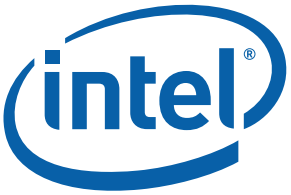


Intel® Xeon Phi™ Coprocessor x100 Product Family

Datasheet

April 2015

Document ID Number: 328209 004EN

By using this document, in addition to any agreements you have with Intel, you accept the terms set forth below.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. 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.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

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.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

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 go to: http://www.intel.com/design/literature.htm

This document contains information on products in the design phase of development.

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: Learn About Intel® Processor Numbers

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
32829	004	• Added a disclaimer regarding the default SMC address of 0x30.	April 2015
328209	003	• Updated product SKU Table 2-1. Added 31S1P, 7120A and 7120D. • Added SKUs 31S1P and 7120A to relevant figures, paragraphs and tables throughout document. • Added thermal throttling due to hot to VRs, Section 3.4.1.1. • Minor changes and clarifications in power management and manageability chapter.	April 2014
328209	002	• Updated product SKU Table 2-1. • Updated mechanical specification Table 3-1 and thermal specification Table 3-2. • Other changes in thermal and mechanical specification chapter. • Added significant information on 5120D in pin list chapter. • Updated power state numbers in Table 5-1 and added Turbo in power management chapter • Changes and clarifications in manageability chapter.	June 2013
328209	001	• First Intel® Xeon Phi™ Coprocessor Datasheet release.	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	Document ID
Intel® Xeon Phi™ Coprocessor Specification Update	328205-006EN [1]
Intel® Xeon Phi™ Coprocessor Safety Compliance Guide	328206-001EN [1]
Intel® Xeon Phi™ Coprocessor System Software Developer's Guide	328207-001EN [1]
Intel® Xeon Phi™ Coprocessor Thermal Mechanical Models	328208-001EN [1]
Intel® Xeon Phi™ Coprocessor Dense Form Factor Models	N/A [1]
Intel® Xeon Phi™ Coprocessor Instruction Set Reference Manual	N/A [2]
Intel® Xeon™ Processor Family External Design Specification (EDS) Volume 1	N/A [1]
PCI Express* Card Electromechanical Specification, Revision 2.0	N/A [3]
PCI Express* 225W/300W High Power Card Electromechanical Specification, Revision 1.0	N/A [3]
Intelligent Platform Management Bus Communications Protocol Specification, v1.0	N/A
Intelligent Platform Management Interface Specification, v2.0	N/A

Note: 1. http://www.intel.com/content/www/us/en/processors/xeon/xeon-technical-resources.html
Note: *2.* http://software.intel.com/en-us/forums/intel-many-integrated-core/
Note: 3. http://www.pcisig.com/

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
FET	Field Effect Transistor
FRU	Field Replaceable Unit
GDDR	Graphics Double Data Rate
IBP	Intel Business Portal
IPMB	Intelligent Platform Management Bus
IPMI	Intelligent Platform Management Interface
ME	Manageability Engine

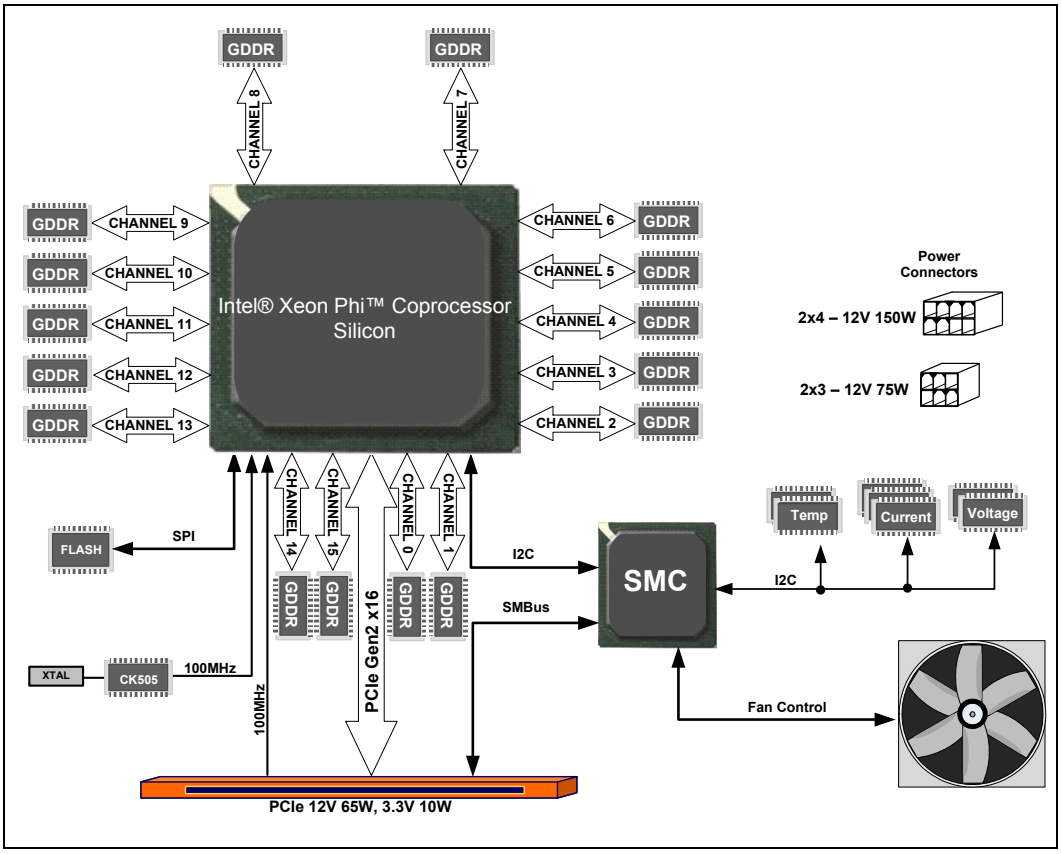
Table 1-2. General Terminology

Terminology	Definition
MPSS	Many Integrated Core Platform Software Stack
PCIe	PCI Express*
PID	Proportional-Integral-Derivative control algorithm
PWM	Pulse Width Modulation
RAS	Reliability Accessibility Serviceability
SCIF	Symmetric Communications Interface
SDK	Software Development Kit
SDR	Sensor Data Record
SEL	System Event Log
SKU	Stock Keeping Unit
SMBus	System Management Bus
SMC	System Management Controller
TDP	Thermal Design Power
VID	Coprocessor Voltage Identification
VR	Voltage Regulator
XTAL	Crystal Oscillator

2 Intel® Xeon Phi™ Coprocessor Architecture

2.1 Intel® Xeon Phi™ Coprocessor Product Overview

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



Notes:

1. On-board fan is available on Intel® Xeon Phi™ Coprocessor 3120A and 7120A SKUs only.

The Intel® Xeon Phi™ coprocessor consists of the following primary subsystems:

- Many Integrated Core (MIC) coprocessor silicon and GDDR5 memory.
- System Management Controller (SMC), thermal sensors (inlet air, outlet air, coprocessor on-die thermal, and single GDDR5 sensor) and fan (only on 3120A and 7120A products; 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) auxiliary 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 including an on-board 100MHz +/ - 50ppm reference clock, and requires only the PCI Express* 100MHz reference clock input from the motherboard.

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* 2.0 x16 interface with optional SMBus management interface.
- Node Power and Thermal Management, including power capping support.
- +12V power monitoring and on-board fan Proportional-Integral-Derivative (PID) controller (3120A and 7120A SKUs).
- 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 high-power add-in product, with an integrated thermal and mechanical solution (see SKU matrix Table 2-1 for exceptions). It supports a maximum of 16 GDDR memory channels, distributed on both sides of the PCB inside the coprocessor package. Each memory channel supports two 16-bit wide GDDR device (for a maximum of 32 devices), 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 not containing the Intel® Xeon Phi™ coprocessor silicon). The VRs are split right and left to help reduce direct current resistance and current density.

The Intel® Xeon Phi™ coprocessor supports 2 power groups for a total of 4 primary low-voltage rails: a group consisting of VDDG, VDDQ, and VSFR power rails, and another group consisting only of 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 in the coprocessor.

Figure 2-2. Intel® Xeon Phi™ Coprocessor Board Top side (for reference only)

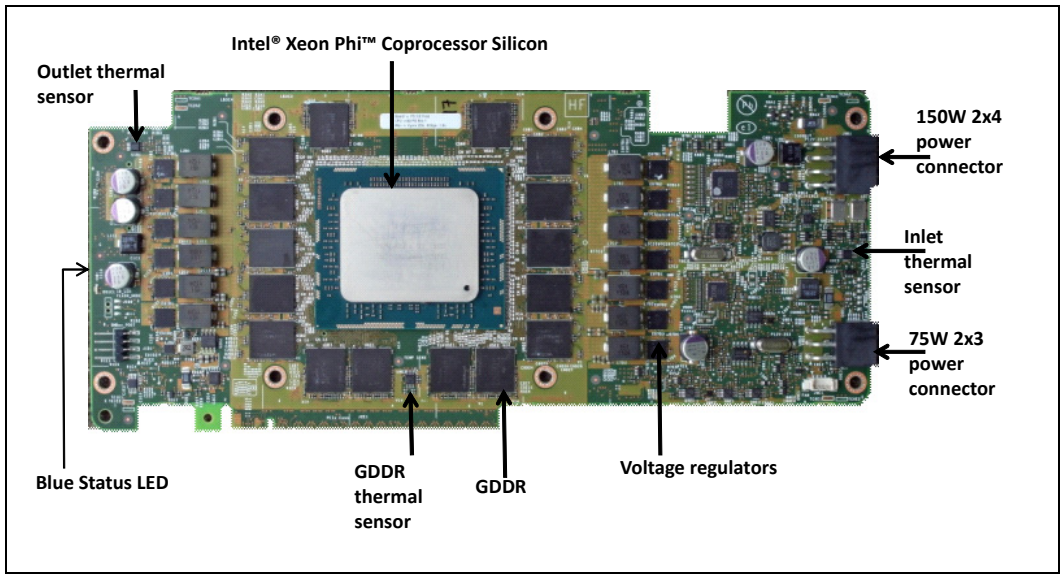
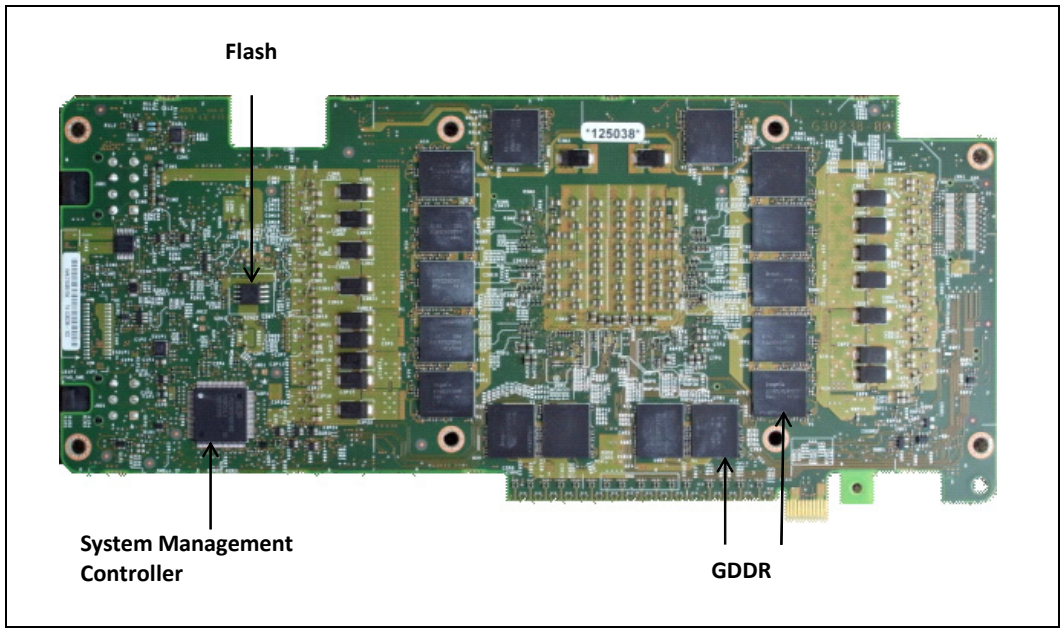


Figure 2-3. Intel® Xeon Phi™ Coprocessor Board, Back side (reference only)



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

2.1.2 System Management Controller (SMC)

The SMC has three I2C interfaces. This allows the SMC to have direct connection to the coprocessor I2C interface, an on-board I2C sensor bus and a third interface through the SMBus pins of the PCI Express* connector to the system management solution. The I2C interface between the SMC and coprocessor is used for polling coprocessor thermal status information. The sensor bus allows board thermal, input power, and current

sense monitoring for system fan and power control. This information is forwarded to the coprocessor for power state control. The SMBus interface can be 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

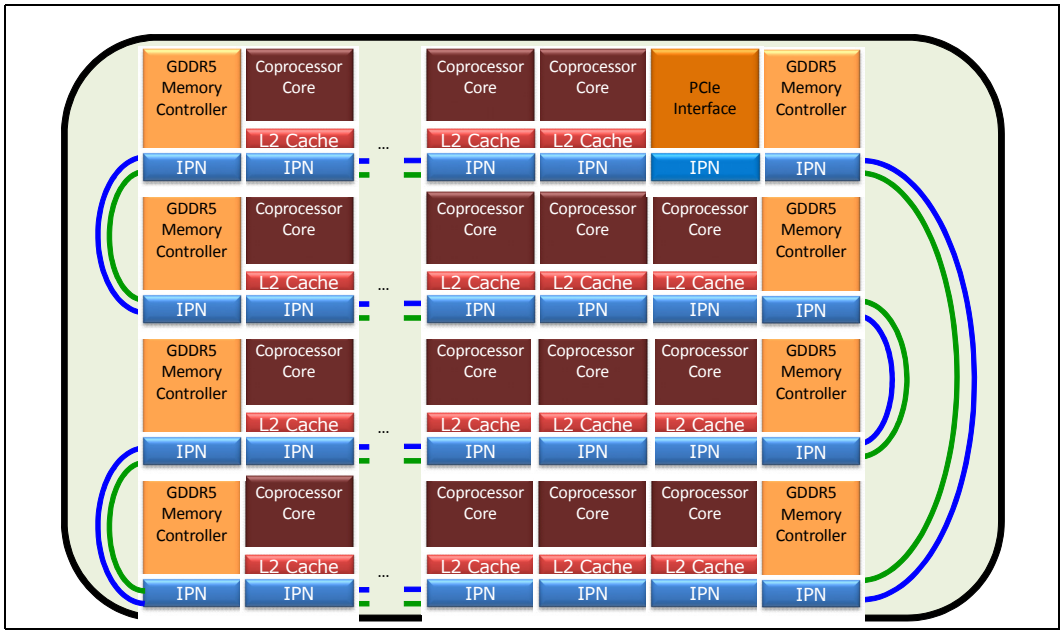


Figure 2-4 is a conceptual drawing of the general structure of the Intel® Xeon Phi™ coprocessor architecture, and does not imply actual distances, latencies, etc. The cores, PCIe Interface logic, and GDDR5 memory controllers are connected via an Interprocessor Network (IPN) ring, which can be thought 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 up to 61 cores (making a 30.5 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™ Coprocessor 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]
3120A	300	Active
3120P / 7120P / SE10P	300	Passive[2,4]
7120X / SE10X	300	None[2,3,4]
31S1P	270	Passive
7120A	270	Active
7120D	270	None[6]
5120D	245	None[5]
5110P[6]	225	Passive

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. SE10P/SE10X are limited edition one-time only SKUs.
3. Same performance and card configuration as the 7120P/SE10P but without Intel heatsink or chassis retention mechanism; allows for custom thermal and mechanical design by users.
4. 7120P/7120X feature Turbo.
5. 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* 2.0 specification compliant.
6. Refer to Section 5.1 for note on total card TDP of 5110P.

2.1.5 Intel® Xeon Phi™ Coprocessor 7120D/5120D(Dense Form Factor)

The Intel® Xeon Phi™ coprocessor 7120D/5120D products, also known as Dense Form Factor (DFF), are derivatives of the standard Intel® Xeon Phi™ coprocessor PCI Express* form-factor card. The high-level features of the DFFs are:

- Maximum TDP of 270W for the 7120D and 245W for the 5120D
- GDDR on both sides of the card.
- 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* 2.0 compliant. The unique edge finger pin definition requires signal routing on baseboard and 12V filter per card.
- All power to the card is supplied through the connector.
- There is no auxiliary 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.

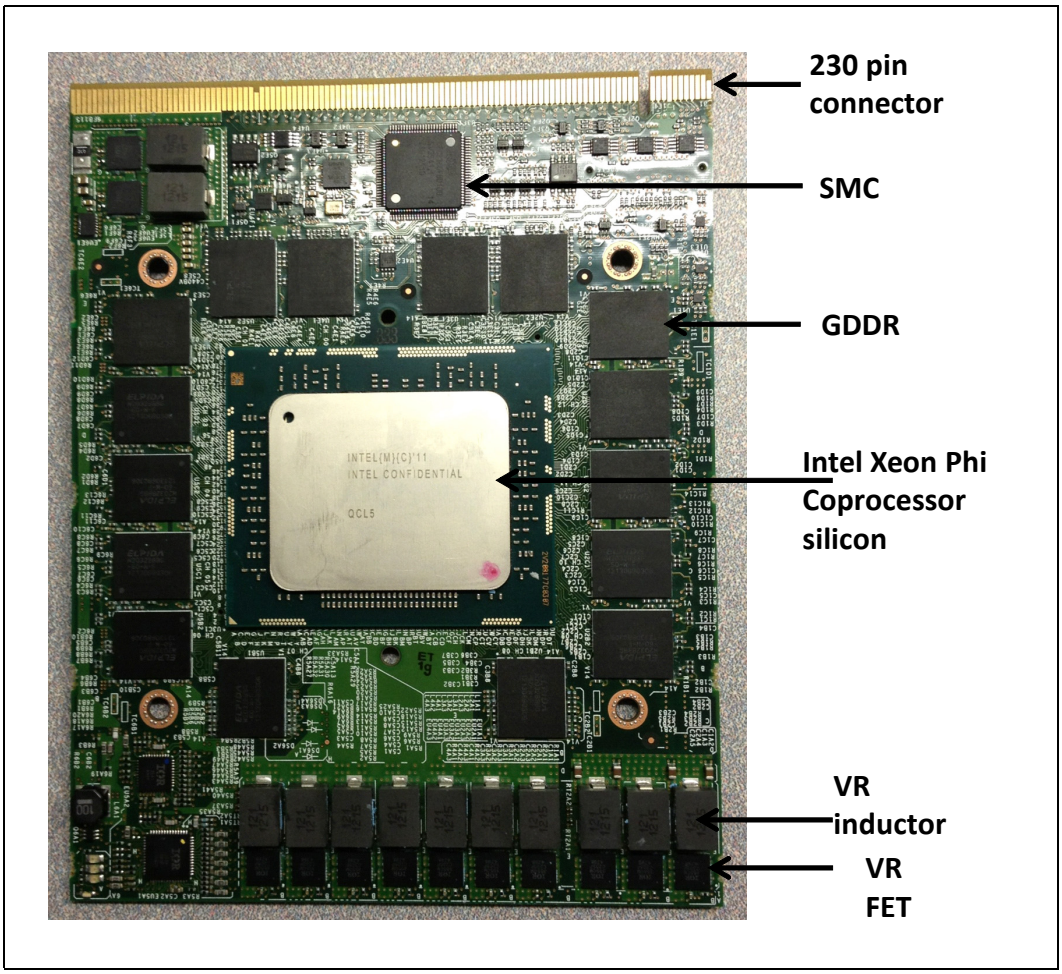


Figure 2-5. 7120D/5120D Dense Form Factor, Topside

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 SKUs.

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

Parameter	Specification
Product Length	247.9mm[1]
Primary Side Height Keep-in	34.8mm
Secondary Side Height Keep-in	2.67mm
3120A/7120A SKU mass	1400g
7120P/SE10P/5110P/3120P/31S1P SKUs mass	1200g
7120D/5120D SKUs mass	183g
7120X/SE10X SKUs mass	225g

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™ Coprocessor Thermal Mechanical Models* for PTC* Creo (previously Pro-Engineer), Icepak and FloTHERM* models.

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

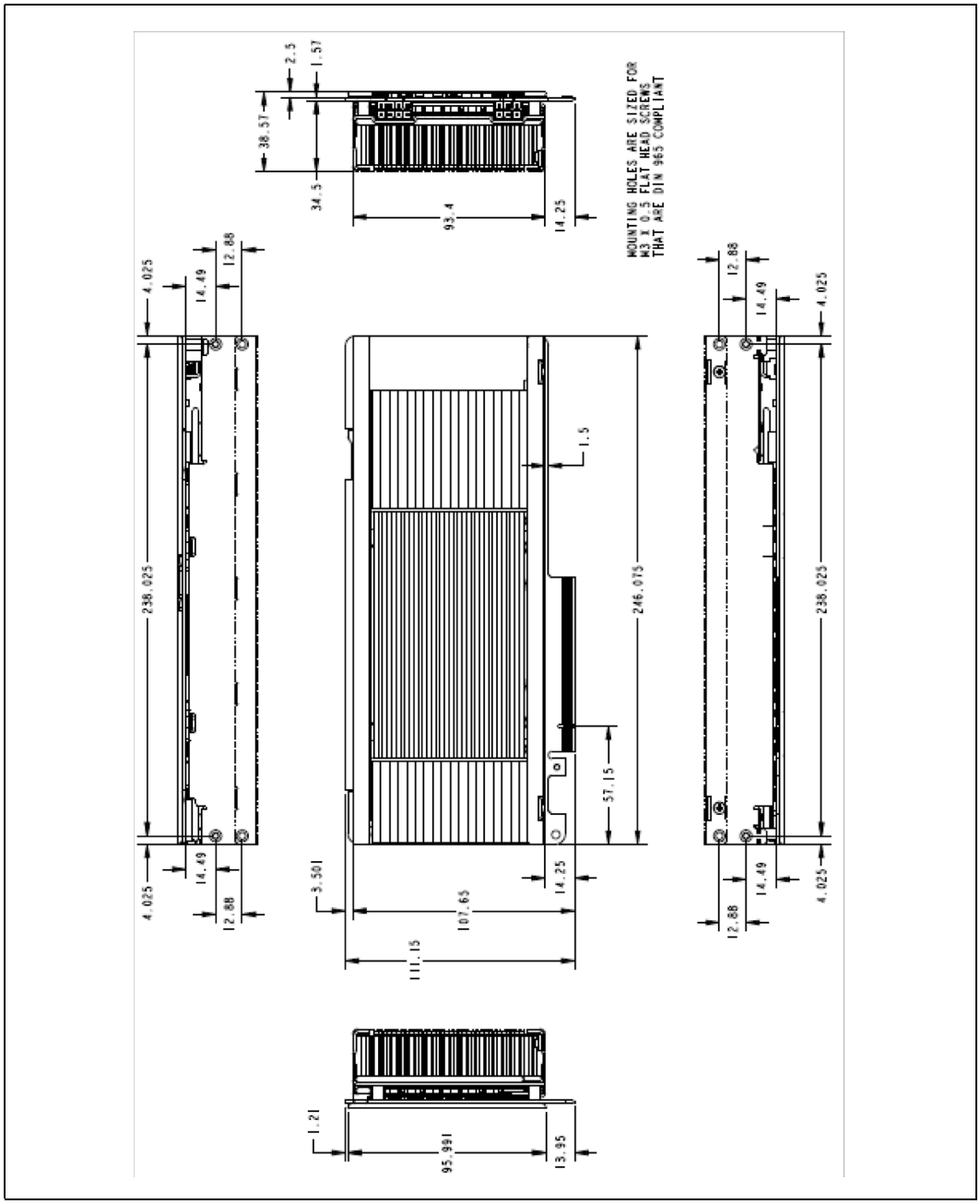
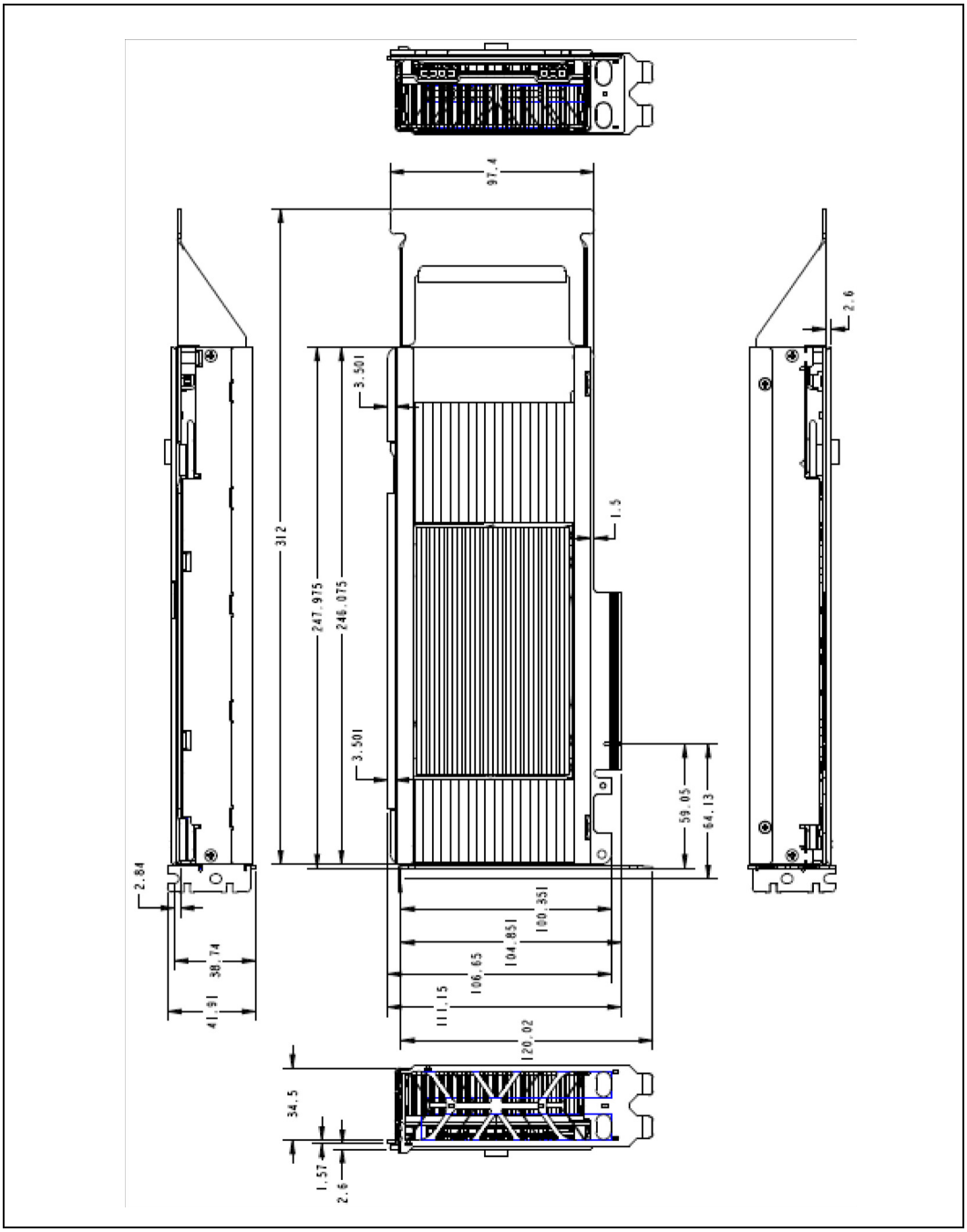


Figure 3-2 Dimensions of the Intel® Xeon Phi™ Coprocessor Card (in mils)



3.2 Intel® Xeon Phi™ Coprocessor Thermal Specification

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

Parameter	Specification
T_{RISE}	10°C
Max T_{INLET}	45°C
Max $T_{EXHAUST}$	70°C
$T_{case\ (processor)}$ min, max	5°C, 95°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 a requirement that the system BMC use IPMB commands to query the SMC on the coprocessor card for accurate $T_{control}$ value as this value can vary between 80°C and 84°C.
2. When the coprocessor junction temperature ($T_{junction}$) reaches $T_{throttle}$, the SMC will force thermal throttle which will drop frequency to lowest supported value and reduce total coprocessor power. It is a requirement that the system BMC query the SMC on the coprocessor card for accurate $T_{throttle}$ value.
3. If the coprocessor temperature reaches $T_{thermtrip}$, the coprocessor OS will take action to shutdown the card to prevent damage to the coprocessor. This includes shutting down the coprocessor VRs, and the only way to restart the coprocessor is by rebooting the host system. $T_{thermtrip}$ should not be considered a specification; it can change between SKUs, and is given here 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, also called junction temperature ($T_{junction}$). 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 read the junction temperature. 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.

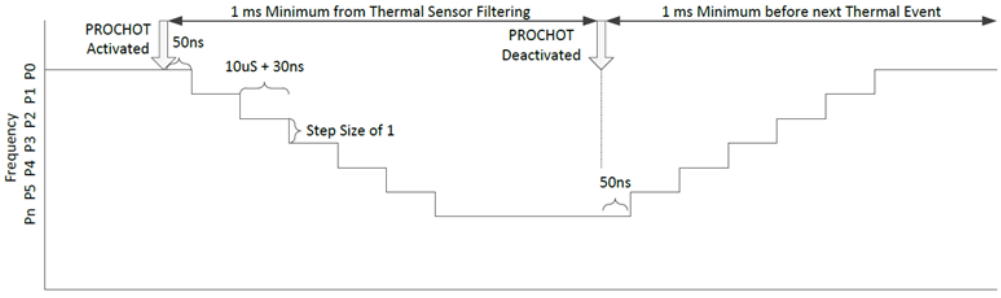
In addition to making thermal information available to system manageability software, the DTS is constantly comparing the coprocessor temperature to the factory set maximum permissible temperature called $T_{throttle}$. If the measured temperature at any time exceeds $T_{throttle}$ (a state also known 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 100ms, 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 to 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

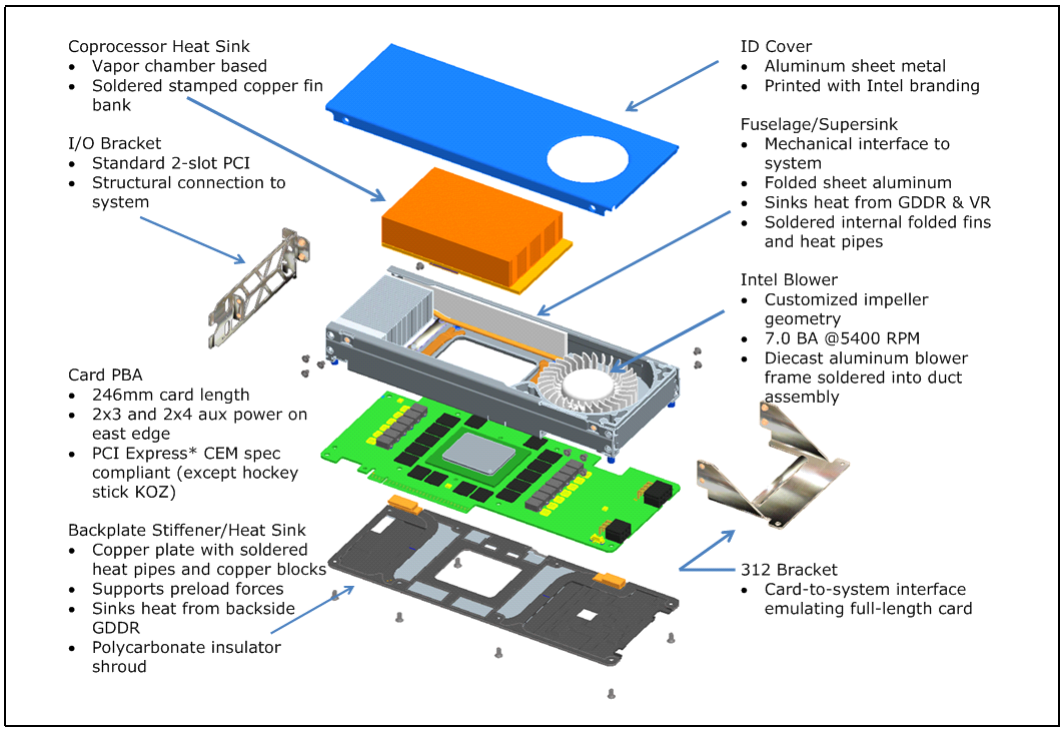
There are two types of thermal solutions to address the Intel® Xeon Phi™ coprocessor power limits: a passive solution for most SKUs as indicated in Table 2-1 (which relies on forced convection airflow provided by the system) and an active solution on the 3120A and 7120A SKUs (which uses a high performance 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 that augment the stiffness of the Intel® Xeon Phi™ coprocessor card by counteracting the preload applied by the primary side (housing the coprocessor). This also protects 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 2.67 mm 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 3120A and 7120A Active Cooling Solution

For the 3120A and 7120A SKUs, 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 3120A / 7120A Active Solution



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.

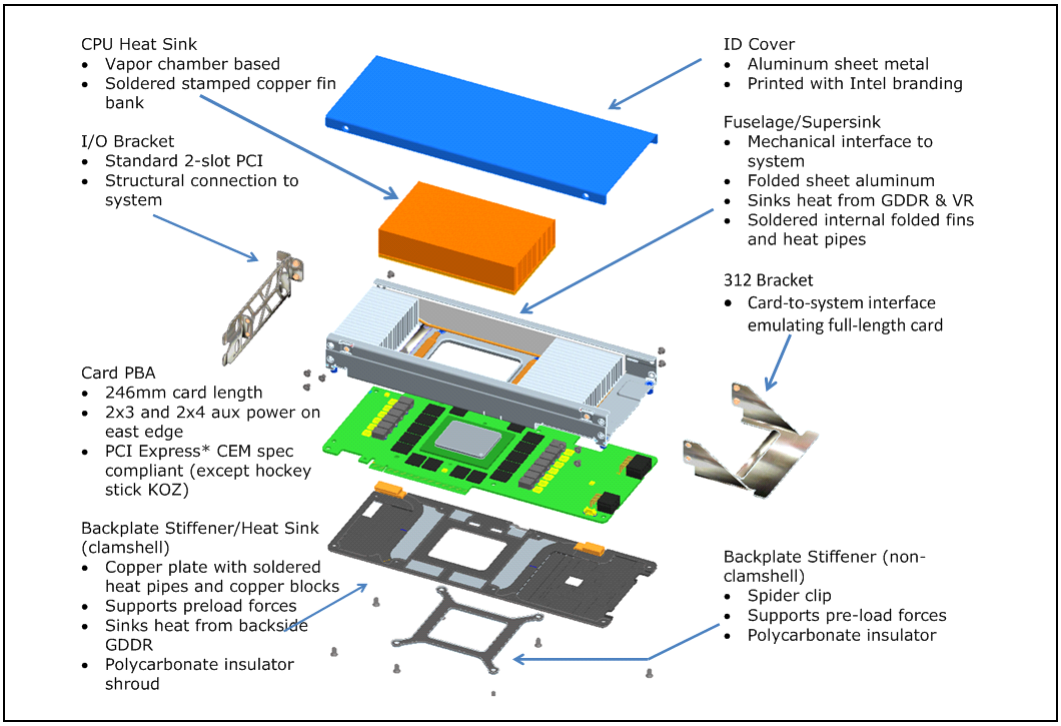
The coprocessor thermal path is separated from the GDDR and VR components, and utilizes a heatsink with parallel plate fins and vapor chamber base.

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 7120P/SE10P/5110P/3120P/31S1P Passive Cooling Solution

For the passive heat sink on the 7120P/SE10P/5110P/3120P/31S1P 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



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 SKUs

In order to ensure adequate cooling of the 5110P SKUs with a 45°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.

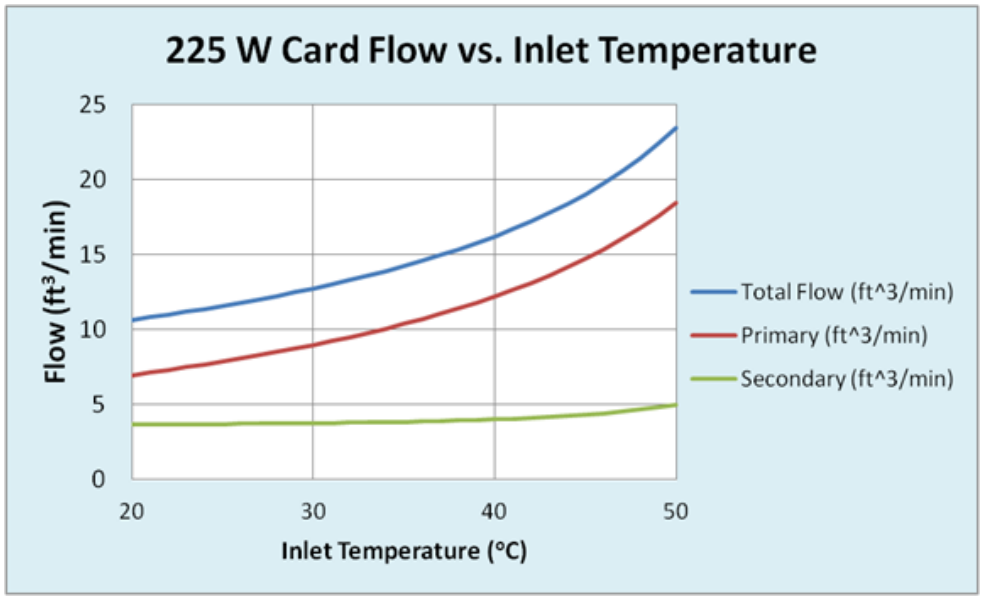
If the 5110P SKU is powered by a 2x4 and a 2x3 connector, the card can support an additional 20W of power for maximum TDP of 245W (see Section 2.1.5 for more details). In this case, the corresponding airflow requirement for cooling the part as a 245W card is shown in Figure 3-8.

3.3.2.2 Airflow Requirement for SE10P/7120P/3120P/31S1P Passive Cooling Solution

In order to ensure adequate cooling of the SE10P/7120P/3120P 300W and 31S1P 270W SKUs 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.

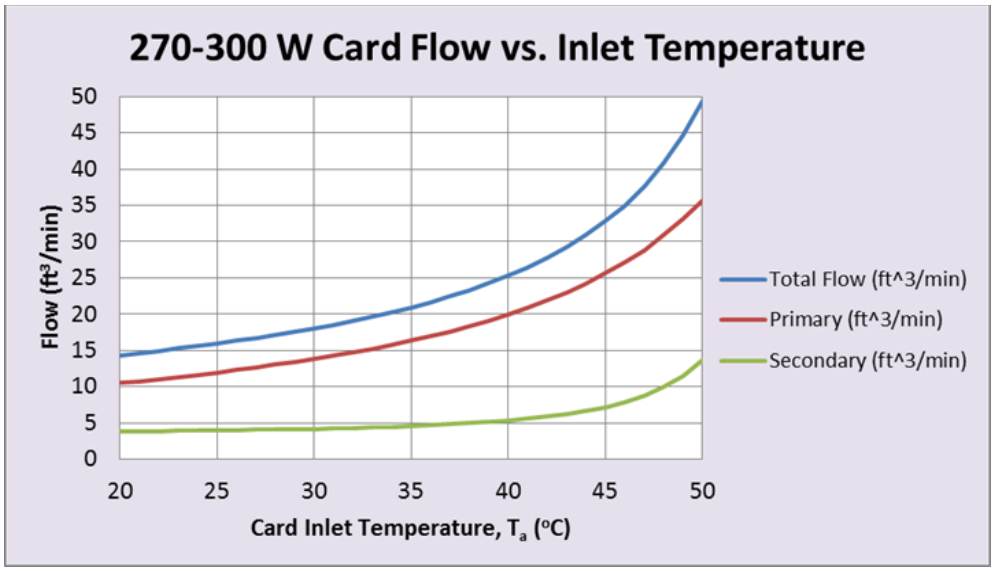
If the system is able to provide a temperature lower than 45^{o}C at the card inlet, then the total airflow can be reduced according to the graph and table in Figure 3-7.

Figure 3-6 Airflow Requirement vs. 45°C Inlet Temperature for the 5110P at 225W TDP



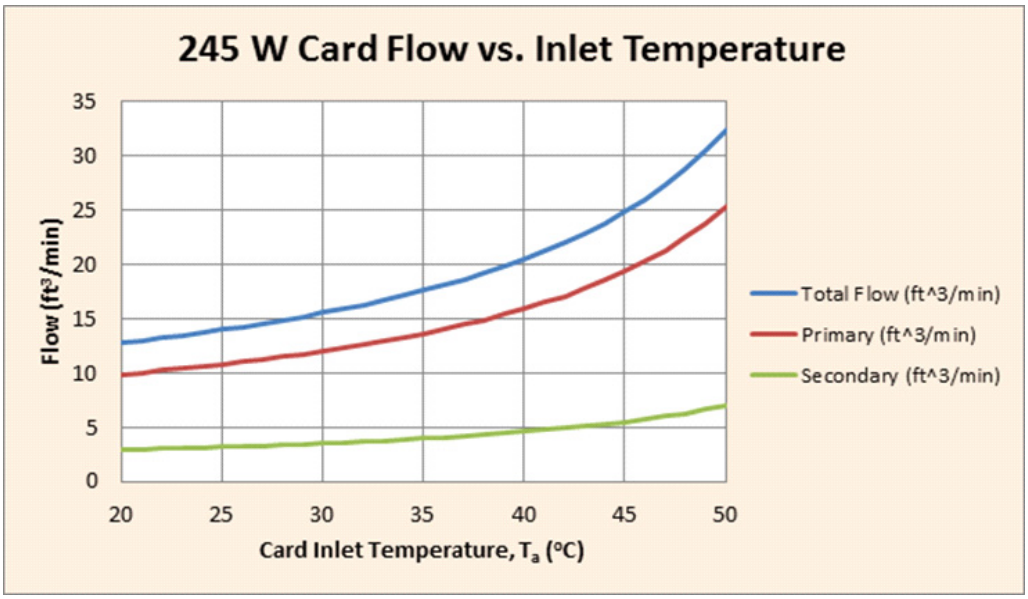
Card Inlet (°C)	Total Flow (ft^3/min)	Primary (ft^3/min)	Secondary (ft^3/min)	Card dP(inch H_2O)
20	10.6	7	3.7	0.08
21	10.8	7.1	3.7	0.08
22	11	7.3	3.7	0.08
23	11.2	7.5	3.7	0.08
24	11.4	7.7	3.7	0.09
25	11.6	7.9	3.7	0.09
26	11.8	8.1	3.7	0.09
27	12	8.3	3.7	0.09
28	12.2	8.5	3.7	0.1
29	12.5	8.7	3.7	0.1
30	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 31S1P at 270W TDP and SE10P/7120P/3120P at 300W TDP



Card Inlet (°C)	Total Flow (ft^3/min)	Primary (ft^3/min)	Secondary (ft^3/min)	Card dP (inH_2O)
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 at 245W TDP[1]



Card Inlet (°C)	Total Flow (ft^3/min)	Primary (ft^3/min)	Secondary (ft^3/min)	Card dP (inH_2O)
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

Notes:

1. Refer to Section 5.1 for note on 5110P TDP.

3.4 Cooling Solution Guidelines for SE10X/7120X and 7120D/5120D

The Intel® Xeon Phi™ coprocessor SE10X/7120X and 7120D/5120D SKUs are shipped without a thermal solution, which gives system designers and integrators an opportunity to fit these SKUs into their custom designed chassis. These SKUs have 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/7120X products. Figure 3-11 to Figure 3-14 show a schematic representation of the power profiles of the 7120D/5120D product.

Figure 3-9 SE10X/7120X Power Profile for Coprocessor Intensive Workload (all values in Watts)

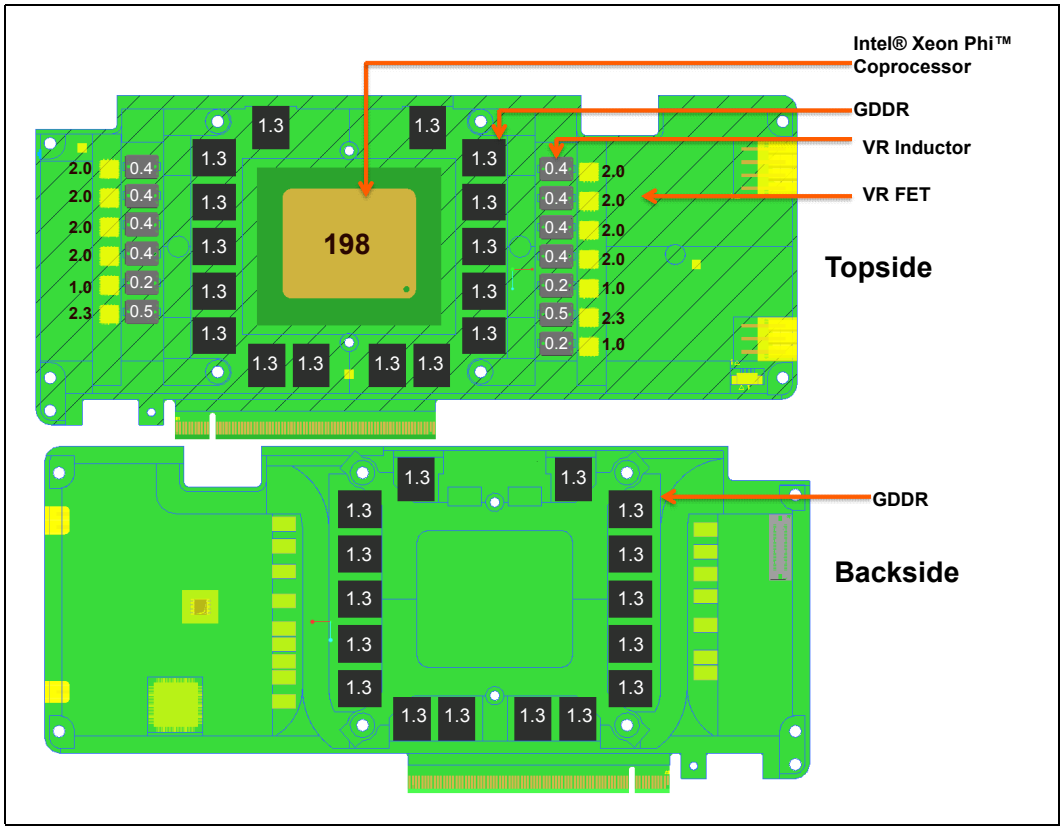
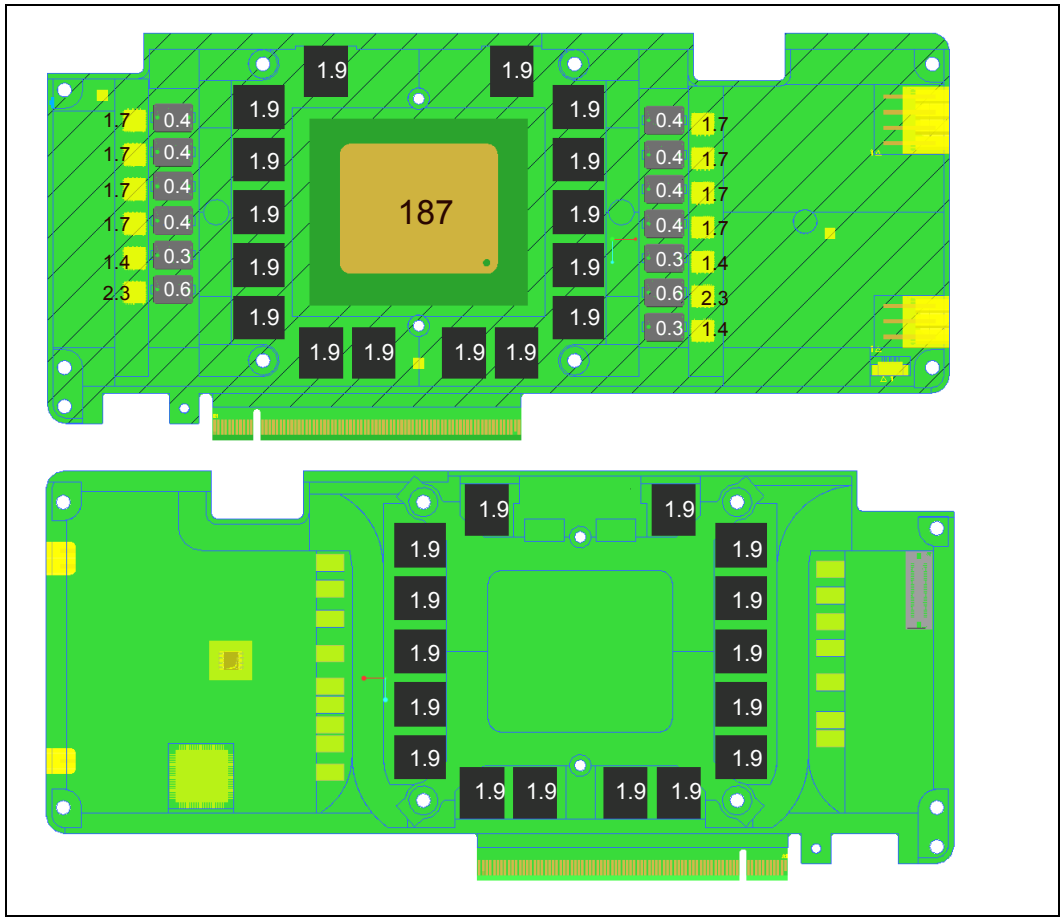


Figure 3-10 SE10X/7120X Power Profile for Memory Intensive Workload (all values in Watts)



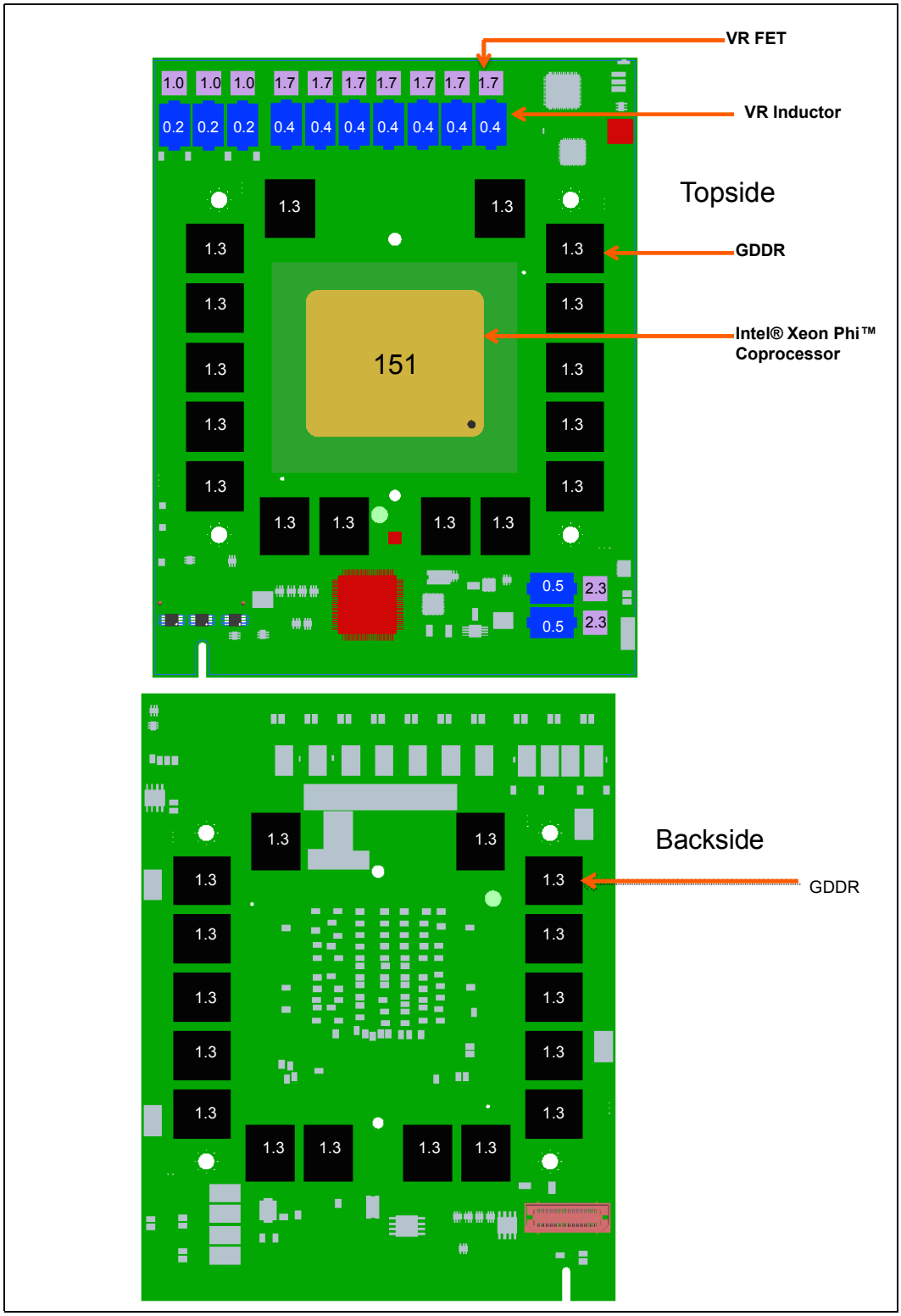


Figure 3-11 5120D Power Profile: Coprocessor Centric (all values in Watts)

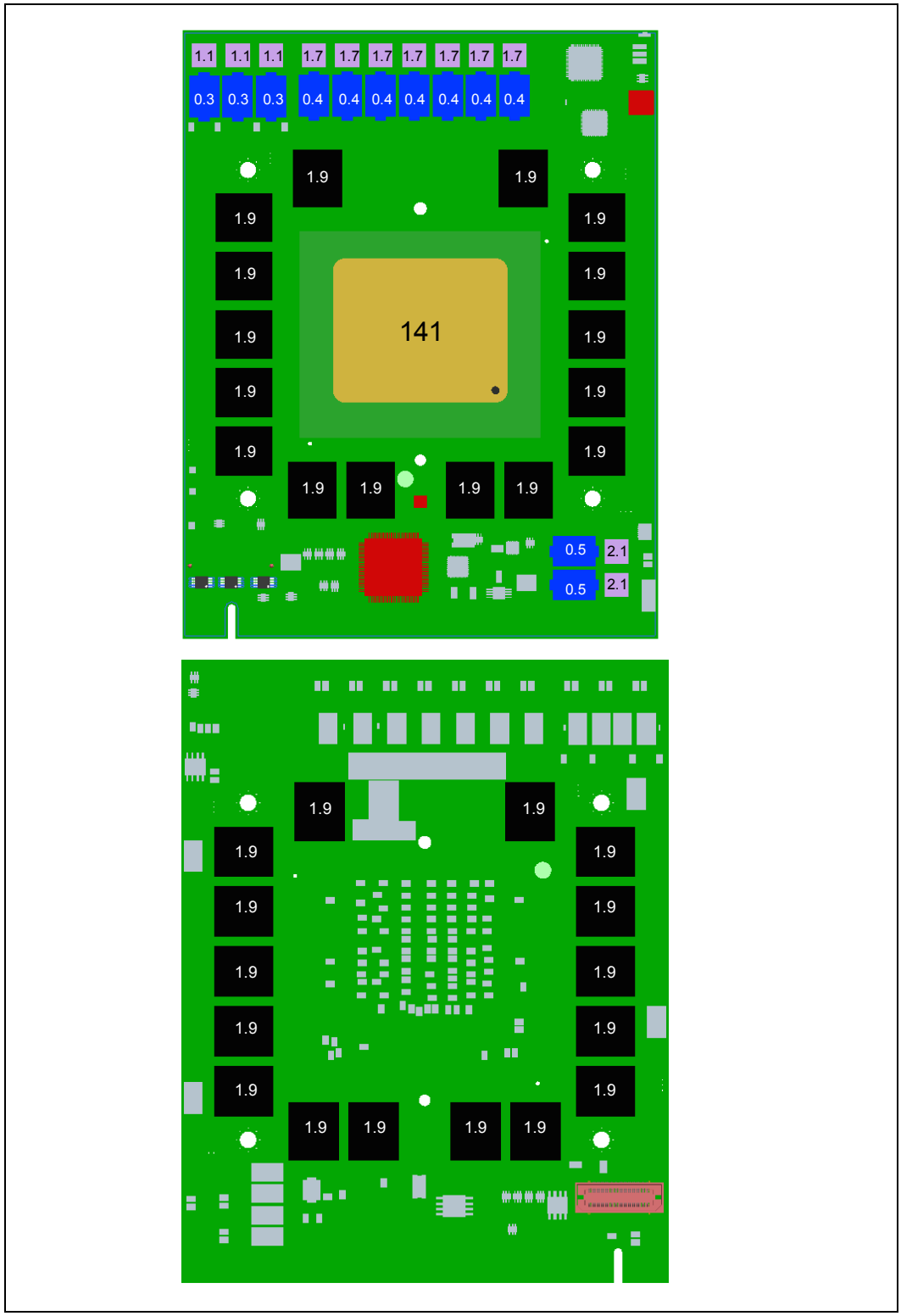


Figure 3-12 5120D Power Profile: Memory Centric (all values in Watts)

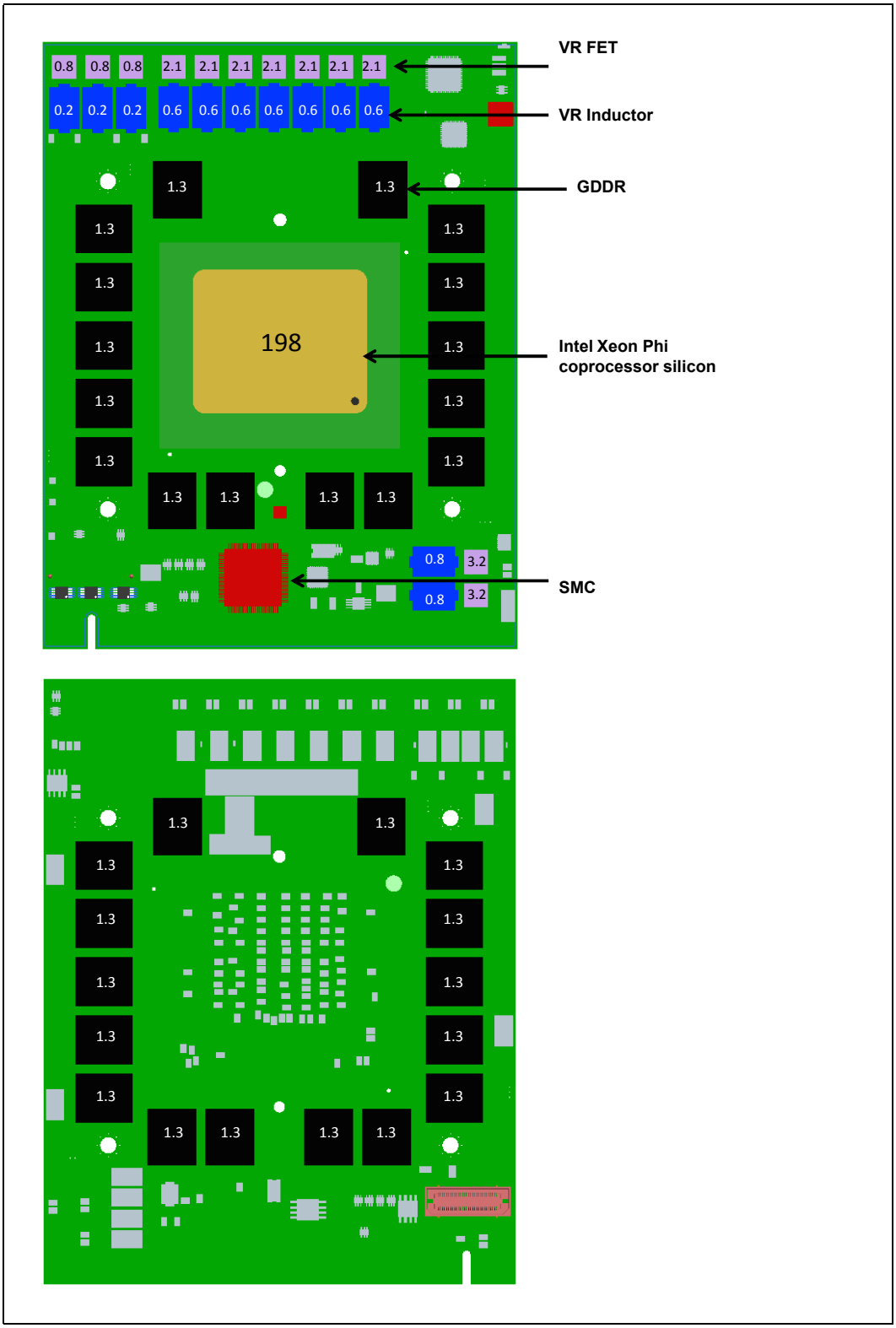


Figure 3-13 7120D Power Profile: Coprocessor Centric (all values in Watts)

Figure 3-14 7120D Power Profile: Memory Centric (all values in Watts)

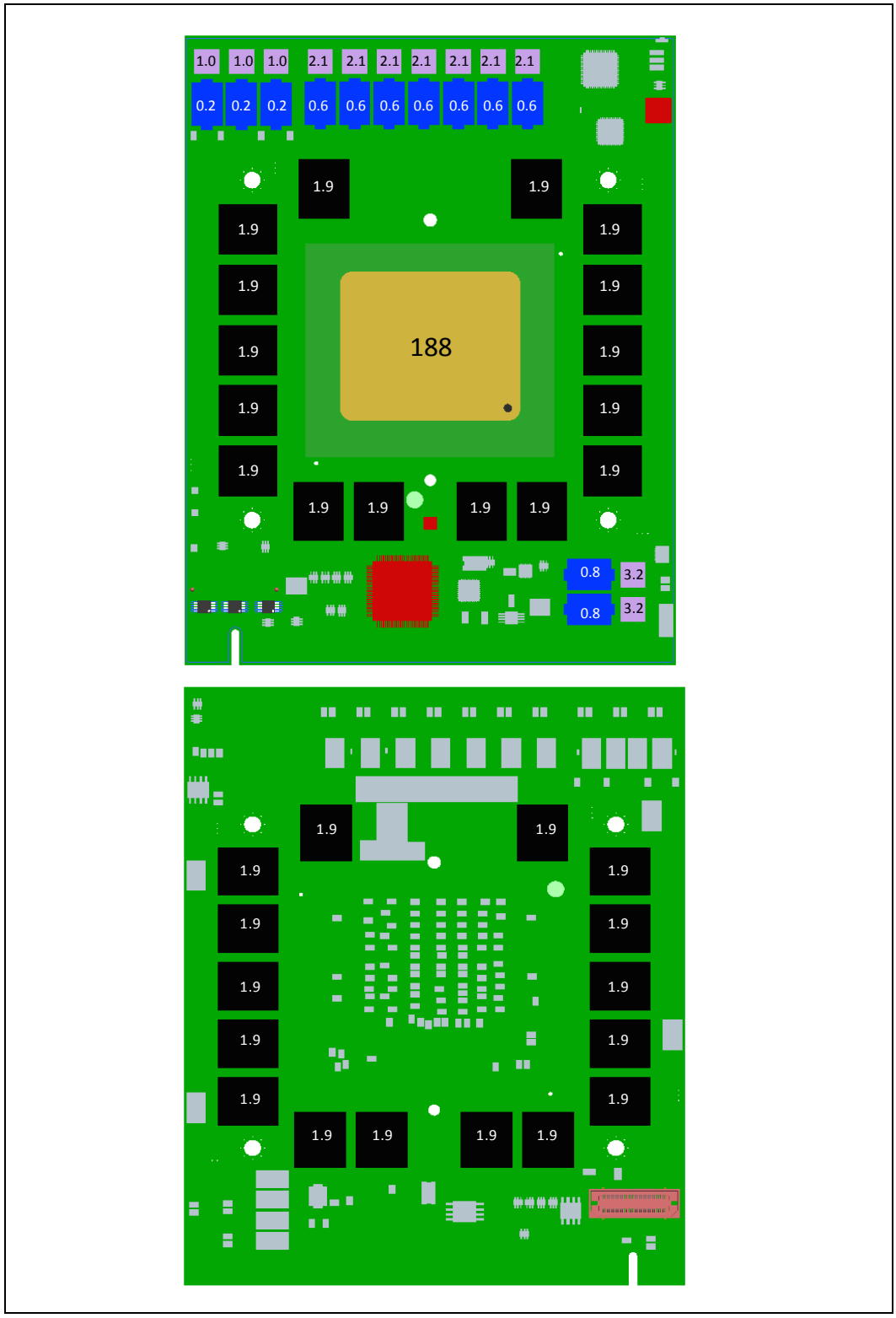


Table 3-3 shows thermal specifications of components present on the SE10X/7120X and 7120D/5120D boards.

Table 3-3. Component Thermal Specification on SE10X/7120X and 7120D/5120D

Component	Thermal specification
Coprocessor	$T_{case} \leq 95°C$
GDDR	$T_{case} \leq 85°C$
VR FET	$T_{junction} \leq 150°C$[1]
VR Inductor	$T_{body} \leq 100°C$

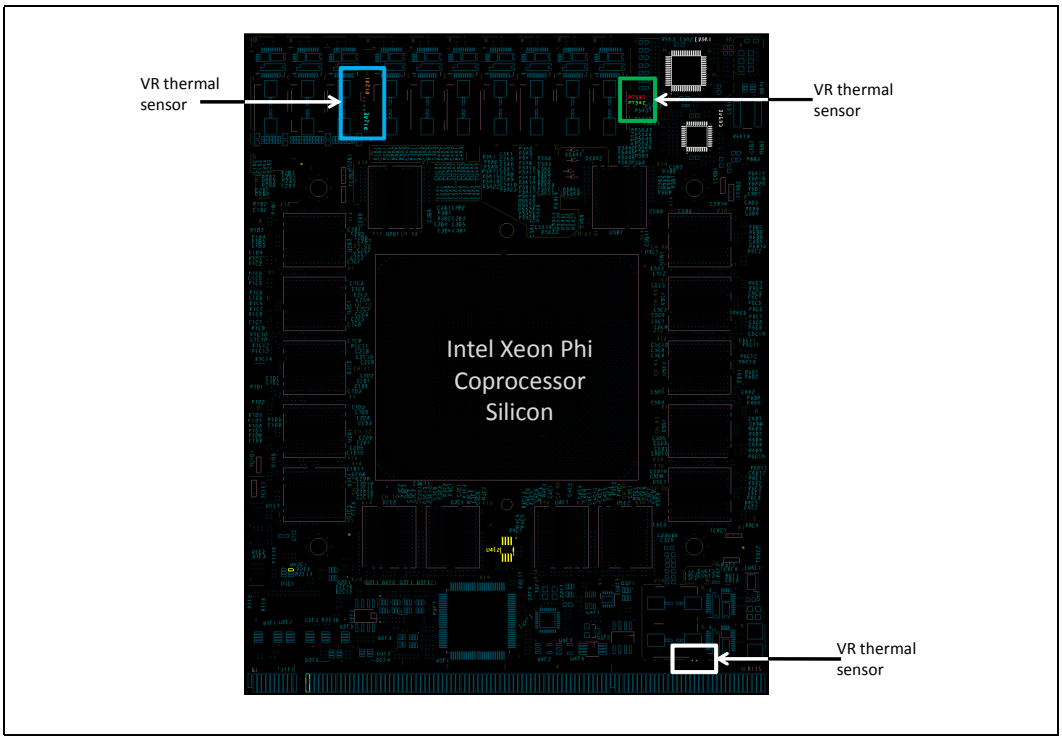
Notes:

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.

3.4.1.1 VR Temperature and Thermal Throttling

Some thermal sensors on the coprocessor are located in the field of the VR inductors and FETs that drive power to the coprocessor silicon, GDDR and other circuitry (Figure 3-15). These sensors continuously monitor the temperature of the circuit board. If the temperature reaches or exceeds 105° C, an interrupt will be sent to the SMC. In response to this interrupt, the SMC will assert PROCHOT to the coprocessor which in turn will force the coprocessor frequency to drop to the minimum supported value (approximately 600MHz). This frequency drop will reduce the power dissipation of the card and should allow the VR components to cool down. Once the temperature has dropped below 105° C, the VR controller will send a message to the SMC to de-assert PROCHOT. This will allow the coprocessor to return to the normal high frequency operating point.

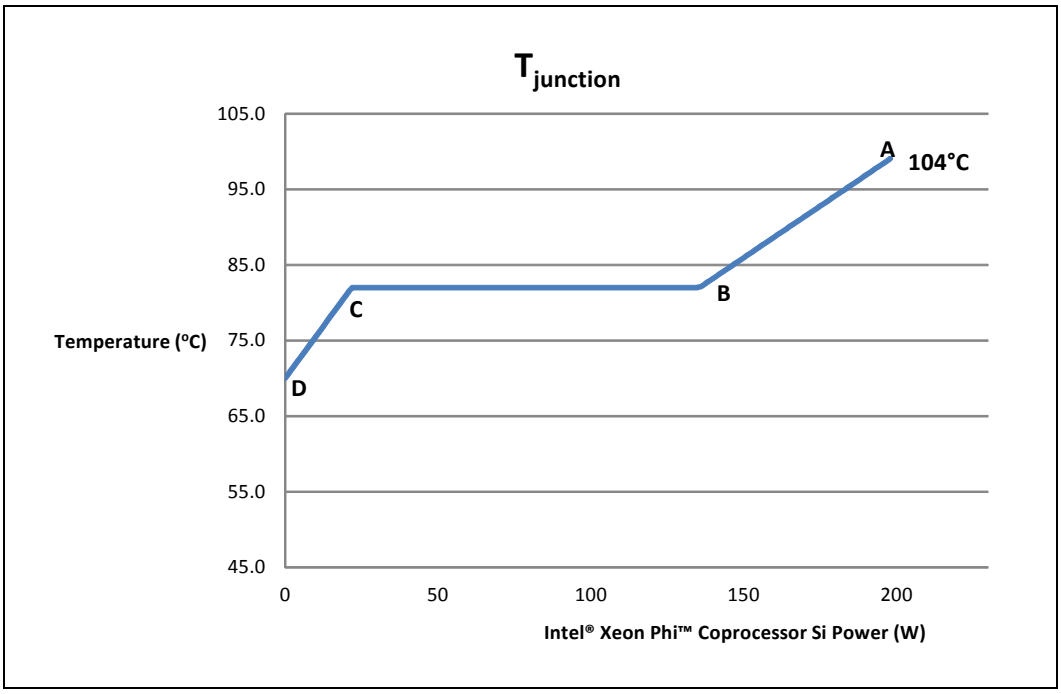
Figure 3-15 7120D/5120D VR Thermal Sensors for Custom Cooling Consideration



3.4.2 Thermal Profile and Cooling

The simplest cooling mechanism would involve running fans at full speed. For those custom air-cooled solutions that intend to be economical in fan power usage and acoustics, Figure 3-16 represents three regions on the SE10X/7120X coprocessor power consumption curve relevant to system fan control.

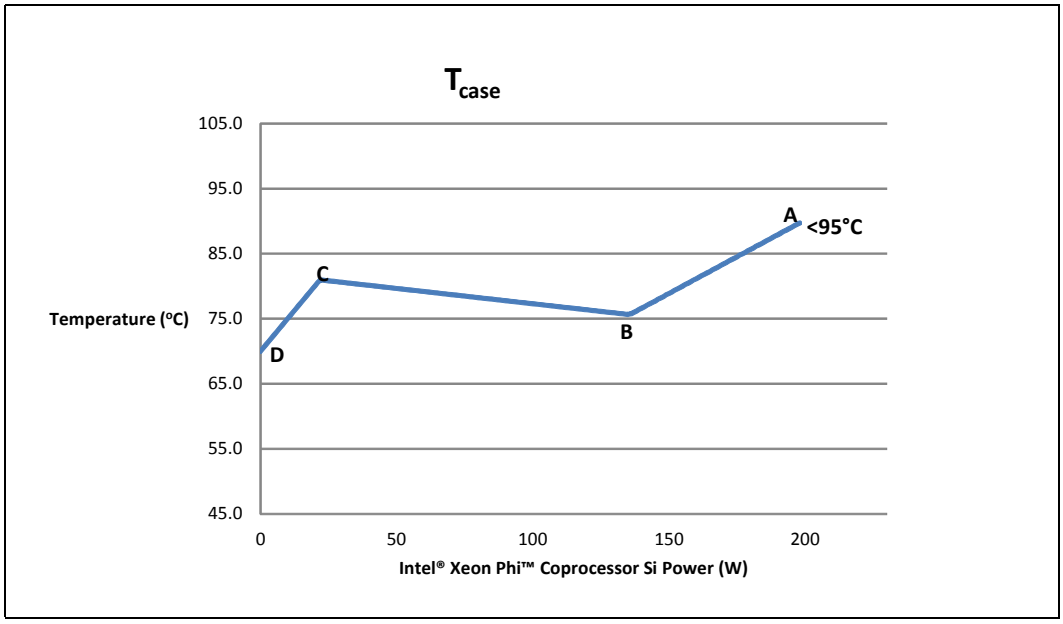
Figure 3-16 SE10X/7120X SKU Coprocessor Junction Temperature ($T_{junction}$) vs Power



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 3-17 shows the analogous thermal behavior of T_{case}.

Figure 3-17 SE10X/7120X 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 Ψ_{ca_req} value adequate to keep the coprocessor junction temperature at or below 104°C. The value for Ψ_{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 Ψ_{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 Ψ_{ca}, but a fixed Ψ_{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 Ψ_{ca}.

3.4.3 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.
- Hitachi Type7 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/7120X and 7120D/5120D products, along with the heights used in the product board design. Figure 3-18 and Figure 3-19 show the front and back sides of the SE10X/7120X SKU. Refer to the *Intel® Xeon Phi™ Coprocessor Thermal Mechanical Models* document for the SE10X/7120X SKU. Refer to the *Intel® Xeon Phi™ Coprocessor Dense Form Factor Models* document for information on 5120D SKU.

Table 3-4. Board Component Heights

Block	Color[1]	Component Height (mils)		
		Min	Typ	Max
PCB Thickness		57	62	70
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 3-18, Figure 3-19, Figure 3-20 and Figure 3-21.

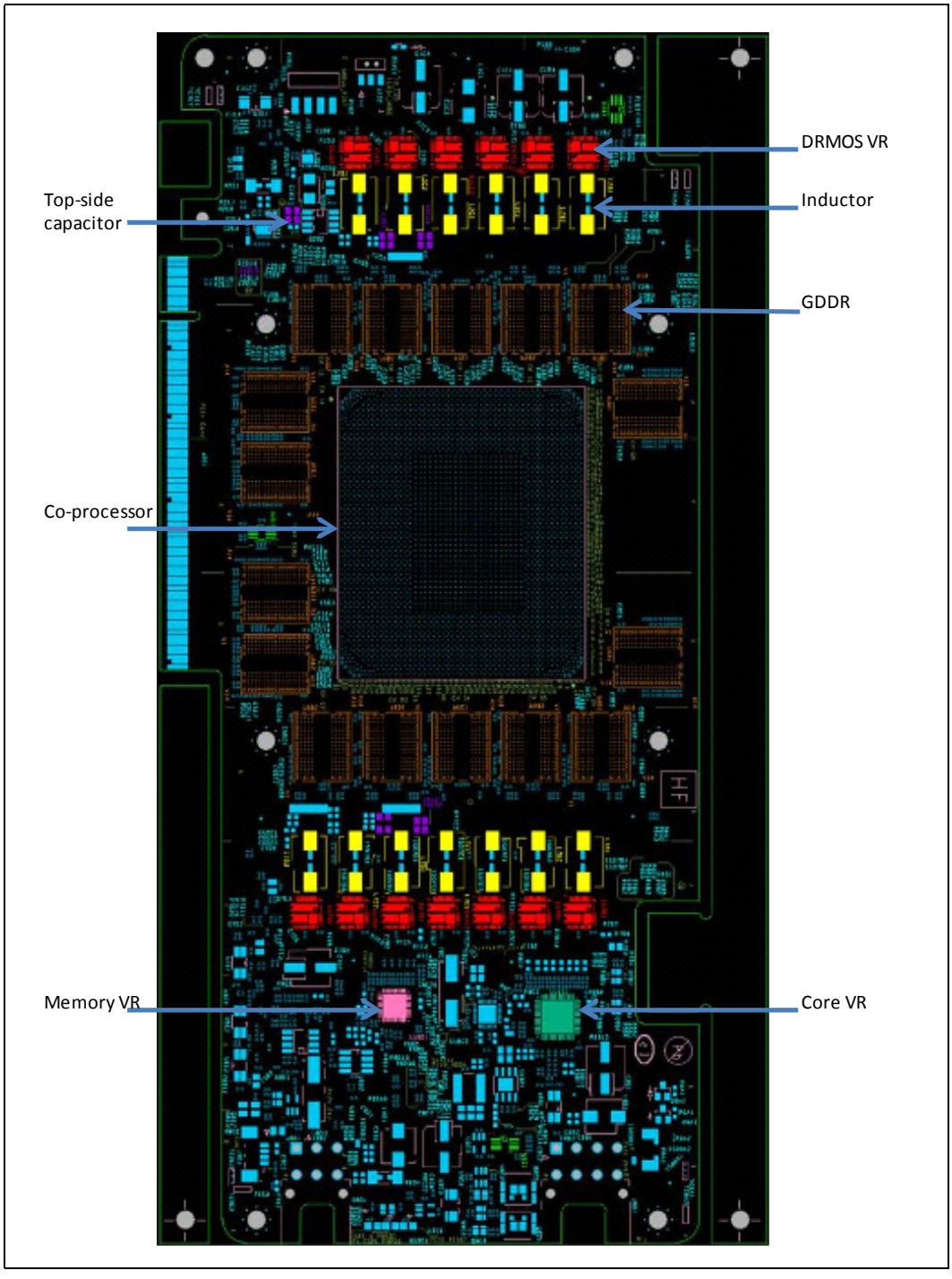
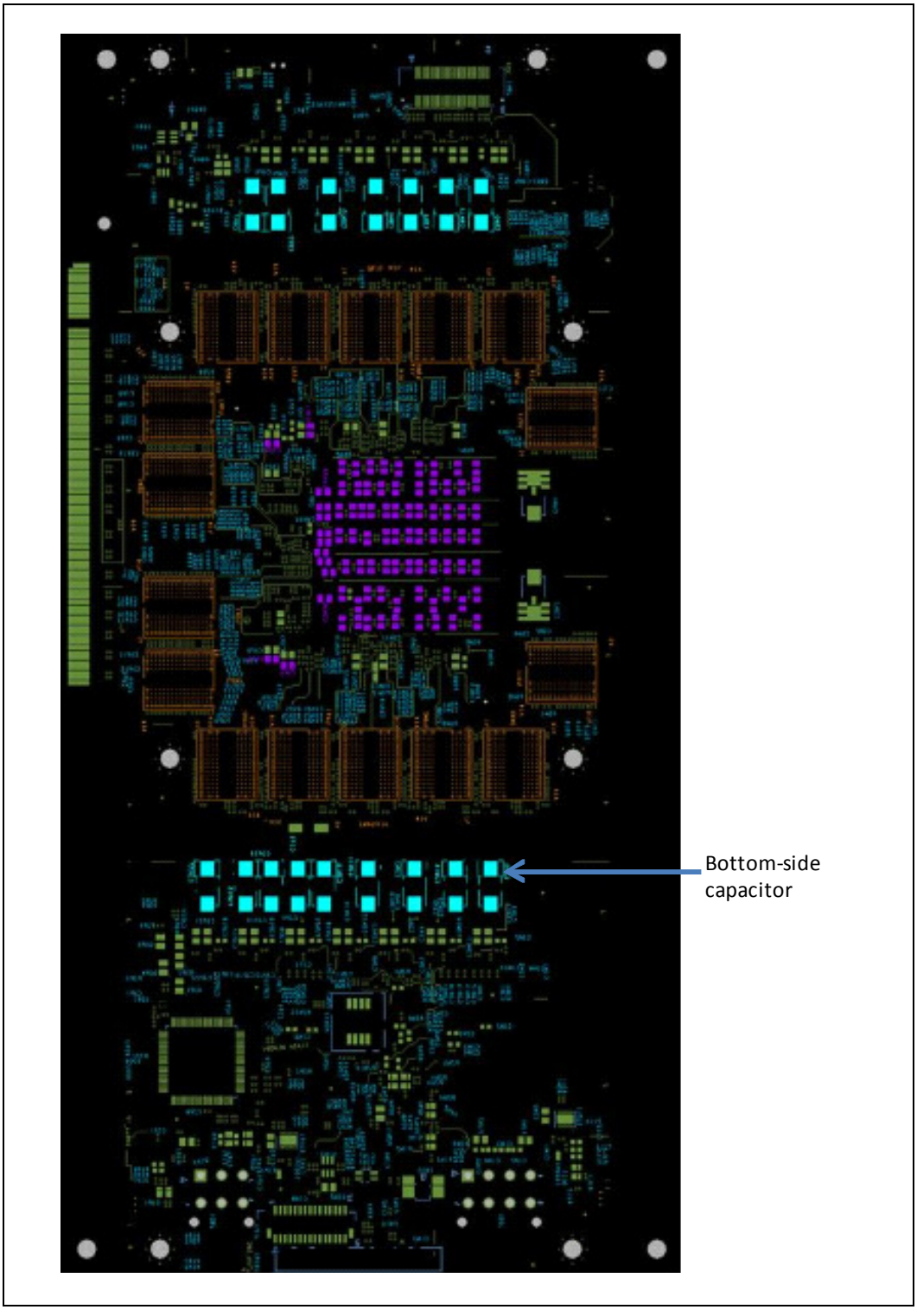


Figure 3-18 SE10X/7120X Board Top Side

Figure 3-19 SE10X/7120X Board Bottom Side



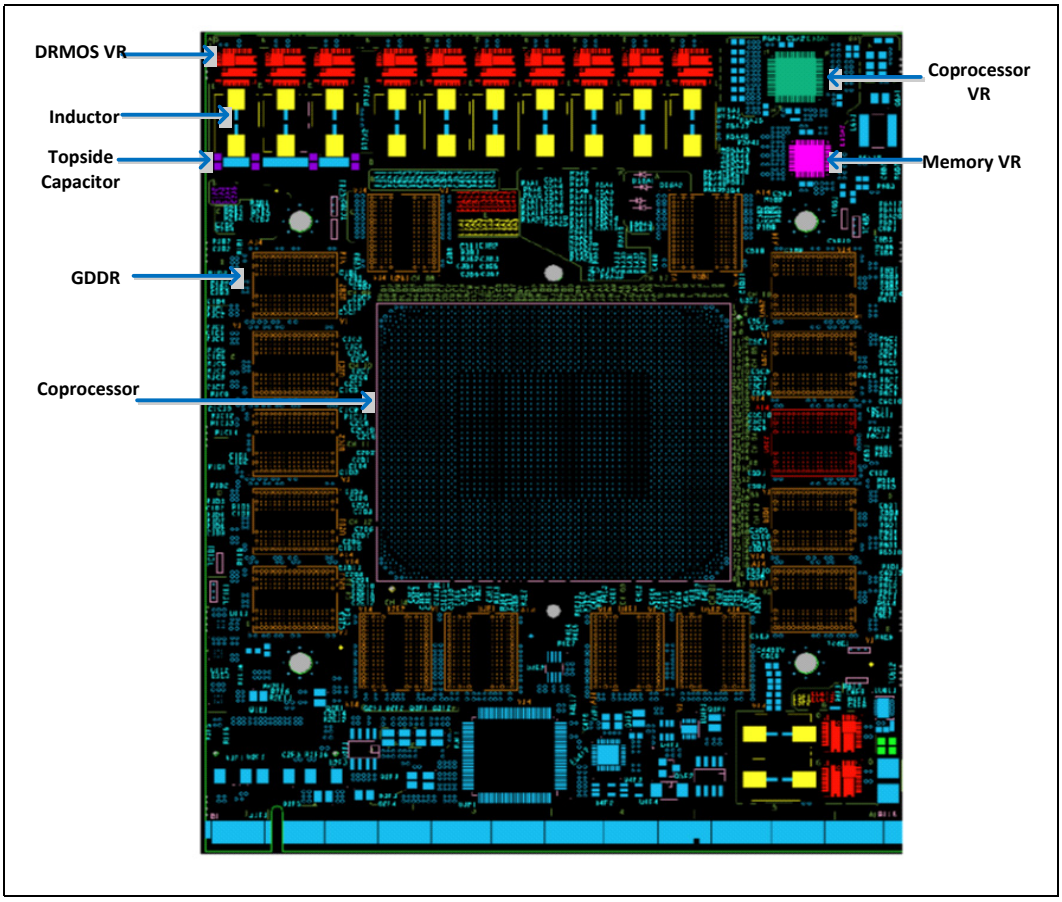
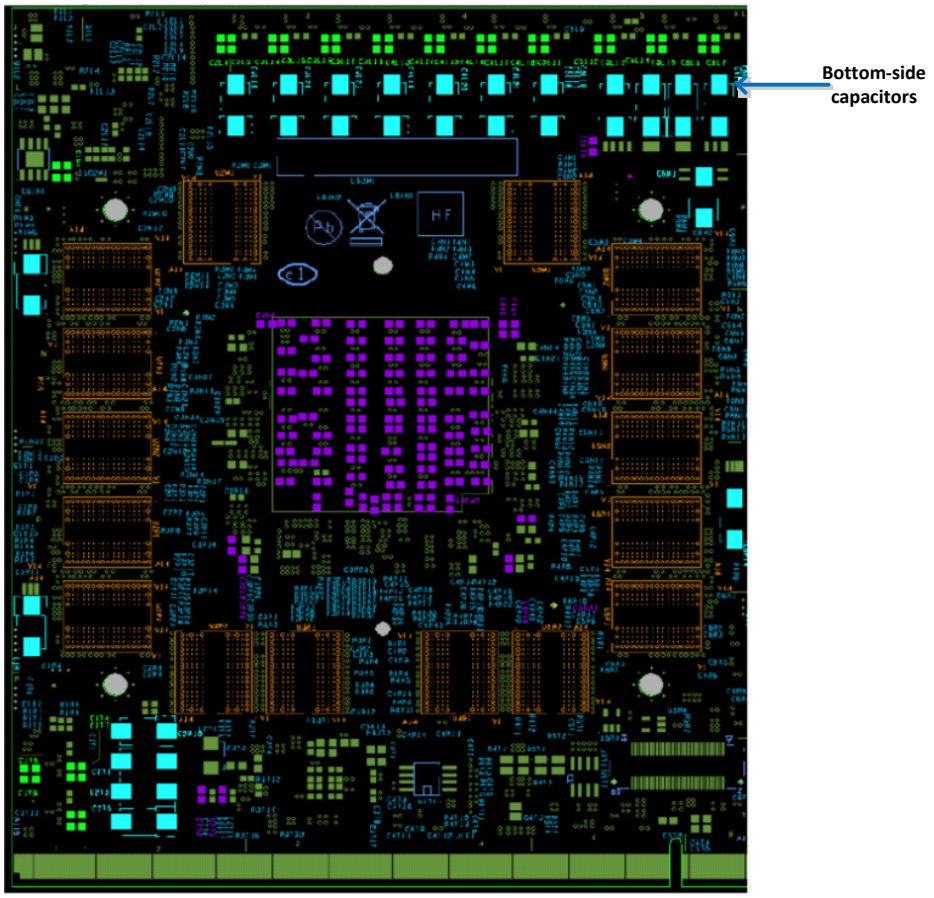


Figure 3-20 7120D/5120D Board Top Side

Figure 3-21 7120D/5120D Board Bottom Side



3.4.4 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 drops: 3x each on 6 faces
Board Unpackaged Random Vibration	5Hz @ $0.01g^2$/Hz to 20Hz @$0.02g^2$/Hz (slope up) 20Hz to 500Hz @ $0.02g^2$/Hz (flat) Input acceleration is 2313g RMS 10mins per axis in all 3 axis

Table 3-5. Dynamic Load Shift Specification

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

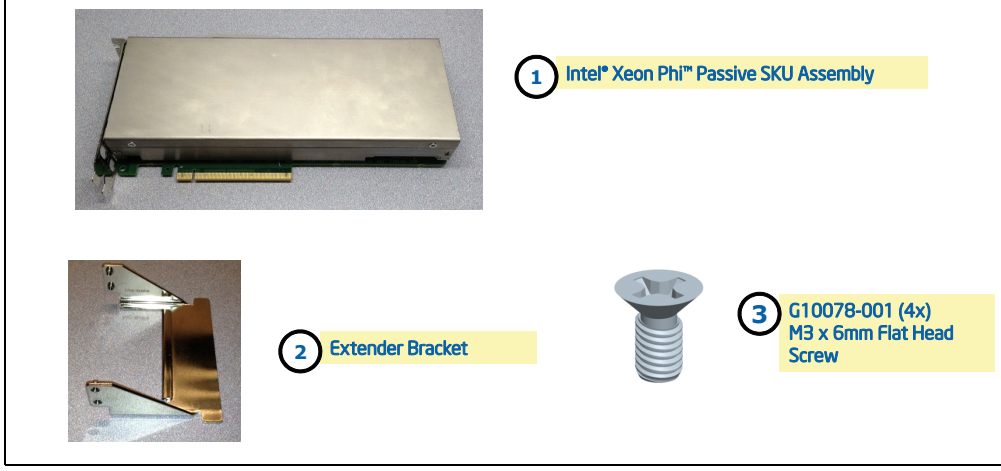
3.5 Intel® Xeon Phi™ Coprocessor PCI Express* Card Extender Bracket Installation

Intel® Xeon Phi™ coprocessor cards are shipped without the PCI Express* bracket (also known as extender bracket) being 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:

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

Note: The SE10X/7120X and 7120D/5120D SKUs are not shipped with the extender bracket.

Figure 3-22 Contents of Intel® Xeon Phi™ Coprocessor Package Shipment



3.5.1 Bracket Installation Steps

1. Determine Lid Type.
   If the lid type is "overlap" where the lid covers the top mounting holes as shown in Figure 3-23, then go to Step 2.
   If the lid type is "clearance" where the lid has cut-outs for mounting holes as shown in Figure 3-24, then go to Step 3.

Figure 3-23 Overlap Lid

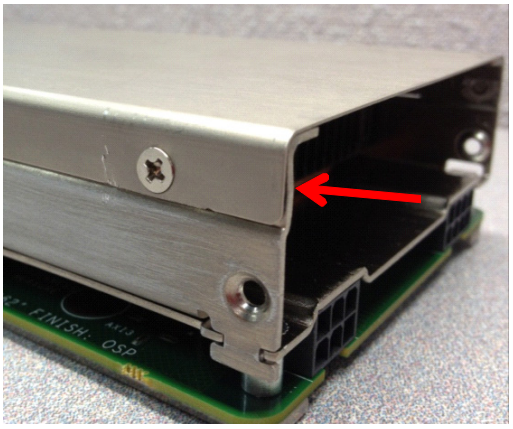
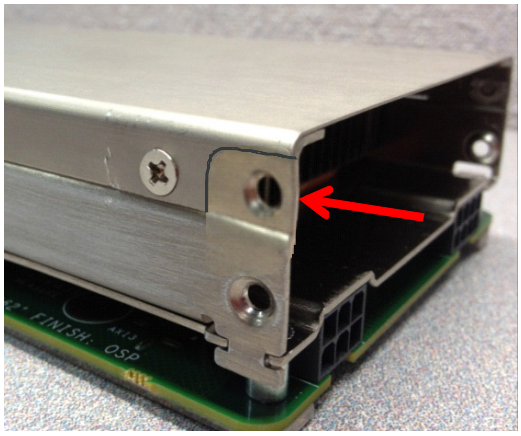
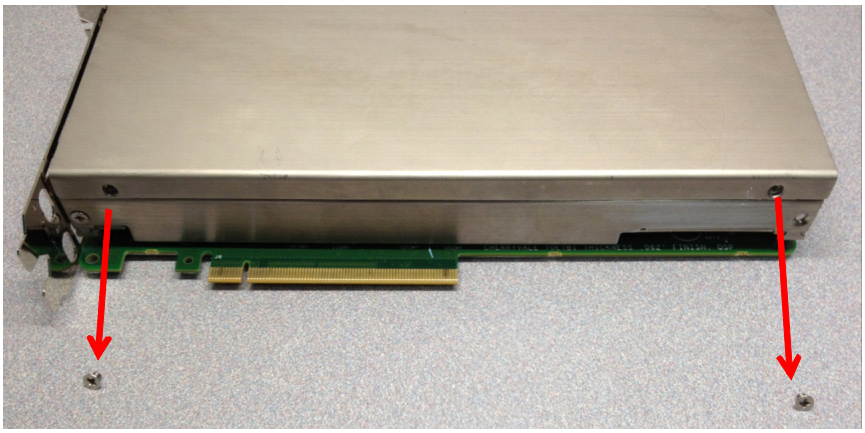


Figure 3-24 Clearance Lid



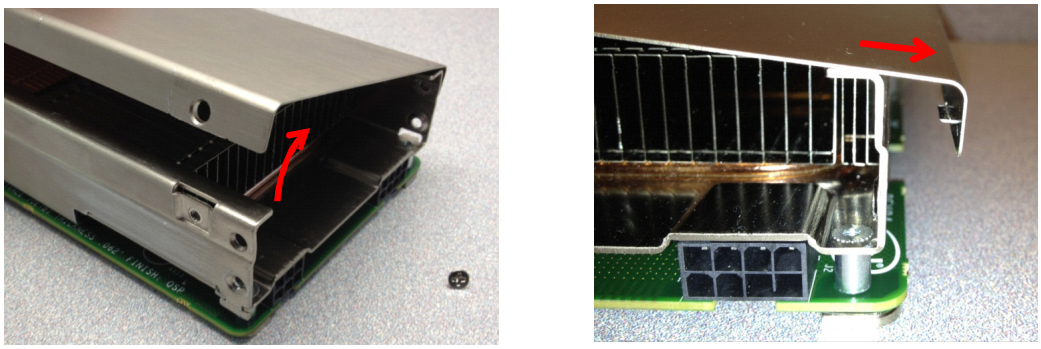
2. Remove Overlap Lid.
   a. Remove 2 of the M3x6mm screws retaining the lid, as shown in Figure 3-25.

Figure 3-25 Overlap Lid Removal



   b. Remove Lid. Take care not to bend tabs, as shown in Figure 3-26.

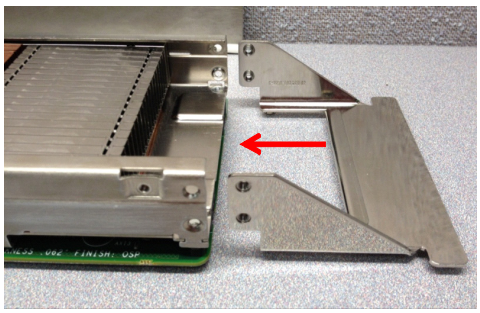
Figure 3-26 Tilt Overlap Lid and Slide as shown to Disengage Tabs



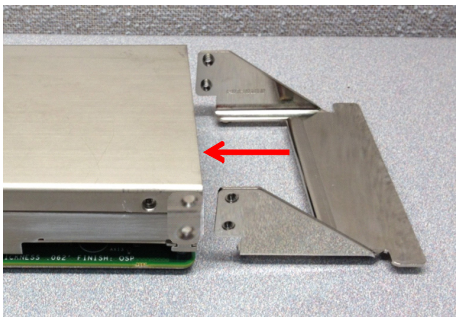
3. Install OEM Bracket.

   a. Insert the OEM bracket into the Intel® Xeon Phi™ coprocessor card assembly, as shown in Figure 3-27.

Figure 3-27 OEM Bracket Installation



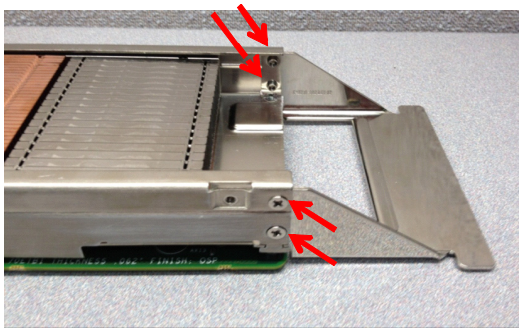
"Overlap Lid" Units



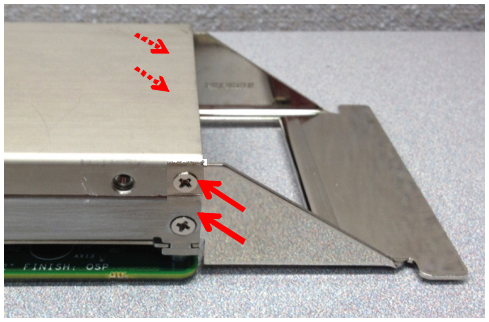
"Clearance Lid" Units

   b. Install (4) M3 x 6mm Flat Head Screws; torque = 6inch-lbs, shown in Figure 3-28.

Figure 3-28 OEM Bracket Installation



"Overlap Lid" Units

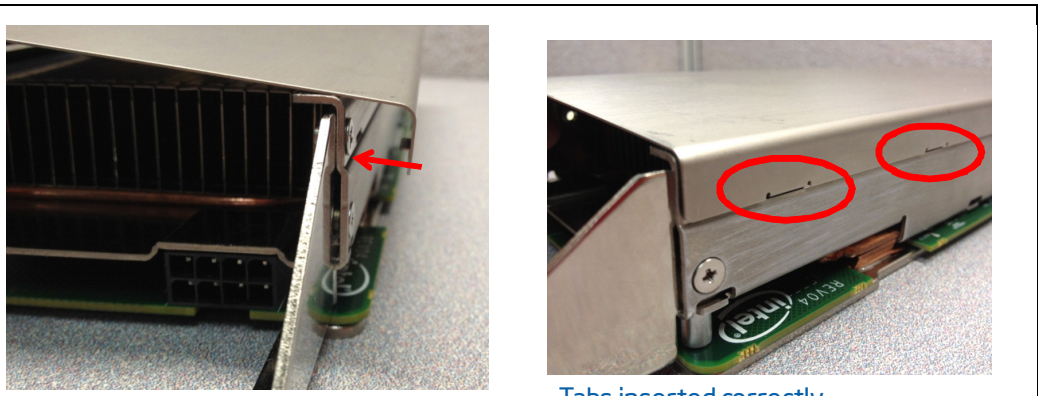


"Clearance Lid" Units

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

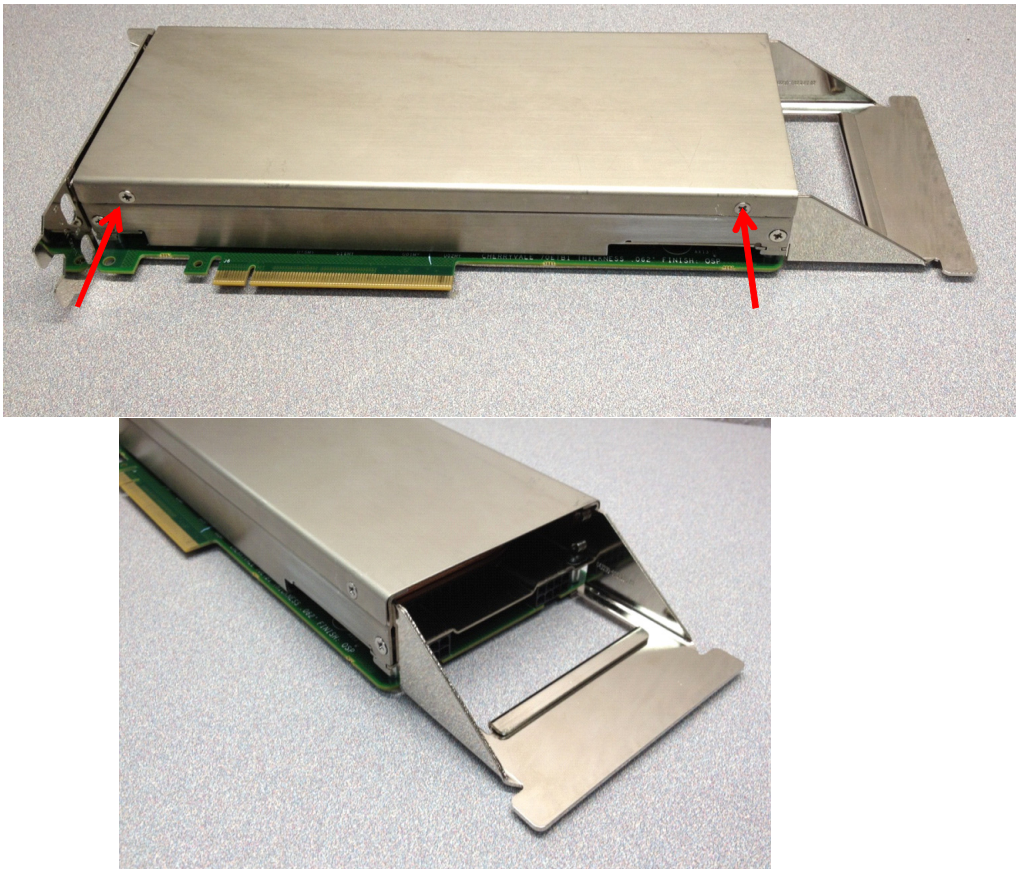
4. Replace Lid on "Overlap Lid" Units Only
   a. Insert tabs into slots in card assembly, shown in Figure 3-29.

Figure 3-29 Replace Lid on "Overlap Lid" Units



   b. Install the lid's screws (M3 x 6mm Flat head); torque = 6 inch-lbs, shown in Figure 3-30.

Figure 3-30 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"*. Signals called out in the PCI Express* specification but not used on the Intel® Xeon Phi™ coprocessor are 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 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 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 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.
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.

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

Signal Name	Signal Type	Description
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 inactive-high state by the card SMC, and must be driven active-low by the baseboard to exert throttling. 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 power supplied 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.

4.3 Dense Form Factor (5120D) Edge Connector Pins

The Intel® Xeon Phi™ coprocessor 5120D SKU (DFF) uses a 230-pin edge finger designed to industry standard x24 PCI Express* connector, PCI Express* Gen2 compliant. Unlike the other SKUs in the Intel® Xeon Phi™ coprocessor family, the 5120D SKU requires PCI Express* signal routing and +12V filter per card on the baseboard.

The symbol _N at the end of a signal name indicates that the active or asserted state occurs when that signal is at a low voltage level.

The following notations are used to describe the signal type:

Type	Details
I	Signal is an Input to the baseboard from 5120D.
O	Signal is an Output from the baseboard to 5120D.
I/O	Bidirectional I/O signal.
Power: +12V, +3.3V, +3.3V_AUX, GND	Power supply and GND inputs to the 5120D are sourced from the baseboard. On the 5120D, the +3.3V_AUX power supply is not electrically routed on the board. The baseboard may choose either to connect +3.3V_AUX supply to this pin or to leave it as No-Connect.
RSVD	Indicates the pin is not defined and may be used for future products. NC next to this type of signal indicates that this pin must not be routed on the baseboard.

Table 4-2. 5120D (DFF) SKU Pinout

Pin#	Signal	Type	Pin#	Signal	Type
B1	RSVD	NC	A1	PRSNT1_N	GND
B2	RSVD	NC	A2	RSVD	NC
B3	RSVD	NC	A3	RSVD	NC
B4	GND		A4	GND	
B5	SMCLK	O	A5	RSVD	NC
B6	SMDAT	I/O	A6	RSVD	NC
B7	GND		A7	RSVD	NC
B8	+3.3V		A8	RSVD	NC
B9	RSVD	NC	A9	+3.3V	
B10	+3.3V_AUX		A10	+3.3V	
B11	RSVD	NC	A11	PERST_N	O
B12	PROCHOT_N	O	A12	GND	
B13	GND		A13	REFCLK+	O
B14	PETp0	O	A14	REFCLK-	O
B15	PETn0	O	A15	GND	
B16	GND		A16	PERp0	I
B17	RSVD	NC	A17	PERn0	I
B18	GND		A18	GND	
B19	PETp1	O	A19	RSVD	NC
B20	PETn1	O	A20	GND	
B21	GND		A21	PERp1	I
B22	GND		A22	PERn1	I
B23	PETp2	O	A23	GND	
B24	PETn2	O	A24	GND	
B25	GND		A25	PERp2	I
B26	GND		A26	PERn2	I
B27	PETp3	O	A27	GND	
B28	PETn3	O	A28	GND	
B29	GND		A29	PERp3	I
B30	RSVD	NC	A30	PERn3	I
B31	RSVD	NC	A31	GND	
B32	GND		A32	RSVD	NC
B33	PETp4	O	A33	RSVD	NC
B34	PETn4	O	A34	GND	
B35	GND		A35	PERp4	I
B36	GND		A36	PERn4	I
B37	PETp5	O	A37	GND	
B38	PETn5	O	A38	GND	
B39	GND		A39	PERp5	I

Table 4-2. 5120D (DFF) SKU Pinout

B40	GND		A40	PERn5	I
B41	PETp6	O	A41	GND	
B42	PETn6	O	A42	GND	
B43	GND		A43	PERp6	I
B44	GND		A44	PERn6	I
B45	PETp7	O	A45	GND	
B46	PETn7	O	A46	GND	
B47	GND		A47	PERp7	I
B48	RSVD	NC	A48	PERn7	I
B49	GND		A49	GND	
B50	PETp8	O	A50	RSVD	NC
B51	PETn8	O	A51	GND	
B52	GND		A52	PERp8	I
B53	GND		A53	PERn8	I
B54	PETp9	O	A54	GND	
B55	PETn9	O	A55	GND	
B56	GND		A56	PERp9	I
B57	GND		A57	PERn9	I
B58	PETp10	O	A58	GND	
B59	PETn10	O	A59	GND	
B60	GND		A60	PERp10	I
B61	GND		A61	PERn10	I
B62	PETp11	O	A62	GND	
B63	PETn11	O	A63	GND	
B64	GND		A64	PERp11	I
B65	GND		A65	PERn11	I
B66	PETp12	O	A66	GND	
B67	PETn12	O	A67	GND	
B68	GND		A68	PERp12	I
B69	GND		A69	PERn12	I
B70	PETp13	O	A70	GND	
B71	PETn13	O	A71	GND	
B72	GND		A72	PERp13	I
B73	GND		A73	PERn13	I
B74	PETp14	O	A74	GND	
B75	PETn14	O	A75	GND	
B76	GND		A76	PERp14	I
B77	GND		A77	PERn14	I
B78	PETp15	O	A78	GND	
B79	PETn15	O	A79	GND	

Table 4-2. 5120D (DFF) SKU Pinout

B80	GND		A80	PERp15	I
B81	PRSNT2_N	I	A81	PERn15	I
B82	RSVD	NC	A82	GND	
B83	RSVD	NC	A83	RSVD	NC
B84	RSVD	NC	A84	RSVD	NC
B85	RSVD	NC	A85	GND	
B86	RSVD	NC	A86	RSVD	NC
B87	RSVD	NC	A87	RSVD	NC
B88	+12V		A88	GND	
B89	+12V		A89	GND	
B90	+12V		A90	GND	
B91	+12V		A91	GND	
B92	+12V		A92	GND	
B93	+12V		A93	GND	
B94	+12V		A94	GND	
B95	+12V		A95	GND	
B96	+12V		A96	GND	
B97	+12V		A97	GND	
B98	+12V		A98	GND	
B99	+12V		A99	GND	
B100	+12V		A100	GND	
B101	+12V		A101	GND	
B102	+12V		A102	GND	
B103	+12V		A103	GND	
B104	+12V		A104	GND	
B105	+12V		A105	GND	
B106	+12V		A106	GND	
B107	+12V		A107	GND	
B108	+12V		A108	GND	
B109	+12V		A109	GND	
B110	+12V		A110	GND	
B111	+12V		A111	GND	
B112	+12V		A112	GND	
B113	+12V		A113	GND	
B114	+12V		A114	GND	
B115	+12V		A115	GND	

Note: The 5120D SKU does not use the +3.3V_AUX power pins, and baseboard designers have the option to either route these pins or leave them unconnected. The PROCHOT_N pin on the 5120D SKU follows the definition and routing requirements listed in Section 4.1.1

4.3.1 Baseboard Requirements of 5120D

Unlike the Intel® Xeon Phi™ coprocessor PCI Express* card, the 5120D SKU requires the baseboard to implement input filter for the 12V and 3.3V power rails. There are no auxiliary or external power connectors on the 5120D and all power is supplied via the 230 pin edge connector.

- Each 5120D product in the system requires a dedicated input filter for the +12V rail.
- The filtering circuitry should be placed as close to the connector pins as possible.
- The +3.3V power rail input must meet the *PCI Express* CEM* specification.

Table 4-3. 51xxD Power Rail Requirements on Baseboard

Requirement	Value
+12V	
Min V_{IN}	11.04V
Max current	22.192A
LC filter	Maximum rated ripple current per capacitor considering 70% of derating= 2.254A Maximum input current slew rate at input inductor= 0.5A/us Maximum operating temperature for the input capacitors= 105°C. A typical implementation of the input filter would use 5 x 150uF (Sanyo, 16SVPC150M) capacitors and 1 x 0.2uH (Pulse, PG0426.201NL) inductor.
+3.3V	
Decoupling caps	100µf
Max Current	3A

4.3.2 AC Coupling on 5120D Data Pins

Each pin on the PETp/n[15:0] and PERp/n[15:0] buses requires a 0.1µF 0402 AC coupling capacitor on the baseboard. The capacitors should be located for differential pair traces at the same location along the differential traces (that is, they should not be staggered from one trace to the other), and should be as close to each other as possible.

5 Power Specification and Management

Power management on 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	[7120A/D] / 7120P/X 3120A/P [31S1P] Power (Watts)	SE10P/X Power (Watts)	5110P / [5120D] Power[1] (Watts)	Wakeup Time
C0[2]	[270] 300	300	225 [245]	N/A
C1	<115	<115	<115	<1µs
PC3	<50	<50	<45	<75ms
PC6	<30	Not supported	Not supported	<525ms

Notes:

1. Refer Section 5.1 for information on TDP.
2. C0 matches the coprocessor TDP; other power states are specifications.

As the Intel® Xeon Phi™ coprocessor is being powered-on, it is expected to draw measurable amounts 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/SE10X 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

The above power-on measurements were taken with a single coprocessor, 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 that draw up to 245W. This should be taken into account when choosing one of the three modes of operation as listed below:

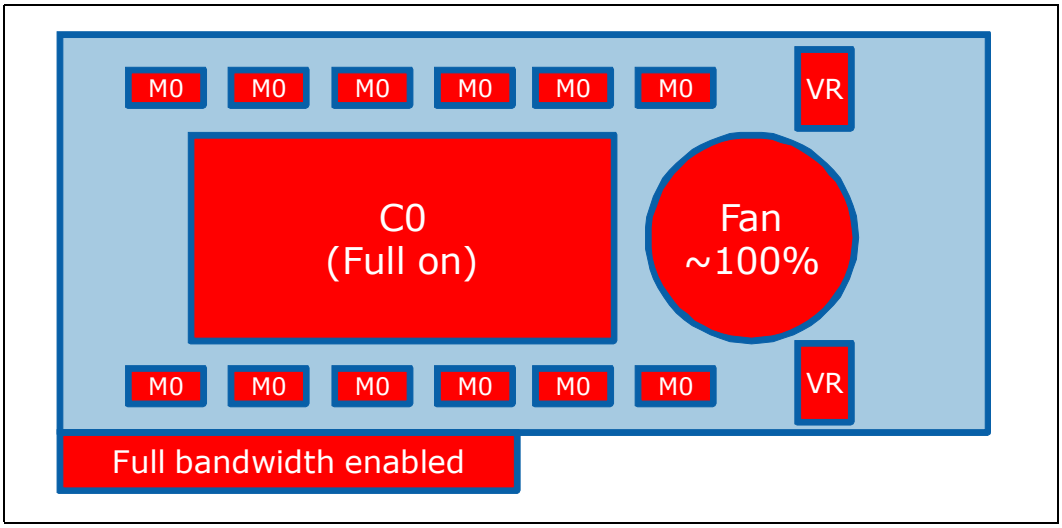
- Users can install both the 2x4 and 2x3 power connectors for total available power of 300W. In this case, the card may draw up to 245W of power depending on the application. This mode ensures sufficient power is available and reduces the risk of throttling. Users may see power dissipation approach 245W, 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).

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 the Intel® Xeon Phi™ coprocessor.

These schematic representations are only for illustrative purposes and do not represent all possible low power states. Cx and Mx refer to coprocessor and memory power states.

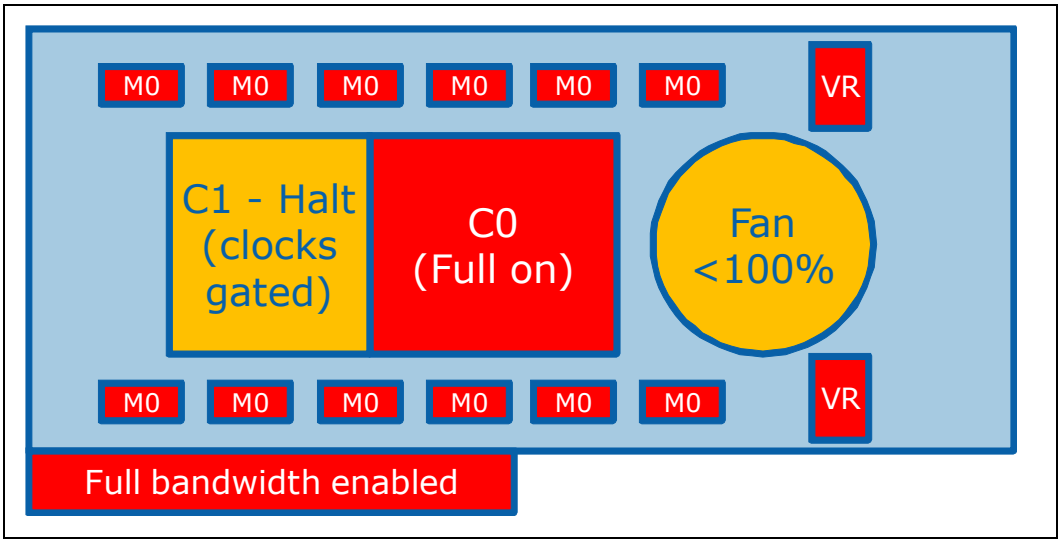
Figure 5-1. Coprocessor in C0-state and Memory in M0-state



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

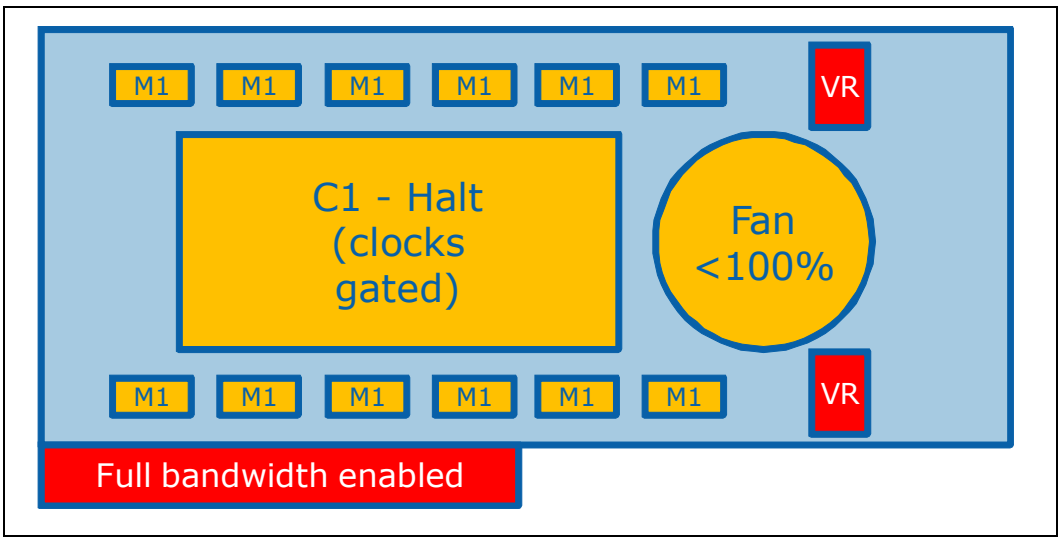
Note: No application is expected to dissipate maximum power from cores and memory simultaneously.

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



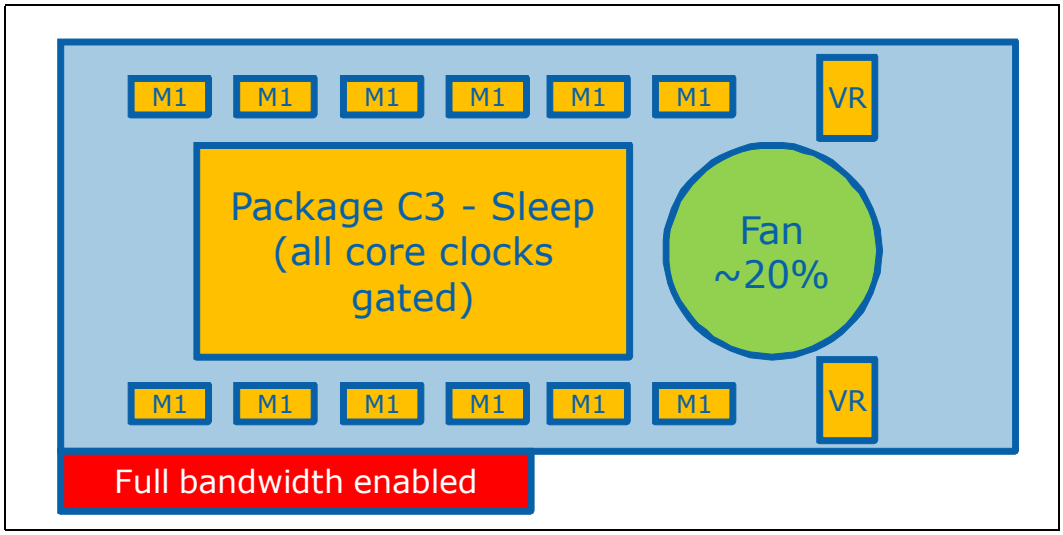
Coprocessor C1 state gates clocks on a core-by-core basis, reducing core power. On the active SKU, the fan slows to an appropriate speed, reducing fan power. If all cores enter C1, the coprocessor automatically enters Auto-pC3 state.

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



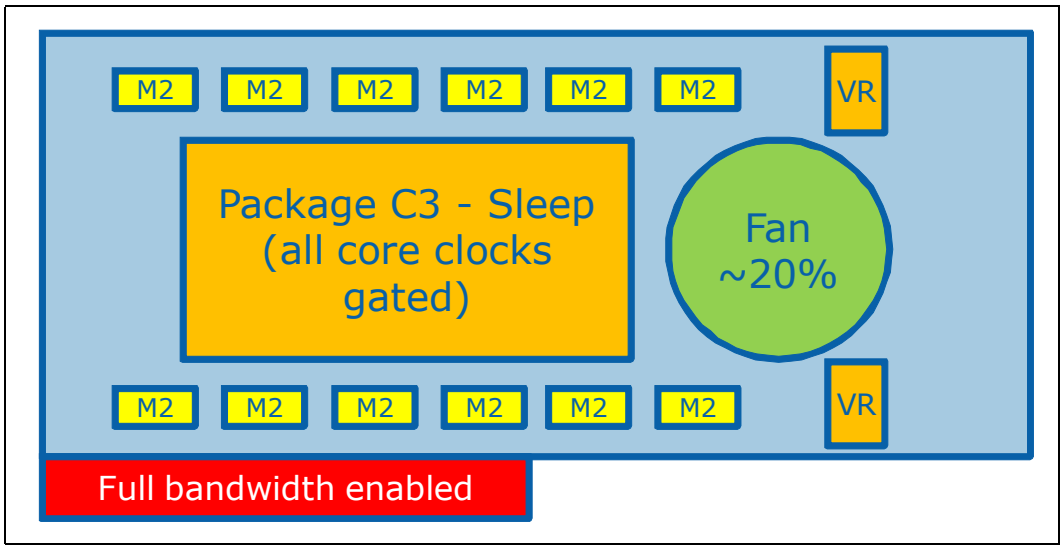
If clock-enable input to memory is pulled high, then memory enters M1 state which reduces memory power.

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



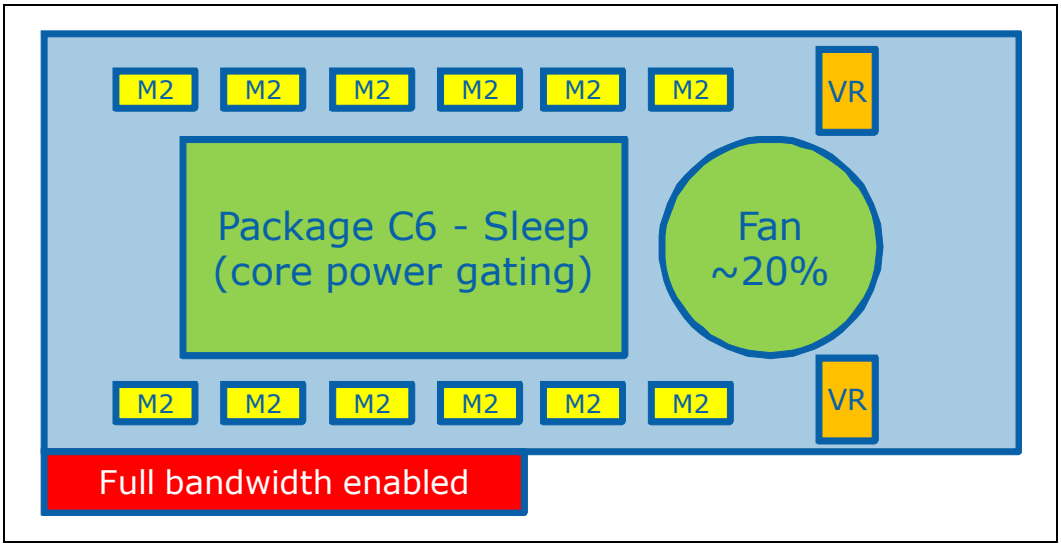
When all cores have entered C1 Halt state, the coprocessor package can reduce the core voltage and enter Deep-pC3. The fan (on active SKUs) can slow to minimum speed. VRs enter low power mode.

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



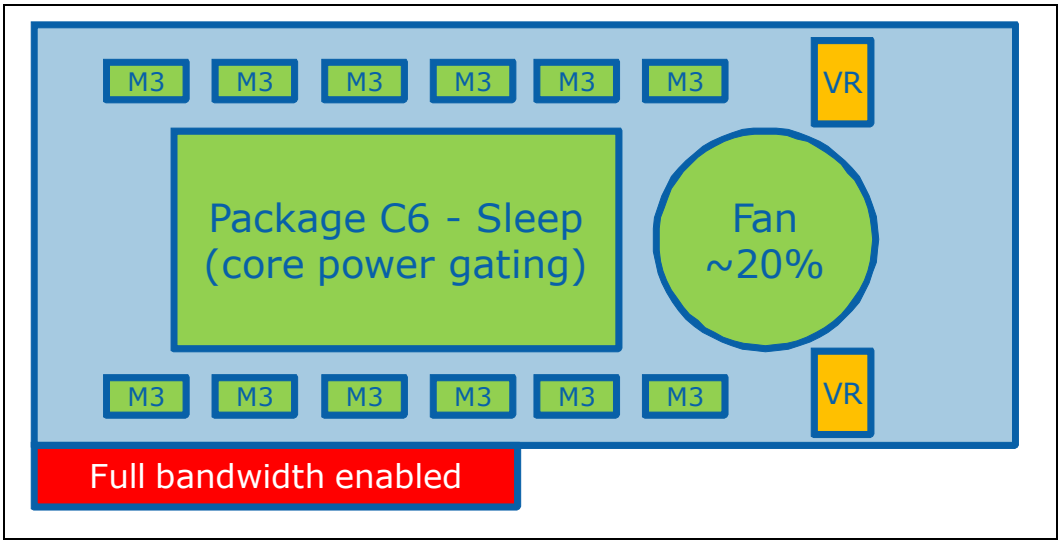
From M1 state, memory can be put in self-refresh mode to enter the M2 state, further reducing memory power.

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.

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.

1. Value may be revised following silicon characterization

Each frequency setting of the coprocessor requires a specific voltage identification (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.

Intel® Xeon Phi™ coprocessor supports Turbo Mode which, at the request of the operating system, will opportunistically and automatically run the coprocessor at a higher frequency than its TDP rated value. When the card is operating below its specified power and temperature limits, the Power Control Unit (PCU) within the card will select the highest possible turbo frequency while still remaining within the power and thermal specifications.

The highest Turbo Mode P-state is P01, followed by sequentially lower frequency states of P02, P03.… on down to the lowest Turbo state of P0n. P-states within the standard frequency range are referred to as P1, P2, P3…. with Pn being the lowest frequency state. Below Pn is one final P-state called LFM or Low Frequency Mode. LFM is only used by the coprocessor when the device is over the PROCHOT trip temperature and is attempting to cool down by reducing power dissipation. LFM reduces the frequency to the absolute lowest possible value, but the voltage will remain the same as P-state Pn. See Figure 5-8. All parts within a given SKU will have the same P-states, but P-states and Turbo frequencies may vary across SKUs.

Once the OS requests turbo operation, by selecting the P01 state, the coprocessor will automatically select the best P0n state that will remain within the specified thermal and power limits. Determination of this P-state is based on the number of active cores, the current draw, the average power consumption and the temperature. If these conditions change, the turbo P-state may also change or even be reduced to the non-turbo P-state of P1. In turbo mode, the coprocessor is free to change the P-state at any time without giving advanced notice to the OS. Although the OS may request P01, there is no guarantee that a turbo frequency will be selected. If the conditions are not sufficient to allow the coprocessor to run above P1, then it will remain in P1. The amount of time the processor can spend in turbo mode may be influenced by the workload and the operating environment.

Turbo mode may be disabled through the SMC Control Panel, or by configuring the operating system such that it never requests the P01 P-state.

Only the 7120A, 7120D, 7120P and 7120X SKUs support Turbo mode.

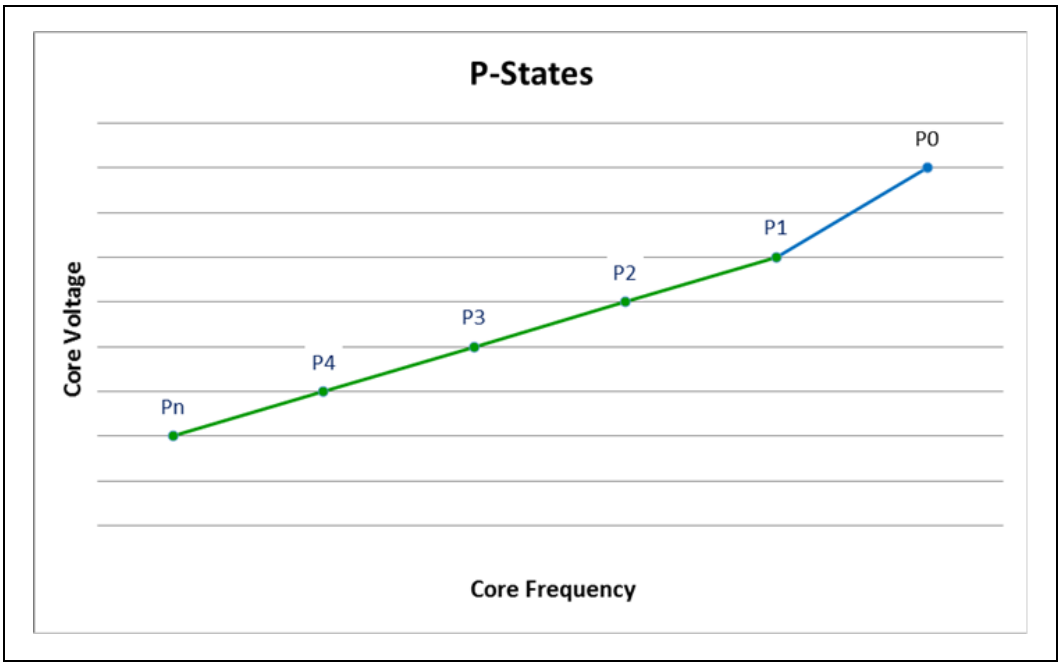


Figure 5-8. Intel® Xeon Phi™ coprocessor P-States and Turbo

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 Symmetric Communications Interface (SCIF) network and the capabilities designed into the coprocessor OS 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 coprocessor OS, but sends the information to the System Management Controller (SMC) using a proprietary protocol. The SMC responds to queries from the platform's BMC using the Intelligent Platform Management Interface (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 Developer's Guide*.

6.2 System Management Controller (SMC)

Intel® Xeon Phi™ coprocessor manageability relies on a SMC on the PCI Express* card. The system 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 PCI Express* card. Card-level power management monitors the card input power and communicates current power conditions to the Intel® Xeon Phi™ coprocessor.

SMC features include:

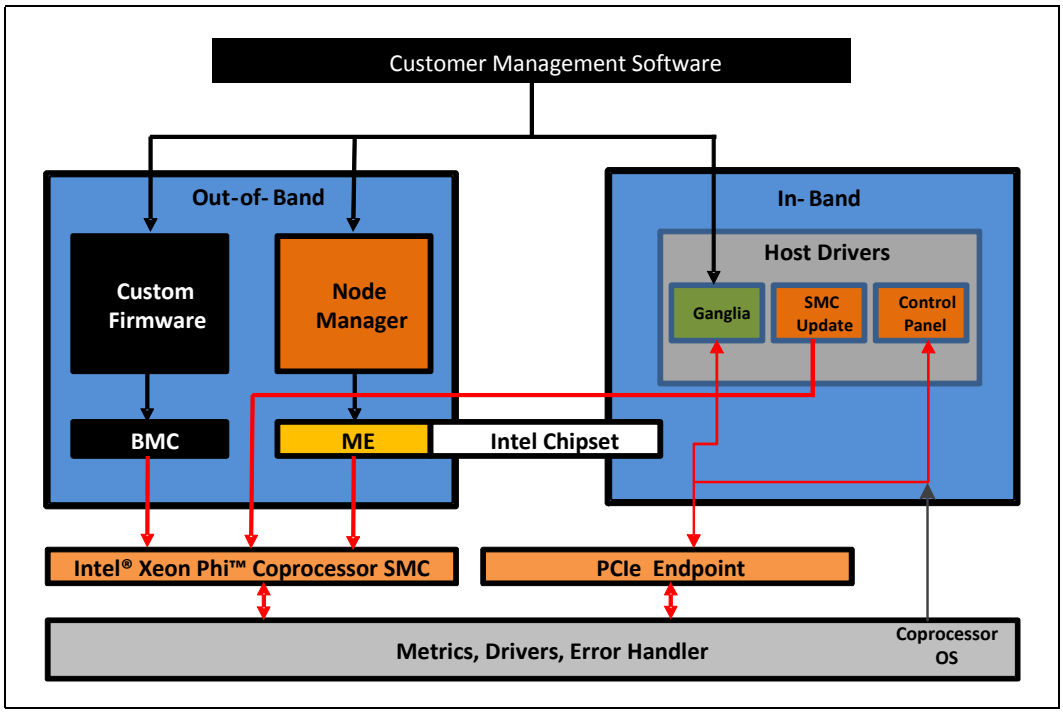
- Four thermal sensor inputs: inlet, outlet, coprocessor die, and GDDR.
- Power alert, thermal throttle, and THERMTRIP# signals.

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 throttle threshold values 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 the PCI Express* card. This information is then used to control the power consumed by the PCI Express* card and the rotating speed of the fan(s) within the PCI Express* card cooling system. 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 throttle threshold throttling vs. over-temperature throttling
- Card-level power throttle threshold/capping
  - Power Throttle Threshold Values 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 the SCIF interface which is part of the MPSS software stack 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 Field Replaceable Unit (FRU) that contains the following information:

- Manufacturer name
- Product name
- Part number / model number
- Universal Unique Identifier (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 SKUs with active cooling solutions. 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. The SMC is the conduit for doing so. 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 in active-high (deasserted) state by the card SMC in the default state, 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.

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 (SCIF)

Manageability, through the SMC, is achievable via the SCIF interface which is part of the MPSS software stack. 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 SCIF interface.

The following SMC information and sensors are accessible over the host-based user mode SCIF interface:

- Hardware strapping pins
- SMC firmware revision number
- UUID
- PCI compliant Memory Mapped Input/Output (MMIO)
- Fan tachometer
- Fan Pulse-Width-Modulation (PWM) to boost 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 throttle threshold (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 SCIF 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 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.

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 adjusts system power using preprogrammed power throttle threshold values. With system-based management, the SMC receives power control inputs via in-band communication from a host application or out-of-band via IPMB commands from a host BMC.

The Intel® Xeon Phi™ coprocessor supports 2 power threshold levels, PL0 and PL1, which determine coprocessor power throttling points. These are not to be confused with setting coprocessor power limits, that is, they do not change the absolute TDP of the product.

PL1 is defined as the first power threshold. When the coprocessor detects that the power consumption stays above PL1 for a specified time period, the coprocessor will begin power throttling. By default, the card's PL1 power threshold is set to 105% of the TDP, and the time duration allowed before throttling starts is 300 ms. When the SMC detects that these conditions have been met it will assert power throttling, causing the frequency to drop by about 100 MHz. Throttling will stop once the power consumption drops 15 W or 20 W (depending on card TDP) below PL1. The difference in the throttling assertion and deassertion thresholds will help prevent the coprocessor from continually cycling between throttling and running normally. For cards with a TDP of 250 W or less, the deassertion point is 15 W lower than PL1. For 300 W cards the deassertion point is 20 W below PL1.

PL0 is normally set to a higher power threshold than PL1. By default, it is set to 125% of TDP, and the time duration allowed at this power level is 50 ms. If these conditions are met, the SMC will use the thermal throttling mechanism to force the coprocessor to the lowest operating frequency which is around 600 MHz. The power reduction from PL0 is expected to be much greater than PL1. The thermal throttling state will continue until the total coprocessor power has dropped 40 W below PL0. When PL0 is exceeded, the initial change in power consumption is a result of the lower operating frequency. The coprocessor will also reduce the CPU core voltage to a value that is appropriate for the lower frequency and this will provide additional power savings. The voltage reduction takes place 3-400 msec after PL0 throttling is asserted.

Note that the PL1, PL0 default thresholds are intended to be percentages of the TDP, and the SMC will dynamically determine actual values for the thresholds during coprocessor boot-up. No user intervention is necessary to enable power threshold throttling. System administrators may program PL1, PL0 thresholds and their respective time durations. The Software Development Kit (SDK) packaged in the coprocessor software stack, or MPSS (http://software.intel.com), contains documentation on programming the power and time registers. The Out-Of-Band mechanism is explained in sections 6.6.3.6.1 and 6.6.3.6.2.

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 Manageability Engine (ME).

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. Due to factors such as noise or traffic on the PCIe SMBus, the address of 0x30 is not guaranteed, and an address in the range of 0x30 to 0x4E may be selected.

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 can be found in the standard MPSS release.

The SMC's PCI Express*/SMBus interface operates as an industry standard 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 hard-coded and not end-user updateable. The SDR can be read in "chunks", suggested size is 16 bytes or the entire SDR can be read passing 'FF' as the number of bytes to read.

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.

The commands supported by the SMC are documented below. The specific information to implement these commands is documented with each command. For byte level details, refer to the *Intelligent Platform Management Bus Communications Protocol Specification, v1.0* and the *Intelligent Platform Management Interface Specification, v2.0*.

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.

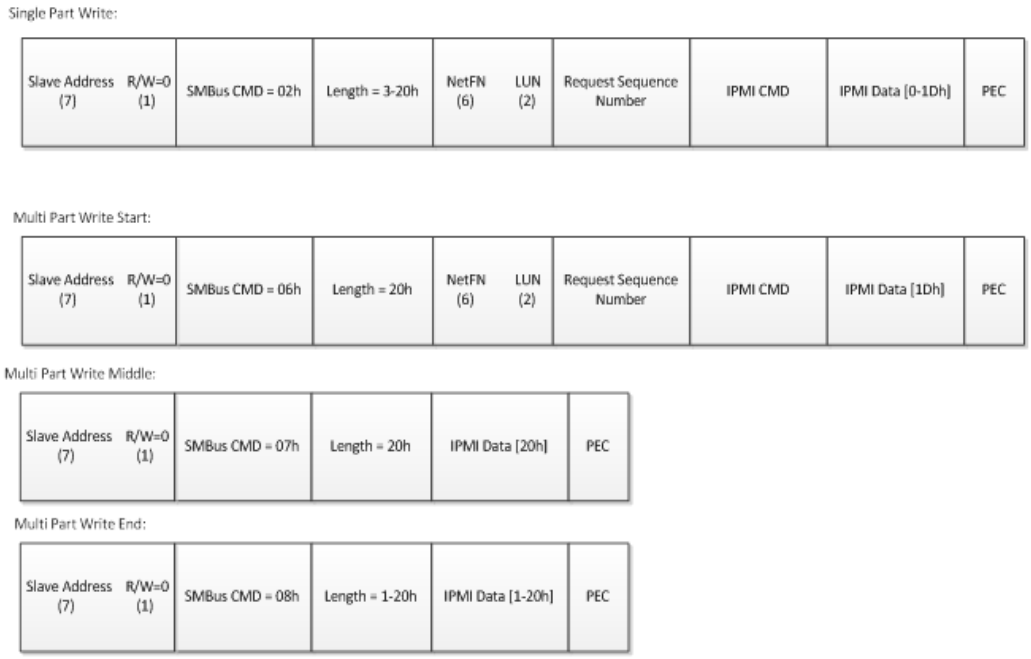
6.6.2.2 SMBus Write and Read Block Command Numbers

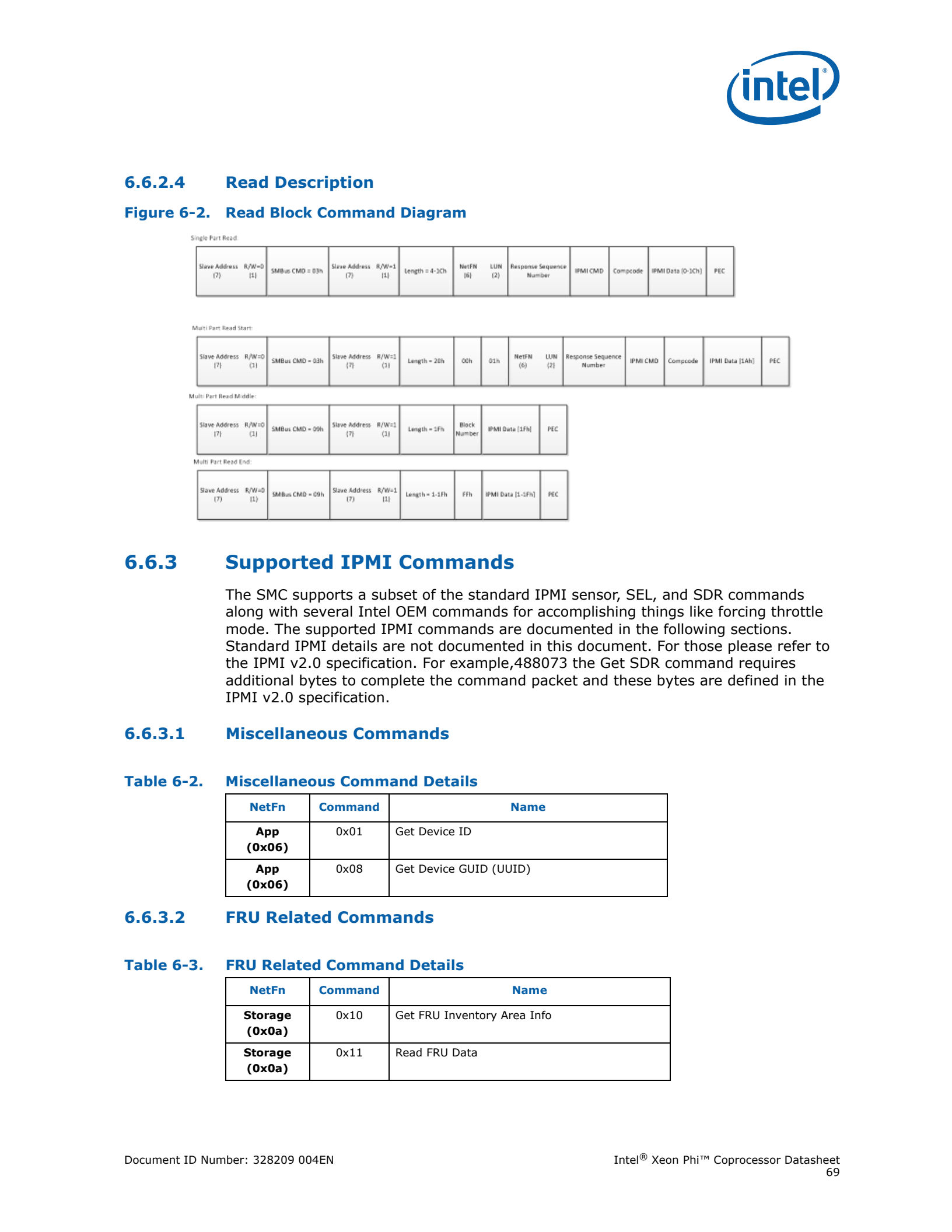
6.6.2.3 Write Description

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 6-1. Write Block Command Diagram





6.6.2.4 Read Description

Figure 6-2. Read Block Command Diagram

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,488073 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: The SDR can be read in "chunks", suggested size is 16 bytes, or the entire SDR can be read by passing 'FF' as the number of bytes to read.

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™ 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 • 3 - Accumulated Energy Status • 11 - Socket Power Throttle Duration • 26 - Package Power Throttle Threshold Value 1 (PL1) • 27 - Package Power Throttle Threshold Value 2 (PL0) • 28 - Package Power SKU A • 29 - Package Power SKU B • 30 - Package Power SKU Unit • 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 • 0x00 - Normal • 0xcc - Invalid field • 0xa1 - Wrong CPU Number • 0xa7 - Wrong Read Length • 0xab - Wrong Command Code • 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*. Figure 2-36 in the EDS shows the format of the data word to effect writing the limits and the time windows. The index referenced below can be correlated to figure 2-36 as SMC-PL0 ==> RAPL-PL2 & SMC-PL1 ==> RAPL-PL1.

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 • 26 - Package Power Throttle Threshold Value 1 (PL1) • 27 - Package Power Throttle Threshold Value 2 (PL0) • All other values reserved
5	0x00	• Parameter LSB •
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 • 0x00 - Normal • 0xc7 - Request Length Invalid • 0xcc - Invalid Field • 0xa1 - Wrong CPU Number • 0xa6 - Wrong Write Length • 0xab - Wrong Command Code • 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 • 1 - Identify • All other values reserved

Table 6-12. Set SM Signal Request Format (Continued)

Byte #	Value	Description
1	0x00	• Instance
2	0x??	• Action • If Signal is 1 • 1 - Assert: Start the identify blink code • 2 - Revert: Return to normal operation • All other values reserved
[3]	0x00	• Value (optional)

Table 6-13. Set SM Signal Response Format

Byte #	Value	Description
0	0x??	• Compcode • 0x00 - Normal • 0xc7 - Request Length Invalid • 0xc9 - Parameter Out of Range • 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
OEM (0x3e)	0x07	• OEM Get Throttle Reason

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 • All other values are reserved.

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

Byte #	Value	Description
0	0x??	• Compcode • 0x00 - Normal • 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 • 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 • 1 - Assert forced throttle • All other values are reserved

Table 6-20. Assert Forced Throttle Response Format

Byte #	Value	Description
0	0x??	• Compcode • 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 (pin B12) allowing the baseboard BMC to directly assert the PROCHOT signal. The baseboard requirements to enable this pin on the baseboard are described in section 4.1.1. The signal to assert emergency throttling via pin B12 is active low on the baseboard and is driven by the BMC. However, the pin must first be enabled by the SMC. This can be accomplished by sending the Enable External Throttle command as described in this section. The pin will need to be enabled each time a reset or power cycle event occurs. Its state is not persistent across these events.

When the baseboard asserts PROCHOT (drives active low signal), the coprocessor OS immediately drops the frequency to lowest rated value (Pn) within 100µs of asserting PROCHOT. If PROCHOT is deasserted in less than 100ms, the coprocessor frequency is restored to the original operational value (either P1 or turbo). if baseboard continues to assert PROCHOT for more than 100ms, the coprocessor OS will respond by reducing the voltage ID (VID) settings to match the lowest frequency, leading to further power savings. Upon subsequent deassertion of PROCHOT, the VID settings are first restored to support operational frequency, followed by the coprocessor frequency itself.

If a baseboard does not support the B12 capability the external throttle signal via pin B12 can be disabled using this command. The card can still be throttled using the SMC by sending the Assert Forced Throttle Command referenced above.

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 • 1 = Enable external throttle signal • All other values are reserved

Table 6-22. Enable External Throttle Response Format

Byte	Value	Description
0	0x??	• Compcode • 0x00 - Normal • 0xc0 - Busy

6.6.3.7.5 OEM Get Throttle Reason

The Get Throttle Reason command returns the state of the three throttle sources, PL1 (RAPL1), PL0 (RAPL2), and the external B12 pin.

Table 6-23. OEM Get Throttle Reason Request Format

Byte	Value	Description
Command	0x07	• OEM Get Throttle Reason
NetFn	0x3e	• NETFN_OEM

Table 6-24. OEM Get Throttle Reason Response Format

Byte	Value	Description
0	0x??	• Compcode • 0x00 - Normal • 0xc0 - Busy
1	0x??	Bitmask of throttle reasons: • 0: RAPL1 (PL1) • 1: RAPL2 (PL0) • 2: External (B12 pin)

6.6.3.8 Other IPMI Related Information

The SMC SEL is a circular log supporting a minimum of 64 log entries. It is resilient to corruption, retaining information across an unexpected power loss.

The sensor names in the IPMI SDR are static and do not change from release to release. The IPMI sensor numbers are not static and may change between releases; hence the sensor number should be discovered during the normal sensor discovery process because additional sensors may be added in the future.

During the normal sensor discovery process, reading the SDR returns the sensors available on the coprocessor. There is a sensor name and sensor number associated with each sensor. Once the sensor number is determined by comparing the sensor name that is desired to be read, the sensor number may be used by the management firmware for reading a particular sensor. It is important that the firmware does not hard code the sensor number as it may change in future releases. It is strongly recommended to use the sensor name if it is "hard-coded" into the management firmware to discover the sensor number, this will ensure the correct sensor is read and will return valid data with future releases of the SMC firmware. Table 6-25 is a list of the current sensor names.

Table 6-25. Table of Sensors

Sensor Type	Sensor Description
POWER	
power_pcie	Power measured at the PCI-e edge fingers input.
power_2x3	Power measured at the 2x3 Aux connector input. (N/A for 5120D)
power_2x4	Power measured at the 2x4 Aux connector input. (N/A for 5120D)
power_pv	Power output reported by the Core VR.
power_vddq	Power output reported by the VDDQ VR.
power_vddg	Power output reported by the VDDG VR.
avg_power0	Average power consumption over Limit Time Window 0.
avg_power1	Average power consumption over Limit Time Window 1.
Instpwr	Instantaneous power consumption reading.
Instpwrmax	Maximum instantaneous power consumption observed.
VOLTAGE	
pv_volt	Voltage reported from the Core VR.
vddq_volt	Voltage reported from the VDDQ VR.
vddg_volt	Voltage reported from the VDDG VR.
TEMP	
east_temp	Temperature sensor on the eastern-most side of the board. (N/A for 5120D)
gddr_temp	Temperature sensor closest to the GDDR memory devices.
west_temp	Temperature sensor on the western-most side of the board. (N/A for 5120D
pv_vrtemp	Temperature reported from the Core VR.
vddq_temp	Temperature reported from the VDDQ VR.
vddg_temp	Temperature reported from the VDDG VR.
proc_temp	Temperature reported by the Intel(R) Xeon Phi(TM) chip DTS
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 (N/A for passive SKUs, 5120D).

Table 6-25. Table of Sensors

Sensor Type	Sensor Description
fan_tach	Fan tach read by SMC (N/A for passive SKUs, 5120D).
OTHER	
status	Critical signal states (described in the datasheet).
Tcritical	Thermal monitoring control value reported by the Intel(R) Xeon Phi(TM) chip
Tcontrol	Fan thermal control value reported by the Intel(R) Xeon Phi(TM) chip

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 6-26. Status Sensor Report Format

Bits	Name	Description
31:7	Reserved	• Reserved
6	P2E_RST	• PCI Express* reset asserted. • Fans boosted.
5	P12V_UVOV	• P12V under-voltage/over-voltage signal asserted. • 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. • Fans boosted and VR output disabled. • This state is latched until power-off.
1	VR1_PHSFLT	• VR1 Phase Fault asserted. • Fans boosted and VR output disabled. • This state is latched until power-off.
0	THERMTRIP	• Coprocessor thermtrip asserted. • Fans boosted and VR output disabled. • 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 6-27. 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. • Initiated by SetSMSignal command.

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