Домашняя ИИ лаборатория

Эволюция от игрушек до командного обучения

DOTNEXT

О докладчике

- Более 20 лет в разработке, архитектуре, построении процессов, управлении и в бесконечной цифровой трансформации банков и ИТ-компаний
- Выигрывал MarksWebb когда за это еще давали премии
- Личный рекорд – проект с 50+ внедрениями в пром в день
- На фоне развития ИИ погрузился в тему и завел собственный ТГ канал «Вселенское зло об ИИ»

@EVIL_ABOUT_IT

Зачем нужна домашняя ИИ лаборатория?

- Чтобы потратить кучу денег на «железо» и получить приглашение выступить на DotNext ☺

Зачем нужна домашняя ИИ лаборатория?

- ~~Чтобы потратить кучу денег на «железо» и получить приглашение выступить на DotNext~~ ☺
- ИИ создает риск потери работы в ИТ
- Если ты профессионал, то нужно иметь ценность выше рынка, а значит знать больше рынка
- Облака используют все, инженерку вокруг моделей изучает почти никто. Пройти путь развертывания кластера с нуля – бесценно
- Дорого. Но потеря уровня дохода за 25 лет до пенсии - дороже
- **Начальная цель – академическая. Надо учиться**

В конце января рак свистнул

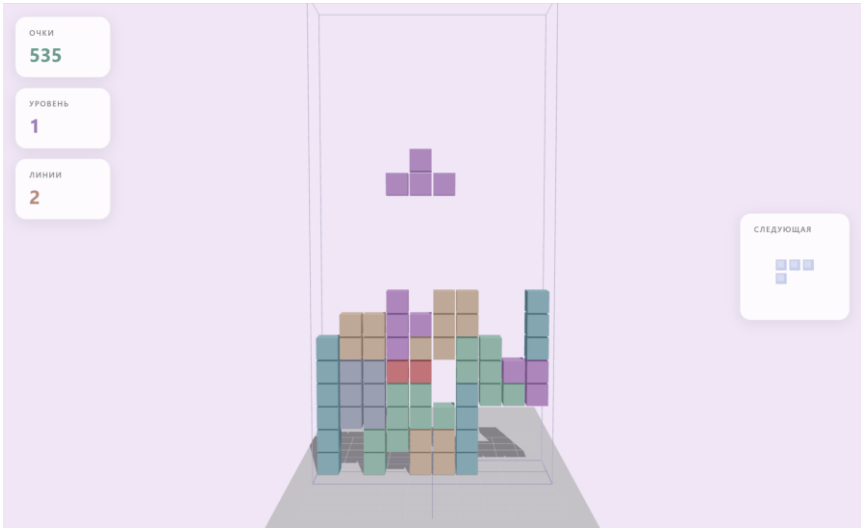
Вовремя для бюджета

Что такое disrupt?

Disrupt - перестать делать однотипную работу, которую ты не можешь не делать

Ради интереса запускаю LLM локально

- Не помню какие модели запускал, но на Хабре была классная статья, где ИИ мог/не мог написать змейку
- Пишу промпт – змейка на месте
- Прошу тетрис – провал
- Ставлю модель побольше – успех
- Прошу микросервис – так себе
- Модель еще больше – уже интересно



Главное выбрать хорошую модель

Внезапно LLM дали такое же ощущение

- Можно увеличить кроссфункциональность в команде и ускорить ее в разы
- Можно получать индивидуальное ПО по требованию, а не искать что-то приемлемое
- Можно делать то, что раньше не умел или на что не имел времени
- Можно убрать совещания

И это у тебя в компьютере, безопасно и приватно

Нужна только локальная модель

Хорошие локальные модели – большие

- На январь 2026 принцип был железобетонный. Больше - лучше
- Первая по времени выхода интересная для кодинга локальная модель GLM-4.7 занимает почти 720ГБ
- Справившаяся с моими локальными задачками Qwen3-Coder-Next:80B – 160ГБ
- Даже в сжатом (квантованном) виде сильно больше 24ГБ у 4090, едва ползает

А что есть на рынке для энтузиастов?

Мини-ПК для локального запуска моделей – ключевая инновация 2025 года

Mac Studio M3 Ultra | M4 Max [6,5-8,5K USD за 256ГБ версию, 4k за 128ГБ]

- 2025-2026 года выпуска (M2 Ultra в 2024)
- До x2 объем памяти (96-512ГБ), до x2 производительности
- Кластер до 5 машин из коробки по Thunderbolt, не расширяется
- Не пригоден для корпоративного внедрения в рамках ИЗ

Nvidia DGX Spark (или альтернативы на GB10 от других вендоров) [5k USD за 128ГБ]

- 2025 год
- 128ГБ памяти, x1 производительность в инференсе, x2-5 в обучении
- Кластер на 2 машины из коробки, не расширяется
- ARM CPU и нетиповой Linux блокирует корпоративное внедрение в рамках ИЗ

AMD Strix Halo (Ryzen AI Max+ 395, от разных вендоров) [2-3k USD за 64-128ГБ]

- 2025 год, официальная поддержка в Linux с 23.10.2026, open source драйверы
- 64-128ГБ памяти, x1 производительность в инференсе, x1 в обучении
- Кластер RDMA с дополнительными картами расширения (Intel, Mellanox) до 5 шт.
- Стандартный x86 позволяет использовать отечественное ПО

И кластер можно

Mac Studio M3 Ultra | M4 Max – счастье из коробки

- 80 Гбит/сек Thunderbolt 5 соединение
- RDMA/RoCE v2* over Ethernet over Thunderbolt начиная с MacOS 26.2 (дек 2025)

*RoCE - RDMA over Converged Ethernet, RDMA – Remote Direct Memory Access, необходимы для эффективной работы в tensor parallel кластере с микросекундными задержками

Nvidia DGX Spark – из пушки по воробьям для инференса

- 2xQSFP по 200 Гбит/сек (до 400 Гбит/с в bond теоретически)
- RDMA/RoCE v2 over Ethernet
- Реальная производительность упирается не в сеть

AMD Strix Halo (Ryzen AI Max+ 395, от разных вендоров) – очумелые ручки

- Из коробки 10G Ethernet или Thunderbolt 4/5 (нет поддержки RDMA на сегодня)
- Отдельные версии (например Minisforum MS-S1 Max) идут с PCI-E 16 разъемом (скорость PCI-E 4.0 x4 до 64 Гбит/с) или можно вывести его от NVME
- Возможна установка карт расширения (Intel E810, Mellanox) 25/100 Гбит/с с соединением устройств попарно с поддержкой RDMA/RoCE v2

Сколько памяти реально нужно

- Qwen/Qwen3.5-0.8B
- Qwen/Qwen3.5-2B
- Qwen/Qwen3.5-4B
- Qwen/Qwen3.5-9B
- Qwen/Qwen3.5-27B
- Qwen/Qwen3.5-35B-A3B
- Qwen/Qwen3.5-122B-A10B
- Qwen/Qwen3.5-397B-A17B

Сколько памяти реально нужно

- Qwen/Qwen3.5-**0.8B**
- Qwen/Qwen3.5-**2B**
- Qwen/Qwen3.5-**4B**
- Qwen/Qwen3.5-**9B**
- Qwen/Qwen3.5-**27B**
- Qwen/Qwen3.5-**35B**-A3B
- Qwen/Qwen3.5-**122B**-A10B
- Qwen/Qwen3.5-**397B**-A17B

namespace (huggingface.co)

- 27B = 27 Billions (миллиардов параметров в модели)

Сколько памяти реально нужно

- Qwen/Qwen3.5-0.8B
- Qwen/Qwen3.5-2B
- Qwen/Qwen3.5-4B
- Qwen/Qwen3.5-9B
- Qwen/Qwen3.5-27B
- Qwen/Qwen3.5-35B-**A3B**
- Qwen/Qwen3.5-122B-**A10B**
- Qwen/Qwen3.5-397B-**A17B**

namespace (huggingface.co)

- 27B = 27 Billions (миллиардов параметров в модели)
- A17B = **A**ctivated **17 B**illion parameters
- 397B-A17B – модель на 397 миллиардов параметров с активацией 17 миллиардов на каждый запрос

Сколько памяти реально нужно

- Qwen/Qwen3.5-0.8B
- Qwen/Qwen3.5-2B
- Qwen/Qwen3.5-4B
- Qwen/Qwen3.5-9B
- Qwen/Qwen3.5-27B

dense

- Qwen/Qwen3.5-35B-**A3B**
- Qwen/Qwen3.5-122B-**A10B**
- Qwen/Qwen3.5-397B-**A17B**

sparse (MoE*)

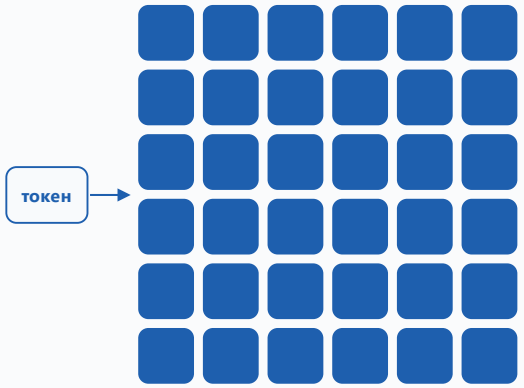
namespace (huggingface.co)

* MoE – mixture of experts

- 27B = 27 Billions (миллиардов параметров в модели)
- A17B = **A**ctivated **17 B**illion parameters
- 397B-A17B – модель на 397 миллиардов параметров с активацией 17 миллиардов на каждый запрос
- В dense моделях активируются всегда все параметры

Разница dense и sparse (MoE)

Dense («плотная» модель)

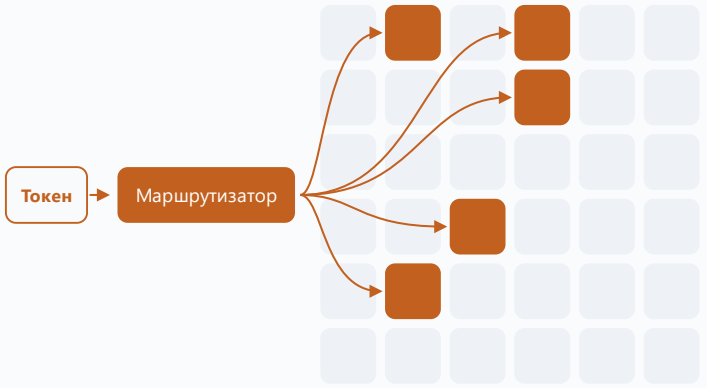


модель не разделена на части:
активны все веса на каждый токен

матрица: 36 блоков, активны все 36

vs

MoE (смесь экспертов)



веса разбиты на 36 экспертов:
на каждый токен активируются подходящие

матрица: 36 экспертов, активны 5 выбранных

Применимость локально

Dense

- Точнее для hard задач, требующих фокуса внимания на многом сразу
- В кодинге сильно лучше
- Минимум галлюцинаций
- Для запуска на dGPU с 24-32ГБ
- KV-кэш* в 2,7 раза больше

MoE (Mixture of experts)

- «Интеллект» сравним с большими моделями
- Быстрее в разы
- Лучше для больших контекстов из-за экономии памяти
- Долгие и исследовательские задачи
- Многоходовые размышления

*KV-кэш - **сохраняет уже вычисленные ключи (K) и значения (V)** для предыдущих токенов, чтобы не пересчитывать их заново при генерации каждого нового токена.

Мысли о качестве моделей в начале

FP16/32

Нативная точность моделей, но крайне ресурсоемко
Для FP16 можно грубо умножить количество параметров на 2 и получить требование к памяти для запуска – 27B требует 56ГБ, FP32 – 112ГБ (в теории)

FP8/FP4

Кратно меньший объем, но требует аппаратной поддержки
Потери точности – адекватные, ускорение вычислений – пропорциональное.
FP4 используется как нативный формат для обучения, что исключает потери.

Q6-Q8

Целочисленное квантование с минимальными потерями
Q8 вдвое легче чем FP16, Q6 – почти втрое. 28ГБ и 23ГБ для Qwen3.5-27B.
Скорость вычислений также существенно выше.

Q4-Q5

Золотая средина для промышленной эксплуатации
Потери точности адекватные для реальной работы.
17ГБ и 20ГБ для Qwen3.5-27B. С учетом KV-кэша пригодно для 24ГБ dGPU.

Q2, Q3 кванты – существенные потери. Только для лабораторных целей

Аналогия «как на самом деле», актуальная

FP16/32	Super HI-RES 192KHZ	
FP8/FP4	CD/DVD	**Мечта аудиофила** Нормальные рабочие форматы
Q4-Q8	MP3 320 kb/s	**Неотличимо от CD вне лаборатории** Даже на дорогом оборудовании
Q2-Q3	MP3 160-192kb/s	**На практике,** Все так и слушали

Q1 – радио на аналоговом приемнике в лесу. Сплошной шум

Цели лаборатории изначально

1. Разобраться как работает инференс моделей «с нуля»
2. Модели надо хотя бы пощупать, пусть и медленно, потом используем по работе
3. Работать на этом железе не планируется, только эксперименты

Начало марта 2026 – первые покупки

Бенчмарки, ожидание и реальность

- <5 t/s— работать невозможно
- 5-20 t/s — пойдет попробовать
- 20-50 t/s — подойдет агентам
- 50-100 t/s – нормальная скорость для чата
- >100 t/s — мечта

Бенчмарки, ожидание и реальность

- <5 t/s— работать невозможно
- 5-20 t/s — пойдет попробовать
- 20-50 t/s — подойдет агентам
- 50-100 t/s – нормальная скорость для чата
- >100 t/s — мечта

На одиночном Ryzen 395 или DGX Spark примерно следующая производительность:

- Qwen3.5-35B-A3B – 40-80 t/s
- Qwen3.5-122B-A10B – 15-25 t/s
- Qwen3.5-27B – 10-12 t/s

Бенчмарки, ожидание и реальность

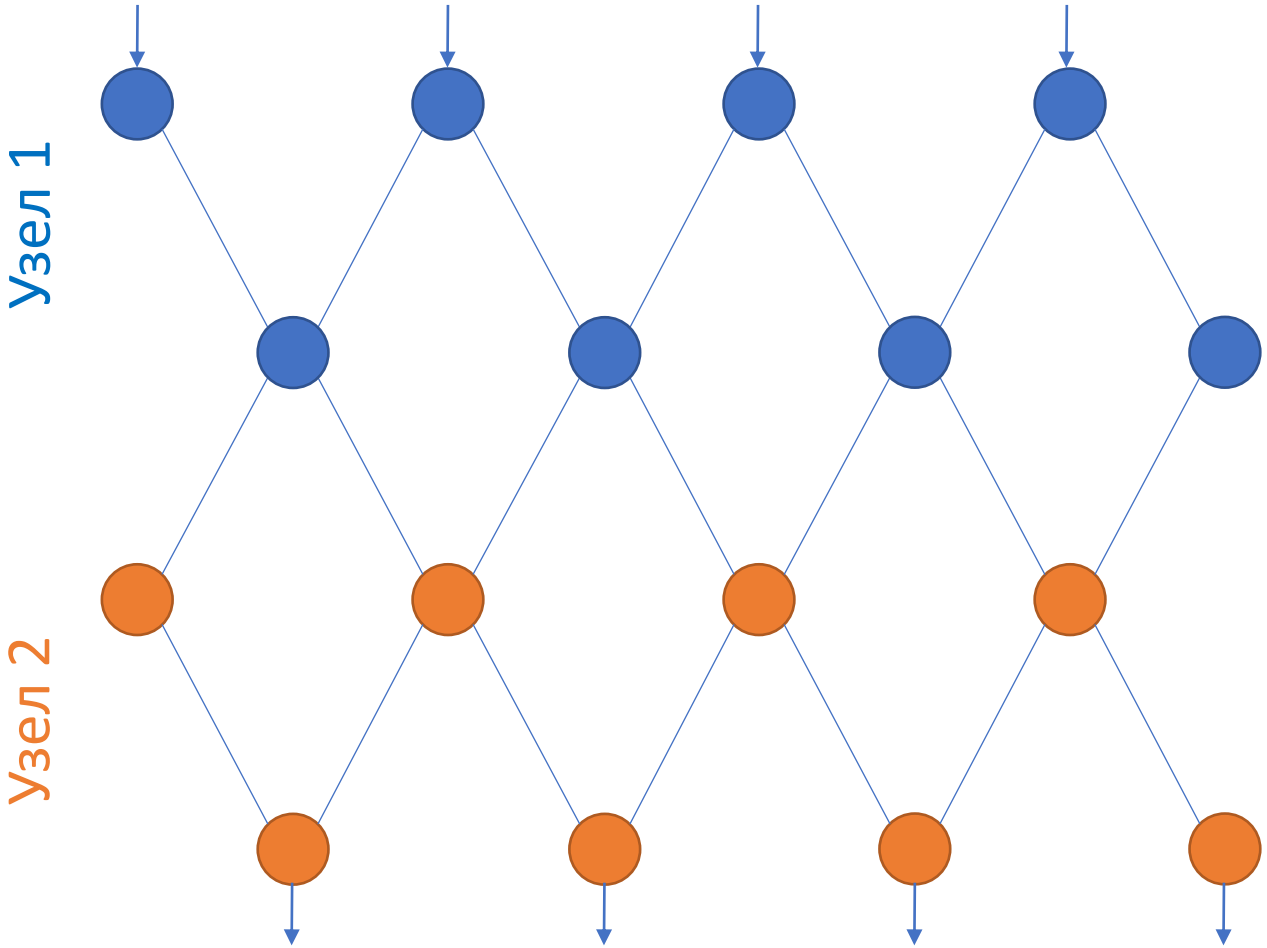
- <5 t/s— работать невозможно
- 5-20 t/s — пойдет попробовать
- 20-50 t/s — подойдет агентам
- 50-100 t/s – нормальная скорость для чата
- >100 t/s — мечта

На одиночном Ryzen 395 или DGX Spark примерно следующая производительность:

- Qwen3.5-35B-A3B – 40-80 t/s
- Qwen3.5-122B-A10B – 15-25 t/s
- Qwen3.5-27B – 10-12 t/s

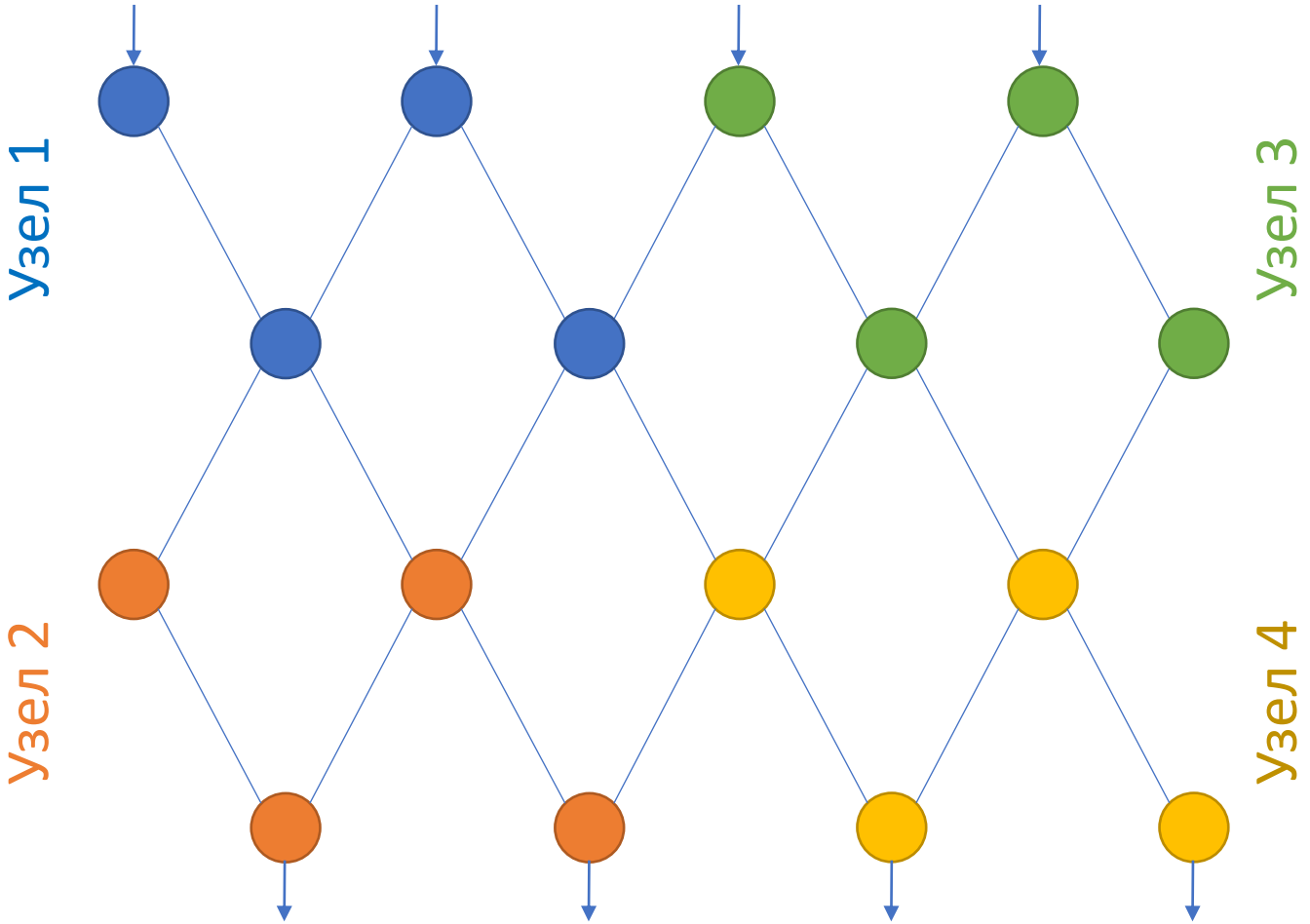
Для лаборатории пойдет, но мало.
Нужен кластер!

Pipeline (конвейерная) кластеризация



- Память общая на кластер
- Вычисления ограничены скоростью одного узла
- Если вычисления на разных слоях, то производительность растет
- Относительно слабо чувствительно к задержкам соединений

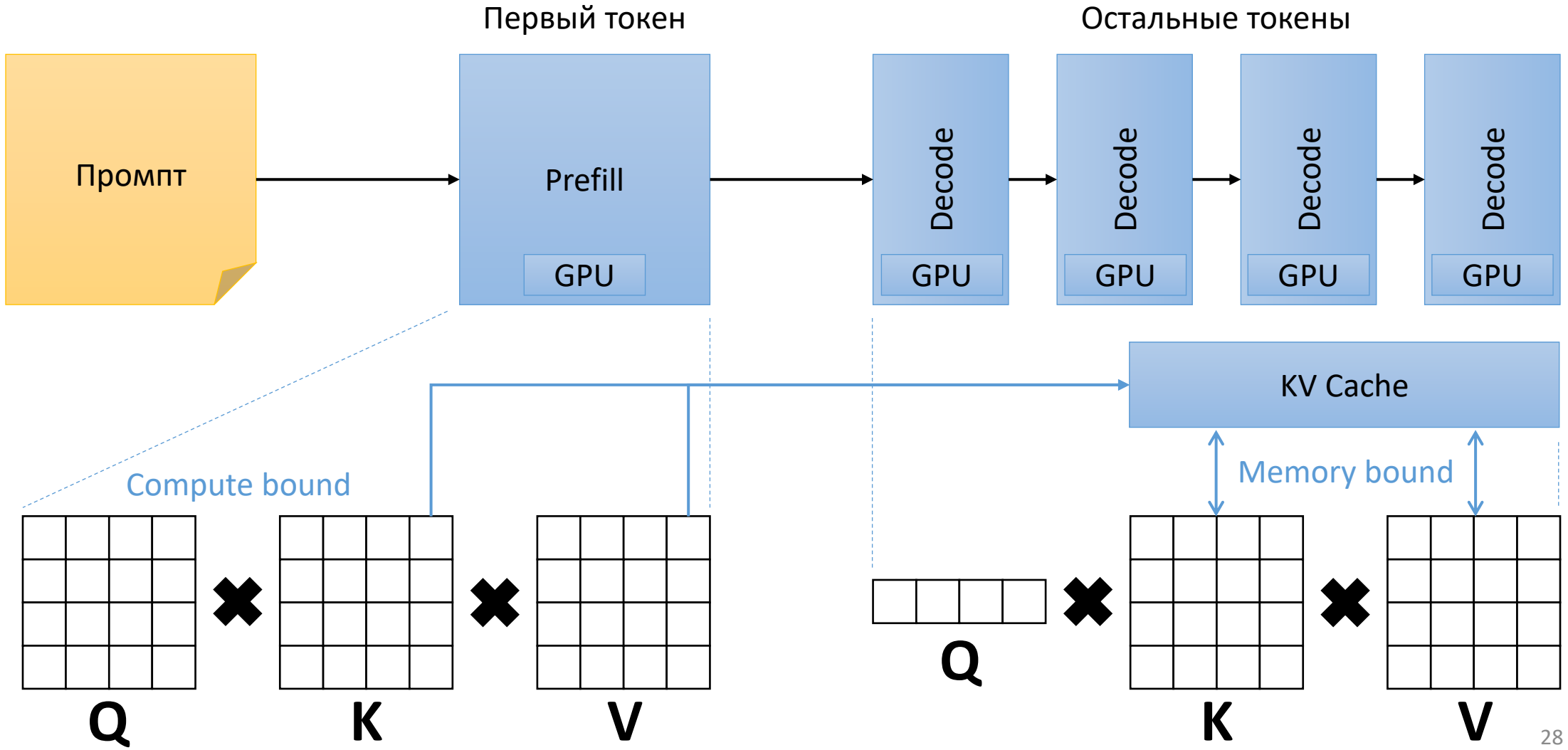
Tensor (параллельная) кластеризация



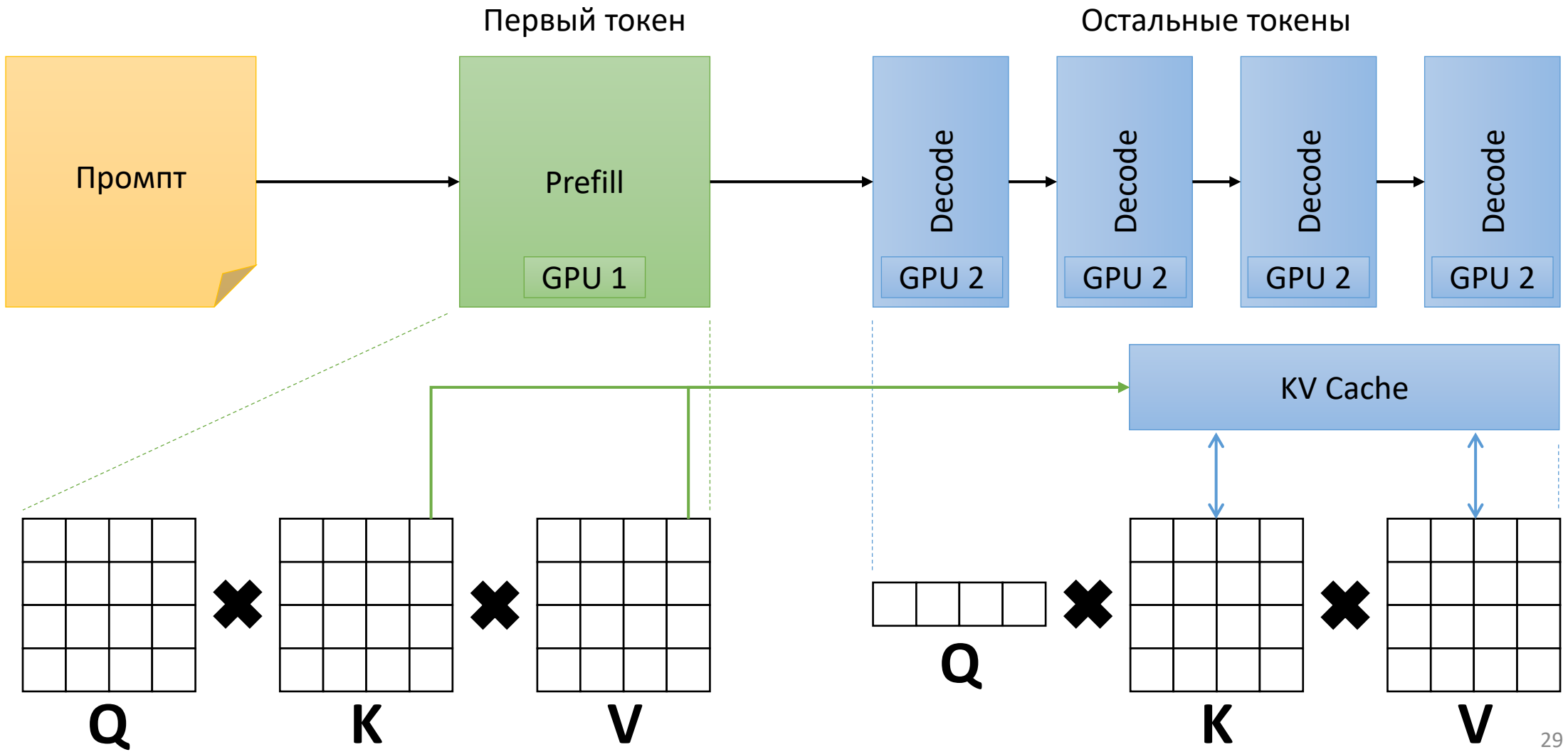
- Память общая на кластер
- Совместимо с pipeline
- Ускорение кратно количеству узлов на слое
- Крайне чувствительно к задержкам соединений
- PCI-E недостаточно в случае dGPU, нужен Infiniband или производные технологии

Ок, докинем еще 2 машины и Mikrotik

Prefill и decode

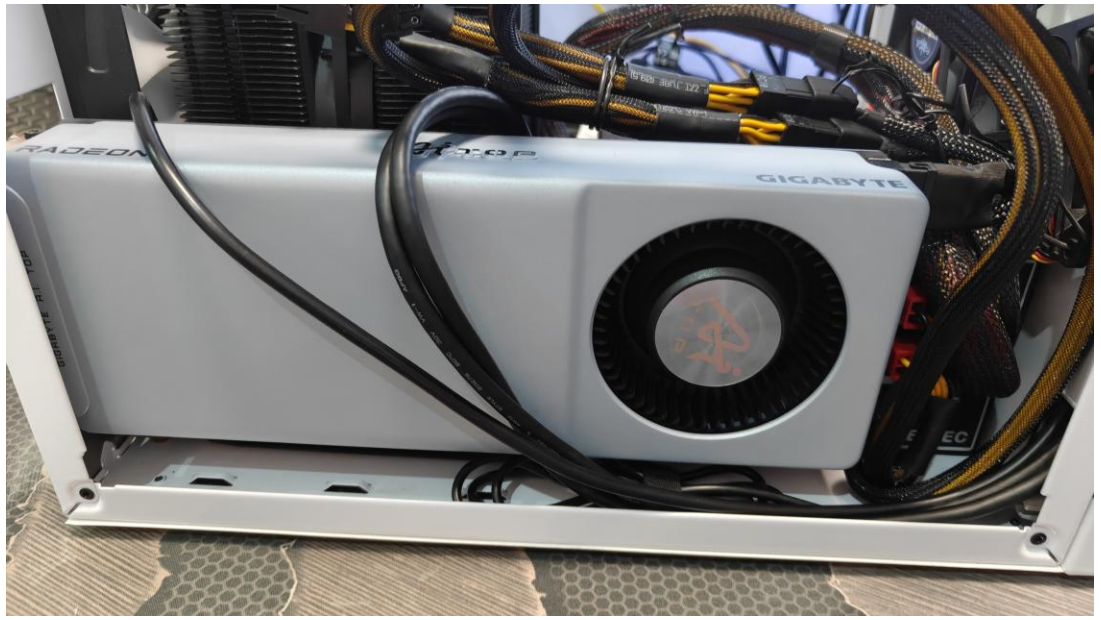


Disaggregated prefill



А еще добавим Radeon для ускорения TTFT

- **Radeon AI PRO R9700 32GB**
- Самые доступные 32GB видеопамяти на рынке
- Уровень поддержки вычислений как у 4090
- Был лишний ПК



Надо было попробовать

Пробуем в 4 машины

Запустилось!

- Qwen3.5-397B-A17B – работает!
- GLM 5 – работает!

Работают, реклама не обманывала!

Llama.cpp vs VLLM

- Все что угодно на чем угодно
- Ежедневный поток релизов
- В один поток, но быстро

- Промышленные модели и GPU
- Если запустилось, то быстро
- Нет поддержки, значит нет

Пробуем в 4 машины

Но!

- Qwen3.5-397B-A17B – 6-7 t/s
- GLM 5 – 6-7 t/s

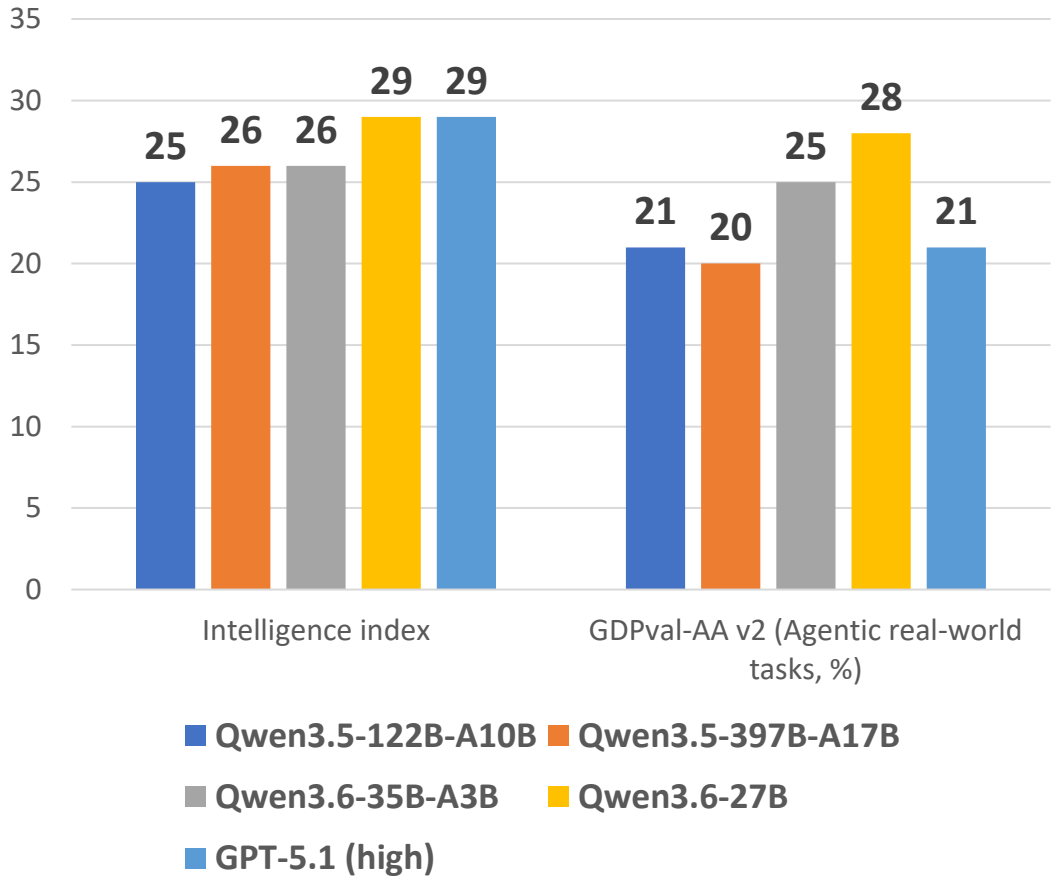
Работают, реклама не обманывала, но скорости нет

А как же ускорение?

- AMD не завезли нормальную поддержку Strix Halo в vLLM
- vLLM работает, но вдвое медленнее llama.cpp
- Llama.cpp не умеет в tensor параллелизм поверх RDMA

Пользы от второй пары машин - нет

Qwen 3.6



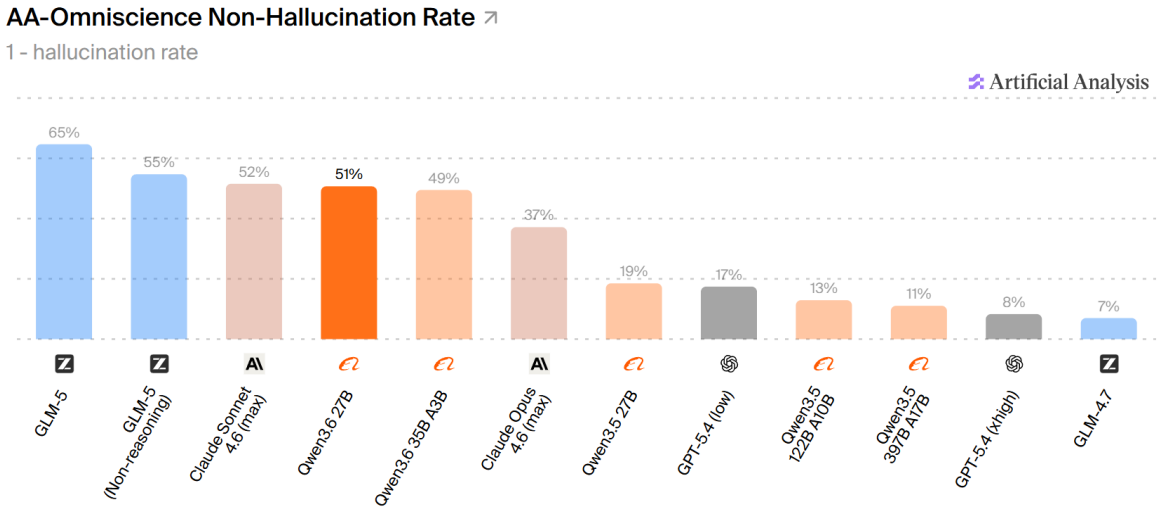
- **Qwen3.6-35B** не хуже чем **Qwen3.5-397B** при том, что он в 11,34 раза меньше
- **Qwen3.6-27B** выдает результат на уровне **GPT-5.1** без каких-то существенных скидок, будучи лучше в части тестов

По данным https://artificialanalysis.ai/

Факты

- Qwen3.6-35B-A3B умнее всех предыдущих крупных open source моделей, но работает на 1 потребительском GPU – **130-150 t/s**

- Qwen3.6-27B – практически не галлюцинирует
- Выдает до **50-60 t/s** с MTP
- Работает на 1 дискретном потребительском GPU



Нам больше не нужно много памяти для того же результата

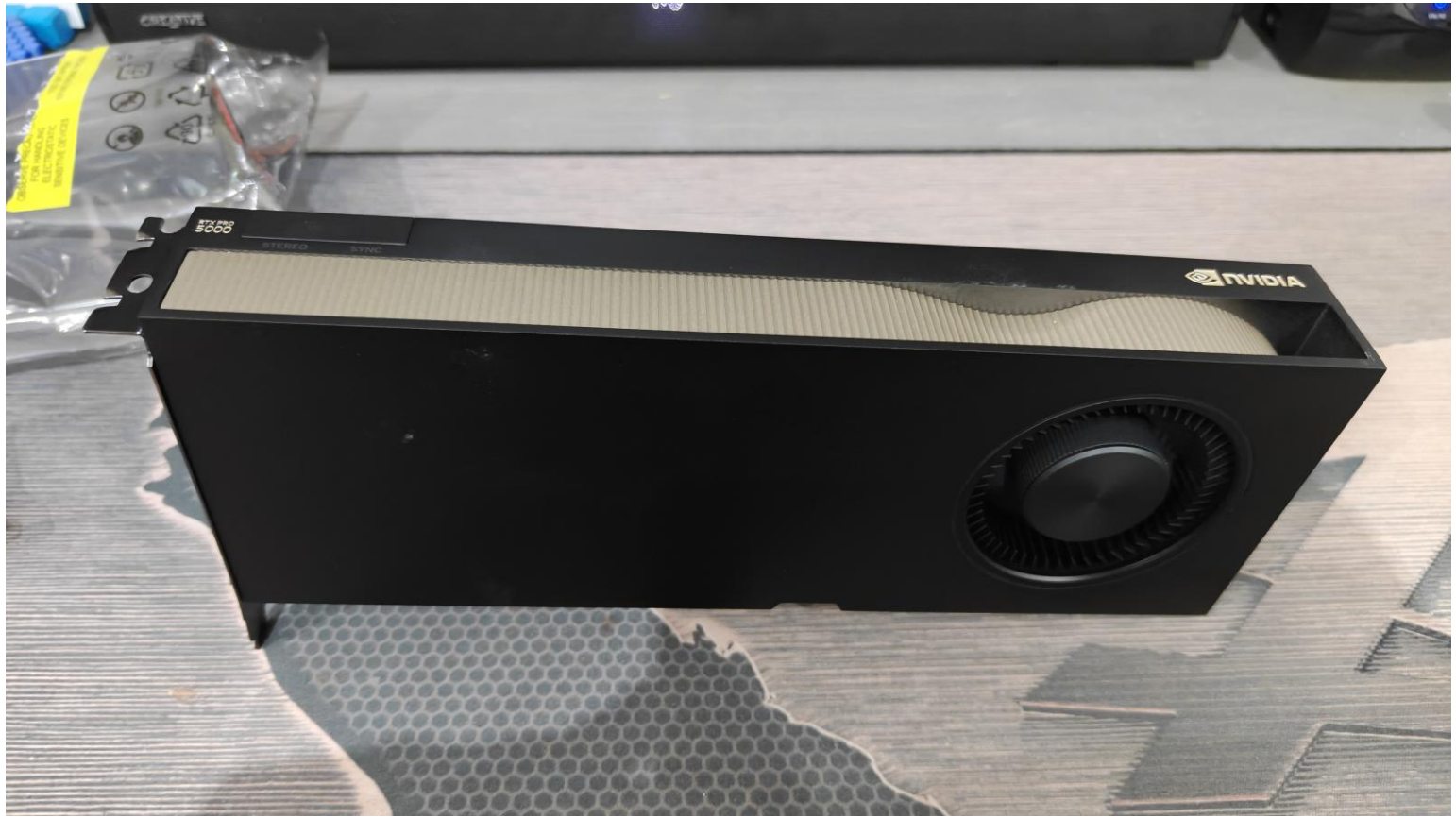
Выводы по лаборатории

- Лаборатория не нужна в имеющемся виде
- Провал не идеи, а прогресс обнулил потребность
- Технология работает, надо идти дальше

- Мини-ПК идут под обучение команды
- Идем в историю с дискретными GPU

Покупается первый RTX PRO

RTX PRO 5000 Blackwell 72GB



Щедрый комплект поставки

Почему RTX PRO Blackwell сегодня лучшие

RTX 4000/ADA и Radeon RDNA4

- Поддержка FP8, нет MXFP
- Программное масштабирование тензоров

Экономия 50% объема и пропускной способности памяти за счет FP8

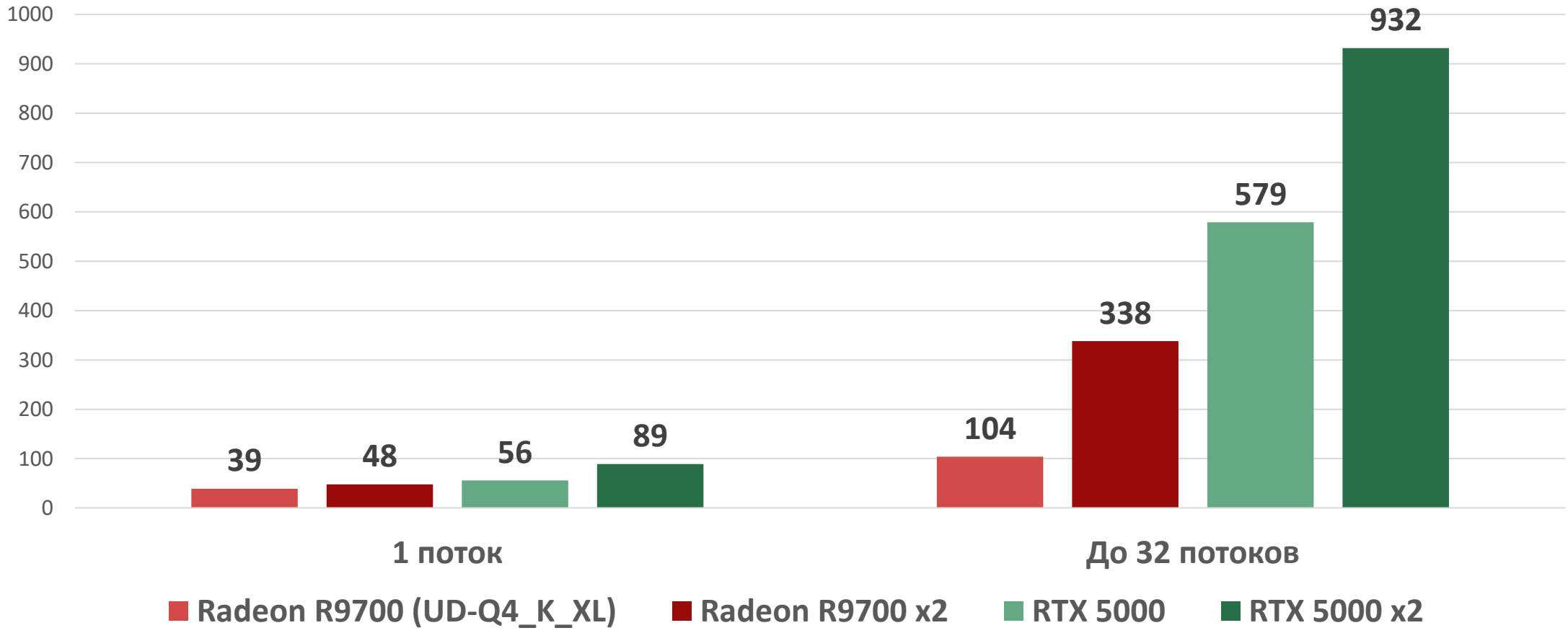
RTX 5000/PRO Blackwell

- Поддержка FP8, MXFP (NVFP4)
- Аппаратный блок масштабирования тензоров

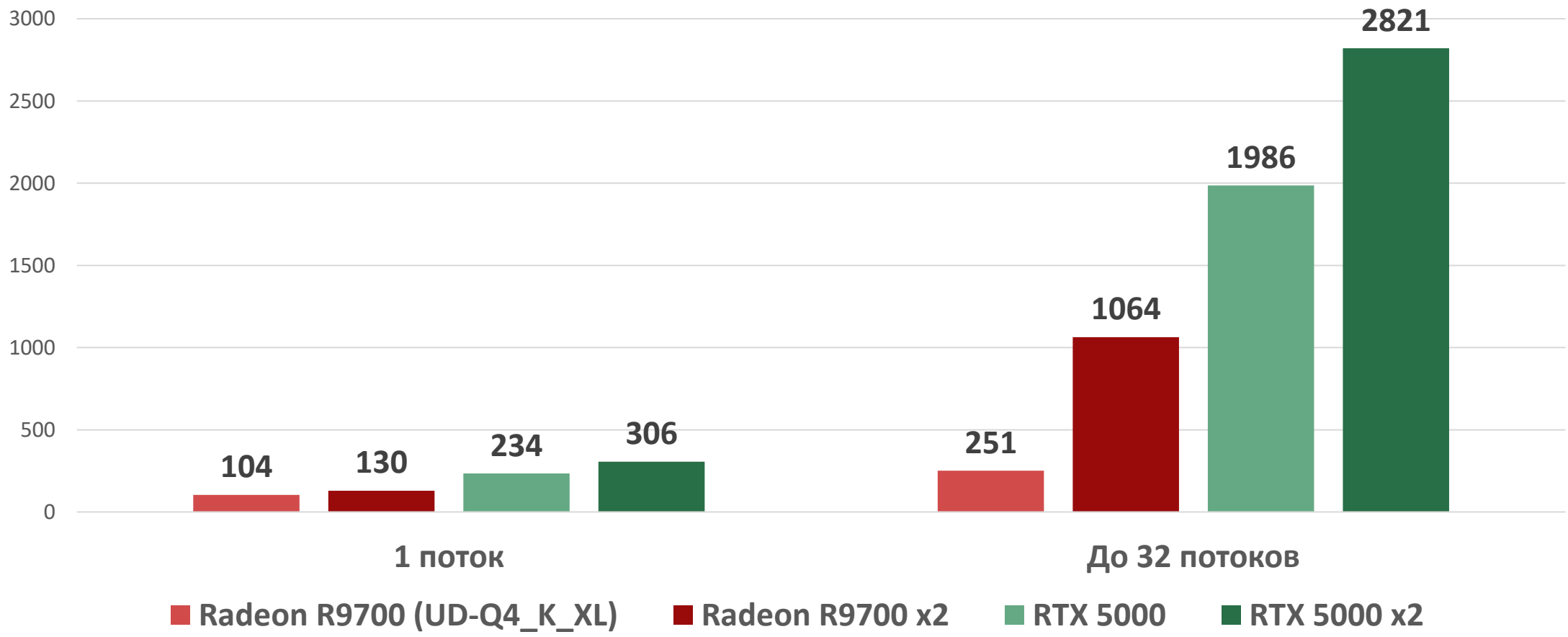
Экономия 75% объема и пропускной способности памяти за счет NVFP4

72GB = 144GB для старых GPU или 108GB относительно Ada/RDNA4

Производительность Qwen3.8-27B-FP8



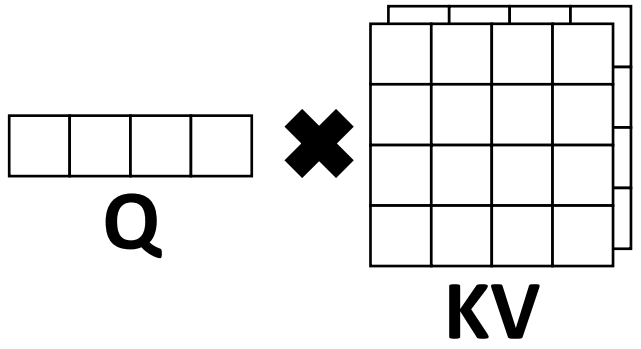
Производительность Qwen3.6-35B-A3B-FP8



Про batching в decode стадии

Одиночный запрос

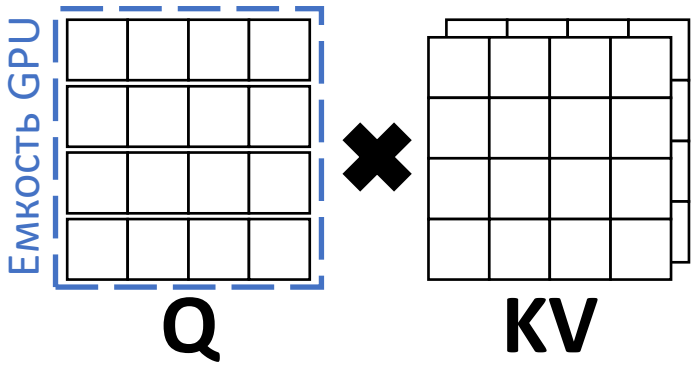
Вектор на матрицу



- Не используется полный ресурс матричных вычислителей
- Максимум скорости 1 потока
- Например, 110 t/s*

Batch n запросов

Искусственная матрица



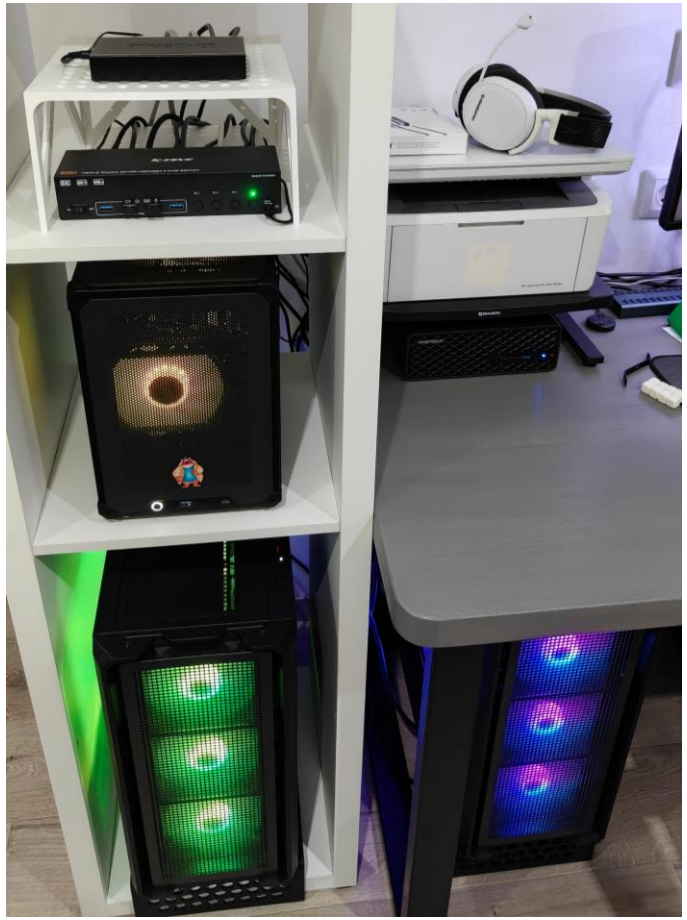
- Задействованы все вычислители
- Максимум общей пропускной способности
- Например, 700 t/s на 12 потоков, 58 t/s на поток*

* Nvidia/Qwen3.8-Flash-Next-NVFP4 на 2x RTX PRO 5000 Blackwell

На сегодня – не лаборатория, опытное производство

Люблю свою жену, она не спрашивала про стоимость

Домашний ИИ-цех



4 машины:

- 2 GPU сервера
- RAG + Application
- Рабочая станция из Strix Halo
- Плюс отдельная подсеть

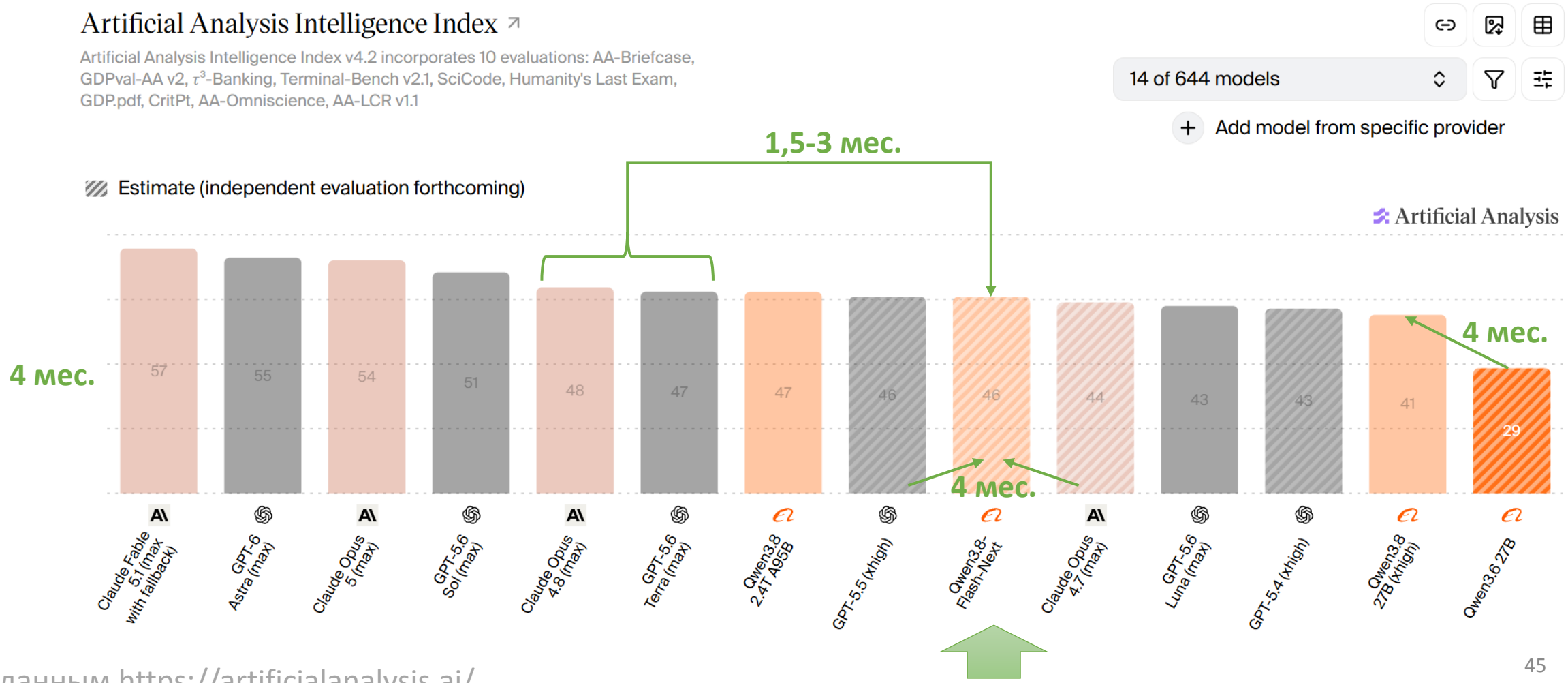
Всего на фото:

- 448ГБ RAM
- 224ГБ VRAM
- 16ТБ SSD

Текущие цели

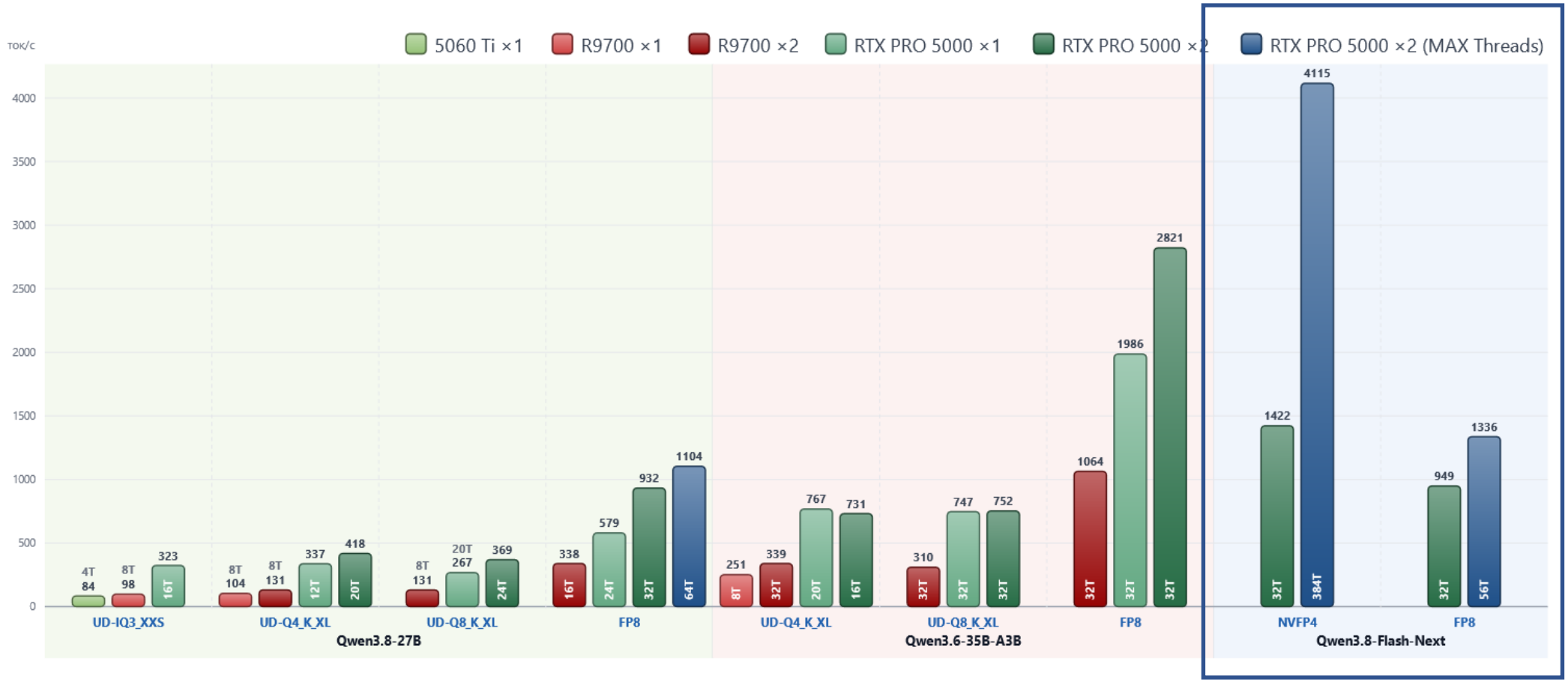
- Задача из академической перешла в практическую плоскость
- Если экспертиза максимально широка, руками работать лень, но ты можешь квалифицированно ставить задачи агентам, то перспективы перевозбуждают фантазию
- Как изменится жизнь, если дома будет фабрика разработки ПО, сопоставимая с мощностью корпоративного управления или департамента разработки?
- Личный вызов – проверить на практике

Qwen3.8-Flash-Next - фронтир локально



По данным https://artificialanalysis.ai/

Qwen3.8-Flash-Next действительно Flash



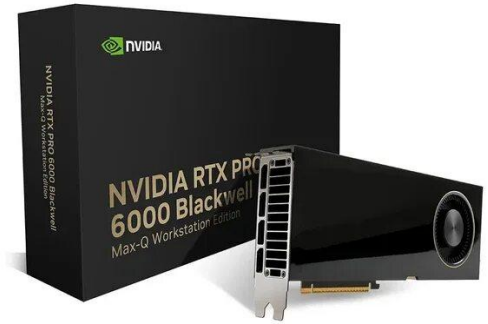
Рекомендации по конфигурациям

AMD Strix Halo – базовый минимум в паре с llama.cpp или DGX Spark, но сильно дороже, уже не минимум

- Производительность для настольного агента или исследований
- Можно использовать как классную рабочую станцию, помимо ИИ
- Киберпанк потянет, ощутимо мощнее Steam Machine в играх

AMD AI PRO R9700 x2 – комфорт для 27B моделей и поддержка FP8

- Самые выгодные 64GB VRAM на рынке – необходимый объем для Qwen3.8-27B-FP8
- FP8 работает под vLLM, давая до x3 производительности от llama.cpp
- Теплопакет и охлаждение подобраны для работы в multi-GPU сборках
- Это все еще близнец игровой 9070 XT. Можно играть в 4K и радоваться жизни



Nvidia RTX PRO 6000 x1-2 или 5000 x1-2 или 4500 x2 – ультимативная сборка

- Быстрее просто ничего нет. Совсем. И пока даже не анонсировано
- NVFP4, MXFP, FP8 с аппаратным масштабированием – лучше только версии для ДЦ
- Теплопакет и охлаждение подобраны для работы в multi-GPU сборках

Альтернативные…

Mac Studio M5 Ultra

- По текущим ценам 12000 USD за 256ГБ версию это шикарное предложение
- Скорость памяти выше только у топовых RTX Blackwell (5090, 6000, 5000)
- Добавлен FP8
- Пока не вышла, нет бенчмарков, но должно быть очень хорошо

…и спорные решения

Nvidia RTX 5090

- Слишком дорого за 32ГБ, RTX PRO 4500 дешевле
- 32ГБ допустимо, но мало для 27B моделей даже в NVFP4, контекст на минимуме
- Две карты не влезут на материнскую плату

Intel ARC Pro B70 – самые дешевые 32ГБ на рынке

- Памяти формально много, но FP8 или MXFP нет
- Хотя две карты позволят работать с 27B в vLLM в FP16, но контекст будет как на 32ГБ у 5090 – не оправдано
- Хотя поддержка ПО обещает быть даже лучше AMD

Take away

- Надо помнить, что ты инженер. Нельзя освоить новую область не работая руками
- Зачем менеджеры, если инженер ускорится в 10-100 раз?
- Работа в облаках не дает понимания глубины. Сложность скрыта
- Уже полученный результат в виде понимания деталей работы нейросетей и направления куда копать дальше оправдал вложения

«Тормозить нельзя отстанешь»

@EVIL_ABOUT_IT

«Тормозить нельзя отстанешь»

Спасибо! Вопросы?