AsteraLabs®

Scaling GPU Clusters & Low Latency Memory Fabrics With Active PCIe / CXL Cabling

Chris Blackburn

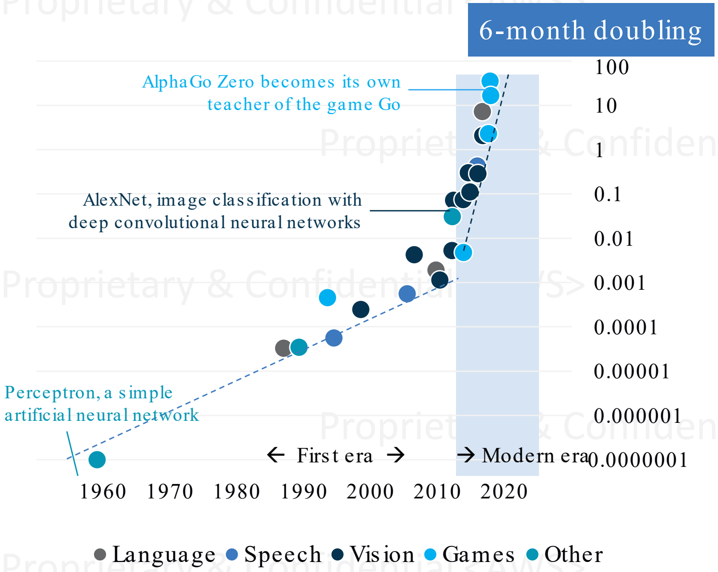
System Architect & Director of Field Applications Engineering

August 2024

FMS
the Future of Memory and Storage

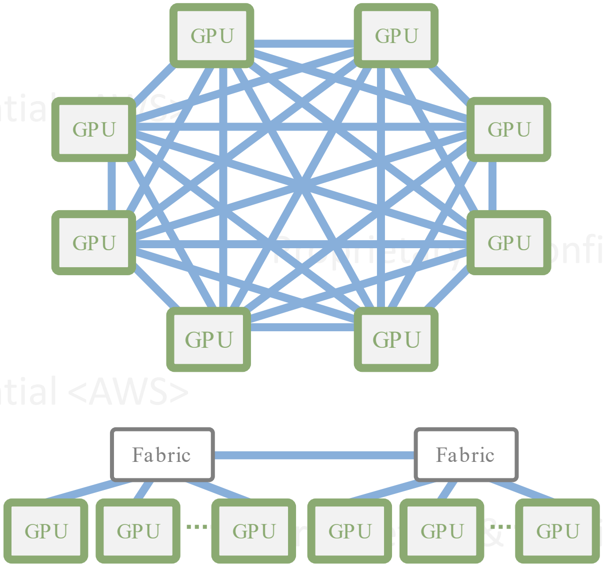
AI Infrastructure Scale Challenges

AI Models Continue to Expand



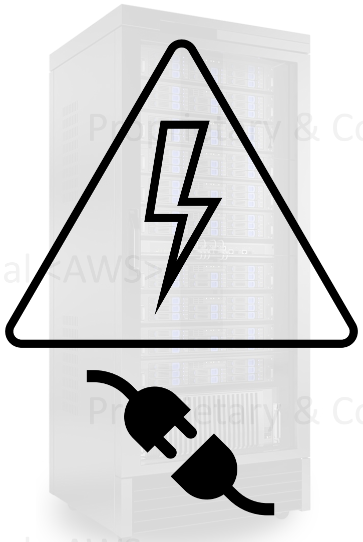
Model sizes have doubled after 6 months*

Larger GPU Clusters Required



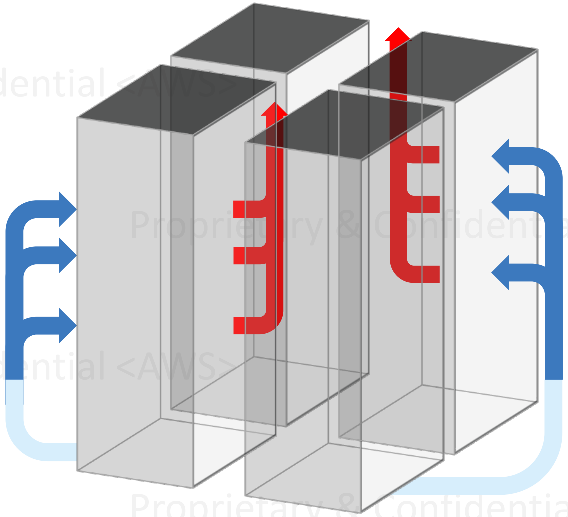
Scale up fabrics connect hundreds of GPUs

Limited Power per Rack



AI servers consume 8X more power than CPU servers**

Thermal Constraints



GPUs transitioning from air to liquid cooling***

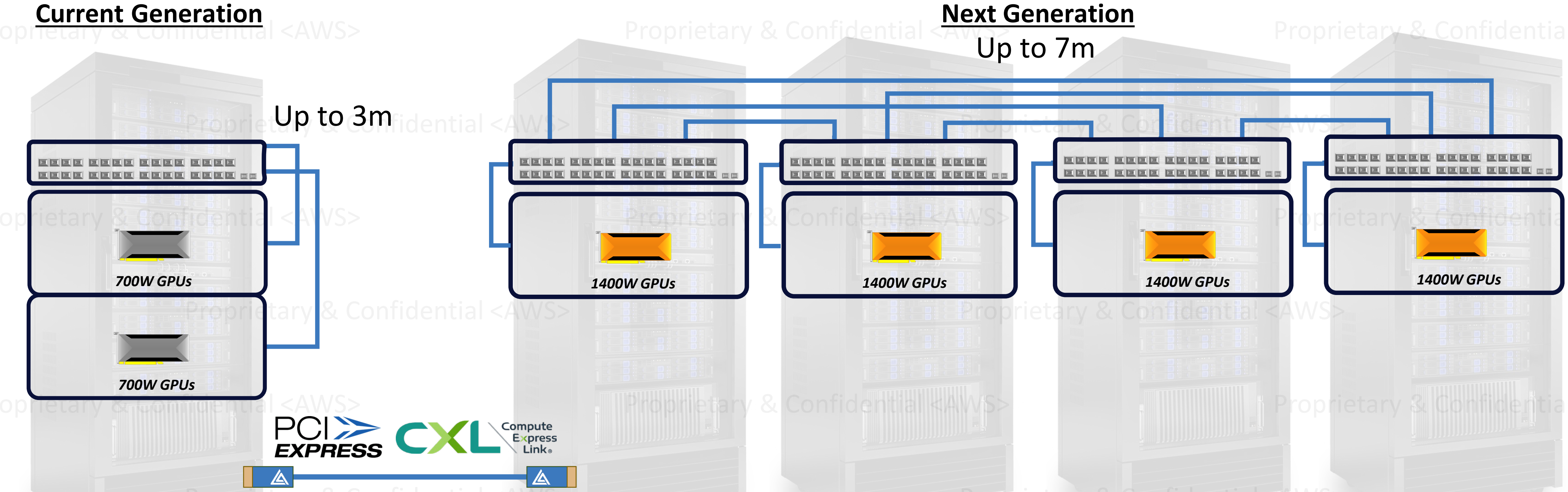
AI infrastructure under heavy pressure to scale clusters across several racks

*https://www.semianalysis.com/p/gpu-cloud-economics-explained-the

**https://www.enterpriseai.news/2023/08/07/the-great-8-bit-debate-of-artificial-intelligence/

***https://www.opencompute.org/documents/practices-and-insights-into-liquid-cooling-on-metas-ai-training-platforms-ocp-format-august-14-2023-pdf

Emerging Application: Multi Rack AI Fabric



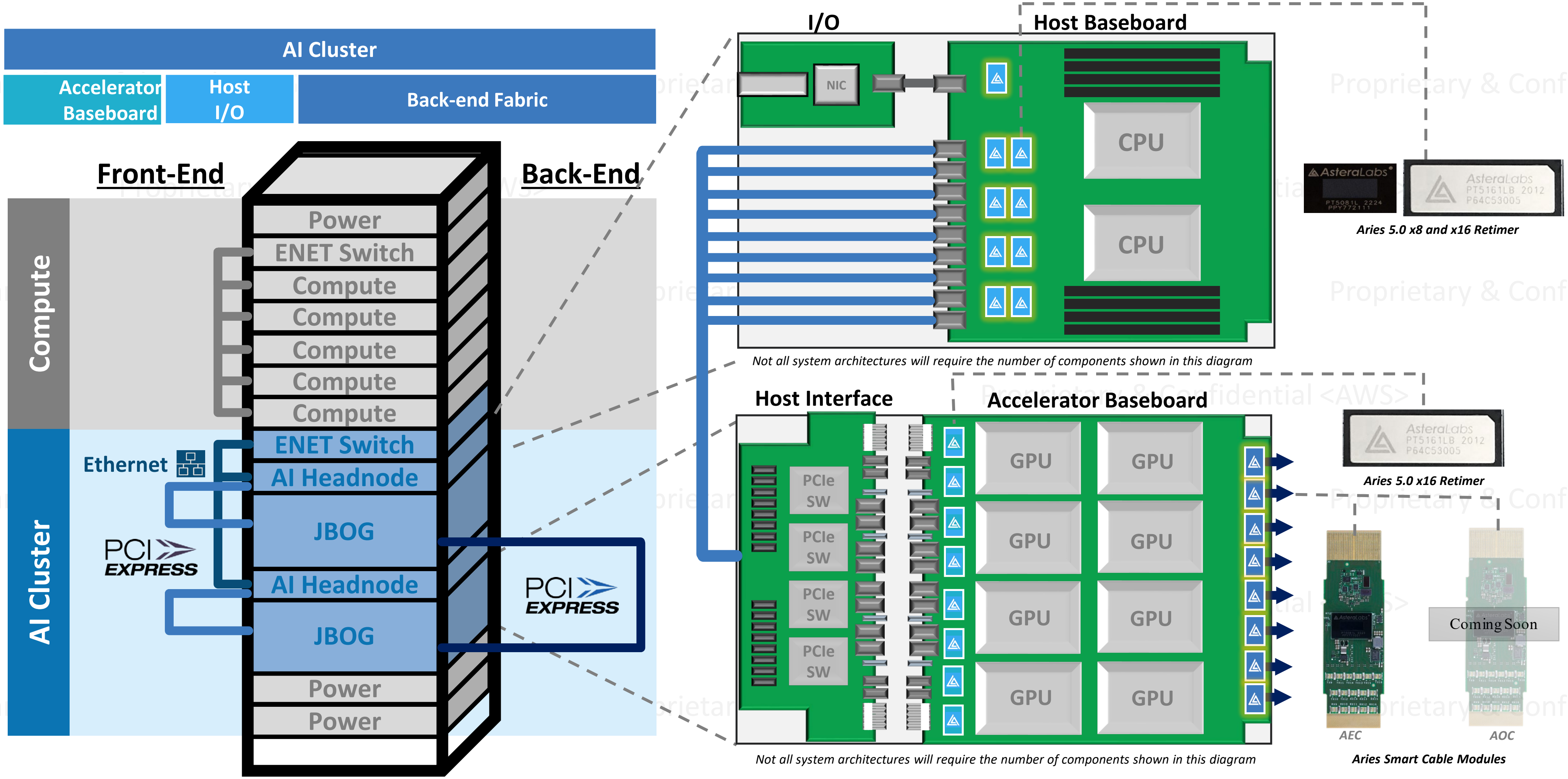
Challenges

- Rack power density
- Rack thermal density
- Increasing number of high-bandwidth links

Advantages

- Accommodates higher-performant GPUs
- Allows for larger clusters while optimizing rack power/thermals

Broadening PCIe Connectivity for the Era of AI



Aries 5.0 x8 and x16 Retimer

Not all system architectures will require the number of components shown in this diagram

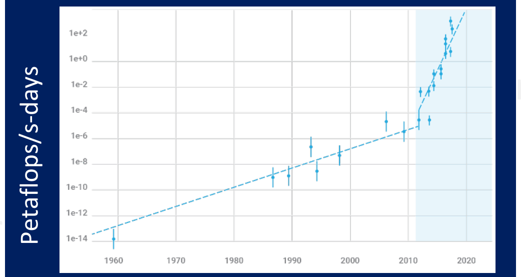
Aries 5.0 x16 Retimer

Not all system architectures will require the number of components shown in this diagram

Aries Smart Cable Modules

Memory Bottlenecks Due to AI / ML Workloads

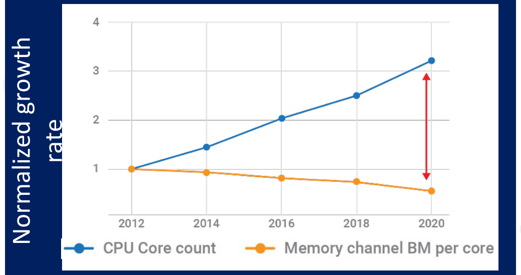
AI Model Complexity Doubling Every ~6 Months



(Source: Open AI)

Workloads need higher CPU efficiency & memory expansion

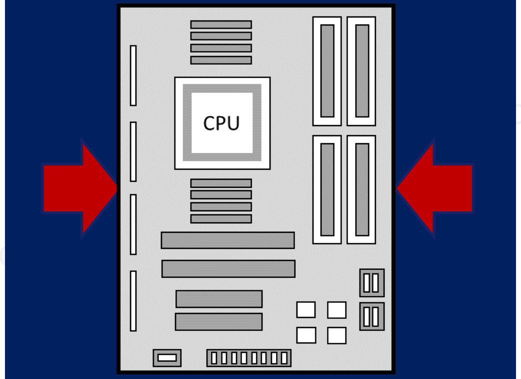
Memory Bandwidth Per Core is Declining



(Source: Meta OCP Presentation, Nov '21)

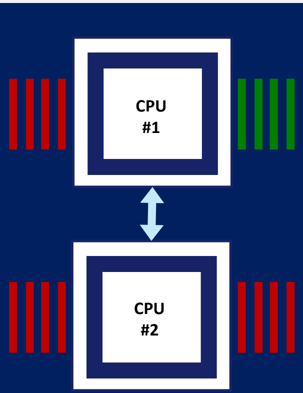
CPU efficiency is declining due to declining memory bandwidth per core

Server CPU Package & Thermal Constraints Limit Memory Channels



Memory bottlenecks caused by CPU Pin and thermal constraints

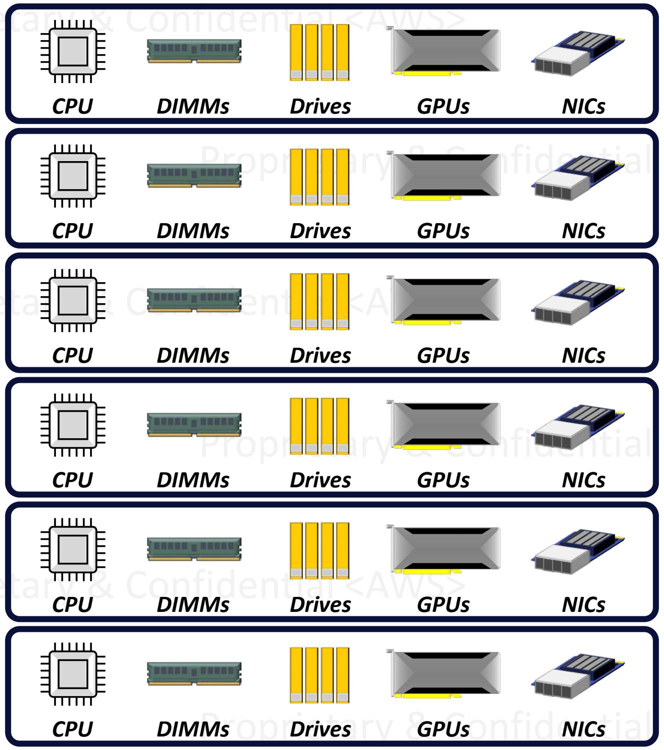
Memory Capacity is Tied to Compute Nodes



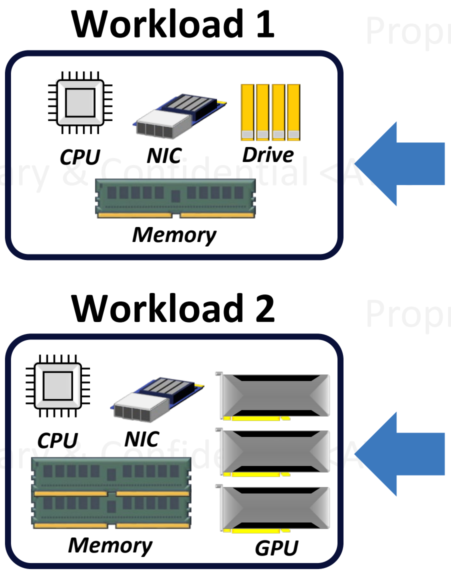
Memory is stranded behind compute leading to over provisioning

Emerging Application: Heterogeneous Infrastructure

Converged Infrastructure

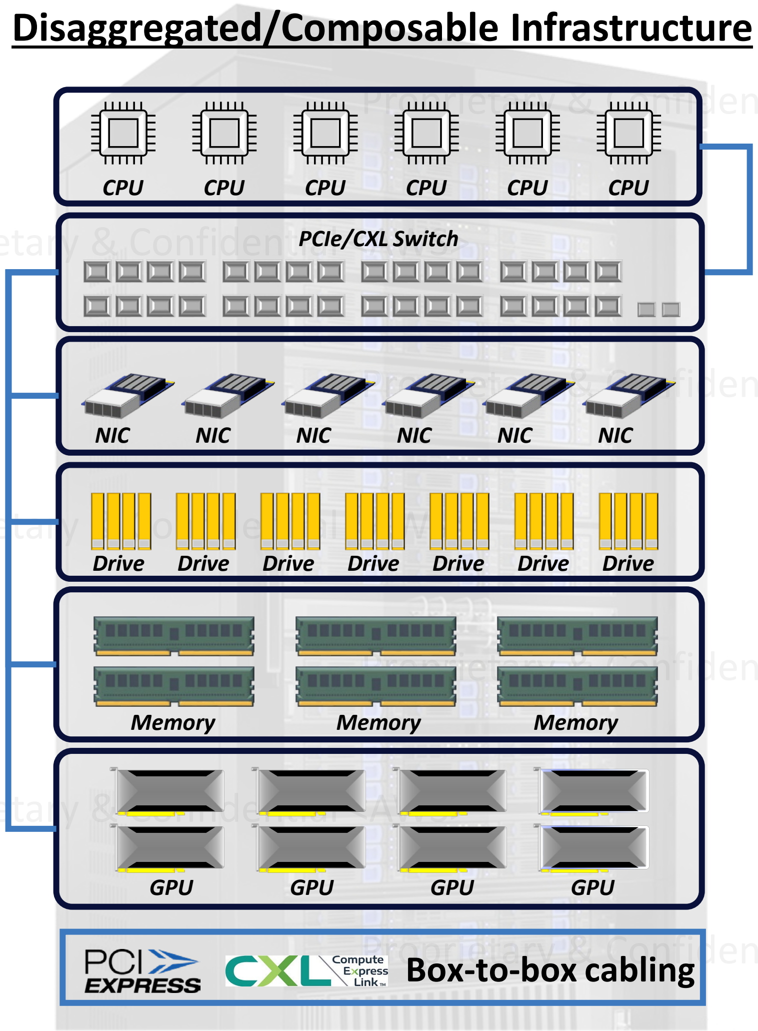


Static Configuration	**Adaptive Environment**
Stranded Resources	*Efficient Performance*
Rigid System Design	*Composable Resources*
Fixed Hosting Costs	*Flexible Cost Model*



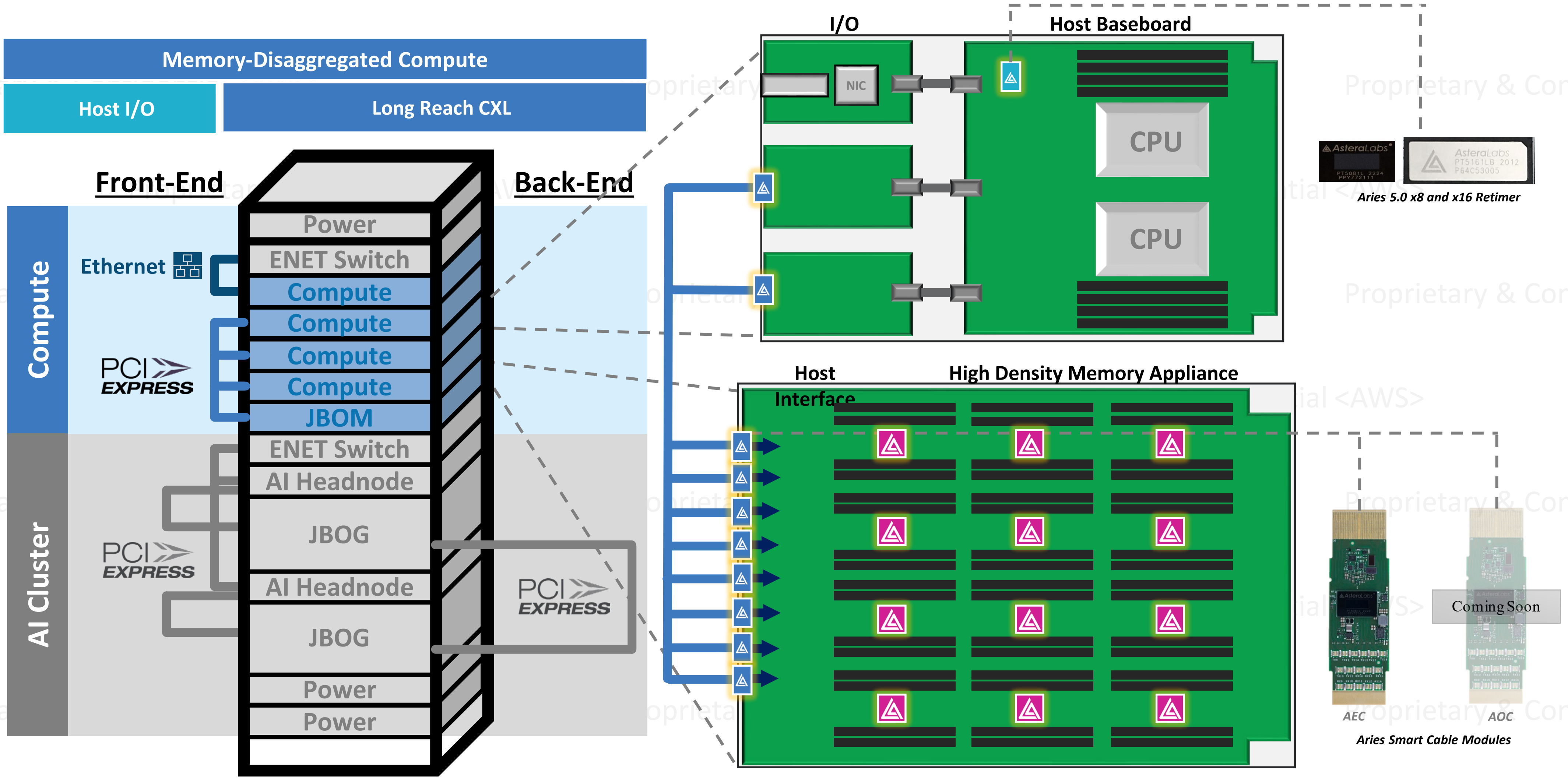
OPEX Challenges	**OPEX Advantages**
High PUE	*Low PUE*
Thermal Management	*Focused Cooling Zones*
SW Performance Tuning	*Bare-Metal Performance*

Disaggregated/Composable Infrastructure



Box-to-box cabling

Broadening PCIe Connectivity for the Era of Compute



Aries 5.0 x8 and x16 Retimer

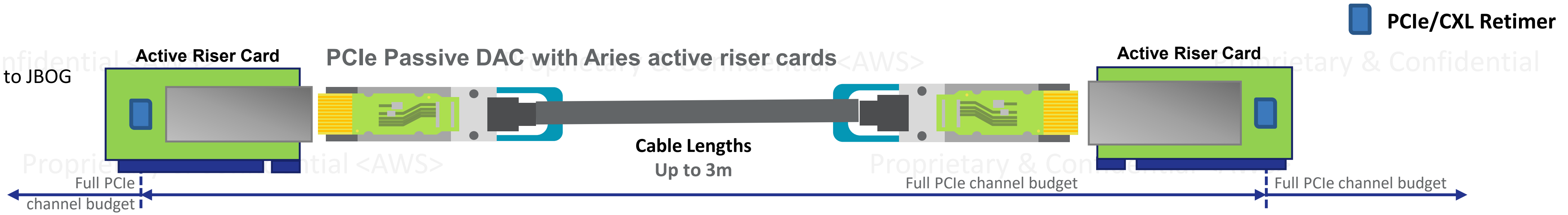
Aries Smart Cable Modules

External Cabling Reach Considerations

Within-the-rack

E.g., AI Server headnode to JBOG

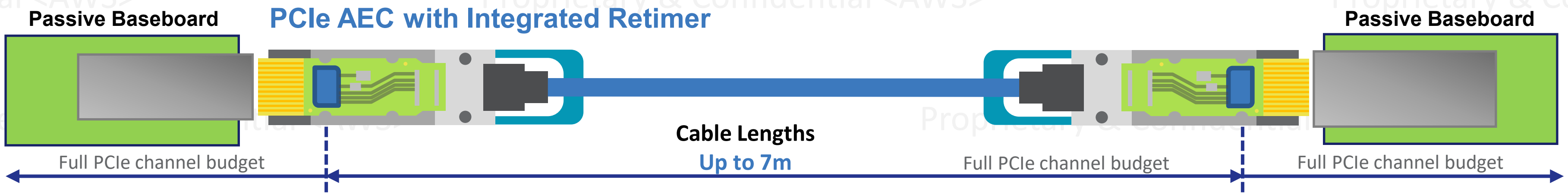
3m reach requirement



Across-racks

E.g., JBOG to JBOG

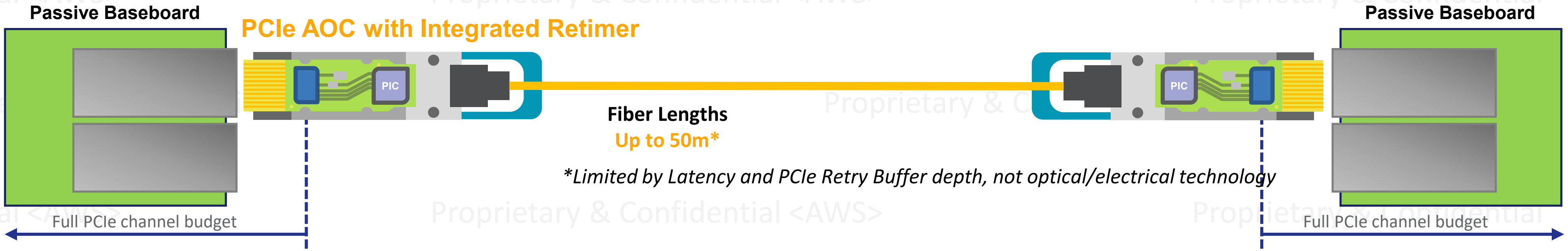
5-7m reach requirement



Across-rows

E.g., Switch to Switch (future)

20-50m reach requirement



**Limited by Latency and PCIe Retry Buffer depth, not optical/electrical technology*

PCIe/CXL AECs: Handling PCIe Side-Band Signals

Three “required” side-band signals defined in PCI-SIG’s Card Electomechanical (CEM) Specification:

PCIe Side-Band Signal	Description	Option for handling within an AEC	Alternative
REFCLK	100 MHz HCSL clock with or without spread-spectrum modulation	Dedicated differential pair to carry REFCLK from one side to the other. Pros: Allows for common clock topologies Cons: Extra cable cost, “asymmetric” cable design	No REFCLK transport in cable: SRNS/SRIS. Pros: lower cost, “symmetric” cable; scalable to multi-link AECs Cons: CC topology requires dedicated side-band cable between systems
PERST#	PCIe Protocol Reset	Dedicated single-ended line to carry PERST#. Pros: Allows PERST# synchronization on a per-link basis Cons: Extra cable cost, “asymmetric” cable design	No PERST# transport in cable. PCIe Reset events are handled through in-band Hot Reset, host-coordinated local reset, side-band management, and/or Hot Plug support. Pros: Lower cost, “symmetric cable”; scalable to multi-link AECs Cons: No dedicated per-link PERST#
PRSNT#	Cable (cable) present indicator	Pluggable cable MSAs (OSFP, OSFP-XD, etc.) include ModPrsL functionality already	N/A

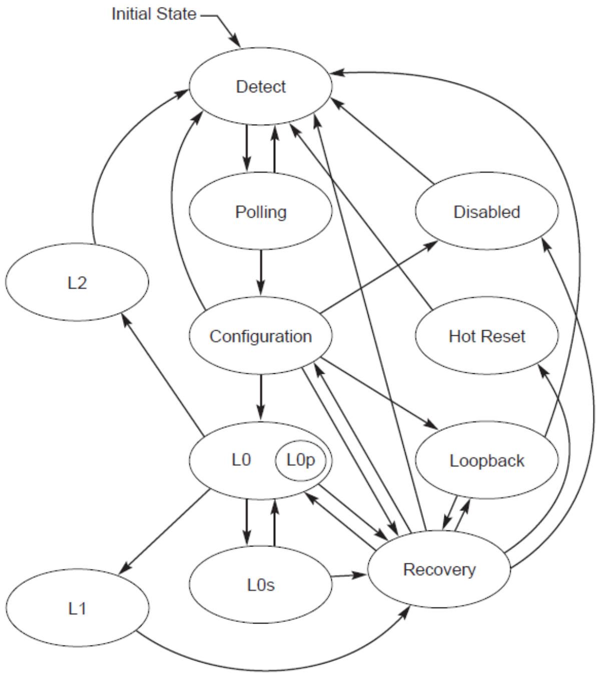
AECs: PCIe VS. Ethernet

- Two main differences:

Protocol complexity: PCIe's backwards compatibility and link training requirements make AECs more complex for PCIe compared to Ethernet

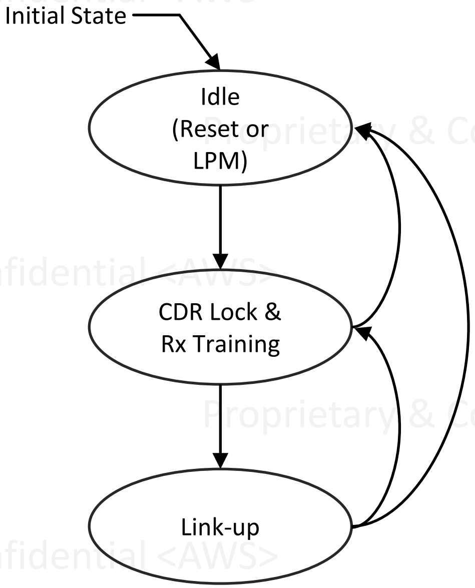
Interoperability: The variety of device types and ecosystem players is significantly more for PCIe compared to Ethernet

PCI Express (and CXL)



- Multiple "training" states for device discovery, link width/speed configuration, protocol negotiation
- Low-power, Reset, and Disabled states
- Recovery states for handling performance degradation

Ethernet



- Essentially two states—Link-Down and Link-Up—with transitory training state
- No dedicated states/protocol for performing link reset / disable
- No low-power states
- No link configuration states

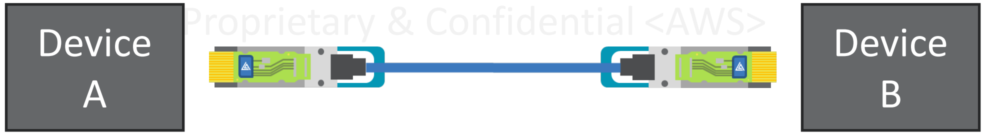
AECs: PCIe VS. Ethernet

- Two main differences:

  Protocol complexity: PCIe's backwards compatibility and link training requirements make AECs more complex for PCIe compared to Ethernet

  Interoperability: The variety of device types and ecosystem players is significantly more for PCIe compared to Ethernet

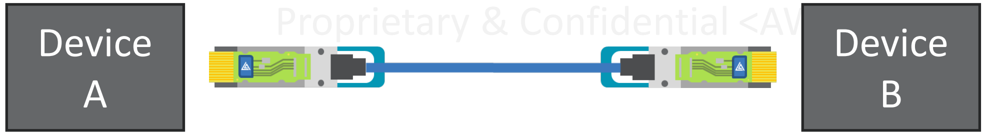
PCI Express (and CXL)



- CPU
- Switch

- Accelerator/FPGA/GPU
- Memory Controller
- NIC
- Storage
- Switch

Ethernet



- Switch

- NIC
- Switch

PCIe Cabling Form Factor Comparison

	OSFP-XD Under consideration for Optical	CDFP (x16) CopprLink	OSFP Under consideration for Optical	QSFP-DD	QSFP
High-speed lane count (full duplex)	16	16	8	8	4
X-Y PCB Size (normalized to x16)	2292 mm^2	1460 mm^2	3989 mm^2	2472 mm^2	3933 mm^2
Connector Contact Pitch	0.60 mm	0.75 mm	0.60 mm	0.80 mm	0.80 mm
Cable Gauge Supported	26-32 AWG	28-32 AWG	26-32 AWG	27-32 AWG	26-32 AWG
32 GT/s Max DAC reach (at max gauge)	4 m	3.0 m	4 m	3.5 m	4 m
32 GT/s Max AEC reach (at max gauge)	7 m	5.5 m	7 m	6 m	7 m
64 GT/s Max DAC reach (at max gauge)	3 m	2.5 m	3 m	2.5 m	3 m
64 GT/s Max AEC reach (at max gauge)	6 m	5 m	6 m	5 m	6 m
Active Copper cable	Yes	No	Yes	Yes	Yes
Active Optical cable	Yes	No	Yes	Yes	Yes
Power Capability per Lane	8x2.5A@3.3V 66W/16 = 4.125W	1x1.5A@12V + 1x1.5A@3.3V 23W/16= 1.44W	4x2.5A@3.3V 33W/8= 4.125W	6x1.5A@3.3V 30W/8= 3.75W	3x1.0A@3.3V 10W/4= 2.5W

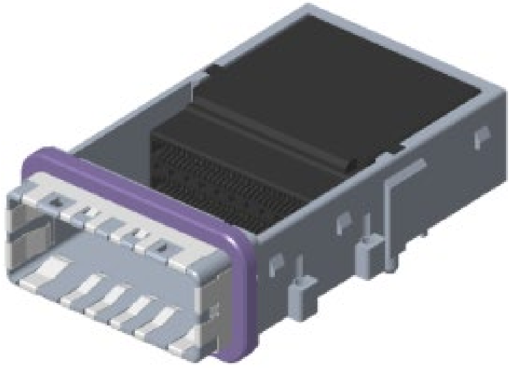
Assumptions:

- Twinax losses: 28/27/26AWG=4.3/4.0/3.6 + 10% dB/m at 16 GHz.
- AEC: Retimer silicon to cable pads: 4 dB @ 16 GHz
- DAC: Retimer silicon (behind cage) to passive DAC cable pads: 9.5 dB @ 16 GHz

Reference: https://drive.google.com/file/d/12ZSTkIgkzESbf4fZzj7WQBi3y4oto-gB/view

CopprLink and Active External Cables

- SFF-TA-1032 (CDFP) uses two physical paddle cards inside a cable assembly
- This presents a significant challenge: **How can you connect Tx and Rx signals from separate paddle cards into a Retimer component?**
- Rx and Tx both terminating in the Retimer is necessary for:
  - Equalization Phase 2/3 training
  - In-band lane margining

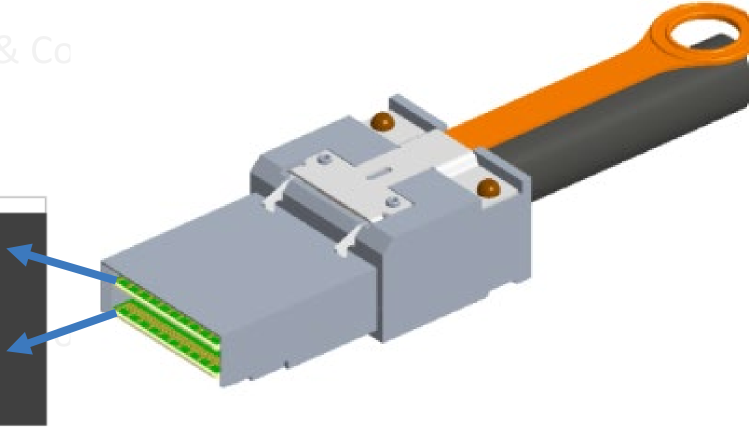


Upper paddle card pinout

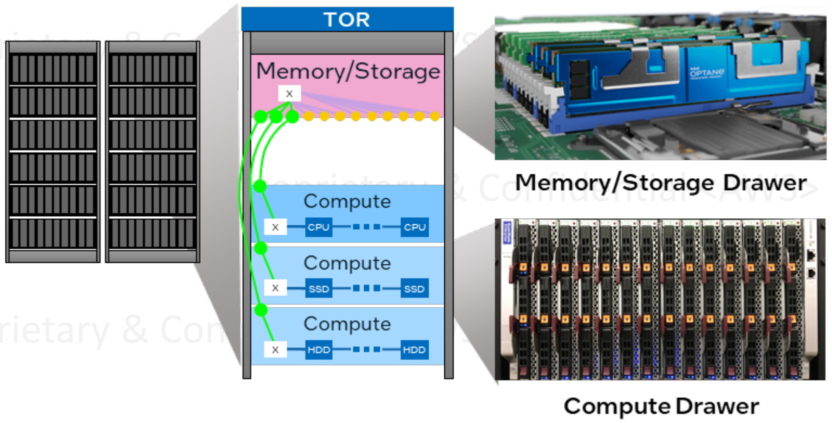
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
D	GND	PERp11	PERn11	GND	PERp10	PERn10	GND	PERp9	PERn9	GND	PERp8	PERn8	GND	FLEXIO7	FLEXIO8
C	GND	PERp12	PERn12	GND	PERp13	PERn13	GND	PERp14	PERn14	GND	PERp15	PERn15	GND	FLEXIO5	FLEXIO6
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15
B	GND	PETp11	PETn11	GND	PETp10	PETn10	GND	PETp9	PETn9	GND	PETp8	PETn8	GND	FLEXIO3	FLEXIO4
A	GND	PETp12	PETn12	GND	PETp13	PETn13	GND	PETp14	PETn14	GND	PETp15	PETn15	GND	FLEXIO1	FLEXIO2
	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30
D	GND	PERp7	PERn7	GND	PERp6	PERn6	GND	PERp5	PERn5	GND	PERp4	PERn4	GND	SCL	SDA
C	GND	PERp0	PERn0	GND	PERp1	PERn1	GND	PERp2	PERn2	GND	PERp3	PERn3	GND	VCC3p3V	PERST#
	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26	C27	C28	C29	C30
	B16	B17	B18	B19	B20	B21	B22	B23	B24	B25	B26	B27	B28	B29	B30
B	GND	PETp7	PETn7	GND	PETp6	PETn6	GND	PETp5	PETn5	GND	PETp4	PETn4	GND	VCC12V	PR/PE
A	GND	PETp0	PETn0	GND	PETp1	PETn1	GND	PETp2	PETn2	GND	PETp3	PETn3	GND	2WCL	2WDA
	A16	A17	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30

Lower paddle card pinout



Wrap Up



PCIe Passive DAC with Aries active riser cards

PCIe AEC with Integrated Retimer

PCIe AOC with Integrated Retimer

- Evolving AI and disaggregated compute system topologies require more **external cabling**
- Reach requirements vary from **2m** (within the rack), to **7m** (rack to rack), and **beyond** (larger clusters)
- Retimer-based AEC and optical solutions enable reach extension while presenting an easy-to-design-to **PCIe compliance point** to the host/device
- Implementing PCIe AEC and optical involves higher design complexity in terms of **protocol and interoperability** as compared to Ethernet
- **OSFP-XD/OSFP** represents an attractive option for PCIe/CXL x16/x8 applications, allowing for passive DAC, AEC, and Optical solutions

FMS

Thank You

AsteraLabs®

Check us out on

www.asteralabs.com