределения ресурса по сигналу [144] при

$V = 2, b = 1, \lambda_1 = 4, \lambda_2 = 3, \mu_1 = \mu_2 = 1, \varepsilon_1 = \varepsilon_2 = 1, r_1 = r_2 = 1$

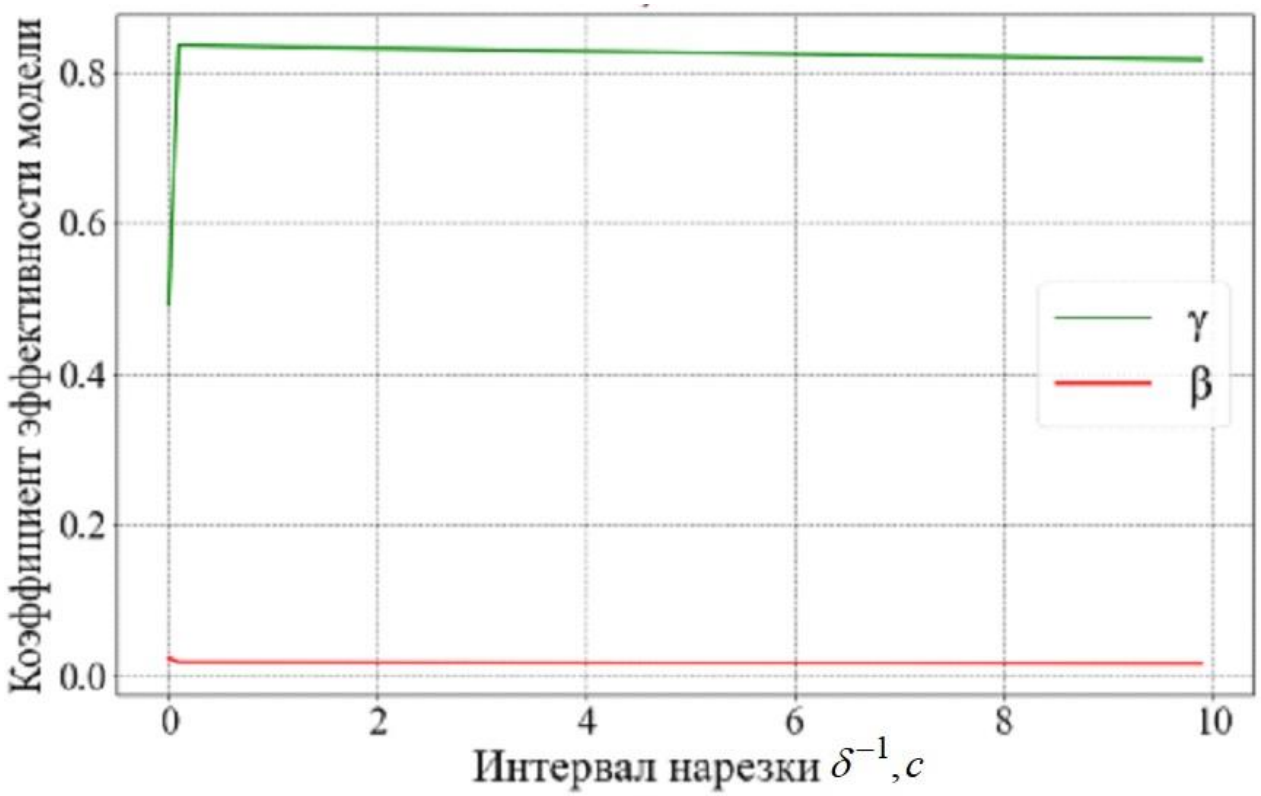


Рис. 2.5. Показатели эффективности для СМО с фиксированной политикой перераспределения ресурса по сигналу [144] при

$$V = 2, b = 1, \lambda_1 = 100, \lambda_2 = 150, \mu_1 = 83, \mu_2 = 111, \varepsilon_1 = \varepsilon_2 = 1, r_1 = r_2 = 1$$

2.5. Задача выбора частоты поступления сигналов

Предложенные в разделе 2.4 показатели эффективности нарезки ресурсов, как со стороны базового оператора – вероятность перераспределения ресурса по сигналу, так и со стороны виртуальных операторов – вероятность блокировки запросов на передачу эластичного трафика, позволяют перейти к задаче выбора частоты поступления сигналов контроллера, которая позволяет максимизировать пропускную способность сети.

Пример 2.3. Пусть виртуальные операторы будут предоставлять доступ к двум типам услуг: просмотр веб-страниц и групповая передача данных. Выбор услуг обуславливается отнесению их к типу обслуживающего ресурса NoN-GBR и услугам, передающим тип «данные». При заданных схеме модуляции и MIMO

схемы определяется объем ресурса базового оператора. Из значений для рекомендуемых и допустимых задержек обслуживания определяем значения скорости передачи данных, а также порог на допустимое время ожидания обслуживания. Исходные данные для численного анализа представлены в табл. 2.2.

Табл. 2.2. Исходные данные для численного анализа

Название	**Групповая передача данных**	**Просмотр веб-страниц**
Базовый оператор		
V	3 МГц, QPSK, MIMO 2X2, 4.688 Мбит/с	
δ	0.000001	
Виртуальный оператор		
μ_k^{-1} – объём данных	1 Мбайт = 8 Мбит	136,58 Кбайт = 1,067 Мбит
	3 Мбайт = 24 Мбит	409,6 Кбайт = 3,2 Мбит
T_1^k , T_2^k - Порог на время задержки	Рекомендуемое – 15 с	Рекомендуемое – 2 с
$b_k = \frac{\mu_k^{-1}}{T_k^1}$ – скорость передачи данных	1,067 Мбит/с 1,6 Мбит/с	
R_k – количество мест в очереди	10	5
Пользователь		
λ_k – интенсивность поступления запросов	0.03	0.6
$(T^2 - T^1)$ - порог на допустимое время задержки в очереди	Допустимое – 60 с 60-15=45с	Допустимое – 4 с
$\varepsilon = \frac{1}{T^2 - T^1}$ – интенсивность ухода по нетерпеливости	$0{,}01\ \text{с}^{-1}$	$0{,}25\ \text{с}^{-1}$

На рис. 2.6 представлена зависимость этих показателей от равных интенсивностей поступления заявок при различной частоте поступления сигналов от контроллера. Далее зафиксируем значение параметра вызовов нарезки сети и построим зависимости параметров от изменения интенсивности поступающих запросов (рис. 2.7). Очевидно, что при увеличении поступающих потоков

вероятности блокировок будут также увеличиваться, причем при заданных исходных данных, блокировки 1-класса трафика будут расти быстрее. Но при этом и коэффициент успеха перераспределения ресурса будет увеличиваться, что свидетельствует о том, что при малой загрузки сети внешние сигналы могут поступать слишком часто и не вносить изменения в распределение ресурсов.

Рис. 2.6. Показатели эффективности для примера 2.3 [145]

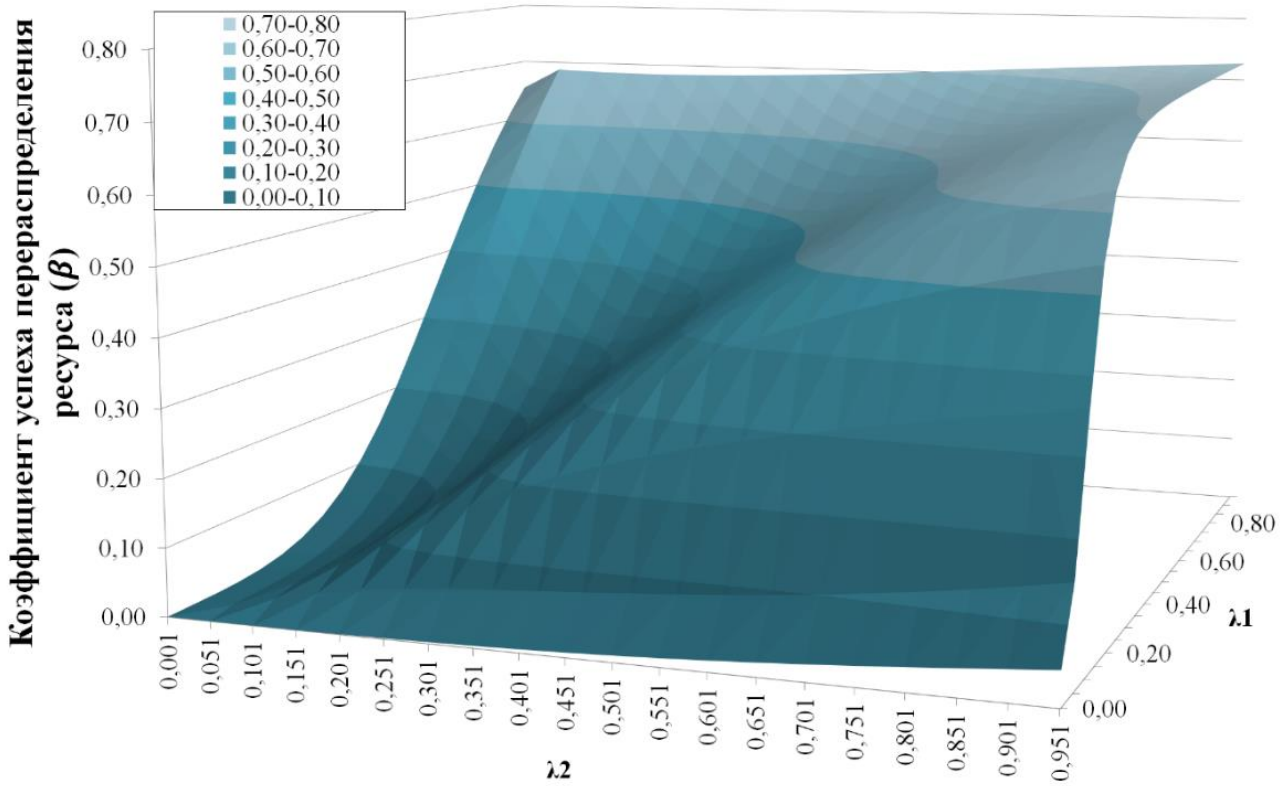


Рис. 2.7. Коэффициент успеха перераспределения ресурса при $\delta = 10^{-6}$

На рис. 2.8 и рис. 2.9 представлено влияние частоты поступления сигналов на поведение системы при фиксировании одной интенсивности поступления заявок и увеличении другой в k раз. Поведение графиков аналогично предыдущим, но при этом увеличение коэффициента успеха перераспределения ресурсов наблюдается при увеличении параметра k. Напомним, что целью анализа является определение оптимального значения частоты поступления внешних сигналов контроллера о перераспределении ресурсов. При определении различных ограничений на вероятность блокировки запросов данный параметр будет изменяться. Так например, для достижения блокировок ниже 0,001 практически вне зависимости от изменения значения k, значение δ будет изменяться начиная от 0,01 (на рис. 2.9 отражено синим цветом). Если же мы ослабим ограничения на блокировки (рис. 2.10), интервалы выбора значения дельта будут уже сильнее зависеть от параметра k, т.е. от того во сколько раз интенсивность поступления одного класса заявок больше другого.

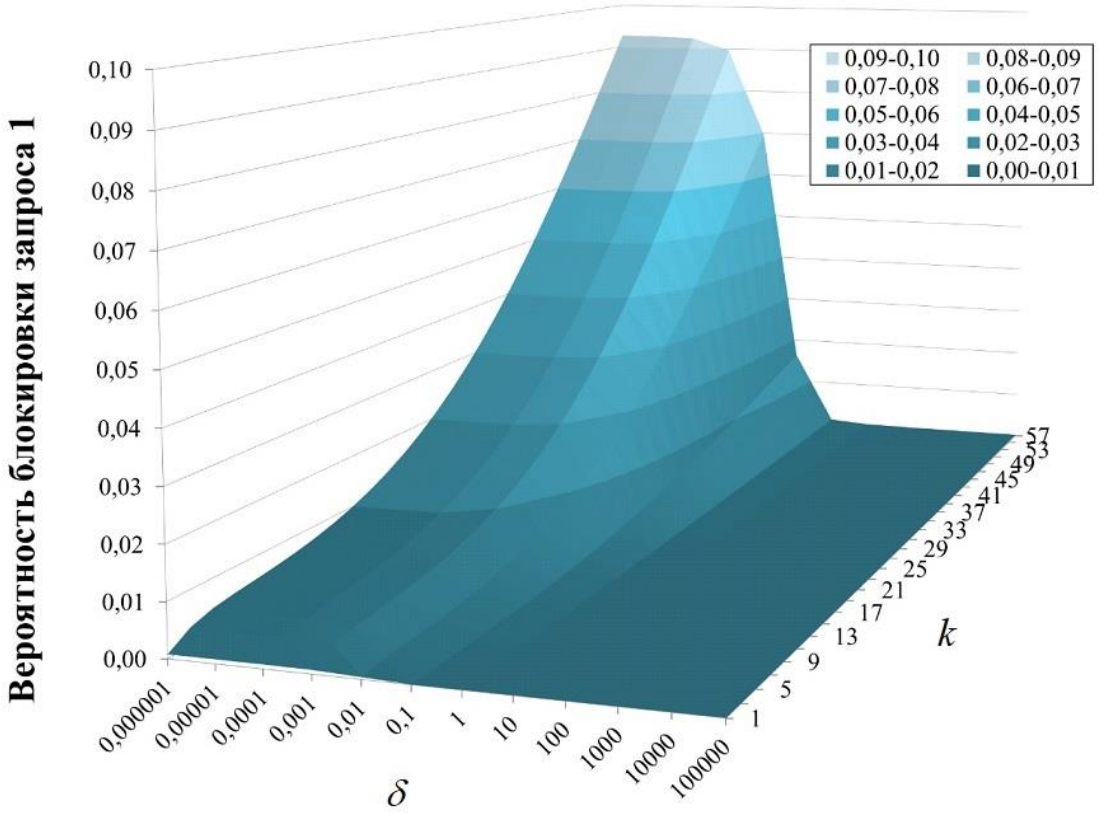


Рис. 2.8. Вероятность блокировки для групповой передачи данных

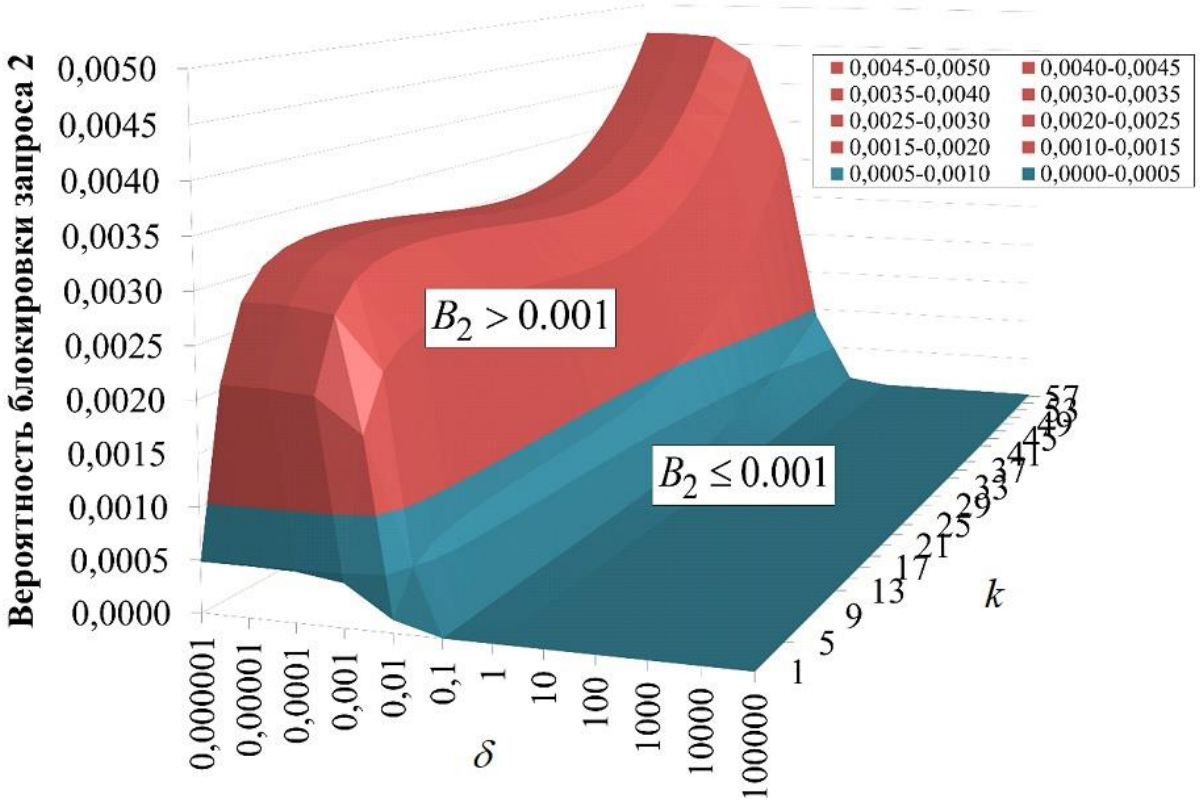


Рис. 2.9. Вероятность блокировки для просмотра веб-страниц и порога $B_2 \le 0,001$

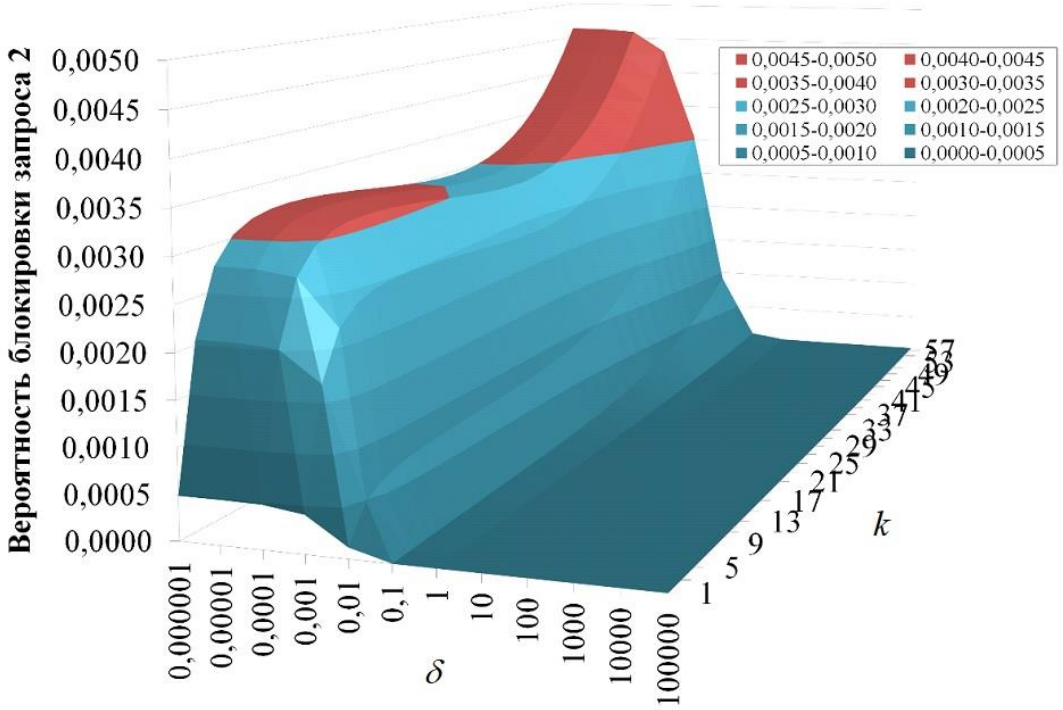


Рис. 2.10. Вероятность блокировки для просмотра веб-страниц и порога $B_2 \le 0,003$

Далее на графиках приведены проекции перечисленных характеристик, по оси X – частота поступления сигналов, а по оси У – соотношение между входящими потоками ($\lambda_2 = k \cdot \lambda_1$, $\lambda_1 = 0,03$). Поведение графиков аналогично предыдущим, но при этом увеличение коэффициента успеха перераспределения ресурсов наблюдается при увеличении параметра k. Более подробно остановимся на этих графиках, на рис. 2.11 синим цветом отражена область допустимых значений при указанных ограничениях для каждого из показателей (вероятностей блокировок и успеха перераспределения ресурса). На рис. 2.12 отражена получаемая область при одновременном учете всех ограничений. Получим, что для разницы в потоках, не превышающей значение 30, нет такой частоты поступления сигналов, удовлетворяющей сразу всем ограничениям. Для значений в интервале от 30 до 40 можно использовать фиксированную нарезку ресурсов. А свыше 40 необходимо настраивать частоту поступления сигналов из заданного набора значений, что обозначено красными отрезками. При этом в зависимости от критерия оптимизации – вероятности блокировок (правая граница) или доли успешных

сигналов (левая граница интервалов) – будет выбрано конкретное значение частоты.

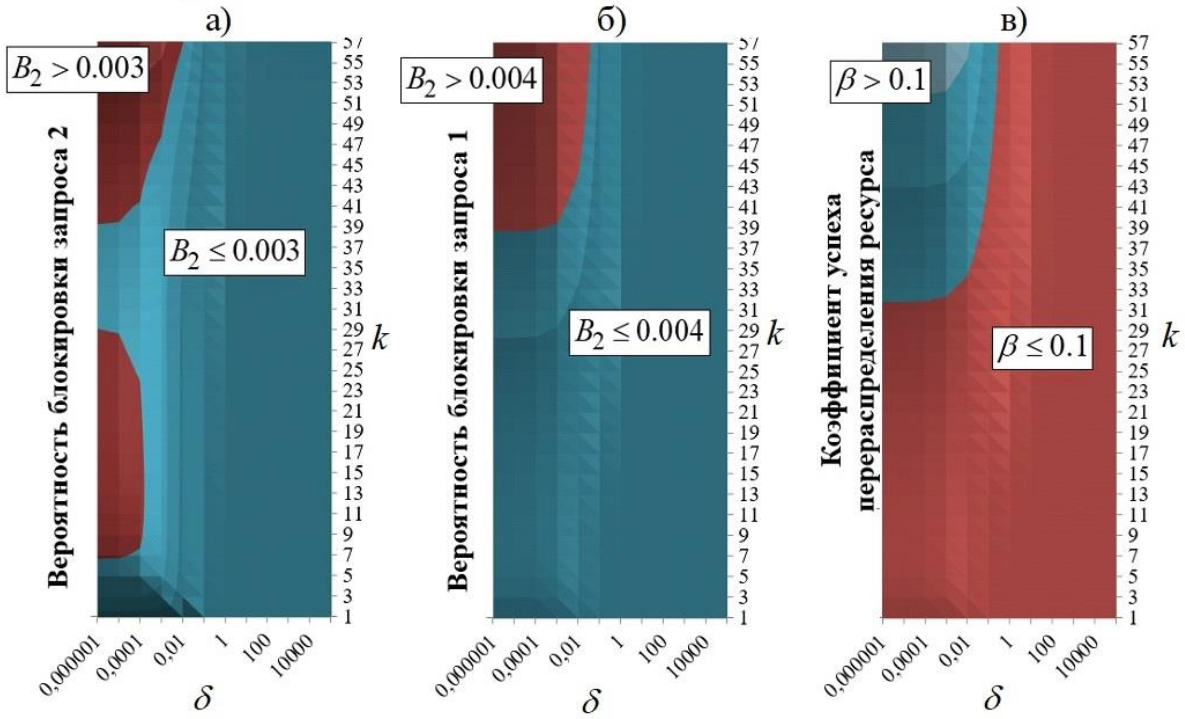


Рис. 2.11. Области допустимых значений δ: а) $B_2 \le 0,003$, б) $B_1 \le 0,004$, в) $\beta \le 0,1$

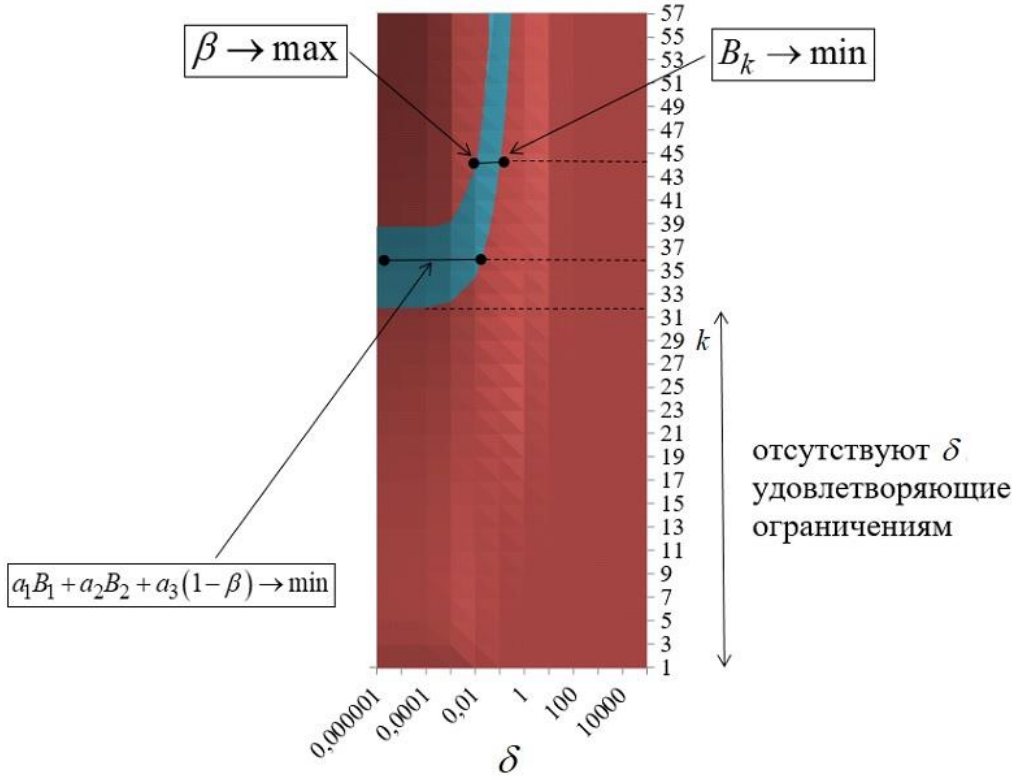


Рис. 2.12. Выбор частоты поступления сигналов δ при $B_2 \le 0,003, B_1 \le 0,004$, $\beta \le 0,1$

Таким образом, во второй главе была построена система массового обслуживания с минимальной скоростью обслуживания нетерпеливого эластичного трафика и перераспределением ресурса по сигналам контроллера. Алгоритм перераспределения ресурса настроен фиксировано для максимального использования имеющихся ресурсов. Полученный матричный алгоритм расчета стационарного распределения позволяет рассчитать показатели эффективности нарезки ресурсов как со стороны базового оператора – вероятность перераспределения ресурса по сигналу, так и со стороны виртуальных операторов – вероятность блокировки запросов на передачу эластичного трафика.

ГЛАВА 3
МОДЕЛЬ С ПОЛИТИКОЙ УПРАВЛЕНИЯ ВЫБОРОМ ОБЪЕМА ПЕРЕРАСПРЕДЕЛЕНИЯ РЕСУРСА

3.1. Построение управляемой системы массового обслуживания

В главе 3 получены результаты №2 и №3, сформулированные в разделе 1.5. В главе 2 анализ системы обслуживания пользователей двумя виртуальными операторами (сегментами) проводился при фиксированной политике перераспределения ресурсов, когда при поступлении сигнала контроллера алгоритм нарезки настроен на максимальное занятие имеющихся ресурсов, т.е. перераспределение всех свободных ресурсов одному из сегментов. В таком случае ожидающие начала обслуживания запросы одного сегмента сети смогут быстрее получать услугу, но до следующего поступления сигнала контроллера другой сегмент будет иметь ресурсов меньше, чем при начальном распределении. Однако, можно использовать не просто фиксированный алгоритм нарезки, а в зависимости от нагрузки применять динамическое перераспределение ресурсов.

В данной главе рассматривается более гибкая модель, в которой возможно применение управляющих воздействий на распределение ресурсов между двумя сегментами при поступлении сигнала контроллера (особенности функционирования контроллера представлены в разделе 2.1). Правилом использования управляющих воздействий будет являться некоторая стратегия, представляющая собой новый объем выделенных ресурсов для первого сегмента в каждом состоянии системы в зависимости от числа занятых ресурсов и числа ожидающих начала обслуживания запросов. Критериями для построения стратегий в зависимости от целей исследования могут служить различные характеристики системы, в данной работе выбор основан на показателях эффективности нарезки, которые представлены ранее в разделе 2.5 и учитывают как максимальное использование ресурсов, так и соответствие начальному распределению и долю

успешных сигналов. Исходя из описания, такая система принадлежит классу УпрСМО, аппарат которого представлен в разделе 1.4.

При описании функционирования системы может быть более удобным переход от случайного процесса $\mathbf{X}(t)$, представленного в главе 2 (рис. 2.1), к процессу уже с тремя состояниями вида $\boldsymbol{s}=(m_1,l_1,l_2)$, где m_1 порог на максимальное число обслуживаемых сессий 1-класса, l_k число сессий k-класса, над пространством состояний

$$\mathcal{S}=\{\boldsymbol{s}=(m_1,l_1,l_2):m_1=0\ldots N;\ l_1=0\ldots m_1+R_1;l_2=0\ldots N-m_1+R_2\}. \quad (3.1)$$

Отметим, что $N-m_1$ порог на максимальное число обслуживаемых сессий 2-класса. Разбиение множества $\mathcal{S}$ на подмножества производится в зависимости от возможности перераспределения ресурсов, а также соотношений между числом обслуживаемых и ожидающих сессий (рис. 3.1). Обозначим $\mathcal{S}_\delta$ состояния, в которых при поступлении сигнала произойдет перераспределение ресурсов, и $\overline{\mathcal{S}_\delta}$ состояния, в которых – не произойдет $\mathcal{S}=\mathcal{S}_\delta\cup\overline{\mathcal{S}_\delta}$. Подмножество состояний с m_1 числом приборов 1-класса $\mathcal{S}=\bigcup_{m_1=0}^{N}\mathcal{S}(m_1),\ \mathcal{S}(m_1)=\{(i,l_1,l_2)\in\mathcal{S}:i=m_1\}$, которое разбивается на те, в которых произойдет перераспределение ресурсов и нет $\mathcal{S}(m_1)=\mathcal{S}_\delta(m_1)\cup\overline{\mathcal{S}_\delta}(m_1)$. В свою очередь $\mathcal{S}_\delta(m_1)=\mathcal{S}_\delta^1(m_1)\cup\mathcal{S}_\delta^2(m_1)$, где $\mathcal{S}_\delta^1(m_1)=\{(i,l_1,l_2)\in\mathcal{S}(m_1):l_1>m_1,\quad l_2<N-m_1\}$ – состояния, в которых при поступлении сигнала будут увеличены ресурсы для 1-класса трафика, а $\mathcal{S}_\delta^2(m_1)=\{(i,l_1,l_2)\in\mathcal{S}(m_1):\quad l_1<m_1,l_2>N-m_1\}$ – ресурсы для 2-класса. Следовательно, в случае наличия необходимого числа свободных ресурсов для обслуживания ожидающих сессий увеличение сегментов осуществляется на число ожидающих сессий $\mathcal{S}_\delta^k(m_1)=\mathcal{S}_\delta^{k1}(m_1)\cup\mathcal{S}_\delta^{k2}(m_1)$, $\mathcal{S}_\delta^{k1}(m_1)=\{(i,l_1,l_2)\in\mathcal{S}_\delta^k(m_1):\ l_1+l_2\le N\}$ или на все свободные ресурсы $\mathcal{S}_\delta^{k2}(m_1)=\{(i,l_1,l_2)\in\mathcal{S}_\delta^k(m_1):\ l_1+l_2>N\}$.

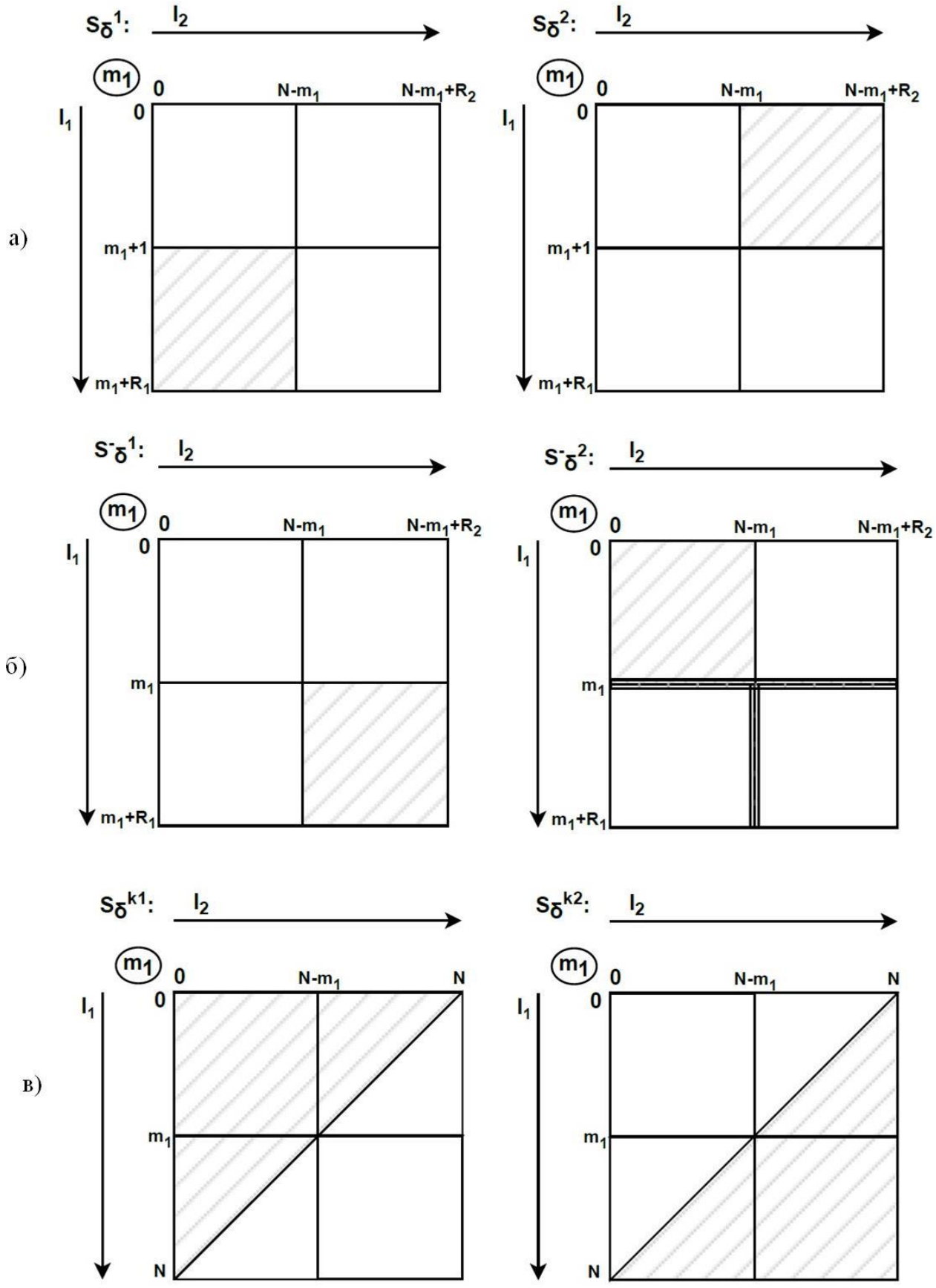


Рис. 3.1. Подмножества пространства состояний $\mathcal{S}$: а) $\mathcal{S}_\delta$, б) $\overline{\mathcal{S}_\delta}$, в) $\mathcal{S}_\delta^k$

Управлением будет являться выбор нового объема ресурса для 1-класса трафика при поступлении сигнала контроллера. При этом специфика рассматриваемой модели накладывает определенные ограничения на множество стратегий, исходя из разбиения пространства состояний. Обозначим $\mathcal{A} = \{0,...,N\}$ множество действий с элементами $a \in \mathcal{A}$, где a означает «выделить a ресурсов под 1-класс сессий». Тогда подмножество допустимых стратегий примет вид, приведенный в утверждении 3.1.

Утверждение 3.1. [146] Множество допустимых стратегий $\mathcal{A}_{\boldsymbol{s}} \subseteq \mathcal{A}$ в состоянии $\boldsymbol{s} = (m_1, l_1, l_2)$, $\boldsymbol{s} \in \mathcal{S}$, когда имеются свободные ресурсы $\{l_1 < m_1\} \cup \{l_2 < N - m_1\}$ определяется в виде

$$\mathcal{A}_{\boldsymbol{s}} = \begin{cases} \{m_1,...,l_1\}, & \boldsymbol{s} \in \mathcal{S}_{\delta}^{11}(m_1), \\ \{m_1,...,N-l_2\}, & \boldsymbol{s} \in \mathcal{S}_{\delta}^{12}(m_1), \\ \{N-l_2,...,m_1\}, & \boldsymbol{s} \in \mathcal{S}_{\delta}^{21}(m_1), . \\ \{l_1,...,m_1\}, & \boldsymbol{s} \in \mathcal{S}_{\delta}^{22}(m_1), \\ \varnothing, & \boldsymbol{s} \notin \mathcal{S}_{\delta}(m_1). \end{cases} \quad (3.2)$$

Доказательство. Стратегией управления для модели с двумя сегментами является выбор нового объема ресурсов для первого сегмента при поступлении сигнала контроллера. Множество допустимых стратегий формируется только на подмножестве состояний системы $\mathcal{S}_{\delta}(m_1)$, где при поступлении этого сигнала перераспределение ресурсов произойдет, т.е. $\mathcal{A}_{\boldsymbol{s}} = \varnothing, \boldsymbol{s} \notin \mathcal{S}_{\delta}(m_1)$. При этом в зависимости от соотношений числа сессий, находящихся на обслуживании и на ожидании, множество допустимых стратегий разбивается на четыре случая. При поступлении сигнала о перераспределении, если обе очереди свободны или заняты, распределение остается неизменным. Если хотя бы одна очередь свободна, в то время как вторая занята, при поступлении сигнала можно инициировать перераспределение ресурсов. При этом для каждого класса трафика возможны еще два случая: когда свободных ресурсов одного класса достаточно для обслуживания ожидающих обслуживания сессий другого класса, и когда недостаточно.

- Если свободных ресурсов 2-сегмента достаточно для обслуживания всех ожидающих сессий $\boldsymbol{s} \in \mathcal{S}_{\delta}^{11}(m_1)$, будет увеличен 1-сегмент от m_1 (останется неизменным) до l_1 (общее число сессий 1-класса) приборов.
- Если свободных ресурсов 2-сегмента недостаточно для обслуживания всех ожидающих сессий $\boldsymbol{s} \in \mathcal{S}_{\delta}^{12}(m_1)$, будет увеличен 1-сегмент от m_1 до $N - l_2$ (все свободные приборы будут предоставлены 1-сегменту).
- Если свободных ресурсов 1-сегмента достаточно для обслуживания всех ожидающих сессий $\boldsymbol{s} \in \mathcal{S}_{\delta}^{21}(m_1)$, будет увеличен 2-сегмент. Т.к. управление состоит в выборе объема ресурсов для 1-сегмента, он и будет уменьшен от $N - l_2$ (1-сегменту выделяется весь ресурс минус число сессий 2-сегмента) до m_1.
- Если свободных ресурсов 1-сегмента недостаточно для обслуживания всех ожидающих сессий $\boldsymbol{s} \in \mathcal{S}_{\delta}^{22}$, будет уменьшен 1-сегмент от l_1 (сегмент уменьшен до числа сессий 1-класса) до m_1.

Утверждение доказано. □

В табл. 3.1 перечислены возможные интенсивности переходов между состоянием $\boldsymbol{s}$ и другими состояниями системы $\boldsymbol{s}'$ при поступлении, обслуживании и уходе сессий из системы. В табл. 3.2 вторая часть переходов из состояния в состояние отражает четыре возможных случая при поступлении сигнала о перераспределении (которые отражены в утверждении 3.1). При этом параметр a на первом месте в состоянии системы, в которое осуществляется переход, будет являться параметром управления системой и отражать новое число выделенных ресурсов под обслуживание 1 класса трафика. Таким образом, при поступлении сигнала важно не просто инициировать перераспределение ресурсов, но и спланировать новое распределение эффективно с учетом нагрузки сеть, чтобы обеспечивать лучшее качество обслуживания.

Табл. 3.1. Интенсивности переходов для модели с политикой управления выбором объема перераспределения ресурса: случайные события с запросами пользователей [148]

№ п/п	Описание	Интенсивность события	Условие на s	Состояние s'
1-1	поступление 1-сессии (либо в очередь, либо на прибор)	λ_1	$l_1+1\le R_1$	(m_1, l_1+1, l_2)
1-2	поступление 2-сессии (либо в очередь, либо на прибор)	λ_2	$l_2+1\le R_2$	(m_1, l_1, l_2+1)
2-1	обслуживание 1-сессии	$\frac{m_1}{N}V\mu_1$	$m_1>0,\ l_1>0$	(m_1, l_1-1, l_2)
2-2	обслуживание 2-сессии, очередь имеется	$\frac{N-m_1}{N}V\mu_2$	$N-m_1>0,\ l_2>0$	(m_1, l_1, l_2-1)
3-1	уход нетерпеливой 1-сессии	$(l_1-m_1)\varepsilon_1$	$l_1\ge m_1$	(m_1, l_1-1, l_2)
3-2	уход нетерпеливой 2-сессии	$(l_2-N+m_1)\varepsilon_2$	$l_2\ge N-m_1$	(m_1, l_1, l_2-1)

Табл. 3.2. Интенсивности переходов для модели с политикой управления выбором объема перераспределения ресурса: случайные события с сигналами контроллера [148]

№ п/п	Описание	Интенсивность события	Условие на s	Условие на a	Состояние s'
4a-1	поступает сигнал, ресурсы 1-класса заняты, есть ожидающие обслуживания запросы 1-класса, ресурсы 2-класса свободны: свободных ресурсов 2-класса достаточно для обслуживания ожидающих сессий 1-класса	δ	$l_1>m_1,\ l_2<N-m_1,$ $l_1-m_1\le N-m_1-l_2$	$m_1\le a\le l_1$	(a, l_1, l_2)

4б-1	поступает сигнал, ресурсы 1-класса заняты, есть ожидающие обслуживания запросы 1-класса, ресурсы 2-класса свободны: свободных ресурсов 2-класса недостаточно для обслуживания ожидающих сессий 1-класса	δ	$l_1 > m_1,\ l_2 < N - m_1,$ $l_1 - m_1 > N - m_1 - l_2$	$m_1 \le a \le N - l_2$	(a, l_1, l_2)
4а-2	поступает сигнал, ресурсы 2-класса заняты, есть ожидающие обслуживания запросы 2-класса, ресурсы 1-класса свободны: свободных ресурсов 1-класса достаточно для обслуживания ожидающих сессий 2-класса	δ	$l_2 > N - m_1,\ l_1 < m_1,$ $l_2 - N + m_1 \le m_1 - l_1$	$N - l_2 \le a \le m_1$	(a, l_1, l_2)
4б-2	поступает сигнал, ресурсы 2-класса заняты, есть ожидающие обслуживания запросы 2-класса, ресурсы 1-класса свободны: свободных ресурсов 1-класса недостаточно для обслуживания ожидающих сессий 2-класса	δ	$l_2 > N - m_1,\ l_1 < m_1,$ $l_2 - N + m_1 > m_1 - l_1$	$l_1 \le a \le m_1$	(a, l_1, l_2)

3.2. Марковский процесс принятия решения в непрерывном времени

Для решения задачи оптимального распределения ресурсов необходимо выбрать наилучшее действие по выбору объема выделяемых ресурсов на основе текущего состояния системы. Такая задача определяется марковским процессом

принятия решения (англ. Markov decision process, MDP). Модель MDP содержит кортеж $\left(\mathcal{S}, A_{\boldsymbol{s}}, \boldsymbol{Q}_a(\boldsymbol{s},\boldsymbol{s}'), R(\boldsymbol{s})\right)$:

- набор возможных состояний системы $\mathcal{S}$, $\boldsymbol{s}=(m_1,l_1,l_2)\in\mathcal{S}$ (определен в (3.1));
- набор возможных действий $\mathcal{A}_{\boldsymbol{s}}$ (определен в утверждении 3.1);
- матрица интенсивностей переходов из состояния $\boldsymbol{s}$ в состояние $\boldsymbol{s}'$ при управляющем действии a, $\boldsymbol{Q}_a(\boldsymbol{s},\boldsymbol{s}')$: $q_a(\boldsymbol{s},\boldsymbol{s}')\geq 0,\ \boldsymbol{s}\neq\boldsymbol{s}'$, $q_a(\boldsymbol{s},\boldsymbol{s})=-q_a(\boldsymbol{s})=-\sum_{\boldsymbol{s}\neq\boldsymbol{s}'} q_a(\boldsymbol{s},\boldsymbol{s}'), q_a(\boldsymbol{s})<\infty$ (определена в соответствии с правилами, определенными в табл. 3.1 и табл. 3.2 и аналогична матрице главы 2):

$$q_a(\boldsymbol{s},\boldsymbol{s}')=\begin{cases}
\lambda_1, & \boldsymbol{s}'=(m_1,l_1+1,l_2),\ \boldsymbol{s}{:}\, l_1+1\leq R_1,\\
\lambda_2, & \boldsymbol{s}'=(m_1,l_1,l_2+1),\ \boldsymbol{s}{:}\, l_2+1\leq R_2,\\
\frac{m_1}{N}V\mu_1, & \boldsymbol{s}'=(m_1,l_1-1,l_2),\ \boldsymbol{s}{:}\, m_1>0,\ l_1>0,\\
\frac{N-m_1}{N}V\mu_2, & \boldsymbol{s}'=(m_1,l_1,l_2-1),\ \boldsymbol{s}{:}\, N-m_1>0,\ l_2>0,\\
(l_1-m_1)\varepsilon_1, & \boldsymbol{s}'=(m_1,l_1-1,l_2),\ \boldsymbol{s}{:}\, l_1\geq m_1,\\
(l_2-N+m_1)\varepsilon_2, & \boldsymbol{s}'=(m_1,l_1,l_2-1),\ \boldsymbol{s}{:}\, l_2\geq N-m_1,\\
\delta, & \boldsymbol{s}'=(a,l_1,l_2),\ m_1\leq a\leq l_1,\ \boldsymbol{s}\in\mathcal{S}_\delta^{11},\\
\delta, & \boldsymbol{s}'=(a,l_1,l_2),\ m_1\leq a\leq N-l_2,\ \boldsymbol{s}\in\mathcal{S}_\delta^{12},\\
\delta, & \boldsymbol{s}'=(a,l_1,l_2),\ N-l_2\leq a\leq m_1,\ \boldsymbol{s}\in\mathcal{S}_\delta^{21},\\
\delta, & \boldsymbol{s}'=(a,l_1,l_2),\ l_1\leq a\leq m_1,\ \boldsymbol{s}\in\mathcal{S}_\delta^{22}.
\end{cases} \qquad (3.3)$$

- функция вознаграждения $R(\boldsymbol{s})$, получаемого в единицу времени в течение пребывания в состоянии $\boldsymbol{s}$.

Таким образом, для построения модели MDP осталось доопределить четвертую компоненту – функцию вознаграждения, которая для рассматриваемой системы будет состоять из трех слагаемых, рассчитываемых для каждого конкретного состояния системы. Каждая из этих трех компонент соответствует принципам

оптимального распределения ресурсов, которые будут представлены ниже (а также соответствуют определенным ранее показателям эффективности нарезки ресурсов, раздел 2.5).

Утверждение 3.2. [146] Функция вознаграждения, получаемого в ед. времени в течение пребывания в системе в состоянии $\boldsymbol{s} \in \mathcal{S}$ вычисляется по формуле

$$R(\boldsymbol{s}) = -\Bigg(c_1 \cdot \chi(\boldsymbol{s}) \cdot 1\Big\{\boldsymbol{s} \in \overline{\mathcal{S}_\delta^1(m_1)}\Big\} \cdot \Bigg[1\Big\{m_1 < \frac{N}{2},\, \boldsymbol{s} \in \mathcal{S}_\delta^1(m_1)\Big\} + \\ + 1\Big\{m_1 > \frac{N}{2},\, \boldsymbol{s} \in \mathcal{S}_\delta^2(m_1)\Big\}\Bigg] + c_2 \cdot \beta_\delta(\boldsymbol{s}) \cdot 1\Big\{\boldsymbol{s} \in \overline{\mathcal{S}_\delta}\Big\} + c_3 \cdot \gamma(\boldsymbol{s}) \cdot 1\{\boldsymbol{s} \in \mathcal{S}_\delta\}\Bigg), \tag{3.4}$$

где $\chi(\boldsymbol{s})$ принцип равного деления ресурсов, $\beta_\delta(\boldsymbol{s})$ принцип «успеха» перераспределения ресурсов и $\gamma(\boldsymbol{s})$ использование ресурсов, а c_1, c_2, c_3 коэффициенты пропорционального соотношения.

$$\chi(\boldsymbol{s}) = \begin{cases} \frac{N}{2} - m_1, & \boldsymbol{s} : m_1 < \frac{N}{2}, l_1 > m_1, l_1 - m_1 > \frac{N}{2} - m_1, \\ l_1 - m_1, & \boldsymbol{s} : m_1 < \frac{N}{2}, l_1 > m_1, l_1 - m_1 \le \frac{N}{2} - m_1, \\ \frac{N}{2} - N + m_1, & \boldsymbol{s} : m_1 > \frac{N}{2}, l_2 > N - m_1, l_2 - N + m_1 > \frac{N}{2} - N + m_1, \\ l_2 - N + m_1, & \boldsymbol{s} : m_1 > \frac{N}{2}, l_2 > N - m_1, l_2 - N + m_1 \le \frac{N}{2} - N + m_1. \end{cases} \tag{3.5}$$

$$\beta_\delta(\boldsymbol{s}) = \delta\Big(\delta + \lambda_1 \cdot 1\{l_1 + 1 \le R_1 + m_1\} + \lambda_2 \cdot 1\{l_2 + 1 \le R_2 + N - m_1\} + \\ + \frac{m_1}{N} V\mu_1 \cdot 1\{m_1 > 0,\, l_1 > 0\} + \frac{N - m_1}{N} V\mu_2 \cdot 1\{N - m_1 > 0,\, l_2 > 0\} + \\ + (l_1 - m_1)\varepsilon_1 \cdot 1\{l_1 > m_1\} + (l_2 - N + m_1)\varepsilon_2 \cdot 1\{l_2 > N - m_1\}\Big)^{-1}, \tag{3.6}$$

$$\gamma(\boldsymbol{s}) = \begin{cases} N - m_1 - l_2, & \boldsymbol{s} \in \mathcal{S}_\delta^{12}(m_1), \\ l_1 - m_1, & \boldsymbol{s} \in \mathcal{S}_\delta^{11}(m_1), \\ m_1 - l_1, & \boldsymbol{s} \in \mathcal{S}_\delta^{22}(m_1), \\ l_2 - N + m_1, & \boldsymbol{s} \in \mathcal{S}_\delta^{21}(m_1). \end{cases} \tag{3.7}$$

Доказательство. Напомним, что параметром управления системы массового обслуживания является число выделяемых ресурсов для 1-класса трафика. В зависимости от критериев оптимизации функция вознаграждения может задаваться разными способами. В данном случае функция вознаграждения состоит из трех слагаемых, которые рассчитываются для каждого конкретного состояния системы. Каждая из трех компонент соответствует определенным ранее принципам оптимального распределения ресурсов (раздел 2.5).

Принцип равного деления ресурсов, который позволяет определять число сессий, которые могли бы обслуживаться, но которым приходится ожидать из-за несправедливого деления ресурсов, задается формулой (3.5), рис. 3.2. Компонента рассчитывается для четырех случаев (определенных в утверждении 3.1) и отражает число сессий, находящихся на ожидании из-за распределения ресурсов, отличного от начального (равного). Второй компонентой является принцип «успеха» перераспределения, который учитывает вероятность того, что поступит сигнал о перераспределении ресурсов, но перераспределения не произойдет из-за состояния системы, и определяется формулой (3.6). Рассчитывается как отношение интенсивностей при поступлении сигнала о перераспределении к сумме всех возможных интенсивностей с учетом условий переходов, которые задаются индикаторной функцией $1(\boldsymbol{s}) = \begin{Bmatrix} 1, \boldsymbol{s} \in \mathcal{S} \\ 0, \boldsymbol{s} \notin \mathcal{S} \end{Bmatrix}$ (матрица интенсивностей переходов задается формулой (3.3)). Третья компонента – использование ресурсов, которая отражает число сессий, которые могли бы обслуживаться, но находятся на ожидании из-за простаивания ресурсов, определяется формулой (3.7), рис. 3.3, рис. 3.4. Если в первой компоненте задержка обслуживания связана с «несправедливым» занятием всех ресурсов, то здесь наоборот с тем, что система выделила недостаточный объем ресурсов для обслуживания трафика.

Утверждение доказано. □

Наконец, функцией вознаграждения будет являться сумма трех этих компонент с весовыми коэффициентами c_1, c_2, c_3. Т.к. функция вознаграждения

фактически представляет собой стоимость или штраф за несправедливое распределение ресурсов, поступление сигнала, который не привел к перераспределению ресурсов и за простаивание ресурсов, то она берется с отрицательным знаком. Таким образом, функция вознаграждения задается формулой (3.4).

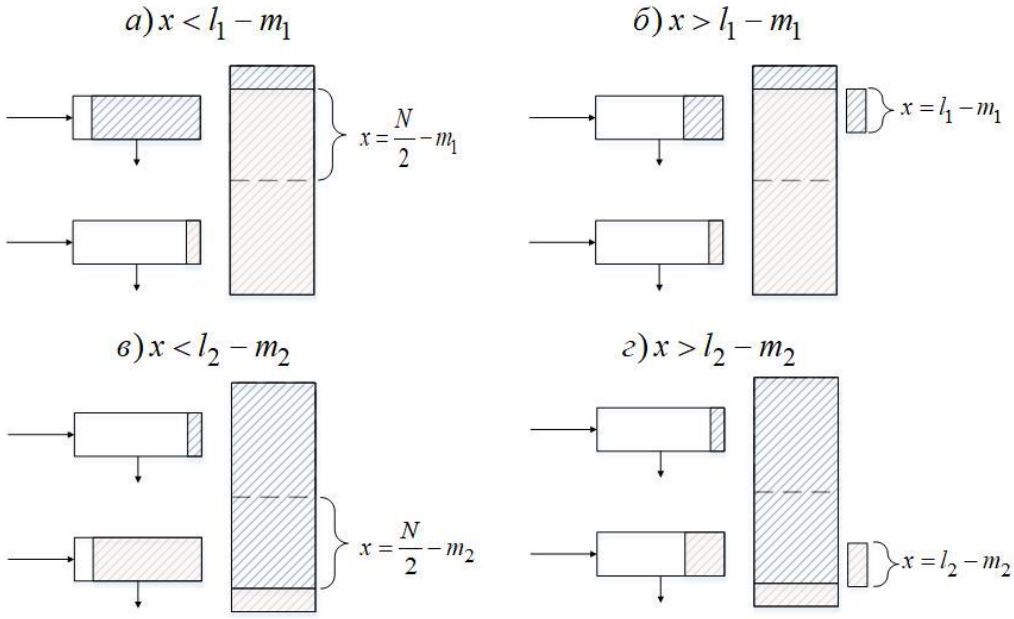


Рис. 3.2. Иллюстрация принципа равного деления ресурсов

при (а-б) $m_1 < \frac{N}{2}, l_1 > m_1, l_1 - m_1 > l_2 - m_2$, (в-г) $m_2 < \frac{N}{2}, l_2 > m_2, l_2 - m_2 > l_1 - m_1$

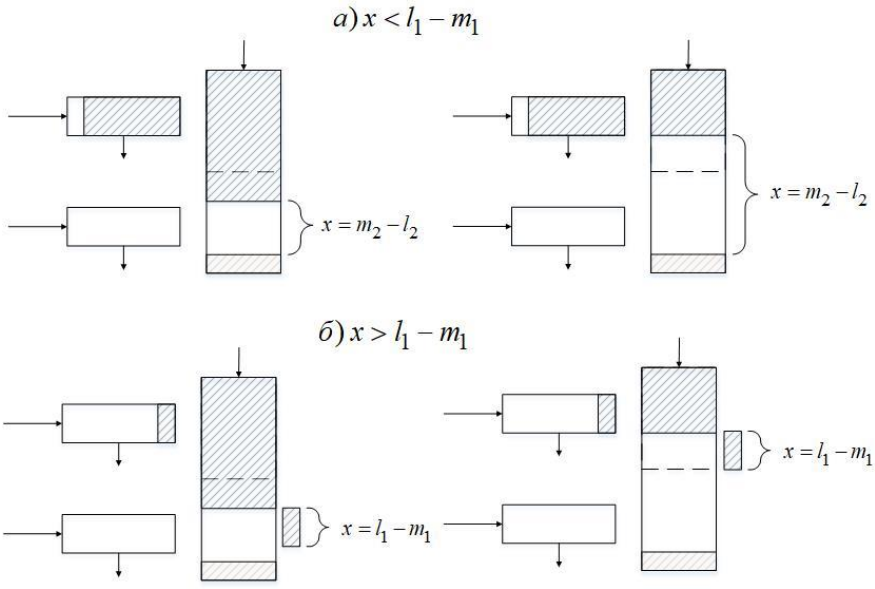


Рис. 3.3. Иллюстрация принципа максимального занятия ресурса

при (а-б) $m_1 \neq \frac{N}{2}, l_1 > m_1, l_2 < m_2$

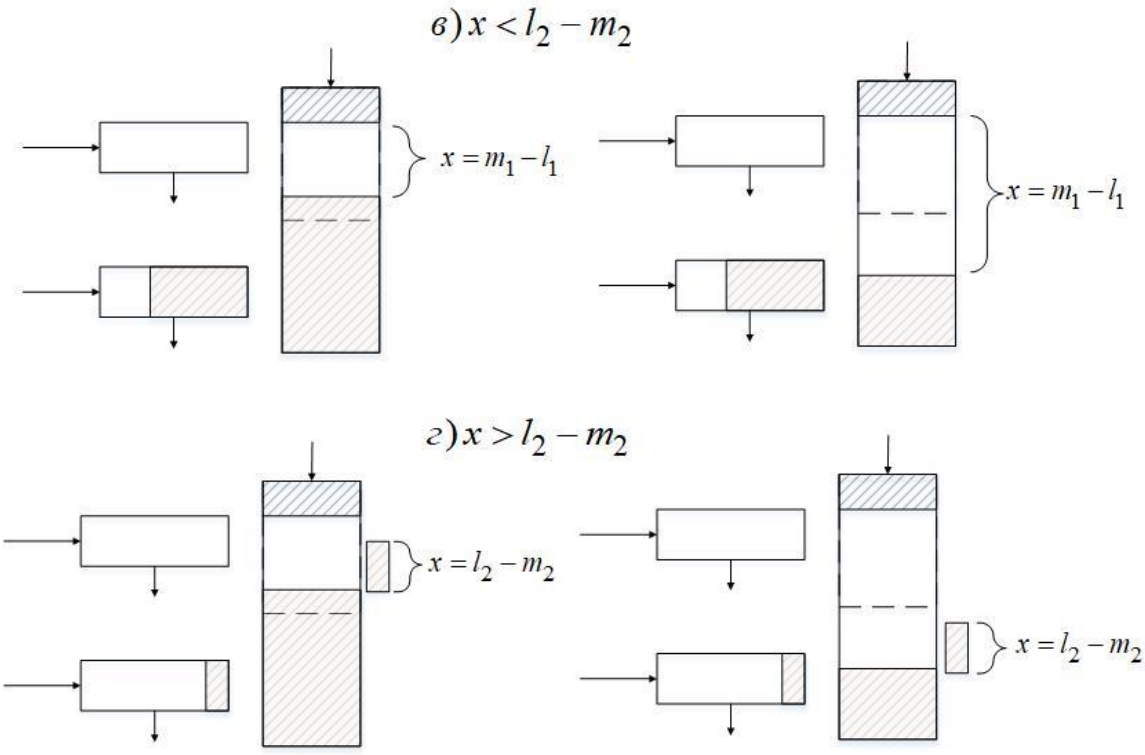


Рис. 3.4. Иллюстрация принципа максимального занятия ресурса при (в-г) $m_2 \neq N/2, l_2 > m_2, l_1 < m_1$

Для нахождения оптимального поведения системы необходимо найти стратегию, которой следует придерживаться в каждом состоянии системы, чтобы добиться максимального вознаграждения. Если воспользоваться стратегией $\boldsymbol{a}$, то функция среднего вознаграждения за ед. времени будет определяться леммой 3.1.

Лемма 3.1. [146] Функция среднего вознаграждения, получаемого в ед. времени, вычисляется по формуле

$$g^{\boldsymbol{a}} = \sum_{\boldsymbol{s} \in \mathcal{S}} R(\boldsymbol{s}) \pi^a(\boldsymbol{s}), \tag{3.8}$$

где $\pi^a(\boldsymbol{s})$ предельная вероятность, что процесс находится в состоянии $\boldsymbol{s}$ при стратегии a.

Доказательство. Функция среднего вознаграждения определяется как

$$g^{\boldsymbol{a}} = \sum_{\boldsymbol{s} \in \mathcal{S}} k_a(\boldsymbol{s}) \pi^a(\boldsymbol{s}), \tag{3.9}$$

где $k_a(\boldsymbol{s})$ непосредственно ожидаемый доход в состоянии $\boldsymbol{s}$ [128, 129]. Предположим, что система приносит вознаграждение в размере $R(\boldsymbol{s})$ ден. единиц

за ед. времени в течение всего периода ее пребывания в состоянии s. Предположим также, что когда система совершает переход из состояния s в состояние $s'(s \neq s')$, она приносит $R(s,s')$ вознаграждения. Отсюда система либо может остаться в состоянии s, либо совершить переход в некоторое другое состояние s'. Если она остается в состоянии s вознаграждение составит $R(s)$. Система также может совершить переход в некоторое состояние $(s \neq s')$ с интенсивностью $q_a(s,s')$, в таком случае вознаграждение составит $R(s,s')$. Следовательно, непосредственно ожидаемый доход в состоянии s при стратегии a: $k_a(s) = R(s) + \sum_{s'=s} q_a(s,s')R(s,s')$ [128].

При этом не обязательно, чтобы система приносила и вознаграждения за ед. времени пребывания в системе, и вознаграждения за переход. В исследуемой модели имеются только вознаграждения за ед. времени пребывания в состоянии, т.е. $R(s,s') = 0$, поэтому непосредственно ожидаемый доход в состоянии s при стратегии a примет вид: $k_a(s) = R(s)$. Тогда формула (3.9) примет вид (3.8).

Лемма доказана. □

Таким образом, для того чтобы принять решение в состоянии s необходимо максимизировать выражение (3.8). Данная задача предполагает рассмотрение большого числа возможных процессов, т.к. для каждого состояния существует свой набор допустимых стратегий, и они могут быть выбраны независимо. Можно найти вознаграждения для каждой из стратегий и выбрать стратегию с наибольшим вознаграждением, однако для более сложных задач сложность вычисления становится высокой.

3.3. Итерационный алгоритм вычисления оптимальной политики

Предложенный Р.А. Ховардом [130] итерационный алгоритм позволяет вычислить оптимальную политику за небольшое число итераций. При этом каждая

итерация представляет собой определение оценок (вознаграждение, получаемое в ед. времени в некотором состоянии при заданной стратегии управления) и улучшение стратегии. Первая часть определена в утверждении 3.3, где получаем систему линейных алгебраических уравнений, аналогично СУР раздела 2.3, но уже относительно оценок для применения итерационного метода. Вторая часть определена в утверждении 3.4.

Утверждение 3.3. [146] Система уравнений относительно функции среднего вознаграждения $g^{\boldsymbol{a}}$ и оценок $v_{\boldsymbol{a}}(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S}$ для итерационного метода решения задается как

$$v_{\boldsymbol{a}}(\boldsymbol{s}) = \frac{1}{\sum\limits_{\boldsymbol{s}' \in \mathcal{S} \setminus \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s}, \boldsymbol{s}')} \left[R(\boldsymbol{s}) + \sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s}, \boldsymbol{s}') v_{\boldsymbol{a}}(\boldsymbol{s}') - g^{\boldsymbol{a}} \right], \boldsymbol{s} \in \mathcal{S}, \tag{3.10}$$

где $q_{a(\boldsymbol{s})}(\boldsymbol{s}, \boldsymbol{s}'), \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{S}$ элемент матрицы интенсивностей переходов, определяемый формулой (3.3), а $R(\boldsymbol{s})$ функция вознаграждения, заданная утверждением 3.2.

Доказательство. Обозначим $v_{\boldsymbol{a}}(\boldsymbol{s})$ – вознаграждение, получаемое в ед. времени в состоянии $\boldsymbol{s}$ при заданной стратегии $\boldsymbol{a} = \left(a(\boldsymbol{s}) \in \mathcal{A}_{\boldsymbol{s}}, \boldsymbol{s} \in \mathcal{S} \right)$. Поэтому $v_{\boldsymbol{a}}(\boldsymbol{s}) = \lim\limits_{t \to \infty} v_{\boldsymbol{a}}(\boldsymbol{s}, t) - t g^{\boldsymbol{a}}$ – разница между вознаграждением, получаемым после выхода из состояния $\boldsymbol{s}$ за время t, и средним значением вознаграждения $g^{\boldsymbol{a}}$ (определяется леммой 3.1), получаемым за время t вне зависимости от начального состояния, т.е. «дельта» вознаграждения за выход из конкретного состояния $\boldsymbol{s}$. Запишем систему уравнений для итерационного метода:

$$v_{\boldsymbol{a}}(\boldsymbol{s}, t + \Delta) =$$

$$= \left(1 - \sum_{\boldsymbol{s}' \neq \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s}, \boldsymbol{s}') \Delta \right) \left[R(\boldsymbol{s}) \Delta + v_{\boldsymbol{a}}(\boldsymbol{s}, t) \right] + \left[\sum_{\boldsymbol{s}' \neq \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s}, \boldsymbol{s}') \Delta \right] v_{\boldsymbol{a}}(\boldsymbol{s}', t) + o(\Delta),$$

вознаграждение, получаемое после выхода из состояния $\boldsymbol{s}$ за время $t + \Delta$ (т.е. в течение интервала времени Δ система может либо остаться в состоянии $\boldsymbol{s}$, либо

совершить переход в некоторое другое состояние s'. Если она остается в состоянии s в течение времени Δ, то вознаграждение составит $R(s)\Delta$ плюс вознаграждение, которое она уже имела $v_a(s,t)$. Интенсивность того, что система останется в состоянии s в течение времени Δ равна 1 минус интенсивность того, что за это время она совершит переход $1-\sum_{s'\neq s} q_{a(s)}(s,s')\Delta$. С другой стороны, за время Δ система может совершить переход в некоторое состояние $s\neq s'$ с интенсивностью $q_{a(s)}(s,s')\Delta$. В этом случае система получит ожидаемое вознаграждение $v_a(s',t)$, которое будет получено за оставшееся время, если бы начальным состоянием было s'. Произведение интенсивностей нужно просуммировать по всем состояниям $s\neq s'$, чтобы получить полное значение ожидаемого вознаграждения.

Т.к. диагональный элемент матрицы интенсивностей переходов $q_{a(s)}(s,s)$ равен сумме недиагональных элементов по строке матрицы, взятой со знаком минус, то $-\sum_{s'\neq s} q_{a(s)}(s,s')=q_{a(s)}(s,s)$:

$$v_a(s,t+\Delta)=\left(1+q_{a(s)}(s,s)\Delta\right)\left[R(s)\Delta+v_a(s,t)\right]+\left[\sum_{s'\neq s} q_{a(s)}(s,s')\Delta\right]v_a(s',t)+o(\Delta)=$$

$$=R(s)\Delta+v_a(s,t)+q_{a(s)}(s,s)\Delta R(s)\Delta+q_{a(s)}(s,s)\Delta v_a(s,t)+$$

$$+\sum_{s'\neq s} q_{a(s)}(s,s')\Delta v_a(s',t)+o(\Delta)=\left\{q_{a(s)}(s,s)\Delta R(s)\Delta=q_{a(s)}(s,s)R(s)\Delta^2=o(\Delta)\right\}=$$

$$=R(s)\Delta+v_a(s,t)+q_{a(s)}(s,s)\Delta v_a(s,t)+\sum_{s'\neq s} q_{a(s)}(s,s')\Delta v_a(s',t)+o(\Delta).$$

Разделим обе части уравнения на Δ

$$\frac{v_a(s,t+\Delta)}{\Delta}=R(s)+\frac{v_a(s,t)}{\Delta}+q_{a(s)}(s,s)v_a(s,t)+\sum_{s'\neq s} q_{a(s)}(s,s')v_a(s',t)+o(1),$$

и перенесем слагаемое $\frac{v_a(s,t)}{\Delta}$ в левую часть уравнения

$$\frac{v_{\boldsymbol{a}}(\boldsymbol{s},t+\Delta)-v_{\boldsymbol{a}}(\boldsymbol{s},t)}{\Delta}=R(\boldsymbol{s})+q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s})v_{\boldsymbol{a}}(\boldsymbol{s},t)+\sum_{\boldsymbol{s}'\neq\boldsymbol{s}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')v_{\boldsymbol{a}}(\boldsymbol{s}',t)+o(1).$$

Устремим интервал времени $\Delta\to 0$ и внесем слагаемое $q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s})v_{\boldsymbol{a}}(\boldsymbol{s},t)$ под знак суммы, получим

$$\lim_{\Delta\to 0}\frac{v_{\boldsymbol{a}}(\boldsymbol{s},t+\Delta)-v_{\boldsymbol{a}}(\boldsymbol{s},t)}{\Delta}=R(\boldsymbol{s})+q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s})v_{\boldsymbol{a}}(\boldsymbol{s},t)+\sum_{\boldsymbol{s}'\neq\boldsymbol{s}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')v_{\boldsymbol{a}}(\boldsymbol{s}',t),$$

$$v'_{\boldsymbol{a}}(\boldsymbol{s},t)=R(\boldsymbol{s})+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')v_{\boldsymbol{a}}(\boldsymbol{s}',t).$$

Далее устремим время $t\to\infty$, получим

$$\lim_{t\to\infty}v'_{\boldsymbol{a}}(\boldsymbol{s},t)=\lim_{t\to\infty}\left[R(\boldsymbol{s})+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')v_{\boldsymbol{a}}(\boldsymbol{s}',t)\right],$$

$$\frac{d\lim\limits_{t\to\infty}v_{\boldsymbol{a}}(\boldsymbol{s},t)}{dt}=R(\boldsymbol{s})+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')\lim_{t\to\infty}v_{\boldsymbol{a}}(\boldsymbol{s}',t).$$

Вернемся к $v_{\boldsymbol{a}}(\boldsymbol{s})=\lim\limits_{t\to\infty}v_{\boldsymbol{a}}(\boldsymbol{s},t)-tg^{\boldsymbol{a}}$, тогда $\lim\limits_{t\to\infty}v_{\boldsymbol{a}}(\boldsymbol{s},t)=v_{\boldsymbol{a}}(\boldsymbol{s})+tg^{\boldsymbol{a}}$:

$$\frac{d\left(v_{\boldsymbol{a}}(\boldsymbol{s})+tg^{\boldsymbol{a}}\right)}{dt}=R(\boldsymbol{s})+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')\left(v_{\boldsymbol{a}}(\boldsymbol{s}')+tg^{\boldsymbol{a}}\right),$$

$$g^{\boldsymbol{a}}=R(\boldsymbol{s})+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')\left(v_{\boldsymbol{a}}(\boldsymbol{s}')+tg^{\boldsymbol{a}}\right)=$$

$$=R(\boldsymbol{s})+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')v_{\boldsymbol{a}}(\boldsymbol{s}')+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')tg^{\boldsymbol{a}}=$$

$$=R(\boldsymbol{s})+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')v_{\boldsymbol{a}}(\boldsymbol{s}')+tg^{\boldsymbol{a}}\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}'),\boldsymbol{s}\in\mathcal{S}.$$

Т.к. суммы элементов по строке матрицы равна $\sum\limits_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')=0$, то система уравнений примет вид

$$g^{\boldsymbol{a}}=R(\boldsymbol{s})+\sum_{\boldsymbol{s}'\in\mathcal{S}}q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')v_{\boldsymbol{a}}(\boldsymbol{s}'),\boldsymbol{s}\in\mathcal{S}. \qquad (3.11)$$

Из системы уравнений (3.11) получим оценки $v_{\boldsymbol{a}}(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S}$ для итерационного метода. Вынесем состояние $\boldsymbol{s} \in \mathcal{S}$ из под знака суммы

$$g^{\boldsymbol{a}} = R(\boldsymbol{s}) + q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}) v_{\boldsymbol{a}}(\boldsymbol{s}) + \sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}') v_{\boldsymbol{a}}(\boldsymbol{s}') =$$

$$= \left\{ q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}) = -\sum_{\boldsymbol{s}' \neq \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}') \right\} =$$

$$= R(\boldsymbol{s}) - v_{\boldsymbol{a}}(\boldsymbol{s}) \sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}') + \sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}') v_{\boldsymbol{a}}(\boldsymbol{s}'), \boldsymbol{s} \in \mathcal{S}.$$

Отсюда получим формулу (3.10):

$$v_{\boldsymbol{a}}(\boldsymbol{s}) = \frac{1}{\sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}')} \left[R(\boldsymbol{s}) + \sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \boldsymbol{s}} q_{a(\boldsymbol{s})}(\boldsymbol{s},\boldsymbol{s}') v_{\boldsymbol{a}}(\boldsymbol{s}') - g^{\boldsymbol{a}} \right], \boldsymbol{s} \in \mathcal{S}.$$

Утверждение доказано. □

Итак, решая систему линейных уравнений, положив на нулевой итерации значения $v_{\boldsymbol{a}}^{[0]}(\boldsymbol{s})$ равными нулю, можно найти оценки и среднее вознаграждение. Далее используя полученные оценки, находится стратегия, которая приводит систему к большему вознаграждению, чем начальная. Исходя из системы уравнений (3.11), для того чтобы принять решение в состоянии $\boldsymbol{s} \in \mathcal{S}$ достаточно максимизировать выражение в утверждении 3.4.

Утверждение 3.4. [146] Целевая функция для улучшения стратегии управления вычисляется по формуле

$$a(\boldsymbol{s}) = \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ v_{\boldsymbol{a}}(a, l_1, l_2) \right\}, \boldsymbol{s} \in \mathcal{S}, \tag{3.12}$$

где $\mathcal{A}_{\boldsymbol{s}}, \boldsymbol{s} \in \mathcal{S}$ множество допустимых стратегий, определенное в утверждении 3.1.

Доказательство. Пусть $a(\boldsymbol{s}) \in \mathcal{A}_{\boldsymbol{s}}$ некоторое действие в состоянии $\boldsymbol{s} \in \mathcal{S}$ из множества допустимых стратегий $\mathcal{A}_{\boldsymbol{s}}$, определенном в утверждении 3.1, и оно известно на n-шаге: $a[n](\boldsymbol{s})$. При этом вектор $\boldsymbol{a}[n] = (a(\boldsymbol{s})[n], \boldsymbol{s} \in \mathcal{S})$ действий, определенных на всем множестве $\boldsymbol{s} \in \mathcal{S}$, является стратегией управления на n-шаге. Тогда система уравнений (3.11) примет вид

$$g^{\boldsymbol{a}[n]} = R(\boldsymbol{s}) + \sum_{\boldsymbol{s}' \in \mathcal{S}} q_{a[n](\boldsymbol{s})}(\boldsymbol{s}, \boldsymbol{s}') v_{\boldsymbol{a}[n]}(\boldsymbol{s}'), \boldsymbol{s} \in \mathcal{S}.$$

Решая систему, получим оценки $v_{\boldsymbol{a}[n]}(\boldsymbol{s}'), \boldsymbol{s}' \in \mathcal{S}$ и среднее вознаграждение $g^{\boldsymbol{a}[n]}$ на n -шаге.

Для поиска стратегии на $n+1$ шаге найдем для каждого состояния $\boldsymbol{s} \in \mathcal{S}$ такое действие $a \in \mathcal{A}_{\boldsymbol{s}}$ из множества допустимых стратегий, которое будет максимизировать среднее вознаграждение $g^{\boldsymbol{a}[n+1]}, \boldsymbol{s} \in \mathcal{S}$:

$$a[n+1](\boldsymbol{s}) = \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ R(\boldsymbol{s}) + \sum_{\boldsymbol{s}' \in \mathcal{S}} q_a(\boldsymbol{s}, \boldsymbol{s}') v_{\boldsymbol{a}[n]}(\boldsymbol{s}') \right\} = \left\{ R(\boldsymbol{s}) = const \right\} =$$

$$= \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ \sum_{\boldsymbol{s}' \in \mathcal{S}} q_a(\boldsymbol{s}, \boldsymbol{s}') v_{\boldsymbol{a}[n]}(\boldsymbol{s}') \right\}, \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{S}.$$

С учетом вида подмножества допустимых стратегий, формула (3.2), действие $a[n+1](\boldsymbol{s})$ может принимать значение n/a, если $\boldsymbol{s} \notin \mathcal{S}_\delta$. Т.е. при поступлении сигнала контроллера действие $a \in \mathcal{A}_{\boldsymbol{s}}$ может быть выбрано только для состояний $\boldsymbol{s} \in \mathcal{S}_\delta$, в которых возможно инициировать перераспределение ресурсов, для остальных состояний $\boldsymbol{s} \notin \mathcal{S}_\delta$ такого перехода осуществить невозможно, следовательно распределение ресурсов будет оставаться неизменным. Состояние $\boldsymbol{s}' \in \mathcal{S}$, в которое будет осуществлен переход из $\boldsymbol{s} \in \mathcal{S}_\delta$ при поступлении сигнала, обозначим (a, l_1, l_2), где $a \in \mathcal{A}_{\boldsymbol{s}}$ новый объем ресурсов для 1-класса сессий (действие) и вынесем из-под знака суммы со своей интенсивностью δ (табл. 3.2). Тогда

$$a[n+1](\boldsymbol{s}) = \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ \sum_{\boldsymbol{s}' \in \mathcal{S}} q_a(\boldsymbol{s}, \boldsymbol{s}') v_{\boldsymbol{a}[n]}(\boldsymbol{s}') \right\} =$$

$$= \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ \delta \cdot v_{\boldsymbol{a}[n]}(a, l_1, l_2) + \sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \{a, l_1, l_2\}} q_a(\boldsymbol{s}, \boldsymbol{s}') v_{\boldsymbol{a}[n]}(\boldsymbol{s}') \right\}, \boldsymbol{s}, \boldsymbol{s}' \in \mathcal{S}.$$

Отсюда состояния $\boldsymbol{s}' \in \mathcal{S} \setminus \{a, l_1, l_2\}$ также не будут влиять на поиск действия, максимизирующего среднее вознаграждение, формула (3.11), поэтому

$$a[n+1](\boldsymbol{s}) = \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ \delta \cdot v_{\boldsymbol{a}[n]}(a, l_1, l_2) + \sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \{a, l_1, l_2\}} q_a(\boldsymbol{s}, \boldsymbol{s}') v_{\boldsymbol{a}[n]}(\boldsymbol{s}') \right\} =$$

$$= \left\{ \sum_{\boldsymbol{s}' \in \mathcal{S} \setminus \{a, l_1, l_2\}} q_a(\boldsymbol{s}, \boldsymbol{s}') v_{\boldsymbol{a}[n]}(\boldsymbol{s}') = const \right\} = \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ \delta \cdot v_{\boldsymbol{a}[n]}(a, l_1, l_2) \right\} =$$

$= \{\delta = const\} = \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ v_{\boldsymbol{a}[n]}(a, l_1, l_2) \right\}, \boldsymbol{s} \in \mathcal{S}$. Опуская индексы итераций $a(\boldsymbol{s}) = \arg\max_{a \in \mathcal{A}_{\boldsymbol{s}}} \left\{ v_{\boldsymbol{a}}(a, l_1, l_2) \right\}, \boldsymbol{s} \in \mathcal{S}$ получим формулу (3.12).

Утверждение доказано. □

Таким образом, был описан метод нахождения оптимальной политики, которая является улучшением некоторой начальной политики и обеспечивает большее вознаграждение. Далее две процедуры определения оценок и улучшения стратегии объединяются в итерационный алгоритм 3.1, с помощью которого находится оптимальная стратегия среди всех возможных стратегий, которая дает наибольшее вознаграждение. Алгоритм начинается с выбора начального решения, положив оценки $v_{\boldsymbol{a}}^{[0]}(\boldsymbol{s}) = 0$. Начальной стратегией будет решение, определенное алгоритмом 2.1. Далее начинается процедура определения оценок (шаг 4) и процедура улучшения стратегии (шаг 5). В случае, если решения двух последних итераций совпадет, оптимальная стратегия, при которой вознаграждение максимизировано, будет найдена.

Алгоритм 3.1. [146] Алгоритм вычисления оптимальной стратегии задается как

1: $n \leftarrow 0$

2: $v_{\boldsymbol{a}}^{[0]}(\boldsymbol{s}) = 0,\ \boldsymbol{s} \in \mathcal{S}$

3: $a[0](\boldsymbol{s}) = \begin{cases} l_1 & \boldsymbol{s} \in \mathcal{S}_{\delta}^{11} \\ N - l_2 & \boldsymbol{s} \in \mathcal{S}_{\delta}^{12} \\ m_1 - l_2 - N + m_1 & \boldsymbol{s} \in \mathcal{S}_{\delta}^{21} \\ l_1 & \boldsymbol{s} \in \mathcal{S}_{\delta}^{22} \end{cases}$ ▷ *начальная стратегия*

4: **решить**$\left(\textit{система}\left(3.10\right)\right)$ ▷ *Утв*.3.3

5: $a[n+1](\boldsymbol{s}) = \left(\textit{по формуле}\left(3.12\right)\right)$ ▷ *улучшение стратегии* (*Утв*.3.4)

6: **если** $a[n+1](\boldsymbol{s}) = a[n](\boldsymbol{s}), \boldsymbol{s} \in \mathcal{S}$ **то**

7: **вернуть** $a[n+1](\boldsymbol{s}), v_{\boldsymbol{a}}^{[n]}(\boldsymbol{s}), g^{\boldsymbol{a}[n]}$

8: **иначе** $n \leftarrow n+1$, **к шагу** 5

Для каждого состояния системы $\boldsymbol{s} \in \mathcal{S}$ выбирается новый объем ресурсов для 1-сегмента, если перераспределение произошло или не произошло (но состояние системы позволяло инициировать перераспределение), элементом матрицы является число – объем ресурсов для 1-сегмента. Если состояние системы заведомо не предполагает перераспределения ресурсов, элементом матрицы является прочерк. Таким образом, при нахождении системы в состоянии $(1,2,7)$, когда имеются свободные ресурсы 2-сегмента и ожидающие начала обслуживания сессии 2-сегмента, согласно выбранной стратегии, система должна инициировать перераспределение ресурса и перейти в состояние $(2,2,7)$ за счет увеличения ресурсов 1-сегмента. Аналогично – из состояния $(1,0,9)$ в состояние $(0,0,9)$ за счет уменьшения ресурсов 1-сегмента. Обратим внимание на состояние $(1,6,2)$, когда переход осуществляется в состояние $(5,6,2)$, т.е. движение границы осуществляется не на все ожидающие обслуживания сессии 1-класса, как при фиксированной стратегии, определенной в главе 2. Наконец, из состояния $(1,0,5)$ не осуществляется переход по изменению числа выделенных ресурсов, а в

состоянии $(1,2,8)$ перераспределение ресурсов не может быть инициировано, исходя из описания функционирования системы и разбиения пространства состояний (отсутствуют свободные ресурсы для движения границы), что в матрице отражено прочерком.

Пример получаемой в результате выполнения алгоритма оптимальной стратегии представлен в виде матрицы на рис. 3.2, где изменение числа сессий 1-класса l_1 отражено по строчкам, числа сессий 2-класса l_2 по столбцам. А на рис. 3.3. показан пример изменения стратегии от итерации к итерации (оптимальной стратегии), который соответствует графическому представлению пространства состояний, определенном в разделе 3.1.

m1 = 1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	1	1	1	1	1	1	1	1	—	0	0	0	0	0
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	2	2	2	2	2	2	2	2	—	—	—	—	—	—
3	2	2	2	2	2	2	2	2	—	—	—	—	—	—
4	2	2	2	2	2	2	2	2	—	—	—	—	—	—
5	5	5	5	5	5	2	2	2	—	—	—	—	—	—
6	5	5	5	5	5	2	2	2	—	—	—	—	—	—

Рис. 3.2. Фрагмент оптимальной стратегии для $V=10$, $R_1=R_2=5$ при $m_1=1$ [147]

Итерация 1

$m_1=0$	0	1	2	3
0	*nan*	*nan*	*nan*	*nan*
1	1	1	*nan*	*nan*

$m_1=1$	0	1	2
0	1	*nan*	0
1	*nan*	*nan*	*nan*
2	2	*nan*	*nan*

$m_1=2$	0	1
0	*nan*	1
1	*nan*	1
2	*nan*	*nan*
3	*nan*	*nan*

Итерация 2

$m_1=0$	0	1	2	3
0	*nan*	*nan*	*nan*	*nan*
1	0	0	*nan*	*nan*

$m_1=1$	0	1	2
0	1	*nan*	0
1	*nan*	*nan*	*nan*
2	1	*nan*	*nan*

$m_1=2$	0	1
0	*nan*	1
1	*nan*	1
2	*nan*	*nan*
3	*nan*	*nan*

Итерация 3 (оптимальная)

$m_1=0$	0	1	2	3
0	*nan*	*nan*	*nan*	*nan*
1	0	0	*nan*	*nan*

$m_1=1$	0	1	2
0	1	*nan*	0
1	*nan*	*nan*	*nan*
2	2	*nan*	*nan*

$m_1=2$	0	1
0	*nan*	2
1	*nan*	2
2	*nan*	*nan*
3	*nan*	*nan*

Рис. 3.3. Иллюстрация изменения стратегии по итерациям для $V=10$, $R_1=R_2=5$

Пример 3.1. На примере 2.3 модели для сценария групповой передачи данных и просмотра веб-страниц была написана программа на языке Python. Для следующих исходных данных проводится сравнение показателей эффективности нарезки ресурсов, определенных в разделе 2.4, коэффициента успеха перераспределения ресурса (рис. 3.4) и коэффициента использования ресурса (рис. 3.5) при использовании фиксированной политики и политики управления выбора объемом перераспределения ресурсов: $R_1 = R_2 = 2$, $c_1 = c_2 = c_3 = 1$, $b = 1{,}067$, $\mu_1 = 0{,}125, \mu_2 = 0{,}937$, $\delta = 0{,}000001$. Объем ресурса базового оператора $V = 8{,}016$ (пропускная способность 5 МГц, схема модуляции QPSK и схема MIMO 2x2).

Полученные графики отражают, что при увеличении общей нагрузки на систему снижается число сигналов, при которых происходит перераспределение ресурсов. Это может быть связано с состояниями, в которых перераспределение ресурсов не может быть инициировано из-за отсутствия свободных ресурсов. Одновременно с этим при использовании оптимальной стратегии управления ресурсов коэффициент растет по сравнению с фиксированной стратегией (рис. 3.4). Что же касается коэффициента использования ресурса (рис. 3.5), для фиксированной стратегии он возрастает при увеличении общей нагрузки на систему. Однако, оптимальная стратегия позволяет равномерно увеличивать занятость ресурса при увеличении нагрузки. При этом стоит отметить, что исходные данные подобраны так, чтобы в равной степени оптимизировать три показателя эффективности нарезки (утверждение 3.2) $c_1 = c_2 = c_3 = 1$, если же какой-то из показателей становится приоритетнее, необходимо скорректировать данные значения.

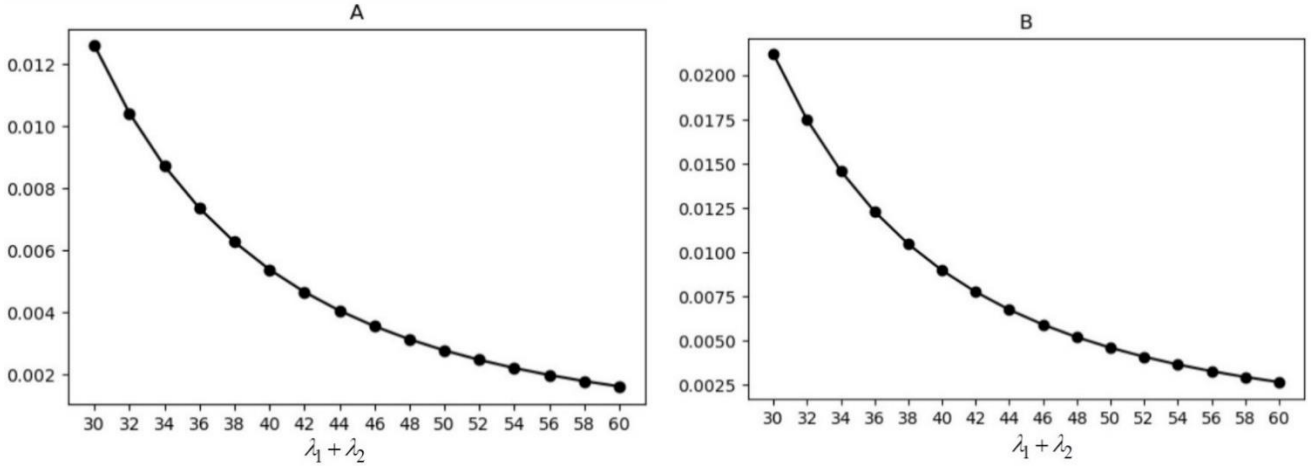


Рис. 3.4. Коэффициент успеха перераспределения ресурса: фиксированная (A) и оптимальная (B) стратегии

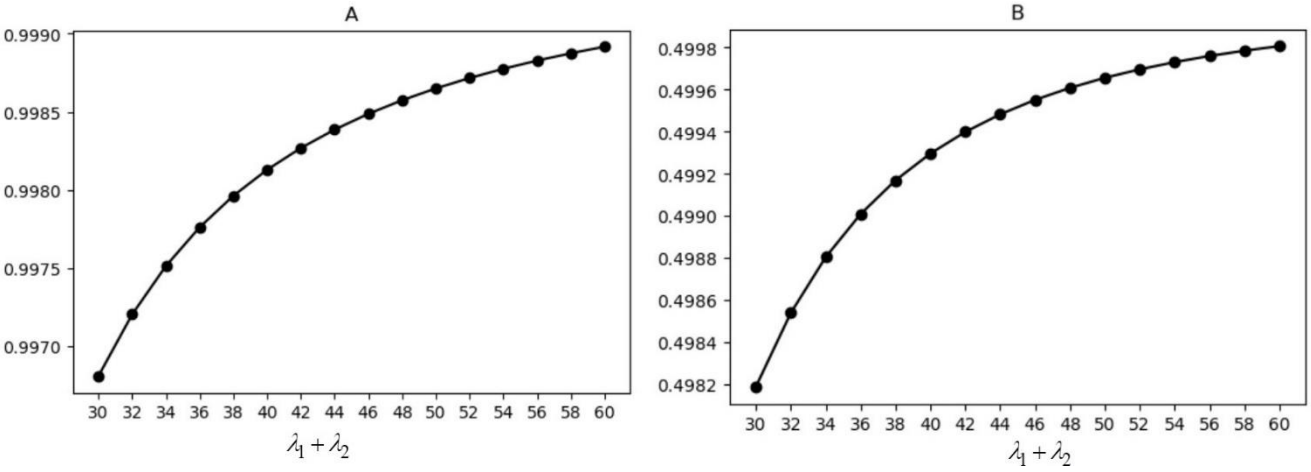


Рис. 3.5. Коэффициент использования ресурса: фиксированная (A) и оптимальная (B) стратегии

3.4. Имитационная модель для произвольного числа сегментов сети

В разделе 1.2 построена математическая модель одного сегмента сети для анализа влияния различных распределений длин передаваемых файлов на показатели эффективности функционирования системы, в главе 2 была построена математическая модель уже для двух сегментов сети с фиксированной политикой перераспределения ресурсов при поступлении сигнала контроллера для анализа влияния частоты поступления таких сигналов на показатели эффективности. В разделах 3.1–3.3 построена модель с управляемым выбором объема

перераспределения ресурсов для анализа влияния выбора стратегии управления на показатели эффективности нарезки ресурсов. Интерес также представляет исследование систем с более, чем двумя сегментами, дальнейшее исследование строится относительно имитационной модели для большего числа сегментов сети. Графическое представление такой модели представлено на рис. 3.6, перераспределение ресурсов также будет осуществляться по сигналу контроллера и иметь фиксированную стратегию управления, что соответствует модели главы 2 и является расширением для K-сегментов, $\mathcal{K} = \{1, 2, \ldots, K\}, k \in \mathcal{K}$.

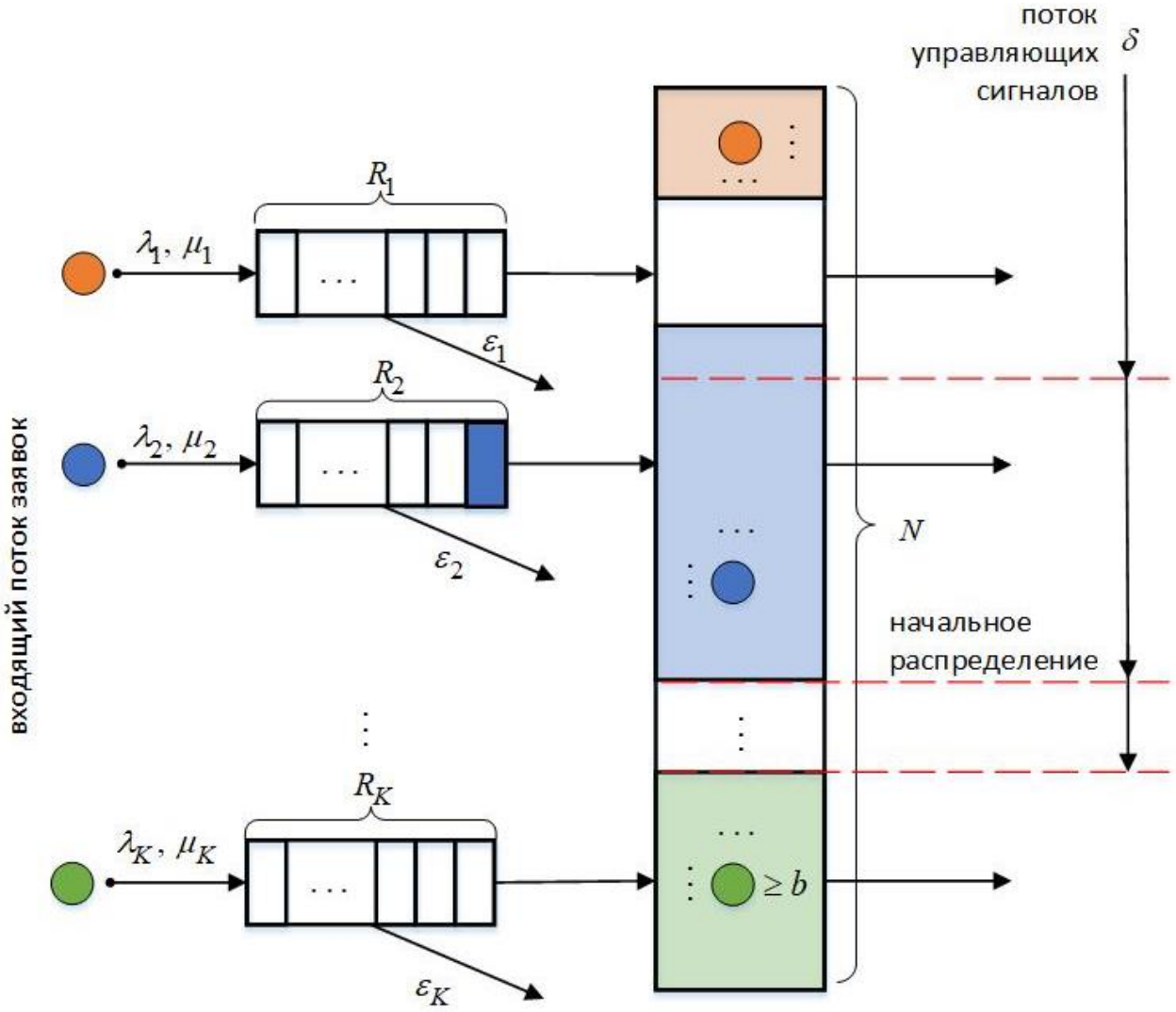


Рис. 3.6. Схема СМО для K-сегментов

Системные параметры аналогичны предыдущим главам: $\lambda_k, k \in \mathcal{K}$ интенсивность входящего потока k-сегмента; $b_k, k \in \mathcal{K}$ минимальная гарантированная скорость обслуживания (введем упрощающее предположение

$b_k = 1$); $\mu_k^{-1}, k \in \mathcal{K}$ объём блока данных k-сегмента; $\varepsilon_k, k \in \mathcal{K}$ интенсивность покидания системы по причине нетерпеливости; $R_k, k \in \mathcal{K}$ определяет максимальное число ожидающих обслуживания запросов; N максимальное число одновременно обслуживаемых запросов, $N_k = [0; N], k \in \mathcal{K}, \sum_{k \in \mathcal{K}} N_k = N$; δ интенсивность сигналов контроллера о перераспределении ресурсов. Распределение ресурсов является динамическим и меняется с течением времени, обозначим $N_k(t)$ максимальное число запросов k-сегмента в некоторый момент времени $t > 0$. Обозначим t_l момент времени поступления l-сигнала о нарезке ресурсов, тогда, исходя из показателя соответствия начальной нарезки, определенном в разделе 2.4, в начальный момент времени t_0 распределение ресурсов является равным, $N_1(t_0) = N_2(t_0) = \ldots = N_K(t_0)$. Отсортируем сегменты так, что $\lambda_1 < \lambda_2 < \ldots < \lambda_K$, тогда для l-нарезки распределение ресурсов (объемы выделенных ресурсов) задается вектором $\boldsymbol{N}(t_l) = (N_1(t_l), N_2(t_l), \ldots, N_K(t_l))$. Пусть $\boldsymbol{n}(t) = (n_1(t), n_2(t), \ldots, n_K(t))$ – число обслуживаемых запросов каждого из K-сегментов в момент времени t, а $\boldsymbol{m}(t) = (m_1(t), m_2(t), \ldots, m_K(t))$ – число ожидающих обслуживания запросов каждого из K-сегментов в момент времени t. Состояние системы описывается парой векторов $(\boldsymbol{n}(t), \boldsymbol{m}(t)) = ((n_1(t), m_1(t)), (n_2(t), m_2(t)), \ldots, (n_K(t), m_K(t)))$ над пространством состояний

$$\mathcal{X} = \prod_{k \in \mathcal{K}} \{(n_k(t), m_k(t)) : n_k(t) \geq 0,\ m_k(t) \geq 0,$$
$$(n_k(t), 0) : n_k(t) = 0, 1, \ldots, N_k(t) - 1,$$
$$(N_k(t), m_k(t)) : m_k(t) = 0, 1, \ldots, R_k\}.$$

Функционирование системы для K-сегментов предполагает три возможных соотношения для числа ожидающих начала обслуживания запросов:

– $\forall\ m_k(t_l) = 0,\ k \in \mathcal{K}$ – отсутсвуют запросы, ожидающие обслуживания $\rightarrow$

распределение ресурсов не меняется, $\boldsymbol{N}(t_l) = \boldsymbol{N}(t_{l-1})$;

– $\forall\, m_k(t_l) > 0,\ k \in \mathcal{K}$ – во всех сегментах имеется, по крайней мере, один запрос, ожидающий обслуживания $\rightarrow$ распределение ресурсов не меняется, $\boldsymbol{N}(t_l) = \boldsymbol{N}(t_{l-1})$;

– $\exists\, k \in \mathcal{K} : m_k(t_l) > 0;\ \exists\, j \in \mathcal{K} : m_j(t_l) = 0$ – по крайней мере один сегмент имеет ожидающие начала обслуживания запросы $\rightarrow$ необходимо перераспределение ресурсов.

Для случая, когда имеется $K = 2$ сегментов, схематическое изображение изменения распределения ресурсов с течением времени представлено на рис. 3.7 и представляет частный случай системы с произвольным числом сегментов сети (такая модель исследовалась в главе 2). Тогда описанные выше соотношения примут вид:

a) $m_1(t_l) > 0, m_2(t_l) > 0 \cup m_1(t_l) = 0, m_2(t_l) = 0$ – если в сегментах отсутствуют ожидающие запросы или все сегменты имеют по крайней мере по одному ожидающему запросу, распределение ресурсов не меняется, $N_1(t_l) = N_1(t_{l-1}), N_2(t_l) = N_2(t_{l-1})$;

b) $m_1(t_l) > 0, m_2(t_l) = 0$ – если в 1-сегменте имеются ожидающие запросы, а во 2-сегменте нет;

c) $m_1(t_l) = 0, m_2(t_l) > 0$ – если в 2-сегменте имеются ожидающие запросы, а в 1-сегменте нет.

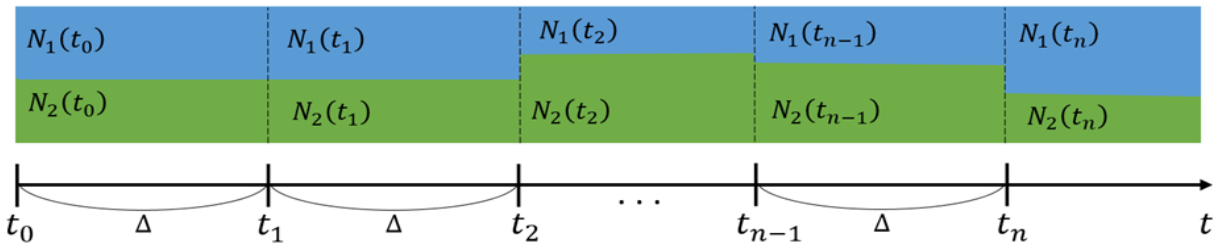


Рис. 3.7. Динамическое перераспределение ресурса для $K = 2$

В случае b (для случая (c) ситуации аналогичны, поэтому здесь не приводятся) в зависимости от состояния системы при перераспределении ресурсов могут возникать две ортогональные ситуации:

I. $m_1(t_l) \le N_2(t_{l-1}) - n_2(t_l)$ – числа свободных ресурсов 2-сегмента достаточно для обслуживания числа ожидающих запросов 1-сегмента;

II. $m_1(t_l) > N_2(t_{l-1}) - n_2(t_l)$ – числа свободных ресурсов 2-сегмента недостаточно для обслуживания числа ожидающих запросов 1-сегмента.

Тогда новый объем ресурсов определяется (данный принцип учтен в главе 2 при построении матрицы интенсивностей переходов) [142]

$$N_1(t_l) = \begin{cases} N_1(t_{l-1}) + m_1(t_l), & m_1 \le N_2(t_{l-1}) - n_2(t_l), \\ N_1(t_{l-1}) + N_2(t_{l-1}) - n_2(t_l), & m_1(t_l) > N_2(t_{l-1}) - n_2(t_l), \end{cases}$$
$$N_2(t_l) = \begin{cases} N_2(t_{l-1}) - m_1, & m_1 \le m_1 \le N_2(t_{l-1}) - n_2, \\ n_2, & m_1 > N_2(t_{l-1}) - n_2. \end{cases} \quad (3.13)$$

В случае, когда в системе имеется более, чем два сегмента необходимо определить не только хватает ли свободных ресурсов для обслуживания ожидающих начала обслуживания запросов, но и как выбрать сегмент, который будет увеличен и определеить сегмент, за счет которого будет увеличен сегмент. Алгоритм перераспределения ресурсов для K-сегментов представлен в алгоритме 3.2.

Алгоритм 3.2. [146] Алгоритм перераспределения ресурса по стратегии управления выбором объема ресурса

1: $\overline{n.} = 0, r. = 0$

2: **for** $k, K \in \mathcal{K}$ **do**

3: $\overline{n_k} = m_k - n_k$

4: $\overline{n.} = \overline{n.} + \overline{n_k}$ ▷ *сумма доступных ресурсов*

5: $r. = r. + r_k$ ▷ *сумма всех ожидающих запросов*

6: **endfor**

7: $i := \min\left(\overline{n.}, r.\right)$

8: **repeat**

9: $s = \arg\max\left(\overline{n_k}\right), k \in \mathcal{K}$

10: $q = \arg\max\left(r_k\right), k \in \mathcal{K}$

11: $m_s := m_s - 1$ ▷ *уменьшение s - сегмента на 1 единицу*

12: $m_q := m_q + 1$ ▷ *увеличение q - сегмента на 1 единицу*

13: $r_q := r_q - 1$

14: $n_q := n_q + 1$

15: $\overline{n_s} := \overline{n_s} - 1$ ▷ *уменьшение числа доступных ресурсов*

16: $i := i - 1$

17: **until** $s > 0 \text{ and } i > 0$

Следовательно, в случае перераспределения ресурсов определяется число свободных единиц ресурса, доступных для перераспределения (шаги 3 и 4), и число ожидающих запросов для каждого сегмента (шаг 5). Затем находится наименьшая разница между числом свободных ресурсов и числом ожидающих запросов (шаг 7). И, наконец, перераспределяется свободный ресурс между самыми загруженными сегментами по числу ожидающих запросов, пока свободный ресурс не исчерпается или все ожидающие запросы не начнут обслуживание.

3.5. Численный анализ показателей эффективности нарезки ресурсов

Принципы эффективности нарезки ресурсов и их показатели уже были определены в разделах 2.4 и 3.2, это коэффициент успеха перераспределения ресурса, коэффициент использования ресурса и коэффициент соответствия начальному распределению. Аналогично определим показатели эффективности для произвольного числа сегментов. Первый показатель – среднее по коэффициентам соответствия начальному распределению вычисляется как

$$\alpha = \frac{1}{K}\sum_{k\in\mathcal{K}} a_k, a_k = \lim_{T\to\infty} \frac{\sum_{l=1}^{L(T)}\left(\left(t_l - t_{l-1}\right)\cdot \min\left(N_k\left(t_l\right), N_k\left(t_0\right)\right)\right)}{T\cdot N_k\left(t_0\right)}, a_k \in [0,1], \qquad (3.14)$$

где T модельное время, $L(T)$ общее количество вызовов нарезки. Используя закон полной вероятности для выбранного ресурса $N_k(t_l)$ для *k*-сегмента с начальной нарезкой $N_k(t_0)$ в момент времени t_l: $P(A|H_i) = P\{N_k(t_0) \le N_k(t) \,|\, t = t_l - t_{l-1}\}$. Это может быть выражено как $\frac{N_k(t_l)}{N_k(t_0)}$, что показывает, насколько выделенное количество ресурсов близко к начальной нарезке. Тем не менее, параметр $N_k(t_l)$ может иметь большее численное значение, чем $N_k(t_0)$. Следовательно, для правильного вычисления этой вероятности, необходимо выбрать минимальное значение из этих двух величин

$$P(A|H_i) = P\left\{N_k(t_l) = N_k(t_0)\right\} = \frac{\min\left(N_k(t_l), N_k(t_0)\right)}{N_k(t_0)}.$$

Вероятность того, что событие произойдет через требуемый интервал времени, может быть записана как $P(H_i) = P\left\{t = t_l - t_{l-1}\right\} = \frac{t_l - t_{l-1}}{T}$, отсюда

$$P(A) = \sum_{l=1}^{L(T)} \cdot \frac{\left(t_l - t_{l-1}\right)}{T} \frac{\min\left(N_k(t_l), N_k(t_0)\right)}{N_k(t_0)}, k \in \mathcal{K}.$$

Далее остается только найти предел этого выражения при $T \to \infty$ и просуммировать его по всем сегментам, затем поделить на количество таких сегментов. Получим формулу для расчета коэффициента соответствия начальному распределению (3.14).

Напомним, что перераспределение ресурсов осуществляется только при поступлении сигнала контроллера, которые направляются с некоторым интервалом Δ. Если же сигналов будет слишком много, это приведет к перегрузке сети. Следовательно, необходимо определить те вызовы нарезки сети, которые

оказались полезными, т.е. привели к перераспределению ресурсов. Второй показатель доля полезных вызовов нарезки $\beta \in [0,1]$ вычисляется как

$$\beta = \lim_{T\to\infty} \frac{\sum_{l=1}^{L(T)} \max_{k\in\mathcal{K}} 1\{N_k(t_l) \neq N_k(t_l - 0)\}}{L(T)+1}. \quad (3.15)$$

Рассмотрим все вызовы алгоритма перераспределения ресурсов и вычислим количество тех, для которых изменилось значение $N_k(t)$. Найдем количество полезных нарезок $\max_{k\in\mathcal{K}} 1\{N_k(t_l) \neq N_k(t_l - 0)\}$, где $t_l - 0$ момент времени перед l-сигналом о нарезке. Результатом этой функции будет 0, если изменений нет, и 1 - в противном случае. Просуммируем эти значения по всем вызовам нарезки сети $L(T)$ и поделим на общее число вызовов нарезки, получим формулу для расчета доли полезных вызовов нарезки (3.15).

Наконец, для увеличения пропускной способности сети необходимо сократить время простоя системы, когда один из сегментов перегружен, а в другом есть свободные ресурсы. Поэтому третий показатель среднее значение по коэффициентам занятости вычисляется как

$$\gamma = \frac{1}{K}\sum_{k\in\mathcal{K}} \gamma_k,\ \gamma_k = \lim_{T\to\infty} \frac{\sum_{i=1}^{I(T)} (t_i - \tau_i)}{T},\ \gamma \in [0,1], \quad (3.16)$$

где τ_i и t_i начало и конец занятости ресурса, а $I(T)$ номер этого интервала. Тогда задача оптимизации выбора частоты поступления сигналов контроллера Δ записывается в виде $\alpha(\Delta) + \beta(\Delta) + \gamma(\Delta) \to \max$.

Для численного анализа системы с $k \in \mathcal{K}$ сегментами построена дискретно-событийная модель в среде Java. Выделим три типа событий (рис. 3.8): A – поступление запроса в систему, B – завершение обслуживания запроса, C – уход запроса из системы по причине нетерпеливости.

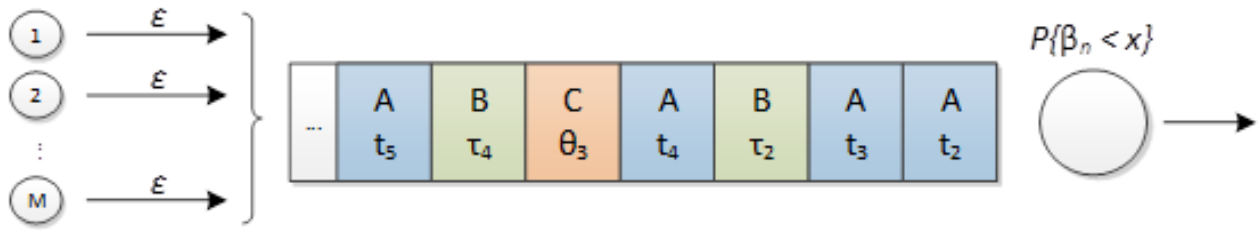


Рис. 3.8. Схема очереди событий в имитационной модели

Схема имитационного моделирования события А представлена на рис. 3.9, когда при поступлении нового запроса в систему он отправляется либо на обслуживание, либо в очередь, если нет мест – блокируется. Особенностью является эластичный трафик, когда скорость обслуживания запроса определяется пропорционально числу запросов на обслуживании, следовательно, и время завершения обслуживания будет изменяться.

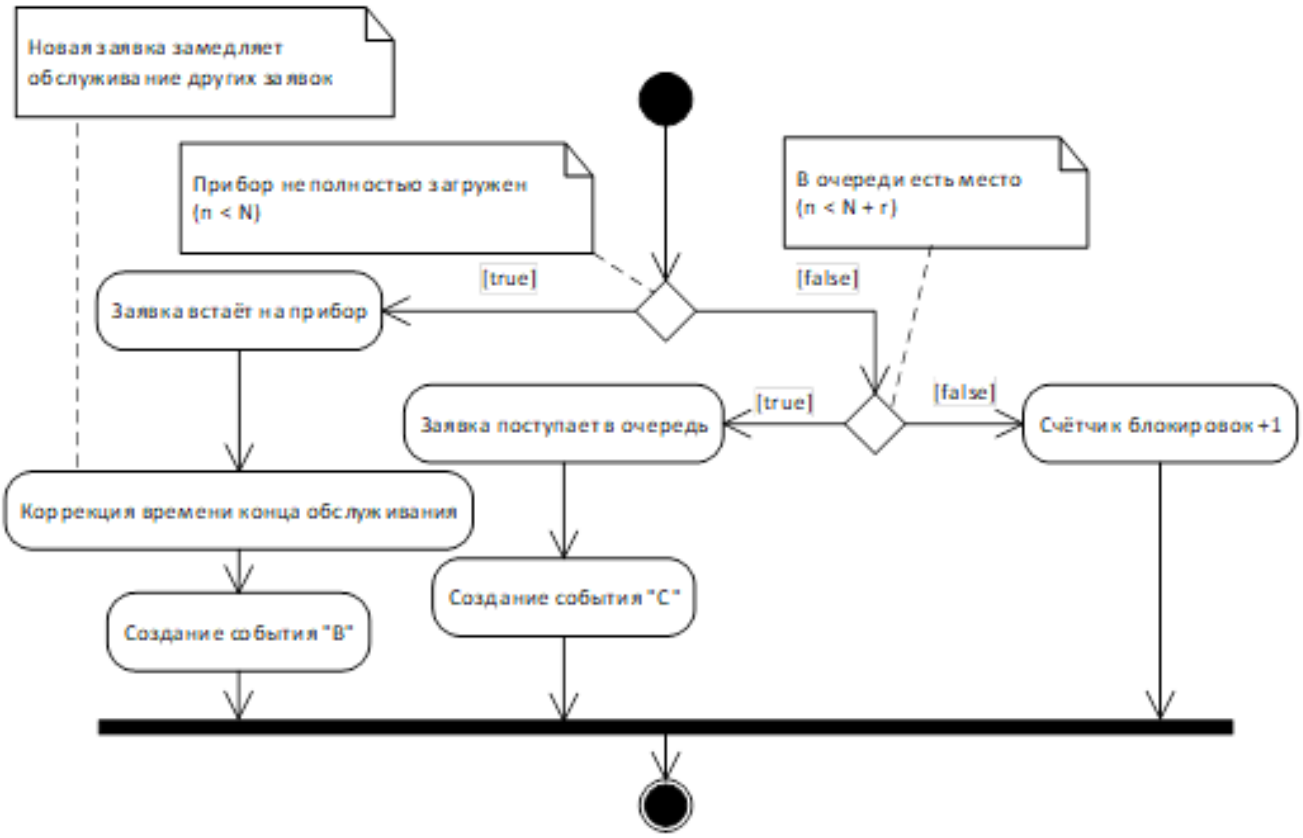


Рис. 3.9. Схема обработки события поступления запроса в систему (А) в имитационной модели

Следующее событие завершения обслуживания В также учитывает особенности эластичного трафика, когда на освободившийся ресурс не поступает нового запроса из очереди. Аналогично событию А необходимо скорректировать

время завершения обслуживания оставшихся обслуживаемых запросов (схема моделирования представлена на рис. 3.10). Схема события ухода запроса из системы по причине нетерпеливости С здесь не приводится, т.к. не влияет на время завершения обслуживания.

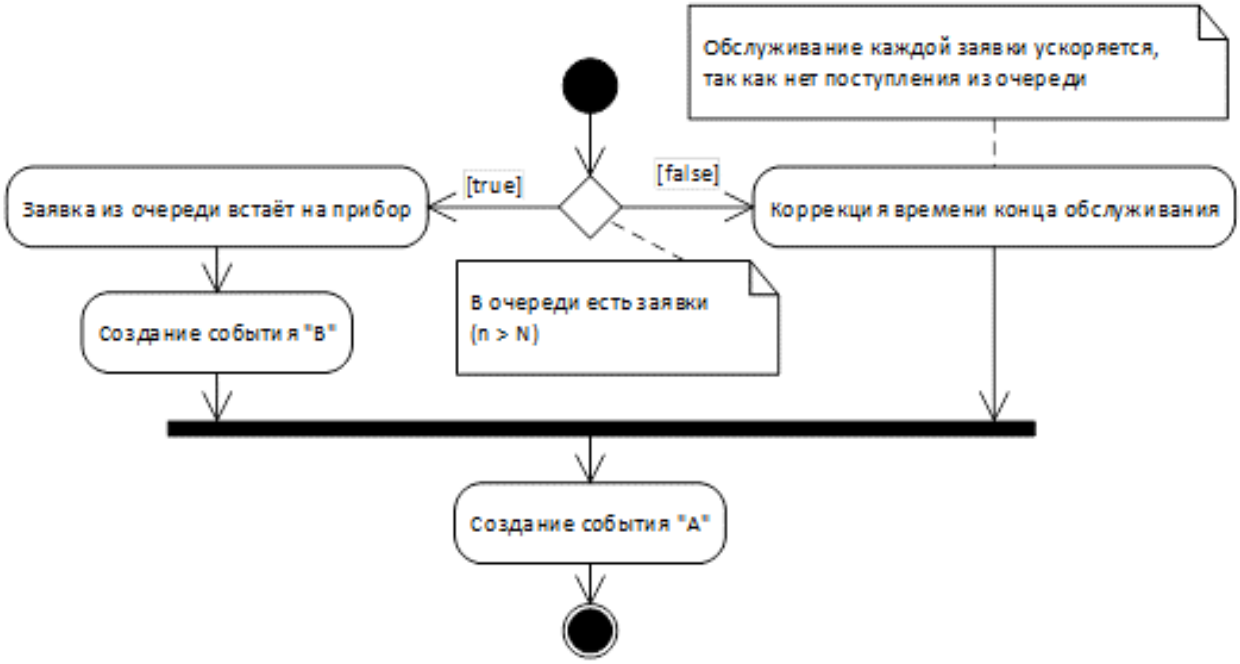


Рис. 3.10. Схема обработки события завершения обслуживания (B) в имитационной модели

Пример 3.2. Для следующих исходных данных проведем имитационное моделирование для анализа частоты поступления сигналов. Пусть имеется $V = 60$ Мбит/с ресурсов базового оператора и $K = 5$ виртуальных операторов, запрашивающих доступ к ресурсам базового оператора. В качестве трафика выбирается услуга передачи файлов, которая характеризуется объемом передаваемого файла (экспоненциальное распределение со средним значением $\mu^{-1} = 2$ Мбит). Предполагается, что интервалы времени между запросами на загрузку файлов распределены экспоненциально со параметрами $\lambda_1 = 2$ для 1-сегмента и $\lambda_2 = 5$, $\lambda_3 = 8$, $\lambda_4 = 10$, $\lambda_5 = 15$ для других сегментов соответсвенно. При этом внутри каждого сегмента обеспечивается одинаковая минимальная гарантированная скорость передачи, $b = 0{,}1$ Мбит/с. Максимальное число

ожидающих начала обслуживания составляет $R = 50$. Запросы могут покинуть очередь по причине нетерпеливости с интенсивностью $\varepsilon = 0,000001$.

Выбор таких параметров системы обуславливается желанием продемонстрировать работу алгоритма выделения ресурсов, описанного выше. На рис. 3.11 показана зависимость каждого из выбранных показателей эффективности от интервала между нарезками сети $\Delta = \delta^{-1}$, а также среднее значение для них, выполненное черной пунктирной линией. Принимая во внимание основные принципы, заложенные в алгоритме, выберем сегмент, в котором зеленая точка и черная пунктирная линия принимают значения, близкие к 1 одновременно. В качестве решения предлагается любое значение Δ на интервале $150 - 200$ мс.

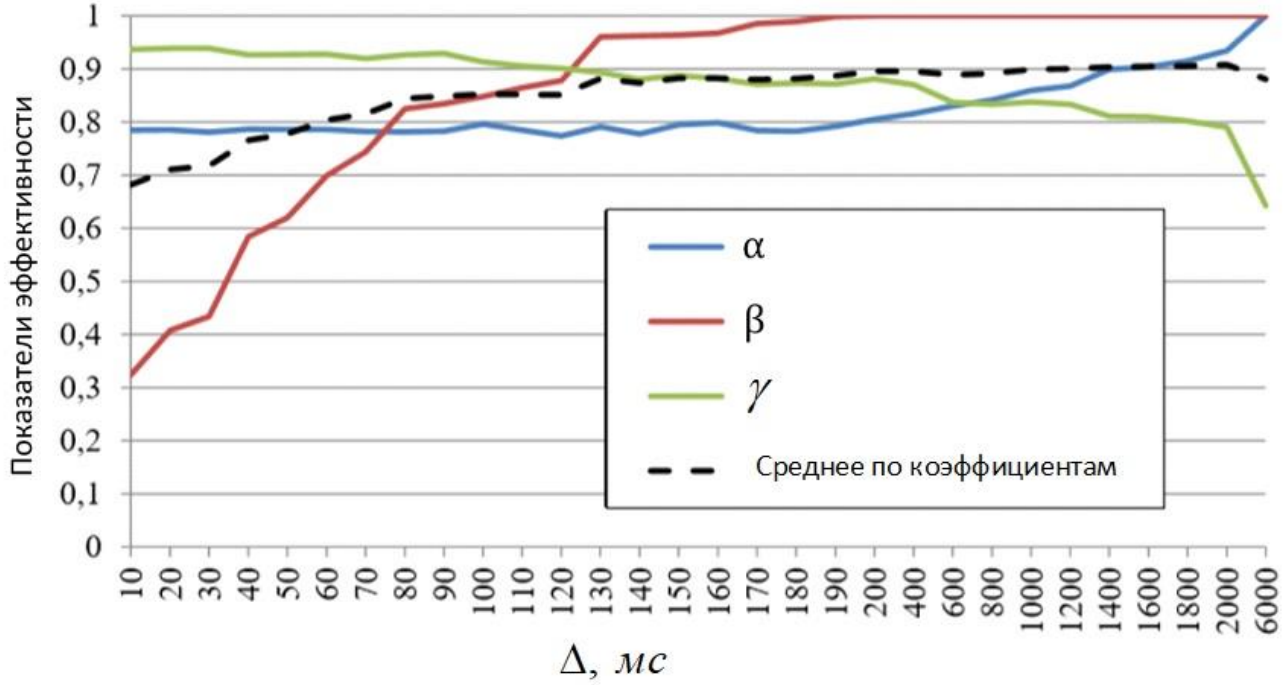


Рис. 3.11. Показатели эффективности нарезки [146]

Схожий результат получен на рис. 3.12, который показывает значения коэффициентов занятости ресурса отдельно для каждого сегмента. Нетрудно заметить, что существенные изменения по сравнению с начальной нарезкой происходят для первых двух сегментов (в первом используется только четверть всего объема, а на 2-сегменте более, чем 40% ресурсов не используется). Соответственно, графики показывают, что значение Δ, которое дает значительное увеличение нагрузки на ресурс, составляет 200 мс (при дальнейшем уменьшении

интервала между нарезками сильных изменений не происходит). Результаты моделирования для Δ=200 мс показаны на рис. 3.13, где можно наблюдать, что значительные скачки в свободных ресурсах происходят для сегментов с более высокой интенсивностью поступления запросов.

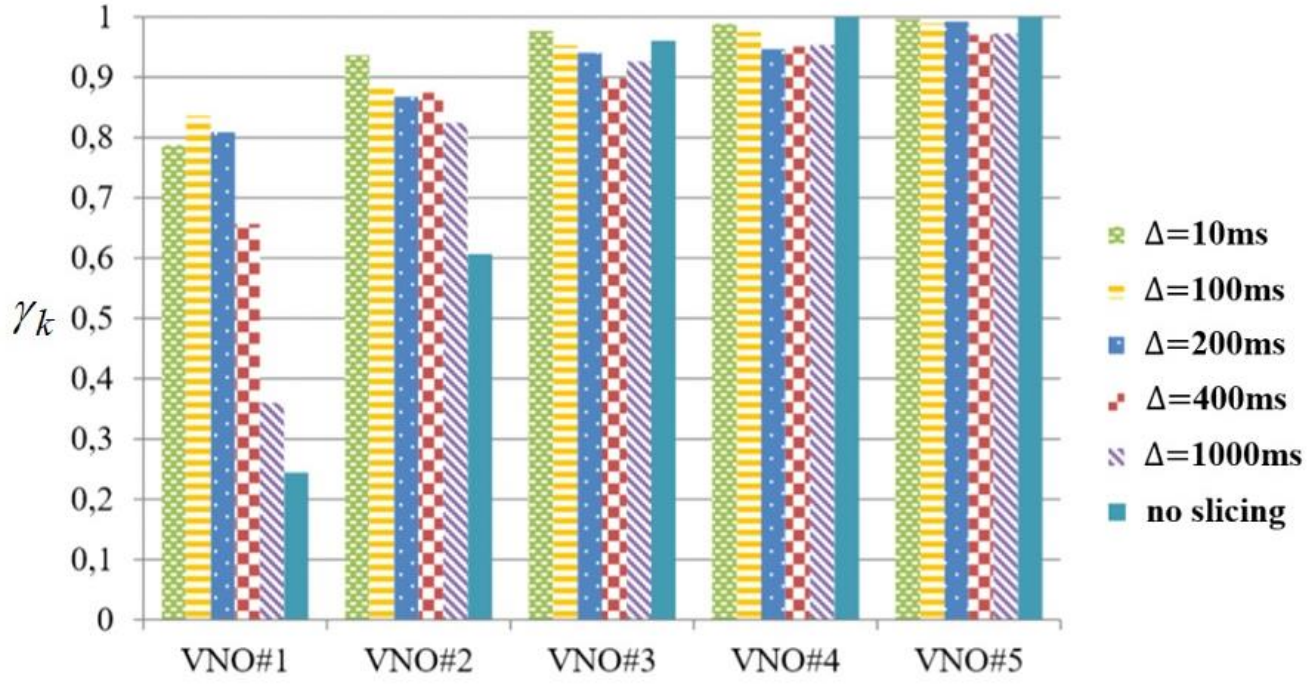


Рис. 3.12. Коэффициент использования ресурса по виртуальным операторам [146]

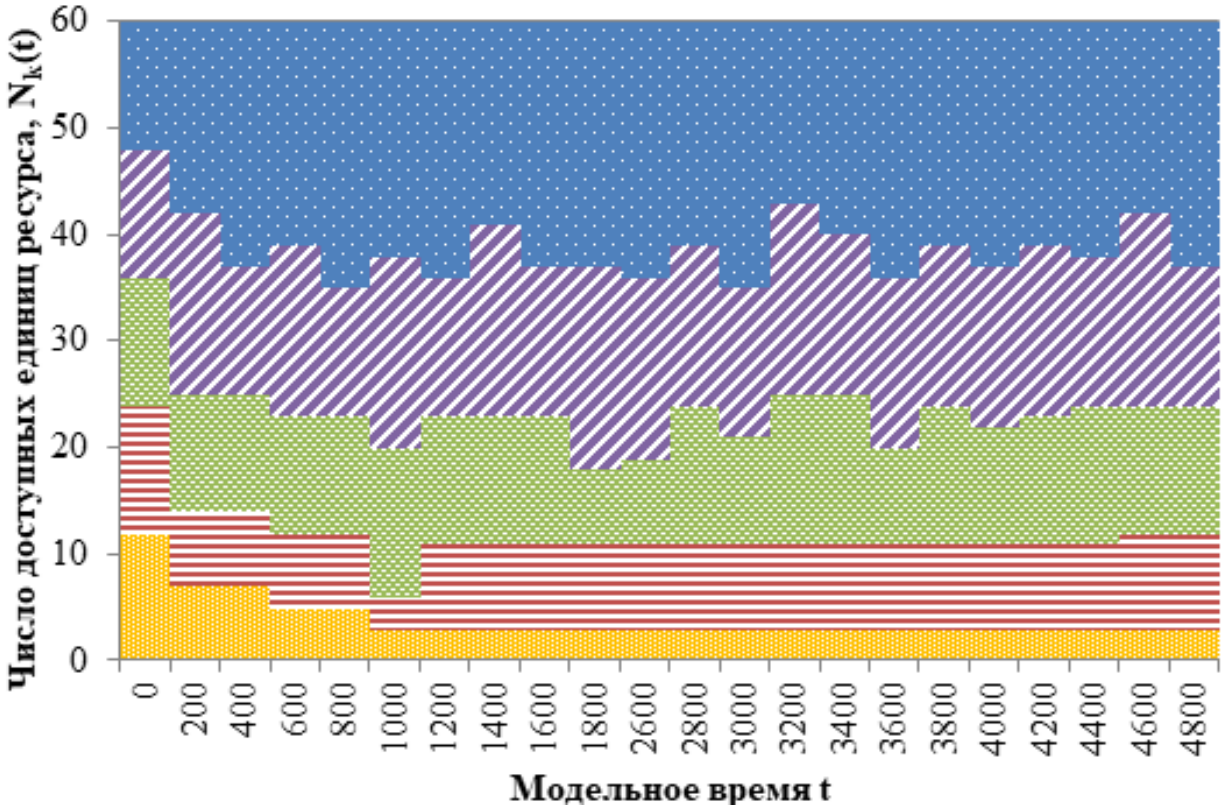


Рис. 3.13. Динамическое перераспределение ресурса для Δ=200 мс [146]

Таким образом, в третьей главе построена управляемая система массового обслуживания с эластичным трафиком и стратегией выбора объема перераспределения ресурса по сигналам, применен итерационный алгоритм вычисления оптимальной стратегии для настройки параметров динамической нарезки сети с учетом простоя ресурса, отклонения распределения ресурса от значений в соглашении о качестве обслуживания и вероятности перераспределения ресурса по сигналу. Для произвольного числа сегментов сети построена дискретно-событийная модель, позволяющая настроить частоту поступления сигналов для максимизации взвешенных коэффициентов использования ресурса и соответствия распределения ресурса соглашению о качестве обслуживания и вероятности перераспределения ресурса по сигналу.

Заключение

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

1. Разработана модель динамической нарезки радиоресурсов сети пятого поколения между виртуальными операторами, передающими данные с минимальной скоростью в виде системы массового обслуживания с эластичным трафиком и сигналами, по которым происходит перераспределение ресурса. Ограничение пользователя на время ожидания начала обслуживания моделируется нетерпеливыми заявками. Для фиксированной стратегии выбора объема перераспределения ресурса, ориентированной на максимальное его использование, матрица интенсивностей переходов записана в блочном трехдиагональном виде. Получен матричный рекуррентный алгоритм расчета стационарного распределения вероятностей.
2. Модель нарезки сети со стратегией управления выбором объема перераспределения ресурса разработана в виде управляемой системы массового обслуживания. Функция вознаграждения отражает простой ресурса, отклонение распределения ресурса от значений в соглашении о качестве обслуживания, вероятность перераспределения ресурса по сигналу. Система уравнений относительно функций среднего вознаграждения и оценок записана в виде для итерационного метода решения. Получен вид целевой функции для улучшения стратегии управления. Применен итерационный алгоритм вычисления оптимальной стратегии.
3. Проведено дискретно-событийное моделирование системы с произвольным числом сегментов сети с алгоритмом перераспределения ресурса, ориентированном на равномерное занятие простаивающих ресурсов и выбором сегментов с большим числом ожидающих начала обслуживания пользователей. Формализована задача максимизации показателей эффективности нарезки ресурсов со стороны базового оператора –

взвешенных коэффициентов использования ресурса и соответствия распределения ресурса соглашению о качестве обслуживания, вероятности перераспределения ресурса по сигналу по частоте поступления сигналов. При ее выборе учитываются также ограничения на вероятности блокировки запросов на передачу эластичного трафика виртуальных операторов.

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

V – радиочастотный ресурс базового оператора

N – число одновременно обслуживаемых сессий

K – число сегментов (для 1 главы – число входящих потоков)

b_k – минимально гарантированная скорость передачи данных k-сегмента

R_k – длина очереди

λ_k – интенсивность входящего потока k-сегмента

μ_k – интенсивность потока обслуживания k-сегмента

ε_k – интенсивность нетерпеливого потока k-сегмента

δ – интенсивность управляющих сигналов

$\boldsymbol{x},\boldsymbol{s}$ – состояние системы массового обслуживания

m_k – порог на максимальное число обслуживаемых сессий k-сегмента

n_k – число обслуживаемых сессий k-сегмента

r_k – число ожидающих начала обслуживания сессий k-сегмента

l_k – число сессий k-сегмента

$\chi(\boldsymbol{s})$ – коэффициент, отражающий принцип равного деления ресурсов

β – коэффициент, отражающий принцип успешного перераспределения ресурсов

γ – коэффициент, отражающий использования ресурса

B_k – вероятность блокировки сессии по заявкам k-сегмента

$\mathcal{A}_{\boldsymbol{s}}$ – множество допустимых стратегий

$\boldsymbol{a}$ – стратегия управления

$R(\boldsymbol{s})$ – функция вознаграждения

c_1, c_2, c_3 – весовые коэффициенты

$g^{\boldsymbol{a}}$ – функция среднего вознаграждения

$v_{\boldsymbol{a}}(\boldsymbol{s})$ – оценки для итерационного алгоритма

Литература

1. Sevastianov, L.A., Vasilyev, S.A. Telecommunication market model and optimal pricing scheme of 5G services // International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, 2019, 2018-November,8631269.
2. Akyildiz, I.F., Kak, A., Khorov, E., Krasilov, A., Kureev, A. ARBAT: A flexible network architecture for QoE-aware communications in 5G systems // Computer Networks, 2018, 147, c. 262-279.
3. Popovski, P., Trillingsgaard, K.F., Simeone, O., Durisi, G. 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view // IEEE Access, 2018, 6,8476595, c. 55765-55779.
4. Vikhrova, O., Suraci, C., Tropeano, A., Pizzi, Sara, Samouylov, K., Araniti, G. Enhanced Radio Access Procedure in Sliced 5G Networks // International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, 2019, 2019-October,8970776.
5. Khan, R., Kumar, P., Jayakody, D.N.K., Liyanage, M. A Survey on Security and Privacy of 5G Technologies: Potential Solutions, Recent Advancements, and Future Directions // IEEE Communications Surveys and Tutorials, 2020, 22(1),8792139, c. 196-248.
6. Ateya, A.A., Alhussan, A.A., Abdallah, H.A., Al duailij, M.A., Khakimov, A., Muthanna, A. Edge Computing Platform with Efficient Migration Scheme for 5G/6G Networks // Computer Systems Science and Engineering, 2023, 45 (2), pp. 1775-1787.
7. Ateya, A.A., Muthanna, A., Koucheryavy, A., Maleh, Y., El-Latif, A.A.A. Energy efficient offloading scheme for MEC-based augmented reality system // Cluster Computing, 2023, 26 (1), pp. 789-806.
8. Пшеничников А. П. Даудов И. М. Концептуальные основы будущих сетей // REDS: Телекоммуникационные устройства и системы. 2022. Т. 12. № 2. С. 24-28.

9. A. Marochkina, A. Paramonov, T. M. Tatarnikova. Ultra-Dense Internet of Things Model Network // Communications in Computer and Information Science. – 2022. – Vol. 1552. – P. 111-122. – DOI 10.1007/978-3-030-97110-6_8.
10. You, X., Wang, C.-X., Huang, J., (...), Fettweis, G.P., Liang, Y.-C. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts // Science China Information Sciences, 2021, 64(1),110301.
11. ITU-T Tec. Spec. FG NET2030 (06/2020) Network 2030 - Terms and Definitions for Network 2030. Режим доступа: https://www.itu.int/en/ITU-T/focusgroups/net2030/Documents/Network_2030_Terms_and_Definitions.pdf (дата обращения: 11.12.2022).
12. Li, X., Samaka, M., Chan, H.A., (...), Guo, C., Jain, R. Network Slicing for 5G: Challenges and Opportunities // IEEE Internet Computing, 2017, 21(5),8039298, с. 20-27.
13. Rost, P., Mannweiler, C., Michalopoulos, D.S., (...), Aziz, D., Bakker, H. Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks, IEEE Communications Magazine, 2017, 55(5),7926920, с. 72-79.
14. Muhizi, S., Ateya, A.A., Muthanna, A., Kirichek, R., Koucheryavy, A. A novel slice-oriented network model // Communications in Computer and Information Science, 2018, 919, с. 421-431.
15. ETSI TS 123 501 V15.2.0 (06/2018) 5G; System Architecture for the 5G System (3GPP TS 23.501 version 15.2.0 Release 15). Режим доступа: http://www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.02.00_60/ts_123501v150200p.pdf (дата обращения: 11.12.2022).
16. ITU-T Rec. Y.3101 (01/2018) Requirements of the IMT-2020 Network. Режим доступа: http://www.itu.int/rec/T-REC-Y.3101-201801-I/en (дата обращения: 11.12.2022).
17. Tikhvinskiy, V.O., Bochechka, G. Prospects and QoS requirements in 5G networks // Journal of Telecommunications and Information Technology, 2015, 2015(1), с. 23-26.

18. Khan, S., Khan, S., Ali, Y., (...), Ullah, Z., Mumtaz, S. Highly Accurate and Reliable Wireless Network Slicing in 5th Generation Networks: A Hybrid Deep Learning Approach, Journal of Network and Systems Management, 2022, 30(2), 29.
19. ITU-T Rec. Y.3112 (12/2018) Framework for the Support of Network Slicing in the IMT-2020 Network. Режим доступа: http://www.itu.int/rec/T-REC-Y.3112-201812-I/en (дата обращения: 11.12.2022).
20. GSM Association Official Document NG.116 (05/2019) Generic Netw. Slice Template, Version 1.0. Режим доступа: http://www.gsma.com/newsroom/wp-Content/uploads//NG.116 (дата обращения: 11.12.2022).
21. Popovski P., Trillingsgaard K.F., Simeone O., Durisi G.: 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view. IEEE Access, vol. 6 (2018)
22. Alessandro Lieto, Ilaria Malanchini , Antonio Capone. Enabling Dynamic Resource Sharing for Slice Customization in 5G Networks // IEEE Globecom. 2018.
23. Yuskov, I.O., Stroganova, E.P. Analysis of neural network model design for telecommunication corporate network monitoring // 2019 Systems of Signal Synchronization, Generating and Processing in Telecommunications, SYNCHROINFO, 2019, 8814111.
24. Shorov, A. 5G testbed development for network slicing evaluation // Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus, 2019, 8656861, c. 39-44.
25. Foukas, X., Patounas, G., Elmokashfi, A., Marina, M.K. Network Slicing in 5G: Survey and Challenges // IEEE Communications Magazine, 2017, 55(5),7926923, c. 94-100.
26. Sina Khatibi. Radio Resource Management Strategies in Virtual Networks: PhD degree in Electrical and Computer Engineering.-Лиссабон, 2016 // Режим доступа: https://sinakhatibi.com/wp-

content/uploads/2016/09/Thesis_sina_khatibi_IST1723601.pdf (дата обращения: 11.12.2022).

27. Zhirnov, N.S., Lyakhov, A.I., Khorov, E.M. Mathematical Model of a Network Slicing Approach for Video and Web Traffic // Journal of Communications Technology and Electronics, 2019, 64(8), с. 890-899.
28. Afolabi, I., Taleb, T., Samdanis, K., Ksentini, A., Flinck, H. Network slicing and softwarization: A survey on principles, enabling technologies, and solutions // IEEE Communications Surveys and Tutorials, 2018, 20(3),8320765, с. 2429-2453.
29. Ordonez-Lucena, J., Ameigeiras, P., Lopez, D., (...), Lorca, J., Folgueira, J. Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges // IEEE Communications Magazine, 2017, 55(5),7926921, с. 80-87.
30. Bega, D.; Gramaglia, M.; Banchs, A.; Sciancalepore, V.; Costa-Pérez, X. A Machine Learning Approach to 5G Infrastructure Market Optimization // IEEE Trans. Mob. Comput. 2019, 19, 498–512, doi:10.1109/TMC.2019.2896950.
31. Stepanov, M.S., Stepanov, S.N., Andrabi, U., Petrov, D., Ndayikunda, J. The Increasing of Resource Sharing Efficiency in Network Slicing Implementation // Communications in Computer and Information Science, 2022, 1552 CCIS, с. 18-35.
32. Muzata, A.R., Pershina, V.A., Stepanov, M.S., Ndimumahoro, F., Ndayikunda, J. The Modeling of Elastic Traffic Transmisson by the Mobile Network with NB-IoT Functionality // 2021 Systems of Signals Generating and Processing in the Field of on Board Communications, Conference Proceedings, 2021, 9416132.
33. Andrabi, U.M., Stepanov, S.N., Stepanov, M.S., Kanishcheva, M.G., Habinshuti, F.X. The Model of Conjoint Servicing of Real Time and Elastic Traffic Streams Through Processor Sharing (PS) Discipline with Access Control // International Conference Engineering and Telecommunication, En and T, 2021.
34. Perepelkin, D., Tsyganov, I. Network slicing algorithm with quality of services in software defined networks // 13th International Conference ELEKTRO 2020, ELEKTRO 2020 – Proceedings, 2020, 2020-May,9130345.

35. Campolo, C., Molinaro, A., Iera, A., Menichella, F. 5G network slicing for vehicle-to-everything services // IEEE Wireless Communications, 2017, 24(6), с. 38-45.
36. Dudin A. N., Klimenok V. I., Vishnevsky V. M. The theory of queuing systems with correlated flows //Cham : Springer International Publishing, 2019. – 410 p. – ISBN 978-3-030-32072-0. – DOI 10.1007/978-3-030-32072-0.
37. Назаров А. А., Моисеева С. П. Метод асимптотического анализа в теории массового обслуживания: монография //Томск : Изд-во Науч.-технической лит., 2006. – ISBN 5-89503-299-0.
38. Сегайер А., Цитович И.И. Построение моделей мультисервисных сетей // Электросвязь. 2009. № 9. С. 54-57.
39. Шнурков П. В., Горшенин А. К., Белоусов В. В. Аналитическое решение задачи оптимального управления полумарковским процессом с конечным множеством состояний / // Информатика и ее применения. – 2016. – Т. 10, № 4. – С. 72-88. – DOI 10.14357/19922264160408.
40. Самуйлов К.Е., Сопин Э.С., Шоргин С.Я. Система массового обслуживания с ограниченными ресурсами и сигналами для анализа показателей эффективности беспроводных сетей // Информатика и ее применения. 2017. Т. 11. № 3. С. 99-105.
41. Ageev K., Sopin E., Chursin A., Shorgin S. The probabilistic measures approximation of a resource queuing system with signals // Lecture Notes in Computer Science. 2021. Т. 13144 LNCS. С. 80-91.
42. Naumov V., Samouylov K., Yarkina N., Sopin E., Andreev S., Samuylov A. LTE performance analysis using queuing systems with finite resources and random requirements // В сборнике: International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. 7. Сер. "2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2015" 2016. С. 100-103.
43. Moiseev, A., Shklennik, M., Polin, E. Infinite-server queueing tandem with Markovian arrival process and service depending on its state // Annals of Operations Research, 326(1), с. 261-279. – DOI 10.1007/s10479-023-05318-1.

44. A. Alzaghir, A. Paramonov, A. Koucheryav. Estimation of Quality of Service in Tactile Internet, Augmented Reality and Internet of Things // Lecture Notes in Computer Science. – 2022. – Vol. 13158 LNCS. – P. 35-45. – DOI 10.1007/978-3-030-97777-1_4.
45. O. A. Mahmood, A. Khakimov, A. Muthanna, A. Paramonov. Effect of Heterogeneous Traffic on Quality of Service in 5G Network // Lecture Notes in Computer Science. – 2019. – Vol. 11965 LNCS. – P. 469-478. – DOI 10.1007/978-3-030-36614-8_36.
46. Borisov, A., Gorshenin, A. Identification of Continuous-Discrete Hidden Markov Models with Multiplicative Observation Noise //Mathematics, 2020, 10 (17), art. no. 3062.
47. Gorshenin, A.K., Belousov, V.V., Shnourkoff, P.V., Ivanov, A.V. Numerical research of the optimal control problem in the semi-Markov inventory model // AIP Conference Proceedings, 2015, 1648, art. no. 250007.
48. Горцев А.М., Назаров А.А., Терпугов А.Ф. Управление и адаптация в системах массового обслуживания // Томск: Национальный исследовательский Томский государственный университет, 1978. – 208 с.
49. Lieto, A.; Malanchini, I.; Capone, A. Enabling Dynamic Resource Sharing for Slice Customization in 5G Networks // GLOCOM 2018, doi:10.1109/GLOCOM.2018.8647249.
50. Vo, P.L.; Nguyen, M.N.H.; Le, T.A.; Tran, N.H. Slicing the Edge: Resource Allocation for RAN Network Slicing // IEEE Wirel. Commun. Lett. 2018, 7, 970–973, doi:10.1109/LWC.2018.2842189.
51. Sun, Y.; Qin, S.; Feng, G.; Zhang L.; Imran, M. Service Provisioning Framework for RAN Slicing: User Admissibility, Slice Association and Bandwidth Allocation // IEEE Trans. Mob. Comput. 2020, doi:10.1109/TMC.2020.3000657.
52. Zhao, G.; Qin, S.; Feng, G.; Sun, Y. Network Slice Selection in Softwarization-Based Mobile Networks // Tran. Emerg. Telecommun. Technol. 2020, 31, doi:10.1002/ett.3617.

53. Andrabi, U.M., Stepanov, S.N. The model of conjoint servicing of real time traffic of surveillance cameras and elastic traffic devices with access control // 2nd International Informatics and Software Engineering Conference, IISEC 2021, 2021.

54. Khatibi, S.; Correia, L.M. Modelling virtual radio resource management in full heterogeneous networks // EURASIP J. Wirel. Commun. Netw. 2017, 73, doi:10.1186/s13638-017-0858-7.

55. Sciancalepore, V., Samdanis, K., Costa-Perez, X., (...), Gramaglia, M., Banchs, A. Mobile traffic forecasting for maximizing 5G network slicing resource utilization // Proceedings - IEEE INFOCOM, 2017, 8057230.

56. Koucheryavy, Y., Lisovskaya, E., Moltchanov, D., Kovalchukov, R., Samuylov, A. Quantifying the millimeter wave new radio base stations density for network slicing with prescribed SLAs // Computer Communications, 2021, 174, c. 13-27.

57. Marabissi, D.; Fantacci, R. Highly Flexible RAN Slicing Approach to ManageIsolation, Priority, Efficiency // IEEE Access 2019, 7, doi:10.1109/ACCESS.2019.2929732.

58. Lee, Y. L.; Loo, J.; Chuah, T.;Wang, L.-C. Dynamic Network Slicing for Multitenant Heterogeneous Cloud Radio Access Networks // IEEE Trans. Wirel. Commun. 2018, doi:10.1109/TWC.2017.2789294.

59. Ozgur Umut Akgul, Ilaria Malanchini, and Antonio Capone. Dynamic Resource Trading in Sliced Mobile Networks // IEEE Trans. on Network and Service Managemen. 2019.

60. Gerasimov, A., Antonenko, V. Slicenomics: How to Provide Cost-Based Intra and Inter Slice Resource Management? // Proceedings - International Conference on Computer Communications and Networks, ICCCN, 2020, 2020-August,9209692.

61. Song, F., Li, J., Ma, C., (...), Shi, L., Jayakody, D.N.K. Dynamic virtual resource allocation for 5g and beyond network slicing // IEEE Open Journal of Vehicular Technology, 2020, 1,2990072, c. 215-226.

62. Tun, Y.K.; Tran, N.H.; Ngo, D.T.; Pandey, S.R.; Han, Z.; Hong, C.S.Wireless Network Slicing: Generalized Kelly Mechanism Based Resource Allocation // IEEE J. Select. Areas in Commun. 2019, 37, 1794–1807.

63. Caballero, P.; Banchs, A.; de Veciana, G.; Costa-Pérez, X. Networkslicing games: Enabling customization in multi-tenant networks // In Proceedings of the IEEE Conference on Computer Communications, Atlanta, GA, USA, 1–4 May 2017; doi:10.1109/INFOCOM.2017.8057046.

64. Caballero, P.; Banchs, A.; de Veciana, G.; Costa-Pérez, X.; Azcorra, A. Network Slicing for Guaranteed Rate Services: Admission Control and Resource Allocation Games // IEEE Trans. Wirel. Commun. 2018, 17, 6419–6432, doi:10.1109/TWC.2018.2859918.

65. Leconte M., Paschos G.S., Mertikopoulos P., Kozat U.C.: A Resource Allocation Framework for Network Slicing // In: IEEE INFOCOM, vol. 2018-April, 2177–2185 (2018).

66. Ksentini, A.; Nikaein, N. Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction // IEEE Commun. Mag. 2017, 55, 102–108, doi:10.1109/MCOM.2017.1601119.

67. Kokku, R.; Mahindra, R.; Zhang, H.; Rangarajan, S. CellSlice: Cellular wireless resource slicing for active RAN sharing // In Proceedings of the 5th International Conference on Communication Systems and Networks, COMSNETS, Bangalore, India, 7–10 January 2013; doi:10.1109/COMSNETS.2013.6465548.

68. Moskaleva, F., Lisovskaya, E., Lapshenkova, L., Shorgin, S., Gaidamaka, Y. Example of Degrading Network Slicing System in Two-Service Retrial Queueing System // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021, 13144 LNCS, c. 279-293.

69. Ravi R. Mazumdar, Christos Douligeris, L.G. Mason. Fairness in Network Optimal Flow Control: Optimality of Product Forms // IEEE Transactions on communications. 1991.

70. Laria Malanchini, Stefan Valentin , Osman Aydin. Wireless resource sharing for multiple operators: Generalization, fairness, and the value of prediction // Computer Networks. 2016.
71. Vlasenko, L., Kulik, V., Kirichek, R., Koucheryavy, A. Development of Models and Methods for Using Heterogeneous Gateways in 5G/IMT-2020 Network Infrastructure // Communications in Computer and Information Science, 2019, 1141 CCIS, c. 636-645.
72. Vassilakis V.G., Moscholios I.D., and Logothetis M.D. Call-level performance modelling of elastic and adaptive service-classes with finite population // IEICE Transactions. – 2008. – Vol. 91-B, No. 1. – P. 151–163.
73. Benameur N., Fredj S.B., Oueslati-Boulahia S., and Roberts J.W. Quality of service and flow level admission control in the Internet // Computer Networks. – 2002. – Vol. 40, No. 1. – P. 57–71.
74. Altman E., Artiges D., and Traore K. On the integration of best-effort and guaranteed performance services // INRIA Rapport de recherche No. 3222. – INRIA. – 1997. – 25 p.
75. Borst S. and Hegde N. Integration of streaming and elastic traffic in wireless networks // Proc. of the 26-th IEEE International Conference on Computer Communications INFOCOM-2007 (May 6–12, 2007, Anchorage, Alaska, USA). – IEEE. – 2007. – P. 1884–1892.
76. Litjens R. and Boucherie R.J. Elastic calls in an integrated services network: the greater the call size variability the better the QoS // Performance Evaluation. – 2003. – Vol. 52, No. 4. – P. 193–220.
77. Fodor G. and Skillermark P. Performance analysis of a reuse partitioning technique for multi-channel cellular systems supporting elastic services // International Journal of Communication Systems. – 2009. –Vol. 22, No. 3. – P. 307–342.
78. 3GPP TS 23.203 V17.2.0 (12/2021) 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and charging control architecture (Release 17) // Режим доступа:

https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=810 (дата обращения: 11.12.2022).

79. Власкина А.С., Поляков Н.А., Самуйлов К.Е., Гудкова И.А. Имитационная модель для анализа качества обслуживания пользователей виртуального мобильного оператора услуг с минимальной скоростью передачи данных // XIII Всероссийское совещание по проблемам управления ВСПУ-2019 : Сборник трудов XIII Всероссийского совещания по проблемам управления ВСПУ-2019, Москва, 17–20 июня 2019 года / Институт проблем управления им. В.А. Трапезникова РАН. – Москва: Институт проблем управления им. В.А. Трапезникова РАН, 2019. – С. 3046-3051. – DOI 10.25728/vspu.2019.3046. – EDN QCFAIZ.
80. Salvat, J.X.; Zanzi, L.; Garcia-Saavedra, A.; Sciancalepore, V.; Costa-Perez, X. Overbooking Network Slices through Yield-Driven End-to-End Orchestration // In Proceedings of the 14th International Conference on emerging Networking EXperiments and Technologies (CoNEXT ’18), New York, NY, USA, 4–7 December 2018; pp. 353–365; doi:10.1145/3281411.3281435.
81. D’Oro, S.; Restuccia, F.; Melodia, T.; Palazzo, S. Low-Complexity Distributed Radio Access Network Slicing: Algorithms and Experimental Results // IEEE/ACM Trans. Netw. 2018, doi:10.1109/TNET.2018.2878965.
82. Bega, D.; Gramaglia, M.; Banchs, A.; Sciancalepore, V.; Costa-Pérez, X. A Machine Learning Approach to 5G Infrastructure Market Optimization // IEEE Trans. Mob. Comput. 2019, 19, 498–512, doi:10.1109/TMC.2019.2896950.
83. Han B., Sciancalepore V., Feng D., Costa-Perez X., Schotten H.D.: A Utility-Driven Multi-Queue Admission Control Solution for Network Slicing // In: IEEE INFOCOM, vol. 2019-April, 55–63 (2019).
84. Vincenzi, M.; Lopez-Aguilera, E.; Garcia-Villegas, E. Maximizing Infrastructure Providers’ Revenue Through Network Slicing in 5G // IEEE Access 2019, 7, doi:10.1109/ACCESS.2019.2939935.
85. Семенова О.В., Власкина А.С., Гудкова И.А., Зарипова Э.Р. Расчёт вероятностно-временных характеристик установления соединения по

радиоканалу случайного доступа (имитационная модель) // Свидетельство о регистрации программы для ЭВМ RU 2018617268, 21.06.2018. Заявка № 2018614349 от 28.04.2018.

86. Семенова О.В., Власкина А.С., Медведева Е.Г., Зарипова Э.Р., Гудкова И.А., Гайдамака Ю.В. Расчёт вероятностно-временных характеристик установления соединения по радиоканалу случайного доступа (аналитическая модель) // Свидетельство о регистрации программы для ЭВМ RU 2018617042, 14.06.2018. Заявка № 2018614315 от 27.04.2018.
87. Medvedeva E., Zaripova E., Semenova O., Vlaskina A., Gudkova I., Gaidamaka Y. Discrete time Markov chain model for analyzing characteristics of RACH procedure under massive machine type communications // В сборнике: ACM International Conference Proceeding Series. 2. Сер. "Proceedings of the 2nd International Conference on Future Networks and Distributed Systems, ICFNDS 2018" 2018. С. 59.
88. Семенова О.В., Власкина А.С., Медведева Е.Г., Зарипова Э.Р., Гудкова И.А. Процедура установления соединения по радиоканалу случайного доступа с возможностью ретрансляции // Вестник Российского университета дружбы народов. Серия: Математика, информатика, физика. 2018. Т. 26. № 3. С. 261-271.
89. Samouylov K.E., Gaidamaka Y.V., Gudkova I.A., Zaripova E.R., Shorgin S.Ya. Baseline Analytical Model for Machine-type Communications over 3GPP RACH in LTE-advanced Networks // 31st International Symposium on Computer and Information Sciences (ISCIS), October 27-28th, 2016, Krakow, Poland. T. Czachórski et al. (Eds.): ISCIS 2016, CCIS 659, pp. 203–213, 2016.
90. Borodakiy, V., Samouylov, K., Gaidamaka, Yu., Abaev, P., Buturlin, I., Etezov, Sh.: Modelling a Random Access Channel with Collisions for M2M Traffic in LTE Networks // In: Balandin, S. et al. (Eds.): NEW2AN/ruSMART 2014. Springer, Heidelberg. LNCS 8638, 301–310 (2014).
91. Vlaskina A., Semenova O., Gudkova I. Algorithm of file transmission between ENODEB and devices through D2D connection and mulicasting // В сборнике:

Applied Problems in Theory of Probabilities and Mathematical Statistics into Telecommunicationsa. Аранити Д., Самуйлов К.Е., Шоргин С.Я. Труды XI Международного семинара. Под редакцией Д. Аранити, К.Е. Самуйлова, С.Я. Шоргина. 2017. С. 10.

92. Власкина А.С., Семенова О.В., Гудкова И.А. Расчет задержки передачи данных по технологиям мультивещания и прямого взаимодействия устройств в беспроводной сети // Свидетельство о регистрации программы для ЭВМ RU 2018617130, 19.06.2018. Заявка № 2018614354 от 28.04.2018.
93. Basharin, G.P., Aterekova, T.V. Analytical model of streaming and elastic traffic with dynamic channel allocation scheme // 2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2010, 5676514, c. 1086-1090.
94. Башарин Г.П. Лекции по математической теории телетрафика. – М.:РУДН, 2009. – с.342 [Basharin, G.P. Lekcii po matematicheskoyj teorii teletrafika. – M.: RUDN, 2009. – c. 342].
95. Samouylov, K.E., Gudkova, I.A. Recursive computation for a multi-rate model with elastic traffic and minimum rate guarantees // 2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2010, 5676509, c. 1065-1072.
96. 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, 2013, 47(2), c. 62-69.
97. Zheng, J., De Veciana, G. Elastic Multi-resource Network Slicing: Can Protection Lead to Improved Performance? // Proceedings - 17th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, WiOpt 2019, 9144138.
98. Gomes, R.L., Bittencourt, L.F., Madeira, E.R.M. Reliability-Aware Network Slicing in Elastic Demand Scenarios // IEEE Communications Magazine, 2020, 58(10),9247519, c. 29-34.

99. Khan, H., Samarakoon, S., Bennis, M. Enhancing Video Streaming in Vehicular Networks via Resource Slicing // IEEE Transactions on Vehicular Technology, 2020, 69(4),9003407, с. 3513-3522.

100. Baena, C., Fortes, S., Baena, E., Barco, R. Estimation of video streaming KQIs for radio access negotiation in network slicing scenarios // IEEE Communications Letters, 2020, 24(6),9031300, с. 1304-1307.

101. 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, 2013, 47(4), с. 202-210.

102. Vlaskina A., Polyakov N., Gudkova I. Modeling and performance analysis of elastic traffic with minimum rate guarantee transmission under network slicing // В сборнике: Internet of Things, Smart Spaces, and Next Generation Networks and Systems. 2019. С. 621-634.

103. Григорьева Т.В., Поляков Н.А., Власкина А.С., Самуйлов К.Е. Вероятностная модель для анализа характеристик обслуживания эластичного трафика в беспроводной сети с нарезкой радиоресурсов // В сборнике: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. 2019. С. 131-136.

104. Поляков Н.А., Власкина А.С., Гудкова И.А., Самуйлов К.Е. Расчет вероятностно-временных характеристик обслуживания эластичного трафика с минимальной скоростью в сегменте беспроводной сети с нарезкой радиоресурсов (имитационная модель) // Свидетельство о регистрации программы для ЭВМ RU 2019664614, 11.11.2019. Заявка № 2019663728 от 01.11.2019.

105. Поляков Н.А., Власкина А.С., Гудкова И.А., Самуйлов К.Е. Расчет вероятностно-временных характеристик обслуживания эластичного трафика с минимальной скоростью в сегменте беспроводной сети с нарезкой радиоресурсов (математическая модель) // Свидетельство о регистрации

программы для ЭВМ RU 2019664613, 11.11.2019. Заявка № 2019663708 от 01.11.2019.

106. Vlaskina A.S., Polyakov N.A., Gudkova I.A., Gaidamaka Yu.V. Performance analysis of elastic traffic with minimum bit rate guarantee transmission in wireless network under network slicing // Izvestiya of Saratov University. New Series. Series: Mathematics. Mechanics. Informatics, 2020, 20(3), pp. 378–387.
107. Ageev, K., Garibyan, A., Golskaya, A., Gaidamaka, Yu., Sopin, E., Samouylov, K., Correia, L.M. Modelling of Virtual Radio Resources Slicing in 5G Networks // Communications in Computer and Information Science, 2019, 1109, с. 150-161.
108. Markova, E., Adou, Y., Ivanova, D., Golskaia, A., Samouylov, K. Queue with Retrial Group for Modeling Best Effort Traffic with Minimum Bit Rate Guarantee Transmission Under Network Slicing // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, 11965 LNCS, с. 432-442.
109. Adou, K.Y., Markova, E.V. Methods for Analyzing Slicing Technology in 5G Wireless Network Described as Queueing System with Unlimited Buffer and Retrial Group // Communications in Computer and Information Science, 2021, 1391 CCIS, с. 264-278.
110. Adou, Y., Markova, E., Chursin, A.A. Analysis of Non-preemptive Scheduling for 5G Network Model Within Slicing Framework // Communications in Computer and Information Science, 2022, 1552 CCIS, с. 36-47.
111. Adou, Y., Markova, E., Gaidamaka, Y. Modeling and Analyzing Preemption-Based Service Prioritization in 5G Networks Slicing Framework // Future Internet, 2022, 14(10),299.
112. Yarkina, N., Correia, L.M., Moltchanov, D., Gaidamaka, Y., Samouylov, K. Multi-tenant resource sharing with equitable-priority-based performance isolation of slices for 5G cellular systems // Computer Communications, 2022, 188, с. 39-51.
113. Ageev, K., Sopin, E., Samouylov, K. Resource Sharing Model with Minimum Allocation for the Performance Analysis of Network Slicing // Communications in Computer and Information Science, 2021, 1391 CCIS, с. 378-389.

114. Яшков С.Ф. The M/D/1 processor sharing queue revisited // Информационные процессы. 2009. Т. 9. № 3. С. 216-223.
115. Polyakov, N., Yarkina, N., Samouylov, K., Koucheryavy, Y. Network Slice Degradation Probability as a Metric for Defining Slice Performance Isolation // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2022, 13158 LNCS, с. 481-492.
116. Bobrikova, E.V., Platonova, A.A., Shorgin, S.Ya., Gaidamaka, Y.V. To the analysis of the dynamic assignment of radio resources in wireless networks with a network slicing mechanism // CEUR Workshop Proceedings, 2020, 2639, с. 83-92.
117. Moskaleva, F.A., Gaidamaka, Yu.V., Shorgin, V.S. Impact of the isolation parameters on resource allocation in the network slicing model // Informatika i ee Primeneniya, 2020, 14(4), с. 9-16.
118. Yarkina, N., Gaidamaka, Y., Correia, L.M., Samouylov, K. An analytical model for 5G network resource sharing with flexible SLA-oriented slice isolation // Mathematics, 2020, 8(7),1177.
119. Moskaleva, F., Lisovskaya, E., Gaidamaka, Y. Resource Queueing System for Analysis of Network Slicing Performance with QoS-Based Isolation // Communications in Computer and Information Science, 2021 1391 CCIS, с. 198-211.
120. Ageev, K.A., Sopin, E.S., Yarkina, N.V., Samouylov, K.E., Shorgin, S.Ya. Analysis of the network slicing mechanisms with guaranteed allocated resources for various traffic types // Informatika i ee Primeneniya, 2020, 14(3), с. 94-100.
121. Савич В.Н., Дымова П.И., Поляков Н.А., Власкина А.С., Гудкова И.А. К анализу системы массового обслуживания с двумя очередями и нетерпеливым эластичным трафиком с минимальной скоростью // В сборнике: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. 2019. С. 103-107.

122. Зубова И.О., Власкина А.С., Кочеткова И.А. Сравнительный анализ схем нарезки радиоресурсов в беспроводной сети с двумя типами услуг // В сборнике: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. Москва, 2020. С. 30-33.

123. Дымова П.И., Савич В.Н., Поляков Н.А., Власкина А.С. Вероятностная модель нарезки ресурсов беспроводной сети между двумя виртуальными операторами // Математическое и программное обеспечение информационных, технических и экономических систем : Материалы VII Международной молодежной научной конференции, Томск, 23–25 мая 2019 года / Под общей редакцией И.С. Шмырина. – Томск: Издательский Дом Томского государственного университета, 2019. – С. 232-236. – EDN ATNDWM.

124. Tusa, F., Clayman, S. End-to-end slices to orchestrate resources and services in the cloud-to-edge continuum // Future Generation Computer Systems, 2023, 141, с. 473-488.

125. Рыков В.В. Управляемые системы массового обслуживания // В кн.: Теория вероятностей. Математическая статистика. Теоретическая кибернетика. Итоги науки и техники. ВИНИТИАН СССР, 1975.

126. M. Yu. Kitaev, V. V. Rykov. Controlled queueing systems // N. Y.: CRC Press, 1995, 304 p.

127. Рыков В.В. Управляемые марковские процессы с конечными пространствами состояний и управлений // Теория вероятностей и ее применения. 1966. Т. 11. № 2. С. 343.

128. Ефросинин, Д. В. Методы анализа управляемых динамических систем: специальность 05.13.01 "Системный анализ, управление и обработка информации (по отраслям)" // Диссертация на соискание ученой степени доктора физико-математических наук, Москва, 2013 – 332 с.

129. Ефросинин, Д. В. Управляемые системы массового обслуживания с неоднородными приборами : специальность 05.13.17 "Теоретические основы информатики" : диссертация на соискание ученой степени кандидата физико-математических наук – Москва, 2005. – 247 с.
130. Howard R.A. Dynamic programming and markov processes // Динамическое программирование и марковские процессы. Перевод с английского В.В. Рыкова под редакцией Н.Я. Бусленко издательство Советское радио: Москва — 1964.
131. Семенова О.В., Дудин А.Н. Система массового обслуживания M|M|N с управляемым режимом обслуживания и катастрофическими сбоями // Автоматика и вычислительная техника. 2007. № 6. С. 72-80.
132. Семенова О.В. Многопороговое управление системой массового обслуживания Bmap/G/1 с MAP-потоком катастрофических сбоев // Автоматика и телемеханика. 2007. № 1. С. 105-120.
133. Efrosinin D., Kochetkova I., Samouylov K., Stepanova N. Algorithmic analysis of a two-class multi-server heterogeneous queueing system with a controllable cross-connectivity // Lecture Notes in Computer Science. 2020. Т. 12023 LNCS. С. 1-17.
134. Avrachenkov, K., Dudin, A., Klimenok, V., Nain, P., Semenova, O. Optimal threshold control by the robots of web search engines with obsolescence of documents // Computer Networks, 2011, 55 (8), pp. 1880-1893.
135. Wang, Y., Chen, W. Adaptive Power and Rate Control for Real-Time Status Updating over Fading Channels // IEEE Transactions on Wireless Communications, 2021, 20(5),9316992, с. 3095-3106.
136. Mandel, A.S., Laptin, V.A. Channel Switching Threshold Strategies for Multichannel Controllable Queuing Systems // Communications in Computer and Information Science, 2020, 1337, с. 259-270.
137. Mandel, A., Laptin, V. Myopic channel switching strategies for stationary mode: Threshold calculation algorithms // Communications in Computer and Information Science, 2018, 919, с. 410-420.

138. Shorgin, S., Samouylov, K., Gudkova, I., Galinina, O., Andreev, S. On the benefits of 5G wireless technology for future mobile cloud computing // SDN and NFV: Next Generation of Computational Infrastructure - 2014 International Science and Technology Conference - Modern Networking Technologies, MoNeTec 2014, Proceedings, 2014, 6995601.
139. Бурцева С.А., Хакимов А.А., Григорьева Т.В., Власкина А.С., Кочеткова И.А. Имитационная модель управляемого занятия ресурсов системы облачных вычислений из двух групп виртуальных машин // В сборнике: Математическое и программное обеспечение информационных, технических и экономических систем. Материалы Международной научной конференции. Сер. "физико-математическая" Томск, 2020. С. 249-254.
140. Kochetkova I.A., Vlaskina A.S., Efrosinin D.V., Khakimov A.A., Burtseva S.A. To analysis of a two-buffer queuing system with cross-type service and additional penalties // Discrete and Continuous Models and Applied Computational Science. 2021. Т. 29. № 2. С. 158-172.
141. Kochetkova I.A., Vlaskina, A.S., Vu, N.N., Shorgin, V.S. Queuing system with signals for dynamic resource allocation for analyzing network slicing in 5G networks // Informatika i ee Primeneniya, 2021, 15(3), pp. 91–97.
142. Ву Н.Н., Бурцева С.А., Власкина А.С. Вероятностная модель для анализа влияния интервала нарезки радиоресурсов на показатели качества обслуживания эластичного трафика // В сборнике: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. материалы Всероссийской конференции с международным участием. Российский университет дружбы народов. Москва, 2021. С. 45-50.
143. Филиппова В.С., Хусайнова Ф.Д., Бурцева С.А., Власкина А.С. Анализ показателей эффективности нарезки радиоресурсов в сети 5G в виде системы массового обслуживания с сигналами // В сборнике: Информационно-телекоммуникационные технологии и математическое моделирование

высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. Москва, 2022. С. 61-65.

144. Филиппова В.С., Хусайнова Ф.Д., Власкина А.С., Кочеткова И.А., Бурцева С.А. Расчет показателей эффективности модели управления нарезкой радиоресурсов беспроводной сети между двумя виртуальными операторами по сигналам контроллера // Свидетельство о регистрации программы для ЭВМ 2022660727, 08.06.2022. Заявка № 2022619563 от 27.05.2022.
145. By Н.Н., Власкина А.С., Кочеткова И.А., Бурцева С.А. Расчет показателей эффективности модели динамической нарезки между двумя сегментами радиоресурсов с управлением по внешнему событию // Свидетельство о регистрации программы для ЭВМ 2021661716, 14.07.2021. Заявка № 2021660847 от 08.07.2021.
146. Vlaskina A.S., Burtseva S.A., Kochetkova I.A., Shorgin S.Ya. Controllable queuing system with elastic traffic and signals for analyzing network slicing // Informatika i ee Primeneniya, 2022, 16(3), pp. 90–96.
147. Леонтьева К.А., Гебриал И.Е.З., Бурдина К.П., Бурцева С.А., Власкина А.С. К анализу политики перераспределения ресурса в управляемой системе массового обслуживания для нарезки сети 5G // В сборнике: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. Москва, 2023. С. 82-86.
148. Власкина А.С. Управляемая система массового обслуживания для анализа динамической нарезки радиоресурсов в сети 5G // В сборнике: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. Москва, 2022. С. 34-38.
149. Kochetkova I., Vlaskina A., Burtseva S., Savich V., Hosek J. Analyzing the effectiveness of dynamic network slicing procedure in 5G network by queuing and simulation models // Lecture Notes in Computer Science. 2020. Т. 12525 LNCS. С. 71-85.