Productizing new technologies for datacenter deployment

Prepared for a Talk at IEEE/ACM International Symposium on Microarchitecture
Cloud@MICRO (10/18/2021)

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

Points expressed here are <u>not</u> reflective of Google plans or views at-large

Abstract

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 **datacenters**.

Outline

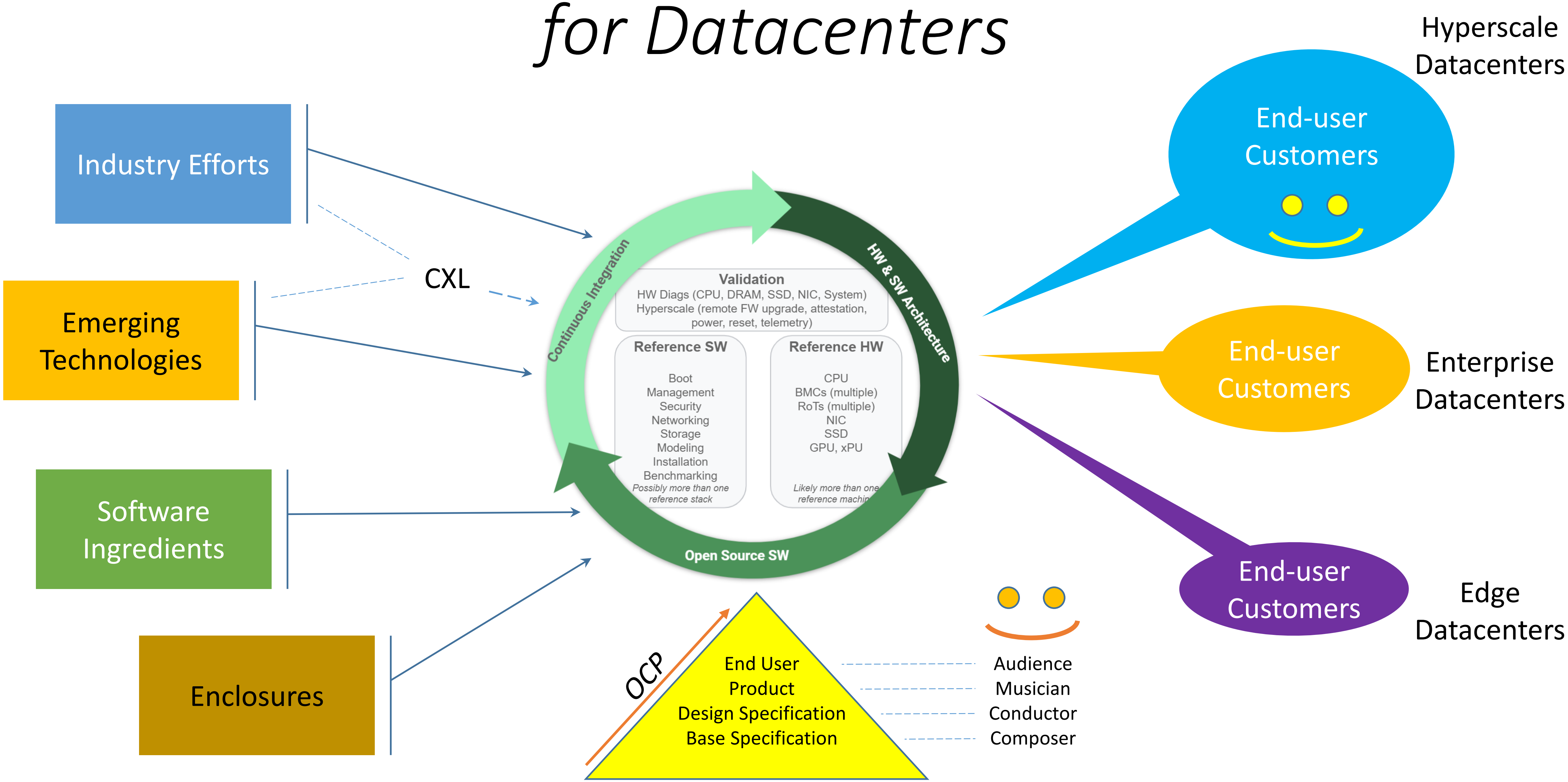
- Earlier presentations discussed workload performance and a few new technologies
- Customer (workloads & operations)

- CXL as a technology example
  - CXL enables technical benefits
  - CXL-based offerings and use-cases of interest

- Fundamental requirements persist
- Solutions

- Enablers
- Industry-wide efforts: CXL, PCIe, SNIA, NVMe, JEDEC, DMTF, and OCP

A Flywheel of Opportunities for Datacenters



Industry Efforts

Emerging Technologies

Software Ingredients

Enclosures

CXL

Continuous Integration

HW & SW Architecture

Open Source SW

Validation
HW Diags (CPU, DRAM, SSD, NIC, System)
Hyperscale (remote FW upgrade, attestation, power, reset, telemetry)

Reference SW
Boot
Management
Security
Networking
Storage
Modeling
Installation
Benchmarking
Possibly more than one reference stack

Reference HW
CPU
BMCs (multiple)
RoTs (multiple)
NIC
SSD
GPU, xPU
Likely more than one reference machin

OCP

End User — Audience
Product — Musician
Design Specification — Conductor
Base Specification — Composer

End-user Customers — Hyperscale Datacenters

End-user Customers — Enterprise Datacenters

End-user Customers — Edge 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-accelerator)

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-accelerator)

Lots of fun for the technologists!

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***

Customers

- ***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**

Do off-the-shelf solutions *(**OTS**)*

meet ***Datacenter*** Requirements?

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

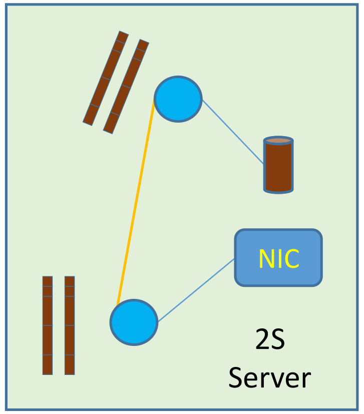
General-purpose

- 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 underused resources
Increased Cost

How can a new technology

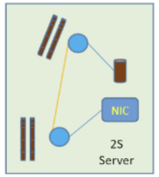
such as CXL

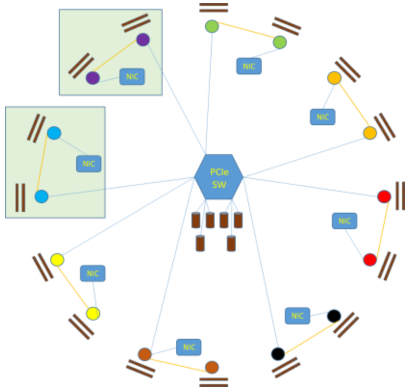
help?

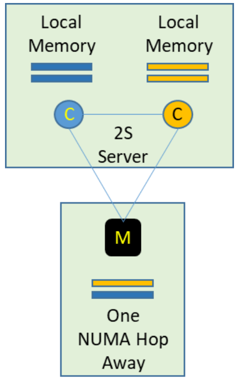
Extensible Solutions

Topology

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





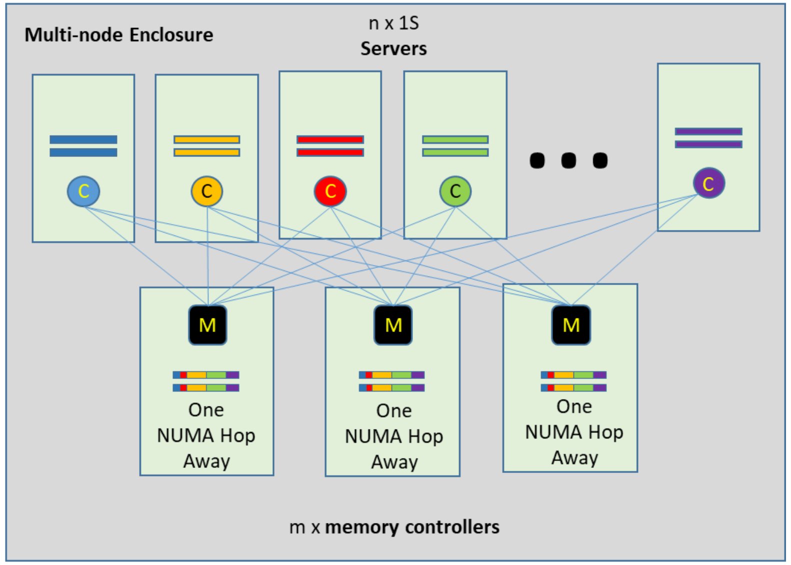


Density (multi-port)

- 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



Capacity-matched
Bandwidth-matched

Extensibility (heterogeneous)

- Compute (xPU), Memory, Storage, Networking

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 Needed

Enablers Needed

A Vibrant Ecosystem

Enablers

Riding on the Coattail of The Giants

Industry Efforts

- CXL Consortium
- PCIe SIG
- SNIA
- NVMe
- JEDEC
- DMTF
- OCP
- …

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

Datacenter-ready Integrated System *(DC-Stack)*

CXL Consortium relies on other standards bodies to provide specifications for suitable mechanical/thermal/electrical **Form Factors**

Newly enabled solutions based on CXL will require new form factors

Reference Designs (open-source hardware at OCP)

PoCs

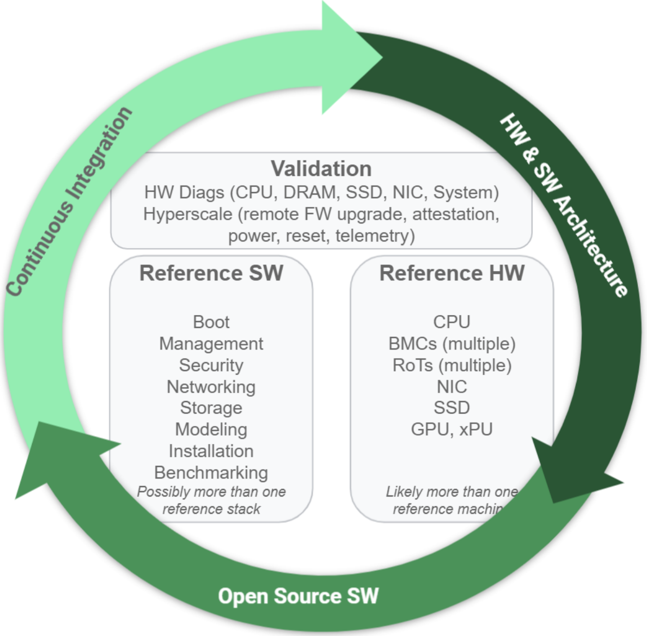
Datacenter-ready Integrated System ***(DC-Stack)***

CXL Consortium relies on other standards bodies to provide specifications for suitable mechanical/thermal/electrical **Form Factors**

Newly enabled solutions based on CXL will require new form factors

Reference Designs (open-source hardware at OCP)

PoCs



Validation

HW Diags (CPU, DRAM, SSD, NIC, System)
Hyperscale (remote FW upgrade, attestation, power, reset, telemetry)

Reference SW

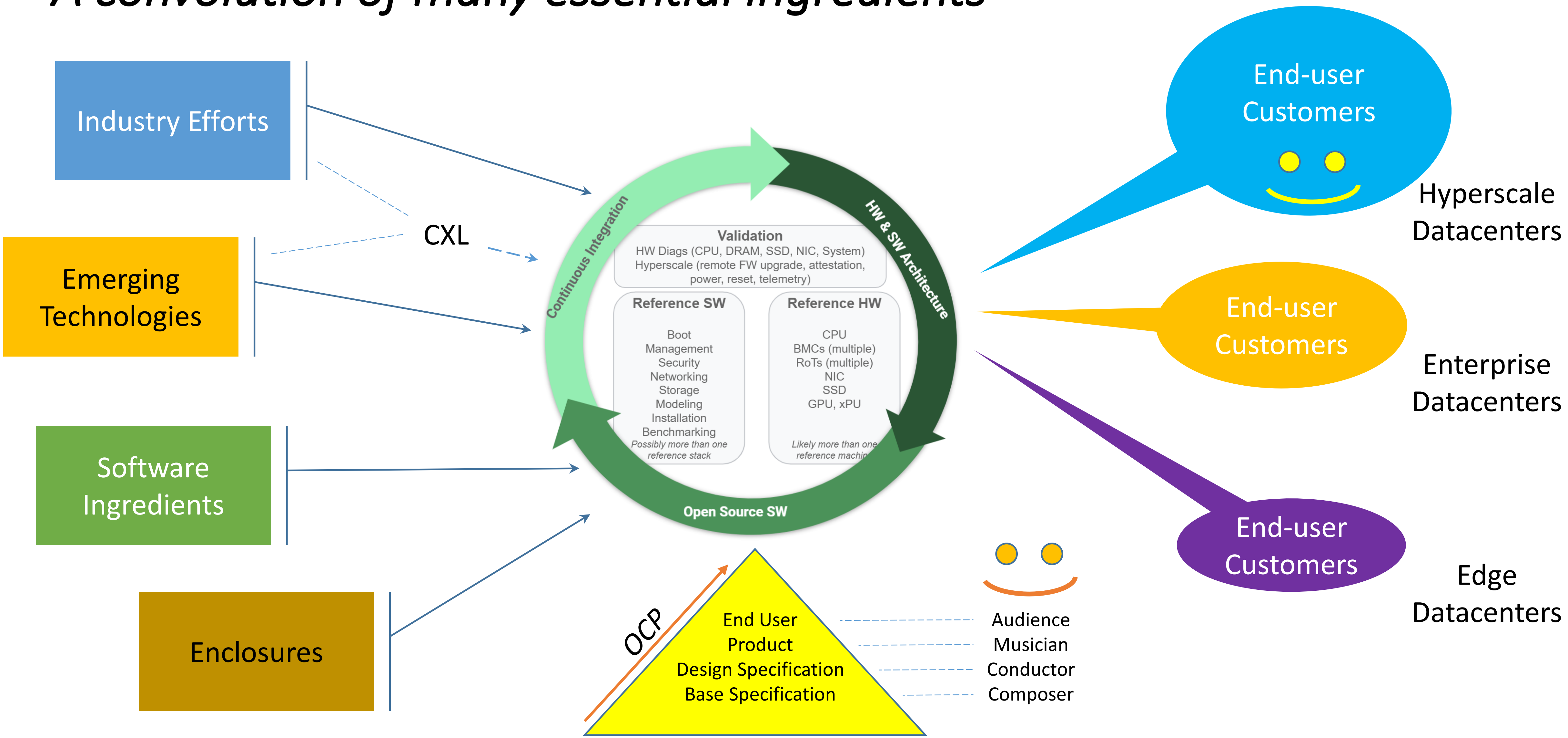
Boot
Management
Security
Networking
Storage
Modeling
Installation
Benchmarking
Possibly more than one reference stack

Reference HW

CPU
BMCs (multiple)
RoTs (multiple)
NIC
SSD
GPU, xPU
Likely more than one reference machin

Datacenter-ready Integrated System *(DC-Stack)*

A convolution of many essential ingredients

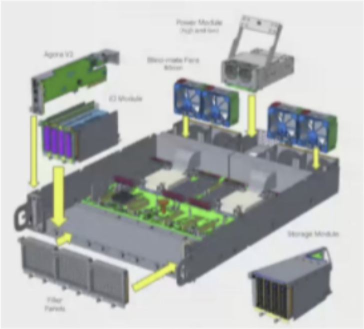


For OCP Discussions

Modular Hardware

Reference designs for three different architectural instances

1S/2S Server



Rack

Power

Cooling

Multi-Blade/Instance Chassis

- HPM (CPUs + Memory) per Blade

Multiple DC-SCMs or Multi-Host DC-SCM

DC-MIO

IO Module/Cage (IO Slots)

Multiple SmartNICs or Multi-Host SmartNIC

Multi-CPU Server

Rack

Power

Cooling

Single Instance Chassis

- Multi-HPM (CPUs + Memory)

IO Module/Cage (IO Slots)

Single-Host DC-SCMS

DC-MIO

Single-host SmartNIC

GPU/Acc. or Storage Expansion Chassis

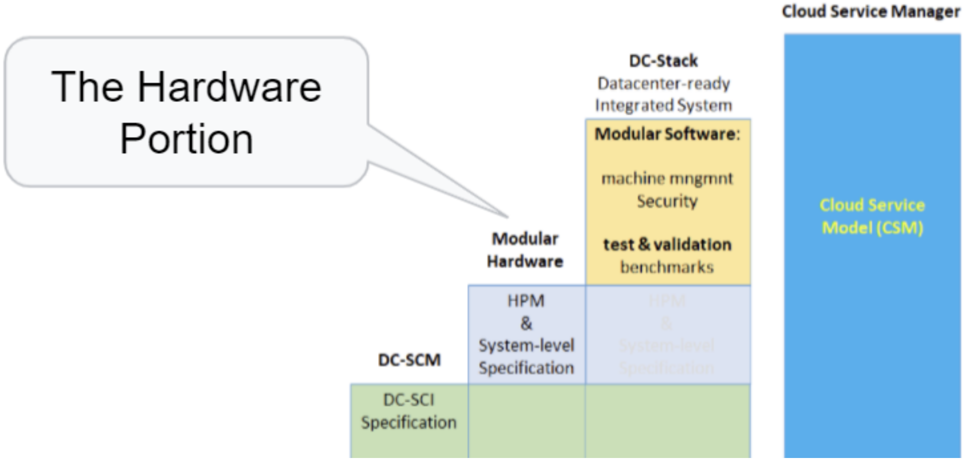
Expansion Chassis, Power, Cooling

Head Node (Server)

Interconnect (cables, retimer?)

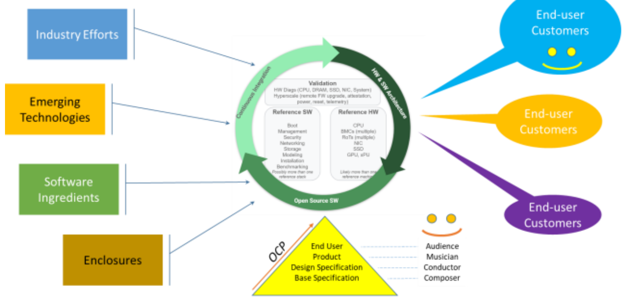
Out-of-band Management

BMC (or a variant of DC-SCM)



Datacenter-ready Integrated System *(DC-Stack)*

A convolution of many essential ingredients



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, along with the staged software stack are the **enablers**

- Taking advantage of the industry-wide efforts, we can deliver CXL-based **PoCs** toward a datacenter-ready integrated system *(**DC-Stack**)* via open-sourced hardware and software ***(OCP)***

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