An AI System View on 448Gbps Signaling

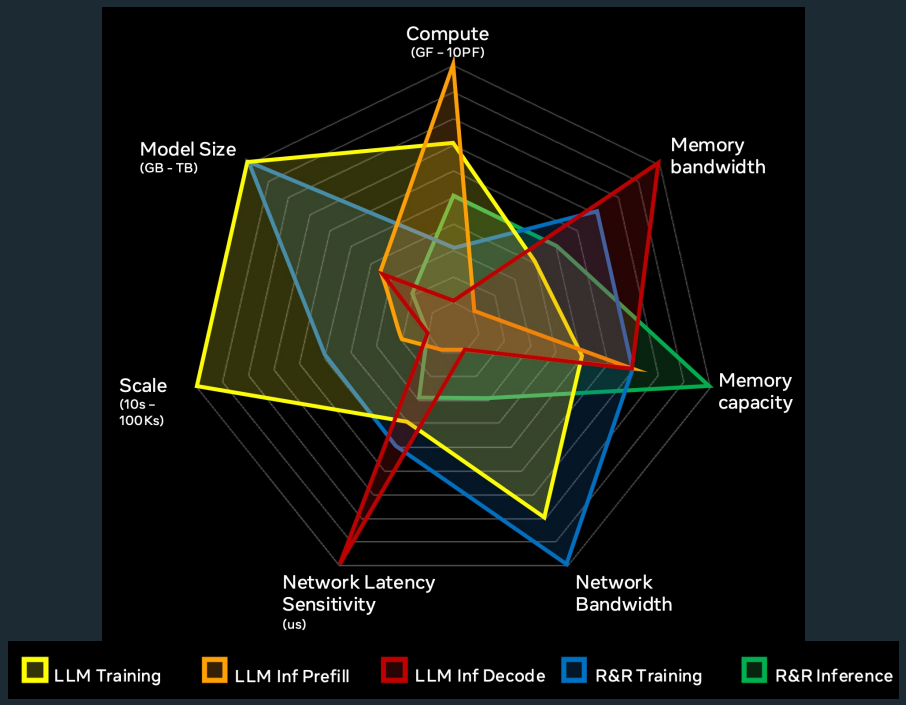
Xu Wang, Srinivas Venkataraman

Meta Platforms, Inc.

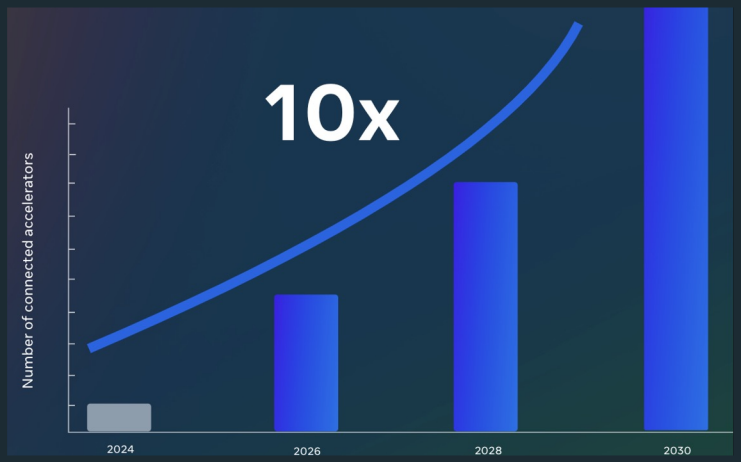
OIF 448Gbps Signaling for AI Workshop
April 15-16, 2025

Meta

AI Systems Workload Demand and Cluster Scaling



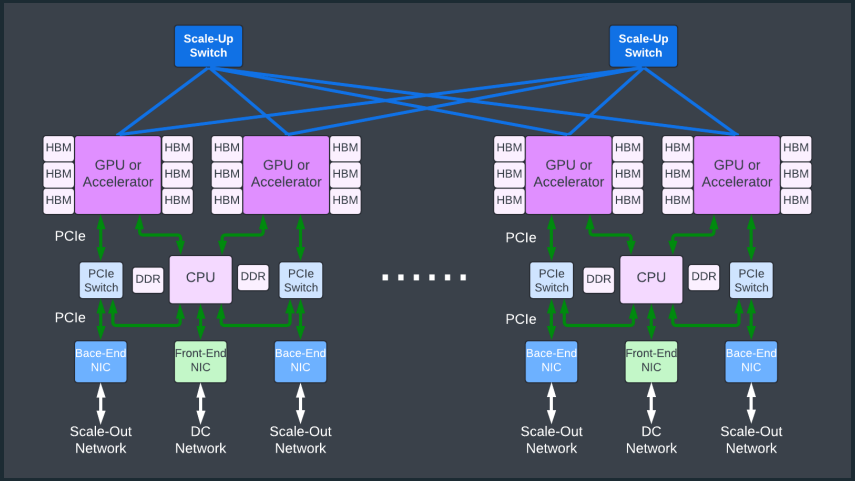
Workload Demand



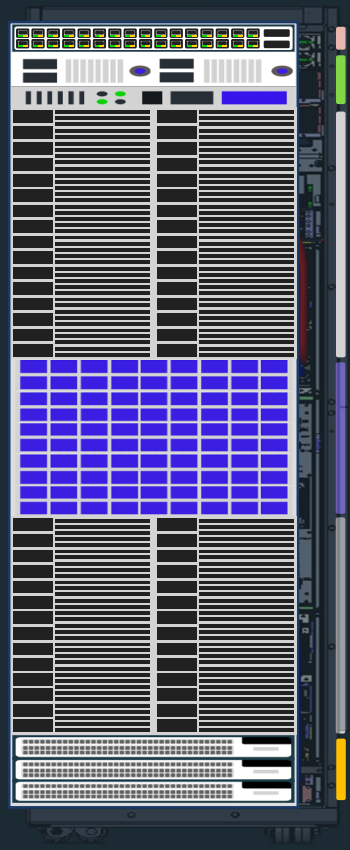
Cluster Size Expansion

The network bandwidth keeps going up along with compute / memory.

Current-Generation Rack-Scale AI Systems



A Generic Rack-Scale AI System
Functional Block Diagram



Top of Rack Switch

Rack Management

Accelerator Bank

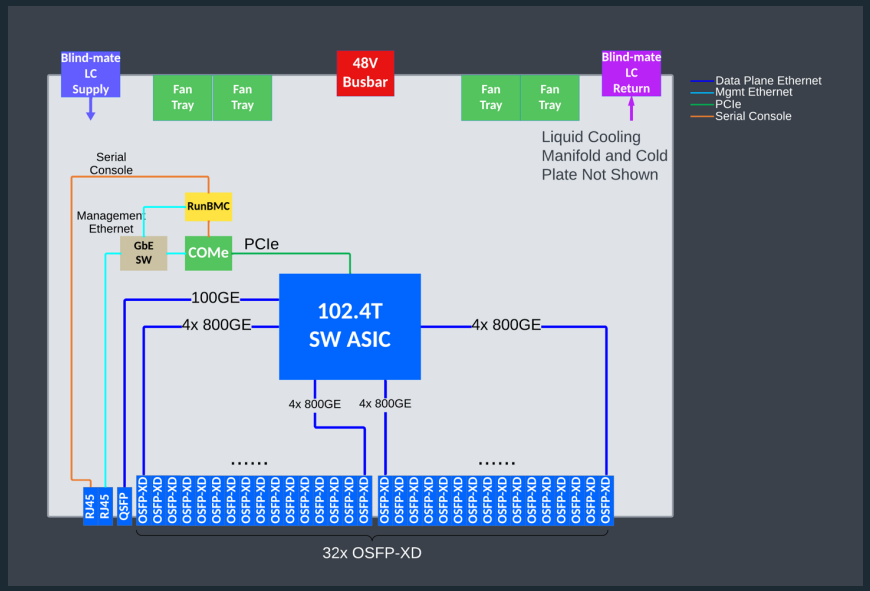
Scale-Up / Scale-Out Switch Bank

Accelerator Bank

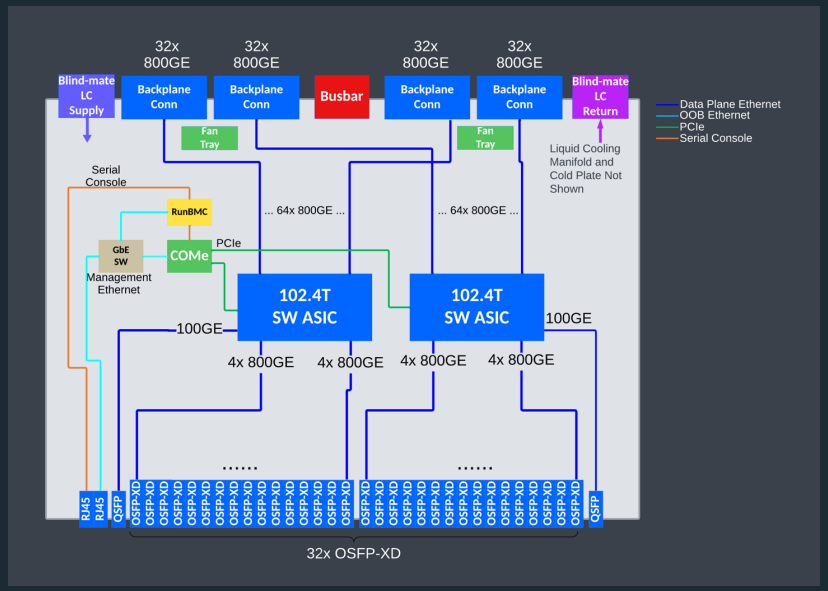
Power and Hold-Up Energy Shelves

A Generic Rack-Scale AI System
Rack Elevation

Scale-Out Network Switches



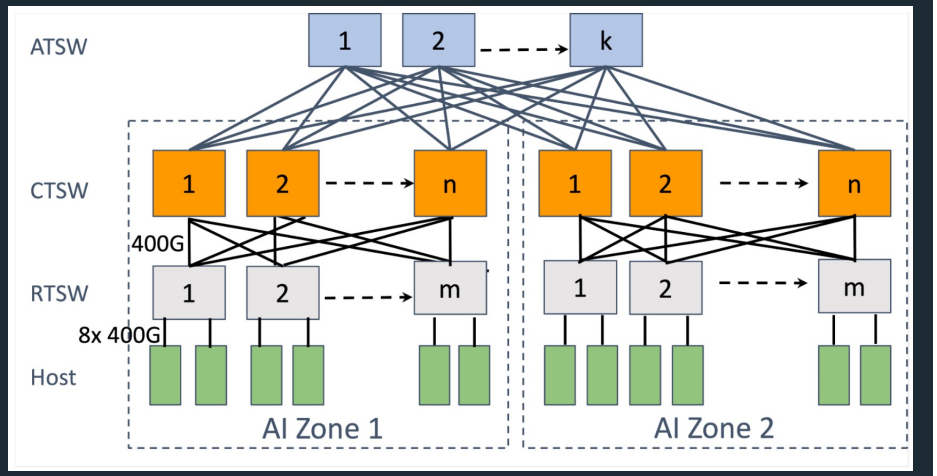
Pizza box-style single switch ASIC system



Chassis-based modular switch
Line cards / fabric cards

Scale-Out Network – Connecting Massive Clusters

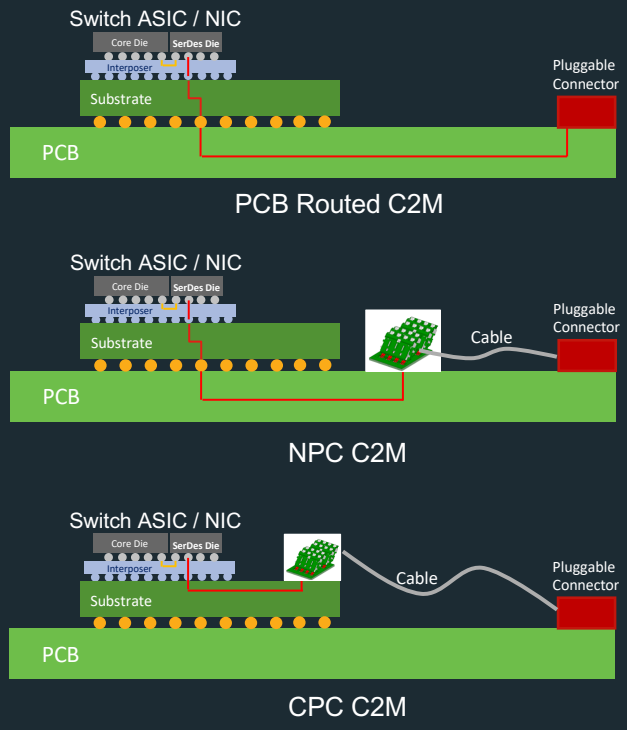
- Characteristics
  - O(K – 100K) accelerators
  - Bandwidth: ~100GBps per accelerator
  - Topology: Hierarchical, fat tree
  - Distance: Kilometers
  - Latency: us ~ 100s of us
  - Energy efficiency: optics technologies, switch form factors
- Benefits of doubling SerDes speed
  - Optics cost and power
  - Switch front-panel bandwidth density



Scale-Out Network Example

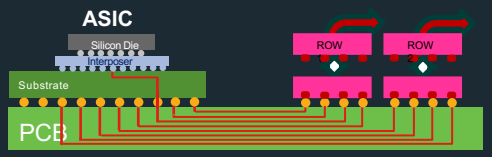
Scale-Out Network Speed Scaling

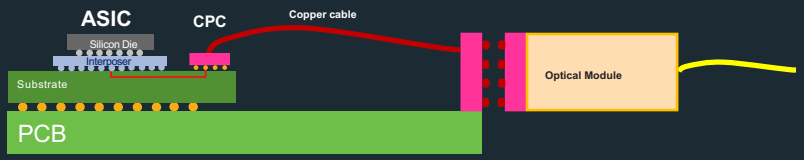
- Pluggable transceiver interface (C2M)
  - 112G PAM4: ~22dB bump to bump
    - PCB routing
    - Retimed optics, TRO, LPO
  - 224G PAM4: 28~32dB bump to bump
    - PCB routing, Near Package Copper (NPC)
    - Retimed optics, TRO
  - 448G: ~40dB bump to bump? (assuming we go with increased channel bandwidth)
    - PCB routing, NPC, Co-Packaged Copper (CPC)
    - Retimed optics, with probably different electrical modulation formats
- CPO
  - Significant power savings over retimed pluggable optics
  - Cost savings
  - Reliability / Availability remains a concern.
  - Vertically integrated ecosystem currently



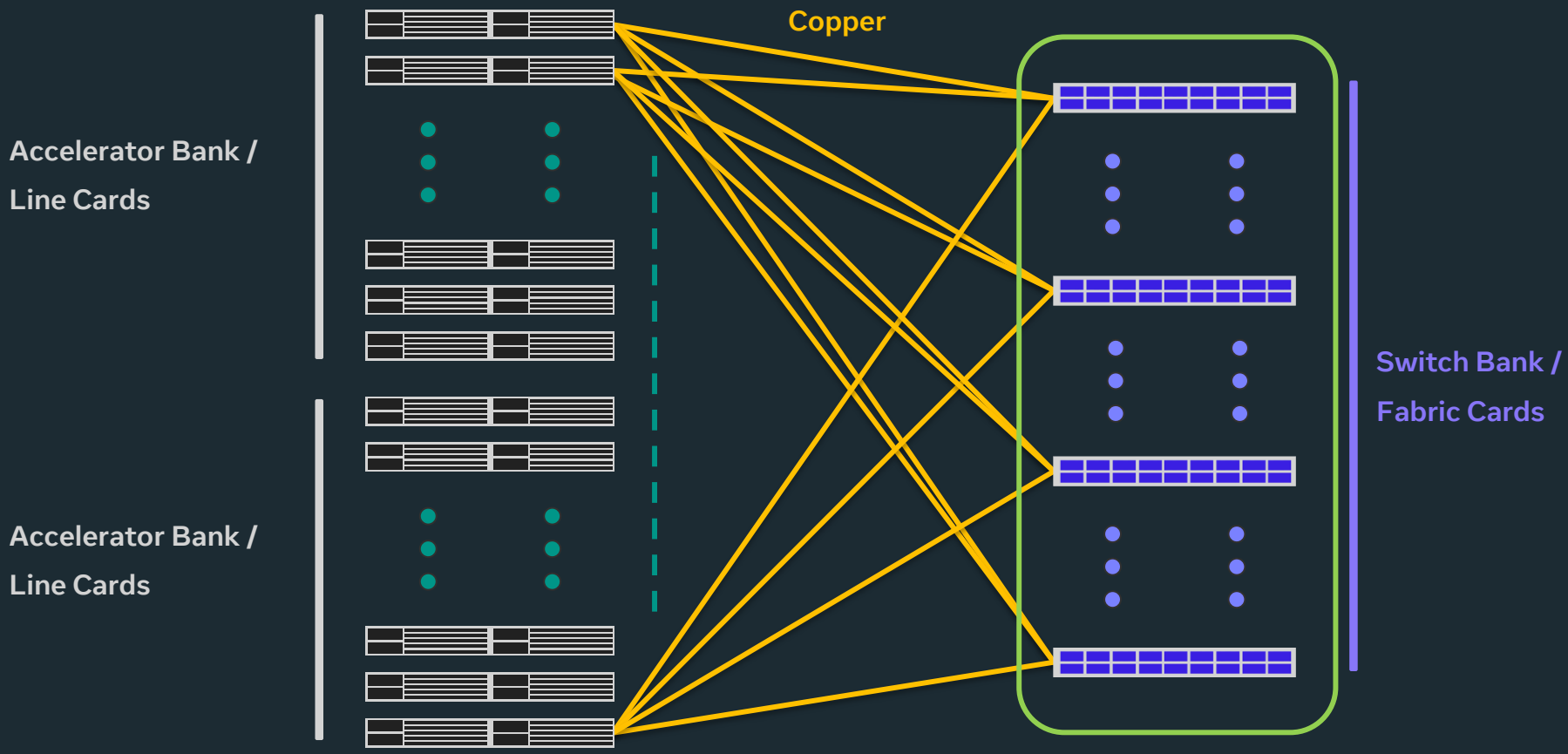
Scale-Out Network Speed Scaling Challenges

- Pluggable connector
- Switch ASIC PCB footprint and break-out via pattern
- Switch ASIC package loss

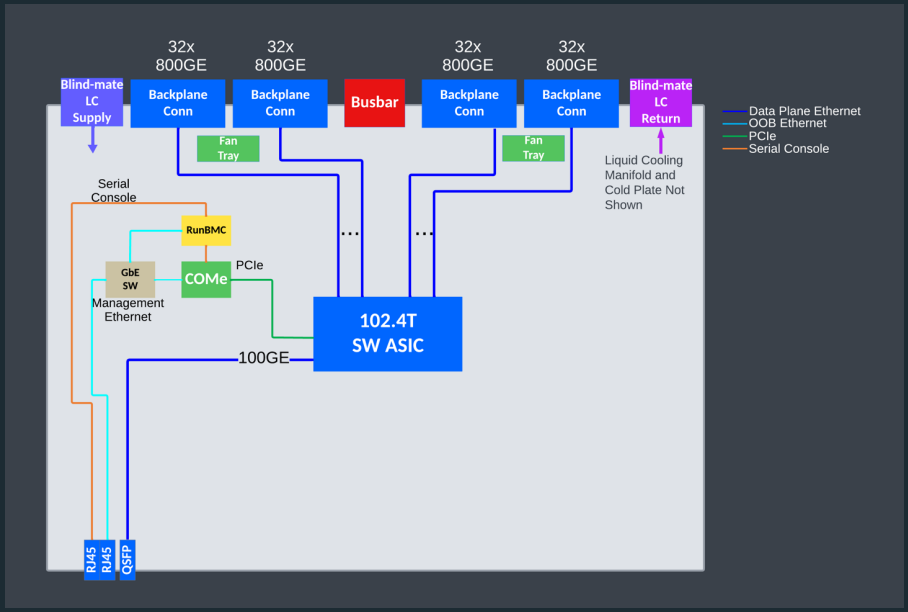




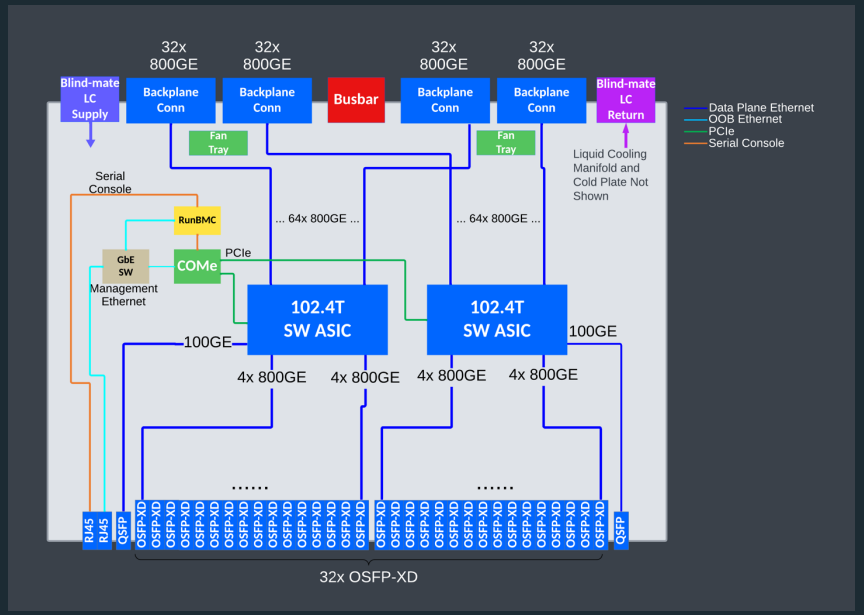
Intra-Rack Connectivity for Scale-Up and Scale-Out Network



Scale-Up Network Switches



Scale-Up Switch without Uplinks



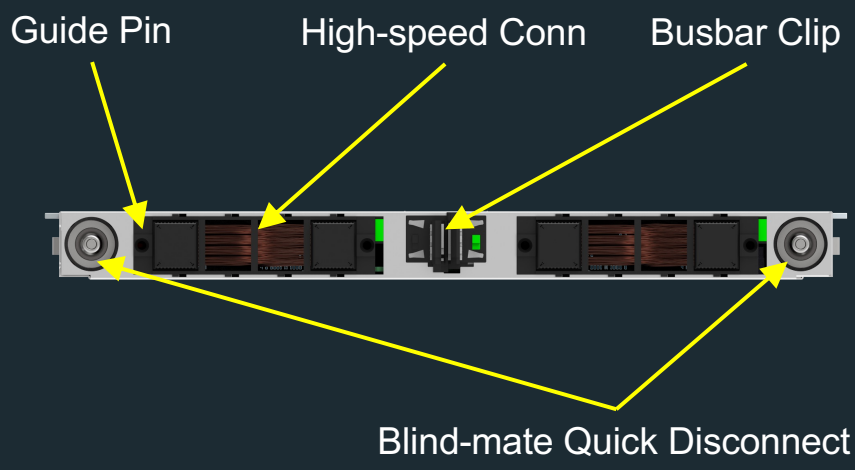
Scale-Up Switch with Uplinks

Scale-Up Network – Delivering Massive Bandwidth

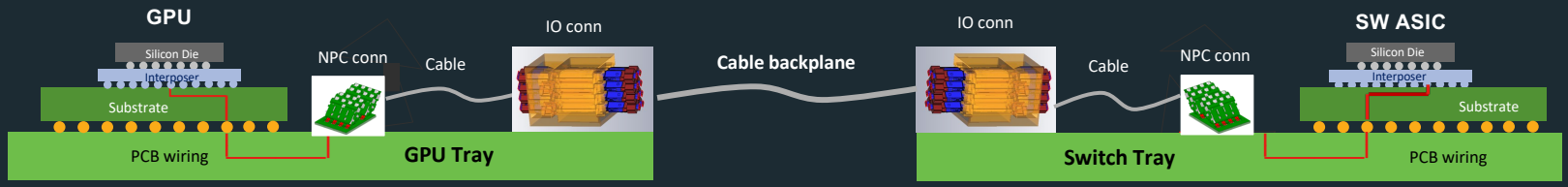
- Characteristics
  - 10s ~ 100s of accelerators
  - Bandwidth: ~1TBps per accelerator (an order of magnitude higher than Scale-Out)
  - Topology: Single-hop star, 2-hop hierarchical
  - Distance: 1 ~ 2 meters
  - Latency: 100s of ns ~ us
  - ***Dominated by copper connectivity: Availability, power, cost advantage over optics***
  - Energy efficiency: system form factors
- Benefit of doubling SerDes speed
  - Silicon beachfront, package size, system boards cross-sectional area
  - Total achievable bandwidth within a scale-up domain by copper connectivity

System Boards Beachfront Bandwidth Density

- Backplane beachfront space
  - High-speed connectors
  - Power delivery – busbar clip
  - Liquid cooling quick disconnect
  - Mechanical guiding pins
  - Air flow opening
- ~1024 differential pairs in 1OU cross-sectional area
- Liquid cooling enabled packing so many chips in a rack, and system boards beachfront became a bandwidth bottleneck.



System Form Factor and Physical Channel Reach Scaling



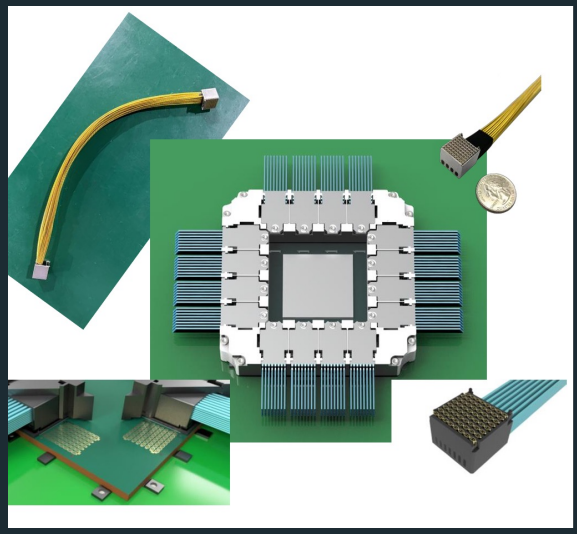
Substrate Wiring	PCB Traces	Flyover Cables	Backplane Cables	Flyover Cales	PCB traces	Substrate wiring
15mm	2.5”	250mm	1,350mm	600mm	5”	50mm

Can we build denser systems with reduced physical channel reach?

- ASIC package sizes increase.
- System boards form factors increase.
- Physical reach and subsequent copper insertion loss increase.
- Some optimization opportunities in system design but may not be enough.

Promise (and Perils) of CPC

- Eliminating the PCB footprint escape and the traces in high-speed SerDes channels
  - Also eliminating the core vias through the thick cores on substrate
- Significant PCB cost reduction
- CPC connector technology readiness
  - Reliability and robustness
  - Mechanical mounting complexity
- Chip substrate size increase
- Relatively high insertion loss with small gauge twinax wires to CPC
  - Solution for gauge transition required
- Ecosystem and supply diversity
  - Where are the possible demarcation points for multi-sources

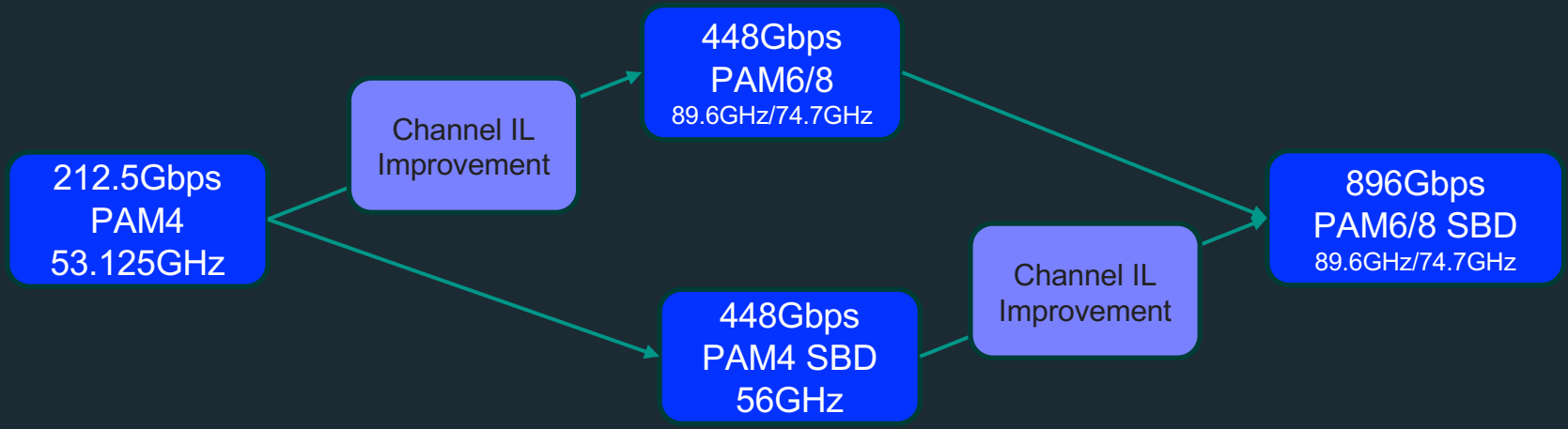


Courtesy of LUXSHARE-ICT, Inc.

Scale-Up Network Power Efficiency – Retimer-less

- Maximize the channel reach even at the expense of pJ/bit
  - Programmable reach vs. power consumption will be preferred, if the savings are significant.
- Preference is retimer-less channels.
  - Double the SerDes power of a given scale-up link
  - Double the SerDes links to manage, tune and monitor
    - Higher risk of link flaps
  - Cost increase
- Where to place the retimers if we must
  - Switch tray where there is longer cable lengths
- Active cable backplane? Or AEC cables inside the chassis?
  - How to service an active cable backplane
  - Where to have a PCB to host the retimers in a cable assembly

SerDes Technology Choices and Roadmap to 896Gbps



- Given the state of interconnect (substrate, PCB, twinax cables), insertion loss will increase from 53.125GHz to 74.7GHz/89.6GHz for the same physical reach.
- If we target 2027/2028, which path is easier to get to 448Gbps with better power efficiency (enabling retimer-less channels)

Optical Innovations for Scale-Up Network

- Characteristics of competitive optical innovations
  - Power efficiency comparable to or better than electrical SerDes with copper
  - Higher beachfront bandwidth density
  - Longer reach at rack row level or even datacenter hall level
  - Comparable availability and failure rate
- What optical innovations may enable
  - Break Free from copper
  - System configuration flexibilities with disaggregated systems

Summary

- The network bandwidth keeps going up along with compute / memory in AI systems.
- For AI systems with copper-based in-rack scale-up network, doubling SerDes speed can enable the bandwidth growth given the system boards beachfront limits.
  - SerDes technology choice should consider the system design benefits.
- Doubling SerDes speed in the scale-out network can reduce the optics power and cost when the technology matures.

Call for Actions

- 448Gbps end-to-end solution is highly desirable for 2027 power on.
- SerDes technology choice for the tradeoff between pj/bit, area, and channel reach
- Interconnect material and technology development targeting higher Nyquist frequencies
  - Package substrate
  - PCB
  - Over the board twinax cables
  - Connectors
- Optical innovations

Thanks

Meta