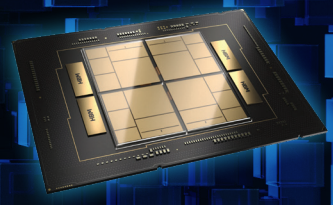
Product Brief

Accelerated Computing Systems and Graphics

Intel® Xeon® CPU Max Series



Over the last decade, peak compute has increased significantly as has the integration of AI, but workload performance hasn't kept pace due to inefficiencies in feeding the cores with data. The Intel® Xeon® CPU Max Series supercharges the Intel® Xeon® platform and is the only x86-based processor with high bandwidth memory architected to unlock and accelerate memory-bound HPC and AI workloads.

From weather forecasting, to human genome mapping and helping to cure the world's deadliest diseases, to designing more energy-efficient materials, high-performance computing (HPC) touches every part of our lives. Advances in HPC and AI drive competitiveness and bring scientific computing demand to new heights, but there is no one-size-fits-all solution. There is incredible diversity in traditional HPC software, and if you look at common workloads by vertical and characteristic, some are memory bound. Others are compute bound. Some have small kernels with a lot of control flow. Others have large, data-parallel kernels. Most involve extremely large data sets.

The Intel® Xeon® CPU Max Series supercharges Intel® Xeon® Scalable processors with high bandwidth memory (HBM) and is architected to unlock performance and speed discoveries in data-intensive workloads, such as modeling, artificial intelligence, deep learning, high performance computing (HPC) and data analytics.

Maximize performance with improved bandwidth

The Intel Xeon CPU Max Series features a new microarchitecture and supports a rich set of platform enhancements, including increased core counts, advanced I/O and memory subsystems, and built-in accelerators that will speed delivery of life-changing discoveries. Intel Max Series CPUs feature:

- Up to **56 performance cores** constructed of four tiles and connected using Intel's embedded multi-die interconnect bridge (EMIB) technology, in a 350-watt envelope.
- **64 GB** of high bandwidth in-package memory, as well as PCI Express 5.0 and CXL 1.1 I/O. Xeon Max CPUs will provide memory (HBM) capacity per core, enough to fit most common HPC workloads.
- Up to **20x performance** speed-up on Numenta AI technology for natural language processing (NLP) with HBM compared to other CPUs.[2]

5x
better performance
for memory bandwidth vs. competition and previous generations[1]

2S Intel Xeon Max CPU vs. 2S AMD EPYC 7773X
2S 3rd Gen Intel® Xeon® 8380

Accelerate scientific innovation

Enable fast discoveries and more effective research. With the Intel Xeon CPU Max Series and 4th Gen Intel Xeon Scalable processors, you gain the performance and power efficiency required for the most challenging workloads and the most built-in accelerators of any CPU on the market. Achieve more efficient CPU utilization, lower electricity consumption and higher ROI with key accelerators for HPC and AI workloads, including:

- **Intel Advanced Matrix Extensions (Intel AMX)** — Significantly accelerate deep learning inference and training on the CPU with Intel® AMX, which boosts AI performance and delivers 8x peak throughput over AVX-512 for INT8 with INT32 accumulation operation.[3]
- **Intel Data Streaming Accelerator (Intel DSA)** — Drive high performance for data-intensive workloads by improving streaming data movement. With Intel® DSA, achieve up to 79% higher storage I/O per second (IOPS) with as much as 45% lower latency when using NVMe over TCP.[4]
- **Intel Advanced Vector Extensions 512 (Intel AVX-512)** — Accelerate performance with vectorization to contribute faster calculations on larger data sets for scientific simulations, AI/deep learning, 3D modeling and analysis, and other intensive workloads. Intel® AVX-512 is the latest x86 vector instruction set to accelerate performance for your most demanding computational tasks.
- **I/O and memory subsystem** advancements including:
  - **DDR5** — Improve compute performance by overcoming data bottlenecks with higher memory bandwidth. DDR5 offers up to 1.5x bandwidth improvement over DDR4.[4]
  - **PCI Express Gen 5 (PCIe 5.0)** — Unlock new I/O speeds with opportunities to enable the highest possible throughput between the CPU and devices. 4th Gen Intel Xeon Scalable and Intel Xeon Max Series processors have up to 80 lanes of PCIe 5.0, double the I/O bandwidth of PCIe 4.0.[4]
  - **Compute Express Link (CXL) 1.1** — Gain support for high-fabric bandwidth and attached accelerator efficiency.
- **Easy integration on Intel Xeon platforms** — Easily add Max Series CPUs to 4th Gen Intel Xeon Scalable platforms by leveraging the same socket configuration resulting in no code changes on most deployments.

Flexibility for all your HPC and AI workloads

Intel Max Series CPUs offer flexibility to run in different memory modes, or configurations, depending on the workload characteristics:

- **HBM-Only Mode** — Enabling workloads that fit in 64GB of capacity and ability to scale at 1-2 GB of memory per core, HBM-Only mode supports system boots with no code changes and no DDR.
- **HBM Flat Mode** — Providing flexibility for applications that require large memory capacity, HBM Flat mode provides a flat memory region with HBM and DRAM and can be applied on workloads requiring >2 GB of memory per core. Code changes may be needed.
- **HBM Cache Mode** — Designed to improve performance for workloads >64GB capacity or requiring >2GB of memory per core. No code changes required, and HBM caches DDR.

Intel® Xeon® Max CPU Series	
Core Count	32-56
HBM2E Memory	64 GB
Peak HBM transfer rate	3200 MT/s
Peak DDR5 transfer rate	4800 MT/s (1 DPC) 4400 MT/s (2 DPC)
Accelerators	AMX, 4 DSA Devices
AI/ML Instructions	INT8 and BFLOAT16

Accelerate HPC and AI Workloads across multiple architectures

The entire Intel Max Series family of products is unified by oneAPI for a common, open, standards-based programming model that unleashes productivity and performance. Developers can build, analyze, optimize and scale general compute, HPC and AI applications across multiple types of architectures more easily using the Intel oneAPI Base Toolkit and Intel oneAPI HPC plus domain-specific toolkits. These resources include state-of-the-art techniques in vectorization, multithreading, multi-node parallelization and memory optimization, so you can easily build high-performance, multiarchitecture software that's ready for HPC. For the latest HPC software developer tools, visit the Software for 4th Gen Intel Xeon & Intel Xeon CPU Max Series Processors and HPC Software and Tools resource pages.

Intel Xeon CPU Max Series Processors

SKU Number	Cores	Base (GHz)	All-Core Turbo (GHz)	Max Turbo (GHz)	Cache (MB)	TDP (Watts)	Maximum Scalability	DDR5 Memory Speed	Intel® SGX Enclave Capacity (Per Processor)
9480	56	1.9	2.6	3.5	112.5	350	2S	4800	512 GB
9470	52	2	2.7	3.5	105	350	2S	4800	512 GB
9468	48	2.1	2.6	3.5	105	350	2S	4800	512 GB
9460	40	2.2	2.7	3.5	97.5	350	2S	4800	128 GB
9462	32	2.7	3.1	3.5	75	350	2S	4800	128 GB

For the most up-to-date information, visit Intel.com/MaxSeriesCPU

[1] Visit intel.com/performanceindex (Events: Supercomputing 22) for workloads and configurations. Results may vary.

[2] Numenta BERT-Large

- AMD Milan: Tested by Numenta as of 11/28/2022. 1-node, 2x AMD EPYC 7R13 on AWS m6a.48xlarge, 768 GB DDR4-3200, Ubuntu 20.04 Kernel 5.15, OpenVINO 2022.3, BERT-Large, Sequence Length 512, Batch Size 1
- Intel® Xeon® 8480+: Tested by Numenta as of 11/28/2022. 1-node, 2x Intel® Xeon® 8480+, 512 GB DDR5-4800, Ubuntu 22.04 Kernel 5.17, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 512, Batch Size 1
- Intel® Xeon® Max 9468: Tested by Numenta as of 11/30/2022. 1-node, 2x Intel® Xeon® Max 9468, 128 GB HBM2e 3200 MT/s, Ubuntu 22.04 Kernel 5.15, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 512, Batch Size 1

[3] Source: AMX Performance-core statement and details at Intel. "Performance Index: Architecture Day 2021." edc.intel.com/content/www/tw/zh/products/performance/benchmarks/architecture-day-2021/.

[4] Intel® DSA. Estimated performance on pre-production configuration: 1-node, 2x 4th Intel® Xeon® Scalable processor (XCC 48C, 350W with 1 Intel® DSA device); pre-production system formerly codenamed Archer City; CPU: SPR E0; memory: 512GB (16x32GB 4800 MT/s [4800 MT/s]) total DDR5 memory; HT on, Turbo on, ucode 0x8e0001a0; BIOS version: EGSDCRB1.86B.0072.D01.2201101353; Ubuntu 21.04; workload: SPDK v22.01 NVMe over TCP (FIO benchmark); block sizes: 4K, 16K, and 128K, random reads; gcc version (Ubuntu 10.3.0-1ubuntu1~21.04) 10.3.0; IDXD driver; IDXD-CONFIG-ACCEL-CONFIG-V3.4.5; NIC: Ethernet Controller E810-C for QSFP; storage – NVMe 10 x Intel® SSD P5510 (5 + 5 balance across socket); test by Intel in March 2022.

Learn more on at intel.com/processorclaims. Performance varies by use, configuration and other factors. Results may vary.

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

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Intel technologies may require enabled hardware, software or service activation.