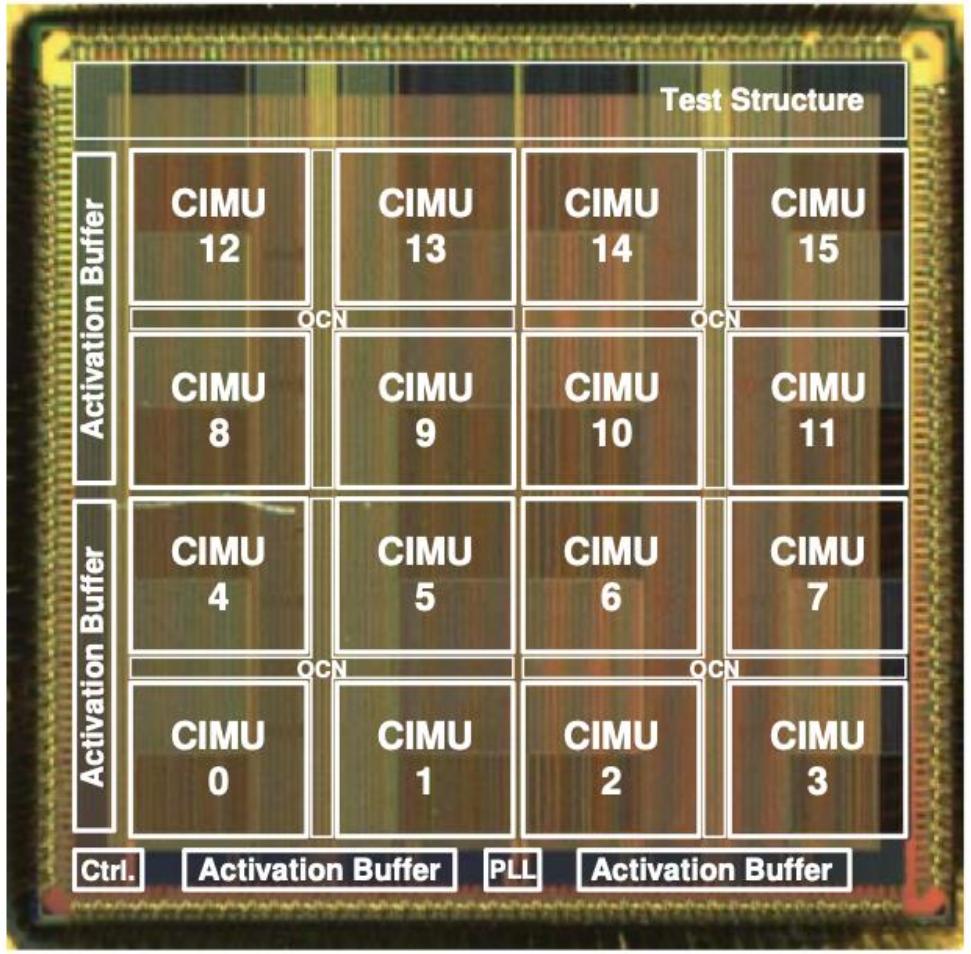
AMA on Semiconductors

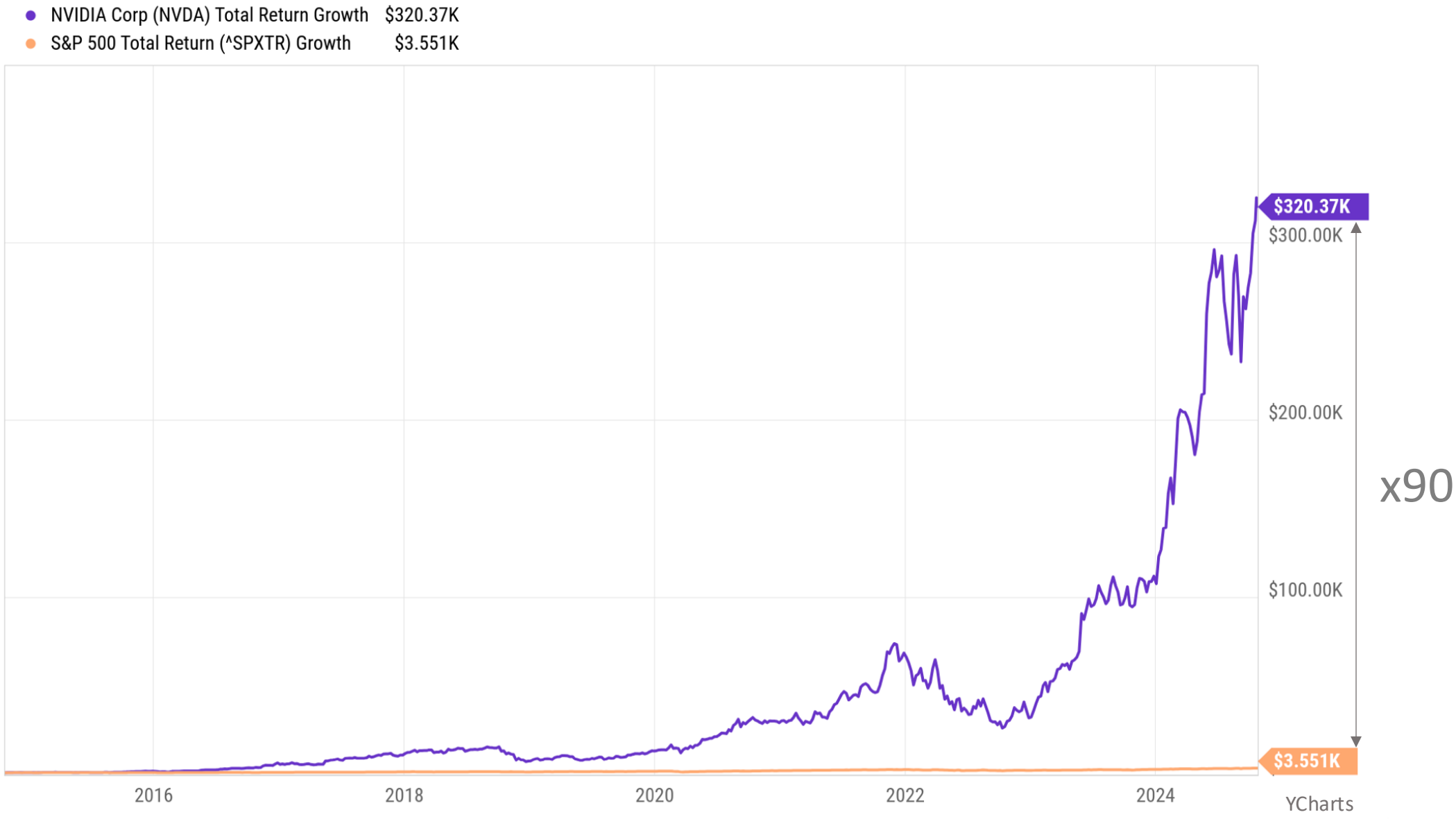
AI Hardware is Capital Intensive. How Can Angels invest?

- Prototyping cost in advanced technology can reach $1M … $10M.
- According to IBS reports R&D and manufacturing costs can reach up to $300M for an SoC in 7nm node and almost double for 5nm.
- Angel investors can easily enter with $250k – $1M rounds at very early stage of building a prototype.
- Look for university spin-offs which already have technology demonstration from prior academia.



EnChargeAI is a spin-off from Princeton University. Chip above is a prototype of an in-memory computing AI accelerator from Princeton published at ISSCC 2021.

NVIDIA Return on $1,000 Investment over 10Y



x90

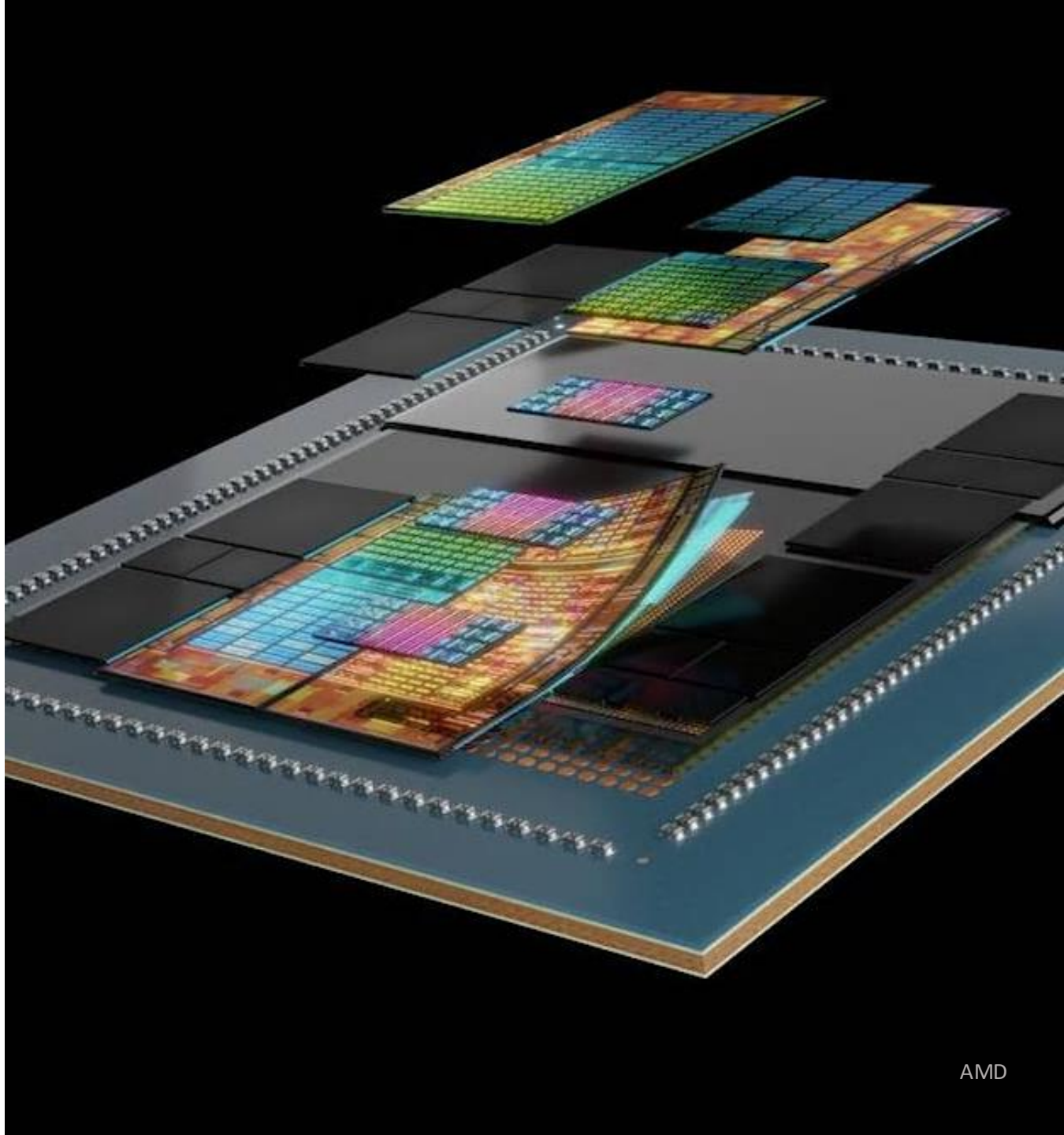
Oversupply of AI Chips

- According to McKinsey “Generative AI” report, computing demand for GenAI tasks will increase 100x over the next 6 years (from 0.2 QFLOPs to 25 QFLOPs).

- 1 QFLOPs = 1 x 10^{30} FLOPs ≈ 10 Trillion most advanced GPUs (NVIDIA B200)

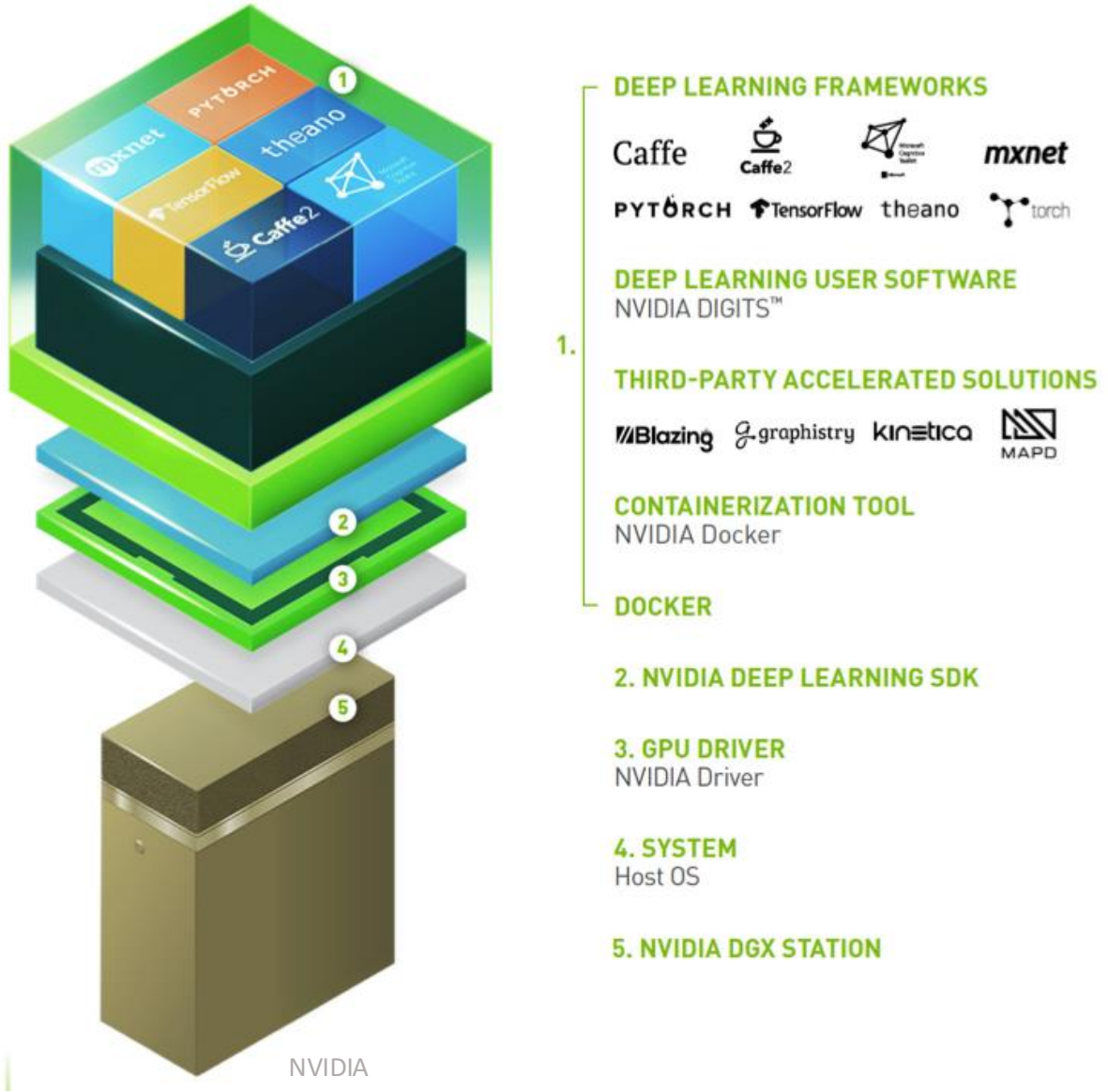
- McKinsey assumes almost quadrupling of computing demand every year.

- In 2024 about 8 Million AI chips were shipped.



Standardization of AI Chips

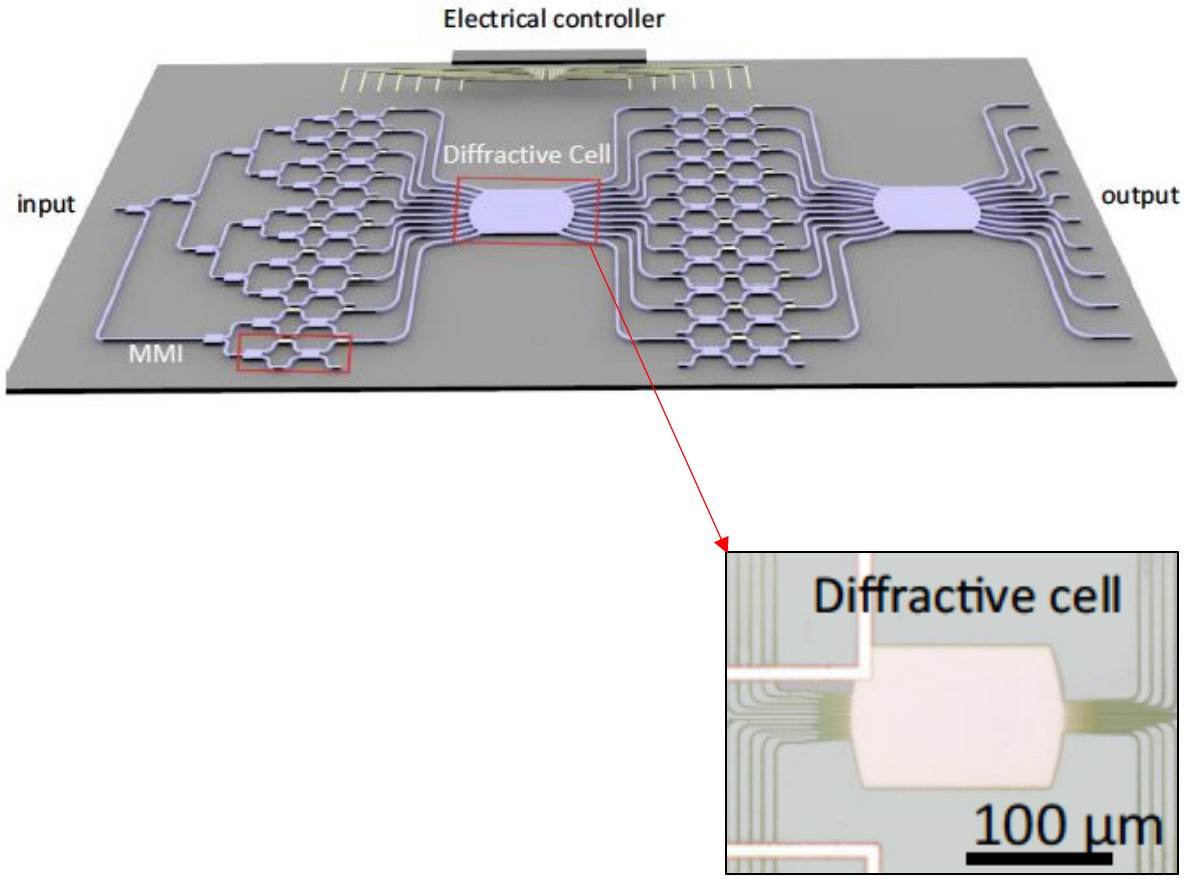
- Shortly, there are no standard like x86 instruction set or RISC-V. These are proprietary chips.
- AI Chips are built with industry-standard tools in standard technology (available to all).
- Standardization is happening on network level (e.g., UEC, UCIe).
- Standardization is also present at high-level software layers (e.g., TensorFlow, etc.)
- Chips developed by hyperscalers have all proprietary stack.



NVIDIA

Photonic Computing Potential

- Photonic computing technology is an alternative to electronics promising 10x … 100x performance.
- The status of the technology is mostly academic research but the first commercially available processor Lightmatter Envise, reportedly demonstrates 10x vs NVIDIA A100 at lower power consumption.
- Computing can be performed at multiple wavelength of light simultaneously.
- Scalability can be achieved with shorter wavelength.
- Photonic computing is at least a decade(s) away. Silicon photonics for heterogenous integration is more promising technology with shorter time span.



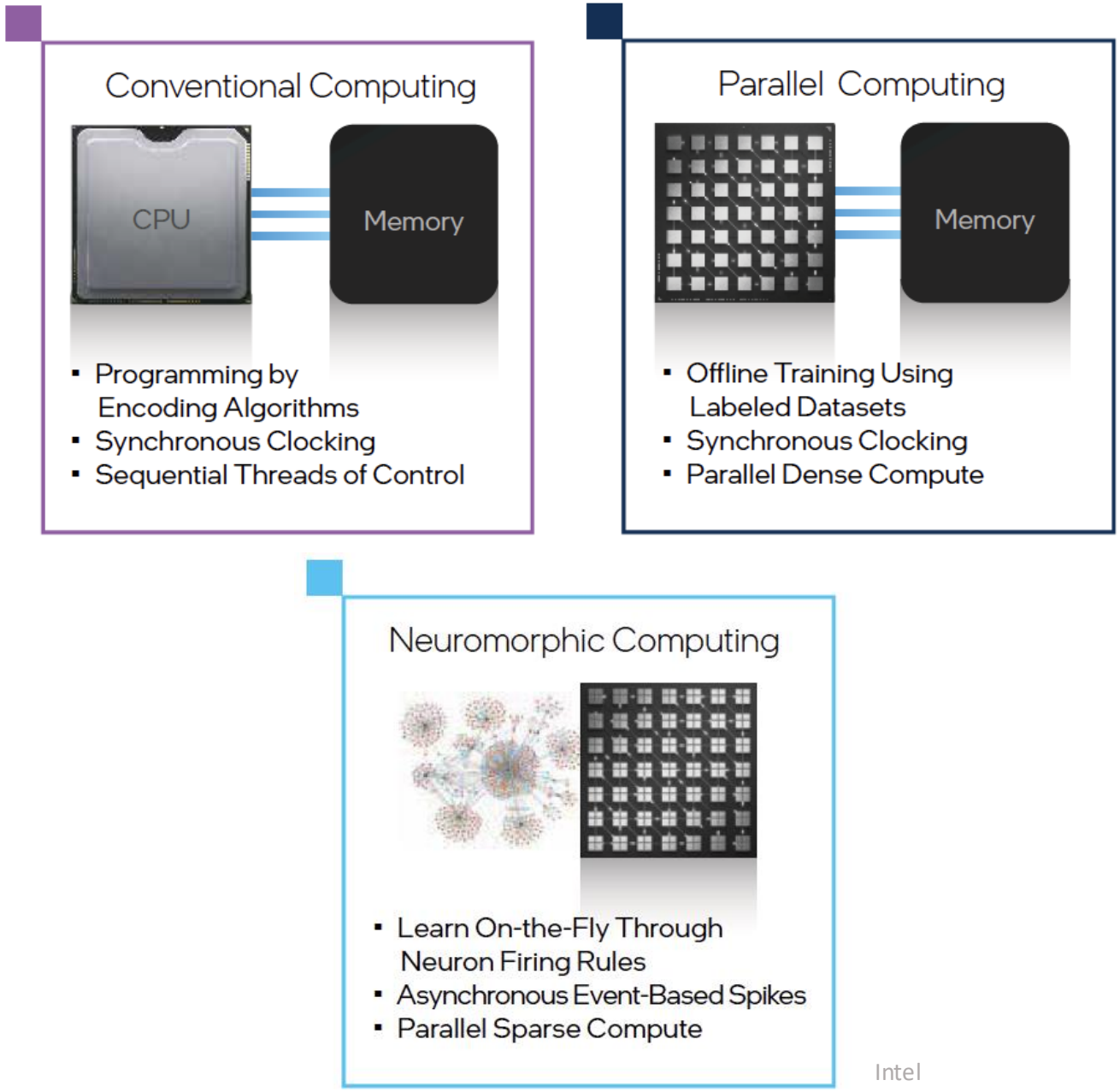
H.H.Zhu, et.al., Nature'22

Neuromorphic Computing

- Neuromorphic computing is an approach that aims to mimic the structure and functionality of human brain:
  - Artificial neurons and synapses;
  - Parallel processing;
  - Event-driven computing;
  - Integrated memory and processing.

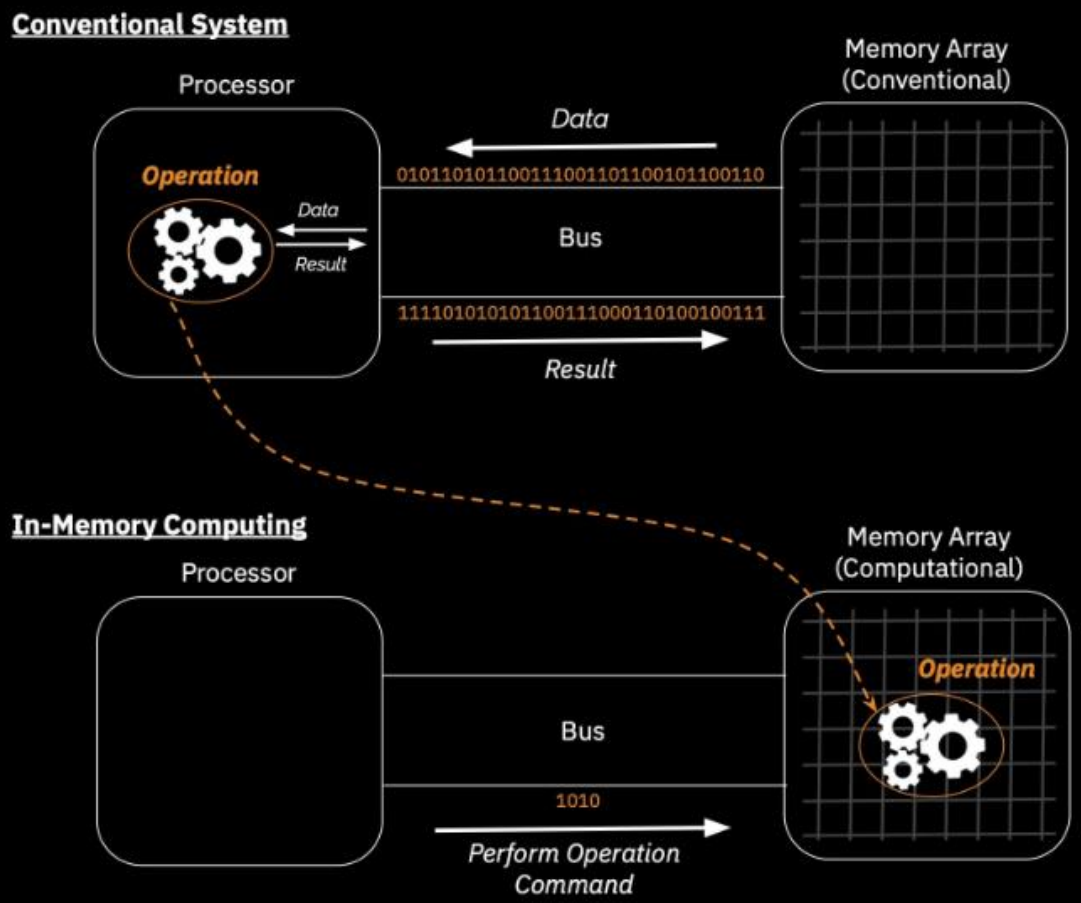
- Examples: Intel Loihi, IBM Hermes, Brainchip Akida.
- Applications: Edge AI.

- Pros: Efficiency, Speed, One-shot learning.
- Cons: Custom software stack, limited flexibility,
- Technology state: Research.



Memory Bottleneck and Solutions

- Processing speed of cores is significantly higher compared to speed of memory access.
- Solutions:
  - High Bandwidth Memory (HBM) and 3D packaging.
  - Near-Memory Computing: Samsung PIM technology demonstrates 3x vs HBM.
  - In-Memory Computing (IMC): potential for 10 – 100x vs HBM.
- In-Memory Computing implementations can be digital or analog.
- Potential for digital IMC is high for AI accelerators designed for specific workloads. Potential for IP companies.



Synthara

Other Questions:

Q: Given the software ecosystem that Nvidia has built, how can another chip vendor unseat them for most applications?
A: AMD recent acquisitions are targeting exactly that.

Q: Where do you think the cell phone processing hardware is headed with regards to integration with more edge AI? Do you think we would go back to analog processing in phones because they are more energy efficient?

A: With the latest push from e.g., Apple to place AI in their hardware, more features over time will receive dedicated hardware acceleration for optimum power and latency.

Apple