SYNOPSYS®

Enabling Programmable Performance: Memory and Interconnect Innovation for AI-Centric Data Planes

Phani Suresh Paladugu
Executive Director-Product Management
P4 Workshop 2025

Legal Disclosure

CONFIDENTIAL INFORMATION

The information contained in this presentation is the confidential and proprietary information of Synopsys. You are not permitted to disseminate or use any of the information provided to you in this presentation outside of Synopsys without prior written authorization.

IMPORTANT NOTICE

This presentation may include information related to Synopsys' future product or business plans. Such plans are as of the date of this presentation and subject to change. Synopsys is not obligated to update this presentation or develop the products with the features and/or functionality discussed in this presentation. Additionally, Synopsys' products and services may only be offered and purchased pursuant to an authorized quote and purchase order or a mutually agreed upon written contract.

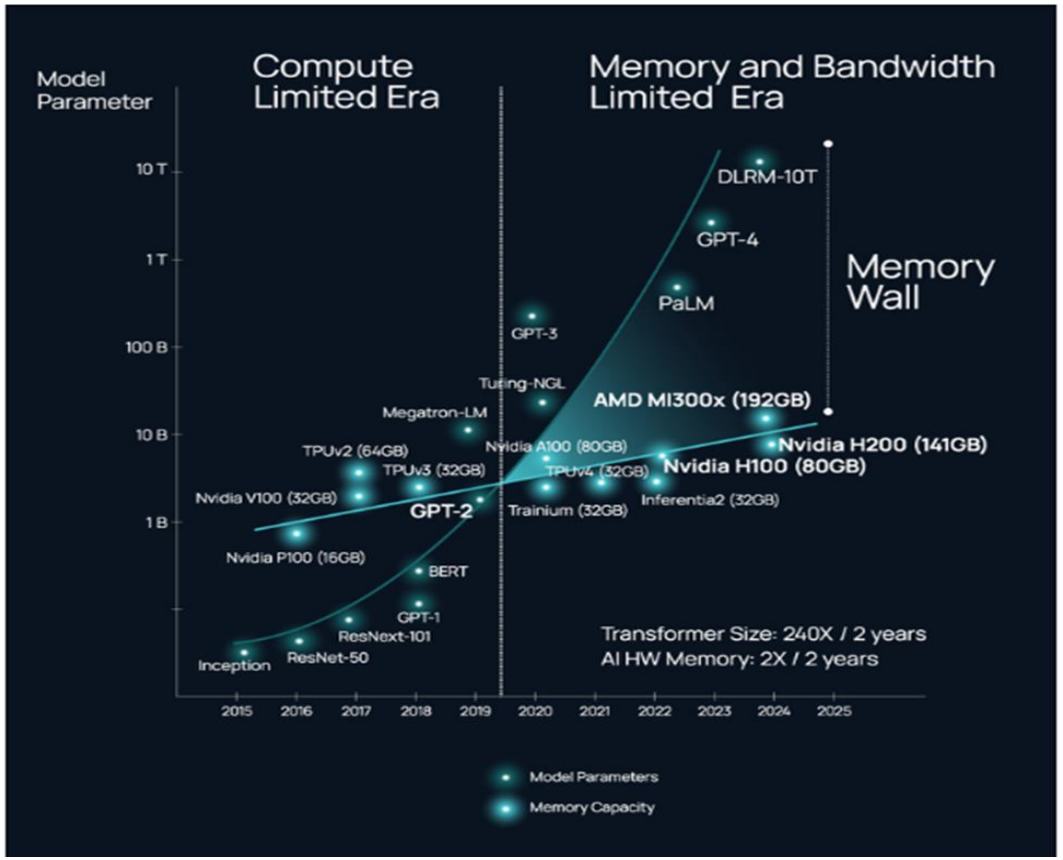
FORWARD LOOKING STATEMENTS

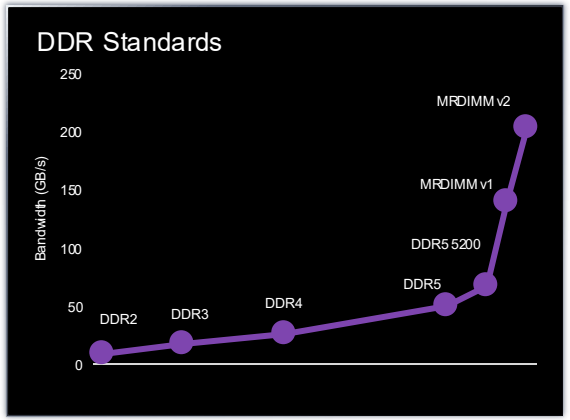
This presentation may include certain statements including, but not limited to, Synopsys' financial targets, expectations and objectives; business and market outlook, business opportunities, strategies and technological trends; and more. These statements are made only as of the date hereof and subject to change. Actual results or events could differ materially from those anticipated in such statements due to a number of factors. Synopsys undertakes no duty to, and does not intend to, update any statement in this presentation, whether as a result of new information, future events or otherwise, unless required by law.

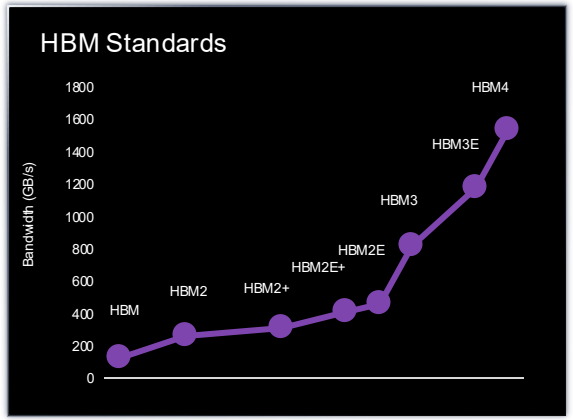
Agenda

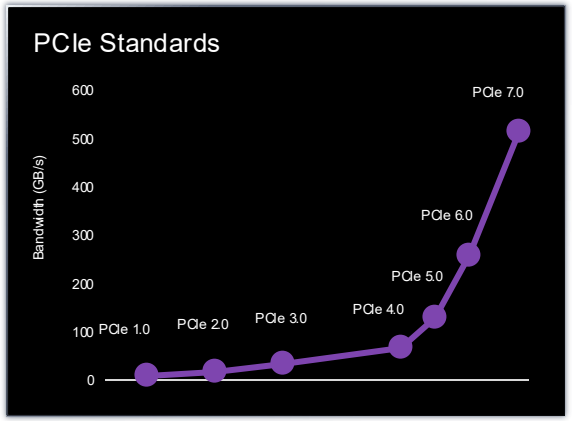
1	2
The AI Revolution and need for data processing	Memory Technologies: Advanced HBM, DDR, and LPDDR
3	**4**
Interconnect Technologies (PCIe, CXL, UALink and Ultra Ethernet)	Chiplet-Based Designs
5	**6**
Future of Programmable AI Infrastructure	Q&A

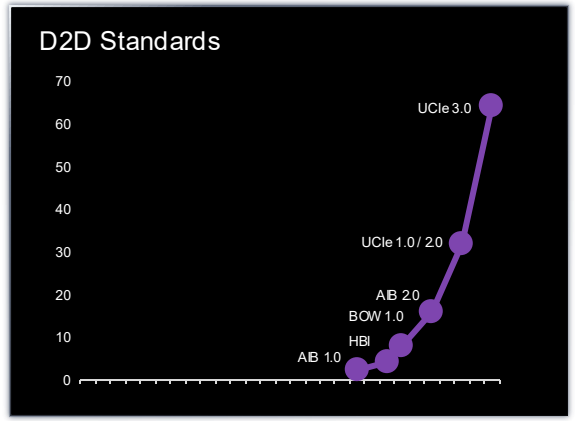
Pervasive AI is Driving Insatiable Need for Data Processing











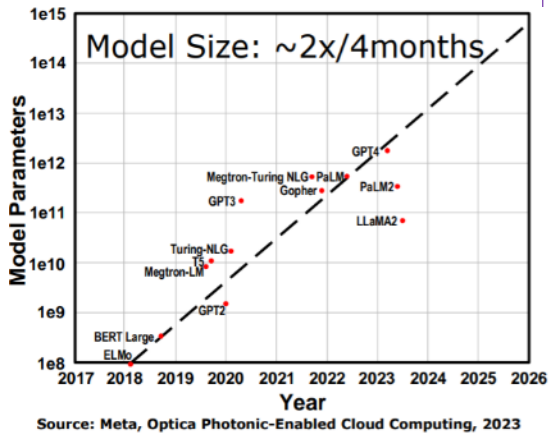
Scaling Up, Scaling Out is Required

Source: Meta, OCP Summit Presentation Nov 2021

AI Workloads Defining Memory and Interface Data Rates

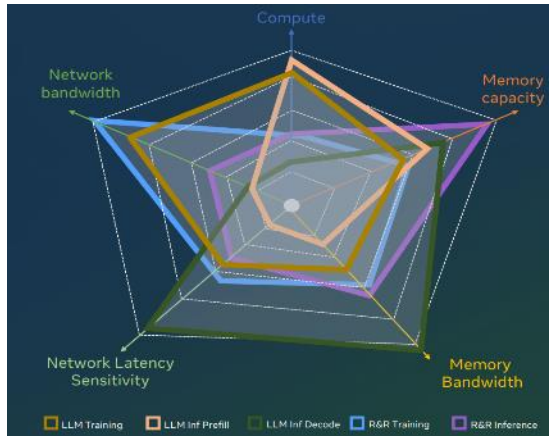
Bandwidth & Throughput are Becoming System Bottlenecks

AI Model Size 2x/4 Months

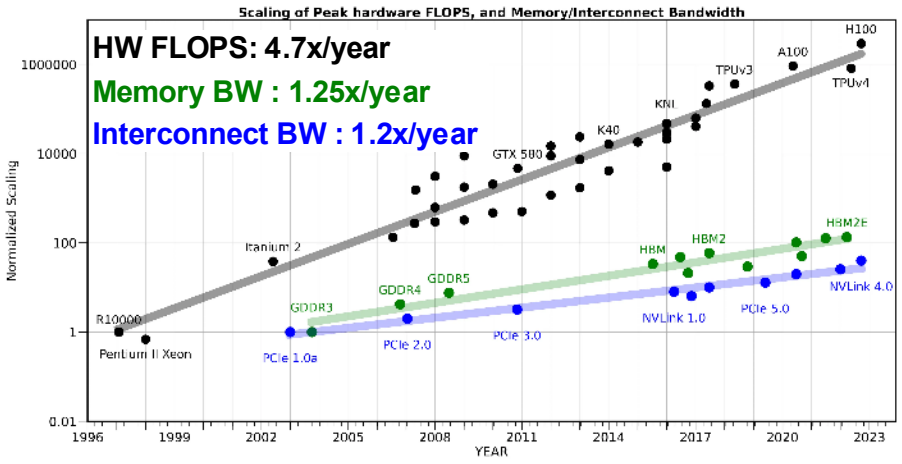


Source: Meta, Optica Photonic-Enabled Cloud Computing, 2023

Multi-Dimensional System Design



Compute Power Demand Outpacing IO & Memory Bandwidth



For Balanced architecture 1PFLOP of Compute requires:

- 3 to 4 TB/s HBM
- 0.5 to 1.0 TB of PCIe & Ethernet

- Memory Interfaces
  - Model needs to fit in the working memory
  - Model's output (tokens/sec) depends on how quickly model weights & KV caches are moved
- PCIe/CXL
  - Data ingress-egress b/w CPU-XPU-NIC-System memory
  - PCIe BW is essential for moving data from Ax & network
- UAL & UEC
  - Scale Up & Out
  - Model architecture decides the scale up/out boundaries
  - The workload is distributed across Accelerators
  - Gradients, model update, data shards are scale out

Memory Interfaces

HBM for AI

- HBM3E → HBM4 → HBM4E
- 3x bandwidth benefit
- Key drivers: Accelerators

DDR for Servers

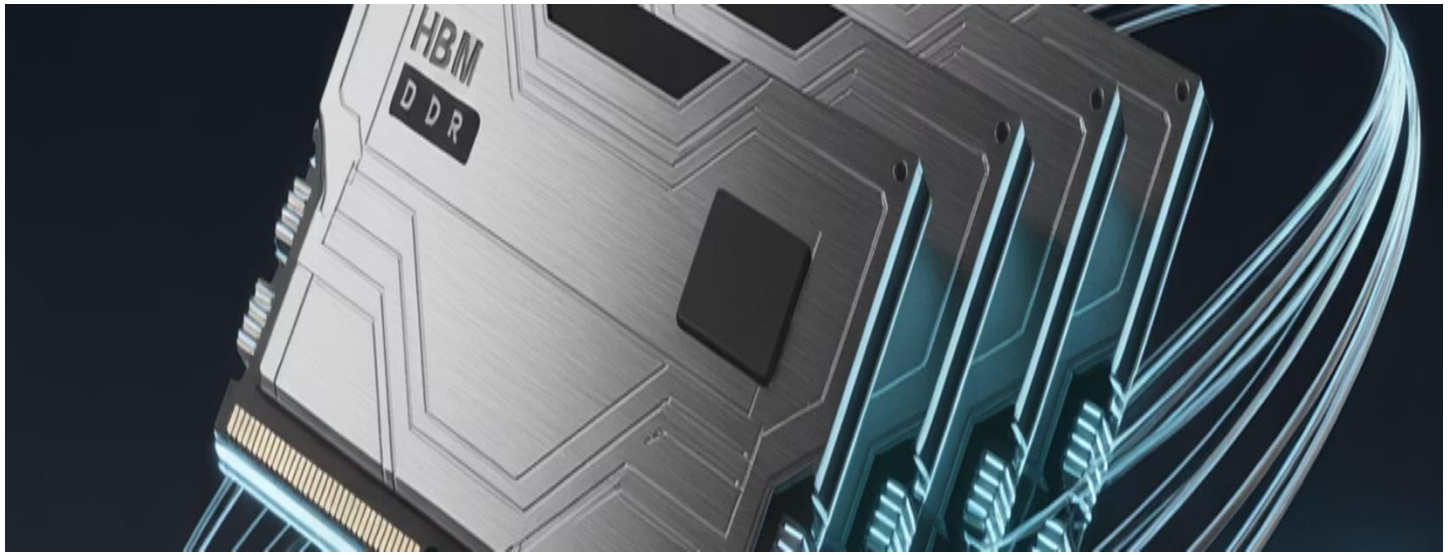
- DDR5 RDIMM → MRDIMM
- 2x bandwidth benefit
- No change to DDR5 SDRAM
- Gen 2 @ 12.8G now

LPDDR for Mobile

- LPDDR5X → LPDDR6
- 2x bandwidth, lower power & enhanced RAS

Unlocking AI Performance with HBM4/3E

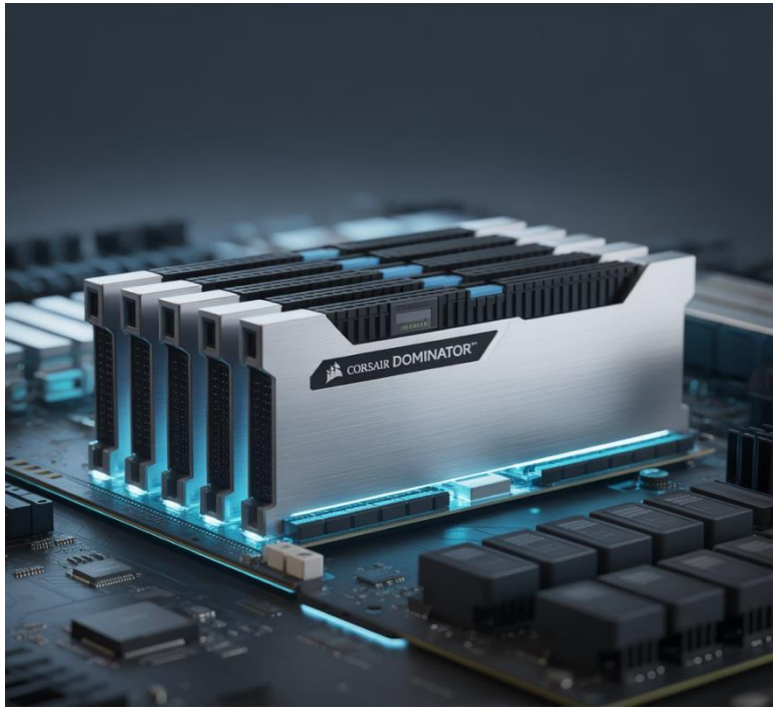
- **Unprecedented Bandwidth**
  - HBM4 sets new benchmarks, delivering unparalleled data throughput per stack
- **Colossal Capacity**
  - Each HBM4 stack provides extensive memory, crucial for handling complex AI models.
- **Revolutionary Power Efficiency**
  - Experience superior performance-per-watt, dramatically reducing operational energy consumption



Collectively, these innovations empower AI models with lightning-fast access to vast training data and complex parameters, effectively dismantling memory bottlenecks and propelling the next generation of AI capabilities.

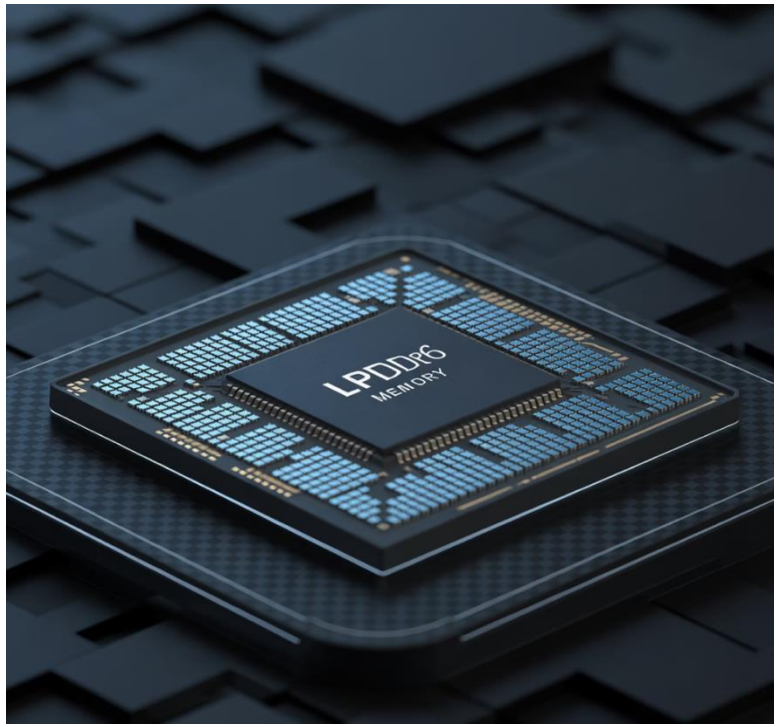
Redefining Cloud & Server Computing with DDR5

- Higher Bandwidth
- On Die ECC and RAS Features
- Improved Power management
- Larger capacities: Enables higher memory density per module for servers & HPC
- Increased Burst length improves data throughput
- Dual Sub Channel Architecture
  - Each DIMM splits into two independent 32-bit channels for efficiency



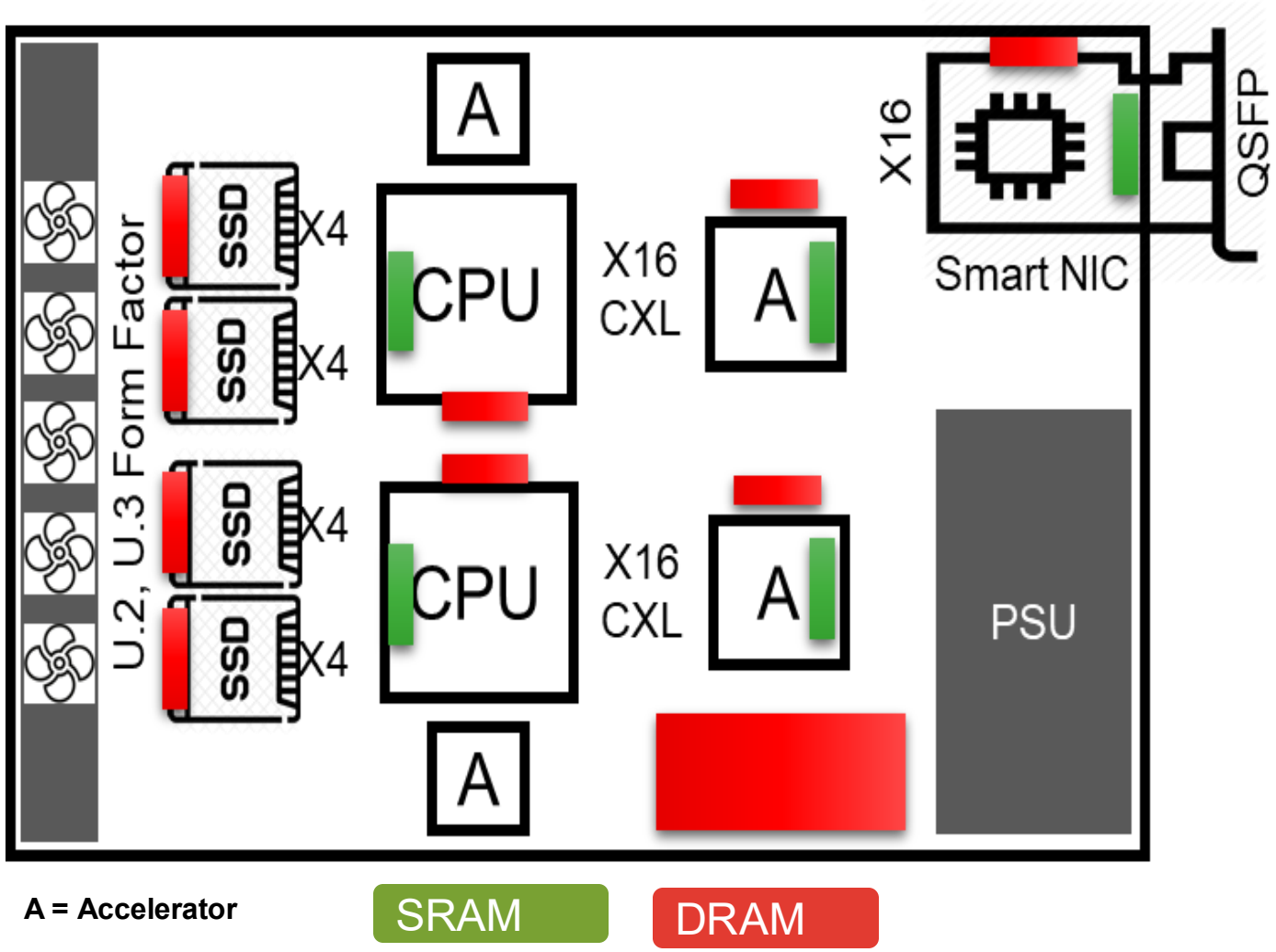
Empowering Edge AI & Mobile Innovation with LPDDR6

- LPDDR6 delivers data rates up to 14.4Gbps
- 30% Power efficiency improvement wrt LPDDR5X
- Adaptive power modes for power efficiency (Dynamic Voltage Scaling)
- Voltage Scaling: Lower IO and Core voltage for reduced energy/bit
- Reliability Enhancements: Stronger ECC and Error mitigation for AI/Automotive
- Deep Sleep and reduced standby currents for always on devices
- Optimized Architecture: Improved bank/row structure for parallelism in AI workloads



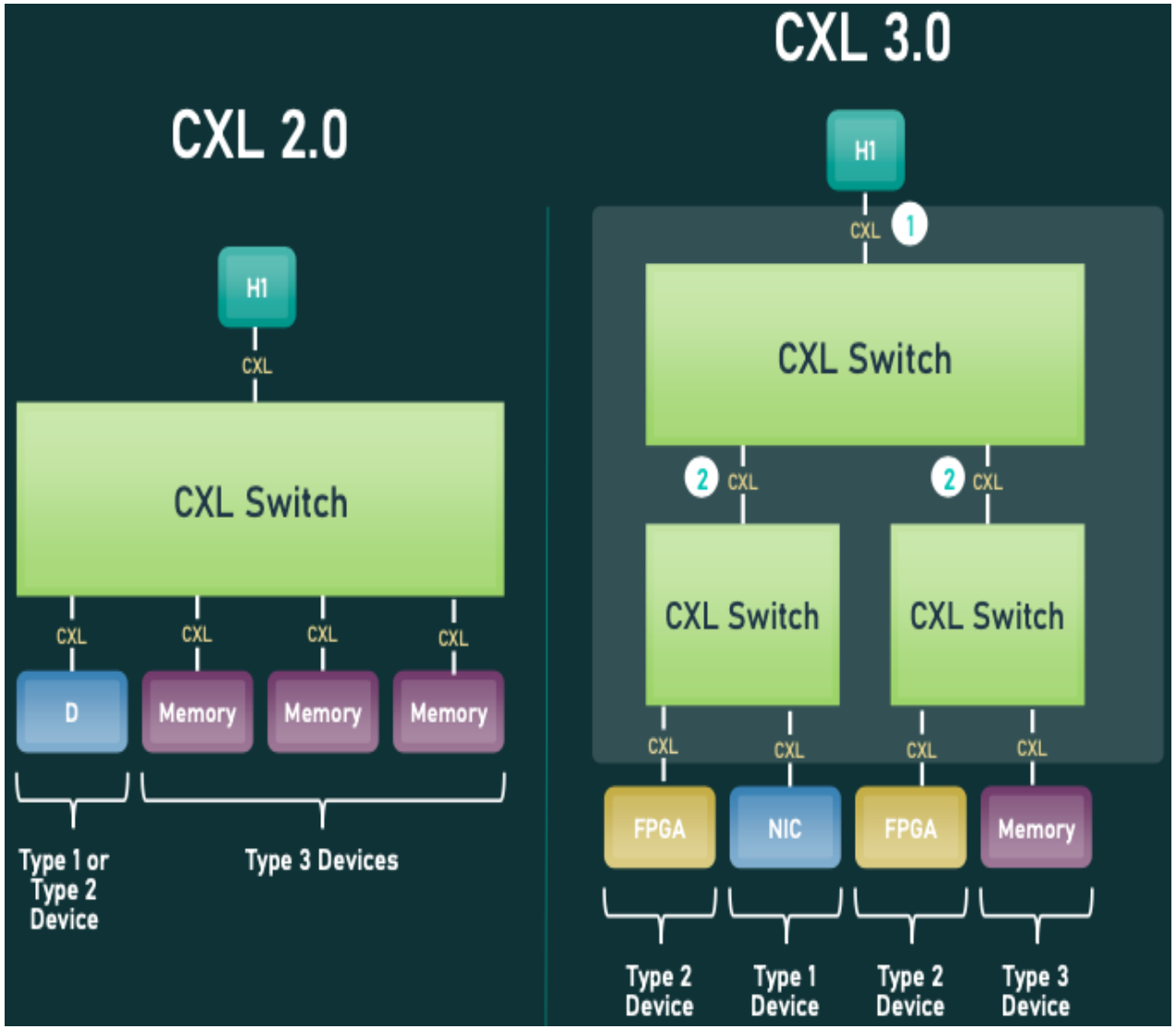
High-Performance Connectivity with PCIe

- High-speed, high-bandwidth
  - 2X speeds to meet AI/HPC requirements
- Low latency
- Scalability and flexibility
- Backward compatibility
- Low-power states



Revolutionizing Data Center Architecture with CXL

- Memory Challenges leading to CXL adoption
- CXL is addressing Memory Bandwidth bottlenecks in Servers
- Build on PCIe Infrastructure
- High-speed, high-bandwidth
- CXL.io/CXL.Cache/CXL.mem
- Optimized Latency



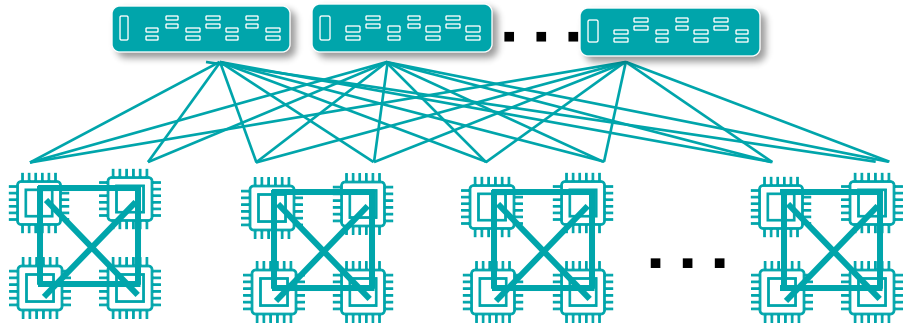
Source: Hotchips tutorial 2022

AI Scale Up architectures with UALink

- Seamless GPU-to-GPU communication
- Enables Memory sharing and Synchronization
- High Bandwidth, Light weight
- Significantly reduces CPU overhead for demanding AI workloads
- Optimized architecture for accelerated ML model training
- UALink_200 - Dedicated for AI Scale Up
  - Lightweight
  - Focused on XPU-to-XPU resource sharing & synchronization between 1,024 accelerators

ULTRA ACCELERATOR LINK™

UALink Switches



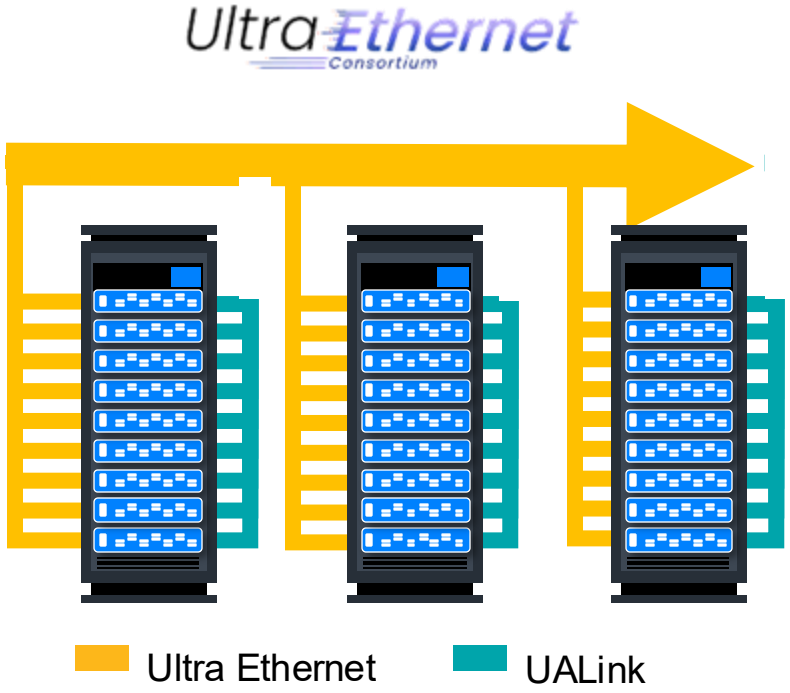
UALink

Promoter Members

AMD, Intel, Meta, HPE, AWS, Astera, Cisco, Google, Microsoft, **Synopsys,** Apple and Alibaba

AI Scale Out architectures with Ultra Ethernet

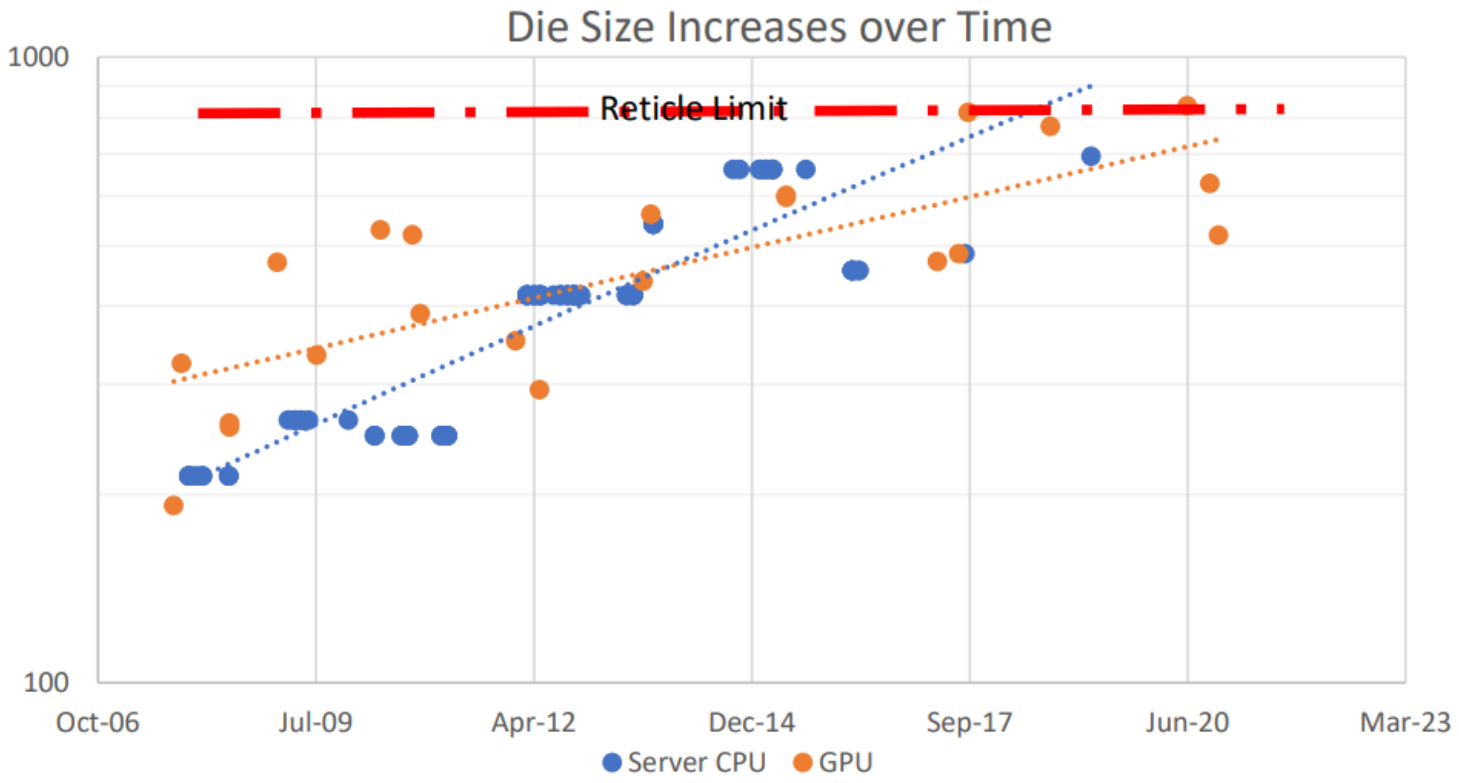
- High BW, Multi-Path, Open Standard, Ease of Configuration
- Delivers unparalleled speeds essential for advanced AI clusters
- Implements intelligent congestion control for managing intense burst traffic
- Lightweight and Low Latency
- Focused on AI workload Resource Sharing & Synchronization between 1M endpoints
- Facilitates flexible and high-performance P4-programmable packet processing



Performance Needs Beyond Reticle Limits

Networking Has Become the Bottleneck of Increasing Performance

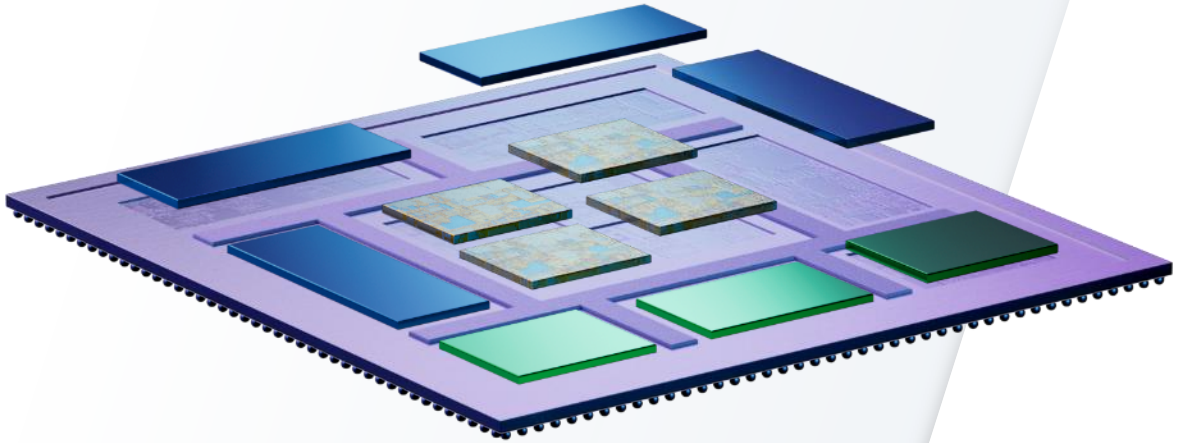
AI Accelerators / GPUs are built to reticle limit



Source: AMD Hot Chips 2022

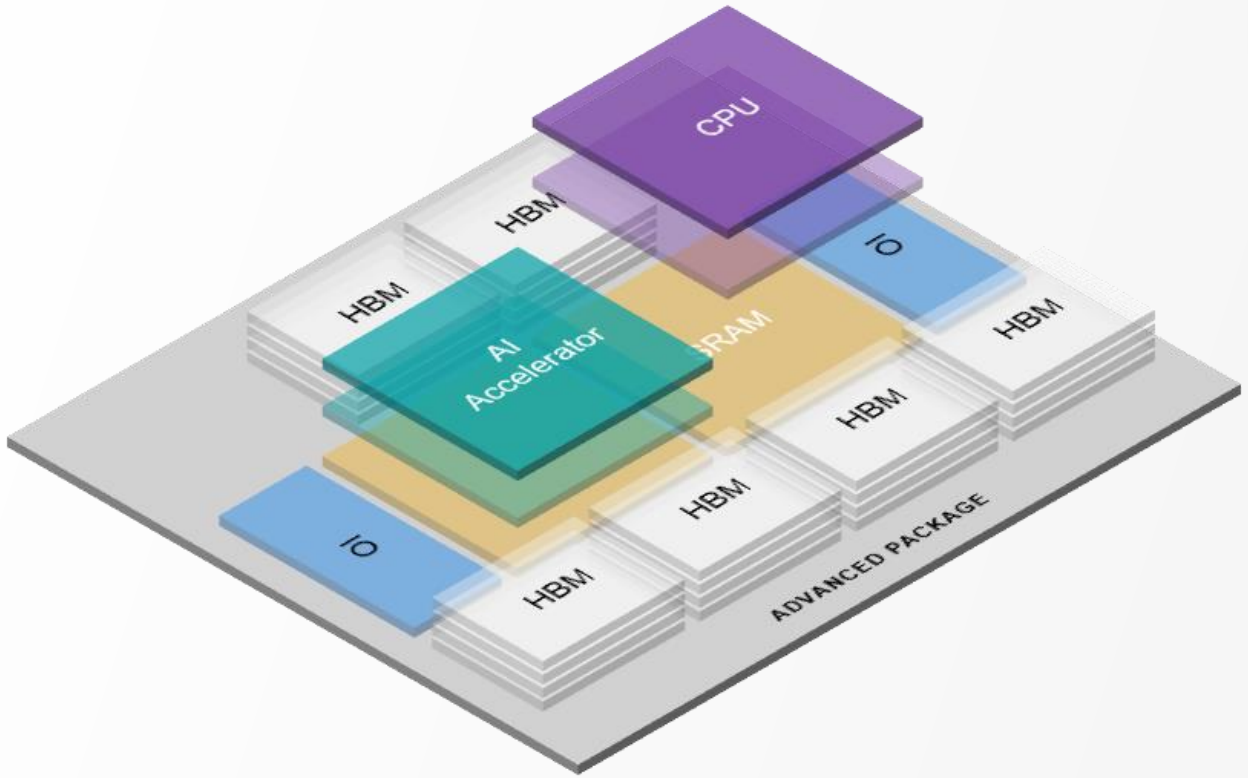
- AI Accelerators / GPUs are built to reticle limit
- XPU to XPU communication requires
  - High Radix / Density
  - Lowest Latency
  - Optimal Energy Efficiency

The Drive to Multi-Die Designs/Chiplets



- Accelerated scaling of system functionality at a cost-effective price (>2X reticle limits)
- Reduced risk & time-to-market by re-using proven designs/die
- Lower system power while increasing throughput (up to 30%)
- Rapid creation of new product variants for flexible portfolio management

Common Chiplets for AI Applications



Common Chiplets

Compute Chiplet: CPUs,GPUs, Arm CSS

AI Accelerator: Extends computational processing for AI tasks

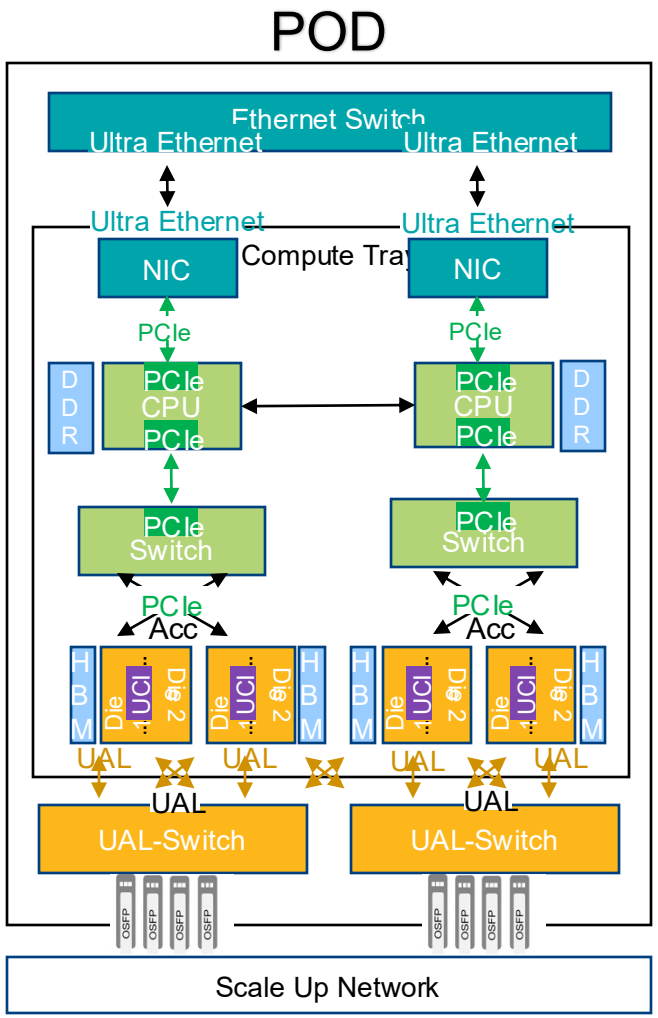
I/O Chiplet: Disaggregate PCIe, Ethernet, UAL for optimal process node & design scalability

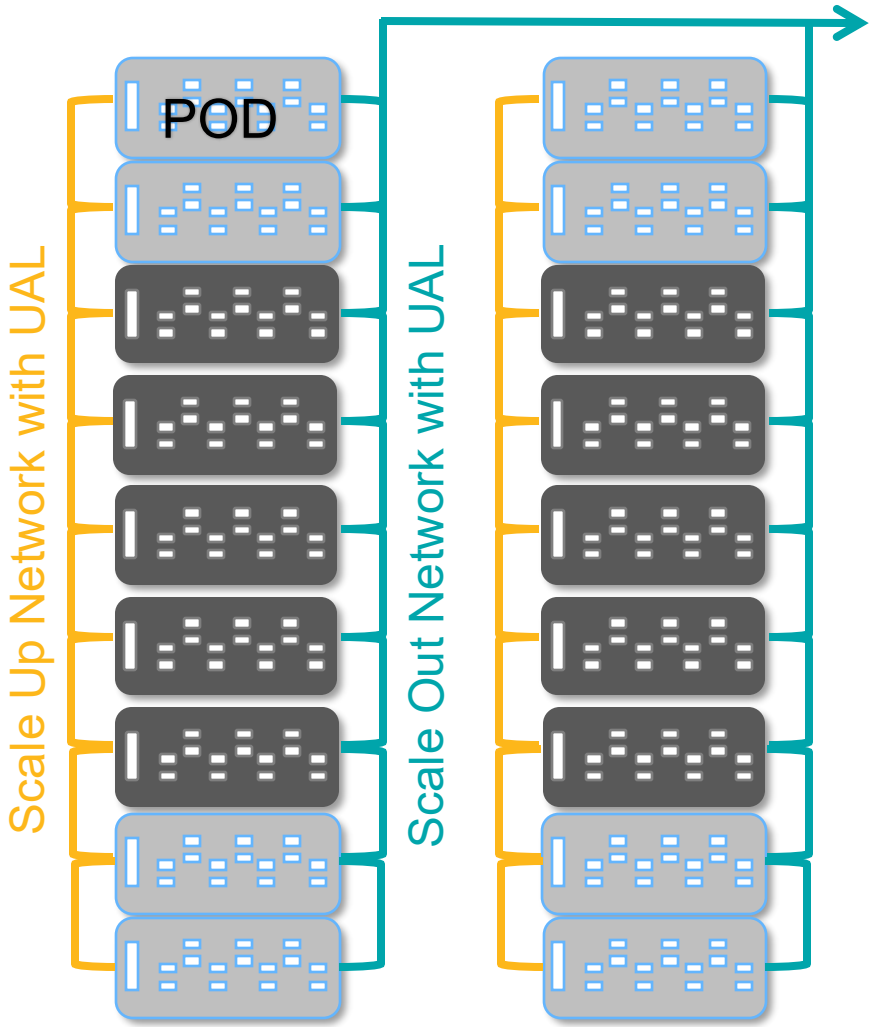
cHBM & Memory Chiplet: Disaggregate memory to increase compute resources on main die; use optimal process node for IO & memory

Chiplet designs naturally align with P4's vision for agile, high-performance networking.

AI Scaling Evolving at an Unprecedented Pace

Next-Gen Solutions for AI Scaling Architectures





PCI-Express

- PCIe 8.0 in 2027
- PCIe over Optics & Cables

Memory Interfaces

- HBM4/4E → HBM5 in 2028
- LPDDR6 @ 14.4G

Die-to-Die

- 32G → 64G in 2026
- 3D-IOs & 3D Compatible IP

UALink

- 224G → 448G in 2027
- UAL Sec for Accelerators

Ethernet (UEC)

- 224G → 448G in 2027
- Linear mode for Optics

SYNOPSYS®

Thank you