Минобрнауки России
Федеральное государственное бюджетное учреждение науки
Федеральный исследовательский центр
«Карельский научный центр Российской академии наук» (КарНЦ РАН)

На правах рукописи

Головин Александр Станиславович

Научный доклад

об основных результатах научно-квалификационной работы на тему: «Вероятностный анализ неоднородных многосерверных систем обслуживания с применением методов точной выборки»

Направление 09.06.01 «Информатика и вычислительная техника»

Научный руководитель:
д.ф.-м.н.
Румянцев Александр Сергеевич

Петрозаводск – 2025

Введение

Актуальность. В современном мире суперкомпьютеры, являющиеся высокопроизводительными вычислительными комплексами (ВК), и центры обработки данных (ЦОД) обязательны для использования в таких областях как наука, искусственный интеллект, изучение климата, облачные сервисы, Big Data, Интернет вещей (IoT, Internet of Things), цифровая экономика и многих других. Высокая стоимость, архитектурная сложность такого специализированного оборудования приводит к необходимости оценивания характеристик его производительности и эффективности.

В данной работе рассмотрены важные задачи, связанные с исследованием ключевых стационарных характеристик производительности и энергоэффективности современных многосерверных систем и разработке методов точного оценивания этих характеристик.

Степень разработанности темы. Анализ характеристик многосерверных систем обслуживания – таких как качество обслуживания, надежность и энергоэффективность – наиболее эффективно осуществляется методами теории массового обслуживания (ТМО). Основные методы прикладного вероятностного анализа многосерверных систем обслуживания изложены в фундаментальных работах, например [9, 44, 18]. Для анализа марковских моделей многосерверных систем достаточно широко применяется матрично-аналитический метод, представленный, например, в работе [59]. Для немарковских моделей, анализ можно проводить с использованием теории регенерирующих процессов, представленной, например, в работе [57]. В то же время при исследовании сложных систем массового обслуживания, часто с нестандартными процессами входного потока или обслуживания, неоднородности заявок или обслуживающих устройств, аналитические выводы характеристик производительности могут стать трудноразрешимыми или невозможными. *Методы точной выборки* (Perfect Sampling) [48] предоставляют альтернативу получения таких характеристик на основе имитационного моделирования. Методы позволяют получать выборки непосредственно из стационарного распределения, избегая необходимости ждать достаточно длительного периода "разогрева" во время моделирования.

Работа [66], посвященная методу каплинга из прошлого ("Coupling-From-The-Past", CFTP), положила начало развитию ряда методов точной выборки (существующие методы и области применения достаточно подробно описаны, например, в работах [25, 50, 48]), которые позволяют исследовать стационарные характеристики многосерверных систем, но обычно эти исследования касались классических систем.

Ключевая особенность современных вычислительных систем в том, что одна задача может одновременно обслуживаться на множестве ресурсов (ядра, узлы, серверы), поэтому эти системы исследуются при помощи модели системы с многосерверными заявками (СМЗ). Начиная с 1960-х годов модель СМЗ изучалась, например, в работах [26, 42]. Такие модели естественным образом возникают во многих приложениях, включая телекоммуникационные и компьютерные системы. В частности, система связи с многоадресными сообщениями, где сообщение передается одновременно многим получателям, была рассмотрена в [80]. Компьютерные системы с конкуренцией за дисковое простран-

ство и программы, имеющие различные требования к размеру памяти и времени, были рассмотрены в статье [8] с использованием модели СМЗ с потерями, в которой задания, которые не могут быть немедленно обслужены, отбрасываются. Аналитические результаты были представлены для модели $M/M/2$-типа в статьях [20, 30]. В [70] с использованием матрично-аналитического подхода был получен критерий устойчивости модели $M/M/c$-типа. Большая часть недавних исследований СМЗ мотивирована необходимостью понимания и оптимизации производительности ЦОД и инфраструктуры облачных вычислений, см., например, [45, 60, 32]. Направления исследований в области СМЗ в последние годы достаточно подробно описаны в работах [71, 43]. Тем не менее, общий случай, характеризуемый произвольным распределением времени обслуживания и более чем двумя серверами, остается в значительной степени неизученным.

Интенсивный рост энергопотребления инфраструктурой информационно-коммуникационных технологий (ИКТ) стимулирует исследования в области повышения энергоэффективности инфраструктуры ИКТ. Данные исследования охватывают широкий спектр устройств ИКТ, начиная от компактных устройств с батарейным питанием, применяемых в экосистеме IoT [27, 61], и заканчивая крупномасштабными ЦОД [31, 76, 63, 36]. Значительная часть работ посвящена вопросам планирования ресурсов, балансировки нагрузки в таких системах. Однако особое внимание уделяется поиску баланса между эксплуатационными затратами, включая энергопотребление, и производительностью системы, оцениваемой через время отклика, например, [17] или исследуется зависимость потребляемой мощности в ВК и ЦОД от рабочей нагрузки, например, [22].

Тенденция к интенсивному росту энергопотребления ВК и ЦОД приводит к необходимости поиска решений, которые могут снизить потребление энергии. Одним из таких решений является внедрение так называемых "зеленых" вычислений (green computing) [82, 11, 55]. Ряд решений по снижению потребления энергии относится к аппаратной части вычислительных комплексов, другая часть решений к управлению нагрузкой [38]. Одним из методов управления нагрузкой являются схемы реагирования на спрос (DR) [29, 6], позволяющие операторам ЦОД координировать корректировки потребления мощности во время дисбаланса спроса и предложения, обеспечивая стабильность, например, из-за автоматизированных операций и гибких рабочих нагрузок, таких как отложенные задачи (резервное копирование, сканирование), которые можно перепланировать без немедленных последствий. За последнее десятилетие схемы управления нагрузкой, в том числе и DR, были предложены и широко изучены в литературе в области ЦОД [40, 39, 13, 14, 5, 3]. Также, что очень важно, в ЦОД можно найти несколько механизмов, позволяющих им участвовать в схемах DR, в том числе, которые влияют на соглашения об уровне обслуживания (SLA). Для таких механизмов (например, динамическое изменение частоты процессора (DVFS), миграция виртуальных машин), которые оказывают влияние на SLA, была предложена альтернатива – Green SLA [15, 12, 21]. Традиционные SLA обеспечивают жесткие гарантии производительности (например, доступность 99,99%), но конфликтуют с гибкостью DR. В то же время, когда в ЦОД применяются Green SLA возможны

два варианта. В режиме обычной работы ЦОД ключевые показатели производительности сохраняются такими же как и при использовании традиционных SLA. Второй вариант, когда ЦОД переходит в режим DR, чтобы снизить потребляемую мощность, клиент ЦОД соглашается, что в эти периоды показатели производительности могут быть ухудшены. Для использования Green SLA важно настроить схемы вознаграждений клиента ЦОД и штрафов для поставщиков услуг, подробности которых выходят за рамки этой работы.

Часто ресурсы ВК и ЦОД (ядра, узлы, серверы), на которых происходит обслуживание задачи, можно представить моделью односерверной системы. В односерверной системе потребление энергии может быть снижено с помощью различных методов, например, с помощью так называемого динамического масштабирования напряжения и частоты [52], методов управления спросом на потребление энергии [16] или с помощью так называемых режимов низкого энергопотребления (спящего режима) [53]. Режим низкого энергопотребления распространен как в системах IoT (например, режимы ACPI Sleep States, когда система, находясь в таком режиме, не обрабатывает/не передает данные активно. Однако часть системы (например, ответственная за обработку прерываний портов) может быть активной, чтобы вызвать пробуждение системы [10]), так и в обычных устройствах с питанием от батареи, таких как ноутбуки/смартфоны (некоторые примеры технически известны как состояния hibernate, sleep, suspend и т. д.). Обычно в таком состоянии система не может обслуживать заявки. Более того, вход и выход из неактивного состояния занимает некоторое время, в течение которого обслуживание также невозможно. Таким образом, происходит снижение производительности системы, что, однако, может привести к экономии энергии.

Таким образом, **цель** работы является вероятностный анализ неоднородных многосерверных систем обслуживания методами точной выборки.

Для достижения поставленной цели были решены следующие **задачи**:

1. Предложена модификация метода точной выборки за счет каплинга из прошлого для модели с многосерверными заявками.

2. Решена задача снижения потребляемой мощности при контролируемом качестве обслуживания в неоднородном пуле серверов.

3. С использованием разработанного комплекса программ проведены численные эксперименты для анализа эффективности предложенных методов и чувствительности полученных решений к параметрам моделей.

Научная новизна работы заключается в адаптации двух методов точной выборки к модели системы с многосерверными заявками; нахождении явного решения задачи оптимизации потребляемой мощности при контролируемом качестве обслуживания.

Методы исследования. В научно-квалификационной работе применяются модифицированные методы точной выборки на основе регенерации и на основе каплинга из прошлого.

Основные положения, выносимые на защиту:

1. Алгоритм получения выборки из распределения стационарной задержки в модели с многосерверными заявками на основе мажорантной односерверной системы с дисциплиной FCFS.

2. Алгоритм получения выборки из распределения стационарной задержки в модели с многосерверными заявками за счет каплинга из прошлого.

3. Явный вид функции распределения времени пребывания в односерверной системе с режимом снижения энергопотребления и произвольным распределением времен обслуживания заявок.

4. Разработаны алгоритмы и программы, реализующие предложенные методы для расчета ключевых характеристик неоднородных многосервеных систем обслуживания.

Практическую значимость составляют адаптированные методы, позволяющие получить стационарные характеристики качества обслуживания в системах с многосерверными заявками, и программные реализации разработанных методов.

Апробация работы Результаты работы докладывались на следующих всероссийских и международных конференциях:

1. ITMM-2021: Information Technologies and Mathematical Modeling, Tomsk, 2021 г.

2. e-Energy '22: The 13th ACM International Conference on Future Energy Systems, 2022 г.

3. RSD-2022: Russian Supercomputing Days. Moscow, 2022 г.

4. CITDS-2022: The 2022 IEEE 2nd Conference on Information Technology and Data Science. Debrecen, Hungary, 2022 г.

5. НСКФ-2022: Национальный Суперкомпьютерный Форум, Переславль-Залесский, 2022 г.

6. DCCN-2022: Distributed Computer and Communication Networks: Control, Computation, Communications, Moscow, 2022 г.

7. e-Energy '23: The 14th ACM International Conference on Future Energy Systems, Orlando, United States, 2023 г.

8. PCI-2023: 5th International Conference on Problems of Cybernetics and Informatics, Baku, Azerbaijan, 2023 г.

9. PMDM-2024: Вероятностные методы в дискретной математике, Петрозаводск, 2024 г.

10. SMARTY-2024: The Fourth International Workshop on Stochastic Modeling and Applied Research of Technology, Петрозаводск, 2024 г.

Публикации

1. Golovin A., Rumyantsev A. Energy Efficiency of a Single-Server with Inactive State by Matrix-Analytic Method // Information Technologies and Mathematical Modelling. Queueing Theory and Applications / ed. Dudin A., Nazarov A., Moiseev A. Cham: Springer International Publishing, 2022.

2. Golovin A., Rumyantsev A., Astafiev S. Distributed Simulation of Supercomputer Model with Heavy Tails, Lecture Notes in Computer Science / под ред. V. Voevodin [и др.]., Cham: Springer International Publishing, 2022.

3. A. Golovin and A. Rumyantsev, On Multiresource Queues with Multiserver Customers, 2023 5th International Conference on Problems of Cybernetics and Informatics (PCI), Baku, Azerbaijan, 2023

4. Alexander Golovin, Robert Basmadjian, Sergey Astafiev, and Alexander Rumyantsev. 2023. Little's Law in a Single-Server System with Inactive State for Demand-Response in Data Centers with Green SLAs. In Companion Proceedings of the 14th ACM International Conference on Future Energy Systems (e-Energy '23 Companion).

5. Alexander Golovin, 2025. Exact Sampling for Heterogeneous Multiserver Job Model. Reliability: Theory and Applications (в печати)

6. Rumyantsev A., Golovin A., Astafiev S., Basmadjian R. A Three-level Modeling Approach for Asynchronous Speed Scaling in High-performance Data Centres. e-Energy 2021: Proceedings of the 2021 12th ACM International Conference on Future Energy Systems. 2021.

7. Rumyantsev A., Golovin A., Astafiev S., Basmadjian R. Evaluating asynchronous speed scaling policies in high-performance data centres with heavy tails. e-Energy '22: Proceedings of the Thirteenth ACM International Conference on Future Energy Systems, 2022

8. Rumyantsev A., Nekrasova R., Astafiev S., Golovin A. Distributed Regenerative Simulation of a Speed Scaling Supercomputer // 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS), 2022.

9. Rumyantsev A., Golovin A., Astafiev S., Basmadjian R. Three-level modeling of a speed-scaling supercomputer // Ann Oper Res. 2022.

10. Румянцев А. С., Долгалева Д. С., Головин А. С. Исследование стационарных характеристик многосерверных моделей с избыточностью // Программные системы: теория и приложения. 2023.

Содержание Главы 1

1. Представлены элементы теории, касающейся классической односерверной системы $M/G/1$, необходимые для анализа в Главах 2, 3, в том числе: рекурсия Линдли, цикл занятости, выражение для получения стационарной нагрузки через незавершенные времена обслуживания, дисциплина Processor Sharing.

2. Рассмотрена система $M/G/1$ с особым обслуживанием первой заявки. Предложена модификация рекурсии Линдли для такой системы. Из выражения для преобразования Лапласа-Стилтьеса, предложенного Велчем, получено выражение для ф.р. стационарной рабочей нагрузки.

   Welch P.D. On a Generalized M/G/1 Queuing Process in Which the First Customer of Each Busy Period Receives Exceptional Service // Operations Research. 1964. Vol. 12, № 5. P. 736–752.

   Предложен способ получения выборки стационарной нагрузки в такой системе.

3. Представлены элементы теории, касающейся классической многосерверной системы $M/G/c$, необходимые для анализа в Главах 2, 3, в том числе: рекурсия Кифера-Вольфовица, стохастическая монотонность.

4. Дано представление методов точной выборки, применимых к многосерверным системам массового обслуживания. Подробно рассмотрены прямой регенеративный метод точной выборки и метод точной выборки на основе каплинга из прошлого.

Содержание Главы 2

1. Дано описание и свойства модели неоднородной системы с многосерверными заявками (СМЗ). В том числе приведена соответствующая модификация рекурсии Кифера–Вольфовица.

2. Предложена модификация прямого регенеративного метода точной выборки с целью применить к модели СМЗ для получения выборки стационарной задержки. Показано что с помощью предложеного метода можно получить стационарную задержку как для типичной заявки, так и для заявки каждого класса. Вместо получения длины незавершенного цикла регенерации при помощи имитационного моделирования предложен способ её расчета в случае, когда время обслуживания имеет распределение Парето второго типа.

3. Предложена модификация метода точной выборки на основе каплинга из прошлого с целью применить к модели СМЗ. Доказано два математических утверждения: о виде распределения накрывающего интервала в системе $M/G/1$, в которой распределение времени обслуживания является конечной смесью, и о свойстве монотонности вектора рабочей нагрузки.

4. Приведены результаты численных экспериментов, показывающих применимость и адекватность предложенных модификаций.

Содержание Главы 3

1. Дано описание модели неоднородного пула серверов, с помощью которой можно описать ЦОД, состоящий из разнородных пулов серверов, где каждый сервер способен снизить среднюю стационарную потребляемую мощность с помощью перехода в режим низкой потребляемой мощности ценой ухудшения производительности системы.

2. Дано описание модели односерверной системы с режимом остановки и разогрева, представляющей отдельный сервер в пуле серверов. В такой модели сервер может экономить потребляемую энергию, переходя в режим низкой потребляемой мощности, но для перехода в это состояние и выхода из него ему требуется некоторое время, в течение которого заявки не обслуживаются. Таким образом, данную модель можно рассматривать как модель с особым обслуживанием первой заявки. Время ожидания до начала перехода в режим низкой потребляемой мощности является настраиваемым параметром γ, с помощью которого необходимо добиться экономии энергии, не выходя за границы допустимого ухудшения обслуживания.

3. Для случая, когда время обслуживания имеет экспоненциальное распределение, получен явный вид для хвоста распределения времени отклика. Для случая, когда время обслуживания имеет распределение Парето второго типа, получено выражение для среднего времени ожидания и времени отклика, выраженного через γ. Получено явное решение для γ, при котором достигается минимум потребления энергии, зависящее от второго момента времени обслуживания и уровня деградации качества обслуживания.

4. Рассмотрена постановка задачи, в которой есть два пула серверов и требуется найти параметры $\gamma_1, \gamma_2, p_1, n_1, n_2$ такие чтобы, не выходя за заданные границы уровня деградации качества обслуживания, минимизировать потребляемую мощность. Где параметр p_1 – вероятность того, что заявка пришедшая в систему попадет в очередь сервера из первого пула (и с вероятностью $1 - p_1$ в очередь сервера из второго пула). При этом часть серверов может быть выключена физически для большей экономии потребляемой мощности и параметры n_1, n_2 обозначают число серверов, доступных

для управления при помощи параметра γ из общего числа серверов соответствующего пула (т.е. серверы, которые физически включены).

5. Представлены результаты численных экспериментов. В качестве серверов рассмотрены ORACLE X9_1 (Intel Xeon Gold 6354 18c at 3.0GHz, 2 x 32 Gb RAM, NVME) и ORACLE X8_1 (Intel Xeon 8260 at 2.4GHz, 32 Gb RAM, SSD), поскольку для них доступны данные о потребляемой ими мощности в различных режимах.

Список литературы

[1] Joseph Abate, Gagan L. Choudhury и Ward Whitt. “Calculating the M/G/1 busy-period density and LIFO waiting-time distribution by direct numerical transform inversion”. en. В: *Operations Research Letters* 18.3 (окт. 1995), с. 113—119. ISSN: 01676377. DOI: 10.1016/0167-6377(95)00049-6.

[2] M. Abramowitz и I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Applied mathematics series. Washington D.C.: U.S. Government Printing Office, 1972.

[3] Dlzar Al Kez и др. “Data Center Potential Flexibilities and Challenges for Demand Response to Facilitate 100% Inverter-Based Resources: A Review”. en. В: *SSRN Electronic Journal* (2022). ISSN: 1556-5068. DOI: 10.2139/ssrn.4269631.

[4] Dlzar Al Kez и др. “Manipulation of static and dynamic data center power responses to support grid operations”. В: *IEEE Access* 8 (2020), с. 182078—182091.

[5] Dlzar Al Kez и др. “Potential of data centers for fast frequency response services in synchronously isolated power systems”. В: *Renewable and Sustainable Energy Reviews* 151 (2021), с. 111547.

[6] OpenADR Alliance. *OpenADR 2.0 Profile Specification B Profile*. Тех. отч. 20120912-1. OpenADR Alliance, 2013.

[7] Abdulhamid Alzaid, Jee Soo Kim и Frank Proschan. “Laplace ordering and its applications”. en. В: *Journal of Applied Probability* 28.1 (март 1991), с. 116—130. ISSN: 0021-9002, 1475-6072. DOI: 10.2307/3214745.

[8] Ed Arthurs и Joseph S. Kaufman. “Sizing a Message Store Subject to Blocking Criteria”. В: *Performance of Computer Systems, Proceedings of the Fifth International Symposium on Modelling and Performance Evaluation of Computer Systems, Vienna, Austria, February 6-8, 1979*. Под ред. M. Arató, A. Butrimenko и Erol Gelenbe. North-Holland, 1979, с. 547—564.

[9] Søren Asmussen. *Applied probability and queues*. 2. ed. Applications of mathematics 51. New York Berlin Heidelberg: Springer, 2003. 438 с. ISBN: 978-0-387-00211-8 978-1-4419-1809-3.

[10] *AT11487: Low Power Consumption Techniques for XMEGA XPLAINED Kits [Application note]*. Тех. отч. Atmel Corporation, 2015. URL: https://ww1.microchip.com/downloads/en/AppNotes/Atmel-42456-Low-Power-Consumption-Techniques-for-XMEGA-XPLAINED-Kits%5C_Application-Note%5C_AT11487.pdf.

[11] Hazril Izan Bahari и Siti Salbiah Mohamed Shariff. "Review on data center issues and challenges: Towards the Green Data Center". В: *2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*. 2016, с. 129—134. DOI: 10.1109/ICCSCE.2016.7893558.

[12] Shahab Bahrami, Vincent WS Wong и Jianwei Huang. "Data center demand response in deregulated electricity markets". В: *IEEE Transactions on Smart Grid* 10.3 (2018), с. 2820—2832.

[13] Robert Basmadjian. "Flexibility-Based Energy and Demand Management in Data Centers: A Case Study for Cloud Computing". В: *Energies* 12.17 (2019). ISSN: 1996-1073. DOI: 10.3390/en12173301.

[14] Robert Basmadjian и др. "Green Data Centers". В: *Large-Scale Distributed Systems and Energy Efficiency*. Hoboken, New Jersey: John Wiley & Sons, Ltd, 2015. Гл. 6, с. 159—196. ISBN: 9781118981122. DOI: 10.1002/9781118981122.ch6.

[15] Robert Basmadjian и др. "Making Data Centers Fit for Demand Response: Introducing GreenSDA and GreenSLA Contracts". В: *IEEE Transactions on Smart Grid* 9.4 (июль 2018), с. 3453—3464. ISSN: 1949-3053, 1949-3061. DOI: 10.1109/TSG.2016.2632526.

[16] Robert Basmadjian и др. "Making Data Centers Fit for Demand Response: Introducing GreenSDA and GreenSLA Contracts". В: *IEEE Transactions on Smart Grid* 9.4 (июль 2018), с. 3453—3464. ISSN: 1949-3053, 1949-3061. DOI: 10.1109/TSG.2016.2632526. URL: https://ieeexplore.ieee.org/document/7755751/.

[17] Julian Bellendorf и Zoltán Ádám Mann. "Classification of optimization problems in fog computing". en. В: *Future Generation Computer Systems* 107 (июнь 2020), с. 158—176. ISSN: 0167739X. DOI: 10.1016/j.future.2020.01.036. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167739X19323568.

[18] P.P. Bocharov, C. D'Apice и A.V. Pechinkin. *Queueing Theory*. Modern Probability and Statistics. De Gruyter, 2011. ISBN: 9783110936025. URL: https://books.google.ru/books?id=2BREaX4EunkC.

[19] Mark Van der Boor и др. "Scalable Load Balancing in Networked Systems: A Survey of Recent Advances". В: *SIAM Review* 64.3 (2022), с. 554—622. DOI: 10.1137/20M1323746.

[20] P. Brill и L. Green. "Queues in which customers receive simultaneous service from a random number of servers: A system point approach". en. В: *Management Science* 30.1 (1984), с. 51—68.

[21] Min Chen и др. “Internet data centers participating in demand response: A comprehensive review”. В: *Renewable and Sustainable Energy Reviews* 117 (2020), с. 109466.

[22] Marta Chinnici, Davide De Chiara и Andrea Quintiliani. “An HPC-Data Center Case Study on the Power Consumption of Workload”. В: *Applied Physics, System Science and Computers II*. Под ред. Klimis Ntalianis и Anca Croitoru. Cham: Springer International Publishing, 2019, с. 183—192. ISBN: 978-3-319-75605-9.

[23] Jacob Willem Cohen. *The single server queue*. Rev. ed. North-Holland series in applied mathematics and mechanics v. 8. Amsterdam ; New York : New York: North-Holland Pub. Co. ; Sole distributors for the U.S.A. и Canada, Elsevier North-Holland, 1982. 694 с. ISBN: 978-0-444-85452-0.

[24] D.J. Daley. “Certain optimality properties of the first-come first-served discipline for G/G/s queues”. В: *Stochastic Processes and their Applications* 25 (1987), с. 301—308. ISSN: 03044149. DOI: 10.1016/0304-4149(87)90208-0. URL: https://linkinghub.elsevier.com/retrieve/pii/0304414987902080.

[25] Xeni K Dimakos. “A Guide to Exact Simulation”. В: *International Statistical Review / Revue Internationale de Statistique* 69.1 (2001), с. 27—48.

[26] Richard V. Evans. “Queuing When Jobs Require Several Services Which Need Not Be Sequenced”. В: *Management Science* 10.2 (1964), с. 298—315. URL: http://www.jstor.org/stable/2627300.

[27] Enver Ever и др. “On the performance, availability and energy consumption modelling of clustered IoT systems”. English. В: *Computing* 101.12 (дек. 2019). Place: Wien Publisher: Springer Wien WOS:000491601200008, с. 1935—1970. ISSN: 0010-485X. DOI: 10.1007/s00607-019-00720-9.

[28] Dror G. Feitelson. *Workload Modeling for Computer Systems Performance Evaluation*. 1-е изд. Cambridge University Press, 23 марта 2015. ISBN: 978-1-107-07823-9 978-1-139-93969-0. DOI: 10.1017/CBO9781139939690. URL: https://www.cambridge.org/core/product/identifier/9781139939690/type/book.

[29] FERC. *Assessment of Demand Response and Advanced Metering*. Тех. отч. Federal Energy Regulatory Commission, 2013. URL: https://www.ferc.gov/sites/default/files/2020-05/oct-demand-response.pdf.

[30] D. Filippopoulos и H. Karatza. “An M / M / 2 parallel system model with pure space sharing among rigid jobs”. В: *Mathematical and Computer Modelling* 45.5 (март 2007), с. 491—530. ISSN: 08957177. DOI: 10.1016/j.mcm.2006.06.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895717706002627.

[31] Jean-Michel Fourneau. “Modeling Green Data-Centers and Jobs Balancing with Energy Packet Networks and Interrupted Poisson Energy Arrivals”. en. В: *SN Computer Science* 1.1 (янв. 2020), с. 28. ISSN: 2662-995X, 2661-8907. DOI: 10.1007/s42979-019-0029-5. URL: http://link.springer.com/10.1007/s42979-019-0029-5.

[32] Eugene Furman и др. *Capacity Allocation for Clouds with Parallel Processing, Batch Arrivals, and Heterogeneous Service Requirements*. Version Number: 2. 2022. DOI: 10.48550/ARXIV.2209.08820. URL: https://arxiv.org/abs/2209.08820.

[33] G. Latouche и V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Англ. Philadelphia: ASA–SIAM, 1999.

[34] Anshul Gandhi и др. "Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward". en. В: *Queueing Systems* 77.2 (июнь 2014), с. 177—209. ISSN: 0257-0130, 1572-9443. DOI: 10.1007/s11134-014-9409-7. URL: http://link.springer.com/10.1007/s11134-014-9409-7.

[35] Anshul Gandhi и др. "Optimality analysis of energy-performance trade-off for server farm management". В: *Performance Evaluation* 67.11 (2010), с. 1155—1171. DOI: 10.1016/j.peva.2010.08.009. URL: http://www.sciencedirect.com/science/article/pii/S0166531610001069.

[36] Lakshmi Ganesh и др. "Integrated Approach to Data Center Power Management". В: *IEEE Transactions on Computers* 62.6 (2013), с. 1086—1096. DOI: 10.1109/TC.2013.32.

[37] Erol Gelenbe и Isi Mitrani. *Analysis and Synthesis of Computer Systems*. 2nd. IMPERIAL COLLEGE PRESS, 2010. DOI: 10.1142/p643. eprint: https://www.worldscientific.com/doi/pdf/10.1142/p643. URL: https://www.worldscientific.com/doi/abs/10.1142/p643.

[38] Clark W. Gellings и John H. Chamberlin. *Demand-side management: concepts and methods*. eng. Lilburn, Ga: Fairmont Press, 1988. ISBN: 978-0-13-198458-5 978-0-88173-027-2.

[39] Girish Ghatikar и др. *Demand Response and Open Automated Demand Response Opportunities for Data Centers*. Тех. отч. Lawrence Berkeley National Laboratory, 2010.

[40] Girish Ghatikar и др. *Demand Response Opportunities and Enabling Technologies for Data Centers: Findings from Field Studies*. Тех. отч. Lawrence Berkeley National Laboratory, 2012.

[41] Fablenne Gillent и Guy Latouche. "Semi-explicit solutions for M/PH/1-like queuing systems". В: *European Journal of Operational Research* 13.2 (июнь 1983), с. 151—160. ISSN: 0377-2217. DOI: 10.1016/0377-2217(83)90077-2. URL: http://www.sciencedirect.com/science/article/pii/0377221783900772.

[42] L. Gimpelson. "Analysis of Mixtures of Wide- and Narrow-Band Traffic". В: *IEEE Transactions on Communication Technology* 13.3 (1965), с. 258—266. DOI: 10.1109/TCOM.1965.1089121.

[43] Mor Harchol-Balter. "Open problems in queueing theory inspired by datacenter computing". В: *Queueing Systems* 97.1 (февр. 2021), с. 3—37. ISSN: 0257-0130, 1572-9443. DOI: 10.1007/s11134-020-09684-6. URL: http://link.springer.com/10.1007/s11134-020-09684-6.

[44] Mor Harchol-Balter. *Performance modeling and design of computer systems: queueing theory in action*. Cambridge: Cambridge University Press, 2013. 548 с. ISBN: 978-1-107-02750-3.

[45] Mor Harchol-Balter. "The multiserver job queueing model". В: *Queueing Systems* 100.3 (апр. 2022), с. 201—203. ISSN: 0257-0130, 1572-9443. DOI: 10.1007/s11134-022-09762-x. URL: https://link.springer.com/10.1007/s11134-022-09762-x.

[46] Qi-Ming He. *Fundamentals of Matrix-Analytic Methods*. Springer New York, 2014. ISBN: 978-1-4614-7329-9.

[47] *HP ProBook 450 G8 Notebook PC, c06907888 - DA16756 - Worldwide - Version 18*. en. 2021. URL: https://h20195.www2.hp.com/v2/GetPDF.aspx/c06907888.pdf.

[48] Mark L. Huber. *Perfect Simulation*. 0-е изд. Chapman и Hall/CRC, 20 янв. 2016. ISBN: 978-0-429-16526-9. DOI: 10.1201/b19235. URL: https://www.taylorfrancis.com/books/9781482232455.

[49] J Keilson и L.D Servi. "A distributional form of Little's Law". В: *Operations Research Letters* 7.5 (1988), с. 223—227. ISSN: 0167-6377. DOI: https://doi.org/10.1016/0167-6377(88)90035-1. URL: https://www.sciencedirect.com/science/article/pii/0167637788900351.

[50] Wilfrid Kendall. "NOTES ON PERFECT SIMULATION". В: W S Kendall, F Liang и J-S Wang. *Markov Chain Monte Carlo*. Т. 7. Series Title: Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore. CO-PUBLISHED WITH SINGAPORE UNIVERSITY PRESS, нояб. 2005, с. 93—146. ISBN: 978-981-256-427-6 978-981-270-091-9. DOI: 10.1142/9789812700919_0003.

[51] J. Kiefer и J. Wolfowitz. "On the theory of queues with many servers". В: *Transactions of the American Mathematical Society* 78.1 (1955), с. 1—18. ISSN: 0002-9947, 1088-6850. DOI: 10.1090/S0002-9947-1955-0066587-3. URL: https://www.ams.org/tran/1955-078-01/S0002-9947-1955-0066587-3/.

[52] Paul J. Kuehn и Maggie Mashaly. "DVFS-Power Management and Performance Engineering of Data Center Server Clusters". В: *2019 15th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*. Wengen, Switzerland: IEEE, янв. 2019, с. 91—98. ISBN: 978-3-903176-13-3. DOI: 10.23919/WONS.2019.8795470. URL: https://ieeexplore.ieee.org/document/8795470/.

[53] Paul J. Kuehn и Maggie Ezzat Mashaly. "Automatic energy efficiency management of data center resources by load-dependent server activation and sleep modes". en. В: *Ad Hoc Networks* 25 (февр. 2015), с. 497—504. ISSN: 15708705. DOI: 10.1016/j.adhoc.2014.11.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S1570870514002650.

[54] Danielle Liu и Y. Quennel Zhao. “Determination of Explicit Solutions for a General Class of Markov Processes”. en. В: *Matrix-Analytic Methods in Stochastic Models*. Под ред. S. Chakravarthy и Attahiru S. Alfa. 1-е изд. CRC Press, 1996, с. 363—378. ISBN: 978-0-429-17626-5. DOI: 10.1201/b17050-21. URL: https://www.taylorfrancis.com/books/9781482292176/chapters/10.1201/b17050-21.

[55] Jorge Marx Gómez и др., ред. *Engineering and Management of Data Centers: An IT Service Management Approach*. Service Science: Research and Innovations in the Service Economy. Cham: Springer International Publishing, 2017. ISBN: 978-3-319-65081-4 978-3-319-65082-1. DOI: 10.1007/978-3-319-65082-1. URL: http://link.springer.com/10.1007/978-3-319-65082-1.

[56] Evsey Morozov, Alexander Rumyantsev и Irina Peshkova. “Monotonicity and stochastic bounds for simultaneous service multiserver systems”. В: *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2016 8th International Congress on*. New York: IEEE, 2016, с. 294—297. ISBN: 978-1-4673-8818-4. DOI: 10.1109/ICUMT.2016.7765374.

[57] Evsey Morozov и Bart Steyaert. *Stability Analysis of Regenerative Queueing Models Mathematical Methods and Applications*. английский. Springer, Cham, 2021. ISBN: 978-3-030-82438-9. URL: https://link.springer.com/book/10.1007/978-3-030-82438-9.

[58] Evsey Morozov и др. “Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking”. en. В: *Performance Evaluation* 134 (окт. 2019), с. 102005. ISSN: 01665316. DOI: 10.1016/j.peva.2019.102005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0166531618303067.

[59] Marcel F. Neuts. “Matrix-geometric solutions in stochastic models an algorithmic approach”. В: *Journal of the American Statistical Association* 77.379 (1982), с. 690—690.

[60] Diletta Olliaro и др. “The Impact of Service Demand Variability on Data Center Performance”. В: *IEEE Transactions on Parallel and Distributed Systems* 36.2 (2025), с. 120—132. DOI: 10.1109/TPDS.2024.3497792.

[61] Ryuji Oma и др. “An energy-efficient model for fog computing in the Internet of Things (IoT)”. en. В: *Internet of Things* 1-2 (сент. 2018), с. 14—26. ISSN: 25426605. DOI: 10/ggw4k5. URL: https://linkinghub.elsevier.com/retrieve/pii/S2542660518300386.

[62] *Oracle’s Power Calculators*. Oracle. 2020. URL: https://www.oracle.com/it-infrastructure/power-calculators/.

[63] Suraj Singh Panwar, M. M. S. Rauthan и Varun Barthwal. “A systematic review on effective energy utilization management strategies in cloud data centers”. В: *Journal of Cloud Computing* 11.1 (17 дек. 2022), с. 95. ISSN: 2192-113X. DOI: 10.1186/s13677-022-00368-5. URL: https://journalofcloudcomputing.springeropen.com/articles/10.1186/s13677-022-00368-5.

[64] John W. Pearson, Sheehan Olver и Mason A. Porter. “Numerical methods for the computation of the confluent and Gauss hypergeometric functions”. en. В: *Numerical Algorithms* 74.3 (март 2017), с. 821—866. ISSN: 1017-1398, 1572-9265. DOI: 10.1007/s11075-016-0173-0.

[65] N. U. Prabhu. “Comments on two papers on queueing theory by J. Kiefer and J. Wolfowitz”. В: *Queueing Systems* 1.3 (март 1987), с. 311—315. ISSN: 0257-0130, 1572-9443. DOI: 10.1007/BF01149541. URL: http://link.springer.com/10.1007/BF01149541.

[66] James Gary Propp и David Bruce Wilson. “Exact sampling with coupled Markov chains and applications to statistical mechanics”. В: *Proceedings of the Seventh International Conference on Random Structures and Algorithms*. Atlanta, Georgia, USA: John Wiley & Sons, Inc., 1996, с. 223—252. DOI: 10.5555/243401.235641.

[67] Colin M. Ramsay. “Exact waiting time and queue size distributions for equilibrium M/G/1 queues with Pareto service”. В: *Queueing Systems* 57.4 (дек. 2007), с. 147—155. ISSN: 1572-9443. DOI: 10.1007/s11134-007-9052-7.

[68] Colin M. Ramsay. “The Distribution of Sums of I.I.D. Pareto Random Variables with Arbitrary Shape Parameter”. В: *Communications in Statistics - Theory and Methods* 37.14 (27 мая 2008), с. 2177—2184. ISSN: 0361-0926, 1532-415X.

[69] Sheldon M. Ross. *Stochastic processes*. 2. ed., [Nachdr.] Wiley series in probability and statistics. New York: Wiley, 2007. 510 с. ISBN: 978-0-471-12062-9.

[70] Alexander Rumyantsev и Evsey Morozov. “Stability criterion of a multiserver model with simultaneous service”. В: *Annals of Operations Research* 252.1 (май 2017), с. 29—39. ISSN: 0254-5330, 1572-9338. DOI: 10.1007/s10479-015-1917-2. URL: http://link.springer.com/10.1007/s10479-015-1917-2.

[71] Alexander Rumyantsev и др. “Three-level modeling of a speed-scaling supercomputer”. В: *Annals of Operations Research* (21 июня 2022). ISSN: 0254-5330, 1572-9338. DOI: 10.1007/s10479-022-04830-0. URL: https://link.springer.com/10.1007/s10479-022-04830-0.

[72] Karl Sigman. “Appendix: A primer on heavy-tailed distributions”. В: *Queueing Systems* 33.1 (1999), с. 261—275. URL: http://www.springerlink.com/index/VJ233Q22M2WK6694.pdf.

[73] Karl Sigman. “Exact simulation of the stationary distribution of the FIFO M/G/c queue”. В: *Journal of Applied Probability* 48.A (2011), с. 209—213. DOI: 10.1239/jap/1318940466.

[74] Karl Sigman. “Exact simulation of the stationary distribution of the FIFO M/G/c queue: The general case for <c”. В: *Queueing Syst.* 70 (нояб. 2012), с. 37—43. DOI: 10.1007/s11134-011-9266-6.

[75] Karl Sigman. “Stationary Marked Point Processes”. В: *Springer Handbook of Engineering Statistics*. Под ред. Hoang Pham. London: Springer London, 2006, с. 137—152. ISBN: 978-1-84628-288-1. DOI: 10.1007/978-1-84628-288-1_8. URL: https://doi.org/10.1007/978-1-84628-288-1_8.

[76] Cheng-Jen Tang и др. “Energy management for the homogeneous server clusters offering web services”. en. В: *Energy Efficiency* 9.5 (окт. 2016), с. 1115—1144. ISSN: 1570-646X, 1570-6478. DOI: 10.1007/s12053-015-9412-9. URL: http://link.springer.com/10.1007/s12053-015-9412-9.

[77] Peter D. Welch. “On a Generalized M/G/1 Queuing Process in Which the First Customer of Each Busy Period Receives Exceptional Service”. В: *Operations Research* 12.5 (1964), с. 736—752. URL: http://www.jstor.org/stable/167778.

[78] Ronald W. Wolff. “An upper bound for multi-channel queues”. В: *Journal of Applied Probability* 14.4 (дек. 1977), с. 884—888. ISSN: 0021-9002, 1475-6072. DOI: 10.2307/3213363.

[79] Ronald W. Wolff. “Upper bounds on work in system for multichannel queues”. В: *Journal of Applied Probability* 24.2 (июнь 1987), с. 547—551. ISSN: 0021-9002, 1475-6072. DOI: 10.2307/3214279. URL: https://www.cambridge.org/core/product/identifier/S0021900200031193/type/journal_article.

[80] Eric Wolman. “The camp-on problem for multiple-address traffic”. В: *The Bell System Technical Journal* 51.6 (1972), с. 1363—1422. DOI: 10.1002/j.1538-7305.1972.tb02657.x.

[81] Yaofei Xiong, Duncan J. Murdoch и David A. Stanford. “Perfect and nearly perfect sampling of work-conserving queues”. В: *Queueing Systems* 80.3 (июль 2015), с. 197—222. ISSN: 0257-0130, 1572-9443. DOI: 10.1007/s11134-015-9437-y. URL: http://link.springer.com/10.1007/s11134-015-9437-y.

[82] Xiao Zhang и др. “Key Technologies for Green Data Center”. В: *2010 Third International Symposium on Information Processing*. 2010, с. 477—480. DOI: 10.1109/ISIP.2010.107.

[83] В.С. Королюк и др. *Справочник по теории вероятностей и математической статистике*. М.: Наука, 1985.