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

На правах рукописи

Сулейманов Алмаз Авхатович

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МЕТОДА ОЦЕНКИ КАЧЕСТВА ИНФОКОММУНИКАЦИОННОЙ ОБЛАЧНОЙ УСЛУГИ «ВИРТУАЛЬНЫЙ РАБОЧИЙ СТОЛ»

Специальность 05.12.13 –
Системы, сети и устройства телекоммуникаций

Диссертация
на соискание учетной степени кандидата технических наук

Научный руководитель:
доктор технических наук
Нетес Виктор Александрович

Москва – 2017

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность темы исследования. На сегодняшний день облачные инфокоммуникационные услуги, предоставляемые пользователям по требованию с удаленных серверов через сеть передачи данных, получают активное развитие. Примером таковых является услуга «виртуальный рабочий стол», основная идея которой заключается в предоставлении пользователю полноценного рабочего места «из облака» на любое устройство, имеющее соответствующий программный агент. Рабочее место разворачивается внутри виртуальной машины на сервере посредством специального программного обеспечения. Доступ к нему предоставляется по сети передачи данных и осуществляется при помощи протокола доставки виртуального рабочего стола.

Изначально эта услуга проектировалась с учетом размещения сервера услуги и клиентского устройства в одной локальной сети. Однако, в условиях нынешнего роста популярности облачной парадигмы размещение данных пользователя не только на локальных серверах, но и на серверах в удаленных центрах обработки данных (ЦОД) породило ряд технических вопросов, связанных с обеспечением требуемого качества для конечных пользователей, которые находятся уже не в локальной сети с сервером, а в глобальных сетях. Эта задача является ключевой для провайдера услуг и ее сложность возрастает с переходом от локальной к глобальной сети. Для решения этой задачи необходимы математические модели, позволяющие прогнозировать состояние качества услуги.

Данная тема недостаточно изучена. В отечественных и зарубежных источниках, не описаны методы оценки и обеспечения качества облачной инфокоммуникационной услуги «виртуальный рабочий стол», которые были бы применяться как на этапе проектирования, так и на этапе эксплуатации сети. Поэтому данная тема является весьма актуальной.

Степень разработанности. Заметный вклад в исследование вопросов качества инфокоммуникационных услуг внесли отечественные и зарубежные ученые Пшеничников А.П., Ефимушкин В.А., Нетес В.А, Лихтциндер Б.Я., Shneiderman B., Emmerich W. и другие.

Вопросам качества облачных услуг посвящены работы Tolia M., Dusi M., Lin K.J, Nieh J. и других. В работах Tolia M. и Dusi M. рассматривается влияние сетевых задержек на работу услуги для одной из устаревших платформ, в работах Lin K.J и Nieh J. исследованы отдельные аспекты работы услуги при различных сетевых условиях (задержки и полоса канала). Полученные результаты оказались недостаточны в современных условиях развития инфраструктуры услуги «виртуальный рабочий стол», поэтому диссертационное исследование продолжилось в направлении всестороннего анализа услуги, рассмотрения современных

облачных платформ, развития новых подходов, учитывающих параметры услуги, определяющих приемлемость ее качества.

Целью диссертационной работы является разработка и исследование метода оценки качества инфокоммуникационной облачной услуги «виртуальный рабочий стол».

Научная новизна исследования состоит в следующем:

1. На основании анализа логики услуги «виртуальный рабочий стол» для разработки математических моделей выделены две фазы ее предоставления для возможности отдельного их исследования. В первой фазе рассмотрено подключение пользователей; во второй фазе предусмотрена их работа с индивидуальными рабочими столами.

2. Для первой фазы разработана аналитическая модель, позволяющая оценить среднее время отклика; получены его зависимости от основных характеристик системы (среднего времени обслуживания одного запроса, числа одновременно обслуживаемых пользователей). Решена задача определения множества допустимых значений характеристик сервера, при которых выполняются ограничения по среднему времени отклика и вероятности отказа в подключении, а также задача определения рациональных вариантов сочетания этих параметров.

3. Для второй фазы разработаны аналитические модели, которые для трех сценариев предоставления услуги позволяют оценить среднее время отклика, а также получить аналитические соотношения между средним временем отклика и интенсивностями обслуживания. Предложена обобщенная модель базового сценария предоставления услуги, которая позволяет оценить среднее время отклика для различных типов потоков и законов распределения времени обслуживания.

4. В результате проведенных натурных экспериментальных исследований получены оценки характеристик инфраструктуры услуги, влияющих на ее качество: среднего времени между запросами к серверу в обеих фазах, среднего времени обработки на пользовательском устройстве, среднего времени обслуживания запросов сервером, среднего время отклика, а также зависимости транспортной задержки от скорости передачи данных.

Теоретическая и практическая значимость диссертации. Разработанный метод оценки качества инфокоммуникационной облачной услуги «виртуальный рабочий стол» может быть использован операторами связи и поставщиками услуг на этапе проектирования и эксплуатации инфраструктуры услуги для ее мониторинга с целью обеспечения требуемого качества. Полученные при помощи предложенных аналитических моделей оценки и рекомендации позволяют управлять уровнем качества услуги путем регулирования параметров ее инфраструктуры, а также учитывать влияние на качество, оказываемое сетью передачи данных.

Методы исследования. При выполнении диссертационной работы были применены методы теории вероятностей, математической статистики, теории массового обслуживания.

Основные положения, выносимые на защиту.

1. Для анализа качества услуги целесообразно для возможности исследования математической модели выделить две фазы, во второй фазе – три сценария предоставления услуги. Основные показатели качества для первой фазы: среднее время отклика, вероятность отказа в подключении; для второй – среднее время отклика.

2. Аналитическая модель первой фазы услуги «виртуальный рабочий стол» на основе СМО M/G/1/K*PS позволяет произвести оценку основных параметров качества и выбрать характеристики инфраструктуры, обеспечивающие требуемый его уровень.

3. Аналитические модели трех сценариев второй фазы облачной услуги «виртуальный рабочий стол», построенные на основе сети Джексона и BCMP-сетей, позволяют в зависимости от сценария предоставления услуги, оценить основные параметры качества и выбрать параметры инфраструктуры, обеспечивающие требуемый его уровень. Обобщенная модель базового сценария второй фазы работы услуги, учитывающая первые два момента случайных величин, описывающих времена между заявками и длительности обслуживания, позволяет оценить среднее время отклика для произвольных законов распределения этих величин.

Достоверность и апробация результатов работы. Полученные результаты обоснованы корректным применением математических методов с использованием теории массового обслуживания, теории вероятностей, математической статистики. Основное содержание диссертационной работы докладывалось и обсуждалось на российских и международных научных конференциях: «Фундаментальные проблемы радиоэлектронного приборостроения INTERMATIC» (Москва, 2013, 2014, 2016 гг.), «Телекоммуникационные и вычислительные системы» (Москва, 2014, 2015, 2016, 2017 гг.), «Перспективные технологии в средствах передачи информации – ПТСПИ» (Владимир-Суздаль, 2015 г.), «Технологии информационного общества» (Москва, 2015, 2016, 2017 гг.). По теме диссертации опубликовано 17 работ, в том числе 4 – в рецензируемых периодических изданиях, входящих в перечень ВАК при Минобрнауки РФ.

Структура и объем диссертации. Диссертационная работа состоит из введения, четырех разделов, заключения, списка литературы, приложения и содержит 130 страниц машинописного текста, 48 рисунков, 32 таблицы. Список литературы содержит 121 наименование.

Раздел 1. Облачные услуги и их качество

1.1 Основные преимущества, понятия и принципы работы облачных услуг

1.1.1 Развитие облачных услуг

Концепция использования вычислительного ресурса по принципу общего доступа восходит к 1960-м годам и встречается в работах Д. Маккарти [92] и Д. Ликлайдера [90]. Следующими этапами ее развития можно считать запуск системы управления взаимоотношениями с клиентами salesforce.com, предоставляемой пользователям по подписке на одноименном веб-сайте, а также начало предоставления доступа к услугам через сеть Интернет магазином Amazon.com в 2002 году [94]. Развитие сервисов Amazon привело к запуску в 2006 году проекта под названием Elastic Computing Cloud. Практически одновременно с этим событием термины «cloud» и «cloud computing» были упомянуты в одном из публичных выступлений Э. Шмидта – главы компании Google [51]. Этот момент можно считать началом упоминаний облачных вычислений в СМИ, публикациях специалистов, научной среде.

Облачные услуги играют существенную роль в развитии инфокоммуникационной среды в наше время. Облачная парадигма, фактически, стала основным направлением развития отрасли, обеспечивая имеющиеся и назревающие потребности пользователя.

В 2016 году российский рынок облачных услуг вырос на 43% с 15,79 до 22,6 млрд руб [11]. По оценке, приведенной в [23], рынок облачных услуг будет расти быстрее, чем ИТ-рынок в целом, и к 2020 году его объем составит 48 млрд руб. То есть при среднегодовом темпе в 21% его объем вырастет в 3 раза по сравнению с 2015 годом.

С момента своего появления модель предоставления услуг «из облака» глубоко проникла в различные информационно-технологические сферы и стала играть весомую роль среди сетевых услуг. Фактически, облачная среда – это удобная площадка для хранения и обработки информации, которая объединяет в себе аппаратные вычислительные ресурсы, средства хранения данных, программное обеспечение, каналы связи и др. Работа в такой среде позволяет снизить расходы и повысить эффективность работы инфраструктуры предприятия, учреждения, ведомства. Особенностями облачных технологий являются отсутствие жесткой географической привязанности к аппаратному обеспечению и масштабируемость (т.е. возможность мгновенного развертывания новых ресурсов). Пользователь может работать в такой среде из

любой точки планеты, с любого устройства. Кроме того, за счет синхронизации данных между различными пользовательскими устройствами, появляется возможность начать работу на одном из них, например, офисном компьютере, а продолжить на другом, например, на планшете или смартфоне во время командировки или совещания.

Перенос проектов в облачные среды позволяет компании сократить расходы на покупку серверов, сетевого оборудования, лицензий, сокращение штата специалистов, обслуживающих ИТ-инфраструктуру. С точки зрения администрирования открываются такие возможности, как централизованность данных, что значительно повышает удобство управления, контроля и учета, а также повышает безопасность за счет уменьшения точек несанкционированного входа пользователей, наличия централизованной технической поддержки, регулярного резервирования пользовательских данных. Немалым достоинством облачных сред является также гибкость в развертывании услуг, которая заключается в том, что администратор по требованию может в течении нескольких минут организовать подключение новых пользователей или расширить ресурсы существующим.

Облачные услуги нашли свое применение во многих сферах: в отраслях народного хозяйства, в бизнес-сегменте, государственном секторе, у обычных пользователей. В зависимости от сферы применения может меняться специфика услуг, которая зависит от деятельности, выполняемой пользователем.

Некоторые из облачных услуг давно заняли прочное место в инфокоммуникационной среде, другие можно отнести к развивающимся. Среди последних можно выделить услугу «виртуальный рабочий стол», которая благодаря особенностям реализации и специфики предоставления сочетает в себе свойства различных облачных услуг, является востребованной и обладает перспективами для дальнейшего развития и внедрения. Перспективы развития и темпы роста услуги «виртуальный рабочий стол» рассмотрены в [4], где, в частности, прогнозируется, что «типовое офисное рабочее место в 2020 году будет организовано в виртуальной среде».

На российском инфокоммуникационном рынке имеется много провайдеров облачных услуг. Ввиду их большого количества рынок является конкурентным, а в условиях конкурентного рынка неизбежно встает вопрос о качестве предоставляемой услуги. Поэтому тематика обеспечения качества облачных услуг является крайне важной и актуальной. Особенно это актуально для услуг, предоставляемых пользователям в реальном времени, в частности, услуги «виртуальный рабочий стол».

Начать исследование вопросов, связанных с обеспечением ее качества следует с рассмотрения определений, присущих облачным услугам в целом.

1.1.2 Определение понятия облачных вычислений и облачных услуг

Согласно определению, приведенному в технических бюллетенях МСЭ-Т и NIST [78, 93], облачные вычисления – это концепция, подразумевающая обеспечение повсеместного сетевого доступа по необходимости к общей группе конфигурируемых вычислительных ресурсов таким, как сети передачи данных, серверы, устройства хранения данных, приложения и сервисы, которые могут быть оперативно предоставлены и изъяты с минимальными эксплуатационными затратами или обращениями к провайдеру.

Согласно [79] облачная услуга может быть определена как услуга, которая поставляется и потребляется по требованию в любое время, через любую сеть доступа с использованием технологий облачных вычислений. Таким образом, под облачной услугой понимается предоставление доступа пользователю к удаленным серверам данных или приложений, организованное техническими средствами и закрепленное юридическими обязательствами (договор, оферта и т.д.).

Под облачной архитектурой понимается принципиальная организация облачной услуги, воплощенная в её элементах, их взаимоотношениях друг с другом и со средой, а также принципы, направляющие её проектирование и развитие. На рисунке 1.1 показаны функциональные уровни облачной архитектуры. Эти уровни разделены и сгруппированы в соответствии с их функциями согласно [68].

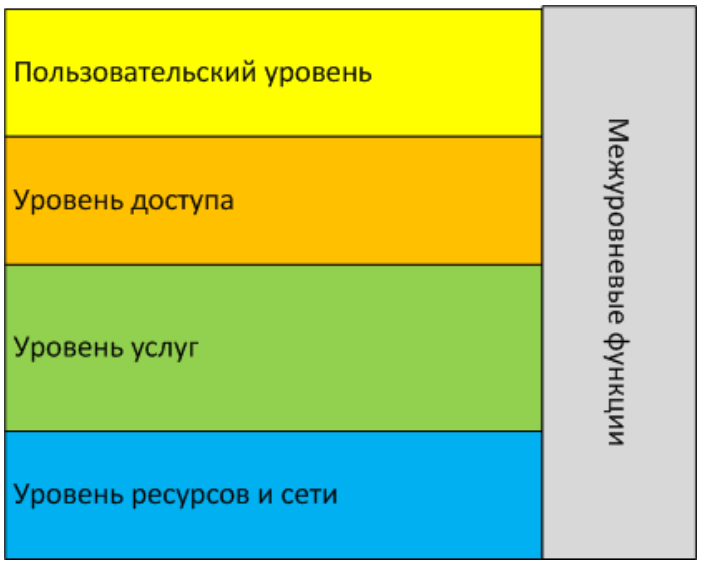

Рис. 1.1. Функциональные уровни облачной архитектуры

Рассмотрим их подробнее в отдельности.

- Пользовательский уровень. Отвечает за взаимодействие пользователя с облачной инфраструктурой, в частности, выполняет следующие функции: отправка запросов на авторизацию, получение и обработка ответов сервера, организация передачи информации от сервера к клиенту, мониторинг ресурсов и состояния сессии клиент-сервер.
- Уровень доступа. Обеспечивает общий интерфейс для функциональной среды облачных услуг посредством подключения пользователей к API облачной инфраструктуры. API (application programming interface, интерфейс программирования приложений) – набор готовых классов, процедур, функций и др., предоставляемых приложением или библиотекой для использования во внешних программных продуктах.
- Уровень услуг. Здесь провайдер размещает все, что необходимо для предоставления одной из категорий услуг (программное обеспечение, инфраструктура, средства хранения и обработки данных и т.д.), а также для ее доставки пользователю.
- Уровень ресурсов и сети. Здесь находится оборудование провайдера облачных услуг: серверы, сетевые коммутаторы и маршрутизаторы, средства хранения данных (СХД).
- Межуровневые функции. Обеспечивают мониторинг, администрирование, безопасность системы.

Таким образом, следует отметить, что архитектура облачных услуг в целом представлена в виде ряда функциональных компонентов, согласованное взаимодействие которых обеспечивает работоспособность услуги.

1.1.3 Классификация и модели развертывания облачных услуг

В Рекомендации Y.3500 МСЭ-Т [76] приведена классификация облачных услуг. Данная классификация построена на понятиях «возможностей» (capabilities) и «категорий» (categories). Возможности – это свойства облачных услуг, предоставляемых провайдером пользователю, в зависимости от назначения и типов предоставляемых ресурсов. Различают три типа возможностей, которые сформированы согласно принципу разделения полномочий:

- Приложения: пользователь получает доступ к приложениям, развернутым на оборудовании провайдера.
- Инфраструктура: пользователю предоставляется возможность обработки и хранения данных, размещенных на стороне провайдера, а также сетевые ресурсы.

- Платформа: пользователю предоставляется доступ к среде отладки, проектирования и запуска приложений, развернутой на оборудовании провайдера с использованием одного или нескольких языков программирования.

Категория облачных услуг – это группа облачных услуг, удовлетворяющих общим наборам свойств. Категория может включать в себя одну или несколько возможностей.

Так, выделяют 7 категорий облачных услуг, самыми распространенными из которых являются следующие три:

1. Инфраструктура как услуга (Infrastructure as a Service, IaaS): категория облачных услуг, в которой пользователю предоставляется возможность использования облачной инфраструктуры: сетевых ресурсов, средств хранения и обработки данных. Пользователю предоставляется возможность самостоятельной установки и запуска необходимого программного обеспечения.

2. Платформа как услуга (Platform as a Service, PaaS): категория облачных услуг, в которой пользователю предоставляется возможность использования облачной платформы, в состав которой входят средства создания, тестирования и выполнения прикладного программного обеспечения, различные среды исполнения языков программирования, которые предоставляются провайдером услуги.

3. Программное обеспечение как услуга (Software as a Service, SaaS): категория облачных услуг, в которой пользователю предоставляется возможность использования приложений, развернутых на сервере провайдера в облаке. Пользователь выбирает необходимые программы из перечня предлагаемых провайдером. Наибольшее распространение в рамках данной категории получили различные офисные и бухгалтерские приложения, почтовые сервисы, антивирусное программное обеспечение и др.

Данные категории услуг являются относительно простыми, проверенными на практике, каждой категории тут соответствует одна возможность. Однако, в [76] приведены, кроме того, еще несколько категорий, относимых к числу «развивающихся». В частности, это услуга типа «виртуальный рабочий стол» (англ. Desktop as a Service, DaaS). Более подробно она описана в Рекомендации Y.3503 МСЭ-Т [77]. Эта услуга представляет большой интерес для исследования: очевидно, что она является более сложной с точки зрения технической реализации, поскольку сочетает в себе сразу несколько возможностей и свойств различных категорий услуг. Для ее предоставления требуется организовать на сервере услуги более сложную инфраструктуру по сравнению с вышеперечисленными категориями облачных услуг. Данная услуга будет рассматриваться в диссертационной работе.

Модели развертывания облачных услуг описывают, как облачные вычисления могут быть организованы на основе управления и обмена физическими или виртуальными ресурсами. Различают следующие модели:

- Публичное облако (Public Cloud): модель развертывания облака, где облачные сервисы являются потенциально доступными любому пользователю облака, ресурсы находятся под контролем поставщика облачных услуг. Публичное облако может быть развернуто и управляемо в бизнес структуре, академическом учреждении или государственной организации, или в каком-либо их сочетании. Оно размещается на территории провайдера облачных услуг. Реальная доступность таких услуг для конкретных клиентов может быть оговорена юридическими нормами и соглашениями. Публичные облака имеют очень широкие границы, где доступ пользователей к облачным сервисам почти не имеет ограничений.
- Частное облако (Private Cloud): модель развертывания облака, где облачные сервисы используются исключительно одним пользователем, и ресурсы находятся только под его контролем. Частное облако может быть развернуто в частном или государственном учреждении, управляться самой организацией или третьей стороной, базироваться на территории самого учреждения или за ее пределами. Частным облаком могут пользоваться только члены организации, прошедшие авторизацию перед подключением.
- Общественное облако (Community Cloud): модель развертывания облака, где облачные сервисы поддерживаются каким-либо сообществом пользователей, которые имеют общие требования к предмету услуги, а также взаимодействуют друг с другом в рамках этой услуги. Ресурсы находятся под контролем, по крайней мере, одного члена этой сообщества. Такое облако может принадлежать и управляется одной или более организациями, входящими в сообщество или третьей стороной. Оно может быть развернуто на территории одной из таких организаций или у третьей стороны. Отличие от модели Частного облака состоит в том, что Общественное облако подразумевает, в частности, подключение пользователей, объединенных общим интересом, а не членством одной в организации. Отличие от Публичного облака состоит в наличии определенного критерия, по которому осуществляется авторизация: политики безопасности, соглашения, членство в сообществе и т.д.
- Гибридное облако (Hybrid Cloud): модель развертывания облака, включающая в себя комбинацию, по меньшей мере, двух других различных моделей развертывания облака. Участвующие в Гибридном облаке модели являются различными по своей сути, но связаны друг с другом соответствующей технологией, что обеспечивает взаимодействие, мобильность данных и переносимость приложений. Гибридное облако может быть

развернуто в частном или государственном учреждении, управляться самой организацией или третьей стороной, и может базироваться на территории самого учреждения или за ее пределами. Такие облака могут быть полезны в ситуациях, когда происходит объединение нескольких структур с сохранением уникальных свойств каждой.

Услуга типа «виртуальный рабочий стол» может быть развернута и предоставлена по модели частного или гибридного облака. Говоря о вариантах развертывания и классификации облачных услуг, следует отметить понятие Соглашения об уровне предоставляемой услуги (Service Level Agreement, SLA). Этот термин обозначает формальный договор между заказчиком услуги и её поставщиком, который содержит описание услуги, права и обязанности сторон, а также согласованный уровень качества предоставления данной услуги. Более подробно особенности применения SLA рассмотрены в [18, 19, 91].

1.1.4 Облачные платформы и принципы их работы

Облачные платформы, разворачиваемые в удаленных Центрах Обработки Данных, представляют собой программно-аппаратные комплексы, составленные из ряда программных продуктов, установленных и специальным образом настроенных на одном или множестве серверов [57]. В целом, облачные системы имеют различные виды реализации в зависимости от области их применения, предполагаемого функционала, ресурса сети и оборудования. Тем не менее, возможно проследить общие принципы формирования таких систем. Так, наибольшее развитие и применение на сегодняшний день получили гипервизоры.

Гипервизор (англ. Hypervisor) – это программный или программно-аппаратный комплекс, обеспечивающий одновременную работу нескольких операционных систем на одном и том же сервере (компьютере) за счет технологии виртуализации серверов. Гипервизор устанавливается на сервер в качестве основной серверной операционной системы (см. рисунок 1.2).

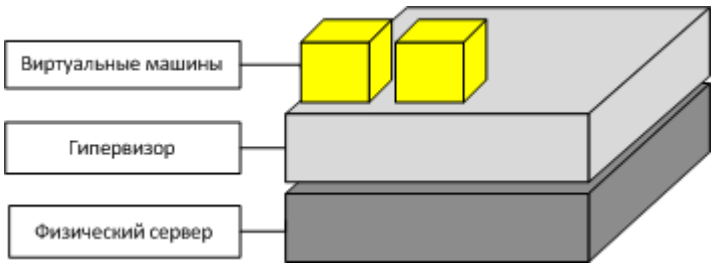

Рис. 1.2. Размещение гипервизора и виртуальных машин на сервере

Виртуализация серверов – это программная имитация с помощью специального ПО, которое устанавливается на сервере и гипервизоре, аппаратного обеспечения компьютера: процессора, памяти, жесткого диска, то есть создание так называемой виртуальной машины (ВМ). На такую виртуальную машину устанавливается операционная система, которая запускается и работает в ней так же, как и на обычном компьютере [5].

Гипервизор обеспечивает полную изоляцию операционных систем (ОС), устанавливаемых в ВМ, друг от друга, однако возможность сетевого взаимодействия между ними, в случае необходимости, может быть организована. Следует также отметить, что использование гипервизора позволяет ощутимо упростить администрирование за счет консолидации ВМ на одном сервере, динамическое разделение ресурсов между виртуальными машинами, автоматическое резервирование пользовательских данных и т.д.

1.2 Архитектура услуги «виртуальный рабочий стол»

1.2.1 Функциональные элементы и анализ их взаимодействия

«Виртуальный рабочий стол» (Desktop as a Service, DaaS) – это облачная услуга, в которой пользователю облачных услуг предоставляются следующие возможности: построение, запуск, настройка и эксплуатация функций удаленного рабочего места [77]. Пользователь работает с интерфейсом, предоставляемым ему по сети.

Рабочее место пользователя организуется следующим образом: вместо установки операционной системы и приложений на пользовательское устройство их размещают на серверах поставщика облачных услуг (Cloud Service Provider, CSP) удаленно в облаке. Для доставки рабочего стола пользователю по IP сети используются протоколы доставки виртуального рабочего стола (протоколы виртуализации). В качестве пользовательских устройств могут выступать смартфоны, планшетные компьютеры, ноутбуки, а также специально предназначенные для подключения к удаленному серверу терминальные устройства – тонкие клиенты. Для пользовательских устройств необходимым условием возможности подключения к виртуальному рабочему столу является наличие в их операционной системе (прошивке) специального программного обеспечения – агента, поддерживающего тот или иной протокол доставки виртуального рабочего стола.

В обобщенном виде инфраструктура услуги «виртуальный рабочий стол» показана на рисунке 1.3.

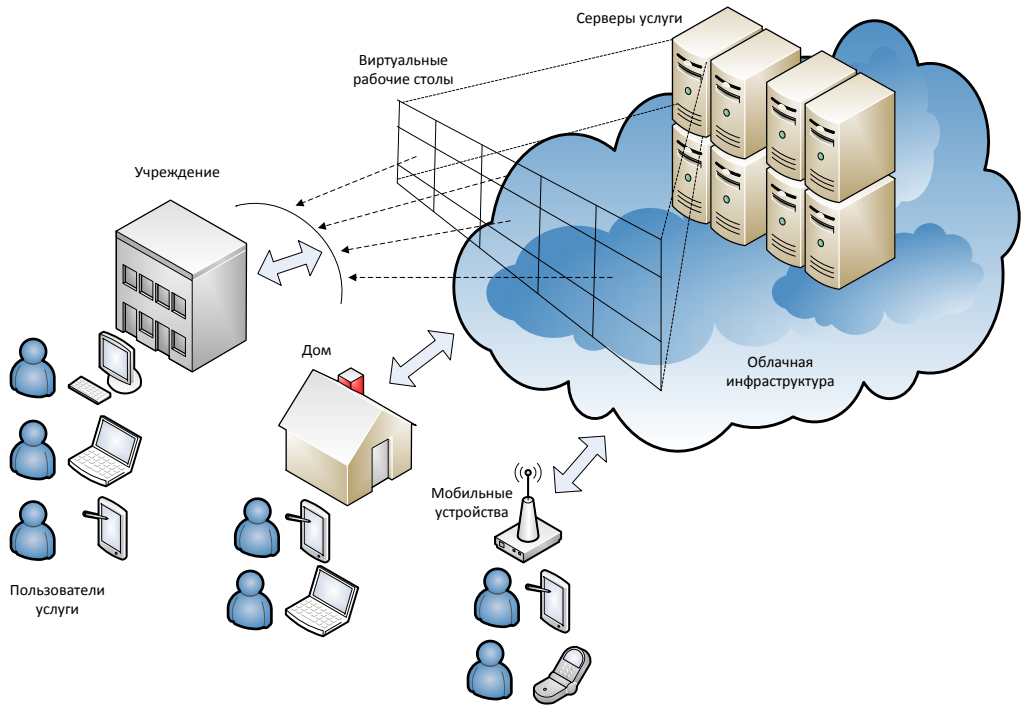

Рис. 1.3. Концепция услуги «виртуальный рабочий стол»

Архитектура услуги типа DaaS основана на клиент-серверной модели. На рисунке 1.4 показаны функциональные блоки архитектуры DaaS и их взаимодействие. В общем виде это: DaaS-клиент, менеджер подключений, система доставки удаленных рабочих столов, виртуальная инфраструктура, пул ресурсов. Рассмотрим их в отдельности.

1. DaaS-клиент. Это программа, позволяющая устройству, в ОС которого она установлена, взаимодействовать с сервером DaaS посредством протокола доставки удаленного рабочего стола. Пользователи инициируют запуск услуги на своем устройстве. Разновидностей таких устройств много (мобильные гаджеты, персональные компьютеры без жесткого диска, тонкие клиенты. На сегодняшний день имеются реализации ряда DaaS-клиентов, разработанных производителями облачных платформ.

Основными функциями DaaS-клиента являются: посылка запросов на инициализацию терминальной сессии, получение IP адреса, получение запросов на аутентификацию, отправка логина и пароля, поддержание сессии в рабочем состоянии, посылка запросов на завершение терминальной сессии с опциональным отключением виртуальной машины.

2. Менеджер подключений. Этот логический блок осуществляет подключение пользователя к виртуальному рабочему столу. В зависимости от конфигурации услуги это может быть индивидуальный доступ к собственной персонализированной ВМ, индивидуальный доступ к случайной доступной ВМ, индивидуальный доступ к общей ВМ.

Основными задачами Менеджера являются:

- аутентификация пользователя и проверка лицензии на доступ к ВМ и приложениям;
- подключение к виртуальному рабочему столу;
- координация протокола виртуализации;
- подключение необходимых ресурсов, в том числе файлового хранилища.

Кроме того, Менеджер отвечает за распределение нагрузки (Load balancing) и управление количеством подключаемых пользователей. Менеджер использует блоки системы доставки удаленных рабочих столов, виртуальной инфраструктуры и пула ресурсов для выделения терминальной сессии необходимых вычислительных ресурсов, ресурсов сети и хранилища.

3. Система доставки удаленных рабочих столов. Основной функцией этого блока является доставка виртуального рабочего стола от сервера к клиенту. Для этого используется протокол доставки рабочего стола (протокол виртуализации), который организует логический канал внутри сетевого канала передачи данных.

4. Виртуальная инфраструктура. Основными функциями этого блока являются: поддержка аппаратного и программного обеспечения, эмуляция виртуальных ресурсов для создания ВМ, создание и настройка ВМ средствами гипервизора, клонирование и перемещение ВМ, репликация данных и др.

5. Пул ресурсов. В этот блок входят: дисковое пространство (которое выделяется на виртуальные жесткие диски пользовательских виртуальных машин, установку пользовательских операционных систем и приложений, хранение данных и т.д.), оперативная память, процессор, сетевая карта. Следует отметить, что пользовательской ВМ может быть эмулировано и выделено различное аппаратное обеспечение в независимости от того, какое установлено на физическом сервере.

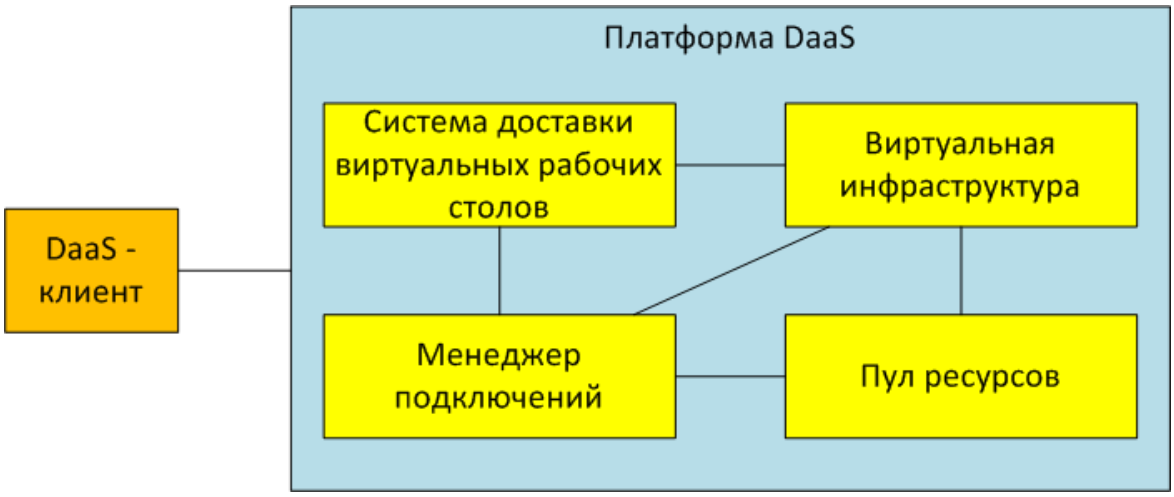

Рис. 1.4. Функциональные блоки архитектуры DaaS

При запуске услуги DaaS перечисленные логические блоки архитектуры согласно [77] должны взаимодействовать следующим образом:

1. DaaS-клиент обращается к Менеджеру подключений посредством защищенного протокола (например, ssh), Менеджер дает доступ пользователю, проверяя логин и пароль.

2. Менеджер подключений активирует профиль пользователя и подключает необходимые ресурсы. Если для пользовательской ВМ по каким-либо причинам не хватает ресурсов, Менеджер подключений обращается к Виртуальной инфраструктуре для выделения дополнительных ресурсов.

3. Менеджер подключений полностью включает пользовательскую ВМ и активизирует виртуальный рабочий стол, после чего посылает сигнал об этом DaaS-клиенту.

4. DaaS-клиент связывается с виртуальным рабочим столом по сети при помощи Системы доставки удаленных рабочих столов, которая использует протокол виртуализации.

5. Когда пользователь завершает работу, DaaS-клиент посылает сигнал о завершении сессии без потери данных (на свое усмотрение пользователь может либо выключить свою ВМ, если ему это позволено политикой услуги, либо просто отключиться от сессии).

6. Менеджер подключений получает сигнал о завершении сессии, затем он либо выключает ВМ либо отключает сессию. После этого он посылает сигнал Пулу ресурсов и Виртуальной инфраструктуре о завершении текущего сеанса и освобождении ресурсов.

Рассмотрим отдельно серверную и клиентскую части архитектуры DaaS, а также сетевое взаимодействие между ними.

1.2.2 Серверная часть архитектуры DaaS

Структуру и принцип работы серверной части услуги DaaS наглядно можно пояснить, опираясь на рисунок 1.5, где показана обобщенная схема одного сервера в «облаке». На практике, когда нужно подключить множество абонентов, применяются одновременно множество серверов, которые в случае необходимости могут быть объединены в кластер. Логически разделим показанное на рисунке 1.5 на четыре слоя.

Первый слой – аппаратный: сервер, его сетевая карта, жесткие диски, процессор, оперативная память и т.д.

Второй слой – платформа (гипервизор). Как было сказано выше, это специализированная операционная система, имеющая целый ряд отличий от обычных пользовательских

операционных систем, самое главное из которых – это возможность создания ВМ, в каждой из которых может быть установлена и запущена отдельная операционная система.

Третий слой – виртуальные машины, создаваемые гипервизором путем эмуляции различного аппаратного обеспечения и ресурсов. Следует отметить, что в этом слое обязательно наличие Управляющей ВМ (Менеджер подключений), которая играет наиболее важную роль в развертывании и функционировании облачной системы. На этапе развертывания с ее помощью осуществляется создание, настройка, администрирование пользовательских ВМ. На этапе эксплуатации она полностью контролирует взаимодействие облачной платформы и пользователя. Более детально процесс создания виртуальных машин описан, в частности, в [83].

Четвертый слой – операционные системы и приложения. В этом слое работает пользователь, запуская приложения, установленные непосредственно на его «собственную» ОС либо в отдельном сервере приложений. Следует отметить, что для того, чтобы ОС, запускаемая в ВМ, могла функционировать в рамках архитектуры DaaS, необходимо установить на нее специальную служебную программу – агент. Для каждой системы виртуализации эта программа своя, она устанавливается автоматически в процессе развертывания ВМ.

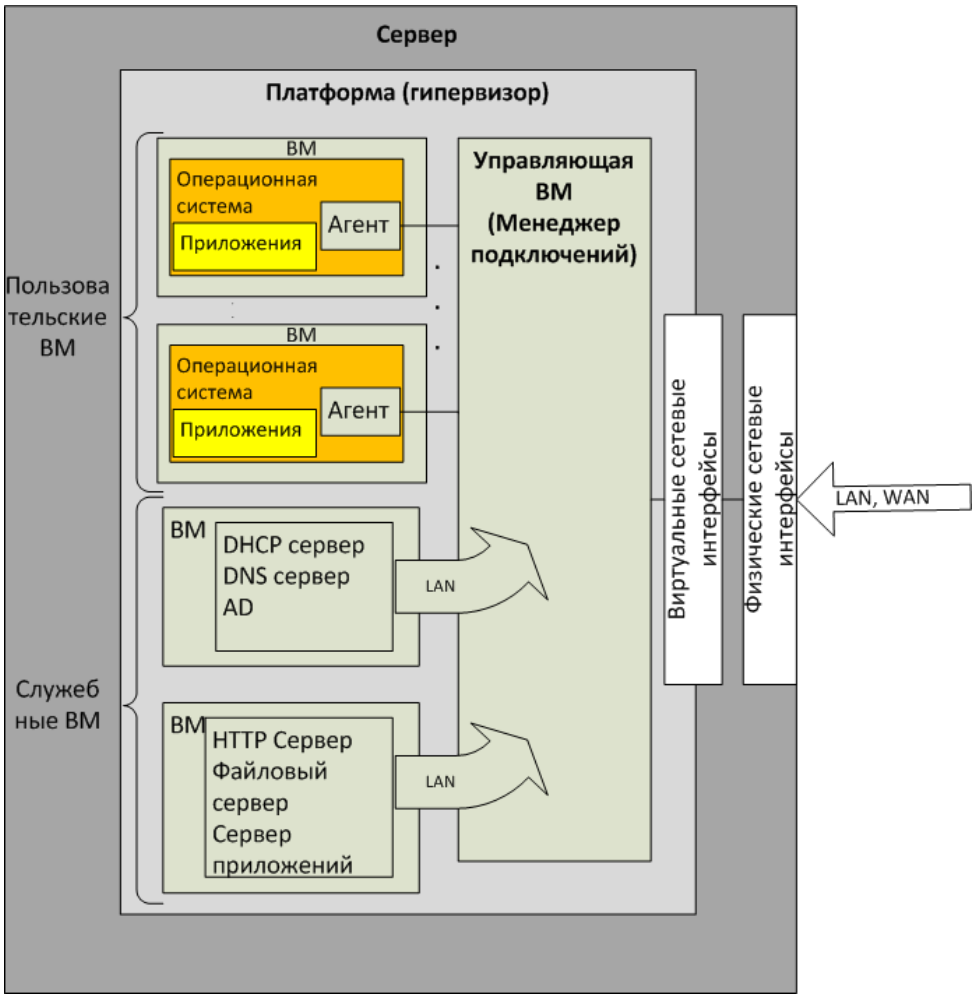

Рис. 1.5. Обобщенная архитектура серверной части услуги DaaS

Следует отметить, что все ВМ, разворачиваемые в облачной среде, можно разделить на служебные и пользовательские. На служебных ВМ выполняются необходимые для работы инфраструктуры программы: базы данных (MySQL, MariaDB, PostgreSQL и др.), сетевые службы и т.д. Наиболее важными и необходимыми для функционирования внутренней облачной инфраструктуры сетевыми службами являются DHCP, DNS, сервис каталогов и групповых политик Microsoft Active Directory (AD). При необходимости, в служебных ВМ могут быть развернуты HTTP-сервер, FTP-сервер, SMTP-сервер и др. Кроме того, в некоторых учреждениях, таких, как например, медицинские центры, пользователи работают со специализированными программами и оборудованием, подключаемым к ним. Тогда, в зависимости, от характера работы, количества задействованных в такой работе пользователей, интенсивности обращения к услуге, такие программы разворачиваются либо в служебных, либо в пользовательских ВМ.

Применение систем виртуализации позволяет сократить количество серверов благодаря консолидации, то есть на одном сервере можно запустить нужное число гостевых ОС в виртуальной среде. Это позволит сэкономить на стоимости оборудования, а также снизить энергопотребление и сократить штат системных администраторов за счет возможности удаленного доступа к консоли виртуальных серверов, что позволяет одному администратору обслуживать достаточно большую инфраструктуру сети, в то время как ранее требовалось несколько специалистов.

Операционная система, запущенная внутри виртуальной машины, «не знает», какое оборудование установлено на физическом сервере. Поэтому, при замене оборудования из-за аварии или при переезде на новый сервер, гостевые ОС сохранят свою конфигурацию, роли в групповых политиках и файлы пользователя.

На сегодняшний день существует несколько систем виртуализации. Рассмотрим наиболее распространенные и функциональные из них: VMware ESXi [27], Citrix XenServer [22], Microsoft Hyper-V [102], Linux KVM [88].

VMware ESXi – это программный продукт для виртуализации уровня предприятия, предлагаемый компанией VMware. ESXi является гипервизором и устанавливается непосредственно на сервер, то есть при установке не требуют наличия на машине установленной операционной системы. Этот гипервизор позволяет разделить физический сервер на логические разделы (создание виртуальных машин), включает в себя средства управления виртуальными ресурсами и предъявляет определённый набор требований к аппаратному обеспечению: например, наличие поддержки виртуализации со стороны архитектуры процессора является обязательным [6].

Citrix XenServer – это гипервизор, являющийся разработкой Кембриджского университета, на данный момент принадлежит компании Citrix. С 2007 года компания опубликовала XenServer в свободном доступе, что подразумевает под собой его полную бесплатность и открытость исходных кодов. На данный момент для скачивания на сайте компании Citrix доступна полноценная версия гипервизора [22].

Microsoft Hyper-V – система виртуализации на основе гипервизора. Ранее была известна как виртуализация Windows Server (Windows Server Virtualization). Имеет широкий набор возможностей для управления ВМ. Отличительной чертой Hyper-V является его интеграция с Active Directory, которая, как и гипервизор, является компонентом Windows Server [102].

KVM (Kernel-based Virtual Machine) – это программное решение, обеспечивающее виртуализацию в среде Linux (на основе встроенного гипервизора) , которая поддерживает аппаратную виртуализацию на базе Intel VT (Virtualization Technology) либо AMD SVM (Secure Virtual Machine). Программное обеспечение KVM состоит из загружаемого модуля ядра, процессорно-специфического загружаемого модуля и компонентов пользовательского режима. Все компоненты ПО KVM имеют открытый исходный код [88].

Как было сказано выше, физические сервера, на которых запускаются ВМ, могут быть объединены в кластер, и в случае отказа одного из серверов автоматически мигрировать на другой. Перемещение ВМ проходит незаметно для пользователей. Такие технологии у разных разработчиков называются по-разному: например, у Microsoft она называется Live Migration [102], у VMware – vMotion [6].

Подобные технологии позволяют проводить работы, связанные с выключением ВМ, прямо в рабочее время и не отключая пользователей от работы. Кроме того, если структура сети построена соответствующим образом, запущенные виртуальные машины могут автоматически перемещаться на менее нагруженные сервера в запланированном режиме. В инфраструктуре сети, построенной на основе технологий от Microsoft, для этого используются службы System Center Virtual Machine Manager (Operations Manager) [102]. В инфраструктуре, построенной на технологиях от VMware и Citrix, используются программы vSphere Center [6] и XenCenter [22], соответственно. Более подробно наиболее особенности платформ виртуализации рассмотрены в [34].

Рассмотренные платформы виртуализации имеют одну общую главную функцию – непосредственно создание ВМ. Они имеют набор сходных функций, однако имеются и различия.

Сравнительная характеристика систем виртуализации приведена в таблице 1.1.

Таблица 1.1. Сравнение систем виртуализации

	VMware ESXi	Citrix XenServer	Microsoft Hyper-V	Linux KVM
Встроенное резервное копирование ВМ	Да	Да	Нет	Нет
Живая миграция виртуальных машин	Да	Да	Да	Да
Миграция ВМ между системами хранения данных	Да	Нет	Нет	Нет
Автоматическое распределение нагрузки между хостами	Да	Да	Нет	Нет
Автоматическое распределение нагрузки между СХД	Да	Нет	Нет	Нет
Распределение ресурсов процессора и памяти между ВМ	Да	Да	Да	Да
Распределение ресурсов сетевых адаптеров (I/O) между ВМ	Да	Нет	Нет	Нет

Отдельно следует рассмотреть различные реализации Менеджера подключений. Каждый из производителей облачных платформ имеет собственную его реализацию. В большинстве случаев она представляет собой сложную программу, устанавливаемую на вспомогательную ОС, которая затем упаковывается в отдельную Управляющую ВМ. У некоторых производителей, таких как Red Hat, Менеджер подключений и гипервизор реализованы совместно в единый программный комплекс. Тем не менее, в большинстве случаев имеет место модульная реализация, описанная выше.

Существует большое множество служебных программ, работающих как в связке с Менеджером подключений, так и отдельно, которые позволяют всевозможным образом расширить услугу DaaS. Например, активация пользователем доступности своего рабочего стола в облаке с различных устройств, всевозможные политики безопасности, конвергенция служб, запущенных в ОС виртуального рабочего стола со службами, запущенными на физическом компьютере и др.

Наиболее популярными программными реализациями Менеджера подключений являются Citrix VDI-IN-A-BOX и Citrix XenDesktop [33], VMware Vsphere [6], Microsoft Hyper-V [102], Red Hat Enterprise Virtualization [106], Huawei FusionCloud Access [82] и др.

Следует, также, отметить, что услуга «виртуальный рабочий стол» реализована различными производителями облачного ПО с небольшими различиями и особенностями, присущими конкретному производителю. Часто, говоря об услуге, используют термин VDI (Virtual Desktop Infrastructure). В [31] приведен обзор рекомендаций МСЭ-Т, относящихся к этой услуге, описаны основные варианты реализации услуги на практике, указаны основные преимущества и недостатки перевода рабочих мест пользователей в облачную среду.

Резюмируя описанное, можно выделить ряд достоинств виртуализации: снижение требований к аппаратному обеспечению на стороне клиентов, повышение безопасности, значительное упрощение администрирования и поддержки. К недостаткам можно отнести следующее: повышение требований к серверам, как по производительности, так и по надежности, относительная дороговизна лицензий и сервера.

1.2.3 Клиентская часть

Клиентская часть архитектуры DaaS может быть организована различными способами, в качестве оконечного оборудования могут выступать: мобильные устройства, устаревшие компьютеры, тонкие клиенты – маломощные малогабаритные рабочие станции, терминалы и т.п. Общим является обязательная установка на клиентское устройство специального программного обеспечения – DaaS-клиента, который полностью отвечает за все виды взаимодействия с платформой, развернутой на сервере.

На практике большинство крупных производителей программных комплексов и инструментов для предоставления услуги DaaS, называют DaaS-клиент Агентом. Это название является устоявшимся в отрасли.

К основным функциями Агента относятся: посылка запросов на инициализацию терминальной сессии, отправка и получение запросов на аутентификацию, верификация логина и пароля, поддержание сессии в рабочем состоянии, посылка запросов на завершение терминальной сессии с опциональным отключением виртуальной машины.

В большинстве случаев Агент реализован в виде программного обеспечения и устанавливается либо в прошивку устройства, либо в его операционную систему. В настоящий

момент существуют реализации Агентов различных систем доставки виртуальных рабочих столов практически для всех известных операционных систем или мобильных платформ.

Следующим важным элементом архитектуры пользовательских устройств является процессор устройства, отвечающий за обработку поступающей от сервера информации и передачу ее на графическую подсистему для отрисовки пользователю. На сегодняшний день существует большое множество процессоров, устанавливаемых на пользовательские устройства. От производительности процессора, в частности, зависит уровень комфорта работы с услугой.

Существует несколько различных типов развертывания услуги DaaS согласно [76]. Рассмотрим их кратко далее.

- Классификация по уровню обслуживания: пользователи могут быть классифицированы согласно различным типам соглашения об уровне предоставления услуги SLA.
- Классификация по типу предоставляемого рабочего стола. Например, некоторые пользователи могут получать на свой терминал Веб-приложение, некоторые полноценную ОС. Последние могут разделяться по способу доступу к ОС: общая ОС на несколько пользователей, индивидуальная ОС, одноразовый доступ к гостевой ОС.
- Классификация по типу размещения виртуальной среды: постоянная инфраструктура или временное развертывание.
- Классификация по типу предоставления ресурсов пользователю: фиксированные ресурсы либо возможность самостоятельно или по запросу расширять рабочее пространство.

Очевидно, что такие классификации не всегда носят строгий характер, зачастую являются взаимодополняющими, а не взаимоисключающими. Наибольшую востребованность на сегодняшний день получило предоставление полноценной индивидуальной ОС каждому пользователю в рамках услуги «виртуальный рабочий стол». Именно такой тип предоставления услуги будет рассмотрен в данной работе.

Сетевое взаимодействие между сервером и клиентами представляет большой интерес в рамках поставленной задачи, поэтому рассмотрим его более подробно в следующем подпараграфе.

1.2.4 Сетевое взаимодействие и протоколы доставки виртуальных рабочих столов

Облачные услуги доставляются пользователю посредствам IP сетей. Следовательно, любое сетевое взаимодействие, связанное с передачей данных, происходит в соответствии с правилами стека TCP/IP. Для доставки виртуального рабочего стола по сети используются так называемые протоколы доставки рабочего стола (протоколы виртуализации) [43]. Эти протоколы проприетарны т.е. права на них принадлежат конкретным компаниям. Как следствие, эти протоколы закрыты, механизмы их работы не разглашаются. Однако, исходя из логики функционирования современных сетей, можно сделать вывод о том, что такие протоколы работают поверх протокола TCP.

На сегодняшний день наибольшее распространение получили следующие протоколы виртуализации: Microsoft Remote Desktop Protocol (RDP) [29], Citrix ICA и Citrix HDX [33], VMware PCoIP [27], Red Hat SPICE [104], VNC [99]. Поясним принцип работы протоколов виртуализации на примерах типичных протоколов RDP, ICA и SPICE.

Remote Desktop Protocol (RDP) является прикладным протоколом, который базируется на TCP и использует по умолчанию порт 3389. Для установки RDP-сессии клиент и сервер обмениваются сообщениями для согласования параметров сессии: шифрования, аутентификации и т.д. После установки соединения на транспортном уровне инициализируется RDP-сессия, в рамках которой происходит передача данных от сервера к клиенту: сервер передает графический данные с пользовательского виртуального рабочего стола и ожидает входные данные от клавиатуры и мыши (команды событий мыши и клавиатуры). В качестве графических данных выступает снимок экрана рабочего стола, передаваемый в одном из форматов сжатия изображений, т.е. в процессе работы сессии в сеть от сервера к клиенту передается непрерывный поток картинок. Обобщенная схема работы протокола показана на рисунке 1.6.

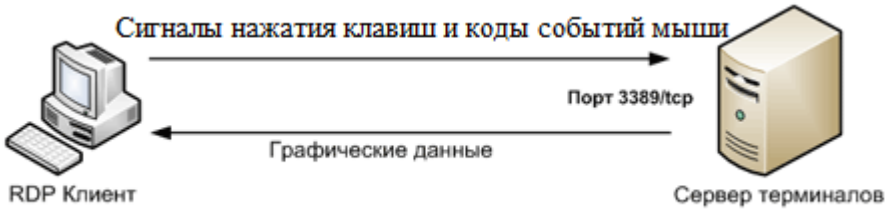

Рис. 1.6. Обобщенная схема работы протокола RDP

RDP клиент обрабатывает полученные команды и выводит изображения с помощью своей графической подсистемы. Пользовательский ввод по умолчанию передается при помощи скан-кодов клавиатуры. Сигнал нажатия и отпускания клавиши передается отдельно при помощи специального флага [28, 29]. RDP поддерживает несколько виртуальных каналов в рамках одного соединения, которые могут использоваться для обеспечения дополнительного

функционала. Характеристики виртуальных каналов согласуются на этапе установки соединения. В упрощенном виде схема обмена сообщениями между клиентом и сервером в рамках RDP-сессии показана на рисунке 1.7.

Рис. 1.7. Упрощенная схема обмена сообщениями между клиентом и сервером в рамках RDP-сессии

Протокол ICA также имеет процедуру согласования параметров сессии, передает копии экрана удаленного рабочего стола в формате JPEG и поддерживает виртуальные каналы, как показано на рисунке 1.8. Под виртуальными каналами производитель подразумевает логическое разнесение данных, передаваемых от сервера к клиенту в тех случаях, когда клиент задействует дополнительный функционал, например, подключение принтера, сканера, кардридера, флеш-накопителя и т.д.

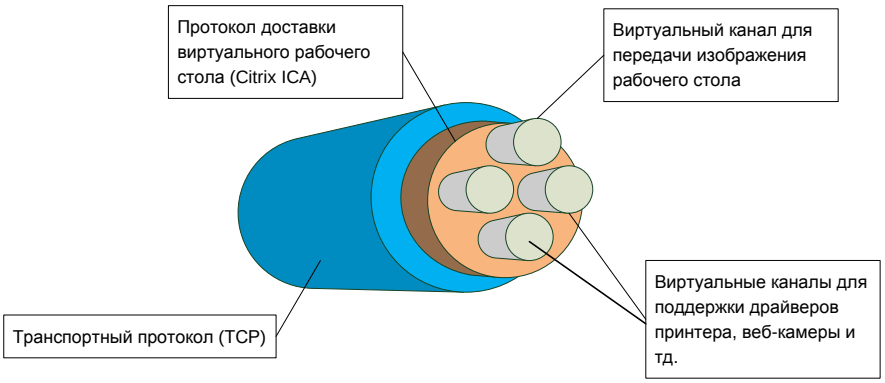

Рис. 1.8. Виртуальные каналы протокола ICA

Протокол SPICE передает изображения рабочего стола, сжатые при помощи проприетарных алгоритмов сжатия Quic и GLZ, которые работают независимо друг от друга [107, 108]. Специальный алгоритм эвристически определяет целесообразность применения

одного из них в каждом конкретном случае. Сжатие изображений осуществляется по схемам без потерь (Lossless), сжатие видео кадров – по схемам с потерями (Lossy).

Данный протокол также поддерживает кэширование изображений рабочего стола на клиентской стороне, целью которого является исключение отправки избыточных изображений, при этом каждое отправляемое изображение помечается уникальной меткой (id) для обеспечения выстраивания правильной последовательности при обработке и отображении пользователю. В случае переполнения буфера агент пользовательского устройства отправляет служебное сообщение на сервер, содержащее номер последнего принятого изображения.

Реализация передачи на сервер событий нажатия клавиш клавиатуры и движений мыши аналогична прочим протоколам виртуализации – агент пользовательского устройства детектирует эти события (движения мыши передаются в виде координат на координатной сетке, нажатия клавиатуры – в виде специальных таблиц), кодирует их и отправляет эти данные на сервер [106].

Протокол SPICE поддерживает возможность доставки аудио (звука) на пользовательское устройство при работе виртуального рабочего стола. В этом случае аудио кадры кодируются кодеком CELT [112]. Кроме того, данный алгоритм имеет возможность эвристического распознавания факта запуска пользователем видео с последующим выделением видео ряда в отдельный поток, кодируемый видео кодеками (M-JPEG согласно [108]) с целью уменьшения объема передаваемых по сети данных, что положительно сказывается на восприятии работы с услугой. Более подробно о способах выделения видео и особенностях работы услуги в таком режиме будет сказано в разделе 3.

Описанные функции рассмотренных протоколов в той или иной степени реализованы в каждом из них, в большинстве случаев техническая информация о них не разглашаются, однако для понимания общих принципов их работы, приведенные данные являются достаточными.

Соотнесем далее расположение компонентов услуги DaaS относительно функциональных уровней облачной архитектуры из [68] (см. рисунок 1.9) и кратко поясним роль протоколов доставки виртуального рабочего стола относительно них.

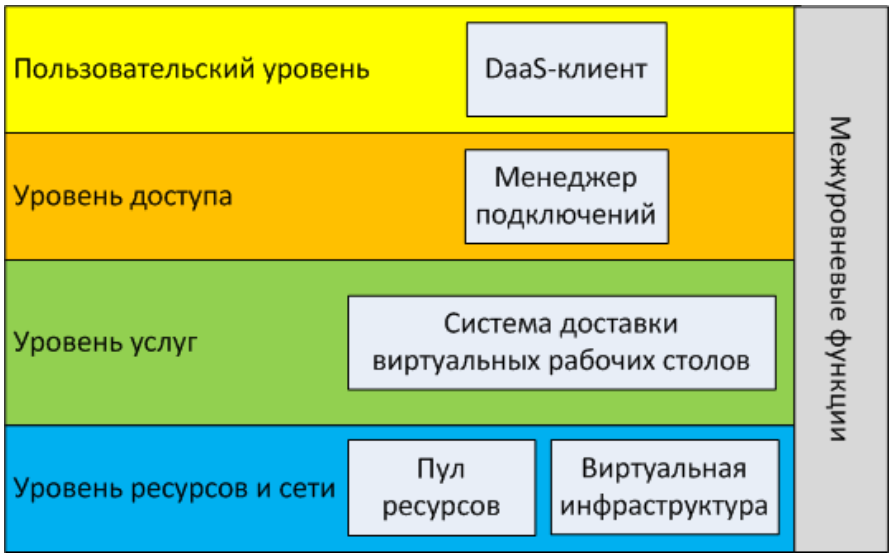

Рис.1.9. Компоненты услуги DaaS и уровни облачной архитектуры

В общем случае рассматриваемые протоколы осуществляют так называемые межуровневые функции: за счет них осуществляется взаимодействие уровней облачной архитектуры, а также решается основная задача – доставка изображений рабочего стола до пользовательского устройства. Номинально ядро протокола локализовано в уровне услуг. На первом этапе менеджер посредствам протокола связывается с клиентской стороной и производит процедуры авторизации и аутентификации. Далее DaaS-клиент виртуальной машины посредствам протокола начинает передачу данных на пользовательское устройство. Уровень ресурсов сети очень тесно связан с уровнем услуг, работа протоколов опирается на данные, предоставляемые им: информации о канале передачи данных, объеме доступной памяти и т.п.

Таким образом, в данном параграфе на примере протоколов RDP, ICA, SPICE были рассмотрены принципы работы протоколов доставки виртуального рабочего стола и некоторые их характерные особенности.

1.3 Качество облачной услуги

1.3.1 Подходы к определению качества облачной услуги

Облачные услуги, доставляемые пользователю через сети передачи данных, должны удовлетворять требованиям к качеству, выраженным в объективных и субъективных параметрах.

Согласно определению, приведенному в Рекомендации Е.800 МСЭ-Т [67], качество обслуживания (Quality of Service, QoS) – это совокупность характеристик услуги электросвязи, которые имеют отношение к ее возможности удовлетворять установленные и предполагаемые потребности пользователя услуги.

Качество обслуживания зависит от сетевых и несетевых параметров [80]. Сетевые параметры в общем виде обозначаются как качество функционирования сети (англ. Network Performance, NP), которое определяется [67], как способность сети или ее сегмента обеспечивать функции, относящиеся к связи между пользователями (в некоторых публикациях также используются другие варианты перевода этого термина: показатели работы сети, производительность сети). В зависимости от типа услуги для обеспечения требований к параметрам QoS, формируется некоторый перечень параметров NP.

В первую очередь, провайдеру услуги или оператору связи необходимо выбрать параметры NP, позволяющие обеспечить требования к QoS [69, 84], а затем убедиться в том, что предоставляемое или рекомендуемое им терминальное оборудование способно довести услугу до пользователя в приемлемом для него виде, поскольку сквозное (End-to-End) QoS зависит как от сети, так и от терминального оборудования [70].

Проанализируем публикации, рассматривающие качество облачных услуг.

В [113] предложен подход обеспечения QoS услуги, построенный на основе динамического сопоставления запросов к облачной системе с одним или несколькими физическими ресурсами. Авторами предложена архитектура системы электронного документооборота, которая обеспечивает качество, характерное для таких систем, для одновременного выполнения нескольких рабочих процессов в облачной среде. Рассмотрены, в частности, следующие параметры: размеры входных и выходных буферов, скорость передачи данных в потоках, время передачи запроса.

В [118] представлены алгоритмы управления доступом пользователей к услуге, развернутой по схеме SaaS, направленные на увеличение прибыли провайдера услуги и удовлетворенности клиентов. Эти алгоритмы основаны на выборе подключения пользователя к одной из виртуальных машин в зависимости от ее состояния, а также на динамическом управлении очередями к виртуальным машинам. Работа [121] посвящена построению модели, предполагающей ранжирование параметров QoS для облачных услуг на основе опыта использования услуги предыдущими ее пользователями. Рассматривался, в частности, параметр время работы с услугой.

В [26] решалась задача контроля изменения рабочей нагрузки путем динамического подключения/отключения обслуживающих серверов (виртуальных машин). В [98] предложен подход к построению алгоритмов балансировки нагрузки в облачных системах на основе учета

априорной информации о потребности исполняемых задач в наборе ресурсов. Схожая задача решалась в [2].

В [25] проведен анализ показателей качества функционирования облачных систем, решалась задача оптимизации распределения ресурсов, для чего применялось гистерезисное управление количеством включенных приборов (виртуальных машин); для получения вероятностно-временных характеристик использовался метод исключения состояний. В работе [8] разработан рекуррентный метод вычисления преобразования Лапласа-Стилтьеса распределения времени пребывания заявки в системе облачных вычислений и времени ожидания начала обслуживания, с помощью полученного преобразования проведен анализ характеристик системы.

В [81] предложен общий подход к анализу работоспособности облачной услуги, проиллюстрированный на примере облака услуги типа IaaS, для которой задержка в обслуживании и задержки ответа на предоставление услуг являются двумя ключевыми показателями QoS. В работе в большей степени рассматриваются задержки, связанные с перенаправлением пользователя к одной из виртуальных машин, ее подготовки к работе и запуску.

В рассмотренных публикациях анализируются параметры, определяющие качество услуги, которые относятся к серверу и терминальному оборудованию. К параметрам сервера относят: размеры входных и выходных буферов, скорость передачи данных в потоках, время передачи запроса, число пользователей, время работы с услугой, количество выделяемых серверов, вероятность подключения к тому или иному серверу, время пребывания заявки в системе, время перенаправления пользователя к виртуальной машине. К параметрам клиента: размеры буферов, время обработки на устройстве.

Следует отметить, что среди работ, посвященных анализу качества облачных услуг, не рассмотрена услуга «виртуальный рабочий стол».

1.3.2 Качество услуги «виртуальный рабочий стол» и классификация ее пользователей

Услуга «виртуальный рабочий стол» рассматривалась в работах [9,12,54,60], однако они носят обзорный характер и не затрагивают вопросов, связанных с оценкой ее качества.

Описание работы услуги, типовые схемы развёртывания, анализ рекомендаций МСЭ-Т, посвященных как облачным услугам в общем, так и услуге «виртуальный рабочий стол» в

отдельности, приведены в [31]. Существующие возможности программно-аппаратной среды для инфраструктуры услуги – облачные платформы, системы виртуализации и доставки виртуального рабочего стола подробно проанализированы в [34].

Анализ качества услуги «виртуальный рабочий стол» следует начать с классификации ее пользователей, поскольку требования к нему могут меняться в зависимости от выполняемой деятельности в рамках рабочего стола. Проанализировав деятельность пользователей, сформируем классификацию целевой аудитории услуги, выполненную согласно специфике работы в рамках рабочего стола, а также с учетом типов используемых приложений. Можно выделить следующие три типа пользователей услуги.

1. По специфике работы и политике компаний мало использующие мультимедиа со своих рабочих мест. Это могут быть офисные работники, программисты, журналисты, аналитики, редакторы. Для такой категории пользователей типична работа с веб-браузером, текстовым редактором, файловым менеджером.
2. Использующие офисные пакеты, а также иногда просматривающие видеоролики. К такому типу пользователей можно отнести работающих на дому (фрилансеров), аналитиков, маркетологов, проектировщиков. Для таких пользователей характерно эпизодическое использование видео.
3. По специфике своей деятельности работающие с мультимедиа, т.е. активно использующие и видео и аудио. Это могут быть маркетологи, аниматоры, графические дизайнеры, сотрудники call-центров или частные лица, заменившие свои ПК на облачный виртуальны рабочий стол. Данная категория пользователей не может обходиться без звука, однако его предоставление сопряжено с техническими сложностями (так называемый «проброс гарнитуры» с пользовательского терминала в виртуальную машину), поэтому данный тип пользователей вынесен в отдельную категорию. Технические аспекты данного случая рассмотрены в разделе 4.

Нужно учитывать, что у разных типов пользователей есть как общие, так и различные особенности. К общим можно отнести необходимость работы с окнами операционной системы, работу с файлами. К различным – необходимость просмотра видео, прослушивания аудио, мультимедиа.

Следующим шагом является выявление параметров, определяющих качество рассматриваемой услуги. На основании анализа работы услуги, представленного выше, выявлены параметры, характеризующие услугу «виртуальный рабочий стол».

К NP относятся параметры, которые могут быть измерены и участвуют в построении, работе и настройке системы и не зависят от действий пользователя [80]. Для услуги

«виртуальный рабочий стол» такими параметрами являются: канальная скорость передачи данных, скорость передачи данных на интерфейсах, транспортная задержка. Следует отметить, что провайдер облачной услуги должен, во-первых, формулировать требования к качеству услуги, а исходя из них заказывать сетевые характеристики у провайдера сети передачи данных; во-вторых, если пользователь выбирает оператора сети самостоятельно, давать рекомендации относительно сетевых характеристик.

Согласно Рекомендации Е.430 МСЭ-Т [66] параметры QoS можно классифицировать в виде матрицы 3х3, содержащей следующие критерии. Скорость (Speed) – временной критерий, характеризующий скорость подключения и функционирования. Точность (Accuracy) характеризует корректность передаваемых данных, а также функционирования услуги в целом. Надежность (Dependability) характеризует степень уверенности в том, что услуга будет предоставляться вне зависимости от скорости и точности, но в пределах заданного интервала наблюдения. Она определяется в рамках заданного интервала наблюдения согласно Рекомендации I.350 МСЭ-Т [71]. В строках матрицы расположены этапы работы услуги. Этот подход был предложен для телефонной сети, а затем распространен и для других услуг. Применим его к услуге «виртуальный рабочий стол» (см. таблицу 1.2).

Таблица 1.2. Параметры качества услуги «виртуальный рабочий стол»

Этапы работы услуги	Критерий		
	Скорость	Точность	Надежность
Установления соединения (подключение к услуге)	Среднее время до подключения	-	Вероятность отказа в подключении
Передача пользовательской информации	Среднее время отклика	Вероятность передачи изображения рабочего стола без искажений	Вероятность успешной передачи изображения рабочего стола
Отключение от услуги	Среднее время до отключения	-	-

Среднее время до подключения – это время с момента отправки пользователем запроса на подключение к виртуальному рабочему столу до момента подключения к услуге, вероятность отказа – вероятность того, что пользователь не будет подключен к услуге из-за переполнения мест в буфере. Более параметры качества будут рассмотрены в разделах 2 – 4.

Следует отметить, что среднее время до отключения т.е. время с момента введения пользователем команды на отключение от рабочего стола до момента его отключения от сервера, мало, поэтому в дальнейшем им будем пренебрегать. Кроме того, поскольку

вероятность того, что изображение не будет доставлено, мала, будем считать вероятность успешной передачи близкой к единице, ее рассматривать не будем.

Ключевыми параметрами, определяющими качество услуги, являются: серверное время обслуживания пользовательских запросов; параметры сервера (количество обслуживаемых пользователей, время обслуживания запросов); время обработки на пользовательском устройстве (время визуализации). Отдельно следует выделить параметр время отклика, объединяющий в себе все временные параметры услуги. На нем следует сосредоточить основное внимание. Анализ этого времени приведен в разделе 2, аналитические выражения для его определения – в разделах 3 и 4.

1.4 Постановка задач диссертационного исследования

Вопрос предоставления облачной услуги «виртуальный рабочий стол» с приемлемым для пользователя качеством в условиях среды Интернет является сложным и актуальным. Ключевой задачей становится планирование инфраструктуры услуги с учетом требуемого для различных задач качества с учетом специфики того или иного пользователя услуги.

Облачные услуги, как правило, предполагают использования соглашений SLA между провайдерами услуги и пользователями [18, 91]. Для формирования SLA провайдер должен оценивать ресурсы, выделяемые для предоставления услуги. Значит, необходим детальный анализ параметров, определяющих качество такой услуги, сценариев ее применения, целевой аудитории, особенностей ее развертывания, требований к ее инфраструктуре. Провайдеру необходимо выбрать показатели качества, нормативы для них, оценить возможность их обеспечения.

Изначально услуга «виртуальный рабочий стол» была разработана для локального развертывания в пределах одного офиса, здания, предприятия, однако с развитием облачной парадигмы началось активное ее применение в облачной среде на серверах в удаленных ЦОД, что породило ряд технических вопросов, связанных с обеспечением требуемого качества для пользователей, которые находятся уже не в локальной сети с сервером услуги, а в глобальных сетях. Обеспечение приемлемого качества услуг, при этом осложненное в связи с переносом в облачную среду, является одной из ключевых задач провайдеров услуг и разработчиков облачного программного обеспечения.

Для разработки метода оценки качества облачной услуги «виртуальный рабочий стол», требуется решить следующие задачи:

1. Проанализировать процесс работы услуги «виртуальный рабочий стол» и разделить его на фазы для возможности отдельного их исследования.

2. Определить сценарии работы услуги для различных категорий пользователей.

3. Разработать аналитические модели для каждой фазы работы услуги «виртуальный рабочий стол» и различных сценариев ее работы.

4. Для этих моделей получить формулы для расчета параметров, определяющих качество.

5. Определить параметры узлов облачной инфраструктуры, сети и пользовательских устройств, влияющие на качество обслуживания, найти допустимые диапазоны их значений и дать соответствующие рекомендации для провайдеров услуги.

1.5 Выводы по результатам первого раздела

В первом разделе представлены общие принципы построения архитектуры облачных услуг, в частности, рассмотрена услуга «виртуальный рабочий стол». В п. 1.1.1 проанализировано текущее состояние рынка облачных услуг, демонстрирующее высокие темпы роста, рассмотрены их возможности и ключевые преимущества, что позволило сделать вывод об их востребованности и актуальности. В п.1.1.2 приведены определения понятий облачных вычислений и облачных услуг согласно международным стандартам, в частности, Рекомендаций МСЭ-Т и NIST, а также данных производителей облачных платформ. В п. 1.1.3 дана классификация и модели развертывания облачных услуг, в 1.1.4 рассмотрены облачные платформы, их компоненты и принципы работы.

В п. 1.2.1 – 1.2.4 рассмотрены основные компоненты услуги «виртуальный рабочий стол»: технологии, протоколы и программное обеспечение, совместное функционирование которых позволяет предоставлять данную услугу пользователям, находящимся как в локальной, так и в глобальной сети.

П. 1.3.1 посвящен анализу подходов к определению качества облачных услуг в целом, показаны критерии, опираясь на которые можно формулировать требования к качеству услуги «виртуальный рабочий стол». В п. 1.3.2 определены параметры, определяющие качество рассматриваемой услуги. Составлена классификация параметров QoS, которые приведены для основных этапов работы услуги. Классификация представлена в виде матрицы, в столбцах которой находятся критерии QoS, в строках – этапы работы услуги. На основании деятельности, выполняемой пользователями в своих виртуальных рабочих столах, проведена классификация пользователей услуги.

Раздел завершается постановкой задач диссертационного исследования (п. 1.4).

Раздел 2. Исследование параметров, влияющих на качество услуги «виртуальный рабочий стол»

2.1 Анализ параметров, определяющих качество услуги

2.1.1 Общие подходы к определению параметров качества

Основной задачей провайдера развивающихся облачных услуг, доставляемых пользователю посредствам IP сетей, в частности услуги «виртуальный рабочий стол», является обеспечение надлежащего качества. Для этого требуется детальное знание требований к характеристикам элементов инфраструктуры услуги и особенностям их функционирования и взаимодействия.

Рядовой пользователь услуги не озабочен тем, как услуга реализуется технически, отправной точкой для формулирования требований к качеству должно быть удовлетворение пользователя, которое является субъективной величиной [37]. При этом приемлемость может зависеть от ожиданий пользователей и контекста [72, 73]. В этой связи воспринимаемое качество определяется как совокупная мера удовлетворенности пользователя, работающего с какой-либо услугой или мультимедиа.

Требования к качеству рассматриваемой услуги можно формулировать, опираясь на следующие критерии:

- Быстрота отклика, отсутствие зависаний, подергиваний изображения.
- Ожидание подключения к услуге не должно быть длительным.

- Время отклика должно быть регламентировано.
- Впечатление от работы с услугой должно описываться пользователем как «хорошее», «приемлемое».

Требования к QoS конечных пользователей для различных приложений раскрываются в Рекомендации G.1010 МСЭ-Т [70], данные которой могут использоваться в качестве отправной точки при проектировании инфраструктуры услуг, в частности, услуги «виртуальный рабочий стол». В ней приведены следующие ключевые параметры, затрагивающие пользователя: задержка (Delay), вариация задержки (Delay Variation), потери информации (Information loss). Рассмотрим их подробнее.

Задержка может быть определена различными способами, однако в общем виде можно привести следующую формулировку: время, проходящее с момента отправки пользователем запроса на получение некоторой информации до момента ее получения. Применительно к рассматриваемой услуге наиболее подходящим термином, выражающим задержку, можно считать время отклика, состоящее из времен обслуживания в узлах инфраструктуры, транспортной задержки, времени обработки информации на пользовательском устройстве. Более детально эти временные параметры раскрываются далее в п. 2.1.2.

Вариация задержки пакетов часто учитывается среди параметров качества ввиду того, что для некоторых услуг при определенных условиях, возникающих в пакетных сетях, может наблюдаться явление изменчивости времени прибытия пакетов [101]. Тем не менее, многие услуги, в особенности, чувствительные к вариации задержки, имеют специальные механизмы борьбы с этим явлением, например, основанные на буферизации пакетов на приемной стороне или на внесении определенной фиксированной задержки. Различные механизмы борьбы с ней в обязательном порядке реализованы в системах доставки удаленного рабочего стола всех протоколов виртуализации.

Потери информации (доля потерь пакетов) также могут учитываться среди параметров качества и могут возникать, в частности, в результате отбрасывания пакетов при переполнении буфера на сервере.

Кроме того, в [70] приведена таблица, описывающая допустимые значения рассматриваемых параметров. Таблица разделена на три категории – аудио, видео, данные в соответствии со спецификой пользовательской работы. Каждая категория имеет несколько подкатегорий. Услуга «виртуальный рабочий стол» может быть причислена к подкатегориям «Command/Control» (величина допустимой временной задержки менее 250 мс, величина вариации задержки пакетов – N.A.) и «Transaction services» (величина допустимой временной задержки менее 2 с, величина вариации задержки пакетов – N.A.).

Согласно [70] параметры услуги, касающиеся пользователя, должны соответствовать следующим критериям:

- Принимать во внимание все аспекты услуги с точки зрения пользователя.
- Не должны зависеть от конкретной архитектуры сети или технологии.
- Могут быть объективно или субъективно измерены в точке доступа к услуге.
- Могут быть легко связаны с сетевыми параметрами.
- Могут быть гарантированы пользователю поставщиком услуги.

Одним из важнейших параметров, определяющих качество рассматриваемой услуги, является время отклика, кроме него целесообразно рассмотреть и другие, в частности, вероятность отказа в подключении к услуге, которые будут рассмотрены далее. Время отклика удовлетворяет приведенным выше критериям, значит выполнение требований по нему является необходимым условием выполнения требований обеспечения приемлемости качества.

2.1.2 Анализ времени отклика

Услуга «виртуальный рабочий стол» является облачной услугой и предполагает работу в реальном времени, следовательно, ключевым фактором, определяющим пользовательское удовлетворение, является суммарная задержка, которая складывается из нескольких составляющих: время, проходящее с момента первоначального запроса пользователя до момента начала предоставления услуги, а также время, проходящее до момента получения пользователем информации после того, как услуга активирована. Задержка имеет самое непосредственное влияние на удовлетворенность пользователя в зависимости от используемого приложения или услуги, и включает в себя задержки в работе терминала, сети и сервера услуги [39, 40].

Суммарное время отклика, т.е. время, проходящее с момента установления услуги до момента начала ее предоставления для услуги «виртуальный рабочий стол» согласно [77], показано на рисунке 2.1. Ниже будет показано и обоснованно, что это время различно для процесса подключения к услуге и для процесса работы услуги.

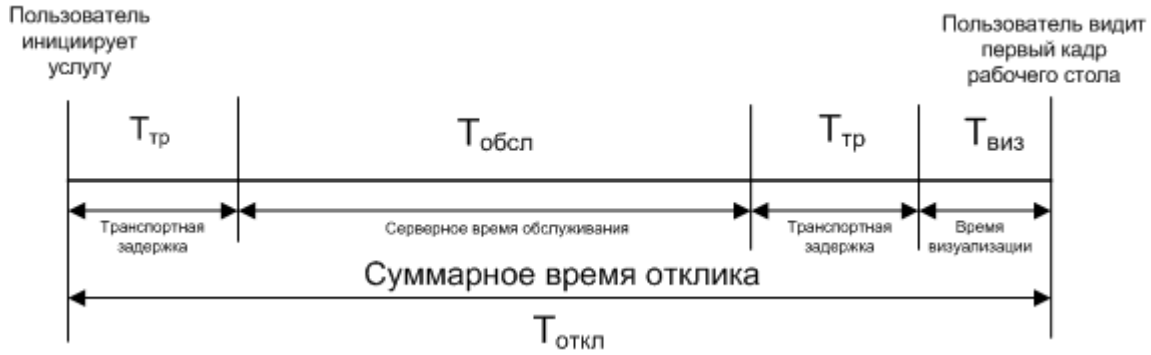

Рис. 2.1. Суммарное время отклика для услуги DaaS согласно рекомендации Y.3503 МСЭ-Т

Рассмотрим приведенные на рисунке 2.1 временные интервалы в отдельности.

T – время отклика, то есть суммарное время взаимодействия сервера и клиента.

$T_{обсл}$ – время, проходящее с момента получения менеджером подключений заявки на формирование снимка рабочего стола до момента отправки этого сообщения, запакованного в пакет, пользователю в процессе функционирования терминальной сессии. Иными словами, это серверное время обслуживания.

$T_{тр}$ – транспортная задержка пакетов в сети. Согласно Рекомендации G.1010 МСЭ-Т и RFC2679 транспортная задержка определяется как односторонняя сетевая задержка [70, 100]. Однако, при описании работы услуги «виртуальный рабочий стол» необходимо учитывать двустороннюю задержку. При дальнейших расчетах делается допущение о симметрии задержки, будем брать ее удвоенное значение. В случаях, когда это не так, вместо удвоенной односторонней задержки следует использовать значение Round Trip Time Delay.

$T_{виз}$ – время визуализации (время обработки на пользовательском устройстве), т.е. время, необходимое для распаковывания пакета, декодирования и отрисовки терминальным оборудованием изображения рабочего стола.

Следует отметить, что среднее время до подключения, описанное в п. 1.3, представляет собой среднее время отклика для фазы установления терминальной сессии.

Проанализировав процесс функционирования рассматриваемой облачной услуги, можно сделать вывод о том, что он имеет две фазы: в первой фазе пользователи подключаются к услуге, происходит их авторизация и аутентификация; во второй фазе подключенные пользователи работают в терминальной сессии со своим виртуальным рабочим столом.

За работу фаз услуги отвечают две подсистемы, логически выделенные из блоков архитектуры услуги [42]. За работу первой фазы услуги отвечают: менеджер подключений, агент пользовательского устройства, служба каталогов (Active Directory). За работу второй фазы отвечают: агент виртуальной машины, агент пользовательского устройства. Фазы предоставления услуги показаны на рисунке 2.2.

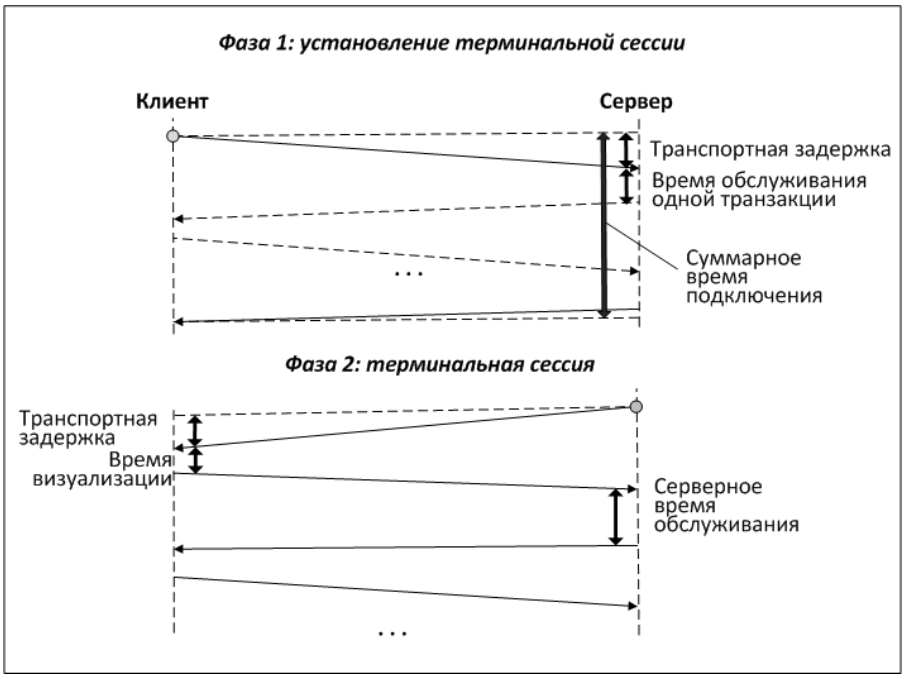

Рис. 2.2. Фазы предоставления услуги «виртуальный рабочий стол»

Таким образом, с учетом сказанного, для фазы установления терминальной сессии справедлива формула (2.1), а для фазы терминальной сессии – формула (2.2):

$$T^{(1)} = T^{(1)}_{\text{обсл}} + 2T_{\text{тр}} . \quad (2.1)$$

$$T^{(2)} = T^{(2)}_{\text{обсл}} + 2T_{\text{тр}} + T^{(2)}_{\text{виз}} . \quad (2.2)$$

Для фазы установления терминальной сессии величина $T^{(1)}_{виз}$ отсутствует, поскольку в процессе установления сессии нет процедуры обработки изображений удаленного рабочего стола, имеется лишь обмен служебными данными для установления соединения.

Поясним приведенные рассуждения на диаграмме сетевого взаимодействия между компонентами архитектуры услуги типа «виртуальный рабочий стол», которая показана на рисунке 2.3.

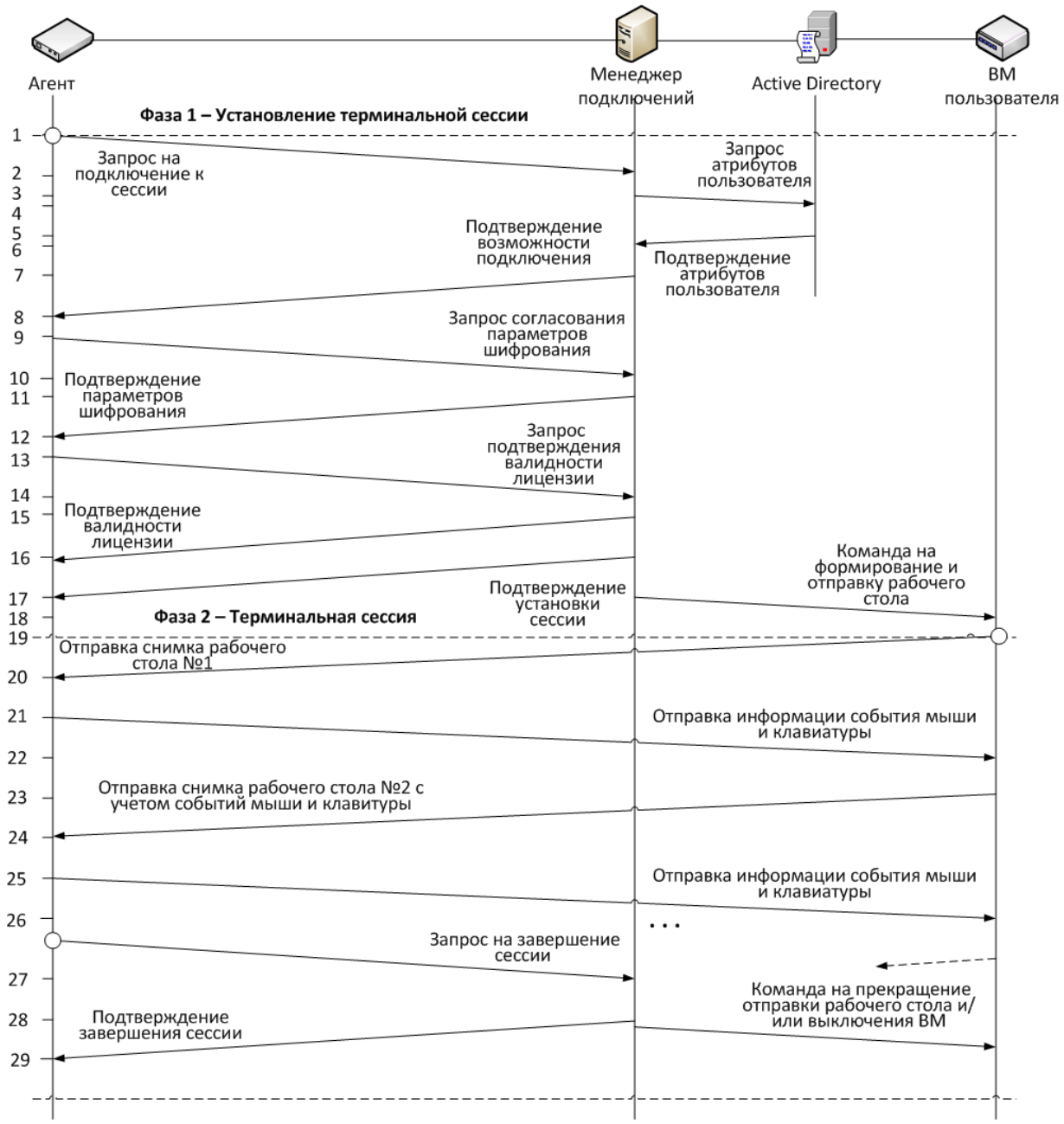

Рис. 2.3. Сетевое взаимодействие между компонентами услуги «виртуальный рабочий стол»

На данной диаграмме цифрами показано поэтапное разбиение процесса работы услуги на интервалы времени, проходящие с момента установления терминальной сессии до момента ее завершения.

Интервалы 1, 7, 9, 11, 13, 15, 16 соответствуют транспортной задержке на участке сети «Агент – Менеджер подключений».

Интервалы 19, 21, 23, 25 соответствуют транспортной задержке на участке сети «Агент – ВМ пользователя». Будем полагать, что величины транспортной задержки на данных участках одинаковы согласно принципу организации инфраструктуры рассматриваемой услуги.

Интервалы 3, 5, 17 соответствуют времени обмена командами программных компонентов архитектуры услуги. Этими интервалы не вносят вклад в суммарное время отклика. Интервалы 2, 4, 6, 8, 10, 14, 17, 20, 22, 24, 28 соответствуют серверному времени обработки.

2.1.3 Анализ транспортной задержки

Рассмотрим данные по транспортным задержкам, приведенные в различных рекомендациях МСЭ-Т.

В Рекомендации Y.1541 МСЭ-Т [75] приведена таблица для сетевых параметров в зависимости от классов QoS (0-5). Так, для класса 0 верхняя граница IPTD составляет 100 мс, для классов 1, 2, 3, 4 – 400 мс. Определение IPTD (IP packet transfer delay) дано в Рекомендации Y.1540 МСЭ-Т [74], согласно которой параметр IPTD определяется как интервал времени $t_2 - t_1$ между двумя событиями – вводом пакета во входную точку сети в момент t_1 и выводом пакета из выходной точки сети в момент t_2, где $(t_2 > t_1)$ и $(t_2 - t_1) \leq T_{max}$.

В Рекомендации МСЭ-Т G.1010 [70], как было сказано выше, приведена таблица сетевых характеристик для различных приложений. В первой фазе работы услуги она может быть причислена к категории «Данные» (подкатегория «Transaction services»), величина допустимой задержки для которой должна быть менее 2 с. Это значение обозначает наихудший вариант, при котором работа все еще возможна. Анализ и соотнесение реально используемых на практике приложений с категориями были даны в [34, 35].

Для более детального анализа перейдем далее к рассмотрению данных, полученных в ходе исследований. Статистический метод определения качества восприятия пользователем работы с услугой типа DaaS, определенный в экспериментальном исследовании, рассмотрен в [64]. Для этого был собран фрагмент сети «сервер – маршрутизатор – модем – тонкий клиент», в который был помещен исследовательский зонд, реализованный в виде ПО на рабочей станции. Результаты оценки разрабатываемого метода были соотнесены с результатами, полученными при оценке параметров QoE группой добровольцев. Итогами исследования стала разработка алгоритма, позволяющего с определенной точностью выявить тип приложения, запущенного пользователем в данный момент времени, а затем, зная тип приложения (аудио, видео, данные и т.д.), применить к нему статистическую оценку показателей QoE. Однако в этой работе рассмотрен только один из возможных протоколов виртуализации – Microsoft Remote Desktop Protocol (RDP) [29]. При помощи опроса целевой группы добровольцев были получены результаты для комфортного времени отклика сервера в терминах RTT (Round Trip Time).

Исследователи добавляли в сеть задержку и фиксировали реакцию пользователей на ее увеличение. При этом предлагали пользователю три вида действий за рабочим столом: прослушивание аудио, обычную работу с окнами, просмотр видео. Для работы с данными «удовлетворительным» приводится значение RTT < 400 мс, «хорошим» (приемлемым) – значение RTT < 100 мс (см. таблицу 2.1).

Таблица 2.1. Значения RTT при подключении по протоколу RDP

Оценка пользователями	1	2-3	4-5
Категория	«Плохо»	«Удовлетворительно»	«Хорошо»
Аудио	RTT≥450 мс	120 мс <RTT< 450 мс	RTT≤120 мс
Данные	RTT≥400 мс	100 мс <RTT< 400 мс	RTT≤100 мс
Видео	RTT≥70 мс	50 мс <RTT< 70 мс	RTT≤50 мс

Заметим, что RTT равно удвоенному значению односторонней транспортной задержки согласно определению RTT, как времени, которое проходит с момента, когда пакет вышел в сеть до момента, когда придет ответ о его получении.

В [89] авторами было проведено экспериментальное исследование, рассматривающее работу нескольких протоколов доставки удаленного рабочего стола в локальных сетях, были проведены измерения канальной скорости передачи данных при работе различных протоколов, а также величины транспортной задержки, диапазон величин которой составила 150 мс..1 с. Следует заметить, что данное исследование затрагивает устаревшие протоколы.

В [95] экспериментально исследуется влияние задержек в WAN-сетях на работу систем сервер – тонкий клиент. Авторами было проведено сравнение ряда протоколов по показателям вносимых задержек и занятия полосы пропускания при передаче различного трафика. Из результатов данной работы следует, что критически важным параметром для обеспечения приемлемого QoE являются задержки в глобальной и локальной сетях, что подтверждает сказанное выше. В данной работе рассмотрены несколько протоколов доставки удаленного рабочего стола, интерес представляют наиболее современные из них – ICA, RDP, VNC. Измерялась величина задержки при выполнении нескольких видов деятельности пользователя: написание письма (текст), пролистывание (скроллинг) страниц, выделение мышью различных областей экрана, операция загрузки изображения в формате JPEG. Заметим, что последняя из перечисленных операция отличается по специфике от предыдущих четырех, поскольку не предусматривает изменений изображения на рабочем столе (которое приводит к необходимости постоянного обновления изображения на серверной стороне и отсылки обновленного). В рассматриваем контексте следует обратить внимание на действие прокручивания (скроллинга),

поскольку это действие вызывает необходимость постоянной передачи нового изображения. Для этого действия задержка составила 150..250 мс, а время отклика составило 400..600 мс.

В [62] рассматривается подход к решению задачи уменьшения промежутков между пакетами в клиент-серверной среде, однако этот подход применяется к системе на основе архитектуры X11, которая на сегодняшний день мало востребована в качестве основной архитектуры облачной среды. Кроме того, работа посвящена проблематике передачи по сети мультимедиа и игровых приложений, в то время как пользователи услуги «виртуальный рабочий стол», в основном, работают с офисными приложениями.

В работе [114] авторы экспериментально исследовали влияние транспортных задержек на работу протокола VNC (Virtual Network Computing [99]). Для этого ими был собран стенд, состоящий из сервера, программного эмулятора транспортных задержек и тонкого клиента. Итогом этого исследования стали выводы о том, что при скоростях канала между сервером и клиентом 10 и 100 Мбит/с задержки менее 150 мс почти не сказываются на восприятии рабочего стола (могут наблюдаться незначительные запаздывания визуального ряда), транспортные задержки 150 мс – 1 с являются терпимыми (наблюдаются различимые пользователю затормаживания), транспортные задержки более 1 с являются крайне неприятными (сильное затормаживание визуального ряда), транспортные задержки от 2 до 5 с являются критичными (наблюдается полное рассыпание изображения, приводящее к остановке работы). Этот вывод согласуется с предыдущими исследованиями, в частности с исследованием другого протокола (RDP) в работе [64].

В работе [58] авторами представлена собственная концепция построения архитектуры удаленного рабочего стола под названием THINC. Отличительной особенностью этой архитектуры является способ доставки видео до клиентского устройства во время терминальной сессии, то есть во время использования услуги «виртуальный рабочий стол». Предлагается детектирование факта запуска видео пользователем и дальнейшая передача видео отдельным от потока изображений видео потоком. Для синхронизации видео потока с потоком изображений рабочего стола вводится специальный набор служебных команд, осуществляющих согласование с пользовательским устройством типа кодека, процессорного ресурса для обработки этого потока, а также команд управления потоком (пауза, остановка, перемотка, изменение битрейта, размера и т.д.) Также применяется поддержка YUV формата пикселей (YUV – цветовая модель, в которой цвет представляется в виде трех компонент - яркость (Y) и две цветоразностных (U и V), при этом конвертация из RGB формата может быть осуществлена по специальным формулам). Преимущества применения YUV модели заключается в том, что ее обработка занимает меньше ресурса пользовательского устройства, что позволяет использовать освободившийся ресурс для ускорения обработки видео потока. Эта

концепция развивалась в [63], где, в частности, приведено сравнение данного подхода с другими протоколами доставки рабочего стола. Применение этой и прочих техник позволяют оптимизировать процесс доставки удаленного рабочего стола, что должно положительно сказаться на пользовательском восприятии. Однако данная концепция направлена по большей части на оптимизацию доставки видео и не нашла широкого применения на практике.

В работе [109] также рассматривается проблема оптимизации доставки видео в рамках терминальной сессии. Авторы представляют свою методику организации терминальной сессии, в основу которой положен принцип разделения потока изображений рабочего стола и видео потока, если пользователь запускает видео во время работы с рабочим столом. При этом сам видеопоток в зависимости от характера видео может делится на два подпотока, которые обрабатываются различными кодеками: протокол RFB [30] обрабатывает «медленные видео» (т.е. потоки с медленно изменяющимся видеорядом), H.264 обрабатывает «быстрые видео» (т.е. потоки с быстро меняющимся видеорядом).

RFB (Remote Frame Protocol) – протокол удаленного доступа к графическому интерфейсу пользовательского компьютера. На основе этого протокола построена система удаленного доступа к графической среде компьютера VNC (Virtual Network Computing) [99]. Среда VNC применяется для удаленного мониторинга рабочих станций и их администрирования, а не как средство доставки виртуально рабочего стола в рамках терминальной сессии услуги «виртуальный рабочий стол».

Данная работа также направлена в основном на оптимизацию доставки видео в рамках терминальной сессии, а также не рассматривает нашедшие применение на сегодняшний день протоколы.

Величина задержки из Рекомендации G.1010 очерчивает верхний предел для задержки, однако специфика услуги «виртуальный рабочий стол» более чувствительна к задержкам. Поэтому исследования [64, 89, 114] сужают круг требований до меньших величин, поскольку затрагивают непосредственно рассматриваемую услугу.

На основании перечисленных исследований восприятия качества работы с услугой при различных задержках, а также на основании данных из рекомендаций МСЭ-Т (к сетям передачи данных по задержкам (Y.1541) и требованиям к категориям QoS из конца в конец (G.1010)), можно оценить диапазон приемлемой транспортной задержки, составляющий 120..150 мс, который используем далее при оценке времени отклика.

Провайдер сети передачи данных может декларировать соответствие сети тому или иному классу QoS, например, классу 1. Значит, задержка будет удовлетворять вышеуказанным требованиям.

2.2 Исследование влияния сетевых и серверных параметров на функционирование услуги

2.2.1 Исследование скорости передачи данных при работе современных протоколов доставки виртуального рабочего стола

Скорость передачи данных является параметром NP, на которых базируются параметры качества. При неудовлетворении этим параметрам невозможно обеспечить требования по качеству.

Для исследования скорости передачи данных при работе современных протоколов доставки виртуального рабочего стола, в частности, для оценки ее минимально допустимых и комфортных значений при работе с удаленным сервером услуги осуществим экспериментальное исследование.

Исследование разделено на два этапа. На первом этапе целью является измерение загруженности канала при работе VDI в «нормальных» стандартных условиях, а именно: облачная платформа развернута в локальной сети, на участке сервер – коммутатор – тонкий клиент используется Fast Ethernet (100 Мбит/с), клиент использует один монитор. На втором этапе целью является измерение загруженности канала при внесении различных величин сетевых задержек [36].

Стенд состоит из «Облака», коммутатора доступа, тонкого клиента и измерительного ПК с установленной ОС FreeBSD [32, 65]. «Облако» включает в себя 4 различных платформы виртуализации (Citrix XenServer 6.2, VMware ESXi 5.1, Microsoft Hyper-V, Red Hat RHEV), которые запускаются поочередно), репозиторий виртуальных машин, сервер авторизации, сервер групповых политик [102].

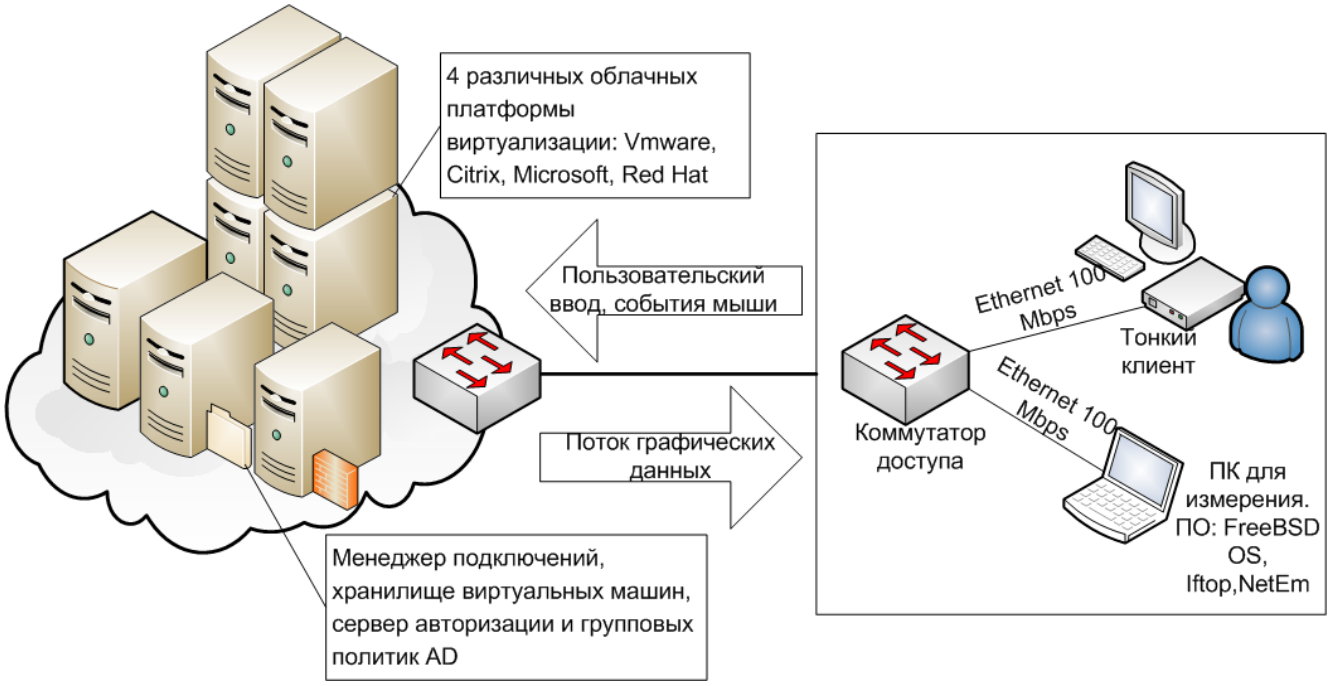

Рис. 2.4. Тестовый стенд

Измерения производились программой Iftop [65]. Результаты представлены в таблице П.1 Приложения 1 и на рисунке 2.5. В таблице третий столбец «Бездействие» означает включенную сессию клиент – сервер без каких - либо действий со стороны клиента; столбец «Работа» означает движение курсора, перемещение окон, открытие и закрытие файлов; столбец «Web-серфинг» означает открытие и закрытие вкладок в браузере, скроллинг страниц, заполненных изображениями; столбец «Просмотр видео в браузере» означает просмотр видео с популярного сайта http://youtube.com; столбец «RemoteFX» означает использование аппаратного ускорения обработки видео (только для Microsoft HyperV), видео имеет качество 1024x768 на экране с разрешением 1366x768; столбец «Версия» введен потому, что производители платформ виртуализации вносят коррективы в механизмы работы своих протоколов, что прямым образом сказывается на качестве.

Следует заметить, что различные платформы, предоставляющие облачные услуги, обладают разной производительностью, поскольку каждая такая платформа работает на основе своих собственных протоколов. Как было сказано ранее, наибольшее распространение получили Microsoft Remote Desktop Protocol (RDP), Citrix High Definition user eXperience (HDX), VMware Personal Computer over IP (PCoIP), Red Hat SPICE. Эти протоколы обладают различными механизмами управления доставкой приложений поверх протокола транспортного уровня TCP/UDP, различными характеристиками и способами организации соединения.

Поэтому при оценке параметров качества необходимо рассматривать этот вопрос отдельно для каждой конкретной платформы виртуализации, которая развернута в облаке.

При работе с облачными услугами пользователи выполняют действия в режиме реального времени через протокол соединения тонкого клиента. Это может быть отправка электронной почты, Веб-серфинг, просмотр видео и флэш-контента и т.п. Получается, что утилизация полосы канала является важным аспектом для определения качества, воспринимаемого конечными пользователями. От того, какой работой занимается пользователь, зависит, насколько сильно будет нагружен канал: видео-потоки имеют наиболее высокое потребление полосы пропускания, особенно, если это потоки HD-качества. Менее всего ресурсов потребуют, к примеру, операции с текстовыми документами. Следует, однако, учитывать, что даже если пользователь не запускает никаких приложений, отрисовка рабочего стола, по сути, есть не что иное, как непрерывный поток изображений того, что отображается на рабочем столе в каждый момент времени, даже если это статичная картинка.

Из диаграммы, показанной на рисунке 2.5, следует, что при идеальных сетевых условиях, т.е. когда в сети присутствует только один клиент и один сервер услуги, скорость при выполнении стандартной работы офисного сотрудника может быть достигать 50 Мбит/с, а в реальных условиях этот показатель может оказаться выше.

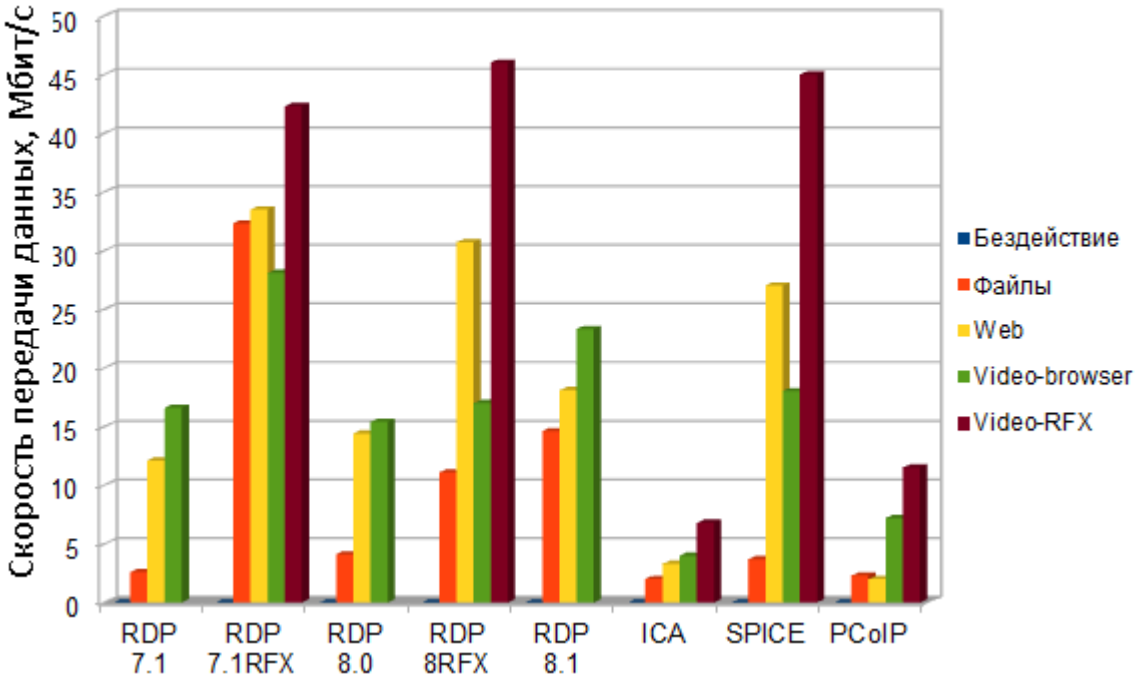

Рис. 2.5. Исследование канальной скорости передачи данных для различных протоколов виртуализации

В [115] приведено исследование загруженности канала при различных сценариях работы пользователей и различных сетевых условиях. Его результаты показаны в таблице 2.2.

Таблица 2.2. Исследование загрузки канала

	Использование видео		Использование канала	
Тип пользователя	Редко	Часто	Минимальная скорость передачи данных	Минимальная скорость передачи данных, необходимая для комфортной работы
Простые задачи			<70 кбит/с	500 кбит/с
			<100 кбит/с	1 Мбит/с
Базовая офисная работа			<150 кбит/с	750 кбит/с
Активная офисная работа			<250 кбит/с	1-3 Мбит/с
	✓		<600 кбит/с	5 Мбит/с
Продвинутая офисная работа	✓		<1.25 Мбит/с	7 Мбит/с
	✓		<2.5 Мбит/с	>10 Мбит/с
Базовая работа с CAD (WAN)	✓		>1 кбит/с	>2 Мбит/с
Базовая работа с CAD (LAN)	✓		>3 кбит/с	>10 Мбит/с
Запуск видео		✓	>7 Мбит/с	>30 Мбит/с
Запуск тяжелого видео		✓	>30 Мбит/с	>50 Мбит/с
Продвинутая работа с графикой и CAD		✓	>30 Мбит/с	>70 Мбит/с
Игры		✓	>80 Мбит/с	>120 Мбит/с

Из проведенного исследования, а также данных, представленных в [115], видно, что скорость передачи данных 100 Мбит/с является недостаточной на сетевом интерфейсе сервера услуги на сети с множеством клиентов для обеспечения их одновременной работы с заданным уровнем качества. Значит, должны применяться различные способы ее увеличения, например, использование серверных сетевых карт, поддерживающих большие скорости передачи данных или применения технологий распределения сетевой нагрузки пользователей на интерфейсы сервера и т.д.

На сегодняшний день скорость передачи данных на сетевом интерфейсе сервера должна составлять минимум 1 Гбит/с в режиме подключения нескольких пользователей услуги. Расчет максимального количества пользователей при заданных условиях будет произведен в разделе 3.

2.2.2 Исследование зависимости транспортной задержки от скорости передачи данных

Услуга «виртуальный рабочим стол», будучи предоставляемой из облака, по своей сути является услугой реального времени (online), следовательно, критически важными параметрами, определяющими качество данной услуги, являются временные задержки и

величина скорости передачи данных. Увеличение величин транспортной задержки и уменьшение скорости передачи данных приводят к ощутимому дискомфорту в работе пользователя: изображение запаздывает, наблюдаются зависания и т.д.

Исследование влияния транспортной задержки на качество услуги было проведено в рамках второго этапа эксперимента, изложенного в 2.2.1. Была установлена терминальная сессия при помощи платформы производства компании Microsoft [102]. Поочередно были инициированы три вида деятельности пользователя за рабочим столом: работа с файлами (оконный менеджер, файловый менеджер); веб-серфинг; просмотр видео качеством 720p. В сеть вносились различные величины сетевой задержки при помощи программы ipfw, входящей в состав ОС FreeBSD, развернутой на измерительном ПК [65]. Измерялась скорость передачи данных на уровне TCP. Результаты измерений сведены в таблицу П.2 Приложения 1. График зависимости транспортной задержки от скорости передачи данных по показан на рисунке 2.6.

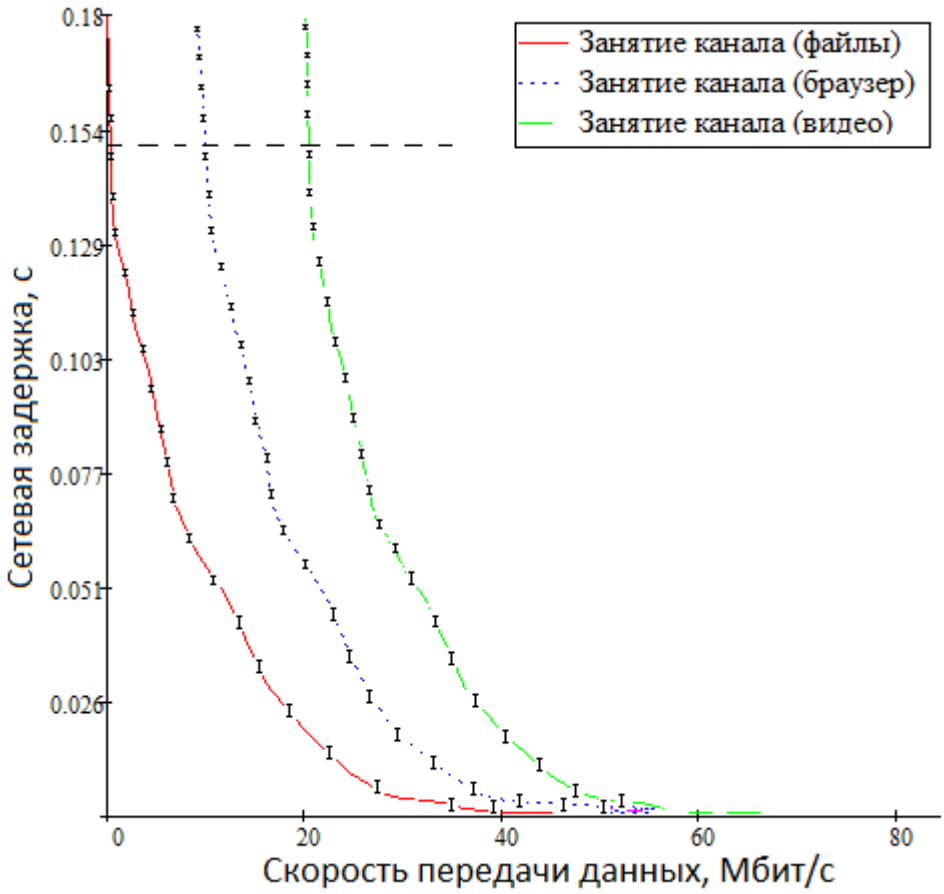

Рис. 2.6. Зависимость транспортной задержки от скорости передачи данных

На экспериментальных кривых показан доверительный интервал, построенный при доверительной вероятности 0.95. Горизонтальной пунктирной линией отмечен уровень 150 мс, описанный в п.2.3.

Результаты исследования подтверждают и дополняют результат эксперимента, показанного в 2.2.1. Полученные результаты можно использовать в качестве некоторой оценки, справедливой для условий эксперимента. Так, для обеспечения приемлемой задержки (на уровне 150 мс), скорость передачи данных должна быть не менее 5 Мбит/с при работе с файлами; не менее 15 Мбит/с при работе с браузером; не менее 25 Мбит/с при работе с видео.

Полученные данные могут быть использованы провайдером услуги, администратором сети на этапе проектирования и эксплуатации облачной и сетевой инфраструктуры в качестве опорных значений.

2.2.3 Исследование временных характеристик услуги

Для исследования временных характеристик услуги в обоих фазах ее работы было собран экспериментальный стенд. Исследуемые параметры: среднее время отклика в фазе 1, средний промежуток времени между соседними требованиями в фазе 1, среднее время отклика в фазе 2, среднее время обслуживания запросов сервером в фазе 2, среднее время обработки терминальным устройством в фазе 2.

В состав стенда входят: серверная платформа (Citrix XenServer 6.2 [22]), менеджер подключений (Citrix VDI-IN-A-BOX [33]), тонкий клиент (TC-20 [32]), сетевые кабели. Скорость на сетевых интерфейсах сервера и клиента – 1 Гбит/с.

Для определения временных параметров в фазе 1 осуществляется многократное подключение терминального устройства к серверу услуги.

Для определения временных параметров в фазе 2 рассмотрена система со ста пользователями. При помощи средств отладки облачной платформы Citrix собрана статистика по временным параметрам. Проведены следующие исследования.

а) Определение среднего времени отклика на этапе установления терминальной сессии. Осуществлялось многократное подключение клиента к серверу с сопутствующим измерением временных показателей: времени, проходящего с момента начала процесса установления до момента появления на экране первого изображения (время отклика). Целью эксперимента являлась экспериментальное исследование величины времени отклика, которое совестно с теоретическими сведениями [17, 103] было использовано для оценки данного параметра. Проведенное исследование позволило оценить величину среднего времени отклика во время установления терминальной сессии (подключение к услуге). Его значение составило 1.5 с.

Построен доверительный интервал, используя ф-лу Стьюдента, с доверительной вероятностью 0.95. Доверительный интервал составил [1.436..1.562].

б) Исследование среднего промежутка времени между соседними требованиями на этапе установления терминальной сессии. Исследован промежуток времени между соседними требованиями потока, поступающего от одного пользователя $\bar{t}$, который понадобится для оценки величины интенсивности потока запросов от одного пользователя λ_0. В результате измерений средний промежуток времени между соседними требованиями составил 0.2 с. Построен доверительный интервал, используя ф-лу Стьюдента, с доверительной вероятностью 0.95. Доверительный интервал составил [0.18..0.22].

в) Исследование среднего промежутка времени между соседними требованиями на этапе терминальной сессии. В результате значение $\overline{T_o}$ составило 0.4 с. Построен доверительный интервал, используя ф-лу Стьюдента, с доверительной вероятностью 0.95. Доверительный интервал составил [0.33..0.46].

г) Определение среднего времени отклика на этапе терминальной сессии. Измерялось время, проходящее с момента отправки с пользовательского устройства на сервер запроса, кодирующего событие движения и нажатия клавиш мыши или клавиатуры, до момента появления на экране пользователя изображения рабочего стола, содержащего реакцию на произведенное действие. В рамках исследуемого вопроса было рассмотрено сто таких временных интервалов. Среднее значение времени отклика составило 1.1 с. Построен доверительный интервал, используя ф-лу Стьюдента, с доверительной вероятностью 0.95. Доверительный интервал составил [0.97..1.18].

д) Определение среднего времени обслуживания запросов сервером на этапе терминальной сессии. Измерялось время, проходящее с момента получения сервером запроса, кодирующего событие движения и нажатия клавиш мыши или клавиатуры, до момента отправки обработанного запроса с сервера. В рамках исследуемого вопроса было рассмотрено сто таких временных интервалов. Среднее значение времени обслуживания составило 0.59 с. Построен доверительный интервал, используя ф-лу Стьюдента, с доверительной вероятностью 0.95. Доверительный интервал составил [0.42..0.76].

е) Определение среднего времени обработки на терминальном устройстве (время визуализации) на этапе терминальной сессии. Измерялось время, проходящее с момента получения устройством сообщения с сервера, содержащего изображение рабочего стола, до момента отрисовки картинки пользователю. В рамках исследуемого вопроса было рассмотрено сто таких временных интервалов. Среднее значение времени обслуживания составило 0.21 с.

Построен доверительный интервал, используя ф-лу Стьюдента, с доверительной вероятностью 0.95. Доверительный интервал составил [0.17..0.31].

Границы доверительных интервалов в проведенных исследованиях отстоят от среднего значения менее чем на 10%, значит объем выборки достаточен для оценки [21].

Проведенные исследования позволили оценить временные характеристики услуги как на этапе установления, так и на этапе работы терминальной сессии. Полученные значения будут использованы в разделе 3 для при аналитическом моделировании.

2.3 Выводы по результатам второго раздела

Во втором разделе были проведены анализ и исследование параметров, совместное влияние которых определяет качество услуги. В п. 2.1.1 рассмотрены общие подходы к определению параметров качества услуг, выявлены критерии, опираясь на которые можно формулировать требования к качеству услуги «виртуальный рабочий стол». Кроме того, в число параметров, влияющих на качество услуги, входят: транспортная задержка, канальная скорость передачи данных, количество обслуживаемых пользователей, серверное время обслуживания, время визуализации. В п. 2.1.2 показано, что ключевым является среднее время отклика. Предложено разделение функционирования услуги на две фазы для возможности отдельного их исследования. Построены диаграммы сетевого взаимодействия компонентов инфраструктуры услуги в обеих фазах.

В п. 2.1.3 проанализированы исследования, экспериментально оценивающие временные характеристики услуги «виртуальный рабочий стол». В результате, на основании данных из Рекомендаций МСЭ-Т и ряда исследований различных авторов, определен диапазон величин транспортной задержки, при котором работа с услугой является приемлемой.

В п. 2.2.1 в результате проведенного экспериментального исследования загрузки полосы канала наиболее распространенными на сегодняшний день протоколами доставки рабочего стола, получена сравнительная характеристика, анализ которой позволил сделать вывод о минимально допустимой скорости сетевого интерфейса сервера услуги для рассмотренных условий. Это исследование дополнено другим, описанным в п. 2.2.2, в результате которого получена зависимость транспортной задержки от скорости передачи данных, которое выявило допустимые величины скорости, при которых уровень транспортной задержки сохраняется в пределах допустимого. Исследование проведено для трех наиболее типичных видов пользовательской деятельности в рамках рабочего стола.

Проведенные в п. 2.2.3 исследования временных характеристик услуги (среднего времени отклика на этапе установления терминальной сессии, среднего промежутка времени между соседними требованиями на этапе установления терминальной сессии, среднего промежутка времени между соседними требованиями на этапе терминальной сессии, среднего времени отклика на этапе терминальной сессии, среднего времени обслуживания запросов сервером на этапе терминальной сессии, среднего времени визуализации на этапе терминальной сессии) позволили получить ряд значений, которые использованы в дальнейших расчетах.

Раздел 3. Математическое моделирование фазы установления терминальной сессии услуги «виртуальный рабочий стол»

3.1 Построение аналитической модели

3.1.1 Введение и постановка задачи

Процесс работы инфокоммуникационной услуги «виртуальный рабочий стол» в разделе 2 был логически разделен на две фазы для возможности их отдельного исследования В первой фазе происходит подключение пользователей к услуге (установление терминальной сессии), во второй – подкаченные пользователи работают в рамках виртуального рабочего стола (терминальная сессия). Для исследования первой фазы необходимо построить аналитическую модель первой фазы работы услуги, получить и проанализировать вероятностно-временные характеристики, оценить среднее время отклика. На основе полученной модели решить задачу поиска множества допустимых значений характеристик сервера услуги, при которых выполняются ограничения по среднему времени отклика и вероятности отказа; и задачу определения вариантов сочетания этих характеристик.

Согласно специфике услуги «виртуальный рабочий стол», в первой фазе работы сервер должен обслуживать всех пользователей, подключающихся к услуге, одновременно. На основании принципа работы облачной платформы услуги «виртуальный рабочий стол», изложенного в разделе 1, механизм работы менеджера подключений уместно описать системой массового обслуживания вида M/G/1/K с дисциплиной обслуживания разделение процессора (processor sharing, PS). В этой дисциплине заявки обслуживаются одновременно со скоростью обратно пропорциональной их числу в системе.

Модель M/G/1 широко используется при моделировании компьютерных систем. Система M/G/1/K с дисциплиной FIFO рассмотрена в [13]. Вопросам, связанным с системой M/G/1, посвящена глава 5 в [14]. Системы M/G/1/K с дисциплиной PS рассмотрены в [55, 56, 119, 120], а также в [86]. В [24] показано, что в определенных случаях дисциплина PS является оптимальной.

Начать построение аналитической модели следует с описания параметров услуги в терминах теории массового обслуживания. Далее введем обозначения, сформулируем основные предположения и допущения.

На основании того, что услуга рассчитана на подключение большого числа пользователей, входящий поток заявок примем пуассоновским, поскольку имеет место суперпозиция множества потоков. Обозначим λ_0 – интенсивность потока заявок от одного пользователя, N – количество всех пользователей услуги. Тогда интенсивность общего потока $\lambda = \lambda_0 N$.

Под заявками будем понимать пакеты, поступающие от клиента к серверу, которые содержат запросы на подключение к терминальной сессии. Под обслуживанием будем понимать их обработку сервером.

Положим, что число заявок в системе ограничено величиной K. Новая заявка, поступившая в момент, когда число заявок в системе равно K, отбрасывается. Время обслуживания одной заявки в системе имеет произвольное распределение со средним b_1. Обозначим p_n вероятность того, что в данный момент обслуживаются n заявок (n = 0,..,K). Подобную СМО можно обозначить M/G/1/K*PS [41].

Величины, которые требуется определить:

p_K – вероятность блокировки из-за заполнения всех мест в системе;

$T_{обсл}$ – среднее время обслуживания на сервере;

T – среднее время отклика.

3.1.2 Оценка времени отклика в фазе установления терминальной сессии

Время ответа классической компьютерной системы рассматривалось в [17, 103]. Архитектура подобных систем подразумевает взаимодействие компьютера с пользователем, при котором изображение, предназначенное для показа пользователю, выводится непосредственно на монитор через системную шину. В случае облачной услуги «виртуальный рабочий стол» доставка рабочего стола происходит по сети передачи данных, следовательно, основным параметром, определяющим качество услуги, будет являться среднее время отклика $T^{(1)}$, складывающееся, как было показано в разделе 2, из нескольких компонент: времени обслуживания сервера $T_{обсл}$, транспортной задержки в сети передачи данных $T_{тр}$.

Время ответа для классических компьютерных систем, рассматриваемое в [17, 103], описывает процесс, который содержит одну транзакцию данных от машины до экрана. Однако процесс установления терминальной сессии услуги подразумевает несколько транзакций обмена служебными данными между сервером и пользовательским устройством, которые необходимы для подключения пользователя к серверу [44, 45]. Следовательно, требуется ввести

некоторые уточнения. Пусть $m_{ср}$ – среднее число запросов в процессе обмена пакетами, необходимыми для организации сессии, $T_{откл1}$ – среднее время отклика для одного запроса [44, 45]. Тогда среднее время до установления сессии будет равно $T_{откл1} \cdot m_{ср}$, откуда вытекает требование ко времени отклика на один запрос: $T_{откл1} \le T^{(1)} / m_{ср}$.

Для того, чтобы пользователи описывали работу, как «приемлемую», среднее время отклика для классической компьютерной системы не должно превышать 2 с [17, 103]. Учитывая наличие нескольких транзакций в процессе установления сессии услуги «виртуальный рабочий стол», значение $m_{ср}$ примем равным 5, что является типичным, исходя из опыта работы с услугой. Например, для протокола RDP, в зависимости от его версии, значение $m_{ср}$ может равняться от 4 до 7, для протокола SPICE – 5 [29, 105]. Таким образом, $T_{откл1} \le 2/5\ c = 0.4\,c$, откуда по формуле (2.1) $T_{обсл} = T_{откл1} - 2 \cdot T_{тр} = 0.4\ \text{с} - 2 \cdot 0.150\ \text{с} = 0.1\ \text{с}$.

3.1.3 Получение вероятностно-временных характеристик

Согласно [50, 97] стационарное распределение числа заявок в рассматриваемой СМО имеет вид:

$$p_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}},\ \text{где}\ \rho = \lambda b_1. \tag{3.1}$$

Поскольку система имеет ограничение по количеству обслуживаемых заявок, при любом значении ρ существует стационарный режим. При ρ = 1 формула (3.1) дает неопределенность вида 0/0, поэтому этот случай необходимо рассмотреть отдельно:

$$p_n = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^{K+1}}, \text{если } \rho \ne 1; \\ \dfrac{1}{K+1}, \text{если } \rho = 1. \end{cases} \tag{3.2}$$

Исходя из формулы (3.2) и того факта, что приходящая заявка, застав в системе K заявок, отбрасывается, можно найти вероятность блокировки:

$$p_K = \begin{cases} \dfrac{(1-\rho)\rho^K}{1-\rho^{K+1}}, \text{если } \rho \ne 1; \\ \dfrac{1}{K+1}, \text{если } \rho = 1. \end{cases} \tag{3.3}$$

Для среднего числа заявок в системе L справедлива формула

$$L = \sum_{n=1}^{K} n \cdot p_n . \tag{3.4}$$

Подставив в выражение (3.4) p_n из формулы (3.2), получим:

$$L = \begin{cases} \dfrac{\rho \cdot \left[1 - (K+1)\rho^K + K \cdot \rho^{K+1}\right]}{(1-\rho^{K+1}) \cdot (1-\rho)}, \text{ если } \rho \neq 1; \\ \dfrac{K}{2}, \text{ если } \rho = 1. \end{cases} \tag{3.5}$$

По формуле Литтла для систем с потерями [14]

$$T_{обсл} = \frac{L}{\lambda \cdot (1 - p_K)} \quad ,$$

откуда, используя выражения для L из формулы (3.5), получим:

$$T_{обсл} = \begin{cases} \dfrac{\rho^{K+1}(K \cdot \rho - K - 1) + \rho}{\lambda \cdot (1-\rho^K) \cdot (1-\rho)}, \text{ если } \rho \neq 1; \\ \dfrac{K(1-\rho^{K+1})}{2\lambda \cdot (1-\rho^K)}, \text{ если } \rho = 1. \end{cases} \tag{3.6}$$

Учитывая, что $\rho = \lambda b_1$, получим:

$$T_{обсл} = \begin{cases} \dfrac{(\lambda \cdot b_1)^{K+1}(K \cdot \lambda \cdot b_1 - K - 1) + \lambda \cdot b_1}{\lambda \cdot \left[1 - (\lambda \cdot b_1)^K \cdot (1 - \lambda \cdot b_1)\right]}, \text{ если } \rho \neq 1; \\ \dfrac{K(1-(\lambda \cdot b_1)^{K+1})}{2\lambda \cdot (1-(\lambda \cdot b_1)^K)}, \text{ если } \rho = 1. \end{cases} \tag{3.7}$$

Эта формула позволяет определить среднее время пребывания заявки в системе, которое является одним из важнейших параметров при определении воспринимаемого пользователем качества услуги, поскольку оно является одной из составляющих суммарного времени отклика.

Учитывая это, можно, приравняв в (3.7) $T_{обсл}$ к значению 0.1 с, найденному выше, решить одним из численных методов (например, методом Ньютона) получившееся уравнение относительно b_1. Далее представлены результаты подобных расчетов.

Графики зависимости среднего времени обслуживания заявки от интенсивности потока заявок показаны на рисунке 3.1.

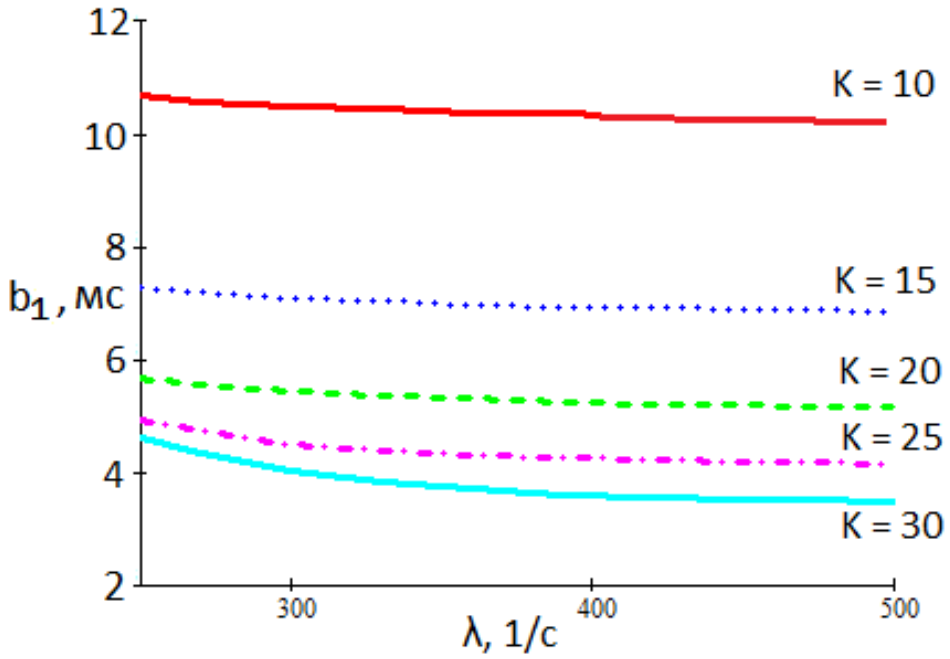

Рис. 3.1. Зависимость среднего времени обслуживания заявки от интенсивности

Естественно, что увеличение интенсивности входящего потока вынуждает систему обслуживать каждую заявку быстрее для сохранения комфортного времени отклика для каждого пользователя, однако при больших значениях интенсивности входящего потока (т.е. при числе пользователей n в системе порядка 100) уменьшение времени обслуживания каждой заявки слабо выражено.

Далее используем значение параметра λ_0, полученное из соотношения $\lambda_0 = 1/\bar{t}$, где $\bar{t}$ – средний промежуток времени между соседними требованиями от одного пользователя , который оценивался экспериментальным путем в п. 2.2.3, где рассматривалась система, имеющая 100 пользователей (это количество пользователей весьма типично для систем, используемых в практике применения услуги). В результате измерений средний промежуток времени между соседними требованиями составил 0.2 с. Таким образом, $\lambda_0 = 5$ с$^{-1}$, откуда $\lambda = 5 \cdot 100 = 500$ с$^{-1}$.

Зависимость среднего времени отклика от среднего времени обслуживания одной заявки при различных значениях K и $\lambda = 500$ с$^{-1}$ показана на рисунке 3.2.

Эта зависимость позволяет найти пороговое значение b_1, которое бы обеспечивало комфортную работу множества пользователей без потери качества с услуги с точки зрения времени отклика системы.

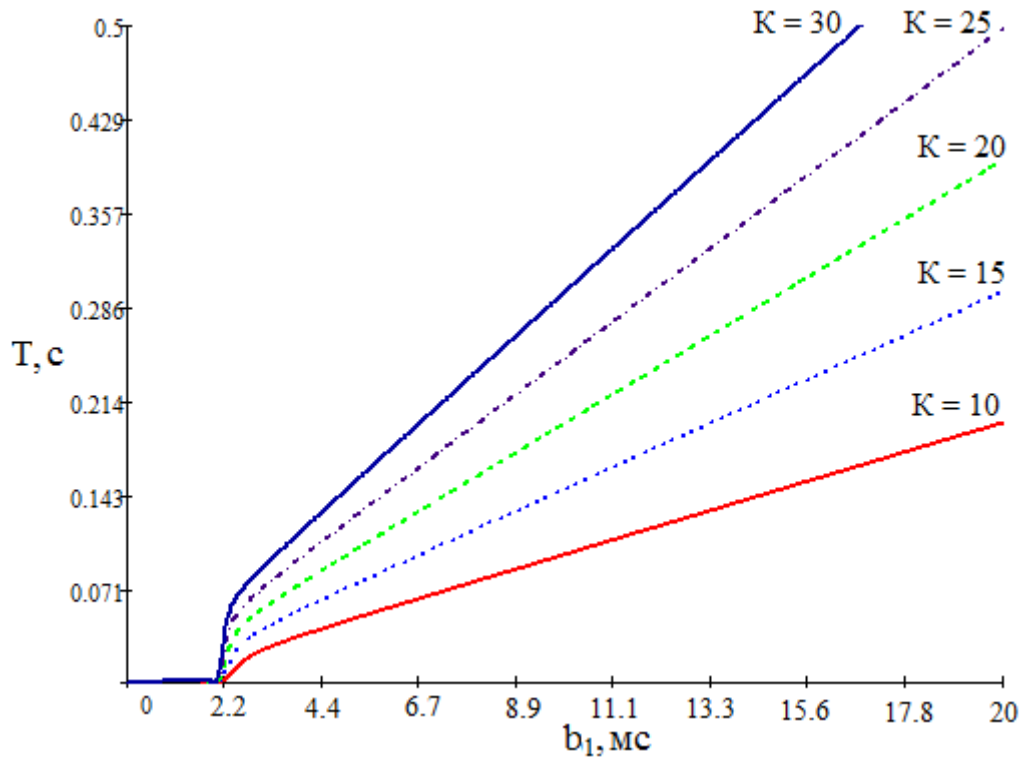

Рис. 3.2. Соотношения между временными характеристиками системы

Зависимость между средним временем отклика и числом одновременно обслуживаемых пользователей в фазе установления терминальной сессии показана на рисунке 3.3.

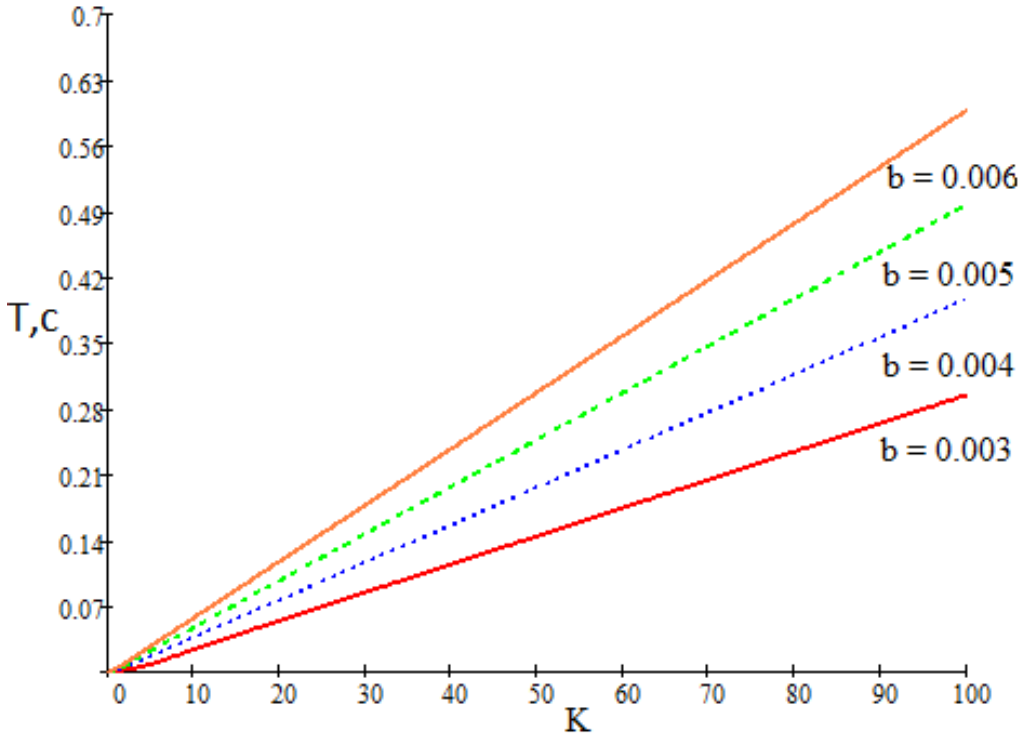

Рис. 3.3. Зависимость среднего времени отклика от максимального числа

3.2 Поиск множества значений параметров услуги, отвечающих заданным требованиям к качеству

3.2.1 Поиск множества допустимых значений параметров услуги

Одной из важных и распространенных задач на практике является расчет параметров, определяющих качество услуги, при которых бы удовлетворялись требования, предъявляемые к качеству. В случае рассматриваемой услуги (ее первой фазы) требования могут предъявляться к среднему времени отклика (T), вероятности отказа (p_K) при некоторой заданной интенсивности входящего потока. Требуется найти такие значения b_1 и K, при которых выполняются ограничения по времени отклика и вероятности отказа.

Провайдеру необходимо знать минимальные требования к серверу (время обслуживания одного запроса, количество подключаемых пользователей услуги), при которых будут выполняться заданные ограничения. Способами уменьшить значение b_1 являются: установка более производительного процессора, использование аппаратного ускорения, задействование графического процессора и др.; увеличить значение K возможно за счет увеличения ресурсов менеджера подключений (оперативная память, быстродействие виртуального процессора). Тогда становится возможным предоставление услуги с приемлемым качеством достаточно большому количеству пользователей (исходя из практического опыта – до 200, хотя на практике редко встречаются серверы, обслуживающие более 150 пользователей).

По условиям задачи нужно, чтобы: $p_K(K, b_1) \le p_0$, $T(K, b_1) \le T_0$, где p_0 и T_0 – требуемые значения.

Требования могут выполняться при различных сочетаниях K и b_1, которые образуют множество допустимых значений D:

$$D = \{(K, b_1): p_K(K, b_1) \le p_0 \text{ и } T(K, b_1) \le T_0\}.$$

Заметим, что если увеличивается b_1, то увеличивается T и увеличивается p_K; если увеличивается K, то увеличивается T и уменьшается p_K. Поэтому для построения множества D будем поступать следующим образом. При фиксированном b_1 определим граничные значения K, при которых выполняются заданные ограничения:

$$r_1(b_1) = \min \{K: p_K(K, b_1) \le p_0 \text{ и } T(K, b_1) \le T_0\},$$
$$r_2(b_1) = \max \{K: p_K(K, b_1) \le p_0 \text{ и } T(K, b_1) \le T_0\}.$$

Тогда $r_1(b_1)$ и $r_2(b_1)$ – есть границы множества *D*. Для расчета p_K используем формулу (3.3), для расчета *T* – формулу (3.7). Поскольку *K* принимает целочисленные значения, множество *D* есть множество отрезков, параллельных оси абсцисс, заключенных между $r_1(b_1)$ и $r_2(b_1)$. На рисунке 3.4 оси абсцисс соответствуют значения b_1, оси ординат – значения *K*.

Заданным условиям можно удовлетворить только сочетанием параметров *K* и b_1, которое показано в виде найденного множества их допустимых значений *D* (см. рисунок 3.4). Из рисунка 3.4 видно, что существует максимальное b_1, при котором возможно удовлетворить заданным условиям. Отсюда вытекает ограничение на b_1. Например, при *T* = 0.2 с, К = 70, $p_K = 10^{-2}$, $b_1 \leq 2.8$ мс.

3.2.2 Поиск множества рационального сочетания параметров услуги

Задачей каждого провайдера услуг является при минимальных затратах удовлетворение заданным требованиям. В данном случае минимальные затраты означают применение более дешевого сервера, дешевизна которого заключается в использовании меньших ресурсов, таких, как ресурс оперативной памяти, процессора, памяти на жестких дисках.

Однозначно определить соотношения параметров сервера, отвечающих желанию удовлетворить требования при минимальных затратах, невозможно: нет такого максимального b_1 и минимального *K* – увеличение одного приводит к уменьшению другого. Однако, существует набор таких значений *K* и b_1, находясь в рамках которого все еще можно удовлетворять заданным требованиям. Математически уместно описать подобную ситуацию при помощи множества Парето.

Множеством Парето принято считать такое множество состояний или параметров системы, в котором значение того или иного параметра не может быть улучшено без ухудшения других [21]. В рассматриваемой ситуации множество Парето представляет собой нижнюю границу множества *D*, т.е. множество $П = \{(K, b_1) \in D: (K - 1, b_1) \notin D\}$. Для построения множества Парето воспользуемся методом, описанном в [52]. Точки множества *П* обозначены на рисунке 3.5 сплошной линией.

Рассмотрим численный пример, где в качестве исходных возьмем данные полученные ранее. Пусть $T_0 \leq 2$ с – допустимое время отклика, $p_K = 10^{-2}$ – допустимая вероятность отказа, $\lambda = 500$ 1/с.

Построим множество П по формулам (3.3) и (3.7), используя исходные параметры из примера (см. рисунок 3.4).

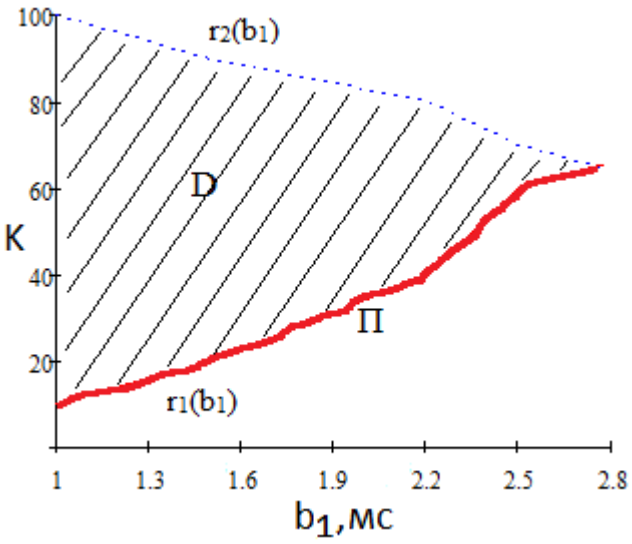

Рис. 3.4. Множество допустимых значений и множество Парето

Окончательный выбор числа обслуживаемых пользователей и времени обслуживания одной заявки производится по экономическим критериям. При сравнении вариантов достаточно ограничиться рассмотрением не всего множества их допустимых значений, а меньшего по числу элементов множества Парето. Полученные соотношения позволяют, исходя из заданных требований к качеству, рассчитать время обслуживания и количество одновременно обслуживаемых пользователей услуги.

3.3 Выводы по результатам третьего раздела

Третий раздел посвящен аналитическому моделированию первой фазы предоставления услуги «виртуальный рабочий стол». В п. 3.1.1 приведено обоснование подбора СМО, наиболее адекватно описывающей работу услуги в первой ее фазе, введены основные положения и допущения, сформулирована постановка задач, решаемых в разделе.

В п. 3.1.2 оценено среднее время отклика в первой фазе. Рассмотрено время отклика для классических компьютерных систем, с учетом специфики работы рассматриваемой облачной услуги внесены поправки (транзакции в процессе скрытого от пользователя обмена служебными данными в процессе подключения к услуге). При помощи формулы для среднего времени отклика, полученной в разделе 2, рассчитано среднее время обслуживания на сервере.

В п. 3.1.3 рассчитаны вероятностно-временные характеристики и получены ключевые зависимости: зависимость среднего времени отклика от среднего времени обслуживания одной

заявки при различных значениях одновременно обслуживаемых пользователей и зависимость среднего времени отклика сервера от числа обслуживаемых заявок в системе.

П. 3.2.1 посвящен решению задачи поиска множества допустимых значений времени обслуживания одной заявки b_1 и числа одновременно обслуживаемых пользователей K, при которых выполняются ограничения по времени отклика и вероятности отказа. Заданным условиям можно удовлетворить только сочетанием параметров K и b_1, которое показано в виде найденного множества их допустимых значений D.

В п. 3.2.2. виде множества Парето представлены рациональные варианты сочетания этих параметров. Полученные результаты могут быть полезны провайдером услуги для обеспечения комфортного для пользователей времени отклика и вероятности отказа. Это может быть достигнуто путем изменения таких параметров, как число одновременно обслуживаемых пользователей (K) и среднее время обслуживания одной заявки (b_1), что может осуществляться за счет увеличения программно-аппаратного ресурса на сервере услуги.

Результаты, полученные в результате анализа построенной в разделе модели, будут использованы далее при формулировании рекомендаций по обеспечению приемлемого качества услуги на этапе ее установления.

Раздел 4. Математическое моделирование фазы работы терминальной сессии услуги «виртуальный рабочий стол»

4.1 Построение аналитических моделей

4.1.1 Введение и постановка задачи

Изначально услуга «виртуальный рабочий стол» проектировалась для решения задачи предоставления пользователю только рабочего стола. Целевой аудиторией услуги были офисные работники, по специфике свой деятельности, работающие только с офисными приложениями. Такой сценарий работы услуги будем назвать базовым. Затем, со временем, по мере накопления опыта эксплуатации услуги, повышения ее привлекательности для более широкого круга пользователей, появились новые сценарии работы: сценарий, при котором пользователю предоставлена возможность просмотра видео, запускаемого внутри рабочего стола и сценарий, при котором предоставлена возможность запуска и видео и аудио.

Таким образом, на основе специфики работы пользователей уместно выделить следующие основные сценарии работы услуги:

1) базовый;

2) сценарий с видео;

3) сценарий с видео и аудио.

Специфика работы услуги в этих сценариях различна, что потребует построения отдельных математических моделей. Таким образом, можно сформулировать постановку задач исследования: составить аналитические модели, описывающие работу пользовательского устройства и сервера в трех сценариях работы, проанализировать временные характеристики этих моделей. Оценить среднее время отклика для каждого сценария.

4.1.2 Аналитическая модель терминальной сессии в базовом сценарии работы

Проанализируем процесс работы терминальной сессии. После того, как пользователь авторизован и подключен к услуге, начинается интерактивный обмен данными между агентом пользовательского устройства и агентом виртуальной машины через сеть передачи данных.

Клиент отправляет команды событий мыши и клавиатуры на сервер, в ответ на них сервер отправляет поток данных клиенту. Более подробно этот процесс был рассмотрен в разделе 1.

Большинство современных пользовательских устройств (в особенности – тонких клиентов, предназначенных специально для услуги «виртуальный рабочий стол») имеют архитектуру с многоядерными процессорами. Операционная система устройства распределяет обрабатываемую информацию между процессорами для увеличения быстродействия обработки данных. На практике большое распространение получили тонкие клиенты, имеющие двухъядерные процессоры.

Схема работы услуги в базовом сценарии работы показана на рисунке 4.1.

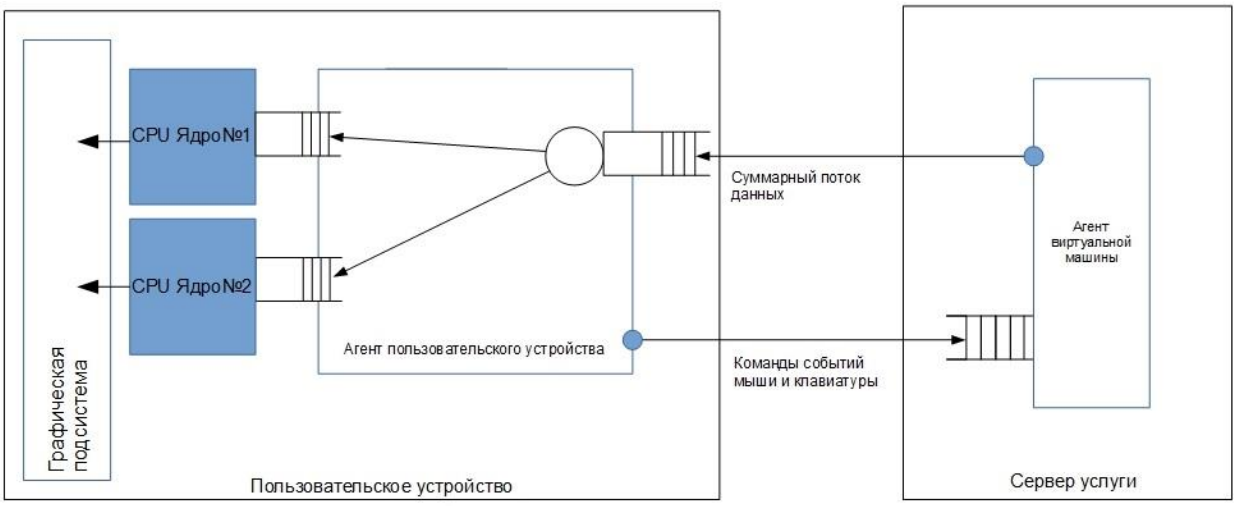

Рис. 4.1. Схема инфраструктуры услуги в базовом сценарии работы

Агент пользовательского устройства распределяет однотипные заявки по ядрам процессора поочередно: одна заявка на ядро №1, вторая – на ядро № 2.

Такая схема реализации использования процессорного ресурса (ядер) подразумевает распределение заявок равномерно на оба ядра. Подобная модель является адекватным описанием работы услуги в базовом сценарии [47]. Процесс работы услуги «виртуальный рабочий стол» в фазе 2 уместно описать сетью массового обслуживания (СеМО), состоящей из четырех СМО: VM — агент ВМ; А — агент пользовательского устройства; С1 и С2 — ядра процессора пользовательского устройства.

Обмен данными в этом случае описывается следующей последовательностью.

1. Агент пользовательского устройства (А) инициирует начало передачи данных с сервера т. е. отправляет поток заявок, который поступает на вход агента ВМ (VM). Поступающие агенту ВМ заявки, обнаружив, что прибор занят, становятся в очередь в буфер.

2. Агент ВМ (VM) обрабатывает поступившие заявки, затем отправляет агенту пользовательского устройства поток данных, содержащие изображения рабочего стола.

3. Агент пользовательского устройства (А) распределяет однотипные заявки по ядрам процессора пользовательского устройства (С1 и С2). Поступающие агенту заявки, обнаружив, что прибор занят, становятся в очередь в буфер.

4. Ядра С1, С2 обрабатывают поступившие заявки. После обработки ядра отправляют обслуженные потоки на графическую подсистему, которая отрисовывает картинку пользователю. Поступающие ядрам заявки, обнаружив, что приборы заняты, становятся в очередь в буферы.

Введем следующие обозначения.

p_{ac1} – вероятность перехода заявки в ядро №1; p_{ac2} – вероятность перехода заявки в ядро №2; λ_1 – интенсивность входящего в узел С1 потока; λ_2 – интенсивность входящего в узел С2 потока; μ_1 – интенсивность обслуживания узла VM; μ_2 – интенсивность обслуживания узла А; μ_3 – интенсивность обслуживания узла С1; μ_4 – интенсивность обслуживания узла С2.

Пусть время обслуживания каждой заявки на любом из приборов сети является случайной величиной (СВ), не зависящей от состояния сети и ее предыстории, а также не зависящей от времени обслуживания этой заявки в других узлах сети. Будем полагать, что эти СВ распределены по экспоненциальному закону с параметром μ_i, $i = \overline{0,M}$. После обслуживания в узле i с вероятностью p_{ij} заявка переходит в узел j. С вероятностью $p_{вых} = 1 - \sum_{i=1}^{M} p_{ij}$ заявка покидает сеть. В каждом узле заявка обслуживается единожды, то есть сеть можно отнести к ациклическим.

Подобные системы, расположенные в узлах сети, могут быть описаны СМО вида M/M/1. Такая СМО часто используется для моделирования различных реальных систем и приборов, поскольку позволяет получить аналитические выражения для многих параметров.

Исходя из логики работы услуги во второй ее фазе (последовательная передача данных от сервера к клиенту), дисциплину обслуживания в очередях всех узлов сети примем FCFS. В рамках модели будем считать все буферы бесконечными.

Показанная на рисунке 4.2 сеть массового обслуживания представляет собой совокупность конечного числа взаимосвязанных узлов обслуживания, в которой циркулируют заявки, переходящие в соответствии с маршрутной матрицей с выхода одного узла на вход другого. Каждый отдельный узел является однолинейной СМО и отображает самостоятельную часть реальной системы.

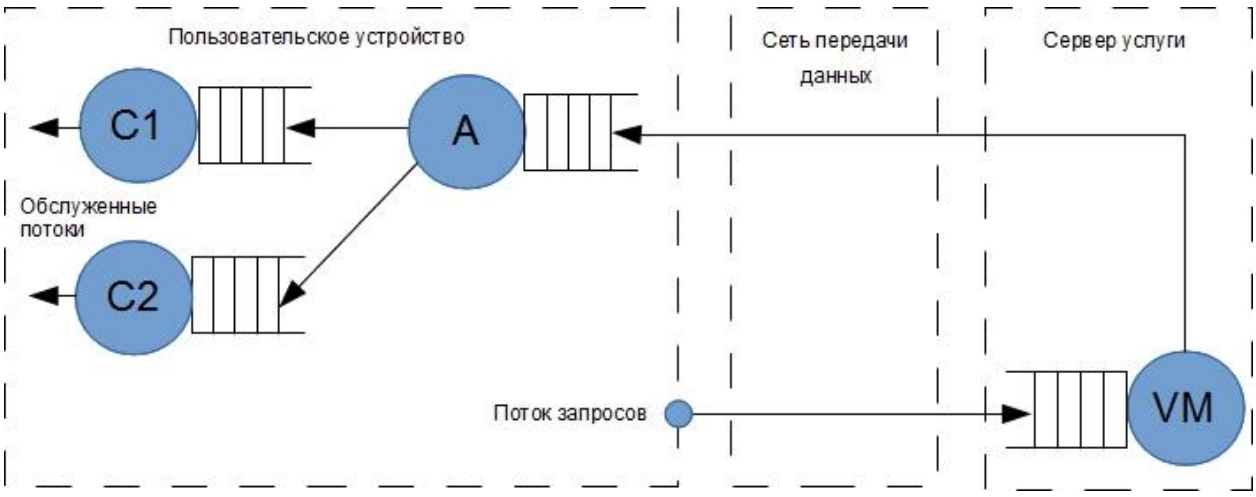

Рис. 4.2. Схема СеМО

Под входным потоком некоторого узла будем понимать поток заявок, приходящих на вход этого узла из предыдущего узла или из внешней среды (для узла VM).

Поток заявок по своей природе является неравномерным во времени: работа пользователя, своей активностью (клавиатура, мышь) генерирующего поток заявок, неравномерна: в какие-то моменты времени пользователь работает (печатает, двигает мышь), а в какие-то моменты читает с экрана. К тому же в течении дня работа чередуется с пассивностью, а после окончания рабочего дня вообще прекращается.

Кроме того, особенность работы протоколов виртуализации такова, что при отсутствии активности пользователя, по сети передаются только вспомогательные данные протокола виртуализации, а при ее наличии происходит передача пользовательских данных. Таким образом, для моделирования услуги нет необходимости рассматривать заявки за весь период, когда устройство включено, достаточно рассматривать лишь период работы пользователя. Будем полагать, что в эти периоды поток заявок можно считать пуассоновским.

Согласно теореме Берка [61] поток на выходе системы M/M/1 тоже является простейшим. Это означает, что при последовательном включении нескольких СМО, их можно рассматривать как независимые СМО.

Таким образом, подобная СеМО удовлетворяет теореме Джексона [14] об открытых сетях и может быть моделирована сетью Джексона. При этом оценку, полученную в результате моделирования, можно считать оценкой сверху, иными словами, модель будет описывать наихудший вариант характеристик работы услуги.

Опишем потоки, циркулирующие в сети. Обозначим λ_{ij} интенсивность потока заявок, следующих из узла i в узел j ($i, j = \overline{0, M}$, $\lambda_{00} = 0$).

Интенсивность суммарного поступающего в узел i потока: $\lambda_{j,i} = \sum_{j=0}^{M} \lambda_{ji}$.

Интенсивность суммарного выходящего из узла i потока: $\lambda_{i,j} = \sum_{j=0}^{M} \lambda_{ij}$.

Приведем далее аналитические соотношения для характеристик сети. Будем полагать, что режим работы сети – стационарный. Следовательно, интенсивности поступающего в узел *i* потока и выходящего из него потока должны совпадать. Система уравнений равновесия согласно [3] будет иметь вид:

$$\lambda_j = \sum_{j=0}^{M} \lambda_i \cdot p_{ij}, j = \overline{0, M}. \quad (4.1)$$

Среднее число заявок в узле *i* обозначим N_i. Его можно определить следующим образом согласно [116]:

$$\overline{N_i} = \frac{\rho_i}{1-\rho_i} \ . \quad (4.2)$$

Среднее время пребывания заявки в узле T_i можно найти из формулы Литтла:

$$\lambda_i \cdot T_i = N_i \ . \quad (4.3)$$

Откуда $T_i = \frac{N_i}{\lambda_i} = \frac{1}{\lambda_i} \cdot \frac{\rho_i}{1-\rho_i} = \frac{1}{1-\rho_i} \cdot \frac{1}{\mu_i} = \frac{1}{\mu_i - \lambda_i}$, где $\rho_i = \frac{\lambda_i}{\mu_i}$. (4.4)

Среднее время ожидания в узле *i*: $W_i = T_i - \frac{1}{\mu_i} = \frac{\rho_i}{1-\rho_i} \cdot \frac{1}{\mu_i}$. (4.5)

Суммарное среднее число заявок во всей сети можно найти из формулы:

$$N = \sum_{i-1}^{M} N_i \ . \quad (4.6)$$

Суммарное среднее время пребывания заявки во всей сети (время отклика) *T* можно выразить из:

$$\lambda_0' \cdot T = N \ . \quad (4.7)$$

Откуда согласно [116]: $$T = \frac{N}{\lambda_0'} = \frac{1}{\lambda_0'} \sum_{i=1}^{M} \frac{\lambda_i}{\mu_i - \lambda_i}. \quad (4.8)$$

Под временем отклика будем понимать среднее время, проходящее с момента вхождения заявки в систему до момента ее выхода обслуженной. Именно это время является один из наиболее важных параметров услуги, поскольку напрямую влияет на воспринимаемое пользователем качество услуги.

Для наглядного представления СеМО построим граф, вершины которого соответствуют отдельным узлам сети, а ребра – связям между узлами. Переход заявок между узлами происходит мгновенно в соответствии с переходными вероятностями, которые обозначают вероятность того, что заявка после обслуживания в узле *i* перейдет в узел *j*. Если узлы *i* и *j* непосредственно между собой не связаны, то $p_{ij} = 0$. Если из узла *i* возможен переход только в узел *j*, то $p_{ij} = 1$. Граф, образованный на основе рассматриваемой СеМО, показан на рисунке 4.3.

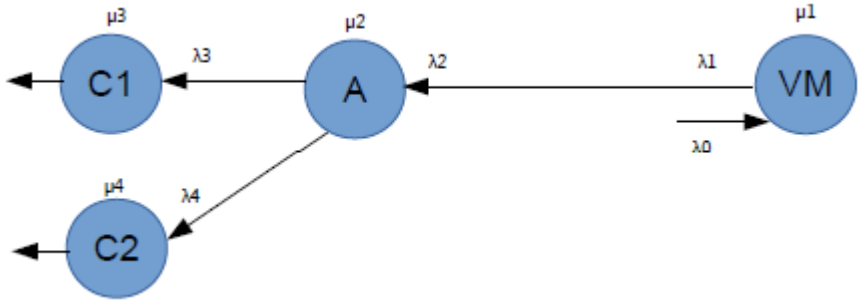

Рис. 4.3. Граф СеМО

Обозначим $P = \|p_{ij}\|$ матрицу переходов (маршрутную матрицу):

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & p_{ac1} & p_{ac2} \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Величина λ'_0 оценивалась из соотношения $\lambda'_0 = \frac{1}{\overline{T_0}}$, где величина $\overline{T_0}$ оценена экспериментально в параграфе 2.2.3 и составила 0.4 с. Таким образом, $\lambda'_0 =$ 2.5 c^{-1}. Из описанного эксперимента значение T_i составило 0.6 с. Тогда из формулы (4.4) получим: $\mu_1 = 10.5\, c^{-1}$.

Воспользуемся полученными результатами для расчета времени отклика без учета транспортной задержки T'. Дано: μ_1 = 10.5 c^{-1}, $\lambda'_0 = 2.5$ c^{-1}, $p_{ac1} = p_{ac2} = 0.5$. Найти: T', оценить T. Решение:

1. Из свойств сети Джексона в узле VM значение $\lambda_i = \lambda_0$.

2. Проверим стационарность: $\frac{2.5}{10.5} = 0.328 \prec 1$.

3. Рассчитаем λ_1 и λ_2: $\lambda_1 = \lambda_2 = p_{ac1} \cdot \lambda_1 = p_{ac2} \cdot \lambda_2 = 0.5 \cdot 2.5 = 1.25$ c^{-1}.

4. Рассчитаем T':

$$T' = \frac{1}{\lambda_0'} \left[\frac{\lambda_i}{\mu_1 \quad \lambda_i} + \frac{\lambda_i}{\mu_2 \quad \lambda_i} + \frac{\lambda_1}{\mu_3 \quad \lambda_1} + \frac{\lambda_2}{\mu_4 \quad \lambda_2} \right] = 0.373\,\text{с}.$$

В рамках описываемой модели процессоры устройства полагали одинаковыми, поэтому интенсивности обслуживания в ядрах процессора (μ_3, μ_4) также полагаем одинаковыми (столбцы 2 и 3 таблицы П.3 Приложения 1).

Для дальнейших расчетов будем поступать следующим образом. Задавая различные значения μ_2, μ_3, μ_4, моделируем различные ситуации: когда узел А недостаточно производителен для комфортной работы с услугой (бюджетное устройство с низкой производительностью), когда узлы С1, С2 недостаточно производительны для комфортной работы (устройство с бюджетным маломощным процессором), смешанный вариант. Для этого:

1) фиксируем интенсивность обслуживания узла А (μ_2), увеличиваем интенсивности в узлах С1 (μ_3), С2 (μ_4);

2) увеличиваем μ_2, фиксируем ее, и снова увеличиваем μ_3, μ_4;

3) повторяем, снова увеличив μ_2.

4) повторяем еще раз, увеличив μ_2.

Моделируем ситуацию маломощных процессоров: увеличиваем μ_2, при этом μ_3, μ_4 варьируем от 5 до 15 $с^{-1}$. Результаты расчетов сведем в таблицу П.3 Приложения 1.

Результаты произведенных расчетов показывают, что если процессоры маломощные (интенсивность обслуживания от 5 до 15 1/с), то обеспечение требуемого времени T' можно достичь только путем увеличения μ_2. В терминах модели требование будет состоять в следующем: интенсивность обслуживания узла А должна быть не менее 25 1/с. Однако, из расчетов видно, что после достижения интенсивностью узла А значения 25 1/с выигрыш в T' становится все меньше, достигая лишь 0.006 с, что не оказывает существенного влияния на пользовательское восприятие.

В терминах прикладного уровня этот вывод трансформируется в рекомендацию к агенту пользовательского устройства. Его быстродействие можно повысить двумя способами или их комбинацией:

- увеличить аппаратный ресурс:

а) оперативная память устройства;

б) чип, обслуживающий агент пользовательского устройства.

- оптимизировать агент пользовательского устройства программно. На практике на этом сосредотачивают основное внимание производители облачного ПО, поскольку выбор чипов для устройств ограничен теми, которые имеют поддержку соответствующих агентов.

Рассмотренный здесь случай моделирует частую ситуацию из практики, когда в качестве терминального выступают слабое по характеристикам оборудование, как правило, это недорогие терминалы. Также такой случай может возникать, когда пользователь подключается к услуге, используя маломощный смартфон или планшетный компьютер, и вообще любое устройство по каким-либо причинам, не обеспечивающее необходимые для комфортной работы

характеристики. При этом в ряде случаев пользователь понимает, что использование такого устройства ухудшает комфорт работы, но готов мириться с этим. Подобная ситуация весьма часто встречается на практике, поэтому ее рассмотрение носит актуальный характер.

На рисунке 4.4. показан график зависимости времени обслуживания в узле А от времени T' при различных вариантах интенсивности обслуживания (времени обслуживания) в узлах С1, С2.

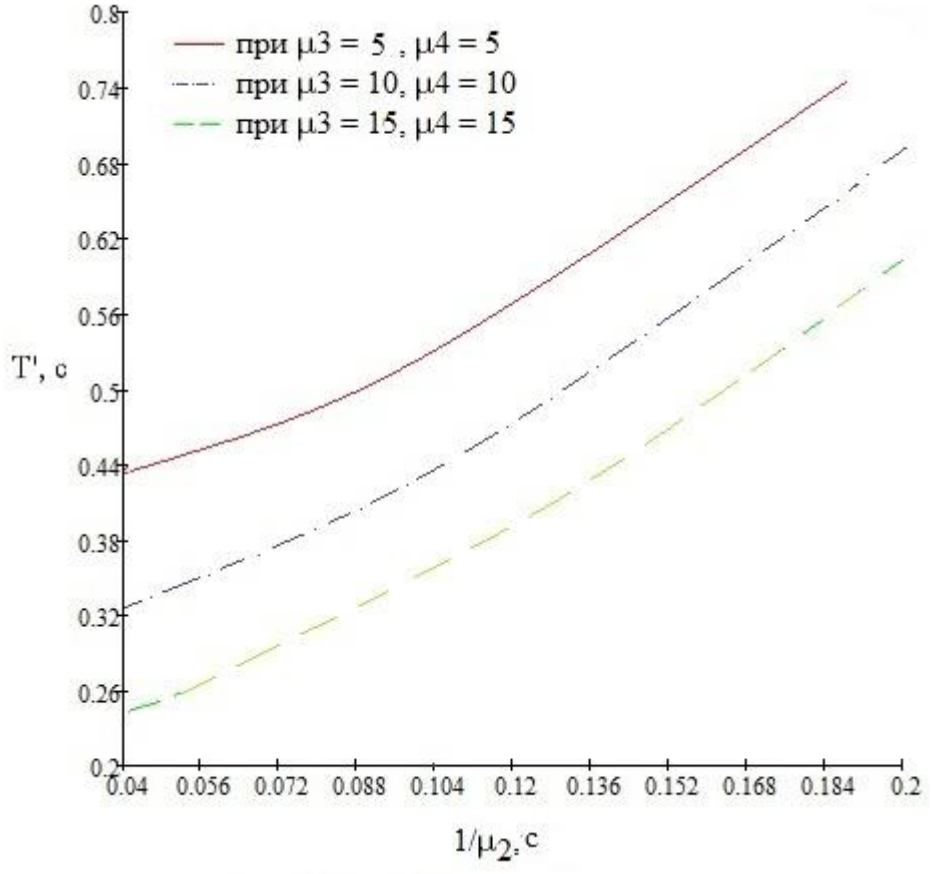

Рис. 4.4. График зависимости времени обслуживания в узле А от времени T'

Моделируем ситуацию более производительных процессоров: увеличиваем μ_2, при этом μ_3, μ_4 варьируем от 15 до 25 $с^{-1}$. Результаты расчетов сведем в таблицу П.3 Приложения 1.

Результаты произведенных расчетов показывают, что если процессоры более мощные (интенсивность обслуживания от 15 до 25 1/с), то уменьшение времени T' также можно достичь только путем увеличения μ_2, однако выигрыш в T' наступает заметно быстрее, чем в предыдущем случае. В терминах прикладного уровня этот вывод трансформируется в рекомендацию к агенту пользовательского устройства. Его интенсивность должна быть не менее 25 1/с. Однако, из расчетов видно, что после 30 1/с выигрыш в T' становится все меньше, достигая лишь 0.005 с, что не существенно для человеческого восприятия.

Ситуация, описанная моделью, представляет собой «классический» случай, когда пользовательский терминал адаптирован к работе с услугой, это устройство среднего ценового диапазона, рассчитанное для тех пользователей, которые занимаются стандартной офисной работой.

На рисунке 4.5 показан график зависимости времени обслуживания в узле А от времени Т’ при различных вариантах интенсивности обслуживания (времени обслуживания) в узлах С1, С2.

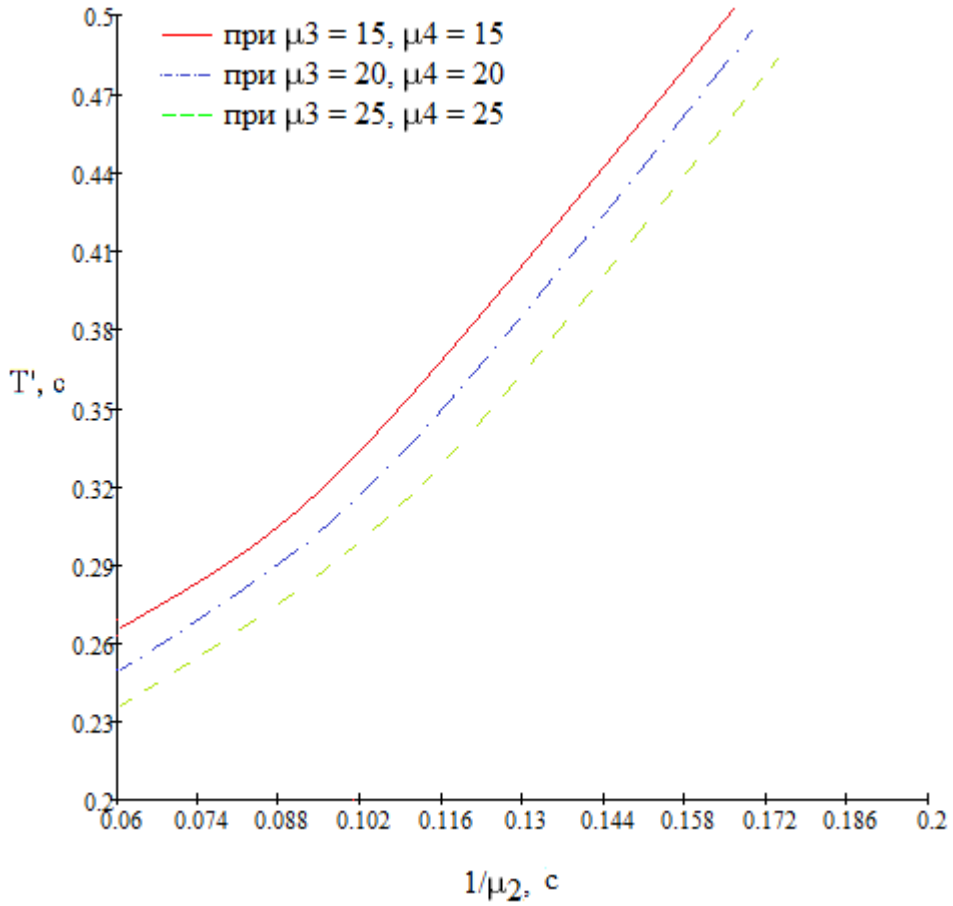

Рис. 4.5. График зависимости времени обслуживания в узле А от времени T'

Оценим время отклика. Как было показано в разделе 2, общее время отклика определяется по формуле (2.1) и состоит из нескольких слагаемых. Моделированию сетью Джексона позволяет учесть первое и третье слагаемые, значит теперь для оценки времени отклика в реальных условиях, необходимо учесть величину транспортной задержки. Будем использовать диапазон приемлемой задержки, определенный в разделе 2 Для расчетов, воспользуемся его нижней границей. Сведем в общую таблицу 4.1 два случая, рассмотренные выше (мало мощный (а) и более производительный (б) процессор).

Таблица 4.1. Среднее время отклика для случаев (а) и (б)

T' (случай а), с	T (случай а), с	T' (случай б), с	T (случай б), с
0.792	0.912	0.598	0.718
0.639	0.759	0.578	0.698
0.598	0.718	0.567	0.687
0.525	0.645	0.331	0.451
0.373	0.493	0.312	0.432
0.331	0.451	0.3	0.42
0.472	0.592	0.278	0.398
0.319	0.439	0.258	0.378
0.278	0.398	0.247	0.367
0.449	0.569	0.255	0.375
0.296	0.416	0.235	0.355
0.255	0.375	0.224	0.344
0.436	0.556	0.242	0.362
0.284	0.404	0.223	0.343
0.242	0.362	0.212	0.332
0.428	0.548	0.234	0.354
0.276	0.396	0.215	0.335
0.234	0.354	0.203	0.323
0.422	0.542	0.228	0.348
0.27	0.39	0.209	0.329
0.228	0.348	0.198	0.318

Сформулируем рекомендации к пользовательскому устройству по итогам моделирования в математических терминах и в физических терминах, присущих устройствам.

Для случая (а) при интенсивности обслуживания узлов С1, С2 от 5 до 15 1/с время отклика составило от 0.912 до 0.348 с. Согласно модели наименьшим (для слабых процессоров) оно становится при интенсивности обслуживания агента пользовательского устройства не менее 25 1/с. Дальнейшее его ускорение (при мало мощных процессорах), например, до 35 1/с, приводит к незначительному уменьшению времени отклика, поэтому, если оно подразумевает большие затраты, то смысла не имеет.

Для варианта №2 (более производительное устройство) при интенсивности обслуживания узлов С1, С2 от 15 до 25 1/с время отклика составило от 0.718 до 0.318 с.

В физических терминах увеличение интенсивности обслуживания агента пользовательского устройства означает увеличение его быстродействия.

4.1.3 Аналитическая модель сценария предоставления услуги при запуске видео внутри рабочего стола

Рассмотрим второй сценарий, при котором пользователю предоставляется возможность просматривать видео на рабочем столе. Предположим, что агент пользовательского устройства (в модели узел A) распределяет заявки по ядрам процессора на основании некоторых заранее известных данных. Агент виртуальной машины (в модели узел VM) имеет механизмы распознавания факта запуска пользователем видео и на основании принадлежности пакета к тому или иному типу, распределяет их по ядрам.

Когда пользователь запускает видео или аудио, в операционной системе запускаются соответствующие программы и наборы кодеков [38]. Информацию об этом считывает агент виртуальной машины, затем формирует выделенные в отдельные от потока изображений рабочего стола подпотоки, которые кодируются видео и аудио кодеками [96, 104, 102] и после передачи по сети декодируются на пользовательском устройстве соответствующим образом. Некоторые фирмы-разработчики протоколов удаленного рабочего стола ввели еще одну новую функцию протоколов виртуализации − эвристические алгоритмы [96, 104], дополнительно распознающие факт того, что пользователь включил видео на своем рабочем столе на основе анализа сегментов экрана, изменяющихся с большой скоростью. Такую модель работы принято называть «аппаратным ускорением» [46], однако различные производители облачного ПО применяют и другие термины.

При этом представляется обоснованным конструирование архитектуры пользовательского устройства (в особенности, специально предназначенного именно для услуги DaaS − тонкого клиента) таким образом, что отдельное ядро процессора предназначено для обработки аудио и видео потоков. Такой подход позволяет разделить передаваемые по сети данные согласно их типу и, соответственно, организовать раздельную обработку этих данных пользовательскими устройствами (видео и аудио данные обрабатываются ядрами процессора, специально подготовленными для этих задач), что положительно сказывается на восприятии качества услуги. Такой подход применяется на практике. Будем оценивать время отклика отдельно для каждого подпотока.

При обычной офисной работе пользователя видео запускается редко и не на полный экран. Случай, когда пользователь откроет видео на полный экран, что будет означать передачу только видео заявок, рассматривать не будем, поскольку это не рекомендовано производителями облачных платформ и систем доставки рабочего стола. Это продиктовано

тем, что качество услуги при этом резко падает: процессор пользовательского терминального устройства не предназначен для работы с такой возросшей нагрузкой, что приводит к проблемам с обработкой, и как следствие, к серьезным нарушениям качества изображения. Кроме того, возрастает загрузка канала связи между сервером и клиентом, что неизбежно приводит к искажениям видеоряда. Эти факторы приводят к характерным эффектам «затормаживания», резким рывкам, артефактам среди видеоряда.

Рассмотрим численный пример. В качестве входных данных для модели будем использовать μ_1 и λ'_0, полученные из эксперимента, описанного в параграфе 2.2.3, и будем задавать различные вероятности появления тех или иных типов заявок.

Пусть $\mu_1 = 10.5\ с^{-1}$, $\lambda'_0 = 2.5\ с^{-1}$, $T_{тр} = 120$ мс. Обозначим p_s – вероятность появления заявки типа видео, p_t – вероятность появления заявки типа изображения.

Требуется составить аналитическую модель, описывающую работу пользовательского устройства и сервера во втором сценарии работы терминальной сессии услуги. Получить и проанализировать временные характеристики компонентов инфраструктуры услуги, оценить среднее время отклика.

На основании специфики его работы процесс работы услуги «виртуальный рабочий стол» во втором сценарии ее работы уместно описать сетью массового обслуживания (СеМО) с двумя типами заявок, состоящей из четырех СМО: VM – агент ВМ; А – агент пользовательского устройства; С1 и С2 – ядра процессора пользовательского устройства.

Обмен данными в этом случае несколько отличается от предыдущей модели и описывается следующей последовательностью. Схема инфраструктуры услуги во втором сценарии ее работы показана на рисунке 4.6.

1. Агент пользовательского устройства (А) инициирует начало передачи данных с сервера т. е. отправляет поток заявок, который поступает на вход агента ВМ (VM). Поступающие агенту ВМ заявки, обнаружив, что прибор занят, становятся в очередь.

2. Агент ВМ (VM) обрабатывает поступившие заявки, затем отправляет агенту пользовательского устройства поток данных, состоящий из двух подпотоков (видео, изображения рабочего стола).

3. Агент пользовательского устройства (А) распознает принадлежность очередной заявки одному из подпотоков и распределяет заявки по ядрам процессора пользовательского устройства (С1 и С2). Поступающие агенту заявки, обнаружив, что прибор занят, становятся в очередь.

4. Ядро С1 обрабатывает заявки, относящиеся к видео, ядро С2 обрабатывает заявки, относящиеся к изображениям. После обработки ядра отправляют обслуженные потоки на

графическую подсистему, которая отрисовывает картинку пользователю. Поступающие ядрам заявки, обнаружив, что приборы заняты, становятся в очередь.

Совокупность перечисленных узлов и соединительных линий между ними будем рассматривать как сеть массового обслуживания, которую можно причислить к классу открытых СеМО, поскольку обслуженные узлами С1 и С2 заявки покидают сеть (отправляются на графическую подсистему пользователя).

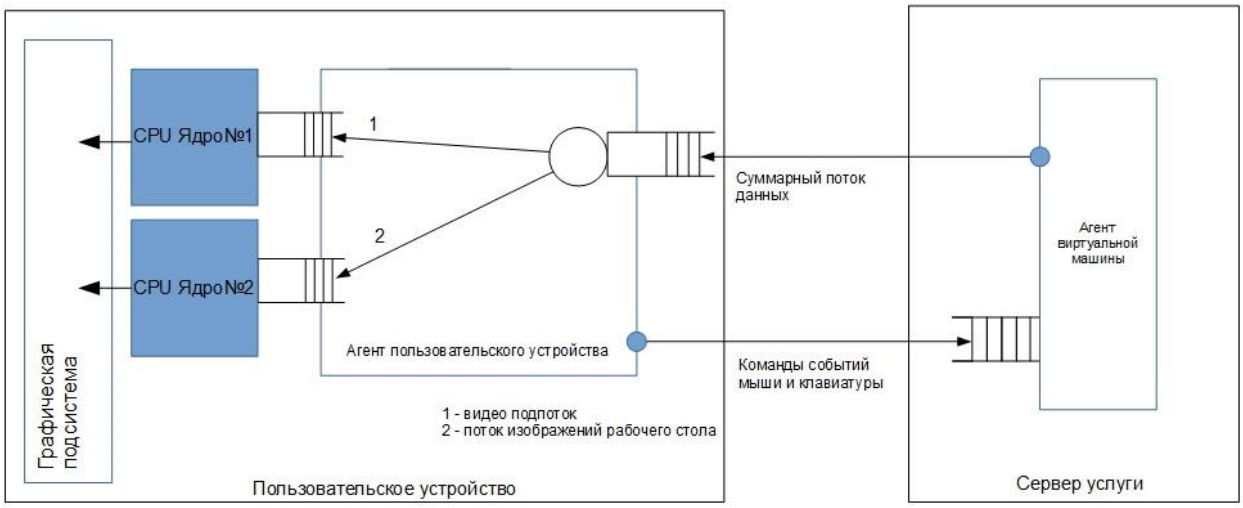

Рис. 4.6. Схема инфраструктуры услуги во втором сценарии работы

На основании приведенного анализа специфики работы услуги, аналитическую модель в этом случае уместно строить на основе BCMP-сети, впервые исследованной в [59] и рассмотренной в [10]. BCMP-сети являются, по сути, расширением сетей Джексона. Согласно аппарату BCMP СеМО состоит из нескольких узлов обслуживания, между которыми в соответствии с матрицей переходов циркулируют несколько классов (типов) заявок. При переходе между узлами заявка может сменить тип.

Согласно BCMP-теореме, лежащей в основе BCMP-сетей, существует четыре типа узлов BCMP. Будем использовать узлы типа 1 как наиболее подходящие по смыслу рассматриваемой услуги.

Узлы типа 1 – это узлы с дисциплиной обслуживания FCFS и экспоненциальным временем обслуживания. Интенсивность обслуживания заявок каждого типа во всех узлах этого типа должна быть одинакова. В узле типа 1 может размещается СМО вида ./M/m (система с различным видом входящего потока). В рамках модели будем использовать систему M/M/1 для сравнения со случаями 1 и 2, рассмотренными ранее. Такой тип узлов подходит для описания работы узла С2, получающего заявки одного типа и обрабатывающего их по порядку поступления.

Согласно [59] одним из условий объединения нескольких узлов в BCMP-сеть состоит в том, что распределение времени обслуживания в каждом узле должно иметь рациональное преобразование Лапласа. Это условие выполняется, в том числе, для узлов типа 1.

Распределение времени обслуживания в различных узлах может быть различным (не одинаковым). По т. BCMP вероятность того, что в узлах системы находится определенное количество заявок, может быть получена как произведение соответствующих вероятностей для отдельных узлов.

Введем следующие дополнительные обозначения и предположения.

1. Обозначим типы заявок следующим образом: *s* – видео, *t* – изображения рабочего стола.

2. Времена обслуживания заявок типов *s, t* в узле *i* распределены экспоненциально со средними $1/\mu_{is}$; $1/\mu_{it}$.

3. Обозначим T_s – время отклика для заявок типа видео; T_t – время отклика для заявок типа изображения.

4. Заявки типов *s* и *t* прибывают в узел *i* извне с интенсивностями $\lambda_{0i,s;}$ $\lambda_{0i,t}$.

При этом $\lambda'_0 = \lambda_{0i,s} + \lambda_{0i,t}$ – интенсивность общего потока.

5. Очереди узлов с дисциплиной FCFS имеют бесконечный буфер.

6. Обозначим интенсивности обслуживания:

- в узле VM типа *s* - μ_{1s}, типа *t* - μ_{1t};
- в узле A типа *s* - μ_{2s}, типа *t* - μ_{2t};
- в узле C1 типа *s* - μ_{3s};
- в узле C2 типа *t* - μ_{4t}.

7. Обозначим *L1, L2, L3* подцепи BCMP-сети. В терминологии BCMP подцепи – это пути следования заявок определенного типа внутри всей сети.

На рисунке 4.7 показан граф исследуемой сети. Для удобства дальнейшего построения уравнений в граф сети внесен источник заявок 0.

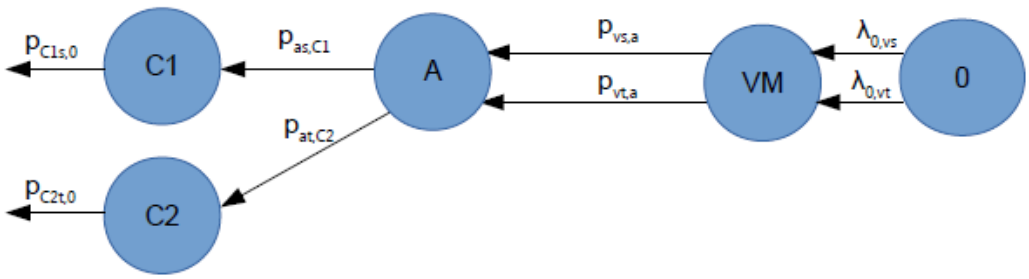

Рис. 4.7. Граф СеМО с двумя типами заявок

Введем также следующие дополнительные обозначения:

λ_{js} – интенсивность потока заявок типа s в узел *j*; $\lambda_{0,js}$ – интенсивность потока заявок типа *s* из узла 0 (источник заявок) в узел *j*; $p_{i,js}$ – вероятность того, что заявка типа *s* выйдет из узла *i* и перейдет в узел *j*; $p_{0,jt}$ – вероятность того, что заявка типа *t* войдет извне в узел *j*; $p_{js,0}$ – вероятность того, что заявка типа *s* выйдет из сети сразу после обслуживания в узле *j* (справедливо только для узлов С1, С2).

Запишем СУР для рассматриваемой сети.

$$\lambda_{js} = \lambda_{0,js} + \sum_{i,s} \lambda_{is} \cdot p_{i,js}\ ,\ \lambda_{jt} = \lambda_{0,jt} + \sum_{i,t} \lambda_{it} \cdot p_{i,jt}\ .$$

$$\begin{cases} \lambda_{vs} = \lambda_{0,vs}; \\ \lambda_{vt} = \lambda_{0,vt}; \\ \lambda_{as} = \lambda_{vs} \cdot p_{vs,a}; \\ \lambda_{at} = \lambda_{vt} \cdot p_{vt,a}. \end{cases}$$

$$\lambda_{C1S} = \lambda_{as} \cdot p_{as,C1};\ \lambda_{C2,t} = \lambda_{at} \cdot p_{at,C2}\ .$$

Для узлов типа 1 (FCFS) справедливы следующие формулы.

Среднее время обслуживания в узле заявки типа *r* можно определить из выражения:

$$t_{ir} = \frac{1}{\mu_{ir}}\ . \tag{4.9}$$

Согласно [22] время ожидания в узле заявки типа *r*определяется по формуле:

$$W_{ir} = \frac{\rho_i}{(1-\rho_i)\cdot \mu_{ir}} (1-\rho_i)^{-1}. \tag{4.10}$$

Среднее время пребывания в узле заявки типа *r* будет равно:

$$\overline{T}_i^r = W_{ir} + \frac{1}{\mu_{ir}}. \tag{4.11}$$

Далее определим среднее время отклика для заявок типа *s*, которое будет определяться в виде суммы средних времен пребывания заявки в узлах всей подцепи (с учетом временных характеристик, полученных для разных типов узлов):

$$t_{is} = \frac{1}{\mu_{is}}\ ,\ t_{it} = \frac{1}{\mu_{it}}.$$

$$T'_s = \overline{T}_1^s + \overline{T}_2^s + \overline{T}_3^s,\ T_t' = \overline{T}_1^t + \overline{T}_2^t + \overline{T}_4^t.$$

Среднее время отклика для заявок типов *s, t* с учетом транспортной задержки:

$T_s = T_s' + T_{mp}$, $T_t = T_t' + T_{mp}$.

Для моделирования различных вариантов интенсивностей поступающего потока, а также для рассмотрения широкого диапазона интенсивностей обслуживания в узлах сети будем придерживаться следующей последовательности действий.

1) Найдем интенсивности входящих в узлы С1, С2 потоков. По условию задачи будем рассматривать три варианта вероятностей появления заявок того или иного типа.

1 вариант:

$\lambda_1 = p_{ar1}\lambda_i = 0.4 \cdot 2.5 = 1\ c^{-1}$;

$\lambda_2 = p_{ar2}\lambda_i = 0.6 \cdot 2.5 = 1.5\ c^{-1}$;

2 вариант:

$\lambda_1 = p_{ar1}\lambda_i = 0.3 \cdot 2.5 = 0.75\ c^{-1}$;

$\lambda_2 = p_{ar2}\lambda_i = 0.7 \cdot 2.5 = 1.75\ c^{-1}$;

3 вариант:

$\lambda_1 = p_{ar1}\lambda_i = 0.2 \cdot 2.5 = 0.5\ c^{-1}$;

$\lambda_2 = p_{ar2}\lambda_i = 0.8 \cdot 2.5 = 2\ c^{-1}$.

Как было показано выше, до узла А доходит один поток интенсивностью λ_i, затем в этом узле он разделяется на два потока интенсивностями λ_1, λ_2. Значит, должно выполняться условие: $\lambda_i = \lambda_1 + \lambda_2$.

2) Рассчитаем T_i, задавая μ_2, μ_3, μ_4.

3) Рассчитаем T_s, T_t.

4) Для каждого варианта моделируем распространенный на практике случай, когда производители облачных услуг, стремясь улучшить воспринимаемое пользователем качество услуги, производят улучшения или замену пользовательских устройств (агент пользовательского устройства и процессорные ядра), поскольку заинтересованы в том, чтобы предоставлять услугу с приемлемым качеством пользователям с как можно более широким спектром устройств – от маломощных и дешевых, до производительных и дорогих, а также специально подготовленных к работе с услугой «виртуальный рабочий стол» (тонких клиентов). Для этого зафиксируем интенсивность обслуживания узла VM на уровне 10 1/с и будем варьировать интенсивности обслуживания узлов на пользовательской стороне. Результаты сведем в таблицы.

Рассмотрим комбинацию вероятностей появления заявок с видео и изображениями, когда $p_s = 0.4$, $p_t = 0.6$. Интенсивности потоков заявок и обслуживания в узлах показаны в таблице П.4, рассчитанные временные характеристики для узлов сети – в таблице П.5 Приложения 1.

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узле VM показаны на рисунке 4.8.

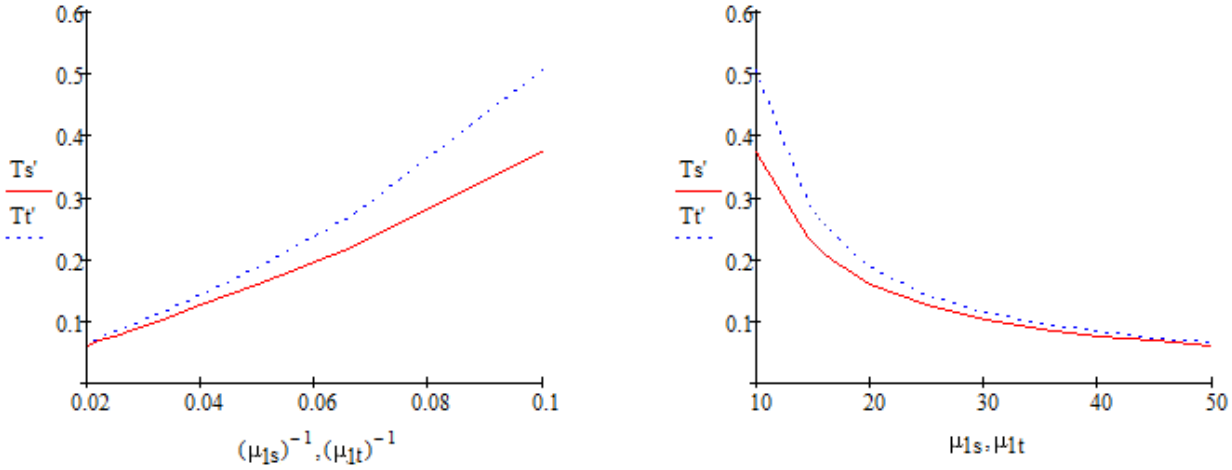

Рис. 4.8. Временные характеристики для узла VM

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узле A показаны на рисунке 4.9.

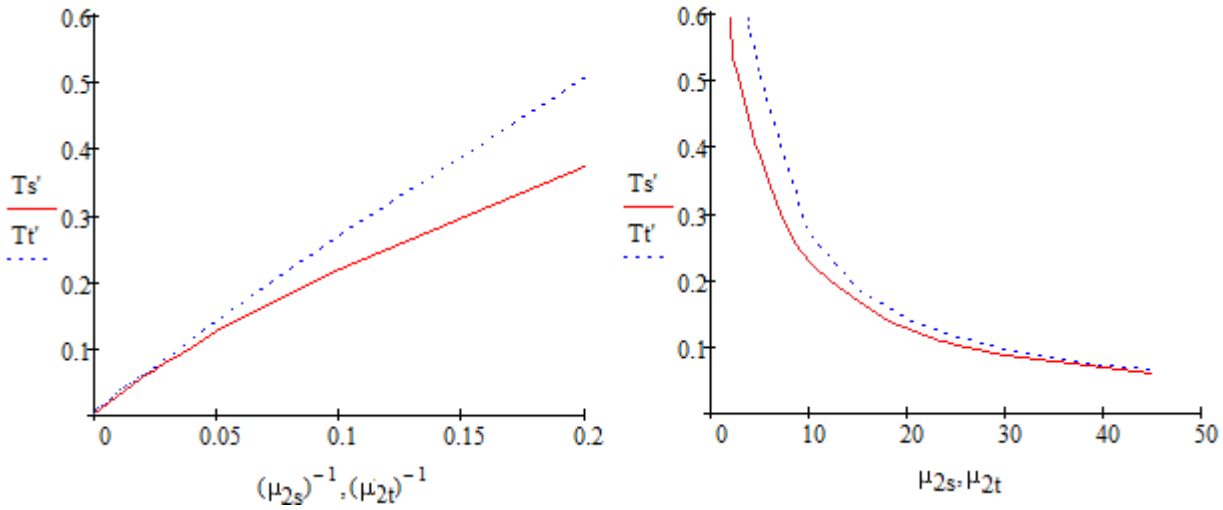

Рис. 4.9. Временные характеристики для узла A

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узлах C1, C2 показаны на рисунке 4.10.

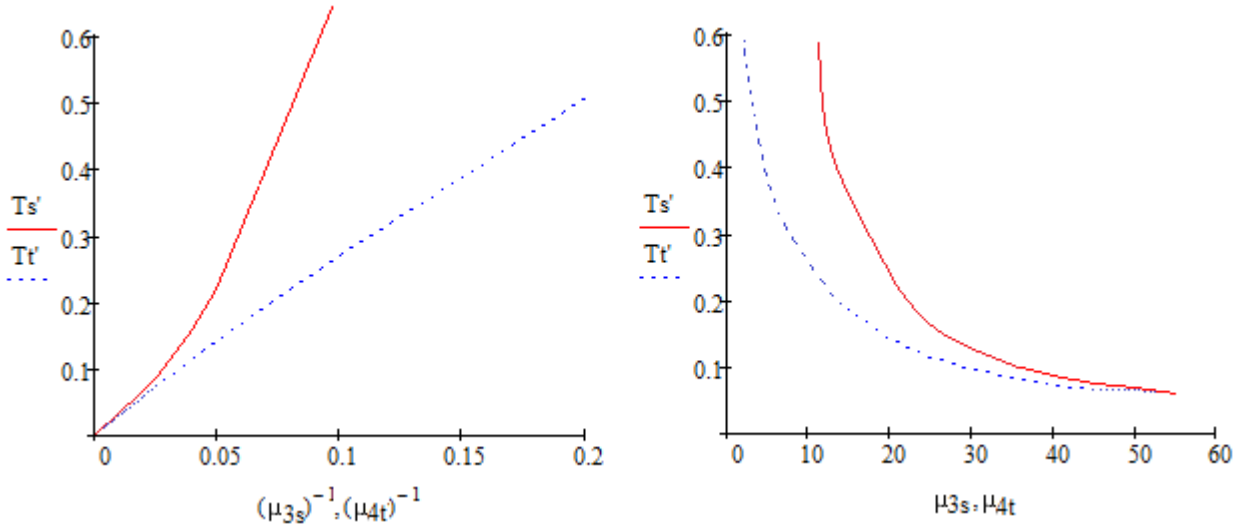

Рис. 4.10. Временные характеристики для узлов С1, С2

Рассмотрим далее распространенный на практике случай, когда производители облачных услуг, стремясь улучшить качество услуги, производят улучшения пользовательских устройств (агент пользовательского устройства и процессорные ядра), поскольку заинтересованы в том, чтобы предоставлять услугу с приемлемым качеством пользователям с как можно более широким спектром устройств – от маломощных и дешевых, до более производительных и дорогих, а также специально подготовленных к работе с услугой «виртуальный рабочий стол» (тонких клиентов). Для этого зафиксируем интенсивность обслуживания узла VM на уровне 10 1/с (это значение было получено экспериментально в параграфе 2) и будем варьировать интенсивности обслуживания узлов на пользовательской стороне. Результаты сведены в таблицу 4.2.

Таблица 4.2. Временные характеристики при фиксированной интенсивности обслуживания узла VM

T'_s	T'_t	T_s	T_t
0.375	0.509	0.495	0.629
0.256	0.307	0.376	0.427
0.213	0.24	0.333	0.36
0.189	0.206	0.309	0.326
0.174	0.186	0.294	0.306
0.164	0.172	0.284	0.292
0.156	0.163	0.276	0.283
0.15	0.155	0.27	0.275
0.146	0.15	0.266	0.27

Рассмотрим комбинацию вероятностей появления заявок с видео и изображениями, когда $p_s = 0.3$, $p_t = 0.7$. Интенсивности потоков заявок и обслуживания в узлах показаны в таблице П.7, рассчитанные временные характеристики для узлов сети сведены в таблицы П.8 Приложения 1.

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узле VM показаны на рисунке 4.11.

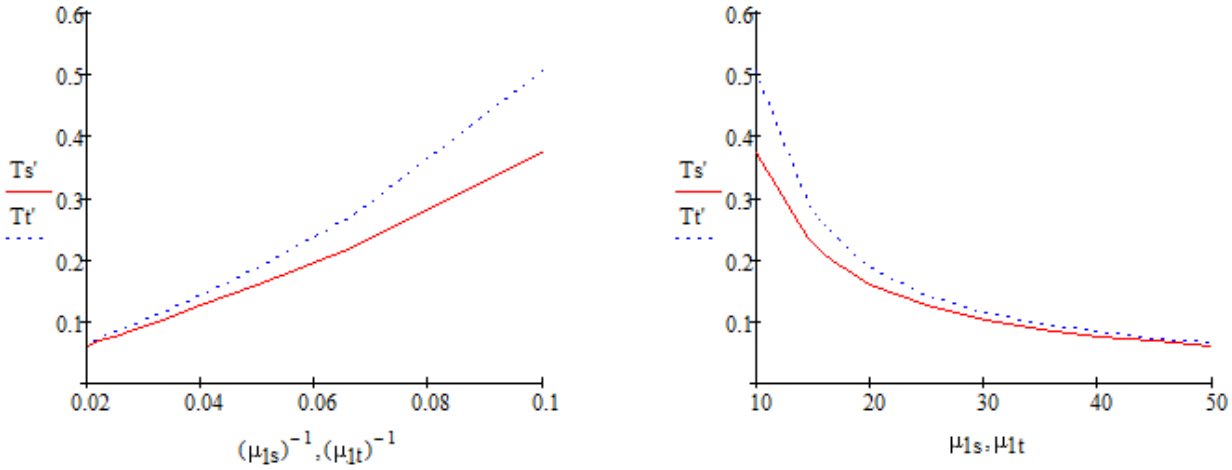

Рис. 4.11. Временные характеристики для узла VM

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узле A показаны на рисунке 4.12.

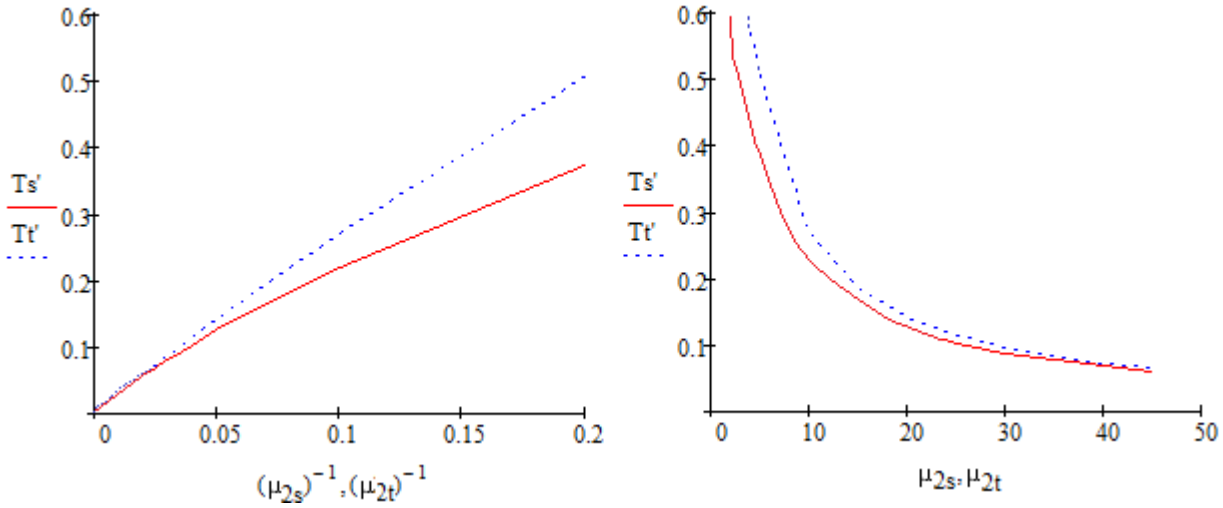

Рис. 4.12. Временные характеристики для узла A

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узлах C1, C2 показаны на рисунке 4.13.

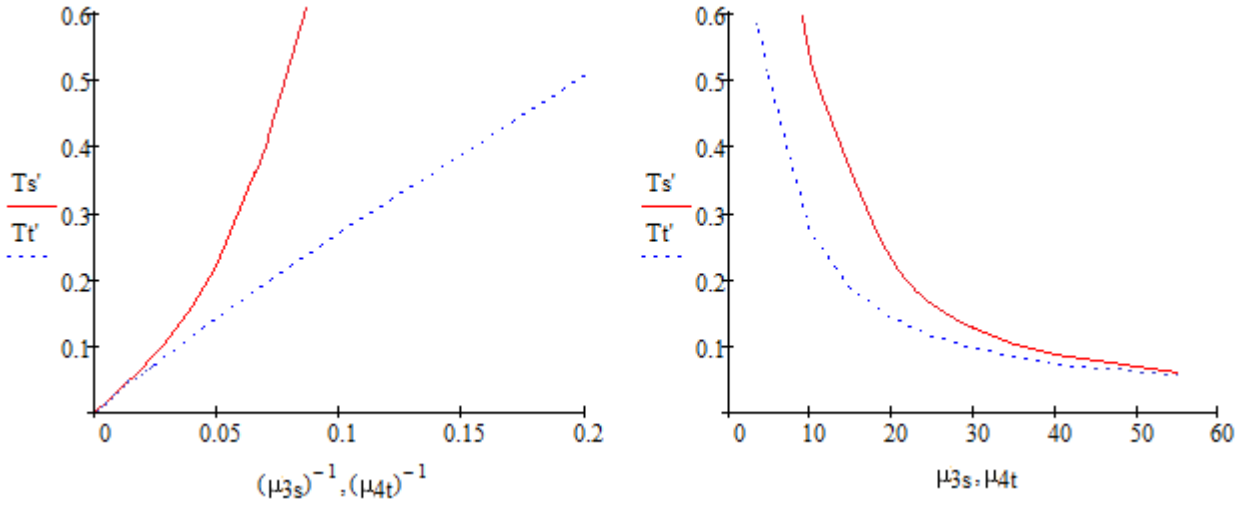

Рис. 4.13. Временные характеристики для узлов C1, C2

Рассмотрим комбинацию вероятностей появления заявок с видео и изображениями, когда p_s = 0.2, p_t = 0.8. Интенсивности потоков заявок и обслуживания в узлах показаны в таблице П.10, рассчитанные временные характеристики для узлов сети сведены в таблицы П.11 Приложения 1.

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узле VM показаны на рисунке 4.14.

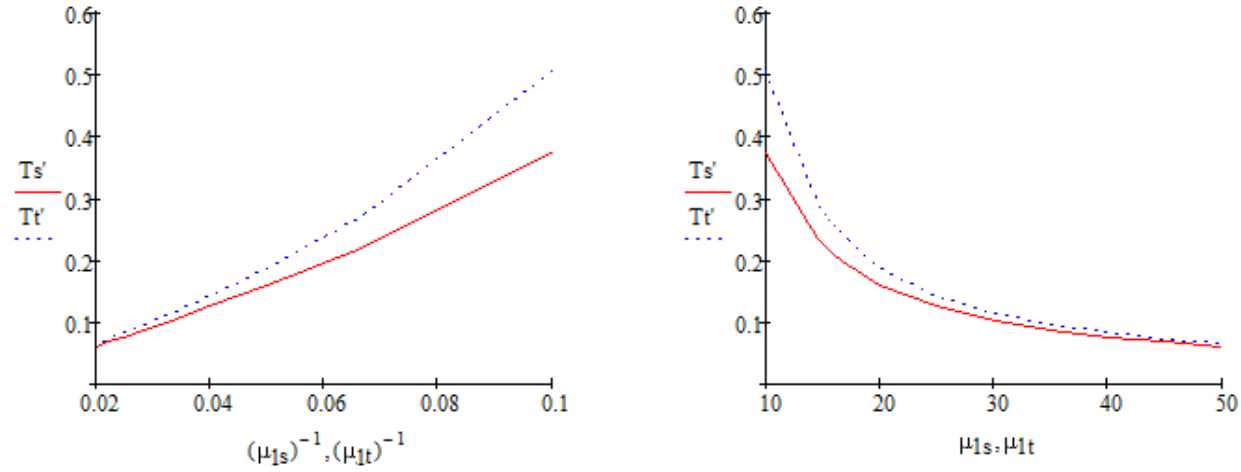

Рис. 4.14. Временные характеристики для узла VM

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узле A показаны на рисунке 4.15.

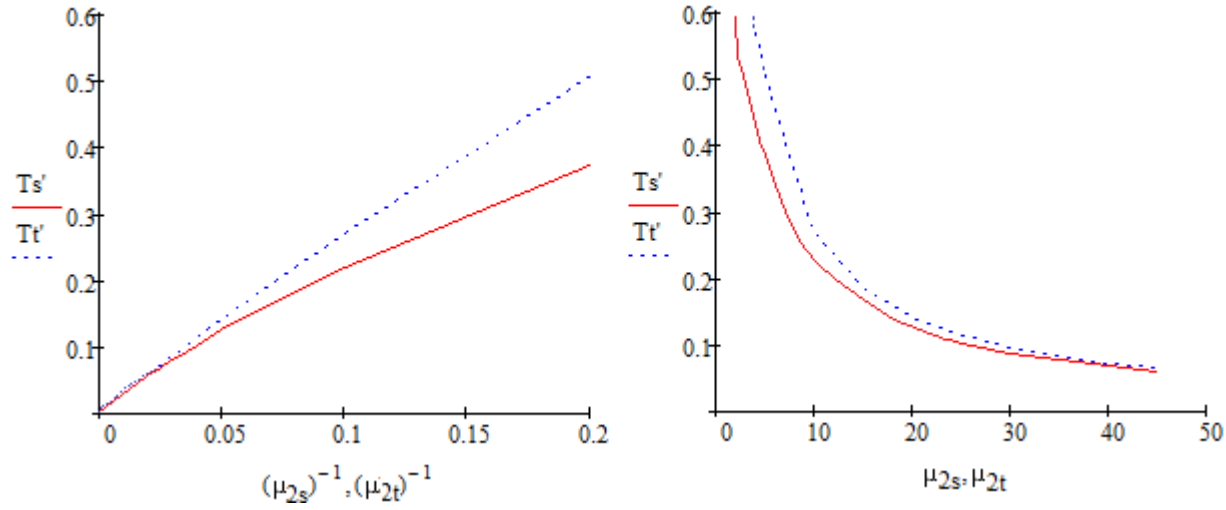

Рис. 4.15. Временные характеристики для узла A

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узлах C1, C2 показаны на рисунке 4.16.

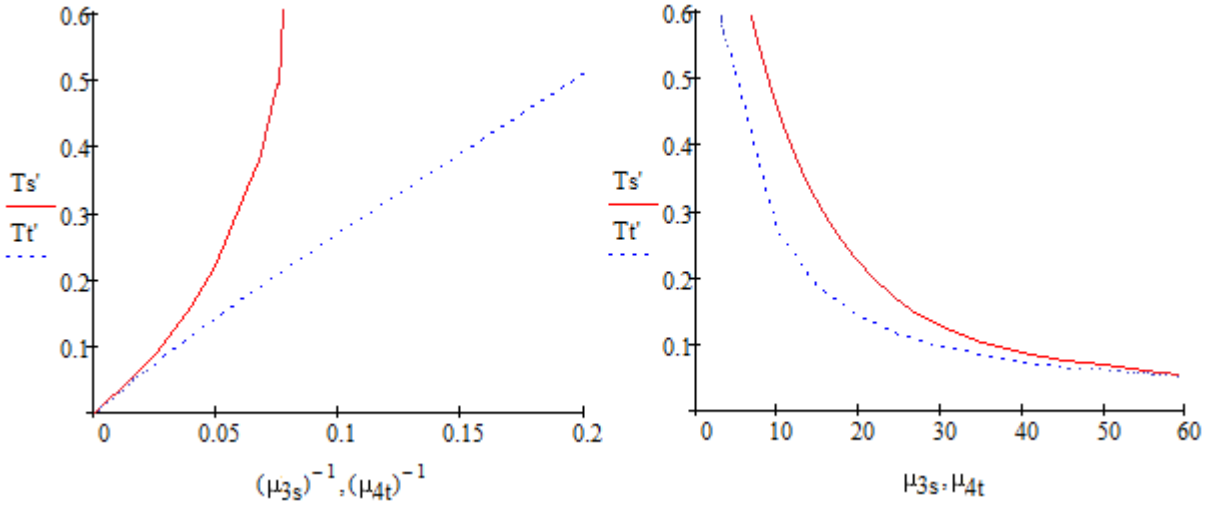

Рис. 4.16. Временные характеристики для узлов C1, C2

Таким образом, в модели было рассмотрено три набора по вероятностям (0.4, 0.6; 0.3, 0.7; 0.2, 0.8). Каждый набор рассчитывался набором значений μ_1 и комбинациями соотношений μ_2, μ_3, μ_4. Такой широкий охват различных условий позволил сделать следующие выводы.

1) Рассмотрен случай, когда производители облачного ПО или провайдеры услуги имеют возможность изменять быстродействие всех компонентов инфраструктуры. Из расчетов видно, что при одновременном увеличении интенсивности обслуживания агента виртуальной машины (VM) на серверной стороне и интенсивности обслуживания агента пользовательского

устройства (А) в три раза, для заявок типа видео время отклика уменьшается практически в 3.5 раза (от 0.371 до 0.103 с), для заявок типа изображения – в 4.4 раза (от 0.505 до 0.115 с). При дальнейшем увеличении интенсивностей обслуживания наблюдается уменьшение времени отклика, составляющее не более 0.02 с. Такое уменьшение оказывает слабое влияние на качество, поэтому дальнейшее увеличение интенсивностей представляется нецелесообразным.

2) Рассмотрен случай, когда производители облачного ПО проводят работы по улучшению быстродействия только ПО пользовательских устройств.

а) При фиксированной интенсивности обслуживания агента виртуальной машины на уровне 10 1/с (найденного экспериментально в разделе 2), и увеличении интенсивностей агента виртуальной машины в восемь раз и ядер процессоров в четыре раза, наблюдается уменьшение времени отклика практически в шесть раз (с 0.371 до 0.061 с).

б) Увеличение интенсивностей обслуживания узлов на пользовательском устройстве дает эффект уменьшения времени отклика, который наиболее выражен при их увеличении в два раза - время отклика уменьшается в 2.5 раза. Однако при дальнейшем увеличении интенсивностей обслуживания узлов этот эффект становится менее выраженным.

В подобных ситуациях, когда увеличение быстродействия на стороне сервера не представляется возможным (ограничен ресурс эмуляции процессора, оперативной памяти и т.д.), эффекта уменьшения времени отклика можно добиться путем увеличения интенсивностей обслуживания на пользовательском устройстве. Это можно осуществить двумя способами: аппаратно (замена блоков на более производительные) и программно (корректировка кода агента его производителем для увеличения быстродействия). Перечисленные действия могут быть выполнены провайдером услуги или производителем облачного ПО.

Эти выводы могут быть использованы производителями пользовательских терминалов, заинтересованных в быстродействии своих устройств; операторами связи и провайдерами облачных услуг при планировании сети и сервера услуги; сотрудникам IT подразделений, готовящих переход офисных или удаленных сотрудников в облачную среду; пользователями, выбирающими устройство для своих нужд.

4.1.4 Аналитическая модель сценария предоставления услуги при запуске видео и аудио

В параграфе 4.1.3 описывался расширенный сценарий с возможностью просмотра видео внутри рабочего стола. Для организации третьего сценария с возможностью прослушивания аудио требуется «проброс» гарнитуры (наушников) с пользовательского устройства в

виртуальную машину. Для этого аудио подпоток также, как и видео, должен передаваться отдельно от потока изображений рабочего стола. За его обработку на пользовательском устройстве отвечает ядро №2 при помощи набора аудио-кодеков, это происходит совместно с видео подпотоком. Схема работы при таком сценарии показана на рисунке 4.17.

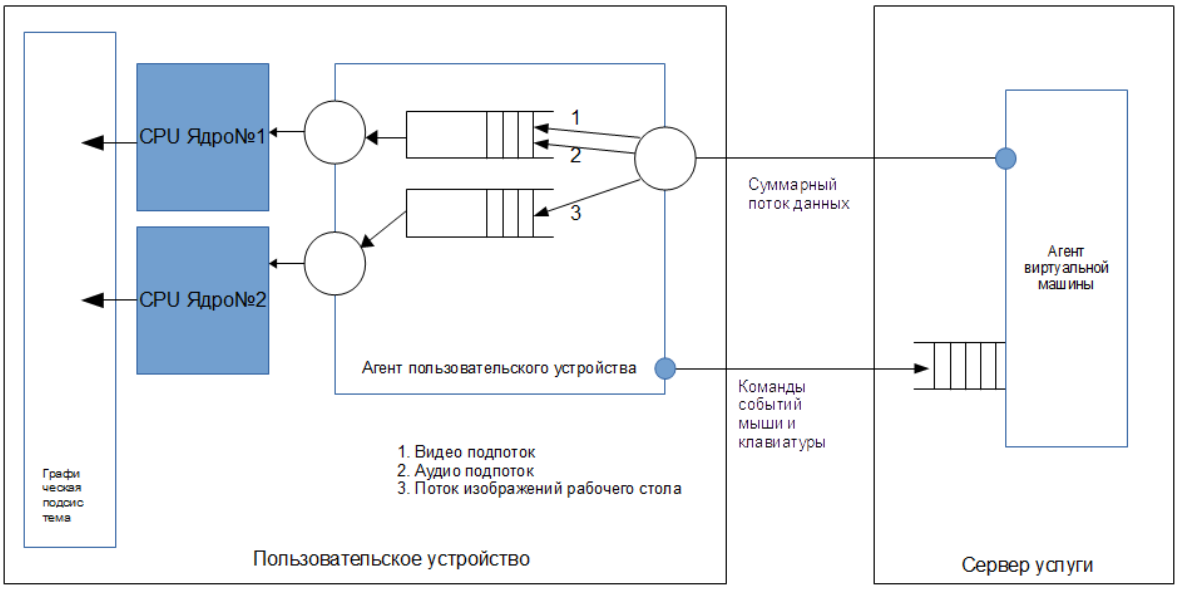

Рис. 4.17. Схема инфраструктуры услуги в третьем сценарии работы

Модель аппаратного ускорения (совместно с дополнительным «пробросом» и обработкой аудио подпотока) реализуется на практике различными производителями программных и аппаратных продуктов для облачных услуг. Например, в техническом бюллетене компании HP [110] показано сравнение производительности чипов с использованием аппаратного ускорения и без. Этот документ примечателен тем, что в нем рассматривается чип Texas Instruments DM 8148 (SoC, system-on-a-chip), который повсеместно используется производителями тонких клиентов, являющимися специальным терминальным оборудованием именно для услуги DaaS. Такой чип устанавливается, в частности, в тонкие клиенты HP T410 [31], Eltex TC-20 [32] и др.

В упомянутом документе показано, что при помощи широко распространенной системы виртуализации Citrix XenDesktop [33] организовывались удаленные рабочие столы, на которых запускалось видео с различным качеством: варьировались разрешение и развертка. В качестве терминального оборудования использовался тонкий клиент модели T410. При этом, при помощи средств разработки и отладки производителя, измерялась производительность чипов двух типов архитектур – с DSP (Digital Signal Processor), осуществляющим аппаратное декодирование видео (аппаратное ускорение) в реальном времени и без DSP. Единицей измерения были приняты fps (кадр/с), которые являются стандартной единицей измерения при оценке производительности видео кодеков. Результаты показаны в таблице 4.3.

Таблица 4.3. Сравнение архитектур пользовательского устройства с аппаратным ускорением и без

Воспроизводимый поток	Архитектура с DSP	Архитектура без DSP
Avi (480p, MPEG-4 v2)	20	20
Mp4 (720p AVC)	23	16
WMV (720p WMV-3)	23	14
WMV (1080p WMV-3)	23	20
Mp4 (1080p AVC)	23	4
Mov (1080p H.264)	24	4

Как видно, при невысоком качестве видео (480p) разницы между двумя типами архитектур нет. Однако, при наиболее часто используемом на практике качестве видео (720p и 1080p) разница становится заметной. Архитектура с аппаратным ускорением существенно опережает в производительности архитектуру с обычным процессором. Этот итог говорит о выигрыше в производительности, который достигается за счет использования специально подготовленного для аппаратного декодирования видео ядра процессора или SoC.

В [111] рассматривается сравнение производительности обработки видео четырьмя моделями тонких клиентов. Согласно приведенным данным, модели, использующие архитектуру с аппаратным ускорением, также опережают обычные по производительности.

Аналогичные рассуждения применимы и для аудио подпотоков: аудио, запускаемое на удаленном рабочем столе, должно передаваться отдельным от изображений рабочего стола потоком. Соответственно, обрабатываться полученный поток должен в выделенном ядре (SoC), имеющим набор определенных видео кодеков, наличие которых автоматически подразумевает и наличие аудио кодеков (например, кодек MPEG-4 несет ряд аудио кодеков, таких, как FLAC; кодек MPEG-2 – AAC). Стандартная офисная работа (случаи 1 и 2) не подразумевает запуска аудио, если это не оговорено отдельно в рамках того или иного программного продукта виртуализации.

Таким образом, становится понятно, что рассматриваемая архитектура с аппаратным ускорением реализована и применяется на практике.

Требуется составить аналитическую модель, описывающую работу пользовательского устройства и сервера в третьем сценарии работы терминальной сессии услуги. Получить и проанализировать временные характеристики компонентов инфраструктуры услуги, оценить среднее время отклика.

Основываясь на описании выше, процесс работы услуги «виртуальный рабочий стол» в третьем сценарии уместно описать сетью массового обслуживания (СеМО) с тремя типами

заявок, состоящей из четырех СМО: VM – агент ВМ; А – агент пользовательского устройства; С1 и С2 – ядра процессора пользовательского устройства.

Обмен данными в этом случае описывается следующей последовательностью.

1. Агент пользовательского устройства (А) инициирует начало передачи данных с сервера т. е. отправляет поток заявок, который поступает на вход агента ВМ (VM). Поступающие агенту ВМ заявки, обнаружив, что прибор занят, становятся в очередь.

2. Агент ВМ (VM) обрабатывает поступившие заявки, затем отправляет агенту пользовательского устройства поток данных, состоящий из трех подпотоков (видео, аудио, изображения рабочего стола).

3. Агент пользовательского устройства (А) распознает принадлежность очередной заявки одному из подпотоков и распределяет заявки по ядрам процессора пользовательского устройства (С1 и С2). Поступающие агенту заявки, обнаружив, что прибор занят, становятся в очередь.

4. Ядро С1 обрабатывает заявки, относящиеся к видео и к аудио. Ядро С2 обрабатывает заявки, относящиеся к изображениям. После обработки ядра отправляют обслуженные потоки на графическую подсистему, которая отрисовывает картинку пользователю. Поступающие ядрам заявки, обнаружив, что приборы заняты, становятся в очередь.

Совокупность перечисленных узлов и соединительных линий между ними будем рассматривать как сеть массового обслуживания, которую можно причислить к классу открытых СеМО, поскольку обслуженные узлами С1 и С2 заявки покидают сеть (отправляются на графическую подсистему пользователя). Схема такой СеМО показана на рисунке 4.18.

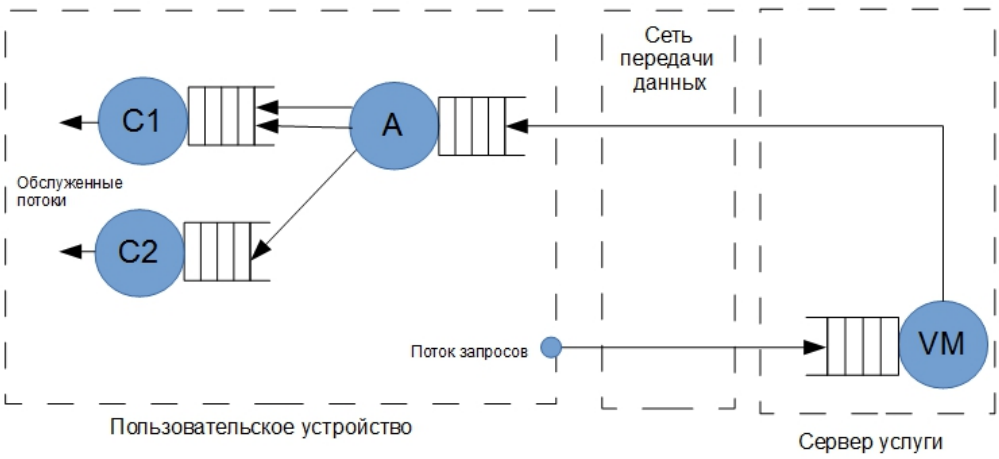

Рис. 4.18. Схема СеМО в третьем сценарии работы услуги

На основании приведенного анализа специфики работы услуги, аналитическую модель в этом случае уместно строить на основе ВСМР-сети с тремя типами заявок. Узел А может быть моделирован ВСМР-узлом типа 1, описание которого приведено в параграфе 4.1.3, либо узлом

типа 2. Узлы типа 2 – это узлы с дисциплиной обслуживания Processor Sharing (PS), в которой заявки обслуживаются одновременно с одинаковой скоростью. В этом узле тип класс заявок обслуживается независимо от других с различными распределениями времени обслуживания. Типы заявок различаются интенсивностями входящих потоков и интенсивностями обслуживания.

Конструирование программной архитектуры агента пользовательского устройства предполагает определения логики (дисциплины) обслуживания поступающих на него заявок. Для выяснения, при моделировании узлом с какой дисциплиной обслуживания время отклика будет меньше рассмотрим два варианта: когда узел А описывается BCMP-узлом типа 1 и узлом типа 2.

Введем следующие обозначения и предположения.

1. Обозначим типы заявок следующим образом: *r* – аудио, *s* – видео, *t* – изображения рабочего стола.

2. Времена обслуживания заявок типов *r, s, t* в узле *i* распределены экспоненциально со средними $1/\mu_{ir}$; $1/\mu_{is}$; $1/\mu_{it}$.

3. Заявки типов *r, s, t* прибывают в узел *i* извне с интенсивностью $\lambda_{0i,r};\lambda_{0i,s};\ \lambda_{0i,t}$.

При этом $\lambda_0 = \lambda_{0i,r} + \lambda_{0i,s} + \lambda_{0i,t}$ – интенсивность общего потока.

4. Очереди узлов с дисциплиной FCFS имеют бесконечный буфер.

5. Обозначим интенсивности обслуживания:

- в узле VM типа *r* – μ_{1r}, типа *s* – μ_{1s}, типа *t* – μ_{1t};
- в узле А типа *r* – μ_{2r}, типа *s* – μ_{2s}, типа *t* – μ_{2t};
- в узле С1 типа *r* – μ_{3r}, типа *s* – μ_{3s};
- в узле С2 типа *t* – μ_{4t}.

На рисунке 4.19 показан граф исследуемой сети. Для удобства дальнейшего построения модели в граф сети внесен источник заявок 0.

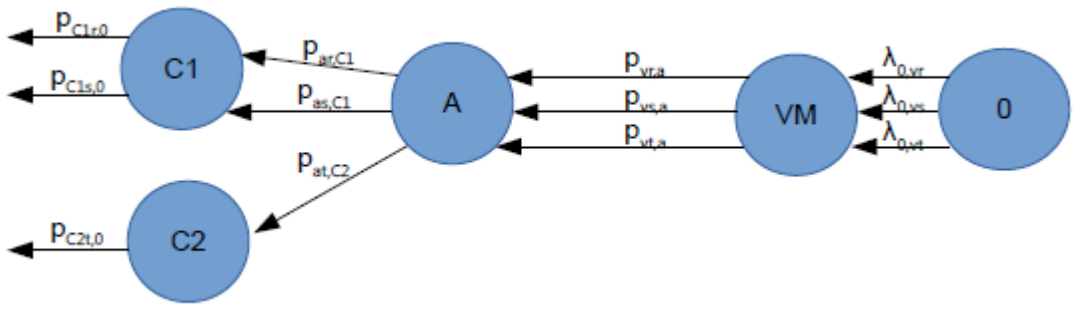

Рис. 4.19. Граф СеМО с тремя типами заявок

На рисунке 4.20 показан граф переходов между состояниями СеМО с тремя подцепями, по которым следуют заявки трех типов. Подцепи обозначены сплошной, пунктирной и штрихпунктирной стрелками. На этом графе перейдем к иным обозначениям узлов, в соответствии с принятым в аппарате BCMP-сетей подходом: узел обозначается индексом (k,x), где *k* – номер узла, *x* – тип заявок, проходящий через этот узел. Если через узел проходят заявки различных типов, такой узел обозначим на графе в виде нескольких объединенных узлов. Источник заявок обозначим точкой.

Заявки трех типов из источника заявок попадают в узел №1, затем общим потоком, состоящим из суммы трех подпотоков, доставляются в узел №2, где разделяются: в узел №3 следуют заявки типов *r* и *s*, в узел №4 следуют заявки типа *t*. Из узлов №3 и 4 заявки покидают сеть.

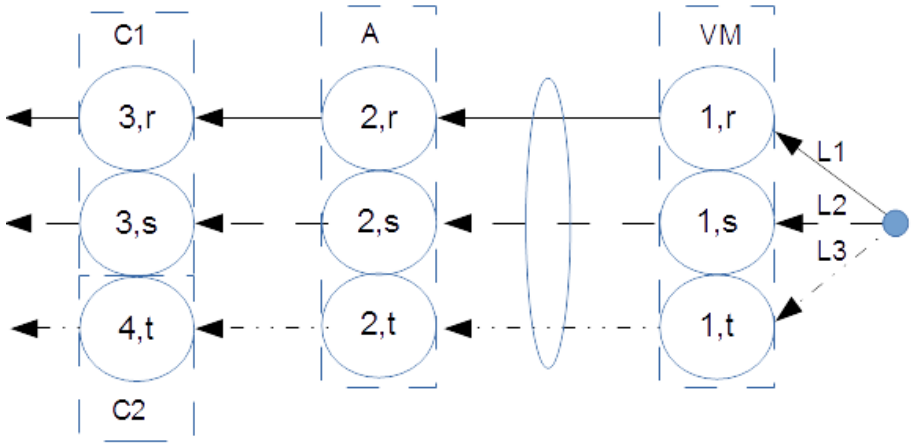

Рис. 4.20. Граф переходов между состояниями СеМО с тремя подцепями

Введем также следующие дополнительные обозначения:

λ_{jr} – интенсивность потока заявок типа *r* в узел *j*.

$\lambda_{0,jr}$ – интенсивность потока заявок типа r из узла 0 (источник заявок) в узел *j*.

$p_{i,jr}$ – вероятность того, что заявка типа *r* выйдет из узла *i* и перейдет в узел *j*.

$p_{0,jr}$ – вероятность того, что заявка типа r войдет извне в узел j.

$p_{jr,0}$ – вероятность того, что заявка типа *r* выйдет из сети сразу после обслуживания в узле *j* (справедливо только для узлов С1, С2).

Запишем СУР для рассматриваемой сети. Для типа *r*: $\lambda_{jr} = \lambda_{0,jr} + \sum_{i,r} \lambda_{ir} \cdot p_{i,jr}$.

Аналогично для типов *s* и *t*: $\lambda_{js} = \lambda_{0,js} + \sum_{i,s} \lambda_{is} \cdot p_{i,js}$, $\lambda_{jt} = \lambda_{0,jt} + \sum_{i,t} \lambda_{it} \cdot p_{i,jt}$.

Для узлов 1 (FCFS) средние времена обслуживания в узле, ожидания в узле и пребывания в узле могут быть вычислены по формулам (4.9 – 4.11). Для узлов типа 2 (PS) справедливы следующие формулы.

Среднее время обслуживания в узле заявки типа *r* можно определить из выражения:

$$t_{ir} = \frac{1}{\mu_{ir}}. \quad (4.12)$$

Среднее количество заявок типа *r* в узле, согласно [85] можно определить из выражения:

$$E[n_i^r] = \frac{\rho_i^r}{1-\rho_i}. \quad (4.13)$$

Среднее время пребывания заявки типа *r* в узле можно определить по формуле Литтла:

$$\overline{T}_i^r = \frac{E[n_i^r]}{\lambda_i^r} = \frac{\frac{1}{\mu_i^r}}{1-\rho_r}. \quad (4.14)$$

Далее определим среднее время отклика для заявок типа *r,* которое будет определяться в виде суммы средних времен пребывания заявки в узлах всей подцепи (с учетом временных характеристик, полученных для разных типов узлов):

$$T'_r = \overline{T}_1^r + \overline{T}_2^r + \overline{T}_3^r. \quad (4.15)$$

С учетом транспортной задержки $T_r = T'_r + T_{mp}$. Для заявок типов *s, t* получим аналогично:

$$t_{is} = \frac{1}{\mu_{is}},\ t_{it} = \frac{1}{\mu_{it}}.$$

$$T'_s = \overline{T}_1^s + \overline{T}_2^s + \overline{T}_3^s,\ T_t' = \overline{T}_1^t + \overline{T}_2^t + \overline{T}_4^t.$$

Среднее время отклика для заявок типов *s, t* с учетом транспортной задержки:

$$T_s = T_s' + T_{mp};\ T_t = T_t' + T_{mp}.$$

В терминах аналитической модели учтем особенность рассматриваемой архитектуры при помощи выбора соотношений интенсивностей обслуживания в узлах сети. Так, чтобы учесть тот факт, что использование второго ядра (SoC) для аппаратного декодирования видео и аудио дает выигрыш в общей производительности примем его интенсивность обслуживания большей интенсивности обслуживания основного ядра с учетом данных таблицы 4.3, однако рассмотрим более широкий по сравнению с представленным охват вариантов.

Зададим значения параметров для рассмотрения численного примера аналогичными со случаями 1 и 2, рассмотренными ранее с целью сравнения. Примем интенсивности обслуживания заявок различных типов узла А равными 10 1/с, интенсивности входящих потоков равными 2.5 1/с (то есть интенсивность суммарного потока будет равна 7.5 1/с). Будем варьировать величины интенсивностей обслуживания узлов сети с учетом особенности работы

рассматриваемого случая для построения численных характеристик. Эти значения также будем рассматривать аналогичными случаям 1 и 2. Исходные данные приведены в таблице 4.4.

Таблица 4.4. Исходные параметры

$\lambda_{0,vr}$	$\lambda_{0,vs}$	$\lambda_{0,vt}$	μ_{1r}	μ_{1s}	μ_{1t}	μ_{2r}	μ_{2s}	μ_{2t}	μ_{3r}	μ_{3s}	μ_{4t}
2.5	2.5	2.5	10	10	10	5	5	5	15	15	5
			15	15	15	10	10	10	20	20	10
			20	20	20	15	15	15	25	25	15
			25	25	25	20	20	20	30	30	20
			30	30	30	25	25	25	35	35	25
			35	35	35	30	30	30	40	40	30
			40	40	40	35	35	35	45	45	35
			45	45	45	40	40	40	50	50	40
			50	50	50	45	45	45	55	55	45

Рассчитанные по формулам (4.9 – 4.14) временные характеристики для узлов VM, A, C1, C4 сведем в таблицы П.13 – П.16 Приложения 1. Рассмотрим два варианта: когда узел А описывается узлом типа 2 (PS) и узлом типа 1 (FCFS). Произведем расчеты, в таблице 4.5 покажем сравнение времени отклика без учета транспортной задержки для заявок типов видео и изображения для рассматриваемых вариантов.

Таблица 4.5. Среднее время отклика для заявок типов *s* и *t* при моделировании узлами типа 1 и 2

Узел А с дисциплиной PS		Узел А с дисциплиной FCFS	
T'_s	T'_t	T'_s	T'_t
0.568	0.703	0.371	0.505
0.251	0.301	0.219	0.27
0.171	0.198	0.158	0.185
0.131	0.148	0.125	0.141
0.107	0.118	0.103	0.115
0.09	0.099	0.088	0.096
0.078	0.085	0.076	0.083
0.069	0.074	0.068	0.073
0.062	0.066	0.061	0.065

При интенсивностях обслуживания $\mu_{1r} = \mu_{1s} = \mu_{1t} = 10$ 1/с, $\mu_{2r} = \mu_{2s} = \mu_{2t} = 5$ 1/с, $\mu_{3r} = \mu_{3s} = 15$ 1/с, $\mu_{4t} = 5$ 1/с средние времена пребывания заявок в системе уменьшилось на 50 %, однако при увеличении интенсивностей на 10 1/с уменьшение времени пребывания стало составлять

всего 2 с, а при дальнейшем увеличении – стало составлять 0.1 % (см. таблицу П.13 Приложения 1). Следовательно, можно сделать вывод о том, что описание узла А узлом с дисциплиной FCFS приводит к уменьшению времени отклика.

Таким образом, можно сформулировать рекомендацию для производителей облачного ПО при конструировании агента пользовательского устройства: согласно проведенным расчетом меньшее время отклика достигается при дисциплине его работы вида FCFS.

Рассчитаем далее временные характеристики узлов С1 и С4. Результаты сведем в таблицу П.15 Приложения 1.

Рассчитаем среднее время отклика. Значение транспортной задержки будем использовать полученное в п. 2.1.3. Сведем в таблицу П.16 результаты расчетов.

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узле VM показаны на рисунке 4.21 Поскольку видео и аудио заявки обрабатываются с одинаковой интенсивностью, кривые для типов *r* и *s* совпадают.

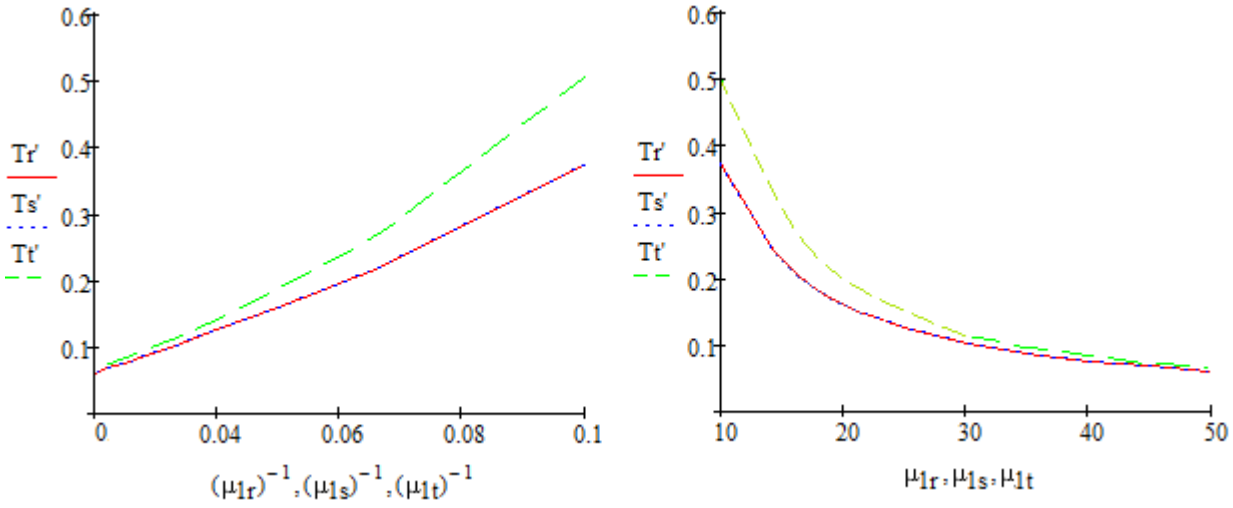

Рис. 4.21. Временные характеристики для узла VM

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узле А показаны на рисунке 4.22. Поскольку видео и аудио заявки обрабатываются с одинаковой интенсивностью, кривые для типов *r* и *s* совпадают.

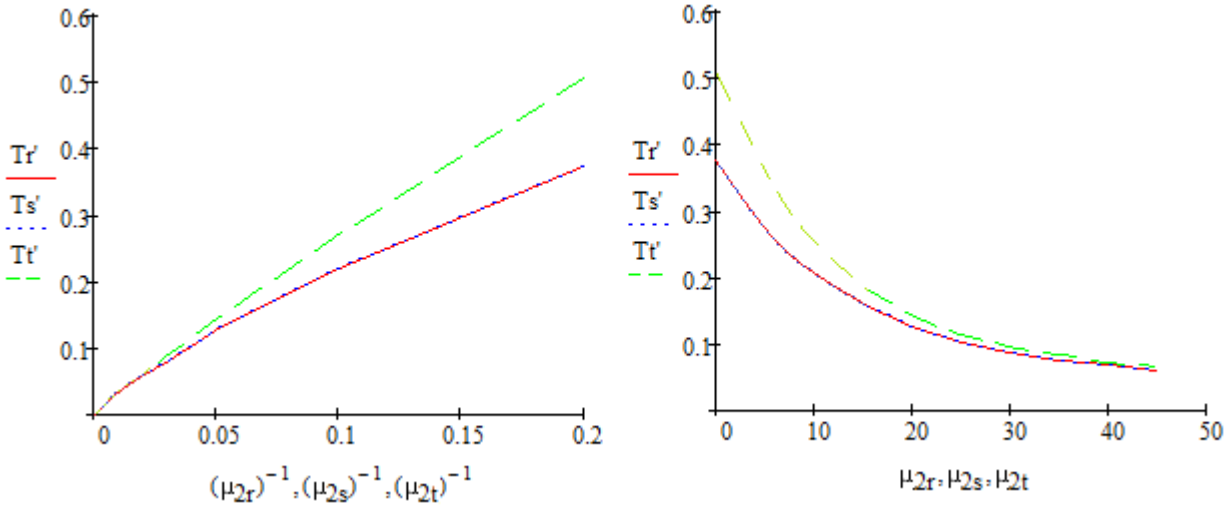

Рис. 4.22. Временные характеристики для узла A

Графики зависимости времени отклика (без учета транспортной задержки) от интенсивности обслуживания заявок и от времени обслуживания заявок в узлах C1 и C2 показаны на рисунке 4.23. Поскольку видео и аудио заявки обрабатываются с одинаковой интенсивностью кривые для типов *r* и *s* совпадают.

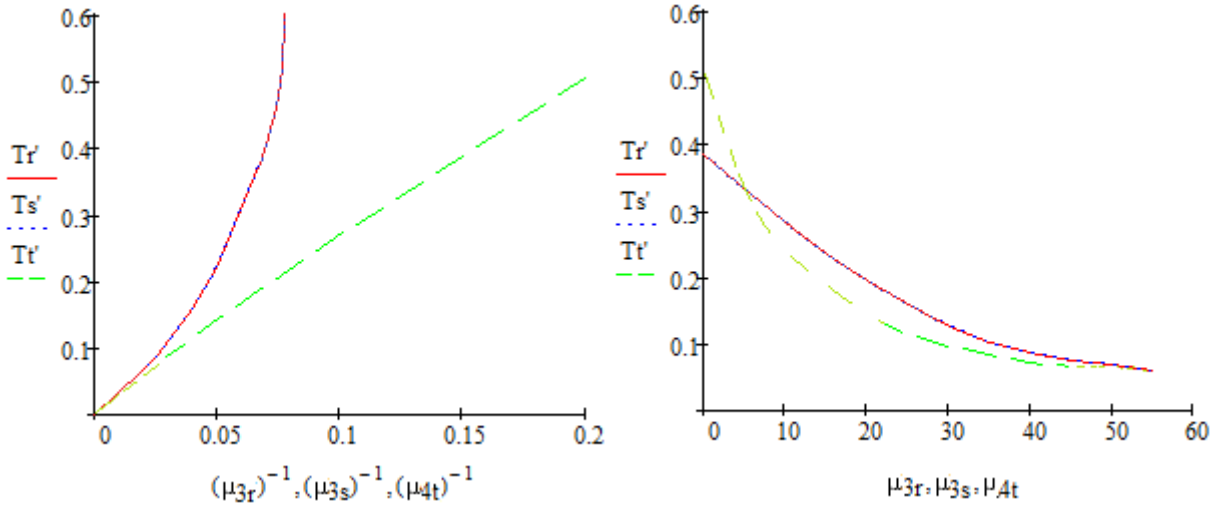

Рис. 4.23. Временные характеристики для узлов C1 и C2

Известно, что разработчики платформ, организующих услугу «виртуальный рабочий стол», стараются улучшить работу агентов пользовательского устройства (например, уменьшить объем передаваемых по сети служебных данных, увеличить интенсивность обработки поступающих с сервера услуги данных, улучшить качество кода агента для конкретной операционной системы пользовательского устройства для сведения к минимуму программных ошибок и т.д.), поскольку эти факторы наряду с временными характеристиками

(время отклика, транспортная задержка и др.) определяют качество работы всей услуги, и как следствие, пользовательское удовлетворение от работы.

Моделируем далее распространенный на практике случай, когда производители облачных услуг, стремясь улучшить воспринимаемое пользователем качество услуги, производят улучшения пользовательских устройств (агент пользовательского устройства и процессорные ядра), поскольку заинтересованы в том, чтобы предоставлять услугу с приемлемым качеством пользователям с как можно более широким спектром устройств – от маломощных и дешевых, до производительных и дорогих, а также специально подготовленных к работе с услугой «виртуальный рабочий стол» (тонких клиентов). Для этого зафиксируем интенсивность обслуживания узла VM на уровне 10 1/с (это значение было получено экспериментально в параграфе 2) и будем варьировать интенсивности обслуживания узлов на пользовательской стороне. Результаты показаны в таблице 4.6.

Таблица 4.6. Временные характеристики при фиксированной интенсивности обслуживания узла VM

T'_r	T'_s	T'_t	T_r	T_s	T_t
0.611	0.611	0.746	0.731	0.731	0.866
0.328	0.328	0.379	0.448	0.448	0.499
0.265	0.265	0.292	0.385	0.385	0.412
0.235	0.235	0.252	0.355	0.355	0.372
0.218	0.218	0.229	0.338	0.338	0.349
0.206	0.206	0.214	0.326	0.326	0.334
0.197	0.197	0.204	0.317	0.317	0.324
0.191	0.191	0.196	0.311	0.311	0.316
0.186	0.186	0.19	0.306	0.306	0.31

Таким образом, проанализировав полученные результаты, сформулируем следующие выводы.

1) При одновременном увеличении интенсивности обслуживания агента виртуальной машины и интенсивности обслуживания агента пользовательского устройства на 50 %, время отклика уменьшается почти вдвое. При дальнейшем увеличении интенсивностей их обслуживания еще на 50% наблюдается более плавное уменьшение времени отклика вплоть до 0.06 с без учета транспортной задержки.

В ситуациях, когда поставщик услуги не имеет возможности увеличивать ресурсы на серверной стороне, оправданно прибегнуть к увеличению быстродействия только пользовательских терминалов. Так, при фиксированной интенсивности обслуживания агента виртуальной машины и увеличении интенсивностей обслуживания агента пользовательского устройства на 50 % можно добиться уменьшения времени отклика практически на 35 %.

Увеличение интенсивности обслуживания узлов на пользовательском устройстве дает эффект уменьшения времени отклика, который наиболее выражен при их увеличении в два раза – время отклика уменьшается в 2.6 раз. Однако при дальнейшем увеличении интенсивностей обслуживания узлов этот эффект становится менее выраженным, в частности, при увеличении в пять раз наблюдается уменьшение времени отклика на 0.2 с меньшее, чем в случае, когда варьировались интенсивности всех узлов архитектуры, в том числе и агента виртуальной машины.

В таблице 4.7 представлены соотношения интенсивностей обслуживания в узлах, удовлетворяющих требованию по времени отклика.

Таблица 4.7. Соотношения интенсивностей обслуживания в узлах и среднее время отклика для модели сценария с видео и аудио

μ_{2s}/μ_{1s}	μ_{3r}/μ_{1r}	μ_{4t}/μ_{1t}	T'_r	T'_s	T'_t
0.5	1.5	0.5	0.371	0.371	0.505
0.66	1.33	0.66	0.219	0.219	0.27
0.75	1.25	0.75	0.158	0.158	0.185

Таким образом, при предоставлении пользователю не только изображений рабочего стола, но и аудио и видео данных, можно сформулировать следующие требования к серверу услуги и пользовательскому устройству. При необходимости обеспечения времени отклика на уровне менее 1.1 с (при этом следует учитывать, что время отклика будет определяться по наибольшему из времен для каждого типа заявок, полученных в модели, поскольку в конечном итоге все они объединяются в единую картинку на экране пользователя):

- планирование серверного ресурса. Должны выполняться следующие соотношения: величины интенсивностей обслуживания в узлах A, C1, C2 должны быть не менее μ_1, $0.6\mu_1$, $1.3\mu_1$ соответственно;
- планирование сетевого ресурса. Если исходить из оговоренного ограничения на время отклика T и при этом T_{mp} находится в пределах 120…150 мс, то среднее время отклика без учета транспортной задержки должно быть для видео и аудио заявок, исходя из расчетов, не более 0.33 с, для изображений – не более 0.29 с для устройств с архитектурой, включающих выделенное ядро для обработки аудио и видео.

4.2 Обобщенная аналитическая модель терминальной сессии

4.2.1 Введение и постановка задачи

Три аналитические модели, предложенные выше, имеют некоторые ограничения, которые на практике не всегда могут выполняться. К таким ограничениям относятся пуассоновский характер потоков, циркулирующих в сети и экспоненциальное распределение времени обслуживания в узлах инфраструктуры.

В ряде случаев некоторую инфокоммуникационную структуру не представляется возможным изучить целиком ввиду отсутствия информации о характерах потоков, циркулирующих в ней, а также о типах распределения длительности обслуживания в ее узлах. Такую структуру можно изучить только приближенно на основе принципа декомпозиции, который состоит в том, что общая сеть разбивается на несколько независимых элементов. Каждая подсистема рассматривается в виде СМО вида G/G/1. Для таких общих систем не существует развитой теории, поэтому используется метод, учитывающий два момента (мат.ожидание и дисперсию) СВ, описывающих расстояния между соседними заявками в потоках и длительность обслуживания в узлах, основанный на формулах Крамера и Лангенбах-Бельца [1, 53]. Для удобства расчетов будем использовать не дисперсию в чистом виде, а коэффициенты вариации (КВ).

Требуется составить математическую модель работы терминальной сессии услуги «виртуальный рабочий стол». Проанализировать характеристики полученной модели, оценить среднее время отклика.

Аналитическая модель услуги, построенная на основе приближенного метода, позволит дать более широкие рекомендации относительно параметров, определяющих качество. В частности, это достигается за счет того, что становится возможным исследовать непуассоновские потоки, а также отличные от экспоненциальных распределения длительности обслуживания в узлах инфраструктуры услуги.

4.2.2 Построение аналитической модели

Составление аналитической модели следует начать с построения сети массового обслуживания и обозначения ее узлов [48]. Для удобства записи уравнений обозначим узлы

следующим образом: 1 – агент ВМ; 2 – агент пользовательского устройства; 3, 4 – ядра процессора пользовательского устройства. Граф сети показан на рисунке 4.24.

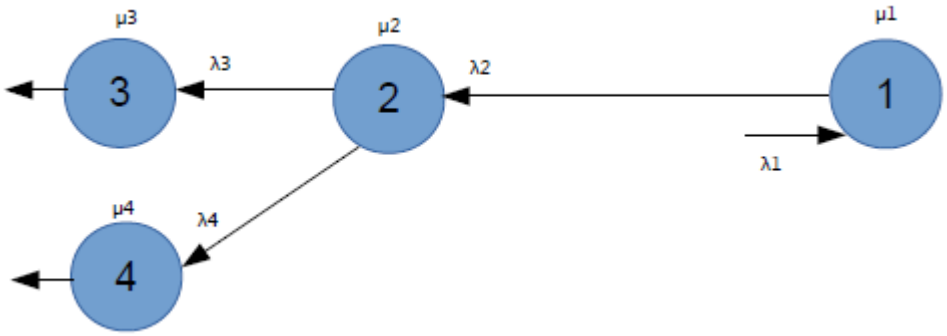

Рис. 4.24. Граф сети

Введем также следующие дополнительные обозначения:

p_{ij} – вероятность перехода заявки из узла i в узел j; λ_{0i} – интенсивность потока заявок извне в узел i; λ_i – интенсивность общего потока заявок в узел i; $M_{cpi} = 1/\lambda_{oi}$ – среднее расстояние между заявками (мат. ожидание расстояния между заявками); $C_A(i)$ – коэффициент вариации распределения расстояний между вызовами поступающего в узел i потока; $C_B(i)$ – коэффициент вариации распределения длительности обслуживания в узле i; μ_i – интенсивность обслуживания заявок в узле i; $h_i = 1/\mu_i$ – средняя длительность обслуживания (мат. ожидание времени обслуживания); T' – среднее время отклика без учета транспортной задержки.

Интенсивности потоков заявок в узел i определяются из СУР

$$\lambda_i = \lambda_{0i} + \sum_{j=1}^{N} \lambda_j p_{ji}, i = 1,.., N.$$

При этом для стационарного режима должно выполняться условие $\rho_i = \frac{\lambda_i}{\mu_i} < 1$.

Будем придерживаться следующей последовательности действий.

1. Зададим исходные параметры: интенсивность и КВ входящего в сеть потока.

2. Оценим коэффициенты вариации, относящиеся к обслуживанию в узлах $C_B(i)$. Определим коэффициенты вариации, относящиеся к потокам, циркулирующим в сети $C_A(i)$.

3. Для каждого узла вычислим среднее время ожидания в узле w_i. Прибавив к h_i, получим время пребывания заявки в узле. Сложив времена пребывания заявки в узлах по пути следования заявок, получим T' сети [49].

Среднее время ожидания в узле согласно [53] определяется по формуле Крамера и Лангенбах-Бельца [87]:

$$w_i = h_i \cdot \frac{\rho_i}{2 \cdot (1 - \rho_i)} \cdot [(c_A(i))^2 + (c_B(i))^2] \cdot g[\rho_i, (c_A(i))^2, (c_B(i))^2], \quad (4.16)$$

где $g[\rho_i,(c_A(i))^2,(c_B(i))^2]=\begin{cases}\exp\left[\dfrac{-2(1-\rho_i)}{3\rho_i}\cdot\dfrac{[1-(c_A(i))^2]^2}{(c_A(i))^2+(c_B(i))^2}\right], \text{если } c_A(i)<1.\\ \exp\left[-(1-\rho_i)\cdot\dfrac{(c_A(i))^2-1}{(c_A(i))^2+4(c_B(i))^2}\right], \text{если } c_A(i)\geq 1.\end{cases}$

Согласно архитектуре услуги, описанной в первом разделе, заявки могут проходить по пути 1-2-3 или 1-2-4.

Суммарное время ожидания заявки на узлах 1-2-3: $w_T = w_1 + w_2 + w_3$, (4.17)

Суммарное время обслуживания на узлах 1-2-3: $h_T = h_1 + h_2 + h_3$, (4.18)

Время отклика (без учета транспортной задержки): $T' = w_T + h_T$. (4.19)

В качестве исходных используем данные из предыдущей модели для сопоставления моделей: λ_1 = 2.5 1/с, μ_1 = 10.5 1/с. Для рассмотрения численного примера будем использовать значения интенсивностей обслуживания в узлах, рассмотренные в п. 4.1.2: μ_2 = 35 1/с, $\mu_3 = \mu_4$ = 15 1/с. Согласно схеме архитектуры с учетом того, что узел 2 распределяет заявки по узлам 3 и 4 равновероятно, определим значения интенсивностей входящих в узлы потоков: $\lambda_1 = \lambda_2$ = 2.5 1/с, $\lambda_3 = \lambda_4$ = 1.25 1/с.

4.2.3 Исследование коэффициентов вариации, относящихся к обслуживанию

Для оценки коэффициентов вариации распределения времени обслуживания в узлах проведем экспериментальное исследование. В состав экспериментального стенда входят: «Облако», коммутатор доступа, тонкий клиент [32], измерительный ПК. «Облако» включает в себя платформу виртуализации Microsoft Windows Server 2016 [102], а также его компоненты: менеджер подключений (Службы удаленных рабочих столов), репозиторий виртуальных машин, сервер авторизации и групповых политик. Измерительный ПК представляет собой компьютер с ОС Ubuntu Linux 14.04 LTS.

Для исследования коэффициентов вариации, относящихся к обслуживанию, рассматривается стенд со ста пользователями. При помощи средств отладки, входящих в комплект облачной платформы, собрана статистика временных интервалов между поступлениями соседних требований, а также временных интервалов между моментом приема требования агентом виртуальной машины и выдачей изображения рабочего стола на подсистему доставки пользователю (то есть процесс обслуживания). Исследуемые интервалы

времени показаны на рисунке 4.25. Также для более детального пояснения на рисунке показана транспортная задержка.

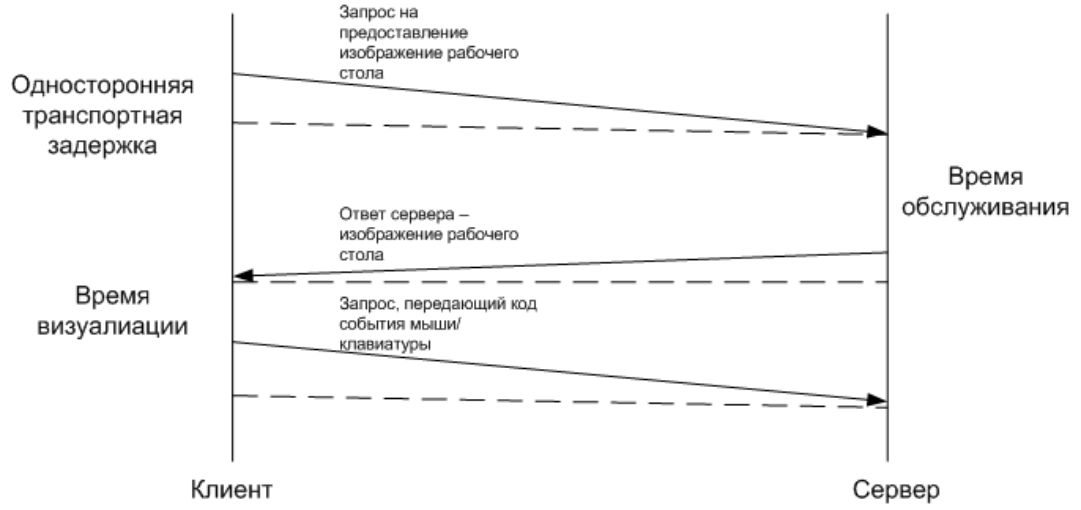

Рис. 4.25. Исследуемые временные интервалы

Коэффициенты вариации распределения длительности обслуживания в узлах 1 и 2 вычислены по формуле: $C_B(i) = \frac{S_i}{\overline{X_i}}$, где S_i – среднеквадратическое отклонение, $\overline{X_i}$ – среднее арифметическое времени обслуживания. В результате получено:

$C_B(1) = \frac{S_1}{\overline{X_1}} = \frac{0.9}{0.6} = 1.5, C_B(2) = \frac{S_2}{\overline{X_2}} = \frac{0.35}{0.2} = 1.75$, где значения 0.6 и 0.2 получены в результате экспериментального исследования, изложенного в 2.4.3.

4.2.4 Аналитический расчет коэффициентов вариации, относящихся к потокам

В [1] приведены алгоритмы расчета КВ интервалов между поступлениями заявок. Кратко перечислим их основные положения.

Алгоритм Райзера и Кобайаши, предполагающий наличие большой нагрузки в сети, основан на исследовании сети методом диффузной аппроксимации. Алгоритм Кюна предлагает собственную формулу для вычисления КВ, включающую в себя известный результат для системы M/GI/1/∞, полученный Макино. Описанные методы могут быть использованы при анализе сети под большой нагрузкой.

Алгоритм Геленбе-Пюжоля основан на трех предположениях: вероятность оставить узел непустым в момент ухода очередной заявки ненулевая; интервалы между поступлениями

заявок на узел *i* образуют рекуррентный поток; эти интервалы не зависят от длительностей обслуживания заявок в этом узле.

Алгоритм УДН для поиска коэффициентов предполагает решение системы линейных уравнений и позволяет исследовать сеть в более широком диапазоне нагрузок.

Воспользуемся методом УДН для аналитического нахождения коэффициентов вариации распределения расстояний между соседними заявками в потоке согласно [1]:

$$\gamma_A(i) - \sum_{k=1}^{M} \gamma_A(k)(1-\rho_k^2)p_{ki}^2 = \gamma_A(0,i) + \sum_{k=1}^{M} \gamma_B(k)\rho_k^2 p_{ki}^2,\ i = \overline{1,M},\ k = \overline{0,M},$$

где $\gamma_A(i) = \lambda_i(C_A{}^2(i)-1)$, $\gamma_A(k,i) = \lambda_{ki}(C_A{}^2(k,i)-1)$, $\gamma_B(i) = \mu_i(C_B{}^2(i)-1)$.

Коэффициент $\gamma_A(i)$ относится к потоку в узел *i*; $\gamma_A(0, i)$ относится к потоку извне в узел *i*; $\gamma_B(k)$ относится ко времени обслуживания в узле *k*.

Если величины $\gamma_A(i)$ известны, то КВ $C_A(i)$ могут быть вычислены по формуле:

$$C_A(i) = \sqrt{1 + \frac{\gamma_A(i)}{\lambda_i}} \quad . \tag{4.19}$$

Составим $P = \|p_{ij}\|$ матрицу переходов (маршрутную матрицу):

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & p_{ac1} & p_{ac2} \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

1. Запишем СЛАУ, сократив обращающиеся в ноль члены суммы, исходя из маршрутной матрицы.

$$\gamma_A(1) = \gamma_A(0,1);$$
$$\gamma_A(2) - \gamma_A(1)(1-\rho_1^2)p_{12}^2 = \gamma_B(1)\rho_1^2 p_{12}^2;$$
$$\gamma_A(3) - \gamma_A(2)(1-\rho_2^2)p_{23}^2 = \gamma_B(2)\rho_2^2 p_{23}^2;$$
$$\gamma_A(4) - \gamma_A(2)(1-\rho_2^2)p_{24}^2 = \gamma_B(2)\rho_2^2 p_{24}^2.$$

2. Используем исходные данные.

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$\gamma_A(1) = \gamma_A(0,1);$
$\gamma_A(2) - 0.943\gamma_A(1) = 0.057\gamma_B(1);$
$\gamma_A(3) - 0.249\gamma_A(2) = 0.00127\gamma_B(2);$
$\gamma_A(4) - 0.249\gamma_A(2) = 0.00127\gamma_B(2).$

3. Решим получившуюся систему методом подстановки. Для этого зададим параметры потока, входящего в узел 1. Как было сказано выше, λ_1 = 2.5 1/с. Пусть $C_A(1)$ = 0.6. Вычислив $\gamma_A(1), \gamma_B(1), \gamma_B(2)$ и подставляя полученные значения в систему, найдем:

$\gamma_A(2) = 0.0625\ \gamma_B(1) + 0.9375\ \gamma_A(1) = 0.095$;
$\gamma_A(3) = 0.001225\ \gamma_B(2) + 0.9987\ \gamma_A(2) = 0.158$;
$\gamma_A(4) = \gamma_A(3) = 0.158$.

Вычислим далее коэффициенты вариации по формуле (4.19) и сведем в таблицу 4.8 найденное.

Таблица 4.8. Коэффициенты вариации, относящиеся к потокам и обслуживанию

Обслуживание в узле 1, $C_B(1)$	Обслуживание в узле 2, $C_B(2)$	Поток в узле 1, $C_A(1)$	Поток в узле 2, $C_A(2)$	Поток в узле 3, $C_A(3)$	Поток в узле 4, $C_A(4)$
1.5	1.75	0.6	0.79	0.94	0.94

Приведенные в таблице 4.8 значения получены для исходных данных с использованием значений, оцененных экспериментально. Для того, чтобы исследовать более широкий диапазон значений, которые могут встречаться на практике, рассмотрим множество КВ, лежащих около полученных. Будем рассматривать множество коэффициентов вариации $C_B(i)$, лежащих в диапазоне 0…5. Этот диапазон включает значения КВ, полученные выше.

Рассмотрим работу сети массового обслуживания более детально. Поток, выходящий из сервера услуги (узел 1), проходит через сеть передачи данных, затем приходит на пользовательское устройство (узел 2).

Из-за наличия в сети передачи данных множества различных устройств характер потока, описываемый коэффициентом вариации распределения расстояния между соседними заявками в нем, может меняться. Иными словами, КВ $C_A(2)$ не равен $C_A(2)'$.

Для оценки влияния СПД на характер потоков, циркулирующих в моделируемой сети введем для описания СПД некоторый виртуальный узел Net (см. рисунок 4.26). Он описывается двумя параметрами: $C_B(Net)$ и $C_A(2)'$. Ввиду того, что в сети передачи данных (Интернет) на пути трафика от сервера до клиента может располагаться множество устройств, обслуживание потока узлом не может быть определено по объективным причинам, к тому же в рамках рассматриваемой задачи оно не представляет необходимости, важно учитывать лишь изменение потока при прохождении через сеть, характеризуемое некоторой $\Delta = C_A(2) - C_A(2)'$. Также будем считать узел Net работающим в стационарном режиме, $\lambda_2 = \lambda_2'$.

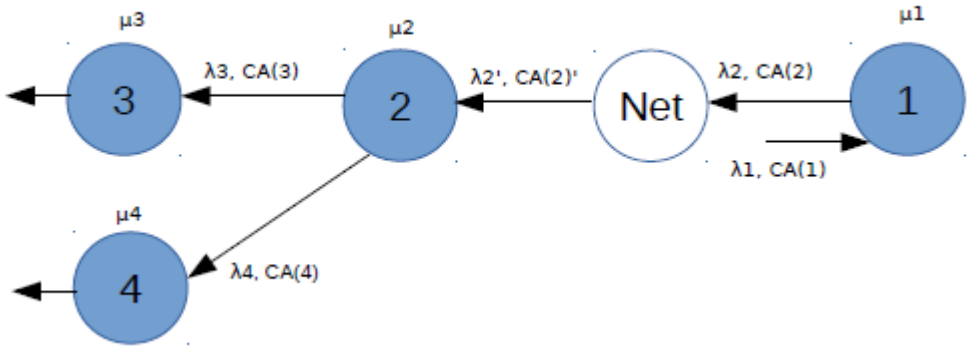

Рис. 4.26. Граф сети при наличии узла, моделирующего СПД

Оценка влияния СПД позволит выявить необходимость учета узла Net при расчетах среднего времени отклика. Если узел мало изменяет поток, при расчетах будем учитывать поток между узлами 1 и 2 с КВ $C_A(2)$, если узел существенно влияет на поток, то при расчетах используем $C_A(2)'$, входящий в узел 2.

Для определения $C_A(2)$ и $C_A(2)'$ проведем экспериментальное исследование.

4.2.5 Исследование коэффициентов вариации, относящихся к потоку, при прохождении через сеть передачи данных

Для исследования изменения характера потока, передаваемого от сервера к клиенту в процессе работы терминальной сессии услуги «виртуальный рабочий стол», при прохождении через сеть передачи данных (Интернет) проведем экспериментальное исследование. Используем стенд, описанный в п. 4.2.3.

Исследуем два наиболее типичных вида пользовательской деятельности: активная работа в рамках рабочего стола (передвижение окон, набор текста, серфинг в интернет-браузере и т.п.)

и отсутствие деятельности (например, чтение с экрана или покидание рабочего места). Подключение к серверу осуществляется при помощи протокола MS RDP [29] через сеть Интернет.

При помощи программы Wireshark [117] проводились измерения расстояний между соседними пакетами в потоках [15]. Измерения $C_A(2)$ производились на выходе с сервера услуги перед входом в сеть передачи данных, измерения $C_A(2)'$ производились на входе в клиентское устройство [16]. Коэффициенты вариации распределения расстояний между соседними заявками вычислены по формуле: $C_A(i) = \frac{\overline{S_i}}{\overline{X_i}}$, где $\overline{S_i}$ – статистическая оценка среднеквадратического отклонения, определяемого по формуле $\overline{S_i} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X_i})^2}{n-1}}$, где $\overline{X_i}$ – среднее арифметическое расстояния между соседними пакетами в потоке. Подставив полученные экспериментально значения, рассчитаем КВ и значения относительной и абсолютной погрешности (δ и Δ), результаты сведем в таблицу 4.9.

Таблица 4.9. Экспериментальные данные

Тип действия	$C_A(2)$	$C_A(2)'$	$\Delta = C_A(2)' - C_A(2)$	$\delta = (C_A(2)' - C_A(2)) / C_A(2)'$	Интервал измерений	Количество пакетов
Работа	0.8205	0.8811	0.0606	0.0687	2 часа	75400
Бездействие	0.7926	0.8259	0.0333	0.0403	2 часа	110

Получены КВ вышедшего с сервера услуги потока $C_A(2)$ и прошедшего через сеть передачи данных (Интернет) и вошедшего в узел 2 потока $C_A(2)'$. Из полученных данных видно, что при прохождении через сеть передачи данных поток изменяется незначительно, следовательно, при расчетах следует учитывать поток между узлами 1 и 2 с КВ $C_A(2)$. Следует уточнить, что полученные результаты являются некоторой оценкой, полученной для описанных выше условий.

Кроме того, из полученных данных видно, что количество пакетов при бездействии мало в сравнении с работой, что обусловлено спецификой работы протокола доставки удаленного рабочего стола: в отсутствии действий пользователя данные от сервера к клиенту практически не передаются. Это говорит о том, что сервер практически не порождает пакеты в отсутствии действий со стороны пользователя.

4.2.6 Исследование зависимостей времени отклика от коэффициентов вариации

Для исследования зависимостей времени отклика от коэффициентов вариации будем поступать следующим образом. Будем варьировать коэффициенты вариации распределения длительности обслуживания в узлах $C_B(i)$, затем по алгоритму УДН рассчитаем коэффициенты вариации распределения расстояний между вызовами входящих в узлы потоков $C_A(i)$. Подставим полученные пары КВ в формулу (4.16), по формулам (4.17) и (4.18) рассчитаем времена ожидания и обслуживания заявки на узлах, затем по (4.19) рассчитаем T' сети.

Для исследования влияния характера входящего в сеть потока проделаем эти расчеты для трех значений интенсивности входящего в узел 1 потока, для каждой рассмотрим по три значения КВ. Результаты сведем в таблицы П.17 – П.19 Приложения 1. На рисунках 4.27 – 4.29 показаны графики полученных зависимостей.

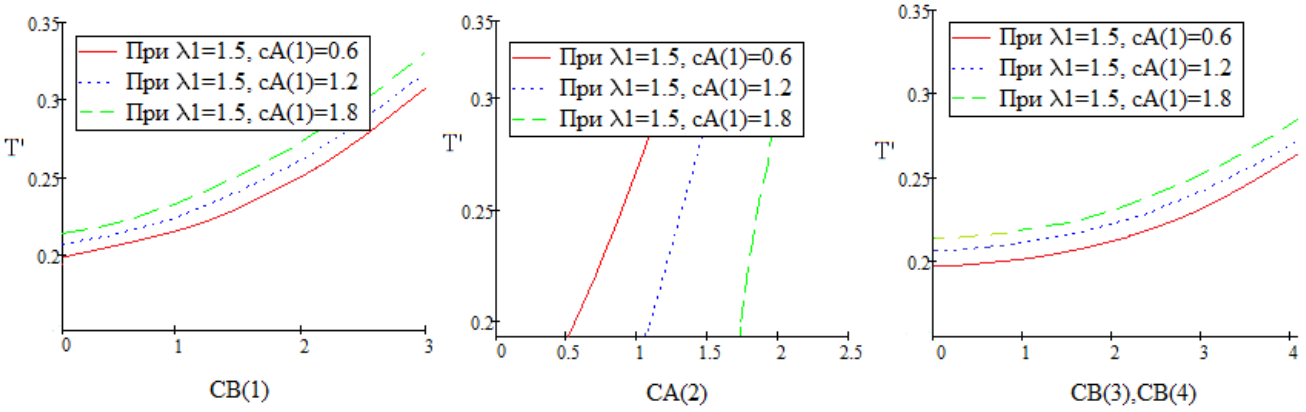

Рис. 4.27. Зависимости времени отклика от КВ при λ_1 = 1.5 1/с и различных КВ входящего в сеть потока

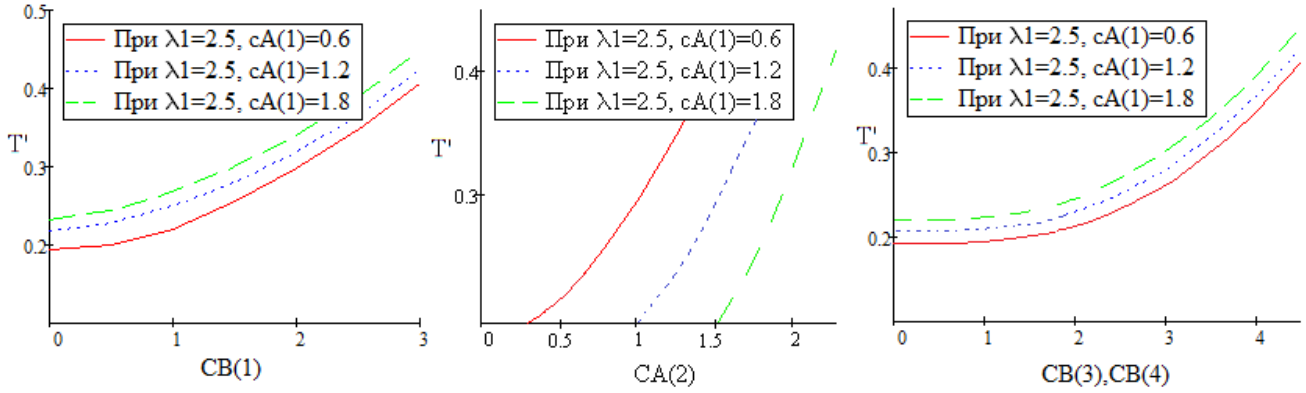

Рис. 4.28. Зависимости времени отклика от КВ при λ_1 = 2.5 1/с и различных КВ входящего в сеть потока

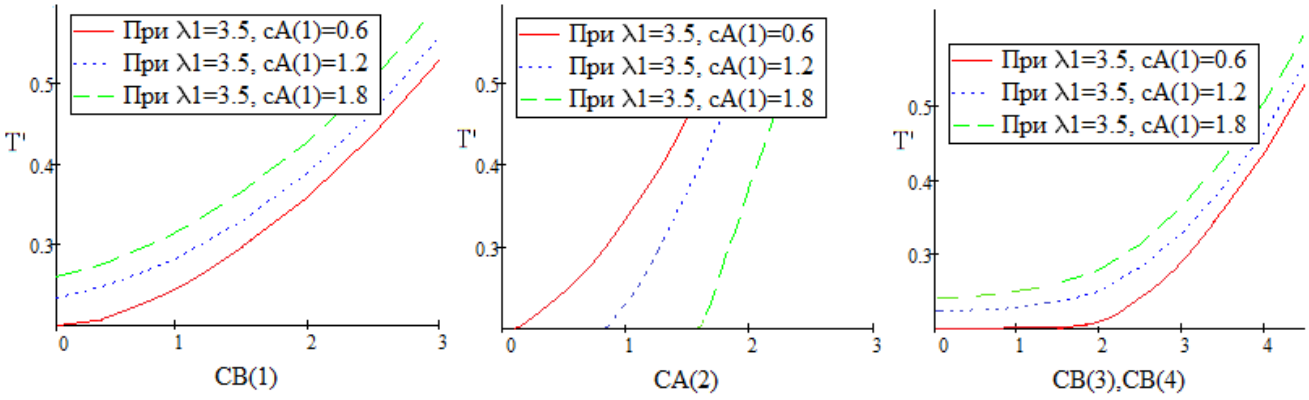

Рис. 4.29. Зависимости времени отклика от КВ при λ_1 = 3.5 1/с и различных КВ входящего в сеть потока

Характер семейства кривых (рисунки 4.27 – 4.29), полученных при фиксированной величине КВ, относящегося к потоку, входящему в сеть, и варьировании КВ, относящихся к обслуживанию в узлах сети, позволяет сделать вывод о том, что среднее время отклика увеличивается с ростом КВ распределения расстояния между соседними заявками в потоках и КВ распределения времени обслуживания в узлах сети. С увеличением интенсивности входящего в сеть потока в пределах 1.5 .. 3.5 $с^{-1}$ величина среднего времени отклика увеличивается на 0.1 с.

Для того, чтобы оценить разницу между моделью на основе сети Джексона, которая предполагает моделирование обслуживания в узлах и характеров потоков экспоненциальными, и более точной, но более трудоемкой в расчетах моделью, рассматриваемой в данном параграфе, проведем следующие вычисления.

Рассчитаем последовательно:

а. среднее время отклика T'_a при КВ $c_A(i) = c_в(i) = 1$.

При КВ $c_A(i) = c_в(i) = 1$ величина T'_a составляет 0.228 с.

б. среднее время отклика $T'_б$ при КВ, оцененными экспериментально в параграфе 4.2.5 и сравним с T'_a.

Вычислим разницу $\Delta_{ба} = T'_б - T'_а$.

При КВ $c_в(1)$, $c_в(2)$, $c_A(2)$, оцененными экспериментально, $T'_б$ составляет 0.256 с. Тогда $\Delta_{ба} = T'_б - T'_а = 0.253 - 0.228 = 0.025$ с, что составляет 10 %.

Объем экспериментально полученных данных, использованных при расчетах, ограничен. В других ситуациях в реальных условиях расхождения могут быть гораздо выше, тогда для оценки среднего времени отклика обобщенная модель оказывается весьма полезной.

4.3 Рекомендации по применению представленных моделей

Аналитическая модель, представленная в параграфе 4.1.2, описывает работу терминальной сессии в базовом режиме. Она моделирует распространенный на практике вариант, когда провайдер разворачивает услугу «из коробки»: базовые настройки на стороне сервера и обычное терминальное устройство. Такой вариант работы услуги, как было описано выше, подходит для той категории пользователей, которые занимаются стандартной офисной работой (работа с документами, браузером, файловым менеджером).

Аналитическая модель, представленная в параграфе 4.1.3, описывает работу терминальной сессии в режиме, при котором пользователю доступен просмотр видео. Такой вариант работы услуги подходит для пользователей, просматривающих видео в рамках рабочего стола. Провайдер может воспользоваться данной моделью при проектировании услуги, если предполагает наличие таких пользователей.

Аналитическая модель, представленная в параграфе 4.1.4, описывает работу терминальной сессии в режиме с аппаратным ускорением. Как было показано выше, в данном режиме пользователь работает с файлами, просматривает видео и имеет возможность прослушивать аудио. На практике этот режим реализуется за счет аппаратного ускорения, принцип которого изложен в описании модели. Работа по такой модели присуща пользователям, работающим с мультимедиа-контентом, графическими редакторами, играми. Провайдеру рекомендовано оценивать сетевые и инфраструктурные ресурсы, используя данную модель, если он предполагает подключение таких пользователей.

Представленная в параграфе 4.2 модель выполнена в соответствии с иным подходом, основанном на приближенном методе с учетом первых двух моментов. Первые три модели предполагают моделирование узлов услуги при помощи марковских СМО, а моделирование потоков, циркулирующих между ними – при помощи пуассоновских потоков. При этом коэффициенты вариации распределения времени обслуживания и расстояния между соседними заявками равны 1.

Математический аппарат первых трех моделей более прост по сравнению с обобщенной моделью. Исходя из этого, перечисленные модели могут быть использованы провайдером услуги в тех ситуациях, когда тому нужно быстро, дешево, но достаточно эффективно оценить характеристики услуги. Аппарат немарковской модели более сложен и требует как дополнительного времени на расчеты, так и более квалифицированных специалистов. Однако, данная модель более точна, поскольку позволяет расширить диапазон охвата первых трех,

рассматривая СМО с произвольным распределением времени обслуживания в узлах и непуассоновские потоки.

4.4 Рекомендации провайдеру по обеспечению приемлемого качества услуги

На основании проведенного аналитического моделирования, а также экспериментальных исследований параметров услуги уместно предложить следующие рекомендации по обеспечению качества услуги «виртуальный рабочий стол».

Анализ параметров качества функционирования сети (NP), которые являются основой для QoS, показал, что для работы пользователя в виртуальном рабочем столе уместно придерживаться следующих рекомендаций. Диапазон приемлемой транспортной задержки составляет 120 .. 150 мс. Для обеспечения приемлемой задержки (на уровне 150 мс), целесообразно использовать следующие значения скорости передачи данных: не менее 5 Мбит/с при работе с файлами; не менее 15 Мбит/с при работе с браузером; не менее 25 Мбит/с при работе с видео.

Обеспечение качества на этапе установления терминальной сессии.

- Оценка времени отклика, полученная в результате анализа предложенной аналитической модели, показала, что для комфортной работы на этапе установления сессии его значение не должно превышать 1.5 с.
- Расчет количества обслуживаемых пользователей услуги показал, что провайдеру достаточно ограничиться рассмотрением не всего множества допустимых значений параметров (времени обслуживания одного запроса и количества одновременно обслуживаемых пользователей), а меньшего по числу элементов множества Парето. Варианты сочетания этих параметров могут быть подобраны на основе этого множества
- Полученные соотношения позволяют, исходя из заданных требований к качеству услуги, рассчитать время обслуживания и количество одновременно обслуживаемых пользователей услуги.

Обеспечение качества на этапе работы терминальной сессии. При рассмотренных условиях можно рекомендовать следующее.

- Оценка времени отклика, полученная в результате анализа предложенных аналитических моделей, показала, что на этапе установления сессии приемлемое среднее время отклика составляет: для режима офисной работы через виртуальный рабочий стол (базовый режим): 0.8 с; для режима с видео: 0.5 с; для режима с видео и аудио: 0.3 с.

- Предложенная обобщенная аналитическая модель позволяет учесть соотношения между временем отклика и коэффициентами вариации, относящимся к обслуживанию в узлах услуги и потоках.

4.5 Выводы по результатам четвертого раздела

Четвертый раздел посвящен анализу и моделированию процесса работы терминальной сессии услуги «виртуальный рабочий стол».

В п. 4.1.1 на основе специфики работы пользователей выделено три наиболее типичных сценария работы терминальной сессии услуги. В п. 4.1.2 предложена аналитическая модель первого (базового) сценария на основе сети Джексона, которая позволяет оценить среднее время отклика для случаев использования различных по производительности пользовательских устройств и облачной платформы.

В п. 4.1.3 предложена аналитическая модель второго сценария, при котором пользователю доступен просмотр видео внутри рабочего стола, выполненная на основе BCMP-сети с двумя типами заявок. Полученная модель позволяет описать работу применяемого в таких случаях на практике аппаратного ускорения – выделения видео в отдельный поток с последующим его аппаратным декодированием в специально приспособленном для этого процессорном ядре пользовательского устройства. Оценено среднее время отклика сети, получены его зависимости от времени обслуживания заявок в узлах инфраструктуры услуги.

В п. 4.1.4 предложена аналитическая модель третьего сценария, при котором пользователю доступен просмотр видео и прослушивание аудио. Модель, построенная на основе BCMP-сети с тремя типами заявок, позволяет оценить среднее время отклика.

П. 4.2 посвящен построению обобщенной модели терминальной сессии услуги. В п. 4.2.1 – 4.2.2 предложена модель, построенная на основе приближенного метода, учитывающего первые два момента случайных величин, описывающих расстояния между соседними заявками в потоках и длительность обслуживания в узлах. Используемый метод основан на формулах Крамера и Лангенбах-Бельца. В п. 4.2.3 экспериментально исследованы коэффициенты вариации, относящиеся к обслуживанию в узлах. В п. 4.2.4 при помощи алгоритма УДН рассчитаны коэффициенты вариации, относящиеся к потокам. В п. 4.2.5 экспериментально исследованы коэффициенты вариации потока, проходящего через сеть передачи данных.

П. 4.2.6 посвящен построению и анализу зависимостей среднего времени отклика от коэффициентов вариации, относящихся к потокам и обслуживанию в узлах. Получено выражение для среднего времени отклика для рассматриваемой сети.

Полученная модель является более трудоемкой, однако позволяет исследовать более широкий диапазон случаев. В п. 4.3 даны рекомендации по применению представленных моделей в зависимости от сетевых и серверных условий, а также потребностей провайдера услуги. В 4.4 сформулированы рекомендации провайдеру услуги.

Таким образом, по итогам четвертого раздела могут быть сформулированы следующие выводы.

1. В результате анализа полученных зависимостей среднего времени отклика (без учета транспортной задержки) от КВ, относящихся к обслуживанию и потокам, установлено следующее: влияние КВ распределения расстояний между соседними заявками в потоке на среднее время отклика более выражено, чем влияние КВ распределения времени обслуживания в узлах инфраструктуры.

2. Полученные данные показывают, что разница между моделью на основе сети Джексона и обобщенной моделью моделью составляет 10 % при следующих условиях: для первой модели все КВ равны единице, для второй – КВ обслуживания в узлах 1, 2 и КВ, относящиеся к потокам между узлами 1 и 2, оценены экспериментально; прочие – вычислены аналитически.

3. Предложенная модель позволяет оценить величины среднего времени отклика в тех случаях, когда характеры потоков между элементами облачной инфраструктуры услуги «виртуальный рабочий стол» являются непуассоновскими. Предлагается производить оценку на основе приближенного метода, учитывающего первый и второй моменты.

Исходя из требований к среднему времени отклика провайдер услуги, оператор связи, разработчик облачного ПО может корректировать обслуживание в узлах инфраструктуры услуги с учетом характера потоков, подаваемых на ее вход и циркулирующих внутри нее. Корректировка может быть осуществлена за счет установки более производительного оборудования на серверной и клиентской сторонах (процессор, оперативная память). Кроме того, в некоторых случаях могут быть проведены работы над кодом программных компонентов, располагаемых в узлах инфраструктуры.

ЗАКЛЮЧЕНИЕ

Основные результаты работы состоят в следующем.

1. На основании анализа логики услуги «виртуальный рабочий стол» для разработки математических моделей выделены две фазы ее предоставления для возможности отдельного их исследования. В первой фазе рассмотрено подключение пользователей; во второй фазе предусмотрена их работа с индивидуальными рабочими столами.
2. Для первой фазы разработана аналитическая модель, позволяющая оценить среднее время отклика; получены его зависимости от основных характеристик системы (среднего времени обслуживания одного запроса, числа одновременно обслуживаемых пользователей). Решена задача определения множества допустимых значений характеристик сервера, при которых выполняются ограничения по среднему времени отклика и вероятности отказа в подключении, а также задача определения рациональных вариантов сочетания этих параметров.
3. Для второй фазы разработаны аналитические модели, которые для трех сценариев предоставления услуги позволяют оценить среднее время отклика, а также получить аналитические соотношения между средним временем отклика и интенсивностями обслуживания. Предложена обобщенная модель базового сценария предоставления услуги, которая позволяет оценить среднее время отклика для различных типов потоков и законов распределения времени обслуживания.
4. В результате проведенных натурных экспериментальных исследований получены оценки характеристик инфраструктуры услуги, влияющих на ее качество: среднего времени между запросами к серверу в обеих фазах, среднего времени обработки на пользовательском устройстве, среднего времени обслуживания запросов сервером, среднего время отклика, а также зависимости транспортной задержки от скорости передачи данных.
5. В общем случае для оценки времени отклика рекомендуется использовать метод, учитывающий первые два момента случайных величин, описывающих расстояния между соседними заявками в потоках и длительность обслуживания в узлах. В случаях, когда коэффициенты вариации распределения времени обслуживания в узлах и расстояния между соседними запросами в потоке отличны от единицы не более, чем на 10%, рекомендуется использовать более простой метод, основанный на модели сети Джексона.
6. Результаты исследования использованы в работе ООО «ЭЛТЕКС-МСК» в виде методики оценки среднего времени отклика в различных сценариях работы услуги, а также в учебном процессе кафедры СС и СК МТУСИ, что подтверждается соответствующими актами.

СПИСОК ЛИТЕРАТУРЫ

1. Башарин, Г. П. Анализ очередей в вычислительных сетях. Теория и методы расчета [Текст] / Г.П. Башарин, П.П. Бочаров, Я.А. Коган. – М.: Наука, 1989. – 336 с.

2. Бойченко, И. В. Управление ресурсами в сервис-ориентированных системах типа «Приложение как сервис» [Текст] / И. В. Бойченко, С. В. Корытников // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2010. – №. 1-2(21). – С. 156–160.

3. Бочаров, П. П. Теория массового обслуживания [Текст] / П. П. Бочаров, А. В. Печинкин. – М.: РУДН, 1995. – 529 с.

4. Вислоцкий, И. Будущее VDI в России: варианты применения, требования к инфраструктуре, преимущества, экономичность, области применения [Текст] / И. Вислоцкий // Connect WIT. – 2017. – № 1-2. – С. 86–89.

5. Вишневский, В.М. Теоретические основы проектирования компьютерных сетей [Текст] / В.М. Вишневский. – М.: Техносфера, 2003. – 512 с.

6. Высокая производительность и доступность приложений [Электронный ресурс] / VMware // – Режим доступа: http://www.vmware.com/ru/products/vsphere/features/vmotion. – (дата обращения 06.10.2016).

7. Гнеденко, Б. В. Введение в теорию массового обслуживания [Текст] / Б.В. Гнеденко, И.Н. Коваленко. – 2-е изд., перераб. и доп. – М.: Наука, 1987. – 336 с.

8. Горбунова, А. В. Преобразование Лапласа-Стилтьеса для времени отклика системы облачных вычислений с гистерезисным управлением и ограничением на одновременное число активаций [Текст] / А. В. Горбунова, К. Е. Самуйлов, Э. С. Сопин // Международный научный журнал «Современные информационные технологии и ИТ-образование». – 2016. – Т. 12. – №. 1. – С. 21–27.

9. Двоеглазов, Д. В. Инфраструктура виртуальных рабочих столов на открытых программных продуктах [Текст] / Д. В. Двоеглазов, И. П. Дешко, К. Г. Кряженков, А. А. Тихонов // Интернет-журнал Науковедение. – 2015. – №4(29) – С. 68.

10. Ефимушкин, В. А. Моделирование процессов управления голосовыми вызовами в сетях LTE с использованием технологии CSFB в GSM [Текст] / В. А. Ефимушкин, И. В. Углов // Труды XII Всероссийского совещания по проблемам управления ИПУ РАН. – Москва: ИПУ РАН. – 2014. – Т. 16. – С. 19.

11. Кантышев, П. Облачные услуги в России принесли 22,6 млрд рублей, показав рост в 43% [Текст] / П. Кантышев // Ведомости. – 2017. – 22 февр.

12. Кауфман, Е. А. Анализ существующих видов терминального доступа и инфраструктуры VDI [Текст] / Е. А. Кауфман // Актуальные вопросы экономических наук. – 2013. – №34. – С. 259–276.

13. Клейнрок, Л. Вычислительные системы с очередями [Текст] / Л. Клейнрок; пер. с англ. под ред. Б. С. Цыбакова. – М.: Мир, 1979. – 600 с.

14. Клейнрок, Л. Теория массового обслуживания [Текст] / Л. Клейнрок; пер. с англ. под. ред. В. И. Неймана. – М.: Машиностроение, 1979. – 432 с.

15. Лихтциндер, Б. Я. Интервальный метод определения задержек в одноприборных СМО с потоками заявок общего вида [Текст] / Б.Я. Лихтциндер // Инфокоммуникационные технологии. – 2016. – Т. 14. – №. 3. – С. 274–278.

16. Лихтциндер, Б. Я. Дистанционный анализ трафика с помощью утилиты Team Viewer и программы WireShark [Текст] / Б.Я. Лихтциндер, Л.А. Сарычев // Материалы V Международной научной конференции «Технические науки: проблемы и перспективы». – Спб: Свое издательство. – 2017. – С. 1–3.

17. Мартин, Дж. Системный анализ передачи данных [Текст] / Дж. Мартин; пер. с англ. под ред. В.С. Лапина. – М.: Мир, 1975. – Т.1. – 256 с.

18. Нетес, В. А. Что нужно для успешного применения SLA [Текст] / В.А. Нетес // T-Comm: Телекоммуникации и транспорт. – 2015. – Т. 9. – №7. – С. 16–20.

19. Нетес, В.А. Виртуализация, облачные услуги и надежность [Текст] / В.А. Нетес // Вестник связи. – 2016. – № 8. – С. 7–9.

20. Нетес, В. А. Услуга «виртуальный рабочий стол» и особенности ее реализации [Текст] / В.А. Нетес, А. А. Сулейманов // Вестник Связи. – 2016. – №9. – С. 12–16.

21. Ногин, В. Д. Принятие решений в многокритериальной среде: количественный подход [Текст] / В.Д. Ногин. – М.:Физматлит. – 2002. – 144 с.

22. Обзор, установка, настройка и использование открытой системы виртуализации Xen [Электронный ресурс] / IBM // – Режим доступа: http://www.ibm.com/developerworks/ru/library/l-xen-citrix. – (дата обращения 06.10.2016).

23. Облачные сервисы. Рынок в России [Электронный ресурс] / Tadviser // – Режим доступа: http://www.tadviser.ru/index.php/Статья%3AОблачные_сервисы_(рынок_России)#2017:_.D0.9F.D1.80.D0.BE.D0.B3.D0.BD.D0.BE.D0.B7_SAP_.D0.B8_Forrester_Russia. – (дата обращения 06.10.2016).

24. Соловьев, А. Д. Анализ системы M/G/1/∞ для различных дисциплин обслуживания [Текст] / А. Д. Соловьев // Теория массового обслуживания. – М.: ВНИИСИ. –1981. – С. 172–178.

25. Сопин, Э. С. Анализ показателей качества функционирования систем облачных вычислений с гистерезисным управлением [Текст] / Э. С. Сопин, М. О. Таланова, Ю. В. Гайдамака // T-Comm: телекоммуникации и транспорт. – 2015. – №9. – С. 54–60.

26. Сопин, Э. С. О задаче минимизации стоимости для системы облачных вычислений с гистерезисным управлением [Текст] / Э. С. Сопин, М. О. Таланова, Ю. В. Гайдамака // Сборник трудов X международной отраслевой научно-технической конференции «Технологии информационного сообщества». – 2016. – С. 65–66.

27. Специализированный аппаратный гипервизор VMware ESXi [Электронный ресурс] / VMware // – Режим доступа: http://www.vmware.com/ru/products/esxi-and-esx.html. – (дата обращения 06.10.2016).

28. Спецификация Microsoft на графические расширения RDP [Электронный ресурс] / Microsoft // – Режим доступа: http://msdn.microsoft.com/en-us/library/cc241537(PROT.10).aspx.– (дата обращения 06.10.2016).

29. Спецификация Microsoft на основные функции RDP [Электронный ресурс] / Microsoft // – Режим доступа: http://msdn.microsoft.com/en-us/library/cc240445(PROT.10).aspx. – (дата обращения 06.10.2016).

30. Спецификация протокола RFB [Электронный ресурс] // – Режим доступа: http://www.realvnc.com/docs/rfbproto.pdf – (дата обращения 06.10.2016).

31. Спецификация тонкого клиента HP T410 // [Электронный ресурс] / Hewlett Packard // – Режим доступа: http://www8.hp.com/us/en/thin-clients/t410.html – (дата обращения 06.10.2016).

32. Спецификация тонкого клиента Eltex TC-20 [Электронный ресурс] / Eltex // – Режим доступа: http://eltex.nsk.ru/upload/iblock/22d/tc_datasheet_16.pdf – (дата обращения 06.10.2016).

33. Спецификация системы Citrix XenDesktop [Электронный ресурс] / Citrix // – Режим доступа: https://www.citrix.ru/products/xenapp-xendesktop/compare.html – (дата обращения 06.10.2016).

34. Сулейманов, А. А. Существующие возможности различных технологий серверной виртуализации [Текст] / А .А. Сулейманов // Материалы научно - технической конференции «INTERMATIC-2013». – М.: МГТУ МИРЭА – ИРЭ РАН. – 2013. – Ч.5. – С. 21–23.

35. Сулейманов, А. А. Воспринимаемое качество при использовании тонкого клиента на базе облачных платформ [Текст] / А. А. Сулейманов // Труды Международной молодежной научно-практической конференции СКФ МТУСИ «ИНФОКОМ-2014». – Ростов-на-Дону: «Университет». – 2014. – Ч.1. – С. 124 –126.

36. Сулейманов, А. А. Качество восприятия услуг на базе тонкого клиента при подключении к различным облачным платформам [Текст] / А. А. Сулейманов // Материалы научно - технической конференции «INTERMATIC-2014». – М.: МГТУ МИРЭА – ИРЭ РАН. – 2014. – Ч.5. – С. 229–232.

37. Сулейманов, А. А. Факторы, определяющие воспринимаемое качество при подключении тонкого клиента к облачным платформам [Текст] / А. А. Сулейманов // Труды конференции «Телекоммуникационные и вычислительные системы» (МФИ-2014). – М: МТУСИ. – 2014. – С. 53–54.
38. Сулейманов, А. А. Качество потокового видео в облачных услугах типа DaaS [Текст] / А. А. Сулейманов // Труды Международной молодежной научно-практической конференции СКФ МТУСИ «ИНФОКОМ-2014». – Ростов-на-Дону: «Университет». – 2015. – Ч.1. – С. 282–284.
39. Сулейманов, А. А. Качество облачных услуг типа «виртуальный рабочий стол» [Текст] / А. А. Сулейманов // Тезисы научно-технических секций IX международной отраслевой научно-технической конференции «Технологии информационного сообщества». – М: ИД Медиа Паблишер. – 2015. – С. 35.
40. Сулейманов, А. А. Качество облачных услуг типа «виртуальный рабочий стол» [Текст] / А. А. Сулейманов // T-Comm: Телекоммуникации и транспорт. – 2015. – Т. 9. – №7. – С. 31–35.
41. Сулейманов, А. А. Аналитическая модель процесса установления сессии облачной услуги типа «виртуальный рабочий стол» [Текст] / А. А. Сулейманов // Материалы 11-ой международной научно-технической конференции. – Владимир: ВлГУ. – 2015. – 376 с.
42. Сулейманов, А. А. Анализ процесса установления терминальной сессии услуги типа «виртуальный рабочий стол» [Текст] / А. А. Сулейманов // Труды конференции «Телекоммуникационные и вычислительные системы» (МФИ-2015). – М: МТУСИ – 2015. – С. 44–45.
43. Сулейманов, А. А. Средства доставки виртуальных рабочих столов на терминальное оборудование [Текст] / А. А. Сулейманов // Сборник тезисов IX Московской научно-практической конференции «Студенческая наука». – М.: «Московский студенческий центр». – 2016. – Т. 3. – С. 702.
44. Сулейманов, А. А. Анализ времени подключения к облачной услуге «виртуальный рабочий стол» [Текст] / А. А. Сулейманов, В. А. Нетес // Сборник трудов X международной отраслевой научно-технической конференции «Технологии информационного сообщества». – М: ИД Медиа Паблишер. – 2016. – С. 69–70.
45. Сулейманов, А. А. Анализ времени подключения к облачной услуге «виртуальный рабочий стол» [Текст] / А. А. Сулейманов, В. А Нетес // T-Comm: Телекоммуникации и транспорт. – 2016. – Т. 10. – №7. – С. 41–46.
46. Сулейманов, А. А. Применение аппаратного ускорения для доставки мультимедиа в терминальную сессию услуги «виртуальный рабочий стол» [Текст] / А. А. Сулейманов // Труды конференции «Телекоммуникационные и вычислительные системы» (МФИ-2016). – М.: МТУСИ. – 2016. – С. 35–36.

47. Сулейманов, А. А. Моделирование подсистемы терминальной сессии услуги типа «виртуальный рабочий стол» [Текст] / А. А. Сулейманов // Материалы научно - технической конференции «INTERMATIC-2016». – М.: МГТУ МИРЭА – ИРЭ РАН. – 2016. – Ч.5. – С. 195–198.

48. Сулейманов, А. А. Немарковская модель терминальной сессии облачной услуги «виртуальный рабочий стол» [Текст] / А. А. Сулейманов // Сборник трудов XI международной отраслевой научно-технической конференции «Технологии информационного сообщества». – М.: МТУСИ. – 2017. – С. 105.

49. Сулейманов, А. А. Немарковская модель терминальной сессии облачной услуги «виртуальный рабочий стол» [Текст] / А. А. Сулейманов // T-Comm: Телекоммуникации и транспорт. – 2017. – Т. 2. – № 4. – С. 72–75.

50. Тихоненко, О. М. Система обслуживания с разделением процессора и ограниченными ресурсами [Текст] / О. М. Тихоненко // Автоматика и телемеханика. – 2010. – №5. – С. 84–98.

51. Черняк Л. Интеграция – основа облака [Текст] / Л. Черняк // Открытые системы. СУБД. – 2011. – Т. 16. – № 7. С. 12–20.

52. Шварцман, О. И. О выборе числа каналов и объема памяти в узле коммутации [Текст] / О. И. Шварцман, В. А. Нетес // Электросвязь. – 1990. – № 2. – С. 20–23.

53. Шнепс, М. А. Системы распределения информации. Методы расчета: Справочное пособие [Текст] / М. А. Шнепс. – М.: Связь, 1979. – 344 с.

54. Шурыгин, В. Н. Технология VDI в информационных системах [Текст] / В. Н. Шурыгин, Н. М. Кабина // Вестник МГУП. – 2015. – №5. – С. 107–108.

55. Яшков, С.Ф. Анализ очередей в ЭВМ [Текст] / С.Ф.Яшков. – М.:Радио и связь, 1989. – 216 с.

56. Яшков, С.Ф. Эгалитарное разделение процессора [Текст] / С.Ф. Яшков, А.С. Яшкова // Информационные системы. – 2006. – Т.6. – №4. – С. 396–444.

57. Antonopoulos, N. Cloud computing / N. Antonopoulos, L. Gillam. – Springer, 2010. – 379 p.

58. Baratto, R. THINC: A remote display architecture for thin-client computing / R. Baratto, J. Nieh, L. Kim // Columbia University. – Department of Computer Science – Technical Report № CUCS-027-04. – 2004.

59. Baskett, F. Open, closed and mixed networks of queues with different classes of customers / F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios // Journal of the ACM. – 1975. – Vol. 22. – No. 2. – P. 248–260.

60. Buck, K. Best Desktop as a Service Review [Электронный ресурс] / K. Buck // Top reviews. – 2016. – Режим доступа: http://www.toptenreviews.com/business/articles/best-desktop-as-a-service-review/ – (дата обращения 06.10.2016).

61. Burke, P. J. The output of a queuing system / P. J. Burke // Operations research. – 1956. – Vol. 4. – No. 6. – P. 699–704.
62. Cao, J. Web server performance modeling using an M/G/1/K* PS queue / J. Cao, M. Andersson, C. Nyberg // IEEE Telecommunications ICT 10th International Conference. – 2003. – Vol. 2. – P. 1501–1506.
63. De Winter, D. A hybrid thin-client protocol for multimedia streaming and interactive gaming applications / D. De Winter, P. Simoens, L. Deboosere, F. De Turck, J. Moreau, B. Dhoedt, P. Demeester // Proceedings of the 2006 international workshop on Network and operating systems support for digital audio and video. – ACM. – 2006. – P. 15.
64. Dusi, M. A closer look at Thin-Client connections: Statistical Application Identification for QoE Detection / M. Dusi, S. Napolitano, S. Longo, S. Niccolini // IEEE Communications Magazine. – 2012. – Vol. 50. – No.11. – P. 195–202.
65. FreeBSD Manual Pages [Электронный ресурс] / FreeBSD // – Режим доступа https://www.freebsd.org/cgi/man.cgi?query=iftop&apropos=0&sektion=8&manpath=FreeBSD+8.1-RELEASE+and+Ports&format=html. – (дата обращения: 06.10.2016).
66. ITU-T Recommendation E.430 Telephone network and ISDN quality of service, network management and traffic engineering. Quality of service framework. – Geneva – 1992. – 3 p.
67. ITU-T Recommendation E.800: Definitions of terms related to quality of service. – Geneva. – 2009. – 32 p.
68. ITU-T FG Cloud Technical Report Part 2. – Geneva. – 2012. – 33 p.
69. ITU-T Recommendation G.1000 Communications quality of service: A framework and definitions. – Geneva. – 2002. – 16 p.
70. ITU-T Recommendation G.1010 End-user multimedia QoS categories. – Geneva. – 2002. – 18 p.
71. ITU-T Recommendation I.350 General aspects of quality of service and network performance in digital networks, including ISDNs. – Geneva. – 1993. – 13 p.
72. ITU-T Recommendation P.10/G.100 Vocabulary for performance and quality of service. Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100. – Geneva. – 2007. – 34 p.
73. ITU-T Recommendation P.10/G.100 Vocabulary for performance and quality of service. Amendment 3: New definitions for inclusion in Recommendation ITU-T P.10/G.100. – Geneva. – 2012. – 3 p.
74. ITU-T Recommendation Y.1540 Internet protocol data communication service – IP packet transfer and availability performance parameters. – Geneva. – 2011. – 40 p.
75. ITU-T Recommendation Y.1541 Network performance objectives for IP-based services. – Geneva. – 2012. – 57 p.

76. ITU-T Recommendation Y.3500 Information technology – Cloud computing – Overview and vocabulary. – Geneva. – 2014. – 10 p.
77. ITU-T Recommendation Y.3503 Requirements for desktop as a service. – Geneva. – 2014. – 25 p.
78. ITU-T Technical Report Focus Group on Cloud Computing Part 1: Introduction to the cloud ecosystem: definitions, taxonomies, use cases and high level requirements. – Geneva. – 2012. – 62 p.
79. ITU-T Technical Report Focus Group on Cloud Computing Part 2: Functional requirements and reference architecture. – Geneva. – 2013. – 33 p.
80. ITU-T Technical Paper How to increase QoS/QoE of IP-based platform(s) to regionally agreed standards – Geneva. – 2013. – 55 p.
81. Ghosh, R. End-to-end performability analysis for infrastructure-as-a-service cloud: An interacting stochastic models approach / R. Ghosh, K.S. Trivedi, V. Naik, D. S. Kim // Dependable Computing (PRDC) – IEEE 16th Pacific Rim International Symposium. – 2010. – P. 125–132.
82. Huawei FusionCloud Desktop Access Software [Электронный ресурс] / Huawei // – Режим доступа: http://e.huawei.com/en/products/cloud-computing-dc/cloud-computing/fusionaccess. – (дата обращения 18.05.2017).
83. Jacob, B. On Demand Operating Environment: Managing the Infrastructure (Virtualization Engine Update) / B. Jacob, S. Mui, J. Pannu, S. Park, H. Raguet, J. Schneider, L. Vanel // IBM Redbooks. – 2005.
84. Jacob, P. Technical challenges in the delivery of interprovider QoS / P. Jacob, B. Davie // IEEE Communications Magazine. – 2005. – Vol. 43. – No. 6. – P. 112–118.
85. Kameda, H. Uniqueness of the solution for optimal static routing in open BCMP queueing networks / H. Kameda, Y. Zhang // Mathematical and Computer Modelling. – 1995. – Vol. 22. – No 10. – P. 119–130.
86. Kleinrock, L. Time-shared system: a theoretical treatment // Journal of the ACM. – 1967. – Vol. 14. – No. 2. – P. 242–251.
87. Kramer, W. Approximate formulae for the delay in the queueing system GI/G/1 / W. Kramer, M. Langenbach-Belz // Proceedings of the 8th International Teletraffic Congress. – 1976. – Vol. 8. – P. 235.1-235.8.
88. KVM (for Kernel-based Virtual Machine) [Электронный ресурс] // – Режим доступа: http://www.linux-kvm.org/page/Main_Page. – (дата обращения 06.10.2016).
89. Lai, A. M. On the performance of wide-area thin-client computing / A. M. Lai, J. Nieh // ACM Transactions on Computer Systems (TOCS). – 2006. – No 24. – P. 175–209.
90. Licklider, J. C. R. On-line man-computer communication / J. C. R Licklider, W. E. Clark // Proceedings of the ACM spring joint computer conference. – 1962. – P. 113–128.

91. Lamanna, D. SLAng: a language for defining service level agreements / D. Lamanna, J. Skene, W. Emmerich // IEEE Computer Society Press. – 2003. – P. 100–106.
92. McCarthy, J. A basis for a mathematical theory of computation // Studies in Logic and the Foundations of Mathematics. – 1963. – Vol. 35. – P. 33–70.
93. Mell, P. The NIST Definition of Cloud Computing [Электронный ресурс] / P. Mell, T. Grance // – Режим доступа http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf – (дата обращения 06.10.2016).
94. Mohamed, A. History of cloud computing [Электронный ресурс] / A. Mohamed // Computer Weekly. – Режим доступа http://www.computerweekly.com/feature/A-history-of-cloud-computing – (дата обращения 06.10.2016).
95. Nieh, J. A comparison of thin-client computing architectures / J. Nieh, S. J. Yang, N. Novik //, Department of Computer Science. – Columbia University. – 2000. – 16 p.
96. PCoIP technology – multimedia redirection [Электронный ресурс] // – Режим доступа: http://www.shopmanda.com/public/PC%20CANS/PCoIP%20Tools/MMR%20Considerations.pdf – (дата обращения 06.10.2016).
97. Philippe, N. Basic elements of queueing theory application to the modelling of computer systems / N. Philippe // Le Chesney. – INRIA. – 1998. – 109 p.
98. Qin, X. Communication-Aware Load Balancing for Parallel Applications on Clusters / X. Qin, H. Jiang, A. Manzanares, X. Ruan, S. Yin // IEEE Transactions on computers. – 2010. – Vol. 59. – No. 1. P. 42–52.
99. Richardson, T. Virtual network computing / T. Richardson, Q. Stafford-Fraser, K. R. Wood, A. Hopper // IEEE Internet Computing. – 1998. – Vol. 2. – No. 1. – P. 33–38.
100. RFC 2679 A One-way Delay Metric for IPPM. – New-York – 1999. – 20 p.
101. RFC 3393 IP Packet Delay Variation Metric for IP Performance Metrics (IPPM). – Ericsson IPI. – 2002. – 21 p.
102. Server and Cloud Computing [Электронный ресурс] / Microsoft // – Режим доступа: http://www.microsoft.com. – (дата обращения 06.10.2016).
103. Shneiderman, B. Response time and display rate in human performance with computers // ACM Computing Surveys (CSUR). – 1984. – Vol. 16. – No 3. – P. 265–285.
104. Spice for newbies [Электронный ресурс] // – Режим доступа: http://www.spice-space.org/static/docs/spice_for_newbies.pdf. – (дата обращения 06.10.2016).
105. Spice remote computing protocol definition v1.0 [Электронный ресурс] // – Режим доступа: https://www.spice-space.org/static/docs/spice_protocol.pdf – (дата обращения 06.10.2016).

106. SPICE: an open remote computing solution [Электронный ресурс] / Red Hat // – Режим доступа: http://www.spice-space.org/static/docs/spice_redhat_summit_2009.pdf (дата обращения 06.10.2016).
107. Starosolski, R. Simple fast and adaptive lossless image compression algorithm/ R. Starosolski // Software: Practice and Experience. – 2007. – Vol. 37. – No. 1. – P. 65–91.
108. Storer, J. A. Data compression via textual substitution / J. A. Storer, T. Szymanski // Journal of the ACM (JACM). – 1982. – Vol. 29. – No. 4. – P. 928–951.
109. Tan, K. A remote thin client system for real time multimedia streaming over VNC / K. Tan, J. Gong, B. Wu, D. Chang // Multimedia and Expo (ICME) IEEE International Conference. – 2010. – P. 992–997.
110. Technical white paper: System-on-a-Chip power for Citrix clients [Электронный ресурс] // – Режим доступа: http://h20195.www2.hp.com/V2/GetPDF.aspx/4AA4-6569ENW.pdf. – (дата обращения 06.10.2016).
111. Technical white paper: Unleash the client HP t310 Zero Client [Электронный ресурс] // – Режим доступа: http://www8.hp.com/us/en/pdf/t310-whitepaper_tcm_245_1389301.pdf. – (дата обращения 06.10.2016).
112. The CELT ultra-low delay audio codec [Электронный ресурс] // – Режим доступа: http://www.celt-codec.org. – (дата обращения 06.10.2016).
113. Tolosana-Calasanz, R. Enforcing QoS in scientific workflow systems enacted over cloud infrastructures / R. Tolosana-Calasanz, J. Banares, C. Pham, O.F. Rana // Journal of Computer and System Sciences. – 2012. – Vol. 78. – No. 5. – P. 1300–1315.
114. Tolia, N. Quantifying interactive user experience on thin clients / N. Tolia, D. Andersen, M. Satyanarayanan // IEEE Computer Society. – Carnegie Mellon University. – 2006. – Vol. 39. –No. 3. – P. 46–52.
115. Tools for planning or troubleshooting virtual desktop or remote workstation deployments using PCoIP protocol [Электронный ресурс] / Teradici // – Режим доступа: https://techsupport.teradici.com/ics/support/DLList.asp?task=download&folderID=6500. – (дата обращения 06.10.2016).
116. Virtamo, J. Queueing Course. Complete lecture notes [Электронный ресурс] / J. Virtamo // – 2005. – Режим доступа: http://www.netlab.tkk.fi/opetus/s383143/kalvot/E_qnets.pdf. – (дата обращения 06.10.2016).
117. Wireshark [Электронный ресурс] // – Режим доступа: https://www.wireshark.org/docs /wsug_html_chunked/ – (дата обращения 01.03.2017).

118. Wu, L. SLA-based admission control for a Software-as-a-Service provider in Cloud computing environments / L. Wu, S. K.Garg, R. Buyya // Journal of Computer and System Sciences. – 2012. – Vol. 78. – No. 5. – P. 1280–1299.

119. Yashkov, S.F. Processor-sharing queues: some progress in analysis/ S. F. Yashkov // Queueing Systems. – 1987. – Vol. 2. – No. 1. – P. 1–17.

120. Yashkov, S.F. The moments of the sojourn time in the M/G/1 processor sharing system / S.F. Yashkov // Информационные системы. – 2006. – Т. 6. – № 3. – С. 237–249.

121. Zheng, Z. QoS ranking prediction for cloud services / Z. Zheng, X. Wu, Y. Zhang // IEEE transactions on parallel and distributed systems. – 2013. – Vol. 24. – No 6. – PP. 1213–1222.

ПРИЛОЖЕНИЕ 1

Результаты численных расчетов

Таблица П1. Измерение загрузки канала в процессе соединения вида клиент-сервер различных облачных платформ и выполнении пользователем различных действий

	Версия	Бездействие, Мбит/с	Работа с файлами, Мбит/с	Web-серфинг, Мбит/с	Просмотр видео в браузере, Мбит/с	Видео или RemoteFX, Мбит/с
Microsoft RDP	7.1	0.000208	2.6	12.1	16.6	-
	7.1(RFX)	0.000208	32.3	33.5	28.1	42.3
	8.0	0.000208	4.1	14.4	15.4	-
	8.0(RFX)	0.000208	11.1	30.7	17.0	46.0
	8.1	0.000208	14.6	18.1	23.3	-
Citrix ICA	-	0.0012	2.0	3.3	4	6.8
Red Hat SPICE	-	0.0015	3.7	27.0	18.0	45.0
VMware PCoIP	-	0.000640	2.3	2.0	7.2	11.5

Таблица П2. Зависимость транспортной задержки от канальной скорости передачи данных

Задержка, с	Работа с файлами, Мбит/с	Веб-серфинг, Мбит/с	Просмотр видео, Мбит/с
0.0008	45.1	54.8	66.1
0.0009	43.2	52.8	63.3
0.001	40.2	50	59.9
0.002	36.1	55.5	56.3
0.003	34.3	43	54
0.004	30.1	39	49.7
0.005	28.1	38	47.9
0.01	24.6	34.4	44.6
0.02	19.8	28.8	39.8
0.03	16	26	36
0.04	14	23.9	34
0.05	12	22.1	32
0.06	9	19	29
0.07	6.8	16.7	26.8
0.08	6	16.2	26
0.09	5	15	25
0.1	4.2	14.1	24.2
0.11	2.8	12.9	22.8
0.12	2	12	22
0.13	0.8	10.6	20.8
0.14	0.5	10.4	20.5
0.15	0.4	9.9	20.4
0.16	0.3	9.7	20.3
0.17	0.2	9.5	20.2
0.18	0.1	8.9	20.1
0.19	0.05	7.1	19.7
0.2	0.02	6.2	19.1

Таблица П3. Расчет среднего времени отклика для двух типов пользовательских устройств

При маломощных процессорах пользовательского устройства				При производительных процессорах пользовательского устройства			
μ_2, c^{-1}	μ_3, c^{-1}	μ_4, c^{-1}	T',c	μ_2, c^{-1}	μ_3, c^{-1}	μ_4, c^{-1}	T',c
5	5	5	0.792	5	15	15	0.598
5	10	10	0.639	5	20	20	0.578
5	15	15	0.598	5	25	25	0.567
10	5	5	0.525	10	15	15	0.331
10	10	10	0.373	10	20	20	0.312
10	15	15	0.331	10	25	25	0.3
15	5	5	0.472	15	15	15	0.278
15	10	10	0.319	15	20	20	0.258
15	15	15	0.278	15	25	25	0.247
20	5	5	0.449	20	15	15	0.255
20	10	10	0.296	20	20	20	0.235
20	15	15	0.255	20	25	25	0.224
25	5	5	0.436	25	15	15	0.242
25	10	10	0.284	25	20	20	0.223
25	15	15	0.242	25	25	25	0.212
30	5	5	0.428	30	15	15	0.234
30	10	10	0.276	30	20	20	0.215
30	15	15	0.234	30	25	25	0.203
35	5	5	0.422	35	15	15	0.228
35	10	10	0.27	35	20	20	0.209
35	15	15	0.228	35	25	25	0.198

Таблица П4. Интенсивности потоков и обслуживания

Вариант 1 (0.4, 0.6)							
$\lambda_{0,vs}$	$\lambda_{0,vt}$	μ_{1s}	μ_{1t}	μ_{2s}	μ_{2t}	μ_{3s}	μ_{4t}
2.5	2.5	10	10	5	5	15	5
		15	15	10	10	20	10
		20	20	15	15	25	15
		25	25	20	20	30	20
		30	30	25	25	35	25
		35	35	30	30	40	30
		40	40	35	35	45	35
		45	45	40	40	50	40
		50	50	45	45	55	45

Таблица П5. Временные характеристики узлов (вариант 1)

Узел VM					
t_1^s	t_1^t	T_1^s	T_1^t	W_1^s	W_1^t
0.1	0.1	0.101	0.101	0.0011	0.0011
0.067	0.067	0.067	0.067	0.00073	0.00073
0.05	0.05	0.051	0.051	0.00055	0.00055
0.04	0.04	0.04	0.04	0.00044	0.00044
0.033	0.033	0.034	0.034	0.00037	0.00037
0.029	0.029	0.029	0.029	0.00031	0.00031

Продолжение таблицы П5

Узел VM					
t_1^s	t_1^s	t_1^s	t_1^s	t_1^s	t_1^s
0.025	0.025	0.025	0.025	0.00027	0.00027
0.022	0.022	0.022	0.022	0.00024	0.00024
0.02	0.02	0.02	0.02	0.00022	0.00022
Узел A (FCFS)					
t_2^s	t_2^t	T_2^s	T_2^t	W_2^s	W_2^t
0.2	0.2	0.2027	0.2027	0.0027	0.0027
0.1	0.1	0.1014	0.1014	0.0014	0.0014
0.067	0.067	0.0676	0.0676	0.0009	0.0009
0.05	0.05	0.0507	0.0507	0.0007	0.0007
0.04	0.04	0.0405	0.0405	0.0005	0.0005
0.033	0.033	0.0338	0.0338	0.0005	0.0005
0.029	0.029	0.029	0.029	0.0004	0.0004
0.025	0.025	0.0253	0.0253	0.0003	0.0003
0.022	0.022	0.0225	0.0225	0.0003	0.0003
Узел C1			Узел C2		
t_3^s	T_3^s	W_3^s	t_4^t	T_4^t	W_4^t
0.067	0.067	0.0002	0.2	0.202	0.0015
0.05	0.05	0.0002	0.1	0.101	0.0008
0.04	0.04	0.0001	0.067	0.067	0.0005
0.033	0.033	0.0001	0.05	0.05	0.0004
0.029	0.029	0.0001	0.04	0.04	0.0003
0.025	0.025	0.0001	0.033	0.034	0.0003
0.022	0.022	0.0001	0.029	0.029	0.0002
0.02	0.02	0.0001	0.025	0.025	0.0002
0.018	0.018	0.0001	0.022	0.022	0.0002

Таблица П6. Времена отклика для заявок различных типов

T'_s	T'_t	T_s	T_t
0.371	0.505	0.491	0.625
0.219	0.27	0.339	0.39
0.158	0.185	0.278	0.305
0.125	0.141	0.245	0.261
0.103	0.115	0.223	0.235
0.088	0.096	0.208	0.216
0.076	0.083	0.196	0.203
0.068	0.073	0.188	0.193
0.061	0.065	0.181	0.185

Таблица П7. Интенсивности потоков и обслуживания

Вариант 2 (0.3, 0.7)							
$\lambda_{0,vs}$	$\lambda_{0,vt}$	μ_{1s}	μ_{1t}	μ_{2s}	μ_{2t}	μ_{3s}	μ_{4t}
2.5	2.5	10	10	5	5	15	5
		15	15	10	10	20	10
		20	20	15	15	25	15
		25	25	20	20	30	20
		30	30	25	25	35	25
		35	35	30	30	40	30
		40	40	35	35	45	35
		45	45	40	40	50	40
		50	50	45	45	55	45

Таблица П8. Временные характеристики узлов (вариант 2)

Узел VM					
t_1^s	t_1^t	T_1^s	T_1^t	W_1^s	W_1^t
0.1	0.1	0.101	0.101	0.0011	0.0011
0.067	0.067	0.067	0.067	0.00073	0.00073
0.05	0.05	0.051	0.051	0.00055	0.00055
0.04	0.04	0.04	0.04	0.00044	0.00044
0.033	0.033	0.034	0.034	0.00037	0.00037
0.029	0.029	0.029	0.029	0.00031	0.00031
0.025	0.025	0.025	0.025	0.00027	0.00027
0.022	0.022	0.022	0.022	0.00024	0.00024
0.02	0.02	0.02	0.02	0.00022	0.00022
Узел A (FCFS)					
t_2^s	t_2^t	T_2^s	T_2^t	W_2^s	W_2^t
0.2	0.2	0.2027	0.2027	0.0027	0.0027
0.1	0.1	0.1014	0.1014	0.0014	0.0014
0.067	0.067	0.0676	0.0676	0.0009	0.0009
0.05	0.05	0.0507	0.0507	0.0007	0.0007
0.04	0.04	0.0405	0.0405	0.0005	0.0005
0.033	0.033	0.0338	0.0338	0.0005	0.0005
0.029	0.029	0.029	0.029	0.0004	0.0004
0.025	0.025	0.0253	0.0253	0.0003	0.0003
0.022	0.022	0.0225	0.0225	0.0003	0.0003
Узел C1			Узел C2		
t_3^s	T_3^s	W_3^s	t_4^t	T_4^t	W_4^t
0.067	0.067	0.0002	0.2	0.202	0.0018
0.05	0.05	0.0002	0.1	0.101	0.0009
0.04	0.04	0.0001	0.067	0.067	0.0006
0.033	0.033	0.0001	0.05	0.05	0.0004
0.029	0.029	0.0001	0.04	0.04	0.0004
0.025	0.025	0.0001	0.033	0.034	0.0003
0.022	0.022	0.0001	0.029	0.029	0.0003
0.02	0.02	0	0.025	0.025	0.0002
0.018	0.018	0	0.022	0.022	0.0002

Таблица П9. Времена отклика для заявок различных типов

T'_s	T'_t	T_s	T_t
0.371	0.506	0.491	0.626
0.219	0.27	0.339	0.39
0.158	0.185	0.278	0.305
0.125	0.142	0.245	0.262
0.103	0.115	0.223	0.235
0.088	0.096	0.208	0.216
0.076	0.083	0.196	0.203
0.068	0.073	0.188	0.193
0.061	0.065	0.181	0.185

Таблица П10. Интенсивности потоков и обслуживания

Вариант 3 (0.2, 0.8)							
$\lambda_{0,vs}$	$\lambda_{0,vt}$	μ_{1s}	μ_{1t}	μ_{2s}	μ_{2t}	μ_{3s}	μ_{4t}
2.5	2.5	10	10	5	5	15	5
		15	15	10	10	20	10
		20	20	15	15	25	15
		25	25	20	20	30	20
		30	30	25	25	35	25
		35	35	30	30	40	30
		40	40	35	35	45	35
		45	45	40	40	50	40
		50	50	45	45	55	45

Таблица П11. Временные характеристики узлов (вариант 3)

Узел VM					
t_1^s	t_1^t	T_1^s	T_1^t	W_1^s	W_1^t
0.1	0.1	0.101	0.101	0.0011	0.0011
0.067	0.067	0.067	0.067	0.00073	0.00073
0.05	0.05	0.051	0.051	0.00055	0.00055
0.04	0.04	0.04	0.04	0.00044	0.00044
0.033	0.033	0.034	0.034	0.00037	0.00037
0.029	0.029	0.029	0.029	0.00031	0.00031
0.025	0.025	0.025	0.025	0.00027	0.00027
0.022	0.022	0.022	0.022	0.00024	0.00024
0.02	0.02	0.02	0.02	0.00022	0.00022
Узел A					
t_2^s	t_2^t	T_2^s	T_2^t	W_2^s	W_2^t
0.2	0.2	0.2027	0.2027	0.0027	0.0027
0.1	0.1	0.1014	0.1014	0.0014	0.0014
0.067	0.067	0.0676	0.0676	0.0009	0.0009
0.05	0.05	0.0507	0.0507	0.0007	0.0007
0.04	0.04	0.0405	0.0405	0.0005	0.0005
0.033	0.033	0.0338	0.0338	0.0005	0.0005
0.029	0.029	0.029	0.029	0.0004	0.0004
0.025	0.025	0.0253	0.0253	0.0003	0.0003
0.022	0.022	0.0225	0.0225	0.0003	0.0003

Продолжение таблицы П11

Узел С1			Узел С2		
t_3^s	T_3^s	t_3^s	T_3^s	t_3^s	T_3^s
0.067	0.067	0.0001	0.2	0.2	0.0005
0.05	0.05	0.0001	0.1	0.1	0.0002
0.04	0.04	0.0001	0.067	0.067	0.0002
0.033	0.033	0.0001	0.05	0.05	0.0001
0.029	0.029	0	0.04	0.04	0.0001
0.025	0.025	0	0.033	0.033	0.0001
0.022	0.022	0	0.029	0.029	0.0001
0.02	0.02	0	0.025	0.025	0.0001
0.018	0.018	0	0.022	0.022	0.0001

Таблица П12. Времена отклика для заявок различных типов

T'_s	T'_t	T_s	T_t
0.371	0.504	0.491	0.624
0.219	0.269	0.339	0.389
0.158	0.185	0.278	0.305
0.125	0.141	0.245	0.261
0.103	0.114	0.223	0.234
0.088	0.096	0.208	0.216
0.076	0.083	0.196	0.203
0.068	0.073	0.188	0.193
0.061	0.065	0.181	0.185

Таблица П13. Временные характеристики узла 1

t_{1r}	t_{1s}	t_{1t}	T_1^r	T_1^s	T_1^t	W_1^r	W_1^s	W_1^t
0.1	0.1	0.1	0.101	0.101	0.101	0.0011	0.0011	0.0011
0.067	0.067	0.067	0.067	0.067	0.067	0.00073	0.00073	0.00073
0.05	0.05	0.05	0.051	0.051	0.051	0.00055	0.00055	0.00055
0.04	0.04	0.04	0.04	0.04	0.04	0.00044	0.00044	0.00044
0.033	0.033	0.033	0.034	0.034	0.034	0.00037	0.00037	0.00037
0.029	0.029	0.029	0.029	0.029	0.029	0.00031	0.00031	0.00031
0.025	0.025	0.025	0.025	0.025	0.025	0.00027	0.00027	0.00027
0.022	0.022	0.022	0.022	0.022	0.022	0.00024	0.00024	0.00024
0.02	0.02	0.02	0.02	0.02	0.02	0.00022	0.00022	0.00022

Таблица П14. Временные характеристики узла 2 при моделировании узлами типа 1 FCFS и типа 2 PS

Узел 2 (тип 1 PS)						Узел 2 (тип 2 FCFS)					
t_{2r}	t_{2s}	t_{2t}	T_2^r	T_2^s	T_2^t	t_{2r}	t_{2s}	t_{2t}	T_2^r	T_2^s	T_2^t
0.2	0.2	0.2	0.4	0.4	0.4	0.2	0.2	0.2	0.2027	0.2027	0.2027
0.1	0.1	0.1	0.133	0.133	0.133	0.1	0.1	0.1	0.1014	0.1014	0.1014
0.067	0.067	0.067	0.08	0.08	0.08	0.067	0.067	0.067	0.0676	0.0676	0.0676
0.05	0.05	0.05	0.057	0.057	0.057	0.05	0.05	0.05	0.0507	0.0507	0.0507
0.04	0.04	0.04	0.044	0.044	0.044	0.04	0.04	0.04	0.0405	0.0405	0.0405
0.033	0.033	0.033	0.036	0.036	0.036	0.033	0.033	0.033	0.0338	0.0338	0.0338
0.029	0.029	0.029	0.031	0.031	0.031	0.029	0.029	0.029	0.029	0.029	0.029
0.025	0.025	0.025	0.027	0.027	0.027	0.025	0.025	0.025	0.0253	0.0253	0.0253
0.022	0.022	0.022	0.024	0.024	0.024	0.022	0.022	0.022	0.0225	0.0225	0.0225

Таблица П15. Временные характеристики узлов 3 и 4

Узел 3						Узел 4		
t_{3r}	t_{3s}	T_3^r	T_3^s	W_3^r	W_3^s	t_{4t}	T_4^t	W_4^t
0.067	0.067	0.067	0.067	0.0001	0.0001	0.2	0.202	0.0015
0.05	0.05	0.05	0.05	0.0001	0.0001	0.1	0.101	0.0008
0.04	0.04	0.04	0.04	0.0001	0.0001	0.067	0.067	0.0005
0.033	0.033	0.033	0.033	0.0001	0.0001	0.05	0.05	0.0004
0.029	0.029	0.029	0.029	0	0	0.04	0.04	0.0003
0.025	0.025	0.025	0.025	0	0	0.033	0.033	0.0003
0.022	0.022	0.022	0.022	0	0	0.029	0.029	0.0002
0.02	0.02	0.02	0.02	0	0	0.025	0.025	0.0002
0.018	0.018	0.018	0.018	0	0	0.022	0.022	0.0002

Таблица П16. Времена отклика для заявок различных типов при моделировании узла 2 узлами типа 1 FCFS и типа 2 PS

Узел 2 (тип 1 PS)					
T'_r	T'_r	T'_r	T'_r	T'_r	T'_r
0.568	0.568	0.568	0.568	0.568	0.568
0.251	0.251	0.251	0.251	0.251	0.251
0.171	0.171	0.171	0.171	0.171	0.171
0.131	0.131	0.131	0.131	0.131	0.131
0.107	0.107	0.107	0.107	0.107	0.107
0.09	0.09	0.09	0.09	0.09	0.09
0.078	0.078	0.078	0.078	0.078	0.078
0.069	0.069	0.069	0.069	0.069	0.069
0.062	0.062	0.062	0.062	0.062	0.062
Узел 2 (тип 2 FCFS)					
T'r	T'r	T'r	T'r	T'r	T'r
0.371	0.371	0.505	0.491	0.491	0.625
0.219	0.219	0.27	0.339	0.339	0.39
0.158	0.158	0.185	0.278	0.278	0.305
0.125	0.125	0.141	0.245	0.245	0.261
0.103	0.103	0.115	0.223	0.223	0.235

Продолжение таблицы П16

Узел 2 (тип 2 FCFS)					
T’r	T’r	T’r	T’r	T’r	T’r
0.088	0.088	0.096	0.208	0.208	0.216
0.076	0.076	0.083	0.196	0.196	0.203
0.068	0.068	0.073	0.188	0.188	0.193
0.061	0.061	0.065	0.181	0.181	0.185

Таблица П17. Расчет T’ при рассмотрении диапазонов КВ (варианты 1а-1в)

1а) При $\lambda_1 = \lambda_2 = 2.5c^{-1}$, $c_A(1) = 0.6$								
Узел 1	$c_A(1)$	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	0.297	0.381	0.569	0.79	1.019	1.254	1.493
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	0.712	0.737	0.812	0.94	1.061	1.215	1.38
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	0.712	0.737	0.812	0.94	1.061	1.215	1.38
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.195	0.201	0.22	0.256	0.298	0.348	0.407
1б) При $\lambda_1 = \lambda_2 = 2.5\ c^{-1}$, $c_A(1) = 1.2$.								
Узел 1	$c_A(1)$	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	1.109	1.134	1.21	1.328	1.476	1.648	1.836
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	1.039	1.057	1.11	1.207	1.304	1.432	1.574
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	1.039	1.057	1.11	1.207	1.304	1.432	1.574
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.219	0.229	0.25	0.28	0.318	0.367	0.425
1в) При $\lambda_1 = \lambda_2 = 2.5\ c^{-1}$, $c_A(1)=1.8$								
Узел 1	$c_A(1)$	1.8	1.8	1.8	1.8	1.8	1.8	1.8
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	1.77	1.786	1.835	1.914	2.02	2.149	2.296
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	1.427	1.44	1.48	1.553	1.63	1.734	1.853
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	1.427	1.44	1.48	1.553	1.63	1.734	1.853
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.233	0.245	0.268	0.301	0.341	0.39	0.449

.

Таблица П18. Расчет T’ при рассмотрении диапазонов КВ (варианты 2а-2в)

2а) При $\lambda_1 = \lambda_2 = 1.5\ c^{-1}$, $c_A(1) = 0.6$								
Узел 1	$c_A(1)$	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	0.453	0.49	0.589	0.725	0.881	1.047	1.221
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	0.762	0.777	0.82	0.898	0.976	1.078	1.19
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	0.762	0.777	0.82	0.898	0.976	1.078	1.19
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.193	0.196	0.204	0.223	0.248	0.276	0.308
2б) При $\lambda_1 = \lambda_2 = 1.5\ c^{-1}$, $c_A(1) = 1.2$								
Узел 1	$c_A(1)$	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	1.143	1.158	1.204	1.276	1.37	1.483	1.61
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	1.064	1.075	1.107	1.166	1.226	1.309	1.403
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	1.064	1.075	1.107	1.166	1.226	1.309	1.403
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.206	0.212	0.223	0.24	0.26	0.287	0.318
2в) При $\lambda_1 = \lambda_2 = 1.5\ c^{-1}$, $c_A(1) = 1.8$								
Узел 1	$c_A(1)$	1.8	1.8	1.8	1.8	1.8	1.8	1.8
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	1.773	1.783	1.813	1.861	1.927	2	2.104
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	1.443	1.441	1.465	1.51	1.557	1.623	1.7
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	1.443	1.441	1.465	1.51	1.557	1.623	1.7
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.213	0.219	0.232	0.25	0.272	0.299	0.331

Таблица П19. Расчет T’ при рассмотрении диапазонов КВ(варианты 3а-3в)

3а) При $\lambda_1 = \lambda_2 = 3.5\ c^{-1}$, $c_A(1) = 0.6$								
Узел 1	$c_A(1)$	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	0.11	0.197	0.537	0.84	1.135	1.428	1.719
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	0.38	0.691	0.801	0.997	1.138	1.337	1.545
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	0.38	0.691	0.801	0.997	1.138	1.337	1.545
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.2	0.209	0.24	0.296	0.358	0.436	0.528

Продолжение таблицы П19

3б) При $\lambda_1 = \lambda_2 = 3.5\ c^{-1}$, $c_A(1) = 1.2$								
Узел 1	$c_A(1)$	1.2	1.2	1.2	1.2	1.2	1.2	1.2
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	1.077	1.113	1.22	1.38	1.578	1.8	2.039
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	1.015	1.041	1.117	1.249	1.379	1.547	1.73
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	1.015	1.041	1.117	1.249	1.379	1.547	1.73
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.235	0.251	0.282	0.329	0.389	0.465	0.557
3в) При $\lambda_1 = \lambda_2 = 3.5\ c^{-1}$, $c_A(1) = 1.8$								
Узел 1	$c_A(1)$	1.8	1.8	1.8	1.8	1.8	1.8	1.8
	$c_B(1)$	0.1	0.5	1	1.5	2	2.5	3
Узел 2	$c_A(2)$	1.777	1.79	1.868	1.976	2.119	2.289	2.481
	$c_B(2)$	0.1	0.5	1	1.75	2	2.5	3
Узел 3	$c_A(3)$	1.429	1.447	1.52	1.603	1.706	1.845	2.001
	$c_B(3)$	1.5	2	2.5	3	3.5	4	4.5
Узел 4	$c_A(4)$	1.429	1.447	1.52	1.603	1.706	1.845	2.001
	$c_B(4)$	1.5	2	2.5	3	3.5	4	4.5
	T’	0.26	0.278	0.314	0.365	0.426	0.504	0.596

ПРИЛОЖЕНИЕ 2

Акты об использовании результатов диссертации

1. Акт об использовании результатов диссертационной работы Сулейманова А.А. на тему «Разработка и исследование метода оценки качества инфокоммуникационной облачной услуги «виртуальный рабочий стол» в госбюджетной научно-исследовательской работе кафедры Сети связи и системы коммутации (ССиСК) МТУСИ.

2. Акт об использовании результатов диссертационной работы Сулейманова А. А. «Разработка и исследование метода оценки качества инфокоммуникационной облачной услуги «виртуальный рабочий стол» в ООО «ЭЛТЕКС-МСК».

ФЕДЕРАЛЬНОЕ АГЕНТСТВО СВЯЗИ

Ордена Трудового Красного Знамени
федеральное государственное
бюджетное образовательное
учреждение высшего образования

«МОСКОВСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ СВЯЗИ И ИНФОРМАТИКИ»
(МТУСИ)

FEDERAL COMMUNICATIONS AGENCY OF THE RUSSIAN FEDERATION

MOSCOW TECHNICAL UNIVERSITY OF COMMUNICATIONS AND INFORMATICS
(MTUCI)

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

17.11. 2017 г. № 2641/02-17

На № _____ от _____

УТВЕРЖДАЮ
Проректор по учебной работе
Московского технического университета
связи и информатики
_____ **Е. В. ТИТОВ**
«17» 11 2017 г.

АКТ

об использовании результатов диссертационной работы Сулейманова А.А. на тему «Разработка и исследование метода оценки качества инфокоммуникационной облачной услуги «виртуальный рабочий стол» в госбюджетной научно-исследовательской работе кафедры Сети связи и системы коммутации (ССиСК) по теме «Разработка учебно-методических материалов по изучению новых телекоммуникационных технологий».

Мы, нижеподписавшиеся, директор Департамента организации и управления учебным процессом (ДО и УУП) МТУСИ Карпушина Н. Д., заведующая Центром планирования и сопровождения учебного процесса (ЦП и СУП) Патенченкова Е. К., заведующий кафедрой ССиСК д.т.н., профессор Степанов С.Н. составили настоящий акт о том, что результаты первой главы диссертационной работы Сулейманова А.А. использованы при выполнении госбюджетной научно-исследовательской работы кафедры ССиСК. Кроме того, материалы используются в учебном процессе в лекционном курсе по дисциплине «Перспективные сетевые телекоммуникационные технологии».

Директор Департамента организации
и управления учебным процессом _____ Карпушина Н. Д.

Заведующая Центром планирования
и сопровождения учебного процесса _____ Патенченкова Е. К.

Заведующий кафедрой ССиСК
д.т.н., профессор _____ Степанов С.Н.

ELTEX-MSK Общество с ограниченной ответственностью "ЭЛТЕКС-МСК"

УТВЕРЖДАЮ
Генеральный директор
ООО «ЭЛТЕКС-МСК»
Юрковский С.А.
«4» ноября 2017 г.

АКТ
об использовании результатов диссертационной работы Сулейманова А. А. «Разработка и исследование метода оценки качества инфокоммуникационной облачной услуги «виртуальный рабочий стол» в ООО «ЭЛТЕКС-МСК»

Настоящим актом подтверждаем, что в проектной деятельности компании, а также в ходе стендовых испытаний были использованы следующие результаты, полученные в диссертационной работе Сулейманова А.А. «Разработка и исследование метода оценки качества облачной инфокоммуникационной услуги «виртуальный рабочий стол», представленной на соискание ученой степени кандидата технических наук:

- аналитические модели услуги «виртуальный рабочий стол», учитывающие параметры узлов облачной инфраструктуры, параметры сети, параметры пользовательских устройств на этапах подключения и работы пользователей с виртуальными рабочими столами;
- методика оценки среднего времени отклика, а также других показателей качества, основанная на анализе аналитических моделей на основе сети Джексона, BCMP-сети, приближенном методе, учитывающем первые два момента (математическое ожидание и дисперсию).

Использование результатов, полученных Сулеймановым А.А., позволило оценить среднее время отклика в условиях функционирования услуги в трех наиболее распространённых сценариях (офисная работа, работа с возможностью просмотра видео через виртуальный рабочий стол, работа с мультимедиа). Кроме того, результаты диссертационной работы были использованы при оценке множества допустимых значений времени обслуживания и числа одновременно обслуживаемых пользователей, при которых выполняются заданные требования к качеству услуги на этапе ее инициализации.

Генеральный директор ООО «ЭЛТЕКС-МСК» Юрковский С.А.

Главный инженер ООО «ЭЛТЕКС-МСК» Попов А.В.

ИНН 7709864010, КПП 771401001
Тел.(495) 357-00-37, факс (495) 979-23-89, e-mail su@eltex-msk.ru