На правах рукописи

Горбунова Анастасия Владимировна

АНАЛИЗ МОДЕЛЕЙ МАССОВОГО ОБСЛУЖИВАНИЯ ДЛЯ ОЦЕНКИ ВРЕМЕНИ ОТКЛИКА В СИСТЕМЕ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ

Специальность 05.13.17 —
«Теоретические основы информатики»

Автореферат

диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2017

Работа выполнена на кафедре прикладной информатики и теории вероятностей факультета физико-математических и естественных наук Российского университета дружбы народов

Научный руководитель: доктор технических наук, профессор
Самуйлов Константин Евгеньевич

Официальные оппоненты: **Моисеева Светлана Петровна,**
доктор физико-математических наук, доцент,
профессор кафедры теории вероятностей и математической статистики Национального исследовательского Томского государственного университета

Нетес Виктор Александрович,
доктор технических наук, старший научный сотрудник,
профессор кафедры сетей связи и систем коммутации Московского технического университета связи и информатики

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В.А. Трапезникова Российской академии наук

Защита состоится 16 июня 2017 г. в ___ ч. ___ мин. на заседании диссертационного совета Д 212.203.28 на базе Российского университета дружбы народов по адресу: 117198, г. Москва, ул. Орджоникидзе, д.3.

С диссертацией можно ознакомиться в научной библиотеке Российского университета дружбы народов по адресу: г. Москва, ул. Миклухо-Маклая, д.6 (отзывы на автореферат просьба отправлять по указанному адресу) или на официальном сайте диссоветов РУДН по адресу: http://dissovet.rudn.ru/.

Автореферат разослан «___» _____ 2017 года.

Ученый секретарь
диссертационного совета  С.А. Васильев

Общая характеристика работы

Актуальность темы исследования. Облачные вычисления (cloud computing) представляют собой технологию, позволяющую удалённому пользователю в режиме реального времени по требованию получать доступ к вычислительным ресурсам: программным приложениям, серверам, устройствам хранения данных, сервисам и др., через Интернет в рамках согласованного качества обслуживания для выбранной ценовой категории. В задачи поставщика таких услуг входит не только обеспечение требуемого уровня оказываемых услуг, но и, как следствие, избежание перегрузки ресурсов при обработке пользовательских требований и снижение энергозатрат.

Существенное значение в системах облачных вычислений имеет технология виртуализации. Благодаря виртуализации ресурсов физического сервера многие облачные провайдеры могут предложить своим пользователям высокопроизводительные сервисы для решения кластерных задач и высокопроизводительных приложений. Подобные задачи характеризуются высоким уровнем параллелизма. В качестве примера можно назвать производственные и бизнес-приложения в таких отраслях как фармацевтика, нефтяная и газовая промышленность, здравоохранение, финансы и обрабатывающая промышленность; научно-исследовательские организации используют возможность применения облачных кластеров при проведении крупномасштабных исследований в различных областях. Таким образом, применение технологии виртуализации для организации параллельных вычислений в рамках одного высокопроизводительного приложения при решении сложной комплексной задачи имеет не меньшую ценность, чем одновременное использование одного сервера несколькими приложениями. Более того, при такой настройке ресурсов, как параллельная обработка данных, очевидно, появляется возможность снижения энергозатрат за счёт уменьшения времени обработки пользовательских запросов, что, вообще говоря, тоже можно отнести к одному из методов повышения энергетической эффективности и энергосбережения.

Наряду с таким важным показателем производительности систем облачных вычислений как время отклика, в контексте решения кластерных задач в последнее время значительное внимание уделяется времени, проведённому подзапросами в буфере синхронизации. Из-за возможного увеличения длительности обслуживания отдельных компонентов запроса в буфере син-

хронизации может накапливаться значительное число подзапросов, превышающее допустимый объем, что может привести если не к сбоям, то к снижению качества оказываемых услуг, если этот фактор не был учтен заранее, уже на этапе проектирования облачного центра. По этой причине наравне с математическим ожиданием времени синхронизации важно оценить и дисперсию этой случайной величины.

Несмотря на то, что современные облачные системы проектируются, как правило, масштабируемыми, всё равно остается проблема недостаточного использования ресурсов системы, что тоже является одной из причин потерь энергии. Одним из способов повышения энергетической эффективности и энергетического сбережения в системах облачных вычислений является динамическая активация виртуальных машин. Иными словами, регулируется количество выделяемых ресурсов для обработки запросов пользователей в зависимости от текущей нагрузки. Одной из проблем анализа подобных систем является вычислительная сложность предложенных решений для анализа вероятностно-временных характеристик, таких как время отклика, вызванная, в частности, увеличением ёмкости системы и количества обслуживающих приборов, что вполне естественно для современных систем облачных вычислений при росте числа пользователей и серверов для их обслуживания. Поэтому применение матрично-геометрических методов для получения стационарных характеристик системы, моделирующей облачный центр, становится затруднительным. В этой связи для анализа современных облачных платформ требуются эффективные вычислительные алгоритмы, позволяющие при этом получить оценку не только для среднего значения времени отклика, но и для его моментов более высоких порядков.

Степень разработанности темы. Для анализа показателей качества обслуживания в системах облачных вычислений в случае обработки сложного комплексного запроса, содержащего в себе несколько задач и требующего высокой готовности и производительности системы, естественно допустить возможность использования модели массового обслуживания с параллельной обработкой данных (fork-join queueing system). К одному из первых упоминаний этой модели можно отнести систему с двумя параллельно функционирующими блоками $D/M/1$, моделирующими ситуацию одновременного прибытия в аэропорт пассажиров и их вещей, последующего разделения

этих двух условных блоков и их соединения через некоторое время в зоне выдачи багажа. Точное выражение для среднего времени отклика было получено только для двух параллельно функционирующих систем $M/M/1$, в остальных же случаях были получены различными методами аппроксимации среднего времени отклика. Сложность в исследовании этой системы объясняется существующей зависимостью между очередями подзапросов из-за общих моментов поступления. Несмотря на широкий спектр задач, которые решаются с помощью систем массового обслуживания с параллельной обработкой запросов, и их популярность среди зарубежных авторов, в нашей стране данная система исследовалась значительно меньше. Среди отечественных учёных, внёсших существенный вклад в анализ подобных систем, можно назвать С.П. Моисееву благодаря проведённым ею исследованиям системы параллельного обслуживания сдвоенных заявок с неограниченным числом приборов, в результате чего были получены точные выражения для двумерного распределения вероятностей числа подзапросов в системе (приборов в каждой подсистеме), характеристики числа занятых приборов в соответствующих подсистемах и коэффициент корреляции между ними.

Одним из методов исследования систем с динамическим управлением подключением виртуальных машин является анализ с помощью моделей массового обслуживания с динамической активацией дополнительных приборов. Стоит отметить, что, как правило, выделяют два типа гистерезисного управления: гистерезисное управление входящим потоком (текущей нагрузкой) и гистерезисное управление обслуживанием, и для анализа этих моделей используются методы теории вероятностей, теории массового обслуживания и теории телетрафика. Среди авторов, внёсших серьёзный вклад в исследования можно назвать: М.А. Красносельского и А.В. Покровского, К.Е. Самуйлова, А.В. Печинкина, Ю.В. Гайдамака, С.Я. Шоргина, В.М. Вишневского, С.П. Моисееву, В.А. Нетеса, R.D. Nelson и A.N. Tantawi, F. Baccelli, S. Balsamo и I. Mura, L. Flatto and S. Hahn, C. Kim и A.K. Agrawala, A. Thomasian, J. Menon, I. Tsimashenka и W.J. Knottenbelt, E. Varki, S. Varma и A.M. Makowski и др.

Цели и задачи исследования. Сформулируем **цель** диссертационной работы — математические модели для систем облачных вычислений в контексте решения кластерных задач, а также с гистерезисным управлением

обслуживанием и разработка методов для анализа вероятностно-временных характеристик показателей качества обслуживания в этих системах. Для достижения цели исследований в диссертации решаются следующие актуальные **задачи**:

1. Построение и исследование системы облачных вычислений в виде системы массового обслуживания с параллельной обработкой заявок.
2. Анализ вероятностно-временных характеристик системы массового обслуживания с параллельной обработкой заявок, таких как среднее время отклика и дисперсия времени синхронизации.
3. Разработка рекуррентного алгоритма расчета оценки времени отклика для модели системы облачных вычислений с гистерезисным управлением подключением виртуальных машин.
4. Разработка рекуррентного алгоритма расчета оценки времени отклика для модели системы облачных вычислений с гистерезисным управлением подключением виртуальных машин с ограничением на одновременное число активаций.

Научная новизна.

1. Для построенной модели системы облачных вычислений в виде системы массового обслуживания с параллельной обработкой заявок получено стационарное распределение маргинальных вероятностей, которое не было представлено в известных источниках.
2. Проанализированы полученные в различных источниках оценки такой вероятностно-временной характеристики системы массового обслуживания с параллельной обработкой заявок, как среднее время отклика, предложена формула для оценки дисперсии времени отклика, а также для оценки дисперсии времени синхронизации, выражения для которых не были представлены в известных источниках.
3. Для модели системы облачных вычислений с гистерезисным управлением подключением виртуальных машин разработан рекуррентный алгоритм вычисления преобразования Лапласа-Стилтьеса времени отклика и времени ожидания начала обслуживания, позволяющий оценить не только математическое ожидание, но и дисперсию,

а также моменты высших порядков указанных случайных величин. Ранее можно было оценить только математическое ожидание непосредственным решением системы уравнений равновесия (СУР).

4. Для модели системы облачных вычислений с гистерезисным управлением подключением виртуальных машин и ограничением на одновременное число активаций разработан рекуррентный алгоритм вычисления преобразования Лапласа-Стилтьеса времени отклика и времени ожидания начала обслуживания, с помощью которого оцениваются не только математическое ожидание, но и дисперсия, а также моменты высших порядков указанных случайных величин. Ранее можно было оценить только математическое ожидание непосредственным решением СУР.

Теоретическая и практическая значимость работы. Теоретическая ценность полученных в диссертации результатов заключается в создании математического аппарата для исследования систем облачных вычислений в контексте решения кластерных задач и высокопроизводительных приложений, а также в контексте повышения энергетической эффективности и энергетического сбережения посредством динамической активации виртуальных машин. Полученные модели системы облачных вычислений в совокупности с разработанными алгоритмами могут использоваться для решения задач подбора оптимальных параметров функционирования систем, способствующих снижению энергозатрат и избежанию ухудшения качества обслуживания пользователей. Полученный математический аппарат может быть расширен с помощью других комбинаций типов входящего трафика и времён обслуживания.

Полученные оценки для дисперсии времени отклика и времени синхронизации, а также рекуррентный алгоритм расчета, позволяющий оценить моменты высших порядков для времени отклика и времени ожидания начала обслуживания, могут использоваться уже на этапе проектирования облачных центров при планировании необходимых ресурсов, которые потребуются для обеспечения соответствующего уровня обслуживания пользователей с учетом целей и задач, поставленных заказчиками перед проектными организациями.

Результаты работы использованы в рамках исследований по грантам РФФИ № 15-07-03051 «Формализация моделей и развитие методов анализа

вероятностных характеристик инфокоммуникационных межмашинных беспроводных сетей пятого поколения» и № 15-07-03608 «Разработка методов решения задач управления доступом в широкополосных беспроводных инфокоммуникационных сетях на основе нелинейного анализа и математической теории телетрафика», а также в учебном процессе при подготовке выпускных работ бакалавров и магистров, обучающихся по направлению «Фундаментальная информатика и информационные технологии».

Методология и методы исследования. В диссертационной работе применяются методология и методы теории массового обслуживания, теории вероятностей, теории марковских случайных процессов, теории порядковых статистик, математической теории телетрафика.

Положения, выносимые на защиту.

1. Для построенной модели системы облачных вычислений в виде системы массового обслуживания с параллельной обработкой заявок стационарное распределение маргинальных вероятностей, а также формула для оценки дисперсии времени синхронизации.
2. Рекуррентный алгоритм вычисления преобразования Лапласа-Стилтьеса времени отклика и времени ожидания начала обслуживания для модели системы облачных вычислений с гистерезисным управлением подключением виртуальных машин.
3. Рекуррентный алгоритм вычисления преобразования Лапласа-Стилтьеса времени отклика и времени ожидания начала обслуживания для модели системы облачных вычислений с гистерезисным управлением подключением виртуальных машин и ограничением на одновременное число активаций, т.е. для модели, аппроксимирующей исходную модель из предыдущего пункта.

Степень достоверности и апробация результатов. Достоверность, полученных в диссертации результатов, вытекает из использования строгих математических методов теории массового обслуживания, теории вероятностей, теории марковских случайных процессов, теории порядковых статистик, а также математической теории телетрафика. Кроме того, об обоснованности результатов свидетельствует численный эксперимент, проведенный на исходных данных, близких к реальным, а также его согласо-

ванность при сопоставлении с известными результатами, полученными для частных случаев.

Результаты диссертационного исследования докладывались на следующих научных конференциях: Вторая молодежная научная конференция «Задачи современной информатики» (Москва, 2015 г.); 9-ая Международная отраслевая научно-техническая конференция «Технологии информационного общества» (Москва, 2015 г.); X Юбилейная Международная практическая конференция «Современные информационные технологии и ИТ-образование» (Москва, 2015 г.); XV Международная конференция имени А.Ф. Терпугова «Информационные технологии и математическое моделирование (ИТТМ–2016)» (Томск, 2016 г.); Девятнадцатая международная научная конференция «Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2016)» (Москва, 2016 г.); I Международная научная конференция «Конвергентные когнитивно-информационные технологии» (Москва, 2016 г.).

Публикации. Основные результаты по теме диссертационного исследования изложены в 8 печатных изданиях [1–8], из которых 2 — изданы в журналах, рекомендованных ВАК РФ [7;8], и получены лично соискателем. В работах, опубликованных в соавторстве, личный вклад соискателя заключается в получении результатов, касающихся разработки моделей и их анализа; при его непосредственном участии разработаны программные средства.

Соответствие паспорту специальности. Диссертационное исследование выполнено в соответствии с паспортом специальности 05.13.17 «Теоретические основы информатики» и включает оригинальные результаты в области исследования информационных процессов и требований их пользователей к показателям эффективности, в области разработки моделей информационных процессов, разработки общих принципов организации телекоммуникационных систем и оценки их эффективности. Таким образом, диссертационное исследование соответствует следующим разделам паспорта специальности 05.13.17 «Теоретические основы информатики»: п. 2 (Исследование информационных структур, разработка и анализ моделей информационных процессов и структур), п. 16 (Общие принципы организации телекоммуникационных систем и оценки их эффективности).

Объем и структура работы. Диссертация состоит из введения, трёх глав, заключения и двух приложений. Полный объём диссертации составляет 97 страниц с 13 рисунками. Список литературы содержит 104 наименования.

Содержание работы

Во **введении** обоснована актуальность темы диссертации, определены цели и задачи исследований, сформулирована теоретическая и практическая ценность работы, представлены выносимые на защиту научные результаты.

В **первой главе** исследуются особенности построения моделей массового обслуживания в контексте решения кластерных задач и высокопроизводительных приложений, а также в контексте повышения энергетической эффективности и энергетического сбережения посредством динамической активации виртуальных машин для анализа системы облачных вычислений, формулируется задача исследования.

В *разделе 1.1* определяются показатели эффективности функционирования облачного центра, такие как время синхронизации и время отклика, определение которого, в частности, может меняться в зависимости от решаемой задачи. Также рассматриваются особенности построения моделей облачного центра с гистерезисным управлением подключением дополнительных приборов.

В *разделе 1.2* приведён обзор существующих типов моделей параллельного обслуживания заявок (fork-join): SPM (splitting and matching) система массового обслуживания, SM (split-merge) система массового обслуживания, FF (fission-fusion) система массового обслуживания, модель независимых серверов (independent server model, ISM), модель группового обслуживания (team service model, TSM), названия которых не имеют утвердившихся в русском языке соответствующих терминологических аналогов.

Также представлены методы анализа времени отклика модели параллельного обслуживания заявок, в контексте решения ресурсоёмких задач: эмпирическая аппроксимация, идея которой появилась из наблюдения за поведением времени отклика при проведении численных экспериментов; интерполяция с помощью предельных значений загрузки системы; анализ с помощью матрично-геометрического подхода; анализ с помощью порядковых

статистик для различных распределений времён пребывания подзапросов в системе.

В *разделе 1.3* формулируется задача исследования.

Во **второй главе** приведён анализ характеристик производительности системы облачных вычислений, представленной в виде системы массового обслуживания с параллельной обработкой запросов, в контексте решения супер-задач.

В *разделе 2.1* рассмотрены особенности построения модели системы облачных вычислений с расщеплением запросов, представлена система уравнений равновесия, а также получено в явном виде маргинальное распределение числа подзапросов в системе облачных вычислений с расщеплением.

Для моделирования облачного центра используется fork-join система массового обслуживания с K ветвями типа $M_\lambda/M_{\mu_k}/1$, $k = \overline{1,K}$. Если множество состояний случайного процесса $\{X(t), t \geq 0\}$, описывающего поведение системы во времени, представить в виде: $X = \{\vec{n} = (n_1,...,n_k,...,n_K), n_1 \geq 0,...,n_k \geq 0,...,n_K \geq 0\}$, n_k — число подзапросов k-го типа, находящихся в системе в некоторый момент времени t, $k = \overline{1,K}$, то стационарное распределение будет удовлетворять следующей системе уравнений равновесия:

$$\left(\lambda + \sum_{k=1}^{K} \mu_k u(n_k)\right) p(\vec{n}) = \lambda p(\vec{n} - \vec{1}) \prod_{k=1}^{K} u(n_k) + \sum_{k=1}^{K} \mu_k p(\vec{n} + \vec{e_k}), \vec{n} \in X, \quad (1)$$

где $\vec{e_k}$ — вектор, все элементы которого равны нулю, кроме k-го, который равен 1, $\vec{1}$ — единичный вектор и

$$u(n_k) = \begin{cases} 0 & \text{если } n_k \leq 0; \\ 1 & \text{если } n_k > 0. \end{cases}$$

Суммирование уравнений системы уравнений равновесия (СУР) (1) по всем индексам, кроме n_k-го, приводит к СУР, соответствующей системе массового обслуживания $M_\lambda/M_{\mu_k}/1$, следовательно, можем сформулировать утверждение.

Утверждение 1. *Маргинальное распределение числа подзапросов в системе облачных вычислений с расщеплением имеет вид:*

$$p_{n_k} = (1 - \rho_k)\rho_k^{n_k}, \quad n_k \geq 0, \tag{2}$$

где $\rho_k = \lambda/\mu_k$.

В *разделе 2.2* проанализировано время отклика системы облачных вычислений, представлены формулы для оценок математического ожидания и дисперсии этой случайной величины, полученные с помощью методов теории порядковых статистик.

Времена пребывания подзапросов в системе с расщеплением являются зависимыми случайными величинами в силу общих моментов поступления. Поскольку точное выражение для времени отклика в fork-join системе в случае $K > 2$ неизвестно, и возможность его получения остаётся под вопросом, то в качестве её аппроксимации будем использовать K независимых параллельно функционирующих систем массового обслуживания $M_\lambda/M_{\mu_k}/1$, $k = \overline{1,K}$.

Использование такого приближения представляется возможным не только исходя из естественных интуитивных соображений, но и благодаря тому, что выражения для маргинальных вероятностей числа подзапросов k-го типа, $k = \overline{1,K}$, в исходной fork-join системе, совпадают с выражениями для стационарных вероятностей в системе $M/M/1$, что и было показано выше.

Итак, считаем, что времена пребывания подзапросов в подсистемах являются независимыми экспоненциально распределёнными с.в. ξ_k, $k = \overline{1,K}$. Тогда для случая неоднородных приборов выражение для оценки дисперсии случайной величины времени отклика $W_{K,\max}$ примет вид:

$$D[W_{K,\max}] \approx \sum_{l=1}^{K} \frac{2}{(\mu - \lambda)^2} - \sum_{1 \leq l < m \leq K} \frac{2}{(\mu_l + \mu_m - 2\lambda)^2} +$$
$$+ \sum_{1 \leq l < m < k \leq K} \frac{2}{(\mu_l + \mu_m + \mu_k - 3\lambda)^2} + \ldots + (-1)^{K-1} \frac{2}{(\mu_1 + \mu_2 + \ldots + \mu_K - K\lambda)^2} -$$
$$- \left(\sum_{l=1}^{K} \frac{1}{\mu - \lambda} - \sum_{1 \leq l < m \leq K} \frac{1}{\mu_l + \mu_m - 2\lambda} + \sum_{1 \leq l < m < k \leq K} \frac{1}{\mu_l + \mu_m + \mu_k - 3\lambda} + \right.$$
$$\left. + \ldots + (-1)^{K-1} \frac{1}{\mu_1 + \mu_2 + \ldots + \mu_K - K\lambda} \right)^2. \tag{3}$$

Раздел 2.3 посвящён анализу времени, проведённому подзапросами в буфере синхронизации, выведена аппроксимирующая формула для дисперсии времени синхронизации.

Время синхронизации определяется как время между поступлением первой и последней частей подзапросов одного запроса в буфер синхронизации, иными словами это разность между максимумом и минимумом из времён пребывания подзапросов в системе. Сформулируем утверждение.

Утверждение 2. *Оценка дисперсии времени синхронизации системы облачных вычислений с расщеплением запросов для случая однородных приборов имеет вид:*

$$D[W_K] \approx \frac{1}{(\mu-\lambda)^2} \sum_{i=1}^{K-1} \frac{1}{i^2}. \tag{4}$$

В *разделе 2.4* проведён численный анализ рассматриваемых характеристик, а также сравниваются аппроксимирующие формулы для среднего времени отклика, представленные в различных источниках.

Относительная погрешность оценки среднего квадратического отклонения времени отклика относительно результатов имитационного моделирования составляет в среднем 7%.

В **третьей главе** предложен рекуррентный алгоритм расчета функции распределения времени ожидания начала обслуживания и времени отклика системы облачных вычислений с гистерезисным управлением в терминах преобразования Лапласа-Стилтьеса (ПЛС), что позволяет определить моменты высших порядков этих случайных величин. Ранее же можно было оценить только математическое ожидание исследуемых характеристик с помощью решения системы уравнений равновесия.

В *разделе 3.1* описан рекуррентный алгоритм в терминах преобразования Лапласа-Стилтьеса, позволяющий определить не только математическое ожидание, но и дисперсию и моменты высших порядков для времени отклика и времени ожидания начала обслуживания.

Рассмотрим систему облачных вычислений с гистерезисным подключением и отключением дополнительных виртуальных машин в виде многолинейной системы массового обслуживания с K приборами, часть которых может быть не активна, и конечной ёмкостью системы R. В систему поступает пуассоновский поток заявок с параметром λ. Считаем, что приборы являют-

ся однородными, время обслуживания распределено по экспоненциальному закону с параметром μ, при поступлении заявок в систему активация приборов происходит не мгновенно, а через случайное время, имеющее экспоненциальное распределение с параметром α. Количество активных приборов определяется числом заявок в очереди, в которой установлены парные пороги, заданные значениями векторов $\mathbf{H} = (H_1,H_2,...,H_{K-1})$, $H_1 < H_2 < ... < H_{K-1}$ и $\mathbf{L} = (L_1,L_2,...,L_{K-1})$, $L_1 < L_2 < ... < L_{K-1}$, где $L_{i+1} < H_i$, $i = \overline{1,K-2}$ и $L_i < H_i$, $i = \overline{1,K-1}$. Функционирование системы описывается марковским процессом $X(t)$ с множеством состояний:

$$S = \left\{ (k,i,n) \left| \begin{array}{ll} 0 \le n \le H_1, & k = 1,i = 1; \\ L_{k-1} \le n \le H_k, & k = \overline{2,K-1}, i = \overline{1,K-1} \\ L_{k-1} \le n \le R, & k = K, i = \overline{1,K} \end{array} \right. \right\}, \qquad (5)$$

где k — необходимое количество приборов для обслуживания заявок; i — количество активированных приборов; n — количество заявок в очереди.

Утверждение 3. *Преобразование Лапласа-Стилтьеса $V(s)$ времени ожидания начала обслуживания и ПЛС $W(s)$ времени пребывания заявки в системе равны:*

$$V(s) = \sum_{(k,i,n) \in S} \pi_{k,i,n} V_{k,i,n}^n, \qquad (6)$$

$$W(s) = \frac{\mu}{\mu + s} V(s), \qquad (7)$$

где $\pi_{k,i,n}$ — стационарные вероятности для соответствующих состояний $(k,i,n) \in S$, а $V_{k,i,n}^n(s)$ — ПЛС времени ожидания n-й в очереди заявки, если система находится в состоянии (k,i,n), определяются рекуррентными выражениями, представленными в разделе 1 главы 3 диссертации, которые в силу громоздкости не приводятся в тексте автореферата.

Благодаря свойствам ПЛС возможно вычислить моменты высших порядков и составить полноценное представление о поведении исследуемых случайных величин.

В *разделе 3.2* представлен рекуррентный алгоритм для вычисления времени ожидания начала обслуживания и времени отклика в терминах преобразования Лапласа-Стилтьеса для системы облачных вычислений с гистерезисным управлением и ограничением на одновременное число активаций.

Рассматриваемая в этом разделе модель отличается от предыдущей уменьшенным пространством состояний из-за ограничения на максимальное количество одновременно возможных активаций, которое вызывает дополнительные вычислительные трудности при увеличении количества приборов и ёмкости системы. Аппроксимация исходной модели упрощённой допустима, поскольку по результатам численного эксперимента максимальная погрешность приближения составляет не более 10%.

Утверждение 4. *Преобразование Лапласа-Стилтьеса* $\widetilde{V}(s)$ *времени ожидания начала обслуживания и ПЛС* $\widetilde{W}(s)$ *времени пребывания заявки в системе равны:*

$$\widetilde{V}(s) = \sum_{(k,i,n)\in\widetilde{S}} \widetilde{\pi}_{k,i,n} \widetilde{V}^n_{k,i,n}, \tag{8}$$

$$\widetilde{W}(s) = \frac{\mu}{\mu + s} \widetilde{V}(s), \tag{9}$$

где $\widetilde{\pi}_{k,i,n}$ *– стационарные вероятности для соответствующих состояний* $(k,i,n) \in \widetilde{S}$*, а* $\widetilde{V}^n_{k,i,n}(s)$ *– ПЛС времени ожидания* n*-й в очереди заявки, если система находится в состоянии* (k,i,n)*, которые определяются выражениями, представленными в разделе 2 главы 3 диссертации.*

Раздел 3.3 посвящен численному анализу полученных результатов, которые согласуются с результатам, представленными в других источниках по данной тематике.

На графике для математического ожидания времени отклика системы (рис. 1) присутствуют локальные минимумы, что объясняется выигрышем во времени при небольших значениях загрузки системы за счет быстрого и своевременного подключения дополнительных приборов (α — интенсивность подключения дополнительных приборов).

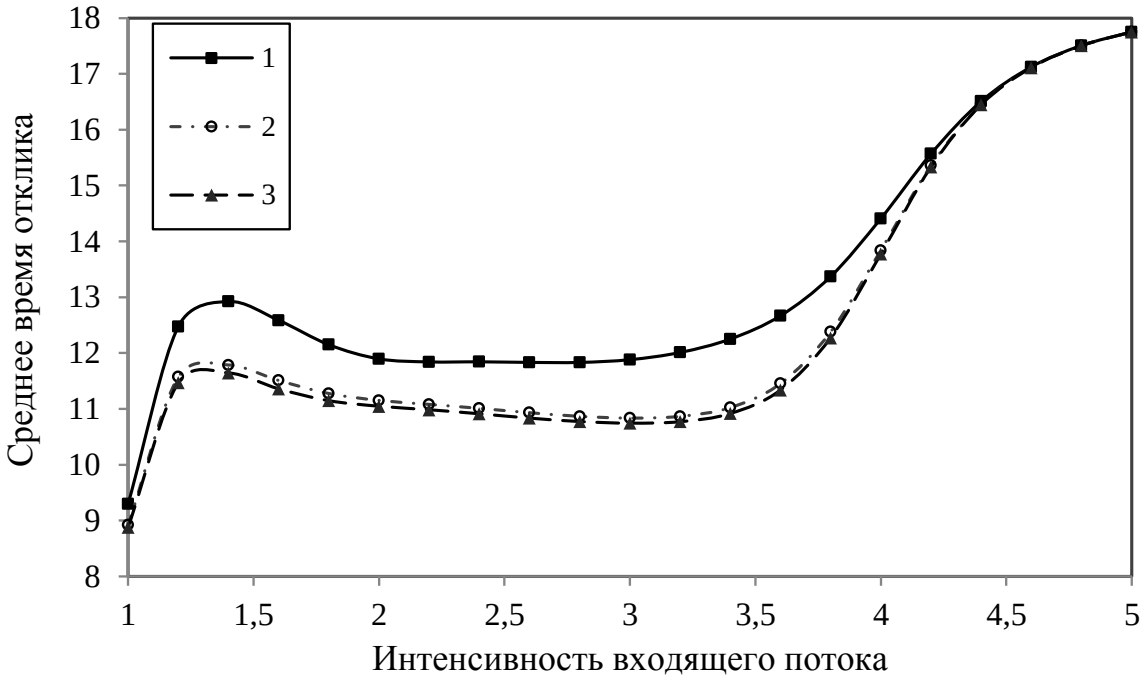

Рис. 1 — Математическое ожидание времени отклика: 1 — $\alpha = 0.1$; 2 — $\alpha = 1$; 3 — $\alpha = 10$

В **заключении** приведены основные результаты работы.

Основные результаты работы

1. Построена модель системы облачных вычислений в виде системы массового обслуживания с параллельной обработкой заявок. Получено стационарное распределение маргинальных вероятностей.
2. Проанализированы полученные в различных источниках оценки такой вероятностно-временной характеристики системы массового обслуживания с параллельной обработкой заявок, как среднее время отклика. Предложена формула для оценки дисперсии времени отклика, а также для оценки дисперсии времени синхронизации.
3. Для модели системы облачных вычислений с гистерезисным управлением подключением виртуальных машин разработан рекуррентный алгоритм вычисления преобразования Лапласа-Стилтьеса времени отклика и времени ожидания начала обслуживания, с помощью которого оцениваются математическое ожидание, дисперсия и моменты высших порядков указанных случайных величин.

4. Для модели системы облачных вычислений с гистерезисным управлением подключением виртуальных машин и ограничением на одновременное число активаций, т.е. для модели, аппроксимирующей исходную модель из предыдущего пункта, разработан рекуррентный алгоритм вычисления преобразования Лапласа-Стилтьеса времени отклика и времени ожидания начала обслуживания, с помощью которого оцениваются математическое ожидание, дисперсия, а также моменты высших порядков указанных величин.

Публикации автора по теме диссертации

1. *Горбунова А.В.* Оценка времени отклика обработки запросов в системе облачных вычислений // Труды Второй молодежной научной конференции «Задачи современной информатики». — 2015. — С. 79–85.

2. Алгоритм вычисления преобразования Лапласа-Стилтьеса для времени отклика системы облачных вычислений с гистерезисным управлением / К.Е. Самуйлов, Ю.В. Гайдамака, Э.С. Сопин, А.В. Горбунова // *Современные информационные технологии и ИТ-образование.* — 2015. — Т. 2, № 11. — С. 172–177.

3. Оценка вероятностных характеристик системы облачных вычислений с расщеплением запросов / А.В. Горбунова, И.С. Зарядов, С.И. Матюшенко, Э.С. Сопин // Информационные технологии и математическое моделирование (ИТММ-2016): Материалы XV Международной конференции имени А.Ф. Терпугова. — 2016. — Ч. 1. — С. 167–172.

4. The Estimation of Probability Characteristics of Cloud Computing Systems with Splitting of Requests / A.V. Gorbunova, I.S. Zaryadov, S.I. Matushenko, E.S. Sopin // Proceedings of the Nineteenth International Scientific Conference Russia: Distributed computer and communication networks: control, computation, communications (DCCN-2016). — Vol. 3: Youth School-Seminar. — 2016. — P. 467–472.

5. *Горбунова А.В., Самуйлов К.Е., Сопин Э.С.* Преобразование Лапласа-Стилтьеса для времени отклика системы облачных вычислений с гисте-

резисным управлением и ограничением на одновременное число активаций // *Современные информационные технологии и ИТ-образование.* — 2016. — Т. 12, № 1. — С. 21–27.

6. *Самуйлов К.Е., Зарядов И.С., Горбунова А.В.* Анализ времени отклика системы облачных вычислений // IX Международная отраслевая научно-техническая конференция «Технологии информационного общества». Тезисы научно-технических секций. — 2015. — С. 29–30.

7. Аппроксимация времени отклика системы облачных вычислений / А.В. Горбунова, И.С. Зарядов, С.И. Матюшенко и др. // *Информатика и её применения.* — 2015. — Т. 9, Вып. 3. — С. 32–38.

8. *Самуйлов К.Е., Зарядов И.С., Горбунова А.В.* Анализ времени отклика системы облачных вычислений // *T-Comm: Телекоммуникации и транспорт.* — 2015. — Т. 9, № 11. — С. 57–61.