FAQ

intel.

The Sustainable Data Center

More energy-efficient artificial intelligence (AI)

How can I execute AI model development, training, and inferencing more sustainably?

Look for hardware solutions that can help deliver resource-intensive AI workloads more efficiently and optimize performance-per-watt. No single processor is the answer for all AI solutions. A heterogeneous portfolio of hardware, optimized for AI, helps ensure your data center is working at optimal efficiency. Only Intel® offers an end-to-end portfolio of AI-optimized hardware, combined with a comprehensive, interoperable suite of AI software tools and framework optimizations to accelerate your AI workflows at every stage. With more acceleration capabilities and optimized performance, Intel® hardware and software solutions can help execute AI functions more quickly and help your data center operate with more energy efficiency.

Our newest Intel® Xeon® scalable processors with built-in accelerator engines provide improved workload results with greater energy efficiency. 4th Gen Intel® Xeon® scalable processors and Intel® Advanced Matrix Extensions (AMX) deliver 8x to 14x higher performance per watt across common AI workloads versus the same workloads on the same processor without acceleration.[1]

Intel® Xeon® scalable processors also deliver better performance and energy efficiency than competitive CPUs. When running a BERTLarge natural language processing workload, a 4th Gen Intel Xeon processor delivered 4.7x better performance/watt than the same workload on AMD's 4th Gen (Genoa) EPYC processor.[2]

Intel dedicated AI accelerators can also provide higher throughput with less power for high-end training and inference performance, including with generative AI and large language models (LLMs). The Habana® Gaudi®2 accelerator delivers 2x higher throughput per watt than the comparable NVIDIA A100, giving you more training with lower power consumption.[3] On inference workloads with BLOOM 176B large language model (LMM), studies show Guadi2 uses 22 percent lower power, with 1.6x power-performance than A100.[4]

Intel can also help you find efficiency with optimized software for the most common AI scenarios. Optimizations in Intel® oneAPI deep learning libraries can provide a 16x gain in image classification inference and a 10x gain for object detection with TensorFlow, as well as a 53x gain for image classification and nearly 5x gain for recommendation systems with PyTorch, giving you more insights for less power.[5]

For additional recommendations on how to design AI projects more sustainably, read our article in MIT Technology Review.

How can Intel help me use AI to reduce energy consumption in the data center?

AI has great untapped potential to find energy saving efficiencies across the enterprise, and the data center is no exception. You can lower your data center's carbon footprint using Intel tools to develop a more dynamic, carbon-aware approach to computing.

Using built-in telemetry capabilities in the latest Xeon processors to help ensure your data center is operating at peak efficiency, you can get real-time insights into power efficiency, thermals, resource utilization, and general system health. By integrating this telemetry data with intelligent data center infrastructure management tools, like server management tools from leading OEMs, you can automatically orchestrate adjustments to optimize energy use and detect anomalies to proactively identify issues before problems arise.

Or pair the rich telemetry capabilities of our latest Xeon® scalable processors with Intel's tools in Kubernetes to create a more proactive approach to data center management to increase carbon-efficiency and reduce energy usage.

Advanced Cluster Systems, Inc. Ex. 2019
Page 1 of 5

- Utilize machine learning to predict peak times for computing and fine tune power use in Kubernetes clusters with Intel Power Manager in Kubernetes. Spin up nodes in advance for rapid response while reducing idle energy use while reducing latency. And at off-peak times, you can easily move nodes to a power saving profile, conserving energy.
- Selectively increase or decrease lower priority workloads based on renewable energy availability. With built-in telemetry tools in Intel® Xeon® Scalable processors and Intel's Telemetry Aware Scheduling (TAS) in Kubernetes, you can ramp up intensive compute tasks when more low- or zero-carbon energy is available.
- Automate and optimize cloud workloads with Kubernetes policy-based software tools and maximize cost and energy efficacy with Intel Cloud Native Orchestration in Kubernetes.

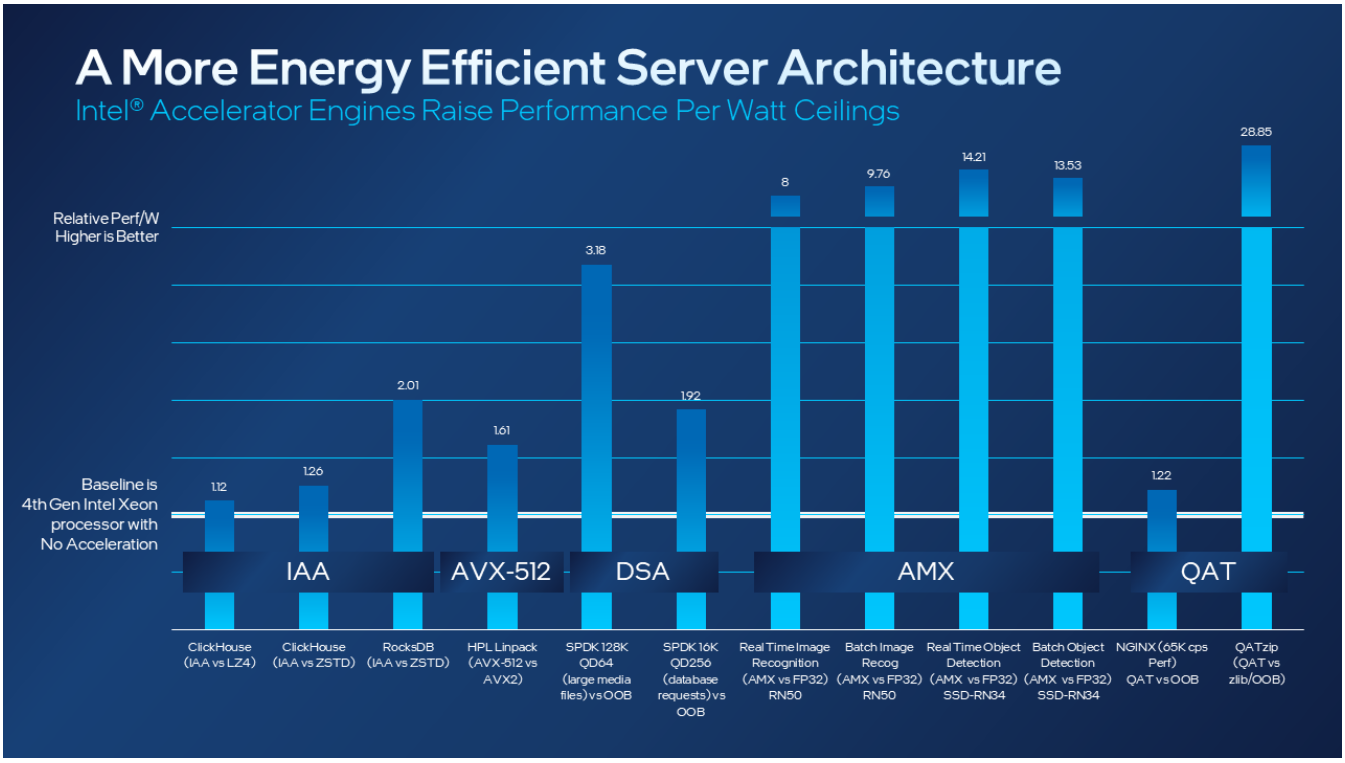
Or simply reap the benefits of AI by automatically optimizing your on-prem, hybrid and/or cloud infrastructure through real-time performance insights and adjustments with Intel® Granulate.™ This application and workload performance optimization solution can improve data center compute performance by up to 60 percent and reduce costs by up to 30 percent without ever having to change application code.6 Granulate also offers a "CO2 Savings Meter" allowing you to easily measure the impact of workload optimization on your data center's carbon footprint alongside cost and resource reductions.

Why should I go through the work of implementing built-in accelerators?

While some initial set-up is required, Intel Accelerator Engines offer massive payoffs overall. These high-bandwidth memory and software optimizations provide an alternative, more efficient way to achieve higher workload performance than simply growing the CPU core count. For example, the Intel® Advanced Matrix Extensions accelerator, built into 4th Gen Xeon processors can quickly accelerate your natural language processing (NLP). You can see 2x-3x performance gains on Bidirectional Encoder Representations from Transformers (BERT) model throughput with 4th Gen Intel® Xeon® Scalable processors and Intel® AMX versus the previous generation CPUs.[7]

The choice to utilize these built-in accelerators can greatly impact the carbon impact and the TCO of your servers. Testing shows 4th Gen Xeon® scalable processors offer as much as 79 percent[8] lower TCO than 4th Gen AMD EPYC (Genoa) while running a BERTLarge natural language processing (NLP) model workload, a difference of 424 kW of energy and 719,546 kg CO2 emissions over 4 years.[9] That's the equivalent of almost 615 round-trip passenger flights between New York City and San Francisco.[10]

And it's not just for NLP. Here's a look at the performance/watt gains you could see by implementing built-in accelerators:



I don't have the expertise in-house to implement a discrete accelerator. Is there another way to run more powerful AI/ machine learning workloads?

Intel® Habana® Gaudi accelerators are also available in cloud environments. You can get started running models on the Intel Developer Cloud or Amazon EC2 DL1 instances and experience low cost-to-train deep learning models for natural language processing, object detection, and image recognition use cases. According to AWS, DL1 instances provide up to 40 percent better price performance for training deep learning models compared to current generation GPU-based EC2 instances.[11] Now you can train more and pay less, accelerating time-to-market with model training.

Aren't AMD CPUs more energy efficient?

While AMD EPYC processors have shown performance per watt advantages in general computing scenarios, Intel® 4th Generation Xeon® processors outperform 4th Gen. EPYC processors (Genoa) when utilizing built-in accelerators. Testing shows 4th Gen Xeon® processors have 50 percent higher performance/watt and offer 20 percent savings in carbon emissions.[12]

And the performance difference becomes even more apparent in compute-intensive workloads like AI. Testing shows 4th Gen Xeon processors have 80 percent higher inference throughput than 4th Gen AMD EPYC CPUs.[13] When running a Deep Learning Recommendation Model (DLRM), 4th Gen Xeon® scalable processors offer up to 61 percent lower TCO than 4th Gen AMD EPYC (Genoa), with just 17 Xeon-based servers doing the work of 50 EPYC servers. This efficiency can save up 361.2 kW of energy and 612,543 kg of CO2 emissions over 4 years.[14] That's the equivalent of more than 520 round-trip passenger flights between New York City and San Francisco.[15]

We also believe a sustainable data center is about more than the performance of one component. Beyond our CPUs, Intel is focused on helping optimize operations across the entire data center, including AI workloads, with our full portfolio of hardware, software, services, tools, and support. This unique combination of hardware and software support drives energy efficiency and lowers costs across network, storage, cooling, and power, as well as server operations.

Why should I consider Intel for AI rather than NVIDIA?

We don't contest that NVIDIA is the market share leader in AI training and certain HPC (High Performance Computing) computing workloads with their GPUs and dedicated accelerator hardware. However, Intel's offerings remain competitive, even surpassing some NVIDIA products in some workloads and configurations, and Intel is leading in AI inferencing. Third-party evaluation shows that the Habana® Gaudi2 server not only consumes less power than the comparable NVIDIA A100, but also delivers up to 2x higher throughput/watt than the A100 on ResNet50 image classification workloads.[3] Gaudi2 was also shown to perform 1.4x faster than A100-80G for BLOOM 176B inference, using 22% lower power, a power-performance 1.6x better than A100.[4]

We also know there is no one-size-fits-all architecture for the diversity of today's AI needs. Intel offers truly heterogeneous computing, giving you the ability to combine general-purpose compute with dedicated, AI-specific resources. Specialized architectures can optimize performance, power, or latency as needed, and only Intel® offers the variety of data center hardware solutions required to customize to business needs from the edge to the cloud.

Intel® is also unique in offering an end-to-end AI software ecosystem, built on an open, interoperable programming model (oneAPI), coupled with an extensible, heterogeneous AI compute infrastructure. OneAPI can help maximize cross-architecture performance and reduce development costs. With a single developer code base across multiple architectures, one AI application can be simply deployed across diverse architectures, opening opportunities from cloud to edge, without leaving any performance behind.

How can more effective server cooling support my company's sustainability goals?

According to Gartner, 40 percent of data center energy consumption is spent on cooling,[16] presenting a huge opportunity to optimize energy efficiency and resource usage, especially for growing AI workloads. While many air-cooling solutions exist, including enhanced system air cooling and AI-assisted automatic cooling, none offer the same ability to optimize power usage effectiveness (PUE) as well as using liquid cooling technology.

Liquid cooling can not only help increase server density, as well as space and energy efficiency, but it can also help extend hardware life. Just a 10°C decrease in average operating temperature can more than double the lifetime of the semiconductor.[17]

There are options for almost any data center scenario: brownfield or greenfield deployments; small- and large-scale operations; and at the server, rack, or system level. And optimized solutions can maintain existing performance levels without increasing costs.

Cold plate solutions work well at the individual component level, offering scalable deployment and easy retrofitting of existing infrastructure without adding weight to the system. Or for warm ambient air environments, or areas with high humidity or pollution, immersion solutions enable system-level cooling with heat recapture and reuse benefits.

Intel is on the leading edge in liquid cooling, introducing the first open IP immersion liquid cooling solution and reference design, enabling partners to accelerate development and improve energy efficiencies.

And we continue to co-innovate with our ecosystem of partners. One Intel partner, Hypertec, has a solution which allows customers to save up to 95 percent on data center cooling OPEX, while also prolonging hardware lifespan 30 percent,[18] offering with a nearly 50 percent reduction in power consumption.[19]

Intel offers processor SKUs optimized for liquid-cooled systems, with an immersion cooling warranty rider available, as well as providing performance validation for cold plate systems to help ensure reliability at scale.

Advanced Cluster Systems, Inc. Ex. 2019
Page 3 of 5

Endnotes

1 Up to 8x and 9.76x higher performance/W using 4th Gen Xeon Scalable w/Advanced Matrix Extensions using AMX vs VNNI instructions on ResNet50 Image Processing. 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), on pre-production Supermicro SYS-221H-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=Resnet 50 v1_5, best scores achieved: BS1 FP32 8 cores/instance (max. 15ms SLA), BS1 INT8 2 cores/instance (max. 15ms SLA), BS1 AMX 1 core/instance (max. 15ms SLA), BS16 FP32 5 cores/instance, BS16 INT8 5 cores/instance, BS16 AMX 5 cores/instance, using physical cores, tested by Intel November 2022. Up to 14.21x and 13.53x higher performance/W using 4th Gen Intel Xeon Scalable w/Advanced Matrix Extensions using AMX vs VNNI instructions on SSD-ResNet34 on Object Detection. 1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), Intel platform with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=SSD-ResNet34, best scores achieved: BS1 FP32 60 cores/instance (max. 100ms SLA), BS1 INT8 4 cores/instance (max. 100ms SLA), BS1 AMX 4 core/instance (max. 100ms SLA), BS8 FP32 8 cores/instance, BS2 INT8 1 cores/instance, BS2 AMX 1 cores/instance, using physical cores, tested by Intel November 2022.

2 4th Gen Xeon Outperforms Competition on Real-World Workloads, slide 4, accessed October 11, 2023.

3 Power performance measurements for power utilization on ResNet performed by Supermicro in their lab (April 2023). Configuration details available at Habana website: https://habana.ai/habana-claims-validation

4 Power performance metrics on BLOOMz 176B: performance evaluated and published by Hugging face here: https://huggingface.co/blog/habana-gaudi-2-bloom#latency and power utilization measured by Habana Labs in April 2023; configuration details published here: https://habana.ai/habana-claims-validation

5 See additional configuration details

6 Intel Granulate | Autonomous Optimization for Intel Processors

7 Improved Machine Learning with Intel AMX

8 4th Gen Xeon Outperforms Competition on Real-World Workloads, slide 5, 17, accessed October 11, 2023

9 4th Gen Xeon Outperforms Competition on Real-World Workloads, slide 5 , accessed October 11, 2023.

10 Calculations based on data from Carbon Footprint Calculator

11 Amazon EC2 DL1 Instances - Amazon Web Services

12 See performance footnotes for workloads and configurations

13 See performance footnotes for workloads and configurations

14 Results are based on DLRM. https://www.intel.com/content/www/us/en/content-details/781683/4th-gen-xeonoutperforms-competition-on-real-world-workloads.html, slide 5, 16, accessed September 6, 2023.

15 Calculations based on data from Carbon Footprint Calculator

16 Gartner, "How can sustainability drive data center infrastructure cost optimization?," November 2022.

17 Device Reliability - How Temperature Affects Mean Time to Failure (jetcool.com)

18 https://hypertec.com/immersion-cooling

19 https://www.intel.com/content/www/us/en/products/docs/processors/xeon/scalable/hypertec-video.html

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration 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.

"A more energy efficient server architecture" chart configurations:

Up to 1.12x and 1.26x higher performance/W using 4th Gen Xeon Scalable w/Intel Analytics Accelerator vs LZ4 and Zstd on ClickHouse

1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=8(2 sockets active), on pre-production Intel platform and software, HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.1.21, accel-config-v3.4.6.4, gcc 11.2, Clickhouse 21.12, Star Schema Benchmark, tested by Intel November 2022.

Up to 2.01x higher performance/W using 4th Gen Xeon Scalable w/Intel Analytics Accelerator vs Zstd on RocksDB

1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), on pre-production Intel platform and software, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, QPL v0.2.1,accel-config-v3.4.6.4, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022.

Up to 1.61 higher performance/W using 4th Gen Xeon Scalable w/AVX-512 vs AVX2 on Linpack

1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core), on pre-production Supermicro SYS-221H-TNR and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC 4, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, One API BaseKit 2022.2.0.262, One API HPC 2022.2.0.191, Linpack ver 2.3, tested by Intel November 2022.

Up to 3.18x and 1.92x higher performance/W using 4th Gen Xeon Scalable w/Data Streaming Accelerator vs out-of-box OS software on SPDK NVMe TCP

1-node, 2x pre-production 4th Gen Intel Xeon Scalable processor (60 core) with integrated Intel Data Streaming Accelerator (Intel DSA), DSA device utilized=1(1 active socket), on pre-production Intel platform and software with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x 1.92TB Intel® SSDSC2KG01, 4x 1.92TB Samsung PM1733, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 2x100GbE, FIO v3.30, SPDK 22.05, tested by Intel November 2022.

Up to 8x and 9.76x higher performance/W using 4th Gen Xeon Scalable w/Advanced Matrix Extensions using AMX vs VNNI instructions on ResNet50 Image Processing

1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), on pre-production Supermicro SYS-221H-TNR with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000c0, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=Resnet 50 v1_5, best scores achieved: BS1 FP32 8 cores/instance (max. 15ms SLA), BS1 INT8 2 cores/instance (max. 15ms SLA), BS1 AMX 1 core/instance (max. 15ms SLA), BS16 FP32 5 cores/instance, BS16 INT8 5 cores/instance, BS16 AMX 5 cores/instance, using physical cores, tested by Intel November 2022.

Up to 14.21x and 13.53x higher performance/W using 4th Gen Intel Xeon Scalable w/Advanced Matrix Extensions using AMX vs VNNI instructions on SSD-ResNet34 on Object Detection

1-node, 2x pre-production 4th Gen Intel® Xeon® Scalable processor (60 core) with Intel® Advanced Matrix Extensions (Intel AMX), Intel platform with 1024GB DDR5 memory (16x64 GB), microcode 0x2b0000a1, HT On, Turbo On, SNC Off, CentOS Stream 8, 5.19.16-301.fc37.x86_64, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Intel TF 2.10, AI Model=SSD-ResNet34, best scores achieved: BS1 FP32 60 cores/instance (max. 100ms SLA), BS1 INT8 4 cores/instance (max. 100ms SLA), BS1 AMX 4 core/instance (max. 100ms SLA), BS8 FP32 8 cores/instance, BS2 INT8 1 cores/instance, BS2 AMX 1 cores/instance, using physical cores, tested by Intel November 2022.

Up to 1.22x higher performance/W using 4th Gen Intel Xeon Scalable w/QuickAssist Accelerator vs out-of-box software on NGINX TLS Handshake.

QAT Accelerator: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel QuickAssist Accelerator (Intel QAT), Number of QAT device utilized=4(1 socket active), on pre-production Intel platform and software with DDR5 memory total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, QAT engine v0.6.14, QAT v20.l.0.9.1, NGINX 1.20.1, OpenSSL 1.1.1l, IPP crypto v2021_5, IPSec v1.1, TLS 1.3 AES_128_GCM_SHA256, ECDHE-X25519-RSA2K, 65K CPS target SLA, tested by Intel November 2022. Out of box configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable Processor (60 cores) with integrated Intel QuickAssist Accelerator (Intel QAT), Number of QAT device utilized=0, on pre-production Intel platform and software with DDR5 memory total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX 1.20.1, OpenSSL 1.1.1l, TLS 1.3 AES_128_GCM_SHA256, ECDHE-X25519-RSA2K, 65K CPS target SLA, tested by Intel November 2022.

Up to 28.85x higher performance/W using 4th Gen Intel Xeon Scalable w/QuickAssist Accelerator vs out-of-box zlib on QATzip compression

1-node, 2x pre-production 4th Gen Intel® Xeon Scalable Processor (60 core) with integrated Intel QuickAssist Accelerator (Intel QAT), QAT device utilized=8(2 sockets active), on pre-production Intel platform and software with DDR5 memory Total 1024GB (16x64 GB), microcode 0x2b0000a1, HT On, Turbo Off, SNC Off, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, QAT v20.l.0.9.1, QATzip v1.0.9, tested by Intel November 2022.