

# EPI Forum

Barcelona, 9–10.10.2024.



## High Performance Computing in the AI era

Jean-Pierre Panziera  
09/10/2024

EVIDEN



# A bit of glossary



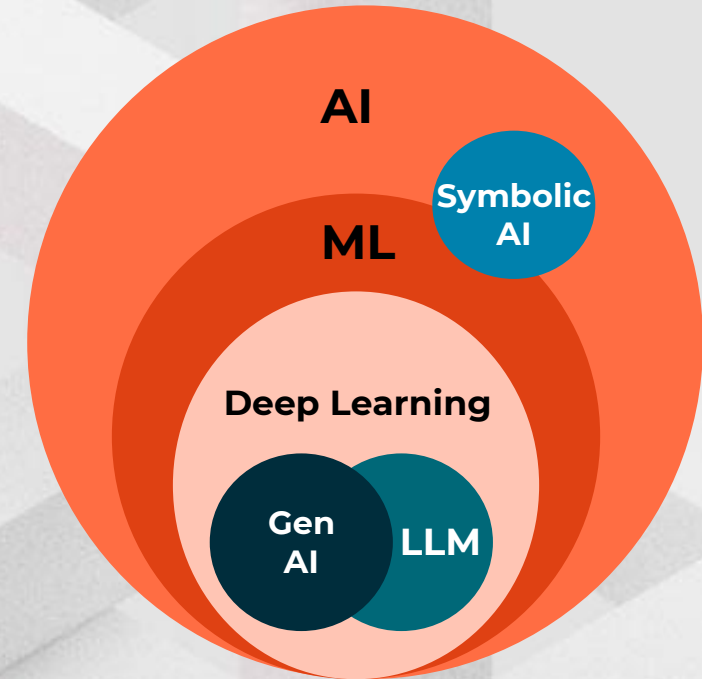
**AI** is theory and development of computer systems that can **perform tasks typically requiring human intelligence**



**Machine learning** allows computers to learn without explicit programming. ML is a subset of AI.



**Deep Learning** Uses Artificial Neural Networks to process more complex patterns than traditional ML. DL is a subset of ML



**Generative AI** refers to computer systems capable of producing new content, such as images, texts, or sounds, that are often indistinguishable from those created by humans.

**LLM** stands for "Large Language Model." It's a class of artificial intelligences specialized in natural language processing, such as ChatGPT. These models are typically trained on vast amounts of textual data to understand and generate text in a contextually coherent manner.

Each **AI domain** applies to **certain type of data & certain use-cases** (e.g sensing vs reasoning).

Each AI domain has different **explainability/intelligibility** characteristics.

An important aspect in front of **rising regulation pressure** in some industries : Healthcare, Defense, Finance along with **Trustworthy AI**

# Manufacturing

## Production

- Automated quality inspection
- Device early fault detection
- Asset predictive maintenance
- Yield optimization for process manufacturing
- IT/OT Data Platforms

## Engineering / R&D

- Root cause analysis based on log data
- *HPC Simulation optimization*
- *Surrogate modeling MLOps*

## Sales & Marketing

- Personalization
- Call center and customer support AI optimization
- Demand forecasting

## IT

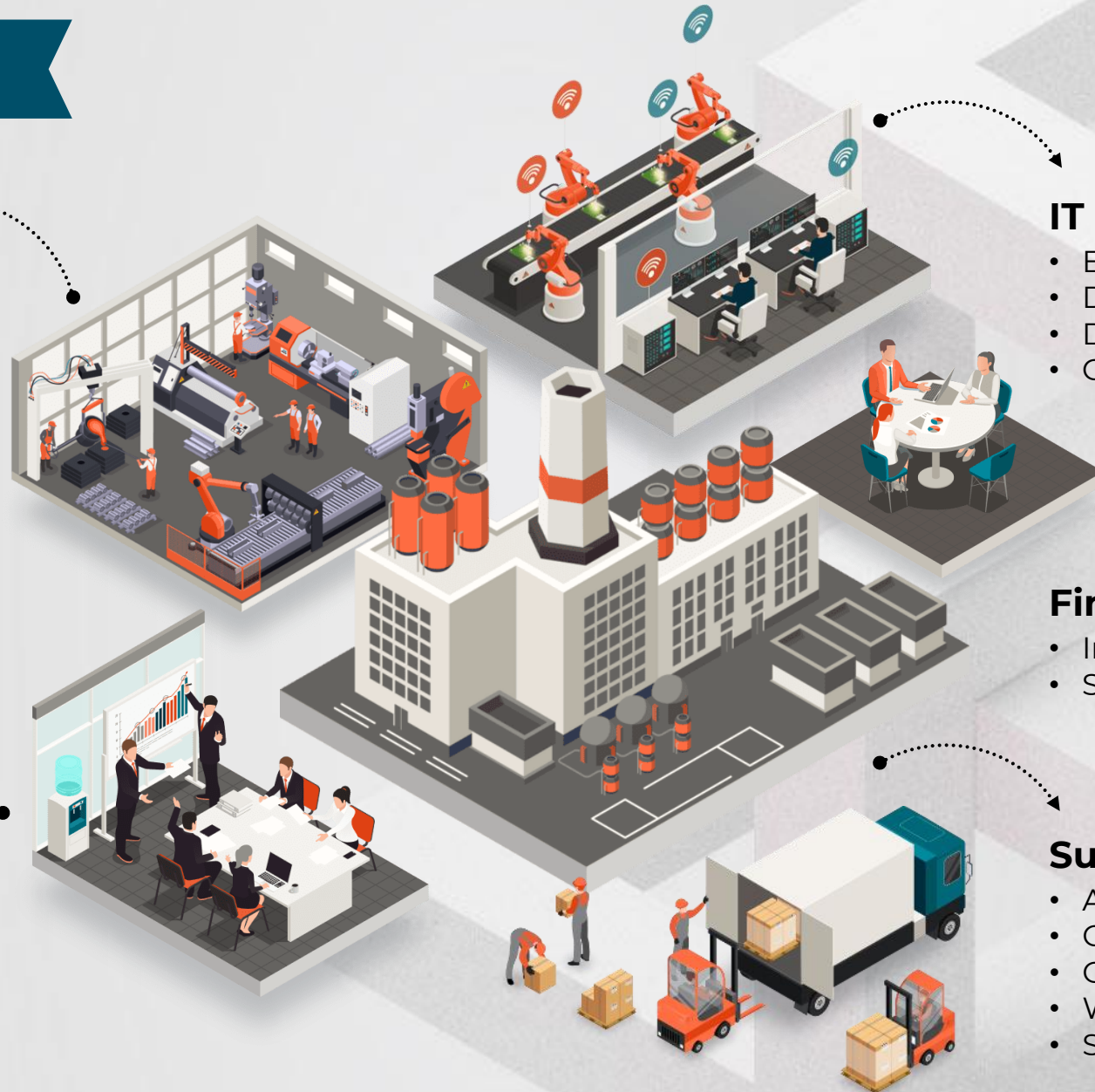
- Enterprise wide MLOps
- Data + AI platforms
- Data Mesh
- Generative AI

## Finance and Backoffice

- Invoice anomaly detection
- Streamlined backoffice

## Supply chain & logistics

- Adaptive planning
- Goods & documents traceability
- Order to cash financial analysis
- Warehouse stock optimization
- Smart distribution



# BullSequana AI range

Full DLC - Highest density - lowest TCO



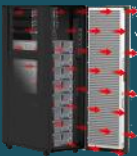
BullSequana AI 1200H

Large memory CPU Scalability



BullSequana AI 800

Hybrid Cooling - Midrange



BullSequana AI 640H HGX H100 DLC  
BullSequana AI 640H MI300X DLC



BullSequana AI 600H series

AI Air/Hybrid Cooled Midrange

BullSequana AI 640 HGX H100 Air  
BullSequana AI 640 MI300X Air



BullSequana AI 620 8 GPUs PCIe Air

BullSequana AI 600 series

AI Air Cooled Mid range

BullSequana AI 200G 2S 2U GPU



BullSequana AI 200 series

Edge computing

BullSequana AI 100R BullSequana AI 100D



BullSequana AI 100 series

Training (Fine-tuning)

Training

Data Ops

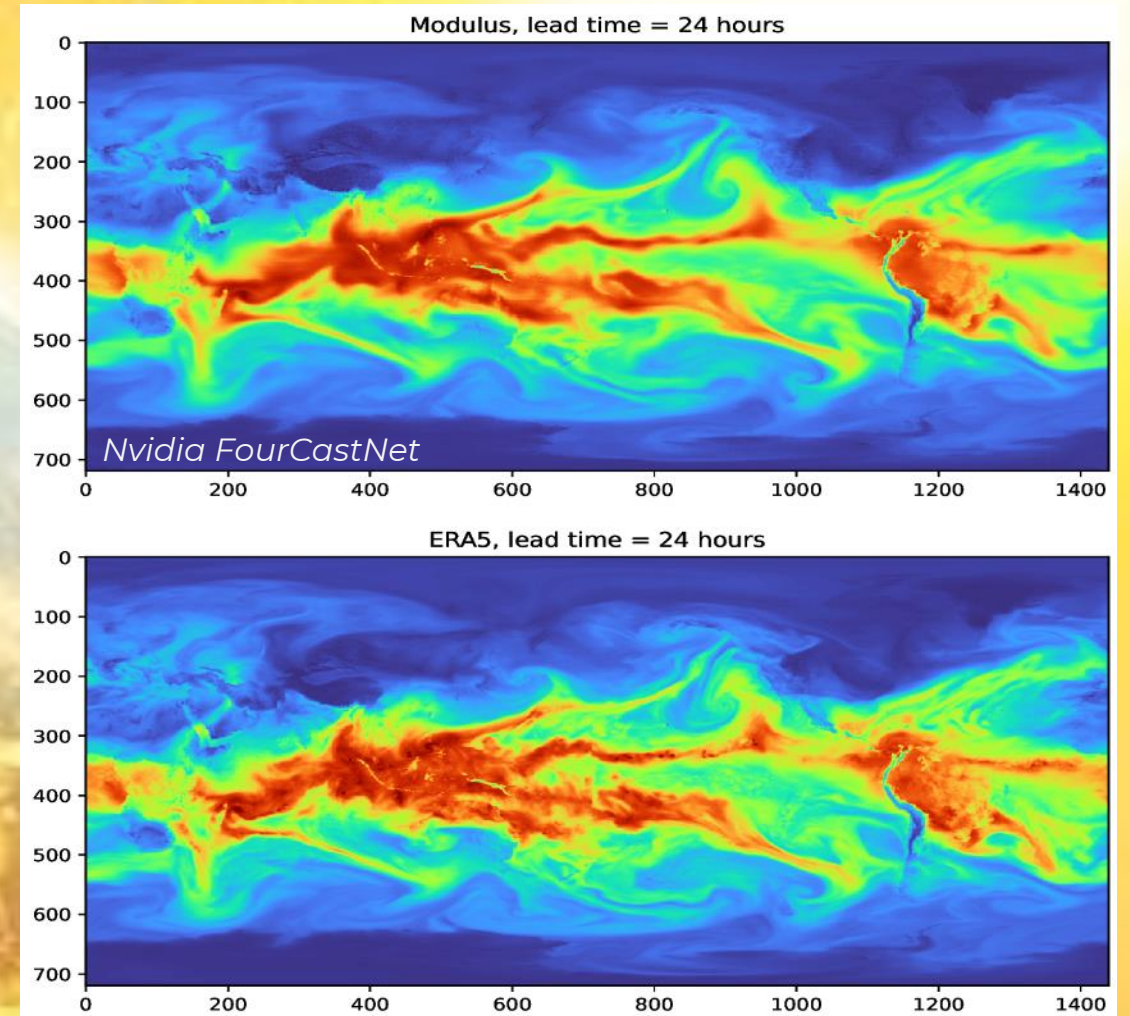
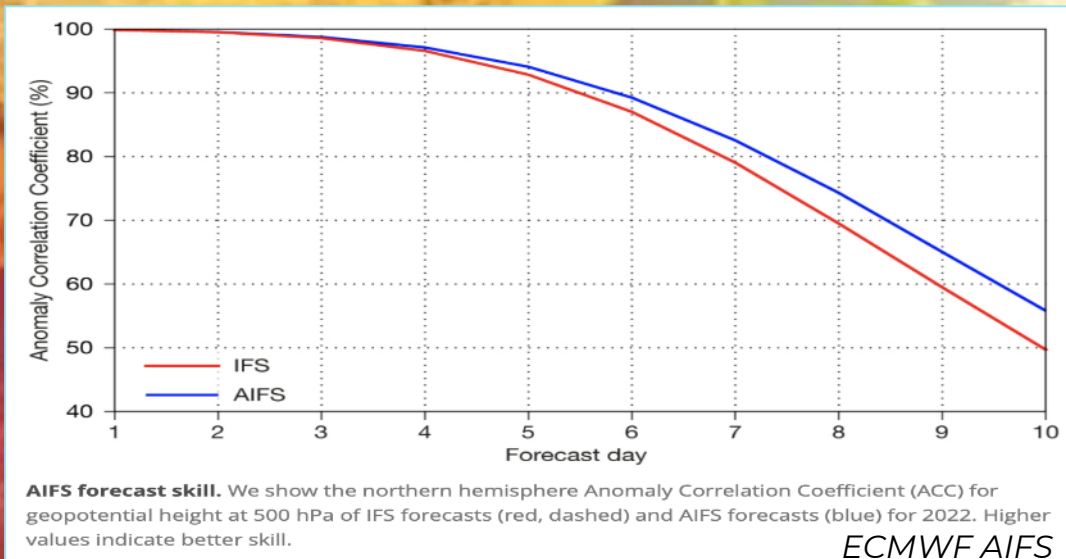
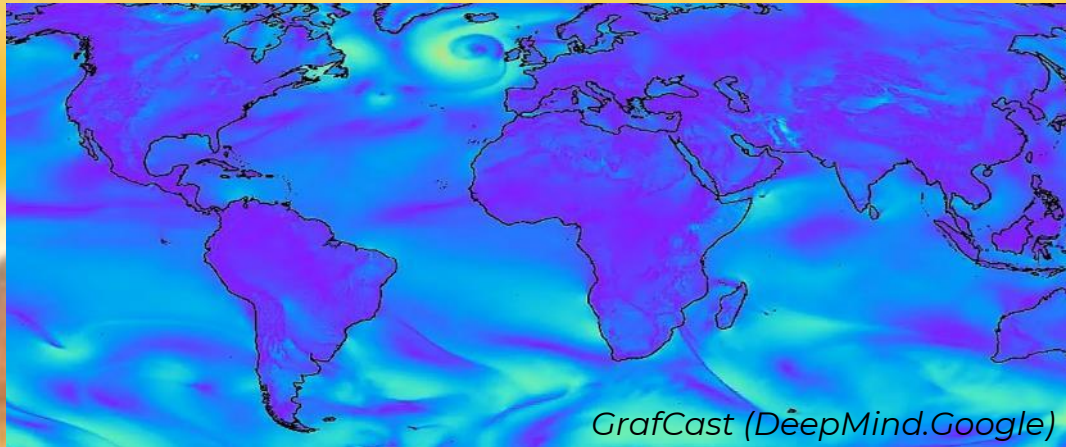
Fine Tuning – LLM

Inferencing

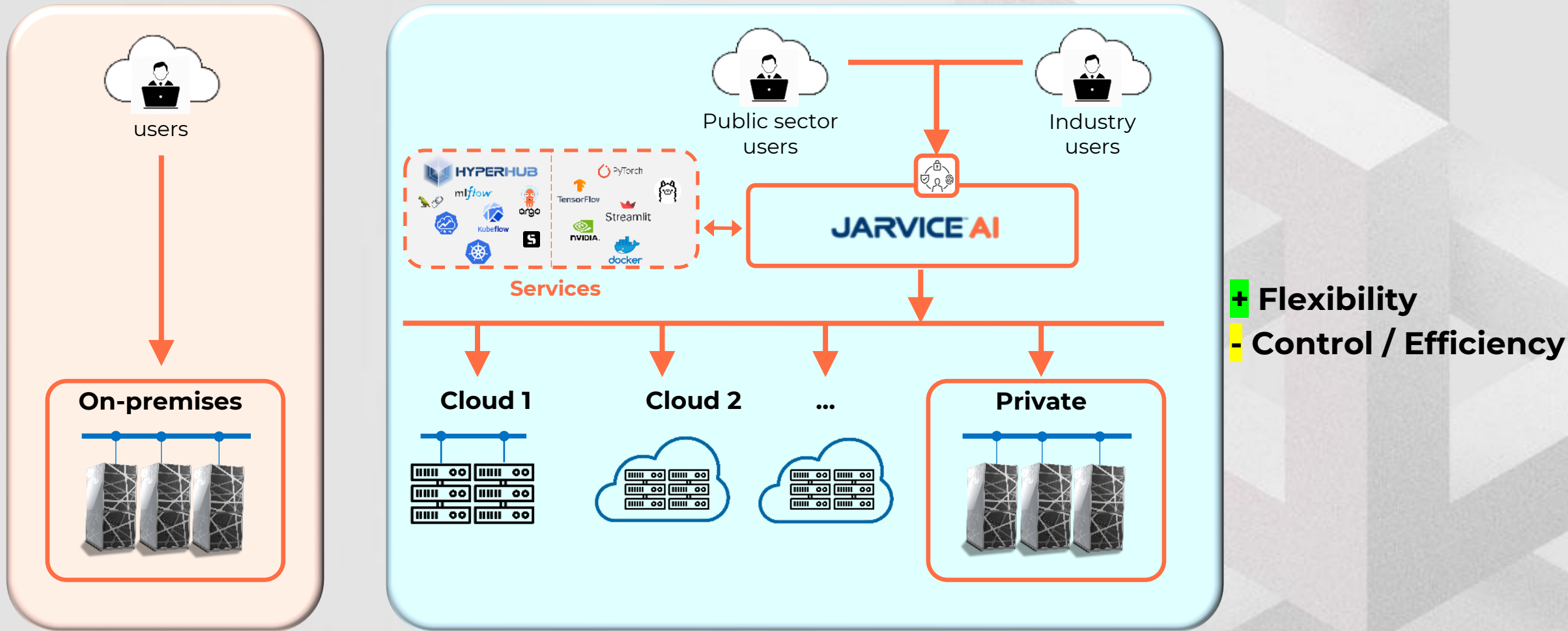


# AI for HPC

## AI acceleration for Weather Forecast and Climate Modeling



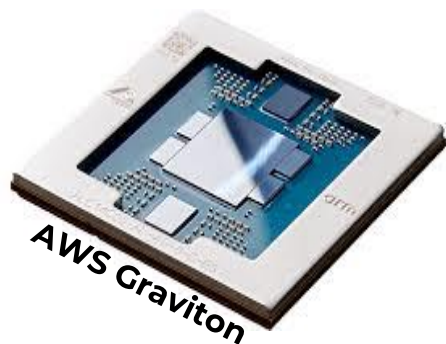
# The Cloud becomes *\_the\_* AI system





# CPUs: x86 Intel, x86 AMD & Arm

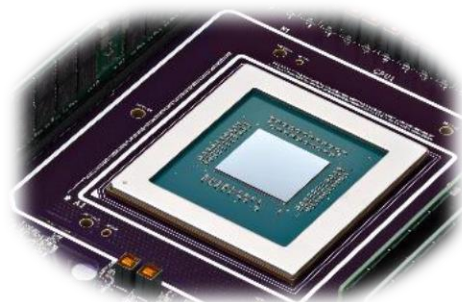
Hyperscalers, Cloud providers are developing / adopting Arm CPUs



AWS Graviton



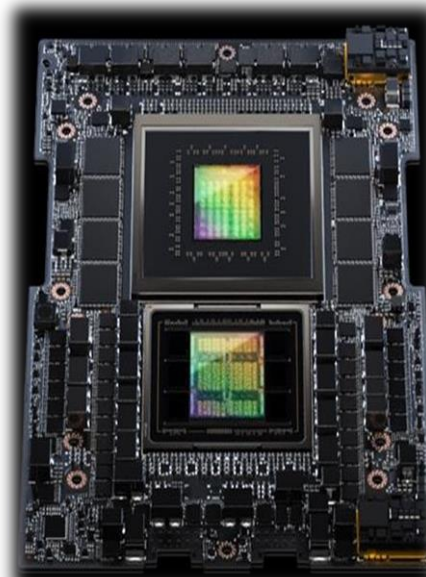
MS Azure Cobalt



Google Axion

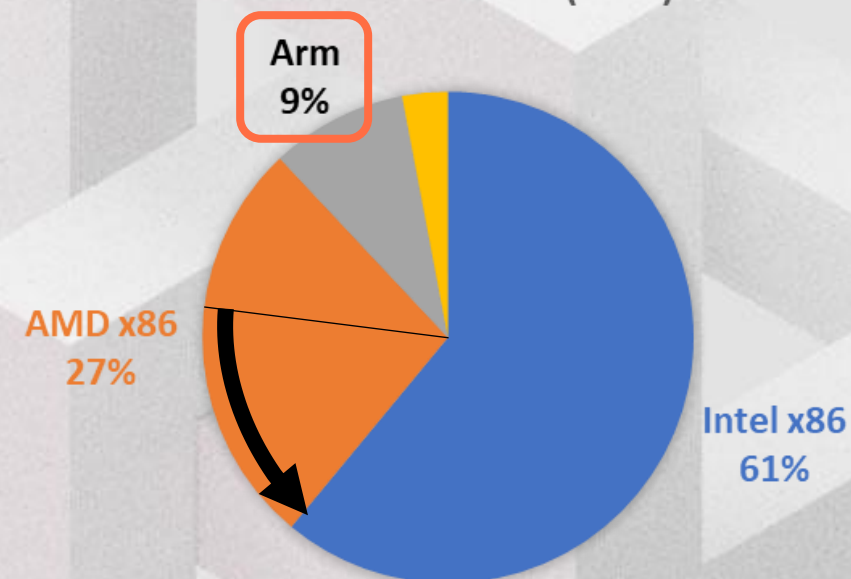


Ampere



Nvidia GH200

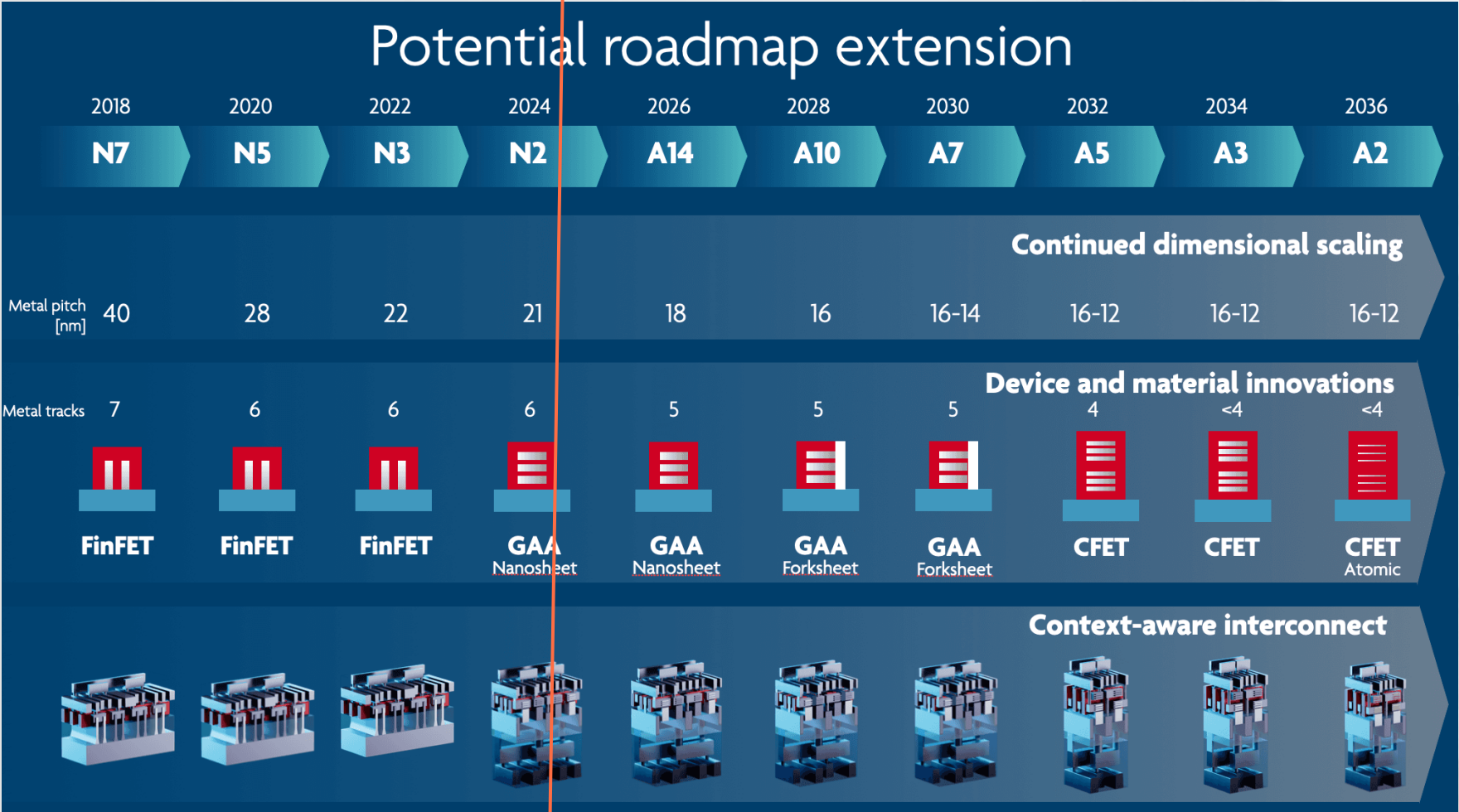
DATACENTERS CPUS (2023)



But no RISC-V CPUs in Datacenters ... yet

# Moore's law not dead ... yet

IMEC's chip scaling roadmap → 2036



<https://www.imec-int.com/en/articles/smaller-better-faster-imec-presents-chip-scaling-roadmap>



# Tflops / Tops evolution

Tflops TOps	RTX 4090	P100	A100	GH200 (1)	GB200 (2)
FP64	1	5	10	34	90
FP64 TC		5	20	67	90
FP32	83	11	20	67	180
TF32 TC	83	11	156	494	5 000
BF16 TC	165		312	989	10 000
FP16 TC	165	21	312	989	10 000
INT8 TC	661		624	1 979	20 000
FP8 TC	661			1 979	20 000
FP6 TC					20 000
INT4 TC	1 321				
FP4 TC					40 000
Watts	450	300	400	1 000	2 700

(1): 1 CPU Grace + 1xGPU H200  
(2): 1 CPU Grace + 2xGPUs B200

Gflops GOPs / W	RTX 4090	P100	A100	GH200 (1)	GB200 (2)
FP64	3	18	24	34	33
FP64 TC		18	49	67	33
FP32	184	35	49	67	67
FP32 TC	184	35	390	494	1 852
BF16 TC	367		780	989	3 704
FP16 TC	367	71	780	989	3 704
INT8 TC	1 469		1 560	1 979	7 407
FP8 TC	1 469			1 979	7 407
FP6 TC					7 407
INT4 TC	2 936				
FP4 TC					14 815

NO real improvement  
for FP64

x50-200

emulate FP64 with  
FP32 TC ? INT8 TC ?

- GPUs consumption going up, **Up, UP** → 2000W next generation → **Liquid Cooling** mandatory
- New generation GPUs are designed for AI workloads optimization
- FP64 performance flatlining
- FP64 emulation would require strong support from toolchain (compilers, libraries...)

# Energy for AI

## Big Computing goes Nuclear

### Amazon buys nuclear-powered data center from Talen

Thu, Mar 7, 2024, 2:01PM | Nuclear News



Susquehanna nuclear plant in Salem Township, Penn., along with the data center in foreground. (Photo: Talen Energy)

Talen Energy announced its sale of a 960-megawatt data center campus to cloud service provider Amazon Web Services (AWS), a subsidiary of Amazon, for \$650 million.

ans.org

### Three Mile Island nuclear plant will reopen to power Microsoft data centers

SEPTEMBER 20, 2024 - 1:40 PM ET



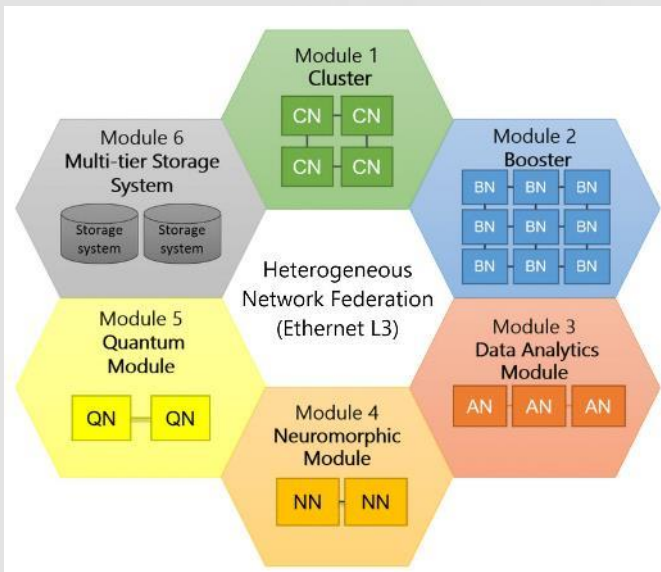
The Three Mile Island nuclear plant is seen in March 2011 in Middletown, Pa. Jeff Fusco/Getty Images

npr.org



# Big Compute / Big System / Big Data

Connect them all ... with BXI v3



## BXI V3 for AI and HPC workloads

- High-bandwidth, Low latency, High message rate, RDMA
- Ethernet +
- Hardware acceleration for Communications offload
- Highly scalable (128modules x64k), efficient and reliable
- Preview of the upcoming UltraEthernet standard

## BXI V3 Key Figures

Samples Q1/2025

Optimized for the Modular / Partitioned Architectures

Target AI and HPC market

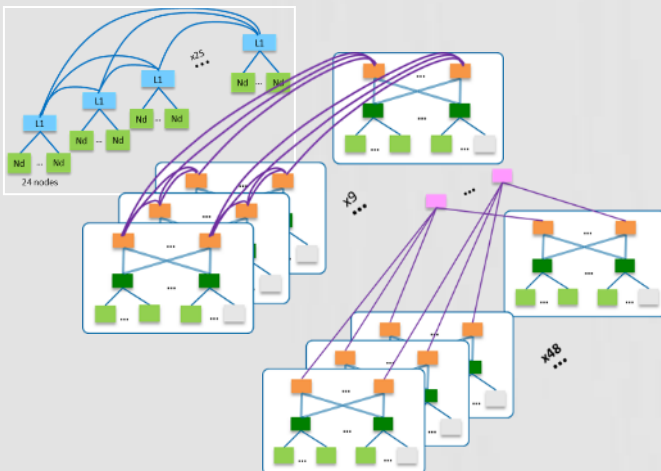
Photonics ready

### NIC:

- Maximum message rate: 220 MMsg/s
- Link Bandwidth: 400 Gb/s
- Sub-micro second latency for intra rack nodes
- Portals/IP offload

### Switch:

- 51.2 TB/s
- 64 ports at 800Gb/s or 128 ports at 400 Gb/s



# HPC in the AI era

Exciting times / Big Challenges ahead:

- AI rules the IT world
- New HW architectures to serve AI
- New usage mode → Public Cloud
- AI acceleration for traditional HPC applications
- Re-engineer HPC applications to take advantage of AI-driven HW
- Computing Components Power Consumption increasing +++
  - Specialized accelerators
  - Optimized SW



EVIDEN

Thank you

<https://eviden.com/>