Master Informatics Eng.

2024/25

A.J.Proença

Multi/Manycore devices and GPU's as computing accelerators

(most images are from the web)

Analysis of critical compute devices in HPC: key issues

Critical performance devices in HPC systems

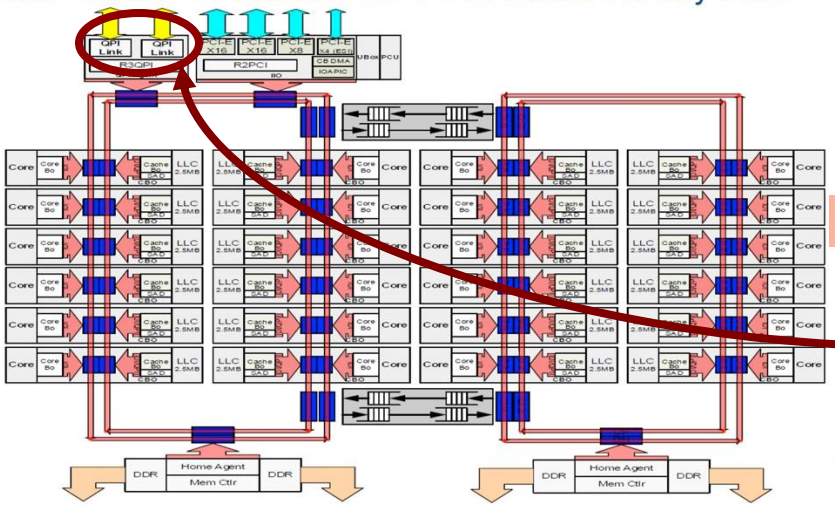
- PUs (*Processing Units*), namely the multi/manycore devices
- Compute accelerators (such as GPUs)
- Memory systems, interconnection fabric, fabrication process, …

Key issues

- on PUs and compute accelerators: see next slides…
- on memory devices
  - latency (technology dependent)
  - bandwidth (#mem channels, bus width, embedded RAM on-chip, …)
- interconnection network fabric
  - NoC (Network-on-Chip): to interconnect cores, caches and accelerators
  - between devices, both PUs and compute accelerators
- fabrication process
  - fabrication technology (5nm, 4nm, 3nm, …)
  - group cores in clusters and move to MCM/dies/chiplets

NoC structure in an HPC server: Intel move from ring to mesh

Intel® Xeon® Processor E5 v4 Product Family HCC

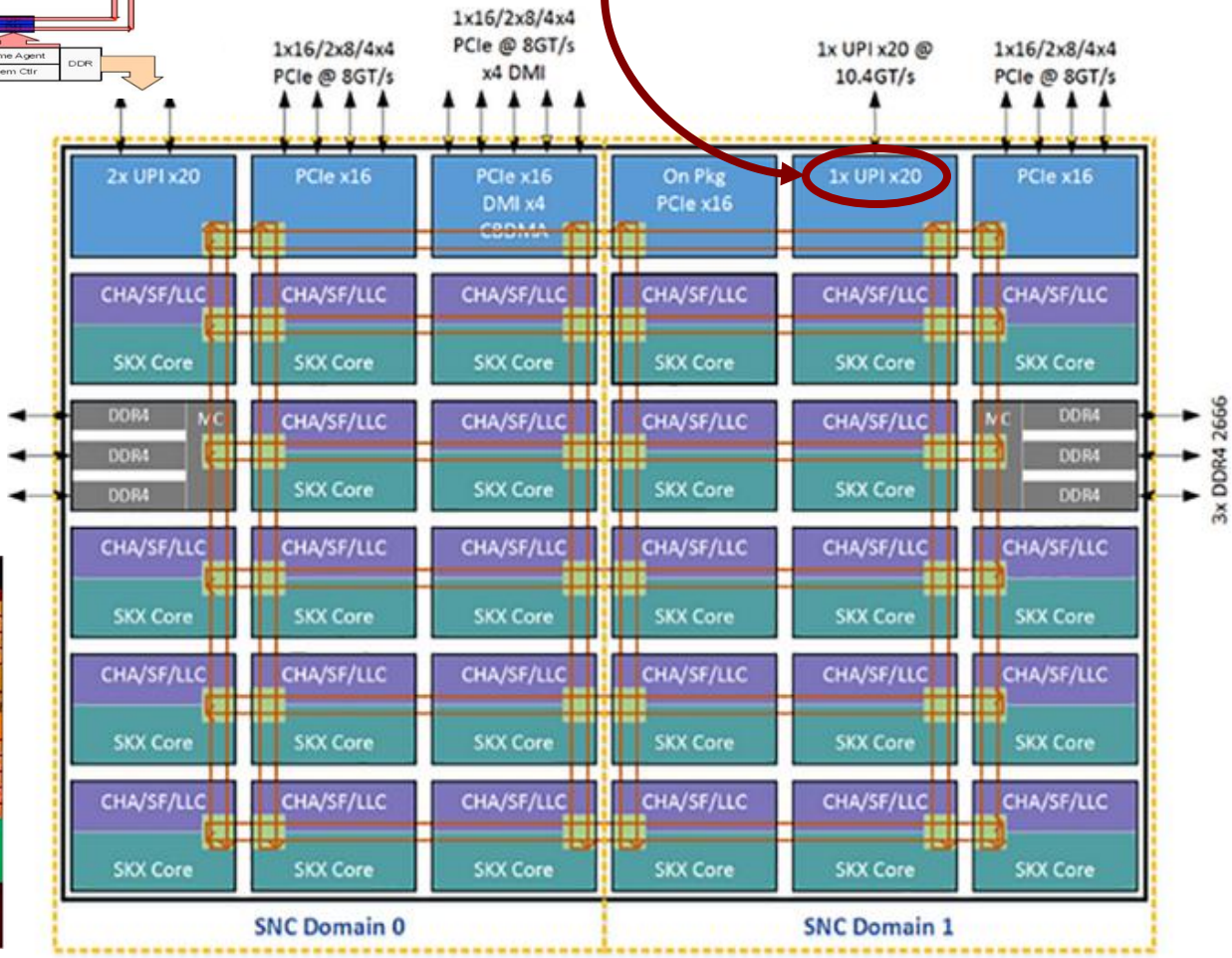


UPI required for dual-socket
*(**U**ltra **P**ath **I**nterconnect)*

Broadwell

ring interconnection does not scale for large #cores

Skylake (server)
(mesh follows KNL)



https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

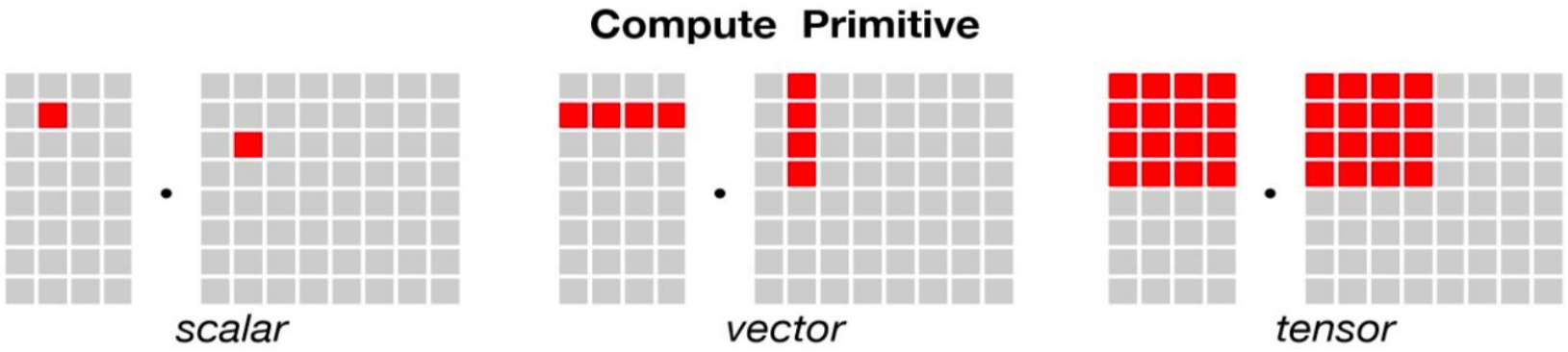
Compute accelerators

Current compute accelerators:

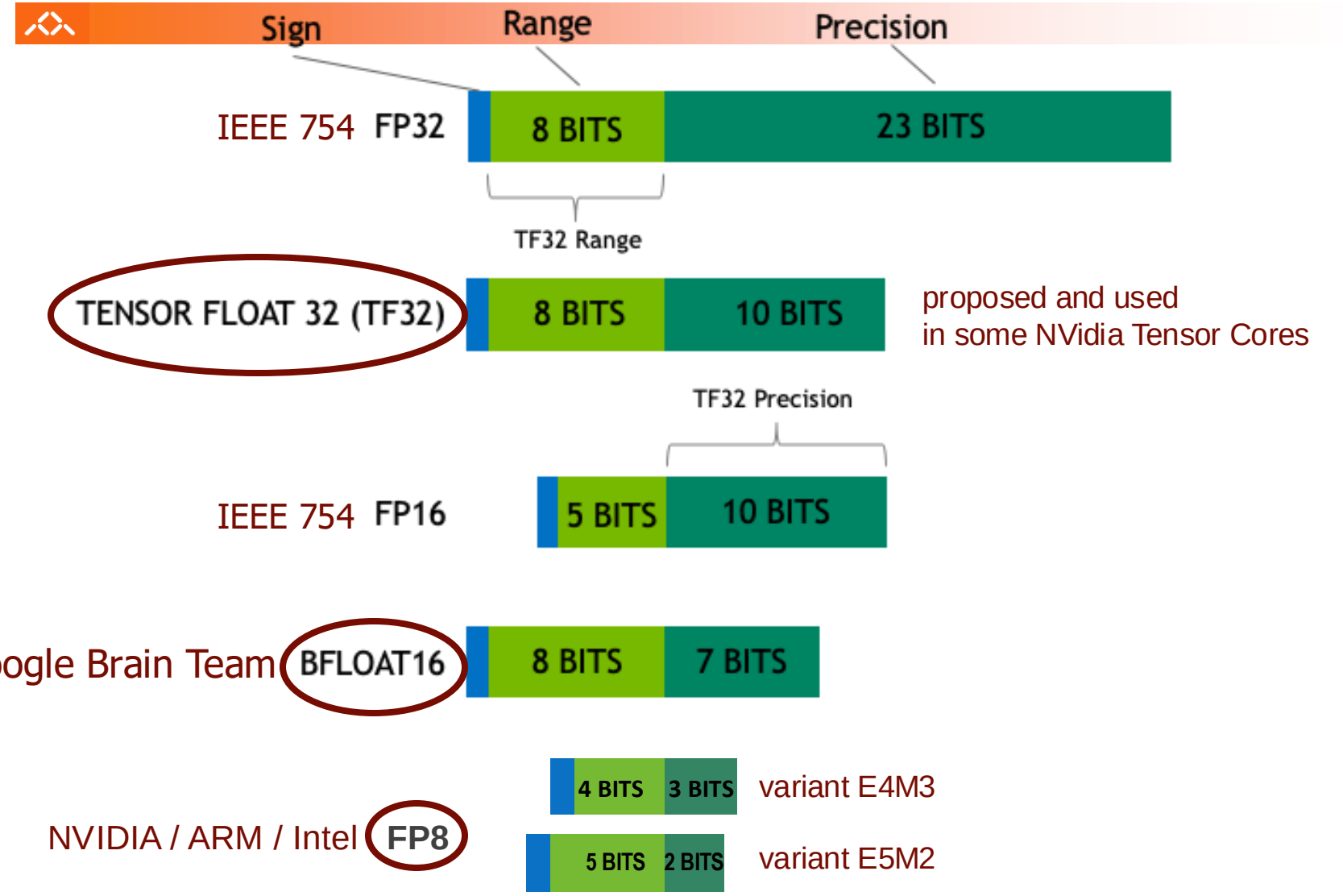
- GPU: Graphical Processor Unit, best accelerator for number crunching, namely intensive vector/matrix computing for scientific applications and AI apps based on machine learning with neural nets
- DSP: Digital Signal Processor, mostly used in telecommunication equipments, from cell phones to radio systems and TVs
- TPU: Tensor Processing Units, optimized for operations with tensors (vector and n-dimensional matrices), popular in AI app's, namely in autonomous driving
- FPGA: Field Programmable Gate Arrays, reconfigurable h/w; can be configured in runtime to behave according to a given spec

Why GPUs are so powerfull as accelerators?

- GPUs in HPC systems have a very large number of FP execution units and are excellent computing accelerators for scientific computing
- Their impact was clear w/ China's Tianhe: **1st in TOP500 List** (Nov'2010)
- Some current figures:
  - A leading AMD x86 CPU (Bergamo) has 128 cores, each can simultaneously perform eight 32-bit FP calculations, i.e., **1,024 32-bit FP execution units** per chip
  - A leading Nvidia's GPU (Hopper, H100) has **18,432 32-bit FP execution units**
- Current GPU devices (e.g., H100) have better support for matrix op's (H100 has 528 4th generation Tensor Cores) and for AI FP-formats



Novel Floating-Point formats for AI



Analysis of PUs and Compute accelerators in HPC: key features that impact performance

Key features in the architecture of a current PU device

- #physical cores and #parallel-multithreads *(SMT)* at each core
- how cores are grouped the chip/package and which fabrication process
- support for vector extensions (128, 256 or 512 bits) and tensor operations
- support for AI app's (including small formats for `int` and `fp` numbers)
- size of L1, L2, L3 caches, #memory channels and which DDRx support
- size of on-chip (or in-package) embedded HBM *(High-Bandwidth Memory)*
- how the PU interconnects with compute accelerators

Key features in the architecture of a current GPU device

- #streaming multiprocessors (SMs), or #GPCs *(Graphics Processing Clusters)*
- #FP-32 cores (NVidia CUDA-cores or AMD Streaming Processors)
- #tensor cores (NVidia) or #matrix cores (AMD)
- size of caches
- size of on-chip (or in-package) embedded HBM

A sample of current compute devices in HPC: manycore and compute accelerators

- Latest generations of manycore PUs used in HPC:
  - Intel Xeon devices: from Cascade Lake to Granite Rapids
  - AMD Epyc devices: the Zen family, from Naples to Turin
  - a wafer-scale device: CEREBRAS (3rd generation)
  - missing: ARM-based devices, Chinese devices, …
- Compute accelerator devices *(GPUs)* used in HPC :
  - NVidia: from Volta architecture to Blackwell
  - missing accelerators: Intel, AMD, …
- APU devices (*Accelerated Processing Units, PU+GPU*) in HPC:
  - NVidia Superchips: from Grace+Hopper to Grace+Blackwell
  - AMD Instinct MI300A
- Popular SoC *(System-on-Chip)* devices: Apple M4 and A17

2nd Gen Intel Xeon Scalable Family: Cascade Lake *(partial view of core architecture)*

Register Alias Table (RAT)
2x4 µOP
µOP µOP µOP µOP µOP µOP
Branch Order Buffer (BOB) (48-entry)
Load
FP
Int Vect.
Int
Store
Common Data Buses (CDBs)
Move Elimination
Rename / Allocate / Retirement
ReOrder Buffer (224 entries)
Ones Idioms
Zeroing Idioms
µOP µOP µOP µOP µOP µOP µOP µOP
Integer Physical Register File
(180 Registers)
Scheduler
Unified Reservation Station (RS)
(97 entries)
Vector Physical Register File
(168 Registers)
Port 0 Port 1 Port 5 Port 6 Port 2 Port 3 Port 4 Port 7
µOP µOP µOP µOP µOP µOP µOP µOP
INT ALU
INT DIV
INT ALU
INT MUL
INT ALU
LEA
INT ALU
Branch
AGU
Load Data
AGU
Load Data
Store Data
AGU
INT Vect ALU
INT Vect MUL
FP FMA
INT Vect ALU
INT Vect MUL
FP FMA
INT Vect ALU
INT Vect MUL
FP FMA
INT Vect ALU
INT Vect MUL
FP FMA
512b fused
512b (zmm only)
(optional)
512bit/cycle
AES
Vect String
FP DIV
Branch
Bit Scan
EUs
Execution Engine
64B/cycle
Unified STLB
L2 Cache
1 MiB 16-Way
64B/cycle
To L3
Store Buffer & Forwarding
(56 entries)
64B/cycle
64B/cycle
Load Buffer
(72 entries)
64B/cycle
64B/cycle
L1 Data Cache
32KiB 8-Way
Data TLB
64B/cycle
Line Fill Buffers (LFB)
(10 entries)
Memory Subsystem

https://en.wikichip.org/wiki/intel/microarchitectures/cascade_lake

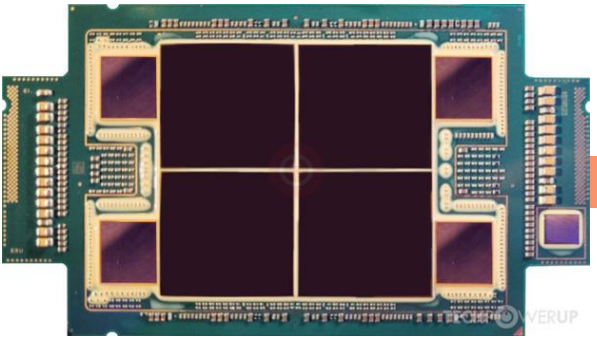
3rd Gen Intel Xeon Scalable Family: Ice Lake *(launched April 2021)*

https://www.anandtech.com/show/16594/intel-3rd-gen-xeon-scalable-review

Key notes:

- **up to 40 cores and 60 MiB L3** (*platinium*)
- **no bronze chips**
- **connects up 2 devices**
- **yet with 10 nm process**

Intel Xeon Comparison: 3rd Gen vs 2nd Gen Peak vs Peak		
Xeon Platinum 8380	AnandTech	Xeon Platinum 8280
40 / 80	**Cores / Threads**	28 / 56
2900 / 3400 / 3000	**Base / ST / MT Freq**	2700 / 4000 / 3300
50 MB + 60 MB	**L2 + L3 Cache**	28 MB + 38.5 MB
270 W	**TDP**	205 W
PCIe 4.0 x64	**PCIe**	PCIe 3.0 x48
8 x DDR4-3200	**DRAM Support**	6 x DDR4-2933
4 TB	**DRAM Capacity**	1 TB

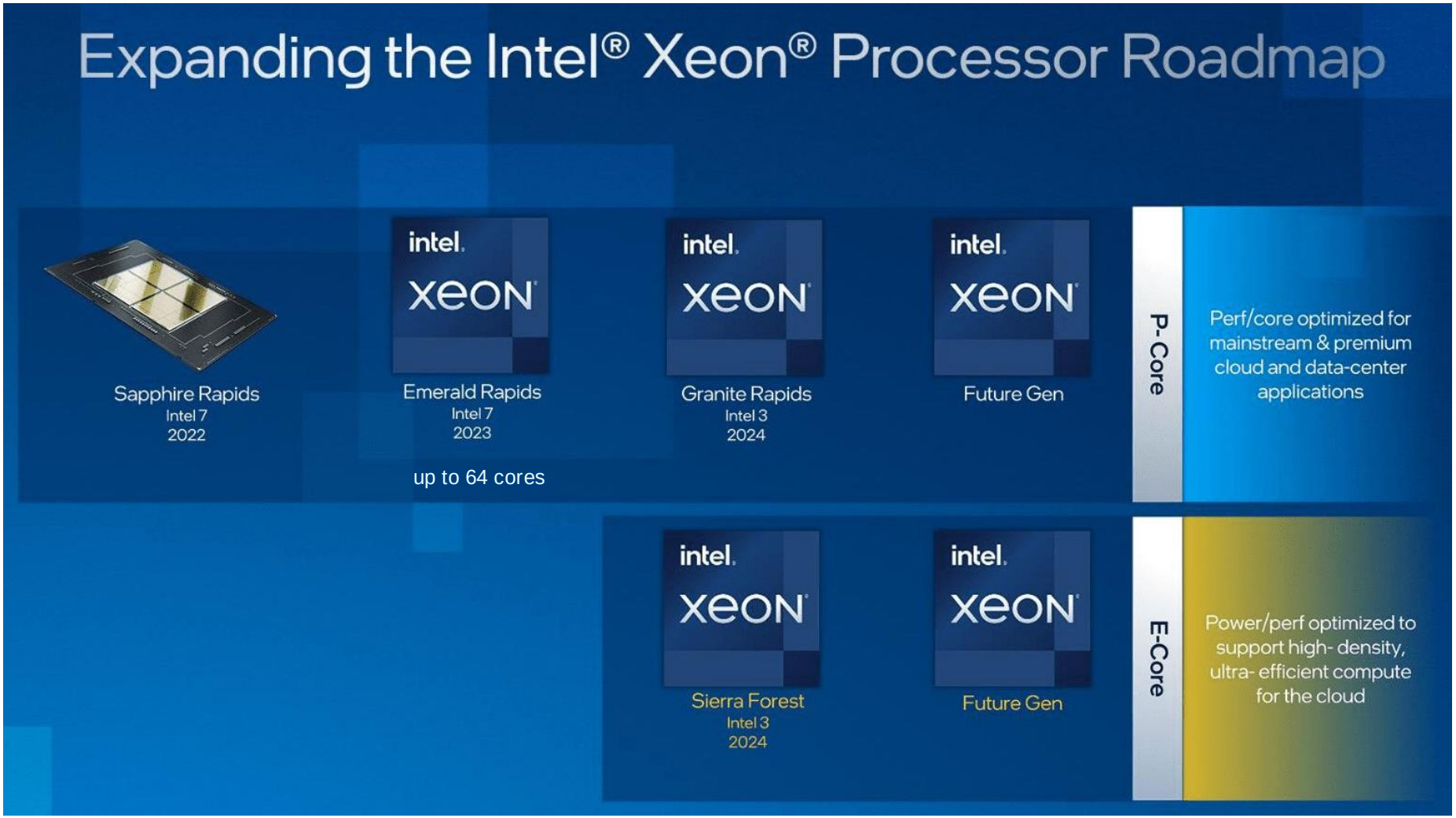


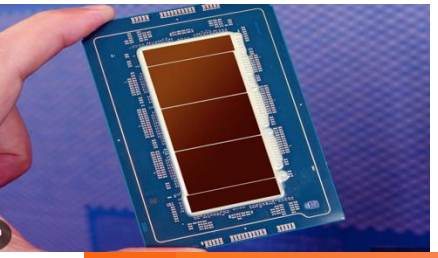
4th Gen Intel Xeon Max Family: Saphire Rapids HBM (Q1'23)

Xeon Max Family => device with HBM

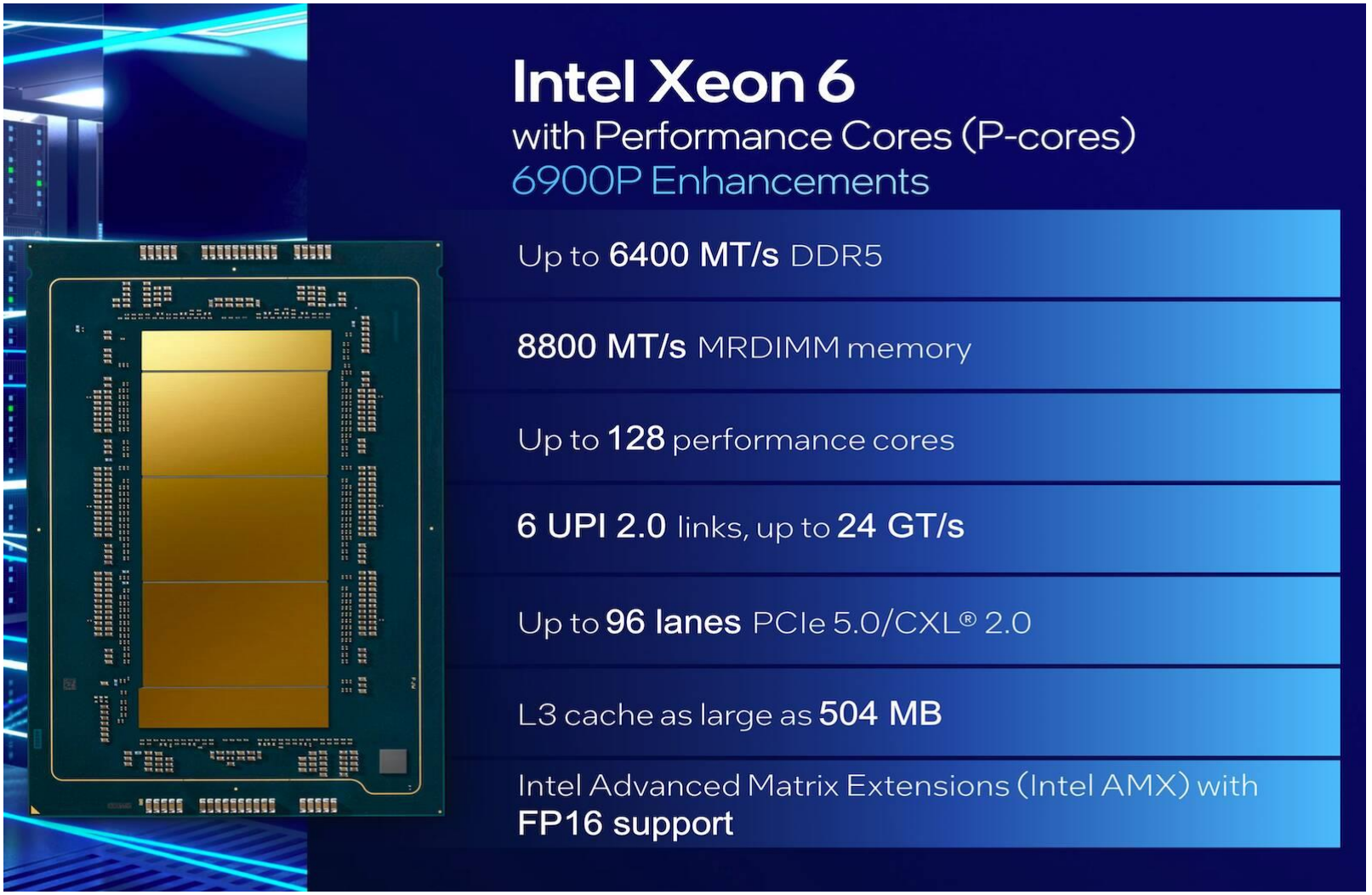
52 cores
AVX-512
AVX-VNNI
AMX
up to 64 GiB HBM

5th Gen Intel Xeon (no-Max) Family: Emerald Rapids HBM *(Q4'23)*

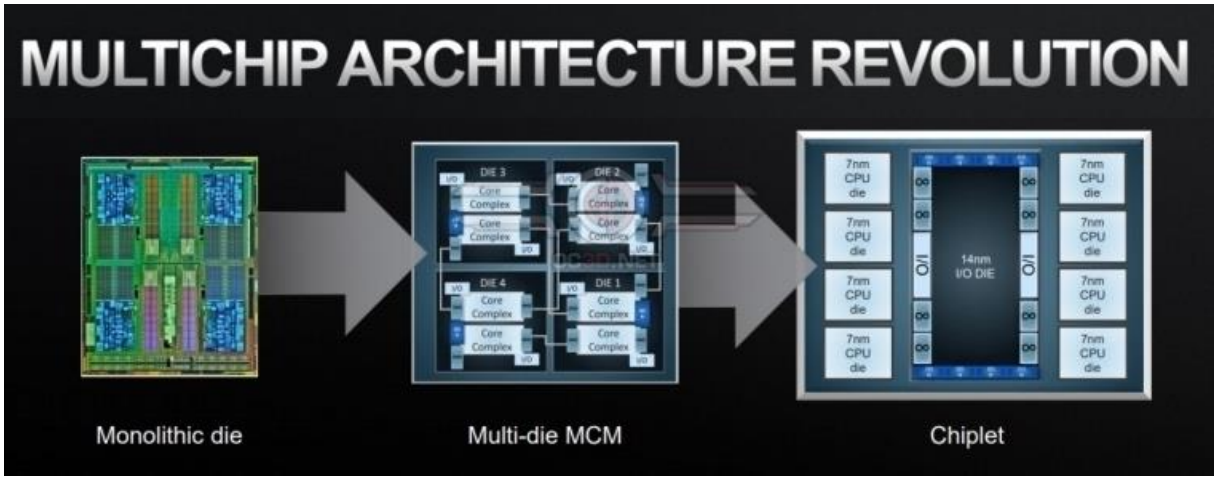




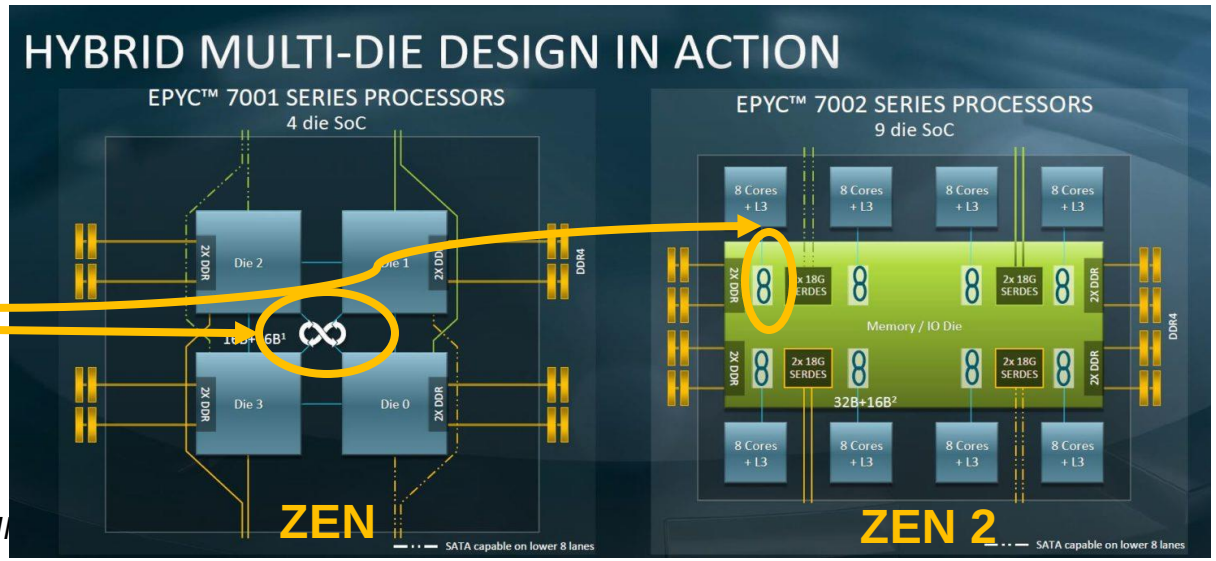
6th Gen Intel Xeon Max Family: Granite Rapids *(Q4'24?)*



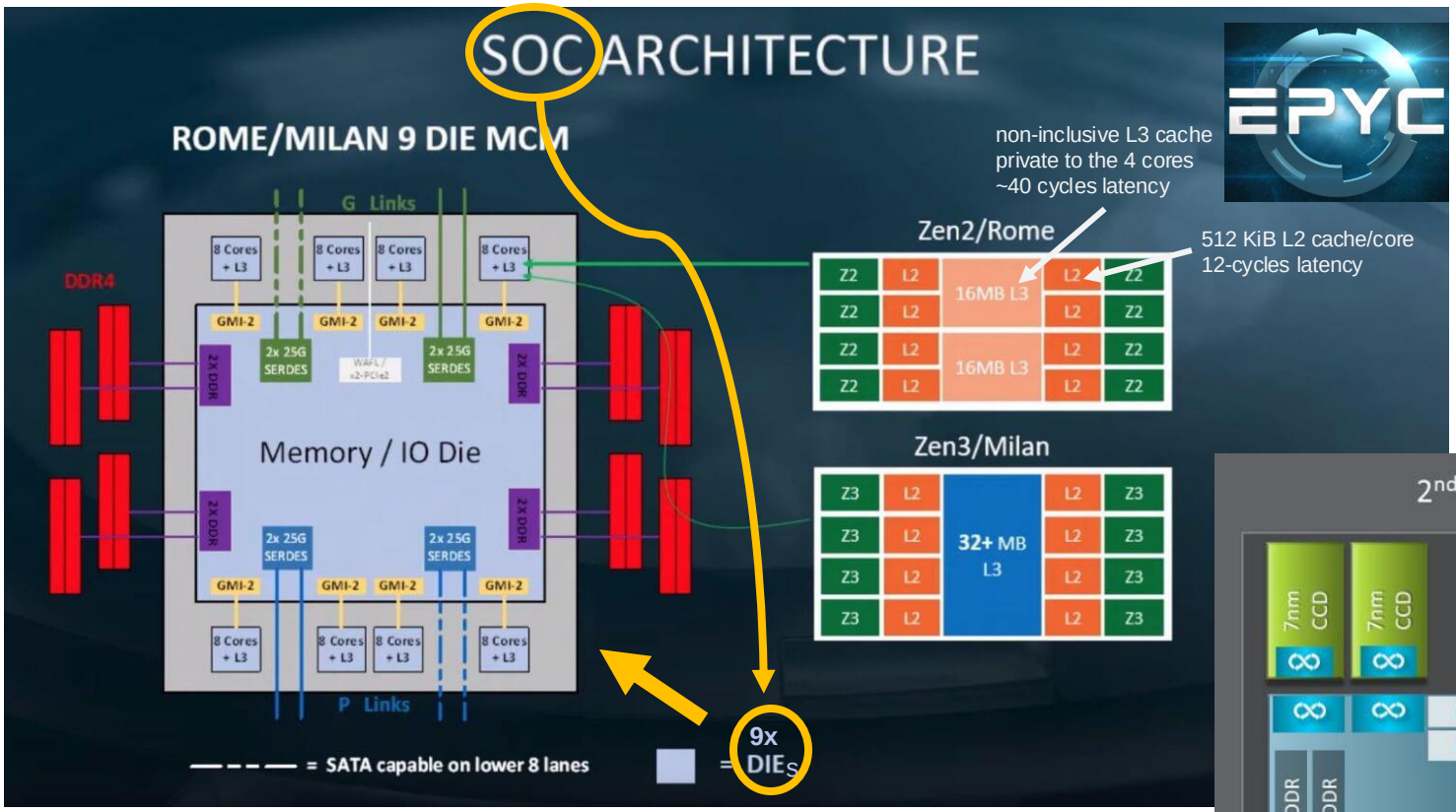
Key Intel Xeon competitor: AMD Epyc (Zen, Zen 2, 3, 4, 5)

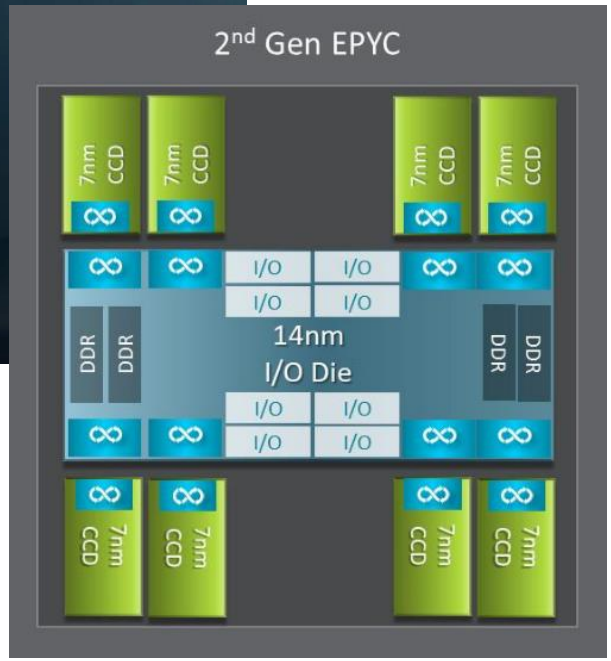


Infinity:
AMD interconnection fabric,
a superset of HyperTransport



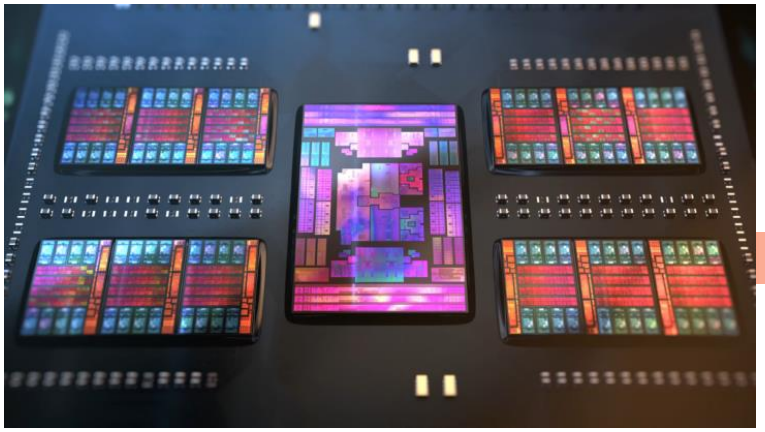
AMD Epyc: from Zen 2 (Rome) to Zen 3 (Milan)





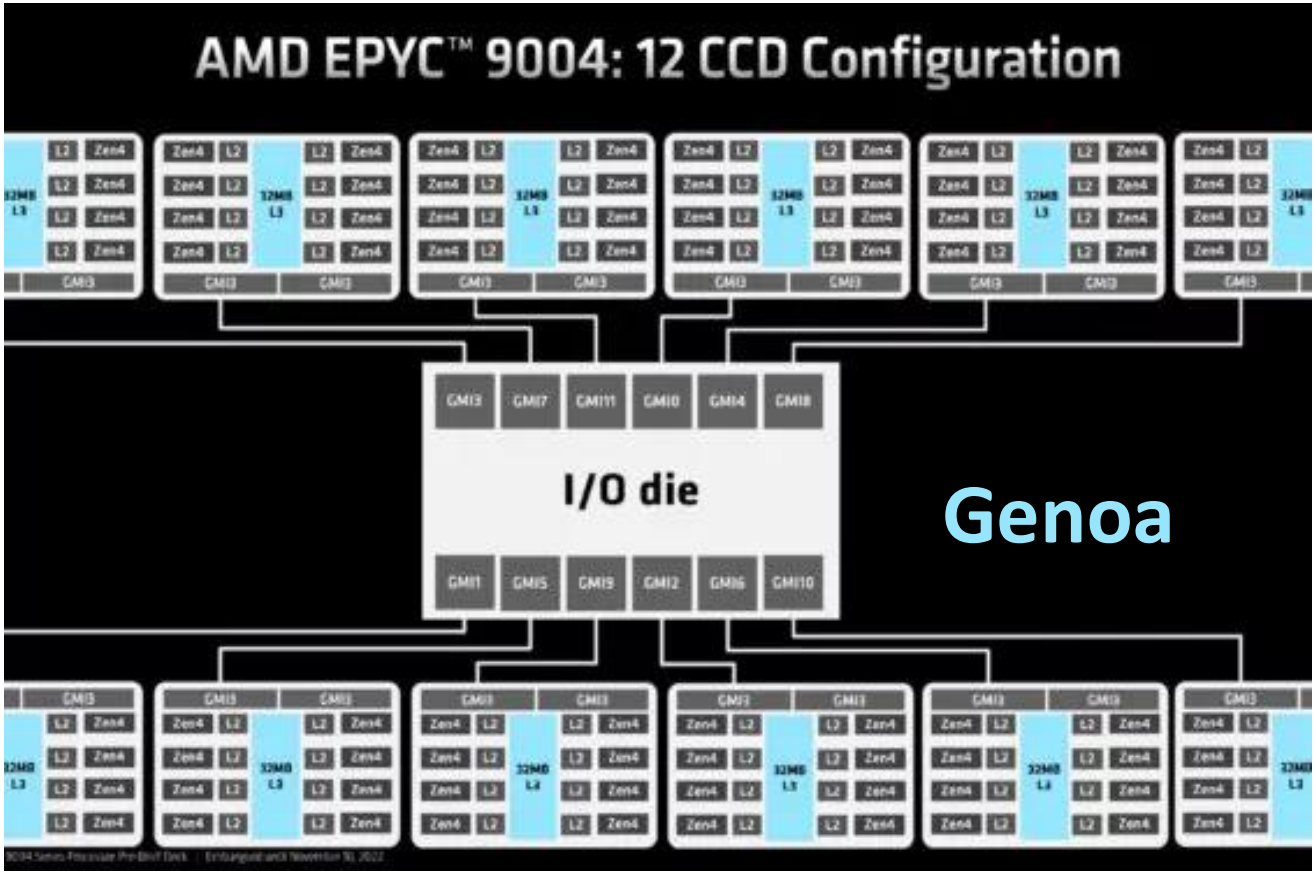
8x 8-core CCD dies at 7 nm
1x I/O die at 14 nm

https://wccftech.com/amd-epyc-rome-7nm-server-cpu-official-launch-64-core-128-thread-128-pcie-gen4/



AMD 4th Gen Epyc: Zen 4 (Genoa & Bergamo)

Genoa: up to **5nm 96 cores** /192 threads
BFLOAT16, VNNU, AVX-512
1 MiB/core L2 cache
384 MiB L3 cache
12 **DDR5** mem chan
128 PCIe Gen 5.0
(launched Nov'22)

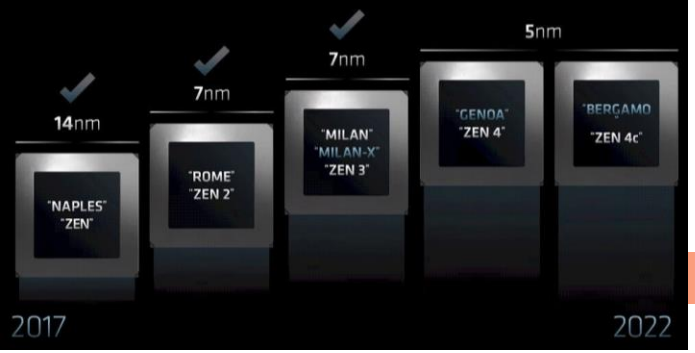


Bergamo:

- up to **128 Zen 4c** cores
- for cloud-native computing

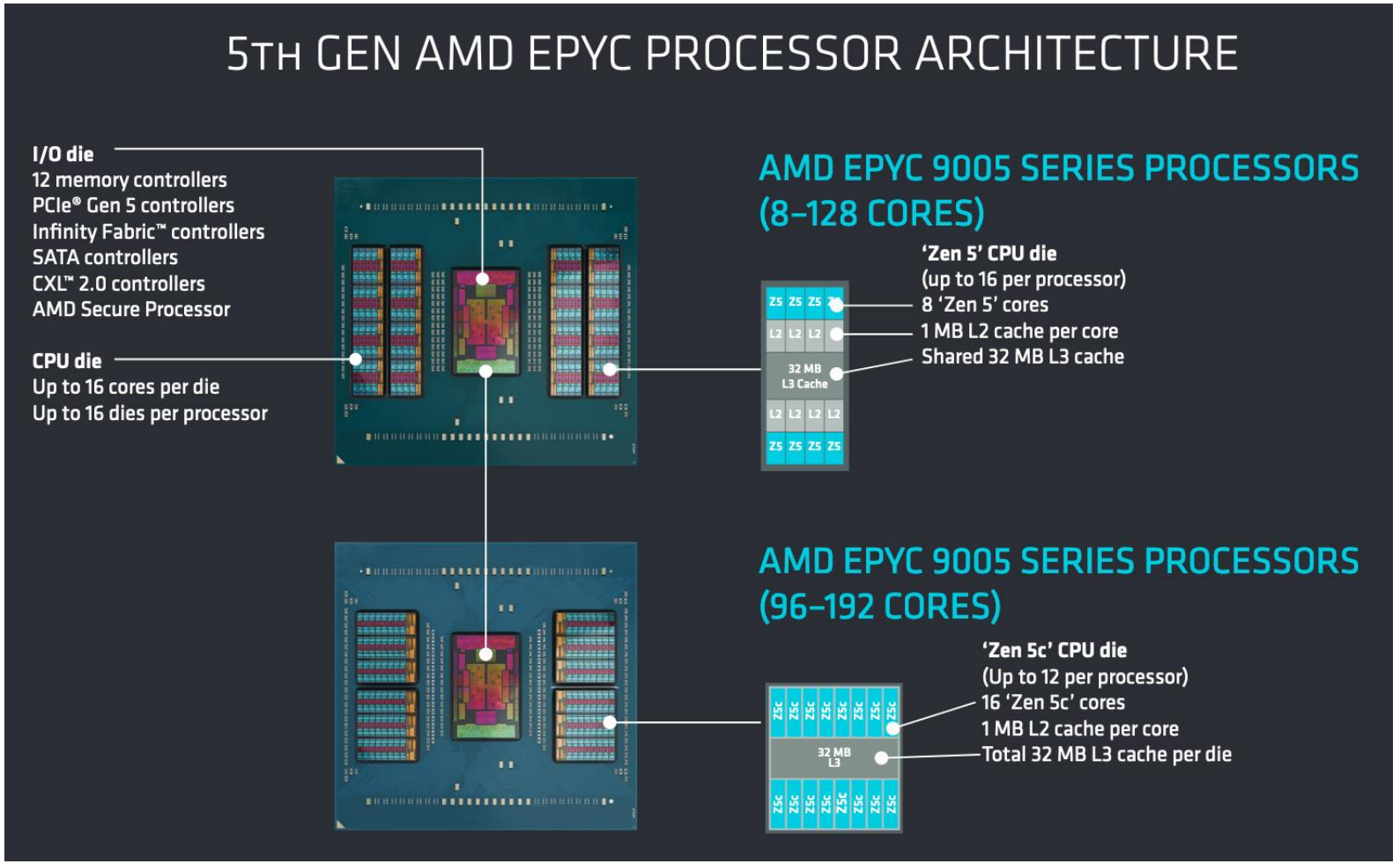
(launched Jun'23)

https://wccftech.com/amd-4th-gen-epyc-genoa-zen-4-cpus-official-launch-96-cores-192-threads-worlds-fastest-server-chips/



AMD 5th Gen Epyc: Zen 5 (Turin)

(launched Oct'24)



https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/5th-gen-amd-epyc-processor-architecture-white-paper.pdf

Wafer-scale devices

billion (10^{12} in Continental Europe)
4 trillion transistors

Keyword!

Cerebras CS-3: the world's largest AI chips

cerebras

Product Solutions Developers Company Resources Contact Us

The future of AI is Wafer-Scale

Cerebras' third-generation wafer-scale engine (WSE-3) is the fastest AI processor on Earth. It surpasses all other processors in AI-optimized cores, memory speed, and on-chip fabric bandwidth.

900,000
AI-Optimized Cores
123x more cores

44GB
On-Chip SRAM
1,000x more on-chip memory

214Pb/s
Interconnect Bandwidth
45,000x more bandwidth

21PB/s
Memory Bandwidth
12,800x more bandwidth

Wafer-Scale Engine (WSE) designed specifically for machine learning and AI applications (FP32)

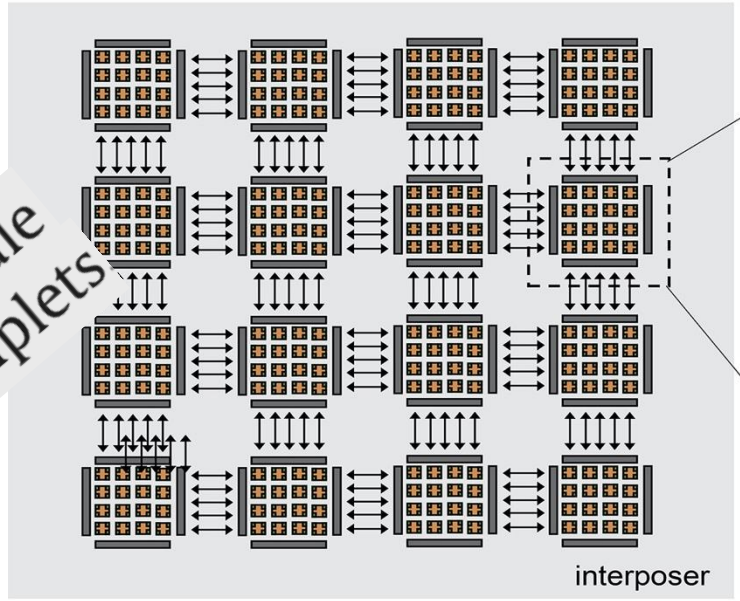
What's going on in China?

The Big Chip: Challenge, Model and Architecture

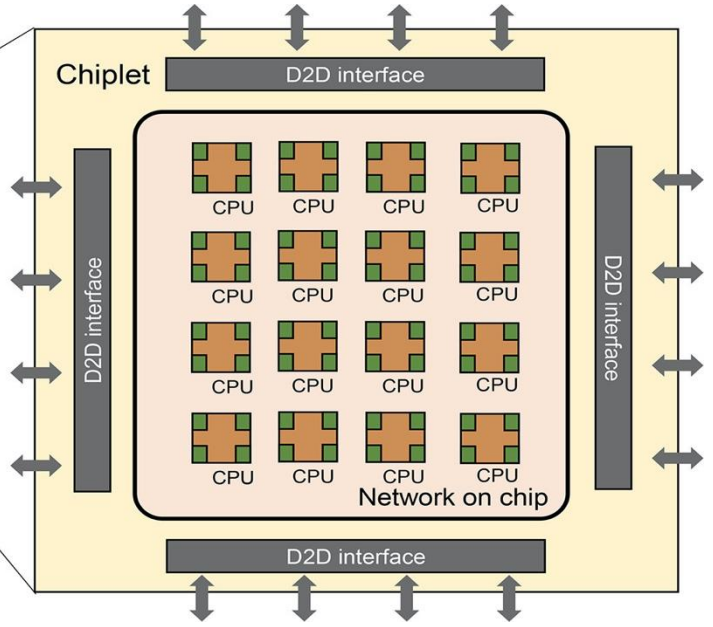
Yinhe Han*†§ Haobo Xu* Meixuan Lu*† Haoran Wang*† Junpei Huang*‡ Ying Wang*†
Yujie Wang*†§ Feng Min* Qi Liu¶ Ming Liu¶ Ninghui Sun*†
*Institute of Computing Technology, Chinese Academy of Sciences
†University of Chinese Academy of Sciences
‡University of Science and Technology of China
§Zhejiang Lab
¶Fudan University

24 October 2023

potential to scale up to 100 chiplets



Zhejiang Big Chip overview



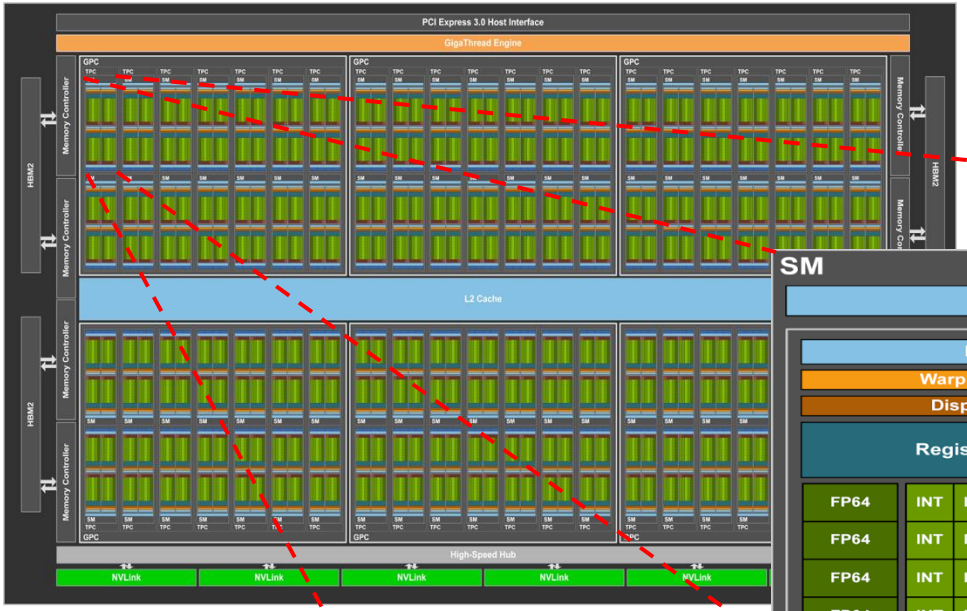
A sample of current compute devices in HPC: manycore and compute accelerators

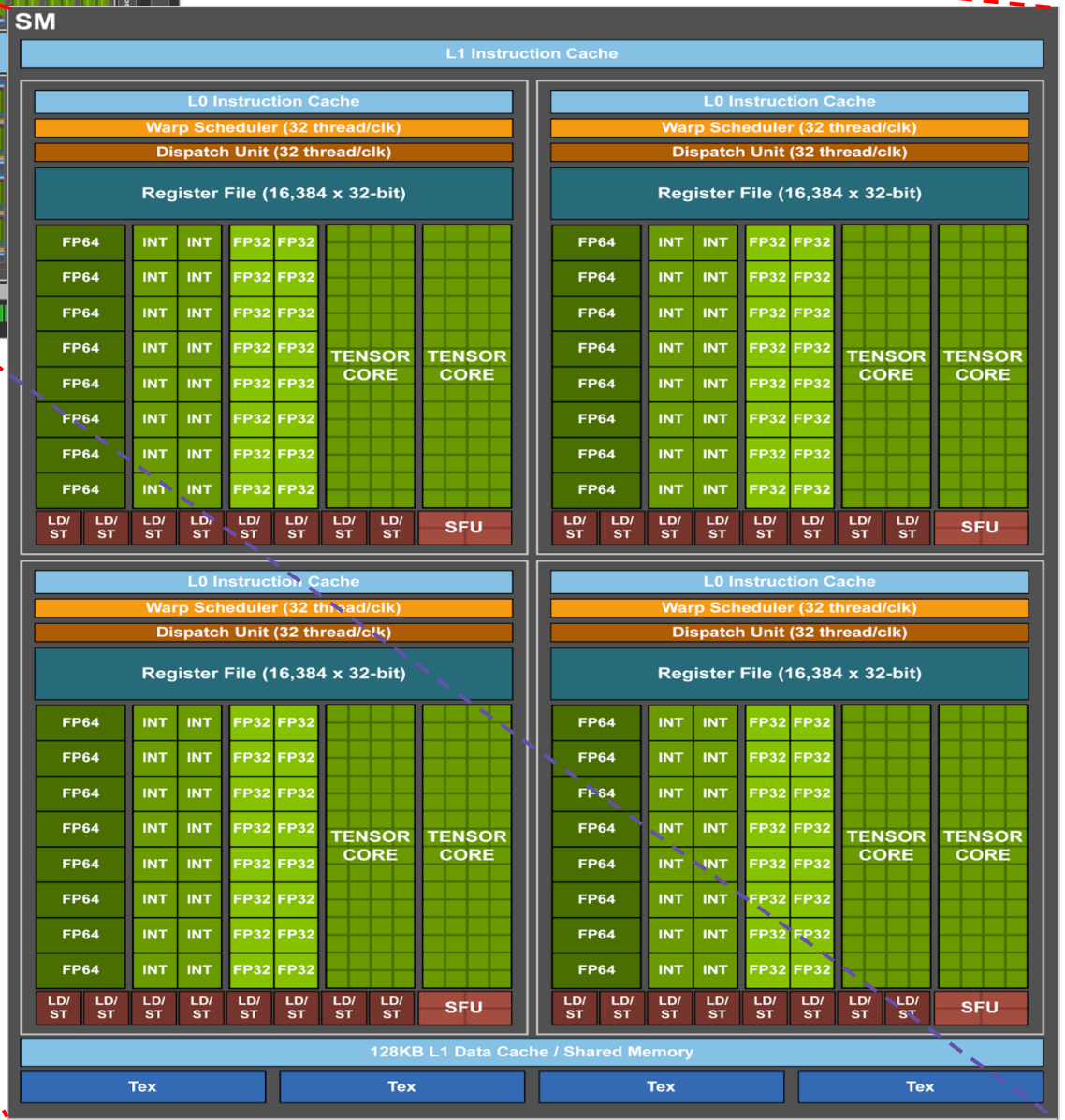
- Latest generations of manycore PUs used in HPC:
  - Intel Xeon devices: from Cascade Lake to Granite Rapids
  - AMD Epyc devices: the Zen family, from Naples to Turin
  - a wafer-scale device: CEREBRAS (3rd generation)
  - missing: ARM-based devices, Chinese devices, …
- Compute accelerator devices *(GPUs)* used in HPC :
  - NVidia: from Volta architecture to Blackwell
  - missing accelerators: Intel, AMD, …
- APU devices (*Accelerated Processing Units, PU+GPU*) in HPC:
  - NVidia Superchips: from Grace+Hopper to Grace+Blackwell
  - AMD Instinct MI300A
- Popular SoC *(System-on-Chip)* devices: Apple M4 and A17

Volta Architecture:

up to 84 SMs and 5 376 FP32-cores

Dec'17





Volta SM:

4x16 (INT+FP32)-cores

New: 4x2 Tensor-cores

Ratio **DP**unit : **SP**unit □> **1 : 2**

Volta V100 w/ **16GiB HBM2**

Ampere Architecture

May'20

up to 128 SMs and 8 192 FP32-cores

Ampere SM:

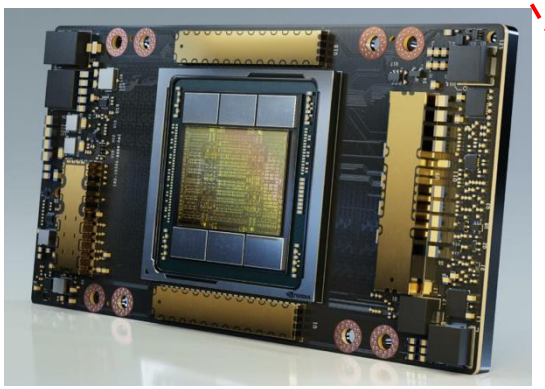
64x FP32 CUDA Cores/SM

32x FP64 CUDA Cores/SM

4x 3rd generation Tensor Cores

Tensor Cores support
FP64, FP32, TF32, FP16, BF16, INT8…

1024 dense FP16/FP32 FMA op's/cycle



Hopper Architecture

Mar'22

up to 132 SMs and 16 896 FP32-cores

Hopper SM:

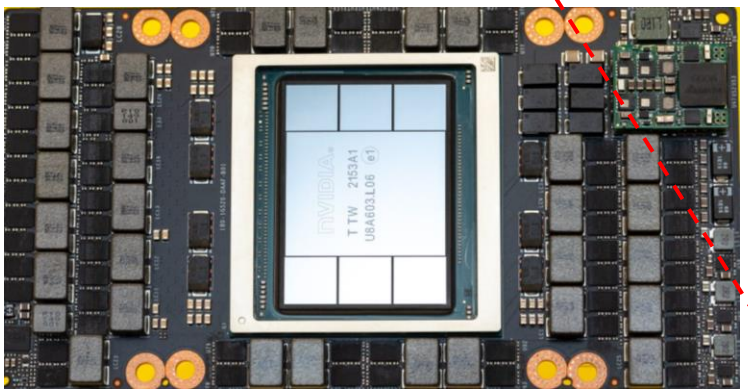
128x FP32 CUDA Cores/SM

64x FP64 CUDA Cores/SM

4x 4th generation Tensor Cores

Tensor Cores support (2x faster)
FP64, FP32, TF32, FP16, BF16, **FP8,** INT8…

1024 dense FP16/FP32 FMA op's/cycle



Key innovation in Hopper Architecture: 4th gen Tensor Core with Transformer Engine

What is a transformer model?

A transformer model is a neural network architecture that can automatically transform one type of input into another type of output. The term was coined in a 2017 Google paper that found a way to train a neural network for translating English to French with more accuracy and a quarter of the training time of other neural networks.

Transformers are crucial in all large language model (LLM) applications, including ChatGPT, Google Search, Dall-E and Microsoft Copilot.

Virtually all applications that use natural language processing now use transformers under the hood

https://www.techtarget.com/searchenterpriseai/definition/transformer-model

New **Transformer Engine** uses a combination of software and custom Hopper Tensor Core technology designed specifically to accelerate Transformer model training and inference. The Transformer Engine intelligently manages and dynamically chooses between FP8 and 16-bit calculations, automatically handling re-casting and scaling

NVIDIA® Transformer Engine is a library for accelerating Transformer models on NVIDIA GPUs, including using 8-bit floating point (FP8) precision on Hopper and Ada GPUs, to

https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

AWS launched new AI acceleration hardware powered by Nvidia H100s

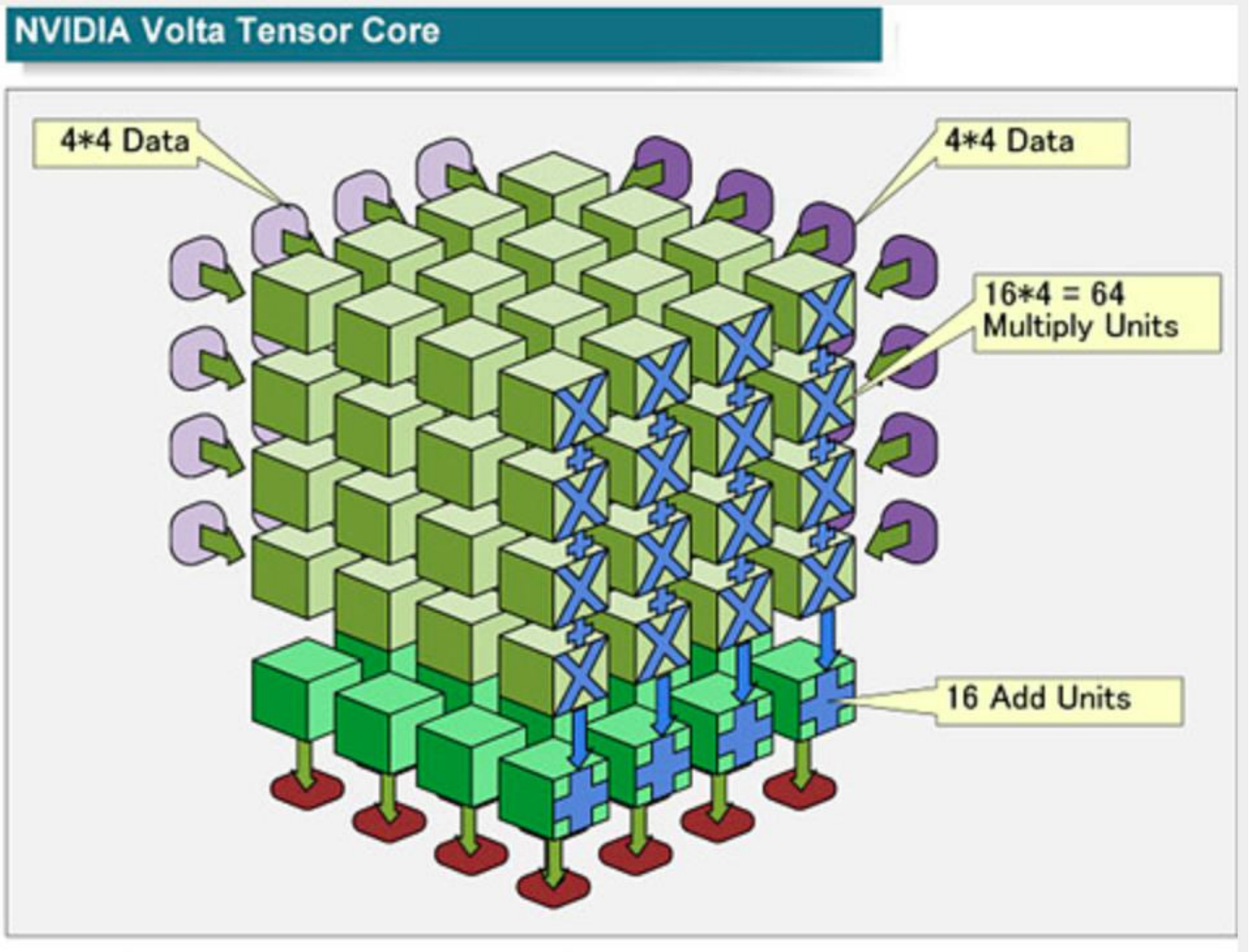
Compute Capability & Precision Support Matrix:
GP100 vs GV100 vs GA100 vs GH100

GPU features	NVIDIA Tesla **P100**	NVIDIA Tesla **V100**	NVIDIA **A100**	NVIDIA **H100**
GPU codename	GP100	GV100	GA100	GH100
GPU architecture	NVIDIA **Pascal**	NVIDIA **Volta**	NVIDIA **Ampere**	NVIDIA **Hopper**
Transistors	15.3 billion	21.1 billion	54.2 billion	80 billion
Process	**16nm**	**12nm**	TSMC **7nm**	TSMC **4nm**
Die size	610 mm^2	828 mm^2	815 mm^2	814 mm^2
Compute capability	6.0	7.0	8.0	9.0
Threads / warp	32	32	32	32

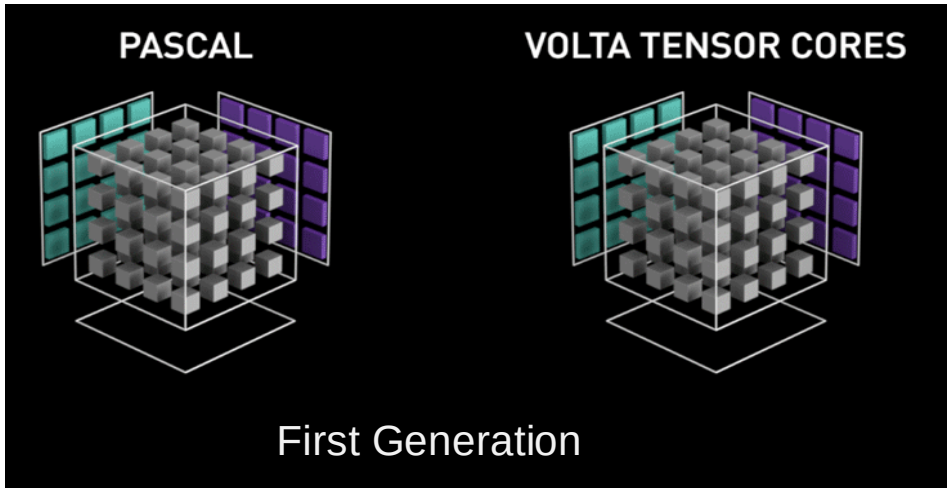
	Supported CUDA Core Precisions									Supported Tensor Core Precisions								
	FP8	FP16	FP32	FP64	INT1	INT4	INT8	TF32	BF16	FP8	FP16	FP32	FP64	INT1	INT4	INT8	TF32	BF16
NVIDIA Tesla P4	No	No	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No
NVIDIA P100	No	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	No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

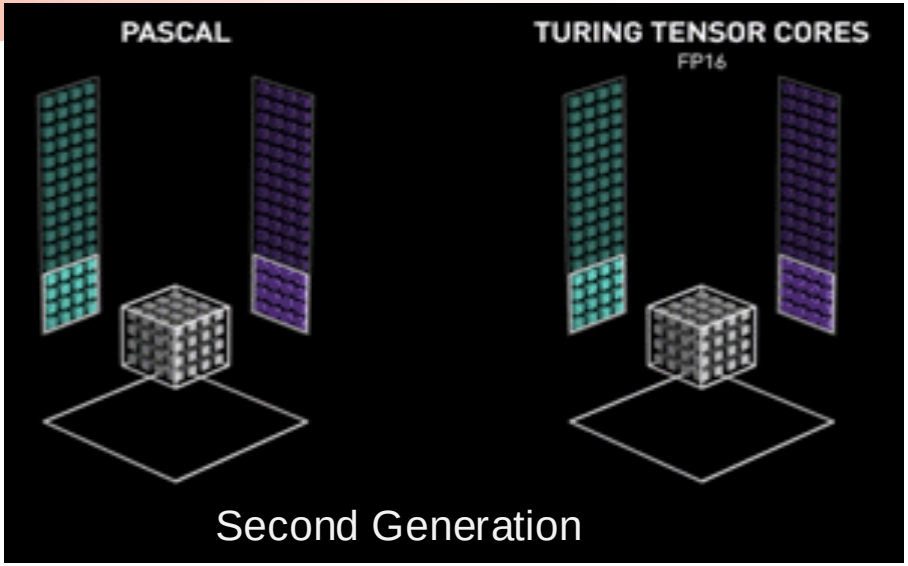
https://en.wikipedia.org/wiki/Hopper_(microarchitecture)

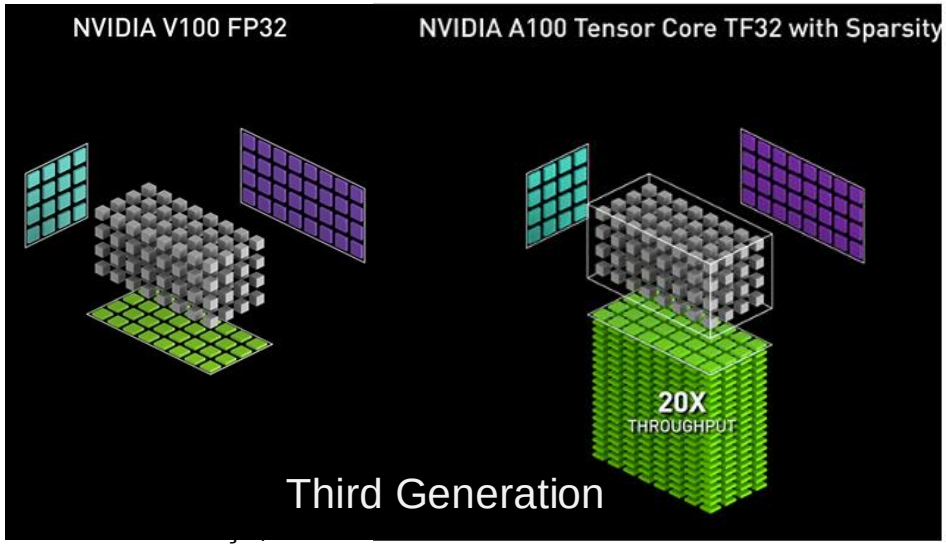
Matrix FMA in a single clock cycle with FP32

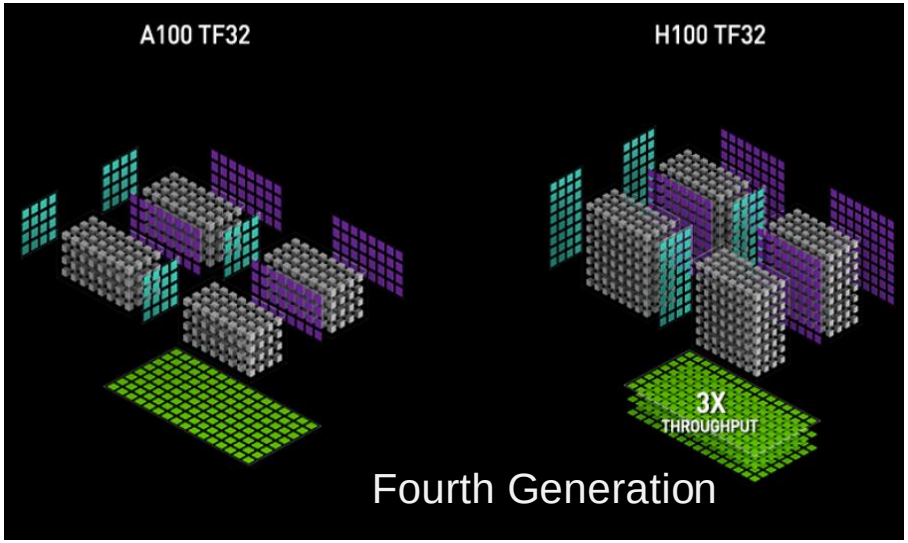


NVidia Tensor cores: 4 generations







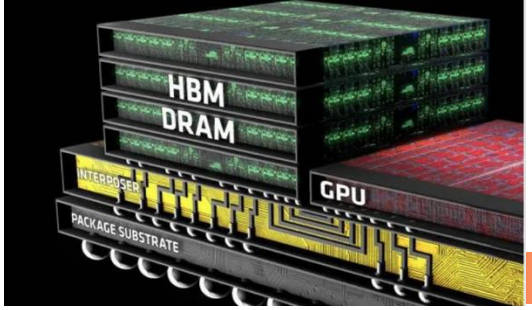


Tesla evolution

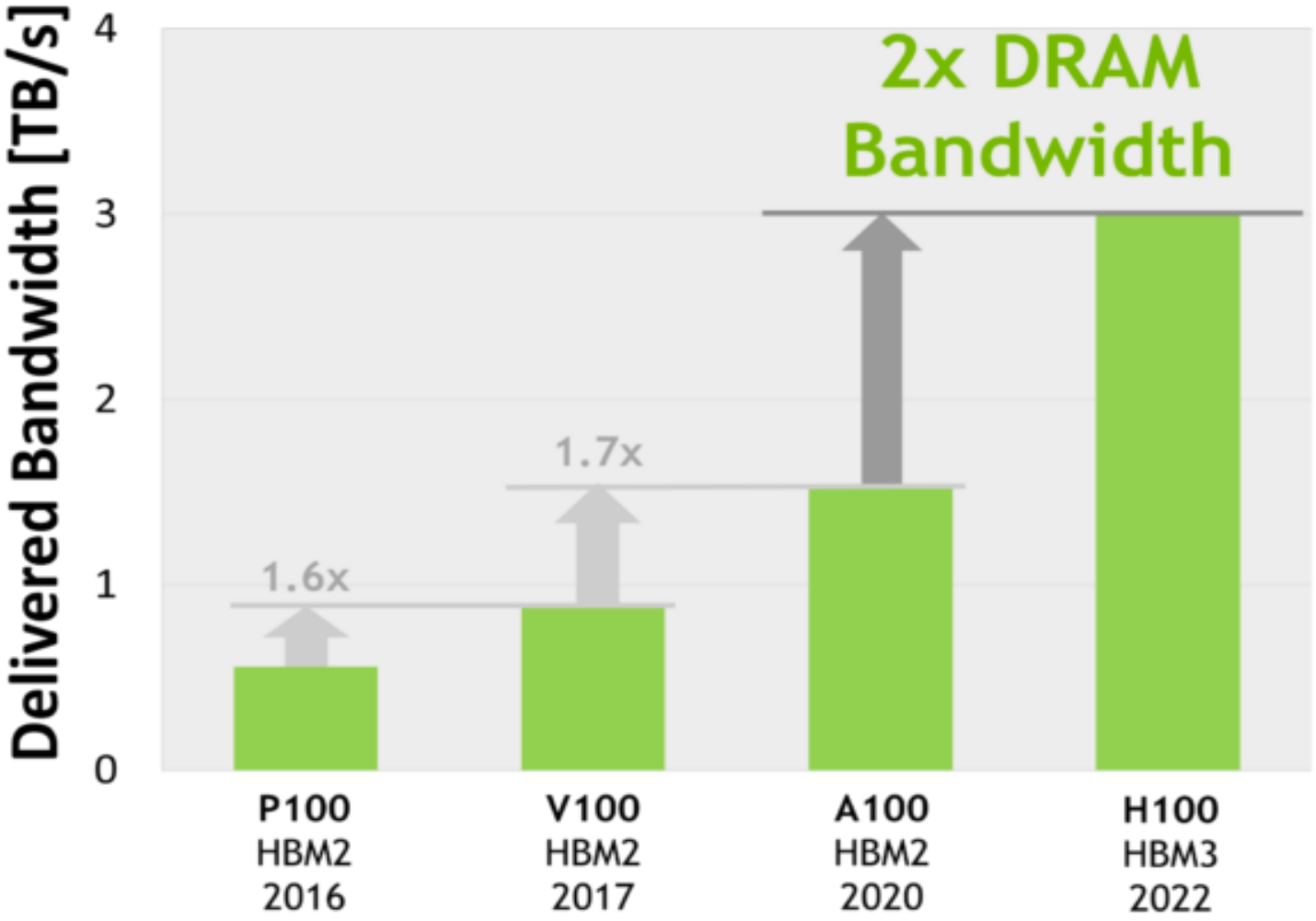
NVIDIA Data-Center GPUs Specifications

VideoCardz.com	NVIDIA H100	NVIDIA A100	NVIDIA Tesla V100	NVIDIA Tesla P100
Picture				
GPU	**GH100**	GA100	GV100	GP100
Transistors	**80B**	54.2B	21.1B	15.3B
Die Size	**814 mm²**	828 mm²	815 mm²	610 mm²
Architecture	**Hopper**	Ampere	Volta	Pascal
Fabrication Node	**TSMC N4**	TSMC N7	12nm FFN	16nm FinFET+
GPU Clusters	**132/114***	108	80	56
CUDA Cores	**16896/14592***	6912	5120	3584
L2 Cache	**50MB**	40MB	6MB	4MB
Tensor Cores	**528/456***	432	320	-
Memory Bus	**5120-bit**	5120-bit	4096-bit	4096-bit
Memory Size	**80 GB HBM3/HBM2e***	40/80GB HBM2e	16/32 HBM2	16GB HBM2
TDP	**700W/350W***	250W/300W/400W	250W/300W/450W	250W/300W
Interface	**SXM5/*PCIe Gen5**	SXM4/PCIe Gen4	SXM2/PCIe Gen3	SXM/PCIe Gen3
Launch Year	**2022**	2020	2017	2016

https://videocardz.com/press-release/nvidia-announces-h100-hopper-gpu-with-up-to-16896-fp32-cores-80gb-hbm3-memory-700w-of-tdp

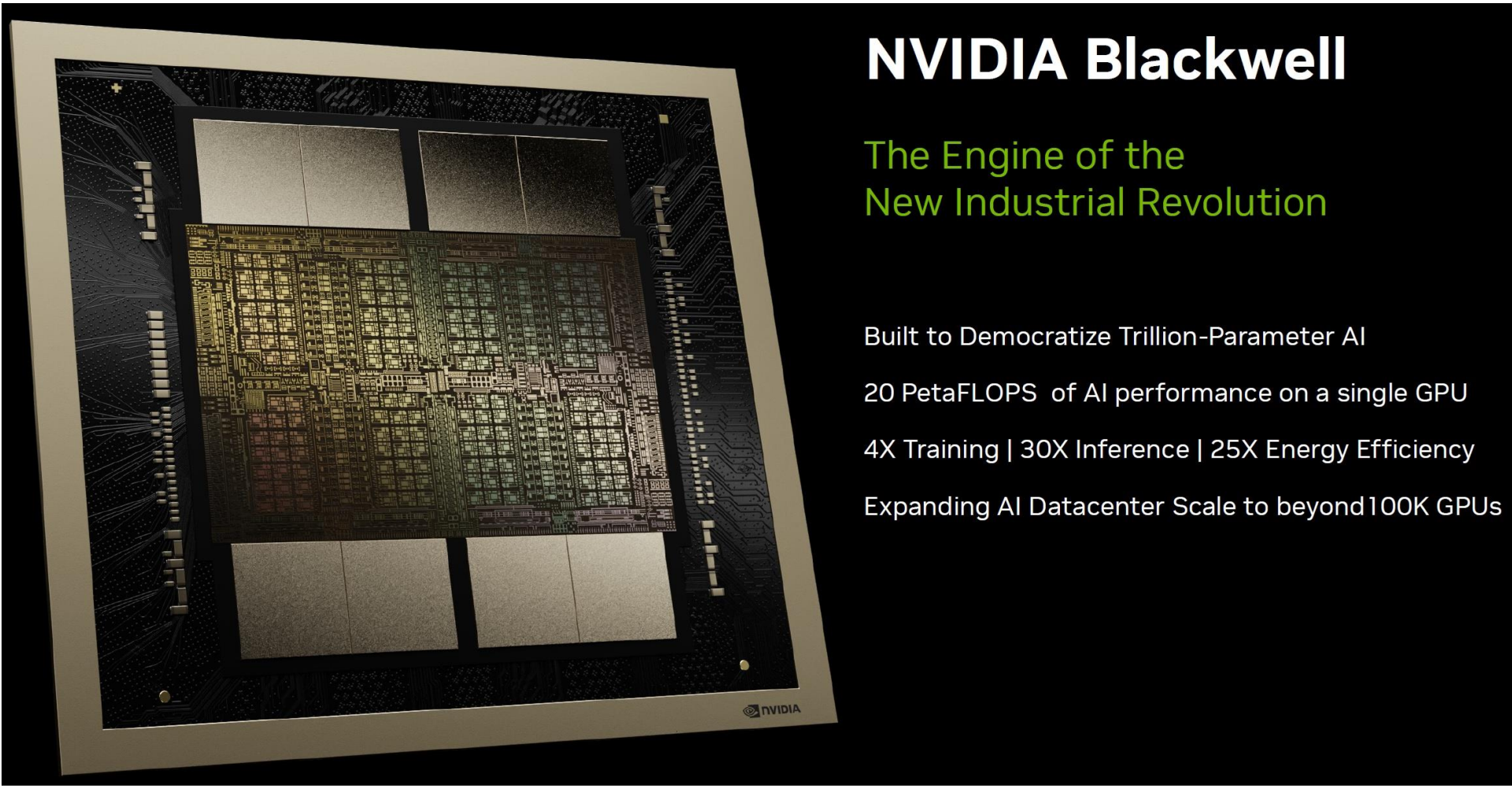


Performance evolution of NVidia stacked HBM



https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/

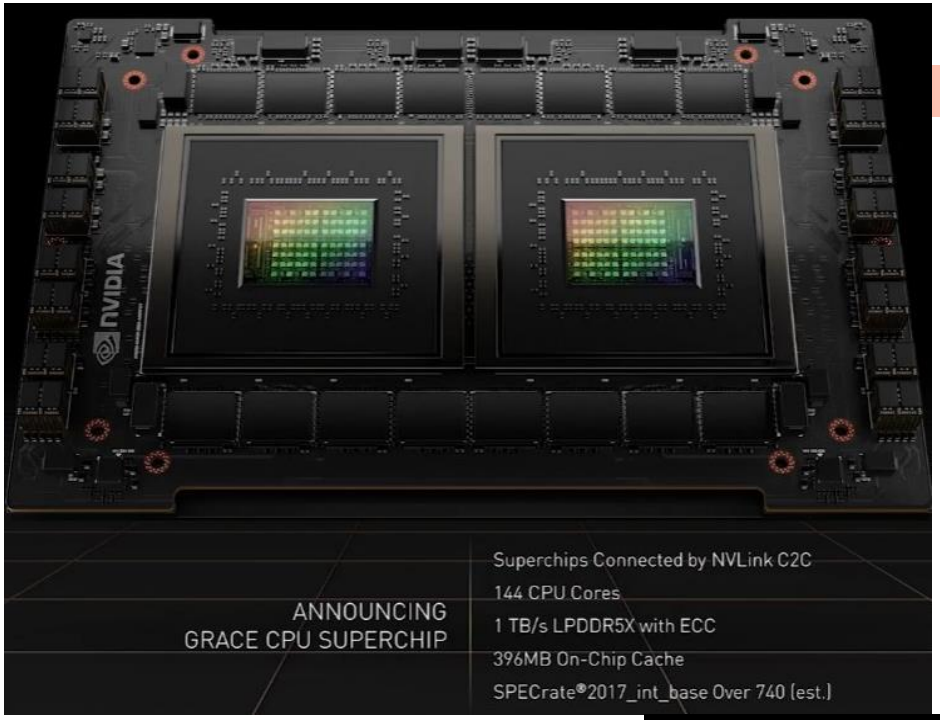
From Hopper to Blackwell

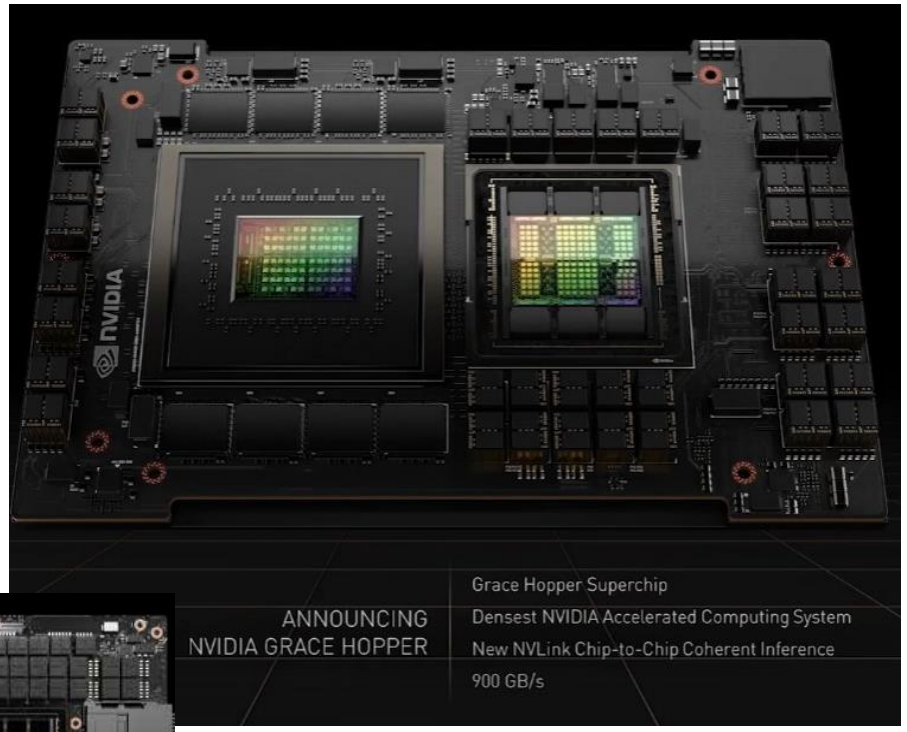


A sample of current compute devices in HPC: manycore and compute accelerators

- Latest generations of manycore PUs used in HPC:
  - Intel Xeon devices: from Cascade Lake to Granite Rapids
  - AMD Epyc devices: the Zen family, from Naples to Turin
  - a wafer-scale device: CEREBRAS (3rd generation)
  - missing: ARM-based devices, Chinese devices, …
- Compute accelerator devices *(GPUs)* used in HPC :
  - NVidia: from Volta architecture to Blackwell
  - missing accelerators: Intel, AMD, …
- APU devices (*Accelerated Processing Units, PU+GPU*) in HPC:
  - NVidia Superchips: from Grace+Hopper to Grace+Blackwell
  - AMD Instinct MI300A
- Popular SoC *(System-on-Chip)* devices: Apple M4 and A17

The 1st NVidia APU: the Grace & Grace-Hopper Superchips





Grace Superchip:

- 2x 72-core Arm v9.0
- 4 nm TSMC
- SVE2 suppport

Grace-Hopper Superchip

https://www.servethehome.com/nvidia-gtc-2022-keynote-coverage-crazy-new-data-center-gear/

NVIDIA GB200 Superchip

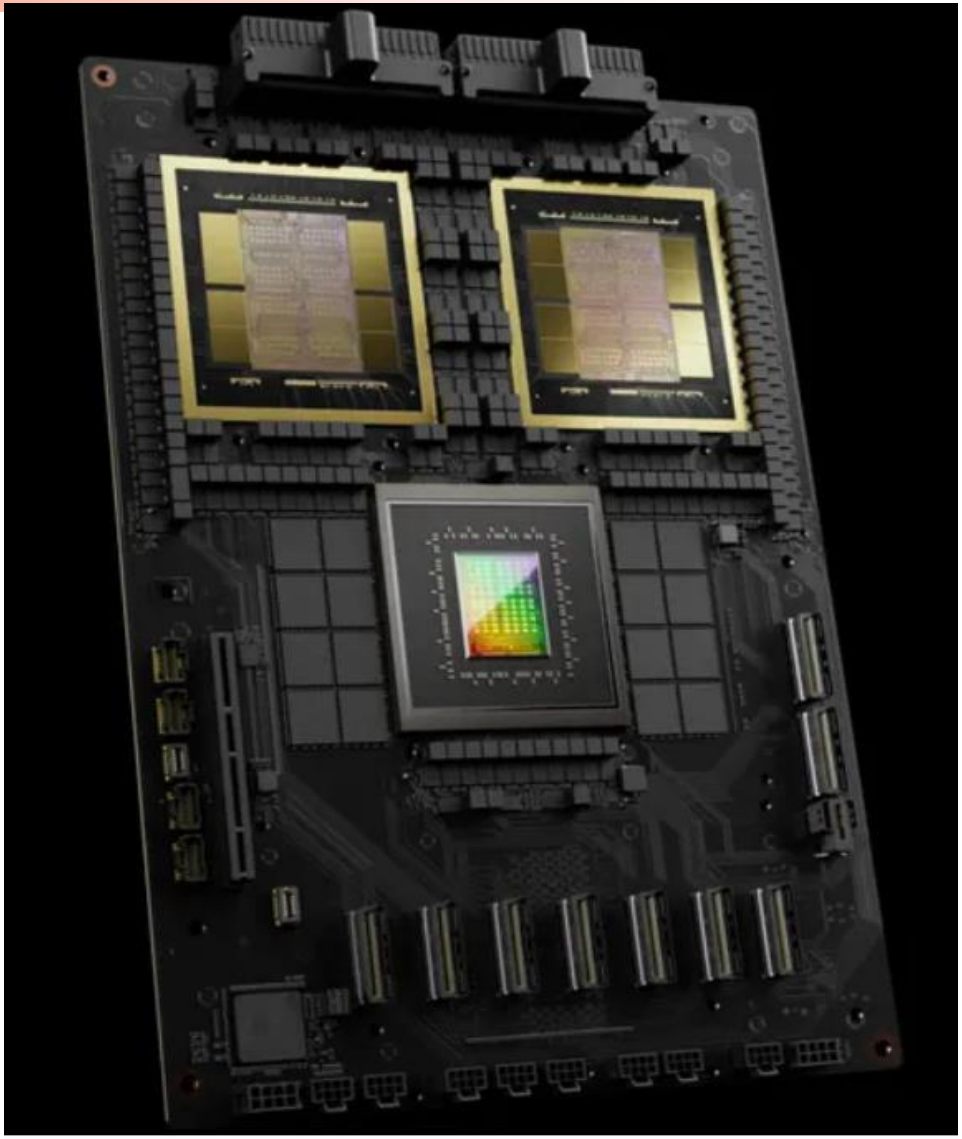
TWO BLACKWELL GPUs AND ONE GRACE CPU

BUILDING BLOCKS OF THE GB200 SUPERCHIP

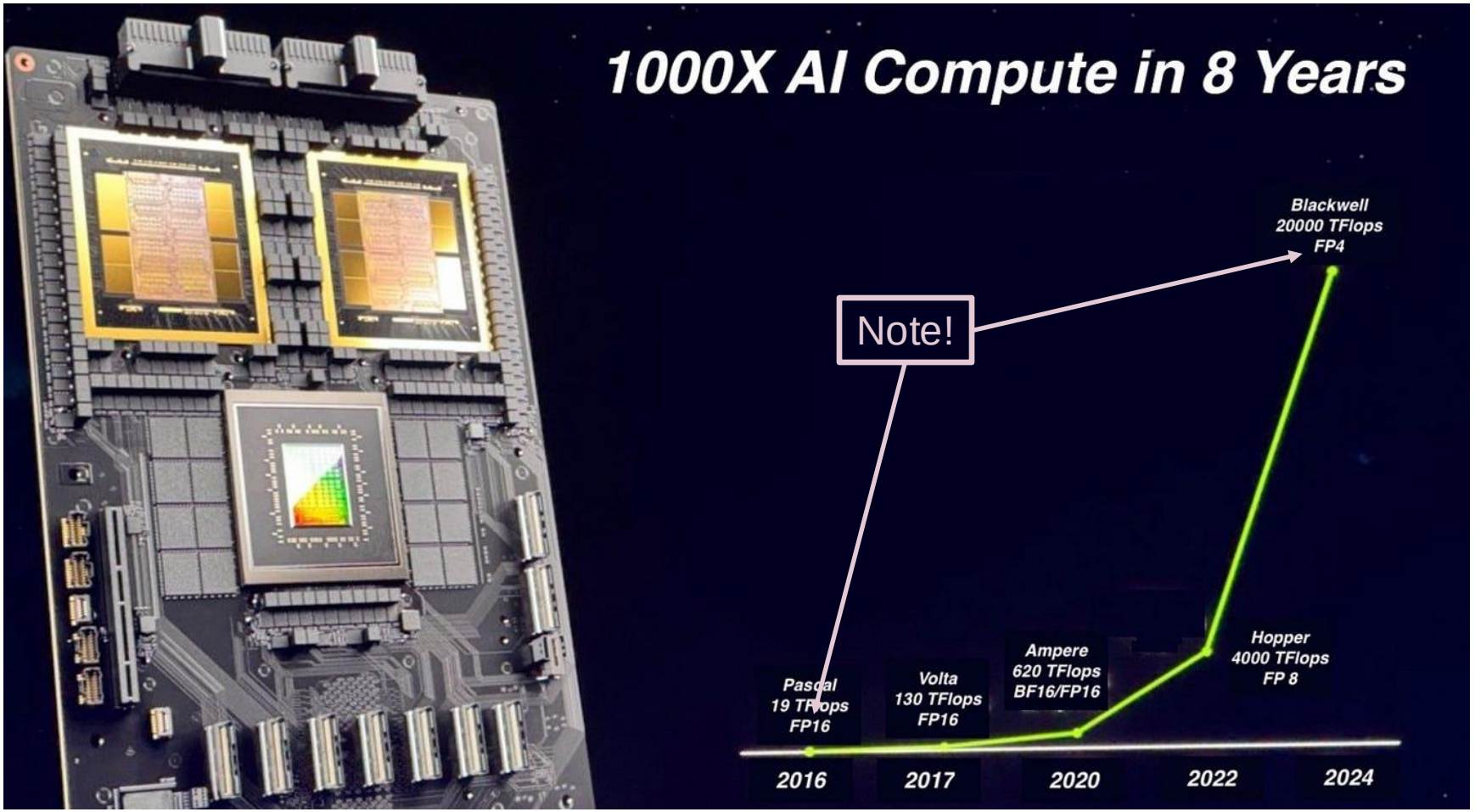
384GB of HBM3e

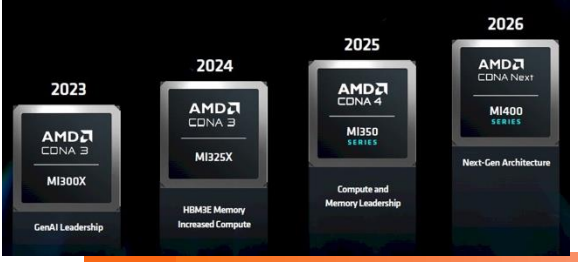
72 Arm Neoverse V2 CPU cores

900GB/s of NVLink-C2C bandwidth



AI performance improvements: NVidia claims with APU GB200





AMD APU: AMD Instinct MI300A

AMD CDNA™ 2
Coherent Memory Architecture

CPU

GPU

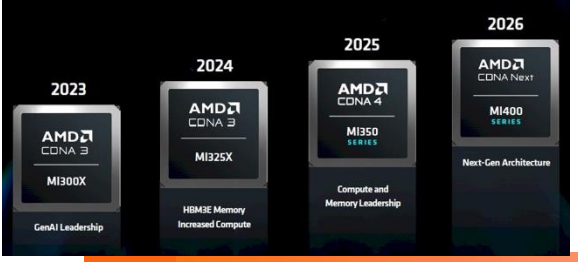
CPU Memory (DDR)

GPU Memory (HBM)

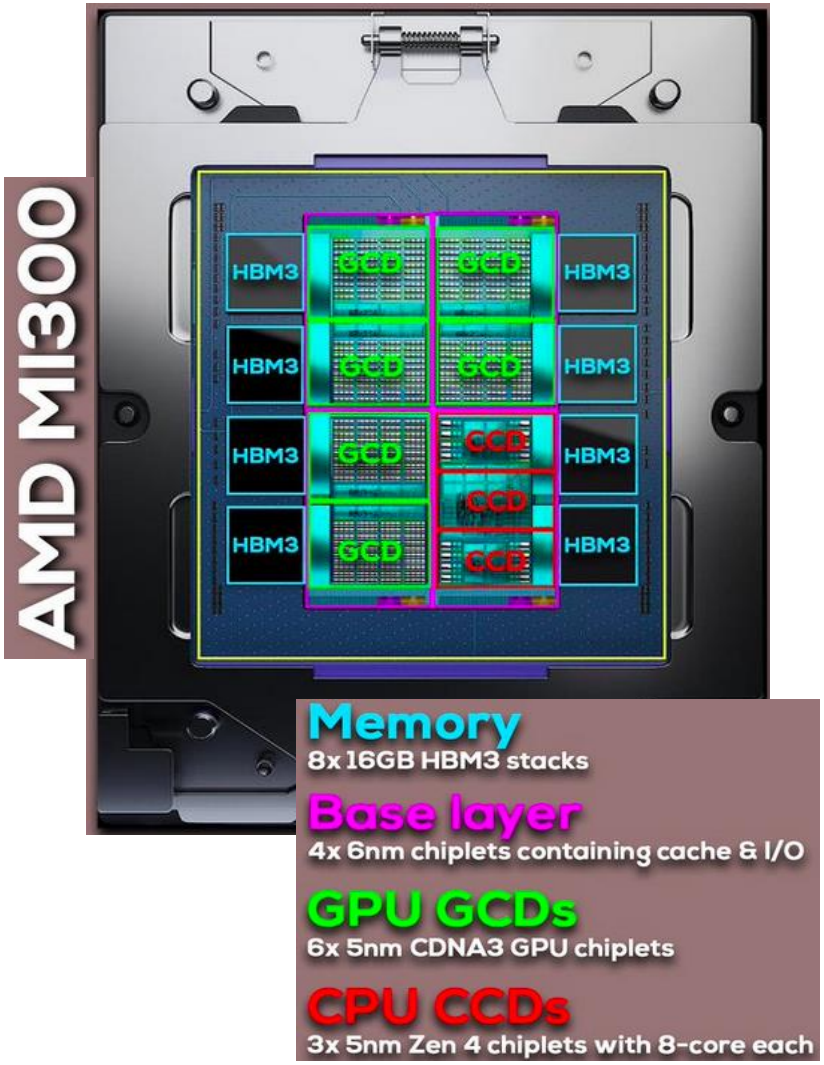
AMD CDNA™ 3
Unified Memory APU Architecture

AMD Instinct™ APU

Unified Memory (HBM)



AMD APU: AMD Instinct MI300A



	MI300A APU
ARCHITECTURE	**AMD CDNA 3**
ACCELERATED COMPLEX DIES (XCD)	6
COMPUTE UNITS	228
STREAM PROCESSORS	14,592
MATRIX CORES	912
MAX ENGINE CLOCK (PEAK)	2,100 MHz
AMD "ZEN 4" CPU CHIPLETS (CCD)	3
TOTAL "ZEN 4" X86 CORES	24
MEMORY	
MEMORY CAPACITY	128GB HBM3
MEMORY INTERFACE	1024-bits x 8 Stacks HBM3
MEMORY CLOCK	5.2 GT/s
MEMORY BANDWIDTH (PEAK)	up to 5.3 TB/sec
L1 CACHE	32 KiB
L2 CACHE	4 MB
AMD INFINITY CACHE™	256 MB

A sample of current compute devices in HPC: manycore and compute accelerators

- Latest generations of manycore PUs used in HPC:
  - Intel Xeon devices: from Cascade Lake to Granite Rapids
  - AMD Epyc devices: the Zen family, from Naples to Turin
  - a wafer-scale device: CEREBRAS (3rd generation)
  - missing: ARM-based devices, Chinese devices, …
- Compute accelerator devices *(GPUs)* used in HPC :
  - NVidia: from Volta architecture to Blackwell
  - missing accelerators: Intel, AMD, …
- APU devices (*Accelerated Processing Units, PU+GPU*) in HPC:
  - NVidia Superchips: from Grace+Hopper to Grace+Blackwell
  - AMD Instinct MI300A
- Popular SoC *(System-on-Chip)* devices: Apple M4 and A17

Previous slides showed the top chips in HPC; what about laptops?

The most popular are the System-on-Chip (SoC) designs from Apple, based on the European ARM processor.

Current chip: **M4**; Apple plans to launch a new chip every year, M5 in 2025, M6 in 2026

Apple M4 chip

The new Apple M4 chip

4 performance cores

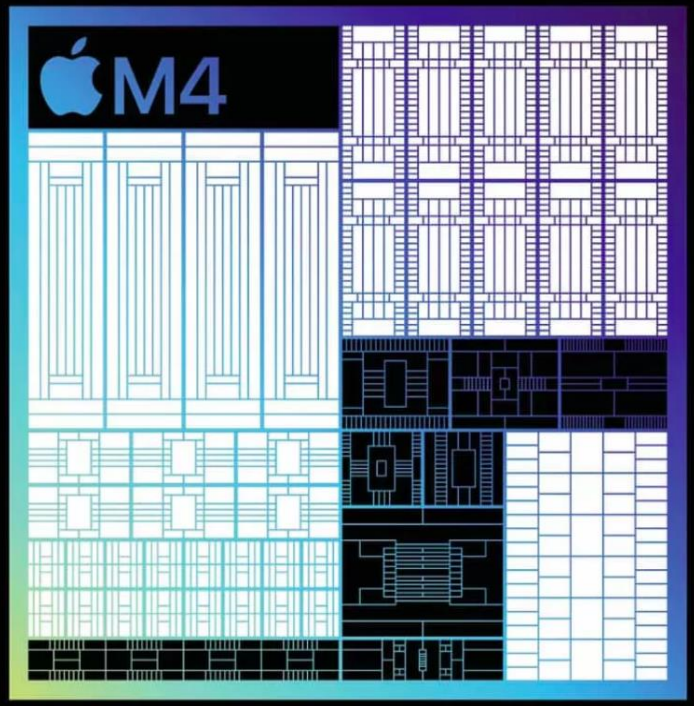
Improved branch prediction
Wider decode and execution engines
Next-generation ML accelerators

6 efficiency cores

Improved branch prediction
Deeper execution engine
Next-generation ML accelerators

Neural Engine

16-core design
Faster and more efficient



10-core GPU

Next-generation architecture
Dynamic Caching
Mesh shading
Ray tracing

Display engine

Tandem OLED support
Brightness and color compensation
10Hz–120Hz ProMotion support

Second-generation
3 nm technology

Apple iPhone chip