2026 OCP GLOBAL SUMMIT

Designing the Next AI Fabric:

When compute meets optics

Celestica™



Tareq Bustami,
VP Market Technology

BROADCOM



Manish Mehta,
VP of Marketing and Operations

OpenAI



Binbin Guan
Member of Technical Staff

AI at Work

AI adoption is growing alongside gains on selected tasks.

OpenAI

ChatGPT weekly users

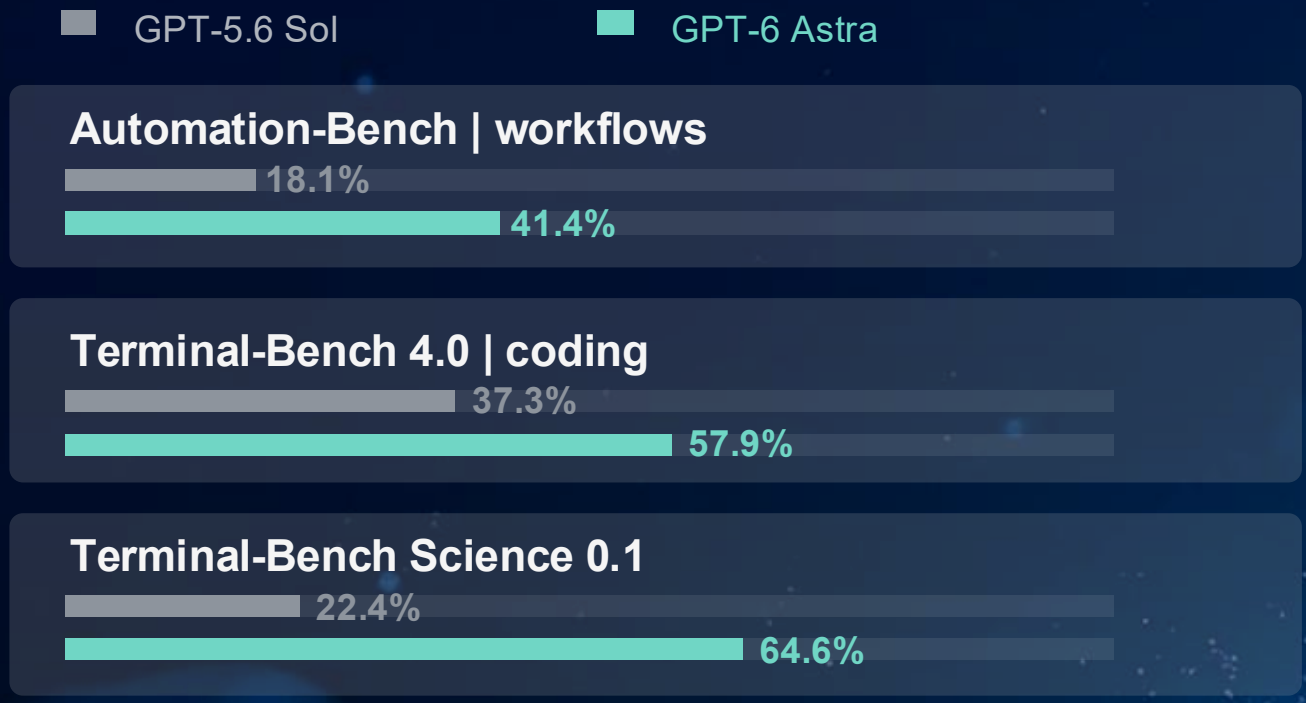


Codex weekly users



Publication dates; separate user scales.

Astra: selected benchmark scores



Higher is better. Maximum across effort settings; not an equal-cost comparison. OpenAI, 3 Sep 2026.

75% of surveyed workers reported better output speed or quality with AI

Metrics that Matter: Latency and Energy

OpenAI

01 · USER EXPERIENCE

Request latency

Time to complete an inference request

TTLT · time to last token

End-to-end latency; TBT is another useful proxy

02 · INFERENCE EFFICIENCY

Energy per request

Energy needed to process an inference request

Energy / token · tokens / joule

Also measured as tokens/s/kW

Single user latency ↔ energy per token

Lower latency usually costs more energy per token.
Energy is a proxy for cost

Compare systems across the full Pareto frontier.



NON GOALS ~~# chips used~~ ~~throughput per chip~~ ~~TTFT~~

Compare efficiency @ matched user experience

A Request Spans Three Distinct Hardware Regimes

OpenAI

01 Prefill
Encode context

FLOPs + attention

Attention-heavy and primarily compute-bound. Low memory-BW demand; communication is easier to schedule smoothly.

Compute	**HIGH**
Memory BW	**LOW**
Comms	**SMOOTH**

02 Draft model
Speculate

Network latency

Small models at ultra-low batch for latency serving—and low batch elsewhere. Low network bandwidth, but extreme latency sensitivity.

Model	**SMALL**
Batch	**ULTRA-LOW**
Comms	**LATENCY BOUND**

03 Spec-verify
Decode

Attention + HBM BW

Attention less memory bound with GQA and spec tokens. At low latency, MoE is memory-bandwidth hungry and communication arrives in bursts.

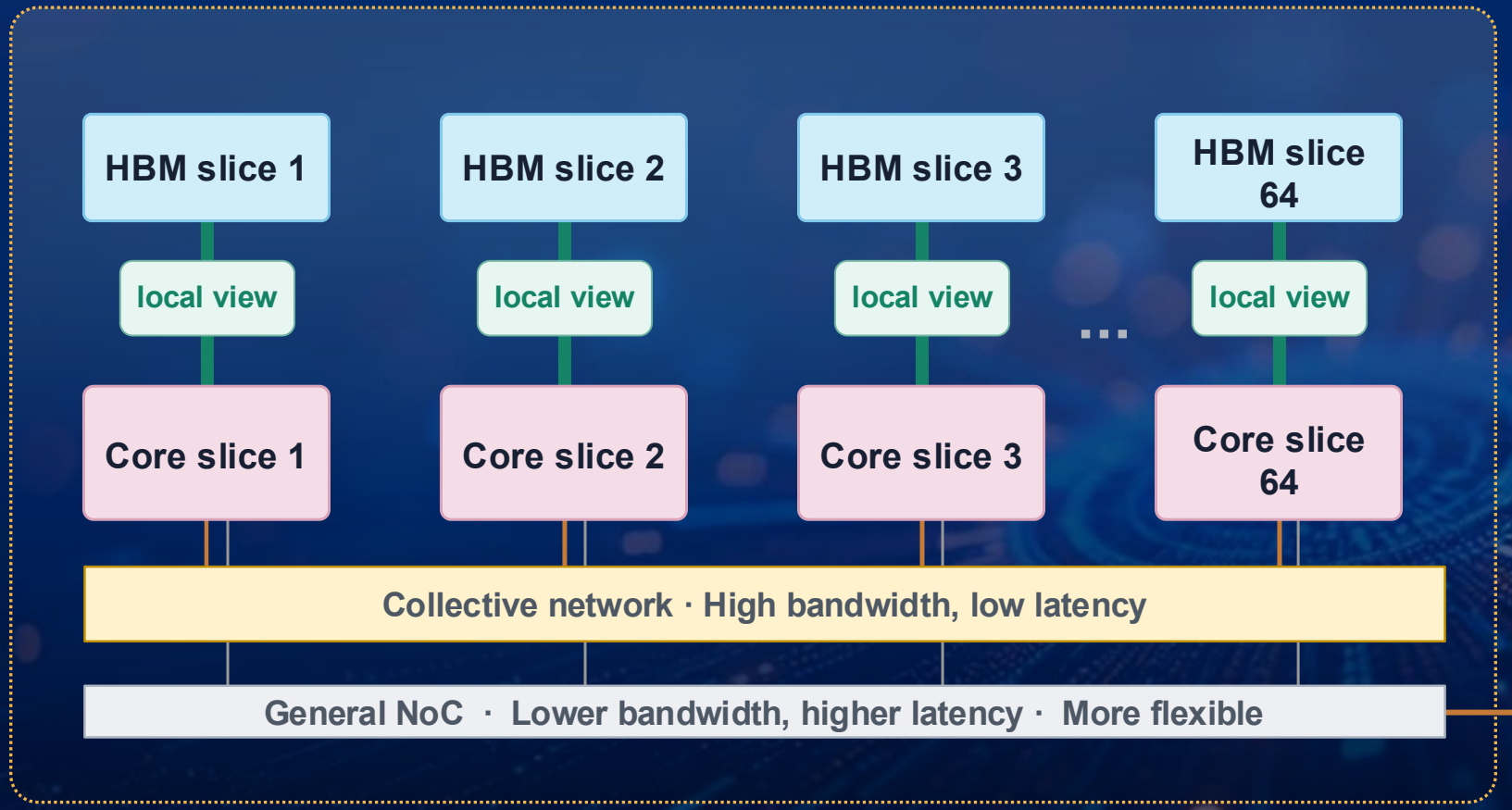
Attention	**COMPUTE**
MoE	**HBM BW**
Comms	**BURSTY**

What matters is requests/second/watt at the required SLA latency

Each phase hits a different bottleneck; efficiency only counts if the complete request remains within its end-to-end latency target.

Co-design the Architecture to Address the Bottleneck

OpenAI



FASTEST LOCAL PATH

Pair each core slice with an HBM slice to expose a low-latency, high-bandwidth local view.

SPECIALIZED CROSS-CORE

Collective network optimizes bandwidth and latency of common comms patterns.

PRESERVE FLEXIBILITY

Shared NoC for general comms and access to scale-up network.

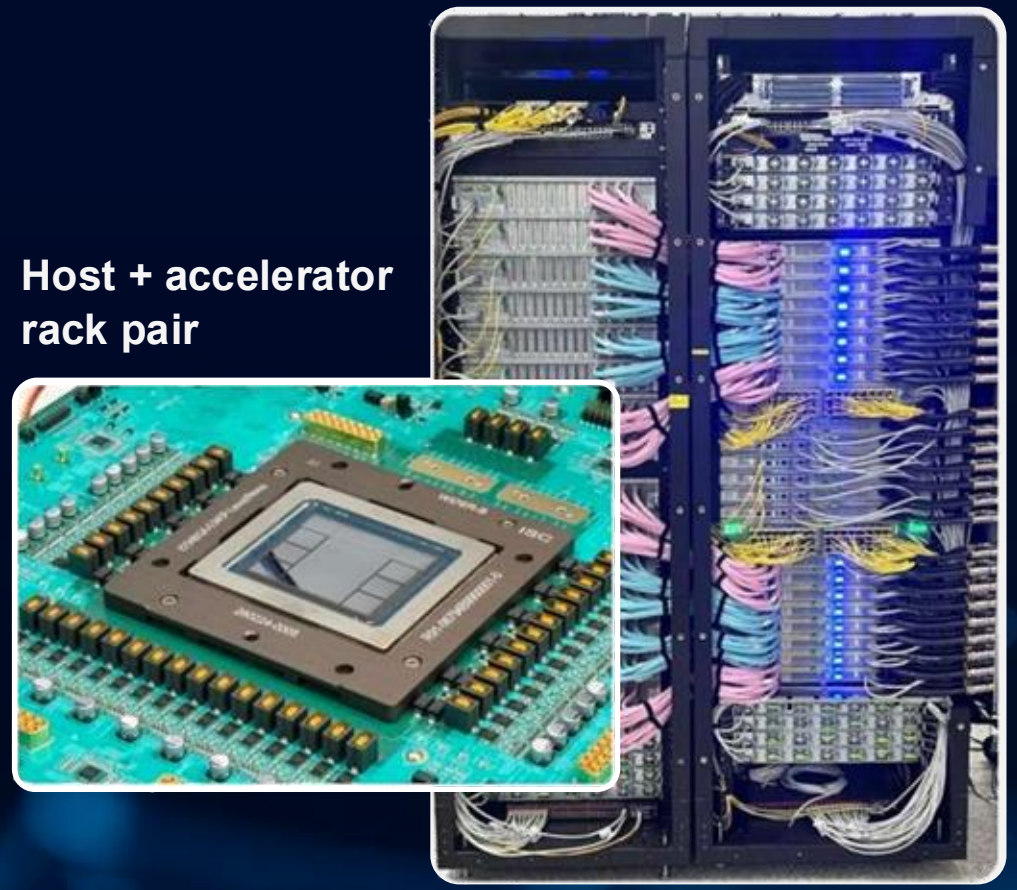
Keep common operands local; use the shared path only when necessary

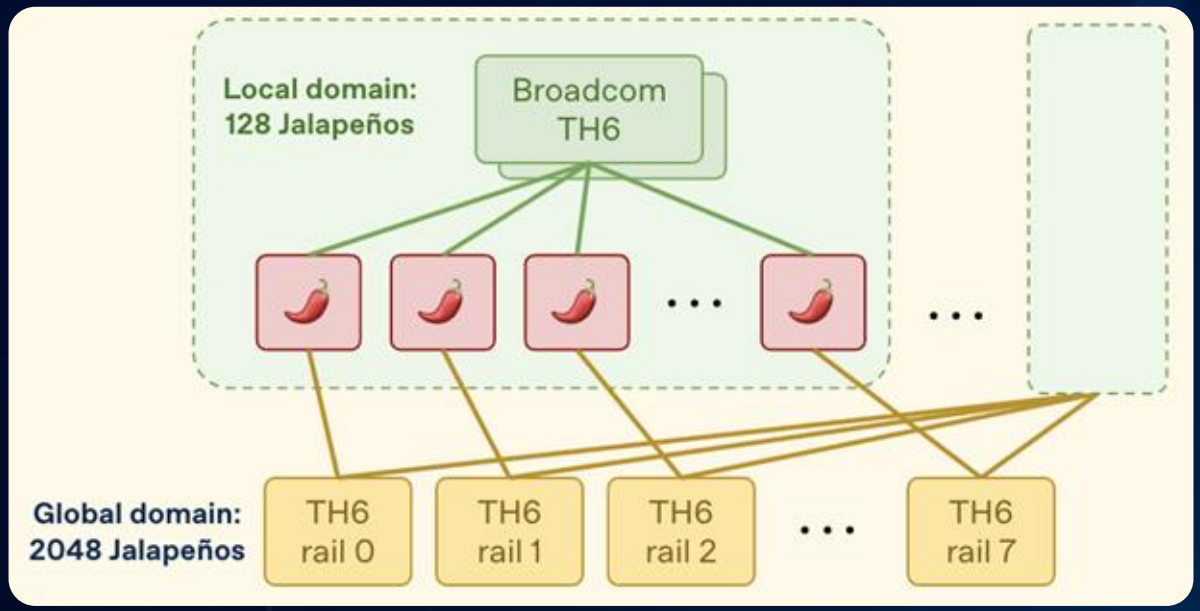
Explicit placement and distributed control prevent data movement/synchronization from dominating execution

OAI Jalapeño: from Chip to Scale-up System

OpenAI

Host + accelerator rack pair



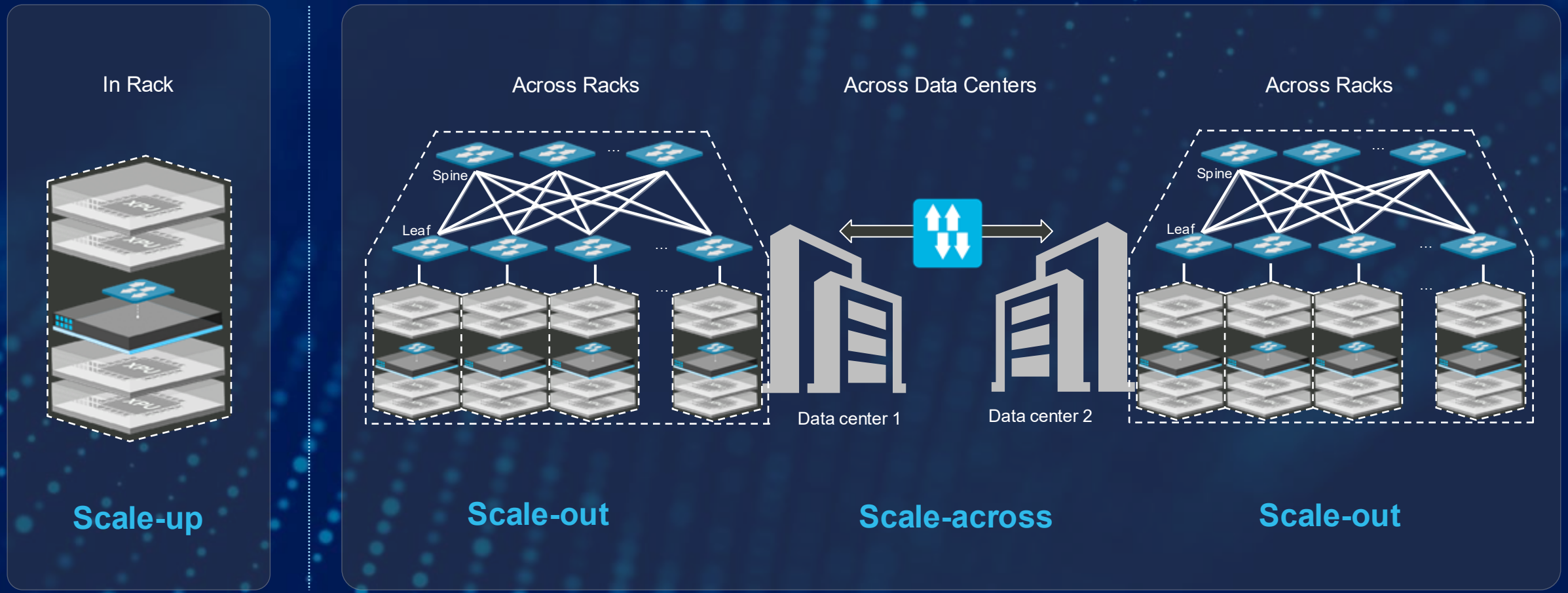


“Half flattened” two-level Clos topology

- Higher bandwidth for tensor parallelism
- Lower bandwidth for expert parallelism
- Low latency for both

AI Scale-up, Scale-out and Scale-across Networking

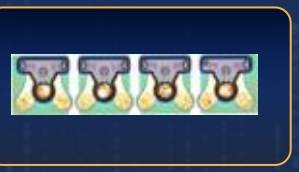
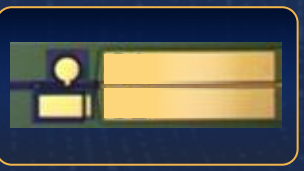
BROADCOM



Critical Optical Interconnect Building Blocks

BROADCOM



	**VCSEL**	Multi-mode, short reach
	**EML**	Single-mode, long reach
	**CWL**	Single-mode for Silicon Photonics and CPO
	**CPO**	Architecture, Reliability, and Power leadership, AI Scale-out
	**OCI**	Architecture, Reliability, Latency and Power leadership, AI Scale-up

Progression of Highly Integrated Photonics Technologies

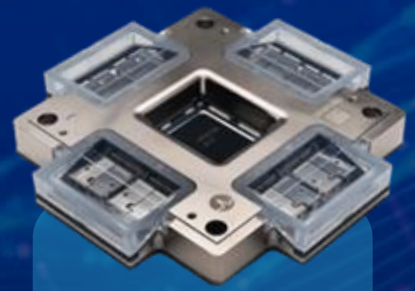
BROADCOM



TH4 Humboldt
100G/lane

2022

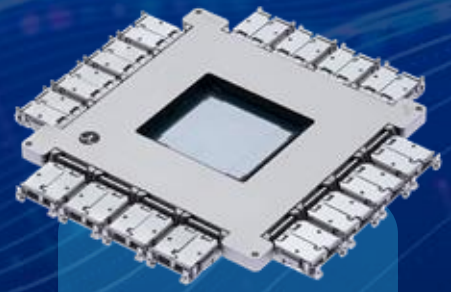
25.6T



TH5 Bailly
100G/lane

2024

51.2T



TH6 Davisson
200G/lane

2026

102.4T



OCI

2028

Next Gen

Scaling Highly Integrated Photonics: Challenges to Overcome

BROADCOM

Advanced Packaging

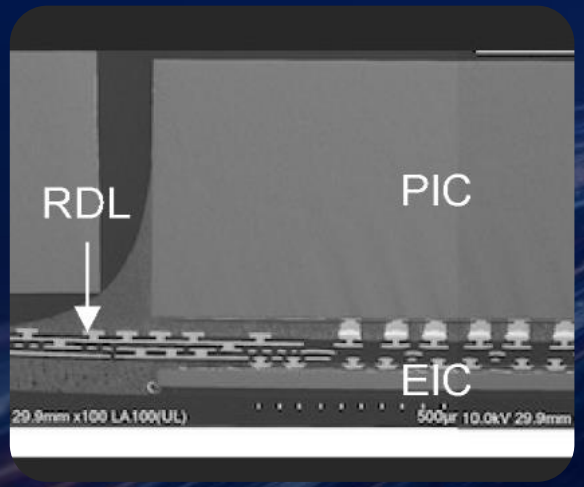
- High-density and high-quality PIC to EIC attach required
- Must optimize signal integrity
- Enablement for 400G direct drive

Optical Connectivity

- More than 1k fibers connecting to each chip package
- Detachable fiber connectivity to enable package reflow

Volume Manufacturing

- Must deliver industry leading yield and reliability through automation
- Enable system partners with automated assembly and test



TH5-Bailly FOWLP structure



Detachable fiber connector



Auto dispenser with loader and unloader

Progress Towards Scale and Automation

Demonstration of purpose-built equipment

Support Frame



In-line OGP



Delatch Z-Springs



Bucket



Introducing Optical Compute Interconnect (OCI)

BROADCOM

Founding Members

Industry alignment around an optical scale-up.

OCI-MSA

Est. March 2026

Future AI infrastructure built on open standards.



Optical Scale-Up

Architecture that delivers the efficiency required to scale AI infrastructure.

Move Scale-up from In-Rack to Multi-Rack & Multi-Row

Does the OCI-MSA Address Scaling?

BROADCOM

Bandwidth	Spec Timeline	Electrical	Optical	Linear	Comment
200Gbps	Today	100Gbaud PAM4	100Gbaud PAM4	YES	
400Gbps	2027?	PAM4/6/8, SBD?	200Gbaud PAM4	Maybe?	IC area & pJ/b diminishing returns Maybe cabled backplane support
800Gbps	>2030?	?	?	Unlikely	
OCI GEN1	Today	NA, wide parallel	4λ x 50Gbps NRZ, DWDM, bidi	NA	GEN2 spec @ 400Gbps = 2027 GEN3 spec @ 800Gbps = 202x

- AI pace of innovation likely faster than “standard” Ethernet AUI interfaces
- Future larger AI cluster scale-up domains will require highest radix switches

The Complete AI Rack System

Four disciplines, engineered in unison

NETWORKING

800G and 1.6T AI fabrics, 112G/224G signaling, qualified optics and connectivity

AI COMPUTE

Advanced liquid cooling, cable backplanes, and fully integrated modular AI racks, including OCP ORv3

STORAGE

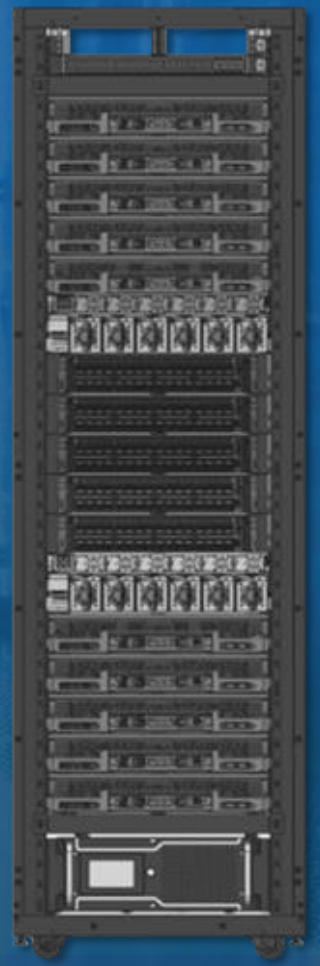
Dense enterprise-class expansion with high-availability PCIe Gen 5 controllers

SOFTWARE & RACK INTEGRATION

SONiC-based software, firmware, and lifecycle services

Celestica™

Designed, integrated and tested as one system



- **Top-of-rack switch**
- **AI fabric switch blades**
- **AI compute servers**
- **NVMe storage**
- **Cooling distribution unit**
- **Power and battery-backup shelves**

The Squeeze: Tech Acceleration Coupled with Faster Qualification

Solution qualification is the new critical path

Qualification windows are shrinking
— no room for iteration loops



- **36 months** · 100G / 400G
- **24 months** · 800G
- **18 months** · 1.6T

Annual

New AI accelerators from major silicon vendors

COMPOUNDING COMPLEXITY

~5,000
Passive copper cables in one scale-up rack

~1 MW
Liquid-cooled racks reshape thermal, mechanical and cable design

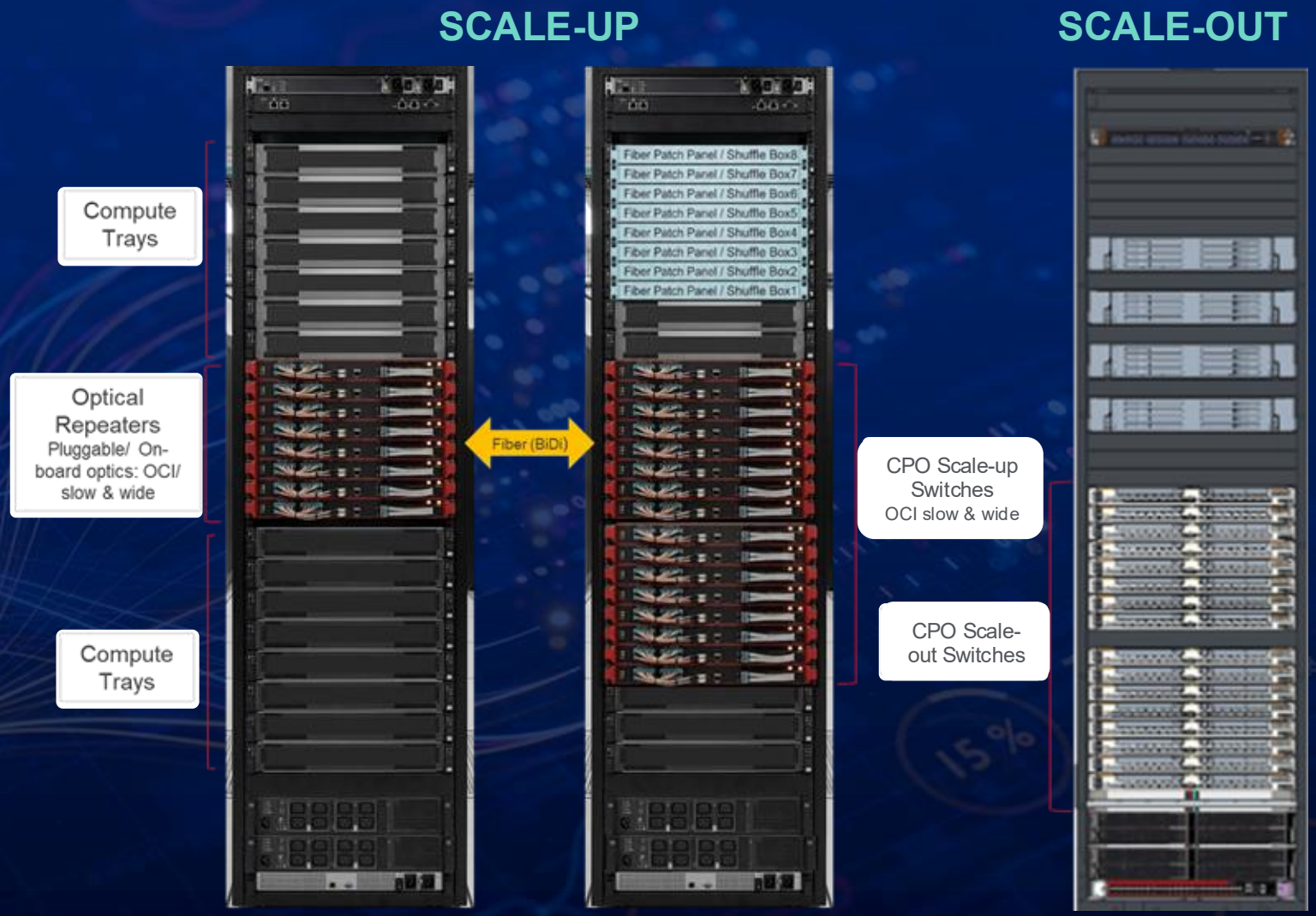
224G
SI challenges at high data rates

1.6T
Copper reach shrinks; transition to optical

~4,000 lb
Rack weight today; 8,000 lb design target ahead

2027 – 2030 Migration

- Transition to all optical
- CPO-based network switches & XPU
- OCI slow & wide optical interfaces in scale-up
- BiDi multi-wavelength channels over fiber
- Compute rack - optics on a separate optical repeater tray or directly on compute trays



Transition from Copper to Optical Interconnects Enables Large Scale-up Domains

At Scale Manufacturing: CPO Development

Category	Key Activity
Design Schedule	Project Kickoff
	Mockup Build
	EVT Prototype Build
Manufacturing Development	FAU / CPO Dev't Samples
	SMT Preparation
	SMT Optimization DOEs
	Standard Connector Inspection & Cleaning
	FAU / CPO Inspection and Cleaning
	FAU Assembly / Loopback Testing
	SMT Optical Validation
	Line Layout
	Test Development
	Hot Gas Rework

EVT Build

Manufacturing Process Readiness

- Front End
- Assembly
- Test









Introducing Celestica's DS6005 Open CPO Platform Solution

AI Fabric solution based on Broadcom's Tomahawk 6 – Davisson CPO chipset



Celestica™

BROADCOM®

SONiC

OPEN Compute Project
SOLUTION PROVIDER®

- Celestica plans to contribute the design to the OCP Community
- For more information: www.celestica.com/DS6005
- Visit Celestica booth C2 to see the DS6005

OCP Community: Let’s partner to design the next AI Fabric