

На правах рукописи

Румянцев Александр Сергеевич

**Методы моделирования, анализа стационарности
и оценивания производительности систем
параллельной обработки**

Специальность 05.13.18 —
«Математическое моделирование, численные методы
и комплексы программ»

Автореферат
диссертации на соискание учёной степени
доктора физико-математических наук

Петрозаводск — 2022

Работа выполнена в Институте прикладных математических исследований – обособленном подразделении Федерального государственного бюджетного учреждения науки Федерального исследовательского центра «Карельский научный центр».

Научный консультант: доктор физико-математических наук, профессор

Морозов Евсей Викторович

Официальные оппоненты: **Дудин Александр Николаевич**, доктор физико-математических наук, профессор, Белорусский государственный университет, заведующий научно-исследовательской лабораторией прикладного вероятностного анализа

Моисеева Светлана Петровна,

доктор физико-математических наук, профессор, федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет», профессор кафедры теории вероятностей и математической статистики

Цициашвили Гурами Шалвович,

доктор физико-математических наук, профессор, Федеральное государственное бюджетное учреждение науки Институт прикладной математики Дальневосточного отделения Российской академии наук, главный научный сотрудник отдела системных исследований

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

Защита состоится «29» сентября 2022 г. в 13 часов на заседании диссертационного совета Д 212.190.03 на базе Петрозаводского государственного университета по адресу: 185910, г. Петрозаводск, пр. Ленина, 33.

С диссертацией можно ознакомиться в библиотеке Петрозаводского государственного университета и на сайте <http://petrsu.ru>.

Автореферат разослан « » июня 2022 года.

Ученый секретарь
диссертационного совета

Воронов Роман Владимирович

Общая характеристика работы

Актуальность темы.

Хранение, передача, обработка данных осуществляется с помощью множества вычислительных, телекоммуникационных систем и систем хранения. Объединяющим свойством таких систем является использование параллельной обработки, целью внедрения которой является обеспечение высокой производительности систем, преодоление фундаментальных физических проблем при производстве микросхем, повышение пропускной способности каналов связи, снижение задержек в облачных системах доставки контента и центрах обработки данных, обеспечение надежности распределенных вычислений и хранения данных. При этом каждый класс таких систем *параллельной обработки* обладает рядом особенностей.

Высокопроизводительные *вычислительные кластеры* (суперкомпьютеры), обеспечивают вычислительную мощность, достаточную для выполнения трудоемких расчетов (моделирование климата, гидро- и газодинамики, анализ больших данных). Ключевой особенностью суперкомпьютеров является возможность такого разделения ресурсов между задачами, при котором одна задача может одновременно использовать множество ресурсов (ядер, узлов, серверов), занятие и освобождение которых происходит одновременно.

Альтернативой суперкомпьютерам являются системы *распределенных вычислений*, такие как вычислительные сети из персональных компьютеров, в которых неоднородные вычислительные ресурсы объединены с помощью локальной/беспроводной/глобальной сети, а для распределения заданий используется управляющий сервер. Баланс между надежностью системы и скоростью получения результата достигается с помощью таких параметров систем, как число одновременно решаемых копий (реплик) одного подзадания и число результатов, необходимых для успешного прохождения валидации и фиксации результата (кворум).

Развитие каналов связи и коммуникационных сетей позволяет осуществлять перенос части задач в так называемые *облачные системы*, в которых вычисления производятся удаленно, на оборудовании поставщика услуг, на гибко выделяемой части ресурсов из общего пула серверов. Пользователю облачных систем необходимо оценивать затраты на работу с ресурсами, в то время как владельца интересует компромисс между энергоэффективностью и обеспечением качества обслуживания. Подобная задача возникает и в малых устройствах, имеющих ограниченный ресурс электропитания, таких как узлы сети интернета вещей, ноутбуки.

Передача данных от облачной системы до конечного потребителя предполагает согласованную работу нескольких устройств, работу с несколькими каналами. Для обеспечения надежной передачи данных

используются сети, предусматривающие избыточность, групповое поступление и групповое обслуживание (например, пакетную передачу), для обеспечения низких задержек, высокой пропускной способности и отказоустойчивости. Параметры таких сетей (скорость передачи, уровень дублирования информации) необходимо выбирать обоснованно. Подобная задача возникает и в современных системах хранения, основанных на технологии твердотельных накопителей, которые, в отличие от традиционных систем, обладают значительным внутренним резервом параллелизма, встроенной избыточностью, но существенно меньшим временем наработки на отказ отдельных модулей хранения.

Анализ и оценивание показателей эффективности (в том числе производительности, качества обслуживания, энергоэффективности) систем параллельной обработки должны проводиться на основе моделирования, поскольку проведение экспериментов на системе в режиме эксплуатации сложно, либо невозможно. Такие модели обладают рядом особенностей: как правило, большим числом взаимодействующих устройств, сложной структурой и случайным характером исследуемых процессов, большим числом конфигурируемых параметров. Это приводит к необходимости развития методов исследования стохастических моделей систем параллельного обслуживания.

Степень разработанности темы.

Исследование характеристик качества обслуживания, надежности, энергоэффективности систем параллельной обработки наиболее целесообразно проводить методами теории массового обслуживания. Модели систем параллельной обработки в рамках теории массового обслуживания относятся, как правило, к классу многосерверных. Современное изложение основных методов прикладного вероятностного анализа многосерверных систем обслуживания представлено, например, в фундаментальных работах С. Асмуссена, П. П. Бочарова и А. В. Печинкина, Х. Арталехо и А. Гомес-Коррала, М. Харколь-Балтер, П. Брилля. Среди используемых методов, для анализа марковских моделей систем параллельной обработки со сложной структурой широко применяется матрично-аналитический метод, представленный, например, в работах М. Ньютса, Г. Латуша и В. Рамасвами, Ч.-М. Хэ, М. Бладта и Б. Нильсена, В. Наумова, Ю. В. Гайдамаки, Н. В. Яркиной и К. Е. Самуйлова. Указанный метод опирается на результаты из алгебры матриц, ключевые для понимания данного метода результаты представлены в книгах Ф. Гантмахера, Е. Сенеты, И. Гохберга. В отсутствие марковского свойства, анализ соответствующих моделей можно проводить с использованием теории регенерирующих процессов, представленной, в частности, в статьях В. Смита, С. Фосса, книгах С. Асмуссена, А. А. Боровкова, Х. Торрисона, Е. В. Морозова и Б. Стюарта. Наконец, в наиболее трудных для анализа случаях необходимо

использовать имитационное моделирование, основы которого приведены в монографиях Ш. Росса, Дж. Шедлера, С. Асмуссена и П. Глинна.

В работах указанных авторов широко представлены классические методы теории массового обслуживания и, более широко, прикладной теории вероятностей, однако во многих случаях для анализа систем параллельной обработки необходима адаптация, модификация существующих, либо разработка новых методов, что объясняется особенностями стохастических моделей таких систем. Остановимся более подробно на особенностях анализа классов моделей, представленных и исследованных в диссертации.

Ключевые особенности высокопроизводительных вычислительных систем учтены в особом подклассе многосерверных моделей. В отличие от классических многосерверных систем (исследованных в фундаментальных статьях Дж. Кифера и Дж. Вольфовица), в указанных моделях предполагается, что клиент одновременно занимает не менее одного сервера. При этом выделяют подклассы систем с потерями (исследованных в работах Дж. Кауфмана, Е. Артурса, В. Витта, Н. Ван Дийка, О. М. Тихоненко) и систем с ожиданием. Среди последних выделяют два подкласса: системы с *независимым* освобождением серверов, в которых требующиеся клиенту сервера начинают обслуживание одновременно, но времена обслуживания на каждом сервере являются независимыми (работы Л. Грин, Э. Сеилы, А. Федергрюна, П. Иттимакина, Э. Као, Х. Шаака, Р. Ларсона, Ф. Жийя, Г. Латуша и др.), а также с *одновременным* освобождением серверов (работы С. Ш. Кима, П. Брилля, Л. Грин, В. Витта, А. Унвина, Г. Флетчера и др.). Последний класс является наиболее трудным для анализа. Отметим, что матрично-аналитическая модель такой системы впервые была предложена в работе С. Ш. Кима. Возобновление интереса к модели с одновременным освобождением серверов мотивировано новыми приложениями к суперкомпьютерным системам. Можно выделить работы Д. Филиппопулоса, Х. Каратца и Ш. Чакраварти, посвященные анализу моделей малой размерности, а также работы Л. Г. Афанасьевой, С. А. Гришуниной и Е. Е. Баштовой, в которых выполнен анализ чувствительности условия стационарности к распределению времени обслуживания клиента. Наконец, открытые проблемы в данном классе моделей представлены в недавних работах М. Харколь-Балтер. Среди таких проблем можно отметить нахождение условий стационарности моделей, а также анализ характеристик в стационарном режиме.

В анализе систем распределенных вычислений, передачи данных и хранения важную роль играют параметры избыточности (репликации), с помощью которых обеспечивается снижение времени отклика, повышение надежности системы. Данный класс систем относится к многосерверным системам с разделяющимися заявками, к которым относятся системы типа Fork–Join, Split–Merge и другие (представленные в работах Л. Липского

и П. Фиорини, Р. Эванса, обзорных работах А. Томасяна, А. В. Горбуновой, И. С. Зарядова и К. Е. Самуйлова). Классические системы Fork-Join предполагают независимое выполнение подзаданий на серверах системы с последующей сборкой результата (такие системы исследованы Ф. Баччелли, К. Гарднера, Э. Хюття, А. Шеллер-Вольфа, Ц. Кима, Р. Нельсона, А. Тантави и др.) и являются наиболее трудными для анализа; в обобщенных Fork-Join системах окончание обслуживания клиента происходит при сборке части результатов выполнения подзаданий (исследованы в работах Г. Джоши); системы типа Fork-Early-Cancel позволяют выполнять раннюю отмену избыточных подзаданий (исследованы Г. Джоши, Э. Сольянин и Г. Ворнелла); наконец, системы обобщенного Split-Merge типа позволяют обслуживать группы подзаданий на серверах, ожидая окончания обслуживания части группы, но не допуская начала обслуживания новой группы ранее окончания обслуживания старой. Последний класс систем можно рассматривать как классические многосерверные системы, в которых времена обслуживания клиентов определяются с помощью порядковых статистик.

В облачных вычислительных системах, центрах обработки данных, а также на малых устройствах компромисс между затратами (в том числе энергетическими) и производительностью исследуется с помощью моделей систем с управлением режимами, скоростью работы серверов. Механизмы управления включают управление частотой работы процессоров и напряжением (работы Ж.-М. Пирсона), пороговое управление скоростью по нагрузке (исследования Р. Беккера), управление доступной вычислительной мощностью (работы Д. Мухерджи, С. Дхары, С. Борста, Й. Леуваардена и др.), управление режимом пониженного энергопотребления (сон, гибернация) (исследования М. Гебрехивота, П. Лассилы, С. Аалто, А. Гандхи и др.), либо комбинацию этих техник (работы Т. Хорвата и К. Скадрона).

Наконец, при анализе систем передачи данных, а также отдельных критических процессов в вычислительных сетях (в частности, процессов обработки результатов вычислений) применяют модели систем с повторными вызовами, систем с групповым обслуживанием и групповым поступлением, систем обслуживания-запасания. В частности, в классе систем беспроводной передачи типа когнитивного радио, допускающей одновременное обслуживание нескольких классов абонентов на одном устройстве, широко применяются методы теории массового обслуживания (работы Ф. Палюнчича, А. Альфы и др.), и, в частности, матрично-аналитический метод (работы А. Н. Дудина, Д. Б. Чжу, Б. Окландер, М. Сиди и др.). Системы с повторными вызовами подразделяют на системы с постоянной интенсивностью орбит (так называемых орбит-очереди), а также с классическими орбитами, в которых интенсивность попыток пропорциональна числу клиентов (исследования Х. Арталехо, Т. Янга, Дж. Темплтона и

др.). Первый из указанных подклассов сложнее с точки зрения анализа стационарности (представлен в статьях Х. Арталехо, К. Авраченкова, Е. В. Морозова и др.). В моделях с групповым обслуживанием (исследованных в работах Н. Бейли, А. Банерджи, Ю. Гупты, Ш. Чакраварти, С. Сасикалы, К. Индиры и др.) важную роль играет размер обслуживаемой группы (варианты моделей исследованы в статьях А. Баника, Ж. Баэтена, Ю. Зенга и др.). Часто указанные виды моделей имеют несколько особенностей, например, повторные вызовы и групповое обслуживание (работы А. Н. Дудина, Ш. Чакраварти, С. А. Дудина, М. Д'Ариенцо и др.), повторные вызовы в системах обслуживания–запасания (исследования А. Кришнамурти, Д. Шаджин, Б. Лакшми и др.), случайный объем поступающих клиентов (работы Е. В. Панкратовой, С. П. Моисеевой, М. П. Фархадова, А. Н. Моисеева), отказы (исследования Г. Ш. Цициашвили, М. А. Осиповой, К. Е. Самуйлова, Ю. В. Гайдамака и др.). Для анализа таких моделей применяют матрично-аналитический, регенеративный (работы Е. В. Морозова), асимптотический (работы Е. Ю. Данилюк, С. П. Моисеевой, А. А. Назарова) и иные (например, работы М. Чодхри, А. Баника, А. Пачеко и др.) методы.

Указанные особенности моделей систем параллельной обработки приводят к необходимости создания новых и развития существующих методов анализа их стационарности, оценивания характеристик производительности и эффективности в стационарном, переходном режиме, а также разработки соответствующих алгоритмов и реализующих их комплексов программ.

Целью диссертационной работы является решение фундаментальной научной проблемы создания теоретических основ, разработки комплекса новых методов анализа и расчета показателей эффективности систем параллельной обработки на основе развития приложений структурированных марковских, обобщенных полумарковских и регенерирующих процессов.

Для достижения поставленной цели в диссертации решены следующие **задачи**:

1. Разработан и исследован комплекс моделей систем параллельной обработки, учитывающий характеристики современных высокопроизводительных и распределенных вычислительных систем и систем хранения;
2. Предложены, разработаны и развиты методы анализа и оценивания эффективности систем массового обслуживания, указанные методы применены к разработанному комплексу моделей;
3. Разработаны алгоритмы и программы, реализующие предложенные методы для расчета характеристик моделей систем параллельной обработки.

Научная новизна диссертации состоит в следующем:

- разработан и исследован комплекс моделей систем параллельной обработки на основе структурированных цепей Маркова с непрерывным временем, стохастических рекуррентных соотношений, положительно возвратных регенерирующих процессов, обобщенных полумарковских процессов, учитывающий ключевые характеристики современных высокопроизводительных и распределенных вычислительных систем и систем хранения, в том числе использование многих серверов/каналов для расчета задания/передачи данных, групповое поступление/обслуживание заданий, избыточность обслуживания/хранения;
- предложены, разработаны, развиты и применены методы анализа и оценивания эффективности систем массового обслуживания, в том числе метод получения явного вида критерия стационарности матрично-аналитических моделей, модифицированный метод расщепления состояний и новый способ совместного применения матрично-аналитического и регенеративного методов для анализа их стационарных характеристик, новый способ регенеративного оценивания характеристик обобщенных полумарковских процессов с использованием расщепления плотности, параллельных и распределенных вычислений;
- разработаны алгоритмы и программы, реализующие предложенные методы для расчета характеристик комплекса моделей систем параллельной обработки, учитывающие сложный характер зависимости характеристик исследуемых процессов, в том числе коррелированный групповой марковский входной поток, распределения времени обслуживания с тяжелым хвостом, либо фазового типа, позволяющие получать характеристики в стационарном и переходном режимах, выполнять оценивание с использованием параллельных и распределенных вычислений, предназначенные для поддержки исследований систем параллельной обработки.

Совокупность предложенных и развитых в диссертации методов составляет теоретическую основу трехуровневого подхода к построению моделей систем обслуживания. При таком подходе, используется согласованная система из модели аналитического (например, на основе матрично-аналитической модели), имитационного (например, на основе обобщенного полумарковского процесса) и технического (например, малая техническая система в контролируемой среде) уровней, различающиеся по степени общности получаемых выводов и степени применимости результатов на практике. Согласованность моделей обеспечивается с помощью проведения попарной валидации моделей всех уровней, а также совместной валидации всех трех уровней моделирования в таких модельных предположениях, при которых модели соответствующих уровней существуют. В такой системе,

высокая степень общности получаемых на аналитическом уровне теоретических результатов дополняется гибкостью имитационных моделей, а также высокой прикладной значимостью модели технического уровня.

Основные положения диссертации, выносимые на защиту:

1. Предложен явный вид критерия стационарности многоклассовых многосерверных моделей систем параллельной обработки. Для класса моделей, обладающих свойством локальной однородности переходов, доказано достаточное условие, позволяющее получать критерий стационарности в явном виде в форме произведения вероятностей. Доказанное условие применено для матрично-аналитической модели суперкомпьютера, в том числе допускающей зависимость интенсивности входного потока от состояния системы (обратную связь) неоднородность классов по интенсивности обслуживания, а также для многосерверной модели системы параллельной обработки с исчерпывающей дисциплиной обслуживания группы однотипных клиентов.
2. Предложен способ совместного применения матрично-аналитического и регенеративного методов для нахождения стационарных вероятностей в моделях систем обслуживания. Способ применен для анализа многоклассовой модели системы с повторными вызовами постоянной интенсивности, модели односерверной системы с управлением скоростью обслуживания.
3. Предложен метод нахождения явного вида матрицы интенсивностей структурированной цепи Маркова с непрерывным временем, обладающей блочной почти верхнетреугольной матрицей, в условиях малого размера блоков, обобщающий известные ранее методы получения явного решения для случая малого фазового пространства. Метод применен для анализа стационарных характеристик ряда стохастических моделей, в том числе матрично-аналитической модели когнитивного радио, модели односерверной системы с управлением скоростью обслуживания, модели вычислительной сети из персональных компьютеров.
4. Предложена модификация метода расщепления состояний для анализа матрично-аналитических моделей систем в стационарном и переходном режиме. Модификация применена для анализа двухсерверной модели суперкомпьютера с неоднородными классами, а также модели когнитивного радио.
5. Разработаны алгоритмы и программы, реализующие предложенные методы для расчета характеристик комплекса моделей систем параллельной обработки.

Теоретическая и практическая значимость Результаты диссертации вносят вклад в развитие фундаментальных основ и прикладных вероятностных методов исследования систем массового обслуживания применительно к системам параллельной обработки. Прикладной характер полученных результатов обеспечивается разработанными и примененными комплексами программ, которые позволяют научно-исследовательским организациям и инновационным компаниям заблаговременно выполнять анализ характеристик эффективности разрабатываемых или исследуемых вычислительных, телекоммуникационных систем, систем хранения и обработки данных.

Основные результаты диссертационной работы получены при выполнении ряда научно-исследовательских проектов, где автор диссертационной работы являлся руководителем и исполнителем, в том числе, при исследованиях по грантам РФФИ, РФФИ, Президента РФ, Московского Центра фундаментальной и прикладной математики МГУ им. М. В. Ломоносова, а также НИР, выполняемых по приоритетным научным направлениям в ИПМИ КарНЦ РАН, Программе стратегического развития ПетрГУ.

Методология и методы исследования. Основу анализа составляют прикладные методы теории вероятностей, в том числе методы теории массового обслуживания. В диссертации предложены новые методы и развиты оригинальные модификации методов и процедур в теории массового обслуживания:

- метод анализа стационарности в матрично-аналитической модели на основе локальной однородности (раздел 1.1.5);
- комбинация регенеративного и матрично-аналитического методов (раздел 4.1.3), а также модификация матрично-аналитического метода (раздел 4.1.1), модификация метода расщепления состояний (раздел 4.1.4) для нахождения стационарных и переходных характеристик систем в явном виде;
- способ регенеративного оценивания стационарных характеристик моделей систем обслуживания на основе обобщенных полумарковских процессов и расщепления плотности, с использованием параллельных и распределенных вычислений (раздел 6.1).

Применяются в диссертации также классические методы исследования, в том числе:

- матрично-аналитический метод для анализа стационарности, а также характеристик в стационарном и переходном режиме структурированных цепей Маркова с непрерывным временем,
- регенеративный метод для анализа стационарности и получения базовых стационарных вероятностей моделей,
- методы теории вероятностей, комбинаторики, линейной алгебры, статистического оценивания.

Разработка программ и программных комплексов, выполнение численных экспериментов и анализ их результатов осуществлялись в программных средах R и Wolfram Cloud. Для выполнения расчетов применялись современные высокопроизводительные и распределенные вычислительные ресурсы (в том числе вычислительная сеть из персональных компьютеров, построенная на основе технологии BOINC), а также методы параллельных многоядерных и распределенных вычислений, доступные в языке R с помощью пакетов расширения, разработчиком которых в том числе является автор диссертации.

Достоверность полученных результатов подтверждается математически корректными выводами и доказательствами теорем и других утверждений, опубликованных в ведущих отечественных и международных журналах, корректностью разработанных численных методов исследования, корректным использованием известных фундаментальных методов, соответствием полученных теоретических результатов данным вычислительных экспериментов.

Апробация работы. Основные результаты работы представлялись на международных и всероссийских научных конференциях и семинарах по тематике исследований:

- International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN): 2010, 2015–2019, 2021 г.
- International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT): 2010, 2015–2017 г.
- Annual International Workshop on Advances in Methods of Information and Communication Technology: 2010, 2011 г.
- International Seminar on Stability Problems for Stochastic Models (ISSPSM) and Workshop «Applied Problems in Theory of Probabilities and Mathematical Statistics related to modeling of information systems»: 2011, 2014, 2015, 2020, 2021 г.
- «Высокопроизводительные параллельные вычисления на кластерных системах»: 2011, 2012 г.
- «Параллельные вычислительные технологии»: 2011, 2013 г.
- «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем»: 2011 г.
- «Вероятностные методы в дискретной математике»: 2012, 2016, 2019 г.
- «Сетевые игры и менеджмент», 2013 г.
- BOINC: Fundamental and Applied Science and Technology (BOINC:FAST): 2013, 2015, 2017 г.
- European Conference on Queueing Theory (ECQT): 2014, 2018 г.
- «Распределенные вычисления и Грид-технологии в науке и образовании» (GRID): 2014 г.

- Национальный Суперкомпьютерный Форум (НСКФ): 2014, 2015, 2016, 2017 г.
- International Conference Parallel Computing Technologies (PaCT): 2015 г.
- Man-Machine Interactions: 2015, 2017 г.
- International Conference on Matrix-Analytic Methods in Stochastic Models (МAM): 2016
- International Conference on Computer Networks (CN): 2016 г.
- European Workshop on Performance Engineering (EPEW): 2016 г.
- European Conference on Modelling and Simulation (ECMS): 2017 г.
- International Scientific Conference on Information Technologies and Mathematical Modelling (ITMM): 2018, 2021 г.
- Conference of the Open Innovations Association (FRUCT): 2018, 2019, 2020 г.
- International Conference on Wireless Sensor Networks, Ubiquitous Computing and Applications (ICWSNUCA): 2018 г.
- Measurement, Modelling and Evaluation of Computing Systems (ММВ): 2018 г.
- International Workshop on Stochastic Modeling and Applied Research of Technology (SMARTY): 2018, 2020 г.
- International Conference on Advances in Applied Probability and Stochastic Processes (ICAAP& SP): 2019 г.
- Russian Supercomputing Days: 2020 г.
- 9th International Workshop on Energy-Efficient Data Centres E2DC Workshop, ACM E-Energy: 2021 г.
- International Conference on Applied Mathematical Models (ICAMM): 2021 г.
- Конференция международных математических центров мирового уровня Сириус: 2021 г.
- научные семинары кафедры математической статистики факультета ВМК МГУ имени М. В. Ломоносова «Теория риска и смежные вопросы», 2016 г.; кафедры теории вероятностей механико-математического факультета МГУ имени М. В. Ломоносова (под руководством Л. Г. Афанасьевой), 2016 г.; межвузовский семинар «Современные телекоммуникации и математическая теория телетрафика», 2016 г.; семинар Aalto University, School of Electrical Engineering, 2015 г.; семинар по стохастическому моделированию Department of Mathematics, CMS College Kottayam, 2018, 2020 г.; семинар Department of Information Engineering and Computer Science, Trento University, 2019 г.; семинар Institute of Data and Decision Analysis, the Chinese University of Hong-Kong, 2019 г.; семинар Department of Mathematics, Anna University Chennai, 2020; семинар Department of Mathematics, Bishop Kurialacherry College

for Women, Kottayam, 2020 г.; семинар института IMEM-CNR Trento, 2021 г.

Результаты диссертационной работы легли в основу учебно-методических комплексов для лекционных курсов, постановок задач для выпускных работ бакалавров (магистерских диссертаций) по направлениям подготовки 01.03.02 (01.04.02) «Прикладная математика и информатика» в Петрозаводском государственном университете, кандидатских диссертаций по специальности 1.2.2 «Математическое моделирование, численные методы и комплексы программ» в ИПМИ КарНЦ РАН, а также апробированы в рамках отдельных тем учебных курсов «Вероятностные модели систем обслуживания», «Многопроцессорные вычислительные системы», «Высокопроизводительный статистический анализ на языке R» в Петрозаводском государственном университете, краткосрочной учебной программы подготовки магистров «Modern Applications of Stochastic Modelling» в Central University of Kerala (Индия).

Личный вклад. Основные результаты диссертации получены лично автором. В работах [54; 1; 4–8; 66; 55; 9–21] автором диссертации выполнены постановка и решение исследовательских задач, анализ полученных результатов. В работах [56; 22; 57; 58; 2; 3; 59–61; 23; 24; 62; 25–28; 63; 29; 30; 67; 64; 31–43; 65; 44] автором развиты математические модели, методы и вычислительные алгоритмы, реализованы программные решения и выполнен численный анализ в прикладных задачах. В программах и комплексах программ [45–53] автором диссертации предложены методы, реализованы алгоритмы анализа данных в виде значимых модулей зарегистрированных инструментов.

Публикации. Основные результаты по теме диссертации изложены в 58 опубликованных работах, 3 из которых опубликованы в журналах, включенных в Перечень российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней доктора и кандидата наук, 41 — в периодических научных изданиях, индексируемых Web of Science и Scopus. Зарегистрированы 9 программ для ЭВМ. Среди работ, в которых изложены основные результаты диссертации, имеются статьи в журналах 1 и 2 квартилей [9; 14; 34; 39–41; 43].

Благодарности

Автор выражает искреннюю признательность своему научному консультанту доктору физико-математических наук, профессору Евсею Викторовичу Морозову за полезные обсуждения, ценные рекомендации и плодотворные совместные исследования.

Содержание работы

Во Введении обоснована актуальность темы диссертации, сформулированы цели, задачи, методы исследования и основные полученные результаты.

В главе 1 получен явный вид критерия стационарности для класса структурированных цепей Маркова с непрерывным временем, в которых переходы являются локально однородными. Результат применен для нахождения явного вида критерия стационарности моделей многосерверных систем обслуживания со многими классами клиентов, в том числе модели с одновременным занятием и одновременным освобождением случайного числа серверов. Такие модели применяются для исследования современных систем параллельного обслуживания, таких как вычислительные кластеры, системы беспроводной передачи данных, системы хранения данных.

В разделе 1.1 приведены основные определения и важные известные результаты, которые потребуются для исследования локально однородных структурированных цепей Маркова. В разделе 1.1.1 кратко приведены определения и известные результаты, позволяющие получить критерии стационарности и стационарное распределение цепей Маркова с непрерывным временем и счетным числом состояний. Результаты данного раздела, в частности, соотношение между стационарными распределениями цепи Маркова с непрерывным временем и вложенной цепи Маркова с дискретным временем (1.15), использованы для получения явного вида критерия стационарности многосерверных многоклассовых систем обслуживания. Модели таких систем относятся к классу структурированных цепей Маркова с непрерывным временем, критерии стационарности которых представлены в разделе 1.1.2. В разделе 1.1.3 представлены свойства группового марковского входного потока и маркированного марковского входного потока, используемых в качестве моделей входного потока клиентов в ряде моделей, рассмотренных в диссертации. В разделе 1.1.4 для подкласса структурированных цепей Маркова, в которых фаза изменяется только в моменты перехода в направлении уменьшения уровня, уточнен вид критерия стационарности с использованием вложенной цепи Маркова с дискретным временем, а также показано, что условие стационарности сохраняется при переходе от пуассоновского к групповому марковскому входному потоку. В частности, для обобщенного процесса рождения и гибели

$$\{(X(t), Y(t))\}_{t \geq 0}, \quad (1)$$

генератор которого имеет блочно-трехдиагональный вид

$$Q = \begin{bmatrix} \mathbf{A}^{0,0} & \mathbf{A}^{0,1} & \mathbf{O} & \mathbf{O} & \dots \\ \mathbf{A}^{1,0} & \mathbf{A}^{1,1} & \mathbf{A}^{(1)} & \mathbf{O} & \dots \\ \mathbf{O} & \mathbf{A}^{(-1)} & \mathbf{A}^{(0)} & \mathbf{A}^{(1)} & \dots \\ \mathbf{O} & \mathbf{O} & \mathbf{A}^{(-1)} & \mathbf{A}^{(0)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad (2)$$

доказана теорема о специальном виде критерия стационарности

$$\alpha \mathbf{A}^{(1)} \mathbf{1} < \alpha \mathbf{A}^{(-1)} \mathbf{1}, \quad (3)$$

где стохастический вектор-строка α является решением системы

$$\alpha(\mathbf{A}^{(-1)} + \mathbf{A}^{(0)} + \mathbf{A}^{(1)}) = \mathbf{0}. \quad (4)$$

Теорема 1. Пусть для неприводимого обобщенного процесса рождения и гибели (1), задаваемого с помощью матрицы (2), матрицы $\mathbf{A}^{(0)}$, $\mathbf{A}^{(1)}$ диагональные. Тогда критерий стационарности (3) имеет вид

$$\sum_{y \in \mathcal{Y}} \pi_y^{(\mathbf{A})} \rho_y < 1, \quad (5)$$

где стохастический вектор $\pi^{(\mathbf{A})}$ является решением системы

$$\pi^{(\mathbf{A})} = \pi^{(\mathbf{A})} \text{diag}(\mathbf{A}^{(-1)} \mathbf{1})^{-1} \mathbf{A}^{(-1)}, \quad (6)$$

а для каждой фазы y из множества фаз \mathcal{Y} величина ρ_y имеет вид

$$\rho_y = \frac{A_{y,y}^{(1)}}{(\mathbf{A}^{(-1)} \mathbf{1})_y}, \quad y \in \mathcal{Y}.$$

В теореме 2 показано, что теорема 1 справедлива для процесса $\{X(t), Y(t), Z(t)\}_{t \geq 0}$, в котором $Z(t)$ есть фаза группового марковского входного потока с интенсивностью, зависящей от фазы $Y(t)$. Далее доказано, что в условиях, когда фаза $Y(t)$ обладает особой структурой (*автоматного типа*), при которой ν -мерный вектор фазы системы $Y(t)$ в моменты уходов испытывает преобразование типа извлечение уходящей компоненты – сдвиг влево – добавление компоненты справа, а в моменты приходов остается неизменным, решение (6) можно получить в форме произведения:

$$\pi_{\mathbf{y}}^{(\mathbf{A})} = \prod_{m=1}^{\nu} p_{y_m}, \quad \mathbf{y} \in \mathcal{Y} = \{1, \dots, \kappa\}^{\nu}. \quad (7)$$

Для того, чтобы получить указанный вид решения, достаточно локальной однородности интенсивностей переходов при уходе клиента, что сформулировано в следующем утверждении.

Утверждение 1. Пусть для любого $\mathbf{y} \in \mathcal{Y}$, $k = 1, \dots, \kappa$, и $\mathbf{u} \in \mathcal{Y}_k^-(\mathbf{y})$, интенсивности обслуживания клиентов равны $\mu_{u_m} = \mu_k$ для всех $m = 1, \dots, \nu$ таких, что $\sigma_m(\mathbf{u}) = 1$, где $\sigma_m(\mathbf{u}) = 1$, если m -й по старшинству (в системе) клиент находится на обслуживании, когда система находится в (фиксированной) фазе \mathbf{u} , иначе $\sigma_m(\mathbf{u}) = 0$. Пусть $\sigma_{k,\mathbf{y}} := \sum_{m=1}^{\nu} \sigma_m(\mathbf{u})$ постоянно для всех $\mathbf{u} \in \mathcal{Y}_k^-(\mathbf{y})$. Тогда (7) является решением (6), если для любых $\mathbf{y} \in \mathcal{Y}$, $k = 1, \dots, \kappa$,

$$\sum_{m=1}^{\nu} \sigma_m(\varphi^{(m,k)}(\mathbf{y})) = \sigma_{k,\mathbf{y}}, \quad (8)$$

где

$$\varphi^{(m,k)}(\mathbf{y}) = (y_1, \dots, y_{m-1}, k, y_m, \dots, y_{\nu-1}).$$

Указанный результат позволил в разделе 1.1.5 получить явный вид критерия стационарности для процессов с локально однородными переходами и привести соответствующие примеры систем обслуживания, обладающих указанным свойством. Результаты раздела 1.1 применены в разделе 1.2 для анализа стационарности многосерверных многоклассовых систем обслуживания. Выполнен анализ модели суперкомпьютера (с одновременным занятием и одновременным освобождением случайного числа серверов на одно и то же случайное время), которая обладает свойством неконсервативности классической дисциплины обслуживания (т.е. допускает возможность простаивания серверов при непустой очереди). Модель суперкомпьютера является системой обслуживания с одной общей очередью и пулом из ν однотипных серверов. В моменты событий пуассоновского потока с интенсивностью λ в систему приходят клиенты ν классов. Клиенту i , поступающему в систему в момент $t_i \geq 0 =: t_1$, требуется обслуживание на случайном числе $N_i \in \{1, \dots, \nu\}$ серверов. Будем называть $N_i \leq \nu$ *классом* клиента i . Класс N типичного клиента имеет дискретное распределение $\mathbf{p} = (p_1, \dots, p_\nu)$, где

$$p_j = P\{N = j\}, \quad j = 1, \dots, \nu. \quad (9)$$

Обслуживание клиента i занимает *одно и то же* случайное время $S_i > 0$ на каждом из N_i серверов, выделенных данному клиенту. Типичное время обслуживания распределено экспоненциально, интенсивность обслуживания μ не зависит от класса клиента. Как занятие, так и освобождение всех выделенных клиенту серверов, происходит одновременно. Для такой системы можно построить цепь Маркова вида (1), где $X(t)$ есть число клиентов в системе в момент $t \geq 0$, а фазой является вектор $Y(t) = (Y_1(t), \dots, Y_{\min(\nu, X(t))}(t))$, где $Y_i(t)$ есть число серверов, требующихся для обслуживания i -го по старшинству (в порядке поступления) клиента среди $X(t)$ находящихся в системе, $i \leq \min(\nu, X(t))$. Доказано, что для модели суперкомпьютера справедливо Утверждение 1.

Теорема 3. Модель суперкомпьютера, на вход которой поступает пуассоновский входной поток с интенсивностью λ , стационарна тогда и только тогда, когда

$$\lambda C_\nu / \mu < 1. \quad (10)$$

Вид константы C_ν уточнен в утверждении 3:

$$C_\nu = \sum_{k=1}^{\nu} \frac{1}{k} \sum_{j=k}^{\nu} p_j^{k*} \sum_{t=\nu+1-j}^{\nu} p_t. \quad (11)$$

Численный метод расчета условий стационарности модели суперкомпьютера с использованием дискретной свертки на основе быстрого преобразования Фурье реализован в программном пакете `hpcwld` для программной среды вычислений R [45]. В разделе 1.2.1 выполнен анализ критерия стационарности при растущем числе серверов, а также предложен численный метод для приближенного расчета константы (11). В разделе 1.2.2 установлен явный вид критерия стационарности модели суперкомпьютера с 2 серверами, изменяемой скоростью обслуживания и изменяемой интенсивностью входного потока. Результаты, представленные в главе, опубликованы в работах [4; 6—9; 12; 18; 39].

В главе 2 модели систем параллельной обработки рассмотрены в более сложном случае, когда возможно несколько неограниченных мест для ожидания, а распределение времени обслуживания произвольно. При этом напрямую применение матрично-аналитического метода затруднительно. (Применение матрично-аналитического метода и метода вложенных цепей Маркова в случае одной неограниченной очереди клиентов и произвольного распределения времени обслуживания выполнено в разделе 5.3.) В наиболее общем случае, состояние системы имеет сложную, немарковскую структуру зависимости от предыстории. В этих условиях применение регенеративного подхода позволяет получить как (достаточные) условия стационарности, так и базовые стационарные характеристики модели. Кроме того, в случае если применение матрично-аналитического метода все же возможно, регенеративный подход позволяет упростить нахождение стационарного распределения.

В разделе 2.1 приведены основные понятия из теории регенерирующих процессов и процессов с регенерирующими приращениями, в том числе предельные теоремы, находящие непосредственное применение при анализе стационарности и оценивании базовых стационарных характеристик. В разделе 2.1.1 сформулировано свойство PASTA, при котором происходит совпадение характеристик исследуемого случайного процесса, наблюдаемых в моменты независимого пуассоновского потока и наблюдаемых усредненно по времени. В разделе 2.2 выполнен анализ стационарности и базовых стационарных характеристик неоднородной многоклассовой односерверной системы обслуживания с повторными вызовами постоянной интенсивности. В системах обслуживания с повторными вызовами в качестве пространства для ожидания клиентов используется так называемая орбита, на которую клиент направляется в случае занятости сервера в момент прихода данного клиента. В отличие от классической очереди, клиенты на орбите не наблюдают состояние сервера и делают повторные попытки попасть на обслуживание. При классической дисциплине повторных вызовов, попытки клиентов не согласованы и интенсивность попыток растет с ростом орбиты. В системах с повторными вызовами постоянной интенсивности, попытки клиентов координируются каким-либо

механизмом (например, попытки совершает только старейший из клиентов), а интенсивность не зависит от числа клиентов на орбите. Анализ стационарности систем последнего класса затруднён, поскольку интенсивность повторных вызовов с орбиты не зависит от числа находящихся на орбите клиентов. Рассмотрена односерверная система, на вход которой поступает пуассоновский поток клиентов с интенсивностью λ , класс K_n входящего клиента n определяется в момент прихода в систему с использованием дискретного распределения $\{p_1, \dots, p_\nu\}$. Время обслуживания S_n клиента n класса $K_n = m \in \{1, \dots, \nu\}$ имеет произвольное распределение с конечной средней длительностью обслуживания $ES^{(m)} = \mu_m^{-1} < \infty$. Для ожидания клиентов каждого класса предусмотрена отдельная орбита, попадание на которую в случае занятости сервера происходит с вероятностью b_m . Интенсивность (пуассоновского) потока повторных вызовов с непустой орбиты m равна $\eta_m > 0$ и не зависит от числа клиентов на данной орбите (постоянная интенсивность повторных вызовов). Анализ базовых стационарных характеристик указанной модели выполнен в разделе 2.2.2. В частности, доказана следующая теорема.

Теорема 4. Если рассматриваемая ν -классовая система стационарна, то

$$P_0 = 1 - \frac{\rho}{C} = 1 - P_b, \quad (12)$$

и для любого $m = 1, \dots, \nu$, справедливы следующие соотношения:

$$P_b^{(m)} = \rho_m \left(1 - \frac{\rho}{C} (1 - b_m) \right); \quad (13)$$

$$P_{0,b}^{(m)} = \frac{\lambda_m b_m}{\eta_m} \frac{\rho}{C}; \quad (14)$$

$$P_{0,0}^{(m)} = 1 - \left(1 + \frac{\lambda_m b_m}{\eta_m} \right) \frac{\rho}{C}, \quad (15)$$

где P_0 есть вероятность простоя сервера, $P_b = 1 - P_0$ вероятность занятости сервера, $P_b^{(m)}$ вероятность занятости сервера клиентом класса m , $P_{0,0}^{(m)}$, вероятность простоя сервера при пустой орбите m , $P_{0,b}^{(m)}$ вероятность простоя сервера при непустой орбите $m = 1, \dots, \nu$,

$$\lambda_m = \lambda p_m, \quad \rho_m = \frac{\lambda_m}{\mu_m}, \quad \rho = \sum_{m=1}^{\nu} \rho_m,$$

а константа C имеет вид

$$C = 1 + \rho - \sum_{m=1}^{\nu} \rho_m b_m.$$

Отметим, что расчет условий стационарности, базовые стационарные характеристики и дискретно-событийная модель системы со многими орбитами реализованы в программном пакете `simulato` для программной

среды вычислений R [53]. Анализ (достаточных) условий стационарности посвящен раздел 2.2.3. В разделе 2.2.4 с помощью матрично-аналитического метода установлен критерий стационарности системы в марковском случае. Результаты, представленные в данной главе, опубликованы в работах [67; 64; 34]

Глава 3 посвящена применению методов регенеративного анализа к моделям систем параллельной обработки, основанным на стохастических рекуррентных соотношениях. При таком анализе, как правило, используются процессы накопления для получения уравнений регенеративного баланса и последующего оценивания требуемой характеристики на одном цикле регенерации. Модель системы обслуживания строится с помощью одношаговых рекурсий. Получаемые при этом траектории случайных процессов являются положительно возвратными регенерирующими, что позволяет осуществить вероятностный анализ стационарных характеристик модели. Как правило, методы, представленные в данной главе, применяются к системам обслуживания в условиях произвольных распределений времен между приходами, времен обслуживания клиентов.

В разделе 3.1 представлены стохастические рекурсии как инструмент анализа, приведены примеры классических рекурсий Линдли и Кифера–Вольфовица. В разделе 3.1.1 рассмотрено понятие стохастического порядка. Это понятие использовано в разделе 3.1.2 для получения неравенства PASTA в классе распределений новое-хуже/лучше-старого. Указанное свойство может быть использовано для оценивания стационарных характеристик моделей на основе стохастических рекуррентных соотношений. В разделе 3.1.3 представлен метод доказательства стационарности моделей на основе регенеративного баланса. Указанные теоретические результаты применены в разделе 3.2 к анализу систем параллельной обработки. В частности, в разделе 3.2.1 модель суперкомпьютера, стационарность которой исследована в разделе 1.2, обобщена на случай произвольных распределений времен между приходами/времен обслуживания, а в разделе 3.2.2 на случай системы с изменяемой скоростью обслуживания (в зависимости от остаточной работы в системе в моменты прихода клиентов). Сформулирована задача снижения средней потребляемой мощности в системе при контролируемой деградации качества обслуживания. Анализ указанной задачи в явном виде проведен в разделе 3.2.3 для модели с одним сервером. Результаты, представленные в главе, опубликованы в работах [5; 66; 55; 58; 2; 3; 60; 61; 23–25; 29]. Модель суперкомпьютера на основе стохастических рекурсий реализована в программном пакете `hrcswld` для программной среды вычислений R [45].

Особенности систем параллельного обслуживания позволяют в ряде случаев выполнить анализ как в стационарном, так и в переходном режиме в явном виде, используя особенности структурированных цепей Маркова и матрично-аналитический метод. В главе 4 представлен метод нахождения

явного вида матрицы интенсивностей структурированной цепи Маркова с непрерывным временем, обладающей блочной почти верхнетреугольной матрицей в условиях малого размера блоков. Этот метод обобщает известные ранее методы получения явного решения для случая малого фазового пространства. Метод применен для анализа стационарных характеристик ряда стохастических моделей, в том числе матрично-аналитической модели когнитивного радио, модели односерверной системы с управлением скоростью обслуживания, модели вычислительной сети из персональных компьютеров. Кроме того, предложена модификация метода расщепления состояний (состоящая в особой упорядоченности состояний цепи после расщепления) для анализа матрично-аналитических моделей систем в стационарном и переходном режиме, модификация применена для анализа двухсерверной модели суперкомпьютера с неоднородными классами, а также модели когнитивного радио.

В разделе 4.1 представлены основные методы анализа стационарных характеристик структурированных цепей Маркова с непрерывным временем, в том числе метод спектрального разложения и методы численного нахождения матрицы интенсивностей, которая используется при нахождении стационарного распределения цепи в матрично-геометрическом виде. В разделе 4.1.1 представлен метод нахождения явного вида указанной матрицы в случае если размерность блоков мала, поскольку при этом с помощью теоремы Гамильтона–Кэли удастся свести систему для нахождения данной матрицы к линейной. Именно, для обобщенного процесса рождения и гибели (1), фазовое пространство которого имеет два элемента, $|\mathcal{Y}| = 2$, матрица \mathbf{R} , используемая в матрично-геометрической форме записи $\boldsymbol{\pi}_i = \boldsymbol{\pi}_{i-1}\mathbf{R}$, $i > 1$, для блоков $\boldsymbol{\pi}_i$ вектора стационарных вероятностей $\boldsymbol{\pi}$, имеет вид

$$\mathbf{R} = - \left[\frac{c_0}{c_3\xi_3} \mathbf{A}^{(-1)} + \mathbf{A}^{(1)} \right] \left[\mathbf{A}^{(0)} - \left(\frac{c_2}{c_3} + \xi_3 \right) \mathbf{A}^{(-1)} \right]^{-1}, \quad (16)$$

где константы $c_i, i = 0, \dots, 3$ являются коэффициентами полинома

$$\det(\xi^2 \mathbf{A}^{(-1)} + \xi \mathbf{A}^{(0)} + \mathbf{A}^{(1)}) / (\xi - 1) = c_3 \xi^3 + c_2 \xi^2 + c_1 \xi + c_0, \quad (17)$$

а вещественное число ξ_3 является корнем полинома (17) вне единичного круга, который можно найти в тригонометрической форме

$$\xi_3 = -\frac{c_2}{3c_3} + 2\sqrt{-p} \cos \left(\frac{1}{3} \arccos \left(\frac{q}{2p\sqrt{-p}} \right) \right),$$

где

$$p = \frac{3c_3c_1 - c_2^2}{9c_3^2}, \quad q = \frac{2c_2^3 - 9c_3c_2c_1 + 27c_3^2c_0}{27c_3^3}.$$

В разделе 4.1.2 предложена оценка спектра матрицы интенсивностей. В разделе 4.1.3 представлены методы нахождения стационарных вероятностей на нулевой границе, а также предложен способ совместного применения матрично-аналитического и регенеративного методов. Именно, для нахождения стационарного вектора $\boldsymbol{\pi}_0$ на уровне 0, применено соотношение $\boldsymbol{\pi}_0 = \boldsymbol{\pi}^{(Y)}(\mathbf{I} - \mathbf{R})$, где маргинальное распределение $\boldsymbol{\pi}^{(Y)}$ фазы Y в стационарном режиме может быть найдено с помощью регенеративного метода. Для использования данного метода необходима особая структура блоков генератора, в частности, для обобщенного процесса рождения и гибели имеющая вид

$$\mathbf{A}^{0,1} \equiv \mathbf{A}^{(1)}, \mathbf{A}^{1,1} \equiv \mathbf{A}^{(0)}, \mathbf{A}^{1,0} \equiv \mathbf{A}^{(-1)}.$$

Наконец, в разделе 4.1.4 предложена модификация метода расщепления состояний для анализа матрично-аналитических моделей систем в стационарном и переходном режиме. Суть модификации состоит в выборе порядка компонент цепи, который позволяет использовать лексикографический порядок, за счет чего возникает блочная структура конструируемых матриц. Предложенные методы применены в разделе 4.2 для анализа характеристик систем параллельной обработки в стационарном и переходном режимах. Именно, в разделе 4.2.1 выполнен анализ системы так называемого когнитивного радио, в которой канал передачи используется двумя категориями пользователей с различными приоритетами доступа к каналу. В разделе 4.2.2 исследована модель малого суперкомпьютера с изменяемой скоростью обслуживания. Наконец, в разделе 4.2.3 выполнен анализ марковской модели многоклассовой системы с повторными вызовами постоянной интенсивности, стационарность которой исследована в разделе 2.2.4. Для данной системы (в случае экспоненциально распределенных времен обслуживания) вектор $\boldsymbol{\pi}^{(Y)}$, с учетом теоремы 4, имеет вид

$$\boldsymbol{\pi}^{(Y)} = \left(\frac{1 - \rho_1}{1 + \rho - \rho_1}, \rho_1, \rho_2 \frac{1 - \rho_1}{1 + \rho - \rho_1}, \dots, \rho_\nu \frac{1 - \rho_1}{1 + \rho - \rho_1} \right),$$

а матрица \mathbf{R} также доступна в явном виде

$$\mathbf{R} = -\mathbf{A}^{(1)}(\mathbf{A}^{(0)} + \mathbf{A}^{(1)}\mathbf{1}e_2')^{-1},$$

где e_2' – второй столбец матрицы \mathbf{I} , соответствующие блоки генератора определены в главе 2 в следующем виде

$$\mathbf{A}^{(1)} = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \lambda_1 \mathbf{I} \end{bmatrix}, \mathbf{A}^{(-1)} = \begin{bmatrix} 0 & \eta_1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \mathbf{A}^{(0)} = \begin{bmatrix} -\lambda - \eta_1 & \boldsymbol{\lambda}' \\ \boldsymbol{\mu} & \text{diag}(-\boldsymbol{\mu} - \lambda_1 \mathbf{1}) \end{bmatrix},$$

а вектора $\boldsymbol{\lambda}$ и $\boldsymbol{\mu}$ содержат интенсивности приходов (обслуживания). Результаты, представленные в главе, опубликованы в работах [14; 20; 21; 30; 32; 38].

В ряде систем обслуживания клиенты приходят или обслуживаются группами случайного или детерминированного размера, что затрудняет анализ. К числу таких систем можно отнести вычислительные сети из персональных компьютеров, системы беспроводной и проводной передачи с разделением канала, системы хранения на основе твердотельных накопителей. Общей чертой указанных систем является сложный характер взаимодействия клиентов и обслуживающих устройств, в том числе допускающий блокировку части устройств в ожидании окончания обслуживания группы (так называемые модели типа разделения–слияния), исчерпывающее обслуживание, а также обслуживание–запасание с различным приоритетом доступа к запасу. Такие модели представлены и исследованы в главе 5.

В разделе 5.2 рассмотрен класс систем, анализ которых сводится, при определенных предположениях, к анализу классических многосерверных систем обслуживания, в которых времена обслуживания являются порядковыми статистиками конечных последовательностей н.о.р.с.в., в том числе в случае если последние имеют распределение с тяжелым хвостом. В этой связи, в разделе 5.1 представлены некоторые свойства порядковых статистик для различных классов распределений, в том числе имеющих фазовый тип, а также моментные свойства незавершенной работы в классических многосерверных системах обслуживания. В разделе 5.3 представлена система с исчерпывающим обслуживанием, занимающим случайное, не зависящее от числа обслуживаемых клиентов, время. Метод вложенной цепи Маркова применен для анализа системы с произвольным распределением времени обслуживания. Анализ для экспоненциального распределения выполнен матрично-аналитическим методом. Наконец, в разделе 5.4 матрично-аналитическим методом выполнен анализ системы обслуживания-запасания с пороговыми приоритетами доступа к запасу. Результаты, представленные в главе, опубликованы в работах [54; 1; 11; 13; 17; 56; 22; 57; 3; 59; 28; 31; 33; 35; 40; 42; 43].

Исследование систем обслуживания в предположении общих распределений времен обслуживания клиентов и входного потока общего типа представляет существенные затруднения. Как правило, для таких систем не удастся построить модель в виде процесса или цепи Маркова. В этой связи, для исследования таких систем применяют методы, относящиеся к категории методов имитационного моделирования. К таким методам формально можно отнести как метод построения стохастических рекурсий (рассмотрен в разделе 3.1), так и метод дискретно-событийного моделирования, рассмотренный в главе 6 (раздел 6.1.1). Общим для указанных классов является предположение о детерминированном характере изменений системы в интервалах между дискретными моментами времени, а также возможность использования доверительного оценивания для получения оценок характеристик системы в стационарном режиме (раздел 6.1.2).

В этой связи, особое внимание уделено построению моментов регенерации в исследуемых моделях. Применение метода дискретно-событийного моделирования к анализу систем параллельного обслуживания представлено в разделе 6.2. В частности, исследована чувствительность 1000-серверной модели кластера к параметру a на множестве $\{0.5, 1, 1.5, 2\}$, задающему форму распределения Вейбулла (интервалов между приходами)

$$P\{T \leq x\} = 1 - e^{-(x/b)^a}, \quad a, b > 0, x \geq 0, \quad (18)$$

где параметр b выбран так, чтобы обеспечить среднее время между приходами равным 10 единицам,

$$b = 10/\Gamma(1 + 1/a), \quad (19)$$

а Γ есть гамма-функция. В качестве дискретного распределения $\mathbf{p} = (p_1, \dots, p_{1000})$ классов клиентов, входящих в систему, использовано так называемое двухчастное лог-равномерное распределение, а распределение времени обслуживания представлено дискретной смесью двух гамма-распределений с плотностями f_1 и f_2 вида

$$f_i(x) = \frac{x^{\alpha_i-1}}{\beta_i^{\alpha_i}\Gamma(\alpha_i)} e^{-x/\beta_i}, \quad x \geq 0, \alpha_i, \beta_i > 0, \quad (20)$$

где соответствующие параметры имеют вид

$$\alpha_1 = 30, \beta_1 = 0.3, \alpha_2 = 330, \beta_2 = 0.03.$$

Плотность указанной смеси имеет вид

$$f(x) = p(c)f_1(x) + (1 - p(c))f_2(x), \quad (21)$$

при этом параметр $p(c)$ является класс-зависимым параметром смеси, определяемым по формуле

$$p(c) = 0.95 - 0.2c/1000.$$

Для каждого значения параметра a расчет произведен по траектории длиной 10^6 событий. Результаты эксперимента, представленные на Рисунке 1, отображают нелинейную зависимость оценки от параметра a . Регенеративное доверительное оценивание и дискретно-событийная модель системы реализованы в программном пакете `simulato` для программной среды вычислений R [53]. Результаты, представленные в главе, опубликованы в работах [10; 15; 16; 19; 62; 26; 27; 63; 36; 37; 41; 65; 44].

Расчеты и вычислительные эксперименты, результаты которых представлены в диссертации, выполнены с помощью **комплекса программ**, основу которого составляют программные модули и пакеты для среды вычислений R. Программный комплекс основан на взаимодействии модулей

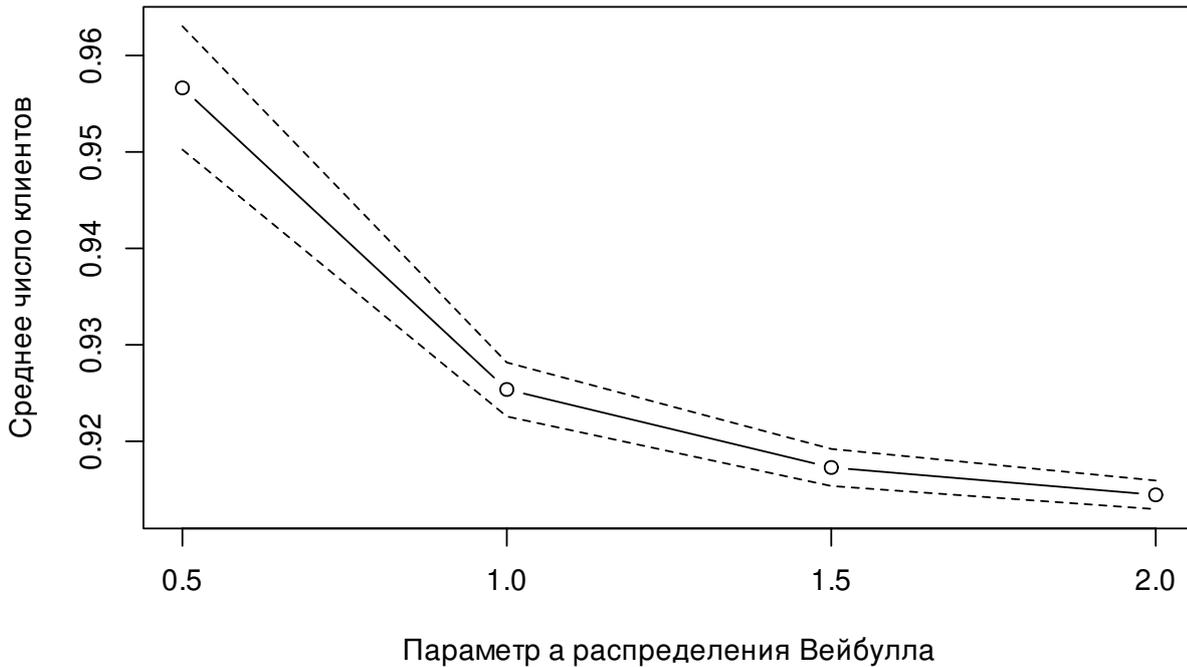


Рис. 1 — Оценка и 95 % доверительный интервал для среднего числа клиентов в системе в 1000-серверной модели суперкомпьютера, в зависимости от параметра a распределения Вейбулла (18), параметр b которого определен в (19).

с использованием общей программной среды и интегрирован в среду R. Такая интеграция позволяет использовать результаты расчетов в качестве основы для последующего анализа и визуализации как встроенными средствами языка R, так и с помощью иных программных пакетов. Следует отметить, что язык и среда вычислений R являются свободным программным обеспечением и доступны для исследователей без дополнительных лицензионных отчислений.

Среди функций и возможностей, предоставляемых комплексом программ, следует отметить

- расчет условий стационарности (11) для модели суперкомпьютера с произвольным числом серверов как точно, так и приближенно (в целях сокращения времени расчета для систем с числом серверов порядка сотен тысяч), а также построение траекторий процесса нагрузки модели суперкомпьютера на основе модифицированной рекурсии Кифера–Вольфовица, в том числе с использованием данных о нагрузке в реальных системах (в структурированном формате) с помощью модулей пакета `hpcwld`;
- выполнение дискретно-событийного моделирования на основе обобщенных полумарковских процессов и доверительного оценивания ключевых характеристик производительности модели суперкомпьютера и многоорбитной многоклассовой модели системы с повторными вызовами постоянной интенсивности регенеративным

- методом, в том числе с использованием искусственной регенерации за счет расщепления плотности, с помощью основных модулей пакета `simulato` и дополнительных модельных модулей;
- построение и численное исследование матрично-аналитических моделей систем параллельной обработки, в том числе при помощи нахождения решения в явном виде (16), либо сходящимся численным методом, в стационарном и переходном режиме, с использованием группового марковского входного потока и распределений фазового типа, а также соответствующих порядковых статистик;
 - выполнение расчетов на многопроцессорных (многоядерных) и распределенных вычислительных системах, в том числе с помощью программного пакета `RBOINC` и вычислительной сети из персональных компьютеров;
 - решение задач по подбору оптимальной конфигурации параметров для снижения потребляемой мощности при контролируемой деградации качества обслуживания в моделях систем с пороговым управлением по нагрузке и рандомизированным управлением скоростью работы системы.

Следует отметить также, что в имитационных моделях могут быть использованы произвольные распределения управляющих последовательностей, в том числе распределения с тяжелым хвостом, при этом генерация требуемых последовательностей производится как штатными средствами языка, так и пакетами, доступными в репозитории пакетов расширения R.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Разработаны и исследованы модели систем параллельной обработки, учитывающие ключевые характеристики современных высокопроизводительных и распределенных вычислительных систем и систем хранения, в том числе использование многих серверов/каналов для расчета задания/передачи данных, групповое поступление/обслуживание заданий, избыточность обслуживания/хранения.
2. Предложены, разработаны, развиты и применены методы анализа и оценивания эффективности систем массового обслуживания, в том числе за счет получения явного вида критериев стационарности, анализа и оценивания стационарных характеристик стохастических моделей.
3. Разработаны алгоритмы и программы, реализующие предложенные методы для расчета характеристик моделей систем параллельной обработки, позволяющие получать характеристики в стационарном и переходном режимах, выполнять оценивание с использованием параллельных и распределенных вычислений.

Таким образом, фундаментальная научная проблема создания теоретических основ, разработки комплекса новых методов анализа и расчета показателей эффективности систем параллельной обработки решена в диссертации путем развития приложений структурированных марковских, обобщенных полумарковских и регенерирующих процессов.

Публикации автора по теме диссертации

В изданиях из списка ВАК РФ

1. *Румянцев, А. С.* Задача оптимизации времени выполнения проекта в вычислительной сети из персональных компьютеров [Текст] / А. С. Румянцев // Программные системы: теория и приложения. — 2014. — Т. 5, № 1. — С. 175—182. — URL: <http://www.mathnet.ru/rus/ps112>.
2. *Морозов, Е. В.* Модели многосерверных систем для анализа вычислительного кластера [Текст] / Е. В. Морозов, А. С. Румянцев // Труды Карельского научного центра Российской академии наук. — 2011. — Т. 5. — С. 75—86.
3. *Морозов, Е. В.* Вероятностные модели многопроцессорных систем: стационарность и моментные свойства [Текст] / Е. В. Морозов, А. С. Румянцев // Информатика и ее применения. — 2012. — Т. 6, № 3. — С. 99—106. — URL: <http://mi.mathnet.ru/ia222>.

В изданиях, входящих в международные базы цитирования Web of Science, Scopus

4. *Rumyantsev, A.* Stabilization of a high performance cluster model [Текст] / A. Rumyantsev // Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2014 6th International Congress on. — IEEE, 2014. — P. 518—521.
5. *Rumyantsev, A.* Simulating Supercomputer Workload with Hpcwld Package for R [Текст] / A. Rumyantsev // Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2014 15th International Conference on. — IEEE, 2014. — P. 138—143.
6. *Rumyantsev, A.* An HPC Upgrade/Downgrade that Provides Workload Stability [Текст] / A. Rumyantsev // Parallel Computing Technologies. Vol. 9251. — Springer International Publishing, 2015. — P. 279—284. — (Lecture Notes in Computer Science).

7. *Rumyantsev, A.* Accelerated Verification of Stability of Simultaneous Service Multiserver Systems [Текст] / A. Rumyantsev, E. Morozov // Proceedings of 2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Brno, Czech Republic, 6-8 October 2015. — IEEE, 2015. — P. 239—242.
8. *Morozov, E.* Stability Analysis of a MAP/M/s Cluster Model by Matrix-Analytic Method [Текст] / E. Morozov, A. Rumyantsev // Computer Performance Engineering: 13th European Workshop, EPEW 2016, Chios, Greece, October 5-7, 2016, Proceedings. Vol. 9951. — Cham : Springer International Publishing, 2016. — P. 63—76. — (Lecture Notes in Computer Science).
9. *Rumyantsev, A.* Stability criterion of a multiserver model with simultaneous service [Текст] / A. Rumyantsev, E. Morozov // Annals of Operations Research. — 2017. — Vol. 252, no. 1. — P. 29—39.
10. On Distributed R Computations over BOINC [Текст] / A. Rumyantsev [et al.] // Proceedings of the Third International Conference BOINC:FAST 2017. Vol. 1973. — CEUR-WS, 2017. — P. 108—113. — URL: <http://ceur-ws.org/Vol-1973/paper14.pdf>.
11. *Rumyantsev, A.* Split-Merge Model of Workunit Replication in Distributed Computing [Текст] / A. Rumyantsev, S. Chakravarthy // Proceedings of the Third International Conference BOINC:FAST 2017. Vol. 1973. — CEUR-WS, 2017. — P. 27—34. — URL: <http://ceur-ws.org/Vol-1973/paper03.pdf>.
12. Evaluating a Single-Server Queue with Asynchronous Speed Scaling [Текст] / A. Rumyantsev [et al.] // Measurement, Modelling and Evaluation of Computing Systems. Vol. 10740. — Springer International Publishing, 2018. — P. 157—172. — (Lecture Notes in Computer Science).
13. Cost and Effect of Replication and Quorum in Desktop Grid Computing [Текст] / A. Rumyantsev [et al.] // Information Technologies and Mathematical Modelling. Queueing Theory and Applications. Vol. 912. — Cham : Springer International Publishing, 2018. — P. 143—156. — (Communications in Computer and Information Science).
14. *Garimella, R. M.* On an exact solution of the rate matrix of G/M/1 -type Markov process with small number of phases [Текст] / R. M. Garimella, A. Rumyantsev // Journal of Parallel and Distributed Computing. — 2018. — Vol. 119. — P. 172—178.
15. *A. Rumyantsev.* Discrete-Event Modeling of a High-Performance Computing Cluster with Service Rate Control [Текст] / A. Rumyantsev, T. Morozova, R. Basmadjian // 2018 22nd Conference of Open Innovations Association (FRUCT). — 2018. — P. 224—231.

16. *Rumyantsev, A.* Artificial Regeneration Based Regenerative Estimation of Multiserver System with Multiple Vacations Policy [Текст] / A. Rumyantsev, I. Peshkova // Information Technologies and Mathematical Modelling. Queueing Theory and Applications. Vol. 1109. — Springer International Publishing, 2019. — P. 38—50. — (Communications in Computer and Information Science).
17. Latency/Wearout in a Flash-Based Storage System with Replication on Write [Текст] / A. Rumyantsev [et al.] // 2019 24th Conference of Open Innovations Association (FRUCT). — 2019. — P. 360—366.
18. *Rumyantsev, A.* Stability of Multiclass Multiserver Models with Automata-type Phase Transitions [Текст] / A. Rumyantsev // Proceedings of the Second International Workshop on Stochastic Modeling and Applied Research of Technology (SMARTY 2020). Vol. 2792. — 2020. — P. 213—225. — URL: <http://ceur-ws.org/Vol-2792/#paper16>.
19. *Ivashko, E.* Discrete Event Simulation Model of a Desktop Grid System [Текст] / E. Ivashko, N. Nikitina, A. Rumyantsev // Supercomputing. Vol. 1331. — Cham : Springer International Publishing, 2020. — (Communications in Computer and Information Science).
20. *Rumyantsev, A.* Steady-State and Transient Analysis of a Single Channel Cognitive Radio Model with Impatience and Balking [Текст] / A. Rumyantsev, R. M. Garimella // Applied Probability and Stochastic Processes. — Singapore : Springer Singapore, 2020. — P. 77—90. — (Infosys Science Foundation Series in Mathematical Sciences).
21. A Three-Level Modelling Approach for Asynchronous Speed Scaling in High-Performance Data Centres [Текст] / A. Rumyantsev [et al.] // Proceedings of the Twelfth ACM International Conference on Future Energy Systems. — New York, NY, USA : Association for Computing Machinery, 2021. — P. 417—423. — (e-Energy '21).
22. *Morozov, E.* Moment properties of queueing systems and networks [Текст] / E. Morozov, A. Rumyantsev // 2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2010. — Moscow : IEEE, 2010. — P. 1056—1061. — (2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT 2010).
23. *Morozov, E.* A State-Dependent Control for Green Computing [Текст] / E. Morozov, A. Rumyantsev // Information Sciences and Systems 2015. Vol. 363. — Springer International Publishing, 2015. — P. 57—67. — (Lecture Notes in Electrical Engineering).

24. *Morozov, E.* Monotonicity and stochastic bounds for simultaneous service multiserver systems [Текст] / E. Morozov, A. Rumyantsev, I. Peshkova // Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2016 8th International Congress on. — IEEE, 2016. — P. 294—297.
25. *Morozov, E.* Stability Analysis and Simulation of a State-Dependent Transmission Rate System [Текст] / E. Morozov, L. Potakhina, A. Rumyantsev // Man-Machine Interactions 4. Vol. 391. — Springer International Publishing, 2016. — P. 673—683. — (Advances in Intelligent Systems and Computing).
26. *Morozov, E.* On Regenerative Envelopes for Cluster Model Simulation [Текст] / E. Morozov, I. Peshkova, A. Rumyantsev // Distributed Computer and Communication Networks: 19th International Conference, DCCN 2016, Moscow, Russia, November 21-25, 2016, Revised Selected Papers. Vol. 678. — Cham : Springer International Publishing, 2016. — P. 222—230. — (Communications in Computer and Information Science).
27. A Regeneration-Based Estimation of High Performance Multiserver Systems [Текст] / E. Morozov [et al.] // Computer Networks: 23rd International Conference, CN 2016, Brunów, Poland, June 14-17, 2016, Proceedings. Vol. 608. — Cham : Springer International Publishing, 2016. — P. 271—282. — (Communications in Computer and Information Science).
28. A Gaussian Approximation of Runtime Estimation in a Desktop Grid Project [Текст] / E. Morozov [et al.] // 2017 9th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). — Munich, Germany : IEEE, 2017. — P. 107—111.
29. *Morozov, E.* Inequalities for Workload Process in Queues with NBU/NWU Input [Текст] / E. Morozov, A. Rumyantsev, K. Kalinina // Man-Machine Interactions 5: 5th International Conference on Man-Machine Interactions, ICMMI 2017 Held at Kraków, Poland, October 3-6, 2017. Vol. 659. — Cham : Springer International Publishing, 2017. — P. 535—544. — (Advances in Intelligent Systems and Computing).
30. *Garimella, R. M.* On an Exact Solution of the Rate Matrix of Quasi-Birth-Death Process With Small Number of Phases [Текст] / R. M. Garimella, A. Rumyantsev // Proceedings: 31st European Conference on Modelling and Simulation ECMS 2017, May 23rd - May 26th, 2017. — Budapest, Hungary, 2017. — P. 713—719.
31. Survey on Deduplication Techniques in Flash-Based Storage [Текст] / I. Chernov [et al.] // 2018 22nd Conference of Open Innovations Association (FRUCT). — 2018. — P. 25—33.

32. *Garimella, R. M.* On Rate Matrix R of G/M/1-type Markov Process [Текст] / R. M. Garimella, A. Rumyantsev // Proceedings of the First International Workshop on Stochastic Modeling and Applied Research of Technology (SMARTY). Vol. 2278. — 2018. — P. 90—95. — (CEUR Workshop Proceedings). — URL: <http://ceur-ws.org/Vol-2278/#paper10>.
33. *Chakravarthy Srinivas R.* Efficient Redundancy Techniques in Cloud and Desktop Grid Systems using MAP/G/c-type Queues [Текст] / Chakravarthy Srinivas R., Rumyantsev Alexander // Open Engineering. — 2018. — Vol. 8, no. 1. — P. 17—31.
34. Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking [Текст] / E. Morozov [et al.] // Performance Evaluation. — 2019. — Vol. 134. — P. 102005.
35. Flash-Based Storage Deduplication Techniques: A Survey [Текст] / Ilya A. Chernov [et al.] // International Journal of Embedded and Real-Time Communication Systems (IJERTCS). — 2019. — Vol. 10, no. 3. — P. 32—48.
36. *Morozov, E.* On Failure Rate Comparison of Finite Multiserver Systems [Текст] / E. Morozov, I. Peshkova, A. Rumyantsev // Distributed Computer and Communication Networks. Vol. 11965. — Cham : Springer International Publishing, 2019. — P. 419—431. — (Lecture Notes in Computer Science).
37. *R. Nekrasova.* Regenerative Estimation of a Simultaneous Service Multiserver System with Speed Scaling [Текст] / R. Nekrasova, A. Rumyantsev // 2020 26th Conference of Open Innovations Association (FRUCT). — 2020. — P. 346—351.
38. *Garimella, R. M.* Autocorrelation Function Characterization of Continuous Time Markov Chains [Текст] / R. M. Garimella, D. G. Down, A. Rumyantsev // Proceedings of the Second International Workshop on Stochastic Modeling and Applied Research of Technology (SMARTY 2020). Vol. 2792. — 2020. — P. 226—234. — URL: <http://ceur-ws.org/Vol-2792/#short2>.
39. *Chakravarthy, S. R.* Analytical and simulation studies of queueing-inventory models with MAP demands in batches and positive phase type services [Текст] / S. R. Chakravarthy, A. Rumyantsev // Simulation Modelling Practice and Theory. — 2020. — Vol. 103. — P. 102092.
40. *Chakravarthy, S. R.* Analysis of a Queueing Model with Batch Markovian Arrival Process and General Distribution for Group Clearance [Текст] / S. R. Chakravarthy, Shruti, A. Rumyantsev // Methodology and Computing in Applied Probability. — 2020. — Vol. 23. — P. 1551—1579.

41. Sensitivity Analysis and Simulation of a Multiserver Queueing System with Mixed Service Time Distribution [Текст] / E. Morozov [et al.] // Mathematics. — 2020. — Vol. 8, no. 8. — P. 1277.
42. *Morozov, E. V.* Far-End-Tail Estimation of Queueing System Performance [Текст] / E. V. Morozov, I. V. Peshkova, A. S. Romyantsev // Journal of Mathematical Sciences. — 2020. — Vol. 248, no. 1. — P. 80—91.
43. A Multi-Server Heterogeneous Queueing-Inventory System with Class-Dependent Inventory Access [Текст] / K. Rasmi [et al.] // Mathematics. — 2021. — Vol. 9, no. 9.
44. *Astafiev, S. N.* Distributed Computing of R Applications Using RBOINC Package with Applications to Parallel Discrete Event Simulation [Текст] / S. N. Astafiev, A. S. Romyantsev // Distributed Computer and Communication Networks. — Cham : Springer International Publishing, 2022. — P. 396—407.

Зарегистрированные программы для ЭВМ

45. *Свидетельство о государственной регистрации программы для ЭВМ.* Пакет hrcwld для программной среды вычислений R [Текст] / А. С. Румянцев ; А. С. Румянцев. — № 2011618397 ; заявл. 02.11.2011 ; опубл. 10.01.2012, 2012610210 (Рос. Федерация).
46. *Свидетельство о государственной регистрации программы для ЭВМ.* Комплекс модулей для оценки энергоэффективности центров обработки данных [Текст] / Е. Е. Ивашко, А. С. Румянцев, А. Л. Чухарев ; И. К. РАН. — № 2014617049 ; заявл. 18.07.2014 ; опубл. 08.09.2014, 2014619073 (Рос. Федерация).
47. *Свидетельство о государственной регистрации программы для ЭВМ.* Система визуализации использования вычислительного кластера [Текст] / Е. Е. Ивашко, А. С. Головин, А. С. Румянцев ; И. К. РАН. — № 2016616145 ; заявл. 14.06.2016 ; опубл. 05.08.2016, 2016618690 (Рос. Федерация).
48. *Свидетельство о государственной регистрации программы для ЭВМ.* Клиент-серверная модель управления энергоэффективностью однопроцессорной системы массового обслуживания [Текст] / П. С. Зуева, А. С. Румянцев, А. С. Головин ; И. К. РАН. — № 2017616705 ; заявл. 10.07.2017 ; опубл. 01.09.2017, 2017619738 (Рос. Федерация).
49. *Свидетельство о государственной регистрации программы для ЭВМ.* Веб-сервис визуализации состояния вычислительного кластера [Текст] / А. Е. Шварц, А. С. Румянцев ; И. К. Р. ПетрГУ. — № 2016663684 ; заявл. 13.12.2016 ; опубл. 01.02.2017, 2017611265 (Рос. Федерация).

50. *Свидетельство о государственной регистрации программы для ЭВМ. Клиент-серверная модель управления энергоэффективностью многосерверной системы массового обслуживания* [Текст] / А. С. Румянцев, А. С. Головин, К. А. Калинина ; И. К. РАН. — № 2018663208 ; заявл. 23.11.2018 ; опубл. 13.12.2018, 2018666193 (Рос. Федерация).
51. *Свидетельство о государственной регистрации программы для ЭВМ. Программный комплекс для стохастического моделирования систем параллельного обслуживания* [Текст] / А. С. Румянцев ; А. С. Румянцев. — № 2020668108 ; заявл. 15.12.2020 ; опубл. 21.01.2021, 2021611099 (Рос. Федерация).
52. *Свидетельство о государственной регистрации программы для ЭВМ. Модуль статистической проверки цепи Маркова на нестационарность InstabilityTest* [Текст] / С. Н. Астафьев, А. С. Румянцев ; И. К. РАН. — № 2021610459 ; заявл. 19.01.2021 ; опубл. 26.01.2021, 2021611329 (Рос. Федерация).
53. *Свидетельство о государственной регистрации программы для ЭВМ. Программный пакет simulate для программной среды вычислений R* [Текст] / А. С. Румянцев [и др.] ; И. К. РАН. — № 2022611099 ; заявл. 28.01.2022 ; опубл. 07.02.2022, 2022612052 (Рос. Федерация).

В сборниках трудов конференций

54. *Potakhina, L. Optimizing performance in a heavy-tailed system: a case study* [Текст] / L. Potakhina, A. S. Rumyantsev // Proceedings of AM-ICT 2010-2011 Advances in Methods of Information and Communication Technology. — Helsinki, Finland : University of Helsinki, 2012. — P. 67—74.
55. *Румянцев, А. С. Стохастическое моделирование вычислительного кластера с пороговым управлением скоростью обслуживания* [Текст] / А. С. Румянцев, К. А. Калинина, Т. Е. Морозова // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2017): материалы Двадцатой междунар. науч. конфер. — Москва : М.: Техносфера, 2017. — С. 286—290.
56. *Morozov, E. Heavy-tailed Distributions with Applications to Broadband Communication Systems* [Текст] / E. Morozov, M. Pagano, A. Rumyantsev // Proceedings of AMICT'2007 Advances in Methods of Information and Communication Technology. Vol. 9. — Petrozavodsk: PetrSU, 2008. — P. 157—174.

57. *Морозов, Е. В.* Регенерация и корреляционные свойства стационарной задержки в одноканальной очереди [Текст] / Е. В. Морозов, А. С. Румянцев // Proceedings of International Workshop Distributed Computer and Communication Networks. Theory and Applications (DCCN-2010). — Moscow : R&D Company «Information, Networking Technologies», 2010. — С. 58—67.
58. *Морозов, Е.* Некоторые модели многопроцессорных систем обслуживания с тяжелыми хвостами [Текст] / Е. Морозов, А. Румянцев // Параллельные вычислительные технологии 2011: сборник трудов Международной научной конференции. — Челябинск : ЮУрГУ, 2011. — С. 555—566.
59. *Morozov, E. V.* Moment properties and long-range dependence of queueing processes [Текст] / E. V. Morozov, A. S. Romyantsev // Proceedings of AMICT 2010-2011 Advances in Methods of Information and Communication Technology. — Helsinki, Finland : University of Helsinki, 2012. — P. 31—38.
60. *Ивашко, Е. Е.* Задача прогнозирования нагрузки для повышения энергетической эффективности вычислительного кластера [Текст] / Е. Е. Ивашко, А. С. Румянцев, А. Л. Чухарев // Параллельные вычислительные технологии (ПаВТ'2013): труды международной научной конференции (г. Челябинск, 1–5 апреля 2013 г.) — Челябинск: Издательский центр ЮУрГУ, 2013. — С. 363—370.
61. Задача повышения энергоэффективности центров обработки данных [Текст] / Е. Е. Ивашко [и др.] // Перспективные информационные технологии (ПИТ 2014): труды Международной научно-технической конференции. — Самара: Издательство Самарского научного центра РАН, 2014. — С. 327—331.
62. *Morozov, E.* On regenerative envelopes for high performance cluster simulation [Текст] / E. Morozov, I. Peshkova, A. Romyantsev // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2016): материалы Девятнадцатой международной научной конференции. Т. 3. — Издательство РУДН, 2016. — С. 343—347.
63. *Morozov, E.* On artificial regeneration for stochastic simulation of the high performance cluster model [Текст] / E. Morozov, I. Peshkova, A. Romyantsev // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2017): материалы Девятнадцатой междунар. науч. конфер. — Москва : М.: Техносфера, 2017. — С. 95—97.

64. Simulation of multi-orbit retrial queue with constant retrial rates [Текст] / E. V. Morozov [и др.] // Распределенные компьютерные и телекоммуникационные сети : управление, вычисление, связь (DCCN-2018): материалы XXI Международной научной конференции. — 2018. — С. 203—207.
65. *Astafiev, S.* Distributed Computing of Embarrassingly Parallel R Applications using RBOINC Package [Текст] / S. Astafiev, A. Rumyantsev // Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь (DCCN-2021): материалы XXIV Междунар. научн. конфер, 20–24 сент. 2021 г., Москва. — М.: Ин-т проблем упр. им. В.А. Трапезникова Рос. акад. наук Минобрнауки РФ, 2021. — С. 155—160.

В прочих изданиях

66. *Румянцев, А. С.* Стохастическое моделирование вычислительного кластера с гистерезисным управлением скоростью обслуживания [Текст] / А. С. Румянцев, К. А. Калинина, Т. Е. Морозова // Труды Карельского научного центра Российской академии наук. — 2017. — № 8. — С. 76—85.
67. *Румянцев, А. С.* Методы регенеративного моделирования для анализа многосерверных систем обслуживания [Текст] / А. С. Румянцев, И. В. Пешкова // Труды Карельского научного центра Российской академии наук. — 2018. — № 7. — С. 68—82.