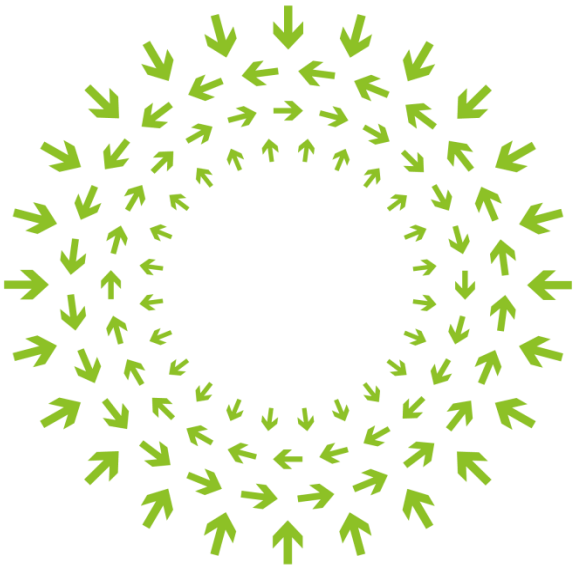


OPEN
Compute Project

Yosemite V3.5 Platform Design Specification 1v2

Authors:
Michael Haken, Mechanical Engineer, Meta
Chenyu Xu, Mechanical Engineer, Meta
Yueming Li, Thermal Engineer, Meta
Ben Wei, Software Engineer, Meta
Kiran Vemuri, Hardware Engineer, Meta
Matt Bowman, Hardware Engineer, Meta
Yeshaswy Rajupalepu, Hardware Engineer, Meta
Abe Garcia, Hardware Engineer, Meta
Shenli Zou, Power Engineer, Meta
Chloe Xu, Mechanical Engineer, Meta
Michael Garner, Software Engineer, Meta
Jayati Athavale, Thermal Engineer, Meta
Zhengyu Yang, Hardware Validation Engineer, Meta
Mehryar Rahmatian, Hardware Systems Engineer, Security, Meta

1 Contents

1 Contents 2
2 License 4
2.1 Copyrights and Trademarks 5
3 Compliance with OCP Tenets 6
3.1 Openness 6
3.2 Efficiency 6
3.3 Impact 6
3.4 Scale 6
3.5 Sustainability 6
Inclusive Terminology 6
4 Version Table 7
5 Scope 7
6 Rack Compatibility 7
7 Overview 8
8 Physical Specifications 13
8.1 Platform Block Diagram 13
8.2 Yosemite V3.5 Subsystems 15
8.3 Yosemite V3.5 Platform Power Delivery 23
8.4 SMBus Block Diagram 25
8.5 1S Server 28
8.6 Network Options 31
9 Thermal 33
9.1 Data Center Environmental Conditions 33
9.2 Server Operational Conditions 34
9.3 Thermal Kit Requirements 36
10 I/O System 37
10.1 Front facing Server Modules 37
10.2 Network 37
10.3 Server Slots Assignment 37
10.4 Front Panel 37
10.5 Fan Module Connector 39
11 Mechanical 40

11.1 Yosemite V3.5 Chassis 40
11.2 Yosemite V3.5 Sled 41
11.3 1S Server Blade 41
11.4 Silkscreen 42
11.5 Retention 42
12 Onboard Power System 43
12.1 Input Voltage Level 43
12.2 48V support 44
12.3 Platform Power Budget 44
12.4 Capacitive Load 45
12.5 Hot Swap Controller Circuit 45
12.6 1S Server Power Management 46
12.7 VR Efficiency 46
12.8 Power Policy 47
12.9 P12V_PSU to GND Clearance 47
13 Environmental Requirements 49
13.1 Vibration and Shock 49
14 Prescribed Materials 49
14.1 Disallowed Components 49
14.2 Capacitors and Inductors 50
14.3 Component De-rating 50
15 Hardware Management 50
15.1 1S Server I^3C Connections 50
15.2 1S Server Serial Connections 51
15.3 1S Server Discovery Process 51
15.4 1S Server Power-on Sequence 51
15.5 Network Interface 52
15.6 BMC Multi-Node Requirements 52
15.7 Local Serial Console and Serial-Over-LAN 52
15.8 Graphics and GUI 53
15.9 Remote Power Control and Power Policy 53
15.10 POST Codes 53
15.11 System LEDs and Buttons 53

15.12 Time Sync 55
15.13 Power and Thermal Monitoring, and Power Limiting 55
15.14 Sensors 56
15.15 Event Log 57
15.16 Fan Speed Control in BMC 58
15.17 BMC Firmware Update 59
15.18 Hot Service Support 59
15.19 OpenBMC 60
15.20 OpenBIC 60
15.21 Security 60
16 Security 60
16.1 TPM (Trusted Platform Module) 60
17 Labels and Markings 60

2 License

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:

Meta, Inc.

Usage of this Specification is governed by the terms and conditions set forth in **Open Web Foundation Modified Final Specification Agreement ("OWFa 1.0")**

Notes:

1) 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.

2.1 Copyrights and Trademarks

Intel® is a trademark of Intel Corporation in the U.S. and/or other countries.

Xeon® is a trademark of Intel Corporation in the U.S. and/or other countries.

3 Compliance with OCP Tenets

3.1 Openness

The Yosemite 3.5 and Crater Lake server specification describes a design which can be used to support many configuration types. It can be leveraged by other OCP partners as a standalone server, or it can drive other use cases with support for expansion options, oBMC, ORv2/ORv3.

3.2 Efficiency

Yosemite 3.5 continues to utilize shared resources such as fans, MHNIC along with support for ORv3 racks, improved thermal efficiency and 4th Gen Intel® Xeon® processors.

3.3 Impact

The Yosemite 3.5 and Crater Lake server have been designed to improve over Yv3.0 and Delta Lake in terms of compute/power density, power/thermal efficiency and IO bandwidth resulting in an overall performance/Watt gains.

3.4 Scale

The Yosemite 3.5 and Crater Lake server are designed for large scale data center deployment and operations. The system is designed with debug hooks, ease of serviceability, and reliability as a focus. YV3.5 sled design continues to be front serviceable, hot-pluggable and support service touch point indicators. The Design continues to support openBMC architecture and the OCP debug options such as OCP debug card and At Scale Debug.

3.5 Sustainability

The Yosemite 3.5 and Crater Lake server have been designed for sustainability. The spec submission will maximize transparency of environmental impacts of the contribution, with the aspiration of improvement over time.

Inclusive Terminology

This specification was written with inclusive terminology to reduce biased terms. The deprecated and new preferred terminology can be found at https://www.opencompute.org/documents/ocp-terminology-guidelines-for-inclusion-and-openness

4 Version Table

Date	Version #	Author	Description
2 Sep 2020	0.1	Matt Bowman	Initial draft based on YV3 OCP specification
14 Dec 2020	0.2	Matt Bowman	Substantial updates in all sections based on early feedback
8 Jan 2021	0.3	Matt Bowman	Fixed Baseboard management cabling diagrams in figure 5-6 and 5-7
25 Feb 2021	0.4	Matt Bowman	Removed references to LFF OCP NIC. Only SFF support is needed Aligned blade slot numbering to YV3
14 Apr 2021	0.5	Kiran Vemuri	Updated fan connector details to support 15k fan Updated overview section to add clarification about gen 5 data rates
1 June 2023	0.6	Yeshaswy Rajupalepu	Updated section 5 with pinouts. Updated overall format and grammar. Updated block diagrams in section 5
2 Oct 2023	1.0	Abe Garcia	Formatting fixes
14 Nov 2023	1.1	Abe Garcia	Formatting fixes
29 Nov 2023	1.2	Abe Garcia	Fixed section referencing

5 Scope

This document defines the technical details of the Yosemite V3.5 base platform.

6 Rack Compatibility

The Yv3.5 server chassis is designed to support both ORv2 and ORv3 open rack standards. Rack density is subject to change based on IO population and workloads and they must be evaluated by adopters on a case-by-case basis. Note that the YV3.5 Medusa and VPDB boards have different SKUs between ORv2 and ORv3 variants of the sled. Ensure you are using the right sled variant. Find more information and details in the Power section of this document.

7 Overview

This document details an incremental update to Meta's multi-node server platform Yosemite Valley v3 (YV3). This new platform has been named Yosemite Valley v3.5 (YV3.5). Most of the existing YV3 platform is leveraged with the YV3.5 changes focused on enabling higher bandwidth interconnects and technology updates.

Changes from YV3 to YV3.5 include:

- The BIC and BMC on the baseboard have been updated to support eSPI, I3C, and secure boot.
- The baseboard supports PCIe Gen 4 data rates. With a PCB material change, a new rev of baseboard could support PCIe Gen 5 data rates.
- New server blades that have 2C connector to support routing up to x8 lanes of PCIe from the servers to the baseboard. The PCIe port is Gen 5 capable but would require an updated sled management cable and baseboard to support Gen 5 rates.
- Fans for Yv3.5 platform have been upgraded to 8038 15k fans from 10k fans used in Yv3 to meet increased impedance and airflow demands
- In addition to ORV2, ORV3 is also supported by YV3.5 with new YV3.5 ORV3 chassis.

The Yosemite V3.5 (YV3.5) platform uses the same blade organization as YV3. Up to four 1U server blades or two 2U server blades can be installed per sled. YV3 and YV3.5 share the same 4OU chassis and sled mechanical design for ORV2, and YV3.5 continues to support ORV3 with a slightly modified 4OU chassis. YV3.5 has an updated baseboard that hosts the baseboard management controller (BMC) and provides network connectivity through an OCP3.0 Network Interface Card (NIC).

The baseboard connects to each individual server module with cables to support various configurations as described in later sections of this document. The baseboard and server blades have independent onboard Hot Swap Controller (HSC) for power regulation, monitoring, and are hot-swap capable.

The BMC on the baseboard is the primary management controller of the platform. Its primary functions include:

1. Control and monitor all thermal sensitive components on the platform
2. Control and monitor power usage of major commodities of the platform such as:
    a. power, voltage and current of server blade and its onboard voltage regulators (VR)
    b. power, voltage and current of the platform
3. Fan control and monitoring
4. Logging of different platform and system events
5. Configuration of any programmable devices for its intended operation
6. Tracking of all Field Replaceable Units (FRU) on the platform

YV3.5 has the same serviceability improvements as YV3. It supports front loading server blades into a powered system eliminating hot service and cable management complexities.

The Yosemite V3.5 Platform is designed to be compatible with the OCP (Open Compute Project) Open Rack V2 (ORv2) and Open Rack V3(ORv3) specification. Please refer to the corresponding OCP Open Rack V2 / Open Rack V3 documentation for more details about the rack. The Yosemite V3.5 Platform is a chassis that can be safely inserted or removed to/from an Open Rack. You can find more details in the mechanical section of this document.

Simplified pictures of the Yosemite V3.5 platform and components can be seen in Figures 7-1 to 7-6.

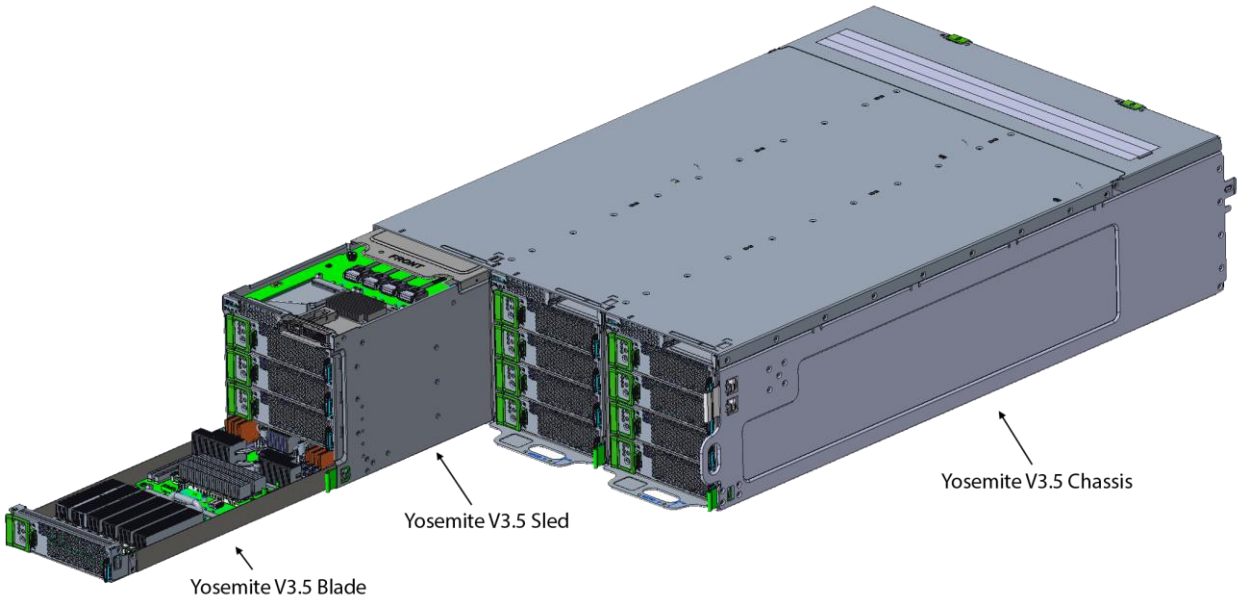


Figure 7-1: Yosemite V3.5, Chassis View

The YV3.5 platform has two types of 4OU chassis to support ORv2 as well as ORv3. Each chassis can hold one, two, or three sleds. Each sled has its own fans and power input. Within the chassis, there is no communication between the sleds.

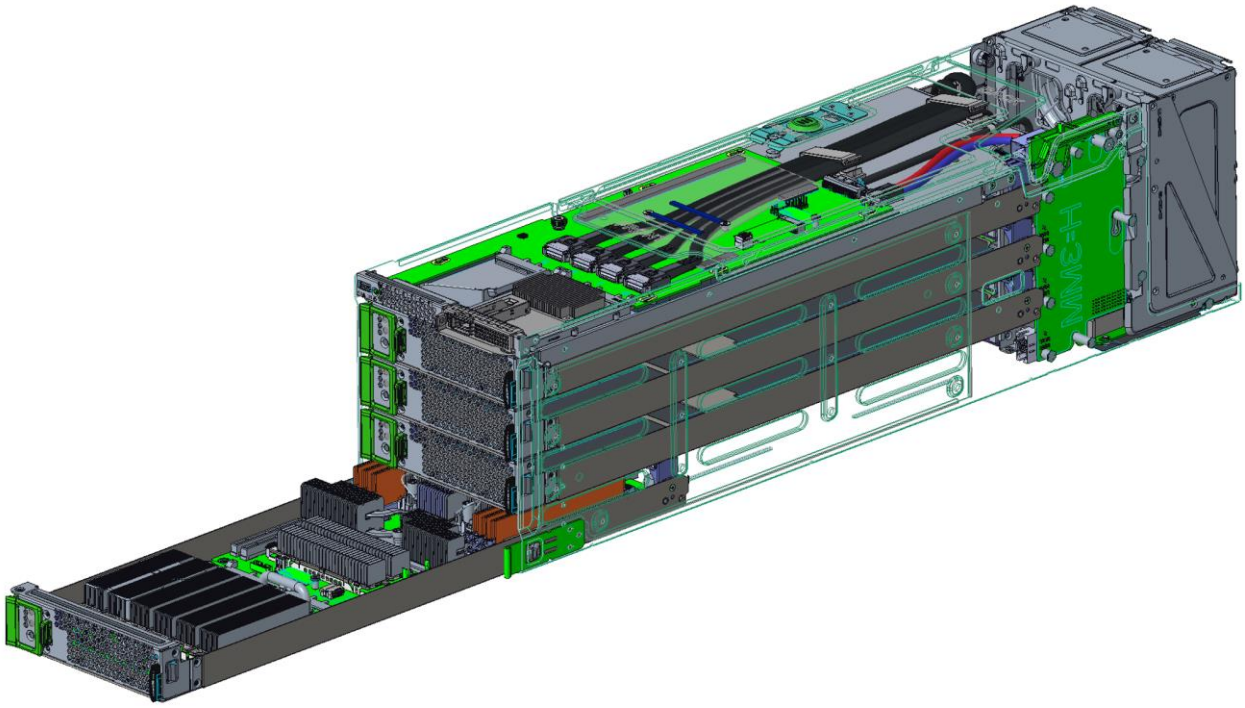


Figure 7-2: Yosemite V3.5, Sled View

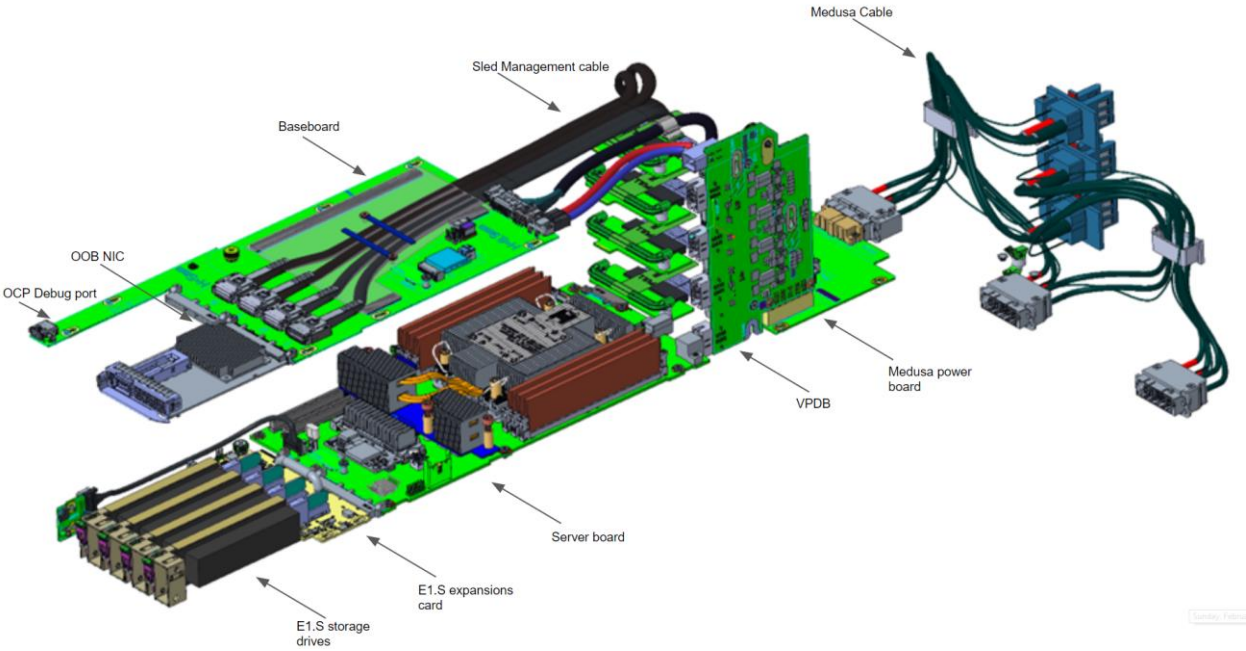


Figure 7-3: Yosemite V3.5, Sled Components

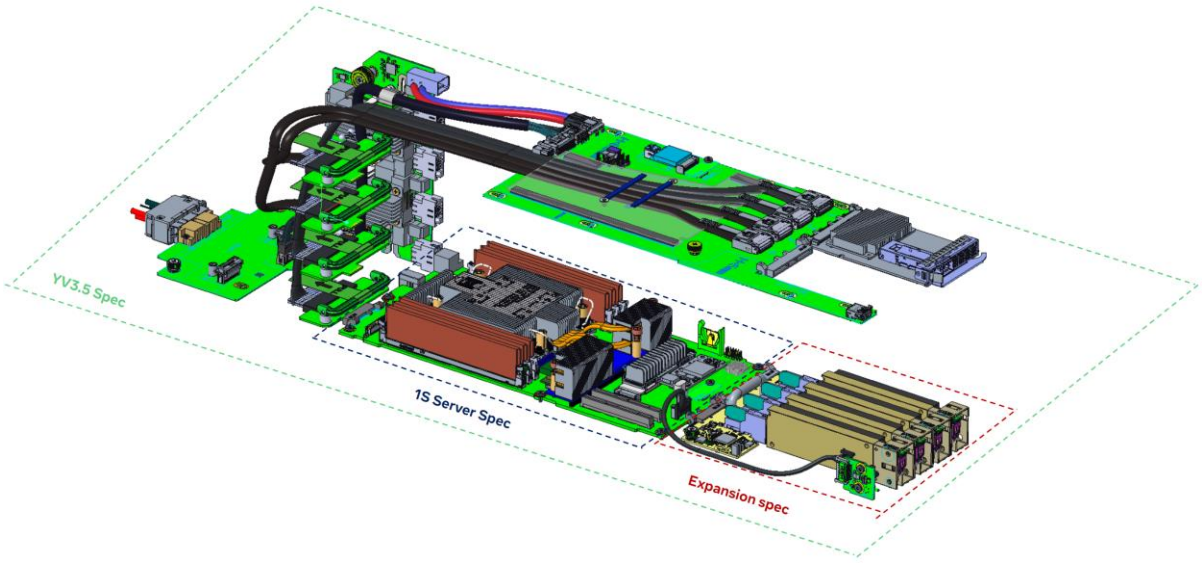


Figure 7-4: Yosemite V3.5, Specification View

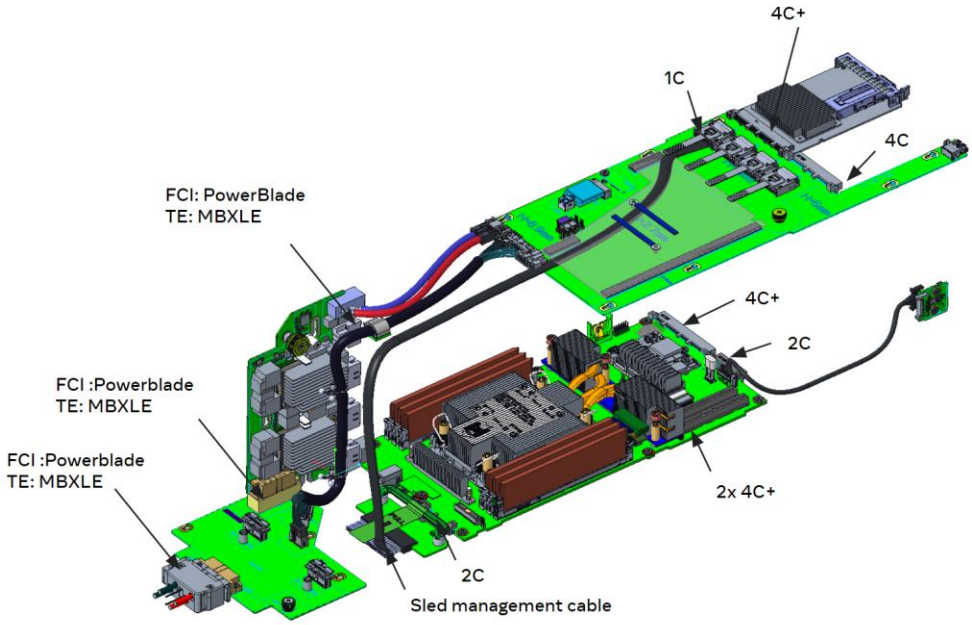


Figure 7-5-1 : Yosemite V3.5, Connector Map, ORV3 –48V Power delivery, Crater Lake

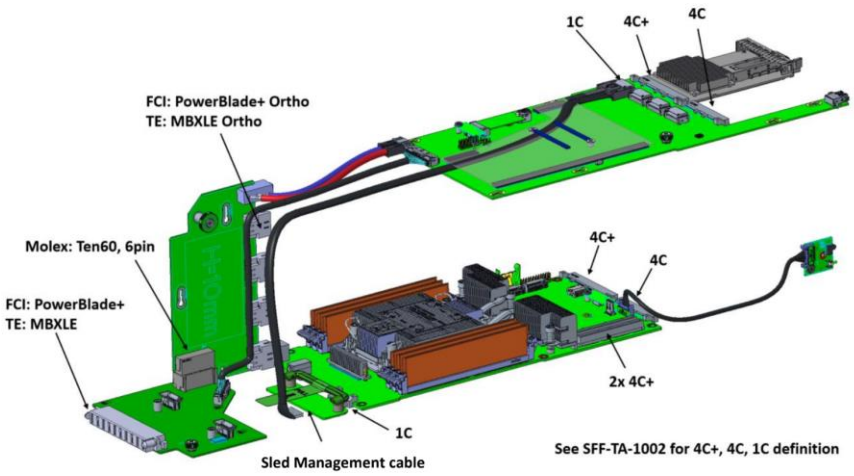


Figure 7-5-2 : Yosemite V3.5, Connector Map, ORV2 –12V Power delivery, Cooper Lake

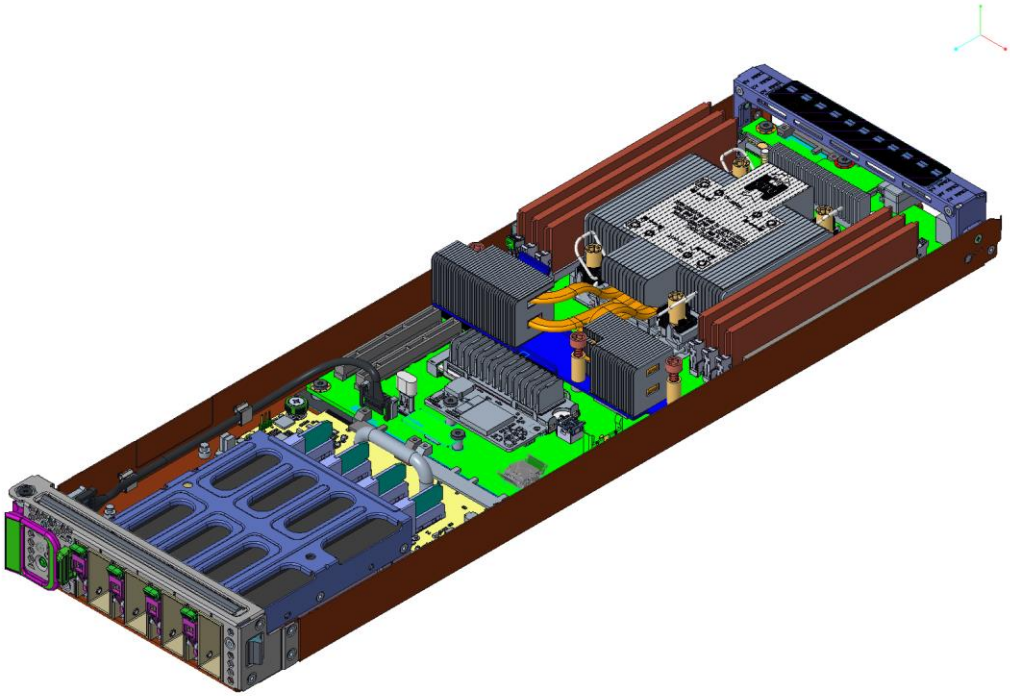


Figure 7-6: Yosemite V3.5, Example Blade Configuration

8 Physical Specifications

8.1 Platform Block Diagram

The following diagrams are examples of Class 1 configurations of YV3.5. Class 1 configurations have a BMC on the baseboard, and the server blades utilize the multihost NIC for network connectivity. Figure 8-1 illustrates a Yosemite V3.5 sled block diagram with four 1OU blades.

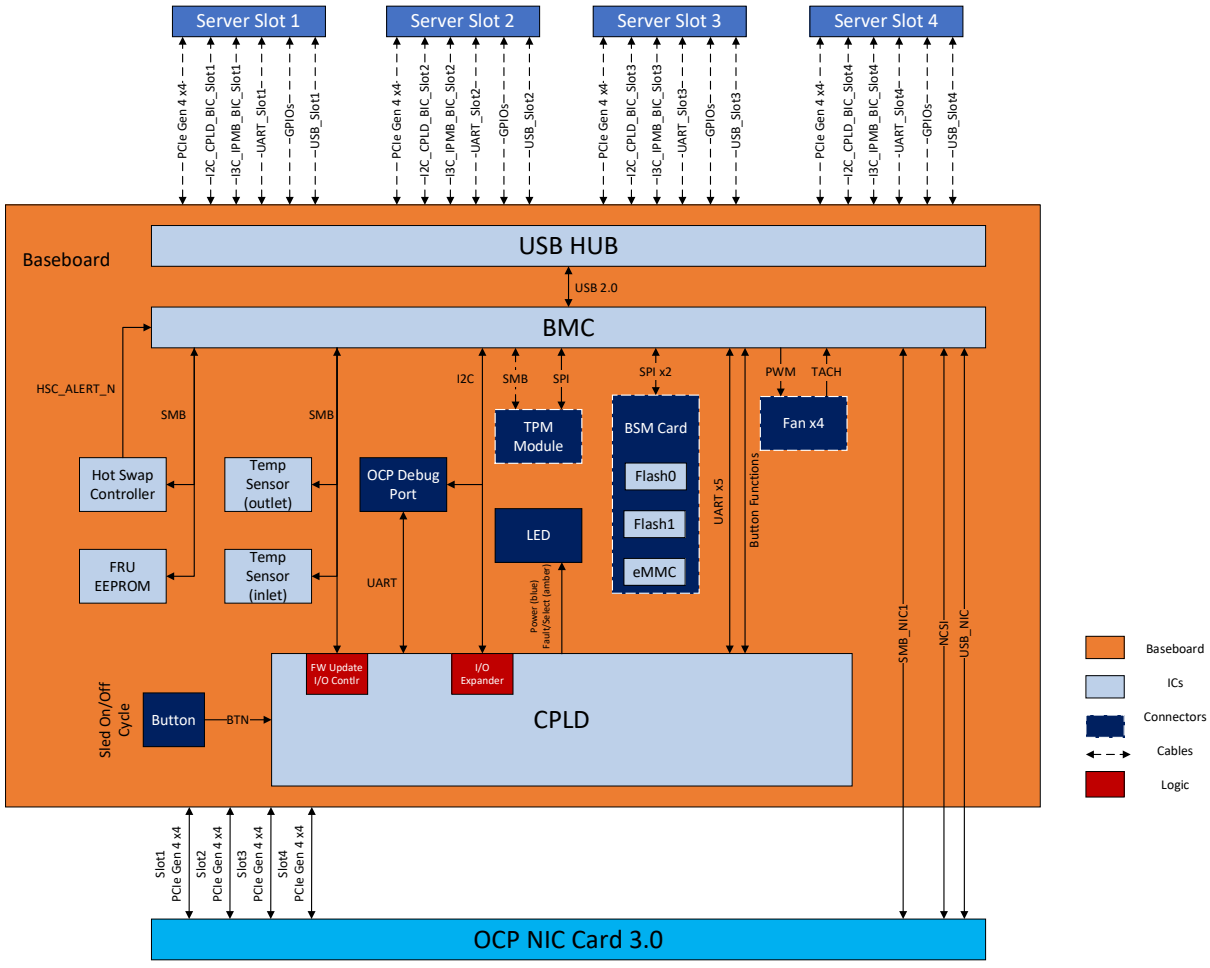


Figure 8-1: Class1 YV3.5 with four 1U blades

Figure 8-2 illustrates another configuration of a YV3.5 sled that has two 2U blades. Each blade has a x8 PCIe Gen 5 connection to the multihost NIC.

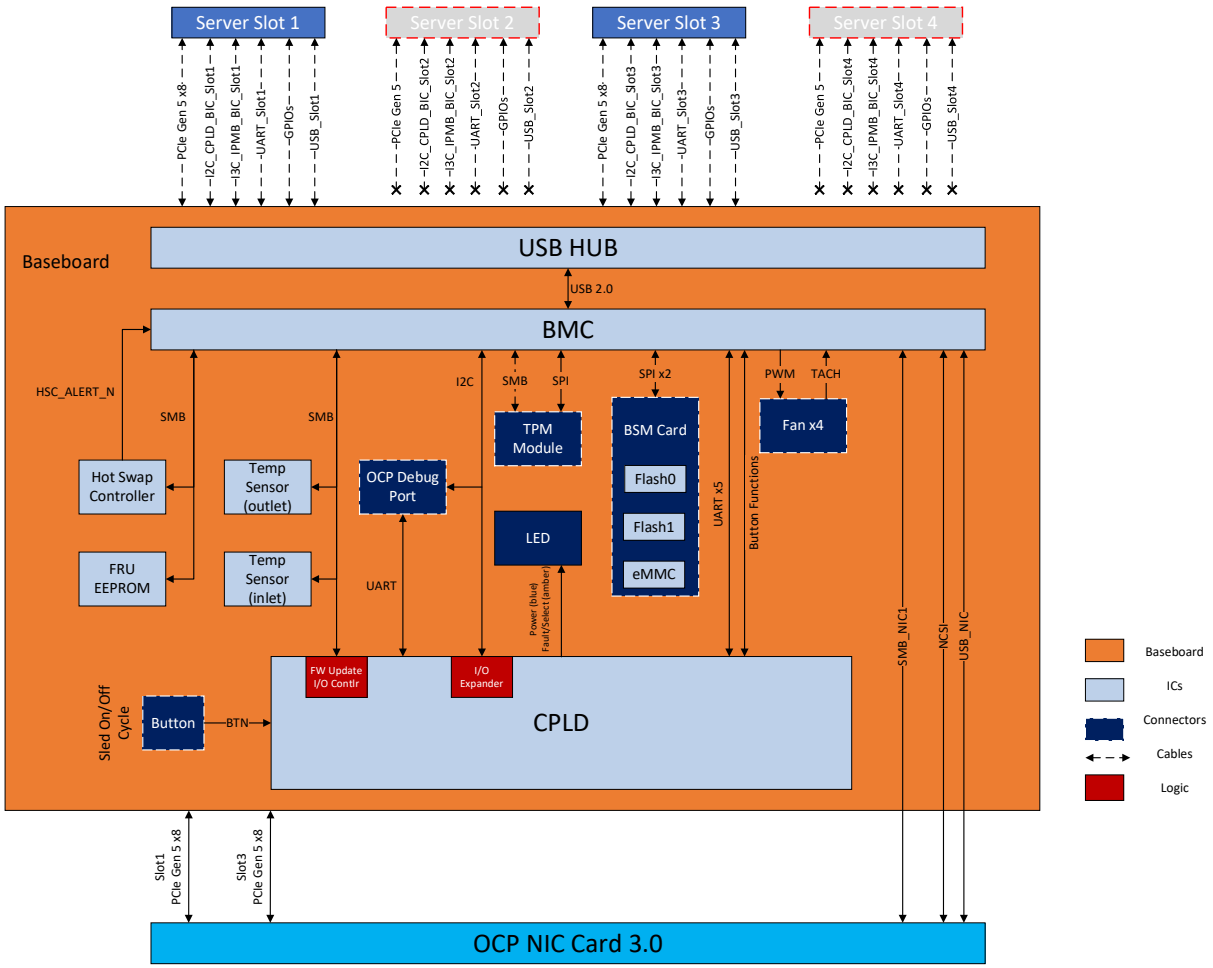


Figure 8-2 YV3.5 with two 2U blades

The YV3.5 baseboard also supports Class 2 configurations. Class 2 configurations have a Bridge IC on the baseboard instead of a BMC, and each server blade uses a dedicated single host NIC. A BMC is inside each blade and is paired with each NIC to provide the management interface. Note that Class 2 functionality on YV3.5 is desired but lower priority behind Class 1. Figure 8-3 shows the block diagram of a Class 2 YV3.5 baseboard.

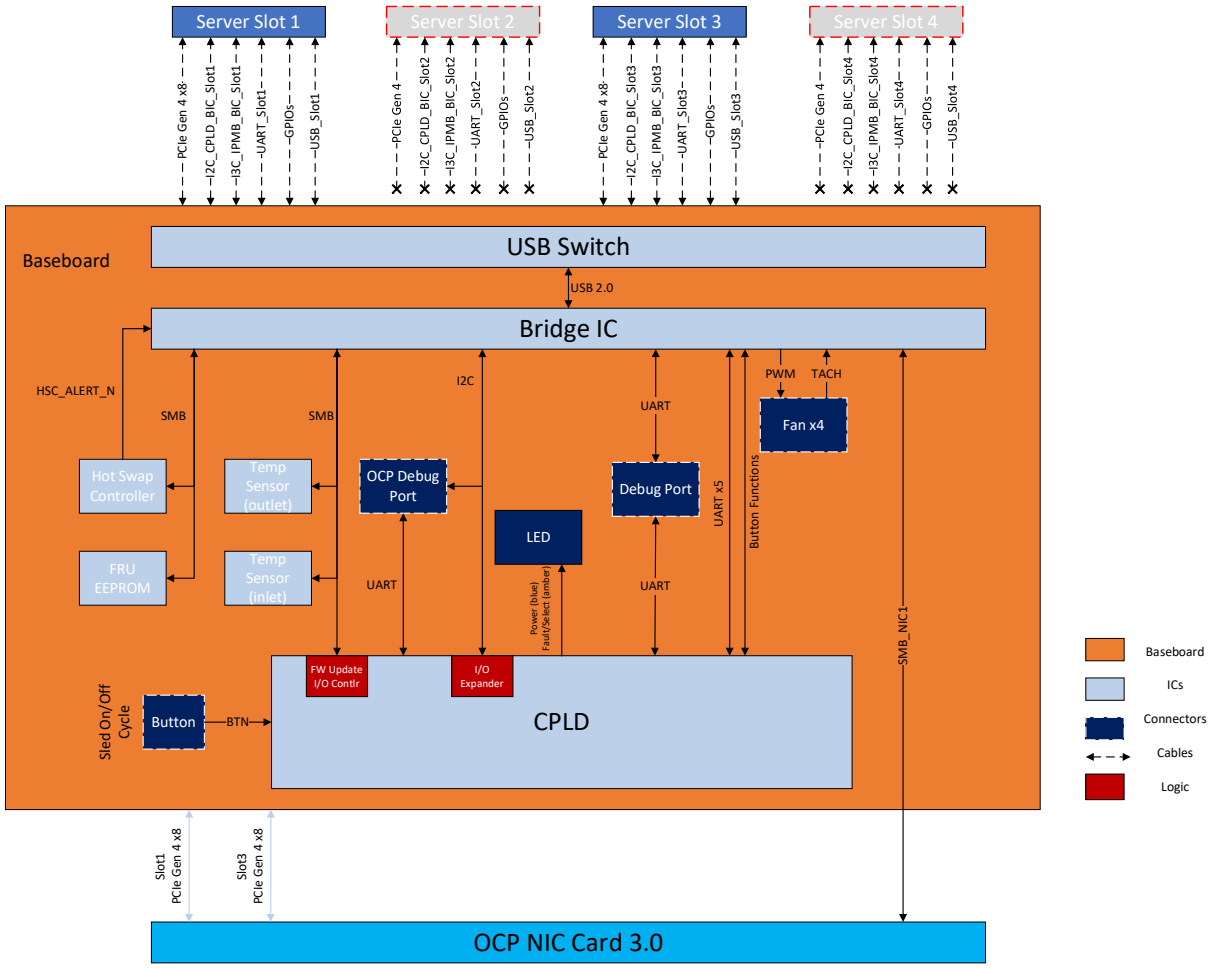


Figure 8-3 YV3.5 Class 2 Baseboard

The PCB design on Class 1 and Class 2 baseboards is the same. Functionality differences come from different BOM stuff options.

8.2 Yosemite V3.5 Subsystems

This section describes the various subsystems of the Yosemite V3.5 platform.

Input Power Delivery

For the Yosemite V3.5 platform power is delivered from the Open Rack V2 bus bar to each subsystem through three segments:

- Medusa cable
- Medusa power board (MPB)
- Vertical power distribution board (VPDB)

These elements of the power distribution system are shown in Figure 8-4. These power distribution components should not adversely impact the voltage that is delivered to the platform integrated baseboard and server blades.

The Medusa power board has a voltage sensor to monitor the input to the sled. The board also includes an eFuse device to protect the system against overcurrent faults.

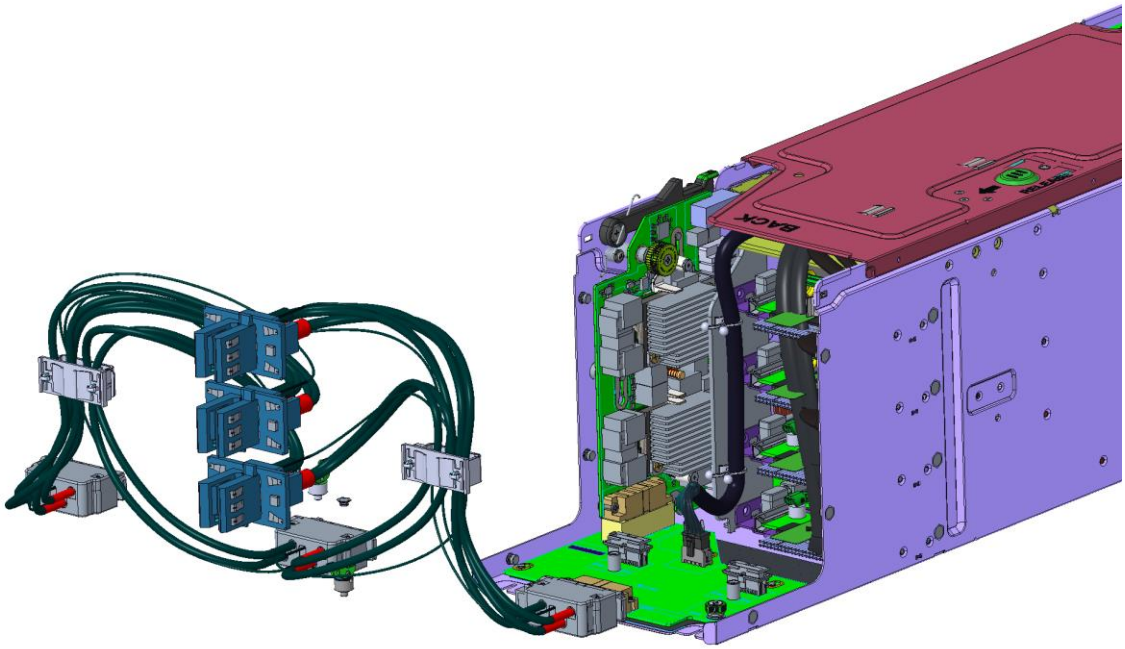


Figure 8-4: Input Power Delivery Topology

Baseboard

The baseboard is the management board of the platform. The BMC on the board serves the following functions:

- OOB access for remote management
- Power management through power monitoring and power control to servers
- Thermal management through temperature sensing and fan speed control
- LED indication and button control
- System event logger
- Responsible for firmware validation and conducting updates.
- FRU data tracking for the platform

The baseboard has a slot for the OCP 3.0 NIC card that supports various OCP NIC 3.0 cards with different PCIe lane configurations. The BMC obtains its network access thru the NIC card via NCSI connection.

NIC Card Slot

The baseboard has a 4C+ and 4C connectors that are intended for OCP NIC installation which are shown in Figure 8-5. Included in these connections are the side band interface of OCP NIC 3.0 as well as 4 x4 PCIe connectors to allow connection from servers to NIC card.

Only small form factor (SFF) OCP NICs (Network Interface Card) are in scope for YV3.5. Additional connectors are present to reserve routing and placement for future expandability. Depending on server's allocation and PCIe cabling arrangements, the design supports x4 or x8 PCIe connections from the server blades to the multi host NIC.

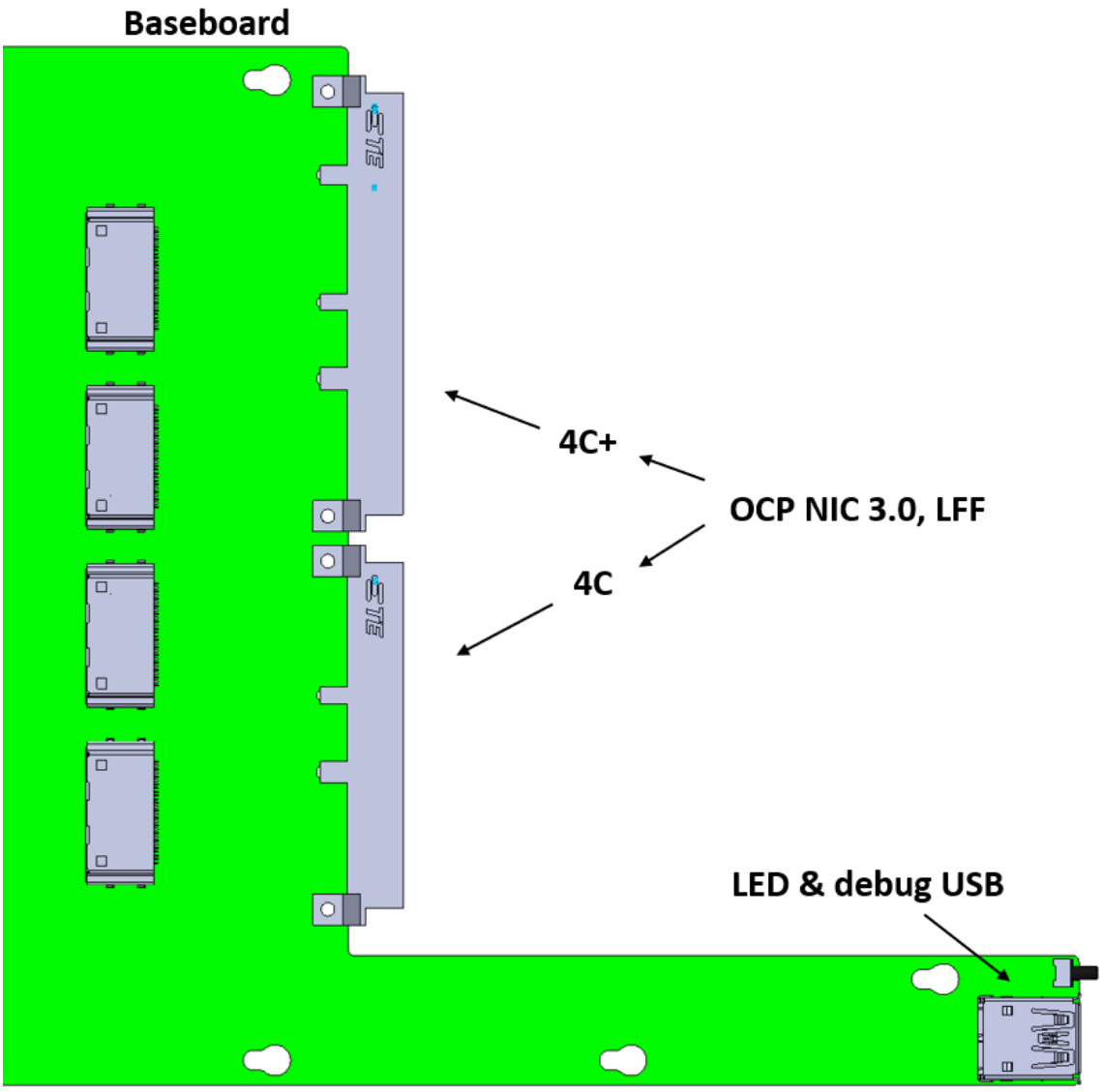


Figure 8-5: 4C and 4C+ Connectors

The configuration of PCIe lanes routed from the server blades to the baseboard is managed by the cable connections within the sled. The PCIe, PCIe refclk, PERST#, and other management signals are sent over the cables while power is provided by the baseboard. The two primary cable configurations are detailed below.

The four x4 configuration has the sled configured with four blades, and each blade has a x4 PCIe connection to the multihost NIC as shown in Figure 8-6.

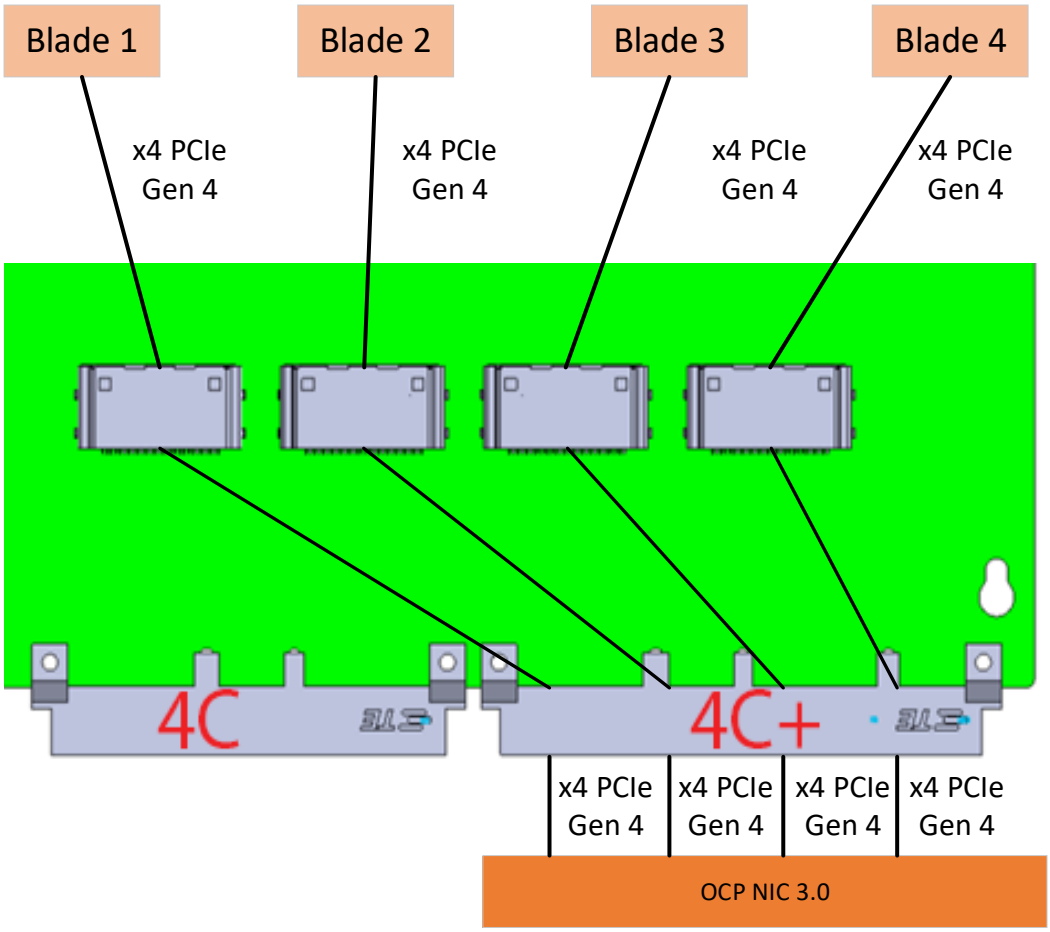


Figure 8-6: Baseboard Cabling with Four x4 Connections

The dual x8 configuration has the sled configured with two blades, and each blade has a x8 PCIe connection to the mulithost NIC as shown in Figure 8-7.

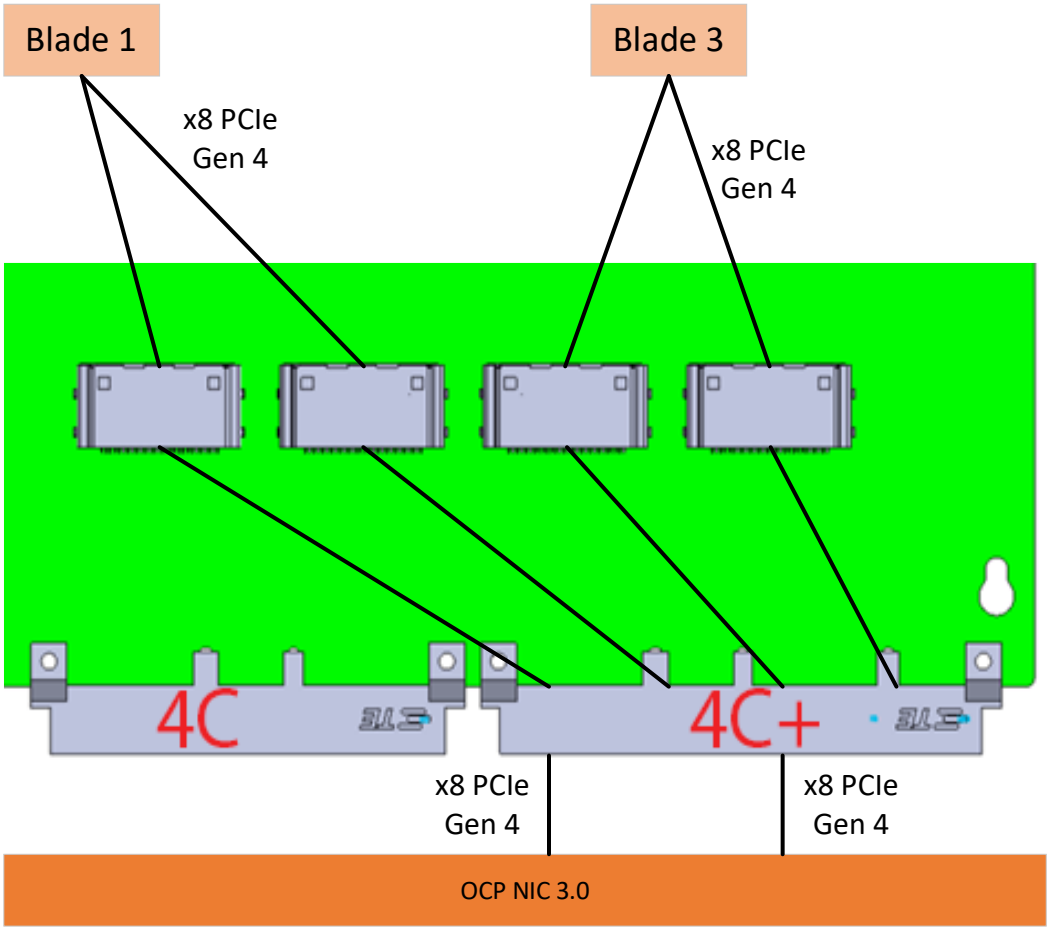


Figure 8-7: Baseboard Cabling with two x8 connections

The single x16 configuration has the sled configured with a single blade with a x16 PCIe connection to the single host NIC as shown in Figure 8-8.

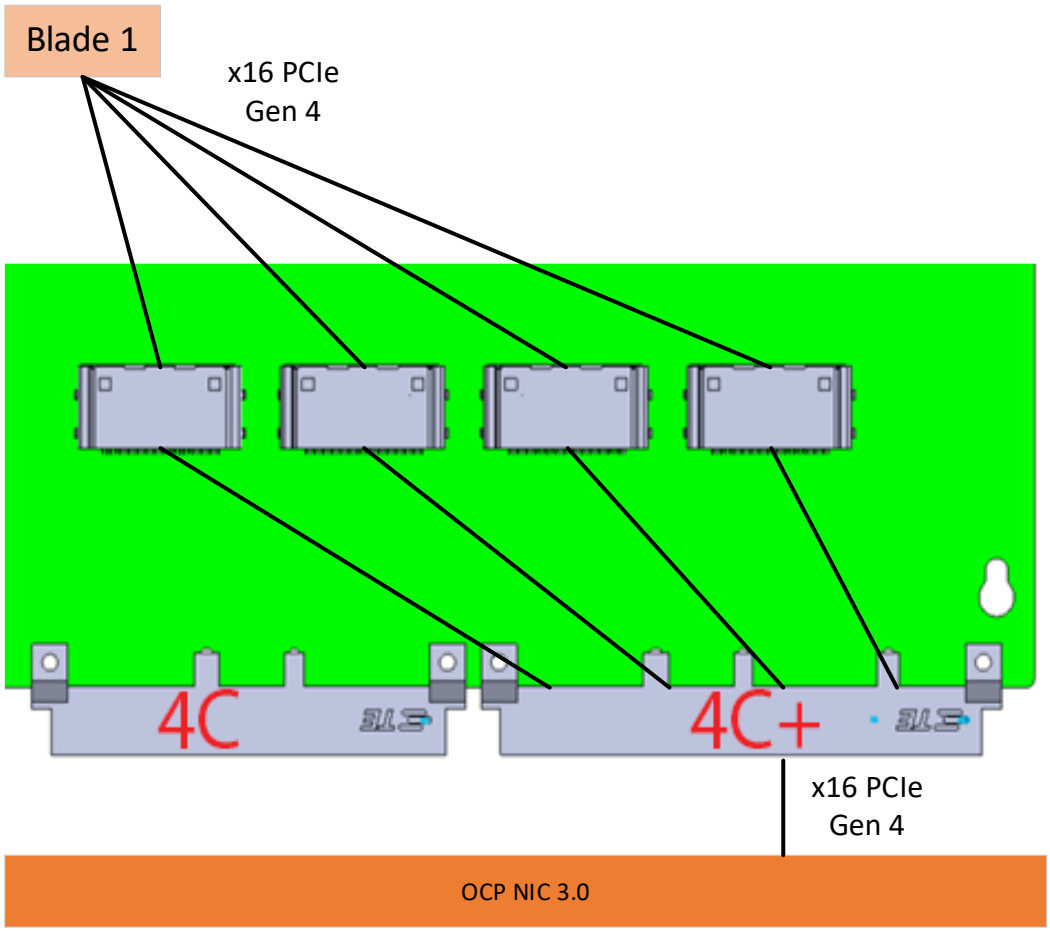


Figure 8-8: Baseboard Cabling with one x16 connection

Other PCIe routing options to the multihost NIC are possible with changes to the cabling. These are being explored for future use cases but are currently not within the scope of this document.

Sled management cable

The sled management cable between the server blades and the baseboard includes low speed signals such as I3Cs, alerts, power enable and fault signals that are used for platform management. The cable also includes high speed PCIe lanes to provide the server blade a connection to the mulithost NIC. This is represented in Figure 8-9.

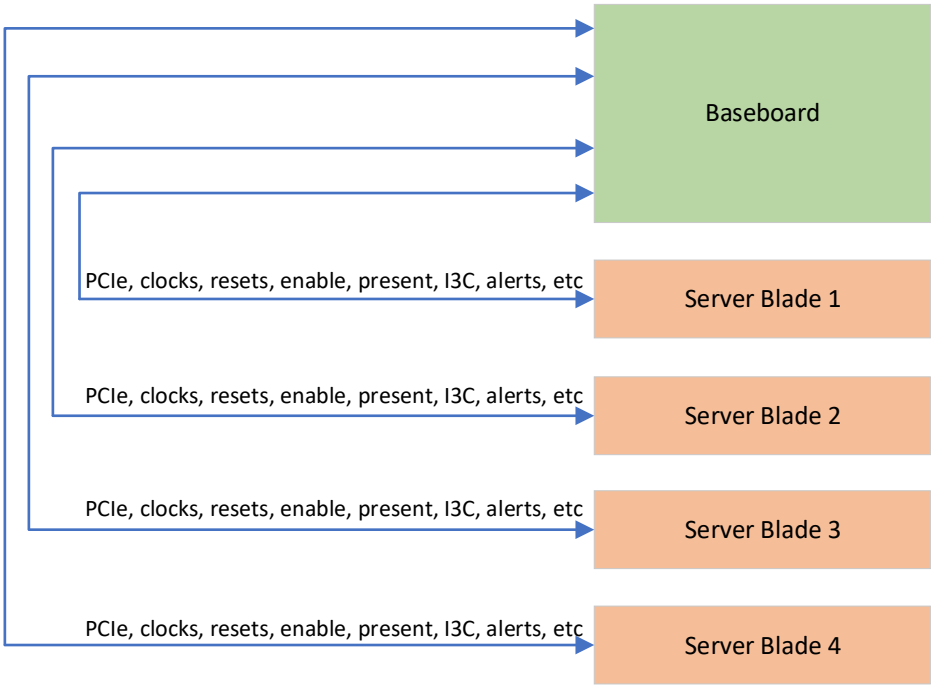


Figure 8-9: Management Cable Signals

Server Blade

The server blade contains the CPU subsystem. There are no CPU-specific features required for a blade to operate in YV3.5. Blades can be hot swapped from the system. It also has the following features:

1. Power is delivered to the server module from the vertical power distribution board through a dedicated power connector.
2. Depending on the available PCIe signals, the board may have x4 to x8 lanes connected to the Multi-Host (MH) NIC.
3. On board Bridge IC (BIC) and CPLD to handle:
    a. power sequencing
    b. thermal and power monitoring and reporting to BMC
    c. system error/event monitoring and reporting to the BMC on the baseboard
    d. configuration/programming of all programmable devices on board (BIOS, VR firmware)
    e. any other sideband communications that exists between the CPU, CPU's internal management engine, and the BMC
    f. providing boot config information for the BIOS when the system is booting up by determining the expansion systems connecting to the server
    g. User panel indicators like switch buttons and LEDs
    h. allowing BMC control with regards to power/thermal
    i. sideband and slow speed signals connect to the platform to allow BMC status report and control.
    j. reset functionality to the BMC in case of a BMC hang
    k. BIOS firmware root of trust and signing verification

The server module may have a front expansion board and/or a 2U expansion board. Various expansion board options are possible to support a variety of use cases. Figure 8-10 shows how a server board would look. Different server boards may take different shapes and sizes based on their needs.

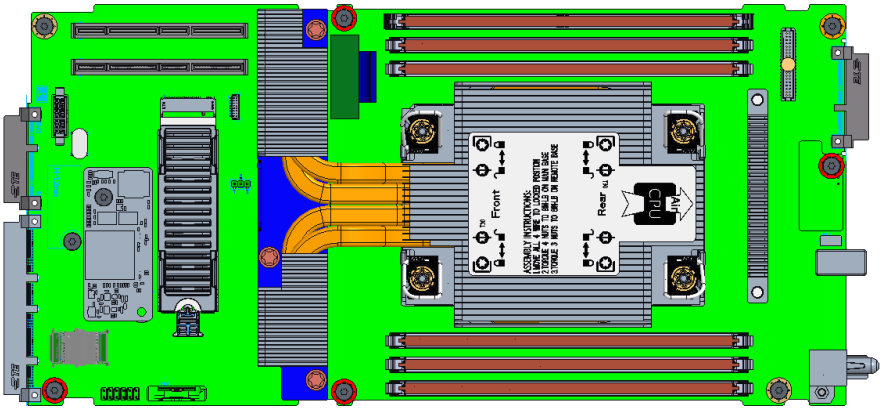


Figure 8-10: Example Server Blade

When using the 2U expansion board, the server blade will take up 2 chassis slots to enable an enhanced cooling solution.

Expansion Options

Yosemite V3.5 platform enables several expansion board options. The expansion board enables YV3.5 to support multiple config Types (T3, T8, T15, T17, etc.). The details are captured in the “Crater Lake 1S Server Expansion Design Specification”. Each expansion board has an expansion BIC. BOARD_ID pins are used to indicate the expansion board type. Note T1/T3/T8/T10 is the current focus, but any necessary design hooks should be included to support other configs.

BOARD_ID
BOARD_ID[3:0]
0000------->Crater Lake Class1
0001-------> Crater Lake Class2
0111------->BMC on Baseboard
1001 -------> NIC Expansion Card
1011 -------> 1U Expansion M.2 Card
1100 -------> 2U Expansion W/O SW
1110 -------> 1U Expansion with EDSFF
1101 -------> 2U Expansion with SW
1111 -------> BIC on Baseboard

8.3 Yosemite V3.5 Platform Power Delivery

Shown in Figure 8-11, is a high-level illustration of the power delivery topology that the Yosemite V3.5 platform implements.

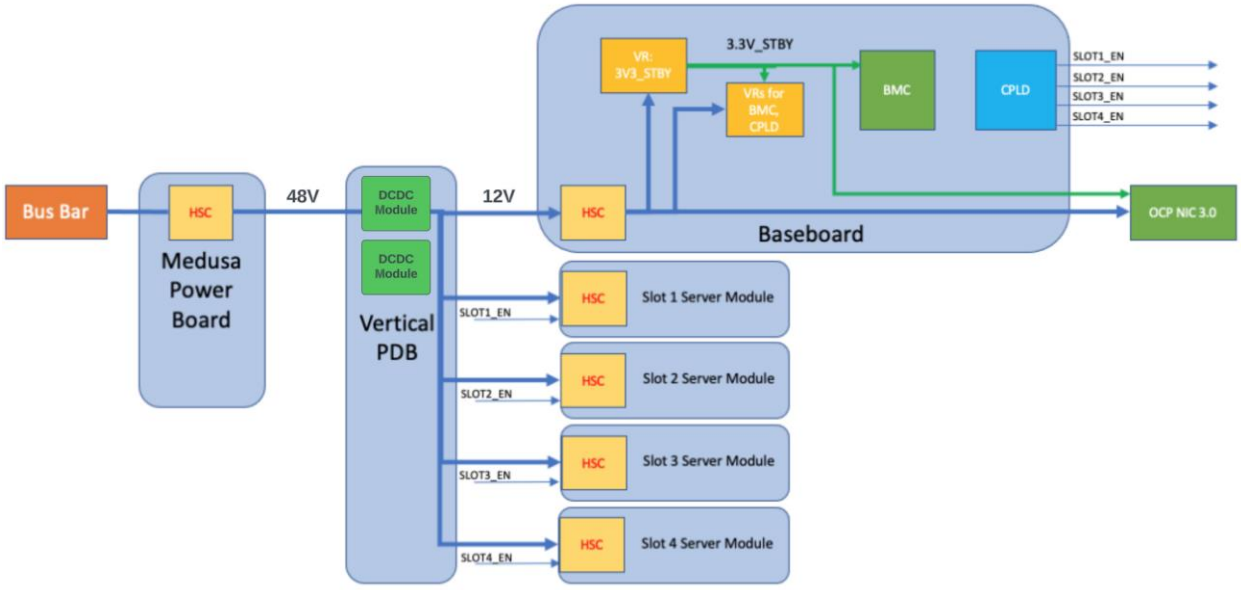


Figure 8-11: Yosemite V3.5 Platform Power Delivery Block Diagram

The Yosemite V3.5 platform uses a shared power distribution topology, as shown in Figure 8-12, where each subsystem draws power from the same input source. A hotswap controller (HSC) is included on each subsystem. This provides greater flexibility to support unique peak power delivery requirements depending on the intended use case. The baseboard's HSC remains responsible for delivery power to devices such as the fan module, BMC, logic circuitry, and OCP NIC 3.0 through the adapter board.

The Yosemite V3.5 platform is intended to support up to 1.6kW of distributed power between the different subsystems. Design consideration should be taken to ensure all aspects of the platform can be sufficiently cooled without exceeding limits based on the power delivery hardware which is defined in Section 9.2.

Upon insertion of the Yosemite V3.5 chassis, the baseboard immediately is energized from the input power delivery path. The baseboard is by design a non-hot swappable board and would be installed within the sled chassis together with the corresponding OCP NIC 3.0 card at the time of assembly. However, the OCP NIC 3.0 card could be cold swapped to enable different NIC options. The electrical design also supports hot swapping of the OCP NIC 3.0, but only with the intention of preventing system damage if it is inadvertently removed. System automates a reboot upon OCP3.0 NIC Hotswap to prevent malfunction.

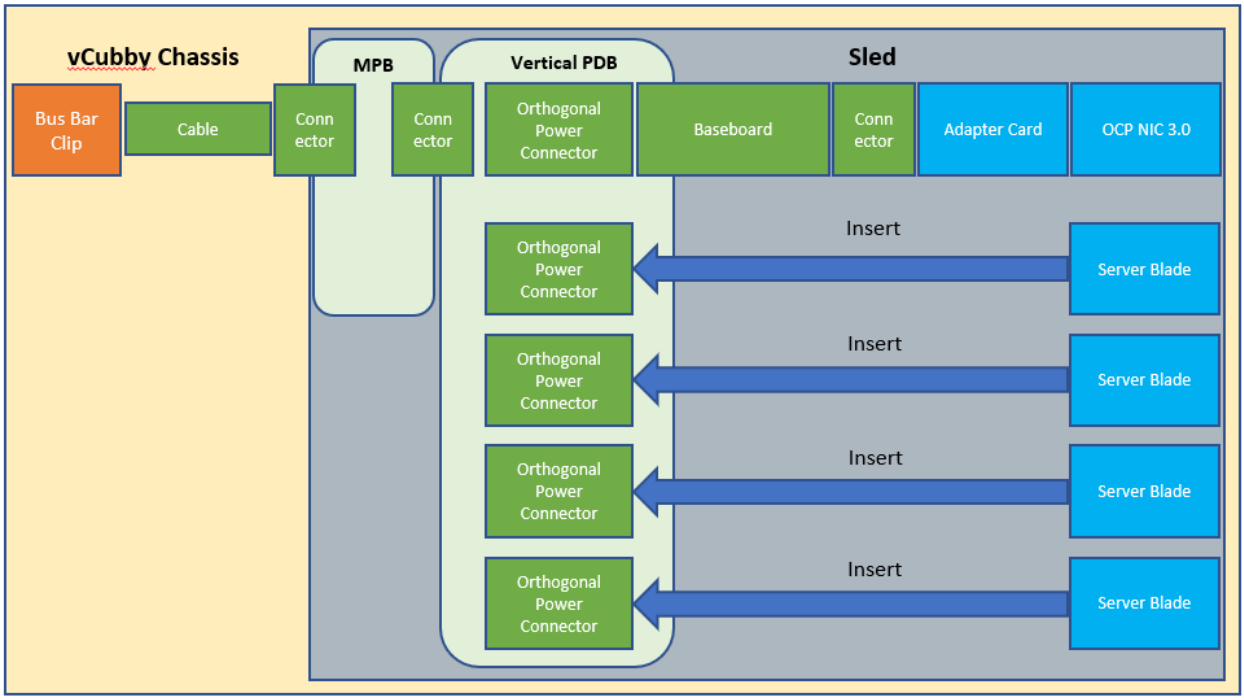


Figure 8-12: Yosemite V3.5 Power Interconnects

As for the server blades, the following power sequence takes place:

1. Upon insertion of a server blade, the card present pin will toggle from high to low to the BMC
2. After 1s delay, the BMC will assert the corresponding SLOTx_PWRON signal to the server blade's HSC which initializes the soft starting function
3. In parallel, the BMC will monitor whether the server blade's HSC pulls up the 12V_PGOOD signal to confirm power is being delivered to the blade

Every server blade is responsible for actively monitoring its own power consumption through the HSC. Either the CPU or Bridge IC of the server blade must query power telemetry and calculate a one second average based on the samples. The BMC must have access to this power sensor through the Bridge IC on the server module. As a development feature, the Bridge IC must be able to sample this power sensor within 10ms. In the event of any faults along the power delivery path, the HSC is expected to assert warning signals to both the BMC and server blades for necessary action.

Although the BMC is responsible for power management of the server blade upon insertion, power delivered to each server should not be interrupted in the event the BMC enters reset to ensure server operation is not impacted. Push buttons for each server module and baseboard is made available to allow operators to power cycle without the need of re-insertion of any blades or sled. Take note that a loss of power to the baseboard would create a loss of power to the entire sled as the power enable signals from the baseboard will be driven low. Design consideration must be made to avoid leakage paths during such a power state.

Depending on the power policy, the BMC enables power to server modules upon request. The BMC shall drive power-on signals as a power button function as defined in the Advanced Configuration and Power Interface (ACPI).

8.4 SMBus Block Diagram

The generic use of SMBUS around the system is to obtain the following:

1. FRU information of the devices
2. Power and temperature readings to understand how the system or subsystems (servers/NIC) are behaving
3. Out of band management of the device
4. Firmware updates

The BMC gains its network access thru the NC-SI interface on the OCP NIC 3.0. Detailed implementation of the server SMbus connections can be referred to in the dedicated server specification.

The BMC can access thermal sensors, the hot-swap controller, and the FRU via a separate SMBus, as shown in Figure 8-13 on the following page.

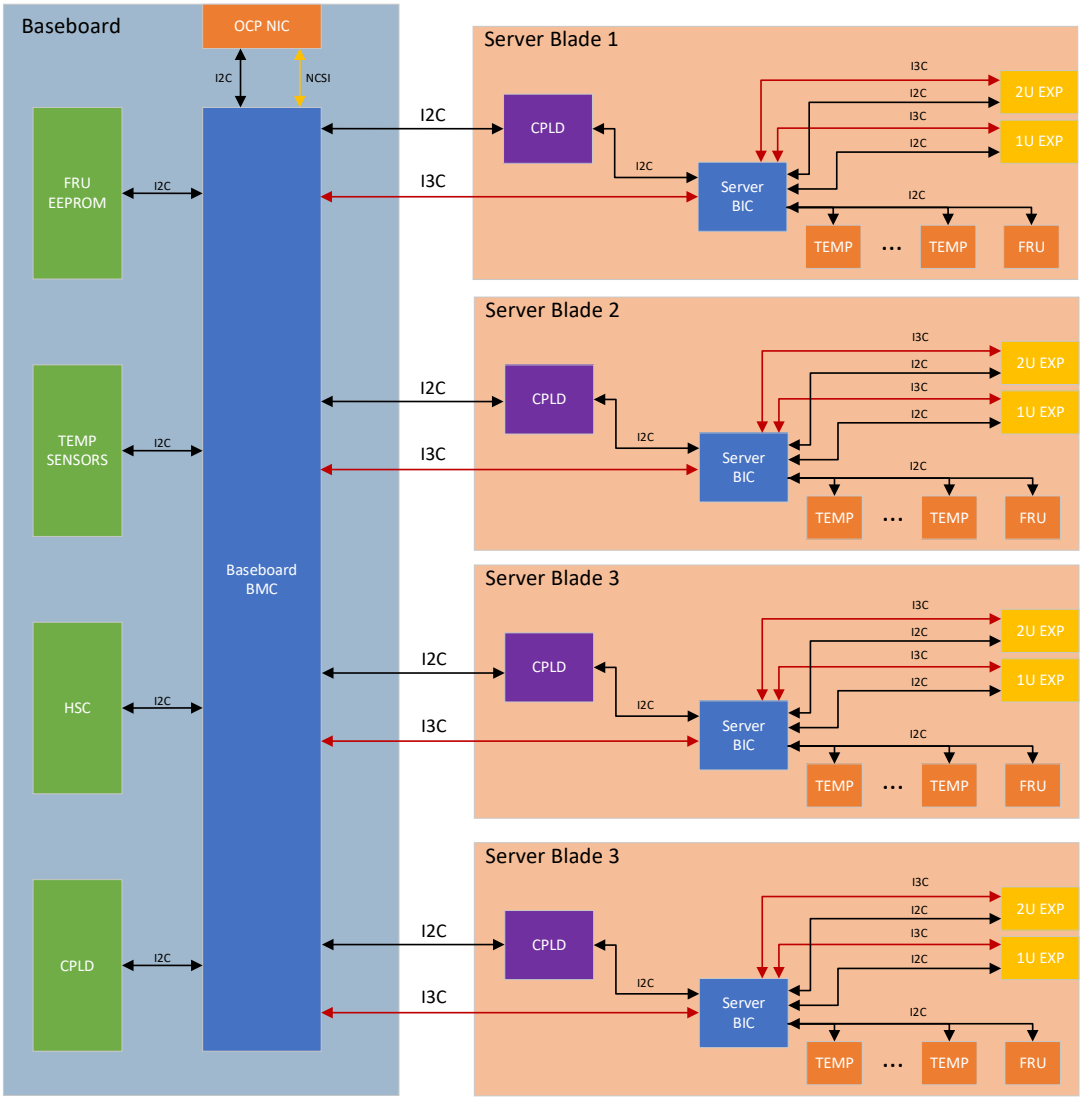


Figure 8-13: Yosemite V3.5 Platform SMBus Block Diagram

For each server, it is possible that expansion modules are connected to them. In such case, the connection from the baseboard to a server would reference to a connection, as shown in Figure 8-14, where BICs on the server board need to relay a message from the BMC to the BICs on the expansion modules.

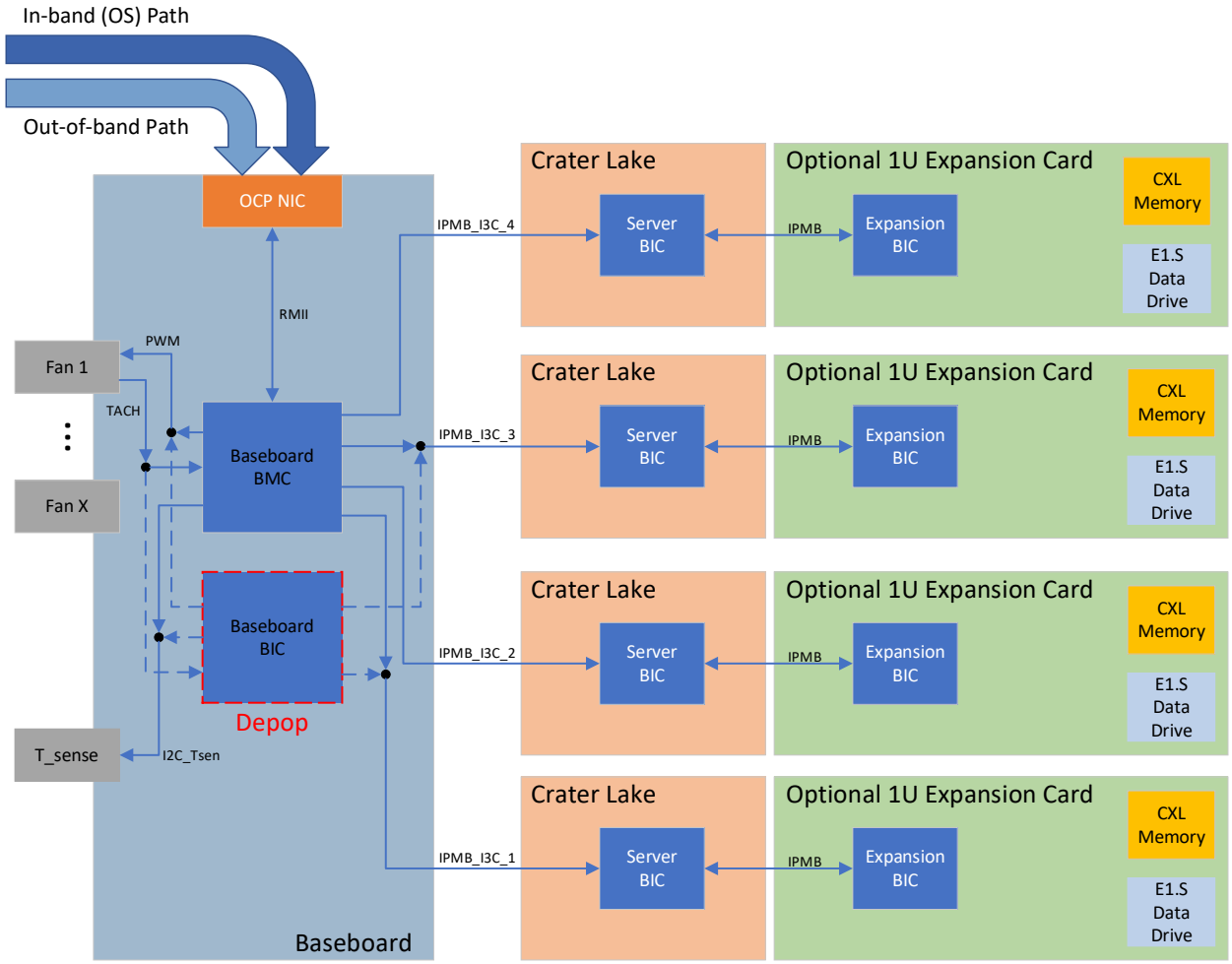


Figure 8-14: IPMB Connectivity with Expansion Cards

8.5 1S Server

Overview

The Yosemite V3.5 Platform has four slots per sled that can host four 1U 1S servers.

1S Server Connectors

The server blade will be powered through PowerBlade+ connector and connected to the system through SFF-TA1002 straddle mount connector. The specific server should have at least a 2C straddle mount connector available for connection.

Depending on if the system configuration has four 1U blades or two 2U blades, the Class 1 configuration may differ in how many PCIe lanes are connected.

- Class 1 routes all the management and sideband signals with at least 4 lanes of PCIe. The upper portion of the 2C connector (pins 29 – 42) may not be needed if the MHNIC connectivity is 4x4. If the configuration has a 2x8 MHNIC connectivity, then all signals must be connected.
- Class 2 systems do not necessarily need the PCIe lanes to the baseboard since each blade has a dedicated NIC. However, all of the management signals are required. Depending on cable costs it may make sense to use a common cable to connect the blades to the baseboard.

1S Server Slot Pinout Definition on Yosemite V3.5 Platform

The pin-out below is a preliminary set of signals for a x8 PCIe connection out of the server blade.

Table 8-1: Crater Lake 1S Server 2C connector to Baseboard

Crater Lake 1S 2C connector			
B1	I3C_BMC_BIC_SB_SDA	BB_BIC_READY	**A1**
B2	I3C_BMC_BIC_SB_SCL	SMB_BMC_CPLD_SDA	**A2**
B3	GND	SMB_BMC_CPLD_SCL	**A3**
B4	AC_ON_OFF_BTN_R	SMB_BMC_CPLD_ALT_N	**A4**
B5	FM_HSC_FAULT_N	GND	**A5**
B6	FM_BMC_HSC_EN_R	PWRBTN_R_N	**A6**
B7	PWRGD_SB_STBY_PWROK	RST_BMC_N	**A7**
B8	RST_PCIE_NIC_PERST_N	FM_BMC_RST_RTCRST_R	**A8**
B9	UART_REQ_R_N	SB_SLOT_ID1	**A9**
B10	GND	SB_SLOT_ID0	**A10**
B11	UART_CPLD_MUX_RXD	RSVD_2C_A11	**A11**

Crater Lake 1S 2C connector			
B12	UART_CPLD_MUX_R_TXD	PRSNT_SB_N	**A12**
B13	GND	GND	**A13**
B14	USB2_BMC_DN	CLK_100M_NIC_PCIE_CLK0_DN	**A14**
B15	USB2_BMC_DP	CLK_100M_NIC_PCIE_CLK0_DP	**A15**
B16	GND	GND	**A16**
B17	P5E_CPU_PE4_NIC_TX_C_DN_0	P5E_CPU_PE4_NIC_RX_DN_0	**A17**
B18	P5E_CPU_PE4_NIC_TX_C_DP_0	P5E_CPU_PE4_NIC_RX_DP_0	**A18**
B19	GND	GND	**A19**
B20	P5E_CPU_PE4_NIC_TX_C_DN_1	P5E_CPU_PE4_NIC_RX_DP_1	**A20**
B21	P5E_CPU_PE4_NIC_TX_C_DP_1	P5E_CPU_PE4_NIC_RX_DN_1	**A21**
B22	GND	GND	**A22**
B23	P5E_CPU_PE4_NIC_TX_C_DN_2	P5E_CPU_PE4_NIC_RX_DN_2	**A23**
B24	P5E_CPU_PE4_NIC_TX_C_DP_2	P5E_CPU_PE4_NIC_RX_DP_2	**A24**
B25	GND	GND	**A25**
B26	P5E_CPU_PE4_NIC_TX_C_DN_3	P5E_CPU_PE4_NIC_RX_DP_3	**A26**
B27	P5E_CPU_PE4_NIC_TX_C_DP_3	P5E_CPU_PE4_NIC_RX_DN_3	**A27**
B28	GND	GND	**A28**
B29	GND	GND	**A29**
B30	P5E_CPU_PE4_NIC_TX_C_DN_4	P5E_CPU_PE4_NIC_RX_DN_4	**A30**
B31	P5E_CPU_PE4_NIC_TX_C_DP_4	P5E_CPU_PE4_NIC_RX_DP_4	**A31**
B32	GND	GND	**A32**
B33	P5E_CPU_PE4_NIC_TX_C_DN_5	P5E_CPU_PE4_NIC_RX_DP_5	**A33**
B34	P5E_CPU_PE4_NIC_TX_C_DP_5	P5E_CPU_PE4_NIC_RX_DN_5	**A34**
B35	GND	GND	**A35**
B36	P5E_CPU_PE4_NIC_TX_C_DN_6	P5E_CPU_PE4_NIC_RX_DN_6	**A36**
B37	P5E_CPU_PE4_NIC_TX_C_DP_6	P5E_CPU_PE4_NIC_RX_DP_6	**A37**
B38	GND	GND	**A38**
B39	P5E_CPU_PE4_NIC_TX_C_DN_7	P5E_CPU_PE4_NIC_RX_DP_7	**A39**
B40	P5E_CPU_PE4_NIC_TX_C_DP_7	P5E_CPU_PE4_NIC_RX_DN_7	**A40**
B41	GND	GND	**A41**
B42	GND	GND	**A42**

Unless otherwise noted, all signal directions are from the perspective of the devices on the server modules.

Table 8-2: Pin Description of Crater Lake <-> Baseboard interface (Class 1)

YV3.5 Baseboard Signals (Class 1)		
Signal Name	**Direction**	**Description (Class 1 BMC on Baseboard)**
I3C_IPMB_SDA	I/O	IPMB between Baseboard management entity (BMC) to BIC on 1S Server Card
I3C_IPMB_SCL	I/O	IPMB between Baseboard management entity (BMC) to BIC on 1S Server Card
SMB_BMC_CPLD_SDA	I/O	400KHz I2C from Baseboard to CPLD+BIC for commands
SMB_BMC_CPLD_SCL	I/O	400KHz I2C from Baseboard to CPLD+BIC for commands
SMB_BMC_CPLD_ALT_N	Output	Alert signal to baseboard, active low OD signal with PU at Baseboard
MB_PRSNT_N	Output	Present signal. Active low. 1S Server card to place 100 ohm to GND
FM_BMC_HSC_EN_R	Input	Signal to enable Server Card HSC Active High push pull. PD at 1S Server Card side Baseboard assert HSC_EN when: 1) 1S Server Card is fully inserted
FM_HSC_FAULT_N	Output	HSC Fault signal; low active; OD with PU on baseboard.
UART_CPLD_MUX_RXD	Input	UART input to 1S Server Card - Source is CPLD on Baseboard - Destination is CPLD on 1S Server Card
UART_CPLD_MUX_R_TXD	Output	UART TX from Server Card - Source is CPLD on 1S Server Card - Destination is CPLD on Baseboard
RST_PCIE_NIC_PERST_N	Output	PCIe reset from Server Card to Baseboard Low active , Push-Pull, 3.3V_STBY domain
PWRGD_SB_STBY_PWROK	Output	Standby Power OK of Server Card High active, push-pull
PWRBTN_R_N	Input	From Baseboard CPLD to CPU card CPLD and BIC to initiate a DC power cycle.
RST_BMC_N	Output	Signal from CPU card BIC to reset BMC on Baseboard

YV3.5 Baseboard Signals (Class 1)		
AC_ON_OFF_BTN_R	Output	AC button on the server module
RSVDx	N/A	Spare signals between server blade and baseboard
CLK_100M_NIC_PCIE_CLK0_DN/DP	Output	Server output of 100MHz clock; 1 clock output
PxE_*_TX*(n/p)X	Output	PCIe Gen5 RX of Server module. Class 1: X ranges from 0 to 3 Class 2: X ranges from 0 to 7
PxE_*_RX* (n/p)X	Input	PCIe Gen5 RX of Server module. Class 1: X ranges from 0 to 3 Class 2: X ranges from 0 to 7
UART_REQ_N	N/A	NC
BB_BIC_READY	Input	Indicates to CPU card CPLD that BMC is ready on Baseboard
USB2_BMC_DN/DP	I/O	USB interface between the baseboard and server Class 1: BMC on Baseboard is host Class 2: Sever blade is host (e.g. NIC expansion BMC)
FM_BMC_RST_RTCRST_R	output	0: CMOS clear 1: NA
SB_SLOT_IDx	output	Server board Slot ID strapping (1-4 server slot)
PRSNT_SB_N	Output	GNDed on Server board to indicate cable presence to Baseboard CPLD. Active Low

8.6 Network Options

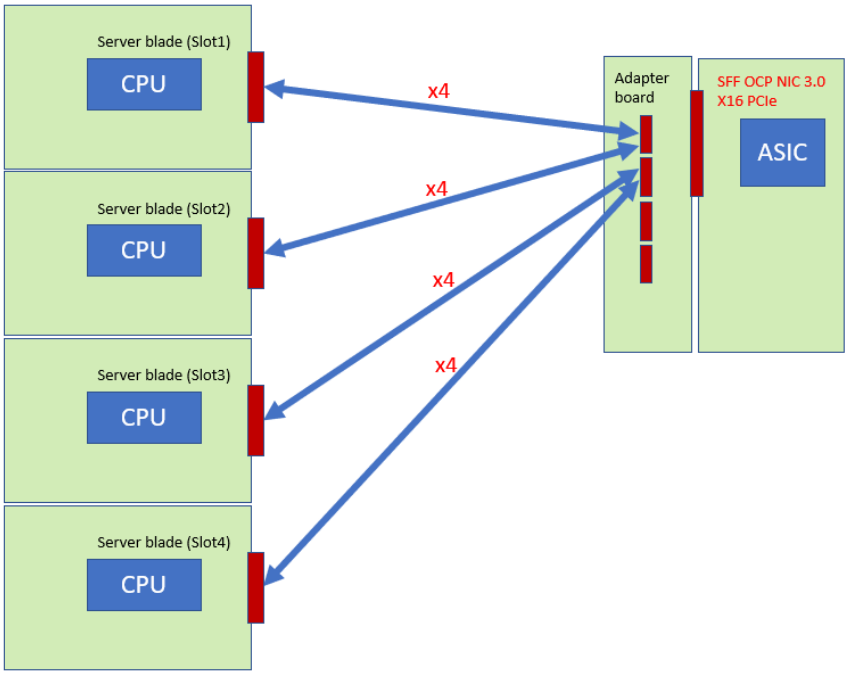
Each sled in YV3.5 can have an OCP NIC 3.0. Depending on the connectivity needs, different cable configurations are supported. The following sections illustrate some of the different NIC cabling options.

The PCIe links can operate up to Gen 4 speeds using the existing baseboard design. Server blade designer may make trade-offs in whether to use ultra-low-loss material in extending the connection reach or choose to use a re-driver/re-timer solution, but at the expense of power and space to achieve Gen5 capabilities.

The BMC on the baseboard connects to the network through the NCSI interface to the OCP NIC 3.0.

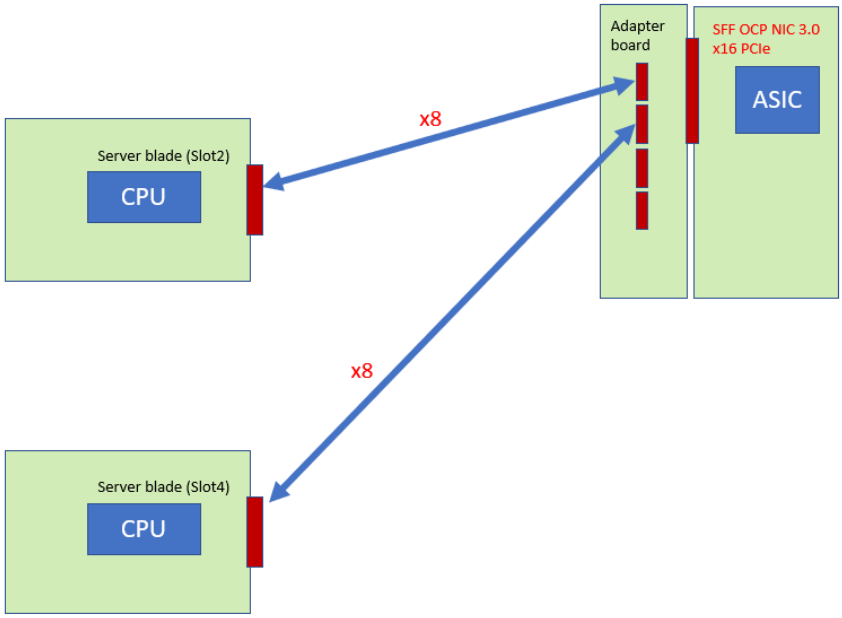
Option 1 (x4 PCIe lanes from Server blade to x16 SFF OCP NIC 3.0 card)

This configuration makes use of a cable set with a custom pin-out that allows 4 lanes from each of the server cards to connect to the x16 SFF OCP NIC 3.0 card.



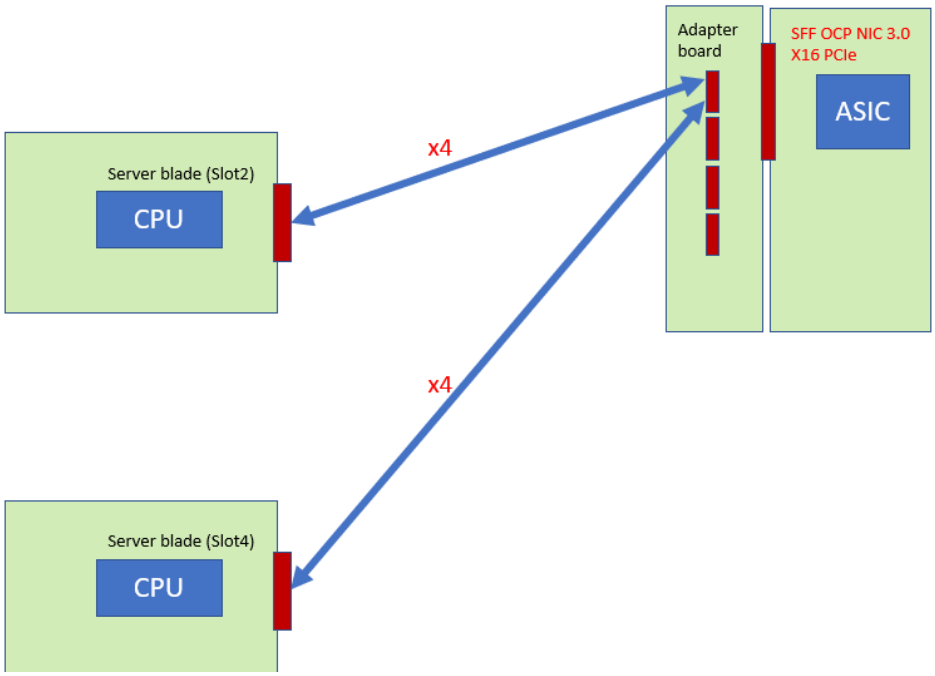
Option 2 (x8 PCIe lanes from Server blade to x16 OCP SFF NIC 3.0 card)

This configuration is used where server blades are occupying two slots due to its 2U expansion card.



Option 3 (x4 PCIe lanes from Server blade to x16 OCP SFF NIC 3.0 card)

This configuration is similar to Option 2, but each blade has only x4 PCIe lanes connecting to the adapter card.



9 Thermal

To meet thermal reliability requirement, the thermal and cooling solution should be designed to dissipate heat from the components when the system is operating at its maximum TDP (Thermal Design Power). The thermal solution should be found by setting a high-power target for initial design in order to avoid redesign of cooling solution; however, the final thermal solution of the system should be most optimized and energy efficient under data center environmental conditions with the lowest capital and operating costs. The thermal solution should not allow any overheating issue for any components in the system under the following environmental conditions -

1. Inlet temperature lower than or equal to 35°C, and 0-inch H2O datacenter pressure with all FANs in each thermal zone running properly.
2. Inlet temperature lower than or equal to 35°C, and 0.001-inch H2O datacenter pressure with one FAN (or one rotor) in each thermal zone having failed

9.1 Data Center Environmental Conditions

This section outlines Meta data center operational conditions.

Location of Data Center/Altitude

Maximum altitude is 6,000 ft above sea level. Any variation of air properties or environmental difference due to the high altitude needs to be deliberated into the thermal design.

Cold-Aisle Temperature

Data centers generally maintain cold aisle temperatures between 18°C and 30°C (65°F to 85°F). The mean temperature in the cold aisle is usually 24°C with 3°C standard deviation. The cold aisle temperature in a data center may fluctuate minutely depending on the outside air temperature. Every component must be cooled and maintain a temperature below its maximum specification temperature in the cold aisle.

Cold-Aisle Pressurization

Data centers generally maintain cold aisle pressure between 0 inches H_2O and 0.005 inches H_2O. The thermal solution of the system should consider the worst operational pressurization possible, which generally is 0 inches H_2O and 0.001 inches H_2O with a single fan (or rotor) failure.

Relative Humidity

Data centers usually maintain a relative humidity between 10% and 90%. The thermal solution must sustain uninterrupted operation of the system across the aforementioned RH range.

9.2 Server Operational Conditions

Inlet Temperature

The inlet air temperature will vary. The cooling system in the Yosemite V3.5 Platform should be able to cover inlet temperatures including 20°C, 25°C, 30°C, and 35°C. Cooling above 30°C is beyond the Meta operational condition but used during validation to demonstrate the thermal reliability and design margin. Any degraded performance is not allowed over the validation range 0°C-35°C.

Pressurization

Except for the condition when one rotor or one fan in a server fan fails, the design of thermal solution should not consider extra airflow available from data center cooling fans due to positive pressurization. If and only if one rotor or one fan in a server fan fails, can the negative or positive DC pressurization be considered in the design of thermal solution in the hot aisle or the cold aisle, respectively. The maximum pressurization on the inlet side is 0.005 inches H_2O.

Fan Redundancy

The server fans at N+1 rotor redundancy should be sufficient for cooling server components to temperatures below their maximum specification to prevent server shut down or to prevent

either CPU or memory throttling. An N+1 rotor redundancy in the Yosemite V3.5 Platform is preferred when the system is operating under normal conditions.

Delta T

The Delta T is the air temperature difference across the system, or the temperature difference between the outlet air temperature and the inlet air temperature. The Delta T must be greater than 13.9°C (25°F) at the rack level when the server is running within the data center operational condition. To meet the rack level delta T target, the desired server level Delta T is greater than 17°C (31°F) when the inlet air to the system is equal to or lower than 30 °C.

System Airflow or Volumetric Flow

The unit of airflow (or volumetric flow) used for this spec is cubic feet per minute (CFM). Airflow (CFM) can be used to determine the thermal expenditure or to calculate the approximate delta T of the system. The thermal expenditure is quantified by the metric CFM/W, which is calculated by the following formula:

$$Thermal\ Expenditure\ = \frac{System\ Airflow}{Total\ System\ Power\ Consumption\ (including\ fans)}\left[\frac{CFM}{[W]}\right]$$

At sea level, the maximum allowable airflow per watt in a Yosemite V3.5 rack is 0.13 at 30 °C inlet temperature under the normal load. The cooling solution at the system level should meet airflow per watt target of 0.1 to account for airflow consumption of the TOR and PSU systems (~20% reduction).

As resource permits, to understand the interaction between the systems and evaluate the performance of in-rack containment, rack-level airflow testing is recommended to ensure rack level CFM/W is meeting data center operational condition.

Thermal Margin

The thermal margin is the difference between the maximum theoretical safe temperature and the actual temperature. Unless specified, the system should operate at an inlet temperature of 35°C (95°F) outside of the system with a minimum 4% thermal margin or 7% thermal margin for inlet temperatures up to 30°C (86°F).

System Loading

The power consumption of individual components in the system motherboard as well as the entire Yosemite V3.5 platform may vary by use. Please see the summary below.

- System loading: idle to 100%
- OCP NIC 3.0: SFF (25W)

A unified thermal solution that can cover up to 100% system loading is preferred. However, an original design manufacturer (ODM) can propose a non-unified thermal solution if there is an alternative way to provide cost benefits.

Fan Speed Controller

The fan speed controller (FSC) must be optimized to provide the necessary cooling for all key components while aiming to maximize thermal efficiency or minimize the CFM/W. The FSC may be a combination of linear, non-linear and PID control and must be set based on sensor readings for all key components. The overshoot of temperature should be minimized and must be less than 2°C from the stabilized temperature. The FSC must be able to maintain the thermal margin for all components while operating within the data center environmental conditions. If required, the FSC must have separate FSC tables to accommodate 0 ft, 3,000 ft and 6,000 ft elevations.

9.3 Thermal Kit Requirements

Thermal testing must be performed at up to 35°C (95°F) inlet temperature to guarantee reliability at high temperatures.

Heat Sinks

Heat sinks must have a thermally optimized design at the lowest cost. Installation must be simple and uncomplicated. Heat sinks must not block debug headers or connectors. Heat sink design for 1U use case and 2U use case can be different to maximize the thermal efficiency.

System Fan

The system fan must be highly power-efficient with dual bearings. The propagation of vibration caused by fan rotation should be minimized and limited. The minimum frame size of a fan is 60mm × 60mm and the maximum frame size is 80mm × 80mm. An ODM can propose a larger frame size than 80mm × 80mm if, and only if, there is an alternative way to provide cost benefits. The maximum fan thickness should be less than 56mm. Each rotor in the fan should have a maximum of five wires. Except for the condition when one fan (or one rotor) fails, the fan power consumption in system should not exceed 5% of total system power, excluding the fan power.

System fans should not have backrush currents in all conditions. System fans should have an inrush current of less than 1A on a 12.5V per fan. When there is a step change on the fan PWM signal from low PWM to high PWM, there should be less than 10% of overshoot or no overshoot for the fan input current. The system should stay within its power envelope per Open Rack V1/V2 power specification in all conditions.

Temperature and Power Sensors

The sensors should be accurate to +/-2°C and desired to be within 2% tolerance across the whole operation temperature range. The goal sensor accuracy is +/-1°C.

10 I/O System

This section describes the Yosemite V3.5 Platform's I/O requirements.

10.1 Front facing Server Modules

The Yosemite V3.5 Platform has a 4-slot system for server modules. It is possible that for certain configurations, one module may occupy two slots.

10.2 Network

Data Network

The Yosemite V3.5 Platform uses an OCP NIC 3.0 at the front panel as its primary data network interface. Please refer to Network Options in Section 8.6 for more details.

Management Network

The management network on the Yosemite V3.5 Platform uses the sideband of the network controller of the data network, either SMBus or NC-SI interface. Please refer to Section 8.2.4 for more details.

10.3 Server Slots Assignment

The server slot assignment for YV3.5 is as shown below, at the time of writing. There could be a server blade occupying 2 "slots" as shown on the left when it has a 2U expansion module, or when system with 4 servers.





10.4 Front Panel

On the adapter card of a Yosemite V3.5 sled, there is a power button, an OCP debug USB connector, and LEDs for the system.

Power Button and LEDs

Please refer to System LEDs and Buttons in Section 15.11.

OCP Debug USB Connector

The Yosemite V3.5 Platform has one OCP debug USB connector located at the front panel of the baseboard. The connector supports the OCP USB3.0 debug card. Information regards to the relevant host and their POST codes/SEL/config are displayed thru the LCD of the OCP USB3 debug card. The select button on the card allows the users to walk through different hosts for their information. The USB port on the debug card is purely meant for BMC. User may insert a thumb drive here to allow data transfer to/from USB etc.

POST Codes

During POST, the BIOS should output POST codes onto the OCP debug card through Bridge IC and the BMC. When a SOL session is available during POST, the remote console should show the POST code.

During the boot sequence, the BIOS shall initialize and test each DIMM module. If a module fails to initialize or fails the BIOS test, the following POST codes should flash on the debug card to indicate which DIMM has failed.

Table 10-1 shows an example of displaying DIMM failure modes on the server modules. Individual server modules will define how codes are being reflected in their respective case.

Table 10-1: DIMM Error Code Table

	Code	Result
CPU (Channel 0 ~ 3)	A0	Channel 0 DIMM 0 (Upper furthest) Failure
	A1	Channel 0 DIMM 1 Failure
	B0	Channel 1 DIMM 0 Failure
	B1	Channel 1 DIMM 1 (Upper closest) Failure
	C0	Channel 2 DIMM 0 (Lower furthest) Failure
	C1	Channel 2 DIMM 1 Failure
	D0	Channel 3 DIMM 0 Failure
	D1	Channel 3 DIMM 1 (Lower closest) Failure

The first hex character indicates the channel of the DIMM module. The second hex character indicates the number of the DIMM module. The POST code will also display the error major code and minor code from the memory reference code. The display sequence will be “00”, DIMM location, Major code and Minor code with a one-second delay for every code displayed. The BIOS shall repeat the display sequence indefinitely. The DIMM number count starts at the furthest DIMM from the CPU.

10.5 Fan Module Connector

Every fan has its own PWM input to control the fan speed and tachometer output so that the BMC can measure the fan speed. All fans are powered by the system's 12V power supply and should be on at full speed before the BMC can control it. The fan module connector is designed to support 10k, 15k and 10k dual rotor fans. The fan cables from BMC is designed to deliver the power for the highest current (15k fan) and control signals for the fan. Signals can be seen in Table 10-2 and Table 10-3

Table 10-2: Fan Connector Pin Definition

Pin	Description	Netname	Pin
2	FAN_PWM_1_BUF_R	FAN_PWM_0	**1**
4	FAN_TACH_2_R	FAN_TACH_0_R	**3**
6	FAN_TACH_1_R	FAN_TACH_3_R	**5**
8	GND	GND	**7**
10	SMB_MEDUSA_SDA	GND	**9**
12	SMB_MEDUSA_SCL	GND	**11**
14	GND	GND	**13**
16	DUAL_FAN0_DETECT_N	DUAL_FAN1_DETECT_N	**15**
18	FAN_TACH_7_R	FAN_TACH_5_R	**17**
20	FAN_TACH_4_R	FAN_TACH_6_R	**19**
22	FAN_PWM_2_BUF_R	FAN_PWM_3_BUF_R	**21**

Table 10-3: Fan Power Connector Pin Definition

Pin	Description	Netname	Pin
5	P12V_FAN3	GND	**1**
6	P12V_FAN2	GND	**2**
7	P12V_FAN1	GND	**3**
8	P12V_FAN0	GND	**4**

11 Mechanical

The Yosemite V3.5 Platform is an Open Rack V2 & Open Rack V3 compatible compute platform which consists of a Yosemite V3.5 chassis, which supports three Yosemite V3.5 sleds in a 4OU space. Each Yosemite V3.5 sled can hold up to 4 server blades.

11.1 Yosemite V3.5 Chassis

Yosemite V3.5 chassis is a power-mechanical chassis distributing power from the rack bus bars to the three sleds. Yv3.5 has two types of chassis – one for ORv2(12V) and one for ORv3(48V), both can be seen in Figure 11-1.

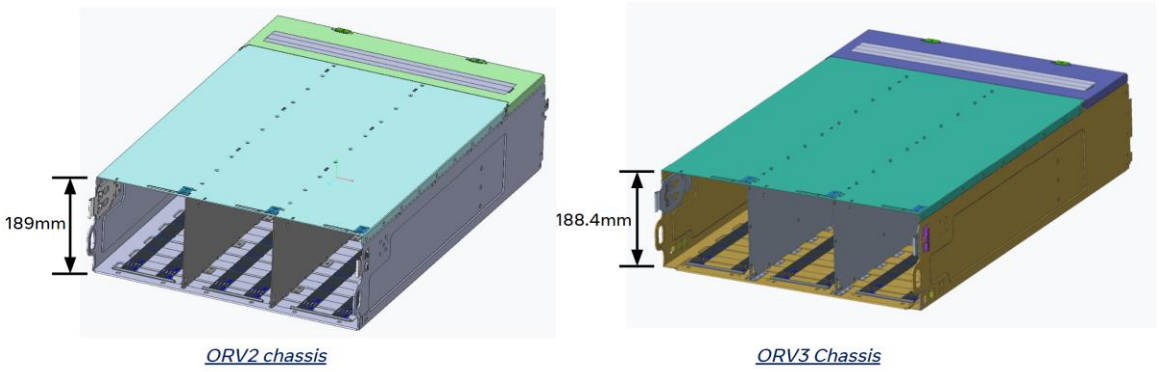


Figure 11-1: Yosemite V3.5 Chassis

11.2 Yosemite V3.5 Sled

A sheet metal and plastic sled serves as the mechanical interface between the Yosemite V3.5 Chassis and the Yosemite V3.5 blade. It is 4OU height, and it provides mechanical retention for the components inside the sled such as the fan bracket, baseboard, vertical power distribution board, Medusa board, and server cards. The combination of sheet metal tray, fan assembly, baseboard, Medusa board, multi-host NIC (OCP NIC 3.0 SFF), and power distribution board makes a Yosemite V3.5 Platform sled which can be seen in Figure 11-2.

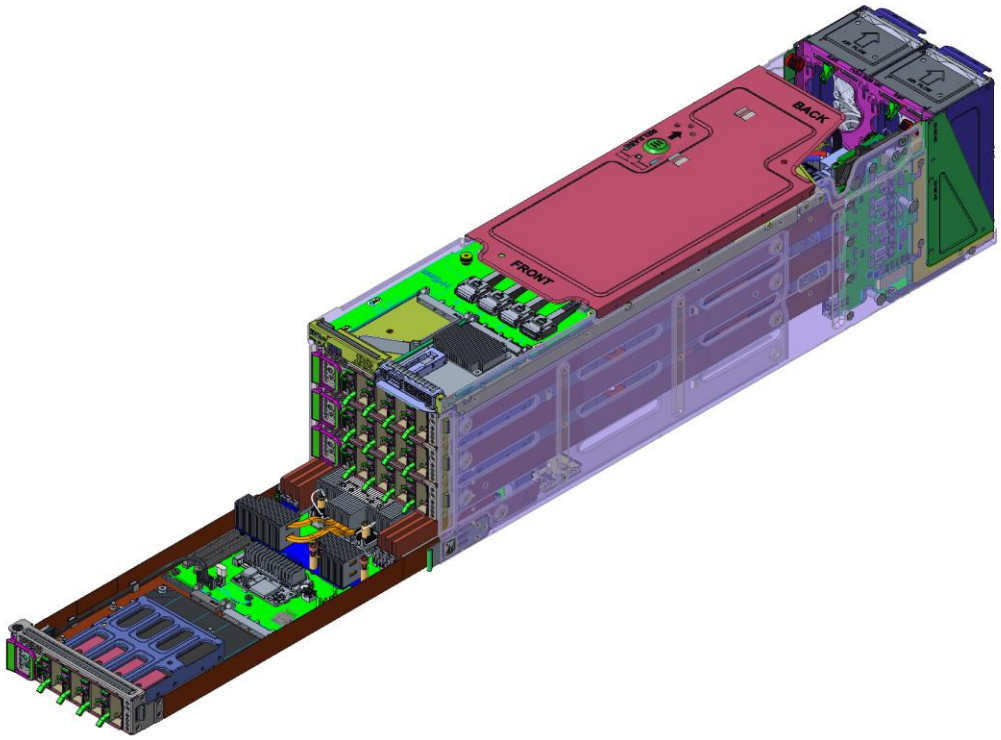


Figure 11-2: Yosemite V3.5, Sled Populated with 4x 1S Server

11.3 1S Server Blade

The server card is mounted to a sheet metal carrier called the blade, which slides on guides mounted inside the sled.

The blade comes in 1U and 2U configurations. Note that these are not a full 1OU (48mm) and 2OU (96mm) height because the sled top base board, NIC, and sheet metal structure takes some space off the 4OU total height. The 1U configuration can be in Figure 11-3 and is a height of 41mm. The 2U configuration is a height of 82.7mm. The blade design is specified in the OCP 1S Server Design Specification.

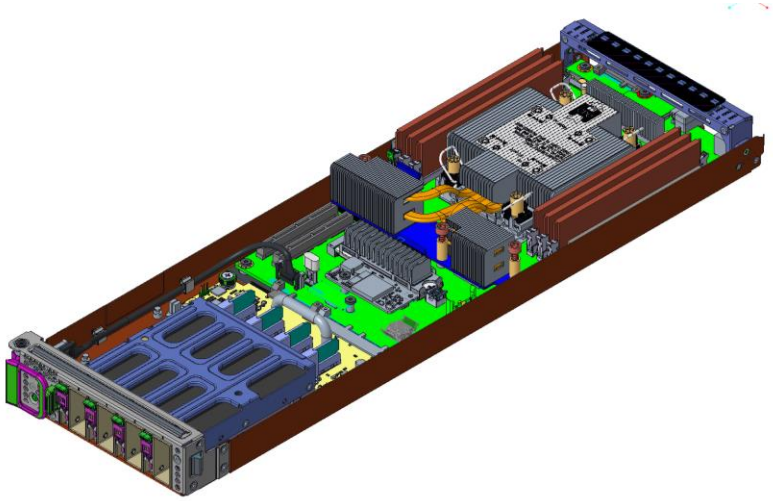


Figure 11-3: Yosemite V3.5 Blade

11.4 Silkscreen

Silkscreens on sheet metal will be black in color. Silkscreens on plastics and PCBs will be white in color and include labels for the components listed below. Additional items required on the silkscreen are listed below:

- Micro-server slots
- Fan connectors
- LEDs
- Switches as PWR and RST.

11.5 Retention

Chassis to Open Rack - The chassis is seated on rack knife and the chassis latch secures its position. A chassis can be removed from a rack by deflecting the chassis latch toward the center of the system, then pulled out using the chassis handle. Note that to remove the chassis from the rack, the sleds need to be removed first so the chassis latch can bend inward.

Sled to Chassis - When the sled is in its “home” position within the chassis, a sled latch engages with the vertical partition of the chassis to ensure it is held firmly in place. Press the sled latch and then pull the sled handle to take the sled out of the chassis.

Blade to Sled - When the blade is fully installed into the sled, the blade ejector rotates into the ejector bottom to lock the blade into place. To remove the blade, push on the ejector button to release the ejector latch and pull the latch to take the blade out.

Figure 11-4 helps illustrate these points.

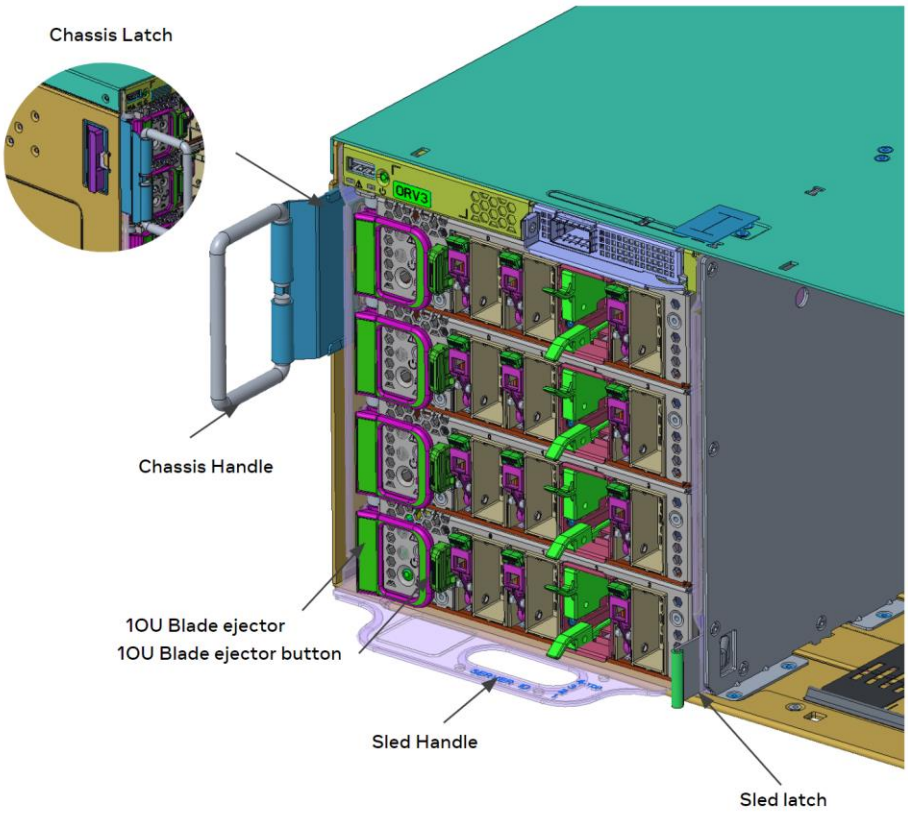


Figure 11-4: Sled Retention, 1OU Blades

12 Onboard Power System

12.1 Input Voltage Level

The YV3.5 platform can be interfaced with both ORv2 and ORv3 rack. If connecting with ORv2 racks, the expected nominal input voltage delivered by the power supply is 12.5 VDC; however, it has a varying range of 11.5V to 13.5V. The motherboard shall accept and operate normally with an input voltage range of 12.5V +/- 7%.

The expected nominal input voltage delivered by the Open Rack V3 Power Shelf is 51VDC. If connecting with ORv3 racks, the YV3.5 input power delivery subsystem shall accept and operate normally with an input voltage tolerance range between 46V and 52V.

12.2 48V support

The YV3.5 platform can be interfaced with both ORv2 and ORv3 rack. If connecting with ORv3 racks, the YV3.5 input power delivery subsystem shall accept and operate normally with an input voltage tolerance range between 46V and 52V. The YV3.5 Platform implements a two-stage conversion topology, meaning rack input bus voltage is stepped down to an intermediate bus voltage of 12V for the CPU and Switch Board whereas direct 51V is provided to the CPU board.

12.3 Platform Power Budget

The Yosemite V3.5 platform shall be designed to support a maximum sustained 1.5kW of distributed power among its subsystems. The BMC shall be responsible for summing the power telemetry reported by the baseboard and each server blade to ensure the entire sled does not operate beyond the platform power budget. The YV3.5 platform can be interfaced with both ORv2 and ORv3 rack. Depending on the Yosemite V3.5 sled configuration, the total power budget is governed by the current carrying capability of various connectors as shown in Figure 12-1 and Figure 12-2.

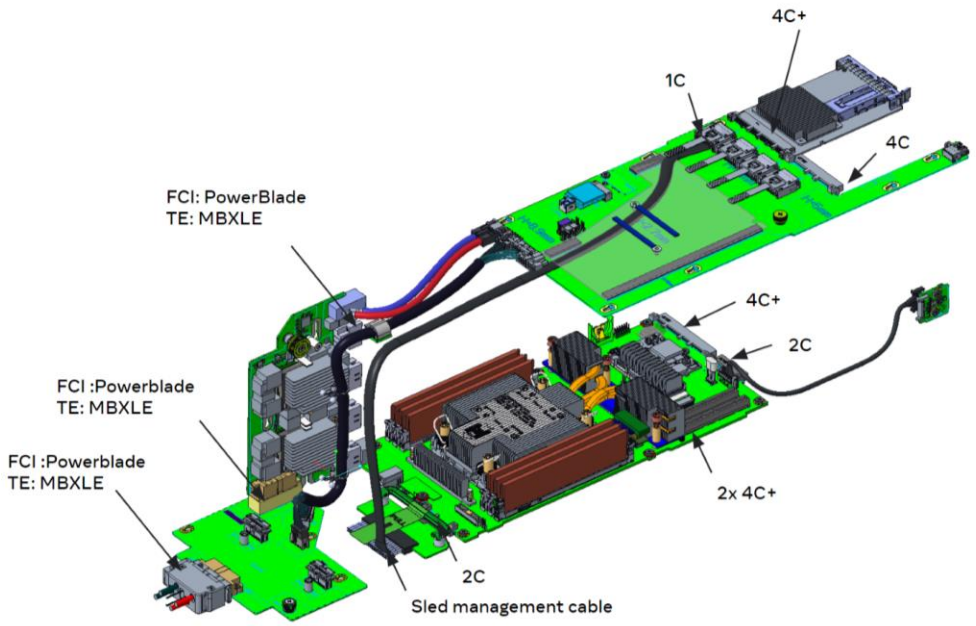


Figure 12-1: Yosemite V3.5, Connector Map, ORV3 –48V Power delivery, Crater Lake

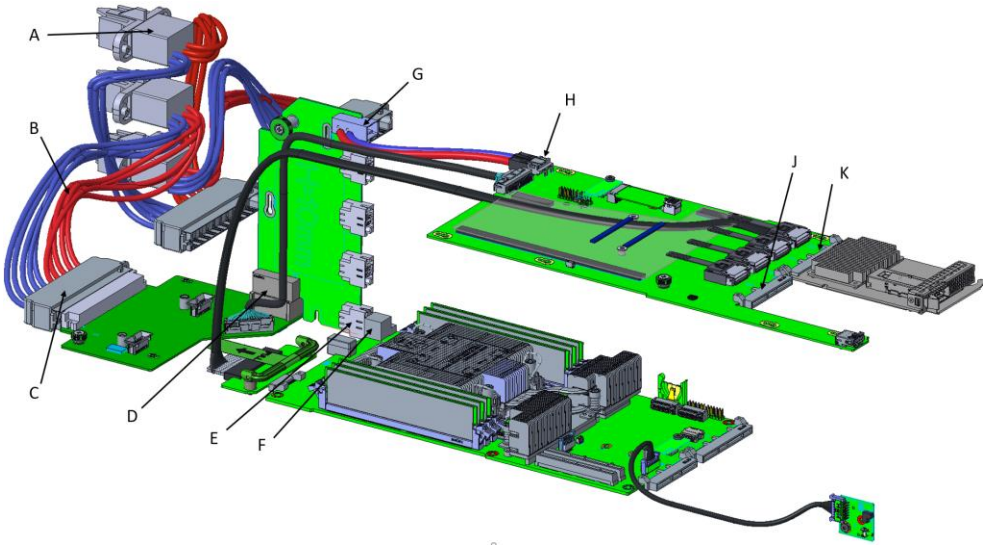


Figure 12-2: Yosemite V3.5, Connector Map, ORV2 –12V Power delivery, Crater Lake

12.4 Capacitive Load

To minimize the inrush current applied to theORv2/ORv3 Power Shelf during initial rack power on and assertion of Yosemite V3.5 sleds into a live bus bar, special design consideration around input capacitance must be considered:

1. Between the bus bar and input of any load switch or Hot Swap Controller (HSC), the placement of input capacitors is disallowed (if truly necessary, this must be thoroughly investigated and validated)

2. The placement of input capacitors at the output of any onboard load switch or HSC is allowed, but the total sum of the capacitance from all payloads within same rack shall not exceed a combined amount of 60mF per power shelf

3. The total simultaneous slew rate from all payloads within same rack shall not exceed 6A/us per ORv3 power shelf under any circumstances (e.g. power on soft start and parallel payload transients)

12.5 Hot Swap Controller Circuit

As mentioned in section 5.3, the Yosemite V3.5 platform shall implement a dedicated HSC on the baseboard and each server blade. The HSC is expected to support the following:

1. In-rush current control when motherboard is inserted and powered up
2. MOSFETs must be kept within safe operating area during all operational conditions such as power on/off and fault conditions
3. Signals that indicate power status, alerts, interrupts are expected to allow rapid response upon impeding fault conditions and/or warnings

4. Current limit protection for over current and short circuit whereby overcurrent threshold should be configured to 62 A for the baseboard
5. Undervoltage and overvoltage protection shall be configured to 10.1 V and 14.5 V for ORv2, and 43.96 V and 64.64 V for ORv3 respectively
6. Default HSC response for fault conditions shall be latch off with retry as stuff option
7. PMBus interface that supports the following features:
    - Report voltage, current, and power (VIP) telemetry with accuracy of +/- 2.0% or better when operating above 10% of the maximum range
    - Status registers that allow the definition of upper and lower critical thresholds for VIP of which are logged upon being triggered
8. Implements a fast (<20us) overcurrent monitoring scheme that generates an alert based on a remotely programmable threshold that triggers system throttling (Fast PROCHOT#) either using the HSC itself or external circuits. The recommended threshold for Fast PROCHOT# shall be slightly lower than the overcurrent limit such that there is no tolerance overlap.

Note: Detailed HSC design requirements for server blades is defined in the separate 1S Server Design Specification, please refer to the corresponding document for appropriate guidelines.

The voltage drop on the HSC current-sense resistor should be less than or equal to 25mV at full loading.

The power reporting of the HSC must be better than 2%, from 50W to full loading at room temperature. Further optimizations to power telemetry accuracy shall be performed through firmware based on characterized results from multiple boards based on entire load range and operating temperature requirements.

12.6 1S Server Power Management

The Yosemite V3.5 Platform supplies single 12V power to all server blade slots. The dedicated HSC on each server blade can be remotely controlled by the BMC. Upon insertion of the 1S server blade, the HSC should be ready by default and await for BMC's HSC enable signal.

The BMC shall implement a sophisticated power management algorithm that monitors the total platform power consumption and individual server blades. Every server blade is responsible for reporting a one-second average power consumption based on samples collected from its HSC as a power sensor reading stored within the Bridge IC whereby it could be accessed by the BMC via SMBus.

A fast throttle feature is implemented on the platform. It enables BMC to throttle an individual server blade or all server blades down to lowest power state in the shortest possible time. The HSC of each server blade may trigger this signal in the event of a blade level over current event occurs, but the BMC is still able to perform throttling on a as needed basis.

12.7 VR Efficiency

High efficiency Voltage Regulators (VRs) shall be used on the YV3.5 platform with at least:

1. 97.5% efficiency over the 30% to 90% load range for 48V to 12V

2. 93.0% efficiency over the 30% to 90% load range for 12V to XPU (CPU, SOC, ASIC, etc)
3. 90.0% efficiency over the 30% to 90% load range for 12V to POL

Note: If higher efficiency VRs are available at additional cost and/or design complexity, then the vendors are encouraged to present the tradeoffs prior to implementation. Beyond meeting design constraints such as efficiency, component size, and total solution cost, special considerations such as enabling multi-sourcing should be considered during solution surveying as a high priority requirement. The support of a digital interface (e.g. PMBus) is not mandatory for VR designs; however, the added benefits of sensor telemetry and configurable VR setting may prove useful as it enables voltage margin, health monitoring, and remote VR firmware update capabilities.

12.8 Power Policy

The power policy of server blades on the Yosemite V3.5 Platform can be set by the BMC to Always On or Last Power State. When the power policy is Always On, the server modules will be powered on automatically regardless of their last power state. When the power policy is Last Power State, the server modules will restore the last power state after AC cycling.

12.9 P12V_PSU to GND Clearance

Design consideration must be taken when routing unprotected power planes, such as P12V_PSU, which is responsible for carrying current from the bus bar to input of HSC. Below are layout recommendations that should be followed when possible with an example shown in Figure 12-3:

- On same and adjacent layers, P12V_PSU shape to all other nets, including GND ≥ 40 mil
- On different layers, from P12V_PSU shape to all other nets, including GND ≥ 2 layers of dielectrics if overlapping
- P12V_PSU traces are typically needed to provide biasing for HSC and related circuitry. Such traces must be ≤ 20 mil and has 40mil clearance to other signals on the same layer. On adjacent layer, it is preferred to generate void in plane to provide clearance to P12V_PSU where there is no other tradeoff.

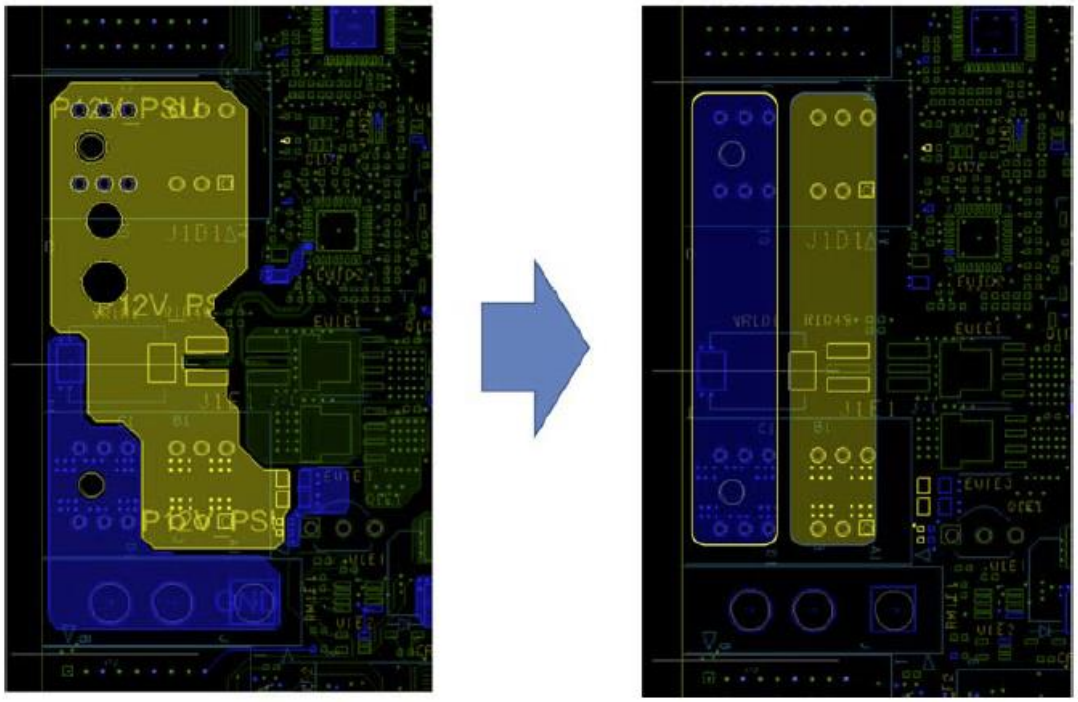


Figure 12-3: Example P12V_PSU and GND Layout

13 Environmental Requirements

The full system with the server card installed meets the following environmental requirements:

- Gaseous contamination: Severity Level G1 per ANSI/ISA 71.04-1985
- Storage relative humidity: 10% to 90% (non-condensing)
- Storage temperature range: -40°C to +70°C (long-term storage) *
- Transportation temperature range: -40°C to +70°C (short-term storage) *
- Operating altitude with no de-rating to 6000 ft

13.1 Vibration and Shock

The motherboard meets shock and vibration requirements according to the following IEC specifications: IEC78-2-(*) and IEC721-3-(*) Standard & Levels. See Table 13-1 for details.

Table 13-1: Vibration and Shock Requirements

	Operating	Non-Operating
Vibration	0.3G, 5 to 500 to 5 Hz per sweep, 10 sweeps at 1 octave/minute, test along three axes 5-20Hz – 6db/Oct 20-200Hz – 0.0003 G^2/Hz 200-500 – -6db/Oct	1G, 5 to 500 to 5 Hz per sweep, 10 sweeps at 1 octave/minute, test along three axes
Shock	6G, half sine, 11ms, 5 shocks, test along three axes	12G, half sine, 11ms, 10 shocks, test along three axes

14 Prescribed Materials

14.1 Disallowed Components

The following components are not used in the design of the motherboard:

- Components disallowed by the European Union's Restriction of Hazardous Substances Directive, **RoHS 2 Directive (2011/65/EU)**
- Trimmers and/or potentiometers
- Dip switches

14.2 Capacitors and Inductors

The following limitations apply to the use of capacitors:

- Only aluminum organic polymer capacitors made by high-quality manufacturers are used; they must be rated 105°C.
- All capacitors have a predicted life of at least 50,000 hours at 45°C inlet air temperature, under the worst conditions.
- Tantalum capacitors using manganese dioxide cathodes are forbidden.
- SMT ceramic capacitors with case size > 1206 are forbidden (size 1206 are still allowed when installed far from the PCB edge and with a correct orientation that minimizes the risk of cracking).
- X7R ceramic material for SMT capacitors should be used by default and at minimum X6S for portions of design subject to thermal hotspots such as CPU and/or DIMM cavities.
- COG or NP0 type should be used for tolerance sensitive portions of design.
- Conditional usage of X5R ceramic material must be based on evaluation of worst-case thermal conditions and upon approval from Meta.

The following limitations apply to the use of inductors:

- Only SMT inductors may be used as the use of through-hole inductors is disallowed.

14.3 Component De-rating

For inductors, capacitors, and FETs, derating analysis is based on at least 20% derating.

15 Hardware Management

The Yosemite V3.5 Platform uses a BMC for various platform management services and interfaces with hardware and BIOS firmware. The BMC utilizes ASPEED's AST2620.

The BMC should be a stand-alone system in parallel to the 1S servers (and/or device carrier cards). The health status of the 1S servers (and/or device carrier cards) should not affect the normal operation and network connectivity of the BMC.

15.1 1S Server I^3C Connections

There is a Bridge IC on each 1S server as a satellite management controller. The Bridge IC utilizes MCTP transactions over an I^3C connection from the Bridge IC on the 1S server to the BMC. The BMC is the primary management interface for the 1S server. Each 1S server's I^3C connection must be to a separate port on the BMC to ensure a dedicated connection with no conflicting traffic. The aspired speed for this communication is to be as fast as the platform/chipsets can support in a reliable manner. The target is a clock of 12.5 MHz with an effective bandwidth of 11Mbps.

The I^3C alert signal from each 1S server slot must be connected to the BMC. It provides an interrupt mechanism for the BMC. If the alert signal is asserted, the BMC must read the 1S server card and determine the source and cause of the interruption. If action is required, the BMC must respond in a timely fashion.

YV3.5 also has an I^2C channel that connects the BMC to the server blade's CPLD. This channel is used for more direct and granular control from the BMC to the server blade's CPLD.

1S Server Command Interface

The BMC and the Bridge IC on the 1S server communicate with each other via IPMI commands transmitted using MCTP transactions over their shared I3C interface.

15.2 1S Server Serial Connections

All serial ports on the 1S server slots are connected to the BMC directly. The BMC shall implement Serial-Over-LAN (SOL) functionality to allow a user to access a 1S server remotely.

The BMC also shall redirect a 1S server's serial port to an OCP debug card on the front panel to allow local debugging. A user will use the select button on the OCP debug card to switch between different hosts in the system. By default, the BMC enables SOL to all 1S servers and expansion cards which contain a Bridge IC. When an OCP debug card is connected and activated on the selected 1S server, the BMC shall provide full access to the serial console for debug purposes and make any existing SOL session to that server a read-only session to avoid possible data collisions.

15.3 1S Server Discovery Process

Initial Discovery

The BMC can detect that a 1S server card is installed using the PRSNT# pin signal coming through the Interconnect board. If the signal is low, it means the BMC has detected a card and it has initiated the discovery process. The discovery sequence is defined as follows:

- The BMC is to validate if the server card's standby power is OK
- The BMC can query the local CPLD to check the card type (passed thru serial stream)
- The BMC collects the FRU information from the Bridge IC
- The BMC sensor tables are updated from the Bridge IC
- The 1S server card is powered on based on the user input or as defined by the power policy configuration (e.g. always-off, always-on, last-power-state)

15.4 1S Server Power-on Sequence

When the server blade is inserted, the BMC will detect its present pin and enable the blade's HSC. Assuming the HSC and standby VRs are good, a STBY_PWROK signal will go high. Thereafter, BMC can assert the PWR_BTN# to the 1S server to initiate main power-on. The BMC

will then poll the Main Power OK status from the server blade to confirm if it has powered on successfully.

It should be noted that if the user removes and re-inserts the server blade, the platform does NOT power on the card immediately but waits for at least 1 second before doing so. In addition, the platform will NOT power all blades at the same time but sequence the power-on process in gaps of 1 second per blade.

15.5 Network Interface

The BMC connects to the network through the PCIe based multi-host OCP NIC 3.0 card. The BMC can use its built-in media access controller (MAC) to transfer management traffic through an NC-SI interface with a TOR switch.

The OCP 3.0 card provides PRSNT pins and BIFUR config pins as per the OCP NIC 3.0 card specification. The BMC shall use this information as well as a known PCIe cable connectivity to configure the NIC. All unused interfaces and devices shall be disabled so that they will not interfere with the activated management interface and device.

Many OCP 3.0 NICs provide multiple NC-SI channels for the BMC to use to access the network. The BMC supports detection of a loss of network access and will attempt to use alternative channels if available.

The BMC FW needs to support both IPv4 and IPv6.

15.6 BMC Multi-Node Requirements

Since there are up to four 1S servers managed by a single physical BMC, the BMC shall provide virtualized BMC (vBMC) functionality to manage each server. The vBMC is responsible for providing local and remote management for each server.

15.7 Local Serial Console and Serial-Over-LAN

The BMC needs to support two paths to access a serial console:

1. A local serial console on a debug header
2. A Serial-Over-LAN (SoL) console

These must be supported through the management network. It is preferred that both interfaces are functional at all stages of system operation. When there is a legacy limitation that allows only one interface to be functional, the default is set to SoL. The BMC needs to be able to switch console connection between SoL and Local on the fly, based on the input of the Serial-Console-Select signal on the front panel.

During system booting, POST (Power-On Self-Test) codes will be sent to Port 80 and decoded by the BMC to drive the LED display. POST codes should be displayed in the SoL console during system POST. Before the system displays the first screen, POST codes are dumped to – and displayed in – the SoL console in sequence (e.g., “[00] [01] [02] [E0],” etc.) After the system shows the first screen in the SoL console, the last POST code received on Port 80 is displayed in

the lower right corner of the console.

15.8 Graphics and GUI

The Yosemite V3.5 Platform does not support the BMC to support graphic, KVM, or GUI features. All the BMC features need to be available in command-line mode by in-band and OOB IPMI command, or by SoL.

15.9 Remote Power Control and Power Policy

The vendor should implement the BMC firmware to support remote 1S server and expansion card (if supported) power on/off/cycle and warm reboot through the in-band or out-of-band interface.

The vendor should implement the BMC firmware to support the power-on policy to be Last State, Always On, and Always Off. The default setting is always on. The change of power policy should be supported and take effect without cold resetting the BMC firmware or rebooting the 1S server systems or associated expansion cards.

If AC power is applied to the BMC, it should take less than three seconds for the BMC to process the Power Button signal and power up the system for POST. It must not wait for the BMC to become ready (which will take about 90 seconds) before processing the Power Button signal.

In order to accommodate the requirement to process the Power Button signal in less than three seconds, the BMC shall enable a pass-through mode in the very early booting stages. This mode must make signals like Power Button, Reset, Universal Asynchronous Receiver/Transmitter (UART), POST Code, etc., available. Once the BMC boots completely (approximately 90 seconds), it shall also take over the control of these signals from the pass-through mode smoothly without any glitches.

15.10 POST Codes

The Bridge IC on the 1S server will pass POST codes to the BMC. The BMC should enable the POST code display to drive 8-bit HEX general-purpose Input/Output (GPIO) data to the OCP debug card on front panel. The BMC post function needs to be ready before the 1S server system BIOS starts to send the first POST code to the corresponding port. The POST codes should also be sent to the SoL interface so that the POST process can be monitored remotely.

15.11 System LEDs and Buttons

The Yosemite V3.5 has the following for the Sled Chassis

1. Power LED (Blue)
2. Fault/LOC/Select LED (Amber)
3. AC cycle

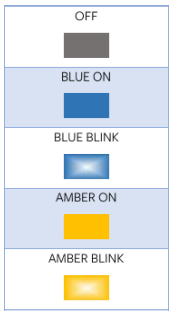
The Power LED and Fault/LOC LED shall follow specifications based on OCP Panel Indicator Specifications which can be found at the following link:

http://files.opencompute.org/oc/public.php?service=files&t=65c02b1c6d59188351357cfb232cbfaa.

A quick reference on the LED form the doc (based on 1.0 specifications) is shown below:

Permitted States	Separate LEDs	
	PWR (Blue)	FAULT/LOC (Amber)
System AC Off/Service Action Allowed		
System On/Status OK		
System Off + Fault	/	
System On + Locate		
System Off + Locate	/	
System On + Fault		
System DC Off		

Legend:



Note that AC off and Standby OFF seems to show same power state if Fault and Locate is not lit.

As for the Select LED, its behavior is based on whether the based board is selected for OCP Debug Cards functions.

For the AC_Cycle button, the intent is as follows:

1. Press less than 4 seconds. No effect. This behavior is to ensure the press of button is INTENTIONAL
2. Press more than 4 seconds, sled will be AC cycled. This is INDEPENDENT of whatever state the hosts/system is.

As for the front panel IO for the server, the following are suggested:

1. Power LED
2. Fault/LOC LED
3. AC_Cycle button

For the LEDs, the server will follow per description earlier for the baseboard.

The AC_Cycle button functions as follows:

1. Press more than 4 seconds, AC power cycle will happen on the server card only.

15.12 Time Sync

Since the Yosemite V3.5 Platform system has no CMOS battery backup, the BMC time sync should be from the Network Time Protocol (NTP) server.

The BMC needs to sync its clock from the NTP server as soon as its network interface is up and running. The BMC should also sync its clock from the NTP server periodically.

Since there is no battery on the BMC, the 1S server BIOS shall not issue IPMI Get System Event Log (SEL) Time command to sync its system clock during POST. The BMC should reject this command if its internal time is not properly synced-up with NTP server.

NTP Time Sync Flow

- BMC first time power on.
- The BMC tries to sync its time with one of the server cards as the server card might have battery backed up RTC.
- BMC firmware image might contain the default NTP IP address and NTP retry configuration.
- The Provisioning server will send the Set NTP IP Address command to the BMC.
- If BMC network interface is down, BMC will wait till network interface is up.
- On seeing BMC network interface is up and running.
    - BMC queries date/time for its clock using the configured NTP IP address.
    - A default date/time (e.g., the time from one of the server cards) will be used for any event log until a date is properly set.
- Once the BMC date/time is synced, the BMC will log the event and start using new time for any events that happen later.
- The BMC will sync its date/time from the NTP server periodically with an interval.

15.13 Power and Thermal Monitoring, and Power Limiting

The BMC firmware supports platform power monitoring. Enabling power monitoring for the 1S servers requires an accurate power sensor on 12.5V to the 1S server. This function is available for users to access through in-band and OOB.

The BMC firmware supports thermal monitoring, including 1S server SOCs, 1S server memory, and inlet/outlet air temperatures. To ensure accuracy, a TI TMP421 with an external PN junction is preferred to detect inlet and outlet temperatures. Take caution when implementing inlet air sensors. It is important to avoid preheating nearby components and to reduce the amount of heat conducted through the printed circuit board (PCB).

The BMC firmware supports a power-limiting feature to make sure the platform is not drawing more power than allocated. The BMC will monitor the power consumption of each 1S server and use an SOC-specific management controller interface to limit the SOC's power consumption (e.g., P-State control).

As the host and baseboard in YV3.5 are tapping power from the vertical power bar, BMC is to periodically query the following:

- The total current flowing through the HSCs of the 1S servers. This is to cap it to ~140 Amps upon sampling total. Upon seeing higher than expected current, BMC should initiate a system wide throttle to all the servers.
- Power, voltage, current and temperature around the system (Baseboard and the server blades).
- BMC also needs to handle cases like HSCs not responding or timeout issues intelligently while at the same time providing a good methodology to ensure data coming into the system is trustworthy and can be used to mathematically sum for a result.

15.14 Sensors

Both analog and discrete sensors may reside on the baseboard, the 1S servers and any associated expansion cards. The BMC must provide a way to read all sensors across the platform, e.g., sensors on the baseboard, sensors on a given server, sensors on a device carrier card, and sensors on the OCP mezzanine card. BMC should also be able to determine a bad or missing sensor condition whereby it is RESILIENT to their misbehavior, while at the same time, will log the anomaly.

Analog Sensors

The BMC has access to all analog sensors on the Yosemite V3.5 Platform directly or through the 1S server management connection.

Some of the required analog sensors include (but are not limited to):

- Outlet Temp
- Inlet Temp
- Slot Current
- SoC Thermal Margin
- SoC VR Temp
- SoC DIMM VR Temp
- Hot Swap Controller's power/current/voltage
- SoC TjMax
- Airflow
- System Fan Speed

BIOS/ME generated Sensors

Sometimes when the BIOS/ME detects a failure, it generates an SEL entry to be logged in the BMC. Some of the required event only sensors include (but are not limited to):

1 Firmware health
2 POST errors
3 Power errors
4 ProcHOT
5 Machine Check errors
6 PCIe errors
7 Memory errors, etc.

15.15 Event Log

The BMC supports storing events/logs from each 1S server, baseboard, device carrier card, and mezzanine card. Errors listed here may not be exhaustive and may not cover different 1S server designs.

Logged Errors

CPU Error

Both correctable ECC errors and uncorrectable ECC errors are logged into the Event log. Error categories include Link and L3 Cache.

Memory Error

Both correctable ECC errors and uncorrectable ECC errors are logged into the Event log. The Error log should indicate the location of the DIMM (if applicable), channel #, and slot #.

PCI-E Error

All errors, which have a status register, are logged into the Event log, including root complex, endpoint devices, and any switch upstream/downstream ports if available. Link disable on errors are also logged. The error classifications Fatal, Non-fatal, or Correctable follow the 1S server vendor's recommendation.

POST Error

All POST errors, which are detected by BIOS during POST, are logged into the Event log.

Power Error

Two power errors are logged. One is a 12.5V DC input power failure that causes all power rails on the baseboard to lose power, including standby power. The other is an unexpected system shutdown during system S0/S1 while the 12.5V DC input is still valid.

MEMHOT# and SOCHOT#

Memory hot errors and processor hot errors are logged. The Error log identifies the error source as internal, coming from the processor or memory, or an external error coming from the voltage regulator.

Fan Failure

Fan failure errors are logged if the fan speed reading is outside expected ranges between the lower and upper critical thresholds. The Error log also identifies which fan fails.

PMBus Status Error

The PMBus status sensors check the PMBus controller's health status and log an error if an abnormal value is detected. The PMBus controller can be a DC Hot Swap Controller (HSC) or a PMBus AC to DC power supply unit.

For all the above error logging and reporting, the user may select to enable or disable each logging option.

Error Threshold Setting

These settings enable the error threshold setting for both correctable and uncorrectable errors. Once a programmed threshold is reached, the system triggers an event and logs it.

- Memory Correctable ECC: Suggest setting the threshold value to be [1,000] in the mass production stage and [4] for the evaluation, development, and pilot run stage, with options of 1, 4, 10, and 1,000. When the threshold is reached, the BIOS should log the event, including DIMM location information and the output DIMM location code, through the debug card.
- ECC Error Event Log Threshold: Defines the maximum number of correctable DIMMs. ECC is logged in the same boot. The default value is 10, with options of Disable, 10, 50, and 100.
- PCIe Error: Follow the 1S server vendor's suggestion.

15.16 Fan Speed Control in BMC

The Fan Speed Control (FSC) support is enabled on the BMC. The BMC samples thermal related analog sensors in real time. The FSC algorithm processes these inputs and drives two pulse width modulation (PWM) outputs in optimized speed.

Fan Speed Control Specification

The FSC implementation in the BMC must refer to the OCP's FSC specification.

Data gathering for FSC

The BMC needs to gather data as input of the FSC. The required data is described in the Table 15-1.

Table 15-1: Required FSC Data

Type of data	Data to be used for FSC input
Temperature	1S server SOC temperature from all slots
Temperature	1S server DIMM temperature from all slots (if available)
Temperature	Inlet and outlet air
Temperature	1S server VR of SOC and DIMM from all slots (if available)
Temperature	Hot Swap Controller
Temperature	Switch temperature
Power	Platform power from HSC
Fan speed	2 Fan tachometer inputs

Fan Speed Controller in BMC

The BMC supports FSC in both proportional–integral–derivative (PID) and step mode. The BMC also supports both in-band and OOB FSC configuration updates. Updates should take effect immediately without rebooting. The BMC supports fan boost during fan failure.

15.17 BMC Firmware Update

Vendors should provide tool(s) to implement a remote BMC firmware update, which will not require any physical input. This remote update can occur either through OOB via the management network or by logging into the local OS via the data network.

A remote BMC firmware update may take five minutes (maximum) to complete. The BMC firmware update process and BMC reset process do not require the host system to reboot or power down. It should have no impact on the normal operation of the host system. The BMC needs to be fully functional, with updated firmware after the update and reset, without any further configuration.

The default update provides a method to recover the BMC to the factory default settings. Options need to be provided to preserve the SEL and configuration. The MAC address should not be cleared with the BMC firmware update.

15.18 Hot Service Support

The Yosemite V3.5 Platform supports hot service of any card in the system while keeping all other cards in service. The BMC shall detect these hot insertions and/or removals and update its database (FRUID information, sensor information, etc.). Since the newly inserted card could possibly be of a different kind, the BMC should be able to detect the new card and configure different services, e.g., sensor monitoring might need to restart for that slot to reflect the new hardware.

15.19 OpenBMC

OpenBMC refers to the open-source implementation of BMC functionality described in the above sections. This specification does not prevent alternate implementations that can meet similar functionality. The source code for OpenBMC is available at https://github.com/facebook/openbmc for reference.

15.20 OpenBIC

OpenBIC refers to the open-source implementation of BIC functionality described in the above sections. This specification does not prevent alternate implementations that can meet similar functionality. The source code for OpenBIC is available at https://github.com/facebook/OpenBIC for reference.

15.21 Security

The Class 1 baseboard configuration utilizes a TPM module connected to the BMC to support measured boot attestation and verified boot. The Class 2 baseboard has a bridge IC that supports secure boot. A platform root of trust (PRoT) implementation is not planned.

16 Security

16.1 TPM (Trusted Platform Module)

CPU host complex (CPU+PCH) shall have access to a TPM 2.0 module on its SPI interface. This TPM will exist as a hardware TPM physically on the Crater Lake board. The TPM will be used for measured boot. It will hold image measurements until the system is fully booted and connected to the network. From there, these measurements will be used to attest the system during provisioning. The TPM will only be used to store measurement values and will not contain any keys used elsewhere in the system.

17 Labels and Markings

The motherboard shall include the following labels shown in Table 17-1 on the component side of the motherboard. The labels shall not be placed in a way that may cause them to disrupt the functionality or the airflow path of the motherboard.

Table 17-1: Labels and Markings

Description	Type	Barcode Required?
Safety Markings	Silkscreen	No
Vendor P/N, S/N, REV (Revision would increment for any approved changes)	Adhesive label	Yes
Vendor Logo, Name & Country of Origin	Silkscreen	No
PCB Vendor Logo, Name	Silkscreen	No
Date Code (Industry Standard: Week / Year)	Adhesive label	Yes
RoHS Compliance	Silkscreen	No
WEEE Symbol. The motherboard will have the crossed out wheeled bin symbol to indicate that the manufacturer will take it back at the end of its useful life. This is defined in the European Union Directive 2002/96/EC of January 27, 2003 on Waste Electrical and Electronic Equipment (WEEE) and any subsequent amendments.	Silkscreen	No
CE Marking	Silkscreen	No
UL Marking	Silkscreen	No