Sapphire Rapids

Arijit Biswas
Intel Senior Principal Engineer

intel®

HOT CHIPS

Sapphire Rapids

Next-Gen Intel Xeon Scalable Processor

New Standard for
Data Center Architecture

Designed for Microservices
& AI Workloads

Pioneering Advanced Memory
& IO Transitions

Node Performance

Data Center Performance

Node Performance

Scalar Performance	Data Parallel Performance	Cache & Memory Sub-System Arch	Intra/Inter Socket Scaling
New Performance Core Microarchitecture	Multiple Integrated Acceleration Engines	Larger Private & Shared Caches	Modular SoC /w Modular Die Fabric
		DDR 5	Wider & Faster UPI
	Increased Core Counts	Next Gen Optane Support	Embedded Silicon Bridge (EMIB)
		PCIe 5.0	

		Next Gen Quality of Service Capabilities	
Fast VM Migration	Low Jitter Architecture	Broad WL/Usage Support and Optimizations	
Better Telemetry	Consistent Caching & Mem Latency	Next Gen Optane Support	Integrated WL Accelerators
IO Virtualization	Inter-Processor Interrupt Virt.	CXL 1.1	Improved Security & RAS
Consolidation & Orchestration	**Performance Consistency**	**Elasticity & Efficient Data Center Utilization**	**Infrastructure & Framework Overhead**

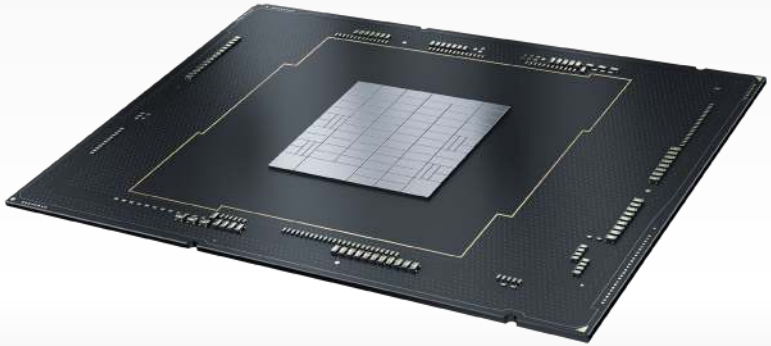
Data Center Performance

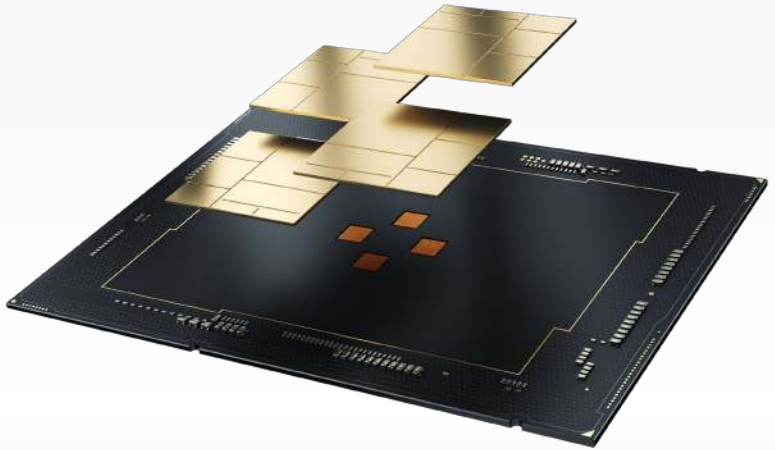
Ice Lake

Single Monolithic Die

Sapphire Rapids

Multi-Tile Design for Increased Scalability





Delivers a scalable, balanced architecture leveraging existing software paradigms for monolithic CPUs via a modular architecture

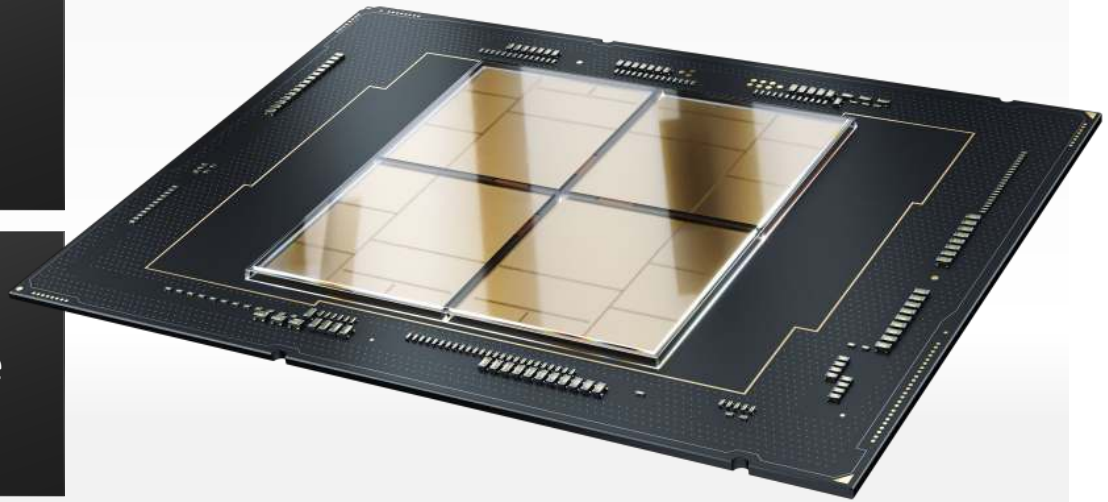
Sapphire Rapids

Multiple Tiles, Single CPU

Every thread has full access to all resources on all tiles

Cache, Memory, IO…

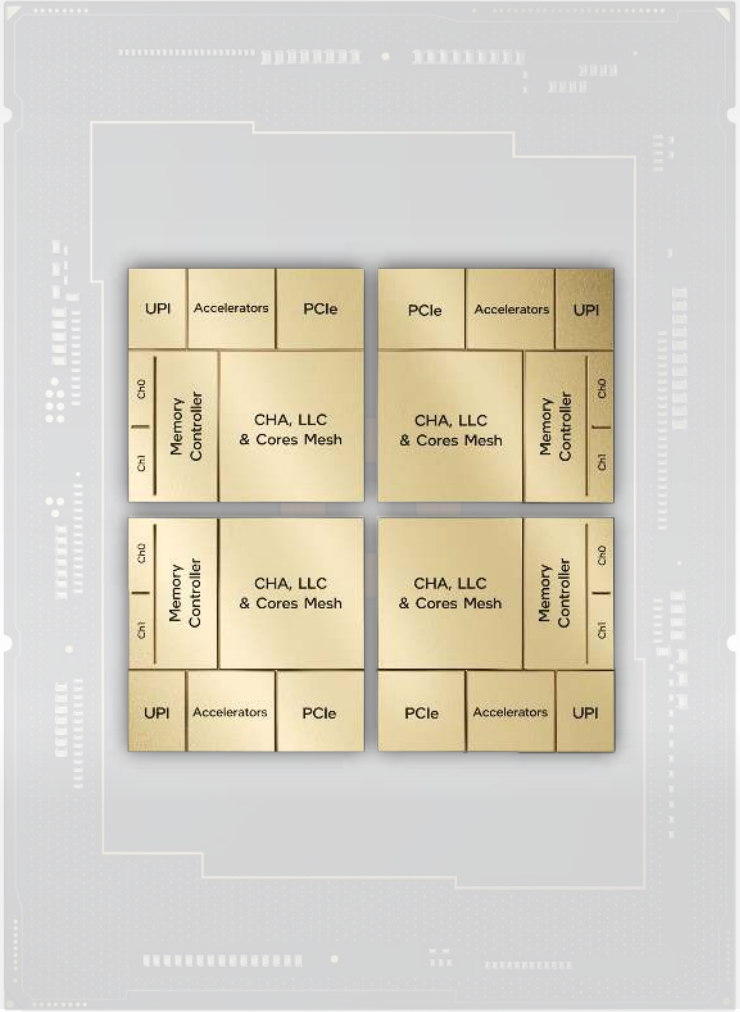
Provides consistent low latency & high cross-section BW across the entire SoC



Sapphire Rapids

Key Building Blocks

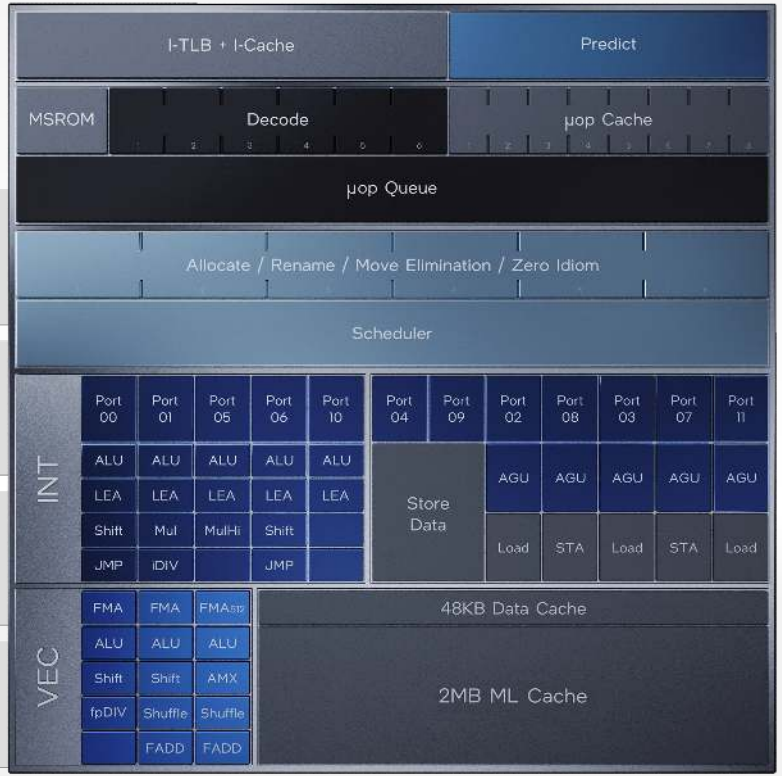
Compute IP Seamless Integration of	Cores	Acceleration Engines	
I/O IP	CXL1.1	PCIe Gen 5	UPI 2.0
Memory IP	DDR 5	Optane	HBM



Performance Core

Built for Data Center

- Major microarchitecture and IPC improvement
- Improved support for large code/data footprint
- Consistent performance for multi-tenant usages
- Autonomous/Fast PM for high freq @ low jitter



Performance Core

Architecture Improvements for DC Workloads & Usages

AI	**Intel® Advanced Matrix Extensions - AMX** Tiled matrix operations for inference & training acceleration
Attached Device	**Accelerator interfacing Architecture - AiA** Efficient dispatch, signaling & synchronization from user level
HFNI	**Half- Precision Float New Instructions** Support for FP16 - higher throughput lower precision
Cache Management	**CLDEMOTE** Proactive placement of cache contents

Sapphire Rapids

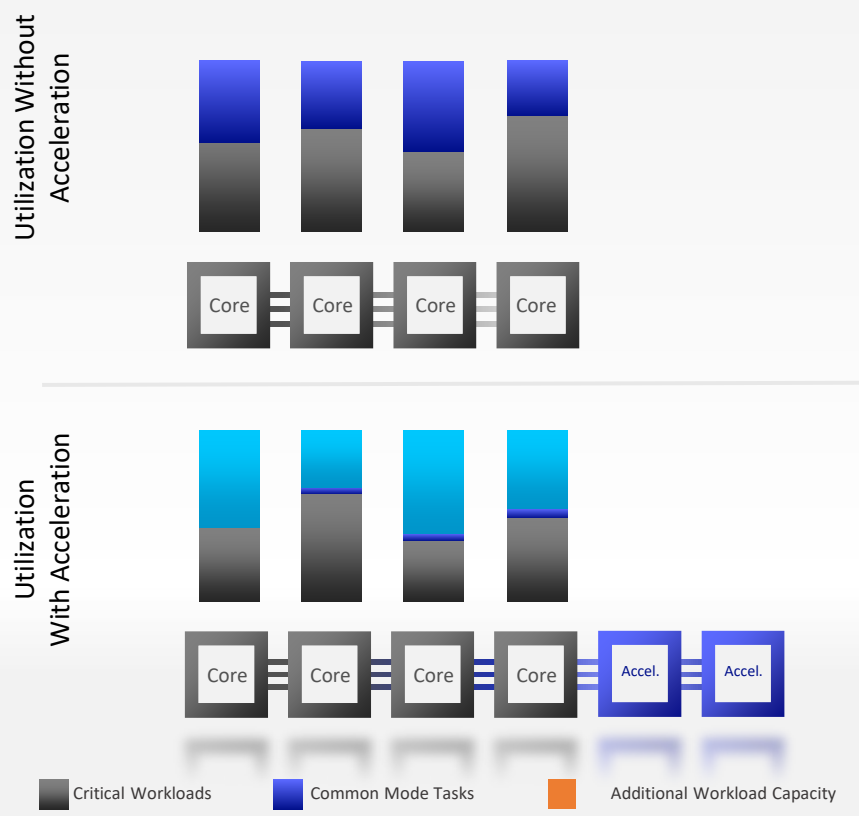
Acceleration Engines

Increasing effectiveness of cores,
by enabling offload of common mode tasks via
seamlessly integrated acceleration engines

Native Dispatch, Signaling & Synchronization from User Space
Accelerator interfacing Architecture

Coherent, Shared Memory Space
Between Cores & Acceleration Engines

Concurrently shareable
Processes, containers and VMs



Intel® Data Streaming

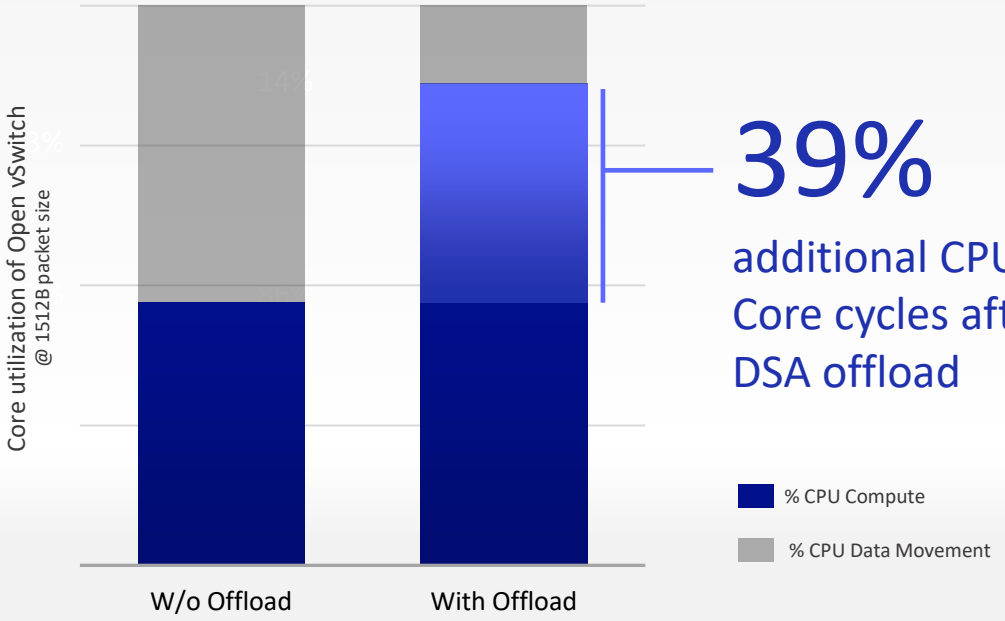
Acceleration Engine

Optimizing streaming data movement and transformation operations

up to
4 Instances per Socket

Low Latency Invocation

No Memory Pinning Overhead



Results have been estimated or simulated based on testing on pre-production hardware and software.
For workloads and configurations visit www.intel.com/ArchDay21claims . Results may vary

Intel® Quick Assist Technology

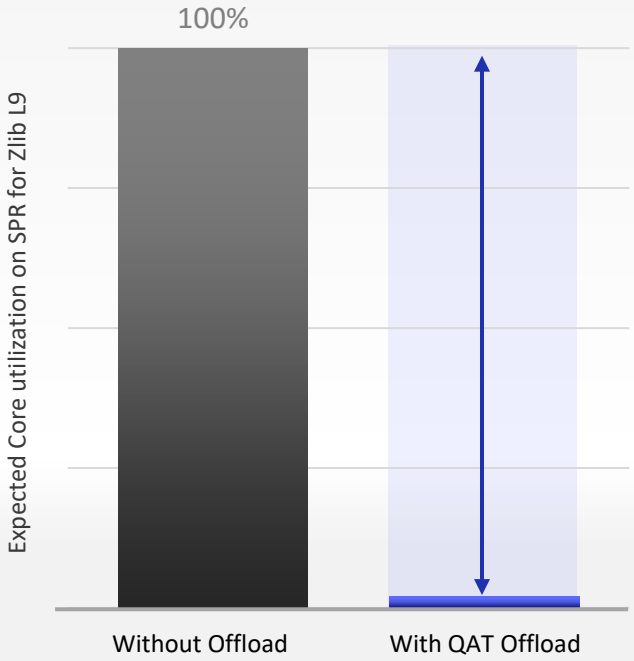
Acceleration Engine

Accelerating Cryptography and Data De/Compression

up to
400Gb/s Symmetric Crypto

up to
160Gb/s Compression +
160Gb/s De-compression

Fused Operations



98%

additional
workload capacity
after QAT offload

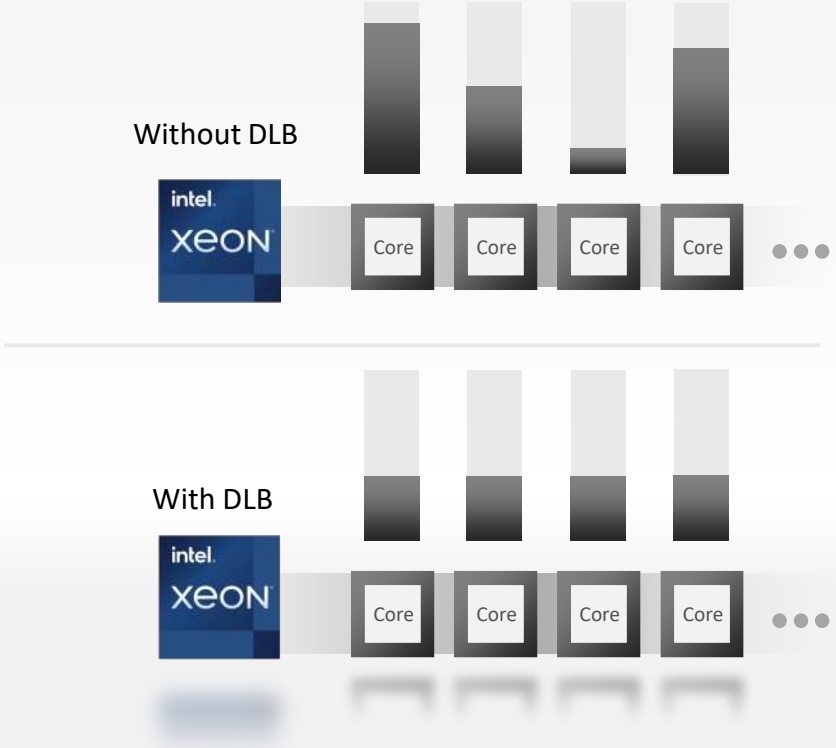
Results have been estimated or simulated. Sapphire Rapids estimation based on architecture models and baseline testing with Ice Lake and Intel QAT. For workloads and configurations visit www.intel.com/ArchDay21claims . Results may vary.

Intel® Dynamic Load Balancer

Acceleration Engine

Efficient Load Balancing across CPU Cores

- **400M** Load Balancing Decisions per Second
- Offloads Software Queue Management
- Dynamic, flow aware load balancing & reordering
- Priority Queuing (up to 8 levels)
- Dynamic, power aware sizing of applications



Sapphire Rapids

I/O Advancements

Introducing Compute eXpress Link (CXL) 1.1

Accelerator and memory expansion in datacenter

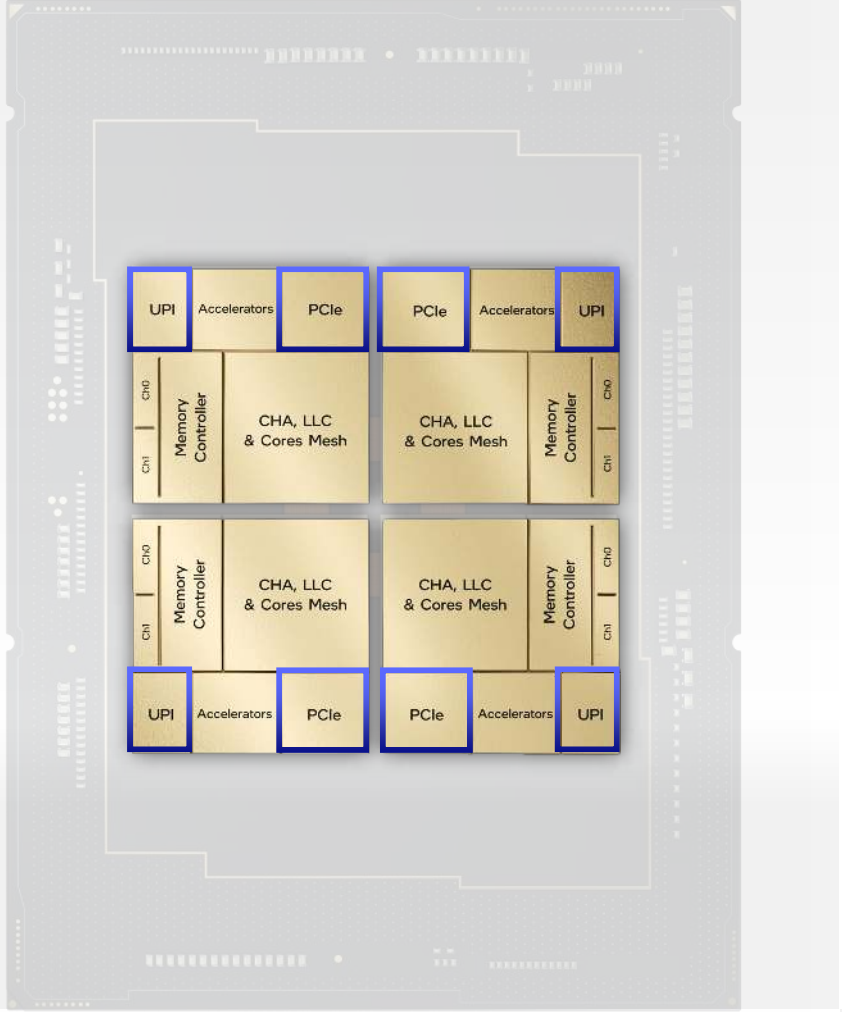
Expanded device performance via PCIe 5.0 & connectivity

Improved DDIO & QoS capabilities

Improved Multi-Socket scaling via Intel Ultra Path Interconnect (UPI) 2.0

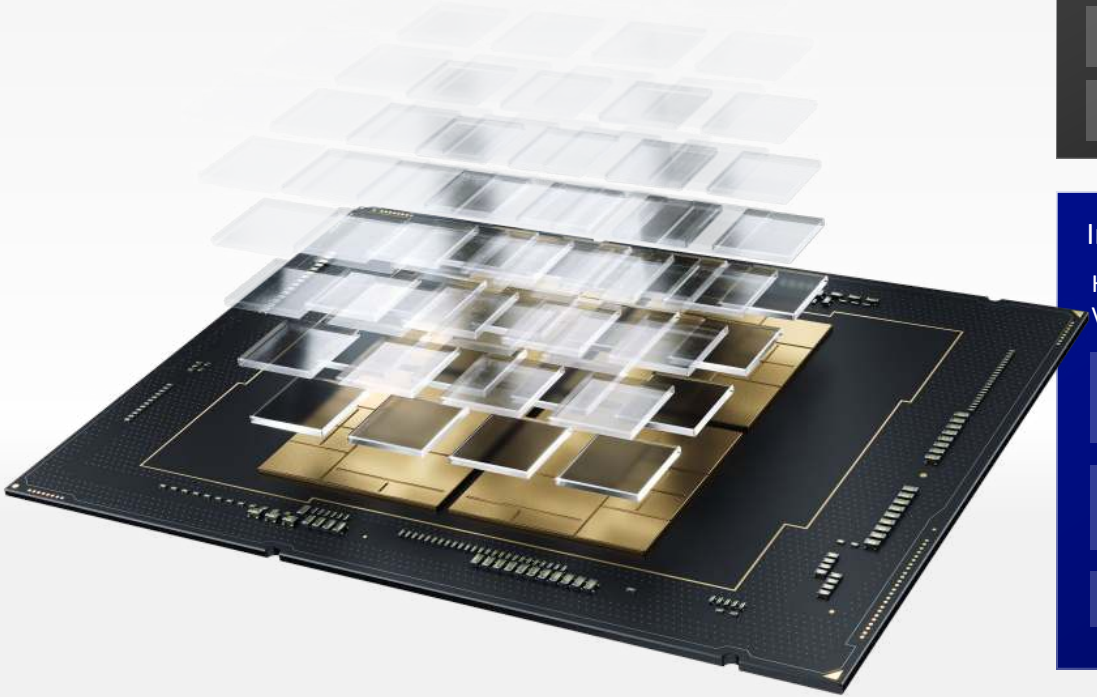
Up to 4 x24 UPI links operating @ 16 GT/s

New 8S-4UPI performance optimized topology



Sapphire Rapids

IO - Virtualization



Intel® Shared Virtual Memory (SVM)

Enabling devices and IA cores to access shared data in CPU virtual address space

- Consistent across host app. and offloaded tasks
- Avoids memory pinning and copying overheads
- Integrated & discrete, bare-metal & VM instances

Intel® Scalable IO Virtualization (S-IOV)

Hardware acceleration for comms between VMs/containers and PCIe devices

- Scalable sharing and direct access to accelerators across 1000s of VMs/containers
- Higher Perf than SW only device scaling, More scalable than SR-IOV
- Supports integrated & discrete devices

Sapphire Rapids

Memory and Last Level Cache

Increased Shared Last Level Cache (LLC)
Up to >100 MB LLC shared across ALL cores

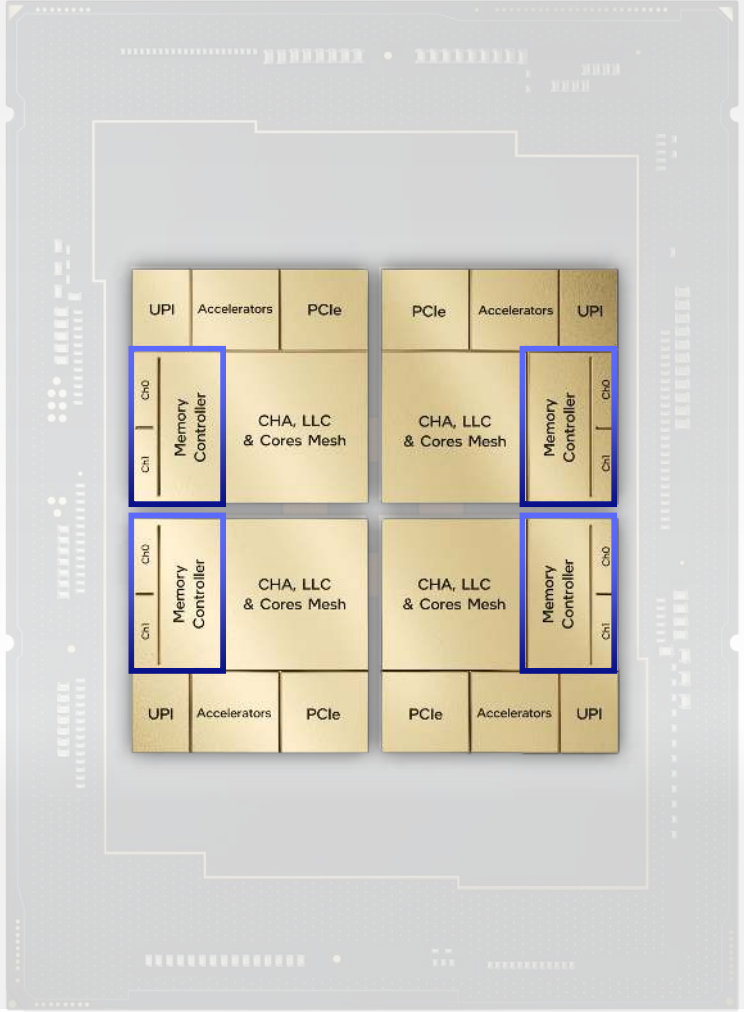
Increased bandwidth, security & reliability
via DDR 5 Memory

4 memory controllers supporting 8 channels

Integrated memory encryption engine

Improved RAS

Intel Optane™ Persistent Memory 300 Series



Sapphire Rapids

High Bandwidth Memory

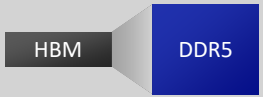
Significantly Higher Memory Bandwidth
vs. baseline Xeon-SP with 8 channels of DDR 5

Increased capacity and Bandwidth
some usages can eliminate need for DDR entirely

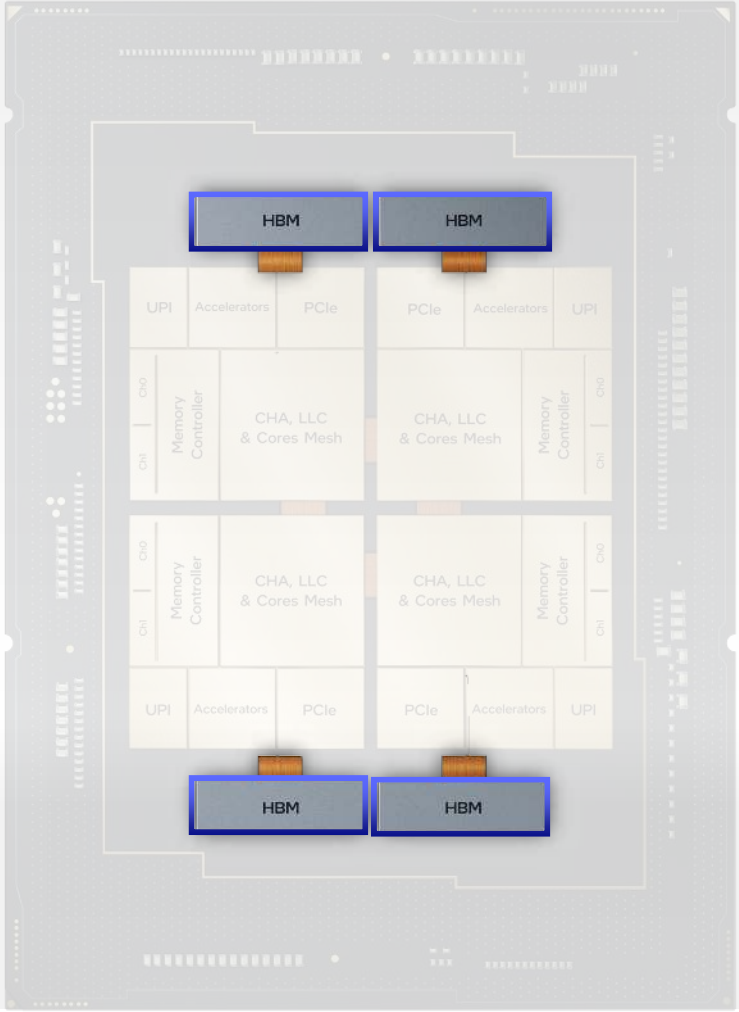
2 Modes

HBM Flat Mode	HBM Caching Mode
Flat Mem Regions w/ HBM & DRAM	DRAM backed cache

HBM Flat Mode enables flat memory regions with HBM & DRAM



HBM Caching Mode allows HBM to act as DRAM backed cache



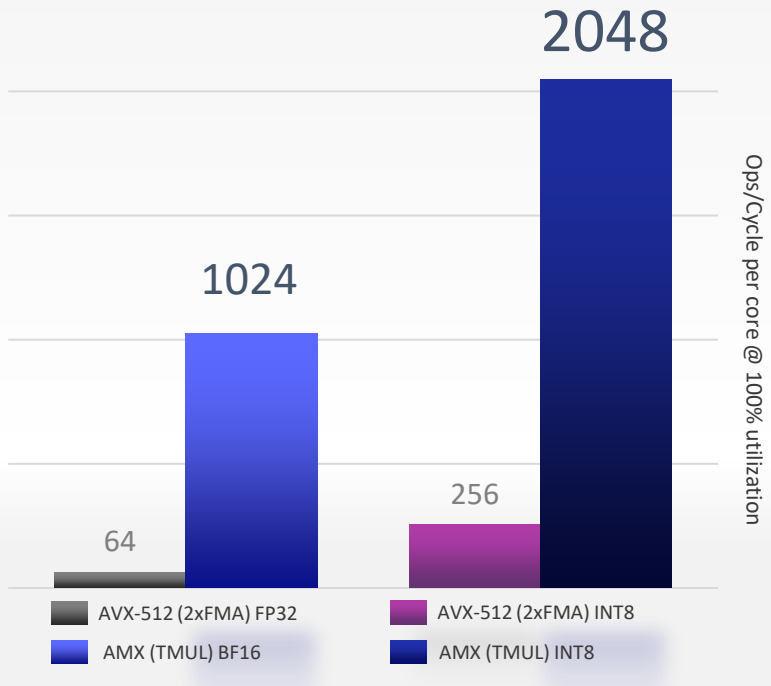
Sapphire Rapids - Architected for AI

AI has become ubiquitous across usages – AI performance required in all tiers of computing

Goal

Enable efficient usage of AI across all services deployed on elastic general-purpose tier by delivering many times more AI performance and lower CPU utilization

For Deep Learning Datatypes	▪ int8 with int32 accumulation ▪ Bfloat16 with IEEE SP accumulation
Acceleration at the ISA Level	▪ Full Intel Arch. programmability ▪ Low Latency
Available and integrated with industry-relevant frameworks & libraries	



Results have been simulated. For workloads and configurations visit www.intel.com/ArchDay21claims . Results may vary

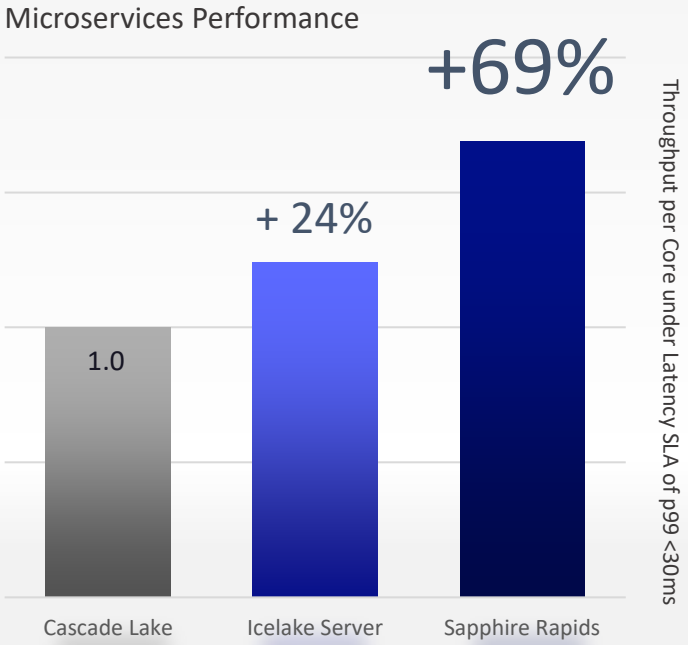
Sapphire Rapids - Built for elastic computing models - microservices

>80% of new cloud-native and SaaS applications are expected to be built as microservices

Goal

Enable higher throughput while meeting latency requirements and reducing infrastructure overhead for execution, monitoring and orchestration thousands of microservices

Improved Performance and Quality of Service	Runtime Languages - lower latency for Runtime Languages AiA ISA's - efficient worker threads, signaling and synch.
Reduced Infrastructure Overhead	Kubernetes – enhanced for scaling, placement and policies Advanced Telemetry - easier analysis & optimization
Better Distributed Communication	Improved latency of Remote procedure calls and service-mesh QAT, DSA etc.- optimized networking and data movement



Results have been simulated. For workloads and configurations visit www.intel.com/ArchDay21claims . Results may vary

Sapphire Rapids

Biggest Leap in Data Center Capabilities in over a Decade

New Standard in Data Center Architecture

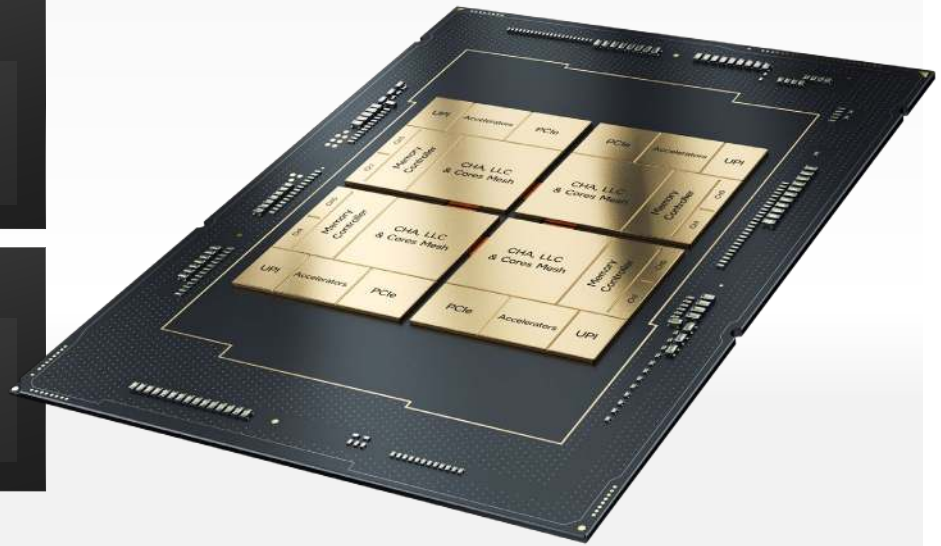
- Multi Tile SoC for Scalability
- Physically Tiled, Logically Monolithic
- General Purpose & Dedicated Acceleration Engines

Designed for Microservices and AI Workloads

- Performance Core Architecture
- Workload Specialized Acceleration

Pioneering Advanced Memory & IO Transitions

- DDR 5 & HBM
- PCIe 5.0
- Enhanced Virtualization Capabilities




intel