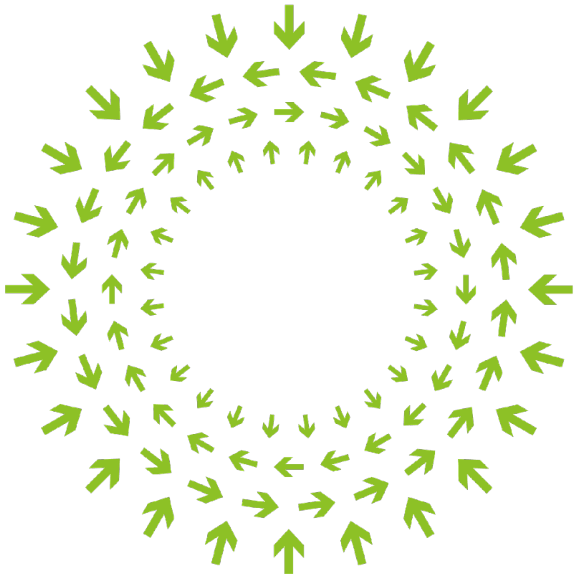


OPEN

Compute Project

Facebook Zion System Specification

1.0

Author: Hao Shen, Hardware Engineer, Facebook
Author: Michael Haken, Mechanical Engineer, Facebook
Author: Harsha Bojja, Product Quality Engineer, Facebook
Author: Tyler Hart, RTP Engineering, Facebook
Author: Garnett Thompson, Hardware Engineer, Facebook
Author: Cheng Chen, Thermal Engineer, Facebook

Table of Contents

1. License

PLEASE PICK EITHER THE OCP CLA OPTION OR THE OWF OPTION. ONLY ONE CAN BE USED. DELETE THE ONE NOT USED.

1.1. OPTION A: OCP CLA

Contributions to this Specification are made under the terms and conditions set forth in Open Compute Project Contribution License Agreement ("OCP CLA") ("Contribution License") by:

Facebook

You can review the signed copies of the applicable Contributor License(s) for this Specification on the OCP website at https://www.opencompute.org/legal-documents
Usage of this Specification is governed by the terms and conditions set forth in
Open Web Foundation Final Specification Agreement ("OWFa 1.0")
also known as a "Specification License".

Notes:

1) The following clarifications, which distinguish technology licensed in the Contribution License and/or Specification License from those technologies merely referenced (but not licensed), were accepted by the Incubation Committee of the OCP:

None

2) The above license does not apply to the Appendix or Appendices. The information in the Appendix or Appendices is for reference only and non-normative in nature.

NOTWITHSTANDING THE FOREGOING LICENSES, THIS SPECIFICATION IS PROVIDED BY OCP "AS IS" AND OCP EXPRESSLY DISCLAIMS ANY WARRANTIES (EXPRESS, IMPLIED, OR OTHERWISE), INCLUDING IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, FITNESS FOR A PARTICULAR PURPOSE, OR TITLE, RELATED TO THE SPECIFICATION. NOTICE IS HEREBY GIVEN, THAT OTHER RIGHTS NOT GRANTED AS SET FORTH ABOVE, INCLUDING WITHOUT LIMITATION, RIGHTS OF THIRD PARTIES WHO DID NOT EXECUTE THE ABOVE LICENSES, MAY BE IMPLICATED BY THE IMPLEMENTATION OF OR COMPLIANCE WITH THIS SPECIFICATION. OCP IS NOT RESPONSIBLE FOR IDENTIFYING RIGHTS FOR WHICH A LICENSE MAY BE REQUIRED IN ORDER TO IMPLEMENT THIS SPECIFICATION. THE ENTIRE RISK AS TO IMPLEMENTING OR OTHERWISE USING THE SPECIFICATION IS ASSUMED BY YOU. IN NO EVENT WILL OCP BE LIABLE TO YOU FOR ANY MONETARY DAMAGES WITH RESPECT TO ANY CLAIMS RELATED TO, OR ARISING OUT OF YOUR USE OF THIS SPECIFICATION, INCLUDING BUT NOT LIMITED TO ANY LIABILITY FOR LOST PROFITS

OR ANY CONSEQUENTIAL, INCIDENTAL, INDIRECT, SPECIAL OR PUNITIVE DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND EVEN IF OCP HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

1.2 Acknowledgements

The Contributors of this Specification would like to acknowledge the following companies for their feedback:

Intel
Quanta

2. OCP Tenets Compliance

2.1. Openness

Zion Design follows the OAI concept. Zion system implements OAM, OpenBMC, OCP NIC technology to support AI workload. Facebook contributes the whole Zion design specification in OCP.

2.2. Efficiency

Zion system is designed with a modular concept where we can support different configurations for different use cases to optimize TCO.

2.3. Impact

Zion system is designed to support large scale AI workloads with OAMs. It has been implemented in a large-scale DC environment. The improvement of production efficiency is significant for AI use cases.

2.4. Scale

Zion System is designed to optimize TCO for large scale implementations. It implements OpenBMC and OCP NIC which benefit large scale monitoring and telemetry requirements in hyper scale datacenter.

3. Revision Table

Date	Revision #	Author	Description
07-28-2021	1.0	Facebook	OCP Contribution

4. Scope

This document describes technical specifications of Facebook's Zion System. Zion System is designed to support AI workload, which implements a modular design concept that contains a server box (Angels Landing), interconnection box (Clear Creek) and Accelerator box (Emerald Pools). Three boxes can be configured in different ways to support different use cases.

4.1 Terminology

Term	Description
2S-MB	2 Socket Motherboard with Intel® Cooper Lake CPU
Angels Landing 2S	2 Socket Server
Angels Landing 4S	4 socket Server
BIC	Bridge IC
BMC	Baseboard management Controller
Clear Creek	Interconnect Box
CPX-6S	Intel® Cooper Lake, 14nm processor
Emerald Pools	Accelerator Box
Zion	Facebook's AI Platform
Zion2S	Zion system with 2 socket host
Zion4S	Zion system with 4 socket host
ZionEx	Zion system with scale up capability

5. Zion Overview

Zion system includes three modules: Angels Landing, Clear Creek and Emerald Pools. Angels Landing is the CPU box; Clear Creek is the interconnect box; Emerald Pools is the accelerator box.

Different AI problems have very different requirements on hardware systems. AI algorithms involve very fast, which creates big challenges to AI platform design. We find out that certain flexibilities in system configuration are very useful to optimize the TCO for different use cases.

These three boxes are connected as the Figure below shows:

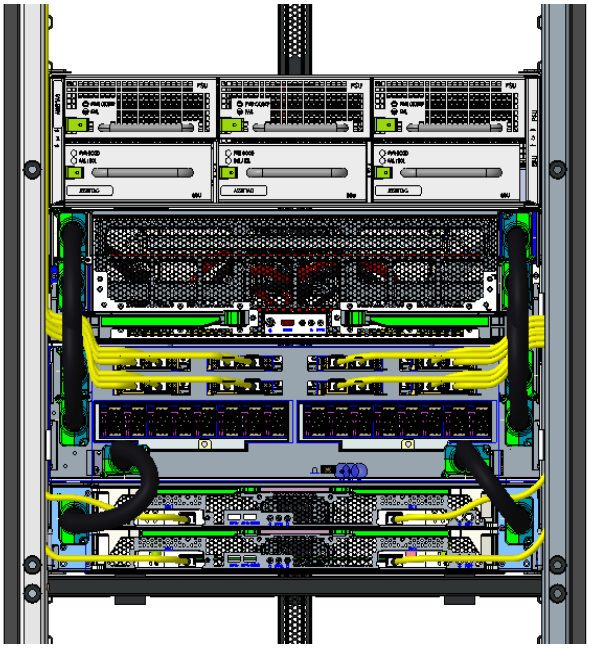


Figure 1: Zion System View

In this document we will introduce three different configurations that Zion system can support. We also contributed Angels Landing, Clear Creek and Emerald Pools specifications to OCP. Vendors can find the details within each box there. The spec document will focus on the system Interconnect.

6. Zion Subsystem Components Overview

I will start with a simple introduction to each subsystem in this chapter.

6.1 Angels Landing System

Angels Landing is the CPU box in the Zion system. It is designed as a 4-socket system based on Intel® CPX-6S processor architecture. The system contains 4 * CPU which are interconnected through Intel® UPI bus in a Fully Connected configuration. This system supports up to 48x DIMMs @3200MT/S and 4x NICs.

Figure. 2 below illustrates the Angels Landing box with the top cover removed. There are 2 trays in the system's front side. Each tray occupies 1 OU space and carries one 2S-MB - dual-socket motherboard. Each tray can be serviced individually.

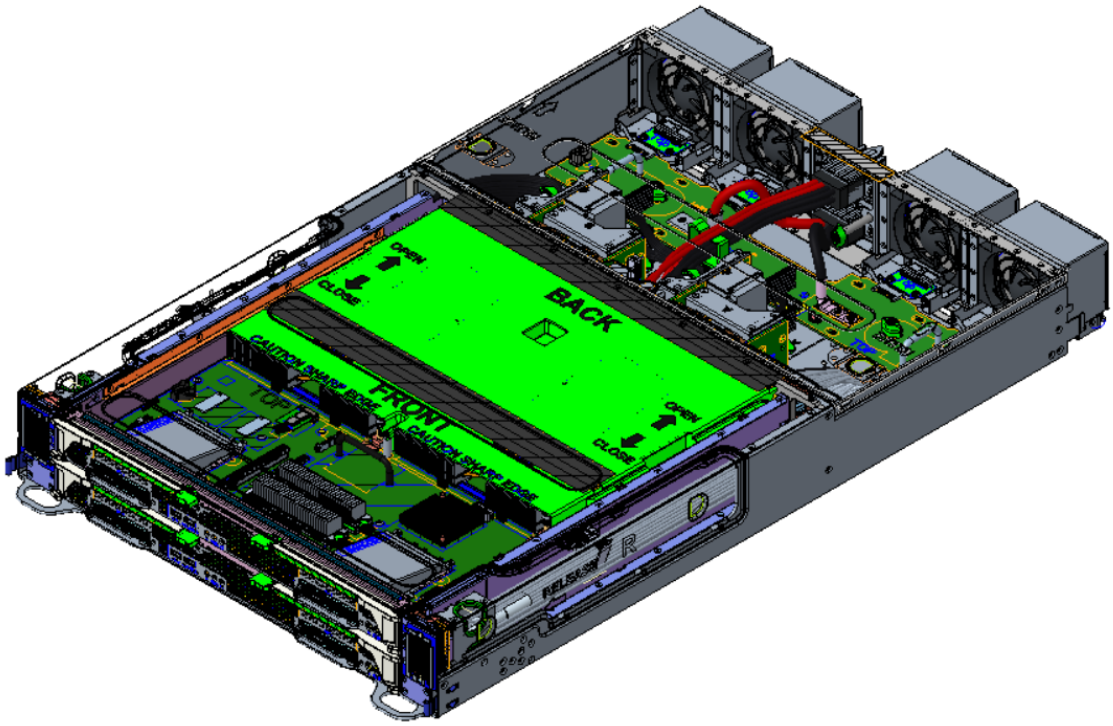


Figure 2: Angels Landing System View, Top cover removed

This box has three different configurations: 1x 4S, 2x 2S and 1x 2S (removing one MB). The detailed configuration method has been introduced in Angels Landing specification.

6.2 Clear Creek System

Clear Creek System is the interconnect box. It contains 4x gen4 PCIe switch, 8x NIC cards and 16x SSDs to handle the complicated workload where it requires hierarchical memory and accelerator scale-out. It can be inserted between the CPU box and Accelerator box.

Figure 3 shows the Clear Creek box. This box supports OCP3.0 NIC form factor. The NICs are connected to the motherboard through the cable to achieve better SI performance under PCIe gen4 speed. For storage, Clear Creek supports both M.2 and E1.S form factor with a removable SSD tray piece. In E1.S case, hot removal and hot add are supported.

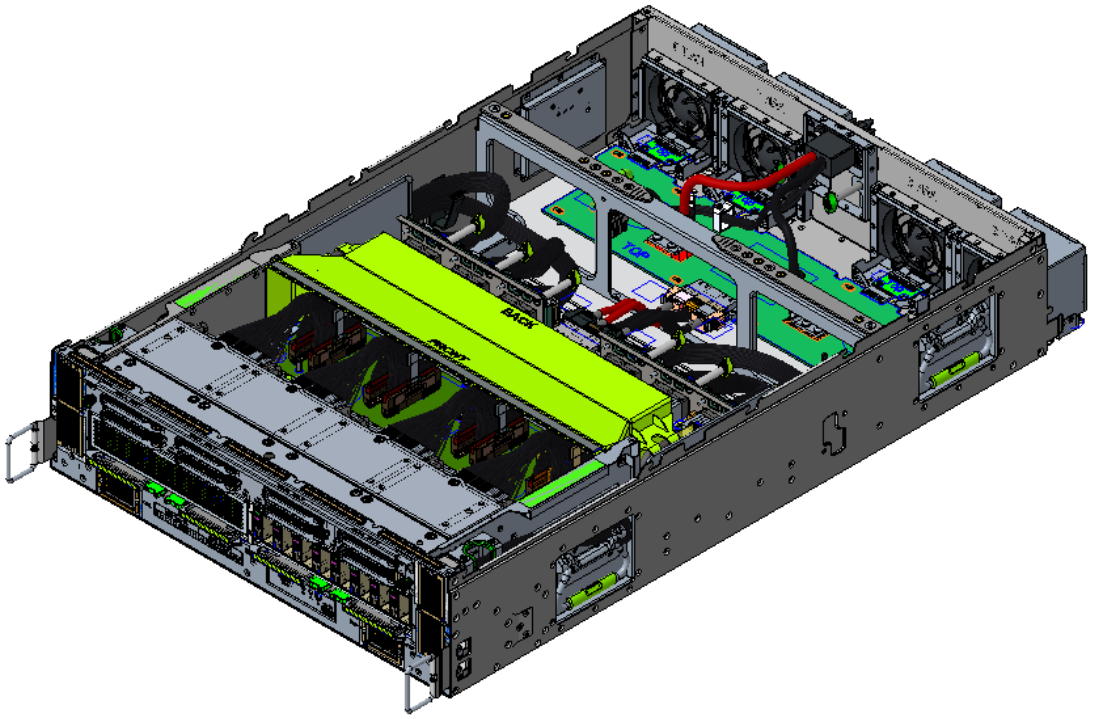


Figure 3: Clear Creek System View, Top cover removed

You can refer to Clear Creek System Specifications for the details of this box.

6.3 Emerald Pools System

Emerald Pools System contains four PCIe switches and eight Open Accelerator Modules (OAM) that support AI workload. We can support up to 400W for each OAM module with air cooling. Those OAMs are also interconnected together with high speed PCIe traces that support up to 50Gb/s speed.

There are 4x PCIe gen4 switches on the board that can config the system with different configuration topology. There is a dedicated BMC chip sitting on Emerald Pools motherboard that can monitor the status of all critical components.

Emerald Pools system contains one motherboard and one power distributed board. The PDB takes 12V input and transforms to 48V to support OAM power.

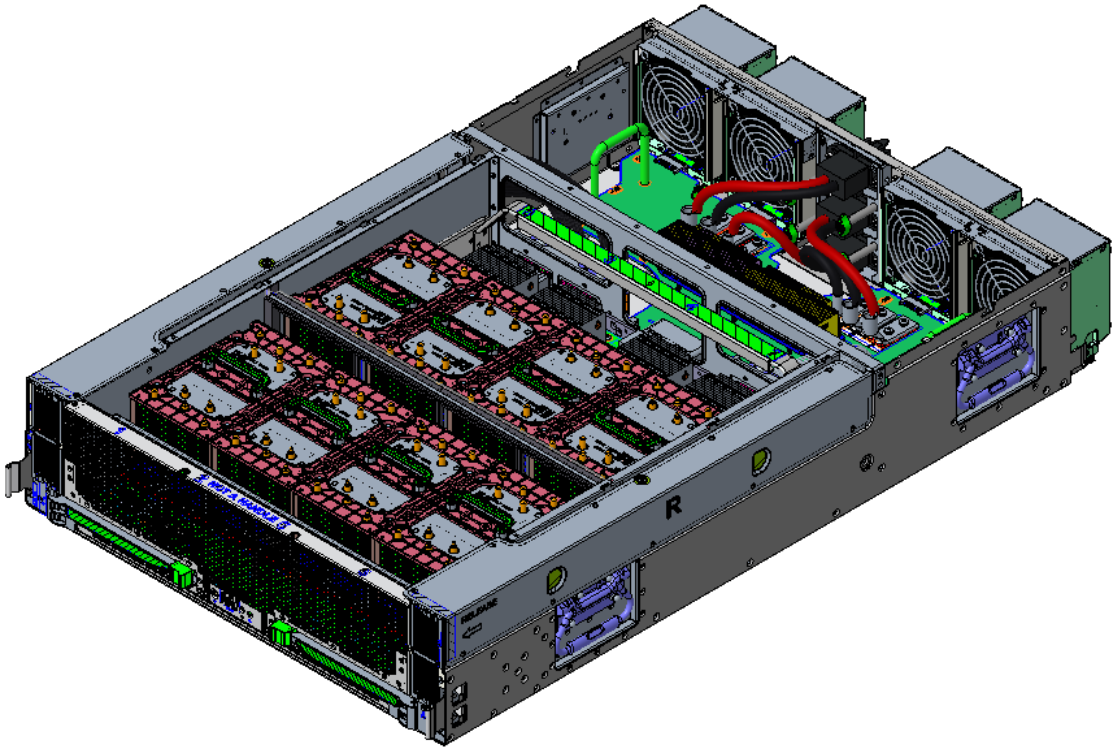


Figure 4: Emerald Pools System View, Top cover removed

6.4 Zion Modular Design Concept

AI problems are complicated. Some of the workloads are compute-intensive and others are IO intensive. Some problems need both. Certain models are small enough that you can batch thousands of them into one system. While other models need a scale-out option to build an accelerator cluster. It is difficult to design one system that fits all the requirements but meanwhile has good power efficiency.

In three boxes in the Zion system, we could configure the system in several different ways to meet the unique requirement of different models. In the next few chapters we will introduce three typical ways that you can configure the Zion system with these three boxes.

7. ZionEx system

ZionEx (Ex means extension) is one of the configurations that allow multiple systems to scale out and build a cluster type of the accelerator pod.

7.1 ZionEx block diagram

Figure 5 describes the block diagram of ZionEx system.

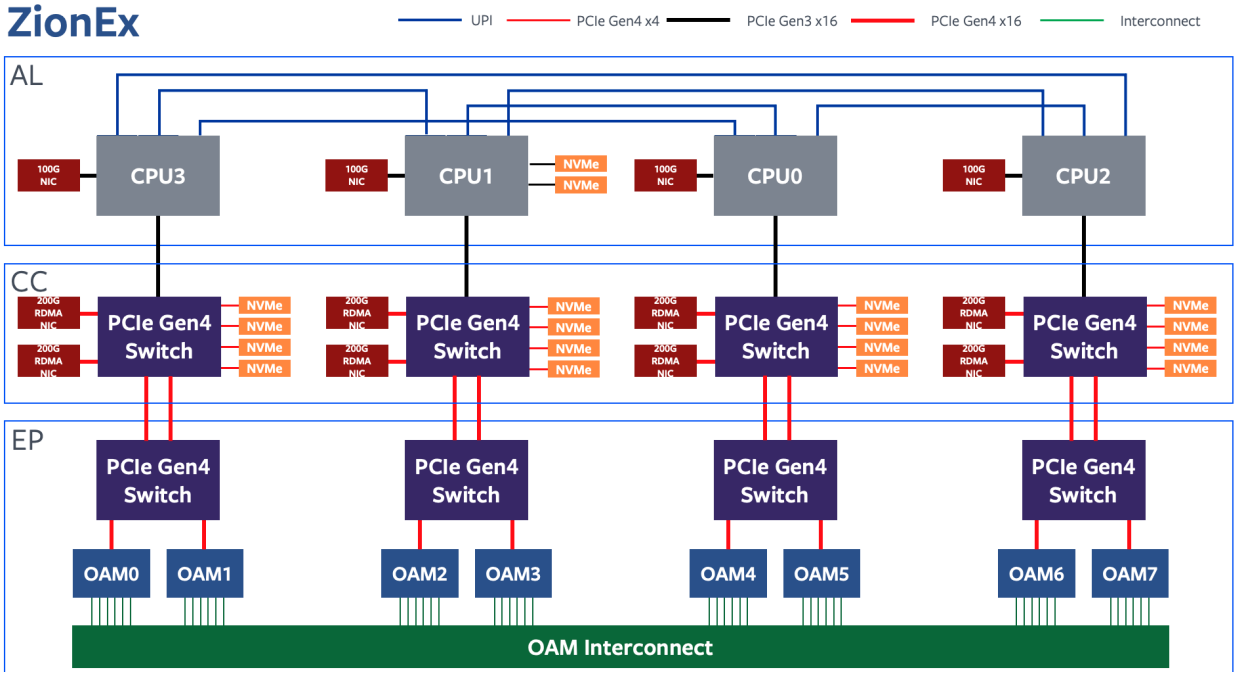


Figure 5: ZionEx Block Diagram

ZionEx system contains a 4-Socket Angels landing host box, a clear creek box to support RDMA NIC based scale out solution, and an Emerald Pools box with 8 OAMs. Three boxes are connected by high speed PCIe cables that can support up to PCIe Gen4 speed.

In the Angels Landing box, we are using 4x Intel Cooperlake CPUs. Each of them supports 1x 100G OCP3 NIC for uplink data transfer. In Figure 5, CPU0 and CPU1 belong to Motherboard 0 and CPU2 and CPU3 belong to Motherboard 1, where MB0 is the main board and MB1 is the secondary board.

In the Clear Creek box, there are 8x 200G NICs to enable scale out configuration. Each OAM module can directly access to 1x NIC and use this interface to share data with other OAMs in the cluster (not just within the box).

Also, we will support 16x NVMe SSDs to enable the offload for huge models. By swapping different SSD trays, we can support either M.2 form factor or E1.S form factor. Please check with Clear Creek Spec for more information.

Multiple ZionEx systems can be connected through the RDMA network and create an accelerator pod. It has the capability to support very large models which require high computational powers, high host power and high DRAM bandwidth.

7.2 ZionEx cable connection

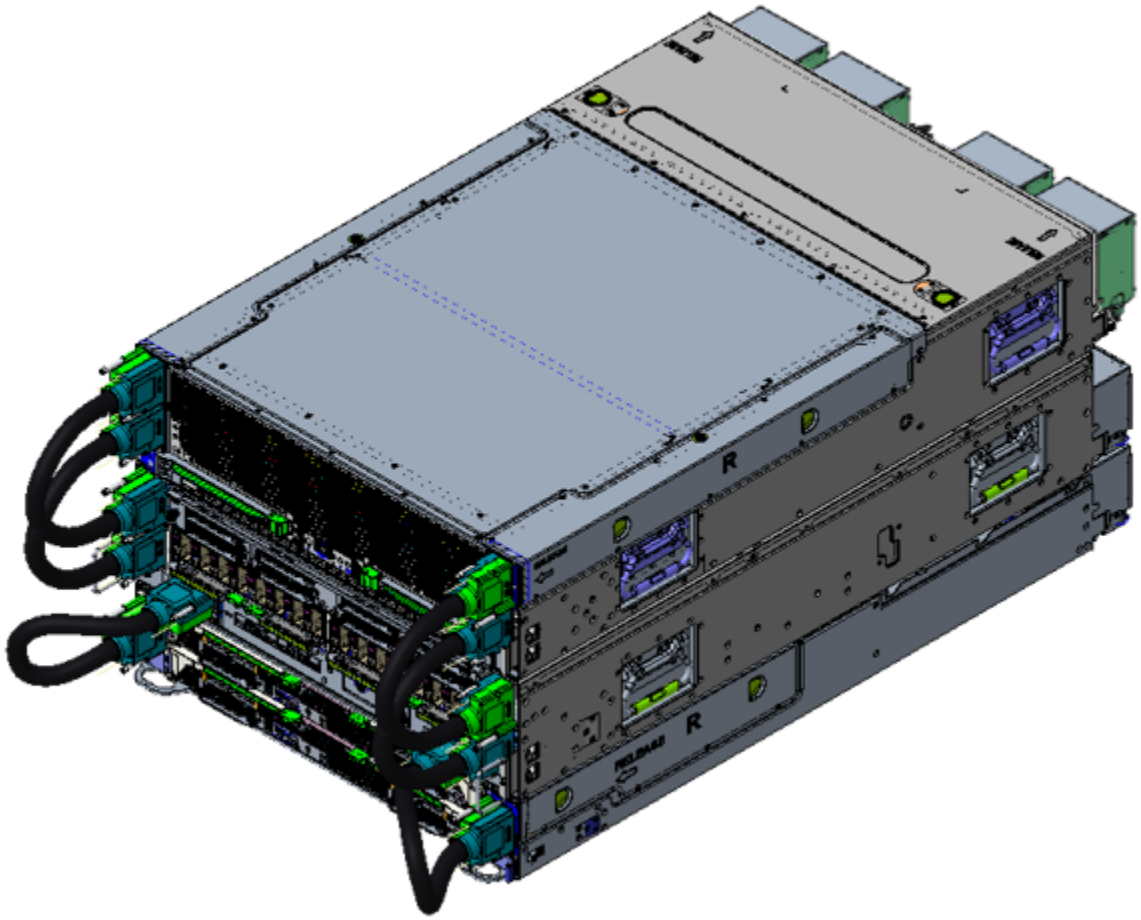
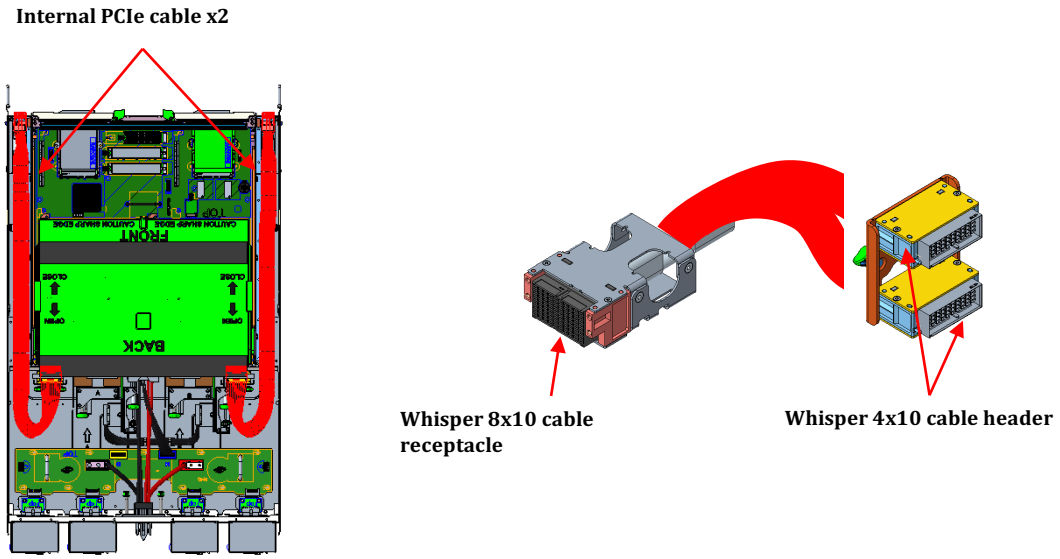


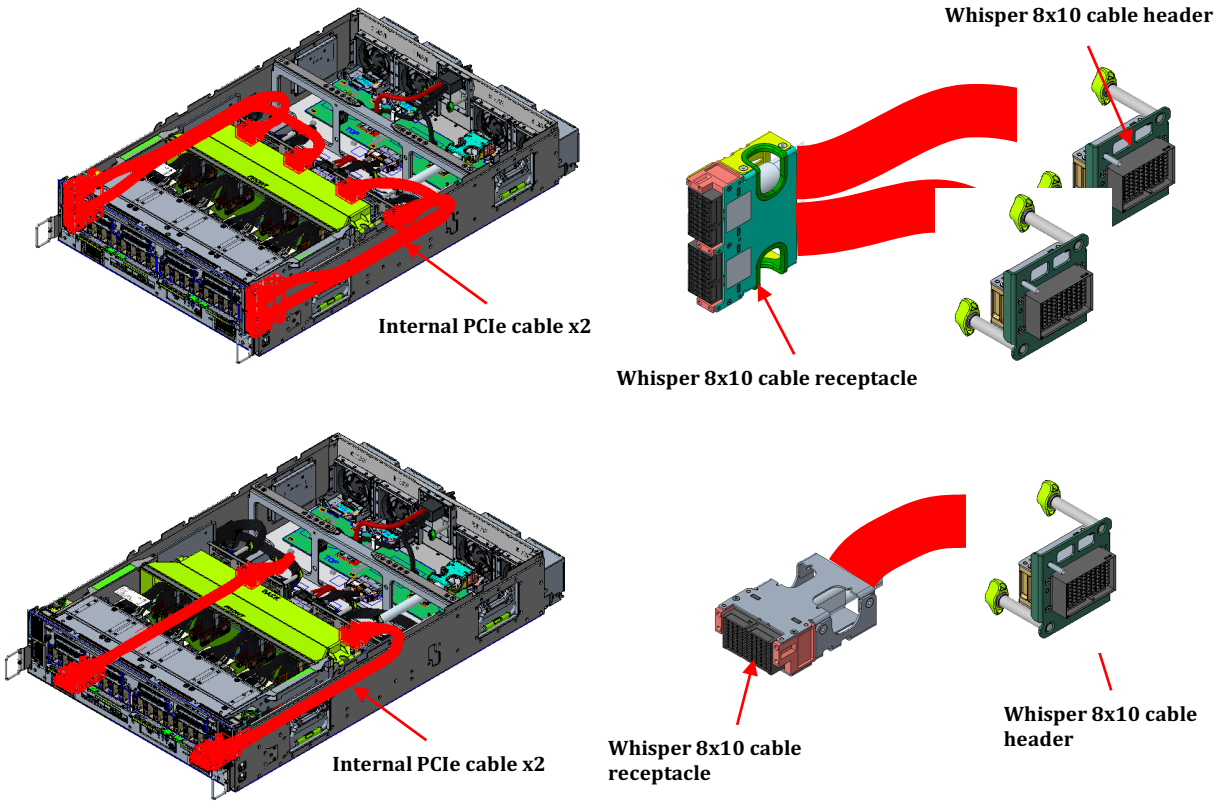
Figure 6: ZionEx Configuration with Cable Connection

Each subsystem is connected by high-speed cables. In Figure 6 you can see from bottom to up there are AL, CC and EP boxes. We are using whisper cables to connect the PCIe and sideband IOs between each box.

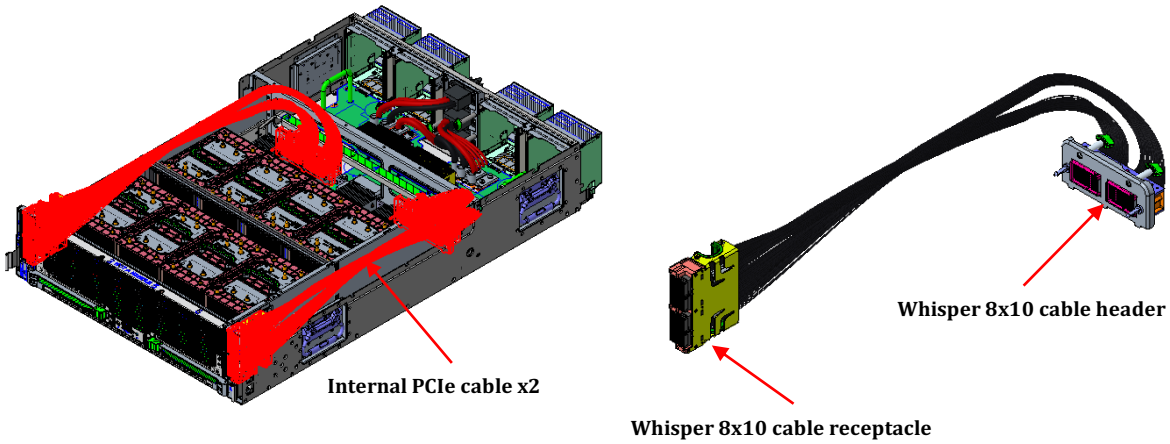
We have several different types of high-speed cables in Zion Design listed below:



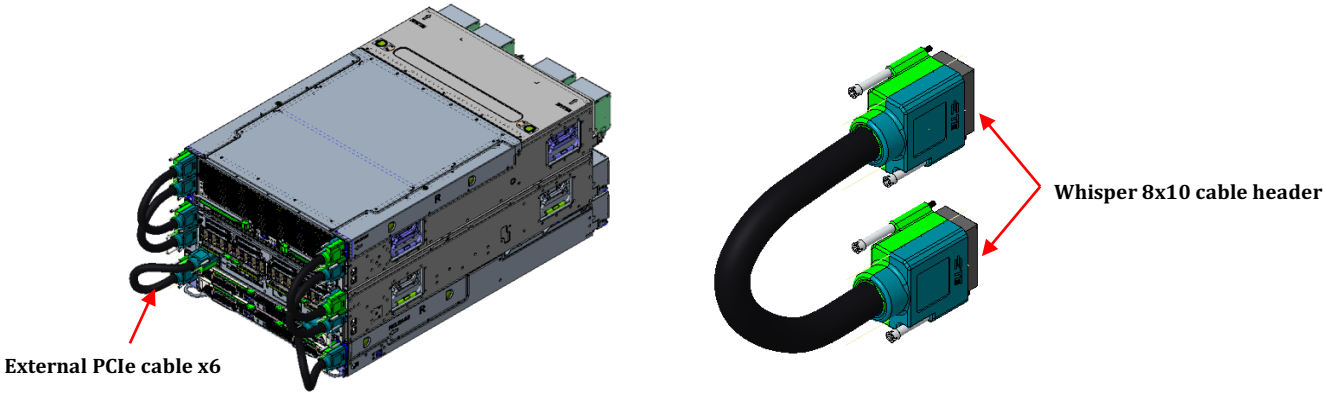
7a. Internal cable used in Angels Landing box



7b. Internal cable used in Clear Creek box



7c. Internal cable used in Emerald Pools box



7d. External cable used in the front panel to connect all three boxes

Figure 7: ZionEx PCIe cable drawing

7.3 ZionEx OOB connection

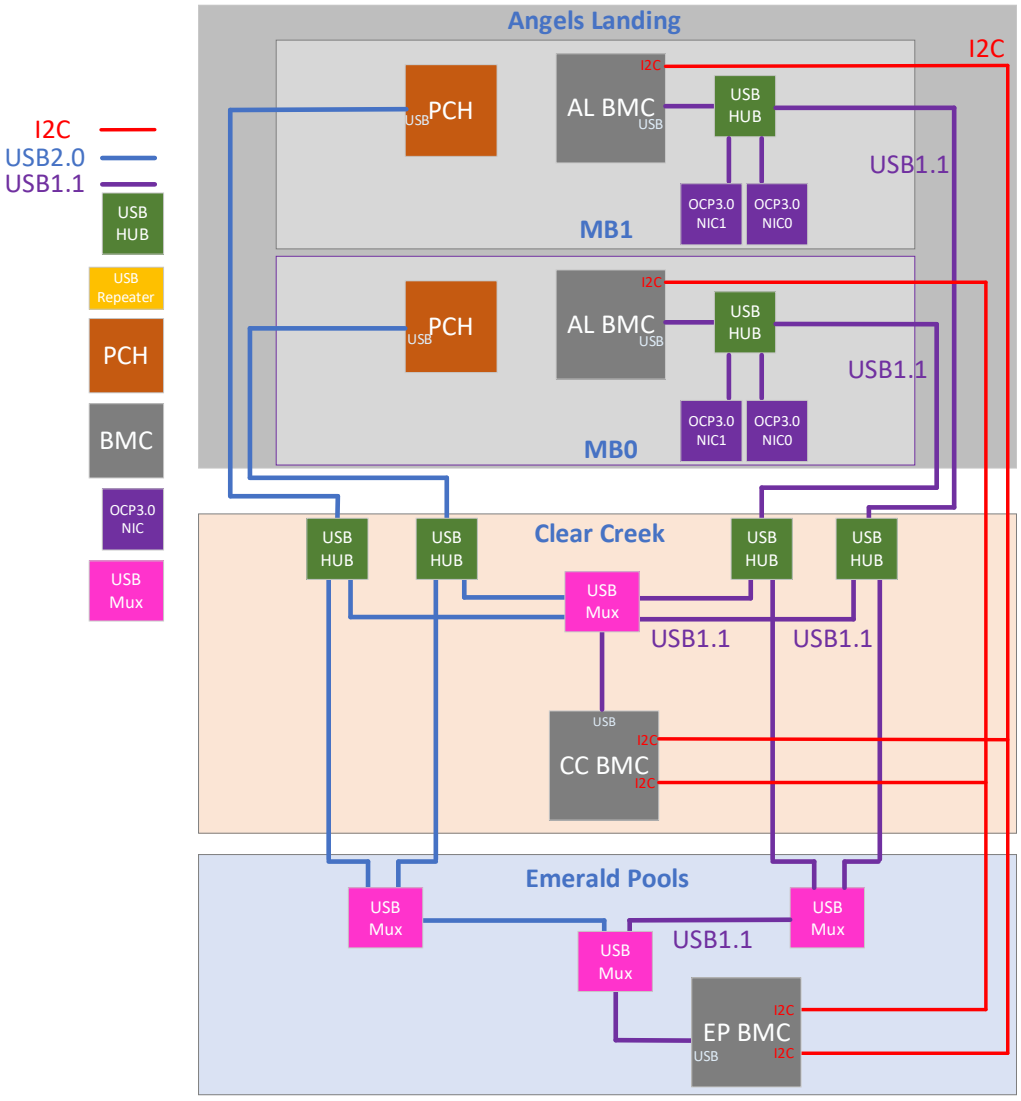


Figure 8: ZionEx BMC connection

Fgiure 8 explains how we connect OOB (out-of-band) components between three boxes. AL has two BMCs and we are using USB links to connect the BMC to CC and EP box.

Ethernet-over-USB stacks are implemented here so users can access the CC/EP BMC chip through AL BMCs. One typical configuration is using AL MB0 BMC to bridge to EP BMC and AL MB1 BMC to bridge CC BMC.

Also we defined a power good pin in EP and CC box to coordinate the power on and power off sequence. The AL system can get this GPIO through the cable.

8. Zion4S system

8.1 Zion4S block diagram

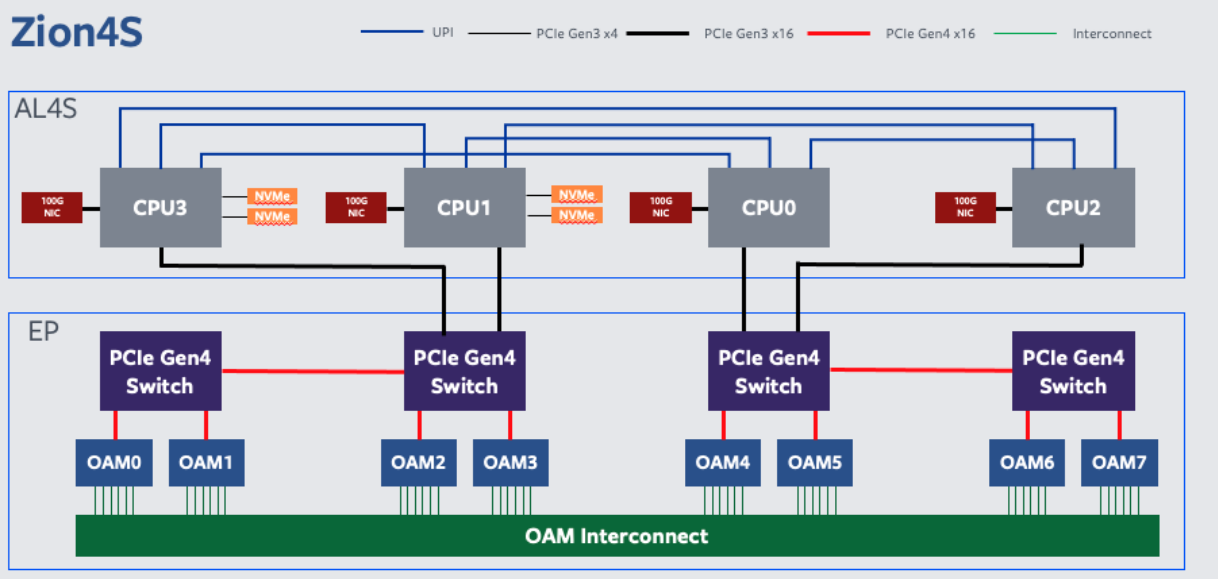


Figure 9: Zion4S Block Diagram

Zion4S is a system configuration that has a 4-socket host and 8x Accelerator box, without a clear creek box. In this setup, the system can provide the same host resource and accelerator compute power, without the capability to scale out. It will be useful in the case where the model size is not that big (100s of GB) and host resource demand is high.

8.2 Zion4S cable connection

The reconfiguration is simple as shown in Figure 10. As we use uniform cable type. In this case, we just need to connect AL box directly to EP box.

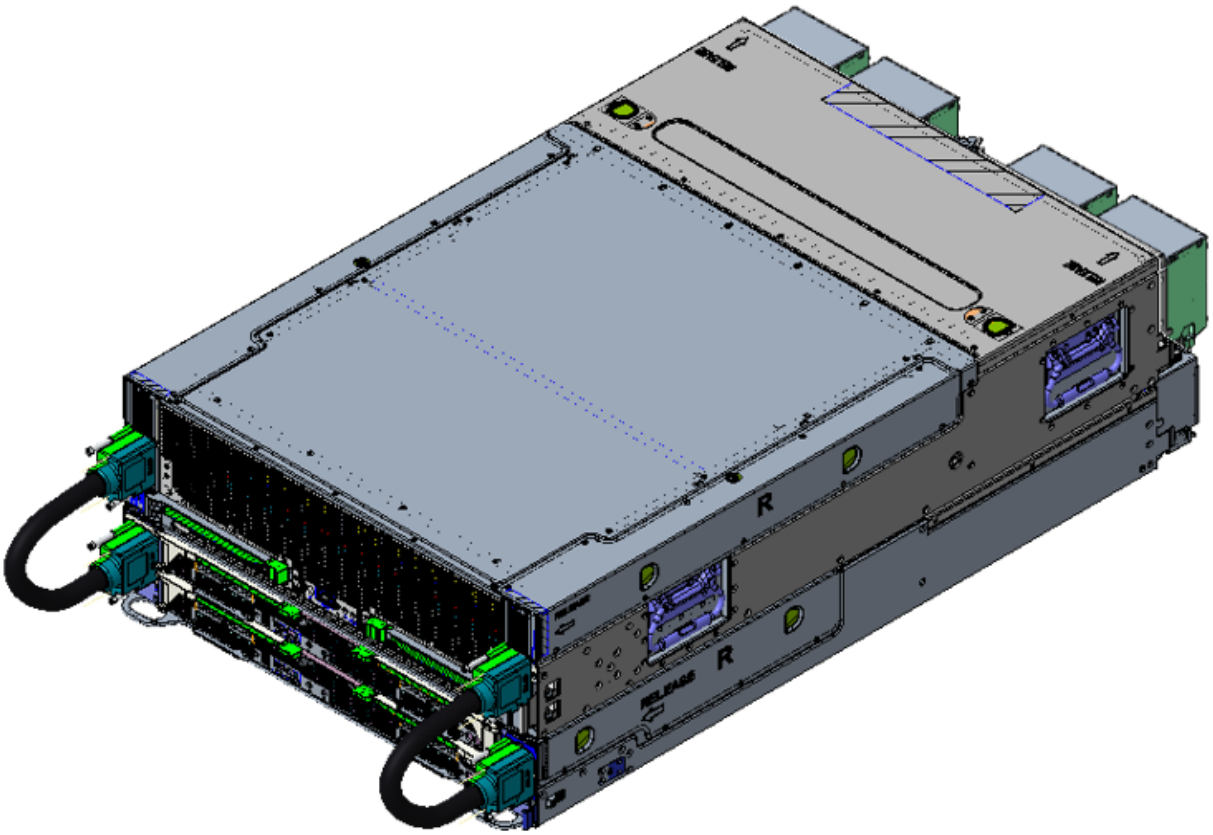


Figure 10: Zion4S Configuration with Cable Connection

8.3 Zion4S OOB connection

In this case, the EP BMC chip is connected to MB0 BMC through the USB interface. System design does have the flexibility to connect either way.

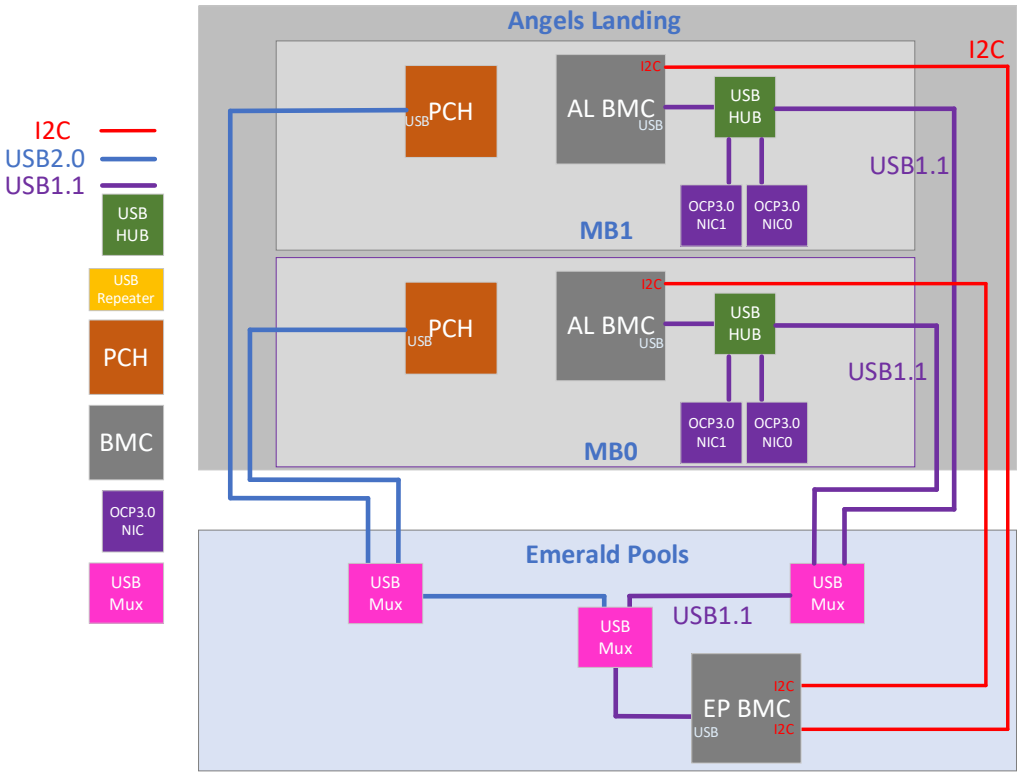


Figure 11: Zion4S BMC connection

9. Zion2S system

9.1 Zion2S block diagram

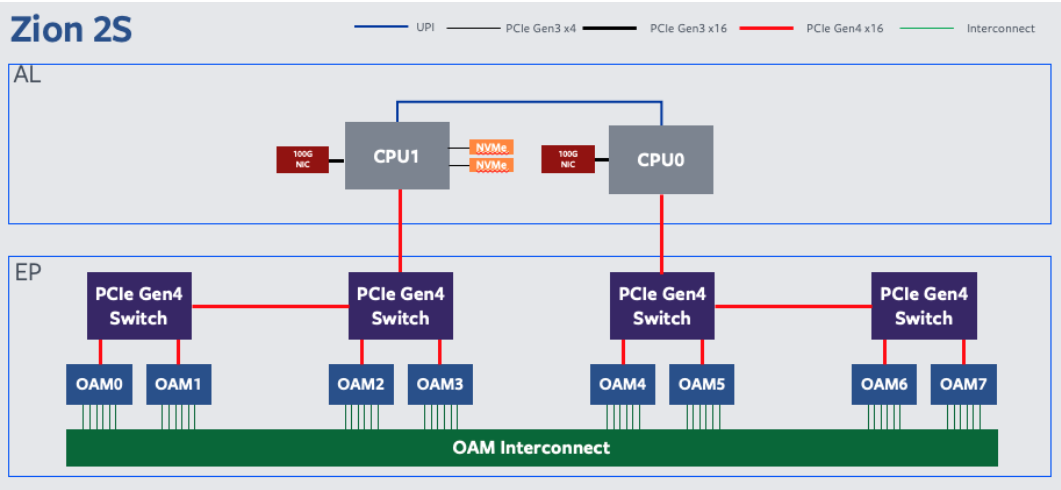


Figure 12: Zion2S Block Diagram

Zion2S is very similar to Zion4S expect we deploy AL box as a 2S host instead of a 4S host. A lot of AI models are very compute heavy while the model size is very small. They don't need a lot of host resources either. In this case, users can use Zion2S configuration to optimize the TCO and improve the reliability of their hardware fleet.

You can check Zion2S configuration and BMC connection in Figure 13 and Figure 14.

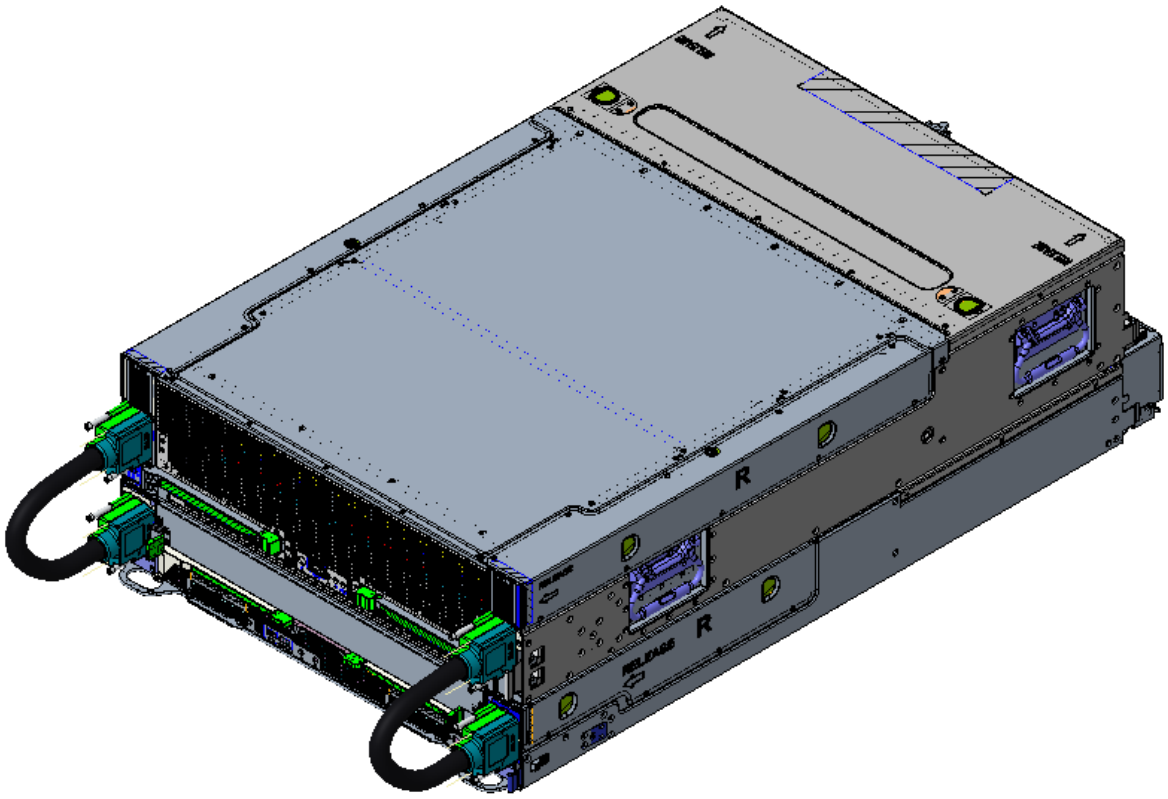


Figure 13: Zion4S Configuration with Cable Connection

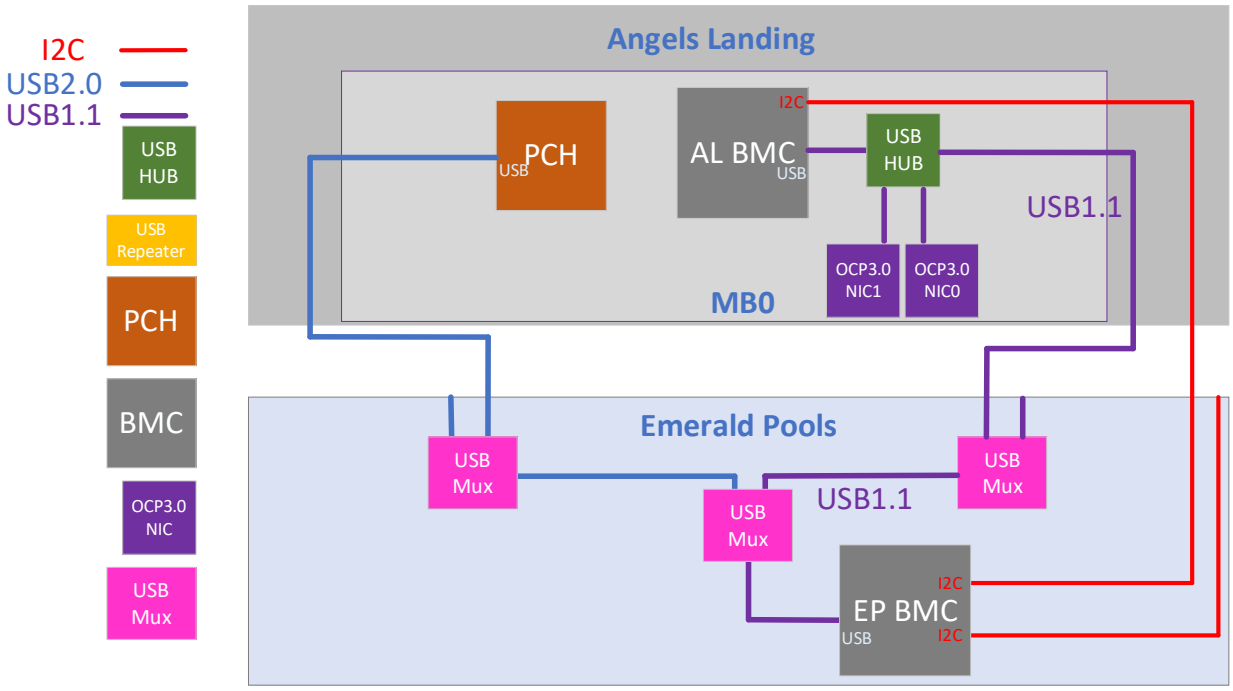


Figure 14: Zion2S BMC connection

10. Rack Compatibility

Zion system fit in the ORv2 rack. Angels Landing box takes 2OU. Clear Creek box takes 4OU and Emerald Pools takes 4OU.

As mentioned before this document introduces system level configuration. More design details are covered in subsystem design specifications.

11. System Firmware

All products seeking OCP Accepted™ Product Recognition must complete the Open System Firmware (OSF) Tab in the 2021 Supplier Requirements Checklist.

If based on an open bios (like AMI's Aptio-OpenEdition), a completed checklist shall be uploaded and made available on the OCP Github.
Note to authors: replace [vendor_name] and [product_name] with actual company name and product identifier.

12. Hardware Management

12.1 Compliance

All products seeking OCP Inspired™ or OCP Accepted ™ Product Recognition shall comply with the OCP Hardware Management Baseline Profile V1.0 and provide such evidence by completing the Hardware Management Tab in the 2021 Supplier Requirements Checklist.

12.2 BMC Source Availability (if applicable)

All Products seeking OCP Accepted™ Product Recognition shall have source code and binary blobs submitted for BMC, if applicable.

The BMC management source code shall be uploaded at:
https://github.com/opencomputeproject/Hardware-Management/[vendor_name]/[product_name]

If the BMC is based on an open source BMC (like AMI's MegaRAC-OpenEdition), the BMC source code shall be uploaded and made available on the OCP Github.

13. Security

All products seeking OCP Inspired™ or OCP Accepted™ Product Recognition shall have a completed Security Profile in the 2021 Supplier Requirements Checklist. Whether the answer is a yes or no, the profile must be completed. For Additional Security Badges (Bronze/Silver/Gold), please fill out the Security Profile in accordance with the requirements for that level. Security Badges will be reassessed on an annual basis as requirements are subject to change.

14. Reference

[1] Facebook Angels Landing System Specification 1.0
[2] Facebook Clear Creek System Specification 1.0
[3] Facebook Emerald Pools System Specification 1.0

Appendix A - Requirements for IC Approval (to be completed Contributor(s) of this Spec)

List all the requirements in one summary table with links from the sections.

Requirements	Details	Link to which Section in Spec
Contribution License Agreement	Ocp Owf-Cla Final	Link to Sec 1
Are All Contributors listed in Sec 1: License?	Yes	
Did All the Contributors sign the appropriate license for this spec? Final Spec Agreement/HW License?	Yes	
Which 3 of the 4 OCP Tenets are supported by this Spec?	Openness Efficiency Impact Scale	List reasons here. Link to presentation if separate.
Is there a Supplier(s) that is building a product based on this Spec? (Supplier must be an OCP Solution Provider)	Yes	List Supplier Name(s)
Will Supplier(s) have the product available for GENERAL AVAILABILITY within 120 days?	Yes	Please have each Supplier fill out Appendix B.

Appendix B-____ - OCP Supplier Information (to be provided by each Supplier of Product)

Company:
Contact Info:

Product Name:
Product SKU#:
Link to Product Landing Page:

Please complete the following 2021 Supplier Requirements. This link will allow you to create a copy for your product-specific requirements.

For OCP Inspired™,

- All Suppliers must be a OCP Solution Provider.
- All Suppliers must run the Hardware Management Conformance Checks and all products must meet the OCP Hardware Baseline Profile v1.0.0.
- All Suppliers must fill out a Security Profile (No Badge Level) for their product.

For OCP Accepted™, Supplier details are required.

- All Suppliers must be a OCP Solution Provider.
- All Suppliers must run the Hardware Management Conformance Checks and all products must meet the OCP Hardware Baseline Profile v1.0.0.
- All Suppliers must fill out a Security Profile (No Badge Level) for their product.
- All Products must meet the Open System Firmware requirements.
- All Products must have source code for BMC, if applicable. This must be in the OCP Github repository.

List all the requirements in one summary table with links from the sections.

Requirements	Details	Links
Which Product recognition?	OCP Accepted™ or OCP Inspired™	Provide Marketplace Link
If OCP Accepted™, who provided the Design Package?		Link
2021 Supplier Requirements for your product(s)		Link