OCP FUTURE TECHNOLOGIES SYMPOSIUM

OCP Global Summit

November 8, 2021 | San Jose, CA

Hierarchical/Tiered Memory for Hyperscale Use-Cases

Software-defined, Flexible Memory Footprint for datacenter deployment

Siamak Tavallaei
Chief Systems Architect, Google
On the Board of Directors, CXL Consortium
Server Project Lead at OCP Incubation Committee

Points expressed here are not reflective of Google plans or views at-large

Abstract

As transistor density grows, the growth in compute capability of SoCs necessitates the growth in memory capacity and bandwidth to maintain a **balanced** system. The physical constraints of delivering increased memory capacity and BW has created a trend for a **hierarchy of memory** characterized by the notion of **Local** Memory, **Extended** Memory, and **Expanded** Memory.

This section explores driving industry-standard efforts and technologies such as **C**ompute e**X**press **L**ink (**CXL**) into Open Compute Project (**OCP**) as a means to realize products in such a way to enable and encourage high-volume adoption into **Hyperscale**, **Enterprise**, and **Edge datacenters**.

An in-chassis pool of CXL-attached memory coupled to a set of multi-ported CXL memory controllers provide locally disaggregated memory pooling capability which enables the notion of **software-defined, flexible memory footprint**.

Via a memory pool, a set of Host Nodes may connect to a set of CXL Pooled Memory Controllers (**CPMC**) to extend the memory footprint of each Host beyond their locally attached memory.

In addition to increasing memory capacity and bandwidth per Host, this flexibility will eliminate much of the memory bubble normally created by overprovisioning memory for each Host to account for fluctuation in general-purpose application needs in datacenters.

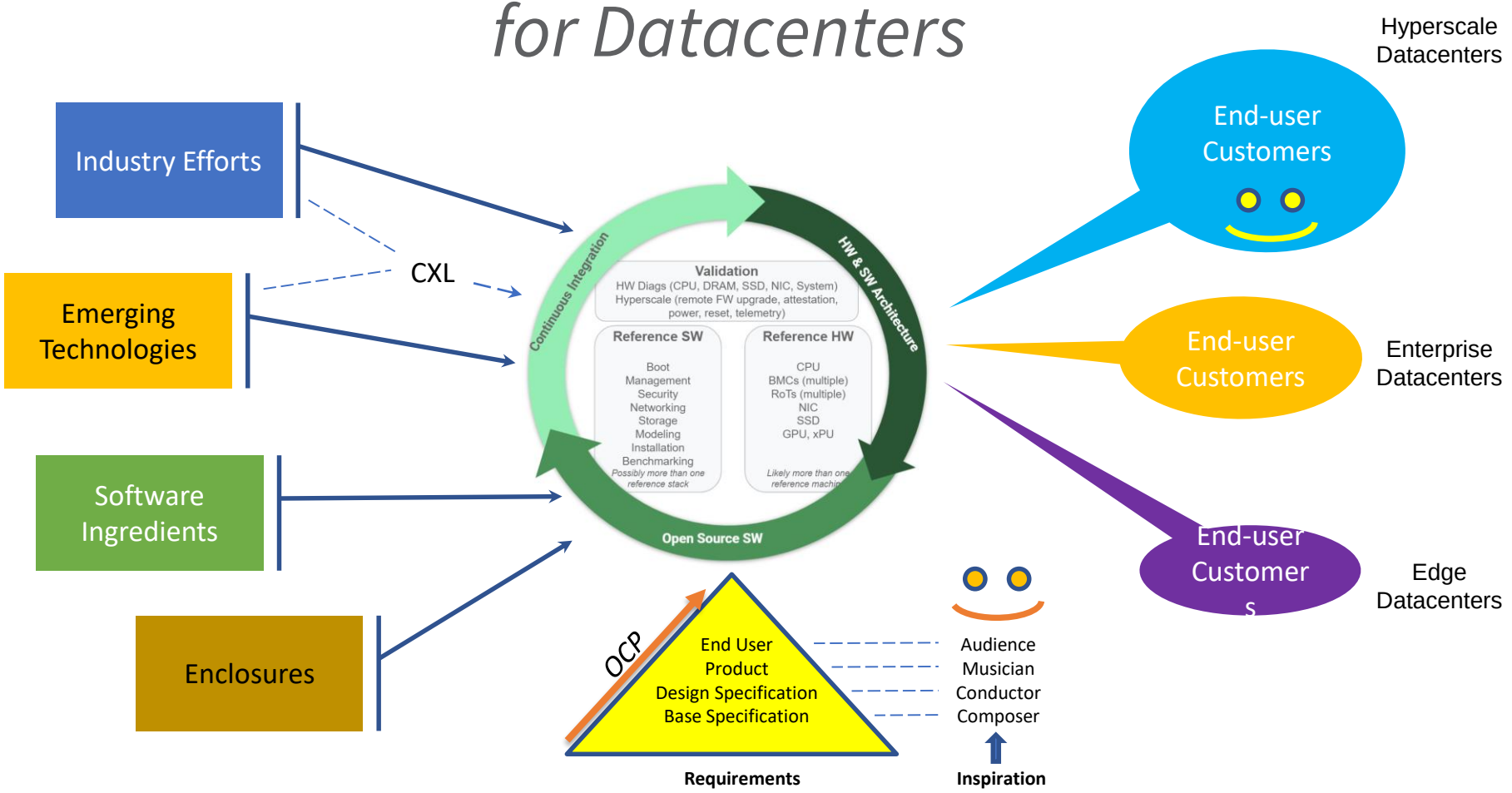
Removing this memory bubble reduces material cost; while adding to the reserved memory pool provides value to applications. These two factors increase **Performance/TCO**.

Outline

Later presentations will discuss the characteristics of memory tiers and their impact on performance, workloads, and operations

- OCP is a great place to integrate various technologies and solutions
- CXL as a technology example
  - CXL enables technical benefits
  - CXL-based offerings and use-cases of interest
- Fundamental requirements persist
- Solutions
- Enablers

A Flywheel of Opportunities for Datacenters



CXL-enabled Opportunities

Interconnect

- Based on PCIe physical layer, **high-speed**
- Optimization for **Coherent**, Load/Store Semantics with **low-latency** for short packets
- Fan-out using a Switch (**large** systems)

Memory

- Memory Capacity and Bandwidth **Expansion**
- Memory **Pooling**
- **Emerging** Memory Technologies

Storage-class Memory

- Architected optimizations for **persistent** memory using Load/Store semantics
- Pooling (**sub-dividing** a Large Device)

Accelerators

- Computational **Off-loading**
- CPU and Accelerator **working on the same coherent memory region**
- **Avoiding** superfluous **data movement** and reducing the associated time and energy (computation applications: in-memory, in-storage, and in-peer-acceleration)

CXL Ecosystem

Major companies have announced product plans around ***CXL***

- SoC suppliers
- Memory controller suppliers
- Storage suppliers
- Network controller suppliers
- Accelerator suppliers
- Switch suppliers

CXL Ecosystem

Major companies have announced product plans around ***CXL***

- SoC suppliers
- Memory controller suppliers
- Storage suppliers
- Network controller suppliers
- Accelerator suppliers
- Switch suppliers

*This **broad** engagement is the major **advantage** we expect of CXL to deliver a **successful** and profitable environment for all **participants***

Customer Requirements

- ***Enterprise*** customers have **diverse** set of needs

- As the *Enterprise* customers move their requirements to the *Cloud Datacenters*, they enjoy the benefits **at-scale** and bring their diverse needs to ***The Cloud***

- ***Edge*** solutions benefit from this Enterprise & Cloud **interplay**

Fundamental Requirements Persist

We still need to deliver integrated **hardware** and **software** solutions which are

Useful

- Desirable

High Quality

- Secure (RoT and Chain of Trust: at-rest, in-transit, and secure execution)
- Safe
- Reliable
- Available

Manageable

- Serviceable
- Diagnosable

Performant

Efficient (power, space, cost, time, complexity, …)

Fundamental Requirements Persist

We still need to deliver integrated **hardware** and **software** solutions which are

Useful
- Desirable

High Quality
- Secure (RoT and Chain of Trust: at-rest, in-transit, and secure execution)
- Safe
- Reliable
- Available

Manageable
- Serviceable
- Diagnosable

Performant

Efficient (power, space, cost, time, complexity, …)

Especially when driving the solutions into large Datacenters

Solution

Balanced **Core** Architecture

- Frameworks, Software, Compute, BW, Capacity, Latency

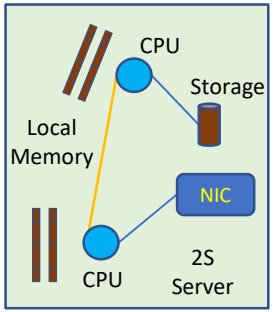
General-purpose

- Modular **Building Blocks**

Extensible

- Allow heterogeneous variants based on the core Building Blocks

Challenges in deploying traditional Servers into Hyperscaled Datacenters



Balanced match of:

- CPU core count
- Memory Capacity, Bandwidth, and Latency
- Storage Capacity, Bandwidth, IOPS, and Tail Latency
- Network Bandwidth

Challenging to meet the above balance in presence of varied workloads and customer VMs in high-volume production

Result:

- **Overprovisioned** resources to meet customer demands
- Unused or underutilized resources
- Increased Cost

Very many specialized Server SKUs in the fleet!

How can a new technology

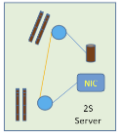
such as CXL

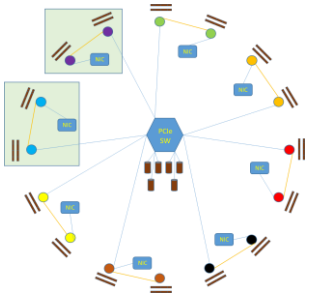
help?

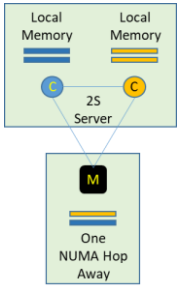
Extensible Solutions

Topology

- Point-to-point
- Multi-port
- Switched







Density (multi-port)

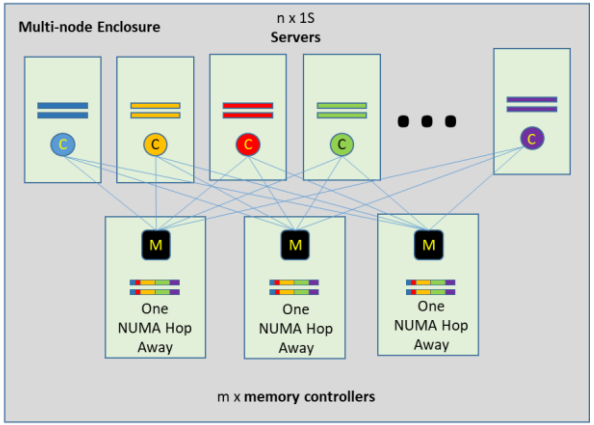
- Dense packaging of (n x m) multi-ported Devices
- Liquid cooling

Reach (SERDES)

- Longer Links *(to all devices including memory!)*
- Modular Enclosures
- Cabled Solutions
- Photonics

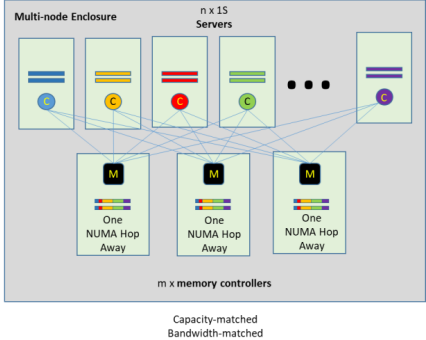
Extensibility (heterogeneous)

- Compute (xPU), Memory, Storage, Networking



Capacity-matched
Bandwidth-matched

Compute Disaggregation Taxonomy



- **Pooling** (dividing a resource to multiple non-overlapping logical units and assigning them to different servers/hosts)
- **Sharing**
  - Serialized Sharing (a device may be fully mapped to a server at one time and to a different server at a different time)
  - Concurrent Sharing (multiple servers/hosts are assigned to the same portion of a device at the same time; coherence and access ordering may be enforced by hardware or software)
- **Borrowing** (as part of its own separate coherence domain, a server may get permission to access a portion of a second server's resource. This resource will leave the second server's coherence domain.)
- **Fabric** (a mechanism for dynamically interconnecting heterogeneous elements to form computing systems)
- **Physical Disaggregation** (interconnected chassis: Server Head-node + Expansion Chassis such as a JBOD, JBOF, JBOG, …)
- **Logical Disaggregation** (composing several servers via a ***Fabric*** to provide access to shared or pooled sets of resources)
- **Local Disaggregation**
  - Multiple servers **in one Chassis** accessing a shared or pooled set of resources
  - Using a multi-host capable *Switch* or via multi-ported *End Devices*
- **Extend** memory via increased memory capacity and range of the same type of medium (e.g., **DRAM** with <3x mem latency)
- **Expand** memory via managing **multiple tiers** of memory (e.g., NAND Flash backing DRAM) & swapping/paging techniques
- Coherence (enforced by HW or maintained by SW via access ordering sequences and appropriate flush mechanisms)
- Scale-up (single host, homogeneous computing, scale via the same type of interconnect protocol)
- Scale-out (networked-based or via changing interconnect protocol, heterogeneous or distributed multi-server computing)

Is CXL the End-all?

Should we move everything to the new technology?

Is CXL the End-all?

*Should we move everything to the **new technology**?*

Putting Things where they Belong!

PCIe

- Software-managed consistency (DMA, RDMA)
- Block Data Moves (Large Payloads)
- Deferred Calls, Interrupts
- Latency-tolerant
- Sequential data access

CXL

- Hardware-managed (Coherence)
- Load/Store (Short Packets)
- In-line codes
- Latency-sensitive
- Concurrent data access

Putting Things where they Belong!

PCIe

- Software-managed consistency (DMA, RDMA)
- Block Data Moves (Large Payloads)
- Deferred Calls, Interrupts
- Latency-tolerant
- Sequential data access

CXL

- Hardware-managed (Coherence)
- Load/Store (Short Packets)
- In-line codes
- Latency-sensitive
- Concurrent data access

- Enabling new optimizations and programing paradigms

There will be a transitional period

from one
to the other

Remember!

All Fundamental Requirements Persist!

Enablers (Software and Firmware Ingredients)

CXL Fabric Manager

- Secure composability, allocation, on-lining/off-lining

Pre-boot Environment

- Discovery, Enumeration

CXL Bus Driver

- Configuration, Resource allocation

CXL Memory Device Driver

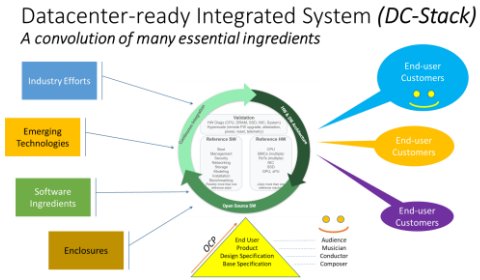
- Interactions with Bus Driver, Fabric Manager, and VMM
- RAS, Security, Fault-isolation, On-lining, Off-lining, …

  ECN: Error Isolation on CXL.mem and CXL.cache (Enabled by the Root Port; requires Software Stack to recover from faults)

OS-specific Software

- VMM, Hypervisor
- VM Allocation, Orchestration, Fault-isolation & Recovery

Summary



- A new technology such as CXL enables special **benefits**, but we still need to deliver the **fundamental requirements**
- Delivering these technical advantages will take major **ecosystem effort** from various industry players
- **You** along with **balanced**, extensible, **modular** solutions, with the staged software stack are the enablers
- Taking advantage of the industry-wide efforts, we can deliver CXL-based **PoCs** toward a integrated system via open-sourced hardware and software

***You** are a*

***Giant** Enabler!*

*Lend your **Coattail**!*

Available presentations on ***C****ompute e****X****press* ***L****ink (CXL)* as an open-standard specification

The material that CXL has published on memory pooling:
Webinar: Compute Express Link™ 2.0 Specification: Memory Pooling
LinkedIn Post on CXL Memory Pooling
Recap Q&A Blog: Part 1
Recap Q&A Blog: Part 2
CXL Memory Pooling Slides
CXL Memory Pooling Video

CXL 2.0 Animated Video
CXL to Gen-Z Use Cases

Presentations on Datacenter-ready Integrated System **(DC-Stack):**

DC-Stack_Datacenter-ready Integrated System_OCP

http://files.opencompute.org/oc/public.php?service=files&t=dd1e012f85ab59a608d758db8357539c


OCP
TOTAL SOLUTIONS

OCP FUTURE TECHNOLOGIES SYMPOSIUM

2021 OCP Global Summit | November 8, 2021, San Jose, CA