INTEL® XEON® SCALABLE PROCESSOR ARCHITECTURE DEEP DIVE

Dr.-Ing. Michael Klemm
Senior Application Engineer

Developer Relations Division
Intel Architecture, Graphics and Software

Notices and Disclaimers

Agenda

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

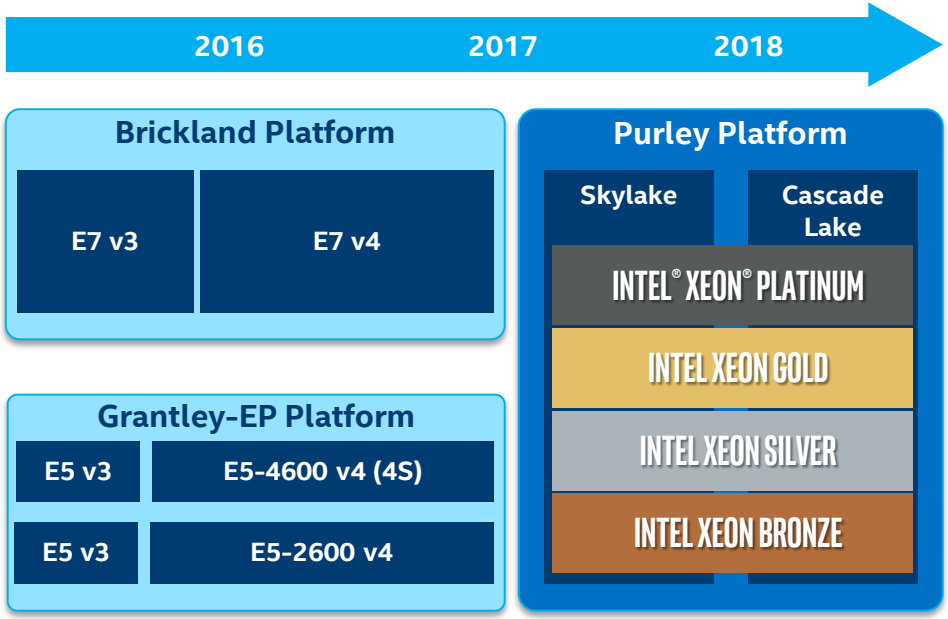
Intel® Xeon® Scalable Processors

Intel® Xeon® Processor E7

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

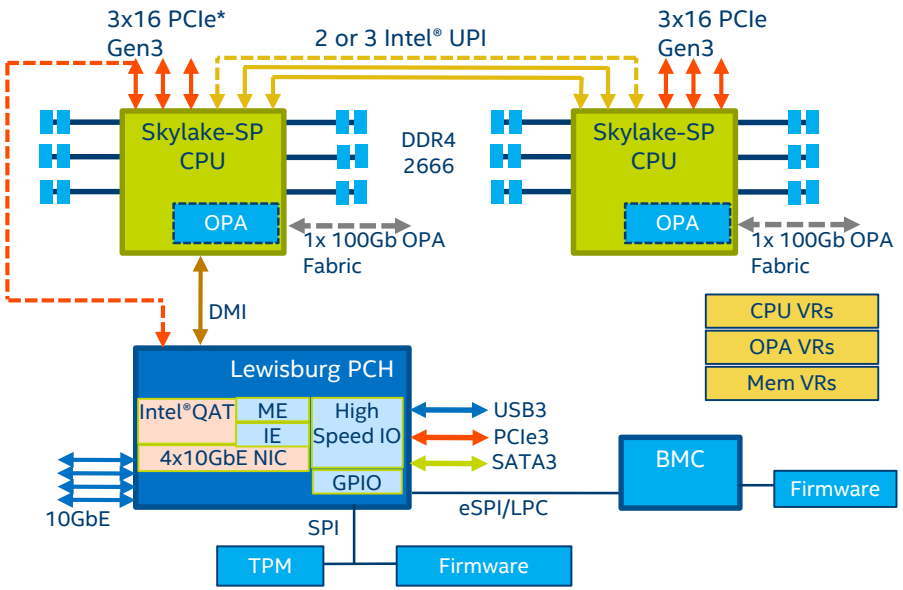
Intel® Xeon® Processor E5

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

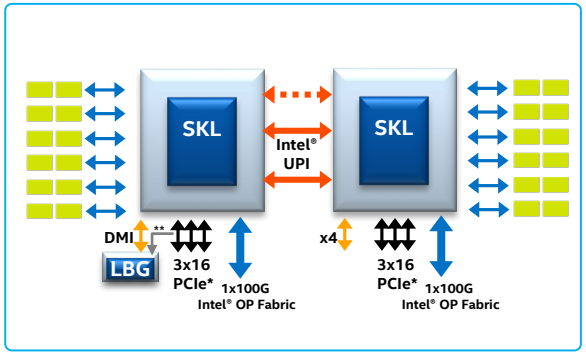


BMC: Baseboard Management Controller	PCH: Intel® Platform Controller Hub	IE: Innovation Engine
Intel® OPA: Intel® Omni-Path Architecture	Intel QAT: Intel® QuickAssist Technology	ME: Manageability Engine
NIC: Network Interface Controller	VMD: Volume Management Device	NTB: Non-Transparent Bridge

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

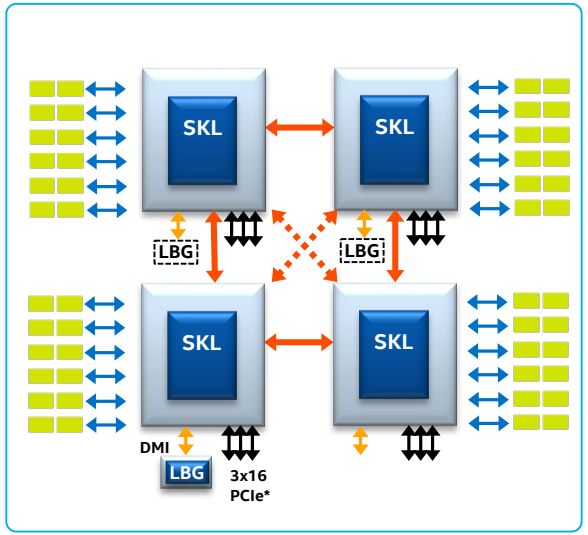
Platform Topologies

2S Configurations



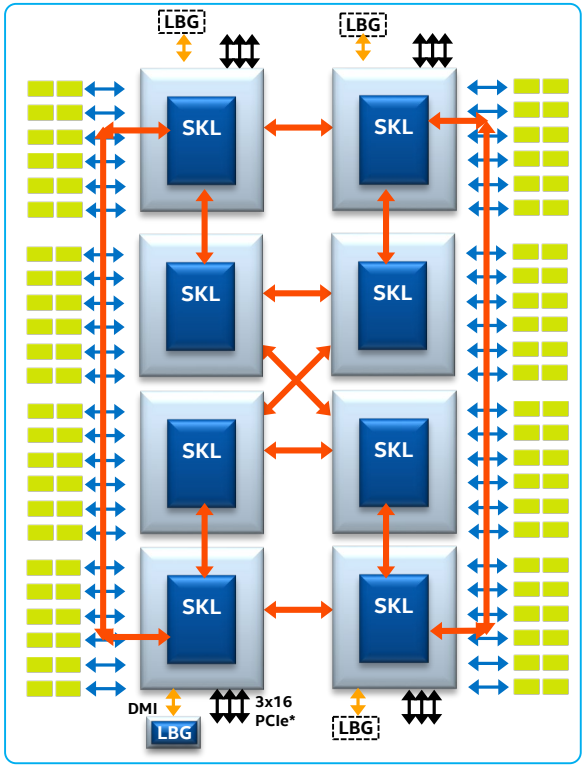
(2S-2UPI & 2S-3UPI shown)

4S Configurations



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

8S Configuration



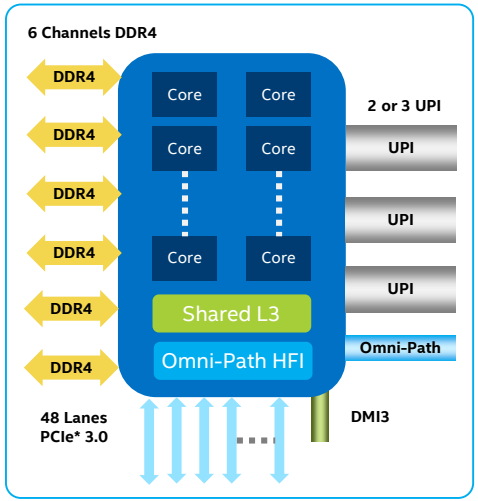
INTEL® XEON® SCALABLE PROCESSOR SUPPORTS CONFIGURATIONS RANGING FROM 2S-2UPI TO 8S

Intel® Xeon® Scalable Processor

Re-architected from the Ground Up

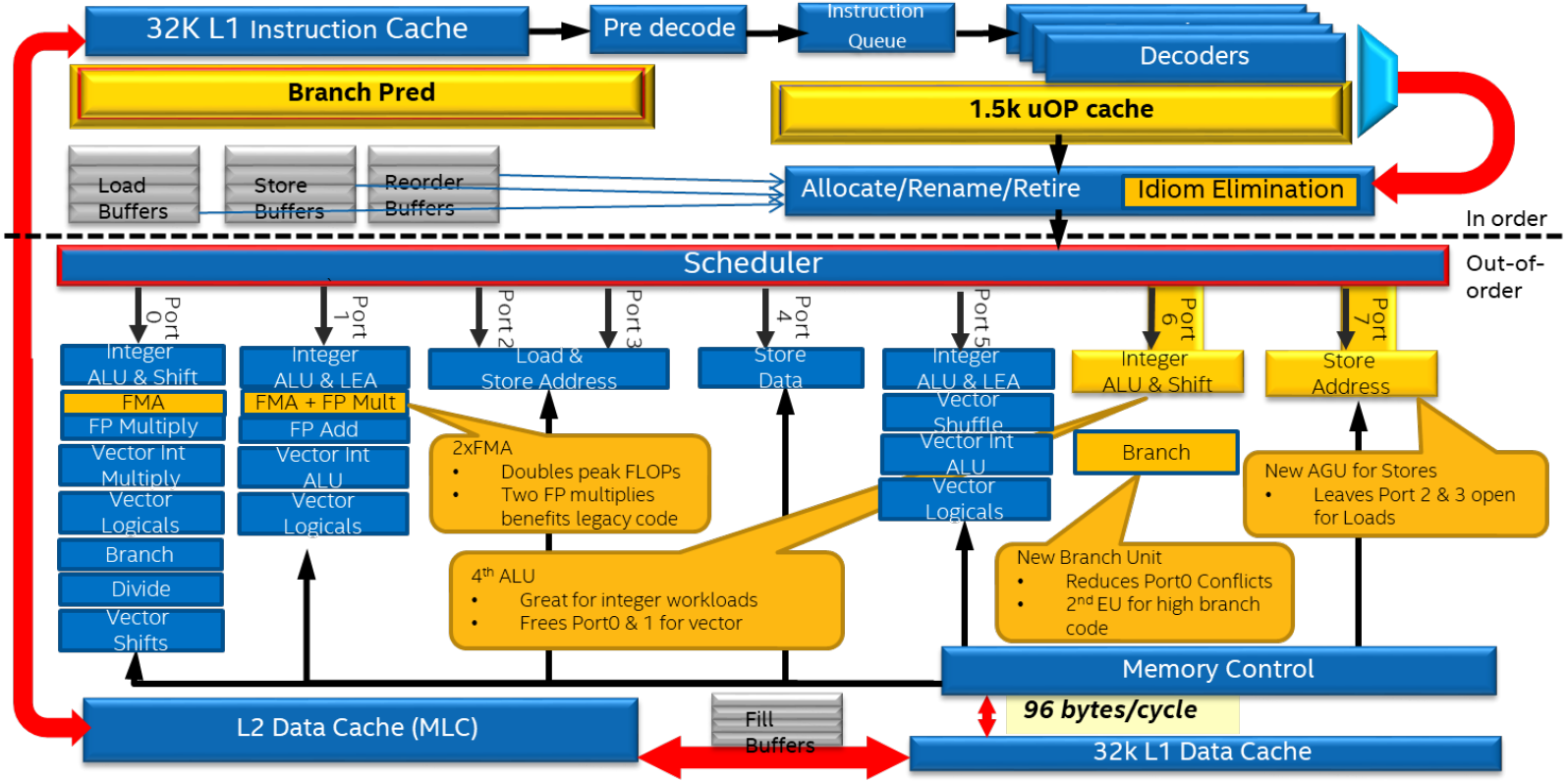
- Skylake core microarchitecture, with data center specific enhancements
- Intel® AVX-512 with 32 DP flops per core
- Data center optimized cache hierarchy – 1MB L2 per core, non-inclusive L3
- New mesh interconnect architecture
- Enhanced memory subsystem
- Modular IO with integrated devices
- New Intel® 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**

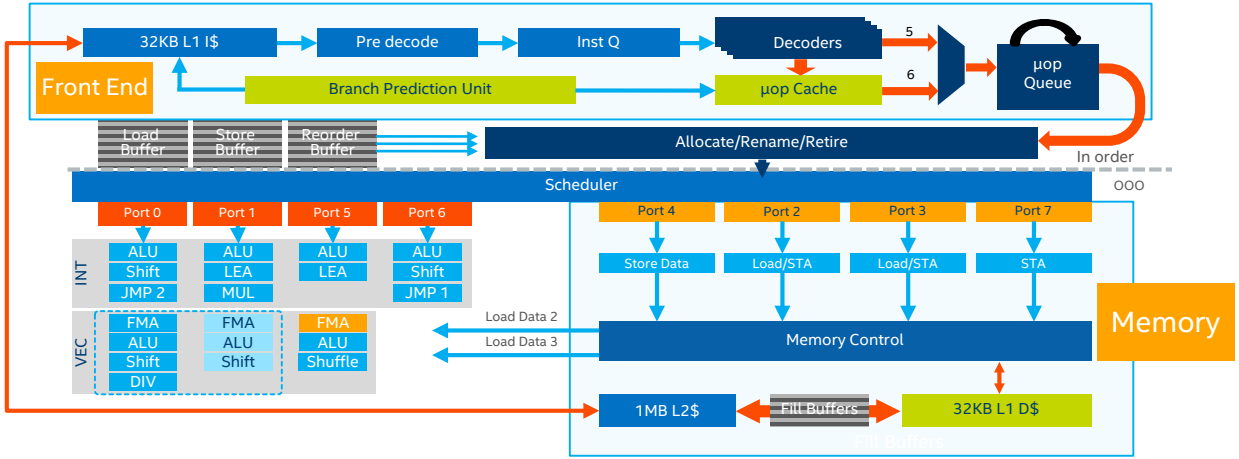


SKYLAKE-SP CORE ARCHITECTURE

Haswell/Broadwell Microarchitecture



Skylake 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

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

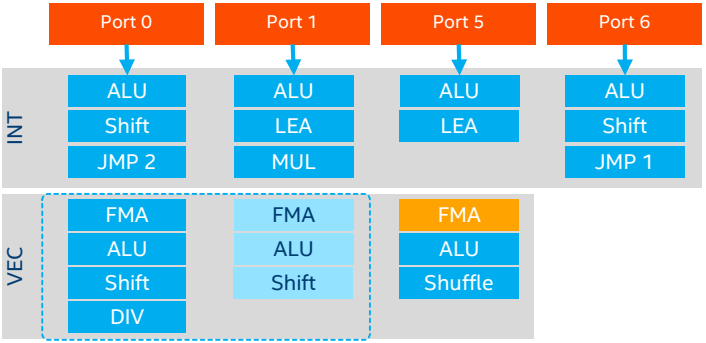
Intel AVX-512 Instruction Types	
AVX-512-F	AVX-512 Foundation Instructions
AVX-512-VL	Vector Length Orthogonality : ability to operate on sub-512 vector sizes
AVX-512-BW	512-bit Byte/Word support
AVX-512-DQ	Additional D/Q/SP/DP instructions (converts, transcendental support, etc.)
AVX-512-CD	Conflict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

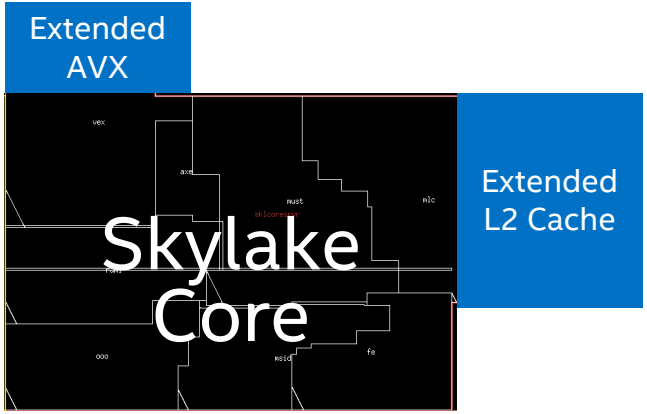
POWERFUL INSTRUCTION SET FOR DATA-PARALLEL COMPUTATION

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



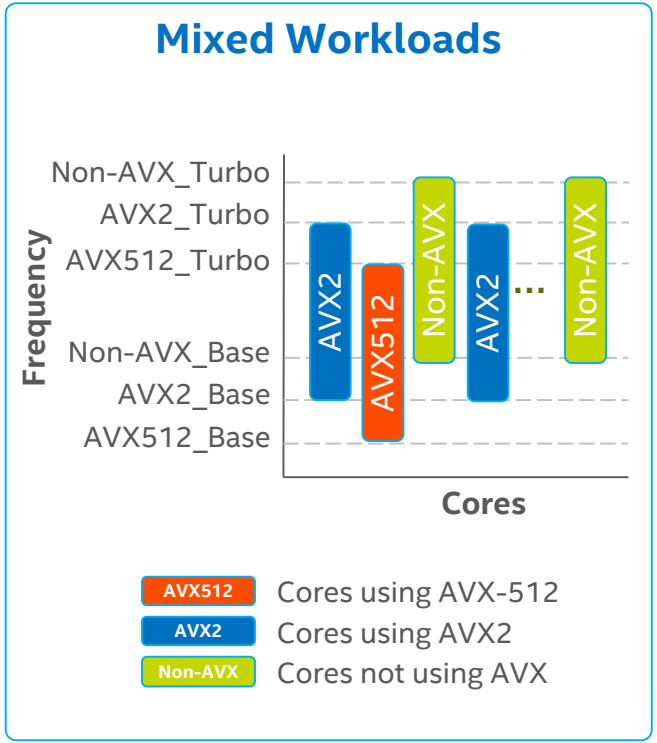


SKYLAKE-SP CORE: OPTIMIZED FOR DATA CENTER WORKLOADS

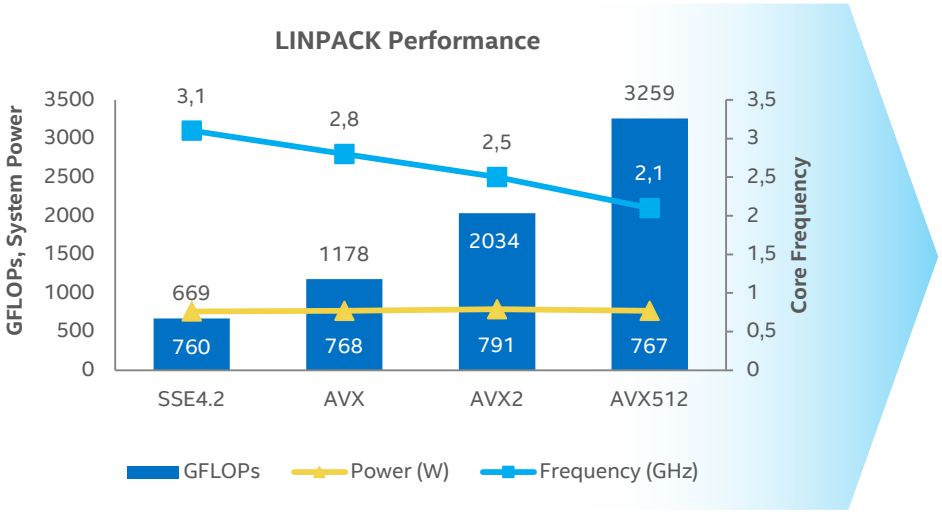
Frequency Behavior While Running Intel® AVX Code

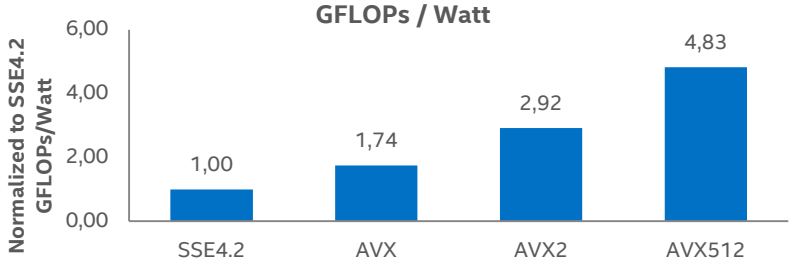
- Cores running non-AVX, Intel® AVX2 light/heavy, and Intel® AVX-512 light/heavy code have different turbo frequency limits
- Frequency of each core is determined independently based on workload demand

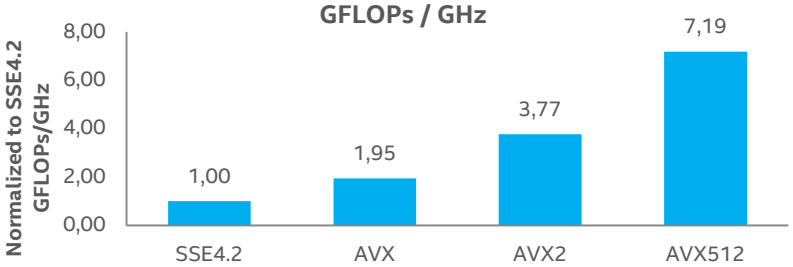
Code Type	All Core Frequency Limit
SSE AVX2-Light (without FP & int-mul)	Non-AVX All Core Turbo
AVX2-Heavy (FP & int-mul) AVX512-Light (without FP & int-mul)	AVX2 All Core Turbo
AVX512-Heavy (FP & int-mul)	AVX512 All Core Turbo



Performance and Efficiency with Intel® AVX-512







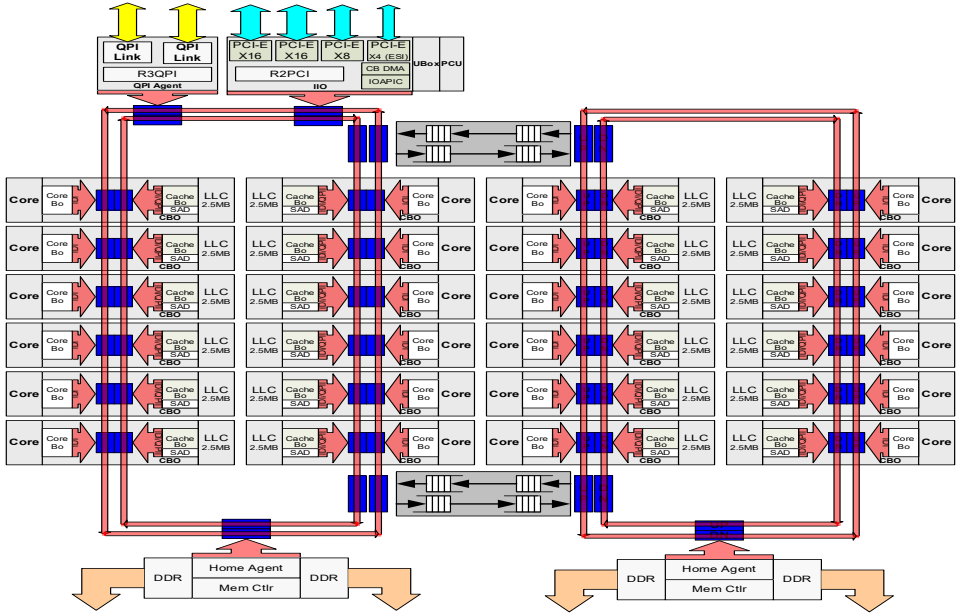
INTEL® AVX-512 DELIVERS SIGNIFICANT PERFORMANCE AND EFFICIENCY GAINS

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.

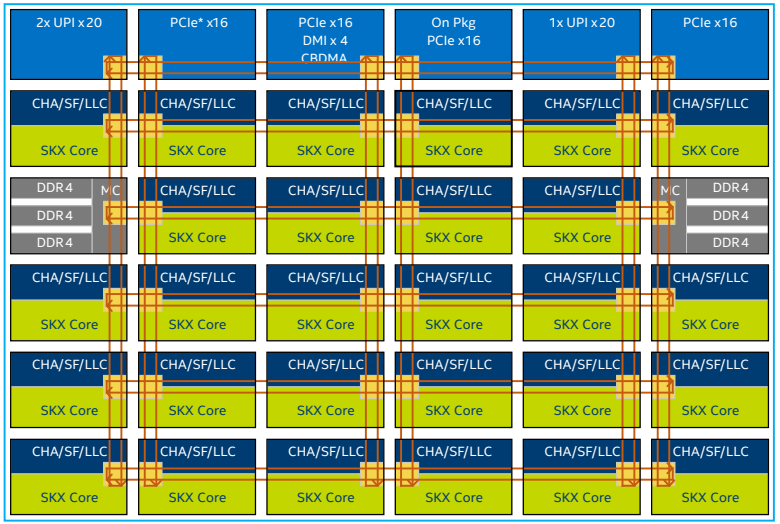
SKYLAKE-SP SOC ARCHITECTURE

New Mesh Interconnect Architecture

Broadwell EX 24-core die



Skylake-SP 28-core die

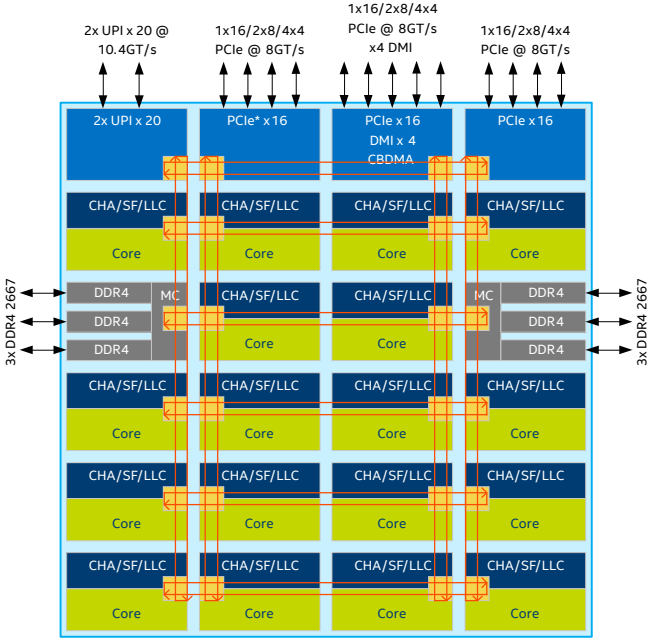


CHA – Caching and Home Agent ; SF – Snoop Filter; LLC – Last Level Cache;
SKX Core – Skylake Server Core; UPI – Intel® UltraPath Interconnect

MESH IMPROVES SCALABILITY WITH HIGHER BANDWIDTH AND REDUCED LATENCIES

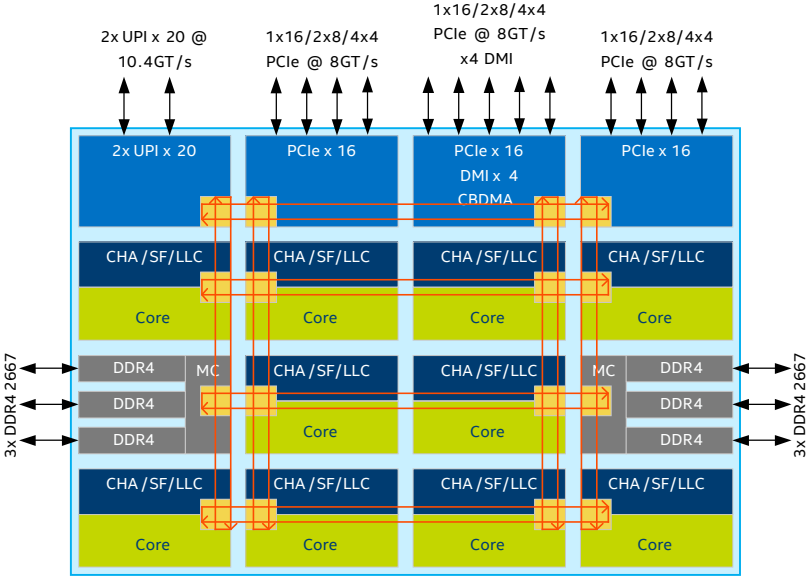
High and Low Core Count Die Configurations

HCC (up to 18 cores)



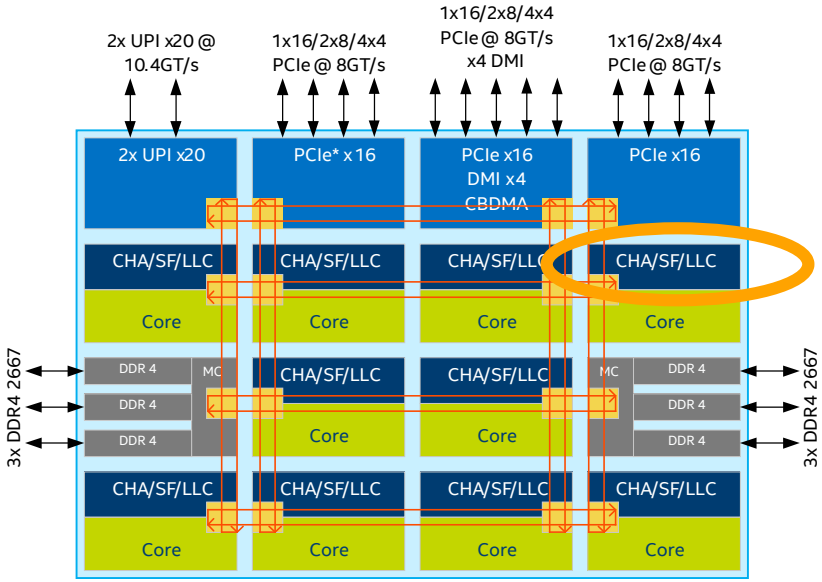
CHA – Caching and Home Agent ; SF – Snoop Filter ; LLC – Last Level Cache ;
Core – Skylake-SP Core; UPI – Intel® UltraPath Interconnect

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

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

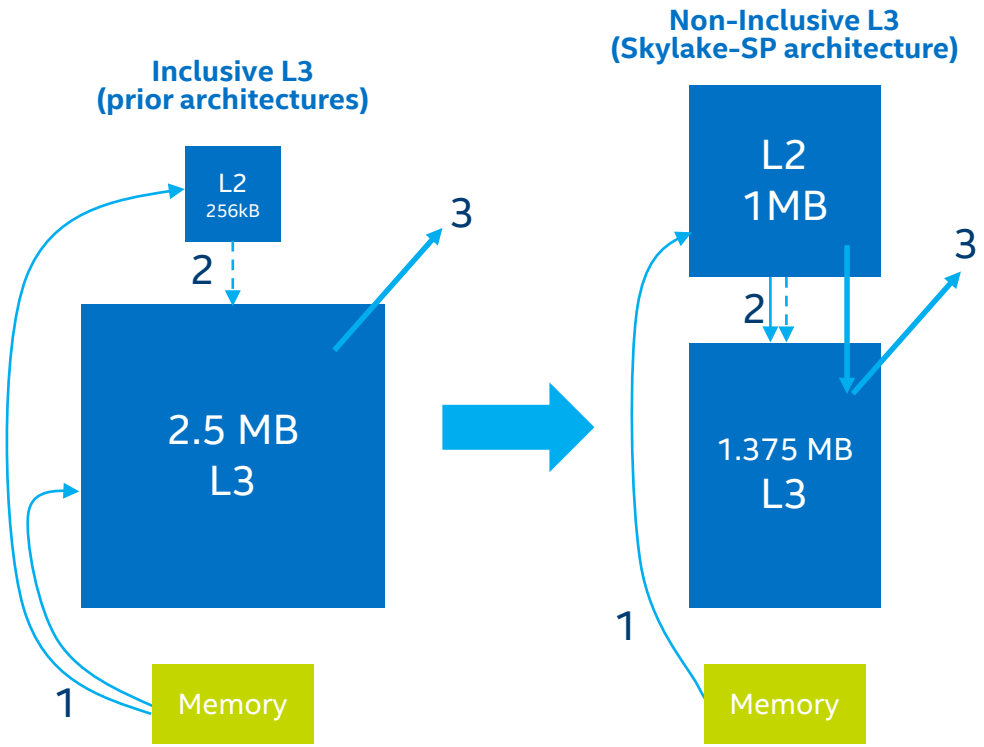
DISTRIBUTED CHA ARCHITECTURE SUSTAINS HIGHER BANDWIDTH AND LOWERS LATENCY

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

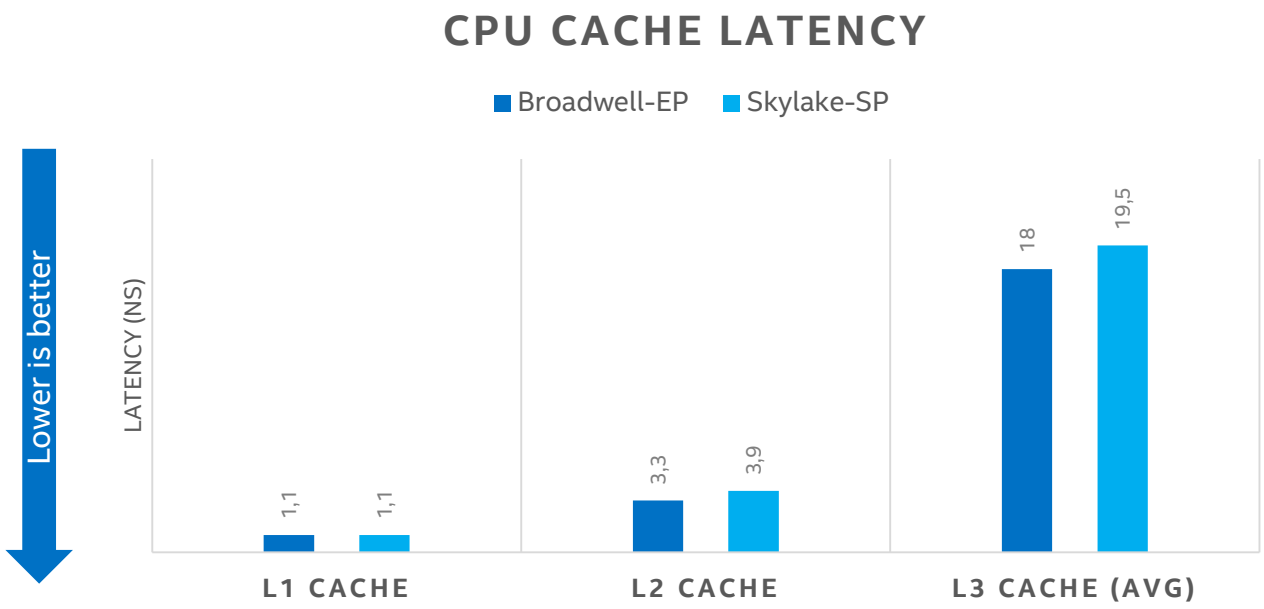


1. Memory reads fill directly to the L2, no longer to both the L2 and L3
2. When a L2 line needs to be removed, both modified and unmodified lines are written back
3. Data shared across cores are copied into the L3 for servicing future L2 misses

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

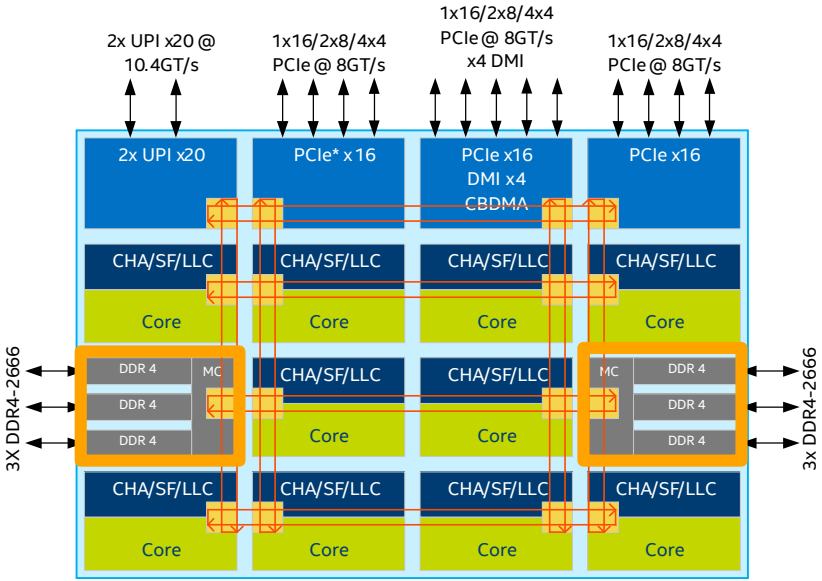


Skylake-SP L2 cache latency has increased by 2 cycles for a 4x larger L2

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)

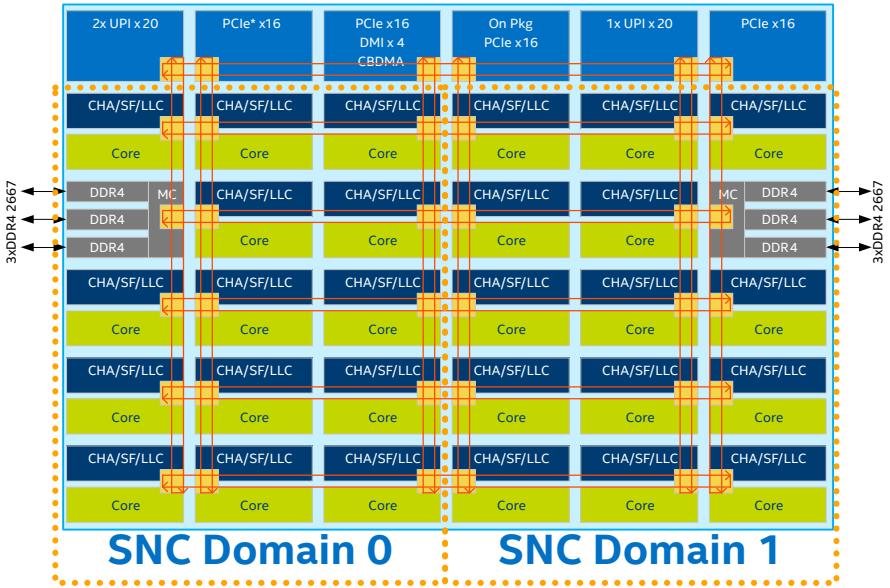
SIGNIFICANT MEMORY BANDWIDTH AND CAPACITY IMPROVEMENTS

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

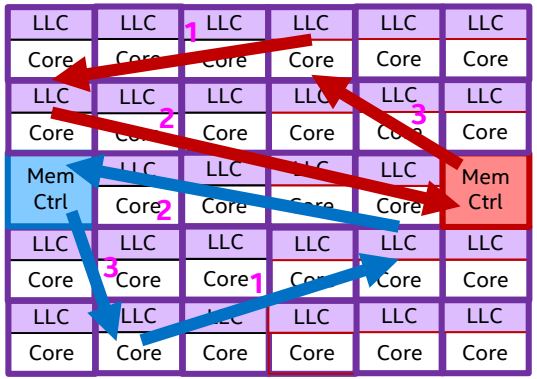


Sub-NUMA Clusters – 2 SNC Example

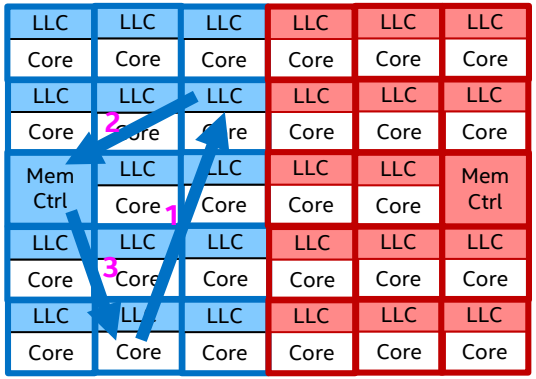
SNC partitions the LLC banks and associates them with memory controller to localize LLC miss traffic

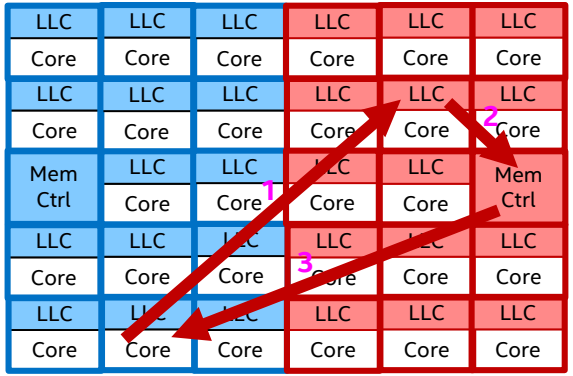
- LLC miss latency to local cluster is smaller
- Mesh traffic is localized, reducing uncore power and sustaining higher BW

Without SNC



Local SNC Access

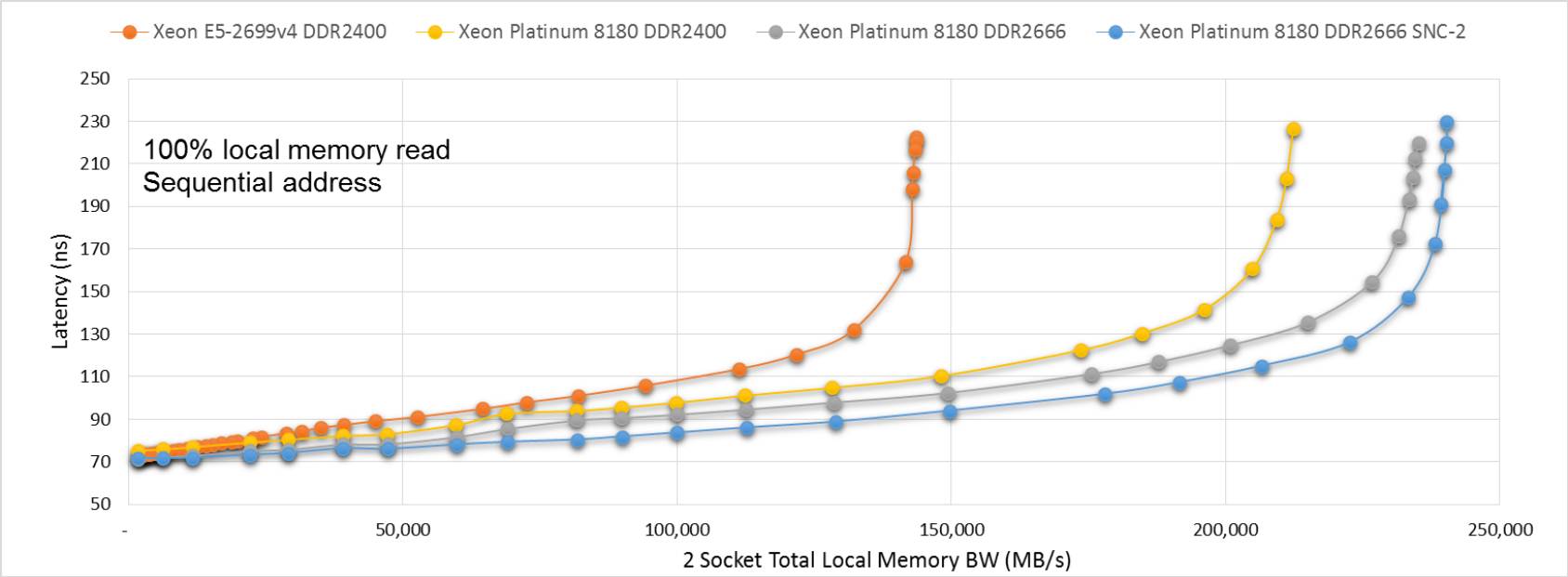




Remote SNC Access

Memory Performance

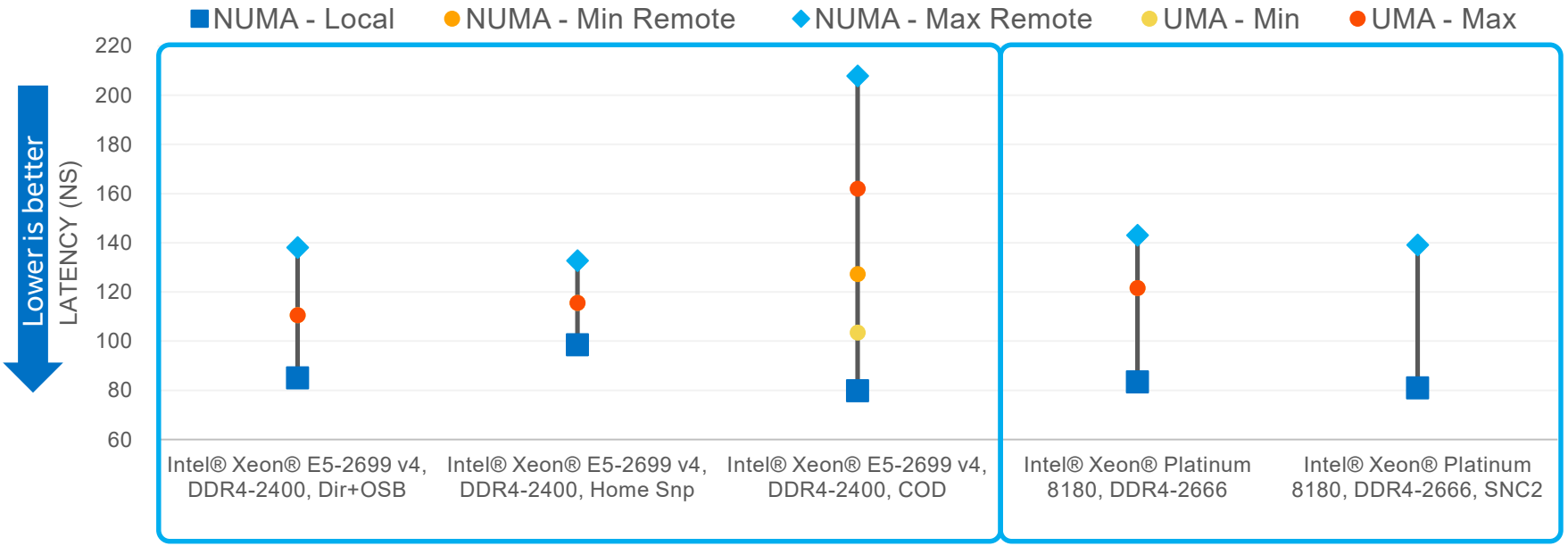
Bandwidth-Latency Profile



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

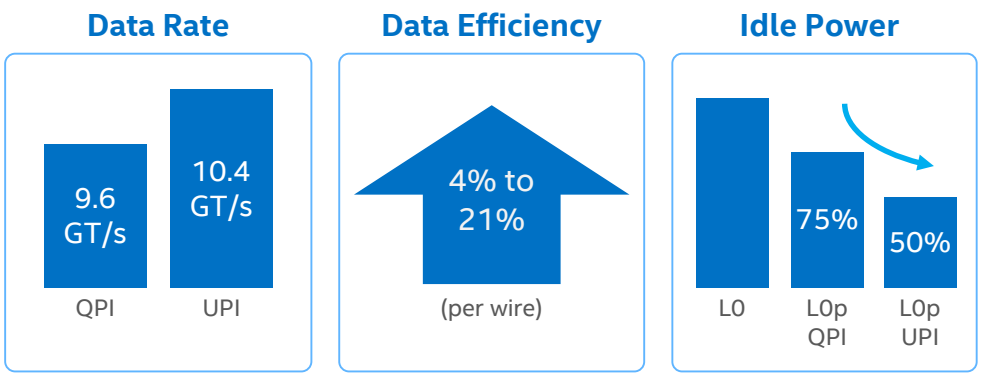
Core to Memory Latency

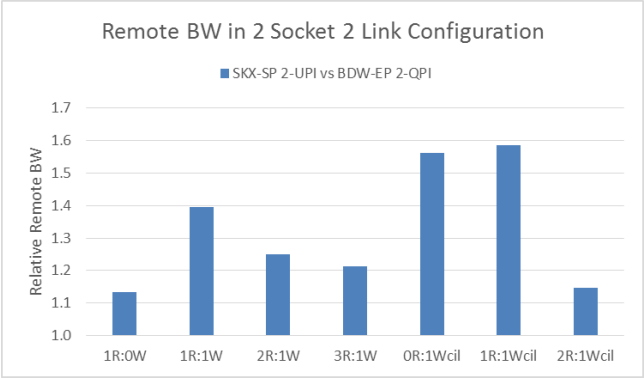


Source as of June 2017: Intel internal measurements on platform with Xeon Platinum 8180, Turbo enabled, UPI=10.4, 6x32GB DDR4-2666, 1 DPC, and platform with E5-2699 v4, Turbo enabled, 4x32GB DDR4-2400, RHEL 7.0. Software and workloads used in performance tests may have been optimized for performance 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.

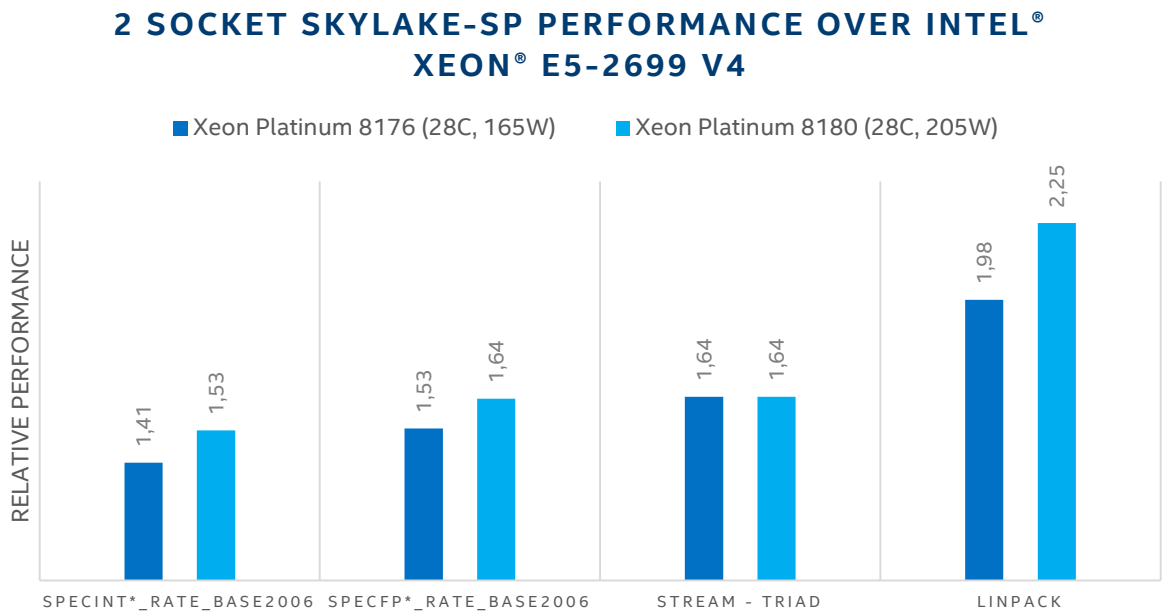
SKYLAKE-SP CPU WRAP UP

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.

intel®