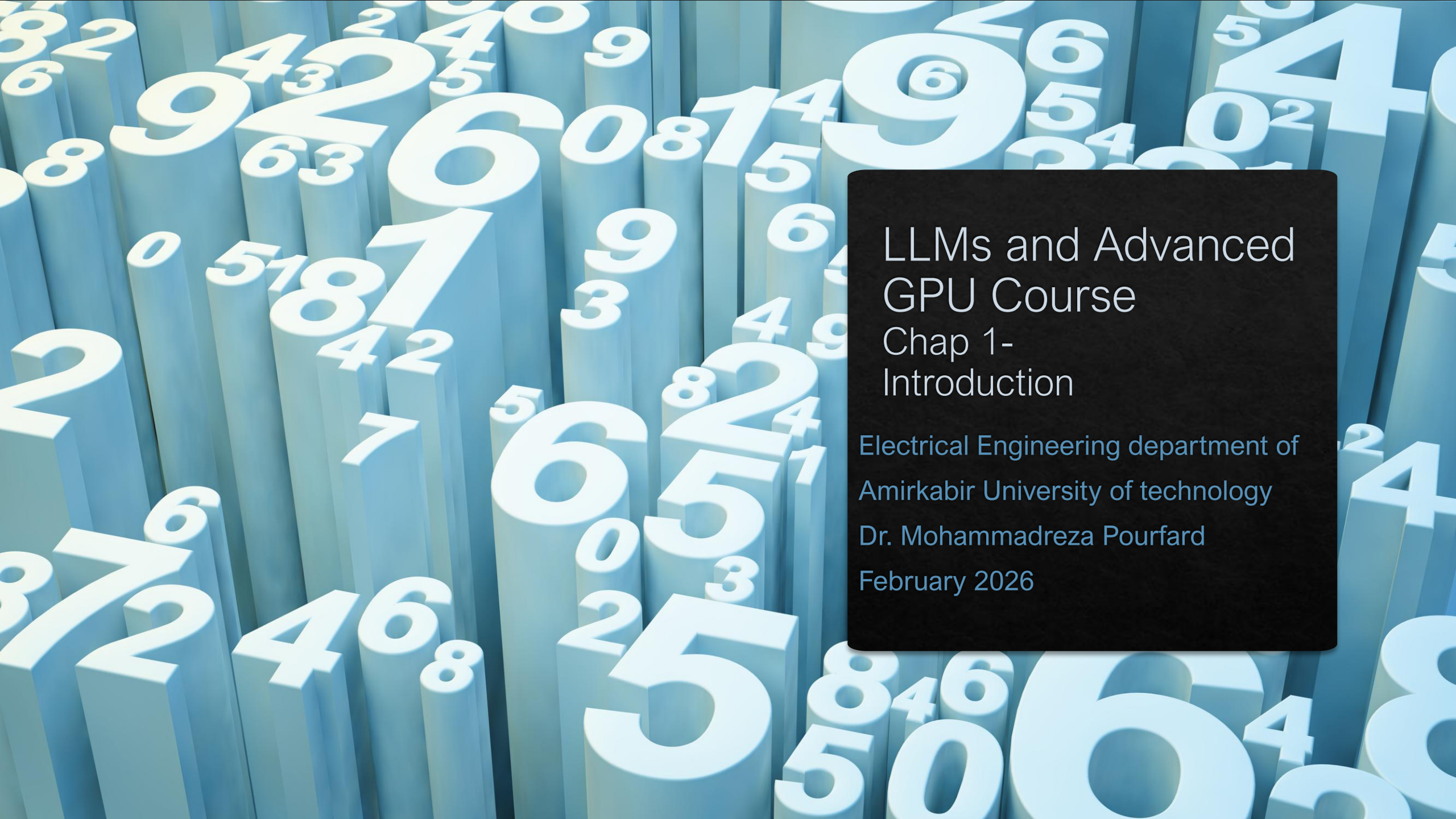


LLMs and Advanced GPU Course

Chap 1- Introduction

Electrical Engineering department of

Amirkabir University of technology

Dr. Mohammadreza Pourfard

February 2026

صفحات درس و استاد

اتاق: طبقه سوم ساختمان ابوریحان، دانشکده مهندسی برق، آزمایشگاه سیستمهای هوشمند دیجیتال، پزشکی شخصی، هوش مصنوعی و GPU

@Electronic_daneshbonyan

pourfardm@gmail.com

+982164543373

http://ssail.aut.ac.ir

ssail.aut.ac.ir

DIGITAL
SMART SYSTEM AND DATA MINING LAB

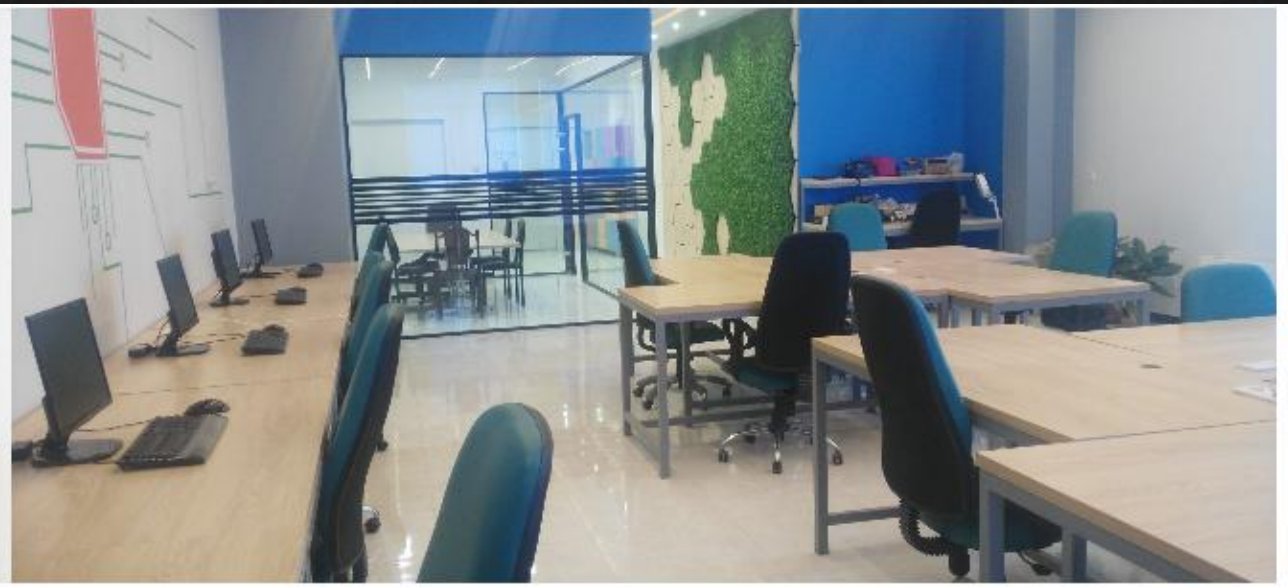
English | Search

Research Groups | Research Projects | People | News | About Us | Admission

Research Groups

In today's rapidly evolving technological landscape, numerous research groups are at the forefront of innovation, each focusing on transformative areas that have the potential to reshape industries and enhance our daily lives. Among these, the Internet of Things (IoT) research group is dedicated to exploring the interconnectedness of devices and sensors, striving to create smart environments that foster efficiency and convenience. Meanwhile, our Artificial Intelligence (AI) team is delving into advanced algorithms and machine learning techniques to develop systems that can learn, adapt, and make informed decisions, pushing the boundaries of automation and intelligence. Additionally, the Big Data research group is

Laboratory page



Profile Contents

Home | Courses

- About
- Theses
- Research Interests
- Research Groups
- Research Projects
- Current Students
- Employment Records
- Contact
- Publications
- News

About

Adjunct Professor of Amirkabir University of Technology

Theses

B.S Degree:
Simulating Multi-Level Cache Memory

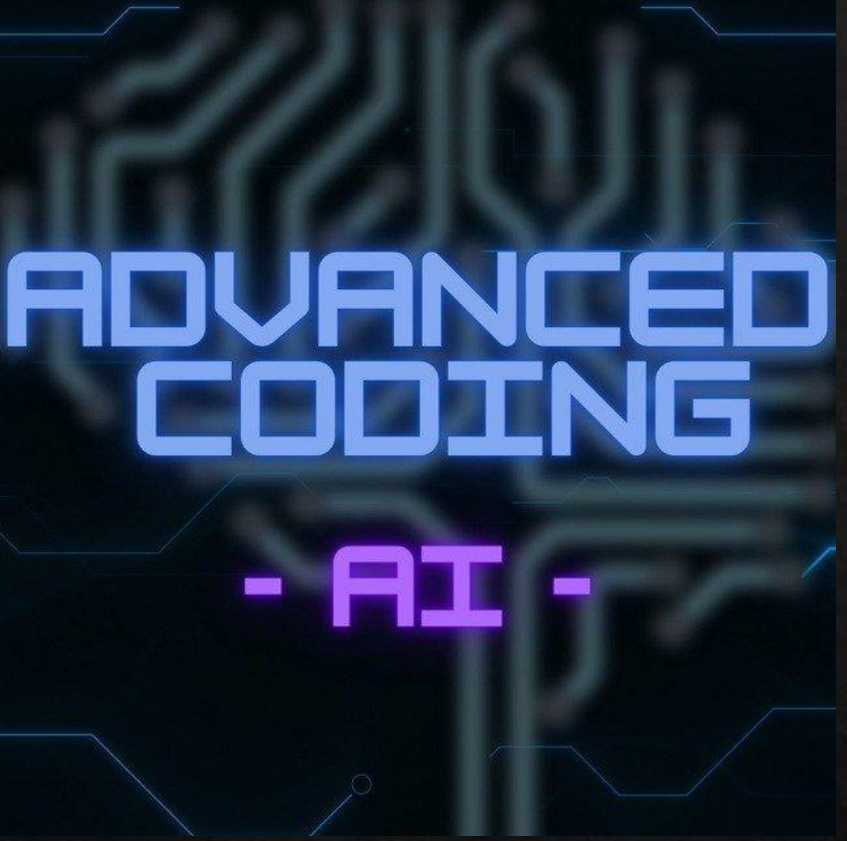
M.S Degree:
Detection and Tracking of a Human in different positions of its body

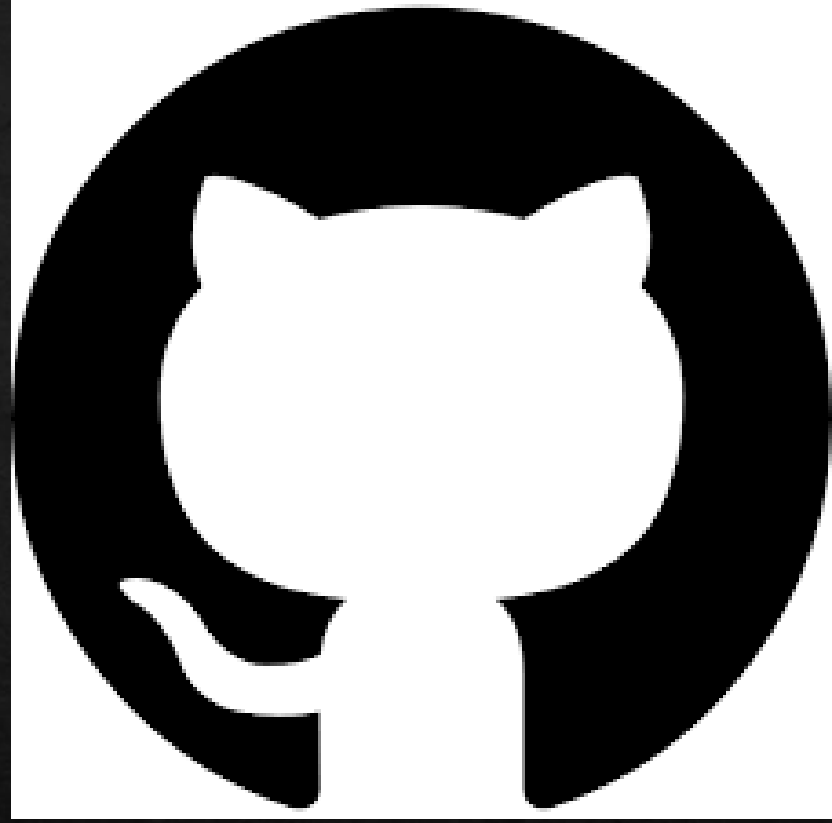
PhD Degree:
Texture Analysis and Separation for Characterization of Material's Structure through

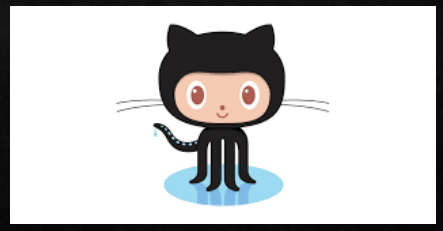
Research Interests

- Genetics data processing,

کانالها و گروههای مرتبط با درس



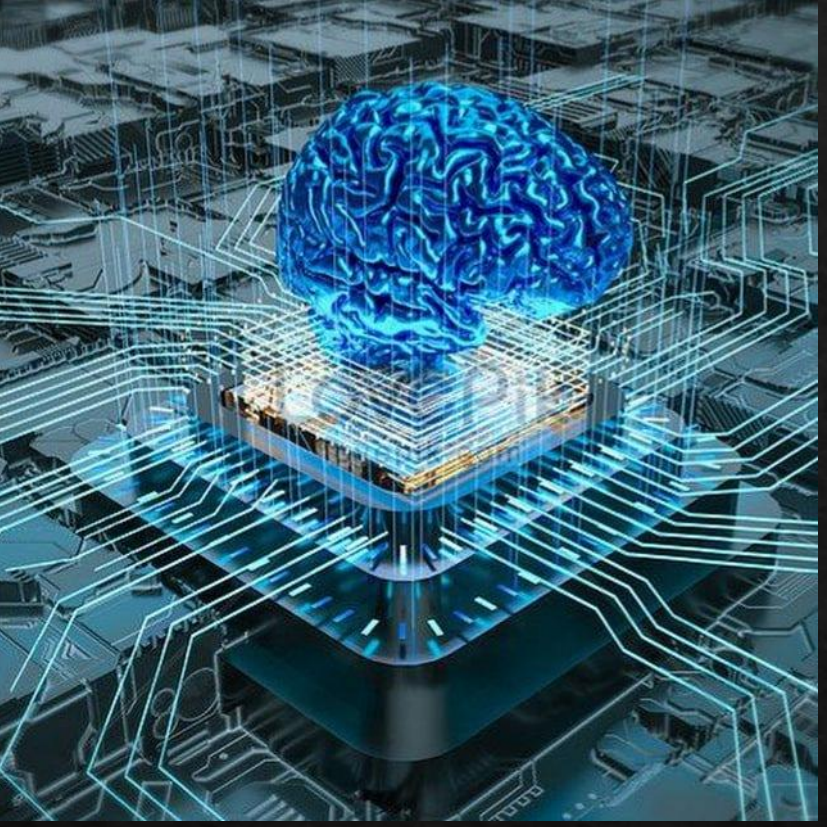


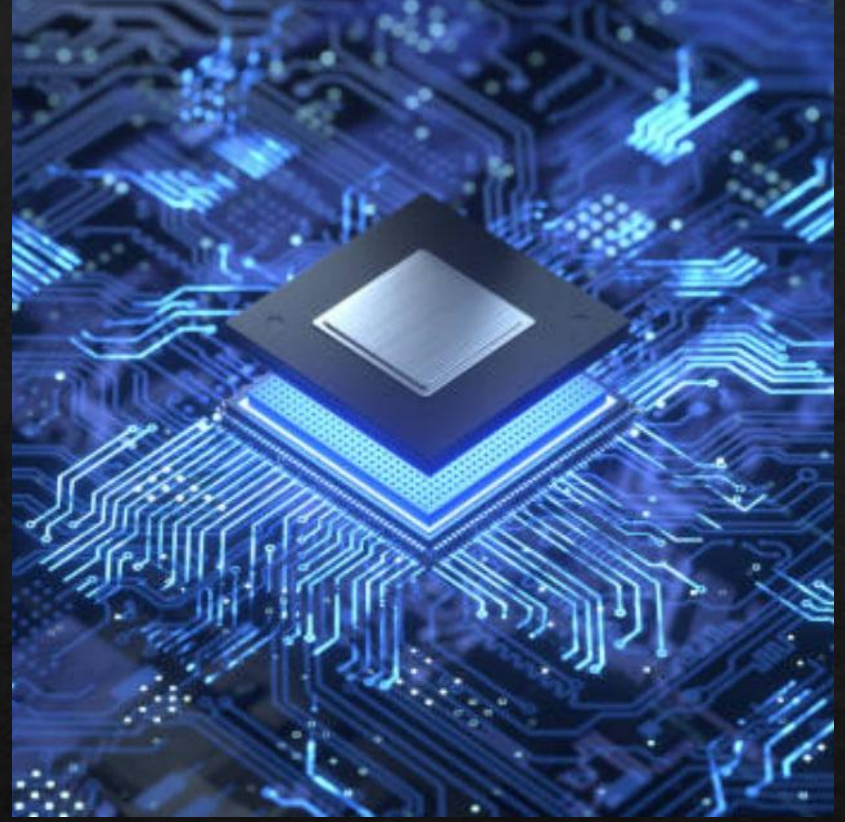


Telegram Channel:
https://t.me/Advanced_programming_algorithm_c

Telegram Group:
https://t.me/Advanced_Programming_Algorithm

GitHub Page:
https://github.com/Pourfardm/

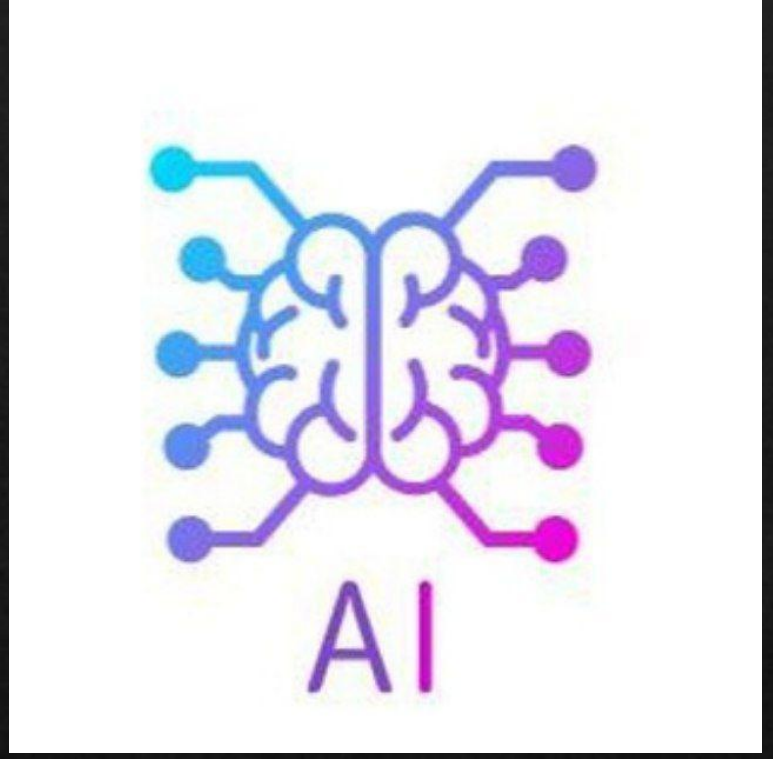




Telegram Channel:
https://t.me/fpga_digital_logic_design

Telegram Group:
https://t.me/Advanced_digital_analog_design

YouTube Channel:
https://www.youtube.com/channel/UCa8ZFUxp37Vy-KEbx7SbbHQ



Telegram group:
https://t.me/AI_Bigdata_amirkabir_group

Telegram Channel:
https://t.me/Deeplearning_BigData_AI

Private Telegram Group:
For class student

Telegram Channel:
https://t.me/AI_Nuroscience_Genomic_Data

YouTube Channel:
https://www.youtube.com/channel/UCa8ZFUxp37Vy-KEbx7SbbHQ

مروری بر نسل دانشگاه‌ها در جهان

اهمیت ورود به دانشگاه نسل سوم و

ارایه درس متناسب با دانشگاه نسل سوم

نسل های مختلف دانشگاهها در جهان

عنوان	نسل اول	نسل دوم	نسل سوم	نسل چهارم
هدف	کاشفیت از حقیقت عالم	توسعه فناوری تحقیق و پژوهش	تولید ثروت و ارزش افزوده	ارزش افزوده قابل توجه ایجاد GDP قابل توجه نسبت به GDP کشور
تاکید	آموزش	آموزش پژوهش محور	پژوهش نیاز محور	توجه به نیازهای کلیدی و مهم کشور
شاخص ارزیابی	کیفیت تدریس ترجمه کتاب	مقالات ISI (امتیاز بالا بدون سقف در آیین نامه ارتقای اساتید)	حجم قراداد دانشگاه با صنعت درآمد زایی دانشگاه ثبت پتنت با ارجاعات بالا تالیف کتاب	نسبت دانشگاه با شرکتهای بزرگ خصوصی با سهم قابل توجه در اقتصاد کشور
مشکلات	عدم آشنایی با مرز علم	پراکنده کاری شناسایی موضوعات مقاله خور تغییر مرتب موضوعات عدم توجه به نیازهای کشور	عدم اثرگذاری قابل توجه در اقتصاد ملی	مخالفت کشورهای غیرهمسو با توسعه صادرات کشورهای رقیب

دانشگاهها در ایران عمدتا نسل هستند.

نکته مهم این درس

به دلیل رویکرد دانشگاه نسل سوم

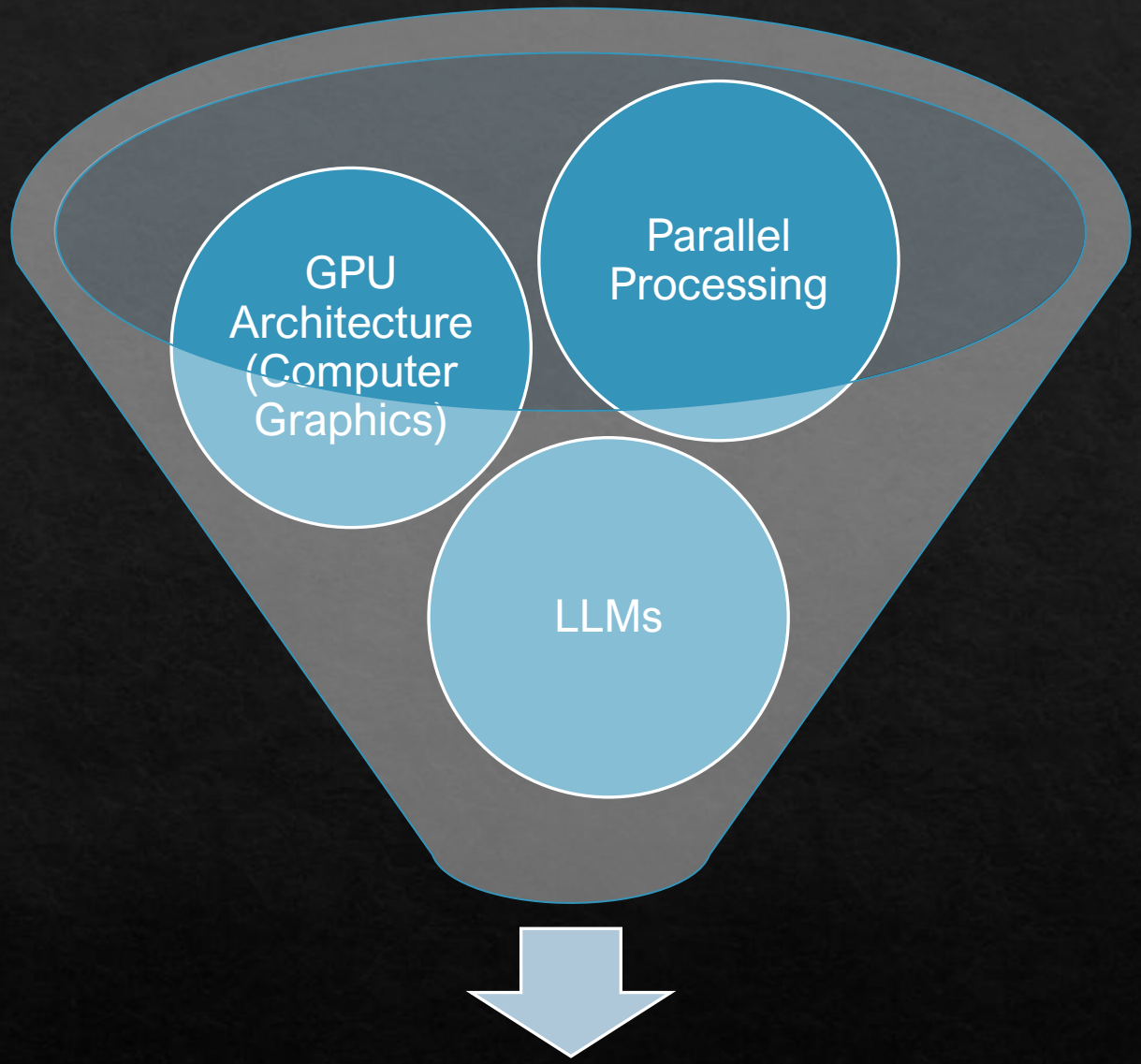
تمرکز این درس فقط روی الگوریتم نیست

اهداف درس

اهداف کلیدی درس

- آشنایی با برنامه نویسی موازی و زبان برنامه نویسی CUDA
- آشنایی با معماری داخلی آخرین نسل GPU کمپانی Nvidia (Blackwell 2024)
- آشنایی با آخرین مدلهای هوش مصنوعی در حوزه LLM
- فهم ارتباط آخرین دستاوردهای پردازنده گرافیکی با آخرین مدلهای هوش مصنوعی
- آشنایی با کاربردهای مفید هوش مصنوعی و LLM ها به منظور راه اندازی کسب و کار
- آشنایی با ابزارهای مورد استفاده در شرکتهای دانش بنیان برای هوش مصنوعی
- بررسی مدلهای LLM به صورت عملی بر روی داده های مختلف (شامل داده های پزشکی)

محتوی درس


GPU Architecture (Computer Graphics)
Parallel Processing
LLMs

This course

جامعہ مخاطب درس


علاقمندان به آخرین پیشرفت های حوزه هوش مصنوعی (LLM) –(هوش مصنوعی)
علاقمندان به معماری –(نسل های آخر GPU (سیستم های دیجیتال)
علاقمندان به برنامه نویسی موازی (نرم افزار)
علاقمندان به پیاده سازی عملی بر روی داده های واقعی بزرگ –(کلان داده)
جامعه مخاطب درس

بارم بندی

بارم بندی درس (۲۴ نمره)

- نمرات اصلی
  - پایان ترم ۸ نمره
  - میان ترم -نداریم
  - تکالیف ۶ نمره
  - پروژه پایانی ۴ نمره
  - ارایه مقاله ۲ نمره
- نمرات امتیازی
  - تکالیف اختیاری ۲ نمره
  - ارزیابی استاد از دانشجو ۱ نمره اضافی (به صورت استثنا)
  - تولید محتوی ۱ نمره اضافی

سیلابس درس

بخش اول) بخش پردازش موازی و معماری GPU های نسل آخر NVidia (۳۳ درصد)

- **الف)سیستم های پردازش موازی**
  - آشنایی با معماری سیستم‌های چندهسته‌ای
  - آشنایی با برنامه‌نویسی چندنخی، مدل‌ها و زبانهای برنامه‌نویسی مرتبط با آن
  - آشنایی با مفاهیم پردازش برداری، SIMD، SSE، AVX و نحوه استفاده از آن
  - پیاده‌سازی الگوریتم‌ها به صورت چندنخی و برداری با استفاده از زبان‌های برنامه‌نویسی چندهسته‌ای (OpenMP)
  - آشنایی با روشهای متداول همگام‌سازی نخ، قفل، مانع

بخش اول) بخش پردازش موازی و معماری GPU های نسل آخر NVidia (۳۳ درصد)

- **آشنایی با معماری پردازنده‌های گرافیکی، سلسله مراتب حافظه در GPU**
- آشنایی با تاریخچه معماری های GPU کمپانی Nvidia شامل Fahrenheit, Kelvin Turing, Hopper, Blackwell, Blackwell Ultra
- آشنایی با بلوک های چیپ گرافیکی CB202 Chipset
- آشنایی با واحدهای GPC, Memory Controller, Cache, AMP & Giga Thread Engine, NVENC / NVDEC, Optical Flow Engine, PCI Express 5.0 Host Interface.
- بررسی ماژول Mixed FP32/INT32 module
- آشنایی با معماری و قابلیت های هسته های پردازشی Ray tracing core, Cude core, Tensor core
- آشنایی با واحد های Warp Schedulers & Dispatch Units
- آشنایی با واحد Texture Units: perform texture fetches and filtering
- آشنایی با بخش Load/Store Units (LD/ST): handle memory access (global/local memory 10-reads/writes)
- آشنایی با بخش Register File: private storage per thread
- آشنایی با حافظه های سریع Shared Memory / L1 Cache: memory accessible by all threads in the SM. و پیاده سازی سریع الگوریتم ها به کمک این حافظه ها
- آشنایی با واحد پردازش هوش مصنوعی AI Management Processor (AI kernels, Multi-GPU Scaling, tensor parallelism, Training massive models (LLMs), data parallelism) و نحوه استفاده از آنها در پیاده سازی مدلهای LLM

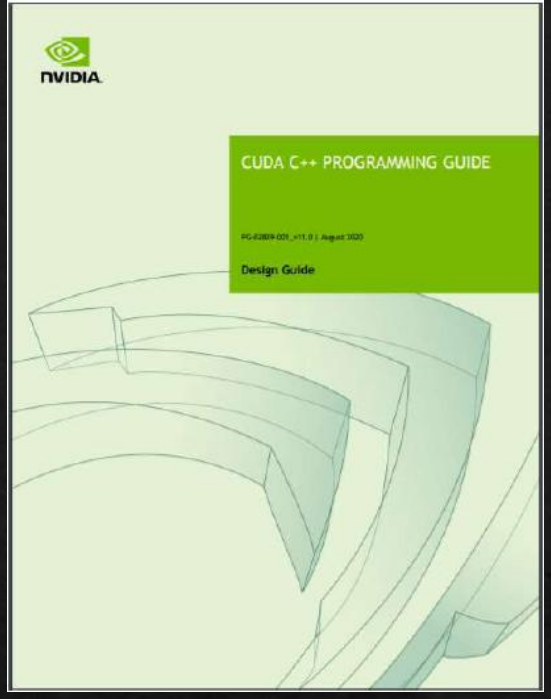
بخش اول) بخش پردازش موازی و معماری GPU های نسل آخر NVidia (۳۳ درصد)

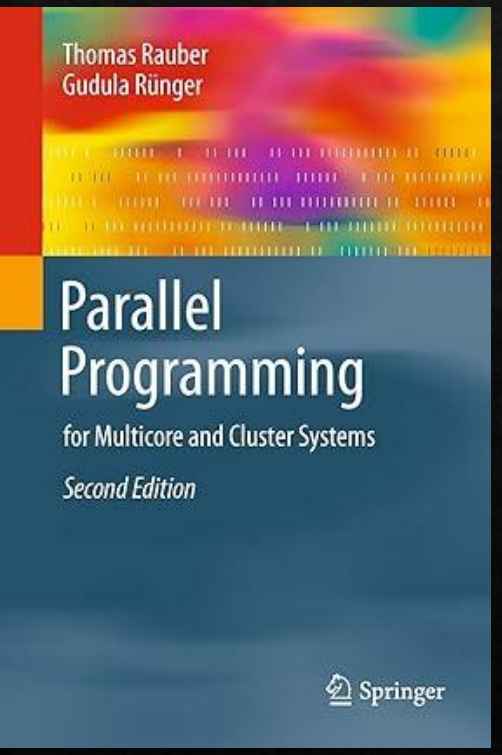
- آشنایی با معماری CUDA و GPU Driver
  - آشنایی با بخش User-Space Driver of GPU (Resource & State Management (Logical View), Build Command Buffers, Interface with Kernel Driver, API Front-End, State Manager, Compiler Stack, Resource Manager, Command Buffer Builder, Caching & Pipeline Database)
  - آشنایی با بخش Kernel-Space Driver of GPU (Context Manager, Memory Manager, Command Processor Interface, Scheduler / Dispatcher, Synchronization Manager, Interrupt Handler, Power & Thermal Control, Virtualization Layer)

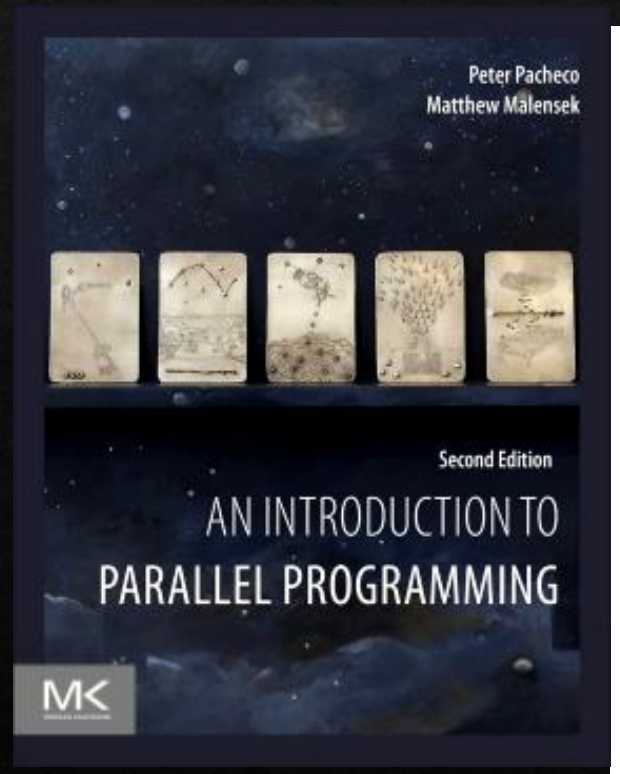
بخش اول) بخش پردازش موازی و معماری GPU های نسل آخر NVidia (۳۳ درصد)

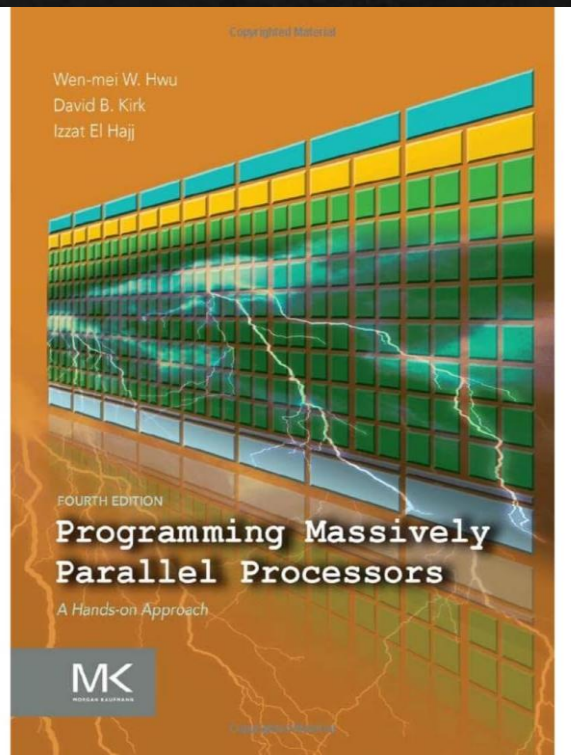
- آشنایی با برنامه‌نویسی موازی GPU و زبان برنامه‌نویسی CUDA
  - ارائه مثال‌هایی از پیاده‌سازی کاربردهای معمول در GPU
  - برنامه نویسی با CUDA
  - تغییر در کرنلهای CUDA برای یک منظور خاص همانند FFT با تعداد نقاط دلخواه

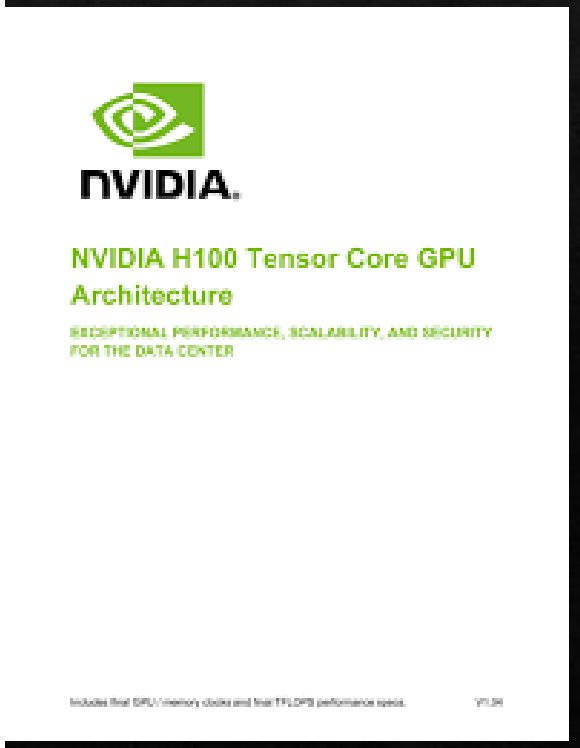
مراجع این بخش











بخش دوم) شبکه های LLM (۳۴ درصد)

- **بررسی شبکه های LLM و آخرین پیشرفت ها در این حوزه**
  - آشنایی با مبانی Large Language Modeling
  - آشنایی با Tokenizationو اثرات پنهان آن
  - آشنایی با مدل‌های پایهGPT، BERT، T5
  - آشنایی با Reasoningدر LLM ها
  - آشنایی با انواع Hallucination و عدم قطعیت
  - آشنایی با انواع روشهای RAGو اتصال به دانش خارجی (همانند Naïve RAG, Advanced RAG, Modular RAG)
  - ارزیابی و بحران Benchmark
  - بررسی معماری آخرین نسل Llama, GPT, Deep Seek و سایر مدلهای بروز جهانی

بخش دوم) شبکه های LLM (۳۴ درصد)

- مراجع درس در این بخش، مقالات بروز جهانی از سال ۲۰۲۲ به بعد خواهد بود.
- Gao, Yunfan, et al. "Retrieval-augmented generation for large language models: A survey." *arXiv preprint arXiv:2312.10997* 2.1 (2023).
- Hoffmann, Jordan, et al. "Training compute-optimal large language models." *arXiv preprint arXiv:2203.15556* (2022).
- Wei, Jason, et al. "Emergent abilities of large language models." *arXiv preprint arXiv:2206.07682* (2022).
- Ji, Ziwei, et al. "Survey of hallucination in natural language generation." ACM computing surveys 55.12 (2023): 1-38.
- اسلاید های درس دانشگاهی CMU – 11-667: Large Language Models
- اسلاید های درس دانشگاهی Stanford – CS324 / CME 295
- اسلاید های درس دانشگاهی Princeton – COS 597G

بخش سوم) کاربرد LLM و GPU بر روی پردازش داده های مختلف (شامل داده های پزشکی) (۳۳ درصد)

- پیاده سازی روشهای LLM به کمک GPU بر روی انواع مختلفی از داده ها
- نحوه استفاده از مفهوم RAG در استفاده از پایگاه های داده متغیر با زمان ژنتیک در مدلهای LLM
- نحوه استفاده از واحدهای AI Kernel , AI Processor در پردازش داده
- پیاده سازی برخی از الگوریتم های سنگین بر روی GPU های نسل آخر کمپانی Nvidia با استفاده از حافظه سریع L1 Cache memory.
- آشنایی با فرمت تصاویر پزشکی همانند MRI, CT-scan, Pet-scan, Radiology, Mammography و نقش هر یک از این داده ها در فهم مسایل پزشکی
- آشنایی با فرمت داده های سیگنال های پزشکی همانند نوار مغز، نواز قلب، نوار عضله و عصب و نقش هر یک از این داده ها در فهم مسایل پزشکی
- آشنایی با مفاهیم پایه ژنتیک همانند DNA, RNA, Protein و Gut Microbiota
- آشنایی با سطوح تحلیل داده های پزشکی شامل ژنومیکس، اپی ژنومیکس، ترنسکریپتومیکس، پرتئیومیکس، متابولومیکس، فنومیکس
- نحوه Fine-tunning مدلهای LLM برای تحلیل داده های مختلف Multi-omics
- نحوه طراحی لایه های Encoder , Decoder در مدلهای LLMبر روی تصاویر، سیگنال و ژنتیک
- آشنایی با پایگاههای داده ژنتیک در پزشکی

بخش سوم) کاربرد LLM و GPU بر روی پردازش داده های مختلف (شامل داده های پزشکی) (۳۳ درصد)

- مراجع درس در این بخش، مقالات بروز جهانی از سال ۲۰۲۴ به بعد خواهد بود.

- 1- Sarumi, Oluwafemi A., and Dominik Heider. "Large language models and their applications in bioinformatics." *Computational and Structural Biotechnology Journal* 23 (2024): 3498-3505.
- 2- Van, Minh–Hao, Prateek Verma, and Xintao Wu. "On large visual language models for medical imaging analysis: An empirical study." *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2024.
- 3- Kao, Jyun-Ping, and Huan-Tang Kao. "Large Language Models in radiology: A technical and clinical perspective." *European Journal of Radiology Artificial Intelligence* (2025): 100021.

موضوعات تکالیف

- ۱- پیاده سازی چند الگوریتم به صورت موازی بر روی بستر GPU با استفاده از CUDA
- ۲- استفاده از Cacle L1 از GPU برای پردازش سریع داد های ژنتیک
- ۳- انجام Fine tunning یک شبکه LLM متن باز
- ۴- پیاده سازی یک سیستم RAG
- ۵- طراحی لایه های انکودری و دیکدری LLM برای روی داده های پزشکی

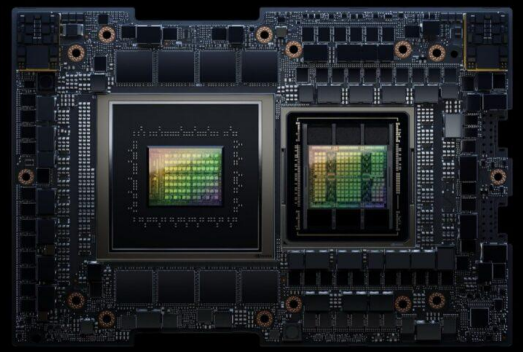
استفاده از ۳۷ گیگابایت رم از ۴۰ گیگابایت رم A100 GPU فقط برای پردازش اطلاعات یک فرد

NVIDIA GH200

Built for the new era
of AI supercomputing

CPU to GPU Bandwidth	Memory Bandwidth
900 GB/s NVLink-C2C	4.9 TB/s HBM3e per GPU

Energy Efficiency	QFT Quantum Simulation	RAG LLM Inference
1.9X Performance vs H100	90X Performance vs dual x86 CPU	100X Performance vs dual x86 CPU



624 GB High-Speed Memory | 4.9 TB/s | 4 PF AI Perf | 72 Arm Cores

مهارت هایی که در درس پیدا خواهید کرد

مهارت هایی که در درس کسب خواهید کرد

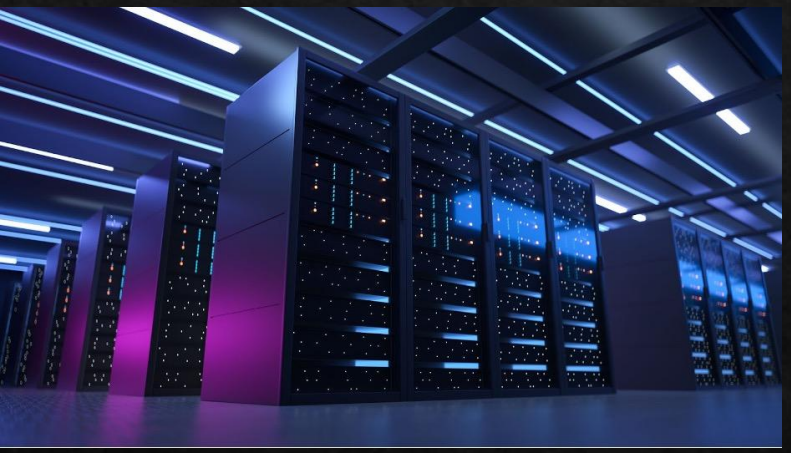
- برنامه نویسی با Python و یادگیری حین انجام پروژه
- کار با پلتفرم برنامه نویسی موازی CUDA و یادگیری حین انجام پروژه
- انجام چند پروژه واقعی در حوزه LLM
- تقویت تفکر موازی و تبدیل الگوریتم های سری به موازی و استفاده بهینه از GPU بجای CPU
- آشنایی با معماری GPU های مختلف به منظور انتخاب GPU بهینه برای حل یک مساله
- کار بر روی یک نمونه داده واقعی بزرگ
- نحوه کار با GPU های نسل آخر Nvidia از طریق کرایه سرور
- کار در بستر Linux و یادگیری حین انجام پروژه
- آشنایی با مقالات جدید حوزه LLM

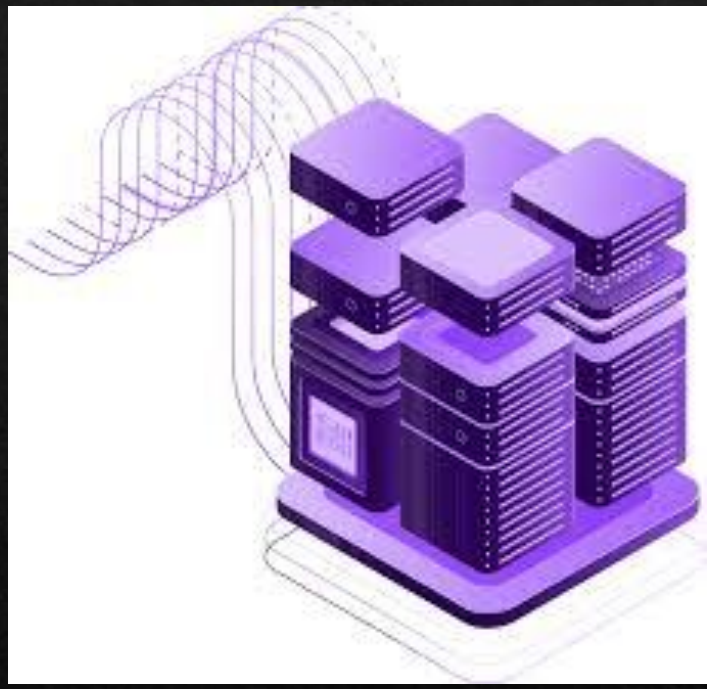
چرایی تعریف درس

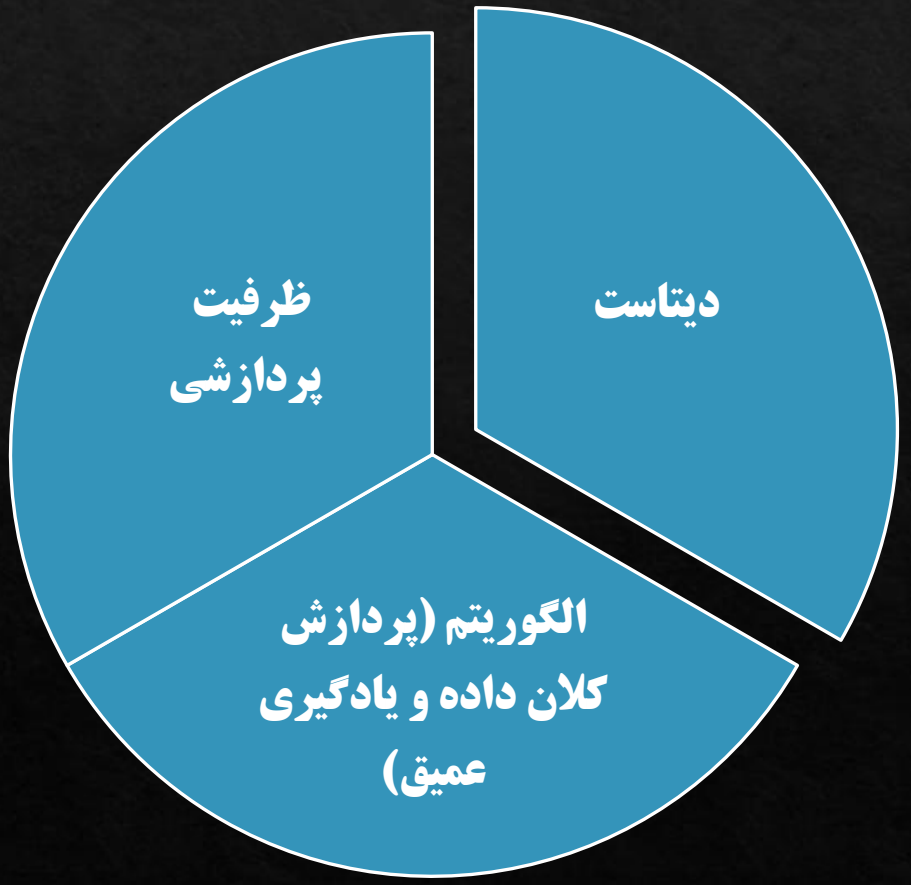
(یکی از جدیدترین دروس دانشگاه

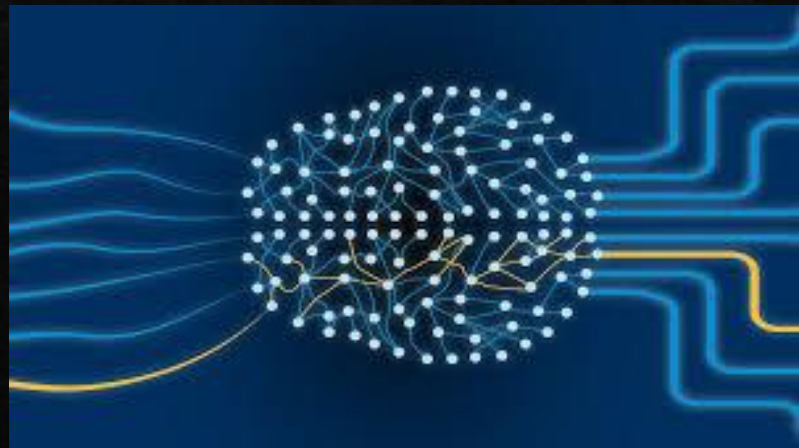
و در مرز علم جهانی)

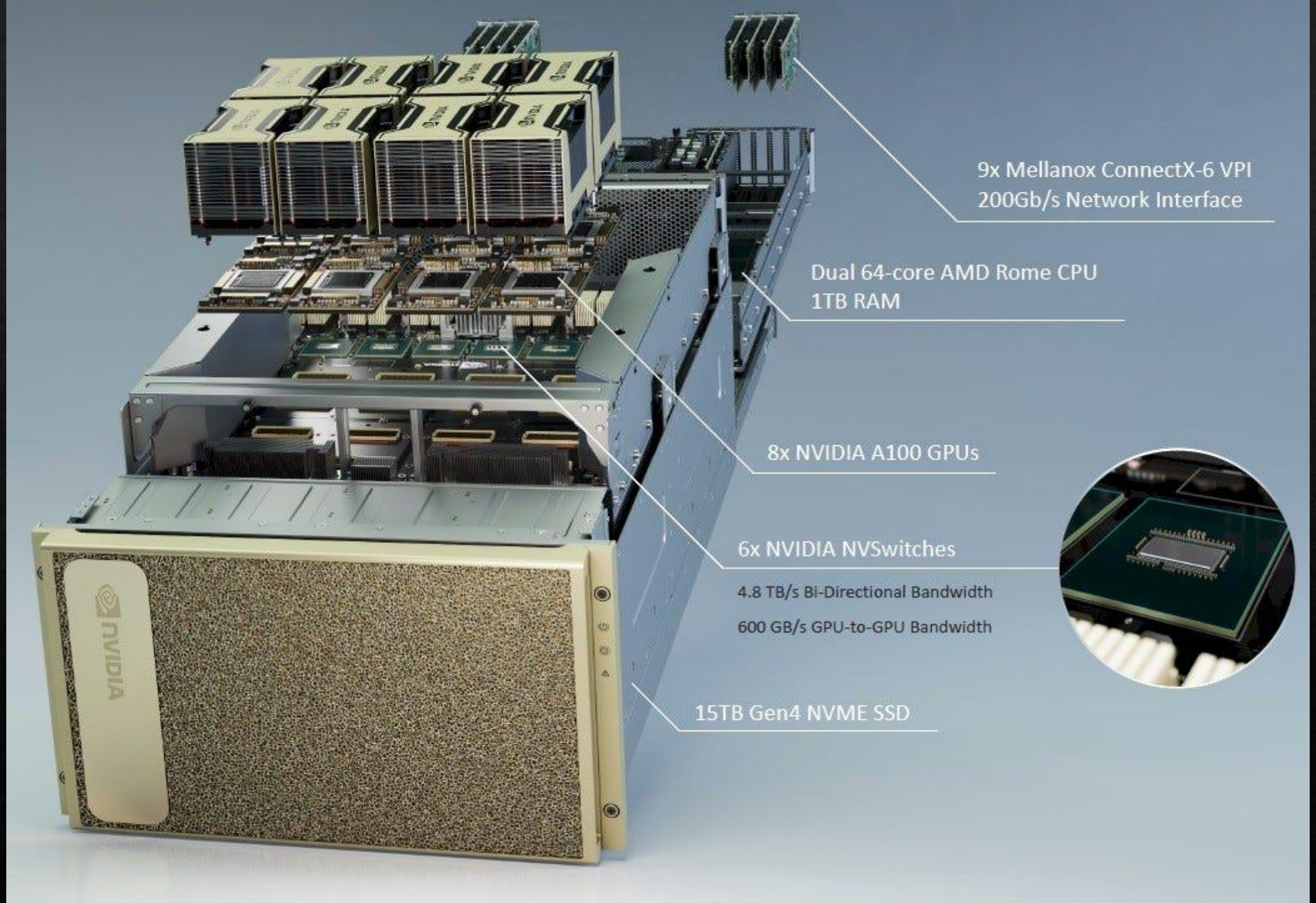
سه راس مهم هوش مصنوعی







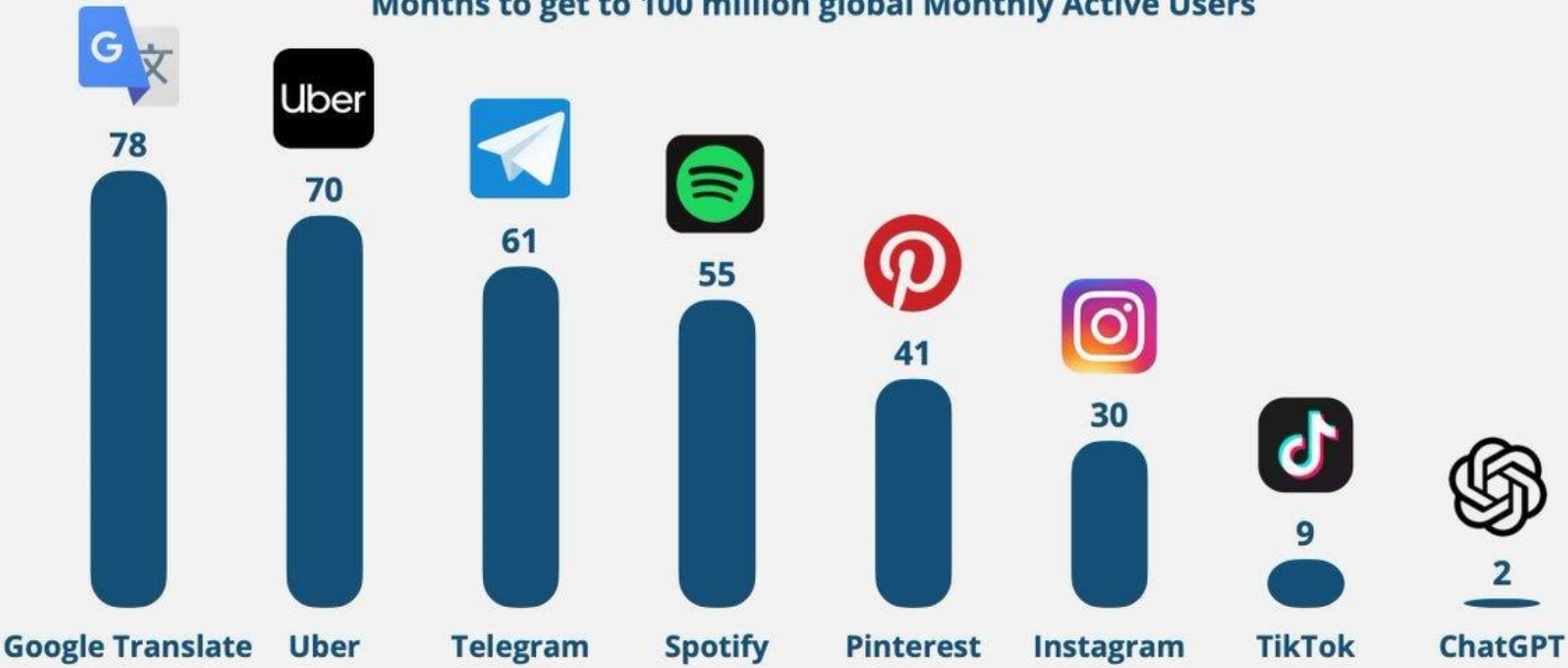



9x Mellanox ConnectX-6 VPI
200Gb/s Network Interface
Dual 64-core AMD Rome CPU
1TB RAM
8x NVIDIA A100 GPUs
6x NVIDIA NVSwitches
4.8 TB/s Bi-Directional Bandwidth
600 GB/s GPU-to-GPU Bandwidth
15TB Gen4 NVME SSD
NVIDIA

اهمیت موضوع درس

Time to Reach 100M Users

Months to get to 100 million global Monthly Active Users



Source: UBS / Yahoo Finance

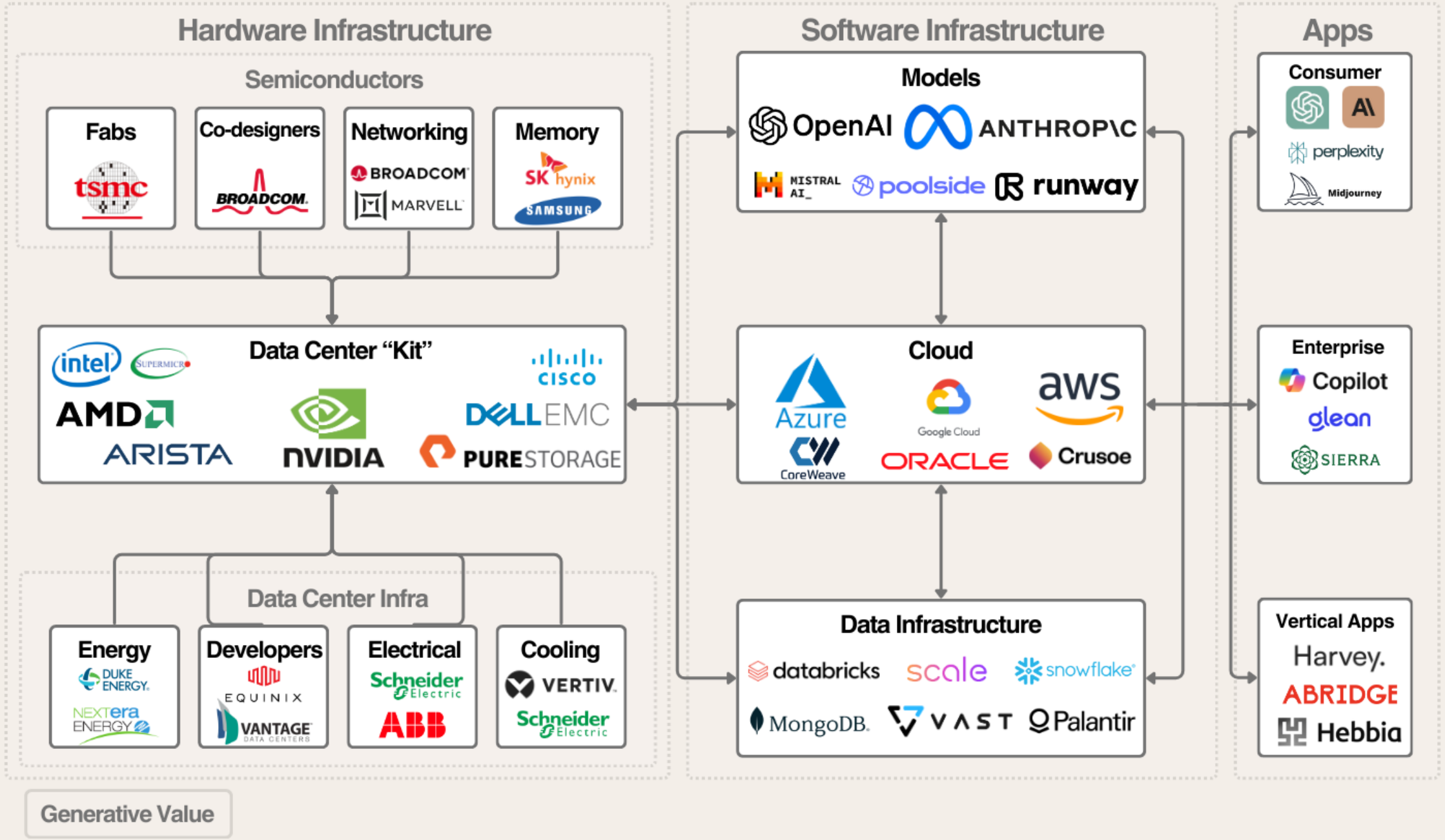
@EconomyApp

APP ECONOMY **INSIGHTS**

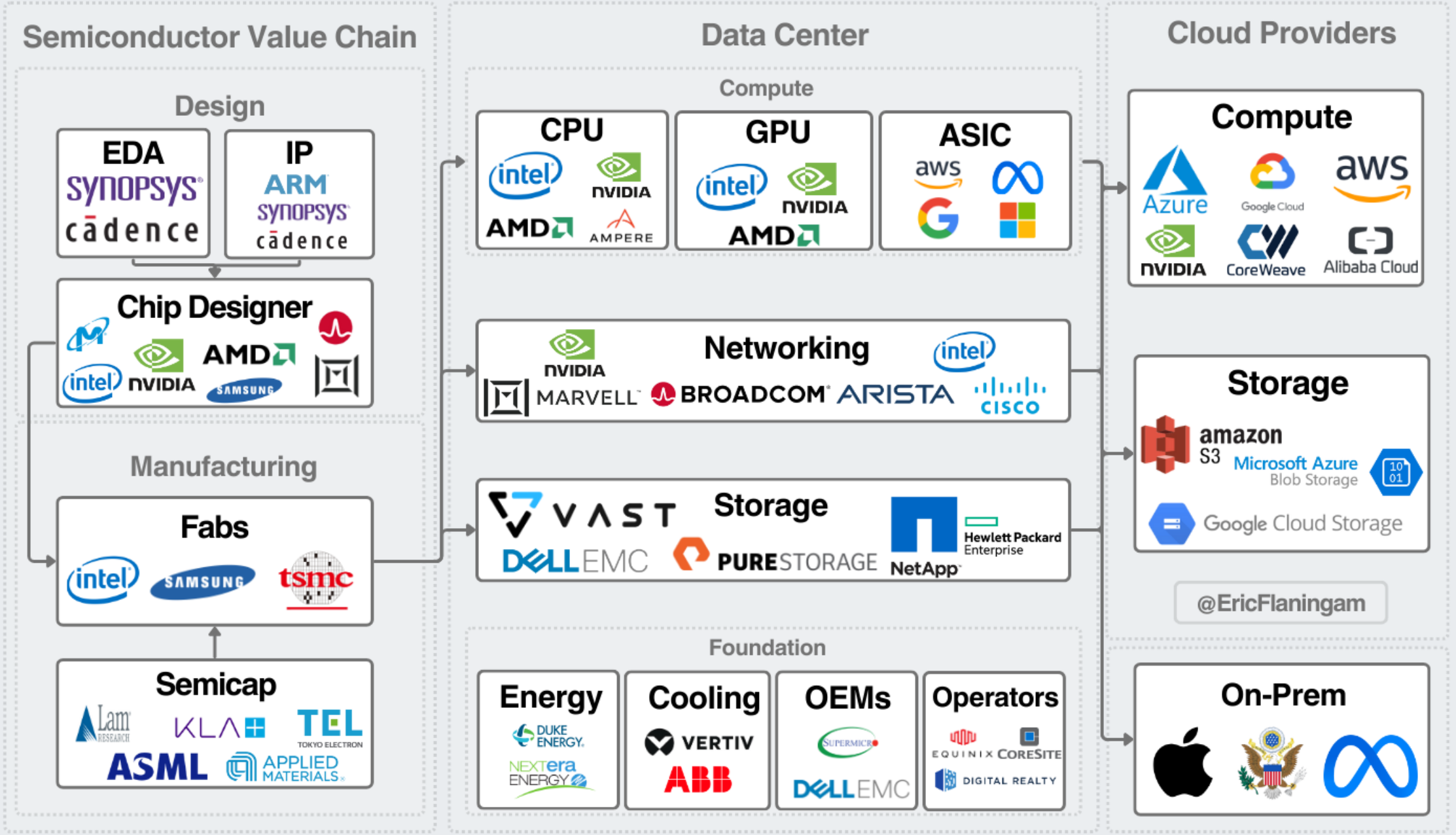
تحولات هوش مصنوعی در علوم مختلف در جهان

- حذف پزشکی سنتی و جایگزینی آن با پزشکی شخصی (Personalized medicine)
- حذف داروسازی سنتی و ایجاد داروی ویژه هر فرد با استفاده از پردازش داده ژنتیک
- از بین رفتن دموکراسی
- حذف رانندگی سنتی و ایجاد خودرو هوشمند و جاده هوشمند و نقش مهم رگولاتوری داده در آینده
- حذف وکالت در آینده
- نقش مهم مهندسان داده در آینده

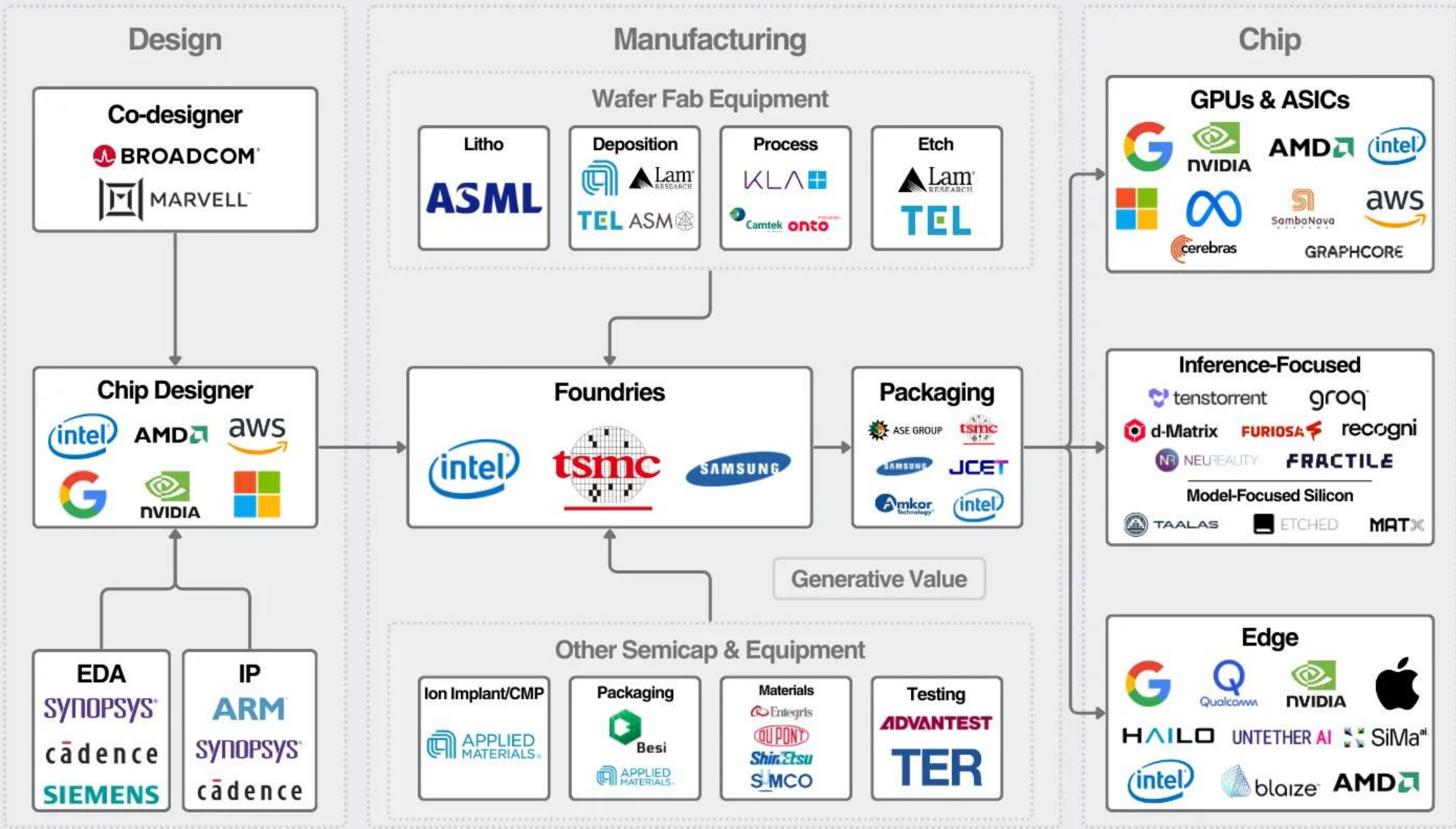
The AI Value Chain



Data Center Value Chain



AI Semiconductor Value Chain



*Not an exhaustive list of companies/segments. Companies may fall into multiple segments. Several inference-focused chips can be used for training.

Semiconductor Companies Ranked by Market Cap

Nvidia's market cap of $1.2T is 2.5x TSMC's

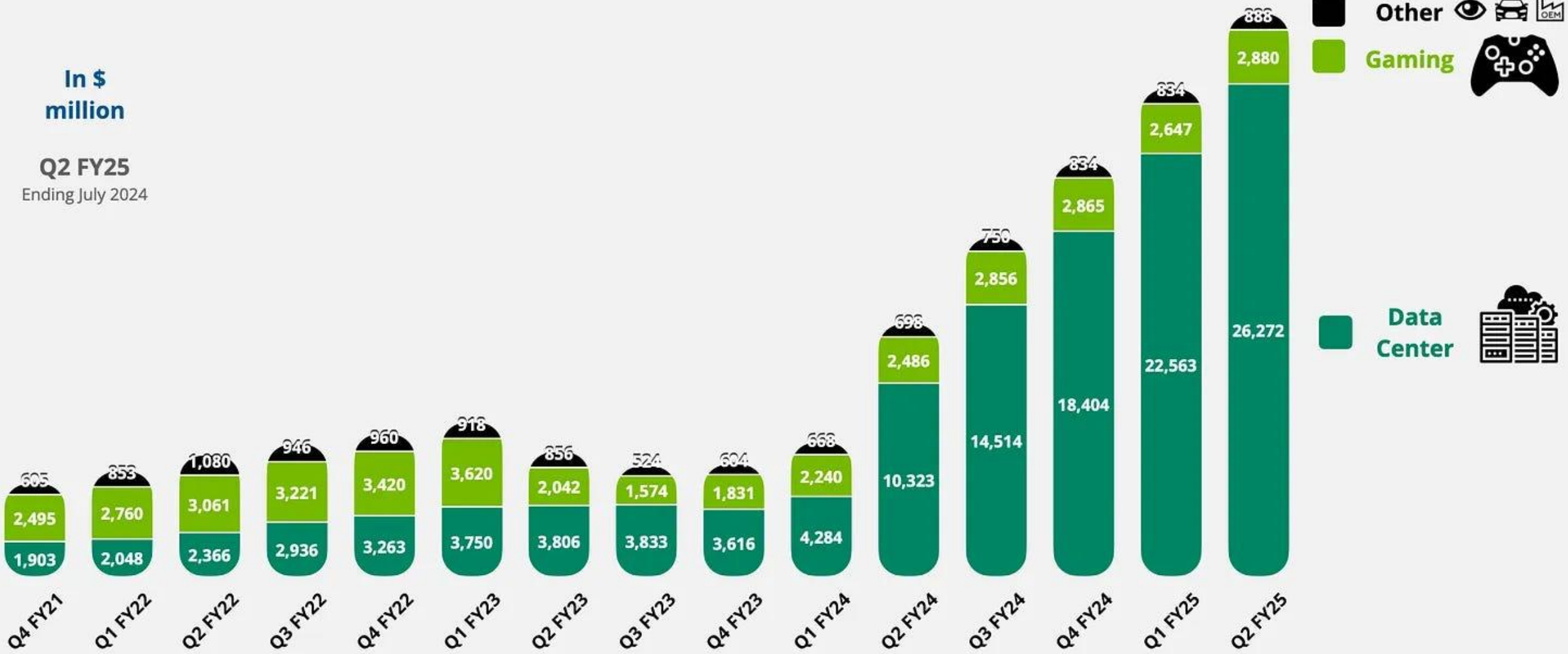
For reference, TEL's market cap is $70B.

Rank	Name	Market Cap	Price	Today	Price (30 days)	Country
1	NVIDIA NVDA	$4.621 T	$189.82	▲ 1.02%		USA
2	Apple AAPL	$3.888 T	$264.58	▲ 1.54%		USA
3	Alphabet (Google) GOOG	$3.809 T	$314.90	▲ 3.74%		USA
4	Microsoft MSFT	$2.952 T	$397.23	▼ 0.31%		USA
5	Amazon AMZN	$2.255 T	$210.11	▲ 2.56%		USA
6	TSMC TSM	$1.921 T	$370.54	▲ 2.82%		Taiwan
7	Meta Platforms (Facebook) META	$1.658 T	$655.66	▲ 1.69%		USA
8	Saudi Aramco 2222.SR	$1.657 T	$6.85	▲ 0.39%		S. Arabia
9	Broadcom AVGO	$1.577 T	$332.65	▼ 0.40%		USA
10	Tesla TSLA	$1.545 T	$411.82	▲ 0.03%		USA

NVIDIA Revenue Breakdown

In $ million

Q2 FY25
Ending July 2024

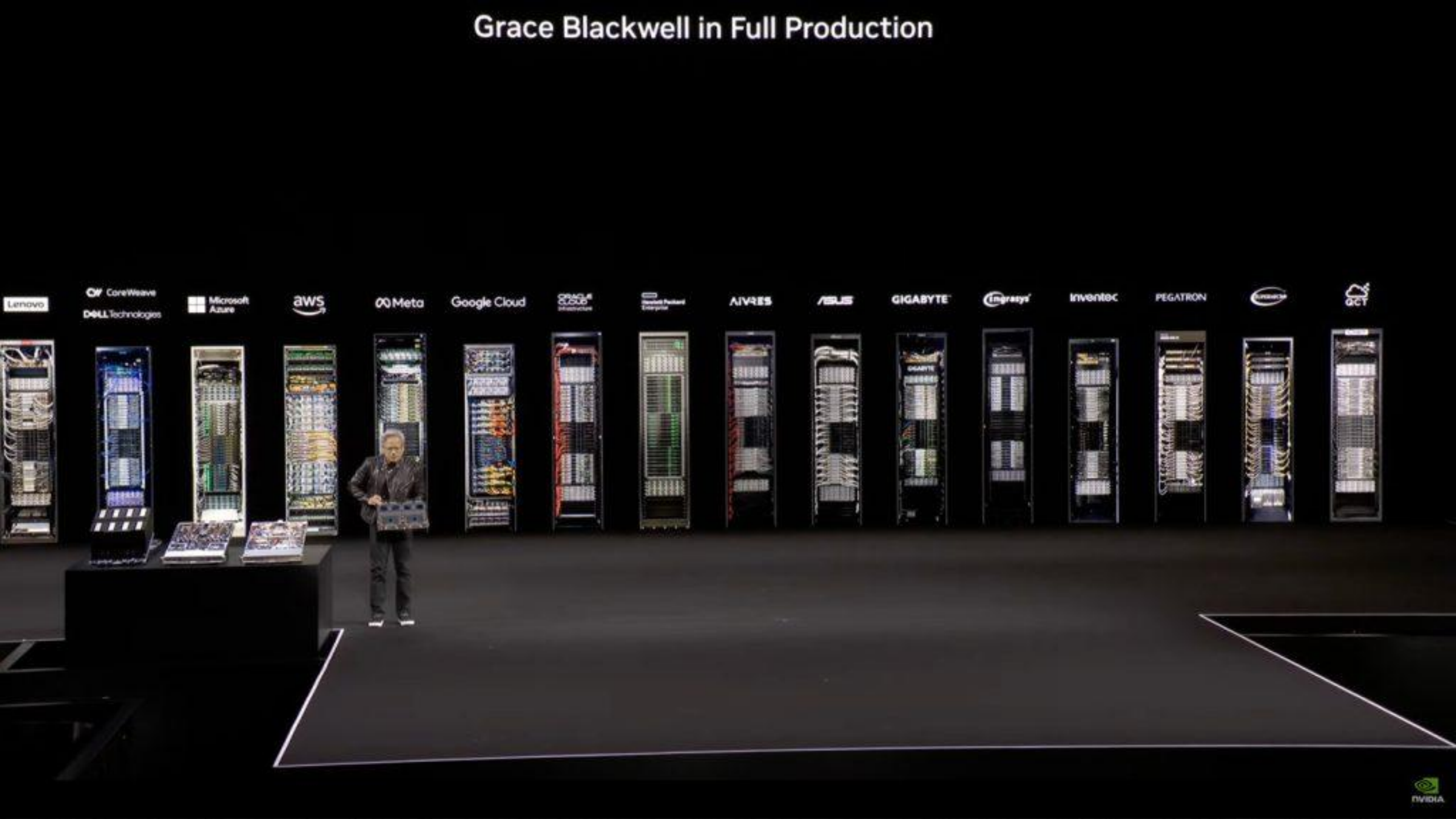


HOW THEY MAKE MONEY

appeconomyinsights.com

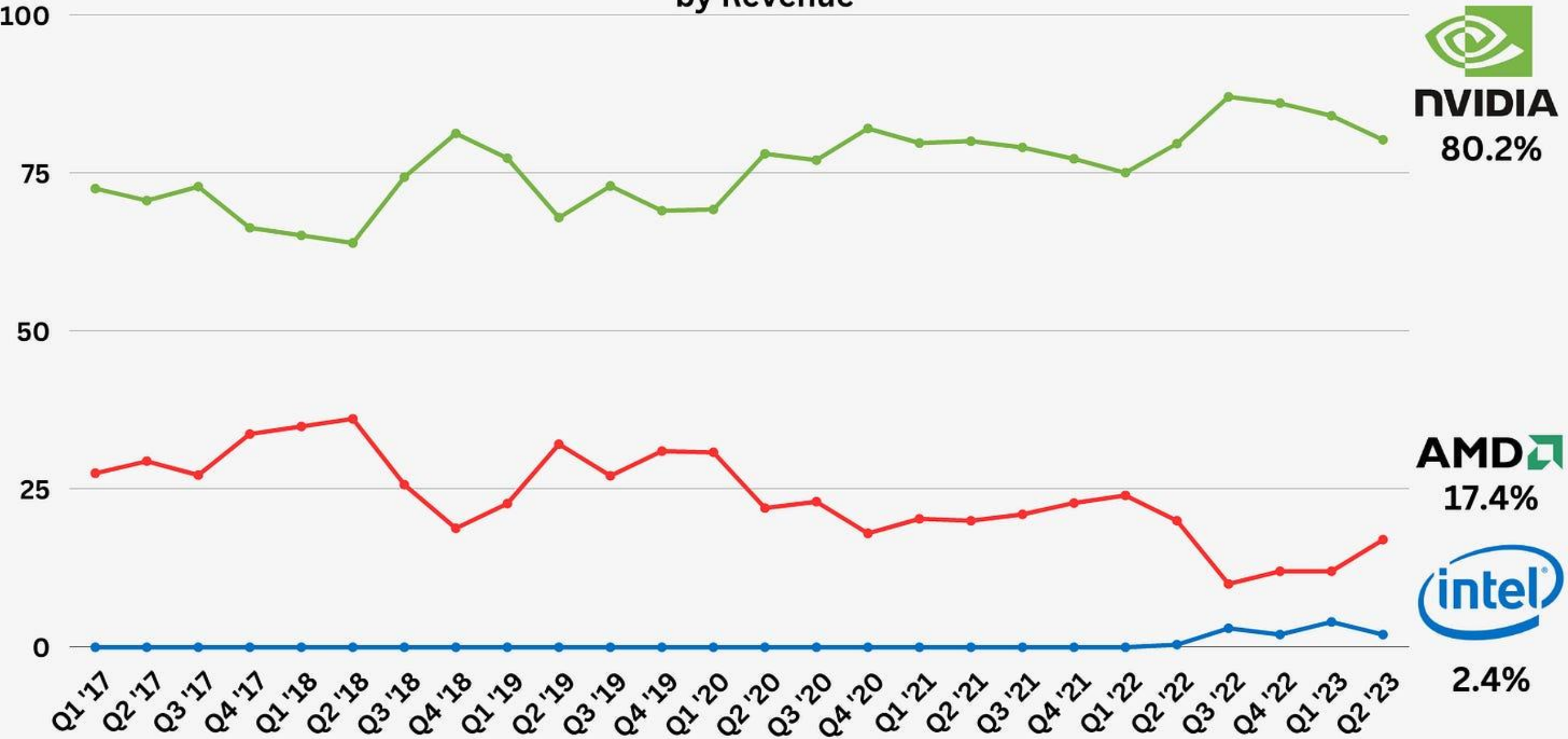
APP ECONOMY INSIGHTS

Grace Blackwell in Full Production



Desktop GPU Market Share

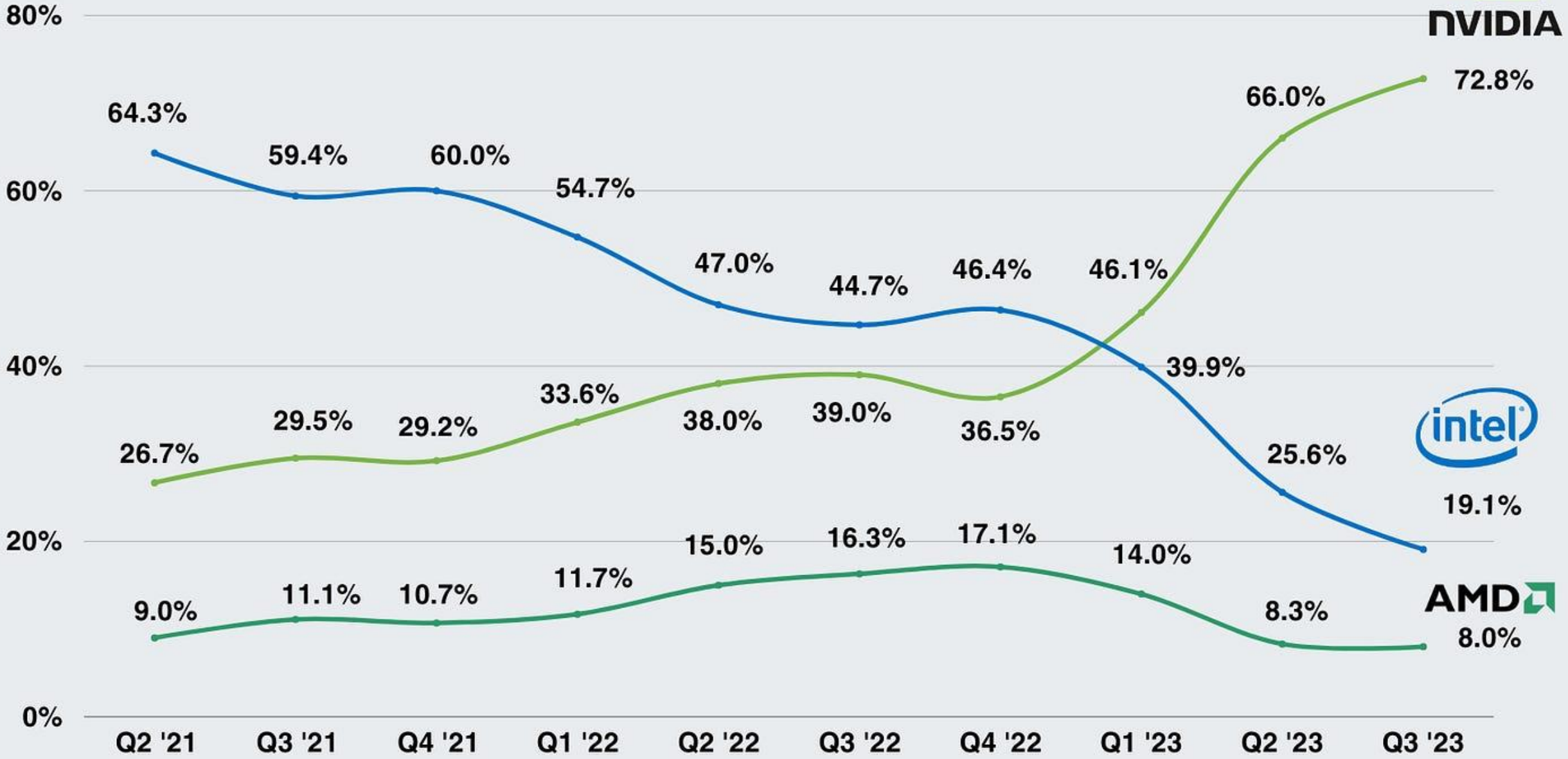
by Revenue



Source: Jon Peddle Research

Data Center Market Share

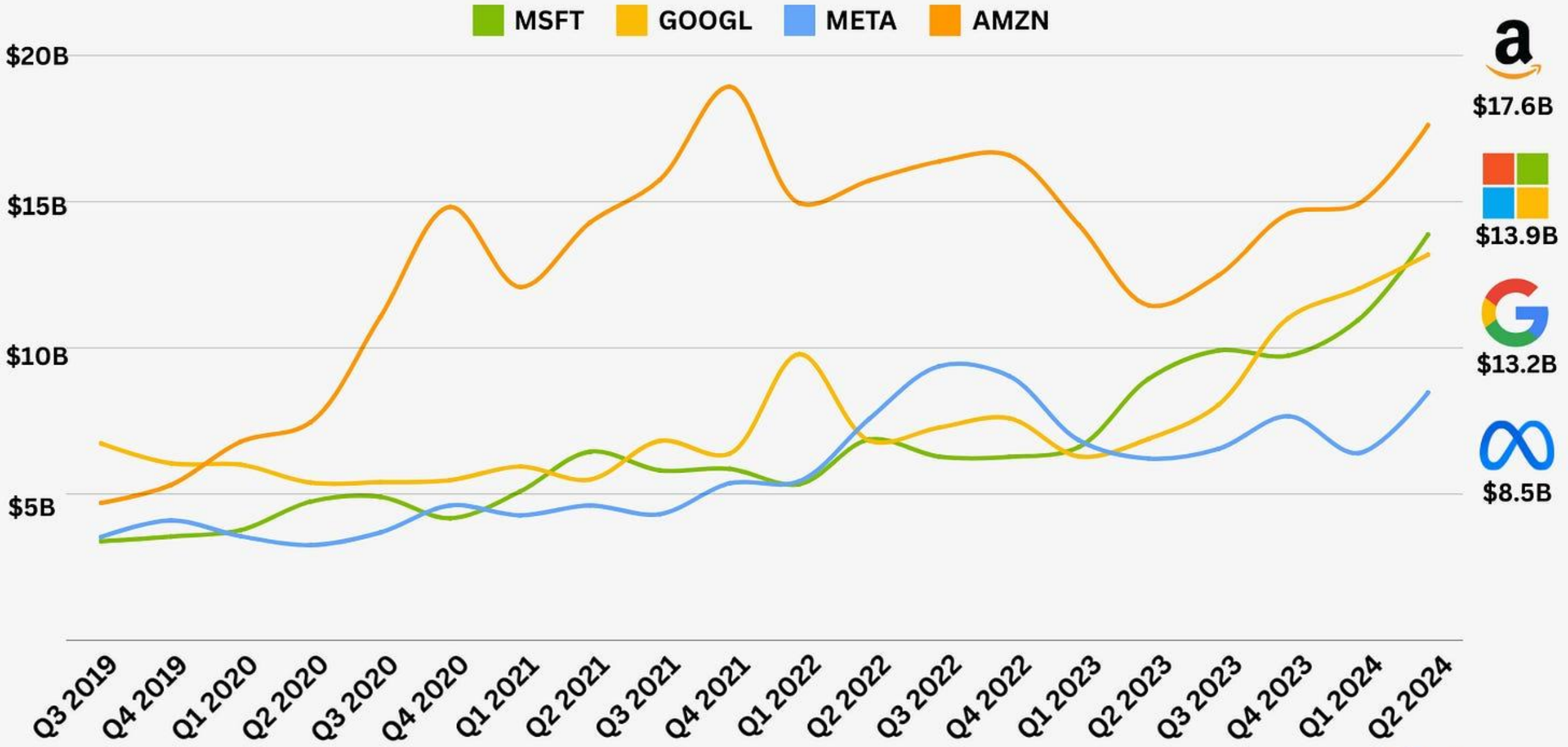
3 largest chip providers



@EricFlaningam

Source: 10Qs, *Nvidia includes networking

Hyperscaler Quarterly Capex



Generative Value @EricFlaningam

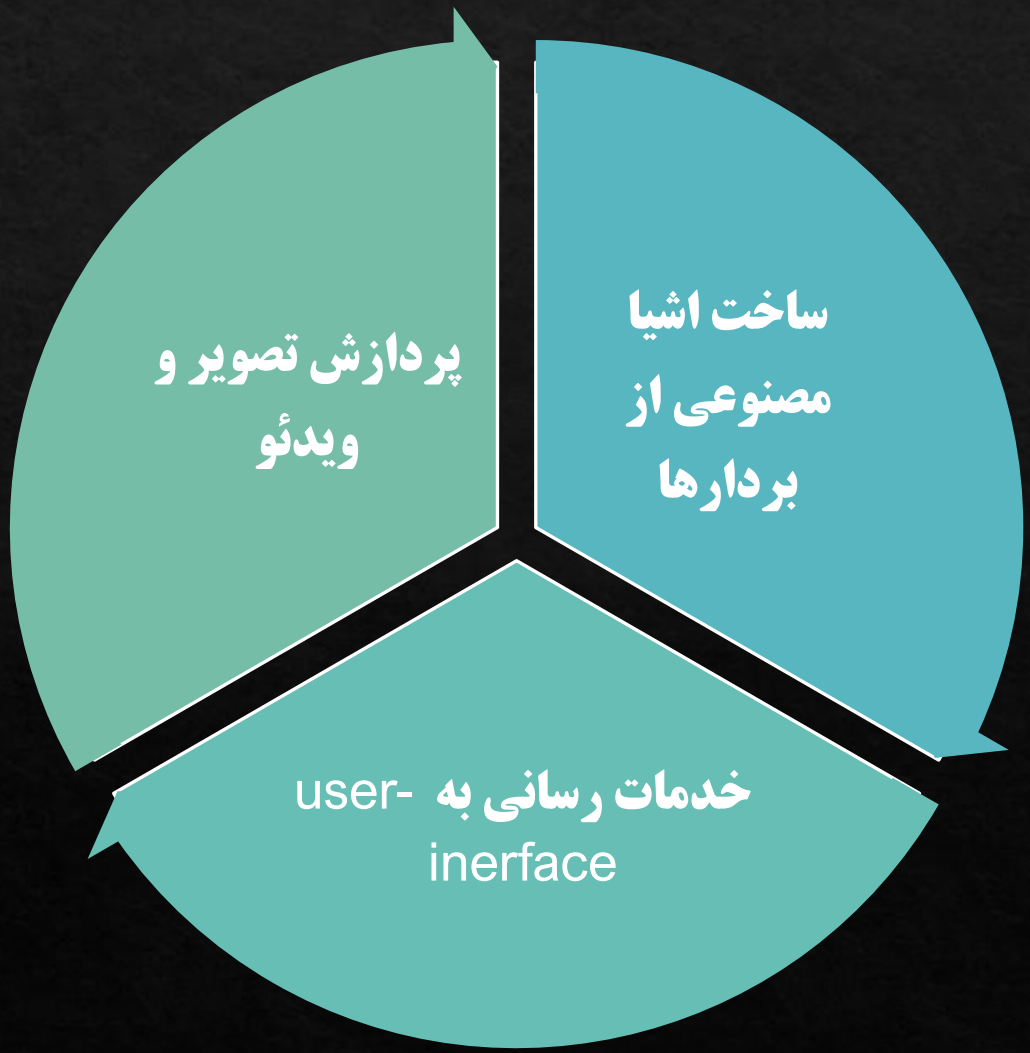
Felicis

تاریخچه پردازنده های گرافیکی

و

اتفاقی مهم در این حوزه!

کارهای GPU های اولیه



The Rise of Nvidia

The key moments leading to its rise from a market cap of $600M to $3.5T.

Nvidia's Quarterly Revenue (since IPO)

Nvidia's RIVA 128 is released, its first commercially successful chip. 1997

Nvidia's first GPU, the GeForce 256, is released. 1999

Nvidia releases Tesla, their first HPC-focused GPU architecture. 2007

AlexNext won the ImageNet Visual Recognition Challenge, using CUDA and Nvidia GPUs. 2012

Nvidia's first chip, the NV1, is released. 1995

Nvidia IPOs at a $600M valuation. 1999

Programmable shaders are released with GeForce 3. 2001

CUDA is released, introducing GPU computing. 2006

The Kepler architecture is released. 2012

The Fermi architecture is released. 2010

Maxwell architecture released. 2014

OpenAI founded. 2015

Pascal architecture released. 2016

Volta architecture released. 2017

Turing architecture released. 2018

Ampere architecture released. 2019

Nvidia acquires Mellanox, the leader in InfiniBand networking. 2020

ChatGPT released. 2022

Hopper architecture released, Grace CPU chip announced, and DGX server released. 2023

DGX Cloud released. 2023

Nvidia surpasses $35B in quarterly revenue, up 5x in just 6 quarters. 2024

Nvidia is Founded 1993

1997

2001 Passed $1B in annual revenue

2005

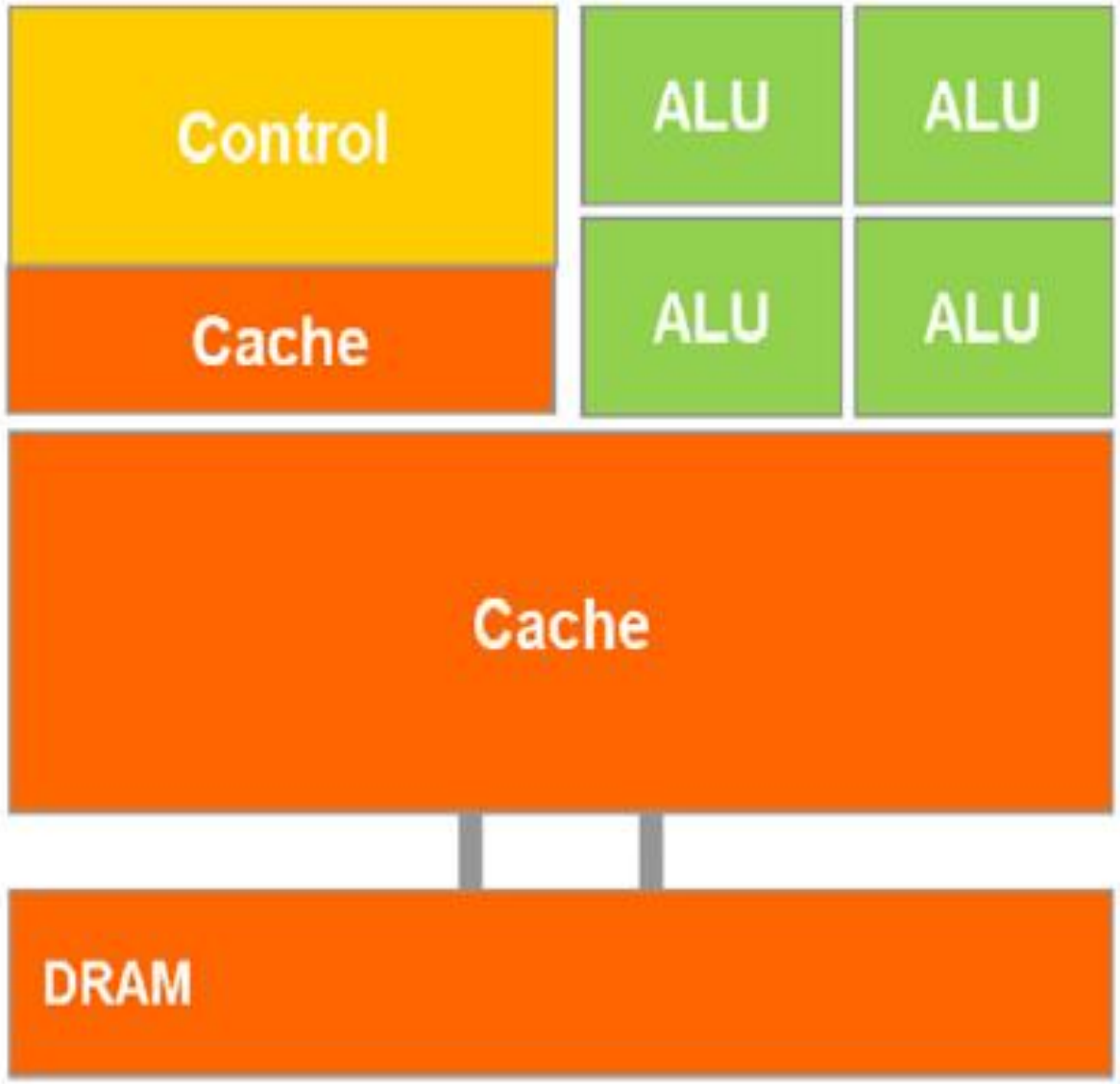
2009

2013

2017 Passed $10B in annual revenue

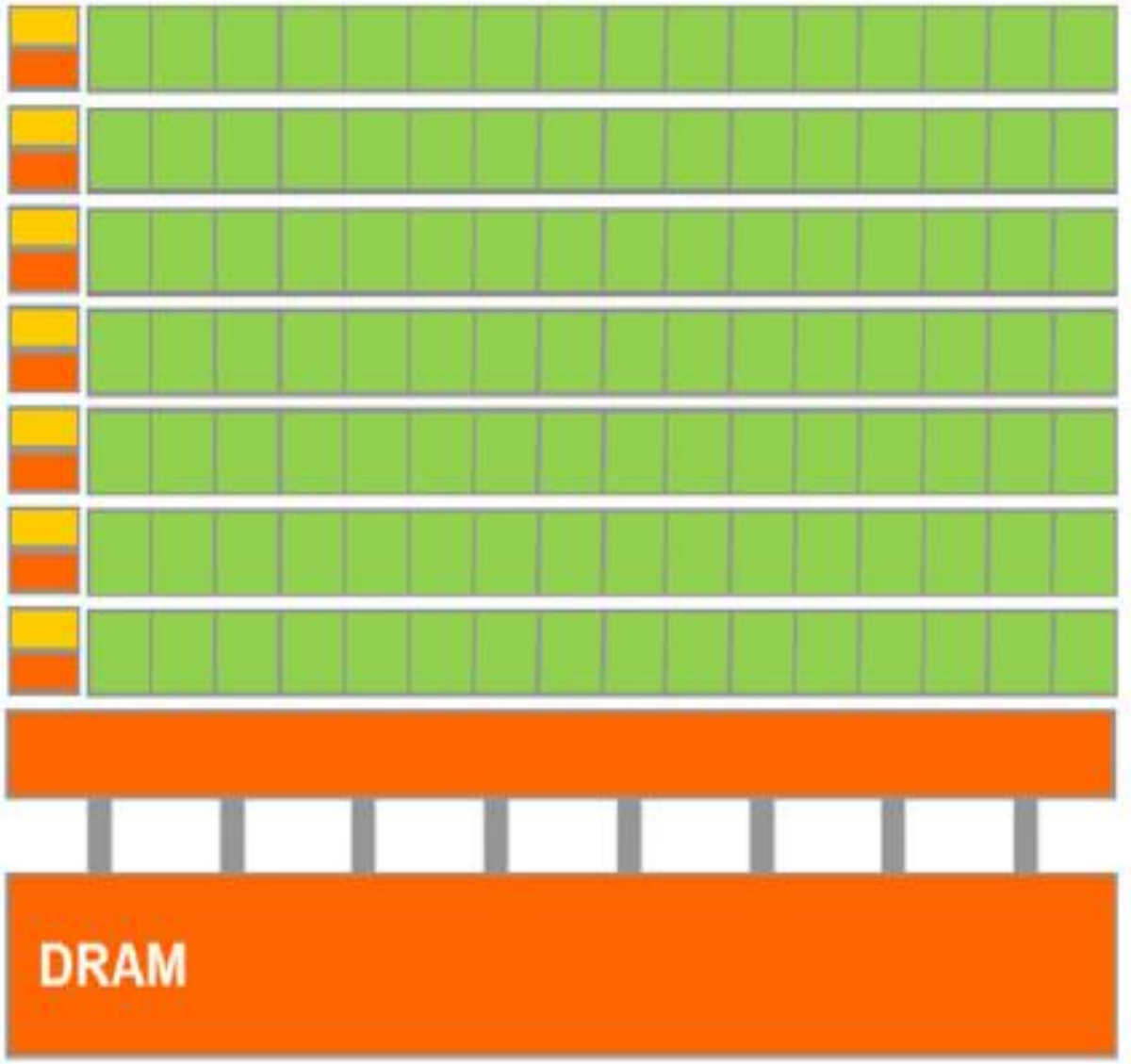
2021

2024



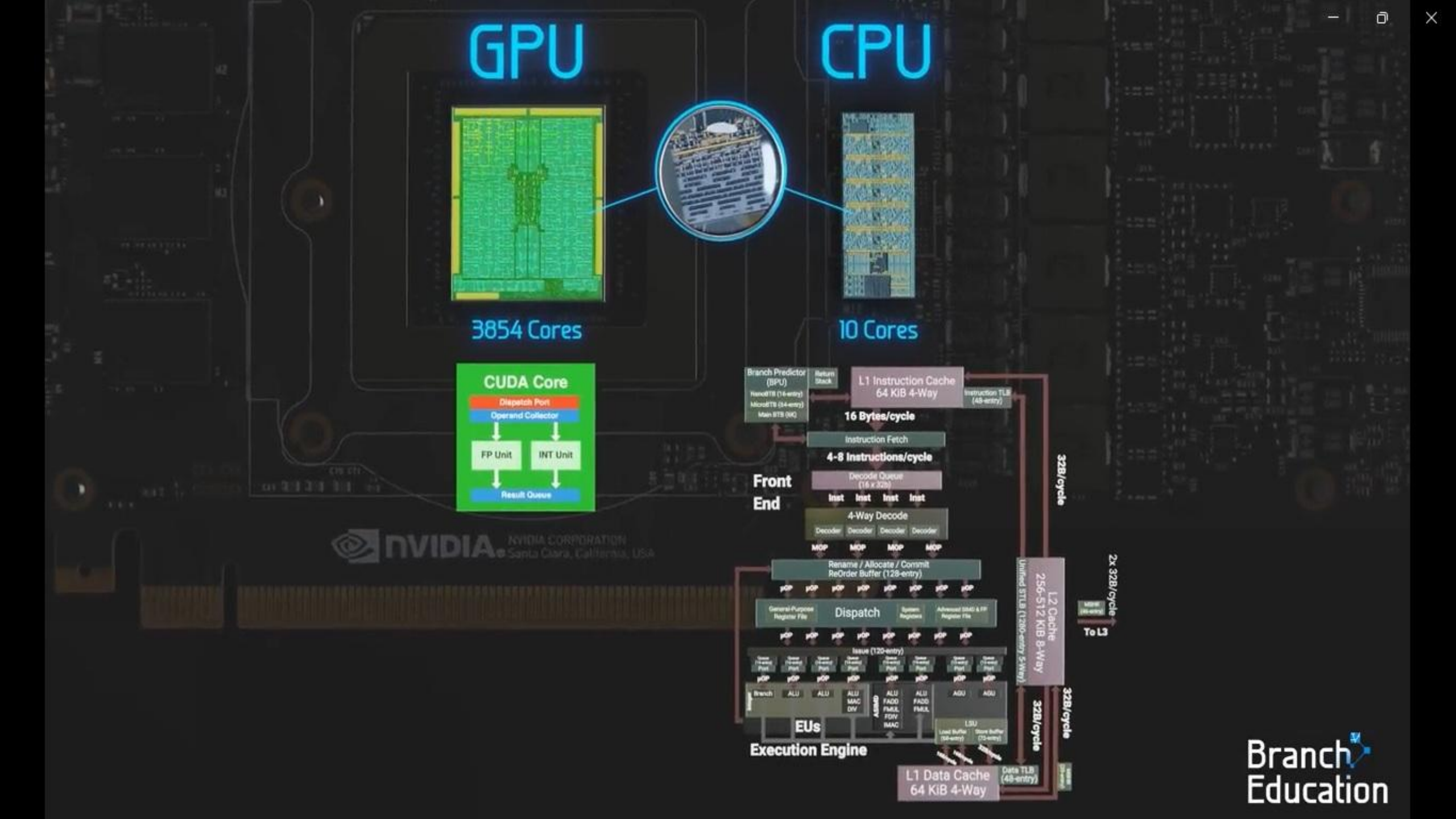
CPU

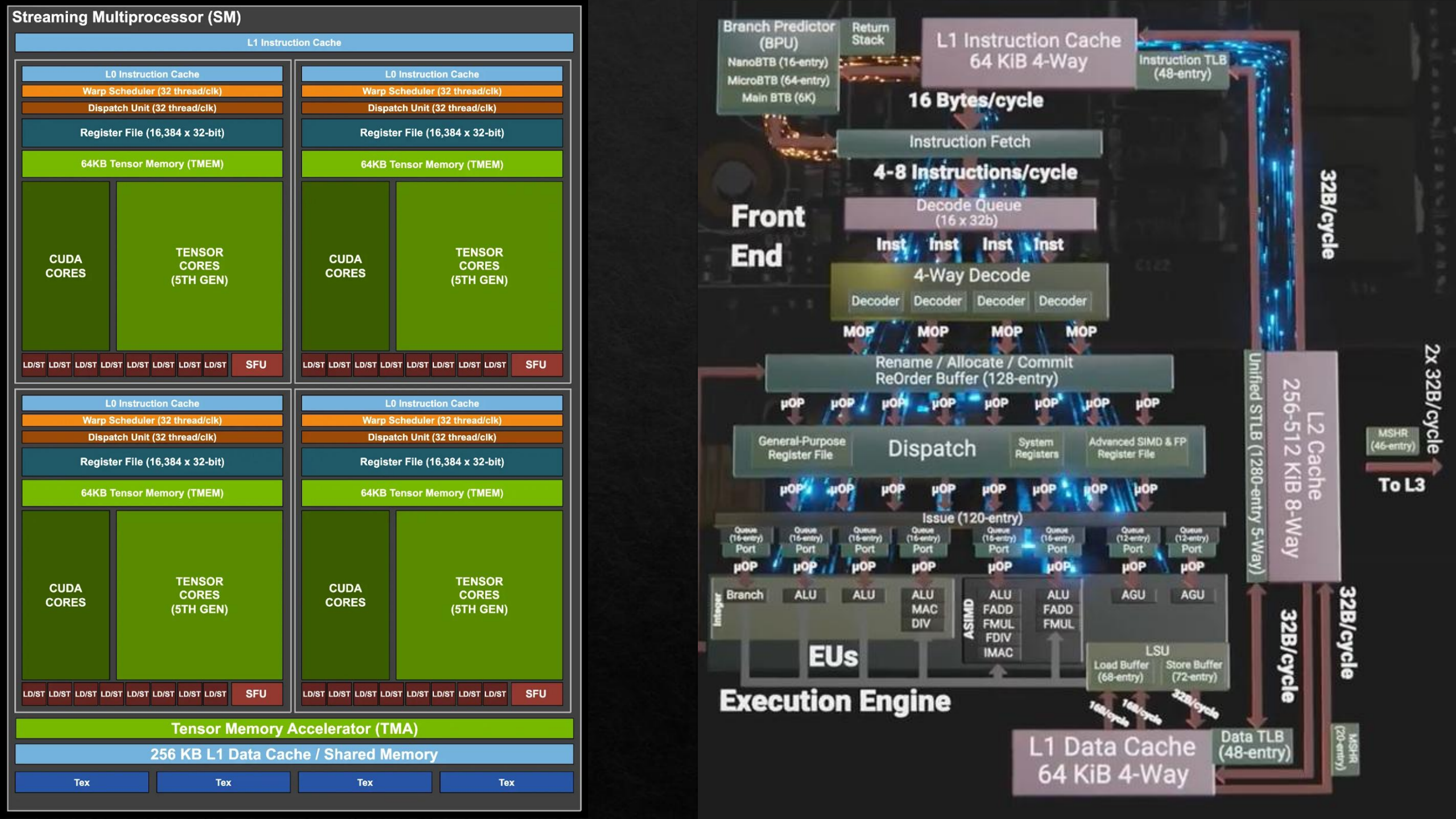
(latency-oriented design)

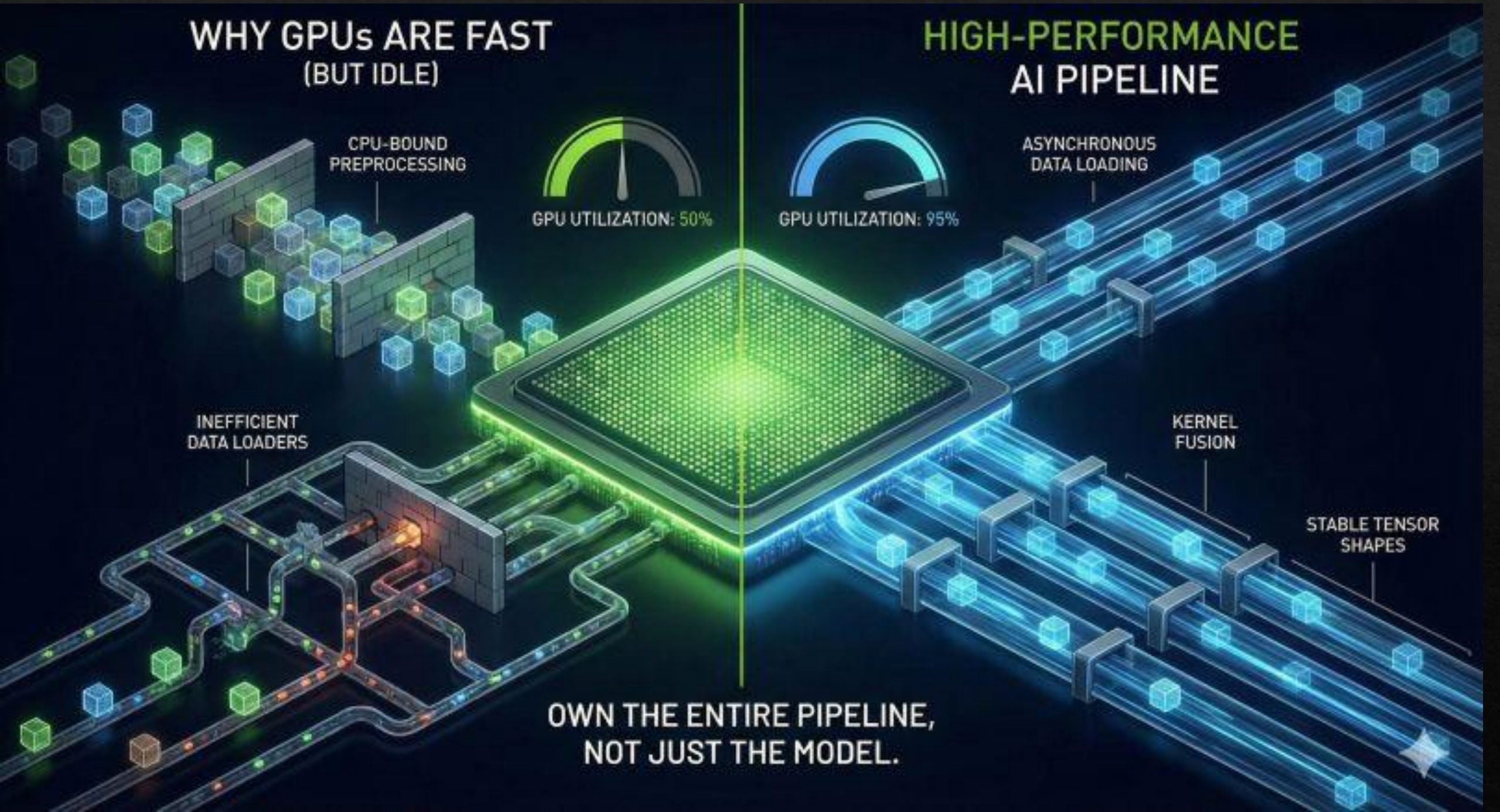


GPU

(throughput-oriented design)


GPU
CPU
3854 Cores
10 Cores
CUDA Core
Dispatch Port
Operand Collector
FP Unit
INT Unit
Result Queue
NVIDIA
NVIDIA CORPORATION
Santa Clara, California, USA
Branch Predictor (BPU)
Return Stack
L1 Instruction Cache 64 KiB 4-Way
Instruction TLB (48-entry)
16 Bytes/cycle
Instruction Fetch
4-8 Instructions/cycle
Front End
Decode Queue
Inst Inst Inst Inst
4-Way Decode
Decoder Decoder Decoder Decoder
MOP MOP MOP MOP
Rename / Allocate / Commit
ReOrder Buffer (128-entry)
pOP
Dispatch
Issue (120-entry)
Branch ALU ALU ALU MAC DIV
ALU FADD FMUL FDIV IMAC
ALU FADD FMUL
AGU AGU
L3U
EUs
Execution Engine
L1 Data Cache 64 KiB 4-Way
Data TLB (48-entry)
L2 Cache 256-512 KiB 8-Way
Unified STLB (1280-entry 5-Way)
32B/cycle
32B/cycle
32B/cycle
2x 32B/cycle
To L3
Branch Education

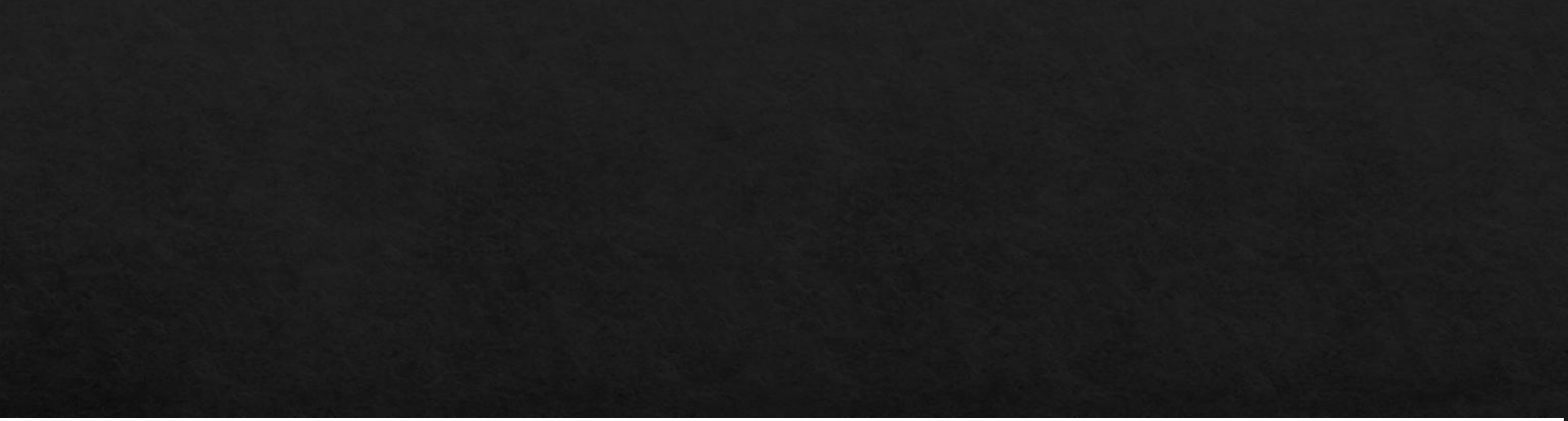

Streaming Multiprocessor (SM)
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
64KB Tensor Memory (TMEM)
CUDA CORES
TENSOR CORES (5TH GEN)
LD/ST
SFU
Tensor Memory Accelerator (TMA)
256 KB L1 Data Cache / Shared Memory
Tex
Branch Predictor (BPU)
NanoBTB (16-entry)
MicroBTB (64-entry)
Main BTB (6K)
Return Stack
L1 Instruction Cache 64 KiB 4-Way
Instruction TLB (48-entry)
16 Bytes/cycle
Instruction Fetch
4-8 Instructions/cycle
Front End
Decode Queue (16 x 32b)
Inst
4-Way Decode
Decoder
MOP
Rename / Allocate / Commit ReOrder Buffer (128-entry)
µOP
General-Purpose Register File
Dispatch
System Registers
Advanced SIMD & FP Register File
Issue (120-entry)
Queue (16-entry)
Queue (12-entry)
Port
Integer
Branch
ALU
ALU MAC DIV
ASIMD
ALU FADD FMUL FDIV IMAC
ALU FADD FMUL
AGU
LSU
Load Buffer (68-entry)
Store Buffer (72-entry)
EUs
Execution Engine
32B/cycle
Unified STLB (1280-entry 5-Way)
L2 Cache 256-512 KiB 8-Way
2x 32B/cycle
MSHR (46-entry)
To L3
16B/cycle
32B/cycle
L1 Data Cache 64 KiB 4-Way
Data TLB (48-entry)
MSHR (20-entry)


WHY GPUs ARE FAST
(BUT IDLE)
CPU-BOUND
PREPROCESSING
GPU UTILIZATION: 50%
INEFFICIENT
DATA LOADERS
HIGH-PERFORMANCE
AI PIPELINE
ASYNCHRONOUS
DATA LOADING
GPU UTILIZATION: 95%
KERNEL
FUSION
STABLE TENSOR
SHAPES
OWN THE ENTIRE PIPELINE,
NOT JUST THE MODEL.

NVIDIA's Evolution: From Graphics to AI Leadership

نسل های برندهای پردازنده های گرافیکی (GPU)

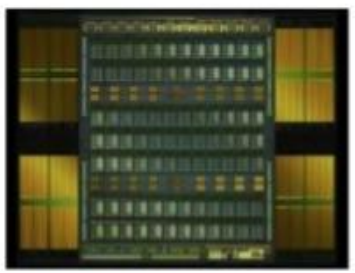

VIDEOCARDZ.COM
GX200
GB200
GH200
GH100
GA100
GV100
GP100



	Nvidia B200	Nvidia H100 SXM5	AMD Radeon Instinct MI300X
Power Target	1000W	700W	750W
Maximum Clock Speed	1.965 GHz	1.98 GHz	2.1 GHz
Core Count	148 SMs	132 SMs	304 Compute Units (CUs)
Last Level Cache	126 MB	50 MB	126 MB
Die Arrangement	2 Dies	Monolithic	8x Compute Dies 4x Base Dies
VRAM	288 GB HBM3E 8 TB/s	80 GB HBM3 3.3 TB/s	192 GB HBM3 5.3 TB/s
Process	TSMC 4NP	TSMC 4N	TSMC 5nm (Compute Dies) TSMC 6nm (Base Dies)

	Supported CUDA Core Precisions									Supported Tensor Core Precisions								
	FP8	**FP16**	**FP32**	**FP64**	**INT1**	**INT4**	**INT8**	**TF32**	**BF16**	**FP8**	**FP16**	**FP32**	**FP64**	**INT1**	**INT4**	**INT8**	**TF32**	**BF16**
NVIDIA Tesla P4	No	No	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No
NVIDIA P100	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
NVIDIA Volta	No	Yes	Yes	Yes	No	No	Yes	No	No	No	Yes	No	No	No	No	No	No	No
NVIDIA Turing	No	Yes	Yes	Yes	No	No	Yes	No	No	No	Yes	No	No	Yes	Yes	Yes	No	No
NVIDIA A100	No	Yes	Yes	Yes	No	No	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
NVIDIA H100	No	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes

نسل های معماری
پردازنده های گرافیکی (GPU)

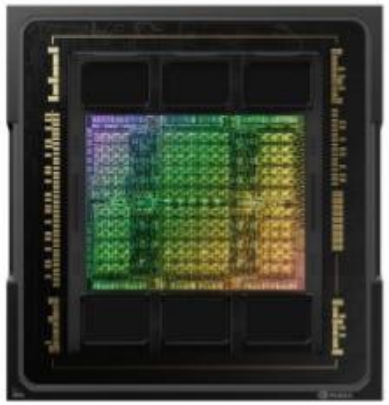


Volta

>21 billion transistors
815mm^2
TSMC 12nm FFN

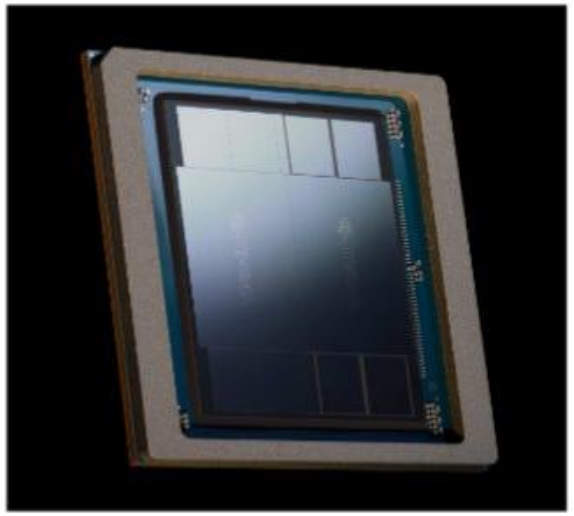
Ampere

>54 billion transistors
826 mm^2
TSMC N7



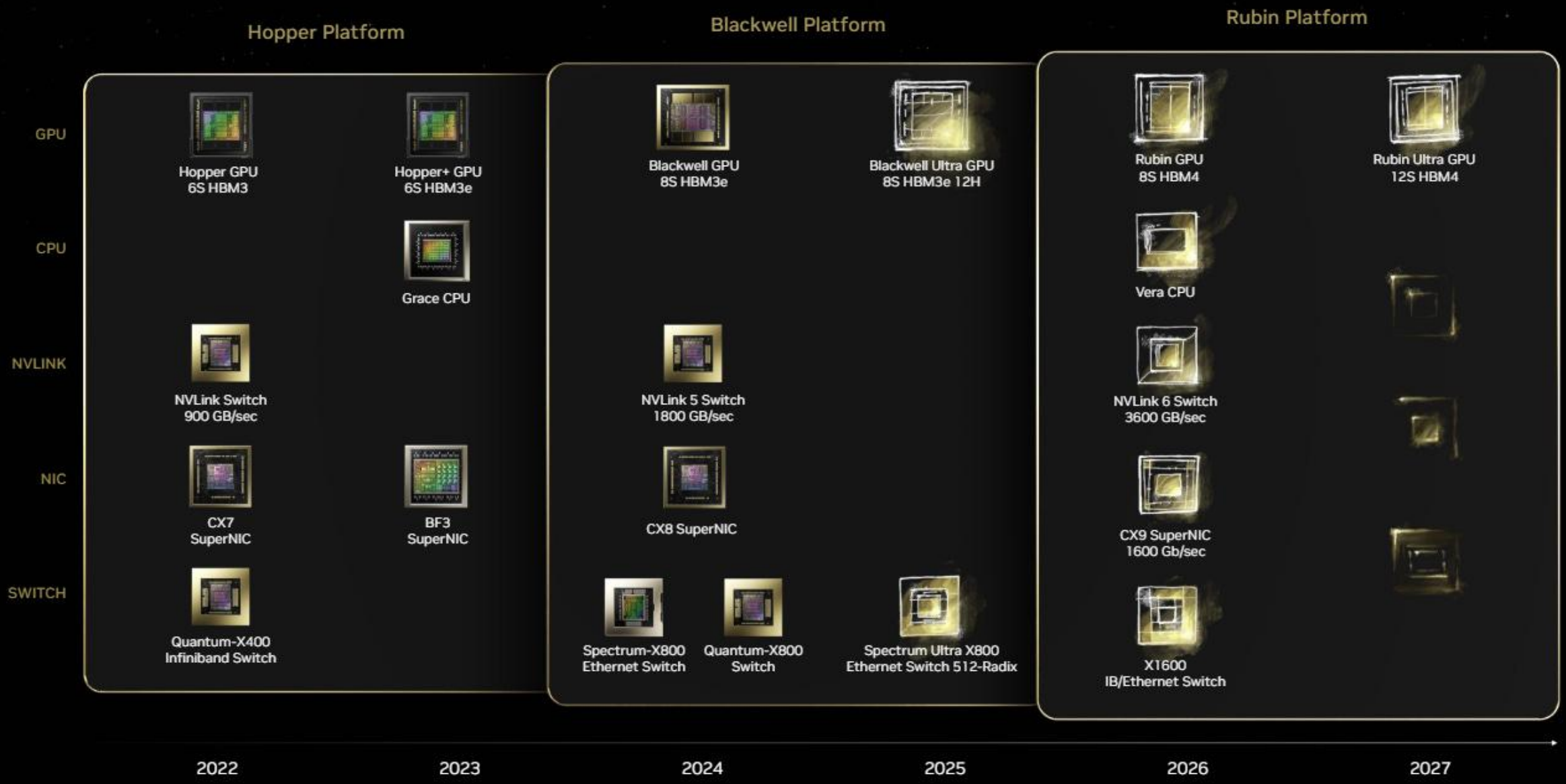
Hopper

>80 billion transistors
814 mm^2
TSMC 4N



Blackwell

>208 billion transistors
>1600 mm^2
TSMC 4NP



資料中心規模 • 一年節奏 • 技術限制 • 一個架構

DATACENTER SCALE • ONE-YEAR RHYTHM • TECHNOLOGY LIMITS • ONE ARCHITECTURE

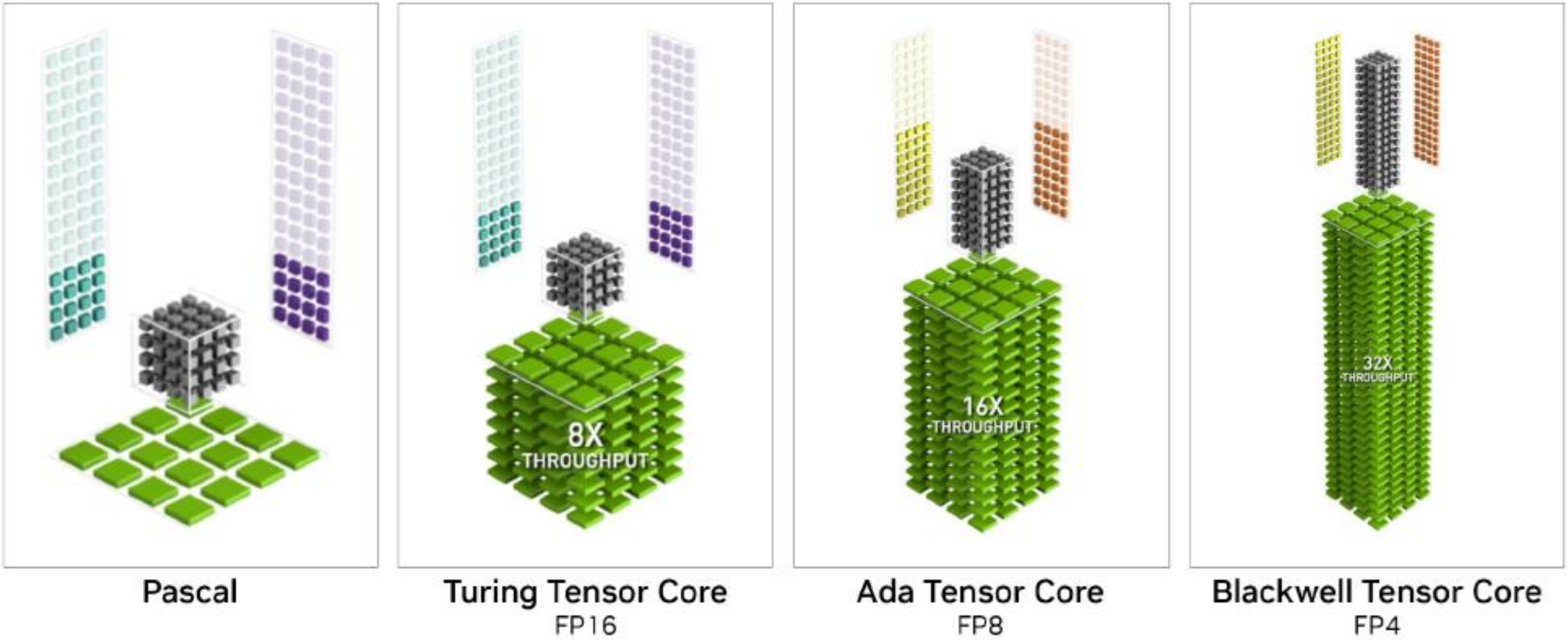
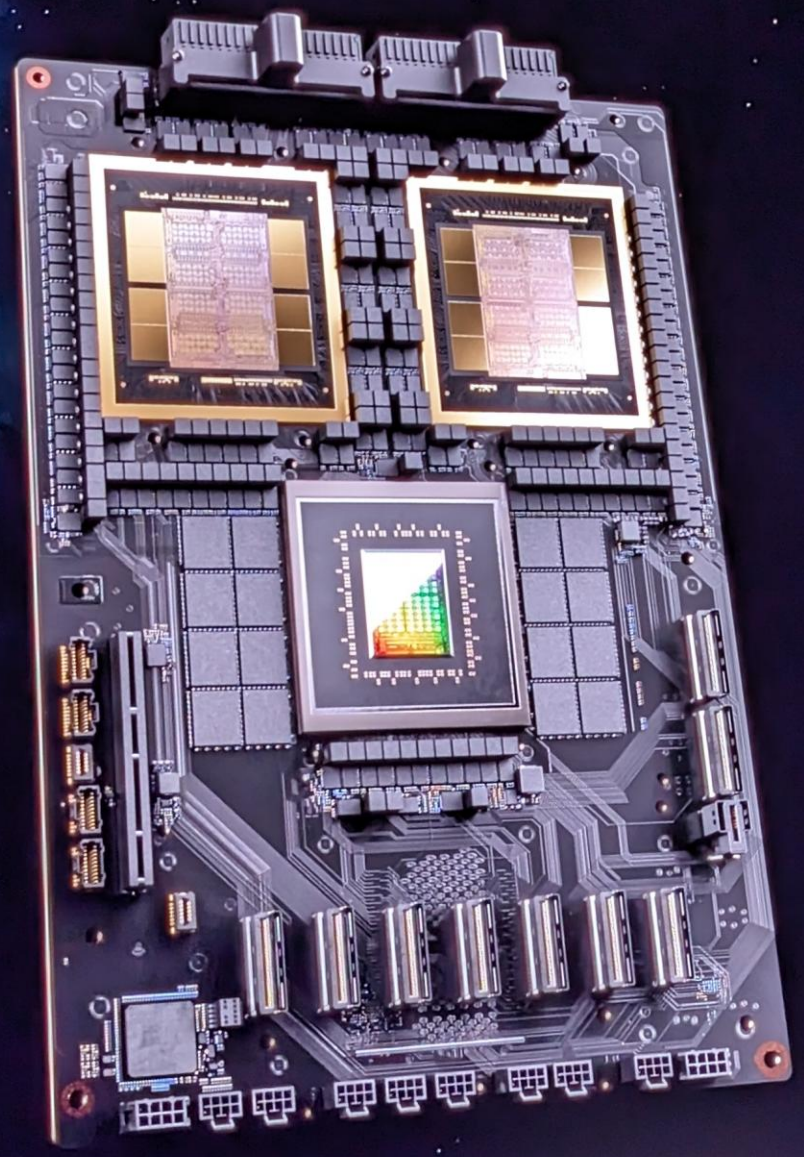


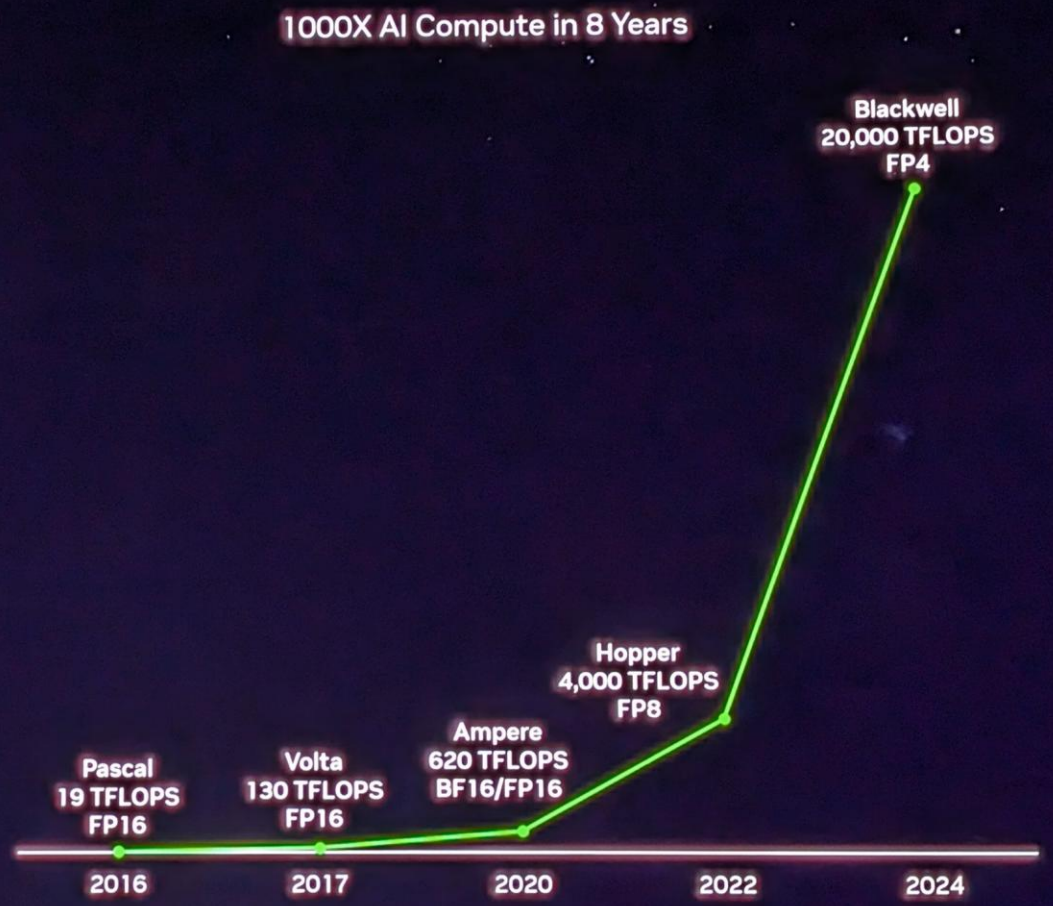
Figure Blackwell 5th Generation Tensor Cores with FP4, double throughput of FP8

	Hopper	Ampere	Turing	Volta
Supported Tensor Core precisions	FP64, TF32, bfloat16, FP16, FP8, INT8	FP64, TF32, bfloat16, FP16, INT8, INT4, INT1	FP16, INT8, INT4, INT1	FP16
Supported CUDA® Core precisions	FP64, FP32, FP16, bfloat16, INT8	FP64, FP32, FP16, bfloat16, INT8	FP64, FP32, FP16, INT8	FP64, FP32, FP16, INT8

Feature	NVIDIA A100 (Ampere)	NVIDIA H100 (Hopper)	NVIDIA Blackwell
Tensor Core	312 TFLOPS (FP16) 3rd Gen	989.4 TFLOPS (FP16) 4th Gen	Higher throughput 5th Gen
Key Precision	BF16, TF32	BF16, TF32, **FP8**	BF16, FP8, **FP4**
Transformer Acceleration	None	Transformer Engine (9x training)	Enhanced Transformer Engine
Max SM Count	108 SMs	132 SMs (SXMS)	Increased SMs
CUDA Cores (FP32)	6,912	16,896	Higher core count
Memory	HBM2 (40 GB)	HBM3 (80 GB, 3 TB/s)	HBM4 (Faster, improved cap)
Asynchronous Execution	Basic support	Tensor Memory Accelerator (TMA)	Enhanced async data transfer
MIG	1st Gen	2nd Gen, Secure MIG	Improved isolation & perf

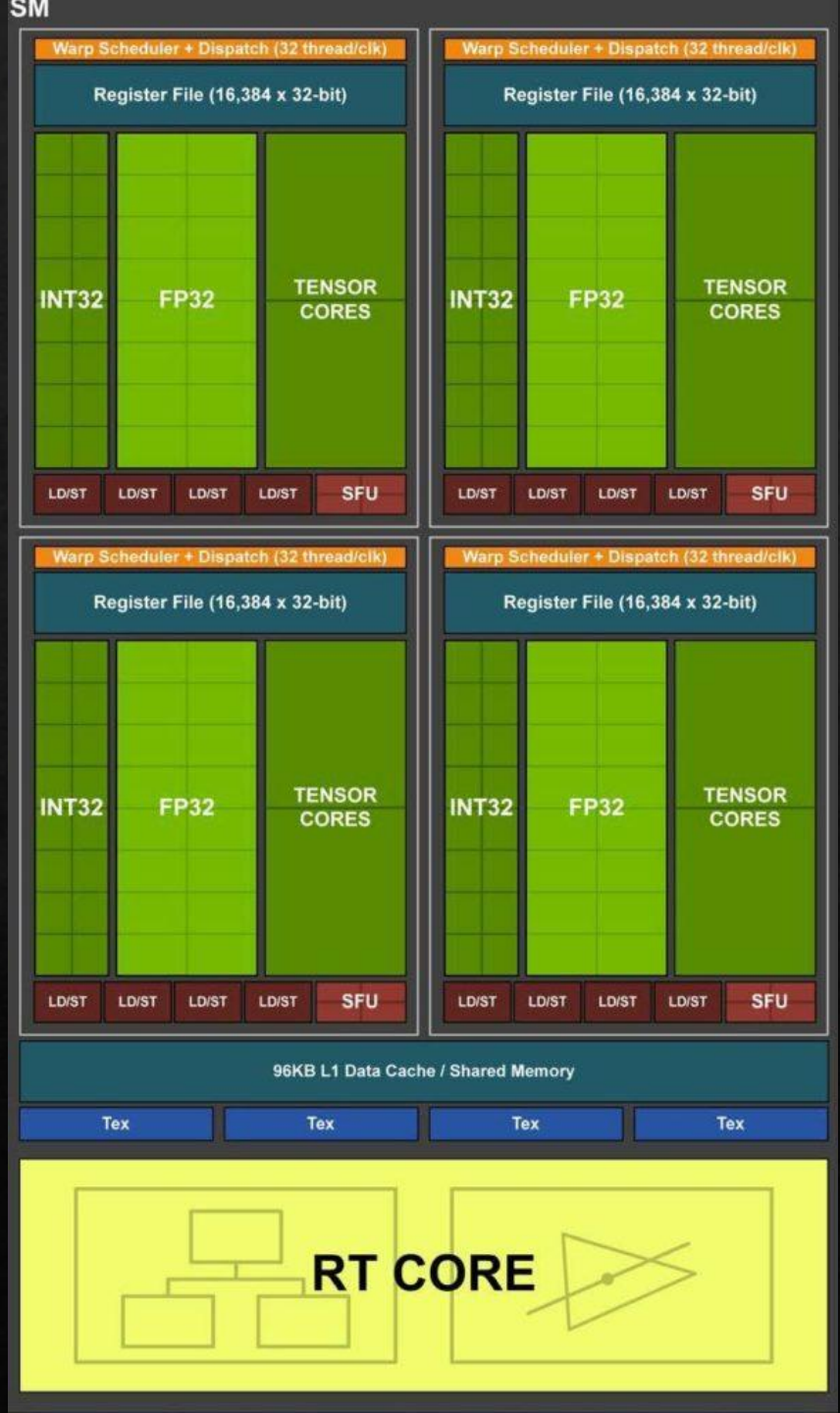
		RTX 40 Series	RTX 30 Series	RTX 20 Series
NVIDIA Architecture	Architecture Name	Ada Lovelace	Ampere	Turing
	Streaming Multiprocessors	2x FP32	2x FP32	1x FP32
	Ray Tracing Cores	Gen 3	Gen 2	Gen 1
	Tensor Cores (AI)	Gen 4	Gen 3	Gen 2
	NVIDIA DLSS	3	2	2




1000X AI Compute in 8 Years
Blackwell
20,000 TFLOPS
FP4
Hopper
4,000 TFLOPS
FP8
Ampere
620 TFLOPS
BF16/FP16
Volta
130 TFLOPS
FP16
Pascal
19 TFLOPS
FP16
2016
2017
2020
2022
2024

نسل های هسته های پردازنده های گرافیکی (GPU)


Streaming Multiprocessor (SM)
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
64KB Tensor Memory (TMEM)
CUDA CORES
TENSOR CORES (5TH GEN)
LD/ST
SFU
Tensor Memory Accelerator (TMA)
256 KB L1 Data Cache / Shared Memory
Tex


SM
Warp Scheduler + Dispatch (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
TENSOR CORES
LD/ST
SFU
96KB L1 Data Cache / Shared Memory
Tex
RT CORE

NVIDIA RTX FULL-STACK INVENTIONS

1ST GEN RTX

DXR

DLSS 1

VkRay

2ND GEN RTX

RTXDI

Caustics

NVIDIA Real-Time Denoisers

Reflex

RTXGI

DLSS 2

3RD GEN RTX

2D SGM Optical Flow

Displaced Micro-Meshes

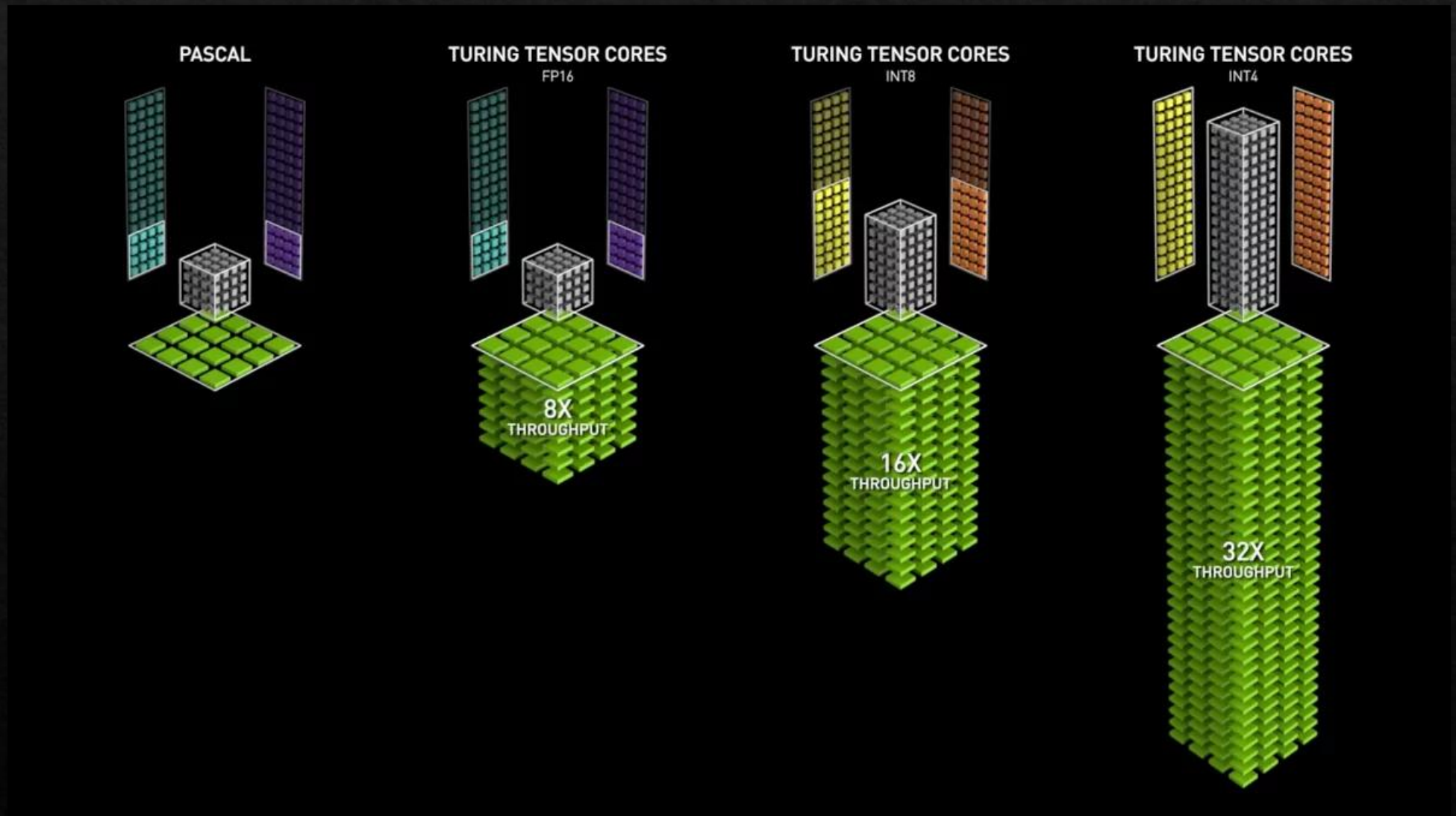
Shader Execution Reordering

Real-Time Path Tracing

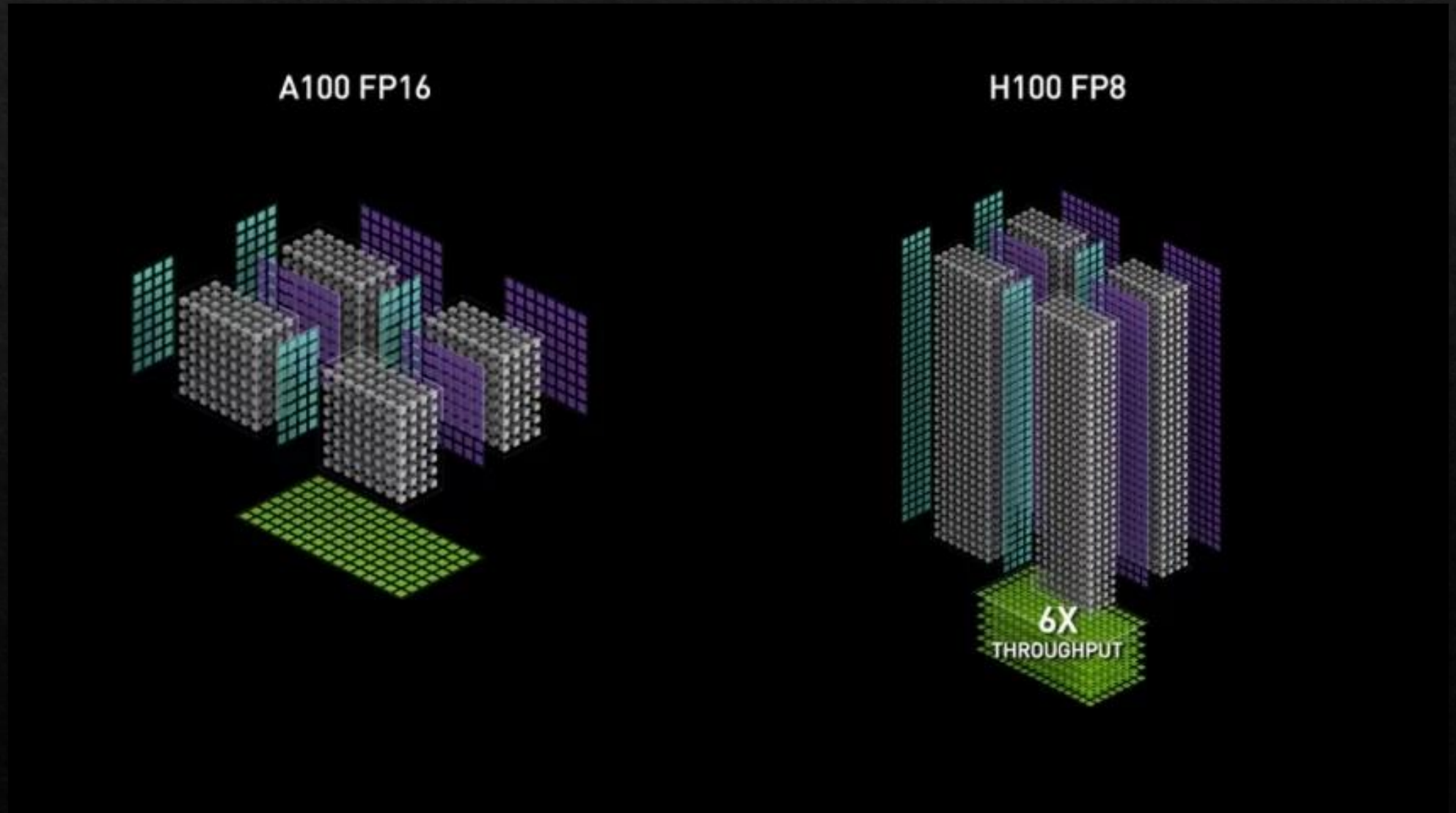
Opacity Micro-Maps

DLSS 3

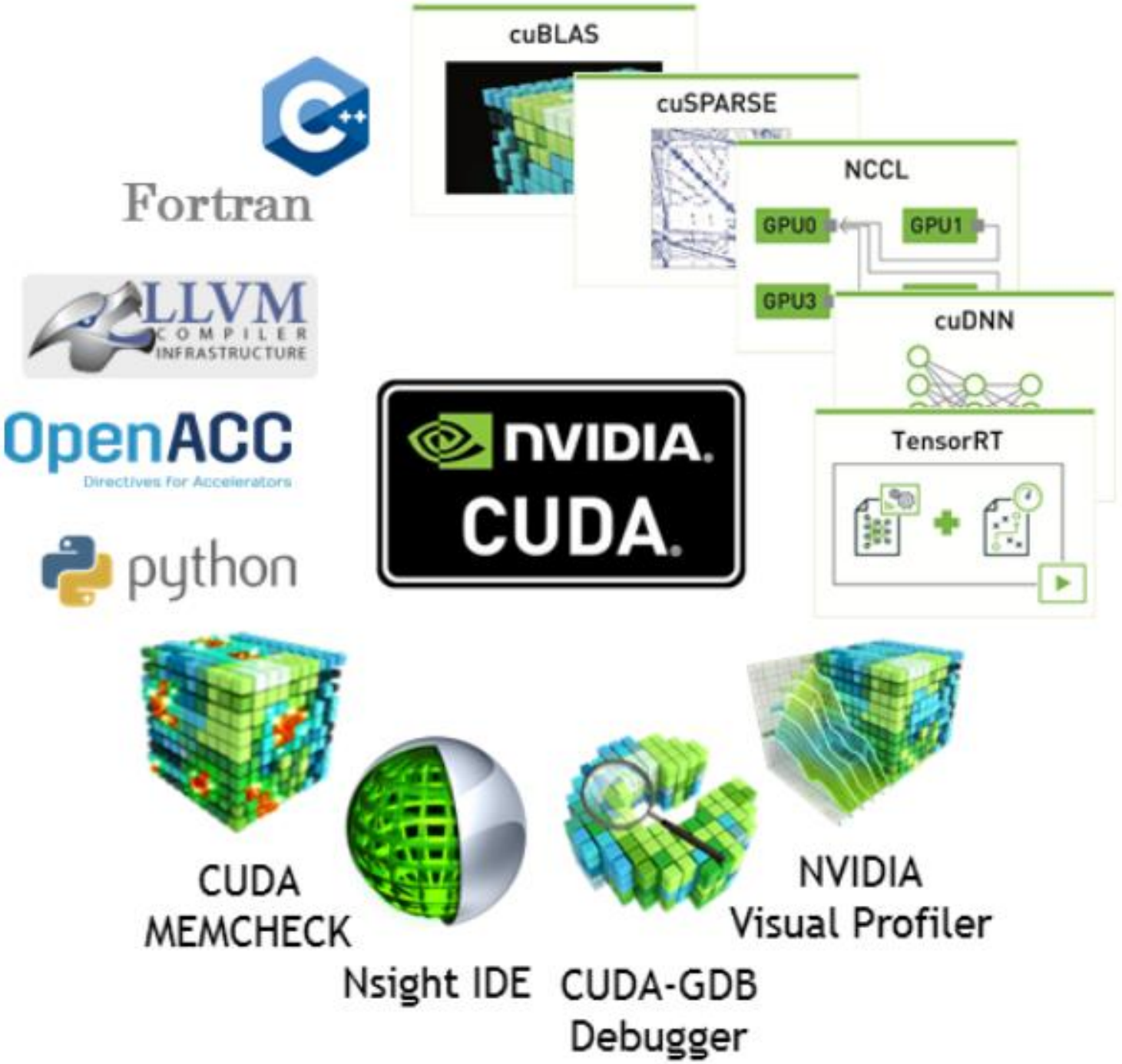
The Evolution of Tensor Cores in NVIDIA GPUs

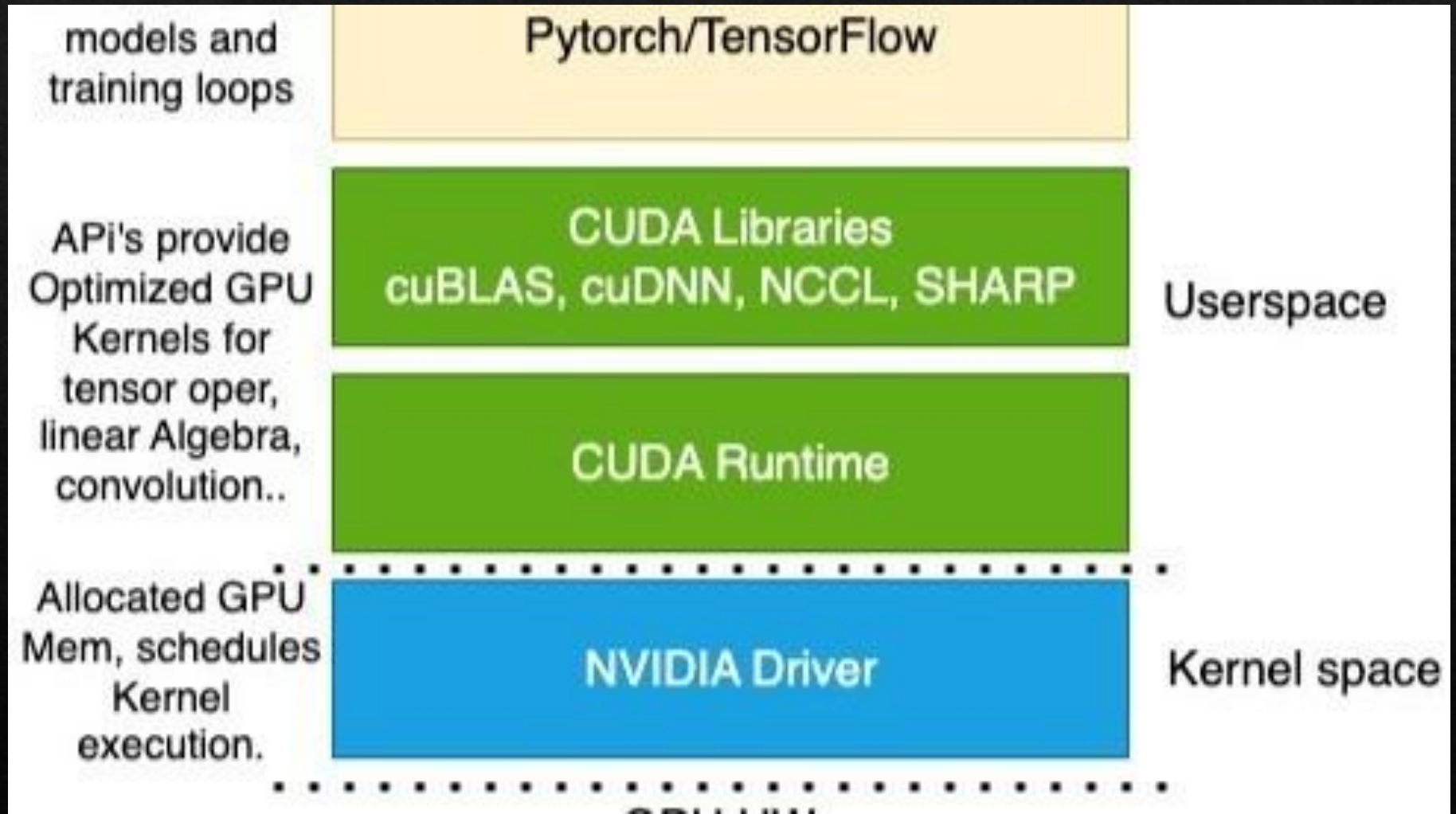


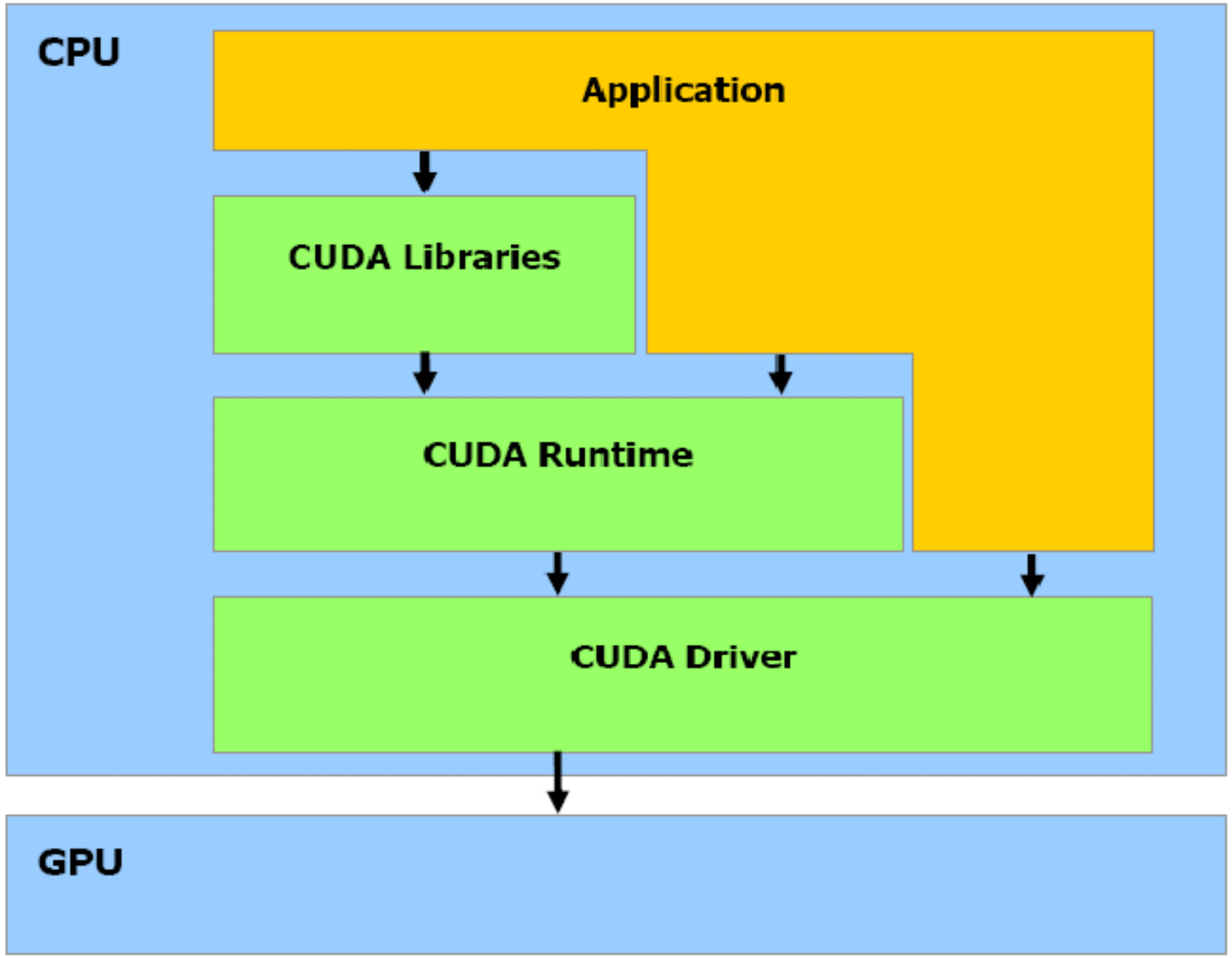
Fourth Generation: Hopper Architecture

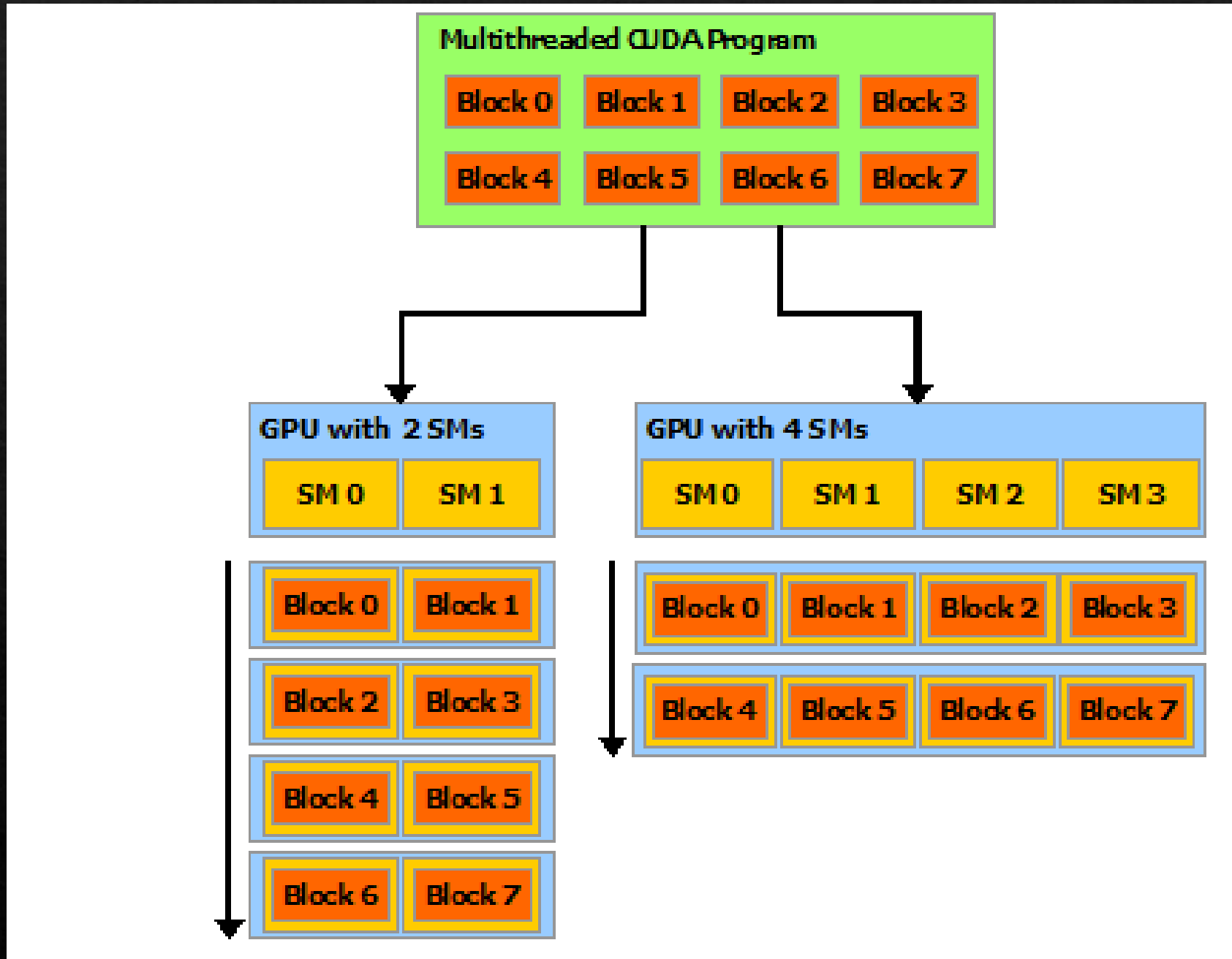


CUDA


cuBLAS
cuSPARSE
NCCL
GPU0
GPU1
GPU3
cuDNN
TensorRT
Fortran
LLVM
COMPILER
INFRASTRUCTURE
OpenACC
Directives for Accelerators
python
NVIDIA
CUDA
CUDA
MEMCHECK
Nsight IDE
CUDA-GDB
Debugger
NVIDIA
Visual Profiler

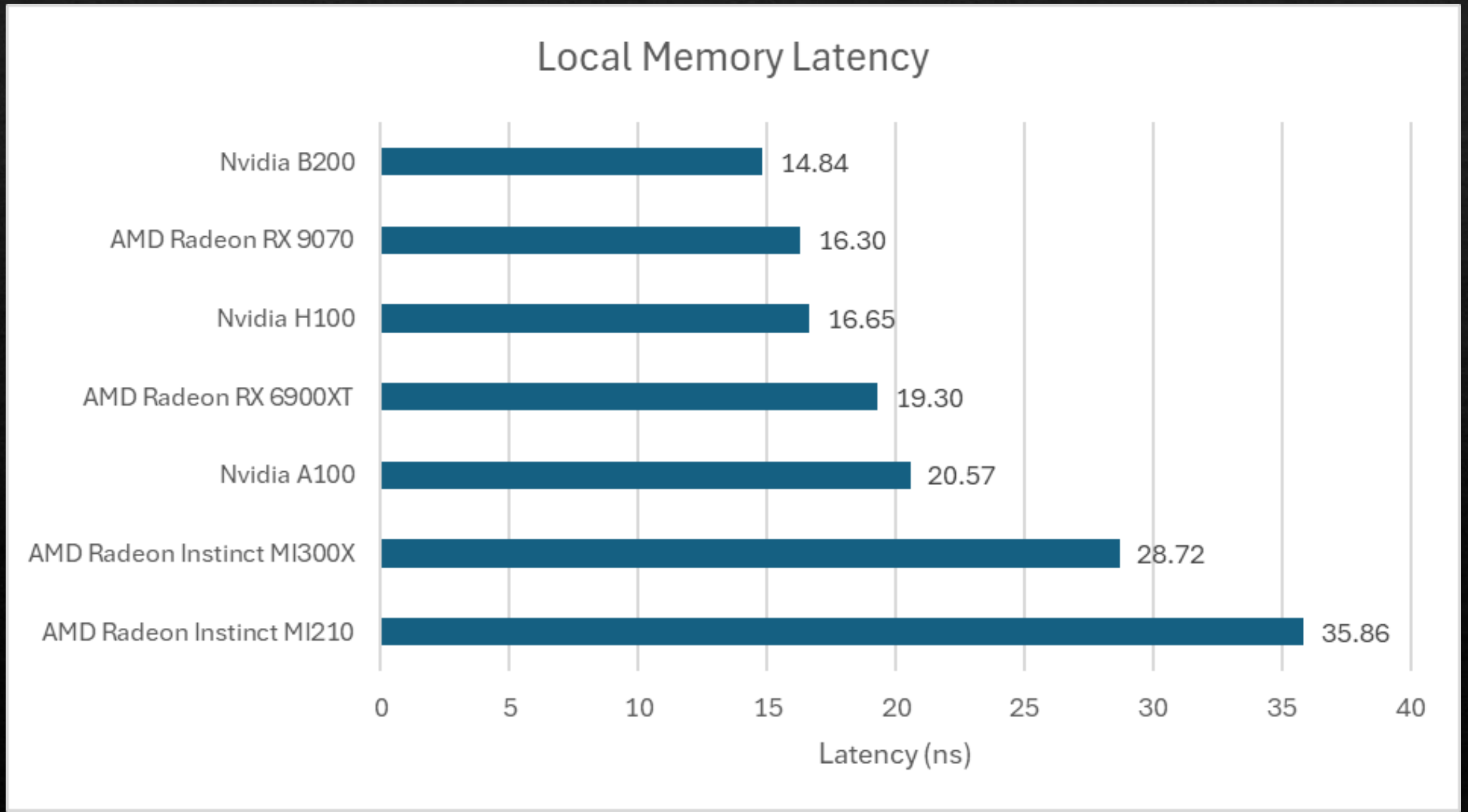

models and training loops
Pytorch/TensorFlow
APi's provide Optimized GPU Kernels for tensor oper, linear Algebra, convolution..
CUDA Libraries
cuBLAS, cuDNN, NCCL, SHARP
Userspace
CUDA Runtime
Allocated GPU Mem, schedules Kernel execution.
NVIDIA Driver
Kernel space

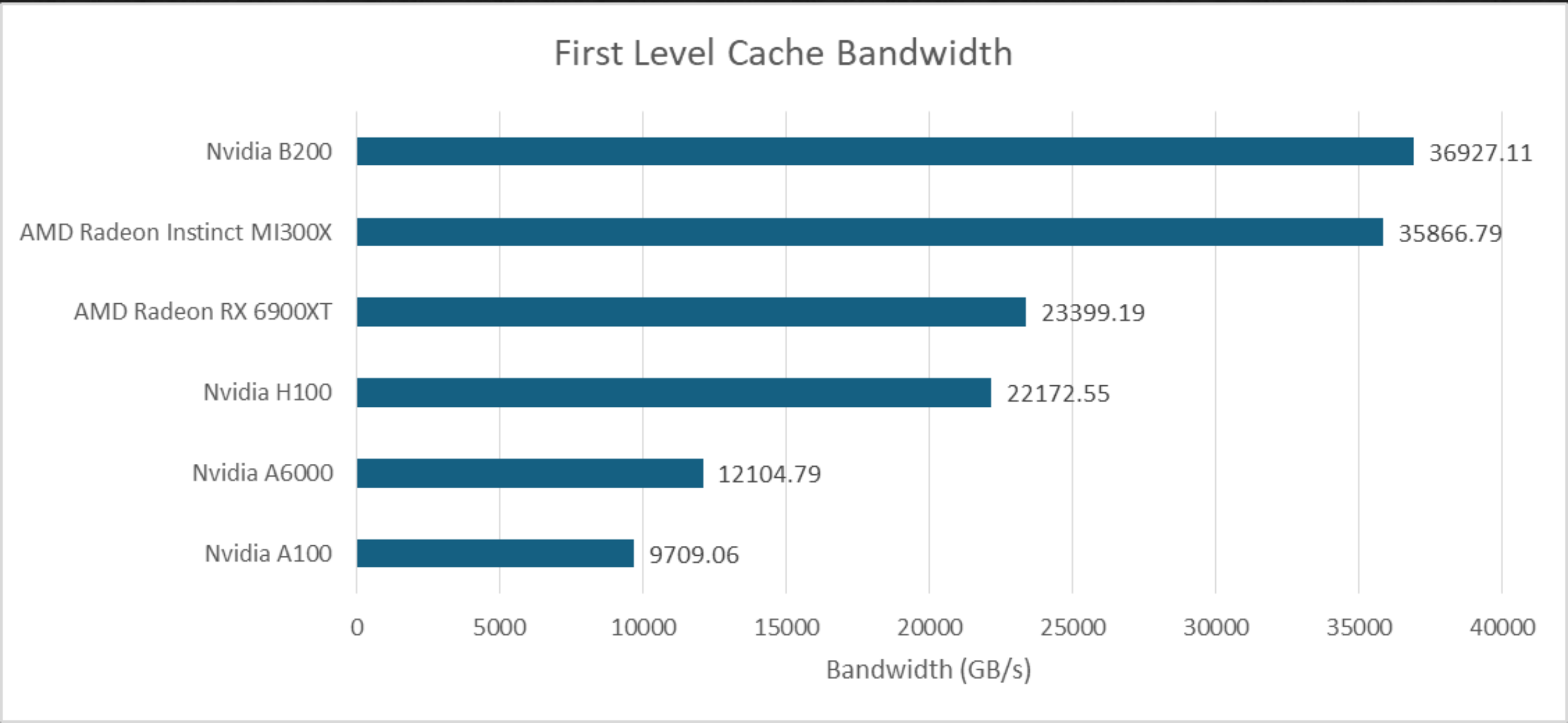

CPU
Application
CUDA Libraries
CUDA Runtime
CUDA Driver
GPU

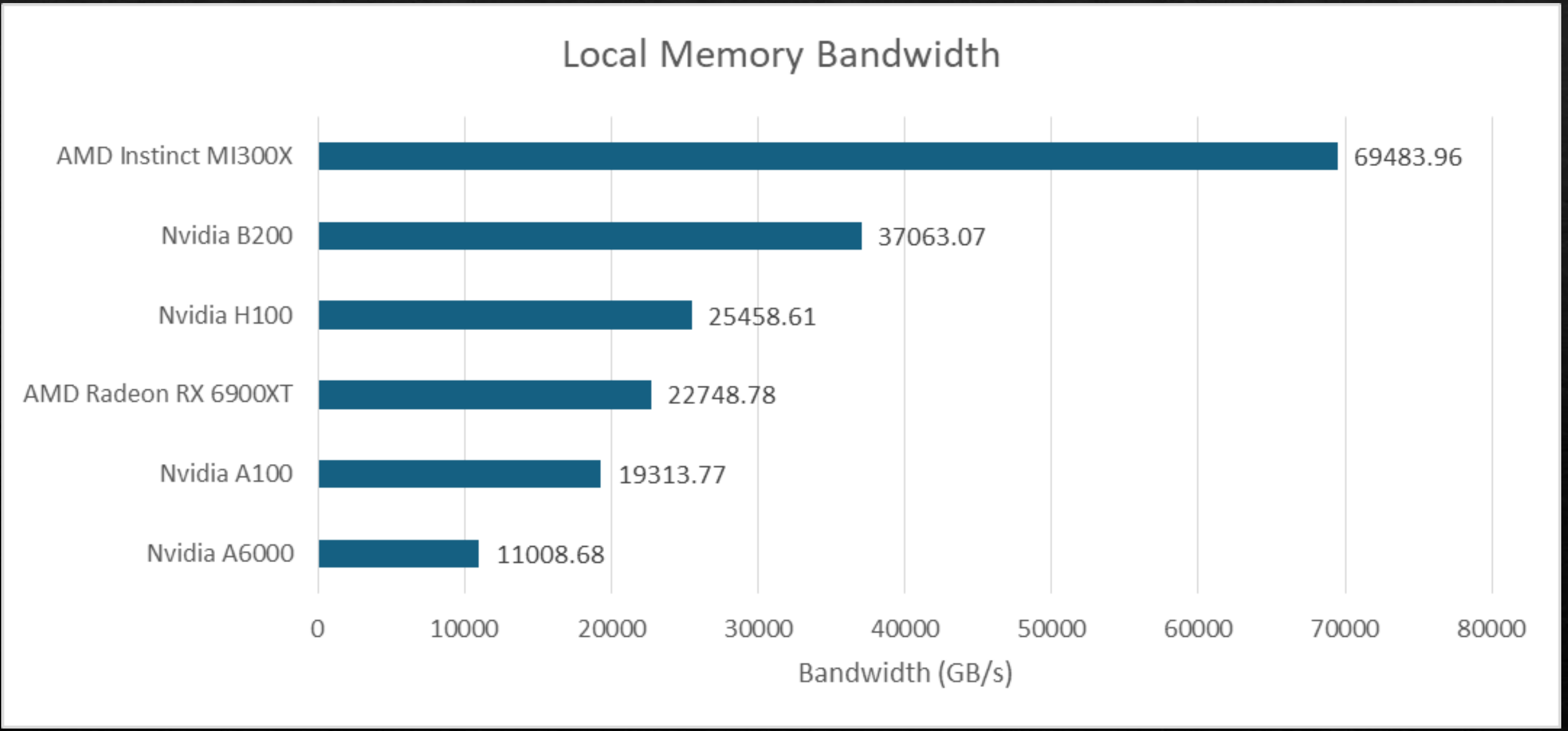

Multithreaded CUDA Program
Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
GPU with 2 SMs
SM 0
SM 1
Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
GPU with 4 SMs
SM 0
SM 1
SM 2
SM 3
Block 0
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7

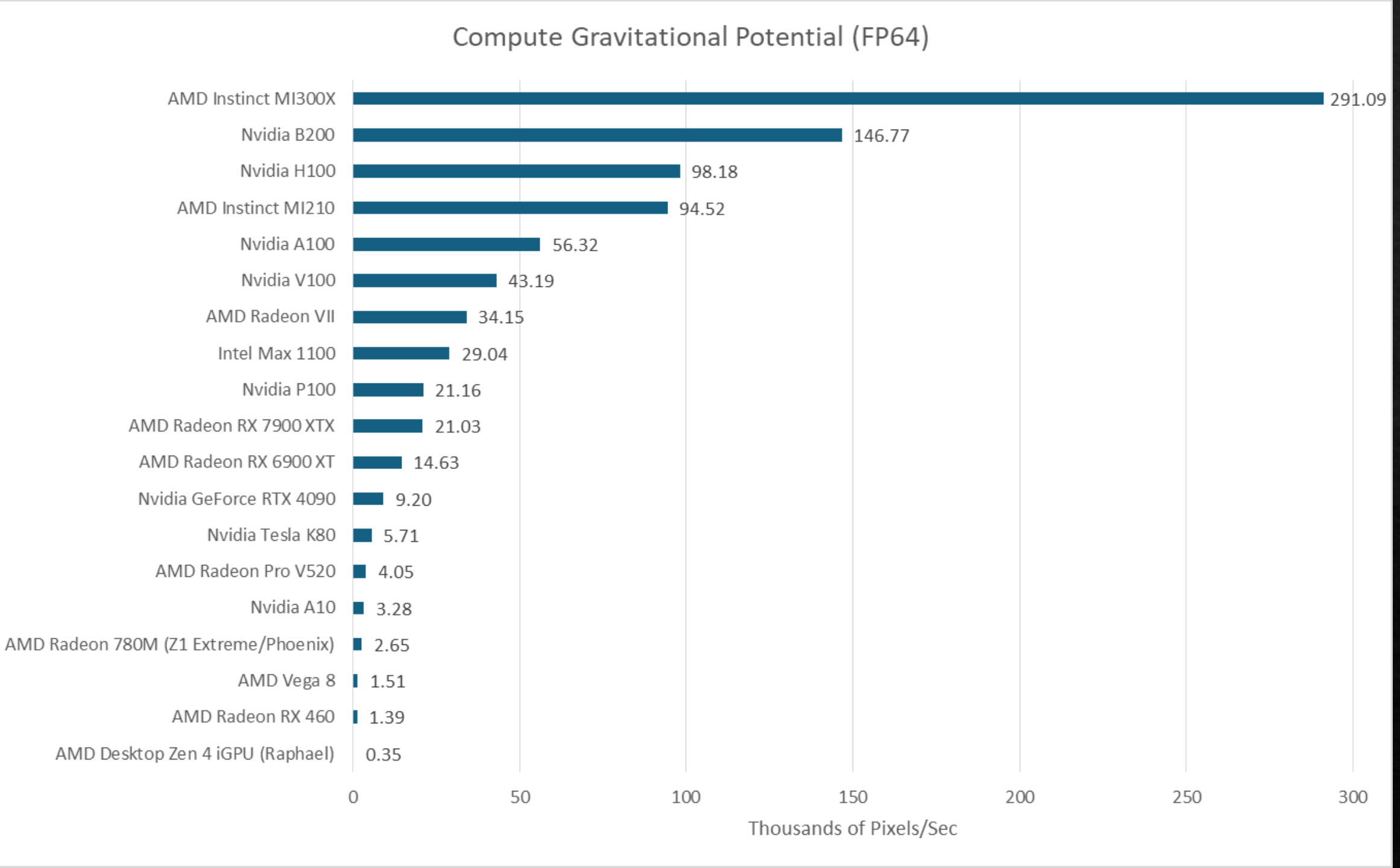
روندهای تکنولوژیک

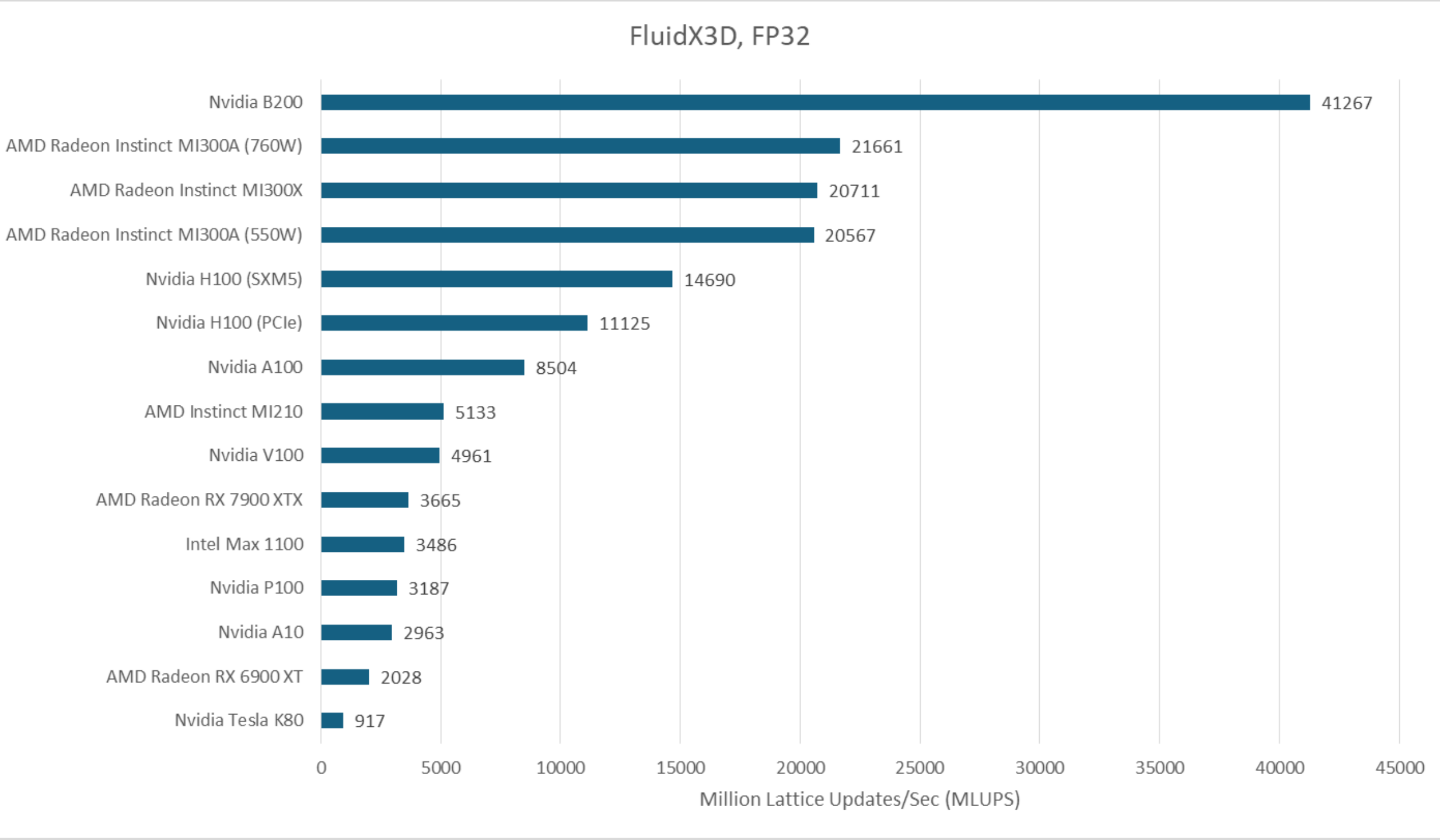
GPU


Local Memory Latency
Nvidia B200
14.84
AMD Radeon RX 9070
16.30
Nvidia H100
16.65
AMD Radeon RX 6900XT
19.30
Nvidia A100
20.57
AMD Radeon Instinct MI300X
28.72
AMD Radeon Instinct MI210
35.86
0
5
10
15
20
25
30
35
40
Latency (ns)


First Level Cache Bandwidth
Nvidia B200
36927.11
AMD Radeon Instinct MI300X
35866.79
AMD Radeon RX 6900XT
23399.19
Nvidia H100
22172.55
Nvidia A6000
12104.79
Nvidia A100
9709.06
0
5000
10000
15000
20000
25000
30000
35000
40000
Bandwidth (GB/s)


Local Memory Bandwidth
AMD Instinct MI300X
69483.96
Nvidia B200
37063.07
Nvidia H100
25458.61
AMD Radeon RX 6900XT
22748.78
Nvidia A100
19313.77
Nvidia A6000
11008.68
0
10000
20000
30000
40000
50000
60000
70000
80000
Bandwidth (GB/s)


Compute Gravitational Potential (FP64)
AMD Instinct MI300X
291.09
Nvidia B200
146.77
Nvidia H100
98.18
AMD Instinct MI210
94.52
Nvidia A100
56.32
Nvidia V100
43.19
AMD Radeon VII
34.15
Intel Max 1100
29.04
Nvidia P100
21.16
AMD Radeon RX 7900 XTX
21.03
AMD Radeon RX 6900 XT
14.63
Nvidia GeForce RTX 4090
9.20
Nvidia Tesla K80
5.71
AMD Radeon Pro V520
4.05
Nvidia A10
3.28
AMD Radeon 780M (Z1 Extreme/Phoenix)
2.65
AMD Vega 8
1.51
AMD Radeon RX 460
1.39
AMD Desktop Zen 4 iGPU (Raphael)
0.35
0
50
100
150
200
250
300
Thousands of Pixels/Sec


FluidX3D, FP32
Nvidia B200
41267
AMD Radeon Instinct MI300A (760W)
21661
AMD Radeon Instinct MI300X
20711
AMD Radeon Instinct MI300A (550W)
20567
Nvidia H100 (SXM5)
14690
Nvidia H100 (PCIe)
11125
Nvidia A100
8504
AMD Instinct MI210
5133
Nvidia V100
4961
AMD Radeon RX 7900 XTX
3665
Intel Max 1100
3486
Nvidia P100
3187
Nvidia A10
2963
AMD Radeon RX 6900 XT
2028
Nvidia Tesla K80
917
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
Million Lattice Updates/Sec (MLUPS)

برنامه آینده
Nvidia

Oberon Rack
Liquid Cooled

Blackwell Ultra NVL72

Second Half 2025

1.1 EF Dense FP4 Inference
0.36 EF FP8 Training
1.5X GB200 NVL72

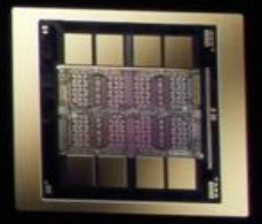
New Attention Instructions
2X

20 TB HBM | 40 TB Fast Memory
1.5X

14.4 TB/s CX8
2X

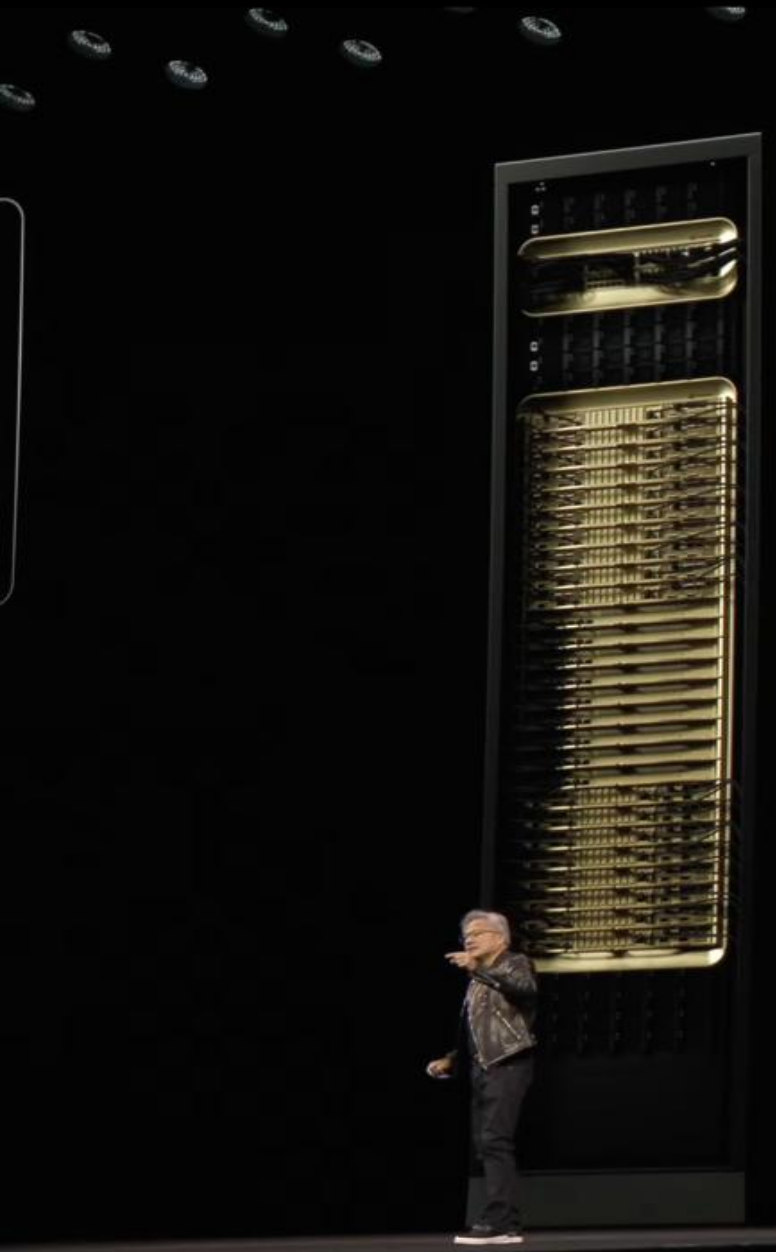
Grace

Blackwell Ultra



2 Reticle-Sized GPUs
Dense FP4 | 288GB HBM3e

NVIDIA



Vera Rubin NVL144

Second Half 2026

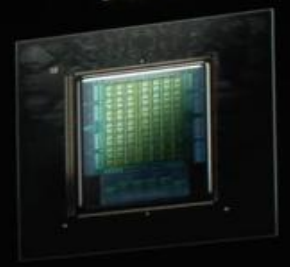
3.6 EF FP4 Inference
1.2 EF FP8 Training
3.3X GB300 NVL72

13 TB/s HBM4
75 TB Fast Memory
1.6X

260 TB/s NVLink6
2X

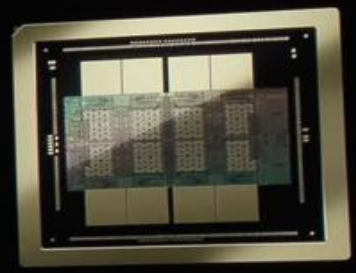
28.8 TB/s CX9
2X

Vera



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

Rubin



2 Reticle-Sized GPUs
50PF FP4 | 288GB HBM4

NVIDIA

Kyber Rack
Liquid Cooled

Rubin Ultra NVL576

Second Half 2027

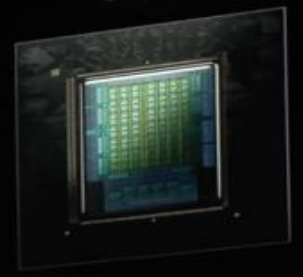
15 EF FP4 Inference
5 EF FP8 Training
14X GB300 NVL72

4.6 PB/s HBM4e
365 TB Fast Memory
8X

1.5 PBs NVLink7
12X

115.2 TB/s CX9
8X

Vera



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

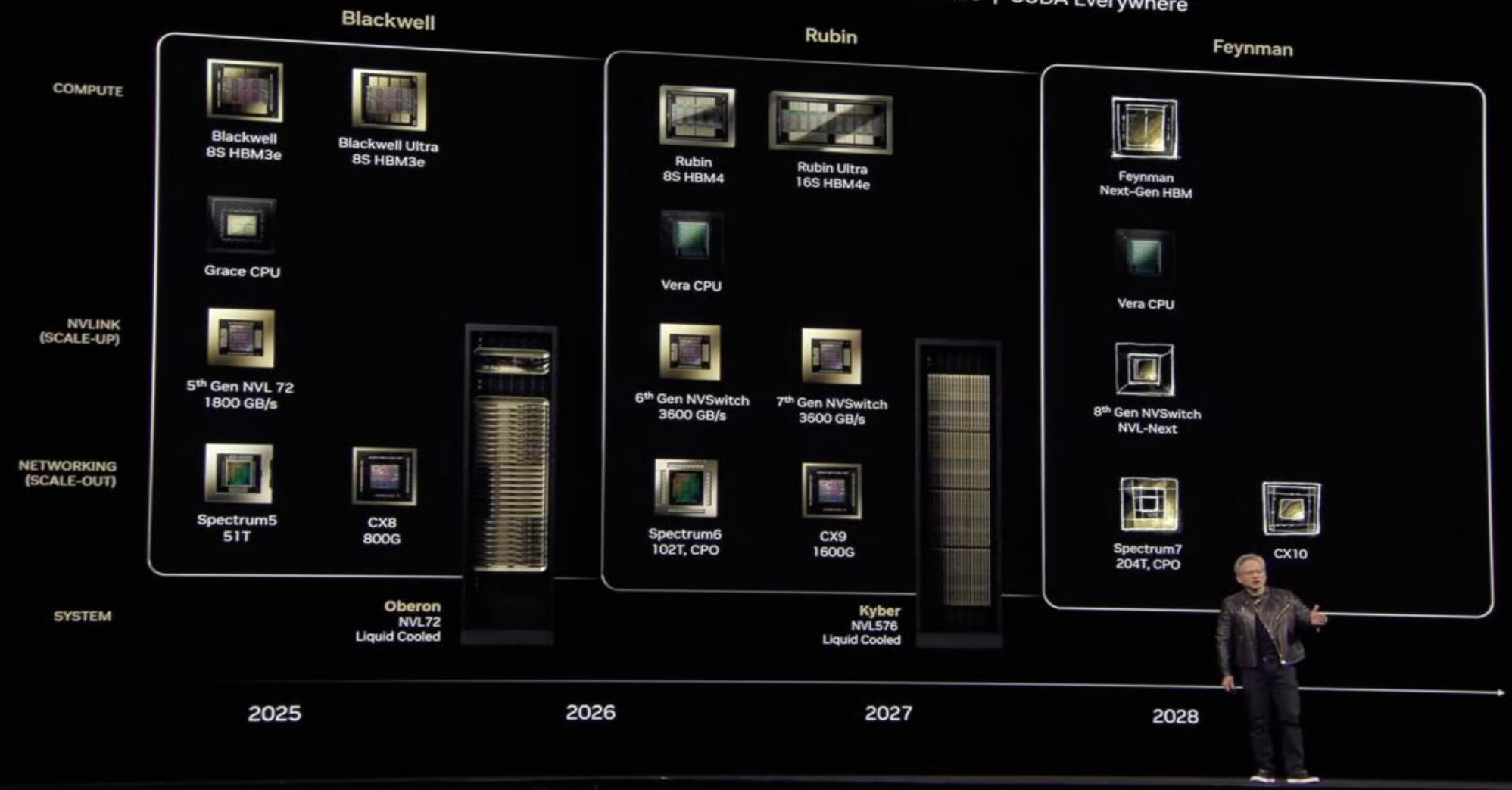
Rubin Ultra

4 Reticle-Sized GPUs
100PF FP4 | 1TB HBM4e

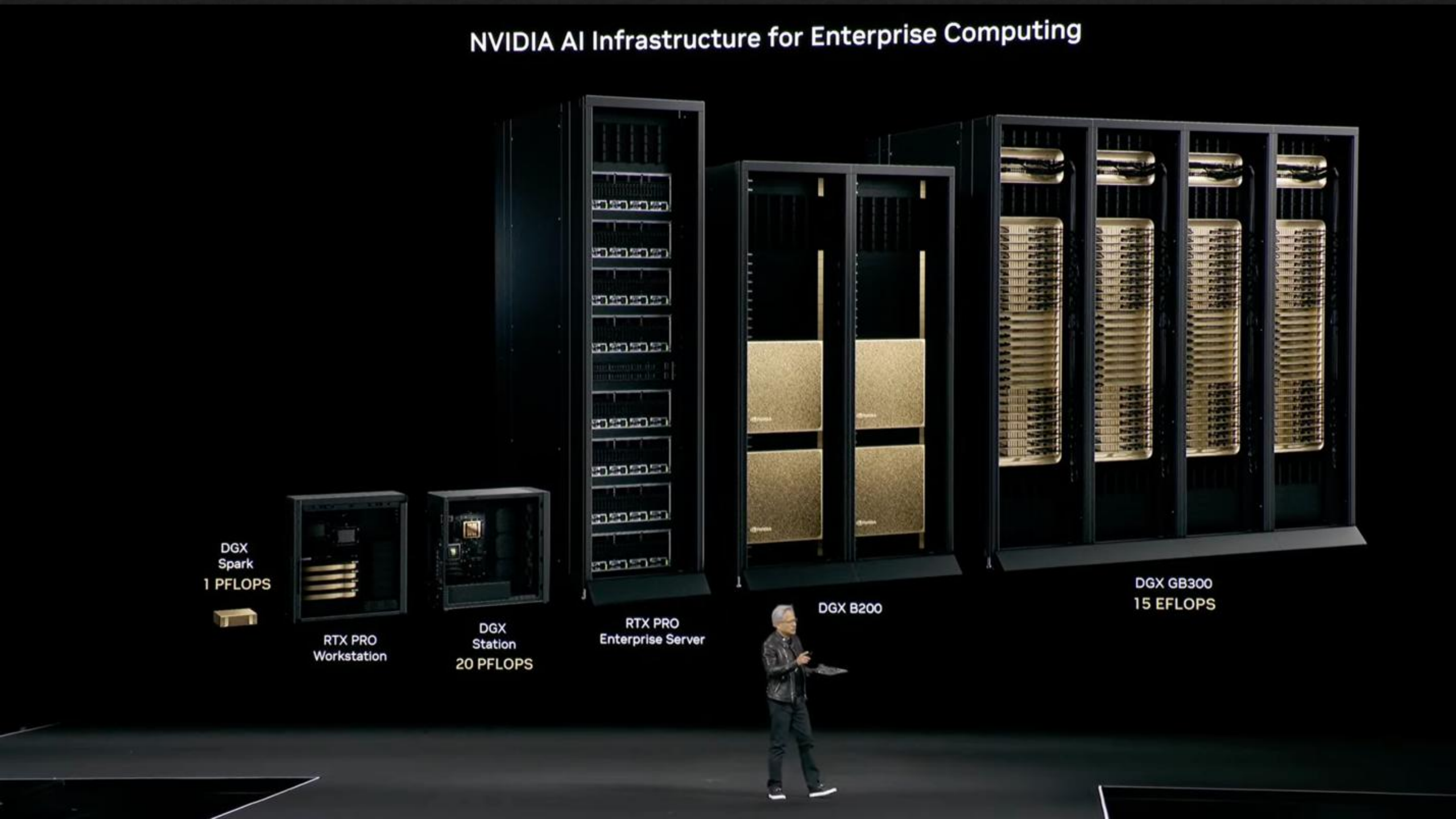
NVIDIA

NVIDIA Paves Road to Gigawatt AI Factories

One-Year Rhythm | Full-Stack | One Architecture | CUDA Everywhere

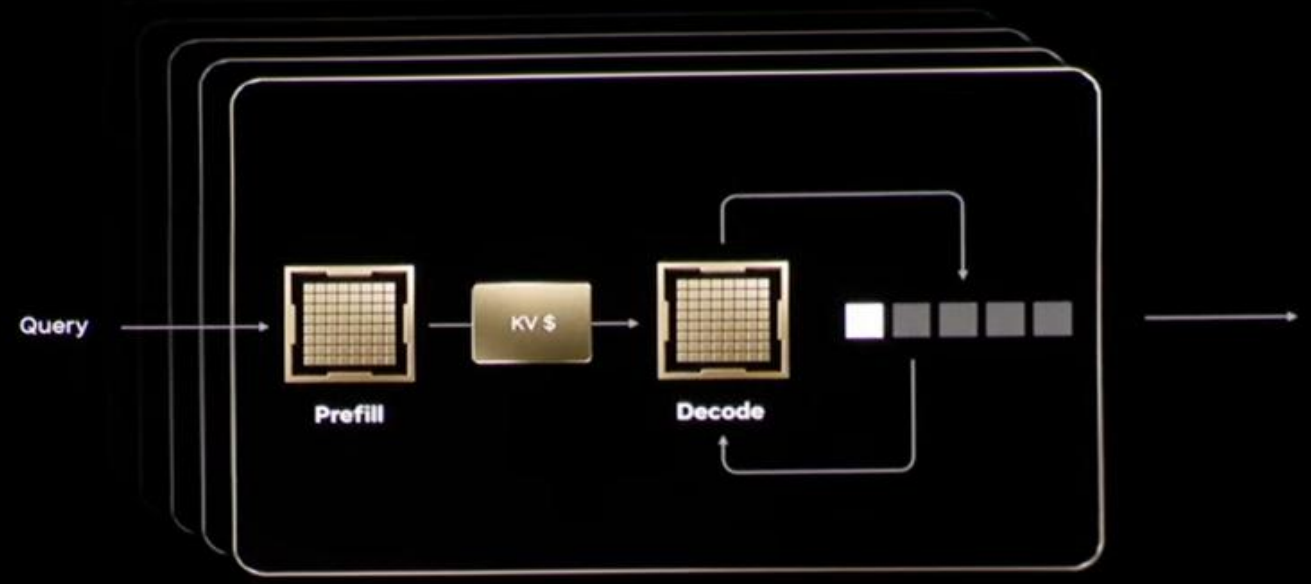


NVIDIA AI Infrastructure for Enterprise Computing



Announcing NVIDIA Dynamo

Distributed Inference Serving Library



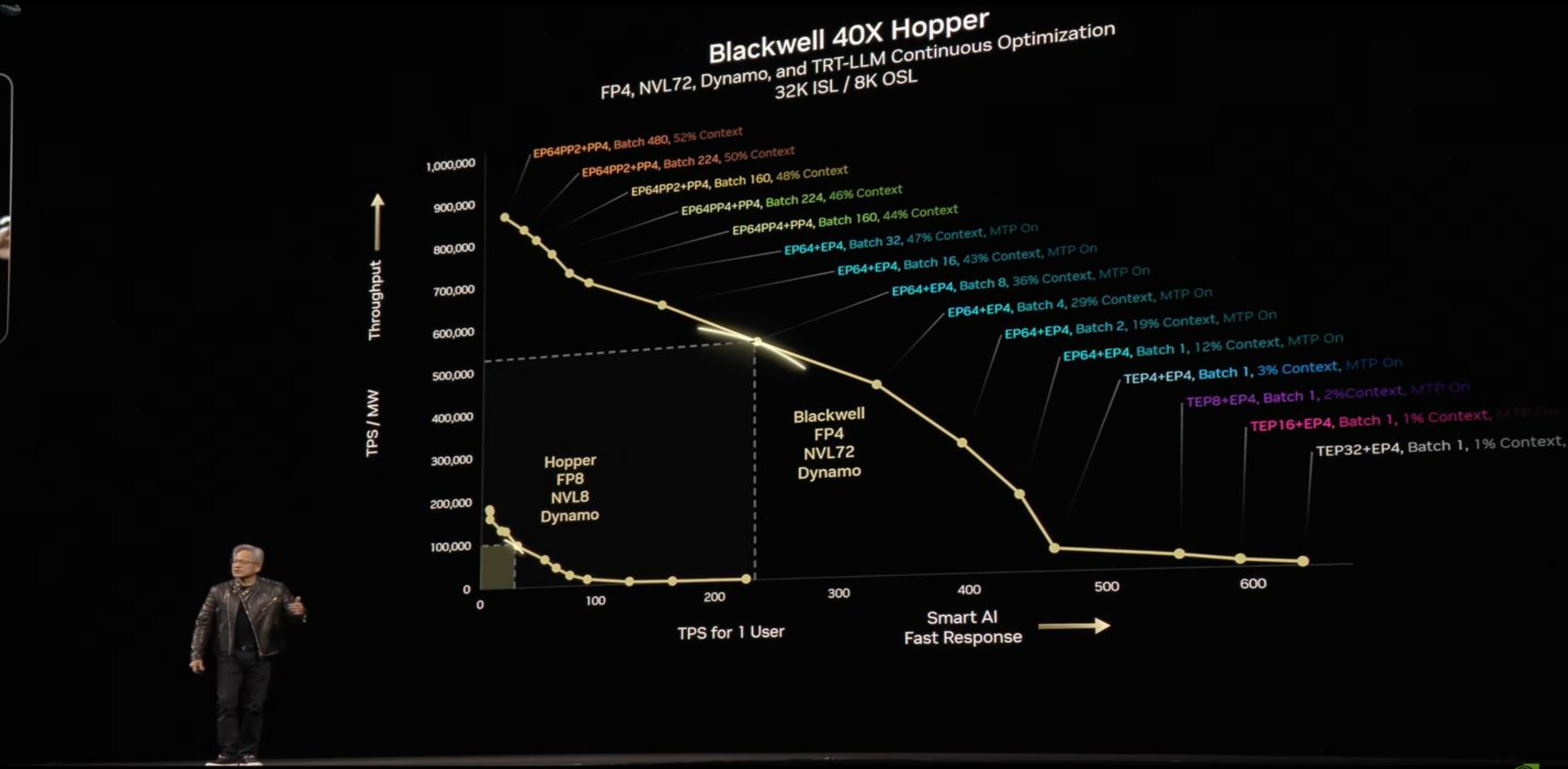
Disaggregated Inference

GPU Resource Allocation

KV Cache Routing

Communication Library (NIXL)

cohere Fireworks AI Meta Microsoft Azure perplexity Py SGL together.ai VAST vLLM


Blackwell 40X Hopper
FP4, NVL72, Dynamo, and TRT-LLM Continuous Optimization
32K ISL / 8K OSL
TPS / MW
Throughput
1,000,000
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
EP64PP2+PP4, Batch 480, 52% Context
EP64PP2+PP4, Batch 224, 50% Context
EP64PP2+PP4, Batch 160, 48% Context
EP64PP4+PP4, Batch 224, 46% Context
EP64PP4+PP4, Batch 160, 44% Context
EP64+EP4, Batch 32, 47% Context, MTP On
EP64+EP4, Batch 16, 43% Context, MTP On
EP64+EP4, Batch 8, 36% Context, MTP On
EP64+EP4, Batch 4, 29% Context, MTP On
EP64+EP4, Batch 2, 19% Context, MTP On
EP64+EP4, Batch 1, 12% Context, MTP On
TEP4+EP4, Batch 1, 3% Context, MTP On
TEP8+EP4, Batch 1, 2%Context, MTP On
TEP16+EP4, Batch 1, 1% Context, MTP On
TEP32+EP4, Batch 1, 1% Context,
Hopper
FP8
NVL8
Dynamo
Blackwell
FP4
NVL72
Dynamo
0
100
200
300
400
500
600
TPS for 1 User
Smart AI
Fast Response

CUDA-X FOR EVERY INDUSTRY

WARP
PHYSICS

cuDF
cuML
DATA SCIENCE AND PROCESSING

cuDSS
cuSPARSE
cuFFT
AMGX
COMPUTER AIDED ENGINEERING

TRT-LLM
MEGATRON
NCCL
cuDDN
CUTLASS
cuBLAS
DEEP LEARNING

cuEQUIVARIANCE
cuTENSOR
QUANTUM CHEMISTRY

cuQUANTUM
CUDA-Q
QUANTUM COMPUTING

EARTH-2
WEATHER ANALYTICS

MONAI
MEDICAL IMAGING

PARABRICKS
GENE SEQUENCING

cuOPT
DECISION OPTIMIZATION

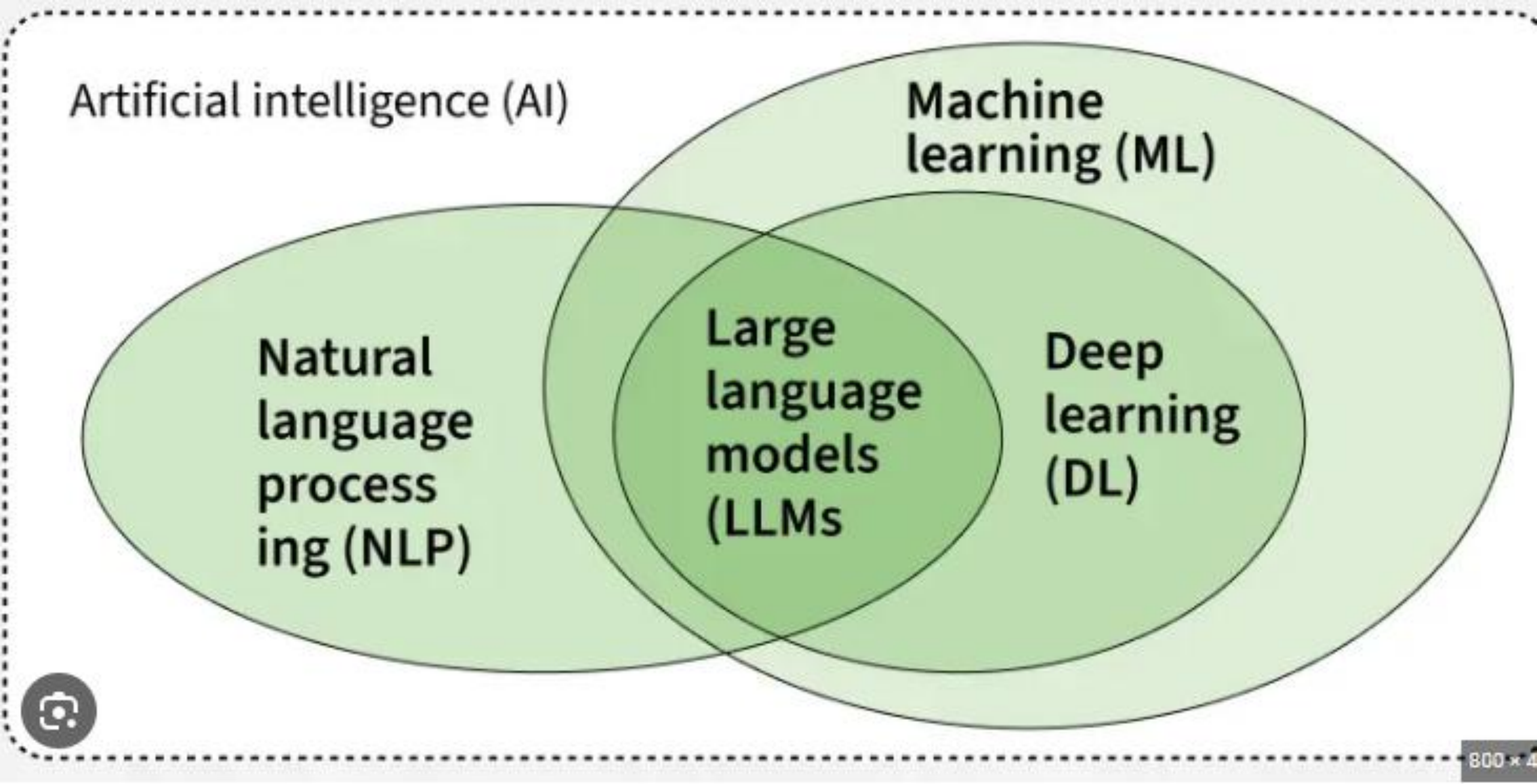
AERIAL
SIONNA
5G/6G SIGNAL PROCESSING

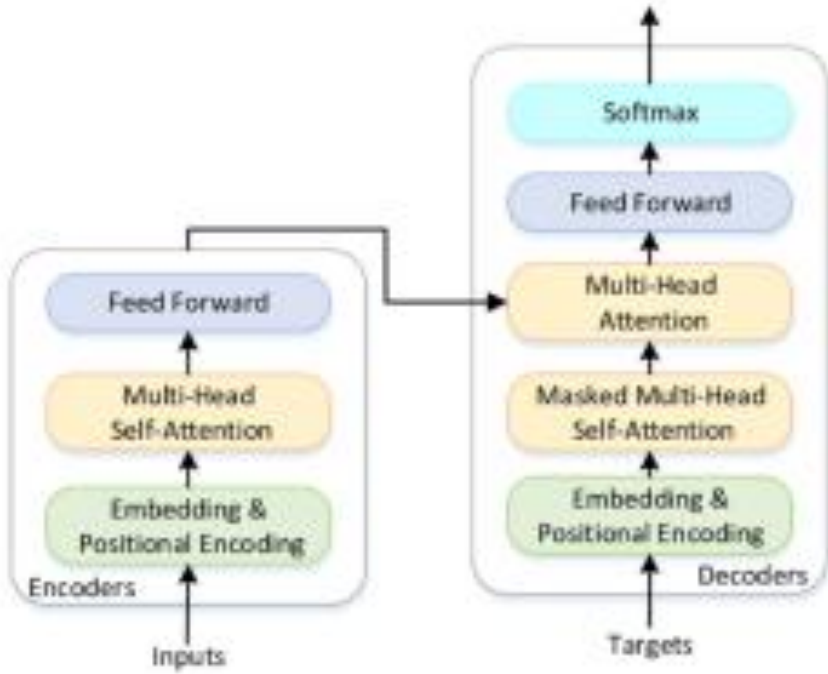
cuLITHO
COMPUTATIONAL LITHOGRAPHY

cuPYNUMERIC
NUMERICAL COMPUTING

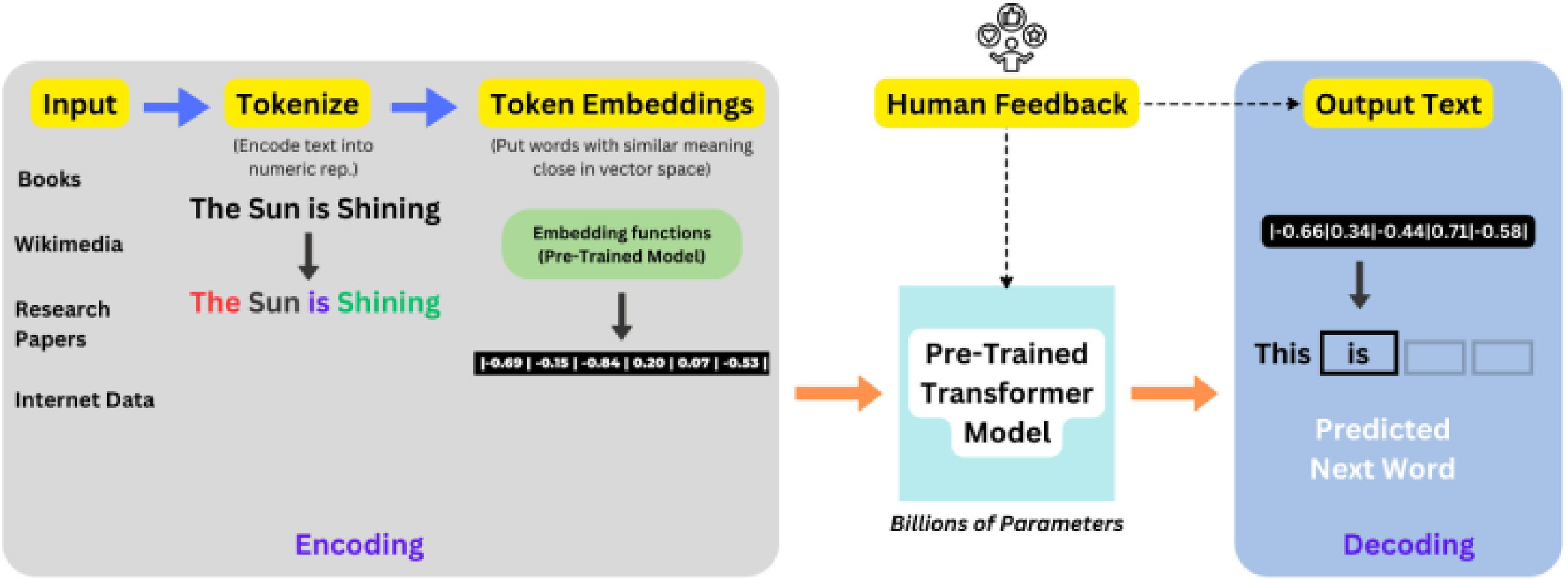
LLM

به زبان ساده


Artificial intelligence (AI)
Machine learning (ML)
Natural language process ing (NLP)
Large language models (LLMs
Deep learning (DL)



Data + Architecture + Training = LLM


Input
Tokenize
Token Embeddings
(Encode text into numeric rep.)
(Put words with similar meaning close in vector space)
Books
Wikimedia
Research Papers
Internet Data
The Sun is Shining
The Sun is Shining
Embedding functions (Pre-Trained Model)
|-0.69 | -0.15 | -0.84 | 0.20 | 0.07 | -0.53 |
Encoding
Human Feedback
Pre-Trained Transformer Model
Billions of Parameters
Output Text
|-0.66|0.34|-0.44|0.71|-0.58|
This
is
Predicted Next Word
Decoding


1 Data Gathering
2 Data Cleaning
3 Data Splitting
4 Model Set-up
LLM Models Training
5 Model Training
6 Model Checking
7 Model Usability
8 Model Enhancement

کاربردهای LLMs

Exceptional Large Language Model Use Cases

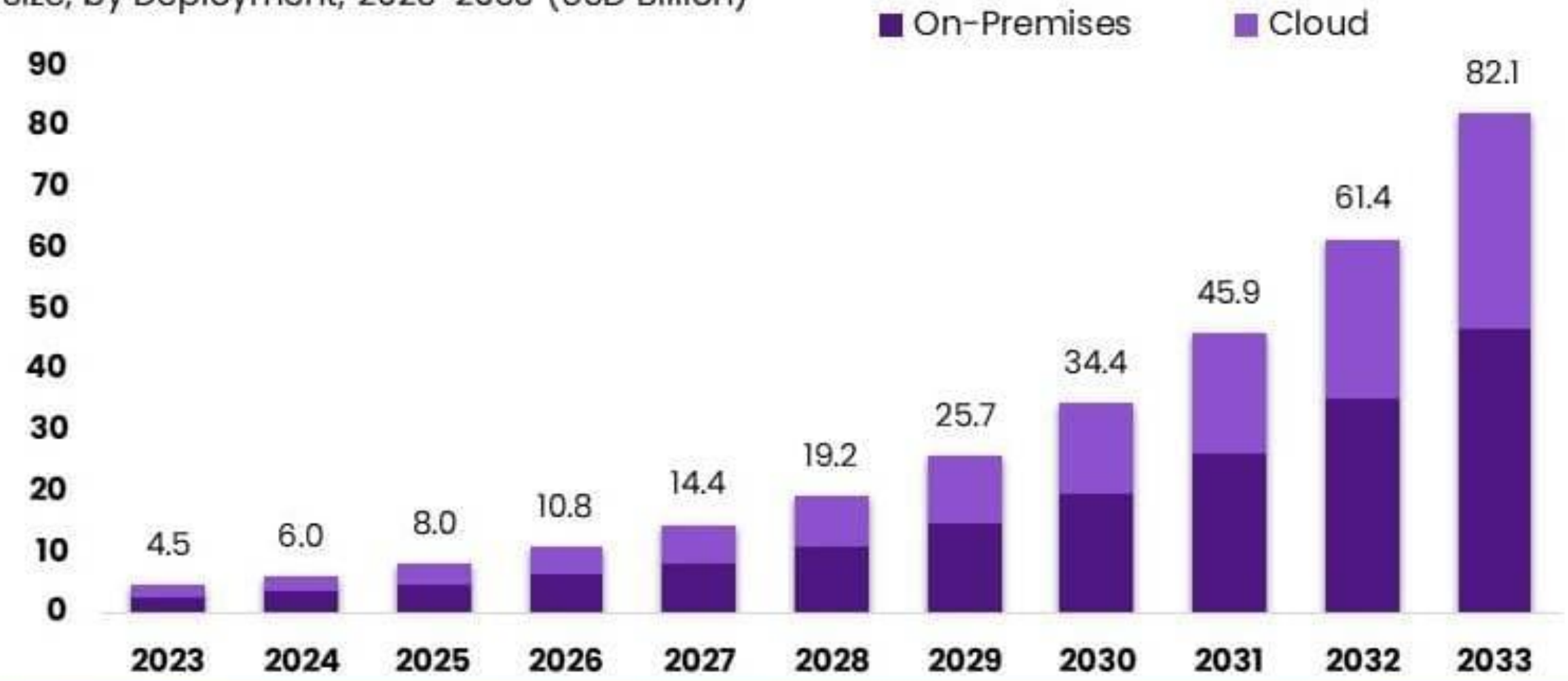
SoluLab

Application of LLMs

اہمیت LLMs

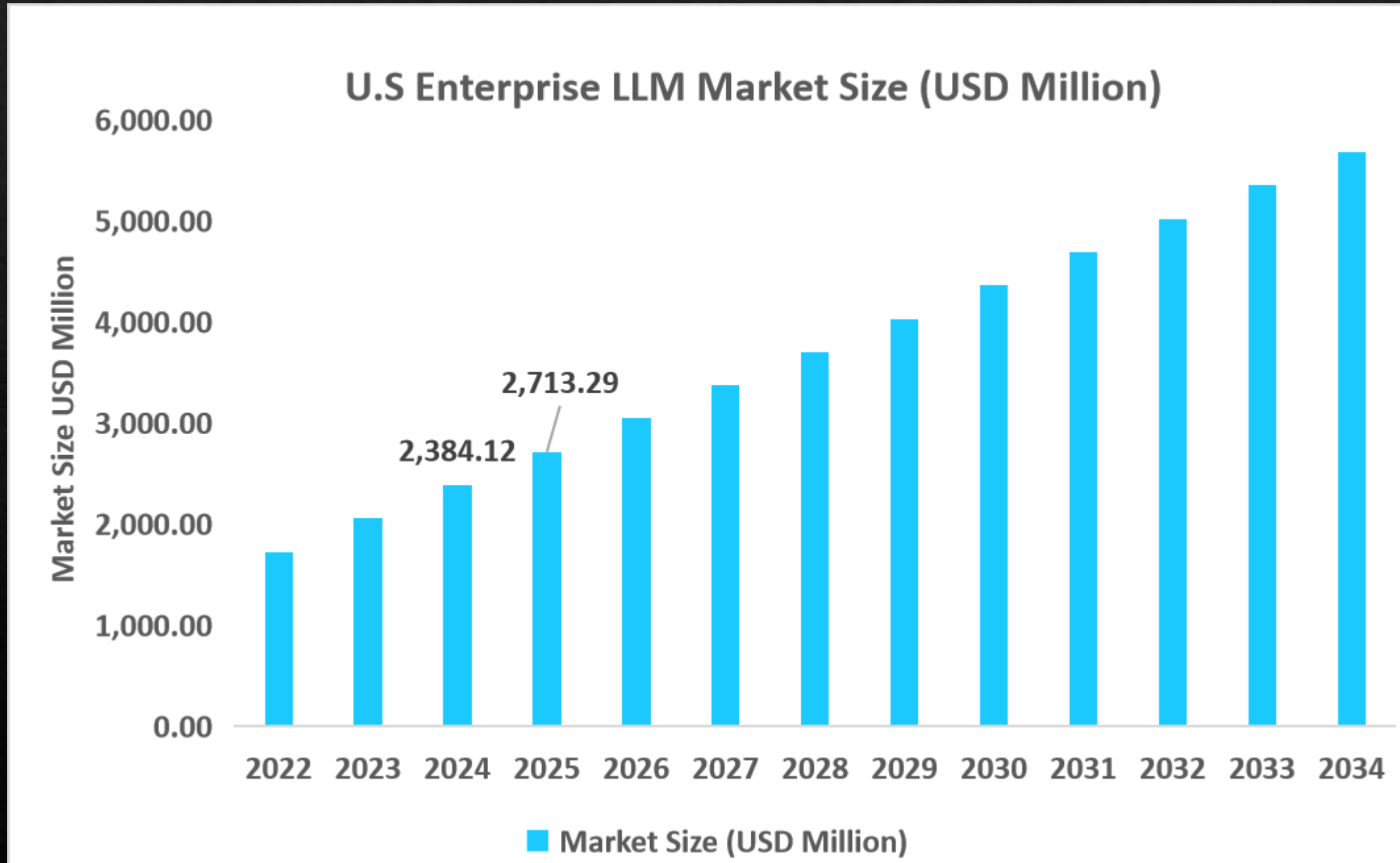
Global Large Language Model (LLM) Market

Size, by Deployment, 2023-2033 (USD Billion)

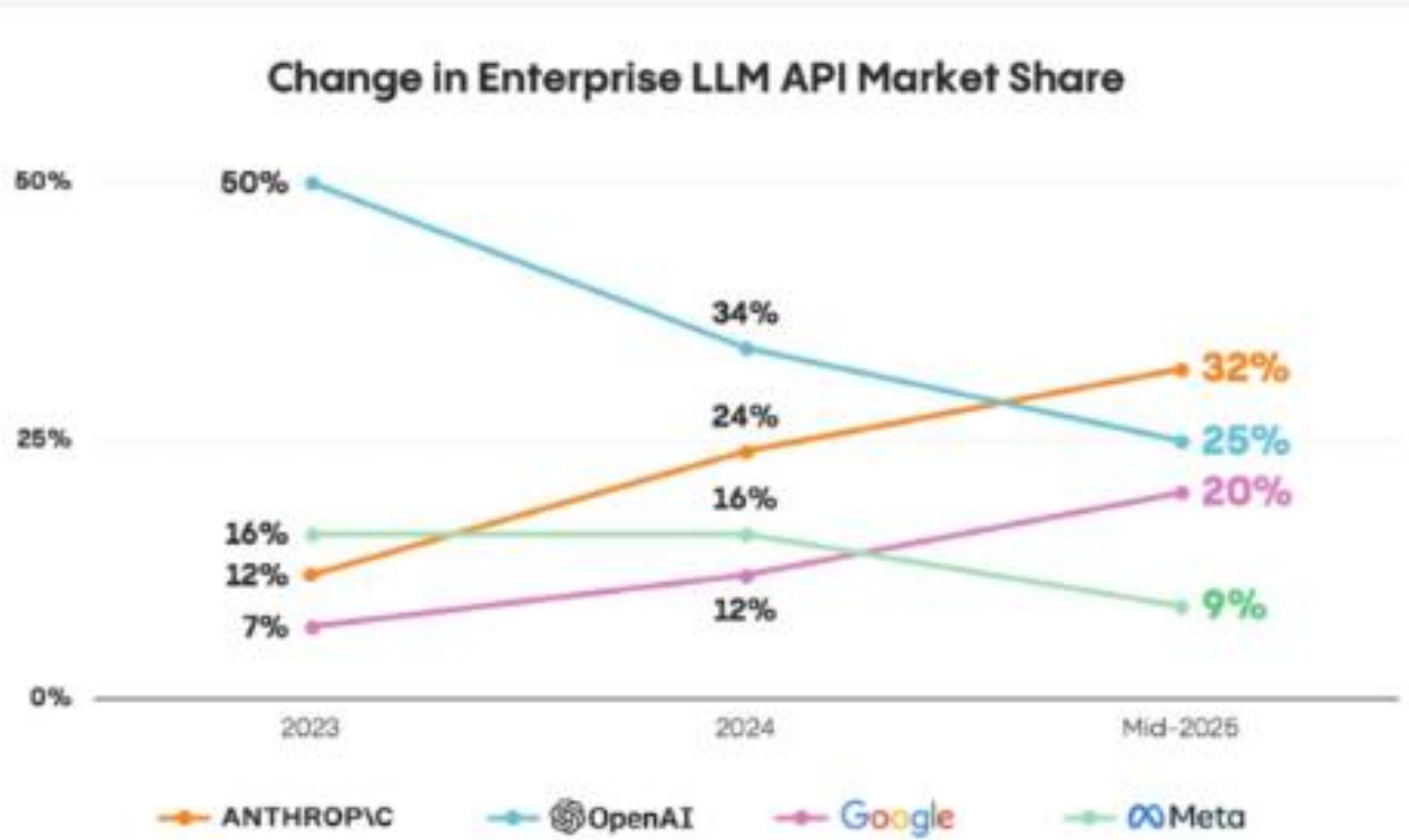


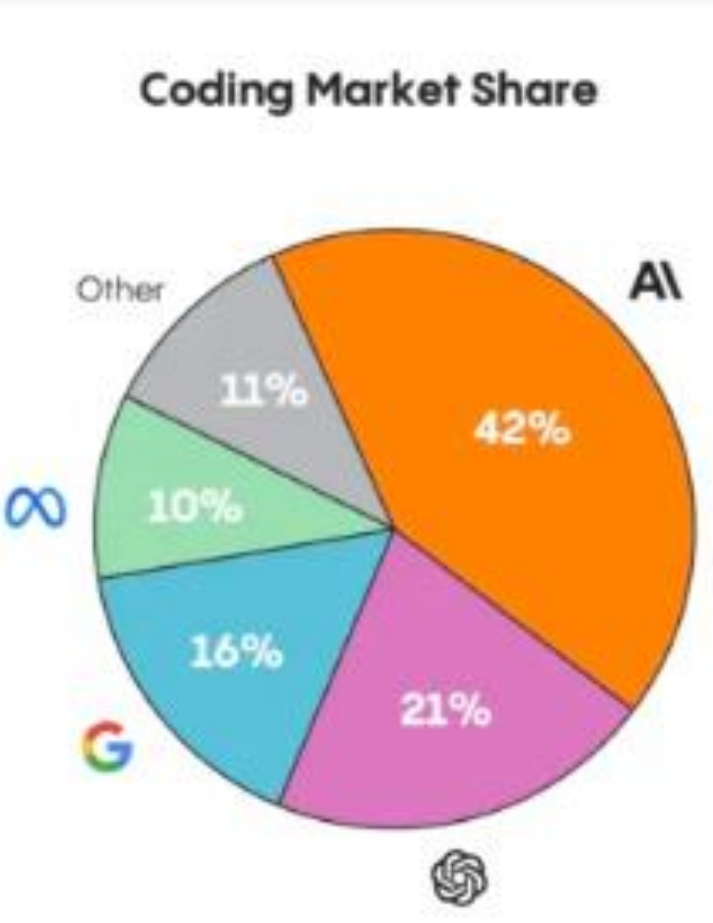
The Market will Grow At the CAGR of: 33.7% **The Forecasted Market Size for 2033 in USD: $82.1 B**

market.us

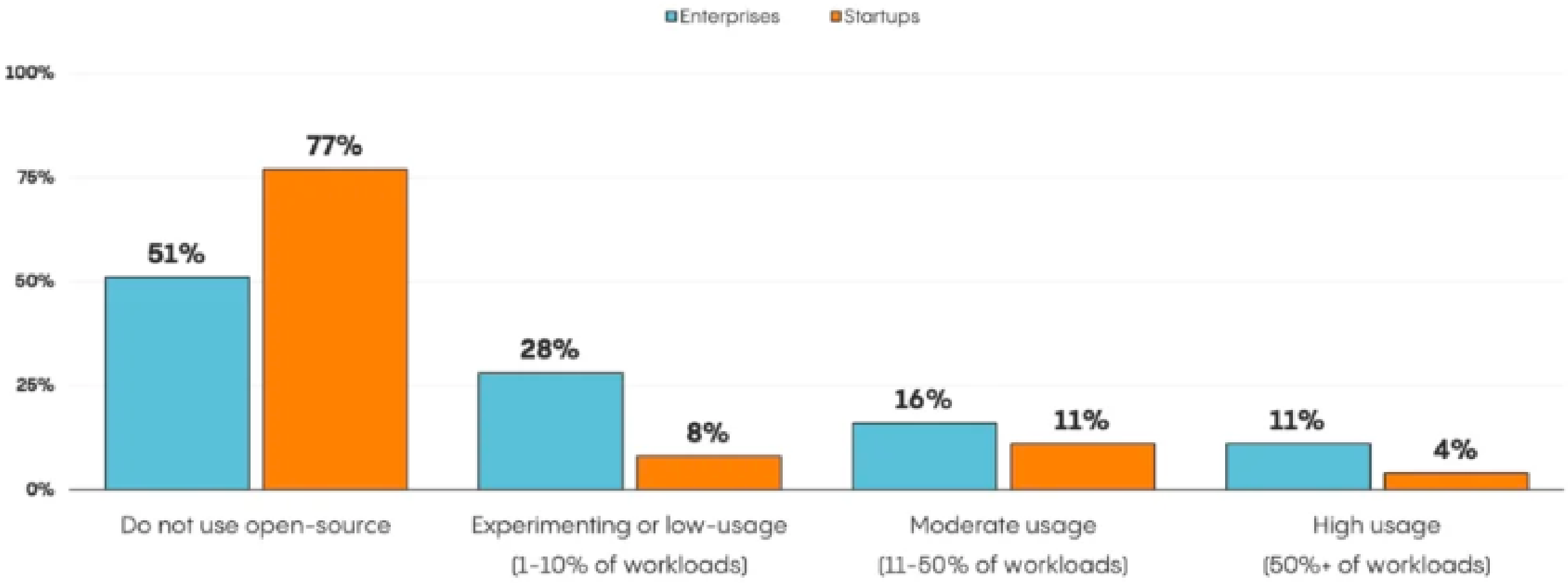

U.S Enterprise LLM Market Size (USD Million)
Market Size USD Million
6,000.00
5,000.00
4,000.00
3,000.00
2,000.00
1,000.00
0.00
2,384.12
2,713.29
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
Market Size (USD Million)

Enterprise LLM API Market Share by Usage



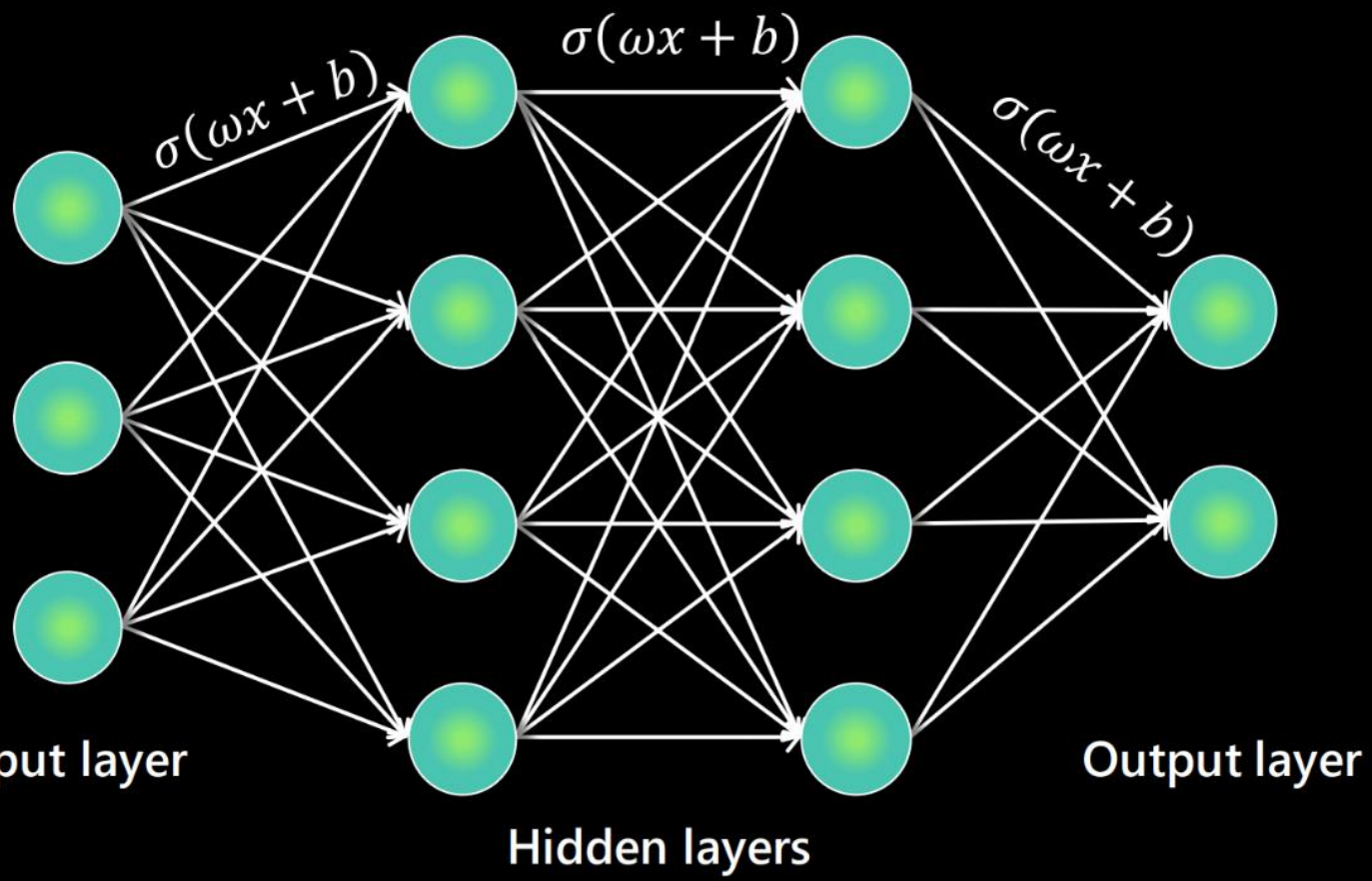


Companies Choose Closed-Source



How large are they?

Transformer model Neural network



Function: weight * input plus bias

BERT Large - 2018
345M

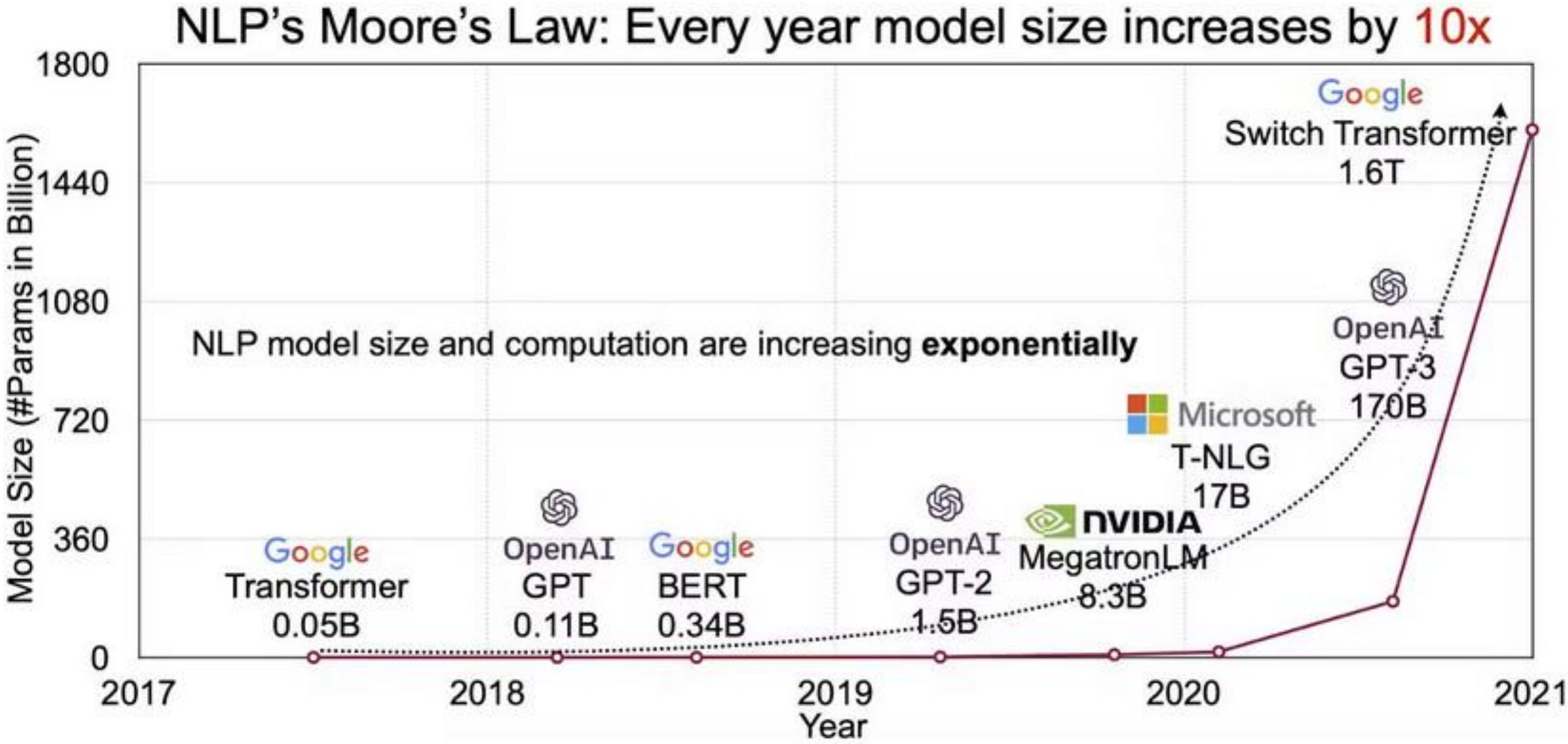
GPT2 - 2019
1.5B

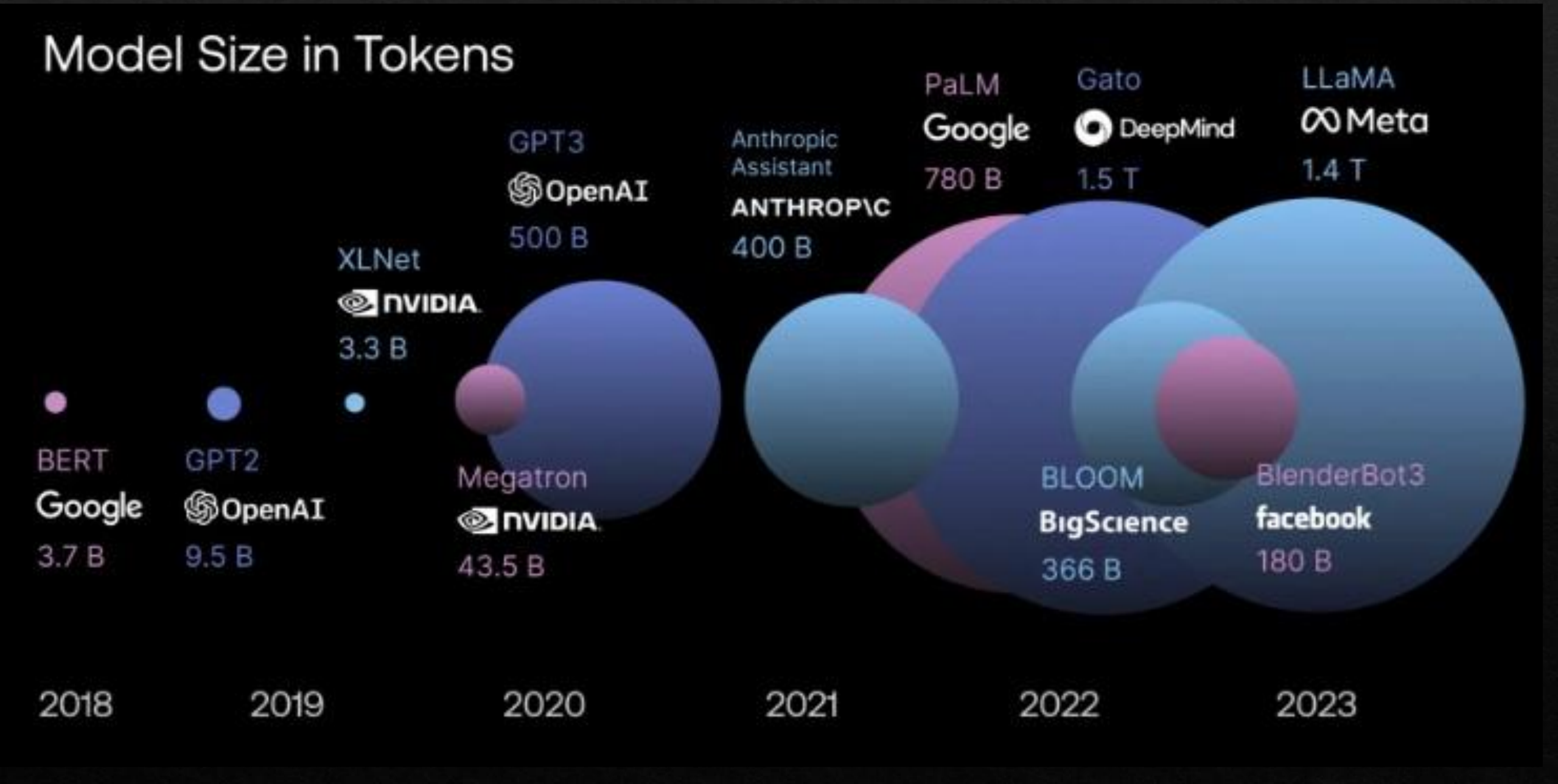
GPT3 - 2020
175B

Turing Megatron NLG 2021
530B

GPT4 – 2023
1.4T (estimated)

Rapid change in model size




Model Size in Tokens
BERT
Google
3.7 B
GPT2
OpenAI
9.5 B
XLNet
NVIDIA
3.3 B
GPT3
OpenAI
500 B
Megatron
NVIDIA
43.5 B
Anthropic
Assistant
ANTHROP\C
400 B
PaLM
Google
780 B
Gato
DeepMind
1.5 T
LLaMA
Meta
1.4 T
BLOOM
BigScience
366 B
BlenderBot3
facebook
180 B
2018
2019
2020
2021
2022
2023

Map of LLMs
October 2023

CodeLlama
- WizardCoder
- Phind-CodeLlama
- OpenAssistant-Codellama
- NexusRaven

Mistral
- Mistralic
- SQLCoder-7B
- Mistral-OpenOrca

StarCoder
- SQLCoder-15B
- WizardCoder-15B

Llama2
- OpenChat
- CodeLlama
- Nous-Hermes
- Vicuna
- NSQL
- Redmond-Puffin
- Llama2-32K
- OpenAssistant-Llama2
- WizardLM

GPT
- GPT-3.5-Turbo
- GPT-4

Stable
- StableLM
- StableCode
- OpenAssistant-StableLM

Pythia
- Pythia-ChatBase
- OpenAssistant-Pythia

Claude
- Claude-2

PaLM
- PaLM-2

Falcon
- Falcon

Qwen
- Qwen

Top 10 Large Language Models in 2025

SEVEN SQUARE

Dextra
delivering 10X transformation

Top 15 LLM to boost

your performance, capabilities and innovation

- GPT-4: OpenAI
- GPT-3: OpenAI
- GPT-3.5: OpenAI
- Gemini: Google DeepMind
- LLaMA: Meta AI
- PaLM 2: Google
- Bard: Google
- Claude v1: Anthropic
- Falcon: TII (UAE)
- Cohere: Cohere.ai
- Orca: Microsoft
- Guanaco: HuggingFace/TUM
- Vicuna: LMSYS
- MPT-30B: MosaicML
- 30B Lazarus: CalderaAI

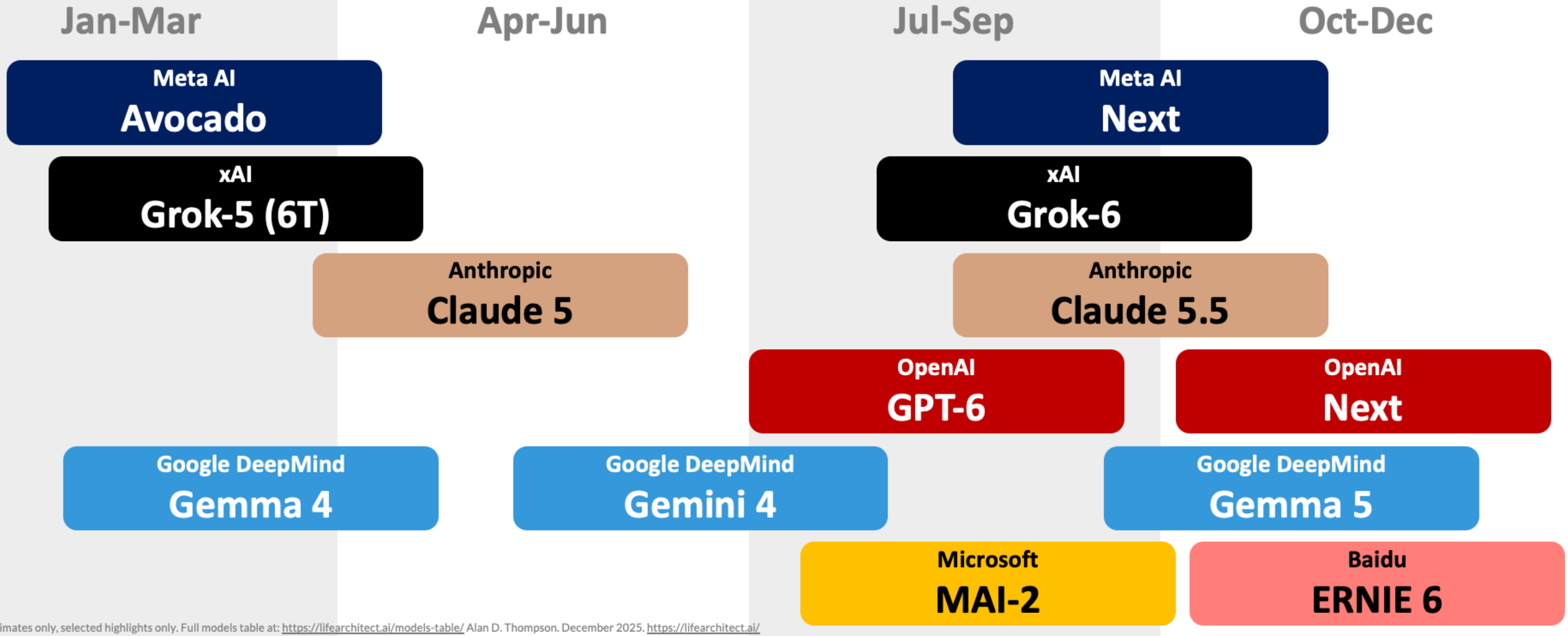
LLM Market Share

LLM	Global	US	India	Germany	France	Taiwan	Hong Kong	Russia
ChatGPT	78%	78%	80%	81%	87%	79%	49%	57%
DeepSeek	8%	4%	5%	6%	2%	2%	24%	25%
Gemini	6%	7%	6%	5%	5%	9%	4%	2%
Grok	2%	2%	4%	1%	0%	3%	6%	3%
Perplexity	2%	3%	2%	3%	2%	2%	11%	5%
Claude	2%	3%	2%	1%	1%	4%	1%	0%
Copilot	1%	2%	1%	2%	2%	1%	4%	0%
Qwen	0%	0%	0%	0%	0%	0%	0%	5%
Mistral	0%	0%	0%	0%	1%	0%	0%	2%
Hey & Go	0%	0%	0%	1%	0%	0%	0%	0%

Sessions (web+mobile), SimilarWeb, without apps, without APIs, Jan-Apr 2025, Malte Landwehr

LLMs: NEXT PUBLIC RELEASES IN 2026

ESTIMATES AS OF DECEMBER 2025



Estimates only, selected highlights only. Full models table at: https://lifearchitect.ai/models-table/ Alan D. Thompson. December 2025. https://lifearchitect.ai/

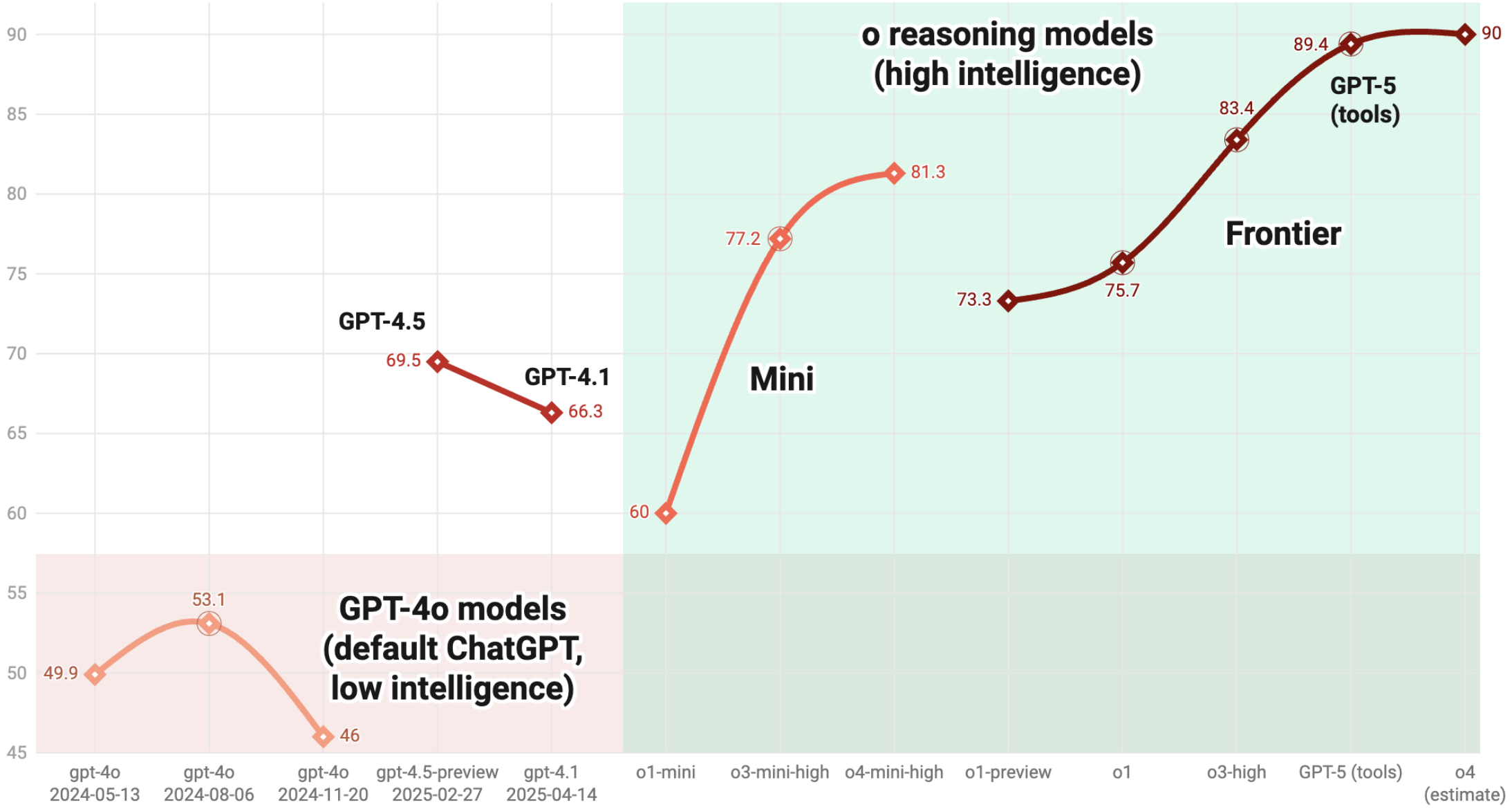
Model	Training end	Chip type	TFLOP/s (max)	Chip count	Wall clock (days)	Total time (years)	Retail (US$)	MMLU
GPT-3 175B	Apr/2020	V100	130	10,000	15 days	405y	$9M	43.9
Llama 1 65B	Jan/2023	A100	312	2,048	21 days	118y	$4M	63.4
Llama 2 70B	Jun/2023	A100	312	*2,048*	35 days	196y	$7M	68.0
Titan 200B	Apr/2023	A100	312	13,760	48 days	1,319y	$45M	70.4
GPT-4 1.7T	Aug/2022	A100	312	25,000	95 days	6,507y	$224M	86.4
Gemini	Nov/2023	TPUv4	275	57,000	100 days	15,616y	$440M	90.0
Llama 3 405B	Apr/2024	H100	989	24,576	50 days	3,366y	$125M	*85+*
GPT-5	*Apr/2024*	H100	989	*50,000*	*120 days*	*16,438y*	$612M	
Grok 2	*Jun/2024*	H100	989	20,000	*50 days*	*6,571y*	$245M	
Olympus	*Aug/2024*	H100	989					
Gemini 2	*Nov/2024*	TPUv6	*1,847*					
Grok 3	*Dec/2024*	H100	989	100,000	*50 days*	*32,855y*	$1.2B	

Alan D. Thompson. May/2024. LifeArchitect.ai

Table. Model training compute (see working, with sources[8]).

The GPT-4x and o Model Family: Varied Intelligence Scores 2024–2025

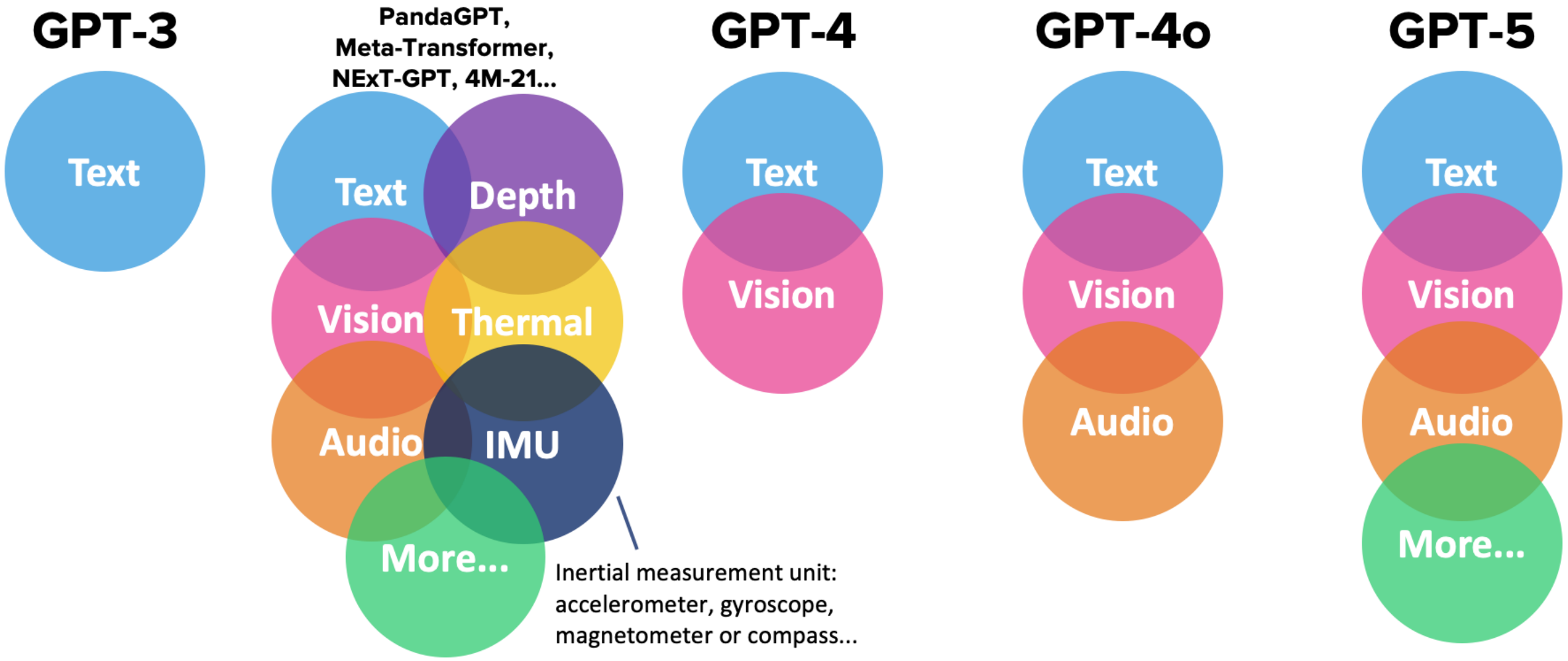
OpenAI's GPT-4x and o models on GPQA scores • LifeArchitect.ai



Source: LifeArchitect.ai/GPT-5 based on official eval data from GitHub.com/openai/simple-evals

Chart: Alan D. Thompson • Aug/2025 • Source: LifeArchitect.ai • Get the data • Created with Datawrapper

GPT MODALITIES (2024)



FRONTIER AI MODELS + HIGHLIGHTS 2026 H1

Google Gemma
Alibaba Qwen/QwQ
Mistral
Microsoft phi

+ hundreds more...

Meta AI Llama
OpenAI GPT-OSS
Kimi
DeepSeek R

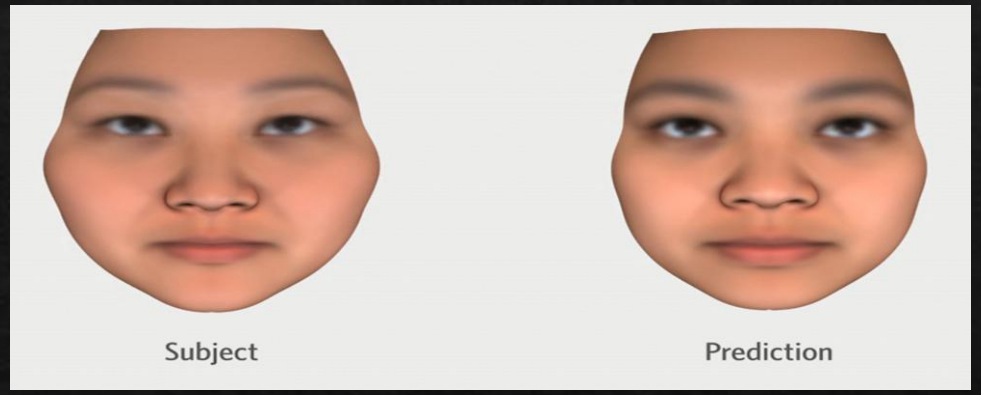
+ hundreds more...

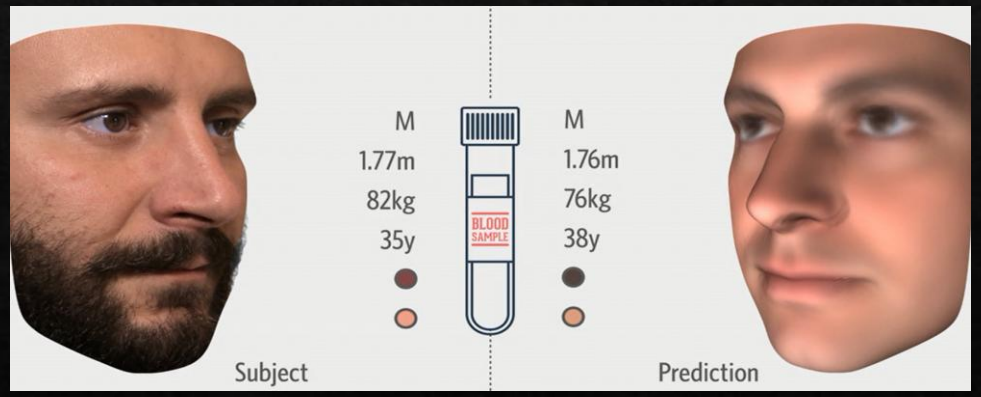
Selected highlights only. Model highlights: https://lifearchitect.ai/models-table/ Total models: https://huggingface.co/models?pipeline_tag=text-generation Some images by Flaticon.com. Alan D. Thompson. 2021–2026.

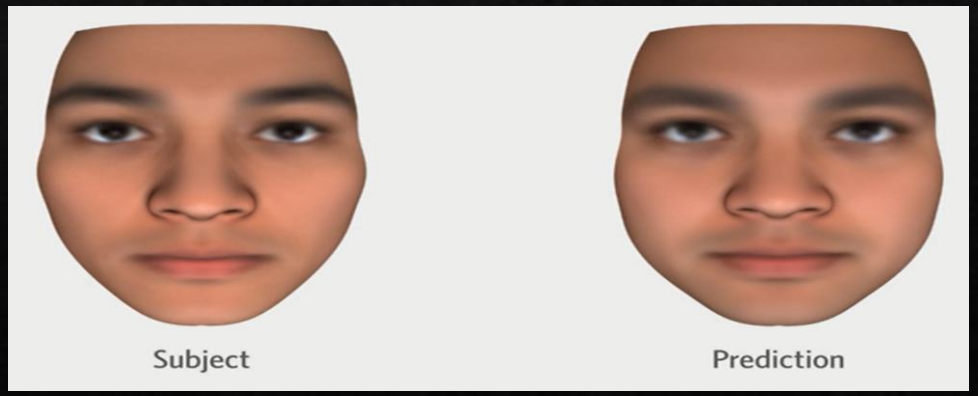

LARGE LANGUAGE MODEL MARKET MAP
LARGE LANGUAGE MODEL SOFTWARE PROVIDERS
LLM APIs
VECTOR DATABASES
LLM FRAMEWORKS
TEXT-TO-SPEECH
LLM MONITORING TOOLS
OpenAI
ANTHROP\C
cohere
aws
Pinecone
Chroma
LlamaIndex
LangChain
FIXIE
RESEMBLE.AI
ElevenLabs
WELLSAID
DISTYL
Guardrails AI
Helicone
LARGE LANGUAGE MODEL SERVICE PROVIDERS
COMPUTE PLATFORM PROVIDERS
MODEL HUBS
FINE-TUNING/CUSTOM MODEL TRAINING FRAMEWORKS
MONITORING/OBSERVABILITY PLATFORM PROVIDERS
HOSTING SERVICE PROVIDERS
Lambda
mosaicML
Azure
Hugging Face
Replicate
PyTorch
TensorFlow
LAMINI
Arthur
arize
WHYLABS
Replicate
Hugging Face
END-USERS
GOVERNMENT & REGULATORY BODIES
edger .finance
BEN
BRANDED ENTERTAINMENT NETWORK
summerhealth
OXIDE.AI
NIST
European Commission
ico.
Information Commissioner's Office
NITDA

چالش های هوش مصنوعی

شرح مساله - پیش بینی چهره در سال۲۰۱۶







دستکاری ژنوم انسانی و ساخت یک موجود جدید!