

ЭНЕРГОПОТРЕБЛЕНИЕ В AI-ДАТА-ЦЕНТРАХ:

ПРИЧИНЫ, ВЫЗОВЫ И РЕШЕНИЯ

Инфраструктура, ориентированная на искусственный интеллект, радикально изменила профиль энергопотребления дата-центров. Если традиционные IT-сервисы — виртуализация, базы данных или корпоративные приложения — использовали процессоры общего назначения и имели умеренные тепловые нагрузки, то обучение и инференс больших AI-моделей потребляют энергию на порядки больше. Это формирует новые требования к электропитанию, охлаждению и общей архитектуре центров обработки данных.

РОСТ ЭНЕРГОПОТРЕБЛЕНИЯ: ОТ КЛАССИЧЕСКИХ ЦОД К AI-КЛАСТЕРАМ



В классическом дата-центре один сервер потребляет 500–800 Вт, а плотность стойки обычно не превышает 10–15 кВт.

В AI-инфраструктуре один вычислительный узел с GPU-ускорителями требует уже 2,5–3,5 кВт, а плотность стойки достигает 50–100 кВт.

В пересчёте на выполняемые операции энергопотребление может быть выше в 300–1000 раз, особенно при обучении крупных языковых моделей (LLM).

Это не означает, что весь дата-центр стал «в тысячу раз прожорливее» — речь идёт о конкретных задачах, где миллионы тензорных операций требуют колоссальных вычислительных и энергетических ресурсов.

ГДЕ РАСХОДУЕТСЯ ЭНЕРГИЯ: ОБУЧЕНИЕ И ИНФЕРЕНС

Основная часть энергии уходит на **обучение моделей**. Этот процесс включает непрерывную работу десятков или сотен GPU, синхронизацию через высокоскоростные сети и использование энергоёмкой памяти. Тренировка крупной модели с триллионами параметров может занимать недели и потреблять мегаватты.

Инференс — применение модели — сам по себе менее энергоёмкий, но он выполняется постоянно. Массовые сценарии (чат-боты, генерация контента, поиск, рекомендации) создают стабильную нагрузку, которая превращает инференс в основной источник постоянного энергопотребления.



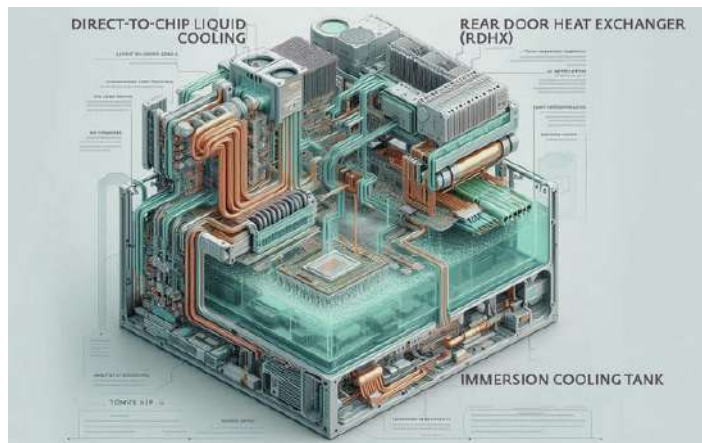
ПРИМЕР ЭНЕРГОПРОФИЛЯ AI-СЕРВЕРА

Типичный AI-узел мощностью ~3 кВт содержит:

КОМПОНЕНТ	КОНФИГУРАЦИЯ	ПОТРЕБЛЕНИЕ
ПРОЦЕССОРЫ	2x AMD EPYC GENOA ИЛИ INTEL XEON SAPPHIRE RAPIDS	~400–500 Вт
ПАМЯТЬ	512 ГБ – 1 ТБ DDR5 ИЛИ HBM	~150–200 Вт
GPU	8x NVIDIA A100 ИЛИ H100 (ПО 400–700 Вт КАЖДЫЙ)	2,4–3,2 кВт
СЕТЬ	INFINIBAND NDR / 200–400 GBE	30–60 Вт
НАКОПИТЕЛИ	NVME SSD (4–8 x 15–20 Вт)	~100–150 Вт
ПЛАТА, PSU, ВМС, ВЕНТИЛЯТОРЫ	---	~200 Вт

Итого: 2,8–3,5 кВт на один узел. Такой профиль характерен для систем уровня NVIDIA DGX H100, Dell XE9680, HPE Cray XD670 и других HPC-платформ.

ИНЖЕНЕРНАЯ ИНФРАСТРУКТУРА И ОХЛАЖДЕНИЕ



Рост плотности тепловой нагрузки меняет сам подход к проектированию дата-центров. Воздушное охлаждение эффективно лишь до 30–40 кВт/стойку.

При большей плотности применяются:

- **Direct-to-Chip** — жидкость подводится напрямую к кристаллу;
- **RDHX (Rear Door Heat Exchanger)** — теплообменник задней двери;
- **Иммерсионное охлаждение** — полное или частичное погружение оборудования в диэлектрическую жидкость.

Стойки с мощностью 50–100 кВт требуют отдельного электропитания и теплоотвода, а также усиленного резервирования (ИБП, распределение фаз, телеметрия и управление нагрузками).

ЭНЕРГОЭФФЕКТИВНОСТЬ И ИСТОЧНИКИ ПИТАНИЯ

Даже с ростом производительности процессоров и GPU (увеличением операций на ватт), суммарное энергопотребление AI-инфраструктуры продолжает расти. Поэтому внимание переносится на эффективность систем в целом.

Внедряются следующие подходы:

- Использование **возобновляемых источников энергии** — солнечных, ветровых, гидро.
- **PPA-контракты** (Power Purchase Agreements) на долгосрочные поставки зелёной энергии.
- **Локальные хранилища энергии** (батареи и тепловые буферы).
- **Интеллектуальное управление нагрузкой** — динамическое перераспределение задач между кластерами.

Ключевыми метриками становятся PUE (Power Usage Effectiveness), углеродный след (Carbon Footprint) и степень повторного использования тепла (например, для отопления зданий или промышленных зон).



ЭКОНОМИКА И УСТОЙЧИВОСТЬ

Энергопотребление — главный фактор, влияющий на OPEX (операционные расходы) дата-центра. Стоимость энергии и охлаждения в ближайшие годы может сравняться со стоимостью оборудования. Поэтому эффективность сегодня измеряется не только в TFLOPS, но и во FLOPS на ватт.

Корпорации и регуляторы требуют отчётности по выбросам CO₂ и перехода к экологически устойчивым решениям. Растёт интерес к рекуперации тепла, системам district heating, водяным контурам и вторичному использованию охлаждающих жидкостей.



ЗАКЛЮЧЕНИЕ: НОВАЯ ЭНЕРГЕТИКА ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

AI-дата-центры формируют новый класс вычислительной инфраструктуры, где энергопотребление становится центральным параметром проектирования.

Высокоплотные серверы, GPU-ускорители, продвинутые системы охлаждения и использование возобновляемых источников энергии — ключевые направления развития отрасли.

Чтобы обеспечить устойчивость и экономическую эффективность, операторам необходимо:

- оптимизировать архитектуру обучения и инференса;
- внедрять энергоэффективные узлы и распределённые системы охлаждения;
- использовать возобновляемую энергетику и технологии рекуперации тепла.

Понимание энергетического профиля AI-нагрузок становится основой для планирования дата-центров следующего поколения — умных, устойчивых и энергетически сбалансированных.

