Хамитов И.А. Многокритериальная модель выбора структуры GPU-вычислительной системы для задач искусственного интеллекта // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 6 (68) Ч.1 с. 57–61

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/



УДК 004.272.2:004.8

МНОГОКРИТЕРИАЛЬНАЯ МОДЕЛЬ ВЫБОРА СТРУКТУРЫ GPU-ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ ДЛЯ ЗАДАЧ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Хамитов И.А.
ФГБОУ ВО «МИРЭА - РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ», Москва, Россия (119454, г. Москва, пр-т Вернадского, д. 78, стр. 4), e-mail: ilnur_khamitov02@mail.ru

В статье рассматривается задача выбора структуры вычислительной системы на базе графических процессоров для прикладных задач искусственного интеллекта. Обосновывается, что экспертное принятие архитектурных решений и сравнение оборудования только по пиковой производительности не позволяют надежно оценить будущую эффективность системы в условиях крупной AI-инфраструктуры. Предлагается многокритериальная модель ранжирования GPU-конфигураций, учитывающая производительность, стоимость, энергопотребление, объем памяти, масштабируемость, задержку обработки и прикладные метрики LLM-нагрузок.

Ключевые слова**:** Kubernetes, GPU, вычислительная система, искусственный интеллект, LLM, энергоэффективность, многокритериальная оценка, дата-центр, стоимость владения.

MULTI-CRITERIA MODEL FOR SELECTING THE STRUCTURE OF A GPU-BASED COMPUTING SYSTEM FOR ARTIFICIAL INTELLIGENCE TASKS

Khamitov I.A.
MIREA - RUSSIAN TECHNOLOGICAL UNIVERSITY, Moscow, Russia (119454, Moscow, avenue Vernadsky, 78, b. 4), e-mail:

The article addresses the problem of selecting the structure of GPU-based computing systems for artificial intelligence workloads. It shows that expert-based decisions and equipment comparison by peak performance alone are insufficient for estimating practical efficiency of large AI infrastructures. A multi-criteria model is proposed for ranking GPU configurations using performance, cost, power consumption, memory capacity, scalability, latency and LLM-oriented workload metrics.

Keywords: GPU, computing system, artificial intelligence, LLM, energy efficiency, multi-criteria evaluation, data center, total cost of ownership.

Введение

Рост практического применения искусственного интеллекта привел к тому, что вычислительная инфраструктура стала одним из ключевых факторов внедрения AI-сервисов. Обучение моделей, инференс больших языковых моделей, мультимодальная обработка данных и корпоративные интеллектуальные помощники требуют специализированных серверов с графическими ускорителями. При этом выбор такой инфраструктуры связан не только с достижением требуемой скорости вычислений, но и с ограничениями по бюджету, энергопотреблению, охлаждению и последующей эксплуатации.

Значимость энергетического фактора усиливается по мере роста дата-центров. Международное энергетическое агентство прогнозирует, что мировое потребление

электроэнергии дата-центрами может увеличиться примерно с 485 ТВт·ч в 2025 году до 950 ТВт·ч к 2030 году, а потребление центров, ориентированных на AI, будет расти еще быстрее [1]. В этих условиях ошибка при выборе структуры вычислительной системы может стать не только технической, но и экономической проблемой: проект получает завышенные расходы на питание и охлаждение, а расчетная окупаемость ухудшается.

В инженерной практике архитектура GPU-системы нередко определяется экспертно: по опыту специалистов, доступности оборудования, рекомендациям поставщика или сравнению отдельных паспортных характеристик. Такой подход удобен при первичном отборе решений, но он плохо формализует компромиссы между производительностью, стоимостью, энергией, памятью и качеством обслуживания запросов. Например, сервер с большей теоретической производительностью может уступить более сбалансированной конфигурации при реальной LLM-нагрузке, если обладает меньшим объемом памяти, высокой задержкой или слабым межпроцессорным обменом.

Целью работы является разработка подхода к выбору структуры GPU-вычислительной системы, при котором альтернативные конфигурации оцениваются не по одному показателю, а по совокупности взаимосвязанных критериев. Такой подход позволяет сделать выбор более прозрачным и снизить риск неэффективных капитальных и эксплуатационных затрат.

Исследование

Проблема экспертного выбора архитектуры

Проектирование вычислительной системы для задач искусственного интеллекта предполагает выбор типа GPU, количества ускорителей, процессорной платформы, объема оперативной и видеопамяти, межсоединения, сетевой инфраструктуры, системы охлаждения и параметров электропитания. Эти элементы образуют единую структуру, поэтому изменение одного параметра влияет на остальные. Увеличение числа GPU повышает потенциальную производительность, но одновременно увеличивает стоимость, энергопотребление и требования к межпроцессорному обмену.

Однокритериальный выбор не отражает этой взаимосвязи. Сравнение по FLOPS или Tensor TFLOPS показывает только верхнюю границу вычислительных возможностей и не описывает, насколько эффективно конкретная нагрузка сможет использовать оборудование. Разработчики TOP500 прямо отмечают, что результат LINPACK не является универсальной характеристикой системы, поскольку один числовой показатель не способен описать все стороны производительности [2]. Для задач машинного обучения более прикладную оценку дают MLPerf-бенчмарки, однако даже они не заменяют технико-экономического выбора конкретной инфраструктуры под заданные ограничения [3].

Следовательно, при выборе GPU-системы важно перейти от экспертного сравнения отдельных характеристик к модели, которая связывает требования задачи с параметрами архитектуры и позволяет ранжировать несколько альтернатив.

Набор критериев для оценки конфигурации

Каждая рассматриваемая конфигурация может быть описана набором технических и экономических параметров: достижимая производительность, среднее энергопотребление, стоимость закупки, объем GPU-памяти, коэффициент масштабируемости, задержка ответа и

скорость генерации токенов. Для сопоставления альтернатив предлагается рассчитывать нормированные частные критерии, представленные в Таблице 1.

Таблица 1 – Частные критерии оценки GPU-конфигурации

Критерий	**Смысл показателя**	**Роль при выборе**
Производительность	Насколько система покрывает требуемый объем вычислений	Позволяет исключить недостаточно мощные решения
Энергоэффективность	Отношение полезной производительности к потребляемой мощности	Снижает эксплуатационные расходы и нагрузку на инженерную инфраструктуру
Стоимость	Закупочная или расчетная стоимость конфигурации	Влияет на доступность проекта и срок окупаемости
Память	Достаточность GPU-памяти для модели, кэша и контекста	Критична для LLM, RAG и мультимодальных задач
Масштабируемость	Сохранение прироста производительности при добавлении GPU	Важна для кластеров и распределенной обработки
Задержка	Время получения ответа на пользовательский запрос	Определяет качество интерактивных AI-сервисов
LLM-эффективность	Скорость генерации токенов и интенсивность обслуживания запросов	Описывает практическую пригодность системы для генеративного ИИ

Нормирование необходимо потому, что исходные параметры измеряются в разных единицах: ваттах, рублях, миллисекундах, гигабайтах, токенах в секунду. После приведения критериев к сопоставимому виду можно рассчитать интегральный показатель пригодности конфигурации для заданной нагрузки.

$I = w_1K_p + w_2K_e + w_3Kc + w_4Km + w_5Ks + w_6Kl + w_7Kllm,$

где K_p — критерий производительности; K_e — критерий энергоэффективности; Kc — критерий стоимости; Km — критерий обеспеченности памятью; Ks — критерий масштабируемости; Kl — критерий задержки; Kllm — критерий эффективности LLM-нагрузки; $w_1 \ldots w_7$ — весовые коэффициенты, сумма которых равна единице.

Идея модели состоит в том, что лучшей считается не обязательно самая дорогая или самая производительная система, а та, которая наиболее полно соответствует прикладному сценарию. Для интерактивного инференса увеличивается значимость задержки и скорости генерации; для пакетной обработки — пропускной способности, энергоэффективности и стоимости; для длинного контекста — объема видеопамяти и устойчивости при росте входных данных.

Алгоритм применения модели

Применение модели включает пять этапов. На первом этапе задаются ограничения: минимальная производительность, максимальный бюджет, допустимая мощность, требуемый объем GPU-памяти, предельная задержка и тип нагрузки. На втором этапе формируется перечень альтернативных конфигураций, которые потенциально могут быть использованы в проекте.

На третьем этапе выполняется предварительная фильтрация. Если конфигурация не удовлетворяет жестким условиям, например не имеет достаточного объема видеопамяти для размещения модели, она исключается до расчета итогового показателя. Такой шаг предотвращает попадание в ранжирование формально дешевых, но непригодных решений.

На четвертом этапе для оставшихся вариантов рассчитываются частные критерии. Для вычислительных показателей могут использоваться паспортные характеристики, результаты бенчмарков, эмпирические коэффициенты использования, а также прикладные метрики: количество токенов в секунду, средняя задержка и расчетная интенсивность обслуживания запросов. На пятом этапе рассчитывается интегральная оценка, после чего конфигурации сортируются по убыванию итогового значения.

Преимущество такого алгоритма заключается в объяснимости результата. Итоговый выбор можно обосновать не общей фразой о большей мощности оборудования, а конкретными причинами: достаточной памятью, лучшим соотношением производительности и мощности, меньшей стоимостью обработки запроса или более высокой устойчивостью к масштабированию.

Практическая значимость

Предложенная модель может использоваться при подготовке технико-экономического обоснования закупки GPU-серверов, модернизации корпоративной AI-платформы, сравнении предложений поставщиков и проектировании дата-центра. Особенно важным является учет энергопотребления, поскольку эксплуатационные затраты становятся значимой частью жизненного цикла вычислительной системы.

Модель также снижает риск неверного масштабирования. Недостаточная конфигурация приводит к невыполнению требований по качеству сервиса и необходимости дополнительной закупки. Избыточная конфигурация создает лишнюю капитальную нагрузку, увеличивает потребление электроэнергии и может сделать проект экономически непривлекательным еще до запуска.

Заключение

В статье рассмотрена проблема выбора структуры GPU-вычислительной системы для задач искусственного интеллекта. Показано, что экспертный подход и сравнение по пиковой производительности не обеспечивают достаточной обоснованности при проектировании крупной AI-инфраструктуры.

Предложена многокритериальная модель, в которой альтернативные конфигурации оцениваются по совокупности производительных, энергетических, стоимостных и прикладных показателей. В отличие от однокритериального выбора модель позволяет учитывать специфику нагрузки: интерактивный инференс, пакетную обработку и задачи с увеличенным контекстом.

Практическая ценность подхода заключается в возможности заранее оценить компромиссы между стоимостью, энергопотреблением, памятью и качеством обслуживания. Это повышает прозрачность архитектурных решений, снижает риск избыточных расходов и позволяет выбирать вычислительную систему с учетом реальных условий эксплуатации.

Список литературы

1. Международное энергетическое агентство. Ключевые вопросы энергетики и искусственного интеллекта: Краткое содержание [Электронный ресурс]. — URL: https://www.iea.org/reports/key-questions-on-energy-and-ai/executive-summary (дата обращения: 10.05.2026).
2. ТОП500. Тест LINPACK [Электронный ресурс]. — URL: https://top500.org/project/linpack/ (дата обращения: 10.05.2026).
3. МЛКоммонс. MLPerf Inference: Датацентр [Электронный ресурс]. — URL: https://mlcommons.org/benchmarks/inference-datacenter/ (дата обращения: 10.05.2026).
4. Уильямс С., Уотерман А., Паттерсон Д. Roofline: содержательная визуальная модель производительности для многоядерных архитектур // Communications of the ACM. — 2009. — Том. 52. — № 4. — С. 65–76.
5. Хеннесси Дж. Л., Паттерсон Д. А. Компьютерная архитектура: количественный подход. - 6-е изд. - Кембридж: Морган Кауфманн, 2019.
6. Саати Т. Л. Процесс аналитической иерархии. - Нью-Йорк: МакГроу-Хилл, 1980.
7. Грин500. Списки Green500 [Электронный ресурс]. — URL: https://top500.org/lists/green500/ (дата обращения: 14.05.2026).
8. МЛКоммонс. MLPerf Training Benchmark [Электронный ресурс]. — URL: https://mlcommons.org/benchmarks/training/ (дата обращения: 15.05.2026).

References

1. International Energy Agency. Key Questions on Energy and AI: Executive summary [Электронный ресурс]. — URL: https://www.iea.org/reports/key-questions-on-energy-and-ai/executive-summary (дата обращения: 10.05.2026).
2. TOP500. The LINPACK Benchmark [Электронный ресурс]. — URL: https://top500.org/project/linpack/ (дата обращения: 10.05.2026).
3. MLCommons. MLPerf Inference: Datacenter [Электронный ресурс]. — URL: https://mlcommons.org/benchmarks/inference-datacenter/ (дата обращения: 10.05.2026).
4. Williams S., Waterman A., Patterson D. Roofline: an insightful visual performance model for multicore architectures // Communications of the ACM. — 2009. — Vol. 52. — № 4. — P. 65–76.
5. Hennessy J. L., Patterson D. A. Computer Architecture: A Quantitative Approach. — 6th ed. — Cambridge: Morgan Kaufmann, 2019.
6. Saaty T. L. The Analytic Hierarchy Process. — New York: McGraw-Hill, 1980.
7. Green500. Green500 Lists [Электронный ресурс]. — URL: https://top500.org/lists/green500/ (дата обращения: 14.05.2026).
8. MLCommons. MLPerf Training Benchmark [Электронный ресурс]. — URL: https://mlcommons.org/benchmarks/training/ (дата обращения: 15.05.2026).