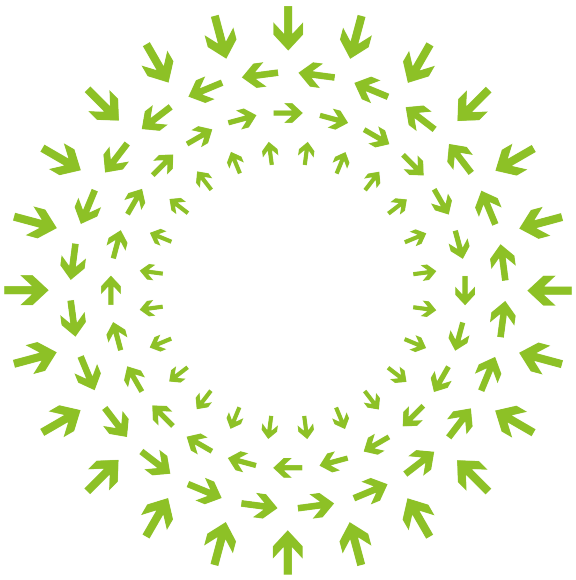

OPEN

Compute Project

Dual M.2 Accelerator Module Hardware Specification

V1.0

Author:
Hao Shen
Pavan Shetty
Yueming Li
Harsha Bojja
Chandra Mallipedi
Chengyu Xu
Ben Wei
Ashwin Narasimha
Chris Petersen

1. License

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

Facebook

Usage of this Specification is governed by the terms and conditions set forth in

Open Compute Project Hardware License – Permissive ("OCPHL Permissive")
Note: The following clarifications, which distinguish technology licensed in the Contribution License and/or Specification License from those technologies merely referenced (but not licensed), were accepted by the Incubation Committee of the OCP:

NOTWITHSTANDING THE FOREGOING LICENSES, THIS SPECIFICATION IS PROVIDED BY OCP "AS IS" AND OCP EXPRESSLY DISCLAIMS ANY WARRANTIES (EXPRESS, IMPLIED, OR OTHERWISE), INCLUDING IMPLIED WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, FITNESS FOR A PARTICULAR PURPOSE, OR TITLE, RELATED TO THE SPECIFICATION. NOTICE IS HEREBY GIVEN, THAT OTHER RIGHTS NOT GRANTED AS SET FORTH ABOVE, INCLUDING WITHOUT LIMITATION, RIGHTS OF THIRD PARTIES WHO DID NOT EXECUTE THE ABOVE LICENSES, MAY BE IMPLICATED BY THE IMPLEMENTATION OF OR COMPLIANCE WITH THIS SPECIFICATION. OCP IS NOT RESPONSIBLE FOR IDENTIFYING RIGHTS FOR WHICH A LICENSE MAY BE REQUIRED IN ORDER TO IMPLEMENT THIS SPECIFICATION. THE ENTIRE RISK AS TO IMPLEMENTING OR OTHERWISE USING THE SPECIFICATION IS ASSUMED BY YOU. IN NO EVENT WILL OCP BE LIABLE TO YOU FOR ANY MONETARY DAMAGES WITH RESPECT TO ANY CLAIMS RELATED TO, OR ARISING OUT OF YOUR USE OF THIS SPECIFICATION, INCLUDING BUT NOT LIMITED TO ANY LIABILITY FOR LOST PROFITS OR ANY CONSEQUENTIAL, INCIDENTAL, INDIRECT, SPECIAL OR PUNITIVE DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND EVEN IF OCP HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

2. Scope

M.2 is a standard format defined by PCI-SIG. In this document we defines an accelerator hardware form factor called Dual M.2 that is derived from M.2 standard form factor. The target use case is for accelerators with high performance and high density.

3. Contents

1. License 2
2. Scope 3
3. Contents 4
4. Revision History 5
5. Overview 6
6. System Block Diagram 7
7. Module Hardware Specification 10
7.1 Pin Definition 10
7.2 Power Specification 14
7.3 IO Description 17
7.4 PCIe Description 18
7.5 FRU specification 20
8. PCB Specification 21
9. Thermal and Heatsink 21
9.1 Thermal Design Guidelines 21
9.2 Integrated Heat Sink Requirements for Dual M.2 22
9.3 Thermal Requirements for Dual M.2 Acceleration Module 23
10. Quality and Reliability 24
10.1 RDT 24
10.2 Compliance 26
11. Prescribed Materials 26
11.1 Disallowed Components 27
11.2 Capacitors and Inductors 27
11.3 Component De-rating 27
12. Labels and Markings 27
12.1 Data required 27
12.2 Data format 28
12.3 Agency Compliance Marks 29

4. Revision History

Ver	Description	Author	Date
0.1	Intial release	Hao	3/12/2019
0.2	5: add preference to use standalone EEPROM in module; update system from YV2 to YV2.50; update the PCIe switch from PM8535 to PM8545. 7.1: Figure 8, change B22 and B28 pin from GND to NC 7.1: Table 2, call out that JTAG pull-ups and isolations should be added on module. call out that UART isolation should be added on module. 7.2.2: Added slew rate and example to read peak power graph. 7.2.4: Add power down handling part in power section 7.3.1: Clarify that alert pin is optional 7.3.2: Request SMBus to provide debug info during the early boot phase 7.4.4: Add maximum non-prefetchable bar size 12.2: Call out minimum label size 5x5mm	Team	12/12/2019
1.0	5. move the change table to chapter 4 7.1 add part number of dual M.2 connector 7.3.4 add USB description 7.5 update FRU content 8 call out latch pad should not be covered by solder	Team	08/11/2020

5. Overview

The high-level hardware specifications are listed below:

Table 1. Hardware Specification Table

Outline	46 x 110mm
Pin Definition	Two sets of M.2 22110 Socket 3 key M Pin Definition with some NC pins redefined.
Board Thickness	0.8mm +- 10%
Board Layer Count	Maximum 12 layers
Component Height	2.0mm on top layer and 1.5mm on bottom layer[1]
Power	20W sustained RMS power, refer to chapter 6.2 for more details.
PCIE[2]	Gen3 or Gen4, 8 lanes and 4 lanes, refer to chapter 6.4 for more details

1. Height keepout just defines the SMT component and does not include the heat sink or thermal materials.
2. PCIE shall be compliant to PCI Express Base Specification Revision 3.1/4.0

The module includes one ASIC, x4 LPDDR4x DRAMs and supporting circuits. Whole system should contain all the circuits within the Dual M.2 standard form factor and should not request any auxiliary circuit on the system mother board to meet the interface spec.

Figure 1 dispalys a recommended placement of the Dual M.2 module. In this case ASIC is located at the center area on top side for the best thermal solution. Vendor can adjust the placement based on the trade off between thermal and routing study.

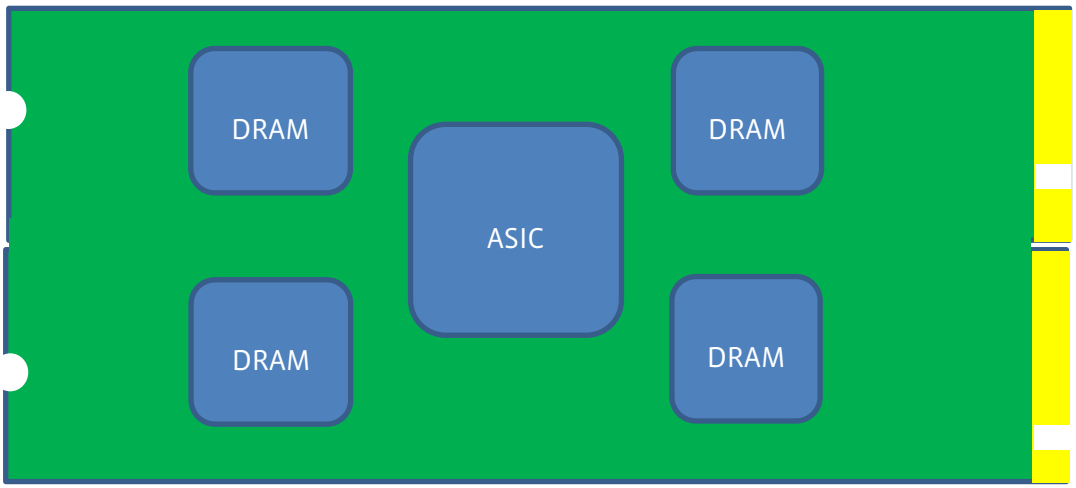

Figure 1: Dual M.2 Module Floor Plan

A module shall include, but not limited to the following components:

1. ASIC to run the workload
2. LPDDR4x Chips

3. Power Circuits
4. SMBus EEPROM or on-chip memory to save FRU information, 8bit SMBus address 0xA6 as defined in NVMe Specification. We prefer standalone EEPROM chip that support at least 400kHz speed. It shall be accessible right way once 3.3V from golden finger is up.
5. Storage for boot firmware if needed but NAND Flash is not allowed.

6. System Block Diagram

In this section we described one use case where Dual M.2 Accelerator modules are plugged into the Glacier Point V2 (GPv2) card. It is a PCIE extension card defined in the Yosemite V2.50 (YV2.50) system. Yosemite V2.50 is a system that is modified from Yosemite V2. For more details you can refer Yosemite V2.50 hardware specification document.

Glacier Point V2 card is the carrier card inserted in the Yosemite V2.50 sled. The following figure shows the carrier card plugged into slots 1 and 3. The carrier card pairs with the twin lake server to form a subsystem. Slot 1 and slot 2 are paird each other. It is the same as slot 3 and slot 4.

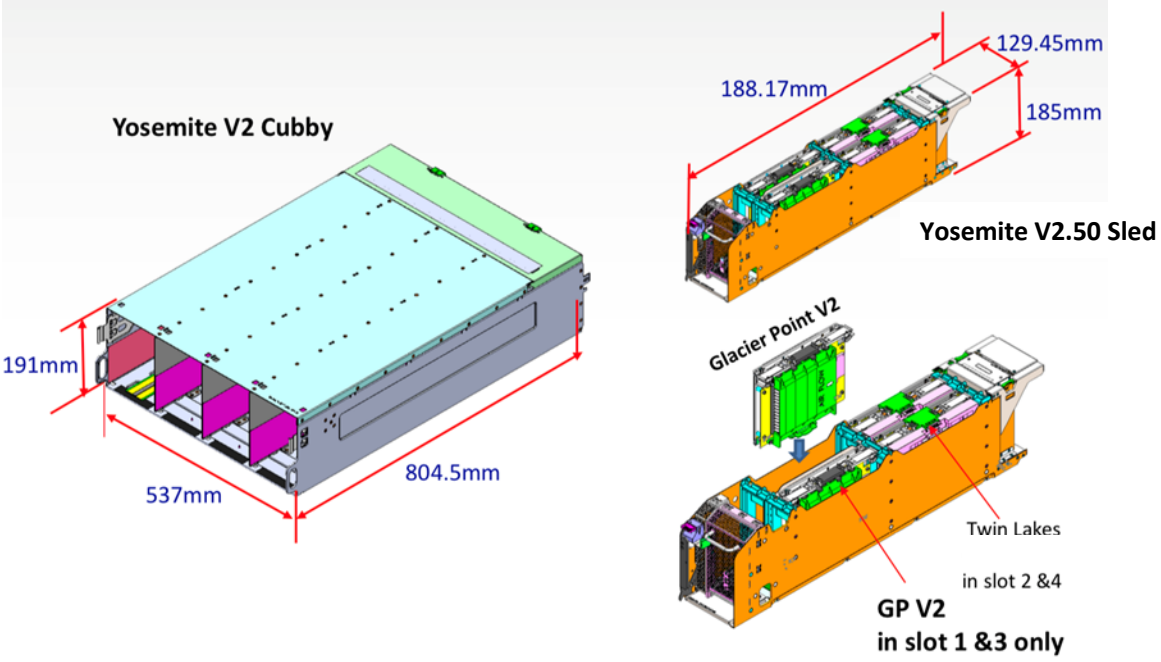

Figure 2: Yosemite V2.50 and Glacier point V2

Previously we designed GPv1 card which is a M.2 carrier card in YV2 system which supports 6x M.2 cards. In GPv2 design, we have made the following changes to better support accelerator workload.

- Support for up to 6x dual-M.2 cards for each GPv2 card
- Add an PCIe Switch to serve fanout function
- Include bridge micro-controller to manage the sideband of each Dual M.2 card
- Include CPLD to mux UART and JTAG of each module to the debug interface
- Add power switch to each Dual M.2 card to allow control software to perform complete power cycle of each cards, when system is in operation mode

PCIe block diagram is listed in Figure 3. There is a Microsemi PM8545 PCIe switch to fanout 16 PCIe lanes from USP to support 6 PCIe Gen3 x8 links at DSP.

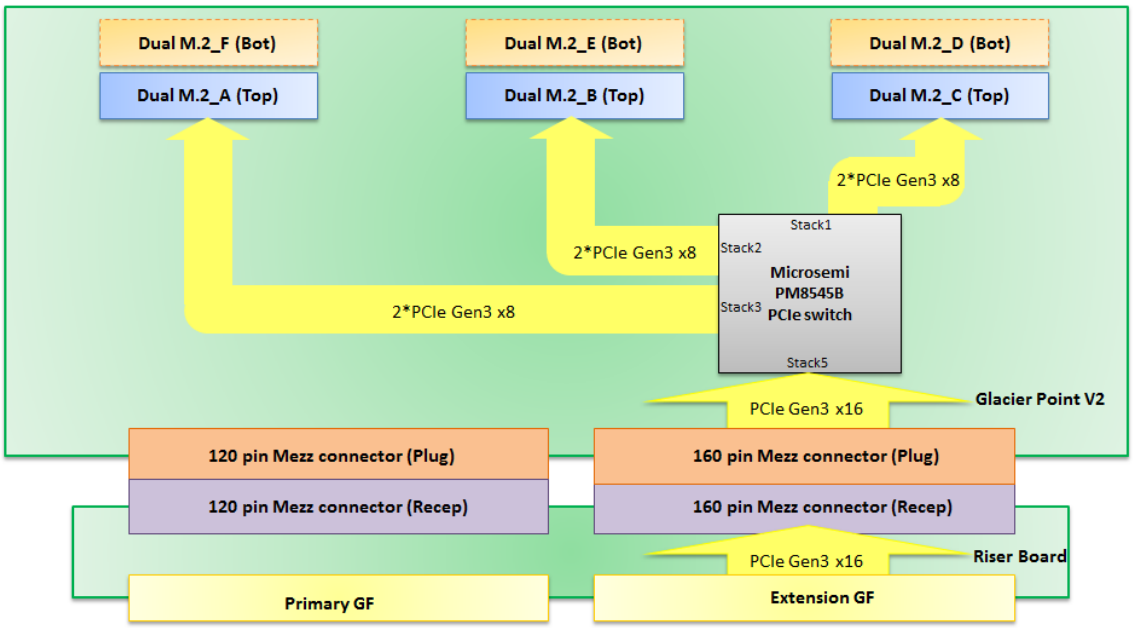

Figure 3. PCIe Block Diagram on GPv2 Card

The block diagram of SMBus topology in GPv2 system is listed below. M.2_A through L represent the M.2 connectors. GPv2 card support both M.2 and Dual M.2 modules. Each Dual M.2 module occupies two adjacent M.2 locations with a Dual M.2 connector. In this graph, SNOWFLAKE represents for bridge IC, which is a micro-controller that manage GPv2 board. INA231 is the voltage and current monitor. PCA9846PW is SMBus Mux.

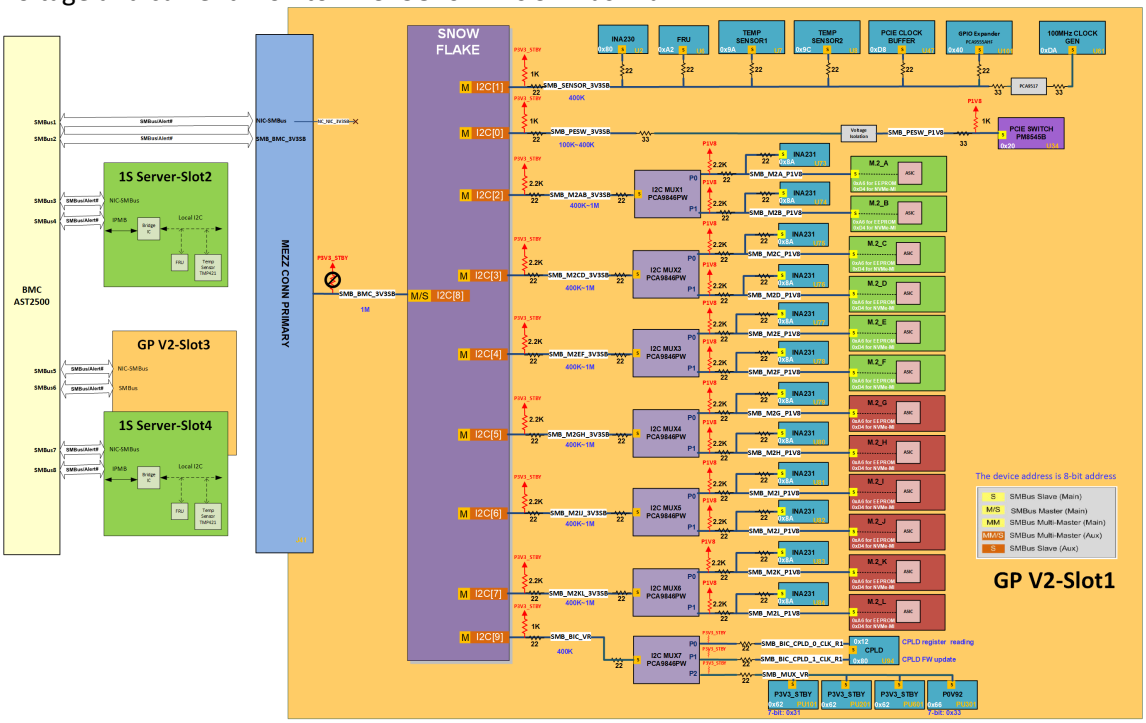

Figure 4: Typical GPv2 SMBus topology

Yosemite V2.50 design has defined the debug port on the front panel to ensure the accessibility. Meanwhile the remote debug capability is a feature that is useful in the fleet so that we can dump the error log once the system fails without the need for operator intervention.

GPv2 is designed with a common clock topology. The host CPU will provide the clock to all Dual M.2 modules through a clock buffer on GPv2 card. PM8545 PCIE switch provides PCIE clock to all M.2 connectors. The dual M.2 module will only use the clock from the primary side.

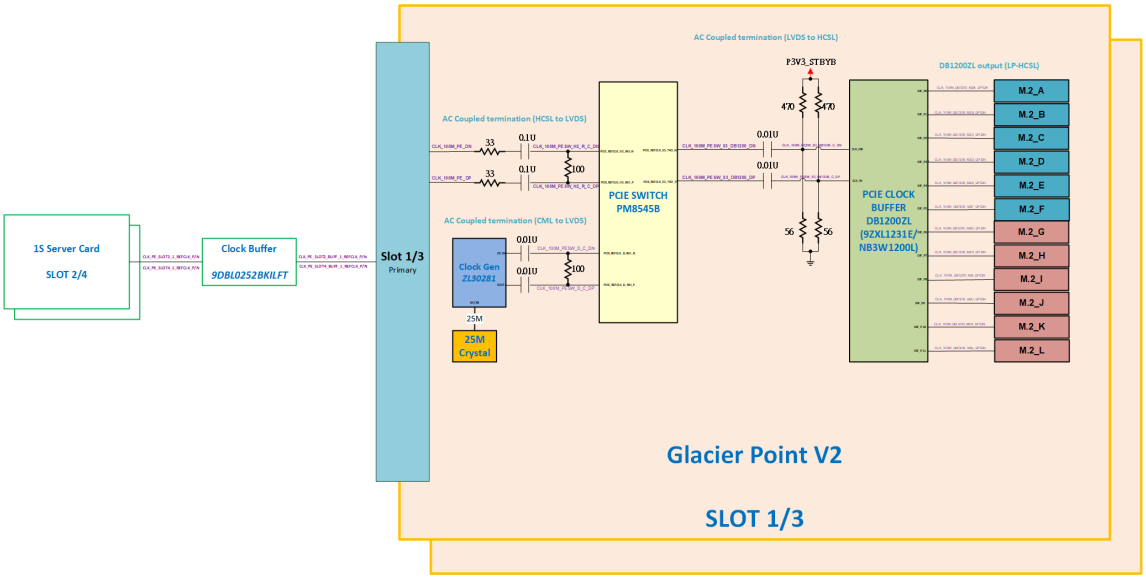

Figure 5: PCIE clock tree

JTAG and UART interfaces are muxed to the system through a CPLD chip. UART ports are connected to both BMC chip and front end debug port on YV2.50 baseboard. The block diagram is listed below:

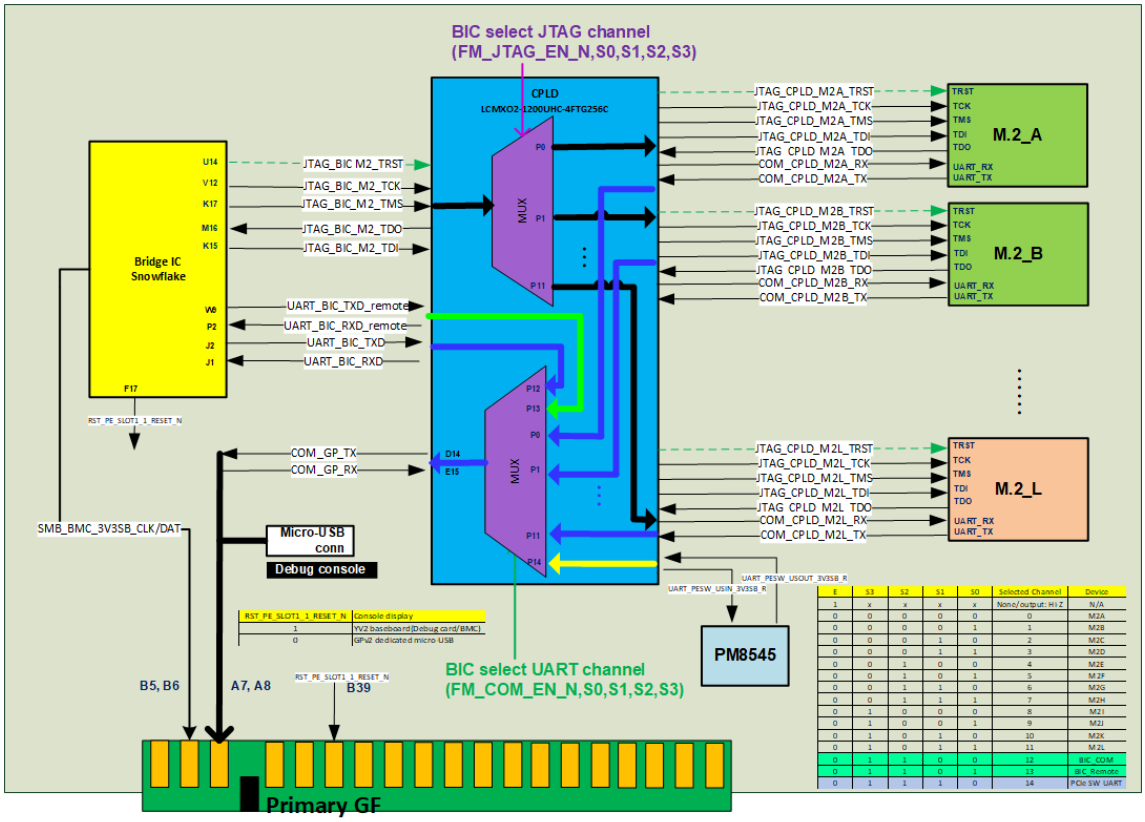

Figure 6: JTAG and UART block on GPv2 board

7. Module Hardware Specification

7.1 Pin Definition

Dual M.2 form factor is simply the combination of two sets of M.2 22110 Socket 3 key M Pin Definition. While the main slot has the full set and the 2nd slot just provide additional power and ground pins. A Dual M.2 connector is designed to support this form factor (Amphenol-Part No: CMDT670M0X005). This connector can support both Dual M.2 and Standard M.2 hardware at the same time. Each Dual M.2 module support 8x PCIe lanes.

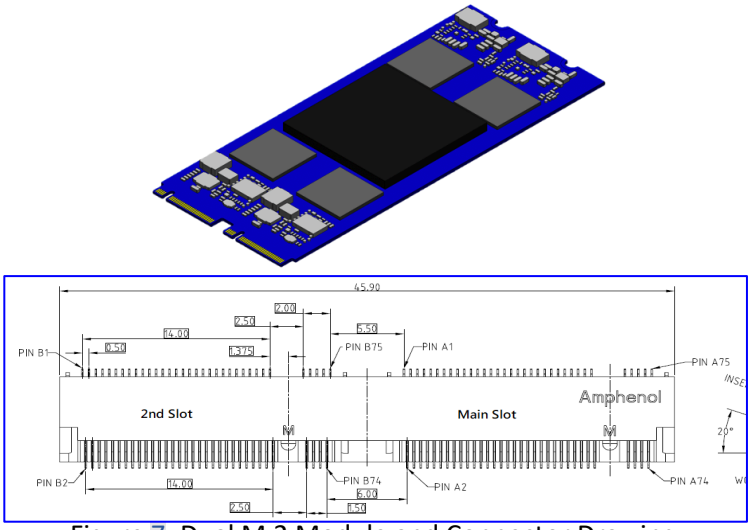

Figure 7: Dual M.2 Module and Connector Drawing

Dual M.2							
Main Slot (SlotA)				2nd Slot (SlotB)			
Signal	Pin	Pin	Signal	Signal	Pin	Pin	Signal
3.3V	A2	A1	GND	3.3V	B2	B1	GND
3.3V	A4	A3	GND	3.3V	B4	B3	GND
PWRDIS	A6	A5	PETn3	NC	B6	B5	PETn7
PLN#	A8	A7	PETp3	NC	B8	B7	PETp7
LED_1#(O)	A10	A9	GND	NC	B10	B9	GND
3.3V	A12	A11	PERn3	3.3V	B12	B11	PERn7
3.3V	A14	A13	PERp3	3.3V	B14	B13	PERp7
3.3V	A16	A15	GND	3.3V	B16	B15	GND
3.3V	A18	A17	PETn2	3.3V	B18	B17	PETn6
TRST	A20	A19	PETp2	NC	B20	B19	PETp6
VIO_1V8	A22	A21	GND	NC	B22	B21	GND
TDI	A24	A23	PERn2	NC	B24	B23	PERn6
TDO	A26	A25	PERp2	NC	B26	B25	PERp6
TCK	A28	A27	GND	NC	B28	B27	GND
PLA_S3#	A30	A29	PETn1	NC	B30	B29	PETn5
GND	A32	A31	PETp1	NC	B32	B31	PETp5
USB_D+	A34	A33	GND	NC	B34	B33	GND
USB_D-	A36	A35	PERn1	NC	B36	B35	PERn5
GND	A38	A37	PERp1	NC	B38	B37	PERp5
SMB CLK (I/O)(0/1.8V)	A40	A39	GND	NC	B40	B39	GND
SMB DATA (I/O)(0/1.8V)	A42	A41	PETn0	NC	B42	B41	PETn4
ALERT# (O)(0/1.8V)	A44	A43	PETp0	NC	B44	B43	PETp4
Reserved_UART_Rx	A46	A45	GND	NC	B46	B45	GND
Reserved_UART_Tx	A48	A47	PERn0	NC	B48	B47	PERn4
PERST# (I)(0/3.3V)	A50	A49	PERp0	NC	B50	B49	PERp4
CLKREQ# (I/O)(0/3.3V)	A52	A51	GND	NC	B52	B51	GND
PEWAKE#(I/O)(0/3.3V)	A54	A53	REFCLKn	NC	B54	B53	NC
Reserved for MFG DATA	A56	A55	REFCLKp	NC	B56	B55	NC
Reserved for MFG CLOCK	A58	A57	GND	NC	B58	B57	GND
ADD IN CARD KEY M			ADD IN CARD KEY M	ADD IN CARD KEY M			ADD IN CARD KEY M
ADD IN CARD KEY M			ADD IN CARD KEY M	ADD IN CARD KEY M			ADD IN CARD KEY M
ADD IN CARD KEY M			ADD IN CARD KEY M	ADD IN CARD KEY M			ADD IN CARD KEY M
ADD IN CARD KEY M			ADD IN CARD KEY M	ADD IN CARD KEY M			ADD IN CARD KEY M
NC	A68	A67	TMS	NC	B68	B67	NC
3.3V	A70	A69	NC	3.3V	B70	B69	NC
3.3V	A72	A71	GND	3.3V	B72	B71	GND
3.3V	A74	A73	VIO_CFG_GND	3.3V	B74	B73	GND
		A75	GND			B75	GND

Figure 8: Dual M.2 Module Pinout Table

This pinout table and I/O direction is defined from the perspective of the module rather than the baseboard. Pin definition is compatible to PCI-SIG M.2 specification though we have re-defined several NC pins mainly for debug purposes. These features are expected to improve the debug capability once hardware is deployed in large scale data center environment.

Table 2. M.2 Module Pinout description

Interface	Signal Name	I/O	Description	Voltage	Requirement
Power Ground	3.3V(18 pins)	I	3.3V running power source	3.3V	Required
	GND(31 pins)		Ground	0V	Required
	VIO_1V8	I	Reserved 1.8V running power source for future PCI-SIG standard.	1.8V	NC in module, GPv2 platform leave this pin open.
PCIe	PETp0/PETn0	O	PCIe TX/RX Differential signals defined by the PCIe 3.1/4.0 specification. The Tx/Rx are defined on module perspective. PET is Tx on module and connects to Rx on host. PER is Rx on module and connects to Tx on host		Required
	PETp1/PETn1	O			
	PETp2/PETn2	O			
	PETp3/PETn3	O			
	PETp4/PETn4	O			
	PETp5/PETn5	O			
	PETp6/PETn6	O			
	PETp7/PETn7	O			
	PERp0/PERn0	I			
	PERp1/PERn1	I			
	PERp2/PERn2	I			
	PERp3/PERn3	I			
	PERp4/PERn4	I			
	PERp5/PERn5	I			
	PERp6/PERn6	I			
	PERp7/PERn7	I			
	REFCLKp/REFCLKn	I	PCIe Reference Clock signals (100 MHz) defined by the PCIe 3.1/4.0 specification		Required
	PERST#	I	PE-Reset is a functional reset to the card as defined by the PCI Express CEM Rev3.0	3.3V	Required
	CLKREQ#	I/O	Clock Request is a reference clock request signal as defined by the PCIe Mini CEM specification; Open Drain with pull up on	3.3V	Optional. GPv2 platform leave this pin open.

			Platform; Active Low; Also used by L1 PM Substates.		
	PEWAKE#	I/O	Vendor do not need to support this feature	3.3V	Optional. GPv2 platform leave this pin open.
Specific Signals	Reserved for MFG DATA		Manufacturing Data line.		Optional. GPv2 platform leave this pin open.
	Reserved for MFG CLOCK		Manufacturing Clock line.		
	LED1# (O)	O	LED pin	3.3V	Optional. GPv2 platform leave this pin open.
	ALERT#	O	Alert notification to master; Open Drain with pull up on Platform; Active Low.	1.8V	Required. Refer to Sec 7.3 for more details.
	SMB_CLK	I/O	SMBus clock; Open Drain with pull up on Platform, slave on module. no pull up on module	1.8V	Required
	SMB_DATA	I/O	SMBus DATA; Open Drain with pull up on Platform, slave on module. no pull up on module	1.8V	Required
USB	USB_D+	I/O	USB 2.0 bus reserved for future application.	N/A	Optional
	USB_D-	I/O	USB 2.0 bus reserved for future application.	N/A	Optional
UART	Reserved_UART_RX	I	UART Receiver Pin based on module perspective. It shall connect to the Tx pin on the host side. IO isolation shall be added on module side to prevent leakage	1.8V	Required. Please refer section 7.3 for details.
	Reserved_UART_TX	O	UART Transmitter Pin based on module perspective. It shall connect to the Rx pin on the host side.	1.8V	Required. Please refer section 7.3 for details.
JTAG	TDI	I	Refer to JTAG Specification (IEEE 1149.1), Test Access	1.8V	
	TDO	O		1.8V	

	TCK	I	Port and Boundary Scan Architecture for definition. The definition is also based on module perspective. All pull-ups and isolations (if needed) should be implemented on module.	1.8V	Required. Please refer section 7.3 for details.
	TMS	I		1.8V	
	TRST	I		1.8V	
Reserved New IOs	PWRDIS	I	Reserved for power disable pin. High: disable power on module. This pin shall be NC on module.	3.3V	GPv2 platform does not support these features.
	PLN#	I	Reserved for Power Loss notification. NC in module.	3.3V	
	PLA_S3#	O	Reserved for Power loss Assert. NC in module.	3.3V	
	VIO_CFG_GND	O	Reserved for IO configure pin. Connected to ground in module.	0V	

7.2 Power Specification

This section defines the power requirements.

7.2.1 Operating (steady state) Conditions

The M.2 module utilizes a single regulated power rail of 3.3V provided by the platform. The following table specifies the power requirements:

Table 3. Operation Mode Rating

Norminal supply Voltage	3.3V
Supply Voltage Tolerance	+/-5%
ASIC Junction Temperature	7% lower than the lowest throttle temperature, specified at Max TDP operating case (e.g. if the throttle temp is 70C then the operating temp would be 65C)

The module shall support Target performance levels in Table 4 by default while take Lower and Highest performance mode as optional setting. Module will determine operating TDP mode during power on stage from firmware. This is a static power level, it is not changed during run time through software, and defines the maximum sustained power drawn from the system during normal operation. The module components should be designed to support the highest power level that the module can support.

Table 4. Module TDP table

	Lower Perf	Target	Highest Perf
Module TDP* power level	15W	20W	25W
Module Absolute Peak Power allowed (For a 20µs load step transient)	25W	30W	40W
Module Performance (assuming DRAM ECC enabled)	>= 80%	100% of the performance listed in the SOW	>= 120%

*TDP: Thermal Design Power, is the sustained average power that the module dissipates while under continuous heavy workload, 50C local ambient temperature and the ASIC junction temperature defined as above in table 3.

7.2.2 Peak (instantaneous) conditions

Transient/Instantaneous power spikes allowed in operation are defined by the absolute peak power specification in Table4. Absolute peak power allowed can vary based on the duration time of the peak power transient load step. Fig.9 represents the curve that defines the peak power vs. load step duration time based on the GPv2 card design and Yosemite V2.50 platform peak power budget. The peak power pulse shorter than 5µs shall be supported by the capacitor on the module. Fig.9 does not specify for power virus condition, but for peak instantaneous power observed while running application workloads and DRAM ECC is enabled.

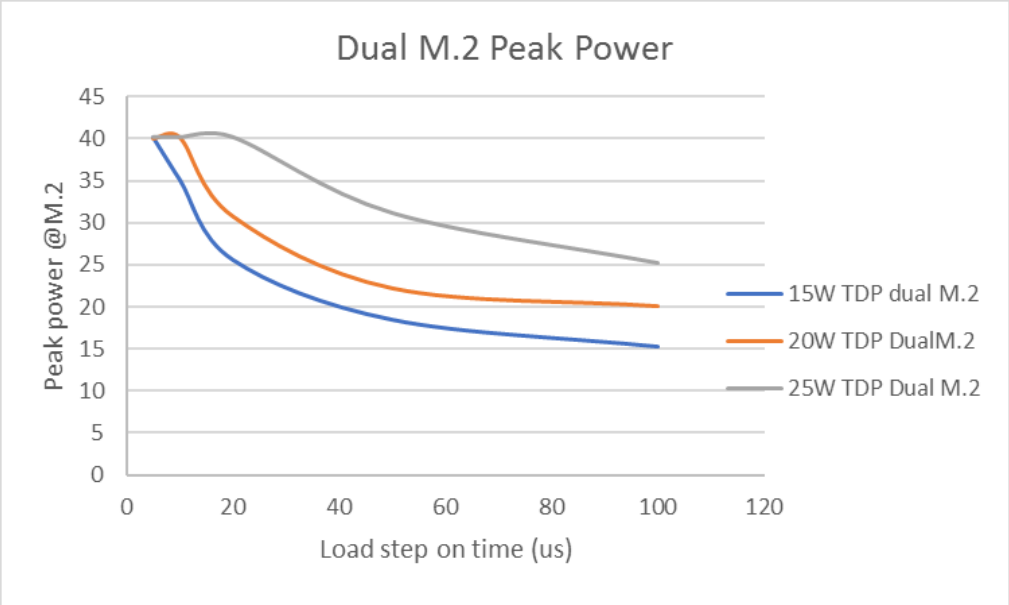

Figure 9: M.2 Peak power specification details across load step

For example if the module is configured to operate at 20W TDP, it is allowed to draw 30W peak power for 20us duration. If the peak power duration is expected to last for 40us, the peak power drawn across this should be limited to 24W.

The peak power here is defined at the 3.3V connector input. The decoupling capacitors and PMIC/VR on the module are expected to suppress transients shorter than 5us to peak power of 40W. If higher transient is expected for these short durations at module input, peak power limiting loop shall be fast enough to limit the instantaneous peak power to less than 40W.

The carrier card is designed to support upto 2A/us slew rate per module on the connector input.

7.2.3 Input Capacitance and undervoltage specification

Input capacitance on the 3.3V connectors should be limited to less than 2mF per dual M.2 module. Device UVLO (minimum voltage for M.2 VR to turn on) should be set to greater than 2.8V to ensure power on ramp time meets PCIE ramp spec and does not stress load switch in addition to capacitive inrush current. Power tree to derive the 3.3V from 12V is described in Figure 10. Power sensors used here are capable of sensing only sustained load currents (>100ms sampling after averaging).

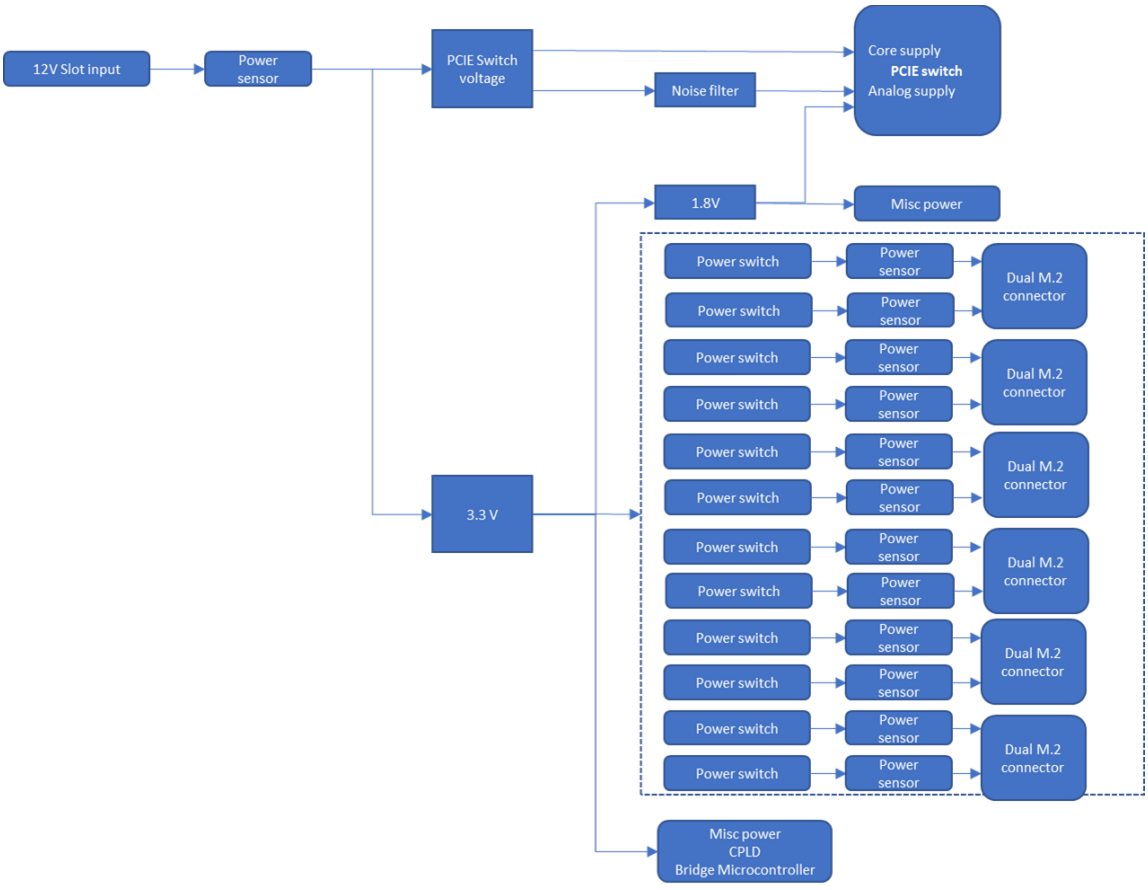

Figure 10: M.2 Power supply configuration on GPv2 carrier card

7.2.4 Power Up/Down Sequence

Module and platform shall follow the standard M.2 power up/down sequence at interface. Module shall manage the power up and power down sequence of SOC, DRAM and other ctirical components by its own. Module shall not expect any special signal from inband or out of band to help manage the power up or power down sequence.

7.3 IO Description

7.3.1 ALERT#

The ALERT# pin connects individually to the Bridge IC as shown in Fig.4. It is optional requirement that ASIC could assert this pin to low once the catastrophic failure happens. Noted that system will not trigger any interrupt event that power off ASIC module based on this pin.

7.3.2 SMBus

Module shall be SMBus/I2C slave to the host with 7 bit address 0x6A which is defined in NVMe base command part. FRU information shall be stored in either an EEPROM on the PCB or in on-chip memory in the ASIC, at 8 bit address 0xA6. This bus shall support 100kHz, 400kHz and 1MHz mode. The ASIC could run at either 400kHz or 1MHz by default depends on system and module design.

ASIC SMBus shall be active at the early boot phase so system BMC can access this interface without needing to wait for driver loading.

7.3.3 UART

We define 1.8V signaling level UART. Baud rate is 57600. Module shall buffer this interface to avoid current leakage once IC is not powered

7.3.4 JTAG

JTAG interface should be compliant to IEEE standard 1149.1. Module shall buffer this interface to avoid current leakage once IC is not powered. All the pull-ups shall be added on Module side.

7.3.5 USB

USB port is an optional debug port in accelerator design. The requirement for USB port is same as M.2 spec. The USB interface supports USB 2.0 in all three modes (Low Speed, Full Speed, and High Speed). Because there is not a separate USB-controlled voltage bus, USB functions implemented on a PCI Express M.2 Adapter are expected to report as self-powered devices. All enumeration, bus protocol, and bus management features for this interface are defined by Universal Serial Bus Specification, Revision 2.0. Module shall isolate this interface to avoid current leakage once IC is not powered.

7.4 PCIe Description

7.4.1 Physical interface

Module PCIe physical interface shall be compliant to PCI-SIG CEM specification 3.0. If the module is capable to run at Gen4 speed, the interface shall be compliant to PCI-SIG CEM Specification 4.0.

Module shall support x8 bifurcation as default status. Module shall also support fail down mode to x4 bifurcation (lane0-3) automatically once plug in to the system where only the primary connector PCIe lanes are connected.

Module shall support PCIe lane and polarity reversal. For x8 case, module should support 0->7 to 7->0 lane reversal. In the fail down (x4) mode, lane reversal is from 0->3 to 3->0.

Module PCIe interface shall support common clock topology with Spread Spectrum Clocking (SSC). SSC's modulation frequency is from 30-33KHz with -0.5%-0% deviation. Separate Reference clock topology support is a preferred but not required.

Module PCIe link loss, including the package loss and trace loss on PCB, shall be less than 5dB at 4GHz for Gen3 case. Link loss for Gen4 case is defined to be less than 7.5dB at 8GHz. PCIe line shall target 85ohm impedance which is the same as platform. Module could be placed very closely to upstream port. So it shall support very short channels as well as long channels.

7.4.2 PCIe power-up timing

The power-up timing of PCIe functions shall follow CEM specification 3.0. Here is the drawing copying from CEM spec:

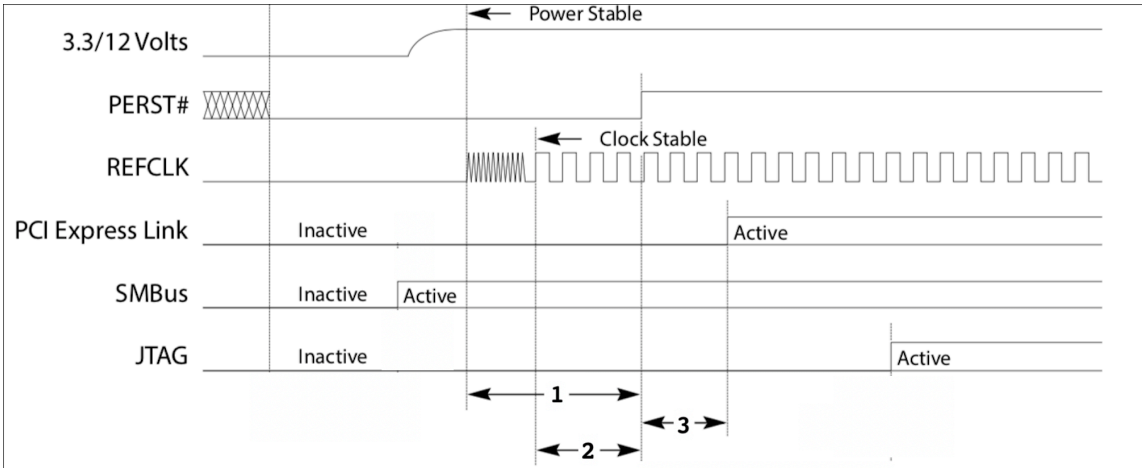

Figure 11: PCIE Power Up Timing

Timing Notes:

1- Minimum time from power rails within specified tolerance to PERST# inactive (T_{PVPERL}): T_{PVPERL} >100ms

2- Minimum clock valid to PERST# inactive($T_{PERST-CLK}$): $T_{PERST-CLK}$ > 100us

3- Minimum PERST# inactive to PCI Express link out of electrical idle:

a. Link LTSSM must enter detect state within 20ms from PERST# being deasserted. This implies that any PHY settings need to be applied to the PCIe PHY before this timing requirement has been exceeded. Caution should be taken here as this may require a boot ROM to execute and/or information to be loaded from SPI flash as well as a reset sequence applied to the logic for it to take effect.

b. Link must be ready for configuration request within 120ms from PERST# being de-asserted.

7.4.3 Reset Mechanisms

The module shall support the following reset mechanisms when the module POWER GOOD indicates that power supplies are stable.

Out of band:

Cold Reset: Signalled by module power good transition from low to high, followed by PCIe PERST# transition from low to high (refer Fig. 11). This is the cold boot scenario for the module and a fundamental reset as defined by the PCIe specification.

Warm Reset: This is signaled by a transition in PERST# without any transition on module POWER GOOD. Warm Reset behavior shall follow PCIe specification.

In Band:

PCIe hot reset : This is signaled in band as per the PCIe specification and the device should respond as per the specification.

PCIe Function level reset : This reset is also signaled in-band and the device should meet all the requirements of the PCIe specification.

In addition, as this module will be used in light out datacenters, we need to have the ability to monitor device health and status using the out of band interface. To meet this requirement the module needs to be able to respond to reads over the OOB interface while it is under all resets (except COLD RESET). The accelerator shall respond in one of two ways:

- Return valid data
- Signal device not ready over the SMBus/i2c protocol

7.4.4 PCIe Configuration

The module shall be configured as a PCIe end-point device. Additionally, the ASIC PCIe controller shall support the following:

1. PCIe class ID shall be set as 0x12
2. A Max Payload Size (MPS) of >= 256 Bytes
3. A Max Read Request Size (MRRS) of >= 512 Bytes
4. At least one BAR shall be pre-fetchable, 64bits, and configurable to be at least 1GB in size
5. Maximum non-prefetchable BAR size shall not exceed 128MB in total
6. The DMA engine shall be capable of saturating at least the PCIe gen4 x4 connection and ideally the full PCIe BW.

7. The DMA engine shall be capable of mapping to all of host memory and the ability to map the majority of the memory on the module with certain memory regions mapped out due to security concerns.
8. The DMA engine shall support a link latency of >= 1us.
9. The DMA engine shall support the ability for software/firmware to enable/disable PCIe MSI/MSI-X interrupts per DMA command and programmatically map the interrupts to either the host or internal CPU cores so that it is possible to chain multiple PCIe commands together.

7.5 FRU specification

Vendor's FRU is stored in an EEPROM or memory area within the ASIC that can be accessed from sideband SMBus line at 8bit address 0xA6. The FRU format should follow *IPMI Platform Management FRU Information Storage Definition 1.0, Version 1.2*. FRU shall support two byte address and FRU content shall start from 0x0000. The FRU template is listed in table 5.

Table 5. FRU Required Fields

Organization	String
Board Info Area	
Language Code	19h (english)
Board Mfg Date	[Generate build time]
Board Mfg	Defined by vendor
Board Product	Defined by vendor
Board Serial Number	Defined by vendor
Board Part Number	Defined by vendor
Fru File ID	Defined by vendor
Custom Field 2	Accelerator Dual M.2
Product Info Area	
Language Code	19h (English)
Product Manufacturer	Defined by vendor
Product Name	Defined by vendor
Part/Model Number	Defined by vendor
Product Version	Defined by vendor
Product Serial Number	Defined by vendor
Product Asset Tag	Defined by vendor
Product Build	EVT (or DVT, PVT)

8. PCB Specification

To maintain the mechanical compliance between the dual M.2 module and connector, a special connector and tighter control on PCB edge finger area is required. Here we disclose the drawing of PCB edge finger area in Fig.12. The A side the primary side and the B side is secondary.

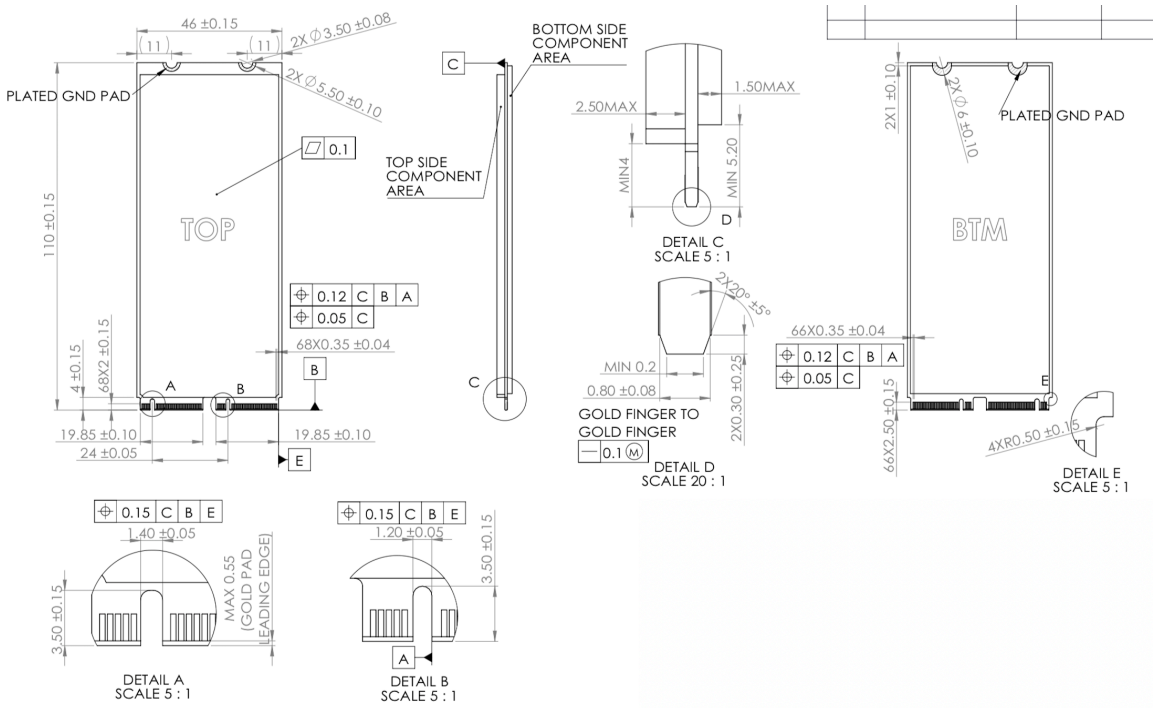

Figure 12: Dual M.2 module PCB Drawing

For the tolerance that has not been called out here, Refer to PCI Express M.2 Specification Revision 1.1 for PCB outline mechanical specification.

HDI type PCB manufacturing is expected here to support high density routing and high density BGA package of ASICs. Vendor could use OSP or ENIG surface finishing on PCB except the golden finger area. Latch pad should be plated. Solder cover is not allowed on latch pad.

9. Thermal and Heatsink

This section defines the thermal and heatsink design guidelines and specifications.

9.1 Thermal Design Guidelines

To improve thermal efficiency, the module shall be fully enclosed by a metal case with a module-level heat sink on the ASIC side. Heat sink dimensions and the associated thermal design requirements need to comply with platform that take the module. Both the module and the heat sink solution for the module shall be provided by the module supplier.

9.2 Integrated Heat Sink Requirements for Dual M.2

This section specifies the dimensions for the integrated heat sink solution for the dual M.2. A reference design is shown in Figure. 13 - Latch material is PA9T. The supplier is encouraged to use their own module design which meets the mechanical dimension requirements. We define a latch design in this spec where the latch actuation must be 2.0±0.22 kfg when the latch is fully pressed with 2.5mm travel distance.

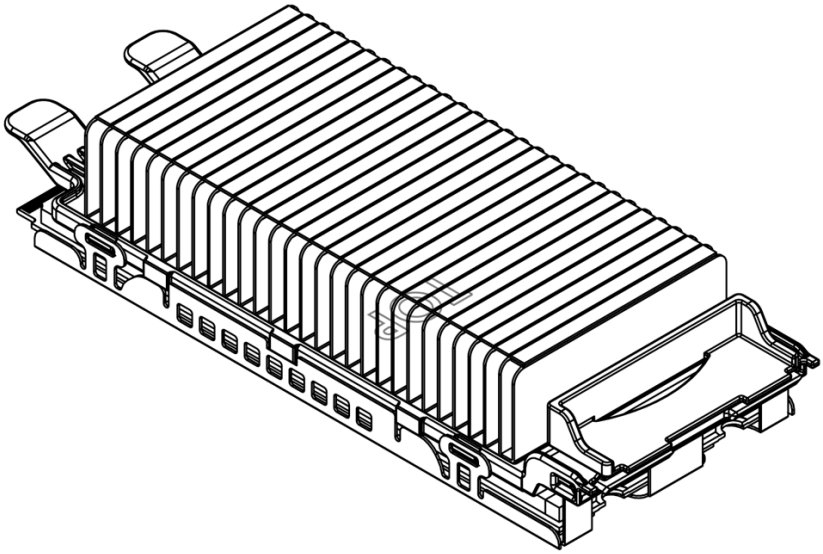

Figure. 13 Reference Design for Dual M.2 Integrated Heatsink

The following list defines the heat sink dimensional requirements for dual M.2 accelerator module are listed as below:

- Nominal height from PCB top surface to heat sink top is 21.4mm, which consists of height of heat sink, TIM (thermal interface material) and SMT components on top side.
- Nominal height from PCB bottom surface to bottom case is 2.4mm, which consists of thickness the metal case, TIM and SMT components on bottom side (if any).
- To provide easy access to connector side and platform integration, the heat sink base of the module shall be die-cast and follow latch design guidance.
- Maximum width of the integrated heatsink shall be kept within 47.4mm for a dual M.2 module.

The heatsink should follow below tolerance requirements to allow easy installation into the chassis as shown in Figure. 14.

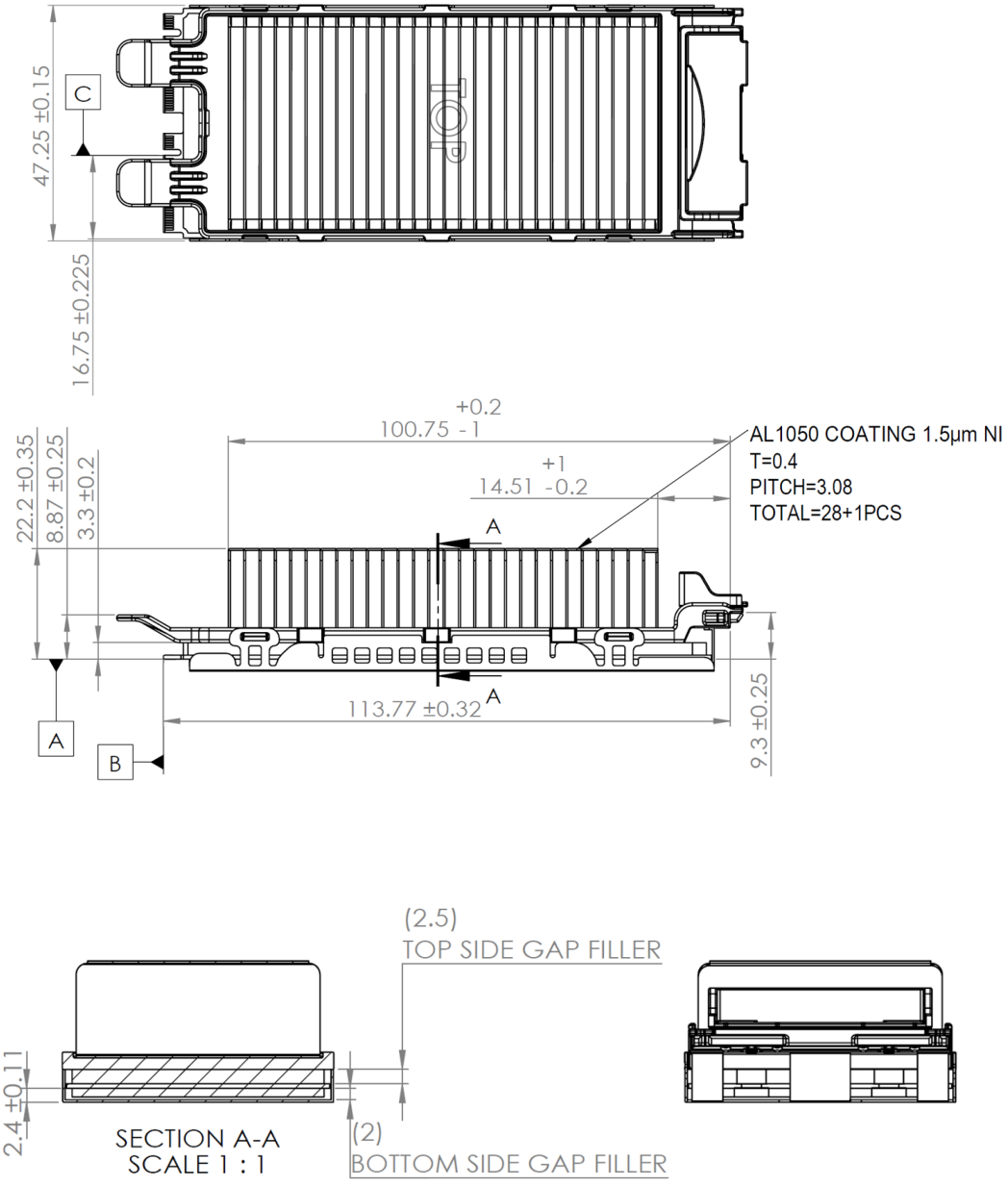

Figure. 14 Dual M.2 Integrated Heatsink Dimensions

9.3 Thermal Requirements for Dual M.2 Acceleration Module

The module shall meet the thermal requirement that as defined by the platform. Specific requirements such as airflow and approach air temperature will be platform specific.

All the temperature values reported by module shall hold at least ±2°C accuracy. ±1°C accuracy is recommended. The ASIC's module-to-module temperature reporting variation shall be ±2°C

with ±1°C recommended. Two different modules shall not report temperature greater than 4 °C apart under the same environmental conditions, slot location, and workload.

10. Quality and Reliability

Module and ASIC are expected to demonstrate Reliabiilty Demonstration Test, Compliance, Robustness, Environmental specifications, and Component qualification. A mix of industry product quality standards (JESD, ASTM, EIA, ISTA etc.) that go above and beyond industry standards are required to be completed and demonstrated as part of product quality reliability, and performance.

10.1 RDT

10.1.1 Quality RDT

Hardware component validation, firmware functionality check, as well as product reliability are the key aspects of Reliability Demonstration Test (RDT). The module is expected to demonstrate MTBF (Mean Time Between Failure) of minimum 2.5 million hours in order to assess product early-life quality and potential failure modes. System integrator is expected to be provided with calculated as well as demonstrated MTBF estimates to include sample size, allowable functional failures (1), stress profile (per JESD218A, JESD219A workloads and temperature conditions) at 60%, and 90% confidence level Weibull modeling accordingly. In-lieu of calculated MTBF availability, component and module level FIT rates are acceptable as well. Planned and Un-planned power loss scenarios are to be accounted as part of MTBF demonstration.

10.1.2 Performance RDT

Module should run work load that can stress the ASIC, DRAM, and PCIE interface. Module vendor would be required to provide the work load for this purpose. Associated test conditions, performance criteria, and duration for this test depend on the project.

10.1.3 Environmental, and Compliance Specifications

Environmental, compliance specifications, and product robustness requirements are listed below so as to ensure that the module/FRU (with Integrated Heat Sink) level requirements are met, and the product functions as expected with no allowable failures post testing.

Table 6. Q&R testing requirement

Module Stress Test	Test Criteria	Standards (As applicable)	Sample Size
Operational Vibration	2.17 G_{rms}, 5-700-5 Hz, all three axes	EIA-364-28	22
Non-Operational Vibration	3.13 G_{rms}, 5-1500-5 Hz, all three axes	EIA-364-28	22
Non-Operational Shock	1250G, 0.5ms, 6 drops, all three axes	EIA-364-27	22

Insertion	300 cycles (plating and power on check every 50 cycles)	EIA-364-09	5
EMC Emission and Immunity		CISPR 22/24, EN55022:2010 +AC:2011, ENTT032:2012 +AC:2013, EN55024:2010 EN 6100-3-2:2014 EN 6100-3-3: 2013 Class B	6
Electrostatic Discharge	± 4kV Contact Discharge ± 8kV Air Discharge	EN55024:2010	6

Module FRU Stress Test (with HIS)	**Test Criteria**	**Standards (Applicable)**	**Sample Size**
Operational Vibration	0.5 G_{rms}, 5-500-5 Hz, all three axes	EIA 364-28	22
Non-Operational Vibration	1.5 G_{rms}, 5-500-5 Hz, all three axes	EIA 364-28	22
Operational Shock	6G, 0.5ms, 6 drops, all three axes	EIA 364-27	22
Non-Operational Shock	70G, 0.5ms, 6 drops, all three axes	MIL-510	22
Package Vibration	1.146 G_{rms}, 2-200-2 Hz, all three axes	ISTA 3E 06-06	10 per tray, 4 trays
Package Drop	8-inch drop	ISTA 3E 06-06	10 per tray, 4 trays
Package Compression	Maximum compression loading on a bulk pack	ASTM D 642-94	1
Thermal Shock **(Non-Operational)**	-40°C to 85°C, 500 cycles (1 Cycle = 5°C to -40°C at ramp 5°C/min, dwell at -40°C for 30 min, -40°C to 85°C at 5°C/min ramp, dwell at 85°C	EIA 364-32	22

	for 30 min, and ramp down to 5°C)		
High Temperature Humidity (Operational)	50°C (local ambient temperature at the module), 90% RH, 500 hours	EIA 364-31	22
Temperature/Voltage Characterization (Operational)	5°C to 50°C (local ambient temperature at the module), Vcc ±x% (per spec), 500 hours		22
Operational Altitude	0 ft to 10000 ft		12
Non-Operational Altitude	0 ft to 30000 ft		12
Power Cycle (AC, DC, Reset)	500 cycles each		22

Note: We recommend using of at least 3 units from the sample size listed above to be subjected to waterfall model reliability testing (using select few tests from above).

10.2 Compliance

North America

- **FCC**: Verification tests only per FCC Part 15 standard. No FCC certification is needed
- **UL:** RU mark is preferred

EU

- **CE mark**: Add CE mark on the accelerator module
- **EMC Directive**: 2014/30/EU- Test partially as applicable
- **ROHS Directive**: 2011/65/EU
- **WEEE mark:** WEEE mark on the accelerator module

APAC

- No specific certification is needed

11. Prescribed Materials

This section defines the required and disallowed components.

11.1 Disallowed Components

The following components shall not be used in the design of PCB board:

- Components disallowed by the European Union's Restriction of Hazardous Substances Directive **RoHS 2 Directive (2011/65/EU)**
- 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 X5R 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

This specification describes label requirements for SSD/Accelerator components used in system. SSD/Accelerator products include PCIE add-in cards, flash drives in 2.5" form factor, flash drives in M.2 form factor and accelerator hardware in M.2/Dual M.2 form factor.

12.1 Data required

- Manufacturer name
- Country of Origin
- Date code of manufacture, includes year & work week
- Product number = same number used to order product from supplier
- Serial number: unique to each product
- Firmware revision
- Hardware revision
- Capacity of card (GB): total data space (system + user), total user space, or user space less OP. (This request is for storage product only)
- PCB Vendor name

We might need to apply TIM (gap pad) material on top of high heat dissipating components such as ASIC, NAND flash and controller on either side of the card. Any standard product labels applied on those parts will be obscured by the TIM, so the supplier should install additional labels that

contain the Serial Number and the Product Number (same PN used for ordering and stored in the SMART data table. The additional labels can use 2D barcodes to save space. 2D bar codes should have human readable text placed in the margin. 2D barcodes should not have any spaces, but dashes are acceptable.

In another case we may have integrated heat sink that encloses module. In this case the label should be attached on bottom of integrated heat sink.

12.2 Data format

- Human-readable. Font size: 10 or larger. Some data on 2D labels can be size 6.
- Barcode. 1D and 2D acceptable. Minimum feature (line width): 10 mils. Minimum bar code size is 5x5mm.
- Electronically readable, e.g. SMART data table.
- Motorola/Zebra readers
    - Motorola CS4070
    - Motorola Symbol DS3578-SR
    - Zebra DS3678-DP

Table 7. Label Requirements

	Requirement		
Data	Human Readable	Barcode	Electronic
Product name	X	X	X
Capacity ()			
Serial Number (Human Readable)	X		X
Serial Number (Barcode)		X	
Sub-assembly No. (Human Readable)			
Sub-assembly number (Barcode)		X	
PCBA Number			
LBA			
Country of Manufacture	X		
Model String	X	Highly Wanted	X
Warranty Disclaimer			
WWN Worldwide Name (human Readable)			
WWN Worldwide Name (Barcode)			
Firmware Version	X		X
Canadian String			

Manufacturer Name	X	X	X
PCB Vendor		X	X
Date code		X	X
PSID Human readable			
PSID Barcode			
Production Date Code	X		X

12.3 Agency Compliance Marks

Module supplier should ensure its products comply with all applicable certificate(s) or verification among the following requirements.

EMC/Safety

- NRTL component level certification
- CE mark based on Directive 2014/35/EC and 2014/30/EU
- FCC verification
- VCCI
- Korea KCC certification
- Taiwan BSMI

Environment regulations

- ROHS – Must be free from hazardous substances prohibited by the RoHS Directives of the EU (European Union)
- China RoHS
- Taiwan RoHS
- WEEE mark
- Halogen Free