INTEL® XEON® SCALABLE PROCESSOR ARCHITECTURE DEEP DIVE

Akhilesh Kumar, Skylake-SP CPU Architect

Malay Trivedi, Lewisburg PCH Architect

June 12th, 2017

Notices and Disclaimers

Agenda

- Intel® Xeon® Scalable Processor Overview
- Skylake-SP CPU Architecture
- Lewisburg PCH Architecture

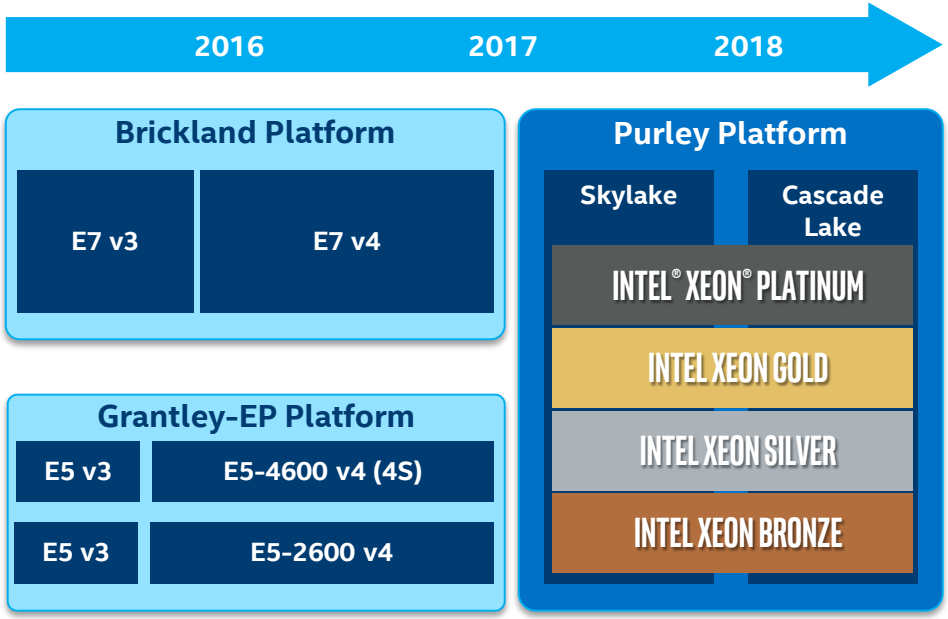
Intel® Xeon® Processor Roadmap

Intel® Xeon® Processor E7

Targeted at **mission critical** applications that value a **scale-up** system with leadership **memory capacity** and **advanced RAS**

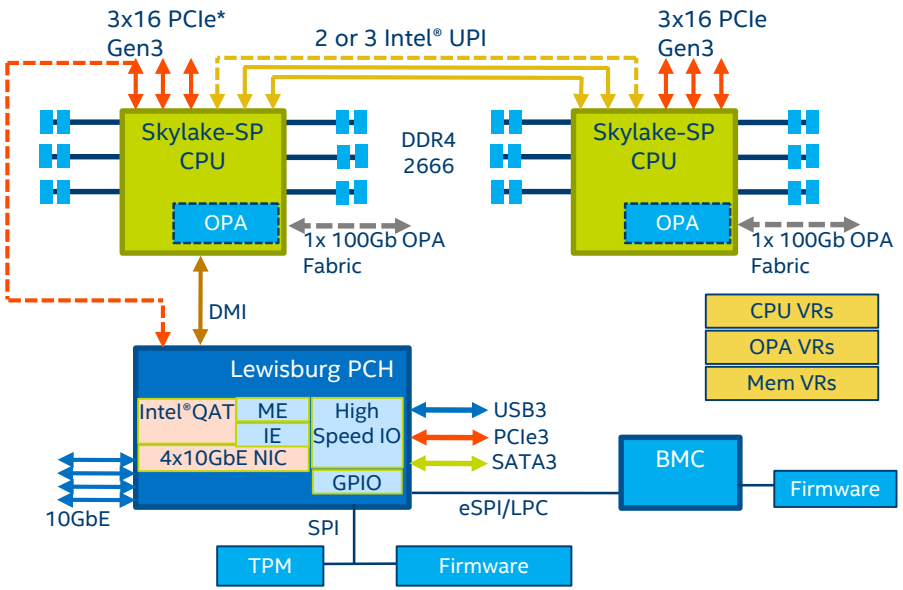
Intel® Xeon® Processor E5

Targeted at a wide variety of applications that value a **balanced system** with **leadership** performance/watt/$



CONVERGED PLATFORM WITH INNOVATIVE SKYLAKE-SP MICROARCHITECTURE

Intel® Xeon® Scalable Processor Feature Overview

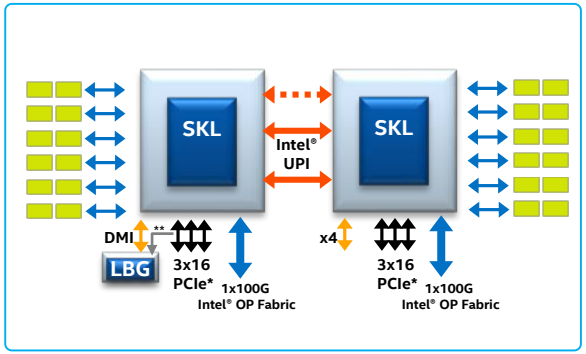


BMC: Baseboard Management Controller	PCH: Intel® Platform Controller Hub	IE: Innovation Engine
Intel® OPA: Intel® Omni-Path Architecture	Intel QAT: Intel® QuickAssist Technology	ME: Manageability Engine
NIC: Network Interface Controller	VMD: Volume Management Device	NTB: Non-Transparent Bridge

Feature	Details
Socket	Socket P
Scalability	2S, 4S, 8S, and >8S (with node controller support)
CPU TDP	70W – 205W
Chipset	Intel® C620 Series (code name Lewisburg)
Networking	Intel® Omni-Path Fabric (integrated or discrete) 4x10GbE (integrated w/ chipset) 100G/40G/25G discrete options
Compression and Crypto Acceleration	Intel® QuickAssist Technology to support 100Gb/s comp/decomp/crypto 100K RSA2K public key
Storage	Integrated QuickData Technology, VMD, and NTB Intel® Optane™ SSD, Intel® 3D-NAND NVMe & SATA SSD
Security	CPU enhancements (MBE, PPK, MPX) Manageability Engine Intel® Platform Trust Technology Intel® Key Protection Technology
Manageability	Innovation Engine (IE) Intel® Node Manager Intel® Datacenter Manager

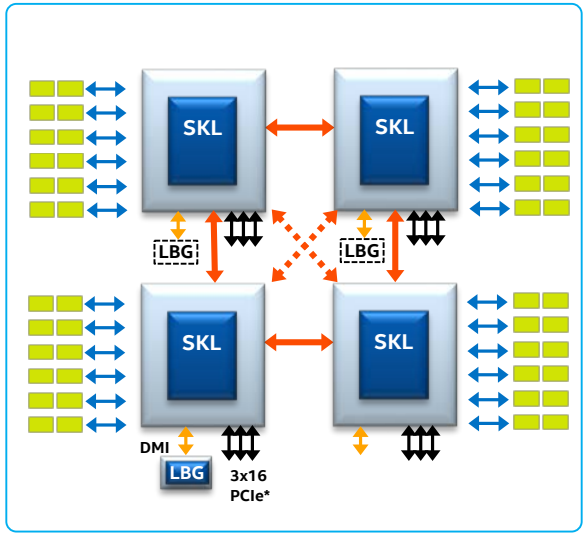
Platform Topologies

2S Configurations



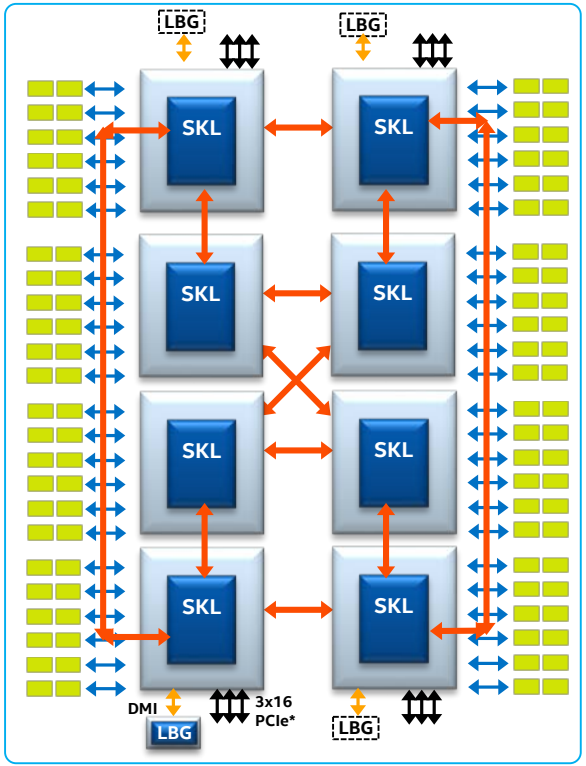
(2S-2UPI & 2S-3UPI shown)

4S Configurations



(4S-2UPI & 4S-3UPI shown)

8S Configuration



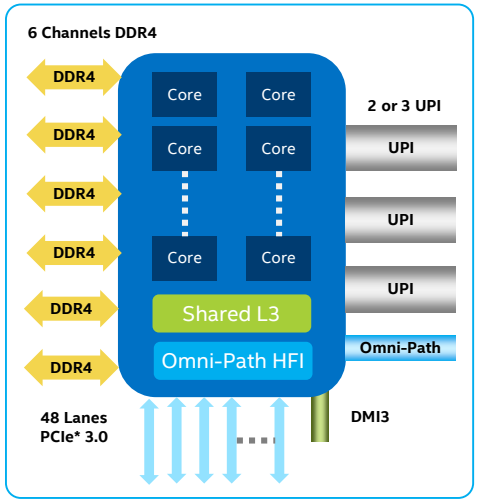
INTEL® XEON® SCALABLE PROCESSOR SUPPORTS CONFIGURATIONS RANGING FROM 2S-2UPI TO 8S

Intel® Xeon® Scalable Processor

Re-architected from the Ground Up

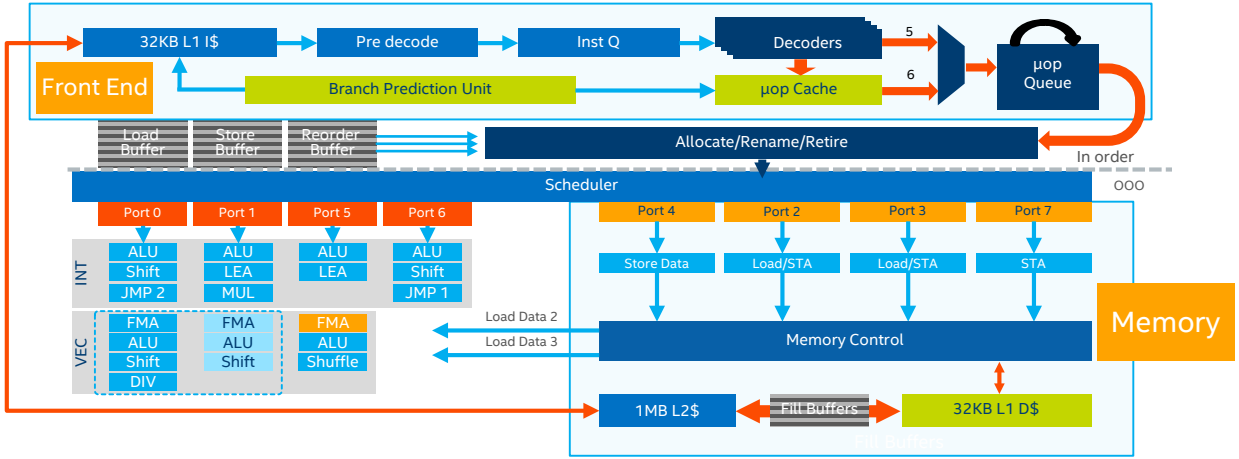
- Skylake core microarchitecture, with data center specific enhancements
- Intel® AVX-512 with 32 DP flops per core
- Data center optimized cache hierarchy – 1MB L2 per core, non-inclusive L3
- New mesh interconnect architecture
- Enhanced memory subsystem
- Modular IO with integrated devices
- New Intel® Ultra Path Interconnect (Intel® UPI)
- Intel® Speed Shift Technology
- Security & Virtualization enhancements (MBE, PPK, MPX)
- Optional Integrated Intel® Omni-Path Fabric (Intel® OPA)

Features	Intel® Xeon® Processor E5-2600 v4	Intel® Xeon® Scalable Processor
Cores Per Socket	Up to 22	**Up to 28**
Threads Per Socket	Up to 44 threads	**Up to 56 threads**
Last-level Cache (LLC)	Up to 55 MB	**Up to 38.5 MB (non-inclusive)**
QPI/UPI Speed (GT/s)	2x QPI channels @ 9.6 GT/s	**Up to 3x UPI @ 10.4 GT/s**
PCIe* Lanes/ Controllers/Speed(GT/s)	40 / 10 / PCIe* 3.0 (2.5, 5, 8 GT/s)	**48 / 12 / PCIe 3.0 (2.5, 5, 8 GT/s)**
Memory Population	4 channels of up to 3 RDIMMs, LRDIMMs, or 3DS LRDIMMs	**6 channels of up to 2 RDIMMs, LRDIMMs, or 3DS LRDIMMs**
Max Memory Speed	Up to 2400	**Up to 2666**
TDP (W)	55W-145W	**70W-205W**



SKYLAKE-SP CORE ARCHITECTURE

Core Microarchitecture Enhancements



	Broadwell uArch	Skylake uArch
Out-of-order Window	192	**224**
In-flight Loads + Stores	72 + 42	72 + **56**
Scheduler Entries	60	**97**
Registers – Integer + FP	168 + 168	**180** + 168
Allocation Queue	56	**64/thread**
L1D BW (B/Cyc) – Load + Store	64 + 32	**128 + 64**
L2 Unified TLB	4K+2M: 1024	4K+2M: **1536** **1G: 16**

- Larger and improved branch predictor, higher throughput decoder, larger window to extract ILP
- Improved scheduler and execution engine, improved throughput and latency of divide/sqrt
- More load/store bandwidth, deeper load/store buffers, improved prefetcher
- **Data center specific enhancements: Intel® AVX-512 with 2 FMAs per core, larger 1MB MLC**

ABOUT 10% PERFORMANCE IMPROVEMENT PER CORE ON INTEGER APPLICATIONS AT SAME FREQUENCY

Key Instruction Set Architecture Enhancements

COMPUTE

Intel® AVX-512

2x compute density per core for vector operations

Cache Management Instructions

CLFLUSHOPT – Lower latency cache line flush

CLWB – Cache line writeback to memory without invalidation

VIRTUALIZATION

Improved Time Stamp Counter Virtualization

Reduced overhead for VMs moving across CPUs with different base frequency

SECURITY

Page Protection Keys (PPK)

Extends paging architecture to provide a page-granular, thread-private user-level memory protection

Mode Based Execution (MBE)

Protects against malicious kernel updates in a virtualized system

MPX (Memory Protection Extension)

Enables bounds checking on data accesses to prevent buffer overflow attacks

INSTRUCTION SET ENHANCEMENT ACROSS COMPUTE, VIRTUALIZATION, AND SECURITY

Intel® Advanced Vector Extensions 512 (Intel® AVX-512)

- 512-bit wide vectors
- 32 operand registers
- 8 64b mask registers
- Embedded broadcast
- Embedded rounding

Microarchitecture	Instruction Set	SP FLOPs / cycle	DP FLOPs / cycle
Skylake	Intel® AVX-512 & FMA	64	32
Haswell / Broadwell	Intel AVX2 & FMA	32	16
Sandybridge	Intel AVX (256b)	16	8
Nehalem	SSE (128b)	8	4

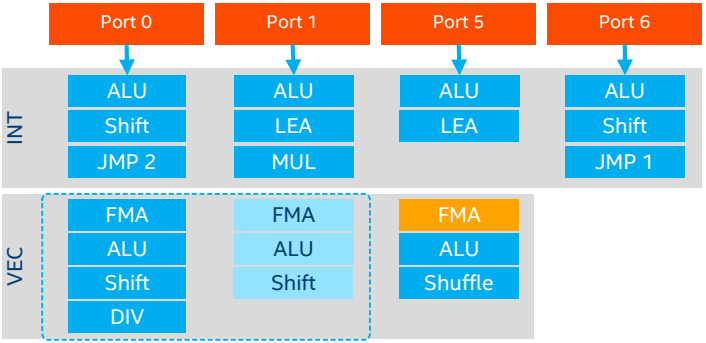
Intel AVX-512 Instruction Types	
AVX-512-F	AVX-512 Foundation Instructions
AVX-512-VL	Vector Length Orthogonality : ability to operate on sub-512 vector sizes
AVX-512-BW	512-bit Byte/Word support
AVX-512-DQ	Additional D/Q/SP/DP instructions (converts, transcendental support, etc.)
AVX-512-CD	Conflict Detect : used in vectorizing loops with potential address conflicts

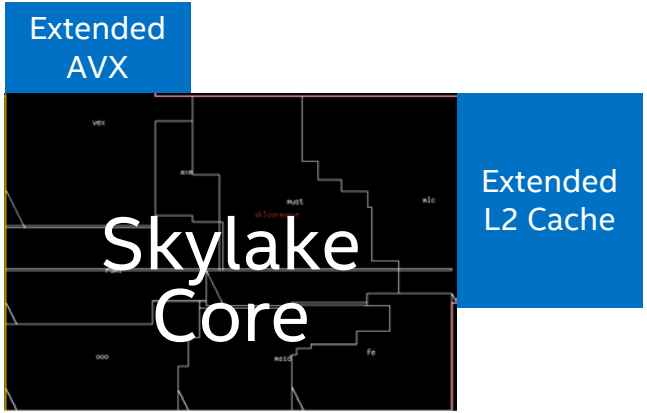
POWERFUL INSTRUCTION SET FOR DATA-PARALLEL COMPUTATION

Skylake-SP Core

Skylake-SP core builds on Skylake core with features architected for data center usage

- Intel® AVX-512 implemented with Port 0/1 fused to a single 512b execution unit
- Port 5 is extended to full 512b to add second FMA outside of Skylake core
- L1-D load and store bandwidth doubled to allow up to 2x64B load and 1x64B store
- Additional 768KB of L2 cache added outside of Skylake core





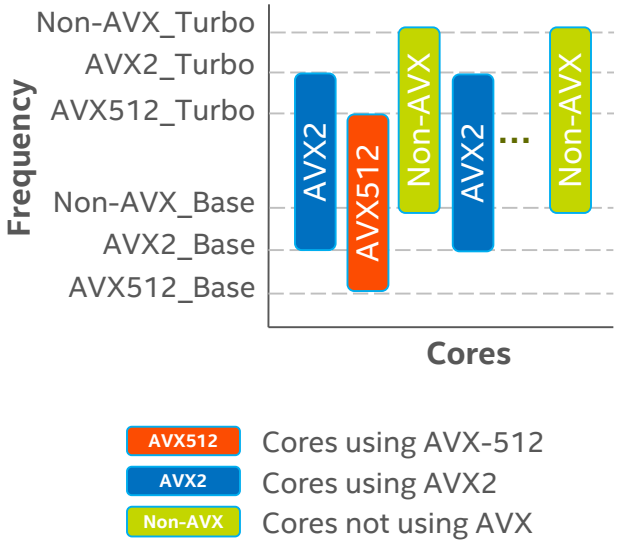
SKYLAKE-SP CORE: OPTIMIZED FOR DATA CENTER WORKLOADS

Frequency Behavior While Running Intel® AVX Code

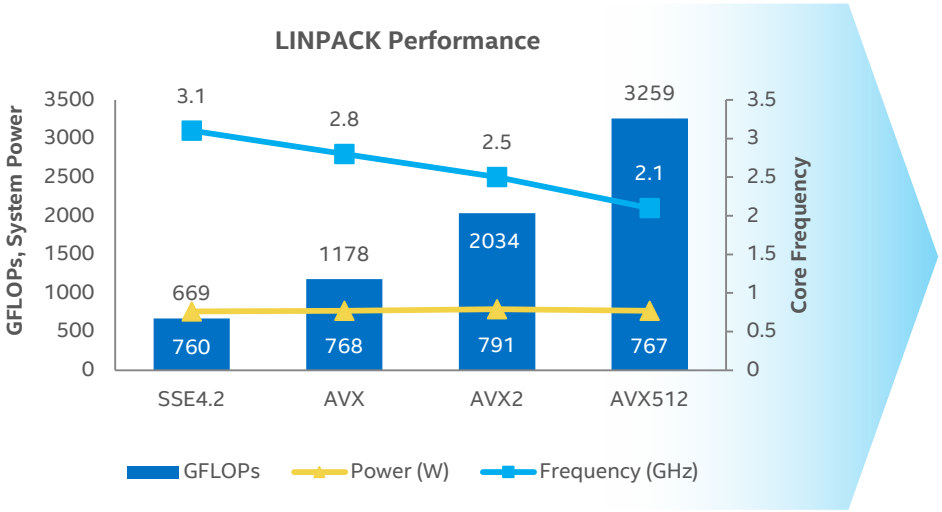
- Cores running non-AVX, Intel® AVX2 light/heavy, and Intel® AVX-512 light/heavy code have different turbo frequency limits
- Frequency of each core is determined independently based on workload demand

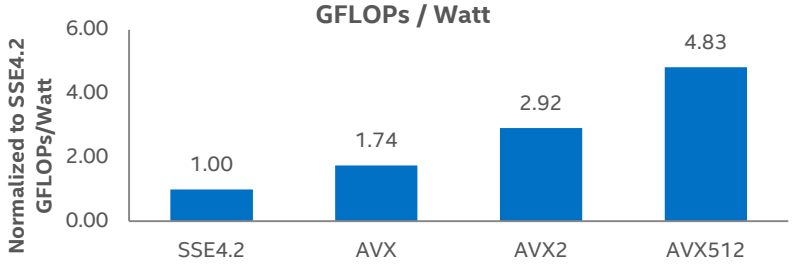
Code Type	All Core Frequency Limit
SSE AVX2-Light (without FP & int-mul)	Non-AVX All Core Turbo
AVX2-Heavy (FP & int-mul) AVX512-Light (without FP & int-mul)	AVX2 All Core Turbo
AVX512-Heavy (FP & int-mul)	AVX512 All Core Turbo

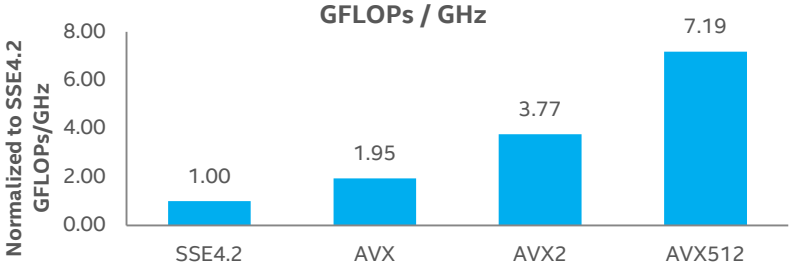
Mixed Workloads



Performance and Efficiency with Intel® AVX-512







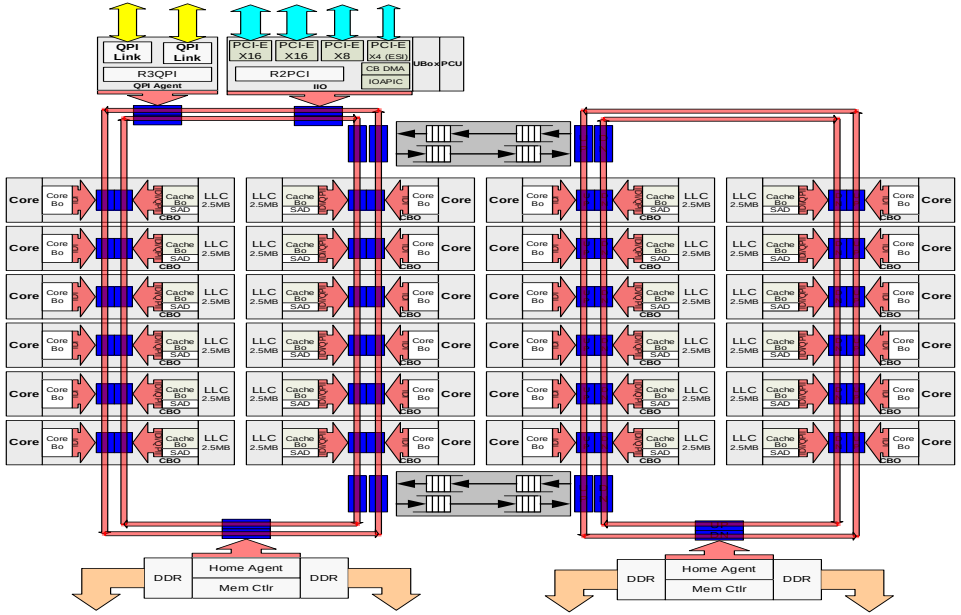
INTEL® AVX-512 DELIVERS SIGNIFICANT PERFORMANCE AND EFFICIENCY GAINS

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2666 per CPU, 1 DPC. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

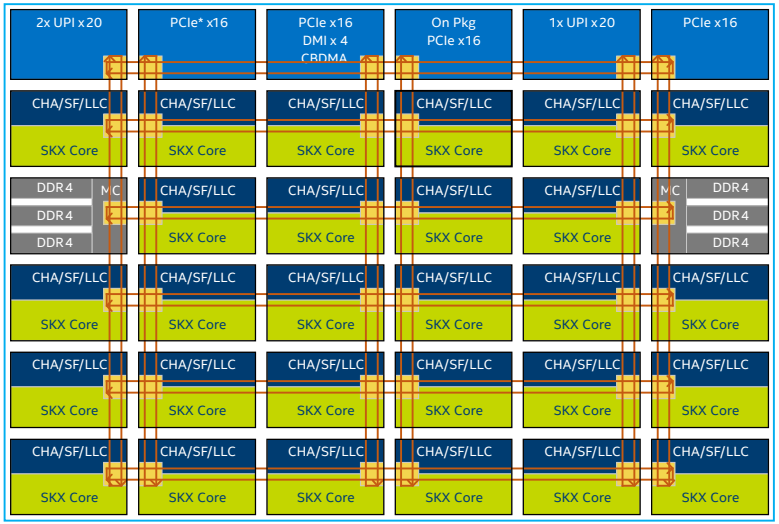
SKYLAKE-SP SOC ARCHITECTURE

New Mesh Interconnect Architecture

Broadwell EX 24-core die



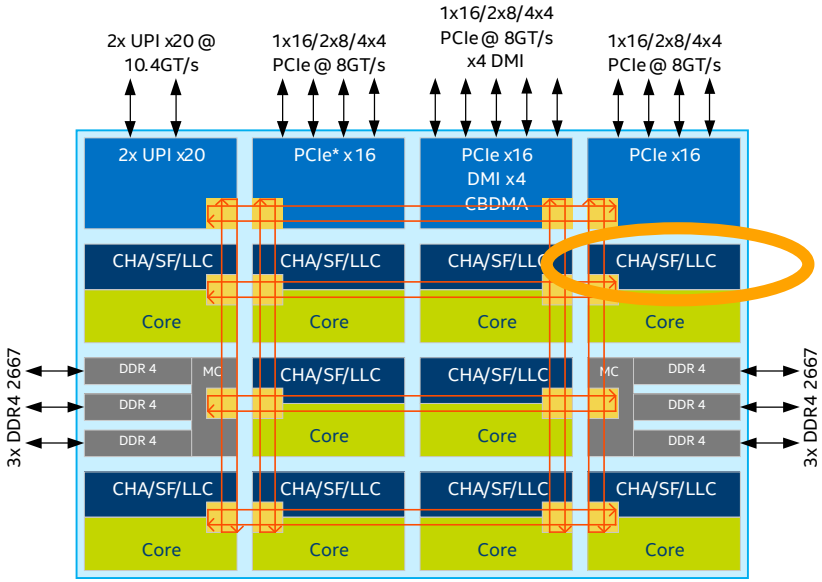
Skylake-SP 28-core die



CHA – Caching and Home Agent ; SF – Snoop Filter; LLC – Last Level Cache;
SKX Core – Skylake Server Core; UPI – Intel® UltraPath Interconnect

MESH IMPROVES SCALABILITY WITH HIGHER BANDWIDTH AND REDUCED LATENCIES

Distributed Caching and Home Agent (CHA)



- Intel® UPI caching and home agents are distributed with each LLC bank
  - Prior generation had a small number of QPI home agents
- Distributed CHA benefits
  - Eliminates large tracker structures at memory controllers, allowing more requests in flight and processes them concurrently
  - Reduces traffic on mesh by eliminating home agent to LLC interaction
  - Reduces latency by launching snoops earlier and obviates need for different snoop modes

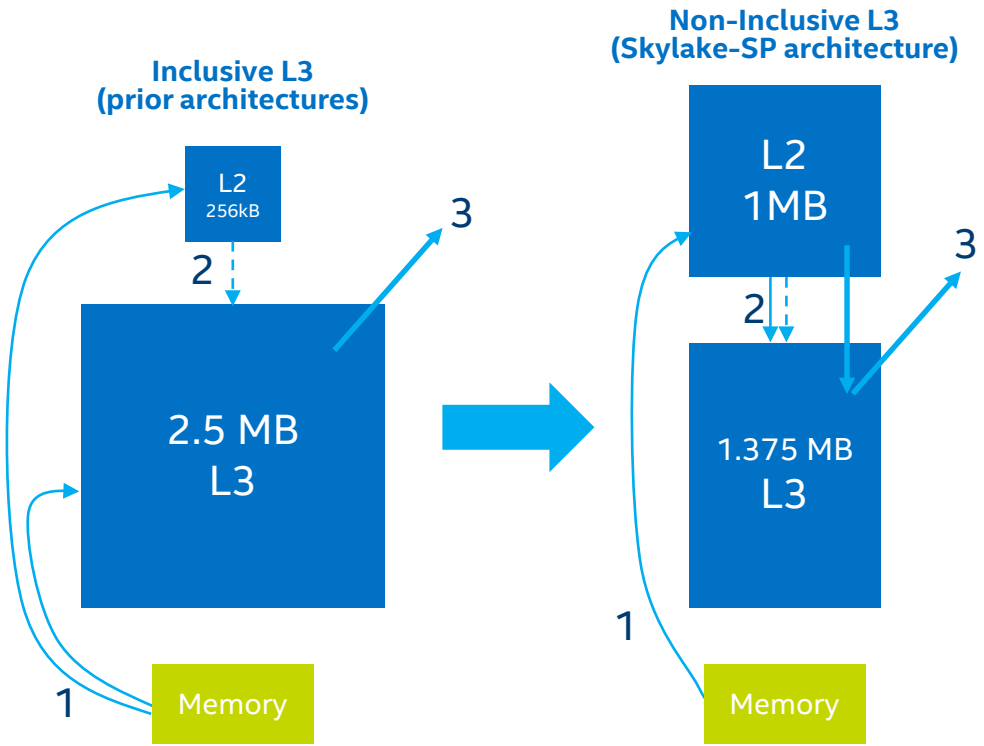
DISTRIBUTED CHA ARCHITECTURE SUSTAINS HIGHER BANDWIDTH AND LOWERS LATENCY

Re-Architected L2 & L3 Cache Hierarchy

- On-chip cache balance shifted from shared-distributed (prior architectures) to private-local (Skylake architecture):
  - Shared-distributed ➔ shared-distributed L3 is primary cache
  - Private-local ➔ private L2 becomes primary cache with shared L3 used as overflow cache
- Shared L3 changed from inclusive to non-inclusive:
  - Inclusive (prior architectures) ➔ L3 has copies of all lines in L2
  - Non-inclusive (Skylake architecture) ➔ lines in L2 ***may not*** exist in L3

SKYLAKE-SP CACHE HIERARCHY ARCHITECTED SPECIFICALLY FOR DATA CENTER USE CASE

Inclusive vs Non-Inclusive L3

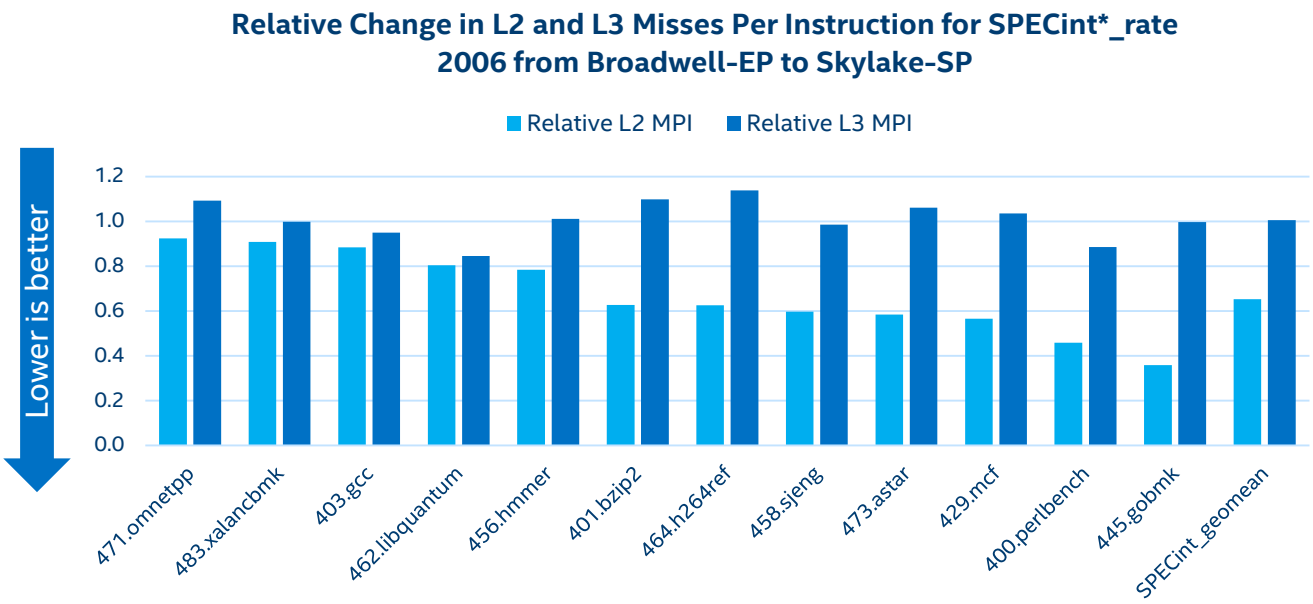


1. Memory reads fill directly to the L2, no longer to both the L2 and L3
2. When a L2 line needs to be removed, both modified and unmodified lines are written back
3. Data shared across cores are copied into the L3 for servicing future L2 misses

Cache hierarchy architected and optimized for data center use cases:

- Virtualized use cases get larger private L2 cache free from interference
- Multithreaded workloads can operate on larger data per thread (due to increased L2 size) and reduce uncore activity

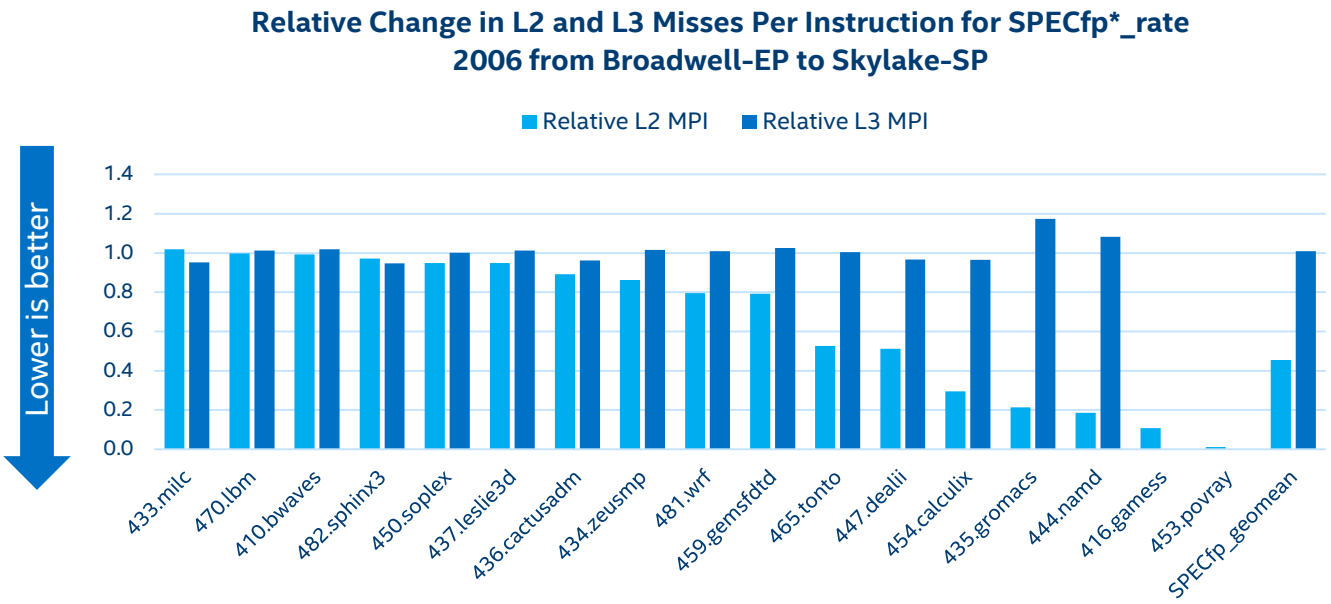
Cache Performance



Skylake-SP cache hierarchy significantly reduces L2 misses without increasing L3 misses compared to Broadwell-EP

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

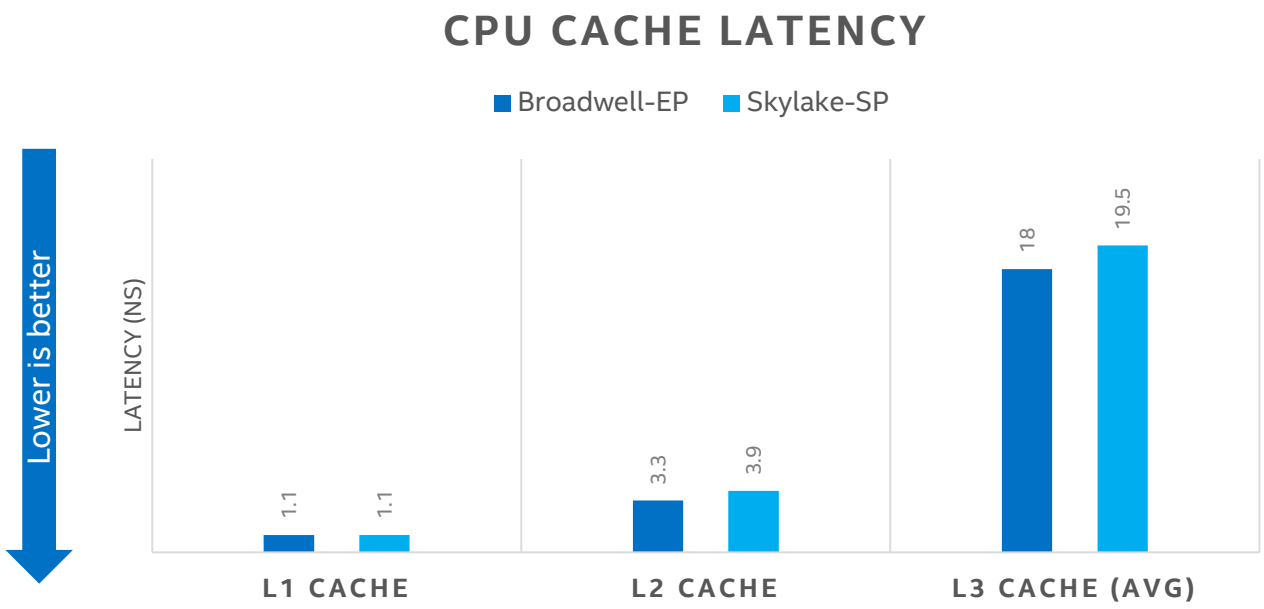
Cache Performance



Skylake-SP cache hierarchy significantly reduces L2 misses without increasing L3 misses compared to Broadwell-EP

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Cache Performance

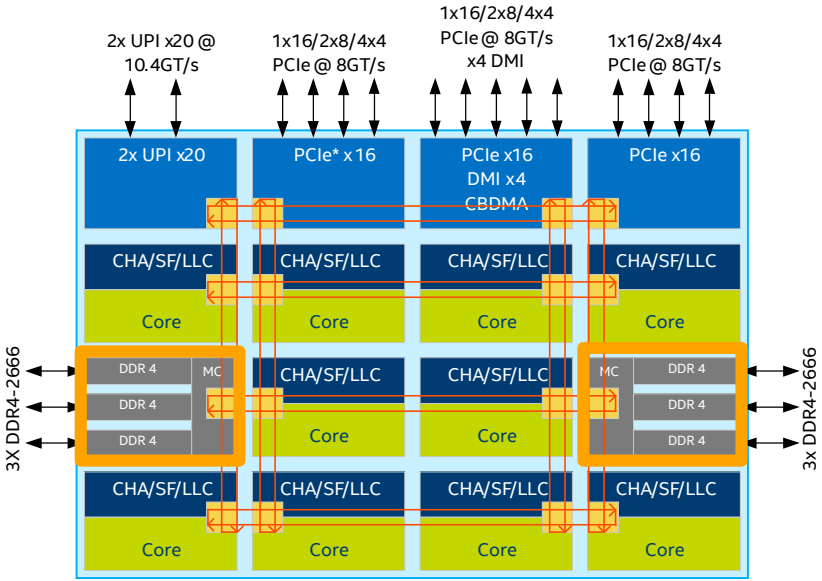


Skylake-SP L2 cache latency has increased by 2 cycles for a 4x larger L2

Skylake-SP achieves good L3 cache latency even with larger core count

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, SNC1, 6x32GB DDR4-2666 per CPU, 1 DPC, and platform with Intel® Xeon® E5-2699 v4, Turbo enabled, without COD, 4x32GB DDR4-2400, RHEL 7.0. Cache latency measurements were done using Intel® Memory Latency Checker (MLC) tool.
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Memory Subsystem



2 Memory Controllers, 3 channels each ➔ total of 6 memory channels

- DDR4 up to 2666, 2 DIMMs per channel
- Support for RDIMM, LRDIMM, and 3DS-LRDIMM
- 1.5TB Max Memory Capacity per Socket (2 DPC with 128GB DIMMs)
- >60% increase in Memory BW per Socket compared to Intel® Xeon® processor E5 v4

Supports XPT prefetch and D2C/D2K to reduce LLC miss latency

Introduces a new memory device failure detection and recovery scheme with Adaptive Double Device Data Correction (ADDDC)

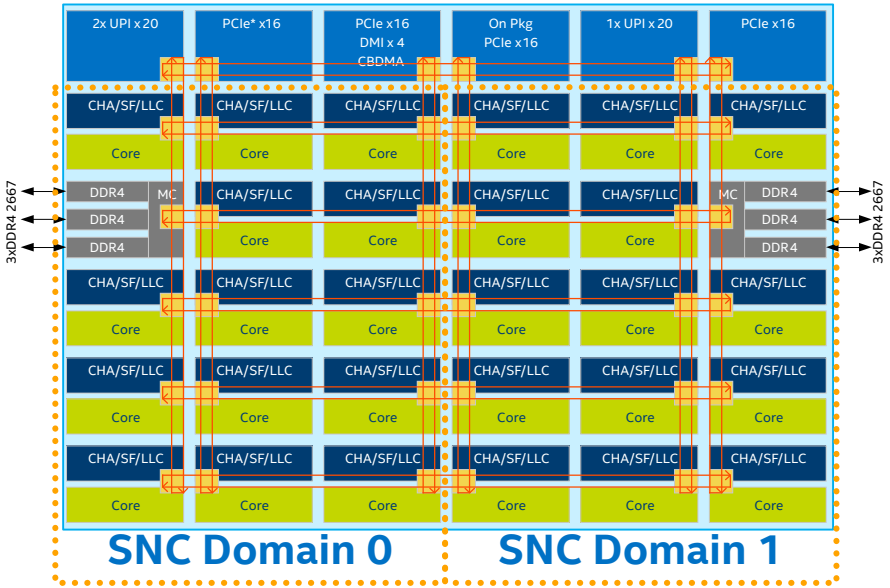
SIGNIFICANT MEMORY BANDWIDTH AND CAPACITY IMPROVEMENTS

Sub-NUMA Cluster (SNC)

Prior generation supported Cluster-On-Die (COD)

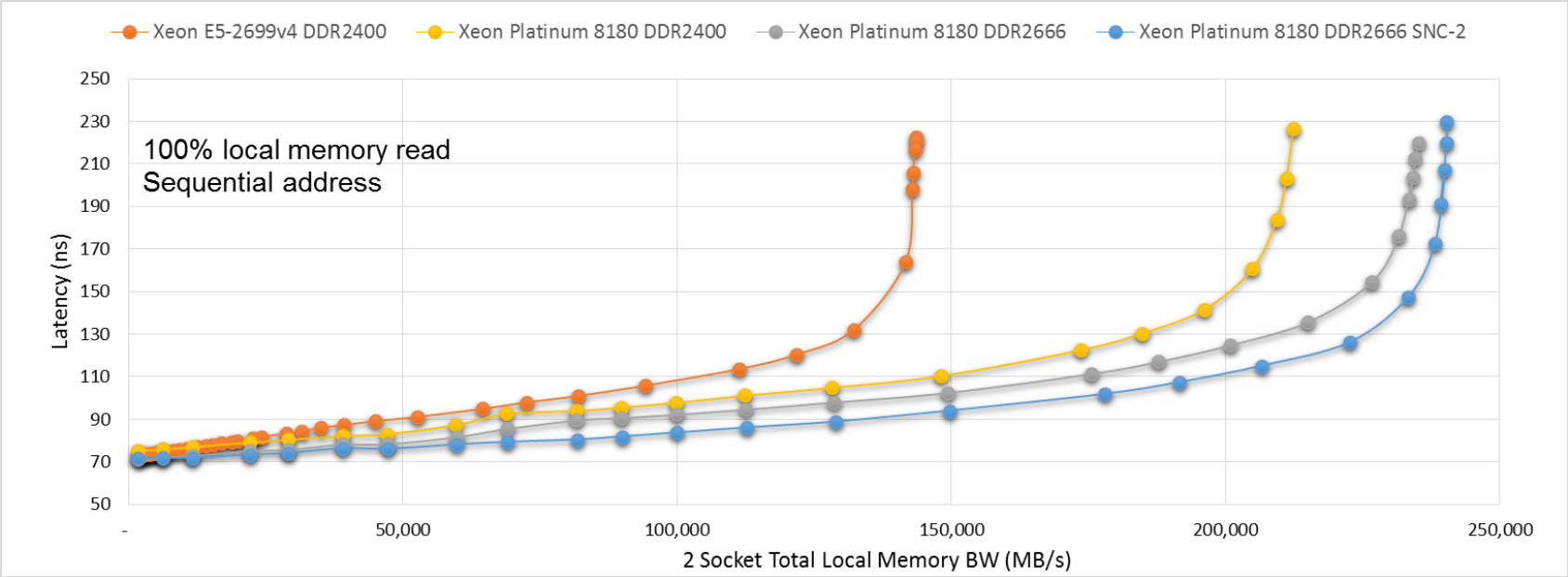
SNC provides similar localization benefits as COD, without some of its downsides

- Only one UPI caching agent required even in 2-SNC mode
- Latency for memory accesses in remote cluster is smaller, no UPI flow
- LLC capacity is utilized more efficiently in 2-cluster mode, no duplication of lines in LLC



Memory Performance

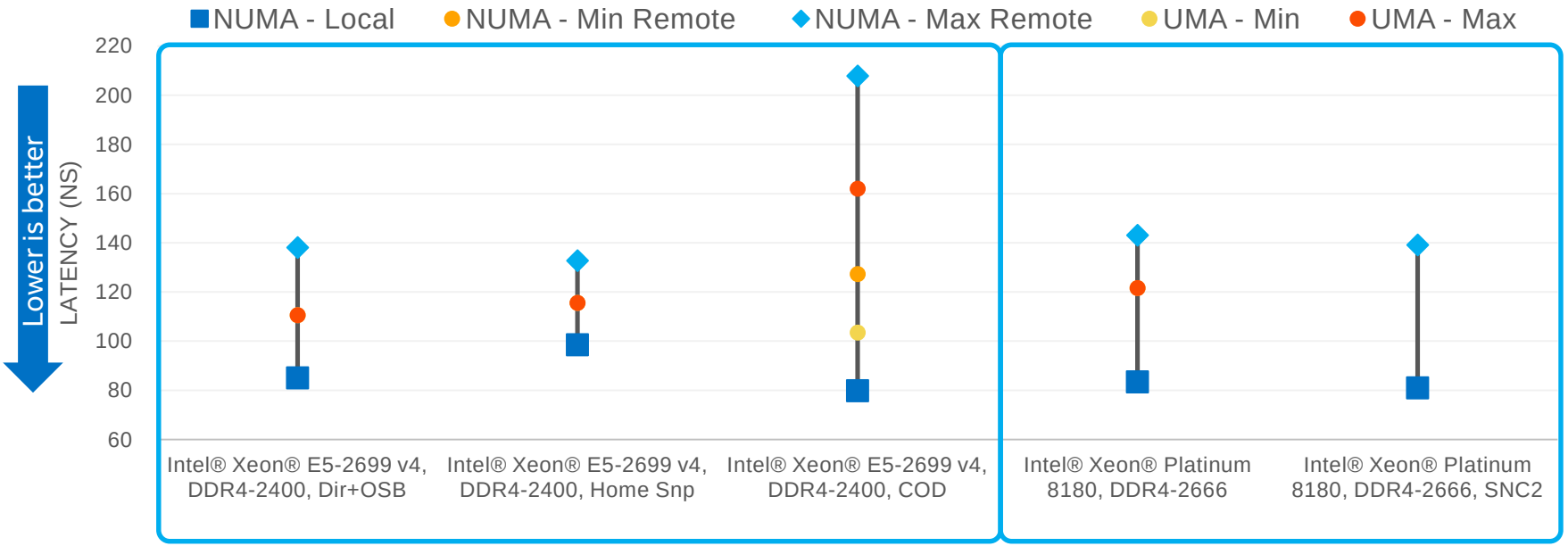
Bandwidth-Latency Profile



Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1/SNC2, 6x32GB DDR4-2400/2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance

Memory Performance

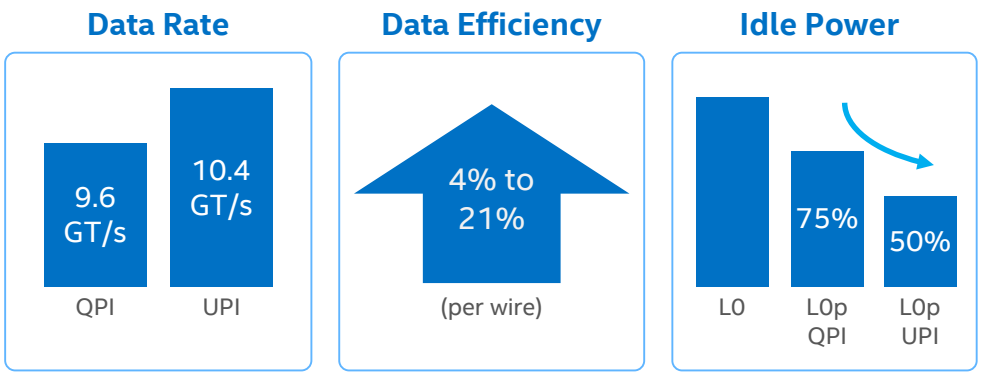
Core to Memory Latency

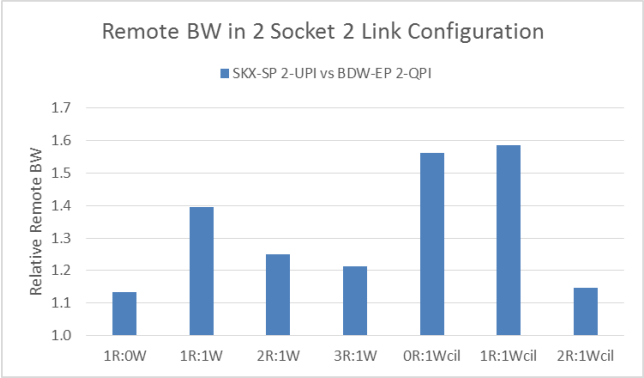


Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, 6x32GB DDR4-2666, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.
Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Intel® Ultra Path Interconnect (Intel® UPI)

- Intel® Ultra Path Interconnect (Intel® UPI), replacing Intel® QPI
- Faster link with improved bandwidth for a balanced system design
  - Improved messaging efficiency per packet
- 3 UPI option for 2 socket – additional inter-socket bandwidth for non-NUMA optimized use-cases





INTEL® UPI ENABLES SYSTEM SCALABILITY WITH HIGHER INTER-SOCKET BANDWIDTH

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, 6x32GB DDR4-2666, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.
Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Processor Integrated I/O

3 independent pipelines of x16 PCIe* Gen3

- Each x16 can be bifurcated into 2x8, 1x8+2x4, or 4x4 root ports
- New traffic controller pipeline improves over prior design
- Additional x16 PCIe for Intel® Omni-Path integration

Non-Transparent Bridging (NTB)

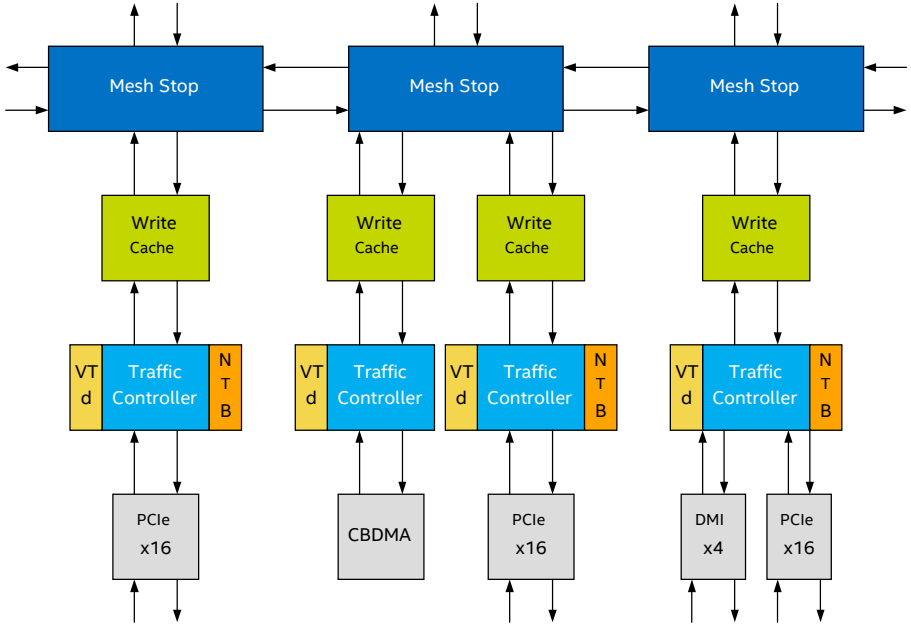
- One NTB per x16 PCIe, which can be configured as 1x8 or 1x4 NTB

Intel® QuickData Technology (CBDMA)

- 2x bandwidth on Mem-Mem copy
- Supports MMIO-Mem copy

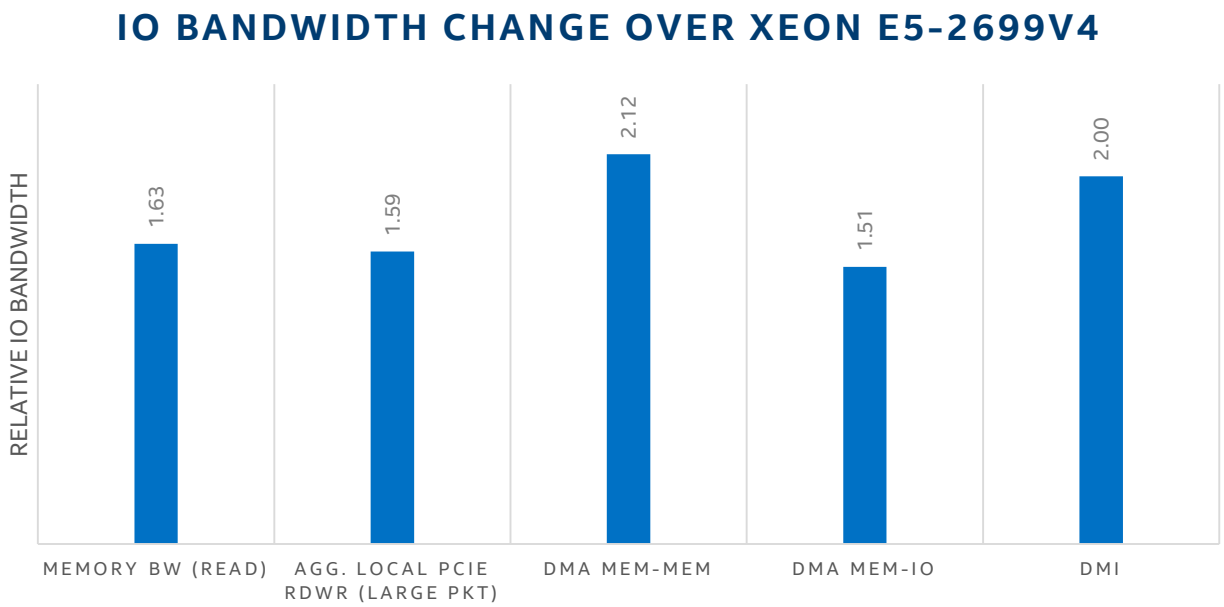
Intel® Volume Management Device (VMD)

- One VMD domain per x16 PCIe



MODULAR IO DESIGN WITH IMPROVED FEATURE SET FOR CONVERGED DATA CENTER

IO Performance



>50% aggregate IO bandwidth improvement in line with memory bandwidth increase for a balanced system performance

Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2400/2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

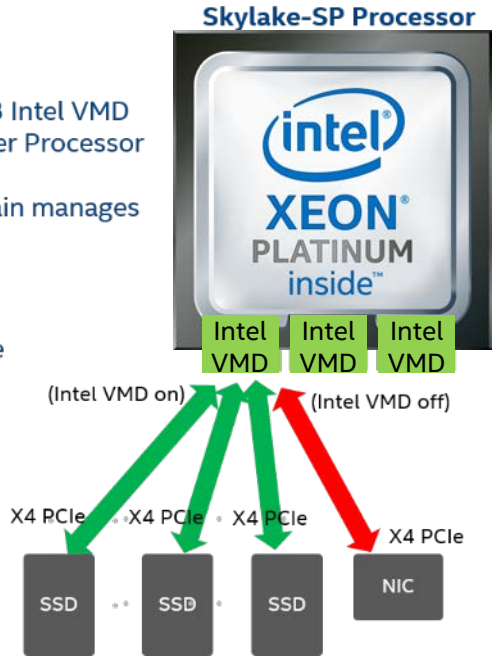
SKYLAKE-SP NEW CAPABILITIES

Intel® Volume Management Device (Intel® VMD)

There are 3 Intel VMD domains per Processor

Each Domain manages x16 lanes

Intel VMD can be turned on/off in BIOS at x4 lane granularity



Intel® VMD is a CPU-integrated device to aggregate NVMe SSDs into a storage volume and enables other storage services such as RAID

- Intel® VMD is an "integrated end point" that stops OS enumeration of devices under it
- Intel® VMD maps entire PCIe* trees into its own address space (a domain)
- Intel® VMD driver sets up and manages the domain (enumerate, event/error handling), but out of fast IO path

ELIMINATES ADDITIONAL COMPONENTS TO PROVIDE A FULL-FEATURE STORAGE SOLUTION

Energy Efficiency and Power Management Enhancements

ENERGY EFFICIENCY

Optimized intermediate core turbo profile

Dynamic power sharing between core, uncore, and fabric HFI

Larger L2 cache for reduced interconnect and coherency activity

Power delivery through integrated VR for core and uncore

POWER MANAGEMENT

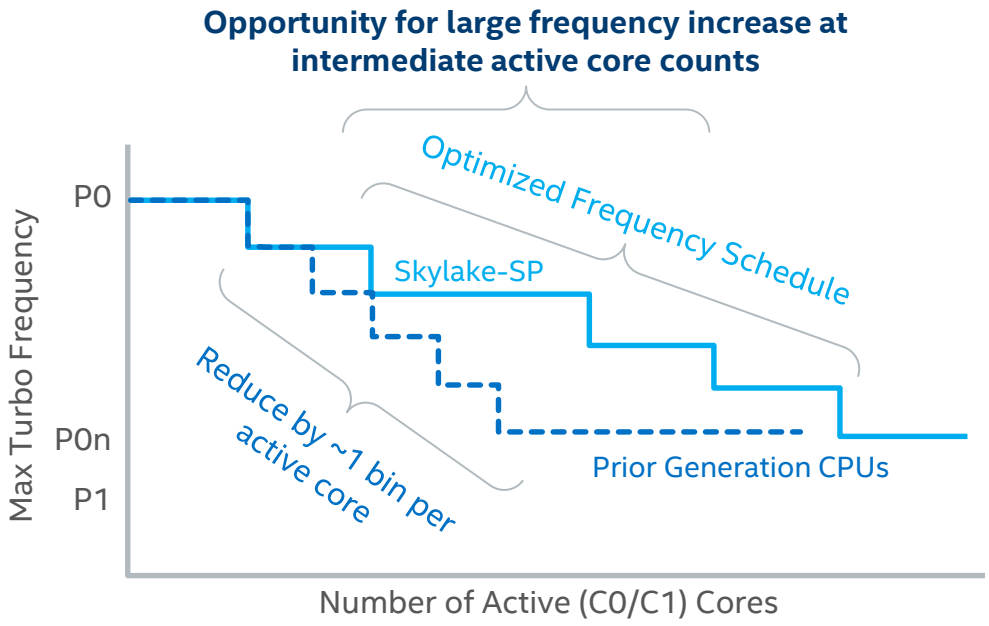
Intel® Speed Shift Technology for autonomous P-state control

Improved core and uncore frequency scaling heuristics

On-die Pmax detector for rapid response to power excursions

Independent per core and CLM (CHA, LLC, and mesh) voltage and frequency domains

Optimized Turbo Profiles



Prior generation data center CPUs typically decreased turbo by 1 bin for each additional active core

Skylake-SP provides higher intermediate turbo points by stepping down in a more optimal manner

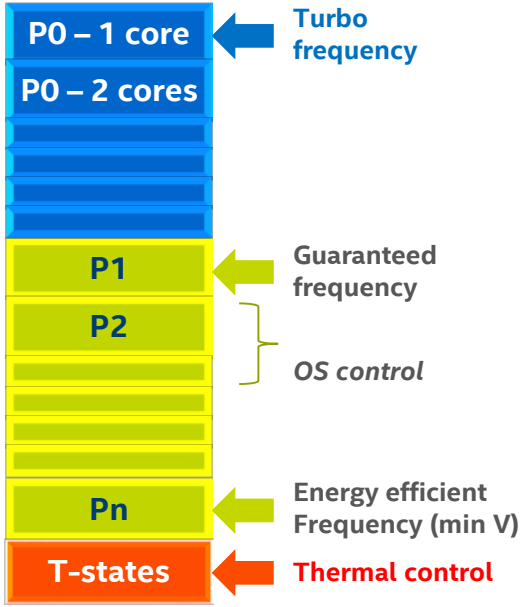
- Higher performance dynamically with C-states
- BIOS/OS core disable can be used to mimic higher frequency SKUs (with some tradeoffs)

Note: there is no guarantee that these frequencies can be achieved for a given workload on all units

*Picture is an illustration only. Not intended to represent any specific SKU or imply any frequency commitments.

Intel® Speed Shift Technology Interface

Legacy Interface



DVFS – Intel SpeedStep® Technology

$P \sim V^2 \cdot f \cdot C_{dynn} + leakage(V) \sim f^3$

Legacy: OS controls P-state

P1-Pn enumerated via ACPI tables

Explicit P-state selection to P1

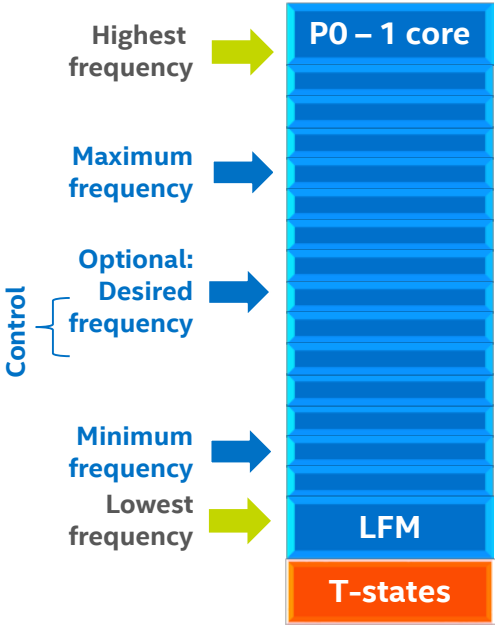
Autonomous control in turbo range

New: guided autonomous control

OS provides min, max and preference

Demand based HW control

HWP i/F



Intel® Xeon® Scalable Processor:

Intel® Run Sure Technology Features

INTEL® RUN SURE TECHNOLOGY

Resilient System Technologies

- Advanced Error Detection and Correction (AEDC)
- MCA 2.0 Recovery (as per eMCA gen2 architecture)
- MCA Recovery-Execution Path
- MCA Recovery-Non Execution Path
- Local Machine Check (LMCE) based recovery

Resilient Memory Technologies

- SDDC + 1, Adaptive DDDC (MR) +1
- Addressed Range/Partial Memory Mirroring

Resilient System Technologies - integrate processor, firmware, and software layers that allow the system to diagnose and/or recover from previously fatal errors

Resilient Memory Technologies - ensure data integrity & enable systems to keep running reliably over a longer period of time, reducing the frequency of service calls

CONTINUED IMPROVEMENT IN DATA CENTER UPTIME WITH INTEL® RUN SURE TECHNOLOGY

Intel® Xeon® Scalable Processor:

New Virtualization Enhancements

GREATER CONSOLIDATION ON A COMPUTE NODE

Improved core performance and larger number of cores

Larger TLBs and per core L2 cache for improved performance and lower variability on virtualized workloads

Improved memory bandwidth and capacity to collocate demanding workloads

SECURE CREATION AND MOVEMENT OF VM ACROSS SYSTEMS

Mode Based Execution Control for hardening VM launch vulnerability

Improved timestamp virtualization to reduce overhead of migrating VMs across different CPU skus

Intel® Xeon® Scalable Processor:

New Security Enhancements

SECURITY PERFORMANCE

Intel® AVX-512

Intel® QuickAssist Technology

HARDENING AGAINST ATTACK SURFACES

Mode Based Execution to protect from attacks during VM creation

Page Protection Keys to create robust applications with differentiated page access rights managed at user level

Intel® Memory Protection Extension (Intel® MPX) to prevent buffer overflow attacks

HARDENING THE PLATFORM

Secure key management within the chipset without a discrete TPM

Authenticated and measured launch and recovery options

Skylake-SP with Integrated Fabric

Single on-package Omni-Path Host Fabric Interface (HFI)

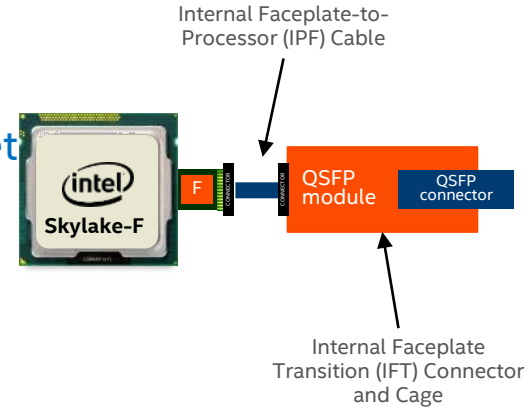
Fabric component interfaces to CPU using x16 PCIe* lanes

Fabric PCIe lanes are additional to the 48 PCIe lanes on the socket

Single cable from SKL-F package connector to QSFP module

Same socket for Skylake-SP and Skylake-F processors

- Purley platform can be designed to support both processors
- Platform design requires an expanded keep-out zone and additional board components to accommodate both processors



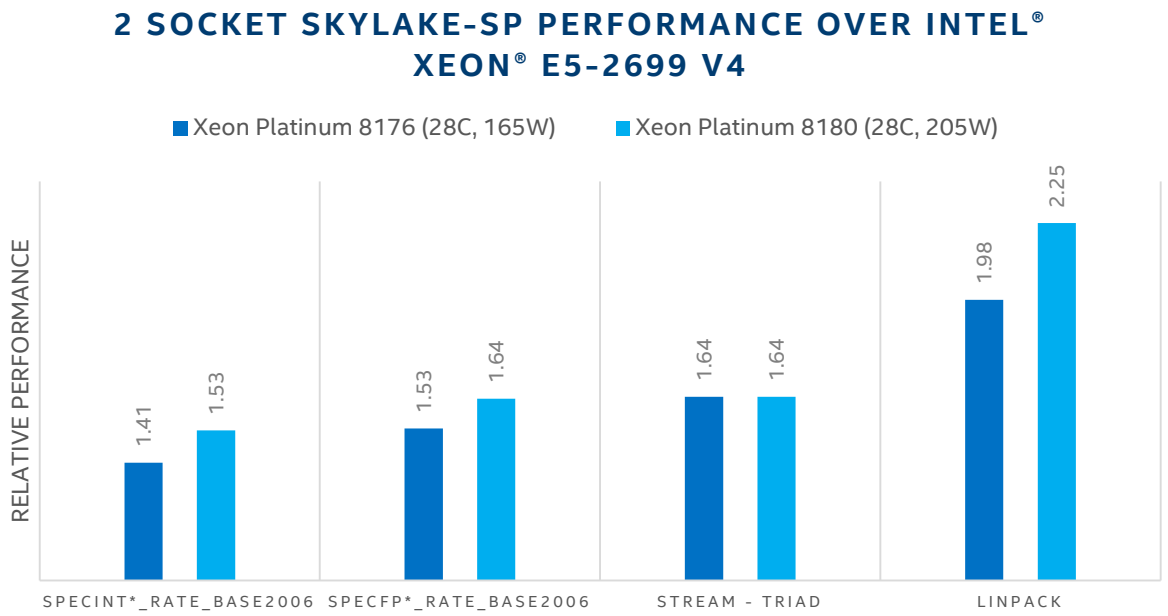
SKYLAKE-SP CPU WRAP UP

Skylake-SP Architecture Summary

New Architectural Innovations for Data Center

- **Up to 60% increase** in compute density with Intel® AVX-512
- **Improved performance and scalability** with Mesh on-chip interconnect
- L2 and L3 cache hierarchy **optimized for data center workloads**
- Improved memory subsystem with **up to 60% higher memory bandwidth**
- Faster and more efficient Intel® UPI interconnect for **improved scalability**
- Improved integrated IO with **up to 50% higher aggregate IO bandwidth**
- **Increased protection** against kernel tampering and user data corruption
- Core, cache, memory and IO improvements for **increased virtual machine performance**
- **Enhanced power management and RAS capability** for improved utilization of resources

Skylake-SP Performance



Skylake-SP CPUs provide significant performance upside compared to prior generation

165W Skylake-SP CPUs provide more than 40% gain on performance

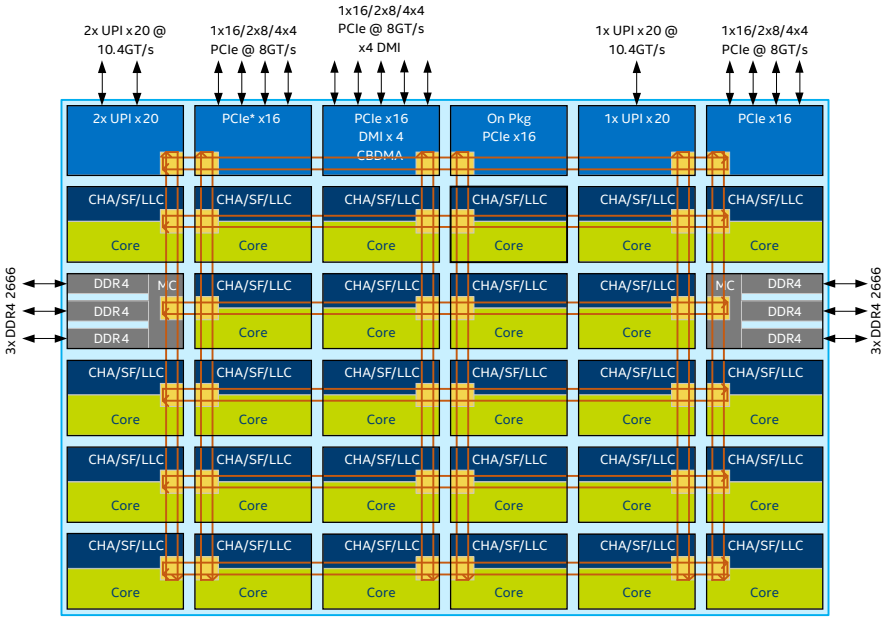
205W Skylake-SP CPUs provide additional boost to core bound workloads

Source as of June 2017: Intel internal measurements with Xeon Platinum 8180 and 8176, Turbo enabled, UPI=10.4, SNC1, 6x32GB DDR4-2666 per CPU, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

Skylake-SP Die Configurations

XCC Die with 28 Cores

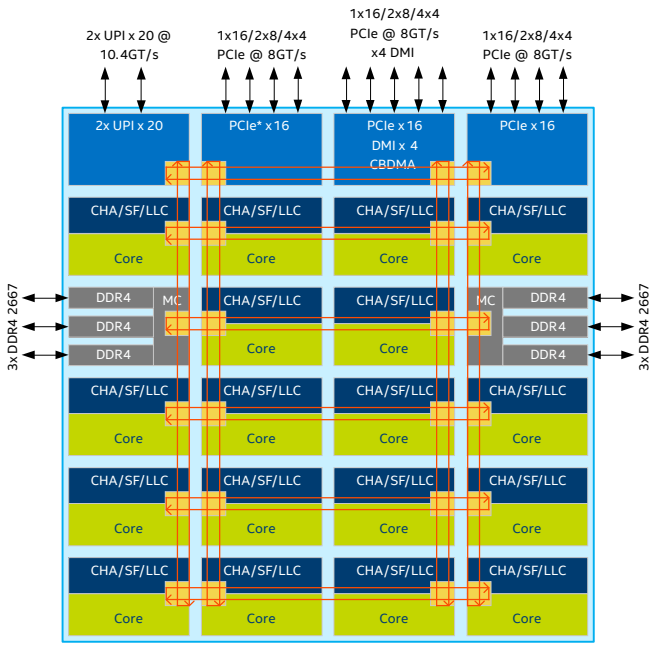
- 6x6 Mesh topology
- 5 rows of core and LLC
- 2 memory controllers, one on each side of die
- All IOs at the top
- 3 x16 PCIe Gen3 stacks
- 1 x16 PCIe for MCP use
- Up to 3 Intel® UPI ports



CHA – Caching and Home Agent ; SF – Snoop Filter; LLC – Last Level Cache;
Core – Skylake-SP Core; UPI – Intel® UltraPath Interconnect

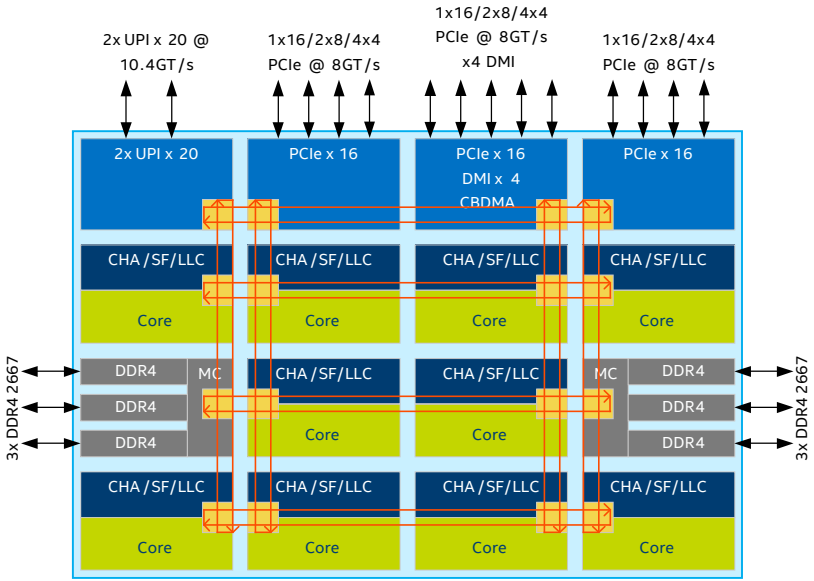
High and Low Core Count Die Configurations

HCC (up to 18 cores)



CHA – Caching and Home Agent ; SF – Snoop Filter ; LLC – Last Level Cache ;
Core – Skylake-SP Core; UPI – Intel® UltraPath Interconnect

LCC (up to 10 Cores)



CHA – Caching and Home Agent ; SF – Snoop Filter ; LLC – Last Level Cache ;
Core – Skylake -SP Core ; UPI – Intel® UltraPath Interconnect

LEWISBURG CHIPSET ARCHITECTURE

Lewisburg: New PCH for the Converged Platform

Ubiquity of network security, efficient data storage and packet manipulation in Cloud, Storage, Enterprise and Network appliances.

Lewisburg is a Common Platform Offering in Purley generation to meet the converged requirement

Data Center PCH provides boot, standard legacy and high-speed IO, manageability and clocking solutions

- Intel® Innovation Engine is a manageability sandbox for system builder

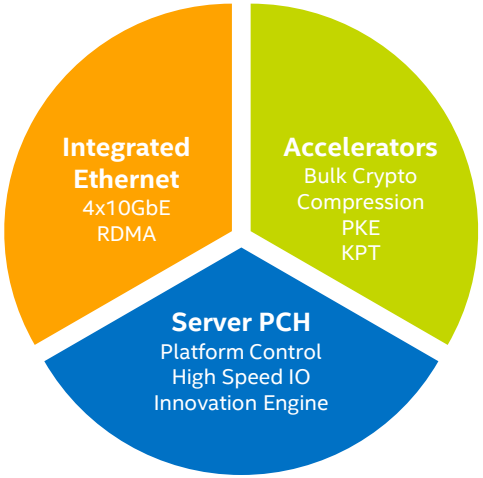
Integrated Intel® Ethernet Controller reduces area, power, cost and provides platform LOM capabilities

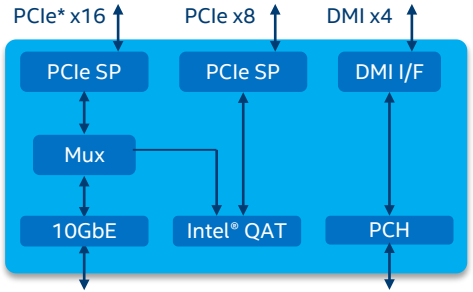
- Intel® Ethernet Connection X722 with up to 4x10Gbps
- Supports Network Virtualization Offloads and iWARP RDMA

Intel® QuickAssist Technology Accelerators provide security and compression for Communications, Storage, and Cloud deplo

- Crypto / compression up to 100Gbps

PKE	Public Key Encryption
KPT	Key Protection Technology
RDMA	Remote Direct Memory Access





PCH Generational Comparison

Features	Intel® C610 series chipset (Wellsburg)	Intel® C620 series chipset (Lewisburg)
Standard Features	Intel® vPro™ technology, AMT, Node Manager 3.0; Server Trace Lengths; 6 SMBus Discrete, Integrated and Hybrid Clocking	Intel® vPro™, AMT, Node Manager 4.0; Server Trace Lengths; 6 SMBus Discrete, Integrated and Hybrid Clocking
SATA Ports	Up to 10 SATA 3 (6 Gb/s)	Up to 14 SATA 3 (6 Gb/s)
USB Ports	Up to 14 USB 2.0; Up to 6 USB 3.0	Up to 14 USB 2.0; Up to 10 USB 3.0
DMI	x4, 2.0 speed	x4, 3.0 speed
Additional CPU Uplink Options	N/A	PCIe* 3.0 at x8 and/or x16;
PCI Express*	Up to 8 PCIe 2.0 (5 GT/s)	Up to 20 ports PCIe 3.0 (8 GT/s)
TPM Support	TPM 1.2	TPM 2.0
SATA Express, NVM Express	No	No / Supported
Intel® RSTe	Yes	Yes
Innovation Engine	No	Yes
LAN	Integrated 1GbE	Integrated Intel® Ethernet Connection X722 with up to 4x10Gb/1Gb ports with iWARP RDMA/ SFI
Intel® QuickAssist Technology - Crypto	N/A	Up to 100 Gb/s IPSec/SSL
Intel QuickAssist Technology - Public Key	N/A	Up to RSA2K 100K Ops
Intel QuickAssist Technology – Compression	N/A	Up to 100Gb/s (Deflate(LZ77))
Enhanced Serial Peripheral Interface (eSPI)	N/A	up to 60MHz, 1.8V

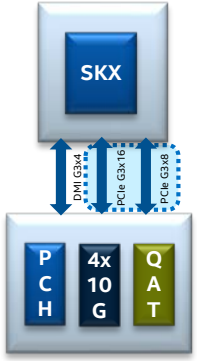
INTEGRATED 10GB INTEL® ETHERNET, ENHANCED I/O AND PLATFORM SECURITY WITH ROBUST CRYPTO AND COMPRESSION SOLUTIONS

Green text is new

Lewisburg Deployment Configurations

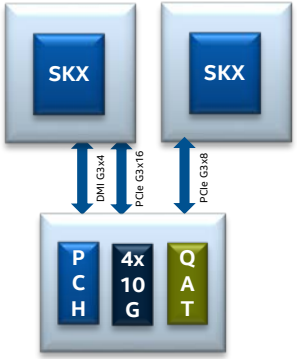
Conventional

On-board PCH. Dual PCIe uplink to single CPU, provisioned for Intel® QAT and 4x10GbE (PCIe Uplink optional)



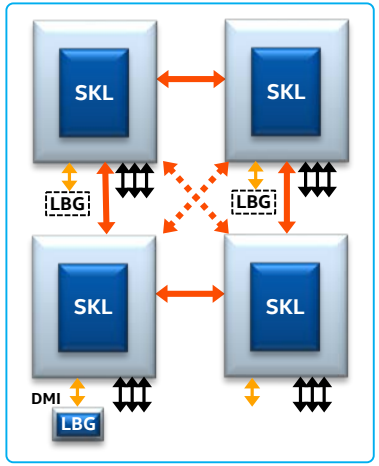
CPU Straddling

On-board PCH PCIe uplink to multiple CPU, provisioned for QAT and 4x10GbE



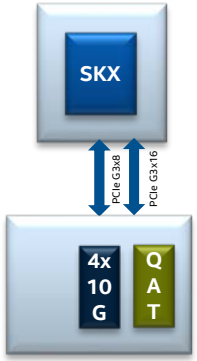
Multi-PCH

Multiple On-board PCH for failover, flexible partition or IO expansion
Optional PCIe uplink for QAT and 4x10GbE



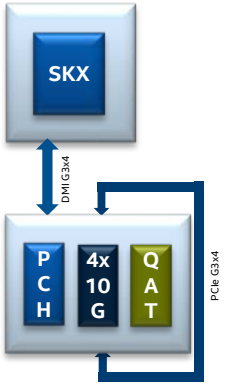
PCIe* Endpoint Mode

On-board or PCIe-card-based acceleration for scalability, provisioned for QAT and 4x10GbE



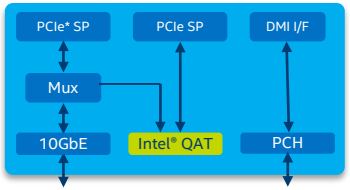
Wrap-Around

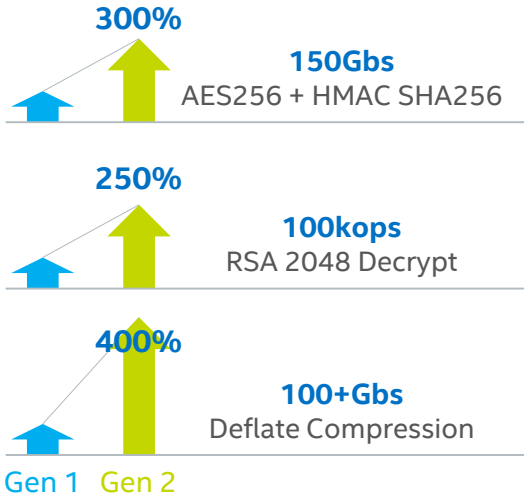
On-board PCH Configure low-BW QAT/Ethernet as PCH-attached downstream device



Intel® QuickAssist Technology

Generation 2





Bulk Crypto → Security for data in flight & rest

Public Key Encryption → Secure Key Establishment

Compression → Lossless compression for data in flight & rest

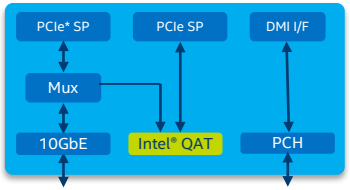
New Platform Capability

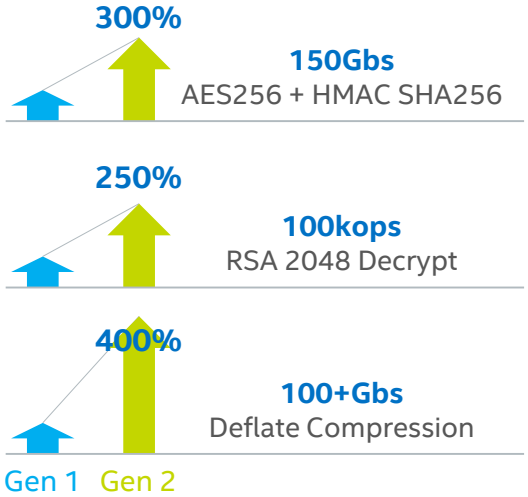
Intel® Key Protection Technology

INTEL® QUICKASSIST TECHNOLOGY IS DESIGNED TO OPTIMIZE THE USE & DEPLOYMENT OF CRYPTO AND COMPRESSION HARDWARE ACCELERATORS ON INTEL® PLATFORMS

Intel® QuickAssist Technology

Generation 2





Bulk Crypto → Security for data in flight & rest

Public Key Encryption → Secure Key Establishment

Compression → Lossless compression for data in flight & rest

New Platform Capability

Intel® Key Protection Technology

INTEL® QUICKASSIST TECHNOLOGY IS DESIGNED TO OPTIMIZE THE USE & DEPLOYMENT OF CRYPTO AND COMPRESSION HARDWARE ACCELERATORS ON INTEL® PLATFORMS

"Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance/datacenter. Configurations: see slide 71

Intel® QuickAssist Technology: Crypto

Usage Model

- Network security (IPsec, SSL/TLS), hashing for data-deduplication, encrypted storage

Symmetric (Bulk) Cryptography

- Ciphers (AES, 3DES/DES, RC4, KASUMI*, Snow 3G)
- Message digest/hash (MD5, SHA1, SHA2x, SHA3) and authentication (HMAC, AES-XCBC)
- Algorithm chaining (one cipher and one hash in a single operation)
- Authenticated encryption (AES-GCM, AES-CCM)
- AES-XTS

Wireless

- KASUMI, Snow 3G and ZUC

Asymmetric (Public Key) Cryptography

- Modular exponentiation for Diffie-Hellman (DH)
- RSA key generation, encryption/decryption and digital signature generation/verification
- DSA parameter generation and digital signature generation/verification
- Elliptic Curve Cryptography: ECDSA, ECDHE

Performance	Coleto Creek	Lewisburg
Network Security Protocols		
TLS @ 16k records	50 Gbs	150 Gbs[1]
IPSec @ 1kB	45 Gbs	100 Gbs[1]
Public Key Encryption		
RSA Decrypt 2K	40k Ops	100k Ops
Wireless Ciphers		
ZUC/Snow 3G/KASUMI* F8[2]	20 Gbs	50Gbs
Cipher or Hash Only		
AES128 CBC @ 4k	50Gbs	150Gbs
SHA1, SHA256, SHA3, MD5 @ 4k	50Gbs	140Gbs

1. Using 16k records using AES-CBC-HMAC SHA1/256, 100Gbs at 1k packet
2. KASUMI-F8 (encryption) at 320B packets, 7Gbs for 64B packets

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Intel measurements as of May 2017; Intel® Customer Reference Board Neon City with one Intel® Xeon® Skylake-SP processor and one Lewisburg Chipset LBG-T B0 ES2* SKU (DCL: #568586). For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

Intel® QuickAssist Technology: Data Compression

Usage Model

- Big data acceleration
- WAN acceleration
- Http compression
- File System
- Databases

Compression and Decompression Algorithm

- DEFLATE: LZ77 compression followed by Huffman coding, with a gzip or zlib header

Other Features

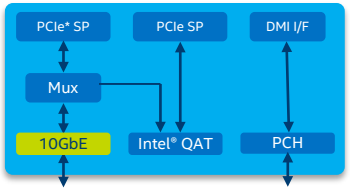
- Engine can be configured to perform either compression or decompression
- Support for stateful (de)compression
- Storage-specific features

Performance	Coleto Creek	Lewisburg
Compression	24 Gbs[1]	100+ Gbs[1]
Decompression	24 Gbs[1]	160 Gbs[1]
Compression + Decompression	24 Gbs[1]	100 Gbs[1]

1. Dynamic Deflate Level 1 using 64KB buffer size
2. Best case compression ratio with Lewisburg is zlib level 4 (Compression L4 performance 24 Gbs)
3. Measured using Calgary Corpus

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
Intel measurements as of May 2017; Intel® Customer Reference Board Neon City with one Intel® Xeon® Skylake-SP processor and one Lewisburg Chipset LBG-T B0 ES2* SKU (DCL: #568586). For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

Integrated Intel® Ethernet Connection X722



Low-cost solution for up to four ports of 10Gb Ethernet

- >50% of server ports will be 10GbE in 2017[1]
- Lower power, less board space: integrated 10GbE saves > 20 cm^2 in board area and > 25% in power consumption [2]
- Available as integrated Ethernet on motherboard or stand-alone NIC cards

Dual 10G BaseT Card

Proven "It Just Works" 10GbE solution

- Based on Intel® Ethernet Converged Network Adapter XL710 (Fortville) IP
- External PHYs in production today support 1GbE and 10GbE 10GBASE-T

Quad-Port SFP Card

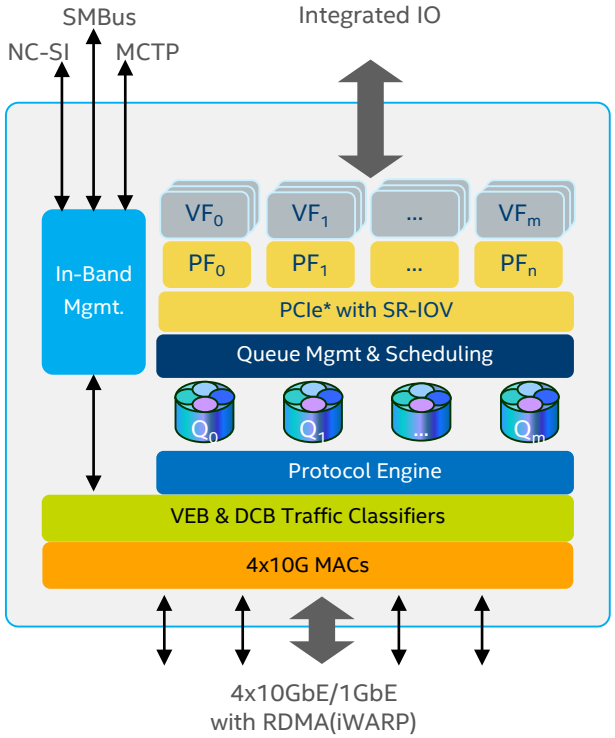
Advanced features to enable Software Defined Infrastructure

- Network Virtualization Offloads to support the move to L3 networks
- iWARP RDMA for increased bandwidth at lower CPU utilization
- Intel® Ethernet Flow Director traffic steering for increased efficiency
- Intel® Data Plane Development Kit (DPDK) for advanced packet forwarding

1. Dell'Oro, "Worldwide Server Market Total Network Metrics" January 2017 2. Intel estimates in comparison to Grantley platforms

Intel® 4x10Gb/1Gb Ethernet Controller

Feature Support	Customer Value
Quad Port 10GbE MAC/PHY • Based on Intel's 10GbE solution • Interfaces: 10G (KR, SFI, XFI), 1G (KX)	• Optimized for networking capability in Cloud, Comms, and Storage • Single network driver on Intel® platform
Remote Direct Memory Access • iWARP	• Routable and scalable RDMA ideal for large segmented networks in private and public clouds
Network Virtualization Offloads • Flexible Filters (ATR, Flow Director) • NVGRE, IPinGRE, VXLAN, MACinUDP	• Abstract the network for cloud flexibility • Enhanced programmability and application affinity
Standards Based Virtualization • SR-IOV: 4 Physical/128 Virtual Function • VEB (Virtual Ethernet Bridge)	• Broad OS enablement
Power Management	• Energy Efficient Ethernet
Manageability • BMC Pass-Through (control & network) • Interfaces: NC-SI, SMBus, and MCTP	• Common transport for LAN-to-BMC



Remote Direct Memory Access (RDMA)

RDMA is a network performance optimization

- Enables direct app-to-app communication across nodes
- Bypasses the OS stack & kernel
- Provides direct channel for remote memory application access

RDMA offers lower latency, high throughput by:

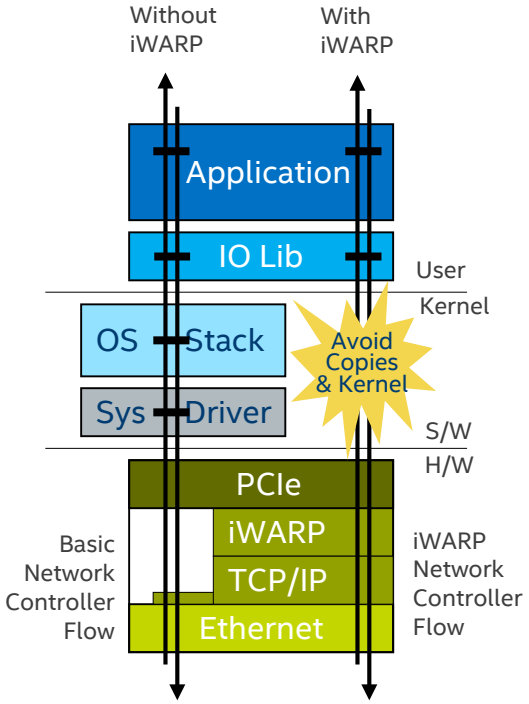
- Avoiding application context switching
- Placing data directly in application buffers (zero-copy DMA)
- Moving protocol processing off the CPU

Intel® Ethernet Connection X722 is featured with iWARP (Internet Wide-area RDMA Protocol) RDMA:

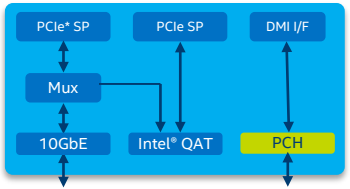
- Recommended for easy deployment and configuration, scalability and congestion control

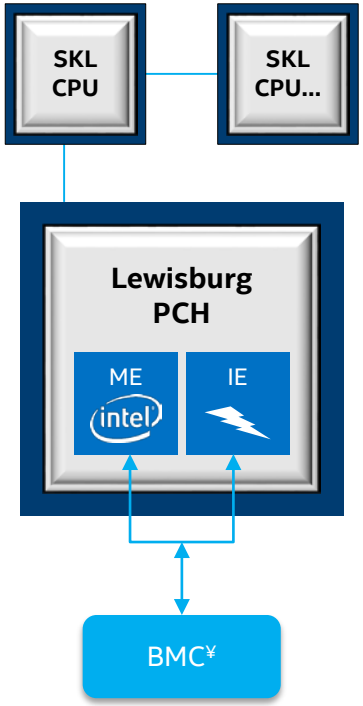
RDMA-aware applications or messaging layers are required

- Windows*/Linux* Drivers supported for Storage, Messaging Middleware and HPC application categories



Introducing the Innovation Engine (IE)





Embedded core in the LBG PCH

- Very small Intel® Architecture core
- Similar to Management Engine (ME) hardware, with some privilege and IO differences

Reserved for system-builder's code, not for Intel firmware

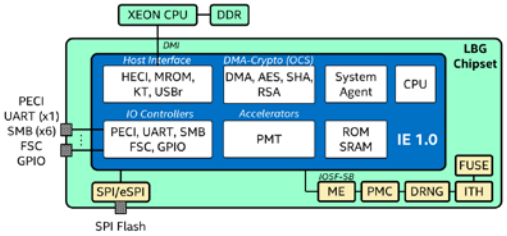
- Intel supplies IE hardware only
- IE code is cryptographically bound to the system-builder
- Code not authenticated by the system-builder will not load

Activation is not required for normal system operation

- Optional feature
- Requires SPS firmware. ME1x firmware does not support

¥BMC is optional if system-builder has implemented a management app in IE that communicates directly with the network

Lewisburg ME vs IE Features

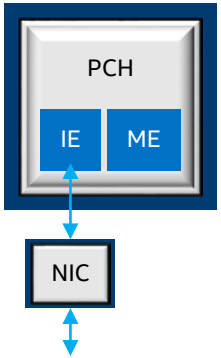


	LBG ME	LBG IE
Processor	Intel® Quark™ x86 (412 DMIPS)	Intel Quark x86 (412 DMIPS)
Memory	1.7MB SRAM	1.4MB SRAM
Crypto Algorithms	Yes	Yes
Host Interface	Yes	Yes
Platform Interface	Yes	Yes + UART
FW	Intel® AMT/vPro, SPS	System Builder FW
Security	TXT, BtG, PTT, KPT	-
Operating States	S0/M0, Sx/M3, Sx-Moff	S0/I0, Sx/I3, S0/Ioff, Sx/Ioff

Red: delta between ME and IE

Broad IE Usage Models

Lite, BMC-less Manageability



Segment

Scale-out Cloud & Embedded/Appliance

Value

Basic manageability without cost, space, power of a BMC

Common manageability solution across vendors

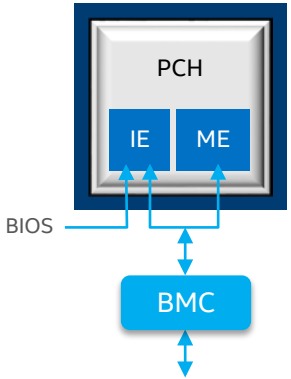
Usages

Simplified platform management (IPMI, Redfish, SoftKVM) without a BMC

Hardware Security Module

Basic node management

BMC/BIOS/ME Assist



Segment

Enterprise & Scale-up Cloud

Value

Improved system manageability through tighter platform integration

Better performance through reduced BMC/CPU interrupts

Usages

In-band PECI over DMI link: higher bandwidth, lower latency internal data paths

Offload tight–loop sensor monitoring

Power sequencing & control

Error & reset handling

OOB telemetry for performance and predictive RAS

Platform NVRAM firmware security (validate BMC/etc)

Crash dump & recovery over PECI

Platform OOB key mgmt

Lewisburg PCH SKU Guidance

Product Name	SKU	10Gb/1Gb Ethernet Ports¥	Compression	Encryption	RSA	Max PCIe* Uplink	Recommended Min Uplink Config	PCIe* Uplink x8 Optional Muxed Link	Est TDP (W)
			Intel® QuickAssist Technology						
Intel® C621 Chipset	LBG-1G	0/4	N/A	N/A	N/A	x1	x1	n/a	~ 15
Intel® C622 Chipset	LBG-2	2/4†	N/A	N/A	N/A	x8	x4	n/a	~ 17
Intel® C624 Chipset	LBG-4	4/4	N/A	N/A	N/A	x16	x8	n/a	~ 19
Intel® C625 Chipset	LBG-E	4/4	20 Gb/s	20 Gbs	20K Ops	x16	x16	n/a	~ 21
Intel® C626 Chipset	LBG-M	4/4	40 Gb/s	40 Gbs	40K Ops	x16	x16	enabled	~ 23
Intel® C627 Chipset	LBG-T	4/4	100 Gb/s	100 Gbs	100K Ops	x16	x16	enabled	~ 26
Intel® C628 Chipset	LBG-L	4/4	100 Gb/s	100 Gbs	100K Ops	x16	x16	enabled	~ 21

†Four ports total: ports 0 & 1 can run up to 10 GbE, while ports 2 & 3 are limited to 1 GbE
Package Size (all SKUs): 34x28mm, Package Pin Count: 1310
Intel recommends two lanes of PCIe3 for each active 10GbE port for networking and x16 if Intel® QuickAssist Technology is active
LBG supports x16, x8, x4 and x1 options, up to the maximum uplink width.
¥These Ethernet ports are in addition to the 1Gb port used by ME11.6
All LBG skus support 7 year production and 10 year use in line with the Unique Extended Supply Life CPU SKUs (10-year use + NEBS-Friendly Thermal Specification)

All SKUs, frequencies and features are PRELIMINARY and can change without notice.

Summary of LBG HSIO Features

DMI
- Dedicated Gen3 x4, 4GB/s raw bidirectional throughput

PCIe* Uplink
- Dedicated Gen3 1 x16 controller
- Configurable Gen3 1 x8 controller

PCIe* Downlink
- Configurable Gen3, 5 x4 controllers, total of 20 lanes

SATA
- Configurable Gen3, 14 ports over 2 controllers
- Soft strap or GPIO based selection of PCIe or SATA

USB
- Configurable 10 USB3.0 lanes, 14 USB2.0 lanes
- Closed-Chassis Debug (DCI)

GbE:
- 1 MAC port, configurable over 5 lanes

Flex IO Lane Muxing

Lane #	PCIe Down			PCIe Up	SATA	USB	GbE
0							
1						USB 3.0	
2							
3							
4							
5							
6		PCIe	x1				
7	PCIe	x2	x1				
8	x4	PCIe	x1				
9		x2	x1				GbE
10		PCIe	x1				GbE
11	PCIe	x2	x1				GbE
12	x4	PCIe	x1				
13		x2	x1				
14		PCIe	x1		sSATA		GbE
15	PCIe	x2	x1		6 x1		
16	x4	PCIe	x1				
17		x2	x1				GbE
18		PCIe	x1				
19	PCIe	x2	x1				
20	x4	PCIe	x1	PCIe			
21		x2	x1	x8/x4/x1			
22		PCIe	x1		SATA		
23	PCIe	x2	x1		8 x1		
24	x4	PCIe	x1				
25		x2	x1				

Other Features

eSPI (Enhanced Serial Peripheral Interface):

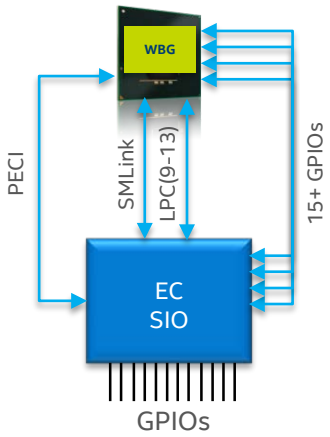
- High-speed, low pin count LPC replacement interface
- Simplified routing, low-cost BMC enabling and BOM reduction
- Supports PCH or BMC-attached Flash sharing, GPIO virtualization

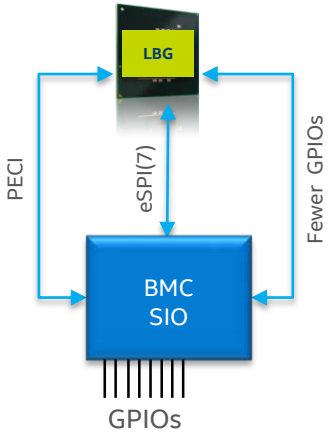
Power management:

- ASPM S-states and DeepS5 support. 2b VID control of logic core voltage
- Independent host, CSME, and IE operating states

RAS:

- ADR: Asynchronous DRAM Self-Refresh
- DWR: Demoted Warm Reset





INTEL® XEON® SCALABLE PROCESSOR

The Secure, Agile, Next-Generation Platform for Multi-Cloud Infrastructures

PERVASIVE PERFORMANCE FOR ACTIONABLE INSIGHTS

Skylake-SP cores

Intel® AVX-512

Feeds: UPI, 6x DDR4, 3x16 PCIe, Intel® SSDs

Integration: Intel® Ethernet / Omni-Path / Intel® QuickAssist / FPGA

SECURITY WITHOUT COMPROMISE

Intel® AVX-512

PPK, MPX, MBE

Intel® QAT w/ Secure Key Management

Intel® Trusted Infrastructure

Intel® Boot Guard

AGILE SERVICE DELIVERY

Intel® Volume Management Device Technology

Intel® RAS

Open Stack Software Optimizations

intel®

BACKUP

New Memory Latency and Bandwidth Optimizations

CACHE & MEMORY

Sub-NUMA Cluster Mode

Associates LLC slice with nearest memory controller

Applications use NUMA primitives to achieve lower LLC/memory latency

XPT Prefetch

Core miss initiates local memory access in parallel with LLC access

Uses history-based prediction to avoid unnecessary prefetches

Local and Remote Direct-to-Core

Data sent directly from memory controller or UPI to requesting core

INTEL® UPI

UPI Prefetch

Similar to XPT prefetch, but for UPI requests from remote socket

Direct-to-UPI

For UPI requests from remote socket, memory controller directly sends data to UPI instead of going through CHA

UPI Optimizations

HitME cache – directory cache for frequently used lines

IO Directory Cache – directory cache for remote IO writes

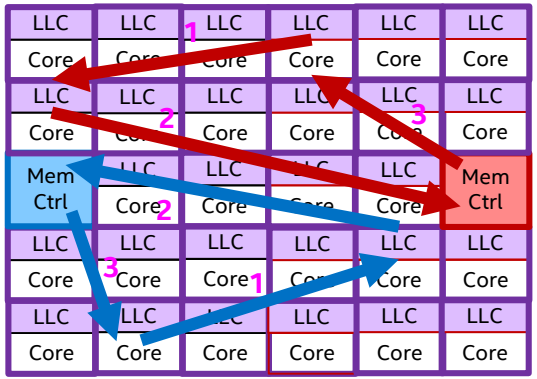
Opportunistic Snoop Broadcast (OSB) – Avoids directory lookup and update for local InvItoE

Sub-NUMA Clusters – 2 SNC Example

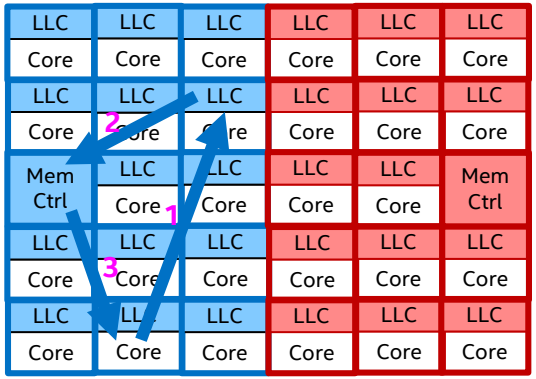
SNC partitions the LLC banks and associates them with memory controller to localize LLC miss traffic

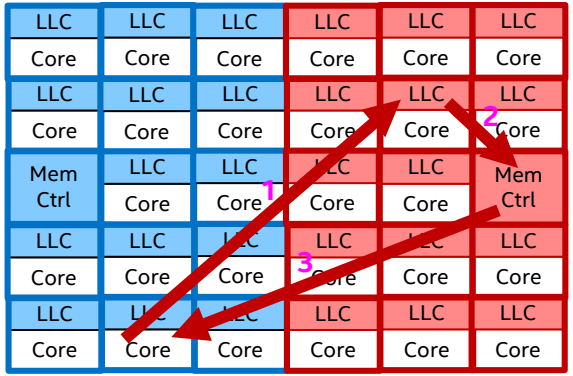
- LLC miss latency to local cluster is smaller
- Mesh traffic is localized, reducing uncore power and sustaining higher BW

Without SNC



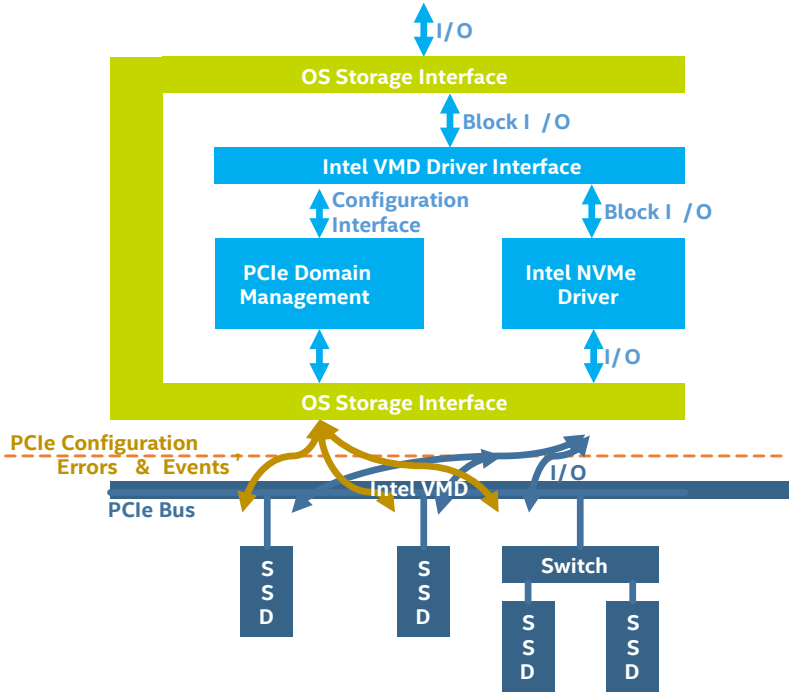
Local SNC Access





Remote SNC Access

Intel® Volume Management Device (Intel® VMD)



Intel® VMD is a CPU-integrated device to aggregate NVMe SSDs into a storage volume and enables other storage services such as RAID

- Intel® VMD is an "integrated end point" that stops OS enumeration of devices under it
- Intel® VMD maps entire PCIe* trees into its own address space (a domain)
- Intel® VMD driver sets up and manages the domain (enumerate, event/error handling), but out of fast IO path

ELIMINATES ADDITIONAL COMPONENTS TO PROVIDE A FULL-FEATURE STORAGE SOLUTION

Intel® Speed Shift Technology

Hardware P-state (HWP) is a new capability for cooperative hardware + software performance control

- Hardware monitors activity / scalability and selects frequency at much faster time scale
- Node Manager or OS provides guidance and constraints to direct hardware
  - Energy Performance Preference: A profile of desired balance between power and performance
  - Minimum Performance Level: A performance floor for meeting QoS requirements
  - Maximum Performance Level: A ceiling to limit low priority applications and services

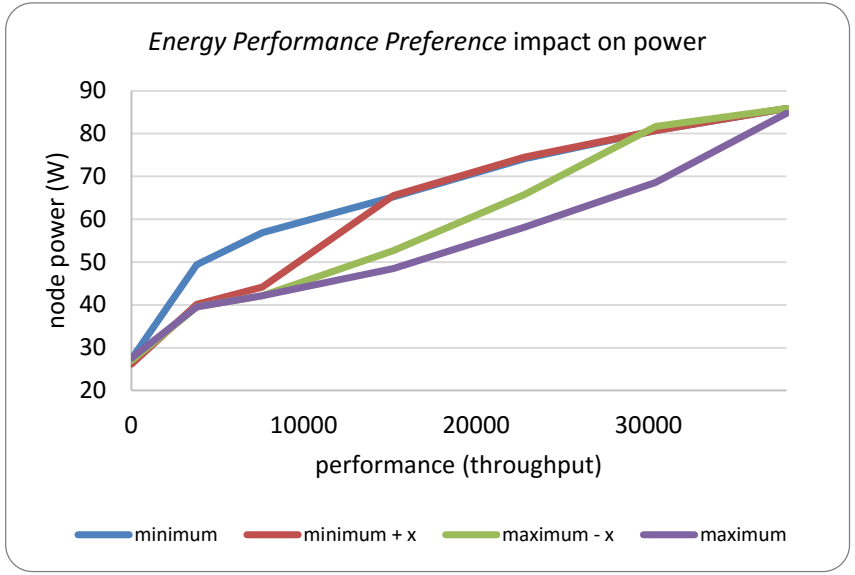
Node Manager or OS is no longer required to monitor activity and update frequency requests at regular intervals

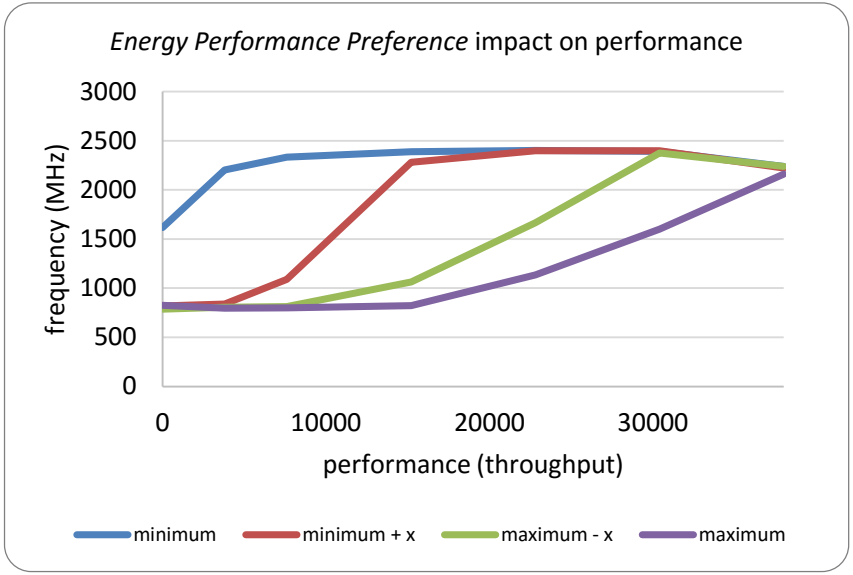
- Node Manager or OS can update HWP guidance based on important events
  - Change in administrator settings, real-time process started, etc.

Intel® Speed Shift Technology

What is the Energy Performance Preference?

- Specifies software preference towards high performance, low power, or some balance between the two





Lewisburg - New Chipset for the Converged Platform

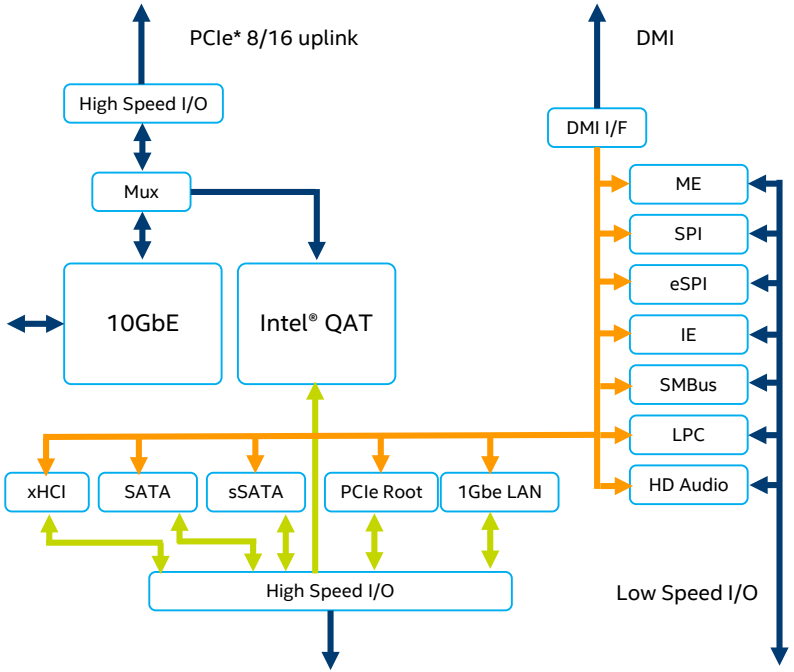
Provides a common footprint from entry level to full configuration with integrated classic server functions

LAN: Intel® Ethernet Connection X722 with up to 4x10GbE

- Supports Network Virtualization Offloads and iWARP RDMA

Comms and Storage accelerator: Intel® QuickAssist Technology (Intel® QAT)

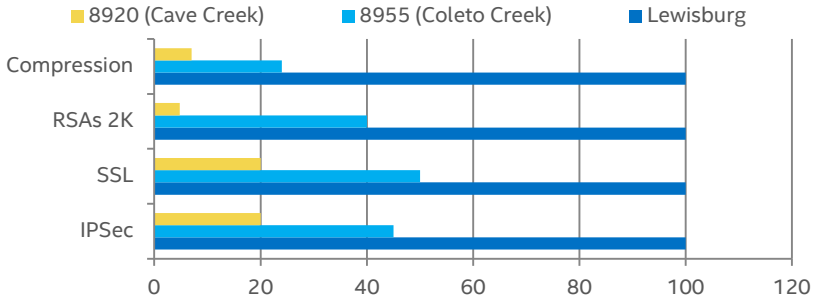
- Crypto / compression up to 100Gbps
- Public key exchange up to 100K Ops



*Other names and brands may be claimed as the property of others.

Lewisburg PCH: Intel® QuickAssist Technology

- Enables standard server platforms to offer ubiquitous compression and security features
- Networking, Storage, Big Data, Cloud, HPC, and Data center applications achieve high performance on:
  - Bulk Ciphers, Authentication
  - Public Key Cryptography
  - Compression



CRYPTOGRAPHIC CIPHERS & AUTHENTICATION

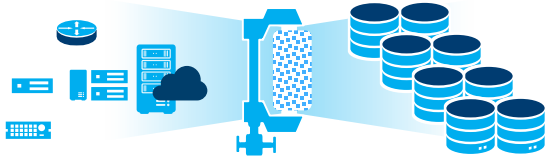
Symmetric Encryption and Authentication

PUBLIC KEY CRYPTOGRAPHY

Secure Key Establishment
(Asymmetric Encryption, Digital Signatures, Key Exchange)

COMPRESSION

Lossless Compression in Flight and at Rest



Configuration: Intel® QAT

QAT API Level tests	1-Node, 1 x Intel® Xeon® Platinum 8180 Processor
Processor	Intel(R) Xeon(R) Gold 6152 H0 (30M Cache, 2.1 GHz)
Vendor	Intel
Nodes	1
Sockets	1
Cores Per Processor	22
Logical Processors	44
Platform	Purley-EP (Lewisburg – B1 stepping)
Accelerator Used	Intel Lewisburg in x24 link mode
Platform Comments	Neon City
Memory DIMMs Slots used/Processor	6
Total Memory	96 GB
Memory DIMM Configuration	16 GB / 2666 MT/s / DDR4 RDIMM
Memory Comments	Kingston 9965662-009.A00G, 16GB, 2Rx8
Network Interface Cards	3x XL710 X710 (Quad Fortville Card) , 12 10Gb ports used
OS	Ubuntu 16.10
OS/Kernel Comments	4.8
Primary / Secondary Software	QAT1.7.Upstream.L.1.0.0-15
Other Configurations	BIOS:PLYDCRB1.86B.0114.R11.16122119, CPU Power&Performance - Perf, P,C and C states disabled, NUMA Enabled
Computer Type	Server
Benchmark	QAT API Level Sample tests

intel®