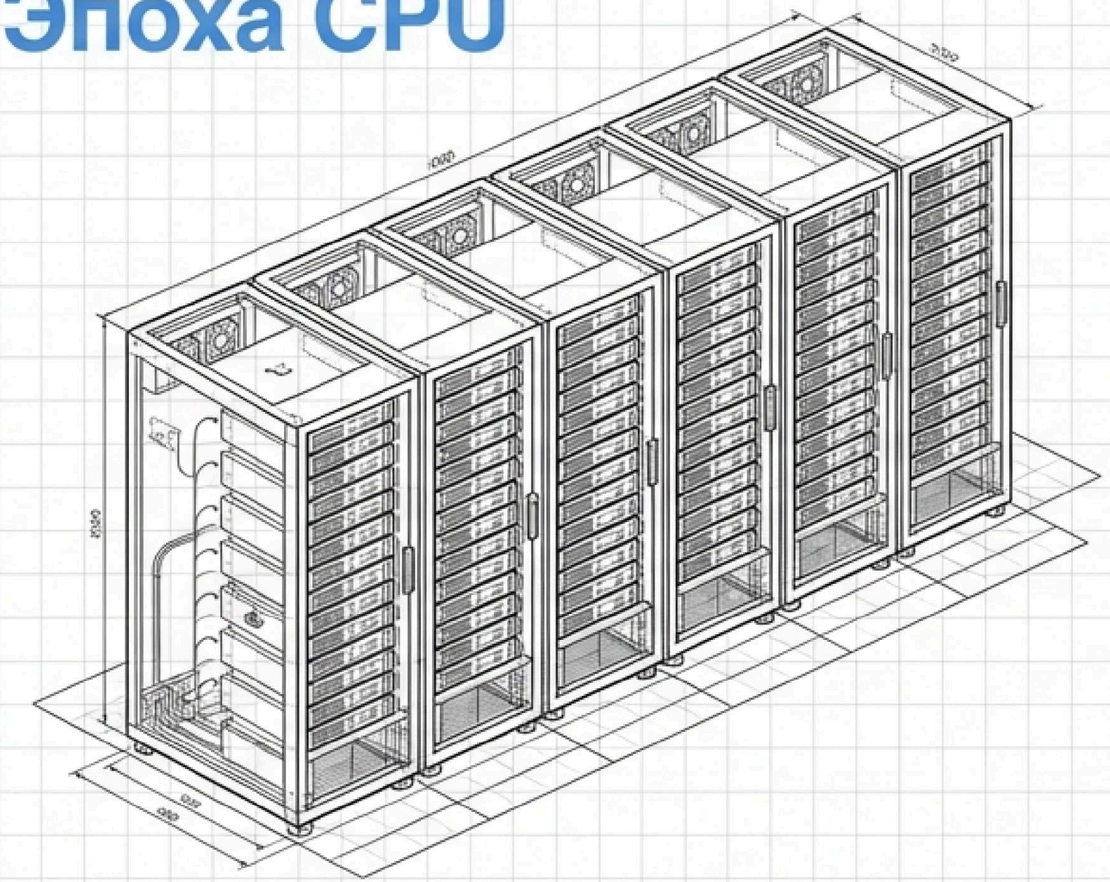
AI перегревает облака: новая экономика GPU-инфраструктуры

Почему модель «бесконечной эластичности» умирает и что приходит ей на смену.

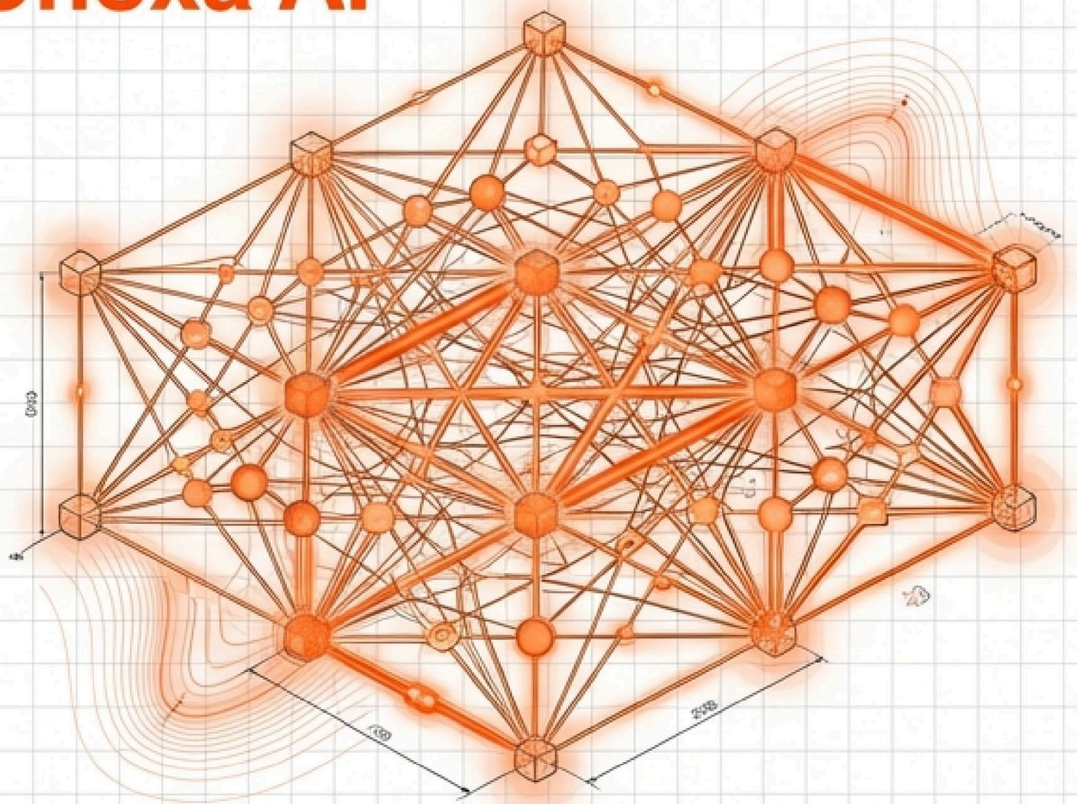
Инженерный взгляд на смену парадигмы.

Эпоха CPU



- Виртуальные машины (VM)
- Burst workloads (всплески нагрузки)
- Утилизация 30–50%
- Оплата по потреблению (Pay-as-you-go)

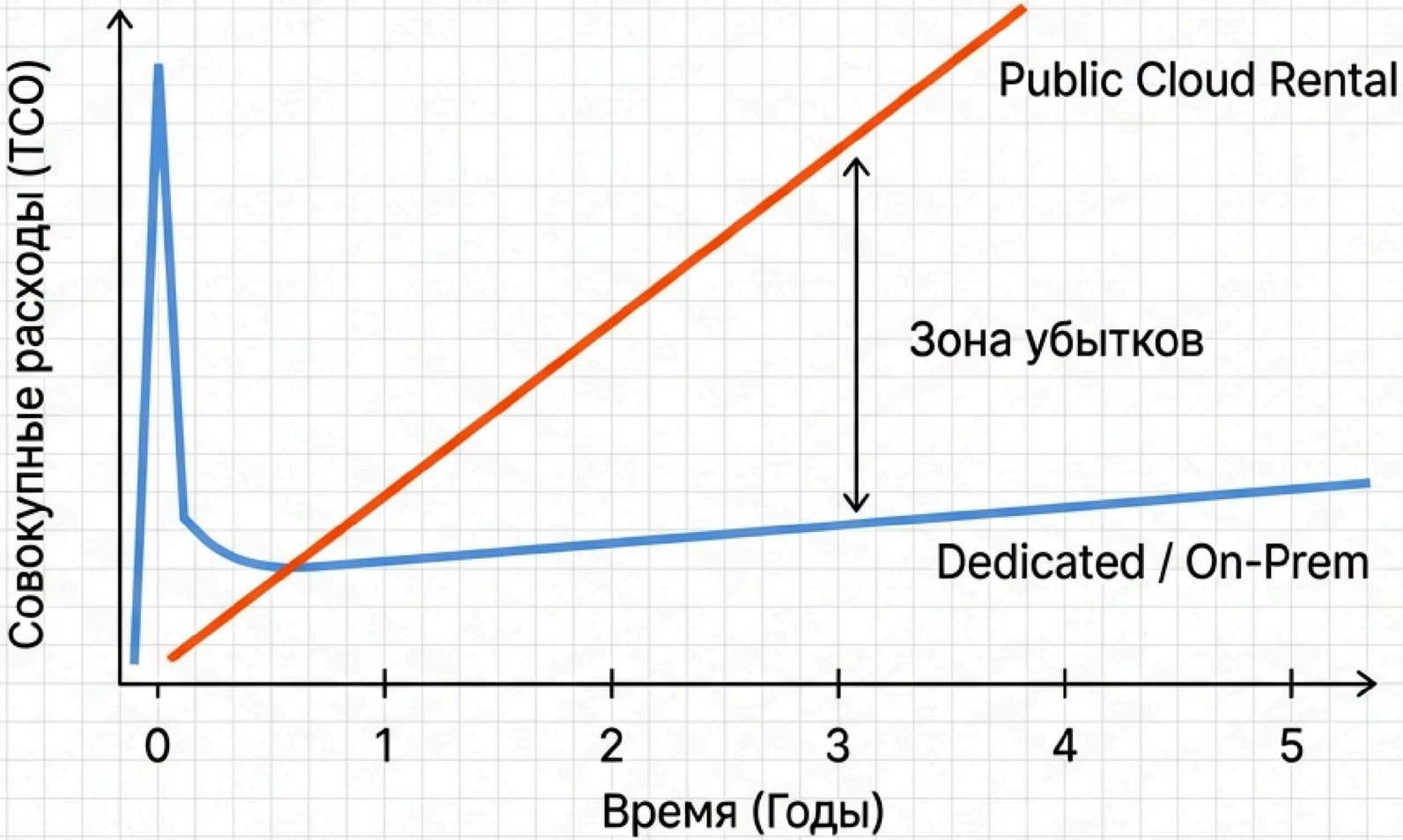
Эпоха AI



- Вычислительные фабрики
- Constant load (постоянное давление)
- Утилизация 70–95%
- Стратегический актив (CAPEX)

«Мы привыкли считать vCPU и гигабайты. AI заставляет считать в ваттах, гигабитах и миллионах токенов.»

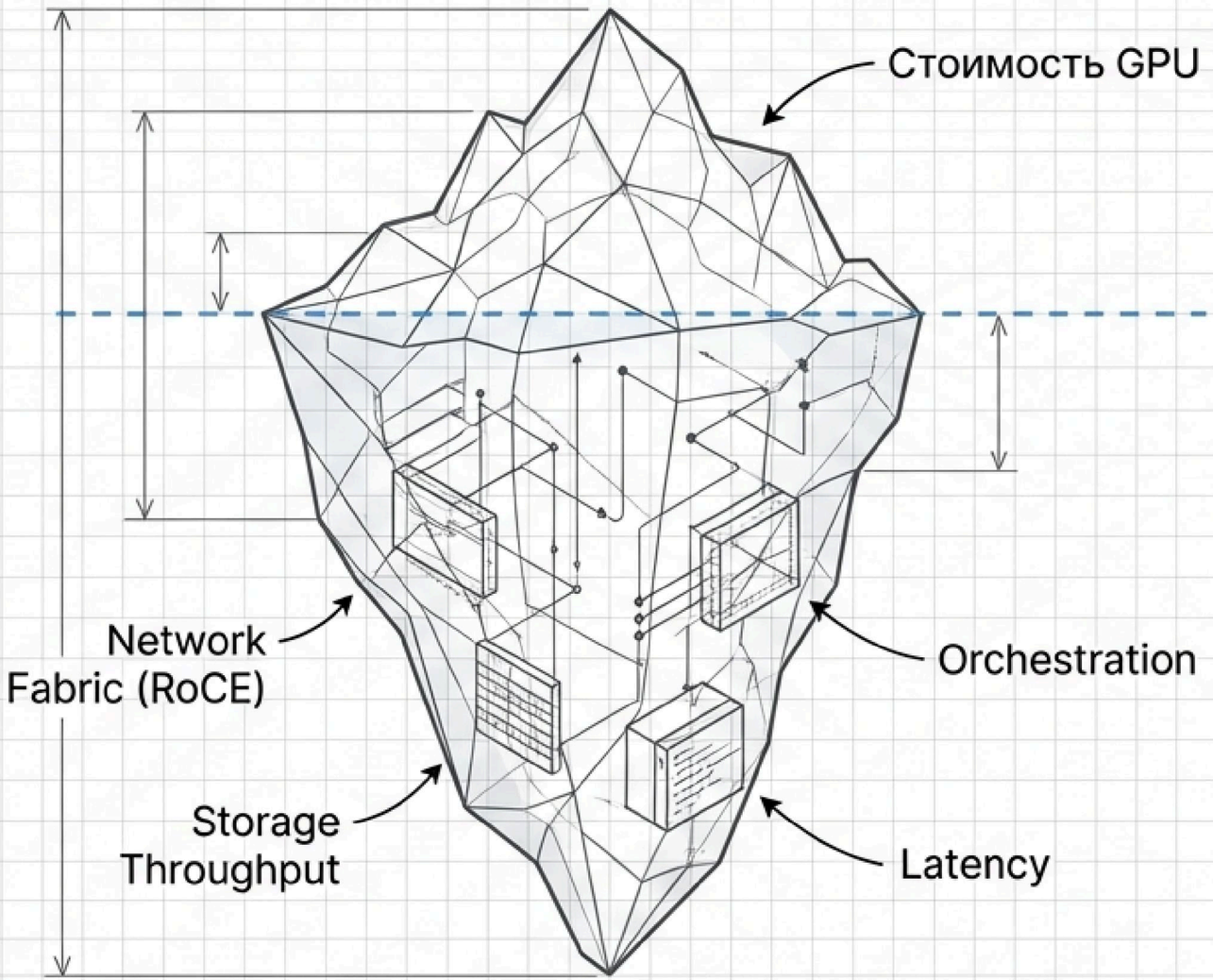
Если AI работает круглосуточно, аренда — это ошибка



- **Public Cloud:**
  Высокая почасовая ставка + дорогой egress трафик. Идеально для экспериментов.
- **Dedicated/On-Prem:**
  Высокий CAPEX, но кардинально низкий TCO на дистанции 3+ лет. Идеально для Production.

При прогнозируемой нагрузке модель «почасовой аренды» становится дороже собственной инфраструктуры. Экспериментальную экономику нельзя применять к рабочим нагрузкам.

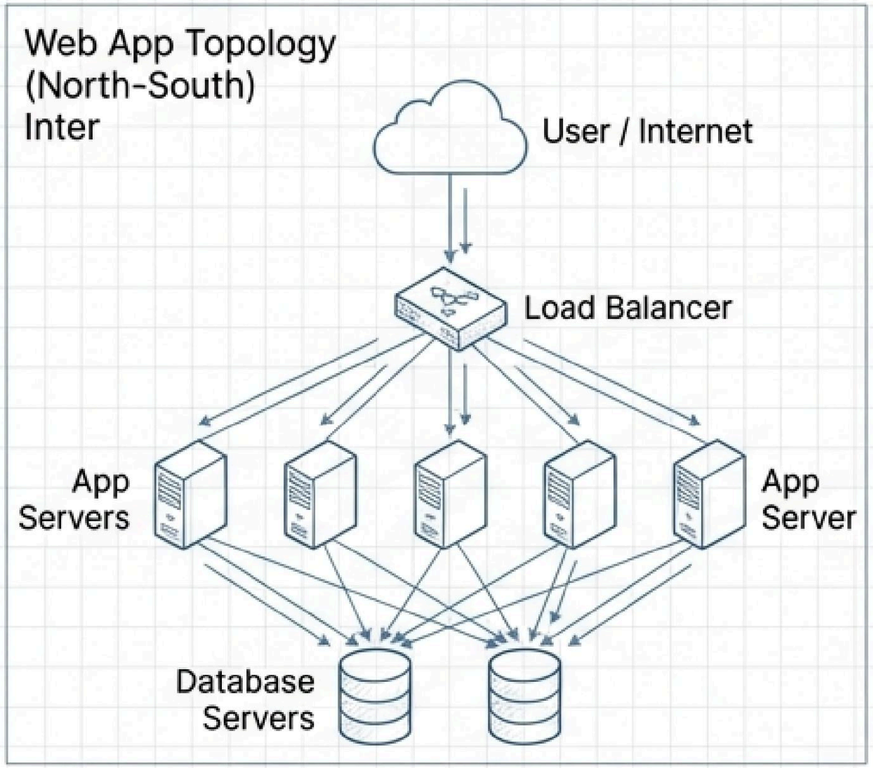
Айсберг неэффективности

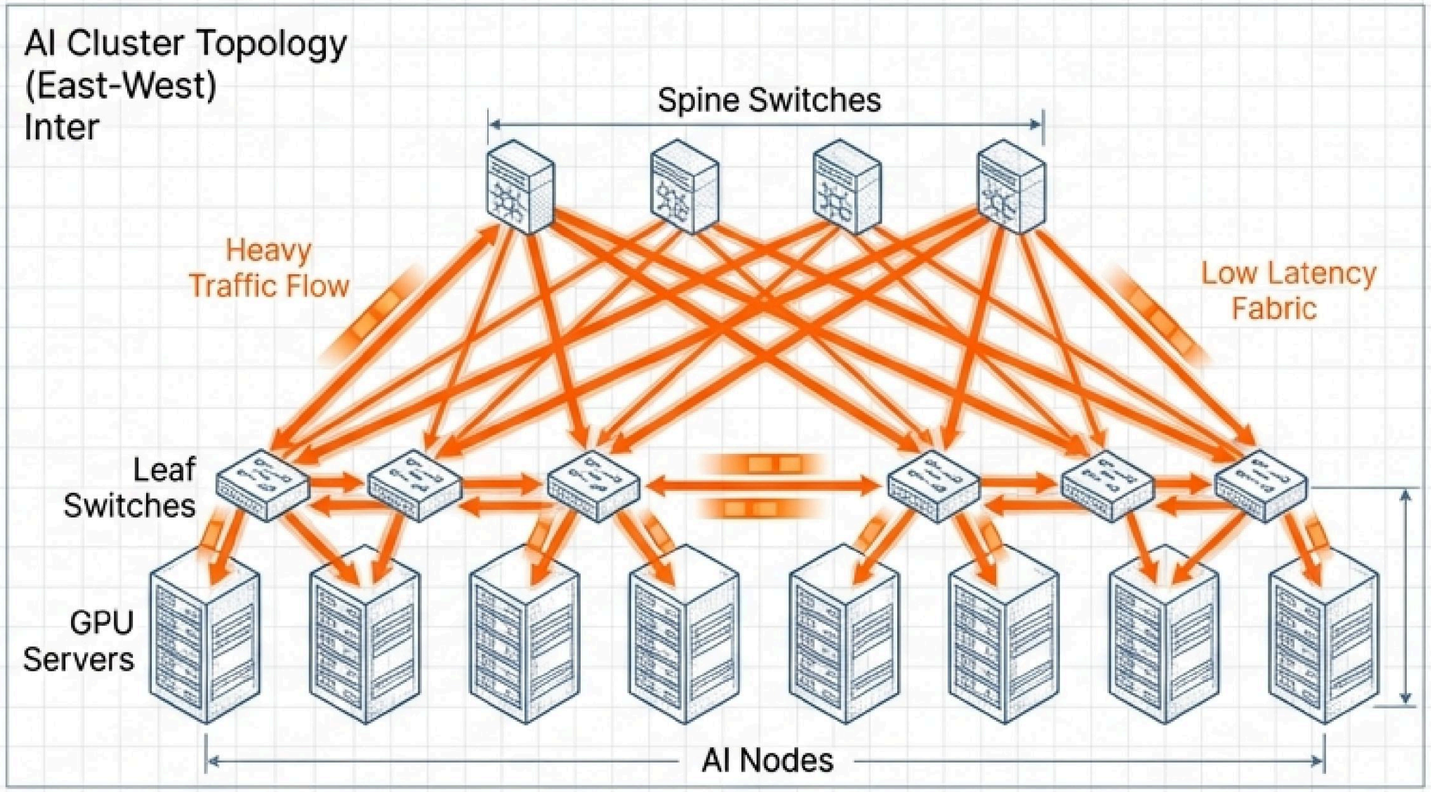


Большинство думает, что проблема в нехватке чипов.
Реальность: GPU простаивают из-за слабой сети и медленных данных.

«Самый дорогой ресурс — это не GPU. Это простаивающий GPU.»

Сеть стала важнее гипервизора

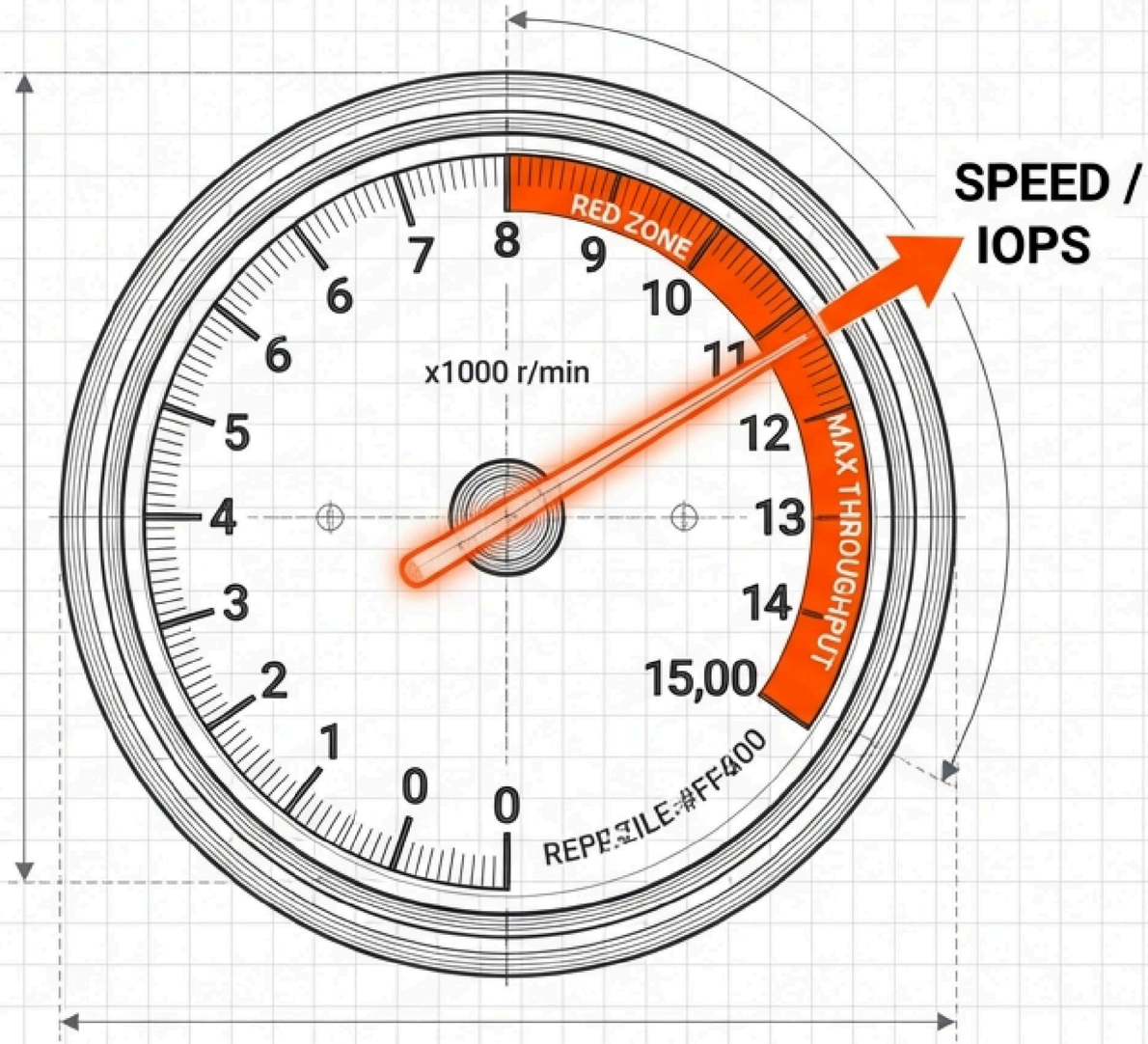




- В классическом облаке трафик идет от пользователя к серверу (North-South).
- В AI-кластере трафик идет между серверами (East-West).
- Рост горизонтального трафика — кратный.

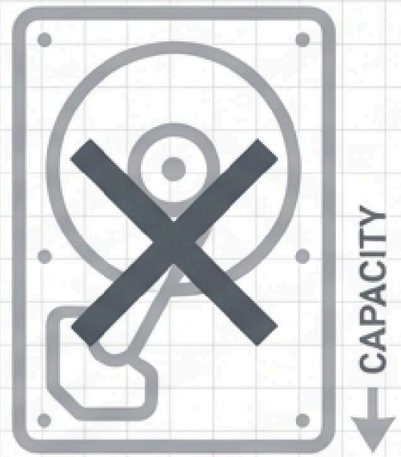
Критичны RoCE, низкая задержка (latency) и отсутствие oversubscription. Если сеть “захлебывается”, суперкомпьютер останавливается.

Storage: Скорость (Throughput) > Емкость (Capacity)



Раньше мы платили за место для хранения (Capacity).

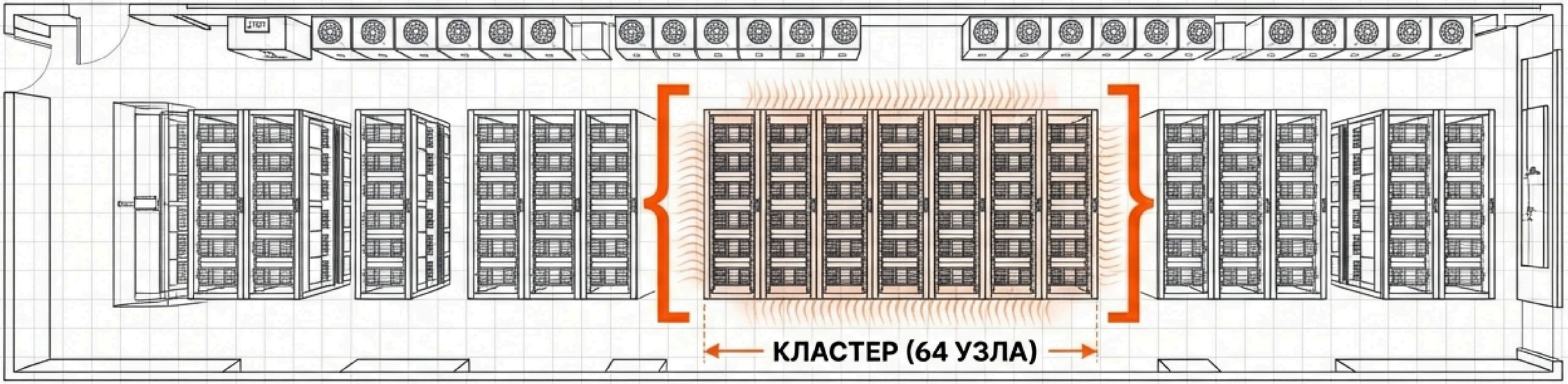
Теперь мы платим за скорость подачи данных (IOPS/Throughput).



Потери эффективности из-за storage bottleneck могут достигать 20—30%. Видеокарта не должна ждать данные.

Кейс: Национальный AI-кластер

Что происходит, когда AI становится реальной инфраструктурной обязанностью.

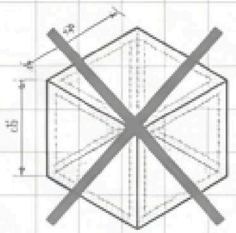


64	Вычислительных узла
512	GPU
LLM-as-a-Service	Модель использования

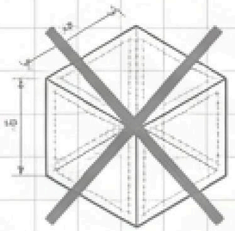
Переход от пилотных проектов к Production-нагрузке для государственных и коммерческих пользователей.

Уроки из траншей: Ожидания vs Реальность

ОЖИДАНИЕ



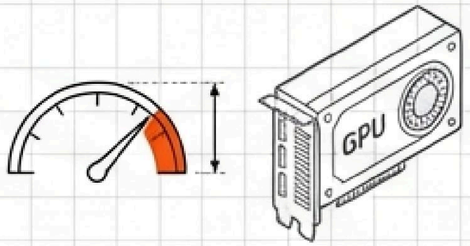
~~**Ожидание:** Burst-нагрузки~~ (то густо, то ~~пусто~~).



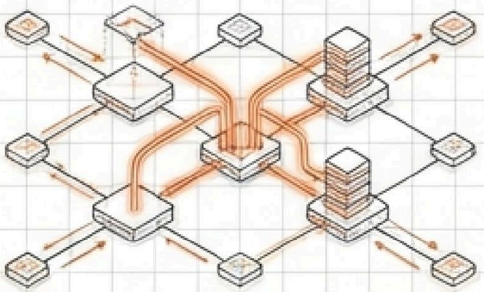
Ожидание:
~~Главный~~ вопрос — «Сколько у нас ~~видеокарт~~?»

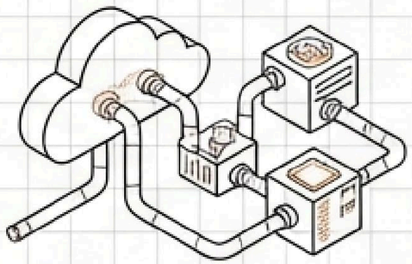
РЕАЛЬНОСТЬ

☑ **Реальность:**
Постоянная утилизация GPU **70–90%**.



☑ **Реальность:**
Главный вопрос — **«Как они связаны между собой?»**





Облако перестало быть набором виртуальных машин. Оно стало единой распределенной фабрикой вычислений.

«Радиоактивные данные»: новая категория риска

AI делает данные более ценными, но и более **опасными** (персданные, финансы, гостайна).

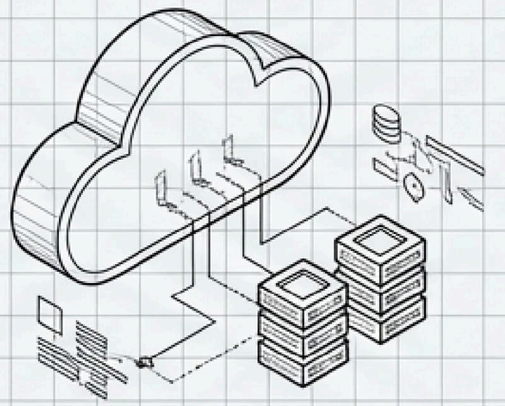
Если ваши данные нельзя показывать конкуренту, их нельзя выносить в публичный inference без контроля.

REQUIREMENT:
Data Locality + Sovereign AI infrastructure.
Изоляция сегментов обязательна.

Гибридная модель будущего

Cool Steel Blue

R&D / Training Experiments

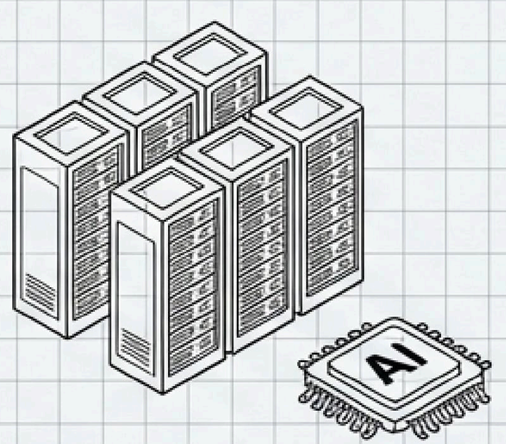


Public Cloud

Flexibility, Scalability

International Orange

Production / Inference



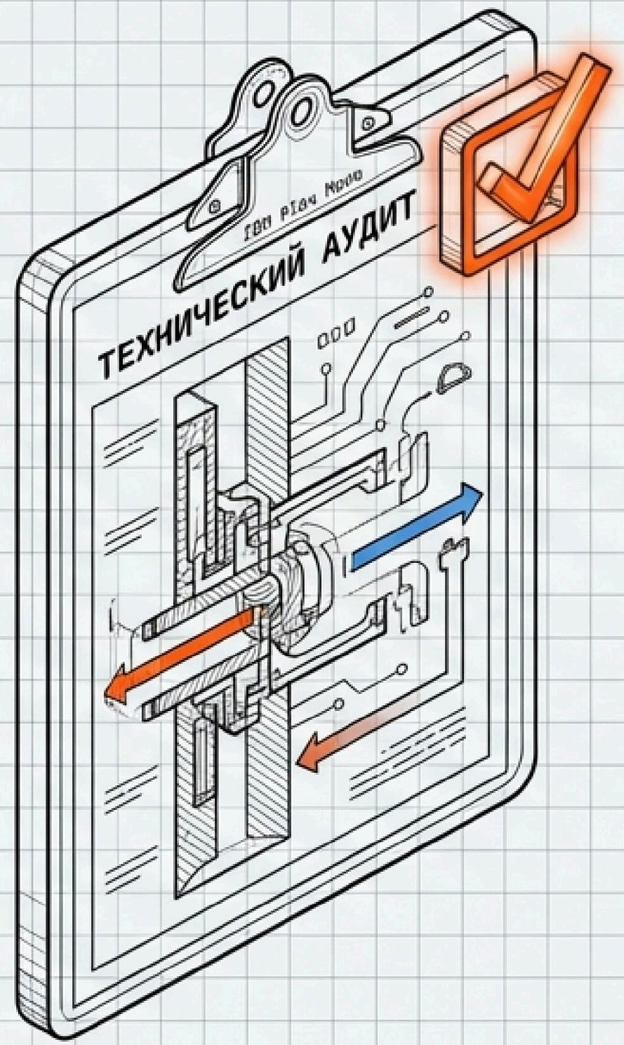
On-Prem / Dedicated

Cost Control, Security, Latency

RULE OF THUMB:

Разделяйте R&D и Production. Не используйте экономику экспериментов для постоянной нагрузки.

Чек-лист для новой экономики

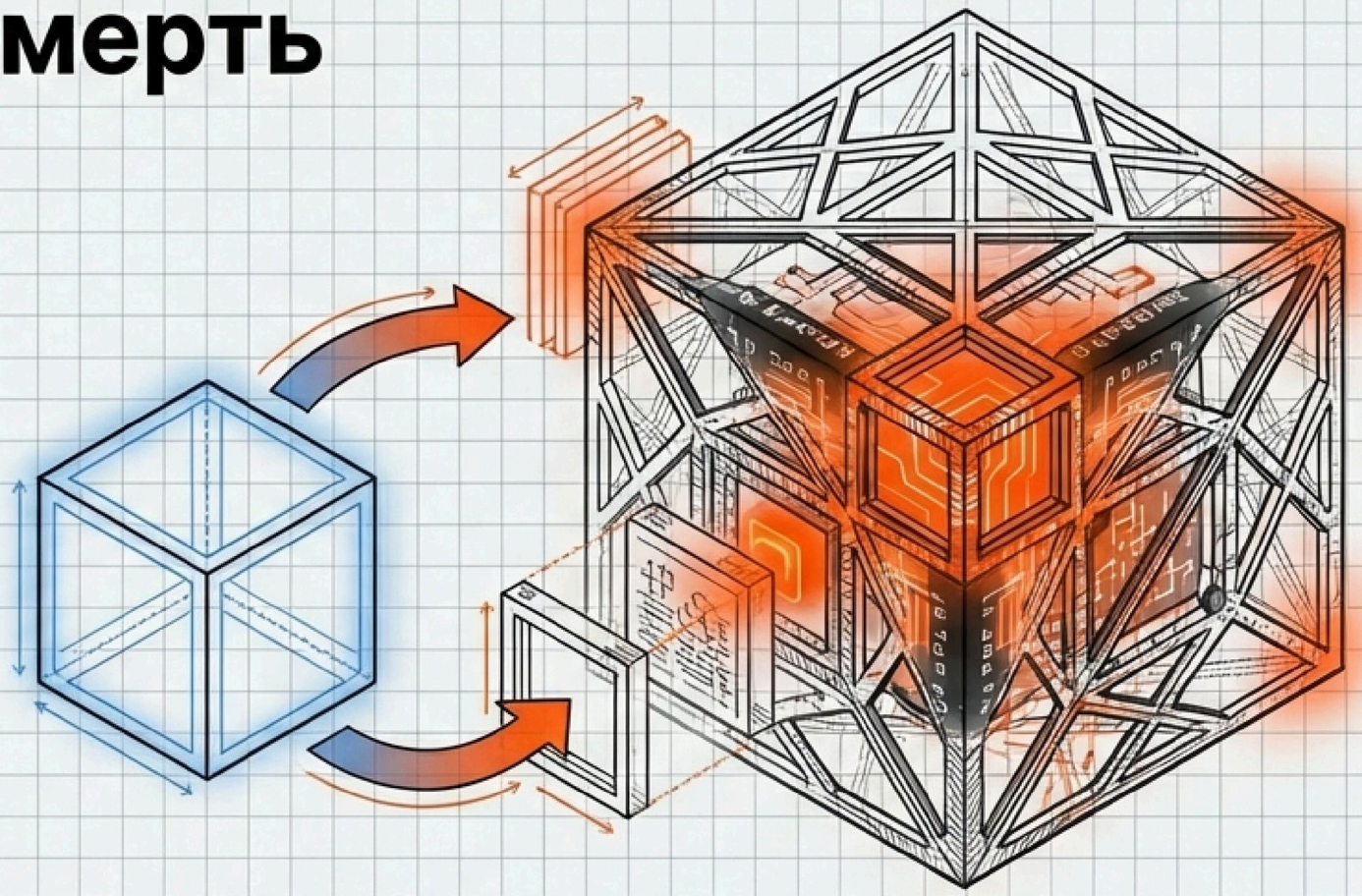


- [x] Считать **TCO** на **3–5 лет**, а не почасовую ставку.
- [x] Относиться к инфраструктуре как к **активу** (CAPEX), а не как к расходу (OPEX).
- [x] Проектировать сеть под **AI**, а не под VM.
- [x] **Open Source** стек: избегайте vendor lock-in для сохранения контроля над архитектурой.

Эволюция, а не смерть

Облако не умирает.
Но эпоха «CPU-first»
архитектуры закончилась.

AI-нагрузка
не масштабируется
по правилам старого
облака.



Хватит строить виртуальные машины.
Начинайте строить фабрики вычислений.

Technical Precision / Engineering Blueprint

Тезисы для обсуждения

«GPU — это уже не ресурс. Это стратегический актив.»

«Самая дорогая ошибка — считать AI по CPU-метрикам.»

«Экономия в облаке без архитектуры превращается в сюрприз в счете.»

«Если AI работает круглосуточно, платить за него как за эксперимент — экономическая ошибка.»