



Cognitar
PER ASPERA AD ASTRA

Топовые вычислительные системы для суперкомпьютинга и ИИ

Валерий Егоршев
КОГНИТАР

Москва, июль 2026

1.1. Планы и прогресс в строительстве ИИ-Дата-центров в мире, мировой рейтинг суперкомпьютеров ТОП-500

Свежий рейтинг Top-500*, июнь 2026 года: первые пять систем и 500-я система

№	Система	Rmax (ПФлопс)	Rpeak (ПФлопс)	GPU (штук)	Мощность (МВт)	Шкафов (шт)
1	LineShine - LingKun, CPU LX2 304C 1.55GHz, Interconnect LingQi, Kylin OS, Shenzhen Cloud Computing Center, NSCS, China	2,198.40	2,735.82	41K (CPU)	42	90
2	El Capitan - HPE Cray EX255a, CPU&GPU AMD MI300A, LLNL, USA	1,809.00	2,821.10	43K	30	87
3	Frontier - HPE Cray EX235a, CPU AMD GPU AMD MI250X, ORNL, USA	1,353.00	2,055.72	38K	25	74
4	Aurora - HPE Cray EX, CPU Intel, Intel GPU Max, ANL, USA	1,012.00	1,980.01	64K	39	166
5	JUPITER Booster - BullSequana XH3000, CPU&GPU NVIDIA GH200 Superchip, Juelich, Germany	1,000.00	1,226.28	24K	16	125
500	ThinkSystem HR650X - Lenovo, CPU Intel, No GPU, China	2.66	5.18	0	Нет данных	Нет данных

* Рейтинг TOP-500 публикуется два раза в год: на конференциях ISC (июнь, Европа) и SC (ноябрь, США)

1.2. Планы и прогресс в строительстве ИИ-Дата-центров в мире, LineShine - суперкомпьютер №1 рейтинга ТОП-500

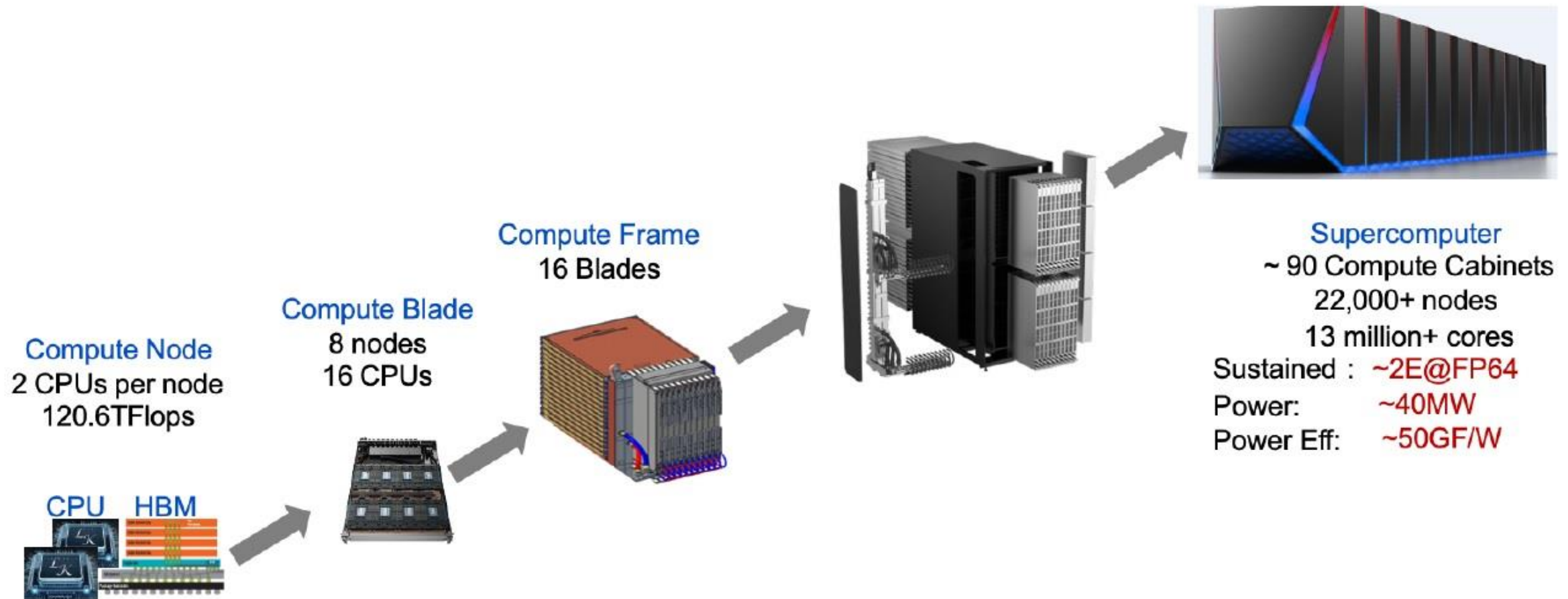


Source: NSCS

1.2a. Планы и прогресс в строительстве ИИ-Дата-центров в мире, LineShine - суперкомпьютер №1, внутреннее устройство



Cognitar



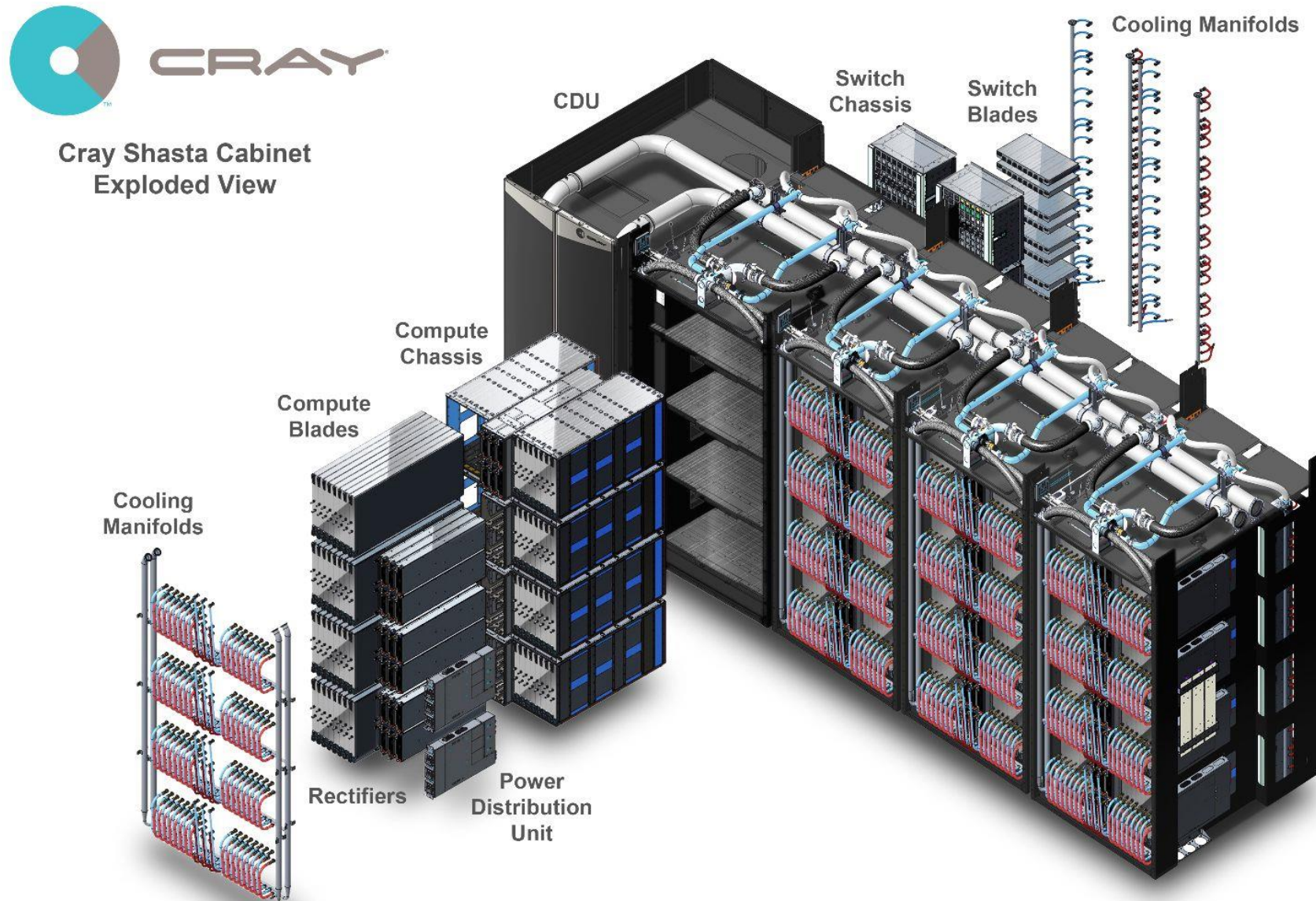
Source: Jack Dongarra

1.3. Планы и прогресс в строительстве ИИ-Дата-центров в мире, El Capitan - суперкомпьютер №2 рейтинга ТОП-500



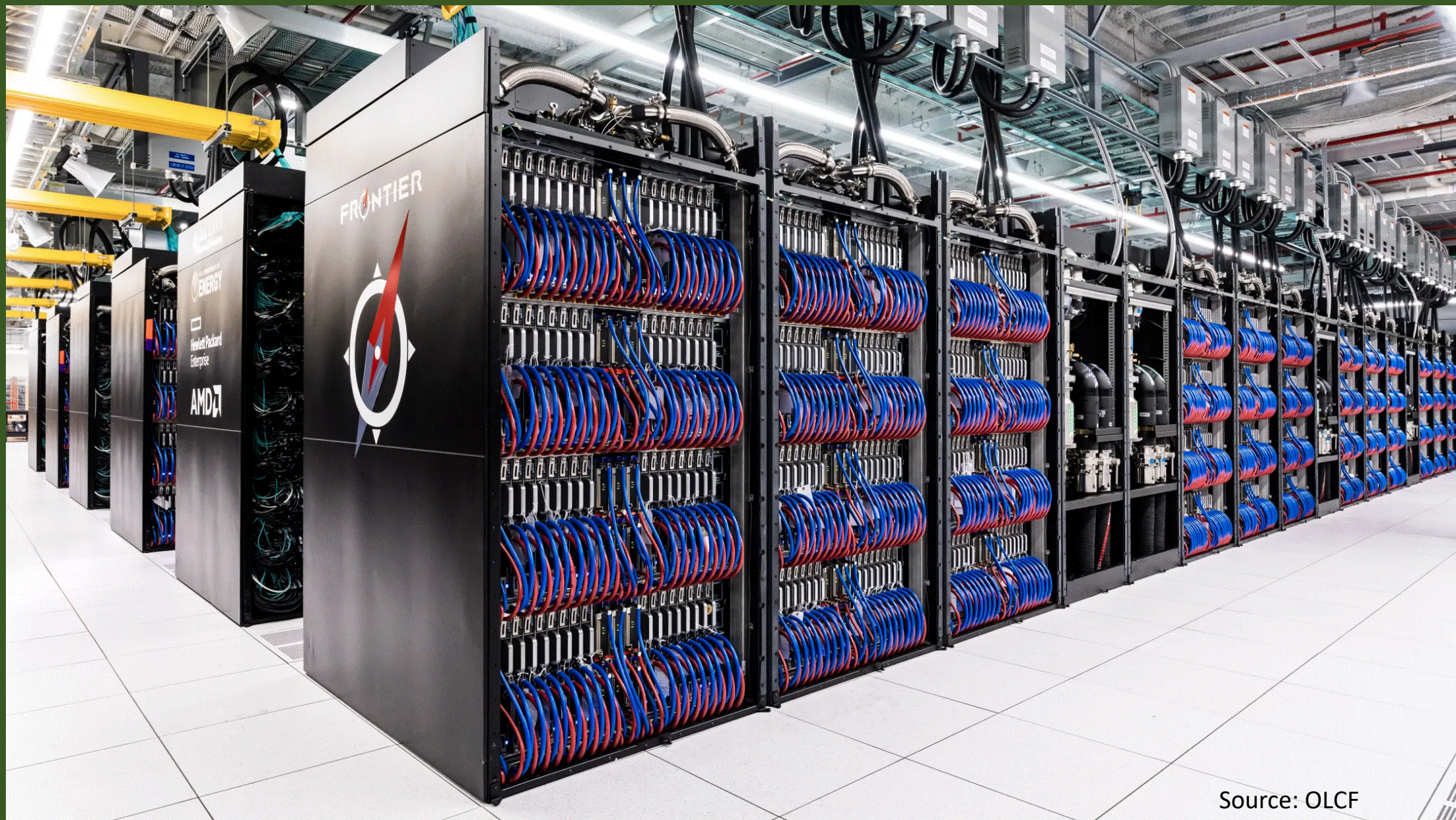
Source: LLNL

1.4. Планы и прогресс в строительстве ИИ-Дата-центров в мире, El Capitan - суперкомпьютер №2, устройство вычислительных шкафов



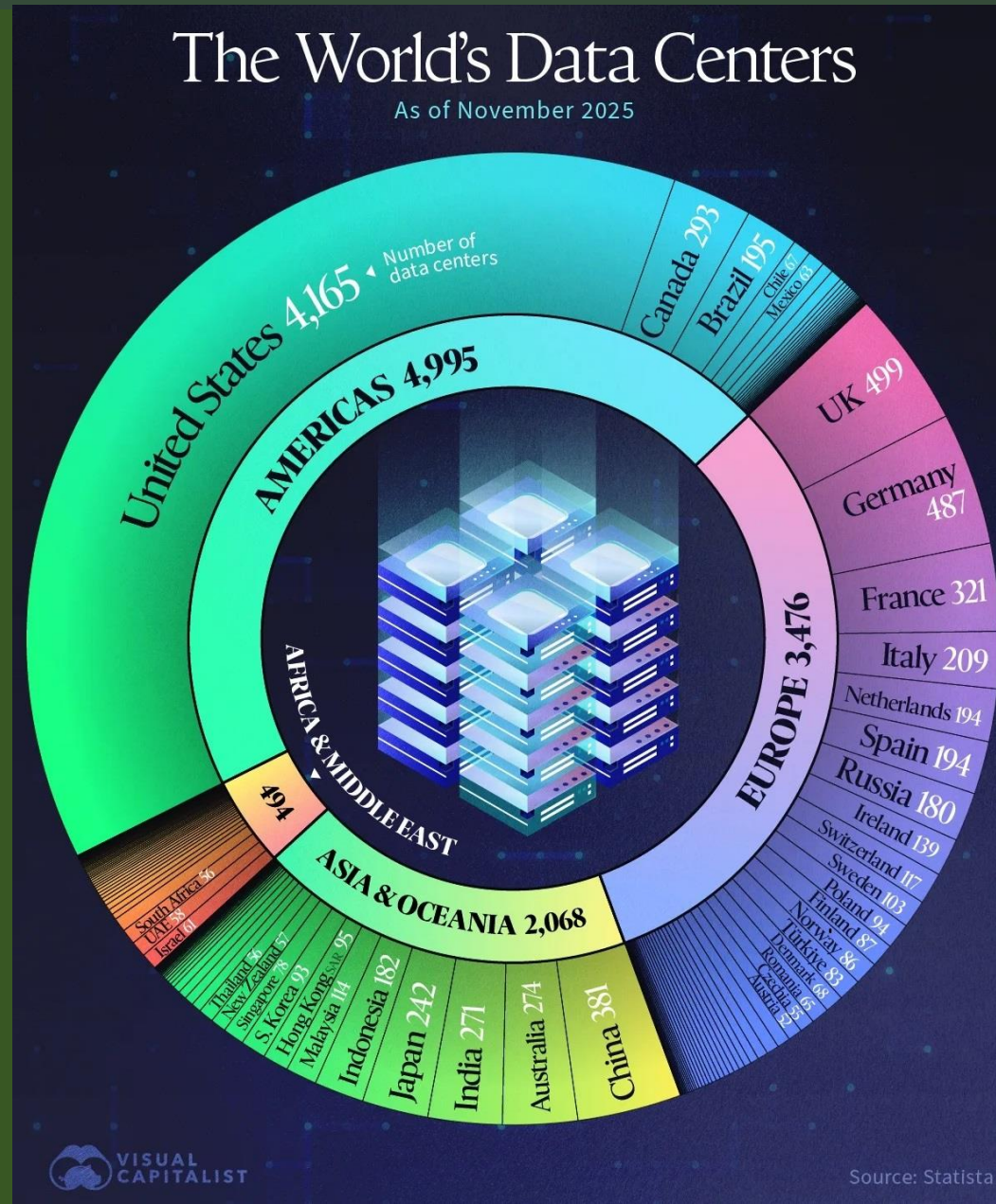
Source: DOE

1.6. Планы и прогресс в строительстве ИИ-Дата-центров в мире, Frontier - суперкомпьютер №3 рейтинга ТОП-500 (без передних панелей)



Source: OLCF

1.6. Планы и прогресс в строительстве ИИ-Дата-центров в мире, количество дата-центров всех типов по странам



1.7. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Соединенные Штаты: Stargate, Colossus и др.

Stargate (OpenAI, Oracle, SoftBank) – уже в 2025г. 7 ГВт - 400 млрд. Долл (как часть 10ГВт – 500млрд.долл.), в Абилене, Техас. И уже в течении ноября OpenAI озвучили новый рубеж - 33ГВт и 1.4 триллион. долл. новых инвестиций (с партнерами: Nvidia, AMD, Broadcom, Oracle.)

Colossus (xAI's Илона Маска) в Мемфисе, Теннесси из двух этапов: Colossus 1 на 100 000 GPU , и затем расширение до 200 000 GPU с энергопотреблением порядка 200МВт. Вторая очередь Colossus 2, будет потреблять 1ГВт , стартует в 2026 году и будет иметь порядка 1 миллиона GPU*.

ИИ-кампус **Fairwater Микрософта** в Маунт-Плезант, Висконсин: стартовые инвестиции 7млрд. Долл., старт в 2026, «сотни тысяч GPU», единая плоская сеть с очень низкой латентностью, стойки nVidia GB200 NVL72, мощность 2.2ГВт.

Project Rainier Амазона – ультра-кластер для ИИ для компании Anthropic на 11 млрд. долл. Кластер на базе 500 000 чипов Trainium2, соединенных в единую систему, мощностью 2.2ГВт. Объявлен уже следующий этап: 15 млрд. долл, и 3ГВт.

Планы **Google** – три дата-центра в Техасе к 2027 г, Google AI Hub в Индии на 15 млрд. долл. и начальной энергетикой в 1ГВт.

*Прим. * Это некоторые «средние GPU типа H100», на практике же используются и более новые поколения.*

1.8. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Оценочные коэффициенты перевода: гигаватты—деньги—производительность



1. Все ИИ-дата-центры базируются в части производительности на «ускорителях» (GPU, TPU etc.)
2. «Средний массовый ускоритель» типа nVidia H200/B200 может быть оценен:
 - a. По приведенной потребляемой мощности - в 1кВт
 - b. По приведенной цене – в 50К долл.
 - c. По «Суперкомпьютерной» производительности FP64 - в 30ТФлопс*
 - d. По «ИИ-тензорной» производительности FP8 (FP4) - в 4-9 ПетаФлопс
3. Следовательно, оценочно, сегмент дата-центра в 1 гигаВатт (например Colossus2 xAI) это:
 - a. Около 1 млн. штук «ускорителей»
 - b. Примерно 50 млрд долл по вложениям
 - c. Примерно 30 ЭкзаФлопс FP64 (например, если сравнивать с суперкомпьютерами ТОП-500)
 - d. (это скорее «рекламная информация») диапазон 4-9 ЗеттаФлопс FP8 (FP4) производительности

*Прим. * Последнее поколения GPU nVidia - B300, Google TPU, Intel GAUDI, ускорители AWS и Microsoft не имеют FP64-блоков, и у них производительность FP64 не нормируется, и примерно равна 1/64 от FP32.*

1.9. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Проект Stargate, Абелин, Техас



Cognitar



Source: Construction is under way on a Project Stargate AI infrastructure site in Texas © Daniel Cole/Reuters

1.10. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Проект Stargate, Абелин, Техас



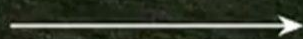
Старгейт:
10 ГВт,
10 млн.GPU

1.11. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Проект Colossus, Мемфис, Теннесси



Colossus: Total GPUs: **200,000**

Phase 1: **122 days** – 100k
GPUs fully training
synchronously. From
scratch.



Phase 2: **92 days** to expand
to 200K GPUs

1.12. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Проект Colossus, Мемфис, Теннесси



Cognitar



Source: SELC

1.13. Планы и прогресс в строительстве ИИ-Дата-центров в мире. ИИ-кампус Fairwater Микрософта в Маунт-Плезант, Висконсин



1.14. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Project Rainier Амазона в Индиане



1.15. Планы и прогресс в строительстве ИИ-Дата-центров в мире. ОАЭ, Китай, Европа

Stargate в ОАЭ (Абу-Даби): G42 из Абу-Даби (совместно с SoftBank, Nvidia, OpenAI, Cisco и Oracle) планирует построить в ОАЭ кампус искусственного интеллекта «Stargate» мощностью 5 ГВт. Первые 200 МВт должны быть введены в эксплуатацию в 2026 году, как начало этапа 1 ГВт.

Китай: Китай ускоряет строительство собственных центров обработки данных ИИ. Например, Шанхай планирует к 2025 году построить пять новых крупных центров обработки данных ИИ, которые добавят более 100 эксафлопс вычислительной мощности. (Все ресурсы сейчас они оценивают в 246 эксафлопс).

Европа: В Великобритании QTS (Blackstone) заложила фундамент кампуса стоимостью 10 млрд. фунтов в Нортумберленде, Microsoft инвестирует 22 млрд. фунтов в инфраструктуру ИИ-дата-центров. Кроме того, крупные проекты по всей Европе: Microsoft Zurich-West за 400млн. долл в Швейцарии, проект Blackstone в Арагоне, Испания будет иметь 300 МВт, Google строит дата-центр ИИ в Бельгии в 1 млрд.евро.

1.16. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Проект Stargate UAE



Cognitar



Source: <https://pressreleasehub.pa.media>

1.17. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Самый большой в Азии Шанхайский ИИ-дата-центр



1.18. Планы и прогресс в строительстве ИИ-Дата-центров в мире.
Будущий ИИ-кампус компании QTS в Нортумберленде, Великобритания



1.19. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Инвестиции в ИИ-дата-центры по гиперскейлерам

Big Five Hyperscaler Annual CapEx

Company	2023	2024	2025	2026 (Forecast)
Google/Alphabet	~32bn	~46bn	91.4bn	175-185bn
Microsoft	~35bn	~65bn	94bn (FY26)	~110bn
Meta *	~28bn	39.2bn	71.8bn	115-135bn
Amazon	~52bn	~77bn	131.8bn	200bn
Oracle	~7bn	~10bn (FY24)	21.2bn (FY25)	45-50bn (FY26)
TOTAL	~154bn	~237bn	~410bn	~645-680bn

1.20. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Обоснование инвестиций: RPO - запланированные доходы на 4-5 лет

Table 1: Estimated RPOs top \$1.5 trillion from a handful of AI labs

Hyperscaler	Estimated RPO	Composition, circularity and timelines
Microsoft	\$625bn	OpenAI accounts for 45% (\$281.3bn) of revenue backlog, including 25% within 12 months (2.5 years weighted duration). Microsoft's \$13bn OpenAI investment has circled back in Azure fees for GPT training.
Oracle	\$523bn	OpenAI represents \$300bn, with the bulk of the balance from Meta, NVIDIA and xAI deals. Around 25% due within 12 months. Hyperscaler investments circle back as Oracle Cloud Infrastructure (OCI) fees for AI training.
Google Cloud	\$155bn	Nine out of 10 top AI labs use Google Cloud to build and run their AI models. Around 20% due within 12 months.
AWS	\$160-180bn	Anthropic is an AWS customer, who is a minority owner. Amazon's \$8bn Anthropic investment circles back as AWS cloud fees for training Anthropic's Claude. Stability AI is also a significant AWS customer. Around 25% due within 12 months.
Total	~\$1.5T	Hyperscalers' RPOs are heavily concentrated on few AI labs, with circular investments amplifying risks.

Source: Earnings reports and calls

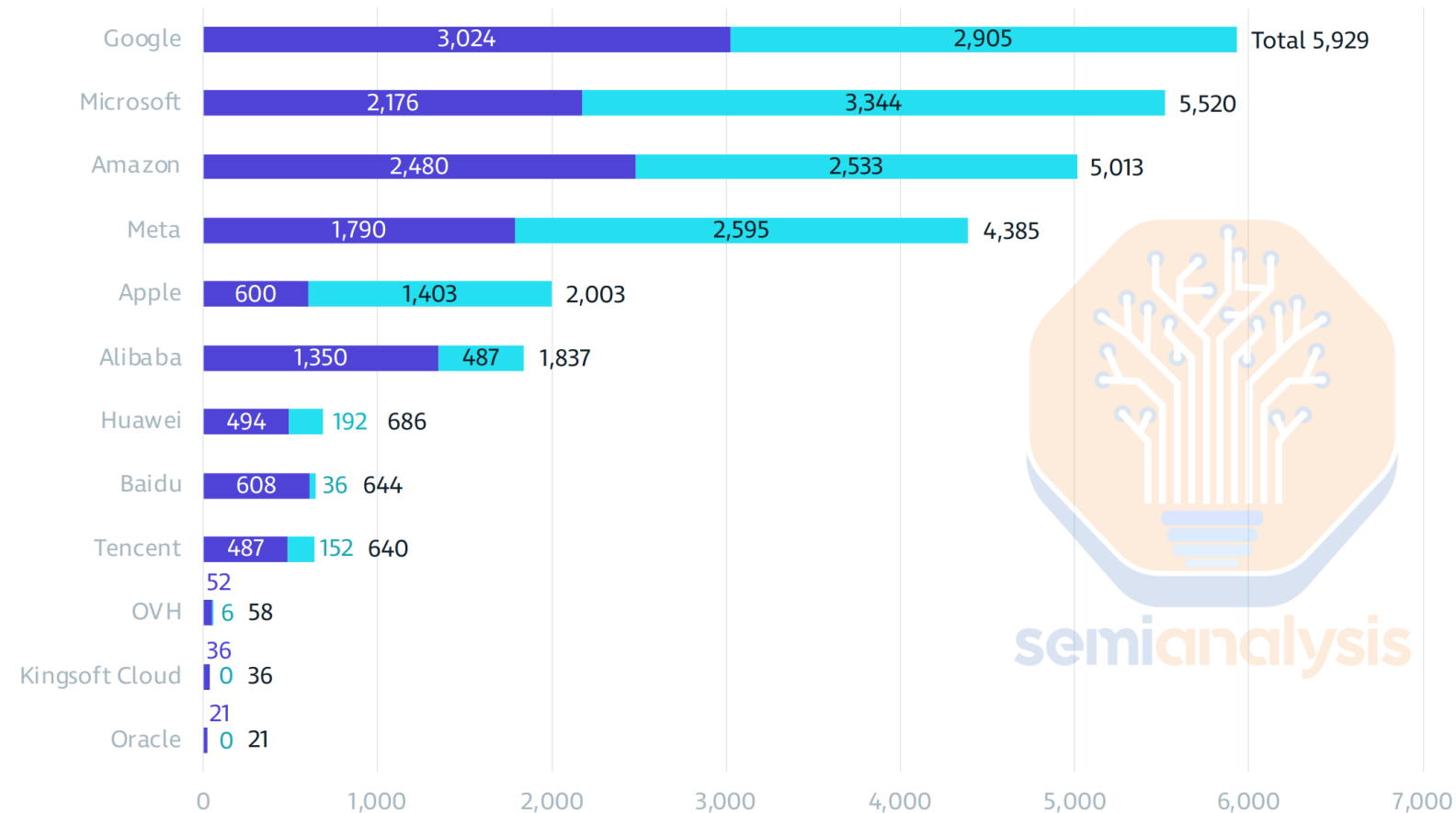
1.21. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Текущие мощности дата-центров по гиперскейлерам

HYPERSCALE SELF BUILD REPORT JANUARY 2023

COPYRIGHT STRUCTURE RESEARCH 2023

GLOBAL: SELF-BUILD DATA CENTRE CAPACITY SPLIT BY HYPERSCALER

IN MEGAWATTS [MW] OF CRITICAL IT POWER CAPACITY



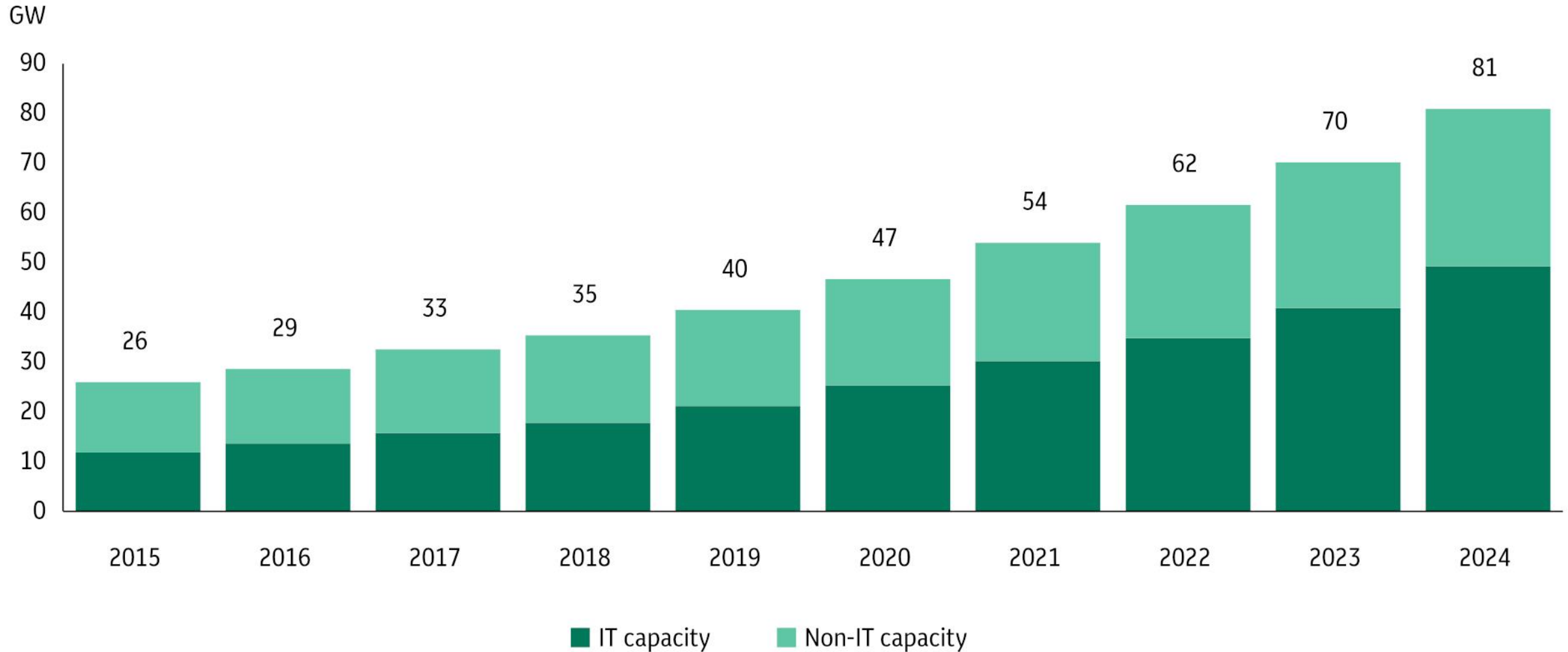
TOP 4 HYPERSCALERS REPRESENT

78%

OF TOTAL GLOBAL SELF-BUILD CAPACITY

- 2022 Operational Data Centre Capacity
- Future - Capacity Under Development and Land Banks

1.22. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Глобальная мощность дата-центров по прошедшим годам

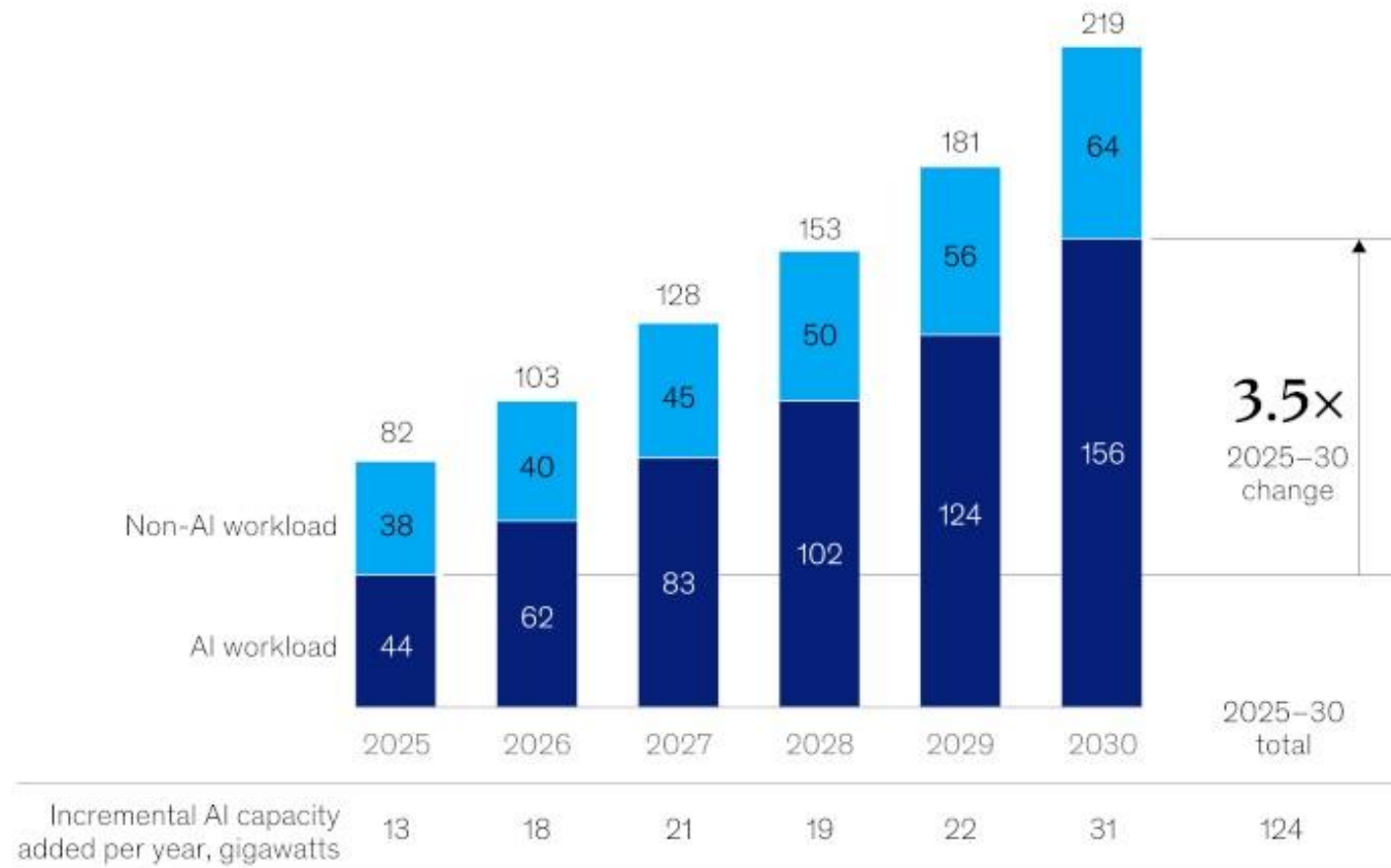


Sources: Uptime Institute, "Rack Density is Rising", December 2020; Omdia, "Data Center Racks Market Analysis – 2023", December 2023; Uptime Institute, "Uptime Institute Global Data Center Survey 2024", July 2024; Macquarie Asset Management analysis.

1.23. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Прогноз роста глобальной мощности ИИ-дата-центров

Both AI and non-AI workloads will be key drivers of global data center capacity demand growth through 2030.

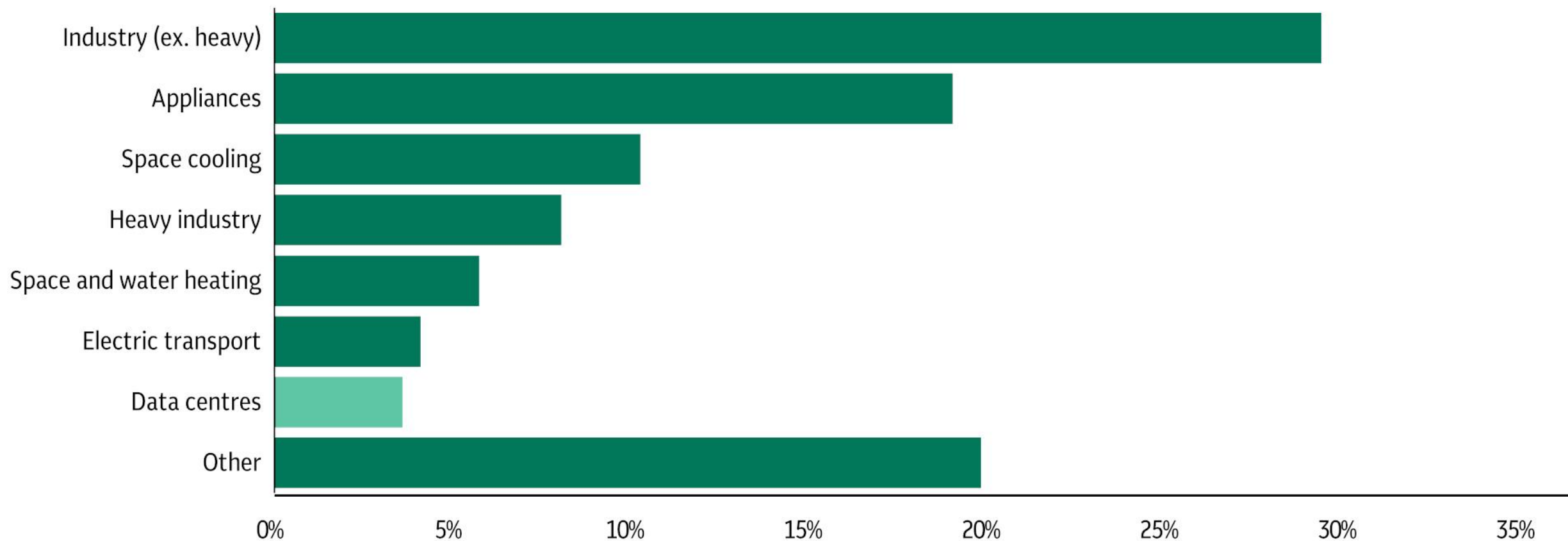
Estimated global data center capacity demand, 'continued momentum' scenario, gigawatts



Note: Figures may not sum to totals, because of rounding.

Source: McKinsey Data Center Demand Model; Gartner reports; IDC reports; Nvidia capital markets reports

1.24. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Сравнение всех потребителей



1.25. Планы и прогресс в строительстве ИИ-Дата-центров в мире. Парадокс Джевонса

**"Что даёт Энди, то отнимает Билл"
(What Andy giveth, Bill taketh away).**

«Закон анти-Мура (The Great Moore's Law Compensator)» или закон Дэвида Мея (David May): "Эффективность программного обеспечения уменьшается вдвое каждые 18 месяцев, что компенсирует закон Мура".

Если строго, то так - механизм действия парадокса Джевонса:

- технологический прогресс повышает эффективность использования ресурса;
- повышение эффективности снижает стоимость использования ресурса;
- снижение стоимости стимулирует рост спроса на ресурс;
- рост спроса приводит к увеличению общего потребления ресурса.

1.26. Планы и прогресс в строительстве ИИ-Дата-центров в мире.

Литература:

1. Top-500, the List: <https://top500.org/lists/top500/2025/11/>
2. AI Data Centers: The Megastructures Powering the AI Revolution:
<https://streamlinefeed.co.ke/news/ai-data-centers-the-megastructures-powering-the-ai-revolution-13c7856b>
3. Data centres: Powering the growth of AI and cloud computing
<https://www.macquarie.com/uk/en/about/company/macquarie-asset-management/institutional-investor/insights/data-centres-powering-the-growth-of-ai-and-cloud-computing.html>
4. Inside the world's most powerful AI datacenter:
<https://blogs.microsoft.com/blog/2025/09/18/inside-the-worlds-most-powerful-ai-datacenter/>
5. The cost of compute: A \$7 trillion race to scale data centers:
<https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers>
6. Part One: AI's \$7 trillion infrastructure build-out and the emerging fault lines:
<https://followtherisk.substack.com/p/part-one-ais-7-trillion-infrastructure>
7. Part Two: Hyperscalers' \$680bn AI capex investment raises the stakes:
<https://followtherisk.substack.com/p/part-two-hyperscalers-680bn-ai-capex>
8. Global Infrastructure And Project Finance: Key Data Center Financing Takeaways From PPIF 2026:
<https://www.spglobal.com/ratings/en/regulatory/article/global-infrastructure-and-project-finance-key-data-center-financing-takeaways-from-ppif-2026-s101668747>
9. The cost of compute: A \$7 trillion race to scale data centers:
<https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers>

2.1. Из чего сделаны ИИ-дата-центры. Идея Дженсена Хуанга – «все есть Generative AI»

GENERATIVE AI
The iPhone Moment of AI



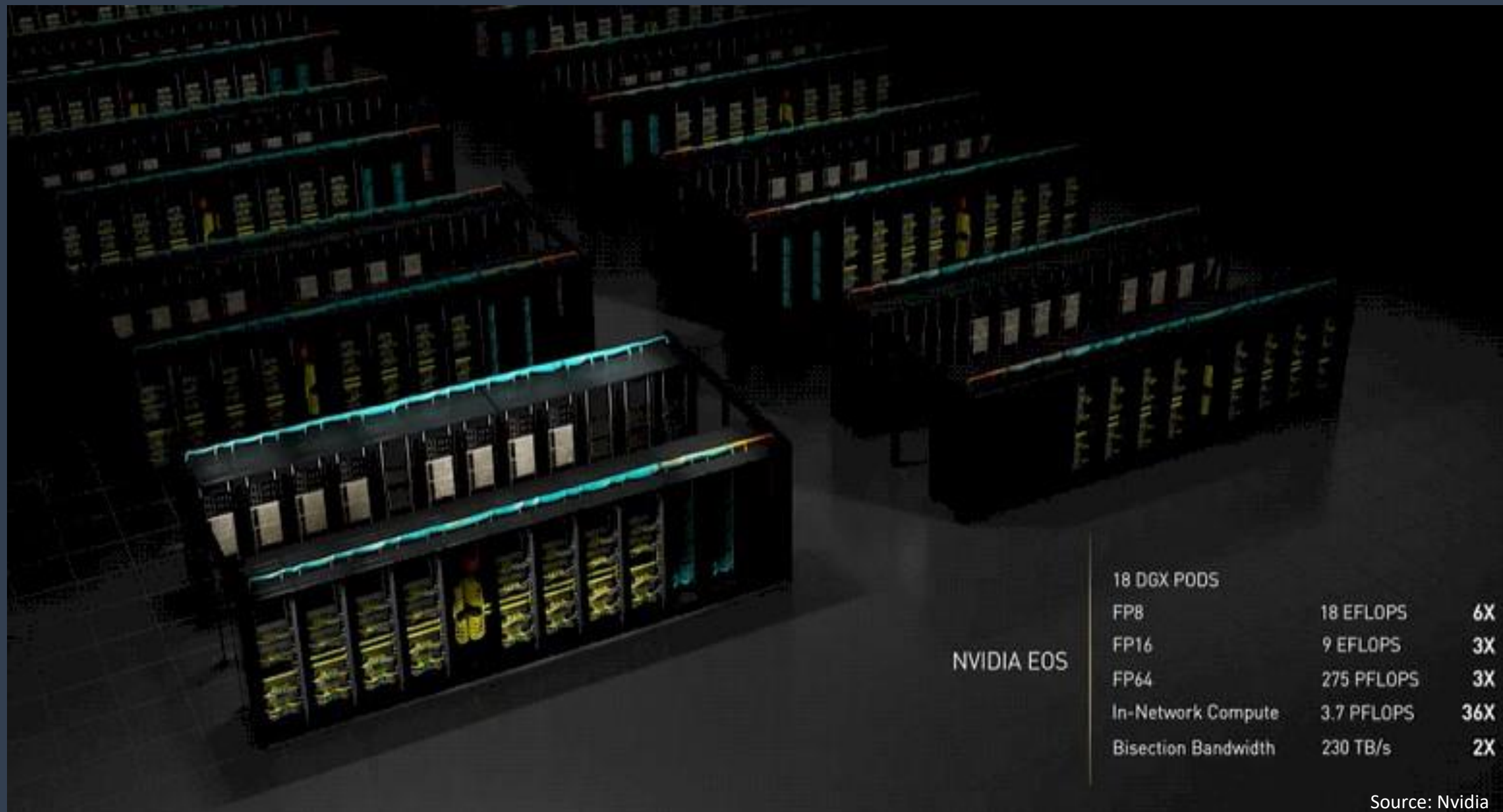
Source: Nvidia

2.2. Из чего сделаны ИИ-дата-центры. «Дата-центры – это фабрики по производству токенов»





2.3. Из чего сделаны ИИ-дата-центры. Масштабируем GPU-системы, чтобы фабрики токенов были эффективными



2.4. Из чего сделаны ИИ-дата-центры. Оценка стоимости обучения Chat GPT 4.0 1.8T

GPT-4 1.8T MoE	2016	2018	2020	2022	2024
10 Day Training	"Pascal"	"Volta"	"Ampere"	"Hopper"	"Blackwell"
GPU	P100	V100	A100	H100	B100
Peak Teraflops	19	130	620	4,000	20,000
<u>Training And Inference Precision</u>	<u>FP16</u>	<u>FP16</u>	<u>FP16</u>	<u>FP8</u>	<u>FP4</u>
Inference Joules/Token	17,000	1,200	150	10	0.4
GPUs to Train In 10 Days	42,105,300	6,153,800	1,290,300	100,000	10,000
GPU Unit Price	\$5,000	\$10,000	\$15,000	\$27,500	\$37,500
<u>GPU Cost For Training In 10 Days</u>	<u>\$210.5 B</u>	<u>\$61.5 B</u>	<u>\$19.4 B</u>	<u>\$2.8 B</u>	<u>\$0.4 B</u>
Power To Train, Gw-hour	1,000	140	40	13	3
Electricity Cost Per Kw-hour	\$0.14	\$0.14	\$0.14	\$0.16	\$0.18
<u>Electricity Cost For Training Run</u>	<u>\$140.0 M</u>	<u>\$19.6 M</u>	<u>\$5.6 M</u>	<u>\$2.1 M</u>	<u>\$0.54 M</u>
<u>Three Year Electricity Cost</u>	<u>\$15.3 B</u>	<u>\$2.1 B</u>	<u>\$613.6 M</u>	<u>\$227.9 M</u>	<u>\$59.2 M</u>

2.5. Из чего сделаны ИИ-дата-центры.

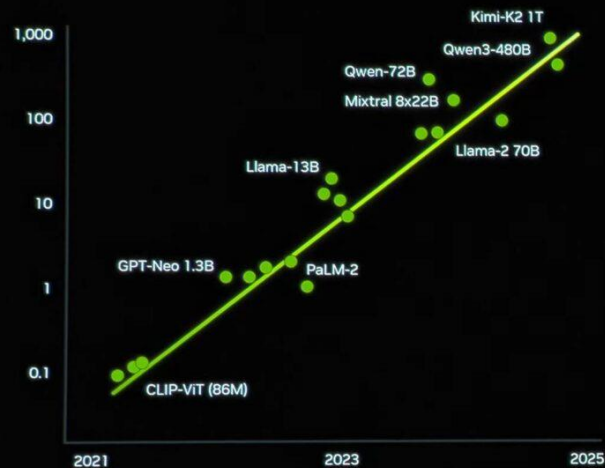
Потребности в вычислениях для ИИ растут экспоненциально



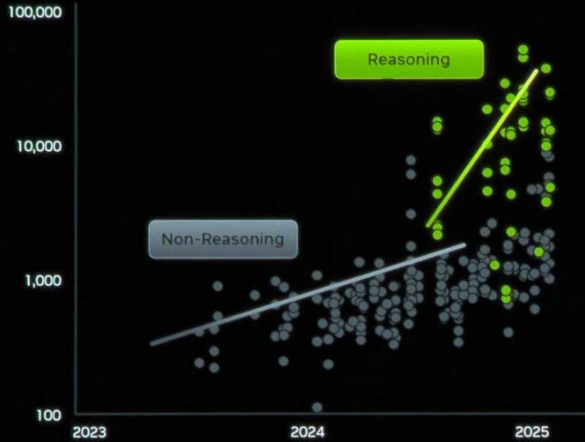
Cognitar

Insane Demand for AI Computing

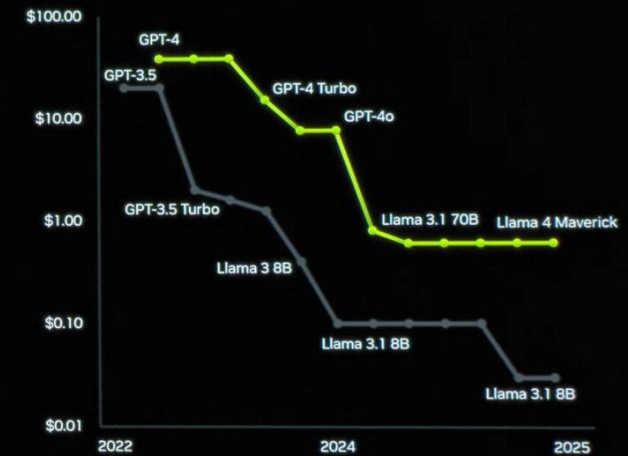
Model Size Growing
10X Parameters Per Year



Test-Time Scaling "Thinking"
5X Tokens Per Year



Token Cost
10X Cheaper Per Year

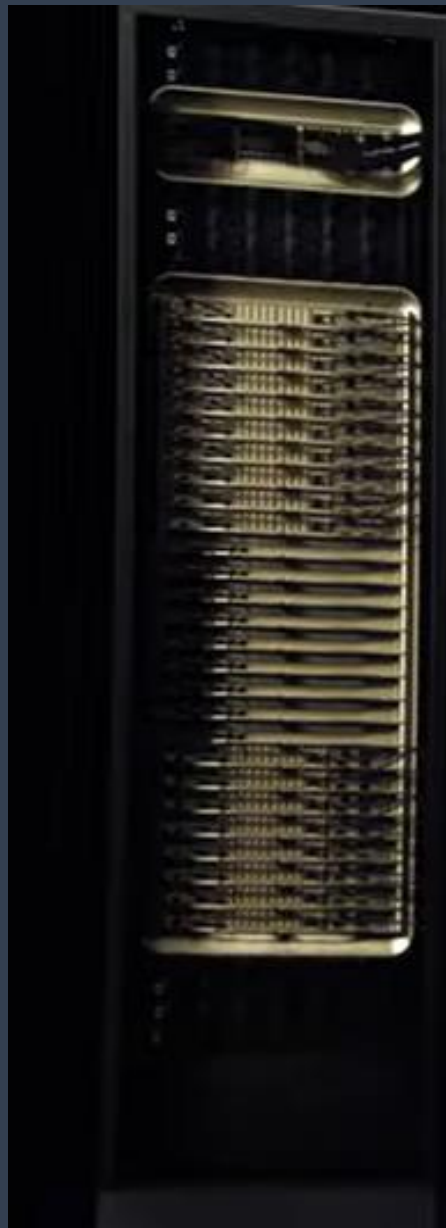


Source : EPOCH AI

Source : Artificial Analysis

Source: Nvidia

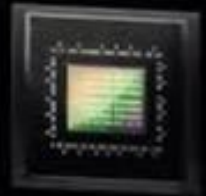
2.6. Из чего сделаны ИИ-дата-центры. Blackwell Ultra NVL72 – единая система из 72 GPU Blackwell Ultra и 36 CPU Grace



Blackwell Ultra NVL72
Second Half 2025

- 1.1 EF Dense FP4 Inference
0.36 EF FP8 Training
1.5X GB200 NVL72
- New Attention Instructions
2X
- 20 TB HBM | 40 TB Fast Memory
1.5X
- 14.4 TB/s CX8
2X

Grace



Blackwell Ultra



2 Reticle-Sized GPUs
Per Dense FP4 | 288GB HBM3e

Oberon Rack
Liquid Cooled

Source: Nvidia

2.7. Из чего сделаны ИИ-дата-центры. Планы 2025 года: Vera Rubin NVL144 – единая система из 144 чипов Rubin и CPU на 36 чипах Vera



Vera Rubin NVL144

Second Half 2026

- 3.6 EF FP4 Inference
- 1.2 EF FP8 Training
- 3.3X GB300 NVL72
- 13 TB/s HBM4
- 75 TB Fast Memory
- 1.6X
- 260 TB/s NVLink6
- 2X
- 28.8 TB/s CX9
- 2X

Vera



88 Custom Arm Cores
176 Threads
1.8 TB/s NVLink-C2C

Rubin



Die-Sized GPUs
FP4 | 288GB HBM4

Oberon Rack
Liquid Cooled

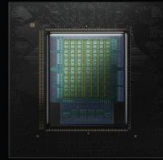
Source: Nvidia



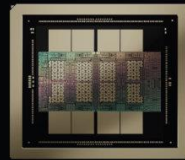
2.8. Из чего сделаны ИИ-дата-центры. Факты 2026 года: Анонсирована Vera Platform – полный набор чипов нового поколения

Announcing the NVIDIA Rubin Platform

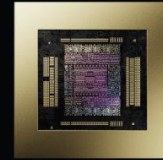
Six new chips. One AI supercomputer.



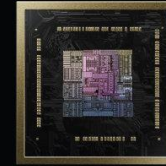
Vera



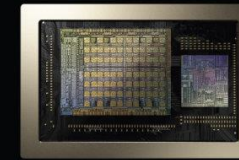
Rubin



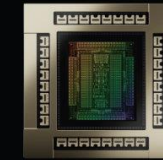
NVLink 6 Switch



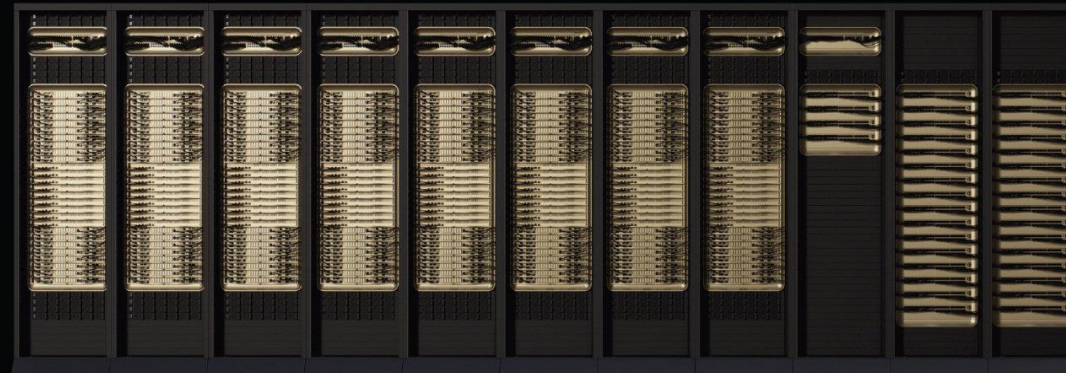
CX9



BF4



Spectrum-X
102.4T CPO



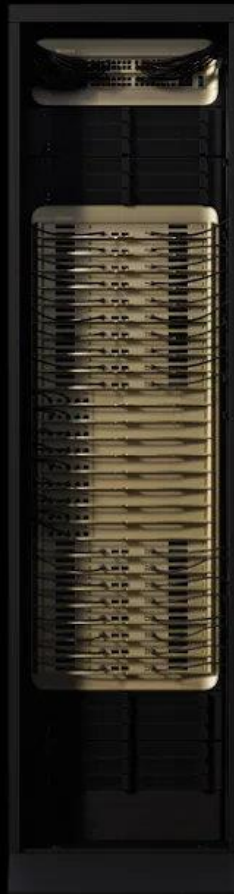
Vera Rubin POD

2.9. Из чего сделаны ИИ-дата-центры.

Факты 2026 года: Vera Rubin NVL72 – единая система
из 72 двойных чипов Rubin и CPU на 36 чипах Vera

NVIDIA Vera Rubin NVL72

AI supercomputer for the next generation of AI



NVFP4 Inference	3.6 EFLOPS	5x Blackwell
NVFP4 Training	2.5 EFLOPS	3.5x
LPDDR5X Capacity	54 TB	2.5x
HBM4 Capacity	20.7 TB	1.5x
HBM4 Bandwidth	1.6 PB/s	2.8x
Scale-Up Bandwidth	260 TB/s	2x

2.10. Из чего сделаны ИИ-дата-центры.

Факты 2026 года: Vera Rubin NVL72 – единая система
из 72 двойных чипов Rubin и CPU на 36 чипах Vera



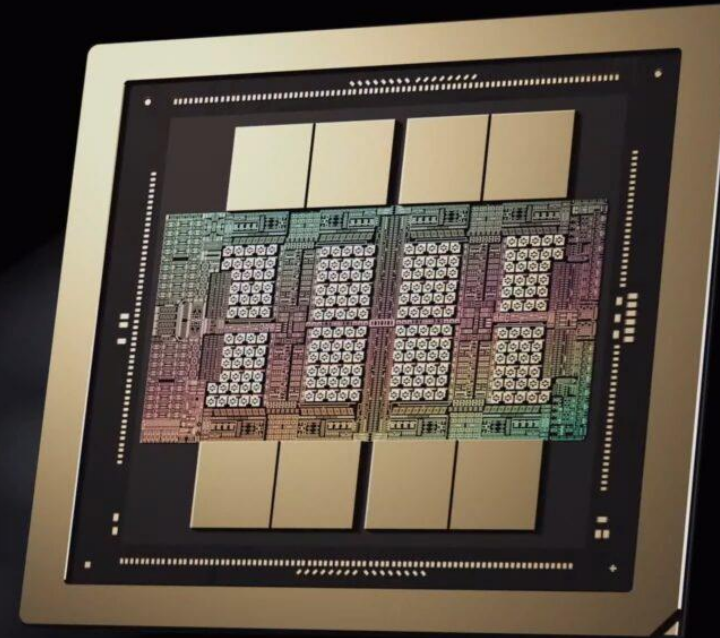
Source: Nvidia

2.11. Из чего сделаны ИИ-дата-центры.

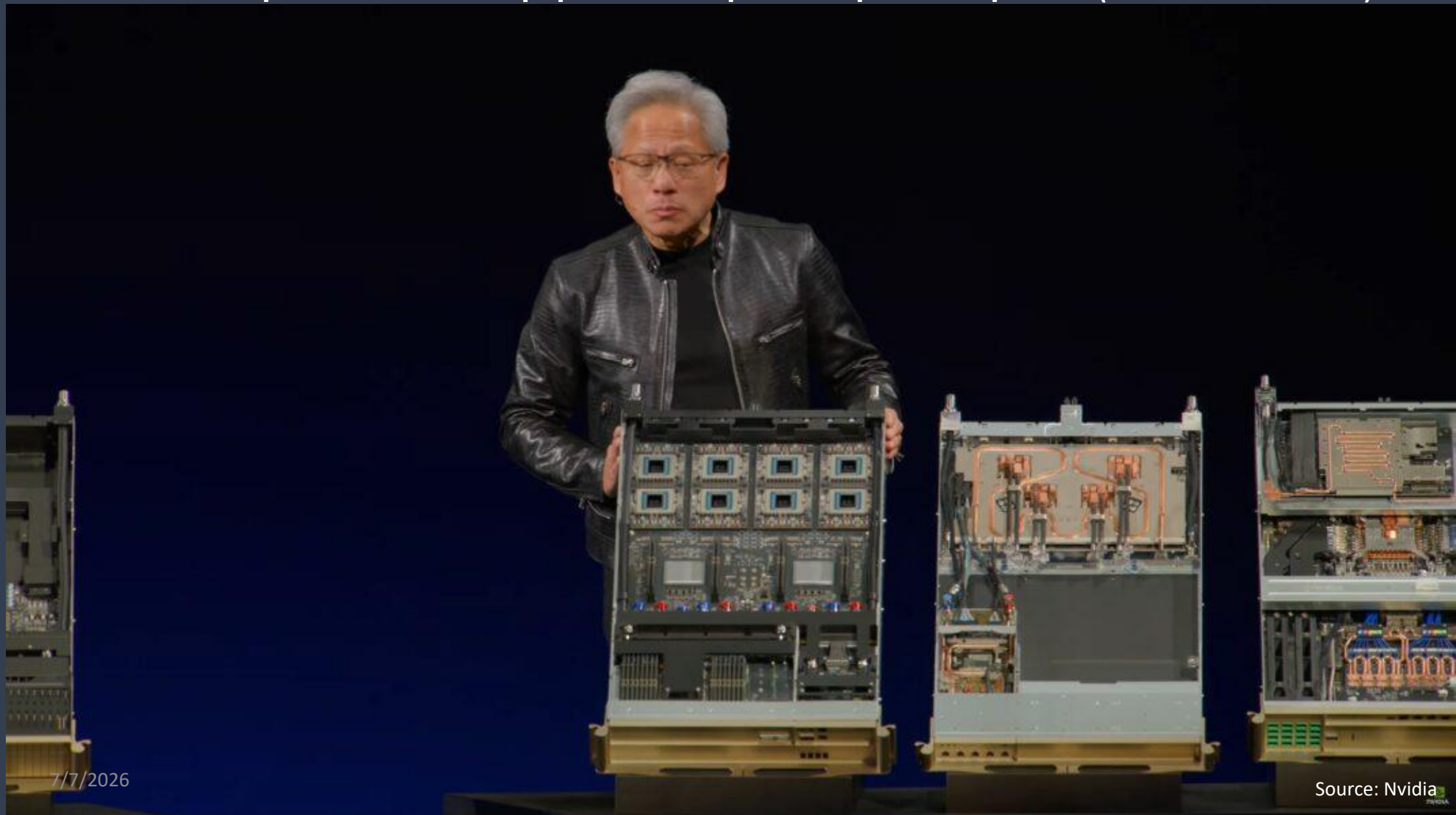
Факты 2026 года: основной чип платформы VR NVL72 – ускоритель GPU Rubin

NVIDIA Rubin GPU

NVFP4 Inference	50 PFLOPS	5X Blackwell
NVFP4 Training	35 PFLOPS	3.5X
HBM4 Bandwidth	22 TB/s	2.8X
NVLink Bandwidth per GPU	3.6 TB/s	2X
Transistors	336 Billion	1.6X



2.11a. Из чего сделаны ИИ-дата-центры. Новый трей LPU: для инференса – с процессором Croq3 LPU (COMPUTEX-2026)



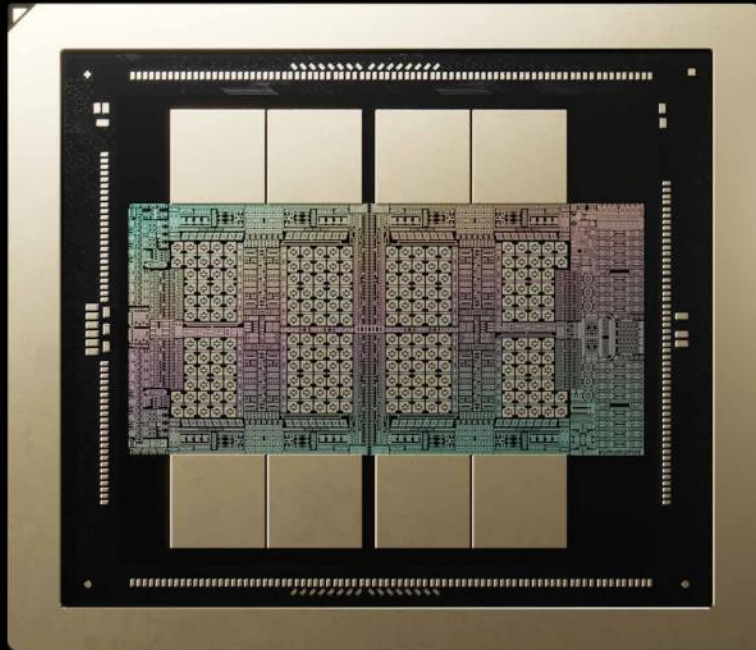
7/7/2026

Source: Nvidia

2.11b. Из чего сделаны ИИ-дата-центры.

Процессор для инференса – Croq3 LPU

Rubin GPU



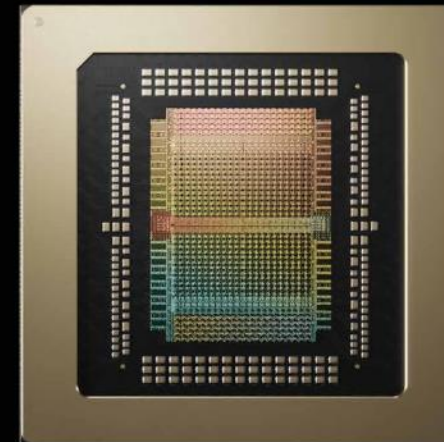
288 GB HBM4

22 TB/s

50 PFLOPs (NVFP4)

336B Transistors
+ 2.5T (HBM4)

Groq 3 LPU



500 MB SRAM

150 TB/s SRAM Bandwidth

1.2 PFLOPs (FP8)

98B Transistors

Uniting Processors of Extreme Performances

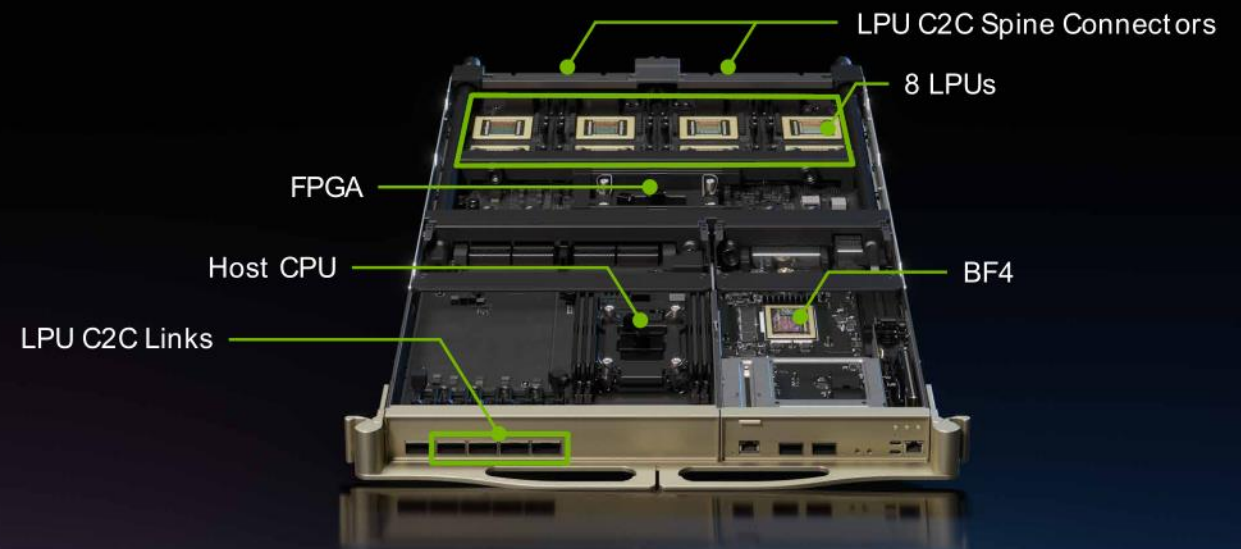
2.11с. Из чего сделаны ИИ-дата-центры. Стойка LPX для инференса с 256 процессорами Croq3 LPU



Available 2H26

NVIDIA Groq 3 LPX

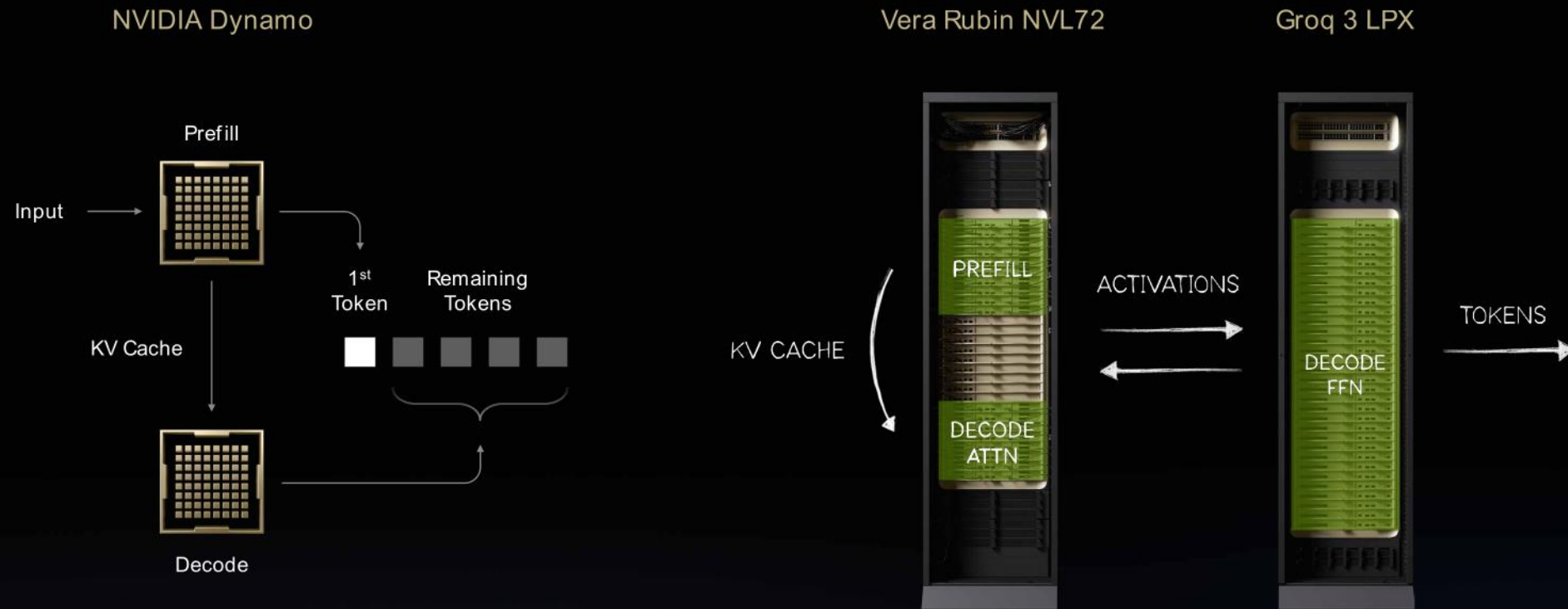
AI Inference Compute	315 PFLOPS
SRAM Capacity	128 GB
Memory Bandwidth	40 PB/s
Scale-Up Density	256 Chips
Scale-Up Bandwidth	640 TB/s



Source: Nvidia

2.11d. Из чего сделаны ИИ-дата-центры.

Как работает эффективный инференс на тандеме стоек: Vera Rubin NVL72 + Croq3 LPX

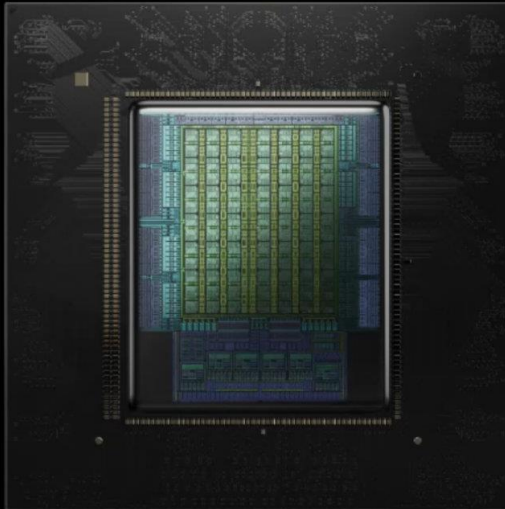


Uniting Processors of Extreme FLOPS and Bandwidth

2.12. Из чего сделаны ИИ-дата-центры. Второй главный чип платформы VR NVL72 – процессор CPU Vera

NVIDIA Vera CPU

Built for data movement and agentic processing



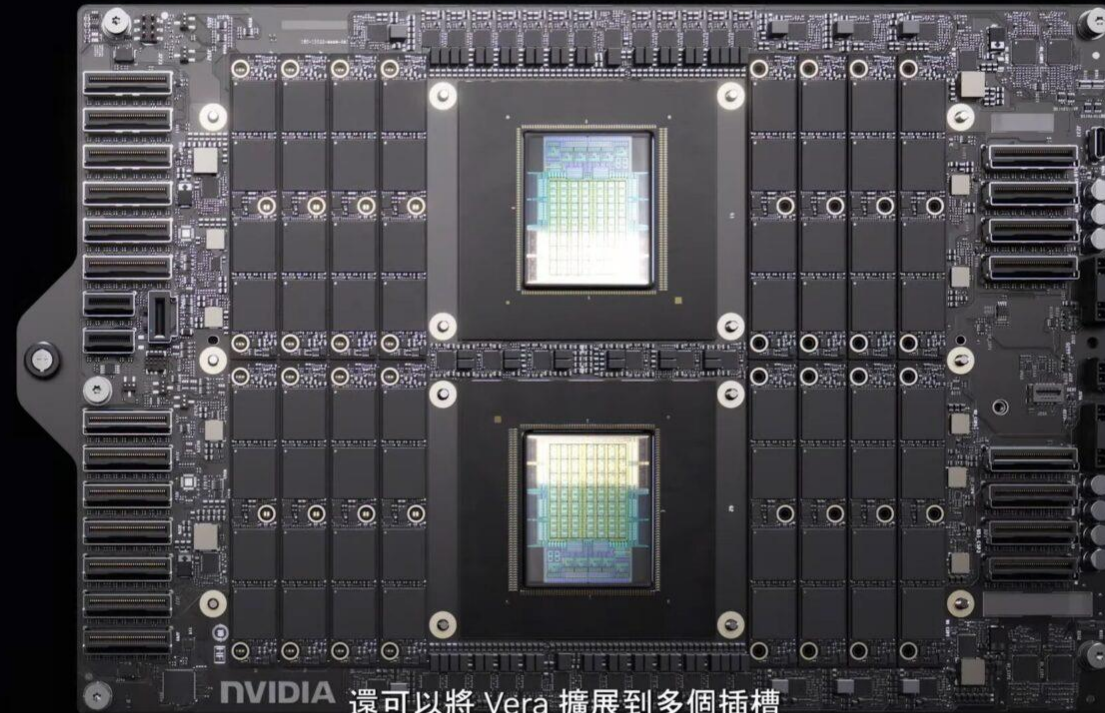
- 88 NVIDIA custom Olympus cores
- 176 threads with NVIDIA Spatial Multi-threading
- 1.8 TB/s NVLink-C2C enabling CPU:GPU coherent memory
- 1.5 TB system memory (3X Grace)
- 1.2 TB/s memory bandwidth with SOCAMM LPDDR5X
- Rack-scale confidential compute

2X Data Processing

2X Compression

2X CI/CD

2.12a. Из чего сделаны ИИ-дата-центры. Узел только на CPU-процессорах Vera (COMPUTEX-2026)

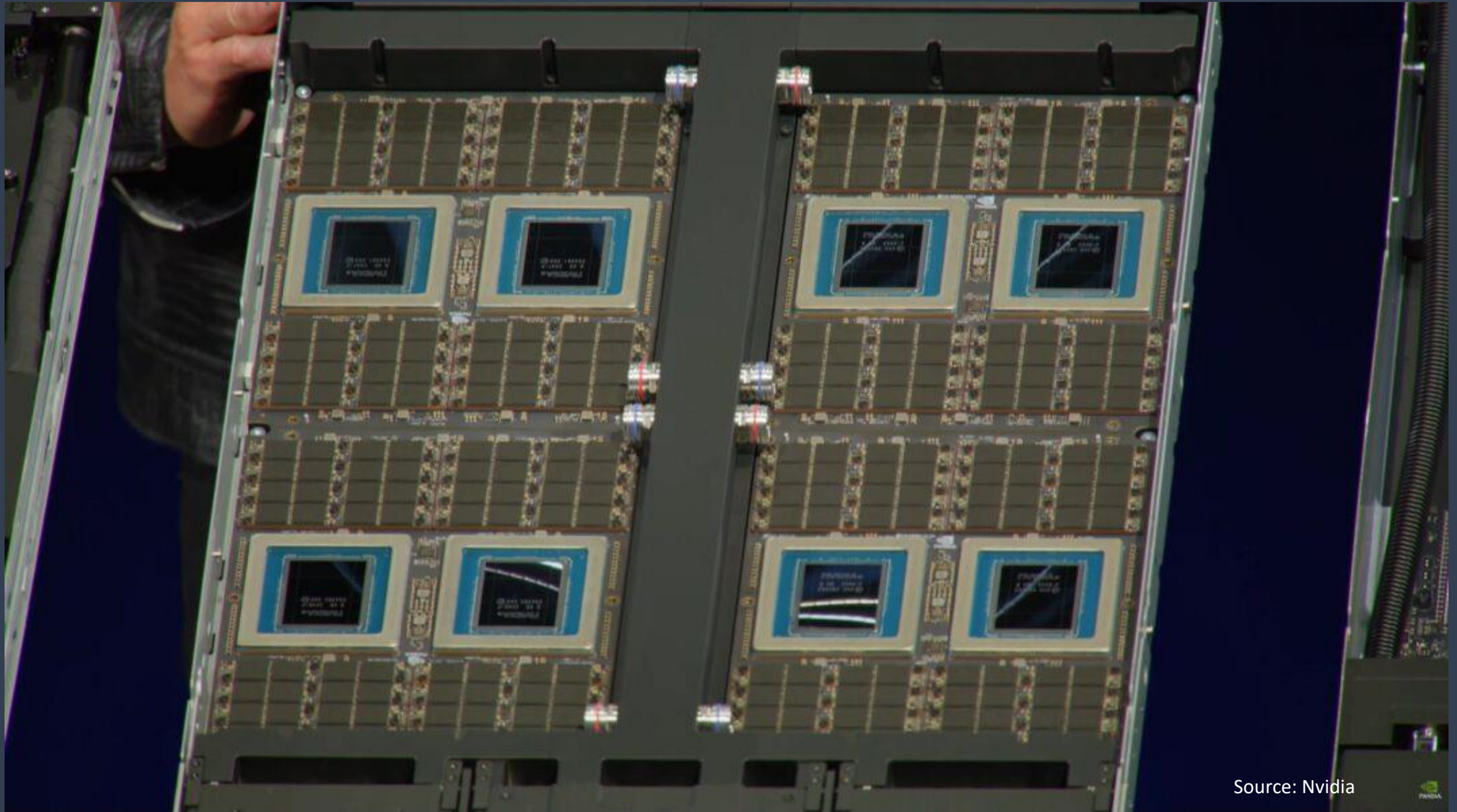


還可以將 Vera 擴展到多個插槽

2.12b. Из чего сделаны ИИ-дата-центры. Трей с узлами только на CPU-процессорах Vera (COMPUTEX-2026)



Cognitar

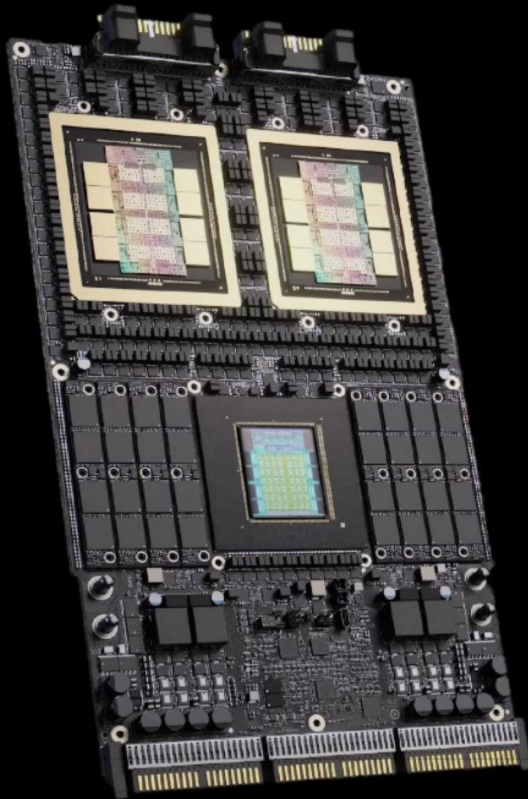


Source: Nvidia

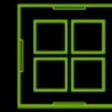


2.13. Из чего сделаны ИИ-дата-центры. Вычислительный модуль платформы VR NVL72

NVIDIA Rubin Technology Breakthroughs



6th Gen NVLink
3.6 TB/s Scale-Up



Vera CPU
Custom Olympus Core



Rubin GPU
50 PF NVFP4 Transformer Engine



3rd Gen Confidential Computing
First Rack-Scale TEE



2nd Gen RAS Engine
Zero Downtime Health Checks

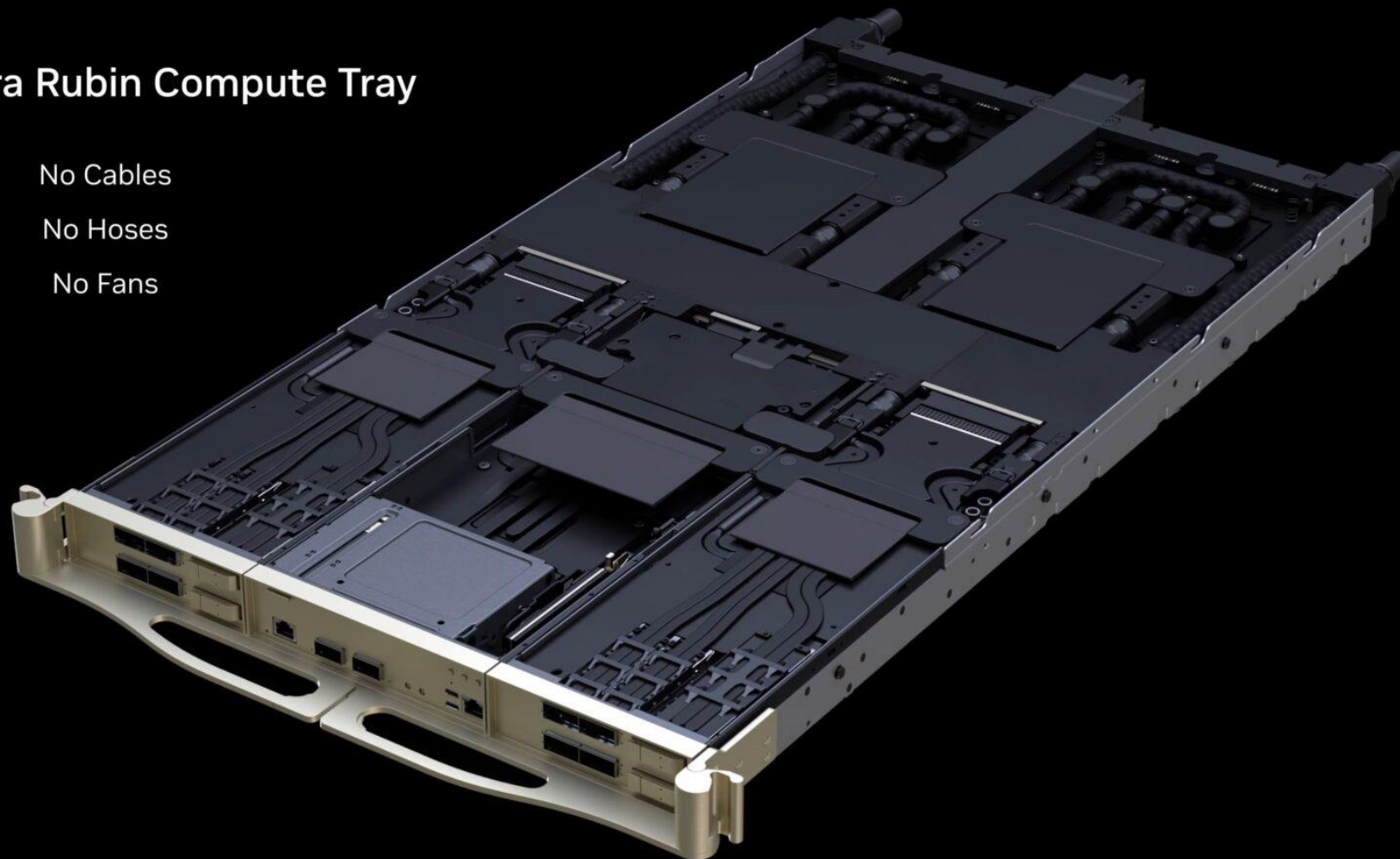
2.14. Из чего сделаны ИИ-дата-центры. Вычислительный модуль платформы VR NVL72

NVIDIA Vera Rubin Compute Tray

No Cables

No Hoses

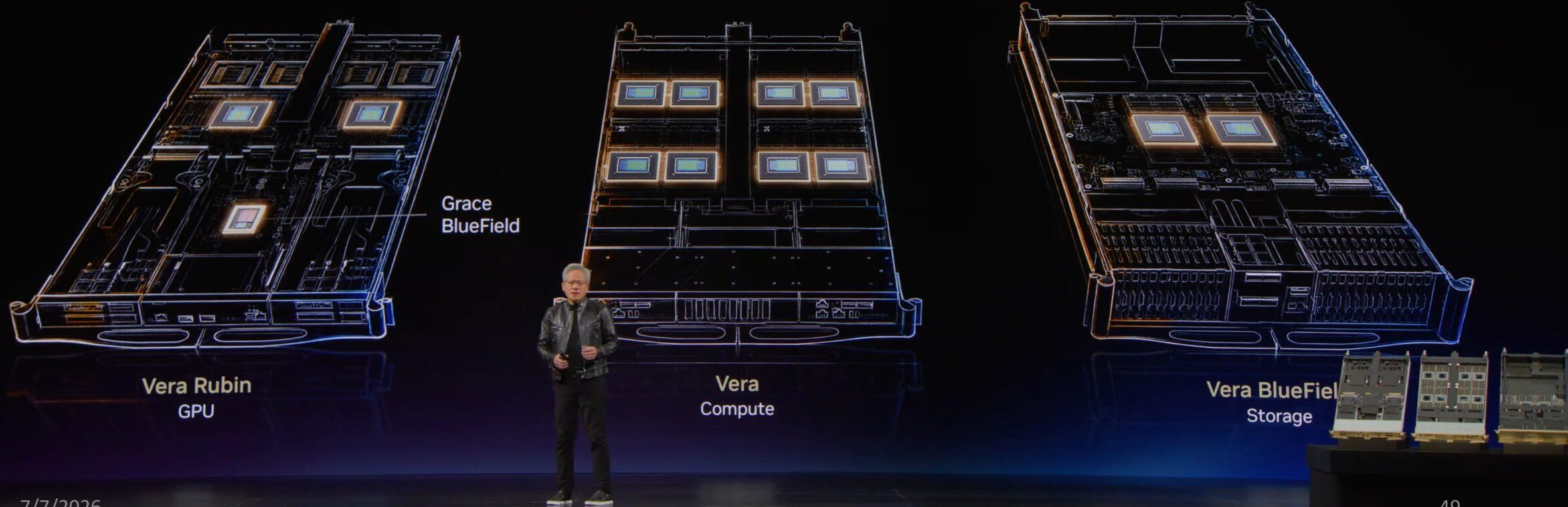
No Fans



Source: Nvidia

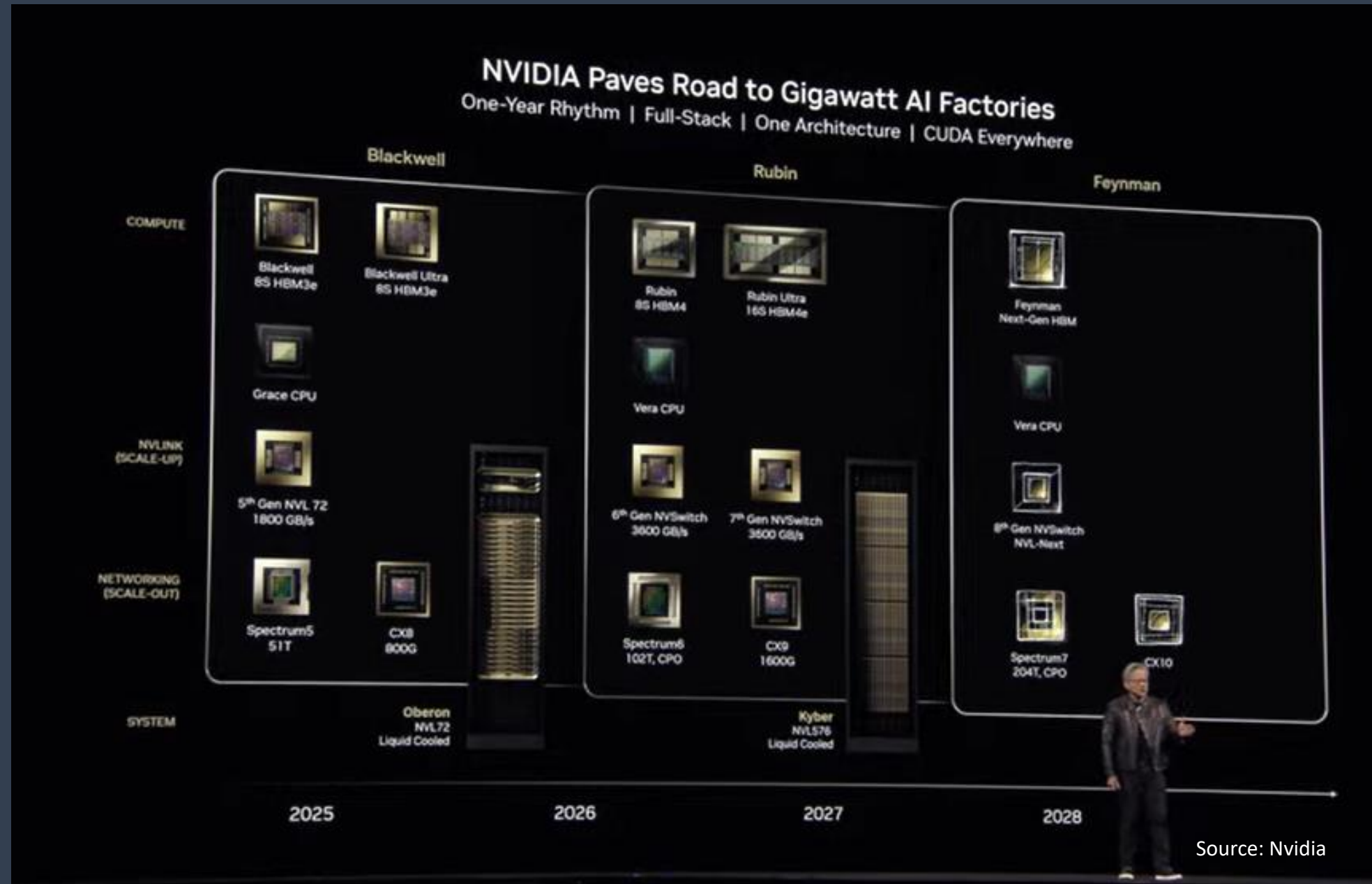
2.14a. Из чего сделаны ИИ-дата-центры. Вся базовая линейка треев Vera Rubin (COMPUTEX-2026)

NVIDIA Vera – CPU for the Age of AI



2.15. Из чего сделаны ИИ-дата-центры.

Рoadmap GPU nVidia – появился GPU чип Feynman (будет в 2028)

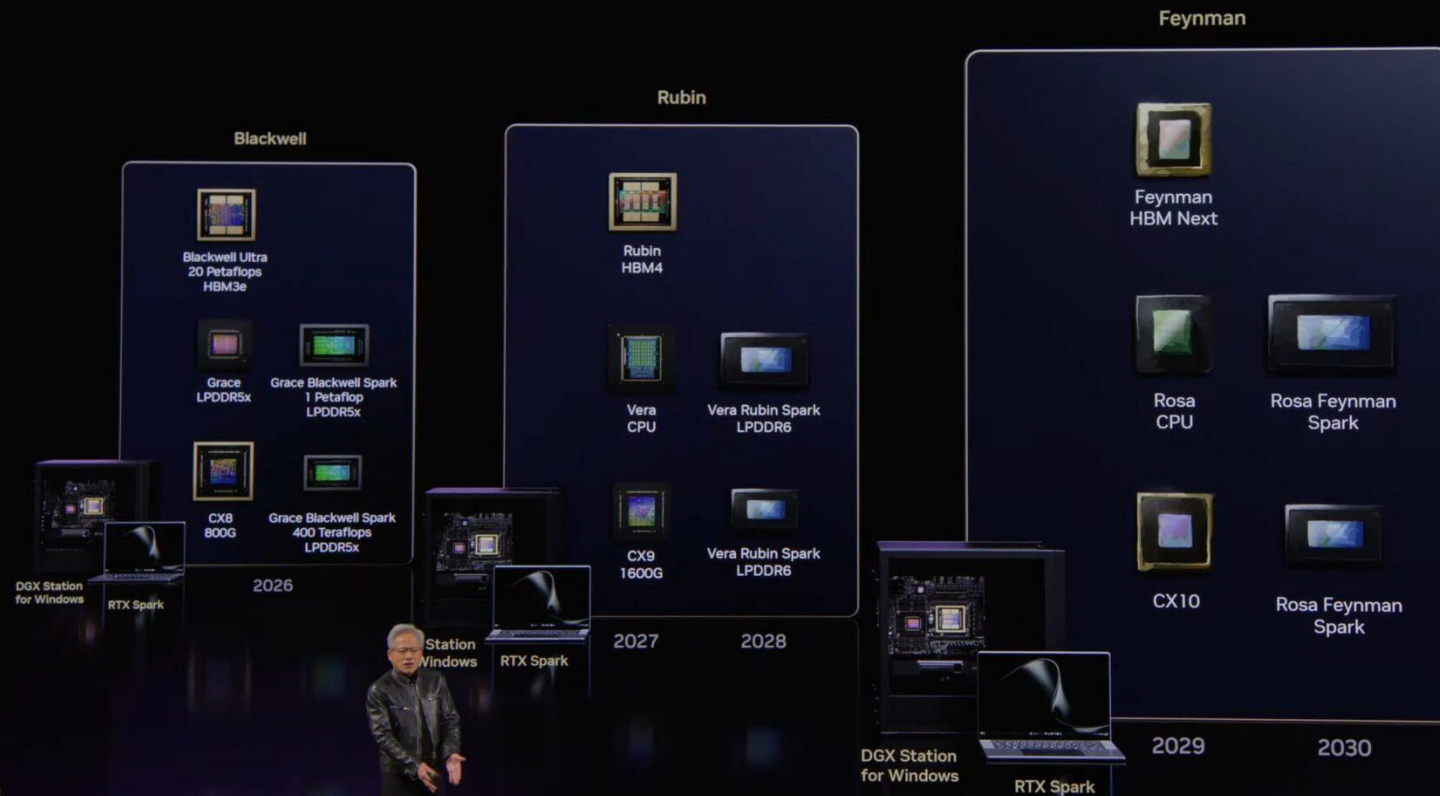


2.15a. Из чего сделаны ИИ-дата-центры. Рoadmap GPU nVidia – обновленный на COMPUTEX-2026, появился чип RTX Sparc



Cognitar

A New Line. A New Beginning.

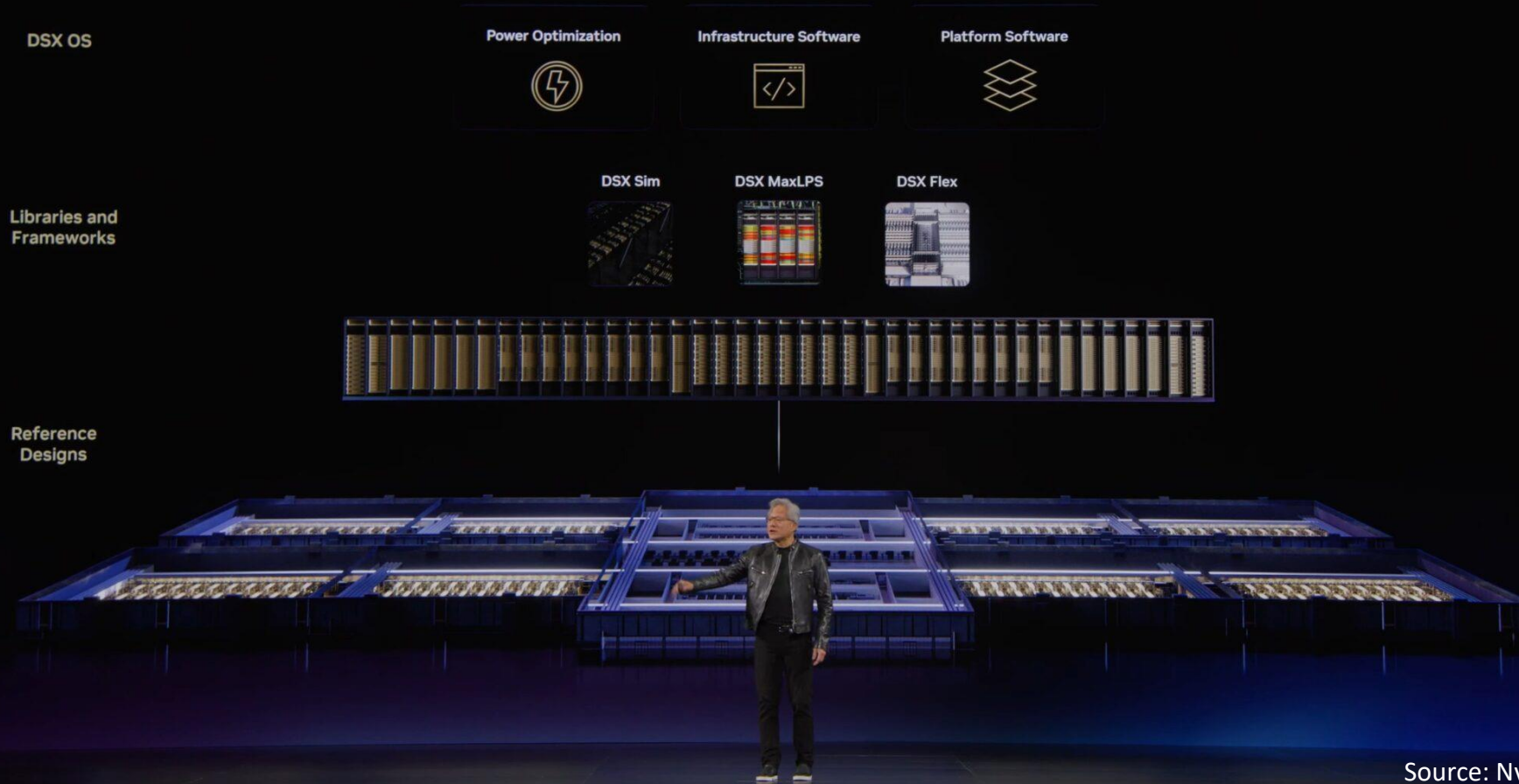


Source: Nvidia



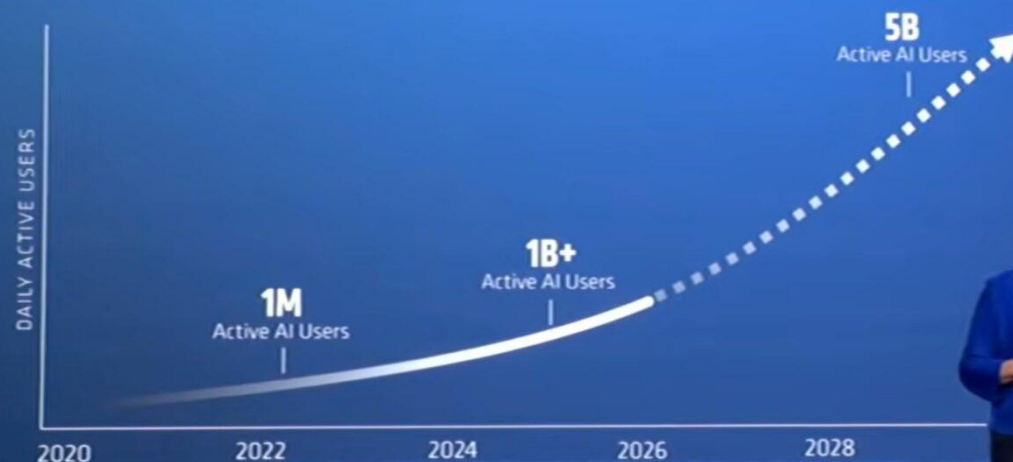
2.15b. Из чего сделаны ИИ-дата-центры. Теперь проектируем целые ИИ-фабрики (COMPUTEX-2026)

NVIDIA DSX AI Factory Platform



Source: Nvidia

2.16. Из чего сделаны ИИ-дата-центры. Взгляд на развитие ИИ и потребности Лизы Су – CEO AMD



Within 5 Years

**5 Billion
People Using AI Daily**

2.17. Из чего сделаны ИИ-дата-центры. АМД. Как закрыть потребности в вычислениях ИИ-фабрик

[AMD Confidential - Distribution Under NDA]

AMD | Advancing AI Infrastructure on an Annual Cadence

2025

AMD EPYC™
"TURIN"

AMD INSTINCT™
MI350 SERIES

AMD PENSANDO™
"POLLARA"



2026

AMD EPYC™
"VENICE"

AMD INSTINCT™
MI400 SERIES

AMD PENSANDO™
"VULCANO"



"Helios"

2027

AMD EPYC™
"VERANO"

AMD INSTINCT™
MI500 SERIES

AMD PENSANDO™
"VULCANO"



Next-Gen AI Rack

2.18. Из чего сделаны ИИ-дата-центры. АМД. Стойка Helios с 72 GPU-ускорителями Instinct Mi-455



2.19. Из чего сделаны ИИ-дата-центры. АМД. Стойка Helios с 72 GPU-ускорителями Instinct Mi-455

Preview

AMD “Helios” Optimized AI Rack Solution

AMD
EPYC

AMD
INSTINCT

AMD
PENSANDO

AMD
ROCm

Available in 2026



2.20. Из чего сделаны ИИ-дата-центры.

АМД. Стойка Helios с 72 GPU-ускорителями Instinct Mi-455

[AMD Confidential - Distribution Under NDA]

AMD “Helios” AI Rack

Rack Scale AI Performance Leadership

	Instinct™ MI400 Series AMD “Helios”	vs. Vera Rubin Oberon
GPU DOMAIN	72	1x
SCALE UP BANDWIDTH	260 TB/s	1x
FP4 • FP8 FLOPS	2.9 EF • 1.4 EF	~1x
HBM4 MEMORY CAPACITY	31 TB	1.5x
MEMORY BANDWIDTH	1.4 PB/s	1.5x
SCALE OUT BANDWIDTH	43 TB/s	1.5x

2.21. Из чего сделаны ИИ-дата-центры. АМД. Стойка Helios – компьютерный трей, в стойке их 18 штук.

AMD Pensando™
"Salina" 400 DPU

AMD Instinct™ MI455X GPUs

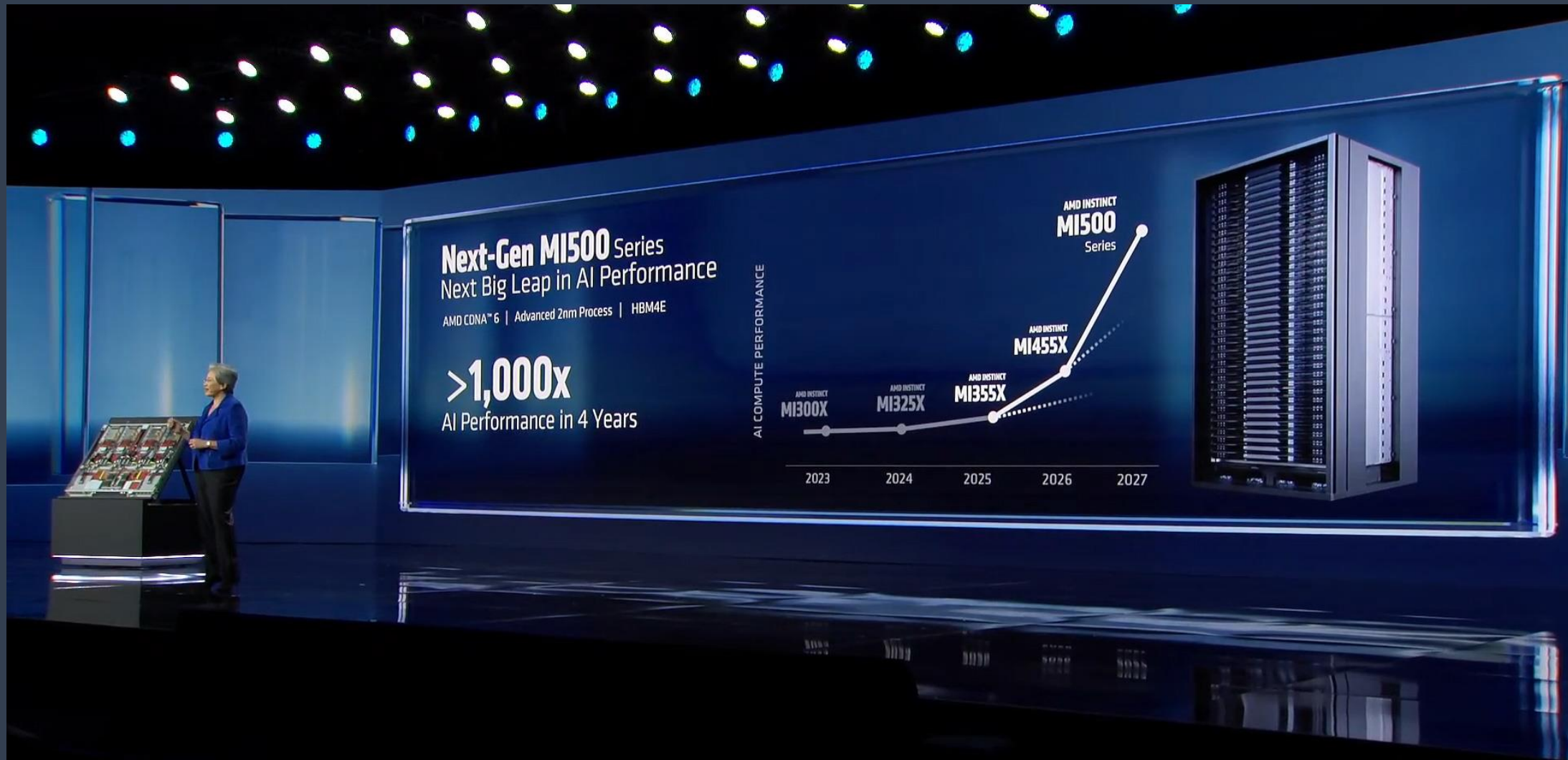
AMD EPYC™ "Venice" CPU

AMD Pensando™
"Vulcano" 800 AI NIC

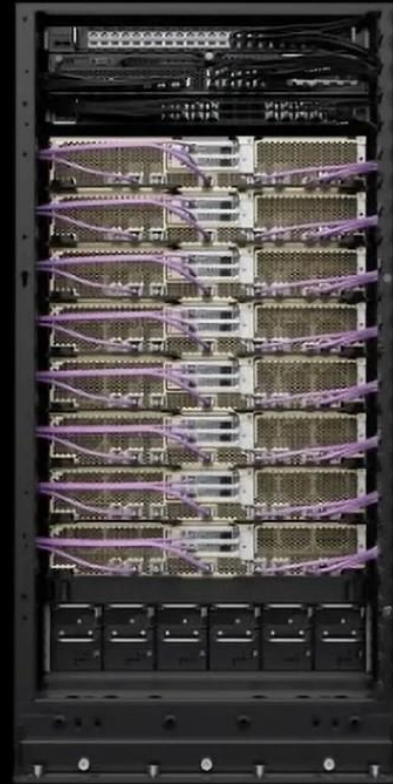
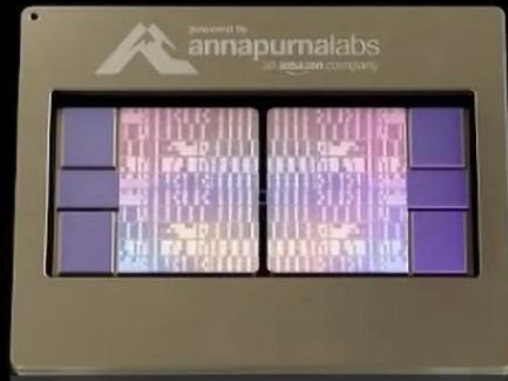
Source: AMD



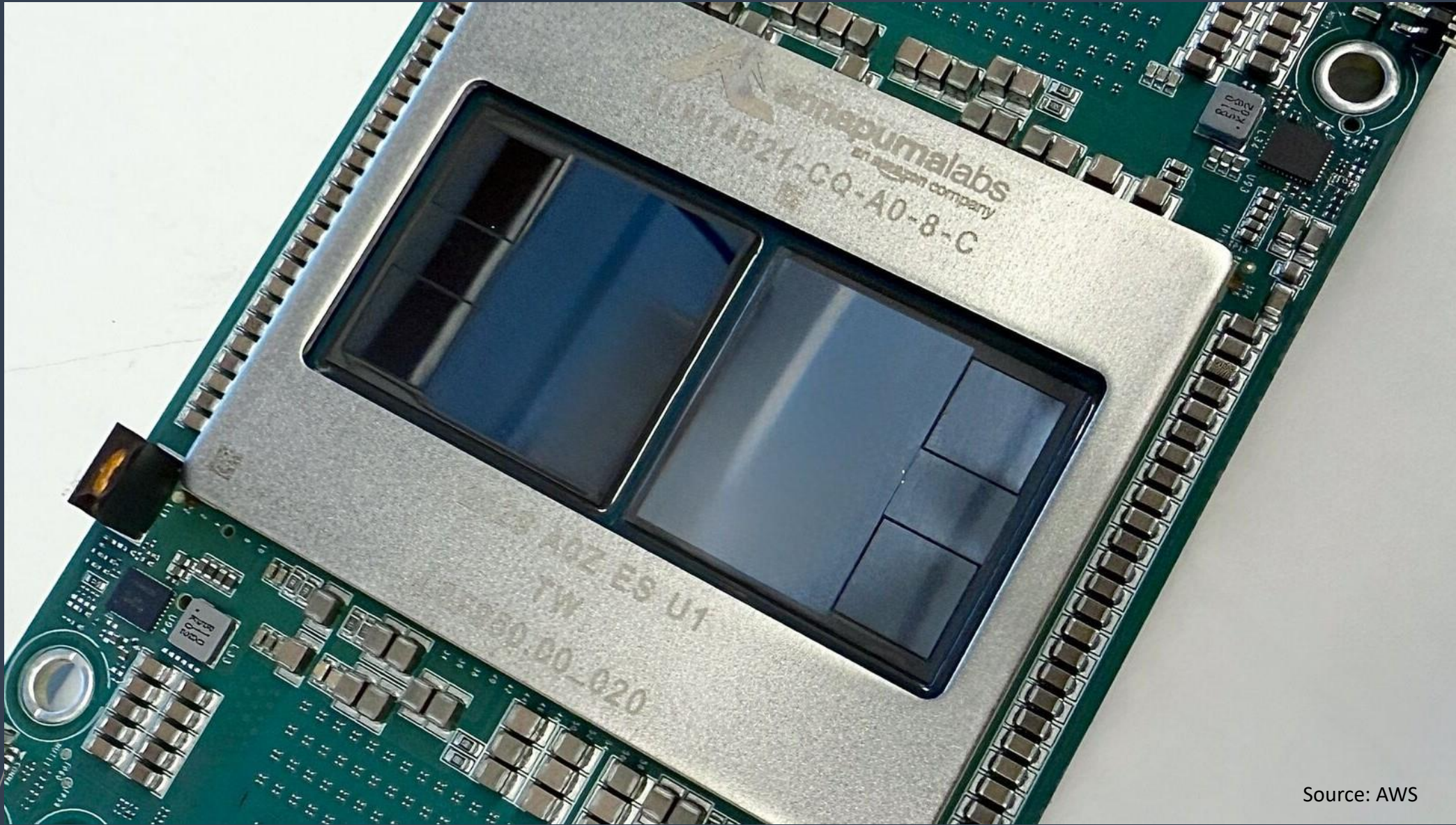
2.22. Из чего сделаны ИИ-дата-центры. АМД. Стойка Helios – дальше последуют ускорители серии MI500



2.23. Из чего сделаны ИИ-дата-центры.
Amazon Web Services (AWS).
Стойка Trainium Ultra с ускорителями Trainium

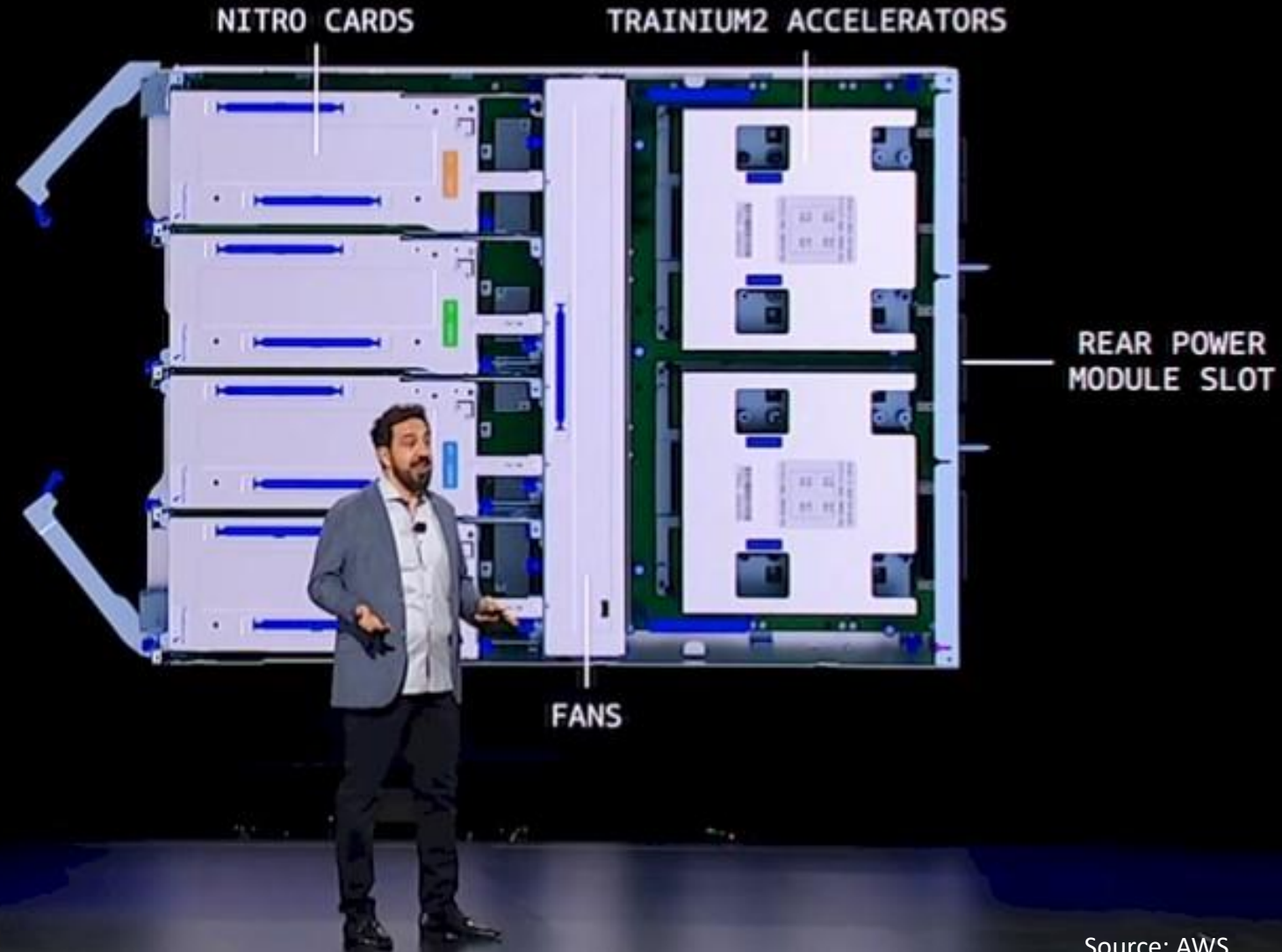


2.24. Из чего сделаны ИИ-дата-центры.
Amazon Web Services (AWS).
Ускоритель Trainium3 (AWS Anapurna labs)



2.25. Из чего сделаны ИИ-дата-центры. AWS. Стойка Trainium Ultra - Компьютерный трей

Simplified, cable-less
chassis design



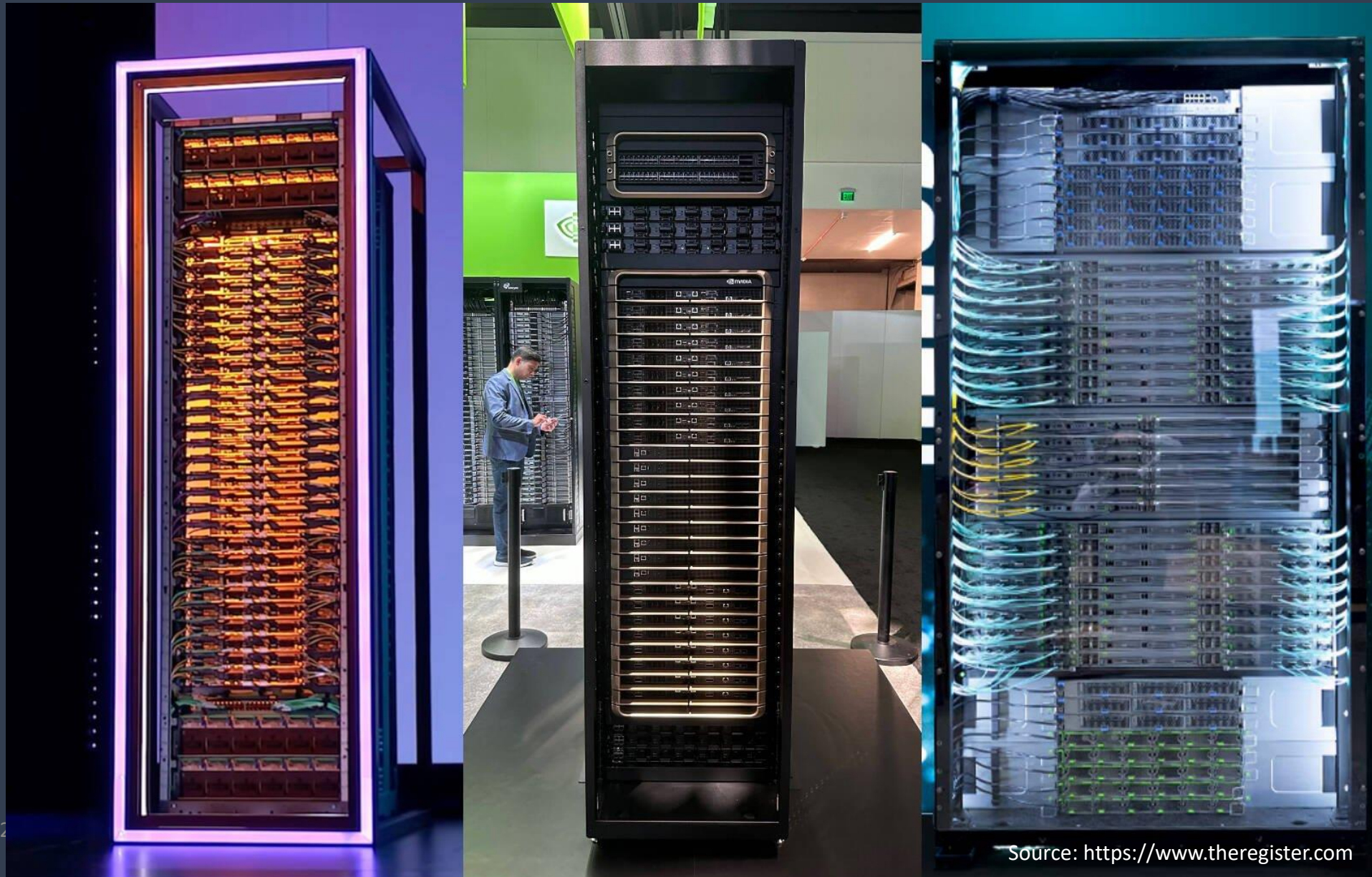
2.26. Из чего сделаны ИИ-дата-центры. AWS. Стойка Trainium Ultra - Компьютерный трей

Trainium2 Server

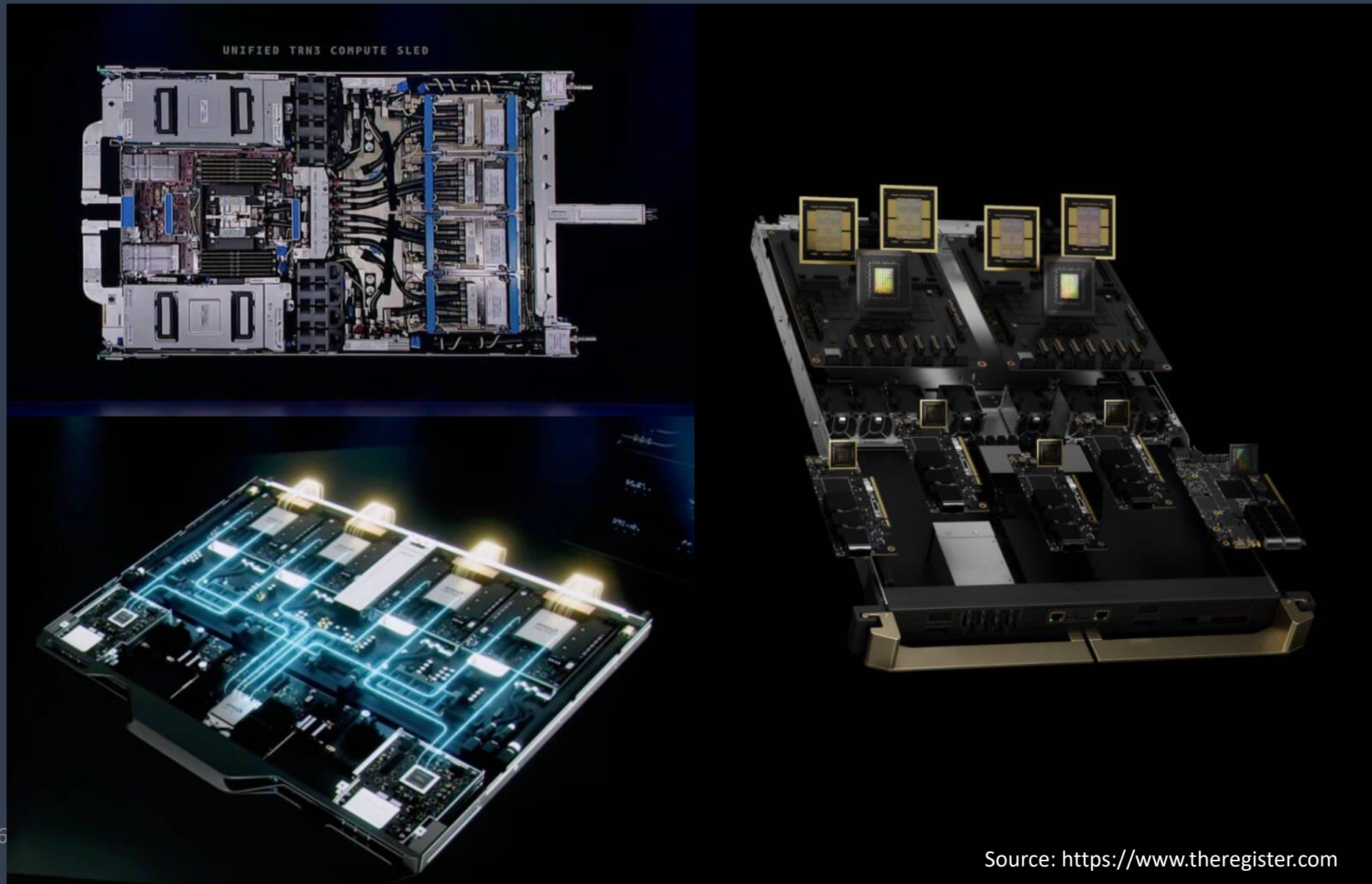
HEAD NODE



2.27. Из чего сделаны ИИ-дата-центры.
И наконец все три стойки для ИИ-фабрик:
AMD Helios, nVidia VR NVL72, AWS Trainium Ultra

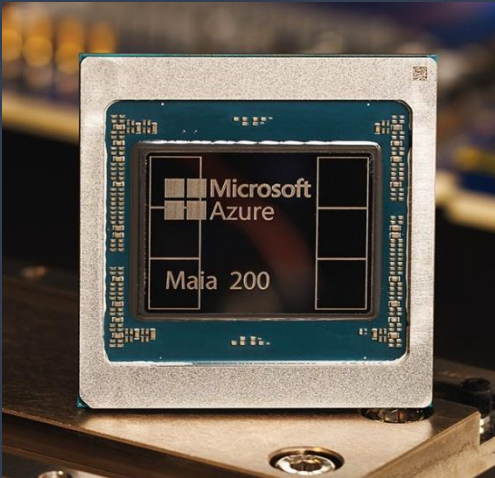


2.28. Из чего сделаны ИИ-дата-центры. Три компьютерных трея трех стоек для ИИ-фабрик: AMD Helios, nVidia VR NVL72, AWS Trainium Ultra



2.29. Из чего сделаны ИИ-дата-центры.

ИИ-ускорители Microsoft Maia 200, AWS Trainium3 , Google TPU v7 Ironwood



Source: Microsoft



Source: AWS



Source: ServeTheHome

Peak specifications	Azure Maia 200	AWS Trainium3	Google TPU v7
Process Node	3nm	3nm	3nm
FP4 TFLOPS	10,145	2,517	n/a
FP8 TFLOPS	5,072	2,517	4,614
BF16 TFLOPS	1,268	671	2,307
HBM technology	HBM3E	HBM3E	HBM3E
HBM BW (TB/s)	7	4.9	7.4
HBM capacity (GB)	216	144	192
Scale-up BW: bidirectional (TB/s)	2.8 TB/s	2.2-2.56 TB/s	1.2 TB/s

Source: Microsoft

2.30. Из чего сделаны ИИ-дата-центры. Литература:

1. GTC March 2025 Keynote with NVIDIA CEO Jensen Huang:

https://www.youtube.com/live/_waPvOwL9Z8

2. CES-2026: NVIDIA Live with CEO Jensen Huang:

<https://www.youtube.com/watch?v=0NBILspM4c4>

3. AMD at CES-2026 (with AMD CEO Lisa SU):

<https://www.amd.com/en/corporate/events/ces.html>

4. AMD's EPYC Venice, Instinct MI455X, & Helios Hardware On Display for First Time at CES 2026:

<https://www.servethehome.com/amds-epyc-venice-instinct-mi455x-helios-hardware-on-display-for-first-time-at-ces-2026/>

5. AWS re:Invent 2025 - Keynote with CEO Matt Garman:

<https://www.youtube.com/watch?v=q3Sb9PemsSo>

6. Amazon launches Trainium3 AI accelerator, competing directly against Blackwell Ultra in FP8 performance...:

<https://www.tomshardware.com/tech-industry/artificial-intelligence/amazon-launches-trainium3-ai-accelerator-competing-directly-against-blackwell-ultra-in-fp8-performance-new-trn3-gen2-ultraserver-takes-vertical-scaling-notes-from-nvidias-playbook>

7. Amazon's Trainium3 is the latest to conform to Nvidia's mold:

https://www.theregister.com/2025/12/07/trainium3_all_nvidia_nvl72_mold/

8. Maia 200: The AI accelerator built for inference:

<https://blogs.microsoft.com/blog/2026/01/26/maia-200-the-ai-accelerator-built-for-inference/>

9. This is the Google TPU v7 Ironwood Chip:

<https://www.servethehome.com/this-is-the-google-tpu-v7-ironwood-chip/>

3. 1. Фаундри-строительство в Америке:

1. TSMC Arizona в Финиксе, Аризона, в 2024 производство N4, инвестиции 165 млрд. долл. В октябре 2025 выпущен первый чип Nvidia Blackwell
2. Intel Fab 52 в Чандлере, Аризона, в октябре 2025 показан и в 2026 стартовал техпроцесс 18А, инвестиции из «домашнего» пакета 100 млрд. долл.
3. Фабрики Texas Instruments в Шермане, Техас, 45-130нм на 300мм пластинах, инвестиции 60 млрд. долл.
4. Фабрики Micron в Айдахо, Нью-Йорке и Вирджинии с техпроцессом 1гамма (10нм) на 300мм пластинах, инвестиции 200 млрд. долл.

3.2. Фаундри-строительство в Америке: TSMC Arizona



Source: TSMC

3.3. Фаундри-строительство в Америке: TSMC Arizona



3.4. Фаундри-строительство в Америке: Intel Fab 52 (18A)



Cognitar



Source: Intel

3. 5. Фаундри-строительство в Америке: 300mm Fab Texas Instruments в Шермане , Техас



3.6. Фаундри-строительство в Америке: 300mm Fab Texas Instruments в Шермане , Техас



Cognitar



Source: Texas Instruments

3.7. Фаундри-строительство в Америке: Fab Micron в Сиракузах, Нью-Йорк



Cognitar



Source: Micron Technology

3.8. Фаундри-строительство в Америке: Fab Micron в Манассасе, Вирджиния



TERAFAB

JOIN US ON OUR JOURNEY

TESLA

| xAI

| SPACEX

WATCH NOW

SHARE YOUR RESUME

3.9. Фаундри-строительство в Америке: Фабрика TeraFab Илона Маска в Остине, Техас

Tesla, SpaceX, and xAI are launching the most epic chip-building effort ever - combining logic, memory and advanced packaging under one roof.

Terafab will close the gap between today's chip production and the future's demand - a future among the stars.

We aspire to be a galactic civilization

The Kardashev Scale

Type I

Harness all the energy
on a planet

Type II

Harness all the energy
from a star

Type III

Harness all the energy
in a galaxy

3.10. Фаундри-строительство в Америке:
Фабрика TeraFab Илона Маска в Остине, Техас

3.11. Фаундри-строительство в Америке: Фабрика TeraFab Илона Маска в Остине, Техас

More than one kind of chip

AI5



Powering FSD
& Optimus

AI6



Powering
Optimus

D3



Powering
Space

...and Beyond



3.12. Фаундри-строительство в Америке: Фабрика TeraFab Илона Маска в Остине, Техас

План: выпуск 1 миллиона пластин в месяц по 2нм техпроцессу,
что эквивалентно 100-200 миллиардов чипов в год,
и что в 30 раз больше производительности любой фабрики Intel или TSMC

0.5 TW

Current annual U.S. consumption

1 TW/year

Terafab output

Source: Tesla Inc.

3.13. Фаундри-строительство в Америке. Литература:

1. Blackwell Rises from Phoenix Fab: <https://www.hpcwire.com/2025/10/17/blackwell-rises-from-phoenix-fab/>

2. TSMC's New Arizona Fab! Apple Will Finally Make Advanced Chips In The U.S.:
https://www.youtube.com/watch?v=WHat_LYrpQE

3. Intel 18A Enters Production in Arizona as Company Previews Xeon 6+ for Scalable AI and Data Center Workloads:
<https://www.hpcwire.com/off-the-wire/intel-18a-enters-production-in-arizona-as-company-previews-xeon-6-for-scalable-ai-and-data-center-workloads/>

4. Inside Intel's \$20 Billion Chip Factory in the US:
<https://www.youtube.com/watch?v=UDmW-pGzsWs>

4. Inside Texas Instruments' \$60 billion U.S. megaproject, where Apple will make iPhone chips:
<https://www.cnbc.com/2025/08/22/apple-will-make-chips-at-texas-instruments-60-billion-us-project.html>

5. Micron announces expanded U.S. investments in manufacturing and R&D:
<https://www.micron.com/us-expansion?srsIid=AfmBOoobqn-XES05820u4EUpmage4QhNIaAQx2HzSUGo1Hj65vfT23TB>

6. TeraFab project site: <https://terafab.ai/>

Fin