

На правах рукописи

Савельева Анастасия Андреевна

**РАЗРАБОТКА МОДЕЛИ, МЕТОДА И АЛГОРИТМА УПРАВЛЕНИЯ
ИНФОРМАЦИОННЫМИ ПОТОКАМИ РАСПРЕДЕЛЕННОЙ
ТЕЛЕКОММУНИКАЦИОННОЙ СИСТЕМЫ
В ИНФРАСТРУКТУРЕ ТУМАННЫХ ВЫЧИСЛЕНИЙ**

2.2.15. Системы, сети и устройства телекоммуникаций

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2026

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича» на кафедре инфокоммуникационных систем.

Научный руководитель: кандидат технических наук, доцент
Елагин Василий Сергеевич

Официальные
оппоненты: **Колбанёв Михаил Олегович,**
доктор технических наук, профессор,
Санкт-Петербургский государственный
экономический университет,
кафедра информационных систем
и технологий, профессор кафедры

Карташевский Игорь Вячеславович,
доктор технических наук, доцент,
Поволжский государственный университет
телекоммуникаций и информатики,
научно-исследовательская лаборатория
«Инновационные проекты»,
заведующий лабораторией

Ведущая организация: Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики», Москва

Защита состоится 09 декабря 2026 года в 16.00 на заседании диссертационного совета 55.2.004.01, созданном на базе федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича», по адресу: Санкт-Петербург, пр. Большевиков, д. 22, корп. 1, ауд. 554/1.

С диссертацией можно ознакомиться в библиотеке СПбГУТ по адресу Санкт-Петербург, пр. Большевиков, д. 22, корп. 1 и на сайте www.sut.ru.

Автореферат разослан 19 октября 2026 года.

Ученый секретарь
диссертационного совета 55.2.004.01,
канд. техн. наук, доцент

А.Г. Владыко

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Развитие современных телекоммуникационных сетей и систем связано с ростом числа сервисов, формирующих и обрабатывающих значительные объемы данных на периферии сети: Интернета вещей, промышленной автоматизации, интеллектуального видеонаблюдения, транспортных систем и приложений реального времени. По оценкам IDC, объем создаваемых и обрабатываемых данных к 2028 г. может достигнуть 393,9 зеттабайт. В этих условиях централизованная обработка и хранение данных исключительно в облачных центрах обработки данных не всегда позволяют обеспечить требуемое качество обслуживания сервисов. Международные документы в области граничных и туманных вычислений рассматривают вычислительные, сетевые и запоминающие ресурсы как взаимосвязанные компоненты инфраструктуры, размещаемые ближе к пользователю и источникам данных. ITU-T Y.3540 определяет инфраструктуру граничных вычислений через ресурсы обработки, сети и хранения, а IEEE 1934-2018 рассматривает туманные вычисления как распределенную архитектуру, размещаемую между облачными центрами обработки данных и конечными устройствами. Перенос операций доступа к данным на уровень туманных вычислений сокращает зависимость времени отклика от состояния транспортных сегментов сети, но усложняет подсистему хранения. Распределенное хранение с репликацией требует межузлового обмена, подтверждения операций записи и восстановления избыточности, использующих коммуникационные и вычислительные ресурсы наряду с пользовательскими операциями. При этом увеличение уровня избыточности повышает готовность хранения, но увеличивает служебный межузловой обмен и может повышать время завершения операции записи. Поэтому актуальность работы определяется необходимостью исследований и разработок в области управления информационными потоками распределенной подсистемы хранения данных телекоммуникационной системы в инфраструктуре туманных вычислений при соблюдении требований к готовности выполнения операций записи и ограничений на использование коммуникационных ресурсов.

Степень разработанности темы. Существенный вклад в формирование и развитие научных исследований перспективных телекоммуникационных сетей и систем, инфраструктуре и распределению ресурсов таких систем и их параметров внесли В.М. Вишневский, А.Н. Волков, Ю.В. Гайдамака, А.Б. Гольдштейн, Б.С. Гольдштейн, В.Г. Карташевский, А.Е. Кучерявый, Е.А. Кучерявый, Р.В. Киричек, М.А. Маколкина, А.И. Парамонов, К.Е. Самуйлов, С.Н. Степанов, В.О. Тихвинский, А.А.А. Ateya, F. Bonomi, E. Brewer, R. Buyya, M. Chiang, D. Gifford, J. Gray, L. Kleinrock, L. Lamport, M. Satyanarayanan и другие ученые. Вместе с тем вопросы управления пользовательским и служебным трафиком распределенной подсистемы хранения, размещаемой на уровне туманных вычислений, при одновременном учете репликации, сетевых задержек и ограничений по готовности требуют дальнейшего исследования.

Научная задача. В диссертации решается научная задача разработки модели, метода и алгоритма управления информационными потоками распределенной подсистемы хранения данных телекоммуникационной системы в инфраструктуре туманных вычислений, направленных на повышение производительности операций записи с репликацией как механизмом обеспечения отказоустойчивости и качества обслуживания при соблюдении ограничений по готовности, размещению реплик и служебному межузловому обмену.

Предмет и объект исследования. Объектом исследования является распределенная подсистема хранения данных телекоммуникационной системы, функционирующей в инфраструктуре туманных вычислений. Предметом исследования являются модели, методы и алгоритмы управления информационными потоками при выполнении операций записи с репликацией данных, обеспечивающие повышение производительности подсистемы хранения при ограничениях по готовности и использованию коммуникационных ресурсов.

Цель работы. Целью работы является разработка модели, метода и алгоритма управления информационными потоками распределенной подсистемы хранения данных телекоммуникационной системы, размещаемой на уровне туманных вычислений, обеспечивающих повышение допустимой интенсивности входного потока операций записи с репликацией данных, измеряемой в заявках в секунду, при соблюдении заданных ограничений на время завершения операции записи, 95-й и 99-й процентиля задержки, готовность выполнения операции записи и объем служебного межузлового обмена.

Задачи работы:

1) выполнить анализ архитектурных особенностей распределенной подсистемы хранения данных, размещаемой на уровне туманных вычислений, и определить факторы, влияющие на выполнение операций записи с репликацией данных,

2) обосновать выбор показателей производительности распределенной подсистемы хранения данных и определить основной показатель оценки, – допустимую интенсивность входного потока операций записи, с учетом ограничений по времени завершения операции, верхним процентилем задержки, готовности и служебному межузловому обмену,

3) разработать модель управления информационными потоками распределенной подсистемы хранения данных, учитывающую пользовательские операции записи, служебные операции репликации и восстановления избыточности, очереди обслуживания на узлах, ограничения обслуживающих ресурсов и пропускной способности межузловых соединений,

4) разработать метод оценки готовности выполнения операции записи с репликацией данных, учитывающий фактор репликации, порог подтверждения записи и доступность элементов цепочки обслуживания операции,

5) разработать алгоритм выбора узлов хранения при выполнении операции записи с репликацией данных, учитывающий пользовательскую очередь, служебную нагрузку, сетевую задержку и ограничение по готовности выполнения операции записи,

б) провести экспериментальную оценку предложенного алгоритма в сравнении с базовыми стратегиями распределения нагрузки и определить его влияние на допустимую интенсивность входного потока операций записи, время завершения операции и верхние проценты задержки.

Теоретическая значимость исследования заключается в развитии модельно-методического аппарата управления информационными потоками распределенной подсистемы хранения данных туманного уровня. Репликация рассматривается одновременно как механизм обеспечения отказоустойчивости и источник служебной межузловой нагрузки. Полученные результаты дополняют существующие подходы к анализу распределенных подсистем хранения за счет совместного учета пользовательской и служебной нагрузки, сетевых задержек и готовности выполнения операции записи.

Практическая значимость исследования заключается в возможности применения полученных результатов при проектировании и настройке распределенных программно-определяемых подсистем хранения данных в составе телекоммуникационных систем и сетей 5G/6G с использованием инфраструктуры периферийных и туманных вычислений. Результаты могут использоваться операторами связи и провайдерами инфраструктурных сервисов при размещении функций хранения и обработки данных на узлах телекоммуникационной сети. Разработанные модель управления информационными потоками, метод оценки готовности и алгоритм выбора узлов хранения позволяют обосновывать параметры репликации, режим подтверждения записи и распределение операций записи.

Методология и методы исследования. В работе использованы методы математического и имитационного моделирования, теория массового обслуживания, моделирования процессов репликации, методы расчета коэффициента готовности, управления потоками и сравнительной оценки.

Научная новизна:

1) Разработана модель управления информационными потоками распределенной подсистемы хранения данных телекоммуникационной системы, размещенной в инфраструктуре туманных вычислений, отличающаяся от известных тем, что репликация представлена одновременно как механизм обеспечения отказоустойчивости хранения и как источник служебного межузлового трафика, конкурирующего с пользовательским трафиком за коммуникационные и вычислительные ресурсы. Что в свою очередь позволяет формализовать управление информационными потоками, представленными пользовательским и служебным трафиком и формирующими нагрузку на узлы хранения и межузловые соединения, при ограничениях по очередям обслуживания, пропускной способности соединений и готовности выполнения операции записи.

2) Разработан метод оценки готовности выполнения операции записи с репликацией данных, отличающийся учетом фактора репликации, порога подтверждения записи и доступности элементов цепочки обслуживания операции записи. Метод позволяет обосновывать параметры репликации как механизма отказоустойчивости не изолированно, а с учетом их влияния на качество обслуживания, время завершения операции записи и объем служебного межузлового обмена.

3) Разработан алгоритм интегрального выбора узлов хранения при выполнении операции записи с репликацией данных, отличающийся учетом пользовательской очереди, служебной нагрузки, сетевой задержки и ограничения по готовности при формировании подмножества узлов размещения реплик. Алгоритм обеспечивает выбор узлов с учетом двойственной роли репликации: как средства отказоустойчивого хранения и как фактора, влияющего на задержки, служебную нагрузку и качество обслуживания операций записи.

Положения, выносимые на защиту:

1) Модель управления информационными потоками распределенной подсистемы хранения данных туманного уровня телекоммуникационной системы, отличающаяся от известных тем, что в ней отдельно учитываются пользовательские операции записи и служебные операции поддержания репликации, что позволяет формализовать их распределение при ограничениях по очередям обслуживания, пропускной способности межузловых соединений и требуемому коэффициенту готовности выполнения операции записи.

2) Метод оценки готовности выполнения операции записи с репликацией данных в распределенной подсистеме хранения, размещаемой на уровне туманных вычислений, позволяет обосновать выбор режима подтверждения записи с учетом числа реплик, порога подтверждения и общей части цепочки обслуживания. Для исследуемой конфигурации при $R = 3$, $MTTR = 30$ мин и $A_{\text{общ}} = 0,99999$ переход от строгого режима подтверждения к пороговому снижает расчетную ожидаемую недоступность функции записи с 22,334 до 5,260 мин/год.

3) Алгоритм выбора узлов хранения при выполнении операции записи с репликацией данных, отличающийся от известных тем, что выбор подмножества узлов выполняется на основе интегральной оценки, учитывающей пользовательскую очередь, служебную нагрузку, сетевую задержку и ограничение по готовности выполнения операции записи, что позволяет увеличить допустимую интенсивность входного потока операций записи на 48,3 % по сравнению с равномерным циклическим распределением при соблюдении заданного ограничения на время завершения операции. При этом дополнительное преимущество относительно стратегии, учитывающей пользовательскую и служебную очереди, составляет 2,4 %.

Степень достоверности. Достоверность результатов подтверждается корректным применением математического аппарата, обсуждением на научных конференциях и публикацией в рецензируемых изданиях.

Апробация результатов. Основные результаты докладывались на конференциях SYNCHROINFO (2018, 2021, 2022), ICUMT (2020, 2021), АПИНО (2019-2021, 2025, 2026), MoNeTeC (2024), и III Международной молодежной конференции по информационной безопасности (2025).

Соискатель является исполнителем НИР шифр «Мульти-ЦОД» по контракту с Минцифры России № ПЗЗ-1-26/5 (2021-2023), связанной с повышением производительности и отказоустойчивости программно-определяемых СХД.

Результаты исследований получили внешнюю конкурсную апробацию: соискатель стала победителем конкурса грантов Комитета по науке и высшей школе Санкт-Петербурга для аспирантов в 2019 и 2021 гг., в составе авторского коллектива – конкурса лучших инновационных проектов в сфере науки и высшего образования Санкт-Петербурга в 2021 г. В 2025 г. соискателю присуждена «Молодежная премия Санкт-Петербурга» в номинации «В области информационных технологий» за результаты, связанные с тематикой исследования.

Публикации по теме диссертации. Материалы работы изложены в полной мере в 27 публикациях, из них: 4 статьи в изданиях, включенных в перечень рецензируемых научных изданиях (перечень ВАК при Минобрнауки России), 6 статей в изданиях, включенных в международные базы цитирования; 3 результата интеллектуальной деятельности; 8 отчетов о НИР; 6 статей в других изданиях и материалах конференций.

Личный вклад автора. Результаты диссертационной работы получены автором самостоятельно. В работах, выполненных в соавторстве, автор принимала непосредственное участие в проведении исследований, анализе и обобщении полученных результатов.

Работа соответствует пп. 1, 3, 4, 5, 12 паспорта специальности 2.2.15, поскольку направлена на разработку моделей и методов исследования и проектирования элементов телекоммуникационных систем и сетей (п. 1), исследование процессов хранения информации (п. 3) и архитектуры распределенной подсистемы хранения данных как элемента телекоммуникационной системы (п. 4), а также на управление информационными потоками (п. 5) и распределение коммуникационных ресурсов в инфраструктуре туманных вычислений (п. 12).

Внедрение результатов работы. Результаты внедрены в АО «ЦВКС «МСК-IX», ООО «ВАС Экспертс», используются в СПбГУТ при проведении лекций и практических занятий по курсам «Основы построения инфокоммуникационных систем и сетей», «Сети связи и системы коммутации».

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения, списка литературы, раздела определений, обозначений и сокращений. Полный объем диссертации составляет 217 стр., приведены 28 рис., 11 табл., 71 формула. Список литературы включает 143 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы исследования, сформулированы научная задача, цель и задачи работы, определены объект и предмет исследования, раскрыты научная новизна, теоретическая и практическая значимость, приведены положения, выносимые на защиту, сведения об апробации, публикациях и личном вкладе автора.

В первой главе рассмотрены архитектурные особенности распределенного хранения данных в телекоммуникационных системах граничных и туманных вычислений. Показано, что размещение функций хранения ближе к источникам и потребителям данных снижает зависимость времени отклика от транспортных сегментов сети, при этом туманный уровень рассматривается как среда выполнения операций чтения и записи.

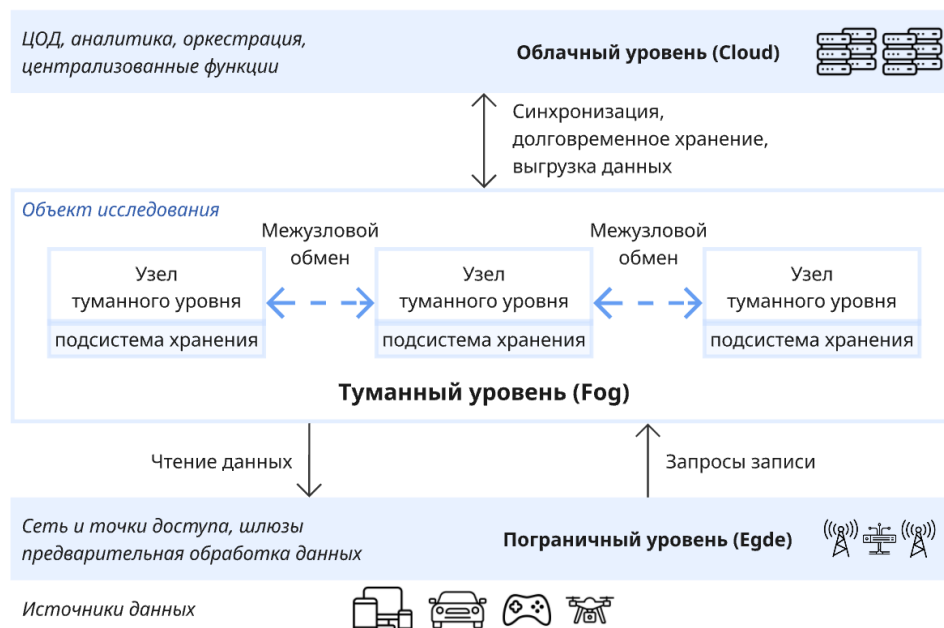


Рисунок 1 – Обобщенная схема взаимодействия уровней распределенной инфраструктуры

Установлено, что наиболее существенное значение имеет операция записи с репликацией (рисунок 2), включающая передачу данных на несколько узлов, фиксацию реплик, получение подтверждений и служебные операции согласования, что определяет необходимость совместного учета задержек, готовности и служебной нагрузки.

В качестве архитектурной основы исследования выбрана распределенная программно-определяемая подсистема хранения, обеспечивающая распределенное размещение узлов, масштабирование и поддержку отказоустойчивых режимов. В качестве механизма обеспечения избыточности обоснован выбор репликации. Показано, что увеличение числа реплик повышает устойчивость хранения, но увеличивает межузловой обмен, нагрузку на коммуникационные ресурсы и время завершения операции записи. Поэтому параметры репликации и режим подтверждения должны выбираться с учетом требований к готовности и производительности.



Рисунок 2 – Схема обслуживания операции записи с репликацией

Анализ подходов к балансировке нагрузки показал недостаточную разработанность схем, обеспечивающих одновременный учет неоднородности узлов, сетевых задержек, пользовательского и служебного трафика, что обосновывает необходимость разработки модели управления информационными потоками с учетом указанных факторов.

Во второй главе разработаны метод оценки готовности выполнения операции записи и модель управления информационными потоками распределенной подсистемы хранения данных телекоммуникационной системы на уровне туманных вычислений. В рекомендациях ITU-T готовность рассматривается как характеристика качества функционирования сетей связи. Предложенный метод учитывает параметры репликации, порог подтверждения записи и доступность инфраструктурных элементов цепочки обслуживания операции.

Распределенная подсистема хранения предоставляет набор целевых функций: запись, чтение и хранение. Работоспособность сервиса задается вектором функциональной готовности (1), включающим готовность сетевой инфраструктуры и выполнения указанных функций. Поскольку запись является технологически наиболее ресурсоемкой операцией, в качестве лимитирующей компоненты рассматривается функция записи (2), где R – фактор репликации, W – порог подтверждения.

$$A_{\text{сервиса}} = (A_{\text{сеть}}, A_{\text{запись}}, A_{\text{чтение}}, A_{\text{хранение}}), \quad (1)$$

$$A_{\text{запись}}(R, W) = \sum_{k=W}^R \binom{R}{k} A_{\text{узел}}^k (1 - A_{\text{узел}})^{R-k}. \quad (2)$$

Выполнена расчетная оценка готовности выполнения операции записи при изменении среднего времени восстановления, параметров репликации и режима подтверждения (рисунки 3, 4).

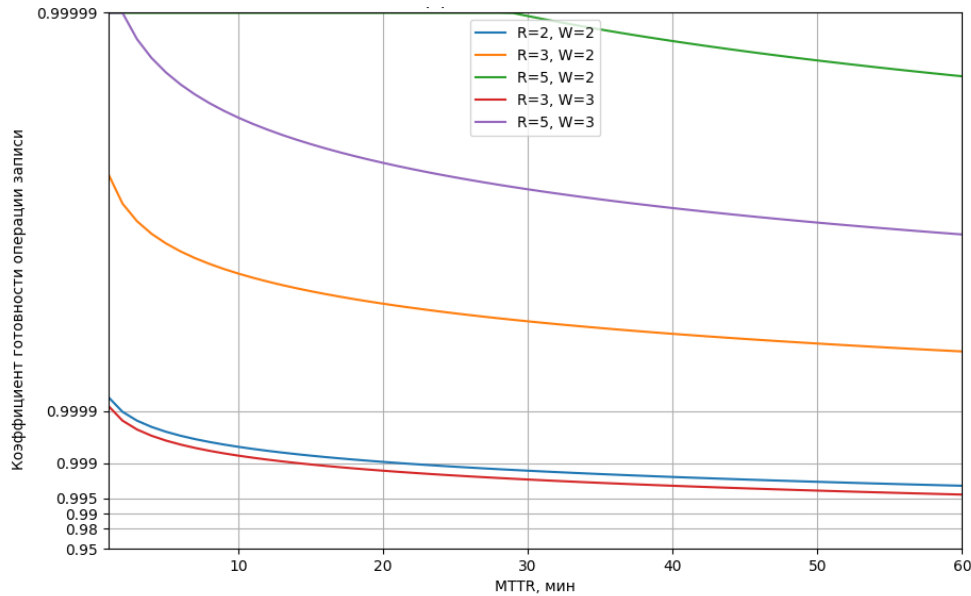


Рисунок 3 – Зависимость коэффициента готовности выполнения операции записи от среднего времени восстановления

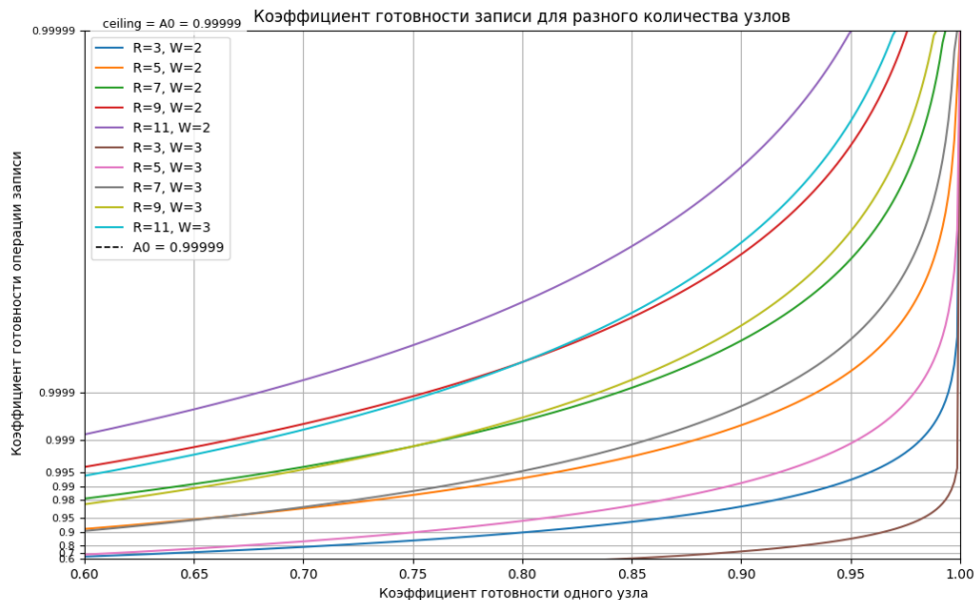


Рисунок 4 – Зависимость коэффициента готовности узла выполнения операции записи от параметров репликации и режима подтверждения

В качестве основного показателя производительности используется допустимая интенсивность входного потока операций записи, определяемая с учетом пропускной способности системы как максимальная интенсивность поступления заявок, при которой сохраняется заданный критерий допустимости по времени завершения операции записи. Дополнительно контролируются 95-й (p95) и 99-й (p99) процентиля задержки операций записи, а также показатели нагрузки от потока служебных операций и время восстановления избыточности.

Построена математическая модель распределенной подсистемы хранения, в которой информационные потоки представлены пользовательским и служебным трафиком, формирующим нагрузку на узлы хранения и межузловые соединения.

Управляющее решение задается выражением $u(k) = \{c_i(k), b_i(k), \pi(k)\}$, где $c_i(k)$ – число пользовательских операций записи, направляемых на i -й узел хранения, $b_i(k)$ – коэффициент регулирования служебной нагрузки, задающий долю обслуживающего ресурса i -го узла, выделяемую служебным операциям, $\pi(k)$ – политика размещения и переразмещения данных, определяющая схему выбора узлов и наборов реплик. При этом допустимая область решений формируется с учетом ресурсных и сетевых ограничений на межузловой обмен $\Phi_e(k) \leq N_e, \forall e \in \mathcal{E}$, ограничения по коэффициенту готовности $A \geq A_{\min}$ и ограничения на служебную нагрузку (3).

$$\limsup_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{i \in \mathcal{N}} B_i(k) \leq B_{\max}, \quad (3)$$

Итоговое представление допустимой области решений приведено в выражении $\mathcal{U}(k) = \mathcal{U}_{\text{модель}}(k) \cap \mathcal{U}_A(k) \cap \mathcal{U}_{\text{восст}}(k)$.

Сформулирована задача управления информационными потоками, целевая функция которой задана выражением (4).

$$\min_{\{u(k)\}} \limsup_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} \left(\sum_{i \in \mathcal{N}} w_Q Q_i(k) + \sum_{i \in \mathcal{N}} w_B B_i(k) + \sum_{i \in \mathcal{N}} w_N \text{Net}(k) \right), \quad (4)$$

где $w_Q, w_B, w_N \geq 0$ – веса, задающие приоритет качества обслуживания пользовательских операций $Q_i(k)$ относительно интенсивности служебных операций $B_i(k)$ и объема сетевого обмена $\text{Net}(k)$.

Таким образом, повышение производительности соответствует увеличению допустимой входной нагрузки при соблюдении ограничений по задержке, готовности и служебному межузловому обмену.

Третья глава. Наличие избыточного множества узлов хранения требует определения правила их выбора при выполнении операции записи с репликацией данных. Модель, сформированная во второй главе, далее используется как основа для построения алгоритма выбора узлов хранения. Переменные состояния характеризуют пользовательскую и служебную составляющие нагрузки, а допустимая область решений $\mathcal{U}(k)$ задает ограничения по готовности, восстановлению избыточности и использованию ресурсов инфокоммуникационной инфраструктуры. На этой основе в третьей главе формируется алгоритм выбора управляющего воздействия $u(k) \in \mathcal{U}(k)$ при выполнении операции записи с репликацией данных.

Предлагаемый алгоритм основан на учете текущего состояния распределенной инфраструктуры (5) и предполагает выбор узлов хранения с учетом пользовательских очередей, служебной межузловой нагрузки, сетевых задержек и ограничений по готовности. Данный подход позволяет снизить вероятность направления реплик на перегруженные или недостаточно надежные элементы системы, а также учитывать влияние служебного трафика на качество обслуживания пользовательских операций.

$$C_i(k) = \alpha \hat{Q}_i(k) + \beta \hat{B}_i(k) + \gamma \hat{d}_i(k), \quad (5)$$

где $\hat{Q}_i(k)$, $\hat{B}_i(k)$, $\hat{d}_i(k)$ – нормированные значения пользовательской очереди, служебной нагрузки и сетевой задержки соответственно, α , β , γ – весовые коэффициенты, задающие относительную значимость указанных факторов. Показатель готовности узла $A_i(k)$ не включается в интегральную оценку $C_i(k)$, а используется как ограничение допустимости при формировании множества узлов размещения.

На основе модели, сформированной во второй главе, предложен алгоритм выбора управляющего воздействия по наблюдаемому состоянию распределенной подсистемы хранения (рисунок 5.а, 5.б). Управляющее воздействие включает выбор множества узлов хранения для размещения реплик операции записи и регулирование доли ресурса, выделяемой служебным операциям.

Формализована процедура формирования множества допустимых узлов с учетом ограничений по готовности, пользовательской и служебной нагрузке. Внутри указанного множества выбор узлов выполняется по критерию, учитывающему пользовательскую очередь, служебную нагрузку и сетевую задержку. Такой подход позволяет учитывать локальное состояние узлов и влияние межузлового обмена, возникающего при репликации, синхронизации и восстановлении данных.

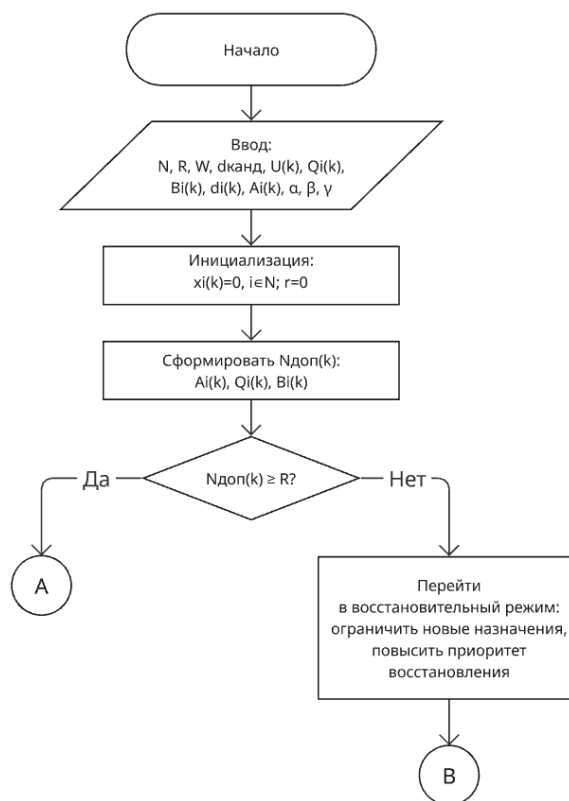


Рисунок 5.а – Алгоритм выбора узлов хранения и регулирования нагрузки при выполнении операции записи с репликацией данных

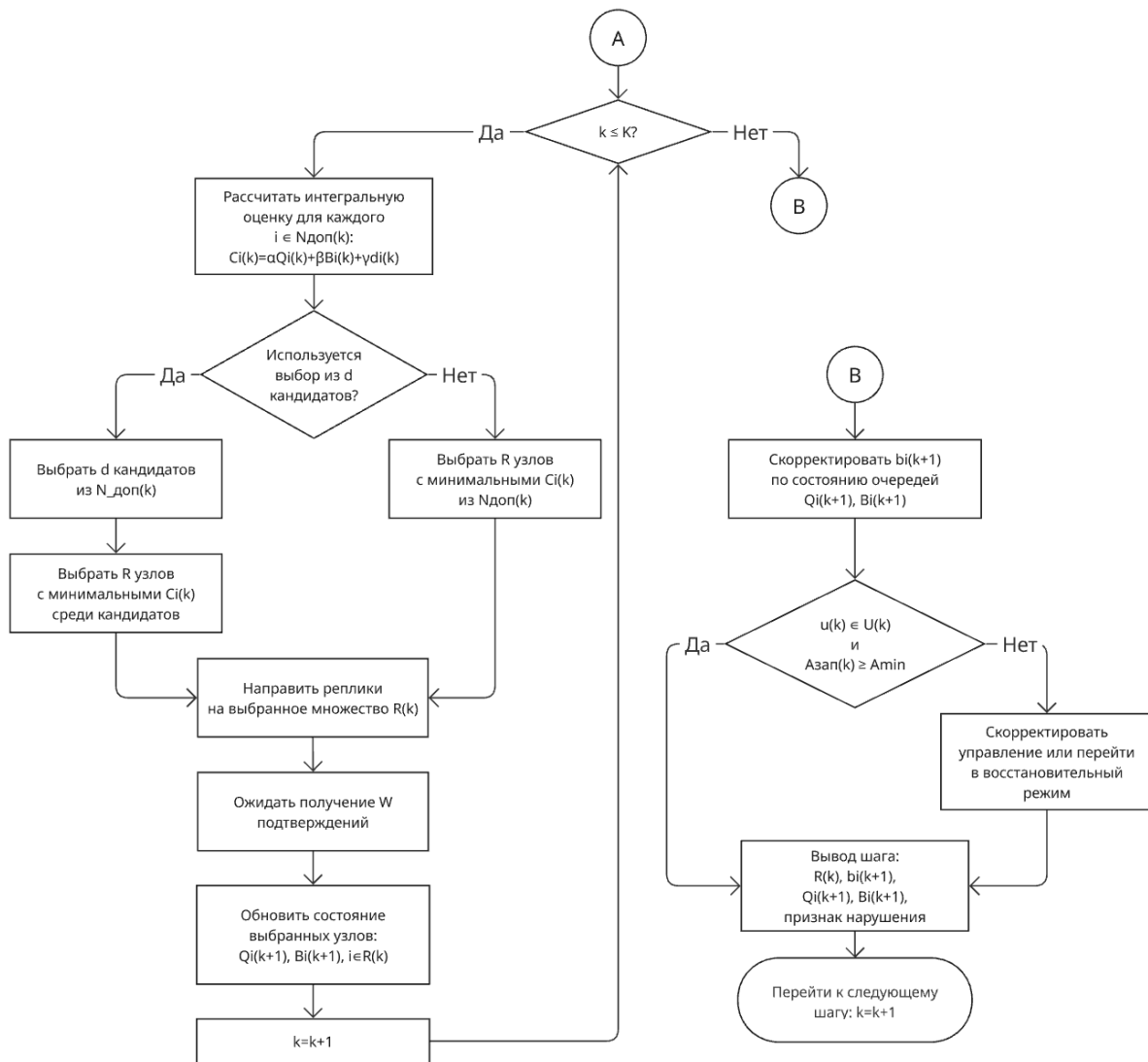


Рисунок 5.6 – Алгоритм выбора узлов хранения и регулирования нагрузки при выполнении операции записи с репликацией данных

Показано, что предложенный алгоритм использует измеримые или оцениваемые эксплуатационные параметры распределенной подсистемы хранения и может быть реализован как при полном рассмотрении множества узлов, так и при выборе из ограниченного числа кандидатов. Выполнено сопоставление предложенного алгоритма с базовыми стратегиями: равномерным распределением, выбором по минимальной пользовательской очереди и выбором с учетом пользовательской и служебной нагрузки. Сформирована методика сопоставительного эксперимента, используемая в четвертой главе для оценки алгоритма при изменении входной нагрузки, неоднородности очередей, сетевых задержек, режимов подтверждения записи и состава доступных узлов.

Четвертая глава. Экспериментальная оценка проводится для четырех стратегий выбора узлов хранения: равномерного циклического распределения, выбора по минимальной пользовательской очереди, выбора с учетом пользовательской и служебной очередей и предложенного алгоритма интегрального выбора узлов. Для всех стратегий учитывается ограничение по готовности выполнения операции записи.

Оценка выполнена в два этапа: на первом сопоставлены строгий и пороговый режимы подтверждения записи, на втором – стратегии выбора узлов при изменении входной и служебной нагрузки и неоднородности сетевых задержек. В предложенном алгоритме показатель готовности учитывается при формировании множества допустимых узлов, после чего выполняется ранжирование по пользовательской очереди, служебной нагрузке и сетевой задержке. Основными экспериментальными показателями являются время завершения операции записи и допустимая интенсивность входного потока; дополнительно анализируются верхние процентиля задержки и состояние служебной нагрузки.

На первом этапе сопоставлены последовательная и параллельная интерпретации операции записи. Параллельный вариант соответствует Fork-Join-представлению (рисунок 6): передача реплик, локальная запись и возврат подтверждений выполняются одновременно по нескольким ветвям обслуживания.

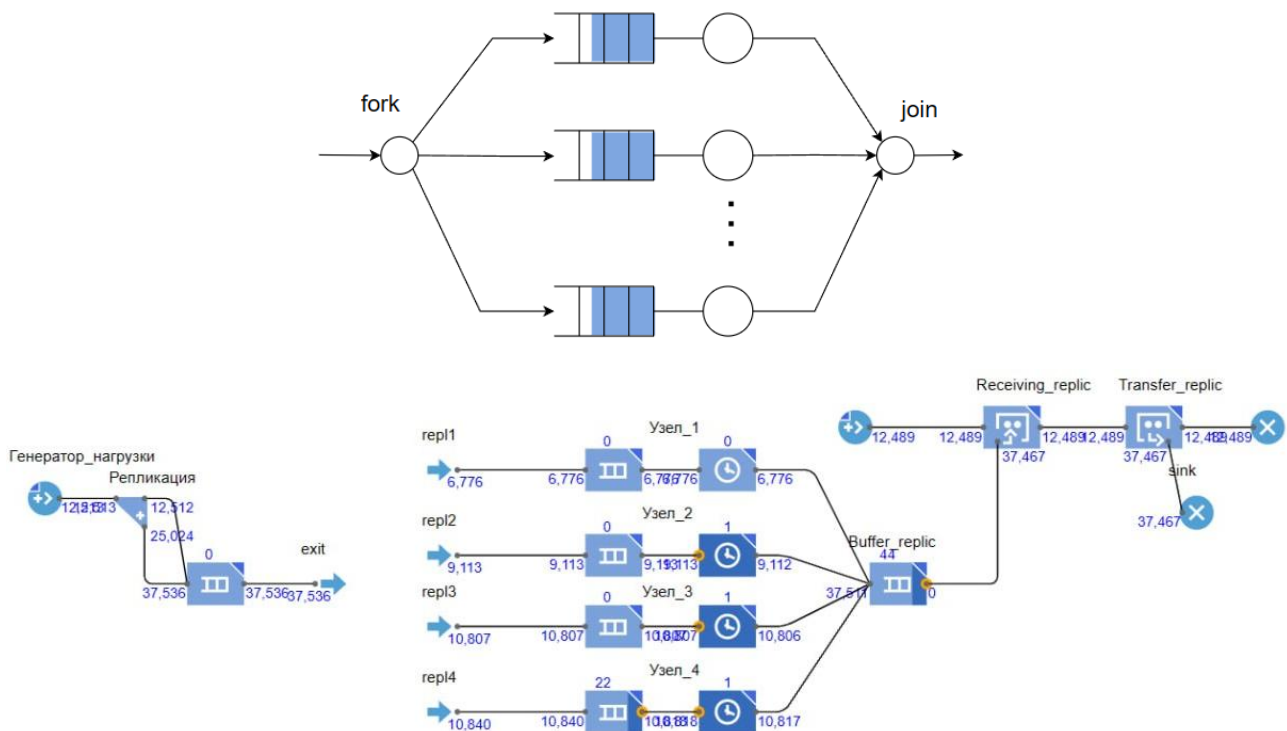


Рисунок 6 – Модель операции записи с репликацией данных (схематично и в программе AnyLogic)

Результаты (рисунок 7) показывают согласованность расчетной и имитационной моделей в области низкой и средней загрузки. При приближении к области высокой загрузки расхождение возрастает, что связано с нелинейным накоплением очередей и усилением влияния случайных задержек обслуживания. При этом расчетная модель сохраняет применимость для сопоставительной оценки сценариев, поскольку основное значение имеют относительные изменения при изменении стратегии выбора узлов.

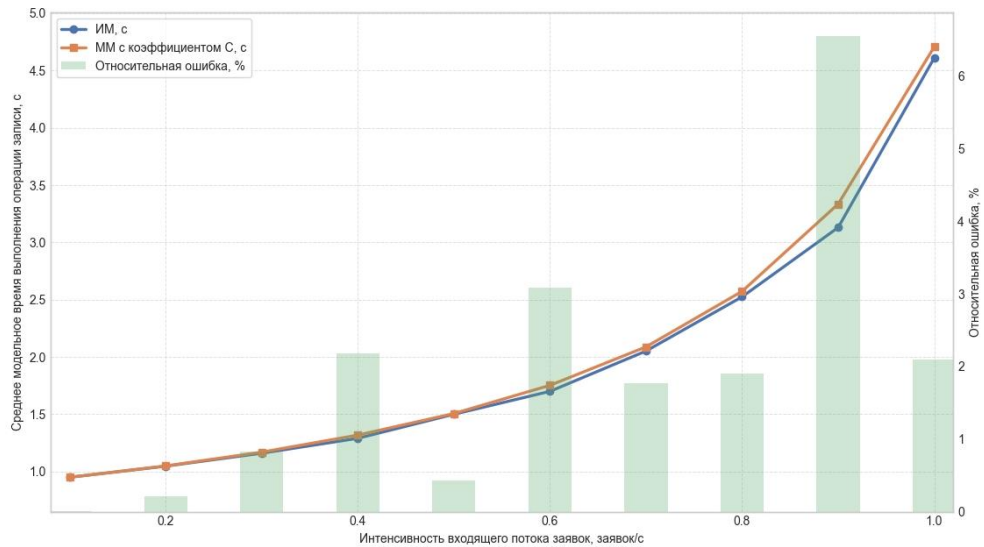


Рисунок 7 – Сопоставление расчетной и имитационной оценок времени записи

Результаты показывают (рисунок 8), что строгий режим подтверждения более чувствителен к локальным перегрузкам и неоднородности сетевых задержек: время завершения операции определяется наиболее длительной ветвью обслуживания. Пороговый режим снижает влияние таких ветвей на пользовательское время ответа, однако не устраняет их вклад в использование ресурсов узлов и формирование служебной нагрузки. Поэтому снижение времени ответа в пороговом режиме должно оцениваться совместно с состоянием служебных очередей и межузлового обмена.

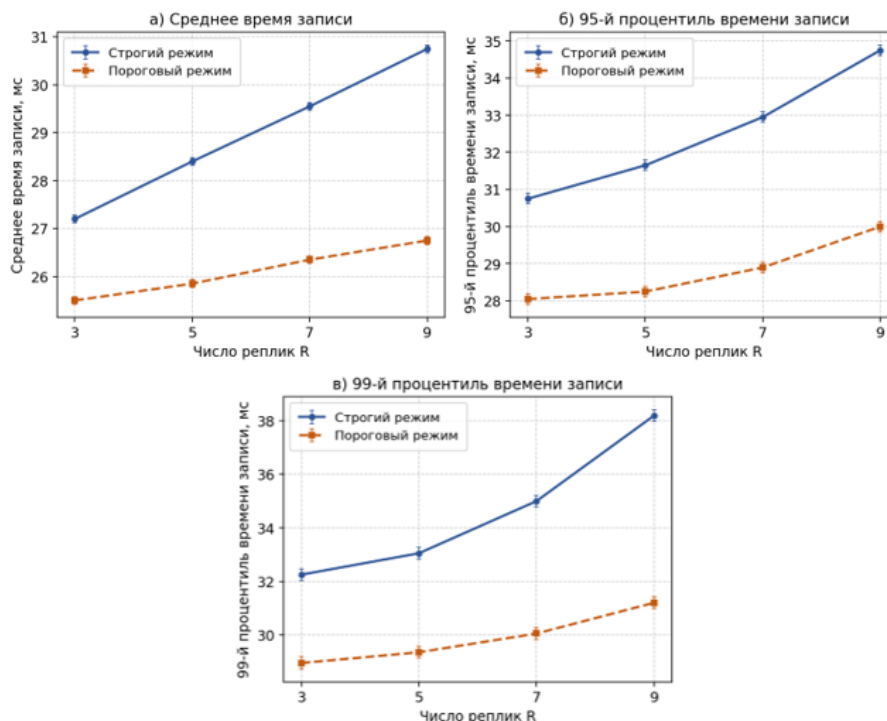


Рисунок 8 – Сравнение времени записи при строгом и пороговом режимах подтверждения

Для оценки влияния стратегий выбора узлов хранения на область допустимых нагрузок рассчитана допустимая интенсивность входного потока λ – последняя проверенная точка нагрузки, при которой выполняется заданный критерий допустимости.

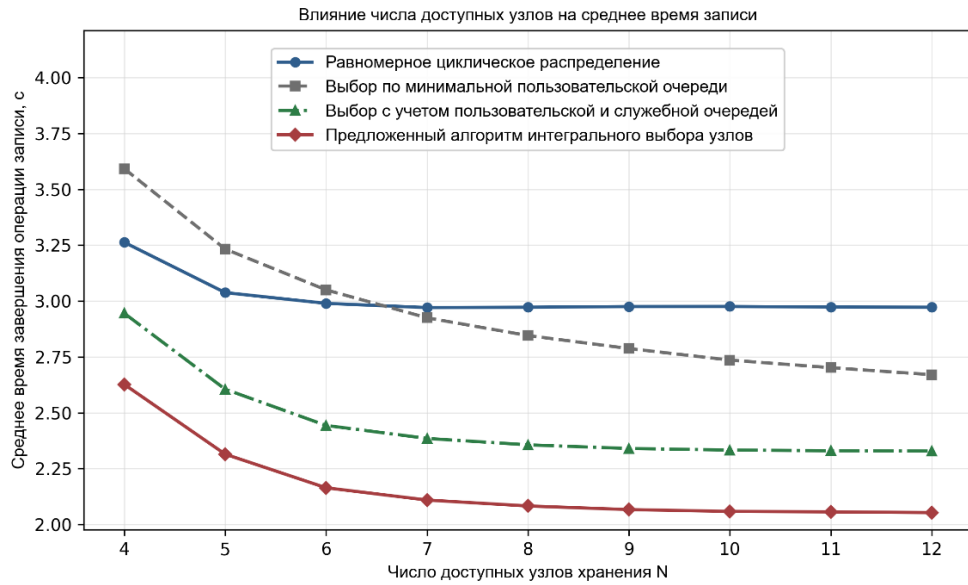


Рисунок 9 – Влияние выбора узлов из избыточного множества на время завершения операции записи при различных стратегиях

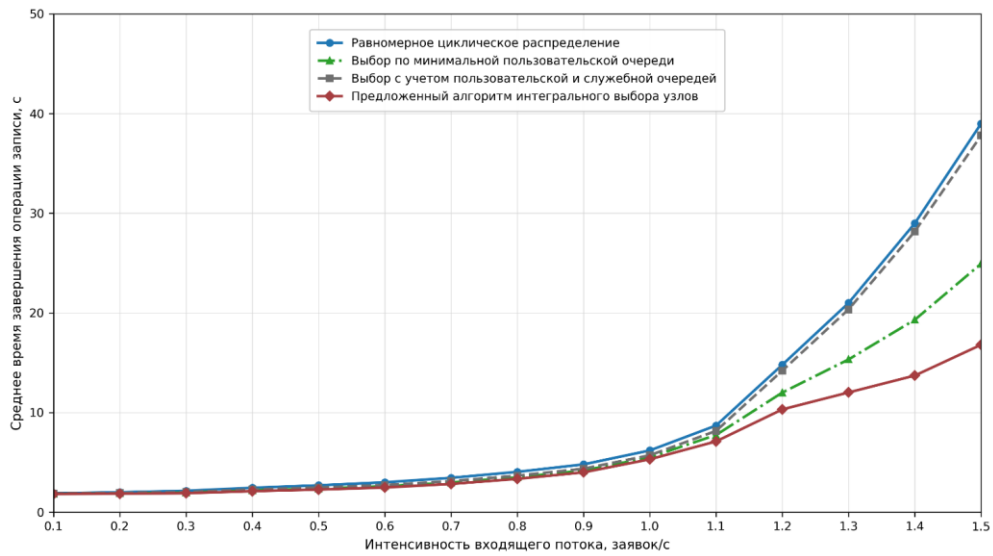


Рисунок 10 – Сравнение стратегий распределения операций записи при увеличении задержки одной ветви обслуживания на 75 %

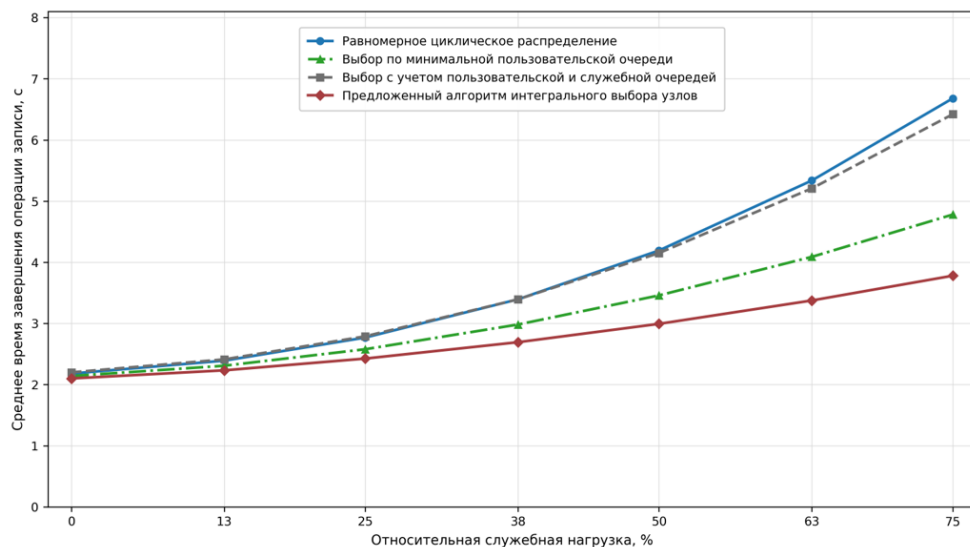


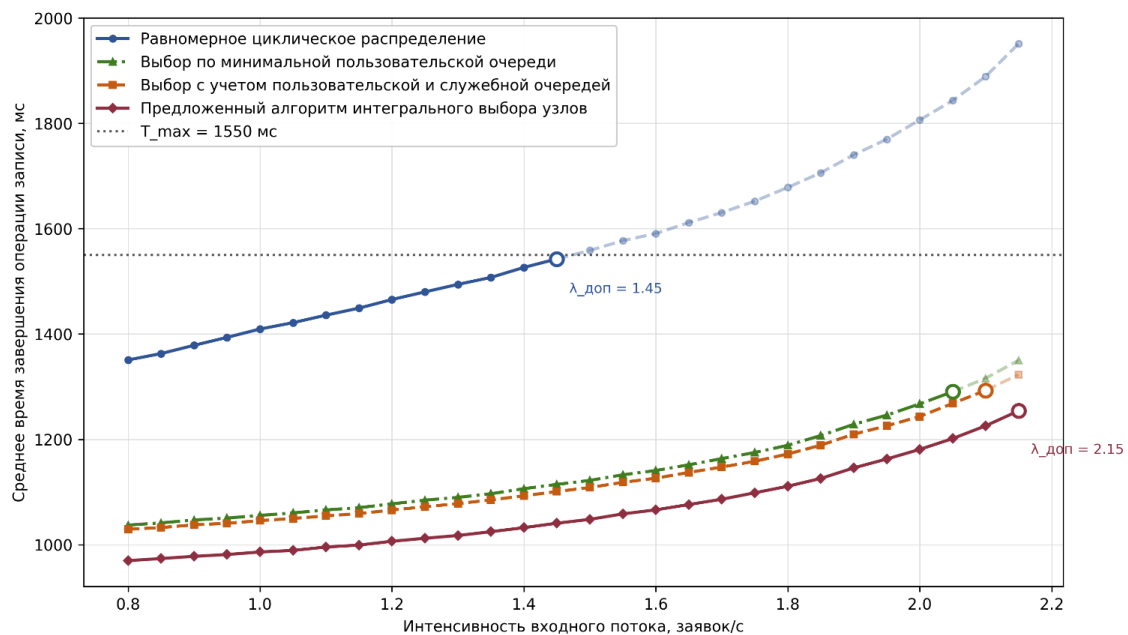
Рисунок 11 – Влияние служебной нагрузки на время завершения операции записи при различных стратегиях выбора узлов при $\lambda = 1,0$ заявок/с

Таблица – Фрагмент оценки допустимой интенсивности входного потока

Стратегия выбора узлов хранения	Учитываемые признаки состояния	$\lambda_{\text{доп}}$, заявок/с	$\bar{T}_{\text{write}}$ при $\lambda_{\text{доп}}$, мс	p^{99} при $\lambda_{\text{доп}}$, мс	Прирост относительно RR, %
Равномерное циклическое распределение	Текущее состояние узлов не учитывается	1,45	1542,49	2973,78	-
Выбор по минимальной пользовательской очереди	$Q_i(k)$	2,05	1290,37	2953,75	41,4
Выбор с учетом пользовательской и служебной очередей	$Q_i(k), B_i(k)$	2,10	1292,99	2982,12	44,8
Предложенный метод интегрального выбора узлов	$Q_i(k), B_i(k), d_i(k), A_i(k) \geq A_{\min}$	$\geq 2,15$	1254,21	2991,68	не менее 48,3

Из данных таблицы следует, что равномерное циклическое распределение достигает границы допустимого режима при 1,45 заявок/с. Стратегии, использующие информацию о состоянии узлов, обеспечивают расширение области допустимых нагрузок. Для стратегии с учетом пользовательской и служебной очередей последняя допустимая точка составляет 2,10 заявок/с, а для предложенного алгоритма интегрального выбора узлов – 2,15 заявок/с.

Как видно из результатов (рисунок 12), при увеличении интенсивности входного потока равномерное циклическое распределение быстрее достигает границы допустимого режима. Для предложенного алгоритма интегрального выбора узлов рост среднего времени завершения операции записи остается более пологим в исследованном диапазоне нагрузок, что обеспечивает увеличение до 2,15 заявок/с.



Участки кривых после $\lambda_{\text{доп}}$ соответствуют расчетным значениям за пределами допустимого режима.

Рисунок 12 – Зависимость среднего времени завершения операции записи от интенсивности входного потока при строгом критерии допустимости

Для повышения статистической устойчивости результатов каждая точка экспериментальных зависимостей рассчитывалась по 30 независимым реализациям имитационной модели. В одной реализации моделировалось 10 000 операций записи, что соответствует 300 000 смоделированных операций записи на одну точку графика по одной стратегии.

По результатам расчета допустимой интенсивности входного потока установлено, что предложенный алгоритм интегрального выбора узлов позволяет увеличить допустимую интенсивность операций записи на 48,3 % по сравнению с равномерным циклическим распределением при соблюдении ограничений на среднее время завершения операции записи, значение p_{99} и готовность выбранного множества узлов. Дополнительное преимущество относительно стратегии, учитывающей пользовательскую и служебную очереди, составляет 2,4 %.

Полученные результаты показывают, что основной вклад в расширение области допустимых нагрузок обеспечивает переход от статического распределения к стратегиям, учитывающим состояние узлов. Сетевая составляющая в предложенном алгоритме выполняет роль дополнительного уточняющего фактора при выборе множества реплик. Расчет допустимой интенсивности входного потока обобщает результаты отдельных сценариев и показывает границу режима, в котором одновременно выполняются ограничения по времени завершения операции записи, 99-му процентилю задержки и готовности выбранного множества узлов.

Экспериментальная оценка подтверждает применимость алгоритма интегрального выбора узлов как механизма управления информационными потоками в неоднородной распределенной подсистеме хранения данных, размещаемой в инфраструктуре туманных вычислений.

Основные результаты экспериментальной оценки заключаются в следующем:

- показано, что операция записи с репликацией данных должна рассматриваться как параллельный процесс с ожиданием подтверждений записи, поскольку последовательная интерпретация завышает оценку времени записи и не отражает одновременную передачу реплик на несколько узлов хранения,

- установлено, что режим подтверждения записи существенно влияет на временные характеристики операции: строгий режим повышает чувствительность к локальным перегрузкам и неоднородности сетевых задержек, а пороговый режим снижает пользовательское время ответа, но сохраняет вклад незавершенных ветвей обслуживания в служебную нагрузку,

- подтверждено, что при наличии избыточного множества доступных узлов стратегия их выбора становится самостоятельным механизмом управления временем выполнения операции записи и объемом служебного межузлового обмена,

– установлено, что учет пользовательской очереди, служебной нагрузки и сетевой задержки позволяет снизить чувствительность времени записи к локальным перегрузкам и неоднородности телекоммуникационной инфраструктуры,

– определена область применимости предложенного алгоритма: наибольший эффект достигается при росте входной нагрузки, наличии служебных операций, различии сетевых задержек и возможности выбора узлов из множества доступных узлов.

ЗАКЛЮЧЕНИЕ

В диссертационной работе разработаны и экспериментально проверены модель управления информационными потоками, метод оценки готовности и алгоритм интегрального выбора узлов хранения для распределенной подсистемы хранения данных телекоммуникационной системы, размещаемой в инфраструктуре туманных вычислений. Основные результаты диссертационной работы заключаются в следующем:

1) Проведен анализ архитектурных особенностей распределенного хранения данных в телекоммуникационных системах граничных и туманных вычислений. Показано, что при переносе операций доступа к данным ближе к источникам и потребителям данных операция записи с репликацией становится одним из ключевых факторов, влияющих на задержку обслуживания, объем межузлового обмена и использование коммуникационных ресурсов.

2) Разработана модель управления информационными потоками распределенной подсистемы хранения данных, в которой пользовательские операции записи и служебные операции поддержания репликации рассматриваются как взаимосвязанные потоки, формирующие нагрузку на узлы хранения и коммуникационные ресурсы. Модель учитывает очереди обслуживания, служебный межузловой обмен, сетевые ограничения и требование готовности выполнения операции записи.

3) Разработан метод оценки готовности выполнения операции записи с репликацией данных. Метод учитывает фактор репликации, порог подтверждения записи и доступность элементов цепочки обслуживания операции. На основе расчетов показано, что для исследуемой конфигурации переход от строгого режима подтверждения к пороговому снижает расчетную ожидаемую недоступность функции записи с 22,334 до 5,260 мин/год.

4) Разработан алгоритм интегрального выбора узлов хранения при выполнении операции записи с репликацией данных. Алгоритм основан на учете пользовательской очереди, служебной нагрузки, сетевой задержки и ограничения по готовности, что позволяет выбирать подмножество узлов хранения с учетом текущего состояния распределенной инфраструктуры.

5) Разработана имитационная модель операции записи с репликацией данных и проведена сопоставительная экспериментальная оценка стратегий распределения нагрузки. В качестве базовых вариантов рассмотрены равномерное циклическое распределение, выбор по минимальной пользовательской очереди, выбор с учетом пользовательской и служебной очередей, а также предложенный алгоритм интегрального выбора узлов.

6) По результатам экспериментального исследования установлено, что предложенный алгоритм повышает допустимую интенсивность входного потока операций записи на 48,3 % по сравнению с равномерным циклическим распределением при соблюдении заданного ограничения на время завершения

операции. Дополнительное преимущество относительно стратегии, учитывающей пользовательскую и служебную очереди, составляет 2,4 %.

7) Определена область применимости предложенного алгоритма интегрального выбора узлов. Наибольший эффект достигается в условиях неоднородности узлов хранения, различий сетевых задержек, накопления служебной нагрузки и наличия альтернативных узлов для размещения реплик. При однородной инфраструктуре и малой входной нагрузке эффект от применения интегрального выбора снижается.

Полученные результаты подтверждают, что совместный учет пользовательской нагрузки, служебного межузлового обмена, сетевой задержки и готовности выполнения операции записи позволяет повысить производительность распределенной подсистемы хранения данных без нарушения заданных эксплуатационных ограничений.

Результаты работы могут быть использованы при проектировании и настройке распределенных программно-определяемых подсистем хранения данных в телекоммуникационных системах, использующих инфраструктуру туманных вычислений.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в рецензируемых научных изданиях, включенных в перечень ВАК при Минобрнауки России

1. **Савельева А.А.** Метод оценки готовности распределенной подсистемы хранения уровня туманных вычислений к обслуживанию операций в телекоммуникационной сети // Труды учебных заведений связи. 2026. Т. 12. № 4. С. 27-39.
2. Елагин В.С., Редругина Н.М., **Савельева А.А.** Экспериментальная оценка алгоритма выбора узлов хранения при выполнении операции записи с репликацией данных в распределенной телекоммуникационной инфраструктуре // Электросвязь. 2026. № 7. С. 24-35.
3. Зарубин А.А., **Савельева А.А.**, Швидкий А.А. Моделирование нагрузки на распределенную систему хранения данных // Вестник связи. 2018. № 7. С. 13-16.
4. Зарубин А.А., Швидкий А.А., **Савельева А.А.** Особенности моделирования нагрузки на РСХД // Вестник связи. 2018. № 8. С. 7-11.

Публикации в изданиях, входящих в международные базы цитирования

5. Elagin V.S., Redrugina N.M., **Saveleva A.A.** The equivalence principle for adaptive load balancing in networks and systems // 2024 International Scientific and Technical Conference Modern Computer Network Technologies (MoNeTeC). 2024. С. 1-6.
6. Zarubin A.A., Shvidkiy A.A., **Savelieva A.A.** Approaches analysis to determining the characteristics of the load on the disk subsystem created by user applications // 2022 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO). 2022. С. 1-8
7. Zarubin A.A., Shvidkiy A.A., **Savelieva A.A.** Analysis of fault tolerance improvement in distributed over network data storage // 2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). 2021. С. 214-218.
8. Zarubin A.A., Shvidkiy A.A., **Savelieva A.A.** Caching methods analysis for improving distributed storage systems performance // 2021 Systems of Signal Synchronization, Generating and Processing in Telecommunications, (SYNCHROINFO). 2021. С. 9488328.
9. Shvidkiy A.A., Spirkina A.V., **Savelieva A.A.**, Tarlykov A.V. Evaluation of the impact the hyper-converged infrastructure storage subsystem synchronization on the overall performance // 2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). 2020. С. 248-252.
10. Zarubin A.A., Elagin V.S., Shvidkiy A.A., **Savelieva A.A.** Approaches to modeling the load on a distributed storage system // 2018 Systems of Signal Synchronization, Generating and Processing in Telecommunications, (SYNCHROINFO). 2018. С. 8456950.

Результаты интеллектуальной деятельности

11. Швидкий А.А., Христофоров Р.О., **Савельева А.А.** Система организации доступа к управлению данными гиперконвергентных центров обработки данных: свидетельство о государственной регистрации программы для ЭВМ № 2023615289, Российская Федерация. 2023.

12. Швидкий А.А., Зарубин А.А., **Савельева А.А.** Система сбора и обработки метрик производительности распределенной системы хранения данных: свидетельство о государственной регистрации программы для ЭВМ № 2022611763, Российская Федерация. 2022.

13. Швидкий А.А., Зарубин А.А., **Савельева А.А.** Система генерации нагрузки на распределенную систему хранения данных с сохранением комбинированных результатов нагрузочного тестирования: свидетельство о государственной регистрации программы для ЭВМ № 2022611764, Российская Федерация. 2022.

Публикации в других журналах и материалах научных конференций

14. Алексеева Н.Н., Редругина Н.М., **Савельева А.А.**, Тарабанов И.Ф. Математическая модель записи информации на узлы распределенной системы хранения данных // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2022): XI Международная научно-техническая и научно-методическая конференция, Санкт-Петербург. 2022. С. 65-70.

15. Зарубин А.А., Швидкий А.А., **Савельева А.А.** Анализ характеристик открытых систем хранения данных // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2022). XI Международная научно-техническая и научно-методическая конференция. Санкт-Петербург, 2022. С. 473-478.

16. Васюткин А.В., Власов Д.В., Швидкий А.А. **Савельева А.А.** Анализ методов отслеживания влияния нагрузки сетевой подсистемы гиперконвергентной инфраструктуры на производительность вычислительного кластера // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2021). Сборник научных статей: в 4-х томах. Санкт-Петербург, 2021. С. 380-384.

17. Казаков Д.Б., Тарлыков А.В., Швидкий А.А. **Савельева А.А.** Оценка производительности работы виртуальных рабочих столов на базе гиперконвергентной инфраструктуры // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2019). Сборник научных статей VIII Международной научно-технической и научно-методической конференции: в 4 томах. 2019. С. 170-174.

18. Зарубин А.А., Елагин В.С., Швидкий А.А., **Савельева А.А.** Подходы к моделированию нагрузки на распределенную систему хранения данных // Системы синхронизации, формирования и обработки сигналов. 2018. Т. 9. № 3. С. 90-95.

19. Долгун В.О., Швидкий А.А., Тарлыков А.В. **Савельева А.А.** Подходы к построению инфраструктуры виртуальных рабочих столов на базе гиперконвергентной инфраструктуры // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2019). сборник научных статей VIII Международной научно-технической и научно-методической конференции: в 4 томах. 2019. С. 158-163.

Отчеты о НИР

20-27. Научно-исследовательская работа по теме: «Разработка методов, моделей и средств повышения производительности и отказоустойчивости программно-определяемых СХД с использованием средств автоматизированного контроля и анализа результатов на базе нейронной сети» шифр «Мульти-ЦОД», государственный контракт с Минцифры России № П33-1-26/5 от «26» февраля 2021 г., период выполнения с 26.02.2021 по 31.12.2023 (8 отчетов согласно количеству этапов).