



# Альтернативные ускорители ИИ: что показали первые тесты

**Глеб Маркевич**

Лидер направления ИИ  
и комплексных решений, YADRO



**<40%**

выполнение летних поставок  
по наблюдению за рынком<sup>1</sup>

<sup>1</sup> Рыночное наблюдение докладчика · лето 2026



## Сроки и доступный объём поставок нестабильны

Китайские производители ориентированы на  
внутренний субсидированный спрос

# Платформу проверяем вместе с площадкой



**7 кВт**



Стойка в ЦОД на Светлой

Пример из каталога Selectel

**до 220 кВт**



Система питания стойки MGX  
для DGX Rubin NVL8

220 кВт — мощность питания стойки.  
Тепловую нагрузку считаем для выбранной конфигурации.

# PCIe, HGX и OAM — разные уровни системы



**PCIe**

Интерфейс

---

Карта в слоте  
сервера

**HGX**

Платформа NVIDIA

---

Модули SXM  
и NVLink / NVSwitch

**OAM**

Формат модуля

---

Модуль на общей  
базовой плате UBB

Возможности системы определяют межсоединение и ПО

# Что можем сейчас: NVIDIA PCIe и MetaX PCIe



Цифры взяты из разных документов и не образуют единый рейтинг платформ

## 01 NVIDIA H200 NVL

4 карты = 564 ГБ суммарной памяти ускорителей  
GLM-4.5-FP8 · 355 млрд параметров  
контекст 128 тыс. токенов — 8 карт протестировано

## 02 NVIDIA RTX PRO 6000

Рабочий расчёт «скорость/цена»  
Результат зависит от модели и формата чисел  
Учтена только текущая цена ускорителей

Статья на habr  
с деталями

## 03 MetaX C500

Промышленная эксплуатация в крупном заказчике  
и тесты производительности

# Что можем сейчас: NVIDIA PCIe и MetaX PCIe



Цифры взяты из разных документов и не образуют единый рейтинг платформ



NVIDIA H200 NVL

4 карты = 564 Гб суммарной памяти ускорителей  
GLM-4.5-FP8 · 355 млрд параметров  
контекст 128 тыс. токенов — 8 карт протестировано



NVIDIA RTX PRO 6000

Рабочий расчёт «скорость/цена»  
Результат зависит от модели и формата чисел  
Учтена только текущая цена ускорителей

Статья на habr  
с деталями



MetaX C500

Промышленная эксплуатация в крупном заказчике  
и тесты производительности

# Наши тесты: скорость на цену GPU



RTX PRO 6000 относительно H100 NVL · серверный режим

Qwen3.6-35B-A3B



**+54%**

Qwen3.6-27B



**-16%**

## Условие

H100 NVL стоит в 1,5 раза дороже RTX



# MetaX C500 уже работает у заказчиков



**Есть коммерческая эксплуатация в серверах YADRO**

**$\approx 2,8 \times$**

дольше до первого токена  
и полного ответа



**$\approx 4 \times$**

ниже пакетная  
скорость генерации



**Тест**

8 C500 против 8 H100, BF16

# ОАМ — открытый формат модуля для плотных систем



Формат задаёт механику и питание; возможности определяет вся конкретная система

## Модуль

---

Ускоритель без PCIe-кронштейна высокая мощность и плотность

- Устанавливается прямо на базовую плату

## Базовая плата UVB

---

UVB — общая базовая плата для модулей

- В рассмотренной системе — 8 модулей общие питание и охлаждение
- Межсоединение зависит от производителя

## Полная система

---

Сервер, стойка и программный контур

- ОАМ ≠ HGX
- Совместимость проверяется целиком

**❗ ОАМ не гарантирует NVLink, общую память или переносимость программ. Прямая работа с производителями карт**



# Скорость засчитываем только после проверки точности

Запуск модели, измерение и эксплуатационная готовность — разные уровни доказательства.

## Шаг 1

### Корректный запуск

- Модель загружается и отвечает
- Точность проходит заданный порог
- Нет зависаний
- NaN — неверное числовое значение — отсутствует

## Шаг 2

### Измерения

- Время до первого токена (TTFT)
- Время следующего токена (TROT)
- Типичная задержка и хвосты p95/p99
- Скорость входных и выходных токенов

## Шаг 3 | Цель

### Приёмка

- 8 ускорителей — целевой режим
- 24–72 часа — целевой прогон устойчивости
- Метрики и ошибки доступны наблюдению
- После обновления тест повторяется

Заявление производителя → наш запуск → измерение → повтор → приёмка

# AIME<sup>1</sup> 2025: что и как мы измеряем



## Олимпиадные математические задачи

**30 задач**

Конечный численный ответ  
сверяем с эталоном



**16 повторов**

Долю верных ответов  
усредняем по запускам



**В наших тестах ищем потерю качества  
относительно реализации на NVIDIA**

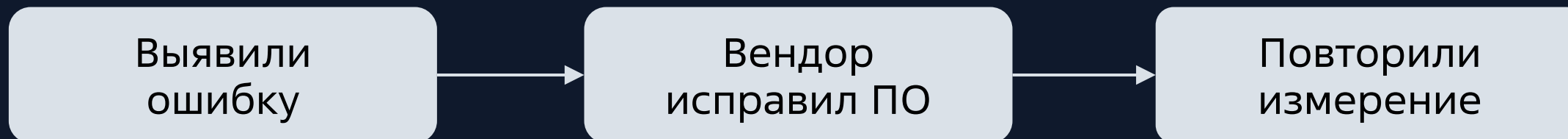
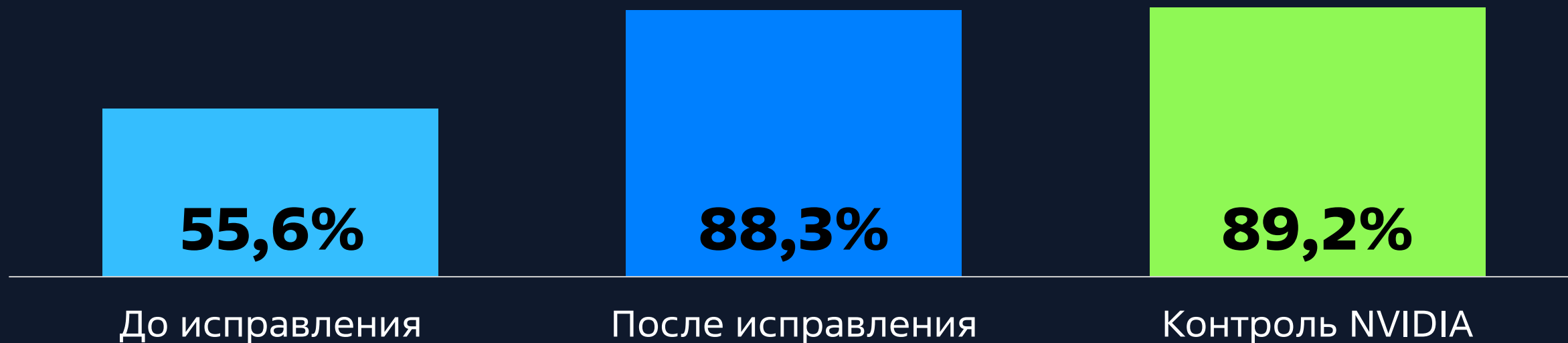


<sup>1</sup> American Invitational Mathematics Examination

# ОАММ: качество после исправления ПО



**Qwen3.6-35B-A3B, FP8 × AIME 2025**



# Длина запроса меняет результат

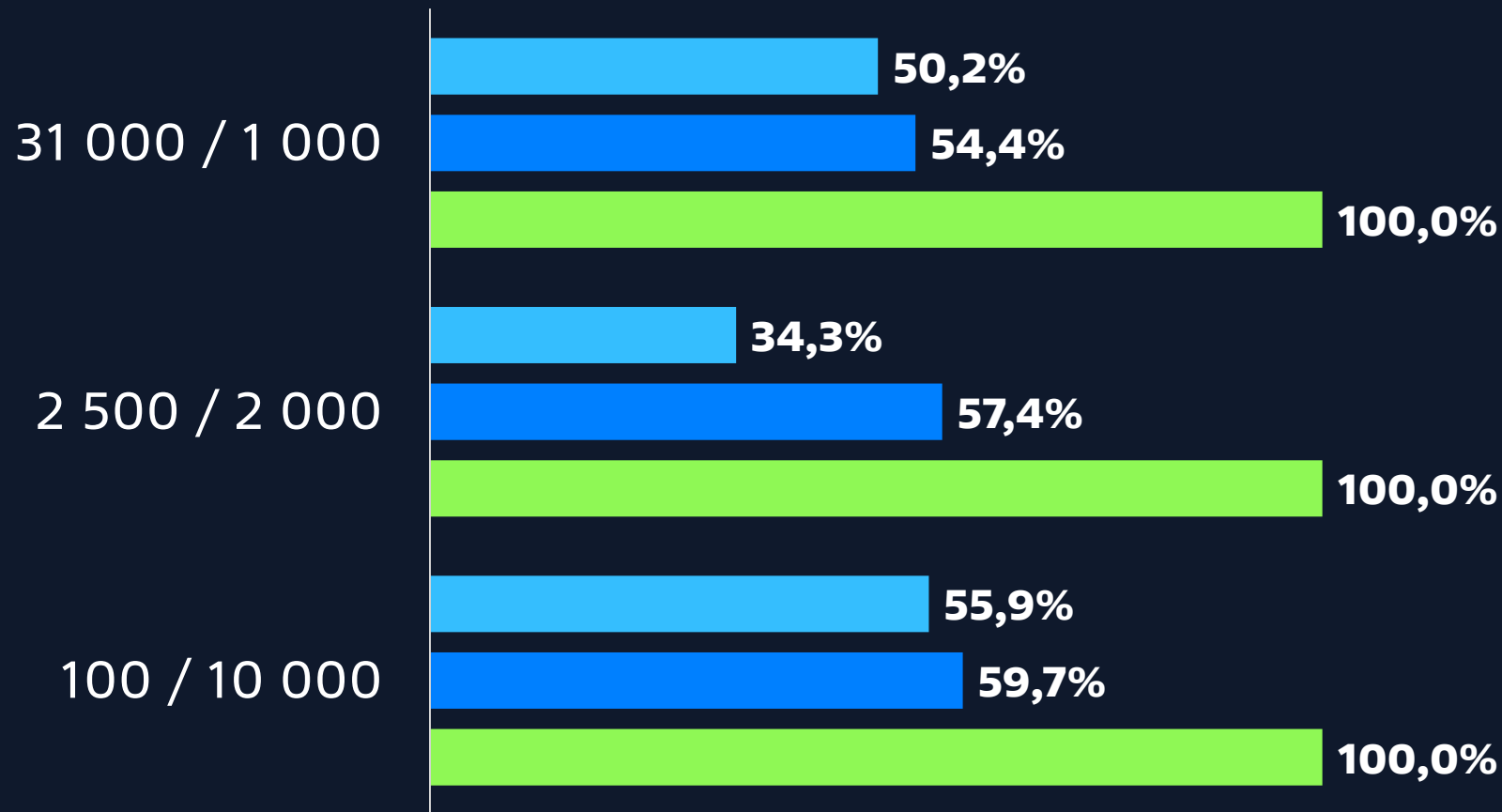


| Профиль                 | Вход, токенов<br>Включает запрос и контекст | Выход, токенов<br>Включает ответ модели |
|-------------------------|---------------------------------------------|-----------------------------------------|
| Короткий вход           | 100                                         | 10 000                                  |
| <b>Сбалансированный</b> | <b>2 500</b>                                | <b>2 000</b>                            |
| Длинный контекст        | 31 000                                      | 1 000                                   |



# Наши тесты: скорость генерации

**Qwen3.6-35B-A3B, FP8 × на 1 GPU × H100 NVL = 100%**



Legend:

- OAM A
- RTX PRO 6000
- H100 NVL

Вход / выход —  
в токенах.  
Пакетный  
режим  
генерации

# Платформу проверяем вместе с площадкой



## Простые примеры

---

Сложение векторов  
и умножение матриц  
прошли проверку

## Программа Ilt.c

---

- Потребовались ручные правки
- Один вариант дал некорректный результат

Успешного запуска мало: **проверяем результат**

# До конца 2026 — ОСНОВА КОМПЛЕКСНОГО КОНТУРА



Фиксируем проверенные слои и пилоты. Готовность каждого слоя подтверждается отдельно

05

## ЦОД-стойка YADRO

Жидкостная стойка уже есть · Питание и охлаждение согласуются с площадкой

Стойка есть

04

## Серверы: 2 GPU server · G4208P · OAM

2 GPU server — начальный класс · G4208P — средний · OAM — старший класс

- 2 GPU server — тестируем текущие сервера в новой конфигурации
- G4208P — базовый сервер
- OAM — ориентир H1 2027 OEM, H2 производства и codesign YADRO

03

## D5264 + RoCE · сеть прямого доступа к памяти

64×800GbE между узлами · Проверка потерь и задержек на рабочей нагрузке

- Ручные пилоты — 2026
- Управление LEF IS — ориентир I пол. 2027
- Сетевой процессор ASIC — 2027

02

## TATLIN.DATACUBE · слой данных

Файлы для агентов · Kafka — поток данных · KV-кэш — контекст модели · Векторное хранилище и предобработка

Первые демонстрации — 2026

01

## LEF IS · управление инфраструктурой

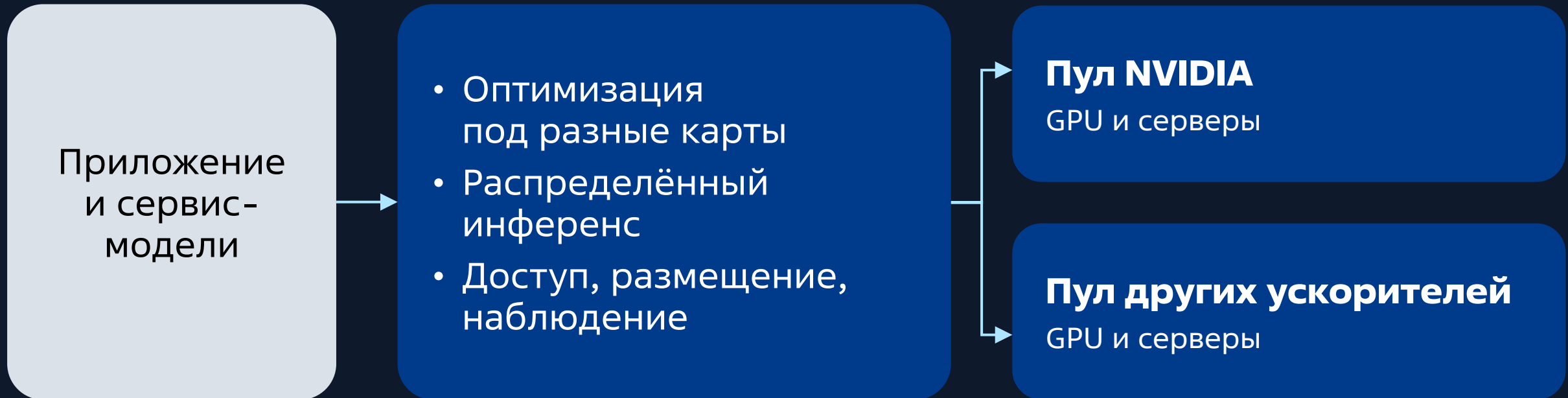
Права · Квоты · Развёртывание · Размещение · Наблюдение · Точки доступа

Первые пилоты — 2026

# LEF IS: работа с разными ускорителями



## Целевая архитектура комплексной системы



План развития: поддерживаемые сочетания проверяем тестами

# От вашей модели — к плану внедрения



Модель и профиль  
запросов



Требования к качеству  
и времени ответа



Обсудим конфигурации, испытания  
и сопровождение внедрения



@GlebMarkevich





Спасибо  
за внимание!