# Yosemite V3: Facebook Multi-Node Server Platform Design Specification

# 1v16

Authors:
**Michael Haken**, Mechanical Engineer, Facebook
**Jarrod Clow**, Thermal Engineer, Facebook
**Yueming Li**, Thermal Engineer, Facebook
**Ben Wei**, Software Engineer, Facebook
**Damien Chong**, Hardware Engineer, Facebook
**Thoon KY**, Hardware Engineer, Facebook

**Kiran Vemuri,** Hardware Engineer, Facebook
**Todd Westhauser,** Hardware Engineer, Facebook
**Pavan Shetty**, Power Engineer, Facebook
**Anthony Chan,** Power Engineer, Facebook
**Haoran Wu,** Power Engineer, Facebook
**Deepa Saxena Mehta**, Hardware System Engineer, Facebook
**Sten Liao**, Hardware System Engineer, Facebook

## Copyrights and Trademarks

Intel® is a trademark of Intel Corporation in the U.S. and/or other countries.
Texas Instruments™ is a trademark of Texas Instruments Incorporated.
Tiva™ is a trademark of Texas Instruments Incorporated.

# 1 License

Contributions to this Specification are made under the terms and conditions set forth in Open Web Foundation Contributor License Agreement (“OWF CLA 1.0”) (“Contribution License”) by:

**Facebook Inc.**

You can review the signed copies of the applicable Contributor License(s) for this Specification on the OCP website at http://www.opencompute.org/products/specsanddesign
Usage of this Specification is governed by the terms and conditions set forth in **Open Web Foundation Final Specification Agreement (“OWFa 1.0”) (“Specification License”).**

You can review the applicable Specification License(s) executed by the above referenced contributors to this Specification on the OCP website at http://www.opencompute.org/participate/legal-documents/

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 OCP Tenets Compliance

## 2.1 Openness

The Yosemite V3 Platform Design Specification exemplifies openness by defining a flexible platform which can support numerous uses cases and providing the requirements for a YV3 compatible PCB. This includes a single sled design with numerous potential use cases and common management interface with allows for any users of the system to implement their own optimizations.

## 2.2 Efficiency

Compared to Yosemite V2, Yosemite V3 offers efficiency gains across the thermal management system due to the form factor modification, performance efficiency gains due to increases perf/w along with serviceability gains since all blades are front serviceable without affecting adjacent blades. There is also a substantial efficiency gain on the R&D side by leveraging the same sled architecture on the web server use cases, storage uses cases and accelerator use cases. This allows for more common parts which allows for reuse in multiple configurations.

## 2.3 Impact

The Yosemite V3 is a demonstration that flexibility does not need to be traded for high performance at scale. This is a system which can provide multi-host NIC interfaces as well as single-host NIC interfaces with minimal changes in the hardware required, all while fitting the in the same sled format factor. This allows industry partners to design only one new card for an entirely different use case instead of starting from the ground up on a new system.

# 3 Scope

This specification describes the design of the Yosemite V3 platform which supports either four One Socket (1S) Server blades or two sets of 1S Server blades with expansion per sled.

# 4 Contents


Copyrights and Trademarks ........ 2

1 License ........ 3

2 OCP Tenets Compliance ........ 4

2.1 Openness ........ 4

2.2 Efficiency ........ 4

2.3 Impact ........ 4

3 Scope ........ 4

4 Contents ........ 5

5 Overview ........ 8

6 Yosemite V3 Platform Features ........ 13

6.1 Platform Block Diagram ........ 13

6.2 Yosemite V3 Subsystems ........ 13

6.3 Yosemite V3 Platform system classes ........ 18

6.4 SMBus Block Diagram ........ 24

6.5 1S Server ........ 25

7 Baseboard Management Controller ........ 29

7.1 1S Server $I^2C$ Connections ........ 29

7.2 1S Server Serial Connections ........ 29

7.3 1S Server Discovery Process ........ 29

7.4 1S Server Power-on Sequence ........ 30

7.5 Network Interface ........ 30

7.6 BMC Multi-Node Requirements ........ 30

7.7 Local Serial Console and Serial-Over-LAN ........ 30

7.8 Graphics and GUI ........ 31

7.9 Remote Power Control and Power Policy ........ 31

7.10 POST Codes ........ 31

7.11 System LEDs and Buttons ........ 31

7.12 Time Sync ........ 33

7.13 Power and Thermal Monitoring, and Power Limiting ........ 33

7.14 Sensors ........ 34

7.15 Event Log ... 35
7.16 Fan Speed Control in BMC ... 36
7.17 BMC Firmware Update ... 40
7.18 Hot Service Support ... 40
7.19 OpenBMC ... 40
7.20 Security ... 40
7.21 Small Form Factor Baseboard Storage Module (SFF BSM) ... 41
8 System Firmware ... 42
9 Mechanical ... 42
9.1 Yosemite V3 Chassis ... 42
9.2 Yosemite V3 Sled ... 42
9.3 1S Server Blade ... 43
9.4 Silkscreen ... 44
9.5 Retention ... 44
10 Thermal ... 47
10.1 Data Center Environmental Conditions ... 47
10.2 Server Operational Conditions ... 48
10.3 Thermal Kit Requirements ... 49
11 I/O System ... 51
11.1 Front facing Server Modules ... 51
11.2 Network ... 51
11.3 Server Slots Assignment ... 52
11.4 Front Panel ... 52
11.5 Fan Connector ... 53
12 Power ... 55
12.1 Input Voltage Level ... 55
12.2 48V support ... 55
12.3 Platform Power Budget ... 55
12.4 Capacitive Load ... 56
12.5 Hot Swap Controller Circuit ... 56
12.6 1S Server Power Management ... 57
12.7 VR Efficiency ... 57
12.8 Power Policy ... 57

12.9 P12V_PSU to GND Clearance ....57
13 Expansion Board Design Requirements ....59
13.1 Overview ....59
13.2 Mechanical ....62
13.3 Thermal ....65
13.4 Electrical ....65
13.5 Power ....66
13.6 Functional ....68
14 Environmental Requirements and Other Regulations ....75
14.1 Environmental Requirements ....75
14.2 Vibration and Shock ....75
14.3 Regulations ....76
15 Prescribed Materials ....77
15.1 Disallowed Components ....77
15.2 Capacitors and Inductors ....77
15.3 Component De-rating ....77
16 Labels and Markings ....78
17 Hardware Management ....78
17.1 Compliance ....78
18 Security ....79
19 Revision History ....79
Appendix A - Requirements for IC Approval ....80

# 5 Overview

This document describes Facebook's next generation multi-node server platform (code name: Yosemite V3) and the design requirements to integrate the platform into OCP Open Rack V2. The Yosemite V3 platform supports up to four single socket(1S) Server blades or two 1S Server blade with expansion pairs per sled which can be installed into the new 4OU chassis. The platform consists of a baseboard that hosts the baseboard management controller (BMC) and provides network connectivity through a compatible OCP3.0 NIC. The baseboard also connects to each individual 1S Server module through Sled Management cables. Those cables carry the management and PCIe signals for NIC. Two or four cables are used to support various configurations as described in later sections of this document. Both the baseboard and 1S Server blade have independent onboard hot swap controllers (HSC) for power regulation and monitoring.

The BMC found on the baseboard is essentially the "brain" of the platform. Its functions include, but are not limited to:

- Control and monitor all thermal sensitive components on the platform
- Control and monitor power usage of major commodities of the platform such as:
  - power, voltage and current of 1S server blade and its onboard voltage regulators (VR)
  - power, voltage and current of the platform
- Fan control and monitoring
- Logging of different platform and system events
- Configuration of any programmable devices for its intended operation
- Tracking of all Field replaceable units (FRU) on the platform
- Any form of platform management needs

Yosemite V3 aims to improve serviceability over prior multi-node server platforms and is intended to support front loading 1S Server blades into a powered system, thus eliminating hot service and cable management complexities.

The Yosemite V3 Platform is designed to be compatible with the OCP Open Rack V2 specification. Please refer to the corresponding OCP Open Rack V2 documentation for more details about the rack. The Yosemite V3 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.

A simplified picture of the Yosemite V3 platform can be seen below.

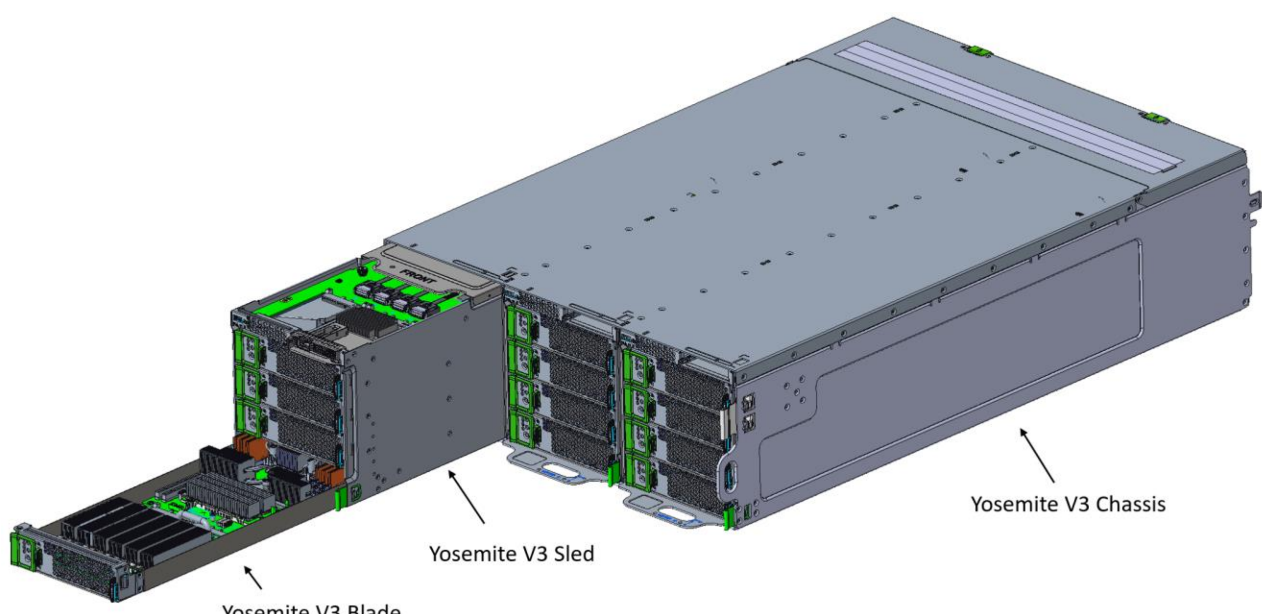


Figure 3-1: Yosemite V3, Chassis View

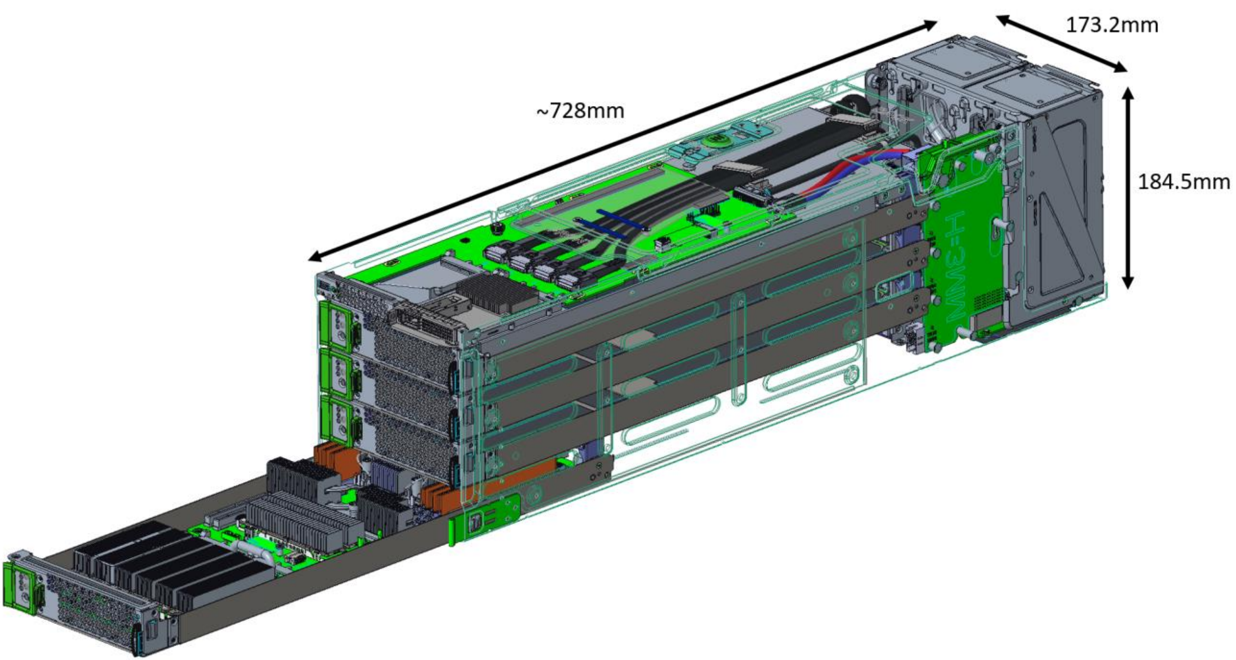


Figure 3-2: Yosemite V3, Sled View

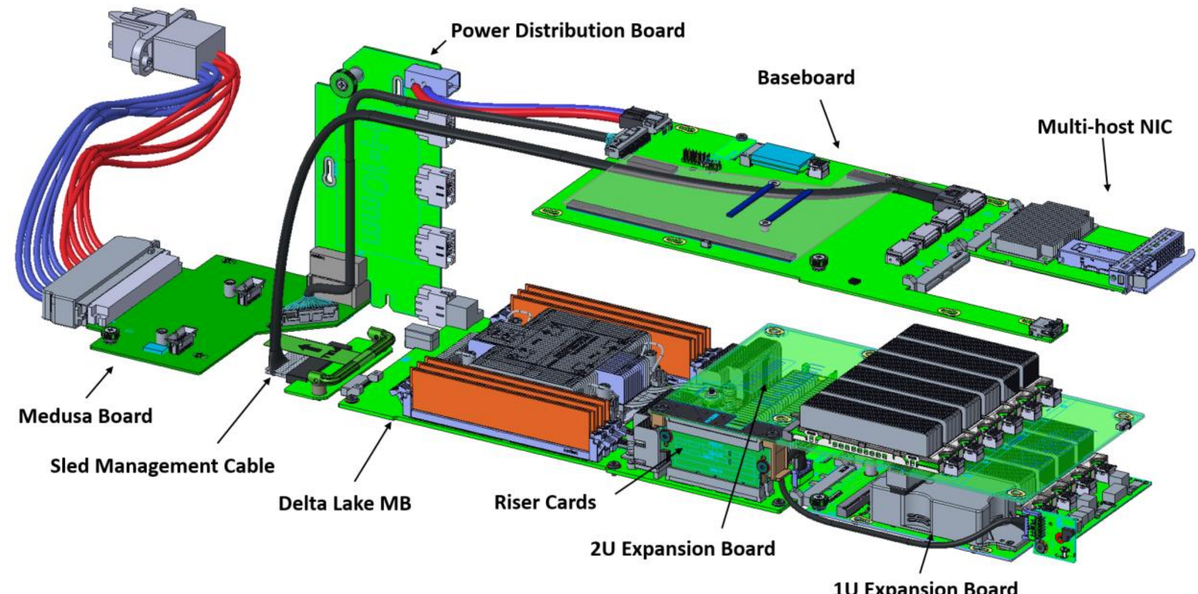


Figure 3-3: Yosemite V3, Sled Components

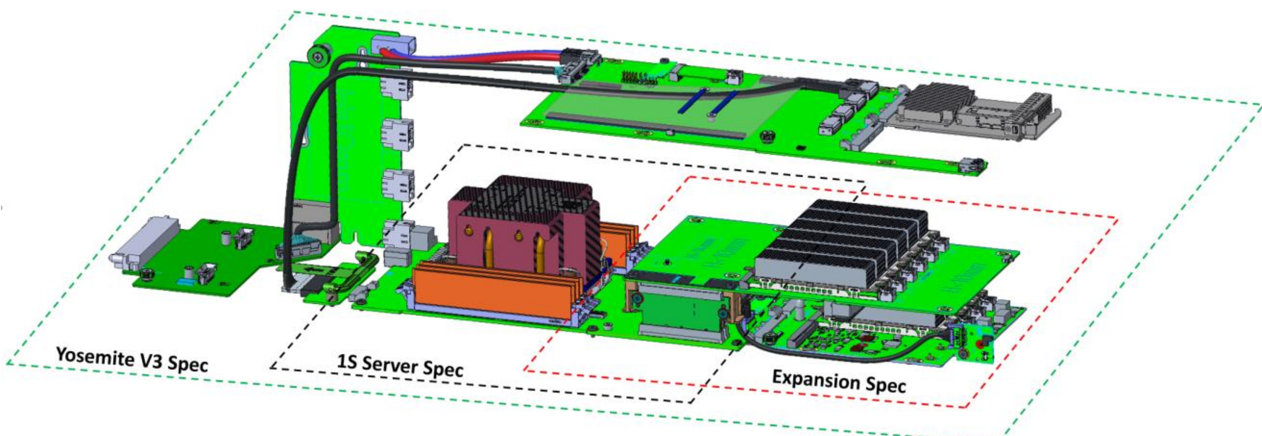


Figure 3-4: Yosemite V3, Specification View

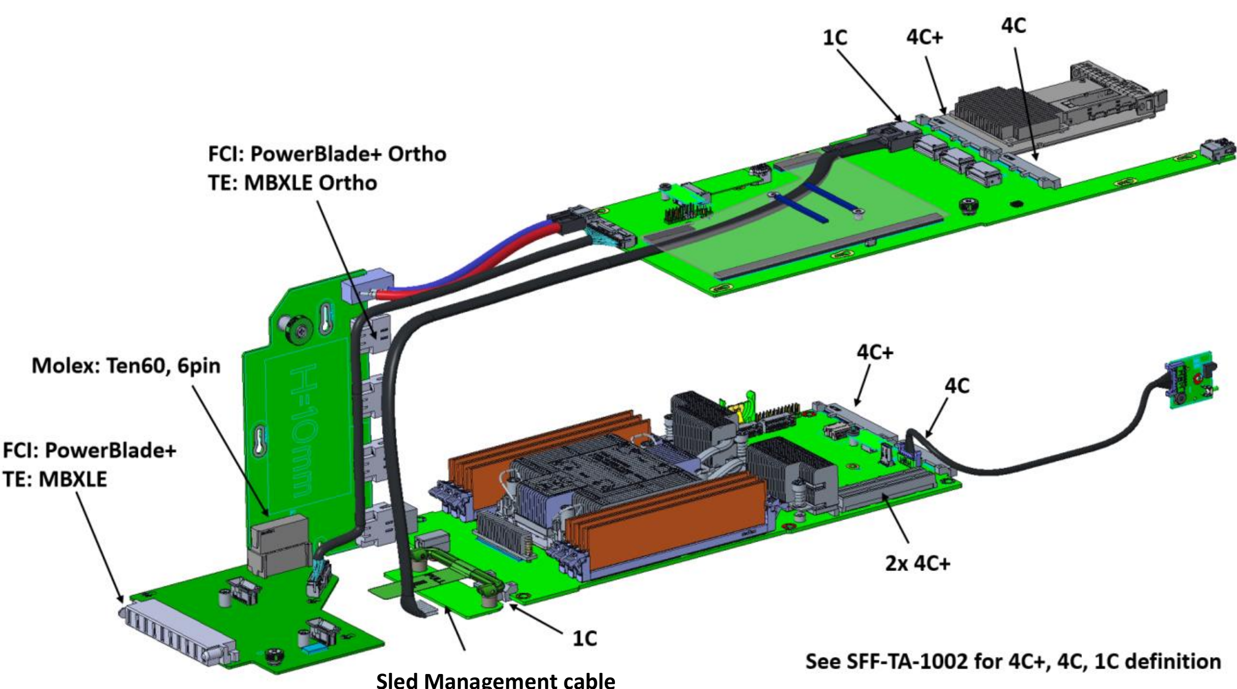


Figure 3-5: Yosemite V3, Connector Map

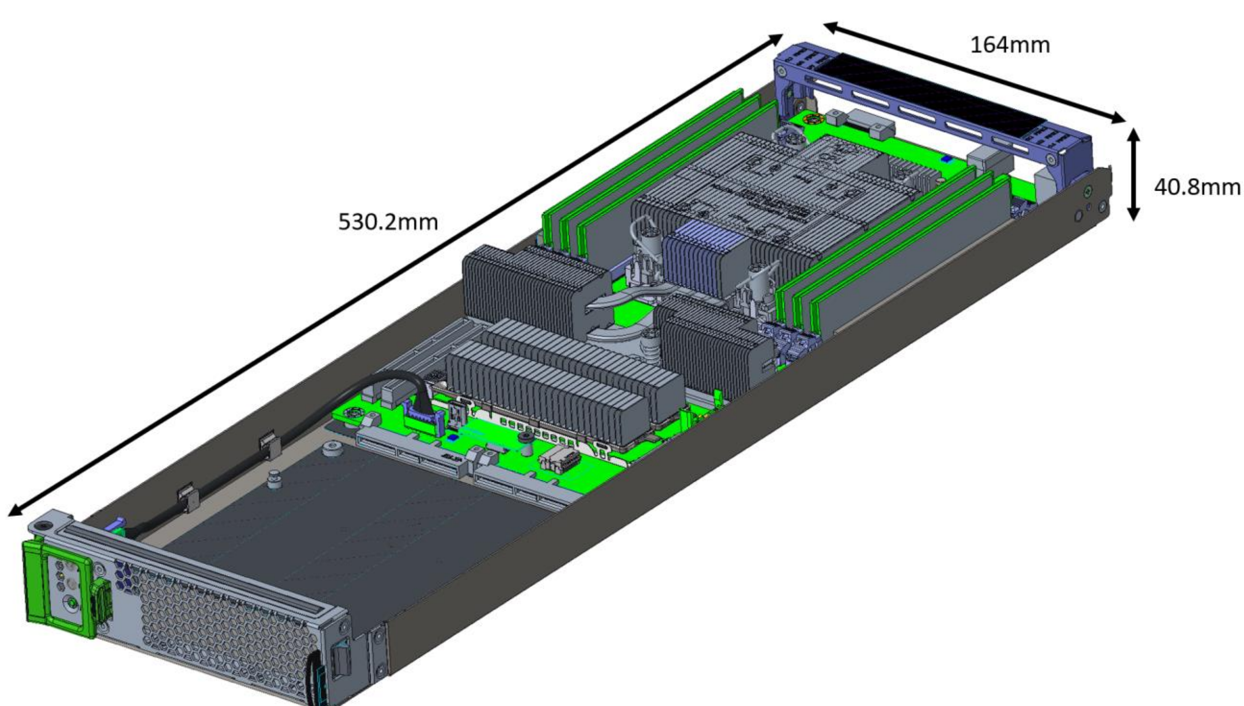


Figure 3-6: Yosemite V3, Example Blade Configuration

# 6 Yosemite V3 Platform Features

## 6.1 Platform Block Diagram

Figures 5-1 2 illustrate the functional block diagram and design details of the Yosemite V3 Platform.

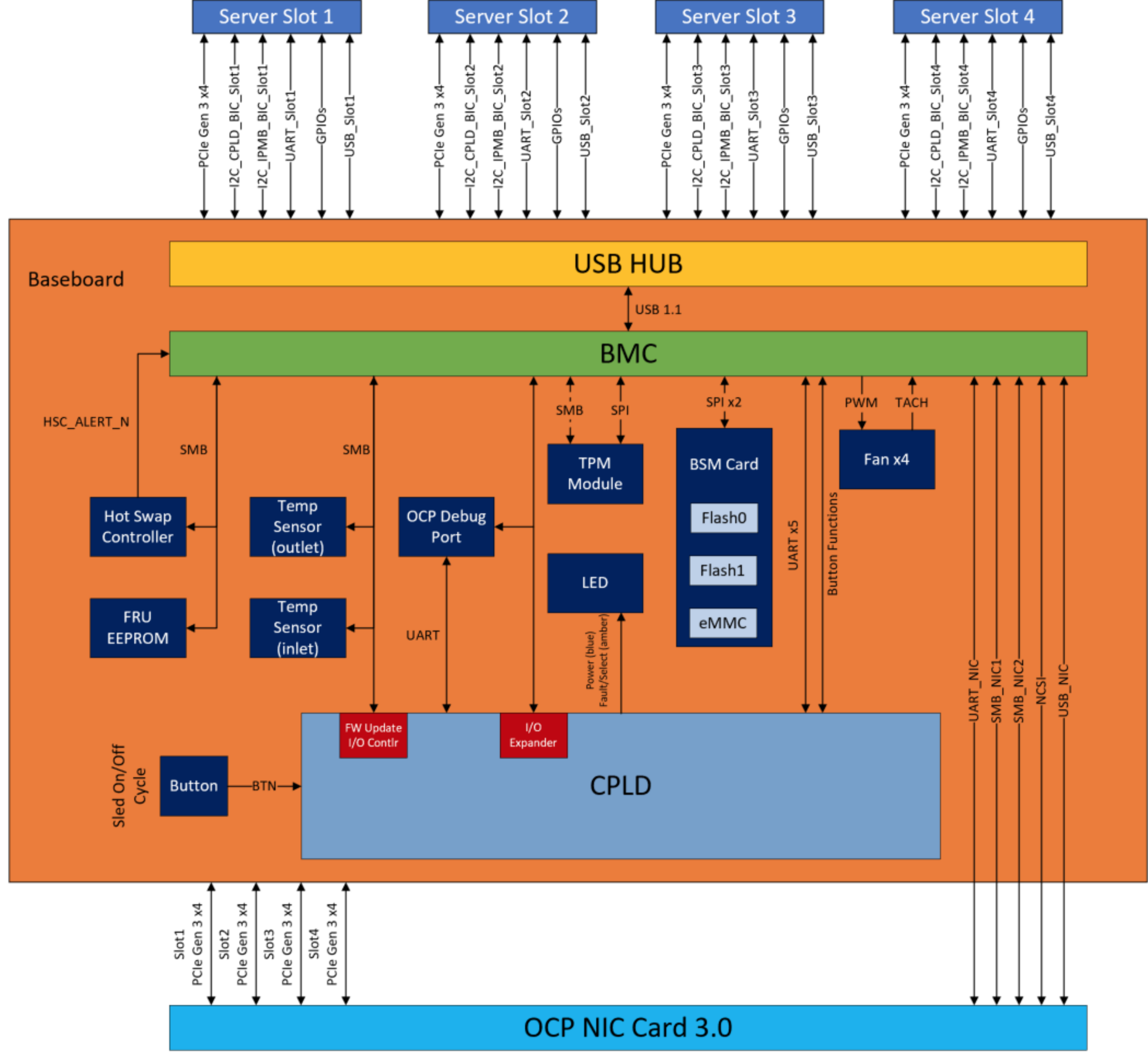


Figure 6-1: Platform Block Diagram

## 6.2 Yosemite V3 Subsystems

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

### 6.2.1 Input Power Delivery

For the Yosemite V3 platform, power is delivered from the Open Rack V2 bus bar to each subsystem through three separate segments namely: medusa cable, medusa power board (MPB) and vertical power distribution board (PDB) as shown in Figure 6-2. The flow of power shall remain uniform as it traverses through each segment until it is distributed to subsystems such as the baseboard and 1S server blades.
It is necessary to have a voltage reading at the Medusa power board to obtain input voltage to the sled. At the same time, an e-Fuse (FETs managed by Hot Swap Controller) should be provided on the Medusa power board to protect the system in event of a short circuit in the sled.

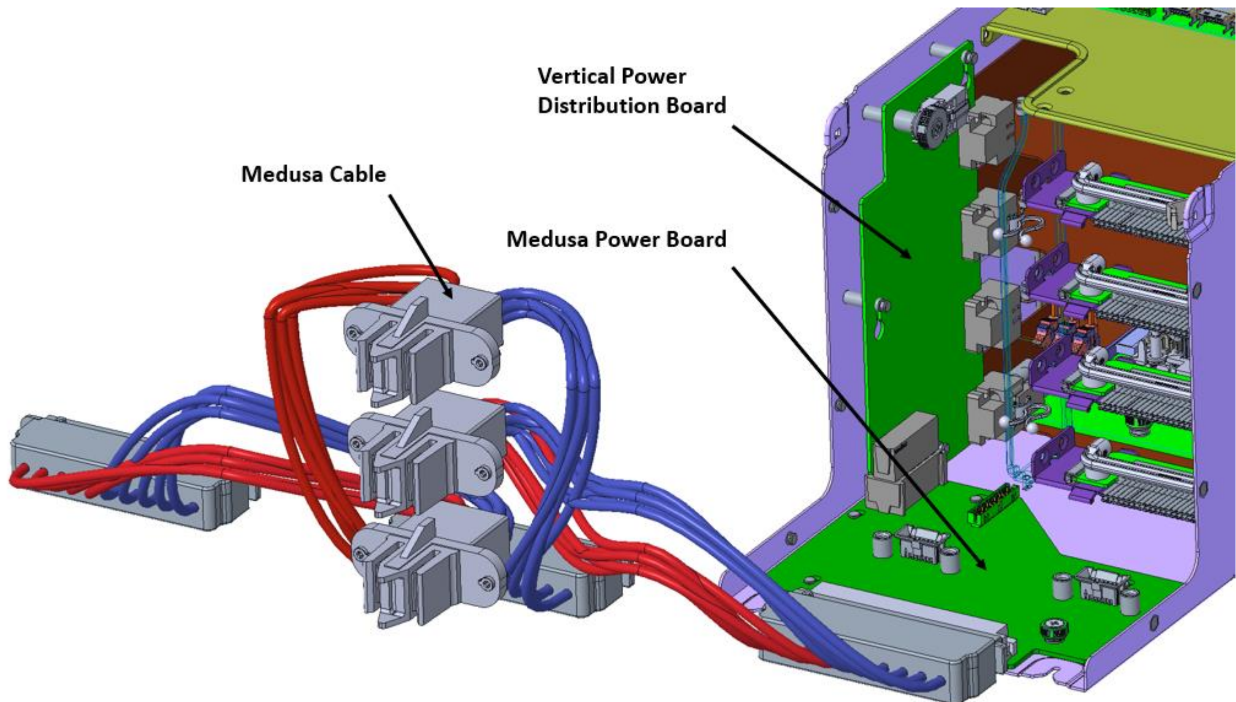


Figure 6-2: Input Power Delivery Topology

### 6.2.2 Baseboard

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

1. OOB access for remote management
    a. Over IPMI and USB to servers
    b. Over NC-SI and USB to NIC
2. Power management through power monitor and control to servers
3. Thermal management through temperature sensing and fan speed control
4. LED indication and button control
5. System event logger
6. Responsible for firmware validation and conduct updates.
7. FRU data tracking for the platform

The baseboard contains a slot for the OCP 3.0 NIC card, that allows different NIC cards with different PCIe lane config to be connected to the server blades. The BMC obtains its network access through the NIC card via NC-SI connection.

### 6.2.3 NIC Card Slot

Shown below is the section of the baseboard that connects to the OCP NIC 3.0. It has connections for the side band interface of OCP NIC 3.0 over NC-SI, as well as 4 x4 PCIe connectors to allow connection from servers to NIC card. Depending on the server's allocation, the design allows the server blades to connect to the Multi-Host NIC with PCIe lanes of x2 or x4.

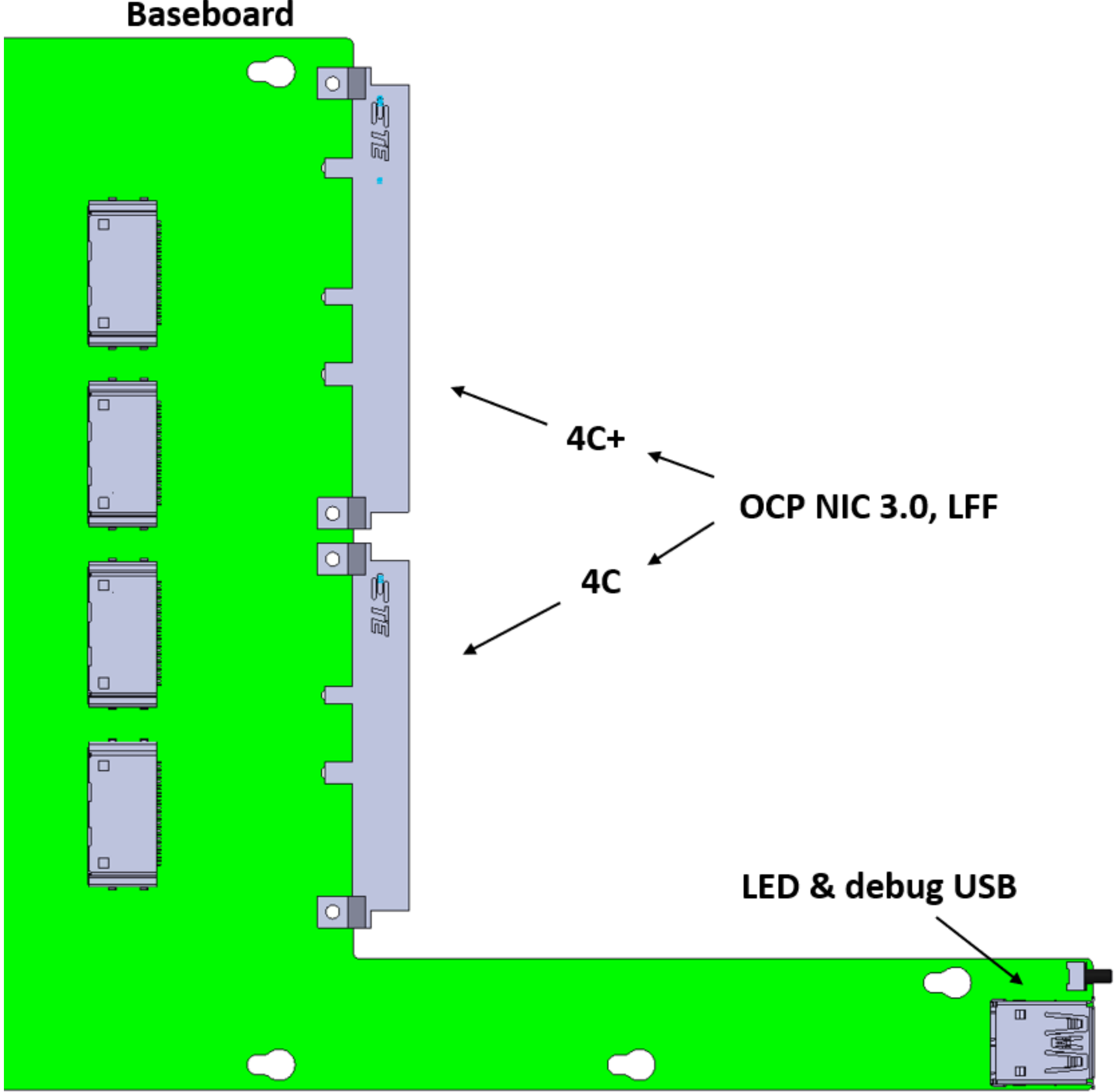


Figure 6-3: OCP NIC Interface

Support for Large Form Factor (LFF) NIC was removed from Yosemite V3 as there was no use case for it. The following picture shows the connection to the Multi-Host NIC on the Baseboard to the servers over Sled Management cable.

Each server blade is allocated an insertion loss of ~**12dB** (@4GHz) from the DIE of the chip to the connector on the server blade. Server blade designer may make trade-offs between using mid-loss or low-loss material in extending the connection reach or using a re-driver/re-timer solution but at the expense of power and space.

Figure 6-4: Connections to multihost OCP NIC

### 6.2.4 Sled Management Cable

The Sled Management cable consists of both low-speed signals and high speed PCIe signals that need to flow from the server blade to the baseboard for management and monitoring as well as the NIC. The low-speed signals on the cable are the side band signals like SMBus, UART, USB, alerts, power enable and fault signals.

### 6.2.5 Server Module

The server module contains the CPU subsystem. The module is to be CPU agnostic and be able to hot swap from the system. It has the following features:

1. Power is delivered to the server module from the vertical power distribution board through a dedicated power connector whereby

2. Depending on the PCIe signals available, the board may have x4 or x2 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 exist between the CPU, the CPU's internal management engine and the BMC.
    f. providing boot config information for the BIOS when system is booting up through determination of the expansion systems connecting to the server.
    g. User panel indicators like switch buttons and LEDs.
    h. allow 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.

The server module may have a front expansion board and a 2U expansion board. Note that the expansions may have different devices with different PCIe lane widths to allow broad usage options. The picture shown below is an illustration of how a server board would look like. Different server boards may take different shapes and sizes based on their needs.

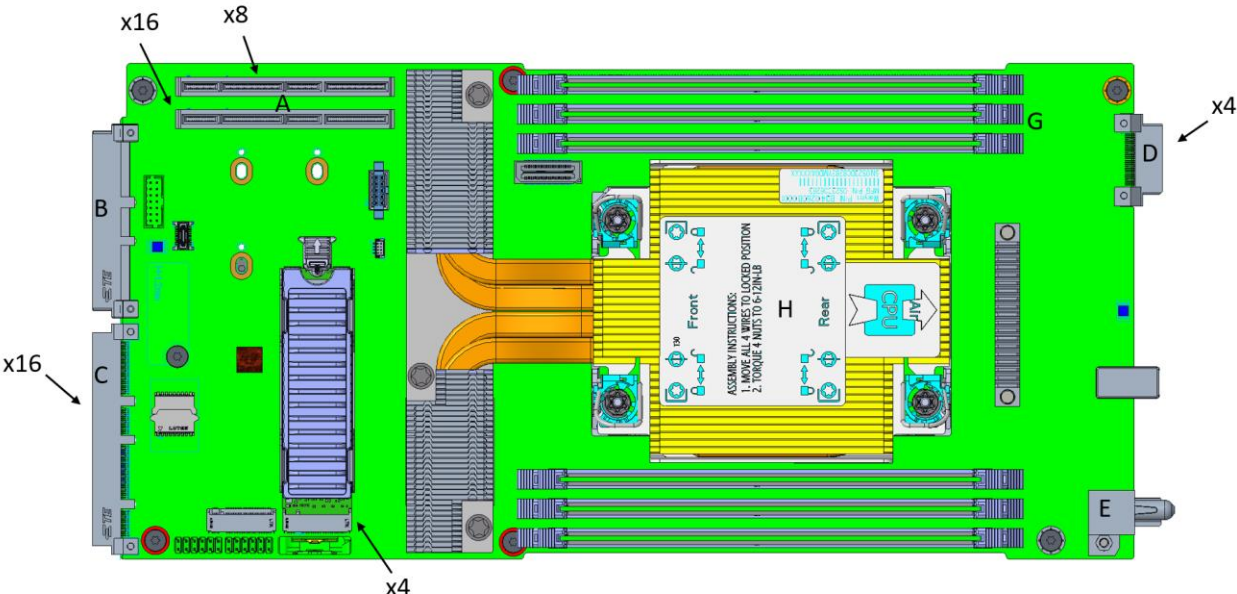


Figure 6-5: Server Board

When using the 2U expansion board, the server blade will take up 2 chassis slots to cater to better space and cooling solution for the system capabilities.

### 6.2.6 Expansion Options

Yosemite V3 platform enables several expansion board options. The expansion board enables YV3 to support multiple config Types (T3, T8, T15, T17., etc). The details are captured in the Delta Lake 1S Server Expansion Design Specification. Each expansion board has a BIC. The BOARD_ID setting for BIC tells what expansion board type it is.

**BOARD_ID**
**BOARD_ID[3:0]**
**0000------->DeltaLakeClass1**
**0001------->DeltaLakeClass2**
**0111------->BMC 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 Baseboard**

## 6.3 Yosemite V3 Platform system classes

Yosemite V3 platform defines 2 classes of system configuration.

### 6.3.1 Class 1

In class 1 system, the Baseboard has the BMC and connects to the multi-host NIC. This is the most common use case. The figure below shows an example of class 1 platform. With 4 PCIe lanes per server running at gen3, this configuration supports 25Gbps of NIC bandwidth per server. Please note that due to PCIe lanes being limited to 4 per server, this configuration cannot exceed 25Gbps NIC bandwidth per server even if one or more servers are de-populated.

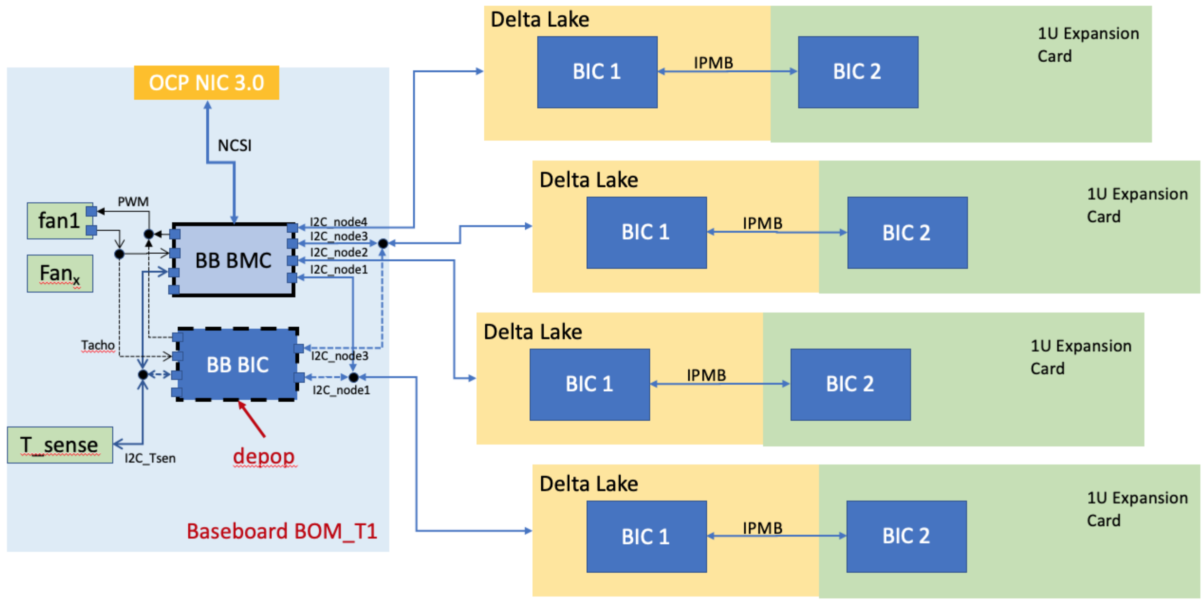


Figure 6-6: Class 1 Yosemite V3 Platform

The figure below shows the detailed connectivity for a class 1 system

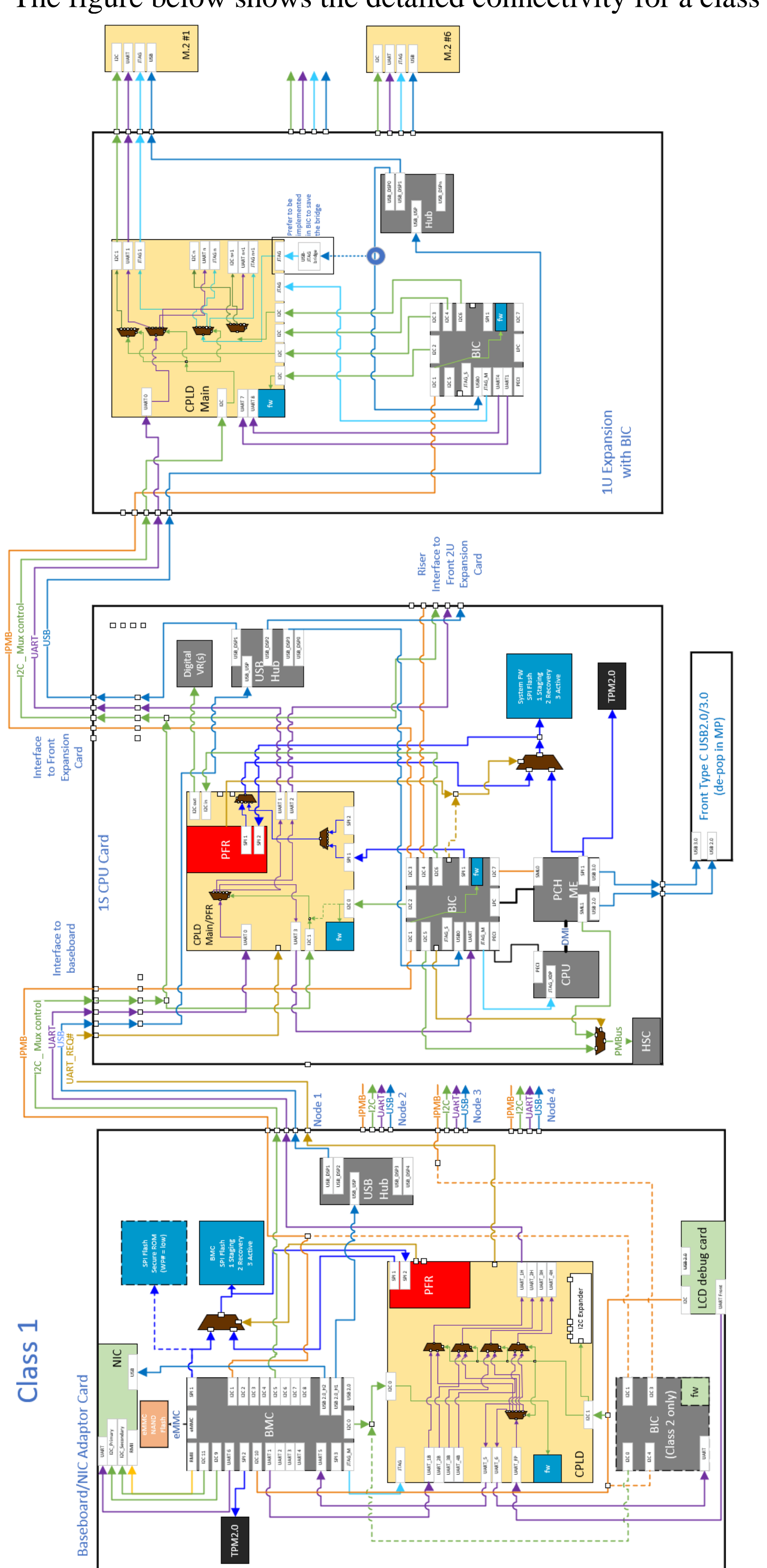


**Figure 6-7: Yosemite V3 class 1 system connectivity**

### 6.3.2 Class 2

In configurations that require more than 25Gbps bandwidth, a NIC can be directly connected to the server through NIC expansion card. In this configuration called class 2, the BMC is moved to the NIC expansion card. BMC is not shared between the servers but is dedicated to a server. The BMC on the Baseboard is de-populated and the functions like power and fan control are managed by BIC (Bridge IC) on Baseboard. The NIC Expansion card is detailed in OCP spec: Yosemite V3 Expansion Design Specification. The figure below shows an example of class 2 system. Class 2 configuration supports only 2 servers in the sled.

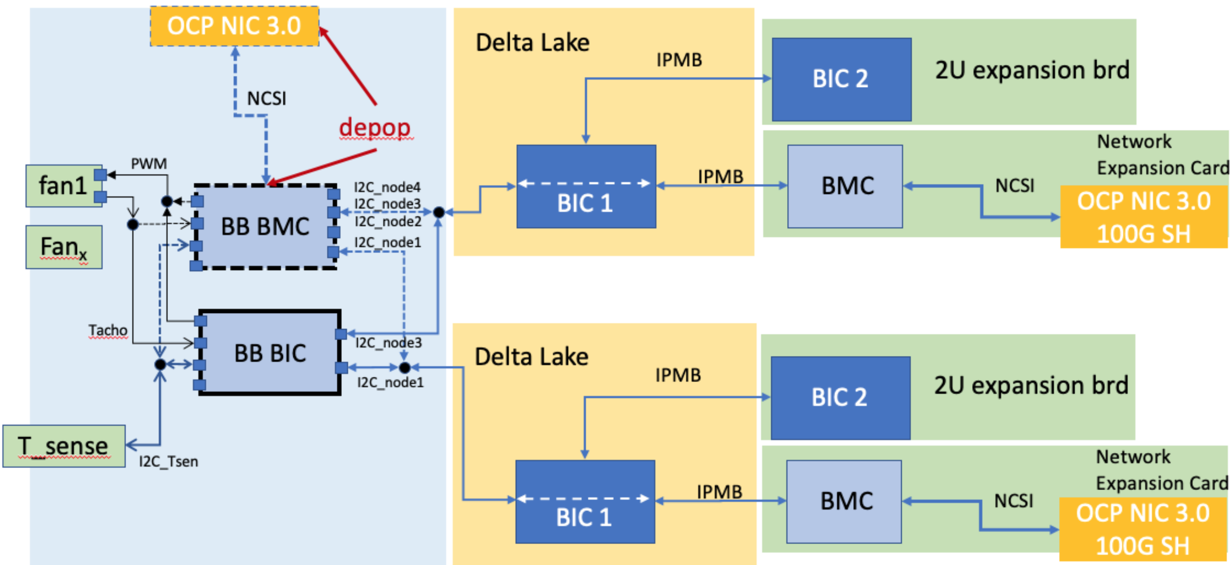


Figure 6-8: Class 2 Yosemite V3 Platform

The figure below shows the detailed connectivity for class 2 system

Figure 6-9: Class 2 Yosemite V3 Platform

Yosemite V3 Platform Power Delivery

Figure 6-10 shows a high-level illustration of the power delivery topology that the Yosemite V3 platform implements.

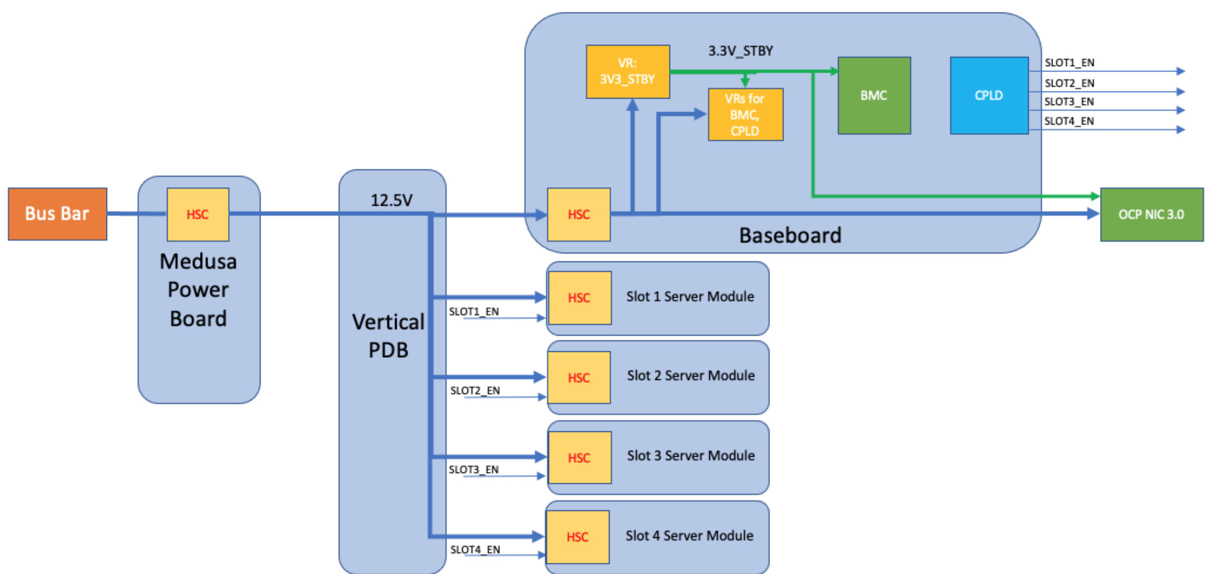


Figure 6-10: Yosemite V3 Platform Power Delivery Block Diagram

The Yosemite V3 platform continues to build upon a shared power distribution topology whereby multiple subsystems draw power from the same input source. However, a dedicated HSC shall be present on the baseboard and each 1S server blade design which deviates from previous Yosemite multi-node platforms. The aim of this change is to allow for greater design flexibility to support unique peak power delivery requirements depending on the intended use case. It should be noted that the baseboard's dedicated 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 platform is intended to support up to 1.5kW of distributed power between the different subsystems. Design consideration should be taken to ensure all aspects of the platform can be sufficiently cooled and without exceeding limits based on the power delivery hardware which is defined in section 12.3.

Shown in Figure 6-11, is a block diagram that highlights the various major power interconnects between the bus bar and Yosemite V3 subsystems.

Upon insertion of the Yosemite V3 chassis, the baseboard immediately obtains power 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.

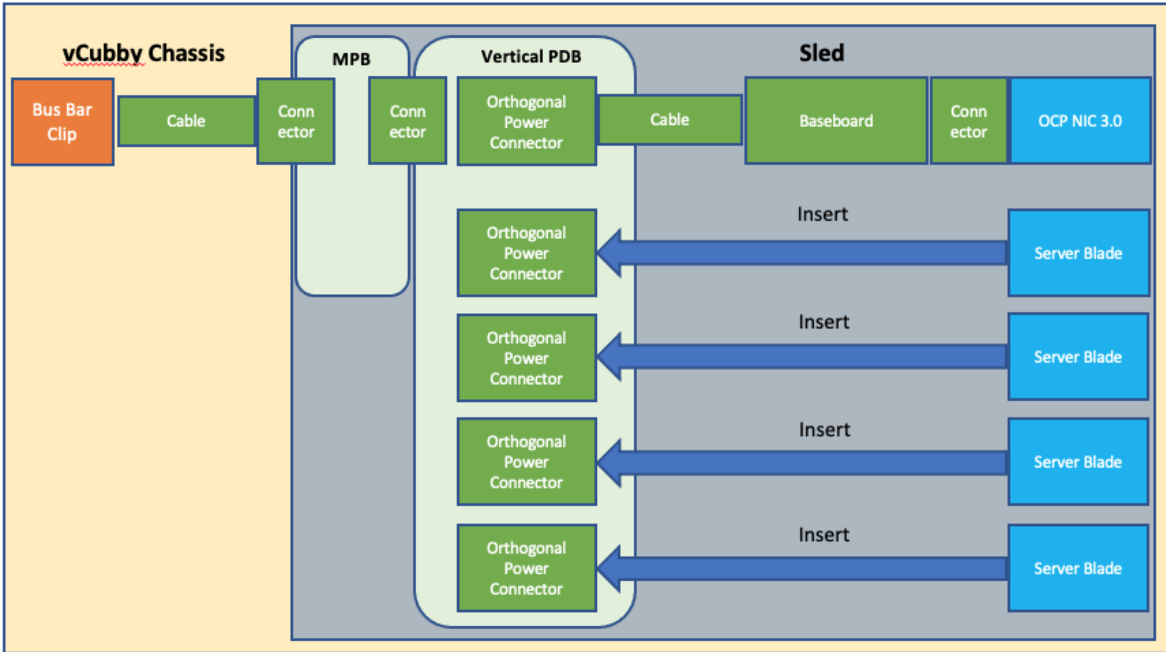


Figure 6-11: Yosemite V3 Power Interconnects

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 a defined 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.

The BMC uses standby rails: P3V3_STBY and P1V2_BMC_STBY and other rails: P1V15_BMC_STBY, P2V5_BMC_STBY, P0V6_DDR_VREF_STBY to power its circuits and DDR4 memory. 12.5V_STBY and P3V3_STBY shall be supplied to the OCP NIC 3.0 network card.

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 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).

## 6.4 SMBus Block Diagram

Figure 6-12 illustrates the Yosemite V3 Platform SMBus block diagram for a class 1 system. Figure 513 illustrates a class 2 system.
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 subsystem (servers/NIC) are behaving.
3. Out of band management of the device.
4. Firmware updates.

The BMC gains its network access through the NC-SI interface on the OCP NIC 3.0.
The BMC can access thermal sensors, the hot-swap controller and the FRU via a separate SMBus, as shown in figure below.

Each server could have expansion modules are connected to them. In such case, the connection from the baseboard to a server would reference to a connection as shown below where BICs on the server board need to relay message from the BMC to the BICs on the expansion modules.

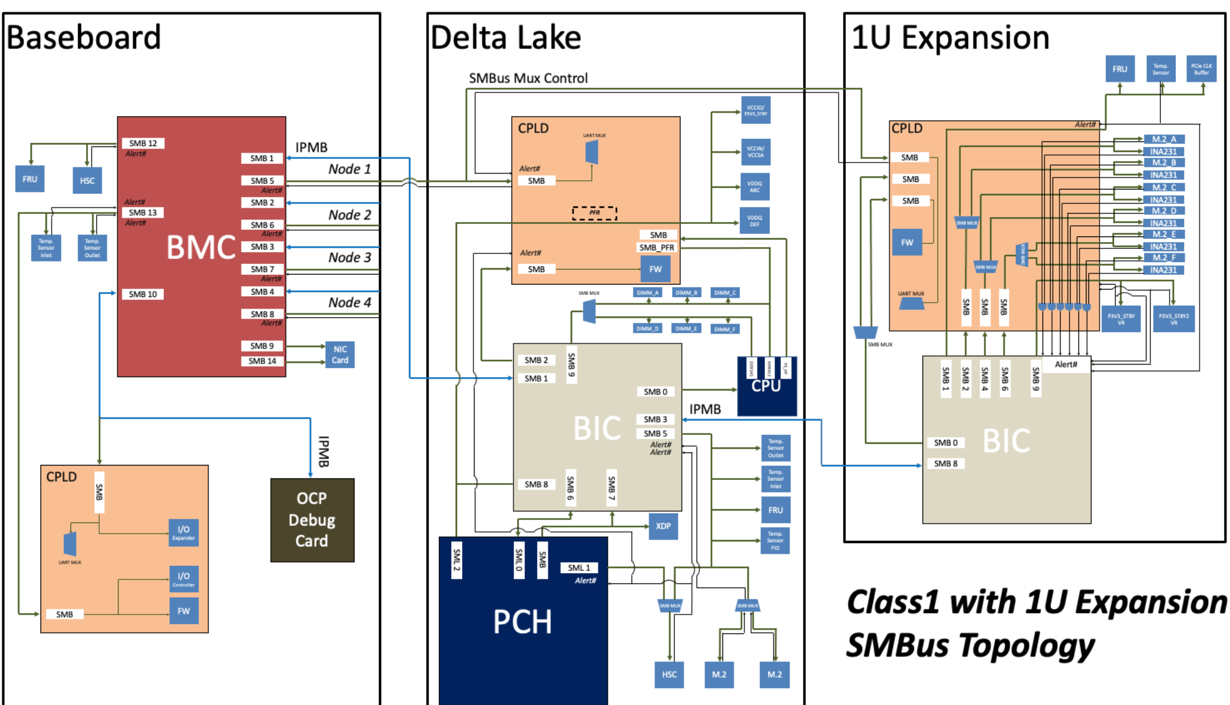


Figure 6-12: Yosemite V3 Platform SMBus Block Diagram

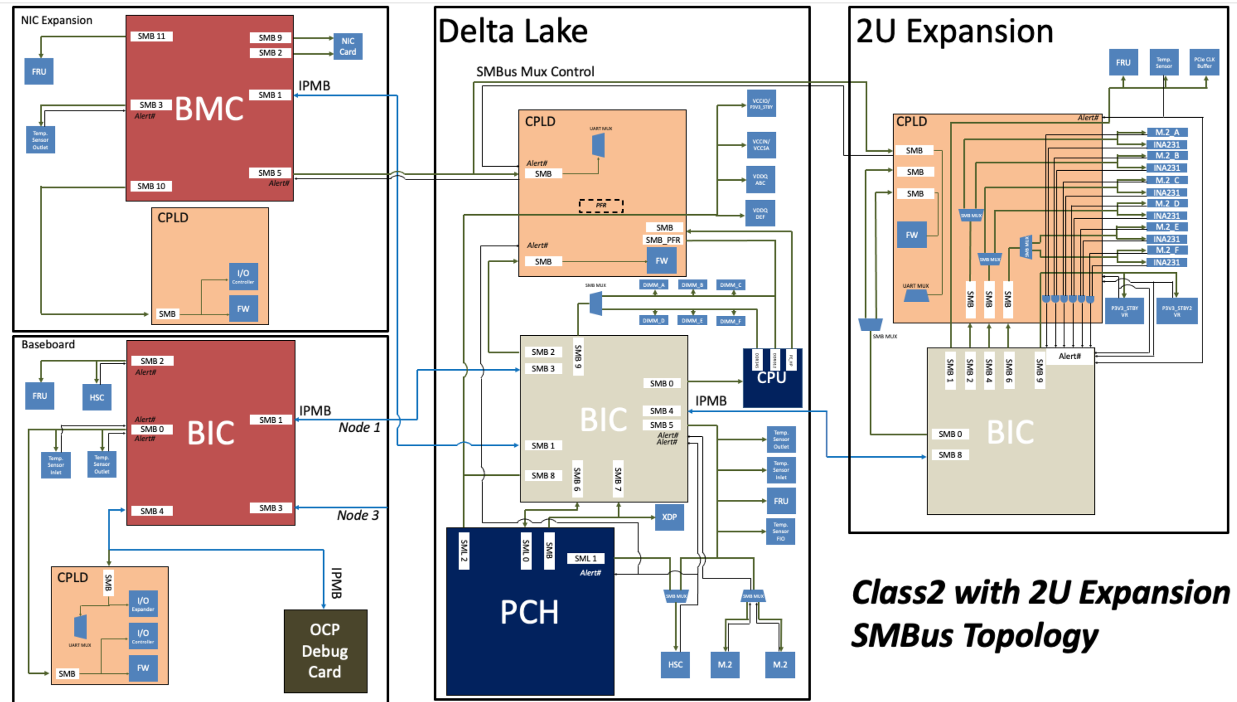


Figure 6-13: Yosemite V3 Platform SMBus Block Diagram

## 6.5 1S Server

### 6.5.1 Overview

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

### 6.5.2 1S Server Connectors

The server blade will be connected to the system through PowerBlade+ connector for power. For signals, it is using SFF-TA1002 straddle mount connector. The specific server should have at least a 1C straddle mount connector available for connection.

### 6.5.3 1S Server Slot Pinout Definition on Yosemite V3 Platform

Table 6-2 shows the set of signals for a x4 PCIe connection out of the 1C connector on server blade.

Table 6-1:Detailed Pin Definitions

| B1 | I2C_IPMB_SDA | BB_BIC_READY | A1 |
|---|---|---|---|
| B2 | I2C_IPMB_SCL | I2C_BIC_CPLD_SDA | A2 |
| B3 | GND | I2C_BIC_CPLD_SCL | A3 |

| B4 | AC_ON_OFF_BTN | I2C_BIC_CPLD_ALT_N | A4 |
|---|---|---|---|
| B5 | HS0_FAULT_N | GND | A5 |
| B6 | HSC_EN | PWRBTN_N | A6 |
| B7 | STBY_PWROK | RST_BMC_N | A7 |
| B8 | PCIE_RESET_N | RSVD | A8 |
| B9 | UART_REQ_N | SB_SLOT_ID1 | A9 |
| B10 | GND | SB_SLOT_ID0 | A10 |
| B11 | UART_RX | RSVD | A11 |
| B12 | UART_TX | MB_PRSNT_N | A12 |
| B13 | GND | GND | A13 |
| B14 | USB- | PERp3 | A14 |
| B15 | USB+ | PERn3 | A15 |
| B16 | GND | GND | A16 |
| B17 | PETp3 | PERp2 | A17 |
| B18 | PETn3 | PERn2 | A18 |
| B19 | GND | GND | A19 |
| B20 | PETp2 | PERp1 | A20 |
| B21 | PETn2 | PERn1 | A21 |
| B22 | GND | GND | A22 |
| B23 | PETp1 | PERp0 | A23 |
| B24 | PETn1 | PERn0 | A24 |
| B25 | GND | GND | A25 |
| B26 | PETp0 | REFCLKp0 | A26 |
| B27 | PETn0 | REFCLKn0 | A27 |
| B28 | GND | GND | A28 |

Table 6-2:Detailed Pin Definitions

| YV3_pinout_table Serverboard <-> Baseboard signals (class 1) | | |
|---|---|---|
| **Signal name** | **Direction** (In perspective of CPU card) | **Description (Class 1 BMC on Baseboard)** |
| I2C_IPMB_SDA | I/O | 1MHz IPMB between Baseboard management entity (BMC) to BIC on 1S Server Card |
| I2C_IPMB_SCL | I/O | 1MHz IPMB between Baseboard management entity (BMC) to BIC on 1S Server Card |

| | | |
|---|---|---|
| I2C_BMC_CPLD_SDA | I/O | 400KHz I2C from Baseboard to CPLD+BIC for commands |
| I2C_BMC_CPLD_SCL | I/O | 400KHz I2C from Baseboard to CPLD+BIC for commands |
| I2C_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 |
| HSC_EN | Input | Signal to enable Server Card HSC<br>Active High push pull.<br>PD at 1S Server Card side<br>Baseboard assert HSC_EN when:<br>1) 1S Server Card is fully inserted |
| HSC_FAULT_N | Output | HSC Fault signal; low active; OD with PU on baseboard. |
| UART_RX | Input | UART input to 1S Server Card<br>- Source is CPLD on Baseboard<br>- Destination is CPLD on 1S Server Card |
| UART_TX | Output | UART TX from Server Card<br>- Source is CPLD on 1S Server Card<br>- Destination is CPLD on Baseboard |
| PCIE_RESET_N | Output | PCIe reset from Server Card to Baseboard<br>Low active, Push-Pull,<br>3.3V_STBY domain |
| STBY_PWROK | Output | Standby Power OK of Server Card<br>High active, push-pull |
| PWRBTN_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 |
| AC_ON_OFF_BTN | Output | AC button on the server module |
| RSVDx | N/A | Spare signals between server blade and baseboard |
| REFCLK(n/p) | Output | Server output of 100MHz clock; 1 clock output |
| PET(n/p)X | Output | PCIe Gen3 RX of Server module. X ranges from 0 to 3 for Delta Lake |

| PER(n/p)X | Input | PCIe Gen3 RX of Server module. X ranges from 0 to 3 |
|---|---|---|
| UART_REQ_N | N/A | NC |
| SB_SLOT_ID0/1 | Input | Strap on connector chassis showing the Baseboard and Server board, which slot in the sled the server board and sled management cable are connected to |
| BB_BIC_READY | Input | Indicates to CPU card CPLD that BMC is ready on Baseboard |
| USB +/- | I/O | USB interface from BMC on Base board to hub on Delta lake to connect BIC on Delta lake and expansion cards. |

### 6.5.4 Server JTAG Access

To support system debug over JTAG interface, the Yv3 platform allows JTAG access of the devices on server and expansion boards through the BIC (Bridge IC) on those boards. USB and IPMI interfaces are used by BMC to communicate to BICs.

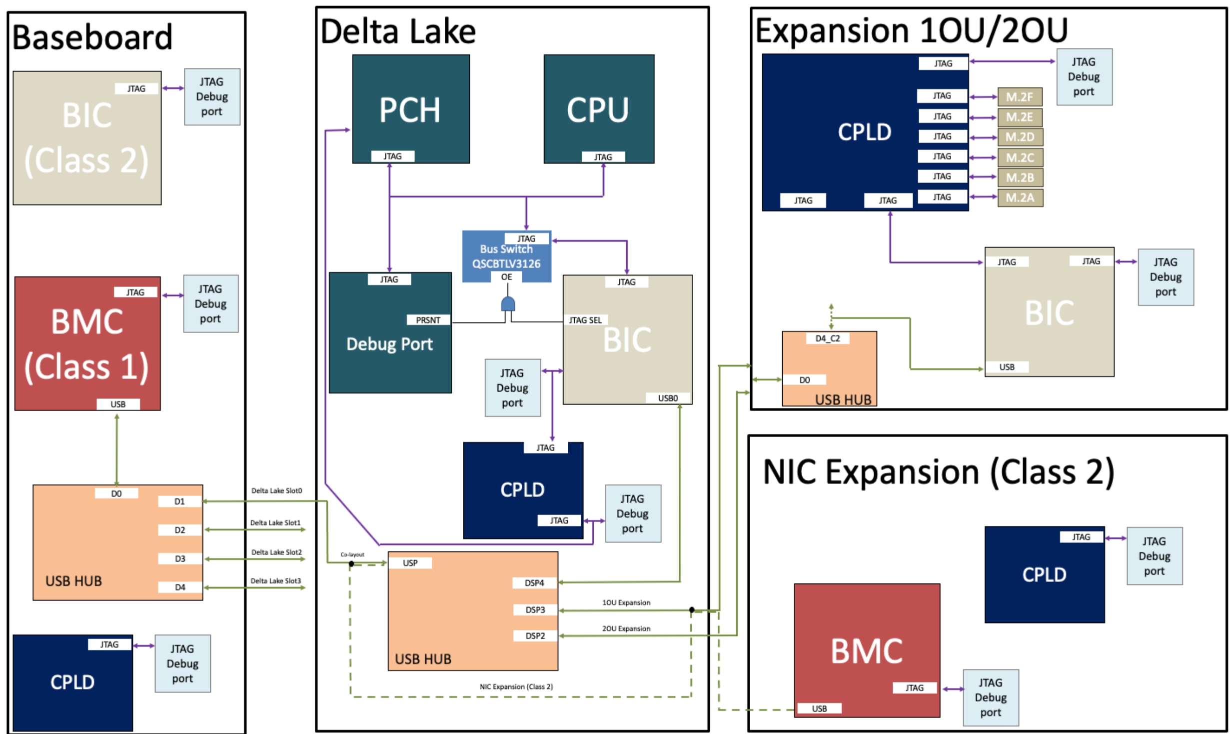


Figure 6-14: Yosemite V3 Server and Expansion Boards JTAG Block Diagram

# 7 Baseboard Management Controller

The Yosemite V3 Platform uses a BMC for various platform management services and interfaces with hardware and BIOS firmware. The proposed BMC is ASPEED's AST2520 given no video and PCIe is used.
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.

## 7.1 1S Server $I^2C$ Connections

There is a Bridge IC (BIC) on each 1S server as a satellite management controller. The Intelligent Platform Management Bus Communications (IPMB) ($I^2C$) connection from the Bridge IC on the 1S server to the BMC is the primary management interface for the 1S server. Each 1S server's $I^2C$ connection must be 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. A good start is to be as fast as 400kHz and preferred to be as fast as 1MHz.
In the new platform, there is an added I2C channel that connects the BMC to the server blade's CPLD and PCIe expansion connectors. This added channel is used to allow granular control from the BMC to the server blade's peripherals directly.
The $I^2C$ 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.

### 7.1.1 1S Server Command Interface

The BMC and the Bridge IC on the 1S server communicate with each other through the Intelligent Platform Management Bus (IPMB) protocol. For the added I2C channel to server CPLD, it would be I2C.

## 7.2 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. 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.

## 7.3 1S Server Discovery Process

### 7.3.1 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 has initiated the discovery process. The discovery sequence is defined as follows:

1. The BMC is to validate if the server card's standby power is OK.
2. The BMC can query the local CPLD to check the card type (passed through serial stream).
3. The BMC collects the FRU information from the Bridge IC.
4. The BMC sensor tables are updated from the Bridge IC.
5. The 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).

## 7.4 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 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 user removes and re-insert 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.

## 7.5 Network Interface

The BMC connects to the network through 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.

The BMC FW needs to support both IPv4 and IPv6.

## 7.6 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.

## 7.7 Local Serial Console and Serial-Over-LAN

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

- A local serial console on a debug header
- An 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.

## 7.8 Graphics and GUI

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

## 7.9 Remote Power Control and Power Policy

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

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 last state. The change of power policy should be supported and take effect without cold resetting the BMC firmware or rebooting the 1S server system.

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.

## 7.10 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 the 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 so that the POST process can be monitored remotely.

## 7.11 System LEDs and Buttons

The Yosemite v3 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 located at 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 Off/Service Action Allowed | | |
| System On/Status OK | | |
| System Off + Fault | | |
| System On + Locate | | |
| System Off + Locate | | |
| System On + Fault | | |

Legend:

Figure 7-1: Indicator LEDs

UART select function for OCP debug card is displayed on the corresponding server Amber LED.

For the AC cycle button, the intent is as follows:

1. Press for less than 4 seconds. No effect. This behavior is to ensure the button press is INTENTIONAL
2. Press for more than 4 seconds, sled will be AC cycled using the following procedure: Baseboard CPLD will pass signal to baseboard BMC/(BIC for class2), then baseboard BMC/(BIC for class2) will use I2C to inform baseboard HSC to do sled AC cycle. 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 for more than 4 seconds, AC power cycle will happen on the server card only.
2. Press for more than 8 seconds, server card will power down.
3. On a powered down server card, pressing the button for >1s, <4s will cause the server card to power up

## 7.12 Time Sync

Since the Yosemite V3 Platform Baseboard 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 the NTP server.

### 7.12.1 NTP Time Sync Flow

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

## 7.13 Power and Thermal Monitoring, and Power Limiting

The BMC firmware shall support platform power monitoring. Enabling power monitoring for the 1S servers requires an accurate power sensor on 12.5V to the 1S server. This function should be able to access through in-band and OOB.
The BMC firmware shall support thermal monitoring, including 1S server SOCs, 1S server memory, and inlet/outlet air temperatures. To ensure accuracy, a TI TMP75 or similar part 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 shall support 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 are tapping power from the vertical power bar, BMC is to periodically query the following

1. The total current flowing through the HSCs of the SLED. The upper threshold is set to 113.4A. Upon seeing higher than expected current, BMC should initiate a system wide throttle to all the servers.
2. Power, voltage, current and temperature around the system (Baseboard and the server blades).
3. BMC needs to handle cases like HSCs not responding or timeout issues intelligently while at the same time provide a good methodology to ensure data coming into the system is trustworthy and can be used to mathematically sum for a result.

## 7.14 Sensors

Both analog and discrete sensors may reside on the baseboard and on the cards. The BMC must provide a way to read all sensors across 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 log the anomaly at the same time.

### 7.14.1 Analog Sensors

The BMC has access to all analog sensors on the Yosemite V3 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

### 7.14.2 BIOS/ME generated Sensors

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

- Firmware health
- POST errors
- Power errors
- ProcHOT
- Machine Check errors
- PCIe errors
- Memory errors, etc.

## 7.15 Event Log

The vendor should implement the BMC to support 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.

### 7.15.1 Logged Errors

#### *7.15.1.1 CPU Error*

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

#### *7.15.1.2 Memory Error*

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

#### *7.15.1.3 PCI-e Error*

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

#### *7.15.1.4 POST Error*

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

#### *7.15.1.5 Power Error*

Two power errors should be 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.

#### *7.15.1.6 MEMHOT# and SOCHOT#*

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

#### *7.15.1.7 Fan Failure*

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

#### *7.15.1.8 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 above error logging and reporting, the user may select to enable or disable each logging option.

#### 7.15.2 Error Threshold Setting

Enable the error threshold setting for both correctable and uncorrectable errors. Once a programmed threshold is reached, the system should trigger an event and log it.

- Memory Correctable ECC: Suggest setting the threshold value to be [1,000] in the mass production stage and [1] for the evaluation, development, and pilot run stage, with options of 1, 4, 10, and 1,000. BIOS could also have the options with range 1 to 32767. 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. BIOS could also have the options with range 0 to 32767.
- PCIe Error: Follow the 1S server vendor's suggestion.

### 7.16 Fan Speed Control in BMC

The vendor should enable Fan Speed Control (FSC) 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.

#### 7.16.1 Fan Speed Control Specification

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

#### 7.16.2 Data gathering for FSC

The BMC needs to gather data as input of the FSC. The required data is described in the table below.

Table 7-1: 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 |

#### 7.16.3 Fan Speed Controller in BMC

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

### 7.16.4 Fan Connection

The fan modules connect to the medusa board through a floating blind mate connector. The Yosemite V3 Platform baseboard has a fan header which connects to the medusa board through a cable.

### 7.16.5 Fan Tray

The system has a cold-swap fan tray, which is comprised of 2x 80mm fans + a cable set to blind-mate interface with the baseboard.

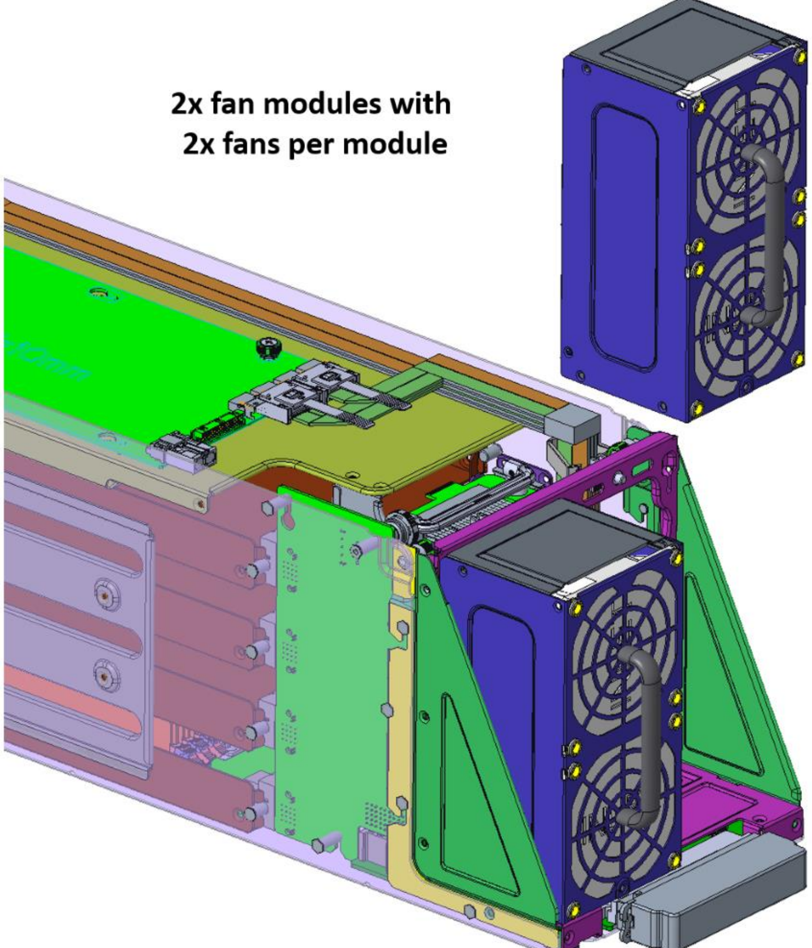


Figure 7-2: Yosemite Fan modules

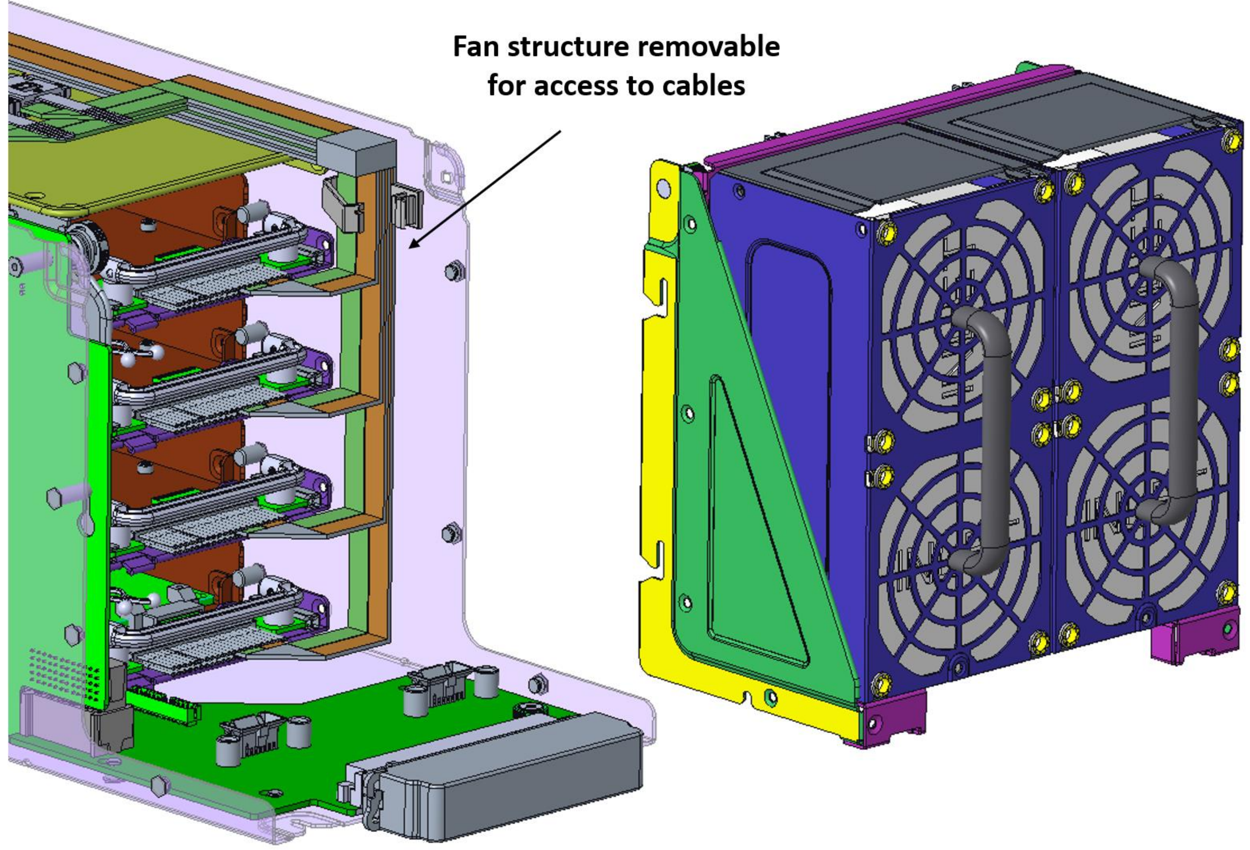


Figure 7-3: Yosemite V3, Cable Access

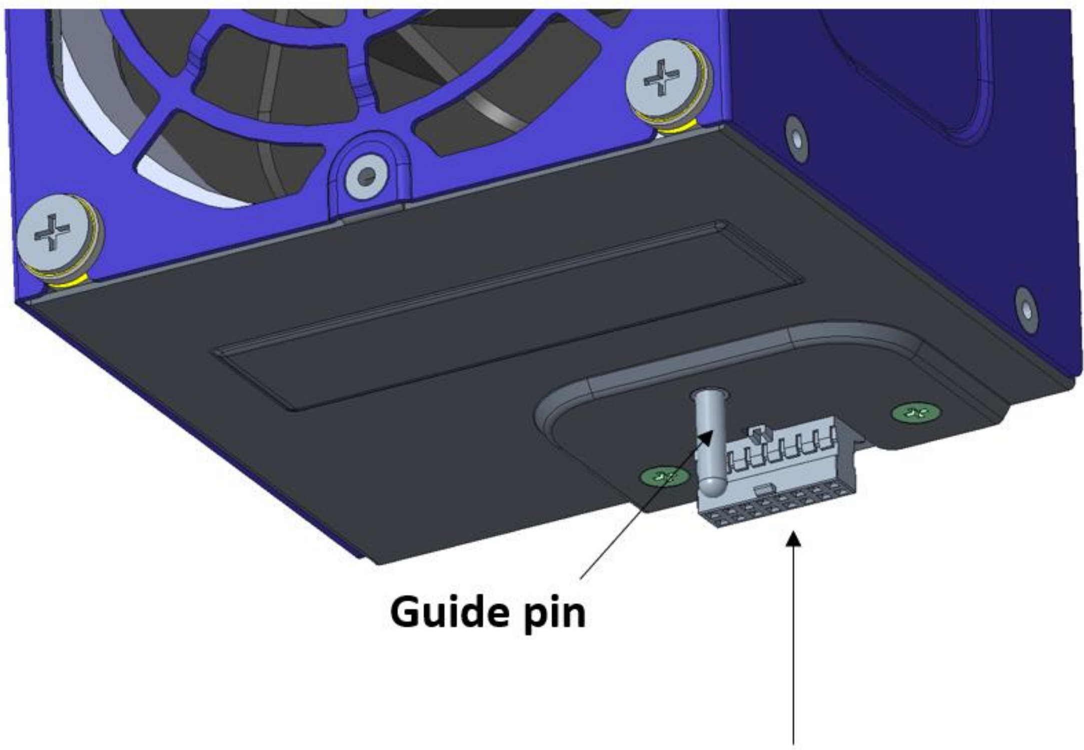

Figure 7-4: Yosemite V3, Fan Connector

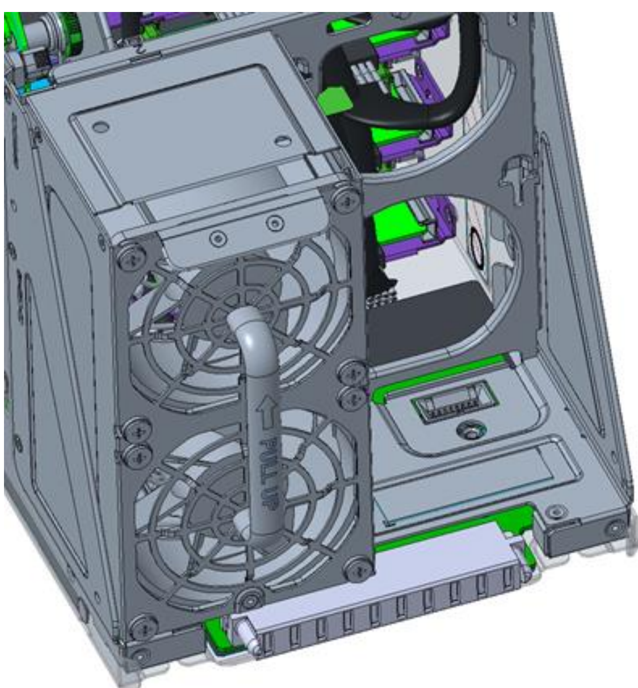

Figure 7-5: Yosemite V3, Fan Connector

Figure 2-5: Yosemite V3, Fan modules

## 7.17 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 (CentOS) via the data network. Tool(s) shall support CentOS.
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 to 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 should 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.

## 7.18 Hot Service Support

The Yosemite V3 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. For example, sensor monitoring might need restart for that slot to reflect the new hardware.

## 7.19 OpenBMC

OpenBMC refers to 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.

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/OpenSystemFirmware/Facebook/YosemiteV3

## 7.20 Security

The implementation of the BMC on the baseboard needs to have a path to allow early PFR evaluation. This would involve a PFR footprint ready CPLD and the necessary connections to the boot flashes involved. The implementation should also have readiness on I2C paths as well. The implementation should be transparent to allow current verified boot process and I2C access prior to availability of the PFR chip.

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.

## 7.21 Small Form Factor Baseboard Storage Module (SFF BSM)

The Baseboard on Yosemite V3 platform shall support a pluggable storage module: Small Form Factor Baseboard Storage Module (SFF BSM). SFF BSM is a M.2 connector pluggable module that has the two BMC Flash chips and eMMC Flash to store the operating system and logs. The purpose of separating the Flash from the Baseboard and mounting it on a pluggable card is to enable ERAD. Since the eMMC will store the log information, it needs to be easily removed and destroyed for security. The SFF BSM spec. has the details of the module board design.

The exact configuration of the components (Flash and eMMC) is dependent on the final test and acceptance of eMMC flash.

The current generation of Yosemite V3 does not mount the eMMC part on the SFF BSM. So, it contains only the BMC NOR Flash.

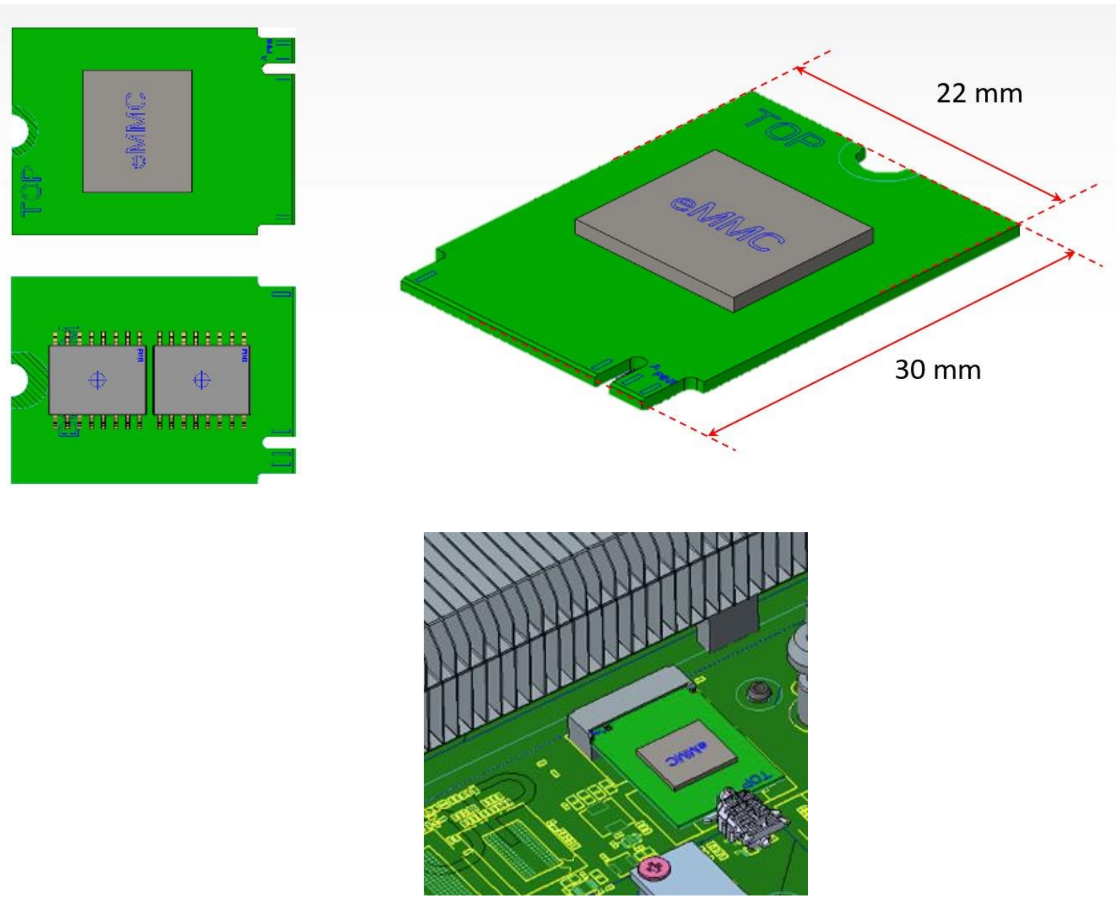


Figure 7-6: Yosemite SFF-BSM

# 8 System Firmware

All products seeking OCP Accepted™ Product Recognition must complete the Open System Firmware (OSF) Tab in the 2021 Supplier Requirements Checklist.
The completed checklist shall be uploaded and available at:
https://github.com/opencomputeproject/OpenSystemFirmware/Facebook/YosemiteV3

# 9 Mechanical

The Yosemite V3 Platform is an Open Rack V2 compatible compute platform which consists of a Yosemite V3 chassis which supports 3 Yosemite V3 sleds in a 4OU space. Each Yosemite V3 sled can hold up to 4 server blades.

## 9.1 Yosemite V3 Chassis

Yosemite V3 chassis is a power-mechanical chassis distributing power from the rack bus bars to the three sleds.

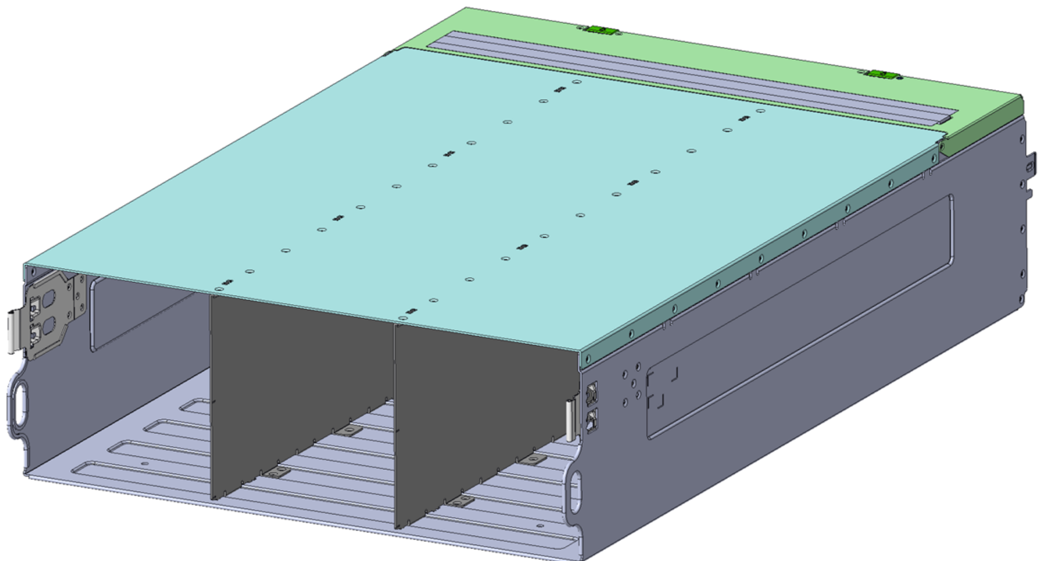

Figure 9-1: Yosemite V3 Chassis

## 9.2 Yosemite V3 Sled

A sheet metal and plastic sled serves as the mechanical interface between the Yosemite V3 Platform and the Yosemite V3 sled. It also provides mechanical retention for the components inside the sled such as the power cable assembly, fan, baseboard, and server cards. The combination of sheet metal tray, baseboard, adapter card, multi-host NIC (OCP NIC 3.0 LFF), and PDB is a Yosemite V3 Platform sled.

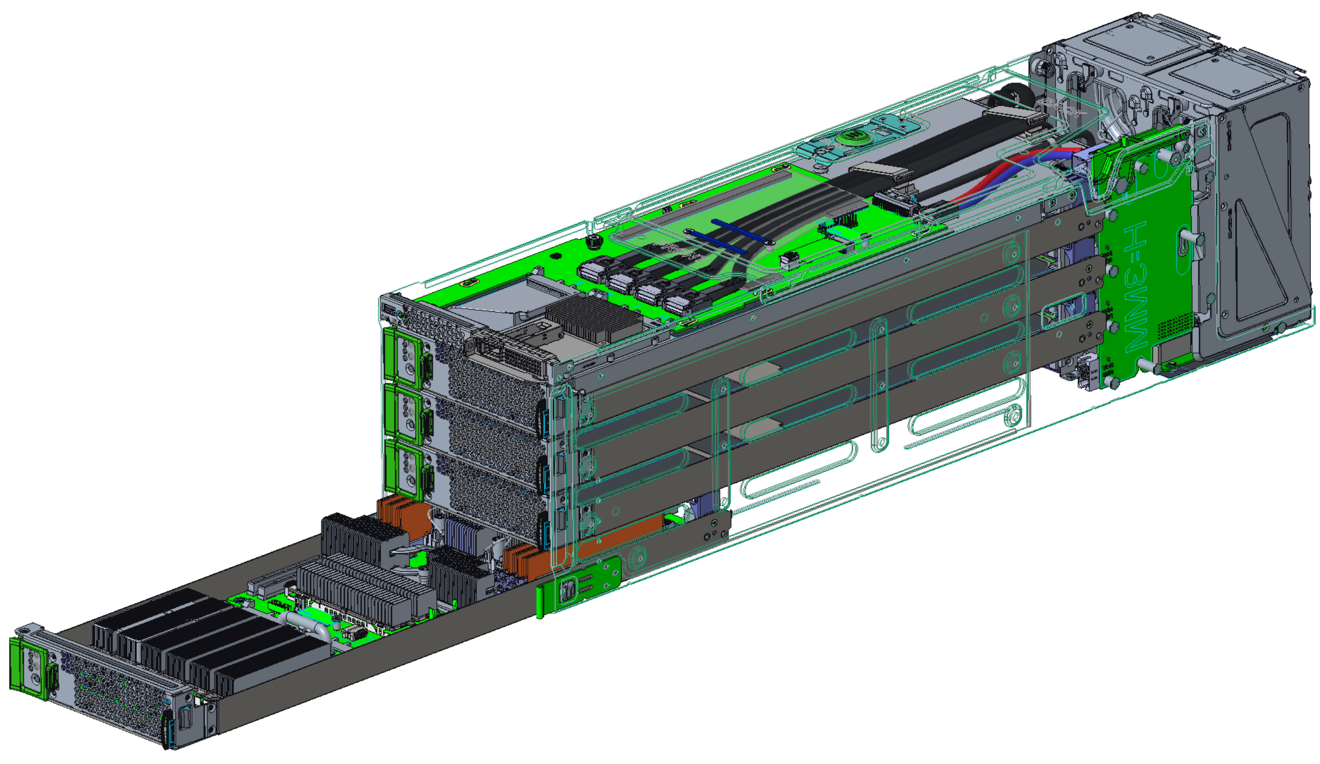


Figure 9-2: Yosemite V3, Sled Populated with 4x 1S Server

### 9.3 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. The 1U configuration can be seen below and is a height of 40.8mm. The 2U configuration is a height of 82.2mm. The blade design is specified in the OCP 1S Server Design Specification.

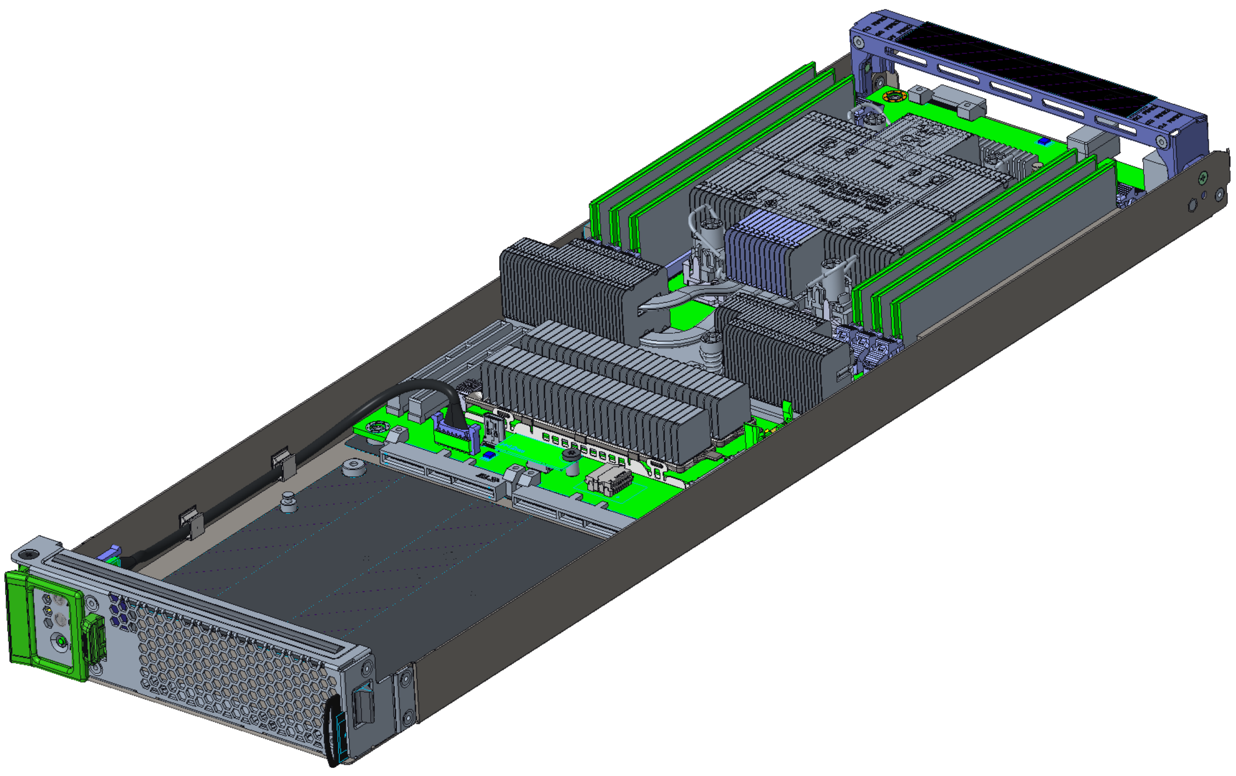

Figure 9-3: Yosemite V3 Blade

## 9.4 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 in Section 12.

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

## 9.5 Retention

When the sled is in its "home" position within the chassis, a rotating combination pull/push handle engages with the side of the chassis to ensure it is held firmly in place. When the blade is fully installed into the sled, the CPU latch rotates into position to lock it into place. To remove the CPU blade, the blade latch can be deflected toward the CPU handle then the handle can be pulled. The 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.

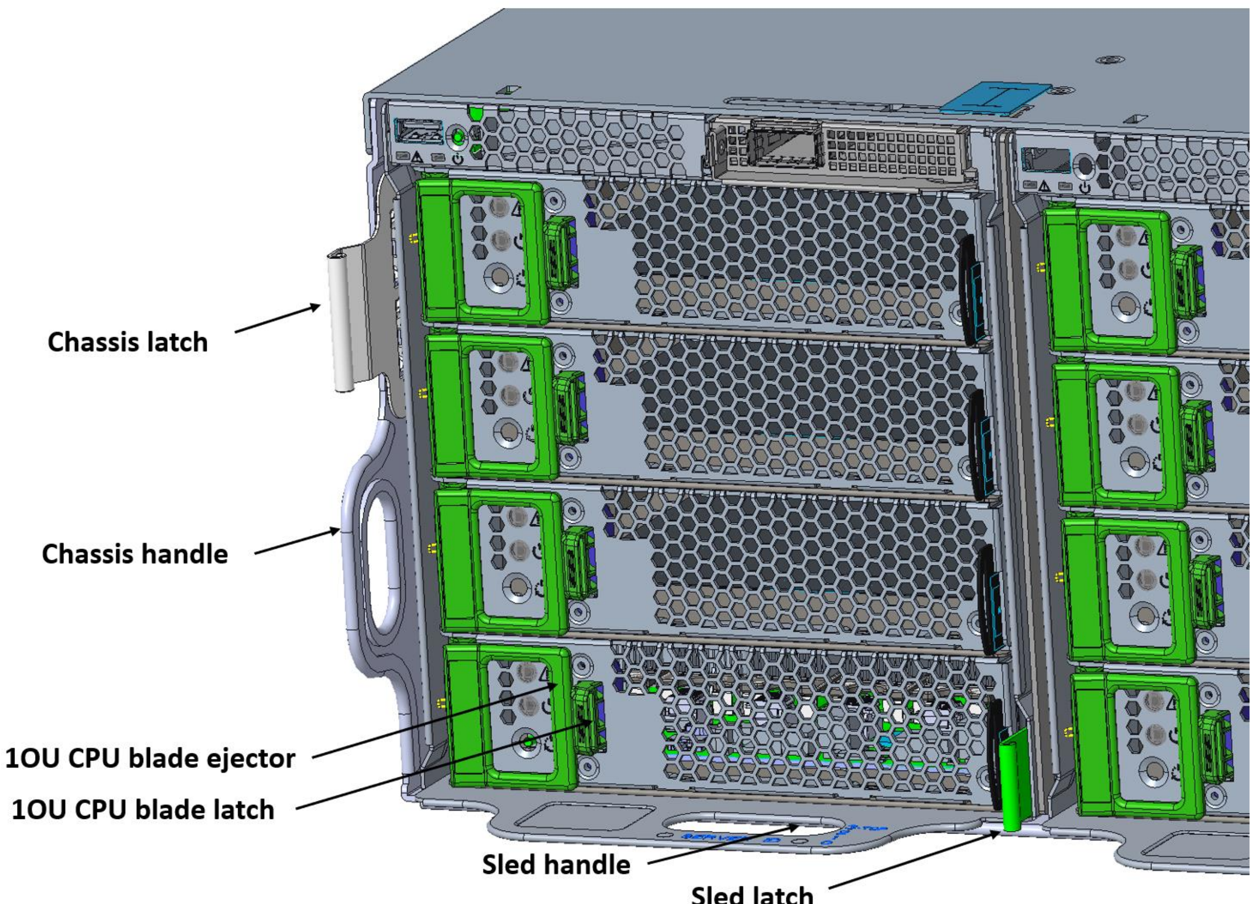


Figure 9-4: Sled Retention, 1OU Blades

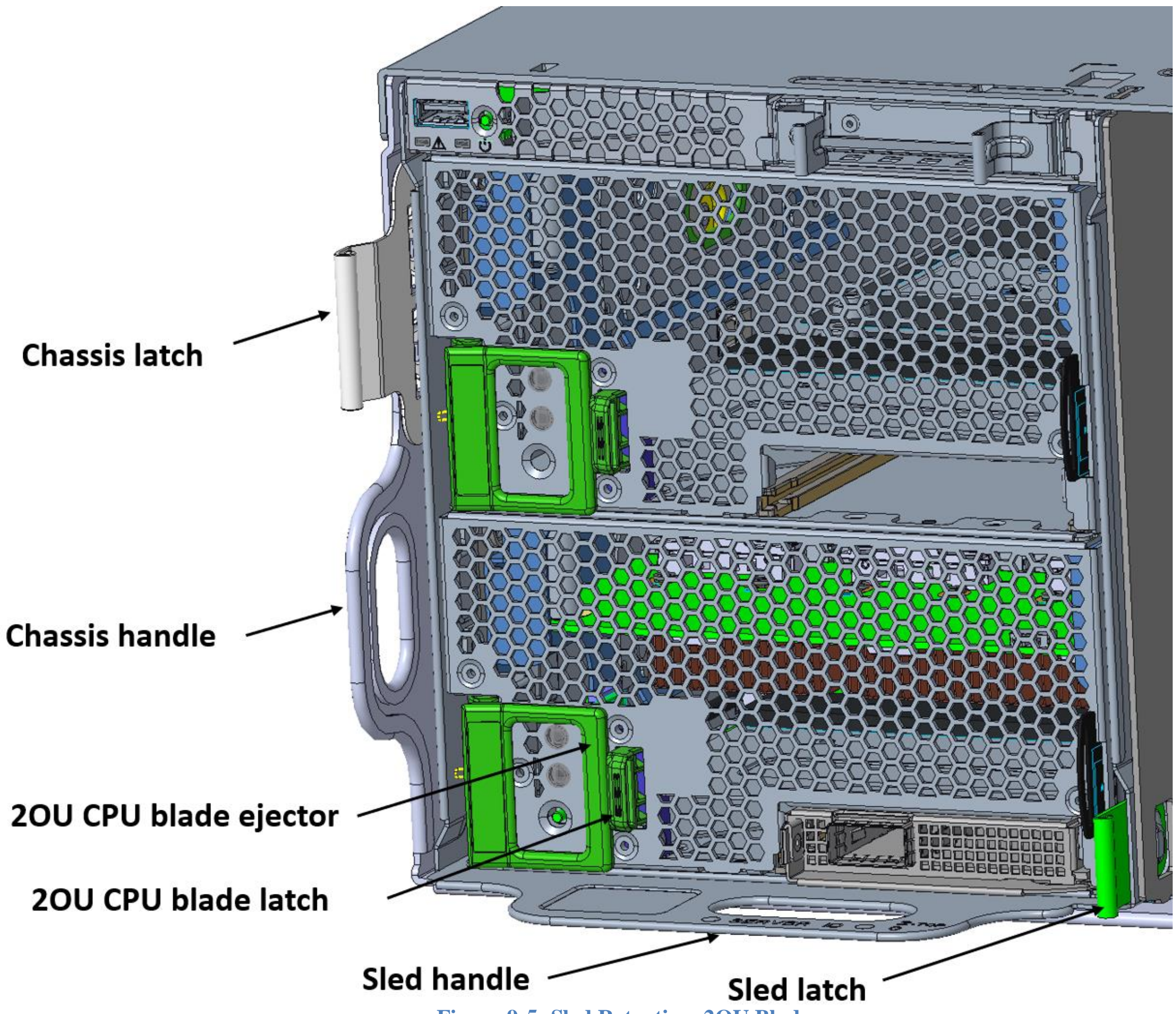


Figure 9-5: Sled Retention, 2OU Blades

# 10 Thermal

To meet thermal reliability requirement, the thermal and cooling solution should 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. Thermal solution should not allow any overheating issue for any components in system. CPU or memory should not throttle due to any thermal issues under the following environments.

- Inlet temperature lower than or equal to 35°C, and 0-inch H2O datacenter pressure with all FANs in each thermal zone running properly
- 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 failed

## 10.1 Data Center Environmental Conditions

This section outlines Facebook data center operational conditions.

### 10.1.1 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.

### 10.1.2 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 must maintain a temperature below its maximum specification temperature in the cold aisle.

### 10.1.3 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.

### 10.1.4 Relative Humidity

Data centers usually maintains a relative humidity between 20% and 90%. The thermal solution must sustain uninterrupted operation of the system across the aforementioned RH range.

## 10.2 Server Operational Conditions

### 10.2.1 Inlet Temperature

The inlet air temperature will vary. The cooling system in the Yosemite V3 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 Facebook operational condition but is 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.

### 10.2.2 Pressurization

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

### 10.2.3 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 Platform is preferred when the system is operating under normal conditions.

### 10.2.4 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. 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.

### 10.2.5 System Airflow or Volumetric Flow

The unit of airflow (or volumetric flow) used for this spec is cubic feet per minute (CFM). The 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:

$$\text{Thermal Expenditure} = \frac{\text{System airflow}}{\text{Total system power consumption, including fans}} \quad [\text{CFM/W}]$$

At sea level, the maximum allowable airflow per watt in a Yosemite V3 rack is 0.13 at 30 °C inlet temperature under the normal load or 9kW rack power. The cooling solution in the system level should consider 20% reduction due to the TOR and PSU. The desired airflow per watt is 0.1 or lower in the system at the mean temperature (plus or minus standard deviation).

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.

### 10.2.6 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).

### 10.2.7 Thermal Sensor

The maximum allowable tolerance of thermal sensors in the Yosemite V3 Platform is ±2°C.

### 10.2.8 System Loading

The power consumption of individual components in the system motherboard varies by use. The total power consumption of the whole Yosemite V3 Platform also may vary with use. Please see the summary below.

- System loading: idle to 100%
- OCP NIC 3.0: SFF (25W) and LFF (45W)

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.

### 10.2.9 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 4°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.

## 10.3 Thermal Kit Requirements

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

### 10.3.1 Heat Sinks

Heat sinks must have a thermally optimized design at the lowest cost. There must be no more than four heat pipes in the heat sink. Installation must be simple and uncomplicated. Heat sinks

must not block debug headers or connectors. The heat sink design for 1U use case and 2U use case can be different to maximize the thermal efficiency.

### 10.3.2 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 of one fan (or one rotor) failing, 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.

# 11 I/O System

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

## 11.1 Front facing Server Modules

The Yosemite V3 Platform has a 4 slot system for server modules. It is possible for certain configurations that 1 module may occupy 2 slots.

## 11.2 Network

### 11.2.1 Data Network

The Yosemite V3 Platform uses an OCP NIC 3.0 at the front panel as its primary data network interface.

### 11.2.2 JTAG Network

JTAG interface is used for debug. The Yosemite v3 platform allows JTAG access of the devices on server and expansion boards through the BIC (Bridge IC) on those boards. Please refer to Server JTAG Access

To support system debug over JTAG interface, the Yv3 platform allows JTAG access of the devices on server and expansion boards through the BIC (Bridge IC) on those boards. USB and IPMI interfaces are used by BMC to communicate to BICs.

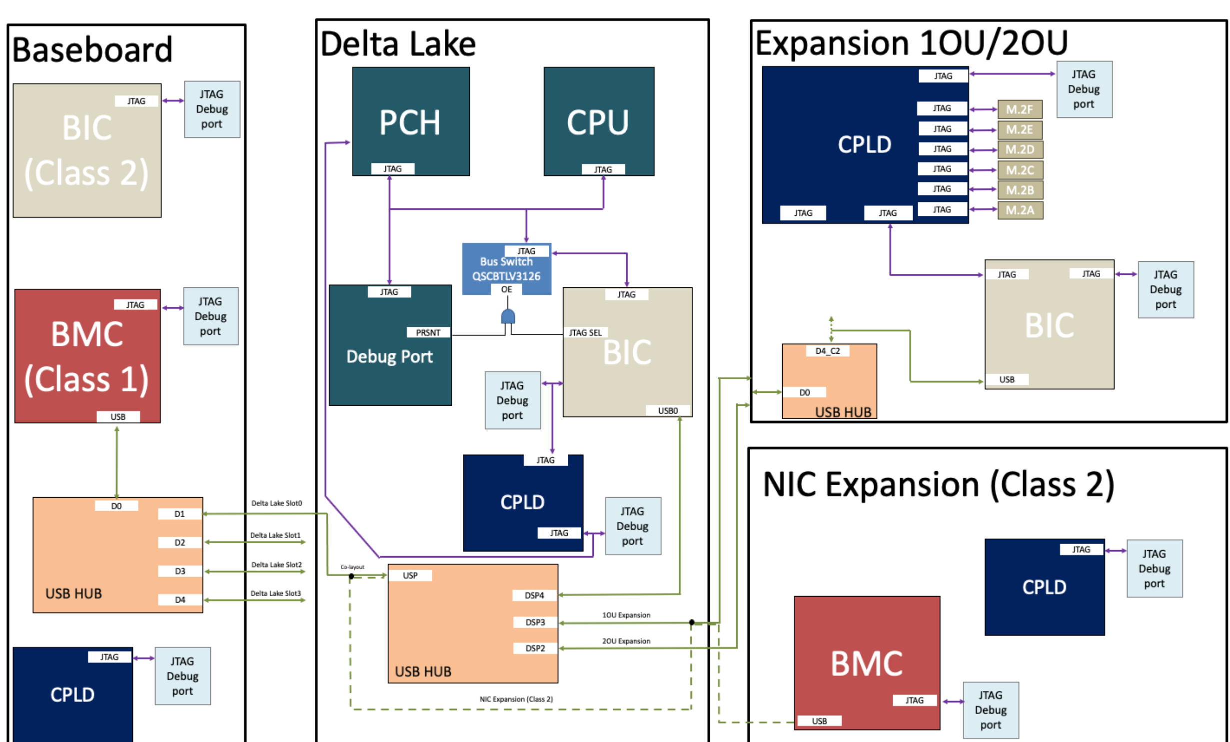


Figure 6-14: Yosemite V3 Server and Expansion Boards JTAG Block Diagram

in Section 5 for more details.

#### 11.2.3 Management Network

The management network on the Yosemite V3 Platform uses the sideband of the network controller of the data network, either SMBus or NC-SI interface. Please refer to Section 5 and 6 for more details.

### 11.3 Server Slots Assignment

The server slot assignment for Yv3 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 as shown on the right in a system with 4 servers.

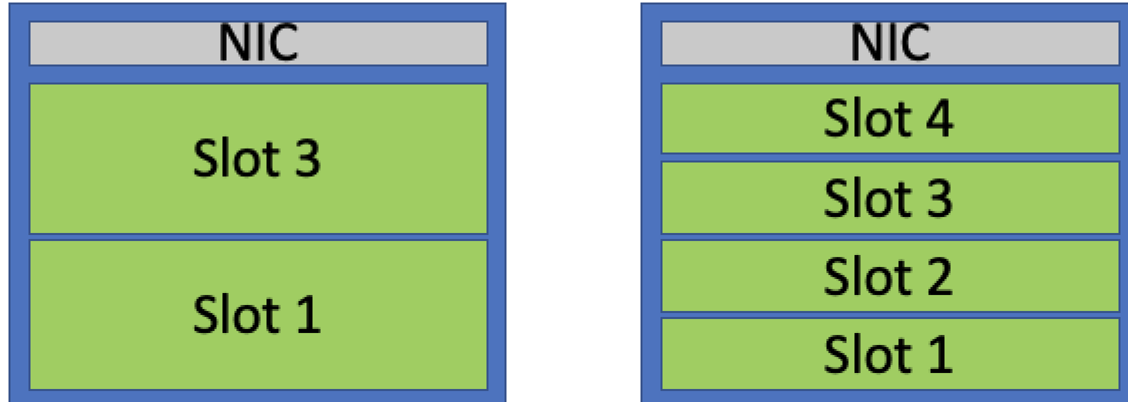


Figure 11-1: Server Slots, Front View

### 11.4 Front Panel

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

#### 11.4.1 Power Button and LEDs

Pls refer to System LEDs and Buttons in Section 6.

#### 11.4.2 OCP debug USB Connector

The Yosemite V3 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 regarding the relevant host and their POST codes/SEL/config are displayed through 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. Although a USB connector is used, the Yosemite V3 platform does not have a USB signal going to that port. The OCP debug card uses the SMBus and UART interface with BMC.

### 11.4.3 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.
The table below 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 11-1: 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 Intel 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.

## 11.5 Fan 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.

Table 11-2: Fan Connector Pin Definition (TBD)

| Pin | Description |
|---|---|
| **1** | Second fan’s PWM input |

| | |
|---|---|
| **2** | First fan's PWM input |
| **3** | Second fan's TACHO output |
| **4** | First fan's TACHO output |
| **5** | Second fan's power 12V |
| **6** | First fan's Power 12V |
| **7** | GND |
| **8** | GND |

# 12 Power

## 12.1 Input Voltage Level

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 tolerance range between 11.25V and 13.75V.

## 12.2 48V support

ORV3 racks are under development which provides 48V input. The Yosemite V3 platform is designed to be able to support the ORV3 with dedicated designed Medusa board and VPDB. Future validation will be conducted when ORV3 racks become available.

## 12.3 Platform Power Budget

The Yosemite V3 platform shall be designed to support a maximum sustained 1.5kW of distributed power among its subsystems in a sled. 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. Depending on the Yosemite V3 sled configuration, the total power budget is governed by the current carrying capability of various connectors as shown in
Figure 12-1 below and limits are summarized in Table 12-1

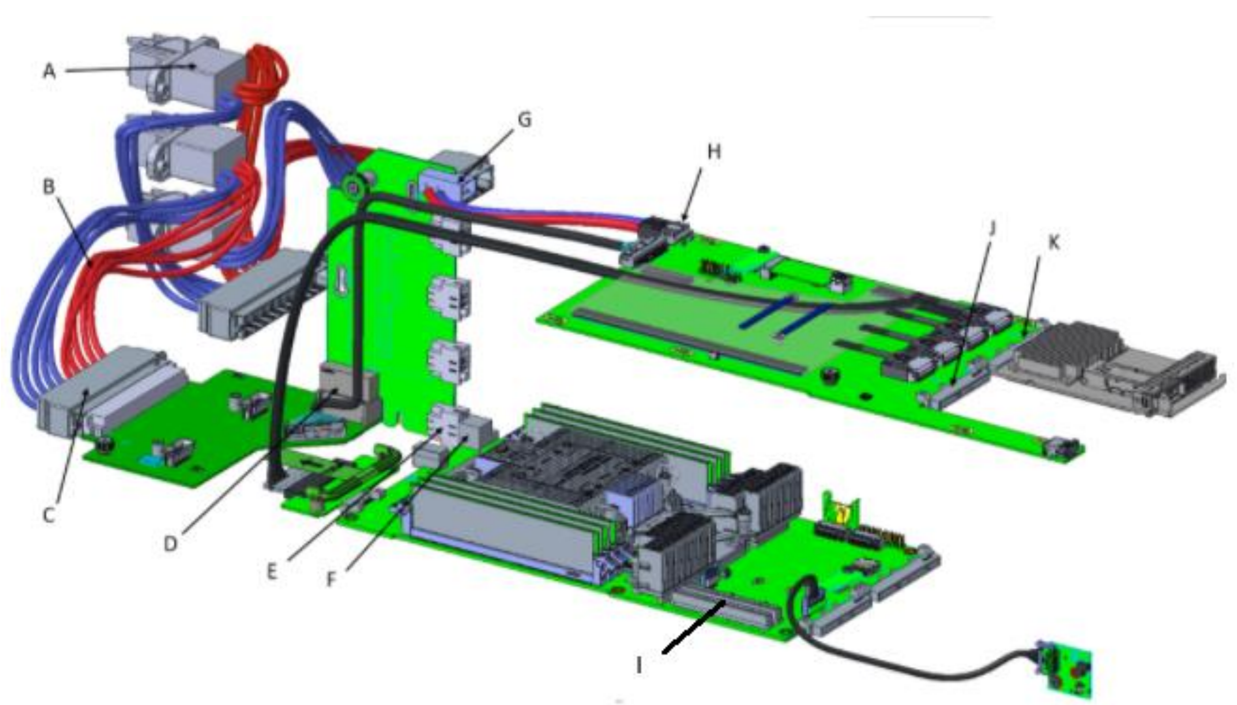


Figure 12-1: Yosemite V3 Power Delivering Connectors

Table 12-1: Maximum Current Rating of Power Connections

| Connector Location | Maximum Current (A) |
|---|---|
| A | 148.48A |

| | |
|---|---|
| B | 148.48A |
| C | 148.48A |
| D | 171A |
| E | 85A |
| F | 85A |
| G | 40A |
| H | 40A |
| I | 39.6A |
| J | 6.6A |
| K | 6.6A |

*Note: Limits defined in Table 12-1 are based on still air measurements taken at 25C ambient and current is increased until temperature rise exceeds 30C.*

## 12.4 Capacitive Load

To minimize the inrush current applied to the Open Rack V2 Power Shelf during initial rack power on and assertion of Yosemite V3 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 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 load switch or HSC is allowed, but the total sum of capacitance from a single Yosemite V3 sled shall not exceed 20.5 mF (maximum of 226mF capacitive load per power zone)

## 12.5 Hot Swap Controller Circuit

As mentioned in section 6.3, the Yosemite V3 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 (SOA) 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 31.9 A for the baseboard.
5. Undervoltage and overvoltage protection shall be configured to 10.09 V and 14.33 V respectively.
6. Default Medusa board HSC response for fault conditions shall be latch off with auto retry.
7. For the server board, the expectation is for the HSC to latch off upon fault condition.
8. PMBus interface that supports the following features:
    a. Report voltage, current, and power (VIP) telemetry with accuracy of +/- 2.0% or better when operating above 10% of the maximum range
    b. Status registers that allow the definition of upper and lower critical thresholds for VIP which are logged upon being triggered

9. 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 Delta lake 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 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 wait 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 Yosemite V3 Platform with at least 89% efficiency over the 30% to 90% load range. If higher efficiency VRs are available at additional cost and/or design complexity, then the vendor is encouraged to present the tradeoffs prior to implementation.

## 12.8 Power Policy

The power policy of server blades on the Yosemite V3 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:

1. On the same and adjacent layers, P12V_PSU shape to all other nets, including GND ≥ 40 mil.

2. On different layers, from P12V_PSU shape to all other nets, including GND ≥ 2 layers of dielectrics if overlapping.
3. 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.

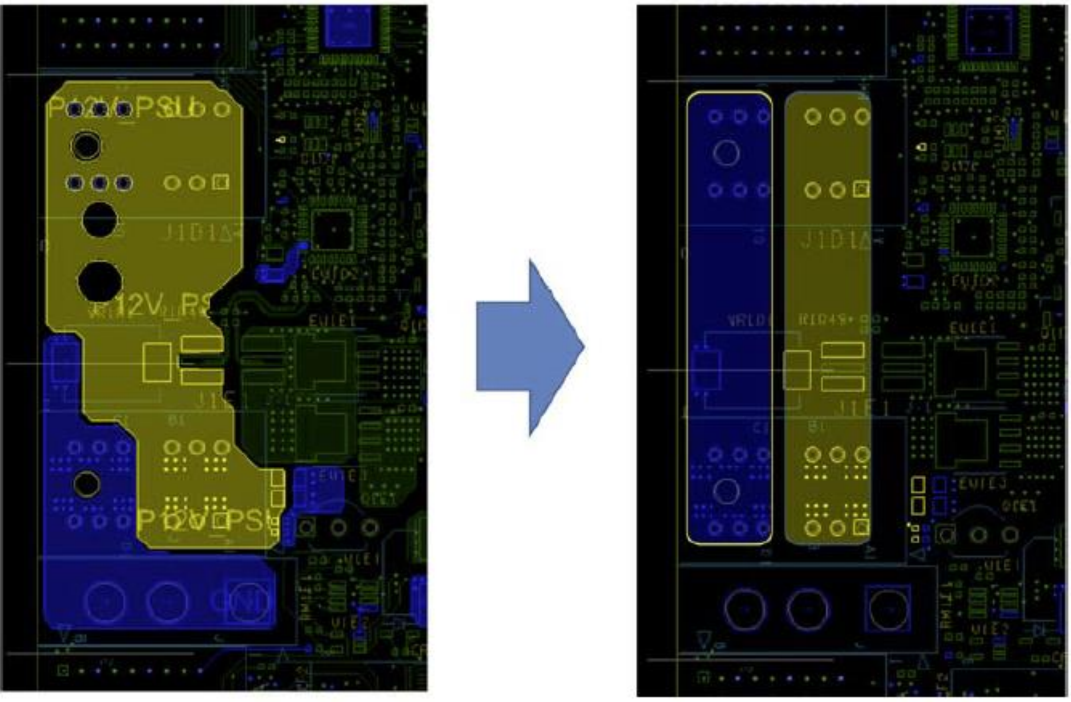

Figure 12-2: Example P12V_PSU and GND Layout

# 13 Expansion Board Design Requirements

## 13.1 Overview

This section describes the expansion system of the next generation 1S server.

The expansion system interfaces with the 1S server blade through the following connectors seen in Figure 4-1.

- 1x Straddle mount 4C+ (16 PCIe lanes)
- 1x Straddle mount 4C (Up to 16 PCIe lanes)
- 2x vertical 4C+ riser connector (2x16 PCIe lanes)

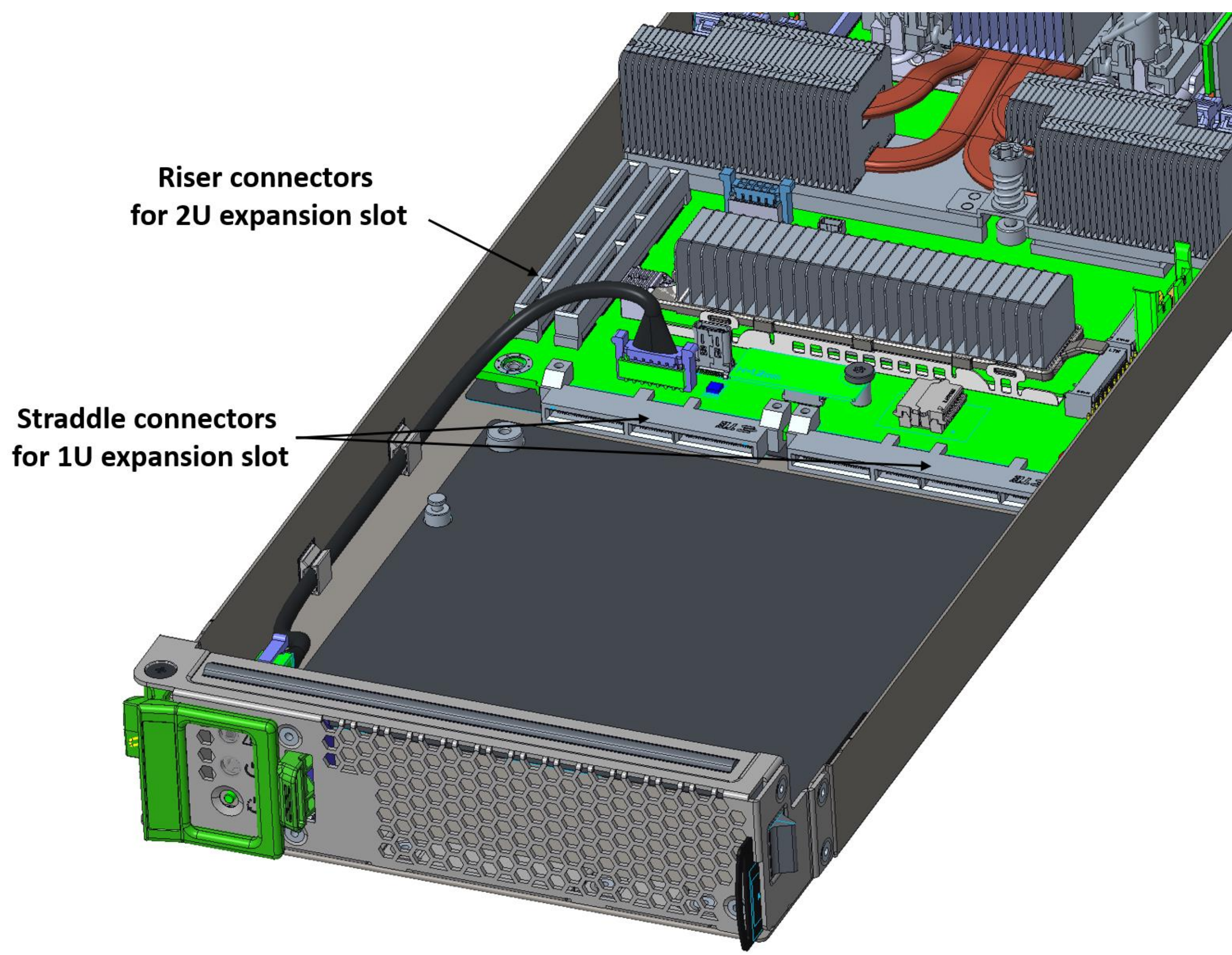


Figure 12-1: 1OU M.2 Expansion Card

There are 6 expansion card options:

1. 1OU M.2 Front Expansion Card, Figure 12-2

2. 1OU NIC Expansion Card, Figure 12-3
3. 1OU EDSFF Front Expansion Card, Figure 12-4
4. 2OU M.2 / Dual M.2 Front Expansion Card, Figure 12-5
5. 2OU EDSFF Front Expansion Card, Figure 12-6
6. 2OU PCIe CEM Expansion Card, Figure 12-7

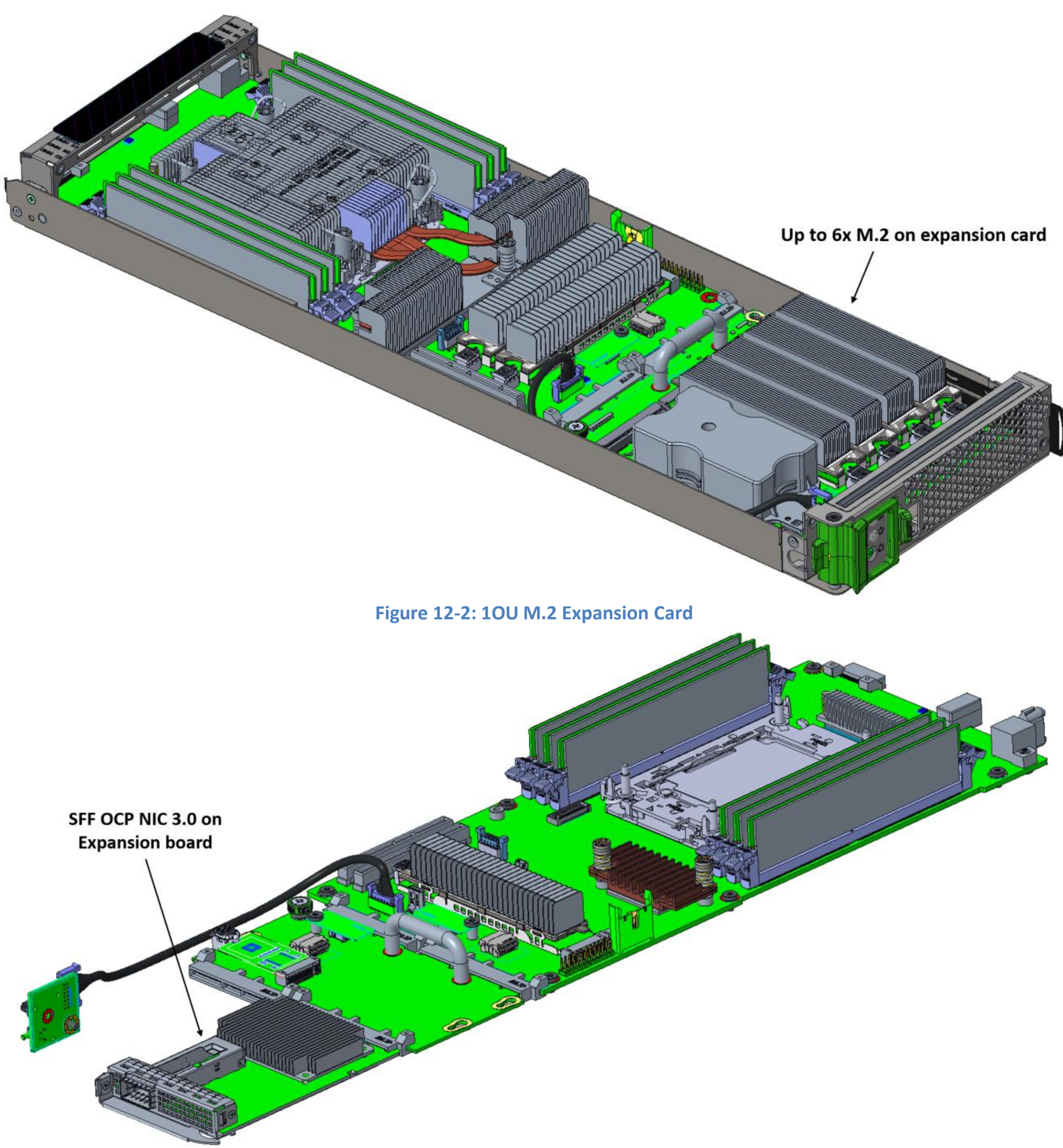


Figure 12-2: 1OU M.2 Expansion Card

Figure 12-3: 1OU NIC Expansion Card

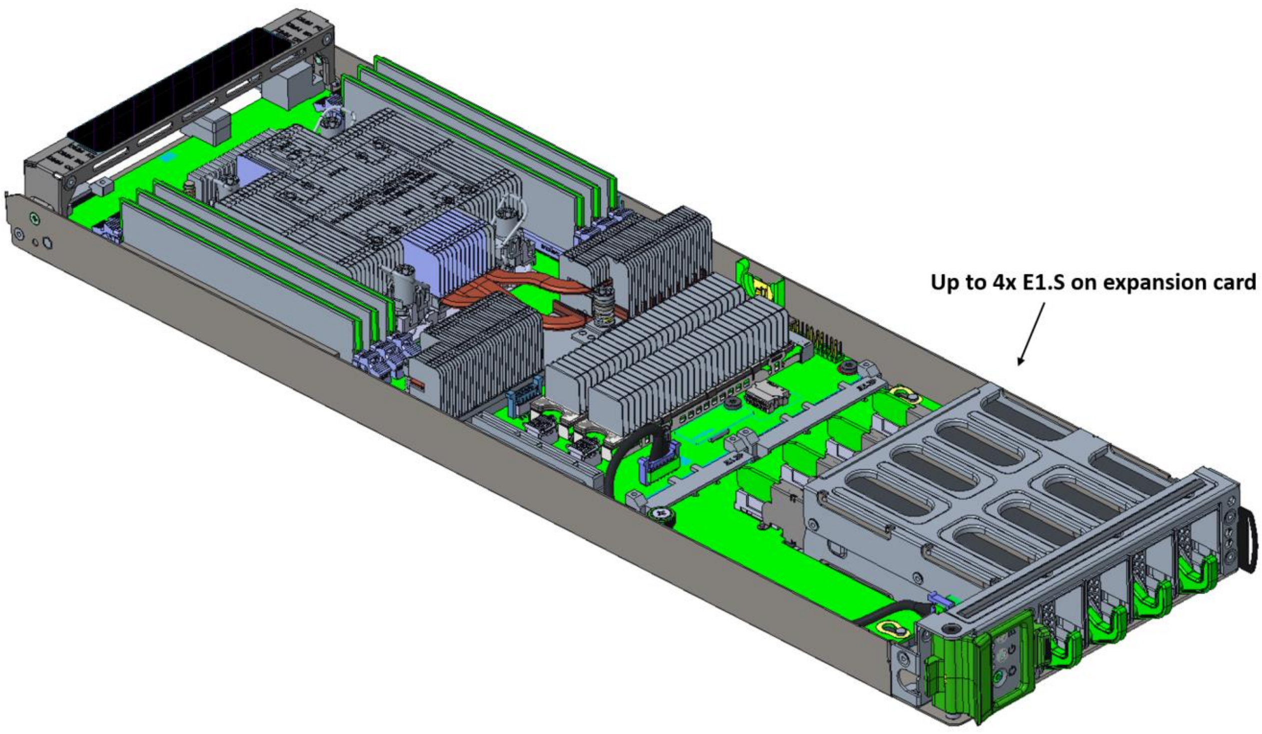


Figure 12-4: 1OU EDSFF Expansion Card

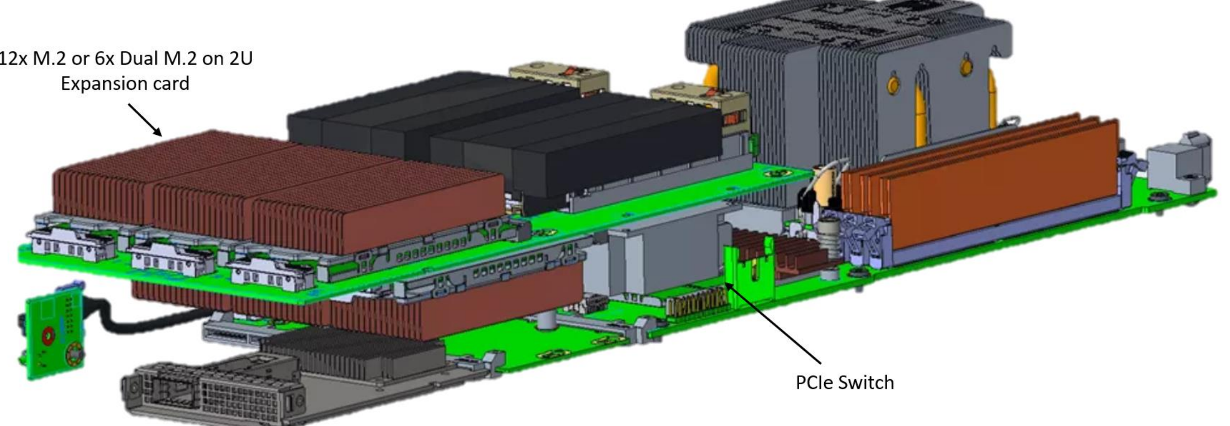


Figure 12-5: 2OU M.2 Expansion Card

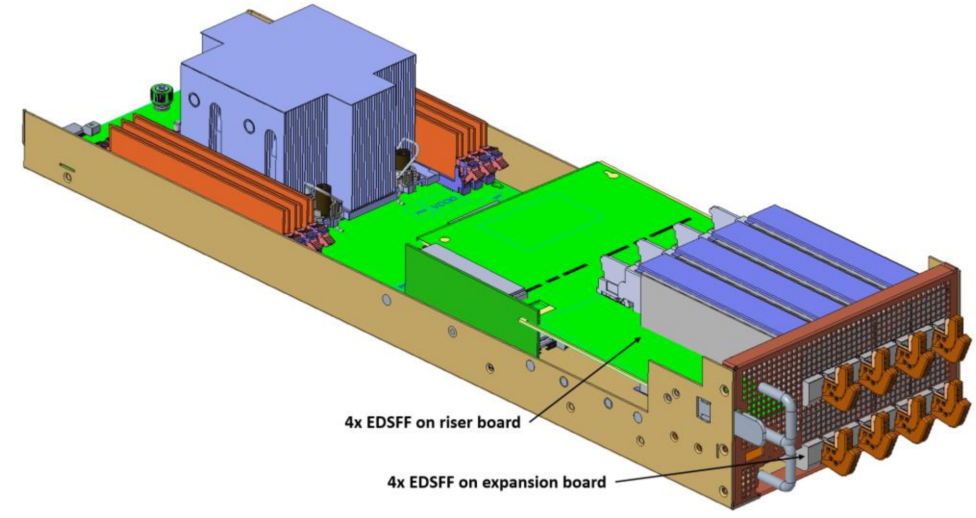


Figure 12-6: 2OU EDSFF Expansion Card

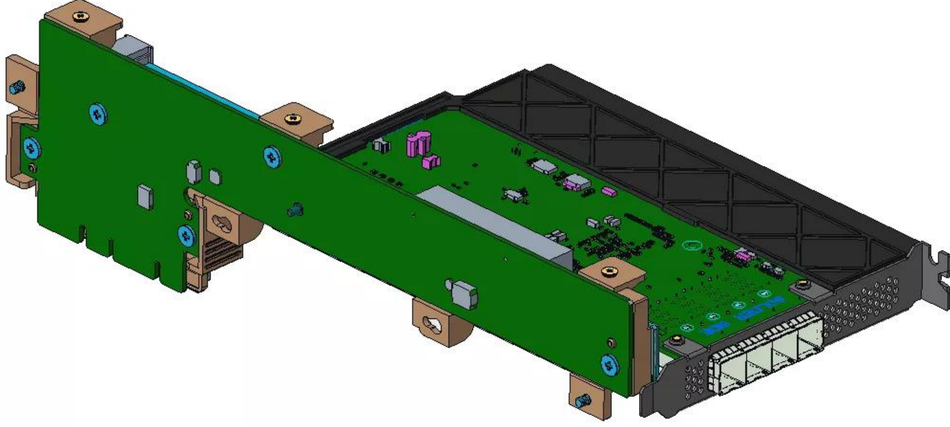

Figure 12-7: 2OU PCIe CEM Expansion Card

The interface of both the expansion system is designed with reference to the OCP NIC 3.0 interface. The use of this interface provides quite an amount of flexibility for different use case for PCIe device cards. However, certain pins which isn't applicable to the use case for the expansion system can be re-defined for what is needed for the platform. These pins could perhaps be connected on PLDs on both the host board and expansion slot to allow different definitions of signal for such purpose. This specification will call specifically which are the pins that could be used for other purposes.

## 13.2 Mechanical

### 13.2.1 Mechanical Outline

The expansion system uses the following connectors found on the 1 socket server.

- Straddle mount 4C+

- Straddle mount 4C
- 2x vertical 4C+ riser connector

Figure 6-1 defines the locations of these connectors. The 1U and 2U expansion system must be confined in the regions defined in Figure 12-6 and 12-7. Note that the 2U expansion region is inclusive of the 1U expansion region.

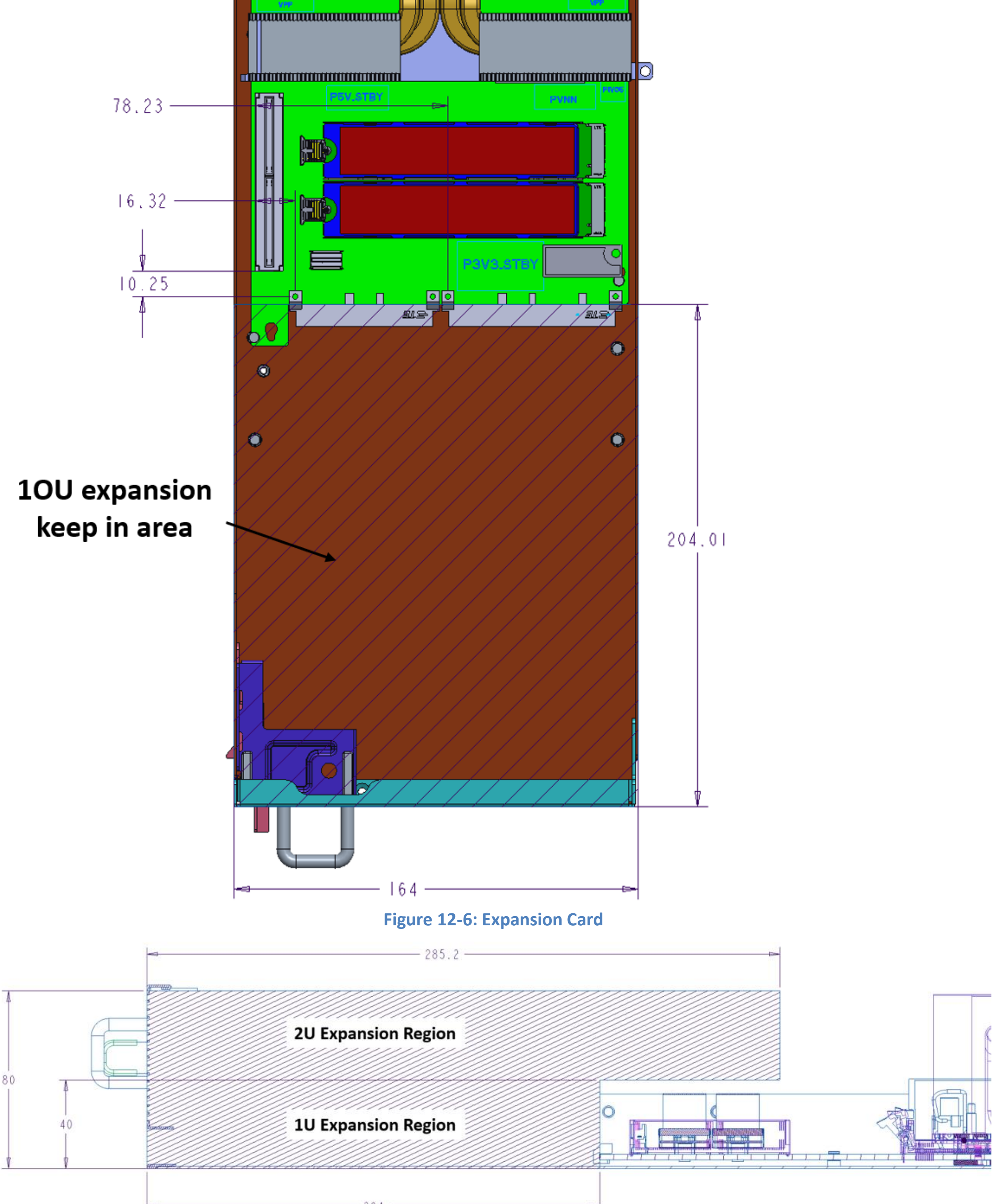


Figure 12-6: Expansion Card

Figure 12-7: Expansion Card, 1U and 2U Envelopes

## 13.3 Thermal

### 13.3.1 M.2 or Dual M.2 Cooling Solutions

#### *13.3.1.1 Orientation*

The M.2 or dual M.2 modules must be placed to avoid pre-heating.

#### *13.3.1.2 Passive Cooling Solution*

The heat sink must be a thermally optimized design at the lowest cost. Heat sink installation must be uncomplicated. Passive cooling is desired. Heat sinks must not block debug headers or connectors. The heat sink fins should be aligned with the airflow direction. Integrated heatsink is desired to enable easy serviceability. The heat sink may require cutouts/pedestals in order to reduce the TIM bond-line thickness as much as possible. The solution must maintain all M.2 component temperatures within their operational limits during all stress conditions. 6.7mm M.2 connectors must be used to allow spacing between integrated heatsink base and expansion board PCB.

## 13.4 Electrical

### 13.4.1 Design Guidelines

For most of the electrical design of the expansion modules would be referred to PDG documents of the CPU platform as well as the OCP NIC 3.0 specifications to provide consistent impedance and power connectivity.

### 13.4.2 Front and 2U Expansion

The connectivity of the front and 2U expansion to the main host board are as follows:

1. x32 PCIe Gen3/Gen4 signals
2. Reference clocks for devices
3. 4-wire serial interface for faster management/debug status
4. SMBUS or I2C interface for sideband management/thermal information
5. Control/alert signals like resets, powerbreak, present etc.
6. Power

An important theme to remember in designing for the expansion system is to ensure that the following are being taken care

1. Side band control and reporting
2. Error logging
3. Device debugging
4. Power delivery from the host system
5. Thermally cool
6. SI for high speed signals

The pinout for the front and riser connectors

### 13.4.3 Pin Definitions

As the connector is referenced to OCP NIC 3.0 interface, this specification will only call out pin definitions that COULD be different for the expansion system

Table 12-2: Expansion interface to NIC potential differences

| Signal name | Direction (with respect to server blade) | Description |
|---|---|---|
| CARD_TYPE_N | I | Found in front expansion module.<br>Used to determine the card type.<br>1.65V -1.75V: Expansion with NIC<br>0.8-0.9V: Expansion with 4 EDSFF<br>0.0-0.2V: Expansion with 6 M.2 |
| BIC_RDY | I | Found in 2U expansion<br>To indicate Bridge IC on expansion is ready |
| CARD_TYPE_Detection[2:0] | I | Found in 2U expansion<br>To identify which type of card for the riser<br>000 – GPV3 with BCM PCIe Switch<br>001 – 2OU Expansion with no PCIe Switch<br>010 – Sierra Point<br>011 – GPV3 with Microchip PCIe Switch<br>110 – Discovery Point, PCIe CEM Card<br>100, 101, 110, 111 -> RFU |

## 13.5 Power

### 13.5.1 Input Voltage Level

The expected nominal input voltage delivered by the Delta Lake 1S server's power delivery subsystem is 12.5VDC; however, has a varying range of 11.5V to 13.5V. The expansion systems shall accept and operation normally with an input voltage range of 12.5V +/- 7%. Electrical

### 13.5.2 Expansion System Power Budget

As described in the Delta Lake 1S Server Design Specification, the Delta Lake 1S Server shall support various expansion configurations and blade chassis sizes to accommodate various future use cases. However, the expansion subsystem must adhere to specific power design requirements:

1. The sustained current draw from the front expansion shall be at most 15A while the riser must be limited to 25A

2. Although the power budget may allow for high power commodities, feasibility is ultimately determined by whether the cooling solution implemented is adequate and does not introduce long term reliability degradation of the hardware
3. The expansion system needs to consume minimal power when it operates in standby mode.

Since there is not require to be a hotswap controller for each expansion board, the transient current limit is defined for the entire server based on the server form factor including the expansion cards. These are defined in terms of over current protection. These include both server level faults as well as sled level faults:

**Table 12-3: 1OU Server Transient Current Peaks**

| HW protection | Design Value (A) | Maximum Time |
|---|---|---|
| Server OCP Fault | 40 | 4 ms |
| Server Severe OCP Fault | 60 | 100 ns |
| Sled OCP Fault | 250 | 4 ms |

**Table 12-4: 2OU Server Transient Current Peaks**

| HW protection | Design Value (A) | Maximum Time |
|---|---|---|
| Server OCP Fault | 70 | 4 ms |
| Server Severe OCP Fault | 105 | 100 ns |
| Sled OCP Fault | 250 | 4 ms |

### 13.5.3 Capacitive Load

The capacitance on the input 12V power rail for Delta Lake 1S Server's expansion system shall be optimized to meet the system's power supply requirements and does not cause instability. In addition, the Delta Lake 1S Server's expansion system shall be designed such that it adheres to the maximum capacitive load allowable as defined in the Yosemite V3 Platform Design Specification document.

### 13.5.4 VR Efficiency

High efficiency Voltage Regulators (VR) shall be used on all Delta Lake 1S server expansion systems with at least 89% efficiency over the 30% to 90% load range. If higher efficiency VRs are available at additional cost and/or design complexity, then the vendor is encouraged to present the tradeoffs prior to implementation.

### 13.5.5 Power Sequence and Standby Power

To ensure the Delta Lake 1S Server host system can identify the expansion systems before turning them on, the existence of standby and main power is necessary. It is the designer's responsibility to derive proper standby power rails from the main P12V_EDGE input to implement the needed power sequencing requirements. Design consideration must be considered to avoid any leakage paths among different power subsystems (e.g. Yosemite V3 platform, 1S Delta Lake Server, and expansion cards).

### 13.5.6 Power Reading and Power Capping

Expansion modules have the freedom to implement power monitors if telemetry readings from various devices are deemed necessary. This shall be reviewed on a case by case basis depending on the use case to ensure tradeoff between functionality, design complexity, and cost have been evaluated as such additional requirements may result in unnecessary power loss.

In the event where power/thermal failure is imminent due to system may make use of the PWRBRK# signal to clamp devices in their power usage.  However, this is entirely dependent on how devices on the expansion system are designed.  It is possible that devices may not have such capability.

## 13.6 Functional

### 13.6.1 PCIe

There are as many as 32 lanes of PCIe available for connecting from the host board to the expansion system (riser or straddle mount).  However, limitations of the host system or the expansion card would ultimately dictate the available connectivity.  Solution designers should keep in mind with regards to how the high-speed lanes are connected.  It should also be noted that typical convention of TX/RX is based off the host board's point of view.

With regards to reference clocks, host system will be at least providing a ref clock at REFCLK0 pins.  More clocks may be connected if need be due to system clock domain.

The PCIe reset signals shall be used to reset the relevant devices on the expansion system.  At minimal, the host needs to provide resets to all the necessary pin outs, while the expansion system will need to handle how these resets are to be used.

### 13.6.2 Low Speed Sideband

SMBus - the both riser and straddle mount expansion system has SMBus to connected to allow host system to obtain management on the devices on board.
RBT – RBT signals for expansion boards is catered for a use case where a NIC is on the expansion or could be re-purpose for other usage.  Server card could choose not to connect for these pins.  This especially applies to the front expansion where the likelihood of a NIC could be

involved. However, the riser expansion could repurpose this for different usage, like power delivery.
Scan chain –These pins could be re-purposed for JTAG signals if devices have JTAG interfaces that need to be mux-ed to the server board.
UART – connect on the host board
USB – no connect on the host board unless there are devices in expansion that will be needing this.

### 13.6.3 Control Signals

The generic control signals for OCP would basically apply here. This is especially if the expansion is catered for an OCP NIC 3.0 use case. However, if the situation does not involve an NIC specifically, they can be re-purpose for other functions. The CARD_TYPE_N data can used to identify which type of card to allow different use of these pins.

The possible signals for re-purposing are listed here
BIF0#, BIF1#, BIF2#
PRSNTB0#, PRSNTB1#, PRSNTB3#
SLOT_ID0, SLOTID1
RFU1, RFU2, RFU3, RFU4

### 13.6.4 Debug

There should be necessary avenue made to allow debug for the devices on the expansion system.

### 13.6.5 EEPROM

There should be an EEPROM that is accessible from the platform via the Bridge IC. The EEPROM contains the Field Replaceable Unit Identification (FRU ID) information and any additional configuration information that may be required. The FRU ID is formatted in accordance with the IPMI Platform Management FRU Information Storage Definition document. The EEPROM must contain the following entries:

- Board Manufacturer
- Board Name
- Board Serial Number
- Board Part Number
- Product Manufacturer
- Product Name
- Product Part Number
- Product Serial Number
- Product Asset Tag
- Product Build: e.g. EVT, DVT, PVT, MP
- Product Version: e.g. C1
- Manufacturing Date and Time
- Manufacturing Lot Code: (preferred, but optional)
- Manufacturing Work Order: (preferred, but optional)

- PCB Revision
- SoC Model Name/Number
- SoC Revision

SoC $Tj_{MAX}$ (Maximum Junction Temperature)

#### 13.6.6 Expansion System Management

The expansion system management philosophy will like Glacier Point v2 methodology with security. The system should

- Include bridge micro-controller (BIC) to manage each I2C/SMBUS devices on expansion system. This includes but not limited to VR, temperature sensors, PCIE switches or clock buffers.
- Include CPLD to mux UART and JTAG of each module to the debug interface
- Add power switch to each device to allow control software to perform complete power cycle of each cards, when system is in operation mode without system crashing
- Devices with external flash would need to be able to support Facebooks security requirements.

##### *13.6.6.1 Expansion Bridge IC*

The Bridge IC on the expansion module will key component on board for external management devices to control the devices. It will communicate with the external management device with IPMI messages. It shall run in high-speed mode with a minimum speed of 400KHz. When possible, a 1MHz or better speed is strongly recommended.
The Bridge IC will have FRU EEPROM, VRs, thermal sensors, clock generators, PCIe switches on a local $I^2C$ bus for it to poll information and manage these devices. On top of that, it also connects to any devices with SMBUS on board. It has the capability to allow external management to update any program on these devices. The generic connection diagram from the baseboard to the expansion modules on the server card is as shown below. Take note that the links that shown are conceptual. Connectivity to the different devices are to be have the following configurations to avoid bus lock or design errors.

1. Devices of different power state should NOT be on the same bus line. Do consider separate channels and disabling of the bus master when it isn't in use.
2. Unique bus address for devices on a bus line
3. Bus multiplexers should be used in when a bus is to access devices of same bus address
4. Design need to cater for separate I2C bus for those connecting to connector.

Diagram can be found in section 5 (Figure 5-13).

##### *13.6.6.2 CPLD v. BIC Functionality*

Some features can't be covered by CPLD.

1. ADC Pin function

2. USB to JTAG function
3. SMBUS master of IPMB (CPLD can't be a SMBUS master to directly output data to server BIC, so server board BIC need polling to get data from CPLD, and it will impact server board BIC performance)

#### 13.6.6.3 GPIO Register

The Bridge IC shall provide a GPIO interface to the BMC through a GPIO register block. In this way, the BMC can control the GPIO behind the Bridge IC (or this hardware abstraction layer) through accessing this register block.
The Bridge IC shall provide a way for BMC to configure GPIO pin direction, interrupt capability, and provide a way to get/set the current status of GPIO signals. It shall send an interrupt message to the BMC when the interrupt enabled GPIO signal changes its state. The GPIO register interface exposed by the Bridge IC shall provide signals that indicate various conditions as shown in Table 4.

**Table 12-6: Bridge IC GPIO Table**

| GPIO offset | GPIO Pin Function | Comments |
|---|---|---|
| Byte 1 -bit [0] | Power Good – CPU Core | Indicates that the CPU's core power input is good |
| Byte 1 - bit [1] | Power Good – PCH core | Indicates that the PCH's core power input is good |
| Byte 1 - bit [2] | DDR Channel A/B Voltage regulator hot | Indicates that the DDR Channel A/B Voltage Regulator is hot |
| Byte 1 - bit [3] | DDR Channel D/E Voltage regulator hot | Indicates that the DDR Channel D/E Voltage Regulator is hot |
| Byte 1 - bit [4] | CPU VccIN Voltage regulator hot | Indicates that the CPU VccIN Voltage Regulator is hot |
| Byte 1 - bit [5] | CPU Throttle | System firmware request CPU throttle |
| Byte 1 - bit [6] | PCH Hot | Indicates PCH temperature is over setpoint |
| Byte 1 - bit [7] | DIMM Hot | Indicates that one or more DIMM is hot |
| Byte 2 - bit [0] | CPU thermal trip | Indicates that CPU experienced over temperature event and shut-down |
| Byte 2 - bit [1] | PCH thermal trip | Indicates that PCH experienced over temperature event and shut-down |
| Byte 2 - bit [2] | CPU FIVR fault | Indicates a CPU internal Voltage Regulator Error condition |
| Byte 2 - bit [3] | CPU Catastrophic Error | Indicates that the CPU experienced catastrophic error |
| Byte 2 - bit [4] | CPU Non-Recoverable Error | Indicates that the CPU experienced non-recoverable error |

| | | |
|---|---|---|
| Byte 2 - bit [5] | CPU Critical Error | Indicates that the CPU experienced critical error |
| Byte 2 - bit [6] | CPU Non-Critical Error | Indicates that the CPU experienced non-critical error |
| Byte 2 - bit [7] | Sleep S4 state | When low, indicates CPU has entered S4 state or lower |
| Byte 3 - bit [0] | Non-maskable interrupt | Non-maskable interrupt |
| Byte 3 - bit [1] | System management interrupt | System management interrupt |
| Byte 3 - bit [2] | Platform Reset | Platform and PCIe Reset |
| Byte 3 - bit [3] | Front Panel Reset Input | Initiate platform reset<br>(front panel reset button being pressed) |
| Byte 3 - bit [4] | Front Panel Reset Output | Host Reset Output from Bridge-IC |
| Byte 3 - bit [5] | Bios Power-on Self Test (POST) complete | Bios Power-on Self Test (POST) complete |
| Byte 3 - bit [6] | Sleep S3 state | When low, indicates CPU has entered S3 state or lower |
| Byte 3 - bit [7] | Power Good – CPU VccIN | Indicates that the VCCIN Voltage Regulator power is good |
| Byte 4 - bit [0] | Boot SPI selection | Select boot SPI0 or SPI1 as boot SPI |
| Byte 4 - bit [1] | Ejector latch detection | Identify ejector fully closed |
| Byte 4 - bit [2] | BMC Reset | Reset BMC from host |
| Byte 4 - bit [3] | At-scale-debug (ASD) TCK selection | Select at-scale-debug (ASD) TCK drive CPU TCK or PCH TCK |
| Byte 4 - bit [4] | BMC ready | Indicated BMC is ready |
| Byte 4 - bit [5] | Host / Bridge-IC UART select | Select Host UART or Bridge-IC Debug UART |
| Byte 4 - bit [6] | I2C MUX reset | Reset I2C MUX |
| Byte 4 - bit [7] | At-scale-debug (ASD) PREQ | CPU probe mode request |
| Byte 5 - bit [0] | At-scale-debug (ASD) JTAG TRST | JTAG Reset |
| Byte 5 - bit [1] | System Throttle | Indicates system is currently throttling |
| Byte 5 - bit [2] | At-scale-debug (ASD) PRDY | CPU probe mode ready |

| Byte 5 - bit [3] | XDP Present | Indicates traditional ITP connected |
|---|---|---|
| Byte 5 - bit [4] | ASD Present | Indicates at-scale-debug connected |
| Byte 5 - bit [5] | CPU power debug | Use for debug CPU integrated VR |
| Byte 5 - bit [6] | JTAG MUX selection | Select JTAG connection to ITP or ASD |

##### 13.6.6.4 Thermal Alerts

The Bridge IC need to have a mechanism to provide thermal alerts and over temperature notifications. The external manager must be able to receive these alerts in a timely fashion to allow it act quickly.

##### 13.6.6.5 Event Log

There should be a logged event whenever the devices on the expansion system encounter events out of the ordinary operations.  These include but not limited to over temperature, over voltage, over current events detected.

### 13.6.7 Expansion System Options

At time of writing, there are essentially a few specific expansion systems available.

1. Front expansion with up to 6 M.2 or up to 3 dual M.2 module slots.
2. Front expansion for an OCP 3.0 NIC (accommodate for LFF)
3. Front expansion with expansion with 4 EDSFF module slots.
4. Riser card with riser expansion module with up to 12 single M.2 or 6 dual M.2 module slots connected to a PCIe switch
5. Riser card with riser expansion module with 6 M.2 or 3 dual M.2 module slots
6. Riser card with 6 EDSFF module slots.

Expansion modules with M.2 modules have the following architecture in general, shown below for case where there are 3 dual M.2 or 6 M.2 modules on the slot.

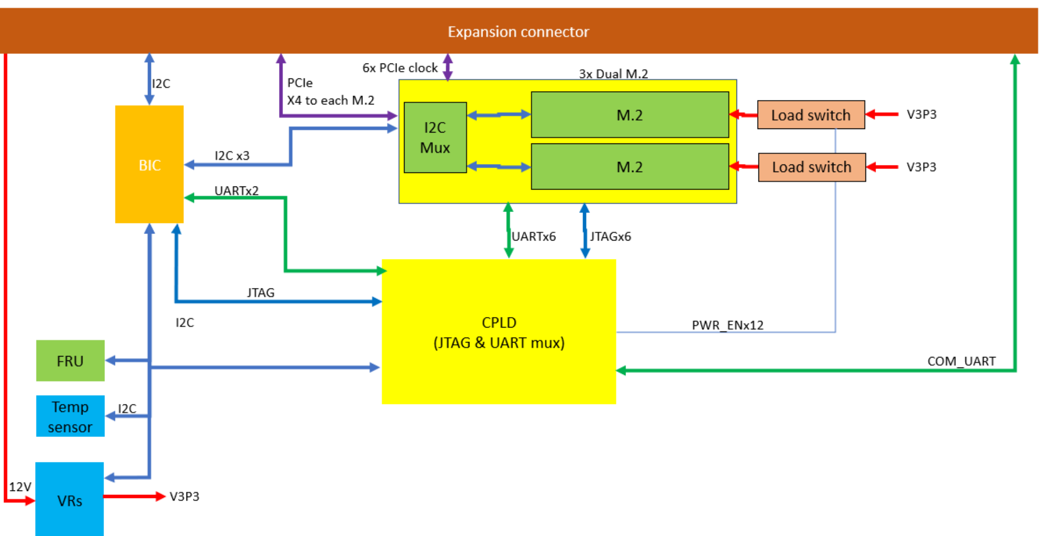


Figure 12-9: M.2 Expansion Architecture

Each M.2 on the module shall have the following connection when there is no PCIe switch present on board.

1. Load switch connecting to its power pins. System can obtain its voltage/power/current information. System OS/BIOS needs to be able to control power to M.2.
2. JTAG connecting to CPLD. CPLD is a mux connecting to BIC. At the same time, The CPLD is also connected to the server blade CPLD thru the Serial IO pins (LD, DATA_OUT, DATA_IN, CK). This allow an alternative path of direction access to the M.2.
3. UART connecting to CPLD mux. CPLD is a mux to connect to the UART to the server blade.
4. X4 PCIe to expansion connector. System is to have the ability determine if there is a M.2 or dual M.2 and bifurcate accordingly. This is with x4 config for each M.2 for initial boot, upon learning if modules are of M.2 or dual M.2, system will reconfigure and reset again to re-run bifurcation.
5. Its reset pin is connected to CPLD and being gated by its load switch’s power good output with ~100mS delay on power on and together on power off.
6. It should be noted that the control of M.2 power is mapped to the relevant Downstream Port’s PCIe config register such that this information can be relayed to the BIC on the expansion board to control the power, reset, SMBUS and clock accordingly.
7. For the system point of view, there needs to be a prevention mechanism for system crashing when the PCIe links could go down without notification, as well as a managed manner when the system conducts AC cycle to the PCIe device.

For modules with PCIe switch, it is required for the PCIe switch to be able to control necessary resets and power of the modules to allow manage hot plug behavior through its sideband signals, with similarity that is implemented in GPv2.

When there EDSFF devices instead of M.2, similar methodology of operations is to be maintained and with addition that their present signal to be involved in the PCIe hot-plug process.

# 14 Environmental Requirements and Other Regulations

## 14.1 Environmental Requirements

All platform boards shall meet the following environmental requirements:

- Gaseous contamination: Severity Level G1 per ANSI/ISA 71.04-1985
- Ambient operating temperature range: 0°C to +35°C
- Operating and storage relative humidity: 10% to 90% (non-condensing)
- Storage temperature range: -40°C to +70°C*
- Transportation temperature range: -40°C to +70°C (short-term storage)

The full system shall meet the following environmental requirements:

- Gaseous contamination: Severity Level G1 per ANSI/ISA 71.04-1985
- Ambient operating temperature range: +15°C to +35°C
- Operating and storage relative humidity: 10% to 90% (non-condensing)
- Storage temperature range: -40°C to +70°C*
- Transportation temperature range: -40°C to +70°C (short-term storage)
- Operating altitude with no de-ratings: 6000 ft

## 14.2 Vibration and Shock

All platform boards shall meet all shock and vibration requirements according to IEC specifications IEC78-2-(*) and IEC721-3-(*) Standard & Levels. Testing requirements are listed in the table below. The motherboard shall comply fully with the specification without any electrical discontinuities during the operating vibration and shock tests. No physical damage or limitation of functional capabilities (as defined in this specification) shall occur to the motherboard during the non-operating vibration and shock tests.

Table 14-1: Vibration and Shock Requirements

| | Operating | Non-Operating |
|---|---|---|
| **Vibration** | 0.5G RMS, 5 to 500 to 5 Hz, Random Vibe, 1 sweep, 20 min along three axes (+/-)<br>5-20Hz – 6db/Oct<br>20-250Hz – 0.0007 $G^2$/Hz<br>250-500Hz – 6db/Oct | 1.2G, 5 to 500 to 5 Hz per sweep, 1 sweep at 0.5 Octave/min, 3 axes.<br>5-10 Hz – 0.5G /10-350 Hz – 1.2G /350-500 Hz – 0.5G |

| **Shock** | 2G-6G, half sine, 11ms, total 6 shocks, test along three axes(+/-) | 12G, half sine, 11ms, total 6 shocks, test along three axes(+/-) |
|---|---|---|

## 14.3 Regulations

Yosemite V3 shall meet the technical requirements in the following EMC, safety and environmental compliance standards.

- UL 62368-1, IEC 62368-1 and EN 62368-1; hazard-based performance standard for Audio video, IT & Communication Technology Equipment.
- FCC CFR47 Part 15, Subpart B, Class A criteria
- EU EMC Directive (2004/108/EC); common broad objectives for EMC regulations, so that electrical equipment approved by any EU member country will be acceptable for use in all other EU countries.
- RoHS Directive (2015/863/EU); aims to reduce the environmental impact of EEE by restricting the use of certain substances during manufacture.
- REACH Regulation (EC) No 1907/2006; registration with the European Chemicals Agency (ECHA), evaluation, authorization and restriction of chemicals.
- WEEE Directive (2012/19/EU); aims to reduce the environmental impact of EEE by restricting the use of certain substances during manufacture.

# 15 Prescribed Materials

## 15.1 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

## 15.2 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 Facebook

The following limitations apply to the use of inductors:

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

## 15.3 Component De-rating

For inductors, capacitors, and FETs, de-rating analysis is based on at least 20% de-rating.

# 16 Labels and Markings

The motherboard shall include the following labels 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 16-1: Lables 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 |

# 17 Hardware Management

## 17.1 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.

# 18 Security

All products shall have a completed Security Tab in the 2021 Supplier Requirements.

# 19 Revision History

| Author | Description | Revision | Date |
|---|---|---|---|
| Thoon KY | ▪ Initial draft | 0.1 | 11/09/2018 |
| Thoon KY | ▪ Multiple updates on various section due to form factor change | 0.2 | 1/16/2019 |
| Kiran/Todd/.. | ▪ PFR and 48V added | 0.13 | 4/24/20 |
| | ▪ | | |
| Kiran/Todd/Haoran.. | ▪ Title changed<br>▪ Block diagrams and Baseboard to server pins updated<br>▪ Updated the interfaces based on the actual implementation<br>▪ Added SFF BSM<br>▪ Power threshold updated | 0.14 | 12/12/20 |
| Kiran/Haoran/Michael | ▪ Updated Mech drawings to show dimensions | 0.15 | 1/11/21 |
| | ▪ Updated power numbers | | |
| | ▪ Fixed typos and correct naming (chassis, sled management cable) | | |
| Todd | Draft for OCP Server Group | 1.00 | 1/21/21 |
| Kiran | Updated table 12.1 for Shock and Vibe | 1.01 | 1/22/21 |
| Todd<br>Kiran | Merged Expansion Card Spec.<br>Minor Error correction | 1.10 | 3/4/21 |
| Todd | Updated 2OU PCIe Switch config image | 1.11 | 3/10/21 |
| Todd | Author update, VR efficiency update | 1.12 | 3/17/21 |

| | | | |
|---|---|---|---|
| Todd | Corrected inconsistencies | 1.13 | 3/18/21 |
| Todd | Updated for OCP template | 1.14 | 3/22/21 |
| Todd | Updated for OCP IC feedback about supplier requirement worksheets | 1.15 | 5/10/21 |
| Todd | Removed Confidential under Facebook NDA from images | 1.16 | 5/19/21 |

# Appendix A - Requirements for IC Approval

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

| **Requirements** | **Details** | **Link to which Section in Spec** |
|---|---|---|
| Contribution License Agreement | OWF CLA<br>OWFa1.0 Final Specification Agreement | License |
| Tenets | Openness<br>Efficiency<br>Impact | OCP Tenets Compliance |
| Supplier available within 120 days | Wiwynn Corporation | |
| Will they apply for OCP product recognition? | Yes | |