Product Solution Brief

4th Gen Intel® Xeon® Scalable processor
High Performance Computing

High Performance Computing Needs High Performing Technology

With 4th Gen Intel® Xeon® Scalable processors, HPC workloads benefit from high-per-core performance, increased core counts, I/O and memory subsystem advances and a range of built-in hardware accelerators.

The growing confluence of AI with HPC workloads adds a new dimension to familiar performance challenges. Both inference and training at enterprise scale place significant demands on system resources that must be met effectively to deliver a favorable end user experience. These requirements are in addition to the traditional and growing needs of HPC workloads in fields such as life and material sciences, manufacturing, simulation/modeling and finance. As a result, the global HPC market is projected to grow at a 7.7% CAGR to reach $59.2 billion by 2026.[1]

7.7% CAGR through 2026 — **GLOBAL HPC MARKET GROWTH**[1]

$59.2B by 2026 — **GLOBAL HPC SPEND**[1]

Balancing the HPC System

4th Gen Intel® Xeon® Scalable processors provide breakthrough performance for HPC workloads, with fast time to value. The platform features a new architecture with higher per-core performance and up to 60 cores per socket, with two, four, or eight sockets per system. That equates to a per core density of up to 120 threads, a 50% increase over its predecessor.

To balance those core count increases, the platform provides accompanying advances in the memory and I/O subsystems. DDR5 memory provides up to 1.5x the bandwidth and speed of DDR4, for 4800 MT/s. The platform also features 80 lanes of PCIe Gen 5 per socket, for dramatically improved I/O compared to earlier platforms. It provides CXL (Compute Express Link 1.1) to support high fabric bandwidth and attached accelerator efficiency.

Accelerate performance across the fastest-growing workloads. 4th Gen Intel Xeon Scalable processors have the most built-in accelerators of any CPU on the market to improve performance in AI, analytics, networking, storage, and HPC, including the following workloads, among many others:

- **Options pricing.** Address tight decision-making timelines, complex applications with varying requirements and changing market demands with an increased use of AI.
- **Life sciences applications.** Enable fast discoveries and more effective research by refining models and performing large-scale calculations for more accurate simulations.
- **Computer-aided engineering.** Get fast results for the computer-aided engineering applications that help you reduce costs, improve product safety and design and speed time to market.

4th Gen Intel Xeon processors introduce a new paradigm for performance delivery based on built-in hardware accelerators, including those for HPC — Intel® HPC Engines.

intel XEON

PERFORMANCE PROOFPOINT

UP TO **1.56X** GEOMEAN HIGHER PERFORMANCE ACROSS 28 POPULAR HPC WORKLOADS VS PREVIOUS GEN[2,3]

Intel® HPC Engines

PERFORMANCE PROOFPOINT

UP TO **1.68X** GEOMEAN HIGHER LAMMPS WORKLOAD PERFORMANCE VS PREVIOUS GEN[2,3]

Advanced capabilities based on built-in accelerators

As workloads increase in complexity and place increased demands on compute resources, there is an opportunity to offload certain functions from the CPU cores to preserve those execution resources for business-critical workload tasks. These functions include AI, security and common storage and networking functions.

Hardware accelerators built directly into the silicon of 4th Gen Intel Xeon Scalable processors offer capabilities that improve data movement and processing within the platform. Because they are built into the processor, they do not incur the latency of going out to the PCIe bus, with a corresponding savings in energy consumption as well, compared to discrete solutions or software-based ones running on the cores. Use cases that draw on these built-in accelerators can achieve better performance as well as CapEx and OpEx savings:

- **Performance.** Specialized, purpose-built accelerators target delivering significant gains in throughput for their targeted workloads.
- **Equipment costs.** Because the accelerators are built into 4th Gen Intel Xeon Scalable processors, they do not require a separate equipment investment.
- **Operating costs.** By reducing the need for additional cores to be added to equipment racks, built-in accelerators may provide significant energy savings.

Intel Advanced Matrix Extensions (Intel AMX): Accelerated deep learning

Machine learning is proving effective at tuning HPC workloads to be more efficient and effective. Intel AMX is a built-in hardware accelerator that provides a significant leap in the performance of inference and training by speeding up the tensor processing at the heart of deep learning algorithms. The technology includes TILEs, a set of up to eight expandable 2D register tiles per core that store larger chunks of data than predecessors, as well as TMUL (Tile Matrix Multiply), a set of matrix multiplication instructions that are the first operators on TILEs. Intel AMX accelerates time to value by enabling deep learning software to complete more inference in a given period of time or to close in on solutions more quickly.

Intel Advanced Vector Extensions 512 (Intel AVX-512): The latest X86 vector instruction set

Progressively more sophisticated vectorization has contributed to faster calculations on larger data sets over many technology generations. Intel AVX-512, the latest X86 vector instruction set, builds on the vector processing power of its predecessors to accelerate the completion of data-intensive workloads. HPC applications can pack 32 double-precision and 64 single-precision floating point operations per clock cycle within the 512-bit vectors, as well as eight 64-bit and 16 32-bit integers with two 512-bit fused multiply-add (FMA) units for the most demanding computational workloads to drive business intelligence. The technology doubles the width of data registers, number of registers and width of FMA units compared to Intel Advanced Vector Extensions 2 (Intel AVX2).

Intel Data Streaming Accelerator (Intel DSA): Optimized streaming data movement

Data movement and transformation operations are critical to the performance of storage, networking and data-intensive workloads, such as analytics in HPC. Intel DSA drives up performance for these functions by offloading the most common data movement tasks that cause overhead in large-scale deployments. By shouldering almost all data movement operations, including checksum, memory compare and checkpointing, Intel relieves the CPU cores of overhead associated with moving data in and out of memory, storage and networking subsystems. Intel DSA optimizes the handling of streaming data across the CPU, memory and caches, as well as all attached memory, storage and network devices.

Intel QuickAssist Technology (Intel QAT): Accelerated encryption and compression

Reducing the overhead associated with encryption and data compression can play a significant role in improving overall cluster performance. Intel QAT is built in as a hardware accelerator in 4th Gen Intel Xeon Scalable processors that enables faster data encrypt and decrypt on the fly, as well as more efficient data compression. This latest version of the technology accelerates cryptographic ciphers, secure hashes, public key encryption and compression/decompression performance relative to prior generations. By offloading these tasks from the processor cores, Intel QAT frees up resources for other work, increasing overall throughput. Intel QAT contributes to zero-trust security strategies that protect data at all stages in any infrastructure — at rest, in flight and in use — without loss of performance for critical workloads.

KEY TECHNOLOGIES

UP TO 60 CORES	UP TO 8 CHANNELS	UP TO 80 LANES
per socket	DDR5, up to 4800 MT/s	PCIe 5.0 per socket
Up to 50% increase;[2] 2, 4 or 8 sockets per system	Up to 50% increase in memory bandwidth & speed[2]	Increased I/O capacity[2]

4th Gen Intel Xeon processors deliver better floating-point performance per core and a range of built-in hardware accelerators

Developer enablement and support

Intel oneAPI toolkits are an evolution of Intel's long-standing commitment to the HPC software ecosystem, providing compilers, libraries and performance tools that streamline the development path to high-quality software optimized for Intel architecture. The toolkits represent a fast path to adoption for developers as they look to take advantage of the accelerators built into 4th Gen Intel Xeon Scalable processors, with an open, standards-based software development stack. Developers can use the Intel oneAPI toolkits to produce code that delivers accelerated performance across Intel architectures, including CPUs with built-in accelerators, GPUs and FPGAs.

Intel® oneAPI Base Toolkit

Core compilers, libraries (including Intel oneAPI Math Kernel Library) and other tools for developing high-performing data-centric applications

Intel oneAPI HPC Toolkit

Intel Fortran Compilers, OpenMP GPU offload, and scalability with message passing interface (MPI)

Intel AI Analytics Toolkit

Optimized frameworks and Python libraries to accelerate data science and analytics pipelines

Intel oneAPI Rendering Toolkit

Rendering and ray-tracing libraries to create high-performance, high-fidelity visual experiences

Open standards code development based on oneAPI benefits from a large open ecosystem that includes open source tools, APIs and drivers. That flexibility helps organizations reduce the complexity, cost and time requirements to bring new services and solutions to market, streamlining adoption of new architectures and enabling engineers and programmers to innovate instead of maintaining code.

Ease of integration with existing implementations

With Intel, businesses can speed up time to deployment with the largest ecosystem of partners they know and use. Hardware and software vendors and solution integrators around the world build their products on Intel Xeon Scalable processors, offering maximum choice and interoperability with the reassurance of thousands of real-world implementations.

PERFORMANCE PROOFPOINT

HIGHER VASP PERFORMANCE[2,3]

UP TO **1.61X** GEOMEAN ON 4TH GEN INTEL XEON SCALABLE PROCESSORS VS PREVIOUS GEN

UP TO **2.01X** GEOMEAN ON INTEL XEON PROCESSOR MAX SERIES VS 2S 3RD GEN INTEL XEON SCALABLE PROCESSOR

Designed for the full spectrum of HPC use cases

4th Gen Intel Xeon processors accelerate a range of real-world use cases with high performance, improved memory bandwidth using DDR5 and advanced I/O using PCIe Gen 5 and CXL 1.1. Developers can build code faster with Intel's leading software libraries and compilers, optimized so that HPC applications perform better out of the box. Code and models can take advantage of powerful Intel AVX-512 technology with two FMA units per core for the most demanding computational workloads. With the Intel MPI Library, workloads can scale across multiple HPC clusters. Add Intel Optane™ Persistent Memory to support large calculations that benefit from a larger memory footprint.

Do more with Intel technology supporting HPC workloads

Maximize bandwidth. The new Intel® Xeon® Processor Max Series is designed to deliver up to a 4x improvement in performance over platforms with just DDR5 by unblocking the bottlenecks for memory-bound workloads such as modeling, AI, HPC and data analytics. This is the first x86 CPU to integrate high-bandwidth memory and accelerators onto the processor package, with up to 64 GB of HBM2e. It improves TCO with reduced DDR dependency, the latest software tools and excellent code reusability.

Maximize impact. The flagship Intel Data Center GPU Max Series uses Intel's most advanced IP and packaging technologies designed to accelerate AI, HPC and advanced analytics workloads for the Exascale era. Based on the Intel Xe HPC architecture, it provides the largest high-bandwidth cache in a GPU. Supported by the oneAPI open ecosystem with SIMT/SIMD flexibility, the GPU integrates multiple IP innovations in-package, including high-bandwidth memory.

Microsecond access to data. DAOS (Distributed Asynchronous Object Storage) is an open-source, software-defined, scale-out object store that cost-effectively provides high bandwidth, low latency, and high I/O operations per second (IOPS) storage containers to HPC and AI applications in a single storage tier. DAOS natively supports structured, semi-structured, and unstructured datasets while eliminating limitations of traditional distributed storage.

Learn More

www.intel.com/xeon/scalable
www.intel.com/hpc

[1] Intersect360 Research, May 20, 2022. "Total HPC Market Revenue Grew 5.2% to $41.0 Billion in 2021, Says Intersect360 Research." https://www.hpcwire.com/off-the-wire/total-hpc-market-revenue-grew-5-2-to-41-0-billion-in-2021-says-intersect360-research/.

[2] Compared to 2S 3rd Gen Intel Xeon Scalable processor.

[3] See intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Availability of accelerators varies depending on SKU. Visit the Intel Product Specifications page for additional product details.

Performance varies by use, configuration, and other factors. Learn more at https://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 configuration disclosure for configuration details. 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.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a nonexclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

1122/JAW/MESH/PDF 350500-003US