Accelerate Critical Edge Workloads with Built-in AI and Security Capabilities

4th Gen Intel® Xeon® Scalable processors for IoT edge deliver incredible performance, memory, I/O, resource manageability features to support workload consolidation, and new AI instructions for deep learning training and inference at the edge.

Deploy new levels of performance, connectivity, and accelerated AI with 4th Gen Intel® Xeon® Scalable processors for IoT. Use up to 52 cores[1] to deliver fast performance and benefit from breakthrough memory and I/O capabilities with PCIe 5.0, DDR5 memory and Compute Express Link (CXL) 1.1. Drive powerful AI with a new acceleration engine under the Intel® Deep Learning Boost (Intel® DL Boost) framework called Intel® Advanced Matrix Extensions (Intel® AMX). Address complex edge-server workloads, including video analytics, medical imaging, and industrial machine vision, with greater efficiency, long-life availability,[2] and high-reliability/long-use-life SKUs[3] that help maximize the value from technology investments.

Accelerate advanced workloads with a new AI engine

Design and deploy AI projects faster with more-efficient and optimized training and inference. Intel AMX brings extensive hardware and software optimizations, with support for INT8 and BF16 data types, to enable fast and efficient AI and deep learning for IoT applications in video analytics, industrial machine vision, and visual data edge servers.

DDR5-driven multitasking and boosted data mobility

Ramp up memory capacity and speed with up to eight channels of DDR5 and up to 16 DIMMs per socket and support for 3rd Gen Intel® Optane™ persistent memory. This platform supports up to 4800 MT/s at one DIMM per channel (DPC) or 4400 MT/s at two DPC. The introduction of hardware-enabled Intel® Data Streaming Accelerator improves data mobility between compute, storage, and networking layers, while PCIe 5.0–enabled Compute Express Link (CXL) 1.1 helps ensure coherent interconnects between CPUs and accelerators working on the same data.

Robust platform hardening

Help reduce attack surfaces while strengthening protections for memory and firmware. The 4th Gen platform integrates many of the key technologies that solution providers depend on to help protect customer IP and data from unauthorized access.

- **Intel® Software Guard Extensions (Intel® SGX)** helps protect data in flight by isolating workloads within memory enclaves, which can be effective against privileged malware attacks.
- **Intel® Total Memory Encryption (Intel® TME)** helps secure total system memory, including user credentials and encryption keys, which helps defend against DIMM removal attacks.
- **Intel® Platform Firmware Resilience (Intel® PFR)** monitors system bus traffic, verifies platform integrity, and restores corrupted firmware.

What's new

- New Intel® DL Boost acceleration engine, Intel® AMX, significantly boosts AI training and inference performance with BF16/INT8 data type support
- Support for DDR5 memory—with up to 4,800 MT/s at 1 DPC, 4,400 MT/s at 1 and 2DPC—and up to 16 DIMMs per socket
- Intel® Data Streaming Accelerator (Intel® DSA) delivers fast data-moving operations for storage or networking, and support for CXL 1.1 delivers efficient and coherent CPU-accelerator interconnects
- Up to 80 lanes of PCIe 5.0 connectivity

4th Gen Intel® Xeon® Scalable processors

Performance compared to 3rd Gen Intel® Xeon® Scalable processors.

1.33x higher performance[4]

3.01x higher AI inference performance with Intel® AMX for image classification[5]

4.25x higher AI inference performance with Intel® AMX for object detection[6]

For workloads and configurations, visit intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary

More flexibility and manageability with high VM density

Drive workload consolidation with more-powerful and effective tools. This platform provides a host of capabilities to streamline resource management in highly virtualized environments, enabling high numbers of VMs on consolidated servers.

- **Intel® Resource Director Technology (Intel® RDT)** provides more visibility into shared resources such as last-level cache to help reduce resource contention between applications and VMs.

- **Intel® Speed Select Technology (Intel® SST)** supports platform tuning with multiple configurations per server and the ability to allocate different CPU frequencies to specific workloads, raising or lowering overall TDP.

- **Intel® Scalable I/O Virtualization (Intel® Scalable IOV)** enables scalable sharing of I/O devices across more VMs and containers.

Connect more devices with 80 lanes of PCIe 5.0

Get expandability to support more devices, add-in cards, and accelerators for flexible solution customization. The PCIe 5.0–ready 4th Gen platform supports future I/O needs with industry-leading bandwidth, featuring up to 80 lanes of PCIe 5.0 on the CPU and up to three Intel® Ultra Path Interconnect (Intel® UPI) 2.0 links of up to 16 GT/s. The PCH provides an additional 20 lanes of PCIe 3.0 for 100 lanes total.

Long-life availability with long-use-life SKUs

Customers in multiple industries, including healthcare, the public sector, and manufacturing, depend on a steady supply of inventory, oftentimes for equipment that needs to operate always-on for many years. IoT SKUs of the 4th Gen Intel Xeon Scalable platform provide long-life manufacturing availability to help customers extend the life of their systems and ensure replacements throughout the equipment life cycle.[2] Select SKUs also support high-reliability, long-life use of up to 10 years, up to 100 percent active (no turbo),[3] with support for extended-use temperatures.

Key features

Performance
- Up to 52 cores/sockets on IoT SKUs with the overall Intel® product lineup offering SKUs up to 60 cores[1]
- Socket scalability with support for 1S or 2S platforms[7]

Accelerated AI
- Intel AMX accelerates a variety of DL inference and training workloads with BF16 and INT8 data type support
- Optimized for industry-relevant frameworks and libraries
- Support for the Intel® Distribution of OpenVINO™ toolkit to accelerate AI inferencing workloads

Memory and I/O
- Up to eight channels DDR5 per CPU, up to 16 DIMMs per socket, and up to 4,800 MT/s (1DPC) or 4,400 MT/s (1 and 2DPC)
- Up to 80 lanes of PCIe 5.0 on the CPU with up to three Intel UPI 2.0 links at up to 16 GT/s
- Up to 20 lanes PCIe 3.0 on the PCH with up to 20x SATA 3 ports and up to 10x USB 3.1 Gen 1 (5 Gbps) ports

- CXL 1.1 for coherent interconnects between CPUs and accelerators or external devices
- Validated for Intel® Optane™ Persistent Memory 300 Series

Virtualization and manageability
- Intel SST available on select SKUs for enhanced control over CPU performance to help optimize TCO
- Intel RDT to enable monitoring and control over shared resources
- Intel Scalable IOV for efficient resource allocation across VMs, containers, and functions
- Support for shared virtual memory

Greater security
- Intel SGX for trusted memory enclaves within applications
- Intel TME to encrypt top-security data in memory with a small performance overhead
- Intel PFR for firmware resilience and platform integrity

Storage
- Intel® Volume Management Device 3.0 (Intel® VMD 3.0) for storage aggregation with robust hot-plug capability and LED management
- Intel® Virtual RAID on CPU (Intel® VROC)

Software
- Cross-architecture support with Intel® oneAPI toolkit
- Multiple OS support: Windows Server 2022/2019 LTSC; Windows 10 (CSME or SPS FW); Red Hat Enterprise Linux 8.4 LTS, 8.5, and later; SUSE Enterprise Linux SLE 15 SP3 and later; Ubuntu 21.10, 22.04 LTS and later; Alibaba Cloud Linux; VMware ESXi

Value-maximizing deployments
- Long-life manufacturing availability,[2] supporting products that have extended certification and life cycles
- High-reliability, long-use SKUs[3]

🗐 Use cases

Video: Get to market faster, support faster analytics and more simultaneous streams

Applications: Video/data analytics server, NVR storage server, management services server

- Benefit from fast AI inference performance with a potentially low BOM cost from consolidated hardware, enabled by Intel AMX—a new acceleration engine in Intel DL Boost.
- Up to 80 PCIe 5.0 lanes allow for up to eight AI accelerators, and fast I/O helps eliminate data transfer bottlenecks.
- DDR5 memory helps move data fast for object recognition and analytics workloads.

Healthcare: Support emerging use cases in AI-enabled drug research, genomics sequencing, and medical imaging

Applications: Ultrasound imaging, medical carts, endoscopy, clinical devices

- High performance with DDR5 memory helps improve processing time for scanned medical images.
- Support potential acceleration of vector- and matrix-heavy applications in fields like drug research, biomanufacturing, and genomics with Intel DL Boost, enhanced with new Intel AMX.
- Long-life availability[2] supports medical system procurements with lengthy certification processes.

Retail, banking, hospitality, and education: Capture more revenue across multiple touchpoints

Applications: Small-format retail for POS, digital security, digital signage, video walls

- More cores[1] and higher performance[4] support more virtual desktops for banking and education.
- 80 lanes of PCIe 5.0 at 32 GT/s open up the possibility for single-socket solutions.
- Intel DL Boost with new Intel AMX enables fast and responsive inference-based solutions.

Industrial manufacturing: Run compute-intensive automation in less time and boost factory output

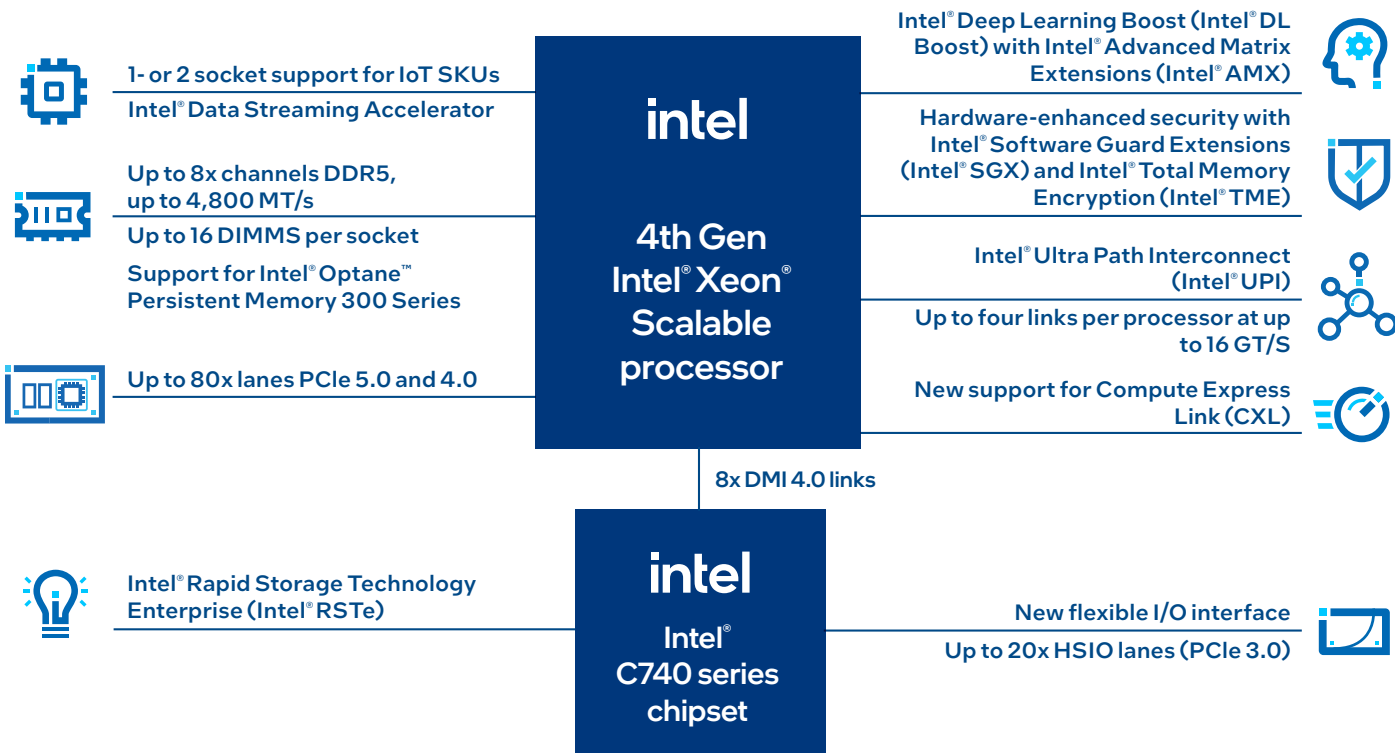**Applications: Assembly line verification, defect detection, human-machine interfaces (HMI)**

- Support more workload convergence in automation and machine vision with up to 52 cores,[1] high-bandwidth DDR5 memory and PCIe 5.0, and enhanced Intel DL Boost with new Intel AMX.
- Optimize control loop timing on x86 hardware with support for Intel RDT and the ability to allocate cache and memory to high-priority tasks.
- Long-life availability[2] and 10-year industrial-commercial temperature use condition[3] (on select SKUs) enhance the reliability of systems.

Public sector: Trusted, consistent experiences for digital safety and reconnaissance applications

Applications: Intelligent surveillance/reconnaissance systems, command and control systems, communication networks

- In recon applications, offload video and AI to the Intel AMX accelerator for rapid processing, and leverage up to 52 cores[1] and high memory bandwidth of up to 4,800 MT/s to enable the creation of high-quality images fast, with clarity and accuracy.
- High DDR5 memory capacity and PCIe 5.0 help deliver fast results for radar data processing with the addition of more accelerators in edge servers and CXL to provide data coherency.
- Hardware-assisted features, such as Intel® TME and Intel SGX, help meet strict security requirements.

Processor block diagram

1- or 2 socket support for IoT SKUs

Intel® Data Streaming Accelerator

Up to 8x channels DDR5, up to 4,800 MT/s

Up to 16 DIMMS per socket

Support for Intel® Optane™ Persistent Memory 300 Series

Up to 80x lanes PCIe 5.0 and 4.0

intel

4th Gen Intel® Xeon® Scalable processor

Intel® Deep Learning Boost (Intel® DL Boost) with Intel® Advanced Matrix Extensions (Intel® AMX)

Hardware-enhanced security with Intel® Software Guard Extensions (Intel® SGX) and Intel® Total Memory Encryption (Intel® TME)

Intel® Ultra Path Interconnect (Intel® UPI)

Up to four links per processor at up to 16 GT/S

New support for Compute Express Link (CXL)

8x DMI 4.0 links

intel

Intel® C740 series chipset

Intel® Rapid Storage Technology Enterprise (Intel® RSTe)

New flexible I/O interface

Up to 20x HSIO lanes (PCIe 3.0)

Software overview

CATEGORY	OPERATING SYSTEMS/SDKs/BOOT LOADERS	IMPLEMENTATION	DISTRIBUTION AND SUPPORT
Operating systems	Red Hat Enterprise Linux 8.4 LTS, 8.5, and later[a]	Red Hat	Red Hat
	SUSE Enterprise Linux SLE 15 SP3 and later[a]	SUSE	SUSE
	Ubuntu 21.10, 22.04 LTS, and later[a]	Canonical, open source	Canonical
	Alibaba Cloud Linux[a]		
	VMware ESXi[b]	VMware	VMware
	Windows Server 2022 Long-Term Servicing Channel (LTSC) – Full compatibility[c]	Intel	Intel, Microsoft
	Windows Server 2019 LTSC/Windows 10 (CSME or SPS FW) – Limited compatibility[d]		
	Azure Stack Hyperconverged Infrastructure (HCI) v21H2 and later – Full compatibility[c]	Intel	Intel, Microsoft
	Azure Stack HCI v20H2 – Limited compatibility[d]		
	Windows 11 (client workstation) (CSME and SPS FW) – Full compatibility[e]	Intel	Intel, Microsoft
SDK	Intel® oneAPI Video Processing Library (Intel® oneVPL)	Intel	Intel
	Intel® Distribution of OpenVINO™ toolkit	Intel	Intel
	Intel® oneAPI Toolkit	Intel	Intel

Not all features are supported in every operating system. Refer to Intel's IoT Solutions Community for partner contact information.

a. Please contact OS vendor for the exact support and version information.

b. Please contact your VMware rep for compatible ESXi versions.

c. This version will have limited support at launch. More features will become available later.

d. Legacy features only. No 4th Gen platform features enabled. Versions will be dropped if the OS vendor discontinues commercial support.

e. Many new OS-supported features will be included out of box.

4th Gen Intel® Xeon® Scalable processors

SKU[A,B,C]	Processor Cores	Base Frequency (GHz)[D]	All-Core Turbo (GHz)	Max Turbo (GHz)	Cache (MB)	TDP (Watts)	Maximum Scalability	DDR5 Memory Speed	Intel® UPI Links Enabled	Default Intel® DSA Devices	Default Intel® QAT Devices	Default Intel® DLB Devices	Default Intel® IAA Devices	Intel® SGX Enclave Capacity (per Processor)	Long-Life Availability[E]	Intel® On Demand Capable
2S Performance General Purpose																
Intel® Xeon® Platinum 8470 processor	52	2	3	3.8	105	350	2S	4800	4	1	0	0	0	512 GB	Yes	Yes
Intel® Xeon® Platinum 8462Y+ processor	32	2.8	3.6	4.1	60	300	2S	4800	3	1	1	1	1	128 GB	Yes	Yes
Intel® Xeon® Gold 6448Y processor	32	2.1	3.0	4.1	60	225	2S	4800	3	1	0	0	0	128 GB	Yes	Yes
Intel® Xeon® Gold 6426Y processor	16	2.5	3.3	4.1	37.5	185	2S	4800	3	1	0	0	0	128 GB	Yes	Yes
Intel® Xeon® Gold 5415+ processor	8	2.9	3.6	4.1	22.5	150	2S	4400	3	1	1	1	1	128 GB	Yes	Yes
2S Mainline General Purpose																
Intel® Xeon® Platinum 8452Y processor	36	2	2.8	3.2	67.5	300	2S	4800	4	1	0	0	0	128 GB	Yes	Yes
Intel® Xeon® Gold 6430 processor	32	2.1	2.6	3.4	60	270	2S	4400	3	1	0	0	0	128 GB	Yes	Yes
Intel® Xeon® Gold 5420+ processor	28	2.0	2.7	4.1	52.5	205	2S	4400	3	1	1	1	1	128 GB	Yes	Yes
Intel® Xeon® Gold 5418Y processor	24	2.0	2.8	3.8	45	185	2S	4400	3	1	0	0	0	128 GB	Yes	Yes
Intel® Xeon® Silver 4416+ processor	20	2.0	2.9	3.9	37.5	165	2S	4000	2	1	1	1	1	64 GB	Yes	Yes
Intel® Xeon® Silver 4410Y processor	12	2.0	2.8	3.9	30	150	2S	4000	2	1	0	0	0	64 GB	Yes	Yes
Single-Socket General Purpose																
Intel® Xeon® Bronze 3408U processor	8	1.8	1.9	1.9	22.5	125	1S	4000	0	1	0	0	0	64 GB	Yes	No
In-Memory Database (IMDB)/Analytics/Virtualization-Optimized and Socket Scalable																
Intel® Xeon® Gold 6418H processor	24	2.1	2.9	4.0	60	185	4S	4800	3	1	0	0	1	512 GB	Yes	No
Long-Life Use (IoT) General Purpose[F]																
Intel® Xeon® Silver 4410T processor	10	2.7	3.4	4.0	26.25	150	2S	4000	2	1	0	0	0	64 GB	Yes	Yes

Intel® UPI: Intel® Ultra Path Interconnect
Intel® DSA: Intel® Data Streaming Accelerator
Intel® QAT: Intel® QuickAssist Technology
Intel® DLB: Intel® Dynamic Load Balancer
Intel® IAA: Intel® In-Memory Analytics Accelerator

Intel® processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families.

All processors are lead free (per EU RoHS directive July 2006) and halogen free (residual amounts of halogens are below November 2007 proposed IPC/JEDEC J-STD-709 standards).

All processors support Intel® Virtualization Technology (Intel® VT-x, VT-d).

A. 6418H long-life-use SKU is also an in-memory database (IMDB) workload–optimized SKU.

B. Y-SKU: Supports Intel® Speed Select Technology Performance Profile (Intel® SST-PP) 2.0.

C. Unless noted, all 8400, 6400, and 5400 processors include support for Intel® Speed Select Technology (Intel® SST) featuring Intel® SST Base Frequency (Intel® SST-BF), Intel® SST Core Power (Intel® SST-CP), and Intel® SST Turbo Frequency (Intel® SST-TF) capabilities. The 5420+ SKU does not have Intel® SST.

D. The frequency of cores and core types varies by workload, power consumption, and other factors. Visit intel.com/content/www/us/en/architecture-and-technology/turbo-boost/intel-turbo-boost-technology.html for more information.

E. Intel does not commit or guarantee product availability or software support by way of road map guidance. Intel reserves the right to change road maps or discontinue products, software, and software support services through standard EOL/PDN processes. Contact your Intel account rep for additional information.

F. High-reliability industrial use supports always-on conditions for up to 10 years.

For more information about Intel® On Demand, visit intel.com/ondemand.

For product specifications, please refer to ark.intel.com.

Get performance, connectivity, and security for video analytics,
edge and industrial servers, and AI/deep learning workload consolidation.
Learn more about 4th Gen Intel Xeon Scalable processors at intel.com/4thgenxeon-iot.