Product Brief

The Intel® Xeon® 6 Product Family

The Intel Xeon 6 processor family introduces a robust computing platform that excels at both performance and efficiency, which are crucial for meeting the evolving demands of modern data centers. From compute-intensive AI to scale-out microservices, the processor family provides versatility for diverse workload requirements. In the claims shown, comparing the Intel Xeon 6 processor to the 2nd Gen Intel Xeon processor is indicative of the gains possible with a 5-year hardware refresh. Gains are also possible compared to 5th Gen Intel Xeon Scalable processors.

Addressing the needs of today's data centers

Data center infrastructure is one of the most important investments an organization makes. IT leaders need to implement resources that foster business growth while balancing security, energy efficiency, manageability, and other factors affecting total cost of ownership (TCO). Perhaps most importantly, IT leaders must be cognizant of market factors driving rapidly evolving data center demands. For example, the **AI market size and opportunity are expected to grow by 4x in the next five years**, fueled by changes like workload automation and generative AI.[1] A growing cohort of enterprise applications is adding inferencing code, which requires processing large vectors of data with data-parallel computing requirements that favor performance per core. In that same time frame, **the cloud microservices market is expected to grow by 5x**, fueled by the redesign of monolithic applications with cloud-native principles.[2] These workloads are task-parallel. Thus, they benefit more from efficient, scalar processing than from added complex compute capabilities.
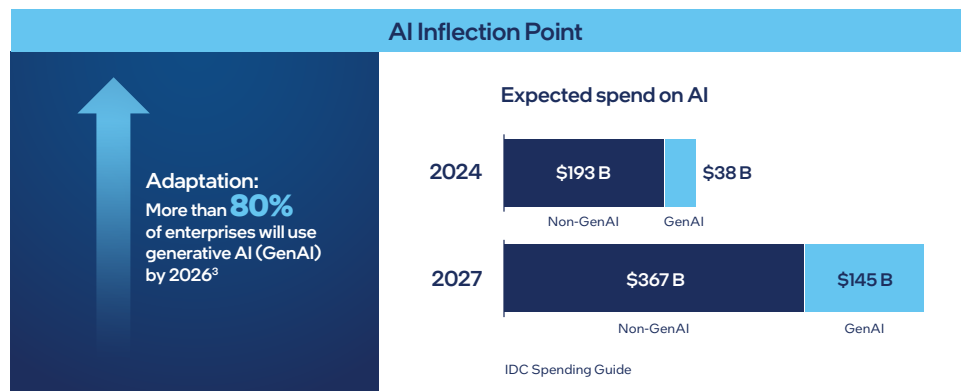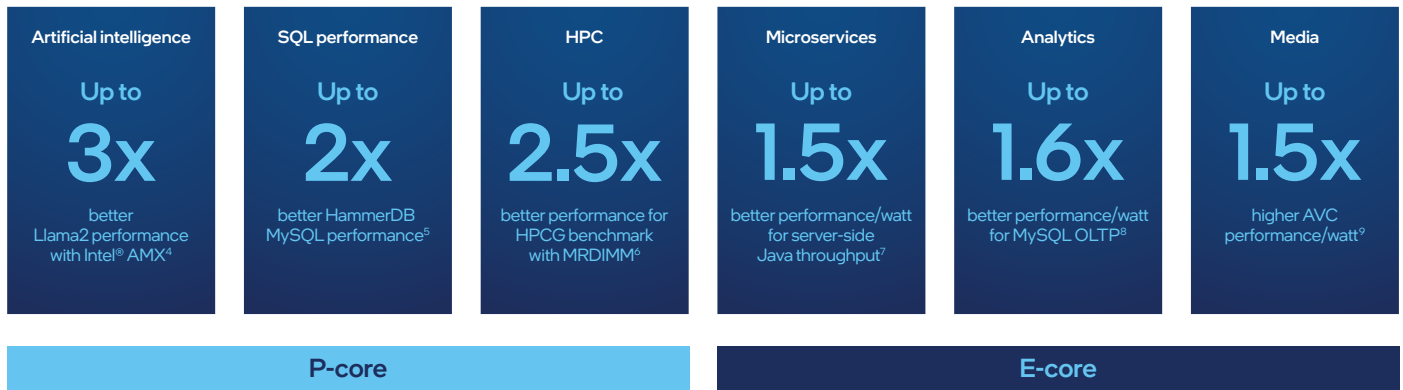

Figure 1. Breakdown of total addressable market (TAM) for non-generative AI (non-GenAI) and generative AI (GenAI)
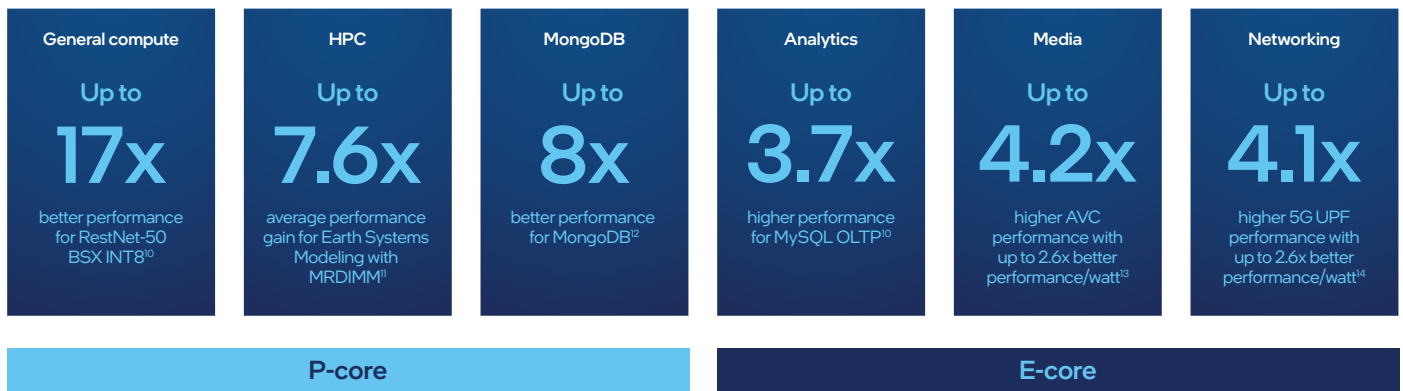
Faster business results across the spectrum of workloads

Intel Xeon 6 processors deliver new degrees of performance with more cores, a choice of microarchitecture, additional memory bandwidth, and exceptional input/output (I/O) across a range of workloads. New features like Multiplexed Rank DIMM (MRDIMM) support, Compute Express Link (CXL) enhancements and many others, in addition to integrated accelerators like Intel® QuickAssist Technology (Intel® QAT), give an additional boost to targeted workloads for even greater performance and efficiency.
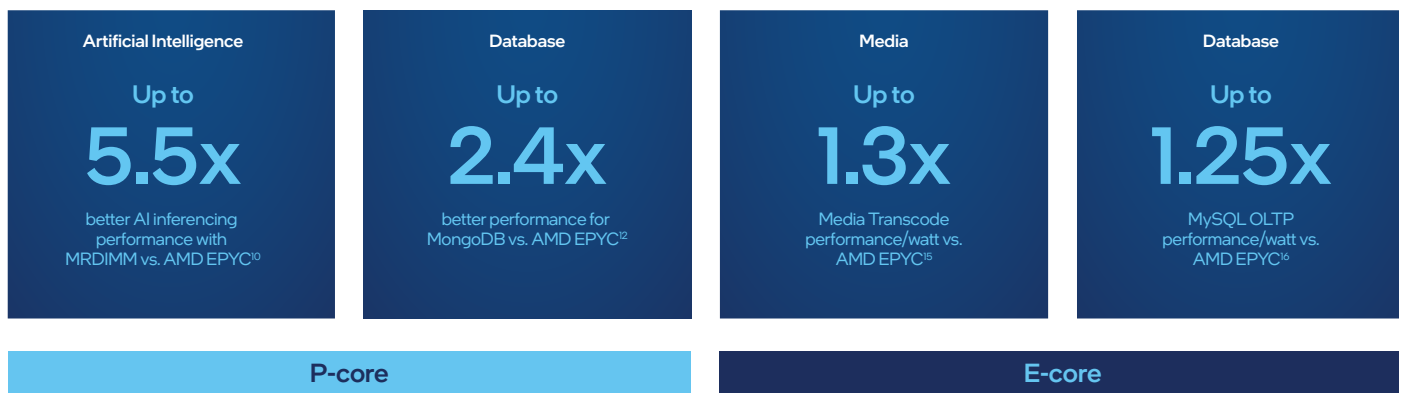
Intel® Xeon® 6 processor compared to 5th Gen Intel® Xeon® Scalable processor

Artificial intelligence

Up to

3x

better
Llama2 performance
with Intel® AMX[4]

SQL performance

Up to

2x

better HammerDB
MySQL performance[5]

HPC

Up to

2.5x

better performance for
HPCG benchmark
with MRDIMM[6]

Microservices

Up to

1.5x

better performance/watt
for server-side
Java throughput[7]

Analytics

Up to

1.6x

better performance/watt
for MySQL OLTP[8]

Media

Up to

1.5x

higher AVC
performance/watt[9]

P-core   **E-core**

Intel® Xeon® 6 processor compared to 2nd Gen Intel® Xeon® Scalable processor

General compute

Up to

17x

better performance
for RestNet-50
BSX INT8[10]

HPC

Up to

7.6x

average performance
gain for Earth Systems
Modeling with
MRDIMM[11]

MongoDB

Up to

8x

better performance
for MongoDB[12]

Analytics

Up to

3.7x

higher performance
for MySQL OLTP[10]

Media

Up to

4.2x

higher AVC
performance with
up to 2.6x better
performance/watt[13]

Networking

Up to

4.1x

higher 5G UPF
performance with
up to 2.6x better
performance/watt[14]

P-core   **E-core**

Intel® Xeon® 6 processor compared to AMD EPYC processor

Artificial Intelligence

Up to

5.5x

better AI inferencing
performance with
MRDIMM vs. AMD EPYC[10]

Database

Up to

2.4x

better performance for
MongoDB vs. AMD EPYC[12]

Media

Up to

1.3x

Media Transcode
performance/watt vs.
AMD EPYC[15]

Database

Up to

1.25x

MySQL OLTP
performance/watt vs.
AMD EPYC[16]

P-core   **E-core**

Performance and efficiency without compromise

The Intel Xeon 6 processor family introduces an innovative modular x86 architecture that allows data center architects to configure and deploy infrastructures that are purpose-built for your unique needs and workloads across private, public, and hybrid clouds. As shown in Table 1, Intel Xeon 6 processors are available in four different series, offering tiered capabilities from entry-level to demanding workloads through options for an increased number of cores, larger cache, faster and higher-capacity memory, and improved I/O over previous generations.

For the ultimate versatility, Intel Xeon 6 processors allow for the choice of two different CPU microarchitectures: Performance-cores (P-cores) and Efficient-cores (E-cores). Both core types use a compatible x86 instruction set architecture (ISA) and a common hardware platform, including CPU socket type. Furthermore, Intel has teamed with industry partners to help ensure seamless use of both core types with common operating systems, compilers, libraries, and frameworks. With this shared software stack and a global ecosystem of hardware and software vendors, solutions can be matched to every business need.

Table 1. The Intel Xeon 6 processor family encompasses four series of processors

Series	Designed for
Intel Xeon 6900-series processors	Maximum performance ideal for the most demanding cloud, AI, and HPC environments
Intel Xeon 6700-series processors	Enhanced performance ideal for a wide array of data center and telco environments
Intel Xeon 6500-series processors	Essential performance ideal for mainstream server and edge environments
Intel Xeon 6300-series processors	Entry-level performance ideal for small/medium business environments

Intel Xeon 6 processors with Performance-cores (P-cores)

Intel Xeon 6 processors with P-cores are optimized for high performance per core. With more cores, double the memory bandwidth, and AI acceleration in every core, Intel Xeon 6 processors provide twice the performance for the widest range of workloads, including AI and high-performance computing (HPC).[4,6] Intel Xeon 6 processors with P-cores excel at a wide range of workloads, delivering better performance than any other general-purpose CPU for compute-intensive workloads like AI inference and machine learning (ML). Intel Xeon 6 processors with P-cores are great for public cloud workloads with improved performance per vCPU for floating point operations, transactional databases and HPC workloads. Through their leadership in AI inferencing, Intel Xeon processors continue to be the host CPU of choice on the world's most powerful AI accelerator platforms for data preprocessing support.

- Enable AI everywhere with AI acceleration in every core. Intel® Advanced Matrix Extensions (Intel® AMX) speeds up inferencing for INT8 and BF16 and offers new support for FP16-trained models with up to 2,048 floating point operations per cycle per core for INT8 and 1,024 floating point operations per cycle per core for BF16/FP16.

- Improve memory throughput with the fastest DDR5 memory available, MRDIMM. These can deliver more than 37 percent greater memory bandwidth than RDIMMs,[17] with an expected data transfer rate of up to 8,800 megatransfers per second (MT/s). Intel Xeon 6 processors (both P-cores and E-cores) also support DDR5 6400 high-speed memory, providing memory bandwidth gains.[17]

- Take advantage of up to 128 cores per socket with up to 504 MB L3 cache and exceptionally low latency at large L3 access sizes. Intel® Advanced Vector Extensions 512 (Intel® AVX-512) is only supported on Xeon 6 processors with P-cores and can be used out of the box, boosting the speed of vector math common to HPC and classical AI workloads.

Up to

2x

better performance

(compared to 5th Gen Intel Xeon Scalable processors)[18]

MRDIMM delivers

37%

greater memory bandwidth

(compared to RDIMMs)[17]

3

Intel Xeon 6 processors with Efficient-cores (E-cores)

Intel Xeon processors with E-cores are optimized for high core density and exceptional performance per watt, delivering distinct advantages for cloud-scale workloads that demand high task-parallel throughput. In comparison to 2nd Gen Intel Xeon Scalable processors, which make up the majority of systems within today's data center footprint and are excellent candidates for performance per watt upgrades in competitive data centers, Intel Xeon 6 processors with E-cores can deliver more than 2.6x better results.[14] This efficient performance is also ideal where power, space, and cooling are limited. Intel Xeon 6 processors with E-cores can:

- Replace four 2nd Gen Intel Xeon Scalable processor–based servers with a single server at similar performance.[19]

- Consolidate three racks of 2nd Gen Intel Xeon Scalable processor–based systems into a single rack.[20]

- Accommodate AI inferencing and vector-oriented operations with Intel® Advanced Vector Extensions 2 (Intel® AVX2) and new enhancements such as Vector Neural Network Instructions (VNNI) and fast-convert for BF16 and FP16.

- Offer up to 288 cores per socket, with as much as 216 MB L3 cache, and with exceptionally low latency at large L3 access sizes.

Consolidate

3
racks

into a
single rack[20]

Up to

2.6x

better
results

(compared to
2nd Gen Intel Xeon
Scalable processors)[14]

The efficiency of Intel Xeon 6 processors with P-cores and E-cores is highlighted by their ability to provide scalable performance per watt as server utilization increases, delivering nearly linear power-performance consumption across the load line. For performance-demanding workloads, this means power is efficiently utilized at high loads to finish jobs faster. For a scalable implementation, common to cloud or shared computing environments, this level of efficiency means that servers are only consuming the power they need when under load, reducing costs when instances are not fully utilized.

The sustainability of these processors is further enhanced through system-wide power management and telemetry capabilities. These capabilities allow for increased performance per watt on a per-application basis to help with lowering overall energy consumption.

Versatility and complementary advantages of Intel Xeon 6 processors

At the extreme end of the workload spectrum, P-cores offer the best solution for compute-intensive, vector-based workloads such as AI. E-cores are best for task-parallel, scalar-based workloads such as microservices. Between these extremes, the two microarchitectures combine to allow for highly versatile and complementary solutions. For example, systems with Intel Xeon 6 processors with E-cores can be used to conserve power so that it is available for AI and scientific workloads running on nodes with Intel Xeon 6 processors with P-cores. Data centers designed with a mix of Intel Xeon 6 processors with P-cores and Intel Xeon 6 processors with E-cores can take advantage of their platform commonality to transition workloads from one core type to the other depending on performance and power needs. The wide mix of options makes it easy for the data center to scale as the business grows.

As another example of the complementary nature of Intel Xeon 6 processors, a data center can easily mix servers with Intel Xeon 6 processors with P-cores and Intel Xeon 6 processors with E-cores to support business needs that require databases of different structures. Relational databases, which are characterized by complex data relationships, complex queries, joins, and aggregations, can benefit from the parallel data processing capabilities of Intel Xeon 6 processors with P-cores. Non-relational databases with numerous small, independent requests for data retrieval, such as key-value stores, can benefit from the task-parallel design of Intel Xeon 6 processors with E-cores.

Highlight technologies

The innovative P-core and E-core microarchitectures of the Intel Xeon 6 processor family deliver the following advanced features and benefits:

- Up to 288 cores in a single socket for Intel Xeon 6 processors with E-cores—or up to 128 cores in a single socket for Intel Xeon 6 processors with P-cores— enabling ultra-high-density compute performance and scalability.

- Intel AMX provides up to 16x more multiply accumulate (MAC) operations than Intel AVX-512 for BF16- and FP16-based models to enhance AI performance (P-core-only feature).

- Intel AVX-512 encompasses unique instructions and two 512-bit fused-multiply add (FMA) units per core, boosting the speed of vector mathematics common to AI, HPC, and database workloads (P-core-only feature).

- Intel AVX2 with VNNI and fast up/down convert for BF16 and FP16 enables better AI compatibility for Intel Xeon 6 processors with E-cores.

- MRDIMMs are capable of providing more than 37 percent additional memory bandwidth compared to standard DDR5 DIMMs, supporting bandwidth-constrained use cases found in AI and HPC (P-core-only feature). Intel Xeon 6 processors (both P-cores and E-cores) also support DDR5 6400 high-speed memory, providing memory bandwidth gains.[17]

- Up to 12 memory channels, further supporting higher memory bandwidth.

- Intel® Ultra Path Interconnect (Intel® UPI) 2.0 provides up to 24 gigatransfers per second (GT/s) of inter-socket bandwidth—a 20 percent increase over the prior generation.

- Up to 192 lanes of PCIe Gen 5 for two-socket servers, with options of up to 136 lanes for one-socket server designs, to allow for significant I/O add-in components including accelerators, network adapters, storage controllers, and storage.

- Up to 64 lanes of CXL 2.0 with data transfer rates up to 32 GT/s per lane, supporting CXL capabilities including memory expansion and sharing, including Type 3 devices.

- Flat Memory Mode helps expand system memory and improve TCO when using lower-cost memory, such as DDR4 with CXL 2.0.

- Intel QAT allows offload of bulk cryptography and compression to accelerate networking and storage.

- Intel® Data Streaming Accelerator (Intel® DSA) 2.0 enables offload of data movement and transform operations such as move, fill, compare, cyclic redundancy checking (CRC), data integrity field (DIF), delta, and flush.

- Intel® In-Memory Analytics Accelerator (Intel® IAA) allows offload of memory compression and decompression, scan and filter functions, and CRC.

- Intel® Dynamic Load Balancer (Intel® DLB) enables the dynamic distribution of network packet processing and offload of reordering operations.

- Intel® Trust Domain Extensions (Intel® TDX) upgrades with AES-256 and 2,048 encryption keys enhance confidential computing for the protection of sensitive business data.

To learn more about Intel Xeon 6 processors, including the listed features above, visit intel.com/xeon.

Overview of the Intel Xeon 6 processor family

Intel Xeon 6900-series processors are delivered in a new class of Intel server platform design, offering customers maximum performance, the highest memory bandwidth, and maximum throughput ideal for cloud, HPC, and AI environments. These processors feature higher core counts, more memory channels, and I/O lanes with thermal design points that are higher than the other series.

Intel Xeon 6700-series and Intel Xeon 6500-series processors are delivered in an updated server platform design featuring high performance with cost- and power-efficient solutions ideal for the widest array of data center environments. These processors come in one-socket to eight-socket options with enhanced I/O and memory within established data center power and cooling footprints.

Addressing Unique Workload Requirements

P-core			Workloads	E-core		
	Modeling and simulation	CAE	HPC			
			Web & microservices	Cloud-native	Consumer digital services	Application DevOps
CRM, ERP	Big data	In-memory analytics	Database & analytics	Unstructured databases	Scale-out analytics	
Generative AI	Deep learning Machine learning	Inference	AI			
HCI	Virtualization	Storage	Infrastructure & storage	Storage		
	CDN	Media & gaming	Networking	Network microservices	Cloud-native CDN	5G core
	Video	Edge analytics	Edge	Virtual protection relay		

Exemplary user experience

Intel Xeon 6 processors provide a high level of quality and reliability that customers appreciate from Intel® products. Maintaining continuous operation and minimizing the time needed to service a system are fundamental to managing a data center's service-level agreements (SLAs) and overall TCO. Intel reliability, availability, and serviceability (RAS) features encompass a suite of capabilities that help increase system uptime, reduce unplanned interruptions, and maintain data integrity. Intel Xeon 6 processors are the first Intel Xeon processors without platform controller hub (PCH) technology, enabling the CPU to self-boot without an assist from PCH. By removing PCH technology, this update supports lifecycle optimization and sustainability and is beneficial for system designers.

Security is important for both the user experience and customer satisfaction. IT teams must protect against a growing number of security threats and remain compliant with privacy regulations, whether on-premises or in the cloud. To protect data in use, Intel Xeon processors allow you to pick the confidential computing technologies that best meet your business and regulatory requirements. Intel TDX offers isolation and confidentiality at the virtual machine (VM) level, while Intel® Software Guard Extensions (Intel® SGX) provides application-level isolation.

For the demands of today's data centers, and for tomorrow's, consider the Intel Xeon 6 processor to foster business growth and manage TCO while balancing security, energy efficiency, and manageability. The Intel Xeon 6 processor family delivers new degrees of performance and efficiency.

Disclaimer #1

Performance varies by use, configuration and other factors. Results may vary. Power used is TDP.

HammerDB MySQL:

6756E: 1-node, pre-production platform, 2x Intel Xeon 6756E processor, 128 cores, 225W, Hyper-threading n/a, Turbo on, total memory 1024 GB (16 x 64 GB DDR5 6,400 MT/s [6,400 MT/s]), microcode 0x13000131, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7 T SAMSUNG MZWLJ1T9HBJR-00007, 2x 1.7 T KIOXIA KCD8XPUG1T92, 2x 1.7 T KIOXIA KCD81PUG1T92, Ubuntu 22.04.4 LTS, 6.5.0-21-generic, test by Intel as of 05/10/24.

9534: 1-node, 2x AMD EPYC 9534 64-core processor, 64 cores, 280W, SMT on, Turbo on, total memory 1536 GB (24 x 64 GB DDR5 4,800 MT/s [4,800 MT/s]), microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 4x 1.7 T KIOXIA KCD8XPUG1T92, 1x 1.7 T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.1 LTS, 6.5.0-21-generic, test by Intel as of 05/09/24.

Socket power for Intel, TDP for AMD

Software: HammerDB TPROC-C, MySQL 8.0.33, multi-instance

intel.
XEON®

1   Fortune Business Insights. "Artificial Intelligence Market Size, Share & Industry Analysis …." April 2024.
    fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114.
2   Fortune Business Insights. "Cloud Microservices Market Size, Share & COVID-19 Impact Analysis …." April 2024. fortunebusinessinsights.com/cloud-microservices-market-107793.
3   Gartner. "Harness the Power of Democratized Generative AI to Transform Your Business." 2024.
    https://www.gartner.com/en/experts/top-tech-trends-unpacked-series/harness-democratized-generative-ai-transform-your-business.
4   See [9A2] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
5   See [9D1] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
6   See [9H10] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
7   See [7W5] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
8   See [7D2] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
9   See [7W2] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
10 See [9A3] at intel.com/processorclaims : Intel® Xeon® 6. Results may vary.
11 See [9H2] at intel.com/processorclaims : Intel® Xeon® 6. Results may vary.
12 See [9D2] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
13 See [7W1] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
14 See [7N1] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
15 See [7W210] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
16 See Disclaimer 1 above for full configuration for comparison.
17 In comparison to DDR5 6,400 RDIMMs.
18 Better performance versus 5th Gen Intel Xeon processors based on Intel architectural projections as of August 2023.
19 Based on MySQL OLTP and server-side Java throughput with SLA. See intel.com/processorclaims: Intel® Xeon® 6. Results may vary.
20 See [7T1] and [7T2] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.