

АППАРАТНОЕ ЯДРО AI:

GPU, TPU, ASIC И
АРХИТЕКТУРА ЧИПЛЕТОВ

В предыдущих статьях мы выяснили, почему традиционные CPU не справляются с AI и какую колоссальную энергию потребляют новые системы. Теперь посмотрим, какие именно аппаратные архитектуры пришли им на смену



Аппаратная революция последних десяти лет радикально изменила путь развития искусственного интеллекта. Традиционные универсальные CPU перестали справляться с углубляющимися требованиями AI—особенно в обучении больших моделей. На смену пришли специализированные ускорители: **GPU, TPU, ASIC**, а также новая парадигма проектирования — **чиплетная архитектура**.

В этой статье мы разберём:

- почему GPU стали доминирующей силой в AI;
- что делают тензорные ядра;
- чем TPU и ASIC отличаются от GPU;
- почему чиплеты — фундамент будущих вычислений.

ДОМИНИРОВАНИЕ GPU: АРХИТЕКТУРНЫЕ ПРИЧИНЫ

Графические процессоры (GPU) не были созданы для AI. Они эволюционировали для рендеринга компьютерной графики, что по своей природе является массово-параллельной задачей. Именно эта архитектурная особенность сделала их идеальными для глубокого обучения.

ПАРАМЕТР	CPU (CENTRAL PROCESSING UNIT)	GPU (GRAPHICS PROCESSING UNIT)
АРХИТЕКТУРА	ПОСЛЕДОВАТЕЛЬНАЯ (ФОН НЕЙМАНА)	ПАРАЛЛЕЛЬНАЯ (SIMD/SIMT)
КОЛИЧЕСТВО ЯДЕР	ДЕСЯТКИ — СОТНИ	СОТНИ — ТЫСЯЧИ
ОСНОВНОЙ ФОКУС	УПРАВЛЕНИЕ, ЛОГИКА, КЭШ	ЧИСТЫЕ ВЫЧИСЛЕНИЯ (ALU)
ОПТИМИЗАЦИЯ	НИЗКАЯ ЗАДЕРЖКА (LATENCY)	ВЫСОКАЯ ПРОПУСКНАЯ СПОСОБНОСТЬ (THROUGHPUT)
ТИП ЗАДАЧ	ПОСЛЕДОВАТЕЛЬНЫЕ	МАССОВО-ПАРАЛЛЕЛЬНЫЕ



Основная операция в нейронных сетях — **перемножение матриц**, разбиваемое на тысячи FMA-операций. GPU, благодаря тысячам одинаковых ALU-ядер, обеспечивают огромный параллелизм и высокую производительность.

ТЕНЗОРНЫЕ ЯДРА: СЕКРЕТНОЕ ОРУЖИЕ УСКОРИТЕЛЕЙ

Следующим эволюционным скачком стало появление тензорных ядер (**Tensor Cores**). Если обычные вычислительные ядра (CUDA Cores) — это универсальные солдаты, то тензорные ядра — это элитные спецназовцы, заточенные под одну, но критически важную операцию.

Что такое тензорные ядро?

Это специализированные блоки внутри современных GPU (начиная с архитектуры NVIDIA Volta), которые выполняют матричные операции с пониженной точностью (FP16, BF16, INT8, INT4) с огромной скоростью.

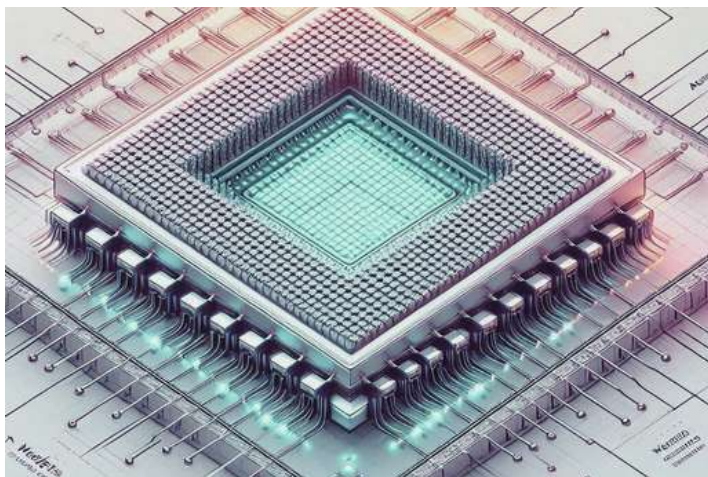
- **Обычное ядро (CUDA Core):** Выполняет 1 операцию FMA за такт.
- **Тензорное ядро:** Оптимизировано для выполнения блочных операций (например, 4x4 матрицы). За один такт оно может выполнять десятки или сотни эквивалентных операций.

Результат: рост производительности на порядки и радикальное улучшение энергоэффективности — ключевого параметра AI-кластеров.

TPU И ASIC:

УЗКАЯ СПЕЦИАЛИЗАЦИЯ ДЛЯ МАКСИМАЛЬНОЙ ЭФФЕКТИВНОСТИ

Пока NVIDIA развивала GPU, другие компании начали создавать полностью специализированные AI-процессоры.



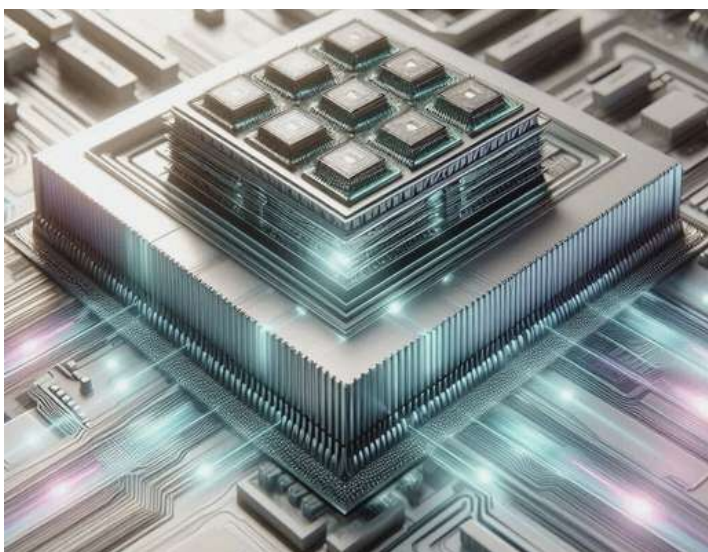
TPU (TENSOR PROCESSING UNIT)

- Разработчик: Google
- Тип: Специализированный ASIC
- Архитектура: systolic array — ритмичная “пульсация” данных между вычислительными элементами

ПРЕИМУЩЕСТВА:

- невероятная энергоэффективность;
- высокая пропускная способность;
- идеальны для матричных операций (свертки, dense-слои).

TPU используются преимущественно внутри Google Cloud.



ASIC (APPLICATION-SPECIFIC INTEGRATED CIRCUIT)

ASIC — это чип, полностью оптимизированный под одну задачу.

Примеры:

- Groq
- Graphcore IPU
- Cerebras WSE (гигантский wafer-scale-чип)

ПРЕИМУЩЕСТВА:

- максимальная эффективность под конкретный тип вычислений;
- альтернативная архитектура (dataflow, графы операций).

НЕДОСТАТКИ:

- высокая стоимость разработки;
- слабая адаптивность к изменению алгоритмов (модели растут слишком быстро).

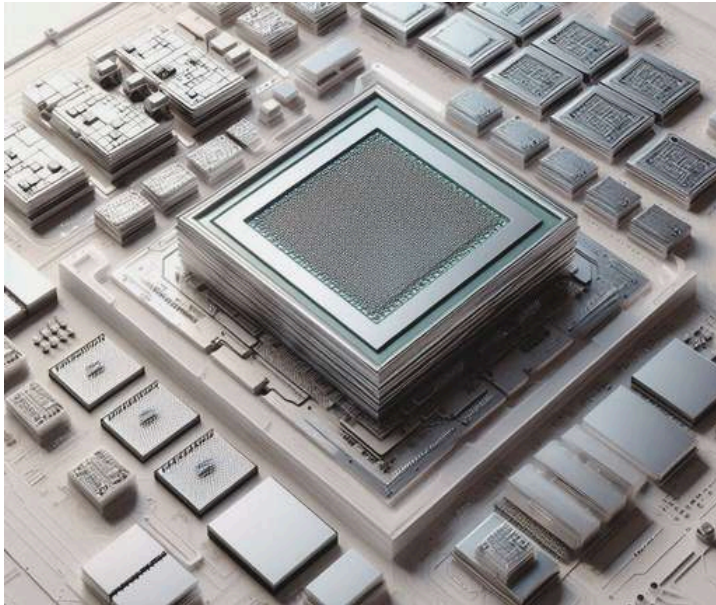


ЧИПЛЕТНАЯ АРХИТЕКТУРА: НОВАЯ ПАРАДИГМА ПРОЕКТИРОВАНИЯ ПРОЦЕССОРОВ

Из-за замедления закона Мура создание огромных монокристаллических чипов становится крайне дорогим. Решение — **чиплеты**.

Что такое чиплеты?

Процессор собирается из нескольких маленьких кристаллов — **чиплетов**, соединённых высокоскоростными интерфейсами (Infinity Fabric, EMIB, NVLink-C2C).



ПРЕИМУЩЕСТВА ЧИПЛЕТОВ

- **Лучший выход годных**

Маленькие чиплеты производить проще — дефект не уничтожает весь процессор.

- **Гетерогенная интеграция**

Разные чиплеты могут использовать разные техпроцессы:

- вычисления — 3 нм,
- I/O — 12 нм,
- память НВМ — свой специализированный процесс.

- **Модульность**

Производитель может собирать множество конфигураций — от серверных CPU до гигантских AI-ускорителей.

- **Масштабируемость**

Чиплеты можно объединять в сложные модули из десятков активных элементов.

Примеры чиплетных архитектур

- AMD EPYC / Ryzen — пионеры массовой чиплетизации CPU.
- NVIDIA Grace Hopper Superchip — CPU + GPU, объединённые NVLink-C2C.
- Intel Ponte Vecchio — 47 чиплетов, интегрированных в единый ускоритель.

Чиплеты — это фундаментальная смена философии: модульная, гибкая, масштабируемая архитектура для будущего AI.

ЗАКЛЮЧЕНИЕ:

Эволюция аппаратных ускорителей искусственного интеллекта прошла путь:

1. От GPU — универсального параллелизма.
2. К тензорным ядрам — ускорению матричных операций.
3. К TPU/ASIC — специализированным, энергоэффективным чипам.
4. К чиплетам — модульной архитектуре будущего.

Ключевые выводы:

- NVIDIA пока сохраняет лидерство благодаря полному стеку «железо + ПО + сеть».
- ASIC и TPU создают сильное давление на нишевые направления.
- Чиплеты открывают путь к гибридным системам, где CPU, GPU, НВМ и специализированные ядра собраны в единое вычислительное пространство.
- Будущее — за гетерогенными системами, адаптируемыми под конкретные AI-нагрузки.

