4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids

Data Sheet Volume 1

Rev. 1.0

August 2024

Note: This document contains information on products in the design phase of development. The information here is subject to change without notice. Do not finalize a design with this information.

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.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

This document contains information on products, services and/or processes in development.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

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

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or visit http://www.intel.com/design/literature.htm.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

No computer system can be absolutely secure. Intel does not assume any liability for lost or stolen data or systems or any damages resulting from such losses.

Contents

Figures

Tables

Revision History

Revision	Description	Date
1.0	Initial release	August 2024

Products this Document Applies To

Eagle Stream platform with 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processors and Intel® C740 Series Chipset PCH.

1.0 Introduction

This document provides functional descriptions of the 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor architecture.

NOTE

Features within this document may not be available on all platform segments, processor types, or processor SKUs.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids is the leadership wave 3 10-nm Intel® Xeon® multi-core server processor. The processor supports up to 52 bits of physical address space and 57 bits of virtual address space. Its design works for a platform that consists of at least one 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor and an Intel® C740 Series Chipset PCH. Including an Integrated Memory Controller (IMC) and an Integrated I/O (IIO) on a single silicon die in this family of processors.

The processor supports up to 80 lanes of PCI Express* 5.0 links capable of 32 GT/s, and eight lanes of DMI. It features four IMC, each IMC supporting up to two channels of DDR5 DIMMs with up to two DIMMs per channel.

1.1 Structure and Scope

The following lists summarize the structure and scope of each specification.

Architecture Specification

- Introduction
- Overview, Features, and Topologies
- Core Functional Description
- Integrated I/O and PCI Express Functional Description
- Intel® Ultra Path Interconnect Version 2.0 (Intel® UPI 2.0) Functional Description
- Integrated Memory Controller (IMC) Functional Description
- Intel® Quick Assist Technology (Intel® QAT)
- Power Controller Unit (PCU) Functional Description
- Reliability, Availability, Serviceability, and Manageability (RAS) Functional Description

Register Specification

- Configuration Process and Registers Overview
- Configuration Space Registers (CSR)
- Model Specific Registers (MSR)

Electrical Specification

- Introduction
- Electrical Specifications
- Processor Land Listing
- Processor Signal Descriptions

1.2 Related Publications

See the following documents for additional information.

Table 1. Public Publications

Document	Document Number/Location
Advanced Configuration and Power Interface Specification 4.0	http://www.acpi.info/
PCI Local Bus Specification 3.0	http://www.pcisig.com/
PCI Express Base Specification, Revision 5.0	http://www.pcisig.com/
PCI Express Base Specification, Revision 4.0	http://www.pcisig.com/
PCI Express Base Specification, Revision 3.0	http://www.pcisig.com/
PCI Express Base Specification, Revision 2.1	http://www.pcisig.com/
PCI Express Base Specification, Revision 1.1	http://www.pcisig.com/
PCIe* Gen 3 Connector High Speed Electrical Test Procedure	325028-001 /http://www.intel.com/content/www/us/en/io/pci-express/pci-express-architecture-devnet-resources.html
DDR5 SDRAM Specification and Register Specification	http://www.jedec.org/
Intel® 64 and IA-32 Architectures Software Developer's Manuals Volume 1: Basic Architecture Volume 2A: Instruction Set Reference, A-M Volume 2B: Instruction Set Reference, N-Z Volume 3A: System Programming Guide Volume 3B: System Programming Guide Intel® 64 and IA-32 Architectures Optimization Reference Manual	325462 / http://www.intel.com/products/processor/manuals/index.htm
Intel® Virtualization Technology Specification for Directed I/O Architecture Specification	http://www.intel.com/content/www/us/en/intelligent-systems/intel-technology/vt-directed-io-spec.html
Intel® Trusted Execution Technology Software Development Guide	http://www.intel.com/technology/security/
Intel® Data Streaming Accelerator Technology Specification	341204-002 / https://software.intel.com/content/www/us/en/develop/articles/intel-data-streaming-accelerator-architecture-specification.html
Intel 4th Gen Intel Xeon Processor Scalable Family (Code name Sapphire Rapids) Data Sheet Vol 2- Register	TBD
Eagle Stream Platform Electrical Datasheet	TBD

1.3 Terminology

Term	Description
ACS	Access Control System
ADI	Assignable Device Interface
AER	Advanced Error Reporting
AES	Advanced Encryption Standard
ARI	Alternative Routing-ID Interpretation
ASPM	Active State Power Management
ATS	Address Translation Services
BE	Byte Enable
BMC	Baseboard Management Controller
CA	Coherency Agent. In some cases this is referred to as a Caching Agent though a CA is not actually required to have a cache. It is a term used for the internal logic providing mesh interface to LLC and Core. The CA is a functional unit in the CHA.
CA	Completer Abort (PCIe context)
CAP	CSR Access Proxy
CHA	The functional module that includes the CA (Coherency Agent) and HA (Home Agent).
CXPSMB	Converged eXPress SMBus Controller
DDR5	Fifth generation Double Data Rate SDRAM Memory technology.
DLLP	Data Link Layer Packet
DMA	Direct Memory Access
DMI3	Direct Media Interface Gen3 operating at PCI Express 3.0 speed.
DTLB	Data Translation Look-aside Buffer. Part of the processor core architecture.
DTS	Digital Thermal Sensor
ECC	Error Correction Code
Enhanced Intel® SpeedStep Technology	Allows the operating system to reduce power consumption when performance is not needed.
ETR	Eagle Tail Rings. Former naming conventions for WQM and WQM Rings
Execute Disable Bit	The Execute Disable bit allows memory to be marked as executable or non-executable, when combined with a supporting operating system. If code attempts to run in non-executable memory the processor raises an error to the operating system. This feature can prevent some classes of viruses or worms that exploit buffer overrun vulnerabilities and can thus help improve the overall security of the system. See the Intel® 64 and IA-32 Architectures Software Developer's Manuals for more detailed information.
F0	Function 0 - Intel® ME/Intel® QAT Cluster
FC	Flow Control
FCC	Flow Control Credits
FLIT	Flow Control Unit. The Intel® UPI Link layer's unit of transfer. A FLIT is made of multiple PHITS. A Flit is always a fixed amount of information (192 bits).
FLR	Function Level Reset

continued...

Term	Description
Functional Operation	Refers to the normal operating conditions in which all processor specifications, including DC, AC, system bus, signal quality, mechanical, and thermal, are satisfied.
GSSE	Extension of the SSE/SSE2 (Streaming SIMD Extensions) floating point instruction set to 256b operands.
HA	A Home Agent (HA) orders read and write requests to a piece of coherent memory. The HA is implemented in the CHA logic.
ICU	Instruction Cache Unit. Part of the processor core architecture.
IFU	Instruction Fetch Unit. Part of the processor core.
IIO	Integrated I/O Controller. An I/O controller that is integrated in the processor die. The IIO consists of the DMI3 module and PCIe modules.
IMC	Integrated Memory Controller. A Memory Controller that is integrated in the processor die.
IMR	Isolated Memory Region. Part of the System Memory
IMS	Interrupt Message Storage. Stores MSI-X vectors for Intel® Scalable IOV
Intel® AVX	Intel® Advanced Vector Extensions (Intel® AVX) promotes legacy 128-bit SIMD instruction sets that operate on XMM register set to use a “vector extension” (VEX) prefix and operates on 256-bit vector registers (YMM).
Intel® AVX-512	The base of the 512-bit SIMD instruction extensions are referred to as Intel® AVX-512 foundation instructions. They include extensions of the Intel® AVX family of SIMD instructions but are encoded using a new encoding scheme with support for 512-bit vector registers, up to 32 vector registers in 64-bit mode, and conditional processing using opmask registers.
Intel® ME	Intel® Management Engine. The processor uses Intel® ME 11 for 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids Workstation 1S, 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids Workstation 2S and 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids HEDT.
Intel® QuickData Technology	Intel® QuickData Technology is a platform solution designed to maximize the throughput of server data traffic across a broader range of configurations and server environments to achieve faster, scalable, and more reliable I/O.
Intel® Ultra Path Interconnect (Intel® UPI2.0)	A cache-coherent, link-based Interconnect specification for Intel® processors. Also known as Intel® UPI.
Intel® 64 Technology	64-bit memory extensions to the IA-32 architecture. Further details on Intel® 64 architecture and programming model can be found at http://developer.intel.com/technology/intel64/.
Intel® Scalable IOV	Intel® Scalable I/O Virtualization
Intel® SPS FW	Intel® Server Platform Services Firmware. The processor uses Intel® SPS FW in server configurations.
Intel® Turbo Boost Technology	A feature that opportunistically enables the processor cores to run at a faster frequency. This results in increased performance of both single and multi-threaded applications.
Intel® TXT	Intel® Trusted Execution Technology
Intel® UPI 2.0	Intel® Ultra Path Interconnect (Intel® UPI) Agent. An internal logic block providing interface between internal mesh and external Intel® UPI.
Intel® Virtualization Technology (Intel® VT)	Processor Virtualization which when used in conjunction with Virtual Machine Monitor software enables multiple, robust independent software environments inside a single platform.
Intel® VT-d	Intel® Virtualization Technology (Intel® VT) for Directed I/O. Intel® VT-d is a hardware assist, under system software (Virtual Machine Manager or OS) control, for enabling I/O device Virtualization. Intel® VT-d also brings robust security by providing protection from errant DMAs by using DMA remapping, a key feature of Intel® VT-d.

continued...

Term	Description
Integrated Heat Spreader (IHS)	A component of the processor package used to enhance the thermal performance of the package. Component thermal solutions interface with the processor at the IHS surface.
IOMMU	I/O Memory Management Unit
IOV	I/O Virtualization
IOVM	I/O Virtual Machine
IVR	Integrated Voltage Regulation (IVR): The processor supports several integrated voltage regulators.
KPT	Intel® Key Protection Technology
LLC	Last Level Cache
LRDIMM	Load Reduced Dual In-line Memory Module
LRU	Least Recently Used. A term used in conjunction with cache allocation policy.
M2M	Mesh to Memory. Logic in the IMC which interfaces the IMC to the mesh.
M2PCIe	The logic in the IIO modules which interface the modules to the mesh.
MESH	The on die interconnect which connects modules in the processor.
MLC	Mid Level Cache
MPS	Maximum Payload Size
MRdLk	Memory Read Lock
MRRS	Maximum Read Request Size
MSI	Message Signal Interrupt
NCTF	Non-Critical to Function: NCTF locations are typically redundant ground or non-critical reserved, so the loss of the solder joint continuity at end of life conditions will not affect the overall product functionality.
NID \ NodeID	Node ID (NID) or NodeID (NID). The processor implements up to 4- bits of NodeID (NID).
Non-Posted Request	A non-posted request is a request which cannot be considered ordered (per PCI rules) until after the completion is received. Non-posted transactions include all reads and some writes (configuration writes).
OOBMSM	Out-of-Band Management Service Module
Pcode	Pcode is microcode which is run on the dedicated microcontroller within the PCU.
PCH	Platform Controller Hub. A chipset with centralized platform capabilities including the main I/O interfaces along with display connectivity, audio features, power management, manageability, security and storage features.
PCU	Power Control Unit.
PCIe	PCI Express
PE	Processing Element: A processing element is a generic term indicating a CPU thread.
PECI	Platform Environment Control Interface
PF	Physical Function (SR-IOV)
Phit	The data transfer unit on Intel® UPI at the Physical layer is called a phit (physical unit). A Phit will be either 20 bits, or 8 bits depending on the number of active lanes.
PME	Power Management Event
Posted Request	A posted request is a request which can be considered ordered (per PCI rules) upon the issue of the request and therefore completions are unnecessary. Example of posted transaction are PCIe memory writes and messages
Processor	Includes the 64-bit cores, uncore, I/Os and package

continued...

Term	Description
Processor Core	The term “processor core” refers to Si die itself which can contain multiple execution cores. Each execution core has an instruction cache and data cache and MLC cache. All execution cores share the L3 cache.
Rank	A unit of DRAM corresponding four to eight devices in parallel, ignoring ECC. These devices are usually, but not always, mounted on a single side of a DDR5 DIMM.
RC	Root Complex. Consists of a host bridge, PCIe port (root ports) and one or more RC integrated endpoints
RCEC	Root Complex Event Collector, collects errors from PCIe RCiEPs, as defined in PCI Express Base Specification
RCiEP	Root Complex Integrated Endpoint
RCRB	Root Complex Register Block as defined in PCI Express Base Specification
RDIMM \ LRDIMM	Registered Dual In-line Memory Module \ Load Reduced DIMM
RID	Ring ID
RLT	Rate Limiter
RP	Root Port. A PCIe port on a root complex
RTID	Request Transaction IDs are credits issued by the CHA to track outstanding transaction, and the RTIDs allocated to a CHA are topology dependent.
S3M	Secure Startup Services Module
SCI	System Control Interrupt. Used in ACPI protocol.
SKU	Stock Keeping Unit (SKU) is a subset of a processor type with specific features, electrical, power and thermal specifications. Not all features are supported on all SKUs. A SKU is based on specific use condition assumption.
SoC	System on Chip
SR-IOV	Single Root IO Virtualization
SSE	Intel® Streaming SIMD Extensions (Intel® SSE)
SMBus	System Management Bus. A two-wire interface through which simple system and power management related devices can communicate with the rest of the system.
SDU	SMM Driver Update
Storage Conditions	A non-operational state. The processor may be installed in a platform, in a tray, or loose. Processors may be sealed in packaging or exposed to free air. Under these conditions, processor landings should not be connected to any supply voltages, have any I/Os biased or receive any clocks. Upon exposure to “free air” (that is, unsealed packaging or a device removed from packaging material) the processor must be handled in accordance with moisture sensitivity labeling (MSL) as indicated on the packaging material.
TAC	Thermal Averaging Constant
TDP	Thermal Design Power
TLP	Transaction Layer Packet
TSOD	Temperature Sensor On DIMM
Tillamook River	The on package component included on 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids with Fabric which provided the second generation of Intel® Omni-Path fabric.
UC	Uncacheable
Uncore	The portion of the processor comprising the shared LLC cache, CHA, IMC, PCU, Ubox, IIO and Intel® UPI modules.

continued...

Term	Description
Unit Interval	Signaling convention that is binary and unidirectional. In this binary signaling, one bit is sent for every edge of the forwarded clock, whether it be a rising edge or a falling edge. If a number of edges are collected at instances t_1, t_2, t_n,...., t_k then the UI at instance "n" is defined as: $UI_n = t_n - t_{n-1}$
UR	Unsupported Request
Volume Management Device (VMD)	Volume Management Device (VMD) is a new technology used to improve PCIe management. VMD maps the PCIe configuration space for child devices/adapters for a particular PCIe x16 module into its own address space, controlled by a VMD driver.
VMM	Virtual Machine Monitor
VRP	Virtual Root Port
VSP	Virtual Switch Port
VF	Virtual Function (SR-IOV)
V_{CCIN}	Primary voltage input to the voltage regulators integrated into the processor.
V_{SS}	Processor ground
V_{CCSA}	System agent supply for Intel® UPI and IIO
V_{CCIO}	IO voltage supply input
x1, x4, x8, x16	Refers to a Link or Port with one, four, eight or sixteen Physical Lanes

2.0 Overview, Features, and Topologies

This section provides an overview of the modules and architectural features of the processor.

The following figures show an example of the 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor. The processor composes of multiple I/O interface and CHA/core modules that are connected to a mesh interconnect. The mesh consists of horizontal and vertical interconnects, each interconnect consisting of bidirectional channels. Different SKUs of the processor may have a different combination of modules. The following figures are for illustration only.

Figure 1. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids XCC Architecture Example

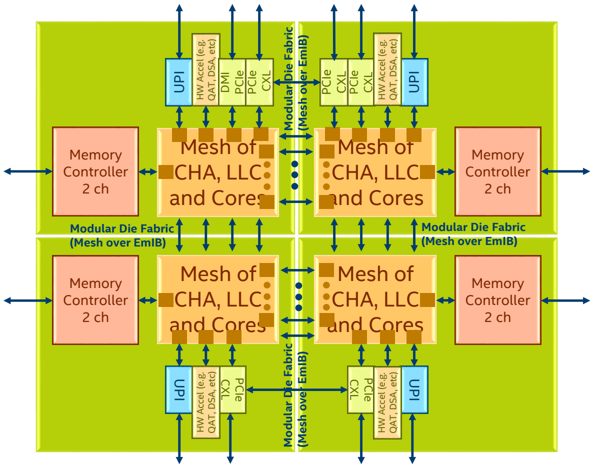
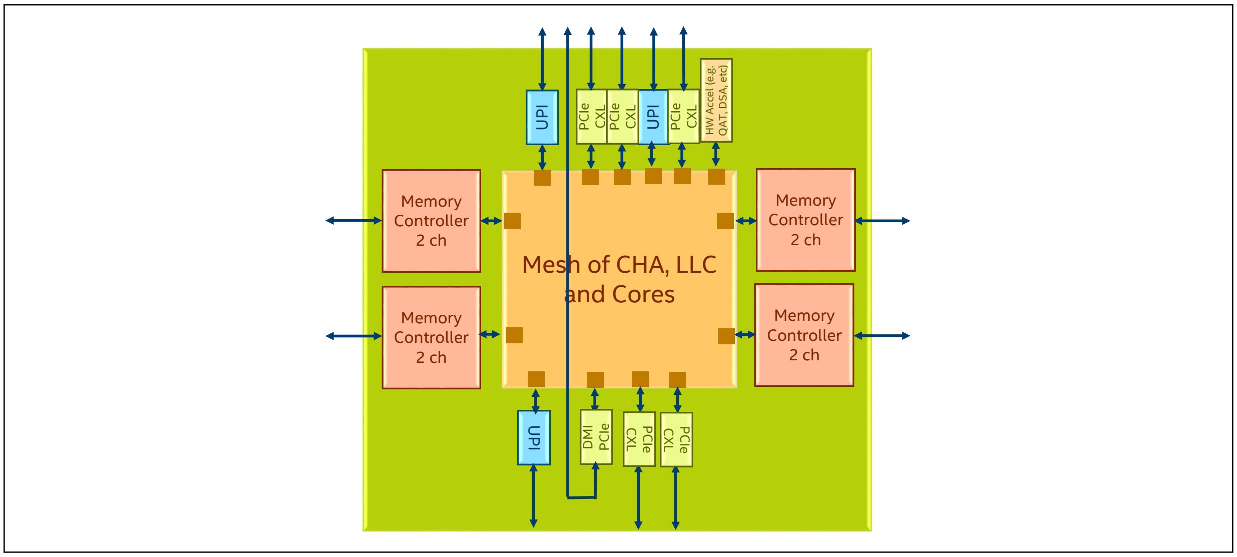

UPI
HW Accel (e.g. QAT, DSA, etc)
DMI
PCIe
PCIe
CXL
Modular Die Fabric (Mesh over EmIB)
PCIe
CXL
PCIe
CXL
HW Accel (e.g. QAT, DSA, etc)
UPI
Memory Controller 2 ch
Mesh of CHA, LLC and Cores
Mesh of CHA, LLC and Cores
Memory Controller 2 ch
Modular Die Fabric (Mesh over EmIB)
Modular Die Fabric (Mesh over EmIB)
Memory Controller 2 ch
Mesh of CHA, LLC and Cores
Mesh of CHA, LLC and Cores
Memory Controller 2 ch
UPI
HW Accel (e.g. QAT, DSA, etc)
CXL
PCIe
Modular Die Fabric (Mesh over EmIB)
CXL
PCIe
HW Accel (e.g. QAT, DSA, etc)
UPI

Figure 2. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids MCC Architecture Example



2.1 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids/5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids Processor Features and Topologies

This section covers the following server platform configurations:

Figure 3. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids XCC

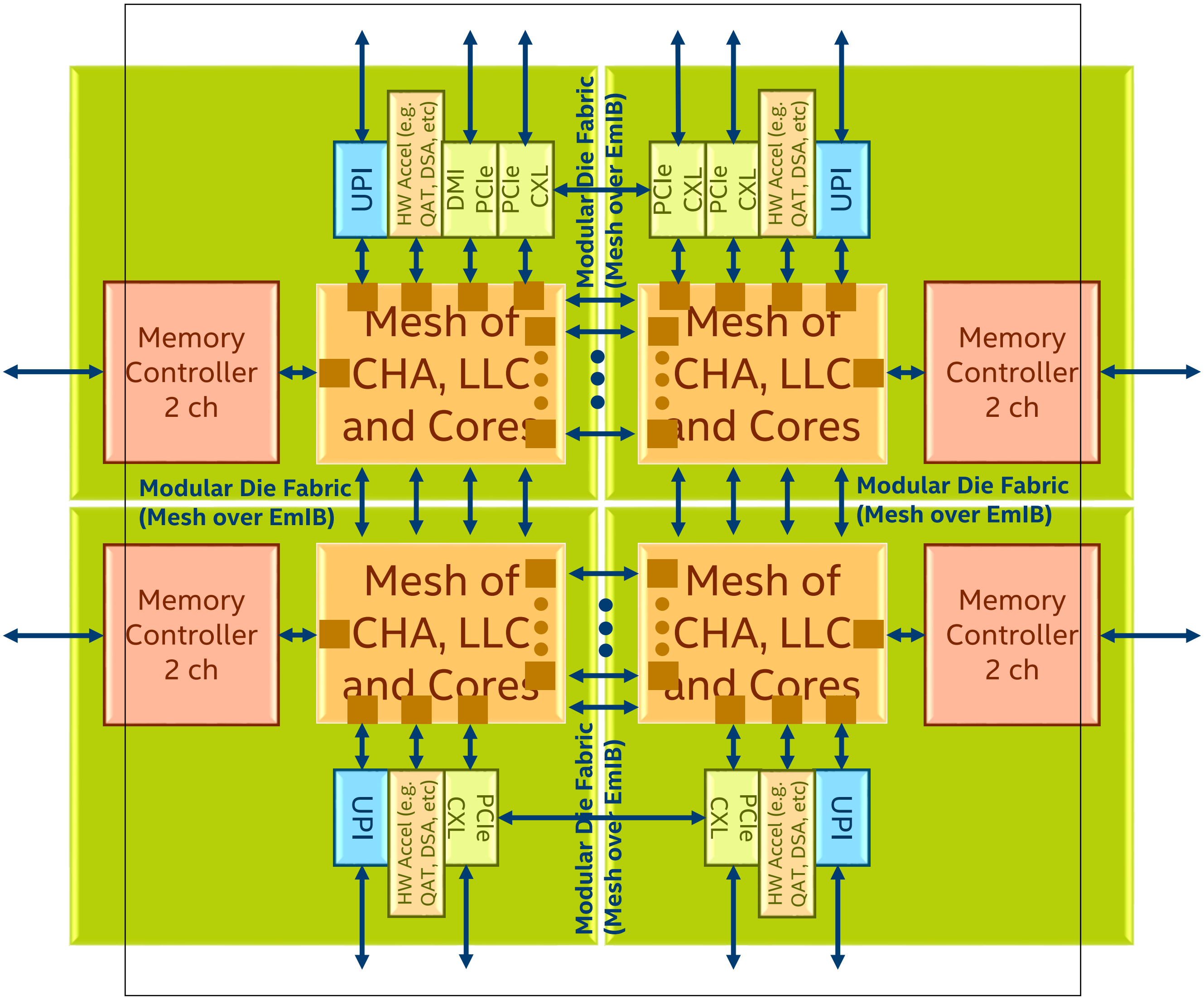
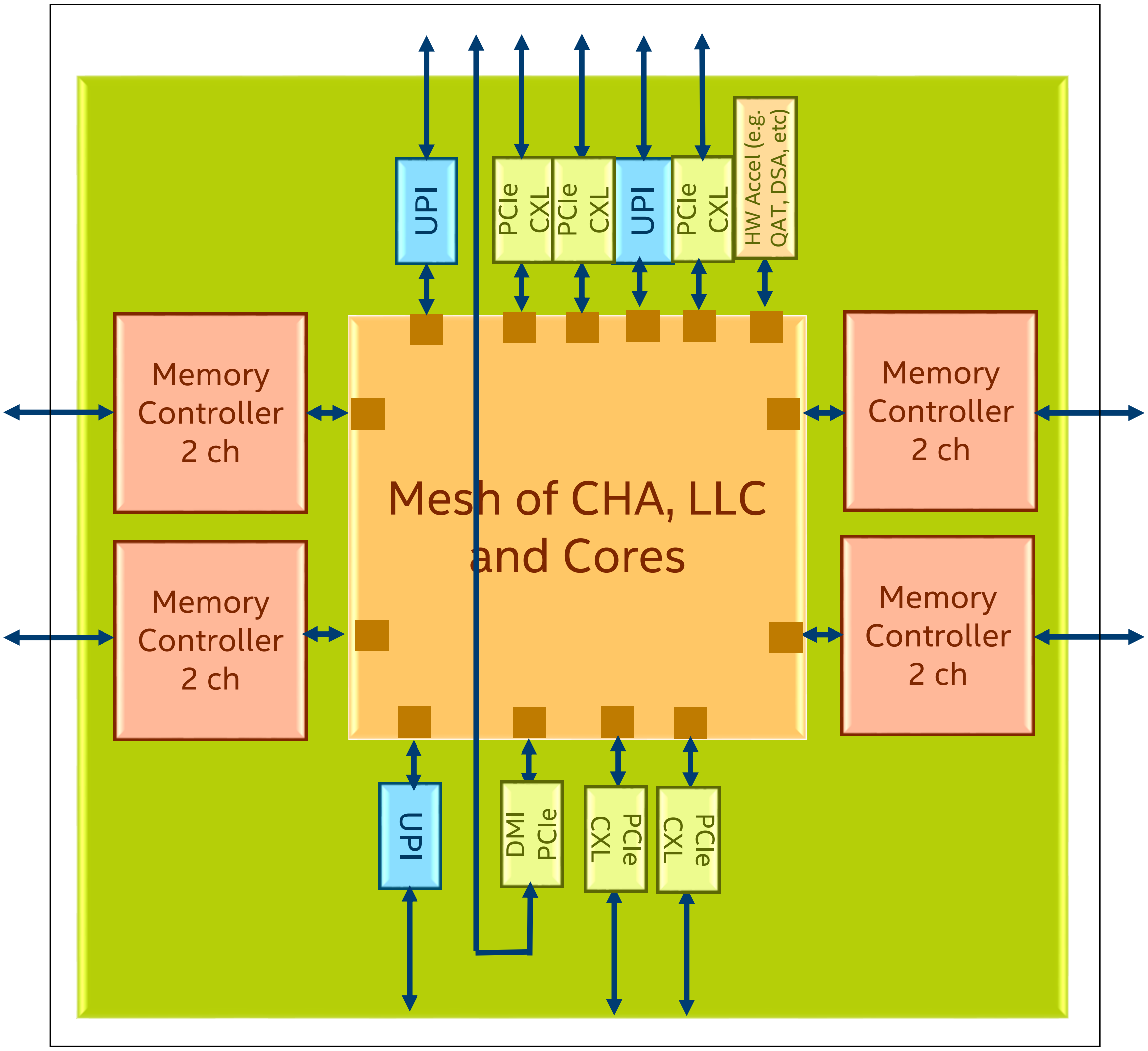

UPI
HW Accel (e.g. QAT, DSA, etc)
DMI
PCIe
PCIe
CXL
Modular Die Fabric (Mesh over EmIB)
PCIe
CXL
PCIe
CXL
HW Accel (e.g. QAT, DSA, etc)
UPI
Memory Controller 2 ch
Mesh of CHA, LLC and Cores
Mesh of CHA, LLC and Cores
Memory Controller 2 ch
Modular Die Fabric (Mesh over EmIB)
Modular Die Fabric (Mesh over EmIB)
Memory Controller 2 ch
Mesh of CHA, LLC and Cores
Mesh of CHA, LLC and Cores
Memory Controller 2 ch
UPI
HW Accel (e.g. QAT, DSA, etc)
CXL
PCIe
Modular Die Fabric (Mesh over EmIB)
CXL
PCIe
HW Accel (e.g. QAT, DSA, etc)
UPI

Figure 4. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids MCC



- 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids with HBM

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids offers various features among its processor types and SKUs to meet the needs of multiple market segments. Features within this document may not be available on all platform segments, processor types, or processor SKUs.

Table 2. Server Features

Features	Comments[1]
Platform	Eagle Stream
Core Count (Up to)	60 Note: Core count varies per processor type and SKU. Value provided is highest core count SKU.

continued...

Features	Comments[1]
Scalability	1S, 2S, 4S and glueless 8S (>8S via xNC support)
Thermal Design Power (TDP)	Varies by processor type and SKU. Refer to the *Eagle Stream Server and Fishhawk Falls Workstation Platforms Thermal Mechanical Specification Design Guidelines (TMSDG)* for details.
OEM External Node Controller Support (XNC)	Supported on select SKUs[2]
Intel® UPI_v2.0 Speeds	12.8 GT/s, 14.4 GT/s, 16 GT/s . Up to 4 links per CPU.
Peripheral Controller Hub (PCH)	Intel® C740 Series Chipsets
Last Level Cache (LLC) per CHA	1.875 MB (non-inclusive with the 2.0 MB Mid Level Cache)
Max Number of PCIe* Lanes	• 80 total PCIe Lanes • 48 lanes North/ 32 lanes South of socket • Five Gen 5 capable x16 PCIe ports • Each port can be subdivided as 2 x8, 4 x4, 8 x2 or any combination there of *Note:* When subdivided to x2, the PCIe port operates up to Gen4 speed.
Memory: DDR5 Technology Support	DDR5 ECC RDIMM 3DS DDR5 ECC RDIMM
Memory: Speeds[2] (Up to)	4400, 4800
Memory: Number of IMC per socket	4
Memory: Channels per IMC (Up to)	2
Memory: Max DIMM per Channel (Up to)	2
Memory: Max DIMMs per Socket	16
Memory: Max Logical Ranks / DDR5 Channel	RDIMM - 4 3DS RDIMM - 16
Memory: DRAM Density	16,24,32 Gb
Memory: Max Capacity[3] (x16 256GB 16Gb 3DS DDR5) per Socket	4 TB
Memory: Max Capacity (x8 256 GB 16 Gb 3DS DDR5) and (x8 512 GB Intel® Optane™ Persistent Memory 300 Series mapped as memory) per Socket	6 TB Note: Requires a SKU which supports Intel® Optane™ Persistent Memory 300 Series
Intel® Advanced Vector Extensions 512 (Intel® AVX-512)	Supported
Intel® AVX-512 with 2nd FMA	Supported on select SKUs
Volume Management Device (VMD)	Supported
RAS	RAS supported by DDR5 memory IEH 2.0
Processor Information ROM (PIROM)	Supported
Non-transparent Bridge Application (NTB)	Supported
Isochronous Application	Not supported
System State S3	Not supported
Multi PCH (Use of multiple PCHs where the PCHs are connected to the system via DMI in order to make use of core and IO PCH functionality in partitionable systems)	Supported on select SKUs[2]

continued...

Features	Comments[1]
Endpoint PCH (One or more Intel® C740 Series Chipsets are used only for the Intel® Quick Assist features. Intel® C740 Series Chipsets are connected only via PCIe and DMI is NOT connected nor used.)	Supported
Trusted Platform Module (TPM)	TPM 2.0
Intel® Platform Trust Technology (Intel® PTT)	Supported
Manageability Engine	Castle Crest
Overclocking	Not Supported
BMCINIT Mode	Yes

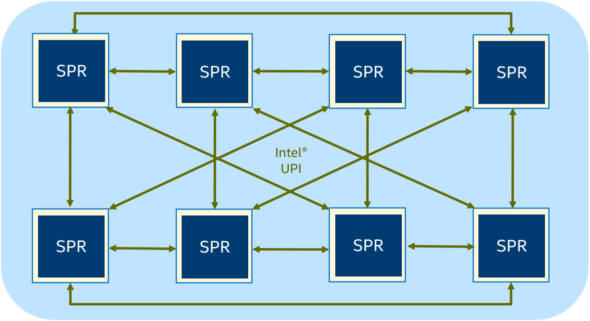
NOTES:

1. Where multiple options are listed feature support varies per SKU, or listed as "Supported on select SKUs" not all SKUs support this feature. A SKU consists of a defined set of features.
2. Contact your Intel representative for further details.
3. Max memory capacity is SKU dependent.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids 1S, 2S, 4S and 8S Configurations

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor enables 1 socket, 2 socket, 4 socket and 8 socket platforms. All Intel® UPI_v2.0 links are used for increased bandwidth between the processors. The following figure show POR topology for 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids on Eagle Stream platforms. The following figure shows an example of the topologies for 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids on Eagle Stream platforms. See the latest *Eagle Stream Platform Design Guide* for the technical details on the supported topologies.

Figure 5. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids Server 8SG



2.1.1 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids and 5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids Processor Overview

This section summarizes the key new feature and technologies in 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids and 5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids processor.

General Purpose performance - increase performance across all market segments relative to 3rd Gen Intel® Xeon® Processor Scalable Family, Codename Cooper Lake, processor baseline.

- Up to eight channels of DDR5.
- Up to four 24-lane (x24) bi-directional Intel® UPI 2.0 ports with speeds of 12.8 GT/s, 14.4 GT/s, and 16 GT/s.

Memory Innovation - Flexible memory subsystem for emerging workloads

- DDR5 - efficient bandwidth, lower latency

Storage and I/O Performance - Platform I/O bandwidth and RAS for storage

- >60% aggregate bandwidth increase
- 16 GT/s PCIe 4.0 on PCH.
- 32 GT/s PCIe 5.0 and Compute Express Link* (CXL*) on CPU via Flex Bus
- x2 PCIe link subdivision capability
- Share Virtual Memory Support
- Intel® Scalable IOV

RAS Capabilities - Increased availability at full performance

- Memory RAS feature support with DDR5 memory
- Preserving RAS state across boot
- Address Translation support in the memory controller to assist BIOS
- Improved error detection, diagnostics and recovery capability
- IEH 2.0

Security - Increased on-chip security

- Intel® Software Guard Extensions (Intel® SGX) with Advanced RAS and 8S support

NOTE

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids and 5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids workstation do not support Intel® SGX .

- Control flow Enforcing Technology (CET)
- IA Key Protection Technology (KPT) and Public Key Encryption (PKE)

Targeted Workload Acceleration - Acceleration for a variety of workloads

- Machine learning acceleration (TMUL)
- Accelerator Interfacing Architecture (AiA) including Interfacing Architecture (AiA) including work submission new instruction (ENQ, MOVDIRI, MOVDIR64GB)
- IA Comms Enhancements (ICE) 1.0
- Bfloat16
- ~2x DSA performance increase
- Intel® QuickAssist Technology (Intel® QAT) for 200-Gb Integrated Crypto/ Compression acceleration
- IMDB Analytics Acceleration
- Intel® Dynamic Load Balancer (Intel® DLB) – Hardware Scheduling Accelerator
- Flex Bus accelerator connect

Power Management - Tuned performance

- Fast C1E
- Platform and Flexible Turbo

2.2 Workstation Features and Topologies

This section covers the following workstation platform configuration:

- 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids and 5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids Workstation Socket E

The workstation offering offers various features among its processor types and SKUs to meet the needs of multiple market segments. Features within this document may not be available on all platform segments, processor types, or processor SKUs.

Table 3. 2S Workstation Features

Features	Comments[1]
Platform	Eagle Stream 2S
Core Count (Up to)	Varies per processor type and SKU
Processor Socket	Socket E: • 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids and 5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids Workstation 2S
Thermal Design Power (TDP)	Varies by processor type and SKU. See the *Eagle Stream Server and Fishhawk Falls Workstation Platforms Thermal Mechanical Specification Design Guidelines (TMSDG)* for details.
OEM External Node Controller Support (XNC)	Not supported
Intel® UPI 2.0: Interfaces per Socket	Varies per processor type. None:

continued...

Features	Comments[1]
	• 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids and 5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids Workstation 1S 2: • 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids and 5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids Workstation 2S 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids 2S
Intel® UPI 2.0 Speeds[2] (Up to)	12.8 GTS, 14.4 GTS, 16 GT/s
Peripheral Controller Hub (PCH)	Intel® C740 Series Chipset
Last Level Cache (LLC) per CHA	1.875 MB (non-inclusive with the 2MB Mid Level Cache)
Max Number of PCIe Lanes	• 80 total PCIe Lanes • 48 lanes North/ 32 lanes South of socket • Five Gen 5 capable x16 PCIe ports • Each port can be subdivided as 2 x8, 4 x4, 8 x2 or any combination there of NOTE: When subdivided to x2, the PCIe port will operate up to Gen4 speed
Memory: DDR5 Technology Support	DDR5 ECC RDIMM 3DS DDR5 ECC RDIMM
Memory: Speeds[2] (Up to)	4000, 4400, 4800
Memory: Number of IMC per socket	4
Memory: Channels per IMC (Up to)	2
Memory: Max DIMM per Channel (Up to)	2
Memory: Max DIMMs per Socket	16
Memory: Max Logical Ranks / DDR Channel	RDIMM - 4 (SR, DR) 3DS RDIMM - 16 (QR, OR)
Memory: DRAM Density	16 Gb
Memory: Max Capacity[3] (x16 256 GB 16 Gb DDR5) per Socket	4 TB
Intel® AVX-512	Supported
Intel® AVX-512 with 2nd FMA	Varies based on SKU support
Volume Management Device (VMD)	Supported
RAS	Standard RAS
Processor Information ROM (PIROM)	Supported[3]
Non-transparent Bridge Application (NTB)	Supported
Isochronous Application	Not supported
System State S3	Supported
Endpoint PCH (One or more Intel® C620 Series Chipset are used only for the Intel® Quick Assist features. Intel® C620 Series Chipset is connected only via PCIe and DMI is NOT connected nor used.)	Not supported
Trusted Platform Module (TPM)	TPM 2.0 (Discrete or via Intel® PTT embedded in PCH)

continued...

Features	Comments[1]
Intel® Converged Security and Management Engine (Intel® CSME)	Intel® CSME FW 15.x. Enables Intel® vProTM and Intel® AMT Note a platform using Intel® CSME 15.0 is defined as a workstation platform.
Overclocking	Not supported
BMCINIT Mode	Not supported
BIOS Guard (formerly PFAT)	Supported
Intel® Platform Trust Technology (Intel® PTT)	Supported
Sun-NUMA Clustering (SNC)	Supported

NOTES:

1. Where multiple options are listed feature support varies per SKU, or listed as “Supported on select SKUs” not all SKUs support this feature. A SKU consists of a defined set of features.
2. Contact your Intel technical representative for further details.
3. Max memory capacity is SKU dependent.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids and 5th Gen Intel® Xeon® Processor Scalable Family, Codename Emerald Rapids Workstation 2S

The 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids offers various features among its processor types and SKUs to meet the needs of multiple market segments. Features within this document may not be available on all platform segments, processor types, or processor SKUs.

3.0 Core Functional Description

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids core is the next generation core architecture with improved Instructions per Cycle (IPC) and other architectural improvements. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids core features are as follows:

- Virtual address space of 57 bits and a physical address space of 52 bits.
- First Level Cache (FLC) 80 KB total. The FLC is comprised of a 32 KB ICU (Instruction Cache) and 48 KB DCU (Data Cache).
- 2 MB Mid Level Cache (MLC) per core (non-inclusive with the 1.875MB LLC).
- Intel® Hyper-Threading Technology (Intel® HT Technology) when enabled allows each core to support two threads. While some execution resources such as caches, execution units, and buses are shared, each logical processor has its own architectural state with its own set of general-purpose registers and control registers. This feature must be enabled via the BIOS and requires operating system support. . For more information on Intel® Hyper-Threading Technology, see http://www.intel.com/products/ht/hyperthreading_more.htm.
- Intel® Turbo Boost Technology allows the processor cores to run faster than its rated operating frequency if it is operating below power, temperature, and current limits. The result is increased performance in multi-threaded and single threaded workloads. It should be enabled in the BIOS for the processor to operate with maximum performance.
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512) extends the Intel® Advanced Vector Extensions 2.0 (Intel® AVX2) with 512-bit integer instructions, floating-point fused multiply add (FMA) instructions and gather operations. The extended integer vectors benefit math, codec, image and digital signal processing software. FMA improves performance in face detection, professional imaging, and high performance computing. Gather operations increase vectorization opportunities for many applications. A 2nd FMA execution unit is enabled in selected processor SKUs. For more information on Intel® AVX, see http://www.intel.com/software/avx.
- Intel® Total Memory Encryption – Multi-Key (Intel® TME-MK) offers protection for data stored in system memory from attackers in physical possession of the system. The Multi-Key provides means of confidentiality protection to customers who use VMM (or bare metal OS) to control different encryption keys and domains thus secure virtual machines.

Refer to the *Intel® 64 and IA-32 Architectures Software Developer's Manuals* for further details on the 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids core.

Other platform technologies supported by 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids Server core are as follows:

- Intel® Virtualization Technology (Intel® VT) for Intel® 64 and IA-32 Intel® Architecture (Intel® VT-x) provides hardware acceleration for virtualization of IA platforms. Virtual Machine Monitor (VMM) can use Intel® VT-x features to provide more reliable and secured virtualized platform.
- Intel® Virtualization Technology (Intel® VT) for Directed I/O (Intel® VT-d) helps the VMM better utilize hardware to improve performance and availability of I/O devices in virtualized environment by direct assignment of devices. It improves reliability and security through device isolation using hardware assisted remapping.
- Intel® Trusted Execution Technology Architecture (Intel® TXT) defines platform-level enhancements with a Trusted Platform Module (TPM) that provides the building blocks for creating trusted platforms. The Intel® TXT platform helps to provide the authenticity of the controlling environment such that those wishing to rely on the platform can make an appropriate trust decision. The Intel® TXT platform determines the identity of the controlling environment by accurately measuring and verifying the controlling software. For more information refer to *the Intel® Trusted Execution Technology Software Development Guide.*

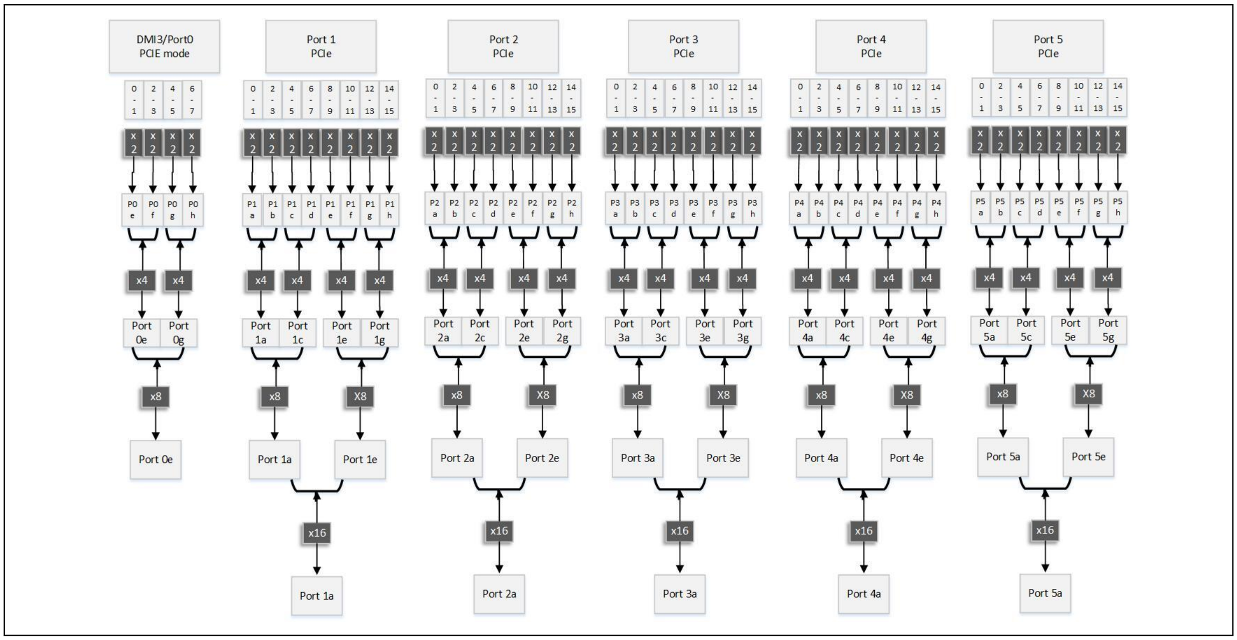
4.0 PCI Express* Modules Functional Description

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor is compliant to *PCI Express Base Specification* Revision 5.0, providing up to 80 Lanes of PCI Express (up to 32 GT/s) and 8 Lanes of DMI (up to 8 GT/s in DMI mode and up to 16 GT/s as PCI Express mode). Processor can be strapped to make use of DMI port as PCI Express port when not connected to a PCH. The 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids CPU also integrates a DMA engine compliant to Intel® Data Streaming Accelerator Technology version 1.0.

The 80 PCI Express Lanes are implemented as five modules (Port 1, Port 2, Port 3, Port 4, Port 5). DMI3 is implemented as Port 0. Each 16 lane can be used as a single x16 Port or Subdivided into x8 or x4 Ports, up to a maximum of 4 Ports @ Gen 5 Speed (32 GT/s). And also the 16 lane module can be subdivided into 8 x2 Ports up to Gen 4 Speed (16 Gbps). The following figure shows the PCI Express Lane partitioning.

The 8 DMI3 lanes are implemented as one single port not subdivided in multiples ports when it is connected to a PCH. Otherwise, When DMI3 lanes are in PCIe mode, those can be used as a single x8 Port or Subdivided into x4 Ports, up to a maximum of two Ports or Subdivided to four x2 Root Ports up to 4.0 Speed (16 GT/s) in all the subdivisions.

Figure 6. PCI Express* Lane Partitioning



SRIS Support

Separate Refclk with Independent Spread-spectrum Clocking (SRIS) is a PCIe* feature that allows two separate clock domains that are not fully synchronous to connect via PCI Express link even when those separate clock domains also have their own Spread Spectrum Clock that is not synchronized.

For more information about SRIS, see the SRIS ECN in the PCI Express Specification.

4.1 Direct Media Interface 3 (DMI3)

The Direct Media Interface 3 is responsible for communication between the PCH and the processor.

The DMI3 Port supports the following:

- Single port only.
- x8 link width (link width downgrade to x4, x2 and x1 supported).
- Gen1 (2.5 GT/s), Gen2 (5 GT/s) and Gen3 (8 GT/s) DMI3 speeds in this mode.
- Lane reversal and Polarity inversion.
  - DMI Polarity inversion is handled automatically by hardware. No soft straps or programming is necessary.
- Additional proprietary messaging with the PCH.
- Intel® Trusted Execution Technology (Intel® TXT) Memory Read and Intel® TXT Memory Write TLPs.
- Address Translation Services (ATS 1.0).
- Access Control Services (ACS).
- Max_Payload_Size of 256 B.
- Completion timeout is disabled by default.
- Features not supported by the DMI3 Port:
  - Hot-Plug
  - DC coupling
  - Multicast/Dualcast
  - Outbound Locks
  - Precision Time Management (PTM)
  - Latency Tolerance Reporting (LTR)
  - Link subdivision
  - Low Power Bus State

DMI3 Port in PCI Express Mode

DMI3 port in PCI Express mode operates as a PCI Express port when is not connected to a PCH.

DMI3 Port as PCI Express mode supports the following:

- It can operate as a standard PCI Express Root Port (RP) compliant with PCI Express Base Specification Revision 5.0.

- DMI3 lanes in PCIe mode supports Gen1 (2.5 GT/s), Gen2 (5 GT/s), Gen3 (8 GT/s) and Gen4 (16 GT/s) PCI Express speeds. DMI3 lanes in PCI Express mode is restricted up to 16 GT/s.
  - — DMI is capable of up to Gen 4 (16 GT/s) speeds but on Emerald Rapids DMI will operate at Gen 3 (8 GT/s) since the PCH is the only device that uses the DMI connection and it can only support Gen 3 speed (8 GT/s).
- In PCI Express mode, the port can be used as a single x8 Port or subdivided into two x4 Ports. And also it can be subdivided to four x2 Root Ports or subdivided into one x4 and two x2 Ports.
- Link width downgrading x1 is supported.
- Hot-Plug is supported.
- VMD is supported.
- For any other PCIe features support, refer to Integrated IO PCI Express Support on page 30 for further details.
- When the boot strap pins sampled on reset are set appropriately and if BIOS clears the DMI_RP_MODE bits, the DMI port will work as a PCI Express port. In this case, the device 0 configuration map is a Type 1 header.
- NTB mode is supported in DMI port when it is not connected to a PCH.
- Features NOT supported by the DMI3 Port in PCIe mode:
  - — CXL mode

4.2 Intel Data Streaming Accelerator Technology

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids Server DMA controller does not support the following features:

- No Backward compatibility to the Intel® QuickData Technology ver 3.4 specifications.
- No DEMS operation via a new Block Fill with NULL mode setting.
- No Hardware model for controlling DMA via NIC hardware.
- No support for Buffer Hint.
- No support for CB_Query message to unlock DMA.
- No support for marker skipping.
- No RAID support except page compare.

4.3 Integrated IO PCI Express Support

PCI Express Link Characteristics:

- Link Subdivision
- Link Training
  - — During link training, link negotiation starts from the highest width of the port (which is set via subdivision control mechanisms) and ramps down to the nearest supported link width that passes negotiation.
- Degraded Mode

- — Degraded mode is supported for x16, x8, x4 and x2 link widths at half or quarter of the original width or as x1.
- Lane Reversal
  - — PCI Express allows for complete reversal of the physical lane ordering between the Root Port and Endpoint (EP). Lane reversal is supported on all PCI Express ports when port subdivision is done, regardless of the link width.
- Max Payload Size
  - — The Max Payload Size is 512 Bytes.
- Isochronous Support and Virtual Channels on PCI Express
  - — PCI Express ports do not support isochrony.
- The processor supports Management Component Transport Protocol (MCTP) per the MCTP specification.
- Volume Management Device (VMD) - An integrated endpoint device that can take control of PCI hierarchies to allow for the aggregation of devices such as PCI Express SSD into a RAID array.
- Write Combining - The PCIe modules does not support any outbound/inbound write combining and neither is any write combining supported on peer-to-peer transactions.
- Relaxed Ordering
- Non-Coherent Transaction Support
- Completion Policy
  - — Read Completion Combining - As a performance optimization, the PCIe modules implements a read completion combining algorithm.
- PCI Express Port Arbitration
  - — The PCIe modules supports a normal weighted (based on port width) round-robin method of arbitration between PCI Express sources.
- 32/64 Bit Addressing
- Hardware autonomous width change, secondary mode only. The processor does not set link width, but responds to link width requests from the link partner.
- Software Link Speed Management
- Link Bandwidth Notification
- Intel® Data Direct I/O Technology (Intel® DDIO)
  - — DDIO makes the processor cache the primary destination and source of I/O data rather than main memory, helping to deliver increased bandwidth, lower latency, and reduced power consumption.
- Security Features
  - — TPM 2.0
  - — Intel® TXT
- Full peer-to-peer support between Uncore PCI Express interfaces.
- Full support for software-initiated PCI Express power management.
- Latency Tolerance Reporting (LTR) support.
- Precision Time Measurement (PTM) support.

- Non-Posted Memory Writes (ENQ) support support for x4, x8 and x16 link-subdivisions.
- RAS Features (see Reliability, Availability, and Serviceability (RAS) Functional Description on page 64).
  - System Management Service Commands to read and write Device registers within the IIO module.
  - Supports PCI Express Base Specification, Revision 2.0, Revision 3.0, Revision 4.0 and Revision 5.0 CRC with link-level retry and link retraining and recovery.
  - Advanced Error Reporting (AER) capability for PCI Express link interfaces Native PCI Express Hot-Plug support.
  - End to End CRC (ECRC).
  - Viral (Advanced RAS only).
  - Enhanced Downstream Port Containment (eDPC).
  - Data Poisoning.
- PCI Express Interface 'Hiding' - The PCIe modules provide the capability to hide a Root Port from OS bus scans. BIOS/FW can set this up via the Port hide functionality.
- Unsupported features in the Root Port:
  - PCI Express locked read requests
  - PHOLD
  - ROL
  - IOSAV is not supported in the processor.

4.4 Intel® Virtualization Technology (Intel® VT) for Directed I/O (Intel® VT-d)

Introduction

Intel® Virtualization Technology for Directed I/O (Intel® VT-d) is the technology that makes a single system appear as multiple independent systems to software. This allows for multiple independent operating systems to be running simultaneously on a single system.

At a high level, support includes:

- DMA remapping
- Device IOTLBs (ATS)
- Interrupt Remapping
- Intel® VT-d Domain Expansion (up to 16 bits of domain ID)

PCI Express Virtualization Support

- Address Translation Services (ATS) - The IIO module allows caching of DMA translations in PCI Express Endpoints. The purpose of having an Address Translation Cache (ATC) in an Endpoint is to minimize time-critical latencies and to provide a way of mitigating the impact on the Root Complex of a Device that does high-bandwidth, widely scattered DMA. The IIO module supports the following ATS requirements:

  - Send a translation in response to a translation request from an endpoint.
  - Issue translation invalidations to endpoint caches.
  - Identify whether the endpoint has completed an invalidation.
- Alternate Request ID (ARI)
  - ARI enables next generation I/O implementations to support an increased number of concurrent users of an Endpoint while providing the same level of isolation and controls found in existing implementations.
- Access Control Services (ACS)
  - ACS can be used to prevent various forms of silent data corruption by preventing PCI Express Requests from being incorrectly routed to a peer Endpoint below a switch. ACS can be used to validate that every Request transaction between two downstream components is allowed. ACS also allows some robustness by checking for the ReqID from a function to be a valid ReqID at a coarse granularity. Refer to the “Other Virtualization Features Supported” section for more details.
  - When ACS is enabled (Upstream Forwarding bit or Request Redirect bit is set in the ACS capability), a PCI Express root port could loopback memory transactions (reads and writes) to the same port if the address map check matched.
- Intel® VT-d Features Supported
  - Root entry, Context entry
  - 57-bit guest address width and 52-bit max host address width for non-ISOCH traffic
  - 4 or 5 level page walks for both non-ISOCH traffic
  - 4K, 2MB, and 1G page sizes for Translation request
  - Register based fault recording and support for MSI interrupts for faults
  - Intel® VT-d on all PCI Express, DMI ports and Intel® Data Streaming Accelerator Technology DMA
  - Intel® VT-d bypass for TCm traffic class on DMI. This is used by Intel® ME to access reserved memory range
  - Leaf and non-leaf caching
  - Domain-specific and device-specific context cache and IOTLB invalidation
  - Boot protection of default page table
  - Non-caching of invalid page table entries
  - Hardware based flushing of translated but pending writes and pending reads, on IOTLB invalidation
  - Page-selective IOTLB invalidation
  - Support for address mask values up to 18 (that is, 1 GB)
  - Support for endpoint caching (ATS)
  - Support for interrupt remapping
  - Support for queue-based invalidation interface
  - Support for Intel® VT-d read prefetching/snarfing, that is, translations within a cacheline are stored in an internal buffer for reuse for subsequent transactions

  - Support for Snoop Control
  - Support for DMA Pass-through
- Other Virtualization Features Supported
  - Support for V, B, C, R and U bits
  - Support SR-IOV devices
- Intel® VT-d New Features supported in 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor
  - Support for PASID-based translation
  - Support for first-level translation structures
  - Support for scalable-mode translation table (enables Intel® Scalable IOV usage model)
  - Support for nested translation (enables Shared Virtual Memory and Nested-I/O Virtual Address usage model)
  - Support for Page Request Service (PRS)
  - Support for per-PASID snoop control for page-walks
- Intel® VT-d Features Not Supported
  - No support for 1, 2 or 3 level walks
  - No support for Intel® VT-d translation bypass address range
  - Second-level Access/Dirty bits
  - Advanced fault log

4.5 Volume Management Device (VMD)

Volume Management Device (VMD) is a new technology used to improve PCI Express management. VMD maps the entire PCI Express trees into its own address space, controlled by a VMD driver, and is enabled in BIOS (to a minimum of x2 Root Port PCI Express granularity). The OS enumerates the VMD device and OS enumeration for the attached child devices ends there. Control of the device domain goes to a VMD device driver which is child device agnostic. A VMD driver sets up the domain (enumerating child devices) and gets out of the way of the fast path to child devices. A VMD driver may load additional child device drivers that are VMD aware, against the respective child devices. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids provides one VMD per PCI Express x16 module, including DMI port in PCIe mode and each VMD can only take ownership of child devices within that PCI Express x16 IIO module. Each VMD can own 0-8 ports on a x16 partition of the respective PCI Express x16 module and can be configured by BIOS.

The 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor supports VMD HW Gen 3.0, providing the ability to span a single volume across PCI Express controllers in PCH as well as in the CPU. Support for Inband PECI configuration access to PCI Express SSDs, and customers need to use their own software to use this HW capability.

VMD HW Gen 3.0 limitations:

- There is no mechanism to aggregate VMDs into a Super VMD
- No MCTP/PCIe messages routing to/from VMD child devices

- Constraints for child devices:
  - MSI/MSI-X interrupts only (no INTx). MSI-X is up to 64 vectors
  - MMIO only (no port-mapped I/O)
- Each VMD on the processor is limited to a max of 128 buses (including the root bus)

Intel's VMD-enabled NVMe driver enables the following functionalities for PCI Express SSDs:

- PCI Express drive can be presented as a target instead of a controller
- Hot-swap of PCI Express SSDs without reboot
- PCI Express errors/events handled by driver
- Aggregates multi-controller SSDs as one bootable volume
- Supports LED indicator enclosure management

Intel's VMD-enabled NVMe driver is scoped to support:

- Mailbox in VMD MEMBAR2 supports MCTP messaging between VMD itself and BMC
- PCI Express NVMe SSDs
- Two level deep of switches (may be internal switch). One discrete switch per VMD, up to 48 SSDs behind switch.

4.6 IIO Power Management

The IIO supports both power management within the module and on the PCI Express links.

IIO Module Power Management

The system power states supported by the IIO module are enumerated in the table below. Note that no "device" power states are explicitly defined except for the Intel® Data Streaming Accelerator Technology integrated device. In general, the IIO module power state may be directly inferred from the system power state. There are two categories of S-State flows:

- OS initiated
  - Targeted to reach S3, S4, or S5
- Non OS initiated
  - Targeted to reach S1_RW

The table below lists the OS Initiated System Power States.

Table 4. OS Initiated System States

System State	Description
S0	Full On: [Supported] Normal operation
S1	Stop-Grant: [Not Supported]
S2	Power-On-Suspend: Not Supported

continued...

System State	Description
S3	Suspend to RAM (STR): [Only Workstation is Supported] CPU and PCI reset. All context can be lost except memory. This state is commonly known as "Suspend."
S4	Suspend to Disk (STD): [Only Workstation is Supported] CPU, PCI and Memory reset. This state is commonly known as "Hibernate." The S4 state is similar to S3 except that the system context is saved to disk rather that main memory. Uses the same sequence as S3.
S5	Soft off: [Supported] Power removed. Uses the S3 sequence.

Exit from the S4 and S5 states requires a full system reset and initialization sequence.

For Non-OS Initiated System Power States, when hardware initiates a system power state transition, Intel® ME/VE must be preserved. Intel® ME/VE requires all cores to stop execution before they can drain their state.

PCI Express Link and Device Power Management

- Supports Autonomous Linkwidth
- Supports L0, L1 and L3 link state (L0s, L0p, L1.substate and L2 not supported)
- Supports Active State Power Management (ASPM) L1
- Support for the L3 link state
- Support for either MSI or GPE event on power manage events internally generated (on a PCI Express port hot-plug event) or received from PCI Express
- Support for D0 and D3 hot power management states per PCI Express port and support for wake event from these D3hot states on a hot-plug event at a PCI Express port. In D3hot, the IIO module will master abort all configuration tx targeting the PCI Express link

Intel® Data Streaming Accelerator Technology Power Management

- Intel® Data Streaming Accelerator Technology device supports the D0 device power state that corresponds to the "fully-on" state and a pseudo D3hot state. Intermediate device power states D1 and D2 are not supported
- Power Management w/ Assistance from OS Level Software
- Support Device Power States
    - The PCIe Modules supports all PCI-PMI and PCI Express messaging required to place any subordinate device on any of its PCI Express ports into any of the defined device low power states.
    - In addition, Intel® Data Streaming Accelerator Technology integrated device (DMA) in the PCIe modules can be placed in D0 or D3hot states by programming the PMCSR register of its power management structure.
    - Directly attached native PCI Express devices are not limited in their available low power states.

4.7 IIO Interrupts

The PCIe modules support both MSI and legacy PCI interrupts from its PCI Express ports.

4.8 PCI Express Hot-Plug/Removal Support

The PCIe module has Hot-Plug/Removal support for PCI Express devices attached directly to the IIO Root Ports. Hot-plug of PCI Express devices are defined in PCI Express/PCI specifications. Four versions of Hot-Plug are supported:

- Standard PCI Express hot-plug using four input pins and four output pins per Root Port. These pins are attached to I/O extenders that are connected to the hot-plug SMBus Controller (HPSMBC) port of the processor.
- Surprise PCI Express hot-plug using the Presence Detect pin on the I/O extenders.
- Surprise PCI Express hot-plug using the link status LinkDown to indicate card presence/absence.
- Storage PCI Express hot-plug uses four input pins and four output pins per Root Port connected through the I/O extenders.

Surprise PCI Express hot-plug is not supported when Presence Detect pin event is used in cards with CEM form factor for 5.0 speed (32 Gbps). For more details, see the PCI Express CEM 5.0 specification.

4.8.1 Hot-Plug SMBus Controller Segment Support

The IO module contains a dedicated SMBus Port which is connected to the Converged eXPress SMBus Controller (CXPSMBC) that serially reads and writes these sideband PCI Express hot-plug signals from I/O expanders. External platform logic is required to implement these I/O expanders to convert the IO module serial stream to parallel.

Summary of the PCI Express Hot-Plug support:

- Support for up to 8 hot-pluggable PCI Express slots per PCI Express x16 module (1 per root port), up to 5 PCI Express x16 modules per processor.
- Support for up to 44 hot-plug slots selectable by BIOS.
- Support any of the 44 IO Expander port to be assigned to any PCI Express port.
- Support for serial mode hot-plug only using I2C/SMBus devices like PCA9555 and PCA9544A.
- Single SMBus is used to control hot-plug slots.
- Support for CEM/Express Module/Cable form factors.
- Support MSI or ACPI paths for hot-plug interrupts.
- A hot-plug event cannot change the number of ports of the PCI Express interface (that is, subdivision).
- Support Surprise PCI Express hot-plug using the Presence Detect pin or link status.
- Surprise PCI Express hot-plug is not supported using Presence Detect pin event in cards with CEM form factor for 5.0 speed (32 Gbps).

4.8.2 Standard PCI Express Hot-plug Interface

The following table describes the different signals supplied by the PCIe Module for each port. These signals are controlled and reflected in the PCI Express root port hot-plug register fields indicated. The I/O Expander Port pin and direction on a given I/O Expander and Port is listed for each signal. For a precise definition of these signals, refer to PCI Express Base Specification, Revision 5.0 and PCI Express Server/ Workstation Module Electromechanical Specification, Revision 1.0.

Table 5. Standard PCI Express Hot-plug Interface

Signal Name	Control or Status Register	I/O Port Pin	I/O Port Dir	Description
ATNLED	SLOTCON [7:6]	Pin [0]	Output	Provides an LED indicator to get the attention of the system operator. This is an active high signal. 00: reserved 01: LED On 10: LED Blink (1 Hz) 11: LED Off
PWRLED	SLOTCON [9:8]	Pin [1]	Output	Provides an LED indicator on the state of the power supplied to the PCI Express slot. This is an active high signal. 00: reserved 01: LED On 10: LED Blink (1 Hz) 11: LED Off
BUTTON#	SLOTSTS [0]	Pin [3]	Input	Input signal from a button for each PCI Express slot that the system operator presses to indicate that the operator wishes to hot-remove or hot-add a PCI Express card or module. If the button is pressed (asserted low), the Attention Button Pressed Event bit is set.
PRSNT#	SLOTSTS [3] SLOTSTS [6]	Pin [4]	Input	Indicates if a hot-pluggable PCI Express card/module is currently plugged into the slot. When a change is detected in this signal, the Presence Detect Event Status bit [3] is set. The current value is stored in the Presence Detect State bit [6].
PWRFLT#	SLOTCON [1]	Pin [5]	Input	A signal from the power controller that when asserted (low), the Power Fault Event bit is set, indicating a power fault has occurred.
PWREN#	SLOTCON [10]	Pin [2]	Output	If the Power Controller Bit is set, this signal is asserted (low), which enables power for the PCI Express slot.
MRL#	SLOTCON [2] SLOTCON [5]	Pin [6]	Input	Manual Retention Latch (MRL#) status or whether the manual retention latch holding the card- edge card is closed or open. This pin is used in this mode if VPPCSR.HPFF= 0. The pin value is reflected in bit 5 in this mode. Bit 2 is set when there is a change in value.
EMILS	SLOTCON [2] SLOTCON [7]	Pin [6]	Input	Electro-Mechanical Interlock Latch Status (EMILS) input indicates whether the electromechanical retention latch holding the SIOM card is closed or open (controlled by software with the EMIL pin). This pin is used in this mode if VPPCSR.HPFF = 1.

continued...

Signal Name	Control or Status Register	I/O Port Pin	I/O Port Dir	Description
				The pin value is reflected in bit 7 in this mode. Bit 2 is set when there is a change in value.
EMIL	SLOTCON [11]	Pin [7]	Output	Electro-Mechanical Interlock Latch control output that opens or closes the retention latch on the board for this slot for SIOM form- factor cards.

4.8.3 Storage PCI Express Hot-plug Interface

Applications in using PCI Express for directly connecting to storage devices like Disk Drives or Solid State Drives may require a different hot-swap solution. Drives are typically hot-swapped from a backplane that is not located on the motherboard. Many storage backplanes often use a different definition for hot-swap signals than those used for PCI Express.

The following solution redefines the existing PCI Express Hot-Plug signals to re-use them for these storage applications. There is no hardware difference between this definition and PCI Express Hot-Plug, only in how they are used. The table below describes this alternate definition.

Table 6. Storage PCI Express Hot-plug Interface

Signal Name	Control or Status Register	I/O Port Pin	I/O Port Dir	Description
Fault or Select 0	SLOTCON [7:6]	Pin [0]	Output	Indicates a fault condition on the drive to the user. 00: reserved 01: LED On 10: LED Blink (1Hz) - unused 11: LED Off It may also be used as part of a 3-bit select for the blink rate of a single LED.
Locate or Select 1	SLOTCON [9:8]	Pin [1]	Output	Provides an LED to allow the user to locate the drive being sought. 00: reserved 01: LED On 10: LED Blink (1Hz) - unused 11: LED Off It may also be used as part of a 3-bit select for the blink rate of a single LED.
Slot Power-On	SLOTSTS [0]	Pin [3]	Input	A signal from a controller that when asserted (low), indicates power-on has occurred (bit 0 is set on a high to low transition) Pin 3 must be tied to an inverted version the same power signal tied to pin 5.
Device Installed	SLOTSTS [3] SLOTSTS [6]	Pin [4]	Input	Indicates if a hot-pluggable PCI Express storage device is currently installed. When a change is detected in this signal, bit [3] is set. The current value of the pin is stored in bit[6].
Slot Power-Loss	SLOTCON [1]	Pin [5]	Input	A signal from the power controller that when asserted (low), indicates power-loss has occurred. Bit 1 is set on a high to low transition. Pin5 must be tied to an inverted version of the same power signal tied to pin 3.

continued...

Signal Name	Control or Status Register	I/O Port Pin	I/O Port Dir	Description
Power Control	SLOTCON [10]	Pin [2]	Output	If the Power Controller Bit is set, this signal is asserted (low), which enables power for the storage device slot.
Slot Power Manageme nt	SLOTCON [2] SLOTCON [5]	Pin [6]	Input	Indicates there is power control circuit on the storage backplane. The value of this pin is reflected in bit 2. A change in value causes bit 5 to become set.
PFA or Select 2	SLOTCON [11]	Pin [7]	Output	This pin may be used as an independent LED indicating Power Fault for the Array. It may also be used as part of a 3-bit select for the blink rate of a single LED.

4.8.4 Surprise Hot-Plug

There are two forms of surprise Hot-Plug that may be implemented:

- Out-of-band Surprise Hot-Plug using the external PRSNT# pin on the slot
- In-band Surprise Hot-Plug using LTSSM Link-Down indication

Out-of-band Surprise Hot-Plug refers to the ability to remove a PCI Express card without prior warning from the push button. The PRSNT# detect pin is used to signal the SLOTSTS register that the card has been removed. The same signal is also used to signal that a card has been added.

Surprise PCI Express hot-plug is not supported using Presence (PRSNT#) Detect pint event in cards with CEM form factor for 5.0 speed (32 Gbps).

In-band Surprise Hot-Plug refers to using the Link-Down indication from the link layer state machine that the card has been removed. This can be used in cases where the PRSNT# pin is not available. Link-Down only occurs after a recovery attempt has failed. The addition of a card would be followed by Link-Up indication to be used to indicate a new card has been added.

4.9 Non-transparent Bridge

The PCI Express Non-Transparent Bridge (NTB) acts as a gateway that enables high performance, low latency communication between two PCI Express Hierarchies, such as a local and remote system. The NTB allows a local processor to independently configure and control the local system and provides isolation of the local Host memory domain from the remote Host memory domain, while enabling status and data exchange between the two domains. The NTB is discovered by the local processor as a Root Complex Integrated Endpoint (RCiEP).

NOTE

If not explicitly stated, the details of NTB mode of operation are the same as the RP mode.

4.9.1 NTB Features Supported

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids supports the following NTB features.

The NTB only supports one configuration/connection model:

- NT Port attached to another NT Port of the same component type and generation. It means that one 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids NT Port can only connect to another 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids server NT Port.

The NTB provides Direct Address Translation between the two PCI Express Hierarchies through two separate regions in Memory Space. Accesses targeting these Memory addresses are allowed to pass through the NTB to the remote system. This mechanism enables the following transactions flows through the NTB:

- Both Posted Mem Writes and Non-Posted Mem Read transactions across the NTB
- Peer-to-Peer Mem Read and Write transactions to and from the NTB

In addition, the NTB provides the ability to interrupt a processor in the remote system through a set of Doorbell registers. A write to a Doorbell register in the local side of the NTB will generate an interrupt to the remote processor. Since the NTB is designed to be symmetric, the converse is also true.

Lastly, the NTB also provides a set of scratchpad registers, most generic with a couple specialized, that facilitate firmware or software communication between the two systems potentially before System Memory configuration and initialization have completed.

Internal Endpoint (iEP) Features

- Root Complex Integrated Endpoint compliant to the PCI Express Base Specification, Revision 4.0.
- Three 64b Memory BAR’s
  - — MEMBAR0 – Registers.
  - — MEMBAR1 – Direct Address Translation.
  - — MEMBAR2 – Direct Address Translation.
- MSI-X Capability
  - — 1 Vector dedicated for HW use.
  - — 32 Vectors for SW use.
- PCI Power Management States: D0 and D3
  - — No_Soft_Reset on D3hot to D0 transition.
- Transaction Descriptor Attributes: RO and NS.
- Max Payload Size: 512 B.
- Max Read Request Size: 4 KiB.
- Multicast (MC) Extended Capability.
- 32 Doorbell registers.
- 18 Scratchpad registers (16 general + semaphore + sticky).

Features Specifically Not Supported by Internal Endpoint (iEP)

- Multi-Function Device.
- 32b Memory BAR's.
- Legacy I/O Space.

- Legacy VGA Memory and I/O Space decode.
- Locked Transactions (MRdLK).
- Function Level Reset (FLR).
- Atomic Operations.
- Power Management Event (PME) generation.
- End-to-end CRC (ECRC).
- Optimized Buffer Flush/Fill (OBFF) Mechanism.
- ID-Based Ordering (IDO).
- TLP Prefix.
- Advanced Error Reporting (AER) Extended Capability.
- Power Budgeting Extended Capability.
- Virtual Channel (VC) Extended Capability.
- Access Control Services (ACS) Extended Capability.
- Resizable BAR Extended Capability.
- Alternative Routing-ID Interpretation (ARI) Extended Capability.
- Latency Tolerance Reporting (LTR) Extended Capability.
- TLP Processing Hints (TPH) Requester Extended Capability.
- Process Address Space ID (PASID) Extended Capability.
- Address Translation Services (ATS).
- Single Root I/O Virtualization (SR-IOV).
- Multi-Root I/O Virtualization (MR-IOV).

NT Link Features for NT Port Mode

- Max Payload Size: 512 B.
- Link Widths: x16, x8, x4, x1.
- Link Speeds: 16.0, 8.0, 5.0, 2.5 Gb/s.
- SW Link Speed Management.
- HW Autonomous Link Width Change.
- Link Bandwidth Notification.
- Proprietary error logging of Advanced Error Reporting (AER) equivalent errors.

Features Specifically Not Supported by NT Link

- Active State Power Management (ASPM) States: L1 and L0s.
- PCI PM L1.
- Power Management Event (PME) generation.
- Hot Reset.
- Virtual Channel (VC).
- Configuration Space Request.
- Legacy I/O Space Request.

- Locked Transactions (MRdLK).
- End-to-end CRC (ECRC).
- Atomic Operations.
- ID-Based Ordering (IDO).
- TLP Prefix.
- Latency Tolerance Reporting (LTR) messages.
- TLP Processing Hints (TPH).
- Process Address Space ID (PASID).
- Address Translation Services (ATS).

4.10 Compute Express Link (CXL)

The 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor is compliant to Compute Express Link (CXL) Specification Revision 1.1. Any x16 PCI Express Port can connect to either a PCI Express device or CXL device, 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor supports up to a maximum of four CXL devices simultaneously. It means that PCI Express Ports can operate as a standard PCI Express Root Port (RP) or as Compute Express Link Port up to 32 Gbps. A maximum 64 of 80 PCI Express lanes can be connected to four CXL devices simultaneously.

- CXL is based on PCI Express PHY infrastructure. It leverages channel, retimer, Physical layer and Protocols:
  - CXL.io is a PCI Express-based non coherent I/O protocol with enhancements for accelerator support. CXL.io includes semantics that are leveraged from PCI Express Base Spec Revision 5.0 definition with some deltas. This protocol is mandatory when IIO Module is used as CXL port.
  - CXL.cache is a Agent coherency protocol that supports device caching of Host memory. This protocol is optional.
  - CXL.memory is a Memory access protocol that supports device-attached memory. This protocol is optional.
- 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor supports caching devices or accelerators defined as Type 1 and accelerator with memory defined as Type 2 devices in CXL 1.1 Spec. Memory expanders can be created using Type 2 with the limitations listed in Type 2 Memory only devices on page 46.
- The processor hardware is capable of auto detect whether the other end is a PCI Express card or a CXL device and dynamically configure the link.
- Single Port only

Any of 5 x16 PCI Express ports (Port 1, Port 2, Port 3, Port 4, Port 5) support CXL devices.

4.10.1 CXL.io

CXL.io is a PCI Express-based non coherent I/O protocol with enhancements for accelerator support. CXL.io provides the load/store interface for I/O devices. This devices operate as a standard PCI Express Root Complex Integrated Endpoint (RCiEP).

CXL.io Characteristics:

- Single Port x16 only
- Port Subdivision for each module must be x16. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids does not support subdivision x8 or x4
- Supported max speed is Gen5 (32 GT/s) speed
- Degraded Mode
    - Supported degraded speeds are 8 GT/s and 16 GT/s.
    - Supported degraded widths: x8. The x16 CXL Port supports degraded widths mode at half (x8) of the original width.
- Lane Reversal
- CXL Link Training
    - Refer to the PCI Express 5.0 Base Specification and CXL 1.1 specification for details.
- 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids server processor can support a maximum of four CXL devices (x16 or x8 in link width degradation) but all five x16 PCIe Ports can support a CXL device. BIOS will enumerate only the first four CXL devices and fail to enumerate any more.
- Max payload Size:
    - The Max Payload size is 512 Bytes.
- Virtual Channels:
    - CXL.io supports the default Virtual Channel 0 (VC0) and VC1.
- Memory Type indication on Address Translation Service (ATS) supported.
- Deferrable Write (a.k.a Non-Posted Memory Write (ENQ)) supported.
- Data Poisoning by transmitter supported.
- CXL.io Features NOT supported:
    - Link Subdivision
    - Hot-Plug
    - Multicast / Dualcast
    - Outbound Locks
    - Enhanced Downstream Port Containment (eDPC)
- CXL.io Transaction and Link Layer follows PCI Express Specification definition but it has some deltas as link training, power management, reset and others. Refer to Compute Express Link Specification revision 1.1 for detail on CXL operation.
- See the following sections to identify other PCI Express features and behaviors supported in CXL.io as well.

4.10.2 CXL.memory

CXL.memory is a Memory access protocol optimized for low latency and high bandwidth that supports device-attached memory. The protocol is transactional in nature and thus, memory type and configuration independent. CXL.memory can be optional in some CXL device implementations.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids is compliance with CXL 1.1 spec supporting all different transactions requests from Master to Subordinate (M2S) and responses from subordinate to master.

CXL.cache support the following RAS Features in 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids server processor:

- Supports Compute Express Link Specification Revision 1.1 Link CRC with link-level retry and link retraining and recovery.
- Data Poisoning
- Viral (Strong Error Containment)

4.10.3 System Architecture

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor supports caching devices or accelerators defined as Type 1 and accelerator with memory defined as Type 2 devices in CXL 1.1 Spec. In 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor, memory expanders using Type 3 devices are not supported. Memory expanders can be created using Type 2 with the limitations listed in Type 2 Memory only devices on page 46.

4.10.3.1 Type 1 Devices

Caching agents and Accelerators without Memory are defined as Type 1 devices. The usages model can be Partition Global Address Space (PGAS) Network Interface Card (NIC) and NIC atomics operations. These devices required a full cache coherency to implement an unlimited number of atomic operations.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor detects these devices and configures control register properly when they are marked as IO_capable, Cache_capable and NOT Mem_capable during CXL training.

4.10.3.2 Type 2 Devices

Accelerators with Memory are defined as Type 2 devices and the usages model are Graphics Processing Unit (GPU) and Field Programmable Gate Arrays (FPGAs).

Type 2 Features:

- 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids supports Host-managed Device Memory (HDM) and traditional IO/PCI Express Private Device Memory (PDM).
- In HDM, Processor fully supports the Bias Based Coherency Model with two defined modes of the bias:
  - Host Bias (device-attached memory accessed by the Host).
  - Device Bias (device-attached memory accessed by the device).
- Support for Bias Mode Management schemes as defined in CXL 1.1 Spec:
  - Software Assisted Mode
  - Hardware Autonomous Mode

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids server processor detects these devices and configures control register properly when they are marked as IO_capable, Mem_capable and Cache_capable during CXL training.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids does not support persistent memory behind CXL link for Type 2 devices. And it only support 1LM configuration.

4.10.3.2.1 Type 2 Memory only devices

In 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor, Type 3 devices are not supported. But Memory expander is feasible with 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor in Eagle Stream platforms using Type 2 Memory only with certain limitations listed next.

Key limitations for Type2 Memory only devices in 4th Gen Intel Xeon Scalable Processor, Codename Sapphire Rapids server:

- No support for Memory RAS flows, including MCA.
- No support QoS for low bandwidth memory, including RDT support.
- No Optimal bandwidth for UMA traffic, including Directory support.
- No Intel® TME-MK support.
- No ADR support.
- No Advanced interleaving support. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids can not interleave addresses across multiple Memory only devices on different links.

This Type 2 memory only devices are marked as IO_capable, Mem_capable and Not Cache_capable. It is essentially behaving as a Type 3 device with limitations since Type 2 Memory only devices are not Cache_Capable. It is not expected to initiate any requests over CXL.Cache channel.

The device operates primarily over CXL.mem to service requests sent from the processor. The CXL.io link is used device discovery, enumeration, error reporting and management.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor does NOT support persistent memory behind CXL link for Type 2 Memory only devices. And it only support 1LM configuration.

4.10.4 CXL.cache

CXL.cache is a Agent coherency protocol optimized for low latency and high bandwidth that supports device caching of Host memory. The protocol operates at 64 bytes cache-line granularity and uses Host Physical Address (HPA) for all transactions and does not preserve ordering for requests. CXL.cache can be optional in some CXL device implementations.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids is fully compliance with CXL 1.1 spec supporting three channels in Device to Host (D2H) and Host to Device (H2D) direction. 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids supports all different semantics for incoming D2H request /response and outgoing H2D request / response defined in CXL 1.1 spec.

CXL.cache support the following RAS Features in 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids server processor:

- Supports Compute Express Link Specification Revision 1.1 Link CRC with link-level retry and link retraining and recovery.

- Data Poisoning
- Viral (Strong Error Containment)

For more specific information relating CXL.cache, see the Compute Express Link Specification revision 1.1.

4.11 PCIe Precision Time Measurement (PTM)

The PCI Express* Precision Time Measurement (PTM) protocol is a standard but optional hardware feature of PCI Express* (and applies equally to protocols derived from PCI Express*, for example, CXL*) that is discoverable through PCI Config Space.

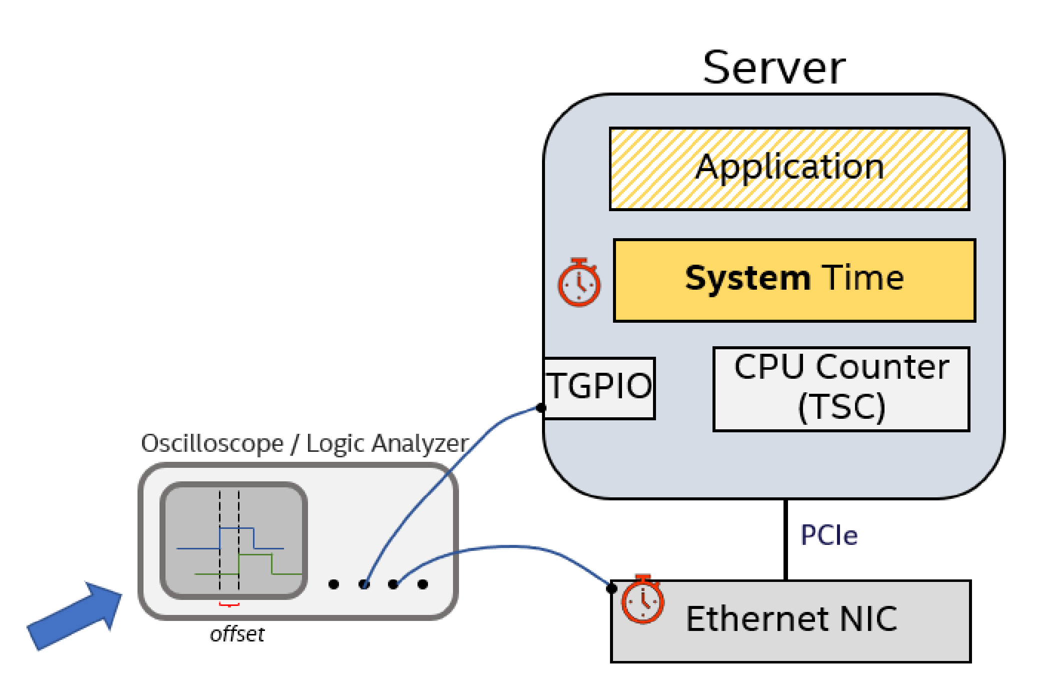
Functionally, it transfers the value of a hardware counter (known as PTM Root Time on the PCI Express* Base Specification) from the PTM-capable PCIe Root Complex to any PTM-capable PCIe device attached to it, on request, with nanosecond accuracy, through a round-trip exchange of PTM TLPs, each of which is timestamped on Rx and Tx.

4.12 Timed-GPIO (TGPIO)

The platform supports one or more Timed-GPIOs (TGPIOs), providing hardware observability of the time of the CPU (see the PCI Express* Precision Time Measurement (PTM) section in this document, and ART and Invariant Timekeeping in the IA SDM).

An example use case would be to use IEEE 1588 (see the following figure), the precision time protocol (PTP) to synchronize time across an Ethernet network from a GPS receiver to the PTP counter on the NIC. Then, cause the NIC to generate a pulse per second (PPS, a rising edge on the second, every second) and also use the TGPIO to generate a PPS output. By observing both PPS signals on an oscilloscope or logic analyzer, the end-customer is able to ascertain the accuracy with which the PTP counter on the NIC has been transferred to the system time of the operating system, which is typically based on TSC (and converted to nanoseconds-since-1970 by a linearity relationship that must be updated periodically due to frequency errors and variations of the crystal). PTM is the preferred method of transferring PTP time from the NIC to System Time.

Figure 7. TGPIO Example - Time-Aware GPIO Use



This component includes two such TGPIOs, called TIME_SYNC_0 and TIME_SYNC_1.

5.0 Intel Ultra Path Interconnect Version 2.0 (Intel UPI 2.0) Functional Description

The 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids Server processor uses a coherent interconnect, called Intel Ultra Path Interconnect Version 2.0, for scaling to multiple sockets. The acronym Intel® UPI 2.0 is used to indicate Intel® Ultra Path Interconnect (Intel® UPI_v2.0) and is used throughout this document. The Intel® UPI 2.0 link is the coherent communication interface between processors. Intel® UPI 2.0 architecture can be used in a wide variety of server platform configuration. The definition is processor agnostic.

Overview of the Intel® UPI 2.0 module

An Intel® UPI_v2.0 is comprised of the following layers for each Intel® UPI link:

- Physical Layer - The Intel® UPI 2.0 Physical layer (PHY) is a hardware layer that lies between the Link layer above it, and the physical wires that connect to other devices. The Physical layer is further sub-divided into the logical and electrical sub-blocks.
- Link Layer - The Intel® UPI 2.0 link layer bi-directionally converts between protocol layer messages and Link layer Flits, passes them through shared buffers, and manages the flow control information per virtual channel. The link layer also detects errors and retransmits packets on errors.
- Routing Layer - The Routing Layer is distributed among all agents that send Intel® UPI messages on the mesh (Intel® UPI, CHA, PCIe, IMC). The Intel® UPI Module provides a routing function to determine the correct mesh stop to which to forward an incoming packet.
- Protocol Layer - The Intel® UPI module does not implement the Protocol Layer. A protocol agent is a proxy for some entity which injects, generates, or services Intel® UPI transactions such as memory requests, interrupts, etc. The Protocol Layer is implemented in the following modules: Coherency Home Agent (CHA), PCIe, Configuration Agent (Ubox). A Coherency agent (CA) in the CHA both generates requests and services snoops. A Home Agent (HA) in the CHA services requests, generates snoops, and resolves conflicts. CHA will sometimes behave as CA, sometimes as a HA, and sometimes both at the same time. The PCIe module handles most IO proxy responsibilities. The Ubox handles internal configuration space and some other interrupt and messaging flows. A HA acts as proxy for DRAM, while the PCIe/Ubox handle all non-DRAM (NCB and NCS) requests.

In 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor, the UPI module contains a cryptography engine. This enables a security feature for ensuring confidentially, integrity and replay-protection of packets transmitted over Intel® UPI 2.0 links that belong to a secure memory range. It provides data encryption, decryption, as well message authentication code generation and verification.

Intel® UPI_v2.0 link features

- Home snoop based coherency protocol

- 12.8 GT/s 14.4 GT/s and 16 GT/s operational speeds
- 52-bit Physical Addressing
- 24 lane physical interface per direction, differential, with embedded clocking
- 4-bit Node ID (only 3 bits used)
- L0 and L1 power state (No support for L0p. No support for L1 in Slow Mode)
- Lane reversal across the link
- Per lane polarity inversion
- VNA and VN0 support
- Intel® UPI protocol protection via CRC 16bit
- Poison handling
- Error detection and logging for both corrected and uncorrected errors

Supported only on processor types which support Advanced RAS:

- Viral mode of error containment
- Port self-healing from full to x8 width

Bifurcation is not supported on Intel® UPI ports.

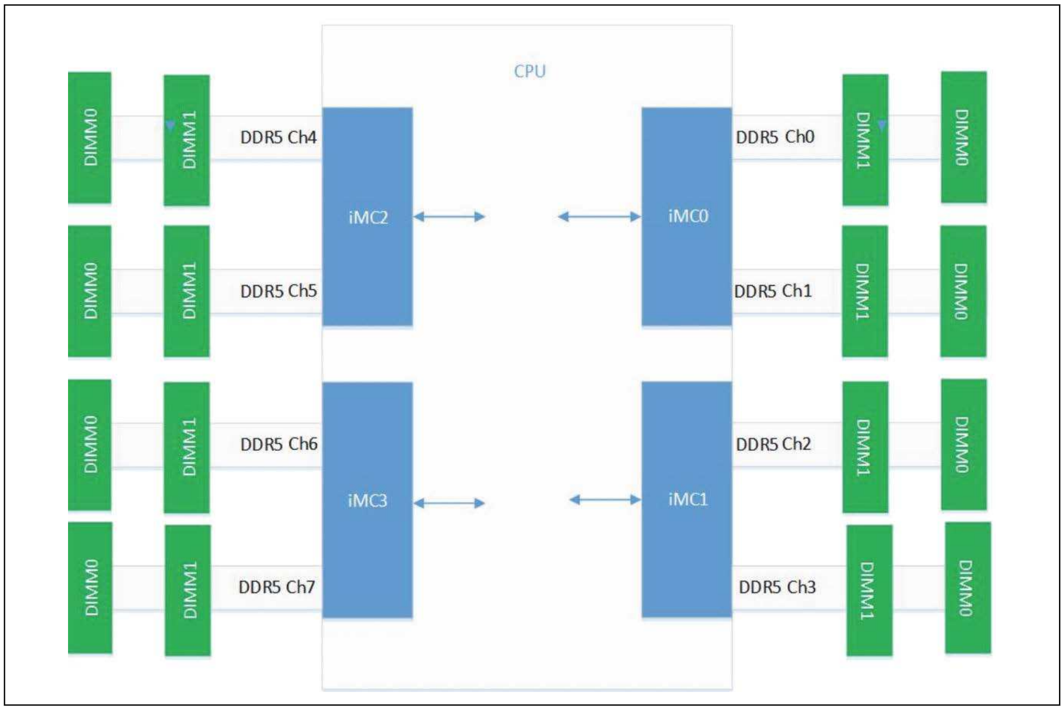
5.1 Link Power Management

Link power management involves managing the Intel® UPI Physical Layer in order to reduce the power consumption by the UPI interface. This handles how the interconnect can be reconfigured to reduce operating power as well as the ways in which power is saved when the interconnect is idle. All link power management features are optional.

6.0 Integrated Memory Controller (IMC) Functional Description

The IMC is the Integrated Memory Controller for the CPU, the 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor supports four IMCs per socket. Each IMC is capable of controlling two DDR5 channels. The IMC interfaces with the rest of the uncore via its own dedicated Mesh2Mem (M2M) logic which interfaces to the mesh. Read and Write addresses from the CHA are steered according to address decode to an IMC, and from there the IMC decodes to a channels or a pair of channels (in the mirroring case). The channels decompose the reads and writes into pre-charge, activate and column commands and schedule their request to the DDR command/address lines. Write data is enqueued in the Write Data Buffers where partial writes are merged to form full line writes. The Channels drive write data to the DRAMs and the DRAMs return read data on the bidirectional data bus. Read returns from the IMC channels are corrected if needed.

Figure 8. IMC Block Diagram for 8 DDR5 Channel Processor Type



IMC features

- Each iMC supports two DDR5 channels and each 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids CPU integrate four iMC’s totaling eight channels with 217 GB/S bandwidth on stream triad.
- Adaptive Data Correction (ADC) and Adaptive Double Device Data Correction (ADDDC) via Adaptive Virtual Lockstep as a RAS option.

- Features map out hard failures.
- Closed Page and Open Page (adaptive idle timer)

DIMM mixing and population rules can be found in the PDG.

The IMC controller contains an encryption/decryption engine to provide , Intel® Total Memory Encryption – Multi-Key (Intel® TME-MK) of data stored in system memory. By encrypting all of the data sent to the system memory, Intel TME-MK prevents cold-boot and other side-channel attacks where the data in system memory or en route to system memory might be stolen. The addition of Intel TME-MK in the processor provides a means of confidentiality protection to cloud customers who use VMM (or bare metal OS) to control different encryption keys and domains thus secure virtual machines.

6.1 Memory Power Features

This section describes the memory power management features supported by the processor for DDR5.

6.1.1 DDR5 Power Optimization Features

Table 7. DDR5 Power Optimization Feature Summary

Memory Power Optimization Feature	Processor Support
Fast CKE: Active Power Down	Per Rank
Fast CKE: Precharge Power Down (DLL On)	Per Rank
Slow CKE: Precharge Power Down (DLL Off)	Not supported. (Not supported on DDR5.)
Opportunistic Sefl-Refresh	Not supported.
RDIMM register IBT/ODT Off	IBT-Off is supported as a static boot-time configuration only, and is supported only for 1 DPC configurations
2X refresh with Optional DIMM feature with Auto Self- Refresh (ASR) based on build-in temperature sensor	Supported
2x refresh without ASR with closed loop	Dynamically entrance if > double refresh threshold
Async DIMM Self-Refresh (AsyncSR/ADR)	Supported
IMC PLL Off	During package C6
Async ODT (needed for per rank PD DLL off when one rank terminates other DIMM)	Not supported
DDR5 (1.1v)	Supported
Close Loop Thermal Throttling (CLTT)	Supported
Open Loop Thermal Throttling (OLTT)	Supported

NOTE

The CKE function is no longer supported by CKE pins explicitly. The CKE function is supported by DDR5 commands and the transitional state of CS_n pin.

The following listdescribes the DDR5 DRAM power features.

- Normal operation

- — Highest power consumption. CKE function is asserted.
- CKE Power Down
    - — CKE power down is the basic DRAM power saving feature. The CKE function to the DRAMs is used to enter and exit power-down modes. When CKE is off, the clock internal to the DRAMs is disabled and DDR power is significantly reduced. The memory controller has an activity timeout for each rank that is configurable per channel. When no reads are present to a given rank for the configured interval, the memory controller will transition the rank to power-down mode.
    - — There are two power down modes:
        - Active Power Down - refers to the power state reached when CKE is lowered without closing pages in the DRAM.
        - Precharge Power Down with Fast Exit (DLL-ON) - refers to a lower power state where all pages are closed before CKE is lowered.
- DLL and PLL Shutdown
    - — The self-refresh may be a trigger for master DLL shut-down and PLL shut-down. This puts DRAM in a deep self refresh. The master DLL shut-down is issued by the IMC after the DRAMs have entered SR.
- Auto Self-Refresh (ASR)
    - — DDR5 DRAM supports autonomous self-refresh rate management based on the temperature of the DRAM. In the default mode, DRAM autonomously manages refresh rate based on an internal temperature sensor. This mode is configured during initialization by writing to the ASR field in MR2 register. In case of high-temperature, the refresh rate should be increased. The IMC monitors the temperature during self-refresh entry and configures refresh rate accordingly.
- DIMM Register Power Management
    - — The buffer on an RDIMM or LRDIMM operates in one of two modes:
        - IBT-On Mode - This is the normal mode of operation. When the CKE function at the RDIMM or LRDIMM level is deasserted, the buffer enters Register Power Down with IBT-On. Input Buffer Terminators (IBTs) are left on, saving power as compared to leaving either of the CKE functions asserted. This mode allows the channel to continue communicating to other DIMMs.
        - IBT-Off Mode - IBT-Off is supported as a static boot-time configuration only, and is supported only for 1 DPC configurations. Dynamic modification of IBT- On / IBT-Off is not supported. In IBT-Off mode, when the CKE function at the RDIMM or LRDIMM level is deasserted, the buffer enters to Register Power Down mode with IBT-Off. Input Buffer Terminators are turned off, saving power per register beyond the savings gained by invoking Register Power Down with IBT-On. This mode prevents the channel from communicating to other ranks. Hence, this mode is only invoked when the entire channel is idle.

7.0 Intel QuickAssist Technology (Intel QAT)

7.1 Overview

The Intel® QuickAssist Technology (Intel® QAT) v2.0 accelerator within the SoC provides acceleration functions that can be used by the IA cores. These acceleration functions include cryptographic functions, compression/decompression, and public key functions.

The IA core(s) access the acceleration services via a standard PCIe* interface and standard PCIe driver. Applications running on the IA core(s) can make use of the acceleration by calling the Intel® QuickAssist Technology Application Programming Interfaces (APIs) in a lookaside coprocessor model. The APIs communicate with the Intel® QAT v2.0 accelerator hardware via PCI configuration space access and assisted rings stored in system memory.

7.2 Intel QuickAssist Technology Features

The Intel® QuickAssist Technology (Intel® QAT) v2.0 supports the following features:

- Symmetric cryptographic functions
  - Cipher operation
  - Hash/authenticate operation
  - Cipher-hash combined operation
  - Key derivation operation
  - Wireless authentication operation
  - Wireless cryptography
- Public Key Functions
  - Rivest–Shamir–Adleman (RSA) operation
  - Diffie-Hellman operation
  - Digital signature standard operation
  - Key derivation operation
  - Elliptic curve cryptography: ECDSA* and ECDH*
- Compression/Decompression
  - Deflate
  - LZ4s
  - LZ4

8.0 Power Controller Unit (PCU) Functional Description

The processor supports the following features for power, thermal and remote management of the processor.

- Power Controller Unit (PCU) dedicated controller that provides power and thermal management for the processor.
- Thermal Management features in both the core and uncore.
- Running Average Power Limit (RAPL)

The processors Power Control Unit (PCU) is a dedicated controller that provides power and thermal management for the processor. The PCU consists of a dedicated micro-controller, ROM and RAM for Pcode (PCU microcode), HW state machines, I/O registers for interfacing to the micro-controller and interfaces to the hardware units in the processor. There are two masters to the power management architecture, the operating system and HW agents in the platform. HW agents can be Intel® Management Engine (Intel® ME) in the PCH, BMC, or the hardware monitor in workstation segment. HW agents may override the OS requests when in conflict. For example; if the OS requests a P-state which exceeds the power limit set by the Intel® ME, it will be ignored. The operating system and platform interact with PCU through a set of registers. Operating system interactions are usually accomplished through MSRs in the core. Platform interactions with PCU will translate to CSRs in the uncore.

Summary of the PCU features:

- Manage ACPI requests for P-State and C-State changes from OS and HW
- Manage S-state requests from PCH
- Uncore Frequency Scaling
- Policy control for Intel® UPI, PCIe, and DMI3
- CPU Thermal and Power Optimization Capabilities
- DRAM Thermal and Power Optimization Capabilities
- Provide uncore CSR access and Machine Check Bank access
- RAPL “Running Average Power Limit” DRAM and Socket requests either via BMC (over PECI electrical interface or MCTP over PCIe) or Node Manager (through management engine on PCH)
- Respond to over power and over current conditions
- Thermal responses to over temperature conditions for processor and memory
- Hardware Power Management (HWP) Support
- Active Idle Efficiency Mode

8.1 PMax Detector

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor implements a maximum power detection circuit on package. The PMax Detector circuit provides faster detection and response to PMax level load events. Instead of relying on the use of the processor's PROCHOT# signal to be asserted by detection circuits that resided in either the PSU or on the system board, this detection circuit will primarily reside on the processor side with some system side implementation required. Refer to *Eagle Stream Platform Design Guide* and *Eagle Stream Platform Electrical Specification* for further details on system side implementation and further details. In general, the PMax detection circuit allows for a (~10x) faster PMax detection and response time as compared to the old PMax detection methods.

8.2 Hardware-Controlled Performance States (HWP)

The processor supports Hardware-Controlled Performance States (HWP), which autonomously selects performance states while utilizing OS supplied performance guidance hints.

HWP is an implementation of the ACPI-defined Collaborative Processor Performance Control (CPPC), which specifies that the platform enumerate a continuous, abstract unit-less, performance value scale that is not tied to a specific performance state / frequency by definition. While the enumerated scale is roughly linear in terms of a delivered integer workload performance result, the OS is required to characterize the performance value range to comprehend the delivered performance for an applied workload.

8.3 Intel Speed Select Technology

Intel® Speed Select Technology (Intel® SST) is a set of performance enhancement/ optimization features comprising of 4 different features to use under various workload scenarios. The four features that make up Intel® Speed Select Technology are:

- Intel® Speed Select Technology – Performance Profile (Intel® SST-PP)
  - — Intel® SST-PP offers up to three different performance profiles within the same CPU. Each performance profile may be defined with different core counts, TDP, base frequency, turbo frequency, and so on.
- Intel® Speed Select Technology – Core Power (Intel® SST-CP)
  - — Intel® SST-CP allows biasing power budget amongst cores. Intel® SST-CP can be used to direct budget to highest priority or bottleneck cores, improving overall performance.
- Intel® Speed Select Technology – Base Frequency (Intel® SST-BF)
  - — Intel® SST-BF enables users to increase base frequency on certain cores (high priority cores) in exchange for lower base frequency on remaining cores (low priority cores). It improves overall performance by boosting frequency on critical cores.
- Intel® Speed Select Technology – Turbo Frequency (Intel® SST-TF)
  - — Intel® SST-TF allows software to assign prioritization; some cores can be granted higher turbo frequencies by restricting other, lower priority cores to a lower frequency, leading to improved overall performance by boosting frequency of the bottleneck task.

8.3.1 Terminology

The next table lists all terminologies used in this document.

Table 8. Acronyms

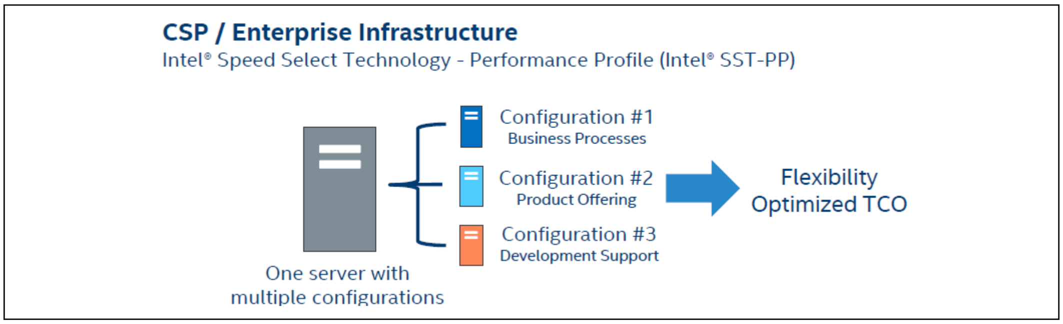
Term	Description
Intel® AVX	Intel® Advanced Vector Extensions
Intel® AVX-512	Intel®Advanced Vector Extensions 512
CLOS	Class of Service
CPU	Central Process Unit
CSP	Cloud Service Provider
CSR	Control and Status Register
HGS	Hardware Guided Scheduling
HP	High Priority cores
HPC	High Performance Computing
HP-TRL	High Priority Core Turbo Ratio Limit
HWP	Hardware-Controlled Performance states
Iaas	Infrastructure as a service
Intel® SST	Intel® Speed Select Technology
Intel® SST-BF	Intel® Speed Select Technology - Base Frequency
Intel® SST-CP	Intel® Speed Select Technology - Core Power
Intel® SST-PP	Intel® Speed Select Technology - Performance Profile
Intel® SST-TF	Intel® Speed Select Technology - Turbo Frequency
LP	Low Priority core
LP-limit	Max P-state for Low Priority core
MSR	Model Specific Register
OS	Operating System
Saas	Silicon as a Service
SSE	Streaming SIMD (Single Instruction Multiple Data) Extensions
SW	Software
TCO	Total Cost of Ownership
TDP	Thermal Design Power
TRL	Turbo Ratio Limits
VM	Virtual Machine
VMM	Virtual Machine Monitor

8.4 Intel Speed Select Technology – Performance Profile

8.4.1 Introduction

The Intel® Speed Select Technology - Performance Profile (Intel® SST-PP) feature was conceived to address flexibility needs in various frequency profiles that the processor enumerates and operates at. Intel® SST-PP allows up to 3 different performance profiles or configurations within the same CPU.

Figure 9. One Server with Multiple Configurations

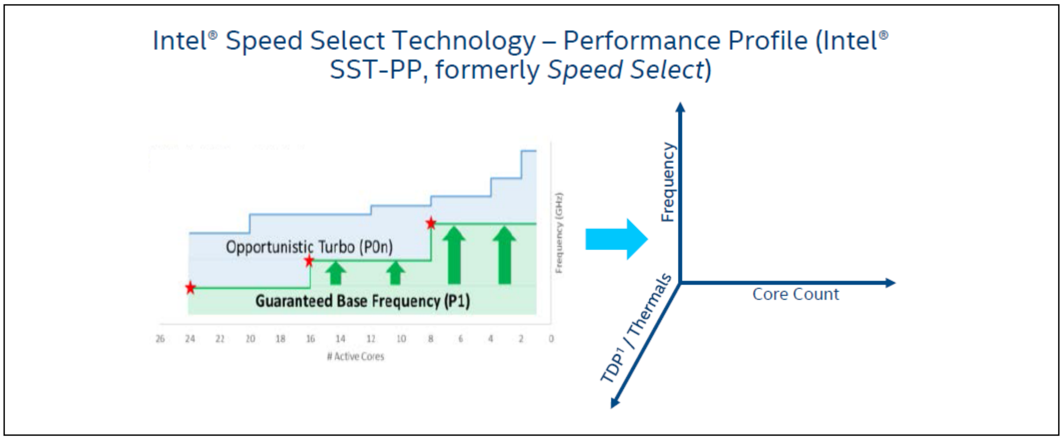


Each performance profile is characterized by any/all of (Base Frequency, Core Count, TDP, Thermals, Turbo Frequency, and so on).

Generally, each configuration can be viewed as a function of various parameters.

Config_n = Fn (Core_mask, TDP_n, Tprochot_n, P1_n, Turbo curves).

Figure 10. Intel® SST-PP with Core Count, TDP/Thermal and Frequency


Intel® Speed Select Technology – Performance Profile (Intel®
SST-PP, formerly *Speed Select*)
Opportunistic Turbo (P0n)
Guaranteed Base Frequency (P1)
Frequency (GHz)
26 24 22 20 18 16 14 12 10 8 6 4 2 0
# Active Cores
Frequency
Core Count
TDP[1] / Thermals

Each profile is discoverable along with all characteristics via the OS mailbox interface provided by the CPU. The OS Mailbox interface is exposed via MSR and CSR interfaces in the processor.

The principal use cases for Intel® SST-PP are:

- Deployment Consolidation.
- Customer buying multiple SKUs to cover deployment space (HPC, Networking, Search, and so forth) and motivated to consolidate.
- IaaS VM Flexibility.

- Ability to offer multiple VMs with different performance levels using a homogenous hardware environment.

8.4.2 Usage models

In terms of deployment, there are two possible options:

- Static - BIOS discovery and configuration. BIOS may lock the configuration to a selected level such that OS level software cannot switch the configuration at runtime.
- Dynamic - Performance levels (also called performance profile) can be switched dynamically. This relies on support from OS/software to discover and off-line the cores according to the core mask.

8.4.3 Hardware Guided Scheduling

Hardware Guided Scheduling (HGS) is a general-purpose infrastructure for abstracting information available to hardware, and making it available as guidance to the OS scheduler. Hardware Guided Scheduling has wide range of applicability across multiple product segments under multiple use cases - Accelerator/FPGA scenarios, Core asymmetry in thermal/speed/reliability etc.

The fundamental idea behind HGS is that:

- Hardware has significantly more "information" about the system, such as, core types, performance characteristics, configuration information, side-band configuration requests (such as Intel® SST-PP), thermal limitations, etc.
- OS has more control over scheduling: number of active cores, which tasks/threads map to which logical processor etc.
- Optimal performance requires combining the above two.

HGS provides a general purpose infrastructure "conduit" for hardware to express abstracted guidance to the OS scheduler. This infrastructure can then be used by multiple features and use cases. Intel® SST-PP is one of the use cases.

8.5 Intel Speed Select Technology - Base Frequency

Intel® SST Base Frequency (Intel® SST-BF) is a feature targeted to support a higher base frequency for some cores as long as power, Iccmax, thermal, and reliability constraints are met. This is achieved by a CSP* software, such as an Orchestrator*, to partition available cores into high and low priority buckets with respective P1 (base frequencies). This allows higher priority VMs or tasks/applications to be assigned to higher-priority cores, for better performance with less performance jitter. Intel® SST-BF has been referred to as Platform Base Frequency in the past.

8.5.1 Intel SST-BF overview

Intel® SST-BF enables users to increase the guaranteed base frequency on certain cores (high priority cores) in exchange for a lower base frequency on the remaining cores (low priority cores). It improves overall performance by boosting frequency on critical cores.

The SW discovers Intel® SST-BF levels and core masks, and is responsible to set min correctly via HWP or CLOS interface. After these settings are completed, the SW enables Intel® SST-BF via mailbox.

8.5.2 Changes over the Second Generation Intel Xeon Processor Scalable Family version of Intel SST-BF

There is no change on the HWP interface. For the CLOS interface, the SW will set the min frequency of each CLOS via PM_CLOS, and assign each core to one of the CLOS groups via PQR_ASSOC.

When using Ordered Throttling prioritization type, the CLOS group number implicitly indicates priority order. CLOS_0 is the highest priority CLOS. The cores subscribed to CLOS_0 are the highest priority ones. It is recommended that the Intel® SST-BF high priority cores should be assigned to the CLOS_0 group.

SW has the option to select a subset of Intel® SST-BF high priority cores, assign P1_Hi frequency to them in the min setting, and leave other Intel® SST-BF able high priority cores running at lower frequency. Intel® SST-BF high frequency is achieved using the right min levels and the bonus is on the software to program the Intel® SST-BF settings as per mailbox discovery.

When Intel® SST-BF is disabled during runtime, the SW is responsible to adjust the min settings.

SW should resample REF_TEMP in MSR 0x1A2 or PECI PCS index 16 when Intel® SST-BF is enabled/disabled during runtime.

8.5.3 Intel SST-BF and Intel SST-PP Interactions

Note that if the Intel® SST-PP level is dynamically changed while Intel® SST-BF is enabled, the Intel®SST-BF is disabled. The Intel® SST-BF discovery and configuration needs to be done again. When there is an Intel® SST-PP config level switching, CPU internally disables Intel® SST-BF if it was enabled. However, the mailbox interface shows Intel® SST-BF is still enabled for the previous level.

8.5.4 Intel SST-BF and Intel SST-TF cross Interactions

Follow the rules to make Intel® SST-TF and Intel® SST-BF work together:

1. Intel® SST-BF High Priority (HP) cores should be selected as Intel® SST-TF HP cores by software. SW discovers the Intel® SST-BF HP cores and aligns the Intel® SST-BF/Intel® SST-TF priorities.
2. Intel® SST-TF Low Priority (LP) clipping value SSE >= P1Lo.
3. Intel® SST-BF LP cores can be Intel® SST-TF HP or LP cores.

Table 9. Intel® SST-BF and Intel® SST-TF Interoperation

Intel® SST-TF/Intel® SST-BF Priority	Low (Intel® SST-BF)	High (Intel® SST-BF)
Low (Intel® SST-TF)	Rule 2 above ensures P1Lo is not violated	Not allowed (rule 1 above) May lead to violation of P1_Hi
High (Intel® SST-TF)	Allowed (rule 3)	Allowed (rule 3)

Essentially, an Intel® SST-BF HP core must not be an Intel® SST-TF LP core; otherwise, P1_Hi cannot be guaranteed. All other combinations are valid as shown in the next table.

Table 10. Intel® SST-BF and Intel® SST-TF Core Priority Setting

Intel® SST-TF	Intel® SST-BF	Valid
HP	HP	Yes
HP	LP	Yes
LP	LP	Yes
LP	HP	No

8.6 Intel Speed Select Technology - Turbo Frequency

8.6.1 Introduction

Intel® Speed Select Technology - Turbo Frequency (Intel® SST-TF) sets different Turbo Ratio Limits (TRL) for cores based on relative priority. Through the CLOS interface, cores are assigned to bins with priority values which determine if each core is high or low priority. High Priority (HP) cores are granted a turbo ratio limit higher than P0n (HP-TRL), based on the number of HP cores (as well as the number of active cores), while Low Priority (LP) cores are restricted to frequencies below a set ratio limit (LP-limit).

8.6.2 Architecture Implementation and Interfaces

Intel® SST-TF operates via a set of SKU configurations and tables with a software interface via OS mailbox:

- Input
  — Number of high priority cores
  — ICCP license level
- Output
  — Turbo Ratio Limit for high-priority cores (HP-TRL)
  — Max P-State for Low Priority cores (LP-limit)

8.6.3 Intel SST-TF Table

The SKU-specific Intel® SST-TF tables determine the ratio limits for High and Low Priority cores. Tables are used for simplification and determinism, reducing turbo variation. Each table is referred to as an Intel® SST-TF configuration, one of which is active at a given time, determined by the Intel® SST-PP configuration that is active. The table assumes that all cores are active (in C0) but is still used independent of the number of active cores.

Every Intel® SST-TF configuration contains buckets identified by number of HP cores. Each bucket is indexed by ICCP license level, and yields the following values:

- Turbo ratio limit for High Priority cores (HP-TRL)
- Clipping value for Low Priority cores (LP-limit)

8.6.4 Software Interface

In the 4th Generation Intel® Xeon® Scalable Processor, Intel® SST-TF has a configuration / discovery interface through the OS mailbox. Intel® SST-TF needs to be enabled both by the SKU and via the OS mailbox user interface, meaning, Intel® SST-TF can only be enabled via user interface if Intel® SST-TF is enabled for the SKU.

Intel® SST-TF will not have any new writable configuration interface, except enable/ disable. Core prioritization will be configured through the CLOS interface. The user can change the TRL through the TRL MSR or PECI interface (Legacy). If the user configures the TRLs, These limits will apply to Intel® SST-TF as well (see the later chapter for details).

8.7 Intel Speed Select Technology – Core Power

8.7.1 Introduction

Intel® Speed Select Technology - Core Power (Intel® SST-CP) is the fundamental building block that enables the Intel® SST-BF and Intel® SST-TF features. It is also a standalone feature on its own that can be used without Intel® SST-BF or Intel® SST-TF. Intel® SST-CP is also known as "RAPL Prioritization".

In the 1st and 2nd Generation Intel® Xeon® Scalable processors, Intel® SST-CP can be enabled/disabled during boot time via BIOS setup menu and is only supported via the HWP interface. Starting from the 3rd Generation Intel Xeon Scalable Processor variants, Intel® SST-CP can be controlled at runtime via OS mailbox interface and supported via both HWP and CLOS interfaces.

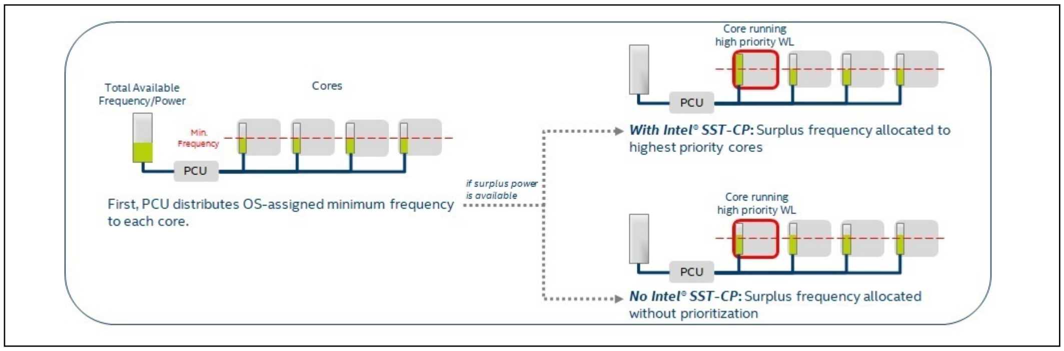
If Intel® SST-BF or Intel® SST-TF is supported by the CPU and enabled by SW/BIOS, the underlying RAPL prioritization flow automatically runs and does not require SW/ BIOS to explicitly enable it.

At a high level, Intel® SST-CP allows OS/VMM to assign a weight (or priority) to each core. When surplus frequency budget is available, CPU distributes surplus frequency according to cores' weights. As a contrast, without the Intel® SST-CP feature, the CPU distributes surplus power equitably among cores.

Intel® SST-CP can be used to direct the frequency budget to highest priority or bottleneck cores, improving the overall performance. Intel® SST-CP can be enabled/ disabled dynamically during runtime, the cores' weight (priority) can be dynamically changed as well.

The picture below shows how frequency budget is allocated in a CPU with and without Intel® SST-CP.

Figure 11. CPU Frequency Budget Distribution


Total Available Frequency/Power
Cores
Min. Frequency
PCU
First, PCU distributes OS-assigned minimum frequency to each core.
if surplus power is available
Core running high priority WL
PCU
With Intel® SST-CP: Surplus frequency allocated to highest priority cores
Core running high priority WL
PCU
No Intel® SST-CP: Surplus frequency allocated without prioritization

Use case examples:

- HPC: Power-aware balancing software, like the Global Extensible Open Power Manager (GEOPM), will identify critical path threads on-the-fly and direct operating power to critical threads. SST-CP enables faster rebalancing and speeds up the overall workload run time.
- SW Encryption: SW encryption often creates a 1-core bottleneck. Intel® SST-CP can be used to direct frequency to this bottleneck, improving overall system performance.

8.7.2 Overview

Starting from the 3rd Generation Intel® Xeon® Scalable Processor variants, a new interface named CLOS (Class Of Service) is introduced that allows frequency prioritization (biasing) between cores. This interface supports three different features that are branded under the Intel® SST.

Intel® SST-CP replaces the legacy PPO-budget name and allows biasing energy/power budget between cores. Intel® SST-BF leverages the Intel® SST-CP capability and provides a higher base, guaranteed frequency for HP cores. Intel® SST-TF allows a higher TRL for HP cores.

Prioritization through the legacy HWP interface is much more limited in capabilities. Legacy functionality is still supported in the 4th Generation Intel® Xeon® Scalable Processor variants; it is however recommended to use CLOS interface for prioritization with these platforms because there are new capabilities introduced through CLOS. Some examples are the new prioritization schemes/types, Turbo Prioritization (Intel® SST-TF), and so on.

Two notions of priority are available: Proportional Prioritization, and Ordered Throttling Prioritization. Ordered Throttling prioritization is supported through CLOS interface only. Proportional prioritization is supported through both the legacy HWP and the CLOS interfaces. The PRIORITY_TYPE field in QOS_CONFIG indicates the preference of priority type (Proportional vs Ordered Throttling).

Intel® SST-TF is a new feature available starting in Cooper Lake, and will only be supported through CLOS interface. This processor supports both Proportional and Ordered Throttling prioritization through CLOS. On the processor, Intel® SST-BF and Intel® SST-TF can operate on both prioritization types.

9.0 Reliability, Availability, and Serviceability (RAS) Functional Description

This chapter describes Reliability, Availability, and Serviceability (RAS) features of 4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor on Eagle Stream platform.

RAS refers to feature sets that are associated with the system resiliency in the presence of hardware faults and is defined as follows.

- Reliability: System capability to detect errors, correct errors, and report errors. Reliability is typically measured in FITs (Failure in Time). 1 FIT = 1 error in 1 Billion Hours. Alternatively it can also be measured as MTTF (Mean Time To Fail). 1FIT is equivalent to approximately 114K Years of MTTF (MTTF= 1/FIT).
- Serviceability: System capability to effectively report a failure with precise location of the faulty component to expedite the servicing efforts. Incorporating serviceability features typically results in more efficient product maintenance, reduces operational costs, and results in minimizing the duration of “service downtime”. It is typically measured as Mean Time To Repair (MTTR).
- Availability: System capability to maintain “service availability” in the presence of system faults. Availability is typically measured as a function of “Reliability” and “Serviceability”: Availability = (1/ (1 + MTTR/MTTF))x100%. A highly available system is capable of redistributing (or reallocating) resources to maintain normal operation if a fault is detected and is expected to have 99.99% or higher Availability, for example, if MTTF= 1 year, then service down-time should be less then 52 min.

In order to meet the target reliability requirements and to minimize the impact of the errors, the processor incorporates various techniques such as parity, ECC, CRC, and redundancy within individual modules. Most of these errors are corrected by the built-in error correction logic and are called as “Corrected Errors”. The processor also incorporates *Corrupt Data Containment* (aka. Data poisoning) feature to help contain the impact of the detected error, and provide a path for the software to recover from the erroneous data. Such Data errors fall into the “Uncorrected Recoverable” (UCR) error category.

In order to meet the target serviceability requirements and to minimize the duration of downtime, the processor incorporates various error reporting techniques such as Machine Check Architecture (MCA) based error reporting within core and system interface logic in the uncore including the Advanced Error Reporting (AER) within the PCI Express* sub-system, and additional processor specific error reporting within memory sub-system, Intel® UPI interface, and IIO sub-system.

4th Gen Intel® Xeon® Processor Scalable Family, Codename Sapphire Rapids processor RAS features are offered within two categories of processor SKUs, “standard RAS” or “advanced RAS” features. Processor CAPID configuration registers reflect the RAS SKU in use.

Standard RAS features are available across all SKUs, feature listed in Table 11 on page 65. Advanced RAS features are only available in the Advanced RAS SKU processors.

Platform RAS is further divided per sub-functions:

1. Embedded RAS within CPU (core, caches, mesh).
2. Memory: RAS features incorporated within the processor's memory subsystem interface to DDR (M2M, IMC, DDR). May require run-time UEFI FW to take advantage of the RAS to react to the detected failing unit, but does not "require" any OS/SW support during run-time.
3. Intel® Ultra Path Interconnect (Intel® UPI): RAS features that are incorporated within the processor's Intel® UPI interconnect sub-system and may require run-time UEFI FW, but does not require any OS/SW support during run-time
4. IIO: RAS features that are incorporated within the processor's PCI Express* , Compute Express Link. Some RAS require run-time UEFI FW, and recovery from eDPC errors requires OS/driver support.
5. System RAS features such as require OS/SW support in addition to the UEFI FW during run-time.

Table 11. Standard RAS Feature List

Category	Platform RAS Feature
CPU	Error Detection and Correction (Coverage at socket level)
CPU	Corrupt Data Containment (aka. Data poisoning)
CPU	Advanced Error Detection and Correction (AEDC)
CPU	DCU/IFU Error Handling Enhancement
CPU	DCU Scrubbing
CPU	LLC DECTED, DBF Reporting
CPU	Time-out timer Schemes
CPU	Error reporting (MCA, AER) – core, Uncore, and IIO
CPU	Error reporting through MCA 2.0 (EMCA Gen2)
CPU	Processor BIST
CPU	Error reporting via IOMCA (PCIe/Compute Express Link)
CPU	MCA Bank Error Control
CPU	First Corrected Error Mode of Error Reporting
CPU	PCI Express Corrected Error Reporting (inclusive of Corrected Error Counter and Leaky-bucket)
CPU	Thresholding for Corrected Errors (Intel® UPI, PCIe, and CSMI threshold for system interface)
CPU	CSR Error Log Cloaking (using DEVHIDE)
Memory	DDR DRAM Memory Single Device Data Correction (SDDC)
Memory	DDR Link Command/Address Parity Check and Retry
Memory	DDR Memory Data Scrambling with Command and Address
Memory	DDR Memory Demand and Patrol Scrubbing
Memory	DDR Memory Thermal Throttling (Thermal event)
Memory	DDR Memory Mirroring – intra IMC
Memory	Adaptive Data Correction - Single Region (ADC - SR)

continued...

Category	Platform RAS Feature
Memory	DDR DRAM Power up Post Package Repair (PPR)
Memory	DIMM Mem SMBus hang recovery
Memory	Partial Cache Line Sparing (PCLS). Supported with HBM only, not supported with DDR5.
Memory	DRAM Memory (Bank/Rank/DIMM) disable/map-out for FRB
Memory	DRAM Memory MEMHOT Pin support for thermal event throttling, (Thermal event)
Memory	DDR Memory corrected error reporting (via Per rank counters, and corrected error counters in McBank.
Memory	DDR Write Data CRC Check and Retry (lab and debug use only)
Memory	HBM- Bank Sparing
Intel® UPI	Intel® UPI Protocol Protection via CRC (16 bit)
Intel® UPI	Intel® UPI Link Level Retry
Intel® UPI	Intel® UPI failing lane isolation
IIO	PCI Express Link Retraining and Recovery
IIO	PCI Express Link CRC Error Check and Retry
IIO	PCI Express Corrupt Data Containment (Data Poisoning)
IIO	PCI Express ECRC
IIO	PCI Express Enhanced Downstream Port Containment
IIO	PCI Express Card Hot-Plug (Add/Remove/Swap)
IIO	PCI Express Card Hot-Plug Surprise
IIO	Compute Express Link Retraining and Recovery
IIO	Compute Express Link CRC Error Check and Retry
IIO	Compute Express Corrupt Data Containment (Data Poisoning)
IIO	Compute Express ECRC
System	Failed DIMM Isolation
System	OOB access to Error logs
System	PECI Access Enhancements
System	Socket Disable for FRB (Not supported)
System	Core Disable for FRB
System	Enhanced SMM (ESMM)
System	Error Injection Capability
System	Predictive Failure Analysis
System	Asynchronous Warm Reset (AWR) for Error Log Capturing

This document primarily discusses RAS pertaining to processor, and refers to platform RAS through supporting collaterals.

10.0 Psys Close Loop Platform Power Management

Psys is a closed loop control of platform power. The entire platform power is sensed by the Psys sensor, sent to the CPU and managed by CPU Firmware. This allows fast, accurate and efficient platform power control.

10.1 Psys Benefits

Modern data center designs cannot tolerate the cost of provisioning power based on peak utilization conditions. Consequently, installed power capability must be protected from exceeding its capability. Data center orchestration software achieves this protection through platform level power capping where CPU and/or Memory is throttled to bring total system power within provisioned power bounds. Essential to power capping is total system level power measurement.

The Psys capability is an improved capability which can replace Intel® Node Manager (NM) Firmware running on the Intel® Manageability Engine or the orchestration SW itself.

Psys provides improved method for server platform power management and Capping by:

- Migrating the control calculation to CPU Hardware/Firmware for significantly more accurate power sensing and faster response time to mitigate overpower.
- Standardizing the power measurement hardware and the flow of information across platform components
- Enabling total platform power budget sharing between the different platform components.
- The above will result in the following enhancements:
- Increase Server Rack Density and Utilization - Tighter control over Server's power consumption allows for higher rack density.
- Potential to reduce TCO – by allowing removing a portion of redundant PSU Capacity

Intel® Server Platforms that implement this feature for power capping will be able to operate under a tighter peak power envelope. This in turn allows data center operators to proportionately increase their node density in their rack or blade designs, and/or decrease overall total cost of ownership (TCO) and provide additional value to Intel® customers.

Glossary

Intel® TME-MK

Intel® Total Memory Encryption - Multi-Key (Intel® TME-MK) is Intel's implementation of full-memory encryption.

IMC

integrated Memory Controller. A Memory Controller that is integrated in the processor die.

Intel® SGX

Intel® Software Guard Extensions (Intel® SGX) is a set of instructions that increases the security of application code and data, giving them more protection from disclosure or modification. Developers can partition sensitive information into enclaves, which are areas of execution in memory with more security protection.

Intel® Turbo Boost Technology

A feature that opportunistically enables the processor to run a faster frequency. This results in increased performance of both single and multi-threaded applications.

Intel® UPI

Intel® Ultra Path Interconnect. A new cache-coherent, link-based Interconnect specification for Intel processors, chipsets, and I/O bridge components.