January 10, 2023

Intel Launches 4th Gen Xeon Scalable Processors, Max Series CPUs and GPUs

Intel highlights broad industry adoption across all major CSPs, OEMs, ODMs and ISVs, and showcases increased performance in AI, networking and high performance computing.

NEWS HIGHLIGHTS

- Expansive customer and partner adoption from AWS, Cisco, Cloudera, CoreWeave, Dell Technologies, Dropbox, Ericsson, Fujitsu, Google Cloud, Hewlett Packard Enterprise, IBM Cloud, Inspur Information, IONOS, Lenovo, Los Alamos National Laboratory, Microsoft Azure, NVIDIA, Oracle Cloud, OVHcloud, phoenixNAP, RedHat, SAP, SuperMicro, Telefonica and VMware, among others.
- With the most built-in accelerators of any CPU in the world for key workloads such as AI, analytics, networking, security, storage and high performance computing (HPC), 4th Gen Intel Xeon Scalable and Intel Max Series families deliver leadership performance in a purpose-built workload-first approach.
- 4th Gen Intel Xeon Scalable processors are Intel's most sustainable data center processors, delivering a range of features for optimizing power and performance, making optimal use of CPU resources to help achieve customers' sustainability goals.
- When compared with prior generations, 4th Gen Xeon customers can expect a 2.9x[1] average performance per watt efficiency improvement for targeted workloads when utilizing built-in accelerators, up to 70-watt[2] power savings per CPU in optimized power mode with minimal performance loss for select workloads and a 52% to 66% lower total cost of ownership (TCO)[3].

SANTA CLARA, Calif.--(BUSINESS WIRE)-- Intel today marked one of the most important product launches in company history with the unveiling of 4th Gen Intel® Xeon® Scalable processors (code-named Sapphire Rapids), the Intel® Xeon® CPU Max Series (code-named Sapphire Rapids HBM) and the Intel® Data Center GPU Max Series (code-named Ponte Vecchio), delivering for its customers a leap in data center performance, efficiency, security and new capabilities for AI, the cloud, the network and edge, and the world's most powerful supercomputers.

This press release features multimedia. View the full release here: https://www.businesswire.com/news/home/20230110005454/en/

Working alongside its customers and partners with 4th Gen Xeon, Intel is delivering differentiated solutions and systems at scale to tackle their biggest computing challenges. Intel's unique approach to providing purpose-built, workload-first acceleration and highly optimized software tuned for specific workloads enables the company to deliver the right performance at the right power for optimal overall total cost of ownership.

On Jan. 10, 2023, Intel introduced 4th Gen Intel Xeon Scalable processors, expanding on its purpose-built, workload-first strategy and approach. (Credit: Intel Corporation)

Press Kit: 4th Gen Xeon Scalable Processors

Additionally, as Intel's most sustainable data center processors, 4th Gen Xeon processors deliver customers a range of features for managing power and performance, making the optimal use of CPU resources to help achieve their sustainability goals.

"The launch of 4th Gen Xeon Scalable processors and the Max Series product family is a pivotal moment in fueling Intel's turnaround, reigniting our path to leadership in the data center and growing our footprint in new arenas," said Sandra Rivera, Intel executive vice president and general manager of the Data Center and AI Group. "Intel's 4th Gen Xeon and the Max Series product family deliver what customers truly want – leadership performance and reliability within a secure environment for their real-world requirements – driving faster time to value and powering their pace of innovation."

Unlike any other data center processor on the market and already in the hands of customers today, the 4th Gen Xeon family greatly expands on Intel's purpose-built, workload-first strategy and approach.

Leading Performance and Sustainability Benefits with the Most Built-In Acceleration

Today, there are over 100 million Xeons installed in the market – from on-prem servers running IT services, including new as-a-service business models, to networking equipment managing Internet traffic, to wireless base station computing at the edge, to cloud services.

Building on decades of data center, network and intelligent edge innovation and leadership, new 4th Gen Xeon processors deliver leading performance with the most built-in accelerators of any CPU in the world to tackle customers' most important computing challenges across AI, analytics, networking, security, storage and HPC.

When comparing with prior generations, 4th Gen Intel Xeon customers can expect a 2.9x[1] average performance per watt efficiency improvement for targeted workloads when utilizing built-in accelerators, up to 70-watt[2] power savings per CPU in optimized power mode with minimal performance loss, and a 52% to 66% lower TCO[3].

Sustainability

The expansiveness of built-in accelerators included in 4th Gen Xeon means Intel delivers platform-level power savings, lessening the need for additional discrete acceleration and helping our customers achieve their sustainability goals. Additionally, the new Optimized Power Mode can deliver up to 20% socket power savings with a less than 5% performance impact for selected workloads[11]. New innovations in air and liquid cooling reduce total data center energy consumption further; and for the manufacturing of 4th Gen Xeon, it's been built with 90% or more renewable electricity at Intel sites with state-of-the-art water reclamation facilities.

Artificial Intelligence

In AI, and compared to previous generation, 4th Gen Xeon processors achieve up to 10x[5,6] higher PyTorch real-time inference and training performance with built-in Intel® Advanced Matrix Extension (Intel® AMX) accelerators. Intel's 4th Gen Xeon unlocks new levels of performance for inference and training across a wide breadth of AI workloads. The Xeon CPU Max Series expands on these capabilities for natural language processing, with customers seeing up to a 20x[12] speed-up on large language models. With the delivery of Intel's AI software suite, developers can use their AI tool of choice, while increasing productivity and speeding time to AI development. The suite is portable from the workstation, enabling it to scale out in the cloud and all the way out to the edge. And it has been validated with over 400 machine learning and deep learning AI models across the most common AI uses cases in every business segment.

Networking

4th Gen Xeon offers a family of processors specificallyoptimized for high-performance, low-latency network and edge workloads. These processors are a critical part of the foundation driving a more software-defined future for industries ranging from telecommunications and retail to manufacturing and smart cities. For 5G core workloads, built-in accelerators help increase throughput and decrease latency, while advances in power management enhance both the responsiveness and the efficiency of the platform. And, when compared to previous generations, 4th Gen Xeon delivers up to twice the virtualized radio access network (vRAN) capacity without increasing power consumption. This enables communications service providers to double the performance-per-watt to meet their critical performance, scaling and energy efficiency needs.

High Performance Computing

4th Gen Xeon and the Intel Max Series product family bring a scalable, balanced architecture that integrates CPU and GPU with oneAPI's open software ecosystem for demanding computing workloads in HPC and AI, solving the world's most challenging problems.

The Xeon CPU Max Series is the first and only x86-based processor with high bandwidth memory, accelerating many HPC workloads without the need for code changes. The Intel Data Center GPU Max Series is Intel's highest-density processor and will be available in several form factors that address different customer needs.

The Xeon CPU Max Series offers 64 gigabytes of high bandwidth memory (HBM2e) on the package, significantly increasing data throughput for HPC and AI workloads. Compared with top-end 3rd Gen Intel® Xeon® Scalable processors, the Xeon CPU Max Series provides up to 3.7 times[10] more performance on a range of real-world applications like energy and earth systems modeling.

Further, the Data Center GPU Max Series packs over 100 billion transistors into a 47-tile package, bringing new levels of throughput to challenging workloads like physics, financial services and life sciences. When paired with the Xeon CPU Max Series, the combined platform achieves up to 12.8 times[13] greater performance than the prior generation when running the LAMMPS molecular dynamics simulator.

Most Feature-Rich and Secure Xeon Platform Yet

Signifying the biggest platform transformation Intel has delivered, not only is 4th Gen Xeon a marvel of acceleration, but it is also an achievement in manufacturing, combining up to four Intel 7-built tiles on a single package, connected using Intel EMIB (embedded multi-die interconnect bridge) packaging technology and delivering new features including increased memory bandwidth with DDR5, increased I/O bandwidth with PCIe5.0 and Compute Express Link (CXL) 1.1 interconnect.

At the foundation of it all is security. With 4th Gen Xeon, Intel is delivering the most comprehensive confidential computing portfolio of any data center silicon provider in the industry, enhancing data security, regulatory compliance and data sovereignty. Intel remains the only silicon provider to offer application isolation for data center computing with Intel® Software Guard Extensions (Intel® SGX), which provides today’s smallest attack surface for confidential computing in private, public and cloud-to-edge environments. Additionally, Intel’s new virtual-machine (VM) isolation technology, Intel® Trust Domain Extensions (Intel® TDX), is ideal for porting existing applications into a confidential environment and will debut with Microsoft Azure, Alibaba Cloud, Google Cloud and IBM Cloud.

Finally, the modular architecture of 4th Gen Xeon allows Intel to offer a wide range of processors across nearly 50 targeted SKUs for customer use cases or applications, from mainstream general-purpose SKUs to purpose-built SKUs for cloud, database and analytics, networking, storage, and single-socket edge use cases. The 4th Gen Xeon processor family is On Demand-capable and varies in core count, frequency, mix of accelerators, power envelope and memory throughput as is appropriate for target use cases and form factors addressing customers’ real-world requirements.

SKU TABLE: SKUs for 4th Gen Xeon and Intel Xeon CPU Max Series

About Intel

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore’s Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers’ greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel’s innovations, go to newsroom.intel.com and intel.com.

[1] Geomean of following workloads: RocksDB (IAA vs ZTD), ClickHouse (IAA vs ZTD), SPDK large media and database request proxies (DSA vs out of box), Image Classification ResNet-50 (AMX vs VNNI), Object Detection SSD-ResNet-34 (AMX vs VNNI), QATzip (QAT vs zlib)

[2] 1-node, Intel Reference Validation Platform, 2x Intel® Xeon 8480+ (56C, 2GHz, 350W TDP), HT On, Turbo ON, Total Memory: 1 TB (16 slots/ 64GB/ 4800 MHz), 1x P4510 3.84TB NVMe PCIe Gen4 drive, BIOS: 0091.D05, (ucode:0x2b0000c0), CentOS Stream 8, 5.15.0-spr.bkc.pc.10.4.11.x86_64, Java Perf/Watt w/ openjdk-11+28_linux-x64_bin, 112 instances, 1550MB Initial/Max heap size, Tested by Intel as of Oct 2022.

[3] ResNet50 Image Classification
New Configuration: 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable 8490H processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), on pre-production SuperMicro SYS-221H-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=Resnet 50 v1_5, best scores achieved: BS1 AMX 1 core/instance (max. 15ms SLA), using physical cores, tested by Intel November 2022. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR, DDR4 memory total 1024GB (16x64 GB), microcode 0xd000375, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=Resnet 50 v1_5, best scores achieved: BS1 INT8 2 cores/instance (max. 15ms SLA), using physical cores, tested by Intel November 2022.
For a 50 server fleet of 3rd Gen Xeon 8380 (RN50 w/DLBoost), estimated as of November 2022:
CapEx costs: $1.64M
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $739.9K
Energy use in kWh (4 year, per server): 44627, PUE 1.6
Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

For a 17 server fleet of 4th Gen Xeon 8490H (RN50 w/AMX), estimated as of November 2022:
CapEx costs: $799.4K
OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $275.3K
Energy use in kWh (4 year, per server): 58581, PUE 1.6

AI -- 55% lower TCO by deploying fewer 4th Gen Intel® Xeon® processor-based servers to meet the same performance requirement. See [E7] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
Database -- 52% lower TCO by deploying fewer 4th Gen Intel® Xeon® processor-based servers to meet the same performance requirement. See [E8] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
HPC -- 66% lower TCO by deploying fewer Intel® Xeon® CPU Max processor-based servers to meet the same performance requirement. See [E9] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

[4] Geomean of HP Linpack, Stream Triad, SPECrate2017_fp_base est, SPECrate2017_int_base est. See [G2, G4, G6] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable.

[5] Up to 10x higher PyTorch real-time inference performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)
PyTorch geomean of ResNet50, Bert-Large, MaskRCNN, SSD-ResNet34, RNN-T, Resnext101.

[6] Up to 10x higher PyTorch training performance with built-in Intel® Advanced Matrix Extensions (Intel® AMX) (BF16) vs. the prior generation (FP32)
PyTorch geomean of ResNet50, Bert-Large, DLRM, MaskRCNN, SSD-ResNet34, RNN-T.

[7] Estimated as of 8/30/2022 based on 4th generation Intel® Xeon® Scalable processor architecture improvements vs 3rd generation Intel® Xeon® Scalable processor at similar core count, socket power and frequency on a test scenario using FlexRAN™ software. Results may vary.

[8] Up to 95% fewer cores and 2x higher level 1 compression throughput with 4th Gen Intel Xeon Platinum 8490H using integrated Intel QAT vs. prior generation.
8490H: 1-node, pre-production platform with 2x 4th Gen Intel® Xeon Scalable Processor (60 core) with integrated Intel QuickAssist Accelerator (Intel QAT), QAT device utilized=8(2 sockets active), with Total 1024GB (16x64 GB) DDR5 memory, microcode 0xf000380, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-47-generic, 1x 1.92TB Intel® SSDSC2KG01, QAT v20.l.0.9.1 , QATzip v1.0.9 , ISA-L v2.3.0, tested by Intel September 2022.
8380: 1-node, 2x 3rd Gen Intel Xeon Scalable Processors(40 cores) on Coyote Pass platform, DDR4 memory total 1024GB (16x64 GB), microcode 0xd000375, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-47-generic, 1x 1.92TB Intel SSDSC2KG01,QAT v1.7.l.4.16, QATzip v1.0.9 , ISA-L v2.3.0, tested by Intel October 2022.

[9] Up to 3x higher RocksDB performance with 4th Gen Intel Xeon Platinum 8490H using integrated Intel IAA vs. prior generation.
8490H: 1-node, pre-production Intel platform with 2x 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0xf000380, 1x 1.92TB INTEL SSDSC2KG01, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.1.21,accel-config-v3.4.6.4, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel September 2022.
8380: 1-node, 2x 3rd Gen Intel Xeon Scalable Processors(40 cores) on Coyote Pass platform, HT On, Turbo On, SNC Off, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x 1.92TB INTEL SSDSC2KG01, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel October 2022.

[10] Intel® Xeon® 8380: Test by Intel as of 10/7/2022. 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, Total Memory 256 GB (16x16GB 3200MT/s DDR4), BIOS Version SE5C620.86B.01.01.0006.2207150335, ucode revision=0xd000375, Rocky Linux 8.6, Linux version 4.18.0-372.26.1.el8_6.crt1.x86_64, YASK v3.05.07
Intel® Xeon® CPU Max Series: Test by Intel as of ww36'22. 1-node, 2x Intel® Xeon® CPU Max SeriesHT On, Turbo On, SNC4, Total Memory 128 GB (8x16GB HBM2 3200MT/s),

BIOS Version SE5C7411.86B.8424.D03.2208100444, ucode revision=0x2c000020, CentOS Stream 8, Linux version 5.19.0-rc6.0712.intel_next.1.x86_64+server, YASK v3.05.07.

[11] Up to 20% system power savings utilizing 4th Gen Xeon Scalable with Optimized Power mode on vs off on select workloads including SpecJBB, SPECINT and NIGNX key handshake.

[12] 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

[13] Intel® Xeon® 8380: Test by Intel as of 10/28/2022. 1-node, 2x Intel® Xeon® 8380 CPU, HT On, Turbo On, Total Memory 256 GB (16x16GB 3200MT/s, Dual-Rank), BIOS Version SE5C6200.86B.0020.P23.2103261309, ucode revision=0xd000270, Rocky Linux 8.6, Linux version 4.18.0-372.19.1.el8_6.crt1.x86_64
Intel® Xeon® CPU Max Series HBM: Test by Intel as of 10/28/2022. 1-node, 2x Intel® Xeon® Max 9480, HT On, Turbo On, Total Memory 128 GB HBM2e, BIOS EGSDCRB1.DWR.0085.D12.2207281916, ucode 0xac000040, SUSE Linux Enterprise Server 15 SP3, Kernel 5.3.18, oneAPI 2022.3.0

Intel® Data Center GPU Max Series with DDR Host: Test by Intel as of 10/28/2022. 1-node, 2x Intel® Xeon® Max 9480, HT On, Turbo On, Total Memory 1024 GB DDR5-4800 + 128 GB HBM2e, Memory Mode: Flat, HBM2e not used, 6x Intel® Data Center GPU Max Series, BIOS EGSDCRB1.DWR.0085.D12.2207281916, ucode 0xac000040, Agama pvc-prq-54, SUSE Linux Enterprise Server 15 SP3, Kernel 5.3.18, oneAPI 2022.3.0

Intel® Data Center GPU Max Series with HBM Host: Test by Intel as of 10/28/2022. 1-node, 2x Intel® Xeon® Max 9480, HT On, Turbo On, Total Memory 128 GB HBM2e, 6x Intel® Data Center GPU Max Series, BIOS EGSDCRB1.DWR.0085.D12.2207281916, ucode 0xac000040, Agama pvc-prq-54, SUSE Linux Enterprise Server 15 SP3, Kernel 5.3.18, oneAPI 2022.3.0

View source version on businesswire.com:
https://www.businesswire.com/news/home/20230110005454/en/

Ann Goldmann
503-702-2412
ann.goldmann@intel.com

Source: Intel Corporation