Windows Cloud Server Systems

A2040 Celestial Peak Integration Requirements

Author:	Microsoft
Document:	M1139986
Revision:	C
Version:	V.1.0
Date	September 4, 2020

Celestial Peak

Distribution:

□ Internal Only

☒ External All

□ External Restricted through Project Access

RESTRICTED

THIS DOCUMENT AND ANY ASSOCIATED DATA CONTAIN RESTRICTED INFORMATION THAT IS MICROSOFT PROPERTY. DO NOT DISCLOSE TO OR DUPLICATE FOR OTHERS EXCEPT AS AUTHORIZED BY MICROSOFT CORPORATION.

This document is provided "AS-IS" with no warranties. Information and views expressed in this document may change without notice. You bear the risk of using it.

Microsoft and/or third parties may have intellectual property rights covering the subject matter in this document. Except as may be expressly provided in a separate agreement, if any, the furnishing of this document does not grant any licenses to any such intellectual property rights. Instead, this document is for your internal reference purposes. This document and its contents are confidential and proprietary to Microsoft. It is disclosed and can be used only pursuant to a non-disclosure agreement between you and Microsoft.

Microsoft Corporation Technical Documentation License Agreement (Standard)

READ THIS! THIS IS A LEGAL AGREEMENT BETWEEN MICROSOFT CORPORATION ("MICROSOFT") AND THE RECIPIENT OF THESE MATERIALS, WHETHER AN INDIVIDUAL OR AN ENTITY ("YOU"). IF YOU HAVE ACCESSED THIS AGREEMENT IN THE PROCESS OF DOWNLOADING MATERIALS ("MATERIALS") FROM A MICROSOFT WEB SITE, BY CLICKING "I ACCEPT", DOWNLOADING, USING OR PROVIDING FEEDBACK ON THE MATERIALS, YOU AGREE TO THESE TERMS. IF THIS AGREEMENT IS ATTACHED TO MATERIALS, BY ACCESSING, USING OR PROVIDING FEEDBACK ON THE ATTACHED MATERIALS, YOU AGREE TO THESE TERMS.

1. For good and valuable consideration, the receipt and sufficiency of which are acknowledged, You and Microsoft agree as follows:

(a) If You are an authorized representative of the corporation or other entity designated below ("**Company**"), and such Company has executed a Microsoft Corporation Non-Disclosure Agreement that is not limited to a specific subject matter or event ("**Microsoft NDA**"), You represent that You have authority to act on behalf of Company and agree that the Confidential Information, as defined in the Microsoft NDA, is subject to the terms and conditions of the Microsoft NDA and that Company will treat the Confidential Information accordingly;

(b) If You are an individual, and have executed a Microsoft NDA, You agree that the Confidential Information, as defined in the Microsoft NDA, is subject to the terms and conditions of the Microsoft NDA and that You will treat the Confidential Information accordingly; or

(c)If a Microsoft NDA has not been executed, You (if You are an individual), or Company (if You are an authorized representative of Company), as applicable, agrees: (a) to refrain from disclosing or distributing the Confidential Information to any third party for five (5) years from the date of disclosure of the Confidential Information by Microsoft to Company/You; (b) to refrain from reproducing or summarizing the Confidential Information; and (c) to take reasonable security precautions, at least as great as the precautions it takes to protect its own confidential information, but no less than reasonable care, to keep confidential the Confidential Information. You/Company, however, may disclose Confidential Information in accordance with a judicial or other governmental order, provided You/Company either (i) gives Microsoft reasonable notice prior to such disclosure and to allow Microsoft a reasonable opportunity to seek a protective order or equivalent, or (ii) obtains written assurance from the applicable judicial or governmental entity that it will afford the Confidential Information the highest level of protection afforded under applicable law or regulation. Confidential Information shall not include any information, however designated, that: (i) is or subsequently becomes publicly available without Your/Company's breach of any obligation owed to Microsoft; (ii) became known to You/Company prior to Microsoft's disclosure of such information to You/Company pursuant to the terms of this Agreement; (iii) became known to You/Company from a source other than Microsoft other than by the breach of an obligation of confidentiality owed to Microsoft; or (iv) is independently developed by You/Company. For purposes of this paragraph, "Confidential Information" means nonpublic information that Microsoft designates as being confidential or which, under the circumstances surrounding disclosure ought to be treated as confidential by Recipient. "Confidential Information" includes, without limitation, information in tangible or intangible form relating to and/or including released or unreleased Microsoft software or hardware products, the marketing or promotion of any Microsoft product, Microsoft's business policies or practices, and information received from others that Microsoft is obligated to treat as confidential.

2. You may review these Materials only (a) as a reference to assist You in planning and designing Your product, service or technology ("Product") to interface with a Microsoft Product as described in these Materials; and (b) to provide feedback on these Materials to Microsoft. All other rights are retained by Microsoft; this agreement does not give You rights under any Microsoft patents. You may not (i) duplicate any part of these Materials, (ii) remove this agreement or any notices from these Materials, or (iii) give any part of these Materials, or assign or otherwise provide Your rights under this agreement, to anyone else.

3. These Materials may contain preliminary information or inaccuracies, and may not correctly represent any associated Microsoft Product as commercially released. All Materials are provided entirely "AS IS." To the extent permitted by law, MICROSOFT MAKES NO WARRANTY OF ANY KIND, DISCLAIMS ALL EXPRESS, IMPLIED AND STATUTORY WARRANTIES, AND ASSUMES NO LIABILITY TO YOU FOR ANY DAMAGES OF ANY TYPE IN CONNECTION WITH THESE MATERIALS OR ANY INTELLECTUAL PROPERTY IN THEM.

4. If You are an entity and (a) merge into another entity or (b) a controlling ownership interest in You changes, Your right to use these Materials automatically terminates and You must destroy them.

5. You have no obligation to give Microsoft any suggestions, comments or other feedback ("Feedback") relating to these Materials. However, any Feedback you voluntarily provide may be used in Microsoft Products and related specifications or other documentation (collectively, "Microsoft Offerings") which in turn may be relied upon by other third parties to develop their own Products. Accordingly, if You do give Microsoft Feedback on any version of these Materials or the Microsoft Offerings to which they apply, You agree: (a) Microsoft may freely use, reproduce, license, distribute, and otherwise commercialize Your Feedback in any Microsoft Offering; (b) You also grant third parties, without charge, only those patent rights necessary to enable other Products to use or interface with any specific parts of a Microsoft Product that incorporate Your Feedback; and (c) You will not give Microsoft any Feedback (i) that You have reason to believe is subject to any patent, copyright or other intellectual property claim or right of any third party; or (ii) subject to license terms which seek to require any Microsoft Offering incorporating or derived from such Feedback, or other Microsoft intellectual property, to be licensed to or otherwise shared with any third party.

6. Microsoft has no obligation to maintain confidentiality of any Microsoft Offering, but otherwise the confidentiality of Your Feedback, including Your identity as the source of such Feedback, is governed by Your NDA.

7. This agreement is governed by the laws of the State of Washington. Any dispute involving it must be brought in the federal or state superior County, Washington, and You waive any defenses allowing the dispute to be litigated elsewhere. If there is litigation, the losing party must pay reasonable attorneys' fees, costs and other expenses. If any part of this agreement is unenforceable, it will be considered modified to the exte

it enforceable, and the remainder shall continue in effect. This agreement is the entire agreement between You and Microsoft concerning these Materials; it may be changed only by a written document signed by both You and Microsoft.

Revision History

Revision	Version/ Date	Notes	Updated by
	0.10 5/24/2019	Initial Abbreviated version: Electrical, Mechanical and Thermal Requirements only	A2040 Development Team
	0.20 9/12/19	Added System FW, Boot sequence, Validation, Reliability and Regulatory Requirements	A2040 Development Team
	0.30 12/19/19	• Added Microsoft Document Number M1139986 • Updated PCIe Resource Requirements • Added "PCIe MaxReadReq BIOS Setting" to BIOS Requirements • Added detailed A2040 firmware and software instructions from SysEng team.	A2040 Development Team
	0.40	• Added Section for Cerberus provisioning • Updated USB/UART connector PN • Updated "Cabling" section to a single USB/UART cable • Added I2C address table • Added SoC Overview • Removed images/mentions of batch scripts in Section 14 to avoid confusion • Updated USB/UART cabling image in Section 14 • Added BIOS requirements provided in document from BIOS engineer	A2040 Development Team
	1.00 9/4/20	• Review and updates to all sections	A2040 Development Team

Contents

List of Figures

List of Tables

1 Integration Requirements for the Celestial Peak A2040 FPGA Card

This document defines the mandatory minimum set of system compatibility requirements to accommodate Microsoft's Celestial Peak A2040 Card. This doc is intended for Server Development teams and System Integrators.

This document shall be used in conjunction with the respective documents listed here:

MSFT Doc#	Title
M1138893	*Celestial Peak A2040 FPGA Card Specification*
M1042129	*Guideline on Correctable PCIe Error Threshold for Integrators*
M1125048	*Microsoft Cloud Platform Server – Common BMC Specification*
M1128582	*Microsoft Cloud Platform Server – Common UEFI Specification*
M1070680	*Microsoft Cloud - Design for Serviceability Specification*
M1162196	*Cerberus Onboarding Playbook*
M1165900	*BMC to Accelerator SoC Specifications*
No MSFT Doc#; Available at http://pcisig.com	*PCI Express® Card Electromechanical Specification Revision 3.0*

All docs with a MSFT Doc# are accessible via https://eci.microsoft.com/#/Documents/Document/Search

2 Quick Overview: Hardware

The Celestial Peak A2040 card includes an Intel Stratix 10 FPGA, a Cerberus Security chip, and an ARM-based SoC device which communicate with each other and the Host Server. The card employs a PCIe interface for communications with the Host server, but unlike a standard PCIe card, the A2040 has a set of additional cabled connections to the host. **Figure 1** shows the connectivity from the card to other devices. Key differences between A2040 and a standard PCIe card include:

1. Custom 12V power cable between Host and Celestial peak card.
2. Card length is 20mm longer than the half-length, full height PCIe card form factor.
3. Custom UART/USB cable from Host BMC to Celestial Peak.
4. Custom rear card retention for shock and vibration.
5. 100GE Cabled connection between NIC card and Celestial peak
6. Required Auxiliary signals as listed under **Table 8** in section 8.1.1

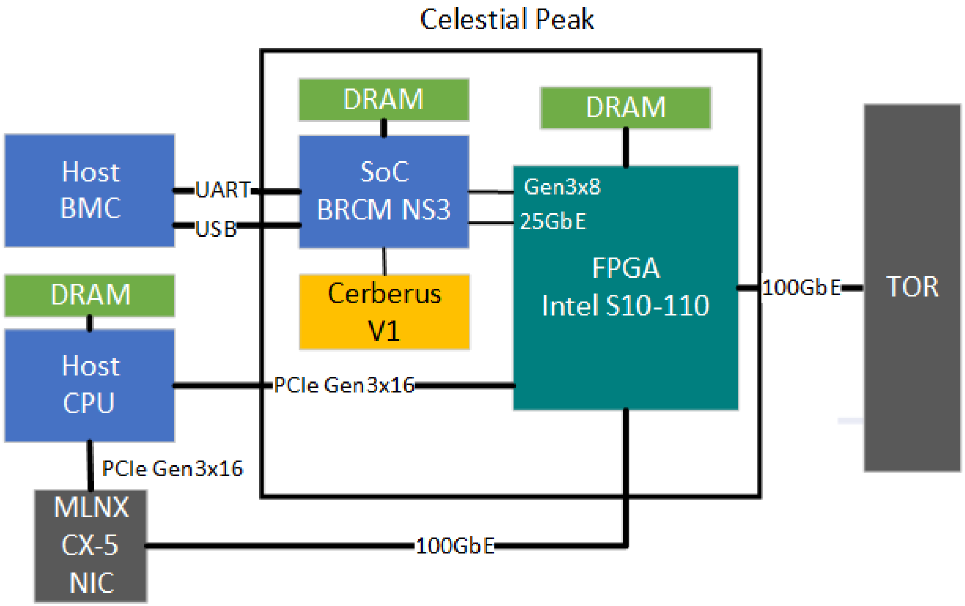


Figure 1: Celestial Peak Connectivity with Host server.

3 Mechanical Requirements

3.1 Card Form Factor

The A2040 is 189.5 mm long (from front surface of IO bracket to the rear surface of the PCB). Additional space is needed for the power cable (mating connector shown in **Figure 2** but not included in the PCA measurement). Height is 111.2 mm. Board thickness complies with PCIe standards for both top and bottom side component height restriction. This is a “single-wide” PCIe card.

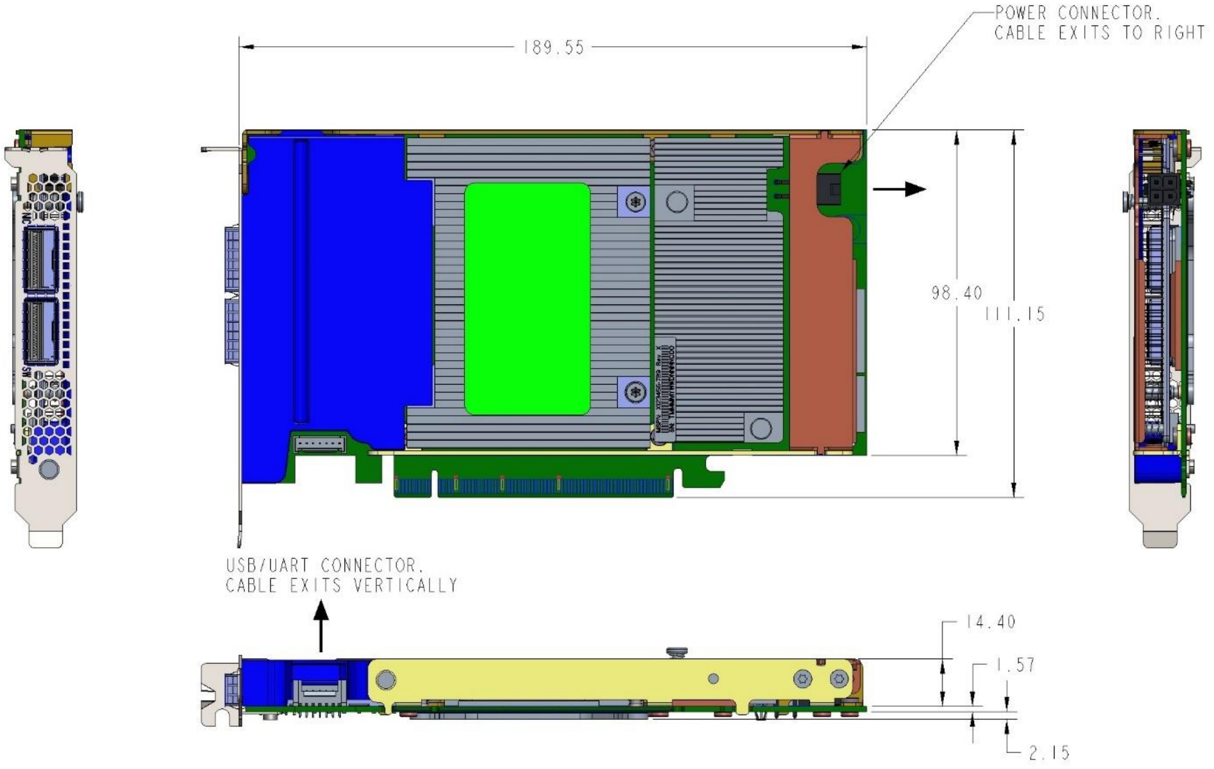


Figure 2: A2040 card dimensions

3.2 Mechanical Stability

Host Servers shall provide adequate mechanical support to stabilize the card during shock and vibration. The A2040 includes two heatsinks, one for the FPGA and one for the SoC, that can act as counterweights during shipping causing the card to flex resulting in solder joint damage. The card design includes a stiffening plate to reduce flex as well as perimeter stiffeners. However, to ensure no damage occurs during both runtime and shock and vibration, server designs shall secure both the front and the rear of the card. System-level FEA modeling is recommended to determine optimal shape and placement of mechanical support. The recommended and tested location for rear card support is shown in **Figure 3** below.

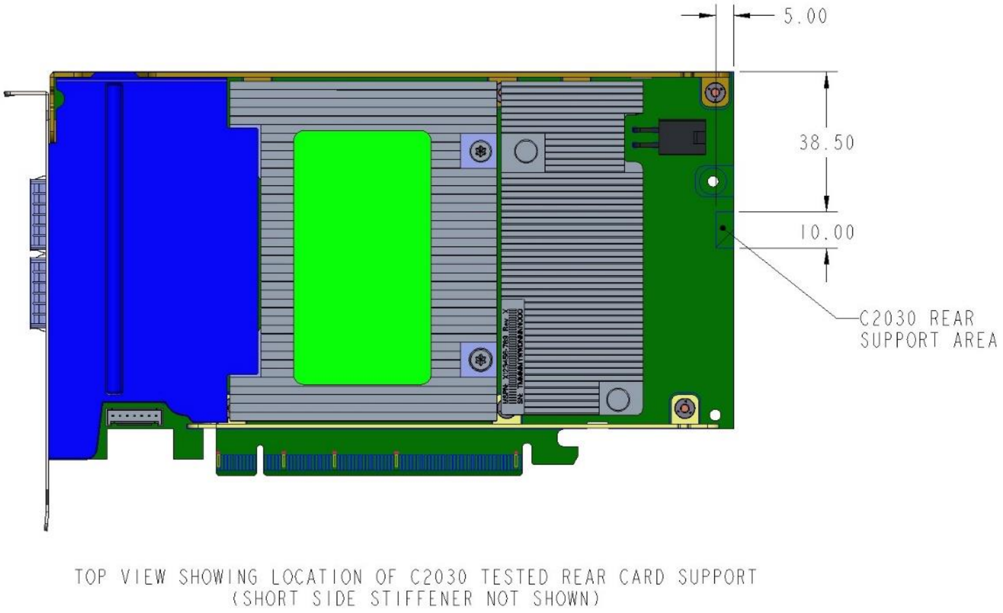

5.00
38.50
10.00
C2030 REAR SUPPORT AREA
TOP VIEW SHOWING LOCATION OF C2030 TESTED REAR CARD SUPPORT
(SHORT SIDE STIFFENER NOT SHOWN)

Figure 3: A2040 location for tested rear card support

4 Cabling

A2040 requires three (3) custom cables described below. Two (2) shall connect the Host Server to the card. The Host Server supplier shall design and qualify cable lengths for these interfaces to accommodate connector location on server relative to A2040 card. The third custom cable connects the A2040 to Mellanox MCX515A-CCAT_C11 NIC card attached to the host.

1. Combined UART and USB cable that connects Host Server BMC to A2040 SoC (cable length determined by blade supplier)
2. 12V Power cable from Host Server main power to A2040 (cable length determined by blade supplier)
3. A custom 0.32 meter, 32AWG QSFP28 cable, **Molex P/N: 1002971321, MSPN: M1132732-001**, connects the A2040 to a separate Mellanox CX5 NIC card

Please consult the Celestial Peak A2040 FPGA Card Specification for all connector pinout details.

4.1 Serial COM Port (UART) and Internal USB

The A2040 includes a 7-pin connector for internal cabling of the USB and UART to the Host server.

Three (3) of the pins provide a serial COM (UART) interface (two signal, one ground) that shall be cabled to the Host Server BMC. This interface provides an out-of-band administrative console to the SoC OS as a serial console session. It is able to provide full-fledged UART redirection via out of band console access for debugging.

Four (4) of the pins provide a USB interface (2 signal, 1 ground, 1 NC/Debug) that shall be cabled to the Host Server BMC. The SoC and Host Server BMC utilize USB2.0 High Speed (480 Mbps). This interface provides backup access to provision the SoC in the event of network or FPGA failure. SoC OS and SoC FW can be provisioned through this USB interface. Additionally, datacenter diagnostics software will rely on this interface for assessing system health.

The connector on the A2040 shall be **Amphenol P/N: 10157575-0730231LF, MSPN: M1163923-001**, or equivalent. The pinout is shown in **Table 2** including both the USB and UART signals. It is recommended that the Host Server side connector use the same connector, but an alternate connector may be selected if necessary.

Manufacturer	Manufacturer Part Number	Microsoft Part Number
Amphenol	10157575-0730231LF	M1163923-001

Table 1: Manufacturer Part Number for UART/USB Connector on A2040

Pin	Signal Name	I/O	Logic	Name/Description
1	No Connect*			No Connect*
2	SOC_USB2_DN	I/O		USB
3	SOC_USB2_DP	I/O		USB
4	GND			Ground
5	SoC UART RX (relative to Celestial Peak)	I/O	3.3V	TX (relative to server/host)
6	SoC UART TX (relative to Celestial Peak)	I/O	3.3V	RX (relative to server/host)
7	GND			Ground

*A2040 supplies 5V to Pin 1. This is ***only*** for debug purposes. Leave this pin floating.

Table 2: Internal USB + Serial COM (UART) Connector Pinout

4.2 12V Power Cabled to Card

The 12V pins on the A2040 PCIe edge fingers are not connected. The 12V input power shall be cabled from the Host Server 12V Main power supply to the 2x2 pin power connector on the A2040. **Table 3**: Manufacturer Part Number for Power Connector on Celestial Peak below provides the manufacturer part number for the power connector. **Figure 4** below shows the pin-out of the 2x2 power connector on Celestial Peak.

Manufacturer	Manufacturer Part Number	Microsoft Part Number
Molex	39-30-1041	M1082431-001
TE Connectivity	1586041-4	

Table 3: Manufacturer Part Number for Power Connector on Celestial Peak

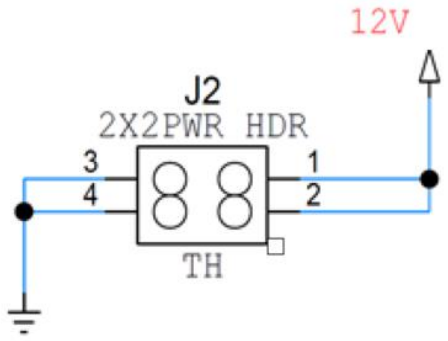


Figure 4: Pinout of 2x2 Power Connector on Celestial Peak

The power connector selected for the Host Server must be capable of meeting the current requirements for the 12V supply to the card. Since the max power enforced by the limit set for the A2040 Hot Swap Controller is 217W (roughly 18 Amps), it is recommended that the cable design and host connector are selected to provide for this worst case condition.

4.3 100GbE Network Cable

The A2040 includes two (2) QSFP28 network ports, each supporting 100Gb Ethernet. One port shall be cabled from the A2040 QSFP1 port to the top-of-rack (TOR) network switch. The other shall be used exclusively to connect the A2040 QSFP0 port to a Mellanox MCX515A-CCAT_C11 NIC card.

The QSFP28 ports on A2040 are not interchangeable. Port mapping details are included in section **8.2 Network Interfaces** within this document.

The A2040 only supports Direct attach copper (DAC) cables for the network interfaces.

Server and Rack configurations shall adhere to the bend radius specifications for Microsoft-approved and Microsoft-qualified DACs.

4.3.1 DAC Cable Requirements from Celestial Peak to Mellanox NIC

The Cable from Celestial peak to Mellanox MCX515A-CCAT_C11 NIC card shall only use cable: **MSPN: M1132732-001** Length: 0.32M Gauge: 32AWG.

4.3.2 DAC Cable requirements from Celestial Peak to Network Switch

Due to the sensitivity of the network interface transceivers on the cards, Server suppliers and Integrators shall use only Microsoft-approved and Microsoft-qualified 100G DACs with Celestial Peak.

5 Power Requirements

The Celestial Peak card accepts conditioned 12V from the host server blade. The card shall derive all required voltages from the 12V source provided by the power cable referenced above in **section 4.2**. The board's over-current protection (OCP) power limit is set higher than the actual expected board power budget and both are summarized in **Table 4** below.

OCP Limits	**Estimated Nominal Operation Power** (full card)	**Server/Host power supply minimum sustained capability**	**Thermal Design Power** (@ 35C & 2.5m/s inlet approach velocity)
167W (min) / 192W (nom) / 217W (max)	105W	130W	130W

Table 4: Card Power Requirements

The nominal **OCP limit** reflects the power level at which the card Hot Swap Controller is intended to cut off its output in order to protect the card from damage. However, the min and max OCP limits reflect the tolerance on this OCP limit where the min is the lowest power at which the HSC output may be disabled and the max is the highest power it may operate to before being disabled. In other words, expect that it may be disabled above the min limit and it will be disabled above the max limit.

Estimated Nominal Operation Power is the total card power draw under typical workload scenarios in the Data Center. However, certain conditions and workloads may result in higher power draw.

Server/Host power supply minimum sustained capability is the power level that the host is required to support continuously under worst case workload conditions for the card. The host system must be designed to provide at least this level of power to the card without any limit on duration.

Thermal Design Power is the maximum power that the system fan control can support under thermal operation boundary conditions defined in **Table 4** and **Table 5**.

5.1 Power monitoring

Power monitoring of total board consumption is available in the Hot Swap Controller via its I2C connection. As mentioned below under system electrical requirements, I2C should be connected to the board I2C resources bus. The blade BMC has access to these I2C devices for monitoring and reporting purposes.

6 Thermal Requirements

6.1 Card Temperature and Server Air flow rates

Server air flow rates shall be adjusted by the BMC based on FPGA and SoC temperature monitoring in combination with the other components within the Host System. System design for PWM and cooling architecture are the responsibilities of the manufacturer. **Table 5** shows the die temperatures that shall govern the BMC controls of system fan speed.

Normal Operational Thermal Conditions					
Component	**Max Operating Overshoot Temperature (C)**	**Throttling Temperature (C)**	**Max Operating Steady State Temperature (C)**	**Through heat sink fin Airflow Target (CFM)**	**Target design power @ airflow rate (W)**
FPGA	88	N/A	84	4.5	75
SoC	97	N/A	93	3.3	30

Table 5: Card Thermal Requirements

The A2040 card also has device thermal thresholds and limits which indicate where system events will be logged or actions will be taken to prevent system malfunction or device damage.

Thermal Thresholds and Limits				
Component	**Temperature warning (SEL)**	**Shutdown / Reset Temp (C)**	**Damage Possible (C)**	**Notes**
FPGA	Upper Non-critical: 89C Upper Critical: 90C Upper Non-Recoverable: 99C	102 Role Reset	120	Data integrity loss potential when card is over 91C
SoC	Upper Non-critical: 98C Upper Critical: 99C Upper Non-Recoverable: 104C	105 OS Shutdown	110	SoC will trigger card malfunction/shutdown at 105C

Table 6: Card Thermal Limits

For clarity, the terminology used in Table 5 and Table 6 associated with normal thermal conditions and thermal thresholds is explained in more detail below in association with **Figure 5**.

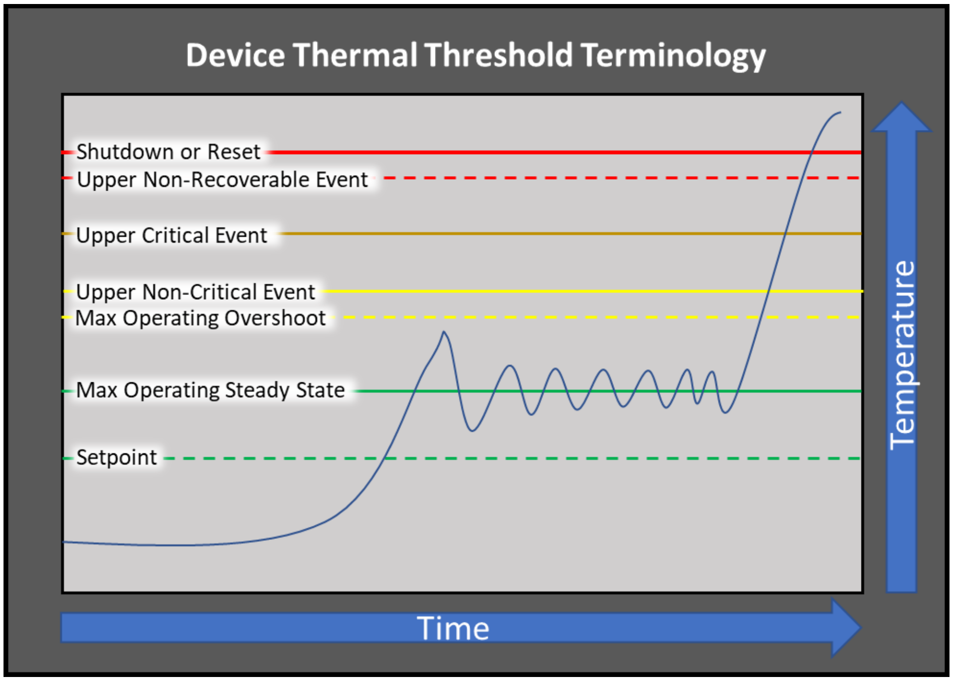


Figure 5: Device Thermal Thresholds

The device **Setpoint** (not shown above) is the first temperature at which the fan control algorithm within the BMC will begin to adjust the system fan speeds in reaction to the device temperature.

Max Operating Steady State is the device temperature that the fan control will target to average around once the Setpoint has been exceeded and assuming no other device in the system is dominating the fan response.

Max Operating Overshoot is the upper limit permitted on the dither of the temperature around the *Max Operating Steady State*. The system fan control must prevent the device temperature from exceeding this level.

The **Upper Non-Critical Event** limit is the device temperature at which the BMC must log an event to the System Event Log (SEL) to indicate that the *Max Operating Overshoot* has been exceeded. It is intended to provide an early warning that the device temperature is at risk of causing issues if it continues to increase.

The **Upper Critical Event** limit corresponds to the temperature at which the device is in danger of malfunctioning. This could translate into bus errors, data corruption, or other device error conditions. It is intended to indicate that the user should consider saving any data and taking the system out of operation for investigation and repair.

The **Upper Non-Recoverable Event** limit is intended to provide a warning that the device is in a thermal runaway state and nearing its Shutdown or Reset temperature. At this stage, data recovery is likely not possible.

The **Shutdown** or **Reset** levels are examples of actions that may be taken to prevent damage to the device by eliminating all or most of the activity which is causing elevated temperatures. These actions will interrupt normal card and system operation in a manner that is likely not recoverable without restarting the system or without restarting certain processes.

6.1.1 Temperature Sensor Calibration and Offset for FPGA Die Temp

The external temperature sensor, TMP461, on the FPGA card needs to be calibrated once upon every power-on to read accurate FPGA die temperature values. Only the remote sensor reading on the TMP461 needs to be calibrated, the ambient temperature (or local sensor) reading from TMP461 does not need calibration.

The BMC must perform the following calibration step after every card power on (12V applied to A2040 card), because the following register write is not non-volatile:

- Write 0xC5 to register 23h
    - This adjusts the n-factor of TMP461. It is changed from its default value (i.e. 1.008) to a calibrated n-factor value (i.e.1.037).
        - System designer must tune this value to meet their system criteria.
    - Every time the board loses its power or goes through a power cycle, register 23h is reset to its default value. Thus, we need to re-write 0xC5 to register 23h after the card is powered on.

Once the calibration is done, an offset of “4 degrees” Celsius must be added to the FPGA die temperature that is read from TMP461’s remote sensor. The temperature value (in degrees C) after adding 4C will be the final/actual FPGA die temperature.

6.2 Temperature and Humidity Conditions

The range of temperatures and humidity levels for both operation and storage/shipping requirements are shown in **Table 7**. Note that the operating relative humidity values are for the inlet to the server and it should be assumed that the highest moisture content of air used for actively cooling the card is equivalent to 80%RH at 35C.

Specification		Requirement
Inlet temperature	Operating	• 18°C to 60°C • Maximum rate of change: 18°F (10°C)/hour
	Non-operating	• -40°C to 75°C • Rate of change less than 36°F (20°C)/hour
Humidity	Operating	• 10%-80% (at a system inlet of 10-37.8C) • Maximum rate of change 20%/hour
	Non-operating	• 5% to 95% non-condensing • 100.4°F (38°C) maximum wet bulb temperature

Table 7: Card Environmental Requirements

7 SoC Overview

The Celestial Peak card uses a Broadcom BCM58732 SoC, see **Figure 6**.

The Broadcom BCM58732 SoC has the following characteristics:

- Octal-Core 64-bit ARMv8 Communication Processor
- Up to 3.0 GHz ARM Cortex-A72 CPU
- 8 CPU cores in 4 clusters of 2 cores, each cluster with 2MB of L2 cache
- Flip Chip BGA (FCBGA), 0.8mm pitch
- 33mm x 33mm, 1579 pin package
- Maximum Junction Temperature is 110°C

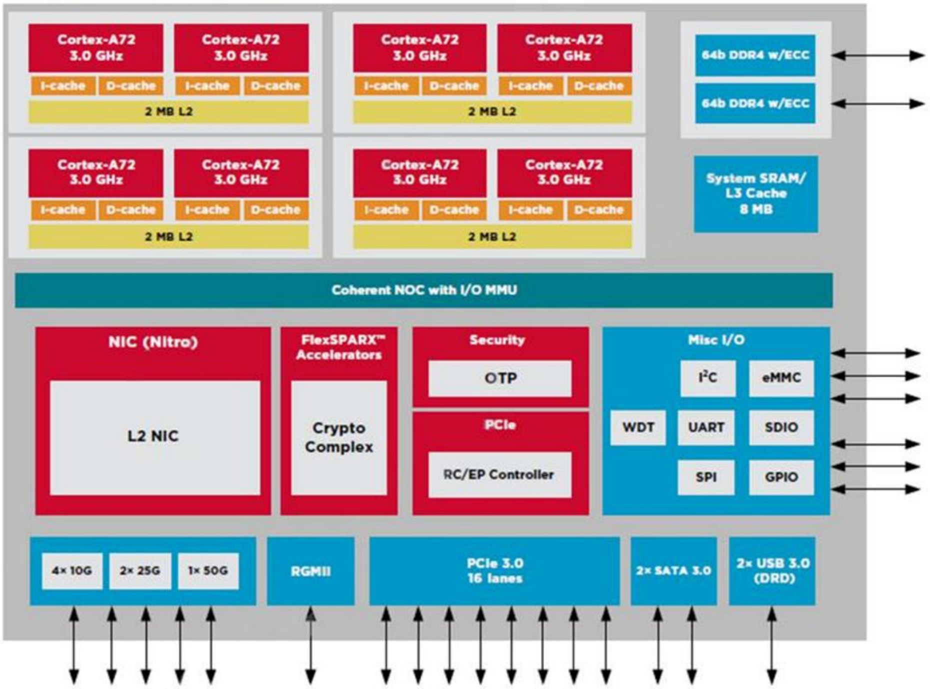


Figure 6: SoC Block Diagram

7.1 SoC Debug

A command line interface for the Celestial Peak SoC device is accessible via the SoC UART. In the standard system configuration for the Data Center, this connection is wired to the Host BMC which is then able to redirect the console to the Rack Manager over the management network. For bench debug where no Rack Manager is available, the serial interface can be connected to another host machine using a debug version of the 7-pin UART/USB cable, plus a serial to USB converter cable.

Similarly, the Celestial Peak SoC USB interface connects to the Host BMC, usually via a USB mux for selecting between SoC and Host USB. In the standard system configuration for the Data Center, this permits the SoC to mount image files stored on the Rack Manager by communicating over the USB interface to the BMC and over the management network from the BMC to the RM. For bench debug where no Rack Manager is available, the USB interface can be connected to a USB thumb drive or other USB storage device using a debug version of the 7-pin UART/USB cable which provides a Type-A USB port.

8 System Electrical Requirements

This section includes detailed requirements for the A2040 electrical interfaces.

8.1 PCI Express Interfaces

The A2040 card includes two (2) PCIe interfaces:

- A PCIe Gen3 x16 primary edge finger PCIe connector interface from the FPGA to the Host Server.
- A PCIe Gen3 x8 interface hard-wired from the FPGA to the on-board ARM® SoC.

The FPGA PCIe endpoints may be re-configured into 1, 2 or 3 physical functions during Server runtime.

8.1.1 PCIe Interface from FPGA to Host Server

The card shall connect to the server through a standard PCIe x16 Gen3 card edge connector. The electrical and mechanical interface must meet the *PCI Express® Card Electromechanical Specification Revision 3.0* requirements.

In addition to the standard PCIe signals defined in the CEM spec including 16 lanes of transmit and receive data, the connector shall include the following auxiliary signals:

FPGA Card	Additional signal enablement on PCIe card edge connector
A2040 - Celestial Peak	• SMCLK (B5) • SMDAT (B6) • 3.3Vaux (B10) • BMC_3V3_PWRBRK_N (B30) • PERST# (A11) • REFCLK+ (A13) • REFCLK- (A14)

Table 8: Additional Signal Enablement Required on PCIe Card Edge Connector

The **BMC_3V3_PWRBRK_N** signal is an input to the A2040 and intended for use as a power throttling signal to the card. It is currently level-translated and connected to the FPGA, but not implemented in the FPGA for the power throttling function. The Host Server design shall provide a 10KOhm pullup resistor to the 3.3V standby rail for this signal to avoid floating. Additionally, the Host Server may connect this signal to a CPLD or other logic device which may be used later to drive the signal for the intended function.

The **SMCLK**, **SMDAT**, **PERST#** and **REFCLK+/-** signals as well as **3.3Vaux** power rail shall be connected to the card and meet the requirements as defined by the PCI Express Card Electromechanical Specification.

The Host system PCIE slot interface shall be programmed by System BIOS for configuration of x16 and 8Gbps gen3 speed. In the event of an issue that affects operation at full width or speed, the Host system must support downtraining to a lower width or speed per the PCIe specification. However, it should be verified in manufacturing and integration testing that the link is running at the max throughput by ensuring that the trained bus width and speed are x16 and 8Gbps.

Server suppliers using Celestial Peak shall demonstrate that the Host Server meets the Intel® guidelines for PCIe margins. For Intel® platforms, this shall include testing and evaluating margins with the Intel® Electrical Validation Test Suite (Intel® EVTS). For other Host Server platforms an equivalent margin evaluation is required.

8.1.2 PCIe Interface from FPGA to ARM® SoC

The internal PCIe Gen3 x8 interface is a chip-to-chip connection between the ARM® SoC and the FPGA. The purpose of this interface is to provide both data transfer/exchange and management capabilities. This is shown in **Figure 7** below.

For management capabilities, the SoC uses this interface to manage and update the FPGA. The ARM® SoC validates images going from its on-board memory to the FPGA and can update the image in the FPGA or the image in the FPGA QSPI NOR flash via PCIe.

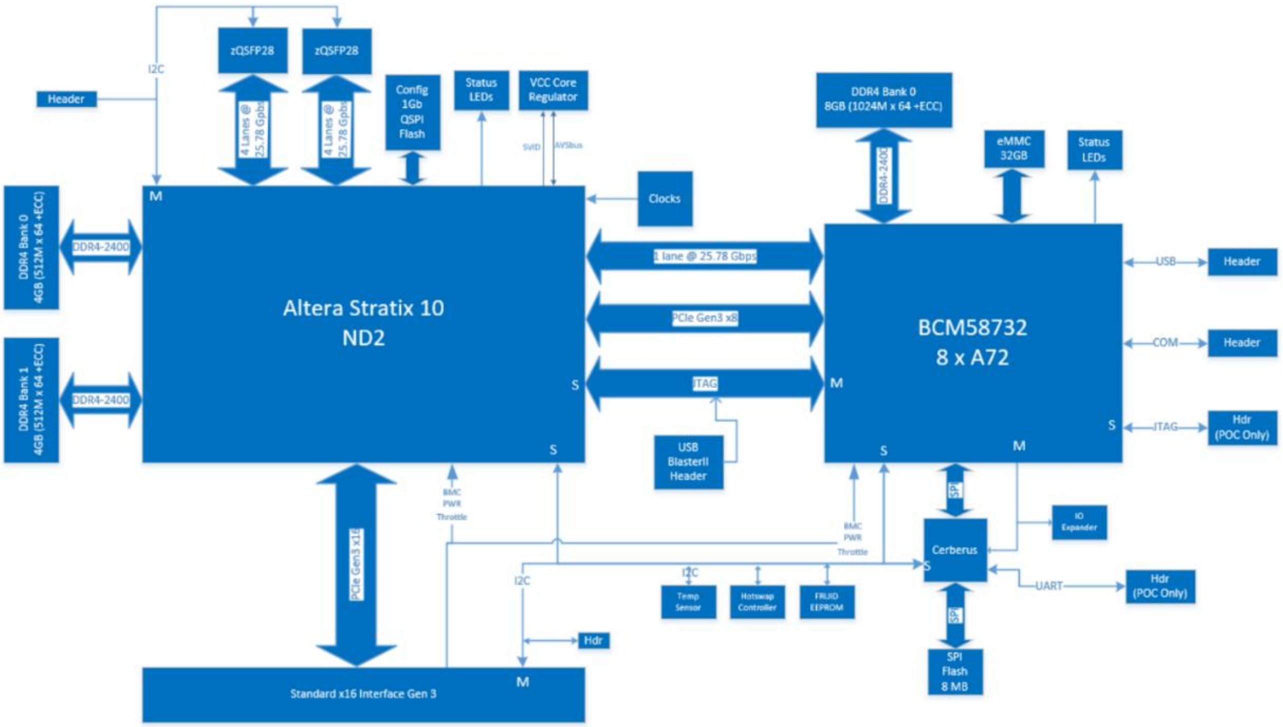


Figure 7: PCIe Interface from FPGA to ARM® SoC

8.2 Network Interfaces

The A2040 has three ethernet network links described in **Table 9** below:

Topology	A2040 ethernet Port	Link Features
Cabled from FPGA to a Mellanox MCX515A-CCAT_C11 NIC card	Port 0 = "NIC" (external)	• 100Gb Ethernet • No FEC • AN/LT not supported
Cabled from FPGA to the top-of-rack (TOR) network switch	Port 1 = "SW" (external)	• 100Gb Ethernet • No FEC • AN/LT supported
Hard-wired between FPGA and on-board SoC	Internal	• 25GbE • No FEC • AN/LT not supported

Table 9: A2040 Network Interfaces Summary

The Stratix10 FPGA on the A2040 is responsible for managing the network traffic for all three interfaces. On power-on, the FPGA will configure using a Golden bootstrap image which includes a simple three-way network passthrough module designed to route all downstream traffic received from the TOR link to either the SoC NIC or Mellanox CX5 NIC connections, and to interleave all upstream traffic received from the SoC NIC and CX5 NIC to transmit to the TOR. A single port on the TOR will receive traffic from both the SoC and the Host Server. Please note that this Golden bootstrap image defaults to have the network connection to the CX5 held in reset. Therefore, to get a link to the host, it is typically required to reconfigure into an application image which has this link out of reset.

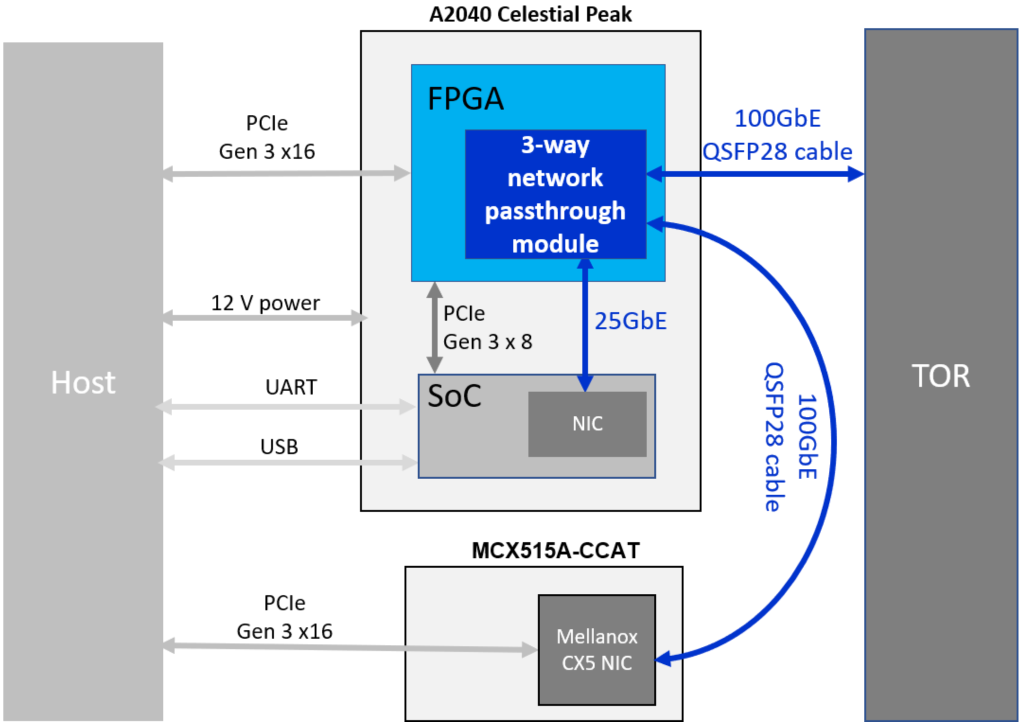


Figure 8: Three-way Network Passthrough Module inside FPGA

8.2.1 Celestial Peak: “bump-in-the-wire” Port Mapping

The two QSFP28 ports on A2040 are not interchangeable. Celestial Peak ports are labeled “NIC” and “SW” on the bulkhead. The ports shall be connected as shown in **Figure 9** below.

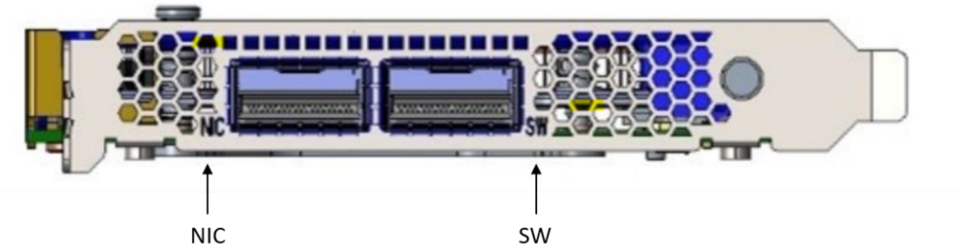


Figure 9: Celestial Peak Network Ports

8.2.2 Network Cable Requirements

Network Cable requirements are covered in **Section 4**.

8.2.3 Verifying Correct Network Port Connectivity

System Integrators shall verify correct FPGA port connectivity prior to shipping racks. The recommended method for doing so is by running the fpgadiagnostics utility to verify the attached cable part numbers. This can be accomplished using the following test sequence, the below commands may be run from the either the SoC or Host:

1) Run "*fpgadiagnostics -dumphealth*"
   i. Verify network status is OK.
   ii. Under FPGA-CABLES, verify the "cables_found" quantity is 2.
   iii. Verify the QSFP-0 cable present=1 and that the part number (pn:) matches the part number expected between the NIC and the FPGA.
   iv. Verify the QSFP-1 cable present=1 and that the part number (pn:) matches the part number expected between the FPGA and the TOR.

Example of FPGA-CABLES from fpgadiagnostics -dumphealth:

```
[FPGA-CABLES    ] OK [cables_found:2] [QSFP-0,present:1,done:1,timeout:0,id:0x11,tech:0xa0,vendor_id:00-
09-3a,pn:1002971321     ,rn:0x2020] [QSFP-1,present:1,done:1,timeout:0,id:0x11,tech:0xa0,vendor_id:78-
a7-14,pn:NDAAFJ-M202     ,rn:0x2042]
```

This can also be accomplished after PXE boot using the following test sequence, the below commands are run from the Host:

1) Run fpgadiagnostics.exe -dumphealth
   i. Verify network status is OK.
   ii. Under FPGA-NETWORK, verify both TOR-MAC and NIC-MAC versions of rx_count and tx_count are non-zero.
   iii. Verify both TOR and NIC lanes_deskew and lanes_stable are set to 1
2) Disconnect cable from NIC
3) Run fpgadiagnostics.exe -dumphealth
   i. Verify TOR lanes_deskew and lanes_stable are 1.
   ii. Verify NIC lanes_deskew and lanes_stable are 0.

4) Reconnect cable to NIC
5) Run fpgadiagnostics.exe -dumphealth
    i. Verify both TOR and NIC lanes_deskew and lanes_stable are set to 1

If any of the above steps fail, verify the cables are physically connected to the correct ports as described in the figures above.

8.3 I2C Path Requirements

Celestial Peak includes five separate 100kHz I2C paths shown in **Figure 10.** The BMC in the chassis that is the host endpoint for the server shall include an I2C master interface for the path to the A2040 card PCIe Edge Fingers. The other four I2C bus interfaces are wholly contained within the A2040 card with one controlled by the SoC and the other three controlled by the FPGA.

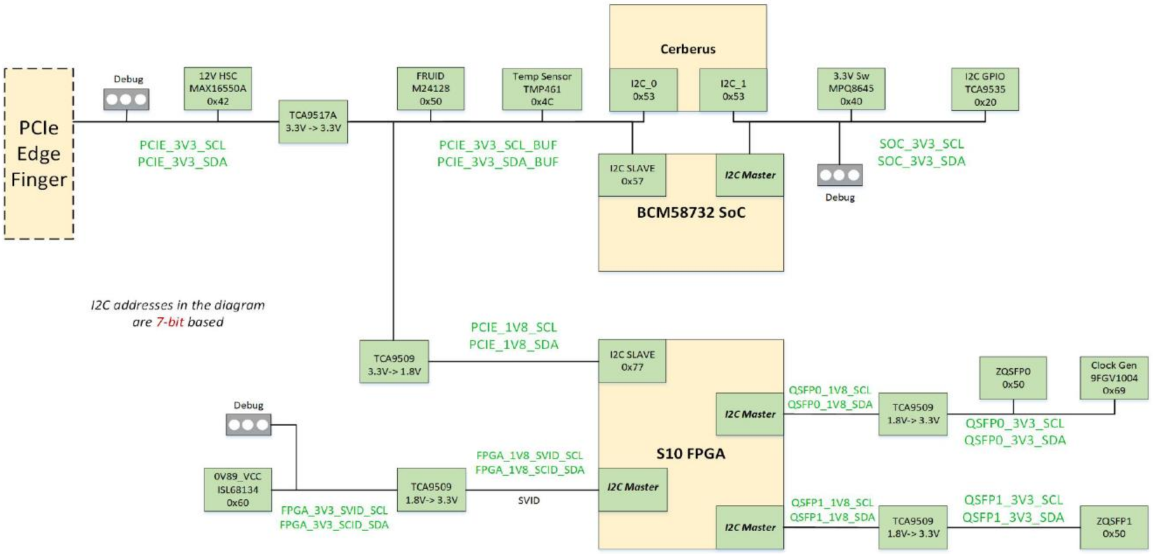


Figure 10: Celestial Peak I2C Paths

I2C Master	Address	Description	Component	BMC Representation
Host BMC	0x42	12V HSC	MAX16550A	HSC Temp & HSC Power
	0x4C	Temp Sensor	TMP461	Die Temp & Ambient Temp
	0x50	FRU EEPROM	M24128	IPMI FRU
	0x53	I2C_0	Cerberus	I2C Pass-through
	0x57	SoC I2C Slave	SoC	SOC Sensor, SOC <-> BMC IPMI Manageability
	0x77	S10 FPGA I2C Slave	S10 FPGA	I2C Pass-through
SoC	0x20	I2C GPIO	TCA9535	--
	0x40	3.3V Sw	MPQ8645	--
	0x53	I2C_1	Cerberus	--
FPGA 0	0x60	0V89_VCC	ISL68134	--

FPGA 1	0x50	QFSP Port 0	ZQSFP0	--
	0x69	Clock Gen	9FGV1004	--
FPGA 2	0x50	QFSP Port 1	ZQSFP1	--

Table 10: I2C Address Table

Host Servers shall comply with these requirements:

1. The path from the Host Server BMC to the FPGA shall have 2 or fewer I2C MUXes.
2. Each FPGA FRU shall be accessible over Host BMC I2C. The I2C command to use and bus address of each FPGA must be discoverable by software.
3. Software shall have the capability to correlate each FPGA I2C address with the card's location on the server.

8.3.1 A2040 I2C is NOT I2C/SMBus Specification Compliant

Celestial Peak includes a I2C level translator repeater component which is a "static offset" I2C buffer. "Static offset" I2C buffers are "compatible" with I2C/SMBus specifications but are not "compliant". This translates into the following important design constraint for the I2C interface from the host:

1. The output voltage low level,$V_{OL,}$ will be around 500mV (450mV – 600mV if we use the TCA9517A as an example), thus the host must be I2C compliant in order to provide margin for the 500mV V_{OL} signal coming from Celestial Peak. (I.e., the host side must have a input voltage low level, $V_{IL,}$ of 0.3*VCC as per the I2C specification, or 0.8V as per the SMBus specification.

8.3.2 Server Compatibility with I2C Level translators on Celestial Peak

The TI TCA9517A component shown in **Figure 11**: Host/Blade facing TCA9517A from Celestial Peak's schematic

has an A-side port and a B-side port. The B-side port is connected to the blade interface. Per TI specification, consecutive TCA9517A parts shall not be connected B-side to B-side, because of the buffered low voltage from the B-side. If the host includes a TCA9517A on this I2C path, the server design shall connect the A-side ports to the PCIE slot connecting the A2040. (Side B connects to the Host/Blade)

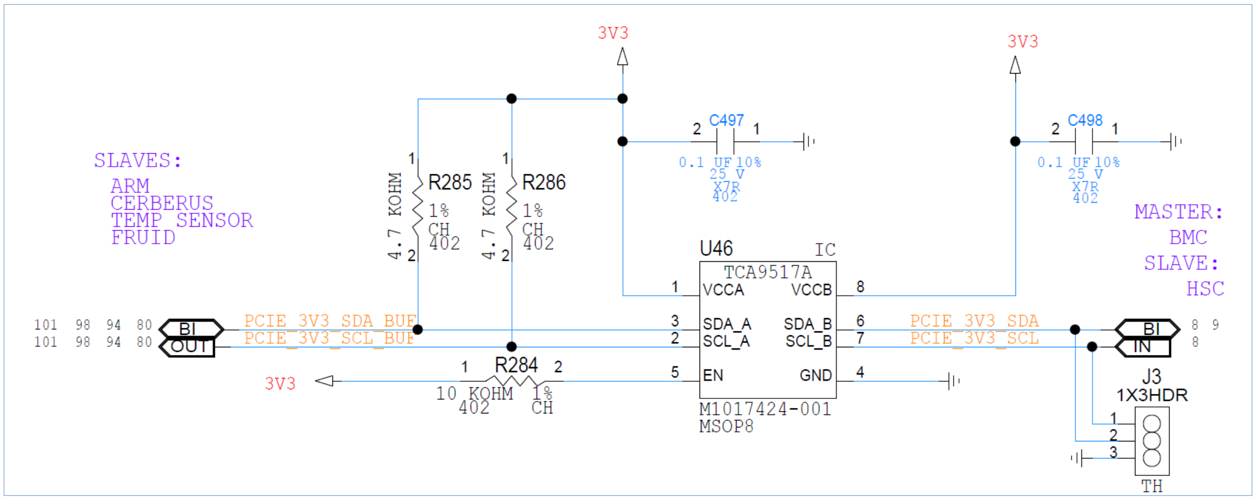


Figure 11: Host/Blade facing TCA9517A from Celestial Peak's schematic

(Side B connects to the Host/Blade)

9 BMC Requirements

The BMC in the chassis that is the manageability endpoint for the server shall provide the following functions for all PCIe cards (including FPGAs) in the System. The I2C devices exposed by the A2040 devices will be accessible to the BMC subsystem via the JSON configuration fed to the BMC. The BMC should consume the JSON configuration and understand the endpoints before providing the below-mentioned functions to external entities.

Required BMC Functions:

1. A2040 Card FRU EEPROM Access
2. Poll A2040 card hot swap controller for total card power consumption
3. Manage PCIe reset requirements for specific cards
4. Poll ambient card temperature, FPGA die temperature, and SoC temperature. Adjust server fan control settings to admissible air flow rates based on readings.
5. Access a subset of the FPGA registers via I2C.
6. Provide IPMI commands to query certain SOC parameters.
7. IPMI commands shall comply with Microsoft IPMI command format defined in the *Microsoft BMC Common-Core specification.* BMC provides Pass-through IPMI commands which basically serves to redirect any raw payload to the I2C endpoints. (Also see **Section 10.1** regarding BIOS requirements for IPMI routing.)
    - As mentioned in **Section 8.3**, the path from the host to the FPGA via BMC must have 2 or fewer I2C MUXes as explained in the BMC Blade API Spec.
8. The UART interface from BMC to SoC provides an out-of-band Administrative Serial Console to the SoC for serviceability.
9. The I2C interface from the BMC to the SoC facilitates thermal monitoring and SoC to BMC messaging to be used for thermal warning events and system cooling mechanisms. This

interface also serves as IPMI communication mechanism between the SOC and BMC for use by agents running on SOC OS.

10. The USB interface from BMC to SoC provides a backup SoC provisioning path to recover the SoC when the FPGA and/or network are unavailable.
11. The I2C interface must be used by the BMC to poll the SoC for IPMI commands issued from the SoC.

9.1 BMC access to FPGA FRU EEPROM

The BMC in the chassis that is the manageability endpoint for the server shall be capable of reading the FRU EEPROMS of all A2040 cards in the system. BMC should present the A2040 FRUs as IPMI FRUs. I.e., accessible via IPMI FRU commands. Each A2040 FRU should be represented as individual IPMI FRU IDs. The FRU EEPROM details will be listed in the inventory JSON file for the Rack Manager or any other devices to consume. To represent the FPGA FRU EEPROM as an IPMI FRU, the BMC FW includes index numbers for each FRU device in the system.

9.2 BMC Requirements for managing PCIe reset

To allow time for the FPGA to load from Flash the FPGA card requires a delay between turning on 12V power to the board and release of PCIe reset, BMC on the host server shall ensure a minimum of 1 second between 12V turn-on to the card and the rising edge (release) of PERST_N.

9.3 BMC Requirements for Checking FPGA Card Power Consumption

The BMC shall enable Power monitoring of total board consumption by accessing the A2040 Hot Swap Controller via I2C. The A2040 uses a Maxim MAX16550A Hot Swap Controller device. Please reference the spec for this device for information on the appropriate registers and offsets to access power readings as these may be different from Hot Swap Controllers used in prior FPGA card designs. **Table 11** below shows the max allowable power draw for the card and the I2C address of the card's hot swap controller. BMC must log a SEL event if it has any problems reading the A2040 HSC via I2C.

Max ASIC Power Consumed	**Server/Host power supply minimum sustained capability**	**Power Monitor I2C Address**
FPGA: 75W SoC: 30W TOTAL: 105W	130W	MAX16550A at I2C address 0x42 (7-bit address). The polling rate is once per second.

Table 11: Power Monitoring Access Points

Note: Please do not confuse the Max ASIC Power Consumed column from Table 11 with the Estimated Nominal Operation Power from Table 4. While the values are the same, they have different meanings. The first is *nominal* operation power for ASICs plus all other devices on the card, and the second is total *max* power draw from *just the ASICs* and no other devices on the card.

9.3.1 PCIe Power States

The FPGA does not take any action based on the PCIe power state. All PCIe devices must support the D3 and D0 power states, so the A2040 Card correctly acknowledges the D3 power state command but does not do anything to act upon it.

The Host Server must not enable D3 cold for the upstream parent port.

Power states D1/D2/L1/L2 are not supported.

9.4 BMC Requirements Polling Card Temperature

The Host Server BMC shall poll the TMP461 remote temp sensor device on the FPGA card (via I2C) to obtain Ambient card temperature and FPGA die temperature for use in the server thermal management algorithm. In addition, BMC must also read die temperature from the A2040 SOC. The BMC shall poll temperatures at a rate no faster than once per second. Please reference the spec for the TMP461 device for information on the appropriate registers and offsets to access temperature readings as these may be different from thermal sensors used in prior FPGA card designs.

The monitors listed in **Table 12** below shall be used for system fan control. This table shows monitoring access points for each of the cards:

1. The TMP461's local sensor provides board ambient temperature.
2. The TMP461's remote sensor, connected to the FPGA integrated thermal diode, provides FPGA die temperature (refer to ***6.1.1 Temperature Sensor Calibration and Offset for FPGA Die Temp*** for calibration steps required).
3. The I2C slave on the SoC provides the SoC die temperature. This must be obtained by BMC-SoC protocol command.

Area Monitored	Temperature Monitor Points
Board Ambient Temperature	TMP461 at I2C address 0x4C (7-bit address)
FPGA die Temperature	TMP461 at I2C address 0x4C (7-bit address)

SoC die Temperature	Directly from the SoC at I2C address 0x57 (7-bit address) using BMC-SoC Protocol command. I.e., not a physical sensor read

Table 12: Temperature Monitoring Access Points

9.5 BMC requirements for System Boot sequence.

Section 11 describes the boot sequence of A2040 and the Host Server. The default boot configuration on the Host Server shall be continual retry of PXE boot. To prevent a retry timeout, the BMC must support the capability to disable the FRB2 timer. The default for this timer in the BMC should be always disabled. The FRB2 Timer will be controlled by the BIOS by using the 'Set Watchdog Timer' standard IPMI commands to the BMC during boot time, as dictated by the BIOS flavor/profile.

9.6 BMC support for SoC serial console access

SoC serial console access in the rack is done from the Rack Manager using SSH protocol over the management network connection from the RM to the BMC. SSH to the standard port will cause the BMC to redirect serial access to the host console; whereas SSH to port 8295 is used to redirect to the SoC. Additional A2040 cards within the system may require additional port designations to differentiate between them.

9.7 BMC support for mounting RM images from the SoC USB

Steps **10** and **17** in **section 15.2** below provide the Rack Manager commands for media redirection to mount and unmount, respectively, an image on the RM to the A2040 SoC. This command uses IPMI over LAN protocol to connect the image as a USB Mass Storage Device. For implementations utilizing a mux to redirect from multiple USB host controllers to a single port on the BMC, it will also be required for the BMC to switch to the appropriate mux channel based on the RM command received. Switching the mux during the mount command will have the effect of disconnecting any currently connected USB devices, similar to pulling a USB thumbdrive, and it is expected that such devices will reconnect when the mux connection is restored to those devices during the unmount command.

Once mounted to an image on the Rack Manager, the SoC may either boot from USB to that image or simply access as another drive for transferring files. Booting to the image is the method used below to enable update of the SoC O/S and/or Firmware.

9.8 BMC support for IPMI commands issued from the SoC

IPMI commands may be entered from the SoC command line which will be stored locally for execution by the BMC. The BMC must poll the SoC over I2C to read any outstanding IPMI commands, process and write back the response. Please see the BMC BladeAPI Spec for additional details.

9.9 FPGA Slot preamble for IPMI commands

Assuming each PCIE slot is behind an I2C mux channel, the device channel number may be different per slot. The preamble for accessing any particular slot will depend on referencing the appropriate channel number. In Microsoft system implementations, this channel number is stored in the inventory JSON file from which the BMC pull the relevant channel number. This file will provide details for master mux read-write (bus, mux, channel, device slave address) which will be used to construct the preamble for accessing any relevant device. Other systems which may not use a muxed architecture may implement this differently.

10 BIOS Requirements

The Server BIOS shall enable the following features:

1. Identification of type and quantity of FPGA cards on the server
2. Memory Space BAR Resource Allocation for FPGA cards
3. PCIe Device management specific to FPGA cards
4. PCIe Power requirements specific to FPGA cards
5. PCIe Error reporting requirements specific to the FPGA cards
6. Continuous PXE boot / FRB2 timer disable

Detailed requirements for each of these topics are included in this section.

10.1 FPGA Device Identification

There are several ways to identify FPGA's in the system either Out-of-band or through the OS. The items below are the requirements for A2040.

10.1.1 Slot Unique Numbers for PCIe slots

The BIOS shall ensure a SUN (slot unique number) assignment method on each PCIe slot to ensure a unique name for each of those ports in the DSDT. This is a common requirement for all FPGA cards.

10.1.2 ACPI BIOS device names

The BIOS shall ACPI device names for each FPGA within the system and their parent root ports or switch ports in the same hierarchy. The ACPI BIOS device name from BIOS is used to populate the “BIOS device name” field in the OS. This is a common requirement for all FPGA cards.

10.1.3 Exposing A2040 ACPI device names to the OS

The BIOS shall expose the ACPI device names for the A2040 FPGA’s in the system through UEFI variables. Since the A2040 card cannot be detected by BIOS at first power-on, the BIOS will be responsible for populating these variables based on hard coding of the variables per system SKU. The OS uses these variables to locate the A2040 endpoints. The two UEFI variables are the following:

- **WcsFpgaSizeInfo** : defines how many FPGA’s are supported in the system
    - i. The GUID required is: {0xBB606E0B, 0x97E7, 0x425B, {0x9C, 0xF8, 0x23, 0x63, 0x95, 0x15, 0x6C, 0x7E}}
- **WcsFpgaPathSetup** : reports the FPGA BIOS device name
    - ii. The GUID required is: {0xCFF3B303, 0x8A67, 0x45BF, {0xB6, 0xA1, 0xEE, 0x87, 0x2A, 0x2E, 0x51, 0x99}}

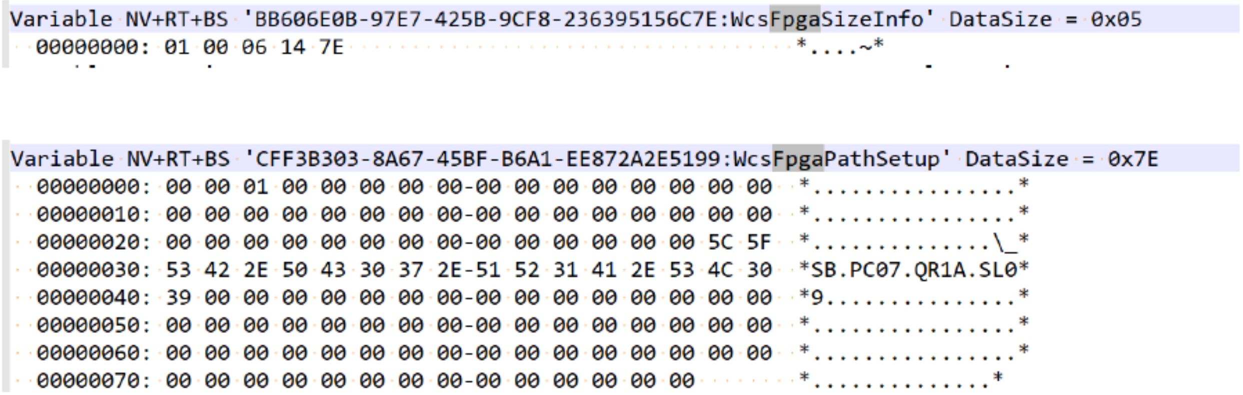
WcsFpgaSizeInfo

```
#define WCS_FPGA_PATH_SIZE_INFO_GUID \
{0xBB606E0B, 0x97E7, 0x425B, {0x9C, 0xF8, 0x23, 0x63, 0x95, 0x15, 0x6C, 0x7E}}
#define WCS_FPGA_PATH_SIZE_INFO_NAME L”WcsFpgaSizeInfo”
#define FPGA_PATH_INFO_MAJOR_VER 0x1
#define FPGA PATH_INFO_MINOR_VER 0x0
#pragma pack (1)
typedef struct _WCS_FPGA_PATH_SIZE_INFO {
UINT8 Version_Major;
UINT8 Version_Minor;
UINT8 NumberOfFPGAs;
UINT8 SizeOfFPGAPathEach;
UINT8 SizeOfFPGAVariable;
} _WCS_FPGA_PATH_SIZE_INFO;
#pragma pack ()
NumberOfFPGAs = 6;
SizeOfFPGAPathEach = MAX_FPGA_PATH_LENGTH;
SizeOfFPGAVariable = sizeof(WCS_FPGA_PATH_SETUP);
Version_Major = FPGA_PATH_INFO_MAJOR_VER;
Version_Minor = FPGA_PATH_INFO_MAJOR_VER;
```

WcsFpgaPathSetup

```
#define WCS_FPGA_PATH_SETUP _GUID \
{0xCFF3B303, 0x8A67, 0x45BF, {0xB6, 0xA1, 0xEE, 0x87, 0x2A, 0x2E, 0x51, 0x99}}
#define WCS_FPGA_PATH_SETUP _NAME L"WcsFpgaPathSetup"
#define MAX_FPGA_PATH_LENGTH 20
#pragma pack (1)
typedef struct _WCS_FPGA_PATH_SETUP {
UINT8 SlotBitmap[NumberOfFPGAs];
UINT8 FpgaAcpiName[NumberOfFPGAs][MAX_FPGA_PATH_LENGTH];
} WCS_FPGA_PATH_SETUP;
#pragma pack ()
```

Variable Definition Example:



```
Variable NV+RT+BS 'BB606E0B-97E7-425B-9CF8-236395156C7E:WcsFpgaSizeInfo' DataSize = 0x05
  00000000: 01 00 06 14 7E                                        *....~*


Variable NV+RT+BS 'CFF3B303-8A67-45BF-B6A1-EE872A2E5199:WcsFpgaPathSetup' DataSize = 0x7E
  00000000: 00 00 01 00 00 00 00 00-00 00 00 00 00 00 00 00  *................*
  00000010: 00 00 00 00 00 00 00 00-00 00 00 00 00 00 00 00  *................*
  00000020: 00 00 00 00 00 00 00 00-00 00 00 00 00 00 5C 5F  *..............\_*
  00000030: 53 42 2E 50 43 30 37 2E-51 52 31 41 2E 53 4C 30  *SB.PC07.QR1A.SL0*
  00000040: 39 00 00 00 00 00 00 00-00 00 00 00 00 00 00 00  *9...............*
  00000050: 00 00 00 00 00 00 00 00-00 00 00 00 00 00 00 00  *................*
  00000060: 00 00 00 00 00 00 00 00-00 00 00 00 00 00 00 00  *................*
  00000070: 00 00 00 00 00 00 00 00-00 00 00 00 00 00        *..............*
```

Figure 12: UEFI Variable Definition Example

10.1.4 Exposing A2040 ACPI device names through BMC

Another method to expose the BIOS ACPI device names is through the BMC inventory JSON configuration file. The ACPI device names and IPMI OEM command for accessing the inventory JSON have not yet been implemented but are the longer-term solution planned and may be part of future integration requirements document versions. This is mentioned for general awareness; however, no specific design accommodations are necessary at this time.

10.2 PCIe Resource Allocation

Even though the Host Server will not detect the A2040 device on AC power in UEFI PCIe enumeration (more on this in **Section 14**), the A2040 does not have full hot plug capability support, so the BIOS by default will need to enable the bridge device to which the A2040 is connected. The typical use case for the A2040 requires recurrent FPGA re-configuration which will cause the device to disappear and reappear to the Host Server at any time. To manage this the Host Server BIOS shall allocate memory space for the device in advance of detecting A2040 PCIe endpoints in accordance with the following requirements.

10.2.1 PCIe Bus Number Resource Reservation

The BIOS shall reserve 5 PCI Bus Numbers for the A2040 FPGA to accommodate up to 1028 PCI Functions (4 Physical Functions + 1024 Virtual Functions). The A2040 supports between 1 and 1028 total Physical and Virtual Functions via PCIe Alternative-Routing ID whose number may change on FPGA reconfiguration. Because the number of functions may change on FPGA reconfiguration, the BIOS must reserve enough busses to be able to support the 1028 Functions.

10.2.2 PCIe 32-bit Memory-Mapped I/O (MMIO) Address Space Reservation

The A2040 requires that the BIOS reserve 256 megabytes of contiguous memory space below 32bits that will be utilized in mapping all 32bit PCIe Base Address Registers. Reserved memory space must be aligned on a 64MB boundary. On re-configuration, the FPGA may change the number of Functions and the number, size, and type of BARs belonging to those functions. As a result, 32-bit address space reservation is required to ensure that all 32-bit BARs whose combined does not exceed the size of the 32-bit memory reservation can be successfully memory mapped. Without the precaution of 32-bit memory space reservation any increases in 32-bit MMIO space requirement may not be mappable as a result of fragmentation in 32-bit address space.

Specific platforms may have additional requirements for larger default apertures. SKU-specific apertures shall be defined by the BIOS configuration settings.

An exception can be requested for platforms without sufficient memory available for all the FPGA compatible PCIe slots, but this will limit the platform’s usability within the Azure fleet.

10.2.3 PCIe 64-bit Memory-Mapped I/O (MMIO) Address Space

The A2040 may use up to 32GB of 64-bit MMIO space depending on the image loaded on the FPGA.

10.2.4 PCIe MaxReadReq BIOS Setting

The FPGA advertises and expects the PCIe MaxReadReq setting in the BIOS to be set to 512 bytes. In some BIOS editions, if MaxReadReq is set to “Auto”, the MaxReadReq may get set to 4096 bytes. The FPGA does not support the value of 4096 bytes, please ensure that the MaxReadReq is set to 512 bytes.

10.2.5 Legacy I/O Space

The A2040 shall not request any I/O address space.

10.2.6 BIOS Interrupt Handling

The A2040 does not issue legacy interrupts. All MSI and MSI-X will be handled by OS and driver software.

10.2.7 PCIE Hot Plug disabled

The reservation of resources should not utilize the PCIE Hot Plug method. It is expected that the PCIE Bridge device will reserve resources without Hot Plug enabled.

10.2.8 Network Card Memory Resource Allocation

It is required that BIOS also restricts the memory resources allocated to the associated network card to be within the first 32 address bits, or 4GB, of memory address space.

10.3 Device Management of the A2040

The following two cases shall be addressed during POST as BIOS enumerates PCIE:

- If the FPGA does not have a Host-side PCIE endpoint exposed, as is the case when the Golden image is configured, the device shall not be enumerated or appear in the PCIe device tree.

- If the FPGA does have a Host-side PCIE endpoint exposed, as is the case when the factory test image is configured, the device shall be enumerated and appear in the PCIe device tree.

In either case, the BIOS must ensure the root port or downstream switch port to the slot is always enumerated and assigned the appropriate memory window range of bus numbers, even if the device does not have the Host-side PCIE endpoint exposed. This allows the OS to rescan the port and detect the endpoint when the FPGA is reconfigured to expose one.

The BIOS must provide ACPI device names for the A2040 and its root port or downstream switch port, as described in **section 10.1** above.

10.4 PCIe Power States

The FPGA does not take any action based on the PCIe power state. All PCIe devices must support the D3 and D0 power states, so the A2040 Card correctly acknowledges the D3 power state command but does not do anything to act upon it.

The Host Server must not enable D3 cold for the upstream parent port.

Power states D1/D2/L1/L2 are not supported.

10.5 BIOS requirements PCIe Error Reporting

The platform shall expose PCIe Advanced Error Reporting (AER) capabilities on both the FPGA device and the PCIe root port immediately upstream from the FPGA. These capabilities shall be discoverable by walking the extended capability list exposed in the configuration space of both devices.

10.6 Continuous PXE boot / disabling FRB2 timer

The FRB2 Timer will be controlled by the BIOS by using the ‘Set Watchdog Timer’ standard IPMI commands to the BMC during boot time. Systems supporting Celestial Peak must disable the FRB2 timer in order to allow continual try of PXE boot until the Host network is enabled. For example, in some implementations this may be controlled via BIOS Flavor profile.

11 Cerberus Requirements

The Cerberus Security device resides on the A2040 card and exists to support SoC FW authentication and attestation. To achieve its purpose, the Cerberus Security device on Celestial Peak must be provisioned during rack integration. The process for Cerberus certificate provisioning is detailed in the document **M1162196 Cerberus Onboarding Playbook**.

The A2040 Cerberus device can be controlled either in-band from the A2040 SoC or the blade Host or out-of-band from the Rack Manager. Any new system design supporting A2040 must provide access to control the Cerberus device from these sources. The Host access is via IPMI over KCS interface to the Host BMC which then communicates with the Cerberus over I2C. SoC access to Cerberus is over a separate SPI bus on the A2040 card which is fully contained on the card. For this access, the system only needs to provide a network path into the SoC as the command line access method to issue Cerberus commands. An alternate path to access the SoC command line for issuing Cerberus utility commands is via serial console access from the Rack Manager; however, Rack Manager direct access to Cerberus is the recommended approach. Rack Manager direct access to Cerberus on A2040 is based on IPMI over the management network to the BMC, which again interfaces to the Cerberus over I2C. Any system implementation must fully support all three mechanisms.

12 Rack Manager Requirements

The Rack Manager must provide the A2040 card with Command Line Interface (CLI), Redfish and OCS Driver support for capabilities including SoC reset, reading card power state, setting of the SoC boot order, harvesting SoC MAC address and serial session to the SoC. In addition, it must provide access to telemetry data such as Soc and FPGA device temperatures, firmware versions, card power consumption, and card Field Replaceable Unit (FRU) data. The telemetry data must also be pushed by Rack Manager to relevant databases for monitoring of fleet health and status.

The Rack Manager interfaces with the A2040 indirectly by communicating through the Host BMC over a management network connection. Please see Section 0 for details on BMC requirements necessary to support Rack Manager operations with the A2040.

In order to access A2040 FRU data from the Rack Manager, the command is:

"set system cmd -i <x> -c fru"   (where <x> is to be replaced with the blade slot number)

An example of the A2040 portion of that command output is:

```
FRU Device Description : Overlake_EEPROM (ID 1)
 Board Mfg Date        : Mon Aug 17 11:47:00 2020
 Board Mfg             : Microsoft
 Board Product         : A2040
 Board Serial          : B65190335020900A
 Board Part Number     : M1096520-001
 Board Extra           :   A
 Product Manufacturer  : Microsoft
 Product Name          : A2040
 Product Part Number   : M1096519-001
 Product Version       : 4.0
 Product Serial        : B65190335020900A
 Product Asset Tag     : 30303030303030
 Product Extra         : A
 Product Extra         : dc98409ddbb2202020202020
    Completion Code: Success
```

The Rack Manager expects that the I2C topology of the system with A2040 will be available through the Host BMC. For IPMI-based BMC designs, it expected that it will utilize an inventory file in JSON format that provides addressing information to access each device. For Redfish-based BMC designs, the above requirement is not applicable. Please contact Microsoft for additional details on requirements for Redfish BMC designs.

The blade profile in the Rack Manager shall support multiple A2040 cards per blade where relevant.

13 OS Compatibility

Drivers for the FPGA cards are validated and compatible with the following OSes:

Drivers for the Host Server:

- Windows Server 2012r2
- Windows Server 2016
- Windows Server 2019 only on OS builds equal to or later than 17763.1.180914-1434

Drivers for the SoC

- Linux 4.19.x

14 Boot Sequencing

The Celestial Peak A2040 delays PXE boot of the Host Server to allow the SoC to boot first. The Host BIOS, BMC, Rack Manager, FPGA images, and FPGA SW and Drivers control this activity.

- On power up, the default "Golden" FPGA image will configure. This image will:
    - enable network traffic to the SoC,
    - hold network to the Host Server in reset
    - has no PCIe interface configured to the Host Server
- The SoC defaults to PXE boot mode, so it will boot after the FPGA is configured providing network traffic to the SoC
- From the SoC, the FPGA can be re-configured from Golden image to the FactoryTest image which enables network and PCIe to the Host Server
- The Host Server defaults to PXE boot mode. It shall continue to retry connection with a PXE server, but will not succeed until FPGA is configured into the FactoryTest image
- The NIC card and the A2040 Card are two separate PCIe devices.
    - Since the FPGA Golden image does not include a PCIe interface to the Host, on power up the Host Server will not discover the A2040 device during UEFI PCIe enumeration, but the NIC card will be discovered.
    - When the A2040 FPGA is re-configured into the FactoryTest image the A2040 will be discoverable by the host.

This timing diagram represents the steps from power on to Host ready to server live traffic. A more detailed description of the activities can be found below the diagram.

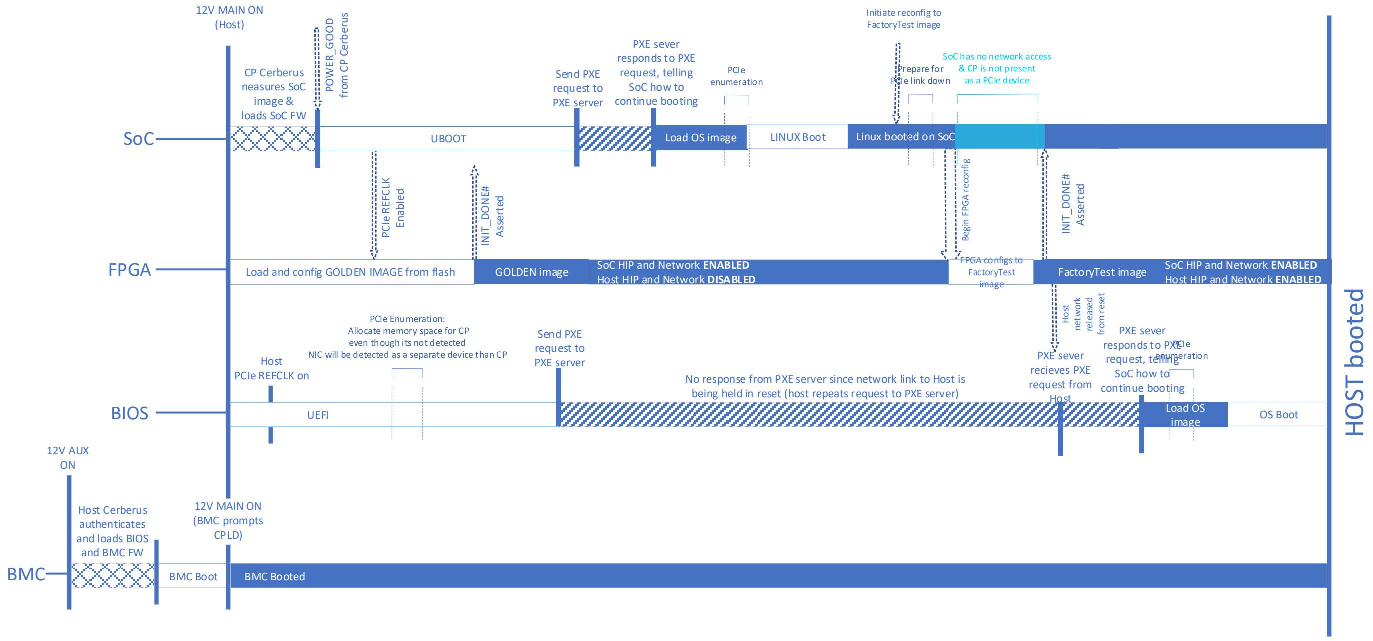

12V MAIN ON
(Host)
CP Cerberus neasures SoC image & loads SoC FW
POWER_GOOD from CP Cerberus
SoC
UBOOT
Send PXE request to PXE server
PXE sever responds to PXE request, telling SoC how to continue booting
Load OS image
PCIe enumeration
LINUX Boot
Linux booted on SoC
Initiate reconfig to FactoryTest image
Prepare for PCIe link down
SoC has no network access & CP is not present as a PCIe device
PCIe REFCLK Enabled
INIT_DONE# Asserted
Begin FPGA reconfig
INIT_DONE# Asserted
FPGA
Load and config GOLDEN IMAGE from flash
GOLDEN image
SoC HIP and Network ENABLED
Host HIP and Network DISABLED
FPGA configs to FactoryTest image
FactoryTest image
SoC HIP and Network ENABLED
Host HIP and Network ENABLED
Host network released from reset
Host
PCIe REFCLK on
PCIe Enumeration:
Allocate memory space for CP
even though its not detected
NIC will be detected as a separate device than CP
Send PXE request to PXE server
No response from PXE server since network link to Host is being held in reset (host repeats request to PXE server)
PXE sever recieves PXE request from Host
PXE sever responds to PXE request, telling SoC how to continue booting
PCIe enumeration
BIOS
UEFI
Load OS image
OS Boot
12V AUX ON
Host Cerberus authenticates and loads BIOS and BMC FW
12V MAIN ON (BMC prompts CPLD)
BMC
BMC Boot
BMC Booted
HOST booted

Figure 13: Boot sequence from AC power on to Host serving live traffic

14.1 Detailed description of Boot flow from AC power on

Power is applied to blade and FPGA, SoC and Host all begin boot sequence in parallel:

1. 12 V Main power is applied to the Host Server
2. Host Cerberus authenticates and loads Host BIOS and BMC

Items 3, 4 and 5 happen in parallel:

3. FPGA automatically loads Golden image from flash,
   a. As soon as the FPGA detects PCIe Refclk on SoC PCIe endpoint it completes configuration into Golden and asserts INIT_DONE signal.
   b. The Golden Image will have the following features:

FPGA Image	Network connect to TOR		PCIe	
	SoC to TOR	Host to TOR	PCIe to SoC (HIP0)	PCIe to Host (HIP1)
Golden image	enabled	Disabled	instantiated	Not instantiated

4. Host Boot:
   a. Host begins boot via UEFI.
      i. Host completes PCIe discovery, but will not detect the A2040 since PCIe to Host is not instantiated in FPGA Golden image

ii. Host boot config defaults to attempt PXE first, then attempt boot from disk, and then retries PXE in a repetitive cycle until one or the other occurs. Generally, with a few exceptions, the SKUs with A2040 will not have any OS image in disk, so the Host will simply retry PXE until response is received.

5. SoC Boot:
    a. A2040 Cerberus authenticates and loads SoC FW (ATF, u-boot)
    b. SoC begins ATF
    c. SOC uBoot sends PXE request
    d. PXE Server responds to PXE request,
    e. Linux image is verified and measured into fTPM PCRs before booting.
    f. PCIe enumeration of FPGA on Linux SOC occurs during SOC OS boot

    Note: The SOC boot mode can be changed using a command from the rack manager or from the SoC OS. During the BSL flow, the BSL controller will change the boot-mode from PXE boot to eMMC boot. The BareMetal controller will change the boot-mode back to PXE boot. In a Live datacenter environment, however, the SoC boot mode will be eMMC and switched to PXE only for recovery flows.

6. Reconfigure FPGA into FactoryTester Image
    a. Network connection to SoC will be temporarily unavailable during reconfig, reconnects when reconfig is complete
    b. PCIe interface between FPGA and SoC will disappear during reconfig then reappears after reconfig is complete.

FPGA Image	Network connect to TOR		PCIe	
	SoC to TOR	Host to TOR	PCIe to SoC (HIP0)	PCIe to Host (HIP1)
App Image	enabled	enabled	instantiated	instantiated

    c. Network to Host will connect after reconfig which will enable host to PXE boot
    d. PCIe interface between Host and FPGA will instantiate. (Note that in order to train up the Host-side PCIE endpoint, it is required to run a command from the host side. More details on this covered in **section 15.4.1**.)
7. Host will receive PXE response and complete boot.

14.2 Enabling Boot Sequence in a Test Environment

In a production Datacenter environment, the Host Server OS boot occurs sequentially after SOC boot and FPGA reconfiguration. The FPGA reconfiguration to release the Host is managed by Agents running on the SoC and Host. In a test environment, a reconfiguration from an FPGA Golden image to an FPGA Factory image will enable connection to the Host Server. FPGA reconfiguration steps can be performed either manually or using test automation such as CSIToolkit. The steps to reconfig to App1 flash slot from the SoC are described in **Section 15.4.1**.

Note: when reconfiguring from the SoC from an image with Host PCIE endpoint enabled, the host may experience a PCIE uncorrectable surprise down error. See section 15.7 for instructions for preventing this error during reconfiguration.

15 Firmware/Software Update Instructions

The instructions in this section are specific to updates and configuration at boot to enable testing.

Please note that the below instructions are for reference only. The version of the recipe and its associated FW/SW will differ from the examples shown below.

There are five firmware/software components which may be updated on the Celestial Peak card for any recipe release. Those components are:

- Celestial Peak Cerberus Firmware
- SoC PFM (FIP + Nitro)
- SoC O/S (including FPGA drivers and utilities)
- SoC FW (FIP + Nitro)
- FPGA Images

In addition to these, there are **also FPGA software drivers and utilities** for Celestial Peak which must be updated on the Host System. This guide is intended to provide instructions for the proper processes to update and verify each of these firmware and software components.

15.1 Celestial Peak Cerberus Firmware Update

The Celestial Peak Cerberus Firmware is physically located on flash within the device itself. Updates can be accomplished via the Host Server O/S, the SoC O/S or the Rack Manager. The instructions below are those associated with updating from the Host O/S. **This Cerberus FW should not be confused with a Cerberus on the Host Server and its associated firmware. A Host System Cerberus requires a different set of firmware. The instructions below pertain only to the Celestial Peak Cerberus device.**

1) Copy the recipe files to the Host. Instructions here assume the entire recipe directory has been copied to C:\. For example: C:\Overlake_System_RevF.2.DV1\

   The Cerberus Utility for verifying firmware versions and performing updates is available at:
   C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows
   (note: please use the latest version of Cerberus Utility available in the Utilities directory)

2) Confirm which firmware version is currently installed by typing the following command in the Host OS Command Prompt: `.\cerberus_utility.exe -s 53 -m 70 -c 5 --slave fwversion`
    - The “-c” parameter is the I2C channel number
    - The “-s” parameter is the slave device address
    - The “-m” parameter is the mux slave address

   The C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\ directory includes a *Cerberus Utility Tool Guide* describing all tool parameters and commands.

   (note: the parameters above may require different values for your system or card slot location. Please verify with the system designer or BMC developer for the appropriate value selections.)

3) After establishing the Cerberus firmware version, there are two main cases to be considered when updating the CP Cerberus – (a) first time update and (b) subsequent updates. Depending on the result of step (2) follow **either** (a) or (b) below. Do not do both.

   a) If the result from the prior step indicated that no Cerberus FW is installed, then program Cerberus for the first time (adjust I2C channel and address parameters to Host Server addressing scheme):

   General command: `.\cerberus_utility.exe -s 10 -m 70 -c 5 isp update_fw 0 <file>`

   For example, from C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows:

```
.\cerberus_utility.exe -s 10 -m 70 -c 5 isp update_fw 0 ..\..\..\v2.2.1.1\cerberus_v2.2.1.1.bin
```

   After the update, the FW will be applied automatically via a Cerberus reset.

   b) If the result from the prior step indicated that Cerberus FW is installed, then program Cerberus with the appropriate utility version according to the FW version currently installed and the document:
   C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\Utility Release Notes.pdf

   This may involve a two-step process depending on the FW version currently installed.

   (e.g. - TO UPDATE TO v2.1.1.5 OR LATER, YOU MUST FIRST BE RUNNING AT LEAST v2.1.1.3.)

   General command: `.\cerberus_utility.exe -s 53 -m 70 -c 5 --slave fwupdate <file>`

   For example, from C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows:

```
.\cerberus_utility.exe -s 53 -m 70 -c 5 –slave fw_update ..\..\..\v2.2.1.1\cerberus_v2.2.1.1.bin
```

   This command should take between a minute to a minute and a half to complete.

If you are running earlier than v2.1.1.3, you will appropriately construct and execute the above general command twice (using the appropriate utility version), first to get to v2.1.1.3 and next to get to the desired version beyond that.

After the update, the FW will be applied automatically via a Cerberus reset.

4) Repeat step 2) after the update to verify the new Cerberus FW version.

15.2 SoC PFM Update

The SoC PFM (Platform Firmware Manifest) is stored on a separate “recovery” flash device connected to the Cerberus device. The default state is to have no PFM loaded so that the Cerberus device is in “bypass” mode for the SoC FIP and SoC NIC (Nitro) images. In later versions of the socflash update, the PFM should be loaded and activated when the SoC FW is updated. In this document, we will provide instructions for updating the SoC FIP and SoC NIC PFM files independently of the SoC FW updates. Each component of the SoC firmware image, FIP and NIC (Nitro), has its own .PFM file to specify the permissible versions of firmware for loading. Each can be updated independently using the same Cerberus Utility as in the first section and from the same sources – Host O/S, SoC O/S or Rack Manager. It is best to check if the PFM’s have been updated after a socflash update using the following procedure. Note: the steps below assume that the recipe has already been copied over to the Host system C: drive root folder.

1. verify if either of the PFM’s are currently loaded and, if so, which version(s). The SoC PFM’s can be verified using the following commands from the Host OS Command Prompt.
General Commands:

➢ Verify whether FIP PFM is active

```
./cerberus_utility.exe -s 53 -m 70 -c 5 --slave checkbypass 0
```

➢ Verify whether Nitro PFM is active

```
./cerberus_utility.exe -s 53 -m 70 -c 5 --slave checkbypass 1
```

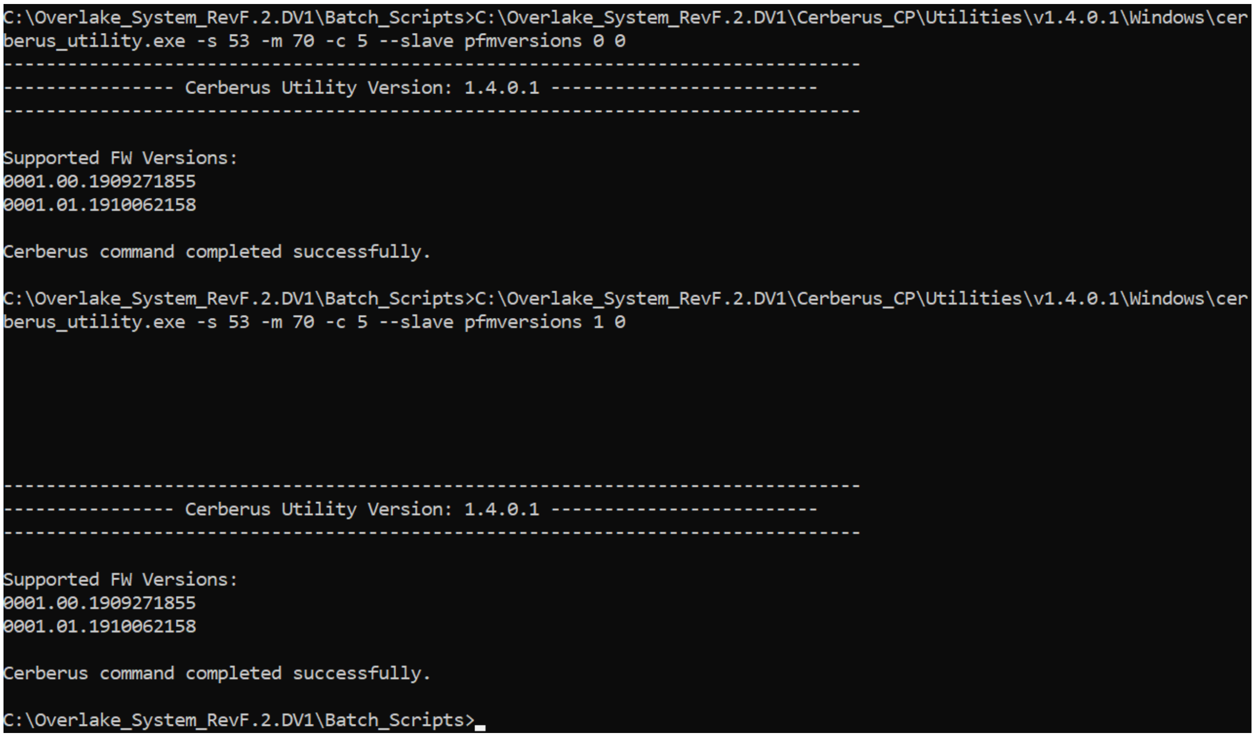
➢ Verify active version of FIP PFM

```
./cerberus_utility.exe -s 53 -m 70 -c 5 --slave pfmversions 0 0
```

➢ Verify active version of Nitro PFM

```
./cerberus_utility.exe -s 53 -m 70 -c 5 --slave pfmversions 1 0
```

Note: The C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\ directory includes a *Cerberus Utility Tool Guide* describing all tool parameters and commands



```
C:\Overlake_System_RevF.2.DV1\Batch_Scripts>C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows\cer
berus_utility.exe -s 53 -m 70 -c 5 --slave pfmversions 0 0
--------------------------------------------------------------------------------
--------------- Cerberus Utility Version: 1.4.0.1 ------------------------
--------------------------------------------------------------------------------

Supported FW Versions:
0001.00.1909271855
0001.01.1910062158

Cerberus command completed successfully.

C:\Overlake_System_RevF.2.DV1\Batch_Scripts>C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows\cer
berus_utility.exe -s 53 -m 70 -c 5 --slave pfmversions 1 0







--------------------------------------------------------------------------------
--------------- Cerberus Utility Version: 1.4.0.1 ------------------------
--------------------------------------------------------------------------------

Supported FW Versions:
0001.00.1909271855
0001.01.1910062158

Cerberus command completed successfully.

C:\Overlake_System_RevF.2.DV1\Batch_Scripts>_
```

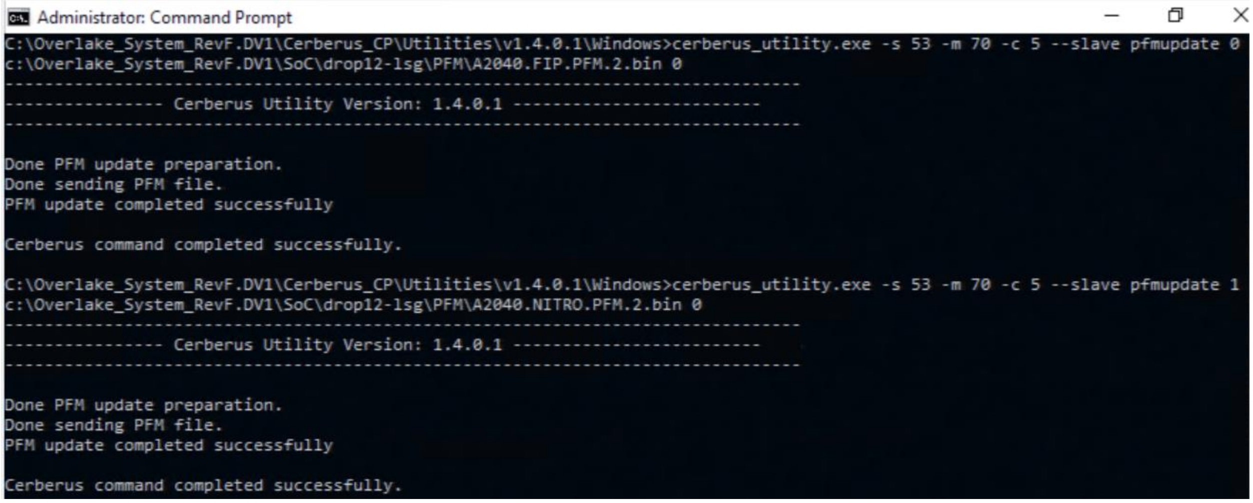
2. Assuming that after observing the SoC PFM version/status, it is deemed appropriate to update, then PFM's can be updated using the following commands from the Host OS Command Prompt. General Commands:

   ➢ Update FIP PFM

   .\ cerberus_utility.exe -s 53 -m 70 -c 5 --slave pfmupdate 0 <FIP pfm file> 0

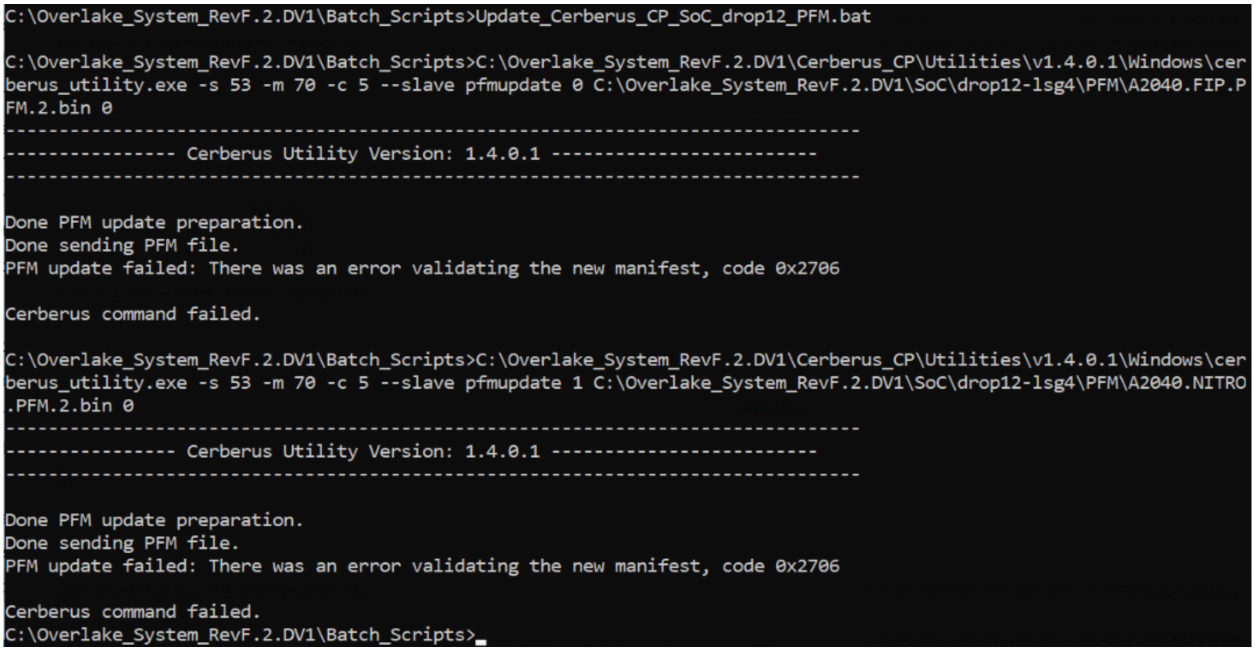
   ➢ Update Nitro PFM

   .\ cerberus_utility.exe -s 53 -m 70 -c 5 --slave pfmupdate 0 <Nitro pfm file> 0



```
C:\Overlake_System_RevF.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows>cerberus_utility.exe -s 53 -m 70 -c 5 --slave pfmupdate 0
c:\Overlake_System_RevF.DV1\SoC\drop12-lsg\PFM\A2040.FIP.PFM.2.bin 0
--------------------------------------------------------------------------------
--------------- Cerberus Utility Version: 1.4.0.1 ------------------------
--------------------------------------------------------------------------------

Done PFM update preparation.
Done sending PFM file.
PFM update completed successfully

Cerberus command completed successfully.

C:\Overlake_System_RevF.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows>cerberus_utility.exe -s 53 -m 70 -c 5 --slave pfmupdate 1
c:\Overlake_System_RevF.DV1\SoC\drop12-lsg\PFM\A2040.NITRO.PFM.2.bin 0
--------------------------------------------------------------------------------
--------------- Cerberus Utility Version: 1.4.0.1 ------------------------
--------------------------------------------------------------------------------

Done PFM update preparation.
Done sending PFM file.
PFM update completed successfully

Cerberus command completed successfully.
```

Note: if an attempt is made to install a previously installed PFM file, the following error message will result:



```
C:\Overlake_System_RevF.2.DV1\Batch_Scripts>Update_Cerberus_CP_SoC_drop12_PFM.bat

C:\Overlake_System_RevF.2.DV1\Batch_Scripts>C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows\cer
berus_utility.exe -s 53 -m 70 -c 5 --slave pfmupdate 0 C:\Overlake_System_RevF.2.DV1\SoC\drop12-lsg4\PFM\A2040.FIP.P
FM.2.bin 0
--------------------------------------------------------------------------------
---------------- Cerberus Utility Version: 1.4.0.1 -------------------------
--------------------------------------------------------------------------------

Done PFM update preparation.
Done sending PFM file.
PFM update failed: There was an error validating the new manifest, code 0x2706

Cerberus command failed.

C:\Overlake_System_RevF.2.DV1\Batch_Scripts>C:\Overlake_System_RevF.2.DV1\Cerberus_CP\Utilities\v1.4.0.1\Windows\cer
berus_utility.exe -s 53 -m 70 -c 5 --slave pfmupdate 1 C:\Overlake_System_RevF.2.DV1\SoC\drop12-lsg4\PFM\A2040.NITRO
.PFM.2.bin 0
--------------------------------------------------------------------------------
---------------- Cerberus Utility Version: 1.4.0.1 -------------------------
--------------------------------------------------------------------------------

Done PFM update preparation.
Done sending PFM file.
PFM update failed: There was an error validating the new manifest, code 0x2706

Cerberus command failed.
C:\Overlake_System_RevF.2.DV1\Batch_Scripts>_
```

3. After the update has completed successfully, to activate the PFM reboot the system. Reboot can be accomplished with an AC power cycle which can be accomplished from the Rack Manager with the commands: "set manager port off -i <x>" followed by
"set manager port on -i <x>"

4. After the reboot, verify the PFM's are active and the version numbers by repeating step (1) above.

15.3 SoC O/S and SoC Firmware Update

The SoC Operating System is installed on the eMMC device on the Celestial Peak card. The SoC firmware is installed on a dedicated QSPI flash device on the card. And the SoC PFM data is stored on a QSPI flash device connected to Cerberus. All may be updated together using a single procedure, or they may each be updated independently. Generally, you will update all at once. There are three methods for updating the SoC O/S, Firmware and PFM's. Those are:

- Update via system Host O/S (recovery)
- Update via Rack Manager (standard or recovery)
- Update via SoC O/S (requires SoC FW and O/S already installed)

Normal updates are expected to happen via the SoC O/S. However, for this document, the SoC O/S and FW update instructions will be from the Rack Manager. Additionally, SoC FW-only recovery instructions will be provided that can be run from the Host O/S.

Note that update to a particular SoC FW version can have a dependency on the Cerberus FW installed. For example, there may be a minimum Cerberus FW version requirement in some cases. When doing upgrades or downgrades of the SoC FW, please first make note of any such dependencies which will be called out within the SoC FW release notes.

SoC OS updates require one of the following interfaces to A2040:

a. The Host Server with A2040 is installed in a Gen6 rack with connection to a rack manager through a management switch into the Host Server's BMC.
b. A connection to a management switch and stand-alone rack manager,
c. A custom USB and serial dongle cables attached to the A2040.

These instructions assume the hardware configuration described in (a) and the install image is already copied to a TFTP server on the same network as the rack manager. (Install image located in the SoC rMedia directory, e.g. C:\ Overlake_System_RevF.2.DV1\SoC\drop12-lsg4\rMedia).

The recovery instructions assume that the system recipe has been copied over to the Host system including the Cerberus utility and SoC FW image files.

15.3.1 Update from Rack Manager

STEPS:

1. Ensure the A2040 internal USB and UART cables are connected to the Host Server. Example connected per **Figure 14**.

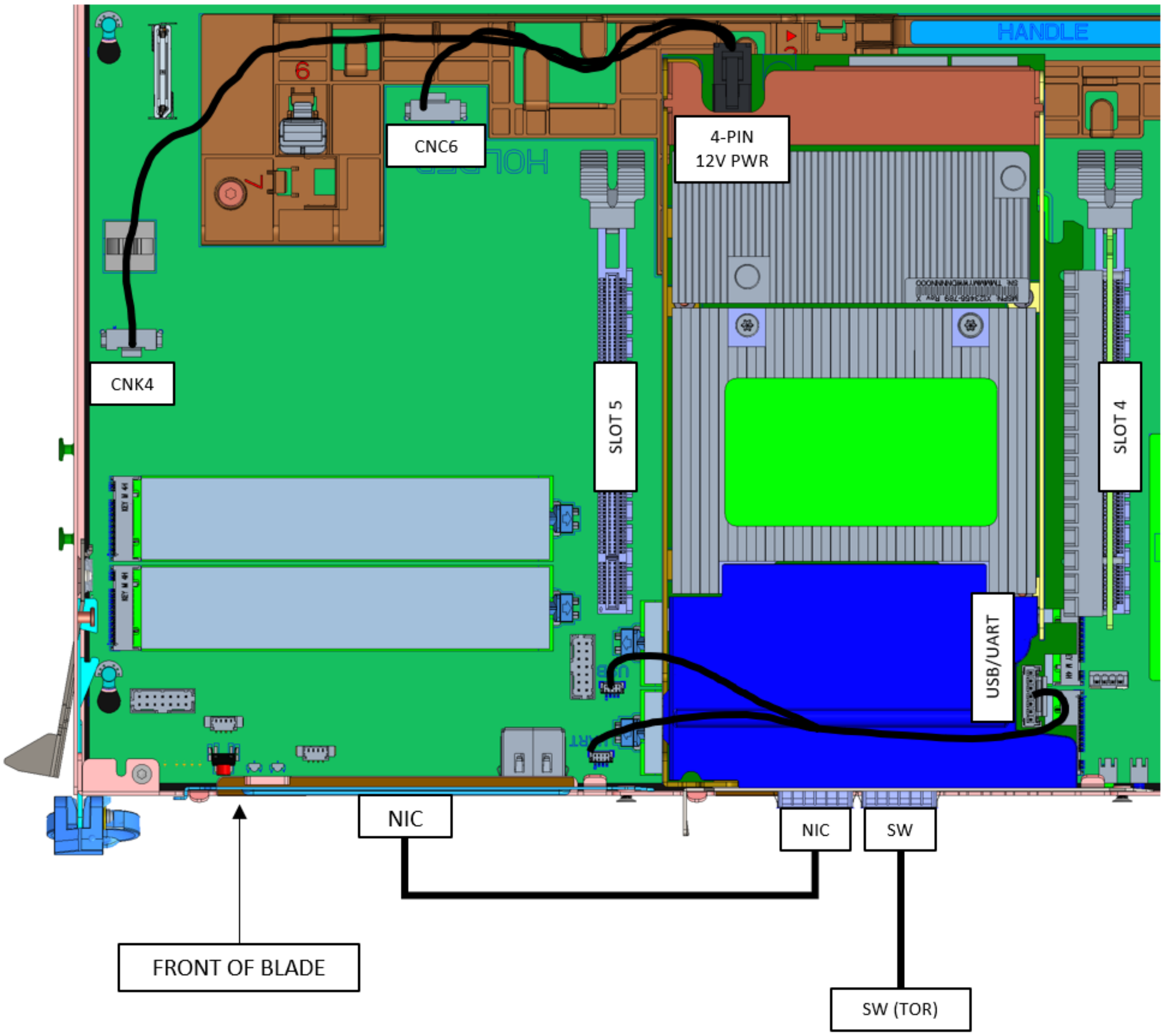


Figure 14: C2030 Host Server USB and UART cabling example

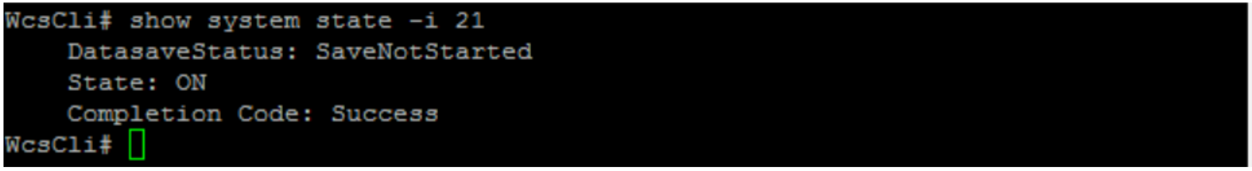
2. Verify that the minimum recipe version pre-requisites have been met:

   - Rack Manager: 1.1.8.5 or later
   - BMC: 2.09.101 or later
   - SoC FW: Drop5 or later

3. Start up two connections to the Rack manager using SSH from Putty or other terminal connection tool. Referred to in these instructions as “window 1”and “window 2.”

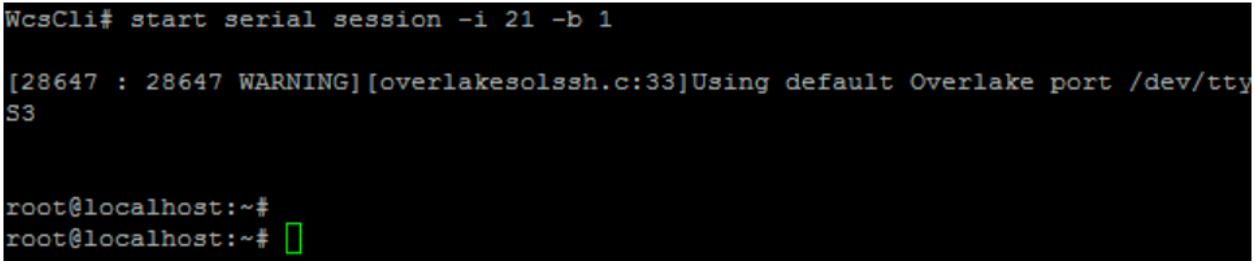
4. From window 1, verify that the blade is powered on:

"show system state -i <x>"
(where <x> is to be replaced with the blade slot number)



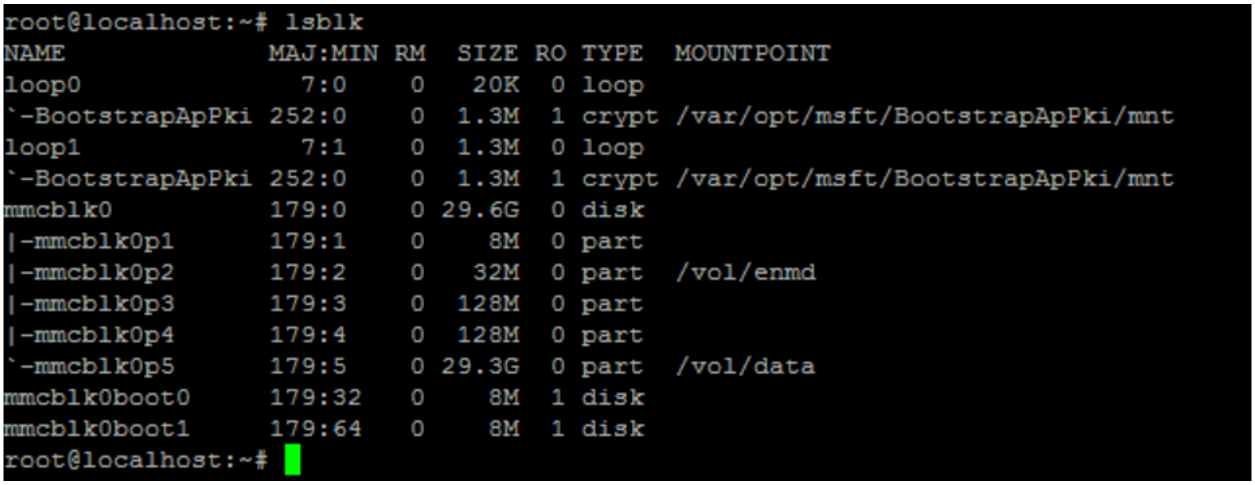
```
WcsCli# show system state -i 21
    DatasaveStatus: SaveNotStarted
    State: ON
    Completion Code: Success
WcsCli# 
```

5. From window 1, start a serial session to the Celestial Peak SoC:
   "start serial session -i <x> -b <y>"
   (where <x> is to be replaced with the blade slot number)
   (where <y> is to be replaced with the device to access, 0 for host CPU, 1 for CP SoC)
   (if the -b option is omitted the command will default to the host CPU)



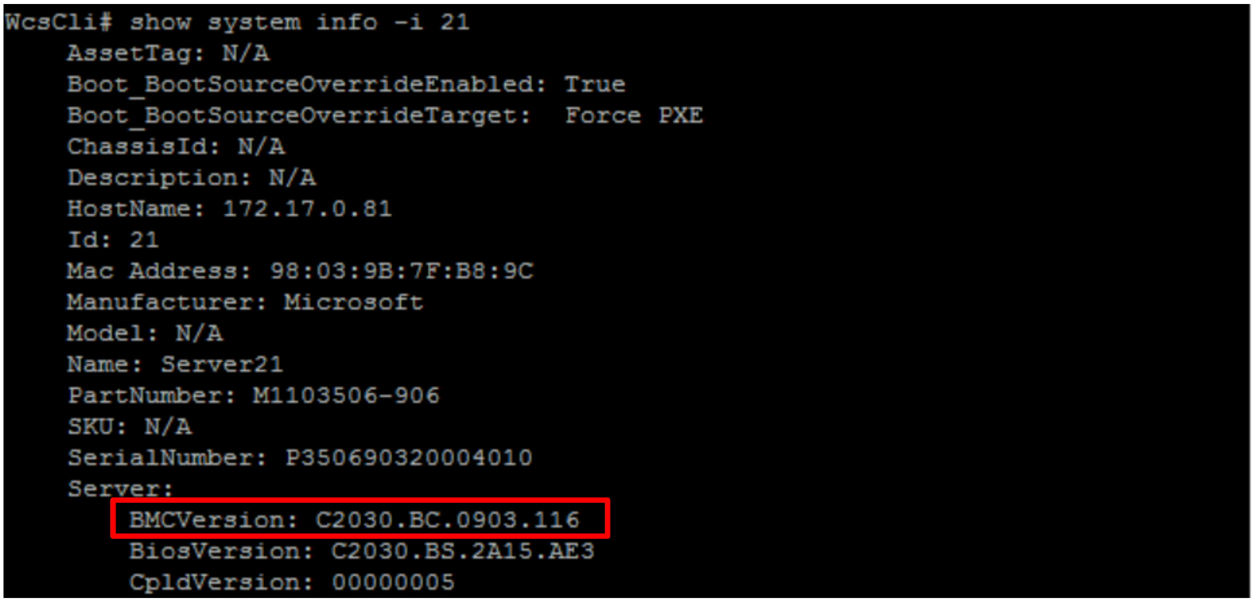
```
WcsCli# start serial session -i 21 -b 1

[28647 : 28647 WARNING][overlakesolssh.c:33]Using default Overlake port /dev/tty
S3


root@localhost:~#
root@localhost:~# 
```

6. From window 1, verify the currently available disk volumes on the SoC are visible:
   "lsblk" command



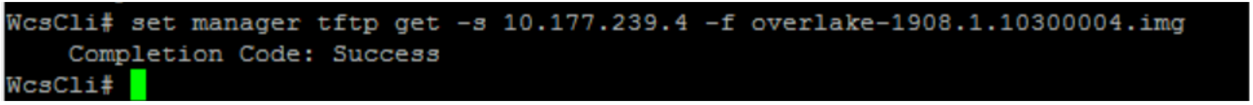
```
root@localhost:~# lsblk
NAME                MAJ:MIN RM  SIZE RO TYPE  MOUNTPOINT
loop0                 7:0    0   20K  0 loop
`-BootstrapApPki   252:0    0  1.3M  1 crypt /var/opt/msft/BootstrapApPki/mnt
loop1                 7:1    0  1.3M  0 loop
`-BootstrapApPki   252:0    0  1.3M  1 crypt /var/opt/msft/BootstrapApPki/mnt
mmcblk0             179:0    0 29.6G  0 disk
|-mmcblk0p1         179:1    0    8M  0 part
|-mmcblk0p2         179:2    0   32M  0 part  /vol/enmd
|-mmcblk0p3         179:3    0  128M  0 part
|-mmcblk0p4         179:4    0  128M  0 part
`-mmcblk0p5         179:5    0 29.3G  0 part  /vol/data
mmcblk0boot0        179:32   0    8M  1 disk
mmcblk0boot1        179:64   0    8M  1 disk
root@localhost:~# 
```

7. From window 2, verify the BMC version:
   "show system info -i <x>"
   verify 2.09.101 or later



```
WcsCli# show system info -i 21
    AssetTag: N/A
    Boot_BootSourceOverrideEnabled: True
    Boot_BootSourceOverrideTarget:  Force PXE
    ChassisId: N/A
    Description: N/A
    HostName: 172.17.0.81
    Id: 21
    Mac Address: 98:03:9B:7F:B8:9C
    Manufacturer: Microsoft
    Model: N/A
    Name: Server21
    PartNumber: M1103506-906
    SKU: N/A
    SerialNumber: P350690320004010
    Server:
        BMCVersion: C2030.BC.0903.116
        BiosVersion: C2030.BS.2A15.AE3
        CpldVersion: 00000005
```

8. From window 2, copy the install image from the TFTP server to the rack manager:
   "set manager tftp get -s <ip address> -f <imagename.img>"

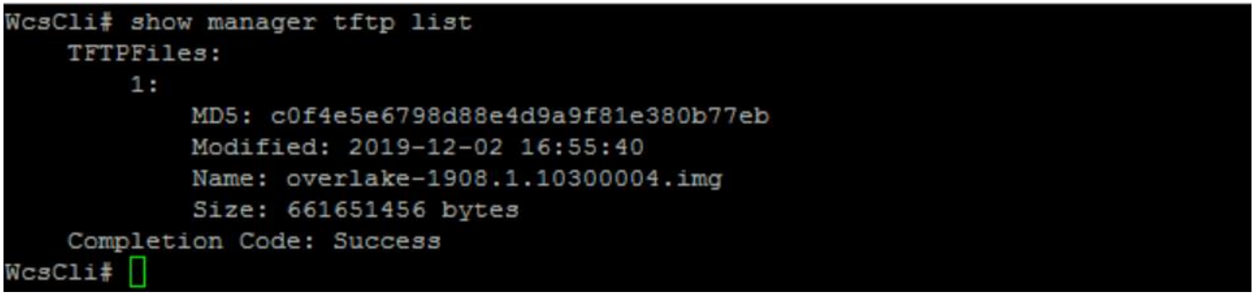


```
WcsCli# set manager tftp get -s 10.177.239.4 -f overlake-1908.1.10300004.img
    Completion Code: Success
WcsCli#
```

Note: this command may take several minutes to complete

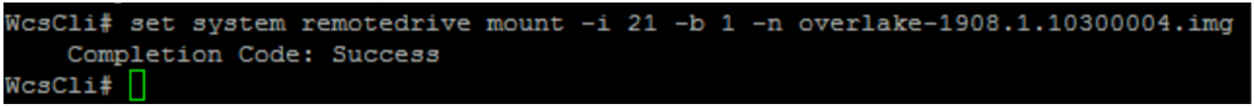
9. From window 2, verify that the file was copied over correctly:

   "show manager tftp list"



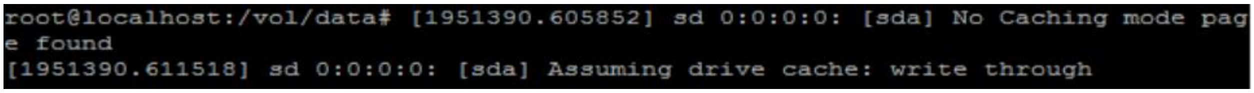
```
WcsCli# show manager tftp list
    TFTPFiles:
        1:
            MD5: c0f4e5e6798d88e4d9a9f81e380b77eb
            Modified: 2019-12-02 16:55:40
            Name: overlake-1908.1.10300004.img
            Size: 661651456 bytes
    Completion Code: Success
WcsCli#
```

10. From window 2, mount image on SoC:
   "set system remotedrive mount -i <x> -b 1 -n <imagename.img>"
   (should see mount activity on window 1)



```
WcsCli# set system remotedrive mount -i 21 -b 1 -n overlake-1908.1.10300004.img
    Completion Code: Success
WcsCli#
```

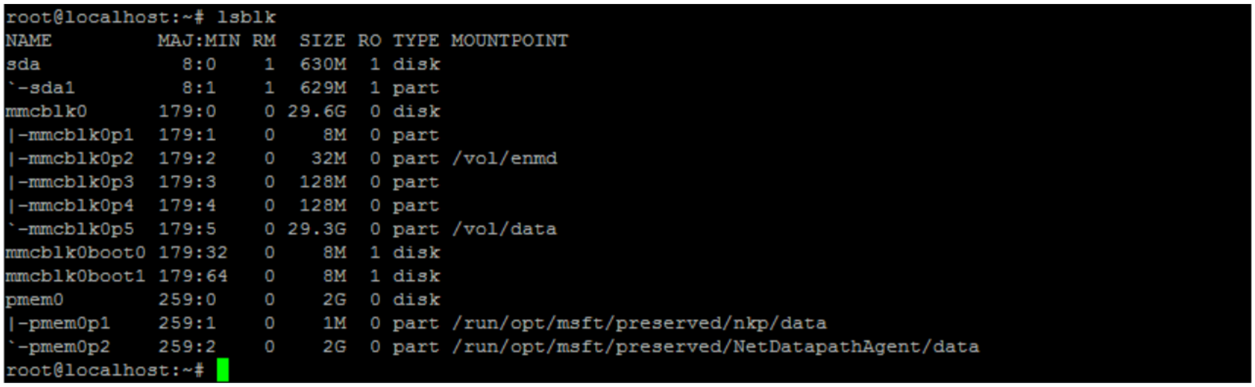
Window 2 ^



```
root@localhost:/vol/data# [1951390.605852] sd 0:0:0:0: [sda] No Caching mode pag
e found
[1951390.611518] sd 0:0:0:0: [sda] Assuming drive cache: write through
```

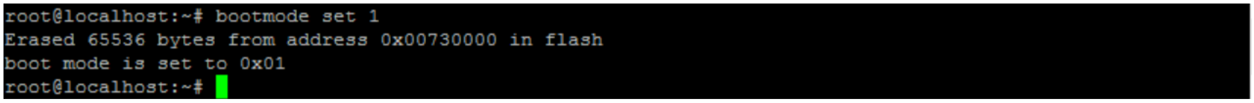
Window 1 ^

11. From window 1, again verify the available disk volumes:
"lsblk" command
Should now see "-sda" drive mounted



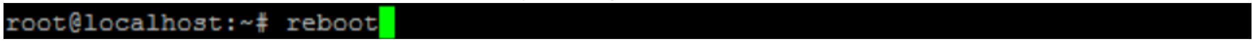
```
root@localhost:~# lsblk
NAME         MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
sda            8:0    1  630M  1 disk
`-sda1         8:1    1  629M  1 part
mmcblk0      179:0    0 29.6G  0 disk
|-mmcblk0p1  179:1    0    8M  0 part
|-mmcblk0p2  179:2    0   32M  0 part /vol/enmd
|-mmcblk0p3  179:3    0  128M  0 part
|-mmcblk0p4  179:4    0  128M  0 part
`-mmcblk0p5  179:5    0 29.3G  0 part /vol/data
mmcblk0boot0 179:32   0    8M  1 disk
mmcblk0boot1 179:64   0    8M  1 disk
pmem0        259:0    0    2G  0 disk
|-pmem0p1    259:1    0    1M  0 part /run/opt/msft/preserved/nkp/data
`-pmem0p2    259:2    0    2G  0 part /run/opt/msft/preserved/NetDatapathAgent/data
root@localhost:~#
```

12. From window 1, issue the command “bootmode set 1” to boot from USB

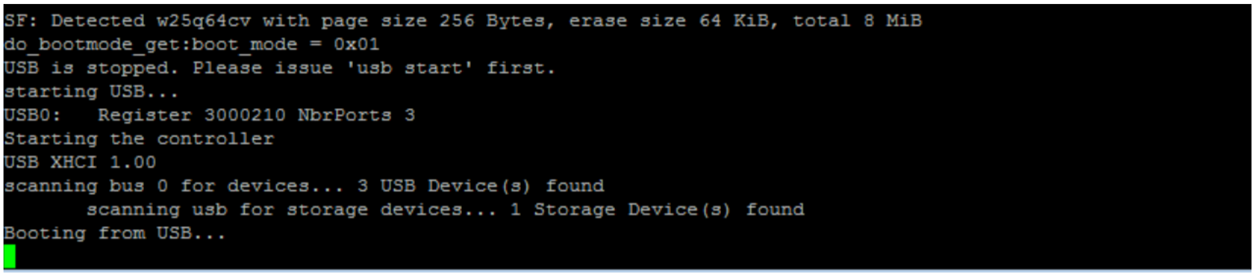


```
root@localhost:~# bootmode set 1
Erased 65536 bytes from address 0x00730000 in flash
boot mode is set to 0x01
root@localhost:~#
```

13. From window 1, reboot the SoC by issuing the “reboot” command



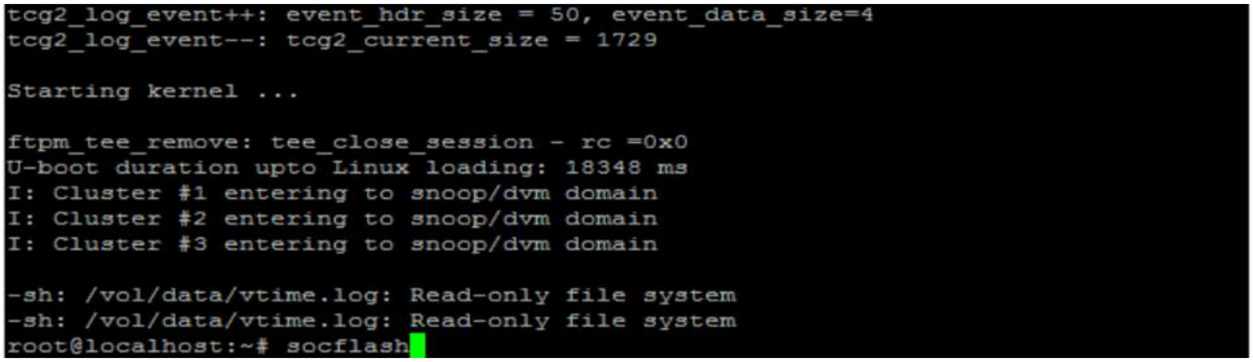
```
root@localhost:~# reboot
```



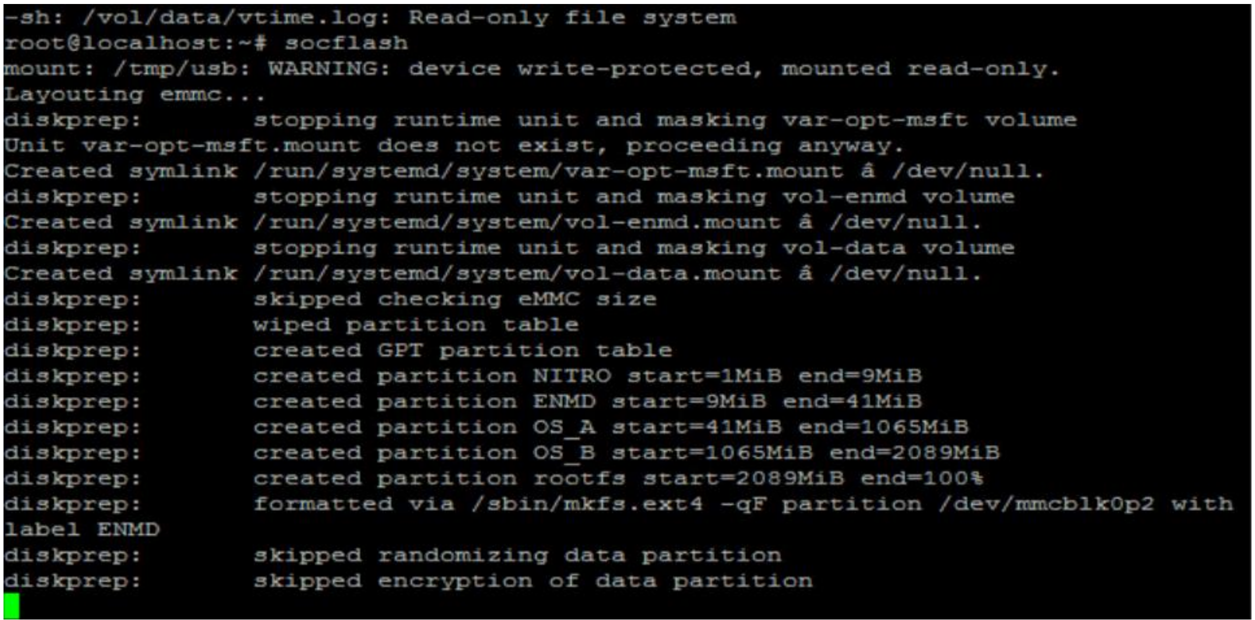
```
SF: Detected w25q64cv with page size 256 Bytes, erase size 64 KiB, total 8 MiB
do_bootmode_get:boot_mode = 0x01
USB is stopped. Please issue 'usb start' first.
starting USB...
USB0:   Register 3000210 NbrPorts 3
Starting the controller
USB XHCI 1.00
scanning bus 0 for devices... 3 USB Device(s) found
       scanning usb for storage devices... 1 Storage Device(s) found
Booting from USB...
```

This step will take a few minutes to get to the shell prompt.

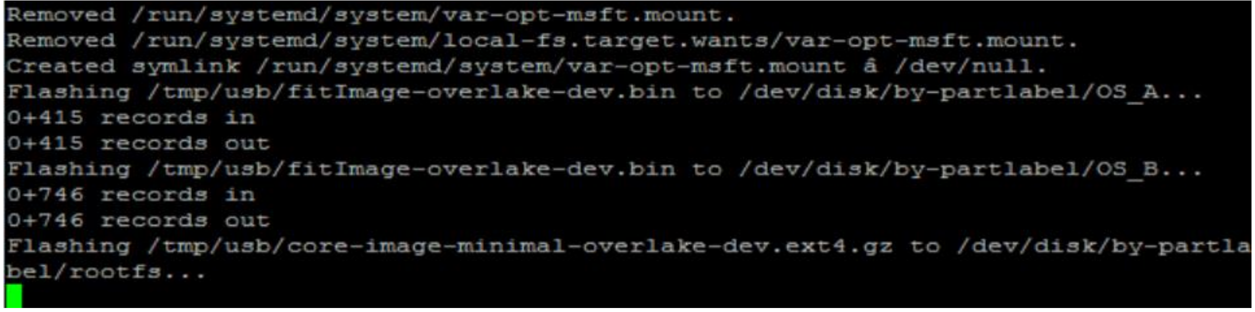
14. From window 1, after the reboot, update the SoC as usual from the booted drive by issuing the “socflash” command



```
tcg2_log_event++: event_hdr_size = 50, event_data_size=4
tcg2_log_event--: tcg2_current_size = 1729

Starting kernel ...

ftpm_tee_remove: tee_close_session - rc =0x0
U-boot duration upto Linux loading: 18348 ms
I: Cluster #1 entering to snoop/dvm domain
I: Cluster #2 entering to snoop/dvm domain
I: Cluster #3 entering to snoop/dvm domain

-sh: /vol/data/vtime.log: Read-only file system
-sh: /vol/data/vtime.log: Read-only file system
root@localhost:~# socflash
```

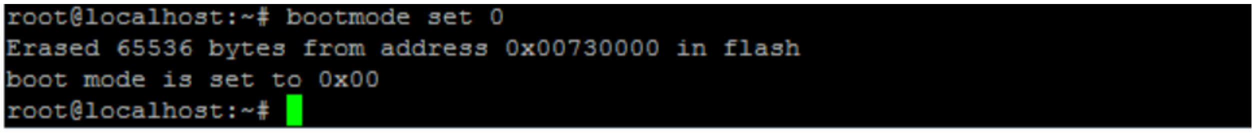


```
-sh: /vol/data/vtime.log: Read-only file system
root@localhost:~# socflash
mount: /tmp/usb: WARNING: device write-protected, mounted read-only.
Layouting emmc...
diskprep:       stopping runtime unit and masking var-opt-msft volume
Unit var-opt-msft.mount does not exist, proceeding anyway.
Created symlink /run/systemd/system/var-opt-msft.mount â /dev/null.
diskprep:       stopping runtime unit and masking vol-enmd volume
Created symlink /run/systemd/system/vol-enmd.mount â /dev/null.
diskprep:       stopping runtime unit and masking vol-data volume
Created symlink /run/systemd/system/vol-data.mount â /dev/null.
diskprep:       skipped checking eMMC size
diskprep:       wiped partition table
diskprep:       created GPT partition table
diskprep:       created partition NITRO start=1MiB end=9MiB
diskprep:       created partition ENMD start=9MiB end=41MiB
diskprep:       created partition OS_A start=41MiB end=1065MiB
diskprep:       created partition OS_B start=1065MiB end=2089MiB
diskprep:       created partition rootfs start=2089MiB end=100%
diskprep:       formatted via /sbin/mkfs.ext4 -qF partition /dev/mmcblk0p2 with
label ENMD
diskprep:       skipped randomizing data partition
diskprep:       skipped encryption of data partition
```



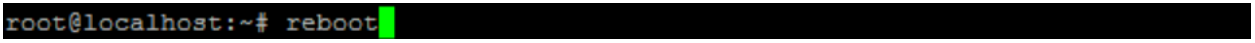
```
Removed /run/systemd/system/var-opt-msft.mount.
Removed /run/systemd/system/local-fs.target.wants/var-opt-msft.mount.
Created symlink /run/systemd/system/var-opt-msft.mount â /dev/null.
Flashing /tmp/usb/fitImage-overlake-dev.bin to /dev/disk/by-partlabel/OS_A...
0+415 records in
0+415 records out
Flashing /tmp/usb/fitImage-overlake-dev.bin to /dev/disk/by-partlabel/OS_B...
0+746 records in
0+746 records out
Flashing /tmp/usb/core-image-minimal-overlake-dev.ext4.gz to /dev/disk/by-partla
bel/rootfs...
```

15. From window 1, issue the command “bootmode set 0” to boot from eMMC



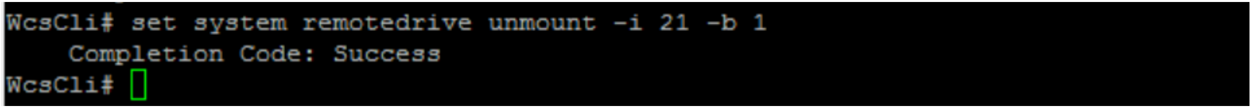
```
root@localhost:~# bootmode set 0
Erased 65536 bytes from address 0x00730000 in flash
boot mode is set to 0x00
root@localhost:~#
```

16. From window 1, reboot the SoC again to the new image, same command as above



```
root@localhost:~# reboot
```

17. From window 1, unmount the RM-based image to the SoC:
"set system remotedrive unmount -i <x> -b 1"



```
WcsCli# set system remotedrive unmount -i 21 -b 1
    Completion Code: Success
WcsCli#
```

15.3.2 Recovery from Host O/S

STEPS:

1. Ensure the A2040 internal USB and UART cables are connected to the Host Server. See **Figure 14** above.

2. Verify that the minimum recipe version pre-requisites have been met:

   - BMC: 2.09.101 or later

3. Run the Cerberus_utility from the Host to update the SoC FW image.
   General Commands:
   - ➢ Update FIP FW

   ```
   .\ cerberus_utility.exe -s 53 -m 70 -c 5 --slave socfwupdate 0 <FIP FW image file>
   ```

   - ➢ Update Nitro FW

   ```
   .\ cerberus_utility.exe -s 53 -m 70 -c 5 --slave socfwupdate 0 <Nitro FW image file>
   ```

4. After the updates have completed, reset the SoC by either AC cycling the system or by issuing a SoC reset from the Cerberus with the following command:
   - ➢ Reset SoC from Cerberus

   ```
   .\ cerberus_utility.exe -s 53 -m 70 -c 5 --slave socreset
   ```

5. Start a serial session from the Rack Manager to the SoC to observe the SoC console output after the firmware update to verify the update was successful.

15.4 FPGA Firmware Update

The FPGA firmware is stored in a dedicated QSPI flash device segmented into four “slots” which can store four distinct FPGA images. The slots are numbered from 0 through 3. Slot 0 is the “Failsafe” image which is what will load if the default (Golden) image fails to load. Slot 1 is the “Golden” image and the default image which the FPGA will load coming out of reset. Slot 2 and Slot 3 are the “Application 1” and “Application 2” images respectively which may be loaded after boot.

The Slot 0 / Failsafe image should not normally need to be updated after cards are received from manufacturing. If it does need to be updated, there are a couple of methods for doing so. One involves use of a USB Blaster cable and Intel Quartus software. The other method, in-system update, uses the standard FPGA tools, but does require that the FPGA is currently configured and operational.

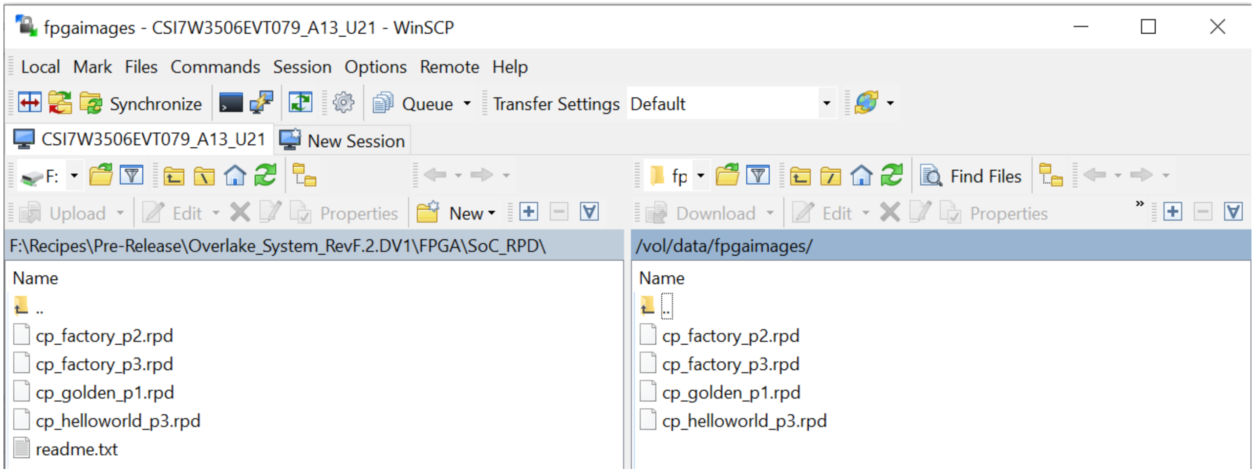
The Slot 1, 2 and/or 3 images may need to be updated when cards are received. These images are typically updated with the in-system method but can also be updated via the USB Blaster method as well.

The in-system method is described below. This update may only be performed from the Host or the SoC command line, depending the currently active firmware.* More specifically, if an older image is installed which supports flash write and reconfig from the Host-only, then the update should be done from the Host OS command line. If a newer image is installed which supports flash write and reconfig from the SoC-only, then the update should be done from the SoC side OS command line.

For the in-system update method, image slots may be updated individually or multiple at once by using the appropriate utility command line options. Both the individual and multiple update command steps are shown below. In addition to having a valid image loaded to the FPGA, the steps below require the FPGA drivers and tools have been loaded to the Host and SoC using the other update instructions included within this document.

15.4.1 FPGA In-System Update

1) Copy all the *.rpd files from the FPGA folder of the system recipe to the file system of either the Host or SoC (as appropriate, see above *).
Note: For the Host, this is intended to be copied over the existing network connection but may also be copied from a thumbdrive. For the SoC, this can be done using a program such as SCP or WinSCP over the network to that device.



fpgaimages - CSI7W3506EVT079_A13_U21 - WinSCP

F:\Recipes\Pre-Release\Overlake_System_RevF.2.DV1\FPGA\SoC_RPD\

Name
..
cp_factory_p2.rpd
cp_factory_p3.rpd
cp_golden_p1.rpd
cp_helloworld_p3.rpd
readme.txt

/vol/data/fpgaimages/

Name
..
cp_factory_p2.rpd
cp_factory_p3.rpd
cp_golden_p1.rpd
cp_helloworld_p3.rpd

Figure 15: Example of WinSCP where .rpd files are copied to SoC

2) Verify the slot and version of FPGA image currently loaded. Compare the values found with those specified in the full system recipe spreadsheet.
Note: Host-side commands are to be run from the command prompt inside Windows on the Host. SoC-side commands are to be run from the SoC Linux prompt which may be accessed via the Rack Manager or SoC network using SSH. If via RM, after SSH into the RM use command "start serial session -i <blade slot number> -b 1". If from SoC network, then direct SSH into the SoC using its IP address and the appropriate credentials. If any of the SoC-side commands fail, it may be due to drivers not yet loaded which may be remedied with the "modprobe catapult" command

General Command:

- Host-side (run from FPGA_Corelib_Drivers_Utilities\<version>\amd64\Tools directory):

```
./fpgadiagnostics.exe -justdumpregs
```

- SoC-side:

```
fpgadiagnostics -justdumpregs
```

Example of the register values to check to see which image slot and version is loaded currently. Look at full system recipe spreadsheet to verify register values for the recipe version you are attempting to check against. Release notes will include the git hash

Shell Registers – Golden Image
Register 010: 0x8273d489 (role git hash)
Register 056: 0x00000011 (DV1 Board ID)
Register 057: 0x00010004 (shell patch=1, board id = 4)
Register 058: 0x0003000b (Shell major=3, shell minor = 11)
Register 059: 0xc0006a18 (build info 10/3/2019)
Register 060: 0xf71a69ec (shell git hash)
Register 061: 0x00020000 (ASL version)
Register 064: 0x00bed70c (shell release id)
Register 065: 0xca7b030b (role version)
Register 101: 0xcaf1601d (role id) -- CAF1

OR

Shell Registers – Factory CAF2 (Host and SoC PCIe + Host and SoC Network)
Register 010: 0xe7f1dbd6 (role git hash)"
Register 056: 0x00000011 (DV1 Board ID)
Register 057: 0x00010004 (shell patch=1, board id = 4)
Register 058: 0x0003000b (Shell major=3, shell minor = 11)
Register 059: 0xc0006a18 (build info 10/3/2019)
Register 060: 0xf71a69ec (shell git hash)
Register 061: 0x00020000 (ASL version)
Register 064: 0x00bed70c (shell release id)
Register 065: 0xaaa1030b (role version)
Register 101: 0xcaf20fac (role id) -- CAF2

3) If appropriate, update the FPGA flash

EITHER

a. Multi-slot update

This can be accomplished by updating to the *.jic.rpd file copied to the system from the recipe. From two to four of the FPGA flash slots will be updated depending on the contents of the image file. Please reference the recipe to determine which slots are targeted by the image.

General Command:

- ➢ Host-side (run from FPGA_Corelib_Drivers_Utilities\<version>\amd64\Tools directory):

./fpgadiagnostics.exe -writeFlashJic <filepath>\<image filename>.jic.rpd

- ➢ SoC-side:

fpgadiagnostics -writeFlashJic <filepath>/<image filename>.jic.rpd

OR

b. Single slot update

This can be accomplished by updating to one of the *.rpd files copied to the system from the recipe. The Golden image can be found under the “FPGA\<version>\Golden-SoC-CAF1” directory as the file ending in “slotp1.rpd”. The Application 1 and 2 images can be found under the “FPGA\<version>\Factory-noECC-SoC-CAF2” as the files ending in “slotp2.rpd” and “slotp3.rpd” respectively.

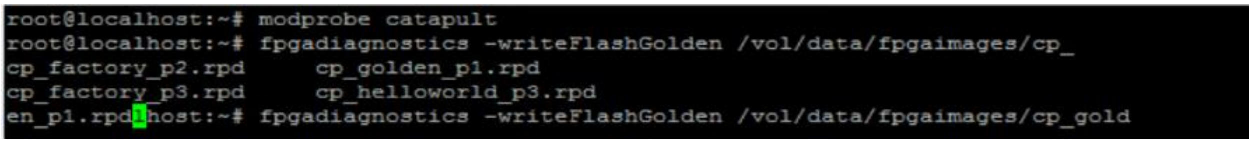
General Command (to update Golden):

- ➢ Host-side (run from FPGA_Corelib_Drivers_Utilities\<version>\amd64\Tools directory):

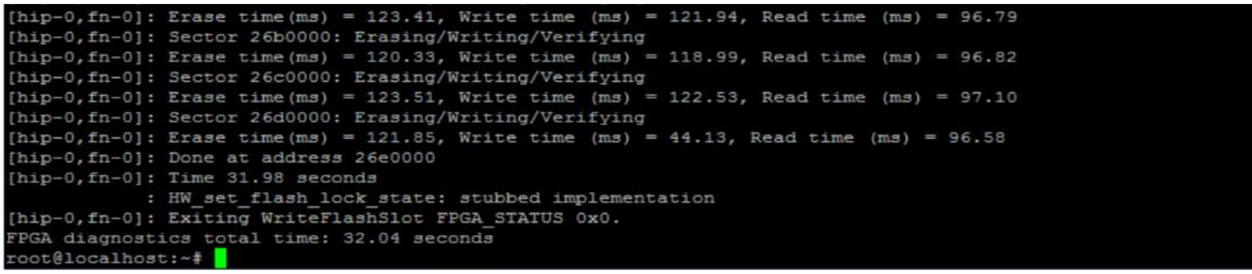
./fpgadiagnostics.exe -writeFlashGolden <filepath>\<image filename>slotp1.rpd

- ➢ SoC-side:

fpgadiagnostics -writeFlashGolden <filepath>/<image filename>slotp1.rpd



```
root@localhost:~# modprobe catapult
root@localhost:~# fpgadiagnostics -writeFlashGolden /vol/data/fpgaimages/cp_
cp_factory_p2.rpd      cp_golden_p1.rpd
cp_factory_p3.rpd      cp_helloworld_p3.rpd
en_p1.rpdlhost:~# fpgadiagnostics -writeFlashGolden /vol/data/fpgaimages/cp_gold
```



```
[hip-0,fn-0]: Erase time(ms) = 123.41, Write time (ms) = 121.94, Read time (ms) = 96.79
[hip-0,fn-0]: Sector 26b0000: Erasing/Writing/Verifying
[hip-0,fn-0]: Erase time(ms) = 120.33, Write time (ms) = 118.99, Read time (ms) = 96.82
[hip-0,fn-0]: Sector 26c0000: Erasing/Writing/Verifying
[hip-0,fn-0]: Erase time(ms) = 123.51, Write time (ms) = 122.53, Read time (ms) = 97.10
[hip-0,fn-0]: Sector 26d0000: Erasing/Writing/Verifying
[hip-0,fn-0]: Erase time(ms) = 121.85, Write time (ms) = 44.13, Read time (ms) = 96.58
[hip-0,fn-0]: Done at address 26e0000
[hip-0,fn-0]: Time 31.98 seconds
            : HW_set_flash_lock_state: stubbed implementation
[hip-0,fn-0]: Exiting WriteFlashSlot FPGA_STATUS 0x0.
FPGA diagnostics total time: 32.04 seconds
root@localhost:~#
```

General Command (to update App1):

- ➢ Host-side (run from FPGA_Corelib_Drivers_Utilities\<version>\amd64\Tools directory):

./fpgadiagnostics.exe -writeFlashSlot 1 <filepath>\<image filename>slotp2.rpd

- ➢ SoC-side:

fpgadiagnostics -writeFlashSlot 1 <filepath>/<image filename>slotp2.rpd

General Command (to update App2):

- ➢ Host-side (run from FPGA_Corelib_Drivers_Utilities\<version>\amd64\Tools directory): *./fpgadiagnostics.exe -writeFlashSlot 2 <filepath>\<image filename>slotp3.rpd*
- ➢ SoC-side: *fpgadiagnostics -writeFlashSlot 2 <filepath>/<image filename>slotp3.rpd*

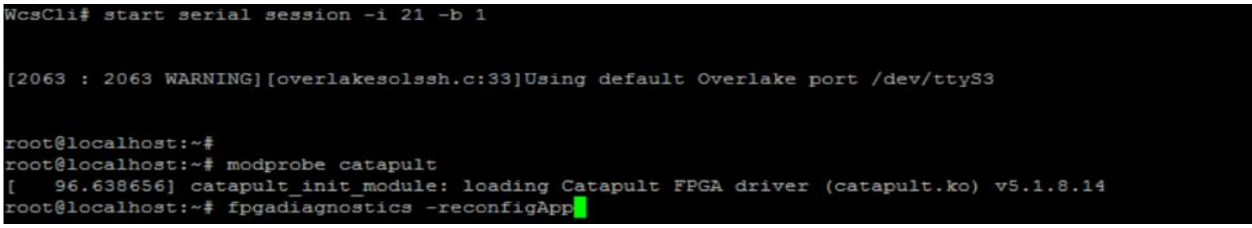
Note: the FPGA will retain its current configuration until the command is issued to reconfig or else the system is power cycled.

4) After the update, power cycle the system to load the new image (will default to Golden).

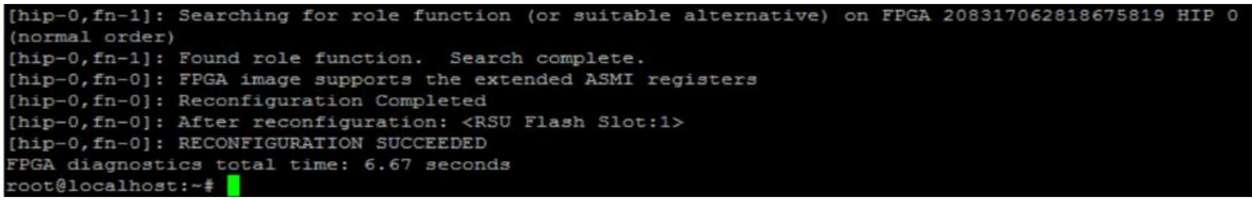
5) Verify the register values for the updated image(s) as in Step (2) above.
   a. Start with the Golden image, if that was part of the update, comparing against expected register values.

   b. Reconfigure to the Application images, if those were part of the update, comparing against expected register values.

   General Command (to reconfig to App1 from the SoC):
   fpgadiagnostics -reconfigApp



```
WcsCli# start serial session -i 21 -b 1


[2063 : 2063 WARNING][overlakesolssh.c:33]Using default Overlake port /dev/ttyS3


root@localhost:~#
root@localhost:~# modprobe catapult
[   96.638656] catapult_init_module: loading Catapult FPGA driver (catapult.ko) v5.1.8.14
root@localhost:~# fpgadiagnostics -reconfigApp
```



```
[hip-0,fn-1]: Searching for role function (or suitable alternative) on FPGA 208317062818675819 HIP 0
(normal order)
[hip-0,fn-1]: Found role function.  Search complete.
[hip-0,fn-0]: FPGA image supports the extended ASMI registers
[hip-0,fn-0]: Reconfiguration Completed
[hip-0,fn-0]: After reconfiguration: <RSU Flash Slot:1>
[hip-0,fn-0]: RECONFIGURATION SUCCEEDED
FPGA diagnostics total time: 6.67 seconds
root@localhost:~#
```

   General Command (to reconfig to App2 from the SoC):
   fpgadiagnostics -reconfigApp2

*Note: reconfiguring from an image which has the FPGA's Host PCIE endpoint configured & connected to the Host can result in an uncorrectable PCIE surprise down error on the Host. See **section 15.7** below to prevent this error during reconfiguration. This condition will not be seen when configuring from an image without a Host PCIE endpoint (e.g. golden)*

6) Recover and verify Host-side PCIE link to FPGA (for reconfig to non-Golden images)
   a. Host-side (run from FPGA_Corelib_Drivers_Utilities\<version>\amd64\Tools directory):
   ./fpgadiagnostics.exe -rootportrevivecp
   b. Host-side (run from FPGA_Corelib_Drivers_Utilities\<version>\amd64\Tools directory):

./fpgadiagnostics.exe -fpgadumphealth

Verify that dumphealth command works to report FPGA health information collected over the Host PCIE interface to the FPGA.

15.4.2 Verifying FPGA Image Checksum Offline

FPGA image checksums can be verified even before the image is loaded using the certutil.exe:

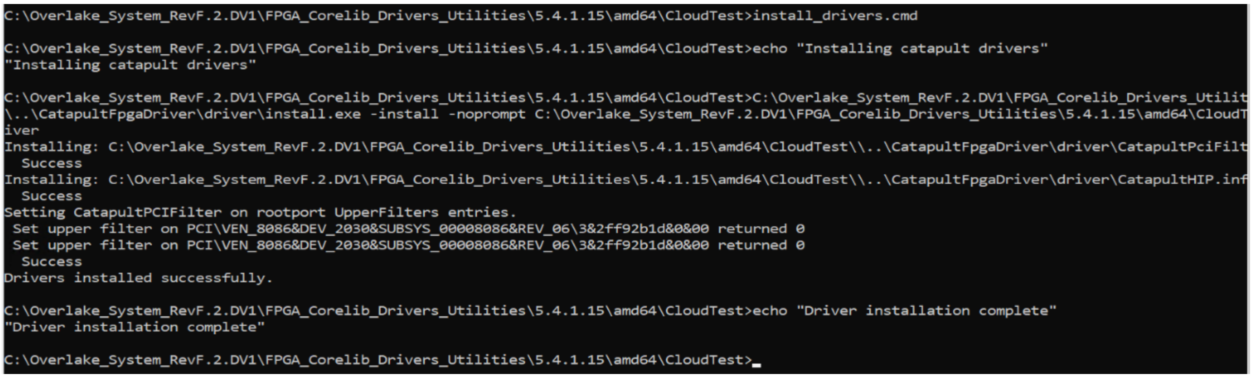
1. Get the JIC.RPD image and the expected SHA256 hash (the hash should be included in the release notes / documentation for the image).
2. Run "certutil.exe -hashfile C:\path\to\the_file.jic.rpd SHA256" in powershell.
3. Check that the hash generated by certutil matches the hash provided with the image. If they match the file is good, if not the file is bad.

15.5 Host Install/Update of FPGA drivers and tools

With the later releases of FPGA firmware and the associated drivers and utilities, the Host side will have the ability to access a subset of functionality to control and query the FPGA. To have this capability, the Windows software must be copied to the Host system and the drivers installed. The following steps describe how to do this.

1) If not already done, start by copying the recipe files to the Host. For this procedure, we will assume the entire recipe directory has been copied to C:\. For example:
   C:\Overlake_System_RevF.2.DV1\

2) The FPGA Catapult driver can be installed by running the following command from the directory FPGA_Corelib_Drivers_Utilities\version\amd64\CloudTest.
   install_drivers.cmd

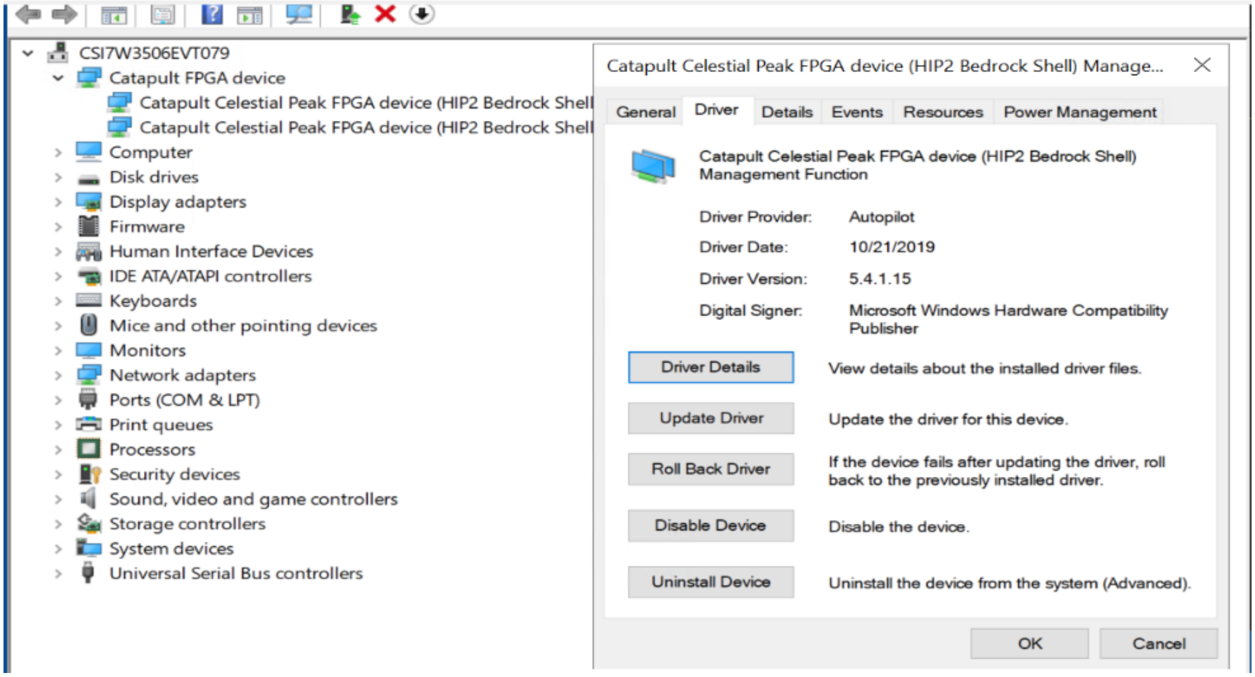


```
C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\CloudTest>install_drivers.cmd

C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\CloudTest>echo "Installing catapult drivers"
"Installing catapult drivers"

C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\CloudTest>C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilit
\..\CatapultFpgaDriver\driver\install.exe -install -noprompt C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\CloudT
iver
Installing: C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\CloudTest\\..\CatapultFpgaDriver\driver\CatapultPciFilt
  Success
Installing: C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\CloudTest\\..\CatapultFpgaDriver\driver\CatapultHIP.inf
  Success
Setting CatapultPCIFilter on rootport UpperFilters entries.
 Set upper filter on PCI\VEN_8086&DEV_2030&SUBSYS_00008086&REV_06\3&2ff92b1d&0&00 returned 0
 Set upper filter on PCI\VEN_8086&DEV_2030&SUBSYS_00008086&REV_06\3&2ff92b1d&0&00 returned 0
  Success
Drivers installed successfully.

C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\CloudTest>echo "Driver installation complete"
"Driver installation complete"

C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\CloudTest>_
```

3) Provided that the FPGA is already configured into a dual HIP image, such as the App1 or App2 slot, and that the PCIE link to the FPGA has been provisioned and re-trained, you can verify that the driver is installed to the correct version and working by verifying from the Device Manager in Windows.
   a. Right click on the Windows Start Icon and select Device Manager

b. From Device Manager, find and right click one of the Catapult devices listed then select properties and the Driver tab.

c. Verify that the Driver version listed corresponds with the one specified from the recipe you are updating to.



- CSI7W3506EVT079
  - Catapult FPGA device
    - Catapult Celestial Peak FPGA device (HIP2 Bedrock Shell
    - Catapult Celestial Peak FPGA device (HIP2 Bedrock Shell
  - Computer
  - Disk drives
  - Display adapters
  - Firmware
  - Human Interface Devices
  - IDE ATA/ATAPI controllers
  - Keyboards
  - Mice and other pointing devices
  - Monitors
  - Network adapters
  - Ports (COM & LPT)
  - Print queues
  - Processors
  - Security devices
  - Sound, video and game controllers
  - Storage controllers
  - System devices
  - Universal Serial Bus controllers

Catapult Celestial Peak FPGA device (HIP2 Bedrock Shell) Manage...

Catapult Celestial Peak FPGA device (HIP2 Bedrock Shell) Management Function

Driver Provider: Autopilot

Driver Date: 10/21/2019

Driver Version: 5.4.1.15

Digital Signer: Microsoft Windows Hardware Compatibility Publisher

Driver Details — View details about the installed driver files.

Update Driver — Update the driver for this device.

Roll Back Driver — If the device fails after updating the driver, roll back to the previously installed driver.

Disable Device — Disable the device.

Uninstall Device — Uninstall the device from the system (Advanced).

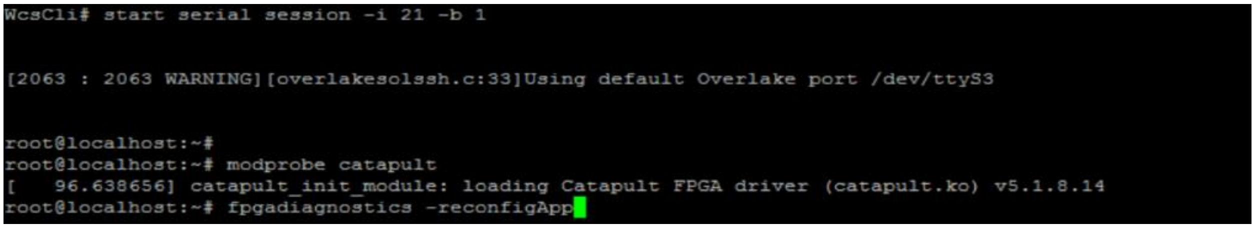
15.6 System Configuration to App Images

The following steps explain how to configure the system to one of the App image slots to enable the Host network access as well as Host access to the associated PCIE endpoint within the FPGA.

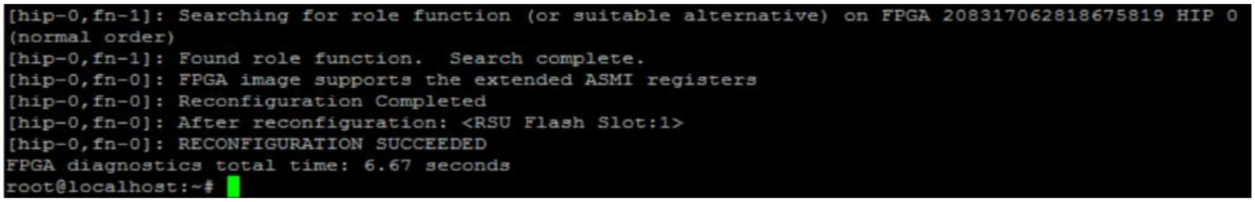
1) After initial boot from AC or DC cycle, reconfigure to one of the two application image slots.

General Command (to reconfig to App1 from the SoC):

fpgadiagnostics -reconfigApp



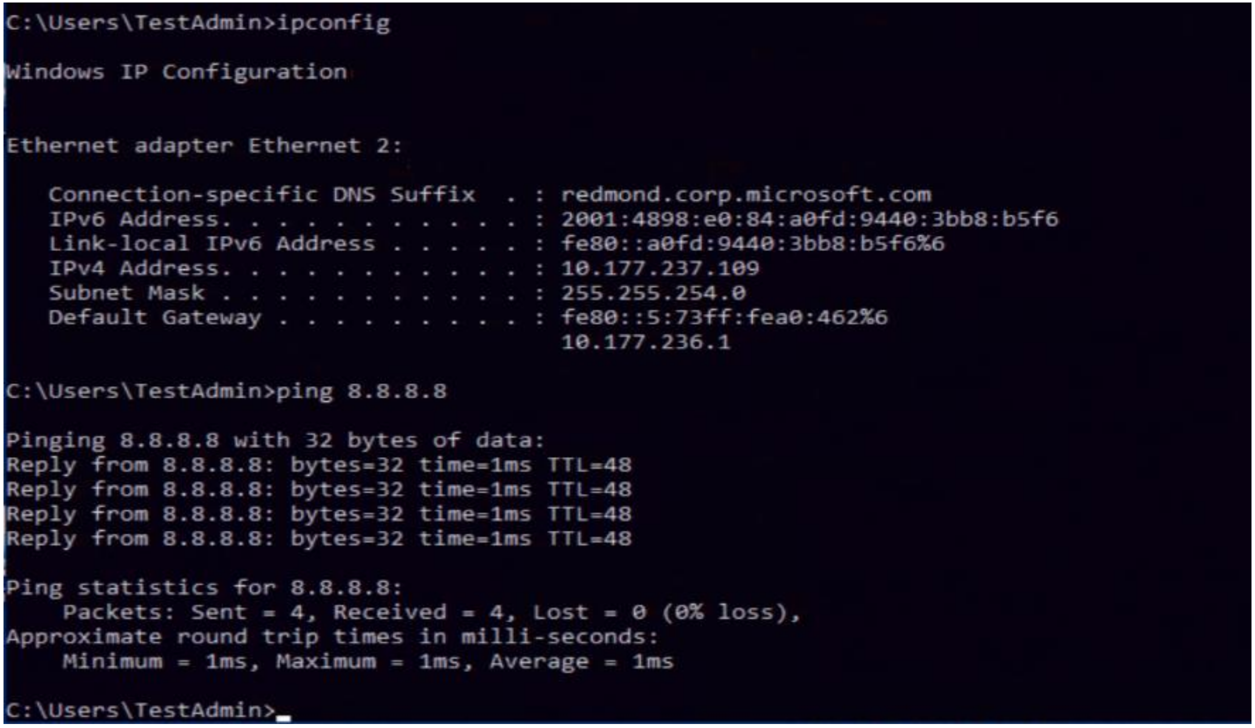
```
WcsCli# start serial session -i 21 -b 1

[2063 : 2063 WARNING][overlakesolssh.c:33]Using default Overlake port /dev/ttyS3

root@localhost:~#
root@localhost:~# modprobe catapult
[   96.638656] catapult_init_module: loading Catapult FPGA driver (catapult.ko) v5.1.8.14
root@localhost:~# fpgadiagnostics -reconfigApp
```



```
[hip-0,fn-1]: Searching for role function (or suitable alternative) on FPGA 208317062818675819 HIP 0
(normal order)
[hip-0,fn-1]: Found role function.  Search complete.
[hip-0,fn-0]: FPGA image supports the extended ASMI registers
[hip-0,fn-0]: Reconfiguration Completed
[hip-0,fn-0]: After reconfiguration: <RSU Flash Slot:1>
[hip-0,fn-0]: RECONFIGURATION SUCCEEDED
FPGA diagnostics total time: 6.67 seconds
root@localhost:~#
```

General Command (to reconfig to App2 from the SoC):
fpgadiagnostics -reconfigApp2

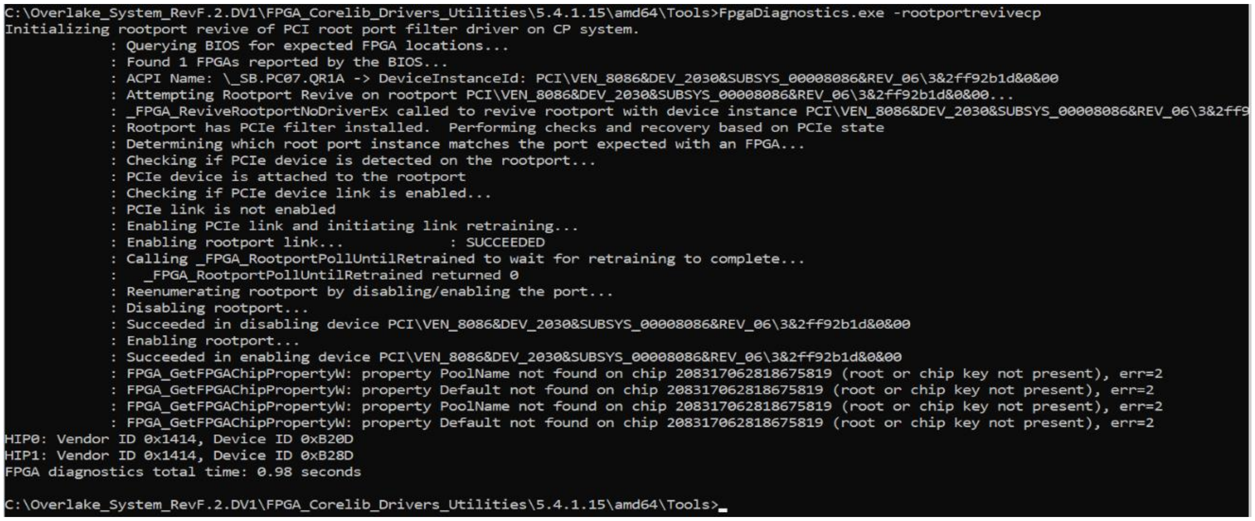
2) After the reconfiguration into one of the app slots, network access from the Host should be operational. You may confirm this by running the "ipconfig" command from a Host command prompt to verify it has an IP address, and then also by trying to ping out.



```
C:\Users\TestAdmin>ipconfig

Windows IP Configuration


Ethernet adapter Ethernet 2:

   Connection-specific DNS Suffix  . : redmond.corp.microsoft.com
   IPv6 Address. . . . . . . . . . . : 2001:4898:e0:84:a0fd:9440:3bb8:b5f6
   Link-local IPv6 Address . . . . . : fe80::a0fd:9440:3bb8:b5f6%6
   IPv4 Address. . . . . . . . . . . : 10.177.237.109
   Subnet Mask . . . . . . . . . . . : 255.255.254.0
   Default Gateway . . . . . . . . . : fe80::5:73ff:fea0:462%6
                                       10.177.236.1

C:\Users\TestAdmin>ping 8.8.8.8

Pinging 8.8.8.8 with 32 bytes of data:
Reply from 8.8.8.8: bytes=32 time=1ms TTL=48
Reply from 8.8.8.8: bytes=32 time=1ms TTL=48
Reply from 8.8.8.8: bytes=32 time=1ms TTL=48
Reply from 8.8.8.8: bytes=32 time=1ms TTL=48

Ping statistics for 8.8.8.8:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 1ms, Maximum = 1ms, Average = 1ms

C:\Users\TestAdmin>_
```

3) Once the Host network is verified, you may also restore access from the Host to the associated PCIE endpoint in the FPGA by retraining the link. This can be accomplished by running the following command from a command prompt on the Host.

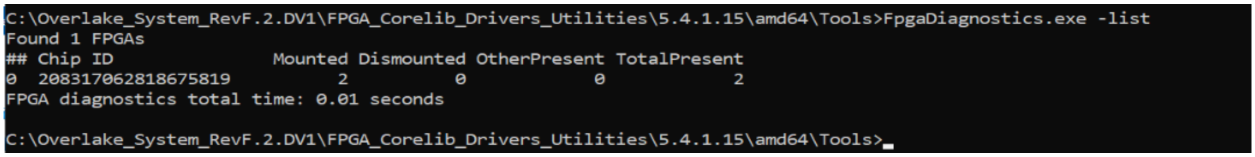
General Command:
fpgadiagnostics.exe -rootportrevivecp



```
C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\Tools>FpgaDiagnostics.exe -rootportrevivecp
Initializing rootport revive of PCI root port filter driver on CP system.
             : Querying BIOS for expected FPGA locations...
             : Found 1 FPGAs reported by the BIOS...
             : ACPI Name: \_SB.PC07.QR1A -> DeviceInstanceId: PCI\VEN_8086&DEV_2030&SUBSYS_00008086&REV_06\3&2ff92b1d&0&00
             : Attempting Rootport Revive on rootport PCI\VEN_8086&DEV_2030&SUBSYS_00008086&REV_06\3&2ff92b1d&0&00...
             : _FPGA_ReviveRootportNoDriverEx called to revive rootport with device instance PCI\VEN_8086&DEV_2030&SUBSYS_00008086&REV_06\3&2ff9
             : Rootport has PCIe filter installed.  Performing checks and recovery based on PCIe state
             : Determining which root port instance matches the port expected with an FPGA...
             : Checking if PCIe device is detected on the rootport...
             : PCIe device is attached to the rootport
             : Checking if PCIe device link is enabled...
             : PCIe link is not enabled
             : Enabling PCIe link and initiating link retraining...
             : Enabling rootport link...             : SUCCEEDED
             : Calling _FPGA_RootportPollUntilRetrained to wait for retraining to complete...
             :   _FPGA_RootportPollUntilRetrained returned 0
             : Reenumerating rootport by disabling/enabling the port...
             : Disabling rootport...
             : Succeeded in disabling device PCI\VEN_8086&DEV_2030&SUBSYS_00008086&REV_06\3&2ff92b1d&0&00
             : Enabling rootport...
             : Succeeded in enabling device PCI\VEN_8086&DEV_2030&SUBSYS_00008086&REV_06\3&2ff92b1d&0&00
             : FPGA_GetFPGAChipPropertyW: property PoolName not found on chip 208317062818675819 (root or chip key not present), err=2
             : FPGA_GetFPGAChipPropertyW: property Default not found on chip 208317062818675819 (root or chip key not present), err=2
             : FPGA_GetFPGAChipPropertyW: property PoolName not found on chip 208317062818675819 (root or chip key not present), err=2
             : FPGA_GetFPGAChipPropertyW: property Default not found on chip 208317062818675819 (root or chip key not present), err=2
HIP0: Vendor ID 0x1414, Device ID 0xB20D
HIP1: Vendor ID 0x1414, Device ID 0xB28D
FPGA diagnostics total time: 0.98 seconds

C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\Tools>_
```

4) After PCIE link to the FPGA is retrained and configured, you may verify access by verifying chip id from the Host side.

   General Command (verify chip id):
   fpgadiagnostics.exe -list



```
C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\Tools>FpgaDiagnostics.exe -list
Found 1 FPGAs
## Chip ID                Mounted Dismounted OtherPresent TotalPresent
0  208317062818675819           2          0            0            2
FPGA diagnostics total time: 0.01 seconds

C:\Overlake_System_RevF.2.DV1\FPGA_Corelib_Drivers_Utilities\5.4.1.15\amd64\Tools>_
```

At this point you should have normal functionality to the FPGA and network from both the Host and SoC side. CSIToolkit Test Requirements

15.7 Preventing Host Crash during Reconfiguration

When reconfiguring the FPGA from the SOC from an image with the Host PCIE endpoint exposed (for example, from the factory tester image) the Host PCIE endpoint hardware is reset. This reset disables the endpoint's PCIE link, which the Host system can report as a PCIE Surprise Link Down condition through AER. Windows (and many editions of Linux) will bugcheck/panic when the platform reports the Surprise Link Down condition. This occurs when reconfiguring the FPGA from the SOC OS or through the JTAG programmer if the FPGA's Host PCIE endpoint is connected to the Host.

You can prevent this error condition on the Host by taking steps to prepare the host for reconfiguration.

1) Before initiating the reconfiguration from the SOC, run the following command on the Host

   Fpgadiagnostics.exe -pgm -wait

This command will prepare the FPGA and then wait for a key press. While it is waiting, you can proceed to reconfig the FPGA as described in **section 15.4.1** above.

2) Once the reconfiguration is complete, press enter in the Host window running fpgadiagnostics. The tool will rescan the FPGA and reenable the Host devices.

This command has a provision to run a caller specified program after preparing the FPGA. When the program exits, the command proceeds as if the user had pressed enter. This can allow integration of the command into an automated workflow, by calling a program which signals automation on the SOC to proceed with reconfiguration and waits for the SOC to report that reconfiguration is complete. The syntax for this is:

Fpgadiagnostics.exe -pgm <command> <command arguments>

For example, to wait for 60 seconds, to allow for the SOC to perform reconfiguration, run:

Fpgadiagnostics.exe -pgm timeout.exe /T 60

In some situations, it may be acceptable and simpler to allow the Host OS to crash and reboot. In the event of the surprise down condition, Windows will crash with Bug Check 0x124: WHEA_UNCORRECTABLE_ERROR.

16 System Validation Requirements

Any new server design intending to support the A2040 card must conduct system validation in order to verify functionality within that system, especially for those features where the host server requires specific design functionality or modifications. Below is a set of test cases associated with features where host server design support is required. While the list below is intended to be fairly comprehensive, it is advised that the server designer evaluate their particular implementation in combination with the requirements from this document to assess for any other testing which may be appropriate. It is ultimately the responsibility of the designer to ensure all required features are functional and error scenarios are handled appropriately.

Additionally, rack integrators for systems containing the A2040 card must conduct testing to ensure that the card and related system features function properly within the rack environment. Rack integration testing should focus on those areas most dependent on correct physical and electrical integration of the hardware and with the correct firmware recipe installed. The list of test cases below also has a column to indicate which tests are appropriate for application within rack integration. Please note that rack integration tests may occur through a combination of SIP and manufacturing testing. Also, test times may be scaled from the test case defaults to suit the rack integration environment with a reasonable time allocation.

To help with both of these areas, Microsoft provides the “CSI ToolKit” software which contains automated system validation tests. The Toolkit helps to provide automation for many of the test cases for the A2040 Celestial Peak card.

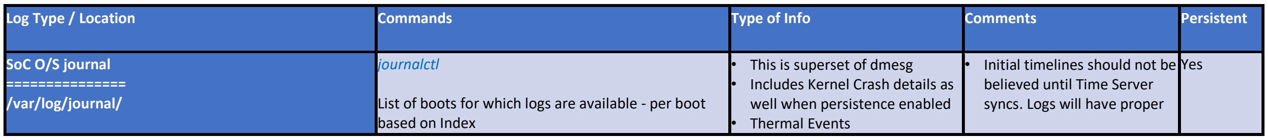
Test Case ID	Test Case Title	Rack Integration	New System Validation
27229	[Cerberus][FW] SoC FW failover		X
27000	[Cerberus][FW][Host] Inband (Host-side) AC loss in middle of CP Cerberus Update		X
58095	[Cerberus][FW][Host] Inband (Host-side) CP Cerberus FW update via Cerberus_utility*	X	X
27232	[Cerberus][FW][SoC] Inband (Host-side) SoC FIP and Nitro PFM update/query via Cerberus_utility*	X	X
100634	[Cerberus][FW][SoC] Inband (SoC-side) SoC FIP and Nitro PFM update/query via Cerberus_utility		X
57989	[Cerberus][FW][SoC][Host][RM] Inband (SoC & Host) and OOB (RM) SoC Boot Order Setting via Cerberus_utility		X
25975	[Cerberus][Reset][Host] Inband (Host-side) SOC reset via Cerberus_utility		X
27226	[EV-SE/SIT] [RM Provisions] Verify RM ability to report total CP card power		X
34847	[Network Test] [EE Test] Arista (TOR) Eye Scan test		X
34853	[Network Test] [EE]Network Electrical Eye at CX5		X
34852	[Network Test][EE] Network Electrical Eye at FPGA (NIC side)		X
34851	[Network Test][EE] Network Electrical Eye at FPGA (TOR side)		X
42618	[RM Provisions] Verify FPGA temp from Rack manager	X	X
27222	[Serviceability][Mechanical] Verify ease of access to install/remove card and associated cables		X
27190	[SoC][Debug][RM] Out-of-band (RM) SoC serial debug output test	X	X
27184	[SoC][FW][Network] PXE install of SoC O/S to eMMC via emmcflash and FW update via socflash*	X	X
92845	Cerberus FW update from in-band SoC O/S level utility	X	X
27191	CP FRU update utility test		X
26999	EMMC image stress - USB/BMC/RM update path	X	X
23481	FPGA Access via RM, Verify integrator can read the FPGA FRU information	X	X
26176	FPGA Configuration from SoC Jtag (JAM Player)		X
26167	FPGA Driver update Host*	X	X
26169	FPGA Device Disable/Enable Host – Devcon*	X	X
26172	FPGA Diagnostics dump health validation*	X	X
26171	FPGA Health and Registers validation*	X	X
42605	FPGA Reconfig from SoC to P0 Slot	X	X
42602	FPGA Reconfig from SoC to P1 Slot*	X	X
42603	FPGA Reconfig from SoC to P2 Slot*	X	X
42604	FPGA Reconfig from SoC to P3 Slot	X	X
26183	FPGA Write Flash from SoC - RPD Image - App - Slot P2*	X	X
26185	FPGA Write Flash from SoC - RPD Image - App2 - Slot P3*	X	X

26181	FPGA Write Flash from SoC - RPD Image - Golden - Slot P1*	X	X
82718	FPGA Writeflash from SoC - RPD Image - Jic all slots	X	X
25974	Full System Stress - FPGA Stress(power virus), SoC Stress and Host Stress at the same time*	X	X
25980	Host PCIe Correctable Error Injection		X
25979	Host PCIe Uncorrectable Error Injection		X
28261	Long Haul network stability tests	X	X
36453	OS Power Cycle - Host + SOC (with Full System Stress between each cycle)		X
57976	PCIE eye margin test for FPGA (from Host CPU)		X
57977	PCIE eye margin test for Host CPU (from FPGA)		X
27106	Power transient stress		X
28616	Power Validation with System Idle		X
27108	RM "AC" Power cycling (with full system stress between cycles)*	X	X
36460	RM "DC" Power cycling - Host + CP (with full system stress between cycles)*	X	X
27183	SoC boot cycling from USB		X
25985	SoC firmware Stress - USB/BMC/RM update path	X	X
25983	SoC Firmware Update Recovery after AC Power Loss - single image corruption		X
42601	SUT Validation against Recipe	X	X
28614	Temperature from ipmi and fpga registers with FPGA Stress from both Host and SoC		X
28617	Temperature from ipmi and fpga registers with System Idle		X
27225	Verify ability to read FPGA, SoC, Cerberus, eMMC FW versions from RM		X
92842	Verify Cerberus_utility update of Cerberus FW from Rack Manager	X	X
27194	Verify fan response and BMC logging with changing FPGA and SoC temps, also verify voltage and current behavior/logging		X
42620	Verify FPGA Asset Info from Rack manager	X	X
42616	Verify FPGA Health from Rack manager	X	X
94814	Verify SEL events for out of range Celestial Peak voltage supplies		X
93793	Verify that CP card can be installed and removed reasonably without tools required		X
92843	Verify update of SoC FW and PFM from RM with Cerberus_utility		X

Table 13: A2040 Validation and Rack Integration Test Requirements

16.1 Error and Informational Logs

For debug purposes, Error logs from the SoC and FPGA on the A2040 can be accessed using the methods described here:



Log Type / Location	Commands	Type of Info	Comments	Persistent
SoC O/S journal =============== /var/log/journal/	*journalctl* List of boots for which logs are available - per boot based on Index	• This is superset of dmesg • Includes Kernel Crash details as well when persistence enabled • Thermal Events	• Initial timelines should not be believed until Time Server syncs. Logs will have proper	Yes

	journalctl --list-boots Use the index value to capture specific boot logs and more after that boot - Uboot logs not included *journalctl -k* *journalctl -b* Give the logs from current boot *journalctl -p err -b* Show only entries logged at the error level or above.		time details only for logs generated after time sync.	
SoC PCIE AER data =============== FPGA: /sys/bus/pci/devices/0000\:01\:00.0/aer_* /sys/bus/pci/devices/0000\:01\:00.1/aer_* SoC NIC: /sys/bus/pci/devices/0008\:01\:00.1/aer_*	Read the File content	• PCIe AERs	• AER_Inject tool to emulate error injection for PCIe devices	No
CRMU Logs	Command to Generate the Dump: *rm -f /tmp/CRMU_Previous.txt && /opt/msft/sochwmon/crmu -m /tmp/memoryfile -b /tmp/crmu.dmp* Parser: */usr/sbin/error_logging/logparser_coarse.py crmu.dmp 0x8f113000 /usr/sbin/error_logging/logparser_fine.py stingray*	• Low Level boot logs from firmware • CRMU Dump • Gives more context of any resets of SoC	• First Use command to dump the file and then parse it either on the SoC Side or on your local machine. • It is good to collect CRMU dump for any failure immediately	No
Cerberus Debug Log	From Host, SoC or RM: *cerberus_utilty debuglogread*	• Cerberus Debug Log • PFM authentication failures		Yes
SoC Network Stats =============== /sys/class/net/enP8p1s0f0np0/statistics/	Read the file content (multiple files)		• Network Counters information from SoC	No
SoC Memory Error Count ============== /sys/devices/system/edac/mc/mc0/ce_count /sys/devices/system/edac/mc/mc0/ue_count	Read the file content	• Memory related errors and details	• DRAM Memory failure logs – correctables and uncorrectables	No
List of all SoC HW and SW devices =============== /dev	*ls /dev*		• Device Tree	Yes
Host BMC System Event Log (SEL)	From RM: *show system log read –i <x>* (where <x> is blade slot number)		• BMC SEL enabled	Yes

Table 14: A2040 Log Locations

16.1.1 Special Considerations for PCIe Correctable Errors reported during Testing

Allowable FPGA PCIe correctable error limits as reported in System Event Logs (SEL) or Windows Hardware Error Architecture (WHEA) logs are covered in the *Guideline on Correctable PCIe Error Threshold for Integrators* document M1042129.

The correctable PCIe error limit is a maximum of 20 entries per 24-hour period under any traffic condition. If test time is less than 24 hours, then use the guidelines below for determining the correctable error limit.

Test/Operating Time	Error Limit
24 hours	20
10 hours	10
7 hours	7
4 hours	4

Table 15: FPGA PCIe Correctable Error Limits

The limits presented above are for single FPGA devices in a system. For Multi-FPGA architectures, the error limit remains the same for each FPGA instance and the limits at PLX switches and root ports becomes the aggregate of the limits for the individual FPGA's on a given bus.   Example: If a Multi-FPGA system has 4 FPGA cards connected to a PLX switch then the PCIe correctable error limit is 4x the single card limit, or 80 FPGA correctable PCIe errors in 24 hours reported by the PLX switch.

16.2 OCSToolkit Test Requirements

Microsoft provides the OCSToolKit test software which includes required test scripts and suites to test the A2040 Celestial Peak card Some FPGA Servers and Appliances have specific functionality and/or features that require additional targeted test cases. All Servers and Appliances shall pass the OCSToolKit testing.

17 Reliability Requirements

All Cards must meet the annual failure rate target of <1% in 3 years' service under the conditions specified in in sections 17.1 and 17.2. Systems shall pass all reliability testing described in this Section 0

17.1 Operational environmental use condition inside data center

Operational data center environmental use conditions are defined in **Table 16** below. Environmental conditions, such as local temperature and relative humidity will vary with the placement and location of the component of interest. Reliability engineer is required to work with the design engineer to determine the local environmental conditions to set up the reliability test plan and to do the AFR calculation accordingly. For external module consumers: server reliability test metrics external to Microsoft should be designed by the criteria of the server supplier.

Environmental Conditions	Distribution	Medium	min/max
Data Center Temperature Distribution	Multimodal	26°C	18/35°C
Temperature rise	When in the bulkhead of C2010, 0°C T_{rise}, When in the exhaust of a JBOD, 25°C $T_{rise.}$		
Data Center Relative Humidity Distribution	multimodal	55%	27%/80%
Altitude	Sea level to 3000 meters		

Table 16: Operational Environmental Use Conditions

17.2 Non-operational environmental use conditions

Storage	-5 to 40 °C, 5% ~ 90% RH for six months
Transportation temperature	-40 <-> 75°C
Transportation temperature gradient	-40 °C at 30K feet to ground in 1 hour
Factory and data center dynamic: **shock** condition for unpackaged components and systems	< 12”(30cm) drop on any surface Or 32G shock
Factory and data center dynamic: **vibration** condition for unpackaged components and systems	< 1.87Grms

Table 17: Non-operational Environmental Use Conditions

17.3 Mechanical Reliability Test Requirements

Test Item	Spec./Criteria	Assy Level
Assembly and Mechanical check	Mechanical parts, Metal/Plastic parts. Assembly check, Operation check, Config check, Fitness check, Safety, Convenience and Maintenance check	Server
Rack Integration check	Install in rack to ensure proper interface with the rack and other components.	Rack
Product weight measurement	Use weight scale equipment to measure product including each module, device, cables, etc. In the maximum shipping configuration.	Server
Acoustic Test	Follow ECMA 74 standard to complete testing.	Server
Rack Level - (Operational) Unpackaged Office vibration test	1. Swept sine survey, 0.1g, 5 – 100 – 5Hz, 0.1 oct/min 2. Test for 3 axes 3. No permanent structural or mechanical damage 4. The equipment shall sustain operation without replacement of components, manual rebooting, or human intervention	Rack
Server Level - (Non-Operational) Packaged Handling drop test Single and hand-held bulk pack	1. 10 drops, 2 times. 2. Refer to ISTA 2A. 3. Shall not sustain any physical damage or deteriorate in functional performance when subjected to free-fall shock levels. 4.Dye and pry BGAs, 3 cards. No damage to BGA solder joints	Server

Server Level - (Non-Operational) Packaged Handling drop test Palleted bulk pack	1. 1 drop from 6". 2. Incline or horizontal impact for the 4 side faces. 3. Refer to ISTA 2B. 4. Shall not sustain any physical damage or deteriorate in functional performance when subjected to free-fall shock levels 5.Dye and pry BGAs, 3 cards. No damage to BGA solder joint.	Server
Server Level - (Non-Operational) Packaged Transportation vibration test Single and hand-held bulk pack	1. Test level 1.15Grms. 2 sequences on 4 faces for 1 hour. 2. Refer to ISTA 2A 3. Equipment shall not sustain any physical damage or deteriorate in functional performance when subjected to vibration levels expected during transportation 4.Dye and pry BGAs, 3 cards. No damage to BGA solder joint	Server
Server Level - (Non-Operational) Packaged Transportation vibration test Palleted bulk pack	1. Test level 1.15Grms. 2 sequences on 1 face for 1 hour. 2. Refer to ISTA 2B 3. Equipment shall not sustain any physical damage or deteriorate in functional performance when subjected to vibration levels expected during transportation 4.Dye and pry BGAs, 3 cards. No damage to BGA solder joint	Server
Bulk Pack Compression	Compression testing of packs per ISTA 2A/2B. Tests the stacking strength of corrugated packaging.	Server
Rack Level - Package Test	Ensure rack packaging fits without modification and will protect equipment.	Rack
Rack Level - (Non-Operational) Packaged Handling drop test	1. Drop test on the normal rest surface for 2 drops 2. Drop height: 100mm (3.9in) 3. Instrument rack with accelerometers to capture Server amplification. Sensor locations to be determined with MS prior to testing. 4. Shall not sustain any physical damage or deteriorate in functional performance when subjected to free-fall shock levels 5. The rack must be populated in the expected shipping configuration with all necessary server, power distribution, networking gear, and cabling. 6.Dye and pry BGAs, 3 cards. No damage to BGA solder joint	Rack
Rack Level - (Non-Operational) Packaged Transportation vibration test	1. Test level - Common carrier: 1.146Grms 2. Test Duration: 1 hour on the normal rest surface (bottom side) 3. Instrument rack with accelerometers to capture Server amplification. Sensor locations to be determined with MS prior to testing. 4. Equipment shall not sustain any physical damage or deteriorate in functional performance when subjected to vibration levels expected during transportation 5. The rack must be populated in the expected shipping configuration with all necessary server, power distribution, networking gear, and cabling. 6.Dye and pry BGAs, 3 cards. No damage to BGA solder joint	Rack

Rack Level - (Non-Operational) Packaged Incline/Horizontal Impact	1. Depending on test equipment available either incline or horizontal impact testing per ISTA 3E. 2. Impact to all 4 vertical surfaces at velocity of 1.1 meter/second. 3. Instrument rack with accelerometers to capture server amplification. 4. Equipment shall not sustain any physical damage or deteriorate in functional performance when subjected to vibration levels expected during transportation. 5. Full functionality test required before and after the test. 6. The rack must be populated in the expected shipping configuration with all necessary server, power distribution, networking gear, and cabling.	Rack

Table 18: Mechanical Test Requirements

17.4 Thermal Reliability Test Requirements

Test Item	Spec./Criteria	Assy Level
Server Impedance Measurement	Collect impedance level of different server configurations and verify the impedance requirement.	Server
IR Scan	Locating thermal hot spot on the motherboard to ensure system cooling can cover the thermal issue. Can be performed in conjunction with the thermal profile test.	Server
Sensor correlation test	Verify the reading sensor accuracy with respect to the thermocouple real time measurement.	Server
Failed Fans	Determine worst fan failure scenario and verify sufficient cooling is available.	Server
Thermal profile test	Collect the critical components temperature data with various fan speed control (manually controlled) and various inlet ambient temperature for fan speed control table reference. Minimum set of temperatures: 25C, 30C, 35C	Server
Thermal verification test	Collect the critical component temperature with auto fan speed control under various inlet ambient temperature.	Server
PSU thermal test	Verify PSU thermal condition in worst case configuration.	Server
Transient test	Step power levels to verify sufficient cooling of components. Ramp ambient temperatures to verify sufficient cooling of components.	Server
Power and Airflow testing	Capture power and airflow at defined inlet temperatures and utilization rates. Refer to the latest version of Microsoft's Power Draw Test Tools document for details.	Server

Table 19: Thermal Test Requirements

17.5 PCBA Reliability Test Plan

PCBA Reliability Test plan and System Reliability Test plan have been divided into the next two sections of this doc. Server suppliers who design and manufacture PCBAs for MSFT Cloud Servers shall complete the Visual and Functional Tests described in **Table 20**. Results shall be reviewed with Microsoft.

The tests in the test plan is grouped into component aging in temperature and humidity, failure due to static and dynamic electrical and mechanical stresses, weakest link by highly accelerated life test (HALT), and design landing zone assessment by calculated mean time between failure (CMTBF). Reliability engineer needs to work with the development engineering team and program management to use failure mode and effect analysis (FMEA) tool to identify high risks in design and manufacturing. Based on the FMEA grading, the reliability engineer will need to define a test plan to address the high risks and to discover the weakest link. This should be done in the early phase of development cycle, such as proof of concept (POC), and EVT to give sufficient lead time to design out the risks. The DV test plan is to verify the product reliability through DV design and early manufacturing.

The reliability tests need 46EV samples, 136 DV samples. The sample size is the quantity of each PCBA the vendors design and manufacture for MSFT CSI. Due to long lead time, MTBF demonstration test shall take the highest priority. PV sample qty depends on the results of DV. Any PV tests is to address, mitigate and close the risks persisting through DV.

#	Reference	Test Description	Sample	Testing Conditions	Pass/Fail Criterion	Sample Quantity	
						EV	DV
1	JESD22-A104	Thermal Cycling	PCBA	1. Non-operating 2. -40C~+100C, ramping 10-15°C/min, peak soaking 15mins. 3. Functional check at 200/300/400/500/600cyc. 4. DnP (FPGA, SoC, DRAM) on DUTs failing functional test or at 600cyc.	1) Cosmetic check including BGA glue delamination. 2) No broken components/parts. 3) Pass fucntional tests. 4) Allowable solder joint cracking area at DnP: - NCTF: Up to 100%. - CTF: Up to 75%.	12	12
2	JESD22-A119/A103 M1071141	Cold & Hot Storage	PCBA	1. Non-operating. 2. -40C for 168hrs then functional check. 3. +70 for 168hrs then functional check.	1) Cosmetic check including BGA glue delamination. 2) No broken components/parts. 3) Pass fucntional tests.	6	6
3	JESDA105/A108 /A122	Cold Soak & Boot Up	PCBA	1. -5C. 2. Soak for 2hrs then power on to full load with data transfer for 10min. 3. Repeat step 2 for 10 cycles. 4. Functional check at room-T.	1) Cosmetic check including BGA glue delamination. 2) No broken components/parts. 3) System can boot normally. 4) Pass fucntional tests.	2	
4	JESD22-A100/A105/A108/A122 M1071141	Cold & Hot Operating	PCBA	1. Power cycling with 3.5min-on & 0.5min-off per cycle and 1 AC power cycling after 1 DC power cycling (reboot). Full load at On with data transfer. 2. Dwelling at 25C for 1hr. 3. Ramping at 10C/hr to 0C w/o condensation. 4. Dwelling at 0C for 24hr. 5. Ramping to 25C/85%RH at 10C/hr & 20%RH/hr w/o condensation. 6. Dwelling at 25C/85%RH for 24hr. 7. Ramping to 40C/20%RH at 10C/hr & 20%RH/hr w/o condensation. 8. Dwelling at 40C/20%RH for 24hrs. 9. Ramping to 40C/85%RH at 20%RH/hr w/o condensation.	1) Cosmetic check including BGA glue delamination. 2) No unwaived running errors during the test. 3) FPGA healthdump before & after the test.		

				10. Dwelling at 40C/85%RH for 24hrs. 11. Ramping to 25C/50%RH at 10C/hr & 20%RH/hr w/o condensation. 12. Functional check at room-T.			
5	JESD22-A101	High Temperature/Humidity	PCBA	1. Non-operating 2. 85C/85%RH/168hrs 3. Functional check at 72/168hrs	1) Cosmetic check including BGA glue delamination. 2) No component/part corrosion. 3) Pass fucntional tests.	8	16
6	JESD22-A101/A105/A108/A122	High Temperature/Humidity with AC/DC Power Cycling	PCBA	1. Power cycling with 3.5min-on & 0.5min-off per cycle and 1 AC power cycling after 1 DC power cycling (reboot) for totally 500cyc (~1.5 day). Full load at On with data transfer. - Ramping to 0C w/o condensation while DUT off. - Power cycling at 0C for 250cyc. - Ramping to 40C/85%RH w/o condensation while DUT power cycling. - Power cycling at 40C/85%RH for 250cyc. 2. Ramping to 60C/85%RH w/o condensation while DUT at full load with data transfer. 3. Continuous run at 60C/85%RH for 1000hrs at full load with data transfer (~6wk). 4. Power cycling with 3.5min-on & 0.5min-off per cycle and 1 AC power cycling after 3 DC power cycling (reboot) for totally 500cyc (~1.5 day). Full load at On with data transfer. - Ramping to 40C/85%RH w/o condensation while DUT at full load with data transfer. - Power cycling at 40C/85%RH for 250cyc. - Ramping to 0C w/o condensation while DUT power cycling. - Power cycling at 0C for 250cyc. 5. Ramp to the room-T for functonal check. 6. Temperature ramping 1~2C/min & moisture ramping 1~2%RH/min.	1) Cosmetic check including BGA glue delamination. 2) Pass all built-in functional check. 3) No unwaived running errors during the test. 4) Reboot allowed but record all log files when DUT down. 5) FPGA healthdump before & after the test.	8	48

7	JESD51 M1071141	Thermal Profiling	PCBA	1. Operating. 2. Acquire temperature of FPGA and SoC from internal sensors. 3. Attach thermal coupons on key components (e.g. DRAM, eMMC, Flash, VR, E-cap, HS, air inlet, air exhaust, PCB, etc.) to meaure their temperature. 4. Apply manually (EV) or automatically (DV) controlled air flow. 5. Air inlet temperature 25C, 35C, 40C, 60C.	1) At the normal opertaing condtion (inlet 25/35C), all components need to meet the derating requirements. 2) At the reliability testing conditons (inlet 40/60C), measurement is for characterization purpose.	2	2
8	IPC-9592	Derating analysis	PCBA	Leverage thermal profiling and electrical measurement.	All components need to meet the derating spec.		
9	Per Supplier	E-capacitor Life	PCBA	Leverage thermal profiling and electrical measurement.	All E-capacitors need to work for 5yr at the worst case.		
10	SR332 Issue 3/4 JESD85	CMTBF	PCBA	1. Leverage thermal profiling and electrical measurement. 2. Critical active components (IC, MOSFET, diodes, VR, etc) and connectors shall use suppliers reliability testing and/or field data at 90% confidence. 3. Non-critical active components can leverage the mean failure rates of SR332. 4. Other passive components shall leverage the mean failure rates of SR332. 5. Inlet temperature 25C and 35C.	Predicted AFR at Inlet temperature of 25C shall meet the design target assuming operating for 8760hr/yr.		
11		Temperatrue Voltage Margin Test	PCBA	1. Tune the 12V VDC, VR output to FPGA/SoC/DDR to be +/-8% or the design spec in the test. 2. For each voltage condition, test the board at both 0C and 40C. 3. All VR component suppliers shall be covered.	Check VR margin shall meet or exceed the design spec.	2	2
12	M1156514	HALT (LOL/UOL/VOL etc.)	PCBA	Operating per HALT spec.	Characteriation to understand the operating limits and weak links in design.	2	
13	IPC-JEDEC-9702/9704A IPC-WP-011	Strain Test	PCBA	1. Non-operating spherical bending. 2. May leverage the strain test for line characterization.	Board strain near BGA corner balls shall be <450ue.	1	1

14	EIA-364	Assembly & Mechanical Check	PCBA	1. Mating/unmating all connectors/cables (power, QSFP, edge finger, etc) 2. Functional check at 15/30/40/50cyc.	1) All mating/unmating cycle >50cyc w/o failure. 2) All mating/unmating force shall meet ergo spec.	2	2
15	MIL-STD-810G 516.6 IPC-JEDEC-9703 JESD22-B104/B110 M1071141	Mechanical Shock of Bare PCBA	PCBA	1. Non-operating. 2. Square wave 32G with pulse width ~25ms velocity 6.85m/s. 3. One drop at each of 6 orthogonal orientations (+X/-X/+Y/-Y/+Z/-Z).	1) Cosmetic check including BGA glue delamination. 2) No broken components/parts. 3) Pass fucntional tests. 4) DnP after S&V and no solder joint cracking allowed.	24	24
16	MIL-STD-810G 514.6 JESD22-B103 M1071141	Random Vibration of Bare PCBA	PCBA	1. Non-operating. 2. 10-500Hz, random 1.87Grms 10Hz $0.13G^2$/Hz 20Hz $0.13G^2$/Hz 70Hz $0.004G^2$/Hz 130Hz $0.004G^2$/Hz 165Hz $0.0018G^2$/Hz 500Hz $0.0018G^2$/Hz 3. Test 15mins at each of 3 orthogonal orientations (X/Y/Z)			
17	ISTA 2A (Drop) ASTM D4169-05/16	Mechanical Shock of Single-Pack Package	PCBA	1. Non-operating. 2. Refer to ISTA 2A (Drop)	1) Cosmetic check including BGA glue delamination. 2) No broken components/parts.		3

18	ISTA 2A (Vibration) ASTM D4169-05/16	Random Vibration of Single-Pack Package	PCBA	1. Non-operating. 2. Test level: 10-250Hz, 1.146Grms 1Hz 0.0001G^2/Hz 4Hz 0.01G^2/Hz 100Hz 0.01G^2/Hz 200Hz 0.001G^2/Hz 3. Test Duration: Total 1 hours Primary sitting package face for 30min The other two orthognal package faces for 10min/ea Opposite to the primary sitting package faces for 10min.	3) Pass fucntional tests. 4) DnP after S&V and no solder joint cracking allowed.		
19	ISTA 2A (Drop) ASTM D4169-05/16	Mechanical Shock of Multi-Package	PCBA	1. Non-operating. 2. Refer to ISTA 2A (Drop)	1) Cosmetic check including BGA glue delamination. 2) No broken components/parts. 3) Pass fucntional tests. 4) DnP after S&V and no solder joint cracking allowed.		6
20	ISTA 2A (Vibration) ASTM D4169-05/16	Random Vibration of Multi-Package	PCBA	1. Non-operating 2. Test level: 10-250Hz, 1.146Grms 1Hz 0.0001G^2/Hz 4Hz 0.01G^2/Hz 100Hz 0.01G^2/Hz 200Hz 0.001G^2/Hz 3. Test Duration: Total 1 hours Primary sitting package face for 30min The other two orthognal package faces for 10min/ea Opposite to the primary sitting package faces for 10min.			

Table 20: PCBA Reliability Test Plan

17.6 System Reliability Test Plan

The System reliability test plan in this section is for suppliers who design and manufacture L10 systems for Microsoft. Suppliers shall complete all visual and functional tests shown in **Table 21**. All results shall be reviewed with Microsoft.

If the PCBAs used are designed by the same system suppliers, the PCBAs shall be subjected to the PCBA Reliability tests specified in **section 17.5**. If the PCBAs are supplied by a third party, the system suppliers shall request the PCBA vendors to successfully complete the PCBA reliability tests specified in **section 17.5**.

Sequential reuse of the system samples is permitted to reduce sample costs. The sample size in parenthesis suggests a reuse scheme. Vendor can decide their own reuse schemes. PV sample size and retest depends on DV results. Suppliers shall confirm with Microsoft the qty of PV samples after providing EV and DV test results

#	Reference	Test Description	Sample	Testing Conditions	Pass/Fail Criterion	Sample Quantity	
						EV	DV
1	ISTA 2A (Drop) ASTM D4169-05/16	Mechanical Shock in a Un-Packaged Blade	L10	1. Square-Wave of 15G/ for 28ms with velocity change of 4.19m/s. 2. Once per side of the 6 orthogonal sides.	1) Cosmetic check including BGA glue delamination 2) No component unseating or damaged. 3) Pass system functional check.		3
2	ISTA 2A (Vibration) ASTM D4169-05/16	Random Vibration in a Un-Packaged Blade	L10	1. 10-250Hz uniform spectrum, 1.54Grms 2. 30 minutes per orientation of the 3 orthogonal orientations (X/Y/Z)			
3	ISTA 2A (Drop) ASTM D4169-05/16	Mechanical Shock in a Single-Blade Package	L10	Refer to ISTA 2A (Drop)	1) Cosmetic check including BGA glue delamination 2) No component unseating or damaged. 3) Pass system functional check.		3
4	ISTA 2A (Vibration) ASTM D4169-05/16	Random Vibration in a Single-Blade Package	L10	1. Test level: 10-250Hz, 1.146Grms 1Hz 0.0001G^2/Hz 4Hz 0.01G^2/Hz 100Hz 0.01G^2/Hz 200Hz 0.001G^2/Hz 2. Test Duration: Total 1 hours Primary sitting package face for 30min The other two orthognal package faces for 10min/ea Opposite to the primary sitting package faces for 10min.			
5	ASTMD4169-05	Drop Test in a Packaged Rack	L11	1. Drop test on the normal rest surface for 2 drops 2. Drop height: 2 inches 3. Instrument rack with accelerometers to capture blade amplification. Sensor locations to be determined with MS prior to testing. 4. Shall not sustain any physical damage or deteriorate in functional performance when subjected to free-fall shock levels	1) Cosmetic check including BGA glue delamination 2) No component unseating or damaged. 3) Pass system functional check.		3

6	ISTA 2B	Random Vibration in a Packaged Rack	L11	1. Test level - 1.146Grms, profile defined in ISTA 2B 2. Test Duration: 1 hour in the vertical direction 3. Instrument rack with accelerometers to capture blade amplification. Sensor locations to be determined with MS prior to testing. 4. Equipment shall not sustain any physical damage or deteriorate in functional performance when subjected to vibration levels expected during transportation			
7	ASTM D880	Inclined Impact in a Packaged Rack	L11	1. Instrument rack with accelerometers to capture blade amplification. 2. Impact to all 4 vertical surfaces at velocity of 1 meter/second. 3. Reliability team may request the addition of strain gauges on the MB during this test.			

Table 21: System Reliability Plan

18 Regulatory Information

L10 systems shall comply with the regulatory requirements of the destination country/region and meet compliance guidelines outlined in Microsoft specification M1039012.

FPGA regulatory model number: **A-2040**

FPGA Card, model A-2040, is:

- Evaluated as Information Technology Equipment (ITE), designed to operate in a typical data center environment. The suitability of this product for other environments may require further evaluation.
- Designed for use with NRTL Listed (UL, CSA, ETL, etc.) and IEC/EN 60950-1 or IEC/EN 62368-1 compliant (CE marked) Information Technology equipment.

NOTICE: Changes or modifications made to the equipment not expressly approved by Microsoft may void the user’s authority to operate the equipment.

CANADA and USA:

NOTICE: This equipment has been tested and found to comply with the limits for a Class A digital device, pursuant to part 15 of the FCC Rules. These limits are designed to provide reasonable protection against harmful interference when the equipment is operated in a commercial environment. This equipment generates, uses, and can radiate radio frequency energy and, if not installed and used in accordance with the instruction manual, may cause harmful interference to radio communications. Operation of this equipment in a residential area is likely to cause harmful interference in which case the user will be required to correct the interference at his own expense.

This device complies with part 15 of the FCC Rules and Industry Canada license-exempt RSS standard(s). Operation is subject to the following two conditions: (1) this device may not cause harmful interference, and (2) this device must accept any interference received, including interference that may cause undesired operation of the device.

Cet appareil numérique de la classe A est conforme à la norme NMB-003 du Canada.

Le présent appareil est conforme aux CNR d'Industrie Canada applicables aux appareils radio exempts de licence. L'exploitation est autorisée aux deux conditions suivantes: (1) l'appareil ne doit pas produire de brouillage, et (2) l'utilisateur de l'appareil doit accepter tout brouillage radioélectrique subi, même si le brouillage est susceptible d'en compromettre le fonctionnement.

CAN ICES-3(A)/NMB-3(A)

EUROPEAN UNION:

Warning: This is a class A product. In a domestic environment, this product may cause radio interference in which case the user may be required to take adequate measures.

Disposal of Electrical & Electronic Equipment: This symbol the product or its packaging means that this product must not be disposed of with your household waste. Instead, it is your responsibility to hand this over to an applicable collection point for the recycling of electrical and electronic equipment. This separate collection and recycling will help to conserve natural resources and prevent potential negative consequences for human health and the environment, which inappropriate disposal could cause due to the possible presence of hazardous substances in electrical and electronic equipment. For more information about where to drop off your electrical and electronic waste, please contact your local city/municipality office, your household waste disposal service, or the shop where you purchased this product. Contact erecycle@microsoft.com for additional information on WEEE.