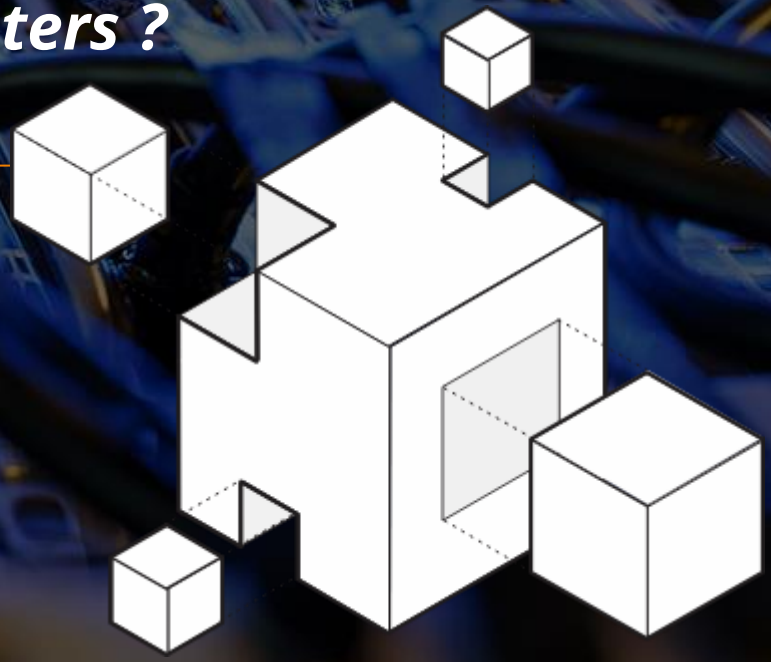
2crsi

Disaggregation : future of data centers ?

FRNOG 38 Meeting – Oct. 6th, 2023



Xavier LE VAILLANT
2crsi.com

Disclaimer: The views and opinions expressed in this presentation are those of the speaker and do not necessarily reflect the views or positions of any entities he represent.

First 1U server supporting up to 32 High-End GPUs !

BREAKING NEWS

Mona 1.12GG

High-end server
based on AMD EPYC™ 4th Gen CPU
with 12x 2.5" SFF Drives

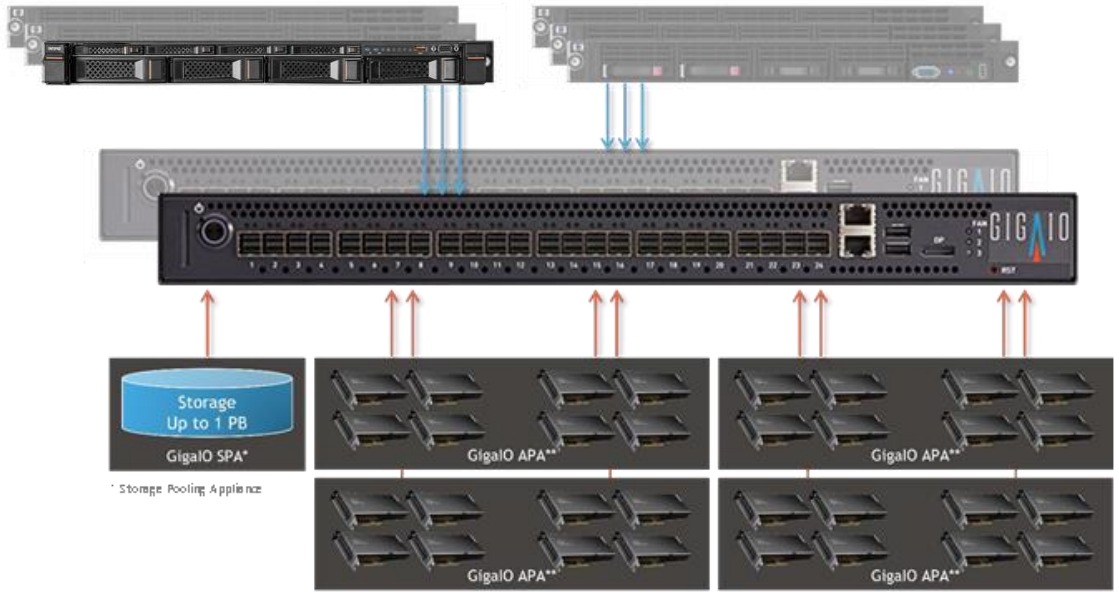
Key Features

- 19-inch - 1U/12x 2.5"
- Dual AMD EPYC™ 4th Gen socket
- 24x DDR5 @ 4800MHz
- 2x PCIe 5.0 x16 HHHL
  1x PCIe 5.0 x16
  2x OCP 3.0 x16
- Air Cooling

SCAN THE CODE!
TO DISCOVER MORE
ABOUT THIS PRODUCT

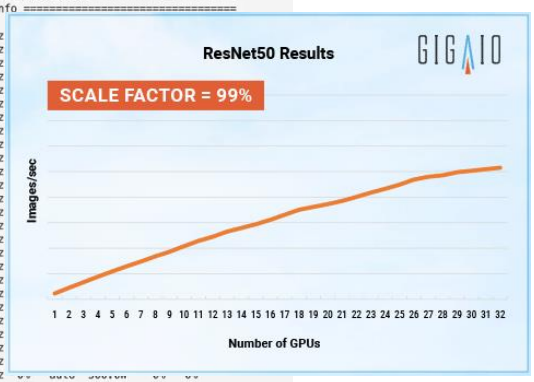
First ~~1U server~~ system supporting up to 32 High-End GPUs !

SuperNode™



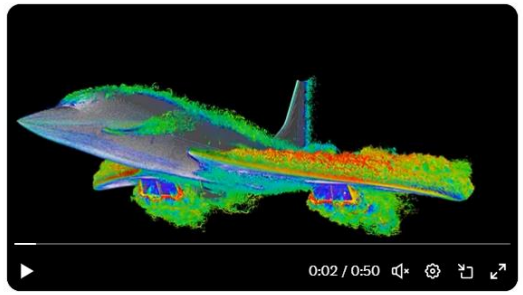
https://gigaio.com/supernode/

```
ssh user@supermilan1
- AMD GPU        32
- Xilinx FPGA    0
- NVME SSD       1
- NVIDIA GPU     0
- Infiniband     0

FabreX Network Adapter, Gen4 (FA4004)

======================== ROCm System Management Interface ========================
================================== Concise Info ==================================
GPU  Temp (DieEdge)  AvgPwr  SCLK    MCLK
0    53.0c           42.0W   800Mhz  1600Mhz
1    57.0c           44.0W   800Mhz  1600Mhz
2    66.0c           44.0W   800Mhz  1600Mhz
3    59.0c           43.0W   800Mhz  1600Mhz
4    62.0c           44.0W   800Mhz  1600Mhz
5    57.0c           42.0W   800Mhz  1600Mhz
6    59.0c           42.0W   800Mhz  1600Mhz
7    61.0c           42.0W   800Mhz  1600Mhz
8    48.0c           42.0W   800Mhz  1600Mhz
9    49.0c           40.0W   800Mhz  1600Mhz
10   53.0c           40.0W   800Mhz  1600Mhz
11   51.0c           42.0W   800Mhz  1600Mhz
12   52.0c           43.0W   800Mhz  1600Mhz
13   50.0c           40.0W   800Mhz  1600Mhz
14   53.0c           42.0W   800Mhz  1600Mhz
15   53.0c           42.0W   800Mhz  1600Mhz
16   48.0c           42.0W   800Mhz  1600Mhz
17   51.0c           42.0W   800Mhz  1600Mhz
18   59.0c           45.0W   800Mhz  1600Mhz
19   55.0c           44.0W   800Mhz  1600Mhz
20   63.0c           44.0W   800Mhz  1600Mhz
21   59.0c           47.0W   800Mhz  1600Mhz
22   55.0c           43.0W   800Mhz  1600Mhz
23   52.0c           44.0W   800Mhz  1600Mhz
24   57.0c           44.0W   800Mhz  1600Mhz
25   55.0c           42.0W   800Mhz  1600Mhz
26   64.0c           44.0W   800Mhz  1600Mhz  0%   auto  300.0W   0%   0%
27   55.0c           42.0W   800Mhz  1600Mhz  0%   auto  300.0W   0%   0%
28   58.0c           42.0W   800Mhz  1600Mhz  0%   auto  300.0W   0%   0%
29   56.0c           45.0W   800Mhz  1600Mhz  0%   auto  300.0W   0%   0%
30   65.0c           45.0W   800Mhz  1600Mhz  0%   auto  300.0W   0%   0%
31   60.0c
```



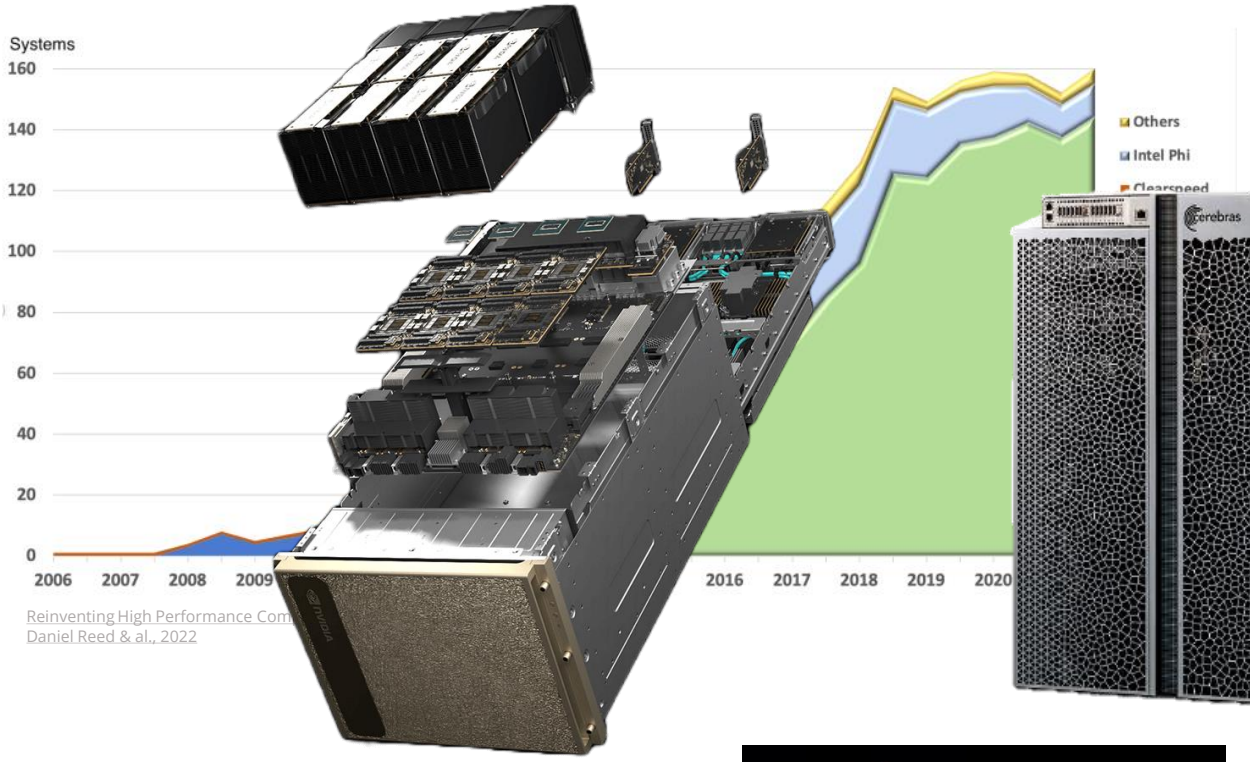
Dr. Moritz Lehmann
@ProjectPhysX

Over the weekend I got to test #FluidX3D on the world's largest #HPC #GPU server, #GigaIOSuperNODE. Here is one of the largest #CFD simulations ever, the Concorde for 1s at 300km/h landing speed. 40 *Billion* cells resolution. 33h runtime on 32 @AMDInstinct MI210, 2TB VRAM.
1/5



More computing power… with Accelerators

Systems Using Accelerators on the TOP500

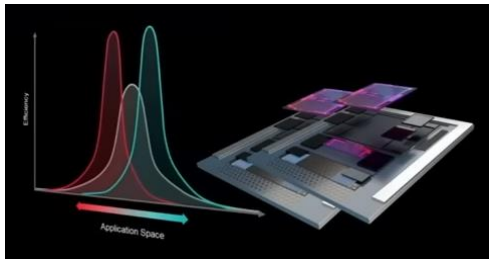


Reinventing High Performance Com
Daniel Reed & al., 2022

Age of Accelerators

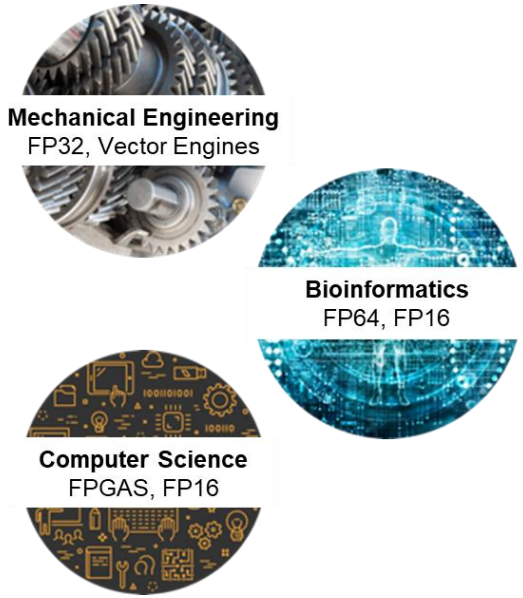
- GPUs, FPGAs, Vector
- rators needed
- nt accelerators
- ntly

Domain-specific computation could drive performance efficiency

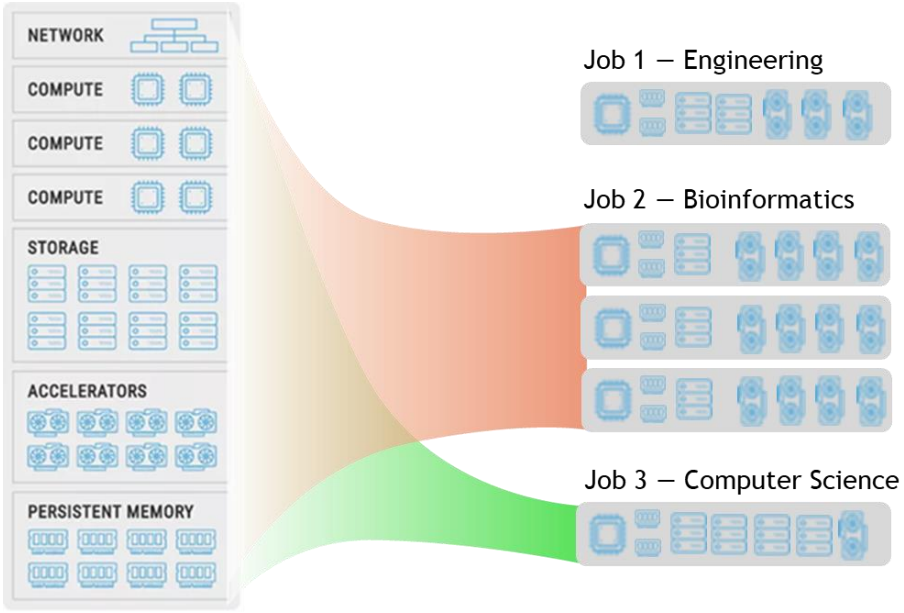


One-size-fits-all architectures are outdated

e.g., different research activities



Defines hardware uniquely for each workload



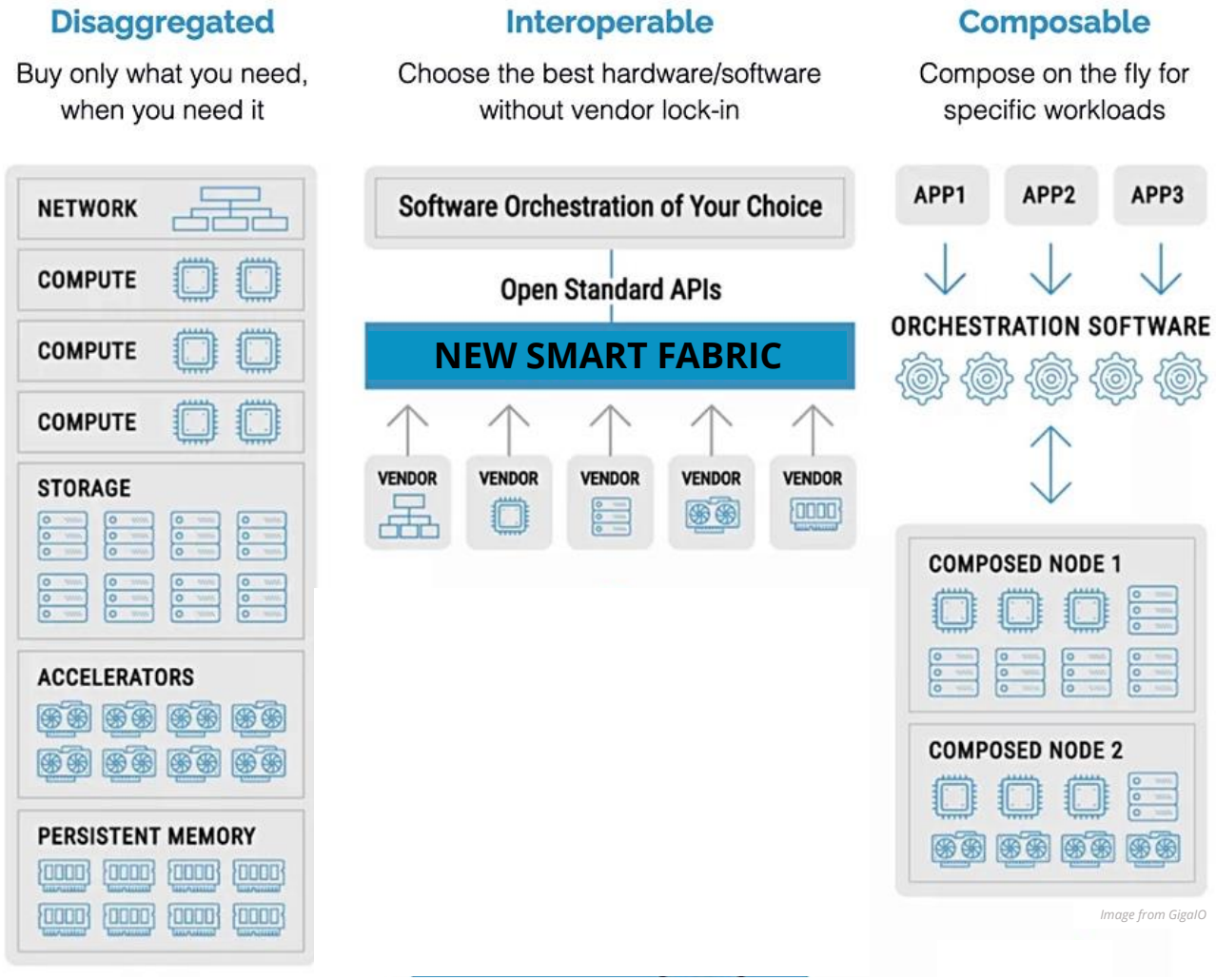
- **The End of Stranded Resources**
  - The right resources in the right place, on demand
- **Scale-Up and Scale-Out as You Grow**
  - In-place scaling and selection of servers and accelerators as requirements evolve
- **True Heterogeneity**
  - The right GPUs, servers and accelerators for the job

Composable Disaggregated Infrastructure

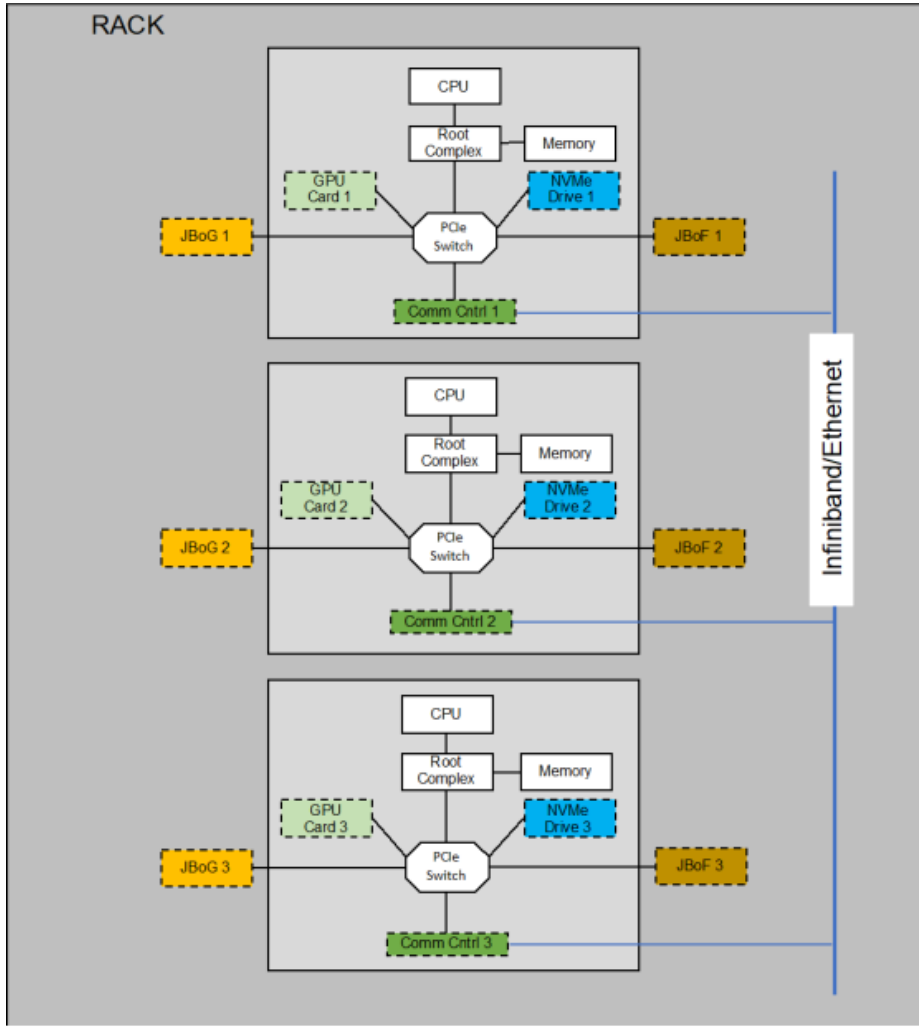
Composable Disaggregated Infrastructure (CDI) brings the agility, savings and efficient resource-sharing of the cloud to the management of on-premises equipment.

Using orchestration and high-bandwidth, low-latency fabrics, shared resources can be combined on-demand for shifting workloads. The goal is to get the right ratio for a specific AI training or inference job, change configurations as the workload pipeline changes, and free up expensive GPUs and other accelerators for additional work.

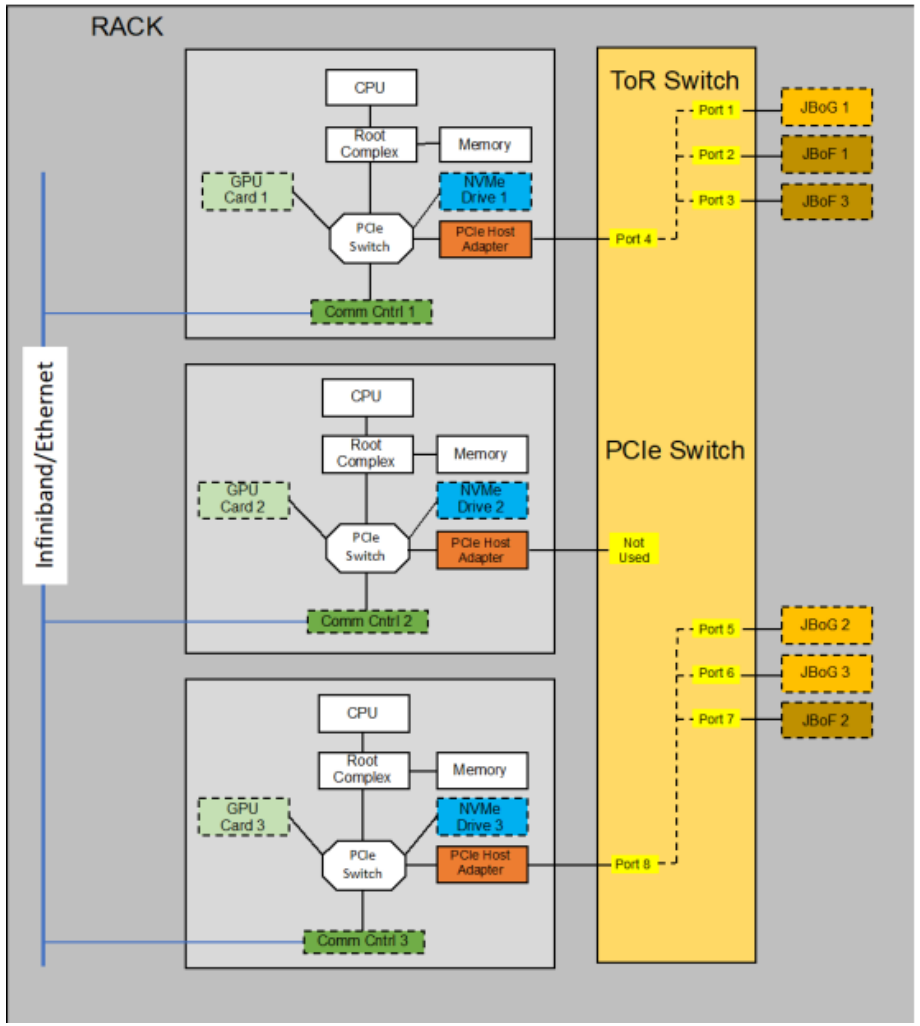
"Software-Defined Hardware"



PCIe-based Fabric

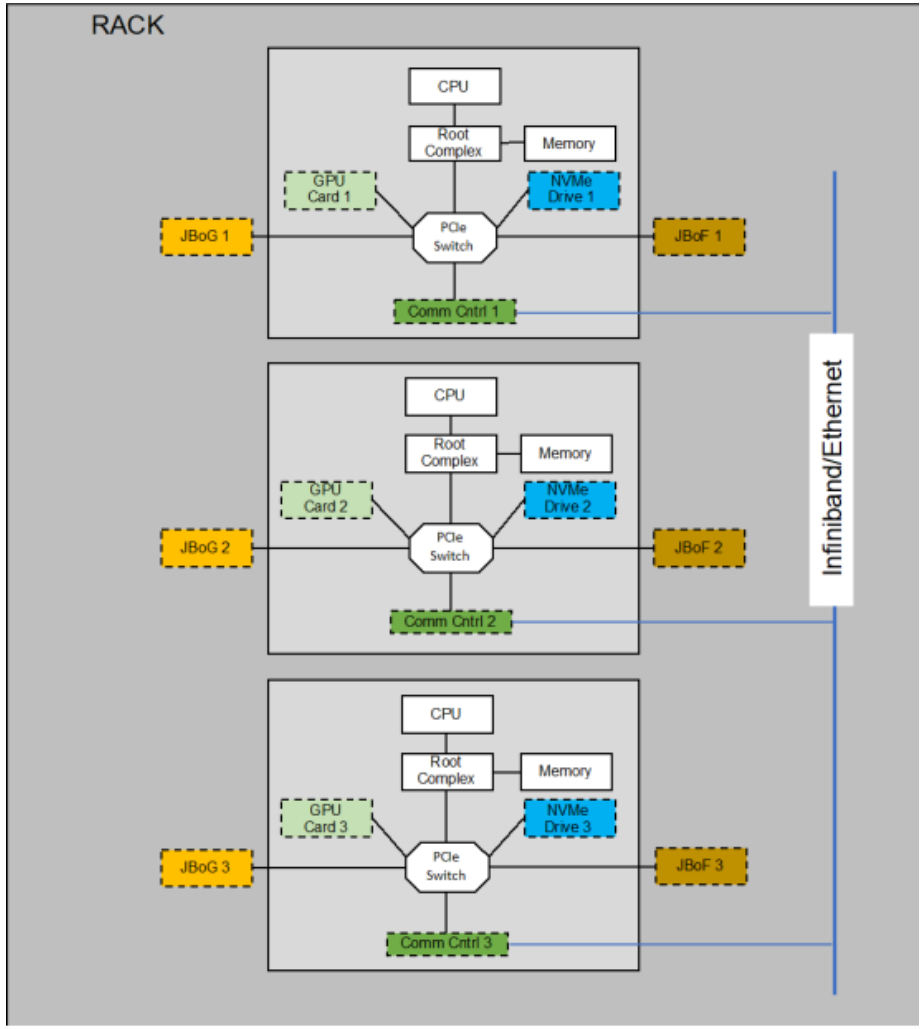


Example of a Rack scale system

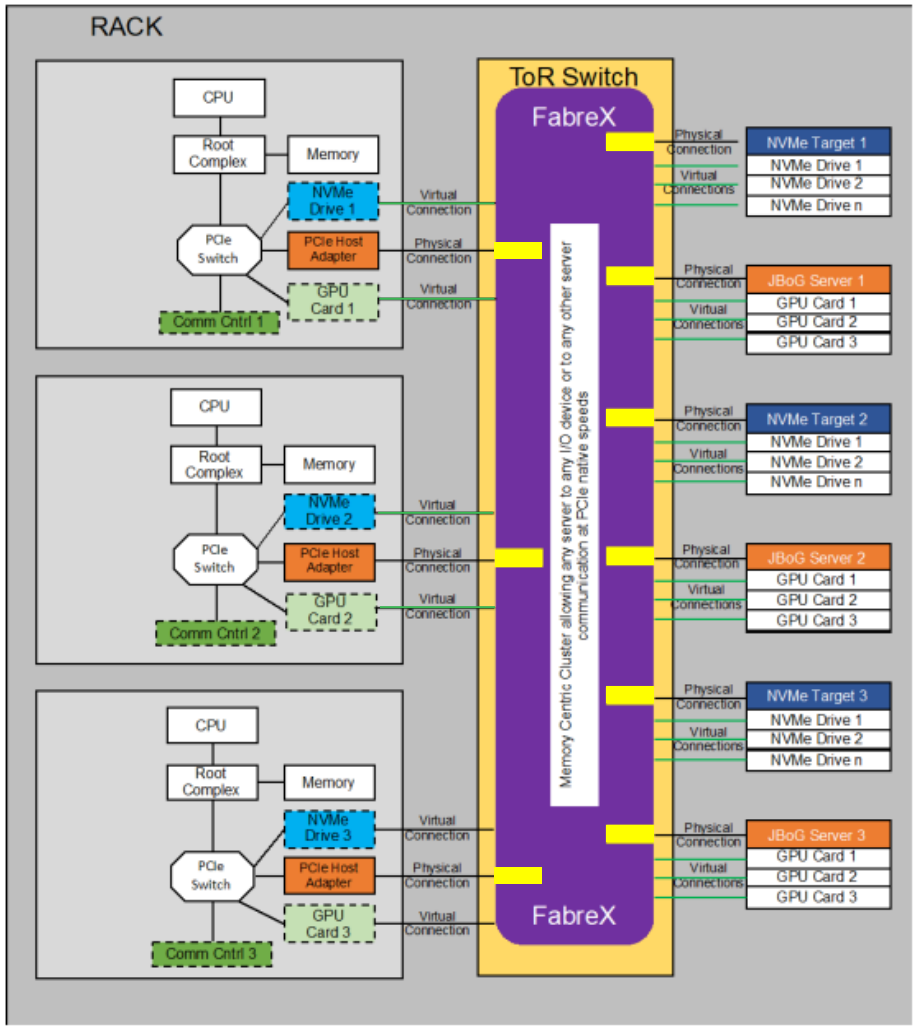


PCIe Switch

PCIe-based Fabric

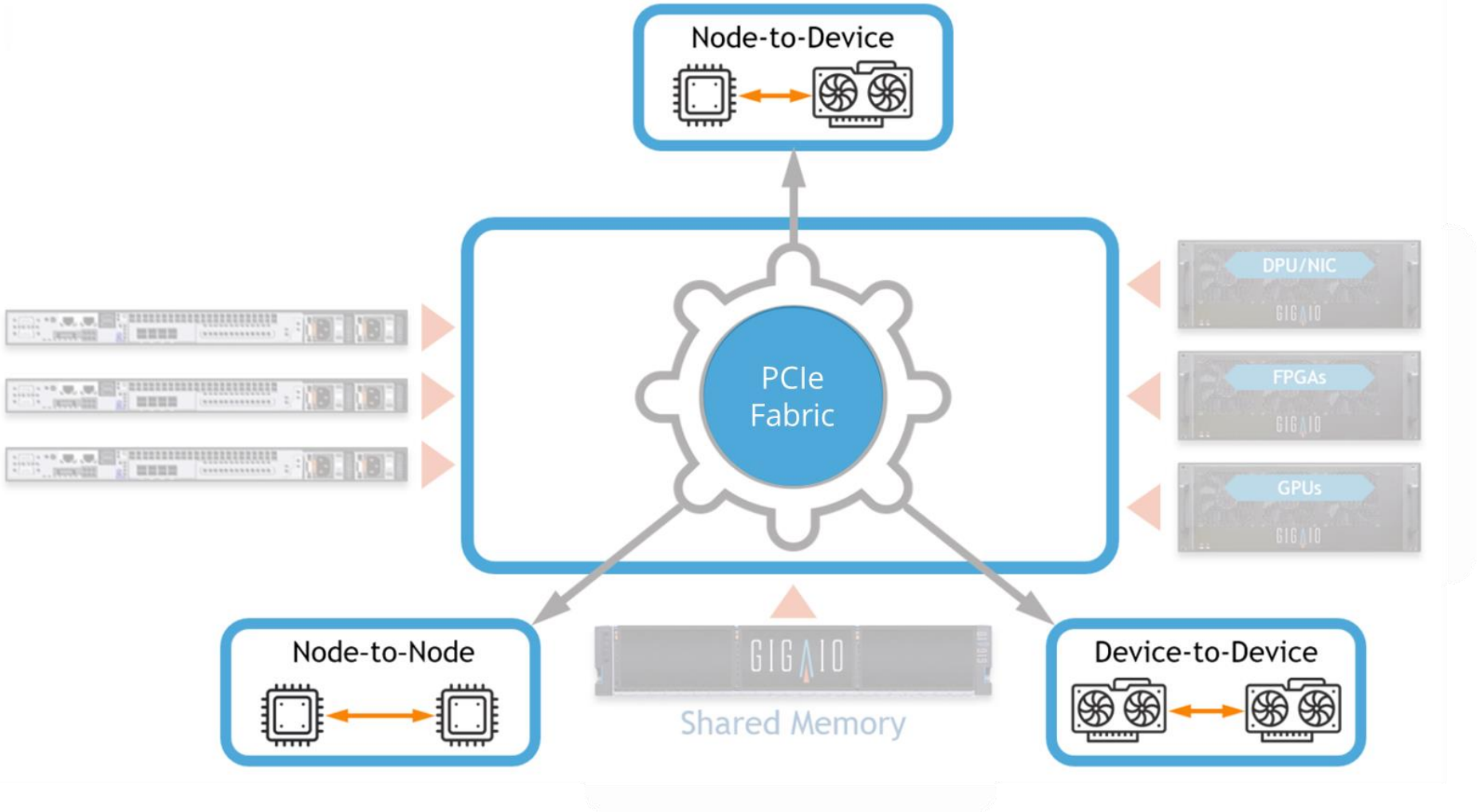


Example of a Rack scale system



PCIe-based Fabric

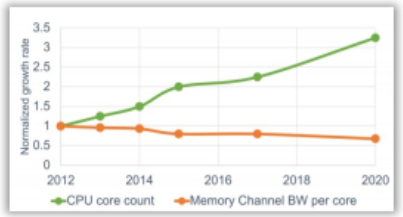
Composable Resources



Summary of Data Center Memory Challenges

BANDWIDTH

Decreasing memory bandwidth per core

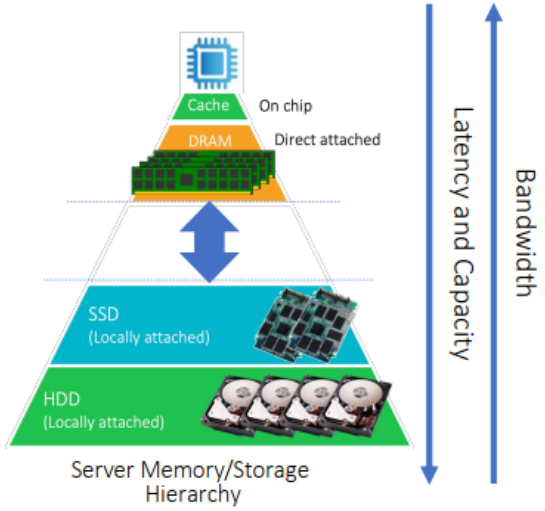


OCP GLOBAL SUMMIT

Source: Meta, OCP Summit Presentation Nov. '21

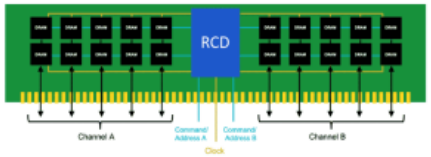
CAPACITY

Huge latency and capacity gap



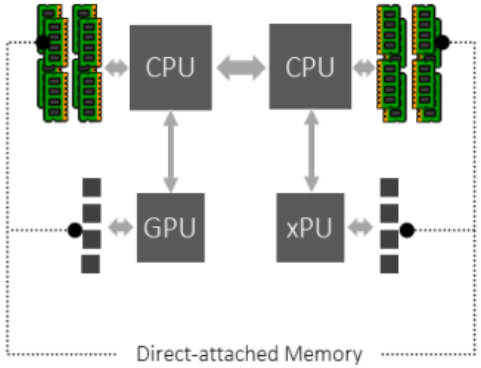
COST

Cost to achieve desired memory density and bandwidth



EFFICIENCY

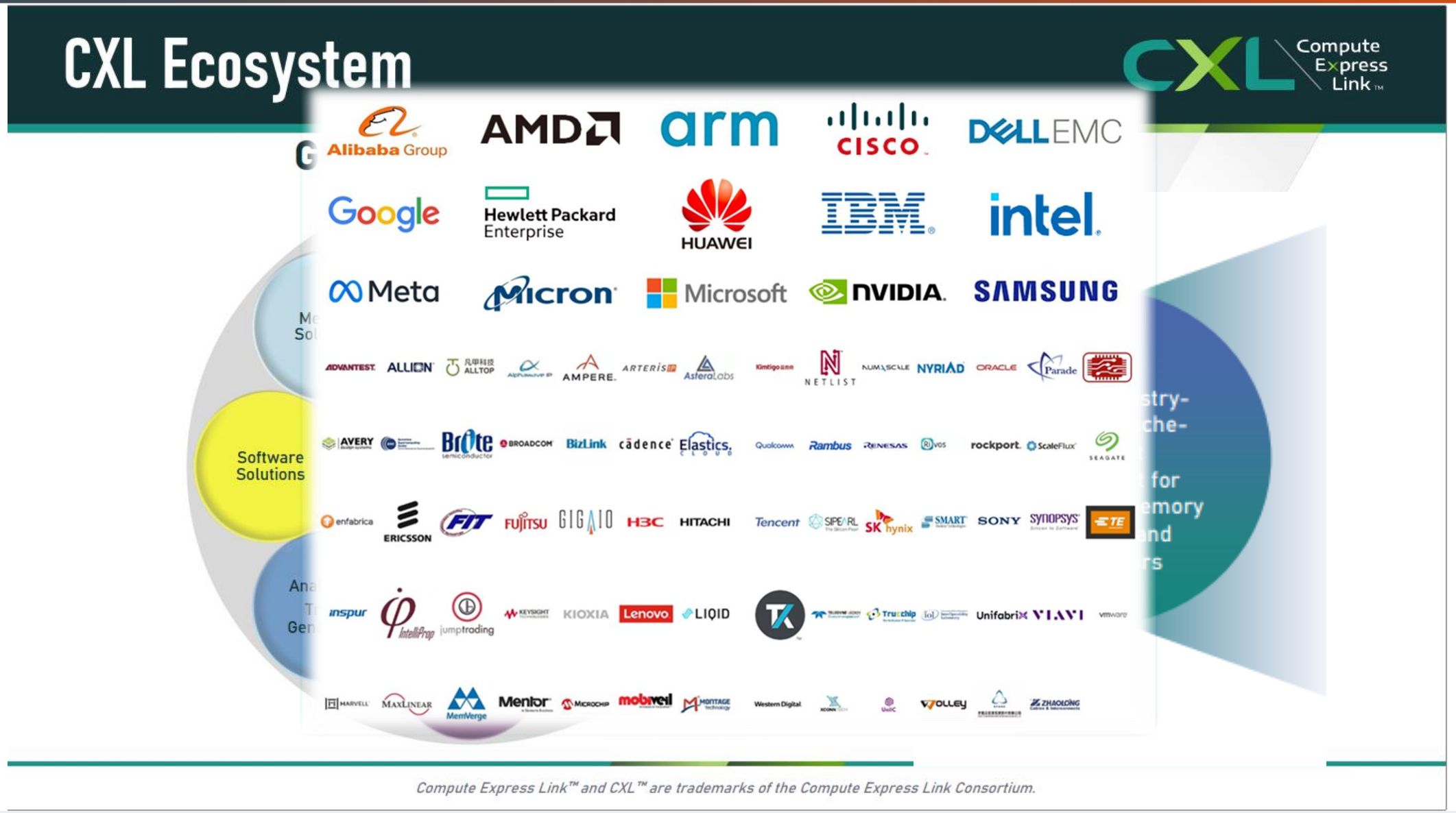
Stranded memory resources and low utilization



Decoupling the Memory Controller from the CPU and Providing Options for New Server Architectures to Address Memory Challenges

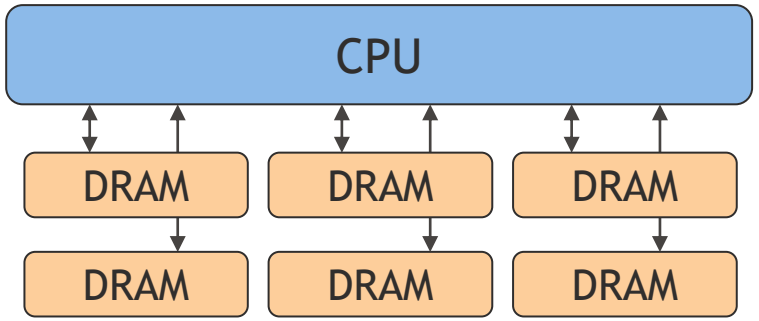
CXL Forum at FMS2023 - Advancing Data Center Architectures with Memory Tiering, Rambus

Scope of CXL

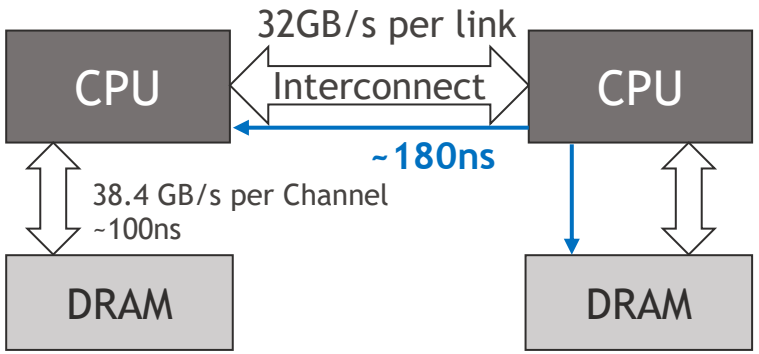


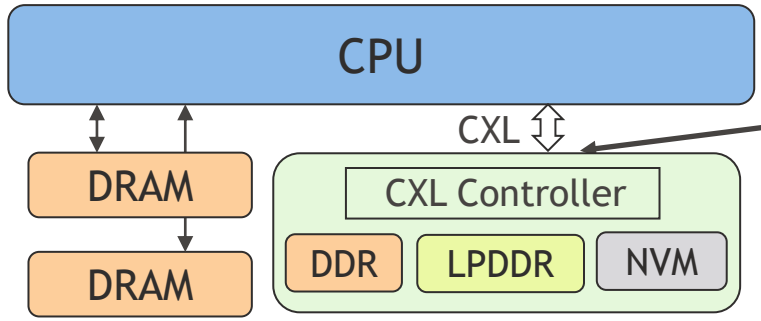
CXL Forum at FMS 2023 - Advancing coherent connectivity. CXL Consortium

Key Design aspects of CXL Memory

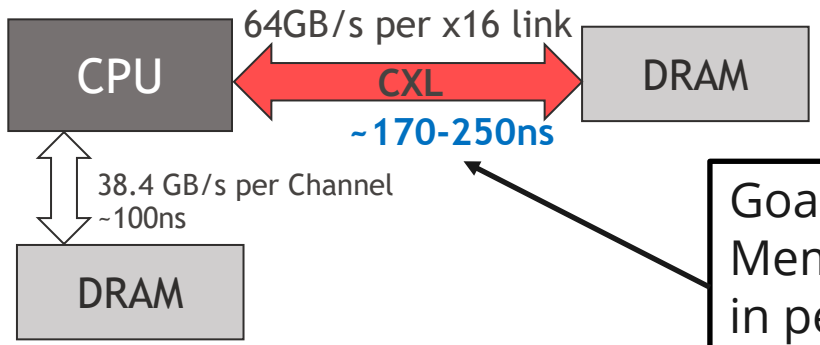


Today Without CXL





With CXL



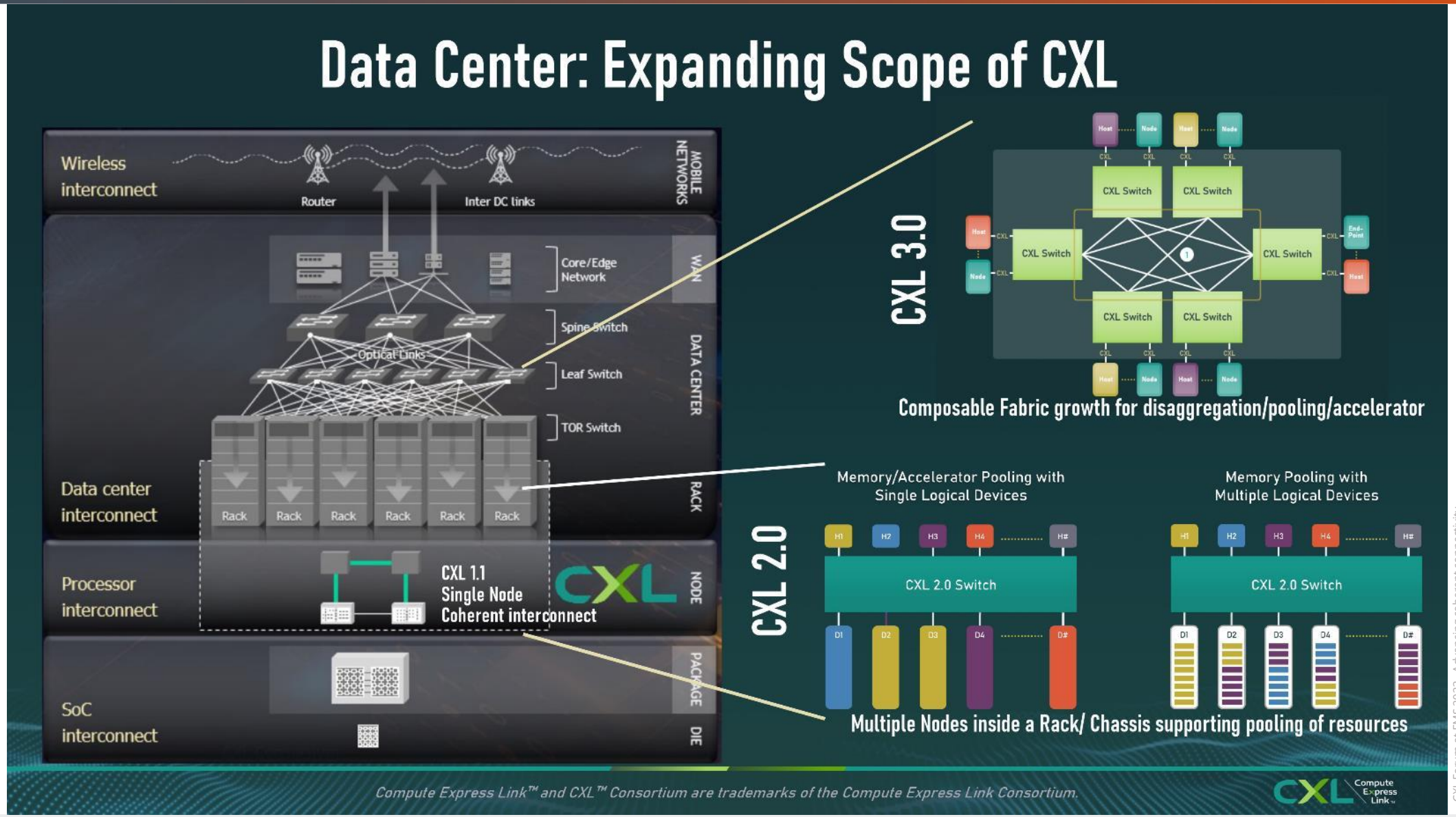
CXL moves memory to a serial comm link and creates Controller

Frees size and type from CPU, potentially allowing much:
- Larger amounts of memory
- Cheaper memory

Goal is for CXL Memory to be equal in perf to 2nd processor memory

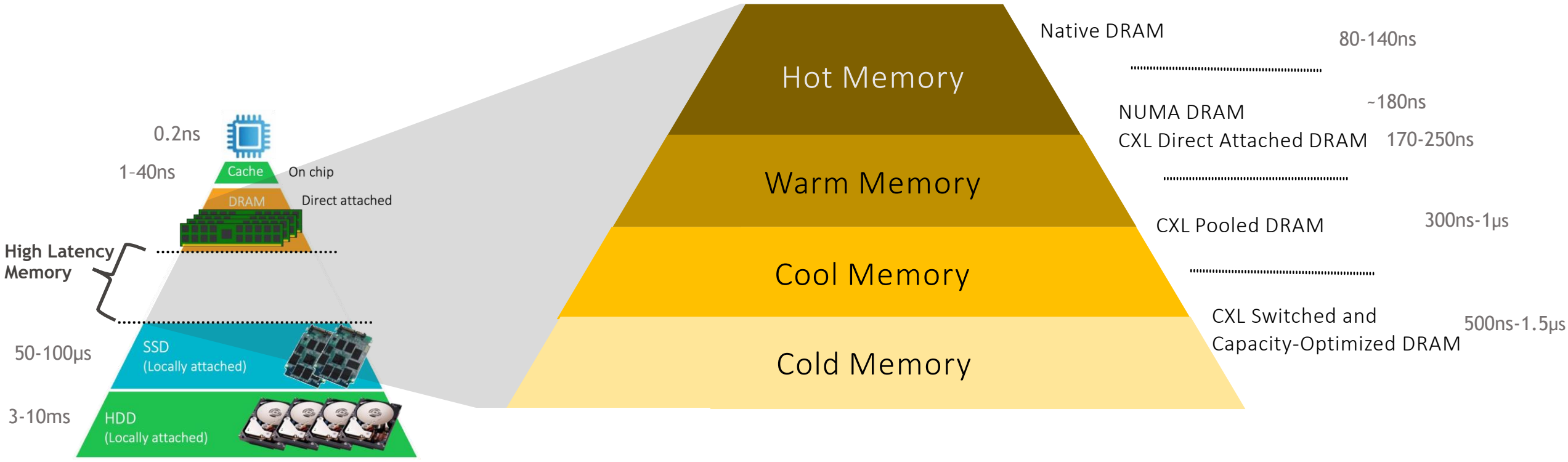
Source: CXL Forum ISC 23 - The March of Composability – Onward to Memory with CXL, GigaIO

Scope of CXL



CXL Forum at FMS 2023 - Advancing coherent connectivity, CXL Consortium

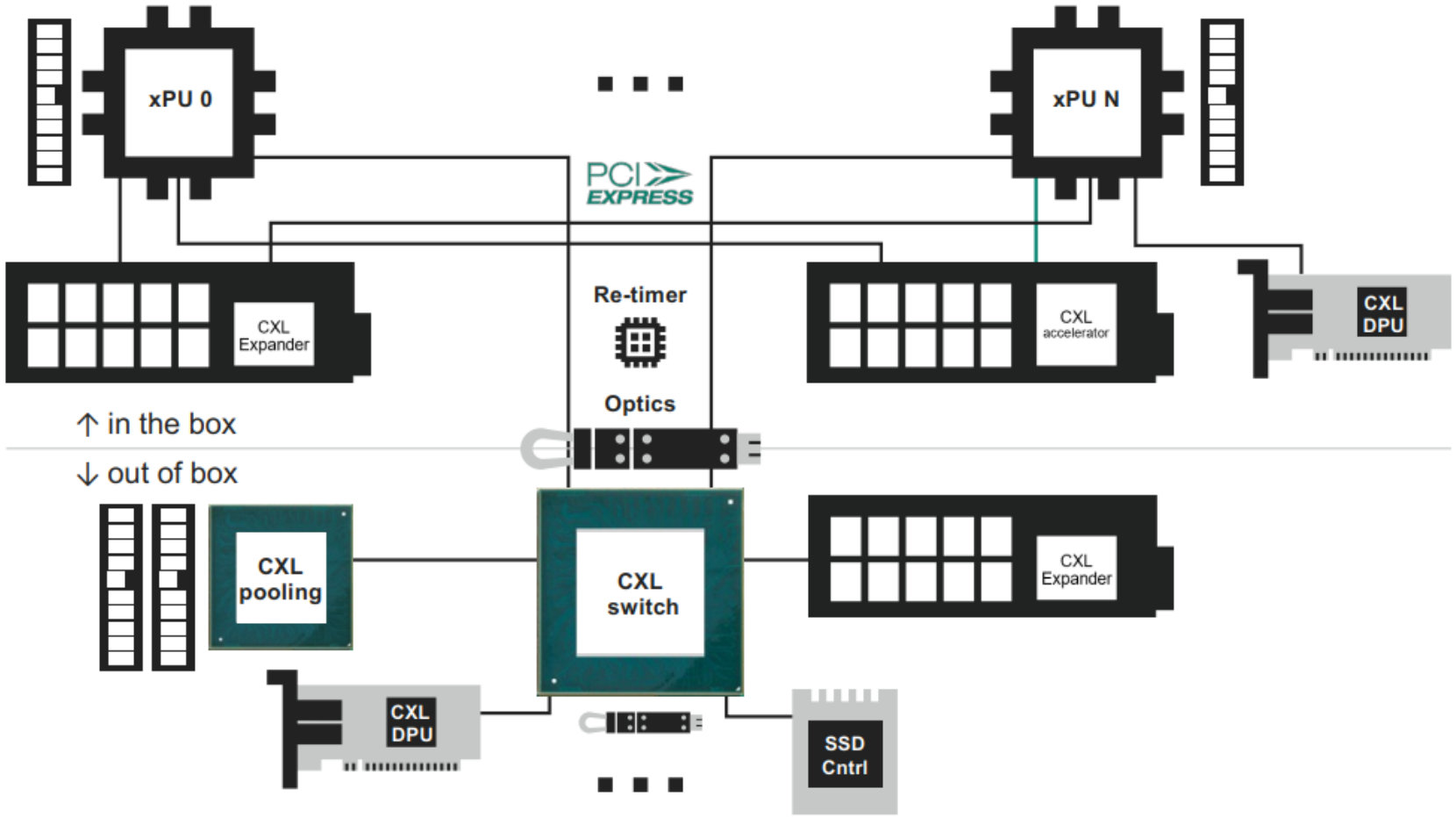
Memory tiers – Span the Latency Gap



- CXL delivers new expansion options for hot DRAM, with no impact to software applications
- CXL also introduces memory tiering, to the Data Center, much like storage tiering before it
- The industry is now working on software infrastructure to take advantage of these new tiers

Source: CXL Forum at FMS2023 - Advancing Data Center Architectures with Memory Tiering, Rambus

Comprehensive end-to-end CXL solutions

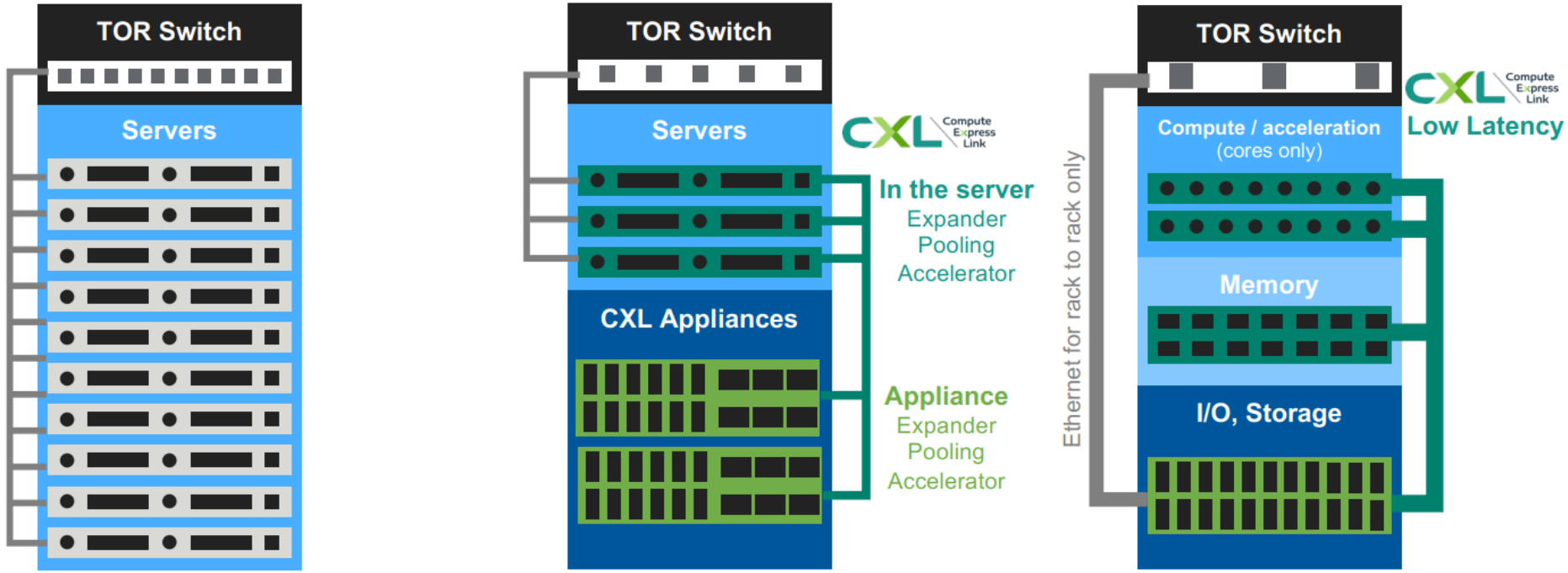


CXL use cases

- Expanders
- Pooling
- Switch
- Accelerators
- Electro-optics
- Re-timers
- Custom Compute
- DPUs / SmartNICs
- SSD Controllers

Source: CXL Forum at FMS 2023 - CXL: Transforming Cloud Infrastructure

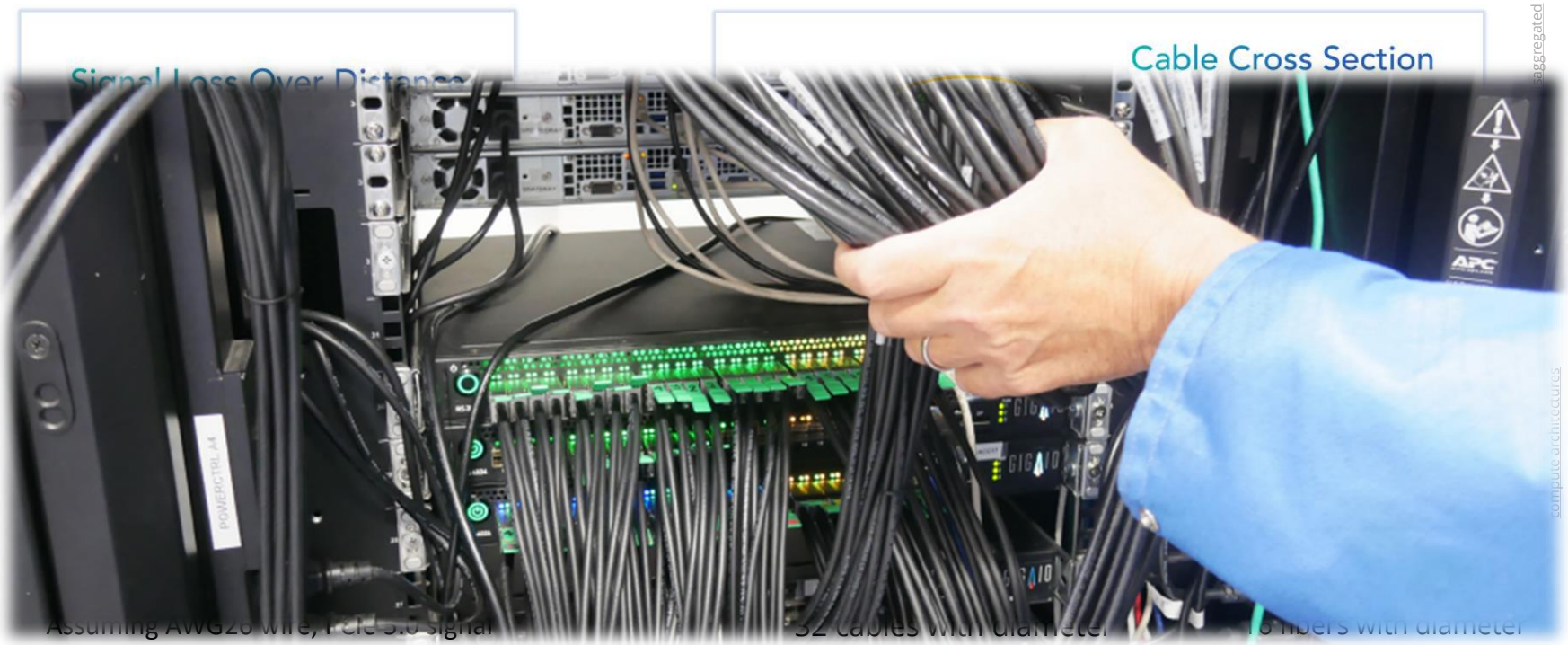
CXL vision: optimal resource utilization



Rack Fabric	Ethernet	Ethernet+CXL	CXL Only
Disaggregation	Storage and Accelerators	+Partial memory	+Full memory, compute

Source: CXL Forum at FMS 2023 - CXL: Transforming Cloud Infrastructure

Optical CXL is Required for Scaling

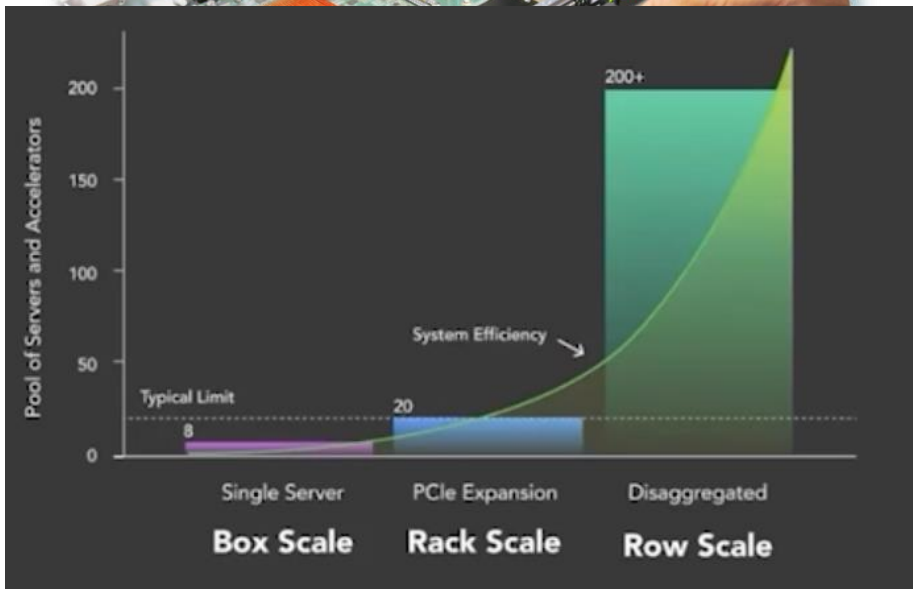


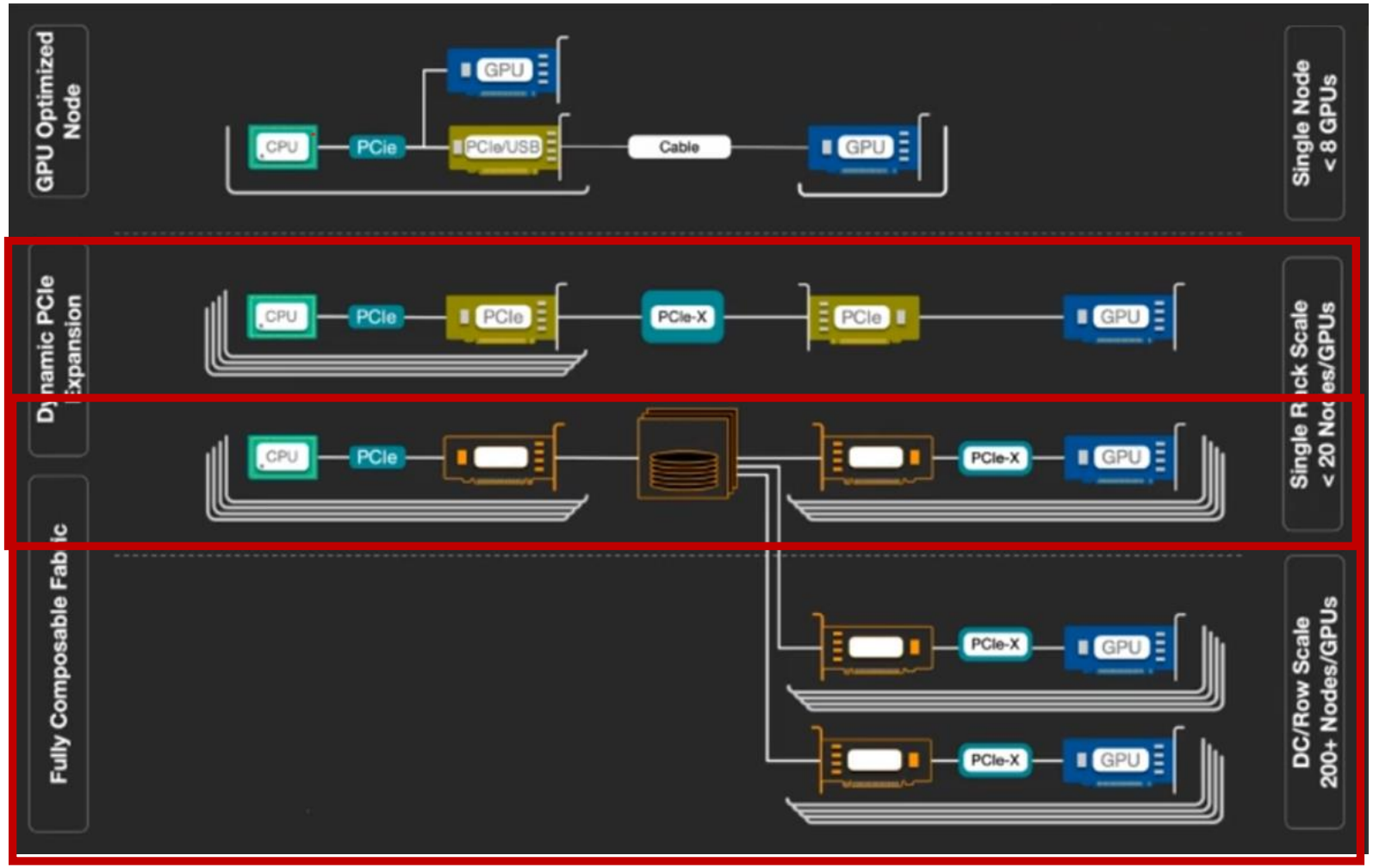
Copper cables struggles to support CXL scaling beyond a few servers

Composability: I/O Fabric (one of) the biggest challenge

Traditional System Model

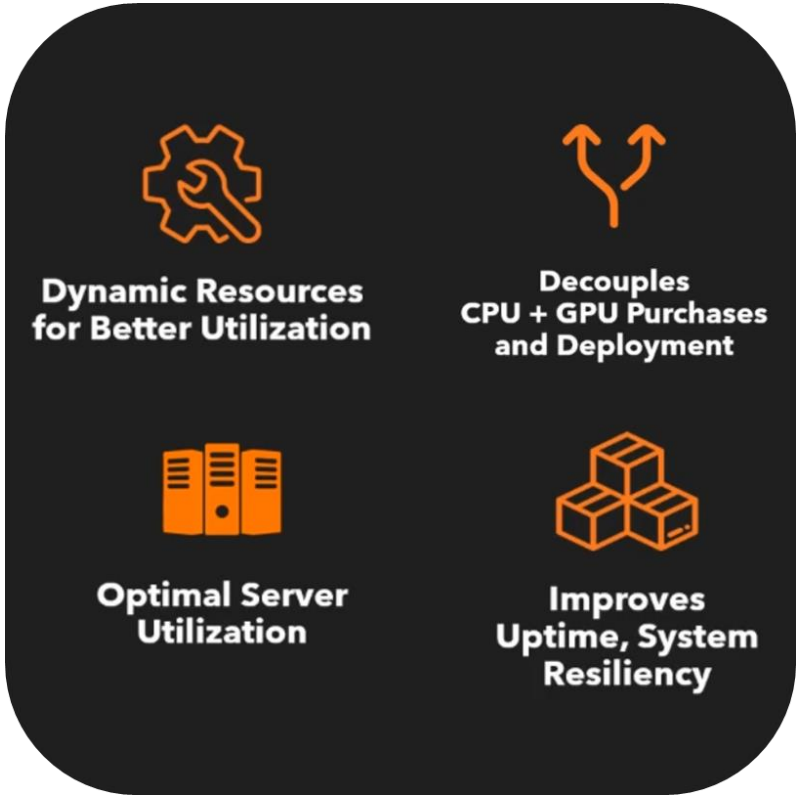
- Inefficient architecture
- Low ratio of GPUs to CPUs





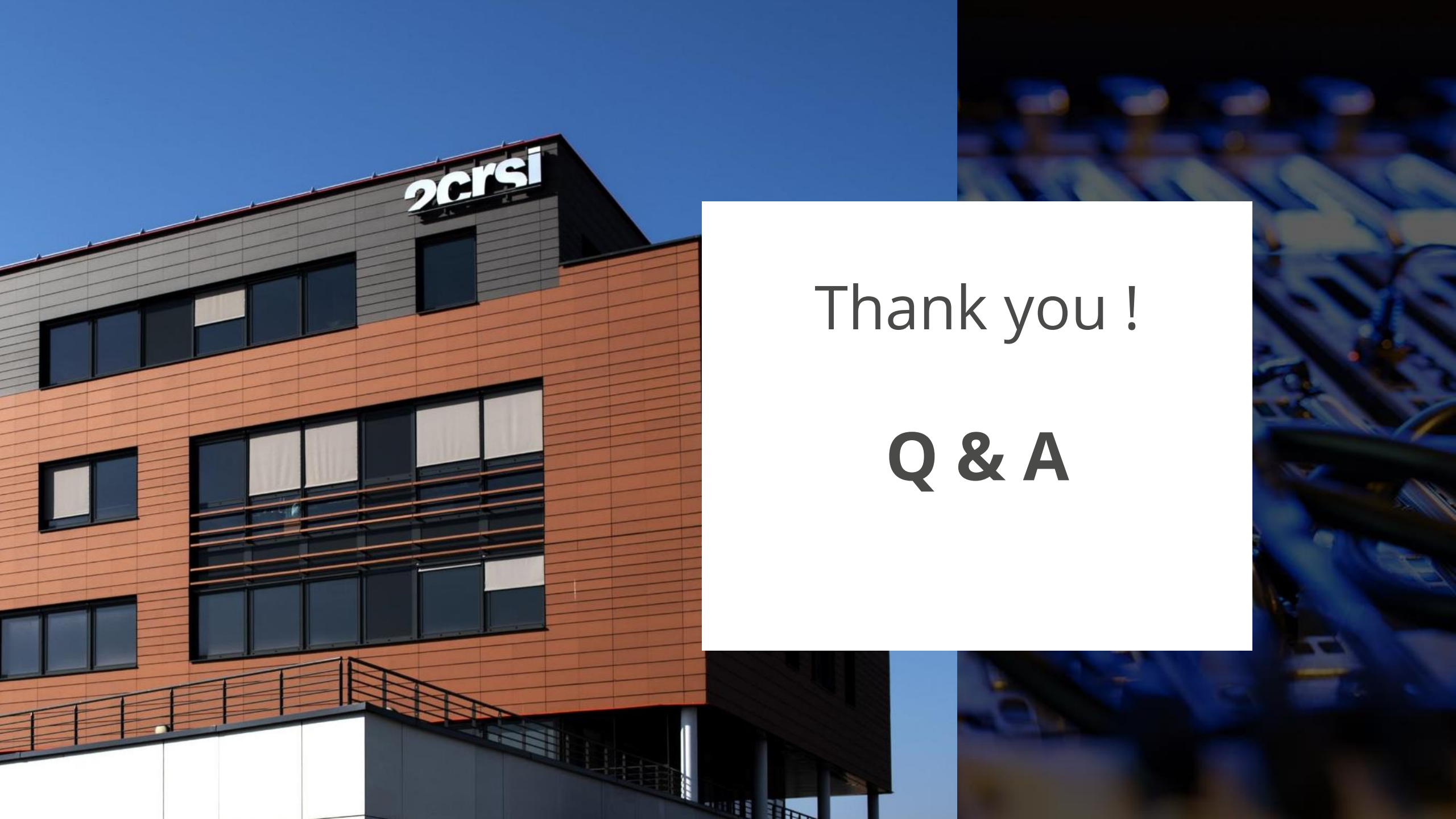
Composability: Benefits

- **Improved system utilization** by more fully leveraging expensive on-premises assets.
- **Flexible hardware profiles** – create the *Impossible server*
- **Pay as you Grow** - Simplified system expansion and reduced system costs via modular resource-specific nodes
- **Reduced Power & Cooling** (Sustainability)
- **Better Managed Life Cycles**



Composability: Challenges

- **Usage and operational impacts**
  - Which workloads are most suitable for composability?
- **Resource impacts**
  - Changes (if any) in application code to support composability?
  - Will it increase or reduce support requirements?
- **Performance impacts**
  - What about the latency to manage, provision, monitor, and re-claim system resources between jobs?
  - Will increased physical distance also add latency?
  - Scaling? How far?
- **Cost impacts**
  - Additional network (MPI, I/O, and now PCIe…)?



Thank you !

Q & A