Emerging Technologies and Future Architecture Improvement Potential in HPC/AI Interconnects

January 27, 2026

Yuichiro Ajima

Global Fujitsu Distinguished Engineer

Principal Architect

Fujitsu Limited

About Me

- Current Role
  - Principal Architect for system and interconnect architecture of supercomputer
- Past Projects and Achievements
  - Development of the K computer system
  - Development of the supercomputer Fugaku
  - Architecture lead of Tofu interconnect series
- Current Projects
  - FugakuNEXT Basic Design
  - Feasibility Study 3.0 for Future HPCI

Agenda

- Today's HPC/AI Interconnects
  - Scale-up network and scale-out network
  - Topology and signal density
- Emerging Technologies
  - Co-packaged optics (CPO) enhances signal density
  - Optical circuit switch (OCS) adds flexibility to topology
- Architectural Opportunities
  - Collective communication-aware network topologies
- Summary

Today's HPC/AI Interconnects

HPC and AI Interconnects: Similarities

- Address-based communication protocols
  - Access user space data directly
    - RDMA Get/Put protocols
    - Read/Write primitives
  - Hardware-based kernel-bypass technologies
    - Contrast with TCP/UDP's software-based processing

- Collective communication is critical in parallel computing
  - Sets of nodes start exchanging large data simultaneously
  - Contrasts with random traffic patterns of traditional data centers

HPC and AI Interconnects: Differences

- HPC interconnects are designed as a single, large-scale fabric to achieve high scalability
  - AI interconnects also incorporate a system-wide network based on proven HPC technology

- Modern AI interconnects add a second, dedicated network known as the scale-up network
  - This is a separate, high-bandwidth network for communication within small node groups
  - The system-wide network is referred to as the scale-out network

Scale-Out and Scale-Up Networks

- Transmission technologies are common
  - Both use PAM4, transferring 2 bits at rates from 50 to 100 Gbaud

- Differences in scalability and signal density
  - Scale-out networks have 2 to 3 orders of magnitude higher scalability than scale-up networks
  - Scale-up networks have an order of magnitude higher signal density than scale-out networks

- Following pages discuss the differences

Difference in Scalability

- The scalability of a multi-tier switch network increases exponentially as the number of switch tiers

- Scale-out network: 10,000 to 100,000 of NICs
  - Typically consists of 2 or 3 tiers of switches

- Scale-up network: 10s to 100s of xPUs
  - Typically connected by a single-tier of switches
  - The scalability is limited to the switch radix or the number of signal lanes per switch

Difference in Signal Density

- The differences in transmission media and implementation lead different bottlenecks for signal density

- Scale-out network: 4 to 8 signal lanes per connector
  - The bottleneck in optical connections between racks is the front panel connectors that support pluggable optical transceivers

- Scale-up network: 24 to 32 signal lanes per connector
  - The bottleneck in electrical connections within a rack is the high-density connectors on backplanes

Topology vs. Signal Density Trade-Off

- Scale-out networks have multiple connection tiers
  - xPU-to-NIC, NIC-to-Switch, and Switch-to-Switch
- Scale-out networks incur additional cost for optical link
  - Inter-rack connections require optical links, whose transceivers are several times more expensive than electrical cables

- Scale-up networks use a simplified topology
  - Only a xPU-to-Switch tier, with no inter-rack connections
- This trade-off enables the 3-8 times higher signal density than scale-out networks

Emerging Technology:
Co-Packaged Optics (CPO)

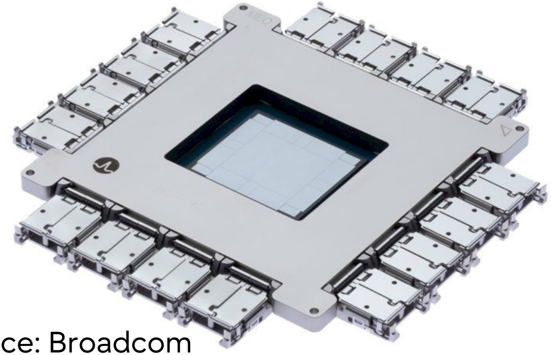
Benefits of Co-Packaged Optics (CPO)

- Conventional pluggable optics
  - Electrical signals driven across the PCB consume high power
  - Large front-panel space occupied by transceiver modules

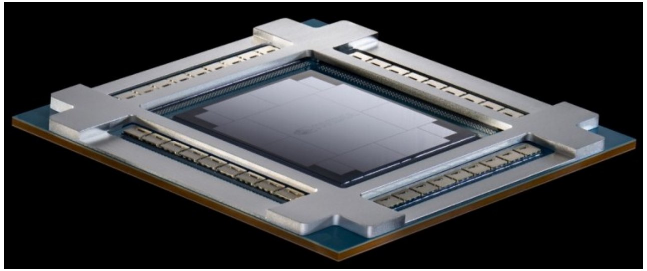
- Co-packaged optics (CPO)
  - Integrated into the same package as the logic chip
  - Low power: Eliminates long, lossy PCB traces
  - Higher density: Enables small, dense fiber connectors (e.g., Multi-fiber Push-On, MPO) on the front panel

Planned Products and Prototypes

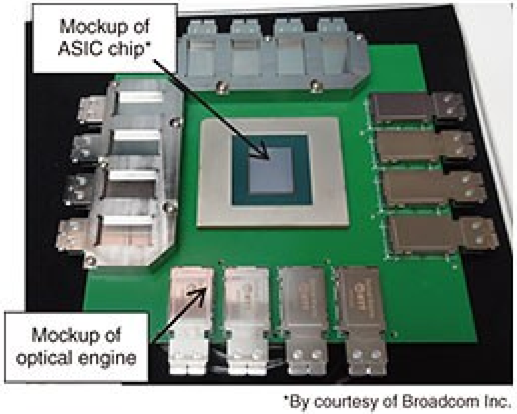
- CPO deployment has started with scale-out network switches... and is extending to xPUs



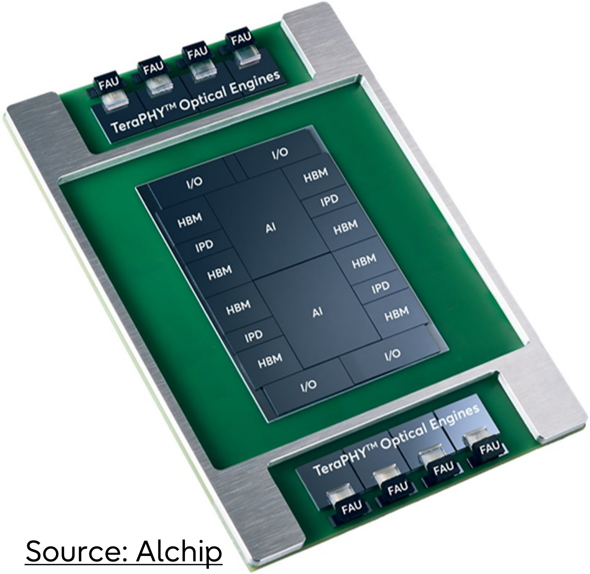
Source: Broadcom



Source: NVIDIA



Source: NTT



Source: Alchip

Future Outlook for CPO Technology

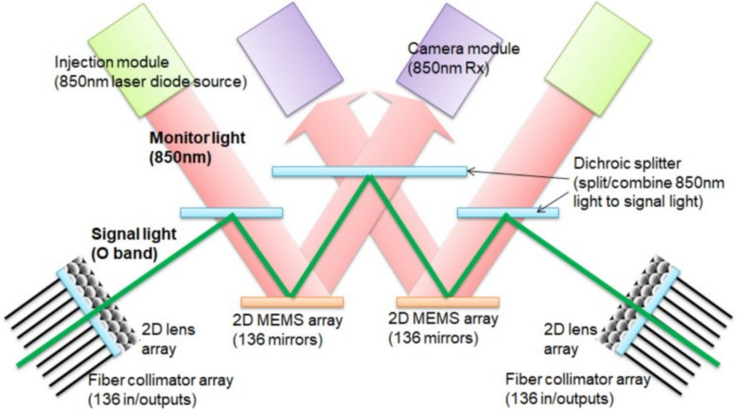
- Widespread adoption for xPUs is anticipated
  - Enables more scalable scale-up networks via optical links
  - Limitation: Fiber count per xPU will still be lower than current electrical I/O count

- Wavelength Division Multiplexing (WDM) will boost density
  - Enabled by silicon photonics for compact transceivers
  - Multiplexing 8-16 wavelengths per fiber
    - e.g., CW-WDM MSA standards

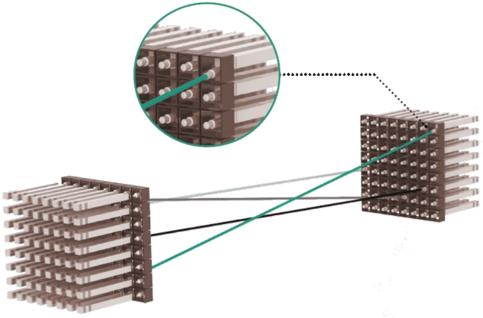
Emerging Technology:
Optical Circuit Switch (OCS)

Emerging Optical Circuit Switches (OCSs)

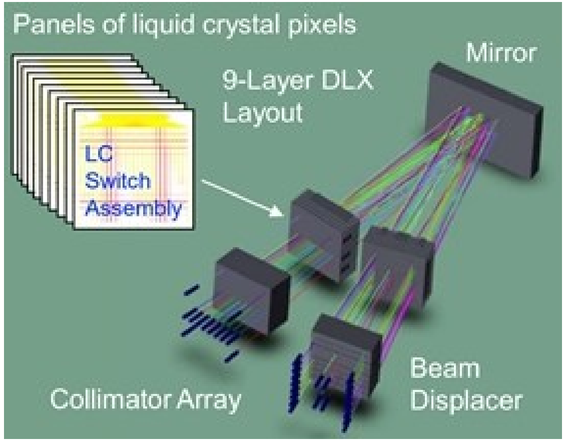
- The modern evolution of the automated patch panel
- New types of OCS have recently emerged
  - MEMS mirrors, piezoelectric beam steering, and liquid-crystal switching



Source: Google



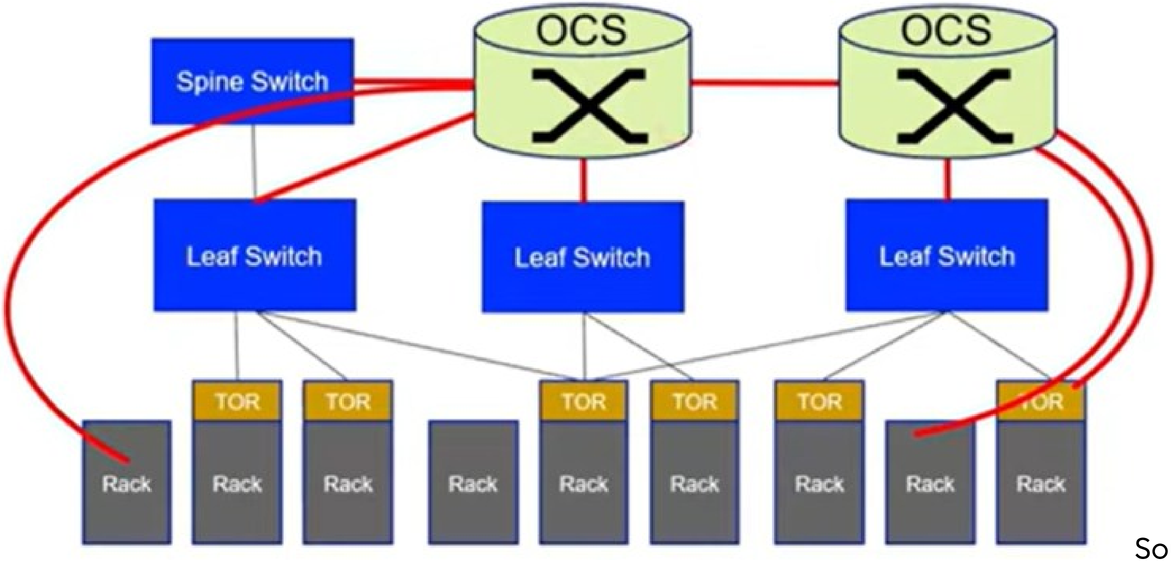
Source: OptiNet China 2024



Source: Coherent

Early Adoption in Data Centers

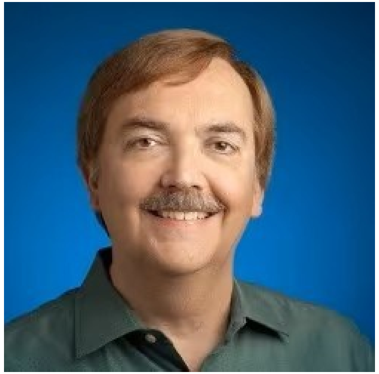
- OCS can replace many, but not all, top-tier switches
  - Eliminating the O-E-O conversion and packet switching



Source: Coherent

OCS Pioneers in HPC and AI Systems

- A visionary proposal for HPC (2005)
  - SC'05 Paper: *"Analyzing Ultra-Scale Application Communication Requirements for a Reconfigurable Hybrid Interconnect"*
  - Led by John Shalf, recipient of the 2025 Seymour Cray Award
- Large-scale deployment for AI (2023)
  - Google TPUv4 System (ISCA'23 Paper): *"TPUv4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings"*
  - Led by Norm Jouppi, recipient of the 2024 Seymour Cray Award



Architectural opportunities

Network Topology Considerations

- With CPO integrated into xPUs:
  - Pro: Enables efficient, high-density connectivity beyond the rack
  - Con: Reduces switch radix, limiting scalability of scale-up network
    - Fiber pairs per switch < Electrical lanes per switch
    - WDM does not compensate for the decrease in switch radix

- New topology designs are required to ensure scalability under lower radix constraints

Strategies for Improving Scalability

- Simple approach: Add switch tiers
  - Effectively creates a second scale-out network
  - Is building two parallel networks justified?

- Cost-effective approach: Optimize topology for traffic
  - Matching the topology to the algorithms is straightforward
    - Example: Ring algorithms naturally suit ring topologies
    - The optimal algorithm for a typical (bidirectional) ring topology is the double-ring algorithm

Collective Communication Algorithms

- Multiple algorithms exist for each collective operation
- The optimal algorithm depends on runtime conditions
  - Factors: Number of xPUs, data size, bandwidth vs. latency
- The suitable topology also varies by algorithm

Collective Communication Algorithm	Suitable Topology
Ring	Ring
Multi-dimensional ring	Torus
Pair-wise exchange	Hypercube
Tree	Tree
Simple spread	Full-mesh

Topology Introduction Strategies

- Integrate as a hybrid hierarchical topology
  - A direct approach to compensate for reduced scalability
  - Example: Integrating a ring topology
    - Approach A: Connect multiple switches in a ring
    - Approach B: Connect xPUs in rings, then link the rings via switches

- Shift to high-scalability topologies (e.g., Torus).
  - The workload's collective communication conditions should be suitable for the topology for optimal performance
  - Otherwise, the algorithm/communication pattern of the workload should be adapted to avoid performance degradation

Reconfiguring Topology with OCS

- A topology optimized for one collective algorithm may underperform for another

- OCS enables flexible, dynamic topology reconfiguration
  - Connect each fiber to a separate OCS, not all to one, for the maximum scalability
  - Example: 48 fibers per xPU and 48 OCSs
    - This enables reconfiguration into any topology at full scale

The Partitioning Issue in Fixed Topologies

FUJITSU

- Partitioning a physical topology is non-trivial
  - Real workloads often deviate from initial assumptions
  - Example: A large physical ring must be partitioned to create smaller logical rings

- Partitioning techniques in production systems:
  - IBM Blue Gene: Link switching via electrical switches
  - Fujitsu K / Fugaku: Ring embedding into high-dimensional networks
  - Google TPU: Dynamic reconfiguration using OCS
- OCS can inherently resolve this partitioning issue

Summary

- Modern AI Interconnects employ dual networks
  - Scale-out network ensures system scalability
  - Scale-up network provides high bandwidth

- Two emerging optical technologies:
  - CPO delivers higher signal density at the cost of scale-up radix
  - OCS introduces topological flexibility

- Optimizing topology for collective communication
  - A cost-effective strategy to restore scalability
  - Enabled by the topological reconfiguration capability of OCS

Thank you