МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра системного анализа и автоматического управления

ИССЛЕДОВАНИЕ МОДЕЛИ ГЕТЕРОГЕННОГО ДАТА-ЦЕНТРА

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета компьютерных наук и информационных технологий
Логинова Артема Юрьевича

Научный руководитель
доцент, к. ф.-м. н. _____ Е. С. Рогачко

Заведующий кафедрой
к. ф.-м. н., доцент _____ И. Е. Тананко

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. Успешное функционирование современных дата-центров является важным аспектом обеспечения стабильной работы облачных сервисов, систем хранения и обработки данных, а также приложений искусственного интеллекта. В последние годы наблюдается стремительный рост объёмов обрабатываемой информации и вычислительных нагрузок, что приводит к необходимости повышения эффективности центров обработки данных. Одним из основных направлений развития современных дата-центров является использование гетерогенной архитектуры, включающей вычислительные узлы с различными характеристиками производительности и специализацией. Такой подход позволяет более эффективно распределять вычислительные ресурсы, однако существенно усложняет процессы анализа, балансировки нагрузки и оптимизации функционирования системы.

Цель бакалаврской работы — разработка и исследование математической модели гетерогенного дата-центра для анализа и оптимизации его ключевых показателей.

Поставленная цель определила **следующие задачи**:

- проанализировать архитектуру и применение гетерогенных дата-центров;
- представить гетерогенный дата-центр в виде модели-сети массового обслуживания;
- описать метод расчета характеристик модели;
- описать метод оптимизации модели;
- разработать программу для расчета характеристик и оптимизации параметров модели;
- провести вычислительные исследования модели и проанализировать полученные результаты.

Методологические основы бакалаврской работы представлены в исследованиях, посвященных анализу и моделированию центров обработки данных и облачных вычислительных систем [1, 2]. Значительная часть работ рассматривает однородные системы обслуживания, в которых все вычислительные узлы обладают одинаковыми характеристиками [3]. Однако для гетерогенных дата-центров, учитывающих различия в интенсивностях обслуживания, количестве серверов и распределении нагрузки, проведено существенно меньше исследований [4]. В связи с этим возникает необходимость разработ-

ки математических моделей, позволяющих исследовать влияние неоднородности вычислительных ресурсов на производительность системы и проводить оптимизацию её параметров [5]. Для построения модели гетерогенного дата-центра в данной работе используются сети массового обслуживания, позволяющие формализовать процессы поступления, маршрутизации и обработки запросов. Применение аппарата теории массового обслуживания даёт возможность определить основные характеристики функционирования системы, включая среднее время отклика, вероятность отказа, загрузку узлов и длины очередей [5].

Практическая значимость бакалаврской работы. Разработанная программа может использоваться для предварительного анализа конфигураций гетерогенных дата-центров, выявления перегруженных узлов, оценки влияния изменения параметров системы на среднее время отклика и выбора более эффективных вариантов маршрутизации и распределения серверов.

Структура и объём работы. Бакалаврская работа состоит из введения, четырех разделов, заключения, списка использованных источников и двух приложений. Общий объем работы – 52 страницы, из них 43 страницы – основное содержание, включая 11 рисунков и 6 таблиц, цифровой носитель в качестве приложения, список использованных источников информации – 22 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Дата-центры и их применение» посвящен рассмотрению основных понятий и компонентов облачного дата-центра, структуры гетерогенного дата-центра и показателей качества его работы. Подраздел 1.1 содержит определение дата-центра, описание его назначения, областей применения и основных понятий: пользователь, планировщик, сервер и отклик. Дата-центр рассматривается как специализированный объект, объединяющий IT-инфраструктуру и инженерную инфраструктуру для обработки, хранения и передачи данных. Упрощённая архитектура взаимодействия пользователей с облачным дата-центром приведена на рисунке 1. Дата-центры применяются государственными организациями, финансовыми учреждениями, промышленными предприятиями, IT-компаниями и пользователями облачных сервисов.

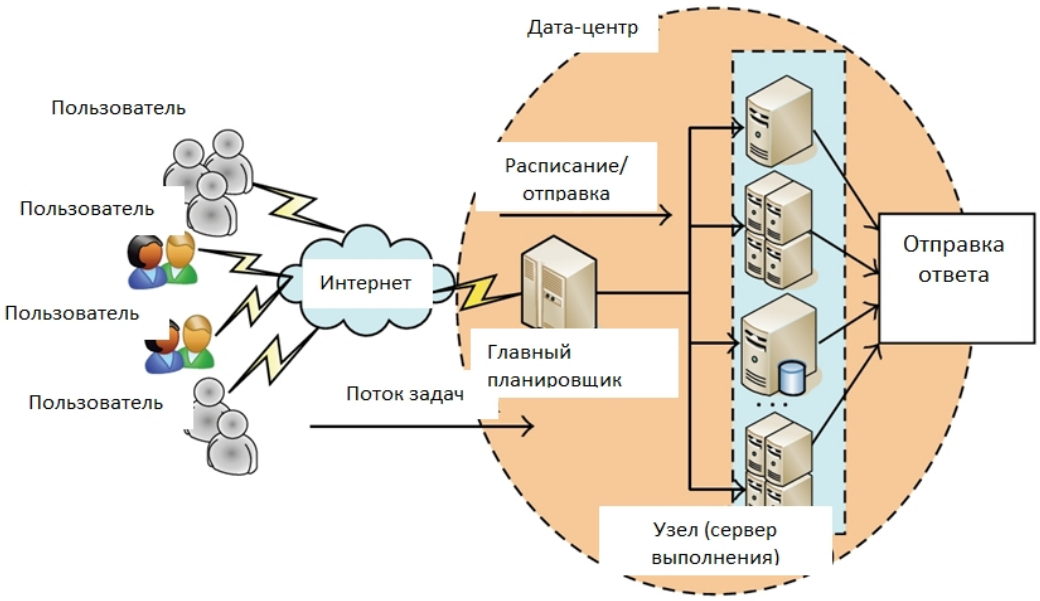

Рисунок 1 – Упрощённая архитектура взаимодействия пользователей с облачным дата-центром

В подразделе 1.2 описываются основные компоненты гетерогенного дата-центра: основной планировщик, обслуживаемые очереди, серверы, контроллер балансировки нагрузки и подсистема мониторинга. Гетерогенность означает, что различные серверы или группы серверов обладают разной производительностью, конфигурацией и интенсивностью обслуживания запросов. В модели используются следующие обозначения: λ — интенсивность входного потока требований, K — ёмкость планировщика, μ_s — интенсивность обслуживания планировщика, n — число обслуживающих систем (узлов), p_i — вероятность направления требования в i-ю систему, S_i — i-я обслуживающая система, c_i — число серверов в i-й группе, c_{ij} — j-й сервер в i-й системе, μ_i — интенсивность обслуживания серверов i-й системе, L_q — средняя длина очереди, W_q — среднее время ожидания требований в очереди, ρ — коэффициент загрузки серверов. В качестве математической модели дата-центра используется сеть массового обслуживания. Модель показана на рисунке 2.

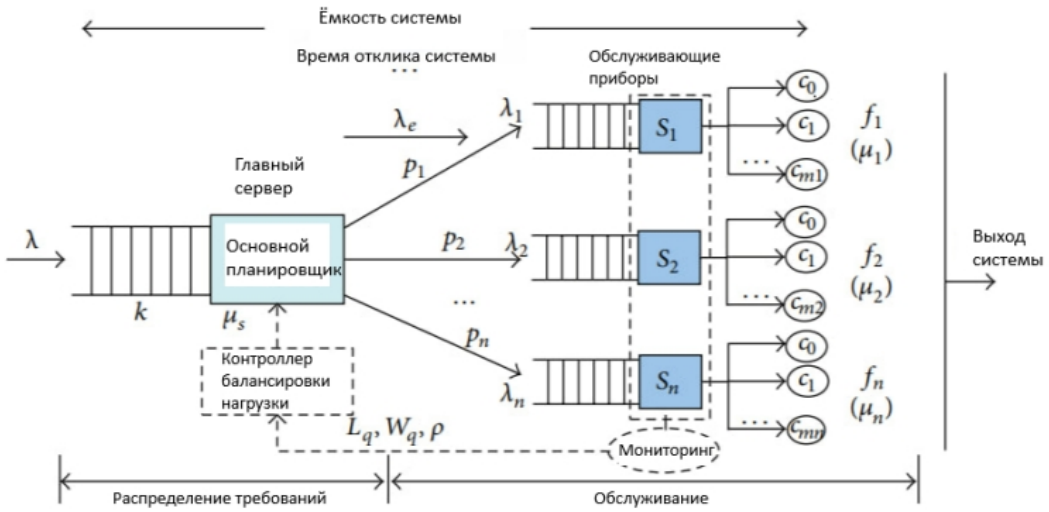

Рисунок 2 – Модель сети массового обслуживания для гетерогенного дата-центра

В подразделе 1.3 перечисляются основные характеристики качества работы дата-центра: среднее время отклика, средняя длина очереди, среднее время ожидания, пропускная способность, коэффициент использования ресурсов, вероятность блокировки и вероятность немедленного обслуживания.

Среднее время отклика рассматривается как основной показатель качества, поскольку именно оно отражает полную задержку от поступления пользовательского запроса до завершения его обработки. Средняя длина очереди и среднее время ожидания позволяют определить, в каких узлах накапливается нагрузка и где возникают задержки. Коэффициент использования ресурсов показывает степень загрузки серверов: слишком малые значения указывают на избыточность ресурсов, а значения, близкие к единице, свидетельствуют о риске перехода системы в критический режим. Вероятность блокировки используется для оценки потерь на входе в систему, если очередь планировщика имеет ограниченную вместимость. Совместное рассмотрение этих характеристик позволяет не только оценить текущее состояние дата-центра, но и определить направление его модернизации.

Второй раздел «Математическая модель гетерогенного дата-центра» посвящен построению математической модели и описанию методов вычисления её характеристик. В подразделе 2.1 гетерогенный дата-центр описывается как открытая сеть массового обслуживания: планировщик моделируется системой типа $M/M/1/K$, а обслуживающие узлы — системами массового обслуживания типа $M/M/c_i$. После обработки в планировщике запрос направляется в один из узлов согласно вероятностям маршрутизации p_i.

Для корректной маршрутизации должны выполняться условия

$$\sum_{i=1}^{n} p_i = 1, \qquad p_i \geq 0.$$

Такая постановка задачи позволяет учитывать как неоднородность вычислительных ресурсов, так и влияние стратегии распределения входного потока запросов.

В модели предполагается, что входной поток запросов является пуассоновским, а времена обслуживания имеют экспоненциальное распределение. Эти предположения позволяют использовать аппарат марковских систем массового обслуживания и получить расчетные характеристики в аналитическом виде. Планировщик выделяется в отдельный узел, так как в реальном дата-центре он выполняет самостоятельную функцию: принимает входящий запрос, учитывает выбранную стратегию маршрутизации и передает запрос в один из обслуживающих узлов. Ограниченная вместимость очереди планировщика отражает ситуацию, при которой входной буфер системы не может принимать неограниченное число запросов.

Каждый обслуживающий узел S_i моделируется как СМО типа $M/M/c_i$, то есть как система с c_i идентичными серверами, неограниченной очередью и экспоненциальным временем обслуживания с интенсивностью μ_i для каждого сервера. Таким образом, гетерогенность дата-центра математически выражается через различные значения параметров c_i и μ_i для разных узлов.

В подразделе 2.2 описывается метод анализа модели. Сначала по параметрам входного потока и планировщика определяется вероятность блокировки P_b. Для планировщика типа $M/M/1/K$ она совпадает с вероятностью нахождения системы в предельном состоянии: $P_b = P_K$. Эффективная интенсивность потока, прошедшего через планировщик, интенсивность поступления в i-й узел и коэффициент загрузки этого узла вычисляются по формулам

$$\lambda_e = \lambda(1 - P_b), \qquad \lambda_i = \lambda_e p_i, \qquad \rho_i = \frac{\lambda_i}{c_i \mu_i}, \qquad i = 1, \ldots, n,$$

Для существования стационарного режима необходимо выполнение условия $\rho_i < 1$ для всех обслуживающих узлов. Если хотя бы один узел перегружен, то устойчивый режим работы всей системы отсутствует, поскольку очередь

в этом узле неограниченно растет.

После проверки стационарности рассчитываются средняя длина очереди, среднее время ожидания и среднее время пребывания запроса в каждом узле. Для обслуживающего узла типа $M/M/c_i$ используются зависимости

$$L_i = L_{qi} + \frac{\lambda_i}{\mu_i}, \qquad \overline{T}_{qi}^{(e)} = \frac{L_{qi}}{\lambda_i}, \qquad \overline{T}_i^{(e)} = \frac{L_i}{\lambda_i},$$

где L_{qi} — средняя длина очереди в i-м узле, L_i — среднее число запросов в узле, $\overline{T}_{qi}^{(e)}$ — среднее время ожидания в очереди, $\overline{T}_i^{(e)}$ — среднее время пребывания запроса в узле. Среднее время отклика всей системы определяется как сумма среднего времени пребывания запроса в планировщике и среднего времени пребывания запроса в узле с учетом вероятностей маршрутизации:

$$\overline{T} = \overline{T}^{(s)} + \sum_{i=1}^{n} p_i \overline{T}_i^{(e)},$$

где $\overline{T}^{(s)}$ — среднее время пребывания запроса в планировщике, $p_i \overline{T}_i^{(e)}$ — вклад i-го обслуживающего узла с учетом вероятности направления запроса в этот узел. Такая последовательность расчета позволяет определить не только итоговый показатель качества, но и компонент системы, который вносит основной вклад в среднее время отклика системы.

В подразделе 2.3 рассматривается задача оптимизации модели. Управляемыми параметрами являются вектор вероятностей маршрутизации $p = (p_1, \ldots, p_n)$ и распределение серверов $c = (c_1, \ldots, c_n)$. Цель оптимизации состоит в минимизации среднего времени отклика:

$$\overline{T}(c, p) \to \min.$$

При этом должны выполняться ограничения на корректность маршрутизации, сохранение общего числа серверов и существование стационарного режима:

$$\sum_{i=1}^{n} p_i = 1, \quad 0 \le p_i < 1, \quad i = 1, \ldots, n, \quad \sum_{i=1}^{n} c_i \le C_{\max}, \quad \rho_i < 1.$$

Для фиксированной конфигурации серверов применяется метод проек-

тивного градиентного спуска [6]: вероятности маршрутизации изменяются в направлении уменьшения среднего времени отклика, после чего полученный вектор возвращается в допустимое множество вероятностных распределений. Для структурных параметров c используется перебор допустимых распределений серверов, а для каждого такого распределения решается задача настройки маршрутизации.

Третий раздел «Описание программной реализации модели гетерогенного дата-центра» посвящен реализации алгоритмов анализа и оптимизации модели гетерогенного дата-центра.

В подразделе 3.1 описывается алгоритм анализа модели, он состоит из пяти блоков. В первом блоке вводятся параметры системы. Во втором блоке проверяется корректность входных данных. В третьем блоке вычисляются характеристики планировщика. В четвертом блоке рассчитываются показатели обслуживающих узлов и проверяется стационарность. В пятом блоке формируется итоговый набор общесистемных характеристик.

Также в подразделе 3.1 представлен алгоритм оптимизации модели, состоящий из четырёх блоков. Сначала выбирается режим оптимизации: настройка вероятностей маршрутизации, распределение серверов или совместная оптимизация. Затем для фиксированного распределения серверов выполняется проективный градиентный спуск по вероятностям маршрутизации. После этого при необходимости перебираются допустимые распределения серверов и сохраняется лучший найденный вариант. На заключительном этапе выводятся оптимальные значения управляемых параметров и соответствующие характеристики системы.

В подразделе 3.2 описывается разработанная программа. Программа написана на языке Python и включает два основных модуля. Первый модуль `app_tk.py` отвечает за графический интерфейс, ввод параметров, запуск расчётов, отображение результатов, формирование графиков и сохранение отчетов. Модуль `datacenter_calculator.py` содержит вычислительное ядро: проверку стационарности, расчет характеристик планировщика и обслуживающих узлов, формирование данных для графиков, оптимизацию вероятностей маршрутизации, оптимизацию распределения серверов и совместную оптимизацию. При разработке программы использованы стандартные библиотеки `tkinter`, `tkinter.ttk`, `json`, `math`, `threading`, `pathlib`, а

также внешние библиотеки `NumPy` для численных расчетов и `Pillow` для сохранения графиков в растровом формате.

В подразделе 3.3 описывается интерфейс программы на примере её использования. Интерфейс программы состоит из вкладок ввода данных, просмотра параметров и характеристик, а также формирования графиков. Пользователь может загрузить конфигурацию модели из JSON-файла, рассчитать характеристики, запустить оптимизацию, сохранить текстовый отчет и экспортировать график.

Четвертый раздел «Результаты исследования модели гетерогенного дата-центра» посвящен исследованию модели гетерогенного дата-центра на основе численных примеров. Базовый пример описывает небольшой облачный дата-центр веб-приложения с тремя обслуживающими узлами: современными, стандартными и старыми серверами. Такая конфигурация отражает типичную ситуацию постепенной модернизации инфраструктуры, когда в одной системе одновременно используются вычислительные узлы разных поколений. В вычислительных экспериментах на основе базового примера изменяются отдельные параметры, что позволяет оценить чувствительность модели и результат оптимизации.

В примере 1 показано, что даже при достаточно производительном планировщике система может оказаться неустойчивой из-за перегрузки отдельного обслуживающего узла. Если третий узел получает слишком большую долю входного потока, то именно он становится ограничением для всей системы.

В примерах 2–5 исследуется влияние изменения нагрузки в течение дня, производительности планировщика, числа серверов в третьем узле и производительности серверов этого узла. Полученные результаты показывают, что рост входной нагрузки прежде всего увеличивает очередь в наименее производительном узле; повышение производительности планировщика эффективно только до момента, когда задержка в нем перестает быть существенной; увеличение числа серверов или производительности третьего узла восстанавливает устойчивость, по достижении некоторого уровня значений характеристик, дополнительный выигрыш становится несущественным. Из результатов следует, что наиболее полезно усиливать тот компонент системы, который фактически ограничивает работу всей системы.

В примере 6 рассматривается оптимизация параметров модели. Сравнение режимов оптимизации показывает, что настройка только вероятностей маршрутизации не всегда приводит к улучшению, если исходная структурная конфигурация плохо распределяет вычислительные ресурсы. Оптимизация распределения серверов дает существенный эффект, а совместная оптимизация распределения серверов и маршрутизации обеспечивает наименьшее значение среднего времени отклика.

Полученные результаты показывают, что разработанная модель полезна как инструмент предварительной оценки конфигурации дата-центра: она позволяет до внесения изменений определить, какой параметр ограничивает работу системы и какой вариант модернизации даст наибольший эффект.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы разработана и исследована математическая модель гетерогенного дата-центра для анализа и оптимизации его ключевых показателей. Были проанализированы особенности архитектуры гетерогенных дата-центров, представлена модель в виде открытой сети массового обслуживания с планировщиком и несколькими обслуживающими узлами, описан метод расчёта основных характеристик системы, а также рассмотрена задача оптимизации вероятностей маршрутизации и распределения серверов между узлами.

На основе разработанной модели была создана программа на языке Python. Программа позволяет задавать параметры дата-центра, проверять существование стационарного режима, вычислять вероятность блокировки на планировщике, коэффициенты загрузки узлов, средние длины очередей и среднее время отклика, а также выполнять оптимизацию параметров модели и строить графики зависимостей характеристик от изменяемых параметров.

Проведённые вычислительные эксперименты показали, что в гетерогенной системе критический режим может возникать из-за перегрузки отдельного обслуживающего узла даже при высокой производительности планировщика. Для рассмотренной конфигурации основным ограничением стал третий узел; восстановление устойчивого режима обеспечивалось увеличением числа его серверов, повышением их интенсивности обслуживания и оптимизацией вероятностей маршрутизации входного потока.

ОСНОВНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ:

1 Furht, B. Cloud computing fundamentals / B. Furht // Handbook of Cloud Computing / ed. by B. Furht, A. Escalante. - New York : Springer, 2010. - P. 3–19.

2 Voorsluys, W. Introduction to cloud computing / W. Voorsluys, J. Broberg, R. Buyya // Cloud Computing: Principles and Paradigms / ed. by R. Buyya, J. Broberg, A. Goscinski. - Hoboken : Wiley, 2011. - P. 1–41.

3 Delimitrou, C. Paragon: QoS-aware scheduling for heterogeneous datacenters / C. Delimitrou, C. Kozyrakis // ACM SIGARCH Computer Architecture News. - 2013. - Vol. 41, № 1. - P. 77–88.

4 Load balancing for heterogeneous traffic in datacenter networks / J. Wang [et al.] // Journal of Network and Computer Applications. - 2023. - Vol. 217. - Article 103692.

5 Ross, S. M. Introduction to Probability Models / S. M. Ross. - 10th ed. - Amsterdam : Academic Press, 2010. - 801 p.

6 Жадан, В. Г. Методы оптимизации. Часть II. Численные алгоритмы : учебное пособие / В. Г. Жадан. - М. : МФТИ, 2015. - 320 с.