Marketing Brief

Transform Your Business and Save Money Using CPUs with Built-In Accelerators

Prowess Consulting concluded that solutions powered by processors with built-in accelerators can deliver better performance, per-core efficiency, and cost savings for real-world workloads, compared to solutions that focus on core numbers and synthetic benchmarks.

Executive Summary

Today's data centers face the challenge of extracting as much value from their data as they can in the least amount of time possible. At the same time, fast-rising energy costs have everyone scrambling to lower the cost of powering and cooling their hardware infrastructures.

In a study sponsored by Intel, Prowess Consulting looked for solutions that could help address both of these challenges. What we found is that a solution powered by CPUs with built-in accelerators can offer the potential to deliver high performance and cost savings. We evaluated how workloads running on CPUs with built-in accelerators performed for the following popular use cases: artificial intelligence (AI)/machine learning (ML), high-performance computing (HPC), secure web services, big data analytics, databases, and storage.

Our analysis revealed that 4th Gen Intel® Xeon® Scalable processors take advantage of their built-in Intel® Accelerator Engines to deliver higher overall performance, performance per watt, and performance per core than previous-generation Intel® Xeon® Scalable processors and AMD EPYC™ processors.

How Successful Businesses Stay on Top

In today's dynamically changing business world, organizations require fast and accurate insights from their data to stay competitive. Business leaders use data-driven insights to understand their customers, respond to changing markets, differentiate themselves from competitors, and recognize new business opportunities.

To get these insights, you need to apply the right workload to your use case, whether it is AI/ML, HPC, secure web services, big data analytics, databases, or storage. For the best performance, efficiency, and total cost of ownership (TCO), you need to power those workloads with the right hardware.

With these considerations in mind, we recommend selecting solutions powered by processors with built-in accelerators that target the workloads you want to run. In general, a built-in accelerator is a workload-acceleration feature built directly onto the CPU that delivers workload performance and other benefits—such as higher performance per watt, space-saving consolidation, and improved TCO—without the need for additional discrete hardware that works outside of the CPU. Today's workloads are demanding in terms of compute, power, memory, and other resources. They are also complex, which means the hardware and software that runs those workloads should be engineered to work well together. The right solution—one that delivers high performance and low TCO—should not only perform optimally now, but it should also support expansion, scale-outs, and upgrades in the future.

To help you receive these benefits, Prowess Consulting examined solutions powered by CPUs with built-in accelerators. We chose not to include GPU-dedicated solutions in our research because TCO is one of our key selection criteria, and GPU hardware can be very expensive. While GPUs are a popular option for powering ML training servers, these ML-specific processors demand astronomical prices. In April 2023, the price for a single NVIDIA® GPU soared up to as much as $45,000.[1] A reliable supply chain is another issue that stands in the way of successful deployment; GPUs have become incredibly difficult to obtain.

Many of the latest-generation processors are capable of powering medium-scale ML training and inference models, without requiring additional hardware. In other words, a CPU with built-in acceleration might be all the hardware you need to meet service-level agreements (SLAs). If you are running large ML models, there are dedicated hardware alternatives you can use, such as the Intel® Gaudi processor, which are less pricey than GPUs.[2]

CPU Hardware and Your Blueprint for Success

The beating heart of your server is the CPU—you should choose one that can deliver the compute cycles you need to run your demanding workloads.

Core counts and frequencies are two of the most common metrics used to measure processor performance. However, relying on these metrics alone can result in tunnel vision that might lead you astray. We suggest that the type of workload you are running should be factored into what CPU you use to power your server. If you upgrade your server simply based on higher core counts or frequencies, you could end up paying for hardware that does not perform as optimally or efficiently as you need.

To get the best value, you should choose CPUs that address your workload needs, help reduce TCO, and can deliver high per-core efficiency. Some compute-intensive workloads might run best with higher CPU core counts, which can naturally drive up power usage. Depending on the workload, however, it might be possible to offload certain tasks from the computing cores to built-in accelerators, as shown in Figure 1. Offloading tasks this way can deliver higher performance because the accelerator is purpose-built to process those tasks. The offloaded tasks can use less power to process because accelerators are often more power-efficient than CPU cores. Offloading also frees the CPU for other tasks, which can help improve overall server efficiency.

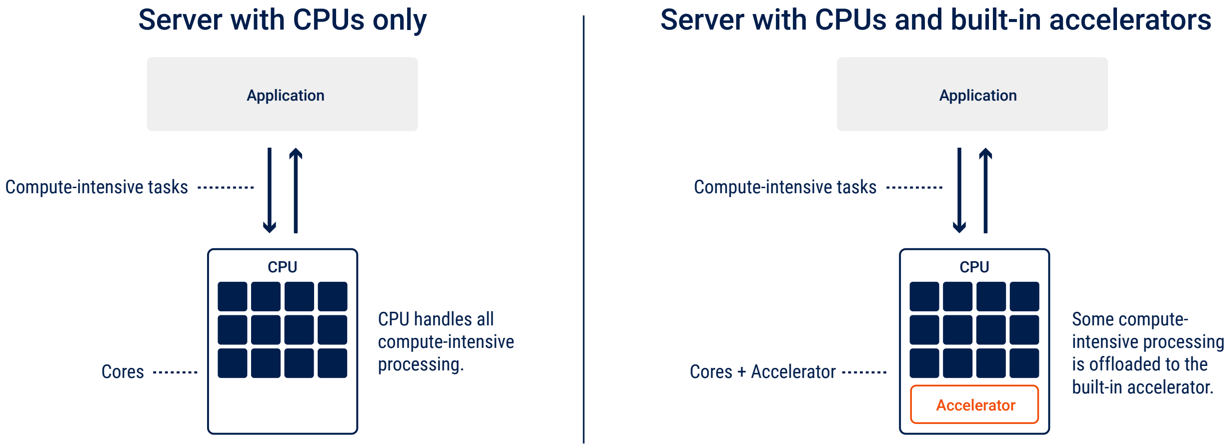
Server with CPUs only
Application
Compute-intensive tasks
CPU
Cores
CPU handles all compute-intensive processing.
Server with CPUs and built-in accelerators
Application
Compute-intensive tasks
CPU
Cores + Accelerator
Accelerator
Some compute-intensive processing is offloaded to the built-in accelerator.

Figure 1 | A built-in accelerator can offload compute-intensive tasks, which can enhance server performance and efficiency

In other words, a server powered by processors with built-in accelerators can help you get the best value from your hardware purchase. Such a right-sized solution can deliver a variety of benefits, compared to one powered by CPU cores alone, including:

- Higher overall performance
- Higher performance per watt
- Lower TCO

A server powered by 4th Gen Intel® Xeon® Scalable processors can achieve up to **3.9x higher performance per watt** with built-in accelerators enabled, versus without accelerators enabled.[3]

With these performance, efficiency, and TCO benefits in mind, we evaluated some examples of solutions powered by processors with built-in accelerators for several popular use cases:

- AI/ML
- HPC
- Secure web services
- Big data analytics/databases
- Storage

For our baseline of comparison, we chose AMD EPYC processors and older-generation Intel Xeon Scalable processors. We used 4th Gen Intel Xeon Scalable processors with built-in accelerators for our latest-generation CPUs. According to Intel, these x86 processors have the most built-in accelerators of any CPU on the market and are purpose-built to improve performance in AI, data analytics, networking, storage, and HPC workloads.[4] Each use case includes benefits and performance claims that you can use to help guide your hardware purchasing decisions.

Accelerate AI/ML Inference with Intel® AMX

Today's AI capabilities give organizations powerful tools to propel them ahead of the competition. Businesses use AI to better understand customers, accurately forecast models, and proactively reduce risk. Smart technologies, such as generative AI, ML, deep learning (DL), natural language processing (NLP), image classification, and recommender systems are changing how businesses relate to their customers and partners. And the rapid adoption of these technologies across all industries challenges data centers to stay current and stay competitive.

Combining CPU power and built-in accelerators can enhance processing for complex and demanding ML/DL workloads. While a CPU-based server might not be as all-out powerful for ML/DL as a GPU-based server, it can deliver the performance needed for SLA requirements without imposing a colossal price tag.

The Intel® Advanced Matrix Extensions (Intel® AMX) accelerator expands ML/DL functionality on the CPU to include training and fine tuning for small- to medium-scale models. An AI/ML inference server powered by 4th Gen Intel Xeon Scalable processors with Intel AMX can outperform servers powered by AMD EPYC processors while using considerably less power. According to Intel, inference workloads powered by the 4th Gen Intel Xeon Platinum 8462Y processor with the Intel AMX accelerator can outperform those running on the AMD EPYC 9354 processor by up to 5.6x for NLP,[5] 2.9x for recommender system,[5] and 5.9x for image-classification.[6]

The 4th Gen Intel® Xeon® Platinum 8462Y processor with the Intel® AMX accelerator engine can deliver up to **2.9x–5.9x higher ML inference performance** than the 4th Gen AMD EPYC™ 9354 processor using the same number of cores.[5,6]

The Intel solution powered by CPUs with built-in accelerators is also more power-efficient than the AMD CPU-based solution, delivering up to 4.7x higher performance per watt for NLP workloads, 2.4x higher performance per watt for recommender system workloads, and 2.4x better performance per watt for image-classification workloads.[5]

E-commerce giant Alibaba offers an example of how Intel AMX works to accelerate NLP in the real world. The company recognizes that faster and more accurate package delivery improves customer satisfaction. It upgraded its back-end NLP servers that were not optimized for AI workloads by installing 4th Gen Intel Xeon Scalable processors with Intel AMX, and it saw ML inference performance improve by up to 2.48x.[7]

For more information on how Intel AMX works, see the tuning guide.

Accelerate HPC with Intel® AVX-512

Stock trading firms rely on stock trackers and transaction handlers to monitor and respond to rapidly changing financial markets. Latency is the enemy on the trading floor, so traders require timely insights that facilitate the swift execution of profitable transactions.

The complex algorithms powering financial-services simulations, such as Monte Carlo and Black-Scholes, can benefit immensely from the speed and performance of an HPC solution powered by CPUs with built-in accelerators. This is another example of getting high performance from a server that targets certain workloads.

The Intel® Advanced Vector Extensions 512 (Intel® AVX-512) instruction set has long been used to accelerate HPC workloads running on Intel Xeon CPUs. The most recent version of Intel AVX-512 built into the 4th Gen Intel Xeon Platinum 8480+ processor has been further optimized so that it delivers up to 1.3x higher Monte Carlo simulation performance, compared to the 3rd Gen Intel Xeon Platinum 8380 processor with an older version of Intel AVX-512.[8]

The 4th Gen Intel® Xeon® Scalable processor with Intel® AVX-512 delivers up to **1.3x higher Monte Carlo performance** than the 3rd Gen Intel Xeon Scalable processor.[8]

Intel AVX-512 is used to accelerate complex workloads for extremely large HPC platforms, also known as supercomputers. For example, the faculty at Kyoto University chose 4th Gen Intel Xeon Platinum 8480+ processors with Intel AVX-512 to refresh their five-year-old HPC supercomputers.[9]

For more information on Intel AVX-512, see the technical brief.

Accelerate Secure Web Services with Intel® QAT

According to the technology usage tracking company W3Techs, 34% of all web servers run on the NGINX® open-source platform in 2023.[10] These servers support websites for organizations such as WordPress, Adobe, Mozilla, DigiCert, the European Union, Tumblr, Tencent QQ, and the National Institutes of Health (NIH).[10] The software developer services company Kinsta reports that NGINX is the most popular web server software for high-traffic websites, in large part because its performance does not degrade when scaled.[11]

Naturally, the more secure connections a web server can establish and maintain, the higher throughput it can deliver. Intel suggests that a combination of CPUs with built-in networking accelerators can deliver better performance than general-purpose CPUs.[5,12]

For example, Intel compared the 4th Gen Intel Xeon Platinum 8462Y processor with Intel® QuickAssist Technology (Intel® QAT) against the 4th Gen AMD EPYC 9354 processor (each with 32 cores) and found that the Intel solution uses up to 83% fewer cores to sustain the same number of NGINX secure key handshakes.[5] Intel testing also shows that a pre-production 4th Gen Intel Xeon Scalable processor with 60 cores requires up to 66% fewer cores to sustain the same number of IPsec connections, compared to the 3rd Gen AMD EPYC 7763 processor with 64 cores.[12]

> The 4th Gen Intel® Xeon® Platinum 8462Y processor with Intel® QAT uses up to **83% fewer cores to sustain the same number of NGINX® key handshakes** than the 4th Gen AMD EPYC™ 9354 processor.[5]

Fewer cores mean lower power usage, which can help you lower your TCO in a couple of ways. You will automatically see a much smaller power bill for the same connection capacity. Or you can spend the power savings on increasing the number of secure connections your web server can support without taking up more space in your server racks.

For more information on Intel QAT, see the white paper.

Accelerate Big Data Analytics and Databases with Intel® IAA

Many organizations rely on big data analytics running on top of powerful databases to process massive volumes of unstructured data. The ridesharing company Uber uses big data analytics to make real-time decisions about driver incentives, accident-avoidance predictions, vehicle positioning, service coverage, and surge pricing.[13]

> “Data is the biggest asset for Uber and its complete business model is based on the big data principle of crowdsourcing.” — ProjectPro[13]

In 2019, Uber migrated its MySQL® databases to MyRocks, an open-source, distributed database engine that integrates with the RocksDB database. The company made the switch to boost database performance, improve resource utilization, and lower its database TCO.[14]

As an example, we suggest that Uber could improve the performance of its big data analytics servers by using CPUs with built-in accelerators. Intel® In-Memory Analytics Accelerator (Intel® IAA) offloads database encryption, analytics, and compression from the CPU, which helps accelerate big data analytics workloads. Intel testing shows that a pre-production 4th Gen Intel Xeon Scalable processor (60 cores) with built-in Intel IAA can deliver up to 1.9x higher throughput and 47% lower latency on RocksDB than the 3rd Gen AMD EPYC 7763 processor (64 cores).[15]

A pre-production 4th Gen Intel® Xeon® Scalable processor with built-in Intel® IAA can deliver up to **1.9x higher throughput** and **47% lower latency** on RocksDB than the 3rd Gen AMD EPYC™ 7763 processor.[15]

In addition to boosting overall performance, the CPU-accelerator combination can deliver higher performance per core than a general-purpose CPU alone, which can help lower TCO. Better performance per core can also help cut costs when it's time to expand server infrastructures, due to using less power and taking up less physical space.

For more information on using Intel IAA, see the enabling guide.

Accelerate Storage with Intel® DSA

Faced with the need to store oceans of data coming from smart edge and Internet of Things (IoT) devices, organizations must figure out how to prevent storage bottlenecks that can slow down data access and delivery. One high-speed storage solution that data centers can implement is an NVM Express® (NVMe®)-over-TCP (NVMe/TCP) flash storage network that runs on industry-standard Ethernet hardware.

The Intel® Data Streaming Accelerator (Intel® DSA) built into 4th Gen Intel Xeon Scalable processors is designed to accelerate high-speed storage performance. Intel DSA offloads data-movement tasks from the CPU cores, accelerating storage performance and freeing up CPU cycles for other tasks. Such CPU offloads can result in higher overall storage performance and higher performance per core compared to having the CPU handle all data-movement tasks.

For example, pre-production 4th Gen Intel Xeon Scalable processors with 60 cores and Intel DSA can outperform 3rd Gen AMD EPYC 7763 processors with 64 cores and without accelerators. Compared to the AMD EPYC CPUs, the 4th Gen Intel Xeon Scalable CPUs with built-in accelerators can deliver up to 2.5x higher performance and 60% lower latency for large media file requests running on NVMe/TCP storage.[16]

For more information on using Intel DSA, see the user guide.

Using fewer cores, a pre-production 4th Gen Intel® Xeon® Scalable processor with Intel® DSA can deliver up to **2.5x higher performance** and **60% lower latency** for NVMe®/TCP storage than the 3rd Gen AMD EPYC™ 7763 processor.[16]

Conclusion

If your organization uses data to drive business decisions, you need solutions that can deliver high speed and throughput. At the same time, server performance must be balanced with finding ways to lower data center TCO.

Based on our findings, we believe that built-in accelerators can make a significant difference when it comes to meeting your performance needs, while also reducing your TCO. 4th Gen Intel Xeon Scalable processors come with a wide variety of built-in accelerators, making them good choices for some of today's most popular and demanding workloads: AI/ML, HPC, web services, data analytics, databases, and storage.

Learn More

Find out more about Intel Xeon Scalable processors with built-in accelerators at www.intel.com/acceleratorengines.

[1] Extreme Tech. "Nvidia's H100 AI Processors Are Selling for Over $40,000 on eBay." April 2023.

[2] StorageReview. "Intel Habana Gaudi2 Accelerators Offer NVIDIA Alternative for Large Language Models." July 2023.

[3] See claim [E2] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[4] Intel. "4th Gen Intel® Xeon® Scalable Processors with Built-In Accelerators." Accessed September 2023.

[5] Intel. "Why Choose Intel for Business? Solutions for Real-World Workloads." July 2023.

[6] See claim [A219] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[7] Intel. "Intel® Advanced Matrix Extensions (Intel® AMX) Enhances AI Inference Performance for Alibaba Cloud Address Purification." Accessed September 2023.

[8] See claim [H16] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[9] Intel. "Kyoto University Extends HPC Capacity." Accessed September 2023.

[10] W3Techs. "Usage statistics of Nginx." Accessed August 2023.

[11] Kinsta. "What Is Nginx? A Basic Look at What It Is and How It Works." January 2022.

[12] See claim [N201] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[13] ProjectPro. "How Uber uses data science to reinvent transportation?" July 2023.

[14] Uber. "MySQL to MyRocks Migration in Uber's Distributed Datastores." September 2022.

[15] See claim [D201] at intel.com/processorclaims: 4thGen Intel® Xeon® Scalable processors. Results may vary.

[16] See claim [N204] at intel.com/processorclaims: 4th Gen Intel® Xeon® Scalable processors. Results may vary.

PROWESS

The analysis in this document was done by Prowess Consulting and commissioned by Intel.
Results have been simulated and are provided for informational purposes only.
Any difference in system hardware or software design or configuration may affect actual performance.
Prowess Consulting and the Prowess logo are trademarks of Prowess Consulting, LLC.

0923/220091