

ОПТИМИЗАЦИЯ РАСПРЕДЕЛЕНИЯ РЕСУРСОВ В СИСТЕМАХ ПЕРИФЕРИЙНЫХ ВЫЧИСЛЕНИЙ ДЛЯ ПРИЛОЖЕНИЙ РЕАЛЬНОГО ВРЕМЕНИ

Забелин Р.Т.

*бакалавр, Московский государственный технический
университет им. Н.Э. Баумана (Москва, Россия)*

Левшиц В.Э.

*бакалавр, Московский государственный технический
университет им. Н.Э. Баумана (Москва, Россия)*

OPTIMIZATION OF RESOURCE ALLOCATION IN EDGE COMPUTING SYSTEMS FOR REAL TIME APPLICATIONS

Zabelin R.T.

*bachelor's degree, Bauman Moscow state technical university
(Moscow, Russia)*

Levshits V.E.

*bachelor's degree, Bauman Moscow state technical university
(Moscow, Russia)*

Аннотация

Оптимизация распределения ресурсов в периферийных вычислительных системах требует учёта специфики приложений реального времени, включая жёсткие ограничения по задержке, устойчивость к перегрузкам и необходимость локальной обработки данных. Представлен анализ требований к архитектурам edge-сетей, классифицированы прикладные сценарии и сформулированы критерии эффективности распределения. Выделены алгоритмические подходы к назначению задач и проанализировано влияние сетевых параметров на стратегию управления. Показана значимость контекстно-адаптивных и гибридных методов, обеспечивающих баланс между производительностью, отказоустойчивостью и масштабируемостью систем edge-компьютинга.

Ключевые слова: периферийные вычисления, распределение ресурсов, приложения реального времени, edge-сети, адаптивные алгоритмы, задержка, балансировка нагрузки.

Abstract

Resource allocation in edge computing systems must account for the strict timing, reliability, and locality requirements of real-time applications. This study provides a structured analysis of edge architectures, application classes, and optimization criteria for task distribution. Several algorithmic strategies are outlined, including priority-based, delay-aware, and predictive methods, with particular attention to network dynamics and heterogeneity. Emphasis is placed on the role of adaptive and hybrid models that enhance resilience and performance in dynamic computing environments.

Keywords: edge computing, resource allocation, real-time applications, task scheduling, adaptive algorithms, latency, load balancing.

Введение

Системы периферийных вычислений (edge computing) становятся ключевым элементом современной распределённой цифровой инфраструктуры, особенно в приложениях, предъявляющих жёсткие требования к задержке, надёжности и локальной обработке данных. В отличие от традиционных облачных архитектур, edge-вычисления обеспечивают выполнение вычислительных задач вблизи источников данных - на уровне сенсоров, мобильных устройств и локальных шлюзов, что значительно снижает время отклика и уменьшает нагрузку на центральные узлы. Однако эффективное функционирование таких систем требует гибких и динамически адаптируемых механизмов управления вычислительными и сетевыми ресурсами.

Особенно остро проблема распределения ресурсов встает в приложениях реального времени, таких как автономное вождение, промышленная автоматизация, дистанционное медицинское наблюдение и тактильный интернет. Здесь даже незначительная задержка или перегрузка одного из узлов может привести к деградации качества обслуживания, нарушению безопасности или полной недоступности критически важной функции. В условиях высокой изменчивости нагрузки, разнообразия вычислительных профилей и ограниченности ресурсов на периферии традиционные алгоритмы управления становятся неэффективными или избыточно затратными.

Цель настоящего исследования - разработка и обоснование подходов к оптимизации распределения ресурсов в периферийных вычислительных системах с учётом требований реального времени. В рамках работы рассматриваются существующие архитектурные модели edge-обработки, классифицируются типовые сценарии приложений, анализируются критерии качества распределения ресурсов и предлагаются алгоритмические решения, направленные на достижение баланса между вычислительной эффективностью, задержкой и устойчивостью системы в условиях ограниченного контекста и высокой динамики среды.

Основная часть

Классификация приложений реального времени и особенности нагрузки на системы edge-вычислений

Оптимизация распределения ресурсов в системах периферийных вычислений невозможна без понимания характера задач, для которых эти системы предназначены. Приложения реального времени существенно различаются по интенсивности трафика, критичности к задержкам, устойчивости к перегрузкам и необходимому уровню обработки данных «на месте». Эти параметры напрямую влияют на стратегию планирования и миграции задач, выбор архитектурных шаблонов и алгоритмов управления ресурсами [1].

В условиях сетей следующего поколения (5G/6G), где периферийные вычисления используются для обслуживания гетерогенных и пространственно распределённых устройств, становится актуальной задача классификации прикладных сценариев по критериям нагрузки и требований к качеству обслуживания (QoS). Такой подход позволяет не только унифицировать принципы архитектурного проектирования, но и формализовать параметры, которые должны учитываться при динамическом распределении вычислений, хранения и пропускной способности на уровне edge-узлов [2].

Таблица 1 содержит обобщённую классификацию типовых приложений реального времени, их характеристики и ключевые требования к системам периферийных вычислений, необходимых для их эффективной поддержки.

Таблица 1

Характеристики приложений реального времени и требования к системам периферийных вычислений

Класс приложения	Характер нагрузки	Ключевые требования
Автономное вождение	Высокочастотная обработка потоков от сенсоров	Максимально допустимая задержка < 10 мс
Класс приложения	Характер нагрузки	Ключевые требования

Промышленный контроль	Циклические сигналы управления и телеметрия	Гарантированная предсказуемость и отказоустойчивость
Дистанционная хирургия	Низкая задержка, высокая точность	Стабильность отклика, медицинская точность
Интернет вещей в умных зданиях	Нерегулярные события, высокая вариативность	Энергоэффективность и адаптивность к событиям
Видеонаблюдение в реальном времени	Потоковое видео, требующее компрессии и предобработки	Обработка больших объёмов видео в реальном времени

Показанная классификация подчёркивает необходимость контекстно-зависимого подхода к распределению ресурсов. Вычислительная архитектура должна адаптироваться не только к объёму входящих задач, но и к специфике конкретного приложения, учитывая особенности сетевого взаимодействия, предсказуемость нагрузки и допустимые границы задержек. Это позволяет проектировать отказоустойчивые, энергоэффективные и масштабируемые решения для критически важных сценариев реального времени.

Критерии оптимизации распределения ресурсов в edge-среде

Для эффективного распределения ресурсов в системах периферийных вычислений необходимо учитывать множество факторов, зависящих как от внутренних параметров системы, так и от внешних условий эксплуатации. Универсальные подходы, ориентированные лишь на один критерий (например, минимизацию задержки), не всегда обеспечивают сбалансированную работу в условиях высокой нагрузки, энергозависимости и нестабильности каналов [3]. В этой связи возникает необходимость в формализации ключевых критериев оптимизации, которые могут использоваться при проектировании систем управления ресурсами.

Критерии выбора включают как классические метрики - задержка, пропускная способность, энергоэффективность, - так и более сложные показатели, такие как адаптивность к изменяющимся условиям и устойчивость к отказам. Их значимость может существенно различаться в зависимости от типа приложения: для телемедицинской диагностики первична надёжность, для систем реального времени важна задержка, для IoT-контроллеров - энергопотребление. Это обуславливает необходимость контекстно-зависимого мультикритериального подхода к управлению ресурсами на периферии.

Таблица 2 систематизирует основные критерии оптимизации, их описание и применимость в различных классах приложений реального времени.

Таблица 2

Критерии оптимизации распределения ресурсов в edge-среде

Критерий оптимизации	Описание	Применимость к приложениям
Минимизация задержки	Сокращение времени от поступления задачи до завершения её обработки	Критически важные системы управления, AR/VR
Максимизация пропускной способности	Эффективное использование каналов связи и узлов передачи данных	Видеонаблюдение, потоковые данные
Снижение энергопотребления	Продление времени автономной работы устройств	Устройства IoT, умные здания
Обеспечение отказоустойчивости	Поддержание работоспособности при сбоях узлов и каналов	Промышленные сети, автономные машины
Балансировка вычислительной нагрузки	Равномерное распределение задач между узлами по мощности и загрузке	Сценарии с переменной интенсивностью нагрузки
Критерий оптимизации	Описание	Применимость к приложениям

Адаптация к динамическим условиям	Корректировка стратегии при изменении характеристик среды	Все классы приложений в изменяющейся среде
-----------------------------------	---	--

Представленные критерии служат основой для построения гибких и адаптивных алгоритмов управления ресурсами. В условиях высокой гетерогенности приложений и ограничений на вычислительные мощности на периферии, выбор критериев оптимизации должен быть обусловлен прикладной задачей, профилем нагрузки и требуемым уровнем отказоустойчивости [4]. Эффективные стратегии должны учитывать динамику среды, возможность приоритизации задач и необходимость поддержки согласованности между узлами.

Алгоритмы распределения задач в edge-средах и их характеристика

В системах периферийных вычислений важнейшей задачей является выбор подходящего алгоритма распределения задач между узлами, обладающими разной вычислительной мощностью, доступом к сети и ограничениями по энергии. От эффективности стратегии назначения задач зависит соблюдение временных ограничений, равномерность загрузки компонентов и устойчивость к изменениям среды [5]. Выбор конкретного алгоритма определяется профилем нагрузки, классом приложения и доступностью системных метрик в момент принятия решений.

На практике применяются как простые эвристические методы, так и сложные алгоритмы, основанные на машинном обучении и прогнозировании поведения среды [6]. Некоторые алгоритмы ориентированы на минимизацию задержек, другие - на распределение нагрузки, третьи - на адаптацию к изменениям и обучение на исторических данных. Важно, чтобы выбранный подход обеспечивал не только производительность, но и предсказуемость выполнения задач, особенно в критически важных приложениях.

Таблица 3 обобщает наиболее распространённые алгоритмы распределения задач в edge-средах, описывает их принципы работы и ключевые преимущества.

Таблица 3

Алгоритмы распределения задач в edge-средах и их характеристики

Алгоритм распределения	Принцип работы	Преимущества
На основе приоритетов (static priority)	Задачи классифицируются по фиксированному приоритету; высокие приоритеты обслуживаются первыми	Простота реализации, предсказуемость выполнения
Распределение по кругу (round-robin)	Задачи распределяются последовательно по доступным узлам без учёта состояния	Равномерность в условиях однородных задач
Минимизация задержки (delay-aware scheduling)	Учитываются задержки между узлами и выбирается оптимальный маршрут	Подходит для систем с ограничениями по времени отклика
Балансировка нагрузки (load balancing)	Распределение задач осуществляется на основе текущей загрузки узлов	Снижает перегрузки, повышает общую производительность
Распределение с предсказанием (predictive allocation)	Прогнозируется будущее состояние нагрузки и ресурсов	Обеспечивает упреждающее распределение, адаптивность к изменениям
Обучаемое распределение (RL-based scheduling)	Используются алгоритмы обучения с подкреплением для динамического выбора узлов	Гибкая адаптация, обучение на опыте, высокая эффективность в нестабильной среде

Выбор алгоритма распределения должен опираться на конкретные цели системы: минимизацию времени отклика, стабильную производительность или адаптацию к условиям. В приложениях реального времени наилучшие результаты часто достигаются при

гибридизации подходов - например, сочетании эвристики и обучаемых моделей, что позволяет достичь компромисса между предсказуемостью, скоростью и адаптивностью [7].

Контекстно-адаптивные подходы к управлению ресурсами в гетерогенных edge-сетях

Современные edge-системы представляют собой высоко гетерогенные вычислительные среды, включающие устройства с различными уровнями производительности, доступом к источникам питания, параметрами сетевого соединения и архитектурными ограничениями. В такой среде статические или универсальные подходы к управлению ресурсами оказываются неэффективными и нередко приводят к локальным перегрузкам, непредсказуемым задержкам или неоптимальному использованию каналов связи [8]. Это особенно критично в приложениях реального времени, где любые отклонения от допустимого окна выполнения могут привести к потере функциональности.

Контекстно-адаптивное управление предполагает, что стратегия распределения ресурсов должна учитывать не только внутренние параметры вычислительных узлов, но и динамику окружающей среды, поведение пользователей и специфику выполняемых задач. Это включает мониторинг текущей загрузки, состояния сети, типов поступающих данных и их критичности [9]. Использование таких параметров позволяет выстраивать более точные и эффективные алгоритмы, которые способны корректировать свои действия в режиме реального времени без вмешательства оператора.

Одним из ключевых компонентов таких подходов является использование моделей предиктивной аналитики и онтологического описания сценариев выполнения задач. Это позволяет формировать не просто реактивную, а проактивную стратегию управления ресурсами, заранее прогнозируя узкие места и адаптируя инфраструктуру под изменяющиеся условия. Дополнительно, внедрение правил приоритизации на основе профилей приложений и уровня важности задач обеспечивает гарантированное обслуживание критически значимых функций даже в условиях перегрузки системы.

Контекстная адаптация также требует распределённого согласования между узлами edge-сети. Это означает, что локальные решения должны быть скоординированы с глобальной политикой, чтобы избежать конфликтов, дублирования вычислений или деградации глобального качества обслуживания. Для реализации такой координации могут применяться децентрализованные алгоритмы консенсуса, синхронизированные политики принятия решений или доверенные агенты, контролирующие состояние соседних узлов.

Таким образом, контекстно-адаптивные подходы представляют собой важнейший вектор развития edge-инфраструктуры [10]. Их применение позволяет обеспечить гибкость, предсказуемость и устойчивость систем в условиях высокой динамичности, ограниченных ресурсов и требований к гарантированному качеству обслуживания, характерных для приложений реального времени.

Влияние параметров сети на стратегию распределения ресурсов

В периферийных вычислительных системах, работающих в реальном времени, качество и стабильность сетевых соединений играют критически важную роль. В отличие от облачных решений, где взаимодействие с инфраструктурой централизовано и относительно предсказуемо, edge-архитектуры характеризуются высокой степенью сетевой фрагментации, наличием беспроводных каналов различной надёжности и постоянными изменениями топологии. Эти факторы существенно усложняют задачу выбора оптимальных путей передачи, миграции задач и синхронизации узлов.

Одним из ключевых параметров, влияющих на распределение ресурсов, является задержка передачи данных. При планировании задач необходимо учитывать не только вычислительные характеристики узлов, но и латентность каналов, через которые происходит передача между источником, вычислительным узлом и конечной точкой. Даже высокопроизводительный узел может оказаться неэффективным, если связан с источником медленным каналом с нестабильной пропускной способностью.

Вторым важным параметром является ширина канала и уровень его загрузки. Периферийные устройства могут использовать каналы общего назначения, подверженные конкуренции за пропускную способность, что делает невозможной надёжную доставку данных при высокой плотности трафика [11]. Для таких случаев необходимо реализовывать адаптивные алгоритмы маршрутизации и перегруппировки задач, способные в реальном времени выбирать не только наименее загруженные узлы, но и узлы, обеспечивающие наилучшую сетевую связность.

Также критичным аспектом является надёжность соединений, особенно в условиях подвижных узлов или нестабильной радиосреды. В таких сценариях ресурсы должны распределяться с учётом возможных обрывов соединения, временной недоступности и необходимости локального буферизации данных. Это требует разработки стратегий с элементами предсказания и резервирования, позволяющих избегать потери данных и поддерживать непрерывность функционирования даже при частичном отказе сети.

Наконец, важным становится учёт географического положения узлов и плотности устройств, особенно в задачах, где критичен физический контекст (например, логистика, транспорт, AR/VR). В таких случаях следует внедрять геопространственные критерии в систему принятия решений, придавая приоритет узлам, находящимся ближе к источникам генерации данных и действия.

Таким образом, параметры сети являются неотъемлемой частью модели распределения ресурсов и требуют включения в оптимизационные алгоритмы наравне с вычислительными и энергетическими характеристиками. Только при комплексном учёте сетевых факторов возможно построение устойчивой, масштабируемой и адаптивной периферийной вычислительной системы, способной стабильно функционировать в реальном времени.

Заключение

В условиях стремительного развития приложений реального времени и повышения требований к скорости отклика, устойчивости и локальной обработке данных, системы периферийных вычислений становятся ключевым компонентом современной вычислительной архитектуры. Однако их высокая гетерогенность, ограниченность ресурсов и нестабильность сетевого окружения предъявляют особые требования к механизмам управления распределением задач и ресурсов.

В рамках настоящей работы проведён структурный анализ основных классов приложений реального времени, обоснованы критерии оптимизации распределения ресурсов, рассмотрены алгоритмические подходы к управлению нагрузкой, а также выделены факторы, влияющие на эффективность в условиях динамичной и неопределённой среды. Показано, что для достижения требуемого уровня качества обслуживания необходимо использовать гибридные, контекстно-адаптивные методы, включающие как эвристики, так и алгоритмы машинного обучения, работающие в тесной связке с мониторингом текущих параметров системы.

Оптимизация распределения ресурсов в edge-сетях требует комплексного подхода, включающего учёт вычислительных, энергетических и сетевых параметров, а также адаптацию к профилю приложений и изменяющимся условиям эксплуатации. Дальнейшие исследования могут быть направлены на разработку предиктивных моделей и самообучающихся систем, способных в реальном времени реагировать на изменения нагрузки и архитектурные перестройки инфраструктуры.

Список литературы

1. Куприяновский В.П., Намиот Д.Е., Покусаев О.Н. Периферийные вычисления для современной мобильности // International Journal of Open Information Technologies. 2025. Т. 13. № 2. С. 94-104.
2. Бердиназарова А., Кулиев Э., Акыев А. Вычисления на периферии: роль edge computing в современных сетях // Вестник науки. 2024. Т. 3. № 9 (78). С. 342-345.

3. Liu J., Li C., Luo Y. Efficient resource allocation for IoT applications in mobile edge computing via dynamic request scheduling optimization // Expert Systems with Applications. 2024. Vol. 255. P. 124716.
4. Alfakih T., Hassan M.M., Al-Razgan M. Multi-objective accelerated particle swarm optimization with dynamic programming technique for resource allocation in mobile edge computing // IEEE Access. 2021. Vol. 9. P. 167503-167520.
5. Haibeh L.A., Yagoub M.C.E., Jarray A. A survey on mobile edge computing infrastructure: Design, resource management, and optimization approaches // IEEE Access. 2022. Vol. 10. P. 27591-27610.
6. Baburao D., Pavankumar T., Prabhu C.S.R. Load balancing in the fog nodes using particle swarm optimization-based enhanced dynamic resource allocation method // Applied Nanoscience. 2023. Vol. 13. No. 2. P. 1045-1054.
7. Abouaomar A., Cherkaoui S., Mlika Z., Kobbane A. Resource provisioning in edge computing for latency-sensitive applications // IEEE Internet of Things Journal. 2021. Vol. 8. No. 14. P. 11088-11099.
8. Bahreini T., Badri H., Grosu D. Mechanisms for resource allocation and pricing in mobile edge computing systems // IEEE Transactions on Parallel and Distributed Systems. 2021. Vol. 33. No. 3. P. 667-682.
9. Smirnov A. Monitoring and logging in distributed systems: application of OpenTelemetry and the ELK stack // Universum: technical sciences : electronic scientific journal. 2025. No. 3(132). P. 30-33.
10. Djigal H., Xu J., Liu L., Zhang Y. Machine and deep learning for resource allocation in multi-access edge computing: A survey // IEEE Communications Surveys & Tutorials. 2022. Vol. 24. No. 4. P. 2449-2494.
11. Tuli S., Mirhakimi F., Pallewatta S., Zawad S., Casale G., Javadi B., Jennings N.R. AI augmented Edge and Fog computing: Trends and challenges // Journal of Network and Computer Applications. 2023. Vol. 216. P. 103648.