На правах рукописи

Разумчик Ростислав Валерьевич

МЕТОДЫ АНАЛИЗА И АЛГОРИТМЫ УПРАВЛЕНИЯ ЧАСТИЧНО НАБЛЮДАЕМЫМИ СТОХАСТИЧЕСКИМИ СИСТЕМАМИ ОБСЛУЖИВАНИЯ

Специальность 2.3.1 —
«Системный анализ, управление и обработка информации, статистика»

Автореферат
диссертации на соискание учёной степени
доктора физико–математических наук

Москва — 2022

Работа выполнена в Работа выполнена в Федеральном государственном учреждении «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН)

Официальные оппоненты:

Зорин Андрей Владимирович

д-р физ.-мат. наук, доцент,
Нижегородский государственный университет им. Н.И. Лобачевского, заведующий кафедрой теории вероятностей и анализа данных

Колногоров Александр Валерианович

д-р физ.-мат. наук, профессор,
Новгородский государственный университет имени Ярослава Мудрого, профессор кафедры прикладной математики и информатики

Хохлов Юрий Степанович

д-р физ.-мат. наук, профессор,
МГУ имени М.В. Ломоносова, профессор кафедры математической статистики

Ведущая организация:

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

Защита состоится 2022 г. в на заседании диссертационного совета 24.1.224.01 на базе ФИЦ ИУ РАН по адресу: 117312, Москва, проспект 60-летия Октября, 9 (конференц-зал, 1-й этаж).

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН по адресу: Москва, ул. Вавилова, д. 40 и на официальном сайте ФИЦ ИУ РАН http://www.frccsc.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, высылать по адресу: 119333, г. Москва, ул. Вавилова, д. 44, кор. 2, ученому секретарю диссертационного совета 24.1.224.01.

Автореферат разослан 2022 года.

Ученый секретарь
диссертационного совета 24.1.224.01,
канд. физ.-мат. наук, доцент И.В. Смирнов

Общая характеристика работы

Для современных суперкомпьютерных систем, систем распределенных вычислений, сетевых и производственных систем типичной является ситуация, когда взаимодействие или работу с ними необходимо организовывать в условиях неполной наблюдаемости (или, что то же, — частичной наблюдаемости, неполного информационного описания и т. п.). Неполнота эта может проявляться по-разному. Это, например, и (частичное или полное) отсутствие априорной информации о системе, и ограниченная возможность наблюдения ее состояний. В подобных ситуациях для анализа и оптимизации системы первостепенное значение приобретает умение воспользоваться теми сведениями о ней, которые имеются в распоряжении.

Если от системы в процессе функционирования поступает какая-либо дополнительная информация, то для достижения цели обычно используются методы теории адаптации[1]. Это направление исследований переживает сейчас большой подъем, что косвенно подтверждается неутихающим из года в год потоком публикаций. Их анализ показывает, что такой интерес вызван как новыми потребностями практики, так и прогрессом в области информационных технологий, который позволил поставить на реальную основу практическую реализацию[2] адаптивных алгоритмов.

Отсутствие дополнительной информации, приобретаемой в ходе взаимодействия или работы с частично наблюдаемой системой, делает фактически невозможным применение адаптивных стратегий. Развиваемое в диссертационной работе направление связано с проблемами именно этого рода, т. е. лежит в русле

[1] *Назин А.В., Позняк А.С.* Адаптивный выбор вариантов: Рекуррентные алгоритмы. — Москва: Наука. Гл. ред. физ.-мат.лит., 1986. — 288 с.; *Sragovich V.G.* Mathematical theory of adaptive control. — Singapore: World Scientific, 2006. — 492 p.

[2] *Коновалов М.Г.* Методы адаптивной обработки информации и их приложения. — М.: ИПИ РАН, 2007. — 212 с.; *Cao X.R.* Stochastic learning and optimization: A sensitivity–based approach. — Springer, 2007. — 566 p.; *Bertsekas D.P.* Dynamic programming and optimal control. 4th ed. — Belmont, MA, USA: Athena Scientific, 2012. — 1270 p.; *Sutton R., Barto A.* Reinforcement learning. 2nd ed. — Cambridge, Massachusetts; London, England: MIT Press, 2018. — 552 p.

фундаментальных исследований неадаптивного характера в области стохастических систем[3] с частичной наблюдаемостью. Сейчас эта проблематика является предметом постоянного внимания научного сообщества как в России, так и за рубежом[4]. Ярким подтверждением этому служит то обстоятельство, что в нее начали проникать идеи[5], тесно связанные с машинным обучением — одной из наиболее активно развивающихся сегодня научных областей[6]. В целом, круг нерешенных и не вполне решенных здесь проблем остается широким. Связано это, во-первых, с большим диктуемым практикой разнообразием постановок: известно большое число обстоятельств, которые фактически могут быть приравнены к условиям частичной наблюдаемости. Во-вторых, зачастую к решениям не удается прийти исключительно математическими методами. Поэтому приходится обращаться к методам статистического моделирования, искать эвристические идеи и разрабатывать инженерные подходы. Таким образом, тематика диссертационной работы находится в одной из актуальных областей современной науки, в которой необходим дальнейший прогресс.

Целью диссертационной работы является решение фундаментальной научной проблемы — разработка комплекса вероятностных моделей и создание на их основе методов анализа и алгоритмов управления для стохастических систем обслуживания с частичной наблюдаемостью.

Для достижения поставленной цели в диссертационной работе решаются следующие **задачи**:

---

[3] *Пугачев В.С., Синицин И.Н.* Теория стохастических систем. — М.: Логос, 2004. — 1000 с.

[4] *Li D., Hu Q., Wang L., Yu D.* Statistical inference for $M_t/G/\infty$ queueing systems under incomplete observations // European Journal of Operational Research, 2019. Vol. 279. Pp. 882–901; *Economou A.* The impact of information structure on strategic behaviour in queueing systems // Queueing theory 2: Advanced trends, 2020. Pp. 137–169; *Cohen A., Saha S.* Asymptotic optimality of the generalized μ rule under model uncertainty // Stochastic Processes and Their Applications, 2021. Vol. 136. Pp. 206–236.

[5] *Mitzenmacher M.* Scheduling with predictions and the price of misprediction // Proceedings of the 11th Innovations in Theoretical Computer Science Conference, 2020. Pp. 1–18.

[6] *Соколов И.А.* Теория и практика применения методов искусственного интеллекта // Вестник РАН, 2019. Т. 89. Вып. 4. С. 365–370.

1. Разработка комплекса вероятностных моделей для анализа стационарных вероятностно–временных характеристик стохастических систем обслуживания, в которых не наблюдаются необходимые для управления очередями фактические времена обслуживания.
2. Разработка метода оценки значений стационарных вероятностно-временных характеристик частично наблюдаемых стохастических систем обслуживания на основе доступной информации о прогнозных временах обслуживания и исследование границ его применимости.
3. Разработка алгоритмов централизованного квазиоптимального управления входящими потоками (диспетчеризации) в стохастических системах с параллельным обслуживанием при полной недоступности динамической информации об их состоянии.
4. Создание для частично наблюдаемых стохастических систем с параллельным обслуживанием простых и эффективных алгоритмов централизованной диспетчеризации, позволяющих решать задачи большой размерности.

Методы исследования. Основным аппаратом для формулировки и изучения теоретических вопросов является математическая теория массового обслуживания (ТМО). Эта область математики, сложившаяся в фундаментальных работах Ф. Поллячека, К. Пальма, Д. Кендалла, Д. Линдли, П. Морана, Л. Такача, Дж. Ф.С. Кингмана, Д. Кокса, Т.Л. Саати, Л. Клейнрока, В.Е. Бенеша, Н.К. Джейсуола, С. Карлина, С. Асмуссена, М. Ньютса и др. за рубежом и А.Я. Хинчина, Б.В. Гнеденко, Б.А. Севастьянова, Ю.В. Прохорова, А.А. Боровкова, Г.П. Башарина, Г.П. Климова, А.Д. Соловьева, В.В. Калашникова, И.Н. Коваленко и многих других в нашей стране, продолжает развиваться и выделяется как разнообразием постановок задач, так и обилием применяемых математических методов исследования. Развитые в диссертации аналитические построения опираются на этот фундамент и относятся к тому направлению[7] исследований в ТМО, которое связано с изучением “неклассических” постановок задач. Для анализа

[7]См. Введение в *Печинкин А.В.* Анализ однолинейных систем массового обслуживания с различными дисциплинами обслуживания // Докт. диссертация. — 1985. — 311 с.

и обоснования эффективности алгоритмов управления широко используются методы системного анализа, оптимизации, метод статистических испытаний и дискретно–событийное имитационное моделирование.

Основные положения, выносимые на защиту:

1. Метод получения оценок значений стационарных вероятностно–временных характеристик изолированно функционирующих стохастических систем обслуживания на основе доступной информации о прогнозных временах обслуживания.
2. Доказательство эффективности предложенного метода для некоторых классов частично наблюдаемых стохастических систем обслуживания, моделируемых немарковскими системами массового обслуживания с пуассоновскими входящими потоками.
3. Методы диспетчеризации по полной предыстории в стохастических системах с параллельным обслуживанием в условиях, когда не наблюдаемы традиционно важные для решения задач оптимизации характеристики, включая показатель, подлежащий минимизации.
4. Алгоритмы квазиоптимальной диспетчеризации в частично наблюдаемых стохастических системах, составленных из параллельно работающих систем массового обслуживания с классическими дисциплинами обработки очередей.
5. Метод создания стратегий управления входящими потоками для некоторых классов стохастических систем с параллельным обслуживанием при отсутствии информации об их динамическом состоянии, основанный на использовании виртуальных вспомогательных процессов и экспериментальное обоснование его эффективности.

Научная новизна:

1. Развит аналитический аппарат анализа стационарных характеристик ранее не изучавшихся классов систем массового обслуживания инверсионного типа, допускающих не сохраняющее работу обслуживание. Показана их связь с классическими результатами в области ТМО.
2. Область применения систем массового обслуживания инверсионного типа расширена на класс задач, связанных

с получением оценок фактических значений стационарных вероятностно–временных характеристик изолированно функционирующих стохастических систем обслуживания с частичной наблюдаемостью.

3. Разработаны новые методы диспетчеризации для широкого класса частично наблюдаемых стохастических систем с параллельным обслуживанием, основанные на общей идее — использовании при управлении входящими потоками полной предыстории наблюдаемых компонент. Базирующиеся на них алгоритмы превосходят ранее известные из научной литературы.
4. Предложен новый простой и эффективный способ создания стратегий управления входящими потоками в стохастических системах с параллельным обслуживанием при отсутствии информации об их динамическом состоянии.

Соответствие паспорту специальности. Диссертационное исследование соответствует следующим пунктам паспорта специальности 2.3.1 — "Системный анализ, управление и обработка информации, статистика":

- Теоретические основы и методы системного анализа, оптимизации, управления, принятия решений и обработки информации;
- Разработка методов и алгоритмов решения задач системного анализа, оптимизации, управления, принятия решений и обработки информации;
- Методы и алгоритмы прогнозирования и оценки эффективности, качества и надежности сложных систем.

Теоретическая и практическая значимость. Теоретические результаты диссертации имеют значение для теории массового обслуживания: область применения специальной методики [8] стационарного анализа СМО расширена на новый класс систем инверсионного типа, допускающих не сохраняющее работу обслуживание. Найдена новая область приложения для задач расчета нестационарных вероятностно–временных характеристик СМО (и СеМО) по заданной последовательности интервалов между поступлениями. Разработан и обоснован новый путь решения задачи

[8] *Pechinkin A. V.* On an invariant queuing system // Math. Operationsforsch. und Statist. Ser. Optimization, 1983. Vol. 14. No. 3. Pp. 433–444.

повышения эффективности диспетчеризации в стохастических системах с параллельным обслуживанием (не требующий изменений их структуры и состава), неотъемлемой чертой которых является полное отсутствие для диспетчера динамической информации об их состоянии. Результаты диссертационной работы используются в учебном процессе Российского университета дружбы народов на факультете физико–математических и естественных наук при преподавании курсов «Имитационное моделирование», «Дискретные вероятностные модели», «Дискретные математические модели» и в Межведомственном Суперкомпьютерном Центре РАН при эксплуатации и развитии ряда суперкомпьютерных систем коллективного пользования. Предложенные в диссертации методы применялись для аналитического исследования ряда стохастических систем, разрабатываемых в рамках проекта 075–15–2020–799 Министерства науки и высшего образования Российской Федерации "Методы построения и моделирования сложных систем на основе интеллектуальных и суперкомпьютерных технологий, направленные на преодоление больших вызовов".

Достоверность полученных результатов обеспечивается строгим применением используемого математического аппарата, правильно подобранными методиками исследования, проведения вычислений и имитационного моделирования, а также согласованностью (при рассмотрении частных случаев) результатов диссертации с известными из научной литературы.

Апробация работы. Основные результаты диссертации докладывались и обсуждались на следующих международных конференциях, симпозиумах, школах и научных семинарах:

1. Европейская конференция по математическому и имитационному моделированию, ECMS (Олесунн, 2013 г.; Регенсбург, 2016 г.; Вильгельмсхафен, 2018 г.; Вильдау, 2020 г.; Эль-Кувейт, 2021 г.);
2. Европейский симпозиум по вопросам системной инженерии, EPEW (Берлин, 2017 г.; Милан, 2019 г.);
3. Первая европейская конференция по теории массового обслуживания, ECQT (Гент, 2014 г.);
4. Международный конгресс по ультрасовременным телекоммуникациям и системам управления, ICUMT (Санкт-Петербург, 2010 г., 2012 г., 2014 г.; Лиссабон, 2016 г.);

5. Международная конференция по матрично-аналитическим методам в стохастических моделях, MAM (Будапешт, 2016 г.);
6. Международный семинар по проблемам устойчивости стохастических моделей (Светлогорск, 2012 г.; Тампере, 2015 г.);
7. Международная конференция “Распределенные компьютерные и телекоммуникационные сети: управление, вычисление, связь” (Москва, 2016–2020 гг.);
8. Международная конференция по стохастическим методам (Геленджик, 2019 г.);
9. XVII и XVIII Международная конференция имени А.Ф. Терпугова “Информационные технологии и математическое моделирование” (Томск, 2018 г.; Саратов, 2019 г.);
10. II–IV школа молодых ученых ИПИ РАН (Москва, 2011–2013 гг.);
11. Научный семинар по теории массового обслуживания кафедры теории вероятностей механико-математического факультета МГУ им. М.В. Ломоносова под руководством проф. Л.Г. Афанасьевой.

Личный вклад. Все выносимые на защиту результаты получены лично автором. В [6; 8; 11; 14; 22] автором предложены методы получения оценок фактических значений стационарных вероятностно–временных характеристик частично наблюдаемых систем; доказана их состоятельность и получены соответствующие условия. В [3; 25] автором развит аналитический аппарат решения задач стационарного анализа введенного нового класса систем массового обслуживания инверсионного типа. В [2; 4; 5] автор предложил методы порождения диспетчеризаций при полном отсутствии динамической информации о состоянии систем и получил экспериментальные обоснования их состоятельности. В [7; 9; 10; 13; 20] автором разработаны алгоритмы диспетчеризации для частично наблюдаемых систем с параллельным обслуживанием и классическими дисциплинами обработки очередей, и получено экспериментальное обоснование их эффективности. В [1; 15] автором предложены квазиградиентные алгоритмы определения оптимальных значений параметров дисцпетчеризаций по наблюдениям за фазовой траекторией. В [21] автору принадлежит подход к диспетчеризации по полной предыстории. В [24] автор описал основные

отмеченные в мировой научной литературе и используемые на практике алгоритмы диспетчеризации, и их ключевые свойства. В [16;26] автором получены основные аналитические результаты и на их основе разработаны алгоритмы расчета оптимальных значений порогов. В [27] автором предложена "имитационно–адаптивная" технология решения задач планирования ресурсов и схема ее реализации.

Публикации. Основные результаты по теме диссертации изложены в 27 печатных изданиях [1–27], 14 из которых изданы в журналах, рекомендованных ВАК [1;3;4;6;9–11;15–18;21;24;25].

Структура и объем диссертации. Диссертация состоит из введения, четырех глав и заключения. Основная часть работы изложена на 274 страницах, включая 17 рисунков, 36 таблиц и список литературы из 548 наименований.

Содержание работы

Введение посвящено краткому изложению содержания диссертации, цель которого — дать общее представление о решаемой проблеме, задачах, их научной новизне, теоретической и практической значимости, а также методах исследования.

К частично наблюдаемым стохастическим системам — системам массового обслуживания (СМО), — являющимся объектом исследования в **первой** и **второй** главах диссертации, относится любая система, для которой выполнены, главным образом, два условия. Во-первых, для каждой поступающей заявки становится известным некоторое положительное число; оно считается ее прогнозным временем обслуживания и имеет смысл работы, которую, как ожидается, необходимо совершить прибору для завершения обработки заявки (см., например, нижний ряд рисунков на рис. 1). Во-вторых, та работа, которую в действительности необходимо совершить прибору для завершения ее обработки — фактическое время обслуживания, — хотя фиксируется в момент поступления заявки в систему, однако ненаблюдаема и может не совпадать с указанным для заявки прогнозным временем обслуживания (см. верхний ряд рисунков на рис. 1). Таким образом, говоря о вероятностно–временных характеристиках частично наблюдаемых СМО, необходимо отличать их прогнозные значения от фактических.

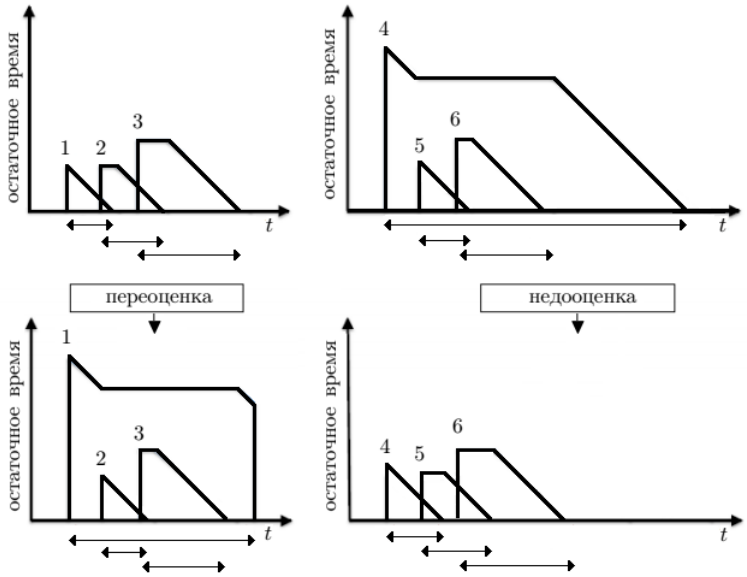


Рис. 1 — Планирование очереди в СМО $M\,|\,GI\,|\,1\,|\,\infty$ с дисциплиной обслуживания наикратчайшей заявки с прерыванием (SRPT)[9]

Поскольку для задач практики первостепенное значение имеют, вообще говоря, лишь последние, то возникает задача оценки[10] фактических значений только на основе наблюдаемых прогнозных времен обслуживания. В **главе 2** диссертации впервые предложен метод, позволяющий получать такие оценки для стационарного режима при определенных, продиктованных практикой ограничениях[11]. Выяснение условий, гарантирующих содержательность получаемых оценок, является одним из центральных результатов этой части диссертации. Идея метода заключается в преобразовании

[9] Рисунок из *Dell'Amico M., Carra D., Michiardi P.* PSBS: Practical size–based scheduling // IEEE Transactions on Computers, 2015. Vol. 99. Pp. 1–15.

[10] Рассматриваемая задача примыкает к характеризационным задачам в теории массового обслуживания (*Золотарев В.М.* Метрические расстояния в пространствах случайных величин и их распределений // Мат. сб., 1976. Т. 143. Вып. 101. С. 416–454; *Калашников В. В., Рачев С. Т.* Математические методы построения стохастических моделей обслуживания. — М.: Наука, 1988. — 311 с.). Однако к известным (прямым) задачам характеризации она не сводится.

[11] Примером одного из них (при поиске оценок сверху) является принадлежность прогнозных времен обслуживания классу случайных величин с убывающей функцией интенсивности.

остаточных прогнозных времен обслуживания заявок некоторым вероятностным механизмом, не сохраняющим работу, причем моменты преобразований синхронизированы с моментами поступления новых заявок в систему. Отсутствие решений в научной литературе, с помощью которых можно было бы объяснить наблюдаемые эффекты, послужили главным поводом для новых теоретических исследований, изложению результатов которых посвящена **глава 1**. Связаны они с развитием аналитического аппарата анализа стационарных характеристик ранее не изучавшихся классов СМО инверсионного типа.

Упомянутый выше вероятностный механизм является разновидностью предложенной в **параграфе 1.1** специальной дисциплины обслуживания — инверсионный порядок обслуживания с обобщенным вероятностным приоритетом (далее — LIFO GPP, Last–In–First–Out with Generalized Probabilistic Priority) — и его содержание посвящено выводу основных стационарных характеристик СМО $M_k \,|\, GI \,|\, 1 \,|\, n$ с этой дисциплиной. В ней постановка в очередь при поступлении новой заявки и сдвиг очереди при изменении числа заявок происходит по следующему алгоритму. Если в системе есть свободные места ожидания, вне зависимости от предыстории функционирования системы, при поступлении новой заявки ее (исходная) длина u сравнивается с (остаточной) длиной v заявки на приборе. С вероятностью $D(x,y|u,v)$, зависящей только от u и v, обслуживавшаяся ранее заявка продолжает обслуживаться, причем ее длина становится меньше y, а вновь поступившая становится на первое место в очереди и ее длина становится меньше x. Кроме того, с вероятностью $D^*(x,y|u,v)$, зависящей только от u и v, вновь поступившая заявка занимает прибор, вытесняя обслуживавшуюся ранее на первое место в очереди, причем длина заявки, бывшей ранее на приборе, становится меньше y, а вновь поступившей — меньше x. Если на приборе находится заявка остаточной длины v и в систему поступает заявка длины u, то с вероятностью $D_0(x|u,v)$ заявка, находящаяся на приборе, покидает систему, а поступившая заявка становится на прибор, причем ее длина становится меньше x. Кроме того, с вероятностью $D_0^*(y|u,v)$ поступившая заявка сразу же покидает систему, а заявка, находящаяся на приборе, продолжает обслуживаться, причем ее длина становится меньше y. Наконец, с вероятностью $d_0(u,v)$ обе заявки покидают

систему, а на прибор становится первая заявка из очереди. При всех u и v имеет место условие нормировки

$$D(\infty,\infty|u,v)+D^*(\infty,\infty|u,v)+D(\infty|u,v)+d_0(u,v)=1,$$

где $D(x|u,v) = D_0(x|u,v)+D_0^*(x|u,v)$. Если длина заявки на приборе становится равной нулю, то она мгновенно покидает систему и на прибор переходит первая заявка из очереди. Остальная очередь сдвигается на единицу. В случае, когда в системе нет свободных мест ожидания, каждая приходящая заявка теряется[12].

Опираясь на развитую в ряде работ других авторов теорию систем со специальными дисциплинами обслуживания, в параграфе 1.1 доказаны теоремы, решающие в общем вопросы расчета стационарного распределения очереди, а также нахождения (в терминах преобразований) стационарных распределений основных временных характеристик поступающих в систему заявок. В частности, получен следующий результат. Введем n-мерный случайный процесс $\eta(t)$, описывающий функционирование системы, как вектор длин заявок, находящихся в системе в момент t и расположенных в порядке, обратном очереди, т. е. если в момент t в системе находится $\nu(t)$ заявок, то $\xi_1(t)$ — длина заявки, находящейся на последнем месте в очереди, $\xi_2(t)$ — длина заявки на предпоследнем месте в очереди, $\ldots$, $\xi_{\nu(t)}(t)$ — длина обслуживаемой заявки, $\xi_{\nu(t)+1}(t) = \cdots = \xi_n(t) = 0$. Положим

$$P_0(t) = \mathsf{P}\{\nu(t)=0\},$$

$$P_k(t;x_1,\ldots,x_k) = \mathsf{P}\{\nu(t)=k, \xi_k(t) < x_1,\ldots,\xi_1(t) < x_k\},\ 1 \le k \le n-1,$$

$$P_k(x_1,\ldots,x_k) = \lim_{t\to\infty} P_k(t;x_1,\ldots,x_k),$$

$$P_k(x) = P_k(x,\infty,\ldots,\infty),\quad P_k = P_k(\infty),\quad p_k(x) = P_k'(x),$$

$$p_k(x_1,\ldots,x_k) = \frac{\partial^k}{\partial x_1\cdots\partial x_k} P_k(x_1,\ldots,x_k),$$

$$Q_n(t;x_1,\ldots,x_n) = \mathsf{P}\{\nu(t)=n, \xi_n(t) < x_1,\ldots,\xi_1(t) < x_n\},$$

[12]Отметим, что введенное правило обслуживания LIFO GPP содержит такие известные специальные дисциплины, как инверсионный порядок обслуживания с прерыванием или без прерывания обслуживания и инвариантную дисциплину обслуживания.

$$Q_n(x_1,\dots,x_n) = \lim_{t\to\infty} Q_n(t;x_1,\dots,x_n),$$
$$Q_n(x) = Q_n(x,\infty,\dots,\infty), \quad Q_n = Q_n(\infty),$$
$$q_n(x_1,\dots,x_n) = \frac{\partial^n}{\partial x_1\cdots\partial x_n} Q_n(x_1,\dots,x_n), \quad q_n(x) = Q_n'(x).$$

При $n=\infty$ величины $Q_n(t;x_1,\dots,x_n)$, $Q_n(x_1,\dots,x_n)$ и $q_n(x_1,\dots,x_n)$ не определяются. Обозначим через λ_k интенсивность пуассоновского потока при наличии k заявок в системе. Пусть длины заявок являются независимыми одинаково распределенными случайными величинами (сл. в.) с произвольной функцией распределения (ф. р.) $B(x)$ и средним значением $\int_0^\infty x\,dB(x) = \mathsf{E}S < \infty$. Предположим, что плотности $b(x) = B'(x)$, p_k и q_n существуют, являются ограниченными и непрерывными, а также что функции D, D^*, D_0 и D_0^* имеют непрерывные ограниченные плотности d, d^*, d_0 и d_0^* соответственно.

Теорема 1[13]. *Для СМО $M_k\,|\,GI\,|\,1\,|\,n\,|\,\mathsf{LIFO\,GPP}$ $(n \le \infty)$ стационарные вероятности состояний определяются рекуррентно из следующей системы уравнений:*

$$-p_1'(x) = \lambda_0 b(x) P_0 - \lambda_1 p_1(x) + \lambda_1\left(\int_0^\infty\int_0^\infty d(x|u,v)b(u)p_1(v)\,du\,dv + \int_0^\infty\int_0^\infty\int_0^\infty (d(x,y|u,v) + d^*(y,x|u,v))b(u)p_1(v)\,dy\,du\,dv\right),$$

$$-p_k'(x_1,\dots,x_k) = -\lambda_k p_k(x_1,\dots,x_k) + \lambda_{k-1}\left(\int_0^\infty\int_0^\infty (d(x_2,x_1|u,v) + d^*(x_1,x_2|u,v))b(u)p_{k-1}(v,x_3\dots,x_k)\,du\,dv\right) + \lambda_k\left(\int_0^\infty\int_0^\infty d(x_1|u,v)b(u)p_k(v,x_2,\dots,x_k)\,du\,dv + \int_0^\infty\int_0^\infty\int_0^\infty (d(x_1,y|u,v) + d^*(y,x_1|u,v))b(u)p_k(v,x_2,\dots,x_k)\,dy\,du\,dv\right),$$

[13]Номера утверждений в автореферате совпадают с номерами в диссертации; формулировки некоторых утверждений изменены для удобства изложения.

при $1 \le k \le n-1$, *и*

$$-q_n'(x_1,\ldots,x_n) = \lambda_{n-1}\Bigg(\int_0^\infty\int_0^\infty (d(x_2,x_1|u,v)b(u)q_{n-1}(v,x_3,\ldots,x_n) + d^*(x_1,x_2|u,v)b(u)q_{n-1}(v,x_3,\ldots,x_n))\,du\,dv\Bigg),$$

с граничными условиями

$$p_1(\infty)=0,\ p_k(\infty,x_2,\ldots,x_k)=0,\ 1\le k\le n-1,\ q_n(\infty,x_2,\ldots,x_n)=0.$$

Постоянная P_0 *определяется из условия* $\sum\limits_{k=0}^{n-1} P_k + Q_n = 1$.

Для СМО рассматриваемого класса нельзя сформулировать общий критерий существования стационарного режима. Это условие зависит от конкретных параметров системы и в каждом отдельном случае нуждается в специальном исследовании. В диссертации получено достаточное условие, заключающееся в выполнении следующих соотношений:

- $\mathsf{E}S < \infty$;
- $d(x,y|u,v) = 0$ при всех u, v и $y > v$ или $x > u$;
- $d^*(x,y|u,v) = 0$ при всех u, v и $y > v$ или $x > u$;
- $d_0(x|u,v) = 0$ при всех u, v и $x > u$;
- $d_0^*(y|u,v) = 0$ при всех u, v и $y > v$,

и $\lim_{k\to\infty} \lambda_k \mathsf{E}S < 1$ при $n = \infty$.

Параграф 1.2 посвящен более подробному изучению важнейшего частного случая дисциплины $\mathsf{LIFO\,GPP}$, когда $d(x,y|u,v) = b(x)b(y)$ и $d^*(x,y|u,v) = d(x|u,v) = d_0(u,v) = 0$ при всех u и v. Привлекая развитый аналитический аппарат, удалось существенно продвинуться в понимании работы СМО $M\,|\,GI\,|\,1\,|\,\infty$ с таким не сохраняющим работу обслуживанием, названным в диссертации инверсионный порядок обслуживания без прерывания и обслуживанием заново с новой реализацией длительности обслуживания (далее — $\mathsf{LIFO\,Re}$). Обозначим через $\beta(s)$ преобразование Лапласа–Стилтьеса (ПЛС) функции $B(x)$.

Теорема 4. *В системе* $M\,|\,GI\,|\,1\,|\,\infty\,|\,\mathsf{LIFO\ Re}$ *стационарные плотности вероятностей состояний* $p_k(x_1,\ldots,x_k)$, $k \geq 1$, *рассчитываются по формуле:*

$$p_k(x_1,\ldots,x_k) = (P_{k-1} + P_k) \int_{x_1}^{\infty} \lambda e^{-\lambda(u-x_1)} dB(u) b(x_2) \cdots b(x_k),$$

где $\{P_k = (2\beta(\lambda) - 1)(1 - \beta(\lambda))^k/\beta(\lambda)^{k+1},\ k \geq 0\}$, *— стационарное распределение общего числа заявок в системе*[14]*;*
ПЛС $u(s)$ *периода занятости имеет вид*

$$u(s) = \frac{\lambda + s - \sqrt{(\lambda+s)^2 - 4\lambda(1-\beta(s+\lambda))\beta(s+\lambda)(\lambda+s)}}{2\lambda(1-\beta(s+\lambda))};$$

ПЛС $\phi(s;x)$ *стационарного распределения времени пребывания в системе заявки длины* x *задается формулой*

$$\phi(s;x) = \frac{P_0 + (1-P_0)u(s)}{\lambda\beta(\lambda+s) + s} \left(\lambda\beta(\lambda+s) - se^{-(\lambda+s)x}\right).$$

Примечательным свойством этой системы[15] является совпадение значений среднего времени пребывания в ней произвольной заявки и средней длины периода занятости. В **параграфе 1.3**, привлекая аппарат матрично–аналитических методов, исследован случай поступления в систему $r > 1$ пуассоновских потоков заявок различных типов. Отличительной особенностью СМО $M_r\,|\,GI_r\,|\,1\,|\,\infty\,|\,\mathsf{LIFO\ Re}$ является отсутствие в ней приоритетов для входящих потоков, т. е. она не относится к известному классу приоритетных СМО. Однако предположение о том, что в системе реализована дисциплина $\mathsf{LIFO\ Re}$ оказывается достаточно сильным для проведения подробного анализа. Приведем для иллюстрации следующий результат. Обозначим

$$\lambda = \sum_{i=1}^{r} \lambda_i,\ \mathbf{A} = diag\left(\frac{\lambda_1}{\lambda}, \ldots, \frac{\lambda_r}{\lambda}\right),$$

[14] Отметим, что стационарное распределение общего числа заявок в системе образует геометрическую прогрессию и при большем числе приборов.

[15] Которое теряется при увеличении числа входящих потоков и/или приборов.

$$\mathbf{B} = diag(1-\beta_1(\lambda),\ldots,1-\beta_r(\lambda)),\ \mathbf{b}^{\mathrm{T}} = (1-\beta_1(\lambda),\ldots,1-\beta_r(\lambda)),$$

$$\kappa_1 = \left(\sum_{i=1}^{r}\frac{\lambda_i}{\lambda}\beta_i(\lambda)\right)^{-1},\ \kappa_2 = \sum_{i=1}^{r}\frac{\lambda_i(1-\beta_i(\lambda))}{\lambda\beta_i(\lambda)},$$

$$\mathbf{p}_k^{\mathrm{T}} = \left(P_k^{(1)},\ldots,P_k^{(r)}\right),\ k \geq 1,$$

где λ_i — интенсивность поступления заявок i-го потока, $\beta_i(s)$ — ПЛС ф. р. $B_i(x)$ длины заявки i-го потока, $P_k^{(i)}$ — стационарная вероятность наличия k заявок в системе и на приборе заявки i-го потока.

Теорема 7. *В системе* $M_r\,|\,GI_r\,|\,1\,|\,\infty\,|\,\mathsf{LIFO\,Re}$ *($r < \infty$) стационарное распределение* $\{\boldsymbol{p}_k^{\mathrm{T}},\ k \geq 1\}$ *имеет модифицированное геометрическое распределение:*

$$\begin{aligned}\mathbf{p}_k^{\mathrm{T}} &= \mathbf{p}_1^{\mathrm{T}}\left(\mathbf{B}+\kappa_1\mathbf{b}\mathbf{b}^{\mathrm{T}}\mathbf{A}\right)^{k-1},\ k \geq 2,\\ \mathbf{p}_1^{\mathrm{T}} &= P_0\kappa_1\mathbf{b}^{\mathrm{T}}\mathbf{A},\end{aligned}$$

где $P_0 = 1-\kappa_2$ *есть стационарная вероятность отсутствия заявок в системе.*

Классический метод анализа систем обслуживания на периодах занятости позволяет получить критерий существования стационарного режима при любом натуральном r: $\kappa_2 \in (0{,}1)$. Обобщениям полученных теоретических результатов на случаи большего числа приборов и более общих входящих потоков посвящен заключительный параграф главы 1 — **параграф 1.4**. Здесь развит аналитический аппарат расчета стационарных характеристик (в терминах преобразований) однолинейных СМО с произвольным обслуживанием и неординарным пуассоновским потоком разнородных заявок, и двумя конкурирующими потоками — основным групповым пуассоновским и потоком насыщения. В обоих случаях предполагается, что в системе реализован инверсионный порядок обслуживания с вероятностным приоритетом[16] — частный случай дисциплины $\mathsf{LIFO\,GPP}$.

Вернемся к методу оценки фактических значений стационарных характеристик частично наблюдаемых СМО. Он заключается

[16] *Нагоненио В.А.* Системы массового обслуживания с инверсионным порядком обслуживания и вероятностным приоритетом // Канд. диссертация. — 1981. — 140 с.

в следующем. Для частично наблюдаемой системы (из данного множества $\mathfrak{M}$) фиксируется интересующая характеристика, стационарное распределение которой существует, и вычисляется ее прогнозное значение. Затем, исходя из имеющейся о системе информации, выбирается СМО с некоторой разновидностью дисциплины LIFO GPP, в которой значение искомой (или другой) характеристики лучше рассчитанного прогнозного значения и близко к неизвестному фактическому. Совершенно ясно, что приблизиться к фактическому значению, не зная его, можно не всегда. В **параграфе 2.1**, имея в виду получение оценок сверху[17], формулируется соответствующее достаточное условие в виде принадлежности частично наблюдаемой СМО некоторому подмножеству $\mathfrak{M}^* \subset \mathfrak{M}$. В качестве $\mathfrak{M}$ рассматривается множество, состоящее из всех возможных СМО типа

$$\Sigma \,|\, B_r(x), r \in \mathcal{R} \,|\, c \,|\, n \,|\, \mathcal{U}, \tag{1}$$

где Σ — суммарный входящий поток, $B_r(x)$ — распределение длины заявки типа r, $\mathcal{R}$ — конечное множество типов заявок, c — число идентичных приборов, n — емкость очереди, $\mathcal{U}$ — дисциплина выбора из очереди и предоставления обслуживания. Тогда $\mathfrak{M}^*$ состоит из элементов $\mathfrak{M}$, для которых выполняются следующие условия:

- найдется вероятностно–временная характеристика, стационарное распределение которой существует, и известны достаточные условия его существования. Пусть X — сл. в., имеющая это распределение;
- найдутся такие два набора $\{B_r(x), r \in \mathcal{R}\}$ и $\{\hat{B}_r(x), r \in \mathcal{R}\}$ ф. р. длин заявок, что справедливо соотношение[18]

$$X_{B_1,\ldots,B_{|\mathcal{R}|}} \overset{d}{\leq} X_{\hat{B}_1,\ldots,\hat{B}_{|\mathcal{R}|}}; \tag{2}$$

[17] Если же для частично наблюдаемых систем выполняются сформулированные ниже условия, в которых неравенства заменены на противоположные, то получаемые по методу оценки являются оценками снизу.

[18] Для двух сл. в. X и Y с ф. р. $F(x)$ и $G(x)$ выполняется соотношение $X \overset{d}{\leq} Y$, если для всех вещественных x выполняется неравенство $F(x) \geq G(x)$. Из-за особенностей предложенного решения, говоря о вероятностно–временных характеристиках, приходится указывать их зависимость от ф. р. длин заявок. Поэтому, например, если сл. в. Q — длина очереди в СМО (1), то сл. в. $Q_{B_1,\ldots,B_{|\mathcal{R}|}}$ имеет тот же смысл.

– найдется вероятностно–временная характеристика, скажем Y, функционирующей в стационарном режиме СМО с тем же входящим потоком, набором ф. р. $\{\hat{B}_r(x), r \in \mathcal{R}\}$, возможно другим числом приборов и емкостью очереди, и некоторым вариантом дисциплины $\mathsf{LIFO\,GPP}$, для которой выполняются соотношения

$$X_{B_1,\ldots,B_{|\mathcal{R}|}} \overset{d}{\leq} Y_{\hat{B}_1,\ldots,\hat{B}_{|\mathcal{R}|}} \overset{d}{\leq} X_{\hat{B}_1,\ldots,\hat{B}_{|\mathcal{R}|}}. \tag{3}$$

В **параграфе 2.2** доказывается ряд теорем, показывающих, что множество $\mathfrak{M}^*$ непусто. Рассмотрим систему $M\,|\,GI\,|\,1\,|\,\infty\,|\,\mathsf{PS}$ с интенсивностью входящего потока λ, в которой длины заявок $\hat{S}$ имеют абсолютно непрерывное распределение $\hat{B}(x)$ (с плотностью $\hat{b}$ и ПЛС $\hat{\beta}$) и конечное среднее. Пусть сл. в. N^{PS} имеет распределение, совпадающее со стационарным распределением общего числа заявок в этой СМО. Рассмотрим также систему $M\,|\,GI\,|\,1\,|\,\infty\,|\,\mathsf{LIFO\,GPP}$ с тем же входящим потоком, той же ф. р. длин заявок $\hat{B}(x)$ и дисциплиной $\mathsf{LIFO\,GPP}$, в которой $D(x,y|u,v) = \hat{B}(x)\hat{B}(y)$, а остальные определяющие дисциплину функции тождественно равны нулю. Сл. в., имеющую стационарное распределение общего числа заявок в этой СМО обозначим через $N^{\mathsf{LIFO\,Re}}$.

Будем говорить, что положительная сл. в. X имеет лог–симметричное распределение с параметром $\sigma > 0$, если $X = e^Y$ и сл. в. Y имеет симметричное (относительно нуля) распределение с плотностью $g(x) = \alpha\left(x^2/\sqrt{\sigma}\right)$, $x \in (-\infty, \infty)$, где α — некоторая положительная при $x > 0$ функция, для которой $\int_0^\infty \sqrt{u}\,\alpha(u)du = 1$. Если же для положительной сл. в. X с ф. р. $F(x)$ при всех $s \geq 0$ выполняется неравенство

$$\int_0^\infty e^{-su}(1 - F(u))du \leq \frac{\mathsf{E}X}{1 + s\mathsf{E}X},$$

то будем говорить, что X принадлежит классу $\bar{\mathscr{L}}$.

Теорема 14. *Если сл. в. $\hat{S}$ принадлежит классу $\bar{\mathscr{L}}$, то выполняется соотношение*

$$N_{\hat{B}}^{\mathsf{LIFO\,Re}} \overset{d}{\leq} N_{\hat{B}}^{\mathsf{PS}}.$$

Если дополнительно известно, что сл. в. $\hat{S}$ представима в виде произведения $S \cdot X$, причем сл. в. S и X независимы, распределение $B(x)$ сл. в. S экспоненциальное и сл. в. X имеет лог–симметричное распределение, то выполняются соотношения

$$N_B^{\mathsf{PS}} \overset{d}{\leq} N_{\hat{B}}^{\mathsf{LIFO\,Re}} \overset{d}{\leq} N_{\hat{B}}^{\mathsf{PS}}.$$

Помимо оценок стационарных характеристик, для частично наблюдаемых СМО из $\mathfrak{M}^*$ возможно и извлечение из ф. р. $\hat{B}(x) = \mathsf{P}\{\hat{S} < x\}$ прогнозных времен обслуживания некоторой информации о характеристиках ф. р. $B(x) = \mathsf{P}\{S < x\}$ ненаблюдаемых фактических времен обслуживания. В диссертации доказано, что, например, в условиях *Теоремы 14*, имеют место оценки

$$\frac{1}{\hat{b}(0)} \leq \mathsf{E}S \leq \frac{1 - \hat{\beta}(\lambda)}{\lambda \hat{\beta}(\lambda)} \leq \mathsf{E}\hat{S}. \tag{4}$$

Вопрос исчерпывающего описания $\mathfrak{M}^*$ остается открытым. В оставшейся части **параграфа 2.2** показывается, что расширение области применения предложенного метода возможно в том случае, когда интересующей характеристикой частично наблюдаемой СМО является стационарное среднее[19] время пребывания заявки в системе. Рассмотрим систему $M_r \,|\, GI_r \,|\, 1 \,|\, \infty \,|\, \mathsf{PS}$, на вход которой поступает $r > 1$ независимых пуассоновских потоков заявок различных типов соответственно с интенсивностями λ_i, $1 \leq i \leq r$. Длины $\hat{S}_i$ поступающих заявок — независимые в совокупности одинаково распределенные абсолютно непрерывные сл. в. с ф. р. $\hat{B}_i(x)$ и конечным средним. Пусть сл. в. V^{PS} имеет распределение, совпадающее со стационарным распределением времени пребывания произвольной заявки в такой СМО. Рассмотрим теперь систему $M_r \,|\, GI_r \,|\, 1 \,|\, \infty \,|\, \mathsf{LIFO\,Re}$ с теми же входящими потоками, тем же набором ф. р. длин заявок $\{\hat{B}_i(x), 1 \leq i \leq r\}$ и дисциплиной $\mathsf{LIFO\,Re}$. Сл. в., имеющую стационарное распределение периода занятости этой СМО, обозначим через $U_{\hat{B}_1,\ldots,\hat{B}_r}^{\mathsf{LIFO\,Re}}$. Тогда, если при каждом $1 \leq i \leq r$ сл. в. $\hat{S}_i$ представима в виде произведения $S_i \cdot X_i$, причем

[19] На примере дисперсии показано, что получение оценок для моментов порядка выше первого возможно.

сл. в. S_i и X_i независимы, и имеют соответственно экспоненциальное распределение (с ф. р. $B_i(x)$) и лог-симметричное распределение, то (по крайней мере) во всей области стационарности частично наблюдаемой системы выполняется двойное неравенство

$$\mathsf{E}\left(V^{\mathsf{PS}}_{B_1,\ldots,B_r}\right) \le \mathsf{E}\left(U^{\mathsf{LIFO\,Re}}_{\hat{B}_1,\ldots,\hat{B}_r}\right) \le \mathsf{E}\left(V^{\mathsf{PS}}_{\hat{B}_1,\ldots,\hat{B}_r}\right). \tag{5}$$

Соотношения (5) иллюстрируют следующее обстоятельство: для того чтобы предложенный метод получения оценок фактических значений стационарных характеристик частично наблюдаемых СМО дал содержательный результат, требуется подобрать не только подходящий вариант дисциплины $\mathsf{LIFO\,GPP}$, но и правильный оценивающий показатель. Так, для системы $M\,|\,GI\,|\,1\,|\,\infty\,|\,\mathsf{PS}$ с несколькими типами заявок (при поиске оценок сверху для стационарного среднего времени пребывания произвольной заявки в системе) подходящей дисциплиной является $\mathsf{LIFO\,Re}$, а показателем — стационарная средняя длина периода занятости.

Основным недостатком результатов, подобных тем, что представлены выше, является предположение о видах распределений сл. в. S и X. В диссертации доказано, что отказаться от этого предположения можно, но в результате возникают ограничения другого рода.

Теорема 17. *Пусть сл. в. $\hat{S}$ с ф. р. $\hat{B}(x)$ принадлежит классу $\mathscr{L}$ и представима в виде произведения $S \cdot X$, где сл. в. S и X независимы, $\mathsf{E}X \ge 1$, и распределение $B(x)$ сл. в. S абсолютно непрерывно. Тогда существует такое $n_0 \in [0{,}1]$, что при $n \in [0{,}n_0]$ во всей области стационарности частично наблюдаемой системы $M\,|\,GI\,|\,1\,|\,\infty\,|\,\mathsf{PS}$ справедливы неравенства*

$$\mathsf{E}\left(V^{\mathsf{PS}}_{B}\right) \le \frac{\mathsf{E}\hat{S}}{1-\lambda\mathsf{E}\hat{S}} - n\lambda\frac{\mathsf{E}\hat{S}^2 - 2(\mathsf{E}\hat{S})^2}{(1-\lambda\mathsf{E}\hat{S})^2} \le \mathsf{E}\left(V^{\mathsf{PS}}_{\hat{B}}\right).$$

Дальнейшие полученные аналитические результаты подсказывают условия[20], при которых предложенный метод пригоден для широкого класса частично наблюдаемых стохастических систем

[20] Это условия на загрузку системы и на распределение(я) прогнозных времен обслуживания.

обслуживания. Обсуждению наиболее интересных случаев посвящен заключительный параграф главы 2 — **параграф 2.3**. Здесь на численных примерах показано, что, если интерес представляет стационарное среднее время пребывания произвольной заявки в системе, то множество $\mathfrak{M}^*$ содержит частично наблюдаемые СМО $M\,|\,GI\,|\,1\,|\,\infty$ с дисциплинами FIFO, LIFO, RANDOM, однолинейные СМО с непуассоновскими входящими потоками и даже частично наблюдаемые системы с неконсервативными дисциплинами обслуживания (см. рис 2).

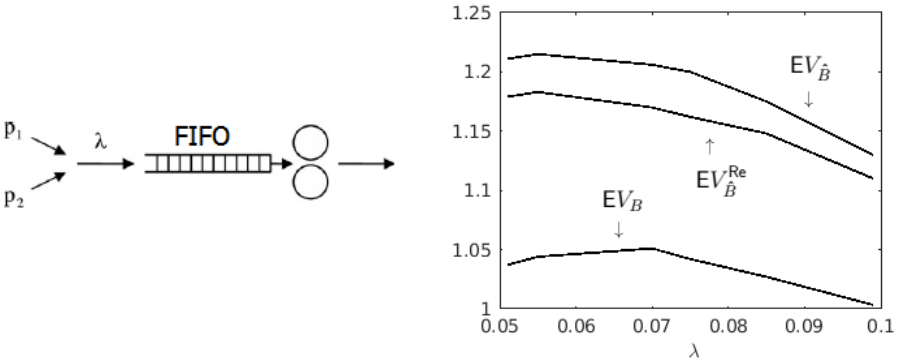


Рис. 2 — Слева — модель[21] вычислительного кластера с двумя процессорами и пуассоновским потоком заявок (p_i — вероятность того, что заявке для выполнения требуется i процессоров). Справа — иллюстрация для нее неравенства типа (5) при малой загрузке

Основным объектом исследований в **третьей** и **четвертой** главах диссертации является следующая математическая модель. В частично наблюдаемую систему из $M \geq 2$ параллельно работающих серверов поступает рекуррентный поток заданий с ф. р. $F(x)$ длины интервала между последовательными поступлениями. Задания поступают по одному, а их размеры являются независимыми одинаково распределенными сл. в. с ф. р. $B(x) = \mathsf{P}\{S < x\}$. Каждое поступившее задание должно быть немедленно направлено диспетчером на один из серверов. Серверы работают независимо, без

[21] Рисунок из *Filippopoulos D., Karatza H.* An $M/M/2$ parallel system model with pure space sharing among rigid jobs // Math. Comput. Model., 2007. Vol. 45. Pp. 491–530.

обмена заданиями и являются абсолютно надежными. В каждом имеется очередь неограниченной емкости для хранения заданий и один процессор для обработки. Индексируя серверы числами, начиная с единицы, производительность сервера m обозначается через $v^{(m)}$, причем предполагается, что хотя бы для одного m значение $v^{(m)}$ отличается от остальных. Выбор на обслуживание в каждом сервере происходит в соответствии с некоторой консервативной дисциплиной обслуживания. Пусть $t_1, t_2, \ldots, t_n, \ldots$ — последовательность моментов поступления заданий в систему, а y_n — решение, принимаемое в момент t_n относительно вновь поступившего задания. Частичная наблюдаемость подразумевает, что при принятии решения в момент t_n диспетчеру доступна только

- априорная информация о системе, включая вид ф. р. $F(x)$, $B(x)$, значения $v^{(1)}, \ldots, v^{(M)}$ и исчерпывающую информацию о состоянии серверов в момент начала функционирования, и
- информация о предыдущих моментах $t_1, t_2, \ldots, t_{n-1}$ поступления заданий в систему и принятых решениях $y_1, y_2, \ldots, y_{n-1}$.

Пусть задание, поступившее в момент t_n и обслуженное согласно правилу y_n, проведет в системе время, равное V_n. Требуется найти такую стратегию (диспетчеризацию) $y = \{y_1, y_2, \ldots\}$, которая минимизировала бы предельное среднее время пребывания задания в системе, определяемое как

$$\mathsf{E}V = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathsf{E}_y V_n, \tag{6}$$

где E_y — интегрирование по мере, порождаемой последовательностью y. Допустимыми являются диспетчеризации, представимые (детерминированной или рандомизированной) функцией вида

$$y_n = f\left(y_1, \ldots, y_{n-1}, t_1, \ldots, t_n\right), \tag{7}$$

а множество доступных наблюдений к моменту поступления n-го задания есть $\{1,2,\ldots,M\}^{n-1} \times (0,\infty)^n$.

В рассматриваемых частично наблюдаемых стохастических системах с параллельным обслуживанием диспетчеризация осуществляется в условиях, когда не наблюдаемы традиционно важные для решения задач оптимизации характеристики. Не наблюдается

даже показатель, подлежащий минимизации. Поэтому большинство диспетчеризаций[22] и приемов решения неприменимы для достижения цели — минимизации (6) на множестве стратегий (7). Из научной литературы известно всего два решения: использовать либо рандомизированную[23], либо программную стратегии.

Рандомизированная стратегия (далее — RND) описывается $(M-1)$ параметрами — вероятностями p_m выбора для очередного задания сервера m, т. е. решающее правило имеет вид

$$y_n = m, \text{ если } \sum_{j=1}^{m-1} p_j \le U < \sum_{j=1}^{m} p_j,$$

где U — равномерно распределенная на [0,1] сл.в. Проблема нахождения оптимального набора $(p_1,\ldots,p_M)$ хорошо известна[24] и наиболее полно решена для полностью марковских систем, систем с входящим пуассоновским потоком заданий и серверов типа $M\,|\,GI\,|\,1\,|\,\infty$. Ключевым обстоятельством, которое позволяет упростить решение задачи является то, что если поток поступающих в систему заданий является рекуррентным (пусть со средним λ^{-1} и коэффициентом вариации C_F), то и прореженный поток на каждый сервер также является рекуррентным. Тогда значение (6) совпадает со значением $\sum_{m=1}^{M} p_m \mathsf{E}V(m)$, где $\mathsf{E}V(m)$ — стационарное среднее время пребывания задания на сервере m, который теперь представляет собой СМО $GI\,|\,GI\,|\,1\,|\,\infty$ средним временем между поступлениями $(\lambda p_m)^{-1}$ и коэффициентом вариации $\sqrt{1+(C_F^2-1)p_m}$.

Программная стратегия (далее — PROG) параметризуется бесконечной последовательностью чисел, пусть $\{a_1,a_2,\ldots,a_{n-1},a_n,\ldots\}$, в которой a_n означает, что n-е задание направляется на сервер с номером a_n (решающее правило: $y_n = a_n$). Для произвольного числа серверов проблема нахождения оптимальной программной стратегии в настоящее время не решена и поиск

[22] Отметим наиболее известные из научной литературы: JSQ, HJSQ(d), MEST, MERL, LWL, Myopic, SITA-E, SITA-V, VITA, C-MU, LAVA, TDP, FPI, TAGS, TAPTF.

[23] Эта стратегия встречается в литературе и под другими названиями: Probabilistic Allocation Policy, Bernoulli Splitting, Random Splitting.

[24] *Ibaraki T.I., Katoh N.* Resource allocation problems: Algorithmic approaches. 2nd ed. — Cambridge: MIT Press, 1988. — 246 p.; *Combe M.B., Boxma O.J.* Optimization of static traffic allocation policies // Theoretical Computer Science, 1994. Vol. 125. Pp. 17–43.

наилучшей обычно осуществляется в два этапа. Сначала находится наилучшее (с точки зрения выбранного критерия) вероятностное распределение $\{d_1,\ldots,d_M\}$, где d_m — доля заданий, направляемых на сервер m. Затем ищется детерминированная последовательность, сохраняющая доли d_m и обеспечивающая максимальное расщепление входящего потока по серверам. В случае двух серверов (каждый из которых представляет собой СМО $G\,|\,GI\,|\,1\,|\,\infty\,|\,\mathsf{FIFO}$) при рациональном d_1 оптимальной программной стратегией является[25]:

$$a_n = \lfloor (n+1)d_1 + \phi \rfloor - \lfloor nd_1 + \phi \rfloor,\ \phi \in (-\infty,\infty),\ n \geq 1. \qquad (8)$$

В общем случае при $M \geq 3$ оптимальных правил, подобных (8), в научной литературе не представлено. Однако имеется ряд процедур для порождения эффективных последовательностей a_n. Судя по вычислительным экспериментам, из доступных в научной литературе программных стратегий, имеющих широкую область применения, наилучшие результаты удается достичь с помощью алгоритма[26]:

$$a_n = \mathrm{argmin}_{1\leq m\leq M}\left(\frac{x_m+\kappa^m(n-1)}{d_m}\right),\ x_1,\ldots,x_M \in [0,1],\ n \geq 1, \quad (9)$$

где $\kappa^m(n-1)$ есть суммарное число заданий (из первых $n-1$), направленных на сервер m.

Частичная наблюдаемость является весьма жестким ограничением на допустимые стратегии диспетчеризации, которое влечет существенный проигрыш в целевой функции по сравнению со стратегиями, использующими максимальную информацию. Оба описанных выше подхода к диспетчеризации в частично наблюдаемых системах с параллельным обслуживанием являются плодотворными, однако обладают и рядом недостатков. Во-первых, качество предоставляемых ими решений сильно зависит от предположений о характерах потоков и процессах обслуживания, и избираемых приемов для преодоления возникающих трудностей. Во-вторых, обе стратегии RND и PROG, ввиду своей универсальности, не дают глубокого понимания того, как должны управляться именно частично

[25] *Altman E., Gaujal B., Hordijk A.* Multimodularity, convexity and optimization properties // Math. Oper. Res., 2000. Vol. 25. Pp. 324–347.

[26] *Hordijk A., van der Laan D.A.* Periodic routing to parallel queues and billiard sequences // Math. Method Oper. Res., 2004. Vol. 59. No. 2. Pp. 173–192.

наблюдаемые системы. Результаты диссертации формируют такое понимание и свободны от упомянутых выше недостатков. Идея, благодаря которой это стало возможным, состоит в использовании при диспетчеризации всей доступной предыстории наблюдаемых компонент. В диссертации показано, что соответствующие стратегии существуют (далее они обозначается AA от англ. Arrival Aware) и предложено несколько способов для их нахождения. В параграфах 3.1 и 3.2 описывается аналитический подход к воплощению идеи диспетчеризации по предыстории. Параграфы 3.3 и 3.4 посвящены изложению более универсального, аналитико–имитационного подхода, значительно расширяющего тот круг систем, который очерчен результатами предыдущих двух параграфов. Наконец, глава 4 посвящена принципиально другому, простому и эффективному подходу к диспетчеризации по предыстории, свободному от тех вычислительных недостатков, которые присущи предыдущим двум подходам. Обычно новые диспетчеризации получаются параметрическими. Из-за предположения о ненаблюдаемости целевого функционала, для оценки их параметров (как и параметров стратегий RND и PROG), необходимо привлекать компьютерную модель исходной системы. Будучи основанными на принципиально иной идее, новые диспетчеризации применимы при общих предположениях о распределениях входящих потоков и размерах заданий, в случае наличия нескольких потоков, многопроцессорных серверов и т. п. В обычных условиях они гарантируют выигрыш, а в наихудших — паритет с лучшими из известных в научной литературе стратегиями.

Остановимся подробнее на результатах третьей и четвертой глав. В **параграфе 3.1** изложено решение задачи диспетчеризации в системах, в которых для величин, связанных с целевым функционалом, можно получить вычислительно реализуемые точные или хорошие приближенные формулы расчета. Полигоном для демонстрации возможностей этого подхода к управлению выбраны системы с дисциплиной FIFO в серверах. Приведем формулировку результата для случая однородных заданий. Пусть $0 < \Delta \ll 1$, $\epsilon \in [0,1)$, $\theta \in [0,1]$ и $W_{n+1}^{(m)}$ — время, необходимое для выполнения всех заданий, имеющихся на сервере m в момент t_{n+1}, без учета задания, поступившего в этот момент. Обозначим через $\{\tilde{s}_m(k), k = 0,1,\dots\}$ и $\{\tilde{w}_n^{(m)}(k), k = 0,1,\dots\}$ распределения на $\{0, \Delta, 2\Delta, \dots\}$, аппроксимирующие соответственно распределения сл. в. $S^{(m)} = S/v^{(m)}$

и $W_n^{(m)} + \mathbf{1}_{(m=y_n)} S^{(m)}$. Тогда диспетчеризация по предыстории предписывает направить поступившее в момент t_{n+1} задание на сервер с номером

$$y_{n+1} = \operatorname{argmin}_{1 \le m \le M} \left(\mathsf{E}\tilde{W}_{n+1}^{(m)} + \theta \cdot \mathsf{E}\tilde{S}^{(m)} \right), \tag{10}$$

где $\tau_n = t_{n+1} - t_n$, $\delta_n^{(m)} = \min \left(\left\lfloor \frac{\tau_n}{\Delta} \right\rfloor, \operatorname{argmax}_{k \ge 0} \left(\tilde{w}_n^{(m)}(k) > \epsilon \right) \right)$ и

$$\mathsf{E}\tilde{W}_{n+1}^{(m)} = \mathsf{E}\tilde{W}_n^{(m)} + \mathbf{1}_{(m=y_n)} \mathsf{E}\tilde{S}_n^{(m)} - \tau_n + \sum_{i=0}^{\delta_n^{(m)}} (\tau_n - i\Delta)\, \tilde{w}_n^{(m)}(i).$$

Варьирование значения ϵ позволяет изменять число компонент аппроксимирующих распределений. Для постоянного коэффициента θ, который зависит, вообще говоря, от исходных параметров, могут привлекаться специальные методы оптимизации на имитируемых траекториях. Центральное место в **параграфе 3.2** занимают результаты вычислительных экспериментов, которые свидетельствуют о том, что для рассматриваемых частично наблюдаемых стохастических систем с параллельным обслуживанием диспетчеризация по предыстории является наилучшей из известных. Например, как видно из табл. 1, в полностью марковской системе из двух серверов при любой загрузке алгоритм (10) наилучшим образом оптимизирует (6); при этом указанные значения функционалов при стратегиях RND и PROG являются неулучшаемыми.

Табл. 1 — Значения стационарного среднего (и стандартного отклонения) времени пребывания задания в системе из двух серверов ($v^{(1)} = 2/3$, $v^{(2)} = 1/3$) при различных стратегиях диспетчеризации и различной загрузке $\rho = \lambda$. Входящий поток — пуассоновский с интенсивностью λ, размер заданий имеет экспоненциальное распределение со средним 1

ρ	RND	PROG	AA
0.1	1.76 (1.76)	1.76 (1.76)	1.738 (1.76)
0.3	2.58 (2.63)	2.41 (2.45)	2.28 (2.36)
0.5	3.77 (3.87)	3.22 (3.87)	3.13 (3.23)
0.7	6.39 (6.56)	5.17 (5.29)	5.1 (5.18)
0.9	19.4 (20)	15.0 (15.17)	14.92 (15.43)

Следующая таблица дает представление о типичном поведении относительного выигрыша в (6) при стратегии AA, относительно лучшей из ранее известных (PROG), для некоторых часто употребительных распределений входящего потока и размера заданий.

Табл. 2 — Зависимость относительного выигрыша от загрузки системы ρ при стратегии AA в системе с $M = 9$ серверами ($v^{(1)} = 0.9$, $v^{(2)} = 1$, $v^{(3)} = 1.1$, $v^{(4)} = 2.9$, $v^{(5)} = 3$, $v^{(6)} = 3.1$, $v^{(7)} = 6.9$, $v^{(8)} = 7$, $v^{(9)} = 7.1$) и заданиями среднего размера 1; E_1 — экспоненциальное распределение, U — равномерное, Par — Парето, H_2 — гиперэкспоненциальное

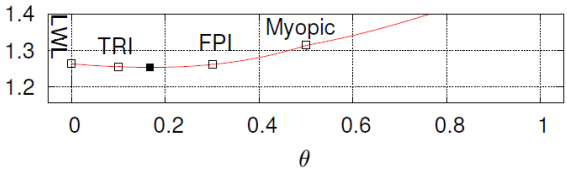
	Входящий поток / Размер задания					
ρ	E_1/E_1	E_1/U	E_1/H_2	Par/E_1	Par/U	Par/H_2
0.125	0	0.6%	0	0	0.7%	0
0.25	3.7%	0.6%	6.7%	6.6%	7.5%	12%
0.375	3.8%	6.4%	3.8%	1.3%	6.7%	6%
0.5	2.1%	4%	3.6%	1%	1.8%	3.3%
0.625	2.8%	4.3%	4.4%	1.3%	4.1%	6.2%
0.75	1.4%	3.4%	2.3%	1.5%	2.3%	1.2%
0.875	0	1.5%	1.7%	0.7%	1%	0.6%
	48%	62%	43%	13%	74%	33%

По-новому раскрывает возможности диспетчеризаций по предыстории смена целевого функционала. Как показывают вычислительные эксперименты, здесь относительный выигрыш от их применения может достигать рекордных значений. Для иллюстрации этого положения в последней строке табл. 2 приведены максимальные значения при оптимизации предельного среднего времени ожидания начала обслуживания. Заключительная часть параграфа 3.2 посвящена обсуждению и всевозможным дополнениям, расширяющим круг систем, для которых применим аналитический подход к диспетчеризации по предыстории. В частности, здесь показано, как может быть видоизменено управление, если в серверах реализована принципиально отличная от FIFO дисциплина обслуживания. Например, при дисциплине PS поступившее в момент t_{n+1} задание следует направить на сервер с номером

$$y_{n+1} = \mathrm{argmin}_{1 \le m \le M} \left(\theta \cdot \mathsf{E} N_{n+1}^{(m)} \right), \qquad (11)$$

где сл. в. $N_{n+1}^{(m)}$ — число заданий на сервере m в момент t_{n+1} (но до прибавления задания к какому-либо серверу).

Цель введения неизвестных постоянных коэффициентов θ (или, по-другому, порогов) в правила y_n — компенсация тех изменений в исходной задаче, которые вызываются различного рода аппроксимациями при расчете величин, связанных с целевым функционалом. Поскольку, судя по вычислительным экспериментам, в каждой постановке существует единственное оптимальное значение порога, то наличие хорошего начального приближения заметно упрощает поиск. Изложение решения соответствующей задачи завершает параграф 3.2. Здесь для двухсерверных систем с рекуррентным потоком заданий фиксированного размера впервые предложен итерационный алгоритм приближенного расчета значения параметра оптимальной стратегии (являющейся пороговой), требующий введения (конечной) неравномерной сетки специального (косоугольного) вида, однако не использующий ни результаты имитационного моделирования, ни вид матрицы переходных вероятностей.



θ	EV
0.160	1.25459
0.162	1.25458
0.164	1.25456
0.166	1.25454
0.168	1.25454
0.172	1.25454
0.174	1.25455

Рис. 3 — Значения стационарного среднего времени пребывания $\mathsf{E}V$ в зависимости от значения порога θ при загрузке 0.8 и различных стратегиях[27] в полностью наблюдаемой системе из двух серверов ($v^{(1)} = 1$, $v^{(2)} = 2$) с пуассоновским потоком заданий размера 1

Как показывают вычислительные эксперименты, предложенное решение (см., например, отмеченное ■ значение на рис. 3) является лучшим из известных в научной литературе.

[27]Значения для стратегий LWL, TRI, FPI, Myopic, отмеченные □, взяты из *Hyytia E.* Optimal routing of fixed size jobs to two parallel servers // INFOR: Information Systems and Operational Research, 2013. Vol. 51. No. 4. Pp. 215–224.

Круг систем, для которых можно разработать диспетчеризации по предыстории, следуя результатам параграфов 3.1 и 3.2, не является широким. Во-первых, фактически неохваченными оказываются системы с серверами, использующими сложные дисциплины обслуживания (например, SRPT). Во-вторых, можно указать условия, в которых выбор очередного действия по трудоемкости выйдет за рамки всякого разумного представления о времени выполнения. Поэтому **параграф 3.3** посвящен изложению более универсального, аналитико–имитационного подхода к задаче диспетчеризации, применимого в любой частично наблюдаемой системе с параллельным обслуживанием. Такое расширения области применения достигается путем замены ранее рассчитываемых значений величин, необходимых диспетчеру для выбора очередного действия, на их статистические оценки, получаемые посредством компьютерной модели, имитирующей процесс поступления и обслуживания заданий в системе. Модель трактуется как преобразование $\mathfrak{F} : \mathfrak{U} \to \mathfrak{B}$ исходных данных $\mathbb{U}$, принимающих значения из некоторого пространства $\mathfrak{U}$, в выходные данные $\mathbb{V}$, возможные значения которых принадлежат пространству $\mathfrak{B}$. Модель $(\mathfrak{U},\mathfrak{B},\mathfrak{F})$ является промежуточным объектом, на котором осуществляется оценка не поддающихся расчету величин, необходимых диспетчеру для выбора управления. Общая схема применения модели такова. Сначала, исходя из основного алгоритма, выбираются выходные данные $\mathbb{V}$. Затем фиксируются входные данные $\mathbb{U}$: для принятия решения y_n в качестве входных данных могут быть выбраны распределения интервалов между поступлениями и размеров заданий, предыстория совершенных действий до момента t_n. Наконец, оценки значений $\mathbb{V}$ строятся по значениям $\mathfrak{F}(\mathbb{U})$. В основу новых алгоритмов диспетчеризации по предыстории положен прием, используемый в теории адаптивного управления и известный под названием идентификационный подход: задавшись какой-нибудь эффективной стратегией, идентифицируются на компьютерной модели необходимые для ее реализации, но недоступные для наблюдения динамические характеристики серверов.

Метод разработки диспетчеризаций по предыстории на основе аналитико–имитационного подхода демонстрируется на тех же постановках, что были рассмотрены в параграфе 3.2. Для систем с двумя FIFO-серверами рассматривается пороговая стратегия

$$y_n = \begin{cases} 1, & \text{если } \mathsf{E}W_n^{(1)} + \theta < \mathsf{E}W_n^{(2)}; \\ 2, & \text{если } \mathsf{E}W_n^{(1)} + \theta \geq \mathsf{E}W_n^{(2)}, \end{cases}$$

в которой вместо точных значений $\mathsf{E}W_n^{(m)}$ используются оценки, полученные по имитируемым на основе наблюденной предыстории траекториям. Обозначим оценку $\mathsf{E}W_n^{(m)}$ к моменту t_n принятия очередного решения через $\hat{W}_n^{(m)}$ и пусть $\hat{W}_1^{(m)} = 0$. Для $n = 2$ наблюдаемая предыстория — это пара $(y_1,\Delta_1 = t_2 - t_1) = h_1$. Оценку $\hat{W}_2^{(m)}$ определим как $\mathsf{E}_{h_1}W_n^{(m)}$, где E_{h_1} — условное математическое ожидание при условии, что предыстория к моменту t_2 была h_1. Продолжая аналогичным образом, для $n = 2,\ldots,k$, определим оценки $\hat{W}_n^{(m)} = \mathsf{E}_{h_n}W_n^{(m)}$, где $h_n = (y_1,\Delta_1,\ldots,y_{n-1},\Delta_{n-1})$. Фиксированное натуральное число k будем называть глубиной памяти. Начиная с номера $n = k+1$ будем строить оценки, исходя из "усеченной" предыстории $h_{n,k} = (y_{n-k},\Delta_{n-k},\ldots,y_{n-1},\Delta_{n-1})$. Полагаем $\hat{W}_n^{(m)} = \mathsf{E}_{h_{n,k}}W_n^{(m)}$, где $\mathsf{E}_{h_{n,k}}$ — условное математическое ожидание при условии, что наблюдаемая часть предыстории на предшествующих k интервалах была $h_{n,k}$ и остаточные времена к моменту t_{n-k} равнялись $W_{n-k}^{(m)} = \hat{W}_{n-k}^{(m)}$. Фиксируя в качестве выходных данных $\mathbb{V}$ поток значений незаконченной работы на каждом сервере в момент поступления n-го задания, а в качестве входных данных $\mathbb{U} = (B,\hat{W}_{n-l}^{(1)},\ldots,\hat{W}_{n-l}^{(m)},h_{n,k})$, $l = \min(k,n-k)$, оценки $\hat{W}_n^{(m)}$ получаем путем усреднения значений $\mathfrak{F}(\mathbb{U})$. Для данного n длина имитируемого отрезка составляет l, начальное значение остаточного времени на сервере m принимается равным $\hat{W}_{n-l}^{(m)}$, а значения действий и промежутков между ними фиксированы и совпадают с наблюденной предысторией $h_{n,k}$. Важно отметить, что, несмотря на конечную глубину предыстории, используемой при расчете оценок $\mathsf{E}W_n^{(m)}$ на каждом шаге, фактически новая диспетчеризация учитывает, хотя и косвенно, всю предысторию. Это является основой ее оптимизационных возможностей (см. табл. 3).

В самом общем виде метод излагается в параграфе 3.3 на примере частично наблюдаемых стохастических систем, состоящих из параллельно работающих серверов с дисциплиной PS. Пусть при некотором (неприменимом в частично наблюдаемой системе) правиле, выбранном диспетчером, динамическое состояние сервера m в момент очередного поступления оценивается с помощью

Табл. 3 — Значения стационарного среднего (и стандартного отклонения) времени пребывания задания в системе из двух серверов ($v^{(1)} = 2/3$, $v^{(2)} = 1/3$) при различных стратегиях диспетчеризации и различной загрузке $\rho = \lambda$. Входящий поток — пуассоновский с интенсивностью λ, размер заданий имеет экспоненциальное распределение со средним 1

ρ	RND	PROG	AA
0.1	1.76 (1.76)	1.76 (1.76)	1.74 (1.76)
0.3	2.58 (2.63)	2.41 (2.45)	2.3 (2.36)
0.5	3.77 (3.87)	3.22 (3.87)	3.18 (3.22)
0.7	6.39 (6.56)	5.17 (5.29)	5.11 (5.2)
0.9	19.4 (20)	15.0 (15.17)	14.91 (15.06)

некоторой количественной оценки $\kappa^{(m)}$ или $\kappa_+^{(m)}$, соответственно без учета или с учетом поступившего задания. Алгоритмы диспетчеризации по предыстории могут быть построены по двум схемам. Первая основана на непосредственном сравнении количественных оценок состояний серверов: поступившее в момент t_n задание направляется на сервер с номером y_n, выбранным равновероятно из множества

$$\left\{ m : v^{(m)} = \max_{j \in \mathcal{J}} v^{(j)} \right\}, \tag{12}$$

где $\mathcal{J} = \{j : \kappa^{(j)} = \min_{m \in \{1,\ldots,M\}} \kappa^{(m)}\}$. В основе второй схемы — сравнение прогнозируемого увеличения количественных оценок состояний серверов: поступившее в момент t_n задание направляется на сервер с номером y_n, выбранным равновероятно из множества (12), но теперь $\mathcal{J} = \{j : \kappa^{(j)} = \min_{m \in \{1,\ldots,M\}} (\kappa_+^{(m)} - \kappa^{(m)})\}$. Реализация таких диспетчеризации по предыстории требует замены отсутствующих значений их статистическими оценками, которые получаются на основе доступных наблюдений в компьютерной модели $(\mathfrak{U},\mathfrak{B},\mathfrak{F})$. С помощью вектора наблюдений $h_{n,k}$ — предыстории к моменту t_n глубины k, а также вектора $s_{n,k} = (s_{n-k},\ldots,s_{n-1})$ независимых реализаций с.в., имеющих функции распределения объема $(n-k)$-го, …, $(n-1)$-го по счету задания, имитируется отрезок траектории системы. В начальный момент в пустую систему поступает задание объемом s_{n-k}, которое направляется на сервер y_{n-k}.

Спустя время Δ_{n-k} поступает задание объемом s_{n-k+1}, которое направляется на сервер y_{n-k+1}, и так далее, вплоть до поступления задания объемом s_{n-1}. Для n-го задания решение выбирается на основе принятого диспетчером правила: обозначим это управление $\hat{y}_1$. Повторением описанной процедуры получается набор управлений $(\hat{y}_1,\hat{y}_2,\dots)$; наиболее часто встречающееся в наборе решение выбирается в качестве искомого управления y_n.

Результаты вычислительных экспериментов, приведенные в **параграф 3.4**, свидетельствуют о том, что, даже несмотря на замену вычислений по формулам вероятностной процедурой, новые алгоритмы диспетчеризации по предыстории позволяют уменьшить значения функционалов в сравнении с ранее известными из научной литературы стратегиями почти во всем диапазоне изменений значений исходных параметров системы. Исключения составляют случаи загрузки, близкой к критической, где наблюдается совпадение результатов с наилучшими из ранее известных. Кроме того, они являются достаточно чувствительными, чтобы подтвердить установленный еще в параграфе 3.2 контринтуитивный факт: опирающиеся на наблюдения диспетчеризации могут уступать стратегиям, вовсе не использующим динамическую информацию о состоянии системы. На рис. 4 можно видеть[28], что в системе из двух серверов ($v^{(1)}=1,\ v^{(2)}=2$) с дисциплиной FIFO, с пуассоновским потоком заданий (фиксированного размера 1) интенсивности λ, диспетчеризация по наикратчайшей очереди (JSQ) уступает диспетчеризации по предыстории (AA).

В **главе 4** излагаются результаты поиска способов реализации идеи диспетчеризации по предыстории, во-первых, в еще более широком, чем ранее, классе частично наблюдаемых стохастических систем с параллельным обслуживанием и, во-вторых, полностью свободных от вычислительных недостатков. В основе нового подхода —

[28] Указанием на то, что значение целевого функционала (6) получено при той или иной диспетчеризации служит ее аббревиатура над символом $\mathsf{E}V$. Через дефис указывается, каким образом выбраны значения неизвестных параметров диспетчеризации: если оптимальным образом, то добавляется -opt; если же значения взяты из другой диспетчеризации, то добавляется ее аббревиатура. Например, $\mathsf{E}V^{\text{PROG-LB}}$ — предельное среднее время пребывания задания в системе при программной стратегии, значения неизвестных параметров которой таковы, нагрузка балансируется между серверами пропорционально их производительности.

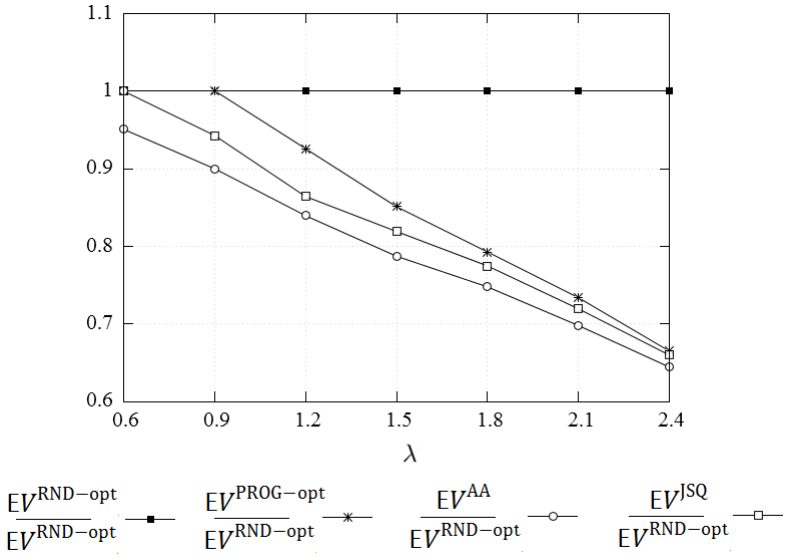


Рис. 4 — Зависимость относительного среднего времени пребывания заявки в системе от интенсивности λ входящего потока в двухсерверной системе при стратегиях PROG-opt, AA и JSQ

идея использования для принятия решений y_1,y_2,... виртуальных вспомогательных процессов, зависящих от небольшого числа неизвестных параметров, и синхронизованных по моментам поступления заданий с основной системой. Ввиду того, что априорная информация дает возможность осуществлять имитацию траектории системы, значения неизвестных параметров могут быть подобраны. Вычислительные эксперименты показывают, что новые алгоритмы успешно конкурируют со всеми ранее известными из научной литературы диспетчеризациями, а в сбалансированных системах часто превосходят их. Эти свойства вкупе с тем обстоятельством, что новые алгоритмы обычно требуют для своей настройки оценки существенно меньшего числа параметров, дают основание назвать их лучшими для частично наблюдаемых систем с параллельным обслуживанием. Рассмотренные серии экспериментов охватывают различные варианты входного потока заданий, различные распределения длины заданий, разное число серверов, различные дисциплины обслуживания. Представление о типичной динамике эффективности новых диспетчеризаций дает табл. 4.

Табл. 4 — Значения функционала (6) в зависимости от загрузки системы ρ при стратегиях RND-opt, PROG-RND-opt, PROG-LB и AA. Система с $M = 128$ серверами ($v^{(m)} = m/(64 \times 129)$) с дисциплиной FIFO, пуассоновским входящим потоком заданий, имеющих экспоненциально распределенный с параметром 1 размер

ρ	RND-opt	PROG-RND-opt	PROG-LB	AA
0.1	92	77	128	74
0.3	136	96	134	96
0.5	206	130	162	133
0.7	362	207	242	219
0.9	1126	592	665	652

При фиксированной дисциплине обслуживания в серверах наличие выигрыша от применения новых стратегий зависит, главным образом, от качества доступных оценок параметров наилучшей из известных стратегий — стратегии PROG. Если нет возможности получить близкие к оптимальным значения или приходится исходить при их выборе из здравого смысла (например, балансируя нагрузку), то новые алгоритмы следует признать наилучшим (см. последние два столбца в табл. 4). В противном случае результат сравнения зависит от соотношений между коэффициентом вариации C_B размера заданий, загрузкой системы ρ и ее размером M. Так, при фиксированном ρ, эффективность новых алгоритмов снижается с увеличением числа серверов; при этом начиная с некоторого M относительный выигрыш стабилизируется. При фиксированном M соотношение между стратегиями зависит от случайности распределения размера заданий. При $C_B \ll 1$ равномерно наилучшими по ρ являются новые стратегии. С увеличением C_B стабильного выигрыша удается добиться только в области малой загрузки (при $C_B = 1$ граница проходит, по-видимому, в районе средней загрузки). Важной отличительной особенностью новых алгоритмов является возможность естественным образом отражения в них структуры и функциональных особенностей системы. Наиболее показательным примером является ситуация с частичной доступностью серверов: при наличии точной информации о моментах выключения/включения серверов новые алгоритмы могут быть лучшими уже во всем диапазоне загрузки; в то же время для других известных стратегий такая информация является бесполезной.

В **заключении** подытоживаются основные результаты диссертации и кратко освещаются вопросы, представляющие интерес при дальнейшей разработке темы.

Список основных работ по теме диссертации

1. *Коновалов М.Г., Разумчик Р.В.* Диспетчеризация в системе с параллельным обслуживанием с помощью распределенного градиентного управления марковской цепью // *Информ. и её примен.* — 2021. — Т. 15, № 3. — С. 41–50.

2. *Konovalov M., Razumchik R.* Minimizing mean response time in batch-arrival non-observable systems with single-server FIFO queues operating in parallel // *Communications of the ECMS.* — 2021. — Vol. 35, no. 1. — Pp. 272–278.

3. *Милованова Т.А., Разумчик Р.В.* Однолинейная система массового обслуживания с инверсионным порядком обслуживания с вероятностным приоритетом, групповым пуассоновским потоком и фоновыми заявками // *Информ. и её примен.* — 2020. — Т. 14, № 3. — С. 26–34.

4. *Коновалов М.Г., Разумчик Р.В.* Об одном новом способе диспетчеризации для ненаблюдаемых систем с параллельным обслуживанием и дисциплиной FIFO в серверах // *Информационные процессы.* — 2020. — Т. 20, № 3. — С. 205–214.

5. *Konovalov M., Razumchik R.* A simple dispatching policy for minimizing mean response time in non-observable queues with SRPT policy operating in parallel // *Communications of the ECMS.* — 2020. — Vol. 34, no. 1. — Pp. 398–402.

6. *Мейханаджян Л.А., Разумчик Р.В.* Стационарные характеристики системы $M/G/2/\infty$ с одним частным случаем дисциплины инверсионного порядка обслуживания с обобщенным вероятностным приоритетом // *Информ. и её примен.* — 2020. — Т. 14, № 2. — С. 10–15.

7. *Konovalov M., Razumchik R.* Minimizing mean response time in non-observable distributed systems with processor sharing nodes //

33rd International ECMS Conference on Modelling and Simulation Proceedings. — 2019. — Vol. 33, no. 1. — Pp. 456–461.

8. *Horvath I., Razumchik R., Telek M.* The resampling $M/G/1$ non–preemptive LIFO queue and its application to systems with uncertain service time // *Perform. Eval.* — 2019. — Vol. 134. — Pp. 102000–1–102000–13 (WoS Q2).

9. *Коновалов М.Г., Разумчик Р.В.* Минимизация среднего времени пребывания в ненаблюдаемых системах с параллельным обслуживанием и дисциплиной справедливого разделения процессора в серверах // *Информационные процессы.* — 2019. — Т. 19, № 3. — С. 327–338.

10. *Коновалов М.Г., Разумчик Р.В.* Комплексное управление в одном классе систем с параллельным обслуживанием // *Информ. и её примен.* — 2019. — Т. 13, № 4. — С. 54–59.

11. *Мейханаджян Л.А., Разумчик Р.В.* Система массового обслуживания $Geo/G/1$ с инверсионным порядком обслуживания и ресамплингом в дискретном времени // *Информ. и её примен.* — 2019. — Т. 13, № 4. — С. 60–67.

12. *Razumchik R.* Two–priority queueing system with LCFS service, probabilistic priority and batch arrivals // *AIP Conference Proceedings.* — 2019. — Vol. 2116. — Pp. 090011–1–090011–3.

13. *Konovalov M., Razumchik R.* Improving routing decisions in parallel non-observable queues // *Computing.* — 2018. — Vol. 100, no. 10. — Pp. 1059–1079 (WoS Q2).

14. *Milovanova T.A., Meykhanadzhyan L.A., Razumchik R.V.* Bounding moments of sojourn time in $M/G/1$ FCFS queue with inaccurate job size information and additive error: Some observations from numerical experiments // *CEUR Workshop Procee.* — 2018. — Vol. 2236. — Pp. 24–30.

15. *Коновалов М.Г., Разумчик Р.В.* Управление случайным блужданием с эталонным стационарным распределением // *Информ. и её примен.* — 2018. — Т. 12, № 3. — С. 2–13.

16. *Коновалов М.Г., Разумчик Р.В.* Об управлении размером очереди в системе с одним сервером // *Системы и средства информ.* — 2017. — Т. 27, № 4. — С. 4–15.

17. *Разумчик Р.В.* Стационарные характеристики системы обслуживания с инверсионным порядком обслуживания, вероятностным приоритетом и групповым поступлением разнородных заявок // *Информ. и её примен.* — 2017. — Т. 11, № 4. — С. 10–18.

18. *Разумчик Р.В.* Стационарные распределения, связанные со временем пребывания в состоянии перегрузки системы $MAP/PH/1/r$ с гистерезисным управлением нагрузкой // *Информ. и её примен.* — 2017. — Т. 11, № 4. — С. 19–25.

19. *Razumchik R.* On $M/G/1$ queue with state-dependent heterogeneous batch arrivals, inverse service order and probabilistic priority // *AIP Conference Proceedings.* — 2017. — Vol. 1863. — Pp. 090006–1–090006–3.

20. *Konovalov M., Razumchik R.* Using inter-arrival times for scheduling in non-observable queues // *31st International ECMS Conference on Modelling and Simulation Proceedings.* — 2017. — Pp. 667–672.

21. *Коновалов М.Г., Разумчик Р.В.* О размещении заданий на двух серверах при неполном наблюдении // *Информ. и её примен.* — 2016. — Т. 10, № 4. — С. 57–67.

22. *Meykhanadzhyan L., Razumchik R.* New scheduling policy for estimation of stationary performance characteristics in single server queues with inaccurate job size information // *30th International ECMS Conference on Modelling and Simulation Proceedings.* — 2016. — Pp. 710–716.

23. *Razumchik R.* Analysis of finite $MAP/PH/1$ queue with hysteretic control of arrivals // *8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings.* — 2016. — Pp. 288–293.

24. *Коновалов М.Г., Разумчик Р.В.* Обзор моделей и алгоритмов размещения заданий в системах с параллельным обслуживанием // *Информ. и её примен.* — 2015. — Т. 9, № 4. — С. 56–67.

25. Стационарные вероятности состояний в системе обслуживания с инверсионным порядком обслуживания и обобщенным вероятностным приоритетом / Л.А. Мейханаджян, Т.А. Милованова, А.В. Печинкин, Р.В. Разумчик // *Информ. и её примен.* — 2015. — Т. 8, № 3. — С. 28–38.

26. *Konovalov M., Razumchik R.* Iterative algorithm for threshold calculation in the problem of routing fixed size jobs to two parallel servers // *Journal of Telecommunications and Information Technology.* — 2015. — no. 3. — Pp. 32–38.

27. *Konovalov M., Razumchik R.* Simulation of job allocation in distributed processing systems // *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops Proceedings.* — 2015. — Pp. 563–569.

28. *Печинкин А. В., Разумчик Р.В.* Системы массового обслуживания в дискретном времени. — М.: Физматлит, 2018. — 432 с.