Technical Research Study

Enhancing Data Center Performance and Efficiency: Intel® Xeon® 6 Processor Family Insights

Sponsored by Intel, this Prowess Consulting analysis shows that the Intel Xeon 6 processor family delivers flexible, scalable infrastructure solutions for a wide range of workload needs.

Executive Summary

To remain competitive in today's dynamic markets, businesses need data centers with agile, flexible, and scalable infrastructures that can support a range of workloads and applications with diverse requirements in terms of performance and energy efficiency. In this Intel-sponsored study, Prowess Consulting examines the Intel® Xeon® 6 processor family, a modular x86 architecture that allows IT staff to configure and deploy purpose-built infrastructures that meet specific business requirements around performance and efficiency. The study then shows how the Intel Xeon 6 processor family can be used to modernize data center infrastructures and resolve many challenges associated with today's computing environments.

The processor family supports two types of core microarchitectures: Intel Xeon 6 processors with Performance-cores (P-cores) and Intel Xeon 6 processors with Efficient-cores (E-cores). Both core types share a common software stack and high-speed input/output (I/O) and memory interfaces, enabling seamless integration and flexibility for a wide variety of use cases. Intel Xeon 6 processors with P-cores are optimized for high-performance deployments, while Intel Xeon 6 processors with E-cores are designed for high-efficiency and scale-out deployments. Demanding workloads such as AI, high-performance computing (HPC), relational databases, and analytics can benefit from the advanced matrix and compute engines and large cache provided with P-cores. Scale-out workloads such as microservices, application DevOps, and cloud-native apps can benefit from the energy efficiency and performance-per-watt capabilities of E-cores. This paper explores use cases for both types of processors to help organizations determine which solution might best suit their needs.

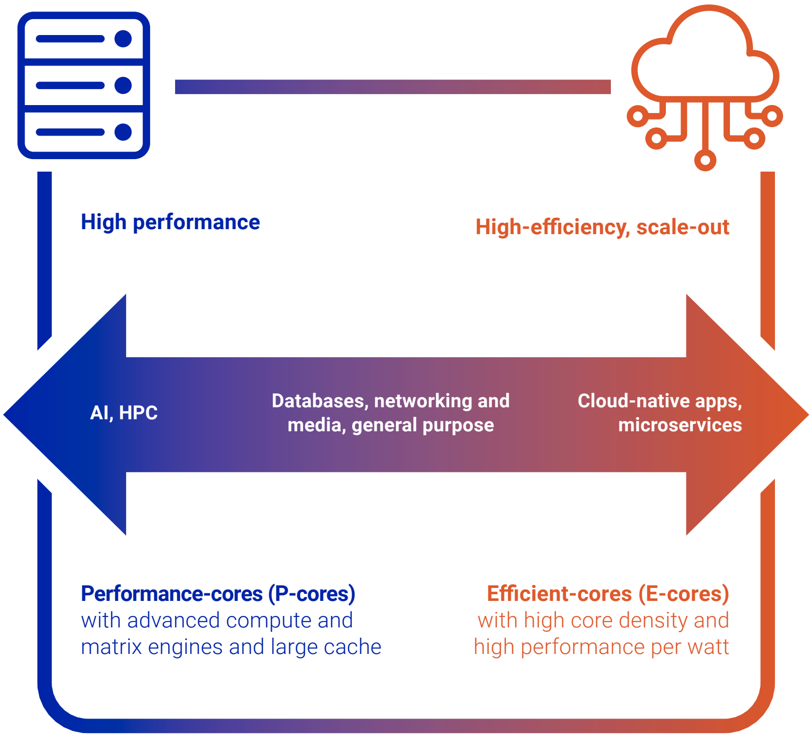
High performance
High-efficiency, scale-out
AI, HPC
Databases, networking and media, general purpose
Cloud-native apps, microservices
Performance-cores (P-cores) with advanced compute and matrix engines and large cache
Efficient-cores (E-cores) with high core density and high performance per watt

Figure 1 | The Intel® Xeon® 6 processor family features two microarchitectures to support a broad range of workload needs

Highlights

Intel® Xeon® 6 processors deliver higher efficiency and performance compared to previous-generation Intel Xeon Scalable processors:

Compared to 2nd Gen Intel Xeon Scalable processors, Intel Xeon 6 processors with Efficient-cores (E-cores) deliver

up to
3x
higher rack density[1]

Intel Xeon 6 processors with E-cores deliver

up to
2.7x
better performance per watt[1]

Compared to 5th Gen Intel Xeon Scalable Processors, Intel Xeon 6 processors with Performance-cores (P-cores) deliver

up to
2x
better performance for GenAI workloads[2]

Today's Market Trends

Data centers today must be scalable, flexible, and ready to evolve in the face of rapidly changing and familiar market imperatives, including the following needs:

- **Embrace AI.** As AI becomes increasingly important as a primary workload or as a new component of a traditional workload, IT staffs need solutions that can help them build AI-optimized infrastructure—with and without graphics processing units (GPUs).
- **Meet performance service-level agreements (SLAs) across a range of data center workloads.** In addition to handling emerging workloads like AI, data centers must continue to support traditional workloads efficiently. Web hosting, database management, virtualization and cloud infrastructure, and email services continue to be critical workloads. IT infrastructure needs to ensure that AI and traditional workhorses complement each other.
- **Reduce total cost of ownership (TCO) and drive sustainability.** The skyrocketing adoption of AI, especially the category running on foundational models and large language models (LLMs) with hundreds of billions of parameters, is sending power consumption in the data center sky-high. Organizations must find ways to increase their general-purpose compute density and lower their power consumption while getting the performance they need from their AI implementations by consolidating older server infrastructure to free up space and power budgets.
- **Protect valuable data.** Cybersecurity cannot be left as an afterthought; it must be a strategic and operational priority. A proactive approach to security includes the purchase of infrastructure solutions that come with advanced security capabilities.

Ensure Data Center Flexibility While Consolidating Infrastructure

The Intel Xeon 6 processor family introduces an innovative modular x86 architecture so that organizations can configure and deploy infrastructures that are purpose-built for their unique needs and workloads. It supports a common software stack that lets engineers develop on an open ecosystem. The solution's modularity offers flexible options for performance, efficiency, sustainability, and scalability.

The processor family's two types of microarchitectures can help tune and optimize investments for performance, power efficiency, and cost without sacrificing security or software interoperability. Intel Xeon 6 processors with P-cores are optimized for compute-intensive deployments, making them suitable for a wide variety of demanding use cases such as AI, HPC, and relational databases. Intel Xeon 6 processors with E-cores are designed for high-efficiency and scale-out deployments, making them ideal for high-traffic, low-intensity tasks such as cloud-native microservices, scale-out analytics, and networking workloads. All Intel Xeon 6 processors have the same built-in I/O accelerators that offload common network and security functions supported in previous generations to help improve overall compute efficiency for targeted workloads. This helps boost performance and performance per watt, reducing overall costs.

Because P-cores and E-cores share the same underlying hardware platform, IT organizations can mix systems with each type to achieve new levels of hardware optimization. The modular architecture can deliver significant cost savings for data centers that need to manage a diverse range of workloads, balance performance and efficiency, stay within a maximum power envelope, and comply with data privacy and sovereignty regulations. The platform commonality and x86 ISA shared by P-core and E-core architectures support the same software and hardware on either processor without having to rewrite code. The underlying platform provides the same high-performance I/O and memory slots, which means data centers can use the same networking, memory, and storage devices for either core type without updating any drivers. This flexibility helps simplify and streamline the DevOps cycle.

The Intel Xeon 6 processor family enables significant infrastructure consolidation by allowing data centers to replace multiple older servers with fewer more powerful and energy-efficient ones, reducing both physical footprint and operational costs. This consolidation is beneficial to all data centers, particularly for infrastructure-as-a-service (IaaS) providers, who can offer enhanced virtualized services with improved resource allocation and system interface commonality, leading to better customer satisfaction. Moreover, the ability to support diverse workloads on a unified architecture simplifies workload transition and scaling as client demands evolve.

Portfolio Ease-of-Use Benefits

A significant benefit of the portfolio's x86 architecture is that it supports some of the largest selections of frameworks, open-source apps, libraries, and container technologies, in addition to legacy cloud and enterprise-class applications. Intel collaborates with open-source developers and independent software vendors (ISVs) to help ensure that Intel's hardware performs as optimally as possible with their software. The Intel Xeon platform is the most prevalent architecture in both cloud and enterprise infrastructures today, showing the widespread acceptance and trust that IT professionals place in its performance and compatibility.

Intel Xeon 6 processors with P-cores and E-cores deliver the reliability, availability, and serviceability (RAS) capabilities that data center professionals rely on. The Intel Xeon 6 CPU advanced RAS features include live diagnostic capabilities and automated policy-based responses to preemptively manage system health and prevent unscheduled downtime.

In addition, the Intel Xeon 6 CPU family lets IT organizations continue to use the software optimizations they have spent decades implementing. Ultimately, IT staff can rely on Intel Xeon 6 processors for performance and ease of use, and they can place their attention on non-CPU challenges. For example, IT staff can adopt new and rapidly evolving workloads such as AI to achieve better and faster business results.

How to Build a Better Data Center

In this section, a purpose-built server infrastructure is configured using the Intel Xeon 6 processor family. The first two scenario subsections address the ends of the workload spectrums for which P-cores and E-cores are uniquely optimized. This is followed by two subsections highlighting scenarios that benefit from the flexibility and choice that P-cores and E-cores provide customers depending on their workload goals. The final two scenarios encompass IaaS and confidential computing. This section shows how the Intel Xeon 6 processor family's flexibility with a common software stack meets diverse operational requirements and optimizes technology investments.

Meeting the High Compute Demands of AI and HPC

Many enterprise organizations use AI-driven features in their business applications. Gartner predicts that by 2026, 80% of enterprises will have deployed generative AI (GenAI)-enabled applications in production environments.[3] GenAI holds the potential to improve efficiency in internal business processes, shorten time to market, and create better user experiences in customer-facing applications.

AI workloads such as GenAI in a retrieval-augmented generation (RAG) implementation require both high-performance compute and high memory bandwidth. Intel Xeon 6 processors address both of these attributes to deliver 2x better performance for GenAI workloads in comparison to the 5th Gen Intel Xeon Scalable processors that currently drive many AI solutions.[2] Intel Xeon 6 processors with P-cores use powerful computational engines and built-in Intel® Advanced Matrix Extensions (Intel® AMX) for the math-intense prefill stage of GenAI. In addition, Intel Xeon 6 processors with P-cores are also capable of using Multiplexer Combined Ranks DIMMs (MCR DIMMs), which can provide as much as 37% higher memory bandwidth than traditional DDR to help alleviate memory access constraints associated with the decode stage of AI. The combination of powerful compute engines with MCR DIMMs offers plenty of horsepower for small to medium-sized GenAI models with RAG.

Imagine a consumer electronics company that would like to use GenAI and RAG to help develop a new line of smart home devices. The team is considering an approximately 7 billion-parameter GenAI model trained about the features, specifications, and target markets for smart home devices. This will let them generate product descriptions and marketing content that resonates well with their expected customers. By adding RAG infrastructure, the team can tie in its internal knowledge base to help improve the AI-generated responses, ensuring that technical details are correct and brand guidelines are met. Intel Xeon 6 processors are ideal for this solution, given that they are capable of meeting response time SLAs for a 7 billion-parameter GenAI model and can also provide sufficient performance for a RAG vector database.

Another workload requiring massive computational power is HPC. Intel Xeon 6 processors with P-cores use another built-in capability, Intel® Advanced Vector Extensions 512 (Intel® AVX-512), to accelerate performance for use cases with vector-based mathematics, such as scientific simulations. As an example of a scenario where Intel AVX-512 would be used, consider a biotechnology company focused on the development of innovative therapeutic treatments. One of the key tools in its arsenal is GROMACS, a highly efficient open-source software package for performing molecular dynamics simulations. Using a cluster of systems with Intel Xeon 6 processors with P-cores to unleash the power of GROMACS, the team can gain valuable insights into structural dynamics, binding affinities, and drug interactions of proposed compounds. Similar to the expectations for AI, Intel Xeon 6 processors with P-cores are capable of boosting HPC performance by more than 2x when compared to 5th Gen Intel Xeon Scalable processors.[4]

Improving Performance while Lowering Infrastructure Costs for Cloud-Native Microservices

Microservices have emerged as a transformative approach to software development, characterized by breaking down complex applications into smaller, independently deployable services. This architectural style has witnessed a remarkable surge in popularity, largely due to its ability to enhance agility, scalability, and maintainability in modern software systems. With a high growth rate fueled by the demands of cloud computing, DevOps practices, and the need for faster innovation cycles, microservices have become the cornerstone of many organizations' digital strategies.

For a scenario such as streaming media services, microservices can be used to independently handle tasks like content delivery, user authentication, and recommendation engines. Intel Xeon 6 processors with E-cores are ideally suited for these microservice solutions. They excel in processing a multitude of service requests through a design that is optimized for single-thread performance, avoiding the intricate management and resource allocation challenges associated with hyperthreading. This streamlined design minimizes overhead and boosts efficiency.

Moreover, the processors' abundance of cores is adept at handling numerous thread-synchronized tasks and can swiftly adapt to changes in service demand.

Complementary Options for General-Purpose Compute

The Intel Xeon 6 processor family provides multiple options to drive general-purpose workloads. Consider a scenario such as Internet of Things (IoT) data processing in a small data center at an edge location where environmental conditions make high performance per watt crucial. This data center's footprint currently consists of 2nd Gen Intel Xeon processor–based servers like many of today's data centers. Upgrading will allow a configuration of server infrastructure that supports greater density and fits a smaller megawatt envelope. Intel Xeon 6 processors with E-cores provide enough compute power to achieve up to 3x higher rack density than 2nd Gen Intel Xeon Scalable processors.[1] At the same time, these processors are optimized to run on less energy, which helps reduce power and cooling costs, delivering up to 2.7x better performance per watt.[1]

For a general-purpose scenario that combines applications such as collaborative software with compute-intense business analysis software such as SAS® solutions, performance per core will be important. For this, Intel Xeon 6 processors with P-cores are the right fit. They excel at the complete spectrum of workloads, with a mainstream series that features core counts ranging from 8 to 86, 176 PCIe® Gen 5 lanes for add-in cards for networking and storage in dual CPU-based systems, and a single-socket product with a remarkable 136 PCIe lanes for single CPU-based systems.

Alternatively, data center administrators could configure the Intel Xeon 6 processor family for a scenario that requires a balance of performance and efficiency. Run demanding inferencing and analytics workloads on Intel Xeon 6 processors with P-cores. Using the same chipset, run low-intensity, high-idle-time workloads—such as system backups, software updates, or non-relational databases—on Intel Xeon 6 processors with E-cores. Workloads such as databases and enterprise apps can be supported by either Intel Xeon 6 processors with P-cores or Intel Xeon 6 processors with E-cores, depending on what is best suited to the specific workload needs in each case. Single instruction multiple data (SIMD) biased workloads align best with Intel Xeon 6 processors with P-cores, given their advanced vector engines and high number of pipeline stages. On the other hand, single instruction single data (SISD) biased workloads align better with Intel Xeon 6 processors with E-cores, given their scalar, throughput-optimized design. The shared Intel Xeon 6 processor family enables IT staff to develop and deploy apps on a common software stack and add network, storage, other I/O peripherals including Compute Express Link® (CXL®) 2.0 capability for Type 3 CXL memory devices, which can take advantage of new features such as memory interleaving and Flat Memory Mode.

Boosting Merchants' Online Sales and Streamlining Operations

Online merchants are increasingly dependent on sophisticated recommender systems to accurately predict and cater to customer preferences, thereby providing a more personalized and engaging shopping experience. These systems are crucial for consumer-facing organizations that need to ensure scalability and effectively manage the ebb and flow of customer demand. In addition to these external pressures, internal operational challenges persist. These include the need to streamline DevOps processes, minimize power and cooling expenditures, and dynamically scale systems to align with the changing demands of an organization's customer base.

Addressing both the external and internal challenges requires a robust, adaptable, and scalable infrastructure that is compatible with a diverse array of software and that supports high-speed I/O connectivity. An optimal solution for online merchants would involve a combination of Intel Xeon 6 processors with P-cores designed for compute-intensive tasks such as machine learning (ML) inference, which is at the heart of recommender systems, and Intel Xeon 6 processors with E-cores, which could be utilized for cloud-native web services that handle customer interactions and require scale-out capabilities. Cutting-edge technologies like Intel AMX and MCR DIMMs are instrumental in delivering high performance for these recommender systems. To address sales scalability, the high-density core counts of Intel Xeon 6 processors with E-cores can facilitate a swift response to spikes in activity. These processors are also adept at managing high-performance non-relational databases, and they are efficient for the less compute-intensive back-end operations that are part of a typical merchant's set of workloads. For all systems, superb I/O response times are achievable through the use of PCIe 5.0 and CXL 2.0 buses, further enhanced by Intel Xeon 6 processors' I/O accelerators such as Intel® QuickAssist Technology (Intel® QAT) and Intel® Data Streaming Accelerator (Intel® DSA).

The modular design of this infrastructure, along with a common software stack, can significantly streamline an IT staff's workload. For instance, repetitive DevOps tasks, such as coding for new deployments, can be greatly reduced. This efficiency is maintained regardless of the hardware configuration, whether it employs Intel Xeon 6 processors with P-cores or E-cores. This approach not only simplifies operations but also helps ensure resources are adaptable to future market demands and technological advancements.

Scaling Resources and Consolidating Hardware for IaaS

IaaS is an important component for both cloud service providers (CSPs) that offer this service and companies that use IaaS to achieve the flexibility to scale computing resources on demand while minimizing the need for physical hardware investment. The Intel Xeon 6 processor family is well suited to power IaaS platforms, providing the performance and efficiency required for a wide range of virtualized environments.

With IaaS, organizations can rapidly deploy virtual machines (VMs), storage, and networking services, paying only for the resources they consume. This model aligns perfectly with the Intel Xeon 6 processor family's capabilities, as it offers a scalable solution that can efficiently handle varying workloads with its mix of P-cores and E-cores.

To illustrate the practical application of the Intel Xeon 6 processor family within an IaaS context, consider a scenario where an IaaS provider caters to a diverse clientele with varying computational demands. For clients running compute-intensive AI models or engaging in real-time data processing, the provider could allocate VMs to servers powered by P-cores, ensuring high throughput and rapid response times. Conversely, for clients with workloads that have lower computational intensity, such as static web hosting or lightweight database services, the provider could utilize servers with E-cores, which offer an optimized balance of performance and power consumption. This strategic deployment not only enhances the provider's ability to deliver customized service levels but also drives down operational costs, contributing to a reduced TCO and more competitive service offerings. The performance and efficiency gains of the Intel Xeon 6 processor family, accompanied by its modular architecture and common software stack, allow for consolidation of hardware and simplified management experience with a built-in path to easily scale resources to meet evolving demands.

Protecting Data Privacy with GenAI

Circling back to AI in this final scenario, let's consider a consulting firm wanting to use GenAI to deliver highly accurate search results to employees at blazing-fast speeds. Its AI foundational models would need to be trained using massive datasets that contain highly confidential information. Thus, the challenge for the IT staff is to use AI to give employees better search results while still maintaining data privacy.

An excellent solution for this challenge would be to use Intel Xeon 6 processors with or without a GPU-based accelerator. Even as a host CPU, plenty of system-level compute power and memory bandwidth are essential. This is a good fit for Intel Xeon 6 processors with Intel AMX, Intel AVX-512, and MCR DIMMs as performance-enhancing features. Furthermore, Intel Xeon 6 processors allow organizations to build a confidential computing environment that safeguards information through either Intel® Software Guard Extensions (Intel® SGX) for application-level isolation or Intel® Trust Domain Extensions (Intel® TDX) for confidentiality at the VM level. These built-in Intel® Security Engines help isolate workloads inside trusted execution environments (TEEs), which can enable enterprises to engage in multi-party, shared analysis without losing privacy.

Learn About Intel Xeon 6 Processors and Consider How Their Flexibility Can Complement Workloads

Our analysis underscores the transformative potential of the Intel Xeon 6 processor family in modernizing data center infrastructures. With its modular x86 architecture, the processor family offers unparalleled flexibility, letting IT staff tailor solutions that precisely meet diverse performance and efficiency requirements. The dual microarchitecture approach, encompassing both P-cores and E-cores, can support a broad spectrum of workloads, from AI and HPC to cloud-native applications, with optimal efficiency. As enterprises navigate the complexities of today's demands on their IT infrastructure, the Intel Xeon 6 processor family could prove to be a pivotal solution. It can help organizations embrace AI, meet stringent performance SLAs, and work toward sustainability goals.

Learn More

Discover how Intel® AI Engines help enhance AI inference and training without adding a separate accelerator.

Discover how Intel® HPC Engines can deliver the performance and speed organizations need without investing in additional hardware beyond the CPU.

Discover how Intel® Security Engines allow organizations to get more value from their data without compromising security.

Discover how Intel® Storage Engines help reduce power consumption and deliver enhanced capabilities, such as on-the-fly compression and encryption, fast data movement, and integrated NVM Express® (NVMe®) device control, at scale.

[1] See [7T1] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

[2] See [9A10] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

[3] Gartner. "Gartner Says More Than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026." October 2023.

[4] See [9H10] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.

PROWESS

The analysis in this document was done by Prowess Consulting and commissioned by Intel.
Prowess Consulting and the Prowess logo are trademarks of Prowess Consulting, LLC.