Facebook Multi-Node Server Platform: Yosemite Design Specification

V 0.4

Authors:

Yan Zhao, Hardware Engineer, Facebook

Renee Chu, Mechanical Engineer, Facebook

Jacob Na, Thermal Engineer, Facebook

Sai Dasari, Software Engineer, Facebook

Copyrights and Trademarks

1 Scope

This specification describes the design of the Yosemite Platform that hosts multiple One Socket (1S) servers.

2 Contents

3 Overview

This document describes the Facebook's multi-node Server platform (code name: Yosemite) and the design requirements to integrate the platform into Open Rack V2.

The Yosemite Platform is a next generation multi-node server platform that hosts 4 Open Compute Platform (OCP) compliant One Socket (1S) server cards in a sled that can be plugged into an OCP Cubby chassis.

There is a vertically installed side plane inside the chassis to hold 1S server cards horizontally in two floors. With this new architecture, we could leverage 1S server cards that use a higher-power, higher-performance System On a Chip (SoC). The Yosemite Platform fully uses space inside the sled to add additional performance and functionality. The modular design makes this platform flexible that it is possible to adapt SoC-agnostic 1S servers from different vendors.

On the network side, the Yosemite Platform has various solutions to provide network access to the 1S servers. All 1S servers have their own independent network interface, virtually. To simplify cabling, only single network cable is allowed to connect a Yosemite Platform to a top-of-rack (TOR) switch.

1S servers with integrated 10GBase-KR Ethernet controllers can use an OCP 2.0 compliant KR Mezzanine card to provide 4x 10G links to a TOR switch through a single QSFP+ cable. If 50Gbps connectivity is required, an aggregated KR Mezzanine card can be used to aggregate 4x 10G KR from 4 1S servers to a single 50Gbps link which is then connected to 100G TOR switch through a QSFP28 cable.

For the 1S servers that do not have or do not use integrated network controllers, a PCI Express (PCIe) based multi-host network mezzanine card can be used on the Yosemite Platform to provide 40Gbps or 50Gbps connectivity on line side to a TOR switch. On the host side, every 1S server connects to the multi-host network mezzanine card through an x2 PCIe Gen3 link and sees its own network interface card.

A Baseboard Management Controller (BMC) on a side plane is used to manage all 1S servers and the Yosemite Platform. There are four virtual BMCs running on the physical BMC so technically every 1S server has its own BMC. The BMC shall support both in-band and out-of-band (OOB) management so that the BMC can be accessed from 1S servers on the same Yosemite platform or from an external server on network.

All 1S servers get single 12V power from the Yosemite Platform, while the Yosemite Platform gets 12V power from the rack. There is a 12V power switch in front of every 1S server slot under the BMC's control. Therefore, the BMC can do full AC power cycling to 1S servers when needed. The BMC monitors the health status of the Yosemite Platform and all 1S servers (e.g., power, voltage, current, temperature, fan speed, etc.) and will take action when failures occur.

The Yosemite Platform is 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 Platform in a Cubby chassis can be safely inserted or removed to/from an Open Rack. You can find more details in the mechanical section of this document.

4 License

5 Yosemite Platform Features

5.1 Platform Block Diagram

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

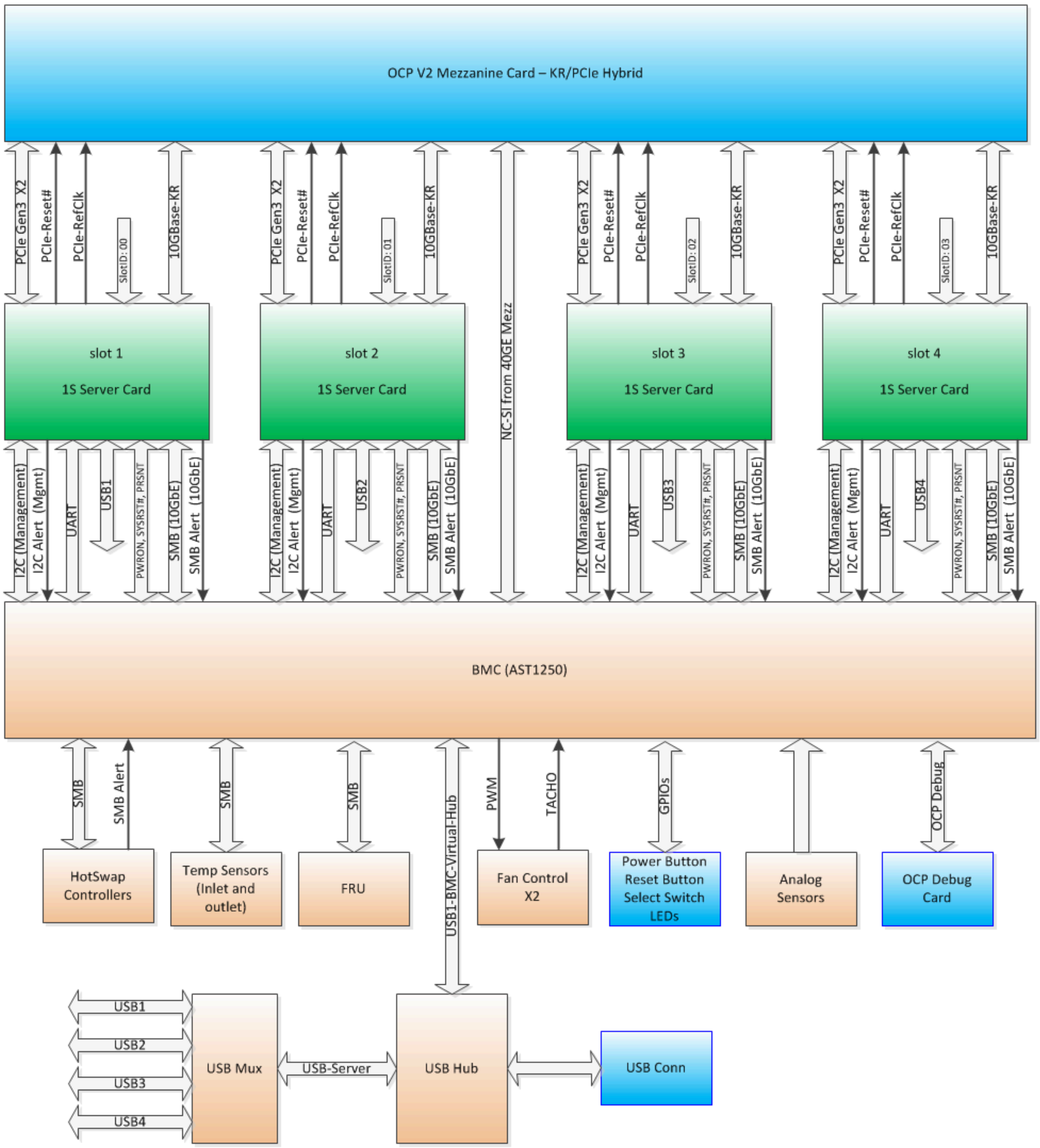


Figure 5-1: Yosemite Platform Block Diagram

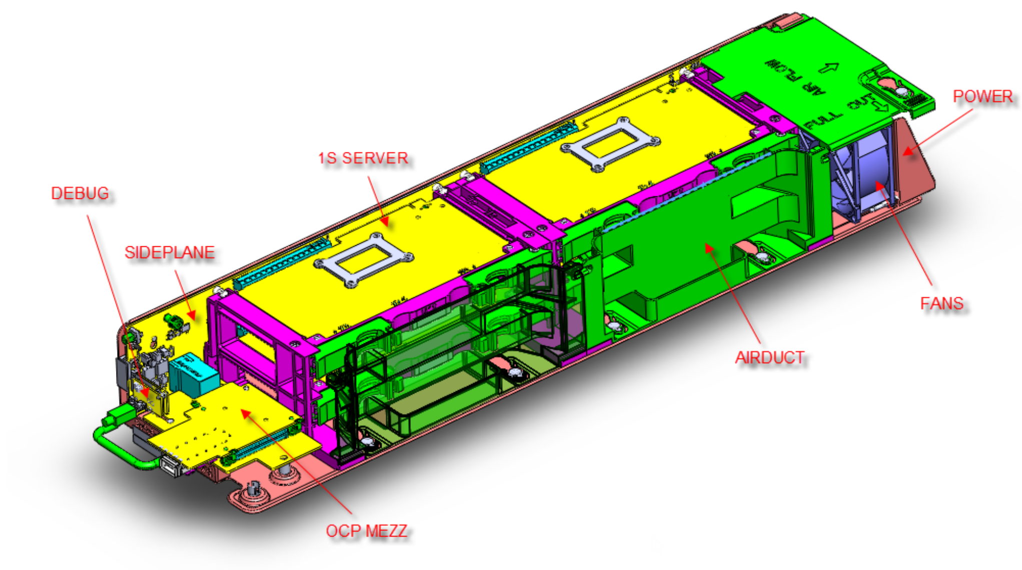


Figure 5-2: Yosemite Platform

5.2 Yosemite Platform Side Plane

The Yosemite Platform contains a side plane to hold all connectors and common infrastructure pieces, including the 1S server card connectors, OCP V2 Mezzanine card adapter card, a 12.5V inlet power connector (from the Cubby chassis), a BMC section, fan connectors, a hot-swap controller, and a front panel.

The side plane is installed vertically on the side of a Cubby chassis. OCP-compliant 1S server cards with height of 110mm or 160mm can be installed horizontally to the side plane.

The side plane shall be implemented as a low-maintenance, robust, platform to reduce the need of service. A BMC (ASPEED AST1250) is the main control unit on the side plane. The Yosemite Platform uses an adapter card at the front of the sled as a carrier board for OCP 2.0 Mezzanine cards. The OCP 2.0 Mezzanine connectors on the adapter cards have been carefully designed in a hybrid way to take a PCIe-based multi-host OCP V2 Mezzanine 40GbE/50GbE card, or a 40Gb capable KR retimer Mezzanine card or a 50GbE KR Aggregation Mezzanine that connects to the 1S server's built-in 10GBase-KR Network Interface Cards (NICs), as the Ethernet interface to the external world. Either way, the NIC will be used as a shared NIC, so that a BMC can be accessed via the OOB of the NIC, Network Controller Sideband Interface (NC-SI), or System Management Bus (SMBus).

By sampling the sensors on the Yosemite Platform periodically, the BMC continuously monitors the system's health status from function, power and thermal perspectives. The BMC shall implement sophisticated algorithms to control fans accordingly.

The BMC is connected to a hot-swap controller thru an Inter-Integrated Circuit (I^2C) bus so that it can get system-wide power consumption and maintain a healthy status. The BMC also controls 12.5V power to each 1S server. It is possible to let the BMC completely shut down 12.5V to a 1S server when the server needs a cold reboot.

A power button, a reset button, a USB connector and an OCP debug card header are provided on the front panel of the side plane and they belong to the current active server. The user could toggle a switch to select a 1S server as the current active server to use the power button, reset button, USB connector and OCP debug card on the front panel.

The USB connections from all 1S servers are connected to the BMC's virtual hub port through a USB multiplexer so that a user could upgrade the BMC firmware via a USB interface from a 1S server. This method is much faster than going through the OOB.

There is one 12.5V inlet power connector and two fan connectors on the backside of the side plane to provide power and cooling to the system.

5.3 Yosemite Platform Power Delivery

Figure 5-3 illustrates the power delivery block diagram of Yosemite Platform.

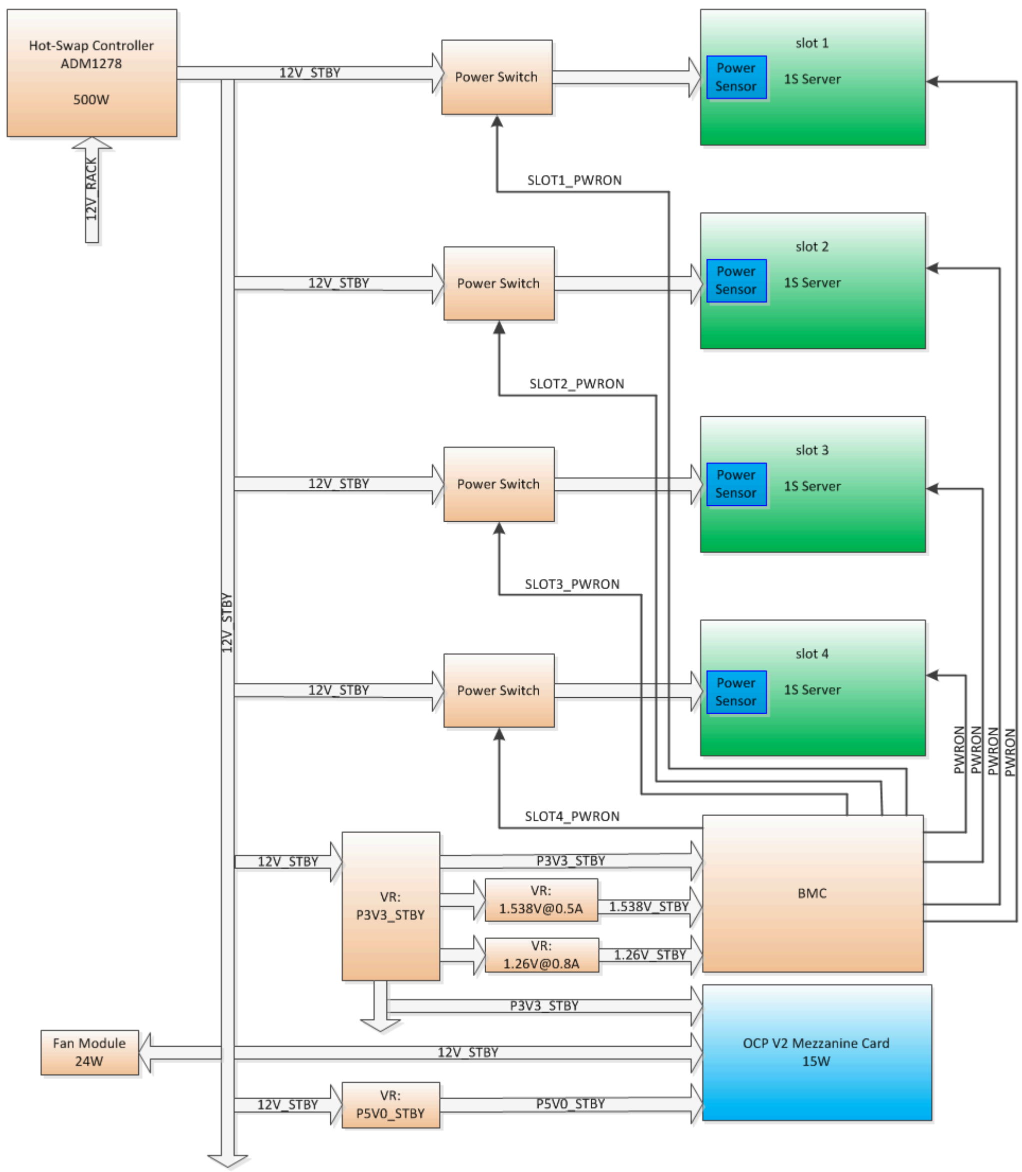


Figure 5-3: Yosemite Platform Power Delivery Block Diagram

The Yosemite Platform is designed to host up to four 1S servers. Each 1S server consumes up to 96W total power. Considering power loss and the side plane's power consumption, the Yosemite Platform could support a maximum of 500W total power.

The Yosemite Platform system gets 12.5V power from the Cubby chassis through a power cable. A hot-swap controller (ADM1278) is used to protect rack power from a voltage drop due to transient current draw when Yosemite Platform is inserted or removed from a live rack. The hot-swap controller shall set the power limit accordingly and shut the platform's 12.5V power down immediately whenever over-current, over-voltage, under-voltage, or over-temperature conditions occur.

The BMC uses standby rails P3V3_STBY, P1V5_STBY, and P1V2_STBY to power BMC circuits and its DDR3 memory. 12.5V_STBY, P5V0_STBY, P3V3_STBY are provided at the OCP V2 Mezzanine card connectors to power an OCP V2 complaint Mezzanine card.

Every 1S server shall have a power sensor right at the 12.5V input side of the card. The SoC or the Bridge IC of the 1S server needs to sample the power sensor periodically and calculate a 1 second average from those samples. The BMC shall be able to access this power sensor via the Bridge IC on the 1S server. As a debugging feature, the Bridge IC shall be able to sample the power sensor as fast as 10ms.

By default, these 12.5V_STBY power switches to 1S servers should always be on unless the BMC turns them off on purpose. Thus, the integrated Ethernet controllers on the 1S server always get standby power to keep side-band traffic alive even when 1S servers are in standby mode.

There is a power switch on the 12.5V_STBY going to every 1S server. The BMC could power cycle a 1S server by toggling the corresponding 12.5V_STBY power switch. This is useful when the 1S server needs a cold reset or AC power cycling. Tri-state buffers on GPIOs between 1S servers and the Yosemite Platform are required to avoid leakage from the Yosemite platform to 1S servers when they are in power-off or stand-by state.

The BMC has a dedicated power-on signal for each 1S server. Depending on the power policy, the BMC enables power to 1S servers upon request. The BMC shall drive power-on signals as a power button function as defined in the Advanced Configuration and Power Interface (ACPI).

5.4 SMBus Block Diagram

Figure 5-4 illustrates the Yosemite Platform SMBus block diagram.

Each 1S server has one SMBus from its integrated NIC connected to the BMC as sideband if a KR Mezzanine card or KR Aggregation card is being used as the network interface for the platform. A Bridge IC on each 1S server is connected to the BMC thru a dedicated I^2C bus as the management interface between a 1S server and the BMC. When a PCIe based multi-host network interface card is being used, the BMC shall use NC-SI as the sideband as it is much faster than SMBus.

Depending on the Mezzanine card type, the BMC could connect to a Management Data Clock (MDC)/Management Data Input/Output (MDIO) or I^2C to LAN_SMB port on the Mezzanine card to configure the Mezzanine card or use the NIC's SMBus as OOB. A dedicated SMBus is also connected to MEZZ_SMB of the Mezzanine card as a management path, so that the BMC can read Field Replaceable Unit (FRU) data from the Mezzanine card or perform other management tasks.

The BMC can access thermal sensors, the hot-swap controller and the FRU via a separate SMBus, as shown in Figure 5-4 below.

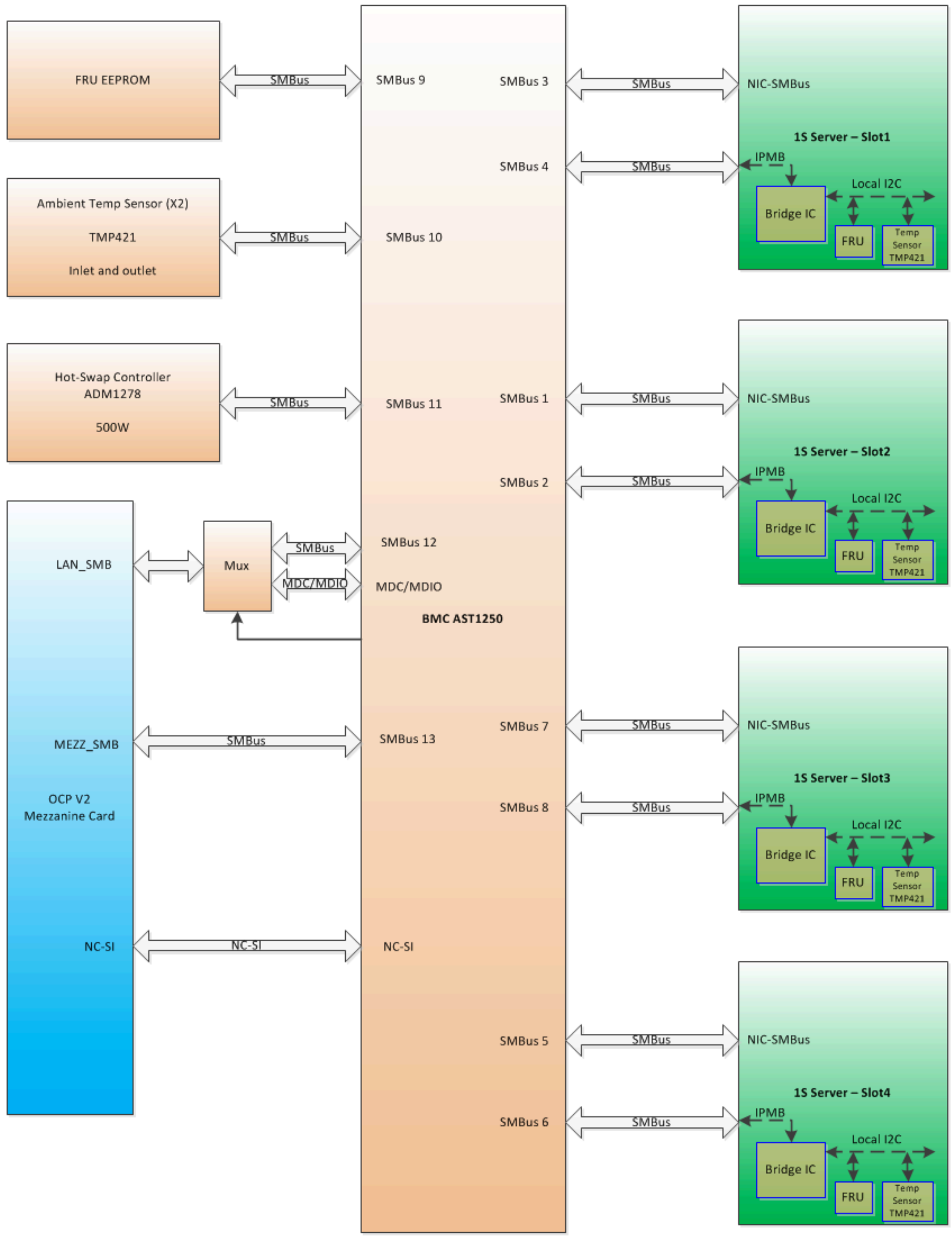


Figure 5-4: Yosemite Platform SMBus Block Diagram

5.5 1S Server Slots

5.5.1 Overview

The Yosemite Platform has four slots that can host four OCP compliant 1S servers.

5.5.2 1S Server Connector

The 1S server specification defined two x16 PCIe-like connectors, primary connector and extension connectors. However, the Yosemite Platform only uses the primary connectors of the 1S servers.

The X16 PCIe connector is defined in the PCIe specification. However, a completely different pinout is used for the Yosemite Platform.

5.5.3 1S Server Slot Pinout

The Yosemite Platform uses the primary x16 pinout as defined in section 6.2 of the OCP 1S Server specification, which you can download from the OCP website through the following link: http://www.opencompute.org/wiki/Motherboard/SpecsAndDesigns.

5.6 Network Options

The Yosemite Platform provides various network options that support a 40GbE or 100GbE top of rack (TOR) switch with a single network cable. With a PCIe/10GBase-KR hybrid design approach, Yosemite Platform supports following OCP 2.0 Mezzanine card compliant network cards:

- 4x10G KR-retimer Mezzanine card
- Multi-Host 40G/50G NIC Mezzanine card
- 50G KR-Aggregation Mezzanine card

Because the real estate on the side plane is limited, an adapter card is designed as a carrier board to host standard OCP 2.0 Mezzanine cards at the front. A PCIe link with two lanes and a 10GBase-KR link from each 1S server, an NC-SI interface, NIC SMBus and management SMBus from BMC, and other related signals are routed to two OCP 2.0 Mezzanine connectors on the adapter card through an AirMax connector on the side plane.

5.6.1 OCP 2.0 Hybrid Mezzanine Connector Pinout

Pin definitions of the hybrid Mezzanine card are shown in the table below. The directions of the signals are from the perspective of the baseboard.

Table 1: OCP 2.0 Hybrid Mezzanine Connector Pinout

Connector A				Connector B			
Signal	**Pin**	**Pin**	**Signal**	**Signal**	**Pin**	**Pin**	**Signal**
P12V_AUX	A61	A1	MEZZ_PRSNTA1_N /BASEBOARD_A_ID	P12V_AUX	B41	B1	MEZZ_PRSNTB1_N /BASEBOARD_B_ID
P12V_AUX	A62	A2	P5V_AUX	P12V_AUX	B42	B2	GND
P12V_AUX	A63	A3	P5V_AUX	NC	B43	B3	KR_RX_DP<0>
GND	A64	A4	P5V_AUX	GND	B44	B4	KR_RX_DN<0>
GND	A65	A5	GND	KR_TX_DP<0>	B45	B5	GND
P3V3_AUX	A66	A6	GND	KR_TX_DN<0>	B46	B6	GND
GND	A67	A7	P3V3_AUX	GND	B47	B7	KR_RX_DP<1>
GND	A68	A8	GND	GND	B48	B8	KR_RX_DN<1>
P3V3	A69	A9	GND	KR_TX_DP<1>	B49	B9	GND
P3V3	A70	A10	P3V3	KR_TX_DN<1>	B50	B10	GND

P3V3	A71	A11	P3V3
P3V3	A72	A12	P3V3
GND	A73	A13	P3V3
LAN_3V3STB_ALERT_N	A74	A14	NCSI_RCSDV
SMB_LAN_3V3STB_CLK	A75	A15	NCSI_RCLK
SMB_LAN_3V3STB_DAT	A76	A16	NCSI_TXEN
PCIE_WAKE_N	A77	A17	PERST_N0
NCSI_RXER	A78	A18	MEZZ_SMCLK
GND	A79	A19	MEZZ_SMDATA
NCSI_TXD0	A80	A20	GND
NCSI_TXD1	A81	A21	GND
GND	A82	A22	NCSI_RXD0
GND	A83	A23	NCSI_RXD1
CLK_100M_MEZZ0_DP	A84	A24	GND
CLK_100M_MEZZ0_DN	A85	A25	GND
GND	A86	A26	CLK_100M_MEZZ1_DP
GND	A87	A27	CLK_100M_MEZZ1_DN
PCIe_MEZZ0_TX_DP_C<0>	A88	A28	GND
PCIe_MEZZ0_TX_DN_C<0>	A89	A29	GND
GND	A90	A30	PCIe_MEZZ0_RX_DP<0>
GND	A91	A31	PCIe_MEZZ0_RX_DN<0>
PCIe_MEZZ0_TX_DP_C<1>	A92	A32	GND
PCIe_MEZZ0_TX_DN_C<1>	A93	A33	GND
GND	A94	A34	PCIe_MEZZ0_RX_DP<1>
GND	A95	A35	PCIe_MEZZ0_RX_DN<1>
PCIe_MEZZ1_TX_DP_C<0>	A96	A36	GND
PCIe_MEZZ1_TX_DN_C<0>	A97	A37	GND
GND	A98	A38	PCIe_MEZZ1_RX_DP<0>
GND	A99	A39	PCIe_MEZZ1_RX_DN<0>
PCIe_MEZZ1_TX_DP_C<1>	A100	A40	GND
PCIe_MEZZ1_TX_DN_C<1>	A101	A41	GND
GND	A102	A42	PCIe_MEZZ1_RX_DP<1>
GND	A103	A43	PCIe_MEZZ1_RX_DN<1>
PCIe_MEZZ2_TX_DP_C<0>	A104	A44	GND
PCIe_MEZZ2_TX_DN_C<0>	A105	A45	GND
GND	A106	A46	PCIe_MEZZ2_RX_DP<0>
GND	A107	A47	PCIe_MEZZ2_RX_DN<0>
PCIe_MEZZ2_TX_DP_C<1>	A108	A48	GND
PCIe_MEZZ2_TX_DN_C<1>	A109	A49	GND
GND	A110	A50	PCIe_MEZZ2_RX_DP<1>
GND	A111	A51	PCIe_MEZZ2_RX_DN<1>
PCIe_MEZZ3_TX_DP_C<0>	A112	A52	GND
PCIe_MEZZ3_TX_DN_C<0>	A113	A53	GND
GND	A114	A54	PCIe_MEZZ3_RX_DP<0>

GND	B51	B11	KR_RX_DP<2>
GND	B52	B12	KR_RX_DN<2>
KR_TX_DP<2>	B53	B13	GND
KR_TX_DN<2>	B54	B14	GND
GND	B55	B15	KR_RX_DP<3>
GND	B56	B16	KR_RX_DN<3>
KR_TX_DP<3>	B57	B17	GND
KR_TX_DN<3>	B58	B18	GND
GND	B59	B19	NC
GND	B60	B20	NC
NC	B61	B21	GND
NC	B62	B22	GND
GND	B63	B23	NC
GND	B64	B24	NC
NC	B65	B25	GND
NC	B66	B26	GND
GND	B67	B27	NC
GND	B68	B28	NC
NC	B69	B29	GND
NC	B70	B30	GND
GND	B71	B31	NC
GND	B72	B32	NC
NC	B73	B33	GND
NC	B74	B34	GND
GND	B75	B35	CLK_100M_MEZZ2_DP
GND	B76	B36	CLK_100M_MEZZ2_DN
CLK_100M_MEZZ3_DP	B77	B37	GND
CLK_100M_MEZZ3_DN	B78	B38	PERST_N1
GND	B79	B39	PERST_N2
MEZZ_PRSNTB2_N	B80	B40	PERST_N3

GND	A115	A55	PCIe_MEZZ3_RX_DN<0>
PCIe_MEZZ3_TX_DP_C<1>	A116	A56	GND
PCIe_MEZZ3_TX_DN_C<1>	A117	A57	GND
GND	A118	A58	PCIe_MEZZ3_RX_DP<1>
GND	A119	A59	PCIe_MEZZ3_RX_DN<1>
MEZZ_PRSNTA2_N	A120	A60	GND

5.6.2 Hybrid Mezzanine Card Pin Description

The Hybrid Mezzanine card pin description is shown in the table below. The signal direction is in the perspective of the baseboard.

Table 2: OCP 2.0 Mezzanine Card Pin Description

Signals on Connector A	Type	Description
GND	Ground	Ground return; total 51 pins on Connector A
P12V_AUX	Power	12V Aux power; total 3 pins on Connector A
P5V_AUX	Power	5V Aux power; total 3 pins on Connector A
P3V3_AUX	Power	P3V3 Aux Power; total 2 pins on Connector A
P3V3	Power	P3V3 power; total 8 pins on Connector A
MEZZ_PRSNTA1_N/BASEBOARD_ID_A	Output	Connector A Present Pin; connect to MEZZ_PRSNTA2_N on Mezz with 0 Ohm; Use as baseboard ID during power up
MEZZ_PRSNTA2_N	Input	Connector A Present Pin; connect to MEZZ_PRSNTA1_N on Mezz with 0 Ohm
LAN_3V3STB_ALERT_N	Input	SMBus Alert for OOB management; 3.3V AUX rail
SMB_LAN_3V3STB_CLK	Output	SMBus Clock for OOB management; 3.3V AUX rail; Share with thermal reporting interface
SMB_LAN_3V3STB_DAT	Bidirectional	SMBus Data for OOB management; 3.3V AUX rail; Share with thermal reporting interface
NCSI_RXER	Input	NC-SI for OOB management
NCSI_RCSDV	Input	NC-SI for OOB management
NCSI_RXD[1..0]	Input	NC-SI for OOB management
NCSI_RCLK	Output	NC-SI for OOB management
NCSI_TXEN	Output	NC-SI for OOB management
NCSI_TXD[1..0]	Output	NC-SI for OOB management
PCIE_WAKE_N	Input	PCIe wake up signal
PERST_N0	Output	PCIe reset signal or Node 0 PCIe reset signal
MEZZ_SMCLK	Output	PCIe SMBus Clock for Mezz slot/EEPROM; 3.3V AUX rail; Share with thermal reporting interface
MEZZ_SMDATA	Bidirectional	PCIe SMBus Data for Mezz slot/EEPROM; 3.3V AUX rail; Share with thermal reporting interface

CLK_100M_MEZZ[1..0]_DP/N	Output	PCIe Reference clock from Node [1..0]
PCIe_MEZZ0_TX_DP/N_C<1..0>	Output	PCIe TX from Node 0; total up to 2 lanes on Connector A
PCIe_MEZZ0_RX_DP/N<1..0>	Input	PCIe RX to Node 0; total up to 2 lanes on Connector A
PCIe_MEZZ1_TX_DP/N_C<1..0>	Output	PCIe TX from Node 1; total up to 2 lanes on Connector A
PCIe_MEZZ1_RX_DP/N<1..0>	Input	PCIe RX to Node 1; total up to 2 lanes on Connector A
PCIe_MEZZ2_TX_DP/N_C<1..0>	Output	PCIe TX from Node 2; total up to 2 lanes on Connector A
PCIe_MEZZ2_RX_DP/N<1..0>	Input	PCIe RX to Node 2; total up to 2 lanes on Connector A
PCIe_MEZZ3_TX_DP/N_C<1..0>	Output	PCIe TX from Node 3; total up to 2 lanes on Connector A
PCIe_MEZZ3_RX_DP/N<1..0>	Input	PCIe RX to Node 3; total up to 2 lanes on Connector A

Signals on Connector B	Type	Description
GND	Ground	Ground return; total 36 pins on Connector B
P12V_AUX	Power	12V Aux power; total 2 pins on Connector B
MEZZ_PRSNTB1_N/ BASEBOARD_ID_B	Output	Connector B Present Pin; connect to MEZZ_PRSNTB2_N on Mezz with 0 Ohm Use as baseboard ID during power up
MEZZ_PRSNTB2_N	Input	Connector B Present Pin; connect to MEZZ_PRSNTB1_N on Mezz with 0 Ohm
PERST_N[3..1]	Output	PCIe reset signal from Node[3..1]
CLK_100M_MEZZ[3..2]_DP/N	Output	PCIe Reference clock from Node [3..2]
KR_TX_DP/N<3..0>	Output	KR TX from Node[3..0]
KR_RX_DP/N<3..0>	Input	KR RX to Node[3..0]
NC	Open	These signals are not connected on adapter card.

5.6.3 OCP 2.0 4x10G KR-Retimer Mezzanine Card Design

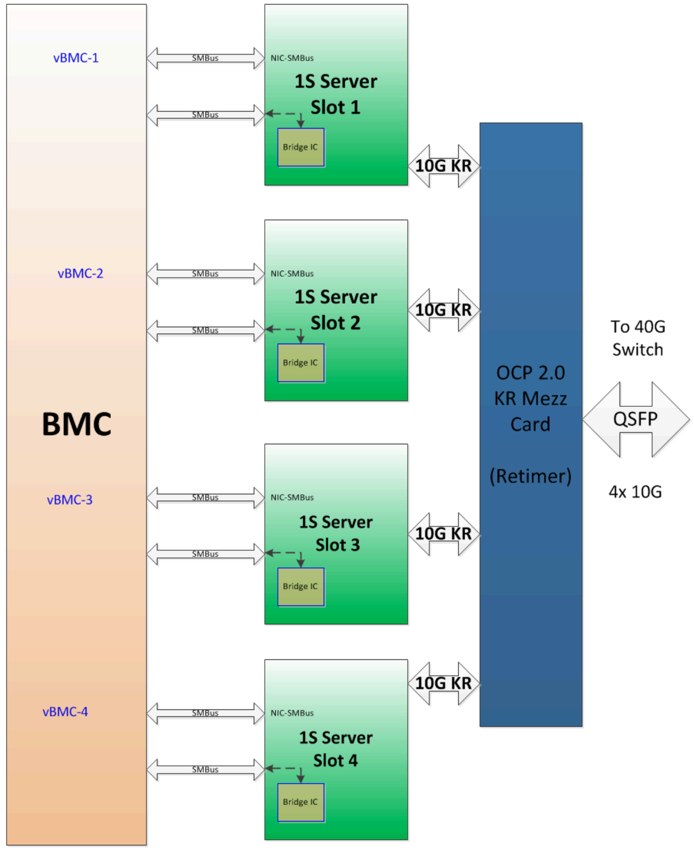


Figure 5-5: Yosemite with 4X10G KR-Retimer Mezzanine Card

For the 1S servers with built-in 10GBase-KR network controllers, a 4x10G KR-retimer Mezzanine card can be used to carry out 4 KR links in a single QSFP+ cable to a 40G TOR switch port that is configured in 4x10G mode. Figure 5-5 above illustrates a Yosemite platform that uses a 4x10G KR-retimer Mezzanine card.

A KR-aware retimer, such as Inphi's CS4223 or Semtech's GN2007, is used to boost signal quality and compensate for the channel loss of four independent 10GbE links from the 1S servers to the Mezzanine connectors. On the host side, the retimer shall fully support auto-negotiation and link training of 10GBase-KR protocol so that it can establish 10GbE links with the 1S servers. On the network side, the retimer shall support a single QSFP port, which can use QSFP copper or fiber cables.

The built-in 10GBase-KR network controller is used as a shared-NIC and its SMBus is connected to the BMC as the sideband. The BMC shall configure the network controller properly so that the network controller can bypass management traffic to the corresponding virtual BMC through the sideband.

Figure 5-6 illustrates the details of a 4x10G KR-retimer Mezzanine card.

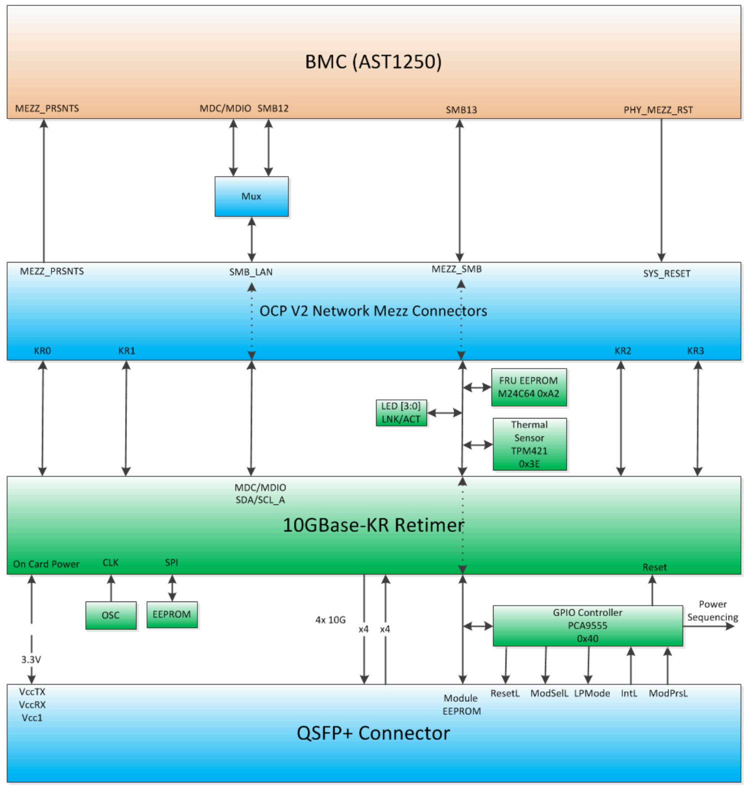


Figure 5-6: 4X10G KR-Retimer Mezzanine Card Block Diagram

The BMC manages the KR-retimer Mezzanine card through MEZZ_SMB. Before powering up a Mezzanine card, the BMC shall read the FRU EEPROM on the card to determine card type, configuration method, and functions. With the information in the FRU EEPROM, the BMC needs to flip the multiplexer on the side plane properly to set SMB_LAN to either SMBus or MDC/MDIO on the BMC. After that, the BMC uses the GPIO expender on MEZZ_SMB to enable power supplies on the card, reset the retimer, and configure the retimer accordingly through SMB_LAN.

The BMC uses the GPIO expender to interact with a QSFP module as illustrated above. When there is a QSFP cable plugged into the card, the BMC shall read the EEPROM on the module to collect configuration information and then adjust the retimer's transceiver parameters accordingly.

A thermal sensor, TPM421 preferred, resides on the MEZZ_SMB or SMB_LAN. The BMC monitors the thermal status of the card through this thermal sensor and controls the fans along with other thermal sensors in the platform. If a retimer offers an on-chip thermal sensor, it shall be accessible to the BMC through MEZZ_SMB or SMB_LAN.

It is possible to configure the retimer through an EEPROM instead of going through a BMC. However, the BMC must be able to update the firmware on the EEPROM if this configuration method is used.

The 4x10 KR-retimer Mezzanine card must be compliant with mechanical and thermal requirements defined in the OCP 2.0 Mezzanine specification. A 15W, or less, maximum total card power is strongly recommended to accommodate the Yosemite Platform's power and thermal restrictions.

There are four LEDs on the card to indicate link and activity status of all 10GBase-KR interfaces. When the link is active, the LED shall be on. Where there is activity, the LED shall blink. It is preferred to have the retimer control these blue LEDs. However, if the retimer does not support LED control function, the BMC can drive the LEDs through an I2C LED controller on MEZZ_SMB, while the BMC gets link and status information from the Bridge IC on the 1S Servers. All LEDs shall be placed on front side of the sled, visible to the operators.

5.6.4 OCP 2.0 Multi-Host 40G/50G NIC Mezzanine Card Design

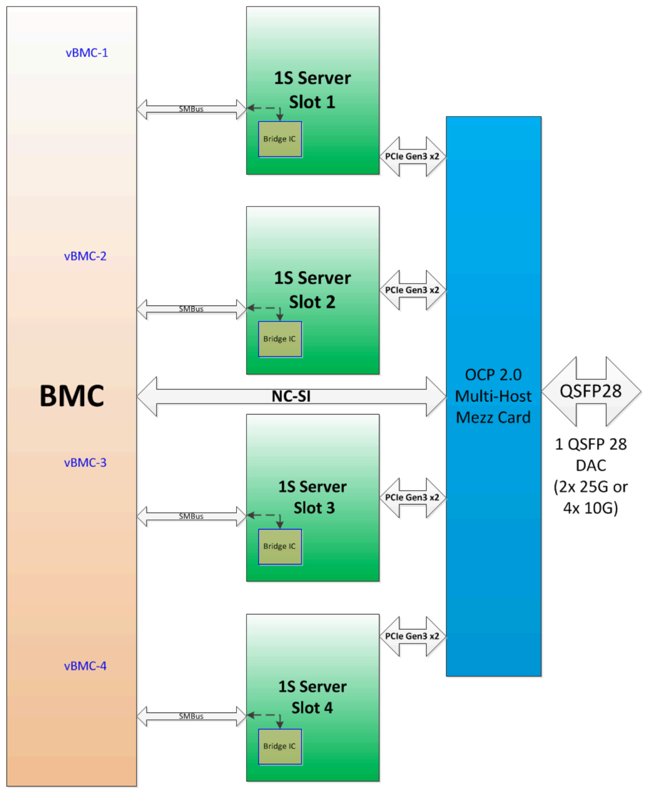


Figure 5-7: Yosemite with multi-host 40G/50G PCIe Mezzanine Card

The network controller vendors offer Network Interface Cards (NIC) that support multi-host functions. These multi-host NIC cards provide network connectivity to multiple servers through a PCIe interface. For example, Mellanox's ConnectX-4 has PCIe interfaces to up to four independent servers with single or dual network interfaces to a TOR switch.

The Yosemite Platform fully supports OCP 2.0 40G/50G multi-host NIC through PCIe interfaces. As shown in Figure 5-8, there is a PCIe Gen3 link with 2 lanes from every 1S Server connected to the multi-host NIC through OCP 2.0 Mezzanine connectors. Every 1S server also provides its own PCIe reference clock, and PCIe reset to the multi-host NIC. With this configuration, every 1S server can operate its portion of the NIC independently regardless of any other 1S Server's status.

On the network side, the multi-host NIC implements a single QSFP28 port, which can be configured automatically to 40G mode with four 10G lanes, or 50G mode with two 25G lanes, through auto-negotiation with the TOR switch. To meet the strict signal integrity requirement of the whole channel between the multi-host NIC and a TOR switch in 25Gbps per lane speed, the channel loss from NIC's network transceivers to the QSFP28 module must be less than 5dB. A channel loss of 3dB or less is strongly recommended. The NIC shall support both copper and fiber cables in various lengths.

There is a BMC on the Yosemite Platform as the management controller. Depends on software implementation, the BMC can be virtualized so that every virtual BMC is assigned to a 1S Server, or the BMC can manage all servers on a per slot basis. The multi-host NIC is used as a shared NIC on the platform and its NC-SI interface is used as the BMC's sideband. As there is only one NC-SI interface between the multi-host NIC and the BMC, it can be virtualized to provide dedicated sideband to every virtual BMC, or just provide one sideband to the whole BMC.

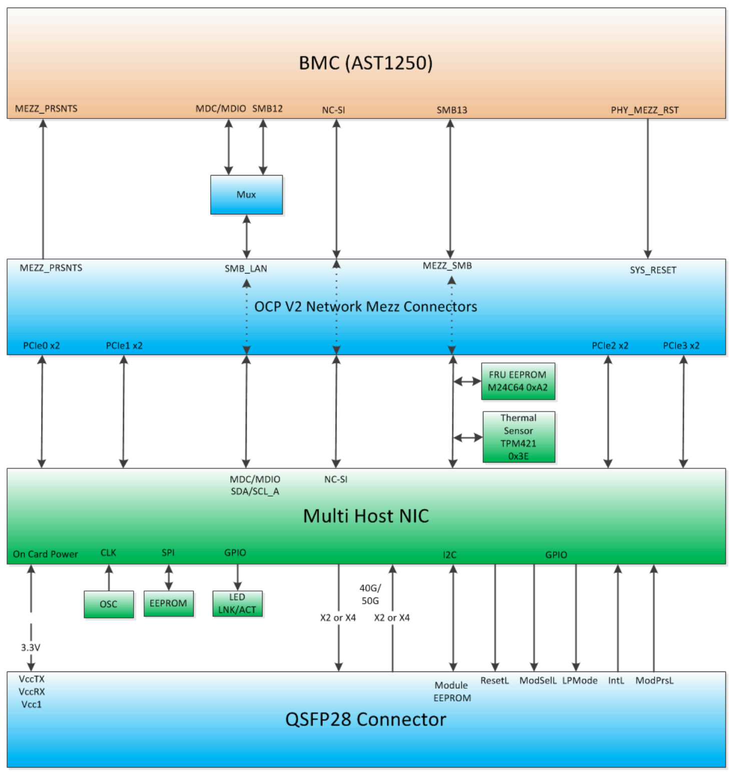


Figure 5-8: Multi-host 40G/50G PCIe Mezzanine Card Block Diagram

Figure 5-8 illustrates the details of a multi-host 40G/50G NIC Mezzanine card.

The BMC manages the multi-host NIC Mezzanine card through MEZZ_SMB. As the multi-host NIC is an intelligent device, it will manage power, reset, and LEDs on its own. The multi-host is powered on when the Yosemite Platform is powered on. The BMC shall read the FRU EEPROM on the card to determine card type, functions and configure the NC-SI interface.

The multi-host NIC also manages the QSFP28 module as illustrated above. If there is a QSFP28 cable plugged into the card, the multi-host NIC shall read the EEPROM on the module to collect configuration information then adjust its transceiver parameters accordingly.

A thermal sensor, TPM421 preferred, resides on MEZZ_SMB or SMB_LAN. The BMC monitors the thermal status of the card through this thermal sensor and control the fans along with other thermal sensors in the platform. If there is an internal thermal sensor on the multi-host network controller IC, it shall be accessible to the BMC through MEZZ_SMB or SMB_LAN.

The multi-host NIC's firmware and configuration EEPROM can be updated by any of the 1S Servers or the BMC. However, the NIC must implement a locking mechanism to eliminate race conditions such as two or more 1S Servers and/or the BMC trying to access the firmware and/or configuration EEPROM at same time.

The multi-host NIC Mezzanine card must be compliant with the mechanical and thermal requirements defined in the OCP 2.0 Mezzanine specification. 15W or less maximum total card power is strongly recommended to accommodate the Yosemite Platform's power and thermal restrictions.

There are two LEDs on the card to indicate link and activity status of the link between the multi-host NIC and TOR switch. When the link is up, the link LED shall be on. Where there is activity, the activity LED shall blink. The NIC shall control this LED independently without BMC's involvement. The LED shall be placed on front side of the sled and visible to the operators.

5.6.5 OCP 2.0 50G KR-Aggregation Mezzanine Card Design

For the 1S servers with built-in 10GBase-KR network controllers, a 50G KR-Aggregation Mezzanine card can be used as an alternative solution for 100G-network environment. As illustrated below, OCP 2.0 50G KR-Aggregation Mezzanine card uses a KR-Aggregator or port expender to aggregate four 10GBase-KR links into a single 50G link to a 100G TOR switch.

A 50G KR-Aggregator on the card is used to boost signal quality and compensate for the channel loss of four independent 10GbE links from the 1S servers to the Mezzanine connectors. It shall fully support auto-negotiation and link training of the 10GBase-KR protocol so that it can establish 10GbE links with the 1S servers. On the network side, the 50G KR-Aggregator shall support a single QSFP28 port, which can use QSFP28-100G-capable copper or fiber cables. To meet the strict signal integrity requirements of the whole channel between the KR-Aggregator and a TOR switch in 25Gbps speed, the channel loss from KR-Aggregator's network transceivers to the QSFP28 module must be less than 5dB., a channel loss of 3dB or less is strongly recommended.

The built-in 10GBase-KR network controller is used as a shared-NIC and its SMBus is connected to the BMC as the sideband. The BMC shall configure the network controller properly so that the network controller can bypass management traffic to the corresponding virtual BMC through the sideband.

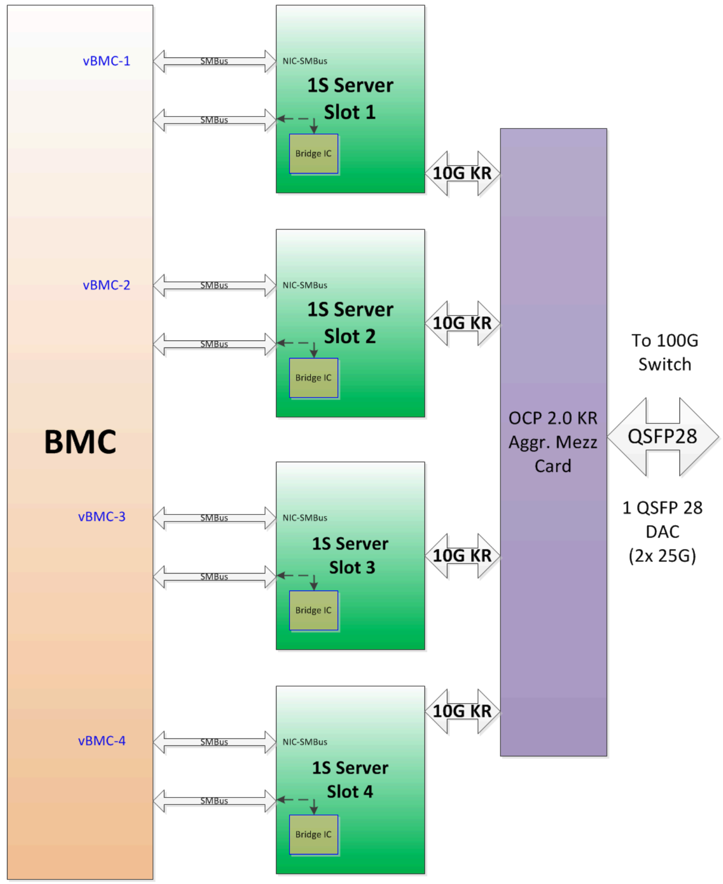


Figure 5-9: Yosemite with 50G KR-Aggregation Mezzanine Card

Figure 5-10 illustrates the details of a 50G KR-Aggregation Mezzanine card.

The BMC manages the KR-Aggregation Mezzanine card through MEZZ_SMB. Before powering up a Mezzanine card, the BMC shall read the FRU EEPROM on the card to determine card type, configuration method and functions. With the information in the FRU EEPROM, the BMC needs to flip the multiplexer on the side plane properly to set SMB_LAN to either SMBus or MDC/MDIO on the BMC. After that, the BMC uses the GPIO expender on the MEZZ_SMB to enable power supplies on the card, reset the retimer, and configure the retimer accordingly through SMB_LAN.

The BMC uses the GPIO expender to interact with a QSFP28 module as illustrated below. When there is a QSFP28 cable plugged into the card, the BMC shall read the EEPROM on the module to collect configuration information and then adjust the KR-Aggregator's transceiver parameters accordingly.

A thermal sensor, TPM421 preferred, resides on the MEZZ_SMB or SMB_LAN. The BMC monitors the thermal status of the card through this sensor and controls the fans along with other thermal sensors in the platform. If a KR-Aggregator offers an on-chip thermal sensor, it shall be accessible to the BMC through the MEZZ_SMB or SMB_LAN.

It is possible to configure the KR-Aggregator through an EEPROM instead of going through a BMC. However, the BMC must be able to update the firmware on EEPROM if this configuration method is used.

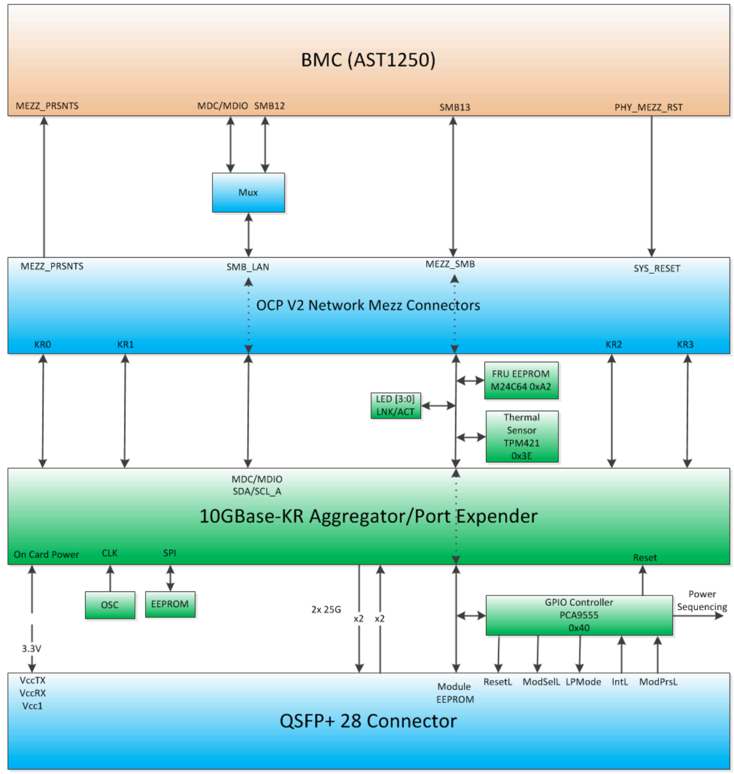


Figure 5-10: 50G KR-Aggregation Mezzanine Card Block Diagram

The 50GKR-Aggregation Mezzanine card must be compliant with all mechanical and thermal requirements defined in the OCP 2.0 Mezzanine specification. 15W or less maximum total card power is strongly recommended to accommodate the Yosemite Platform's power and thermal restrictions.

There are five LEDs on the card to indicate a link and the activity status of the link between the KR-Aggregation Card and the TOR switch and the links from 1S Server's KR network controller to the KR-Aggradation Card. When the link is up, the LED shall be on. Where there is activity, the LED shall blink. For the link between KR-Aggregation Card and the TOR switch, the KR-Aggregator shall control this LED independently, without the BMC's involvement. For the link between KR-Aggregation Card and 1S Server's KR network controllers, the BMC shall control these LEDs and it gets the link and status information from the Bridge IC on the 1S Servers.

These LEDs shall be placed on front side of the sled and shall be visible to the operators.

6 Baseboard Management Controller

The Yosemite Platform uses a BMC for various platform management services and interfaces with hardware and BIOS firmware. The proposed BMC is ASPEED's AST1250.

The BMC should be a stand-alone system in parallel to the 1S servers. The health status of the 1S servers should not affect the normal operation and network connectivity of the BMC.

6.1 1S Server I^2C Connections

There is a Bridge IC on each 1S server as a satellite management controller. The proposed Bridge IC is TI's Tiva micro-controller. The Intelligent Platform Management Bus Communications (IPMB) (I^2C) connection from the Bridge IC on 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 speed of the I^2C connection is 1.0MHz if an AST1250 used as the BMC and Texas Instruments' Tiva Microcontroller is used as Bridge IC. If a different BMC or Bridge IC is used, a maximum speed of 400kHz is expected, as per the I^2C specification.

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.

6.1.1 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.

6.2 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 can use a switch to select which 1S server is connected to the debug card. By default, the BMC enables SOL to all 1S servers. When a SOL session is activated on the selected 1S server, the BMC shall keep SOL alive all the time and only use serial port on front panel as a snooping port to avoid possible data collisions. If a full duplex serial port access from the OCP debug card to a 1S server is preferred, the user has to deactivate the SOL session associated to this particular 1S server first.

6.3 1S Server Discovery Process

6.3.1 Initial Discovery

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

1. The BMC collects the FRU information from Bridge-IC
2. The BMC sensor tables are updated from Bridge-IC.

3. The pin assignment tables are loaded in to the 1S server via Bridge-IC.
4. The card is powered on based on the user input or as defined by the last AC power state.

6.3.2 Pin Assignment Tables

As defined in the 1S server specification, (http://www.opencompute.org/wiki/Motherboard/SpecsAndDesigns, V0.7), a table of pin assignments must be provided to each 1S server. The table below describes the capabilities of Yosemite Platform.

Table 3: Yosemite Platform Pin Assignment Table

Byte #	Byte value	Note
0	0x03	A13/A14 PCIe0 RefClk
1	0x02	A17/A18 PCIe0 Lane 0, Gen3
2	0x02	A21/A22 PCIe0 Lane 1, Gen3
3	0x02	A25/A26 PCIe0 Lane 2, Gen3
4	0x02	A29/A30 PCIe0 Lane 3, Gen3
5	0x05	A33/A34 SATA0, Gen3
6	0x03	A37/A38 PCIe2 RefClk
7	0x02	A49/A50 PCIe1 Lane 0 Gen3
8	0x02	A53/A54 PCIe1 Lane 1 Gen3
9	0x02	A57/A58 PCIe1 Lane 2 Gen3
10	0x02	A61/A62 PCIe1 Lane 3 Gen3
11	0x02	A65/A66 PCIe2 Lane 0 Gen3
12	0x02	A69/A70 PCIe2 Lane 1 Gen3
13	0x02	A73/A74 PCIe2 Lane 2 Gen3
14	0x02	A77/A78 PCIe2 Lane 3 Gen3
15	0x02	B15/B16 PCIe0 Lane 0, Gen3
16	0x02	B19/B20 PCIe0 Lane 1, Gen3
17	0x02	B23/B24 PCIe0 Lane 2, Gen3
18	0x02	B27/B28 PCIe0 Lane 3, Gen3
19	0x05	B31/B32 SATA0, Gen3
20	0x03	B35/B36 PCIe1 RefClk
21	0x02	B51/B52 PCIe1 Lane 0 Gen3
22	0x02	B55/B56 PCIe1 Lane 1 Gen3
23	0x02	B59/B60 PCIe1 Lane 2 Gen3
24	0x02	B63/B64 PCIe1 Lane 3 Gen3
25	0x02	B67/B68 PCIe2 Lane 0 Gen3
26	0x02	B71/B72 PCIe2 Lane 1 Gen3
27	0x02	B75/B76 PCIe2 Lane 2 Gen3
28	0x02	B79/B80 PCIe2 Lane 3 Gen3

The BMC must write the pin assignment to the EEPROM on the 1S server card via the Bridge IC. This ensures the BIOS can properly configure the SOC or turn off I/O that is incompatible. Before writing the pin assignment table, the Bridge IC shall check if the new table is different than the current one to avoid unnecessary updating.

6.4 1S Server Power-on Sequence

The BMC will de-assert the PWR_BTN# and SYS_RESET# signals to the 1S server to initiate power-on. The BMC will then poll the Power Good status from the 1S server to confirm if the 1S server card has powered-on successfully. It will then update the power status.

6.5 Network Interface

The BMC has two possible network paths. First, if a PCIe based multi-host 40/50GbE NIC mezzanine card is used, the BMC can use its built-in media access controller (MAC) to transfer management traffic through an NC-SI interface with TOR switch.

Second, the Yosemite Platform may only have a KR PHY card on the Mezzanine slot and use the 1S server's built-in NICs. In this case, the BMC will use the NIC SMBus connections going from the BMC to each 1S server slot for OOB management traffic.

The Mezzanine card needs to provide a Field Replaceable Unit ID (FRUID) as per the OCP 2.0 Mezzanine card specification. The BMC shall use this FRUID to identify the card type and configure network paths accordingly. 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.

6.6 BMC Multi-node Requirements

Since there are 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 a particular server. Depending on software implementation, each vBMC may have a unique IP address so that the remote clients can access and manage the server, or they may share a single IP address from the sideband.

6.7 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
- A SOL console

These must be supported through the management network described in Section 6.5. 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. For example, display as "[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.

6.8 Graphics and GUI

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

6.9 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 OOB Intelligent Platform Management Interface (IPMI) command.

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 by an IPMI command and take effect without cold resetting the BMC firmware or rebooting the 1S Server system.

If AC power is flowing 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.

6.10 POST Codes

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

6.11 Power and System Identification LEDs

The Yosemite Platform combines a Power LED and a System Identification LED into a single bicolor blue-yellow LED per 1S server. A total of 4×LEDs will be placed along the front edge of the board in a grid. The grid will be 2×rows of 2×LEDs to match the layout of the 1S server slots.

When the Power LED is on, it defines the readiness of all power rails on the 1S server card. It also indicates the status of the card (overall health).

When the Yosemite Platform is being identified by the BMC, all four yellow LEDs blink at 2.5Hz simultaneously, with 50% duty cycle. All four blue LEDs shall be off, regardless of the power status of the 1S Servers. During this operation, all identification requests for an individual 1S

Server inside the sled are ignored by the BMC. The identification operation shall continue until the user withdraws the identification request. This operation has the highest priority.

When an individual 1S Server is being identified by the BMC, its corresponding yellow LED blinks at 2.5Hz, with 50% duty cycle. Its blue LED shall be off regardless of its power state. The identification operation shall continue until the user withdraws the identification request. This operation has second level priority.

When the selector knob is turned to the BMC position, all four blue LEDs blink at 1HZ simultaneously, with 50% duty cycle. All four yellow LEDs shall be off. This operation has third level priority. All requests other than identification are ignored by the BMC.

If a 1S Server is not selected as the current server and is not being identified by the BMC, its corresponding LED shall operate according to the server's status. When the 1S Server is powered off, both the blue LED and the yellow LED shall be off. If the 1S Server is powered on and the server is operating normally, the blue LED shall be on and not blinking. The yellow LED shall be off. If the 1S Server is powered on and the server operates abnormally, such as a bad power state or if critical errors have been logged, the yellow LED shall be on and not blinking. The blue LED shall be off. This operation has the lowest priority.

If a 1S Server is selected as the current server but is not being identified by the BMC, its corresponding LED shall operate according to the server's status. When the 1S Server is powered off, the blue LED blinks at 1Hz with 10% duty cycle. The yellow LED shall be off. If the 1S Server is powered on and the server operates normally, the blue LED blinks at 1Hz with 90% duty cycle. The yellow LED shall be off. If the 1S Server is powered on and the server operates abnormally, such as a bad power state or critical errors have been logged, the yellow LED blinks at 1Hz with 90% duty cycle. The blue LED shall be off. This operation has lowest priority.

The Power LED blinks in different ways (varying colors and times) to convey various system statuses. There are fix different situations are defined as below to make the LED operation more user friendly.

Situation 1: when the Yosemite sled is being identified by the BMC, the BMC shall simultaneously blink all 4 yellow LED in 2.5 HZ (200ms on, 200ms off) while keeping all blue LEDs off. In this situation, the BMC shall ignore identification requests for individual servers inside this platform.

Situation 2: when the selector knob is turned to the BMC position, the BMC shall blink all 4 blue LEDs in 1HZ and 50% duty cycle (500ms on, 500ms off) simultaneously while keeping all yellow LEDs off.

Situation 3: when a 1S server is not selected as the current server by the knob, but it is being identified by the BMC, regardless of its power status, the BMC shall blink this server's yellow LED in 2.5HZ frequency and 50% duty cycle while keeping this server's blue LED off.

Situation 4: when a 1S server is not selected as the current server by the knob, and it is not being identified by BMC, the BMC shall controls the LEDs according to 1S server's power and health status as below:

If the 1S server is in power off state, the BMC shall turn off both this server's blue and yellow LEDs.

If the 1S server is in power on state and operates, the BMC shall turn this server's blue LED solid on while keeping the yellow LED off.

If the 1S server is in power on state but operates abnormally, the BMC shall turn this server's yellow LED solid on but keep the blue LED off.

Situation 5: when a 1S server is the current server selected by the knob and it is being identified by the BMC, regardless of this server's power status, the BMC shall blink this server's yellow LED in 2.5HZ and 50% duty cycle while keeping the blue LED off.

Situation 6: when a 1S server is the current server selected by the knob and it is not being identified by the BMC, the BMC shall control the LEDs according to the 1S server's power and health status as below:

If the server is in the power off state, the BMC shall blink this server's blue LED in 1HZ frequency 10% duty cycle while keeping the yellow LED off. (100ms on, 900ms off)

If the server is in the power on state and operates normally, the BMC shall blink this server's blue LED in 1HZ frequency 90% duty cycle while keeping the yellow LED off. (900ms on, 100ms off)

If the server is in the power on state but operates abnormally, the BMC shall blink this server's yellow LED in 1HZ frequency 90% duty cycle while keeping the blue LED off. (900ms on, 100ms off)

6.12 Time Sync

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

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

The BMC should sync its clock from the NTP server periodically.

The 1S server BIOS will issue an IPMI Get System Event Log (SEL) Time command to sync its system clock during POST. Afterwards, no additional time sync comes from BIOS.

6.12.1 NTP Time Sync Flow

1. BMC first time power on.
2. BMC firmware image contains the default NTP IP address and NTP retry timeout.
3. Provisioning server will also send a Set NTP IP Address command to the BMC via a Set NTP Server OEM command. The command, at the same time, configures the NTP retry timeout.
4. BMC network interface up and running.
5. BMC queries date/time for the BMC clock using the configured NTP IP address.
   - If successful, the IPMI Set SEL Time command updates the BMC system clock. The BMC should not enable the SEL log until its date/time has been synced with the NTP server.
   - Failure triggers a three-retry mechanism. The retry timeout interval could be incremental (the first retry is 30 seconds, the second retry is 60 seconds, etc.).
6. If retries fail three times, a SEL log will be enabled and an NTP date/time sync event will be generated. A default hardware date/time (Jan 1. 1970) will be used for any event log.
7. Meanwhile, the BMC will try to re-sync its date/time for a longer time interval.

8. Once the BMC date/time is synced, the BMC will adjust the SEL log entries since the NTP date/time sync event (with hardware date/time (Jan 1, 1970) with the actual time difference between the BMC clock and the synced date/time from the NTP server.
9. The BMC will sync its date/time from the NTP server periodically with an interval (in hours). The interval is configured via a Set NTP Server OEM command.

6.13 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 TMP421 with an external PN junction is preferred to detect inlet and outlet temperatures. Take caution when implementing inlet air sensors. It is important to avoid preheating nearby components and to reduce the amount of heat conducted through the printed circuit board (PCB).

The BMC firmware 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 a SOC-specific management controller interface to limit the SOC’s power consumption (e.g., P-State control).

6.14 Sensors

Both analog and discrete sensors may reside on the side plane and on the 1S server cards. The BMC must consolidate the sensor information such that an IPMI command returns the sensor information for a specific 1S server plus information from the common sensors.

6.14.1 Analog Sensors

The BMC has access to all analog sensors on the Yosemite Platform directly or through the 1S server management connections. The sensor data record (SDR) repository must display all analog sensors.

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

6.14.2 Discrete Sensors

The BMC firmware shall provide discrete sensors in the SDR. The BMC should log abnormal sensor readings to the SEL.

Some of the required discrete sensors include but are not limited to:

- System Status
- Power Threshold Event
- SEL Status
- DCMI Watchdog
- Processor Failure
- Chassis Power Status
- Thermal Limits
- NTP Status
- PMBUS Status

6.14.3 Event Only Sensors

The event only sensors are not shown in SDR. These sensors trigger SEL if an abnormal value is detected. Some of the required event only sensors included but not limited to:

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

6.15 SEL

The vendor should implement the BMC to support SEL for each 1S server. All errors that are logged must include the slot ID of the 1S server.

6.15.1 Logged Errors

6.15.1.1 CPU Error

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

6.15.1.2 Memory Error

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

6.15.1.3 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.

6.15.1.4 POST Error

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

6.15.1.5 Power Error

Two power errors should be logged. One is a 12.5V DC input power failure that causes all power rails on the side plane 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.

6.15.1.6 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.

6.15.1.7 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.

6.15.1.8 PMBus Status Error

The PMBus status sensors check the PMBus controllers' 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.

6.15.2 Error Threshold Setting

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

6.16 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.

6.16.1 Fan Speed Control Specification

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

6.16.2 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 4: Required FSC Data

Type of data	Data used 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

6.16.3 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.

6.16.4 Fan Speed Controller Update

The BMC must implement the FSC update commands as described in the OPC IPMI-FSC update commands specification.

6.16.5 Fan Connection

The Yosemite Platform side plane has 2× fan headers on the motherboard.

6.17 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 5 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.

As shown in the system block diagram, the BMC firmware update should be possible through a USB path from one of the 1S server to the BMC's virtual USB hub port. The BMC needs to toggle select pins on the USB mux to set up the USB path right before updating.

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

7 Mechanical

Yosemite Platform is an Open Rack V2 compatible compute platform via the Cubby three–bay shelf for Open Rack V2. Cubby has 2x16 OpenU power zones. Each power zone can hold 8x2 OpenU Cubbies, and each Cubby can hold up to 3x Yosemite Platform sleds. Each Yosemite Platform sled consists of a sheet metal chassis, one Yosemite Platform side plane, and other components, such as guide features for PCBs. The chassis enables up to four PCBs mounted via PCIe to the side plane (see the general card specification).

7.1 Cubby Chassis

Cubby is a power-mechanical shelf distributing power from the rack bus bars to three sled bays per shelf. Figure 7-1 shows a Cubby with the maximum available sled volume (yellow). One Yosemite Platform chassis will occupy the space of this sled volume.

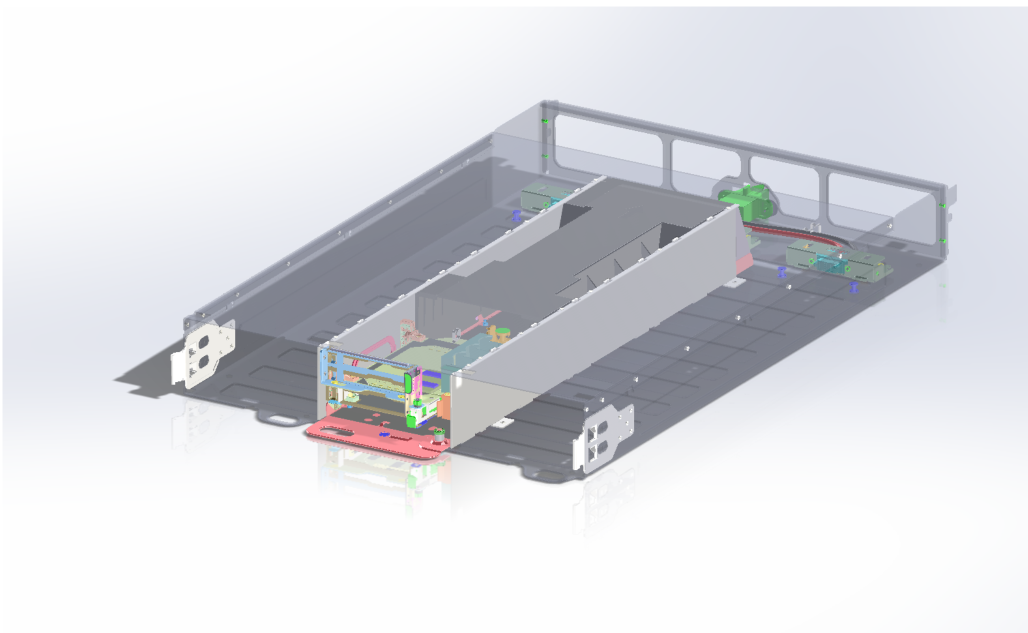


Figure 7-1: Cubby Chassis (Single Sled in Middle Bay)

7.2 Sled Chassis

A sheet metal and plastic sled serves as the mechanical interface between the Yosemite platform and the Cubby chassis. It also provides mechanical retention for the components inside the sled, such as the power cable assembly, fan, mezzanine card, side plane, and 1S-server cards. The combination of sled, side plane and other components assembled in the chassis is a Yosemite Platform sled.

A reference sled 3D model is attached to this specification for reference (Figure 7-2). Interfaces such as keyed slots (Figure 7-3), panel detail and panel connector location for power connector retention, and plunger location are critical for compatibility with cubby and should strictly adhere to the 3D database. See Figure 7-4. (Two cards are depopulated for clarity).

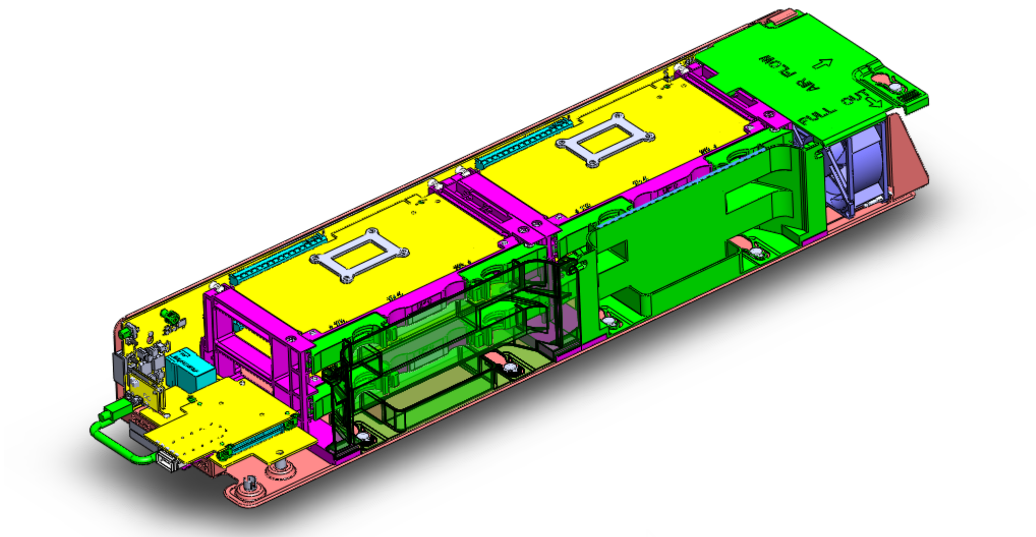


Figure 7-2: Yosemite Platform Chassis, Populated with 4x 1S Server

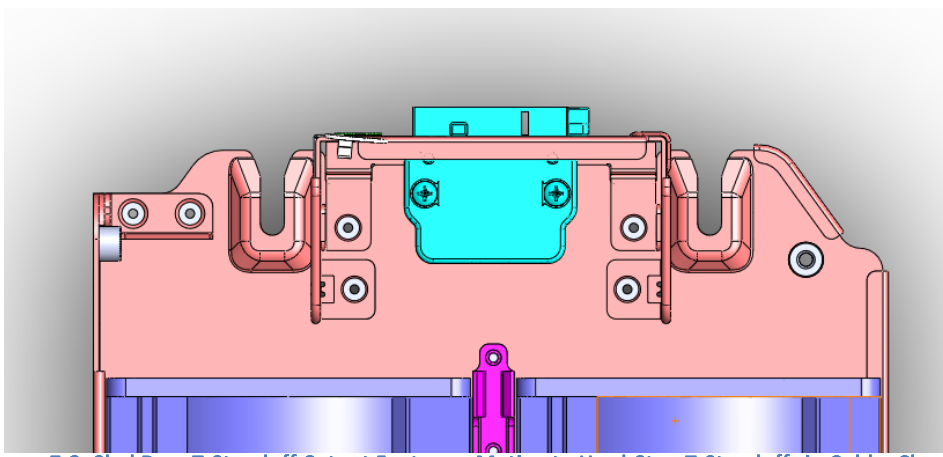


Figure 7-3: Sled Rear T-Standoff Cutout Feature – Mating to Hard-Stop T-Standoffs in Cubby Chassis

See Side Plane & Power Distribution section for breakout of side plane to panel-mount connector.

REVISIONS			
REV	DESCRIPTION	DATE	APPROVED
01	INITIAL RELEASE	03.05.15	RLCHU

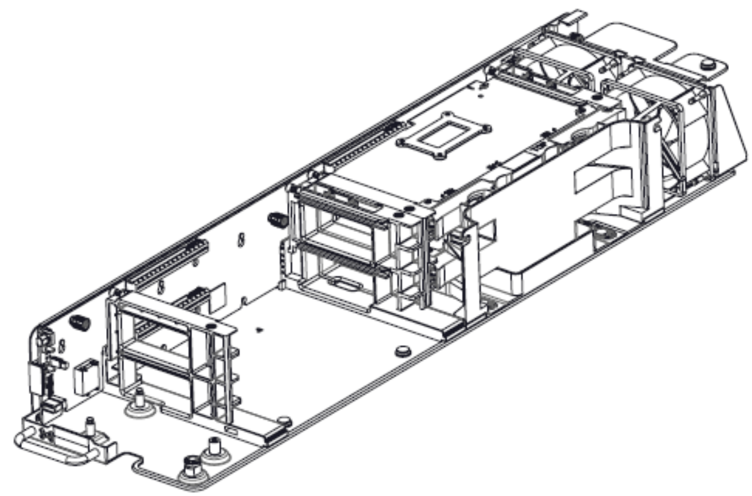
NOTES:
1 INTERPRET DIMENSIONS AND TOLERANCES PER ASME Y14.5M-2009
2 COMPONENT KEEPIN DIMENSIONS ARE DENOTED AS MAXIMUM
3 ALL DIMENSIONS ARE IN MM UNLESS OTHERWISE SPECIFIED
4 ALL SHOWN DIMENSIONS ARE CRITICAL TO FUNCTION AND WILL BE INCLUDED IN THE QUALIFICATION PROCESS.
5 ALL COMPONENTS SHALL BE UL 94V-0 RECOGNIZED AND ROHS COMPLIANT
6 CHASSIS MATERIAL: ELECTRO GALVANIZED STEEL
MATERIAL TYPE: SGCC
MATERIAL THICKNESS:

MATERIAL SPECIFICATION:

ASTM A-591/A 591M ELECTROLYTIC ZINC-COATED, COMMERCIAL STEEL, TYPE B, PHOSPHATE TREATED, CHEMICALLY TREATED DRY -(NO OIL), 42-60 ROCKWEL B OR JIS G3313 SECC, CHROMATE-FREE ANTI-FINGER PRINT TREATED, 42-60 ROCKWELL B, NO OIL.

CONDUCTIVITY-SURFACE RESISTANCE <1.0 m-OHM

7 MATERIAL MUST BE COMPLIANT WITH ENVIRONMENTAL SPECIFICATION
8 ALL CUT/SHARP EDGES WILL BE TREATED FOR OXIDATION PREVENTION. TREATMENT PROCESS WILL BE INCLUDED IN THE QUALIFICATION PROCESS
9 COMPONENTS SHALL BE CLEAN AND FREE FROM FOREGIN MATERIAL. OXIDATION, DIRT, OIL, GREASE AND OTHER CONTAMINANTS ARE NOT ALLOWED.
10 BURR SIDE AS INDICATED
11 COINED EDGES WILL HAVE A CHAMFER DEPTH OF AT LEAST 15% MATERIALTHICKNSS UNELSS OTHERWISE SPECIFIED. BREAKS IN COINING MUST BE APPROVED BY FACEBOOK ENGINEERING
12 EXTERNAL EDGES NOT CONFORMING TO NOTE 8 WILL BE DEBURRED
13 ALL EDGES NOT CONFORMING TO NOTES10 AND 11 SHALL BE CONDITIONED FOR HANDLING WITH A BURR NO GREATER THAN 10% OF MATERIAL THICKNESS AND SHALL CONFORM TO UL SAFETY SPECIFICATION 1439.
14 ALL DIMENSIONS NOT SHOWN MUST COMPLY WITH TOLERANCE TABLE. DIMENSIONS EXCEEDING TOLERANCE LIMIT MUST BE APPROVED BY FACEBOOK ENGINEERING
15 MUST BE COMPLIANT WITH THE FOLLOWING REFERENCE DOCUMENTS:
  i. ORV2 POWER DISTRIBUTION SPECIFICATION
  ii. CUBBY SELD REFERENCE SPECIFICATION
16 EXTERNAL COMPOMENTS MUST COMPLY WITH COMPONENT TABLE UNLES AUTHORIZED BY FACEBOOK ENGINEERING



14

FEATURE	TOL± mm/in
HOLE TO HOLE	0.13/.005
HOLE TO BEND	.25/.01
SHEAR EDGE TO BEND	0.25/.01
BEND TO BEND	0.40/.016
HOLD TO SHEAR EDGE	0.25/.01
CS DIAMETER	0.20/.008
CS FEATURE TO DATUM	0.10/.004
CS FEATURE TO CS FEATURE	0.15/.006
CS TO HOLE CONCENTRICITY	0.15/.006
ANGLES	0.5°

	NAME	DATE
DRAWN	RLCHU	03MAR15
ENG APPR.	RLCHU	03MAR15

PROPRIETARY AND CONFIDENTIAL

THE INFORMATION CONTAINED IN THIS DRAWING IS THE SOLE PROPERTY OF FACEBOOK. ANY REPRODUCTION IN PART OR AS A WHOLE WITHOUT THE WRITTEN PERMISSION OF FACEBOOK IS PROHIBITED.

facebook

TITLE: MULTI-NODE PLATFORM

SIZE B SPECIFICATION REV 1

SCALE: 1:10 WEIGHT: SHEET 1 OF 2

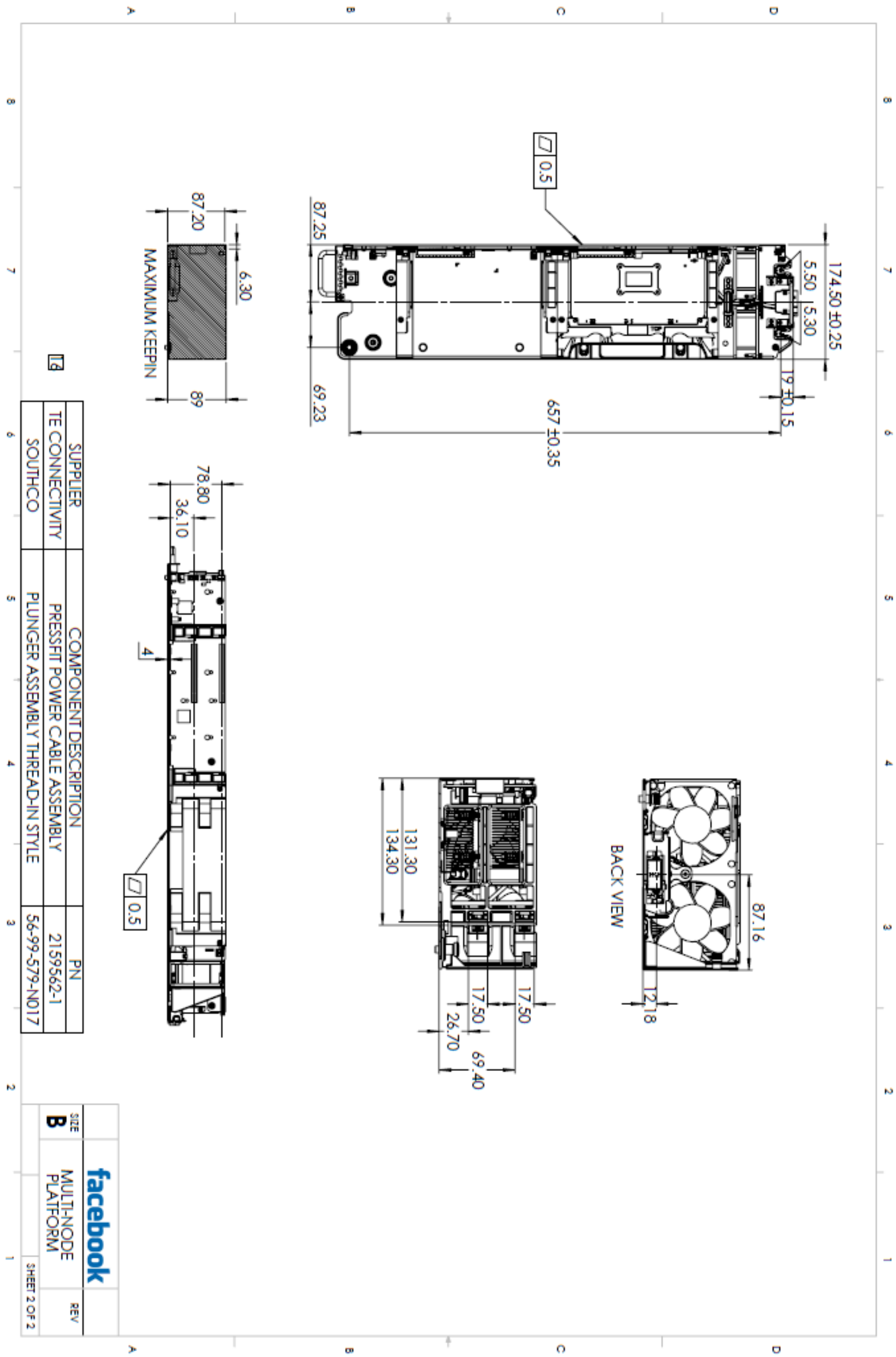


Figure 7-4: Yosemite Platform Specification Drawing

7.3 1S Server Card Retention/Extraction

See the General 1S Server Card specification drawing (Figure 7-8) for card dimensions, keepouts, goldfinger placement, and critical hole definitions. Some critical hole definitions correspond to the latching/retention mechanism between the 1S server card and the Yosemite Platform chassis. See Figure 7-8 for critical definitions. The retention mechanism compatible with the Yosemite Platform (Figure 7-5) is included in the 3D reference model.

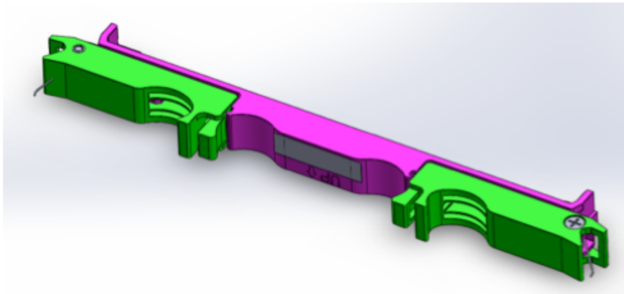


Figure 7-5: Card Latching Mechanism

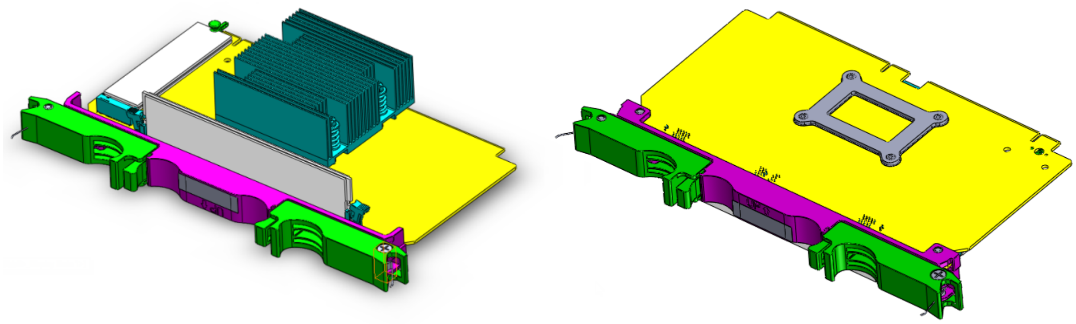


Figure 7-6: Card Retention Mechanism Mounted on 1S Server Card

The overall dimensions of the general card [mm] are 210 x L where L=110 or L=160. Major components (of significant height) will be placed on Side A. Small components will be placed on Side B. CPU placement must correspond to the drawing in Figure 7-8 if CPU placement is modified (only possible for L=160). The available area for guide features is a 5mm keepout on the sides of the card. This area will be silkscreened white on both sides. The Yosemite Platform chassis will support up to 4 cards via vertical PCIe to the side plane. Cards are supported by plastic rails that lock/unlock with a retention mechanism (Figure 7-7).

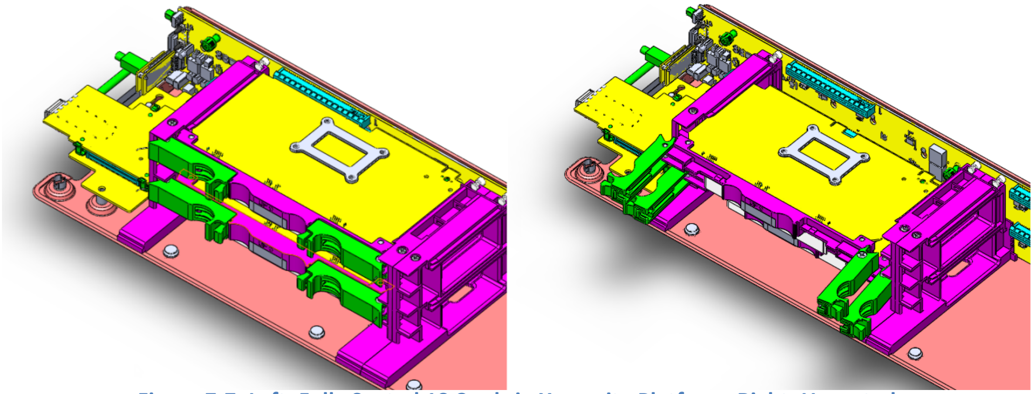


Figure 7-7: Left: Fully Seated 1S Cards in Yosemite Platform; Right: Unseated

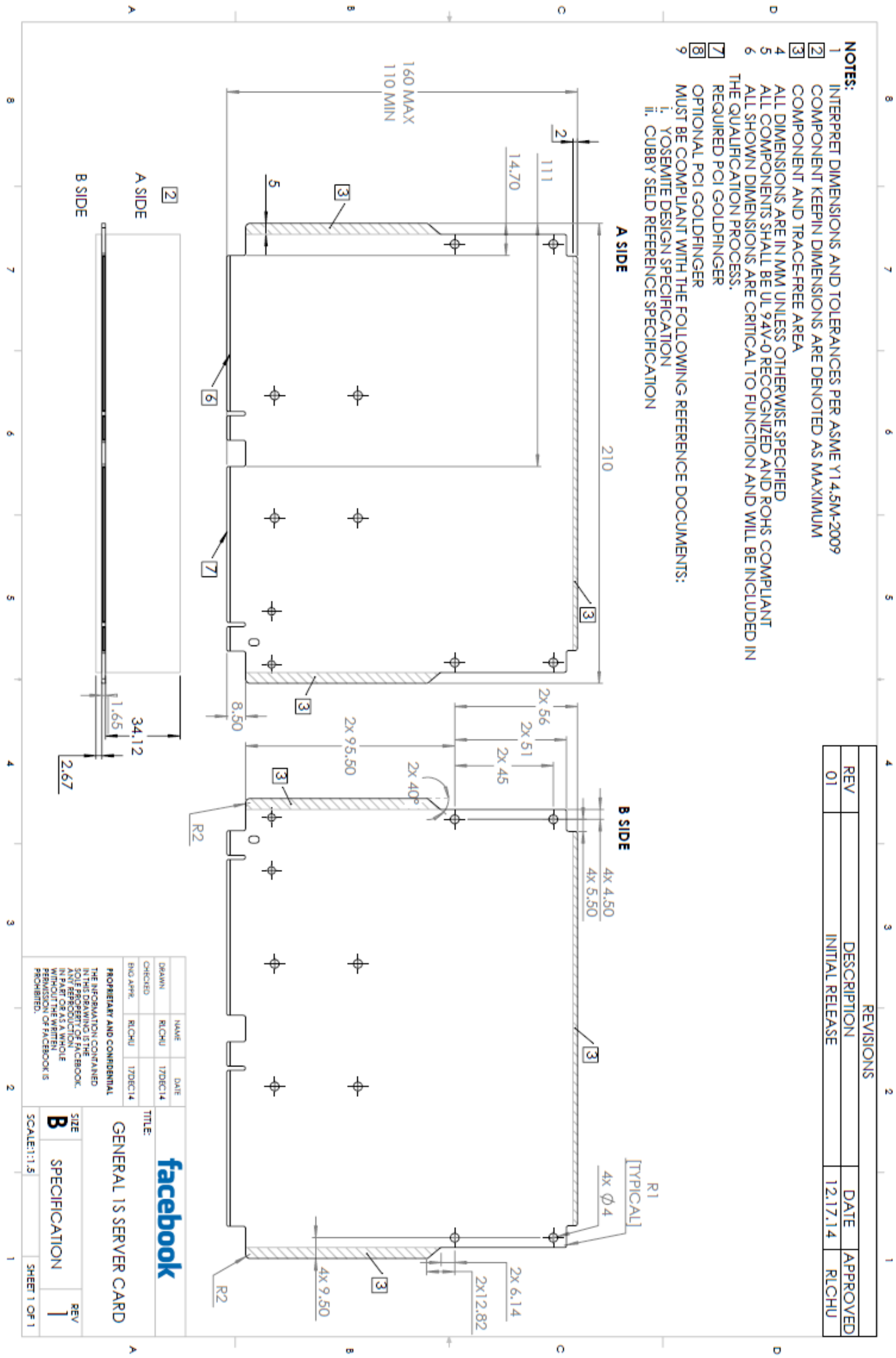


Figure 7-8: General 1S Server Card Specification Drawing

7.4 Silkscreen

Silkscreens 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.
- Keep-out area (see the General Card Specification drawing above)

7.5 Side Plane & Power Distribution

At the interface between The Cubby and the sled, a floating slide-to-lock panel-mounted connector will be affixed to the sled panel (Figure 7-9). The panel-mounted connector will be cabled to either a SQR (squeeze to release) connector (mating to a board mount vertical/right angle connector) or to connectors directly soldered to the side plane (Figure 7-10).

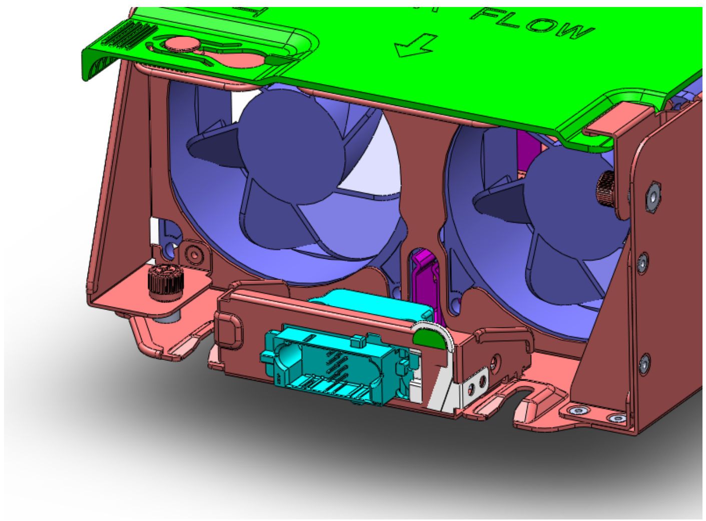


Figure 7-9: Panel mount power connector nested in sled panel

Power distributed from the side plane can terminate to a panel connector in two ways through a directly soldered press-fit connector. Both ends are depicted in Figure 7-10. Fans are suppressed for clarity.

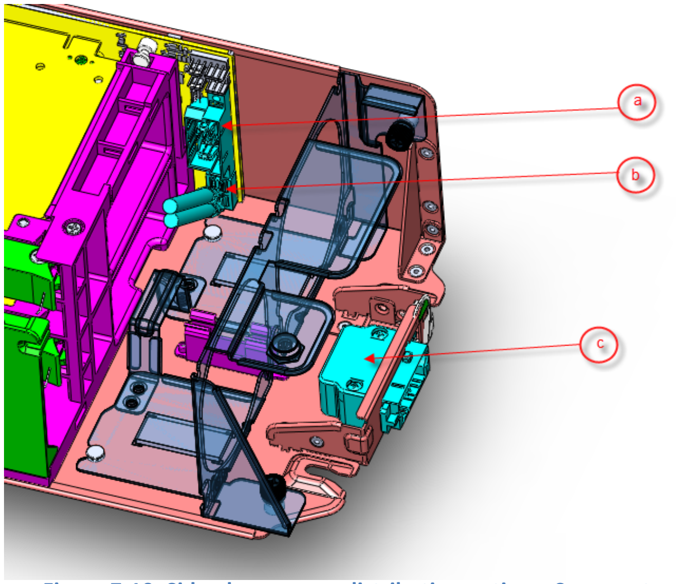


Figure 7-10: Side plane power distribution options: Squeeze to release (a) and pressfit (b) both terminating to a panel-mount power connector (c)

The family of connectors enabled for this application is FCI PowerBlade Plus and Tyco Multibeam XLE in the 1P 8S 1P configuration (Figure 7-11). The signal wires will not be used in Yosemite but spare cable assemblies (from connectors 2 -3) will be populated to maintain compatibility.

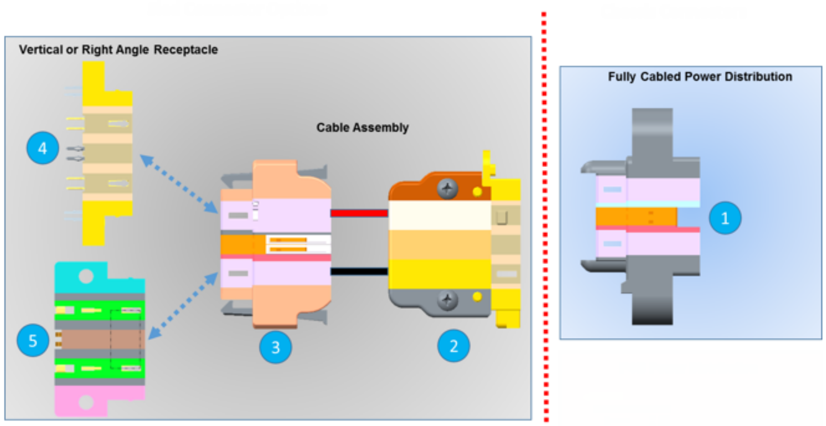


Figure 7-11: Power connector options

Connector 1 is the Cubby chassis connector. Connectors 2-3 are the cabled sled connectors. Connectors 4 or 5 will be mounted on the side-plane. The table below shows the part numbers. The table is subject to expansion as more connectors are qualified.

Table 5: Connector Part Numbers

Connector	Part Number
1	1-1892903-2
2	1-1892933-1
3	1-1892820-1
4	6450824-5
5	6450844-2
6	3-6450840-6

The side plane will support up to 4 1S Server cards of fixed width (210mm) and discrete length (110mm and 160mm) via generic vertical PCIe x16 connectors (Figure 7-12). A sample acceptable reference part number is TE PN 7-1734774-3.

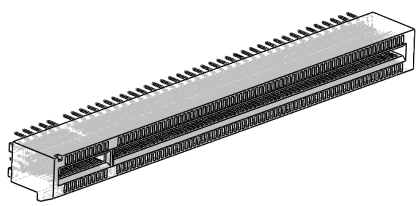


Figure 7-12: Generic vertical x16 PCIe connector

The side plane will support the OCP Mezzanine 2.0 form factor (PCB with keepouts and connectors attached). The I/O port(s) (at least 1 QSFP or QSFP+) will face the front of the sled (Figure 7-13).

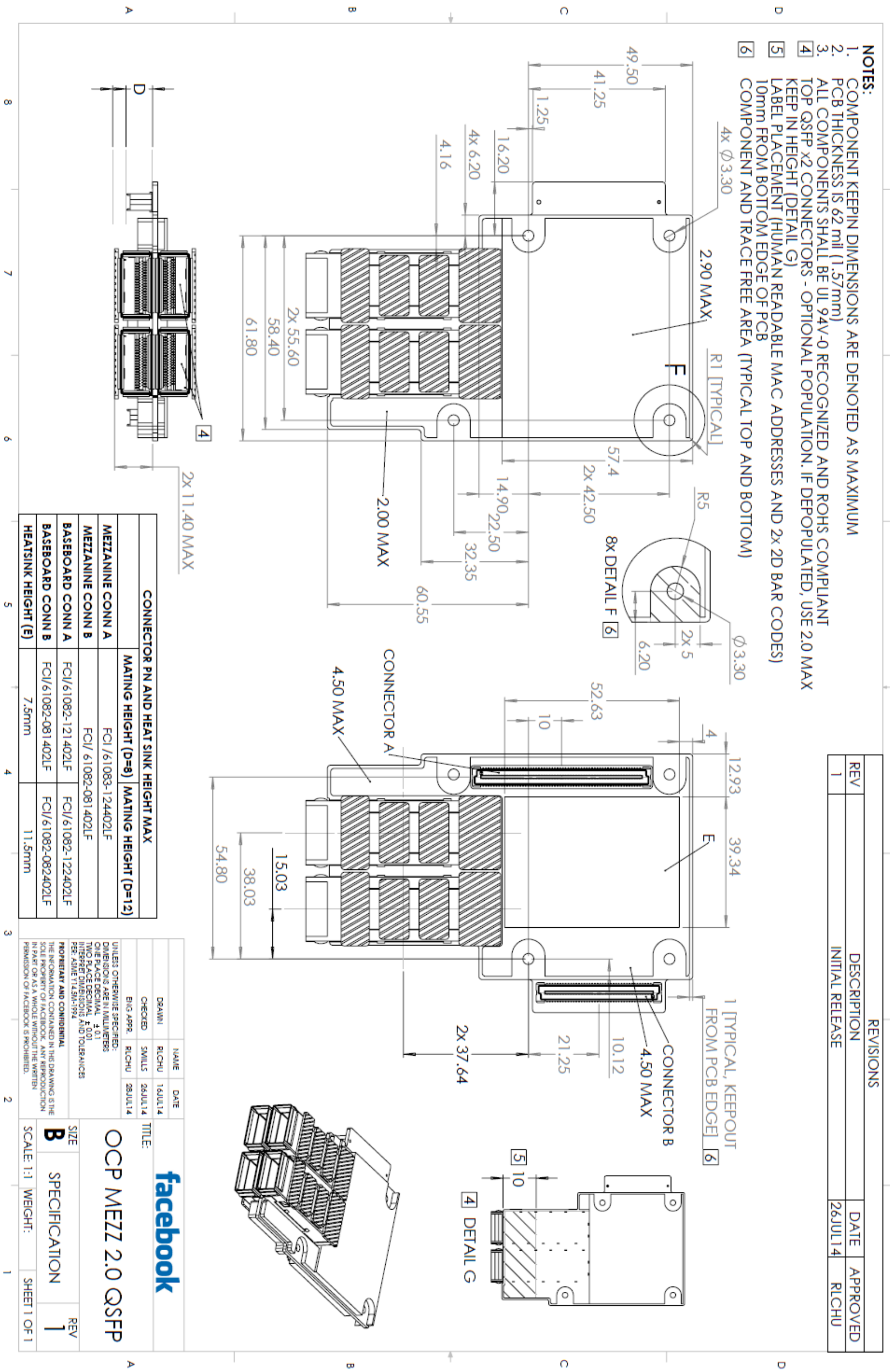


Figure 7-13: OCP mezzanine 2.0 QSFP specification drawing

8 Thermal

To meet thermal reliability requirements, the thermal and cooling solution should dissipate heat from the components when the system is operating at its maximum thermal power. Find the best thermal solution by setting a high power target for the initial design. This will enable you to avoid redesigning your cooling solution. The final thermal solution of system that you deliver should be the most optimized and energy efficient for data center environmental conditions with the lowest capital and operating costs. The thermal solution should not allow Yosemite Platform components to overheat. The CPU or memory should not throttle due to any thermal issue under the following environmental conditions.

- Inlet temperature lower than or equal to 35°C, and 0 inches H_2O datacenter pressure with all fans in each thermal zone running.
- Inlet temperature lower than or equal to 35°C, and 0.005 inches H_2O datacenter pressure with one failed fan (or one rotor) in each thermal zone.

8.1 Data Center Environmental Conditions

The thermal design for the Yosemite Platform needs to satisfy the data center operational conditions described below.

8.1.1 Location of Data Center/Altitude

Maximum altitude is 1,000m above sea level. Any variation of air properties or environmental difference due to the high altitude needs to be deliberated into the thermal design.

8.1.2 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 to the outside air temperature. Every component must be cooled and must maintain a temperature below its maximum specification temperature in the cold aisle.

8.1.3 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 0.005 inches H_2O with a single fan (or rotor) failure.

8.1.4 Relative Humidity

A data center usually maintains a relative humidity between 20% and 85%. When the thermal design can meet the requirement with a maximum relative humidity (85%), you may not need to consider the environmental condition changes due to high altitude.

8.2 Server Operational Conditions

8.2.1 System Loading

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

- System loading: idle to 100%
- Mezzanine card: 90W maximum

The Yosemite Platform can consume up to 500W per node on a DC bus bar with 14 configured nodes per power zone in an Open Rack V2 6.3KW power zone. The system design should support up to 500W per node for future configurations.

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 alternative way to provide cost benefits. At minimum, the air-duct design should be unified for all SKUs.

8.2.2 DDR DIMM DRAM Operation

The thermal design should meet the 1S server DIMM (or DRAM) max operating temperature (85°C with a single refresh rate). 1S server vendors should implement BIOS and memory subsystems to have an optimized refresh rate and to use optional DIMM Auto-Self-Refresh (ASR) based on the DIMM temperature. The implementation should follow all updated DDR3/DDR4 memory controller and DIMM vendor specifications.

8.2.3 Inlet Temperature

The inlet air temperature will vary. The cooling system in Yosemite Platform should be able to cover inlet temperatures including 20°C, 25°C, 30°C, and 35°C.

8.2.4 Pressurization

Except for the condition when one rotor in a server fan fails, the thermal solution should not consider extra airflow from data center cooling fans. If and only if one rotor 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.

8.2.5 Fan Redundancy

The server fans at N+1 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. A N+1 fan redundancy in the Yosemite Platform is preferred when the system is operating under normal conditions.

8.2.6 System Airflow or Volumetric Flow

The unit of airflow (or volumetric flow) used for this spec is cubic feet per minute (CFM). The maximum allowable airflow per watt in the system must be 0.13. The desired airflow per watt is 0.1 or lower in the system at the mean temperature (plus or minus the standard deviation).

8.2.7 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 11.7°C (21°F). The desired Delta T is 20°C (36°F) when the inlet air temperature to the system is lower than 30°C.

8.2.8 Thermal Margin

The thermal margin is the difference between the maximum theoretical safe temperature and the actual temperature. The board design operates at an inlet temperature of 35°C (95°F) outside of the system with a minimum 4% thermal margin for every component on the card. Otherwise, the thermal margin for every component in the system is at least 7% for temperatures up to 30°C.

8.3 Thermal Kit Requirements

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

8.3.1 Heat Sinks

Heat sinks must have a thermally optimized design at the lowest cost. There must be no more than three heat pipes in the heat sink. Installation must be simple and uncomplicated. Heat sinks must not block debug headers or connectors.

8.3.2 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 alternative way to provide cost benefits. The maximum fan thickness should be less than 38mm. Each rotor in the fan should have a maximum of five wires. Except for the condition when one fan (or one rotor) fails, the fan power consumption in system should not exceed 5% of total system power, excluding the fan power.

System fans should not have backrush currents in all conditions. System fans should have an inrush current of less than 1A on 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 (300W for Open Rack V1/V2) in all conditions.

8.3.3 Air Duct

The air duct design must be the most energy efficient possible. The air-duct design should be simple and easily serviceable. A unified air-duct design is preferred for all SKUs in Yosemite Platform. A highly green material or reusable material is also preferred.

8.3.4 Thermal Sensor

The maximum allowable tolerance of thermal sensors in the Yosemite Platform is ±1°C.

8.4 Fan Speed Control (FSC)

The fan speed control should control the system fan's RPM/PWM to maintain server components at their desired operational conditions. Non-linear algorithms, such as a PID algorithm, must be used for major high-power components. Linear algorithms can be used for the minor thermal components in the system.

8.4.1 Fan Speed Control Update

The FSC table should be human-readable and be easy-to-update to the BMC with an open-source IPMI. Please read the OCP Fan Speed Control Interface specification available on the Open Compute Platform website.

9 I/O System

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

9.1 PCI-E Slots

The Yosemite Platform does not have PCI-E slots for external I/O cards.

9.2 Network

9.2.1 Data Network

The Yosemite Platform uses an OCP 2.0 Hybrid Mezzanine card on the front panel as its primary data network interface. It could be a 4x10G KR-retimer card, or a multi-host 40G/40G network interface card, or a 50G KR-Aggregation card. Please refer to Section 5 for more details.

9.2.2 Management Network

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

9.3 1S Server slots assignment

The Yosemite Platform defines 1S Server slot ID assignment and order in the table below, which is a side view of a Yosemite sled.

Table 6: Slot ID Assignment and Order

Front Side	Slot 1 (top)	Slot 2 (top)	Rear Side
Cold Aisle	Slot 3 (Bottom)	Slot 4 (Bottom)	Hot Aisle

9.4 Front Panel

On the front panel of a Yosemite sled, there is a power button, a reset button, an OCP debug card and a USB port attached to the current selected 1S Server,. The selected server is determined by the position indicated on the selector knob. There are four blue LEDs placed on the front panel in the same order as 1S Server slots to indicate server status.

9.4.1 Selector Knob

A user can turn the selector knob to select a 1S Server and a BMC. When a 1S Server is selected, it owns the power button, reset button, OCP debug card and USB port on the front panel. The LED associated with the active 1S Server blinks as visual feedback to the user. When a BMC is selected, all four LEDs blink as visual feedback to the user. The BMC owns the OCP debug card, but not the power button, reset button or USB port.

9.4.2 Power Button and Reset Button

A red power button and a black reset button are on the front panel. They belong to the currently selected 1S server.

When the power button is pressed for less than four seconds and then released, the currently selected 1S server receives a Power Management event. This event will power on the 1S Server (if it was off). However, if the current selected 1S Server is already on but a user presses the power button for more than four seconds, the 1S Server will perform a hard power off.

If the reset button switch is pressed for any duration of time, and the currently selected 1S server is on, it shall perform a hard reset.

A label on the side plane's silkscreen will indicate the functionality of each button.

9.4.3 LED

There are four dual colored Blue/Yellow LEDs on the front panel. These LEDs are used to indicate power and to identify which 1S server is currently selected. These LEDs are placed in a grid (2 rows of 2 LEDs each) and represent each 1S server's power status. The placement and silkscreen label must match the 1S server slot ID assignment.

9.4.4 USB Connector

The Yosemite Platform has one USB 2.0 port located at the front panel of the side plane. It belongs to the currently selected 1S Server.

The BIOS should support the following devices attached to the USB port:

- USB Keyboard and mouse
- USB flash drive (bootable)
- USB hard drive (bootable)
- USB optical drive (bootable)

On the side plane a USB mux is used to connect all four 1S servers to the USB port. The BMC will control the mux based on the position of the selector knob. In addition, a BMC's virtual hub port is connected to the 1S Server through a hub. This enables any 1S Server to update the BMC firmware through this path.

9.4.5 OCP Debug Header

A standard OCP debug header is at the front panel of the side plane. Through this debug header, an OCP debug card can provide serial port access to 1S Servers and the BMC, as well as to the POST code display. The Reset button on the OCP debug card behaves exactly like the Reset button on the front panel.

The debug header is a 14-pin, shrouded, vertical, 2mm pitch connector. Figure 9.1 is an illustration of the header. The debug card should have a key to match with the notch to avoid pin shift when plugging in.

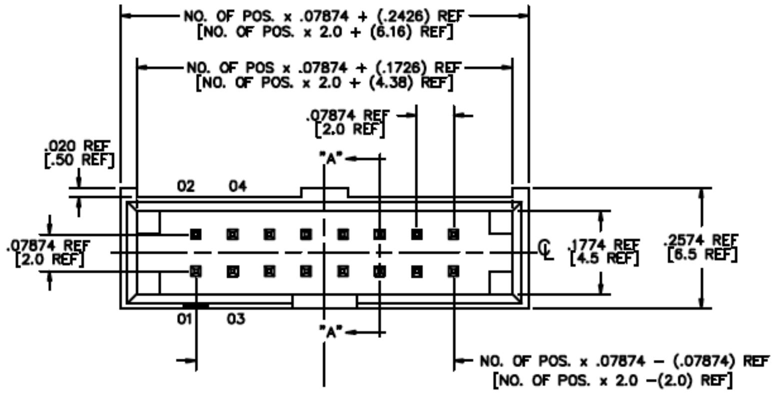


Figure 9-1: Debug Header

Table 7: Debug Header Pin Definitions

Pin (CKT)	Function
1	Low HEX Character [0] Least Significant Bit
2	Low HEX Character [1]
3	Low HEX Character [2]
4	Low HEX Character [3] Most Significant Bit
5	High HEX Character [0] Least Significant Bit
6	High HEX Character [1]
7	High HEX Character [2]
8	High HEX Character [3] Most Significant Bit
9	Serial Transmit (Motherboard Transmit)
10	Serial Receive (Motherboard Receive)
11	System Reset
12	UART Channel Selection
13	GND
14	VCC (+5VDC)

9.4.6 POST Codes

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

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

Table 8: DIMM Error Code Table

	Code	Result
CPU0 (Channel 0 & 1)	A0	Channel 0 DIMM 0 (furthest) Failure
	A1	Channel 0 DIMM 1 Failure
	B0	Channel 1 DIMM 0 Failure
	B1	Channel 1 DIMM 1 (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.

9.4.7 Serial Console

The output stage of the system's serial console shall be contained on the debug card. The TX and RX signals from the system UART shall be brought to the debug header at the chip logic levels (+3.3V). The debug card will contain a mini-USB connector with the pin definition shown in the table below. A separate convertor is needed to provide a RS-232 transceiver and a DB9 connector.

Table 9: Debug Card Mini-USB UART Pin Definitions

Pin	Function
1	VCC (+5VDC)
2	Serial Transmit (motherboard transmit)
3	Serial Receive (motherboard receive)
4	NC
5	GND

By default, the Yosemite Platform performs console redirection through the SOL. When the debug card is plugged in, debug card pin 12 shall be used to select console redirection between the SOL and the local serial port on the card, as described above.

9.5 Fan Connector

The Yosemite Platform motherboard has an 8-pin fan connector that supports two fans. Every fan has their 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 4: Fan Connector Pin Definition

Pin	Description
1	Second fan's PWN 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

9.6 Power

9.6.1 Input Voltage Level

The nominal input voltage delivered by the power supply is 12.5 VDC. The voltage has a 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.

The total power of the Yosemite Platform shall be 500W or lower.

9.6.2 Capacitive Load

To ensure compatibility with the system power supply, the side plane may not have a capacitive load greater than 4000µF. The capacitive load of the platform should not exceed the maximum value of 4000µF under any operating condition as defined in Section 10.

9.7 Hot Swap Controller Circuit

In order to have better control of the 12.5V DC power input to each platform, a HSC (ADI ADM1278) is used on the side plane. A HSC circuit provides the following functions:

- Inrush current control when the Yosemite Platform is inserted and powered up.
- Current limiting protection for over current and short circuit. The over current trip point should be able to be set to 45A.
- Safe operating area protection when the MOSFET turns on and off.
- PMBus interface to enable the following BMC actions
    - Report server input power and log an event if it triggers the upper critical threshold.

  - Report input voltage (up to 1 decimal point) and log an event if it triggers either the lower or the upper critical threshold.
  - Log a status event based on the hot swap controller's status register.
- Provide a fast overcurrent sense alert with a resistor option to disable.

The voltage drop on the HSC current-sense resistor should be less than or equal to 25mV at full loading. The hot-swap controller should have the SMBus address set to 0x20 (7-bit format).

The power reporting of the hot-swap controller must be better than 2%, from 50W to full loading at room temperature.

9.8 1S Server Power Management

The Yosemite Platform supplies single 12V power to all 1S Server slots. There is a power switch for each 1S Server slot under the BMC's control. These 12V power switches should be on by default unless the BMC turns them off on purpose. It is a useful feature to implement AC on, off, or cycling through the BMC.

The BMC can sample total platform power consumption from the hot-swap controller via an SMBus. As specified in the OCP 1S Server specification, every 1S Server shall implement a power sensor to monitor total 1S Server power consumption. These power sensors are accessible to the BMC via the Bridge IC on the 1S Server card. The BMC shall implement a sophisticated power management algorithm based on total platform power consumption and the power consumption of individual 1S Servers.

A fast throttle feature is implemented on the platform. It enables you to throttle an individual 1S Server or all 1S Servers down to lowest power state in the shortest possible time period. The hot-swap controller could trigger this signal when a platform-level over-current condition happens, which will throttle down all the 1S Servers. The BMC can also throttle particular 1S Servers as needed.

9.9 System VRM Efficiency

High efficiency VRMs shall be used for the Yosemite Platform with 91% efficiency over the 30% to 90% load range.

9.10 Power Policy

The power policy of 1S server cards on Yosemite Platform can be set by the BMC to Always On or Last Power State. When the power policy is Always On, the 1S Servers will be powered on automatically regardless of their last power state. When the power policy is Last Power State, the 1S Servers will restore the last power state before AC cycling.

10 Environmental Requirements and Other Regulations

10.1 Environmental Requirements

The motherboard shall meet the following environmental requirements:

- Gaseous contamination: Severity Level G1 per ANSI/ISA 71.04-1985
- Ambient operating temperature range: -5°C to +45°C
- Operating and storage relative humidity: 10% to 90% (non-condensing)
- Storage temperature range: -40°C to +70°C
- Transportation temperature range: -55°C to +85°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: -5°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: -55°C to +85°C (short-term storage)
- Operating altitude with no de-ratings: 1,000m (3,300 feet)

10.2 Vibration and Shock

The motherboard 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 5: Vibration and Shock Requirements

	Operating	Non-Operating
Vibration	0.5g acceleration, 1.5mm amplitude, 5 to 500 Hz, 10 sweeps at 1 octave / minute per each of the three axes (one sweep is 5 to 500 to 5 Hz)	1g acceleration, 3mm amplitude, 5 to 500 Hz, 10 sweeps at 1 octave / minute per each of the three axes (one sweep is 5 to 500 to 5 Hz)
Shock	6g, half-sine 11mS, 5 shocks per each of the three axes	12g, half-sine 11mS, 10 shocks per each of the three axes

10.3 Regulations

The vendor needs to provide certification body reports of the Yosemite Platform motherboard and tray at the component level.

11 Prescribed Materials

11.1 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 6)
- Trimmers and/or potentiometers
- Dip switches

11.2 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).
- Ceramic material for SMT capacitors must be X7R or better (COG or NP0 type are used in critical portions of the design). Only SMT inductors may be used. The use of through-hole inductors is disallowed.

11.3 Component De-rating

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

12 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, which may cause them to disrupt the functionality or the airflow path of the motherboard.

Table 6: 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

13 Revision History

Author	Description	Revision	Date
Yan Zhao	▪ Initial draft.	0.1	2/11/2014
Yan Zhao Renee Chu Jacob Na Sai Dasari	▪ Incorporated review comments.	0.2	02/06/2015
Yan Zhao	▪ Updated to OCP format.	0.3	5/12/2015
Yan Zhao	▪ Incorporated review comments.	0.4	1/12/2016