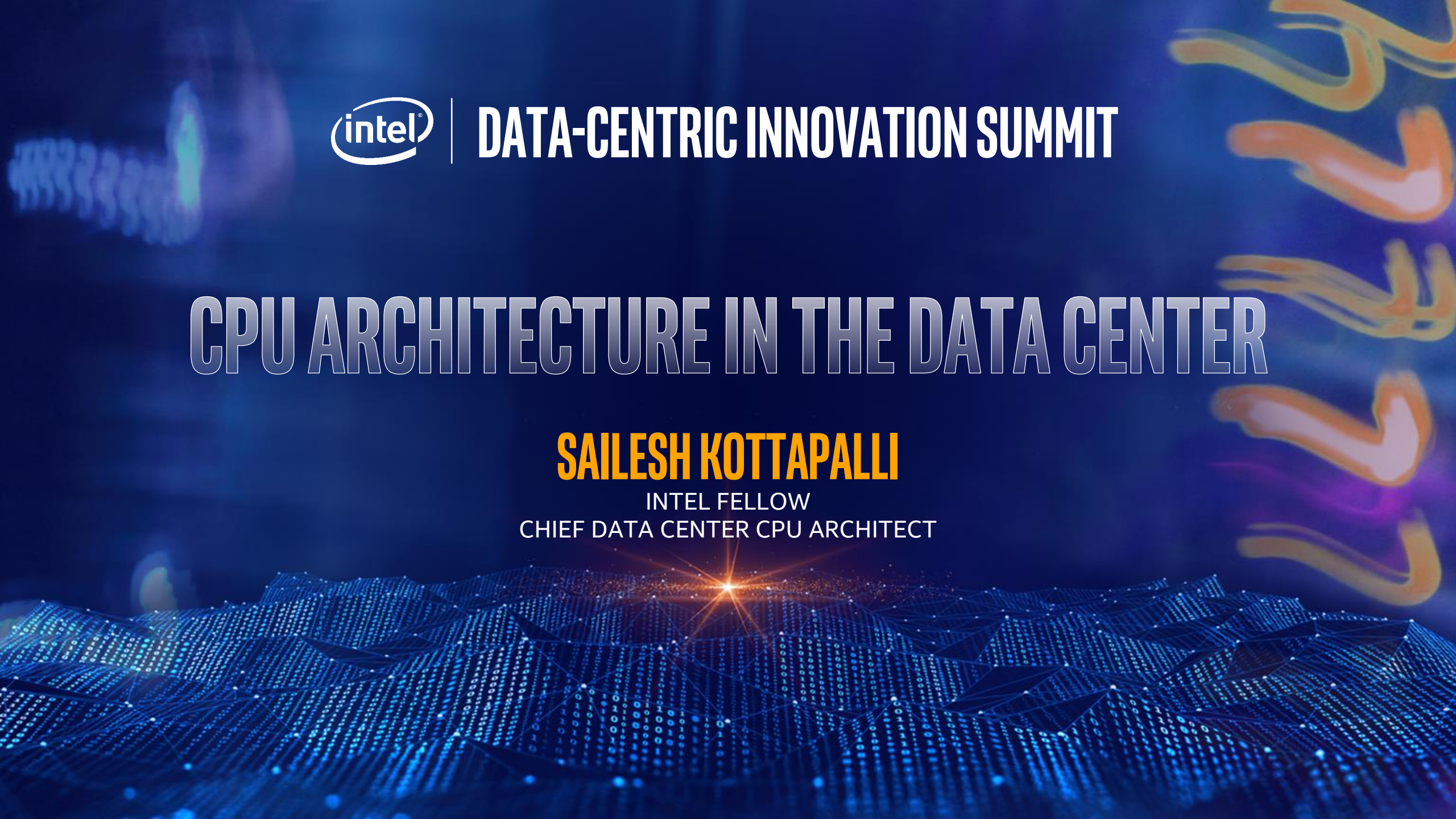

DATA-CENTRIC INNOVATION SUMMIT
AUGUST 8, 2018 | SANTA CLARA, CA


intel
DATA-CENTRIC INNOVATION SUMMIT
CPU ARCHITECTURE IN THE DATA CENTER
SAILESH KOTTAPALLI
INTEL FELLOW
CHIEF DATA CENTER CPU ARCHITECT

THE FUNDAMENTALS OF DATA CENTER COMPUTING

PER-CORE PERFORMANCE (PCP)

Performance of a software context running on each physical or virtual core when the entire server is running.

Minimum PCP:
PCP needed to process a single web transaction within the latency SLA.

THROUGHPUT PERFORMANCE (TPT)

Cumulative throughput performance that we can achieve from the processor or server.

Effective TPT:
Rate of processing multiple transactions while meeting the latency SLA on each web transaction.

TPT = PCP x (number of cores)

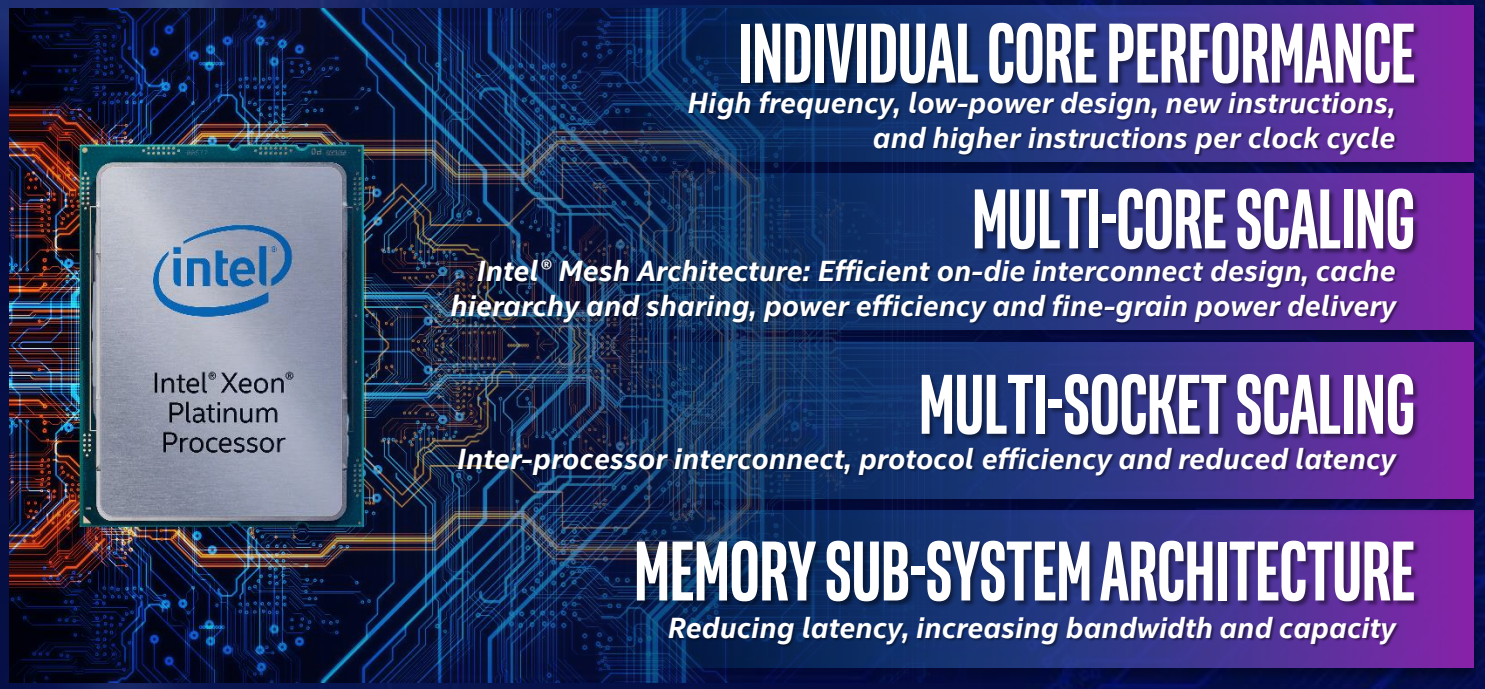
DATA CENTER REQUIRES AN OPTIMAL BALANCE

RELEVANCE OF PER-CORE PERFORMANCE

- PERFORMANCE & RESPONSE TIME
- FLEXIBILITY AND ELASTIC COMPUTE
- PERFORMANCE AT SCALE
- AMDAHL'S LAW
- SOFTWARE TCO

LEADERSHIP PCP CRITICAL TO OVERALL DATA CENTER PERFORMANCE

SCALING PCP AND TPT



INDIVIDUAL CORE PERFORMANCE

High frequency, low-power design, new instructions, and higher instructions per clock cycle

MULTI-CORE SCALING

Intel® Mesh Architecture: Efficient on-die interconnect design, cache hierarchy and sharing, power efficiency and fine-grain power delivery

MULTI-SOCKET SCALING

Inter-processor interconnect, protocol efficiency and reduced latency

MEMORY SUB-SYSTEM ARCHITECTURE

Reducing latency, increasing bandwidth and capacity

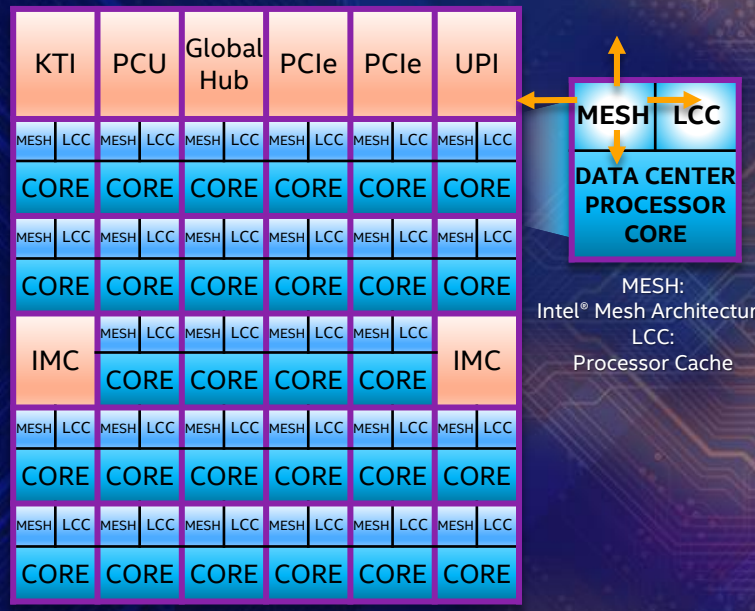
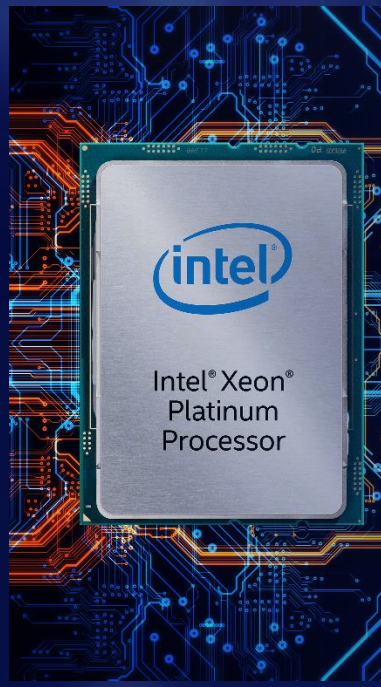


Illustration: Intel® Xeon® Scalable Processor

HIGH PER-CORE PERFORMANCE → LEADERSHIP EFFECTIVE THROUGHPUT

SCALING PCP AND TPT



INDIVIDUAL CORE PERFORMANCE

High frequency, low-power design, new instructions, and higher instructions per clock cycle

MULTI-CORE SCALING

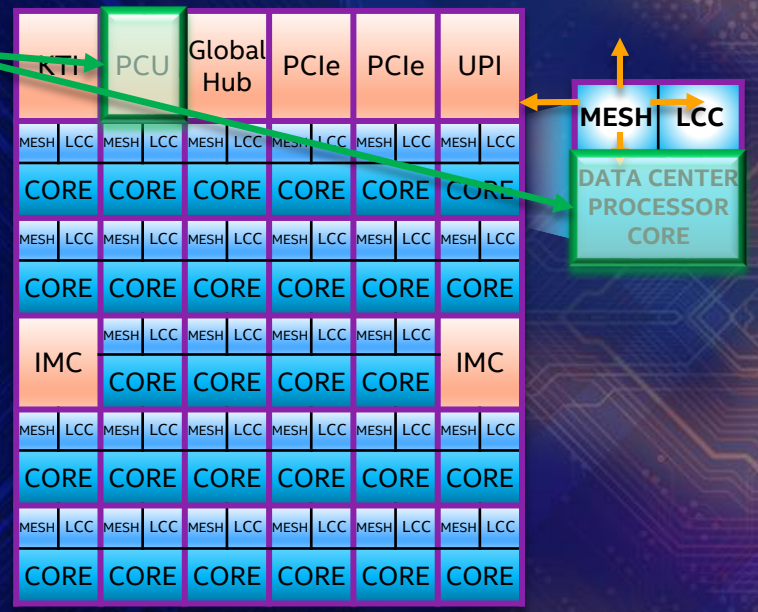
Intel® Mesh Architecture: Efficient on-die interconnect design, cache hierarchy and sharing, power efficiency and fine-grain power delivery

MULTI-SOCKET SCALING

Inter-processor interconnect, protocol efficiency and reduced latency

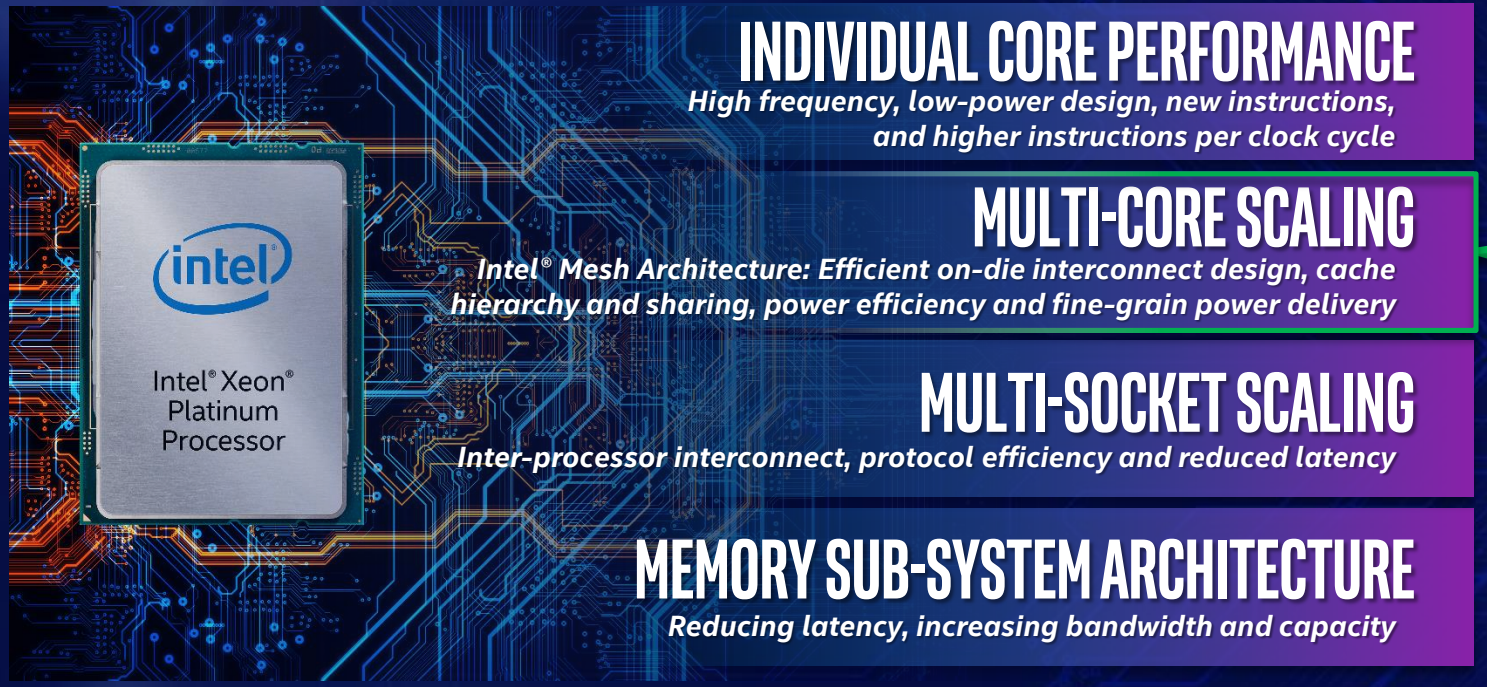
MEMORY SUB-SYSTEM ARCHITECTURE

Reducing latency, increasing bandwidth and capacity



HIGH PER-CORE PERFORMANCE → LEADERSHIP EFFECTIVE THROUGHPUT

SCALING PCP AND TPT



INDIVIDUAL CORE PERFORMANCE

High frequency, low-power design, new instructions, and higher instructions per clock cycle

MULTI-CORE SCALING

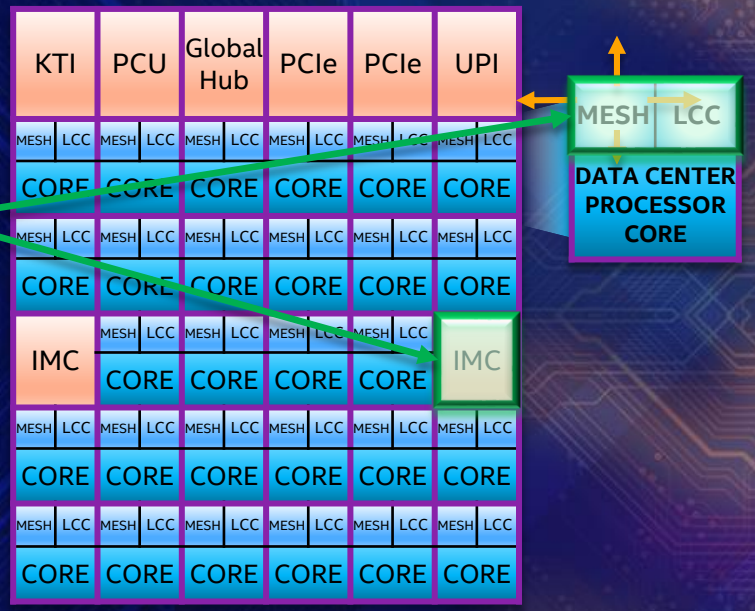
Intel® Mesh Architecture: Efficient on-die interconnect design, cache hierarchy and sharing, power efficiency and fine-grain power delivery

MULTI-SOCKET SCALING

Inter-processor interconnect, protocol efficiency and reduced latency

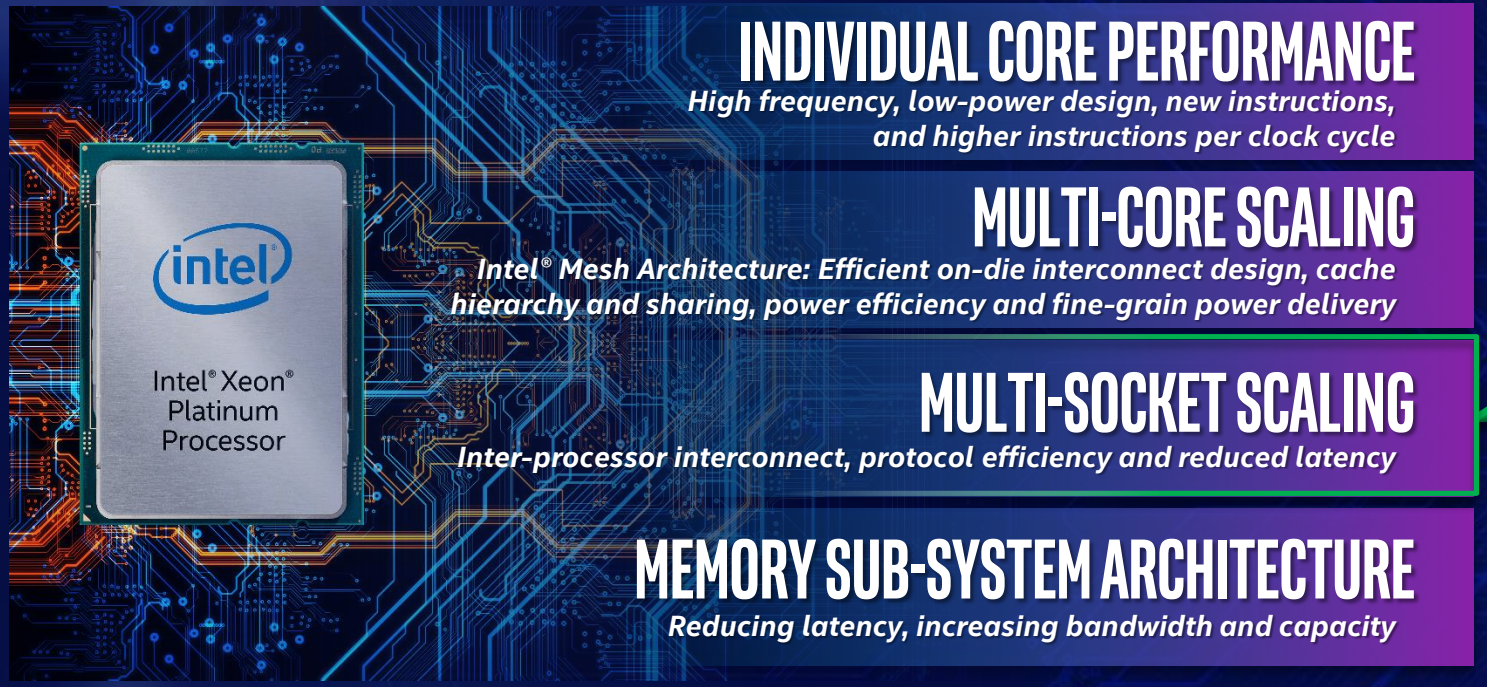
MEMORY SUB-SYSTEM ARCHITECTURE

Reducing latency, increasing bandwidth and capacity



HIGH PER-CORE PERFORMANCE → LEADERSHIP EFFECTIVE THROUGHPUT

SCALING PCP AND TPT



INDIVIDUAL CORE PERFORMANCE

High frequency, low-power design, new instructions, and higher instructions per clock cycle

MULTI-CORE SCALING

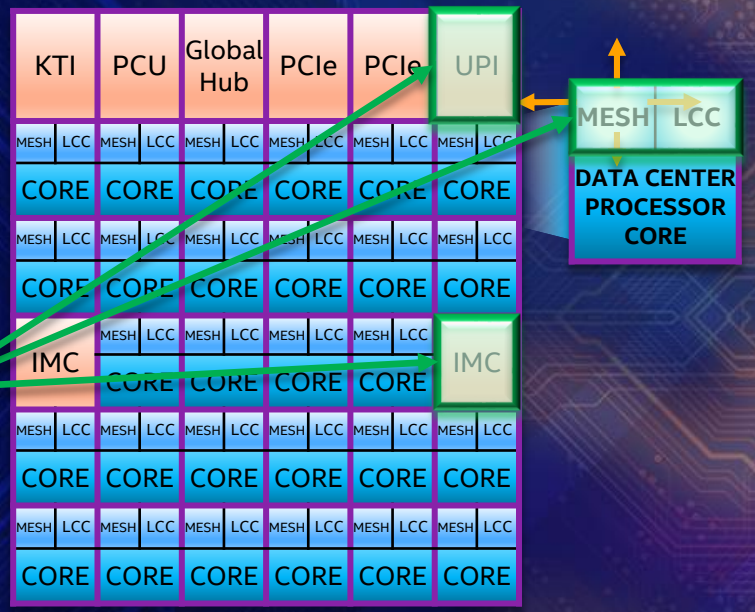
Intel® Mesh Architecture: Efficient on-die interconnect design, cache hierarchy and sharing, power efficiency and fine-grain power delivery

MULTI-SOCKET SCALING

Inter-processor interconnect, protocol efficiency and reduced latency

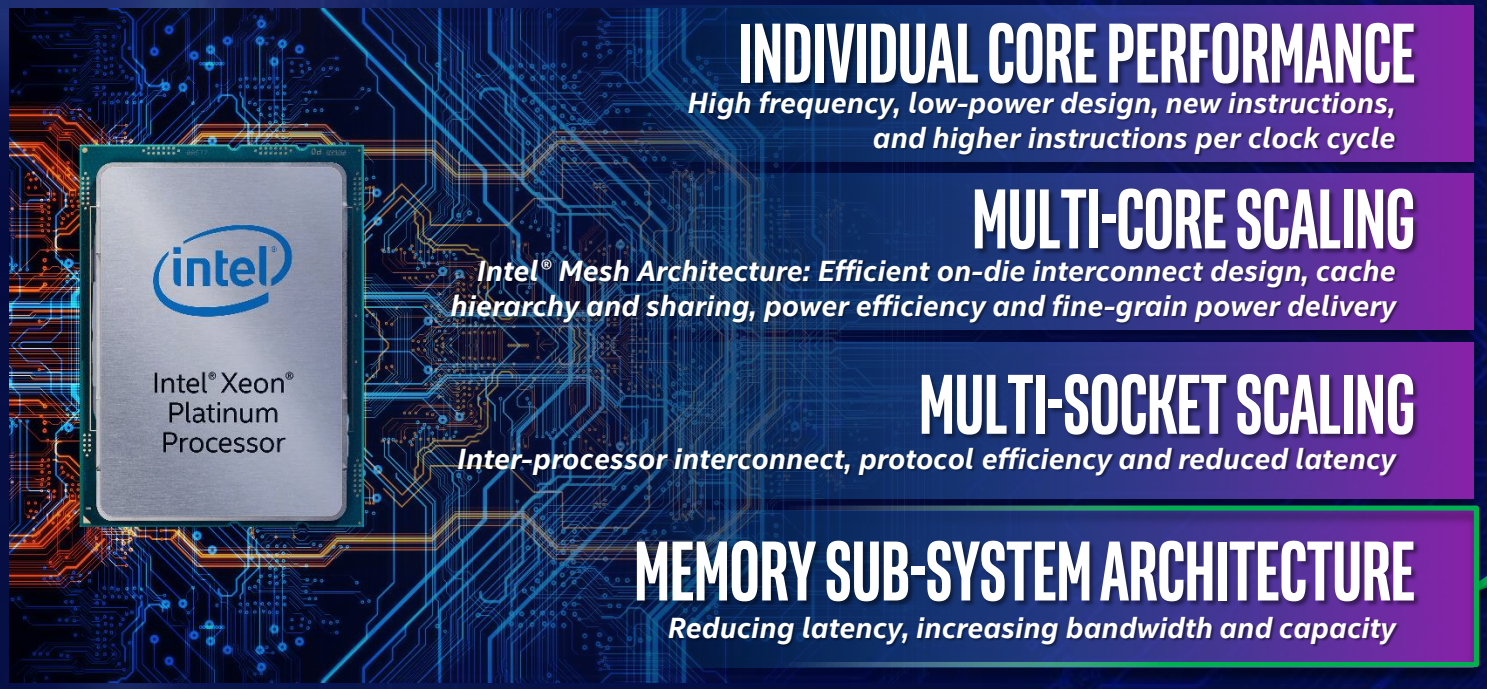
MEMORY SUB-SYSTEM ARCHITECTURE

Reducing latency, increasing bandwidth and capacity



HIGH PER-CORE PERFORMANCE → LEADERSHIP EFFECTIVE THROUGHPUT

SCALING PCP AND TPT



INDIVIDUAL CORE PERFORMANCE

High frequency, low-power design, new instructions, and higher instructions per clock cycle

MULTI-CORE SCALING

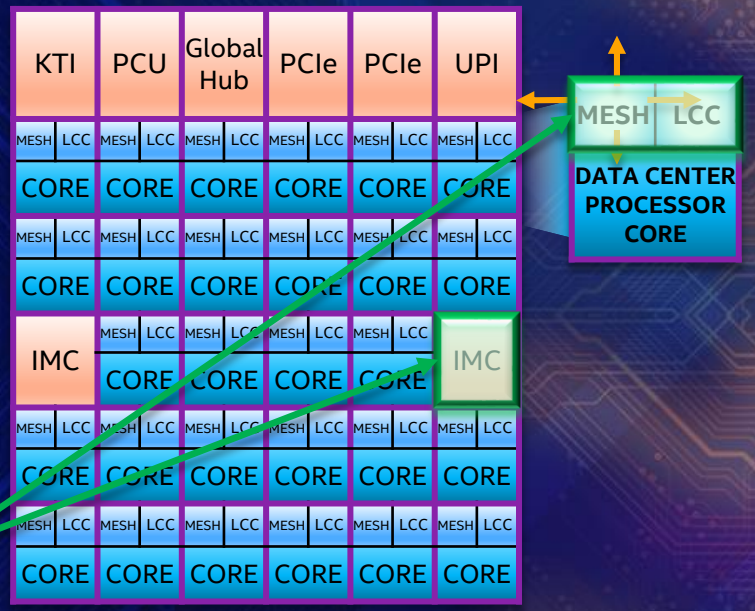
Intel® Mesh Architecture: Efficient on-die interconnect design, cache hierarchy and sharing, power efficiency and fine-grain power delivery

MULTI-SOCKET SCALING

Inter-processor interconnect, protocol efficiency and reduced latency

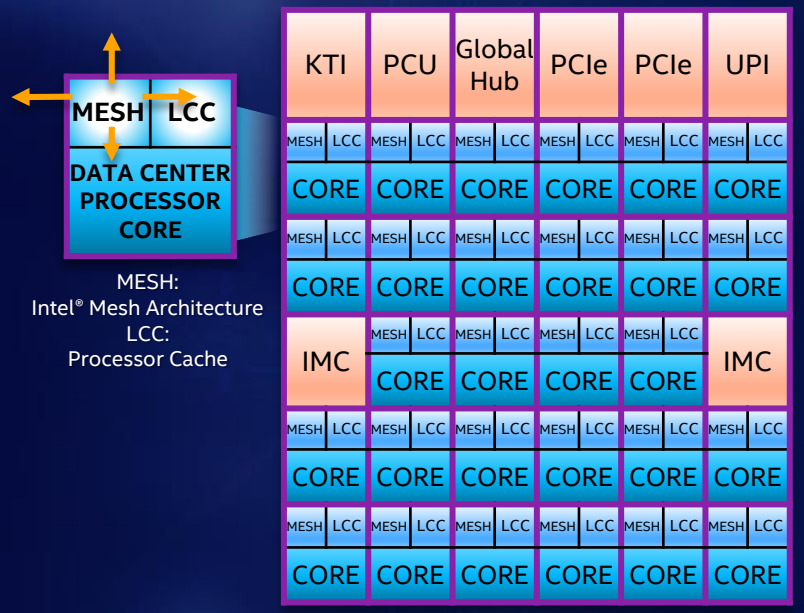
MEMORY SUB-SYSTEM ARCHITECTURE

Reducing latency, increasing bandwidth and capacity



HIGH PER-CORE PERFORMANCE → LEADERSHIP EFFECTIVE THROUGHPUT

UTILIZATION IN THE DATA CENTER "COMPUTER"



MESH:
Intel® Mesh Architecture
LCC:
Processor Cache

Illustration: Intel® Xeon® Scalable Processor

CONSOLIDATION

Virtualization: VT performance, VM migration, handling large VM sizes. Profiling: Cache and memory bandwidth monitoring, instruction tracing. Availability: Intel® Run Sure Technology

PERFORMANCE CONSISTENCY

Cache capacity enforcement, memory bandwidth allocation, turbo/throttle isolation, code and data isolation

LOW JITTER

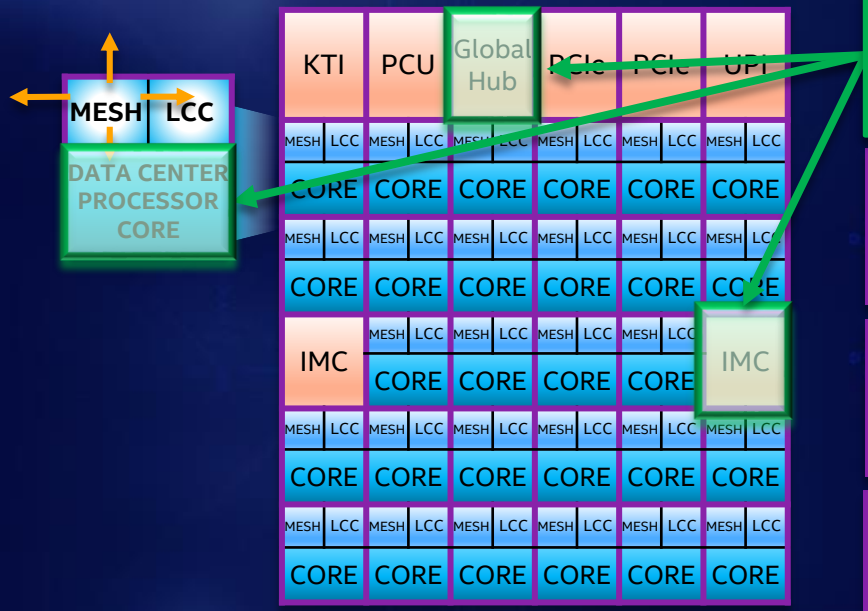
Low performance variability: I/O, memory, compute = consistently lower latency
Non-blocking frequency transitions

HIGH DATA CENTER EFFICIENCY

Intel® Data Direct I/O Technology, CB-DMA, DPDK, Crypto/Compression with Intel® AVX instructions and Intel® QuickAssist Technology, Erasure Coding, RAID

PROCESSOR ARCHITECTURE DESIGNED FOR HIGH UTILIZATION

UTILIZATION IN THE DATA CENTER "COMPUTER"



CONSOLIDATION

Virtualization: VT performance, VM migration, handling large VM sizes. Profiling: Cache and memory bandwidth monitoring, instruction tracing. Availability: Intel® Run Sure Technology

PERFORMANCE CONSISTENCY

Cache capacity enforcement, memory bandwidth allocation, turbo/throttle isolation, code and data isolation

LOW JITTER

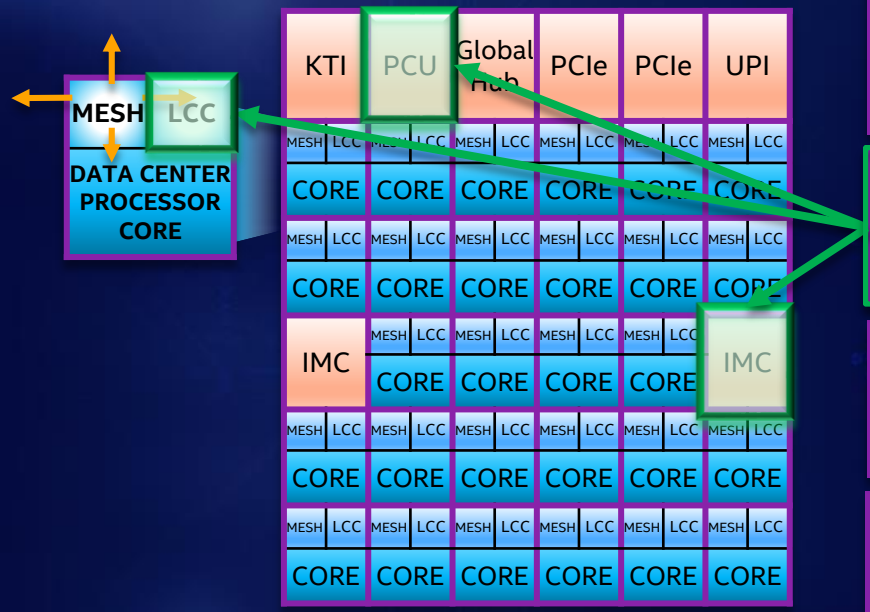
Low performance variability: I/O, memory, compute = consistently lower latency
Non-blocking frequency transitions

HIGH DATA CENTER EFFICIENCY

Intel® Data Direct I/O Technology, CB-DMA, DPDK, Crypto/Compression with Intel® AVX instructions and Intel® QuickAssist Technology, Erasure Coding, RAID

PROCESSOR ARCHITECTURE DESIGNED FOR HIGH UTILIZATION

UTILIZATION IN THE DATA CENTER "COMPUTER"



CONSOLIDATION

Virtualization: VT performance, VM migration, handling large VM sizes. Profiling: Cache and memory bandwidth monitoring, instruction tracing. Availability: Intel® Run Sure Technology

PERFORMANCE CONSISTENCY

Cache capacity enforcement, memory bandwidth allocation, turbo/throttle isolation, code and data isolation

LOW JITTER

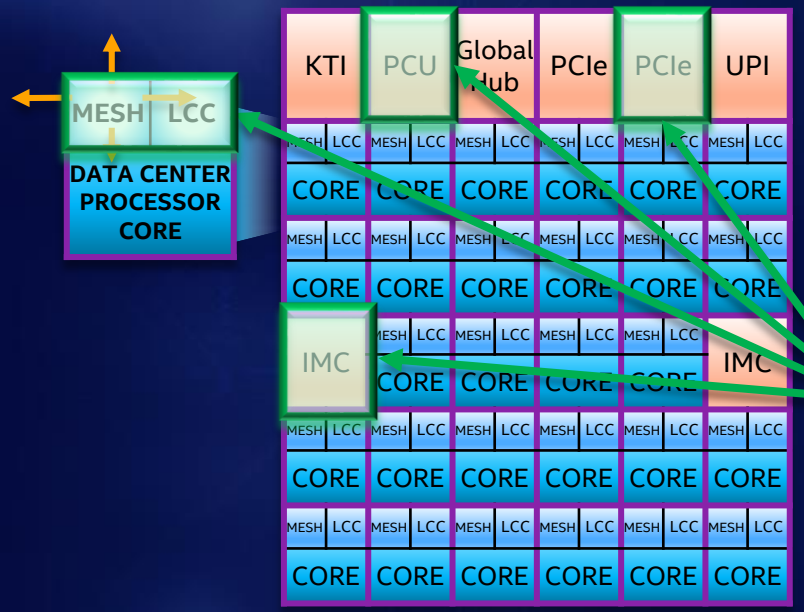
Low performance variability: I/O, memory, compute = consistently lower latency
Non-blocking frequency transitions

HIGH DATA CENTER EFFICIENCY

Intel® Data Direct I/O Technology, CB-DMA, DPDK, Crypto/Compression with Intel® AVX instructions and Intel® QuickAssist Technology, Erasure Coding, RAID

PROCESSOR ARCHITECTURE DESIGNED FOR HIGH UTILIZATION

UTILIZATION IN THE DATA CENTER "COMPUTER"



CONSOLIDATION

Virtualization: VT performance, VM migration, handling large VM sizes. Profiling: Cache and memory bandwidth monitoring, instruction tracing. Availability: Intel® Run Sure Technology

PERFORMANCE CONSISTENCY

Cache capacity enforcement, memory bandwidth allocation, turbo/throttle isolation, code and data isolation

LOW JITTER

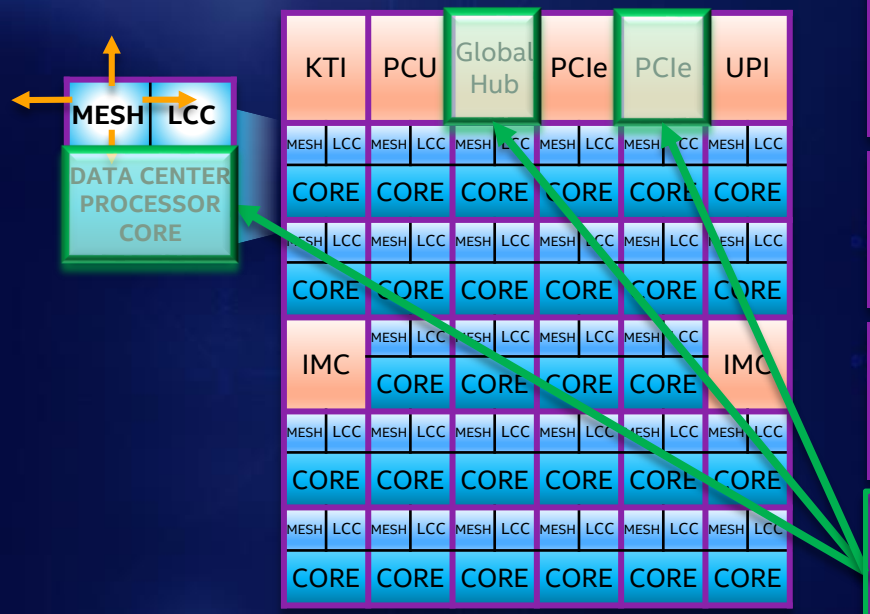
Low performance variability: I/O, memory, compute = consistently lower latency
Non-blocking frequency transitions

HIGH DATA CENTER EFFICIENCY

Intel® Data Direct I/O Technology, CB-DMA, DPDK, Crypto/Compression with Intel® AVX instructions and Intel® QuickAssist Technology, Erasure Coding, RAID

PROCESSOR ARCHITECTURE DESIGNED FOR HIGH UTILIZATION

UTILIZATION IN THE DATA CENTER "COMPUTER"



CONSOLIDATION

Virtualization: VT performance, VM migration, handling large VM sizes. Profiling: Cache and memory bandwidth monitoring, instruction tracing. Availability: Intel® Run Sure Technology

PERFORMANCE CONSISTENCY

Cache capacity enforcement, memory bandwidth allocation, turbo/throttle isolation, code and data isolation

LOW JITTER

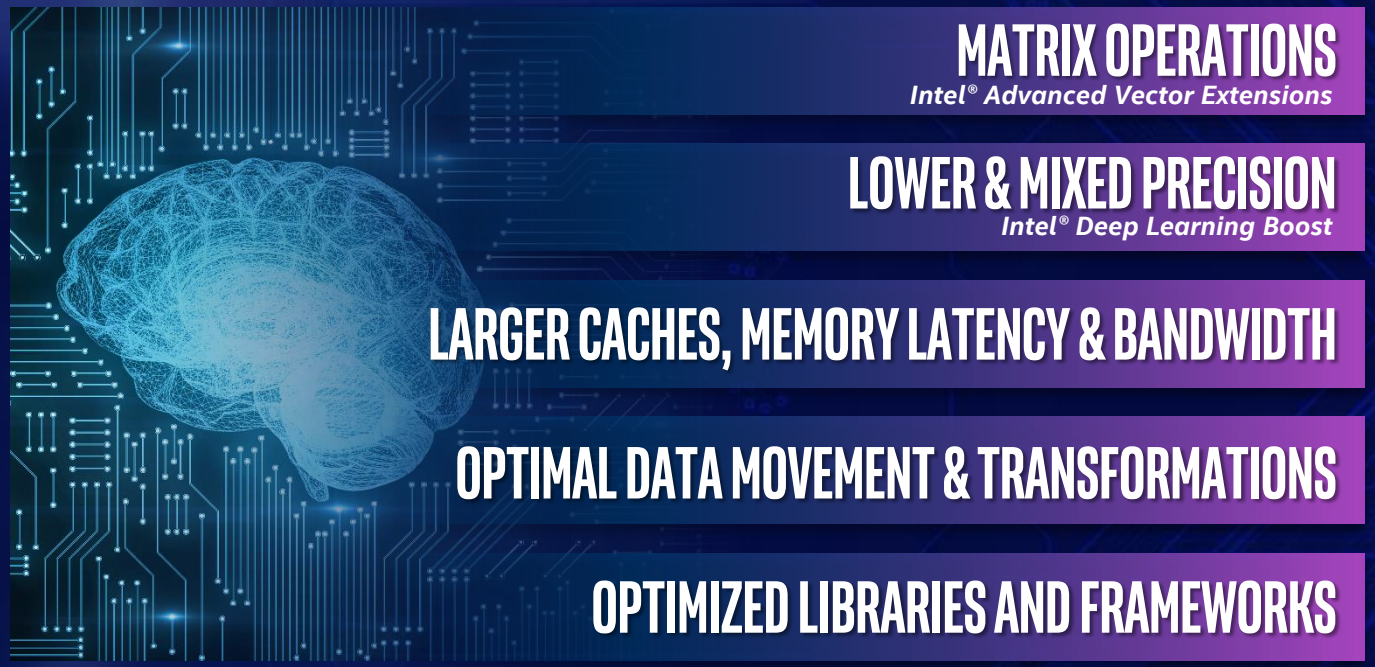
Low performance variability: I/O, memory, compute = consistently lower latency
Non-blocking frequency transitions

HIGH DATA CENTER EFFICIENCY

Intel® Data Direct I/O Technology, CB-DMA, DPDK, Crypto/Compression with Intel® AVX instructions and Intel® QuickAssist Technology, Erasure Coding, RAID

PROCESSOR ARCHITECTURE DESIGNED FOR HIGH UTILIZATION

CPU FOUNDATION FOR ARTIFICIAL INTELLIGENCE



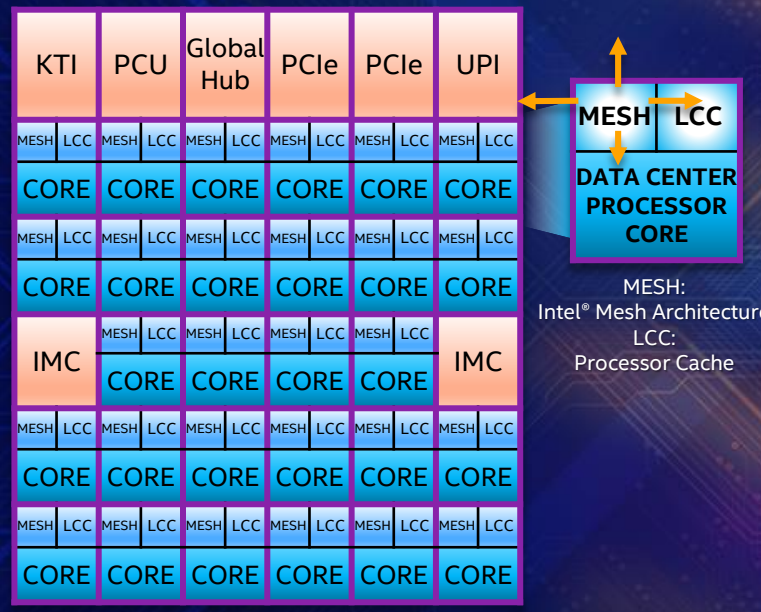
MATRIX OPERATIONS
Intel® Advanced Vector Extensions

LOWER & MIXED PRECISION
Intel® Deep Learning Boost

LARGER CACHES, MEMORY LATENCY & BANDWIDTH

OPTIMAL DATA MOVEMENT & TRANSFORMATIONS

OPTIMIZED LIBRARIES AND FRAMEWORKS

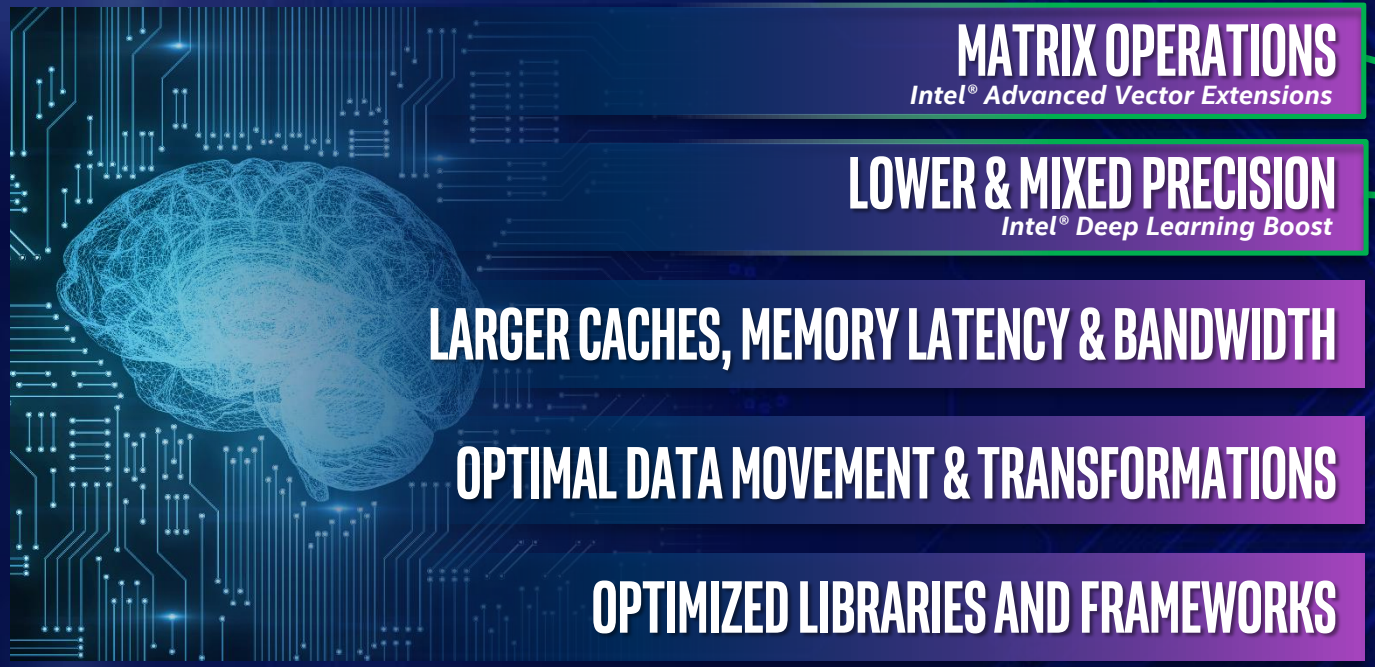


MESH:
Intel® Mesh Architecture
LCC:
Processor Cache

Illustration: Intel® Xeon® Scalable Processor

INTEL® XEON® SCALABLE PROCESSOR: ENABLES INFRASTRUCTURE-WIDE AI READINESS

CPU FOUNDATION FOR ARTIFICIAL INTELLIGENCE



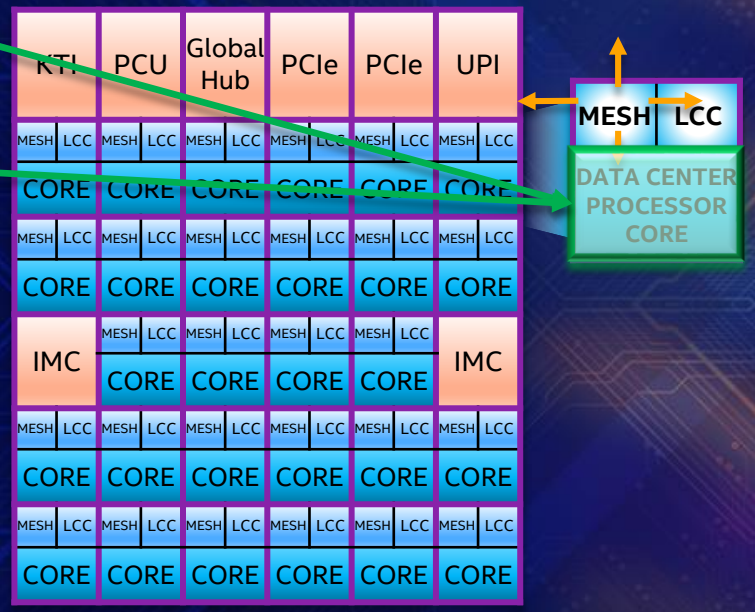
MATRIX OPERATIONS
Intel® Advanced Vector Extensions

LOWER & MIXED PRECISION
Intel® Deep Learning Boost

LARGER CACHES, MEMORY LATENCY & BANDWIDTH

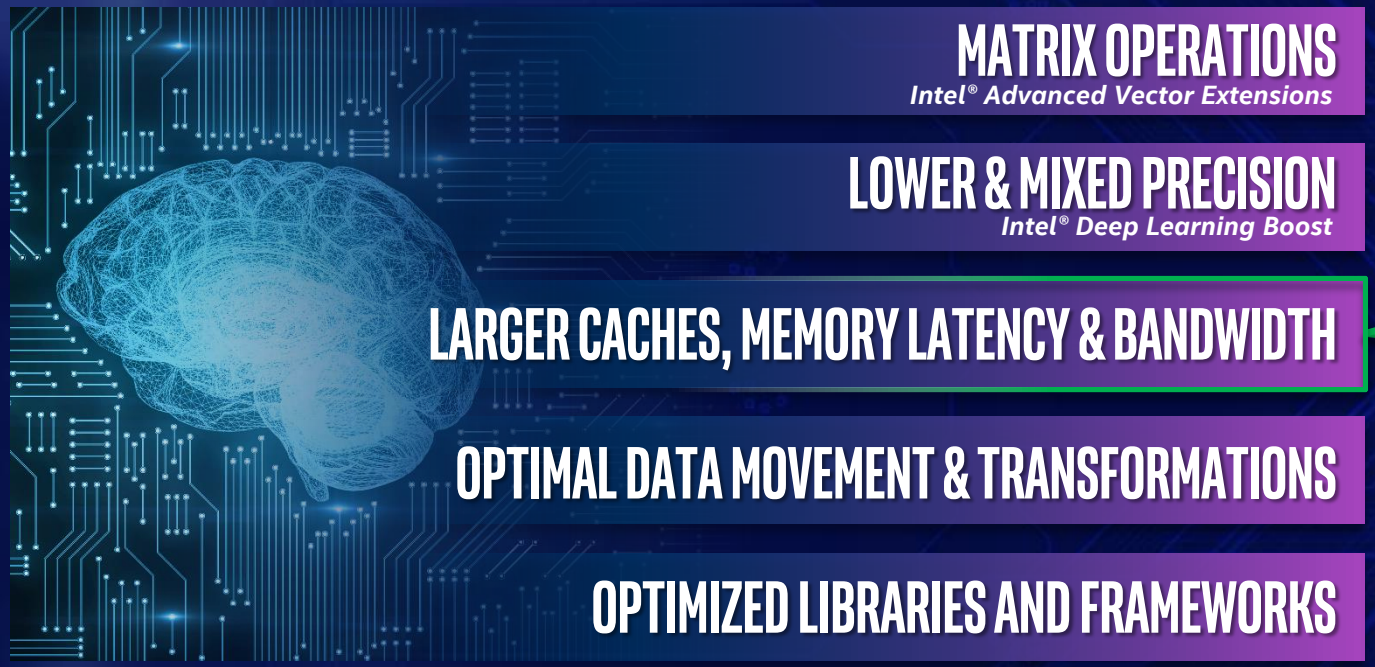
OPTIMAL DATA MOVEMENT & TRANSFORMATIONS

OPTIMIZED LIBRARIES AND FRAMEWORKS



INTEL® XEON® SCALABLE PROCESSOR: ENABLES INFRASTRUCTURE-WIDE AI READINESS

CPU FOUNDATION FOR ARTIFICIAL INTELLIGENCE



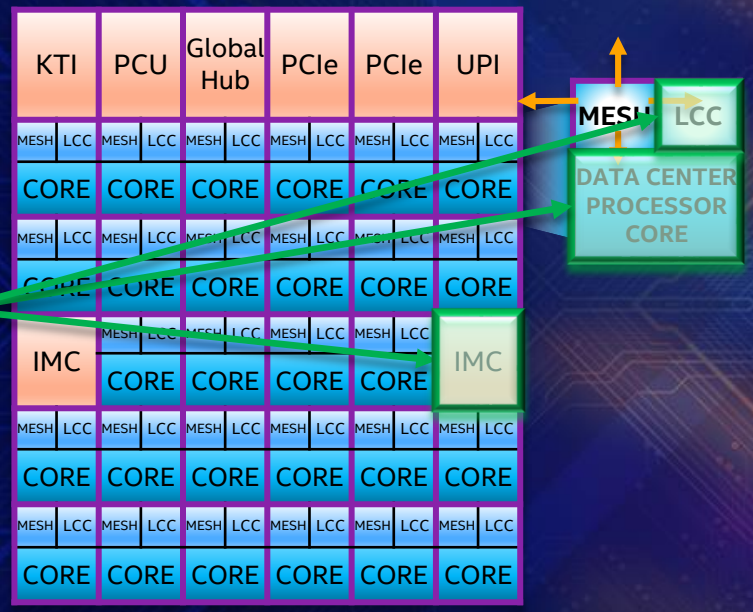
MATRIX OPERATIONS
Intel® Advanced Vector Extensions

LOWER & MIXED PRECISION
Intel® Deep Learning Boost

LARGER CACHES, MEMORY LATENCY & BANDWIDTH

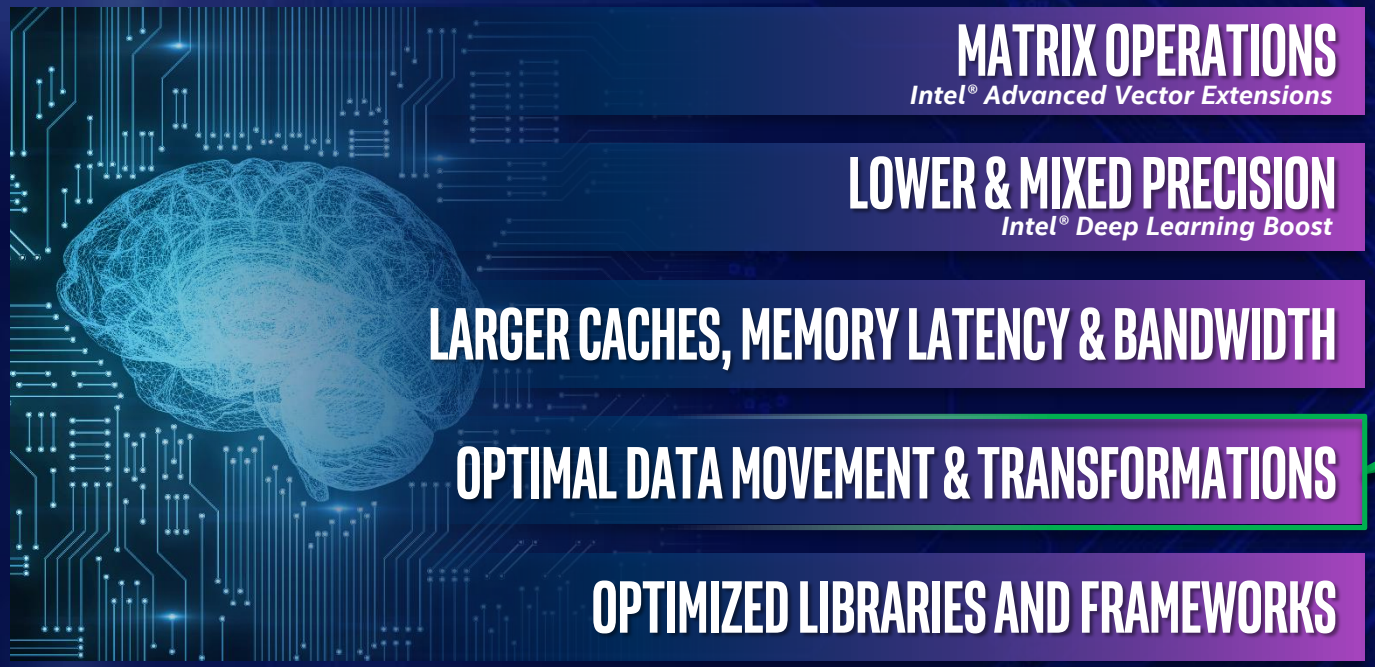
OPTIMAL DATA MOVEMENT & TRANSFORMATIONS

OPTIMIZED LIBRARIES AND FRAMEWORKS



INTEL® XEON® SCALABLE PROCESSOR: ENABLES INFRASTRUCTURE-WIDE AI READINESS

CPU FOUNDATION FOR ARTIFICIAL INTELLIGENCE



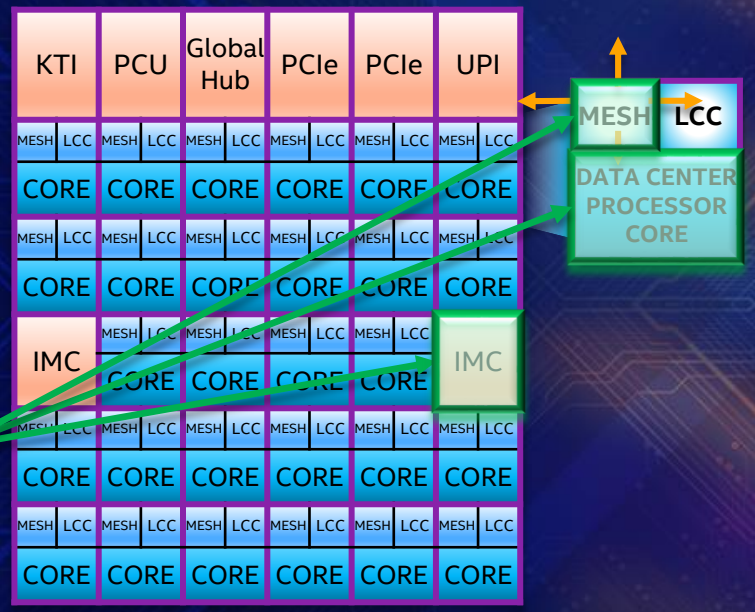
MATRIX OPERATIONS
Intel® Advanced Vector Extensions

LOWER & MIXED PRECISION
Intel® Deep Learning Boost

LARGER CACHES, MEMORY LATENCY & BANDWIDTH

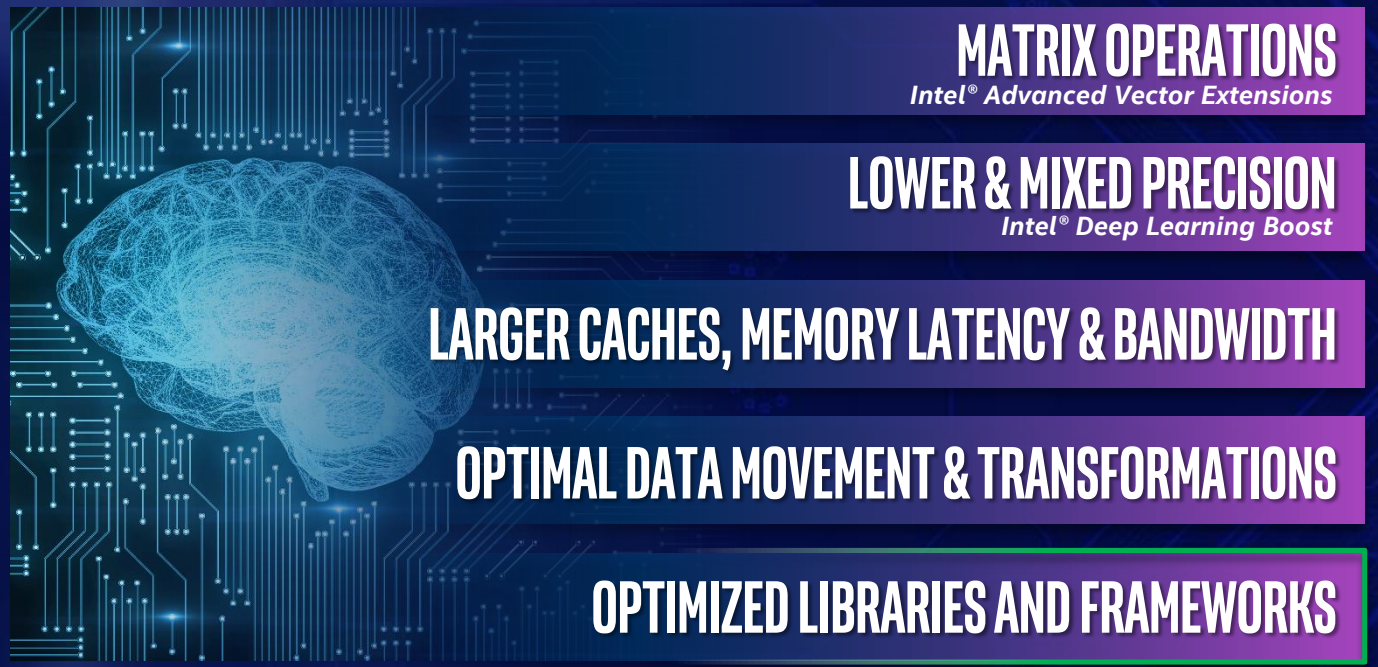
OPTIMAL DATA MOVEMENT & TRANSFORMATIONS

OPTIMIZED LIBRARIES AND FRAMEWORKS



INTEL® XEON® SCALABLE PROCESSOR: ENABLES INFRASTRUCTURE-WIDE AI READINESS

CPU FOUNDATION FOR ARTIFICIAL INTELLIGENCE



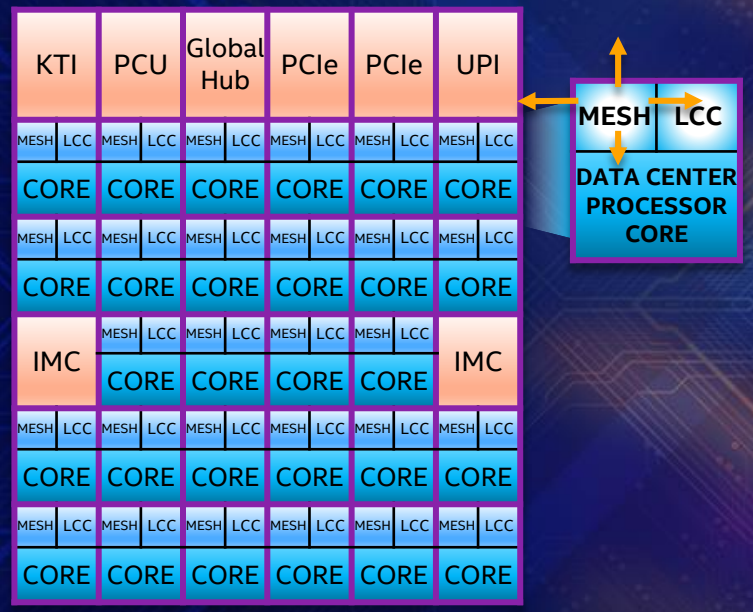
MATRIX OPERATIONS
Intel® Advanced Vector Extensions

LOWER & MIXED PRECISION
Intel® Deep Learning Boost

LARGER CACHES, MEMORY LATENCY & BANDWIDTH

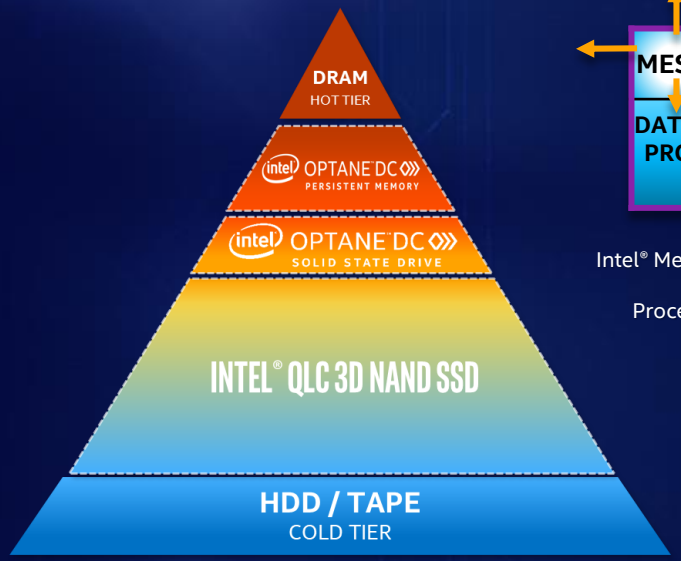
OPTIMAL DATA MOVEMENT & TRANSFORMATIONS

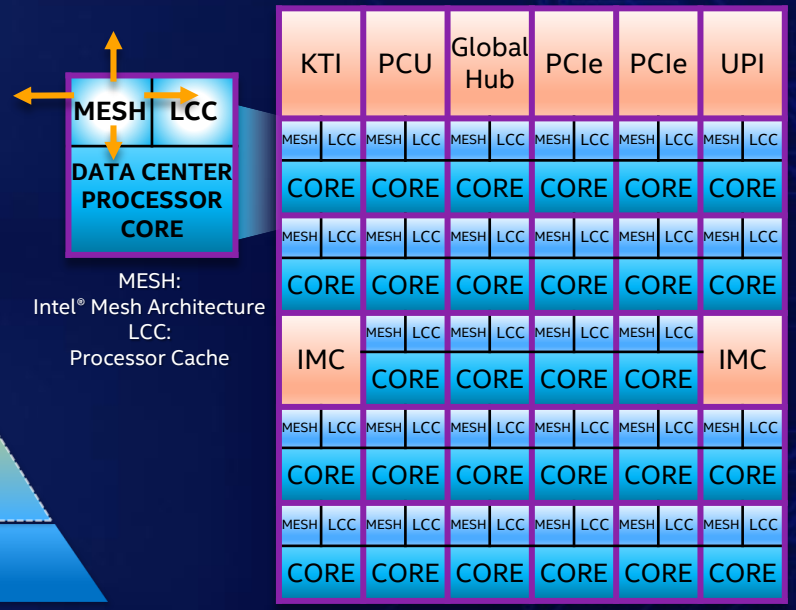
OPTIMIZED LIBRARIES AND FRAMEWORKS



INTEL® XEON® SCALABLE PROCESSOR: ENABLES INFRASTRUCTURE-WIDE AI READINESS

INTEL® OPTANE™ DC MEMORY ARCHITECTURE





MESH:
Intel® Mesh Architecture
LCC:
Processor Cache

Illustration: Intel® Xeon® Scalable Processor

MEMORY SYSTEM ARCHITECTURE

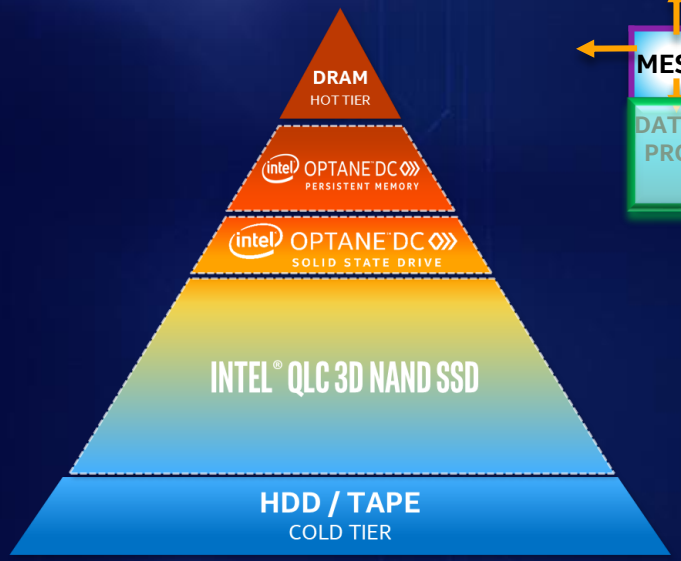
LATENCY AND BANDWIDTH SUPPORT

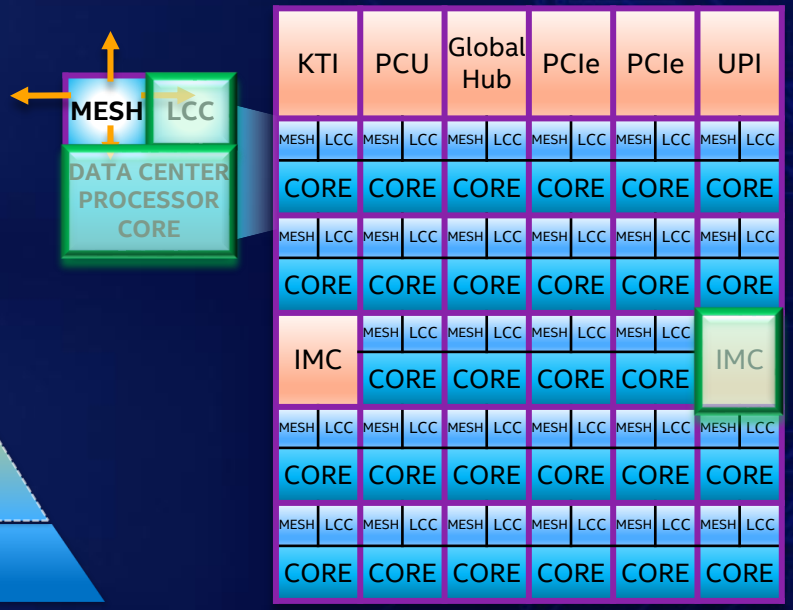
INTERFACES TO SUPPORT PERSISTENCE

QUALITY OF SERVICE

ARCHITECTURE TO TRANSFORM THE MEMORY AND STORAGE HIERARCHY

INTEL® OPTANE™ DC MEMORY ARCHITECTURE





MEMORY SYSTEM ARCHITECTURE

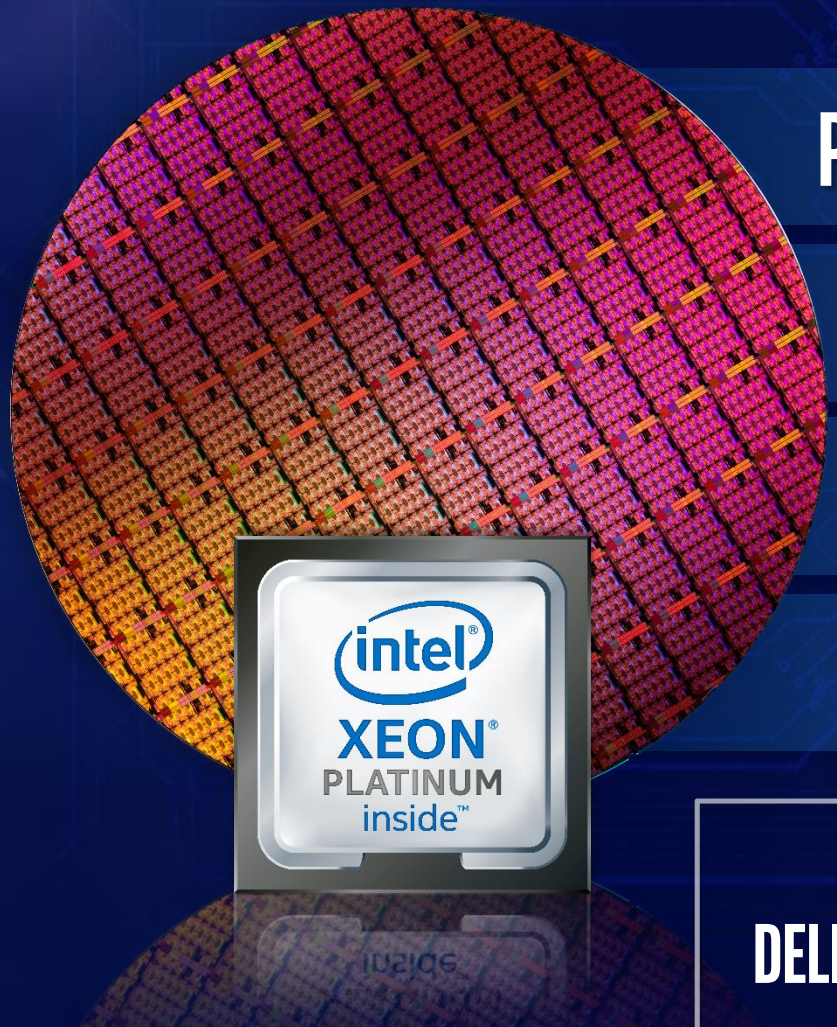
LATENCY AND BANDWIDTH SUPPORT

INTERFACES TO SUPPORT PERSISTENCE

QUALITY OF SERVICE

ARCHITECTURE TO TRANSFORM THE MEMORY AND STORAGE HIERARCHY

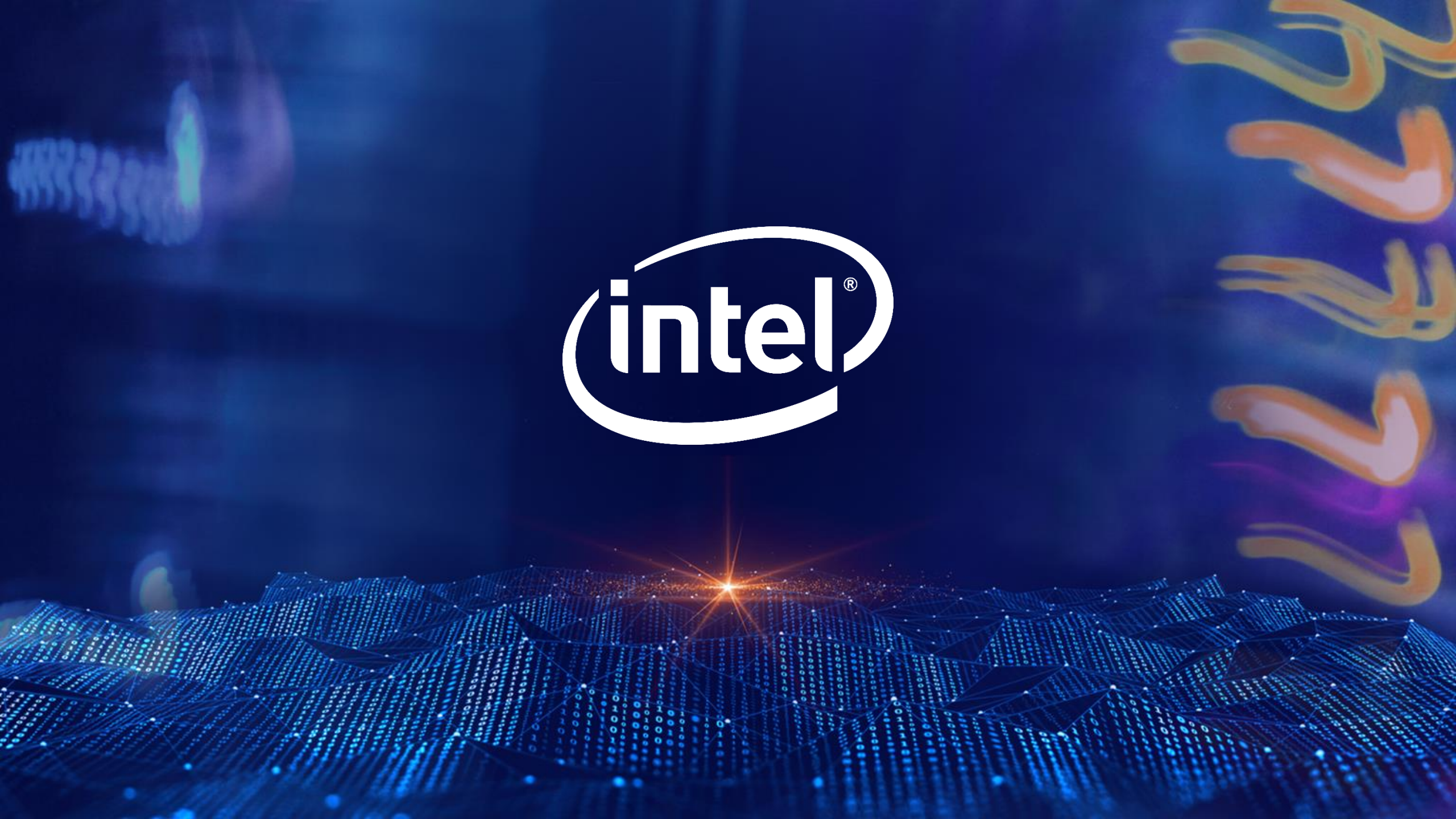
LEADERSHIP CPU ARCHITECTURE FOR A DATA-CENTRIC FUTURE



- PCP AND TPT LEADERSHIP ✓
- HIGH UTILIZATION ✓
- FOUNDATION FOR AI ✓
- MEMORY INNOVATION ✓

DESIGNED. TESTED. TRUSTED.
DELIVERING INNOVATION AND CUSTOMER VALUE


Q&A
DATA-CENTRIC
INNOVATION SUMMIT
#IntelDCISummit
intel


intel

DISCLOSURES

Statements in this presentation that refer to business outlook, future plans and expectations are forward-looking statements that involve a number of risks and uncertainties. Words such as "anticipates," "expects," "intends," "goals," "plans," "believes," "seeks," "estimates," "continues," "may," "will," “would,” "should," “could,” and variations of such words and similar expressions are intended to identify such forward-looking statements. Statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Such statements are based on management's current expectations, unless an earlier date is indicated, and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in these forward-looking statements. Important factors that could cause actual results to differ materially from the company's expectations are set forth in Intel's earnings release dated July 26, 2018, which is included as an exhibit to Intel’s Form 8-K furnished to the SEC on such date. Additional information regarding these and other factors that could affect Intel's results is included in Intel's SEC filings, including the company's most recent reports on Forms 10-K and 10-Q. Copies of Intel's Form 10-K, 10-Q and 8-K reports may be obtained by visiting our Investor Relations website at www.intc.com or the SEC's website at www.sec.gov.

All information in this presentation reflects management’s views as of the date of this presentation, unless an earlier date is indicated. Intel does not undertake, and expressly disclaims any duty, to update any statement made in this presentation, whether as a result of new information, new developments or otherwise, except to the extent that disclosure may be required by law.