

Data Centric Processor Roadmap



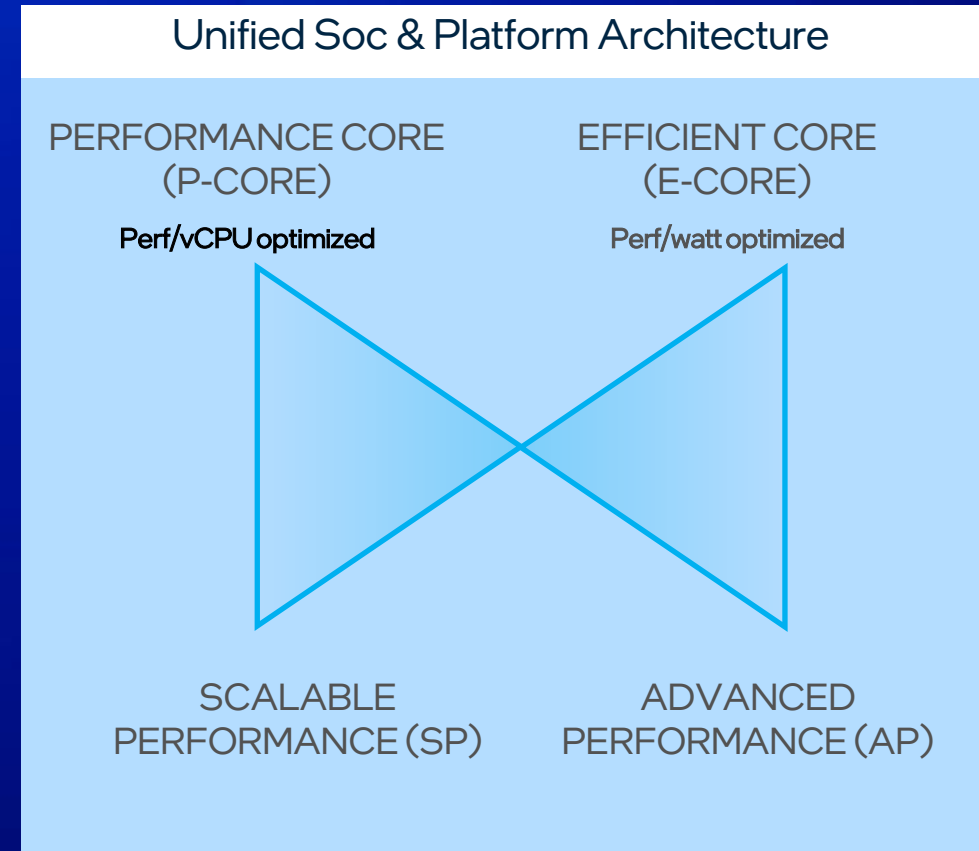
Intel® Xeon® Products Strategy

One unified SoC & Platform Architecture

Two optimization points (P-core & E-core)

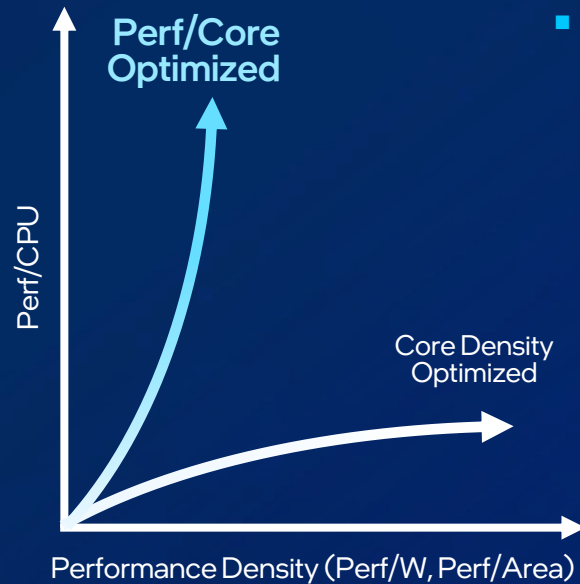
Two Platform design points (SP & AP)

Shared software stack and eco-system



CPUs Optimized for Mainstream Compute

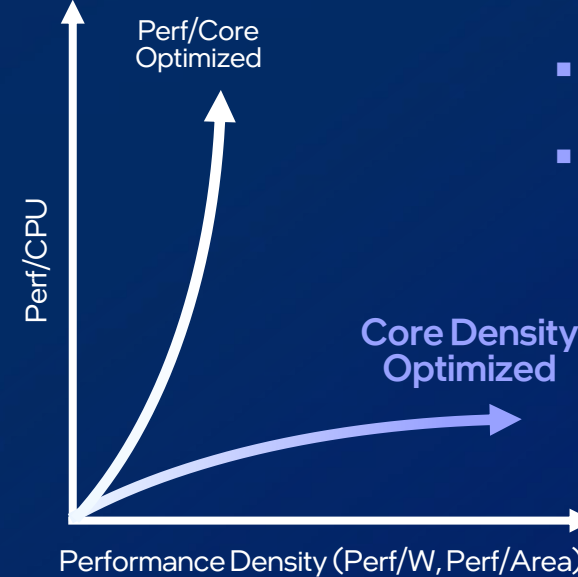
P.core



- High-core performance
- Workload-optimized performance with built-in accelerators

Optimized for Performance

E.core



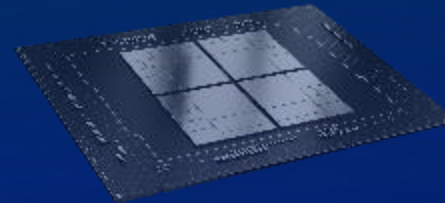
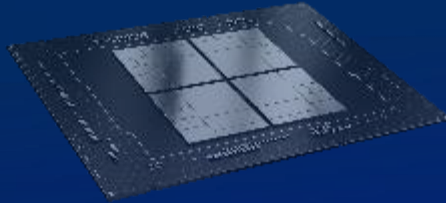
- Performance-per-watt optimized
- High-core density
- High-throughput performance

Optimized for Efficiency

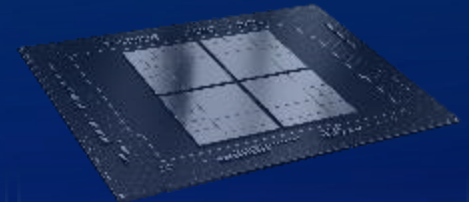
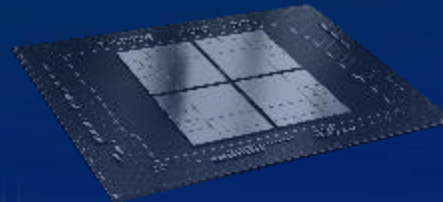
Executing on Our Xeon Roadmap

intel.
XEON

CPU P-Core



CPU E-Core



4th Gen Intel® Xeon® Scalable processors

5th Gen Intel® Xeon® codenamed Emerald Rapids

Next-Gen Intel® Xeon® codenamed Sierra Forest

Next-Gen Intel® Xeon® codenamed Granite Rapids

Next-Gen Intel® Xeon® codenamed Clearwater Forest

Today

Q4 2023

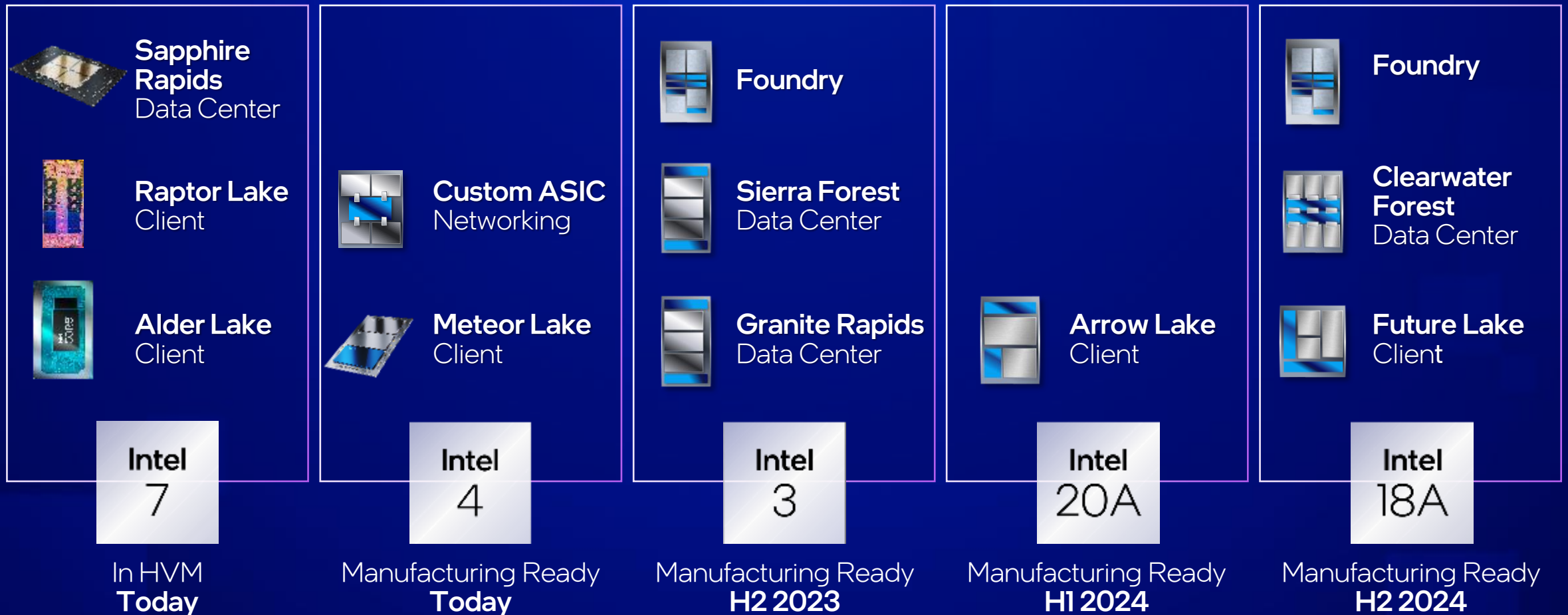
2024
(First Half)

2024
(closely following
Sierra Forest)

2025

intel.

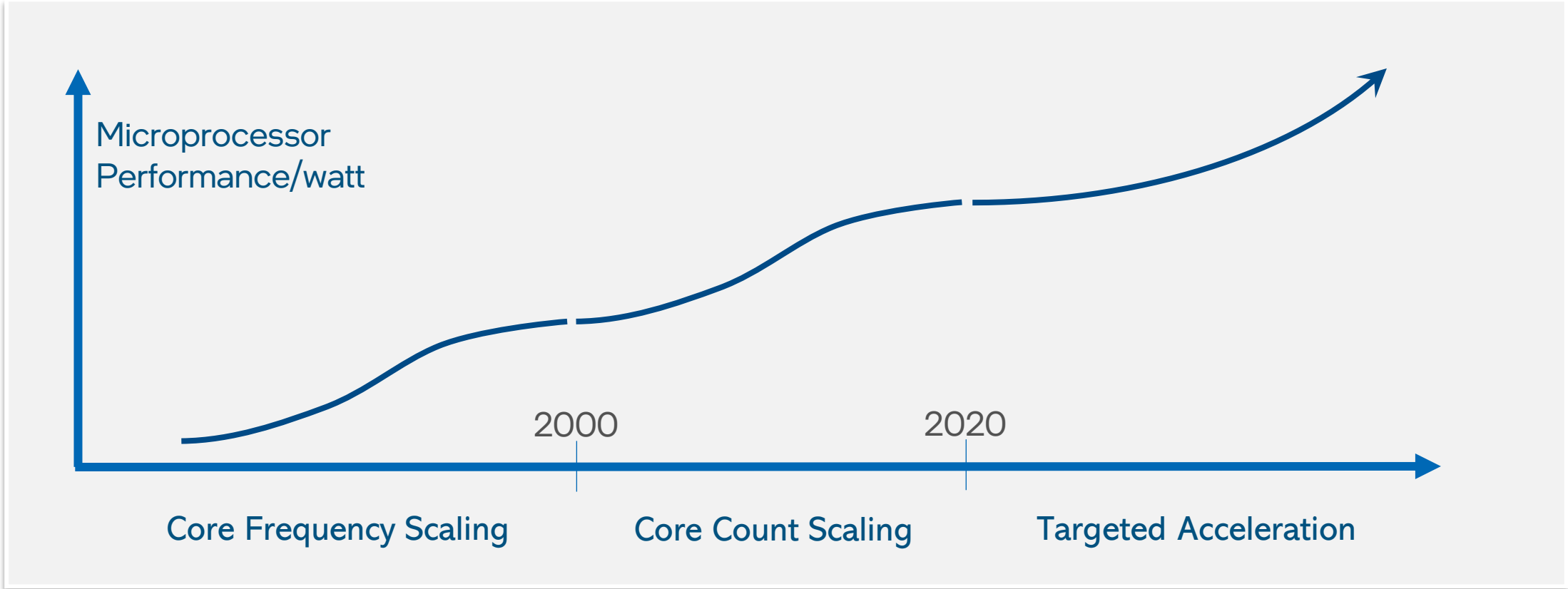
Intel & Moore's Law : Silicon Technology



Between Now & 2025

* Select Products Shown. Based on internal estimates. Technology readiness timing does not necessarily indicate product production timing. Learn more at www.intel.com/PerformanceIndex.

Achieving gains in performance and efficiency requires looking beyond CPU cores performance



Note: Constraints not Intel specific – learnings from ASIC teams at multiple CSPs

Benefits of Intel® Accelerator Engines

A Higher Performance Server Architecture

Intel® Advanced Matrix Extensions
(Intel® AMX)

Up to

8.6x

higher speech recognition inference performance with built-in AMX BF16 vs. FP32

Intel® Dynamic Load Balancer
(Intel® DLB)

Up to

96%

lower latency at the same throughput for Istio-Envoy Ingress with Intel® DLB vs. software for Istio Ingress gateway

Intel® Data Streaming Accelerator
(Intel® DSA)

Up to

1.7x

higher IOPs for SPDK-NVMe with built-in Intel® DSA vs. ISA-L software

Intel® In-Memory Analytics Accelerator
(Intel® IAA)

Up to

2.1x

higher RocksDB performance with Intel® IAA vs Ztsd software

Intel® QuickAssist Technology
(Intel® QAT)

Up to

84%

fewer cores to achieve same connections/s on NGINX with built-in QAT vs. out-of-box software

Accelerators Enable Step Function Performance Beyond Base Architecture

Developer Tools for 4th Gen Intel® Xeon® Scalable Processors

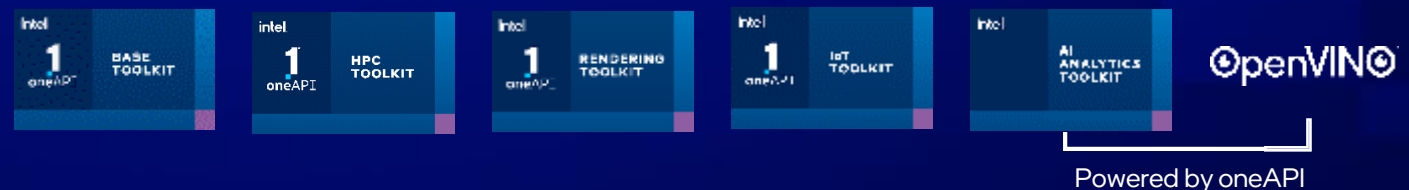
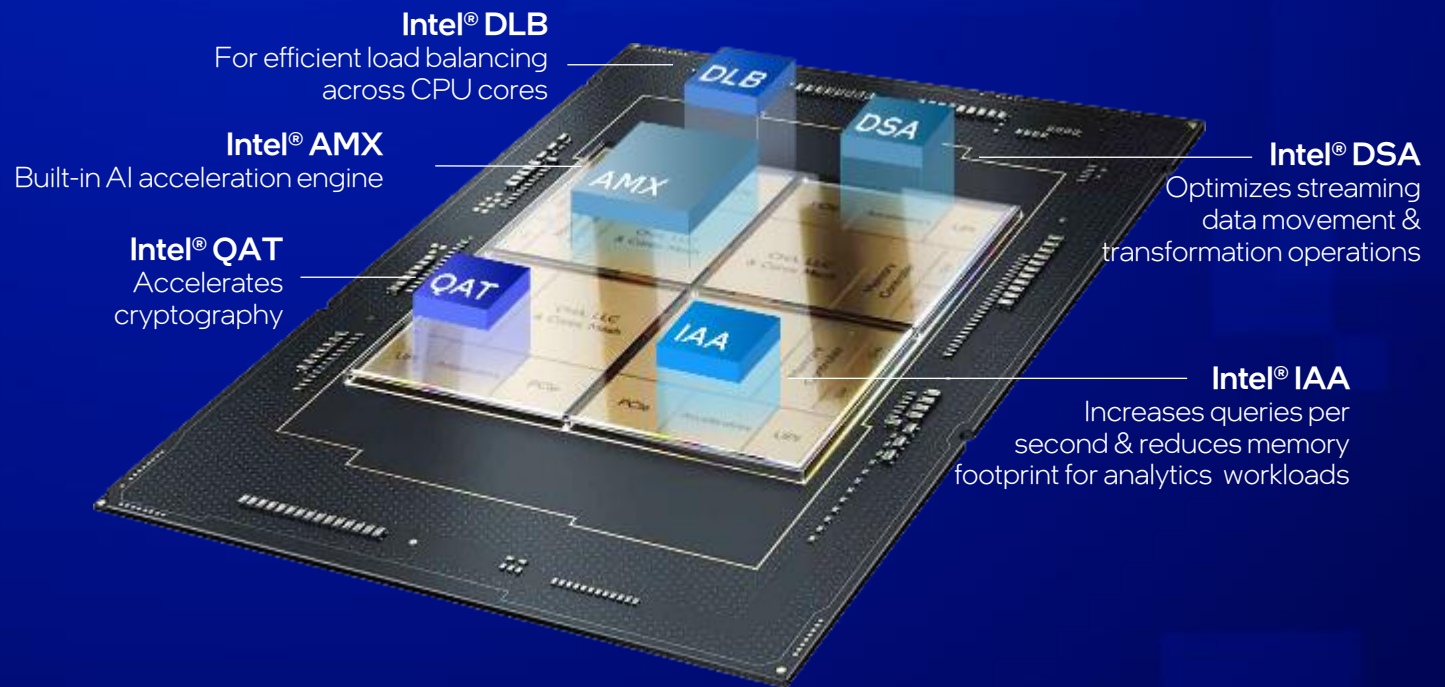
Intel® oneAPI, AI tools and optimized AI frameworks help developers maximize application performance by activating advanced capabilities of 4th Gen Intel® Xeon® Scalable processors and Intel® Max Series processors. In multiarchitecture systems with Intel Xeon processors and Intel GPUs, using a single codebase through [oneAPI](#) delivers productivity and performance.

[Compilers, libraries & analysis tools](#) support built-in accelerators to unleash performance, and fast training and inference for AI workloads.

- **Intel® oneAPI Math Kernel Library**
for HPC and technical compute
- **Intel® oneAPI Deep Neural Network Library**
for deep learning training + inference
- **Intel® Query Processing & Intel® Data Mover Library***
for query processing, compression and data movement
- **Intel® VTune™ Profiler**
helps locate time-consuming parts of code and identify significant issues affecting application performance

Learn more: [Software for 4th Gen Intel Xeon & Max Series Processors](#)

*Intel® OPL is open source. Open source Intel® DML in beta, v1 coming soon



5th Gen Intel[®] Xeon[®] Scalable processor

Grow and excel with workload-optimized performance

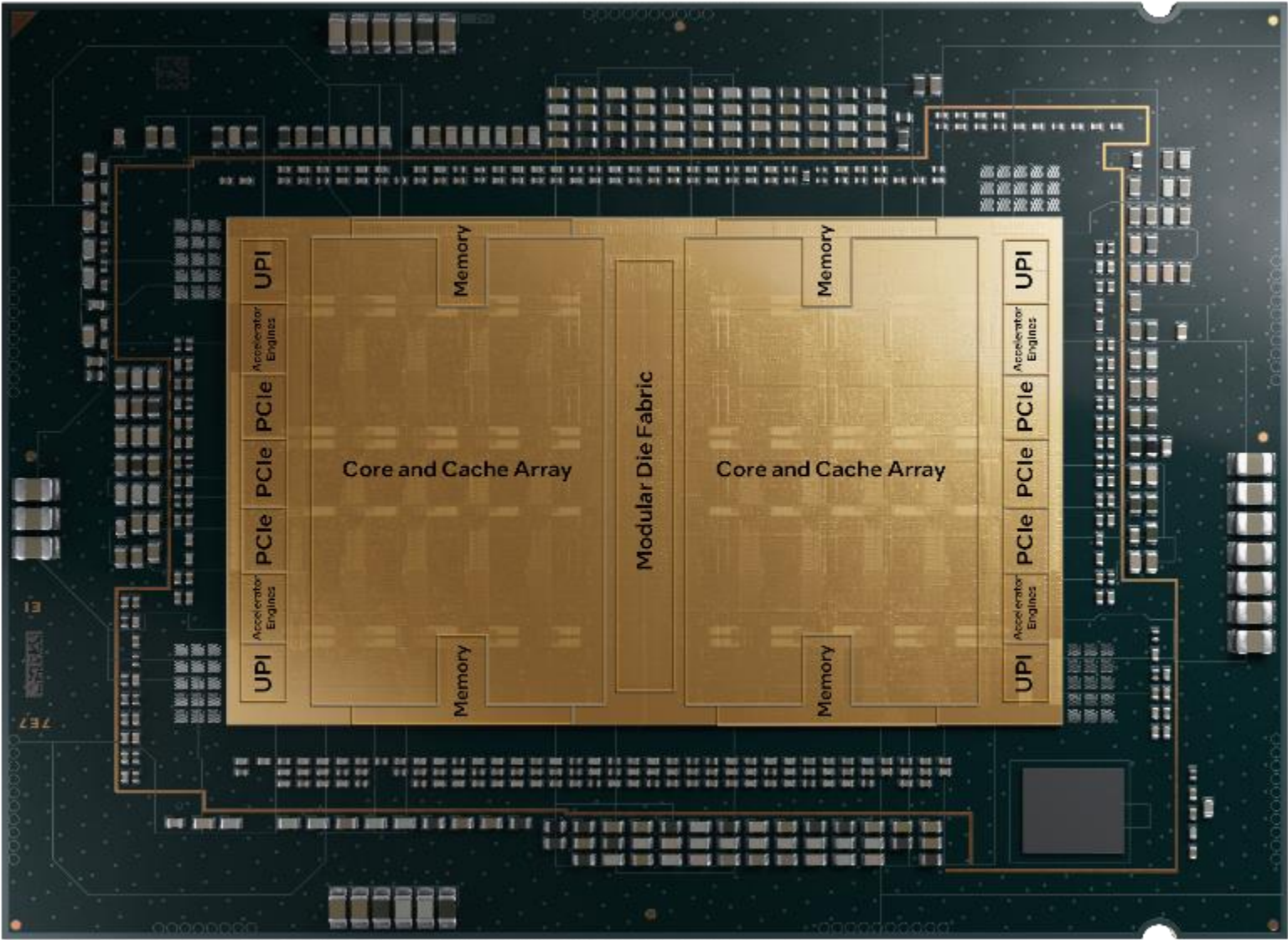
Run AI everywhere with the best CPU for AI

Lower your costs and carbon footprint with increased energy-efficient compute^{1a}

Trusted, quality solutions and security features



5th Gen Intel[®] Xeon[®] processor | Key architecture blocks



5th Gen Intel[®] Xeon[®] Scalable processors deliver

Growth with workload-optimized performance

Increased performance^{1a} & perf/watt^{1a} with higher core counts & larger shared L3cache

Compute Express Link (CXL)
Type 1, 2 & 3*

Increased Intel[®] Ultra Path Interconnect (Intel[®] UPI) speeds

PCIe 5
80 lanes

Intel[®] Accelerator Engines

Increased memory speeds

Bringing AI Everywhere

Built in AI accelerators on every core, Intel[®] Advanced Matrix Extensions (Intel[®] AMX)

Intel AI software suite of optimized open-source frameworks and tools

Out of the box AI performance and E2E productivity with 300+ models validated

Lowering costs and carbon footprint with energy-efficient compute

Optimized Power Mode (OPM) 2.0

Improved TCO and Perf/\$ gains^{1a}

Server refresh opportunities

Workload Optimized SKUs

Built-in accelerators for efficient compute

Reduced downtime with Seamless firmware updates

Trusted quality solutions and security features

Application isolation with Intel[®] Software Guard Extensions (Intel[®] SGX)

Virtual machine isolation with Intel[®] Trust Domain Extensions (Intel[®] TDX)

Achieve performance gains with 5th Gen Intel[®] Xeon[®] CPUs

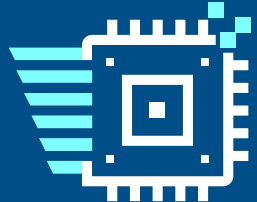
Platform enhancements driving generational gain



Web
Server-side Java

Up to
1.2x²


higher Java Throughput
on 5th Gen Xeon 8592+
vs 4th Gen Xeon 8480+



HPC
LAMMPS - Copper

Up to
1.3x⁴

higher performance
on 5th Gen Xeon 8592+
vs 4th Gen Xeon 8480+



Media
Transcode (FFMPEG)

Up to
1.2x⁶

aggregate FPS
on 5th Gen Xeon 8592+
vs 4th Gen Xeon 8480+

Based on preproduction estimates comparing 4th Gen Intel Xeon processors to 5th Gen Intel Xeon processors. See backup for workloads and configurations. Results may vary

Achieve Real-Workload Performance Gains with Intel® Accelerator Engines

5th Gen Intel® Xeon® processors + built-in accelerators



AI

Natural Language Processing

Up to

1.4x^{7a}

higher throughput on 5th Gen Xeon 8592+ vs 4th Gen Xeon 8480+ With Intel® Advanced Matrix Extensions (Intel® AMX)



AI

Recommendation Systems

Up to

1.4x^{7b}

higher throughput on 5th Gen Xeon 8592+ vs 4th Gen Xeon 8480+ With Intel® Advanced Matrix Extensions (Intel® AMX)

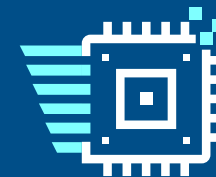


Networks
Secure Gateway

Up to

2.4x⁸

higher performance on 5th Gen Xeon 8592+ with built-in Intel® QuickAssist Technology (Intel® QAT) vs. Native NGINX



Data Movement
Large transfer size

Up to

1.7x⁹

higher performance offload with 5th Gen Xeon 8592+ With Intel® Data Streaming Accelerator (Intel® DSA) vs. SW digest



Databases
Cassandra

Up to

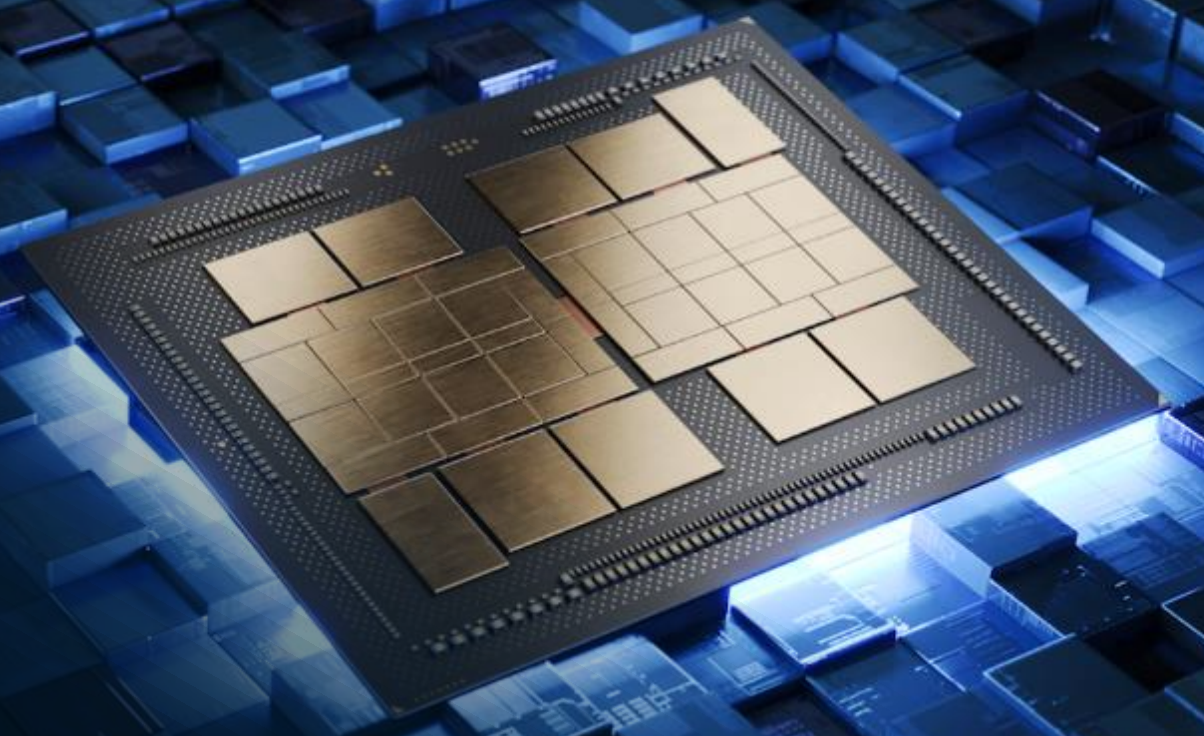
1.3x¹⁰

higher performance with 80/20 read-write on 5th Gen Xeon 8592+ With Intel® In-memory Advanced Accelerator (Intel® IAA) vs. SW compression




Intel® Data Center GPU Max Series

formerly codenamed Ponte Vecchio



Constructed with
EMIB and Foveros

Up to
128
Xe HPC
Cores



52TF
Peak FP64
Throughput

16
Xe Links for
GPU-to-GPU
communication



Highest
Compute Density in a
Socket



Maximize Bandwidth. Maximize Capacity. Maximize Memory.

Up to
128GB
HBM2e

Up to
408MB
Rambo L2
Cache

Up to
64MB
L1 cache

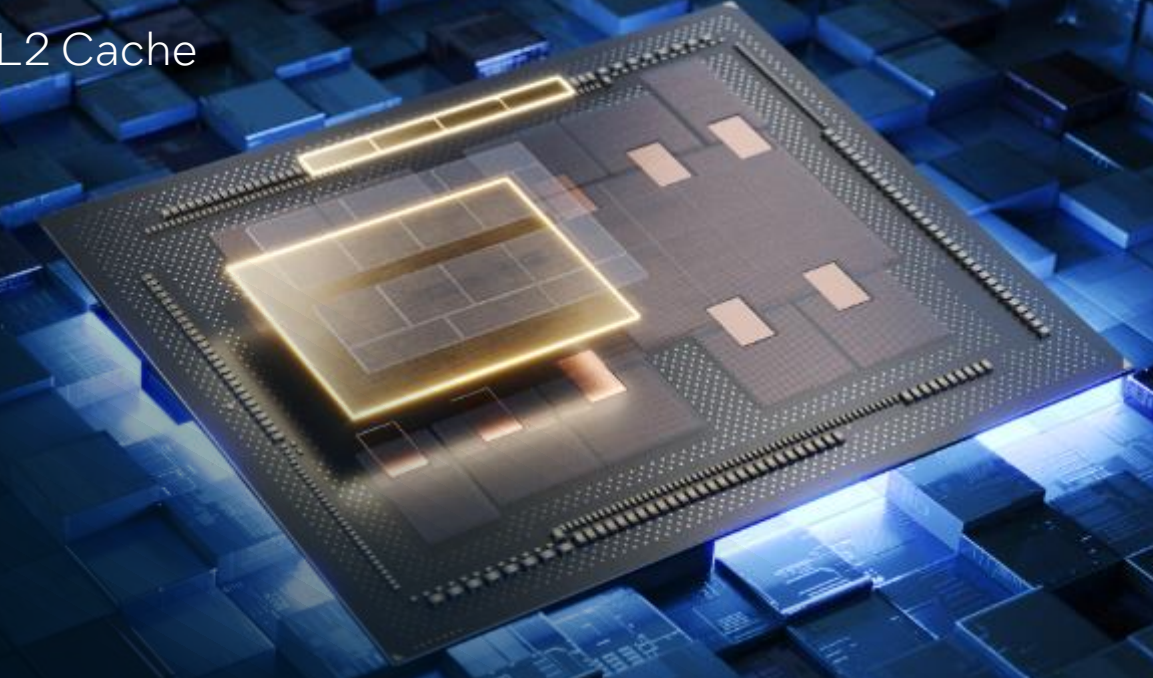
Rambo Cache
(Random Access Memory, Bandwidth
Optimized)

Base Tile



Up to 2x performance

On 4Kx4K 2D-FFT DP due to large L2 Cache



Intel Xe Matrix Extensions

Dedicated built-in AI functionality
Accelerating most AI Data Types

Vector

OPS/CLK/Xe core

256
FP32

256
FP64

512
FP16

Matrix

OPS/CLK/Xe core

2048
TF32

4096
FP16/BF16

8192
INT8



HPC apps run out-of-the-box on Intel® Xeon Max CPU



Growing breadth of applications on Intel® Data Center GPU Max Series



Benchmarks

xGEMMs	STREAM
GUPS	MLBench
HPL	SPECACCEL
SPEChpc	SpMV

Libraries & Frameworks

ELPA	HeFFTe
PETSc	Ginkgo
HYPRE	

Oil & Gas

ISO3DFD
SPECFEM3D GLOBE

Performance Portability

AMReX
Kokkos
RAJA
OCCA
YAKL

Physics & Manufacturing

MFIX-Exa	Gem
HOTQCD	miniFE
Chroma	NEKBONE
MILC	OpenMC
NekRS	CloverLeaf
Shift	BookLeaf
PHASTA	TeaLeaf
NAQMD	GRID QCD
HACC	QUDA
GENE	XGC

Financial Services

Binomial Options
Black-Scholes
Monte Carlo
STAC-A2
Riskfuel Credit Opts Pricing

Earth System Model.

SW4	SeisSol
E3SM	

Bio & Chem Sciences

Autodock	NWChemEx
CP2K	miniBUDE
LAMMPS	OpenMM
GROMACS	ParSplice
NAMD	QMCPack
NWChem	Relion

AI & Emerging WLS

Candle ML
Computational-Catalysis ML
Atlas ML
MatSci ML
Fusion Energy ML
Connectomics ML

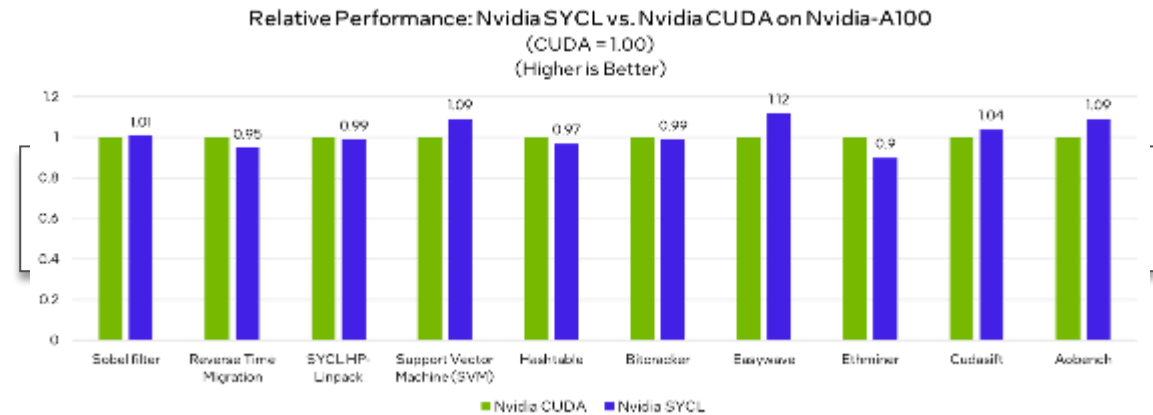
Visualization & Rendering

Blender
OSPRay
Embree 4
Open Shading Library

Accelerating Choice with SYCL

Khronos Group Standard

- Open, standards-based
- Multiarchitecture performance
- Freedom from vendor lock-in
- Comparable performance to native CUDA on Nvidia GPUs
- Extension of widely used C++ language
- Speed code migration via open source [SYCLomatic](#) or Intel® DPC++ Compatibility Tool



Testing Date: Performance results are based on testing by Intel as of August 15, 2022 and may not reflect all publicly available updates.

Configuration Details and Workload Setup: Intel® Xeon® Platinum 8360Y CPU @ 2.4GHz, 2 socket, Hyper Thread On, Turbo On, 256GB Hynix DDR4-3200, ucode 0x000363. GPU: Nvidia A100 PCIe 80GB GPU memory. Software: SYCL open source/CLANG 15.0.0, CUDA SDK 11.7 with NVIDIA-NVCC 11.7.64, cuMath 11.7, cuDNN 11.7, Ubuntu 22.04.1. SYCL open source/CLANG compiler switches: -fsycl-targets=nvptx64-nvidia-cuda, NVIDIA NVCC compiler switches: -O3 -gencode arch=compute_80,code=sm_80. Represented workloads with Intel optimizations.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure.

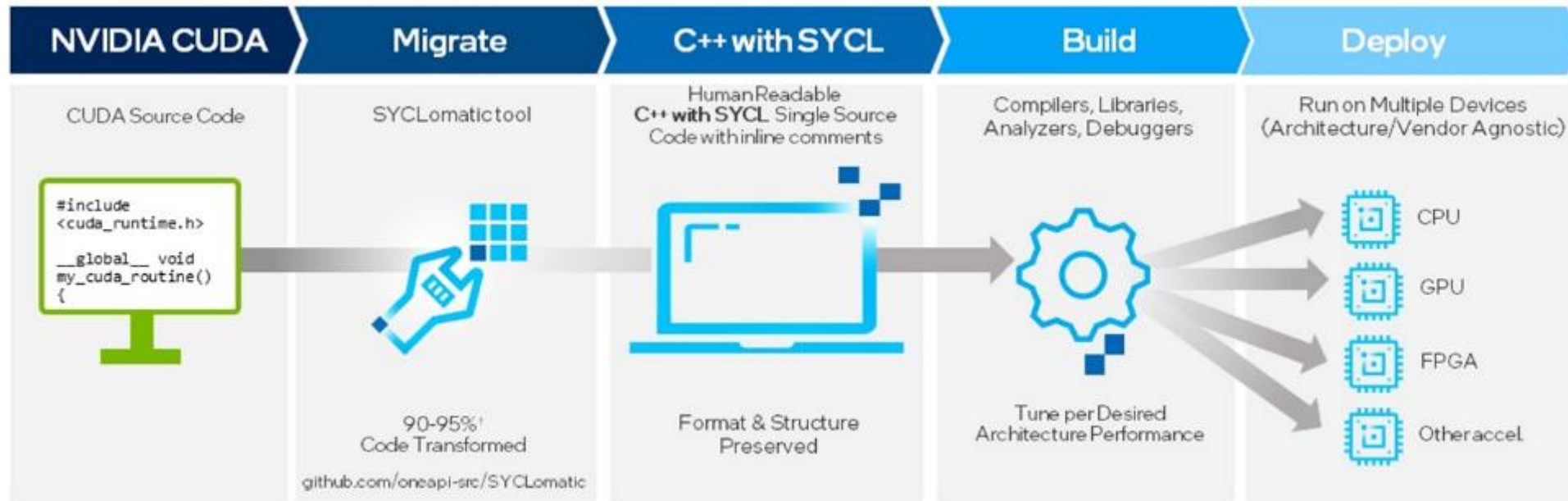
Performance varies by use, configuration, and other factors. Learn more at www.intel.com/PerformanceIndex. Your costs and results may vary.

Architectures

Intel | Nvidia | AMD CPU/GPU | RISC-V | ARM Mali | PowerVR | Xilinx

CUDA to SYCL Migration Made Easy

Open Source SYCLomatic Tool Reduces Code Migration Time



Assists developers migrating code written in CUDA to C++ with SYCL, generating **human readable** code wherever possible

~90-95% of code typically migrates automatically¹

Inline comments are provided to help developers finish porting the application

Intel® DPC++/C++ Compatibility Tool is Intel's implementation, available in the Base Toolkit

¹Intel estimates as of September 2021. Based on measurements on a set of 70 HPC benchmarks and samples, with examples like Rodinia, SHOC, PENNANT. Results may vary.

Codeplay Compiler Plug-ins for Nvidia and AMD GPUs

Adding support for NVIDIA and AMD GPUs to the Intel® oneAPI Base Toolkit

oneAPI for NVIDIA & AMD GPUs

- Free Codeplay download of latest binary plugins to the Intel DPC++/C++ compiler:
 - Nvidia GPU
 - AMD Beta GPU
- Availability at the same time as the Intel oneAPI Base Toolkit
- Plug-ins updated quarterly in-sync with oneAPI

Priority Support

- Sold by Intel and Codeplay and our channel
- Requires Intel Priority support for Intel DPC++/C++ compiler
- Intel takes first call and Codeplay delivers backend support
- Codeplay access to older versions of plugins

[Nvidia GPU plug-in](#)

[AMD GPU plug-in](#)

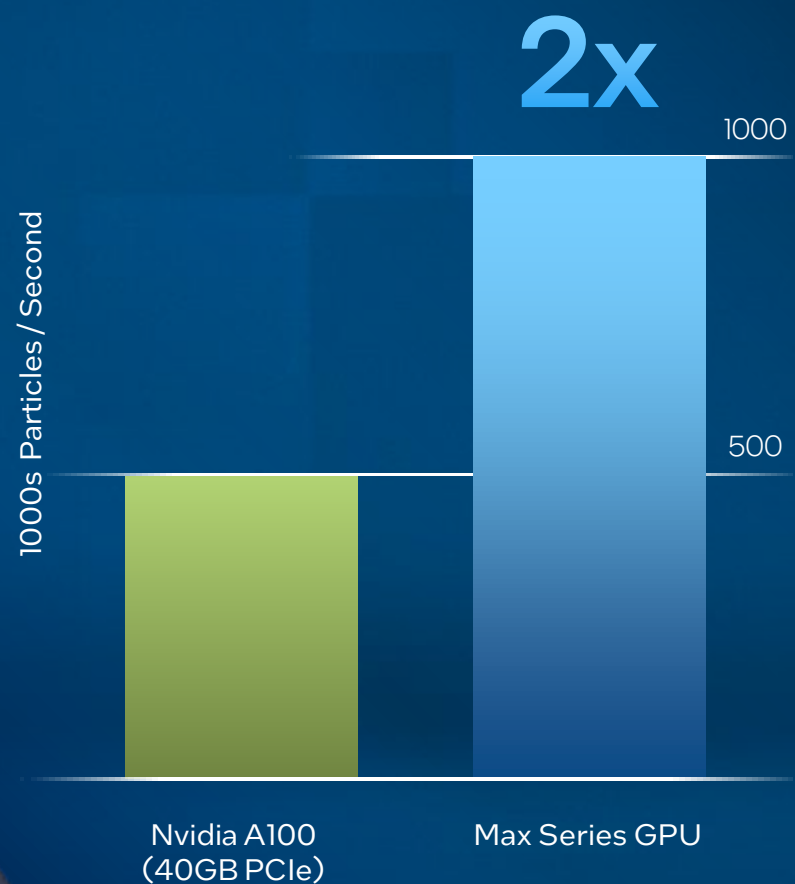
[Codeplay blog](#)

[Codeplay press release](#)



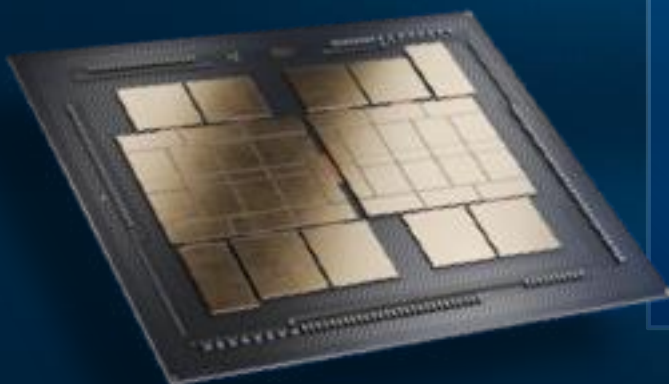
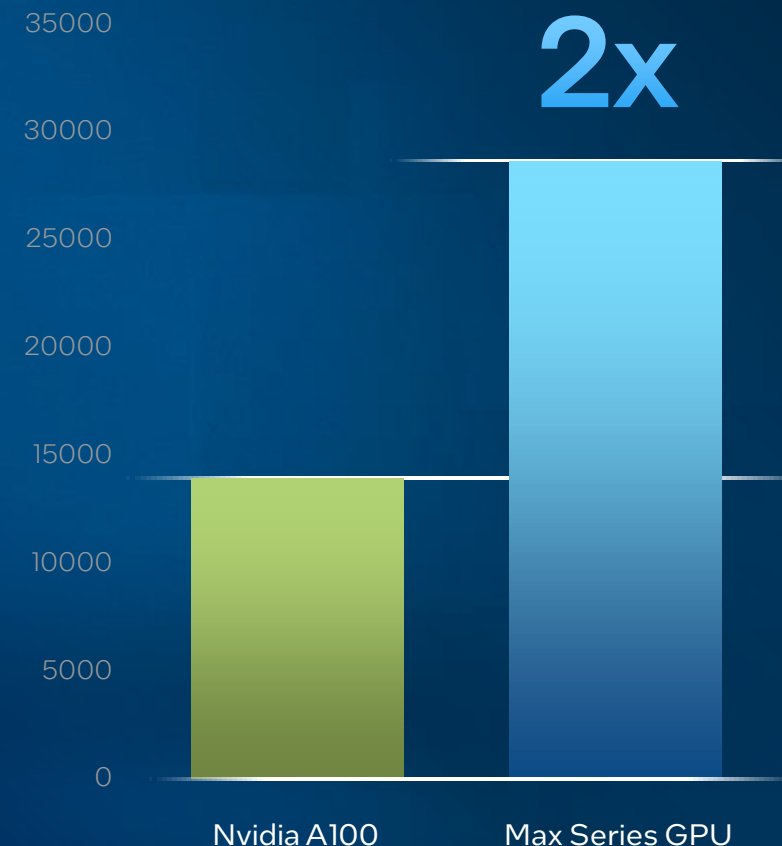
OpenMC

Monte Carlo particle transport code for exascale computations



miniBUDE

Core computation of the Bristol University Engine

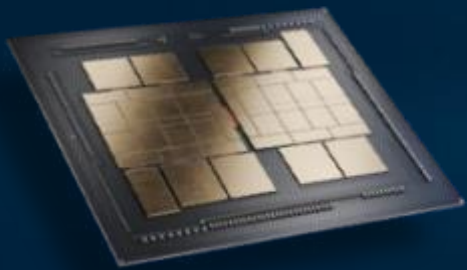


Based on pre-production measurements. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

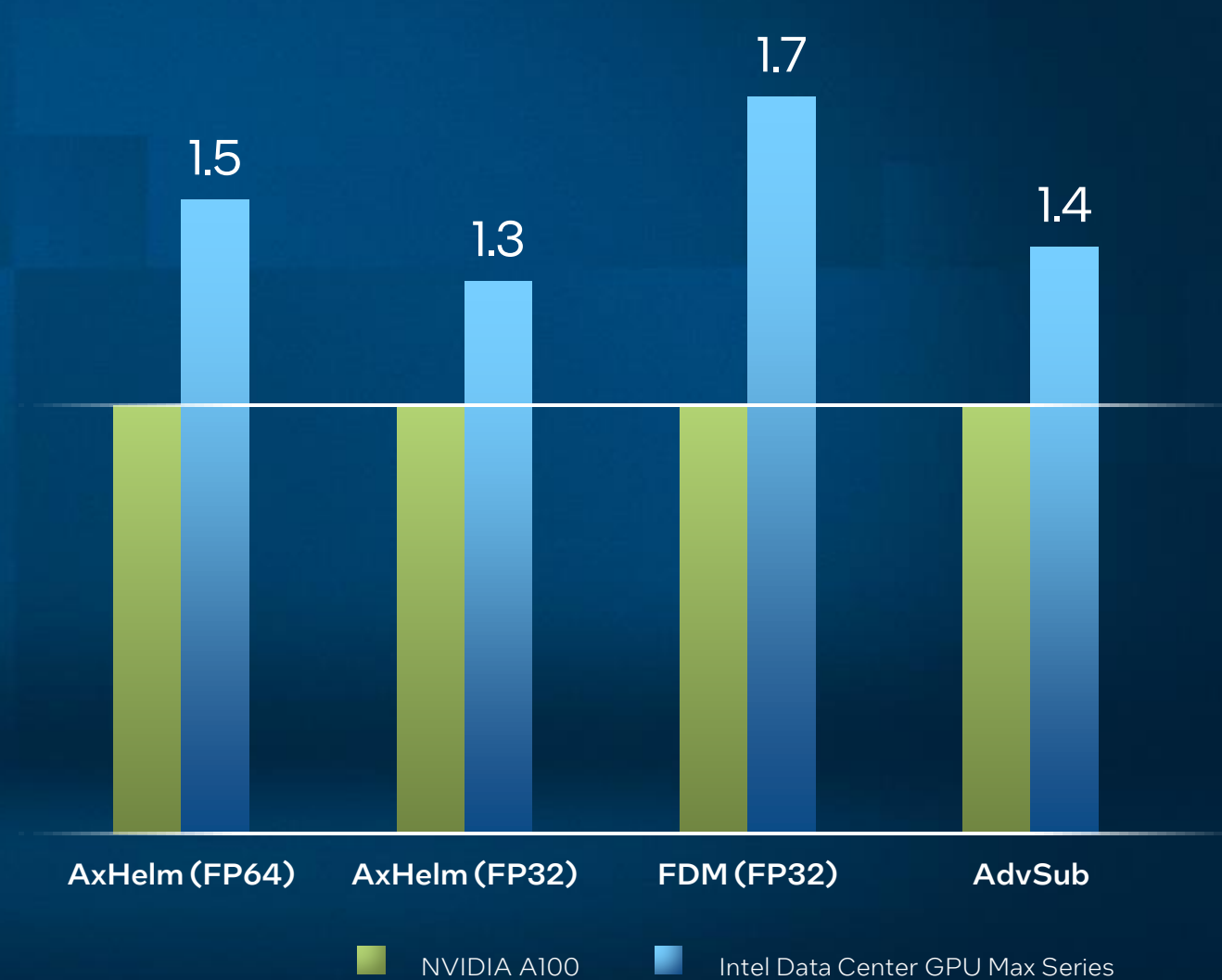


Speeding Up Virtual Reactor Simulation

ExaSMR - NekRS Performance



1.5x performance lead



Relative Performance ((Averaged throughput, higher is better)

Relative Performance of NekRS Benchmarks w/ problem size of 8196. With Intel oneAPI SYCL implementation

Visit the SuperComputing 22 page at [intel.com/performanceindex](https://www.intel.com/performanceindex) for workloads and configurations. Results may vary



The Intel logo is centered on a dark blue background. It features the word "intel" in a white, lowercase, sans-serif font. A small, light blue square is positioned above the letter "i". To the right of the word "intel" is a registered trademark symbol (®).

intel®