# Intel® Xeon Phi™ Coprocessor

## Datasheet

*November, 2012*

Reference Number: 328209-001EN

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS.  EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY RELATING TO SALE AND/OR USE OF INTEL PRODUCTS, INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT, OR OTHER INTELLECTUAL PROPERTY RIGHT.

UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

**This document contains information on products in the design phase of development. The information here is subject to change without notice.** Do not finalize a design with this information. Contact your local Intel sales office or your distributor to obtain the latest specification before placing your product order.

All products, dates, and figures are preliminary for planning purposes and are subject to change without notice.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document, or other Intel literature may be obtained by calling 1-800-548-4725 or visiting Intel's website at http://www.intel.com. Confidential documents are only available via Intel Business Link at http://www.intel.com/ibl.

**Intel, Xeon, Xeon Phi,** and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.

# Table of Contents

# List of Figures

# List of Tables

# Revision History

| Document Number | Revision Number | Description | Date |
|---|---|---|---|
| 328209-001 | 001 | - First release of datasheet | November 2012 |

§

# 1 Introduction

## 1.1 Reference Documentation

Table 1-1 lists most of the applicable documents. For complete list of documentation, contact your local Intel representative or go to www.intel.com.

Table 1-1. Related Documents

| Document | www.intel.com |
|---|---|
| PCI Express* Card Electromechanical Specification, Revision 2.0, April 11, 2007<br>*www.pcisig.com* | N/A |
| Intel® Xeon Phi™ Coprocessor Specification Update | 328205-001EN |
| Intel® Xeon Phi™ Coprocessor Safety Compliance Guide | 328206-001EN |
| Intel® Xeon Phi™ Coprocessor System Software Developer’s Guide | 328207-001EN |
| Intel® Xeon Phi™ Coprocessor Thermal Mechanical Models | 328208-001EN |
| Intel® Xeon Phi™ Coprocessor Instruction Set Reference Manual<br>*http://software.intel.com/en-us/forums/intel-many-integrated-core/* | N/A |

## 1.2 Conventions and Terminology

### 1.2.1 Terminology

This section provides the definitions of some of the terms used in this document.

Table 1-2. General Terminology

| Terminology | Definition |
|---|---|
| BGA | Ball Grid Array |
| BMC | Baseboard Management Controller |
| DFF | Dense Form Factor |
| ECC | Error Correction Code |
| GDDR | Graphics Double Data Rate |
| IBP | Intel Business Portal |
| IPMB | Intelligent Platform Management Bus |
| IPMI | Intelligent Platform Management Interface |
| ME | Manageability Engine |
| PCIe | PCI Express* |
| RAS | Reliability Accessibility Serviceability |
| SKU | Stock Keeping Unit |
| SMBus | System Management Bus |

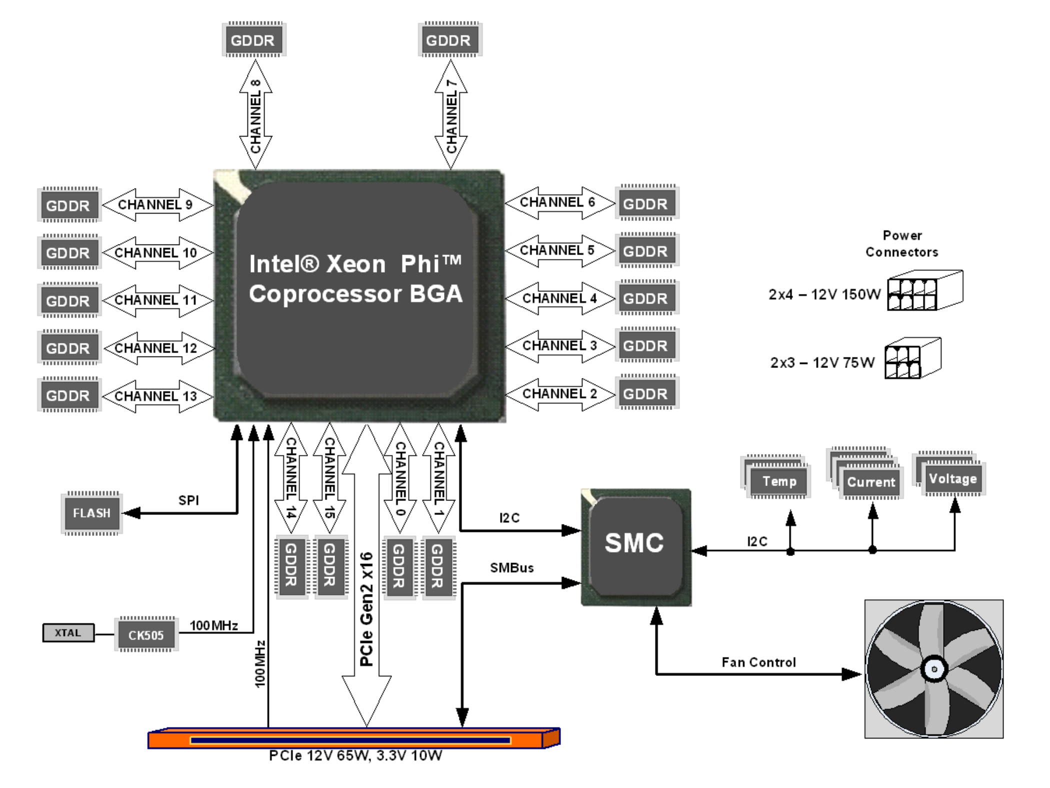
## Table 1-2. General Terminology

| Terminology | Definition |
|---|---|
| SMC | System Management Controller |
| TDP | Thermal Design Power |
| VR | Voltage Regulator |

# 2 Intel® Xeon Phi™ Coprocessor Architecture

## 2.1 Intel® Xeon Phi™ Coprocessor Board Overview

Figure 2-1. Intel® Xeon Phi™ Coprocessor Board Schematic[1]



***Notes:***

1. On-board fan is available on Intel® Xeon Phi™ Coprocessor product with 3100 series active SKU only.

The Intel® Xeon Phi™ Coprocessor consists of the following primary subsystems.

- Many Integrated Core (MIC) coprocessor and GDDR5 memory.
- System Management Controller (SMC), on-board thermal sensors (inlet air, outlet air, coprocessor and single GDDR5 sensor) and fan (only on the 3100 series; see SKU matrix Table 2-1).
- Voltage regulators (VRs) powered by the motherboard through the PCI Express* connector, a 2x4 (150W) and a 2x3 (75W) power connector on the east edge of the card. Along with power through the PCI Express* connector, the 300W SKUs need both 2x4 and 2x3 connectors to be driven by system power supplies. The 225W SKU may be powered only through the PCI Express* connector and the 2x4 connector.

- PCI Express* connections.
- The clock system is integrated in the coprocessor and requires only the PCI Express* 100MHz reference clock and an on-board 100MHz +/- 50ppm reference.

The Intel® Xeon Phi™ coprocessor provides the following high-level features:

- A many-core coprocessor.
- Maximum 16-channel GDDR memory interface with an option to enable ECC.
- PCI Express* x16 lane Gen2 interface with optional SMBus management interface.
- Node Power and Thermal Management, including power capping support.
- +12V power monitoring and on-board fan PID controller on the 3100 series active SKU.
- On-board flash device that loads the coprocessor OS on boot.
- Card level RAS features and recovery capabilities.

### 2.1.1 Intel® Xeon Phi™ Coprocessor Board Design

The Intel® Xeon Phi™ coprocessor is a PCI Express* compliant, 246mm x 111mm, high-power add-in card. It supports a maximum of 16 GDDR memory channels, distributed on both sides of the PCB. Each memory channel supports two 16-bit wide GDDR device (for a maximum of 32 devices in the card), combining to give 32-bit wide data. Figure 2-2 and Figure 2-3 show the front and back sides of the PCB. The two notches along the top edge of the card are used to attach the cooling plate for the GDDR devices on the backside of the PCB (side without the Intel® Xeon Phi™ coprocessor silicon). The VRs are split right and left to help reduce direct current resistance and current density.

Intel® Xeon Phi™ coprocessor supports 2 power groups for a total of 4 primary low-voltage rails: GROUP A contains VDDG, VDDQ, and VSFR, while GROUP B contains VCCP. The VCCP, VDDG, and VDDQ rails are powered from the PCI Express* edge connector and the auxiliary 12V inputs. The VSFR rail is powered from the PCI Express* edge connector 3.3V input (~5W). VCCP is the coprocessor core voltage rail, while VDDQ, VDDG and VSFR supply power to memory, portions of the coprocessor and miscellaneous circuitry on the card.

**Figure 2-2. Intel® Xeon Phi™ Coprocessor Board Top side**

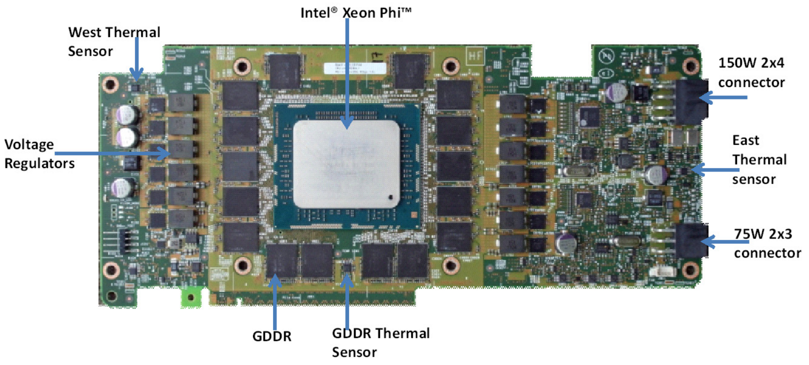
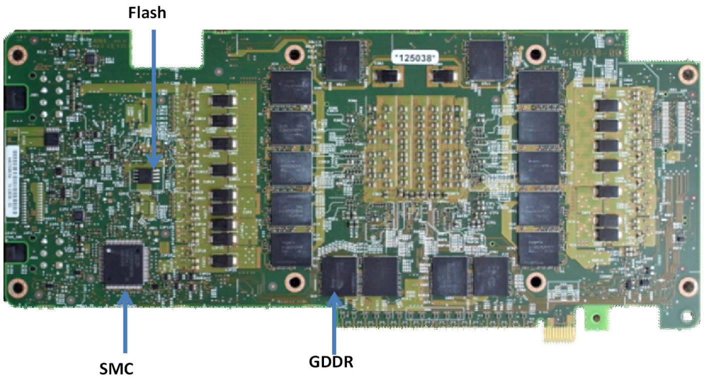


Figure 2-3. Intel® Xeon Phi™ Coprocessor Board, Back side



***Note:*** Figure 2-2 and 2-3 are representative of the final Intel® Xeon Phi™ Coprocessor board without the package thermal and mechanical elements.

### 2.1.2 System Management Controller (SMC)

The SMC has three I2C interfaces. This allows a direct connection to the coprocessor I2C interface, an intracard I2C sensor bus and a system SMBus for clean integration with platform and power management systems. The interface between the SMC and coprocessor interface is used for coprocessor thermal and status information exchange. The sensor bus allows board thermal, input power, and current sense monitoring for fan and power control. This information can be forwarded to the coprocessor for power state control. The SMBus is used by system for chassis fan control with the passive heat sink card and for integration with the Node Management controller in the platform. Communication with the system baseboard management controller (BMC) or peripheral control hub (PCH) occurs over the SMBus using the standard IPMB protocol. See chapter on manageability for more details

### 2.1.3 Intel® Xeon Phi™ Coprocessor Silicon

Figure 2-4. Intel® Xeon Phi™ Coprocessor Silicon Layout

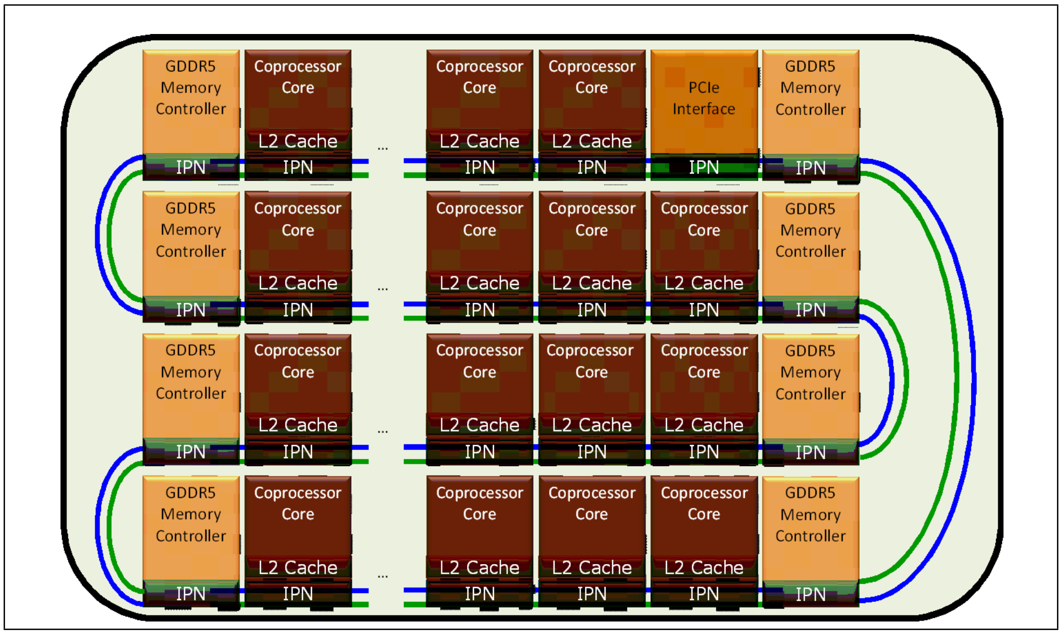


Figure 2-4 is a conceptual drawing of the general structure of the Intel® Xeon Phi™ coprocessor architecture, and does not imply actual distances, latencies, etc. The cores, PCIe Interface logic, and GDDR5 memory controllers are connected via an Interprocessor Network (IPN) ring, which can be thought of as independent bidirectional ring.

The L2 caches are shown here as slices per core, but can also be thought of as a fully coherent cache, with a total size equal to the sum of the slices. Information can be copied to each core that uses it to provide the fastest possible local access, or a single copy can be present for all cores to provide maximum cache capacity.

The Intel® Xeon Phi™ coprocessor can support upto 61 cores (making a 31 MB L2) cache) and 8 memory controllers with 2 GDDR5 channels each. The maximum number of cores and total card memory varies with Intel® Xeon Phi™ coprocessor SKU; refer to the *Intel® Xeon Phi™ PCI Express* Card Specification Update* for information.

Communication around the ring follows a Shortest Distance Algorithm (SDA). Co-resident with each core structure is a portion of a distributed tag directory. These tags are hashed to distribute workloads across the enabled cores. Physical addresses are also hashed to distribute memory accesses across the memory controllers.

### 2.1.4 Intel® Xeon Phi™ Coprocessor Product Family

**Table 2-1. Intel® Xeon Phi™ Coprocessor Product Family**

| SKU | Card TDP (Watts) | Cooling Solution[1] |
|---|---|---|
| SE10P | 300 | Passive |
| SE10X | 300 | None[2] |
| 5110P[3] | 225 | Passive |
| 3100 Series | 300 | Passive, Active |
| Dense Form Factor | TBD | None[4] |

***Notes:***

1. Passive cooling solution uses topside heatsink (vapor chamber and copper fins) and backside aluminum plate. Active cooling uses on-card dual-intake blower.
2. Same performace and card configuration as the SE10P but without Intel heatsink or chassis retention mechanism; allows for custom thermal and mechanical design by users.
3. Refer to Section 5.1.
4. Dense Form Factor (DFF): Smaller physical footprint than the other Intel® Xeon Phi™ coprocessor products, for innovative platform designs with unique PCI Express* interface, PCI Express* Gen2 specification compliant.

### 2.1.5 Intel® Xeon Phi™ Coprocessor Dense Form Factor

The Intel® Xeon Phi™ coprocessor Dense Form Factor (DFF) is a derivative of the standard Intel® Xeon Phi™ coprocessor PCI Express* form factor card. The high-level features of DFF are:

- 117.35mm(4.62”) x 149.86mm(5.9”) PCB.
- 230-pin unique edge finger designed to industry standard x24 PCI Express* connector, PCI Express* Gen2 compliant.
- All power to the card is supplied through the connector.
- There is no auxilliary 2x4 or 2x3 power connector on the card
- Supports vertical, straddle or right-angle mating connectors.
- On board SMC. The manageability features and software capabilities remain the same as for other Intel® Xeon Phi™ coprocessor products.
- To allow for system design innovation and differentiation, Intel will ship only the assembled and fully functional PCB, without heatsink or chassis retention mechanism. This allows system designers to implement their own cooling solution and connector of choice. Due to presence of GDDR5 memory components on the backside of the DFF board, a custom cooling design must comprehend both sides of the DFF product.
- Baseboard designers must ensure the signal integrity of all PCI Express* signals as they pass the connector of choice and reach the connector fingers of the DFF product.

§

# 3 Thermal and Mechanical Specification

## 3.1 Mechanical Specifications

The mechanical features of the Intel® Xeon Phi™ coprocessor are compliant with the *PCI Express* 225W/300W High Power Card Electromechanical Specification 1.0.*

Table 3-1 shows the mechanical specifications of Intel® Xeon Phi™ coprocessor passive and active cards.

Table 3-1. Intel® Xeon Phi™ Coprocessor Mechanical Specification

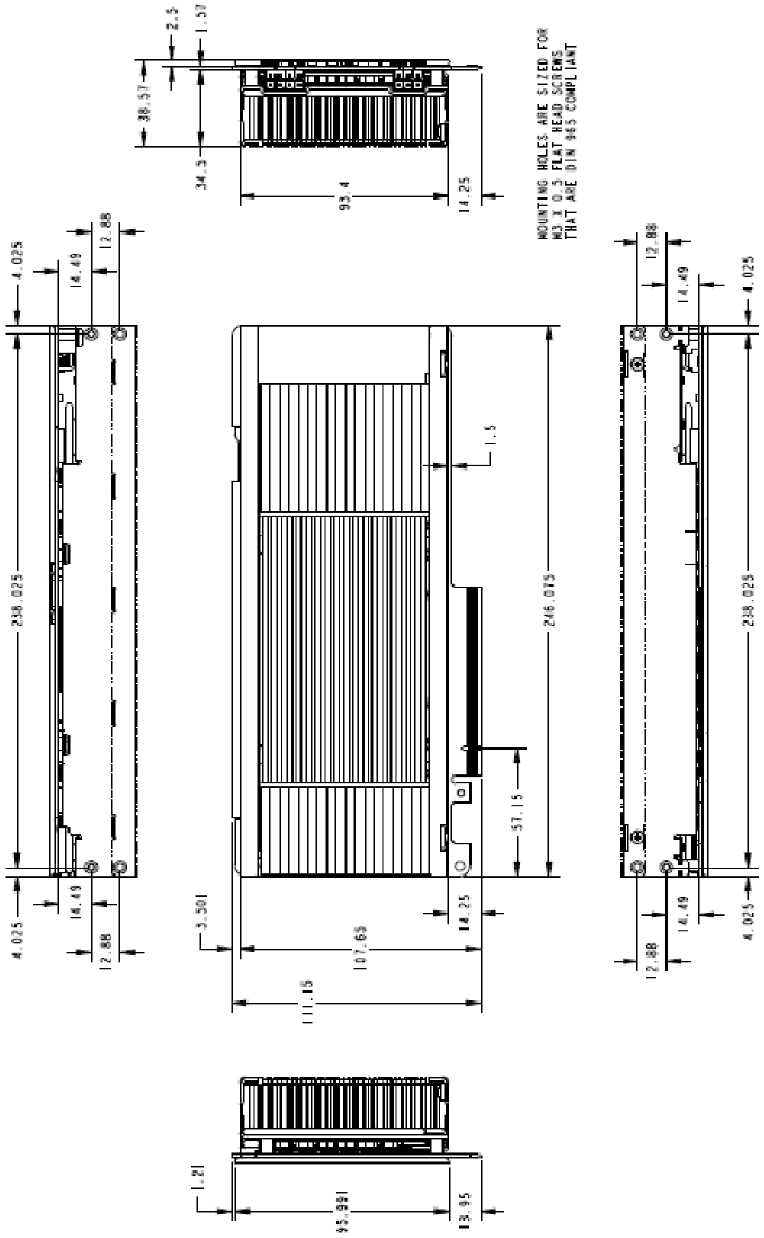
| Parameter | Specification |
|---|---|
| Card Length | 247.9mm[1] |
| Primary Side Height Keep-in | 34.8mm |
| Secondary Side Height Keep-in | 2.67mm |
| Total Card Mass (3100 series active SKU)<br>Total Card Mass (all passive SKUs) | 1400g<br>1200g |

***Notes:***

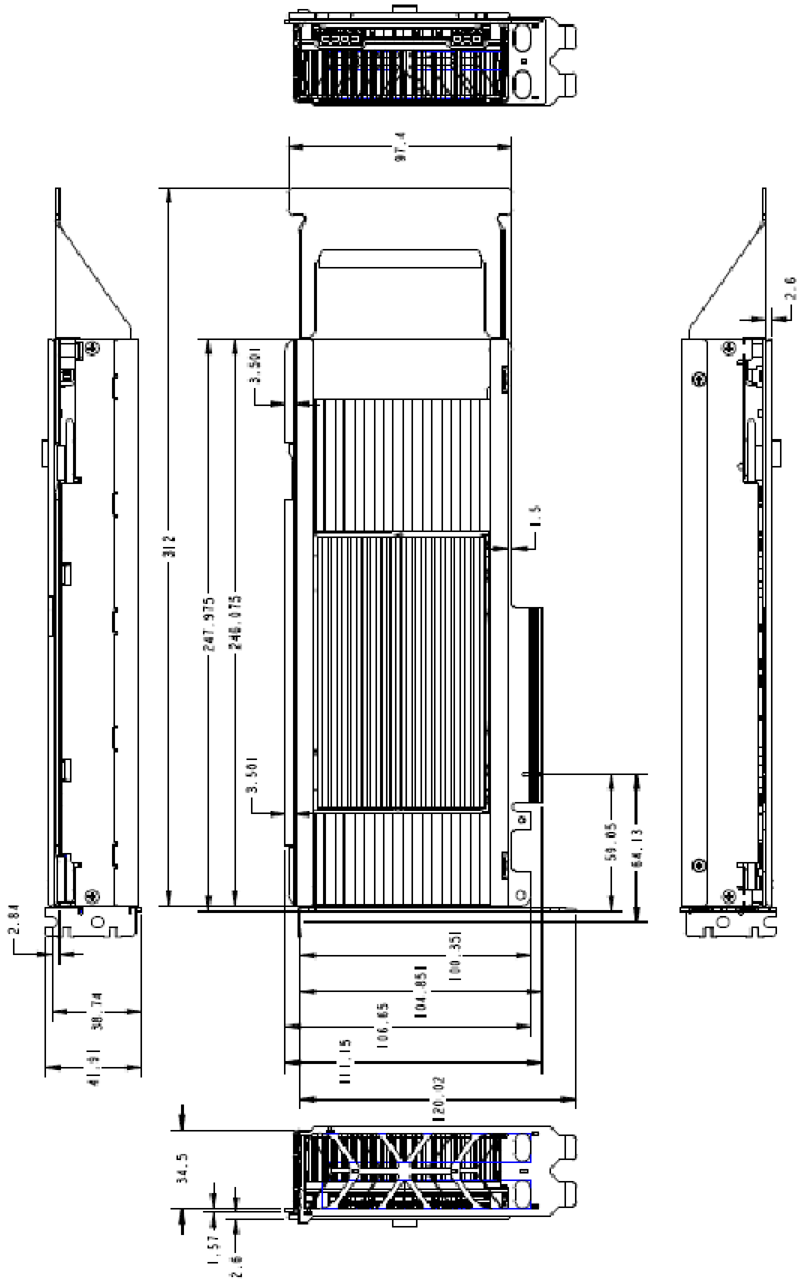
1. Inclusive of I/O bracket

Figure 3-1 shows the mounting holes and Figure 3-2 shows the relevant dimensions of the Intel® Xeon Phi™ coprocessor passive and active cards for chassis retention. Refer to the *Intel® Xeon Phi™ PCI Express* Card Thermal and Mechanical Models* for Pro-Engineering, Icepak and Flotherm models.

Figure 3-1 Location of Mounting Holes on the Intel® Xeon Phi™ Coprocessor (in mils)



### Figure 3-2 Dimensions of Intel® Xeon Phi™ Coprocessor (dimension in mils)



## 3.2 Intel® Xeon Phi™ Coprocessor Thermal Specification

**Table 3-2. Intel® Xeon Phi™ Coprocessor Thermal Specification**

| Parameter | Specification |
|---|---|
| Assumed $T_{AMB}$ (chassis external) | 35°C |
| Assumed $T_{RISE}$ to card inlet | 10°C |
| Assumed card $T_{INLET}$ | 20°C |
| Max card $T_{INLET}$ | 45°C |
| Max card $T_{EXHAUST}$ | 70°C |
| $T_{case}$ of coprocessor | 92°C |
| $T_{case}$ of GDDR | 85°C |
| $T_{control}$ | ~82°C[1] |
| $T_{throttle}$ | ~104°C[2] |
| $T_{thermtrip}$ | ~$T_{throttle}$ + 20°C[3] |

***Notes:***

1. $T_{control}$ is the setpoint at which the system fans must ramp up towards full power (or RPM) to maintain the Intel® Xeon Phi™ coprocessor temperature around $T_{control}$ and prevent throttling. It is highly recommended that the system BMC query the SMC on the coprocessor card for accurate Tcontrol value.
2. When the coprocessor junction temperature ($T_{junction}$) reaches $T_{throttle}$ due to insufficient cooling, the SMC will force thermal throttle resulting in the lowest frequency of 800MHz in an attempt to reduce the power and cool down the coprocessor.
3. If the coprocessor temperature continues to rise beyond $T_{throttle}$ and approached $T_{thermtrip}$, it will result in card shutdown to prevent damage to the coprocessor. The VRs will be shut down and power to the card must be recycled. $T_{thermtrip}$ should not be considered a specification; it can change between SKUs, and is given here to as guidance.

### 3.2.1 Intel® Xeon Phi™ Coprocessor Thermal Management

Thermal management on the Intel® Xeon Phi™ coprocessor card is achieved through a combination of coprocessor based sensors, card level sensors and inputs, and a coprocessor frequency control circuit. Reducing card temperature is accomplished by adjusting the frequency of the coprocessor. Lowering the coprocessor frequency will reduce the power dissipation and consequently the temperature.

The coprocessor carries in it a factory calibrated Digital Temperature Sensor (DTS) that monitors coprocessor temperature. Data from this sensor is available to the BMC or other system software via both in-band (direct software reads) and out-of-band (over the PCI Express* SMBus) interface. Refer to chapter titled "Manageability" for more information on how to access DTS information. System management software can use this data to monitor the silicon temperature and take any appropriate actions. Systems that adjust airflow based on component temperatures must monitor the coprocessor's DTS to ensure sufficient cooling is always available.

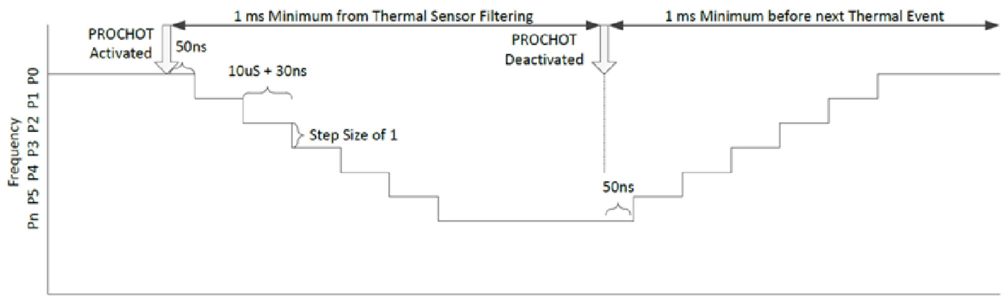
In addition to making thermal information available to system manageability software, the DTS is constantly comparing the coprocessor temperature to the factory set maximum permissible temperature called $T_{throttle}$. If the measured temperature at any time exceeds $T_{throttle}$ (a state also known as PROCHOT), then the coprocessor will automatically step down the operating frequency (or Pstate) in an attempt to reduce the temperature (this is often referred to as "thermal throttling"). Once the temperature has dropped below $T_{throttle}$, the frequency will be brought back up to the original setting. See Figure 3-3 below.

Within 50ns of detecting $T_{throttle}$, the DTS circuit begins stepping down the P-states until Pn is reached.  Each frequency step is approximately 100MHz; the exact value will depend on the starting frequency.  After each step, the DTS will wait 10uS before taking the next step. The number of steps, or P-states, depends on the starting frequency and the minimum frequency supported by the processor.  Once Pn is reached, the frequency will be held at that level for approximately 1ms, or until the temperature has dropped below Tprochot, whichever is longer.

If throttling continues for more than 100mSec, the coprocessor OS will reduce the voltage setting in order to further decrease the power dissipation. The voltage settings are pre-programmed at the factory and cannot be reconfigured.

Upon removal of the thermal event, the process reverses and the voltage and frequency are stepped back up the P1 state. Although the process to reduce frequency is managed by the coprocessor circuits, the sequence to bring the coprocessor back to P1 is controlled by the coprocessor OS. As a result, the precise timings of the step changes may be slightly longer than 10uS.

**Figure 3-3 Entering and Exiting Thermal Throttling (PROCHOT)**



Activation of PROCHOT by the on-die Digital Temperature Sensor (DTS) initiates the sequence of stepping down the processor frequency from P1 to Pn. Deactivation of PROCHOT reverses the process.

## 3.3 Intel® Xeon Phi™ Coprocessor Thermal Solutions

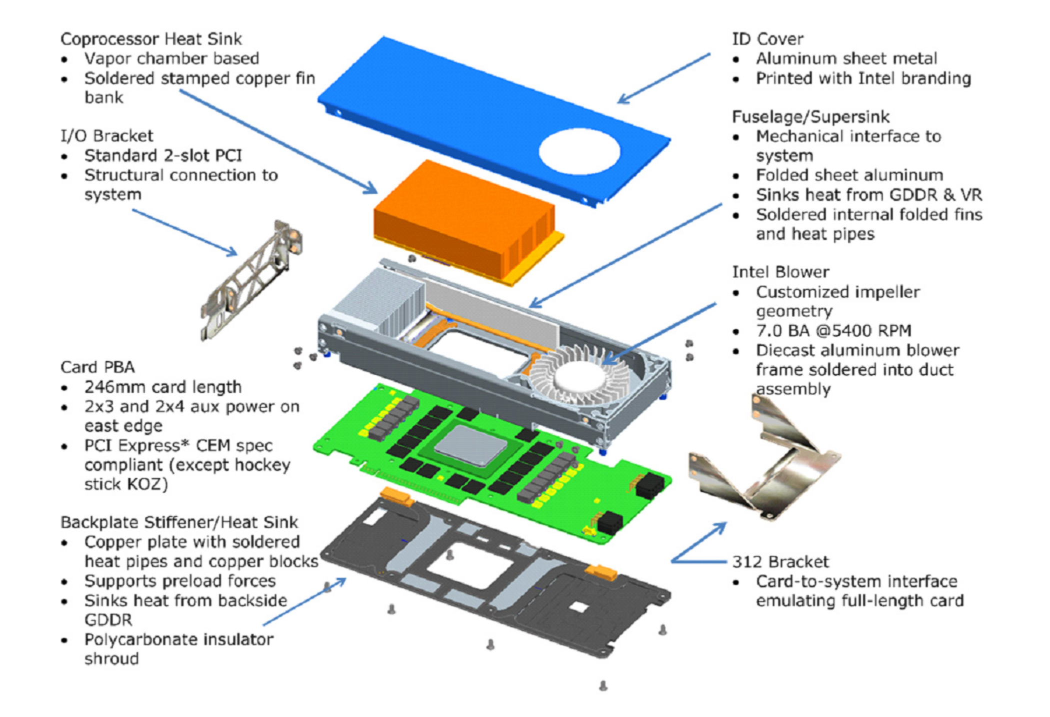
There are two types of thermal solutions to address the Intel® Xeon Phi™ coprocessor power limits: a passive solution for most SKUs as indicated in table 2-1 (which relies on forced convection airflow provided by the system) and an active solution on the 3100 series active  SKU (which uses a high performance blower.) The active solution is designed to operate in an 'adjacent card configuration' such that the impedance from a nearby flow blockage is accounted for within the design. Both passive and active solutions come with cooling backplates, which are required to augment the stiffness of the Intel® Xeon Phi™ coprocessor in order to counteract the preload applied by the primary side (housing the coprocessor) when assembled, to protect the structural integrity of the coprocessor and GDDR packages during a shock event, and to provide a protective cover.

Given the requirement to dissipate backside GDDR heat within the 2.67mm keep-in height prescribed by the PCI Express* specification, the backplate is designed to transfer the GDDR heat from the secondary side via heat pipes to the primary side thermal solution.

## 3.3.1 3100 Series Active Cooling Solution

For the 3100 series active SKU, the Intel® Xeon Phi™ coprocessor thermal-mechanical solution utilizes a supersink approach in which a primary heatsink is used to cool the coprocessor while a metallic fuselage/supersink cools the VR and GDDR components. Figure 3-4 illustrates the key components of the active cooling design.

**Figure 3-4 Exploded View of 3100 Series Active Solution**


Coprocessor Heat Sink
• Vapor chamber based
• Soldered stamped copper fin bank
I/O Bracket
• Standard 2-slot PCI
• Structural connection to system
Card PBA
• 246mm card length
• 2x3 and 2x4 aux power on east edge
• PCI Express* CEM spec compliant (except hockey stick KOZ)
Backplate Stiffener/Heat Sink
• Copper plate with soldered heat pipes and copper blocks
• Supports preload forces
• Sinks heat from backside GDDR
• Polycarbonate insulator shroud
ID Cover
• Aluminum sheet metal
• Printed with Intel branding
Fuselage/Supersink
• Mechanical interface to system
• Folded sheet aluminum
• Sinks heat from GDDR & VR
• Soldered internal folded fins and heat pipes
Intel Blower
• Customized impeller geometry
• 7.0 BA @5400 RPM
• Diecast aluminum blower frame soldered into duct assembly
312 Bracket
• Card-to-system interface emulating full-length card

In the fuselage/supersink approach, the duct is metallic and performs both structural and thermal roles. In its 'fuselage' function, the duct provides structural support for the forces generated by the coprocessor thermal interface, protects against shock events, and channels airflow through the card. In its 'supersink' function, the duct contains internal fins, heat pipes, and diecast blower frame. The internal heat pipes serve to transmit heat from GDDR (both top- and bottom-side) and VR components to the internal fin banks, diecast blower frame, and metal fuselage structure where it can be effectively transferred to the airstream. The duct also contains horizontal webs which interface to the east and west GDDR as well as to the VR FETs. Together, these structures dissipate heat lost from the GDDR and VR components into the air.

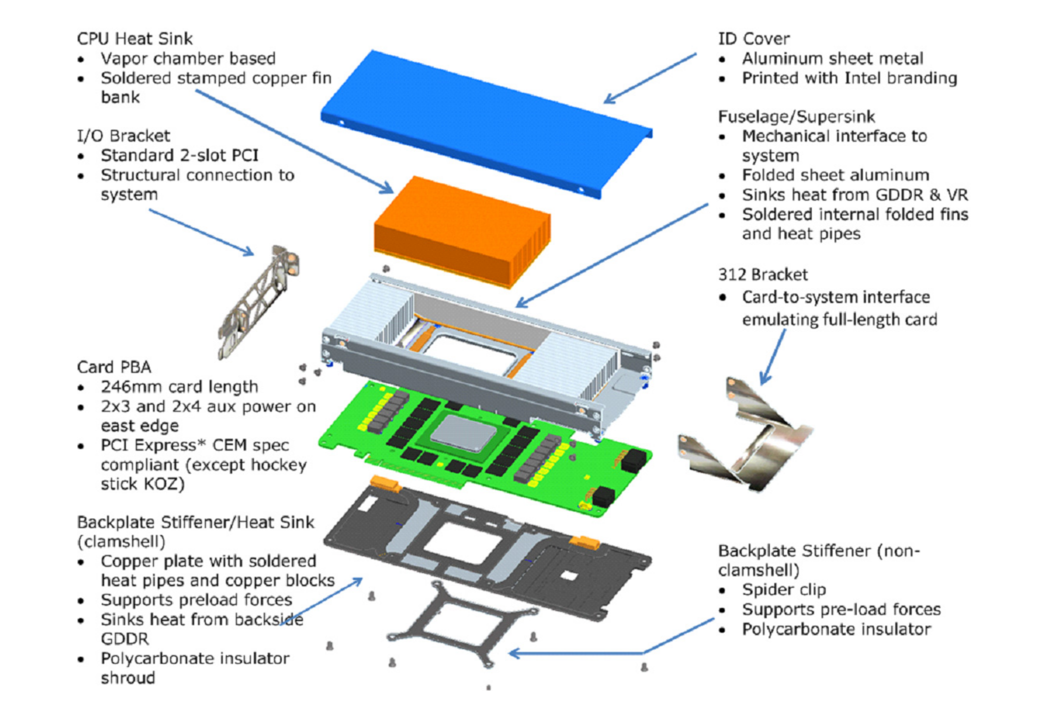
The coprocessor thermal path is separated from the GDDR and VR components, and utilizes a heatsink with parallel plate fins and vapor chamber base.

The active solution also contains a high-performance dual-intake blower that operates up to 5400 rpm at 20W of motor power. The blower has been designed to maximize the pressure drop capability and is able to deliver up to 35 $ft^3$/min in an open airflow environment. When installed on the card, the blower delivers 31 $ft^3$/min with no adjacent blockage. When an adjacent card is considered, the resultant impedance loss causes the flow rate to drop to 23 ft$^3$/min. The active thermal solution is designed to provide sufficient cooling even in the latter scenario.

### 3.3.2 SE10P/5110P/3100 Series Passive Cooling Solution

For the passive heatsink on the SE10P/5110P/3100 Series SKUs, the Intel® Xeon Phi™ coprocessor thermal & mechanical solution also utilizes a 'fuselage/supersink' approach. Figure 3-5 illustrates the key components of the passive design.

Figure 3-5 Exploded View of Passive Thermal Solution


CPU Heat Sink
• Vapor chamber based
• Soldered stamped copper fin bank
I/O Bracket
• Standard 2-slot PCI
• Structural connection to system
Card PBA
• 246mm card length
• 2x3 and 2x4 aux power on east edge
• PCI Express* CEM spec compliant (except hockey stick KOZ)
Backplate Stiffener/Heat Sink (clamshell)
• Copper plate with soldered heat pipes and copper blocks
• Supports preload forces
• Sinks heat from backside GDDR
• Polycarbonate insulator shroud
ID Cover
• Aluminum sheet metal
• Printed with Intel branding
Fuselage/Supersink
• Mechanical interface to system
• Folded sheet aluminum
• Sinks heat from GDDR & VR
• Soldered internal folded fins and heat pipes
312 Bracket
• Card-to-system interface emulating full-length card
Backplate Stiffener (non-clamshell)
• Spider clip
• Supports pre-load forces
• Polycarbonate insulator

As in the active thermal solution, the duct is metallic and performs both structural and thermal roles. In its 'fuselage' function, the duct provides structural support for the forces generated by the CPU thermal interface, protects against shock events, and channels airflow through the card. In its 'supersink' function, the duct contains internal fins and heat pipes. The internal heat pipes serve to transmit heat from GDDR (both top- and bottom-side) and VR components to the internal fin banks, diecast blower frame, and metal fuselage structure where it can be effectively transferred to the airstream. The passive solution does not have a diecast blower frame as it relies upon forced airflow from the host system. In place of the blower and frame, an additional fin bank is added to dissipate waste heat from GDDR and VR components. The fin spacing of all fin banks as well as of the CPU heat sink fin bank have been optimized for receiving system-supplied airflow. A backplate stiffener/heat sink is used.

### 3.3.2.1 System Airflow for 5110P/3100 Series Passive SKUs

In order to ensure adequate cooling of the 5110P/3100 Series Passive SKUs with a 45$^{o}$C inlet temperature, the system must be able to provide 20 ft$^3$/min of airflow to the card with 4.3 ft$^3$/min on the secondary side and the remainder on the primary side. The total pressure drop (assuming a multi-card installation conforming to the PCI Express* mechanical specification) is 0.21 inch $H_2O$ at this flow rate.

***Note:*** For systems with reversed airflow, the corresponding airflow requirement is expected to be within +/-5% tolerance of the values shown in the following tables.

If the system is able to provide a temperature lower than 45$^{o}$C at the card inlet, then the total airflow can be reduced according to the graph and table in Figure 3-6.

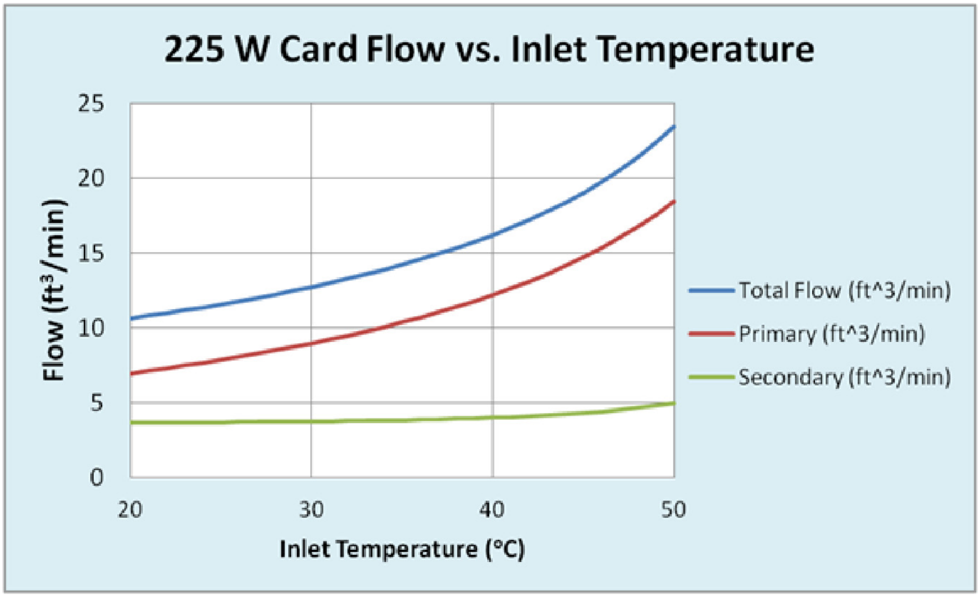
If the 5110P SKU is powered by a 2x4 and a 2x3 connector, the card can support an additional 20W of power for maximum TDP of 245W (see Section 2.1.5 for more details). In this case, the corresponding airflow requirement for cooling the part as a 245W card is shown in Figure 3-8.

### 3.3.2.2 Airflow Requirement for SE10P Passive Cooling Solution

In order to ensure adequate cooling of the SE10P 300W SKU with a 45$^{o}$C inlet temperature, the system must be able to provide 33 ft$^3$/min of airflow to the card with 7.2 ft$^3$/min on the secondary side and the remainder on the primary side. The total pressure drop (assuming a multi-card installation conforming to the PCI Express* mechanical specification) is 0.54 in $H_2O$ at this flow rate.

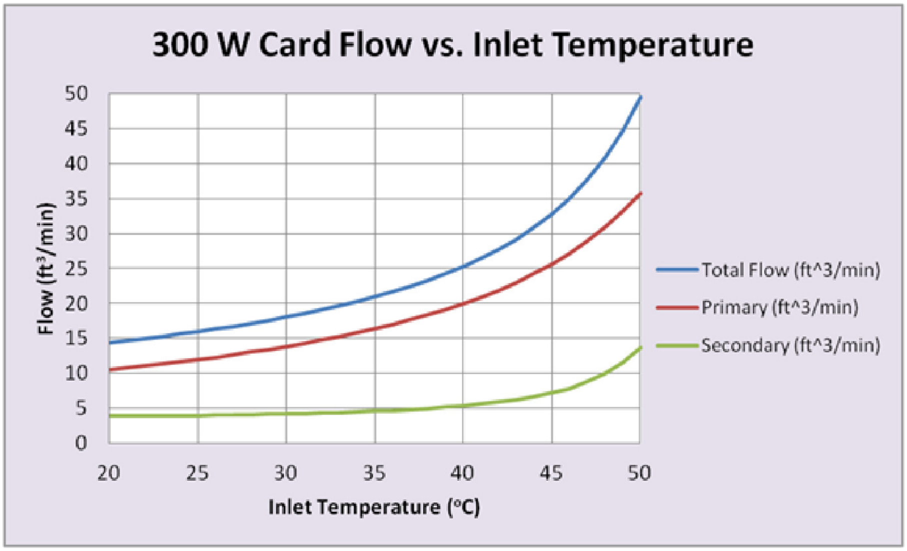
If the system is able to provide a temperature lower than 45$^{o}$C at the card inlet, then the total airflow can be reduced according to the graph and table in Figure 3-7.

### Figure 3-6 Airflow Requirement vs. Inlet Temperature for the 5110P/3100 Series Passive Cards



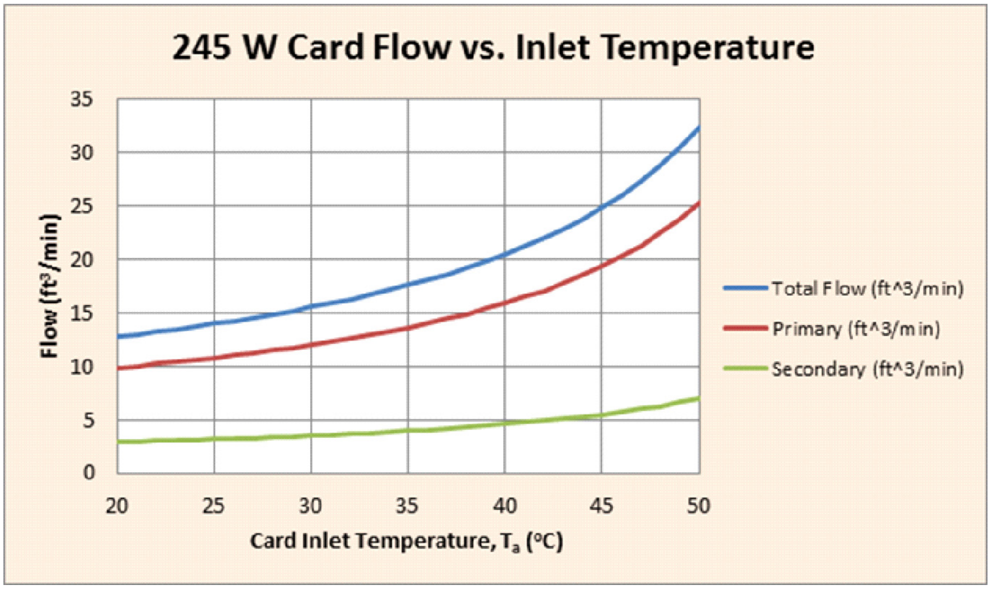
| Card Inlet (°C) | Total Flow (ft^3/min) | Primary (ft^3/min) | Secondary (ft^3/min) | Card dP(inch $H_2O$) |
|---|---|---|---|---|
| 20 | 10.6 | 7 | 3.7 | 0.08 |
| 21 | 10.8 | 7.1 | 3.7 | 0.08 |
| 22 | 11 | 7.3 | 3.7 | 0.08 |
| 23 | 11.2 | 7.5 | 3.7 | 0.08 |
| 24 | 11.4 | 7.7 | 3.7 | 0.09 |
| 25 | 11.6 | 7.9 | 3.7 | 0.09 |
| 26 | 11.8 | 8.1 | 3.7 | 0.09 |
| 27 | 12 | 8.3 | 3.7 | 0.09 |
| 28 | 12.2 | 8.5 | 3.7 | 0.1 |
| 29 | 12.5 | 8.7 | 3.7 | 0.1 |
| 30 | 12.7 | 9 | 3.8 | 0.1 |
| 31 | 13 | 9.2 | 3.8 | 0.11 |
| 32 | 13.3 | 9.5 | 3.8 | 0.11 |
| 33 | 13.6 | 9.8 | 3.8 | 0.12 |
| 34 | 13.9 | 10.1 | 3.8 | 0.12 |
| 35 | 14.2 | 10.4 | 3.8 | 0.12 |
| 36 | 14.6 | 10.7 | 3.9 | 0.13 |
| 37 | 15 | 11.1 | 3.9 | 0.14 |
| 38 | 15.4 | 11.4 | 3.9 | 0.14 |
| 39 | 15.8 | 11.8 | 4 | 0.15 |
| 40 | 16.2 | 12.2 | 4 | 0.16 |
| 41 | 16.7 | 12.7 | 4.1 | 0.16 |
| 42 | 17.2 | 13.1 | 4.1 | 0.17 |
| 43 | 17.8 | 13.6 | 4.2 | 0.18 |
| 44 | 18.4 | 14.2 | 4.2 | 0.19 |
| 45 | 19.1 | 14.7 | 4.3 | 0.21 |

### Figure 3-7 Airflow Requirement vs. Inlet Temperature for the SE10P Passive Card



| Card Inlet (°C) | Total Flow (ft^3/min) | Primary (ft^3/min) | Secondary (ft^3/min) | Card dP (inH$_2$O) |
|---|---|---|---|---|
| 20 | 14.4 | 10.5 | 3.9 | 0.13 |
| 21 | 14.7 | 10.8 | 3.9 | 0.13 |
| 22 | 15 | 11.1 | 3.9 | 0.14 |
| 23 | 15.3 | 11.3 | 3.9 | 0.14 |
| 24 | 15.6 | 11.7 | 4 | 0.15 |
| 25 | 16 | 12 | 4 | 0.15 |
| 26 | 16.3 | 12.3 | 4 | 0.16 |
| 27 | 16.7 | 12.7 | 4.1 | 0.16 |
| 28 | 17.1 | 13 | 4.1 | 0.17 |
| 29 | 17.6 | 13.4 | 4.1 | 0.18 |
| 30 | 18 | 13.8 | 4.2 | 0.19 |
| 31 | 18.5 | 14.3 | 4.3 | 0.2 |
| 32 | 19.1 | 14.8 | 4.3 | 0.21 |
| 33 | 19.7 | 15.3 | 4.4 | 0.22 |
| 34 | 20.3 | 15.8 | 4.5 | 0.23 |
| 35 | 21 | 16.4 | 4.6 | 0.24 |
| 36 | 21.7 | 17 | 4.7 | 0.26 |
| 37 | 22.5 | 17.6 | 4.8 | 0.27 |
| 38 | 23.3 | 18.4 | 5 | 0.29 |
| 39 | 24.3 | 19.1 | 5.1 | 0.31 |
| 40 | 25.3 | 20 | 5.3 | 0.34 |
| 41 | 26.5 | 20.9 | 5.6 | 0.37 |
| 42 | 27.8 | 21.9 | 5.9 | 0.4 |
| 43 | 29.2 | 23 | 6.2 | 0.44 |
| 44 | 30.9 | 24.2 | 6.7 | 0.48 |
| 45 | 32.8 | 25.6 | 7.2 | 0.54 |

### Figure 3-8 Airflow Requirement vs. Inlet Temperature for the 5110P Card with 245W TDP



| Card Inlet (°C) | Total Flow (ft^3/min) | Primary (ft^3/min) | Secondary (ft^3/min) | Card dP (inH$_2$O) |
|---|---|---|---|---|
| 20 | 12.8 | 9.9 | 2.9 | 0.1 |
| 21 | 13 | 10 | 3 | 0.11 |
| 22 | 13.2 | 10.2 | 3 | 0.11 |
| 23 | 13.5 | 10.4 | 3.1 | 0.11 |
| 24 | 13.7 | 10.6 | 3.1 | 0.12 |
| 25 | 14 | 10.8 | 3.2 | 0.12 |
| 26 | 14.3 | 11 | 3.2 | 0.13 |
| 27 | 14.6 | 11.3 | 3.3 | 0.13 |
| 28 | 14.9 | 11.5 | 3.4 | 0.13 |
| 29 | 15.2 | 11.8 | 3.5 | 0.14 |
| 30 | 15.6 | 12 | 3.5 | 0.15 |
| 31 | 15.9 | 12.3 | 3.6 | 0.15 |
| 32 | 16.3 | 12.6 | 3.7 | 0.16 |
| 33 | 16.7 | 13 | 3.8 | 0.16 |
| 34 | 17.2 | 13.3 | 3.9 | 0.17 |
| 35 | 17.6 | 13.7 | 4 | 0.18 |
| 36 | 18.1 | 14.1 | 4.1 | 0.19 |
| 37 | 18.7 | 14.5 | 4.2 | 0.2 |
| 38 | 19.2 | 14.9 | 4.3 | 0.21 |
| 39 | 19.9 | 15.4 | 4.4 | 0.22 |
| 40 | 20.5 | 15.9 | 4.6 | 0.23 |
| 41 | 21.2 | 16.5 | 4.7 | 0.25 |
| 42 | 22 | 17.1 | 4.9 | 0.26 |
| 43 | 22.9 | 17.8 | 5.1 | 0.28 |
| 44 | 23.8 | 18.5 | 5.3 | 0.3 |
| 45 | 24.9 | 19.4 | 5.5 | 0.33 |

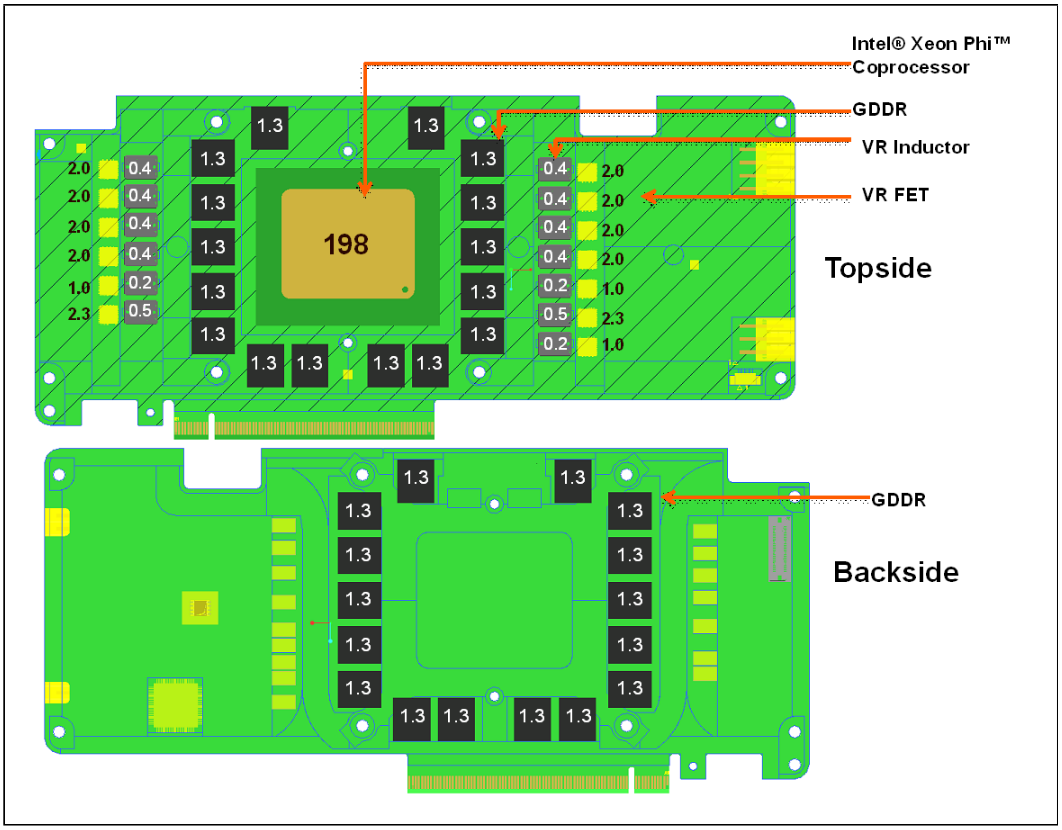
## 3.4 Cooling Solution Guidelines for SE10X

The Intel® Xeon Phi™ coprocessor SE10X SKU is shipped without a thermal solution, which gives system designers and integrators an opportunity to fit this SKU into their custom designed chassis. This SKU has GDDR components on the back side that must be cooled, in addition to the front side where the coprocessor resides.This section documents thermal and mechanical specifications and guidelines that would be useful to developers of custom designs.

### 3.4.1 Thermal Considerations

Figure 3-9 and Figure 3-10 show a schematic representation of the power profiles of the Intel® Xeon Phi™ coprocessor SE10X product.

**Figure 3-9 SE10X Power Profile for Coprocessor Intensive Workload (all values in Watts)**



### Figure 3-10 SE10X Power Profile for Memory Intensive Workload (all values in Watts)

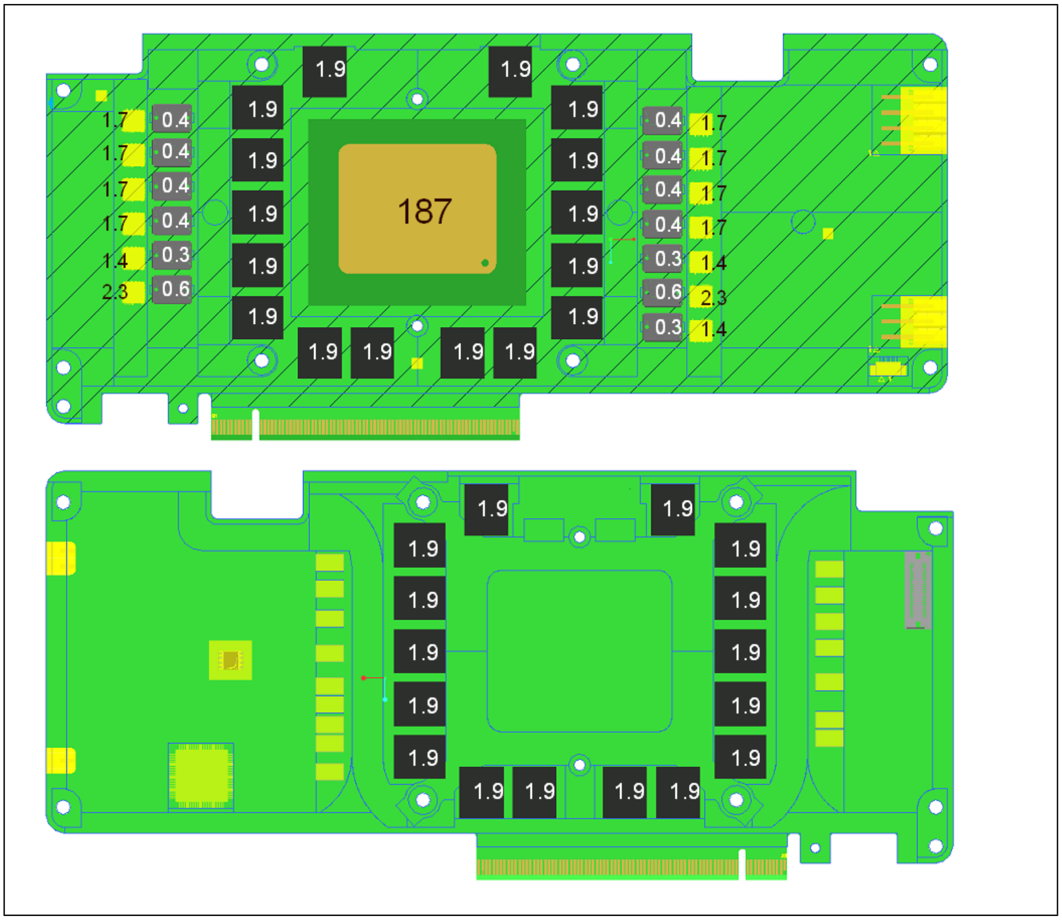


Table 3_3 shows thermal specifications of components present on the SE10X.

### Table 3_3. Component Thermal Specification on SE10X

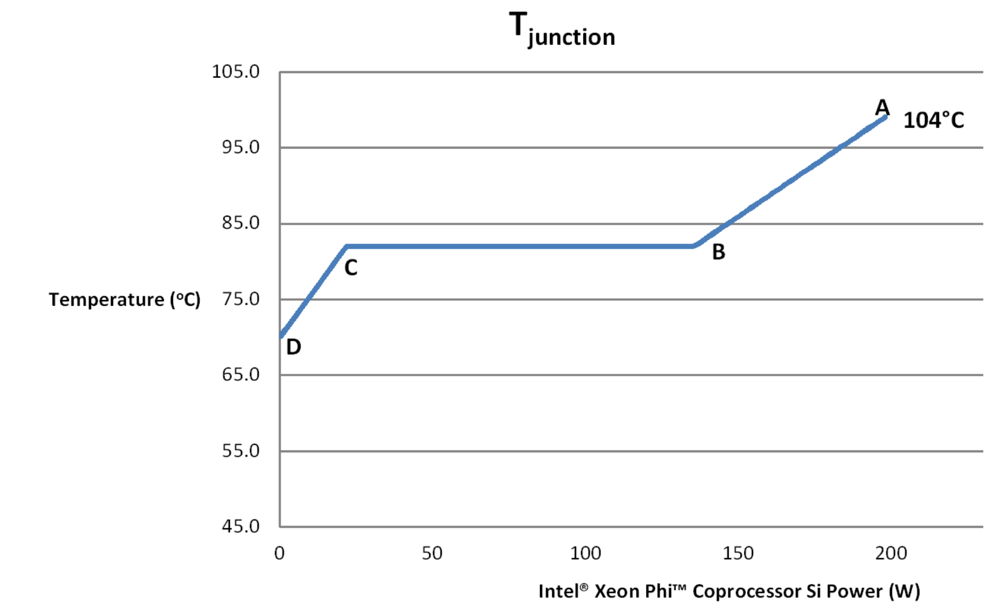
| Component | Thermal specification |
|---|---|
| Coprocessor | $T_{case} \leq$ 92°C |
| GDDR | $T_{case} \leq$ 85°C |
| VR FET | $T_{junction} \leq$ 150°C[1] |
| VR Inductor | $T_{body} \leq$ 100°C |

***Notes:***

1. While this is the component specification, on the passive and active Intel® Xeon Phi™ coprocessor products, the junction temperature is limited to 135°C in order to prevent damage to the PCB.

The simplest cooling mechanism would involve running fans at full speed. For those custom aircooled solutions that intend to be economical in fan power usage and acoustics, Figure 3-11 represents three regions on the SE10X coprocessor power consumption curve relevant to system fan control.

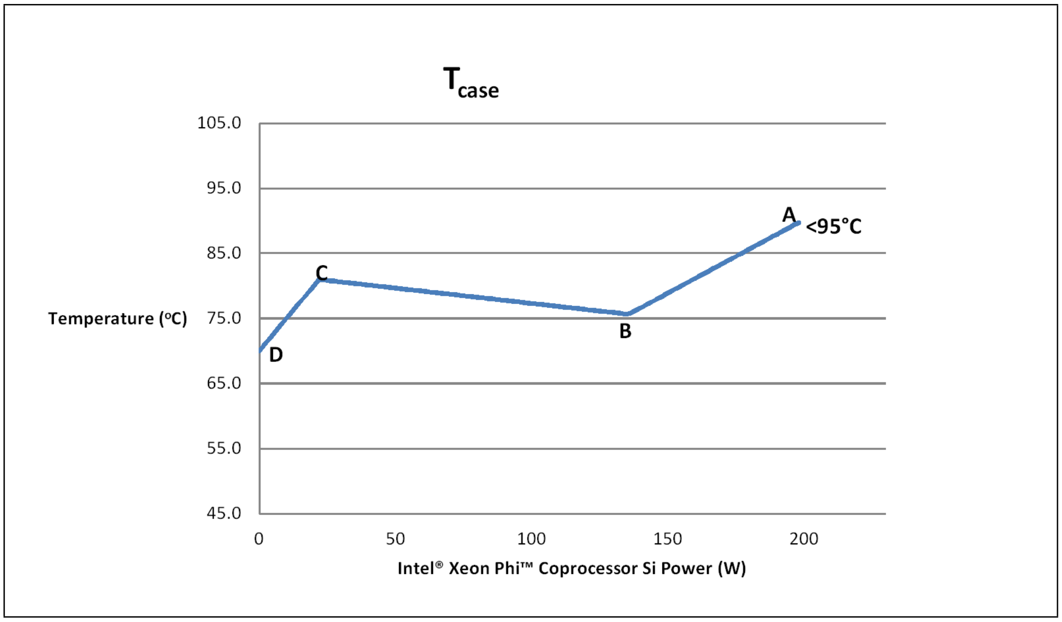
### Figure 3-11 SE10X SKU Coprocessor Junction Temperature ($T_{junction}$) vs Power



Region (A-B) on the line represents the minimum necessary performance of a cooling solution to keep the coprocessor silicon temperature ($T_{junction}$) below $T_{throttle}$ of 104°C (Table 3-2), during high power dissipation. In this region, a cooling solution based on airflow would ensure the fans are operating at 100% capacity. In region B-C, the coprocessor power consumption is low enough that the cooling solution may be set to maintain the junction temperature at a target temperature. Finally, in region C-D, the coprocessor may need to be cooled to below the target temperature to maintain a reasonable exhaust air temperature.

Figure Figure 3-12 shows the analogous thermal behavior of $T_{case}$.

### Figure 3-12 SE10X SKU Coprocessor Case Temperature ($T_{case)}$ vs Power



For the region A-B, the cooling solution must maintain the case temperature below 95°C which will in turn maintain the coprocessor silicon junction temperature below 104°C. Assuming an air-cooled heat sink, at a maximum coprocessor power dissipation of 198W (Figure 3-9) and an inlet air temperature of 45°C, the following equation between coprocessor junction-to-case and case-to-air heat sink rating can be used to determine the minimum necessary performance of a cooling system:

$T_{junction}=\Psi_{jc} * CPU_{power} + \Psi_{ca\_req} * CPU_{power} + T_{ambient}$

The heat sink must have a $\Psi_{ca\_req}$ value adequate to keep the coprocessor junction temperature at or below 104°C. The value for $\Psi_{jc}$ is a characteristic of the Intel® Xeon Phi™ coprocessor and may be treated as 0.047, a constant.

As the coprocessor power level goes down (region B-C), it is desirable to keep the junction temperature at or below a target temperature, here shown at 82°C. Since each coprocessor is programmed at the factory with the actual control temperature ($T_{control}$), a sophisticated cooling system may continuously read the junction temperature from the card SMC and compare it to the programmed $T_{control}$ to adjust airflow. The change in airflow over an air cooled heat sink affects the $\Psi_{ca}$ value. It is common to reduce fan speed when maximum airflow is not needed to save power, reduce noise, or both.

In the B-C region, even though $T_{junction}$ is at a constant value, $T_{case}$ actually goes up a little bit at lower power consumption levels. This is because a variable fan speed results in a variable $\Psi_{ca}$, but a fixed $\Psi_{jc}$.

Finally, in the C-D region where the coprocessor consumes very little power, an air cooled heat sink using a variable fan speed to maintain a target junction temperature may slow the airflow down too much. If the airflow is too low, the junction temperature may be maintained properly, but the exhaust air temperature approaches the junction temperature.  Data center design considerations, including safety, may dictate that a maximum allowable exhaust air temperature, such as 70°C, which in turn will set a maximum limit on $\Psi_{ca}$.

## 3.4.2 Mechanical Considerations

- In the passive Intel® Xeon Phi™ coprocessor products, the only component on the card with IHS load is the coprocessor. The compressive load is assumed to be approximately uniformly distributed over the IHS. The minimum load is 23lbf and maximum load is 75lbf. The mean pressure on the IHS is 33lbf.
- Honeywell PTM3180 is recommended as the thermal interface material (TIM).
- The gap filler used is the Bergquist 3500S35.
- The Intel passive heat sink is designed to nominal gaps of
  - — GDDR: 0.3 +/- 0.1225 mm
  - — VR FETs: 0.511 +/- 0.1225 mm
  - — VR Inductors: 0.5 +/- 0.2 mm

Table 3_4 shows the maximum heights of the different components on the SE10X product, along with the heights used in the product board design. Figure 3-13 and Figure 3-14 show the front and back sides of the SE10XSKU. Refer to the *Intel® Xeon Phi™ Coprocessor Thermal and Mechanical Models* document for the SE10X SKU.

### Table 3_4. Board Component Heights

| Block | Color[1] | Component Height (mils) | | |
|---|---|---|---|---|
| | | Min | Typ | Max |
| Coprocessor | | 171.221 | 177.992 | 184.763 |
| GDDR | Orange | | 47 | 47.25 |
| VR Inductor | Yellow | | 217 | 217 |
| VR phase controller | Red | | 35 | 39.37 |
| Coprocessor VR controller | Green | | 37 | 37 |
| GDDR VR controller | Pink | | 35 | 35.43 |
| Capacitor topside | Purple | | 49 | 49 |
| Capacitor backside | Light blue | | 83 | 83 |

***Notes:***

1. Colors are in reference to Figure 3-13 and Figure 3-14.

### Figure 3-13 SE10X Board Top Side

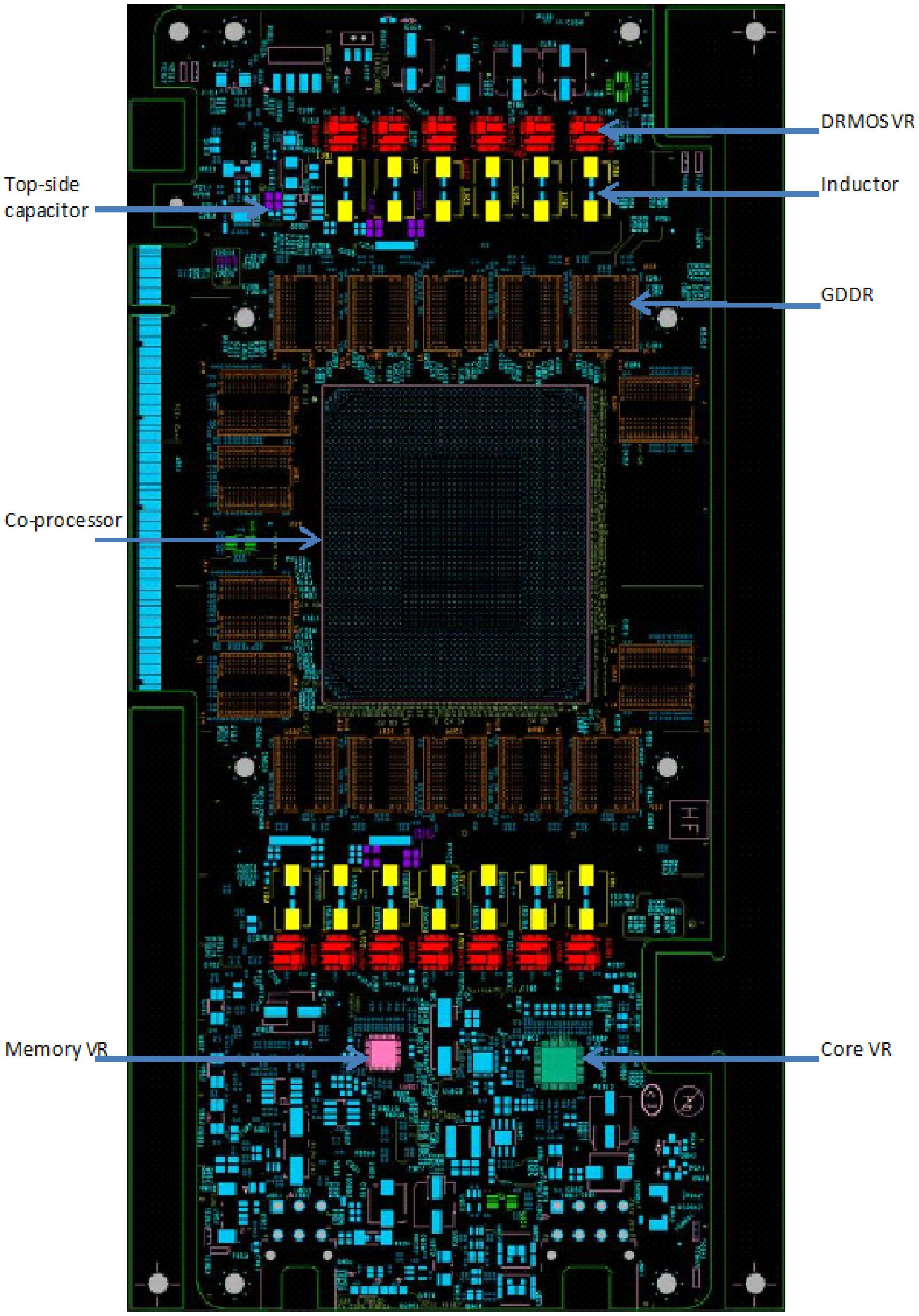
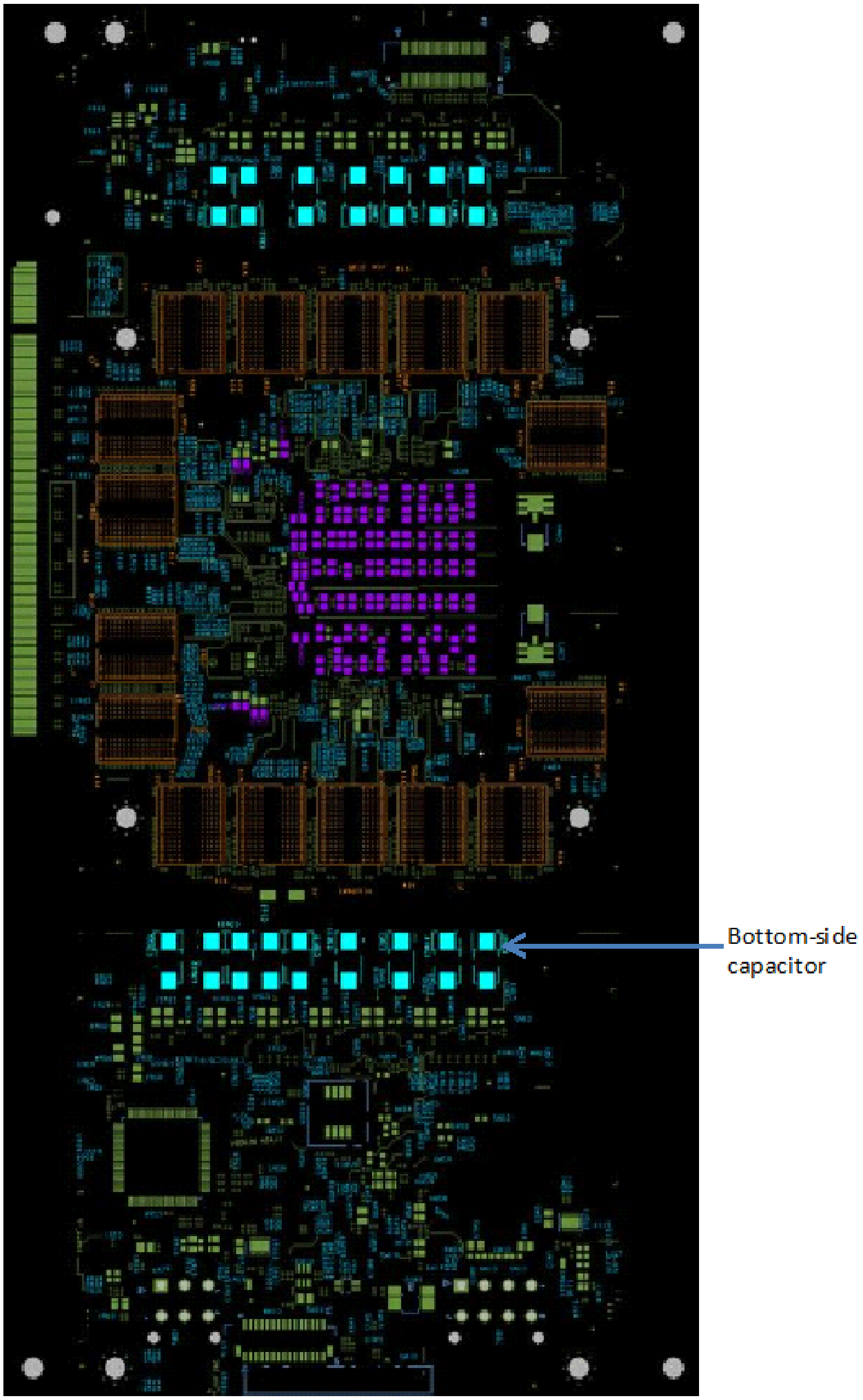


Figure 3-14 SE10X Board Bottom Side



### 3.4.3 Mechanical Shock and Vibration Testing

Table 3_5 shows the recommended shock and vibration guidelines, and dynamic load shift specifications.

Table 3_5. Dynamic Load Shift Specification

| Test | Specification and Guidelines |
|---|---|
| Board Unpackaged Shock | 50g trapezoidal; V:170in/s<br>drops: 3x each on 6 faces |
| Board Unpackaged Random Vibration | 5Hz @ 0.01g$^2$/Hz to 20Hz @0.02g$^2$/Hz (slope up)<br>20Hz to 500Hz @ 0.02g$^2$/Hz (flat)<br>Input acceleration is 2313g RMS<br>10mins per axis in all 3 axis |
| System Unpackaged Shock | 25g trapezoidal; Varies by system weight (20-39lbs: 225 in/sec; 40-79lbs: 205 in/sec)<br>drops: 2x each of 6 faces |
| System Unpackaged Random Vibration | 5Hz @ 0.001g$^2$/Hz to 20Hz @0.001g$^2$/Hz (slope up)<br>20Hz to 500Hz @ 0.001g$^2$/Hz (flat)<br>Input acceleration is 2.20g RMS<br>10mins per axis in all 3 axis |

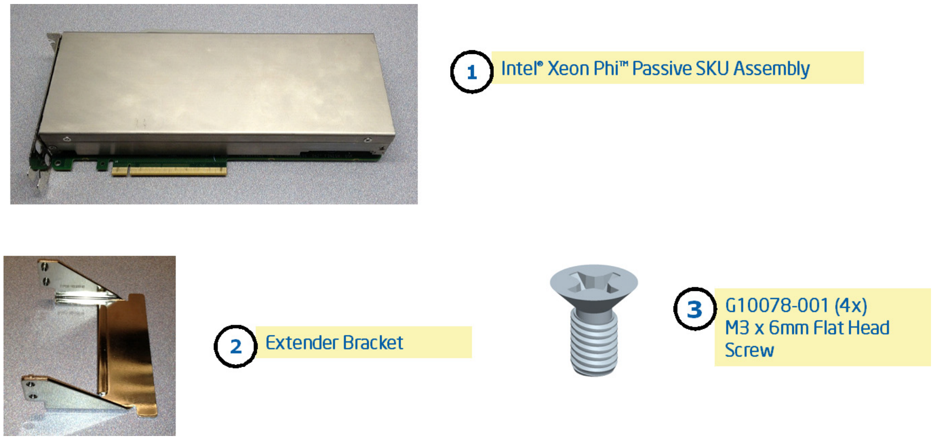
## 3.5 Intel® Xeon Phi™ Coprocessor PCI Express* Card Extender Bracket Installation

Intel® Xeon Phi™ coprocessors are shipped without the PCI Express* bracket (also known as extender bracket) installed on the card. The purpose of this bracket is to interface with the chassis mechanical card guides for standard full-length PCI Express* cards. In the shipped package, customers should expect to find:

- 1 Intel® Xeon Phi™ coprocessor with assembled thermal solution.
- 1 Intel® Xeon Phi™ coprocessor card extender bracket.
- 4 M3 x6mm flat head screws.

***Note:*** The SE10X SKU is not shipped with the extender bracket.

Figure 3-15 Contents of Intel® Xeon Phi™ Coprocessor Package Shipment



### 3.5.1 Step 0: Determine Lid Type

If lid type is "overlap" where lid covers top mounting holes as shown in Figure 3-16, then go to Step 1.

If lid type is "clearance" where lid has cut-out for mounting holes as shown in Figure 3-17, then go to Step 2.

Figure 3-16 Overlap Lid

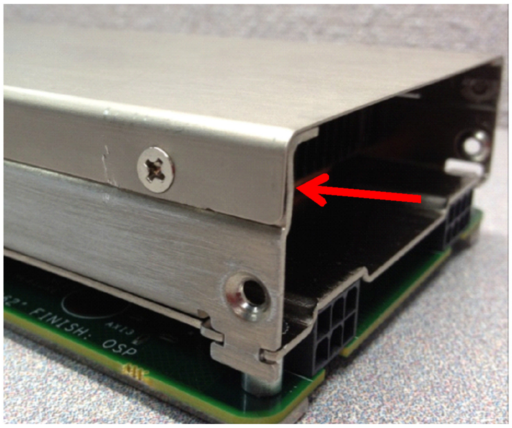
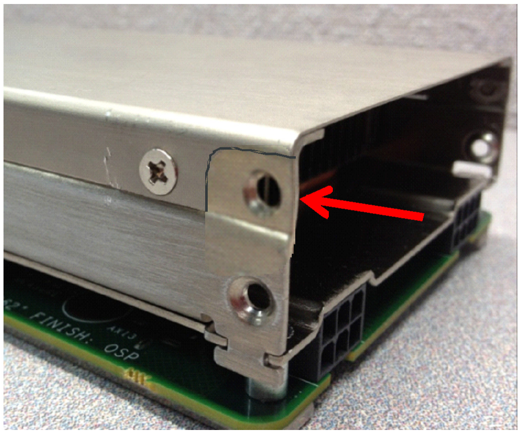


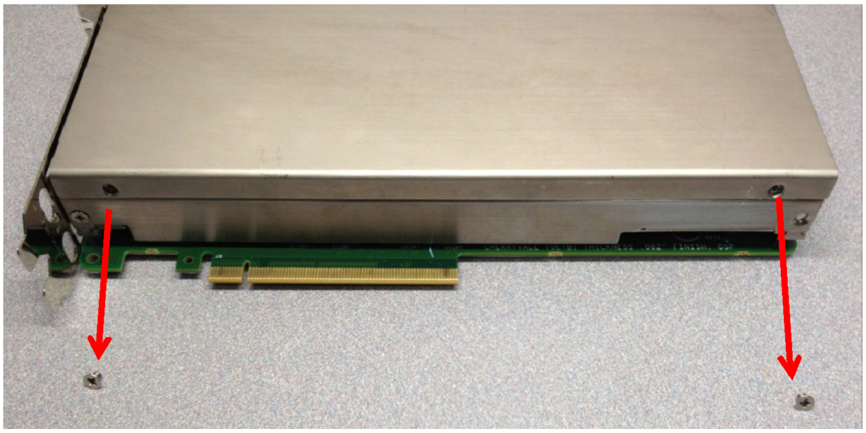
Figure 3-17 Clearance Lid



## 3.5.2 Step 1: Overlap Lid Removal

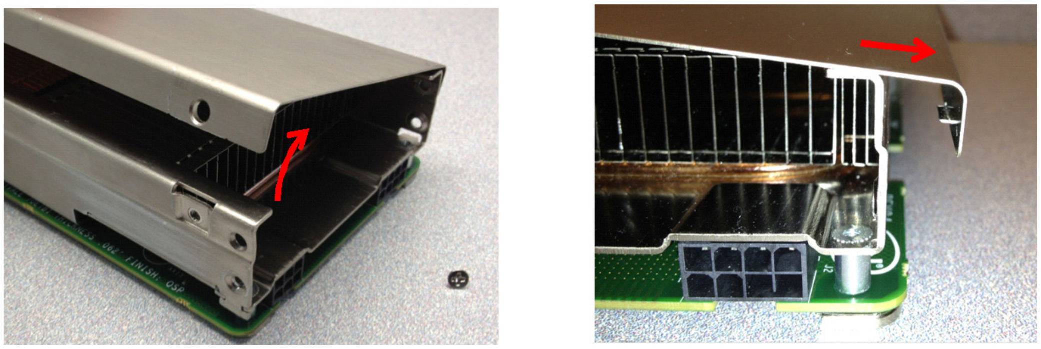
a. Remove 2 of the M3x6mm screws retaining the lid, as shown in Figure 3-18.

Figure 3-18 Overlap Lid Removal



b. Remove Lid. Take care not to bend tabs, as shown in Figure 3-19.

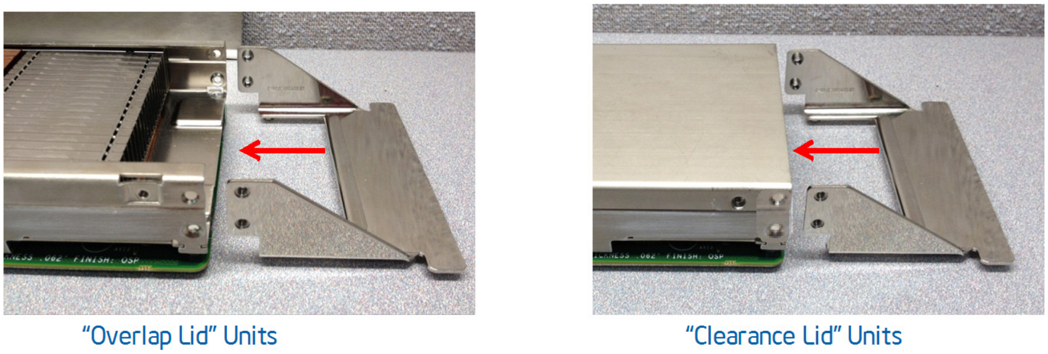
Figure 3-19 Tilt Overlap Lid and Slide as shown to Disengage Tabs



### 3.5.3 Step 2: OEM Bracket Installation

a. Insert the OEM bracket into the Intel® Xeon Phi™ coprocessor card assembly, as shown in Figure 3-20.

Figure 3-20 OEM Bracket Installation

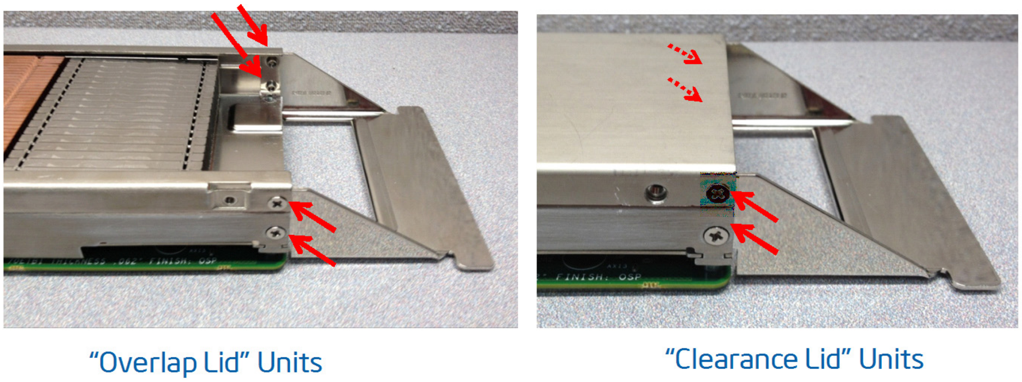


"Overlap Lid" Units

"Clearance Lid" Units

b. Install (4) M3 x 6mm Flat Head Screws; torque = 6inch-lbs, shown in Figure 3-21.

Figure 3-21 OEM Bracket Installation



"Overlap Lid" Units

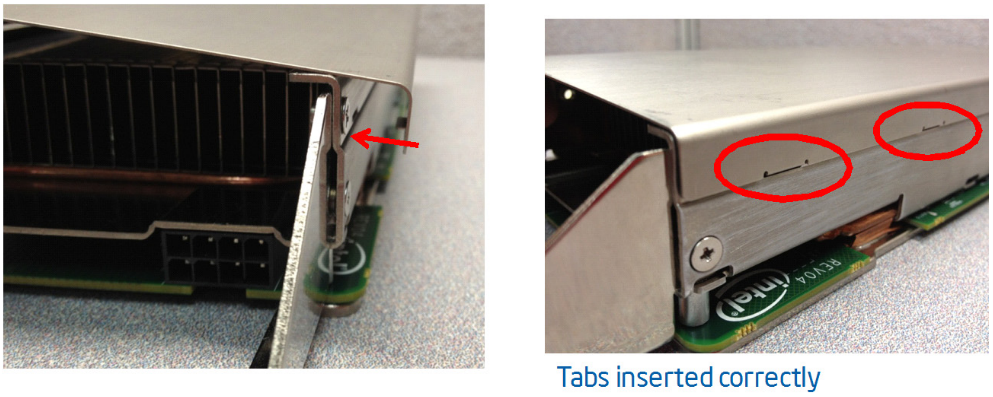
"Clearance Lid" Units

At this point, "clearance lid" units are ready to be mounted in the chassis.

## 3.5.4 Step 3: Replace Lid on "Overlap Lid" Units Only

a. Insert tabs into slots in card assembly, shown in Figure 3-22.

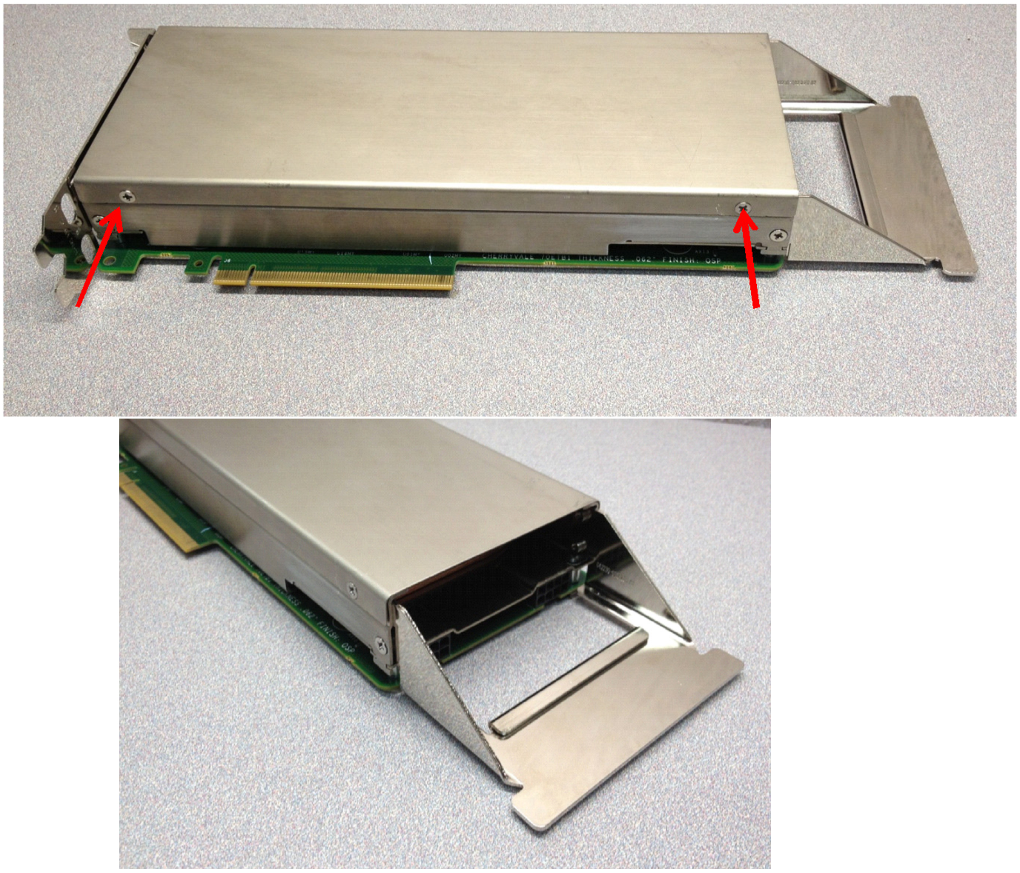
Figure 3-22 Replace Lid on "Overlap Lid" Units



Tabs inserted correctly

b. Install the lid's screws (M3 x 6mm Flat head); torque = 6 inch-lbs, shown in Figure 3-23.

Figure 3-23 Replace Lid on “Overlap Lid” Units (cont.)



# 4 Intel® Xeon Phi™ Coprocessor Pin Descriptions

## 4.1 PCI Express* Signals

The PCI Express* connector for the Intel® Xeon Phi™ coprocessor is a x16 interface and supports signals defined in the "*PCI Express® Card Electromechanical Specification"*. Not all signals called out in the PCI Express* specification are utilized on the Intel® Xeon Phi™ coprocessor, and listed as "not used" in Table 4-1.

The symbol _N at the end of a signal name indicates that the active or asserted state occurs when the signal is at a low voltage level. When _N is not present after the signal name, the signal is asserted when at the high voltage level.

The following notations are used to describe the signal type:

I Signal is an Input to the Intel® Xeon Phi™ coprocessor

O Signal is an Output from the Intel® Xeon Phi™ coprocessor

I/O Bidirectional Input/Output signal

S Sense pin

P Power supply signal, sourced from the PCI Express* edge fingers or supplemental power connectors.

**Table 4-1. PCI Express* Connector Signals on the Intel® Xeon Phi™ coprocessor**

| Signal Name | Signal Type | Description |
|---|---|---|
| EXP_A_TX_[15:0]_DP<br>EXP_A_TX_[15:0]_DN | O | PCI Express* Differential Transmit Pairs: 16-channel differential transmit pairs, referenced to the Intel® Xeon Phi™ coprocessor. The EXP_A_TX_[15:0]_DP and EXP_A_TX_[15:0]_DP are connected to the PCI Express* device transmit pairs on the Intel® Xeon Phi™ coprocessor. |
| EXP_A_RX_[15:0]_DP<br>EXP_A_RX_[15:0]_DN | I | PCI Express* Differential Receive Pairs: 16-channel differential receive pairs referenced to the Intel® Xeon Phi™ coprocessor. The EXP_A_RX_[15:0]_DP and EXP_A_RX_[15:0]_DP are connected to the PCI Express* device receive pairs on the Intel® Xeon Phi™ coprocessor. |
| CK_PE_100M_16PORT_DP<br>CK_PE_100M_16PORT_DN | I | PCI Express* Reference Clock: 100MHz differential clock I to Intel® Xeon Phi™ coprocessor for use by the coprocessor to properly recover data from the PCI Express* Interface. |

### Table 4-1. PCI Express* Connector Signals on the Intel® Xeon Phi™ coprocessor

| Signal Name | Signal Type | Description |
|---|---|---|
| RST_PCIE_N | I | PCI Express* Reset Signal: RST_PCIE_N is a 3.3-volt active-low signal that when deasserted (high) indicates that the +12V and VCC3 power supplies are stable and within their specified tolerance. |
| SMB_PCI_CLK | I/O | PCI Express* System Management Bus Clock: SMB_PCI_CLK is the 3.3-volt clock signal for the SMBus Interface, which is normally used for power and/or thermal management and for monitoring the card. |
| SMB_PCI_DAT | I/O | PCI Express* System Management Bus Data: SMB_PCI_DAT is the 3.3-volt data signal for the SMBus Interface, which is normally used for power and/or thermal management and for monitoring the card. |
| PRSNT1_N, PRSNT2_N | S | Following PCI Express* specification, PRSNT1_N# (pin A1) is connected on the coprocessor card to PRSNT2_N (pin B81). Remaining PRSNT2_N pins (17, B31, B48) must be unconnected on the baseboard. |
| VCC3 | P | +3.3V Supply: The positive 3.3-volt power supply to the PCI Express* card. |
| +12V | P | +12V Supply: The positive 12-volt power supply to the PCI Express* card. |
| V_3P3_PCIAUX | P | +3.3VAux Supply. |
| PROCHOT_N (pin B12) | I | On the Intel® Xeon Phi™ coprocessor, the SMC supports an external path from the baseboard to the card's B12 pin, which allows system agents such as BMC or ME to throttle the card in response to card thermal events (thermal throttling). Pin B12, defined as reserved in the PCI Express* specification, has been renamed PROCHOT_N on the Intel® Xeon Phi™ coprocessor and is driven by 3.3V power. This pin is held in active-high state by the card SMC, and must be driven active-low by the baseboard to exert throttling. This feature is not available on the 3100 series active SKU. See Section 4.1.1 and Chapter 6 for details. |
| WAKE_N | Not Used | PCI Express* Wake Signal. |
| EXP_JTAG[5:1] | Not Used | PCI Express* JTAG Interface. |

### 4.1.1 PROCHOT_N (Pin B12)

System baseboard routing to the PROCHOT_N pin must take into consideration the following details:

- PROCHOT_N pin is driven by the +3.3V power rail.
- PROCHOT_N pin is connected to a pull-down of 100k-ohm on the card.
- The input signal arriving at the pin from the baseboard must meet the following characteristics:

  - — VIH(min)= 2.7V
  - — VIL(max)= 0.5V
  - — Rise/Fall times(max)= 240ns
- The baseboard implementation can choose to be either push-pull or open-drain. In particular, an open-drain implementation must provide a pull-up resistor of 10k-ohm or less on the baseboard to counteract the pull-down on the card.

## 4.2 Supplemental Power Connector(s)

The Intel® Xeon Phi™ coprocessor gets only maximum 75W from the PCI Express* connector, per the PCI Express* specification. The 2x4 and 2x3 supplemental power connectors on the coprocessor card provide the additional +12-volt power needed by the coprocessor. Per the PCI Express* specifications, the 2x4 connector must be capable of maximum 150W power draw by the coprocessor, and the 2x3 must be capable of maximum 75W power. The 300W TDP products of the Intel® Xeon Phi™ coprocessor family must have provide power to the 2x4 and the 2x3 connectors. The 225W products can have either a single 2x4 connector connected to a power supply, or two 2x3 connectors (each capable of maximum 75W power draw). Within the coprocessor, the power rails from the three sources are not connected to each other. Instead, the Intel® Xeon Phi™ coprocessor is designed to draw power proportionally from the three power sources. During coprocessor power-up, sensors on the coprocessor card detect presence of power supplies to the supplemental connectors, and depending on the maximum TDP of the coprocessor, can determine if sufficient power is available to power up the card. For example, sensors on a 300W coprocessor card must detect both 2x4 and 2x3 power supplies in order for the card to be powered up and function properly.

# 5 Power Specification and Management

Power management in the Intel® Xeon Phi™ coprocessor is primarily managed via the on-board resident coprocessor OS with hardware-controlled functionality. Table 5-1 shows estimates for coprocessor power states and respective memory power states, along with estimates of corresponding card power and wakeup times.

**Table 5-1. Intel® Xeon Phi™ Coprocessor Power States**

| Coprocessor Power State | Memory Power State[1] | SE10P/SE10X Card Power (Watts) | 5110P Card Power[2] (Watts) | Wakeup Time[3] |
|---|---|---|---|---|
| C0 | M0 | 300 | 225 | N/A |
| C1 | M1/M2 | <115 | | <1µs (<5ms M2) |
| Auto-pC3 | M3 | <105 | | ~10µs (<5ms M3) |
| Deep-pC3 | M3 | <40 | | ~10-40µs (<5ms M3) |
| pC6 | M3 | Not supported | | ~4-12 ms (<5ms M3) |

***Notes:***

1. In many cases, memory power state latencies will govern the overall card wakeup time.
2. Refer Section 5.1.
3. Wakeup times are shown to provide orders of magnitude comparisons only, and may change based on part characterization.

As the Intel® Xeon Phi™ coprocessor is being powered-on, it is expected to draw measurable amount of current from each of the power rails connected to the coprocessor card. Below is the current and power drawn from each source during the power-on phase of the SE10P SKU:

- +12V 2x4 connector: 5.4A (~64W)
  A peak current of 7.1A for duration of 40µs
- +12V 2x3 connector: 3A (~36W)
- +12V PCI Express* slot pins: 1.6A (~20W)
- +3.3V PCI Express* slot pins: 1.3A (~5W)
- Measured total power consumption: ~120W

**Note:** *The above power-on measurements were taken with a single coprocessor card, using a specific open chassis system. It is not indicative of the coprocessor behavior in all types of systems in which the coprocessor will be used. The current and power values are meant to be guidelines for system power planning, and not specification of the Intel® Xeon Phi™ coprocessor.*

## 5.1 5110P SKU Power Options

Most HPC applications running on the 5110P SKU are expected to draw less than 225W, but the card is designed to support power surges above 225W. If the power surge goes above 236W for more than 300ms, then the SMC on the card will instruct the Intel® Xeon Phi™ coprocessor to drop its operating frequency by approximately 100MHz, thus reducing power dissipation by approximately 10W. If power surge goes above 245W for more than 50ms, then the SMC will assert the PROCHOT_N signal to the coprocessor, which will cause the frequency to drop to the minimum possible value (refer to Section 3.2.1). The level and duration of the power surge are programmable by the end user (refer chapter on manageability for more details).

Additionally, there may be applications such as HPC Linpack that may draw up to 245W (Table 5-2). This should be taken into account when choosing one of the three modes of operation as listed below:

- — Users can install both the 2x4 and 2x3 power connectors for total available power of 300W. In this case, the card may draw up to 245W of power depending on the application. This mode ensures sufficient power is available and reduces the risk of throttling. Users may see power dissipation approach 245W, as applications become more highly tuned to take advantage of the Intel® Xeon Phi™ coprocessor architecture.
- — Users can install either the 2x4 connector only or two 2x3 connectors for total available power of 225W. The card is designed to support power surges of up to 236W. If the power surge goes above 236W for more than 300ms, then the SMC on the card will instruct the Intel® Xeon Phi™ coprocessor to drop its operating frequency by approximately 100MHz, thus reducing power dissipation by approximately 10W.
- — If a greater card power limitation is desired, users can configure the SMC to further limit the power draw of the 5110P SKU, ensuring compatibility with less capable power delivery systems (refer to Section 6.5).

**Table 5-2. HPC Linpack Power Guidelines**

| HPC Workload | Card SKU | Card Power on Workload (W) |
|---|---|---|
| Linpack DP | SE10P/SE10X | 300 |
| Linpack DP | 5110P | 245 |

***Note:*** Results may change with Intel® Xeon Phi™ coprocessor steppings, frequencies, cluster configurations, changes in the card coprocessor OS and workload binaries.

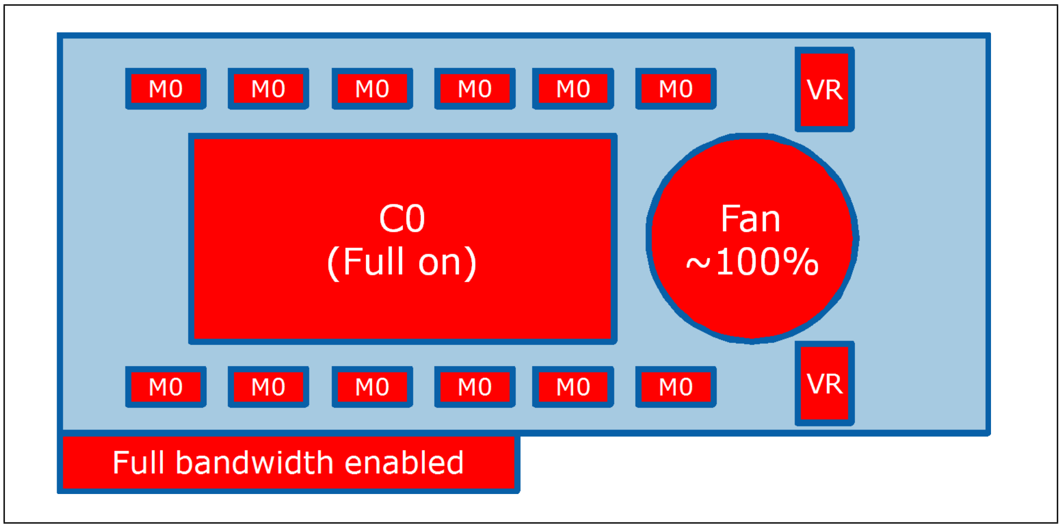
- — All workloads provided by Intel and run as a native offload application.
- — Test system setup:  dual Intel® Xeon® E5-2680 CPUs, 32GB RAM; two passive Intel® Xeon Phi™ coprocessor cards (61 core) per Intel® Xeon® processor
- — All measurements taken with LabVIEW® Signal Express @ 25C ambient temperature.
- — HPC Linpack was run for about 5 minutes to reach thermal saturation before measuring power.

## 5.2 Intel® Xeon Phi™ Coprocessor Power States

Figure 5-1 to Figure 5-8 are a schematic representation of the inter-relationship between the different coprocessor and memory power states on Intel® Xeon Phi™ coprocessor.

*These schematic representations are only for illustrative purposes and do not represent all possible low power states.*

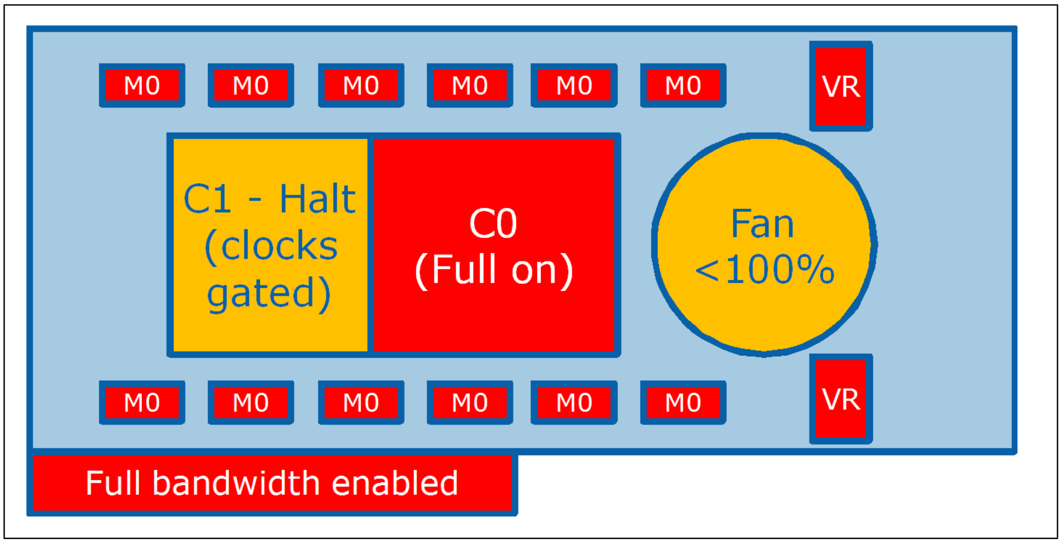
### Figure 5-1. Coprocessor in C0-state and Memory in M0-state



In this power state, the card is expected to operate at its maximum TDP rating.

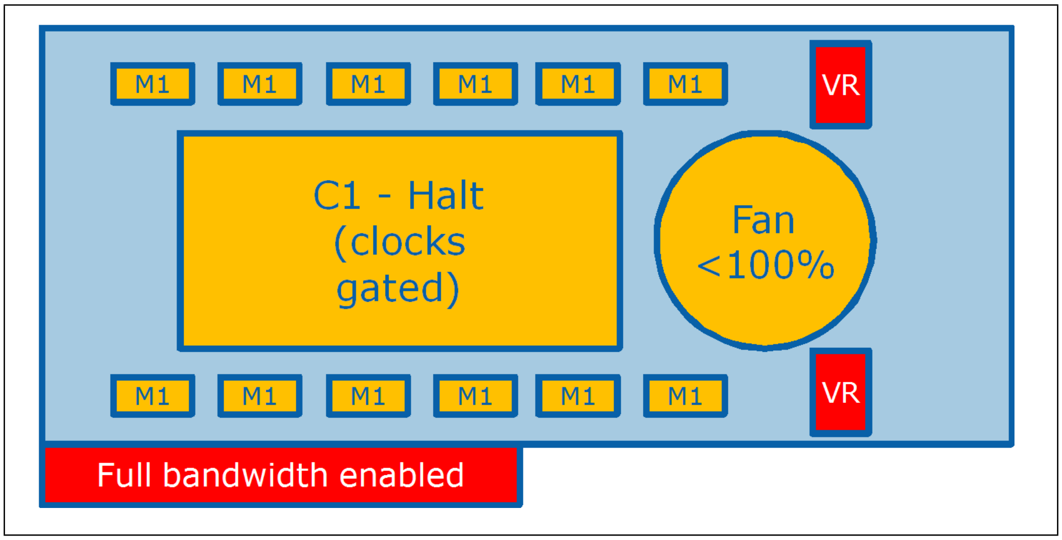
***Note:*** No application is expected to dissipate maximum power from cores and memory simultaneously.

### Figure 5-2. Some cores are in C0-state and other cores in C1-state; Memory in M0-state



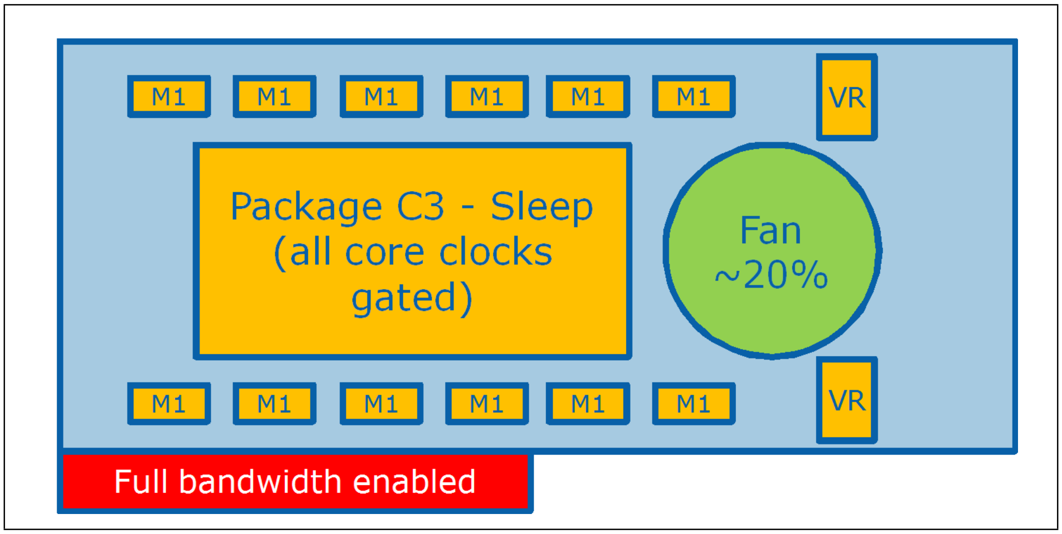
Coprocessor C1 state gates clocks on a core-by-core basis, reducing core power. On the active SKU, the fan slows to an appropriate speed, reducing fan power. If all cores enter C1, the coprocessor automatically enters Auto-pC3 state.

**Figure 5-3. All Cores In C1 state; Memory In M1 state**



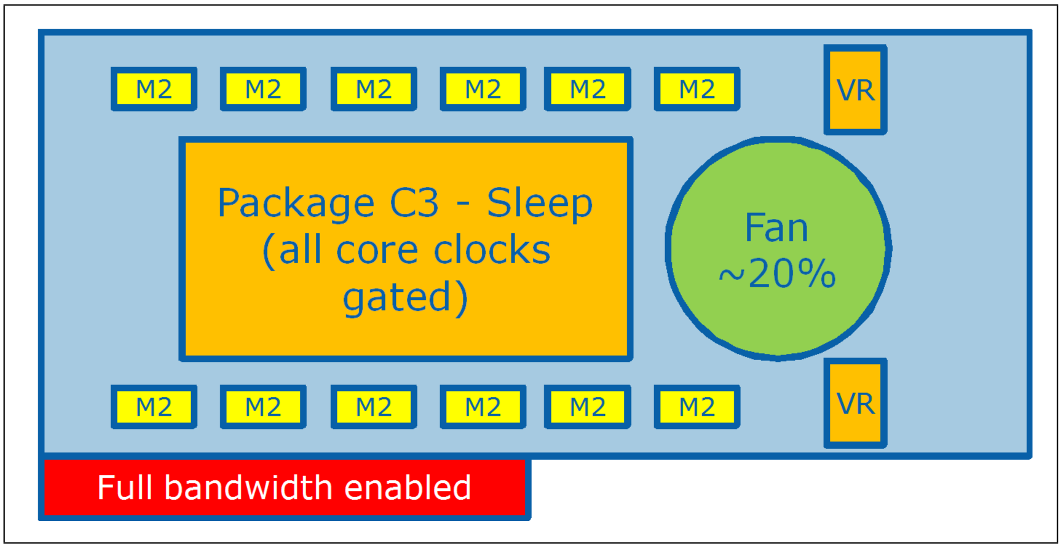
If clock-enable input to memory is pulled high, then memory enters M1 state which reduces memory power.

**Figure 5-4. All Cores In Package-C3 State; Memory In M1**



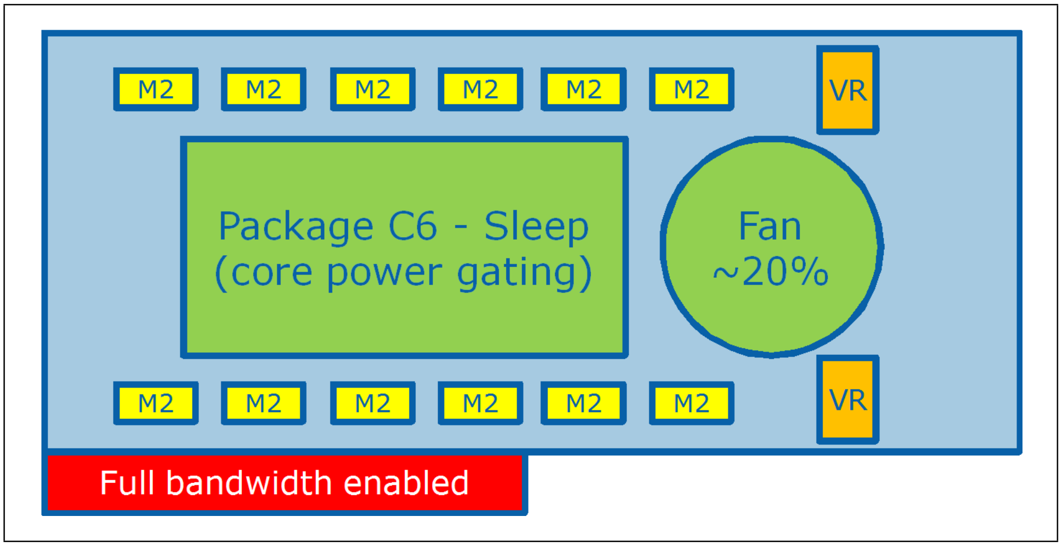
When all cores have entered C1 Halt state, the coprocessor package can reduce the core voltage and enter Deep-pC3. The fan (on active SKUs) can slow to minimum speed. VRs enter low power mode.

Figure 5-5. Package-C3 and Memory M2 state



From M1 state, memory can be put in self-refresh mode to enter the M2 state, further reducing memory power.

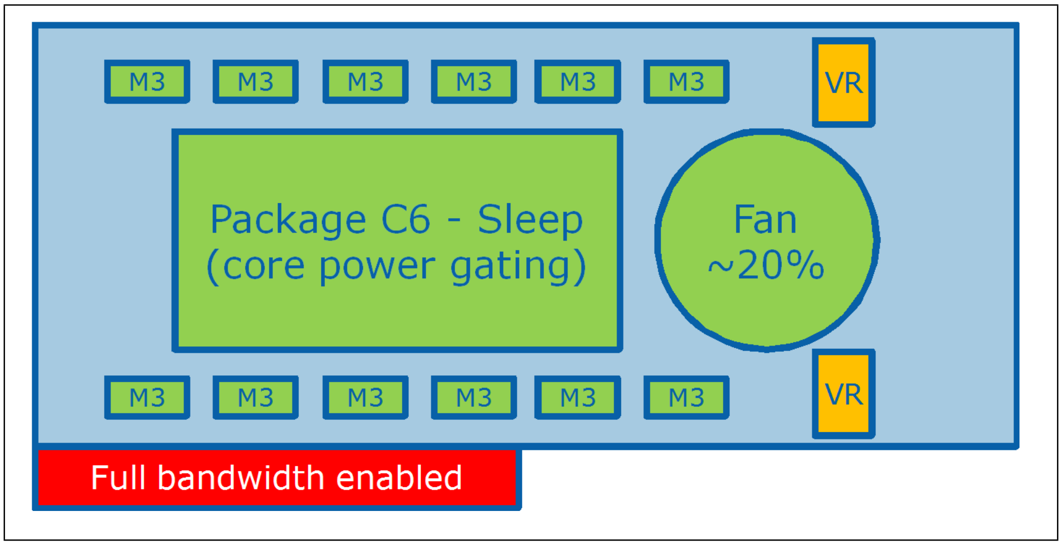
Figure 5-6. Package-C6 and Memory M2 state



The coprocessor OS can request that the coprocessor enter package C6 state. Core voltage is shut down. Coprocessor power is <10W[1] in this state.

[1] Value may be revised following silicon characterization

Figure 5-7. Package-C6 and Memory M3 state



The memory clock can be fully stopped, reducing memory power to its minimum state.

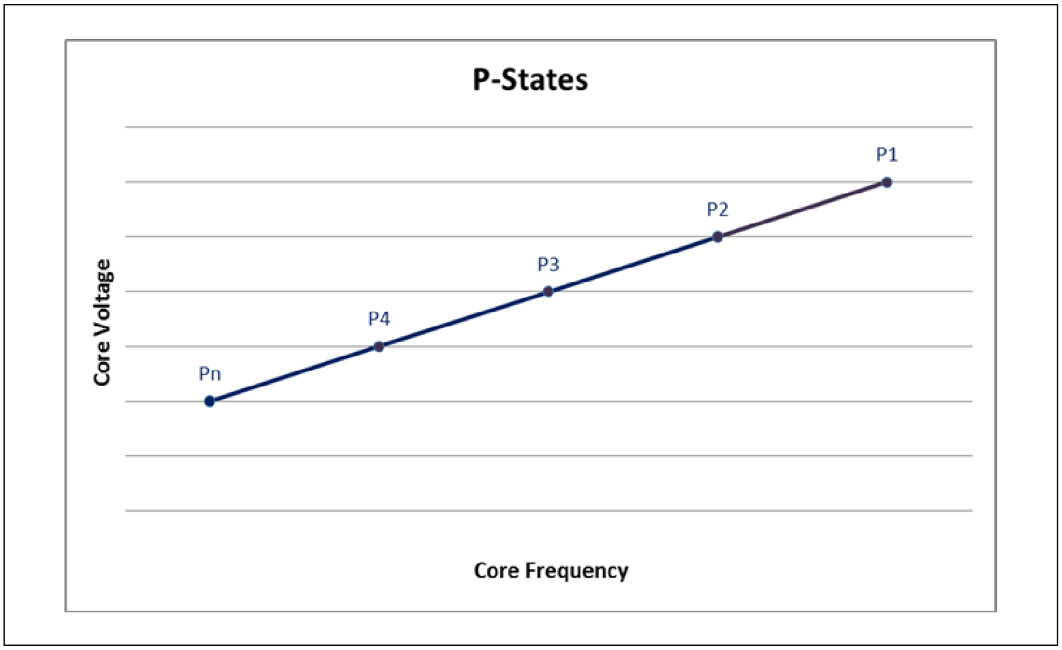
## 5.3 P-states and Turbo Mode

P-states, or Performance states, are different frequency settings requested by the host OS or application when the cores are in the C0 active/executing state. Switching between P-states is done by the coprocessor when the OS or application determines that more or less performance is needed. All active cores run at the same P-state frequency as there is only one clock source in the coprocessor.

Each frequency setting of the coprocessor requires a specific VID voltage setting in order to guarantee proper operation, and each P-state corresponds to one of these frequency and voltage pairs. Each device is uniquely calibrated and programmed at the factory with its appropriate frequency and voltage pairs. As a result, it is possible that two devices with the same frequency specification may have different voltage settings.

The highest P-state is P1, followed by sequentially lower frequency states of P2, P3…. with Pn being the lowest frequency state. All parts within a given SKU will have the same P-state settings, but P-state frequencies may vary across SKUs.

Figure 5-8. Intel® Xeon Phi™ Coprocessor P-states



# 6 Manageability

## 6.1 Intel® Xeon Phi™ Coprocessor Manageability Architecture

The server management and control panel component of the Intel® Xeon Phi™ coprocessor architecture provides a system administrator with the runtime status of the Intel® Xeon Phi™ coprocessor installed in a given system. There are two access methods by which the server management and control panel component may obtain status information from the Intel® Xeon Phi™ coprocessor. The "in-band" method utilizes the SCIF network and the capabilities designed into the μOS and the host driver to deliver the Intel® Xeon Phi™ coprocessor status. It also provides a limited ability to set specific parameters that control hardware behavior. The same information can be obtained using the "out-of-band" method. This method starts with the same capabilities in the μOS, but sends the information to the System Management Controller (SMC) using a proprietary protocol. The SMC can then respond to queries from the platform's BMC using the IPMI protocol to pass the information upstream to the administrator or user. For more information on the tools available for management see the *Intel® Xeon Phi™ Coprocessor System Software Developers Guide*.

## 6.2 System Management Controller (SMC)

Intel® Xeon Phi™ coprocessor manageability relies on a System Management Controller (SMC) on the card. The SMC provides sensor telemetry information for management by in-band (host) software and out-of-band software via the PCI Express* SMBus. The SMC also provides additional functionality as described in this chapter.

The SMC is a microcontroller-based thermal management and communications system that provides card-level control and monitoring of the Intel® Xeon Phi™ coprocessor. Thermal management is achieved through monitoring the Intel® Xeon Phi™ coprocessor and the various temperature sensors located on the coprocessor card. Card-level power management monitors the card input power and communicates current power conditions to the Intel® Xeon Phi™ coprocessor.

SMC features include:

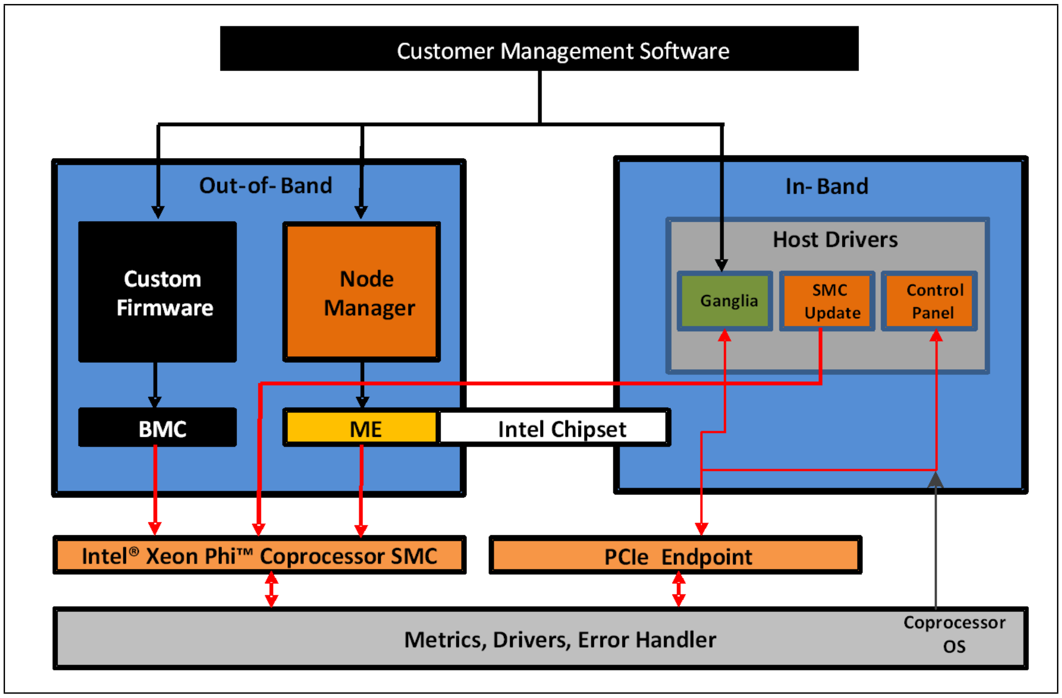
- Four thermal sensor inputs: inlet, outlet, coprocessor die, and GDDR.
- Power alert, thermal throttle, and THERMTRIP# signals.

The SMC connects to coprocessor silicon via the following I2C and out-of-band signals:

- In-band Communication
  - Software access to thermal and power metrics via Ganglia
  - gmond exposed via standard Ethernet port
  - Accessible via Control Panel GUI and API
- Out-of-band Communication
  - Access to the SMC via the PCI Express* SMBus using the IPMI IPMB protocol
  - 50ms sampling rate for power data

The manageability architecture also provides support for the Intel® Xeon Phi™ coprocessor in Node Manager mode, which adds functionality such as setting power limits and time windows.

**Figure 6-1. Intel® Xeon Phi™ coprocessor System Manageability Architecture**



In operational mode, the SMC monitors power and temperatures within the Intel® Xeon Phi™ coprocessor and through sensors located on coprocessor card. This information is then used to control the power consumed by the PCI Express* card and, in the case of the 3100 series active SKU, the rotating speed of the fan on the card. The SMC provides status information (temperature, fan speed, and voltage levels) to the Intel® Xeon Phi™ coprocessor drivers, which then can be provided to the end user via a GUI. The SMC provides a master/slave SMBus (using the IPMI IPMB protocol) so that a platform BMC or ME can control the SMC.

The SMC on the Intel® Xeon Phi™ coprocessor has the following capabilities:

- General manageability features
- Board ID and SKU definition
- Unique identifying number
- Fan Control
  - Read fan RPM
- Thermal throttling and throttle monitoring
  - Force throttling of the coprocessor
  - Monitor time in throttled state
  - Separated status if power limited throttling vs. overtemperature throttling
- Card-level power limiting/capping
  - Power Limit 0 and 1, tracked over separate time windows

  - — P-state clamping if the P-state requested is not possible within the set power envelope
- Power/energy measurement
  - — Can choose to include or preclude 3.3V power

## 6.3 General SMC Features and Capabilities

The Intel® Xeon Phi™ coprocessor supports the PCI Express* 2.0 standard. The SMC located on the card has direct access to information about the card operation (such as fan speeds, power usage, etc.) that must be managed from host-based software.

The SMC supports manageability interfaces via MMIO and the preferred PCI Express* SMBus (IPMI IPMB protocol) as well as with polled master only IPMI protocol.

The SMC firmware update process is resilient against unexpected power loss and resets.

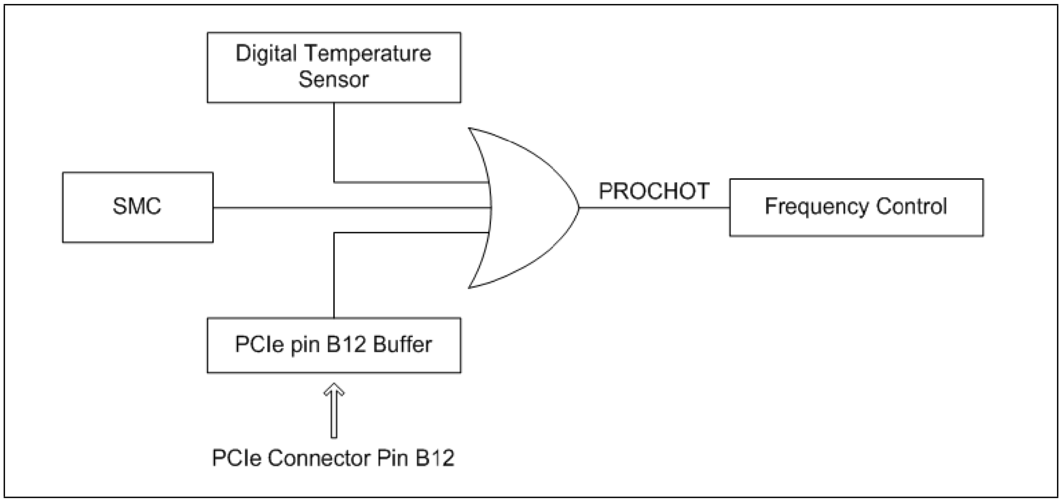
The SMC supports a read only IPMI compliant FRU that contains the following information:

- Manufacturer name
- Product name
- Part number / model number
- UUID
- Manufacturer’s IPMI ID
- Product IPMI ID
- Manufacturing time / date stamp
- Serial number (12 ASCII bytes)

To keep the Intel® Xeon Phi™ coprocessor within the operational temperature range, the SMC boosts the fan to full speed when either PERST or THERMTRIP_N are asserted on the 3100 series active SKUs with on-board fan. On SKUs with passive cooling solutions, the SMC will sample a GPIO pin on startup to determine if the closed loop fan control algorithm and monitoring should be disabled on certain SKUs.

Additionally the SMC supports enabling and disabling an external assertion path from the baseboard to the card pin B12. This allows an external agent, such as a BMC or ME, to force throttle the Intel® Xeon Phi™ coprocessor during thermal events. Pin B12, defined as *reserved* in the PCI Express* specification, has been renamed PROCHOT_N on Intel® Xeon Phi™ coprocessor and is driven by 3.3V power. This pin is held active-high (deasserted) by design, and must be driven active-low by the baseboard to exert throttling. An OEM IPMB message from the baseboard to the SMC is required to enable the external throttling mechanism. See Section 4.1.1 for baseboard implementation details.

Figure 6-2. Schematic Representation of PROCHOT_N on the Intel® Xeon Phi™ coprocessor



### 6.3.1 Catastrophic Shutdown Detection

Catastrophic shutdown is the act of the Intel® Xeon Phi™ coprocessor silicon shutting itself down to prevent damage to the device caused by overheating. The SMC monitors THERMTRIP_N to detect this event. When THERMTRIP_N is asserted (low), the SMC detects this and immediately forces the fan(s) to full speed and shuts down the VRs. Removal of power is required to reset the microcontroller to a known start point.

## 6.4 Host / In-Band Management Interface (MMIO)

Manageability, through the SMC, is achievable via the PCI MMIO interface. This allows host programs to obtain MIC telemetry and other information from the SMC managed features of the Intel® Xeon Phi™ coprocessor itself, as well as control SMC enabled functions. The SMC supports a host based MMIO based interface.

The following SMC information and sensors are accessible over the MMIO-based interface:

- Hardware strapping pins
- SMC firmware revision number
- UUID
- PCI compliant MMIO (required for PCI compliancy)
- Fan tachometer
- Fan PWM adder for boosting the fan speed for additional cooling
- SMC System Event Log (SEL)
- All registers mentioned in the Ganglia support section
- Voltage rail discrete monitoring
- All discrete temperature sensors
- $T_{critical}$
- $T_{control}$

- $T_{current}$
- $T_{control\ offset\ adder}$
- Thermal throttle duration due to card power limit (in ms), free running counter that overflows at 60 seconds
- $T_{inlet}$ (derived numbers)
- $T_{outlet}$ (derived numbers)
- PERF_Status_Thermal
- 32-bit POST register
- SMC SEL Entry select and data registers (read only)
- SMC SDR Entry select and data registers (read only - required to interpret the SEL)

Each SMC sensor that is exposed over MMIO indicates one of four states in a consistent manner, returned in the same register value as the sensor reading itself, regardless of sensor type. These states do not apply to non-sensor information:

- Normal
- Upper critical
- Lower critical
- Inaccessible (sensor not available)

This minimizes the complexity of host-driven software and SMC firmware implementations.

The sensors available from the SMC vary within the Intel® Xeon Phi™ coprocessor family of products. However, the IPMI SDR sensor names will not change from release to release.

$T_{inlet}$ and $T_{outlet}$ are derived numbers based on the Inlet and Outlet temperature sensors.

The sensors located on the Intel® Xeon Phi™ coprocessor relate information about the CPU temperature as well as the temperature from three locations on the Intel® Xeon Phi™ coprocessor. Currently, one sensor is located between memory chips near the PCI Express* slot while the other two are located on the east and west sides of the card. These are sometimes referred to as the "inlet" and "outlet" air temperature sensors but they do not actually indicate airflow temperature but rather the temperature of the board. The sensors are attached to the 12 inputs from the PCI Express* slot, the 2x3 connector, and the 2x4 connector. Input power can be estimated by summing the currents over these three connections. For an actively cooled card, the SMC can also provide the fan percentage pulse width modulation (PWM) being used. Fan speed is a simple PID control with setpoints set rather high to keep the sound level low when max cooling is not needed.

There are two programmable power limits. One defaults to 105% of the design power consumption. If this is exceeded, the SMC notifies the Intel® Xeon Phi™ coprocessor silicon through an interrupt to reduce power. The second power limit forces throttling of the Intel® Xeon Phi™ coprocessor silicon when the power consumption is above 125% of design.

## 6.5 System and Power Management

The Intel® Xeon Phi™ coprocessor PCI card supports both on-card power management and an option for system-based management. With on-card power management, the SMC controls system power using preprogrammed power limits. The µOS can change the preprogrammed power limits if necessary. With system-based management, the SMC receives power control inputs via in-band communication from a host application.

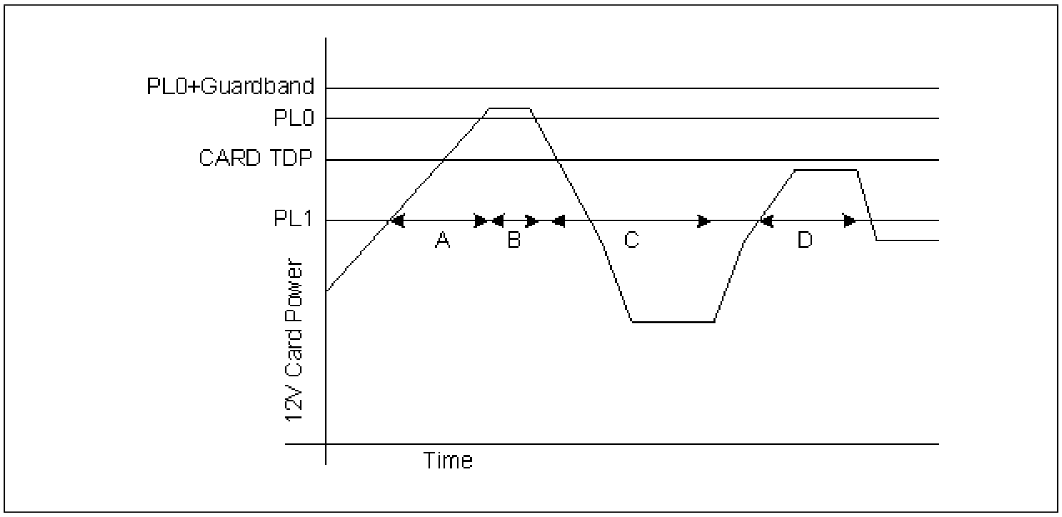
If the server administrator knows that there is sufficient power in the server infrastructure to support greater than TDP for short time frames, then PL0 can be set to a value greater than the card TDP. If the server administrator knows the that the power infrastructure cannot support power greater than PL1 indefinitely, then PL1 can be set to a value less than TDP.

There is no relationship between PL0, PL1, and card TDP other than the fact that PL1 must be less than or equal to PL0. PL0 sets the peak power limit for the card level. This is the moving average power (Watts) that can be consumed in Time Window 0 (set in increments of milliseconds). PL1 sets the peak power limit for the card level. This is the moving average power (Watts) that can be consumed in Time Window 1 (set in increments of milliseconds). PL0 and PL1 are set in increments of 1W.

Typically, the time window for PL0 is set to less than the time window for PL1. This means that the SMC carries a running average for PL0 and PL1. The PL0 running average is calculated over time window 0 and the PL1 average is calculated over time window 1. The SMC collects power sensor data at an update rate of time window 0 (or faster) so PL0 is the last data collected from power sensors. The PL1 average is a moving average over time window 1. A power excursion over PL1 for time window 1 may not cause the SMC to assert a power alert, assuming the PL1 time window is sufficiently long.

Figure 6-2 and the following discussion is a representation of how to use the power limits.

**Figure 6-2 Example of Using Power Limits**



In this scenario PL0 is greater than the card TDP because the server operator knows that there is enough power in the server infrastructure to support greater than TDP for short time frames, even after reviewing the PL0 guardband. PL1 is less than TDP because the server operator knows that the power infrastructure cannot support greater than PL1 indefinitely.

The server operator set the time window for PL0 at 50ms and the time window for PL1 at 300ms. This means that the SMC carries a running average of PL0 and PL1.

The PL0 running average is calculated over a 50ms window and the PL1 average is calculated over a 300ms window. The SMC collects power sensor data at an update rate of 50ms, so PL0 is the last data collected from power sensors. The PL1 average is a moving average over the last 300ms. A 50ms power excursion over PL1 might not cause the SMC to assert power alert if the PL1 time window is longer than 300ms.

Figure 6-2 shows a possible 12V Card Power profile. In time window A the server started to run a high-power application that consumes more power than PL1. Power_Alert does not assert because the 12V power's 300ms moving average is less than PL1, even though the 50ms moving average is greater than PL1. During time window B, the 50ms moving average is greater than PL0.

When there is a PL0 violation, the SMC will immediately assert thermal throttling (also known as PROCHOT event) for 1µs. The µOS will manage clearing the Intel® Xeon Phi™ coprocessor silicon and reducing the card operating frequency to the minimum value. The µOS has a routine that checks the source of the minimum frequency state and then takes actions to minimize future assertions of PL0.

During time C the µOS services an interrupt and sets the maximum performance state so that future PL0 violations are unlikely to occur. The µOS then restores the card to its rated operating frequency and the power starts to ramp.

During time D, the application does not ramp to PL0 because the µOS sets the maximum performance state to a lower state than in time A. However, the power draw is still greater than PL1. The SMC asserts Power_Alert near the end of time window D because the 12V power 300ms moving average is greater than PL1. The µOS services the Power_Alert interrupt and takes appropriate action to lower the Intel® Xeon Phi™ coprocessor power consumption. Eventually the card power is under PL1.

## 6.6 Out of Band / PCI Express* SMBus / IPMB Management Capabilities

The Intel® Xeon Phi™ coprocessor PCI Express* card exists as part of a system-level ecosystem. In order for this system to manage its cooling and power demands, the Intel® Xeon Phi™ coprocessor telemetry must be exposed to ensure that the system is adequately cooled and that proper power is maintained. Manageability code running elsewhere in the chassis, through the SMC, can retrieve SMC sensor logs, sensor data, and vital information required for robust server management. Note that logging, in this context, is completely separate from and has nothing to do with the MCA error log.

The SMC public interface (SMBus) is a compliant IPMB interface. It supports a minimal IPMB command set in order to interact with manageability devices such as BMCs and the ME (Manageability Engine).

The IPMB implementation on the SMC can receive additional incoming requests while responses are being processed. This enables the interleaving of requests and responses from multiple sources using the SMC's IPMB, thus minimizing latency.

Upon initial power-on or restart, the SMC selects an IPMB slave address from the range 0x30 - 0x4e in increments of 2 (e.g., 0x30, 0x32, 0x34, etc.). The IPMB slave address self-select starting address is nonvolatile, starting at the last selected slave address. This ensures that the card doesn't move nondeterministically in a static system. To determine the address of the Intel® Xeon Phi™ coprocessor card scan the range of addresses issuing the Get Device ID command for each address. A valid response indicates the address used is a valid address.

For the Intel® Xeon Phi™ coprocessor cards, the IPMB slave address will be found at 0x30 if only a single card is installed. If the motherboard has an exclusive connection to the SMBus on each PCI Express* connection, then the Intel® Xeon Phi™ coprocessor will assign itself a default address (0x30). If the SMBus connections are shared, each Intel® Xeon Phi™ coprocessor in a chassis will negotiate with each other and select addresses in the range from 0x30 to 0x4e. If a mux is incorporated into the design to isolate devices on a shared link the address negotiation process should result in each card having address 0x30. However, if the mux in use allows for the channels to be merged, i.e., creating a shared bus scenario, the address negotiation may result in each card having a unique address behind the mux.

Power management and power control are performed through the host driver interface (in-band). An SDK is provided as part of the Intel® Xeon Phi™ coprocessor software stack and is named - MICSYSSW_OEM.tar.

The SMC's PCI Express*/SMBus interface operates as an industry standard IPMI IPMB with a reduced IPMI command implementation. The SMC supports a system event log (SEL) via the IPMI interface.

The SMC supports a read only IPMI SDR. It is hardcoded and not end-user updateable. The SDR must be read in "chunks", suggested size is 16 bytes. The request to read the entire buffer will result in an error due to the buffer size is insufficient to return the complete SDR.

### 6.6.1 IPMB Protocol

The IPMB protocol is a symmetrical byte-level transport for transferring IPMI messages between intelligent I2C devices. It is a worldwide standard widely used in the server management industry. In this case, the client requests are sent to the SMC with a master I2C write.

Although both devices are a master on the bus at different times, the SMC only responds to requests. With the exception of the address selection algorithm, it does not initiate master transactions on the bus at any other time during normal operation.

For byte level details, refer to the *Intelligent Platform Management Bus Communications Protocol Specification, v1.0*.

### 6.6.2 Polled Master-Only Protocol

The polled master-only protocol may be used in the event IPMB is not feasible. The client sends requests to the SMC using one or more SMC SMBus Write Block commands then, at a later time, reads the response using one or more SMBus Read Block commands.

### 6.6.2.1 Polled Master-Only Protocol Clarifications

The polled master-only protocol is loosely based on the IPMI defined SSIF protocol; however, there have been a few changes made and ambiguities clarified in order to make the protocol more reliable:

- The I2C address for the polled master-only protocol and the IPMB protocol are the same and work together transparently.
- PEC bytes are required for all write commands and are returned with all valid read responses.
- The maximum SMBus data length is restricted to 32 bytes.
- The SMC ignores write commands that occur while it is internally processing a previous command.
- The SMC does not return valid data while busy internally processing a command.
- A sequence number has been added to help identify the condition where a new write command (using the same NetFn and command as the last command sent) was corrupted during transit. Without this precaution, two sequential requests of the same type (i.e., Get Sensor Reading) could result in one sensor’s reading being mistaken for the other’s.
- SMBAlert is not supported.

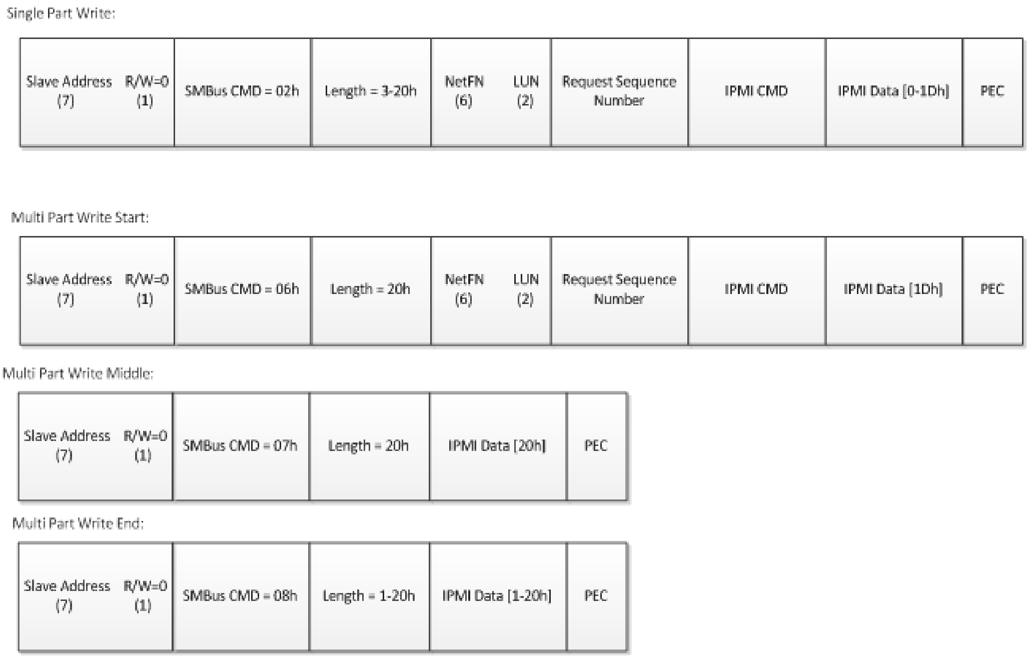
### 6.6.2.2 SMBus Write and Read Block Command Numbers

### 6.6.2.3 Write Description

**Table 6-1. SMBus Write Commands**

| Command | Name | Command Type |
|---|---|---|
| 02h | Single Part Write | Write Block |
| 06h | Multi-Part Write Start | Write Block |
| 07h | Multi-Part Write Middle | Write Block |
| 08h | Multi-Part Write End | Write Block |
| 03h | Single Read Start | Read Block |
| 03h | Multi-Part Read Start | Read Block |
| 09h | Multi-Part Read Middle | Read Block |
| 09h | Multi-Part Read End | Read Block |

**Figure 6-3. Write Block Command Diagram**



Single Part Write:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Slave Address<br>(7) | R/W=0<br>(1) | SMBus CMD = 02h | Length = 3-20h | NetFN<br>(6) | LUN<br>(2) | Request Sequence Number | IPMI CMD | IPMI Data [0-1Dh] | PEC |

Multi Part Write Start:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Slave Address<br>(7) | R/W=0<br>(1) | SMBus CMD = 06h | Length = 20h | NetFN<br>(6) | LUN<br>(2) | Request Sequence Number | IPMI CMD | IPMI Data [1Dh] | PEC |

Multi Part Write Middle:

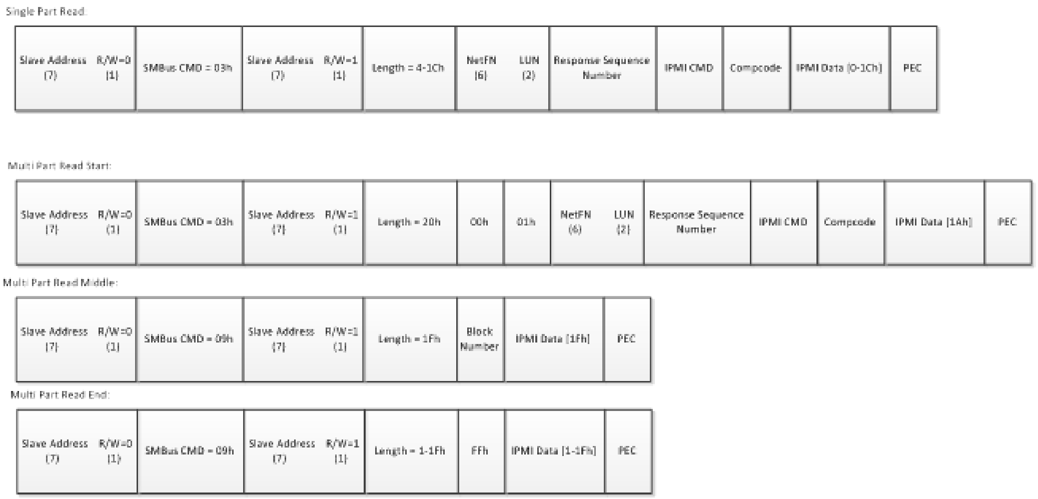
| | | | | | |
|---|---|---|---|---|---|
| Slave Address<br>(7) | R/W=0<br>(1) | SMBus CMD = 07h | Length = 20h | IPMI Data [20h] | PEC |

Multi Part Write End:

| | | | | | |
|---|---|---|---|---|---|
| Slave Address<br>(7) | R/W=0<br>(1) | SMBus CMD = 08h | Length = 1-20h | IPMI Data [1-20h] | PEC |

### 6.6.2.4 Read Description

Figure 6-4. Read Block Command Diagram



Single Part Read:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Slave Address (7) R/W=0 (1) | SMBus CMD = 03h | Slave Address (7) R/W=1 (1) | Length = 4-1Ch | NetFN (6) LUN (2) | Response Sequence Number | IPMI CMD | Compcode | IPMI Data [0-1Ch] | PEC |

Multi Part Read Start:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Slave Address (7) R/W=0 (1) | SMBus CMD = 03h | Slave Address (7) R/W=1 (1) | Length = 20h | 00h | 01h | NetFN (6) LUN (2) | Response Sequence Number | IPMI CMD | Compcode | IPMI Data [1Ah] | PEC |

Multi Part Read Middle:

| | | | | | | |
|---|---|---|---|---|---|---|
| Slave Address (7) R/W=0 (1) | SMBus CMD = 09h | Slave Address (7) R/W=1 (1) | Length = 1Fh | Block Number | IPMI Data [1Fh] | PEC |

Multi Part Read End:

| | | | | | | |
|---|---|---|---|---|---|---|
| Slave Address (7) R/W=0 (1) | SMBus CMD = 09h | Slave Address (7) R/W=1 (1) | Length = 1-1Fh | FFh | IPMI Data [1-1Fh] | PEC |

## 6.6.3 Supported IPMI Commands

The SMC supports a subset of the standard IPMI sensor, SEL, and SDR commands along with several Intel OEM commands for accomplishing things like forcing throttle mode. The supported IPMI commands are documented in the following sections. Standard IPMI details are not documented in this document. For those please refer to the IPMI v2.0 specification. For example the Get SDR command requires additional bytes to complete the command packet and these bytes are defined in the IPMI v2.0 specification.

### 6.6.3.1 Miscellaneous Commands

Table 6-2. Miscellaneous Command Details

| NetFn | Command | Name |
|---|---|---|
| **App (0x06)** | 0x01 | Get Device ID |
| **App (0x06)** | 0x08 | Get Device GUID (UUID) |

#### 6.6.3.2 FRU Related Commands

**Table 6-3. FRU Related Command Details**

| NetFn | Command | Name |
|---|---|---|
| **Storage (0x0a)** | 0x10 | Get FRU Inventory Area Info |
| **Storage (0x0a)** | 0x11 | Read FRU Data |

#### 6.6.3.3 SDR Related Commands

**Table 6-4. SDR Related Command Details**

| NetFn | Command | Name |
|---|---|---|
| **Storage (0x0a)** | 0x20 | Get SDR Repository Info |
| **Storage (0x0a)** | 0x21 | Get SDR Repository Allocation Info |
| **Storage (0x0a)** | 0x23 | Get SDR |

***Note:*** It is recommended to read the SDR in "chunks" rather than request to read the entire record. See Section 6.6 for more information.

#### 6.6.3.4 SEL Related Commands

**Table 6-5. SEL Related Command Details**

| NetFn | Command | Name |
|---|---|---|
| **Storage (0x0a)** | 0x40 | Get SEL Info |
| **Storage (0x0a)** | 0x41 | Get SEL Allocation Info |
| **Storage (0x0a)** | 0x43 | Get SEL Entry |
| **Storage (0x0a)** | 0x47 | Clear SEL |
| **Storage (0x0a)** | 0x48 | Get SEL Time |
| **Storage (0x0a)** | 0x49 | Set SEL Time |

### 6.6.3.5 Sensor Related Commands

**Table 6-6. Sensor Related Command Details**

| NetFn | Command | Name |
|---|---|---|
| **Sensor (0x04)** | 0x2b | Get Sensor Event Status |
| **Sensor (0x04)** | 0x2d | Get Sensor Reading |

### 6.6.3.6 General Commands

**Table 6-7. General Command Details**

| NetFn | Command | Name |
|---|---|---|
| **Intel (0x2e)** | 0x42 | CPU Package Config Read |
| **Intel (0x2e)** | 0x43 | CPU Package Config Write |
| **Intel General App (0x30)** | 0x15 | Set SM Signal |

#### 6.6.3.6.1 CPU Package Configuration Read

The CPU Package Config Read command reads power control data. For the parameter byte formats, refer to the *Intel Xeon*$^{TM}$ *Processor Family External Design Specification (EDS) Volume 1*.

**Table 6-8. CPU Package Config Read Request Format**

| Byte # | Value | Description |
|---|---|---|
| Command | 0x42 | • CPU Package Config Read |
| NetFn | 0x2e | • NETFN_INTEL |
| 0-2 | | • Manufacturer ID (LSB format): 0x57, 0x01, 0x00 |
| 3 | 0x00 | • CPU Number |
| 4 | 0x?? | • PCS Index<br>• 3 - Accumulated Energy Status<br>• 11 - Socket Power Throttle Duration<br>• 26 - Package Power Limit 1 (PL1)<br>• 27 - Package Power Limit 2 (PL0)<br>• 28 - Package Power SKU A<br>• 29 - Package Power SKU B<br>• 30 - Package Power SKU Unit<br>• All other values reserved |
| 5 | 0x00 | • Parameter LSB |
| 6 | 0x00 | • Parameter MSB |
| 7 | 0x?? | • Number of Bytes to Read |

### Table 6-9. CPU Package Config Read Response Format

| Byte # | Value | Description |
|---|---|---|
| 0 | 0x?? | - Compcode<br>- 0x00 - Normal<br>- 0xcc - Invalid field<br>- 0xa1 - Wrong CPU Number<br>- 0xa7 - Wrong Read Length<br>- 0xab - Wrong Command Code<br>- 0xff - Unspecified Error |
| 1-3 | | - Manufacturer ID (LSB format): 0x57, 0x01, 0x00 |
| 4[-7] | 0x?? | - Data bytes read, up to 4 bytes |

### 6.6.3.6.2 CPU Package Configuration Write

The CPU Package Config Write command allows the setting of power control data. For the parameter byte formats, refer to the *Intel Xeon Processor Family External Design Specification (EDS) Volume 1*.

### Table 6-10. CPU Package Config Write Request Format

| Byte # | Value | Description |
|---|---|---|
| Command | 0x43 | - CPU Package Config Write |
| NetFn | 0x2e | - NETFN_INTEL |
| 0-2 | | - Manufacturer ID (LSB format): 0x57, 0x01, 0x00 |
| 3 | 0x00 | - CPU Number |
| 4 | 0x?? | - PCS Index<br>- 26 - Package Power Limit 1 (PL1)<br>- 27 - Package Power Limit 2 (PL0)<br>- All other values reserved |
| 5 | 0x00 | - Parameter LSB<br>- |
| 6 | 0x00 | - Parameter MSB |
| 7 | 0x?? | - Number of Bytes to Write |
| 8[-11] | 0x?? | - Data bytes to write |

### Table 6-11. CPU Package Config Write Response Format

| Byte # | Value | Description |
|---|---|---|
| 0 | 0x?? | - Compcode<br>- 0x00 - Normal<br>- 0xc7 - Request Length Invalid<br>- 0xcc - Invalid Field<br>- 0xa1 - Wrong CPU Number<br>- 0xa6 - Wrong Write Length<br>- 0xab - Wrong Command Code<br>- 0xff - Unspecified Error |
| 1-3 | | - Manufacturer ID (LSB format): 0x57, 0x01, 0x00 |

### 6.6.3.6.3 Set SM Signal

The Set SM Signal command gives you control of firmware signals. The primary use of this command is to set the status LED into identify mode. In identify mode the status LED flashes on for a short period twice every 2 seconds. This allows an administrator to locate the card in a system that has multiple cards.

**Table 6-12. Set SM Signal Request Format**

| Byte # | Value | Description |
|---|---|---|
| Command | 0x15 | • Set SM Signal |
| NetFn | 0x30 | • NETFN_INTEL_GENERAL_APP |
| 0 | 0x?? | • Signal<br>• 1 - Identify<br>• All other values reserved |
| 1 | 0x00 | • Instance |
| 2 | 0x?? | • Action<br>• If Signal is 1<br>• 1 - Assert: Start the identify blink code<br>• 2 - Revert: Return to normal operation<br>• All other values reserved |
| [3] | 0x00 | • Value (optional) |

**Table 6-13. Set SM Signal Response Format**

| Byte # | Value | Description |
|---|---|---|
| 0 | 0x?? | • Compcode<br>• 0x00 - Normal<br>• 0xc7 - Request Length Invalid<br>• 0xc9 - Parameter Out of Range<br>• 0xcc - Invalid Field |

## 6.6.3.7 OEM Commands

**Table 6-14. OEM Command Details**

| NetFn | Command | Name |
|---|---|---|
| **OEM (0x3e)** | 0x00 | • OEM Set Fan PWM Adder |
| **OEM (0x3e)** | 0x04 | • OEM Get POST Register |
| **OEM (0x3e)** | 0x05 | • OEM Assert Forced Throttle |
| **OEM (0x3e)** | 0x06 | • OEM Enable External Throttle |

#### 6.6.3.7.1 OEM Set Fan PWM Adder

The Set Fan PWM Adder command allows a PWM percentage to be added to the final fan cooling algorithm for additional cooling based on chassis requirements.

**Table 6-15. Set Fan PWM Adder Command Request Format**

| Byte # | Value | Description |
|---|---|---|
| Command | 0x00 | • OEM Set Fan PWM Adder |
| NetFn | 0x3e | • NETFN_OEM |
| 0 | 0x?? | • PWM percent to add to standard cooling: 0x00 - 0x64<br>• All other values are reserved. |

**Table 6-16. Set Fan PWM Adder Command Response Format**

| Byte # | Value | Description |
|---|---|---|
| 0 | 0x?? | • Compcode<br>• 0x00 - Normal<br>• 0xc9 - Parameter out of range |

#### 6.6.3.7.2 OEM Get POST Register

The Get POST Register command allows the BMC to obtain the last POST code written to the SMC by the coprocessor. The SMC does not modify this value in any way.

**Table 6-17. Get POST Register Request Format**

| Byte # | Value | Description |
|---|---|---|
| Command | 0x04 | • OEM Get POST Register |
| NetFn | 0x3e | • NETFN_OEM |

**Table 6-18. Get POST Register Response Format**

| Byte # | Value | Description |
|---|---|---|
| 0 | 0x?? | • Compcode<br>• 0x00 - Normal |
| 1-4 | 0x?? | • 32 bit POST code in little endian format |

#### 6.6.3.7.3 OEM Assert Forced Throttle

The Assert Forced Throttle command allows the BMC to cause the SMC to assert the PROCHOT pin to the coprocessor.

### Table 6-19. Assert Forced Throttle Request Format

| Byte # | Value | Description |
|---|---|---|
| Command | 0x05 | • OEM Assert Forced Throttle |
| NetFn | 0x3e | • NETFN_OEM |
| 0 | 0x?? | • 0 = Deassert forced throttle<br>• 1 - Assert forced throttle<br>• All other values are reserved |

### Table 6-20. Assert Forced Throttle Response Format

| Byte # | Value | Description |
|---|---|---|
| 0 | 0x?? | • Compcode<br>• 0x00 - Normal |

### 6.6.3.7.4 OEM Enable External Throttle

The Enable External Throttle command causes the SMC to enable a pin on the baseboard connector allowing the baseboard to directly assert the PROCHOT signal.

### Table 6-21. Enable External Throttle Request Format

| Byte | Value | Description |
|---|---|---|
| Command | 0x06 | • OEM Enable External Throttle |
| NetFn | 0x3e | • NETFN_OEM |
| 0 | 0x?? | • 0 = Disable external throttle signal<br>• 1 = Enable external throttle signal<br>• All other values are reserved |

### Table 6-22. Enable External Throttle Response Format

| Byte | Value | Description |
|---|---|---|
| 0 | 0x?? | • Compcode<br>• 0x00 - Normal<br>• 0xc0 - Busy |

## 6.6.3.8 Other IPMI Related Information

The SMC supports a read only IPMI FRU.The SMC System Event Log is a circular log supporting a minimum of 64 log entries. It is resilient to corruption, retaining information across unexpected power loss.

The sensor names in the IPMI Sensor Data Record are static and do not change from release to release. The IPMI sensor numbers may change and hence should be discovered during the normal sensor discovery process. Sensors may be added in the future, but the previously defined sensor names will not change.

Reading the SDR returns the sensors available on the card. There will be a sensor number and sensor name associated with each sensor. Once the sensor name is determined it can be used in the management firmware for reading a particular sensor by discovering the sensor number associated with the sensor name. It is strongly recommended to use the sensor name if it is "hard coded" into the management firmware. Sensor numbers should not be hard coded as the sensor numbers are subject to change. Incorporating a sensor number in the management firmware as a hard coded value could result in incorrect values with subsequent releases of the SMC firmware. Using the sensor name and discovering the sensor number associated with a sensor name will ensure the correct sensor is read and returns valid data with each future release of the SMC firmware. The following table is a list of the current sensor names.

### Table 6-23. List of Sensor Names on the Intel® Xeon Phi™ coprocessor

| Sensor Name | Sensor Function |
|---|---|
| Power | |
| power_pcie | Power measured at the PCI Express* edge finger input |
| power_2x3 | Power measured at the 2x3 auxiliary connector input |
| power_2x4 | Power measured at the 2x4 auxiliary connector input |
| power_pv | Power output reported by VR supplying power to coprocessor |
| power_vddq | Power output reported by VR supplying power to coprocessor |
| power_vddg | Power output reported by VR supplying power to memory and other circuitry |
| avg_power0 | Average power consumption over Limit Time Window 0 |
| avg_power1 | Average power consumption over Limit Time Window 0 |
| Instpwr | Instantaneous power consumption reading |
| Instpwrmax | Maximum instantaneous power consumption observed |
| Voltage | |
| pv_volt | Voltage reported from VR supplying power to coprocessor |
| vddq_volt | Voltage reported from VR supplying power to coprocessor |
| vddg_volt | Voltage reported from VR supplying power to memory and other circuitry |
| Temperature | |
| east_temp | Temperature sensor on the eastern-most side of the board |
| gddr_temp | Temperature sensor closest to the GDDR memory devices |
| west_temp | Temperature sensor on the western-most side of the board |
| pv_vrtemp | Temperature reported from VR supplying power to coprocessor |
| vddq_temp | Temperature reported from VR supplying power to coprocessor |
| vddg_temp | Temperature reported from VR supplying power to memory and other circuitry |
| proc_temp | Temperature reported by the coprocessor (junction temperature) |
| exhst_temp | Highest of discrete temperature sensors on the board |
| inlet_temp | Lowest of discrete temperature sensors on the board |
| Fan | |
| fan_pwm | Fan PWM driven by SMC software (only on active SKU) |
| fan_tach | Fan tach read by SMC (only on active SKU.) |

Table 6-23. List of Sensor Names on the Intel® Xeon Phi™ coprocessor

| Sensor Name | Sensor Function |
|---|---|
| Other | |
| status | Critical signal states (Section 6.6.3.9.) |
| tcritical | Value reported by coprocessor for thermal monitoring |
| Tcontrol | Value reported by coprocessor for stem fan control |

The SMC implements the ability to read all SMC-based sensors via the Get Sensor Reading command. The sensor number to be sent with the command must be discovered and not hard coded in the firmware as this can lead to incorrect readings or returning invalid sensor errors.

### 6.6.3.9 SMC IPMI Discrete Sensors

The SMC's IPMI discrete sensors are defined here because the meaning of each discrete bit cannot be easily derived from the SDR definition.

#### 6.6.3.9.1 Sensor Status

The status sensor reports the state of several critical signals on the card such as thermtrip, VR phase, fault, VR hot, UV/OV Alert, and PCI Express* Reset. The sensor is not mirrored as a register on the in-band register interface.

Table 6-24. Status Sensor Report Format

| Bits | Name | Description |
|---|---|---|
| 31:7 | Reserved | • Reserved |
| 6 | P2E_RST | • PCI Express* reset asserted.<br>• Fans boosted. |
| 5 | P12V_UVOV | • P12V under-voltage/over-voltage signal asserted.<br>• Fans boosted and VR output disabled. |
| 4 | VR2_HOT | • VR2 Hot signal asserted. Fans boosted and PROCHOT asserted. |
| 3 | VR1_HOT | • VR1 Hot signal asserted. Fans boosted and PROCHOT asserted. |
| 2 | VR2_PHSFLT | • VR2 Phase Fault asserted.<br>• Fans boosted and VR output disabled.<br>• This state is latched until power-off. |
| 1 | VR1_PHSFLT | • VR1 Phase Fault asserted.<br>• Fans boosted and VR output disabled.<br>• This state is latched until power-off. |
| 0 | THERMTRIP | • Coprocessor thermtrip asserted.<br>• Fans boosted and VR output disabled.<br>• This state is latched until power-off. |

## 6.7 SMC LED_ERROR and Fan PWM

The SMC firmware drives the LED_ERROR pin as follows

:

### Table 6-1. LED Indicators

| Blink Frequency | Condition |
|---|---|
| 0.5HZ Blink | • In boot loader mode |
| 2HZ Blink | • Firmware update in progress |
| 8HZ Blink | • Operational code executing |
| Identify Blink | • 2 short blinks every 2 seconds.<br>• Initiated by SetSMSignal command. |

The SMC drives the fan PWM to the static rate provided in the IPMI FRU while in boot loader mode.