OPEN
Compute Project

Big Basin-JBOG Specifications

Rev 0.05

Author:
Kevin LEE, Technical Program Manager, Facebook
Jia NING, Hardware Engineer, Facebook
K.Y. THOON, Hardware Engineer, Facebook

1 Revision History

Table 1-1

Date	Name	Description
0.01	5/11/2016	Initial release
0.02	7/31/2016	Update Project name to Big Basin
0.03	11/21/2016	Update for review
0.04	12/21/2016	Minor corrections.
0.05	01/19/2017	Change factory default power policy to always-on

2 Scope

This document describes technical specifications for Facebook's Big Basin-JBOG for use in Open Rack V2[1].

3 Contents

1 http://files.opencompute.org/oc/public.php?service=files&t=348f3df2cc4ce573397fcc4424f68ca6&download

4 Overview

4.1 Introduction

JBOG (Just-a-bunch-of-GPUs) supports 8x GPUs with flexible interconnect internally, and externally to host. There is no server head node in JBOG design. JBOG needs to work with a server head node such as Leopard.
JBOG shall be compatible with Open Rack V2 (ORv2) and occupies 3x Open U(OU) of space.
JBOG supports 8x NVIDIA Tesla P100 GPUs in SXM2 form factor. The baseboard design is to be able to support next gen SXM2 GPU as well.

Below is a list of important items that form the JBOG system.

Table 4-1

Item	Description/Major components
Baseboard	8 NVIDIA Tesla P100 SXM2 GPUs 4 PCIE switches 8 x16 PCIE Gen3 slots
Middle Plane Board	BMC and CPLD Power connectors from Rack Bus bar Power/Signal connectors for Baseboard Power/Signal connectors for IO board
IO board	BMC network port Debug support User LEDs
Host Side Bridge Card	4 mini-SAS HD connectors x16 PCIE Gen3 retimer config jumpers to configure retimer in different modes
JBOG Side Bridge Card	4 mini-SAS HD connector
Fans	4x System fans for cooling
Power Supply	Open Rack V2
Chassis	3 OU BOX for Open Rack v2

4.2 PCIe Block Diagram

There are 2x topologies for the JBOG system PCIe connection shown below, using a Leopard head node (also known as: Facebook server Intel Next Generation Xeon motherboard v3.1[2]).

User shall be able to configure the system in these topologies with only jumper change to retimer cards.

[2] http://files.opencompute.org/oc/public.php?service=files&t=9dec25e49d573cd781bbee4a63e1cf79&download

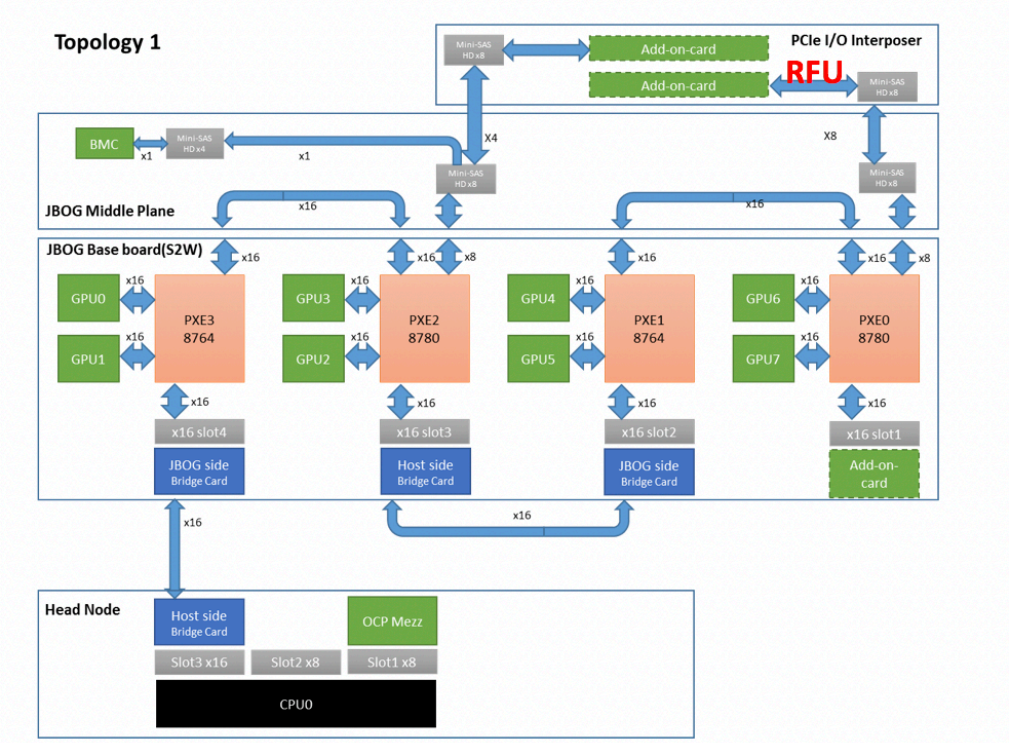


Figure 4-1

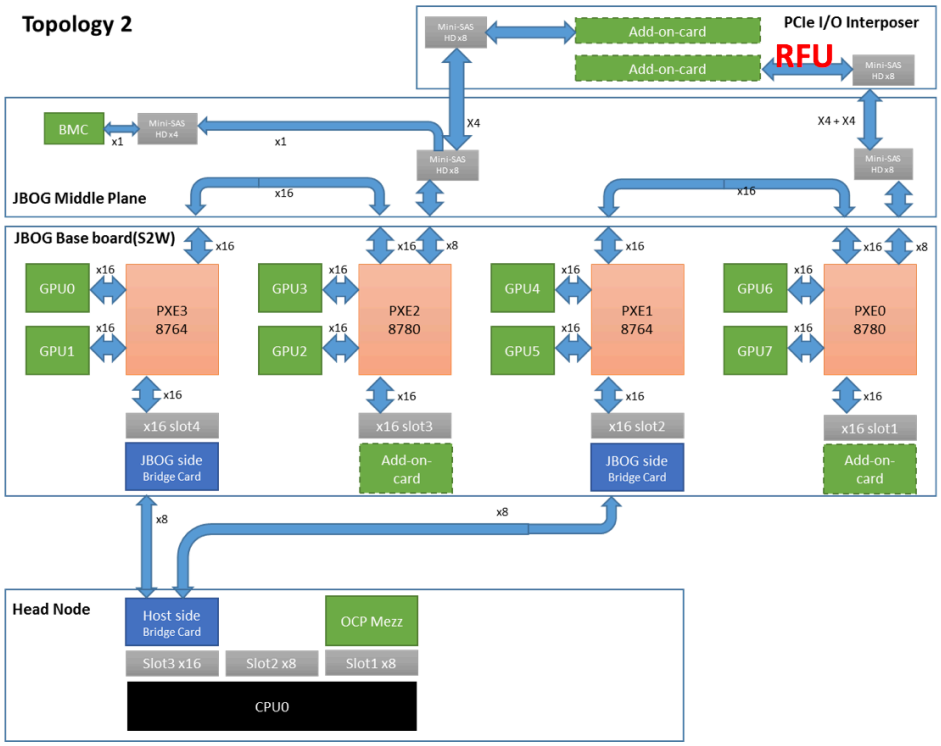


Figure 4-2

4.3 NVLINK topology

For JBOG with 8x GPUs in SXM2 form factor, it shall support NVLINK interconnection shown below. The current generation Tesla P100 GPU have SERDES running at 20 Gbps. Next Gen GPU could have SERDES running up to 25.78125 Gbps.

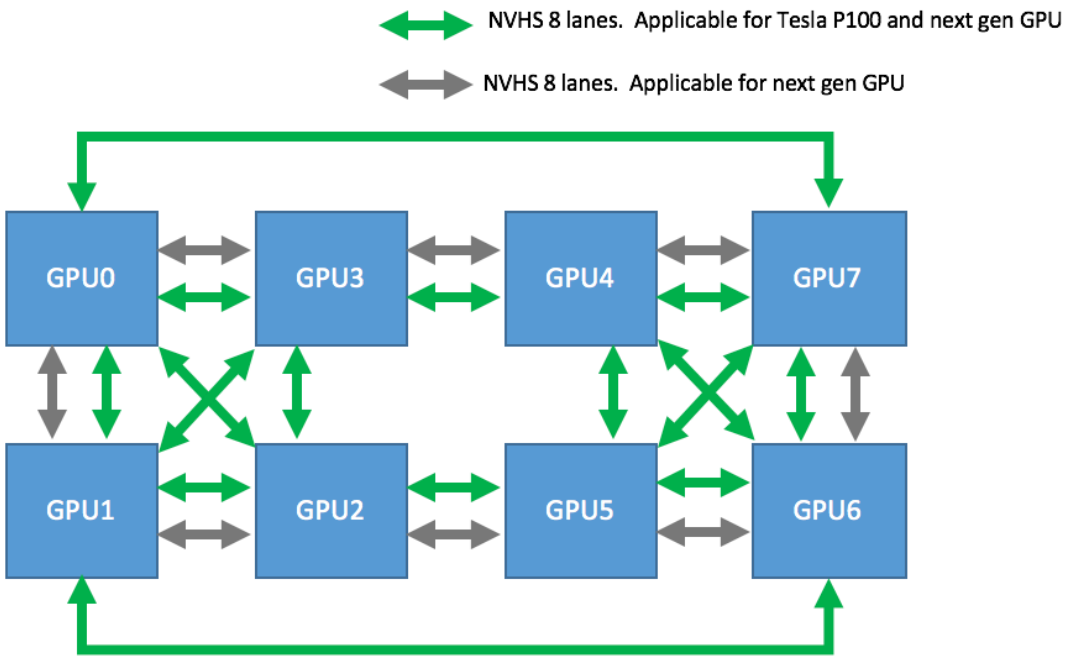


Figure 4-3

4.4 GPU Baseboard Design

JBOG for SXM2 GPUs leverages from the GPU Baseboard design of NVIDIA DGX-1 system. (http://www.nvidia.com/object/deeplearningsystem.html?mboxSession=8742eb98c6df4426b172ß881c7c6f7c3a)

JBOG for PCIe form factor GPU card will need a new PCIe GPU baseboard design. The detailed design is not in the scope of this specification. The vendor shall provision this use case when designing the rest of the system.

5 JBOG Interface to Host

This chapter defines the interface between JBOG and Host, including retimer cards, cables, and Host to JBOG side band connection for JBOG BMC.

5.1 PCIe Cable Interface

A MiniSAS-HD cable is used to provide connection from host node to JBOG, from JBOG to another JBOG, and within JBOG itself. Each cable has an x8 PCIe lane with side band signals defined in this section. Cables shall be plugged into PCIe bridge cards, which are plugged into the host node and JBOG enclosure. More details of PCIe bridge cards are described in section 5.2.

5.1.1 Mini-SAS HD Connector (SFF-8644)

The pinout of host connectors has been modified from a standard Mini-SAS HD connection to allow for a sideband signaling. An external Mini-SAS HD connector is SFF-8644 standard form factor.

Please note the recent swap of PCIe Lane 0 and Lane 1 to be compatible with PCIe external cabling spec (PCI_Express_External_Cabling_Rev3 0_v0 6 20151006)

Table 5-1

PCI Ext cable Dual layou	HOST			END POINT		PCI Ext cable Dual layout
	Name	Pin		Pin	Name	
	GND	D9	→	B9	GND	
	PETn2	D8	→	B8	PERn2	
	PETp2	D7	→	B7	PERp2	
	GND	D6	→	B6	GND	
	PETn1	D5	→	B5	PERn1	
	PETp1	D4	→	B4	PERp1	
	GND	D3	→	B3	GND	
MGTPWR	USB_DP	D2	→	D2	USB_DP	MGTPWR
PWR	USB_DN	D1	→	D1	USB_DN	PWR
	SCL	C1	→	C1	SCL	
	SDA	C2	→	C2	SDA	
	GND	C3	→	A3	GND	
	PETp0	C4	→	A4	PERp0	
	PETn0	C5	→	A5	PERn0	
	GND	C6	→	A6	GND	
	PETp3	C7	→	A7	PERp3	
	PETn3	C8	→	A8	PERn3	
	GND	C9	→	A9	GND	
	GND	B9	→	D9	GND	
	PERn2	B8	←	D8	PETn2	
	PERp2	B7	←	D7	PETp2	
	GND	B6	→	D6	GND	
	PERn1	B5	←	D5	PETn1	
	PERp1	B4	←	D4	PETp1	
	GND	B3	→	D3	GND	
	CBLPRSNT#	B2	→	B2	CBLPRSNT#	
PWR	PCIE_RESET_N	B1	→	B1	PCIE_RESET_N	PWR
	RFU				RFU	
Caddr	PCIE_CLOCK_DP	A1	→	A1	PCIE_CLOCK_DP	Caddr
CINT#	PCIE_CLOCK_DN	A2	→	A2	PCIE_CLOCK_DN	CINT#
	GND	A3	→	C3	GND	
	PERp0	A4	←	C4	PETp0	
	PERn0	A5	←	C5	PETn0	
	GND	A6	→	C6	GND	
	PERp3	A7	←	C7	PETp3	
	PERn3	A8	←	C8	PETn3	
	GND	A9	→	C9	GND	
		Shell	→	Shell		

5.1.2 Cable Selection and Wiring

The vendor shall only make use of the cables listed in the table below for different topology connections described in Section 4.2.

Table 5-2

Supplier	Length	Type	Part Number
Amphenol	520mm	X8	NEETCT-F402
Amphenol	350mm	X8	NEETCT-F403
JPC	520mm	X8	P4103B250550-1
JPC	350mm	X8	P4103B250400-1

See Figure 5-1 and Figure 5-2 how the cables are used for Topology 1 and 2 described in Section 4.2 for 2 JBOG systems in the Rack.

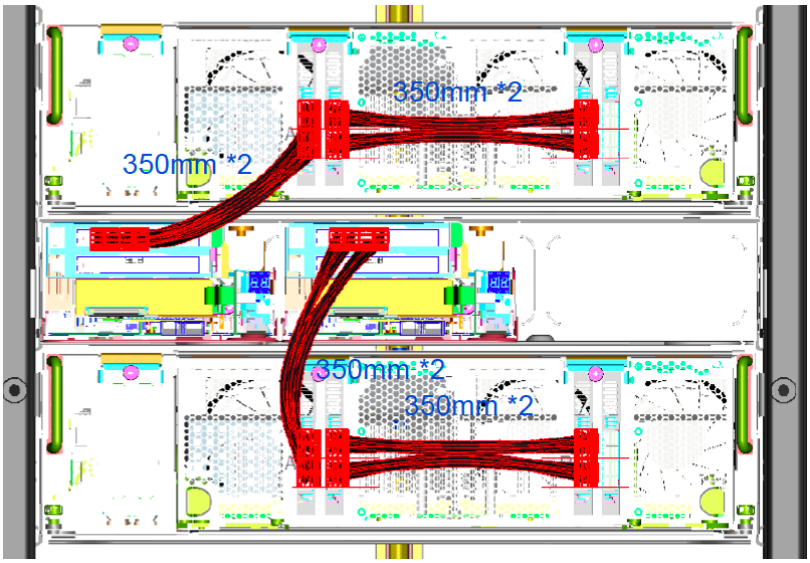


Figure 5-1

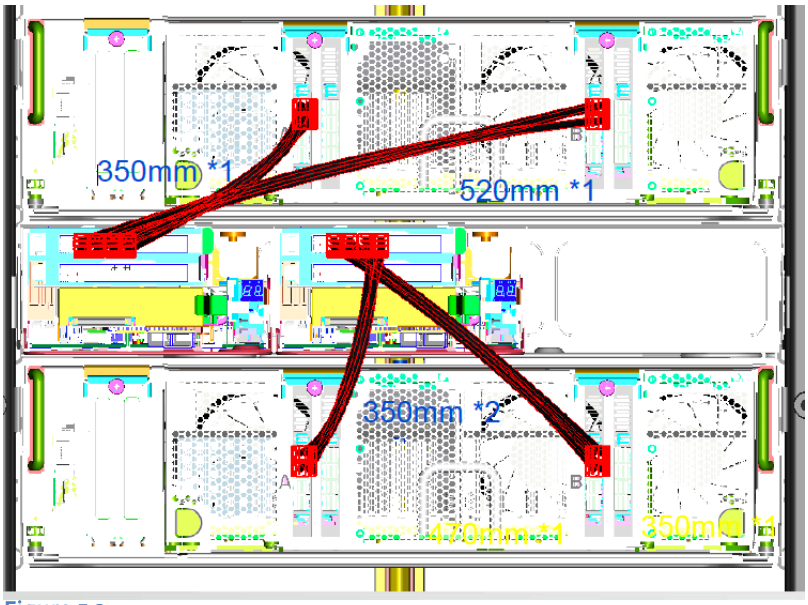


Figure 5-2

5.2 PCIe Bridge Cards

PCIe bridge cards are used in both the host node and JBOG side as an adaptor between the MiniSAS HD cable, and the standard PCIe slot in the host node and JBOG. There are two types of PCIe bridge cards:

- Host side PCIe bridge card with retimer
- JBOG side PCIe bridge card without retimer

Each PCIe bridge card follows low profile, half-length PCIe form factor
This section defines the connection block diagram and requirements to PCIe bridge cards.

5.2.1 Golden Finger Definition

The PCIe bridge card has a standard x16 PCIe golden finger design with latch. The vendor shall make use of Table 5-2 and Table 5-4 for modification of the golden fingers for JBOG's use.

Table 5-3

	JBOG_Pin_Definition_Table_0v3_20160510	
Pin #	Side B Golden Finger Name	Side A Golden Finger Name
1	+12v	PRSNT#1
2	+12v	+12v
3	+12v	+12v
4	GND	GND
5	SMCLK_BMC	NC
6	SMDAT_BMC	NC
7	GND	SMBCLK_PCH
8	+3.3v	SMBDAT_PCH
9	NC	+3.3v
10	P3V3_AUX	+3.3v
11	WAKE#	PERST#
	Mechanical Key	
12	SMB_ALERT_N	GND
13	GND	REFCLK+
14	PETp(0)	REFCLK-
15	PETn(0)	GND
16	GND	PERp(0)
17	PRSNT#2-1	PERn(0)
18	GND	GND
19	PETp(1)	RSVD
20	PETn(1)	GND
21	GND	PERp(1)
22	GND	PERn(1)
23	PETp(2)	GND
24	PETn(2)	GND
25	GND	PERp(2)
26	GND	PERn(2)
27	PETp(3)	GND
28	PETn(3)	GND
29	GND	PERp(3)
RFU 30	PWR_BRK#	PERn(3)
31	PRSNT#2-2	GND
32	GND	USB2.0+P2
33	PETp(4)	USB2.0+N2
34	PETn(4)	GND
35	GND	PERp(4)
36	GND	PERn(4)
37	PETp(5)	GND
38	PETn(5)	GND
39	GND	PERp(5)
40	GND	PERn(5)
41	PETp(6)	GND
42	PETn(6)	GND
43	GND	PERp(6)
44	GND	PERn(6)
45	PETp(7)	GND
46	PETn(7)	GND
47	GND	PERp(7)
48	PRSNT#2-3	PERn(7)
49	GND	GND
50	PETp(8)	RSVD/JBOG_PERST#
51	PETn(8)	GND
52	GND	PERp(8)
53	GND	PERn(8)
54	PETp(9)	GND
55	PETn(9)	GND
56	GND	PERp(9)
57	GND	PERn(9)
58	PETp(10)	GND
59	PETn(10)	GND
60	GND	PERp(10)
61	GND	PERn(10)
62	PETp(11)	GND
63	PETn(11)	GND
64	GND	PERp(11)
65	GND	PERn(11)
66	PETp(12)	GND
67	PETn(12)	GND
68	GND	PERp(12)
69	GND	PERn(12)
70	PETp(13)	GND
71	PETn(13)	GND
72	GND	PERp(13)
73	GND	PERn(13)
74	PETp(14)	GND
75	PETn(14)	GND
76	GND	PERp(14)
77	GND	PERn(14)
78	PETp(15)	GND
79	PETn(15)	GND
80	GND	PERp(15)
81	PRSNT#2-4	PERn(15)
82	RSVD	GND

Table 5-4

JBOG_Pin_Definition_Table_0v3_20160510						
	Host Bridge Card	Dir.	Description	JBOG Bridge Card	Dir.	Description
A7 A8	SMBCLK_PCH SMBDATA_PCH	N/A	SMBus for Host main processor; typical usage is to allow system FW to access bridge card FRU, and other GPIO without dependency of the status of Management Controller.	RSVD	N/A	N/A
B30	PWR_BRK# RFU	In	3.3V level GPIO from Head node to JBOG for power throttling	PWR_BRK#	Out	3.3V level GPIO from host to JBOG for power throttling
A32 A33	USB2.0 P/N	N/A	USB2.0; Head node SOC or southbridge/PCH as host device	USB2.0 P/N	N/A	USB2.0; JBOG BMC as device
A50	RSVD	N/A	N/A	JBOG_PERST#	Out	GPIO to pass host reset from MiniSAS HD cable to JBOG

5.2.2 PCIe Connection

The vendor shall follow the diagram below for PCIe connection for two types of PCIe bridge cards.
A PCIe bridge card with x16 retimer has bifurcation set by jumper. The status of the jumper shall be mapped to an LED visible to the user when the PCIe bridge card is installed in the system. The vendor shall implement the LED color shown in the diagrams below.

PCIe Bridge Cards - PCIe connection diagram	
Host side – with Retimer	X
JBOG side – without Retimer	

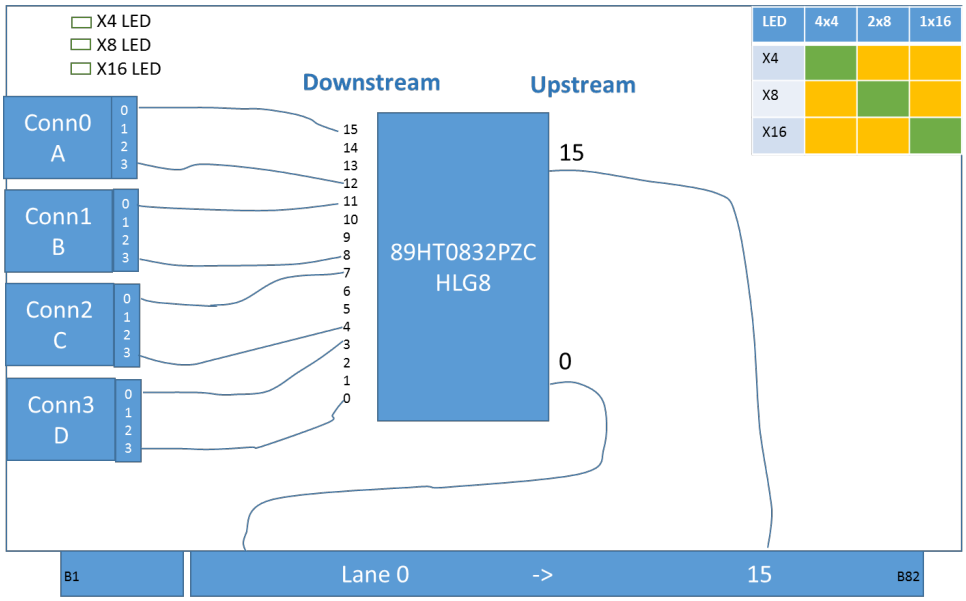


Figure 5-3

PCIe Bridge Cards - PCIe connection diagram	
Host side – with Retimer	
JBOG Side – without Retimer	X

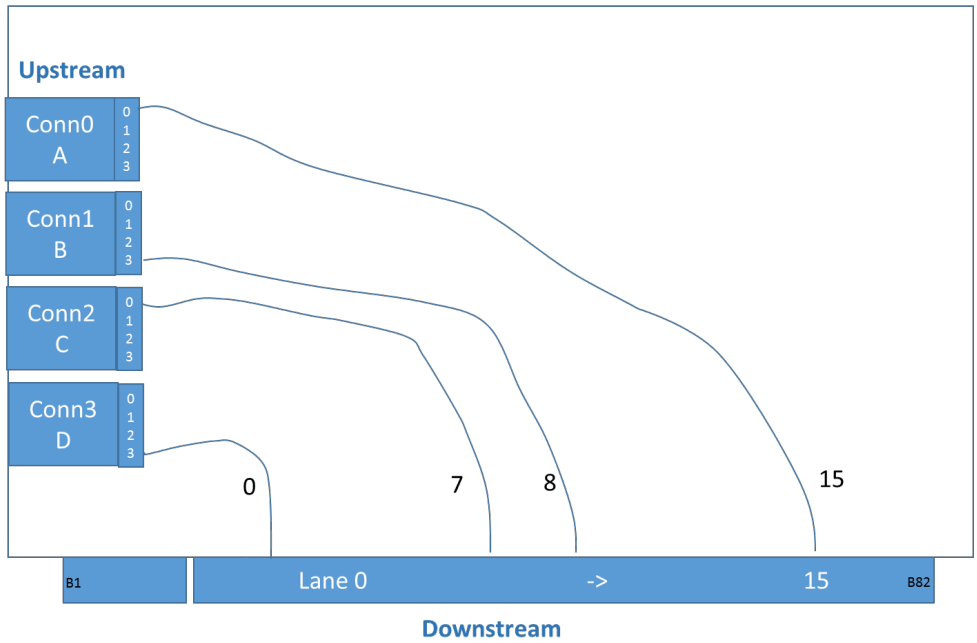


Figure 5-4

5.2.3 Retimer Configuration

The retimer shall have 3x jumpers to set a binary combination to the golden finger for host to be aware of the retimer card PCIe bifurcation configuration.
The same jumper will trigger logic to light the LED and configure the IDT retimer accordingly

Table 5-5

	Retimer		
	4x4	2x8	1x16
GF_PRSNT3_N (b31)	0	1	0
GF_PRSNT4_N (b48)	0	0	0
GF_PRSNT6_N (b81)	0	0	1
Merge[0]	0	1	X
Merge[1]	0	1	X
Merge[2]	0	0	1
4X4 LED	Green	Yellow	Yellow
2x8 LED	Yellow	Green	Yellow
1x16 LED	Yellow	Yellow	Green

The figure and table below shows the configuration jumper on the board for different bifurcation configuration.

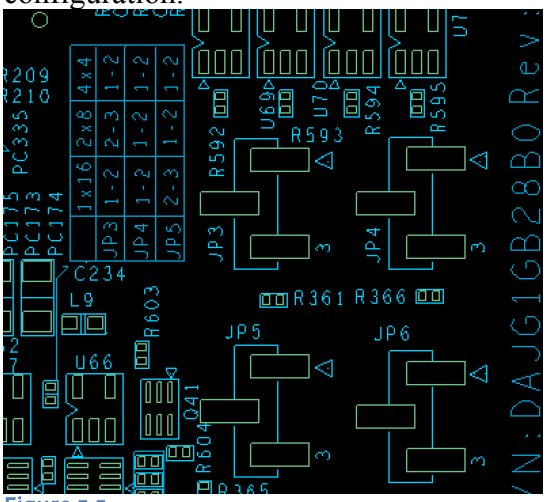


Figure 5-5

Table 5-6

Bifurcation configuration	JP3	JP4	JP5
1x16	Short 1-2	Short 1-2	Short 2-3
2x8	Short 2-3	Short 1-2	Short 1-2
4x4	Short 1-2	Short 1-2	Short 1-2

5.2.4 PCIe Clock, Reset, USB Connection

The vendor shall follow the diagram below to connect the PCIe clock (Clock buffer is RFU), PCIe reset, and USB on bridge cards. The PCIe clock from Head node to JBOG will use the SRNS (Separate Reference clocks with No SSC) scheme.

PCIe Bridge Cards – Clock/Reset/USB diagram	
Host side – with Retimer	X
JBOG side – without Retimer	

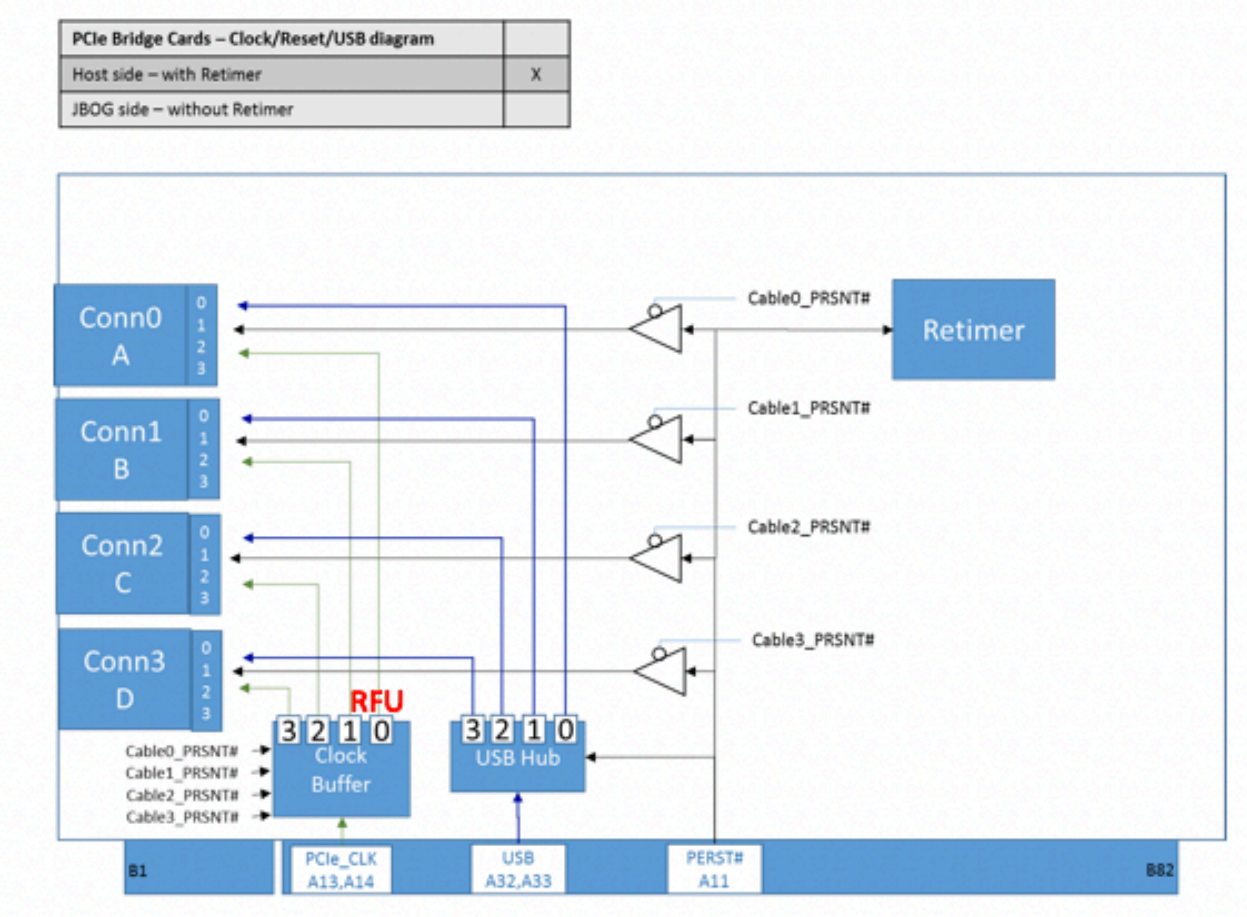


Figure 5-6

5.2.5 I2C Connection

The vendor shall follow the diagrams below for I2C connection.
There is an SMBus connection between Host side and JBOG side PCIe bridge card. This is to allow IPMB communication between BMC controllers on each side. The Pull-up of the SMBus on the interface shall be done on the host side only to avoid leakage.

PCIe Bridge Cards – I2C diagram	
Host side – with Retimer	X
JBOG side – without Retimer	

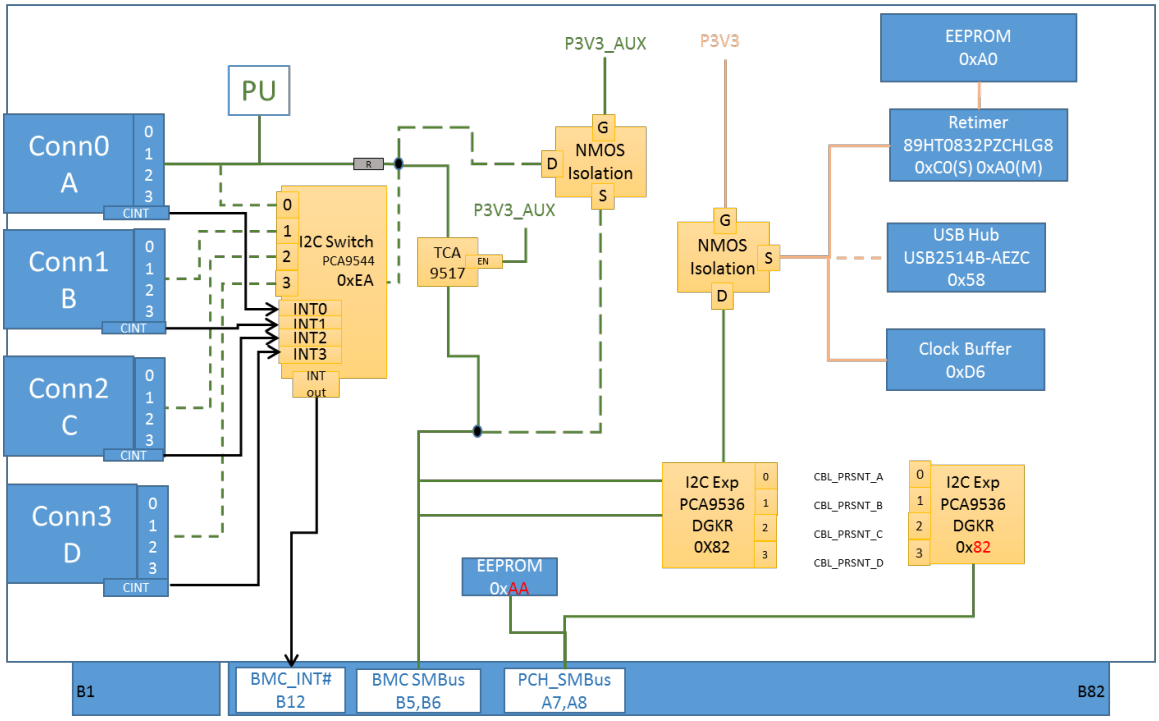


Figure 5-7

PCIe Bridge Cards – I2C diagram	
Host side – with Retimer	
JBOG Side – without Retimer	X

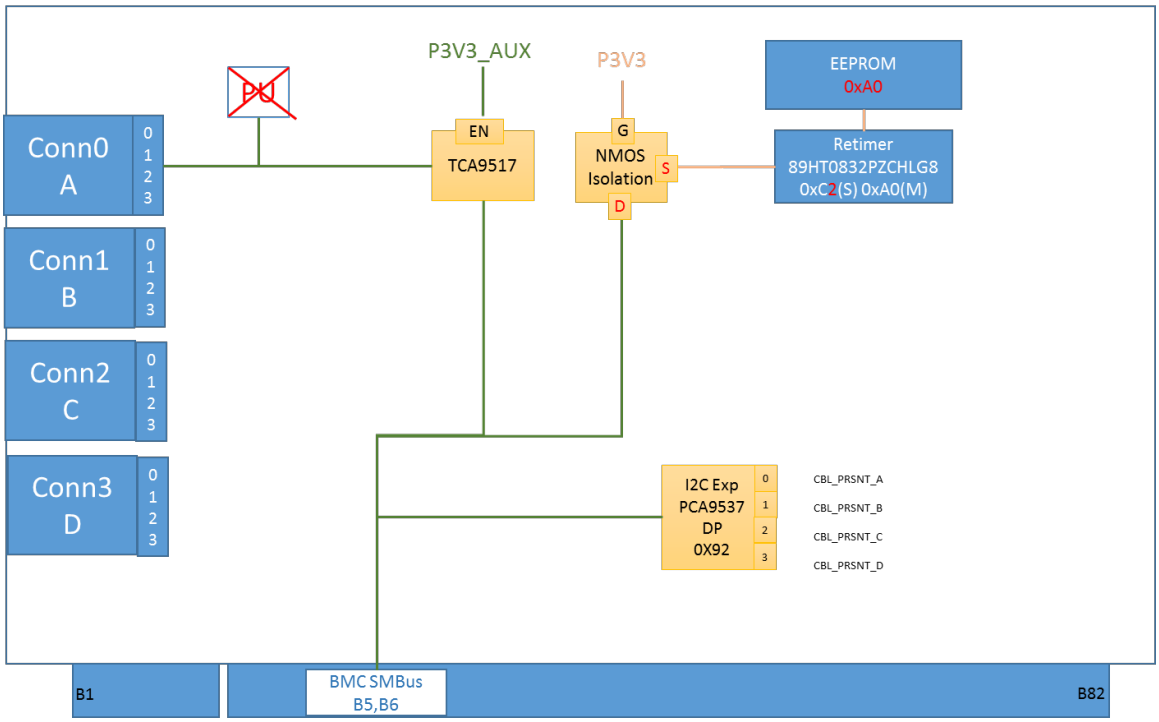
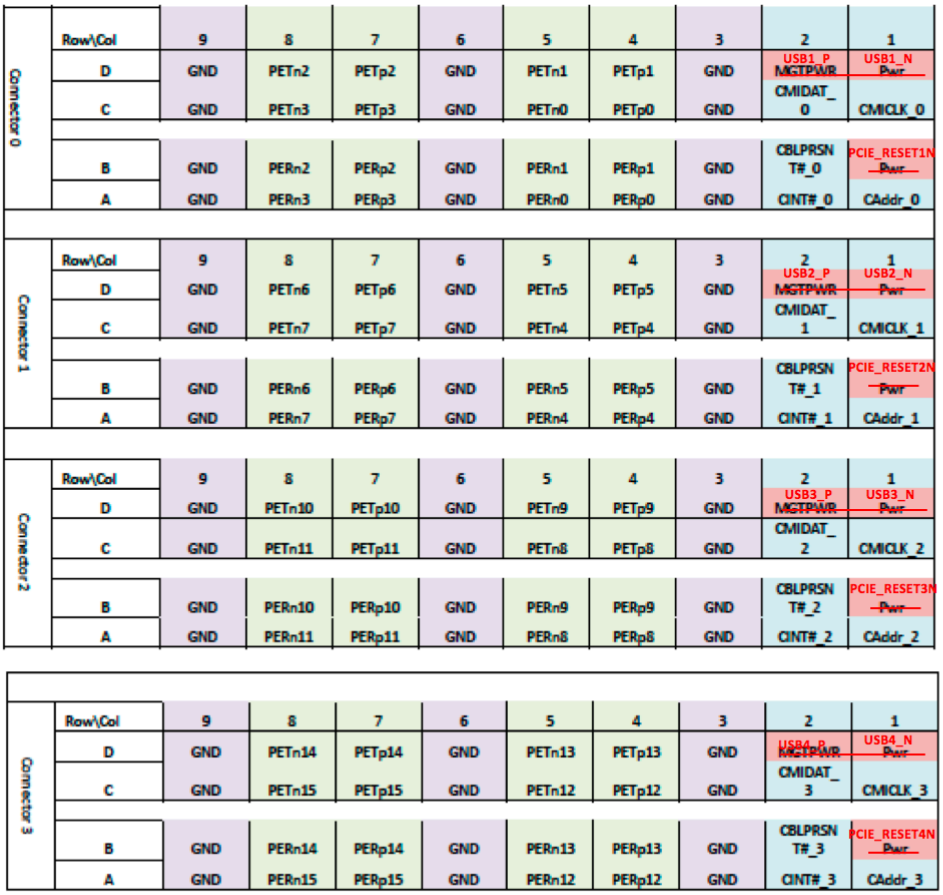


Figure 5-8

5.2.6 Retimer Card Dual Layout for PCIe Standard External Cable

The Facebook cable definition has five pins different from the PCIe standard external cabling spec (PCI Express® External Cabling Specification 3.0, REV0.6). The difference is highlighted in the cable interface diagram as shown in Figure 5-9, where power pins are re-assigned to handle USB and RESET down the cable. It should also be noted that the PCIE lane as described on the diagram is based on the specifications indicated. The actual implementation has PCIE lanes connected in **reverse** (lane 0 to 15 map to lane 15 to 0) but no polarity reversal for ease of routing.



	Row\Col	9	8	7	6	5	4	3	2	1
Connector 0	D	GND	PETn2	PETp2	GND	PETn1	PETp1	GND	USB1_P MGTPWR	USB1_N Pwr
	C	GND	PETn3	PETp3	GND	PETn0	PETp0	GND	CMIDAT_0	CMICLK_0
	B	GND	PERn2	PERp2	GND	PERn1	PERp1	GND	CBLPRSNT#_0	PCIE_RESET1N Pwr
	A	GND	PERn3	PERp3	GND	PERn0	PERp0	GND	CINT#_0	CAddr_0
Connector 1	Row\Col	9	8	7	6	5	4	3	2	1
	D	GND	PETn6	PETp6	GND	PETn5	PETp5	GND	USB2_P MGTPWR	USB2_N Pwr
	C	GND	PETn7	PETp7	GND	PETn4	PETp4	GND	CMIDAT_1	CMICLK_1
	B	GND	PERn6	PERp6	GND	PERn5	PERp5	GND	CBLPRSNT#_1	PCIE_RESET2N Pwr
	A	GND	PERn7	PERp7	GND	PERn4	PERp4	GND	CINT#_1	CAddr_1
Connector 2	Row\Col	9	8	7	6	5	4	3	2	1
	D	GND	PETn10	PETp10	GND	PETn9	PETp9	GND	USB3_P MGTPWR	USB3_N Pwr
	C	GND	PETn11	PETp11	GND	PETn8	PETp8	GND	CMIDAT_2	CMICLK_2
	B	GND	PERn10	PERp10	GND	PERn9	PERp9	GND	CBLPRSNT#_2	PCIE_RESET3N Pwr
	A	GND	PERn11	PERp11	GND	PERn8	PERp8	GND	CINT#_2	CAddr_2
Connector 3	Row\Col	9	8	7	6	5	4	3	2	1
	D	GND	PETn14	PETp14	GND	PETn13	PETp13	GND	USB4_P MGTPWR	USB4_N Pwr
	C	GND	PETn15	PETp15	GND	PETn12	PETp12	GND	CMIDAT_3	CMICLK_3
	B	GND	PERn14	PERp14	GND	PERn13	PERp13	GND	CBLPRSNT#_3	PCIE_RESET4N Pwr
	A	GND	PERn15	PERp15	GND	PERn12	PERp12	GND	CINT#_3	CAddr_3

Take note that PCIE lane numbers shown above are based on PCIE external cabling specifications.
On FB's retimer card, due to ease of routing, lane reversal for PCIE is happening, mapping lane 0 to lane 15 to lane 15 to lane 0. There is no polarity reversal.

Figure 5-9

The vendor shall implement the BOM option to make the retimer card compatible with PCIe standard external cabling spec. The vendor shall consider the schematic and CAD to have proper design for both cases. This includes but is not limited to:

- Power delivery need for trace width and decoupling cap for power to SFF connectors
- Impedance control on which signal has such a need

5.3 Host to JBOG BMC Interface

JBOG BMC AST2500 has two connections to host as below:

1. PCIe x1 to Host CPU
2. I2C/IPMB to Host BMC (BMC Message Bridging per IPMI spec)

JBOG BMC FW update support has 2 paths as below in Figure 5-10.

1. Host CPU through PCIe (doesn't require BMC firmware to be functional)

2. Host CPU through USB (require BMC firmware to be functional)

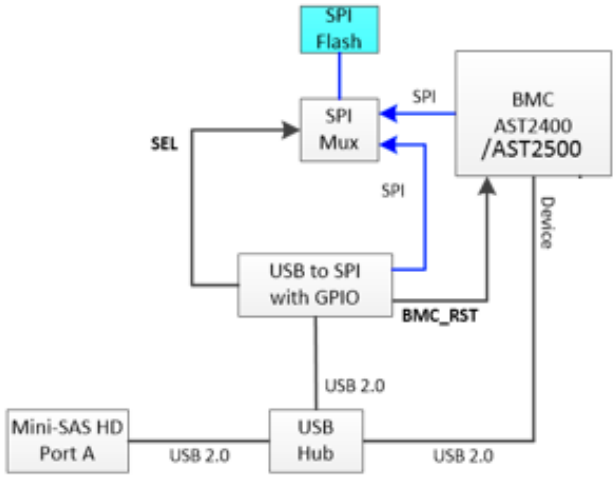


Figure 5-10

6 User LED and Buttons

This section describes the various LEDs and buttons around the JBOG system.

6.1.1 Power Button/LED and Status LED on IO Board

There is a power button on the IO board for toggling system power. Embedded on the button is the blue Power LED. Its behavior on different modes is as follows:

Table 6-1

Behavior	DC Power State	System Locator State
Always OFF	OFF	NO
ON for 0.1s, OFF for 0.9s, loop	OFF	YES
Always ON	ON	NO
ON for 0.9s, OFF for 0.1s, loop	ON	YES

The blue/orange STATUS LED beside the power button has the following behavior:

Table 6-2

Behavior	BMC Ready	System Log
Always OFF	NO	X
Blue ON	YES	NO
Orange Blinking	YES	YES

6.1.2 Network Port LED at RJ45

The network port has two LEDs as shown below.

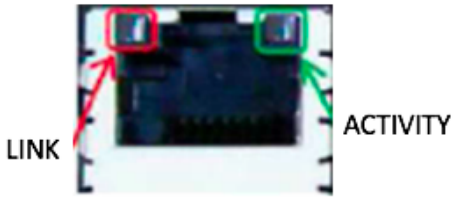


Figure 6-1

These LEDs behavior are as follows:

Table 6-3

Network port mode	Link (Upper Left) LED	Activity (Upper Right) LED
No link	OFF	OFF
1Gbps Link	Amber ON	Green Blink when activity
100Mbps Link	Green ON	Green Blink when activity
10Mbps Link	OFF	Green Blink when activity

6.1.3 FAN LED

There are FAN LEDs on fan modules, visible from back of the system.
The fan LED on the fan module behaves as follows:

Table 6-4

LED Behavior	Fan State
OFF	No Power
BLUE	FAN OK
RED	FAN Speed Lower than set threshold or Fault

7 BMC on JBOG

Similar to any other chassis implementation, the BMC on JBOG functions is able to have the ability of accessing system health information, and log and initiate preventive measures for system failure.

The BMC is implemented as follows:

1. ASPEED AST2500 chip with 4G DDR4-1600 RAM (800MHz)
2. Connected to CPLD LCMXO2-2000HC-4TG100CTR which allows:
    a. System power control and monitoring
    b. Reset control
    c. Fan LED control

7.1 Management Network Interface

JBOG's BMC OOB network interface is accessible in the front of the chassis. The RJ45 network jack that resides on the IO board is capable of 10/100/1000 IEEE802.3 Ethernet connectivity. The implementation to bring BMC's network connectivity to the chassis front end is shown below. Networking access to JBOG's BMC is optional here as it can be accessed via the Head Node's BMC (In-Band/Out-of-Band)

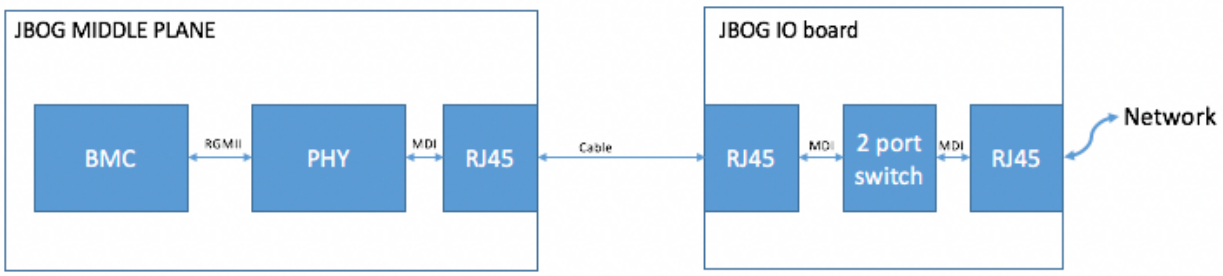


Figure 7-1

7.2 FRU

FRU EEPROMs are connected to BMC (see Section 7.3). The FRU data in EEPROM includes information such as serial number, part number, model, and asset tag. BMC would access FRU EEPROM as per IPMI specifications.

There are two FRUs in the JBOG system for the BMC to access.

Table 7-1

FRU device ID	FRU device location
0	Middle plane board
1	Baseboard

7.3 I2C/SMB/IPMB Connection

The I2C/SMB connection of the BMC implemented on Big Basin is shown below in the simplified I2C system map.

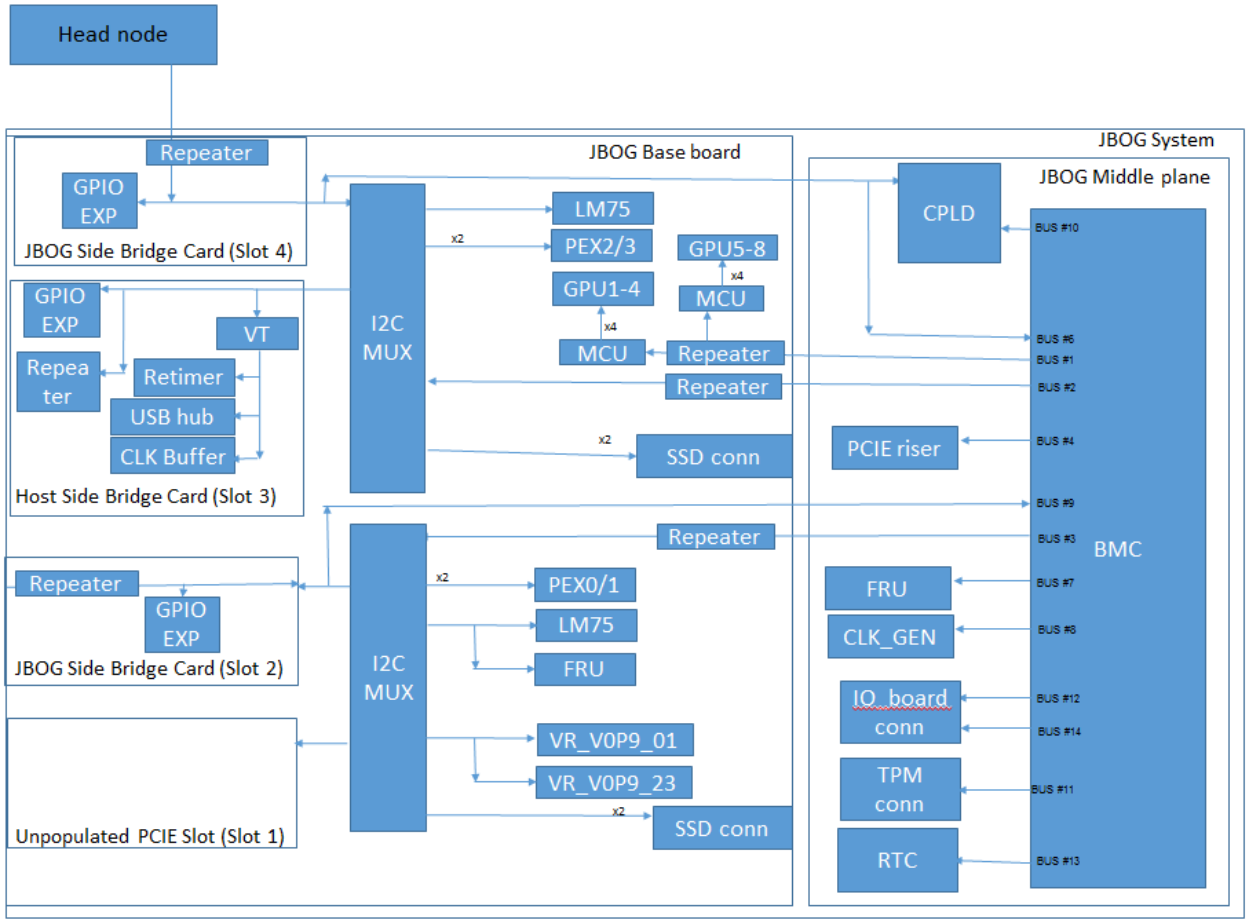


Figure 7-2

Connection of the BMC's I2C/SMBchannels on JBOG are as follows:

Channel Number	Connected to	Description
1	Baseboard MCUs accessing GPUs	The MCUs on baseboard each connects to 4 GPUs.
2	Baseboard 8 port I2C mux (0xE0)	8 port Mux is connected as follows 0: Reserved for Future Use (RFU) 1: RFU 2: PCIE Slot 3 3: PCIE Slot 4 4: PCIE switch 2 5: PCIE switch 3 6: Not used 7: Thermal Sensor (LM75,0x96)

3	Baseboard 8 port I2C mux (0xE0)	8 port Mux is connected as follows 0: Reserved for Future Use (RFU) 1: RFU 2: PCIE Slot 1 3: PCIE Slot 2 4: PCIE switch 0 5: PCIE switch 1 6: Voltage regulators for 0.9V and main board (VR1, 0xC4 and VR2, 0xC6) 7: FRU(EEPROM, 0xA8) and Thermal Sensor (LM75, 0x96)
4	RFU	For PCIE Riser, not used at the moment.
5	RFU	For extending to 2nd Chassis, connected to Slot 1
6	Bridge card at PCIE Slot 4	This is where BMC is the slave IPMB channel for Head node access.
7	JBOG middle plane FRU (0xA8)	FRU EEPROM
8	JBOG Clock Generator (0xD0)	100MHz Clock generator from mid plane to PCIE devices on main board.
9	Bridge Card at PCIE Slot 2	Another slave channel from PCIE slot. RFU.
10	CPLD (TBD)	Running PMBus on this channel. The CPLD serves as an arbiter for PMBUS access to 3 hotswap controllers on the middle plane. In addition, the CPLD More information on Section 7.4
11	TPM chip (TBD)	TPM chip is actually on a TPM module on a separate board. Connected at J1084.
12	LM421 temperature sensor on IO board (0x98)	Temperature sensor to measure chassis ambient. Has local sensor (on chip) and remote sensor. The remote sensor connected to a thru hole BJT as air temperature sensor.
13	On board RTC chip (0xA2)	RTC is a PCF85263A chip.
14	GPIO expander on IO board (0x40)	RFU. GPIO expander mainly for OCP v2 debug.

7.4 CPLD

The CPLD, implemented using LCMXO2-2000HC-4TG100CTR on Big Basin, serves as the following functions that supplements the BMC's need to monitor and control around the system.

1. 12V HSCs PGOOD, ALERT monitoring and their ENABLE control
2. PMBUS access to HSCs. (three PMBUSes to 3 HSCs; one to each)
3. Power and Reset Sequencing for the system.
4. Fan PGOOD, present monitoring and FAN LED control
5. GPU throttling
6. To behave like a sensor for BMC to query and take note of its status, thus able to log into SEL when are anomalies observed.

The diagram below shows major connectivity of the CPU around the system.

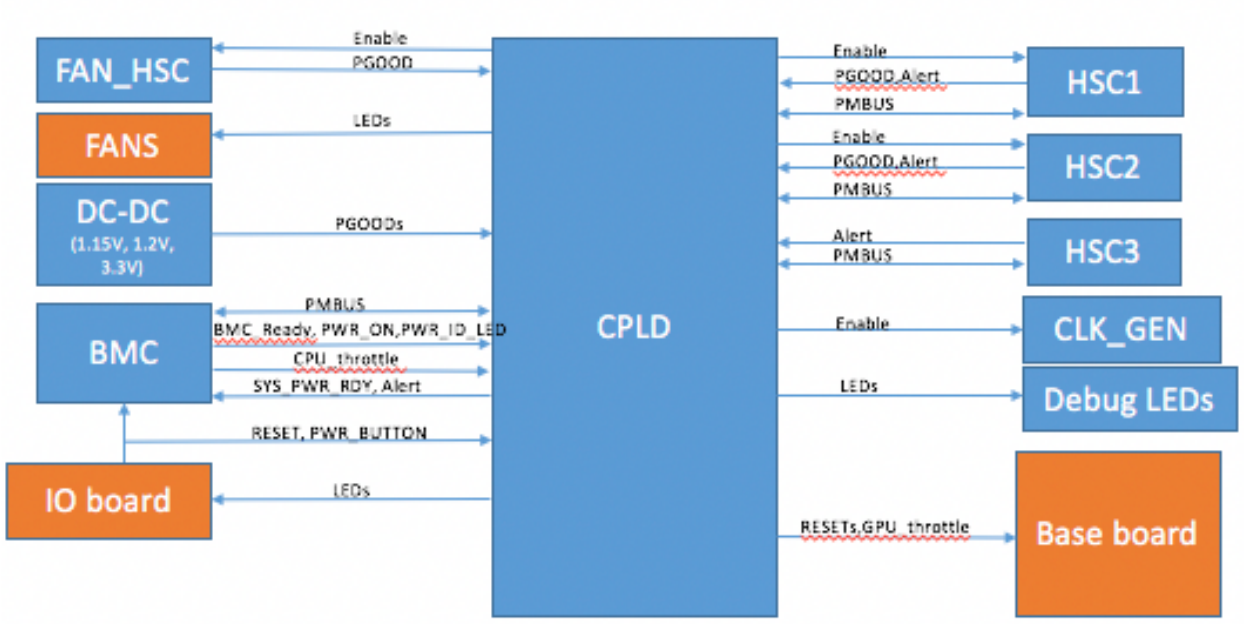


Figure 7-3

7.5 Debug support

BMC is to be implemented with the support of two types of debug support connectors.
The first is a 14-pin 2x7 header is accessible at the front of the chassis. Pin arrangement is shown below:

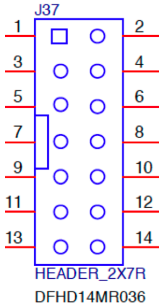


Figure 7-4

Its pin descriptions are shown in the table below:

Pin	I/O/IO with respect to system	Description
1	O	Low HEX character [0], LSB
2	O	Low HEX character [1]
3	O	Low HEX character [2]
4	O	Low HEX character [3], MSB
5	O	High HEX character [0], LSB
6	O	High HEX character [1]
7	O	High HEX character [2]
8	O	High HEX character [3], MSB
9	O	UART transmit from system
10	I	UART receive into system
11	I	Reset signal from button. Active low
12	I	Present/UART channel select
13	GND	Ground reference pin
14	O	5V from system to OCP v1 debug board.

The second debug support connector uses a modified USB3.0 connector.

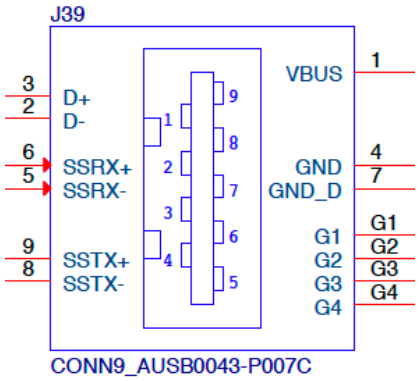


Figure 7-5

Pin definition for this debug header is as follows:

Pin	I/O/IO with respect to system	Description
1	O	5V from system
2	NA	Not used
3	NA	Not used
4	O	GND return
5	IO	SMBUS clock
6	IO	SMBUS Data
7	I	Present/UART channel select
8	O	UART transmit from system
9	I	UART receive into system

The relevant POST codes implemented for the JBOG can be found in Section 17.1, which is served as a reference.

It should be noted that concurrent serial UART for communication from the two debug connectors is possible as TX from the system is broadcasted and RX from debug support are AND-ed. Therefore, to avoid data corruption, when using one of the UARTs, the other UART should not be used.
The modified USB3.0 debug connector is to be implemented such that it will not damage any USB3.0 devices when plugged in.

7.6 Local Serial Console

UART access to BMC (also known as Local Serial Console) is to be implemented with the following requirements

- The local serial console is to be functional in all stages of system operation. During booting, POST codes sent to the 7-segment display should also be available on the Local Serial Console to allow console logging and debugging. For example, codes like “[00]”, “[01]”, “[02]”, “[E0]”, etc, should be displayed on the console upon power up.
- The last POST code received upon entering normal BMC console is to be displayed in the lower right hand corner of the console screen.

- The serial console buffer for the last five console screens are to be stored in volatile media; each screen is of standard 80 columns by 24 rows. Screens are to be completely erased within five seconds of standby power removal for security and privacy reasons.
- The OOB raw command is to have the ability to extract and displayed what is shown in the console screen buffer for remote debugging purposes and can be used for scaling data collection. This is done using a buffer dumping script with OEM commands.

7.7 Remote Power Control and Power Policy

The BMC implemented needs to have the following capabilities

- Support remote power on/off/cycle and warm reboot either through In-Band or Out-of-Band IPMI command.
- Support power on policy to be last-state, always-on and always-off upon recovery from an AC power loss event. The default setting is to be always-on. Power policy changes can be done through IPMI command and take effect without a cold reset or system reboot.
- Take no more than three seconds upon AC power on for BMC to power up and display system POST on debug support ports and local serial console.
- Take no more than three seconds to be able to process a power button signal upon AC power on.

7.8 Power/Thermal Monitoring and Limits

The BMC needs to support power and thermal monitoring around the system. Limits are also set to ensure system health is good. In the event of the power/temperature exceeding limits set by the user, BMC is able to log the event in SEL log, and initiate preventive measures to avoid system meltdown. Ability to monitor the system power and thermal information is to be made available in In-Band and Out-of-Band manners.

7.9 Sensors Used

Various sensors are deployed around in order to support BMC functions like power/thermal monitoring, throttling, fan-speed control. Sensors are classified as Analog, Discrete and Event-only type sensors.

7.9.1.1 Analog Sensors

The BMC implemented is to have access to all analog sensors placed in the system and ensure that they are displayed in a sensor data record repository. The tables below list different types of analog sensors and their set limits around the JBOG system.

7.9.1.1.1 Temperature Monitoring Sensors

Sensor Name	Sensor Number	Lower Critical Threshold	Upper Non-Critical Threshold	Upper Critical Threshold	Availability
Temp_Outlet	0x20	na	na	na	DC on

Temp_Inlet	0x21	na	40	42	Standby
Temp_PCIE_Inlet1	0x22	na	65	70	DC on
Temp_PCIE_Inlet2	0x23	na	65	70	DC on
Temp_PLX0	0x24	na	92	93	DC on
Temp_PLX1	0x25	na	92	93	DC on
Temp_PLX2	0x26	na	92	93	DC on
Temp_PLX3	0x27	na	92	93	DC on
Temp_GPU0	0x90	na	81	82	DC on
Temp_GPU1	0x91	na	81	82	DC on
Temp_GPU2	0x92	na	81	82	DC on
Temp_GPU3	0x93	na	81	82	DC on
Temp_GPU4	0x94	na	81	82	DC on
Temp_GPU5	0x95	na	81	82	DC on
Temp_GPU6	0x96	na	81	82	DC on
Temp_GPU7	0x97	na	81	82	DC on
HSC1 Temp	0x51	na	80	85	DC on
HSC2 Temp	0x58	na	80	85	DC on
HSC3 Temp	0x5F	na	80	85	DC on

7.9.1.1.2 Voltage Monitoring Sensors

Sensor Name	Sensor Number	Lower Critical Threshold	Upper Critical Threshold	Availability
P12V_AUX	0x70	11.24	13	Standby
P3V3_AUX	0x71	3.04	3.55	Standby
P5V_AUX	0x72	4.6	5.4	Standby
P12V_1	0x73	11.24	13	DC on
P12V_2	0x74	11.24	13	DC on
P12V_FAN	0x75	11.24	13	DC on
P12V_SSD	0x76	11.24	13	DC on
P3V3	0x77	3.04	3.55	DC on
P3V_BAT	0x78	2.74	3.56	Standby
HSC1 Input Voltage	0x52	10.98	12.96	DC on
HSC1 Output Voltage	0x54	10.98	12.96	DC on

HSC2 Input Volt	0x59	10.98	12.96	DC on
HSC2 Output Volt	0x5B	10.98	12.96	DC on
HSC3 Input Volt	0x60	10.98	12.96	DC on
HSC3 Output Volt	0x62	10.98	12.96	DC on

7.9.1.1.3 Power Monitoring Sensors

Sensor Name	Sensor Number	Lower Critical Threshold	Upper Critical Threshold	Availability
HSC1 Input Power	0x53	na	1920	DC on
HSC2 Input Power	0x5A	na	2220	DC on
HSC3 Input Power	0x61	na	480	DC on
HSC total PIN	0x50	na	na	DC on
PWCS_GPU0	0x98	na	420	DC on
PWCS_GPU1	0x99	na	420	DC on
PWCS_GPU2	0x9A	na	420	DC on
PWCS_GPU3	0x9B	na	420	DC on
PWCS_GPU4	0x9C	na	420	DC on
PWCS_GPU5	0x9D	na	420	DC on
PWCS_GPU6	0x9E	na	420	DC on
PWCS_GPU7	0x9F	na	420	DC on

7.9.1.1.4 Current Monitoring Sensors

Sensor Name	Sensor Number	Lower Critical Threshold	Upper Critical Threshold	Availability
HSC1 Output Current	0x55	na	160	DC on
HSC2 Output Current	0x5C	na	185	DC on
HSC3 Output Current	0x63	na	40	DC on

7.9.1.2 Discrete Sensors

The list of discrete sensors being deployed around the JBOG system are as follows:

Sensor Name	Sensor #	Entity ID	Instance	Sensor Type	Event/Reading Type
Fan Present	40h	1Eh	00h	Fan – 04h	Sensor Specific – 6Fh
HSC1 Sts Low	56h	14h	01h	OEM – C1h	Sensor Specific – 6Fh
HSC2 Sts Low	5Dh	14h	02h		
HSC3 Sts Low	64h	14h	03h		
HSC1 Sts High	57h	14h	01h	OEM – C1h	Sensor Specific – 6Fh
HSC2 Sts High	5Eh	14h	02h		
HSC3 Sts High	65h	14h	03h		
Event Log	D0h	06h	01h	10h	Sensor Specific – 6Fh
Watchdog	D1h	06h	02h	Watchdog 2 – 23h	Sensor Specific – 6Fh

7.9.1.3 Event-Only sensors

Sensors that are not listed in the SDR are classified as event only sensors. These sensors trigger the SEL log if an abnormal value is detected. The following are the event-only sensors used in the system.

Sensor	Sensor #	Entity ID	Instance	Sensor Type	Event/Reading Type
Power State	01h	13h	01h	Power Unit - 09h	Sensor Specific – 6Fh
Button	02h	0Ch	01h	Button/Switch – 14h	Sensor Specific – 6Fh
BMC FW Health	03h	90h	01h	Management Subsystem Health - 28h	Sensor Specific – 6Fh
Power Policy	04h	13h	01h	Power Unit – 09h	Sensor Specific – 6Fh
HSC1 Alert	05h	14h	01h	Power Supply – 08h	Digital Discrete – 03h
HSC1 CSOUT Alert	06h	14h	01h	Power Supply – 08h	Digital Discrete – 03h
HSC1 TIMER Alert	07h	14h	01h	Power Supply – 08h	Digital Discrete – 03h
HSC2 Alert	08h	14h	02h	Power Supply – 08h	Digital Discrete – 03h

HSC2 CSOUT Alert	09h	14h	02h	Power Supply – 08h	Digital Discrete – 03h
HSC2 TIMER Alert	0Ah	14h	02h	Power Supply – 08h	Digital Discrete – 03h
HSC3 Alert	0Bh	14h	03h	Power Supply – 08h	Digital Discrete – 03h
Power Error	0Ch	15h	01h	Power Supply – 08h	Digital Discrete – 03h
GPU N0 PSU Alert	0Dh	15h	01h	Power Supply – 08h	Digital Discrete – 03h
GPU N0 Throttle	0Eh	15h	01h	Microcontroller/Coprocessor – 16h	Digital Discrete – 03h
GPU N1 PSU Alert	0Fh	15h	01h	Power Supply – 08h	Digital Discrete – 03h
GPU N1 Throttle	10h	15h	01h	Microcontroller/Coprocessor – 16h	Digital Discrete – 03h
GPU N0 Over Temp	11h	12h	01h	Temperature – 01h	Digital Discrete – 03h
GPU N1 Over Temp	12h	12h	01h	Temperature – 01h	Digital Discrete – 03h

7.10 System Event Log (SEL)

The BMC needs to support SEL capabilities. The following items are to be logged in the SEL

- Analog sensors exceeding thresholds that are being set.
- Discrete sensors that indicates Assertion and De-assertion in Section 7.9.1.2
- Event-only sensors like
  - All types of PCIE errors, status change
  - All types of NVLINK errors, status change
  - All types of POST error
  - All types of Machine Check Error
  - All types of Fan failures (like fan speed out of range), and identify which fan failed
  - I2C/SMA/PMBUS transactions errors or device NACK
  - Any other system anomalies

7.10.1.1 Error thresholds Settings

Programmable threshold settings for both correctable and uncorrectable errors are implemented in the BMC. Logging is only done when the number of errors hits its threshold. This feature is catered to minimize the number of log entries in the SEL. Threshold can be set low (or to zero) in the product development phase to allow debug and higher in mass production (MP) phase.

- PCIE errors – Follow threshold settings per chipset vendor recommendation in MP phase.
- NVLink errors - Follow threshold settings per chipset vendor recommendation in MP phase
- BMC Memory Correctable ECC – [1000] in MP phase.

7.10.1.2 Critical SEL filter

OEM commands are required to set and display two different level of SEL filtering.

The default is to log all errors during EVT/DVT/PVT with the option to log only critical SEL that needs servicing or indicates power cycle state change. Ability to log when user clear the SEL log or when it has overflowed.

7.11 Fan Speed Control (FSC) on JBOG

7.11.1 Data gathering for FSC

To have an effective fan speed control mechanism in the system, the following data would be needed:

Type of data	Data used for FSC input
Temperature	All sensors listed in Section 7.9.1.1.1
Power	All sensors listed in Section 7.9.1.1.3
Fan speed	4 FAN tachometer inputs

The sampling rate of the these should be ≥ 1 sample/s.

7.11.2 FSC in BMC

The BMC needs to implement FSC and an FSC update interface per *Facebook Server Fan Speed Control Interface*[3]. The control can be done through both In-Band and Out-of-Band FSC configuration updates. Updates should take effect immediately without requiring a reboot.

7.11.3 Fan Connection

The fan connector for the fan modules uses a re-mapped pin-out for PCIEx1 PCB edge connector located on the Mid plane.

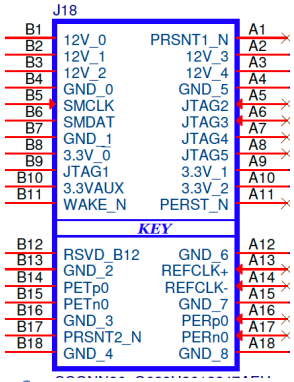


Figure 7-6

The re-defined connector pin-out as follows:

Pin	I/O/IO with respect to system	Description
A1	NA	RFU
A2	O	12V
A3	O	12V
A4	GND	Power Return
A5	NA	RFU

[3] http://files.opencompute.org/oc/public.php?service=files&t=d48482b8b87a596dd93ac5b80e9fa3a2&download

A6	O	12V
A7	NA	RFU
A8	NA	RFU
A9	NA	RFU
A10	NA	RFU
A11	NA	RFU
A12	GND	Power Return
A13	NA	RFU
A14	NA	RFU
A15	NA	RFU
A16	NA	RFU
A17	NA	RFU
A18	NA	RFU
B1	O	12V
B2	O	12V
B3	O	12V
B4	GND	Power return
B5	O	FAN_PWM
B6	NA	RFU
B7	GND	RFU
B8	O	FAN_PWM
B9	NA	RFU
B10	NA	RFU
B11	NA	RFU
B12	O	FAN_LED
B13	GND	Power return
B14	I	FAN_TACH<1>
B15	I	FAN_TACH<0>
B16	GND	Power return
B17	I	FAN_PRESENT
B18	GND	Power return

7.12 OEM Commands

BMC also includes OEM features that are control using the OEM command listed in the table below

Table 7-2

Command	NetFn /LUN	Cmd	Description
Get UBOOT ETHADDR	0x2E/00	0x24	Get MAC address
Get CPLD Info	0x30/00	0x17	Get CPLD information **Request:** Byte 1 – Access Type 01h – CPLD checksum 02h – CPLD device ID (IDCODE) 03h – CPLD UES (User Electronic Signature Data) **Response:** Byte 1 – Completion Code If access type is 01h: Byte 2:3 – checksum data If access type is 02h: Byte 2:5 –device ID (IDCODE) If access type is 03h: Byte 2:5 –UES (User Electronic Signature Data)

7.13 BMC Firmware Update

A remote BMC firmware update needs to be supported due to possible lack of any physical access to the system. A remote update happens either through Out-of-Band by the management network or through In-Band by logging into local OS (CentOS) with with data network. Tool(s) shall support 64-bit CentOS version 6.4 and above with an updated Kernel specified by customer.

A remote BMC firmware update may take a maximum of five minutes to complete. The I2C sideband has a bottleneck to achieve this requirement and an NC-SI interface is needed. The BMC firmware update process and BMC reset process require no reboot or power down of the host system and should have no impact to normal operation of the host system. The BMC needs to be fully functional with updated firmware after the update and reset without further configuration.

An In-band BMC firmware update can go through KCS or USB. USB is the preferred interface due to higher update speed.

A default update should recovery BMC to factory default; an option needs to be provided to preserve SEL, configuration. The MAC address is based on NIC MAC, so it would not be cleared with a BMC firmware update.

7.14 CPLD Firmware Update

The vendor should implement the BMC to access JBOG CPLD through JTAG and perform a CPLD code upload from the remote-control server to the BMC, update code from the BMC to the CPLD through a CPLD JTAG interface, and verify CPLD code. All the steps above shall be done from the OOB command line mode. The vendor shall provide an update utility that supports 64-bit CentOS version 6.4 and above with an updated Kernel specified by the customer.

The BMC shall implement an OEM command to read CPLD FW checksum, device ID, and FW version.

7.15 BMC Average Power Reporting

The BMC shall record READ_EIN_EXT every 0.1 seconds in a ring buffer. The BMC shall support the OEM command Get_PIN or similar to calculate and report the average power between the current and a duration defined in the Get_PIN command. The ring buffer size shall support any duration from 0.1 second to 60 seconds, in 0.1-second increments. The return shall have a resolution of 0.1W.

8 Thermal Design Requirements

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

- Inlet temperature lower than or equal to 35°C, and 0" H_2O datacenter pressure with all FANs in each thermal zone running properly
- Inlet temperature lower than or equal to 35°C, and 0.05" H_2O datacenter pressure with one FAN (or one rotor) in each thermal zone failed

8.1 Data Center Environmental Conditions

The thermal design for JBOG needs to satisfy the data center operational conditions as described below.

8.1.1 Location of Data Center/Altitude

Data centers may be located 6000 feet above sea level or higher. Any variation of air properties or environmental difference due to the high altitude needs to be considered when creating the thermal design.

8.1.2 Cold-Aisle Temperature

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

8.1.3 Cold-Aisle Pressurization

Data centers will maintain the cold aisle pressure to be between 0" H_2O and 0.05" H_2O. The thermal solution of the system should be considered the worst operational pressurization in a data center, which it is 0" H2O, and 0.05" H2O with a single fan (or rotor) failure.

8.1.4 Relative Humidity

Most data centers will maintain a relative humidity of between 20% and 90%. In the thermal design, the environmental condition changes due to the high altitude may not be considered when the thermal design can meet the requirement with maximum relative humidity, 90%.

8.2 Operational Condition

8.2.1 Pressurization

Except for the condition when one rotor in a fan module fails, the thermal solution should not be found with considering extra airflow from a data center cooling system. If and only if one rotor in the fan module fails, the negative or positive DC pressurization can be considered in the thermal solution in the hot aisle or in the cold aisle respectively.

8.2.2 System Fan Rotor Redundancy

The system supports N+1 fan rotor redundancy, which should be sufficient for cooling server components to temperatures below their maximum spec to prevent server shut down or to prevent either CPU or memory throttling.

When a rotor fails, all other fan rotors will be set to run at 100% and be in this mode without thermal risk. Warning messages will be triggered for a fan module replacement. The replacement of the failed fan module (one rotor failed) to a new fan into the system should be kept within 40 seconds to avoid thermal risk to the system.

8.2.3 System Airflow or Volumetric Flow

The unit of airflow (or volumetric flow) used for this spec is CFM (cubic feet per minute). 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 standard deviation). See section 8.1.2 for the temperature definitions.

8.2.4 Delta T

The delta T is the air temperature difference across the system or the temperature difference between outlet air temperature of the system and inlet air temperature of the system. The delta T must be greater than 22°F. The desired delta T is 20°C (36°F) when the inlet air temperature to the system is lower than 30°C.

8.2.5 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 2% 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 Sink

The heat sink design should choose to be the most optimized design with lowest cost. The heat sink design should be reliable and the most energy efficient design that satisfies all the conditions described above. The number of heat pipes in the heat sink should not be more than three. The ODM can always propose for a different heat sink type if there is an alternative way to provide cost benefits. The heat sink should be without complex installation guidance, such as air-flow direction.

8.3.2 System Fans

The system fan must be highly power-efficient with dual bearing. The propagation of vibration caused by fan rotation should be minimized and limited. Fans of 92mmx92mmx38mm are used in JBOG systems. Hot swap controller design for system fans is to be maintained at less than 1A on the 12V rails. Change in PWM from low percentage to high percentage is to be kept at less than 10% overshoot of the additional current to the fans.

8.3.3 Air-flow Design

Any air-flow design for the JBOG system needs to be the most simple, energy efficient and easily serviceable. The design should be unified for all SKUs. Use of green material or reusable material for parts is preferred.

8.3.4 Thermal Sensor

The maximum allowable tolerance of thermal sensors in the JBOG is ±3°C.

9 Mechanical

9.1 JBOG Mechanical

9.1.1 JBOG Sled View

Shown in Figure 9-1 is the isometric view of the JBOG system.

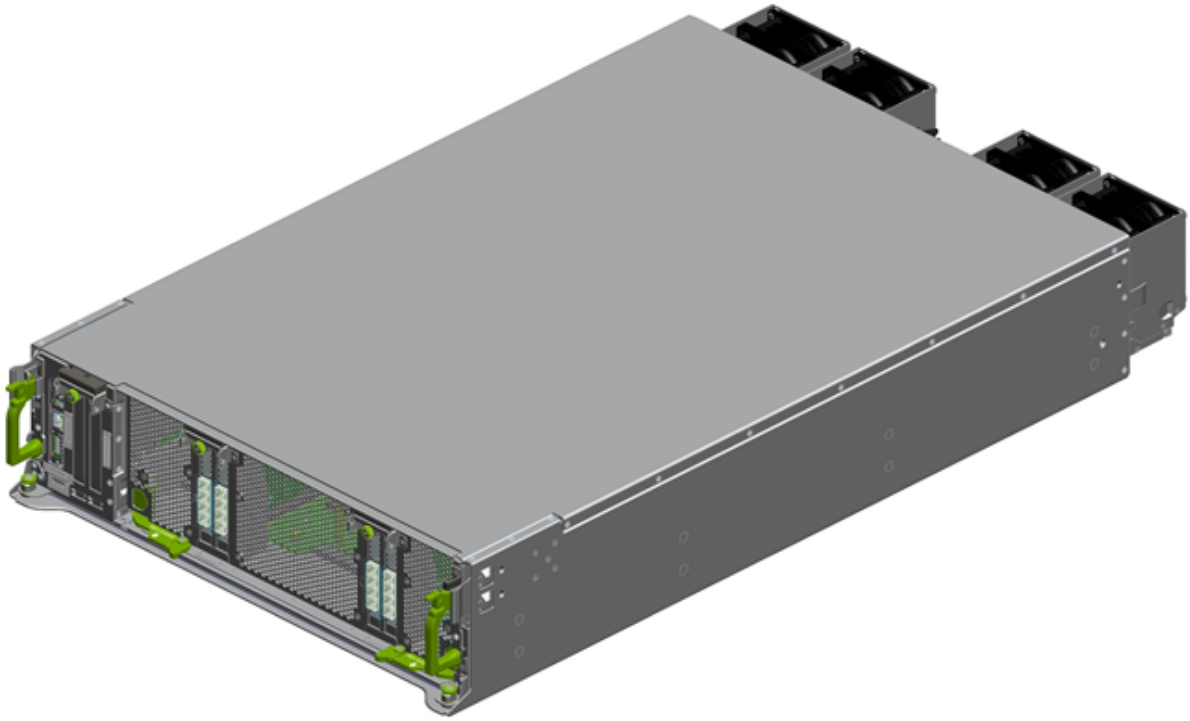


Figure 9-1

9.1.2 JBOG Top View

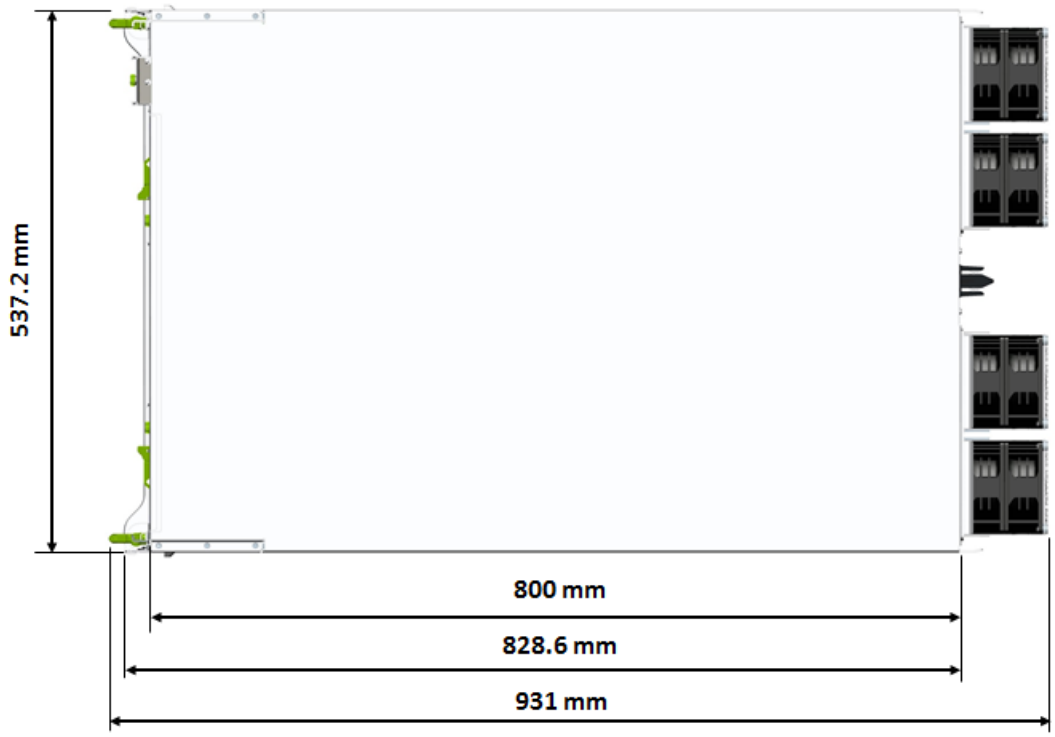


Figure 9-2

Top view with cover opened

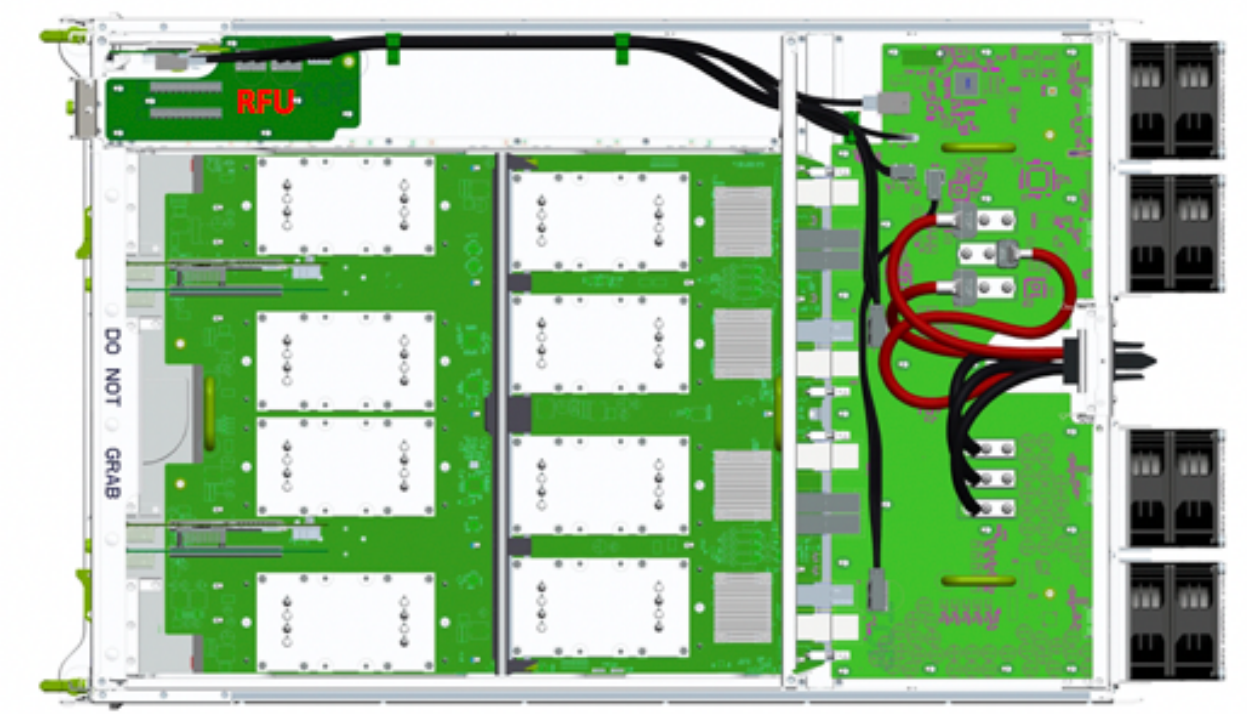


Figure 9-3

9.1.3 JBOG Front View and Rear View

Front and Rear view of the JBOG system are shown below with Topology 1 configured as a reference.

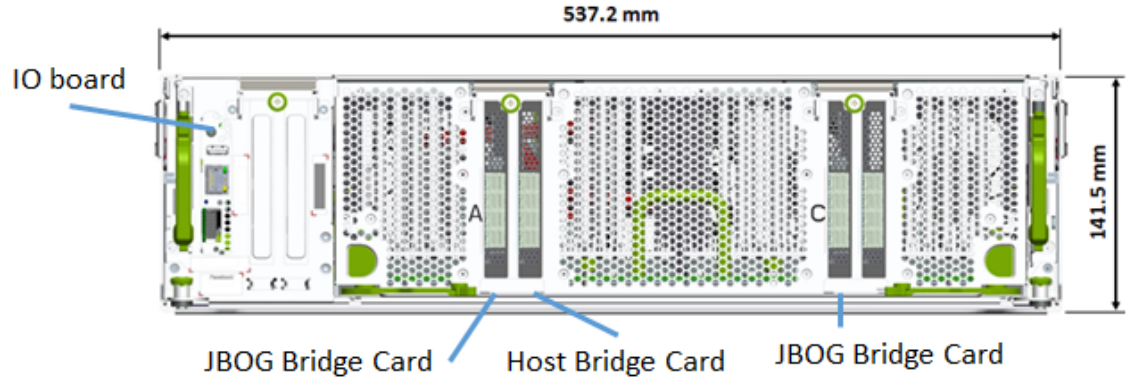


Figure 9-4

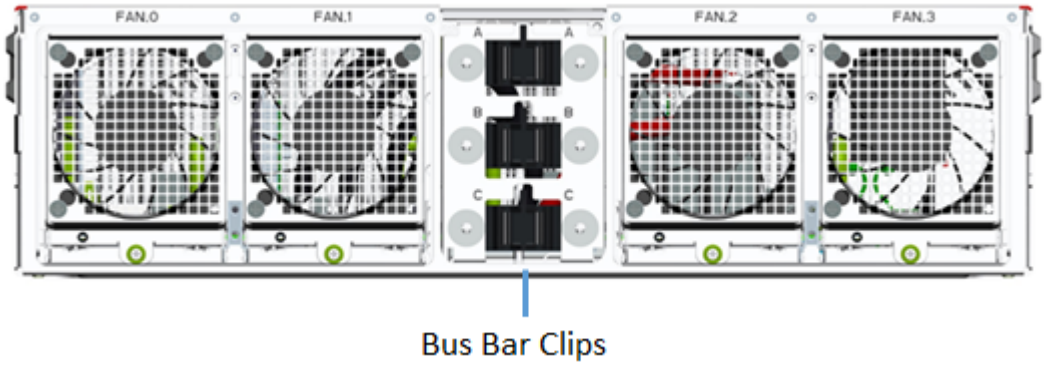


Figure 9-5

9.1.4 Baseboard

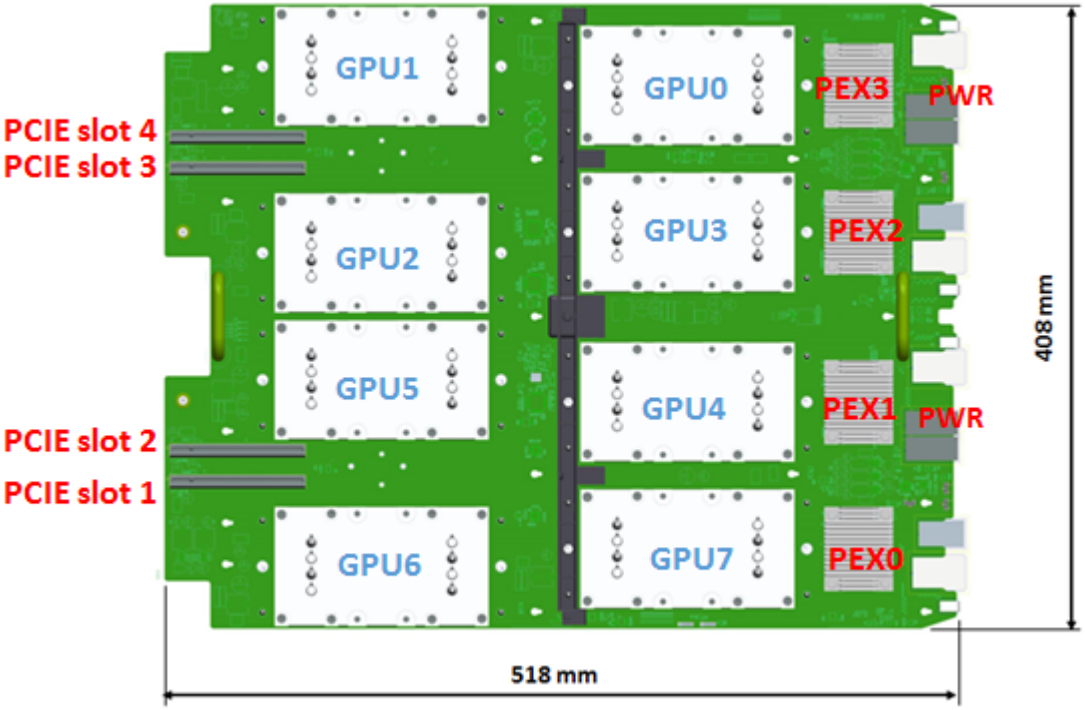


Figure 9-6

9.1.5 Middle plane board

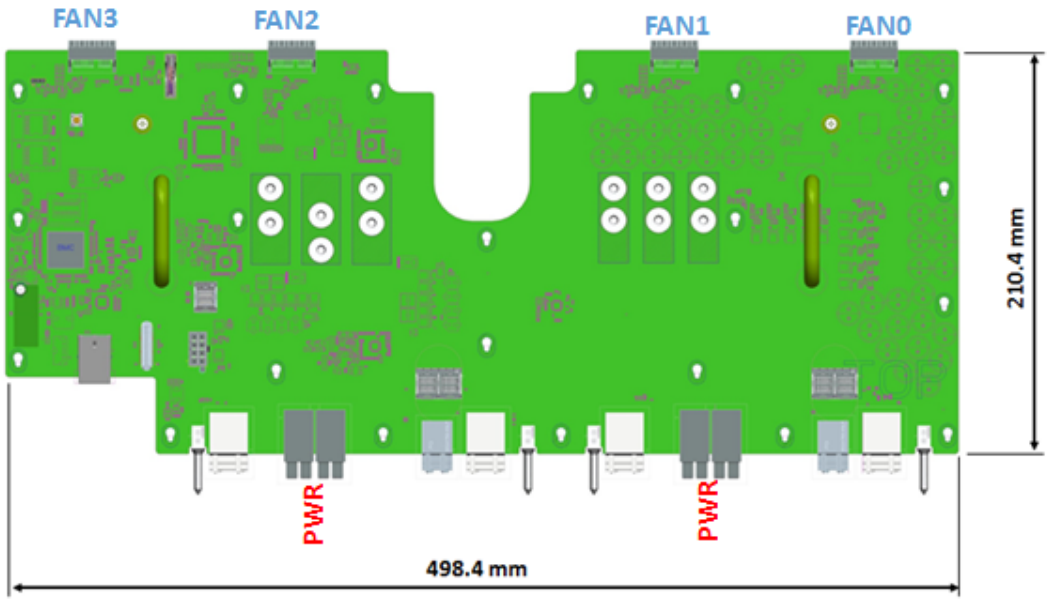


Figure 9-7

9.1.6 Host Side Bridge Card

The host side bridge card has LEDs and configuration jumpers as mentioned in Section 5.2.2 and Section 0. They are shown below in Figure 9-8.

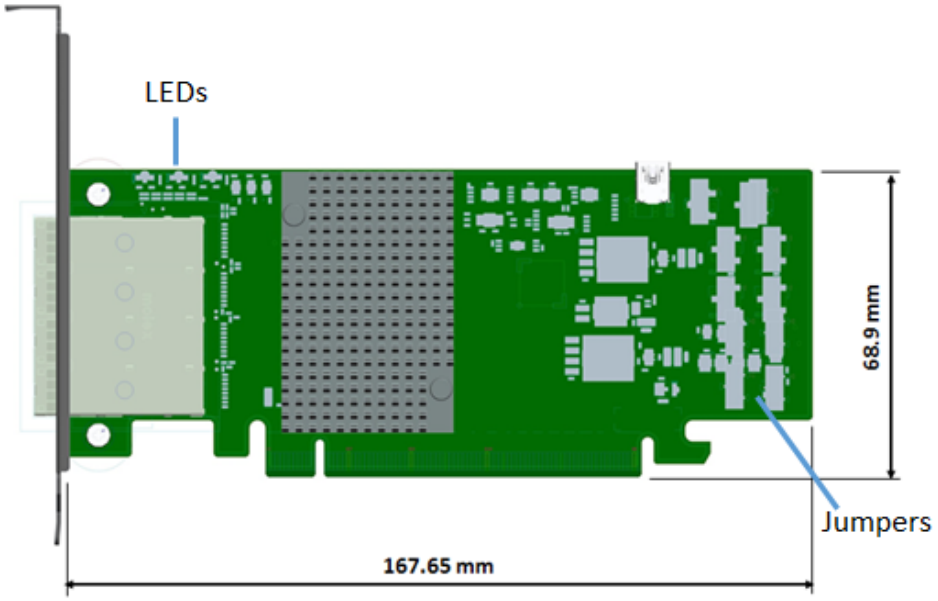


Figure 9-8

9.1.7 JBOG Side Bridge Card

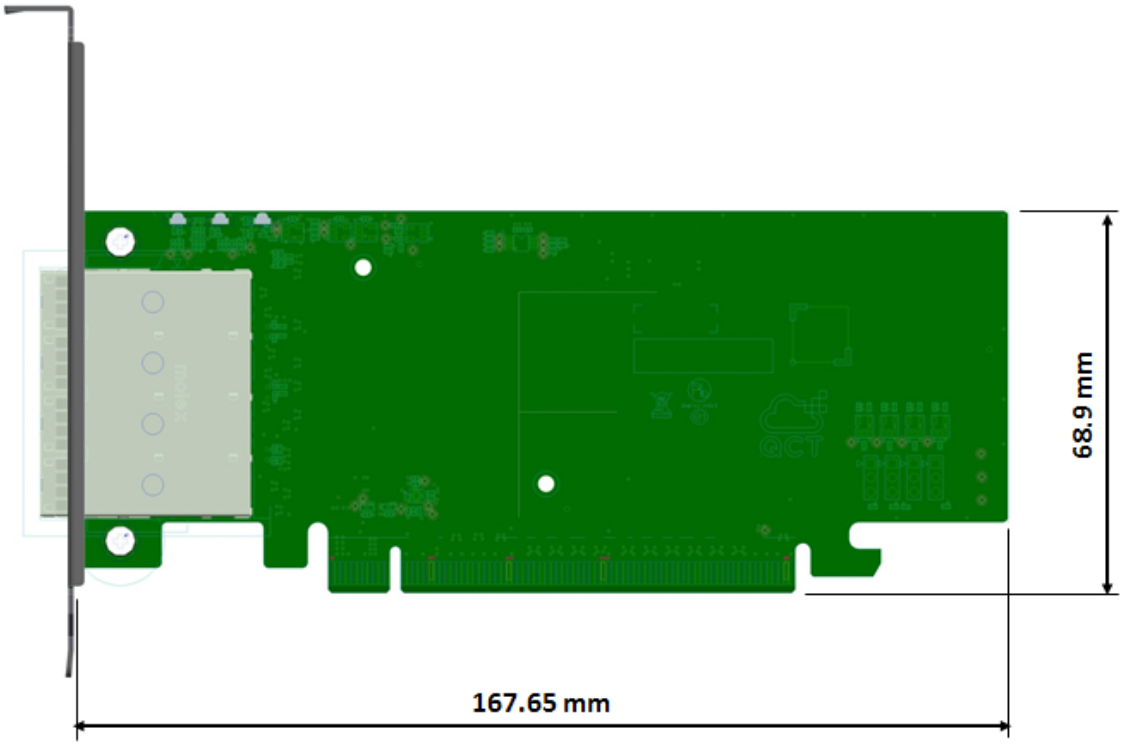


Figure 9-9

9.1.8 IO Board

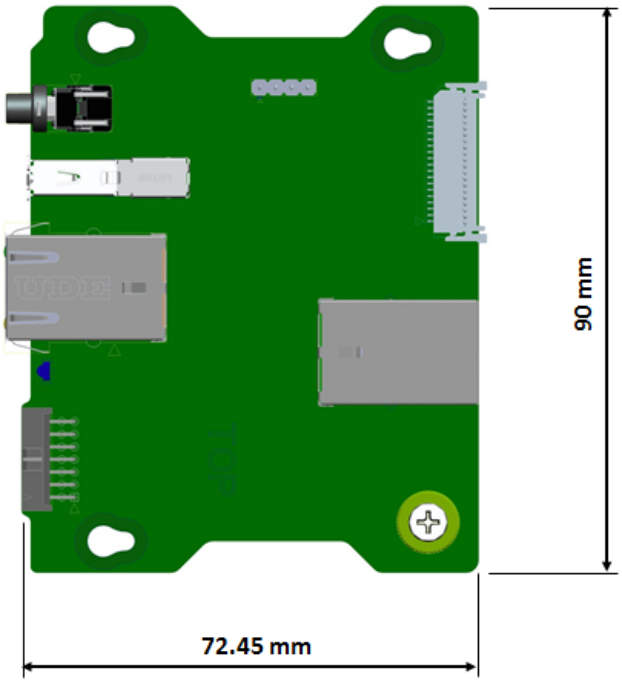


Figure 9-10

9.2 PCB Stackup Information

9.2.1 Baseboard and Middle Plane Board Stackup

Layer#	Description	Copper Weight (oz)	Thickness (mil)
	Soldermask		0.60
1	TOP	0.5+plating	1.95
	PP		3.00
2	GND	1	1.30
	CORE		3.50
3	IN1	1	1.30
	PP		3.30
4	GND1	1	1.30
	CORE		3.50
5	IN2	1	1.30
	PP		3.30
6	GND2	1	1.30
	CORE		3.50
7	IN3	1	1.30
	PP		3.30
8	GND3	1	1.30
	CORE		4.00
9	VCC	2 (RTF)	2.60
	PP		9.00
10	VCC1	2 (RTF)	2.60
	CORE		4.00
11	GND4	1	1.30
	PP		3.30
12	IN4	1	1.30
	CORE		3.50
13	GND5	1	1.30
	PP		3.30
14	IN5	1	1.30
	CORE		3.50
15	GND6	1	1.30
	PP		3.30
16	IN6	1	1.30
	CORE		3.50

17	GND7	1	1.30
	PP		3.00
18	BOTTOM	0.5+plating	1.95
	Soldermask		0.6
		Total	**92.3±10%**

PCB Material: IT-968BS SE/IT-968TC SE with H-LVP

Delta L requirements: -0.46 dB/inch at 4GHz for stripline routing
-0.49 dB/inch at 4GHz for micro-stripline routing

9.2.2 Host Side and JBOG Side Bridge Card

Layer#	Description	Copper Weight (oz)	Thickness (mil)
	Soldermask		0.60
1	TOP	0.5+plating	1.95
	PP		2.70
2	GND	1	1.30
	CORE		4.00
3	IN1	1	1.30
	PP		15.00
4	VCC	2 (RTF)	2.60
	CORE		4.00
5	VCC1	2 (RTF)	2.60
	PP		15.00
6	IN2	1	1.30
	CORE		4.00
7	GND1	1	1.30
	PP		2.70
8	BOTTOM	0.5+plating	1.95
	Soldermask		0.6
		Total	**62.9±10%**

PCB Material: IT-170GRA1BS/IT-170GRA1TC

9.2.3 IO Board

Layer#	Description	Copper Weight (oz)	Thickness (mil)
	Soldermask		0.60
1	TOP	0.5+plating	1.95
	PP		3.70
2	VCC	2 (RTF)	2.60
	CORE		45.30
3	GND	2 (RTF)	2.60
	PP		3.70
4	BOTTOM	0.5+plating	1.95
	Soldermask		0.6
		Total	**63±10%**

PCB Material: IT-170GRA1BS/IT-170GRA1TC

10 Power Delivery

10.1 Input Voltage

The nominal input voltage delivered by the power supply is 12.5 VDC nominal at light loading with a range of 11V to 13V. The JBOG shall accept and operate normally with an input voltage tolerance range between 10.8V and 13.2V when all under-voltage related throttling features are disabled in section 10.2. The JBOG under-voltage protection level should be less than 10.1V.

10.2 Hot Swap Controller (HSC) Circuit on 12V

Three HSCs (ADI/ADM1278) are used on the JBOG. HSC circuit provides the following functions:

- Inrush current control when JBOG is inserted and powered up.
- Current limiting protection for over current and short circuit. Over current trip point should be able to set to ~200% of the maximum current the system will be experiencing.
- HSC UV protection shall be set to 10V~10.1V
- SOA protection during MOSFET turning on and off
- HSC fault protection is set to latch off (default) with retry as stuff option
- PMBUS interface to enable BMC the following actions
    - Report input power and log event if it triggers upper critical threshold
    - Report input voltage (up to one decimal point) and log event if it triggers either lower or upper critical threshold
    - Log status event based on hot swap controller's status register

- Use HSC or external circuit to provide fast (<20us) over current sense alert to trigger system throttling; features need to be controlled by BMC GPIO directly. BMC sets it to disable as the default. Before BMC is ready, the hardware POR state is enable.
- Use HSC or external circuit to provide fast (<20us) under-voltage alert to trigger system throttling and CPU fast PROCHOT#. This feature is enabled by default with resistor option to disable.
- Use HSC or external circuit to provide fast (<20us) under-voltage alert to trigger system FAN throttling. This feature is disabled by default with resistor option to enable.
- Use HSC or external circuit to provide HSC timer alert to trigger system throttling before HSC OCP happens
- The voltage drop on HSC current sense resistor should be less or equal to 25mV at full loading
- The power reporting of hot swap controller needs to be better than 2% from 50W to full loading in room temperature

10.3 VRM Design Guideline

For VRM, the vendor should list the current budget for each power rail based on the worst case loading case in all possible operation conditions. General requirements for VR component selection and VR design should meet 150% of this budget, and OCP should set to 200% of this budget. Vendors should do design check, inform purchasers about the actual OCP setting chosen for VRM and explain the reason if it cannot meet this general requirement above.

For VRM that requires firmware or power code or a configuration file, vendors should maintain version control to track all the releases and changes between each version, and provide a method to retrieve versions through application software during system run time. This software method should run under 64-bit CentOS version 6.4 or above with an updated Kernel specified by the customer.

All switching VRs should reserve testing hooks for Bode plot measurement.

10.4 System VRM Efficiency

The vendors shall supply high efficiency VRMs for all other voltage regulators over 20W not defined in this specification. All other voltage regulation modules shall be 91% efficiency over the 30% to 90% load range. Vendors are encouraged to deliver systems with higher efficiencies. If higher efficiencies are available at additional cost vendors shall present those options.

10.5 System Power On

JBOG is a system that will power up completely with all rails turned on after AC on/off.
Only when JBOG is set to ALWAYS OFF or LAST POWER STATE, is the system allowed to keep power off after AC on/off event.

It should be noted that the power button at the chassis shall have complete control over system power to turn DC on, even when there is a bad BMC chip.

11 Environmental and Regulations

11.1 Environmental Requirements

The JBOG 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: 1000m (3300 feet)

System would be deployed into datacenter with following environment.
Site 1:

- Temperature: 65F to 85F
- Humidity: 30% to 85%
- Altitude: 1000m (3300 feet)

Site 2:

- Temperature: 65F to 85F
- Humidity: 30% to 85%
- Altitude: 300m (1000 feet)

11.2 Vibration & Shock

The JBOG shall meet shock and vibration requirements according to the following IEC specifications: IEC78-2-(*) & IEC721-3-(*) Standard & Levels; the testing requirements are listed in Table 11-1. The JBOG shall exhibit fully compliance to the specification without any electrical discontinuities during the vibration and shock tests. No physical damage or limitation of functional capabilities (as defined in this specification) shall occur to the JBOG during the non-operational vibration and shock tests.

Table 11-1 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

11.3 Regulations

The vendor must provide CB reports of the JBOG and tray in component level. These documents are needed to have rack level CE. The sled should be compliant with RoHS and WEEE. The JBOG PCB should have a UL 94V-0 certificate. The vendor should design an EMI panel kit and pass FCC Class A.

12 Labels and Markings

The JBOG shall include the following labels on the component side of the JBOG. The labels shall not be placed in a way that may cause them to disrupt the functionality or the air flow path of the JBOG.

Description	Type	Barcode Required?
MAC Address. One per network interface[4]	Adhesive label	Yes
Vendor P/N, S/N, REV (Revision would increment for any approved changes)	Adhesive label	Yes
Vendor Logo, Name & Country of Origin	Silk Screen	No
PCB Vendor Logo, Name	Silk Screen	No
Purchaser P/N	Adhesive label	Yes
Date Code (Industry Standard: WEEK / YEAR)	Adhesive label	Yes
RoHS Compliance	Silk Screen	No
WEEE symbol:  The JBOG will have the crossed out wheeled bin symbol to indicate that it will be taken back by the Manufacturer for recycle 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.	Silk Screen	No
CE Marking	Silkscreen	No
UL Marking	Silkscreen	No
Vendor Asset Tag[5]	Adhesive label	Yes

[4] MAC label for LOM is on JBOG; MAC label for NIC is on NIC.

[5] Work with purchaser to determine proper placement (if an asset tag is necessary)

13 Prescribed Materials

13.1 Disallowed Components

The following components shall not be used in the design of the JBOG.

- Components disallowed by the European Union's Restriction of Hazardous Substances Directive (RoHS)
- Trimmers and/or Potentiometers
- Dip Switches

13.2 Capacitors & Inductors

The following limitations shall be applied to the use of capacitors.

- Only Aluminum Organic Polymer Capacitors shall all be used they must be rated 105C, and shall be selected only from Japanese Manufacturers.
- All capacitors will have a predicted life of at least 50,000 hours at 45C inlet air temperature, under worst conditions.
- Tantalum capacitor using manganese dioxide cathode is not allowed.
- SMT Ceramic Capacitors with case size > 1206 are not preferred. Vendor shall discuss with Facebook before using MLCC > 1206 case by case. Size 1206 still allowed when installed far from PCB edge, and with a correct orientation that minimizes risks of cracks.
- X7R Ceramics material shall be used for SMT capacitors by default. COG or NP0 type should be used in critical portions of the design. X6S can be used in CPU Cage area. Vendor shall discuss with Facebook before using X5R with evaluation of worst case temperature of the location.
- Only SMT inductors may be used. The use of through-hole inductors is disallowed.

13.3 Component De-rating

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

14 Reliability and Quality

14.1 Specification Compliance

Vendors must ensure that the JBOG meets these specifications as a stand-alone unit and while functioning in a complete server system. The vendor is ultimately responsible for assuring that the production JBOGs conform to this specification with no deviations. The vendor shall exceed the quality standards demonstrated during the pilot build (PVT) while the JBOG is in mass production. The customer must be notified if any changes are made which may impact product quality.

14.2 Change Orders

Vendors must notify the customer any time a change is made to the JBOG. A Specification Compliance Matrix will be submitted to the customer for each revision of the JBOG including prototype samples.

14.3 Failure Analysis

Vendors shall perform failure analysis on defective units, which are returned to the vendor. Feedback shall be provided to the customer with a Corrective Action plan within two weeks from the date when the units were received at the vendor's facility.

14.4 Warranty

The Vendor shall warrant the JBOG against defects and workmanship for a period of two years from the date of initial deployment at the customer's facility. The warranty is fully transferable to any end user.

14.5 MTBF Requirements

The JBOG shall have a minimum calculated MTBF of 300K hours at 90% confidence level at 45°C ambient temperature. The JBOG shall also demonstrate the MTBF requirement above by running at full load and 50% of time and performing an AC cycling test 50% of time at 45°C. Typical alternation period is 1 week for the stress test and one week for the AC cycling test. This MTBF demonstration shall finish prior to First Customer Shipment (Pilots samples, Mass Production units).

The JBOG shall have a minimum Service Life of 5 years (24 Hours / day, Full Load, at 45C ambient temperature).

Vendors shall provide a calculated MTBF number based on expected component life.

14.6 Quality Control

Below is a list of manufacturing requirements to ensure ongoing product quality:

- Incoming product must have less than 0.1% rejections
- Cpk values will exceed 1.33 (Pilot Build & Production)

- Vendors will implement a quality control procedure during Production, by sampling JBOGs at random from the production line and running full test to prove ongoing compliance to the requirements. This process shall be documented and submitted prior to Production. The relative reports shall be submitted on an ongoing basis.
- Vendors will conduct an ongoing burn-in procedure for use in Production (Production will not start without an agreement on some sort of burn-in procedure). Vendors shall submit documentation detailing the burn in procedure.

14.7 Change Authorization and Revision Control

After the JBOG is released to mass production, no design changes, AVL changes, manufacturing process or materials changes are allowed without prior written authorization from customer. The AVL (Approved Vendor List) is defined by the list of components specified in the BOM (Bill of Materials).

Any request for changes must be submitted to the customer with proper documentation showing details of the changes, and reason for the changes. This includes changes affecting form, fit, function, safety, or serviceability of the product. Major changes in the product (or in the manufacturing process) will require re-qualification and/or re-certification to the product. A new set of First Article Samples may be required to complete the Engineering Change Order (ECO) process. Any modifications after approval shall phase-in during production without causing any delays or shift of the current production schedule. Vendors shall provide enough advance notice to the customer to prevent any discontinuation of production.

All changes, beginning with the pilot run must go through a formal ECO process. The revision number (on the JBOG label) will increment accordingly. Revision Control: copies of all ECOs affecting the product will be provided to the customer for approval.

14.8 PCB Tests

The server ODM should arrange an independent third party lab testing on Delta-L, IST, and IPC-6012C for each baseboard and middle plane PCB from every PCB vendor. The server ODM cannot use the PCB vendor for these tests. The server ODM should submit reports for review and approval before a PCB vendor can be used in mass production. The testing lots should be manufactured at the same facility of a PCB vendor with the same process that is planned to be used by mass production.

SET2DIL/Delta-L requires 5x different PCB fabrication lots from the PCB vendor. Environmental shipping, packaging, and handling of this board is vital to test success; overnight shipping direct from the PCB vendor to SET2DIL/Delta-L independent lab is recommended.

IST is done once. It is required to be tested on a board manufactured at the same time as a board that completely passes SET2DIL/Delta-L. (Run SET2DIL/Delta-L, and if it passes then ask the IST lab to run IST on the board they receive.) The IST test profile is 3x cycles to 250°C and up to 1000x cycle to 150°C. Passing criteria is 150x cycles average, and 100x cycles minimum for 35x coupons.

IPC-6012C is done when 2x of the 5x SET2DIL tests pass from a PCB vendor. (Passing at the independent test lab)
The ODM should work with then PCB house to implement IST and SET2DIL/Delta-L coupon to break off panel without increasing unit cost of PCB.

14.9 Secondary Component

Secondary component planning should start from EVT and reach 80% of total number of BOM items in PCBA BOM in EVT. The rest of the secondary component should be included in DVT. It is recommended that PCB is planned with three vendors at EVT. The EVT and DVT build plan should cover all possible combinations of key components of DC to DC VR including output inductor, MOSFETs and driver.

The ODM should provide second source plan and specification comparison before each build stage.

15 Deliverables

15.1 OS Support

The head node that needs to work with JBOG shall support 64-bit CentOS version 6.4 and above with an updated Kernel specified by customer, and pass Red Hat certification tests.

15.2 Accessories

All related accessories, including heat sink, back-plate and CPU socket protectors, should be provided and installed at the vendor's factory. All accessory boards including debug cards and PCIe riser cards, should be provided by the vendor.

15.3 Documentation

The vendor shall supply the following documentation to the customer:

- Projection Action Tracker
- Bug Tracker
- Testing Status Tracker
- Design documents
    - Schematics for EVT, DVT and PVT(Cadence and PDF)
    - Board Layout EVT, DVT and PVT (Cadence and Gerber RS-274)
    - Board Design Support Documents:
        - System Block Diagram
        - Power distribution Diagram
        - Power and Reset Sequence Diagram
        - High Speed Signal Integrity Simulation, especially for DDR4 memory
        - Power Integrity Simulation, for important power rails such as CPU and DDR4 memory
        - SMBUS and JTAG Topology

    - GPIO Table for BMC and PCH
    - Hardware Monitor Topology
    - Clock Topology
    - Error Management Block Diagram
- BIOS Version Plan, Version Tracker, and Specification
- BMC Version Plan, Version Tracker, and Specification
- BMC Sensor Table
- Mechanical 2D Drawings (DXF and PDF)
- Mechanical 3D Model (IGS or STEP, and EASM)
- BOM with MFG Name, MFG P/N, Quantity, Reference Designators, Cost
- BOM in customer's defined format, whose definition is provided in separate file.
- Validation documents
    - Server Hardware Validation Items: Test Plan and Report
    - FAI Test plan and Report
    - VR Test Plan and Report
    - Signal Integrity Test Plan and Report
    - Functional Test Report
    - MTBF Test Plan and Report, including calculation
    - System AVL(CPU, DIMM, PCIe cards, Mezzanine Cards, SSD) Qualification Test Plan and Report
    - Reliability Test Plan and Report
    - De-rating Report (worst conditions)
    - 2nd Source Component Plan and Test Report
    - Thermal Test Plan and Report (with indication of critical de-ratings, if any)
    - Mechanical Test Plan and Report

15.4 Mass Production First Article Samples

Prior to final project release and mass production, the vendor will submit the following samples and documentation:

- All the pertinent documentation described in section 15.3 and any other documents and reports, necessary for customer to release the product to mass production
- Pilot samples that are built in the allocated facility for mass production
- A full Specification Compliance Matrix
- A full Test/Validation Report
- Production line final Test 'PASS' tickets
- Samples that have passed the production burn-in process
- Samples shipped using the approved-for-production shipping box described in section 16.

16 Shipping

The JBOG shall be shipped using a custom packaging containing multiple JBOGs in each package. The quality of the packing assembly will be such that the JBOG will not get damaged during transportation. The units shall arrive in optimum condition and will be suitable for

immediate use. A Shock Test for the shipping box shall be conducted by the vendor and submitted to the customer for audit and approval.

17 Appendix

17.1 JBOG Debug POST codes.

Error Codes	Description
00	No error
01	No battery detect
02	No TPM detect
03	RTC fail
04	NA
05	HSC 1 alert
06	HSC 2 alert
07	HSC 3 alert
08	Fan 1 not detect
09	Fan 2 not detect
0A	Fan 3 not detect
0B	Fan 4 not detect
0C	NA
0D	No IO cable detect
0E	No SSD01 cable detect
0F	No SSD23 cable detect
10	I2C bus 1 fail *Note2
11	I2C bus 2 fail
12	I2C bus 3 fail
13	I2C bus 4 fail
14	I2C bus 5 fail
15	I2C bus 6 fail
16	I2C bus 7 fail
17	I2C bus 8 fail
18	I2C bus 9 fail
19	I2C bus 10 fail
1A	I2C bus 11 fail
1B	I2C bus 12 fail
1C	I2C bus 13 fail
1D	NA

1E	NA
1F	NA
20	I2C bus 1 recovery *Note3
21	I2C bus 2 recovery
22	I2C bus 3 recovery
23	I2C bus 4 recovery
24	I2C bus 5 recovery
25	I2C bus 6 recovery
26	I2C bus 7 recovery
27	I2C bus 8 recovery
28	I2C bus 9 recovery
29	I2C bus 10 recovery
2A	I2C bus 11 recovery
2B	I2C bus 12 recovery
2C	I2C bus 13 recovery
2D	HSC0 SMBUS FAIL
2E	HSC1 SMBUS FAIL
2F	HSC2 SMBUS FAIL
30	GPGPU0 overheat
31	GPGPU1 overheat
32	GPGPU2 overheat
33	GPGPU3 overheat
34	GPGPU4 overheat
35	GPGPU5 overheat
36	GPGPU6 overheat
37	GPGPU7 overheat
38	NA
39	NA
3A	NA
3B	NA
3C	NA
3D	NA
3E	NA
3F	NA
40	P12V_AUX voltage sensor critical
41	P3V3_AUX voltage sensor critical
42	P5V_AUX voltage sensor critical
43	P12V_1 voltage sensor critical
44	P12V_2 voltage sensor critical

45	P12V_FAN voltage sensor critical
46	P12V_SSD voltage sensor critical
47	P3V3 voltage sensor critical
48	P3V_BAT voltage sensor critical
49	NA
4A	NA
4B	NA
4C	NA
4D	NA
4E	NA
4F	NA
50	PCIe switch PEX0 uncorrectable error
51	PCIe switch PEX1 uncorrectable error
52	PCIe switch PEX2 uncorrectable error
53	PCIe switch PEX3 uncorrectable error
54	PCIe switch PEX0 internal Temp critical
55	PCIe switch PEX1 internal Temp critical
56	PCIe switch PEX2 internal Temp critical
57	NA
58	NA
59	NA
5A	NA
5B	NA
5C	NA
5D	NA
5E	NA
5F	NA
60	SSD0 SMART temp critical
61	SSD1 SMART temp critical
62	SSD2 SMART temp critical
63	SSD3 SMART temp critical
64	SSD0 fault
65	SSD1 fault
66	SSD2 fault
67	SSD3 fault
68	NA
69	NA
6A	NA

6B	NA
6C	NA
6D	NA
6E	NA
6F	NA
70	Inlet Temp sensorUNC (IO board) *Note1
71	Inlet Temp sensorUC (IO board) *Note1
72	PCIE inlet temperature sensor 1 UNC(baseboard)
73	PCIE inlet temperature sensor 1 UC(baseboard)
74	PCIE inlet temperature sensor 2 UNC(baseboard)
75	PCIE inlet temperature sensor 2 UC(baseboard)
76	NA
77	NA
78	NA
79	NA
7A	NA
7B	NA
7C	NA
7D	NA
7E	NA
7F	NA
80	SEL full
81	NA
82	NA
83	NA
84	NA
85	NA
86	NA
87	NA
88	NA
89	NA
8A	NA
8B	NA
8C	NA
8D	NA
8E	NA
8F	NA

90	Fan 1 lower critical
91	Fan 2 lower critical
92	Fan 3 lower critical
93	Fan 4 lower critical
94	NA
95	NA
96	NA
97	NA
98	NA
99	NA
9A	NA
9B	NA
9C	NA
9D	NA
9E	NA
9F	NA
A0	HSC0 Temp UNC
A1	HSC0 Temp UC
A2	HSC0 input voltage UNC
A3	HSC0 input power UNC
A4	HSC0 input power UC
A5	HSC0 output voltage UNC
A6	HSC0 output current UNC
A7	HSC0 output current UC
A8	HSC1 Temp UNC
A9	HSC1 Temp UC
AA	HSC1 input voltage UNC
AB	HSC1 input power UNC
AC	HSC1 input power UC
AD	HSC1 output voltage UNC
AE	HSC1 output current UNC
AF	HSC1 output current UC
B0	HSC2 Temp UNC
B1	HSC2 Temp UC
B2	HSC2 input voltage UNC
B3	HSC2 input power UNC
B4	HSC2 input power UC
B5	HSC2 output voltage UNC

B6	HSC2 output current UNC
B7	HSC2 output current UC
B8	Throttle Alert
B9-FF	NA

17.2 Commonly Used Acronyms

This section provides definitions of acronyms used in the system specifications.

ANSI – American National Standards Institute
BIOS – basic input/output system
BMC – Baseboard Management Controller
CFM – cubic feet per minute (measure of volume flow rate)
CMOS – complementary metal–oxide–semiconductor
DCMI – Data Center Manageability Interface
DDR4 – double data rate type 4
DHCP – dynamic host configuration protocol
DIMM – dual inline memory module
DPC - DIMMs per memory channel
DRAM – dynamic random access memory
ECC – error-correcting code
EEPROM - electrically erasable programmable read-only memory
EMI – electromagnetic interference
FRU – field replaceable unit
FSC – Fan Speed Control
GPIO – general purpose input output
I²C – inter-integrated circuit
IPMI – intelligent plattform management interface
KCS – keyboard controller style
LAN – local area network
LPC – low pin count
LUN – logical unit number
MAC – media access control
MP – mass production
MTBF – mean time between failures
MUX – multiplexer
NA – Not Applicable
NIC – network interface card
OOB – out of band
ORv1 – Open Rack Version One
ORv2 – Open Rack Version Two
OU – Open Compute Rack Unit (48mm)
PCB – printed circuit board
PCIe – peripheral component interconnect express
PCH – platform control hub
POST – power-on self-test
PSU – power supply unit
PWM – pulse-width modulation
PXE – preboot execution environment
QPI – Intel® QuickPath Interconnect
QSFP – Quad small form-factor pluggable
RU – rack unit (1.75”)
RFU – Reserved for Future Use
SAS – serial-attached small computer system interface (SCSI)
SATA – serial AT attachment
SCK – serial clock
SDA – serial data signal
SDR – sensor data record
SFP - small form-factor pluggable
SMBUS – systems management bus
SMBIOS – systems management BIOS
SOL – serial over LAN
SPI – serial peripheral interface
SSD – solid-state drive
SSH – Secure Shell
TDP – thermal design power
TOR – top of rack
TPM – trusted platform module
U – Rack unit
UART – universal asynchronous receiver/transmitter
UEFI – unified extensible firmware interface
UL – Underwriters Laboratories