

АРХИТЕКТУРА СЕТЕЙ ДЛЯ AI:

ТОПОЛОГИИ, ПРОТОКОЛЫ И ПРАКТИЧЕСКИЕ СТРАТЕГИИ

В предыдущих статьях мы выяснили, почему традиционные CPU не справляются с AI и какую колоссальную энергию потребляют новые системы. В этой статье мы разберём два фундаментальных кита сетевой инфраструктуры AI: организацию путей передачи (топологии) и технологии обмена данными (протоколы).

Эпоха больших моделей сделала сеть не просто компонентом дата-центра, а ключевым фактором производительности.

Обучение распределённых нейросетей включает постоянную синхронизацию данных между сотнями и тысячами GPU. Если сеть не справляется — падает скорость обучения, растут задержки, а иногда процесс полностью останавливается.

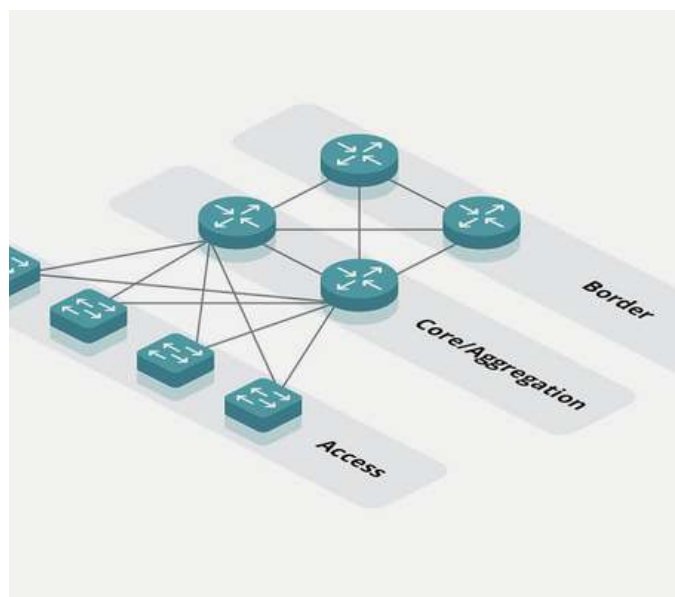
Чтобы избежать узких мест, инфраструктура AI требует неблокирующих топологий и протоколов, которые обеспечивают высокую пропускную способность, стабильные задержки и потерю нуля пакетов.

ПОЧЕМУ AI НУЖДАЕТСЯ В НЕБЛОКИРУЮЩИХ СЕТЯХ

Традиционные иерархические сети ЦОД, построенные по принципу «корень — листья», неизбежно сталкиваются с проблемой блокировок. Представьте многополосную дорогу, которая сужается до одной полосы перед мостом — именно так выглядит сетевое соединение, когда десятки серверов пытаются одновременно обмениваться данными через ограниченное количество uplink-портов.

Но AI — это непрерывная двусторонняя передача больших объёмов данных между GPU.

Если канал узкий, а потребителей много, возникает «эффект моста»: несколько потоков заполняют uplink, и новые соединения блокируются.



Что такое блокирующая сеть

Среда, в которой новое соединение может быть не установлено из-за занятости ресурса существующими потоками.

Почему это критично

При распределённом обучении тысячи GPU должны синхронно обмениваться градиентами. Любая задержка одного узла приводит к задержке всех.

Решение

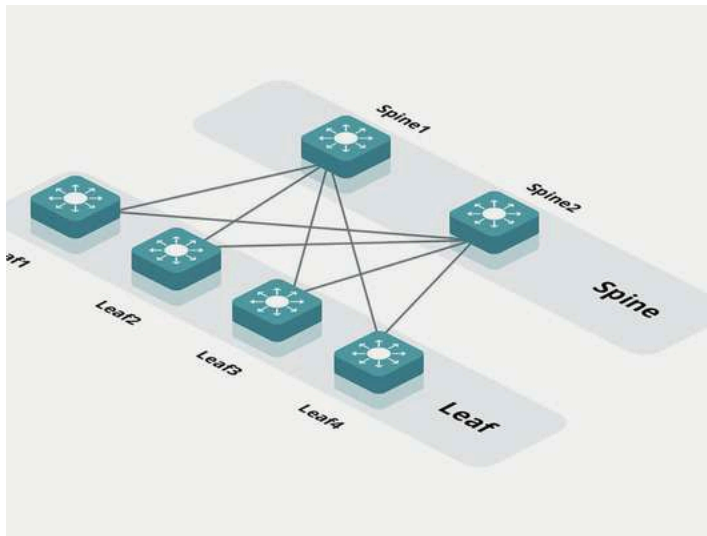
Неблокирующие сети, основанные на теории Клоза (Clos Network), где у каждого потока всегда есть доступный путь. Теория неблокирующих сетей (строго неблокирующая, перенастраиваемая неблокирующая и пр.) — это фундамент, на котором строится современная сетевое проектирование для кластеров. Практические расчёты и классификации см. в **Руководстве Patchwork/Mycelium — Раздел «Сетевые топологии», п. 1.2 и п. 4.3.**



Ключевой вывод: для AI-кластеров неприемлемы традиционные иерархические сети — необходима архитектура, гарантирующая одновременную передачу данных между множеством узлов без конфликтов.

FAT-TREE (SPINE-LEAF): ПРАКТИЧЕСКАЯ ОСНОВА СЕТЕЙ AI

Fat-Tree — это реальная реализация идей Клоза, оптимальная для больших высокоплотных AI-кластеров. Концептуально это «утолщённое дерево», где пропускная способность каналов увеличивается по мере приближения к корню, сводя к минимуму переподписку (oversubscription) и обеспечивая равномерную пропускную способность между узлами.



ПРЕИМУЩЕСТВА FAT-TREE / SPINE-LEAF

- Отсутствие переподписки — uplink и downlink сбалансированы.
- Предсказуемая задержка — одинаковое количество хопов между любыми серверами.
- Множественные маршруты — выше отказоустойчивость.
- Лёгкая масштабируемость — добавление новых уровней по формальным правилам.

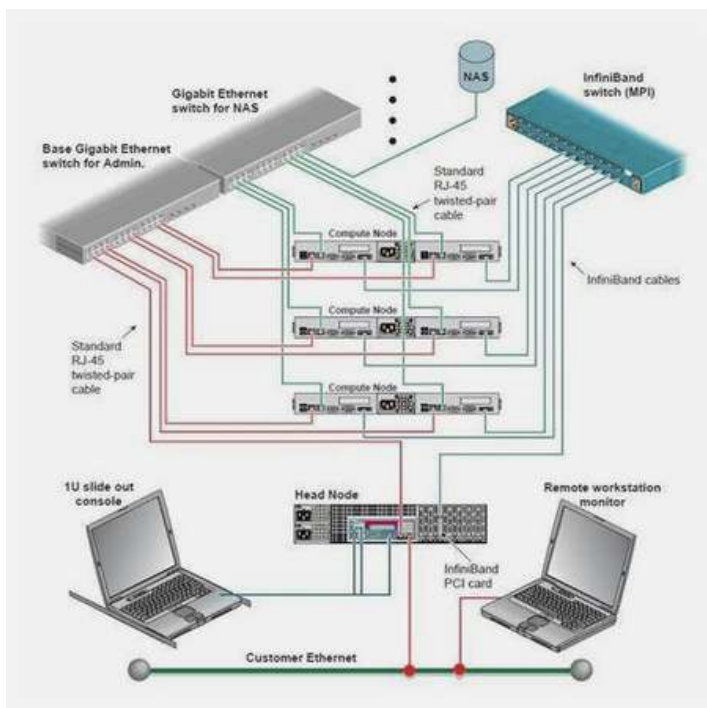
Fat-Tree сегодня является де-факто стандартом при проектировании AI-кластеров на десятки и сотни тысяч GPU.

Подробные расчёты (число серверов, портов, коммутаторов при radix 32–64) приведены в руководстве Patchwork/Mycelium.

ПРОТОКОЛЫ ПЕРЕДАЧИ ДАННЫХ: INFINIBAND VS ROCE

Даже идеальная топология бесполезна без эффективного протокола.

В мире AI доминируют две технологии, каждая со своей философией и областью применения: InfiniBand и RoCE (RDMA over Converged Ethernet).



INFINIBAND: ЭТАЛОННАЯ СЕТЬ ДЛЯ HPC И AI

Коммуникационная экосистема, спроектированная для экстремально низкой задержки и высокой пропускной способности. Этот протокол создавался как высокоскоростной, детерминированный транспорт.

ПРЕИМУЩЕСТВА:

- RDMA — прямой доступ к памяти без участия CPU.
- Lossless Fabric — отсутствие потерь пакетов.
- Аппаратные транспортные сервисы — задержки < 1 мкс.
- In-Network Computing (SHARP) — суммирование градиентов прямо в коммутаторе.

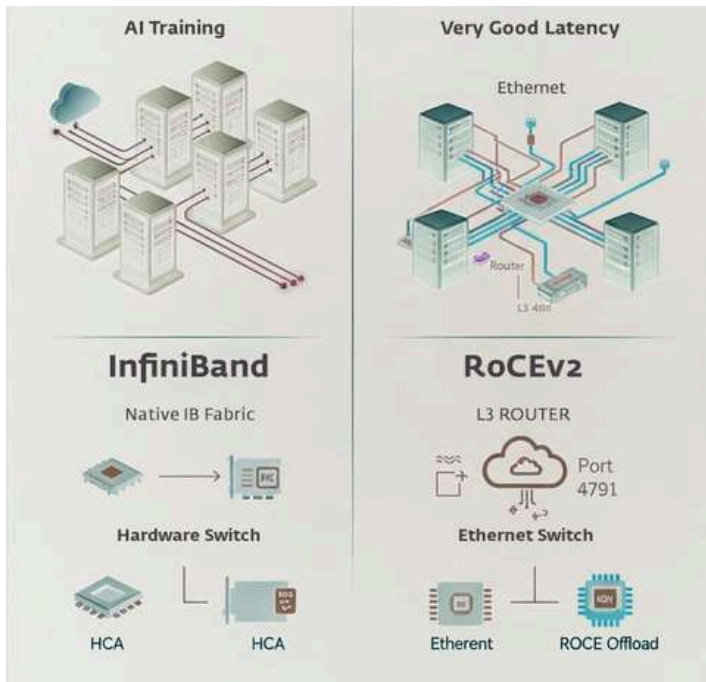
Современные поколения:

- HDR — 200 Гбит/с
- NDR — 400 Гбит/с

InfiniBand — лучший выбор для обучения больших моделей (256+ GPU).

Детальное описание архитектуры InfiniBand, включая роль Subnet Manager, HCA, коммутаторов и шлюзов, а также сравнение с TCP/IP, см. Руководство Patchwork/ Mycelium — Глава «Что такое InfiniBand», п. 3.





RoCE (RDMA OVER ETHERNET): ГИБКОСТЬ И МАССОВОСТЬ

RoCEv2 работает поверх UDP/IP и позволяет использовать RDMA внутри Ethernet-инфраструктуры.

Когда RoCE подходит:

- Для инференса.
- Для гибридных сред (AI + Storage + Management).
- Для средних кластеров, где задержка не критична.
- Когда команда сильна в Ethernet-экспертизе.

Требования:

- PFC — предотвращение потерь пакетов.
- ECN — предупреждение перегрузки.
- QoS / DSCP — маркировка классов трафика.

RoCE позволяет снизить стоимость владения и использовать массовый Ethernet-серверный стек

INFINIBAND VS ROCE: КАК ВЫБРАТЬ

Выбор между InfiniBand и Ethernet — это не вопрос «что лучше», а «что оптимальнее для вашей задачи». Ниже — сводная таблица критериев:

КРИТЕРИЙ	INFINIBAND	HIGH-PERFORMANCE ETHERNET (ROCE)
ЗАДЕРЖКА	СВЕРХНИЗКАЯ (< 1 МКС)	НИЗКАЯ (НЕСКОЛЬКО МКС)
ДЕТЕРМИНИЗМ	ОЧЕНЬ ВЫСОКИЙ	ЗАВИСИТ ОТ НАСТРОЙКИ
ЭКОСИСТЕМА ДЛЯ AI	ПОЛНЫЙ СТЕК, SHARP	БЫСТРО РАЗВИВАЕТСЯ
УНИВЕРСАЛЬНОСТЬ	УЗКОСПЕЦИАЛИЗИРОВАННЫЙ	ШИРОКОЕ ПРИМЕНЕНИЕ
СТОИМОСТЬ	ВЫШЕ	НИЖЕ
ЭКСПЕРТИЗА	ТРЕБУЕТСЯ СПЕЦИФИЧЕСКАЯ	БОЛЕЕ ДОСТУПНАЯ

ПРАКТИЧЕСКИЕ РЕКОМЕНДАЦИИ:

- **InfiniBand:** Масштабное обучение (> 256 GPU) — за счёт детерминизма и низкой задержки.
- **Ethernet + RoCE:** Инференс и средние кластеры — лучше с точки зрения гибкости и стоимости.
- **Гибридные Fabrics:** смешанная среда ЦОД или разделение слоёв по целям — оптимальный путь.

Дополнительный сравнительный анализ оборудования и реализации (Cisco, Arista, Juniper, H3C и т. д.) см. в Руководстве Patchwork/Mycelium — Раздел «Сетевые топологии», п. 7.



ЗАКЛЮЧЕНИЕ:

Сетевой стек современного AI-кластера — это тщательно спроектированная система. Топология Fat-Tree (Spine-Leaf) обеспечивает предсказуемую, неблокирующую и масштабируемую физическую основу. Протоколы InfiniBand и RoCE предоставляют семантику удалённого доступа к памяти, необходимую для эффективной работы распределённых алгоритмов. Выбор зависит от сценария, бюджета и долгосрочной стратегии.

На практике часто применяют комбинированные сети:

- ядро обучения — InfiniBand,
- периферия и сервисы — Ethernet + RoCE.

Такой подход обеспечивает баланс стоимости, предсказуемости задержек и совместимости.

