Disaggregation: future of data centers?

FRNOG 38 Meeting – Oct. 6\textsuperscript{th}, 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!

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

Mona 1.12GG

High-end server based on AMD EPYC™ 4th Gen CPU with 12x 2.5” SFF Drives
First 4U server system supporting up to 32 High-End GPUs!

https://gigaio.com/supernode/
More computing power... with Accelerators

Age of Accelerators
- GPUs, FPGAs, Vector engines, DPUs
- Accelerators needed to get to value
- Different accelerators process data differently

Systems Using Accelerators on the TOP500

Domain-specific computation could drive performance efficiency
One-size-fits-all architectures are outdated

e.g., different research activities

- 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 (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”
PCle-based Fabric

Example of a Rack scale system

PCle Switch
PCIe-based Fabric

Example of a Rack scale system

PCIe-based Fabric
Composable Resources

Diagram showing a PCIe fabric with connections to Node-to-Device, Node-to-Node, and Device-to-Device components, as well as shared memory.
Summary of Data Center Memory Challenges

**BANDWIDTH**

Decreasing memory bandwidth per core

**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
Scope of CXL

CXL Ecosystem

Compute Express Link™ and CXL™ are trademarks of the Compute Express Link Consortium.
Key Design aspects of CXL Memory

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

With CXL

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
Data Center: Expanding Scope of CXL

CXL 3.0
Composable Fabric growth for disaggregation/pooling/accelerator

CXL 2.0
Memory/Accelerator Pooling with Single Logical Devices
Multiple Nodes inside a Rack/Chassis supporting pooling of resources

Memory Pooling with Multiple Logical Devices

Scope of CXL
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

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

Source: CXL Forum at FMS 2023

Assuming AWG26 on PCIe 5.0 32 cables with diameter > 6mm (CAT8)
16 fibers with diameter of 0.125mm
Composability: I/O Fabric (one of) the biggest challenge

**Traditional System Model**
- Inefficient architecture
- Low ratio of GPUs to CPUs

**Fully Composable Fabric**
- Data center scale
- Highly resilient
- No single point of failure
- Very high radix
- Fully dynamic path selection
- Accelerates CLX memory scalability
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