FAQ

The Sustainable Data Center

Going beyond server refresh and consolidation

How can a refresh to servers with the latest Intel® Xeon® processors impact my data center's energy use and carbon footprint?

Our 4th Gen Intel® Xeon® scalable processors are Intel's most sustainable data center processors ever, with the most built-in accelerators ever offered in an Intel processor, giving you more performance-per-watt improvements than ever.

By replacing servers built on 1st and 2nd Gen Intel Xeon processors, you can significantly reduce your operational carbon footprint—and your power bill. Refreshing and consolidating from 1st Gen Intel Xeon processors to 4th Gen Intel Xeon processors can result in up to a 60 percent reduction in CO2 emissions and power, as well as providing up to 75 percent reduction in total cost of ownership (TCO). With these savings, you can recover your costs in just 4 months.[1]

Even compared to servers with 3rd Gen Xeon processors, 4th Gen Xeon hardware offers significant energy savings. For example, instead of deploying 50 3rd Gen Intel Xeon® servers for a RocksDB database workload, you can see the same performance with just 18 4th Gen Xeon processor-based servers with our new Intel® In-Memory Analytics Accelerator (IAA). In this case, using the latest generation can save 15.4 kW of energy over 4 years, reducing carbon emissions by 366 metric tons while also giving you 52 percent lower TCO.[2]

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.[3]

And when you compare performance on key, compute-intensive workloads like AI (Artificial Intelligence), Intel's leadership is clear. Our 4th Gen Xeon® scalable processors offer up to 79 percent lower TCO than 4th Gen AMD EPYC (Genoa) while running a BERTLarge natural language processing (NLP) model workload, a difference of 424.3 kW of energy and 719,546g of CO2 emissions over 4 years.[4] That's the equivalent of almost 615 round-trip passenger flights between New York City and San Francisco.[5]

We also believe a sustainable data center is about more than the performance of one component. Intel is focused on helping optimize operations across your data center with our portfolio of hardware, software, services, tools, and support. This unique combination of hardware and software support will drive energy efficiency and lower costs across network, storage, cooling, and power, as well as server operations.

How can a workload-first approach benefit the efficiency of my data center?

Intel's strategy of aligning CPU cores with built-in accelerators, alongside optimized software optimized for specific workloads delivers superior performance at higher efficiency, helping to optimize your total cost of ownership across the data center.

By choosing the newest Xeon® scalable processors, you gain more flexibility to shift from general compute needs to optimizing for more specialized needs, accelerating workloads like AI, data analytics, networking, security, storage, and HPC (High Performance Computing) with Intel accelerator engines. These accelerator engines target key real-world workloads to help raise performance per watt ceilings, offering you improved performance and energy efficiency. For example, 4th Gen Intel® Xeon® scalable processors and the built-in Intel® Advanced Matrix Extensions (AMX) accelerator deliver 8x to 14x higher performance per watt across common AI workloads versus the same processor without acceleration.[6]

Intel software tools can also help ensure efficiency on some of your most compute- and energy-intensive emerging workloads with optimized software for the most common scenarios. For AI, Intel® optimizations based in Intel® oneAPI deep learning libraries can provide 16x gain in image classification inference, a 10x gain for object detection, a 53x gain for image classification and a nearly 5x gain for recommendation systems, giving you more insights for less power.[7]

Across workloads, the Intel software stack can help accelerate your data center, free from proprietary lock-in. The oneAPI initiative and other Intel open-source projects encourage collaboration on the specification and compatible implementations across the ecosystem, including oneAPI plugins for NVIDIA and AMD hardware.

How can a server refresh with Intel help increase reliability in my data center?

The latest Intel hardware can accelerate network infrastructure and free up CPU cores for value-generating compute. Our 4th Generation Intel® Xeon® processors offer more integrated IP accelerators on chip than ever, helping increase the effectiveness of cores by enabling offload of common mode tasks via seamlessly integrated acceleration engines like Intel Data Streaming Accelerator (DSA), Intel Quick Assist Technology (QAT), and Intel Dynamic Load Balancer (DLB). These integrated acceleration engines help free up cores for more general-purpose compute tasks, increasing your overall workload performance and energy efficiency (performance/watt), as well as reducing latency and overall service jitter.

With the newest Xeon® scalable processors you can have more efficient database back-ups than with the competition. Just 29 4th Gen Intel Xeon servers running a Microsoft SQL 2022 + QAT Backup could do the same work of 50 4th Gen AMD EPYC (Genoa) servers, an energy difference of 145.1 kW and almost 246 metric tons of carbon emissions over 4 years. Not to mention 35 percent savings in TCO.[8]

The latest Intel Xeon servers can also help ensure faster data processing and reduce the risk of downtime due to memory failure. 4th Gen Xeon® scalable processors offer up to 50 percent higher memory bandwidth (DDR5)[9] and a 2x PCIe bandwidth improvement[10] over the previous generation.

Incorporating other Intel hardware technologies in your refresh and consolidation process can likewise increase the efficiency of your data center. For example, you can use Intel Infrastructure Processing Units (IPUs) to offload networking tasks from your CPU and manage infrastructure services such as virtual switching, security, and storage, saving you CPU cycles and giving you greater compute capacity.

How can I implement more effective server cooling to reduce the power use and carbon footprint of my data center?

According to Gartner, 40 percent of data center energy consumption is spent on cooling,[11] presenting a huge opportunity to optimize energy efficiency and resource usage. 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 liquid cooling technology.

Liquid cooling can not only help increase server density, as well as physical 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.[12]

There are viable liquid cooling options for almost any data center scenario – both brownfield and greenfield deployments, small- and large-scale operations, and at the server, rack, or system level. You can create an optimized solution that maintains 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.
- Immersion solutions offer efficient cooling in warm ambient air environments, or areas with high humidity or pollution, while also offering efficient system-level cooling with the bonus of heat recapture and reuse benefits. Upfront capital investments are manageable, and density advantages are clear.

Intel is on the leading edge in liquid cooling, co-innovating 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,[13] offering with a nearly 50 percent reduction in power consumption.[14]

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.

What other Intel resources can support my hardware refresh in bringing more efficiency to my data center?

In addition to the energy efficiency gains inherent in deploying new hardware with the latest Xeon® scalable processors, you can lower your data center's carbon footprint by implementing more tools for dynamic, carbon-aware computing.

Take advantage of the telemetry capabilities of 4th generation Xeon® scalable processors and Intel's tools in Kubernetes to develop a more proactive approach to data center management to not only reduce downtime but also increase carbon-efficiency and data center reliability.

- 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 automatically optimize your on-prem, hybrid, and cloud infrastructure with real-time, AI-based 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.15 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.

How can Intel help me reduce my e-waste footprint?

By choosing the 4th Gen Xeon® scalable processor, you gain more flexibility to shift and scale from today's compute needs to more specialized needs, like AI, HPC, and data analytics. And with Intel® OnDemand, you can activate built-in accelerators and other hardware-enhanced features when you need them, now or in the future, to ensure your workloads meet your performance needs.

At Intel, we also strongly believe in open-standards foundations and strive to deliver hardware and software interoperability across the data center—including with our competitors' solutions—helping ensure our customer's newest solutions are compatible with their existing infrastructure. Intel is also aiming to reduce waste and extend the life of your hardware by advancing modular architectures, enabling components to be easily replaced or upgraded without discarding the entire system. The first specification for this, Data Center Modular Hardware System (DC MHS) is available now through the Open Compute Project.

Intel IT has implemented a similar disaggregated server architecture approach within our own data centers. By decoupling the CPU/DRAM and NIC/Drives modules from other server components, we can independently refresh servers' CPU and memory without replacing other server components that are not yet ready for end-of-life, such as fans, power supplies, cables, network switches, drives, add-on modules/accelerators, and chassis. This results in faster technology adoption, which in turn puts new, innovative technology at our Intel users' fingertips faster. With this approach, Intel IT has seen[16]:

- Refresh costs cut by a minimum of 44 percent
- Reduced technician time spent on refresh by 77 percent
- Decreased shipping weight for refresh materials by 82 percent

Read how Intel IT is approaching its data center transformation strategy

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.

Endnotes

1 Refresh and Consolidate: 1st Gen to 4th Gen Intel Xeon Processor-based Servers, slide 3, Accessed October 11, 2023.

2 New Configuration: 1-node, 2x pre-production 4th Gen Intel Xeon Scalable 8490H 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. Baseline: 1-node, 2x production 3rd Gen Intel Xeon Scalable 8380 Processor (40 cores) on SuperMicro SYS-220U-TNR , HT On, Turbo On, SNC Off, Total Memory 1024GB (16x64GB DDR4 3200), microcode 0xd000375, 1x3.84TB P5510 NVMe, 10GbE x540-AT2, Ubuntu 22.04.1 LTS, 5.18.12-051812-generic, ZSTD v1.5.2, RocksDB v6.4.6 (db_bench), tested by Intel November 2022. For a 50 server fleet of 3rd Gen Xeon 8380 (RocksDB), estimated as of November 2022: CapEx costs: $1.64M, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $677.7K, Energy use in kWh (4 year, per server): 32181, PUE 1.6, Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394. For an 18-server fleet of 4th Gen Xeon 8490H (RockDB w/IAA), estimated as of November 2022: CapEx costs: $846.4K, OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $260.6K, Energy use in kWh (4 year, per server): 41444, PUE 1.6, Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

3 Geomean of 5 workloads [PostgreSQL, Microsoft SQL Database with a Backup Function w/QAT, BlackScholes, DLRM PyTorch BF16, DeathStarBench – Hotel Reservation] benefitting from acceleration. See performance footnotes for workloads and configurations

4 Results are based on BERT Large. 4th Gen Xeon Outperforms Competition on Real-World Workloads, slide 5, 17, accessed July 6, 2023.

5 Calculations based on data from Carbon Footprint Calculator

6 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.

7 See additional configuration details

8 Intel 8462Y+: 1-node, 2x 4th Gen Intel Xeon Platinum 8462Y+ w/Integrated Quick Assist Accelerator, 32 cores, HT On, Turbo On, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.1, microcode 0x2b000161, 2x Ethernet Controller 10-Gigabit X540-AT2, 7x 3.5T INTEL SSDPF2KX038TZ, Windows Server 2022, HammerDB 4.5, Microsoft SQL 2022-SSEI-Eval/ SQL Server Management Studio 19.0.1, QATZip 2.0.W.2.0.4-0004. Tested by Intel March 2023.

9 See [G2] at intel.com/processorclaims

10 See [G8] at intel.com/processorclaims

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

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

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

14 Hypertec Immersion Cooling for FSI and M&E

15 Intel Granulate | Autonomous Optimization for Intel Processors

16 PDF: IT@Intel: Data Center Strategy Leading Intel's Business Transformation