Right-sizing AI for energy efficiency from edge to cloud with Intel® Xeon® processors

Considerations for developing an environmentally responsible approach to AI transformation

Introduction—The double-edged sword of AI

From breakthrough advancements in disease diagnosis to optimizing cities' traffic flow to helping boost farms' crop yields, the potential for revolutionary applications of artificial intelligence (AI) is nearly endless.

Addressing climate change is no exception. A recent report from the Boston Consulting Group estimates that AI has the potential to mitigate 5-10 percent of global greenhouse gas emissions by 2030.[1] AI can offer tech positive transformation: helping measure and reduce greenhouse gas emissions, furthering climate-related research efforts, as well as boosting climate resilience with new early warning systems for extreme weather and adaptive solutions to help mitigate the impacts of climate change.

At the same time, AI is compute-intensive requiring large amounts of energy to train and run workloads, thereby increasing carbon emissions and natural resource consumption in computing environments. For example, training the BLOOM model, one of the most carbon-efficient major large language models (LLM) with 176 billion parameters, emitted 25 metric tons of CO2 emissions.[2] That's the equivalent carbon footprint 15 round-trip flights between New York and San Francisco. Meanwhile, it's estimated that GPT-3, with 175 billion parameters, emitted more than 500 metric tons of CO2.[2] At the same time, LLM model size has been increasing 10x every year for the last several years,[3] and a model-size "arms race" is underway with the GPT-4 model reportedly containing over 1 trillion parameters. Increased model size typically means increased complexity and training time, leading to higher energy and carbon footprints.

the ever-growing size of these models, AI is certainly not being confined to data centers. AI inferencing is proliferating from cloud to edge, and more and more intelligent systems, like self-driving cars, are requiring low-latency inference close to the data source.

No matter where these AI workloads are generating insight, they're also generating heat. With water-based evaporative cooling commonly being both the cheapest and the most energy efficient method to manage data center temperatures, higher computing requirements also means increased water consumption. Depending on the location of the data center this can be problematic. Studies estimate that one-fifth of US data centers draw their water from moderately to highly stressed watersheds,[4] and increased consumption would only exacerbate critical water management challenges.

There's a two-fold challenge for IT in creating sustainable change for their organization:

1. Reaching tech zero: Reducing the carbon footprint of an organization's IT function
2. Facilitating tech positive: Using technology as a lever for the whole organization to reach its net-zero goals and to have a positive overall impact, driving business growth and accelerating innovation

Read more about tech positive transformation and the role of technology in sustainable transformation here

Building an AI environment tailored for efficiency

At Intel, we believe environmental responsibility need not conflict with the cutting-edge transformation driven by AI. It is possible to responsibly deliver high-impact AI transformation across your organization with a lower environmental footprint. But it requires an intentional, right-sized approach to AI initiatives. Carefully selecting the right tools, software, hardware, and design is critical to achieving more sustainable AI.

Model Selection & Tuning

One of the primary methods for optimizing AI projects is building from the best model for your need. While there are certainly application use cases requiring some of the biggest LLMs, leveraging domain-specific models can often produce the required results with greater efficiency. Smaller datasets require less energy for training and have lighter ongoing compute and storage requirements, thereby producing fewer carbon emissions.

Beyond selecting smaller out-of-the-box datasets, optimizing domain-specific models with popular compression techniques such as quantization, pruning, and knowledge distillation can also foster further efficiency. Software developers can use tools and libraries like Intel® Neural Compressor and OpenVINO Runtime, it's possible to implement automatic accuracy-driven tuning strategies to easily generate quantized model and speed up inference. Studies show that many of the parameters within a trained neural network can be pruned by as much as 99 percent, yielding much smaller, more sparse networks.[5] These right-sized datasets are also better suited to edge computing scenarios, where size and power constraints create limitations, and make it possible to deploy AI workloads almost anywhere they're needed.

By also assessing the level of accuracy truly required for the AI use case, you can drive significant energy and compute-time savings. By fine-tuning models and utilizing lower 16-bit precision or mixed precision during training and inference, models can run faster and use less memory. With Intel optimizations built into Intel® oneAPI as well as the default of AI application libraries like TensorFlow, accelerating transfer learning and inference with automatic mixed precision is easy.

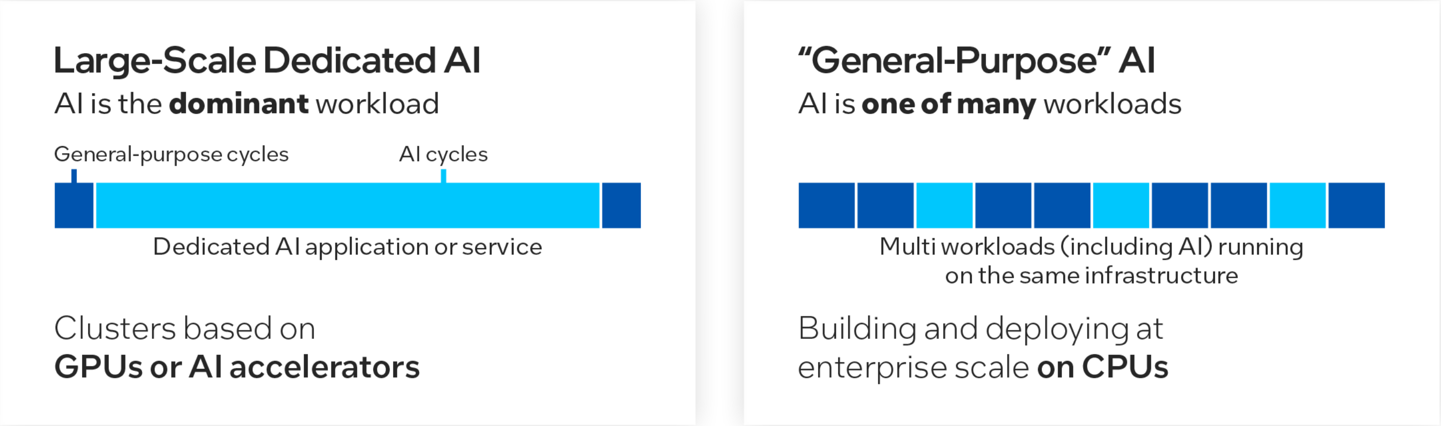
Learn more about getting started with automatic mixed precision

Hardware selection

While GPUs and dedicated accelerators, like Intel® Gaudi®, have become synonymous with AI, they should not have a monopoly on AI workloads. GPUs will continue to deliver the best performance/watt for dedicated AI training and some inference, but CPUs like our Intel® Xeon® Scalable processors offer integrated features that make advanced AI possible anywhere—including the network and/or IoT edge—no GPU or dedicated accelerator required.

For example, VMware recently took an off-the-shelf LLM (Llama2- 7B) and was able to stand up a financial services-specific chatbot running exclusively on a 5th Gen Intel® Xeon® Scalable processor with the built in Intel® Advanced Matrix Extensions (Intel AMX) accelerator. After just 3.5 hours fine-tuning with financial services information, the chatbot was able to answer basic questions about finance and financial terms.[6] No GPUs were used to do any of this.

For most companies, AI execution is not an around-the-clock workload. Discrete accelerators for dedicated AI only deliver the best TCO when clusters are being fully utilized, since the hardware will be idle when not being used for AI. Intel Xeon processors offer the flexibility to scale out AI workloads when required, while also running other general-purpose computing. CPU-based AI execution on Intel Xeon processors allows you to co-locate AI with the rest of your workloads and get the most out of your existing servers with better server utilization—helping your IT budgets stretch further.



Large-Scale Dedicated AI

AI is the **dominant** workload

Dedicated AI application or service

Clusters based on **GPUs or AI accelerators**

"General-Purpose" AI

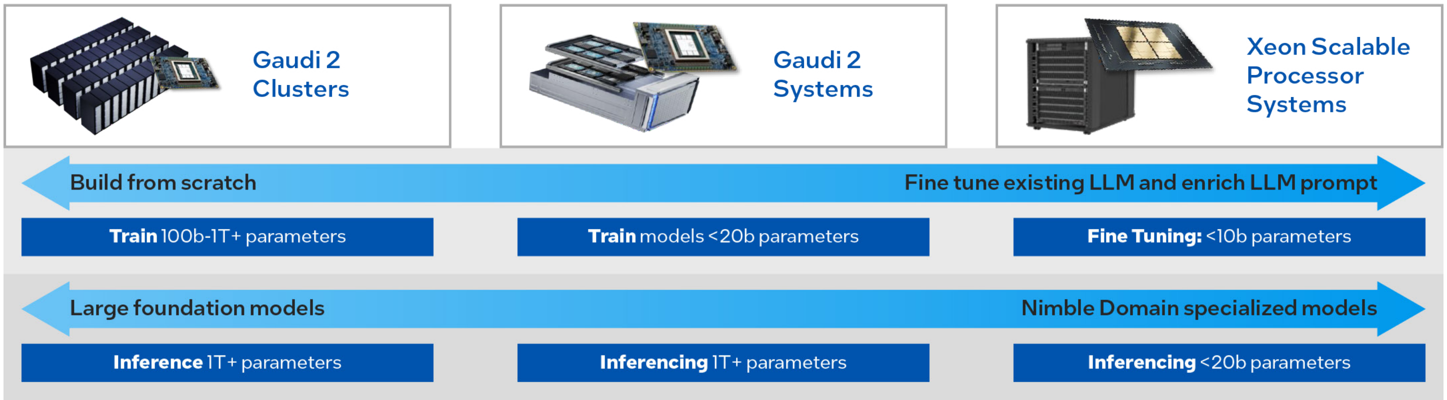
AI is **one of many** workloads

Multi workloads (including AI) running on the same infrastructure

Building and deploying at enterprise scale **on CPUs**

As AI applications span edge, data center, and cloud locations, the energy efficiency and smaller physical footprint of CPU-based processing is key. For example, on the IoT edge, CPUs offer both a more efficient model, while also delivering TCO benefits for mixing general purpose and AI workloads within one server. Intel Xeon processors offer additional flexibility and agility, whether in your own data center, a cloud service provider environment, or at the network and IoT edge.

Using a more heterogenous AI infrastructure, with a combination of chipsets selected to suit your application and compute location, will offer significant overall energy savings. While building or training large foundation models for generative AI may require dedicated accelerators like the Intel® Gaudi®2 system, fine-tuning existing LLMs and enriching LLM prompts or working with domain-specialized models are best suited to the more ubiquitous and flexible CPU of Intel Xeon Scalable processors.



Gaudi 2 Clusters

Gaudi 2 Systems

Xeon Scalable Processor Systems

Build from scratch

Fine tune existing LLM and enrich LLM prompt

Train 100b-1T+ parameters

Train models <20b parameters

Fine Tuning: <10b parameters

Large foundation models

Nimble Domain specialized models

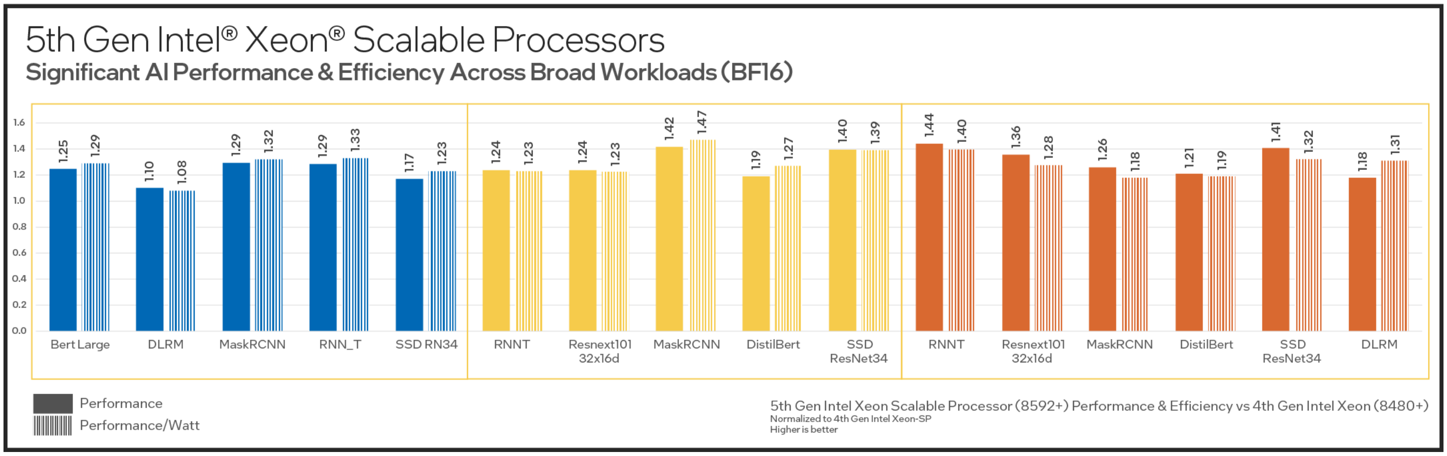
Inference 1T+ parameters

Inferencing 1T+ parameters

Inferencing <20b parameters

Intel Xeon processors are particularly well-suited for inference tasks using with nimble, domain-specialized models. For example, using our latest 5th Gen Intel® Xeon® Scalable processors, we fine-tuned BioGPT— a 1.5 billion parameter biomedical generative Transformer language model pre-trained on large-scale biomedical literature—on PubMedQA dataset using HuggingFace APIs. After just 24.8 minutes of fine-tuning with a 4-node 5th Gen Xeon processor, the model was able to generate a response text in 'yes/no/maybe' to answer for a given question and context. For comparison, the same fine-tuning to SOTA (state-of-the-art) accuracy of 79.4 percent takes 20.2 minutes on an Nvidia A100 GPU.[7]

Designed to address the new environment where AI is everywhere, 5th Gen Intel Xeon Scalable processors continue to deliver the top performance expected of the Xeon family while leveraging architectural advancements to allow for more energy efficient compute. Offering higher compute power with larger core count and better Intel AMX and Intel® AVX (Advanced Vector Extensions) frequencies, plus better cache and memory capabilities and additional software optimizations, the newest Intel Xeon processors can help take efficient AI compute to the next level.



5th Gen Intel® Xeon® Scalable Processors

Significant AI Performance & Efficiency Across Broad Workloads (BF16)

5th Gen Intel Xeon Scalable Processor (8592+) Performance & Efficiency vs 4th Gen Intel Xeon (8480+)
Normalized to 4th Gen Intel Xeon-SP
Higher is better

For full configuration details, see Footnote 8

For compute-intensive AI workloads, utilizing the latest hardware and replacing aging servers can also significantly reduce energy and costs. Just 3 servers running 5th Gen Intel Xeon Scalable processors can do the work of 50 1st Gen Intel Xeon processor-based servers on a natural language processing (Bert-Large) workload. This efficiency not only saves rack space but can save up to 1,697 MWh of fleet energy – reducing CO2 emissions by more than 719 metric tons—and offer up to $541,000 in TCO savings over 4 years.[9]

Intelligent Power Management

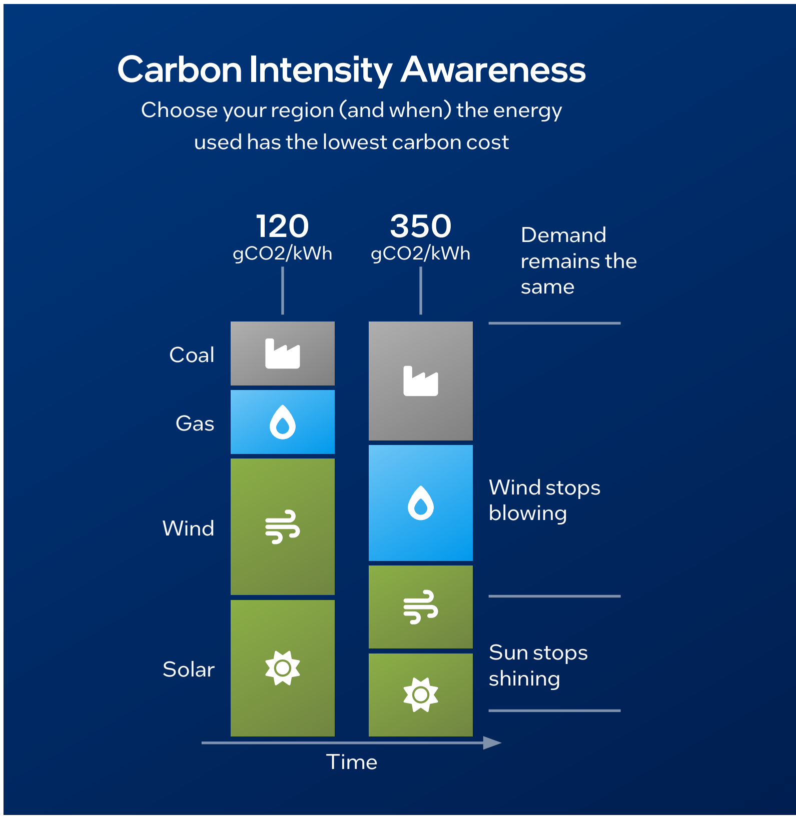
Another way to make AI more sustainable can also apply to other workloads: power management. While the latest Intel Xeon processors offer BIOS-level settings like Optimized Power Mode and Active Idle to offer power savings at low utilization levels, taking power management to the next level by pairing AI learnings with Intel tools can both optimize the energy efficiency of hardware and increase server utilization. Gartner inquiries show that data center infrastructure utilization is often far less than 50 percent, if not as low as 20 percent.[10] Yet, these data centers consume resources 24/7. With predictive insights and intelligent controls, target server utilization and rack density can be increased with minimal performance impact, thereby reducing data center footprint—both physical and environmental.

Intel® Granulate™ utilizes AI to help automatically monitor and improve server performance across your on-prem, hybrid and/or cloud infrastructure. Granulate can autonomously improve your application's performance by adapting OS resource management to your individual workloads. By learning usage patterns and optimizing runtime, Granulate can improve compute performance by up to 60 percent and reduce costs by up to 30 percent without ever having to change application code.[11] Plus, Granulate also offers a "CO2 Savings Meter" allowing you to easily measure the impact of your workload optimization on your IT carbon footprint alongside cost and resource reductions.

Telemetry capabilities built into the latest Xeon® processors can help generate real-time insights into system performance: offering feedback on power efficiency, thermals, resource utilization, and general health. By integrating this telemetry with intelligent data center infrastructure management tools, like the server management tools from our leading OEM partners, you can automatically orchestrate adjustments to optimize energy use and detect anomalies, proactively identifying problems before they arise.

Intel's tools in Kubernetes also allow you to automate management tasks in containerized environments, increasing efficiency and reducing energy usage. Paired with machine learning to predict peak compute times and fine-tune power use in Kubernetes clusters, Intel® Power Manager in Kubernetes can spin up nodes in advance to help ensure rapid response times while also reducing idle energy use and latency. And at off-peak times, Intel Power Manager can allow you to easily move nodes to a power saving profile, further conserving energy. Or use AI to predict factors like usage trends and easily act on insights with Intel's Telemetry Aware Scheduling (TAS) in Kubernetes, selectively increasing or decreasing lower priority workloads accordingly and increase server utilization.

For those workloads that are not time sensitive, intelligently managing the timing and location of a workload execution can have significant impacts on overall carbon intensity, and when optimized can help reduce emissions. The chart below shows the impact that a carbon-aware computing approach can have on the overall footprint of a computing workload. While compute requirements and power consumption remain the same, the variable of the electricity grid's energy mix can vary greatly. When greater fossil fuels are used to generate electricity, the associated carbon emissions from energy use can also vary dramatically, almost tripling the carbon cost (CO2 equivalent emissions) in this example.



With carbon aware software, like Carbon Aware SDK available from the Green Software Foundation, it is possible to control when workloads are run, optimizing for times when more renewable energy is available or a location/region where the energy is cleaner. And by harnessing AI insights, you can better predict renewable energy availability to dynamically decrease your data center's carbon footprint.

AI-enhanced liquid cooling

With these increasingly complex, compute-intensive AI workloads, greater attention to data center cooling strategy is needed. According to Gartner, an average of almost 40 percent of data center energy consumption is used in cooling today,[12] so effective cooling is fundamental to achieving greater energy efficiency. While AI-assisted automatic cooling can increase the efficacy of air-cooling methods, the most effective way to optimize cooling-related power usage today is liquid cooling technology: whether cold plate or liquid immersion.

One Intel® partner, Hypertec, shown customers save up to 95 percent on data center cooling OPEX, while also prolonging hardware lifespan 30 percent with their immersion cooling solution.[13] At the same time, liquid cooling can eliminate cooling-related water consumption, creating a closed loop system preventing any water evaporation.[14] Intel has also been collaborating with Vertiv to offer a liquid cooling solution compatible with both the upcoming Gaudi 3 AI accelerator and the latest Xeon processor-based systems. This solution has designed to remove up to 160kW of heat, even using relatively 'warm' facility water from 17°C up to 45°C (62.6°F to 113°F), thereby lowering overall system power requirements. With this system, customers have flexibility to implement heat reuse, warm water cooling, and/or free air cooling, all of which can help support reductions in power usage effectiveness (PUE) and water usage effectiveness (WUE), as well as total cost of ownership (TCO).

More effective cooling can also increase density potential of your racks up to 10x,[15] meaning more liquid cooling can quite literally offer more room for AI.

Take AI everywhere more sustainably with right-sized solutions

Today, you can capitalize on the opportunities of AI modernization while helping to ensure responsible energy- and resource-use aligned with your net-zero carbon goals. With an intentional approach to project design and supported by Intel technologies, you can execute AI initiatives – at the edge, in your data center, or on the cloud – more sustainably and optimize your IT operations. AI-enabled insights can also unveil new opportunities for other improvements across your business, including reducing resource inefficiencies, optimizing operations, and opening new, tech positive possibilities to support your sustainable transformation. While there will be a place for GPUs and discrete accelerators in some AI tasks, Intel Xeon CPUs offer an effective and flexible platform for right-sizing your AI and building more energy efficient operations.

With Intel as your trusted partner, you can unlock the potential of AI everywhere more sustainably and drive meaningful business outcomes.

While you may achieve better performance/watt for dedicated AI workloads with discrete accelerators, Intel Xeon CPUs offer an efficient and flexible platform for right-sizing your AI and building more energy efficient operations as you scale your AI capabilities your business.

With Intel as your trusted partner, you can unlock the potential of AI everywhere more sustainably and drive meaningful business outcomes.

[1] Accelerating Climate Action with AI Report (gstatic.com)

[2] We're getting a better idea of AI's true carbon footprint | MIT Technology Review

[3] Large Language Models: A New Moore's Law? (huggingface.co)

[4] The environmental footprint of data centers in the United States (vt.edu)

[5] Understanding deep learning requires rethinking generalization (arxiv.org)

[6] AI without GPUs: Accessing Sapphire Rapids AMX instructions on vSphere | VMware

[7] Fine-tuning BioGPT model to SOTA accuracy of 79.4% on Nvidia A100 vs. Intel® Xeon® Platinum 8480+ processor. Results may vary.

Test by Intel on 09/04/2023

8480+: 1-4 nodes, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024 GB memory(16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS SE5C7411.86B.9525.D21.2303221016, microcode 0x2b0001b0, OS Rocky Linux 8.8 (Green Obsidian), 1x Ethernet Controller X710 for 10GBASE-T, 9x Ethernet Controller E810-C for QSFP, 2x 1.5T INTEL SSDPF21Q016TB, 2x 931.5G CT1000P2SSD8, Multiple nodes connected with 100Gbps OmniPath. PyTorch 2.0.1, IPEX 2.0.1, Transformers 4.31.0, Datasets 2.14.0, Peft 0.4.0, Evaluate 0.4.0.

Nvidia A100: Test by Intel on 09/04/2023, 1-node, GPU NVIDIA A100-PCIE-40GB, 2x Intel(R) Xeon(R) Platinum 8280, 56 cores, HT On, Turbo On, 768GB (12x64GB DDR4 3200 MT/s [2934 MT/s]) , BIOS SE5C620.86B.02.01.0013.121520200651, microcode 0x5003302, OS Ubuntu 20.04.2 LTS, kernel 5.4.0-124-generic, compiler gcc version 9.4.0 (Ubuntu 9.4.0-1ubuntu1~20.04.1), BioGPT(1.5B) fine-tuning, Pytorch 2.0.1+cu118, Transformers 4.31.0, Datasets 2.14.0, Peft 0.4.0, Evaluate 0.4.0.

[8] Configurations:

Performance varies by use, configuration and other factors. See backup for configuration details. Results may vary.

PT Training :

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

BS=x, 1 instance/ numa node [Using only one socket]. SSD Resnet34: COCO 2017, BERT-large: Wikipedia 2020/01/01 (seq len =512), DLRM: Criteo Terabyte Dataset, RNNT: LibriSpeech, Mask RCNN: COCO 2017

BERT-Large Inference

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x

1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, Test by Intel as of 10/10/23.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

Software configuration: BERT-large, Intel Model Zoo:https://github.com/IntelAI/models, gcc=12.3, OneDNN3.2, Python 4.1, PyTorch 2.0, IPEX 2.0, physical cores only.

DLRM Inference :

8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x

Ethernet interface, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, Test by COMPANY as of 10/10/23.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 2.0, microcode 0x2b0004d0, 1x Ethernet interface, 2x Ethernet Controller X710 for

10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, Test by COMPANY as of 10/25/23.

Software configuration: DLRM, Intel Model Zoo:DLBoost package - WW35 release, gcc=12.3, OneDNN3.2, Python 3.8, PyTorch2.1, IPEX2.1, physical cores only.

SSD-ResNet34 Inference

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

SSD-ResNet34, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; SSD-Resnet34 coco 2017 (1200 x1200)

Performance varies by use, configuration and other factors. See backup for configuration details. Results may vary.

ResNeXT101_32x16d Inference:

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

ResNeXT101_32x16d, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; Resnext101: ImageNet

RNN-T Inference :

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

RNN-T, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; RNNT: LibriSpeech

DistilBERT Inference

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1.SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

DistilBERT, BS=1: 4cores/instance, BS=x: 1 instance/ numa node

MaskRCNN Inference:

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1.SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1. SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

MaskRCNN, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; Mask RCNN: COCO 2017

Performance varies by use, configuration and other factors. See backup for configuration details. Results may vary.

ResNeXT101_32x16d Inference:

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1. SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1. SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

ResNeXT101_32x16d, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; Resnext101: ImageNet

RNN-T Inference :

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1. SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1. SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

RNN-T, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; RNNT: LibriSpeech

DistilBERT Inference

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1. SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1. SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

DistilBERT, BS=1: 4cores/instance, BS=x: 1 instance/ numa node

MaskRCNN Inference:

8592+: 1-node, 2x Intel(R) Xeon(R) Platinum 8592+, 64 cores, HT On, Turbo On, NUMA 4, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS EGSDCRB1. SYS.0105.D74.2308261933, microcode 0x21000161, 1x Ethernet

Controller I225-LM, 1x 3.6T INTEL SSDPE2KX040T8, 1x 931.5G INTEL, CentOS Stream 9, 6.2.0-emr.bkc.6.2.3.6.31.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu, oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1, Test by INTEL as of 09/07/2023.

8480+: 1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDCRB1. SYS.0102.D37.2305081420, microcode 0x2b0004b1, 1x Ethernet

Controller I225-LM, 1x 7.2G Cruzer Blade, 2x 3.7T INTEL SSDPE2KX040T8, 1x 894.3G INTEL SSDSC2KG96, CentOS Stream 8, 5.15.0-spr.bkc.pc.16.4.24.x86_64. PyTorch: torch:2.1.0.dev20230825+cpu oneDNN:v3.2.1 IPEX:2.1.0+git31b5ee1.

Test by INTEL as of 09/05/2023.

MaskRCNN, BS=1: 4cores/instance, BS=x: 1 instance/ numa node; Mask RCNN: COCO 2017

[9] See [T7] at intel.com/processorclaims: 5th Gen Intel Xeon Scalable Processors. Results may vary.

[10] Gartner, How Can Sustainability Drive Data Center Infrastructure Cost Optimization?, November 2022

[11] https://www.intel.com/content/www/us/en/architecture-and-technology/granulate/overview.html

[12] Gartner, How Can Sustainability Drive Data Center Infrastructure Cost Optimization?, November 2022

[13] Hypertec Immersion Cooling

[14] Hypertec Immersion Cooling

[15] Hypertec Immersion Cooling

Intel is committed to the continued development of more sustainable products, processes, and supply chain as we strive to prioritize greenhouse gas reduction and improve our global environmental impact. Where applicable, environmental attributes of a product family or specific SKU will be stated with specificity. Refer to the 2022 Corporate Responsibility Report (p. 64) for further information.

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.