

# Reduce Costs with Improved Performance per Dollar in the Cloud with Intel Xeon 6 Processors

**Intel Xeon 6 processors with Performance-cores (P-cores) deliver up to 2x higher performance compared to the latest AMD EPYC CPU, enabling greater performance in the cloud.<sup>1</sup>**



Up to

## 2x higher

INT8 inference performance with the Intel Xeon 6980P processor, compared to the AMD EPYC 9755 processor.<sup>1</sup>

In today's fast-paced digital landscape, enterprises across industries are navigating a complex array of challenges. They must balance capital and operating expenses (CapEx and OpEx) with substantial investments in AI, prioritizing cost-effectiveness and energy efficiency for sustainable operations, and upholding stringent data security standards in an evolving environment of risks and regulatory requirements.

Amidst these challenges, an increasing number of organizations are using cloud-based solutions for database and AI applications. Running these applications in the cloud provides unlimited resources and allows you to scale your AI models and infrastructure up or down based on demand, enabling faster innovation.

### Performance where it's needed in the cloud

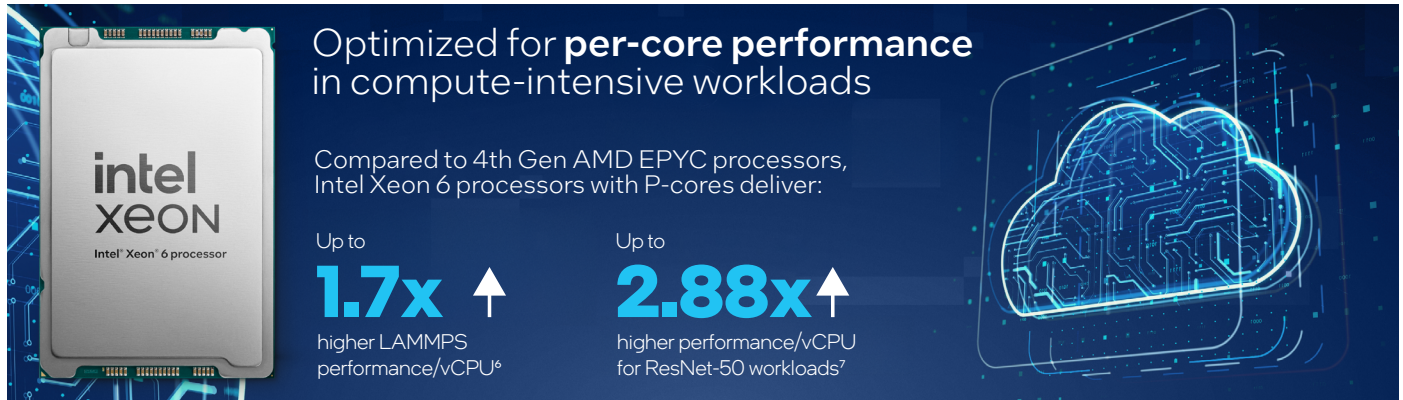
Intel Xeon 6 processors with Performance-cores (P-cores) are designed to address the multifaceted demands that organizations place upon their cloud-based solutions. With up to 128 cores, these processors are architected for compute-intensive tasks that benefit from multiple data elements being processed in parallel, and they deliver increases in socket performance across workloads. They are optimized for high performance per core, and they excel at a wide range of workloads, including providing better AI performance than any other general-purpose CPU.<sup>2</sup> In comparison to

5th Gen Intel Xeon processors, which are the CPUs commonly referenced for newer compute-intensive solutions, Intel Xeon 6 processors with P-cores can provide more than 2x better performance.<sup>3</sup> You can take advantage of this development to improve performance per dollar by running twice as many workloads on the same number of cores, or by finishing workloads twice as fast. In addition, an extensive ecosystem of ISV apps, operating system tools, libraries, and frameworks that support these processors means greater ease of use for developers.

The following features contribute to these performance gains:

- Intel® Advanced Vector Extensions 512 (Intel® AVX-512) speeds up vector processing-intensive workloads. This Intel® Accelerator Engine provides enhanced performance for demanding applications such as scientific simulations, computer-aided engineering (CAE), financial analytics, and 3D modeling. 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, high-performance computing (HPC), and database workloads.
- Intel® Advanced Matrix Extensions (Intel® AMX) improves the performance of deep learning (DL) training and inference, making it ideal for workloads like natural language processing (NLP), recommender systems, and image recognition. Intel Xeon 6 processors with Intel AMX deliver up to 2x better generative AI (GenAI) performance with BF16 data types compared to 5th Gen Intel Xeon processors,<sup>4</sup> and up to 2x better AI inferencing performance compared to the latest AMD EPYC processors.<sup>1</sup> In addition, Intel Xeon 6 processors with P-cores support the FP16 data type, which allows for more compact data storage. Due to their smaller size, FP16 calculations can be significantly faster than FP32 calculations, making FP16 ideal for image processing, AI training, and graphics rendering.
- Intel Xeon 6 processors with P-cores improve memory throughput with the fastest DDR5 memory available, Multiplexed Rank DIMMs (MRDIMMs). These processors support bandwidth-constrained and memory-bound workloads in AI and HPC, and they deliver more than 37 percent greater memory bandwidth than RDIMMs, with an expected data transfer rate of up to 8,800 megatransfers per second (MT/s).<sup>5</sup>
- These processors also feature up to 64 lanes of Compute Express Link (CXL) 2.0, with data transfer rates of up to 32 gigatransfers per second (GT/s) per lane, supporting CXL capabilities such as memory expansion and sharing, even for Type-3 devices.

Relational databases in the cloud can benefit from the parallel data processing capabilities of Intel Xeon 6 processors with P-cores. These workloads are characterized by complex data relationships, queries, joins, and aggregations. Advanced vector engines in Intel Xeon 6 processors with P-cores allow the single instruction, multiple data (SIMD)-biased workloads common to advanced database and analytics use cases to run effectively. In the cloud, compared to 4th Generation AMD EPYC processors, Intel Xeon 6 processors with P-cores deliver up to 1.7x higher performance per vCPU for Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) workloads,<sup>6</sup> and up to 2.88x higher performance per vCPU for ResNet-50 workloads.<sup>7</sup>



Optimized for **per-core performance** in compute-intensive workloads

Compared to 4th Gen AMD EPYC processors, Intel Xeon 6 processors with P-cores deliver:

Up to <b>1.7x</b> ↑ higher LAMMPS performance/vCPU <sup>6</sup>	Up to <b>2.88x</b> ↑ higher performance/vCPU for ResNet-50 workloads <sup>7</sup>
--	--

The graphic features an image of an Intel Xeon 6 processor on the left and a stylized cloud with circuitry on the right.

## Intel Xeon 6 processors deliver unparalleled performance for AI workloads in the cloud

With these advanced features and capabilities, Intel Xeon 6 processors are keeping pace with the increasing prevalence of AI workloads. AI-assisted workloads, such as data analysis and predictive modeling, are integral to making informed business decisions and driving innovation. The greater demand that these workloads place on processors for intensive calculations and throughput is more than a trend—it is a necessity, as these workloads play a pivotal role in an enterprise’s ability to stay competitive.

Intel Xeon 6 processors with P-cores deliver unparalleled performance for AI workloads in the cloud—such as data preparation and pre-processing, NLP, and GenAI—outperforming any other general-purpose CPU in the market.<sup>8</sup> In fact, Intel Xeon 6 processors with P-cores provide up to 3x higher AI inferencing performance compared to the prior generation.<sup>3</sup> This makes them an ideal choice for enterprises looking to take advantage of AI acceleration. They feature L3 cache as large as 504 MB and have exceptionally low latency at large L3 access sizes. The built-in Intel AMX accelerator 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. For help navigating the world of AI development in the cloud, see “[Field Guide to AI Developers in the Cloud](#).”

Up to **3x higher** AI inferencing performance with Intel Xeon 6 processors with P-cores, compared to the prior generation.<sup>3</sup>

## Experience the benefits of the cloud while enhancing your enterprise’s data security

Intel Xeon 6 processors include features that enhance security so you can deploy to the cloud with confidence. These features include:

- Intel® Trust Domain Extensions (Intel® TDX) helps narrow the attack surface and increase data and application protection and confidentiality in the cloud through hardware-level isolation within a virtual machine (VM). VM isolation with Intel TDX simplifies the porting and migration of existing applications to a confidential computing environment. In most cases, no application code changes are required to activate a trusted domain enabled by Intel TDX inside a VM. This accelerator includes Advanced Encryption Standard 256 (AES-256) and 2,048 encryption keys to enhance confidential computing for the protection of sensitive business data. New support for Intel TDX Connect enables encrypted communications with connected PCIe devices.
- Intel® Software Guard Extensions (Intel® SGX) helps businesses stay in control of their data while taking advantage of the cloud. Intel SGX helps protect data that is actively being used in the processor and memory by creating a trusted execution environment (TEE) called an enclave. Users can scale the amount of trusted code inside an enclave from an entire application with thousands of lines of code to a single function with just a few dozen, reducing the attack surface and restricting access to sensitive data.

### Enhance your security and privacy in the cloud

Intel TDX helps ensure data and applications are secure within a trusted domain, even if your underlying infrastructure is compromised. Intel SGX further strengthens security through security-enabled enclaves, where data can be processed in an encrypted format, helping protect sensitive information even from malicious software running on the same host.

## Benefits of cloud computing on Intel Xeon 6 processors with P-cores

Running cloud-based workloads on Intel Xeon 6 processors offers several key benefits that can directly influence business costs and outcomes:

- **Better efficiency for sustainability initiatives.** Intel Xeon 6 processors with P-cores deliver more than 2x better performance per watt compared to the prior generation.<sup>3</sup> This efficiency can help cloud service providers (CSPs) decrease their energy consumption and can help your organization meet its sustainability goals as a result.

More than **2x** better performance per watt<sup>3</sup>

- **Better performance per dollar.** Servers powered by Intel Xeon 6 processors with P-cores contain twice as many cores per socket and deliver up to 1.2x higher average performance per core compared to the previous generation, delivering better performance per dollar and improved output.<sup>9</sup>
- **Reduced latency for better customer experience.** With up to 128 cores per socket and an L3 cache as large as 504 MB, Intel Xeon 6 processors offer exceptionally low latency at large L3 access sizes. This translates into faster data processing and improved response times, enhancing customers' experiences and operational efficiency.

## Take advantage of performant and cost-effective instances

With their superior performance for data-intensive operations, advanced memory capabilities, and integrated security features, Intel Xeon 6 processors provide a robust platform for enterprises to navigate the complexities of the modern cloud environment. Their ability to excel in AI workloads and relational databases further underscores their value in today's data-driven world. By choosing Intel Xeon 6 processors, enterprises can confidently scale their operations in the cloud, harnessing the power of advanced technology to drive their success.

Visit [intel.com/xeon](https://intel.com/xeon) for more information about Intel Xeon 6 processors for cloud-based workloads.



<sup>1</sup> See [9A221] at [intel.com/processorclaims](https://intel.com/processorclaims): Intel Xeon 6. Results may vary.

<sup>2</sup> See [9A3] at [intel.com/processorclaims](https://intel.com/processorclaims): Intel Xeon 6. Results may vary.

<sup>3</sup> See [9A2] at [intel.com/processorclaims](https://intel.com/processorclaims): Intel Xeon 6. Results may vary.

<sup>4</sup> See [9A10] at [intel.com/processorclaims](https://intel.com/processorclaims): Intel Xeon 6. Results may vary.

<sup>5</sup> In comparison to DDR5 6,400 MT/s RDIMMs.

<sup>6</sup> See [9H5] at [intel.com/processorclaims](https://intel.com/processorclaims): Intel Xeon 6. Results may vary.

<sup>7</sup> See [intel.com/processorclaims](https://intel.com/processorclaims): Intel Xeon 6. Results may vary.

<sup>8</sup> See [9A3] and [9A7] at [intel.com/processorclaims](https://intel.com/processorclaims): Intel Xeon 6. Results may vary.

<sup>9</sup> See [9G4] at [intel.com/processorclaims](https://intel.com/processorclaims): Intel Xeon 6. Results may vary.

Performance varies by use, configuration and other factors. Learn more at [www.intel.com/PerformanceIndex](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 additional details.

No product or component can be absolutely secure.

Your costs and results may vary.

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

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.