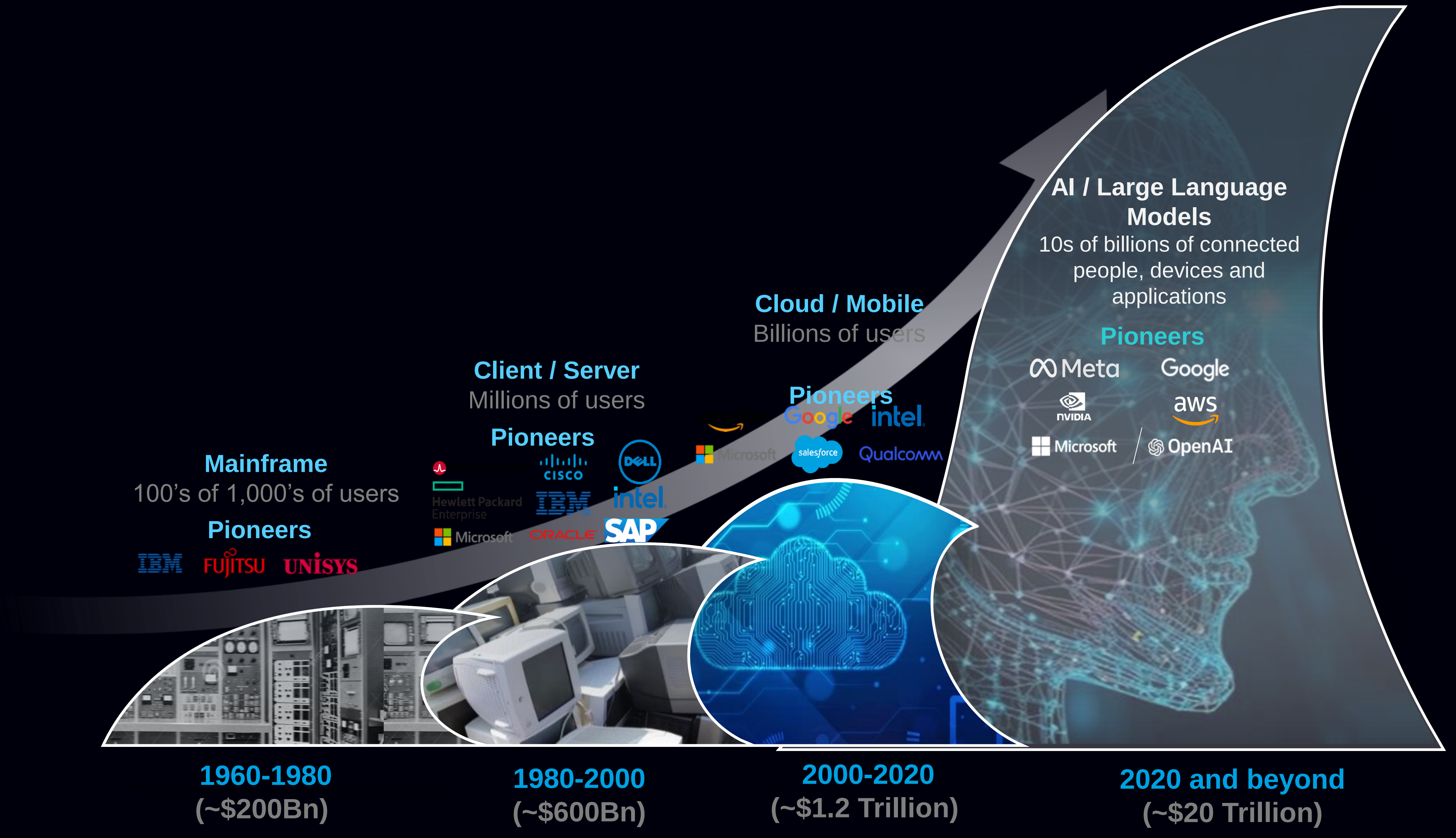
celestial AI™

Photonic Fabric™ based Scale-Up Network for Chip-to-Chip & Chip-to-Memory Connectivity

Preet Virk
Co-Founder & COO, Celestial AI

AI Hardware Summit, San Jose
11th Sep 2024

AI – The Largest Technological Wave We Have Ever Seen



Source: Morgan Stanley

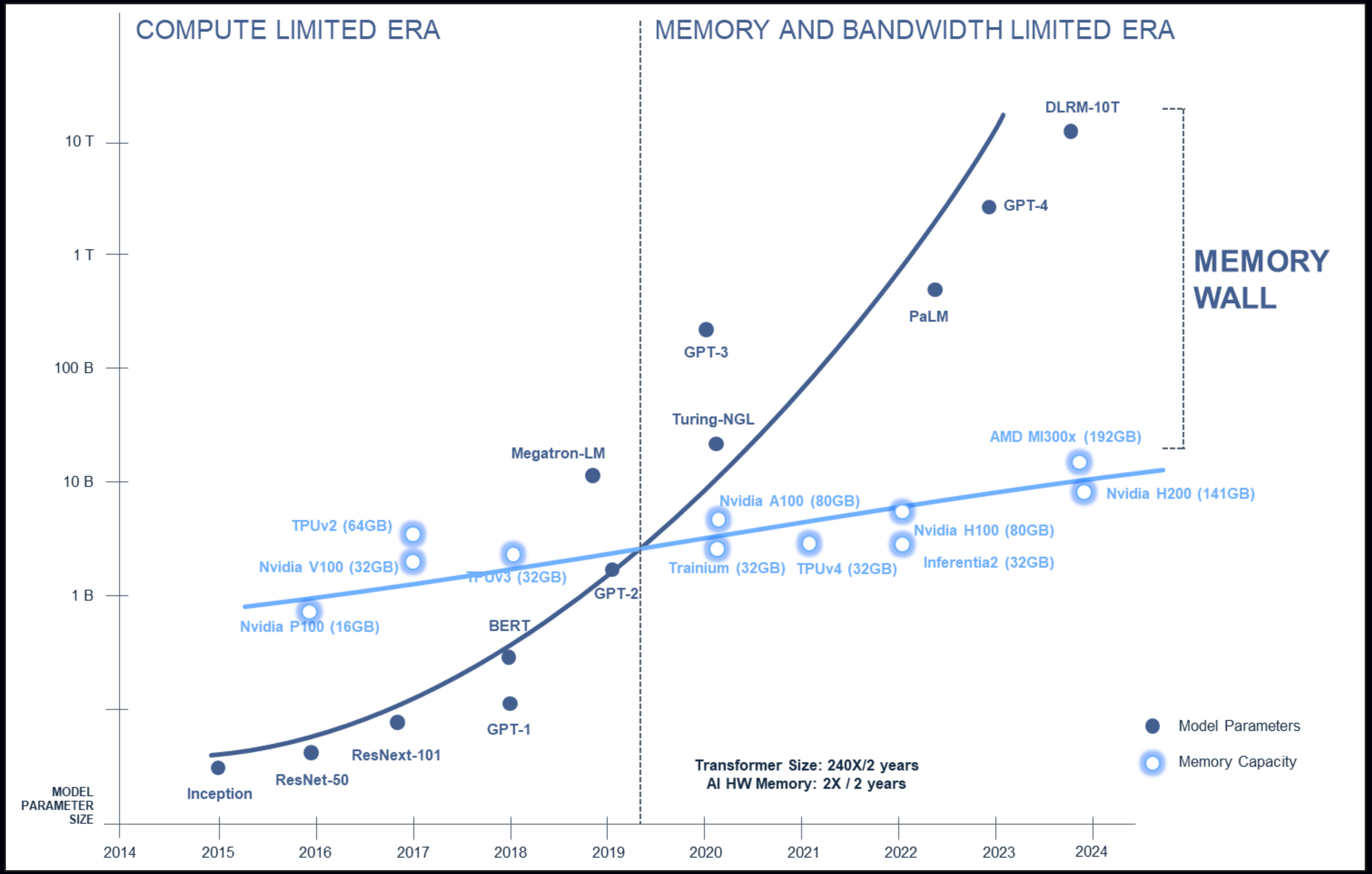
Unprecedented Growth & Market Concentration

Four Hyperscalers Represent
<u>70% of Data Center Market</u>

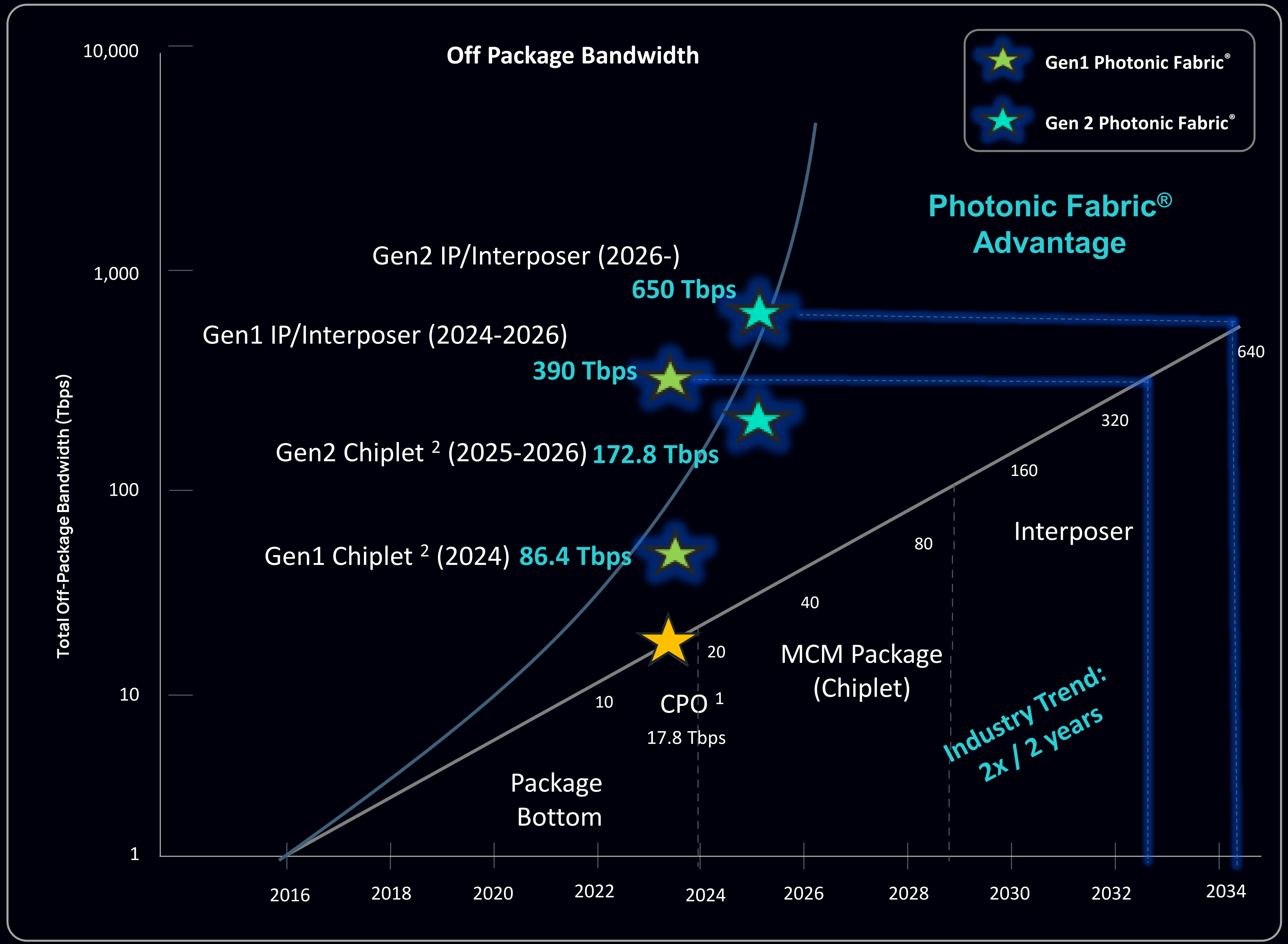
aws Meta
Google Microsoft

All are building their own AI Processors and Data Center Architectures

Artificial Intelligence is Changing Everything!



Celestial AI's Photonic Fabric - Optical Interconnect for Accelerated Computing



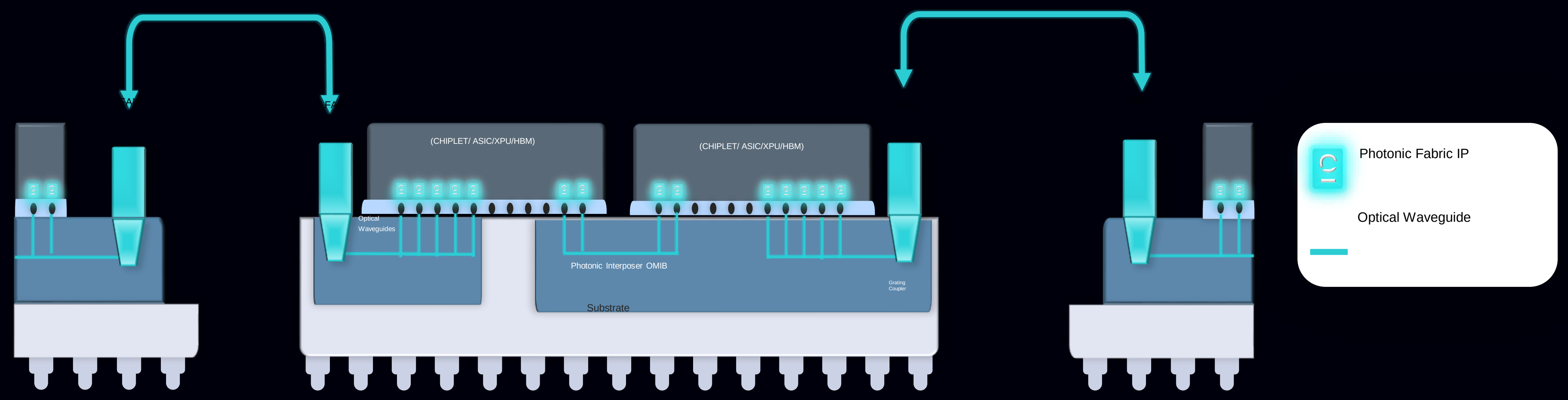
1. Industry Trend based on data by John Wilson, Nvidia Research: *"High Bandwidth Density, Energy Efficient Short Reach Signaling that Enables Massively Scalable Parallelism"*
2. Assuming 6 chiplets per XPU

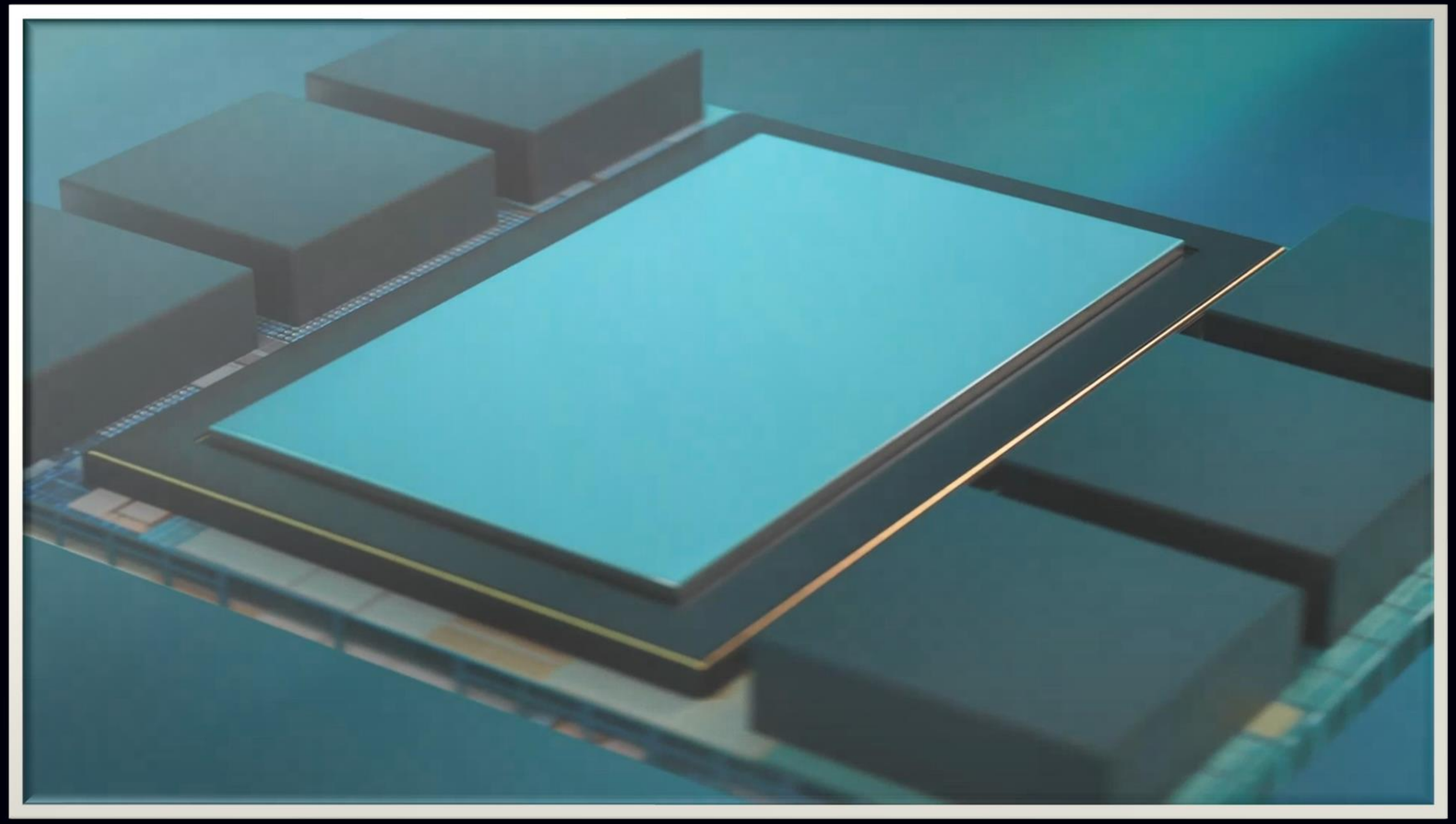
HBM3 → HBM4 Equivalent Bandwidth : The Minimum Requirement for Accelerated Computing Optical Interconnect

Scale-Up Network for XPU-XPU Connectivity: Enabling Cluster Scale Processing of AI models

AI Driving a New Generation of Optical Compute Interconnect

Revolutionary Integration of Silicon Photonics for Accelerated Computing





Compact, Thermally Stable Optical Modulator

✦Chip-to-chip packaging with XPUs dissipating 100s of watts

No DSP – Linear Drive Optics

✦High SNR, low BER, close proximity of optics to electronics: Eliminates need for DSP

Multiple Packaging Options Tailored for Customer Applications

✦Integration of Advanced CMOS with Si Photonics

Full Stack E-O-E Link Optimization

✦Protocol adaptive Network Convergence Layer (NCL)

✦Full Electrical to Optical to Electrical (EOE) link management

FEC, CRC, FLIT Replay

celestial AI

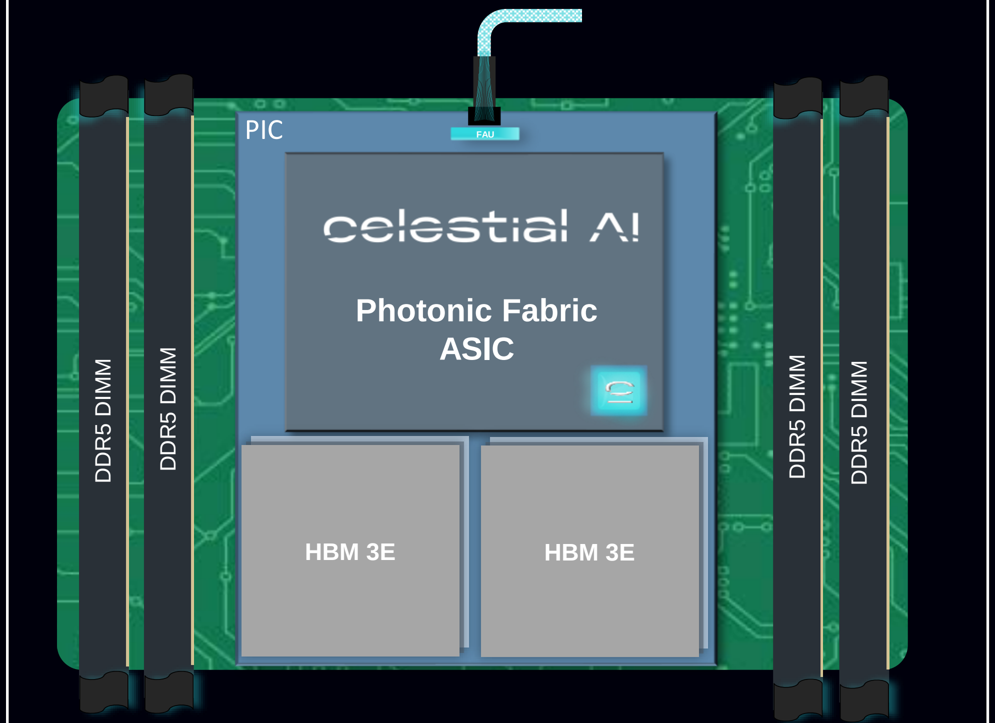
Why Photonic Fabric™ vs. Copper?

Higher Off-Package Bandwidth	100s of Tbps BW, unrestricted by beachfront
Efficient Data Movement	Zero mass photons vs I^2R losses from electrons
Lower Latency & Power	Eliminate DSPs, Deep FEC and re-timers
Longer Reach	Connect multiple racks with less power
More efficient remote memory transactions(RDMA)	10pJ/bit (incl. 2.4 pJ/bit for optics) for up to ~50m 60 pJ/bit for copper for up to ~ 1m
Fewer System Connections	Higher overall system reliability

Compute–Compute & Compute–Memory Photonic Connectivity for Accelerated Computing

celestial AI

Photonic Fabric™ Link: Module & Appliance



Photonic Fabric Module

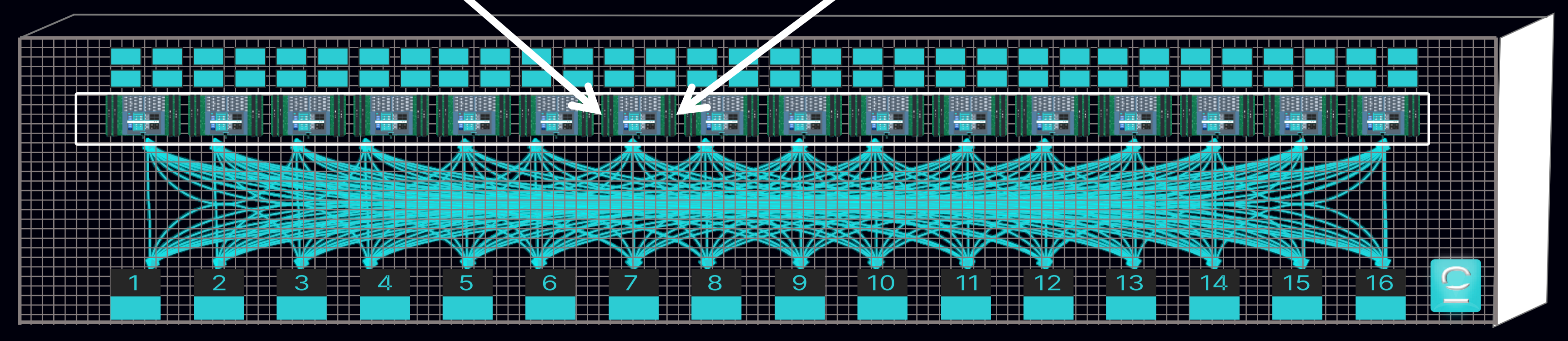
Photonic Fabric Module

Celestial AI Designed *System in Package (SIP)*

- ✦Memory Controllers + Photonic Fabric™ Link + Network Switch

2.07 TB Memory Capacity at 7.2 Tbps Bandwidth with ~ 100ns Latency

- ✦HBM3E Operates as Write-through Cache for DDR
- ✦Bandwidth & Latency of HBM3E with Capacity & Cost of DDR5



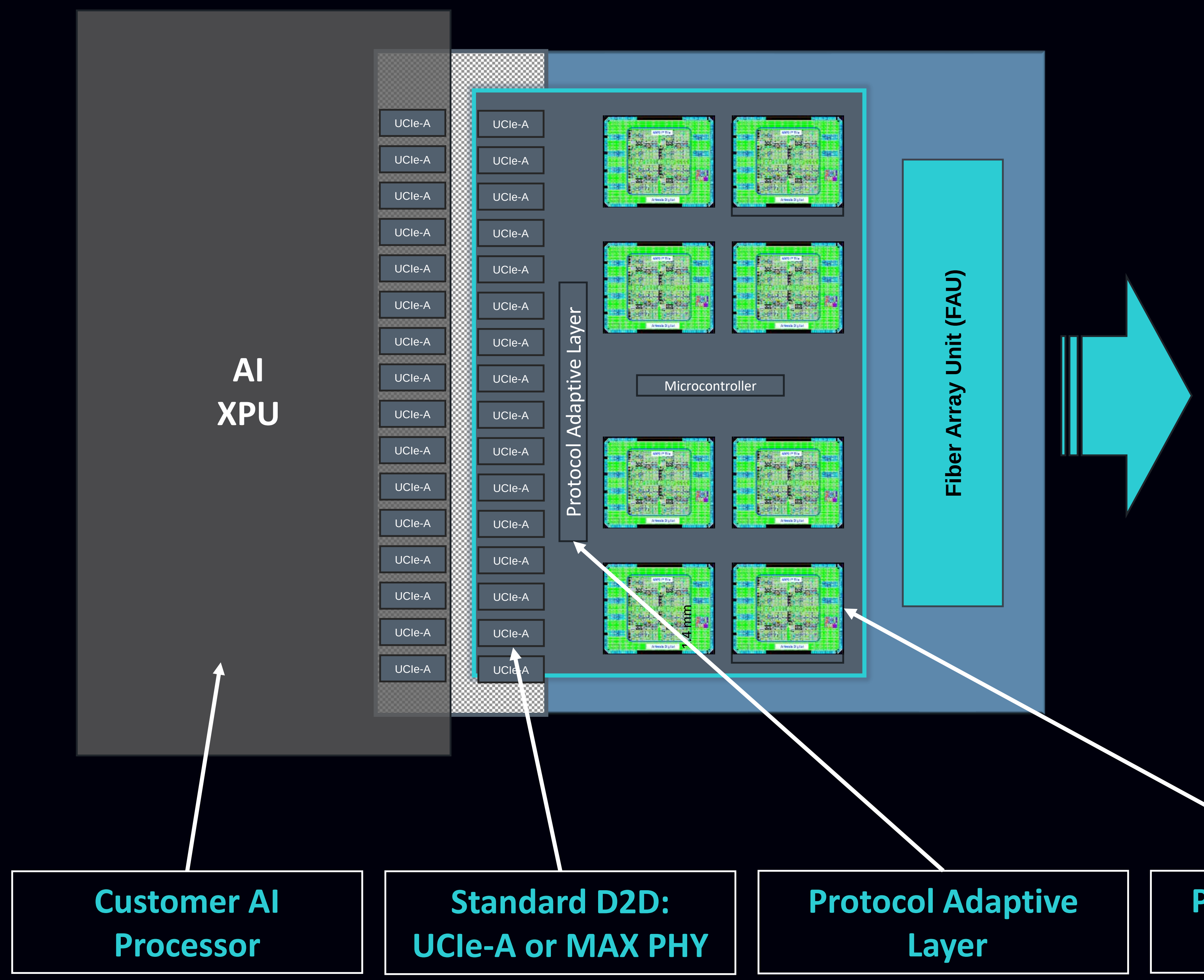
Photonic Fabric Appliance

Photonic Fabric Appliance

- ✦16x Photonic Fabric Memory Modules in a 2U Appliance
- ✦33TB Memory Capacity
- ✦115Tbps Network Switch Enabling a Backend/Scale-Up AI Fabric

Shattering the Memory & Interconnect Bandwidth Wall for Accelerated Computing

Photonic Fabric™ Link Implementation: The Chiplet Approach



14.4 Tbps (Gen1) Per Chiplet
28.8 Tbps (Gen2) Per Chiplet
> Full HBM3E bandwidth

3.6X more bandwidth than CPO chiplet
10x Gen1 chiplets per XPU package = 144 Tbps
10x Gen2 chiplets per XPU package = 288 Tbps

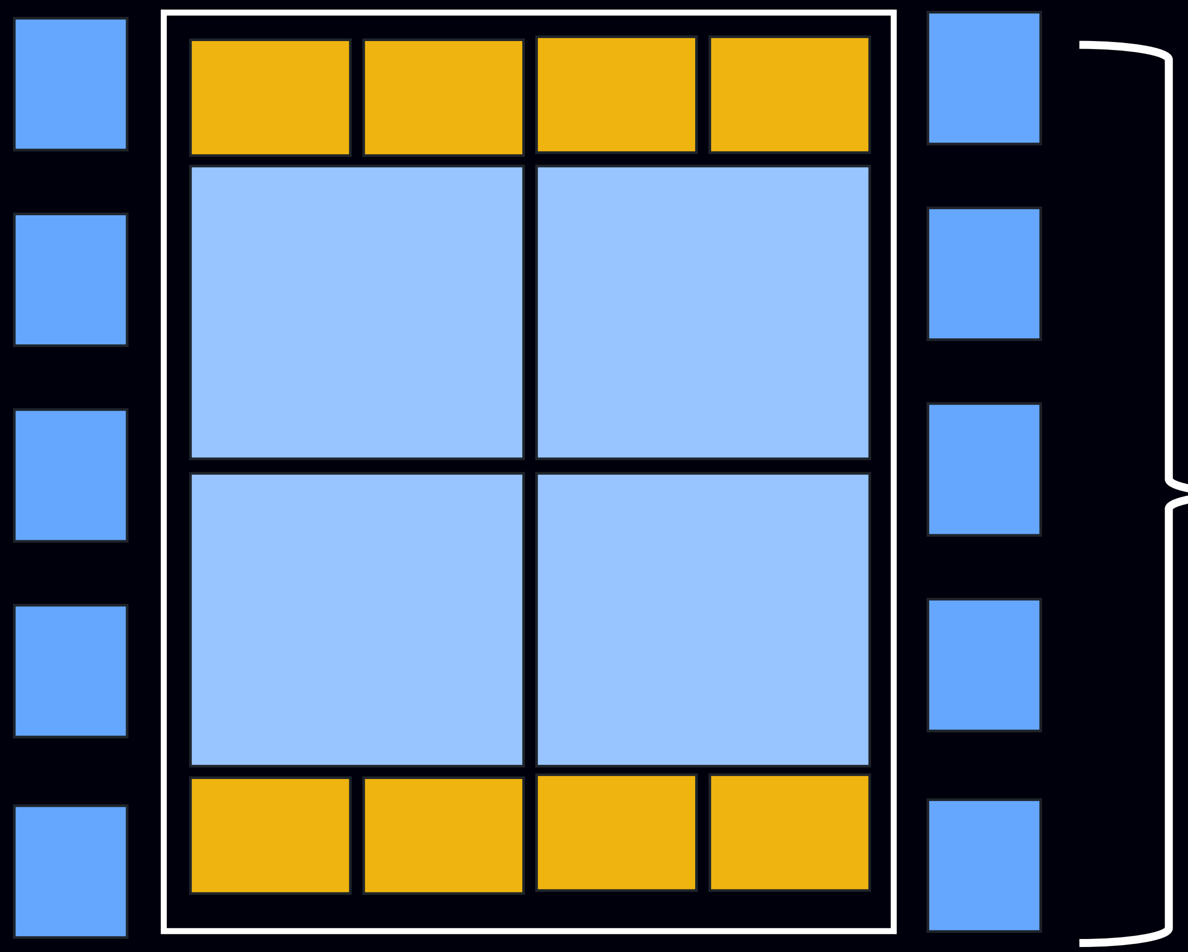
Protocol Adaptive :
AXI, HBM/DDR, UAL, CXL etc.

2.4 pJ/bit photonic link power (Gen1)

Full Stack High-Bandwidth Optical Interconnect Solution for Accelerated Computing

Photonic Fabric Delivers Superior Off-Package Bandwidth

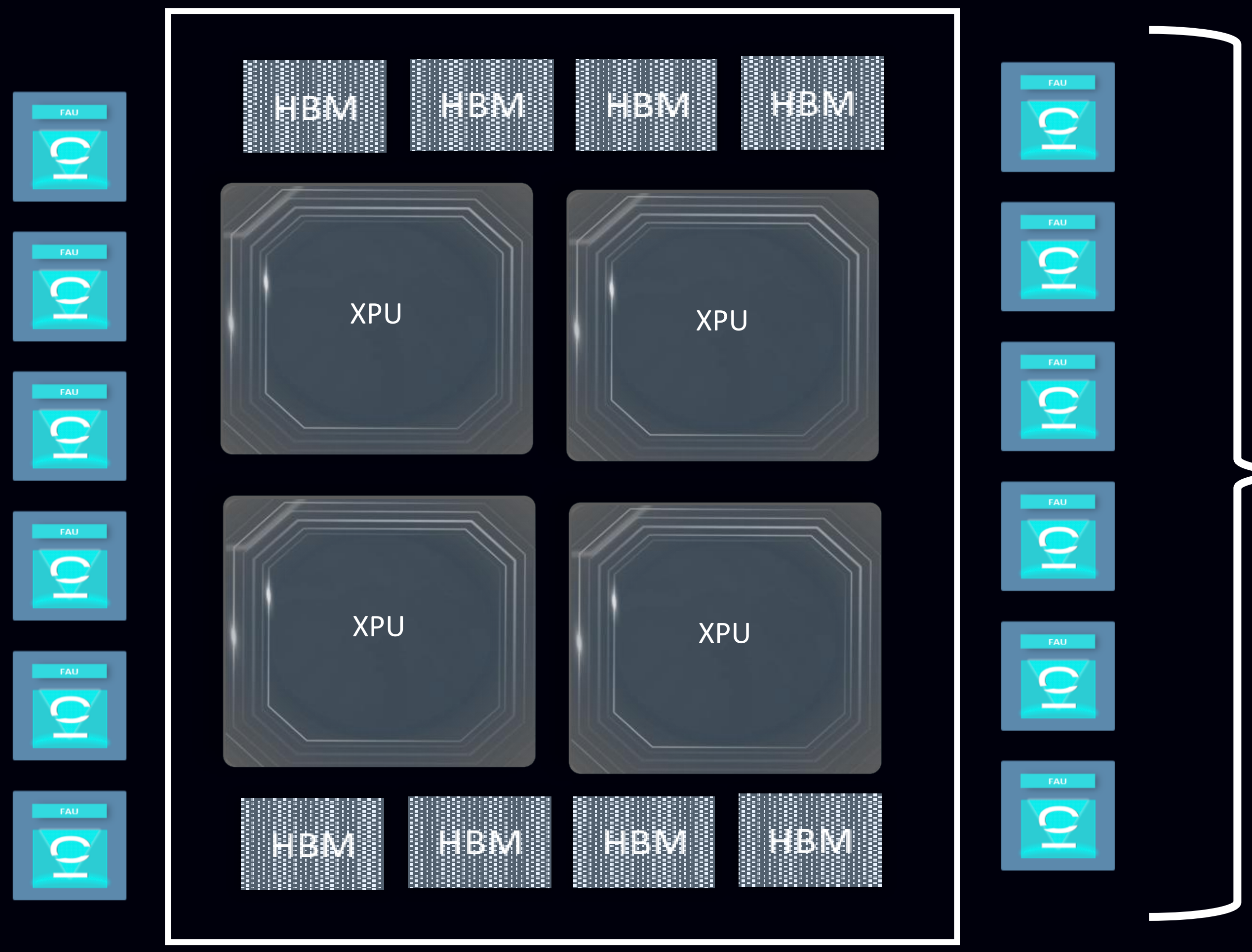
Current SoTA



4 Tbps per Chiplet
40 Tbps per Package

4x reticle package with 10 chiplets

Photonic Fabric



14.4 Tbps (Gen1) /
28.8 Tbps (Gen2)
per chiplet

144 Tbps (Gen1)/
288 Tbps (Gen2)
per package

4x reticle package with 10 chiplets

Photonic Fabric Based Chiplets Offer More Package Bandwidth With Full Link Management

Photonic Fabric™ 14.4 Tbps Photonic Interface Chiplet

Less beachfront than 1 HBM stack

Standard 2.5D packaging flows from multiple large OSATs

Automated high-volume, high-throughput fiber-attach process

Standard D2D Interface

- **UCIe**
- **MAX PHY**

Unlocks Photonic Connectivity (E-O-E) for AI XPUs

- **Compute-to-Compute**
  - ✦Scale-up/backend networks
- **Compute-to-Memory**
  - ✦Photonically scalable disaggregated memory
- **Full HBM3E bandwidth**

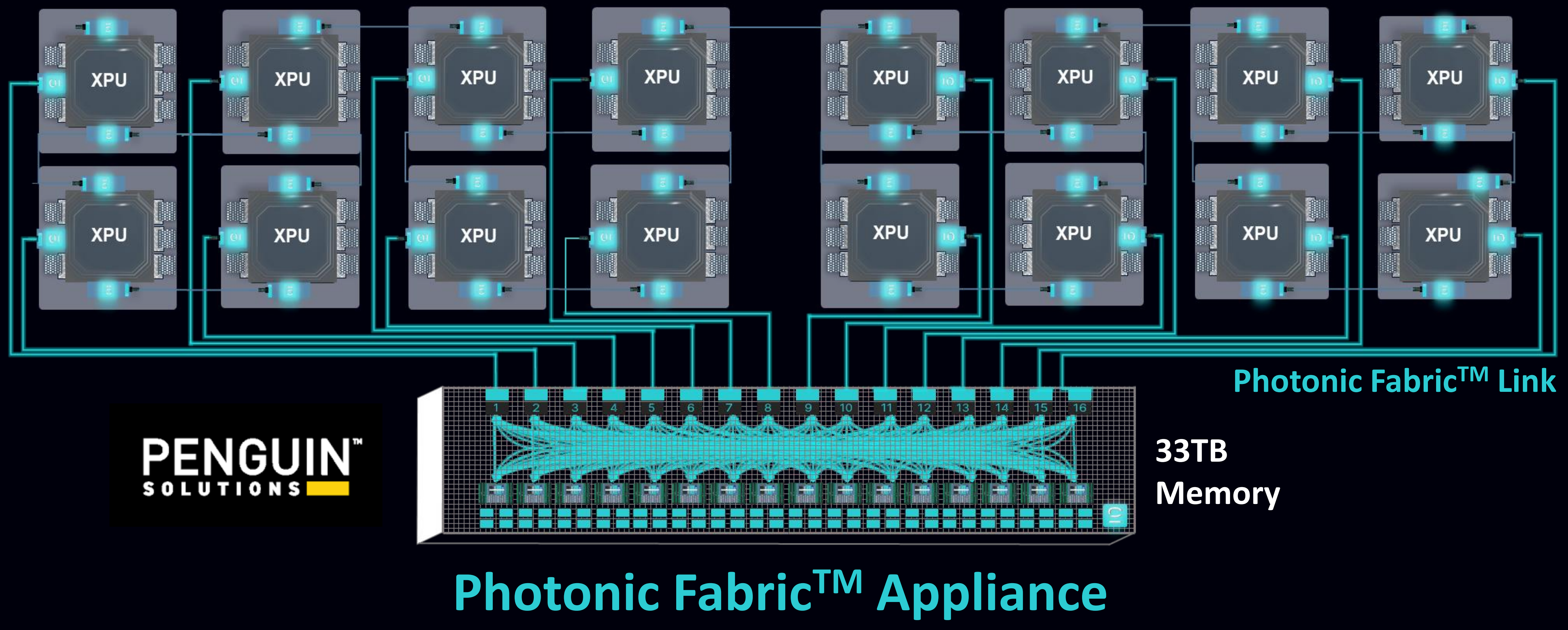
Protocol Adaptive

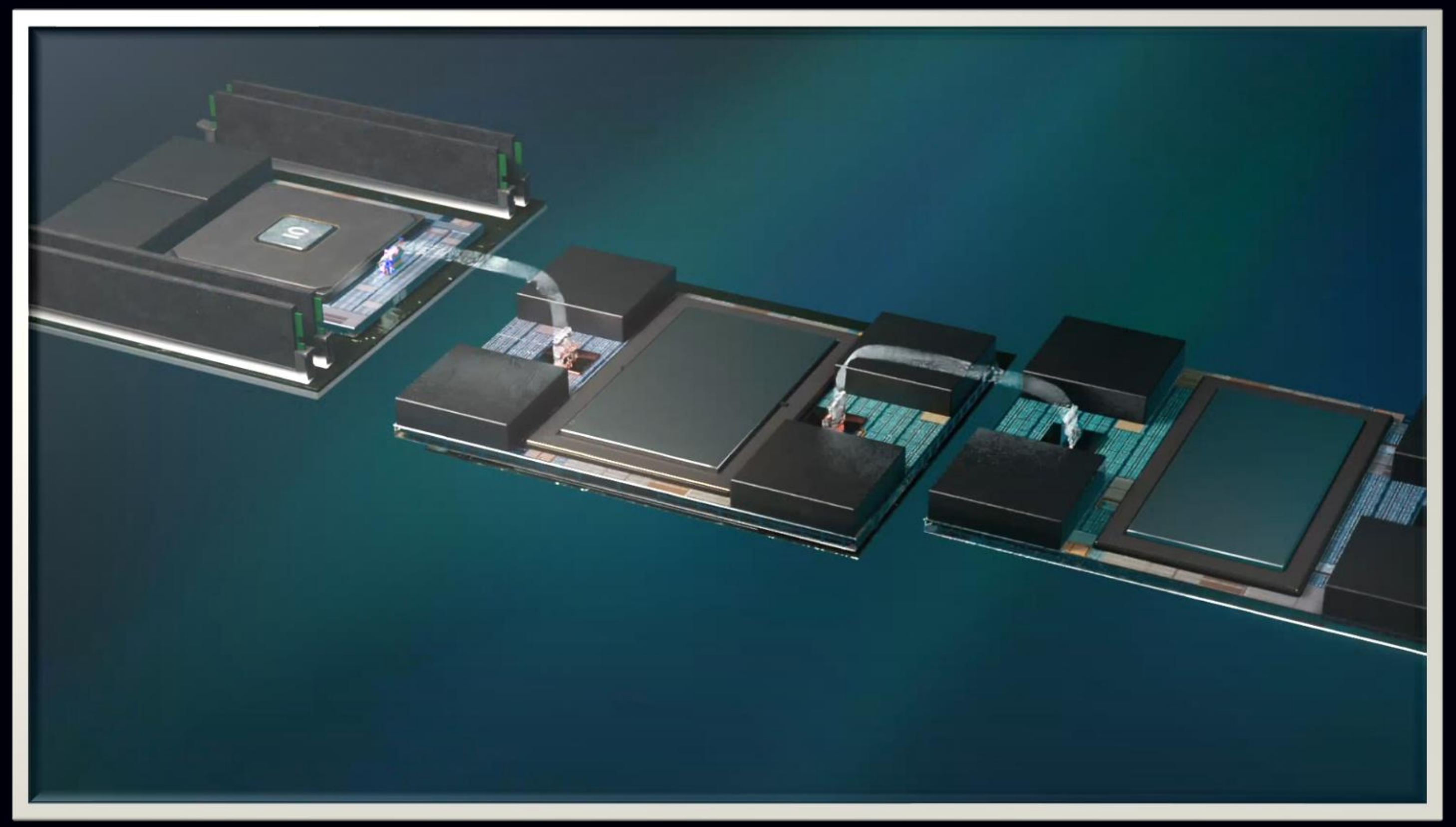
- **Standard Protocols**
  - ✦AXI, HBM/DDR, CXL
- **Emerging Protocols**
  - ✦UAL
- **Proprietary Protocols**

Seamless Integration with Existing AI Accelerators & XPUs

celestial AI

Compute-to-Compute & Compute-to-Memory Photonic Connectivity



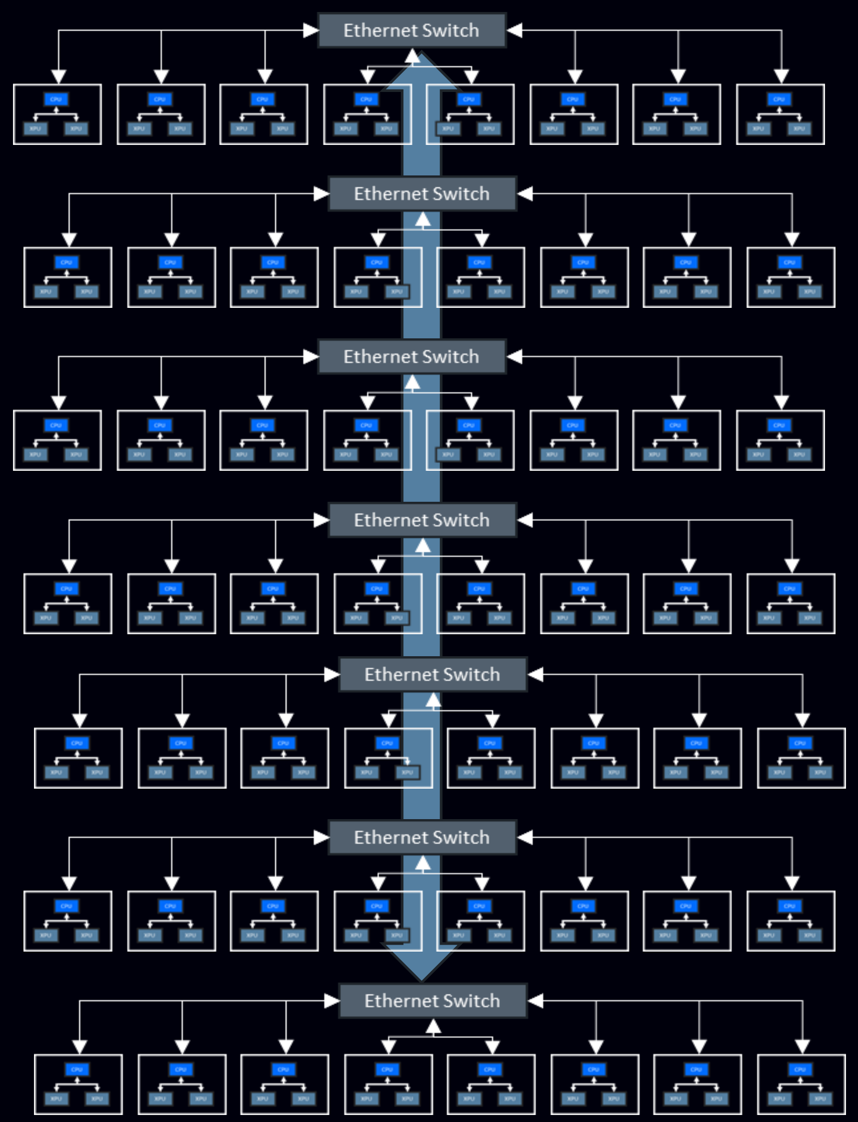


- 115 Tbps Switch for backend/scale-up network
- All-to-All Connectivity for efficient Collective Comms
- 33TB Unified/Shared Memory Space
- Broadcast & Reduce across all connected XPUs

Photonic Fabric™ Technology Platform Enables Cluster Scale AI Processing

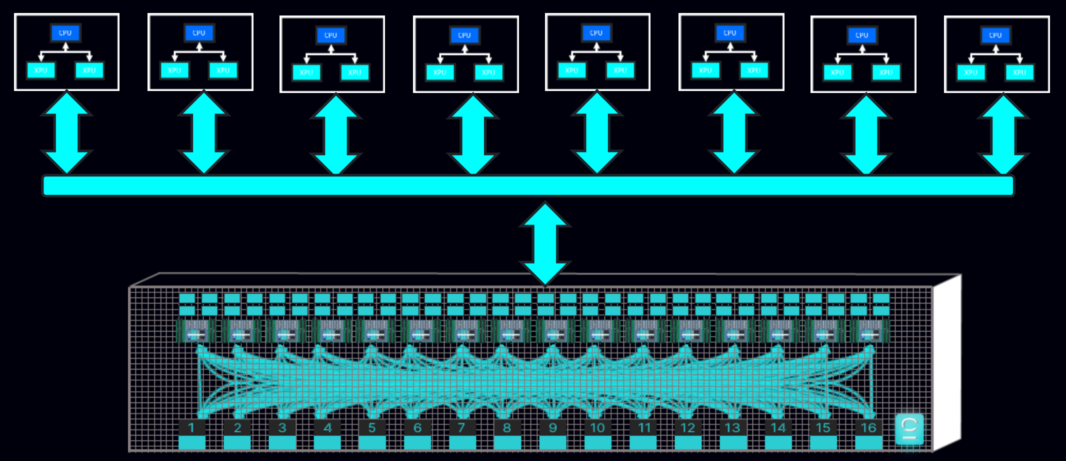
Photonic Fabric Link & Appliance: AI Efficiency & Performance

56 Conventional XPU with 192GB each of HBM3 required to process a 10T parameter DLRM model



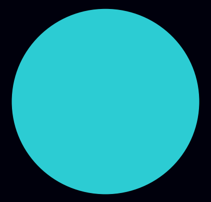
Each conventional XPU holds 1/56th (1.79%) of the 10T model in its HBM memory (192GB)

16 XPU-Optical connected to Photonic Fabric Appliance to process a 10T parameter DLRM model



- 115Tbps Switch
- 33TB Memory Capacity
- Full HBM3 Bandwidth

✦Up to 71% XPU CapEx and Power Reduction
✦Higher Compute Density
✦Memory Resources Scalable Independent of Compute
✦12.5X DLRM Performance Speed-up



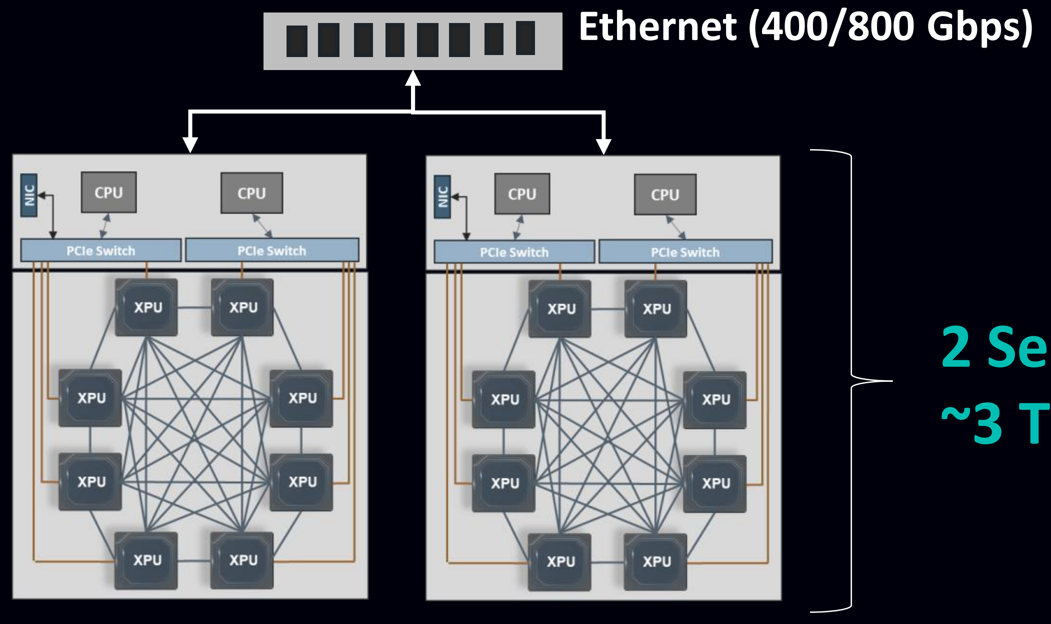
Each XPU-Optical has access to the full model stored in the PFMA

Photonic Fabric Link & Appliance Delivers Compelling Benefits for AI Workloads

celestial AI

Photonic Memory Fabric: Delivers Higher Throughput on GPT4

Conventional 16-XPU/2 Server Rack

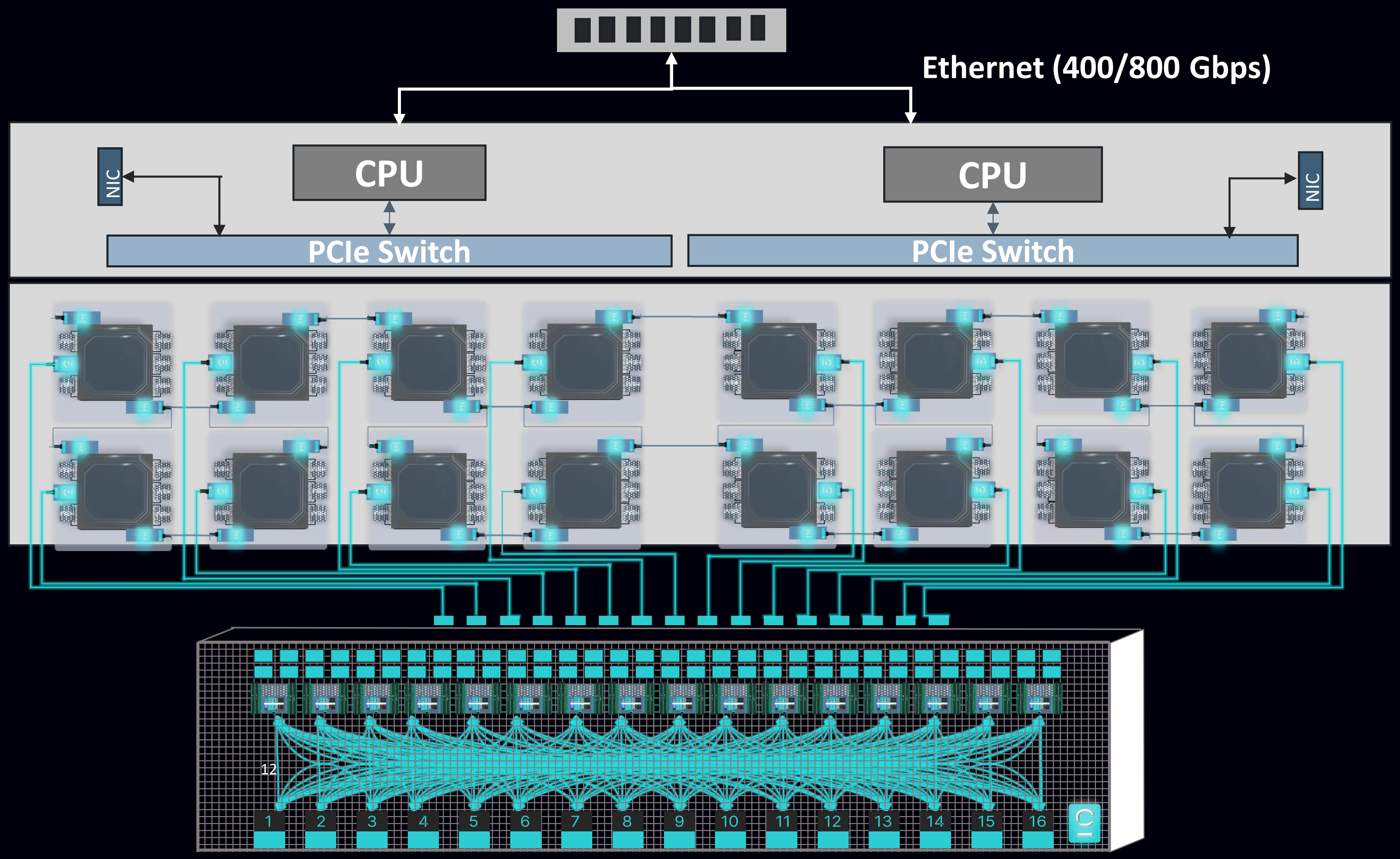


2 Servers :
~3 TB of HBM

Scale-up / Backend Network Constrained

16-XPU/2-Server + Photonic Fabric Appliance

36 TB (12x) More Fast Memory + **7.2 Tbps** Scale-Up Network per XPU (Backend Network)



- 115Tbps Switch
- 33TB Memory Capacity
- Full HBM3 Bandwidth

Memory Restricted: GPT4 (1.8T) MOE models

~3TB

1.8TB | 1.2T

GPT4-1.8T MoE
Model Weights (fp8)

Memory Available for Inferencing
(KV Cache, Context Lengths , Batch Size)

Easily serve GPT4 (1.8T) MOE models with large context length and batch sizes

36TB

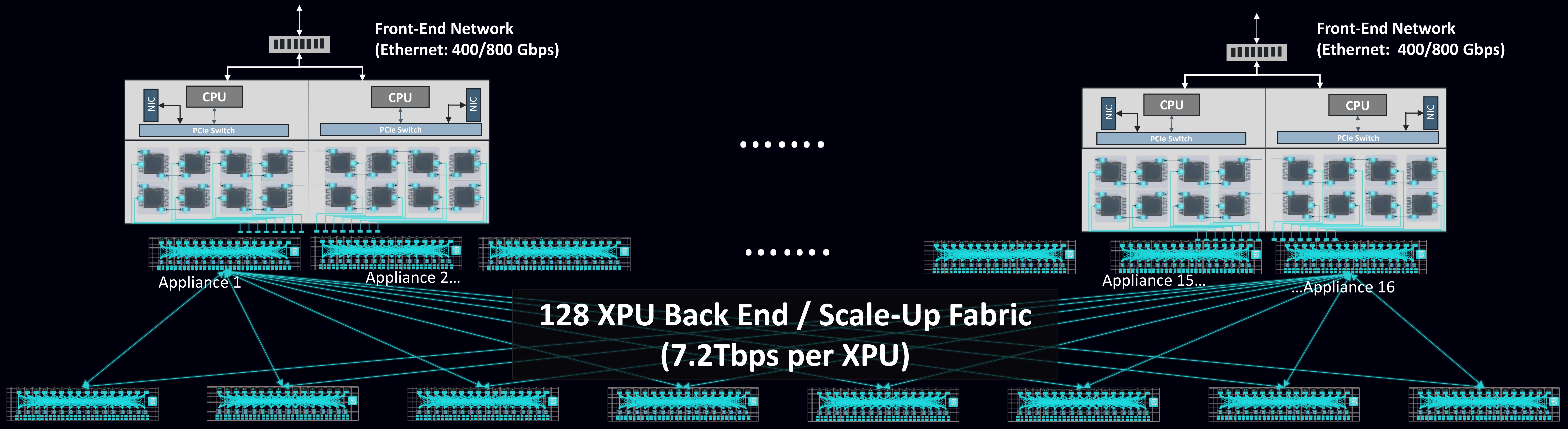
1.8TB | 34.2TB

GPT4-1.8T MoE
Model Weights (fp8)

Memory Available for Inferencing
(KV Cache, Context Lengths , Batch Size)

celestial AI

128 XPU Cluster Scale Photonic Fabric for Accelerated Computing

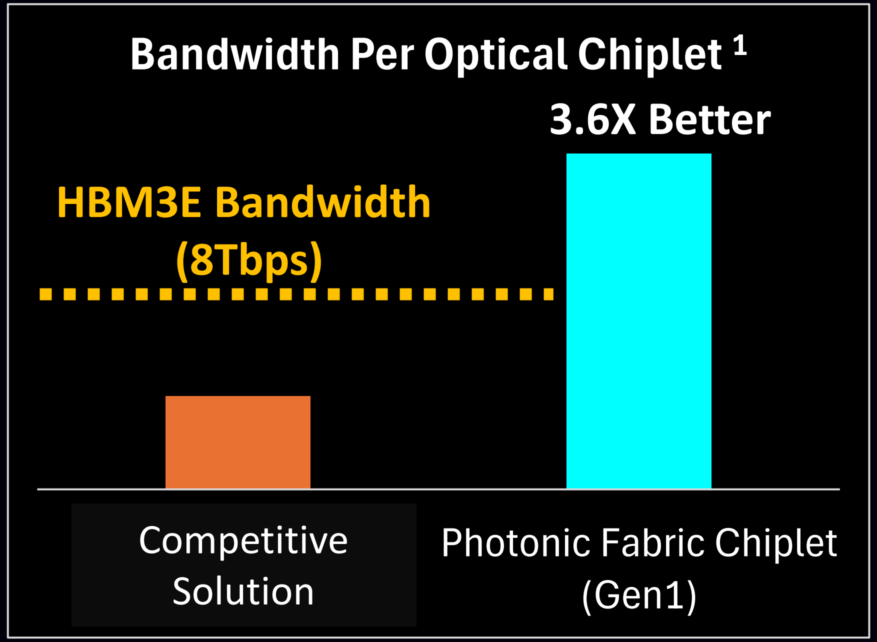


Illustrative Picture
Not all photonic links shown

Up to 792 TB System Memory Capacity

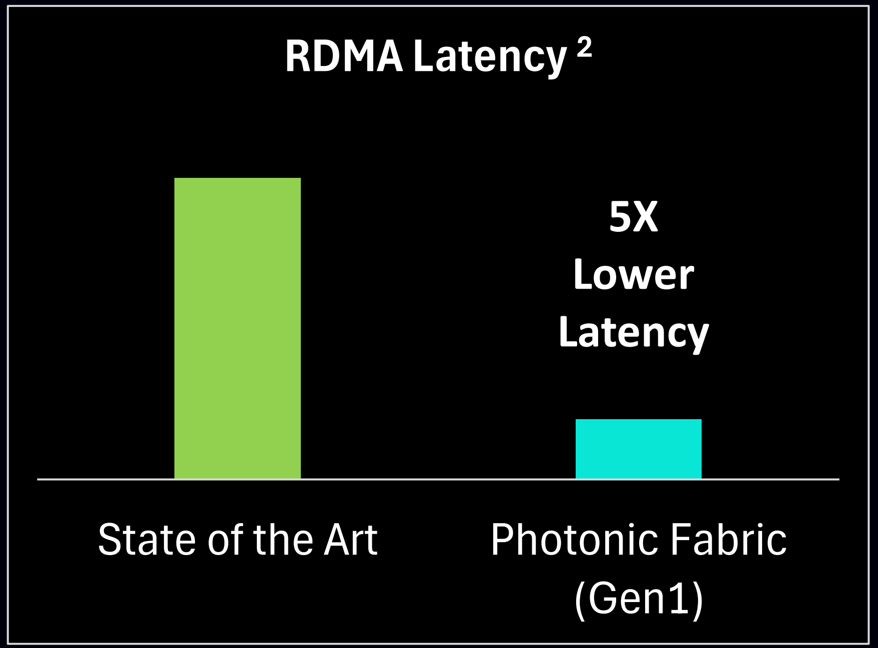
Enabling Very Large Clusters With Significantly Lower Carbon/Energy (TCO_2) Impact

Celestial AI Photonic Fabric™ vs. Current State-of-the-Art



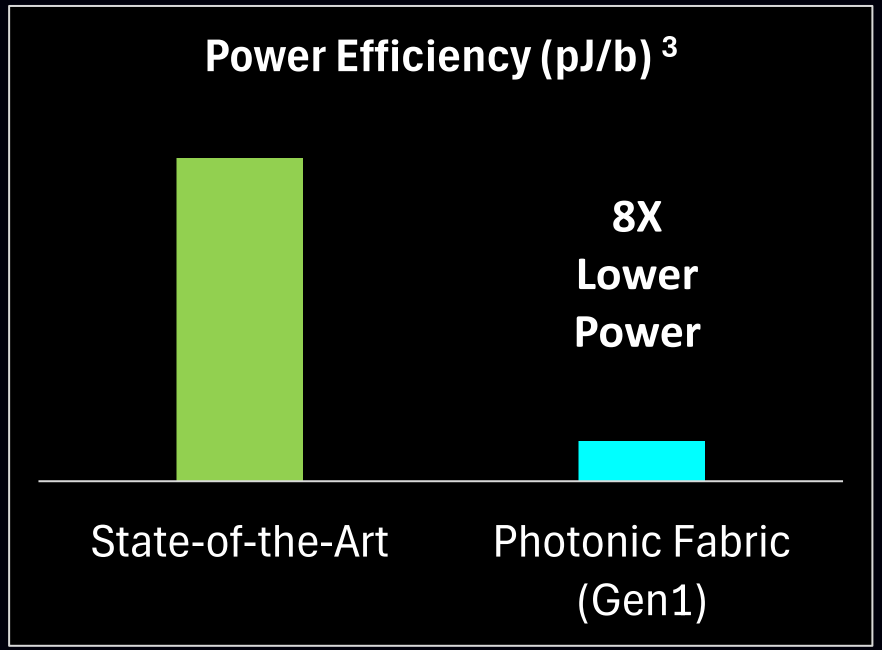
Photonic Fabric offers 16x off-package bandwidth

[1] SOTA Benchmarks: Ayar Labs TeraPHY



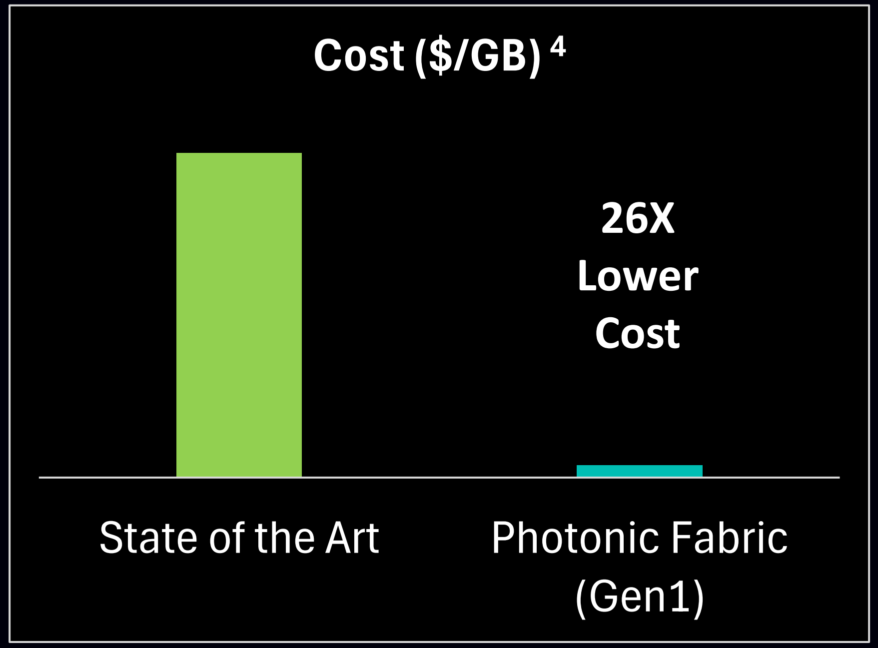
Photonic Fabric offers 5X better latency

[2] SOTA Benchmark: DGX H200 Remote Direct Memory Access (RDMA)



Photonic Fabric offers 8X better power efficiency

[3] SOTA Benchmark: RDMA Power for Fourth-generation NVLink



Photonic Fabric offers 26X better cost efficiency

[4] SOTA Benchmark: NVIDIA DGX H200

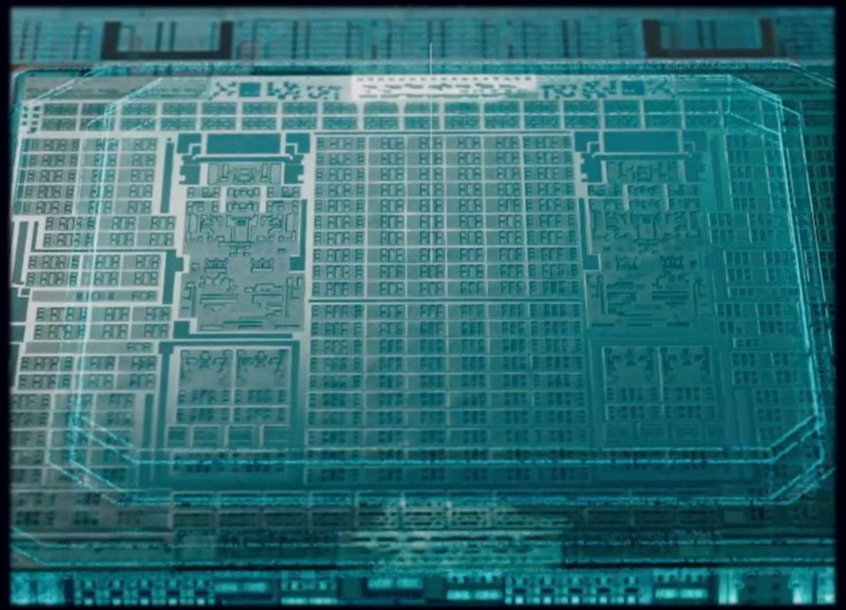
Photonic Fabric Delivers Disruptive Performance, Lowest Latency, Improved Energy Efficiency for Data Movement & Compute and Lower TCO

The Photonic Fabric™ – Designed for Volume Deployment

Silicon Photonics

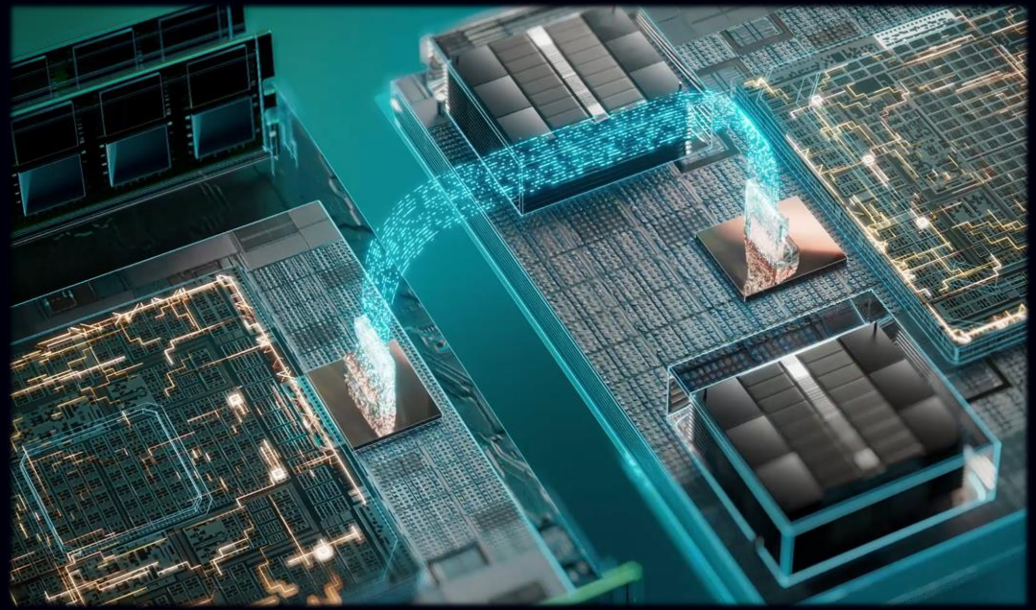
Volume Manufacturing Si Photonics
Driven by Data Communications

Advanced CMOS



Silicon Photonics Control Circuitry,
SERDES, Router/Switch

Opto-Electronic Systems-in-Package



Wafer-Scale Assembly & Test
Leveraging Established Supply Chains

The Right Technology at the Right Time – Full-Stack Optical Interconnectivity Platform

celestial AI

How Does the Photonic Fabric Transform AI?

Democratizes AI

- Lowers cost of AI

Efficiently Process Very Large LLMs

- Train and Serve 1T+ Models like :
  - ✦ Open AI GPT4-1.8T MoE
  - ✦ Anthropic Claude Opus (2T)

Support Larger Context Lengths

- Larger Context Lengths activate new use cases

Accelerates Multi-Modality

- Text-To-Image
- Text-To-Video
  - ✦ Open AI Sora
  - ✦ RunwayML

Simplify AI Software Stack

- All-to-All/Broadcast Interconnect simplifies collective comms
- Reduces needs for sharding

Mitigate AI Carbon Impact

- Helps meet carbon neutrality targets even with increasing AI usage

A New Vision of AI infrastructure Unconstrained By Fast Memory Capacity & Scale-Up Bandwidth

celestial AI

info@celestial.ai

www.celestial.ai

FAST COMPANY

100 BEST WORKPLACES FOR INNOVATORS 2024

celestial AI