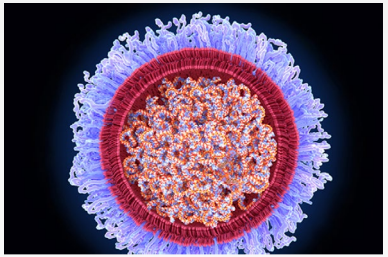
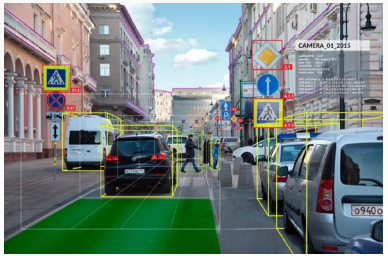
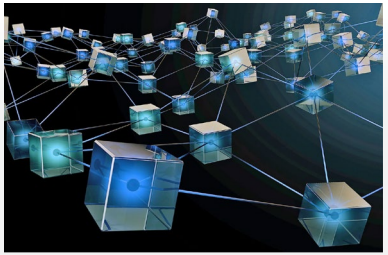
This event includes forward-looking statements about future products and other topics, which are based on our current expectations and subject to risks and uncertainties. Please refer to the press release for this event and our SEC filings at intc.com for more information on the risk factors that could cause actual results to differ materially.

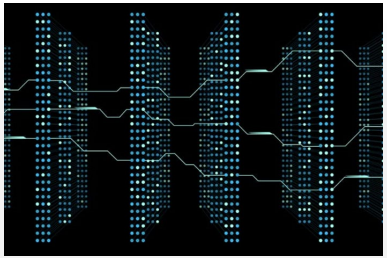
intel

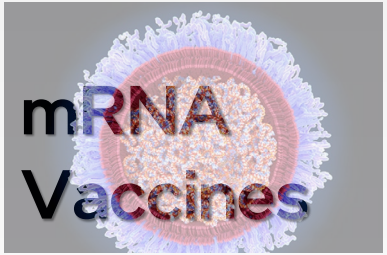
Architecture Day 2021

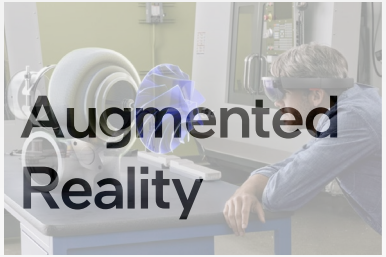



CAMERA_01_2015



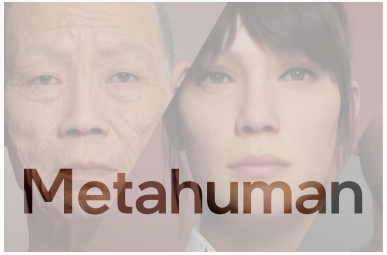



mRNA
Vaccines


Augmented
Reality


Github
Co-Pilot


Computer
Vision


Metahuman


Crypto
Currencies

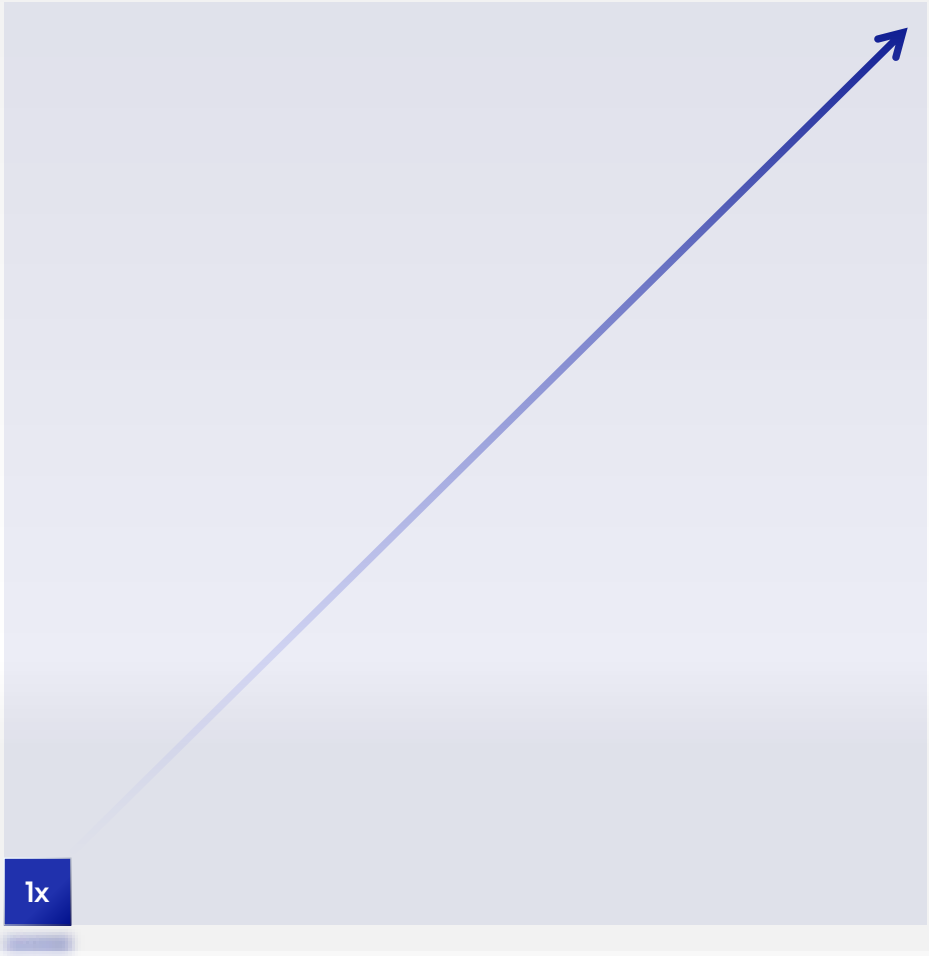

Decentralized
Finance


Space
2.0


Artificial
Intelligence

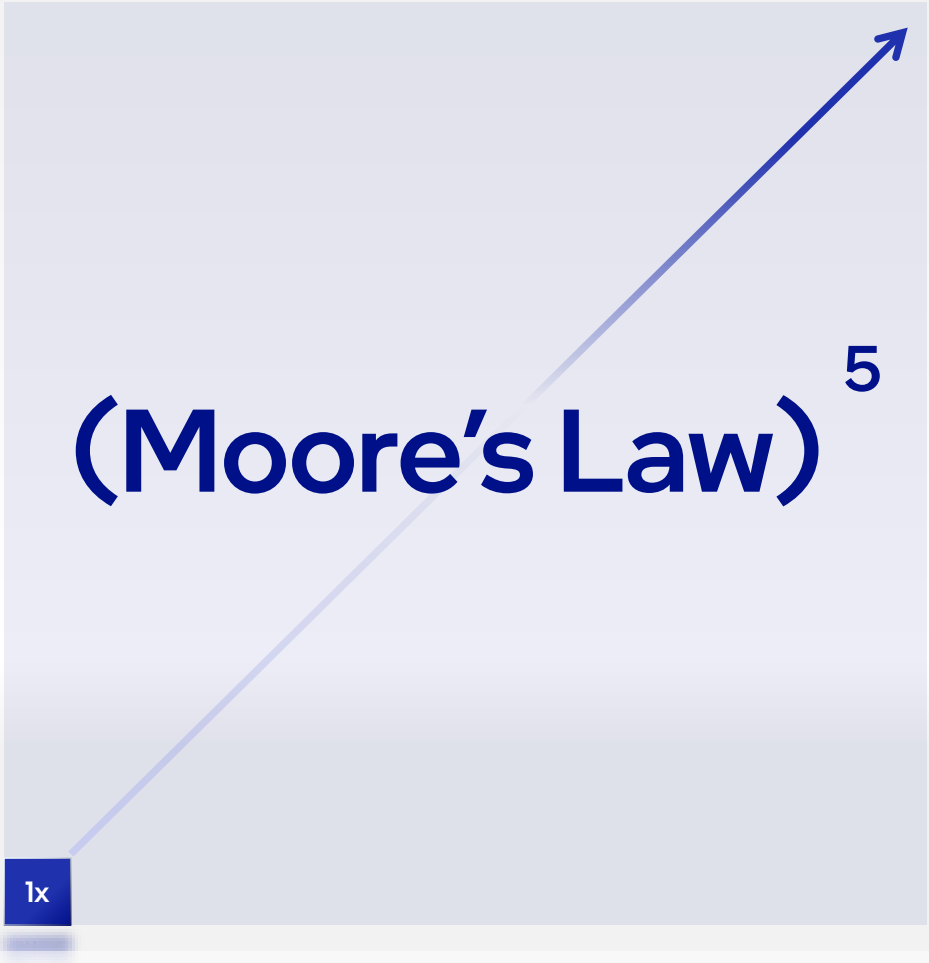
1000x
by 2025

1000x
by 2025

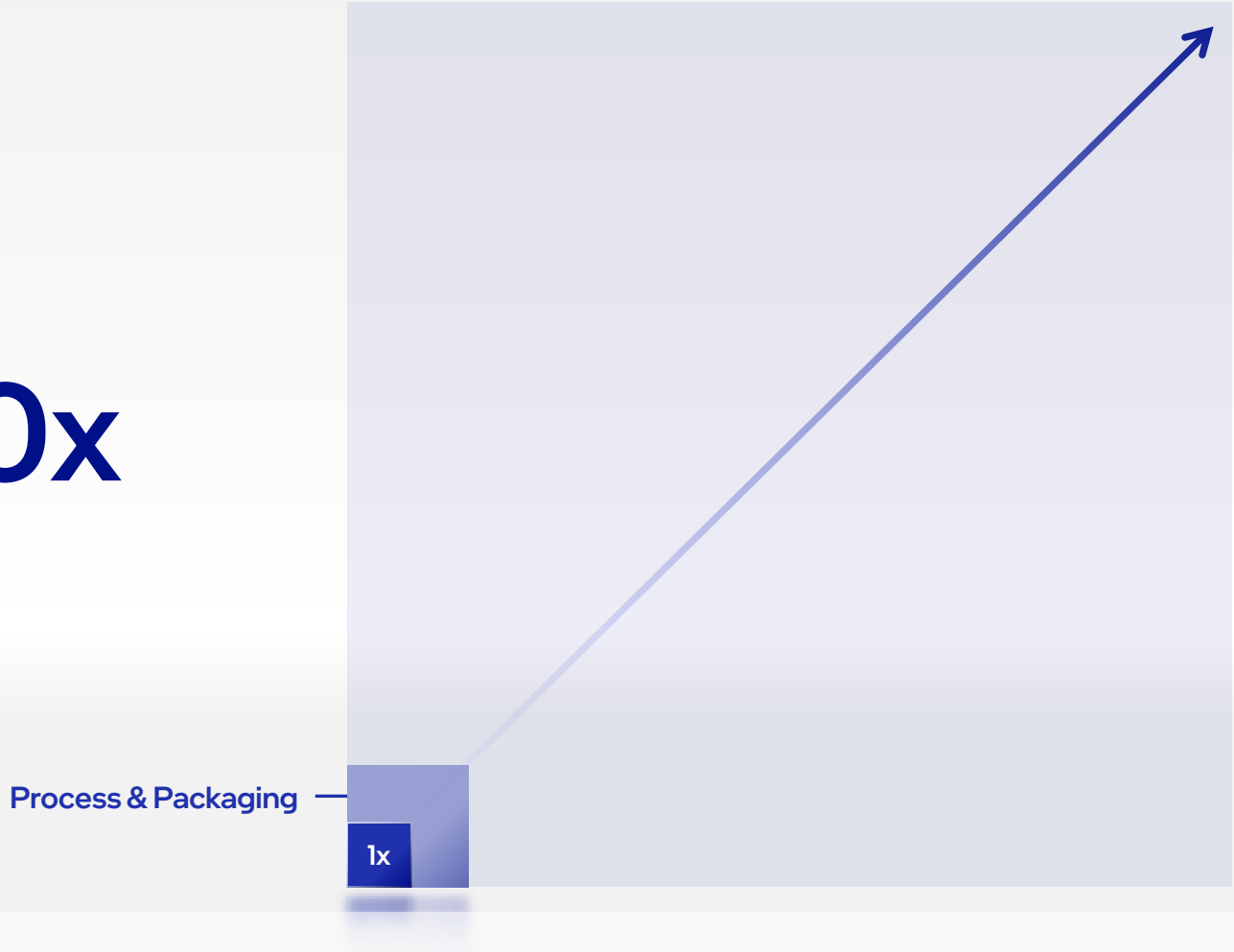


1000x
by 2025

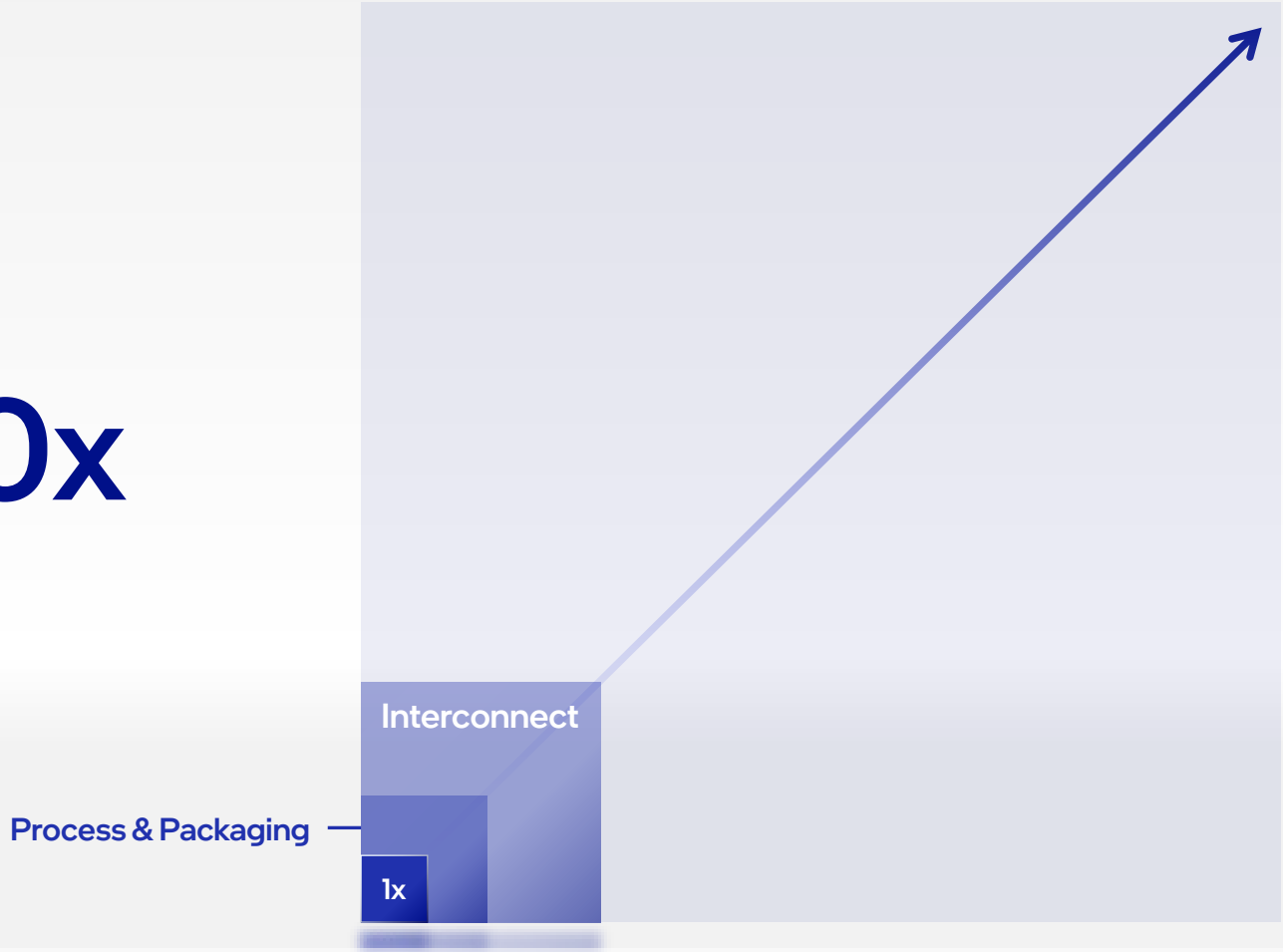
=



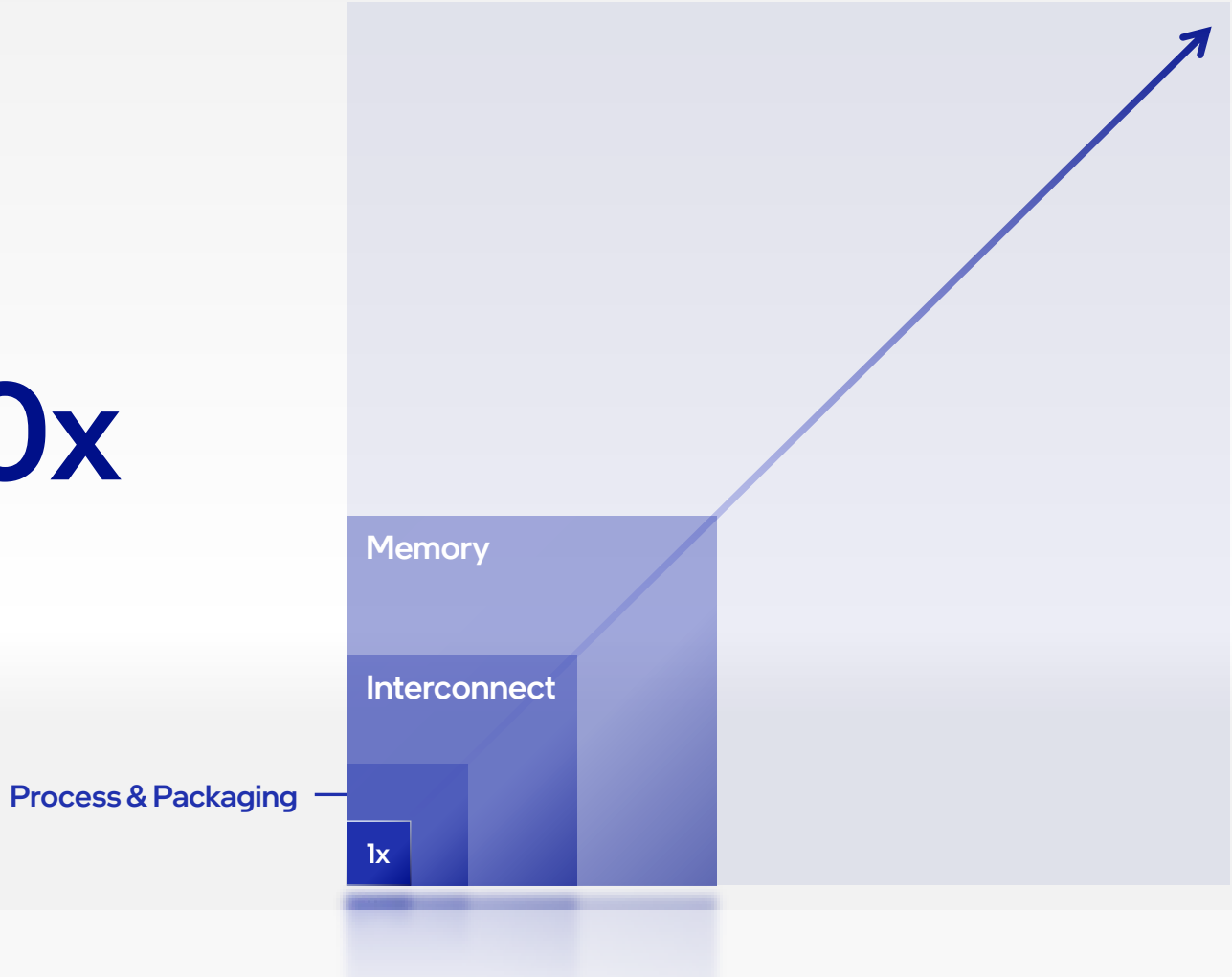
1000x
by 2025

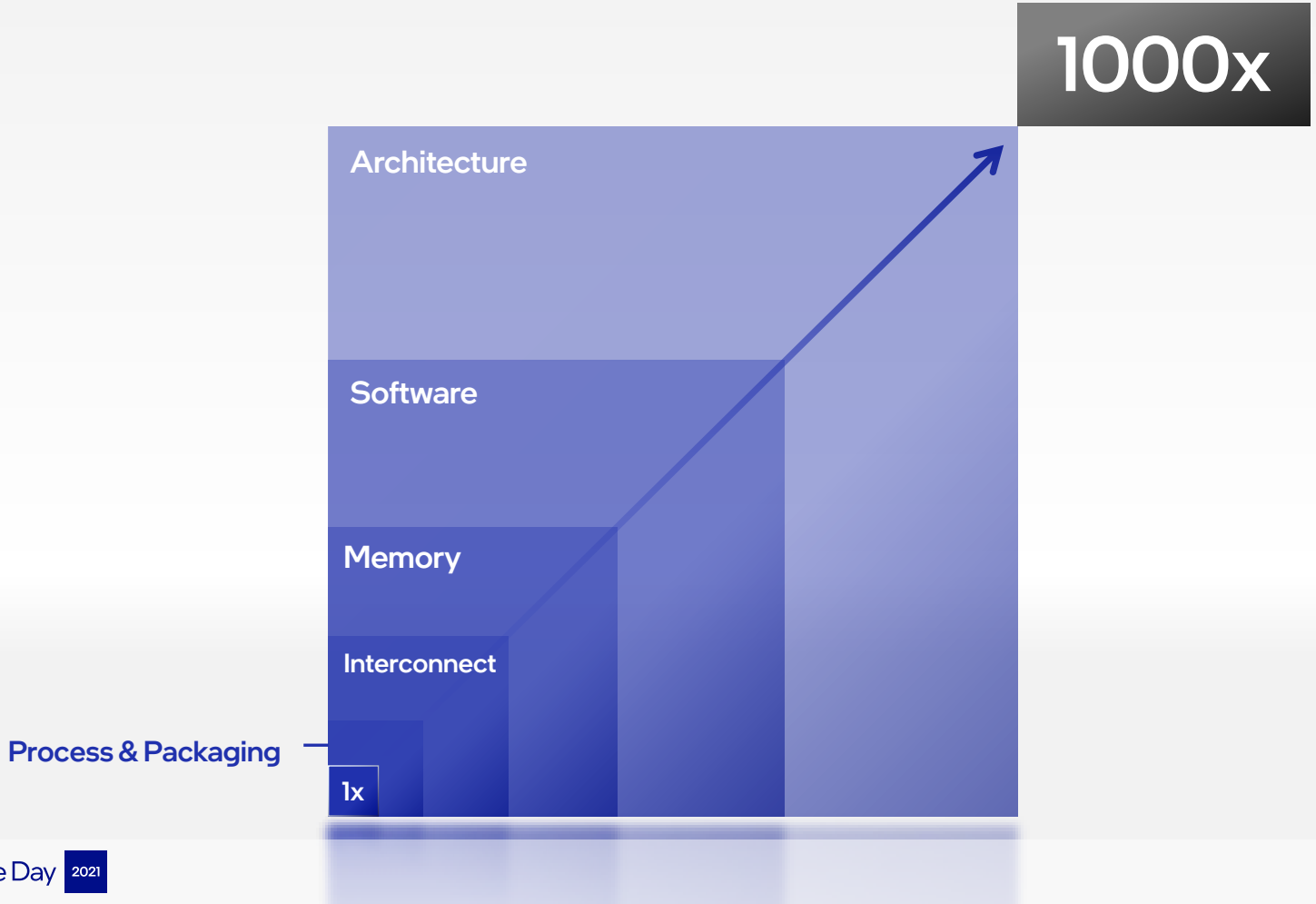


1000x
by 2025



1000x
by 2025




1000x
Architecture
Software
Memory
Interconnect
Process & Packaging
1x

Scalar

Scalar

Vector

Scalar

Vector

Matrix

Scalar

Vector

Matrix

Spatial

Scalar Vector Matrix Spatial

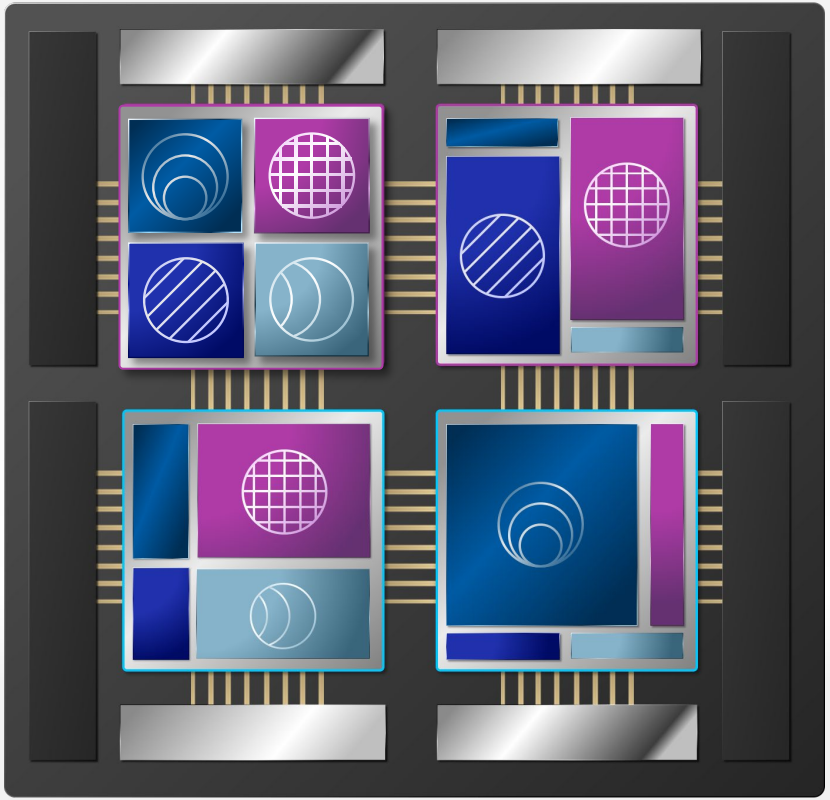
Hybrid Computing Architectures

Hybrid Computing Architectures

- Process
- Packaging
- Caches
- Memory
- Interconnect

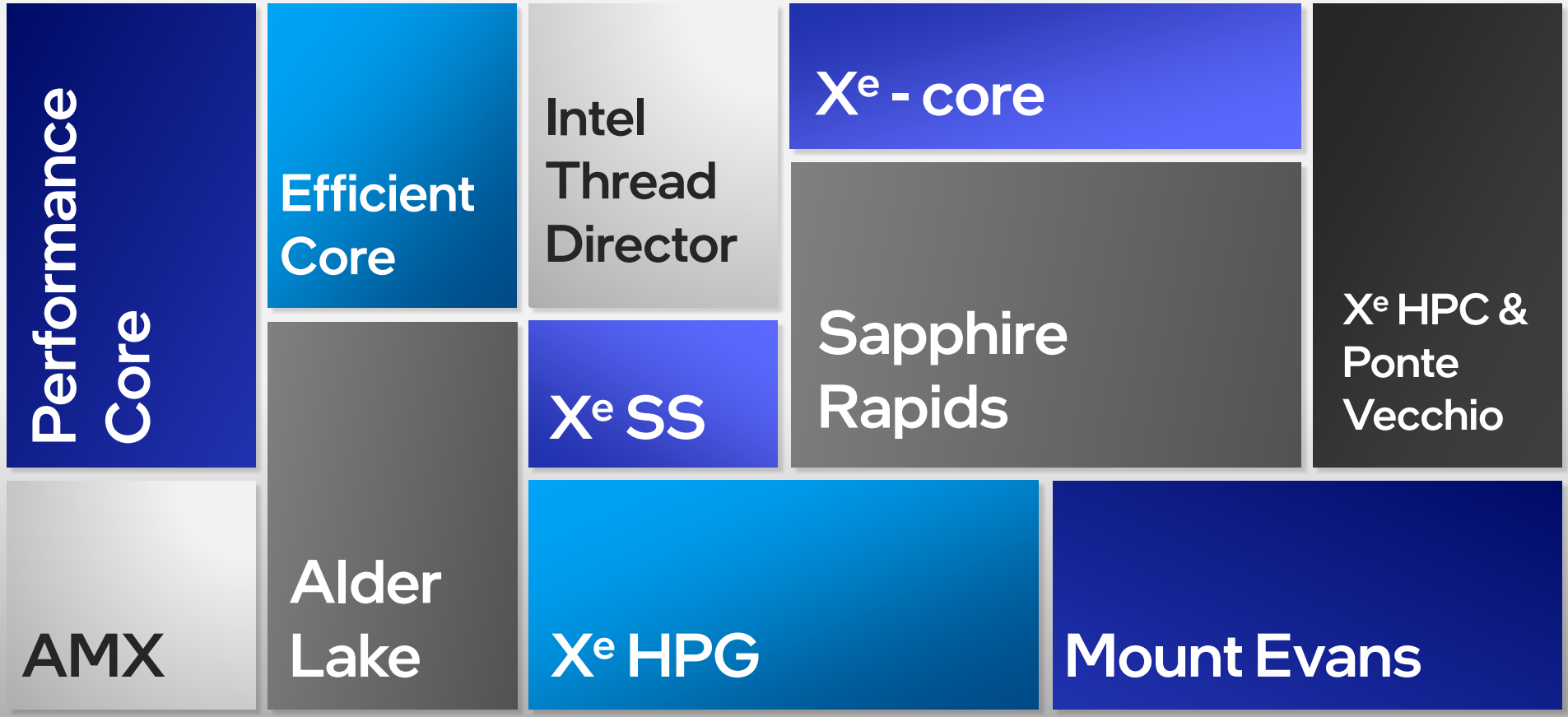
Hybrid Compute Cluster

in a Package



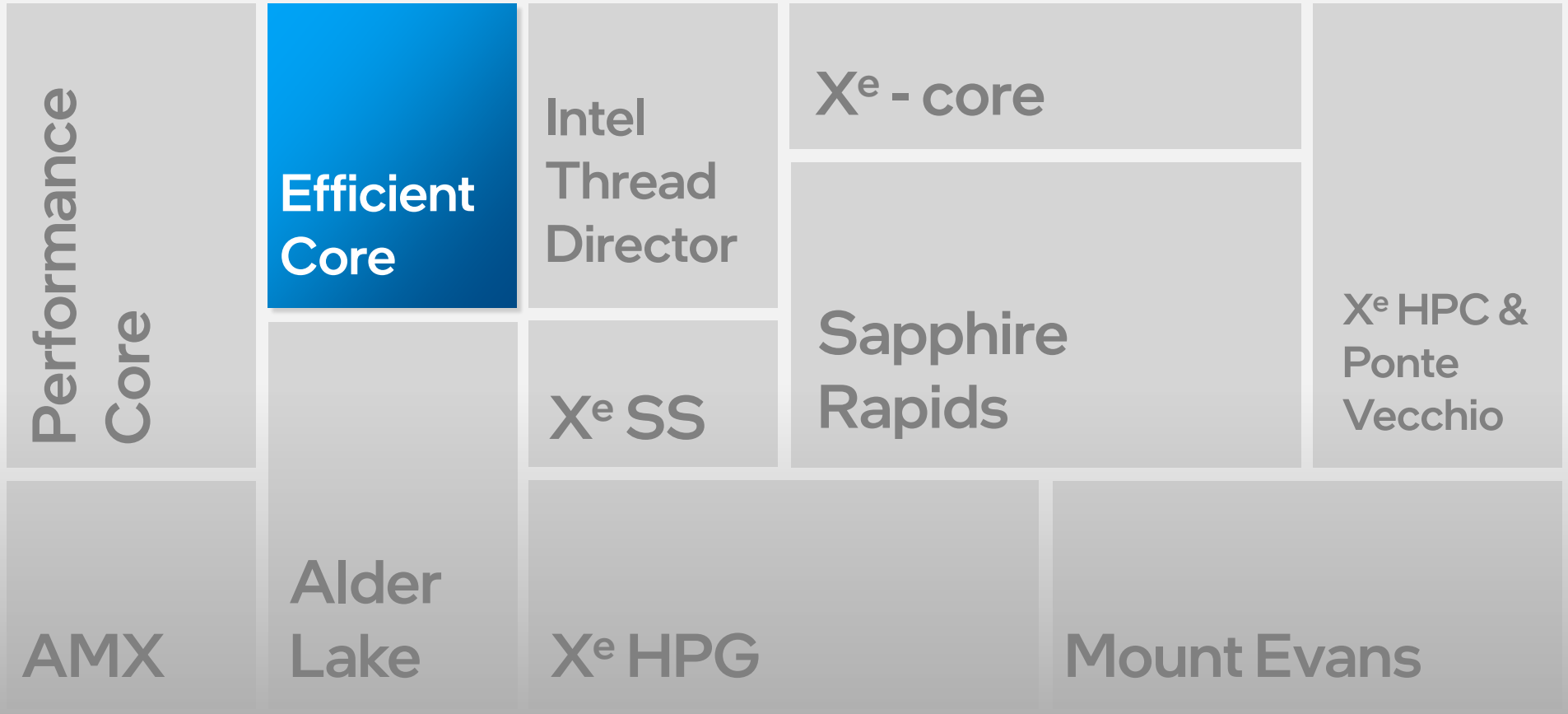
Architecture Day 2021

New Architectural Foundations



Architecture Day 2021

New Architectural Foundations



Efficient x86 Core

Stephen Robinson

Microarchitecture Goals

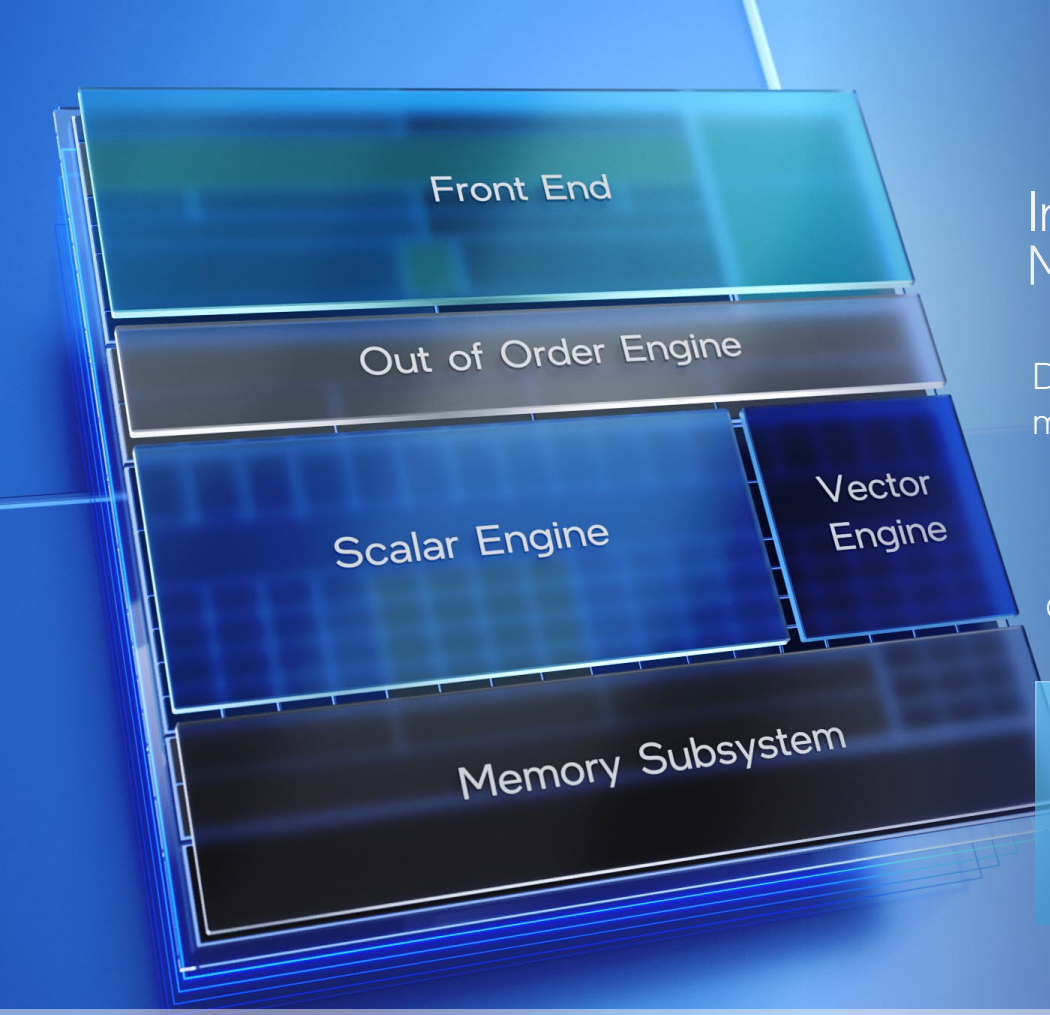
Highly Scalable Architecture To Address the Throughput Efficiency Needs For the Next Decade of Compute

Intel's Most Efficient Performant CPU

Dense & Highly Scalable

Vector and AI Instruction Support

Wide Dynamic Range



Intel's **New** Efficient x86 Core Microarchitecture

Designed for throughput, enabling scalable multi-threaded performance for modern multi-tasking

Optimized for power and density efficient throughput with:

Deep Front-End
with on-demand length decode

Wide Back-End
with many execution ports

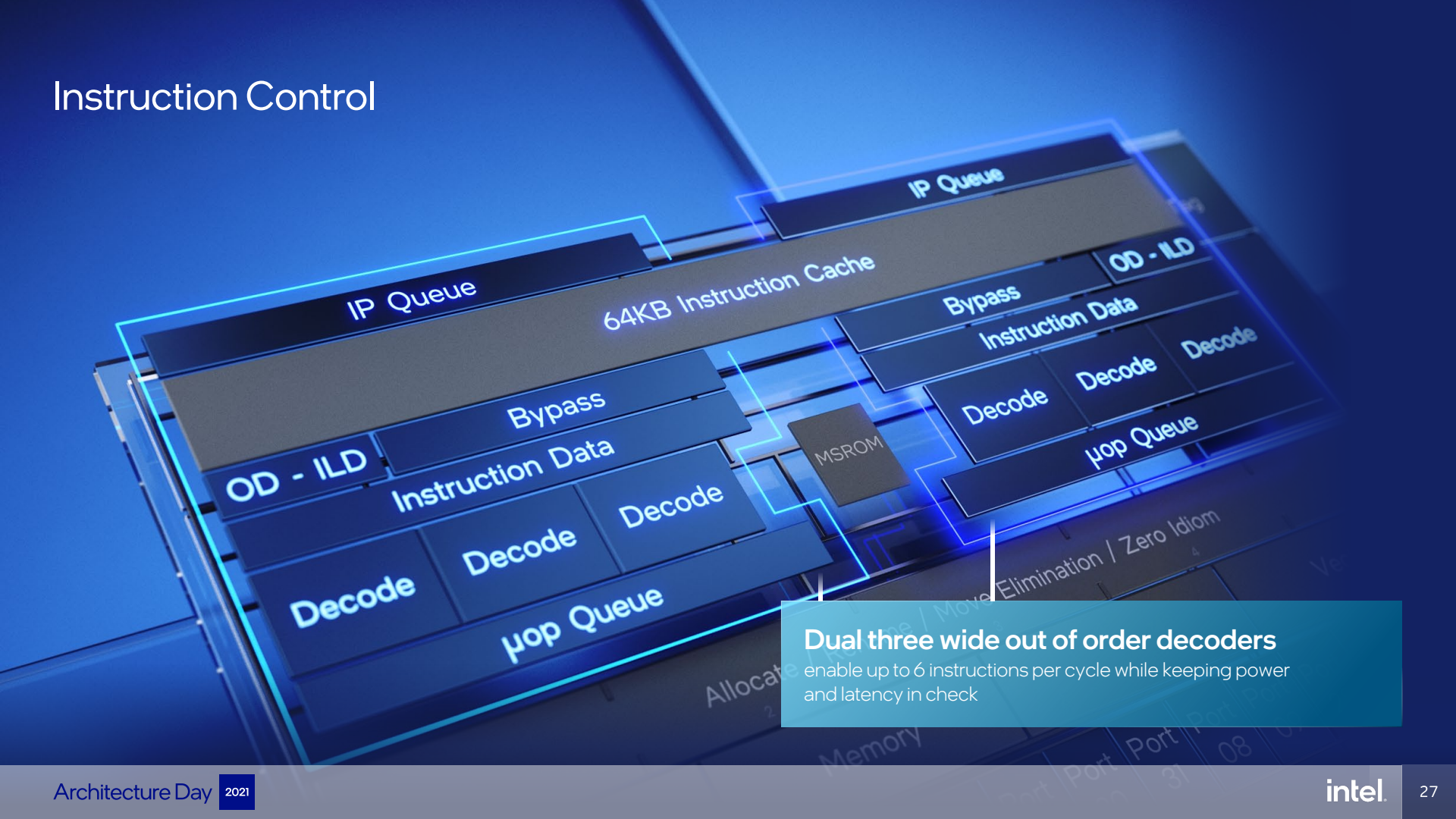
Optimized Design
for latest transistor technologies

Instruction Control

Large Instruction Cache (64KB)
with an on-demand instruction length decoder accelerates modern workloads with large code footprints

Accurate branch prediction
through deep branch history and large structure sizes

Instruction Control



Dual three wide out of order decoders
enable up to 6 instructions per cycle while keeping power and latency in check

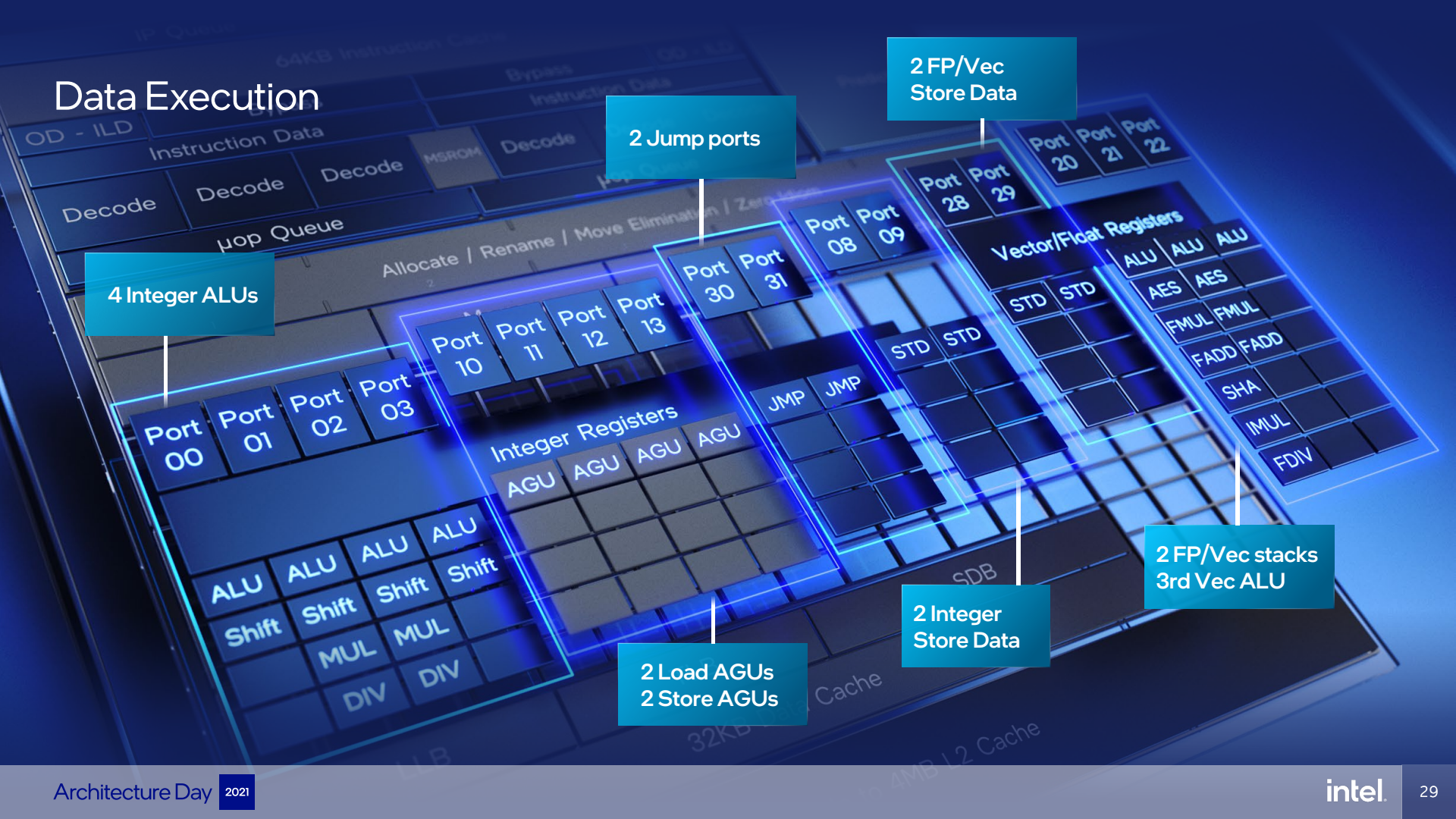
Data Execution

Five-wide allocation with eight-wide retire

256 entry out of order window
Discovers data parallelism

Seventeen execution ports
Executes data parallelism

Data Execution



Memory Subsystem

Dual Load + Dual Store

Up to 4MB L2
shared among four cores
with 64 Bytes/cycle bandwidth in 17 cycles of latency

Deep buffering
supporting 64 outstanding misses

Advanced Prefetchers
at all cache levels to detect a wide variety of streams

Intel® Resource Director Technology
enables software to control fairness among the cores
and between different software threads

Modern Instruction Set

Security

Intel® Control-flow Enforcement Technology designed to improve defense in depth

Intel® VT-rp (Virtualization Technology redirect protection) Supported

Advanced speculative execution validation methodology

Support for **Advanced Vector Instructions** with AI extensions

Wide Vector Instruction Set Architecture

Floating point multiply-accumulate (FMA) instructions for 2x throughput

Key instruction additions to enable integer AI throughput (VNNI)

Efficiency in Both Power and Performance per Transistor

Intense focus on feature selection and design implementation costs

to maximize area efficiency, which in turns enables core count scaling

Low switching energy per instruction

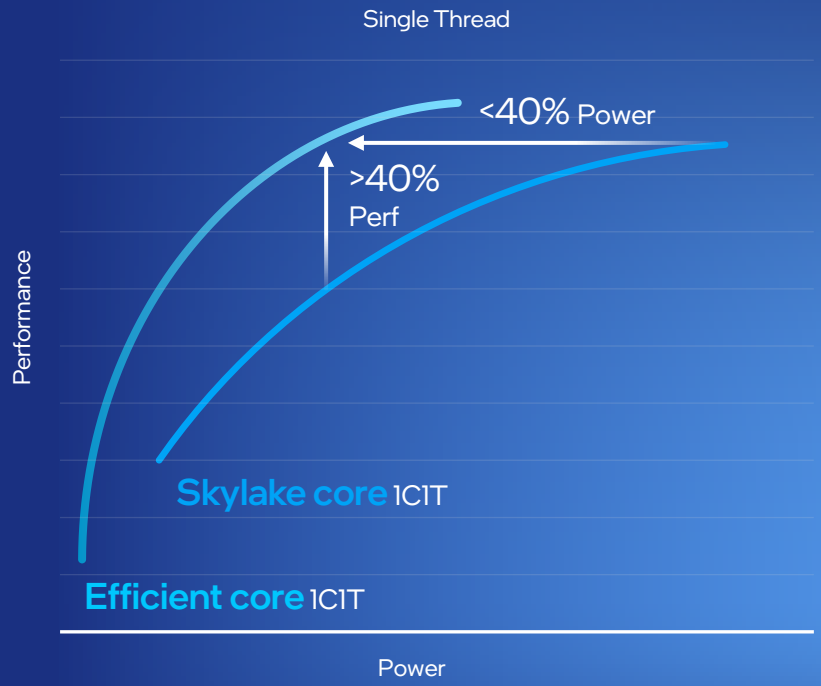
to maximize power constrained throughput, key for today's throughput-driven workloads

Reduced operating voltage required for all frequencies

saving power while extending the performance range

$$P = C \times F \times V^2$$

Latency Performance

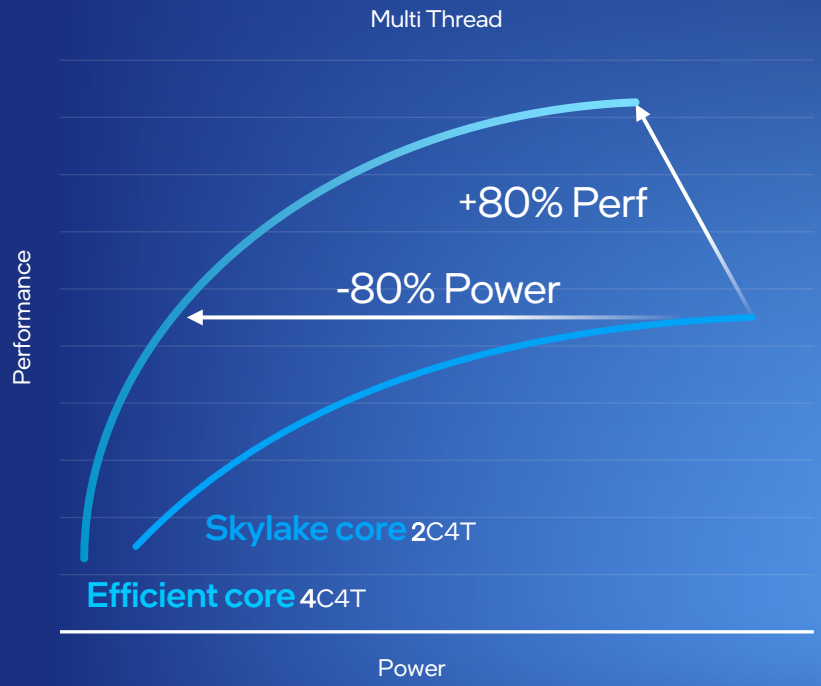


>40% Performance at ISO Power

<40% Power at ISO performance

SPECrate2017_int_base estimates using an open source compiler, iso-binary.
For workloads and configurations visit www.intel.com/ArchDay21claims. Results may vary.

Throughput Performance



+80% More Performance

-80% Power at ISO performance

SPECrate2017_int_base estimates using an open source compiler, iso-binary
For workloads and configurations visit www.intel.com/ArchDay21claims. Results may vary.

Intel's **New** Efficient x86 Core Microarchitecture

Designed for throughput, enabling scalable multi-threaded performance for modern multi-tasking

Optimized for power and density efficient throughput with:

Deep Front-End
with on-demand length decode

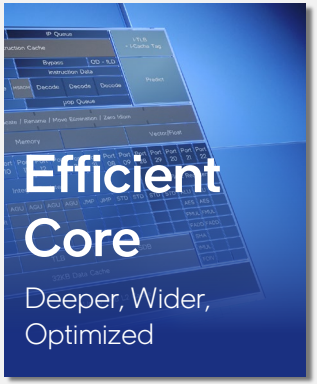
Wide Back-End
with many execution ports

Optimized Design
for latest transistor technologies

Architecture Day 2021

New Architectural Foundations

Performance Core



Intel Thread Director

Xe - core

Xe HPC & Ponte Vecchio

Xe SS

Sapphire Rapids

AMX

Alder Lake

Xe HPG

Mount Evans

Performance x86 Core

Adi Yoaz

Performance
x86 Core

Architecture Goals

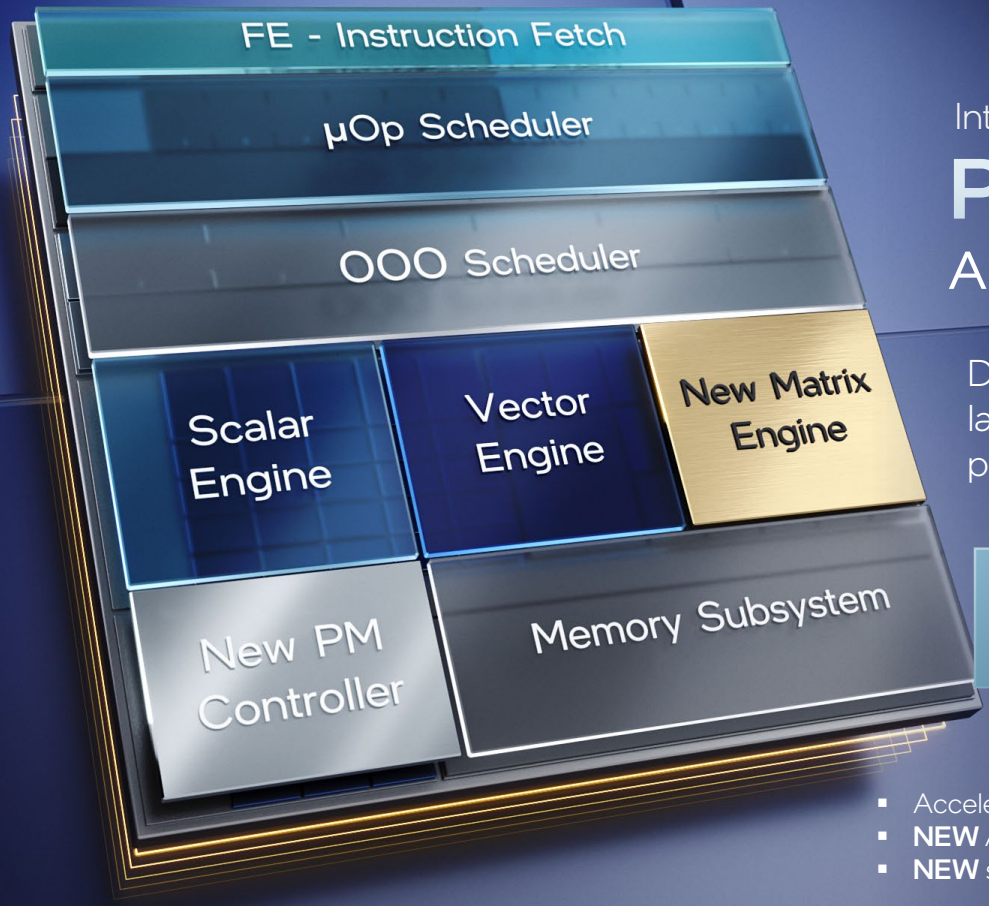
A Step Function in CPU Architecture Performance For the Next Decade of Compute

All in a tailored scalable architecture to serve the full range of Laptops to Desktops to Data Centers

Deliver a step function
in general purpose CPU performance

Advance the Arch/uArch with new features
for evolving trends of workload patterns

Innovate with next disruption in
AI performance acceleration



Intel's New

Performance x86 Core Architecture

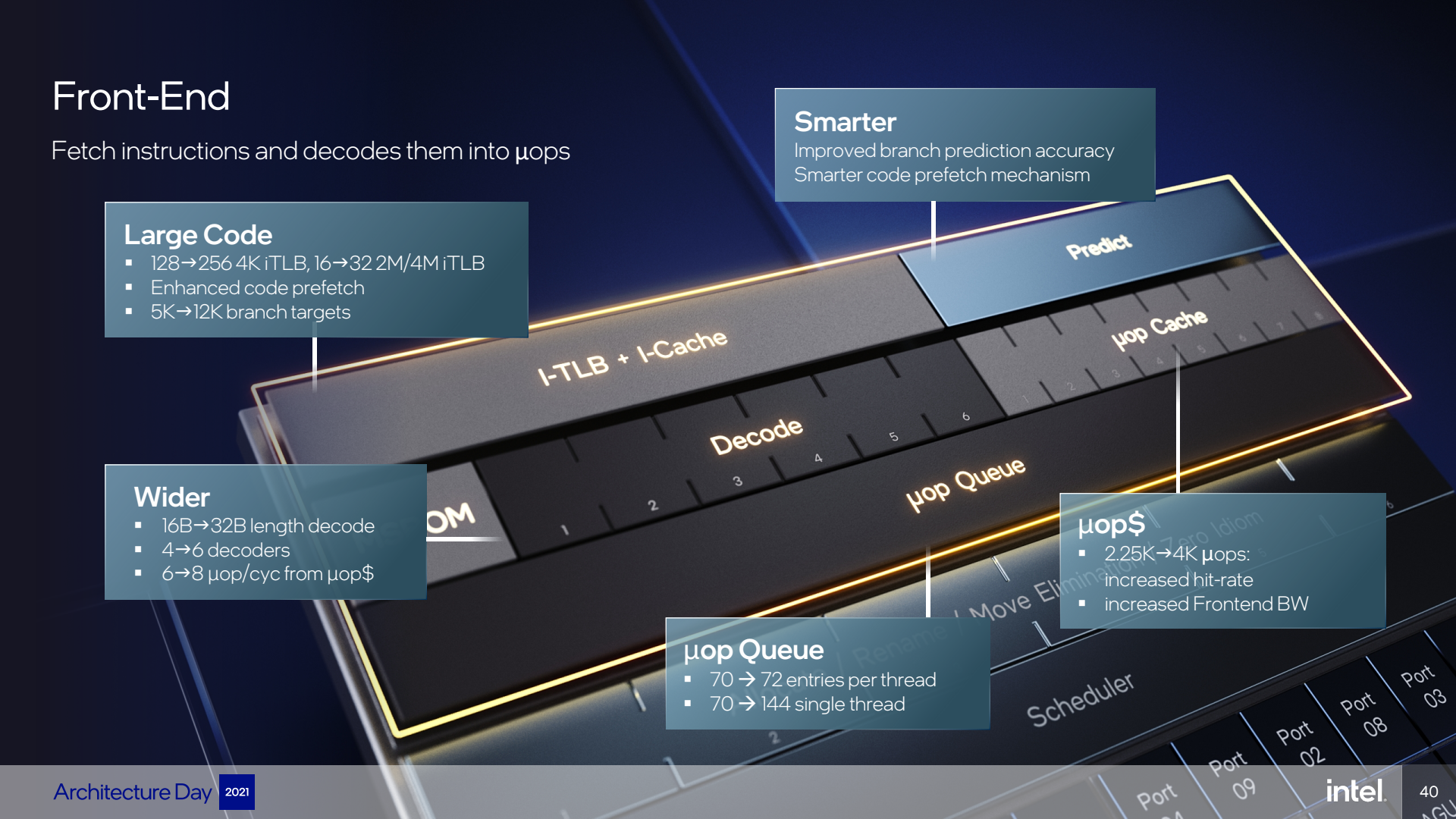
Designed for speed, pushing the limits of low latency and single threaded application performance via:

Wider | Deeper | Smarter

- Acceleration of workloads with large code footprint & large data sets
- **NEW** AI acceleration technology via coprocessor for matrix multiplication
- **NEW** smart PM controller for fine grain power budget management

Front-End

Fetch instructions and decodes them into µops



Out of Order Engine

Track µop dependencies and dispatch ready µops to execution units

Wider
5 → 6 wide allocation
10 → 12 execution ports

Deeper
512-entry Reorder-Buffer and larger Scheduler sizes

Smarter
More instructions "executed" at rename / allocation stage

Integer Execution Units

5th Integer execution port / ALU added

1-cycle LEA on all 5 ports
Used also for arithmetic calculations

Vector Execution Units

New Fast Adder (FADD):
Power efficient, low latency

FMA units support FP16 data type
FP16 added to Intel® AVX512 including complex numbers support

L1 Cache & Memory Subsystem

Wider
2 → 3 load ports:
3×256bit loads
2×512bit loads

Smarter
- Reduced effective Load Latency
- Faster Memory Disambiguation resolution

Deeper
Deeper Load Buffer and Store Buffer expose more memory parallelism

Large Data
- DTLB 64 → 96
- L1 D$: 12 → 16 fill buffers
- L1D$ enhanced prefetcher
- 2 → 4 page walkers

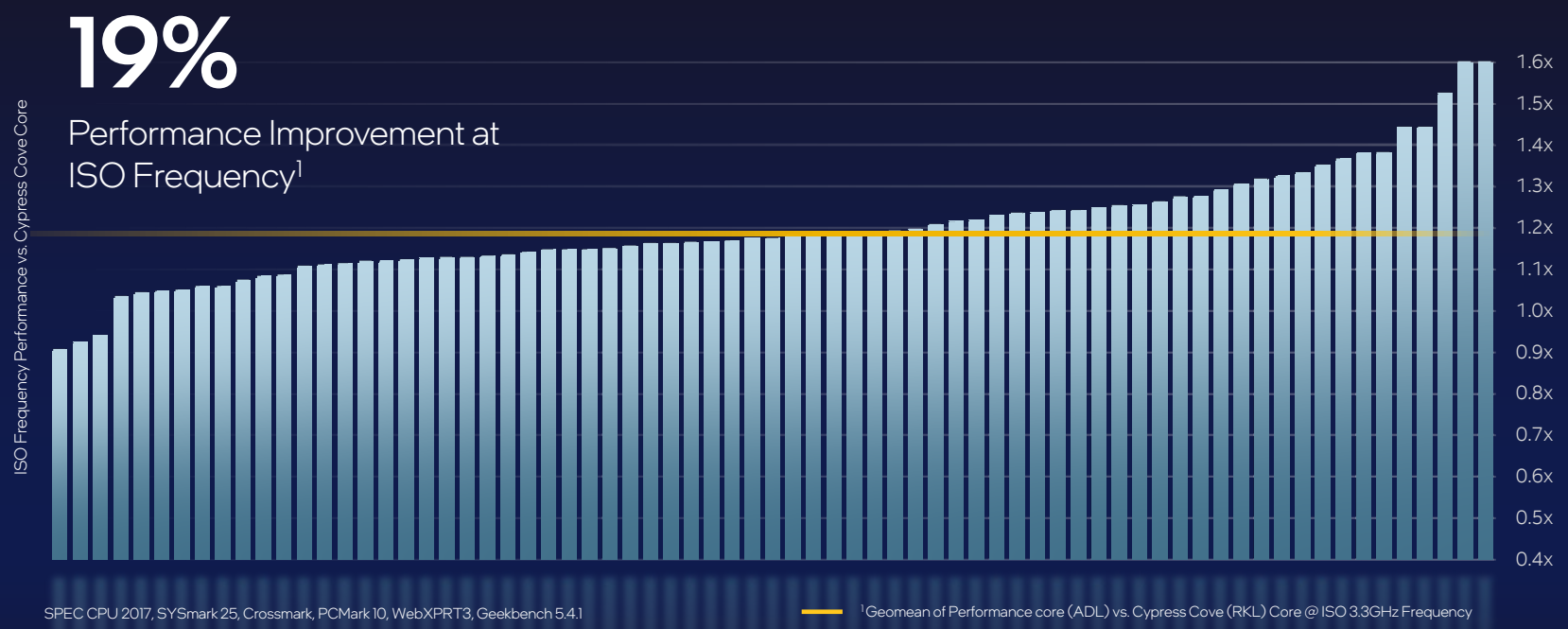
L2 Cache & Memory Subsystem

Bigger
L2$: 1.25MB (client) or 2MB (data center)

Faster
Max demand misses 32→48

Smarter
- L2$ pattern-based multi-path prefetcher
- Feedback-based prefetch throttling
- Full-line-write predictive bandwidth optimization – reduces DRAM reads

General-Purpose Performance Vs. 11th Gen Intel® Core™



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

Intel® Advanced Matrix Extensions (Intel® AMX)

Tiled Matrix Multiplication Accelerator - Data Center

AMX
2048 int8

8x

operations / cycle / core

VNNI
256 int8

Intel® Advanced Matrix Extensions (Intel® AMX)

Tiled Matrix Multiplication Accelerator - Data Center

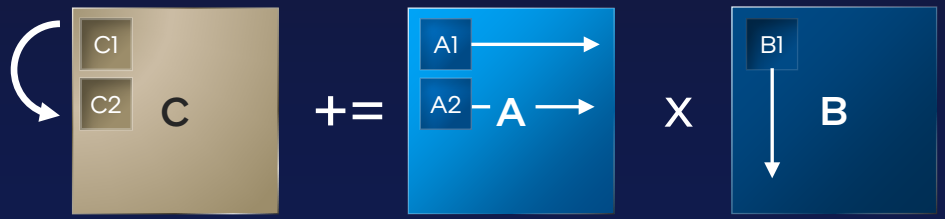
AMX architecture has two components:

Tiles

- A new expandable 2D register file – 8 new registers, 1Kb each: T0-T7
- Register file supports basic data operators – load/store, clear, set to constant, etc.
- TILES declares the state and is OS-managed by XSAVE architecture

TMUL

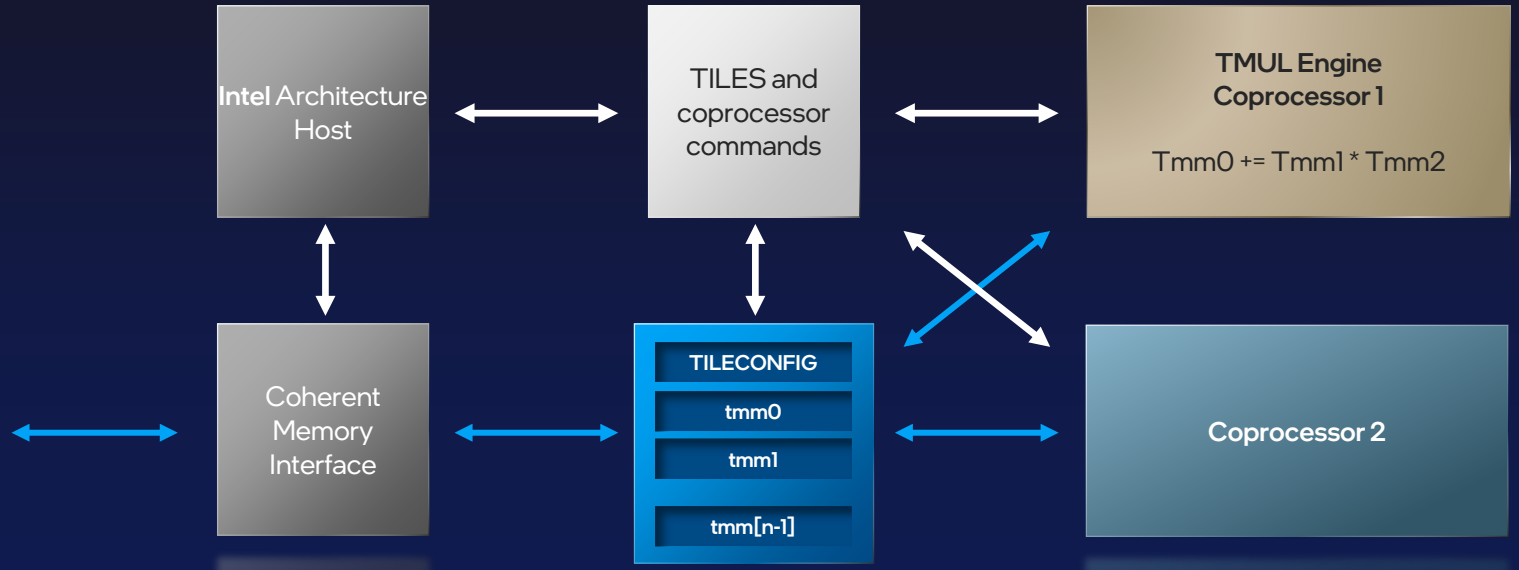
- Set of matrix multiplication instructions, the first operators on TILEs
- A MAC computation grid calculates 'tiles' of data
- TMUL – performs Matrix ADD-Multiplication (C=+A*C) using three Tile registers (T2=+T1*T0)
- TMUL requires TILE to be present



Express more work per instruction and per µop – save power for fetch/decode/OOO

Intel® Advanced Matrix Extensions (Intel® AMX)

Architecture



- New state to be managed by OS
- Commands and status delivered synchronously via TILE/accelerator instructions
- Dataflow – accelerators communicate to host through memory

New

Performance

x86 Core

A Step Function in CPU Architecture Performance For the Next Decade of Compute

A significant IPC boost at high power efficiency

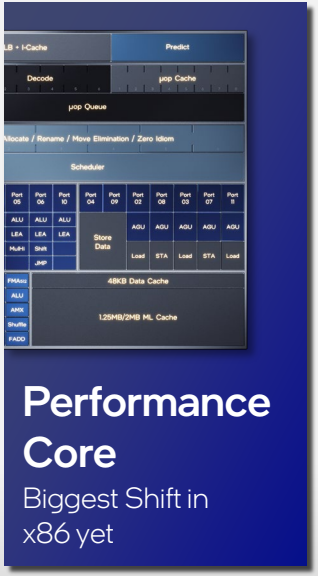
Wider | **Deeper** | **Smarter**

- Better supports large data set and large code footprint applications
- Enhanced power management improves frequency and power
- **Machine Learning Technology**: Intel® AMX – Tile Multiplication

All in a tailored scalable architecture to serve the full range of Laptops to Desktops to Data Centers

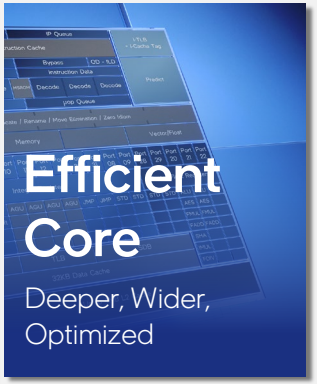
Architecture Day 2021

New Architectural Foundations



Performance Core

Biggest Shift in x86 yet



Efficient Core

Deeper, Wider, Optimized

Intel Thread Director

Xe - core

Xe HPC & Ponte Vecchio

Sapphire Rapids

Xe SS

AMX

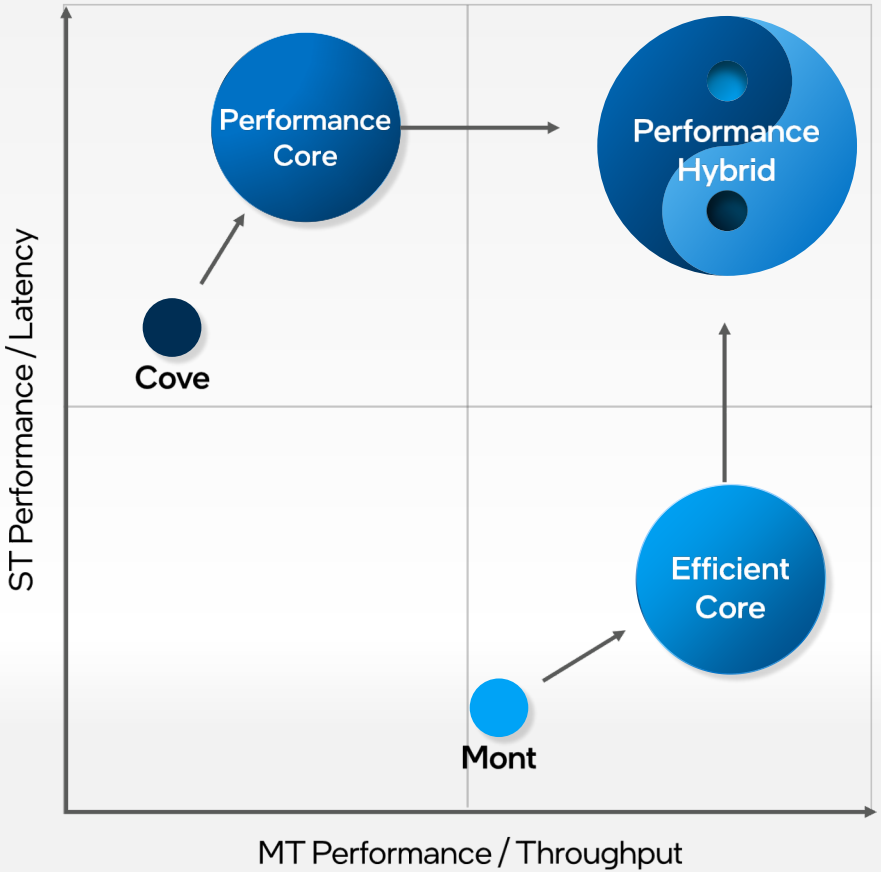
Advanced Matrix Extension - Engine

Alder Lake

Xe HPG

Mount Evans

Scalar Architecture Roadmap



Graph is for conceptual illustration purposes only.

Intel Thread Director

Rajshree Chabukswar

Performance Hybrid

Scheduling Goals

- Software Transparent
- Real -Time Adaptive
- Scalable from Mobile to Desktop

Introducing

Intel **Thread Director**

Intelligence built directly into the core

Monitors the runtime instruction mix
of each thread and as well as the state of each core – with nanosecond precision

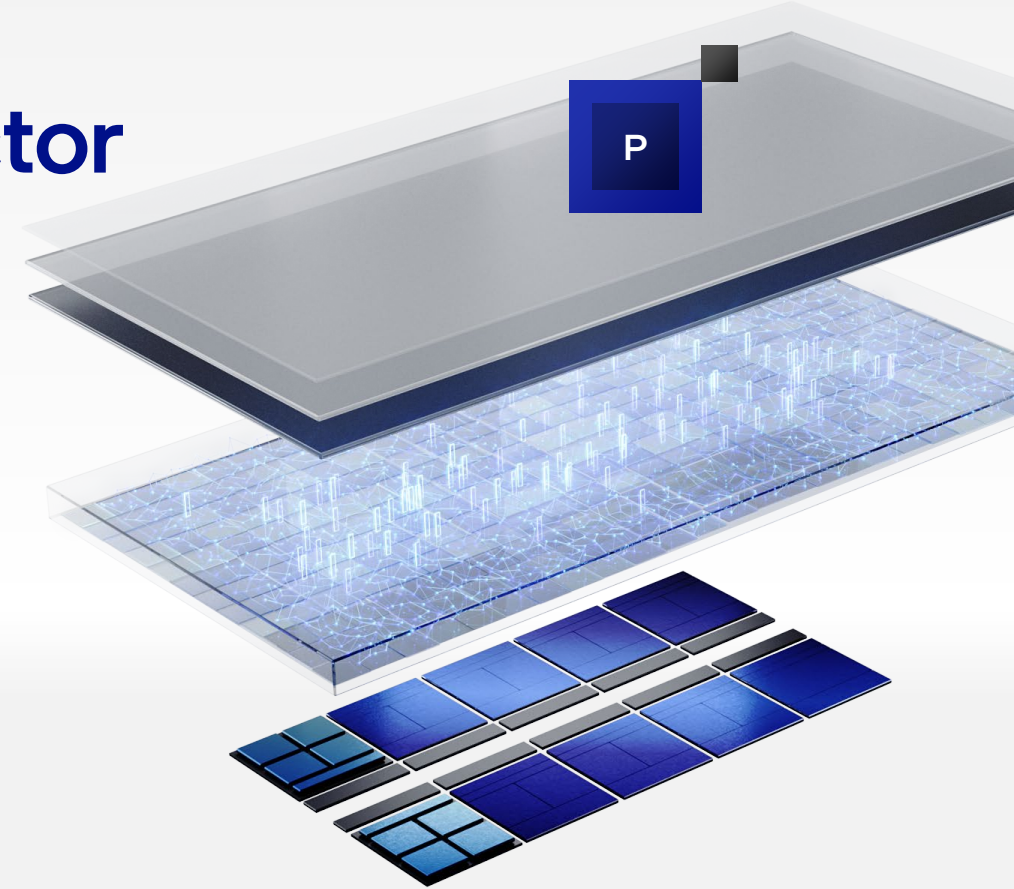
Provides runtime feedback to the OS
to make the optimal scheduling decision for any workload or workflow

Dynamically adapts guidance
based on the thermal design point, operating conditions, and power settings – without any user input

Introducing

Intel **Thread Director**

Scheduling Examples

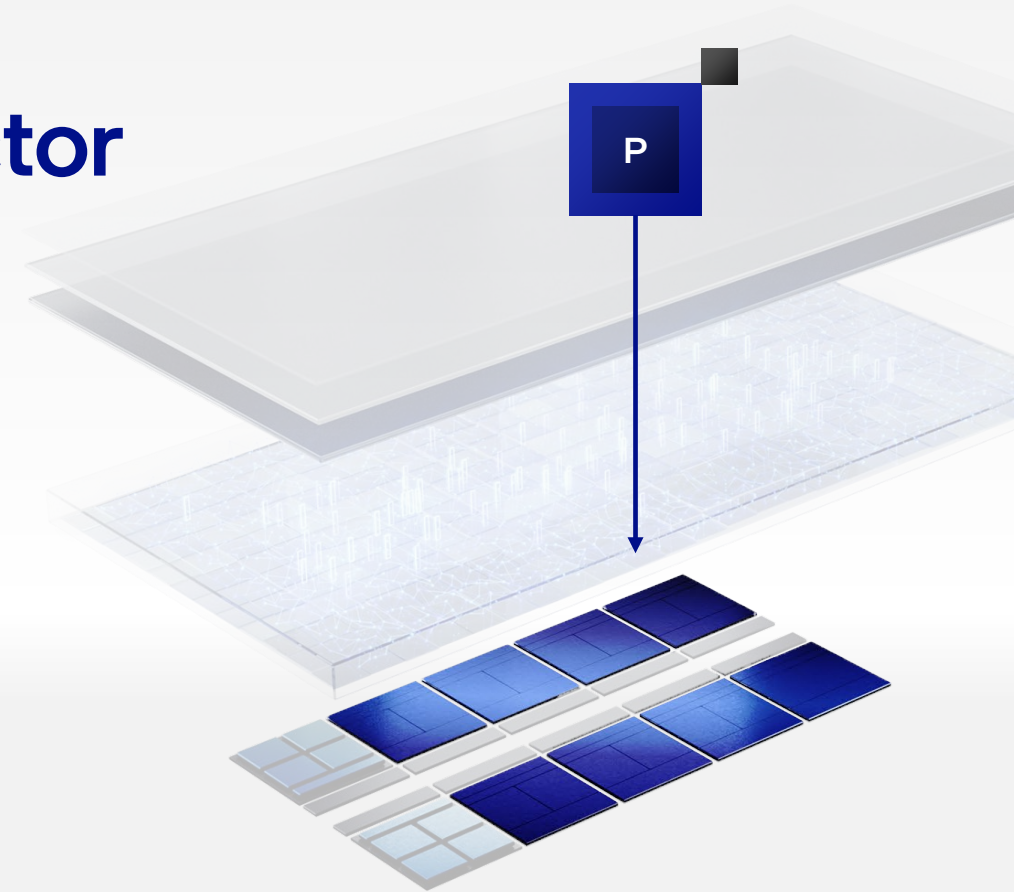


Introducing

Intel **Thread Director**

Scheduling Examples

1 Priority tasks scheduled on P-cores

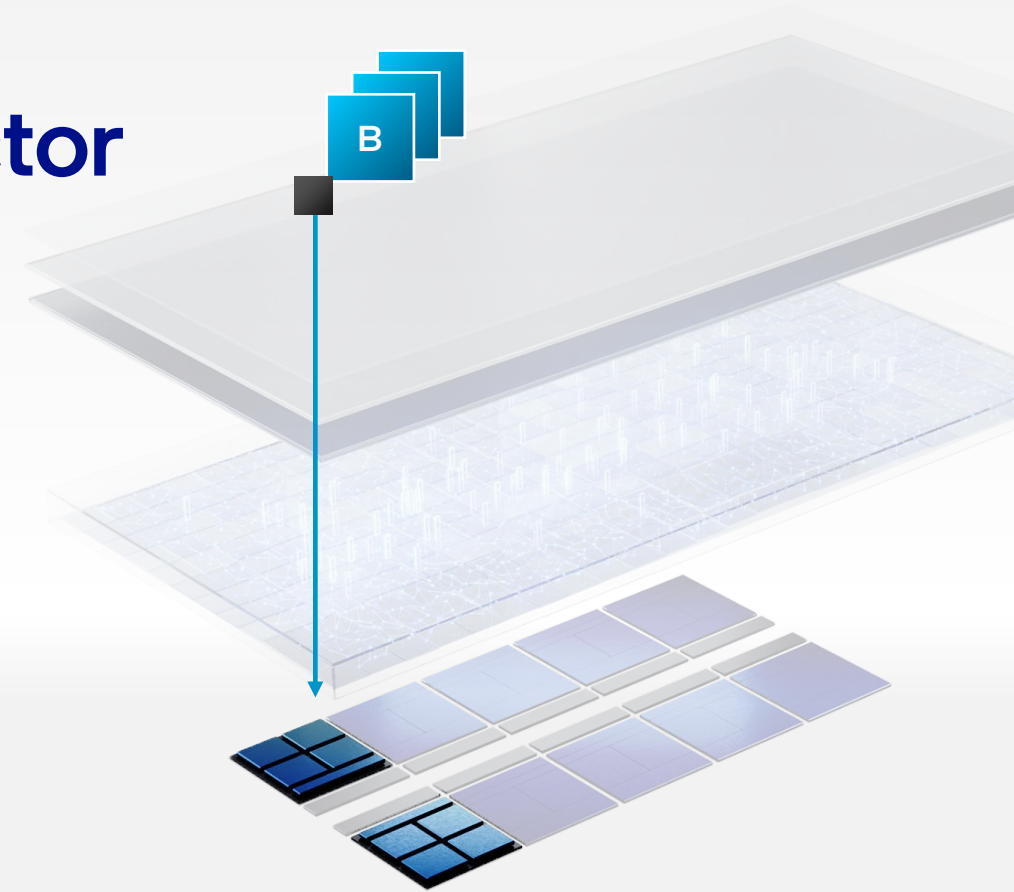


Introducing

Intel **Thread Director**

Scheduling Examples

1. Priority tasks scheduled on P-cores
2. Background tasks scheduled on E-cores

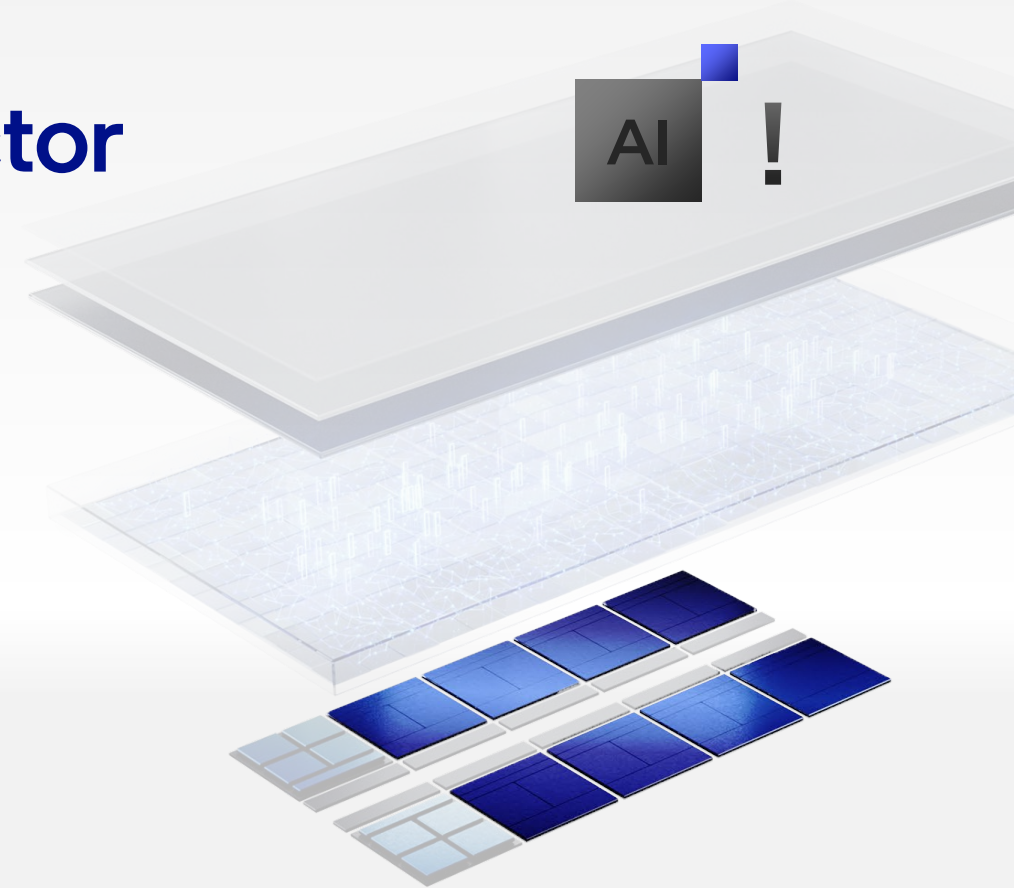


Introducing

Intel **Thread Director**

Scheduling Examples

1. Priority tasks scheduled on P-cores
2. Background tasks scheduled on E-cores
3. New AI thread ready

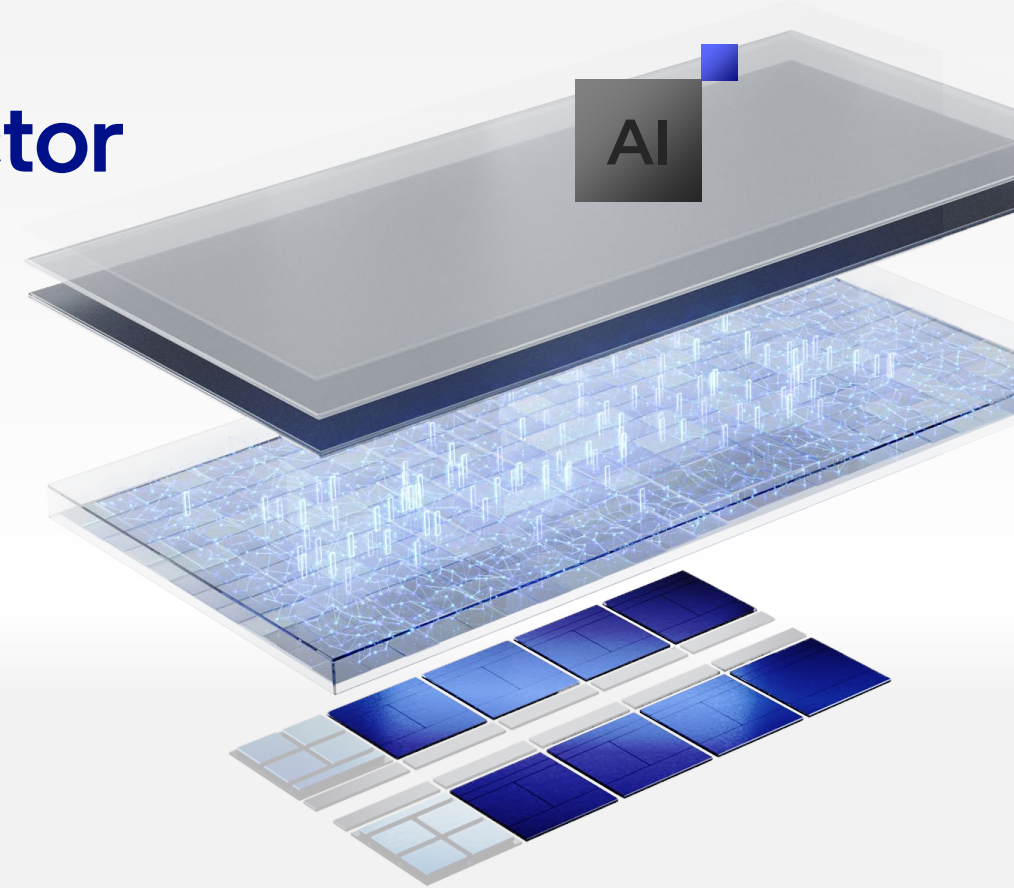


Introducing

Intel **Thread Director**

Scheduling Examples

1. Priority tasks scheduled on P-cores
2. Background tasks scheduled on E-cores
3. New AI thread ready

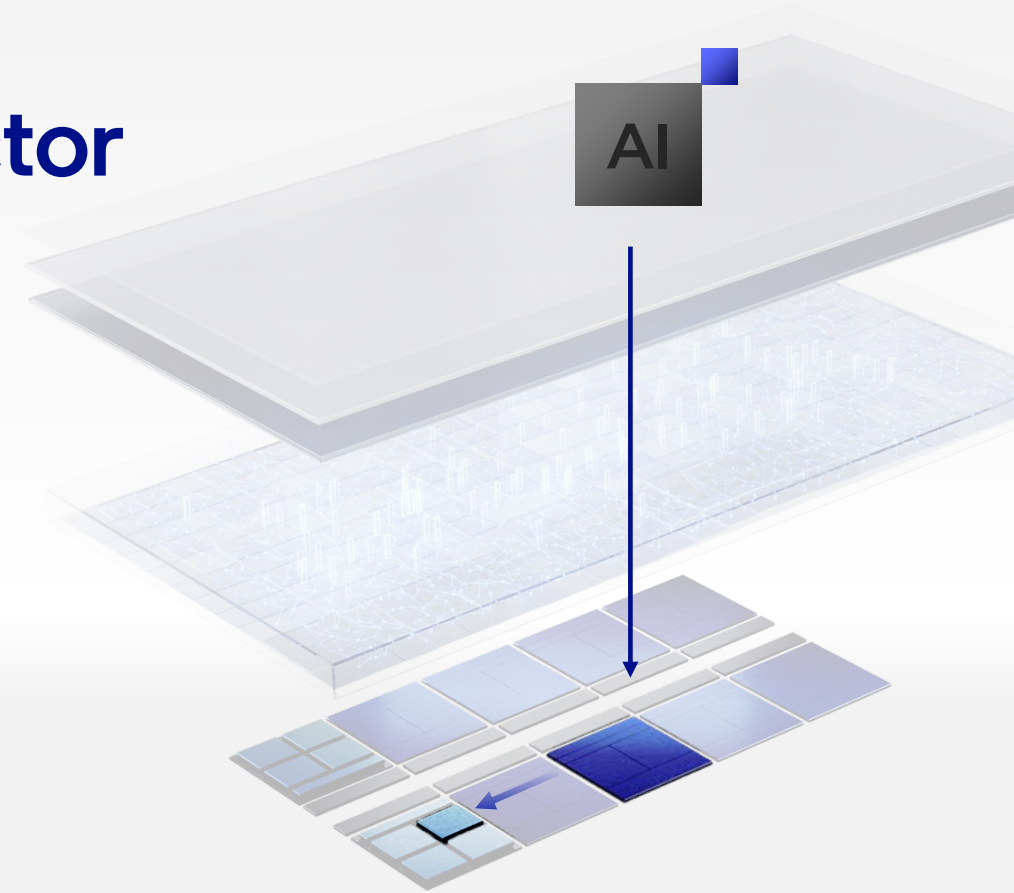


Introducing

Intel **Thread Director**

Scheduling Examples

1. Priority tasks scheduled on P-cores
2. Background tasks scheduled on E-cores
3. AI thread prioritized on P-core

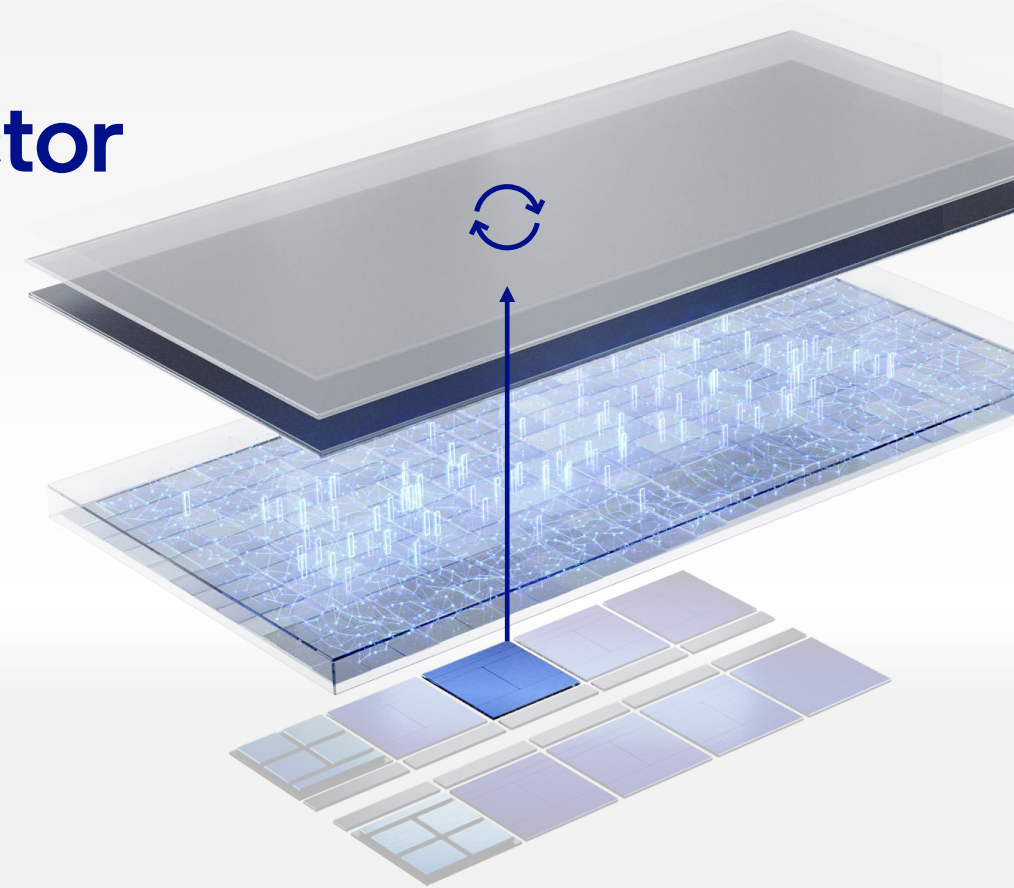


Introducing

Intel **Thread Director**

Scheduling Examples

1	Priority tasks scheduled on P-cores
2	Background tasks scheduled on E-cores
3	AI thread prioritized on P-core
4	Spin loop wait moved from P to E-core



Introducing

Intel **Thread Director**

Scheduling Examples

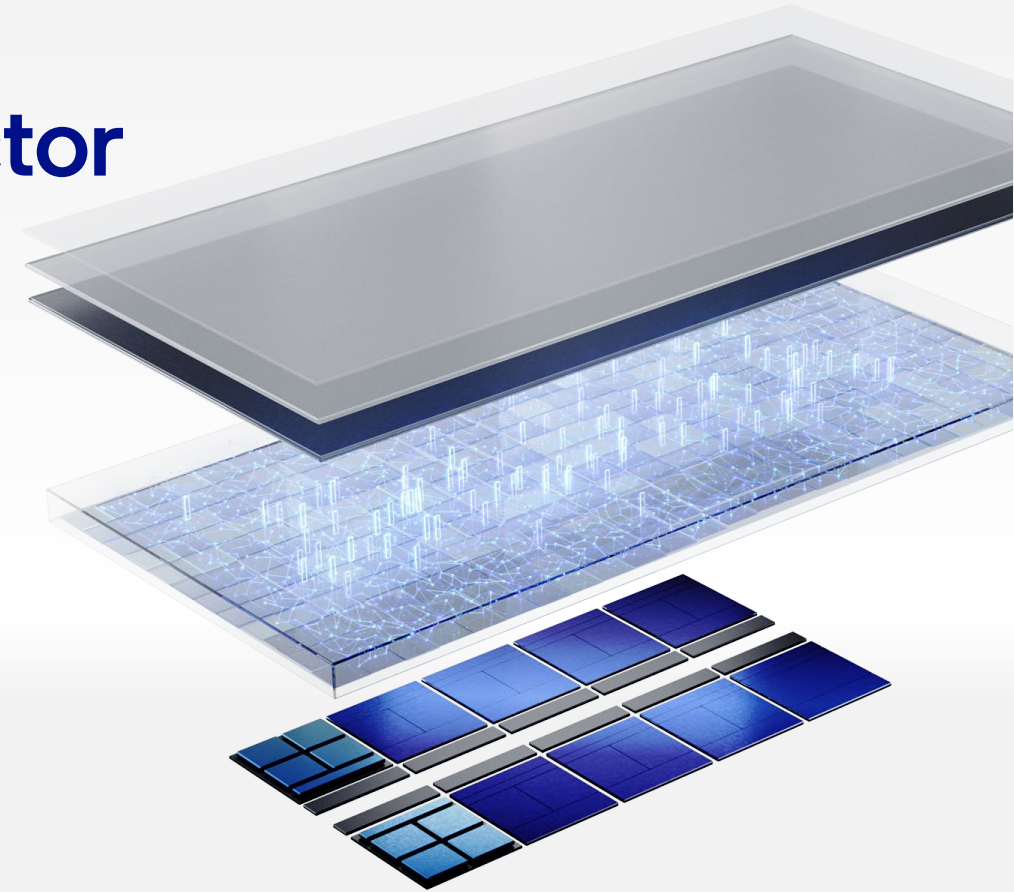
1. Priority tasks scheduled on P-cores
2. Background tasks scheduled on E-cores
3. AI thread prioritized on P-core
4. Spin loop wait moved from P to E-core

Introducing

Intel **Thread Director**

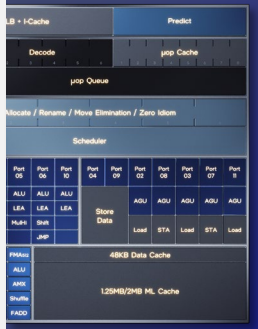
Scheduling Examples

1	Priority tasks scheduled on P-cores
2	Background tasks scheduled on E-cores
3	AI thread prioritized on P-core
4	Spin loop wait moved from P to E-core



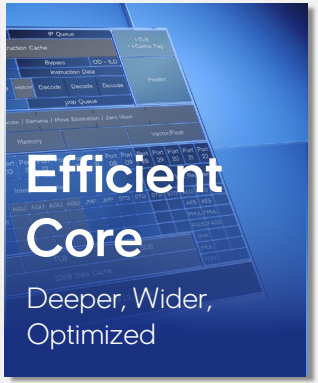
Architecture Day 2021

New Architectural Foundations



Performance Core

Biggest Shift in x86 yet

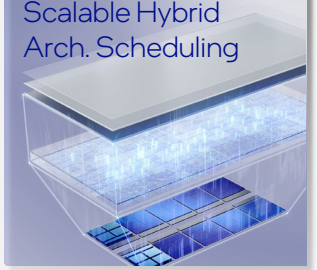


Efficient Core

Deeper, Wider, Optimized

Intel Thread Director

Scalable Hybrid Arch. Scheduling



Xᵉ - core

Xᵉ SS

Sapphire Rapids

Xᵉ HPC & Ponte Vecchio

AMX

Advanced Matrix Extension - Engine

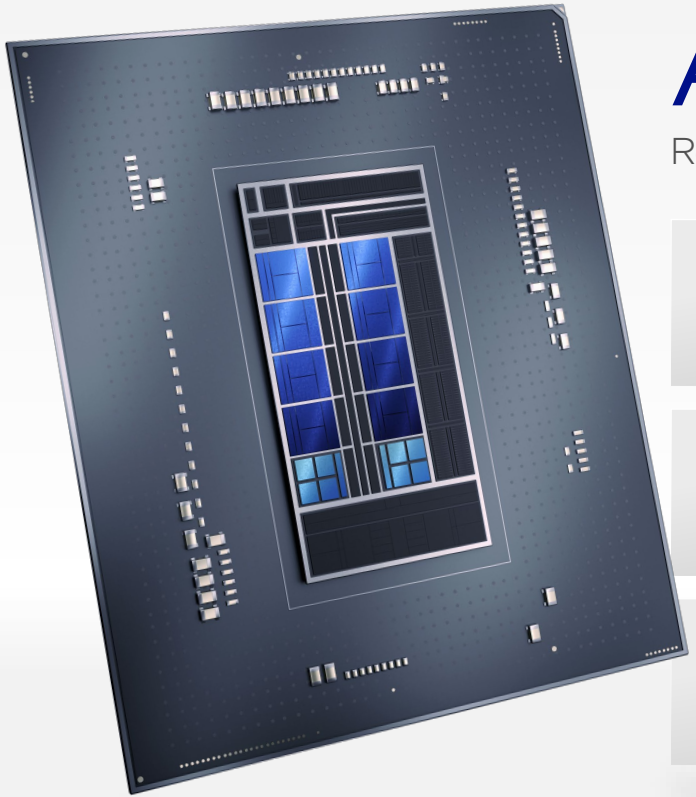
Alder Lake

Xᵉ HPG

Mount Evans

Alder Lake

Arik Gihon



Introducing

Alder Lake

Reinventing Multi Core Architecture

Single, Scalable SoC Architecture

All Client Segments – 9W to 125W – built on Intel 7 process

All-New Core Design

Performance Hybrid with Intel Thread Director

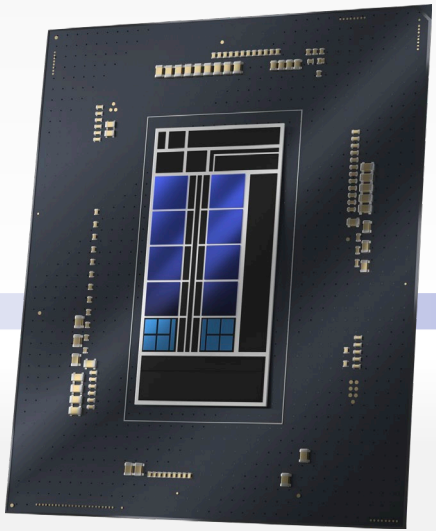
Industry-Leading Memory & I/O

DDR5, PCIe Gen5, Thunderbolt™ 4, Wi-Fi 6E

Scalable Client Architecture

Desktop

LGA 1700
Socket



Mobile

BGA Type3
50 x 25 x 1.3 mm

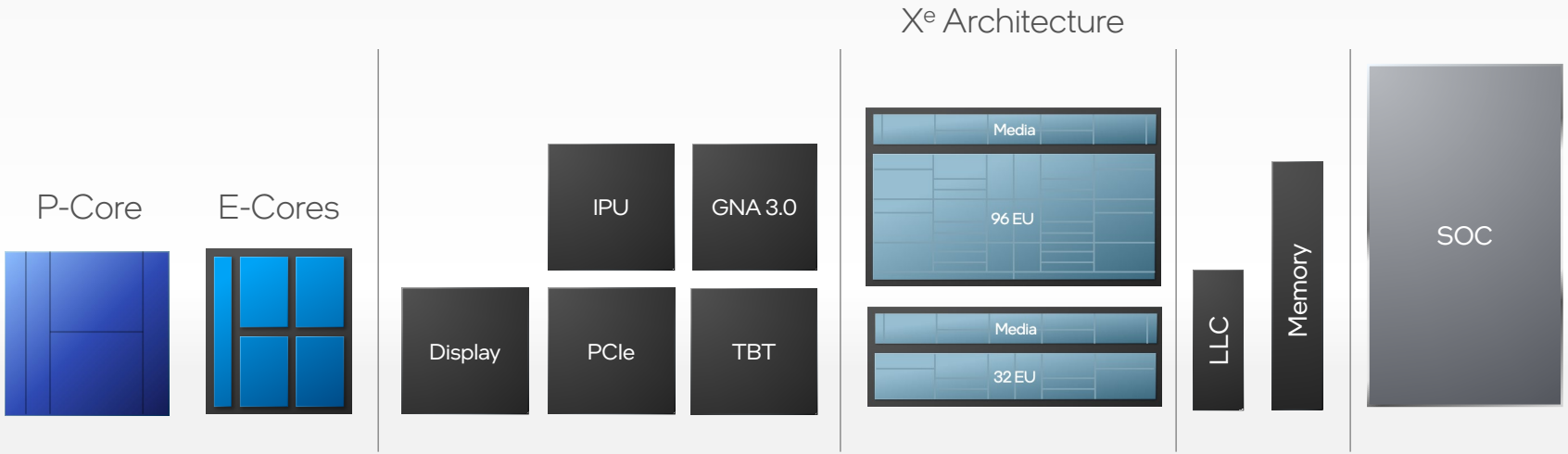
Ultra Mobile

BGA Type4 HDI
28.5 x 19 x 1.1 mm

Visit www.intel.com/ArchDay21claims for details

Alder Lake

Building Blocks

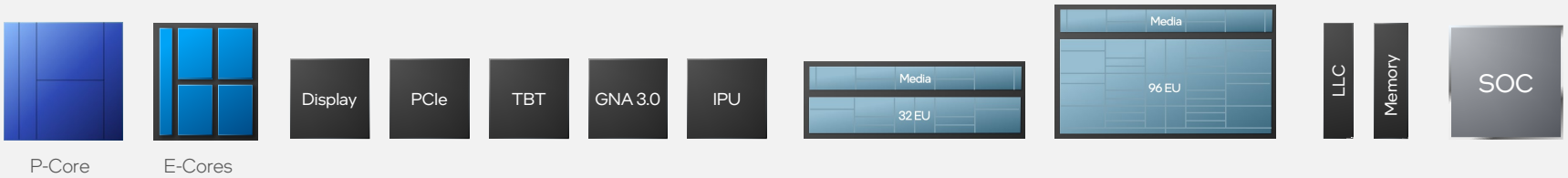


Desktop

Mobile

Ultra Mobile

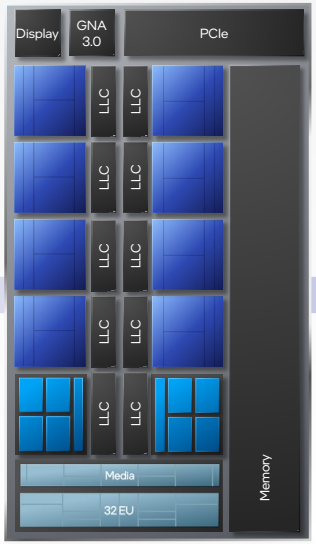
Building Blocks



Desktop

Mobile

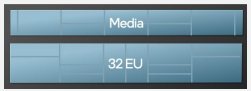
Ultra Mobile

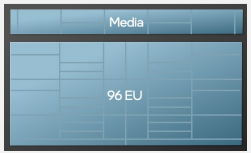


Building Blocks

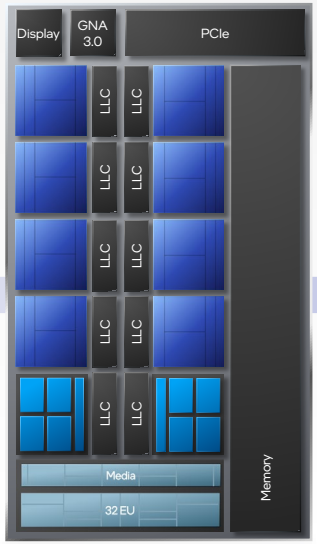
P-Core

E-Cores

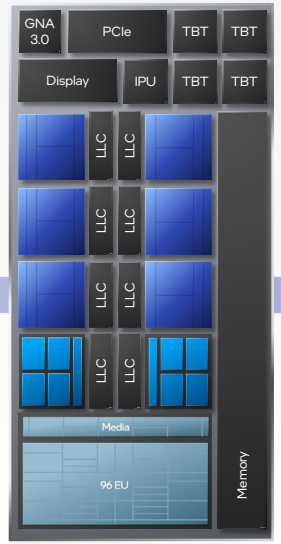




Desktop



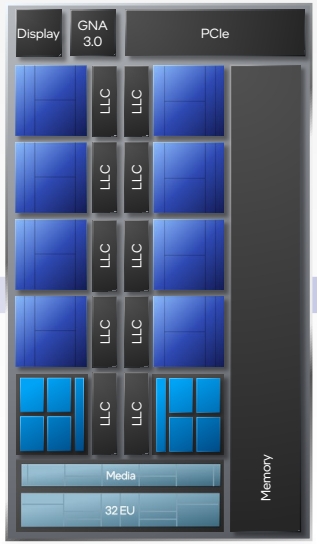
Mobile



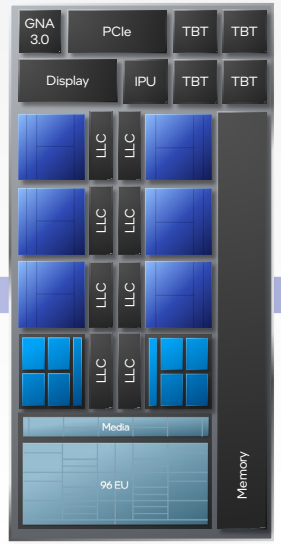
Ultra Mobile

Building Blocks

Desktop

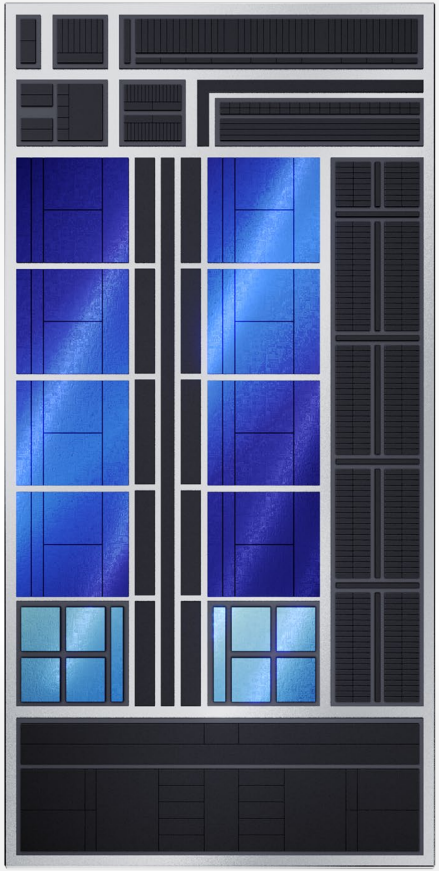


Mobile



Ultra Mobile

Building Blocks



Alder Lake

Core/Cache

Up To
16 Cores

8 Performance
8 Efficient

Up To
24 Threads

2T per P-core
1T per E-core

Up to
30MB

Non-inclusive
LL Cache

Alder Lake

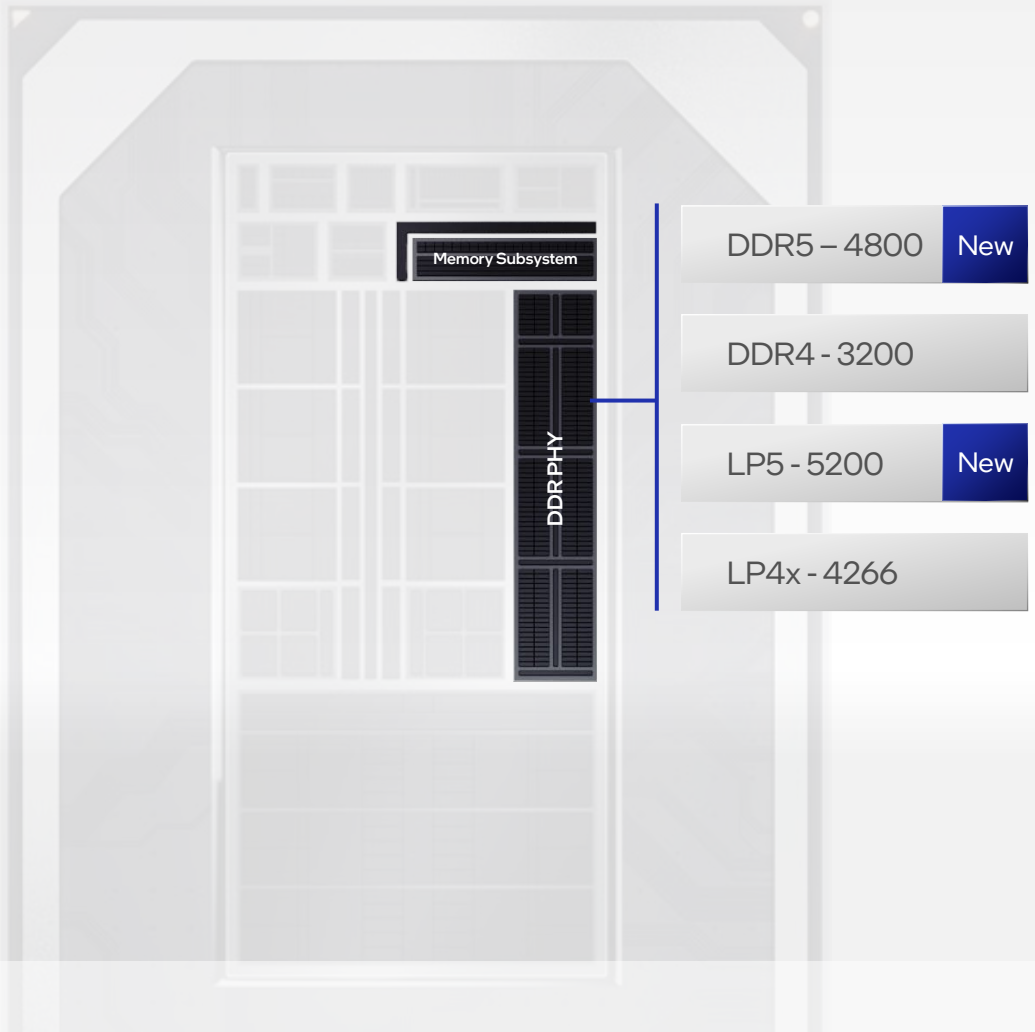
Memory

Leading the industry transition to DDR5

Support for all four major memory technologies

Dynamic voltage-frequency scaling

Enhanced overclocking support

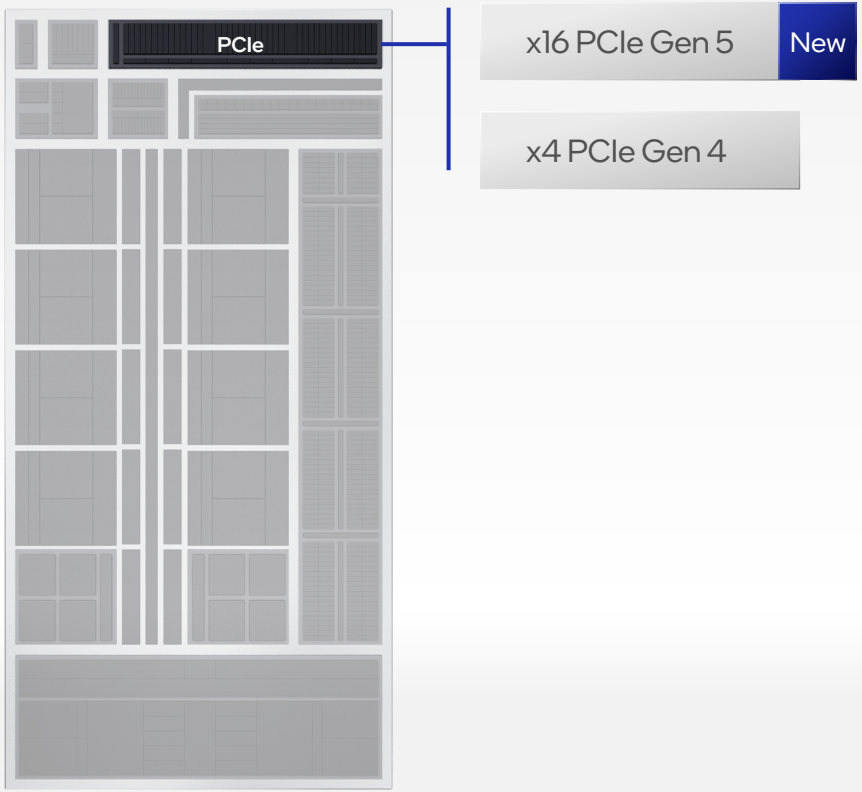


Alder Lake

PCIe

Leading the industry transition to PCIe Gen5

Up to 2X bandwidth vs. Gen4
Up to 64GB/s with x16 lanes



Visit www.intel.com/ArchDay21claims for details

Alder Lake

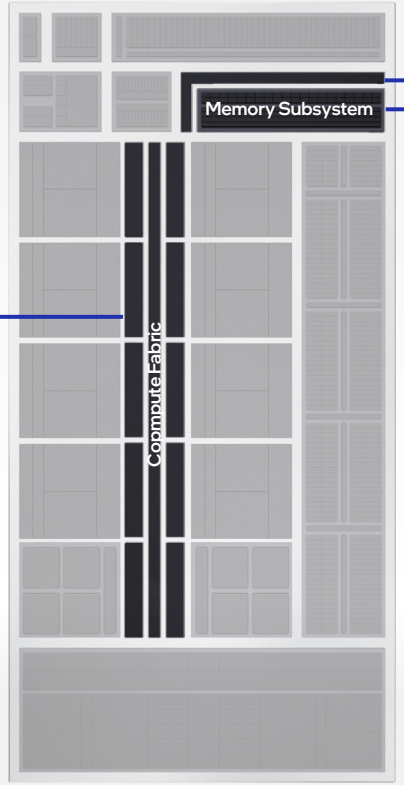
Interconnect

Compute Fabric

Up to

1000 GB/s

Dynamic Latency Optimization



I/O Fabric

Up to

64 GB/s

Real-time, demand-based BW control

Memory Fabric

Up To

204 GB/s

Dynamic Bus Width & Frequency

Visit www.intel.com/ArchDay21claims for details

Beginning Fall 2021

Alder Lake

Reinventing Multi Core Architecture

Single, Scalable SoC Architecture

All Client Segments – 9W to 125W – built on Intel 7 process

All-New Core Design

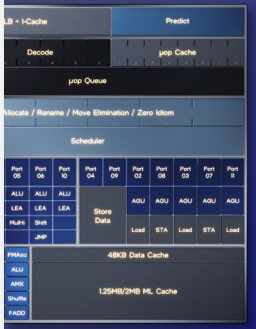
Performance Hybrid with Intel Thread Director

Industry-Leading Memory & I/O

DDR5, PCIe Gen5, Thunderbolt™ 4, Wi-Fi 6E

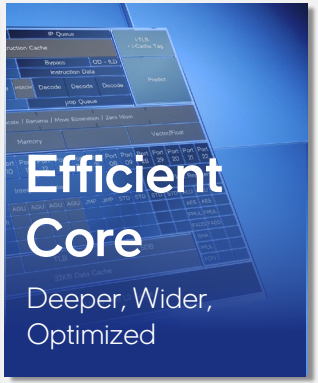
Architecture Day 2021

New Architectural Foundations



Performance Core

Biggest Shift in x86 yet



Efficient Core

Deeper, Wider, Optimized

Intel Thread Director

Scalable Hybrid Arch. Scheduling

Xe - core

Xe HPC & Ponte Vecchio

Sapphire Rapids

Xe SS

Alder Lake

Performance Hybrid

AMX

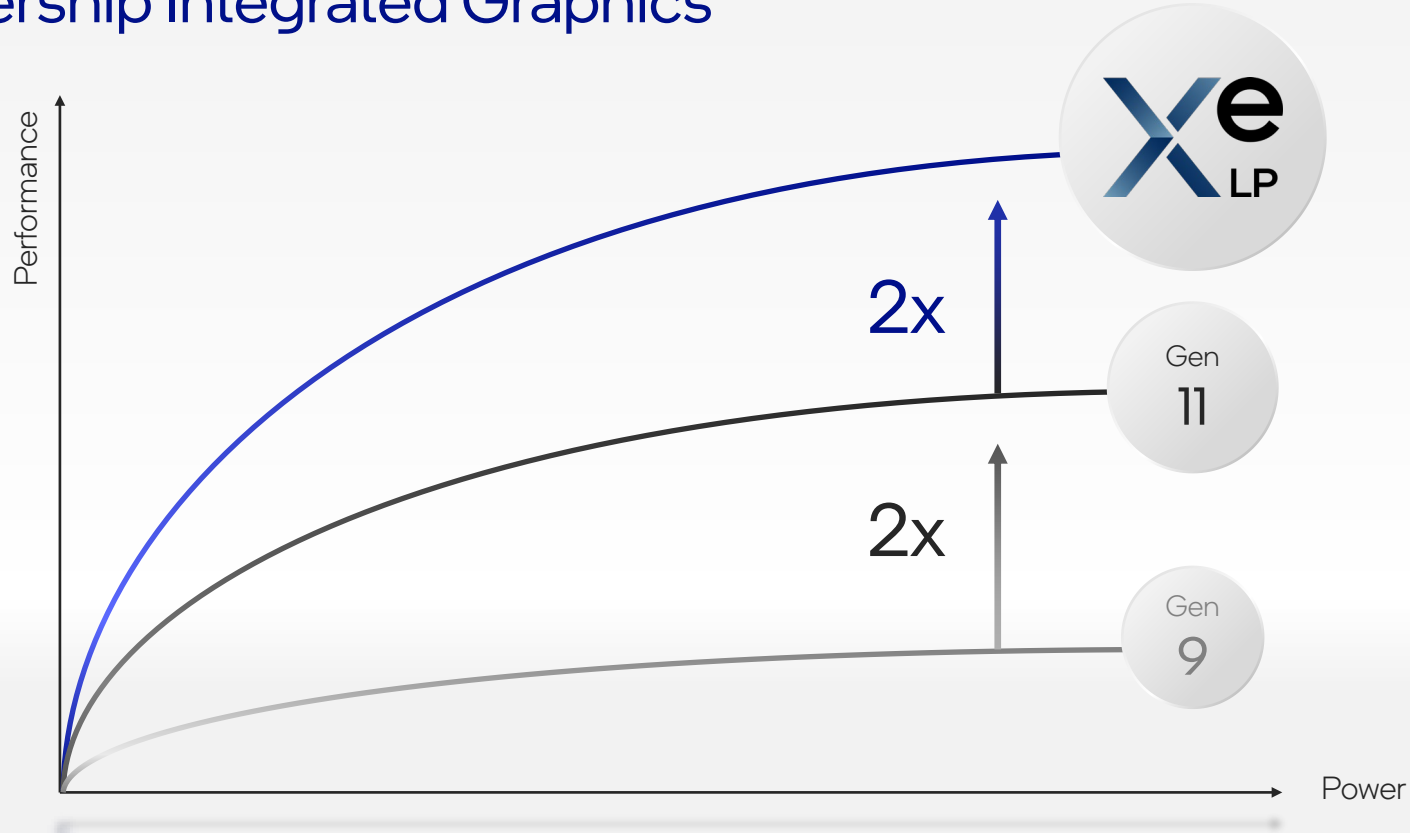
Advanced Matrix Extension - Engine

Xe HPG

Mount Evans

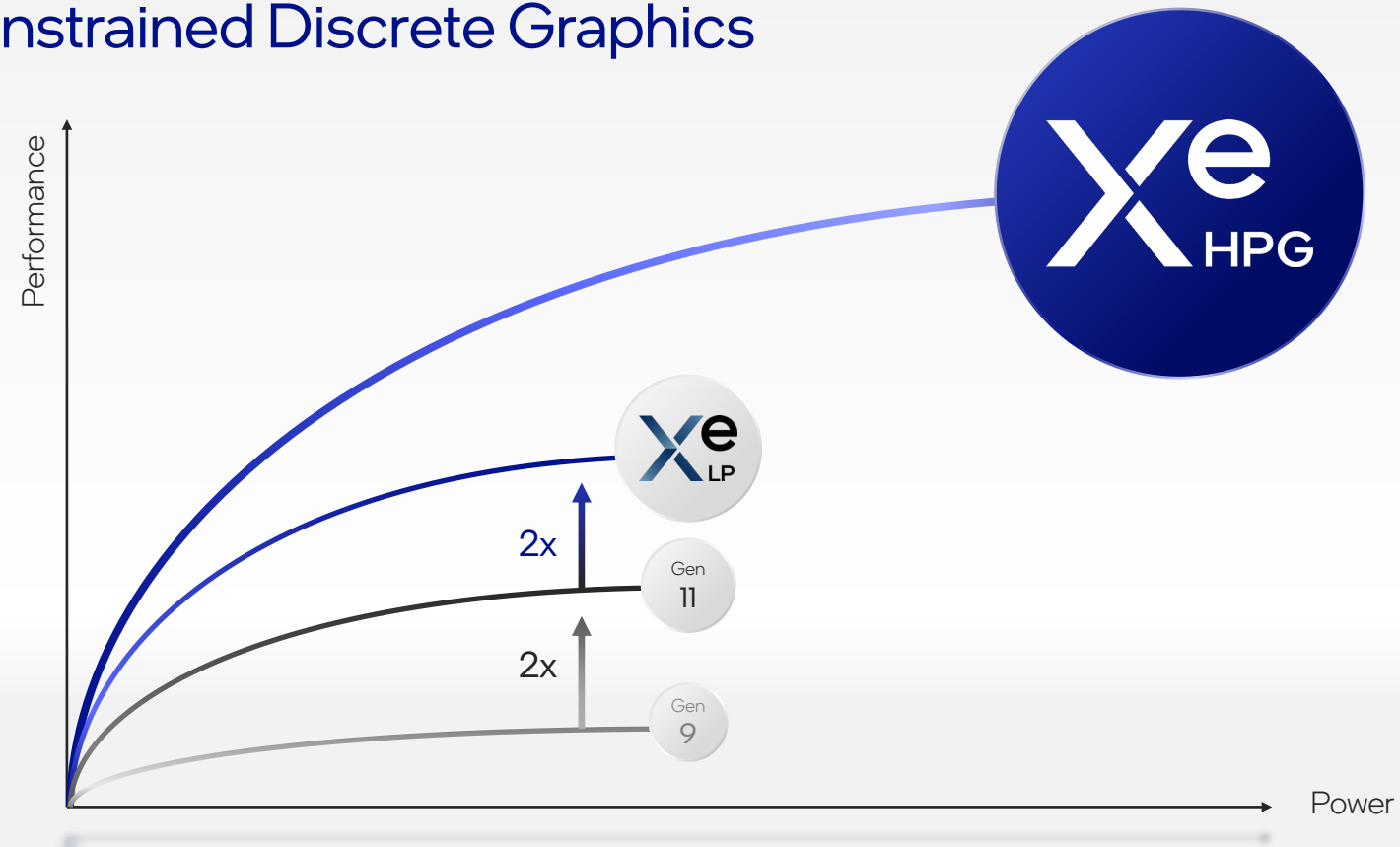
Xe HPG architecture

Leadership Integrated Graphics



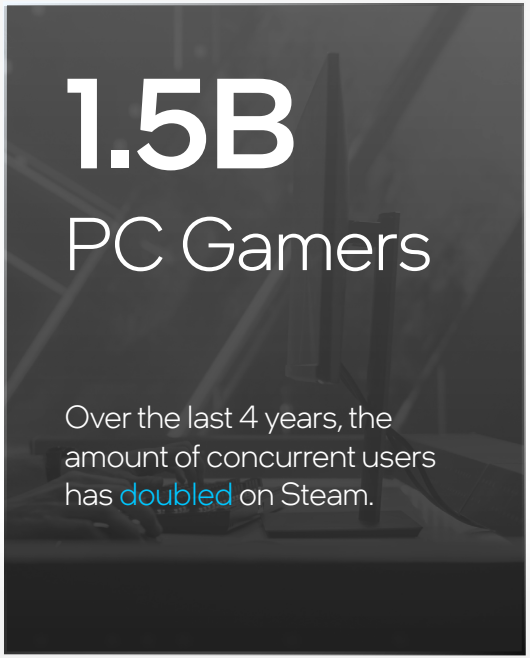
For workloads and configurations visit www.intel.com/ArchDay21claims. Results may vary.

Unconstrained Discrete Graphics



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

Vivid PC Graphics Market

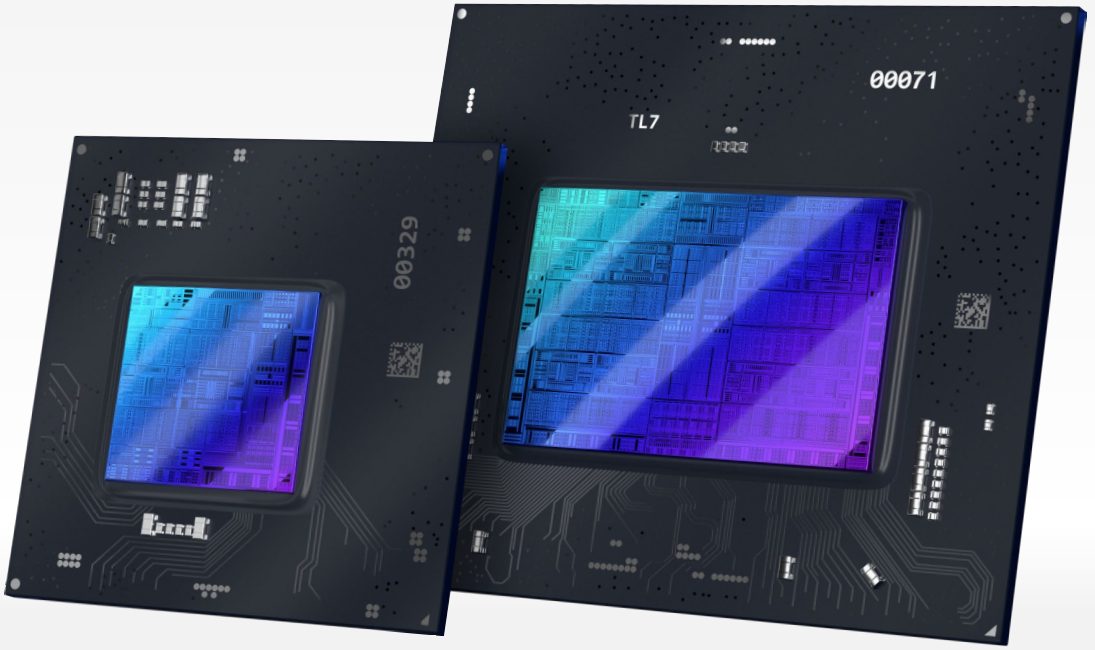


1. Source: https://www.pcgamesn.com/pc-gaming-study
2. Source: https://blog.streamlabs.com/streamlabs-stream-hatchet-q1-2021-live-streaming-industry-report-eaba2143f492
3. Source: Part 1 : Game Developer Population Forecast 2020, April 2020, SlashData

intel® ARC™

Powered by

Alchemist SoC

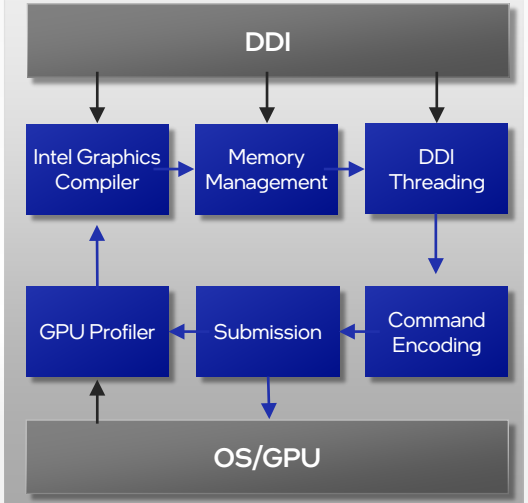


Xe HPG Sneak Peek

Lisa Pearce

Software First

Core Driver



APIs and Engines

UNREAL ENGINE

DirectX XII ULTIMATE

Vulkan™

unity

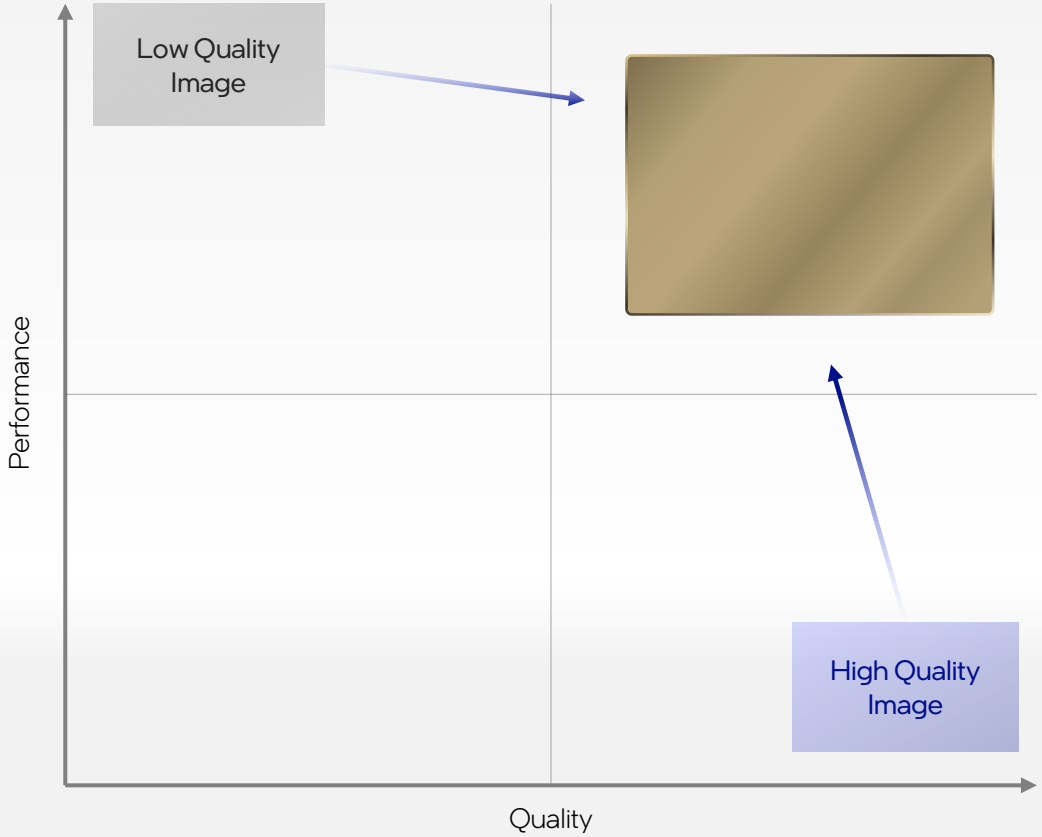
Experiences

- Smooth Gaming
- Live Game Streaming
- Modern User Interface
- Performance Tuning

Render Quality Trade-off

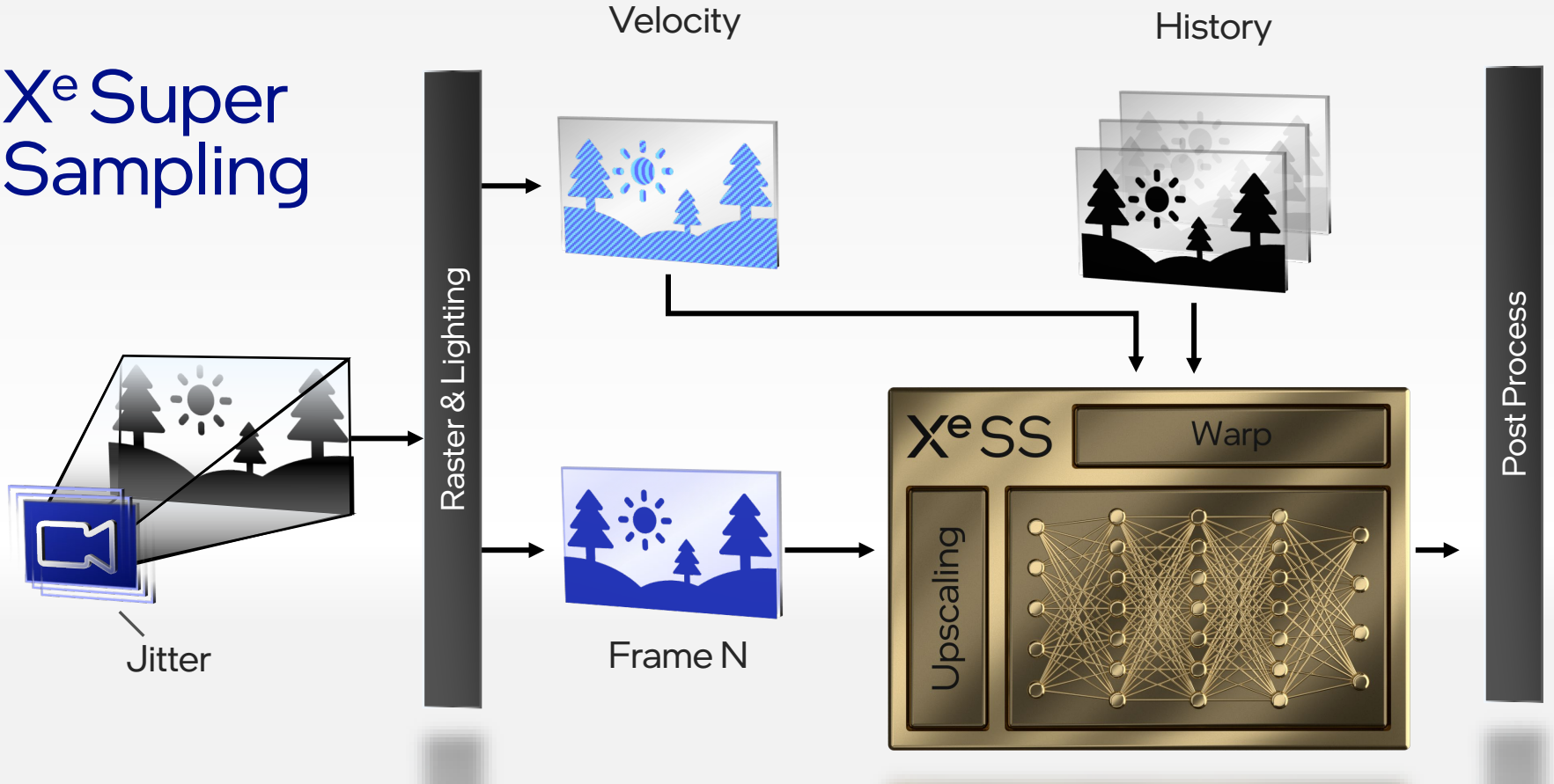
Render Quality Trade-off

Render Quality Trade-off

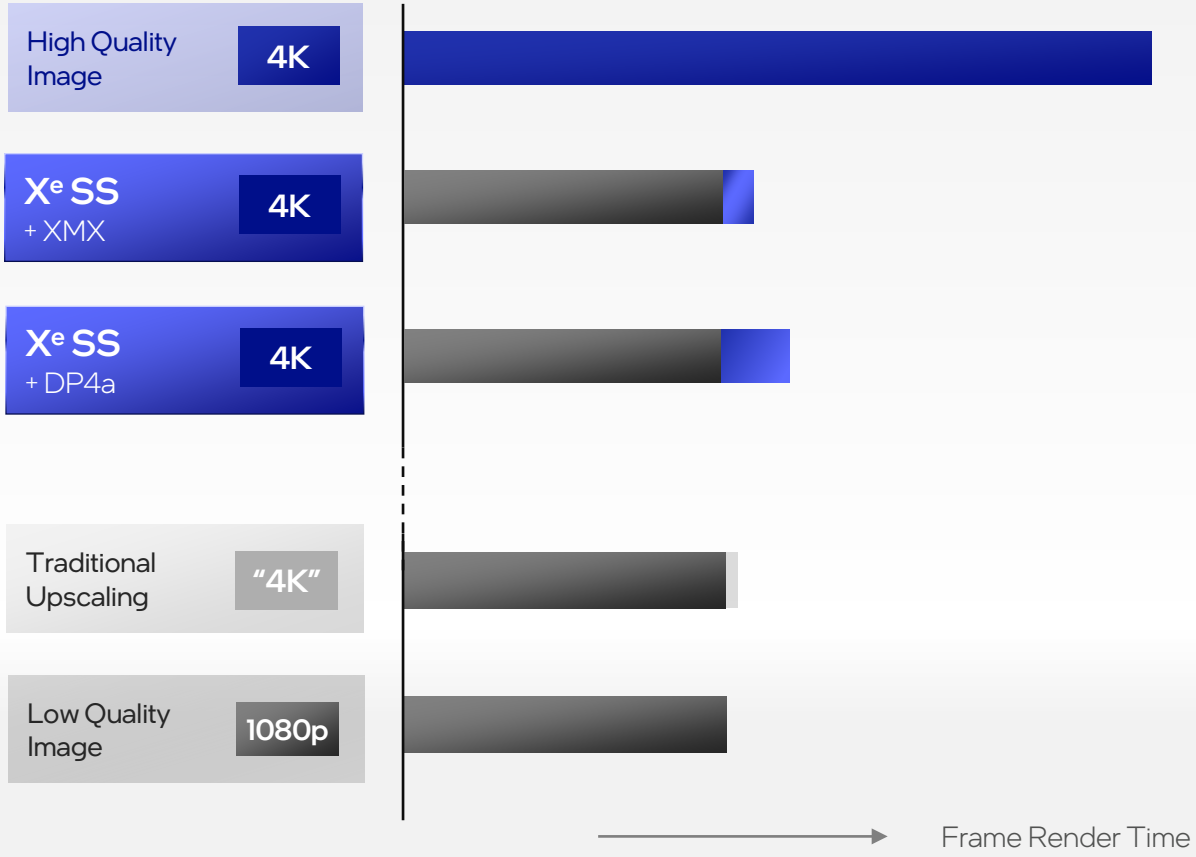


Render Quality Trade-off

Xe Super Sampling



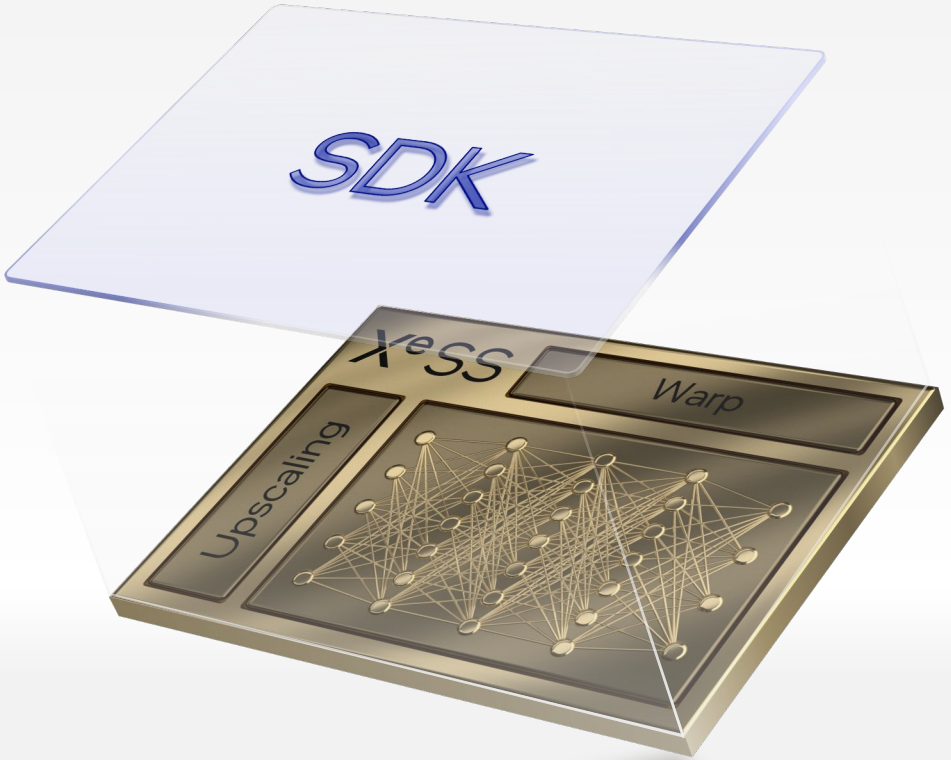
Xe SS Hits the Sweet Spot



Graph is for conceptual illustration purposes only. Subject to revision with further testing.

XeSS SDK

Available this month



Xe HPG Sneak Peek

David Blythe


Xe HPC
Xe HP
Xe LP
Compute Efficiency
Scalability
Graphics Efficiency
Xe HPG
High Performance
Gaming Optimized

Xe-core

Compute Building Block of Xe HPG-based GPUs

16
Vector Engines

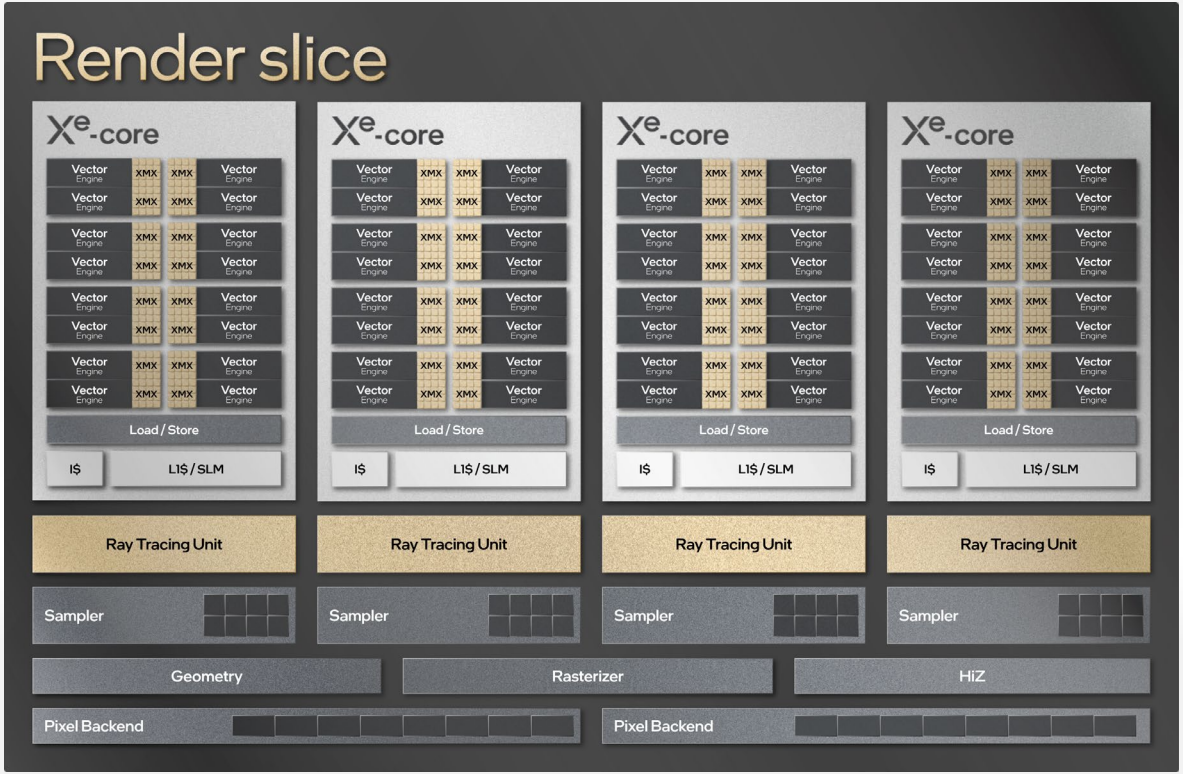
256 bit
per engine

16
Matrix Engines

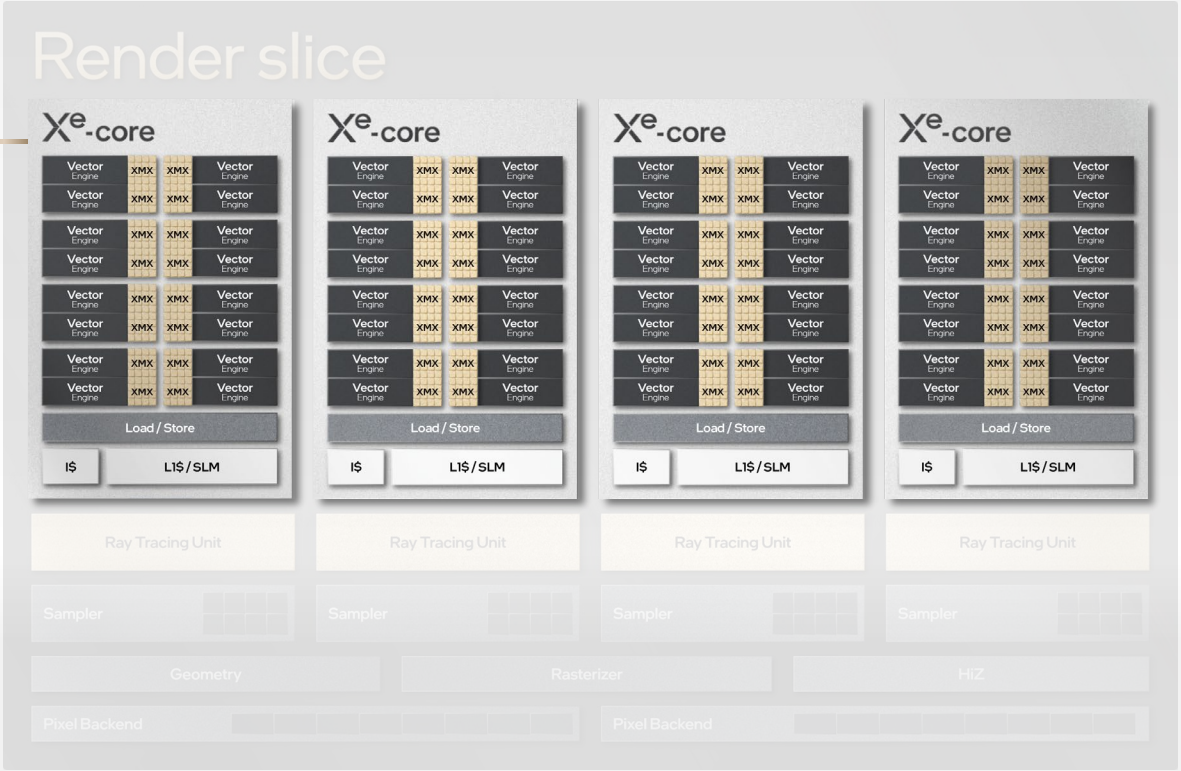
1024 bit
per engine

Xe HPG

Render Slice

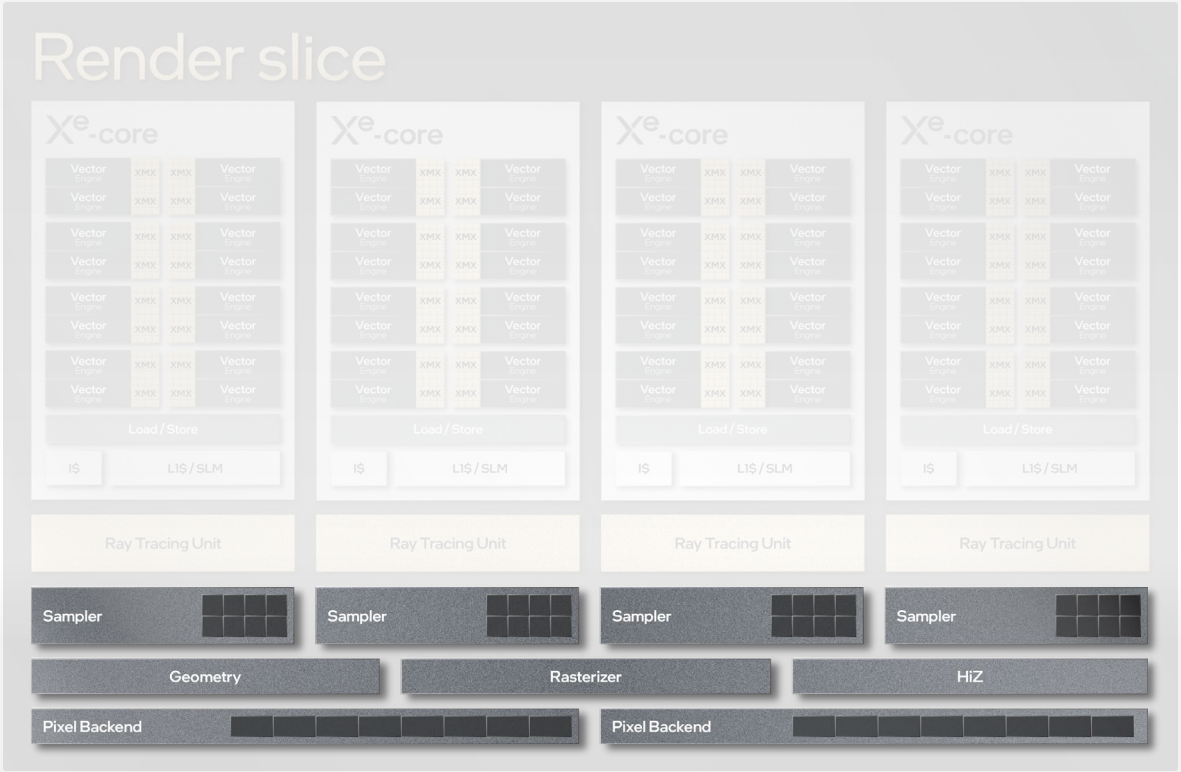


4 Xe-cores with XMX



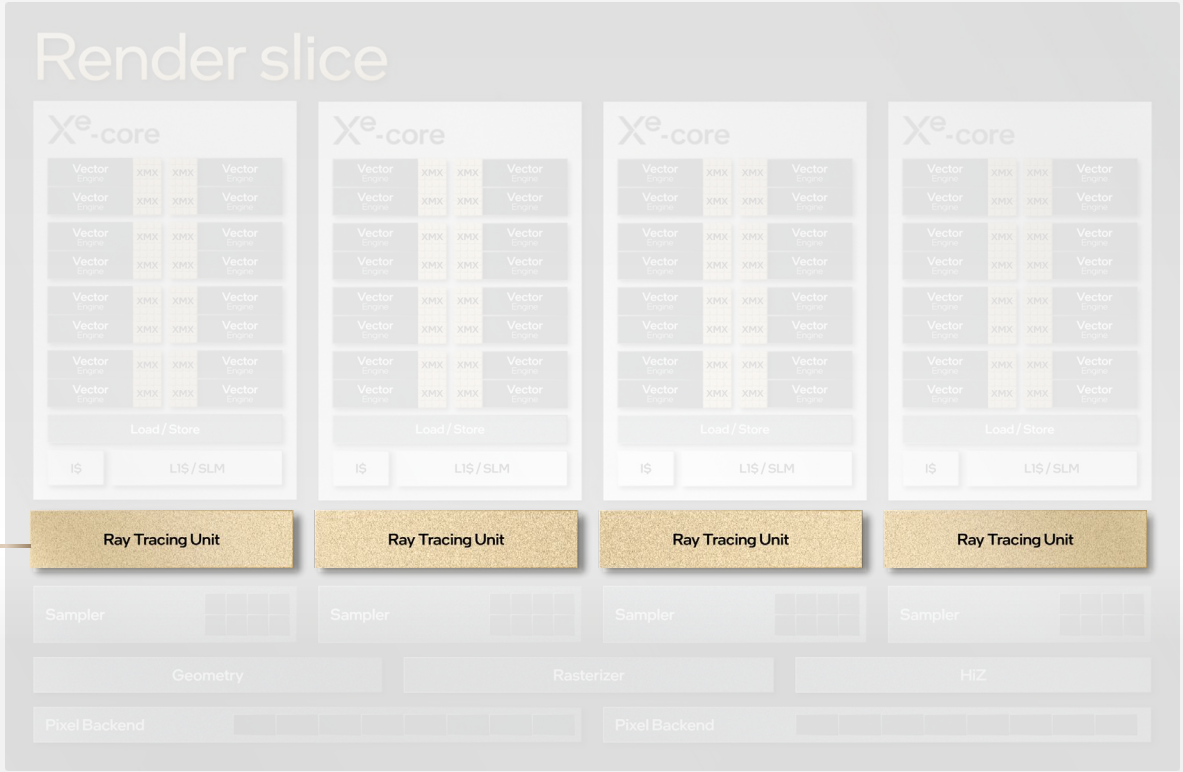
Fixed Function optimized for DX12 Ultimate Gaming

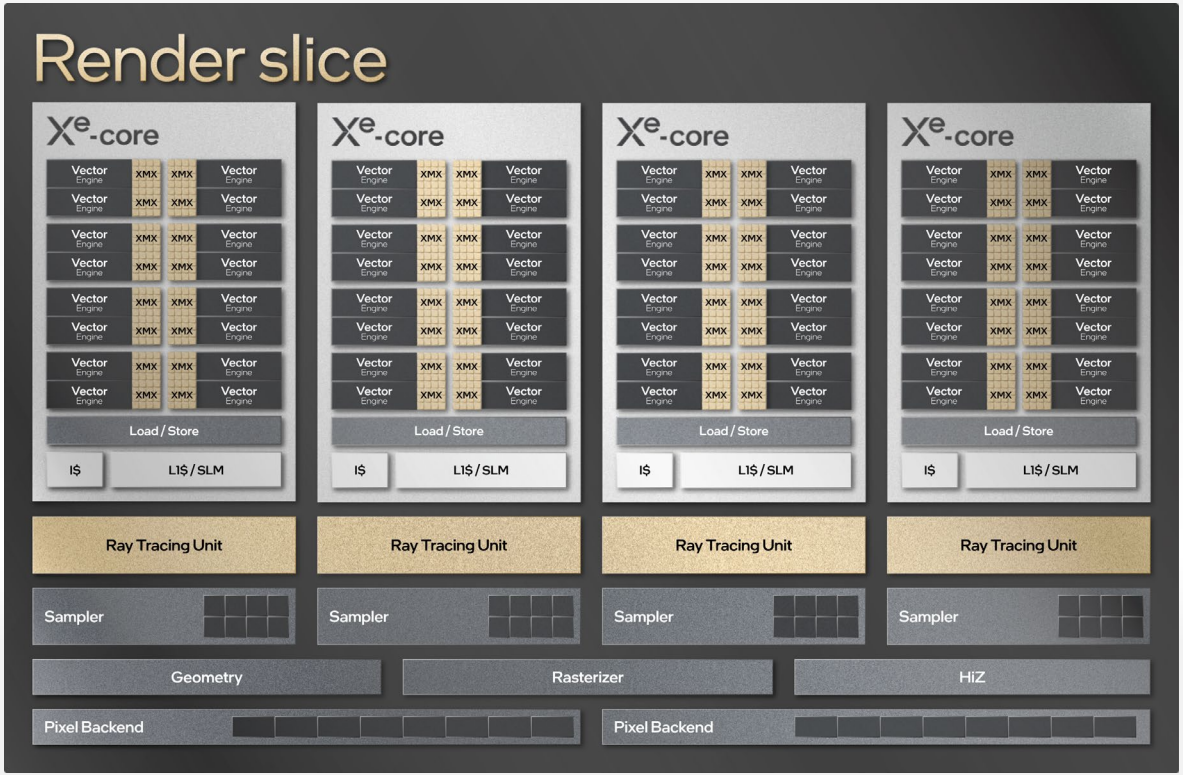
- Geometry Pipeline
- Rasterization Pipeline
- Samplers
- Pixel Backends



4 Ray Tracing Units

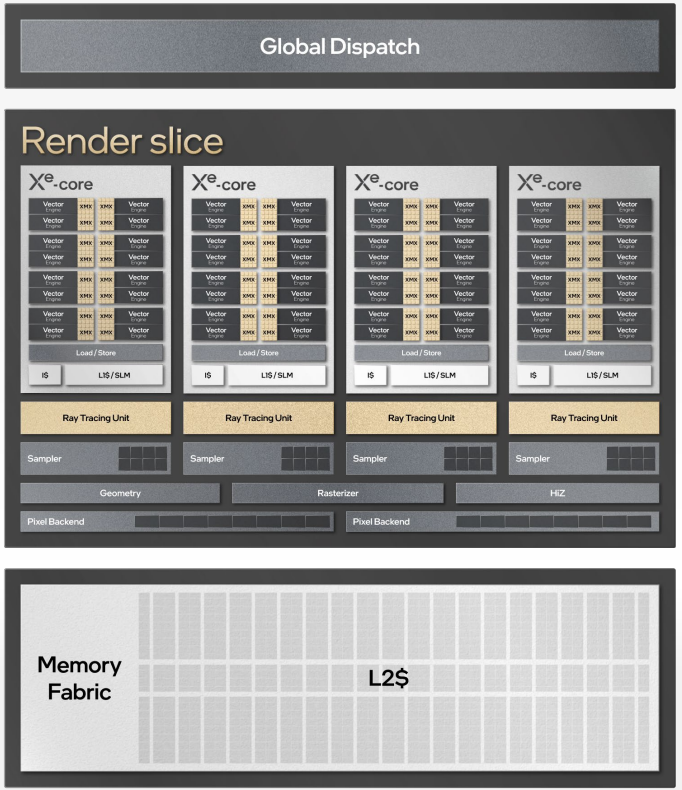
- Ray Traversal
- Bounding Box Intersection
- Triangle Intersection




Render slice
Xe-core
Vector Engine
XMX
Load / Store
I$
L1$ / SLM
Ray Tracing Unit
Sampler
Geometry
Rasterizer
HiZ
Pixel Backend

Xe HPG

Scaling the Graphics Engine

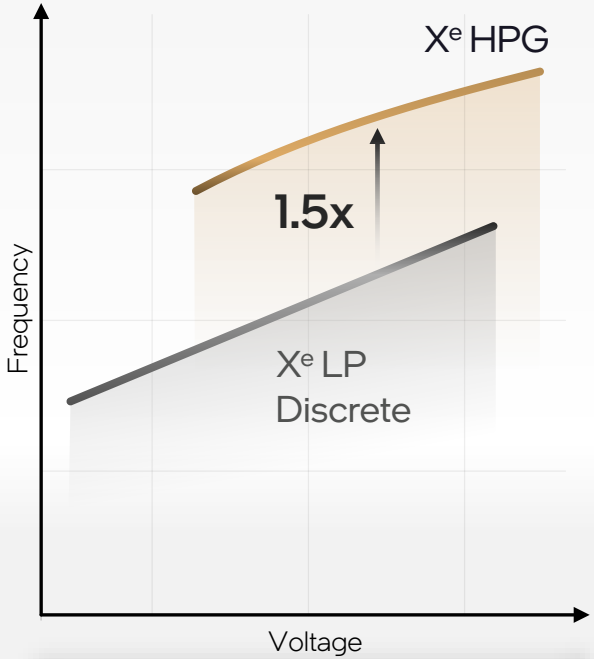


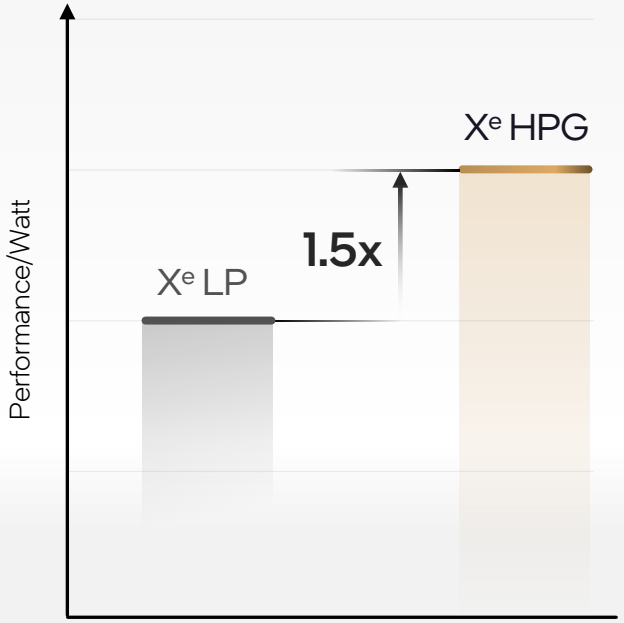
Xe HPG

Scaling the Graphics Engine

Leadership IP Performance/Watt

- Architecture
- Logic Design
- Circuit Design
- Process Technology
- Software





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

"In the world of graphics, there is an insatiable demand for better performance and more realism. TSMC is excited that **Intel has chosen our N6 technology for their Alchemist family of discrete graphics solutions".**

"There are many ingredients to a successful graphics product including the semiconductor technology. With N6, TSMC provides an optimal balance of performance, density and power efficiency that are ideal for modern GPUs. We are pleased with the **collaboration with Intel on the Alchemist family of discrete GPUs".**

Dr. Kevin Zhang,
Senior Vice President of Business Development at TSMC

tsmc

N6

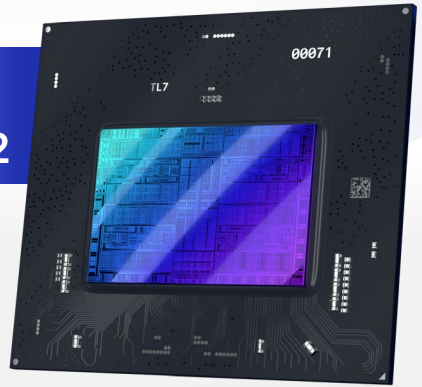
Multi-Year **Roadmap**

Performance

Alchemist

Xe HPG

Q1 2022



Battlemage

Xe2 HPG

Celestial

Xe3 HPG

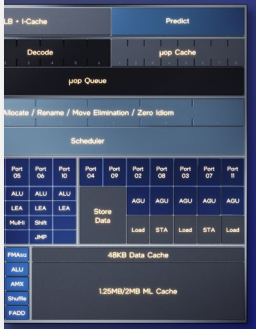
Druid

Xe Next Architecture

intel® ARC™

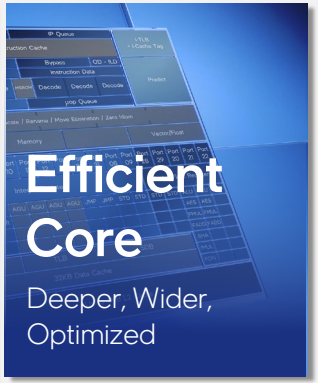
Architecture Day 2021

New Architectural Foundations



Performance Core

Biggest Shift in x86 yet

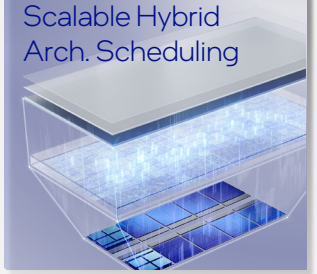


Efficient Core

Deeper, Wider, Optimized

Intel Thread Director

Scalable Hybrid Arch. Scheduling



Xe - core

Foundational Building Block for Xe With Xe Matrix Extensions

Sapphire Rapids

Xe HPC & Ponte Vecchio

AMX

Advanced Matrix Extension - Engine

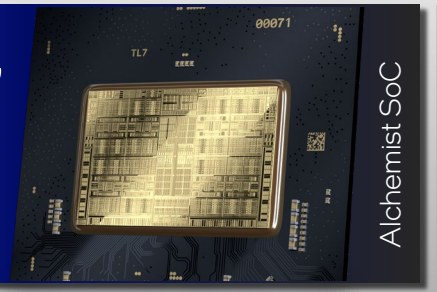
Alder Lake

Performance Hybrid

Xe SS

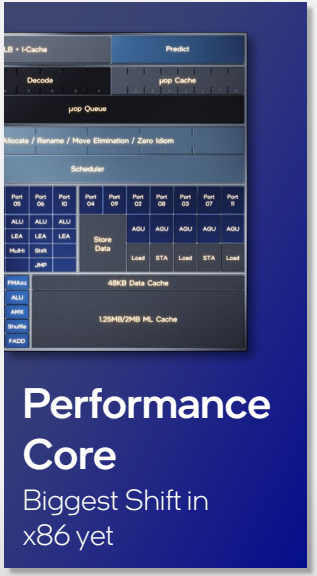
Xe HPG

Gaming & Creation First Architecture



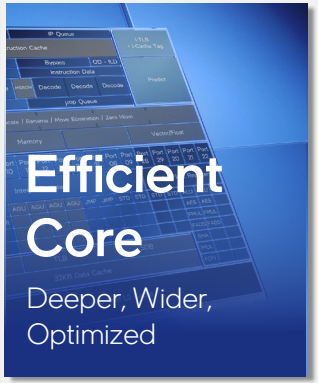
Mount Evans

Architecture Day 2021 Part 1 Recap



Performance Core

Biggest Shift in x86 yet

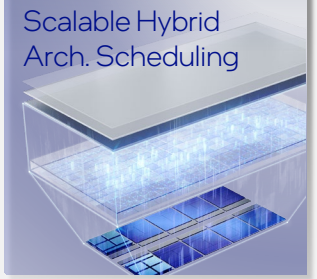


Efficient Core

Deeper, Wider, Optimized

Intel Thread Director

Scalable Hybrid Arch. Scheduling



Xe - core

Foundational Building Block for Xe With Xe Matrix Extensions

Sapphire Rapids

Xe HPC & Ponte Vecchio

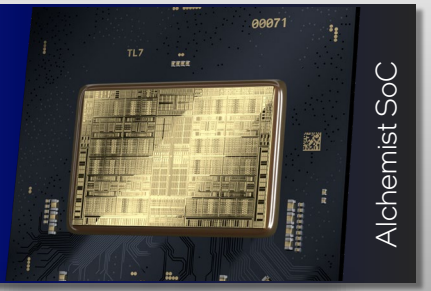
Xe SS

AMX

Advanced Matrix Extension - Engine

Alder Lake

Performance Hybrid



Xe HPG

Gaming & Creation First Architecture

Alchemist SoC

Mount Evans

Sapphire Rapids

Sailesh Kottapalli

Introducing

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
	Increased Core Counts	DDR 5	Wider & Faster UPI
		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**

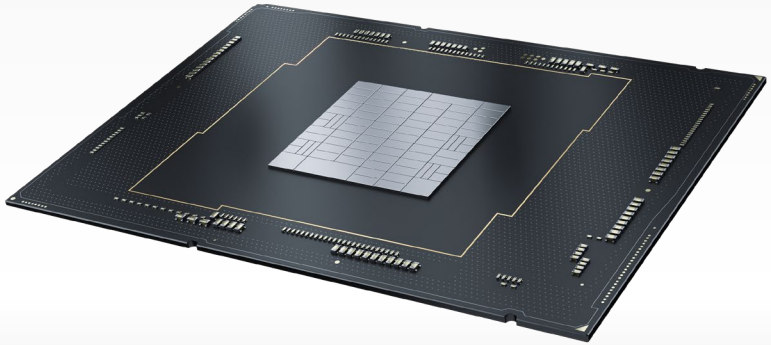
Data Center Performance

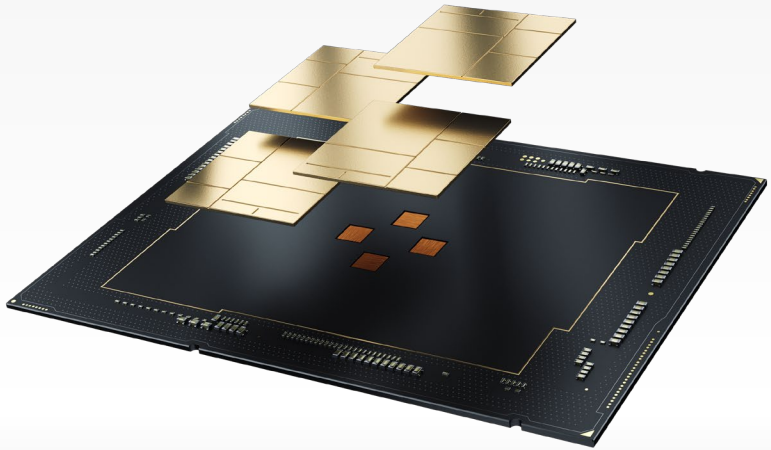
Ice Lake

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

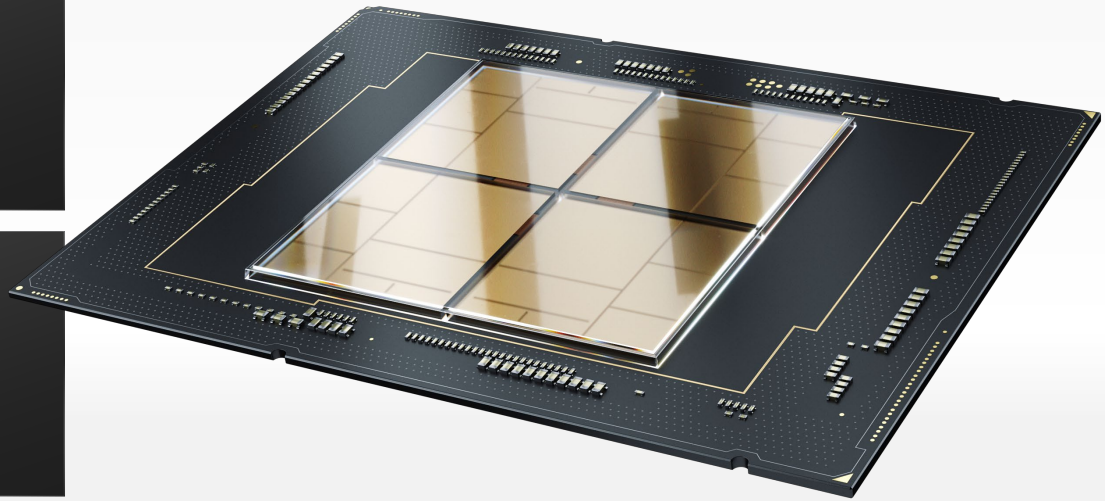
Sapphire Rapids

Multiple Tiles, Single CPU

Every thread has full access to all resources on all tiles

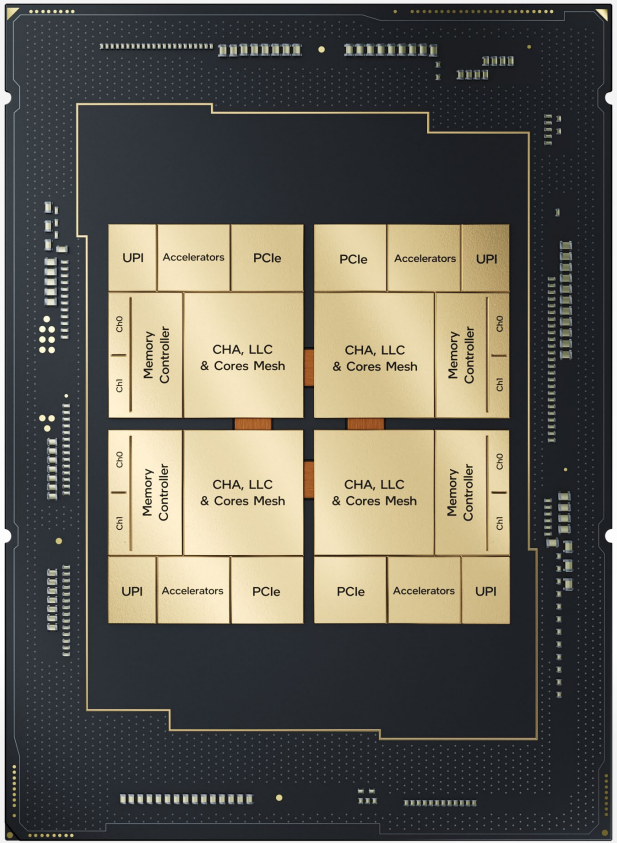
Cache, Memory, IO…

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



Sapphire Rapids

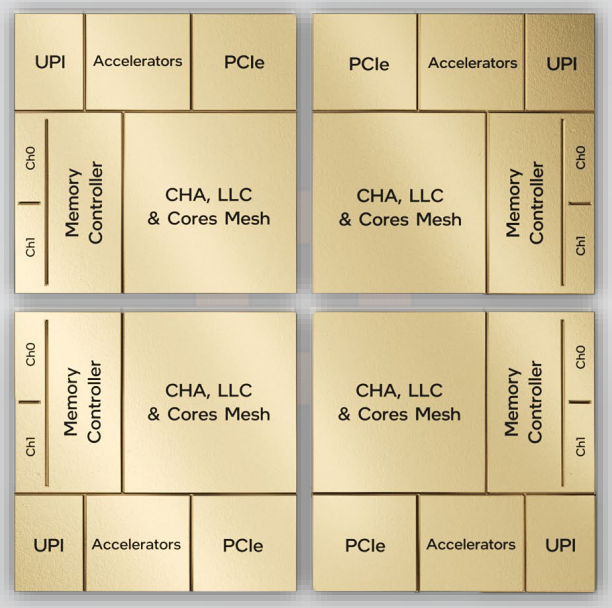
SoC



Sapphire Rapids

Key Building Blocks

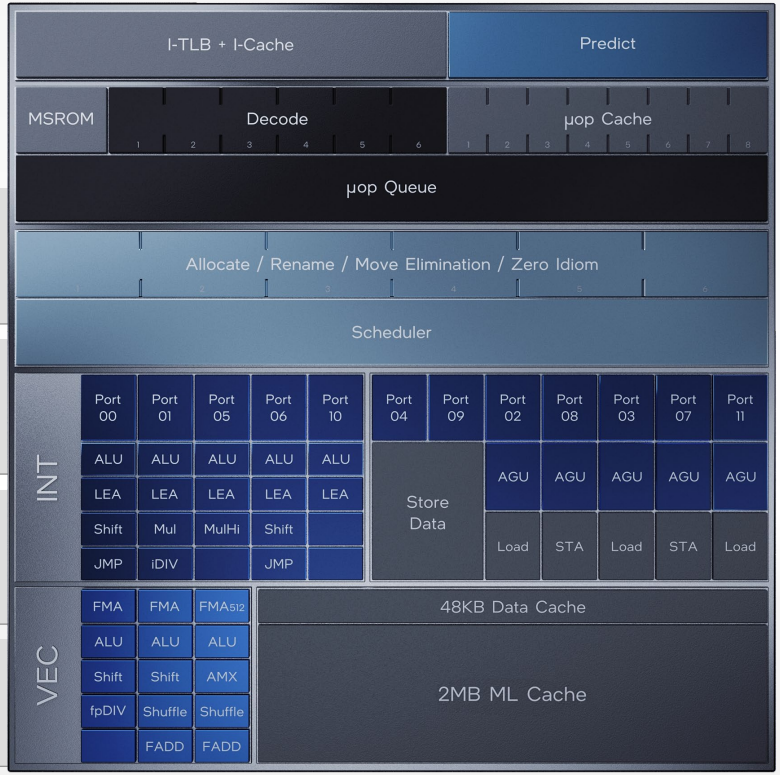
Compute IP Seamless Integration of	Cores	Acceleration Engines	
I/O IP	CXL1.1	PCIe Gen 5	UPI 2.0
Memory IP	DDR5	Optane	HBM



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
FP16	**Half- Precision** Support for higher throughput lower precision
Cache Management	**CLDEMOTE** Proactive placement of cache contents

Sapphire Rapids

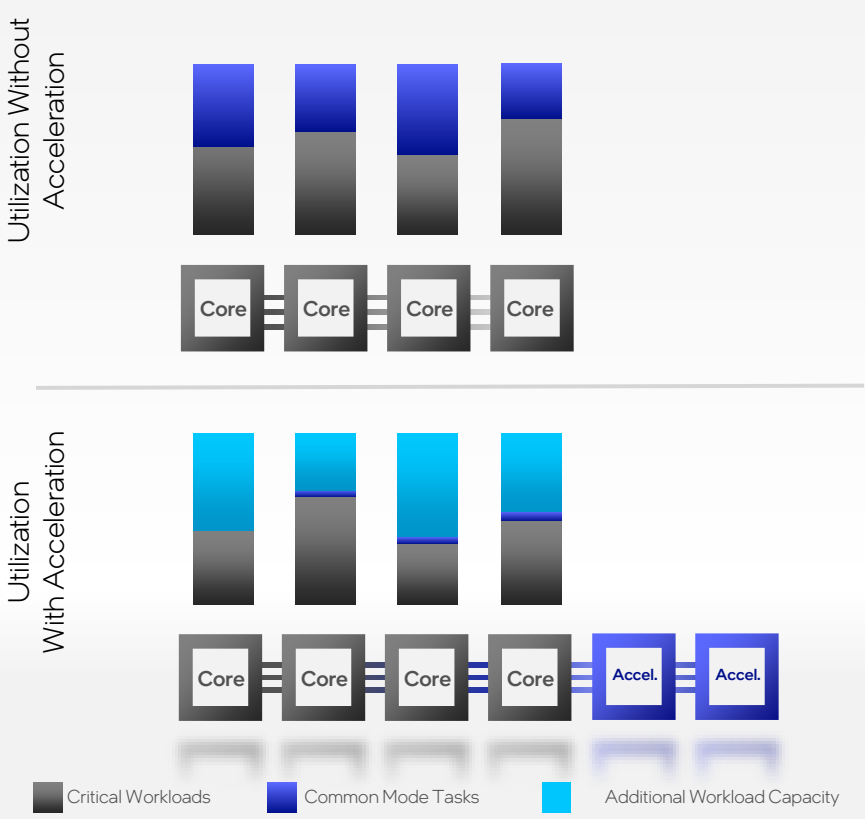
Acceleration Engines

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



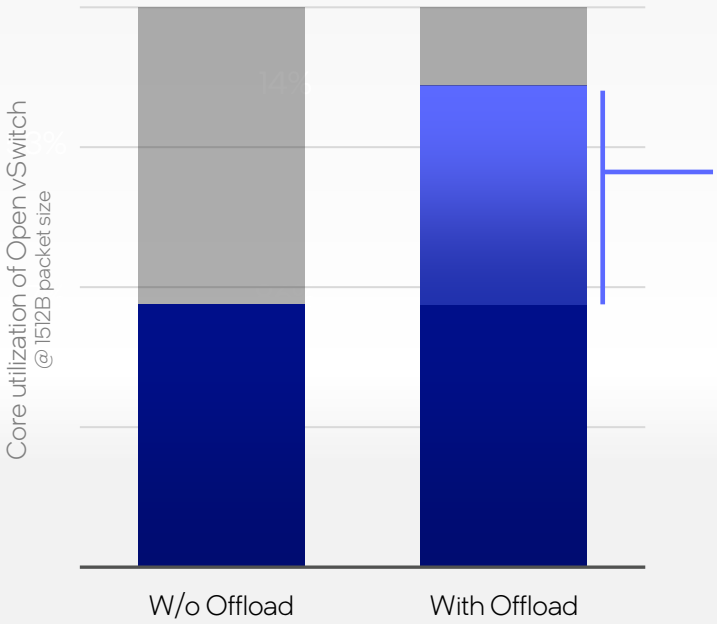
Intel® Data Streaming Acceleration Engine

Optimizing streaming data movement and transformation operations

up to
4 Instances per Socket

Low Latency Invocation

No Memory Pinning Overhead



39%

additional CPU Core cycles after DSA offload

% CPU Compute

% CPU Data Movement

Results have been estimated or simulated and based on tests with Ice Lake with Intel QAT For workloads and configurations visit www.intel.com/ArchDay21claims. Results may vary.

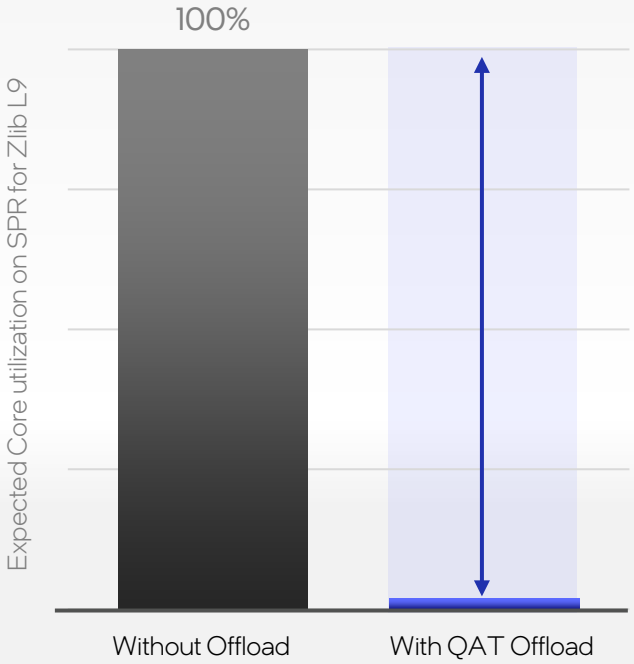
Intel® Quick Assist Technology Acceleration Engine

Accelerating Cryptography and Data De/Compression

up to
400Gb/s Symmetric Crypto

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

Fused Operations



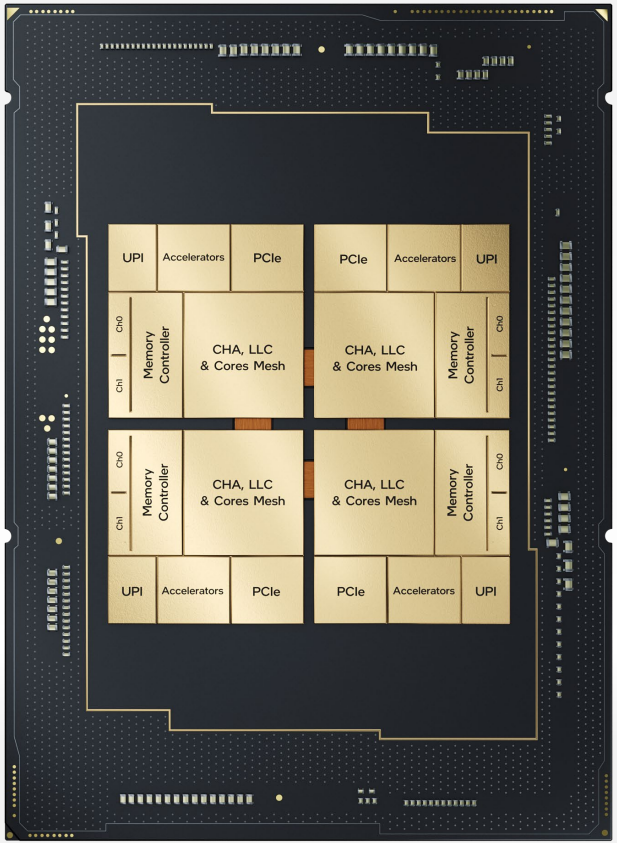
98%

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.

Sapphire Rapids

SoC



Sapphire Rapids

I/O Advancements

Introducing Compute eXpress Link (CXL) 1.1

Accelerator and memory expansion in datacenter

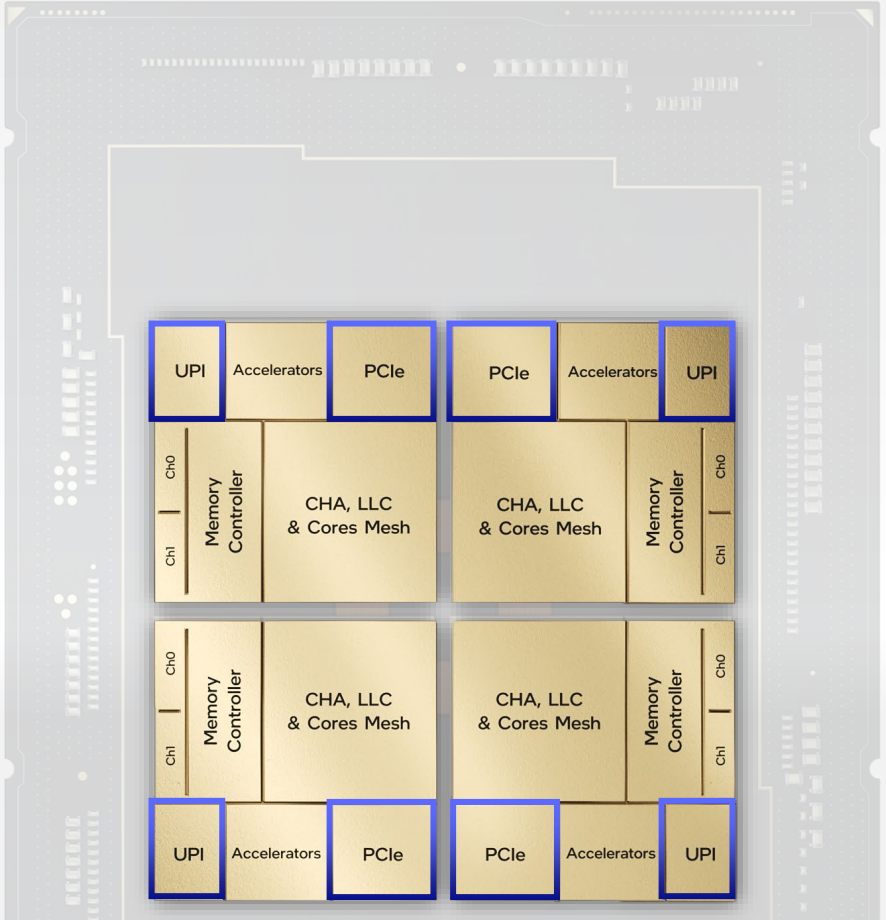
Expanded device performance via PCIe 5.0 & connectivity

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

Memory and Last Level Cache

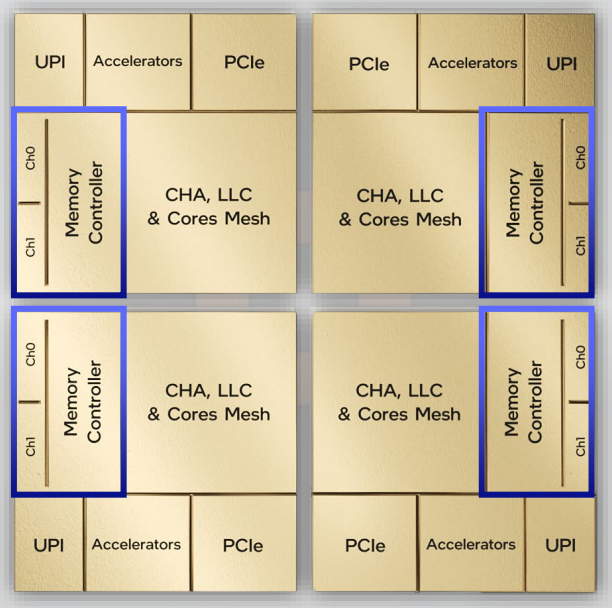
Increased Shared Last Level Cache (LLC)

Up to >100 MB LLC shared across ALL cores

Increased bandwidth, security & reliability via DDR 5 Memory

4 memory controllers supporting 8 channels

Intel® Optane™ Persistent Memory 300 Series



Sapphire Rapids

High Bandwidth Memory

Significantly Higher Memory Bandwidth

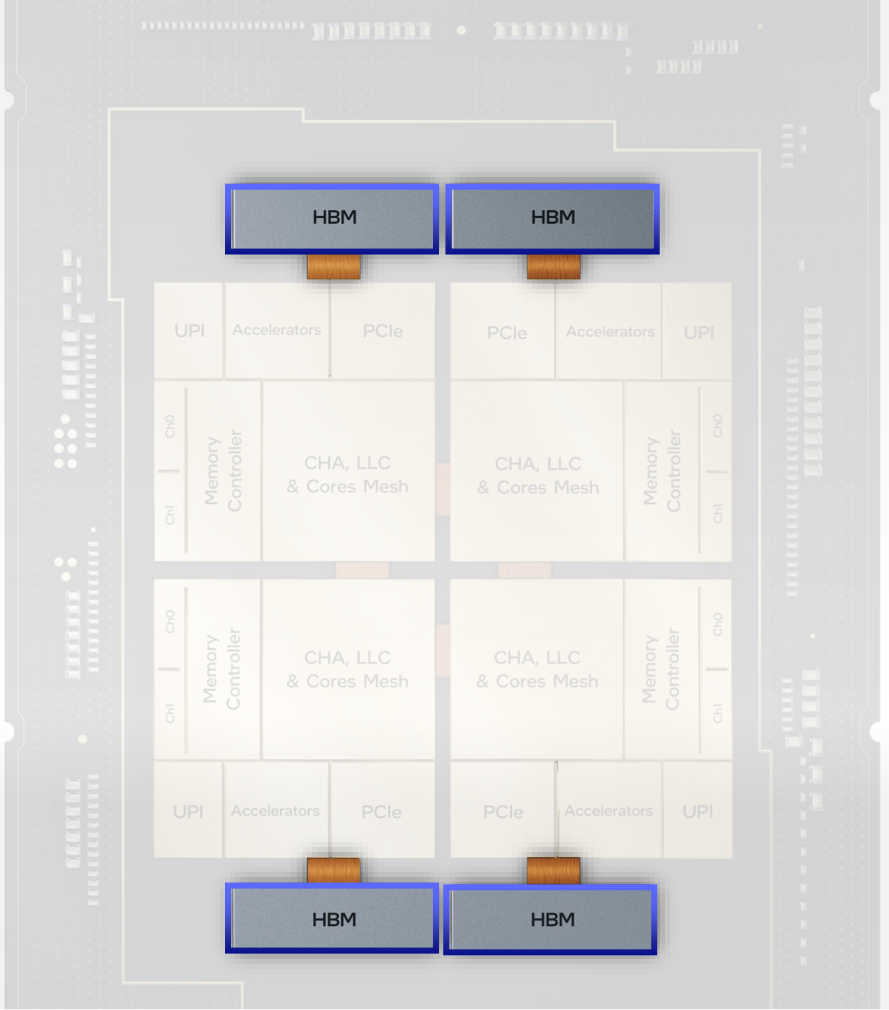
vs. baseline Xeon-SP with 8 channels of DDR 5

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



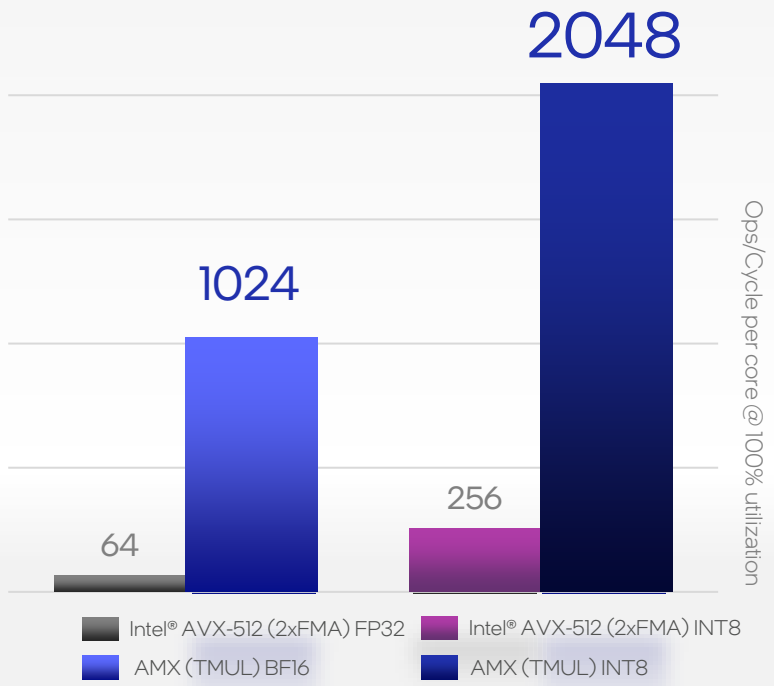
Sapphire Rapids - **Architected for AI**

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.

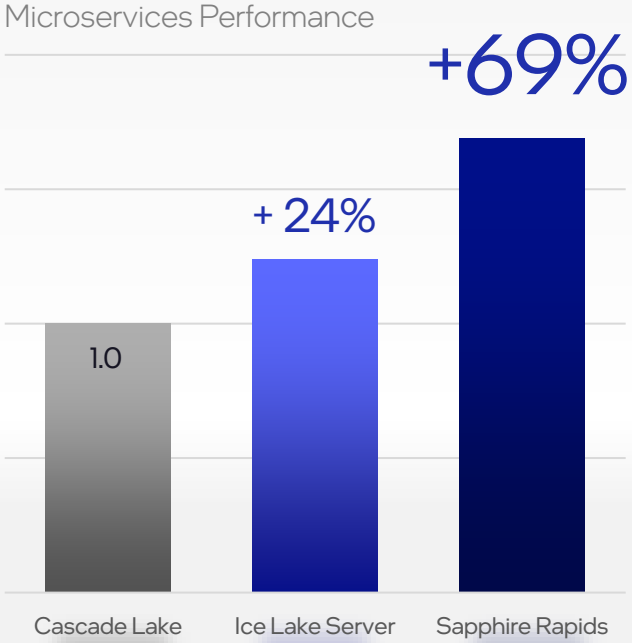
Sapphire Rapids - **Built for elastic computing models - microservices**

>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

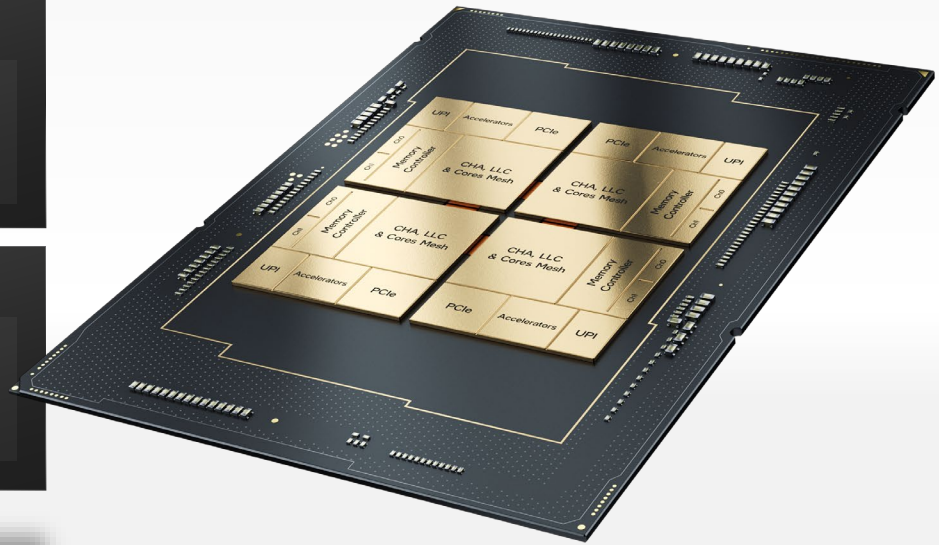
- Multi Tile SoC for Scalability
- Physically Tiled, Logically Monolithic
- General Purpose & Dedicated Acceleration Engines

Designed for Microservices and AI Workloads

- Performance Core Architecture
- Workload Specialized Acceleration

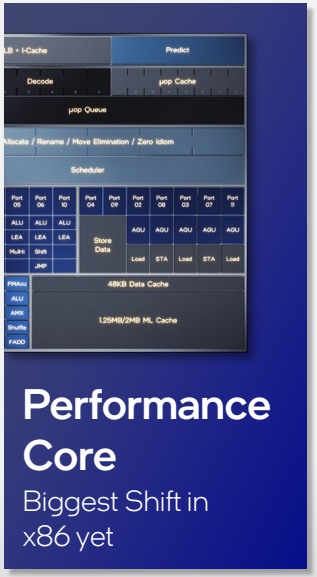
Pioneering Advanced Memory & IO Transitions

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



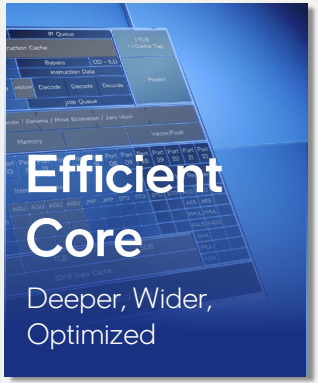
Architecture Day 2021

New Architectural Foundations



Performance Core

Biggest Shift in x86 yet

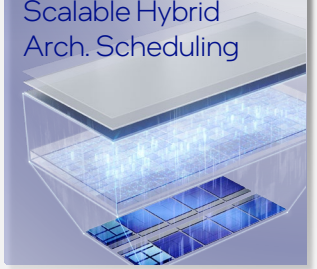


Efficient Core

Deeper, Wider, Optimized

Intel Thread Director

Scalable Hybrid Arch. Scheduling

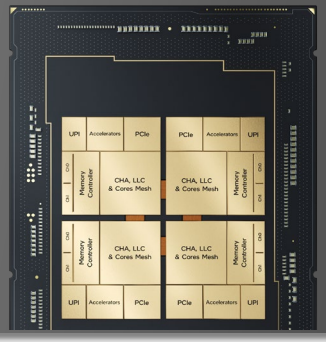


Xe - core

Foundational Building Block for Xe With Xe Matrix Extensions

Sapphire Rapids

Biggest Leap in DC Capabilities in a decade



Xe HPC & Ponte Vecchio

AMX

Advanced Matrix Extension - Engine

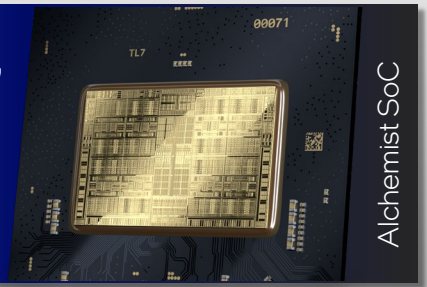
Alder Lake

Performance Hybrid

Xe SS

Xe HPG

Gaming & Creation First Architecture



Alchemist SoC

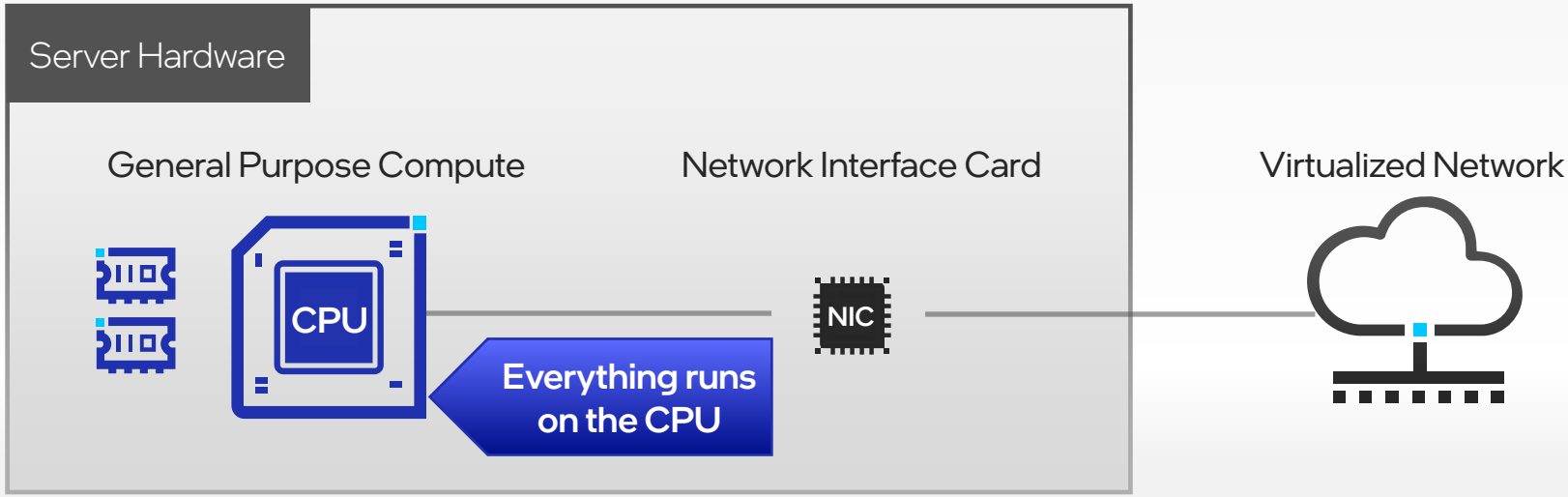
Mount Evans

Infrastructure Processing Unit

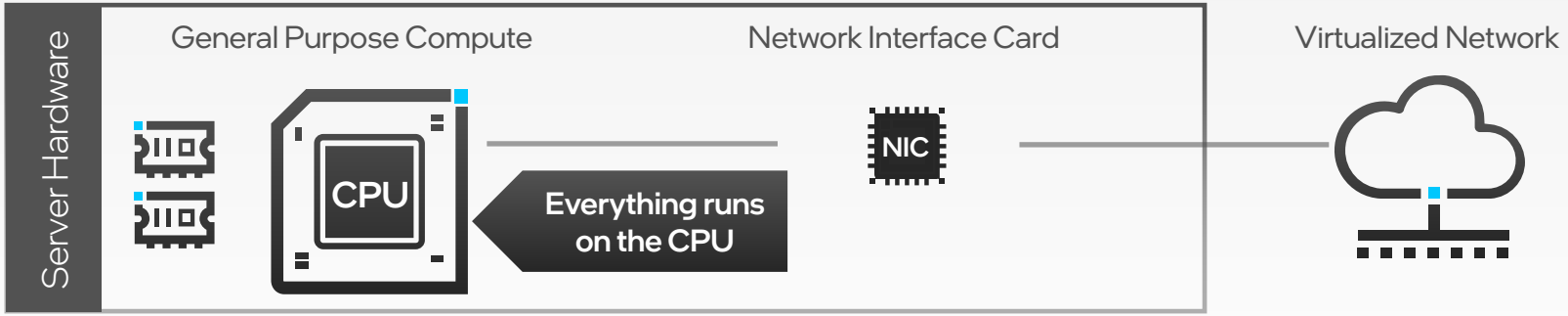
Guido Appenzeller

Server Architecture in a classic Data Center

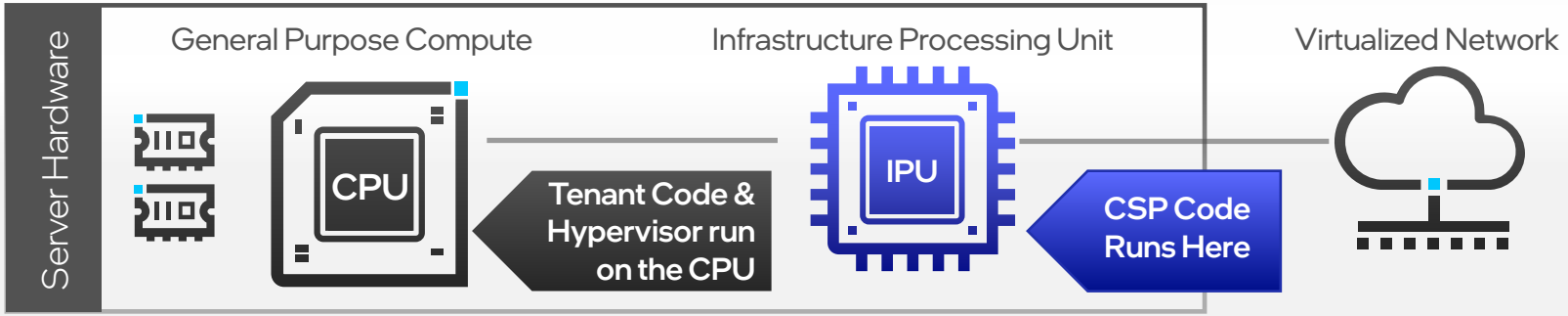
Software and Infrastructure are all controlled by One Entity

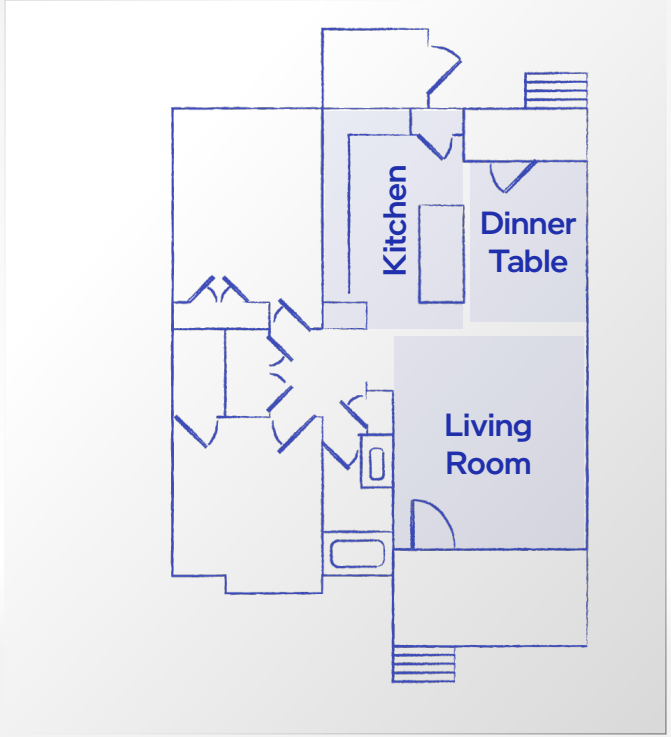


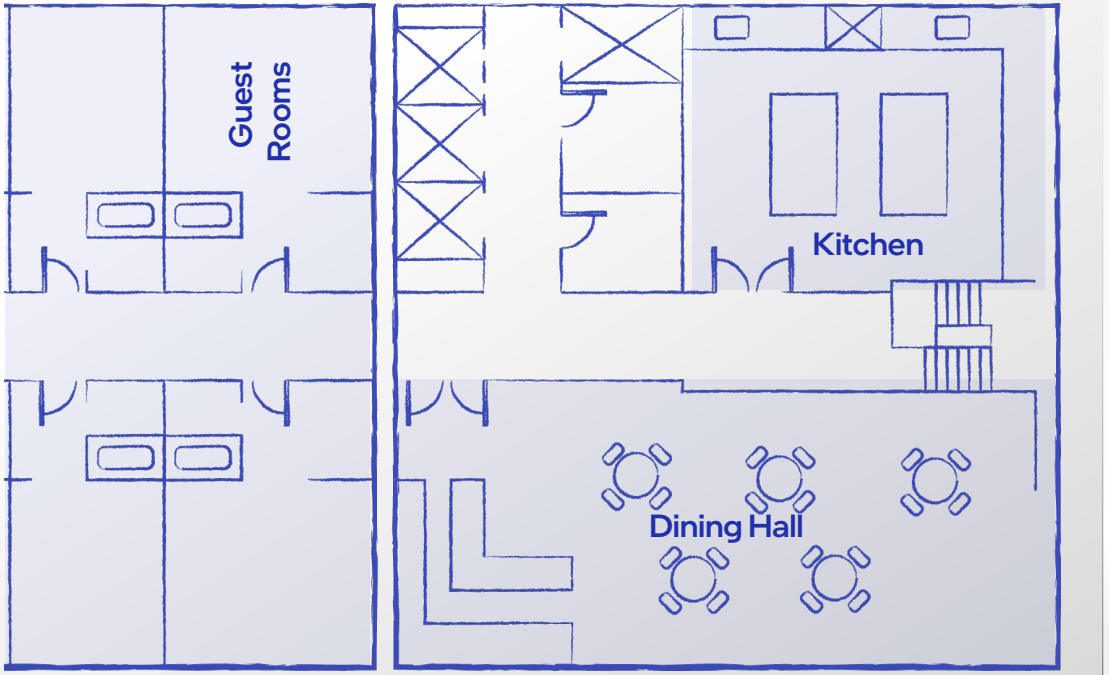
Classic Server Architecture



Cloud Server Architecture




Kitchen
Dinner Table
Living Room


Guest Rooms
Kitchen
Dining Hall

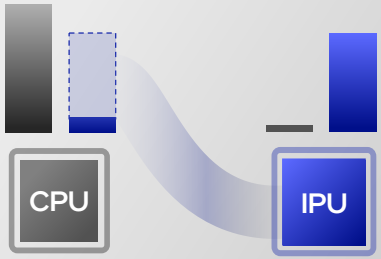
Major Advantages of IPUs

1

Separation of Infrastructure & Tenant

Guest can fully control the CPU with their SW, while CSP maintains control of the infrastructure and Root of Trust

2



Infrastructure Offload

Accelerators help process these task efficiently. Minimize latency and jitter and maximize revenue from CPU

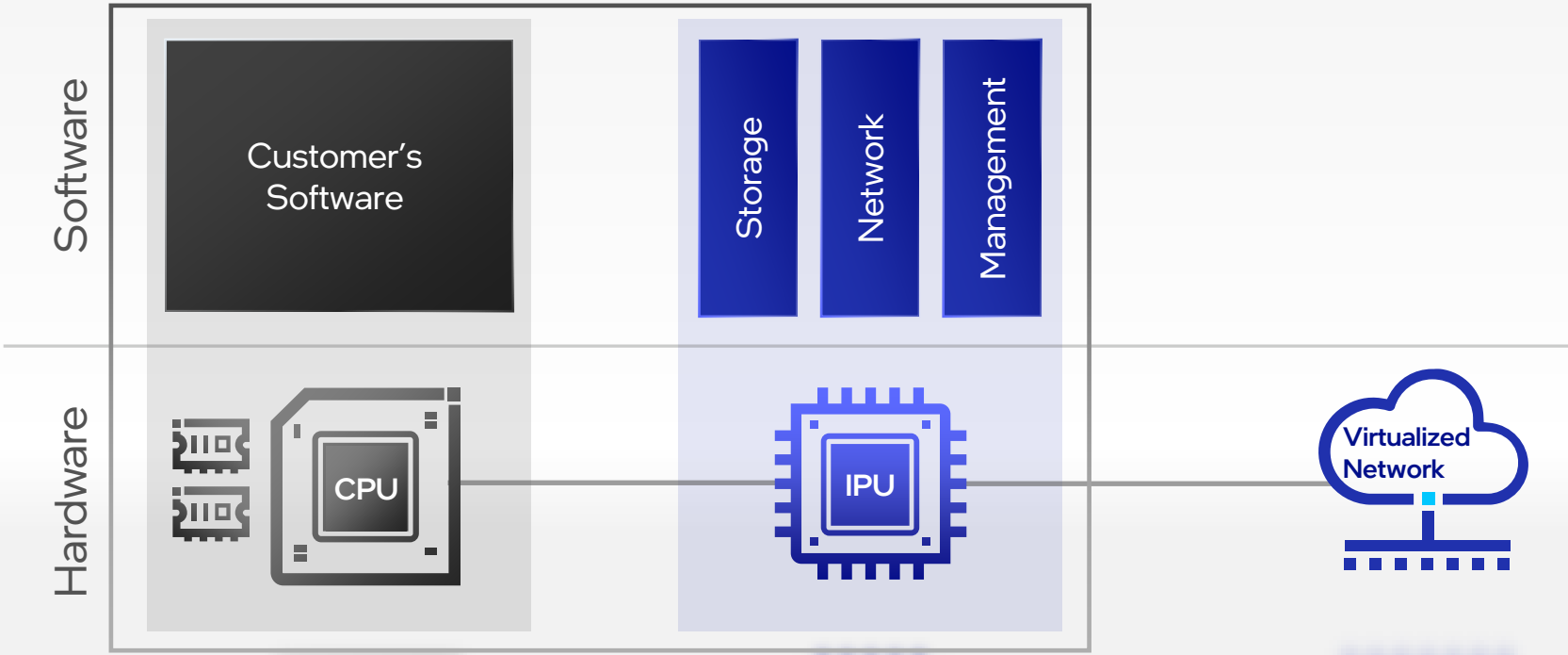
3

Diskless Server Architecture

Simplifies data center architecture while adding flexibility for the CSP

Advantage 1 - Separation of Infrastructure and Tenant

Maximum Control and Isolation for the Tenant

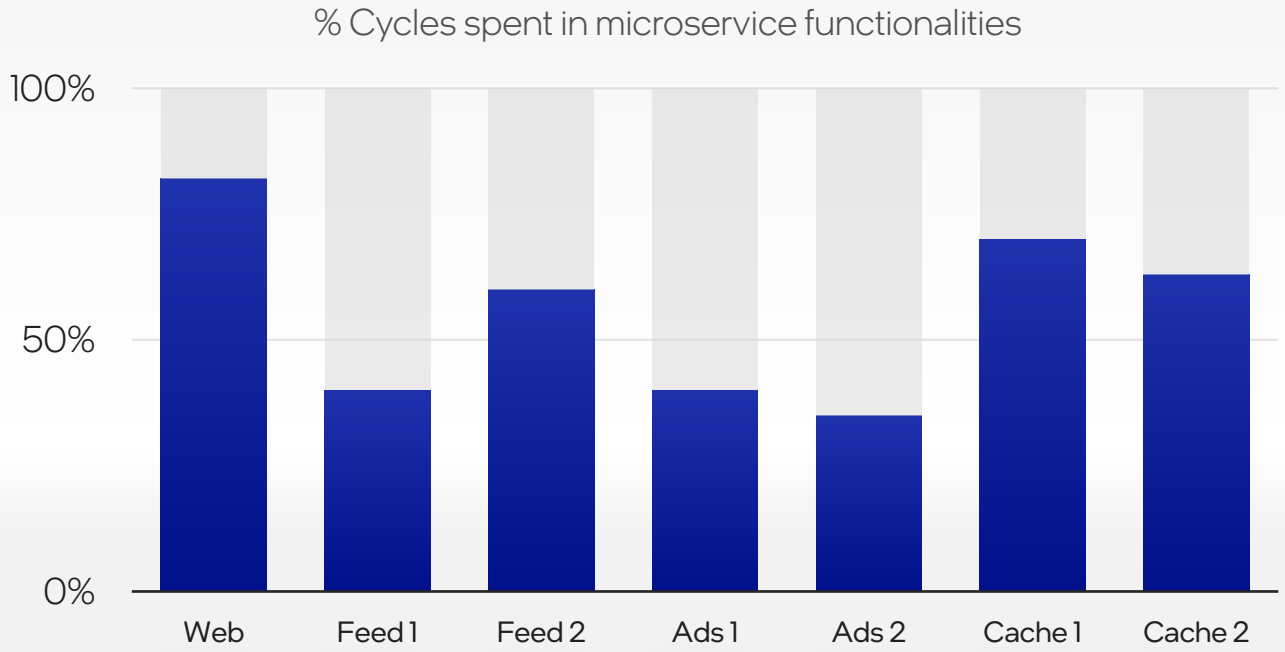


Advantage 2 - Infrastructure Offload

In some cases, the majority of CPU cycles are spent on overhead

31% to 83%

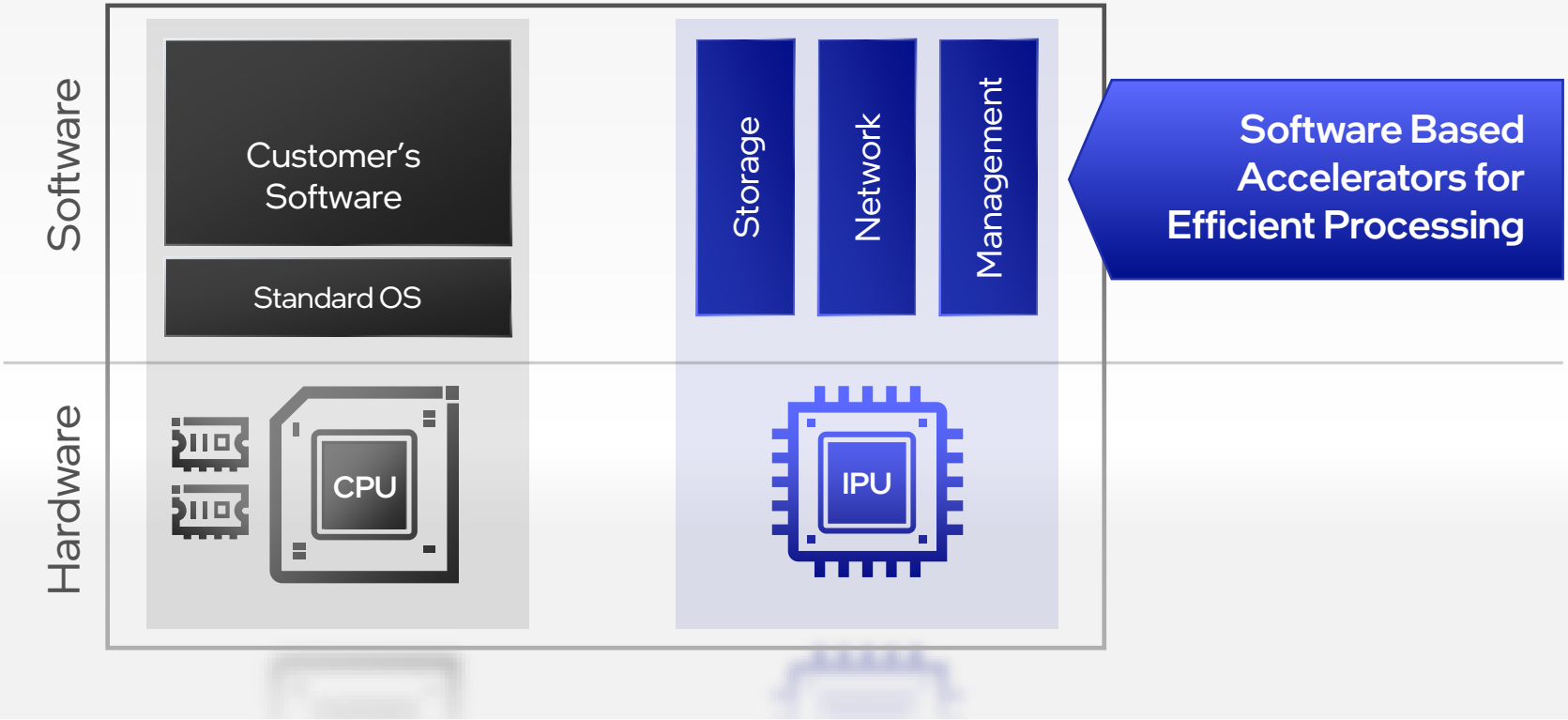
Microservice Overhead at Facebook



Source: From Accelerometer: Understanding Acceleration Opportunities for Data Center Overheads at Hyperscale. Akshitha Srirama, Abhishek Dhanotia. Facebook.

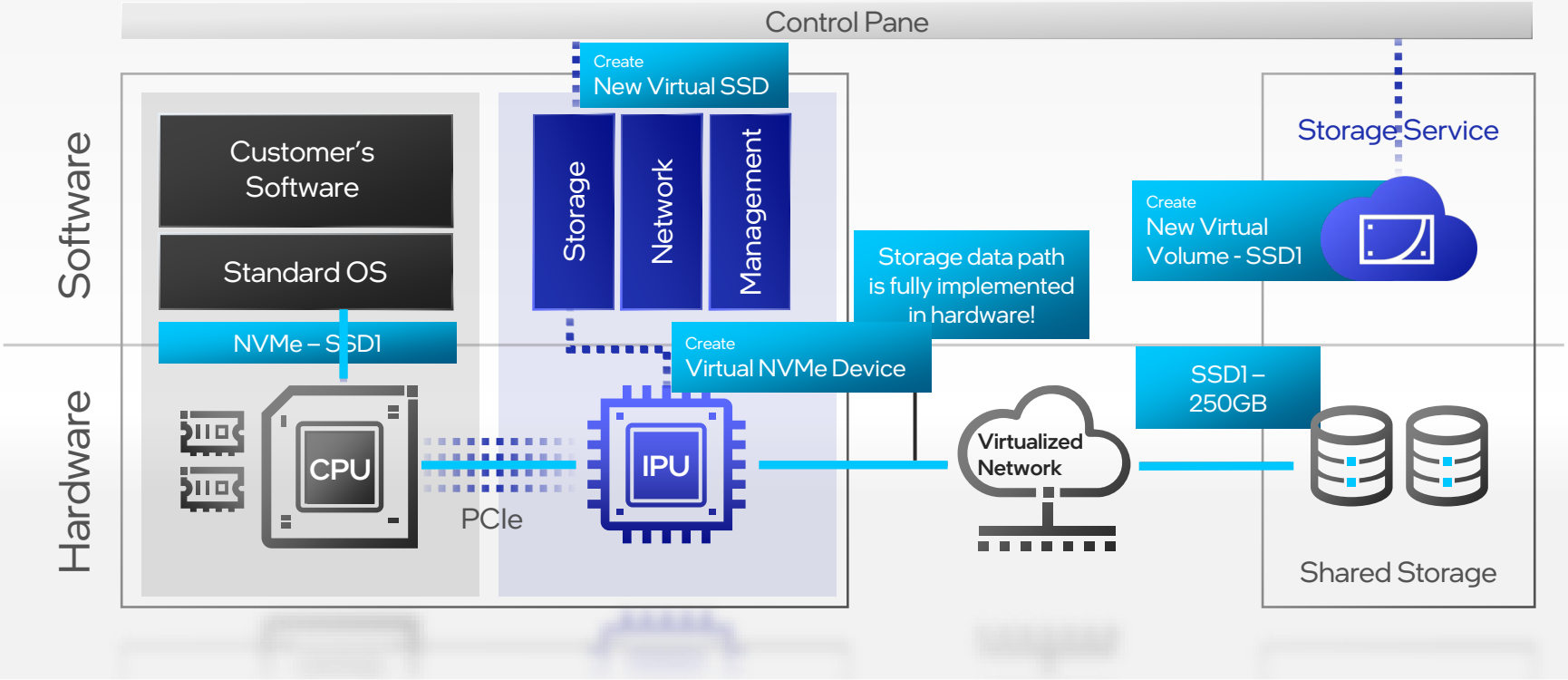
Advantage 2 - Infrastructure Offload

Dedicated Accelerators Free up CPU Capacity



Advantage 3 - Diskless Server Architecture

Scale with Virtual Storage via Network



Broad Infrastructure Acceleration Portfolio

Dedicated ASIC IPU

Performance and power optimized

Optimized secure networking and storage pipeline

FPGA-based Acceleration

IPU Platforms & Adapters

SmartNICs

Faster time to market for evolving standards

Re-programmable Secure Datapath enables flexible/customizable workload offload (future proof)

Programmable accelerated infrastructure workloads with customizable packet processing

Onboard Intel® Xeon® processor

Intel Ethernet NIC with DPDK support

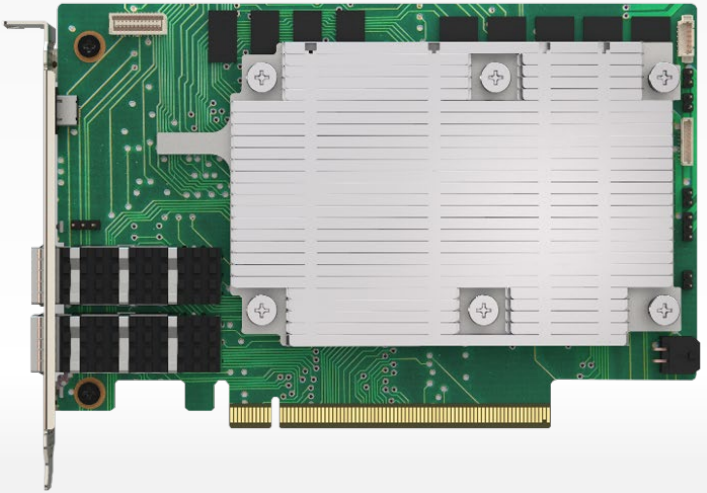
Note: Future Intel IPUs may integrate both ASIC and FPGA

Introducing

Oak Springs Canyon

High perf networking and storage acceleration for Cloud Service Providers

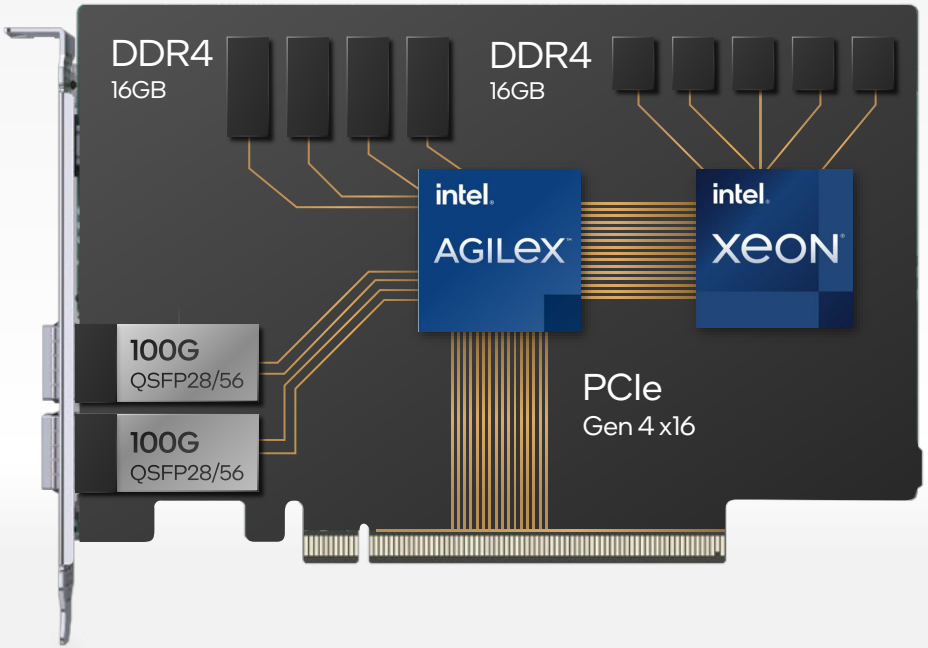
- OVS, NVMe over Fabric, and RoCE solutions
- Programmable through Intel OFS, DPDK, and SPDK
- Customizable solutions with FPGA



Oak Springs Canyon

Built with Intel® Agilex FPGA and Xeon-D SoC

- High speed Ethernet support - 2x100G
- PCIe Gen 4 x16
- Hardware crypto block enables security at line rate

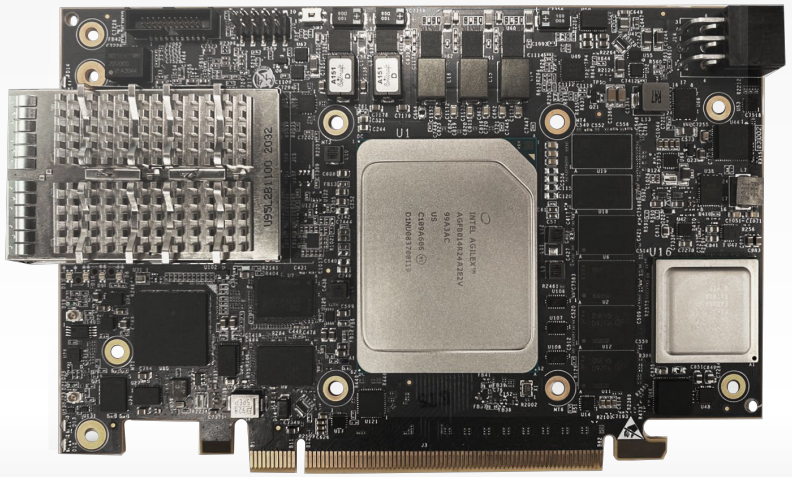


Introducing

Arrow Creek

Acceleration Development Platform (ADP) for High Performance 100G networking acceleration

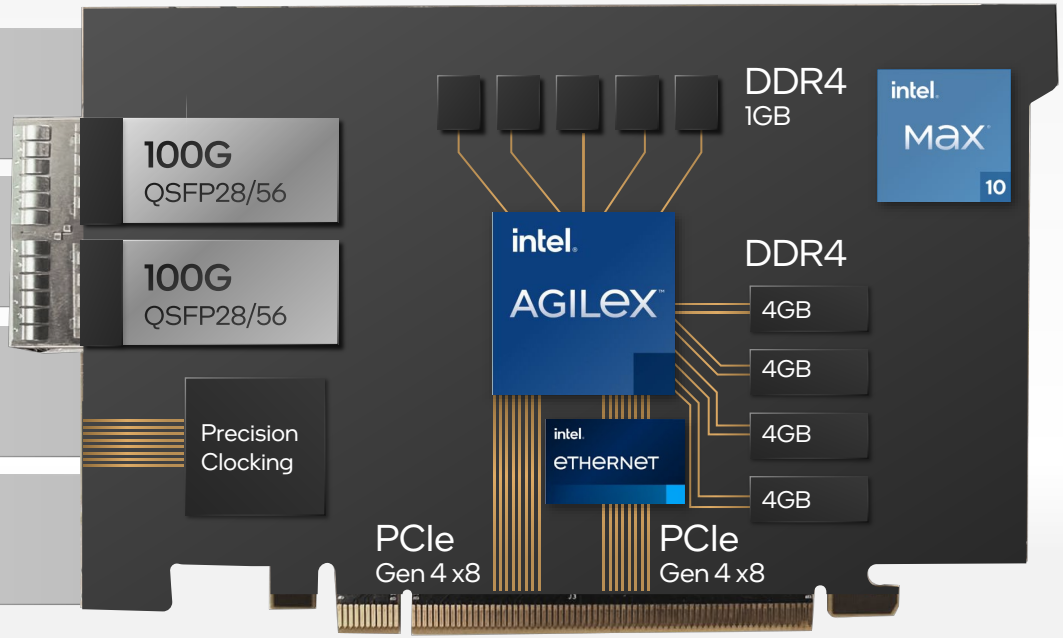
- Customizable packet processing including bridging and networking services
- Programmable through Intel OFS and DPDK
- Accelerated infrastructure workloads Juniper Contrail , OVS, SRv6, vFW
- Secure Remote Update of FPGA and Firmware over PCIe
- On-board root of trust



Arrow Creek

Built with Intel® Agilex FPGA and Ethernet E810 Controller

- 2x QSFP56 supporting up to 2x100GbE networking
- PCIe Gen4 x16
- 16GB DDR4 memory
  1GB DDR4 to integrated CPU
- Full height, half length PCIe form factor; passively cooled



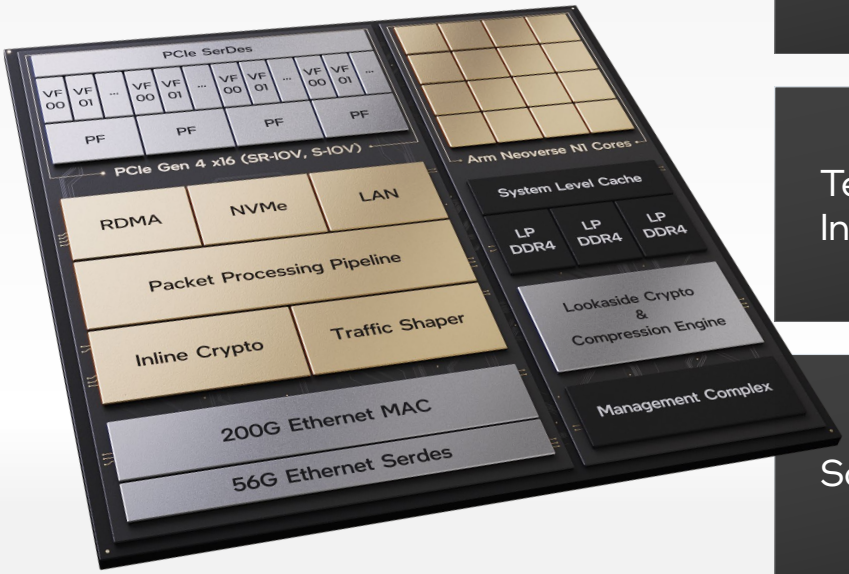
Mount Evans

Naru Sundar

Introducing

Mount Evans

Intel's 200G IPU



Hyperscale Ready

- Co-designed with a top cloud provider
- Integrated learnings from multiple gen. of FPGA sNICs
- High performance under real world load
- Security and isolation from the ground up

Technology Innovation

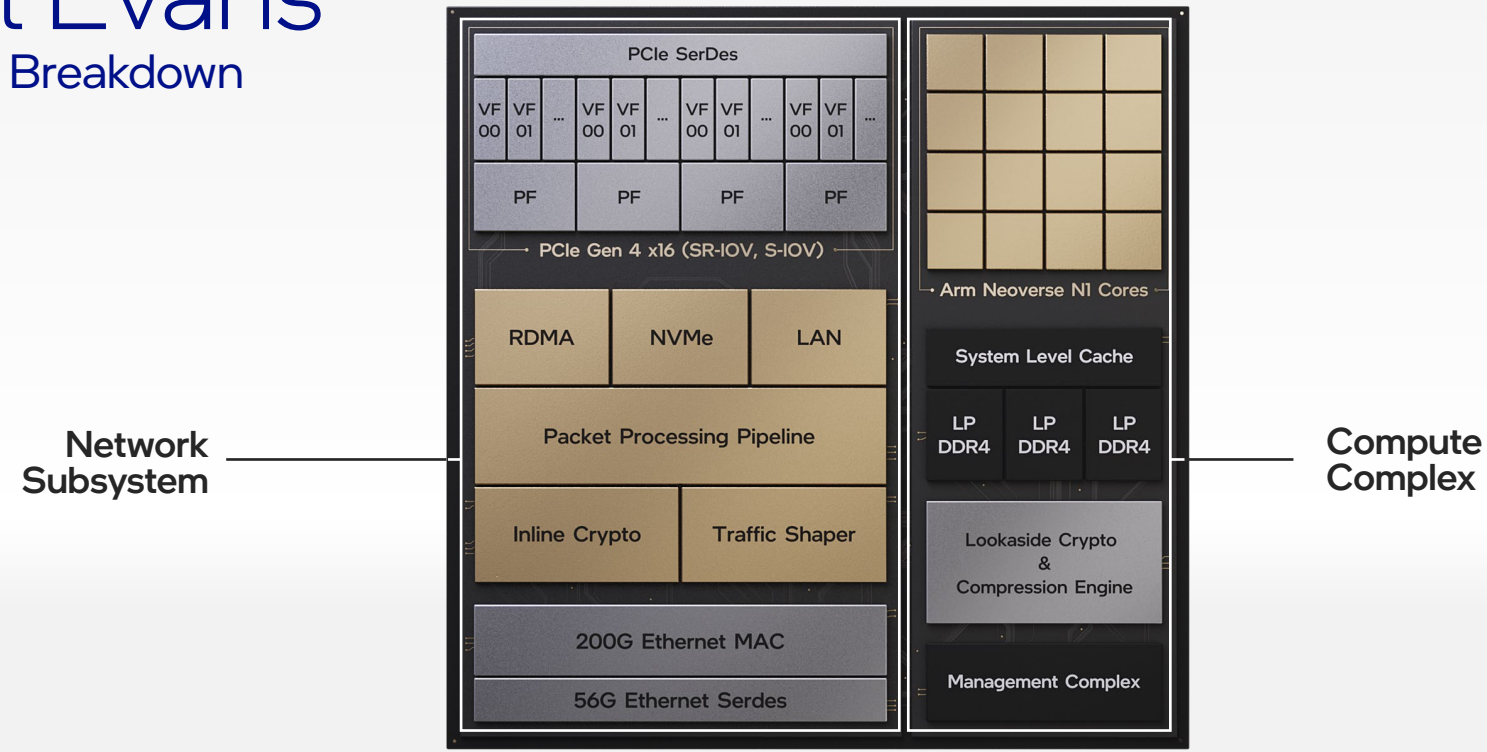
- Best-in-Class Programmable Packet Processing Engine
- NVMe storage interface scaled up from Intel Optane Tech
- Next Generation Reliable Transport
- Advanced crypto and compression accel.

Software

- SW/HW/Accel co-design
- P4 Studio based on Barefoot
- Leverage and extend DPDK and SPDK

Mount Evans

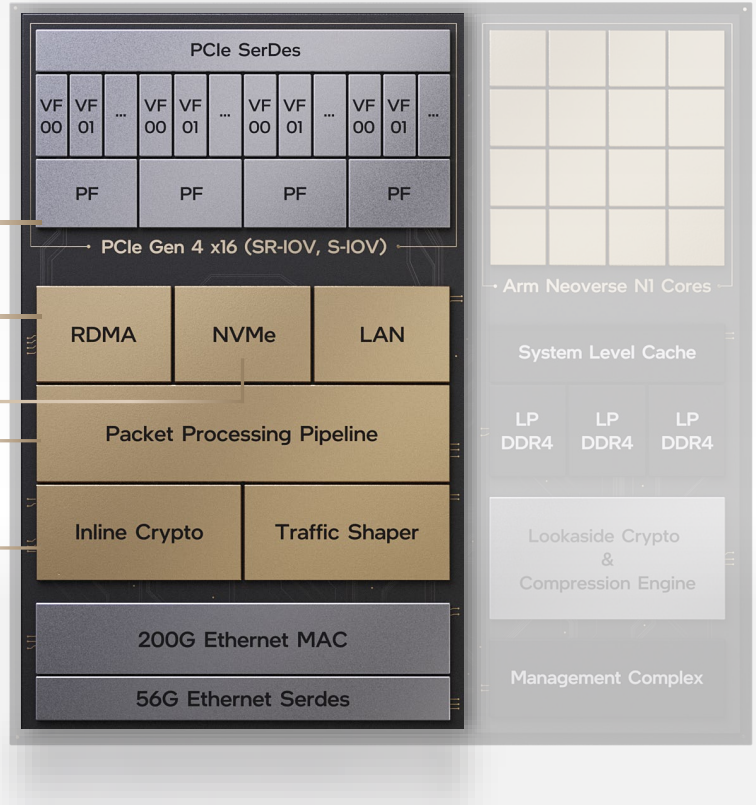
Architectural Breakdown



Mount Evans

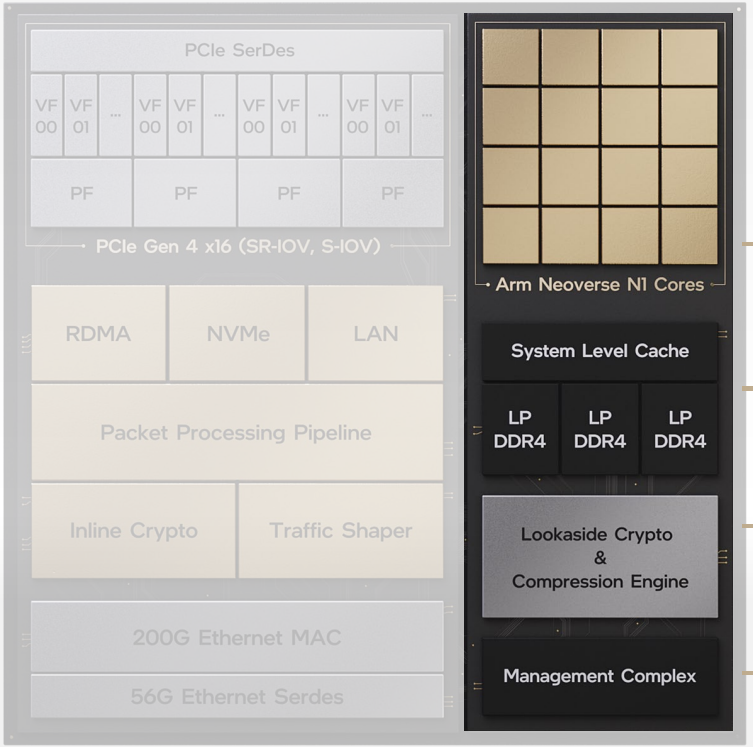
Architectural Breakdown

- Support for up to 4 host Xeons with 200Gb/s full duplex
- High-performance ROCEv2
- NVMe offload engine
- Programmable packet pipeline with QoS and telemetry capabilities
- Inline IPSec



Mount Evans

Compute Complex



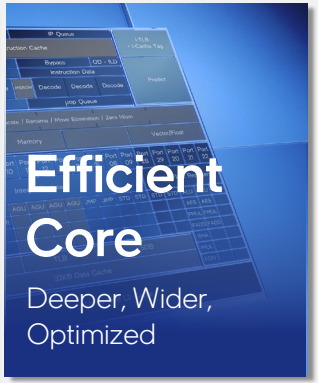
- Up to 16 Arm Neoverse® N1 Cores
- Dedicated compute and cache with up to 3 memory channels
- Lookaside crypto and compression
- Dedicated management processor

Architecture Day 2021

New Architectural Foundations

Performance Core

Biggest Shift in x86 yet



Efficient Core

Deeper, Wider, Optimized

Intel Thread Director

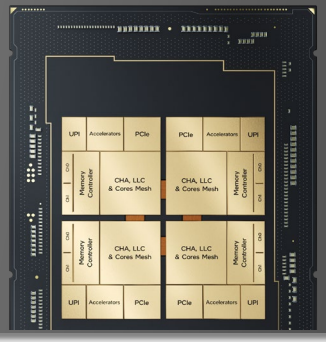
Scalable Hybrid Arch. Scheduling

Xe - core

Foundational Building Block for Xe With Xe Matrix Extensions

Sapphire Rapids

Biggest Leap in DC Capabilities in a decade



Xe HPC & Ponte Vecchio

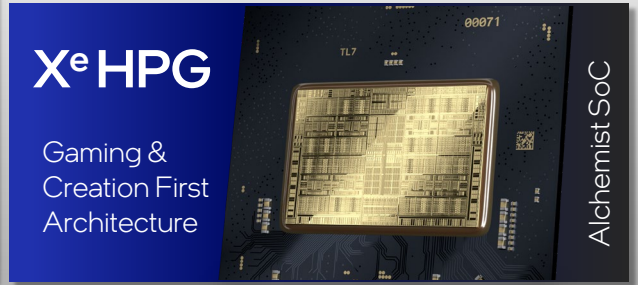
AMX

Advanced Matrix Extension - Engine

Alder Lake

Performance Hybrid

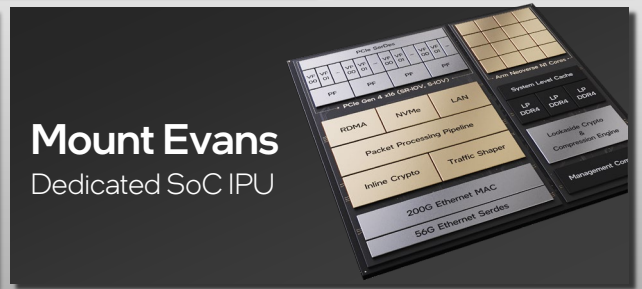
Xe SS



Xe HPG

Gaming & Creation First Architecture

Alchemist SoC

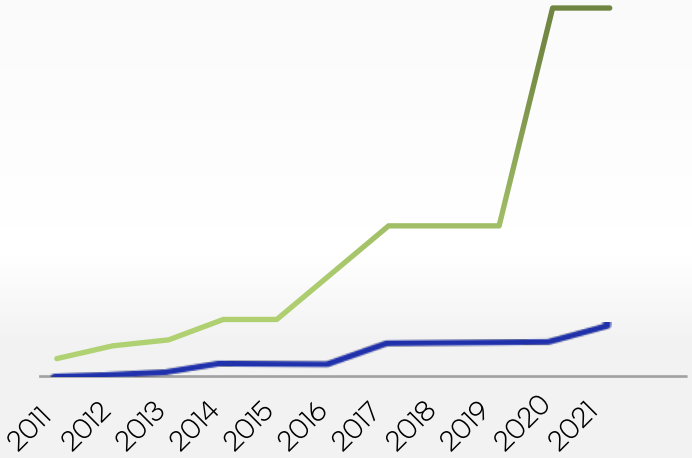


Mount Evans

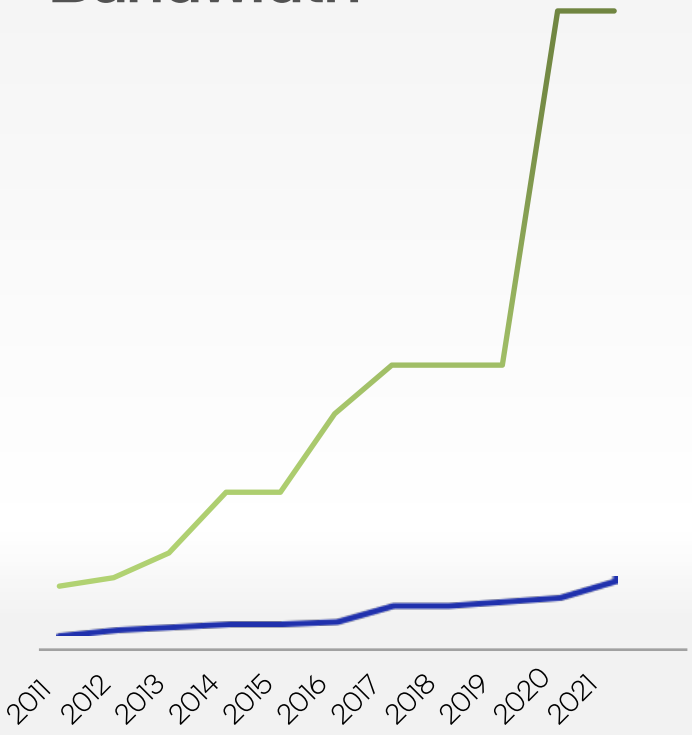
Dedicated SoC IPU

Xe HPC architecture

HPC FP64

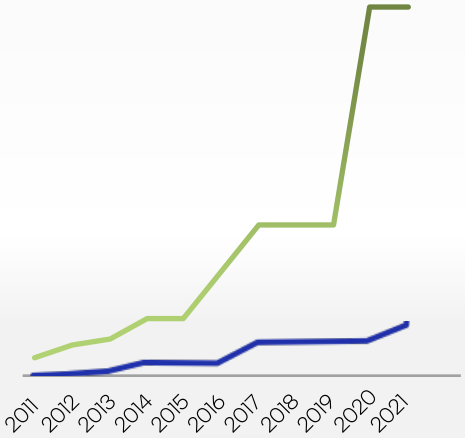


Bandwidth GB/s

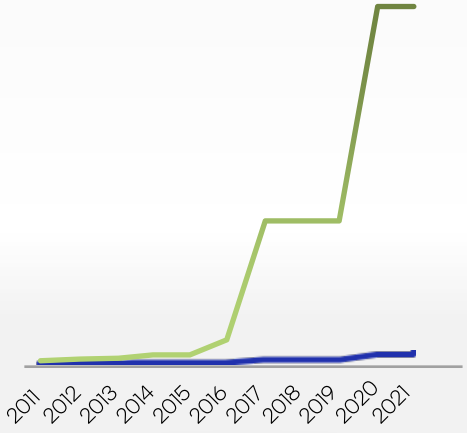


For workloads and configurations visit www.intel.com/ArchDay21claims . Results may vary.

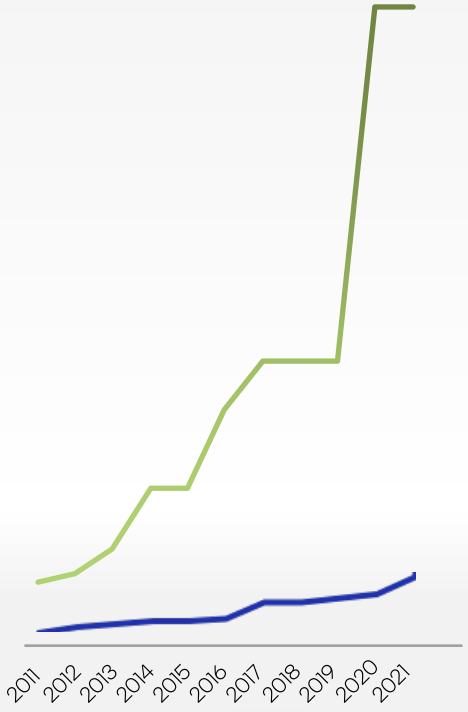
HPC FP64



AI FP16/BF16

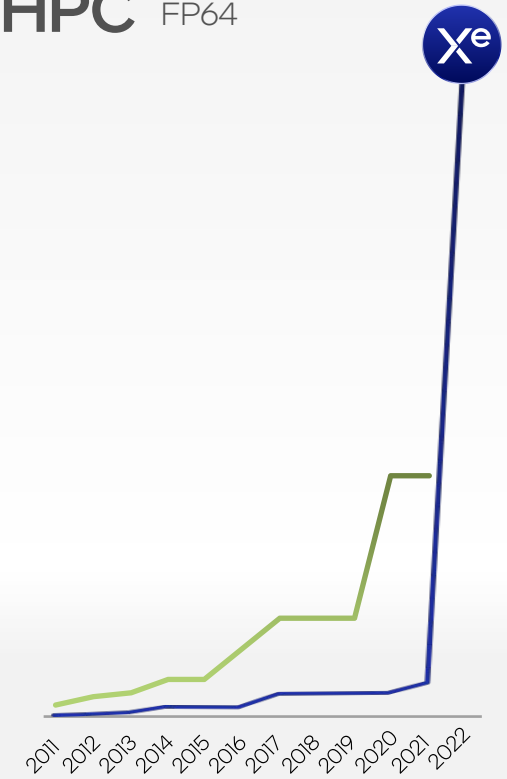


Bandwidth GB/s

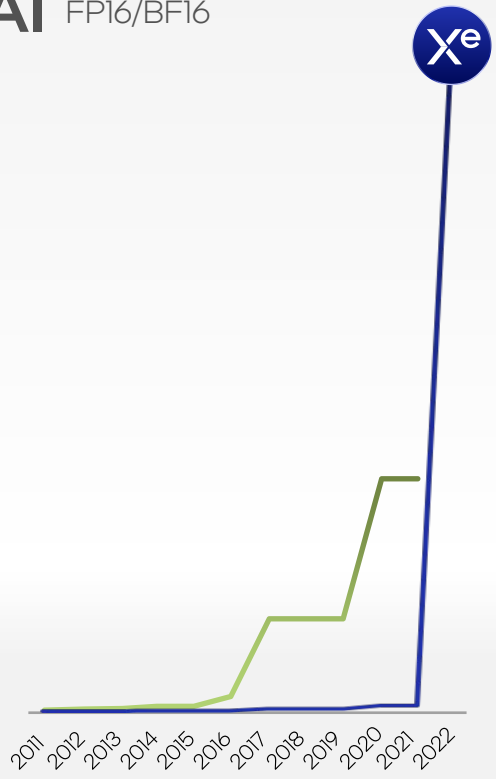


For workloads and configurations visit www.intel.com/ArchDay21claims . Results may vary.

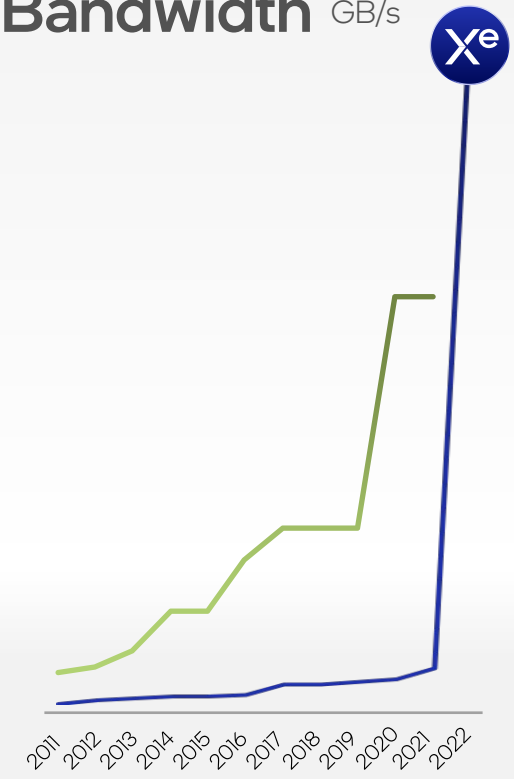
HPC FP64



AI FP16/BF16



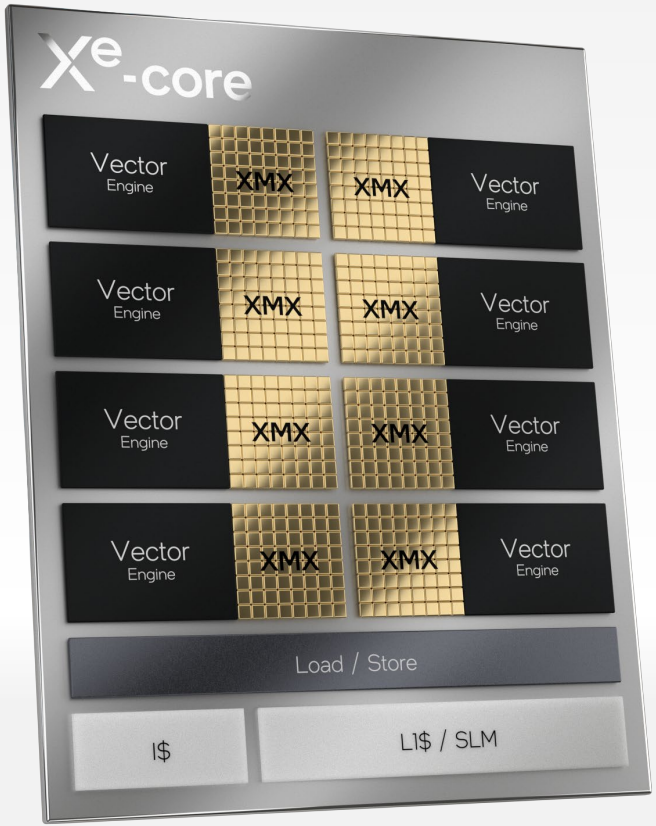
Bandwidth GB/s



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

Architecture

Hong Jiang



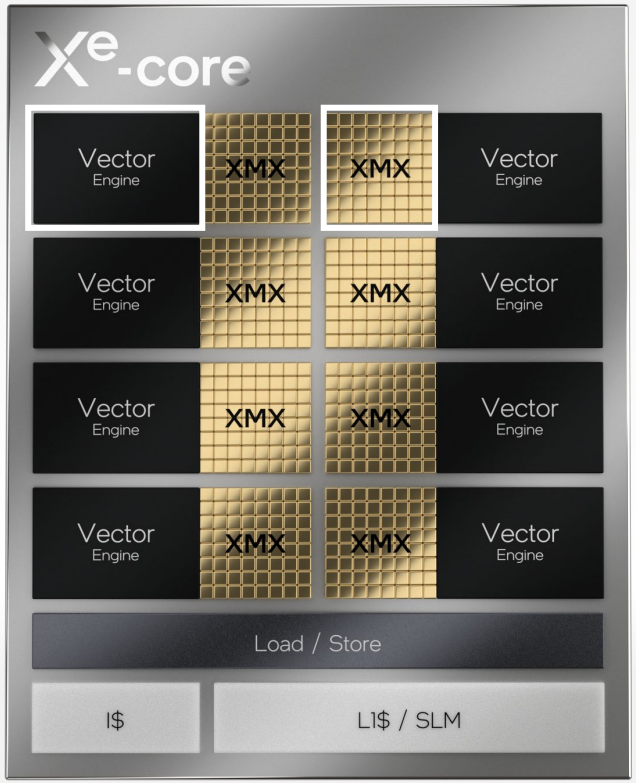
Xe-core

Compute Building Block of Xe HPC-based GPUs

8 Vector Engines 512 bit per engine	8 Matrix Engines 4096 bit per engine	Load / Store 512 B/CLK	Cache L1$/ SLM (512KB), I$

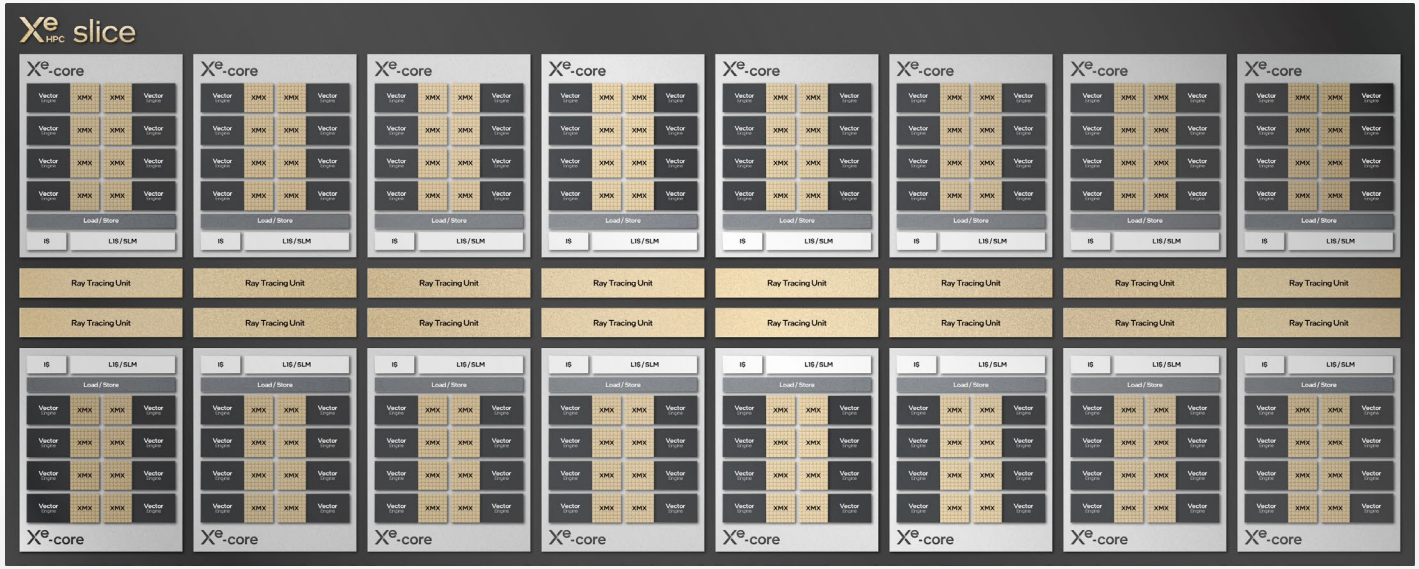
Xe HPC Core

Vector Engine (ops/clk)
256 FP32
256 FP64
512 FP16



Matrix Engine (ops/clk)
2048 TF32
4096 FP16
4096 BF16
8192 INT8

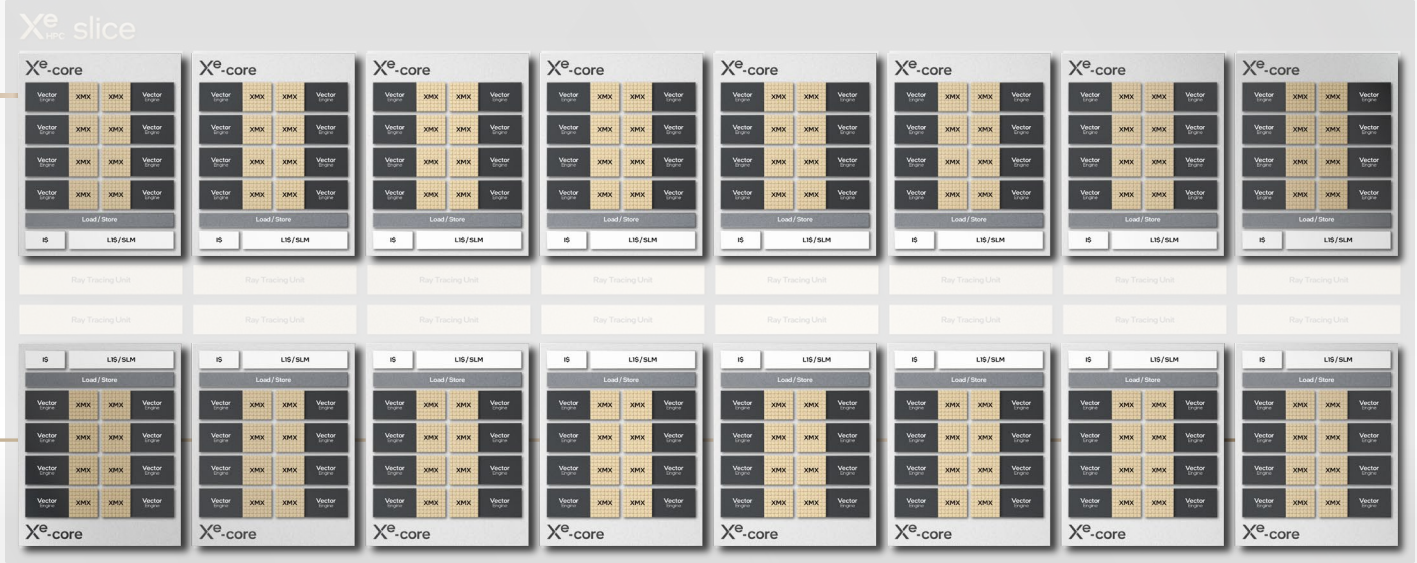
Xe HPC Slice


Xe HPC slice
Xe-core
Vector Engine
XMX
Load / Store
IS
L1S / SLM
Ray Tracing Unit

Xe HPC Slice

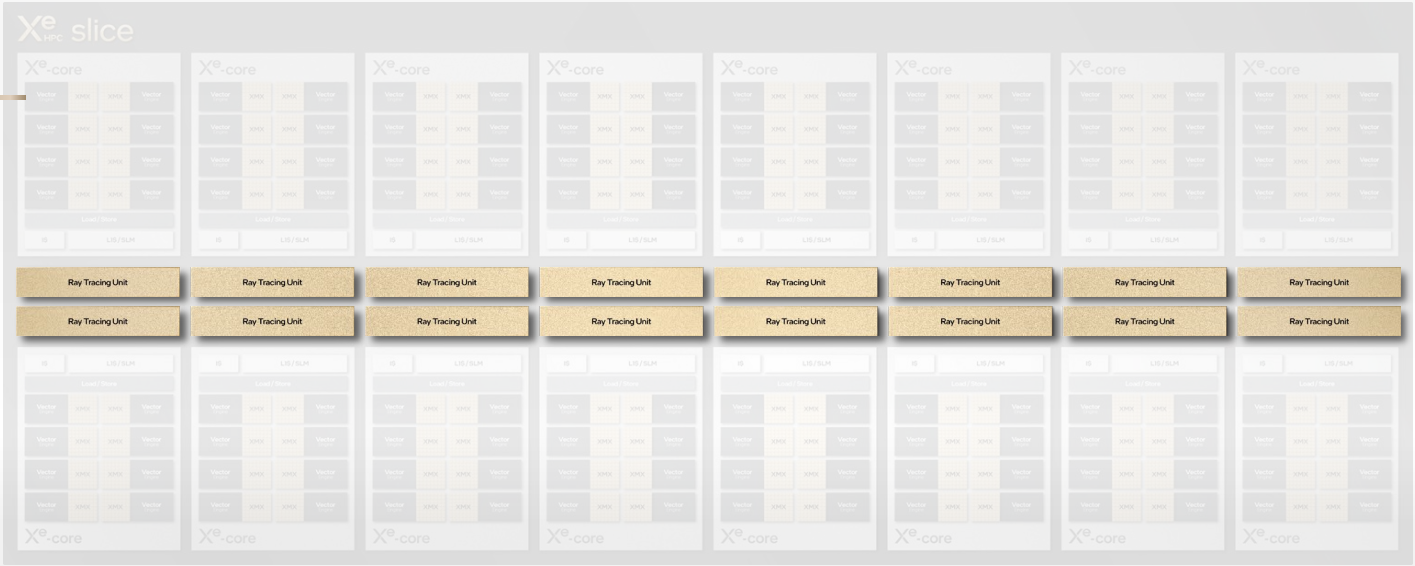
16 Xe – cores

8MB L1 Cache



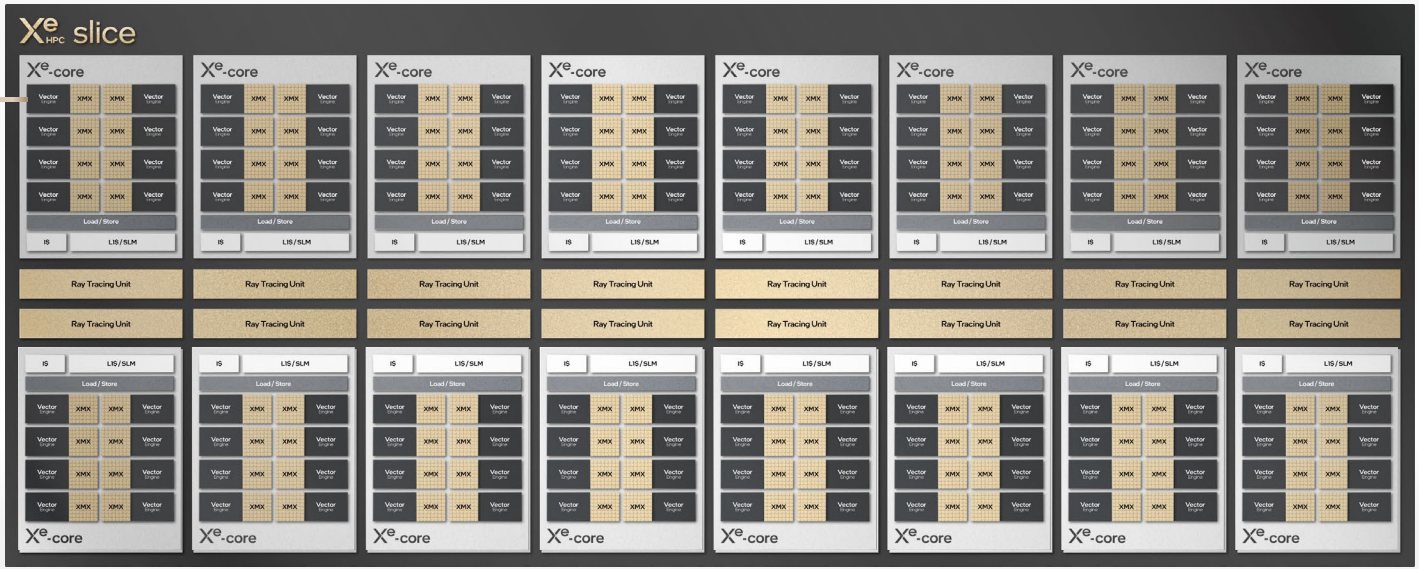
Xe HPC Slice

- 16 Xe – cores
  - 8MB L1 Cache
- 16 Ray Tracing Units
  - Ray Traversal
  - Triangle Intersection
  - Bounding Box Intersect.



Xe HPC Slice

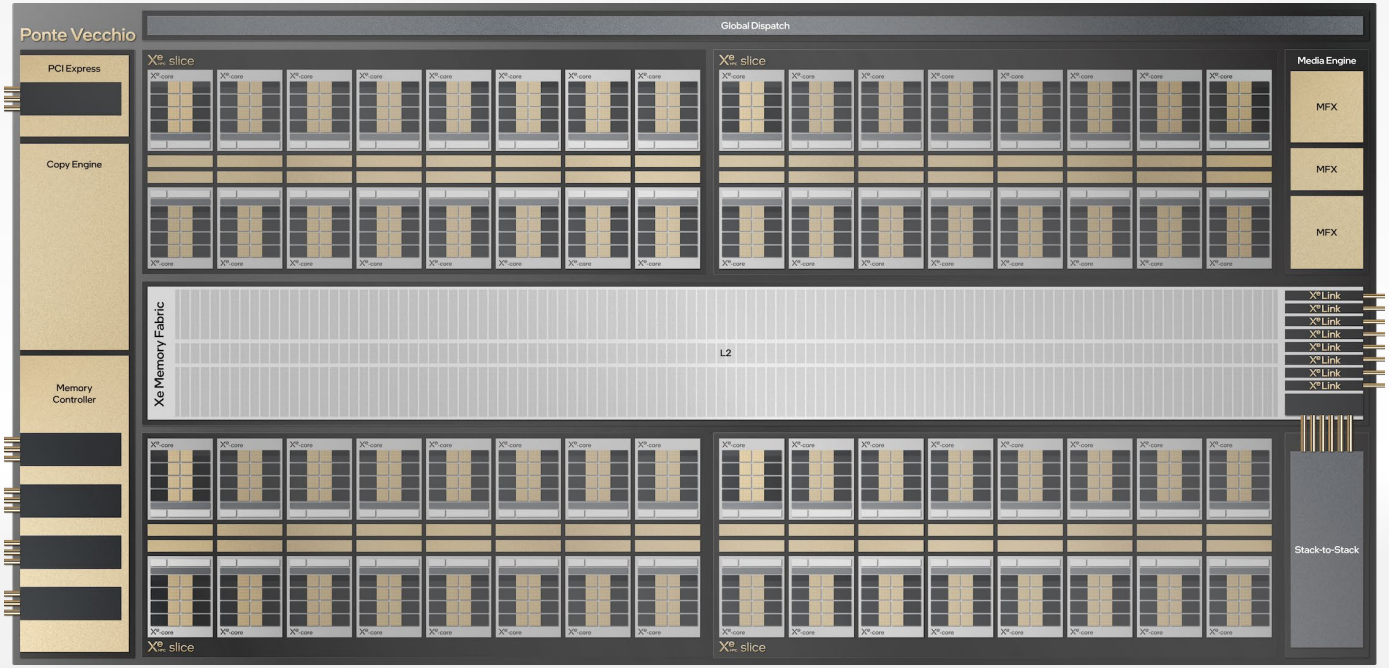
- 16 Xe – cores
  - 8MB L1 Cache
- 16 Ray Tracing Units
  - Ray Traversal
  - Triangle Intersection
  - Bounding Box Intersect.
- 1 Hardware Context



Xe HPC Stack

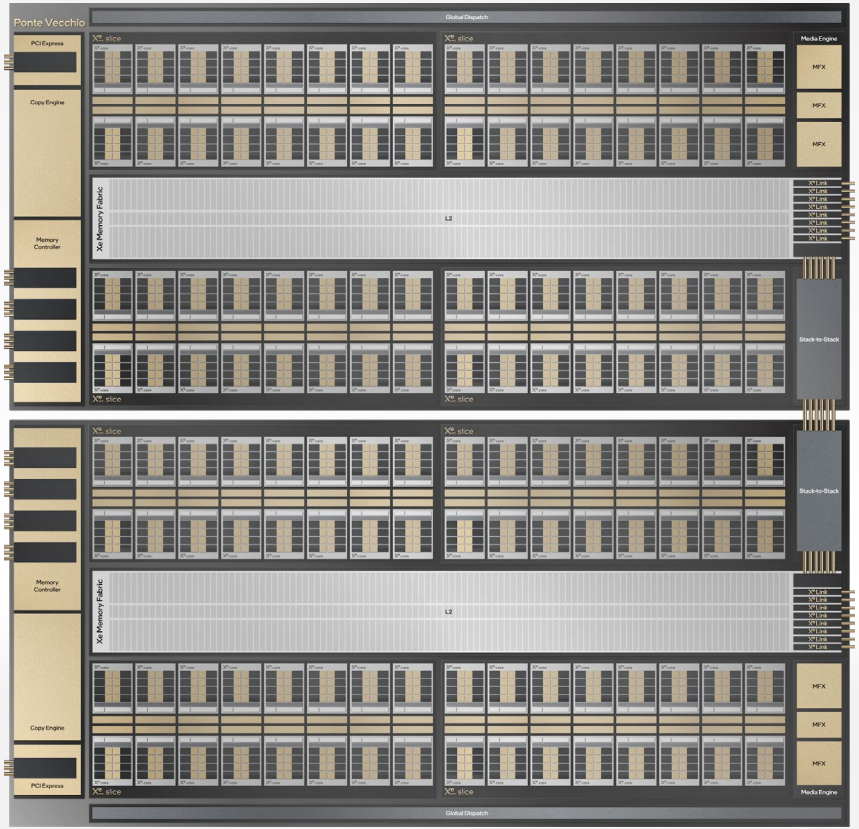
Up to

- 4 Slices
  - 64 Xe - cores
  - 64 Ray Tracing Units
  - 4 Hardware Contexts
- L2 Cache
- 4 HBM2e controllers
- 1 Media Engine
- 8 Xe Links



Xe HPC 2 - Stack

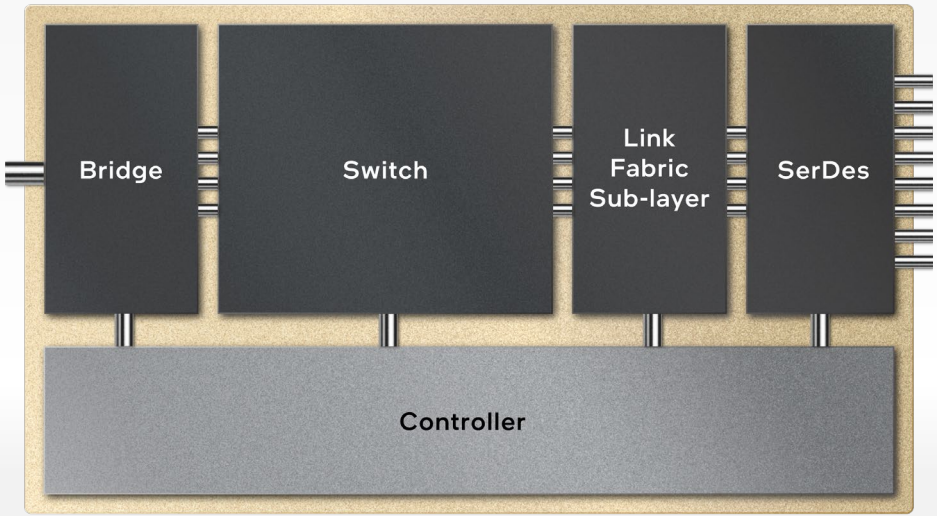
- 8 Slices
  - 128 Xe - cores
  - 128 Ray Tracing Units
  - 8 Hardware Contexts
- 2 Media Engines
- 8 HBM2e controllers
- 16 Xe Links



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

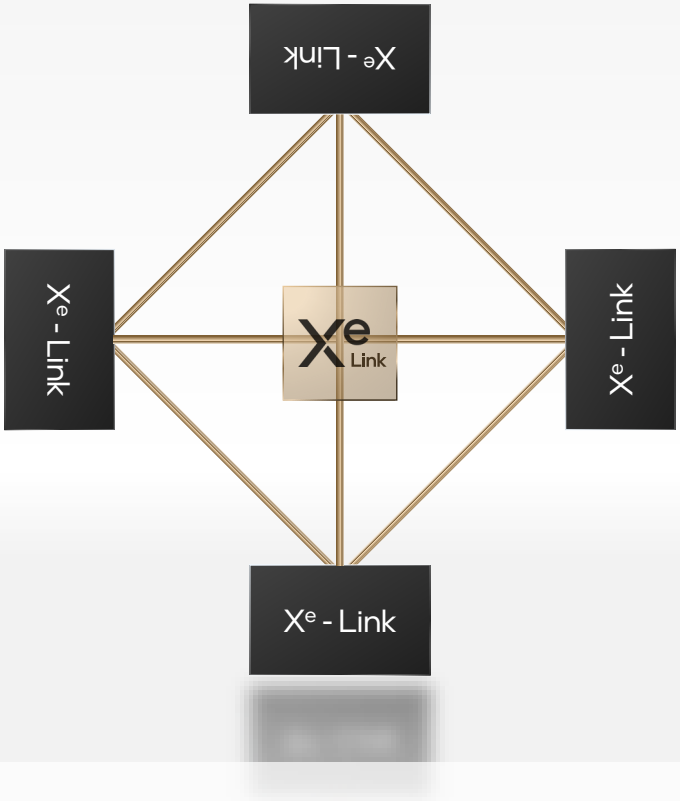
Xe Link

- High Speed Coherent Unified Fabric (GPU to GPU)
- Load/Store, Bulk Data Transfer & Sync Semantics
- Up to 8 Fully Connected GPUs through Embedded Switch

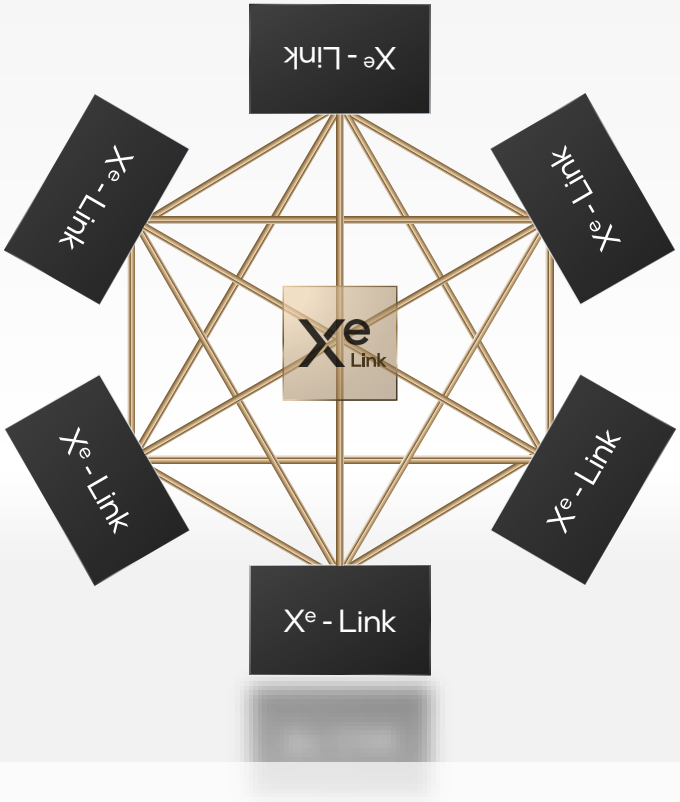


Link for Scalability

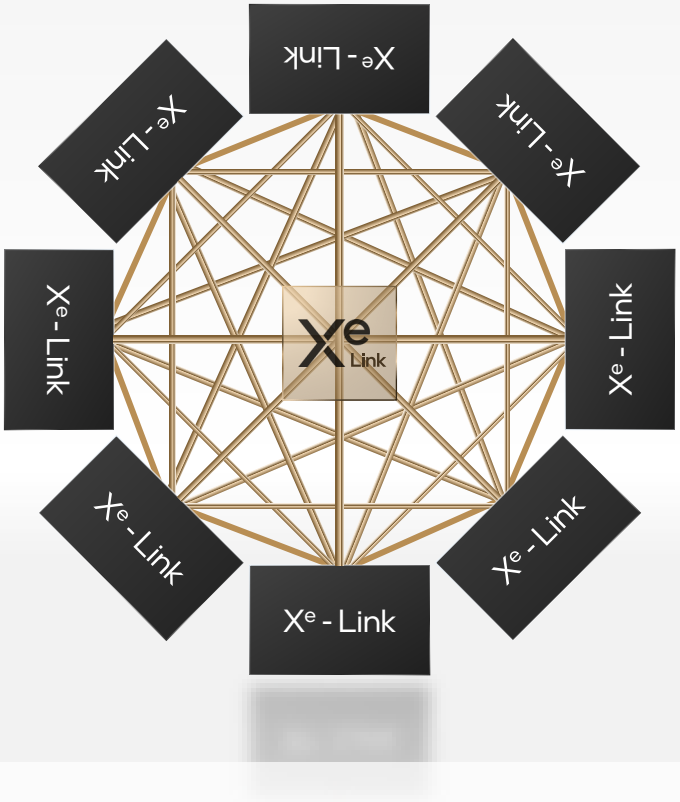
Link for Scalability

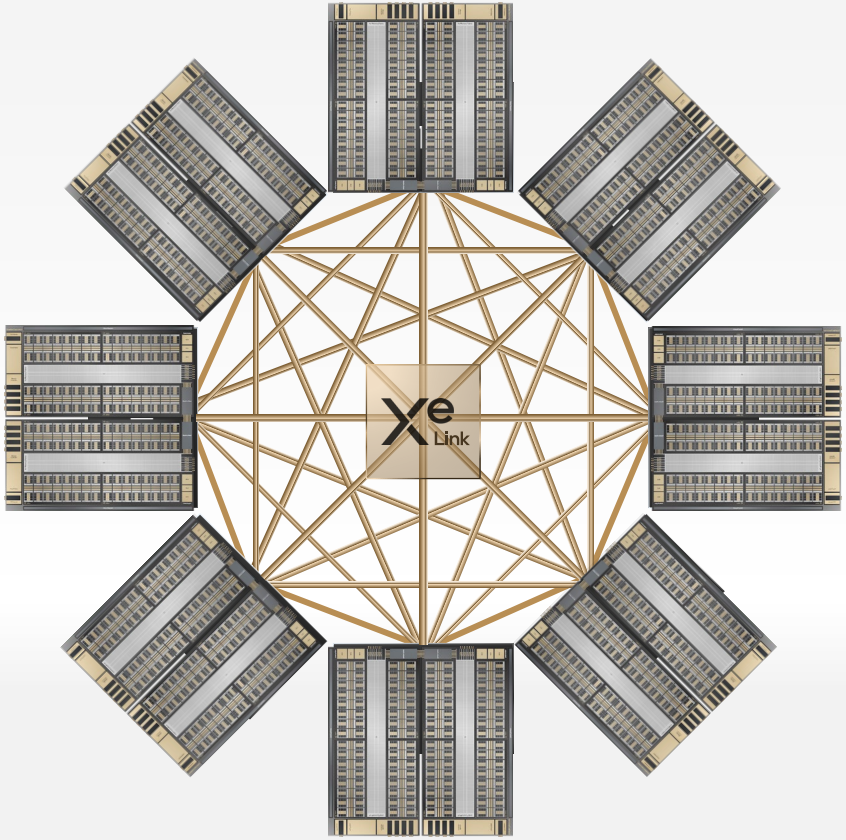


Link for Scalability



Link for Scalability





Xe HPC

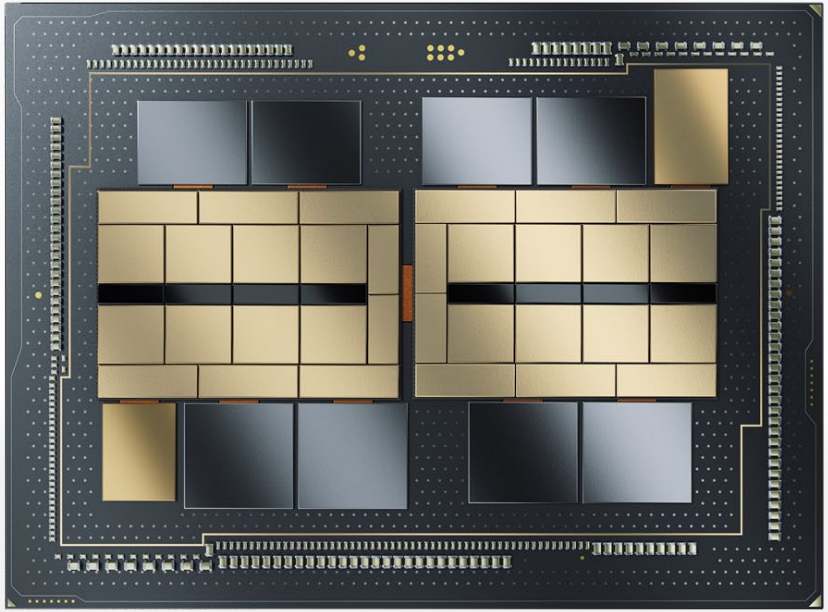
8x System Compute Rates

	Vector		Matrix	
8x	Up to 32,768 FP64 Ops/CLK	8x	Up to 262,144 TF32 Ops/CLK	
8x	Up to 32,768 FP32 Ops/CLK	8x	Up to 524,288 BF16 Ops/CLK	
		8x	Up to 1,048,576 INT8 Ops/CLK	

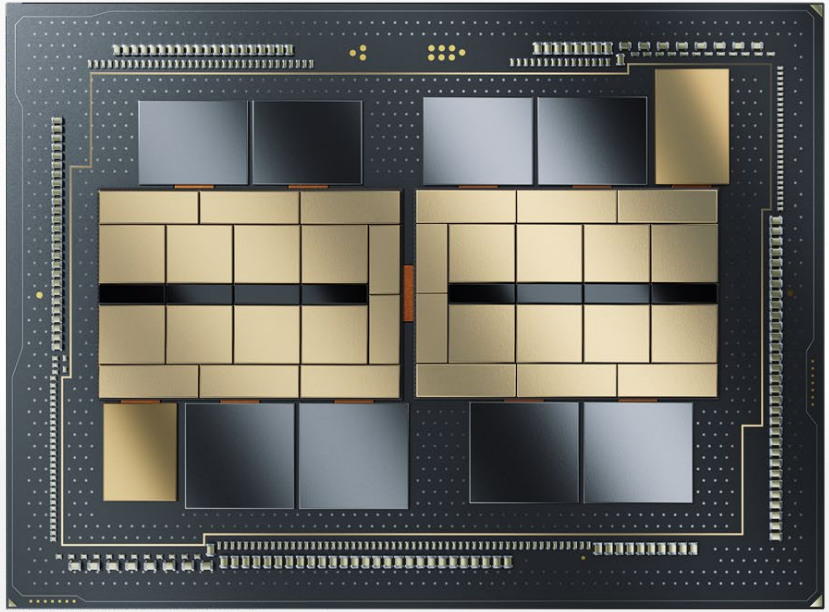
Ponte Vecchio

Masooma Bhaiwala

Ponte Vecchio



Ponte Vecchio



- New **Verification Methodology**
- New **Software**
- New **Reliability Methodology**
- New **Signal Integrity Techniques**
- New **Interconnects**
- New **Power Delivery Technology**
- New **Packaging Technology**
- New **I/O Architecture**
- New **Memory Architecture**
- New **IP Architecture**
- New **SOC Architecture**

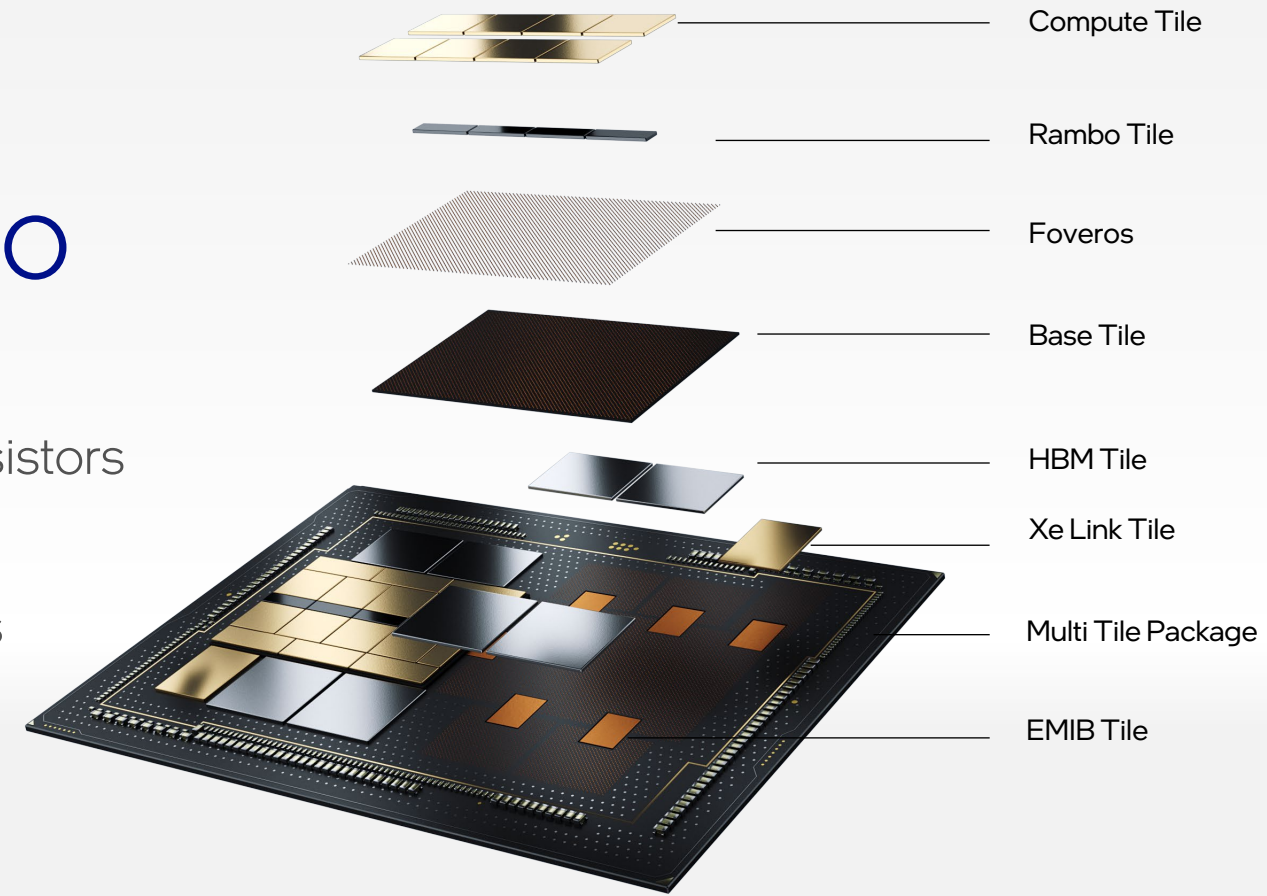
Ponte Vecchio

SOC

>100 Billion Transistors

47 Active Tiles

5 Process Nodes



Ponte Vecchio

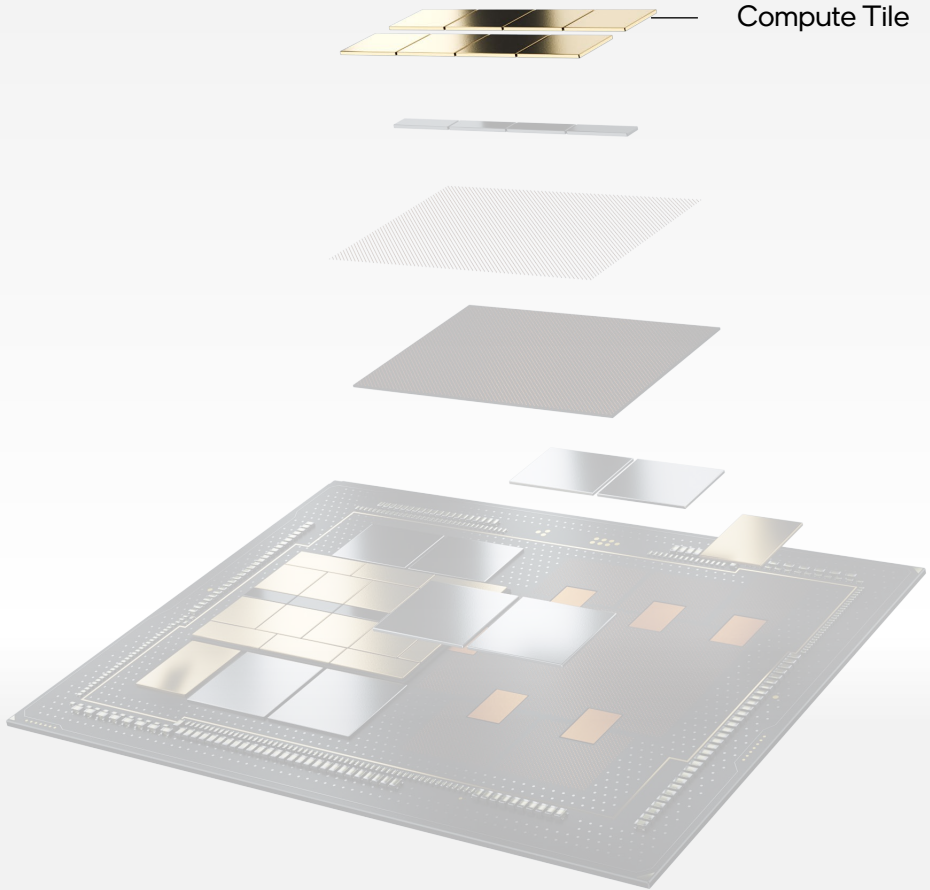
Key Challenges

- Scale of Integration
- Foveros Implementation
- Verification Tools & Methods
- Signal Integrity, Reliability & Power Delivery

Ponte Vecchio

Compute Tiles

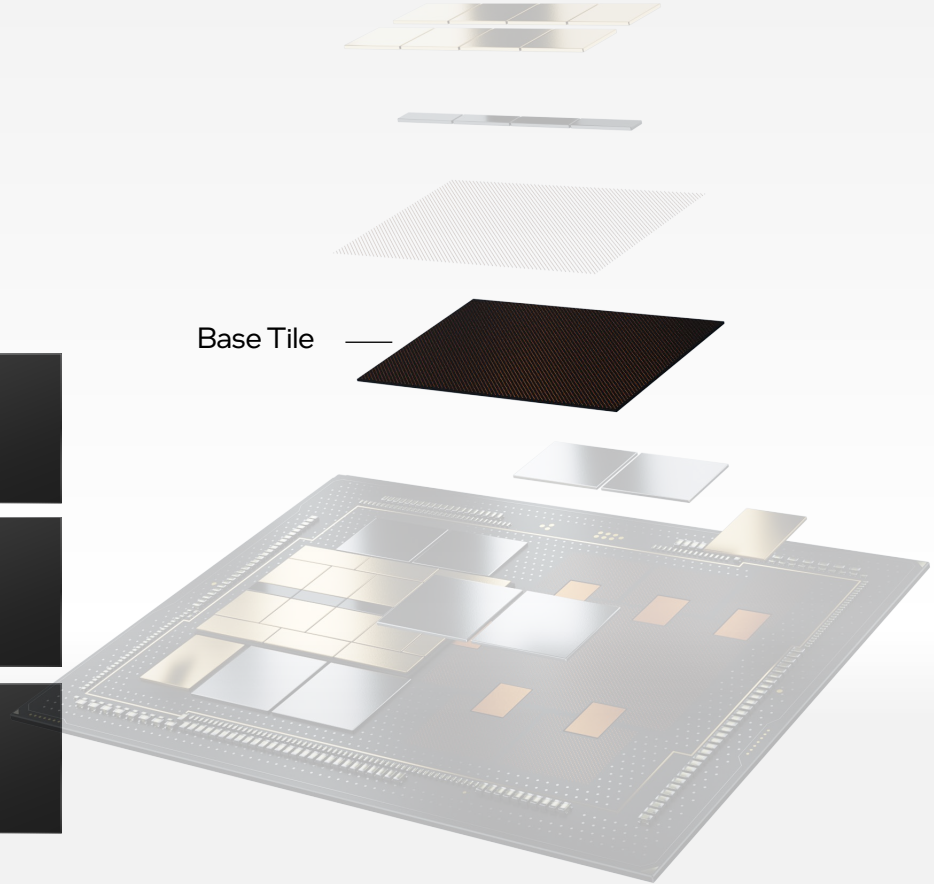
Per Tile **8** **Xᵉ - cores**	L1 Cache **4MB** **Per Tile**
Built on **TSMC** **N5**	Bump Pitch **36um** Foveros



Ponte Vecchio

Base Tile

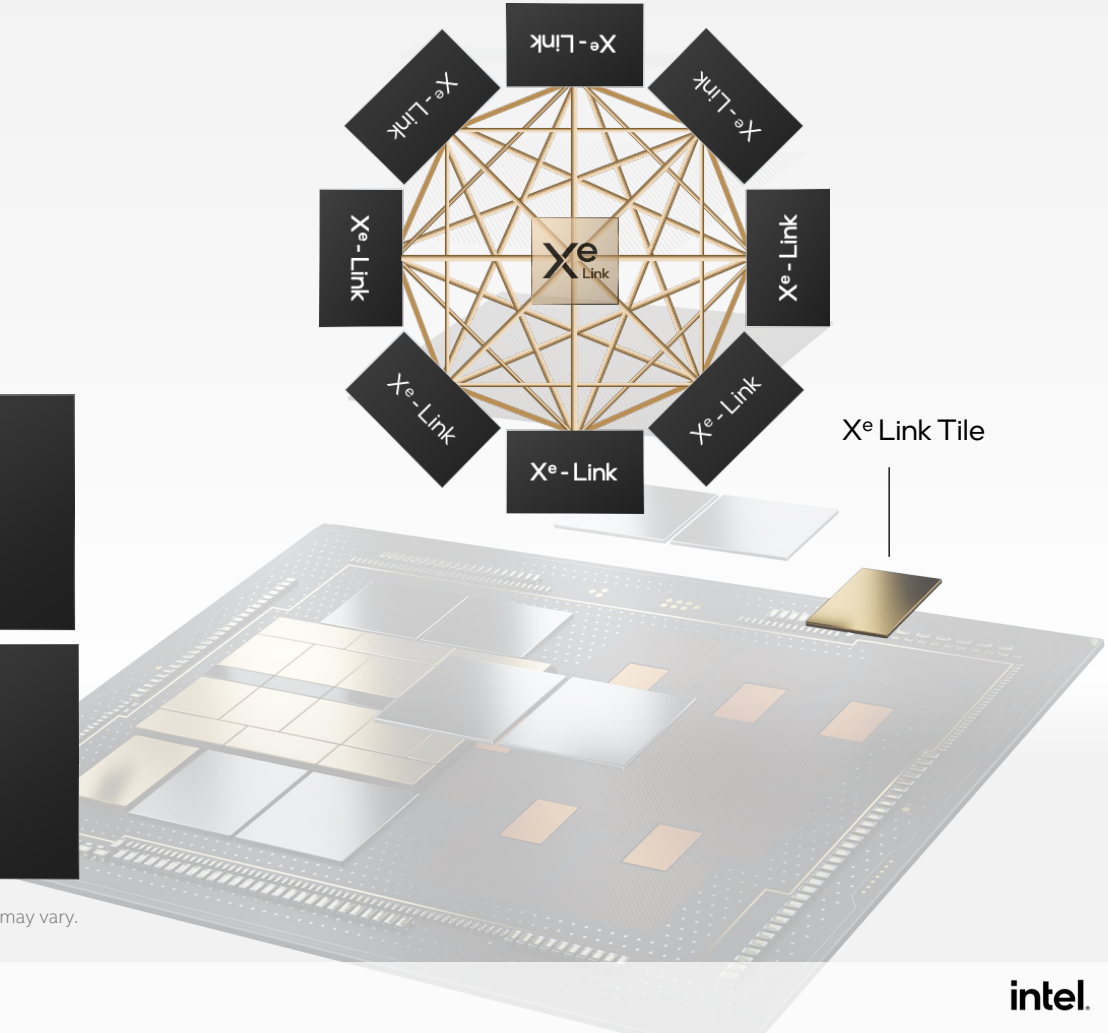
Built on **Intel 7** FOVEROS	Area **640mm²**	**HBM2e**
L2 Cache **144MB**	Host Interface **PCIe Gen5**	**MDFI**
		EMIB



Ponte Vecchio

Xe Link Tile

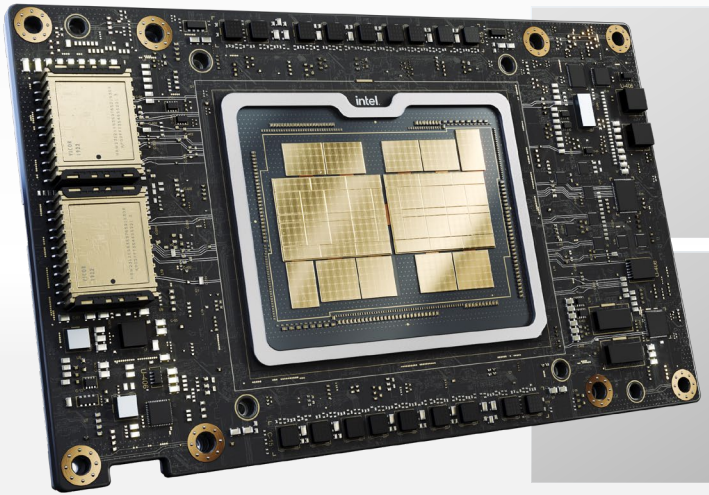
Per Tile **8 Xe Links**	8 ports **Embedded Switch**
Built on **TSMC N7**	Up to **90G** Serdes



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

Ponte Vecchio

Execution Progress



A0 Silicon Current Status

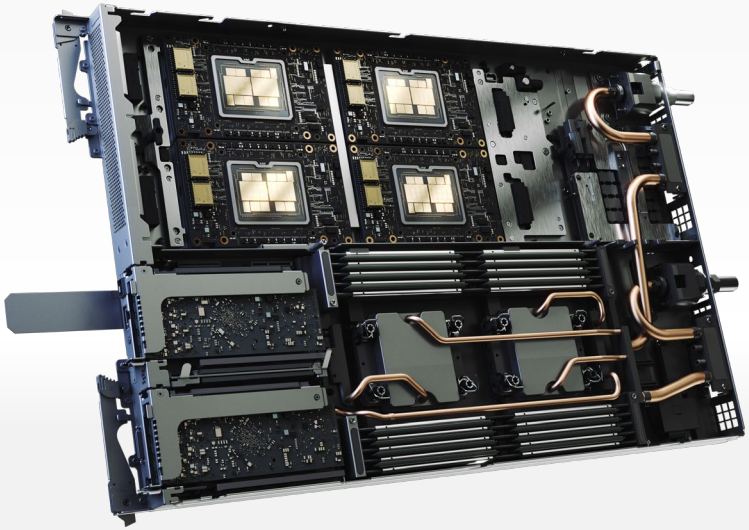
> 45 TFLOPS	FP32 Throughput
> 5 TBps	Memory Fabric Bandwidth
> 2 TBps	Connectivity Bandwidth

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

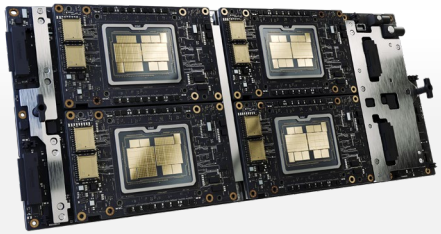
Accelerated Compute Systems

Ponte Vecchio x4 Subsystem
with Xe Links

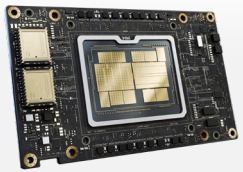
+ 2S Sapphire Rapids



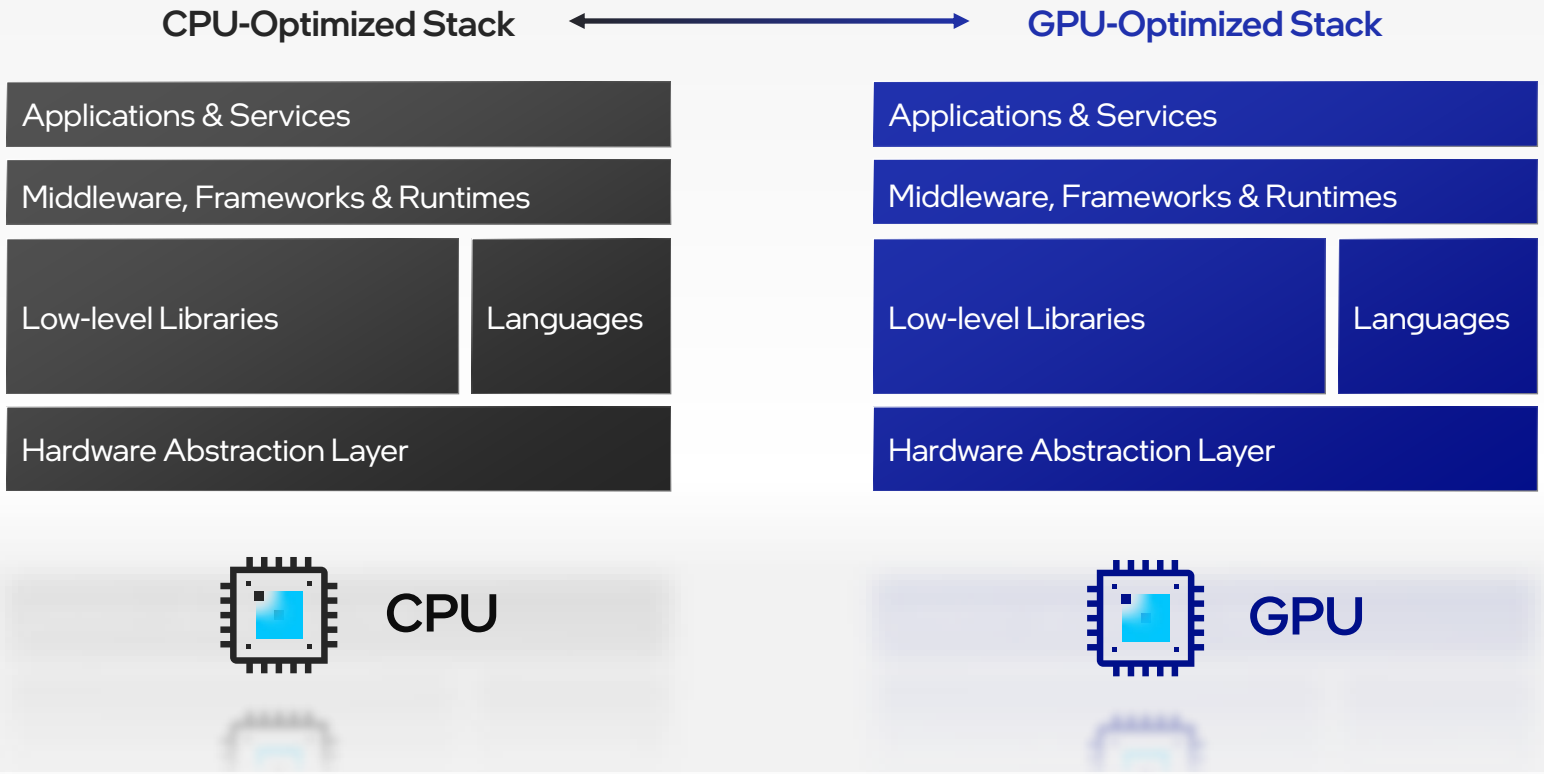
Ponte Vecchio
x4 Subsystem
with Xe Links



Ponte Vecchio
OAM



Overcoming Separate CPU and GPU Software Stacks



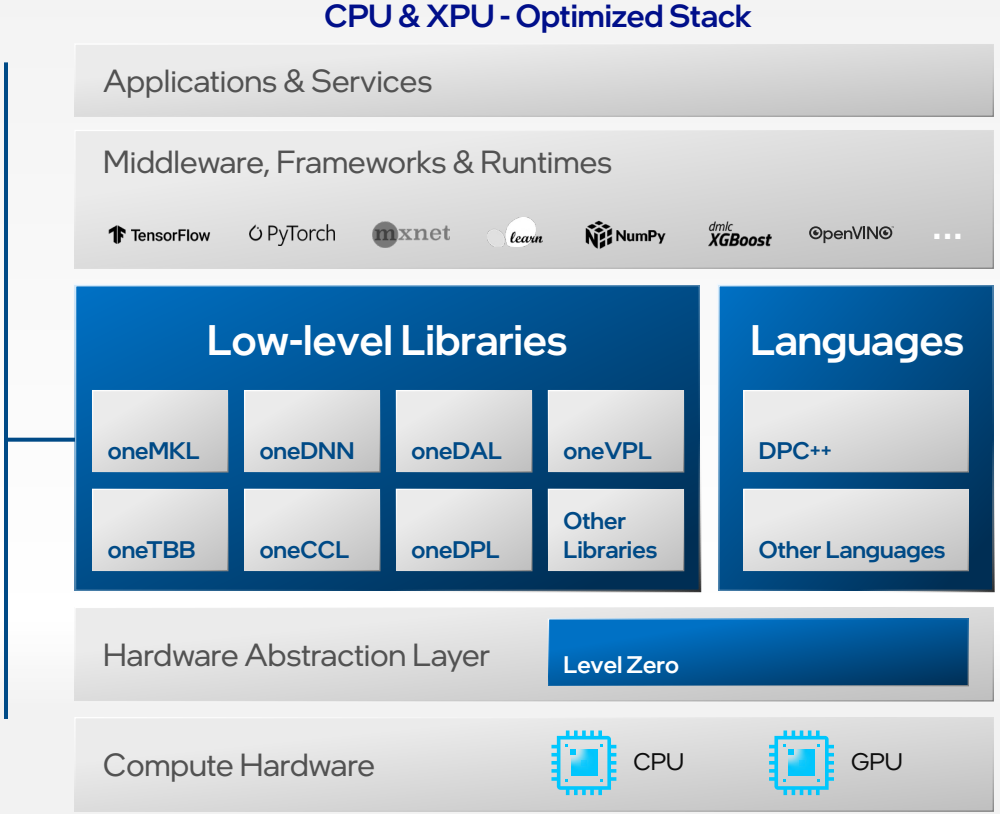
oneAPI

Open, Standards-Based Unified Software Stack

Freedom from proprietary programming models

Full performance from the hardware

Piece of mind for developers



oneAPI Industry Momentum

Cross-Vendor

3rd-party implementations on

- Nvidia GPU
- Arm CPU
- Huawei ASIC
- AMD GPU

Evolving Spec

Provisional spec v1.1 released May'21 with deep industry leader involvement

- + Graph interfaces for Deep Learning workloads
- + Advanced raytracing libraries

oneAPI

Industry Momentum

End Users

BioDataAnalysis GmbH
BENTLEY
Accrad Accelerated Radiology
archanan
LAIKA
EURECOM
GE
Brightskies
PHILIPS
大势智慧 DASPATIAL
SAMSUNG MEDISON
TANGENTSTUDIOS
allegro.ai
ILLUMINATION MACGUFF

National Labs

Argonne NATIONAL LABORATORY
CINECA
CERN openlab
Laboratório Nacional de Computação Científica

ISVs & OSVs

AI SINGAPORE
codeplay
CANONICAL
CERTIFACE
CGG
SAP
ANSYS
sas
FOUNDRY.
mercenaries engineering Guerilla
SENAI CIMATEC FIEB
CHAOSGROUP
MAXON
GIGASPACES
MEGH COMPUTING
KFBIO
Tech Mahindra
AsiaInfo 亚信科技
SUSE
UNITED IMAGING

OEMs & SIs

DELL Technologies
Atos
Hewlett Packard Enterprise
HCL
BittWare a molex company
Lenovo
RENIAC
MEGWARE

Universities & Research Institutes

Ben-Gurion University of the Negev
UNIVERSITY OF OREGON
TÉCNICO LISBOA
inesc id lisboa
ZUSE INSTITUTE BERLIN
Stockholm University
lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften
LOBACHEVSKY UNIVERSITY
सी डैक CDAC
UNIVERSIDAD DE MÁLAGA
UNIVERSITY OF CAMBRIDGE
University College London
ILLINOIS Theoretical and computational Biophysics Group
KTH
Indian Institutes of Technology Delhi / Kharagpur / Roorkee
Indian Institute of Science Bangalore
IISER PUNE
Indian Institute of Science Education & Research Pune

CSPs & Frameworks

Microsoft Azure
Alibaba Cloud
Google Cloud
TensorFlow
Taboola
Tencent 腾讯
SBER BANK
PyTorch

>200K Developers

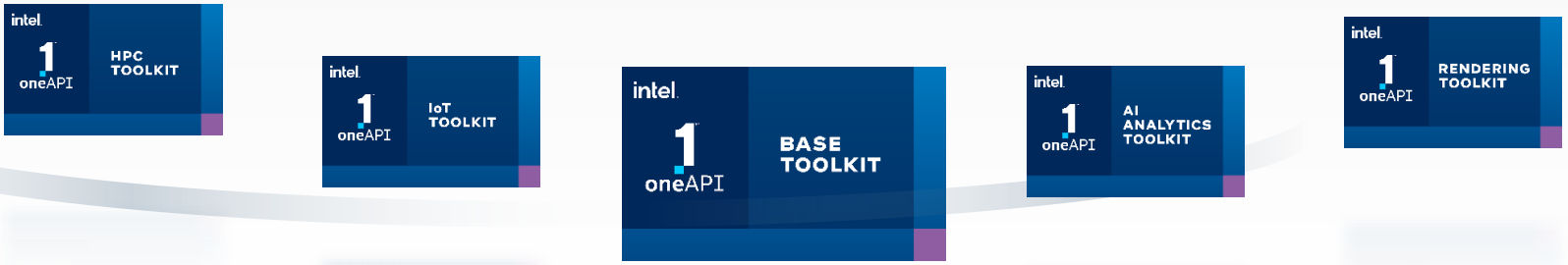
Unique installs of Intel® oneAPI product since Dec'20 release

>300 Applications

Deployed in market using Intel® oneAPI language & libraries

>80 HPC & AI Applications

Functional on Intel's Xᵉ HPC architecture using Intel® oneAPI



oneAPI

Toolkits v2021.3
Available Now

>200K Developers

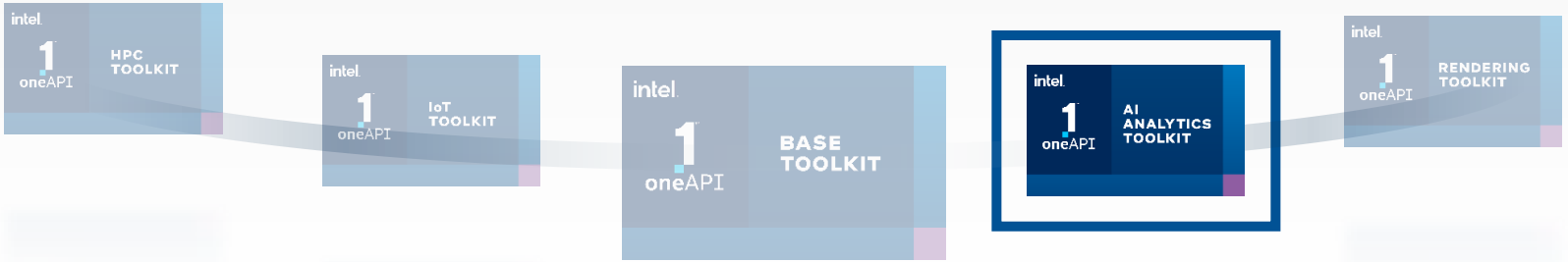
Unique installs of Intel® oneAPI product since Dec'20 release

>300 Applications

Deployed in market using Intel® oneAPI language & libraries

>80 HPC & AI Applications

Functional on Intel's Xᵉ HPC architecture using Intel® oneAPI



oneAPI

>200K Developers

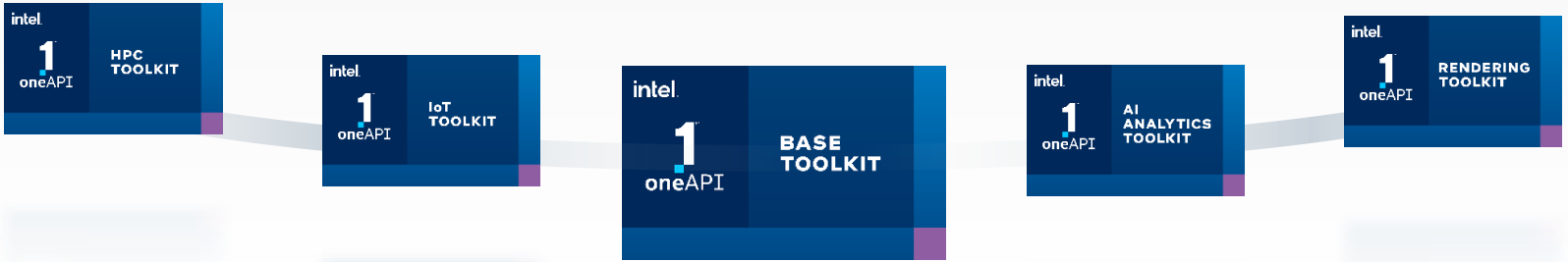
Unique installs of Intel® oneAPI product since Dec'20 release

>300 Applications

Deployed in market using Intel® oneAPI language & libraries

>80 HPC & AI Applications

Functional on Intel's Xe HPC architecture using Intel® oneAPI



oneAPI

>200K Developers

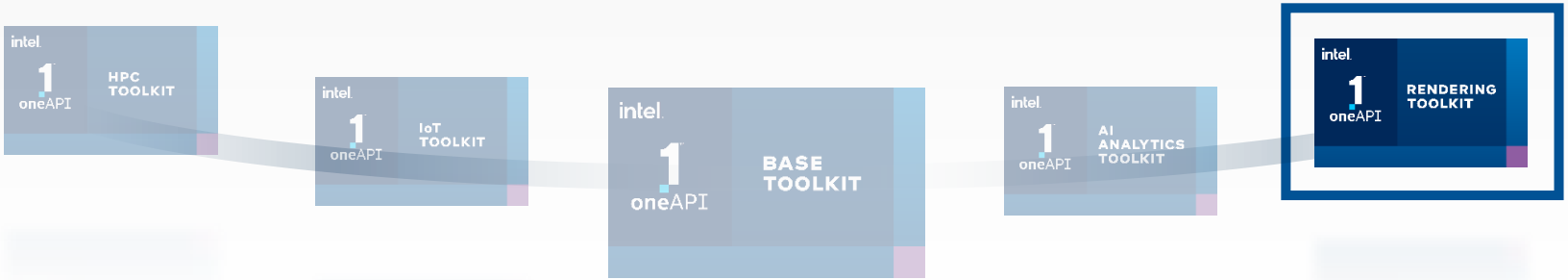
Unique installs of Intel® oneAPI product since Dec'20 release

>300 Applications

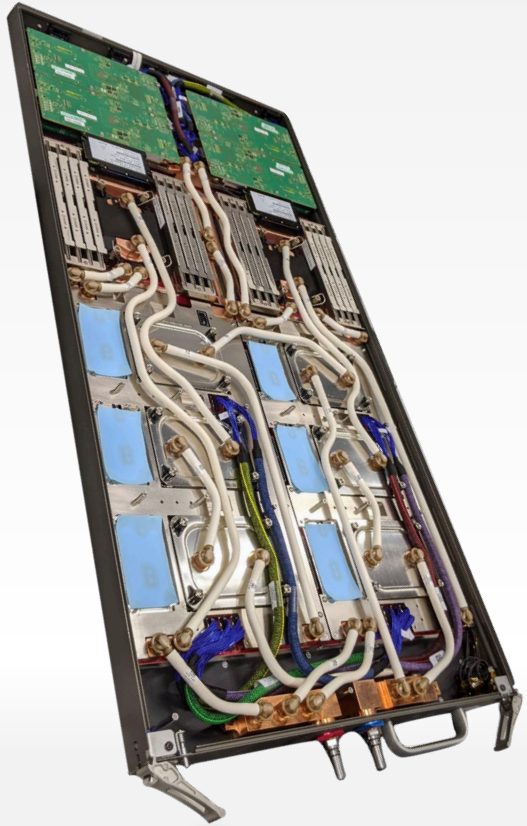
Deployed in market using Intel® oneAPI language & libraries

>80 HPC & AI Applications

Functional on Intel's Xᵉ HPC architecture using Intel® oneAPI



oneAPI



Aurora Blade

Building Block for the ExaScale Supercomputer

oneAPI

Argonne NATIONAL LABORATORY

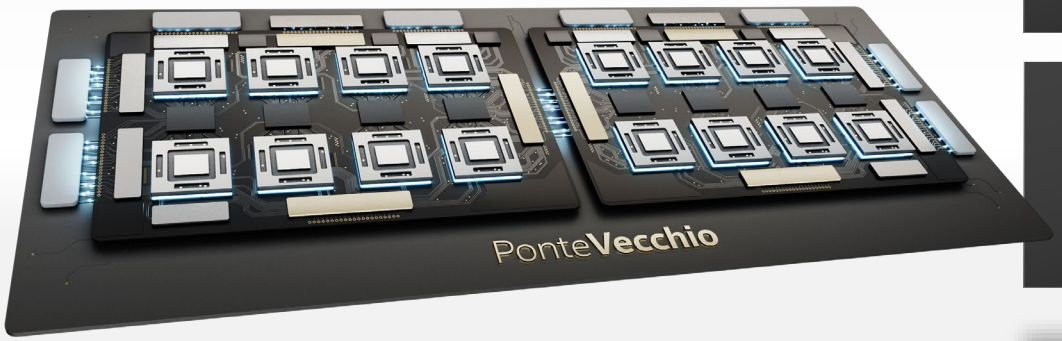
U.S. DEPARTMENT OF ENERGY

Hewlett Packard Enterprise

intel

Ponte Vecchio

The vision 2 years ago...



Leadership Performance for HPC/AI

Connectivity to drive scaleup and scale out

Unified Programming Model powered with oneAPI

intel

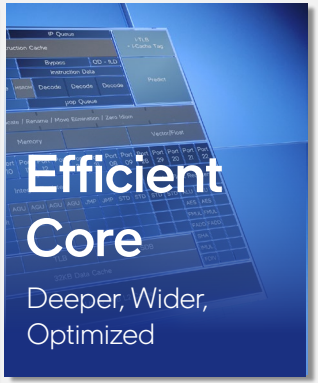
Architecture Day 2021

Architecture Day 2021

New Architectural Foundations

Performance Core

Biggest Shift in x86 yet

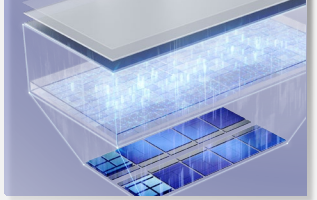


Efficient Core

Deeper, Wider, Optimized

Intel Thread Director

Scalable Hybrid Arch. Scheduling

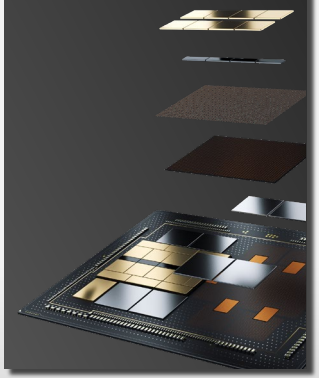


Xe - core

Foundational Building Block for Xe With Xe Matrix Extensions

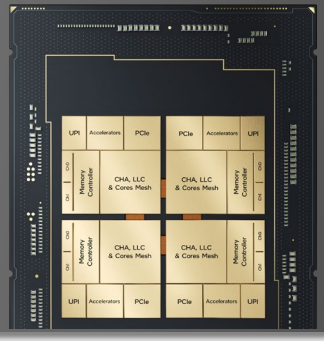
Xe HPC & Ponte Vecchio

Back to FLOP Leadership

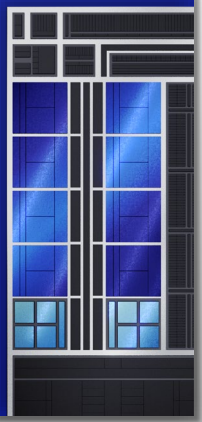


Sapphire Rapids

Biggest Leap in DC Capabilities in a decade



Xe SS

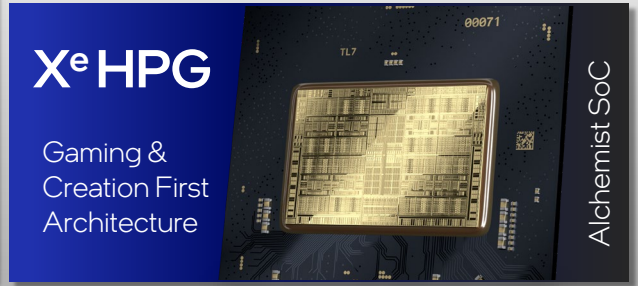


Alder Lake

Performance Hybrid

AMX

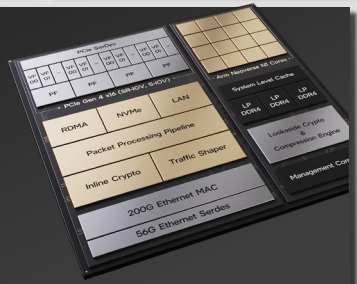
Advanced Matrix Extension - Engine



Xe HPG

Gaming & Creation First Architecture

Alchemist SoC



Mount Evans

Dedicated SoC IPU

See you at

intel

Architecture Day 2021

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See www.Intel.com/ArchDay21claims for configuration details. No product or component can be absolutely secure.

All product plans and roadmaps are subject to change without notice. Results that are based on pre-production systems and components as well as results that have been estimated or simulated using an Intel Reference Platform (an internal example new system), internal Intel analysis or architecture simulation or modeling are provided to you for informational purposes only. Results may vary based on future changes to any systems, components, specifications, or configurations. Intel technologies may require enabled hardware, software or service activation.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

Intel contributes to the development of benchmarks by participating in, sponsoring, and/or contributing technical support to various benchmarking groups, including the BenchmarkXPRT Development Community administered by Principled Technologies.

Statements in this presentation that refer to 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 estimates, forecasts, projections, uncertain events or assumptions, including statements relating to future products and technology and the expected availability and benefits of such products and technology, market opportunity, and anticipated trends in our businesses or the markets relevant to them, also identify forward-looking statements. Such statements are based on management's current expectations 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 reports filed or furnished with the Securities and Exchange Commission (SEC), including Intel's most recent reports on Form 10-K and Form 10-Q, available at Intel's investor relations website at www.intc.com and the SEC's website at www.sec.gov. 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.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.