Bringing AI Everywhere

5th Gen Intel® Xeon® Scalable Processors

Walter Riviera – EMEA AI Tech LEAD

intel®

Notices and 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.

Intel® Xeon® - The Processor Designed for AI

Architected for general AI and Large Language Models (LLMs)

Up to 64 cores per CPU

Intel® Advanced Matrix Extensions (Intel® AMX)

Better AMX Frequencies

Increased Memory BW

Up to 5600 MT/s

CXL Memory BW expansion

Large Last Level Cache (LLC)

Up to 3x

Compared to 4th Gen Intel® Xeon® processors

Intel® AI Software

Optimizations up-streamed
300+ DL Models
50+ optimized ML and Graph Models
Intel® AI Developer Tools

PyTorch containers

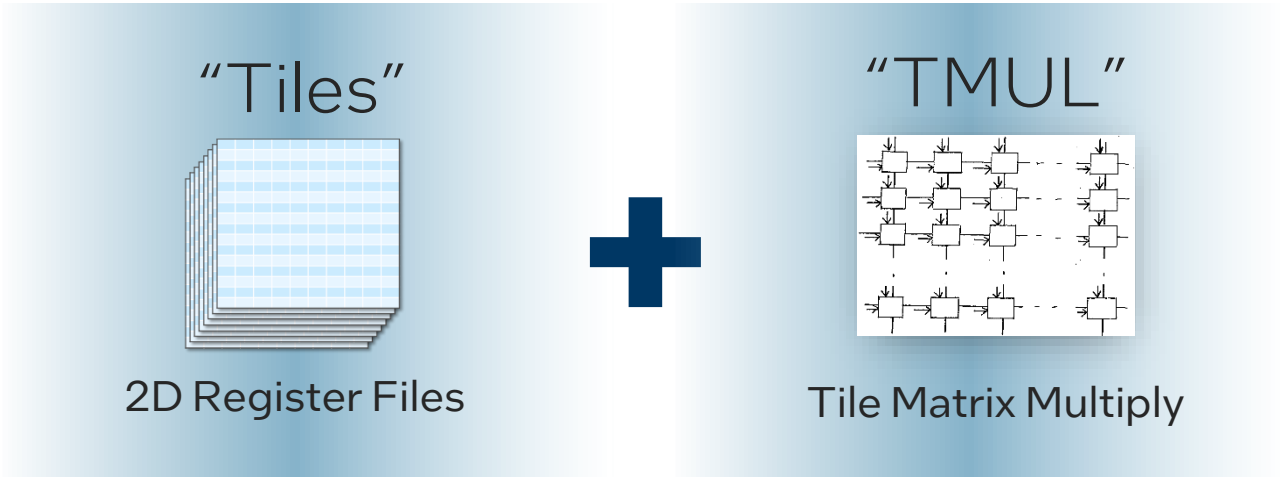
https://hub.docker.com/r/intel/intel-optimized-pytorch

TensorFlow containers

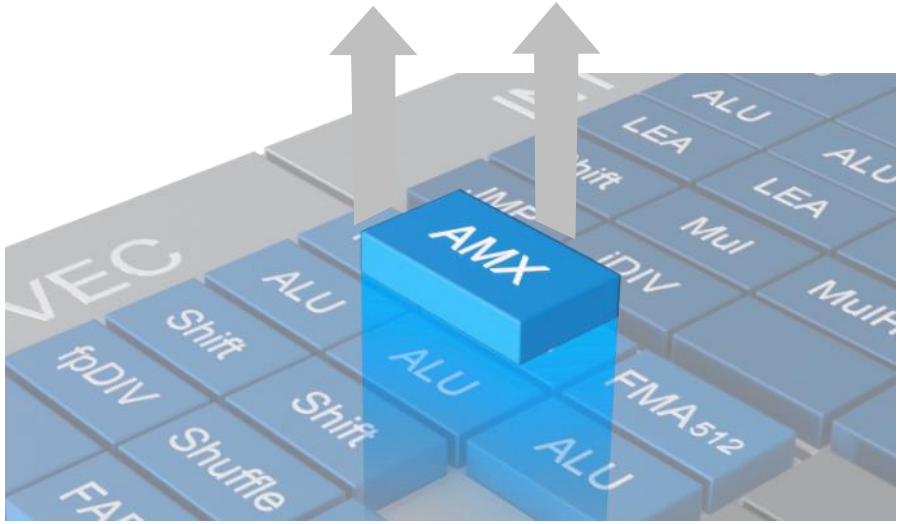
https://hub.docker.com/r/intel/intel-optimized-tensorflow

Intel® Advanced Matrix Extensions (Intel® AMX)

DL Accelerator Performance Built Into Every Core

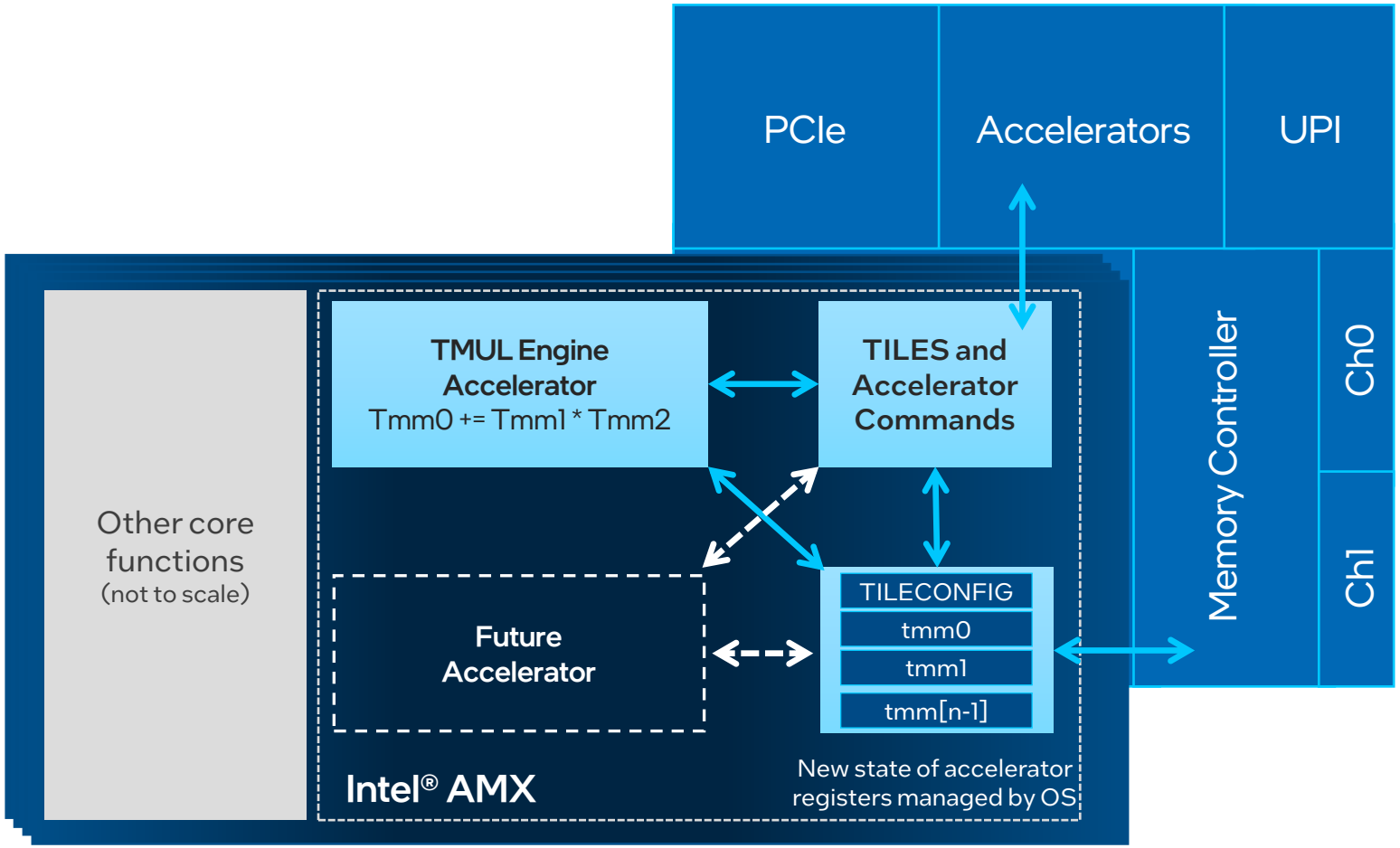


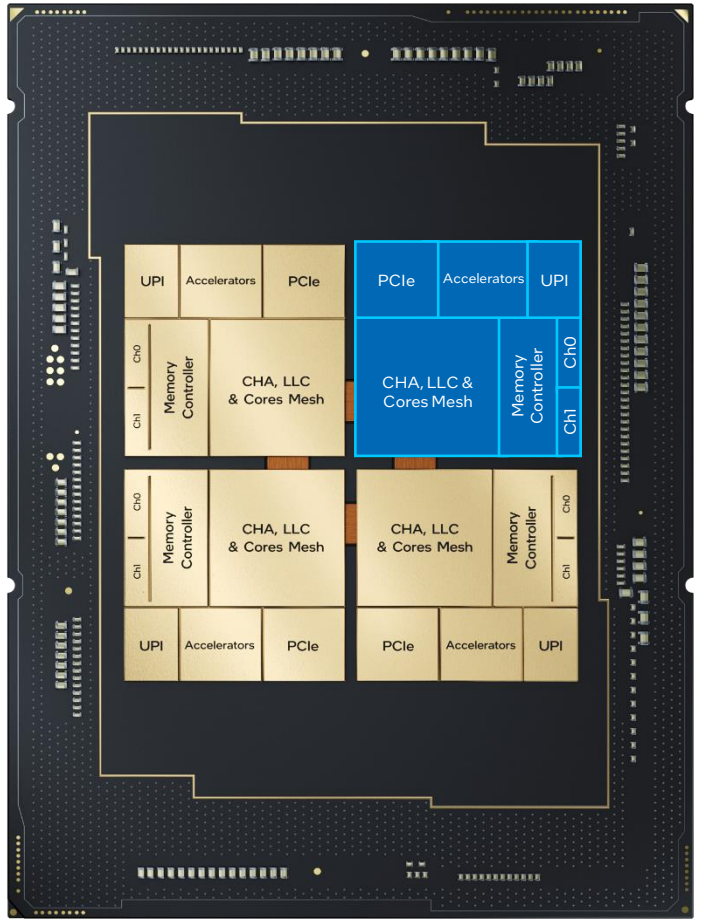
Store bigger chunks of **data**



Instructions that compute larger matrices in a single operation

Intel® AMX Works Directly on the Data





TILEs/TMUL coding

Configuration:

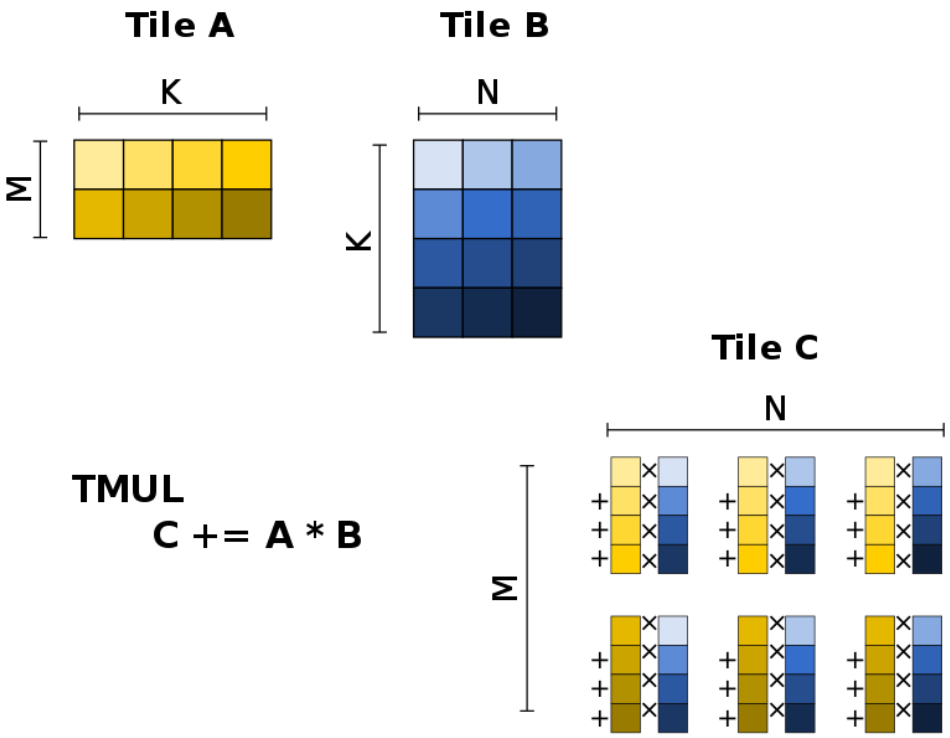
- `LDTILECFG` - Load tile configuration, loads the tile configuration from the 64-byte memory location specified.
- `STTILECFG` - Store tile configuration, stores the tile configuration in the 64-byte memory location specified.

Data:

- `TILELOADD` / `TILELOADDT1` - Load tile
- `TILESTORED` - Store tile
- `TILERELEASE` - Release tile, returns TILECFG and TILEDATA to the INIT state
- `TILEZERO` - Zero tile, zeroes the destination tile

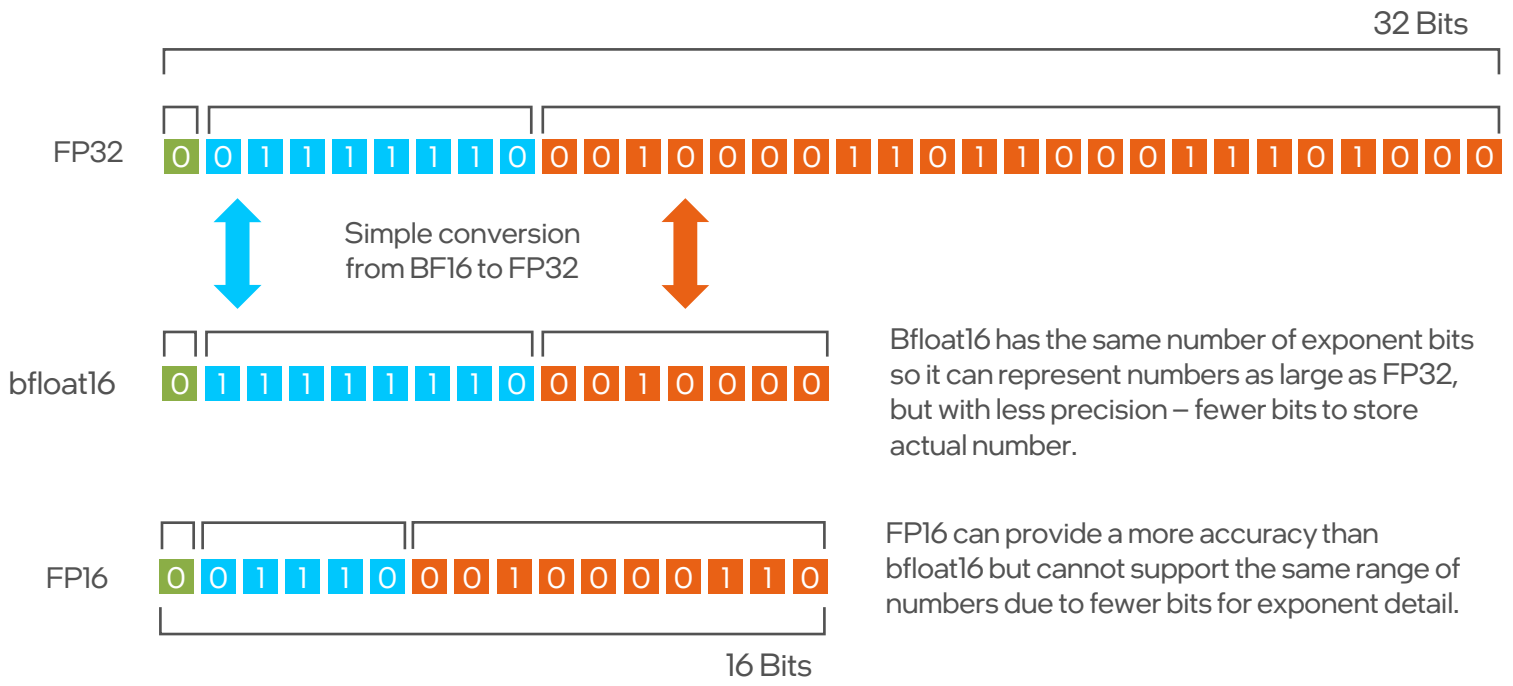
Operation:

- `TDPBF16PS` - Perform a dot-product of BF16 tiles and accumulate the result. Packed Single Accumulation.
- `TDPB[XX]D` - Perform a dot-product of Int8 tiles and accumulate the result. Dword Accumulation.



Intel Deep Learning Boost
Brain Floating-point Format with 16 Bits (bfloat16)

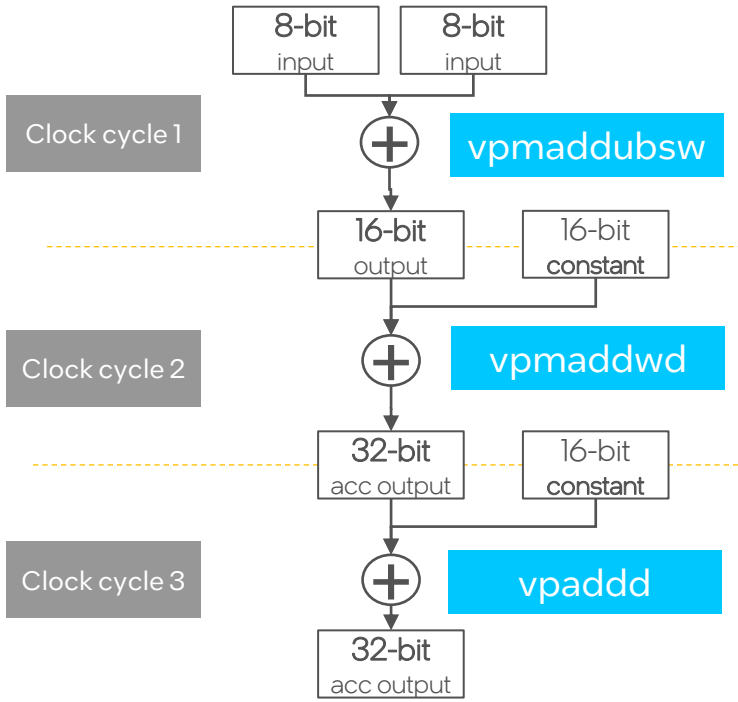
- Floating Point 32 (FP32) provides high precision based on the number of bits used to represent a number
- Many AI functions do not require the level of accuracy provided by FP32
- Bfloat16 supports the same range of numbers based on the same exponent field but with lower precision
- Conversion between bfloat16 and FP32 is simpler than FP16
- Twice the throughput per cycle can be achieved with bfloat16 when comparing FP32



One Processor for Scalar, Vector, and Matrix

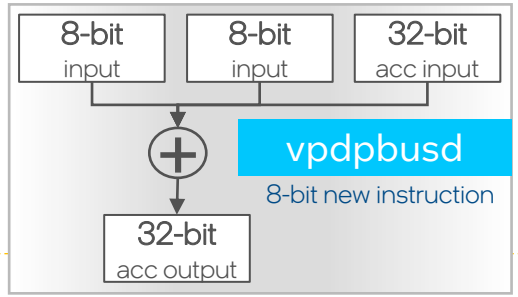
Intel® AVX-512

85 int8 ops/cycle/core
with 2 FMA



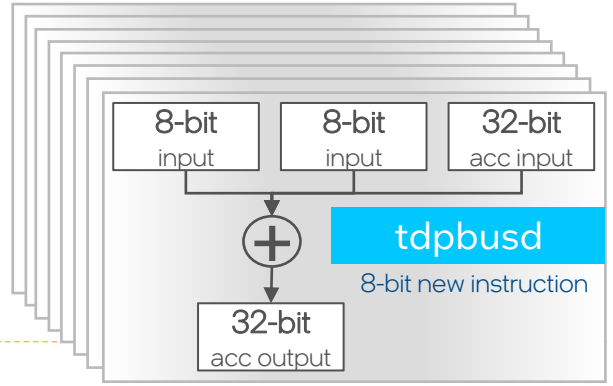
Intel® AVX-512 (VNNI)

256 int8 ops/cycle/core
with 2 FMAs



Intel® AMX

2048 INT8 ops/cycle/core
Multi-fold MACs in one instruction



Intel® AMX supports BF16 and INT8 data types

	3rd Gen Intel® Xeon® Scalable processor		4th Gen Intel® Xeon® Scalable processor
	Cooper Lake	Ice Lake	Sapphire Rapids
AVX512	FP64, FP32, bfloat16	FP64, FP32	FP64, FP32
VNNI	INT8	INT8	INT8
AMX			**bfloat16, INT8**

FP32 supported through AVX512 instructions

Intel® AMX supports BF16 and INT8

5th Gen Xeon TCO advantages over AMD

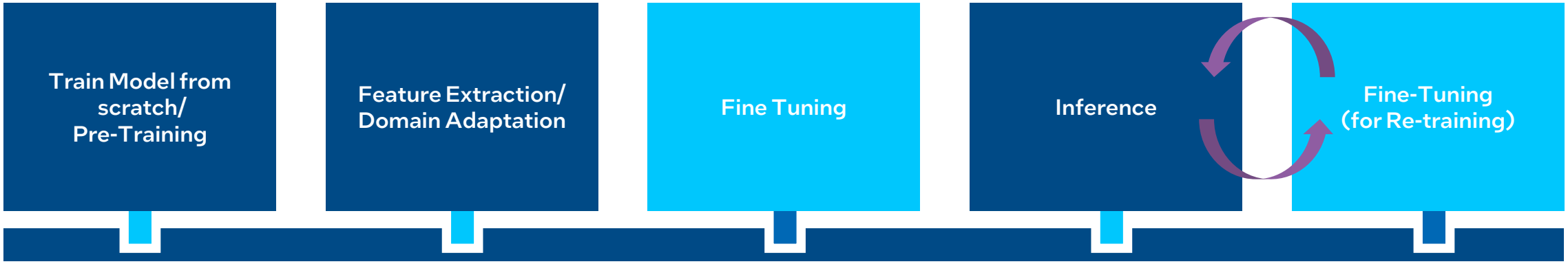
A comparison against 50 4th Gen AMD EPYC 9554 servers

	Web NGINX TLS	Data Services RocksDB	Data Services MySQL	HPC Monte Carlo	AI - NLP DistilBERT
5th Gen Xeon Servers	31 servers	31 servers	30 servers	28 servers	15 servers
Fleet Energy Saved*	489.7 MWh	1218.1 MWh	684.0 MWh	585.8 MWh	1496.5 MWh
Reduced CO2 Emissions*	207,611 kg	516,402 kg	289,967 kg	248,352 kg	634,428 kg
TCO Savings*	$444K	$471K	$509K	$561K	$1,300K
TCO Delta	21% savings	22% savings	24% savings	27% savings	62% savings

*Estimated over 4 years
See backup for workloads and configurations. Results may vary.

Generative AI

The New Generative AI Deep Learning Pipeline



Fine tuning is the new training

"Transfer Learning is quickly making training a thing of the past for 99% of the organizations"

—Julien Simon – Hugging Face Chief Evangelist

Intel Generative AI Products for the Enterprise

Developer Clouds

10bs -1T+ parameters

100s of customers

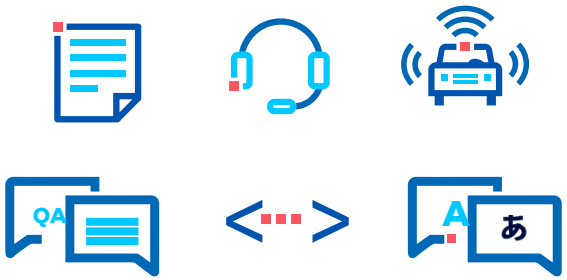
Large foundational model pre-training, batch inferencing and AIaaS

Multi-Cloud Enterprise IT: On-premise, public cloud, and the edge

<10b parameters

10,000s of customers

Fine tuning 10,000s of expert nimble models in minutes to hours on workstation and servers

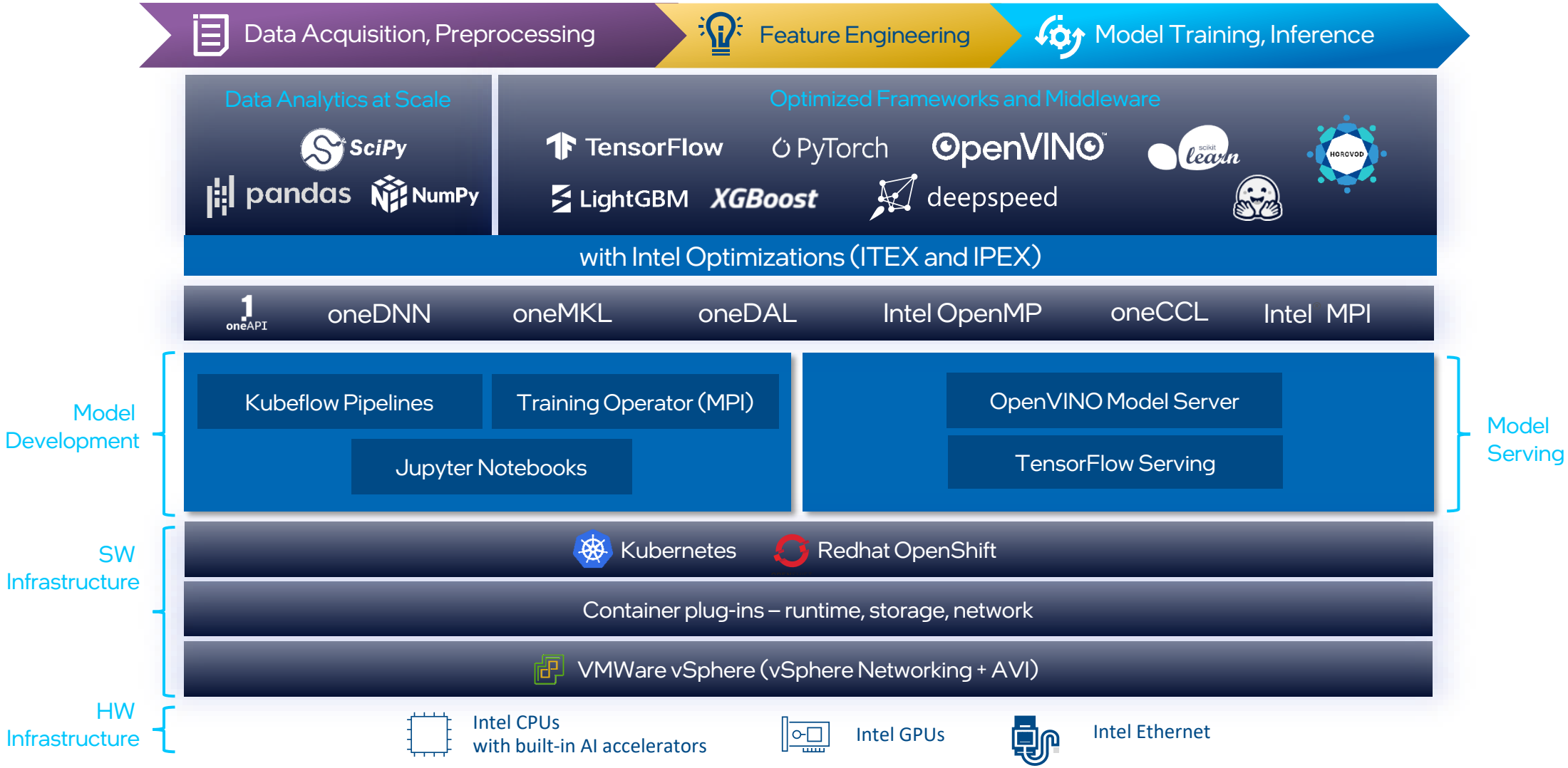


<20b parameters
Starting from Intel® Xeon®

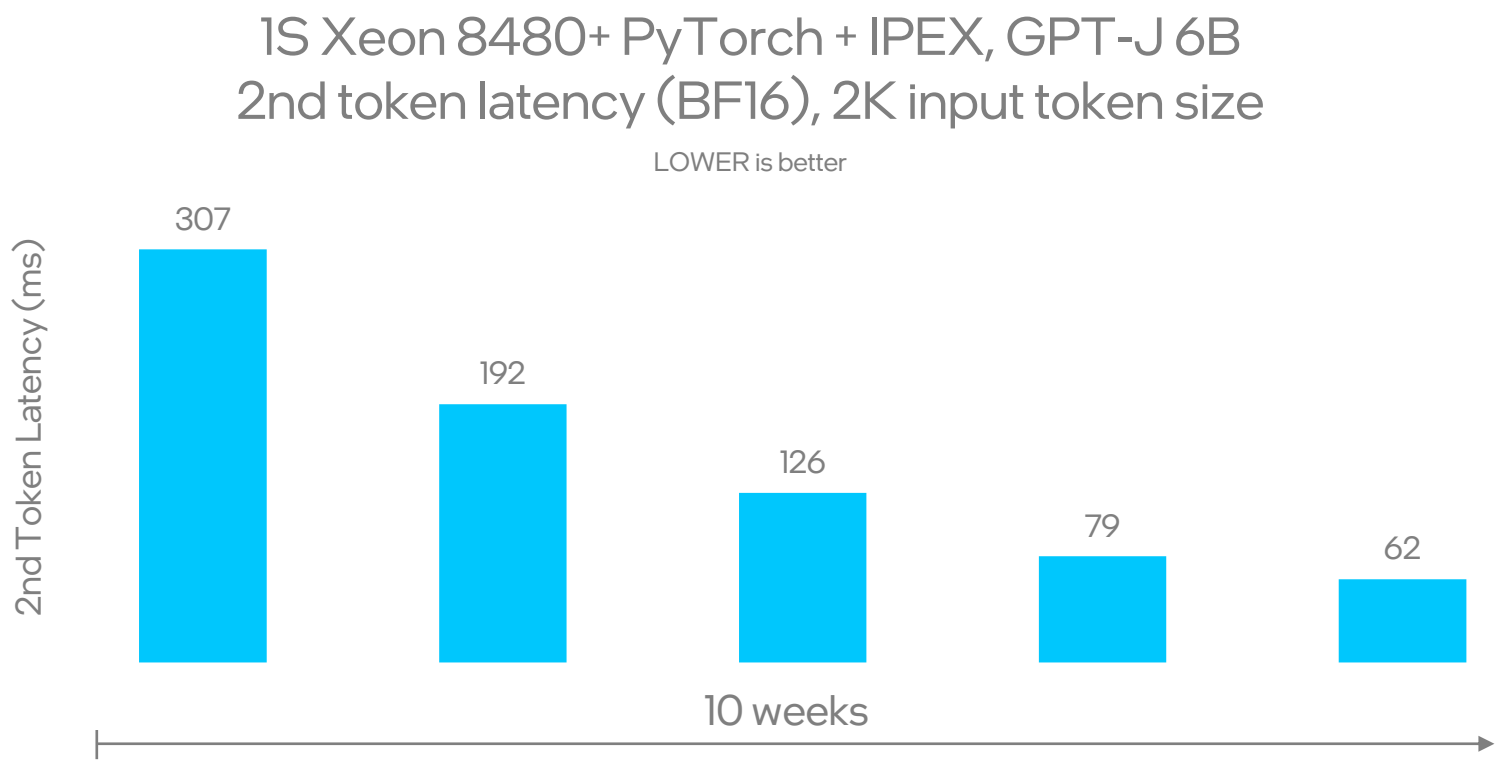
10,000s of customers

Inference deployment meeting customers latency SLA by Inferencing at <100ms token latency

Intel® AI Software is Enterprise Ready



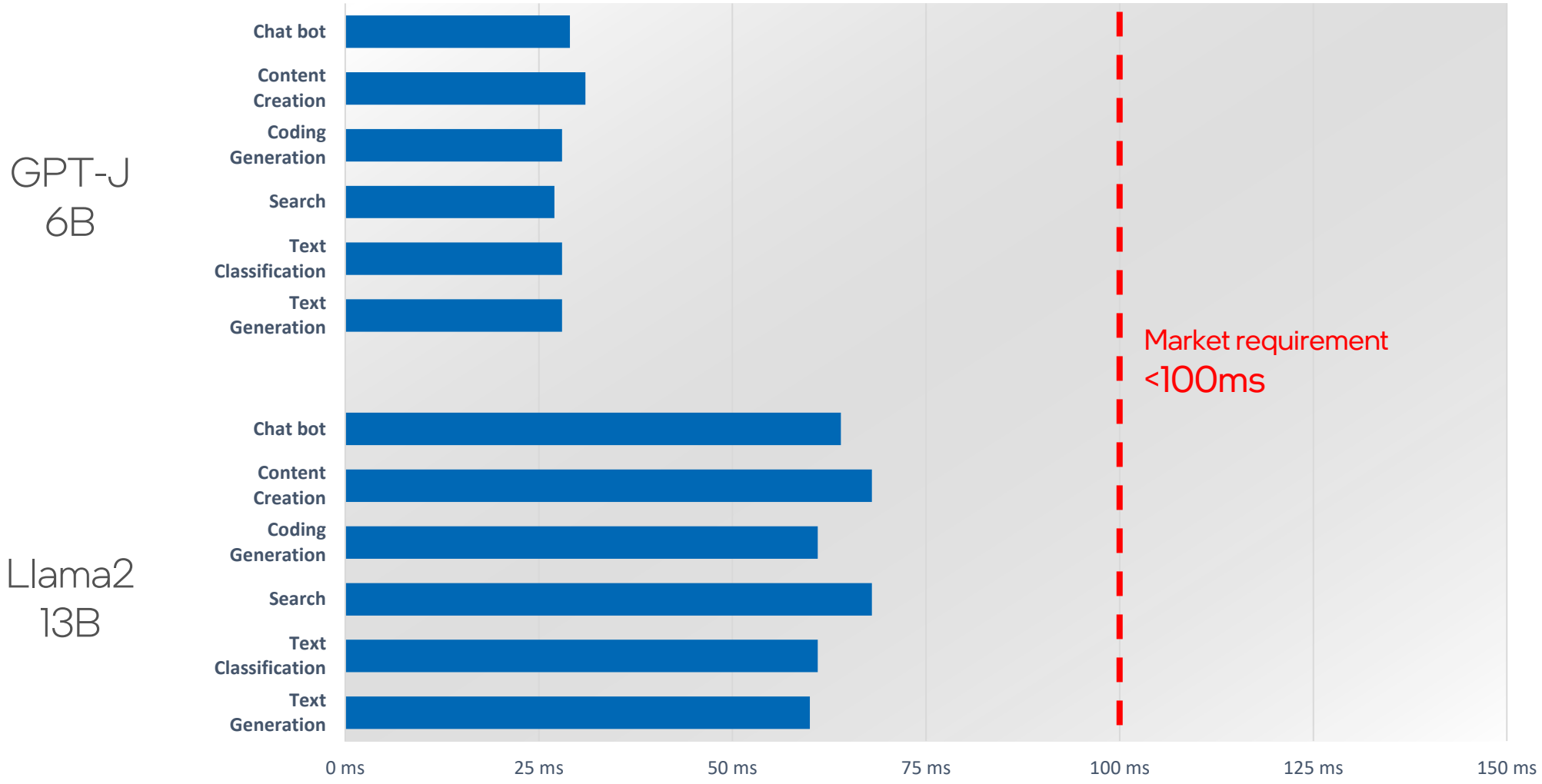
We're keeping pace with optimizing Large Language Models (LLMs)



~5X improvement on the same single-socket 4th Gen Intel® Xeon® Scalable processor

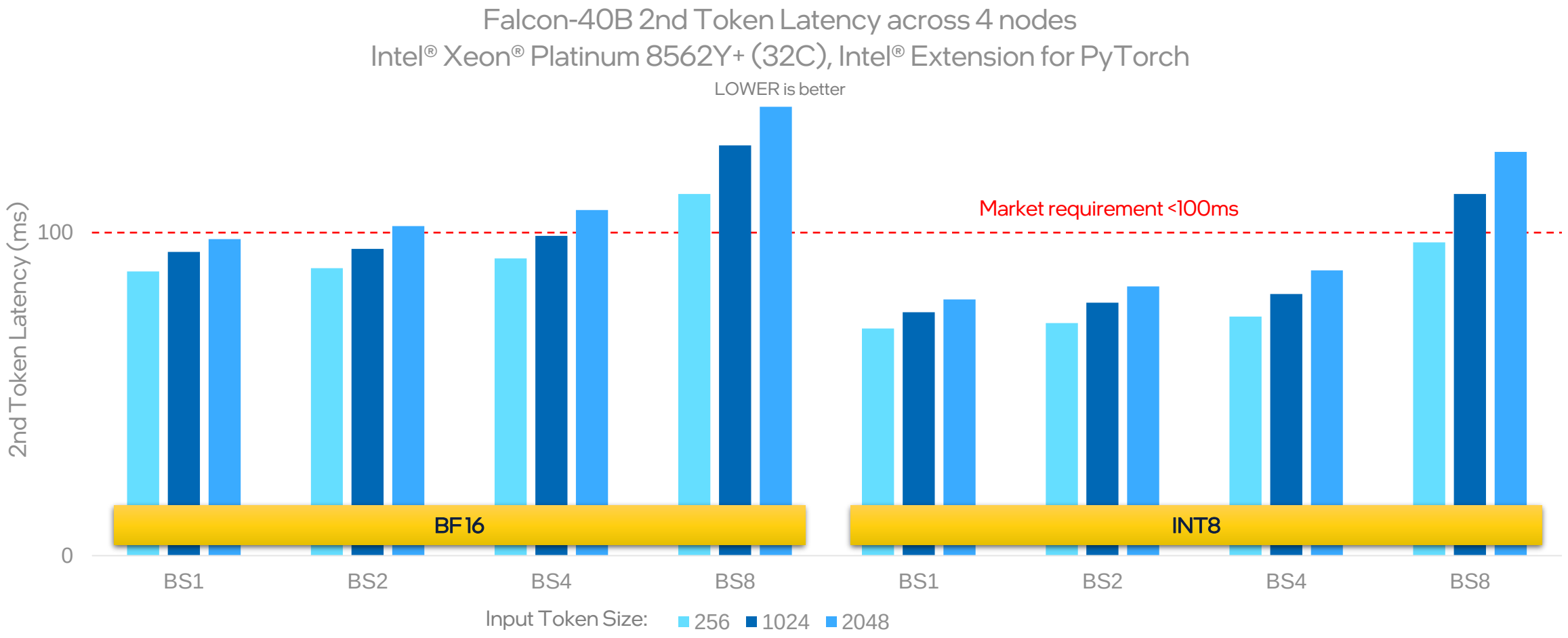
5th Gen Xeon bests market requirements on LLM latencies

Single node 2S 5th Gen Xeon 8592+ (64C) Large Language Model Next Token Latency



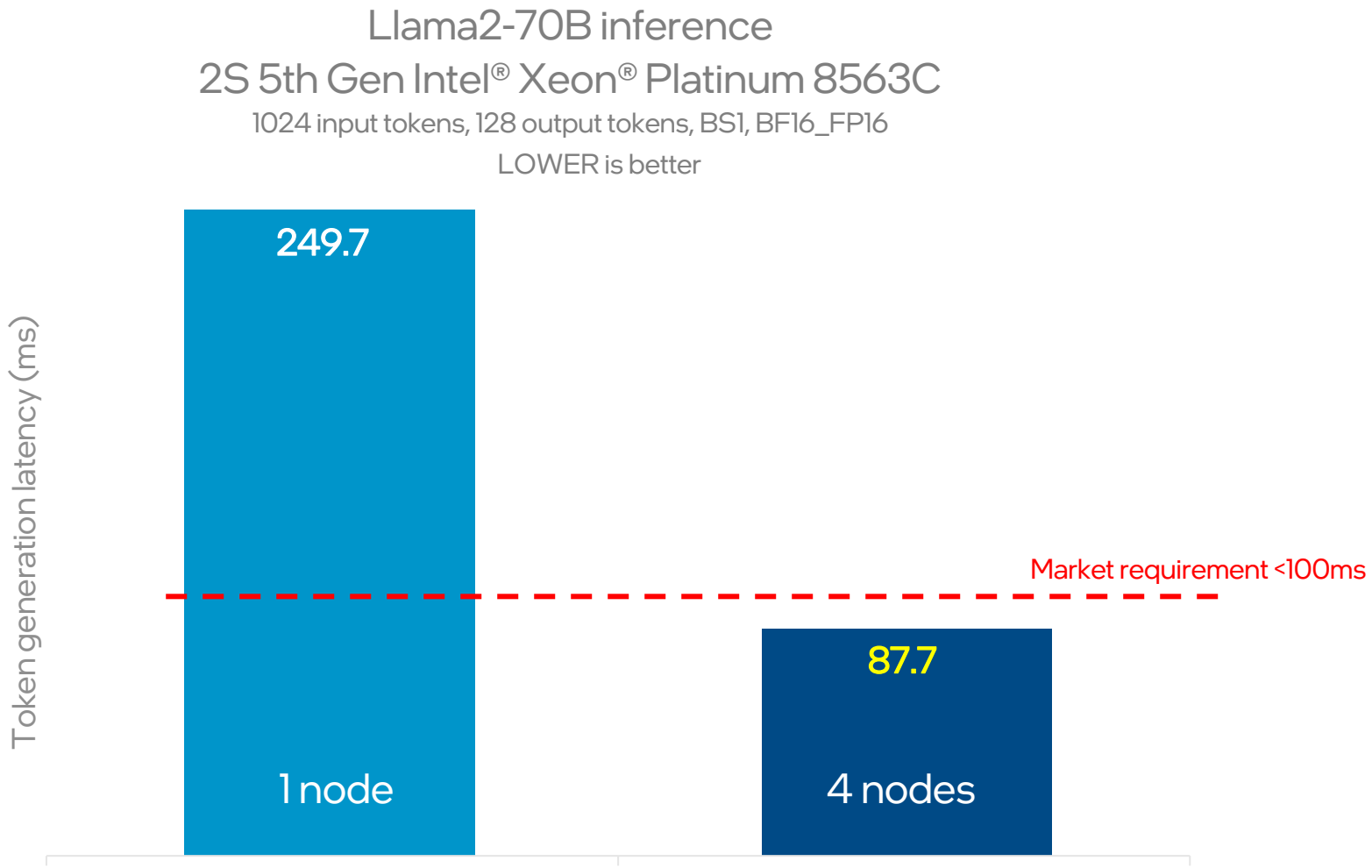
See backup configuration for workload and configurations. Results may vary.

Xeon LLM optimizations continue on larger models using multiple nodes

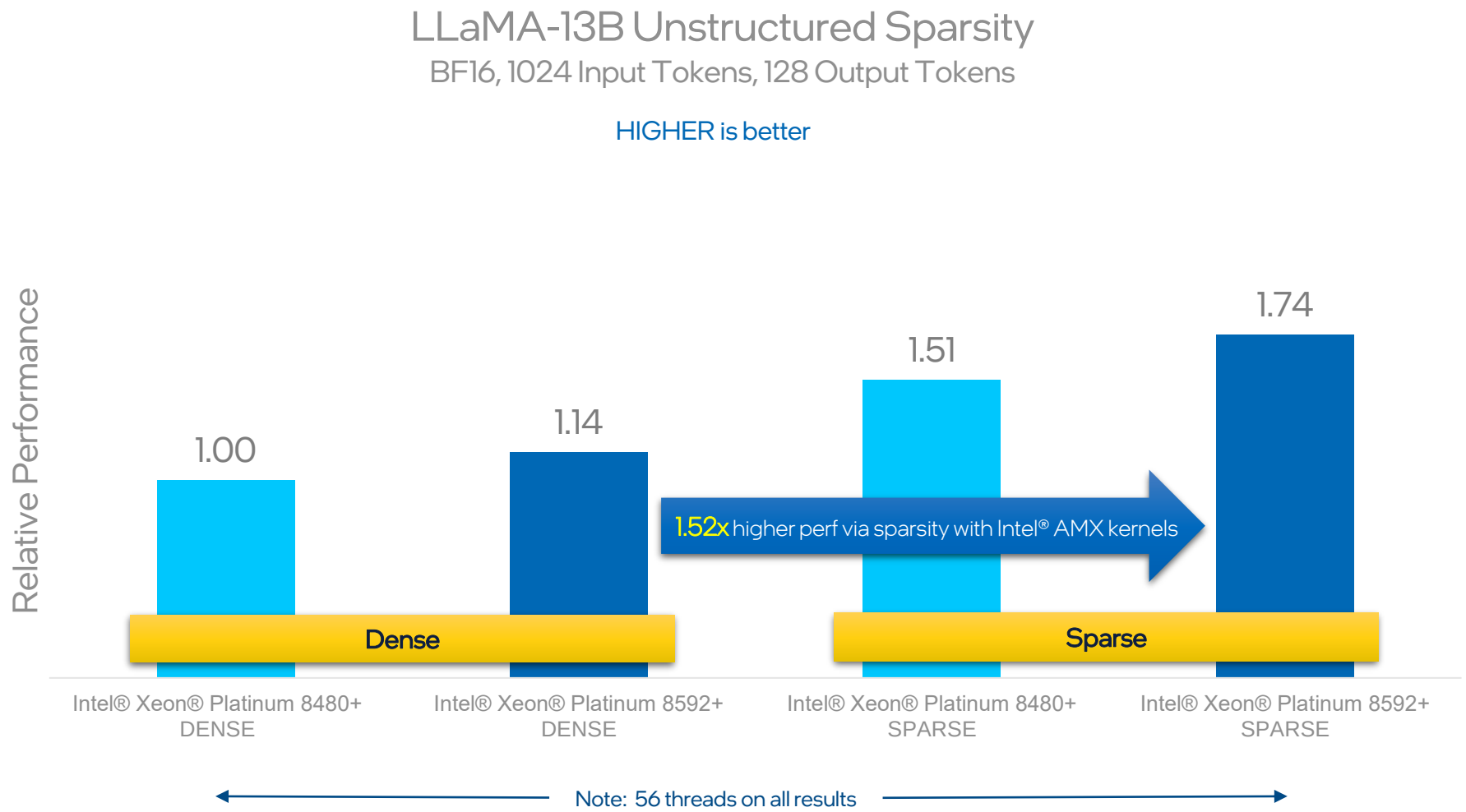


See backup for workload and configurations. Results may vary.

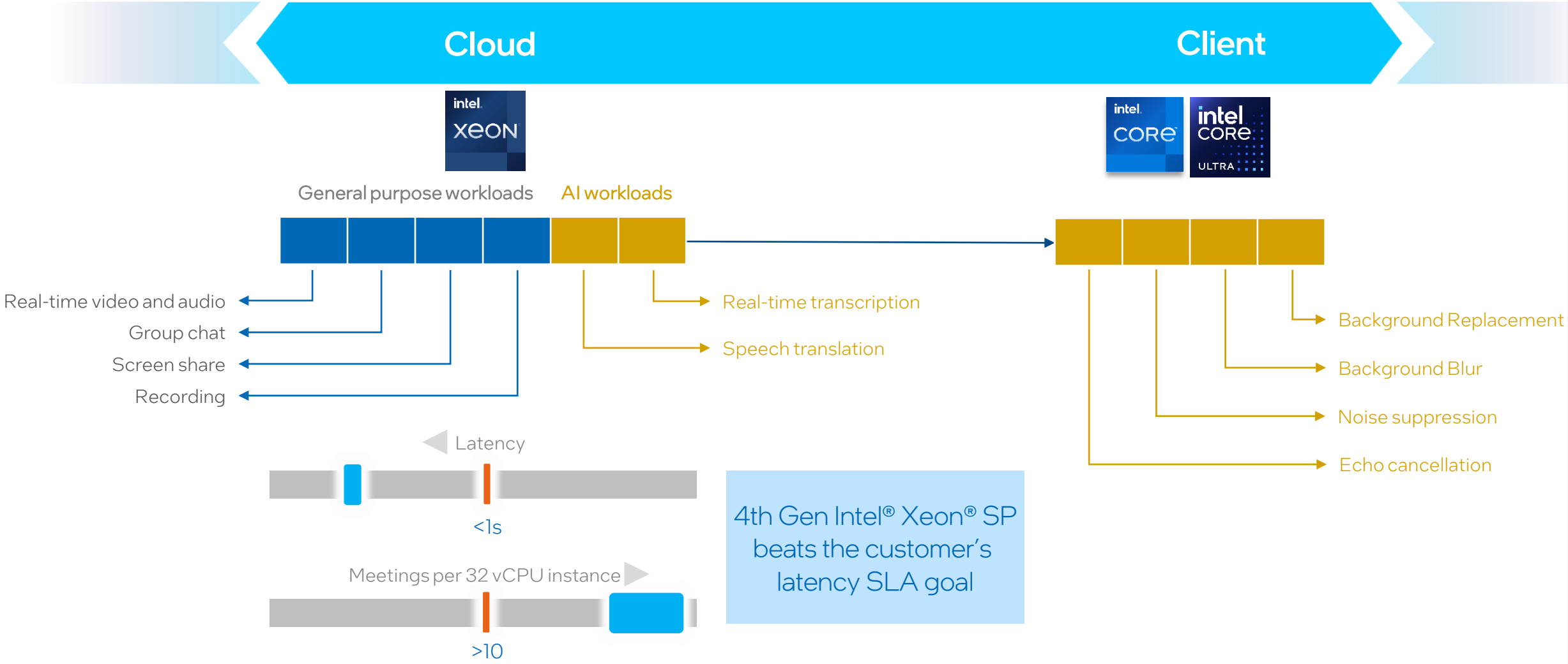
Llama2-70B multi-node inference on 5th Gen Intel® Xeon® Scalable processor



In the Intel labs: Unstructured sparsity-based acceleration of LLaMA-13B inferencing



Leading video collaboration customer using the full Intel AI Roadmap

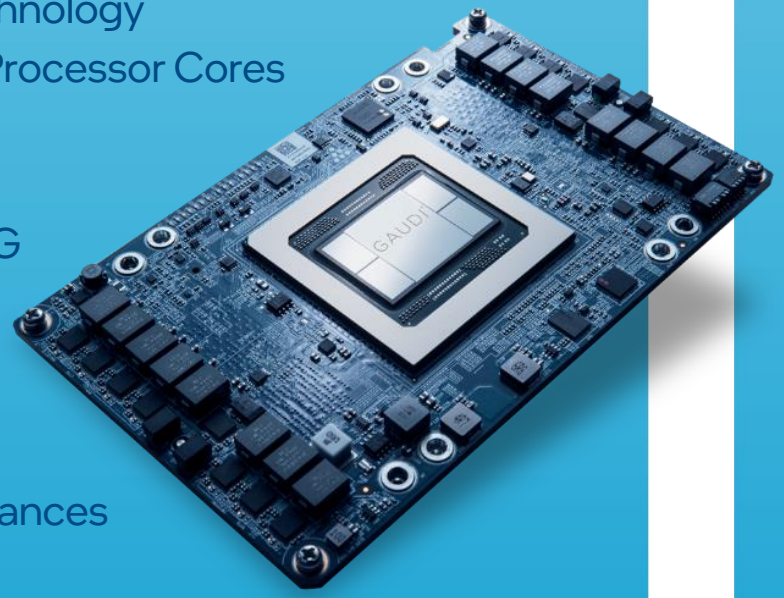


Habana Deep Learning Solutions

GAUDI®

High-performance, high-efficiency (price/performance)

- 16nm process technology
- 32 GB 8 Tensor Processor Cores
- on-board HBM2
- 24 SRAM
- 10 integrated 100G Ethernet ports



In the cloud:

- Amazon EC2 DL1 Instances

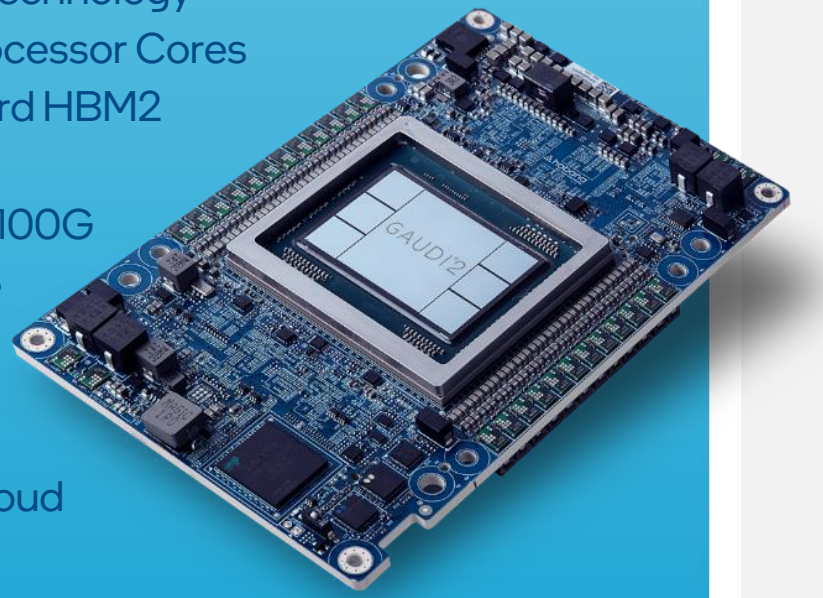
On-premises:

- Supermicro X12 Gaudi Server with 3rd Gen Xeon CPU

GAUDI®2

Higher performance, high-efficiency; optimized speed, memory, scalability for large scale models

- 7nm process technology
- 24 Tensor Processor Cores
- 96 GB on-board HBM2
- 48 SRAM
- 24 integrated 100G Ethernet ports



In the cloud:

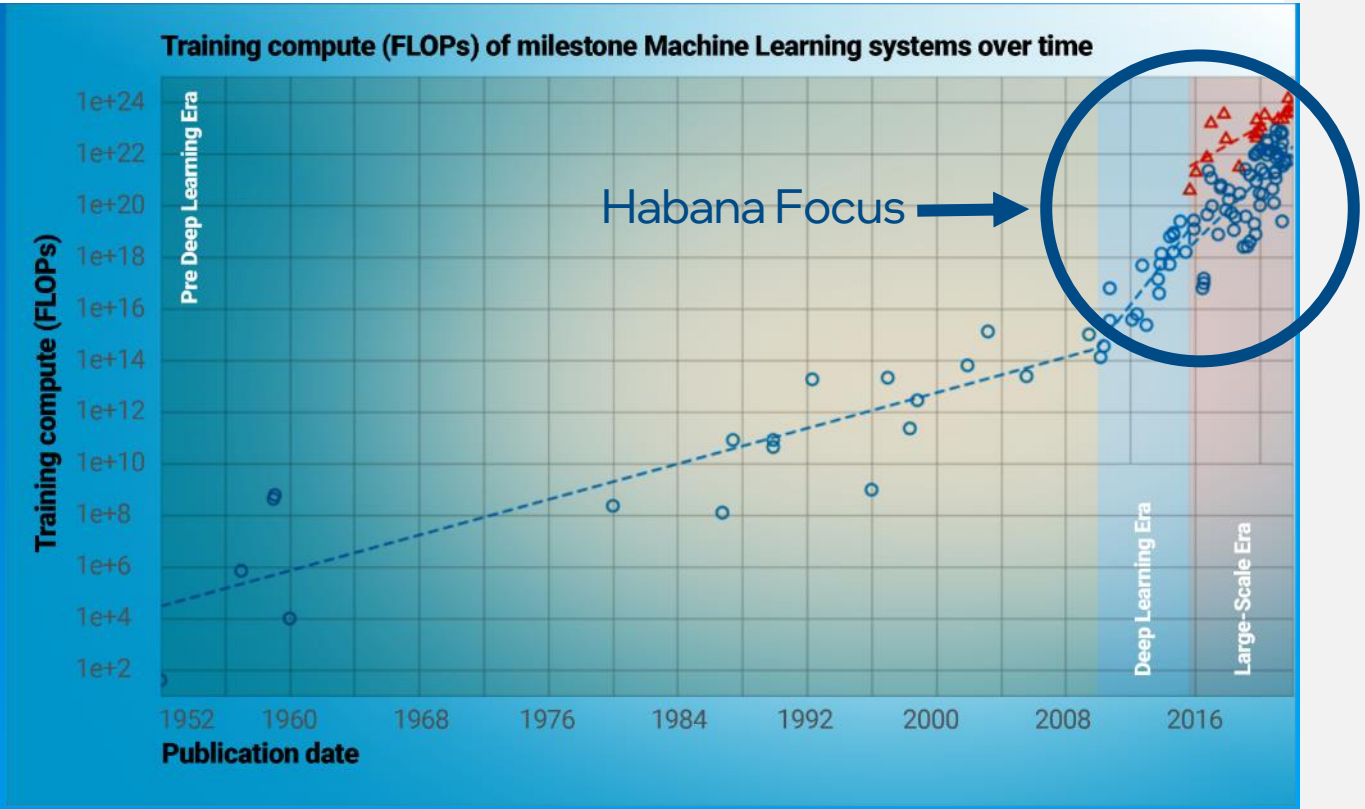
- Intel Developer Cloud

On-premises:

- Supermicro Gaudi2 Server with 3rd Gen Xeon CPU

Gaudi2: Keeping Pace with Foundation Models

- Training
  - GPT3 175B training on 384x Gaudi2
  - Stable diffusion training on 64x Gaudi2
- Inference
  - Stable Diffusion
  - BLOOM 176B & variety of LLMs



Gaudi2 is IDEALLY designed to address large-scale, complex models

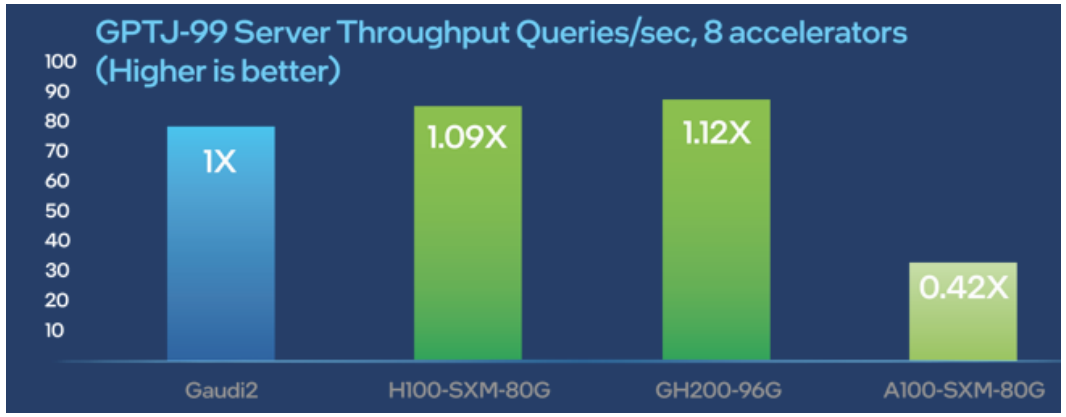
Study by Epoch, University of Aberdeen, Center for the Governance of AI, University of St. Andrews, MIT, Eberhard Karls Universitat Tubingen, Universidad Complutense

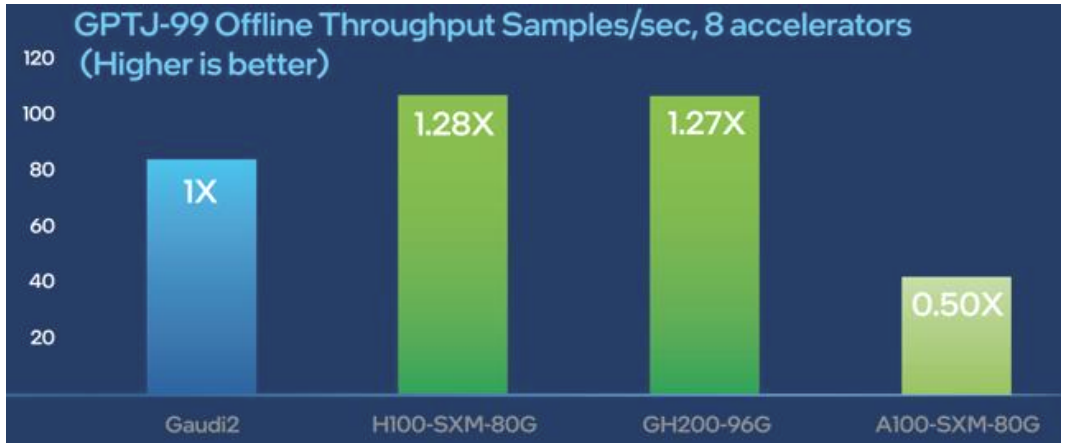
Near-Parity GPT-J Inference Performance vs. H100

Intel® Gaudi® 2 with FP8 achieved accuracy of

99.9%

- Intel Gaudi 2 throughput:
  -9% (server) and
  -28% (offline) vs H100
- Vs. A100: 2.4x (Server) and 2x (Offline)



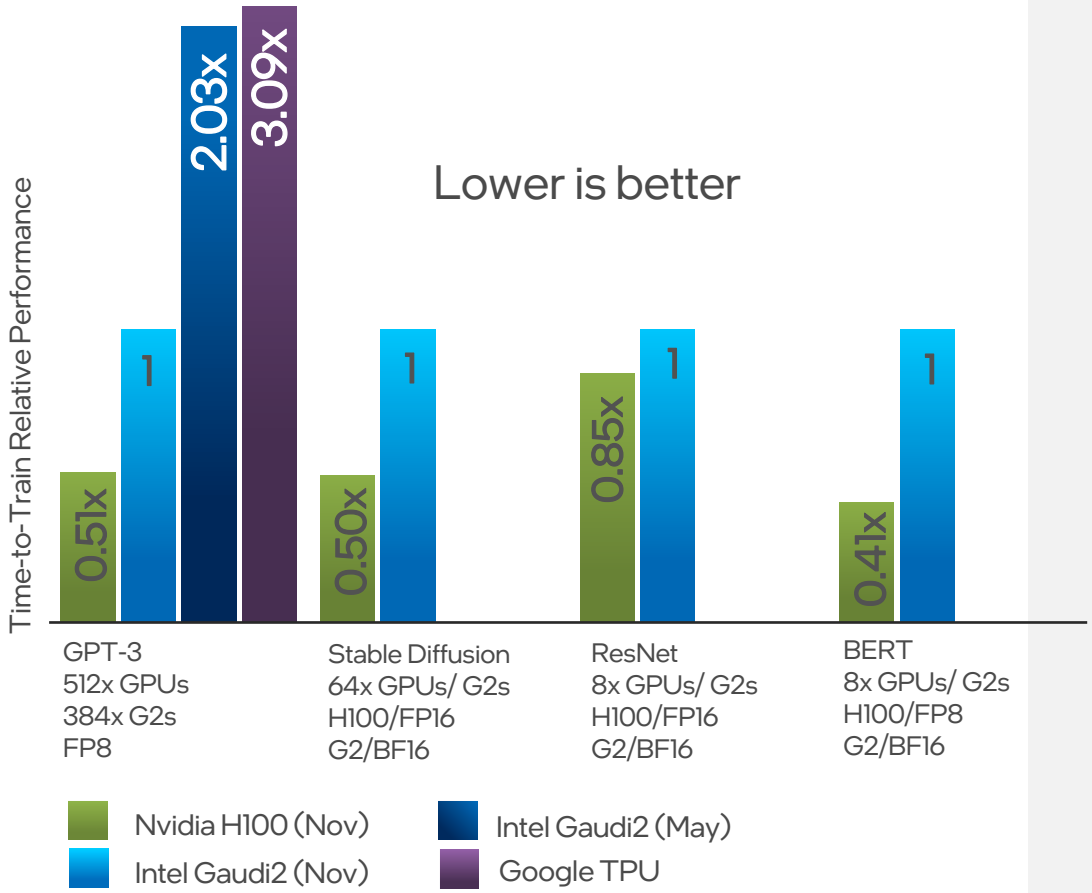


Performance source: MLPerf Inference 3.1 results https://mlcommons.org/benchmarks/inference-datacenter/

2x Performance = 2x Better Price-Performance

1 of only 2 merchant silicon submissions for GPT-3

- H100/FP8 outperformed Gaudi2/BF16 on BERT
- Intel® Gaudi® 2 ResNet result near H100 submission



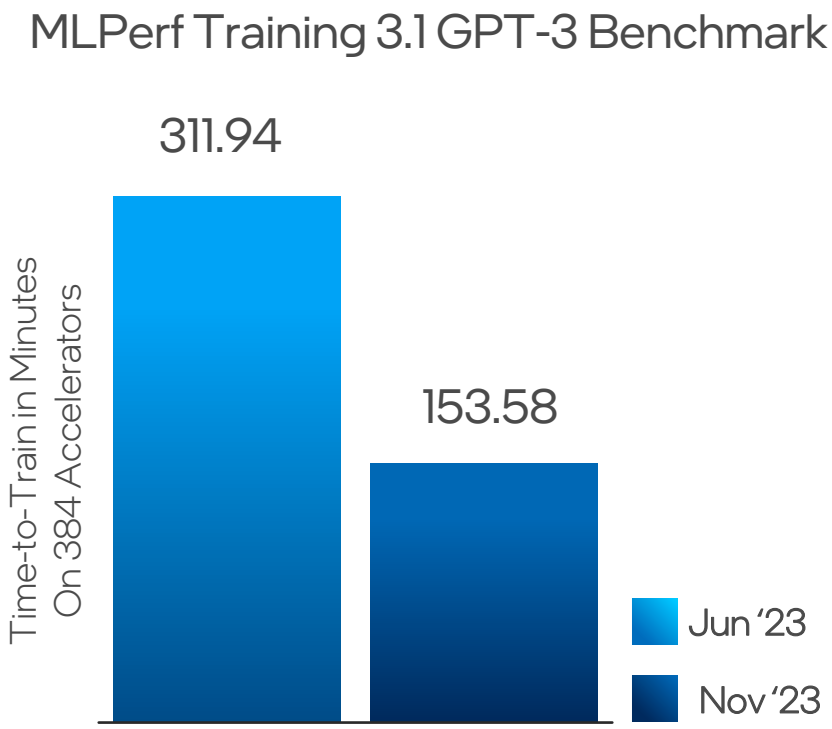
Performance source: MLPerf Training 3.1 results https://mlcommons.org/benchmarks/training/

Intel® Gaudi® 2 Accelerator Performance Doubled with FP8

We projected for customers +90% performance gain with FP8

Delivered more than 2x

Committed to credible, reliable performance projections and delivering on them



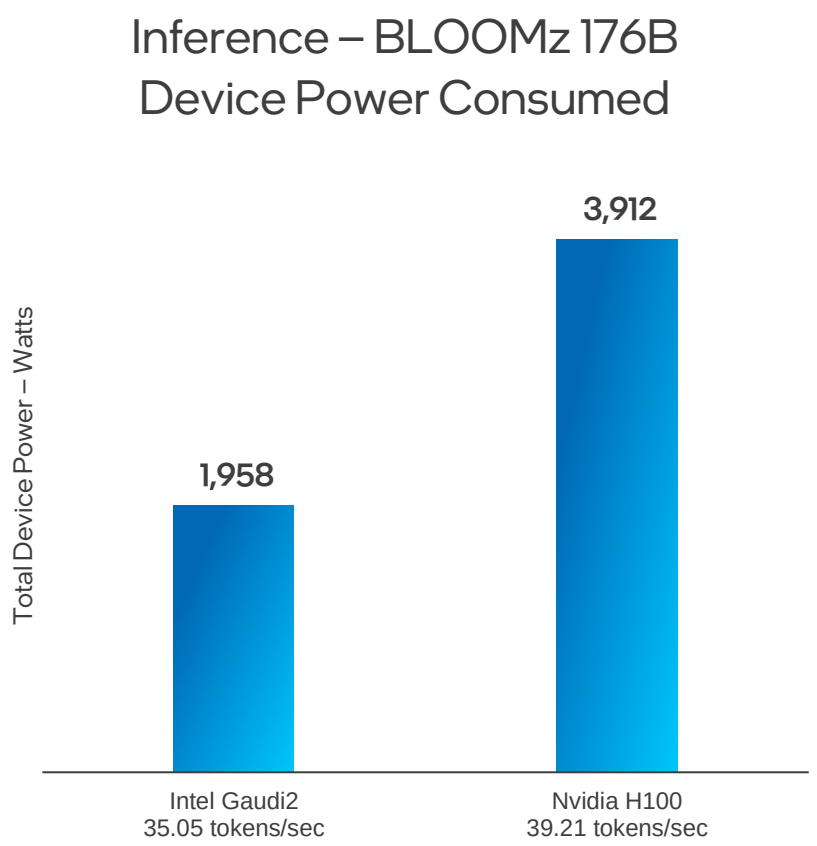
Performance source: MLPerf Training 3.1 results https://mlcommons.org/benchmarks/training/

Lower Power Means Better TCO, Lower CO_2
Intel® Gaudi® 2 Power on LLM: BLOOMZ 176B

Intel Gaudi 2 AI accelerator vs. Nvidia H100

50% lower device power

79% higher throughput-per-watt

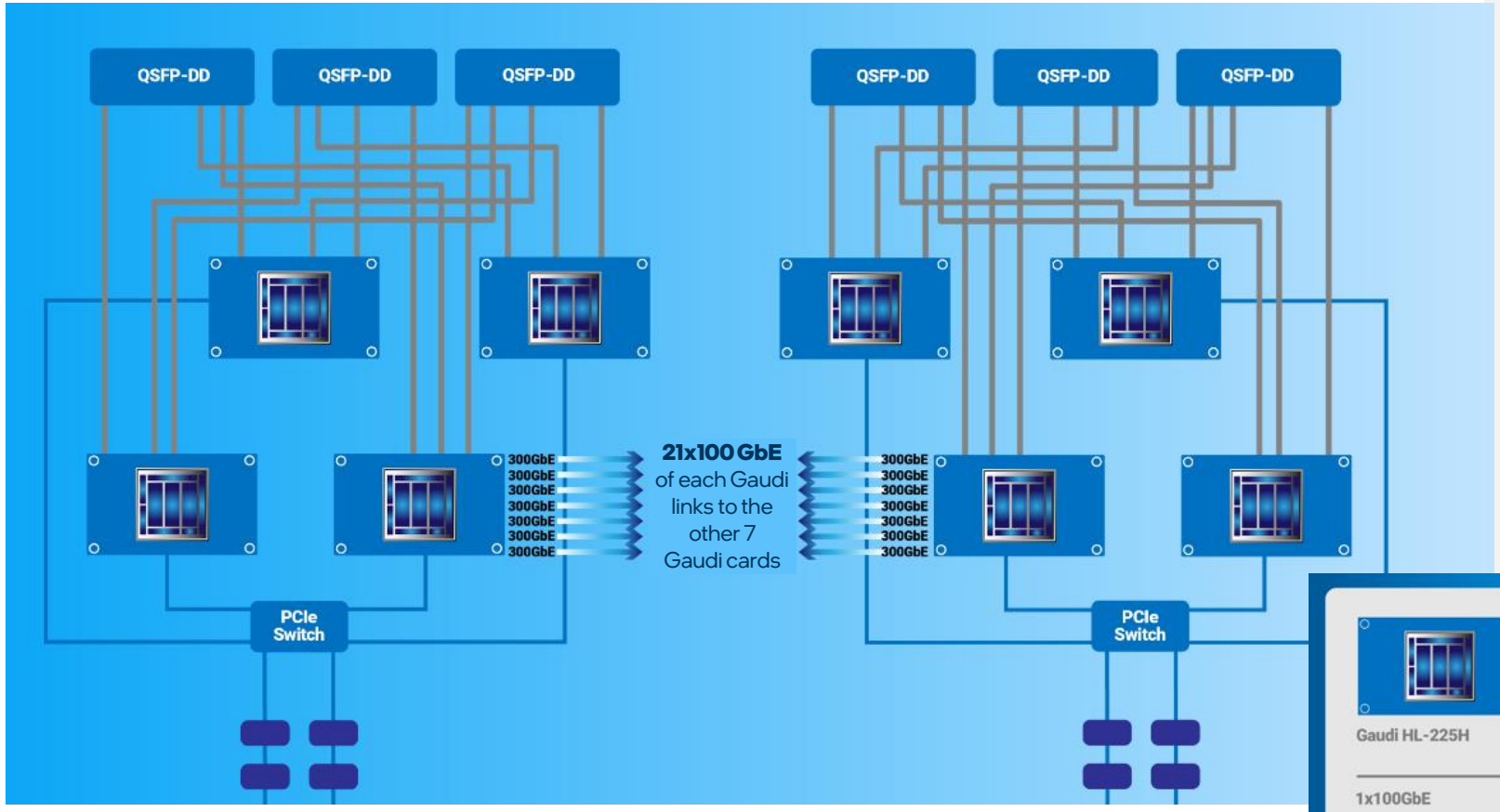


For test details and configuration, please see the Appendix or https://habana.ai/habana-claims-validation/. Results may vary.

Intel® Gaudi® 2 Server: Designed for Flexible, Efficient Scalability

HLS-Gaudi 2 Reference Server featuring...

- 8 Intel Gaudi 2 mezzanine cards
- 24x 100 GbE ports per card
  - 21 for all-to-all connectivity to other 7 Intel Gaudi processors within the server
  - Three to scale out
  - Through six QSFP-DD ports
- Dual-socket Host CPU: Intel® Xeon® Scalable processor

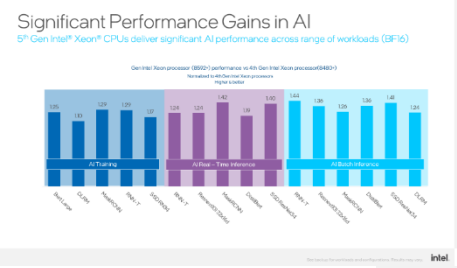


Dual-socket Intel Xeon Scalable Processor Host CPU

Thank you

Configuration details

Configuration details: AI performance



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

ResNeXT101_32x16d Inference: BS1, BSx

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 : BS1, BSx

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 : BS1, BSx

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 2.0, microcode 0x2b0004d0, 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 10/25/23.

Software configuration: DistilBERT, Intel Model Zoo:https://github.com/IntelAI/models, gcc=12.3, OneDNN3.2, Python 3.9, PyTorch 2.0, IPEX 2.0, Transformer version 4.18.0, physical cores only.

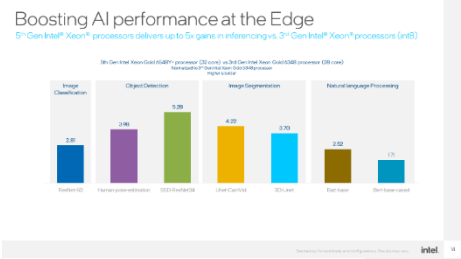
MaskRCNN Inference: : BS1, BSx

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

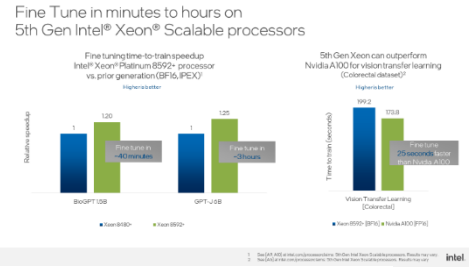
Configuration Details: Boosting AI performance at the Edge



- 6548Y+: 1-node, 2x INTEL(R) XEON(R) GOLD 6548Y+, 32 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 2 [0], DSA 2 [0], IAA 2 [0], QAT 2 [0], Total Memory 512GB (16x32GB DDR5 4800 MT/s [4800 MT/s]), BIOS EGSDREL1.SYS.0105.D74.2308261931, microcode 0x21000161, 1x Ethernet Controller I225-LM, 1x 931.5G WD Green 2.5 1000GB, 1x 931.5G CT1000MX500SSD1, CentOS Stream 9, 6.2.0, OpenVino Ver 2023.1.0 benchmark_app for Resnet50, SSD-Resnet34, 3D-UNET, ssd-mobilenet, unet-camvid, bart-base, bert-base-cased, human-pose-estimation. INT8, BS=1 Test by Intel as of 10/19/23.
- 6348: 1-node, 2x Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz, 28 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 0 [0], DSA 0 [0], IAA 0 [0], QAT 0 [0], Total Memory 512GB (16x32GB DDR4 3200 MT/s [3200 MT/s]), BIOS SE5C620.86B.01.01.0008.2305172341, microcode 0xd0003a5, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZ1L21T9HCLS-00A07, CentOS Stream 9, 6.2.0, OpenVino Ver 2023.1.0 benchmark_app for Resnet50, SSD-Resnet34, 3D-UNET, ssd-mobilenet, unet-camvid, bart-base, bert-base-cased, human-pose-estimation. INT8, BS=1 Test by Intel as of 10/19/23.

Fine tune in minutes to hours – Time to-train

Configuration details



Vision Transfer Learning for colorectal cancer detection: 25 seconds less time to train ResNet model for colorectal cancer classification via transfer learning on a dual socket 5th Gen Intel Xeon Platinum 8592+ with AMX BF16 compared to Nvidia A100 GPU

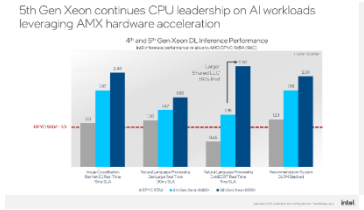
BioGPT: Fine-tuning to SOTA accuracy of 79.4% on NV A100 takes 20.2 minutes, single node EMR takes 40.1 minutes with 8 DDP instances

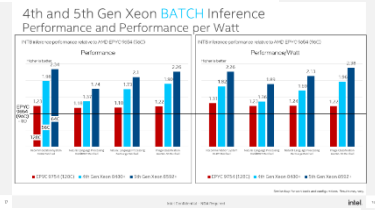
GPT-J: Fine-tune the 6 billion parameter GPT-J model in ~3 hrs on a single node 5th Gen Intel Xeon scalable processor.

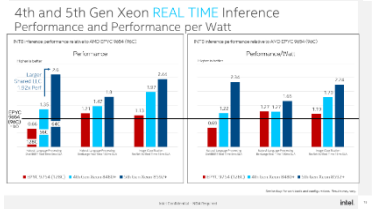
Hardware configuration for Intel® Xeon® Platinum 8592+ processor (formerly code named Emerald Rapids): 2 sockets, 64 cores, 350 watts, 16 x 64 GB DDR5 5600 memory, BIOS version 3B05.TEL4P1, operating system: CentOS stream 8, using Intel® Advanced Matrix Extensions (Intel® AMX) int8 and bf16 with Intel® oneAPI Deep Neural Network Library (oneDNN) v2.6.0 optimized kernels integrated into Intel® Extension for PyTorch* v2.0.1, Intel® Extension for TensorFlow* v2.14. Measurements may vary.

Hardware configuration for Intel® Xeon® Platinum 8480+ processor (formerly code named Sapphire Rapids): 2 sockets, 56 cores, 350 watts, 16 x 64 GB DDR5 4800 memory, BIOS version EGSDCRB1.SYS.0102.D37.2305081420, operating system: CentOS stream 8, using Intel® Advanced Matrix Extensions (Intel® AMX) int8 and bf16 with Intel® oneAPI Deep Neural Network Library (oneDNN) v2.6.0 optimized kernels integrated into Intel® Extension for PyTorch* v2.0.1, Intel® Extension for TensorFlow* v2.14. Measurements may vary.

Resources and Configurations





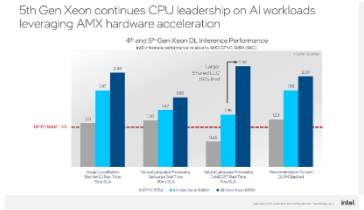


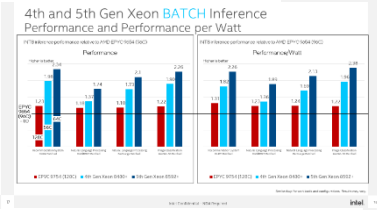
4th and 5th Gen Xeon DL Inference Performance

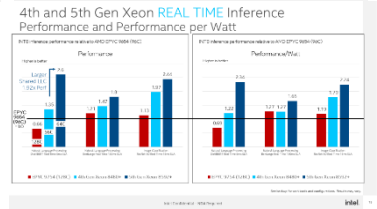
ResNet50v1.5

- Intel Xeon 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. TensorFlow= Intel TF 2.13, OneDNN=3.2, Python 3.8, AI Model=ResNet50v1.5 (https://github.com/IntelAI/models/), INT8-AMX, Real Time (BS=1) results while maintaining 15ms latency SLA, Test by INTEL as of 10/10/2023.
- Intel Xeon 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, TensorFlow= Intel TF 2.13, OneDNN=3.2, Python 3.8, AI Model=ResNet50v1.5 (https://github.com/IntelAI/models/), INT8-AMX, Real Time (BS=1) results while maintaining 15ms latency SLA, Test by Intel as of 10/25/23.
- AMD EPYC 9654: 1-node, 2x AMD EPYC 9654 96-Core Processor, 96 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, ZenDNN 4.1, TensorFlow= 2.12.1, Python 3.8, AI Model=ResNet50v1.5 (https://github.com/IntelAI/models/) ,, INT8, Real Time (BS=1) results while maintaining 15ms latency SLA. Test by INTEL as of 09/11/23.
- AMD EPYC 9754: 1-node, 2x AMD EPYC 9754 128-Core Processor, 128 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]),BIOS 1.5., microcode 0xaa00212, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, ZenDNN 4.1, TensorFlow= 2.12.1, Python 3.8, AI Model=ResNet50v1.5 (https://github.com/IntelAI/models/) , INT8, Real Time (BS=1) results while maintaining 15ms latency SLA. Test by INTEL as of 10/26/23.

Resources and Configurations





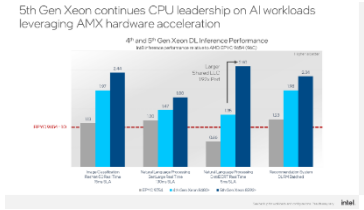


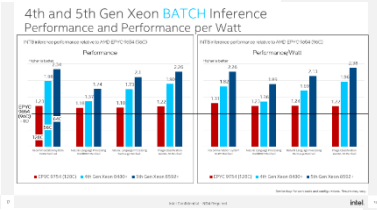
4th and 5th Gen Xeon DL Inference Performance

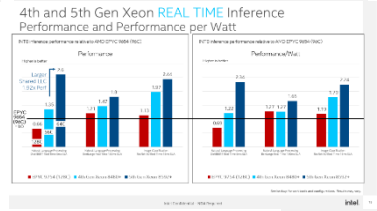
BERT-Large

- Intel Xeon 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. Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=BERT Large(https://github.com/IntelAI/models/) , INT8-AMX , Real Time (BS=1) results while maintaining 130ms latency SLA, Test by INTEL as of 10/10/2023.
- Intel Xeon 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 Controller I225-LM,11x 1.7T SAMSUNG MZQL21T9HCJR-00A07,Ubuntu 22.04.2 LTS, 5.15.0-78-generic. Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=BERT Large(https://github.com/IntelAI/models/) , INT8-AMX, Real Time (BS=1) results while maintaining 130ms latency SLA . Test by INTEL as of 09/05/2023.
- AMD EPYC 9654: 1-node, 2x AMD EPYC 9654 96-Core Processor, 96 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=BERT Large(https://github.com/IntelAI/models/) , INT8, Real Time (BS=1) results while maintaining 130ms latency SLA. Test by INTEL as of 09/11/23.
- AMD EPYC 9754: 1-node, 2x AMD EPYC 9754 128-Core Processor, 128 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xaa00212, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=BERT Large(https://github.com/IntelAI/models/) , INT8, Real Time (BS=1) results while maintaining 130ms latency SLA. Test by INTEL as of 10/26/23.

Resources and Configurations





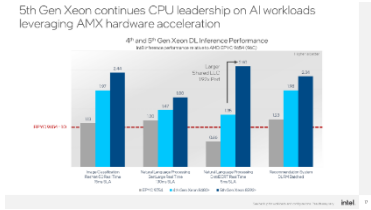


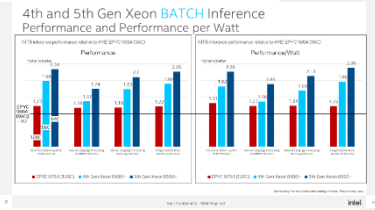
4th and 5th Gen Xeon DL Inference Performance

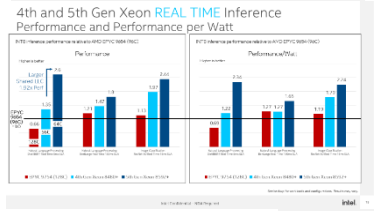
DistilBERT

- Intel Xeon 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 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. Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=DistilBERT (https://github.com/IntelAI/models/), INT8-AMX, Real Time (BS=1) results while maintaining 5ms latency SLA, Test by INTEL as of 10/10/2023.
- Intel Xeon 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 Controller I225-LM,1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic. Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=DistilBERT (https://github.com/IntelAI/models/) , INT8-AMX, Real Time (BS=1) results while maintaining 5ms latency SLA . Test by INTEL as of 09/05/2023.
- AMD EPYC 9654: 1-node, 2x AMD EPYC 9654 96-Core Processor, 96 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=DistilBERT (https://github.com/IntelAI/models/) , INT8 Real Time (BS=1) results while maintaining 5ms latency SLA. Test by INTEL as of 09/11/23.
- AMD EPYC 9754: 1-node, 2x AMD EPYC 9754 128-Core Processor, 128 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xaa00212, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=DistilBERT (https://github.com/IntelAI/models/) , INT8 Real Time (BS=1) results while maintaining 5ms latency SLA. Test by INTEL as of 10/26/23.

Resources and Configurations





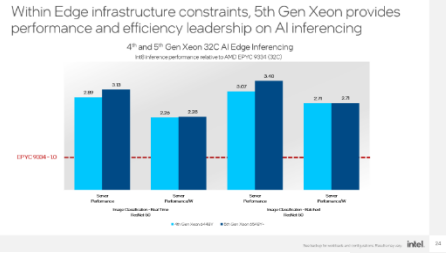


4th and 5th Gen Xeon DL Inference Performance

DLRM

- Intel Xeon 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. Framework=Pytorch 2.1, IPEX=2.1, oneDNN:v3.2.1, Python 3.8, AI Model= DLRM(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BS>1, Precision=INT8-AMX, Test by INTEL as of 10/10/2023.
- Intel Xeon 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 0x2b0004b1, 1x Ethernet Controller I225-LM, 1x 894.3G INTEL SSDSC2KG96, Ubuntu 22.04.2 LTS,5.15.0-78-generic, Framework=PyTorch 2.1, IPEX=2.1, oneDNN:v3.2.1, Python 3.8, AI Model= DLRM(https://github.com/IntelAI/models/), Batched Results: best scores achieved using BS>1, Precision=INT8-AMX, Test by INTEL as of 09/05/2023.
- AMD EPYC 9654: 1-node, 2x AMD EPYC 9654 96-Core Processor, 96 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.1, IPEX=2.1, oneDNN: v3.2.1, Python 3.8, AI Model= DLRM(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BS>1, Precision=INT8. Test by INTEL as of 09/11/23.
- AMD EPYC 9754: 1-node, 2x AMD EPYC 9754 128-Core Processor, 128 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.1, IPEX=2.1, oneDNN: v3.2.1, Python 3.8, AI Model= DLRM(https://github.com/IntelAI/models/) , Batched Results: best scores achieved using BS>1, Precision=INT8. Test by INTEL as of 09/11/23.

Resources and Configurations

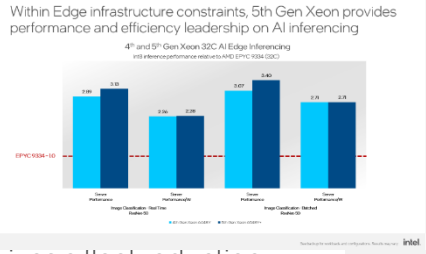


5th Gen Xeon 32C AI Inferencing for Real-World Scale

ResNet50v1.5

- Intel Xeon 6548Y+: 1-node, 2x INTEL(R) XEON(R) GOLD 6548Y+, 32 cores, HT On, Turbo On, NUMA 2,, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5200 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T Micron_7450_MTFDKCB3T8TFR, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, ResNet50 v1.5 Intel TensorFlow 2.13, OneDNN=3.2, INT8-AMX, Real Time (BS=1) results while maintaining 15ms latency SLA, Batched Results: best scores achieved using (BS>1), Test by INTEL as of 11/20/23.
- Intel Xeon 6448Y: 1-node, 2x Intel(R) Xeon(R) Gold 6448Y, 32 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 2.0, microcode 0x2b0004d0, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T Micron_7450_MTFDKCB3T8TFR, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, ResNet50 v1.5 Intel TensorFlow 2.13, OneDNN=3.2. INT8-AMX, Real Time (BS=1) results while maintaining 15ms latency SLA, Batched Results: best scores achieved using (BS>1), Test by INTEL as of 11/20/23.
- AMD EPYC 9334: 1-node, 2x AMD EPYC 9334 32-Core Processor, 32 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, ResNet50 v1.5, ZenDNN 4.1 TensorFlow 2.12.1. INT8, Real Time (BS=1) results while maintaining 15ms latency SLA, Batched Results: best scores achieved using (BS>1), Test by INTEL as of 10/23/23.

Configurations: 5th Gen Xeon TCO Advantages (1 of 6)



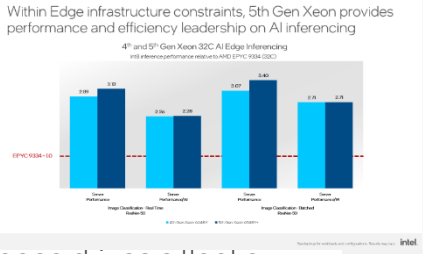
Claim: 5th Gen Intel Xeon delivers up to 24% lower TCO than the 4th Gen AMD Epyc while running a HammerDB MySQL OLTP database workload.

- Based on 5th Gen Intel Xeon delivers up to a 1.70x faster than the 4th Gen AMD Epyc while running a HammerDB MySQL OLTP workload. This performance drives a fleet reduction from 50 to 30
- servers which, over 4 years, saves: 684.0 MWH of energy, 289,967 kgCo2 emissions, and $508.9k of cost.
- 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 8 [0], DSA 8 [0], IAX 8 [0], QAT 8 [0], 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, 2x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.3 LTS, 5.15.0-84-generic, HammerDB Mv4.4, MySQL 8.0.33. Test by Intel as of 10/04/23.
- 1-node, 2x AMD EPYC 9554 64-Core Processor, 64 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 0 [0], DSA 0 [0], IAX 0 [0], QAT 0 [0], Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, 2x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, HammerDB v4.4, MySQL 8.0.33. Test by Intel as of 10/05/23.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.40M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $755.1K
- Energy use in kWh (4 year, per server): 47654, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 30 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $1.17M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $480.0K
- Energy use in kWh (4 year, per server): 58625, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates andinformation from thinkmate.com as of October 2023.

Configurations: 5th Gen Xeon TCO Advantages (2 of 6)



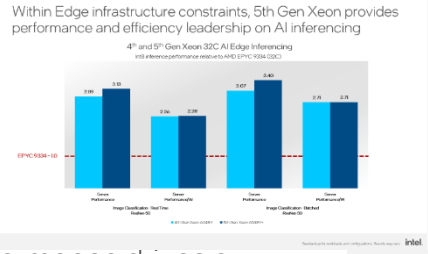
Claim: 5th Gen Intel Xeon delivers up to 22% lower TCO than the 4th Gen AMD Epyc while running a RocksDB database workload.

- Based on 5th Gen Intel Xeon delivers up to a 1.62x faster than the 4th Gen AMD Epyc while running a RocksDB database workload. This performance drives a fleet a reduction from 50 to 31 servers which, over 4 years, saves: 1,218 MWH of energy, 516,402 kgCo2 emissions, and $471.8k of cost.
- 1-node, 2x 5th Gen Intel Xeon Scalable processor 8592+ (64 cores) with integrated Intel In-Memory Analytics Accelerator (Intel IAA), Number of IAA device utilized=8(2 sockets active), HT On, Turbo On, SNC off, Total Memory 1024GB (16x64GB DDR5 5600), microcode 0x21000161, 2x Ethernet Controller 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 6.5.0-060500-generic, QPL v1.2.0, accel-config-v4.0, iaa_compressor plugin v0.3.0, ZSTD v1.5.5, gcc 10.4.0, RocksDB v8.3.0 trunk (commit 62fc15f) (db_bench), 4 threads per instance, 64 RocksDB instances, tested by Intel October 2023.
- 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost On, NPS1, Total Memory 1024GB (16x64GB DDR5 4800), microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 6.5.0-060500-generic, ZSTD v1.5.5, gcc 10.4.0, RocksDB v8.3.0 trunk (commit 62fc15f) (db_bench), 4 threads per instance, 28 RocksDB instances, tested by Intel October 2023.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.36M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $809.5K
- Energy use in kWh (4 year, per server): 58531, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 31 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $1.21M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $491.3K
- Energy use in kWh (4 year, per server): 55111, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023.

Configurations: 5th Gen Xeon TCO Advantages (3 of 6)



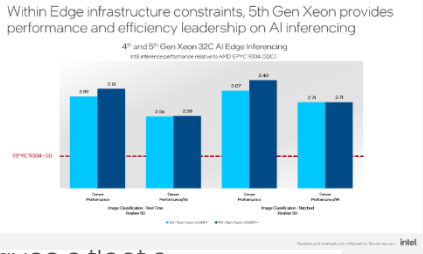
Claim: 5th Gen Intel Xeon delivers up to 21% lower TCO than the 4th Gen AMD Epyc while running a NGINX TLS Handshake workload.

- Based on 5th Gen Intel Xeon delivers up to a 1.66x faster than the 4th Gen AMD Epyc while running a NGINX TLS Handshake workload. This performance drives a fleet a reduction from 50 to 31 servers which, over 4 years, saves: 489.7 MWH of energy, 207,611 kgCo2 emissions, and $443.5k of cost.
- 1-node, 2x 5th Gen Intel Xeon Scalable processor (64 core) with integrated Intel Quick Assist Technology (Intel QAT), Integrated Accelerators Available [used]: DLB 2 [0], DSA 2 [0], IAA 2 [0], QAT 2 [0], HT On, Turbo Off, SNC On, with 1024GB DDR5 memory (16x64 GB 5600), microcode 0x21000161, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX Async v0.5.1, OpenSSL 3.1.3, IPP Crypto 2021.8, IPsec MB v 1.4, QAT _Engine v 1.4.0, QAT Driver 20.I.1.1..20-00030, TLS 1.3 Webserver: ECDHE-X25519-RSA2K, tested by Intel October 2023.
- 1-node, 9554: 1-node, AMD platform with 2x 4th Gen AMD EPYC processor (64 cores), SMT On, Core Performance Boost Off, NPS1, Total Memory 1536GB (24x64GB DDR5 4800), microcode 0xa10113e, Ubuntu 22.04.3 LTS, 5.15.0-78-generic, 1x 1.7T SAMSUNG MZWLJ1T9HBJR-00007, 1x Intel® Ethernet Network Adapter E810-2CQDA2, 1x100GbE, NGINX Async v0.5.1, OpenSSL 3.1.3, TLS 1.3 Webserver: ECDHE-X25519-RSA2K,tested by Intel October 2023.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.41M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $698.7K
- Energy use in kWh (4 year, per server): 36386, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 31 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $1.21M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $453.4K
- Energy use in kWh (4 year, per server): 42889, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023

Configurations: 5th Gen Xeon TCO Advantages (4 of 6)



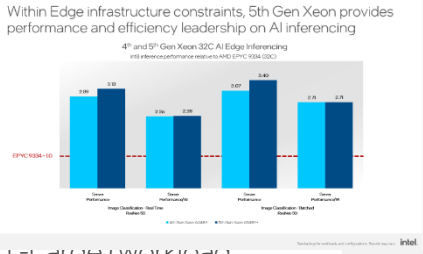
Claim: 5th Gen Intel Xeon delivers up to 27% lower TCO than the 4th Gen AMD Epyc while running Monte Carlo workload.

- Based on 5th Gen Intel Xeon delivers up to a 1.83x faster than the 4th Gen AMD Epyc while running a Monte Carlo workload. This performance drives a fleet a reduction from 50 to 28 servers which, over 4 years, saves: 585.8 MWH of energy, 248,352 kgCo2 emissions, and $561.0k of cost.
- 1-node 2x Intel Xeon 8592+, HT On, Turbo On, SNC2, 1024 GB DDR5-5600, ucode 0x21000161, Red Hat Enterprise Linux 8.7, 4.18.0-425.10.1.el8_7.x86_64, Monte Carlo v1.2, cmkl:2023.2.0 icc:2023.2.0 tbb:2021.10.0. Test by Intel as of October 2023.
- 1-node, 2x AMD EPYC 9554, SMT On, Turbo On, CTDP=360W, NPS=4, 1536GB DDR5-4800, ucode= 0xa101111, Red Hat Enterprise Linux 8.7, Kernel 4.18,Monte Carlo v1.2, cmkl:2023.2.0 icc:2023.2.0 tbb:2021.10.0. Test by Intel as of March 2023

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.36M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $739.3K
- Energy use in kWh (4 year, per server): 44505, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 28 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $1.09M
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $453.3K
- Energy use in kWh (4 year, per server): 58550, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023.

Configurations: 5th Gen Xeon TCO Advantages (5 of 6)



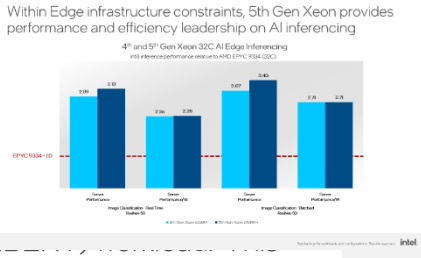
Claim: 5th Gen Intel Xeon delivers up to 46% lower TCO than the 4th Gen AMD Epyc while running a real-time Natural Language Processing inference (BERT-La

- Based on 5th Gen Intel Xeon delivers up to a 2.44x faster than the 4th Gen AMD Epyc while running a real-time Natural Language Processing inference (BERT-Large) workload. This performance drives a fleet a reduction from 50 to 21 servers which, over 4 years, saves: 1231.2 MWH of energy, 521,941 kgCo2 emissions, and $982.9k of cost.
- Intel Xeon 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. Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=BERT Large(https://github.com/IntelAI/models/) , INT8-AMX , Real Time (BS=1) results while maintaining 130ms latency SLA, Test by INTEL as of 10/10/2023.
- 9554: 1-node, 2x AMD EPYC 9554 64-Core Processor, 64 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=BERT Large(https://github.com/IntelAI/models/) , INT8, Real Time (BS=1) results while maintaining 130ms latency SLA. Test by INTEL as of 09/11/23.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.36
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $766.9K
- Energy use in kWh (4 year, per server): 50021, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

- For a 21 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $801K
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $344.0K
- Energy use in kWh (4 year, per server): 60472, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023.

Configurations: 5th Gen Xeon TCO Advantages (6 of 6)



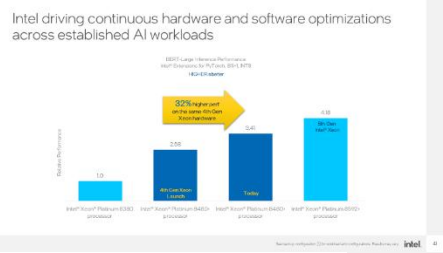
Claim: 5th Gen Intel Xeon delivers up to 62% lower TCO than the 4th Gen AMD Epyc while running real-time Natural Language Processing inference (DistilBERT

- Based on 5th Gen Intel Xeon delivers up to a 3.49x faster than the 4th Gen AMD Epyc while running a real-time Natural Language Processing inference (Disti… performance drives a fleet a reduction from 50 to 15 servers which, over 4 years, saves: 1496.5 MWH of energy, 634,428 kgCo2 emissions, and $1,300k of cost.
- 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. Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=DistilBERT (https://github.com/IntelAI/models/) , INT8-AMX, Real Time (BS=1) results while maintaining 5ms latency SLA, Test by INTEL as of 10/10/2023.
- 1-node, 2x AMD EPYC 9554 64-Core Processor, 64 cores, SMT On, Turbo On, NUMA 2, Total Memory 1536GB (24x64GB DDR5 4800 MT/s [4800 MT/s]), BIOS 1.5, microcode 0xa10113e, 2x Ethernet Controller 10G X550T, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.125-0515125-generic, Framework=Pytorch 2.0, IPEX=2.0, Python 3.8, AI Model=DistilBERT Large(https://github.com/IntelAI/models/) , INT8, Real Time (BS=1) results while maintaining 5ms latency SLA. Test by INTEL as of 09/11/23.

- For a 50 server fleet of AMD EPYC 9554, estimated as of October 2023:
- CapEx costs: $1.36
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $749.7K
- Energy use in kWh (4 year, per server): 46573, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394

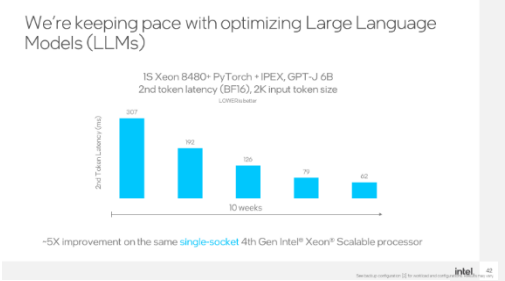
- For a 15 server fleet of 5th Gen Xeon 8592+ as of October 2023
- CapEx costs: $572K
- OpEx costs (4 year, includes power and cooling utility costs, infrastructure and hardware maintenance costs): $238.3K
- Energy use in kWh (4 year, per server): 55475, PUE 1.6
- Other assumptions: utility cost $0.1/kWh, kWh to kg CO2 factor 0.42394
- Costs based on Intel estimates and information from thinkmate.com as of October 2023.

Gen-to-gen hardware and software optimizations



- Hardware configuration for Intel® Xeon® Platinum 8380 processor (formerly code named Ice Lake): 2 sockets, 40 cores, 270 watts, 16 x 64 GB DDR5 3200 memory, BIOS version SE5C620.86B.01.01.0005.2202160810, operating system: Ubuntu 22.04.1 LTS, INT8 with Intel® oneAPI Deep Neural Network Library (oneDNN) v2.6.0 optimized kernels integrated into Intel® Extension for PyTorch* v1.13, Intel® Extension for TensorFlow* v2.11, and Intel® Distribution of OpenVINO™ toolkit v2022.3, BS=1, Test by Intel as of Oct 2022. Results may vary.
- Hardware configuration for Intel® Xeon® Platinum 8480+ processor (formerly code named Sapphire Rapids) (Oct 2022): 2 sockets, 56 cores, 350 watts, 16 x 64 GB DDR5 4800 memory, operating system CentOS* Stream 8 with BIOS version EGSDCRB1.SYS.0093.D22.2211170057 and Ubuntu 22.04.1 with BIOS version SE5C7411.86B.9525.D07.2301120334. Using Intel® Advanced Matrix Extensions (Intel® AMX) INT8 with Intel® oneAPI Deep Neural Network Library (oneDNN) optimized kernels integrated into Intel® Extension for PyTorch v1.13*, Intel® Extension for TensorFlow v2.11*, and Intel® Distribution of OpenVINO™ toolkit v2022.3, BS=1. Test by Intel as of Oct 2022. Results may vary.
- Hardware configuration for Intel® Xeon® Platinum 8480+ processor (code named Sapphire Rapids) (Oct 2023): 2 sockets for inference, 1 socket for training, 56 cores, 350 watts, 1024GB 16 x 64GB DDR5 4800 MT/s memory, operating system CentOS* Stream 8. Using Intel® Advanced Matrix Extensions (Intel® AMX) INT8 with Intel® oneAPI Deep Neural Network Library (oneDNN) optimized kernels integrated into Intel® Extension for PyTorch* v2.1, Intel® Extension for TensorFlow* v2.14, and Intel® Distribution of OpenVINO™ toolkit v2023.0, BS=1. Test by Intel as of Oct 2023. Results may vary.
- Hardware configuration for Intel® Xeon® Platinum 8592+ processor (code named Emerald Rapids): 2 sockets for inference, 1 socket for training, 64 cores, 350 watts, 1024GB 16 x 64GB DDR55600 MT/s memory, operating system CentOS* Stream 9. Using Intel® Advanced Matrix Extensions (Intel® AMX) INT8 with Intel® oneAPI Deep Neural Network Library (oneDNN) optimized kernels integrated into Intel® Extension for PyTorch* v2.1, Intel® Extension for TensorFlow* v2.14, and Intel® Distribution of OpenVINO™ toolkit v2023.0, BS=1. Test by Intel as of Oct 2023. Results may vary.

4th Gen Intel® Xeon® Scalable processor GPT-J optimizations over time

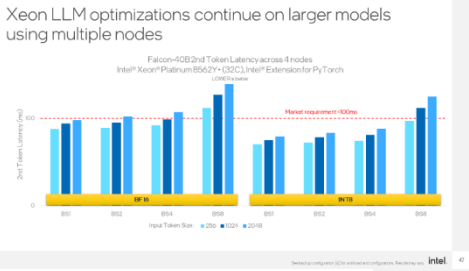


1-node, 2x Intel(R) Xeon(R) Platinum 8480+, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 512GB (16x64GB DDR5 4800 MT/s [4800 MT/s]), AMI BIOS, microcode 3A06, 2x Ethernet Controller X710 for 10GBASE-T, 1x 1.8T Samsung SSD 970 EVO Plus, CentOS Stream9, 5.14.0-378.el9.x86_64, Text generation on GPT-J 6B , gcc 12.3, IPEX CPU + PyTorch 2.1. BF16, BS1 CPI 56, Input token size = 1016, Output token size = 32. Test by Intel as of 09/27/23. Results may vary.

5th Gen Xeon Large Language Model Latency

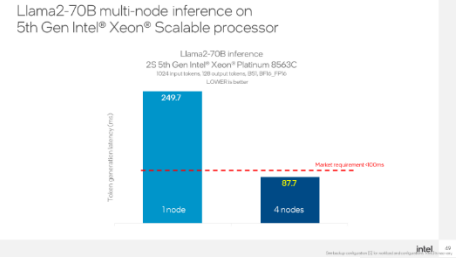
- Llama2-13B
- 8592+: 1-node, 2x INTEL(R) XEON(R) PLATINUM 8592+, 64 cores, HT On, Turbo On, NUMA 2, Total Memory 1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s]), BIOS 2.0, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x Ethernet interface, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.2 LTS, 5.15.0-78-generic, Test by Intel as of 10/10/23.
- Llama2, gcc=12.3 gxx=12.3, python 3.9, intel-extension-for-pytorch 2.2.0+,git305fc52, torch 2.2.0.dev20230914+cpu, transformers 4.31.0, protobuf 3.20.3, neural-compressor 2.2, test on 1 socket, Batch Size(BS)=1, Corest per instance (CPI)=#cores/socket, search algorithm: BEAM4.
- GPT-J 6B
- 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 10-Gigabit X540-AT2, 1x 1.7T SAMSUNG MZQL21T9HCJR-00A07, Ubuntu 22.04.3 LTS, 5.15.0-83-generic, GPT-J 6B, gcc=12.3, python 3.9intel-extension-for-pytorch 2.2.0+git305fc52torch 2.2.0.dev20230914+cputransformers 4.31.0protobuf 3.20.3neural-compressor 2.2, Test on 1 socket, BS1 CPI 64. Test by Intel as of 09/21/23.

Xeon LLM optimizations continue on larger models using multiple nodes (Falcon-40B)



4-nodes, 2x INTEL(R) XEON(R) PLATINUM 8562Y+, 32 cores, HT On, Turbo On, NUMA 2, Total Memory 512GB (16x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS 3B05.TEL4P1, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 2x Ethernet Controller E810-C for QSFP, 1x 894.3G INTEL SSDSC2KG96, 1x 3.5T SAMSUNG MZQL23T8HCLS-00A07, 3x 3.5T SAMSUNG MZQL23T8HCLS-00B7C, Red Hat Enterprise Linux 8.8 (Ootpa), 4.18.0-477.21.1.el8_8.x86_64, gcc=12.3, gxx=12.3, Torch 2.2.0.dev20230911+cpu, IPEX 2.2.0+git880fda9/llm_feature_branch, Deepspeed 0.10.2+f15e6d48, Transformers 4.31.0, torch-ccl (ccl_torch_dev_0905), Pytorch Falcon-40b model BF16/INT8 precision, Batch Size: 1,2,4,8, Input Token size: 32,256,1024,2048, Output Token size: 32, 99ms with BS4/1024 Input Tokens with BFloat16 precision, 97ms with BS8/256 Input Tokens with INT8 precision, Test by Intel as of 11/08/23.

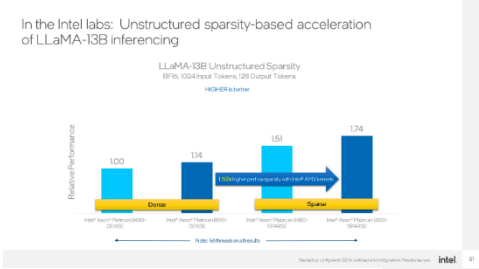
Llama2-70B multi-node



Single node: 1 node, 2x INTEL(R) XEON(R) PLATINUM 8563C, 52 cores, HT On, Turbo On, NUMA 2, Total Memory 512GB (16x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS 3B05.TEL4P1, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 2x MT27800 Family [ConnectX-5], 1x INTEL SSDPF2KX038TZ (3.5T), CentOS Stream 9 kernel 6.2.15, llama-2-70B distributed inference task, xFasterTransformer 1.0.0, gcc-11.3.1/g++, oneDNN-3.2, MKLML, oneCCL, max number of threads = 52, dataset = boolq/piqa/lambada,BS=1, BF16_FP16, number of instances = 2, Input Token size: 1024, Output Token size: 128. Test by Intel as of 11/28/23. Results may vary.

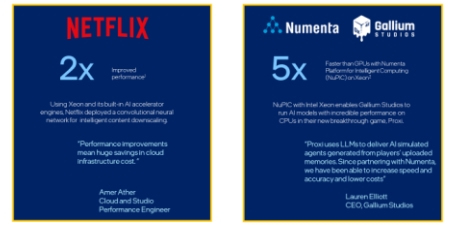
4 nodes: 4 nodes, each node with 2x INTEL(R) XEON(R) PLATINUM 8563C, 52 cores, HT On, Turbo On, NUMA 2, Total Memory 512GB (16x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS 3B05.TEL4P1, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 2x MT27800 Family [ConnectX-5], 1x INTEL SSDPF2KX038TZ (3.5T), CentOS Stream 9 kernel 6.2.15, llama-2-70B distributed inference task, xFasterTransformer 1.0.0, gcc-11.3.1/g++, oneDNN-3.2, MKLML, oneCCL, max number of threads = 52, dataset = boolq/piqa/lambada,BS=1, BF16_FP16, number of instances = 8, Input Token size: 1024, Output Token size: 128. Test by Intel as of 11/28/23. Results may vary.

Unstructured sparsity-based acceleration of LLaMA-13B inferencing



4th Gen Intel® Xeon® Scalable processor configuration (dense): 1-node, 2x Intel(R) Xeon(R) Platinum 8480L, 56 cores, HT On, Turbo On, NUMA 2, Total Memory 512GB (16x32GB DDR5 4800 MT/s [4800 MT/s]), BIOS 05.01.00, microcode 0x2b0004b1, 4x I350 Gigabit Network Connection, 2x Omni-Path HFI Silicon 100 Series [discrete], 1x Ethernet interface, 1x 894.3G Micron_5400_MTFDDAV960TGA, Rocky Linux 9.2 (Blue Onyx), 5.14.0-284.30.1.el9_2.x86_64, LLM inference with Llama 13B model, Pytorch 2.0.0, libxsmm, Intel TPP extensions for Pytorch, max number of threads = 56, Llama-13B dense, Input token size = 1024, Output token size = 128, BS1, BF16, number of instances = 2. Tested by Intel as of 11/29/23. Results may vary.

5th Gen Intel® Xeon® Scalable processor configuration (sparse): 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 3B05.TEL4P1, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 1x 3.5T Micron_7450_MTFDKCB3T8TFR, Red Hat Enterprise Linux 9.3 (Plow), 5.14.0-362.8.1.el9_3.x86_64, LLM inference with Llama 13B model, Pytorch 2.0.0, libxsmm, Intel TPP extensions for Pytorch, max number of threads = 56, Llama-13B with 50% unstructured sparsity, Input token size = 1024, Output token size = 128, BS1, BF16, number of instances = 2. Tested by Intel as of 11/29/23. Results may vary.

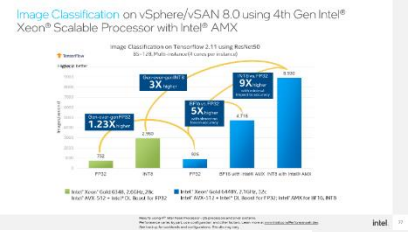


Numenta and Gallium Studios

NuPIC on Xeon is 5x faster than GPUs

- Numenta: BERT-Large: Sequence Length 64, Batch Size 1, throughput optimized
  - NVIDIA Tesla M60 GPU: Tested by Gallium as of 11/17/2023. 1-node, 1x GPU on AWS g3s.xlarge, 4vCPU, 30.5 GB memory, 8 GB GPU memory, Ubuntu 22.04 Kernel 5.17, Gallium custom Word2vec, Sequence Length 64, Batch Size 64.
  - 4th Gen Intel® Xeon® Scalable processor: Tested by Gallium as of 11/28/2023. 1-node, 2x Intel® Xeon® Platinum 8488C on AWS m7i.xlarge, 4vCPU, 16GB memory, Ubuntu 22.04 Kernel 5.17, Numenta Platform for Intelligent Computing V1.0, Numenta-Optimized BERT-Large, Sequence Length 64, BF16, Batch Size 64.

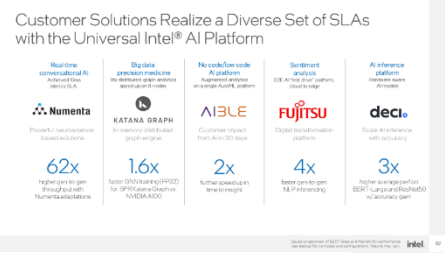
Configuration details: AI on vSphere/vSAN 8.0 (Gen-to-Gen)



3rd Gen Intel® Xeon® Scalable platform configuration: 4-node cluster, Each node: 2x Intel® Xeon® Gold 6348 Processor, 1x Server Board M50CYP2UR, Total Memory 512 GB (16x 32GB DDR4 3200MHz), HyperThreading: Enable, Turbo: Enabled, NUMA noSNC, Intel VMD: Enabled, BIOS: SE5C620.86B.01.01.0006.2207150335 (ucode:0xd000375), Storage (boot): 2x 80 GB Solidigm SSD P1600X, Storage (cache): 2x 400 GB Intel® Optane™ DC SSD P5800X Series, Storage (capacity): 6x 3.84 TB Solidigm SSD DC P5510 Series PCIe NVMe , Network devices: 1x Intel Ethernet E810CQDA2 E810-CQDA2, FW 4.0, at 100 GbE RoCE, Network speed: 100 GbE, OS/Software: VMware/vSAN 8.0, 20513097, Test by Intel as of 03/08/2023 using Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN default policy (RAID-1, 2DG), Kernel 5.15, Intel-optimized-tensorflow:2.11.0, ResNet50v1.5, Batch size=128, VM=56vCPU+64GBRAM, Multi-instance scenario (4 cores per instance), BERT-Large, SQuAD 1.1, Batch size=128, VM=56vCPU+64GBRAM

4th Gen Intel® Xeon® Scalable platform configuration: 4-node cluster, Each node: 2x Intel® Xeon® Gold 6448Y Processor QS pre-production, 1x Server Board M50FCP2SBSTD, Total Memory 512 GB (16x DDR5 32GB 4800MHz), HyperThreading: Enable, Turbo: Enabled, NUMA noSNC, Intel VMD: Enabled, BIOS: SE5C741.86B.01.01.0002.2212220608 (ucode:0x2b000161), Storage (boot): 2x240GB Solidigm S4520, Storage (data): 6x 3.84 TB Solidigm SSD DC P5510 Series PCIe NVMe , Network devices: 1x Intel Ethernet E810CQDA2 E810-CQDA2, FW 4.0, at 100 GbE RoCE, Network speed: 100 GbE, OS/Software: VMware/vSAN 8.0, 20513097, Test by Intel as of 03/13/2023 using Ubuntu Server 22.04 VM (vHW=20, vmxnet3), vSAN ESA – Optimal default policy (RAID-5, flat), Kernel 5.15, intel-optimized-tensorflow:2.11.0, ResNet50v1.5, Batch size=128, VM=64vCPU+64GBRAM, Multi-instance scenario (4 cores per instance), BERT-Large, SQuAD 1.1, Batch size=128, VM=64vCPU+64GBRAM

Customer Solutions
Numenta configuration details



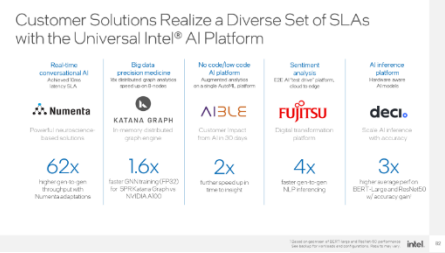
62x higher gen-to-gen throughput with Numenta value-add based on 3rd Gen Intel® Xeon® Scalable Processor (AVX512)) without Numenta optimization compared to 4th Gen Intel® Xeon® Scalable Processor with Numenta optimizations.

Numenta: BERT-Large: Sequence Length 64, Batch Size 1, throughput optimized

3rd Gen Intel® Xeon® Scalable: Tested by Numenta as of 11/28/2022. 1-node, 2x Intel® Xeon®8375C on AWS m6i.32xlarge, 512 GB DDR4-3200, Ubuntu 20.04 Kernel 5.15, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 64, Batch Size 1

Intel® Xeon® 8480+: Tested by Numenta as of 11/28/2022. 1-node, 2x Intel® Xeon® 8480+, 512 GB DDR5-4800, Ubuntu 22.04 Kernel 5.17, OpenVINO 2022.3, Numenta-Optimized BERT-Large, Sequence Length 64, Batch Size 1

Customer Solutions
Aible configuration details



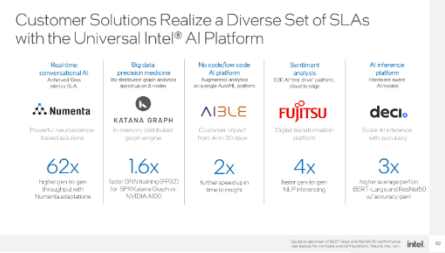
Claims:

- 2x further speed up in time to insight based on average performance on training mixed shallow and deep neural networks and ML models
- Intel® Xeon® Platinum 8480+ w/AMX delivers up to 2.79x faster NN model training than Intel® Xeon® Platinum 8380
- Intel® Xeon® Platinum 8480+ delivers up to 1.56x faster LightGBM model training than Intel® Xeon® Platinum 8380

Configurations:

- 3rd Gen Xeon Scalable Processor: Test by Intel as of November 10, 2022. 1-node with 2x Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz, 40 cores/socket, 2 sockets, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 3200 MHz [run @ 3200 MHz]), Dell® PowerEdge R750, 1.6.5, 0xd000375, Rocky Linux 8.6 (Green Obsidian), 4.18.0-372.32.1.el8_6.x86_64, gcc 8.5.0, Sapphire Rapids AI DL Software Package Customer Preview III (NDA Release) Tensorflow 2.10, intel/intel-optimized-ml:xgboost 1.4.2, Python 3.8.10 [NN Models], Intel® Distribution for Python 3.7.10 [LightGBM Models] Intel Numpy1.22.4, LightGBM 3.3.3, Kubespray 2.20.0, Multus 3.8, Calico 3.23.3, containerd 1.6.8, Docker Registry 2.8.1, Kubernetes 1.24.6 (TopologyManager-Enabled), Kubeflow 1.6.1, DirectPV 3.2.0, Minio 4.5.2, Prometheus 2.39.1, Aible's Propietary AI Workload for Enterprise Insights – NN Models [HiddenLayers/Batchsize/Epochs=5/Probability=0.5], LightGBM Models [Num_Estimators/Probability]. Model Training Time for Aible's NN Models [FP32]: 519s.
- 4th Gen Xeon Scalable Processor: Test by Intel as of November 10, 2022. 1-node with 2x Intel(R) Xeon(R) Platinum 8480+ CPU @ 2.00GHz, 56 cores/socket, 2 sockets, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 4800 M Hz [run @ 4800 MHz]), Quanta Cloud Technology Inc., QuantaGrid D54Q-2U, 3A06, 0x2b000081, Rocky Linux 8.6 (Green Obsidian), 4.18.0-372.32.1.el8_6.x86_64, gcc 8.5.0, Sapphire Rapids AI DL Software Package Customer Preview III (NDA Release) Tensorflow 2.10, intel/intel-optimized-ml:xgboost 1.4.2, Python 3.8.10 [NN Models], Intel® Distribution for Python 3.7.10 [LightGBM Models] Intel Numpy1.22.4, LightGBM 3.3.3, Kubespray 2.20.0, Multus 3.8, Calico 3.23.3, containerd 1.6.8, Docker Registry 2.8.1, Kubernetes 1.24.6 (TopologyManager-Enabled), Kubeflow 1.6.1, DirectPV 3.2.0, Minio 4.5.2, Prometheus 2.39.1, Aible's Propietary AI Workload for Enterprise Insights – NN Models [HiddenLayers/Batchsize/Epochs=5/Probability=0.5], LightGBM Models [Num_Estimators/Probability]. Model Training Time for Aible's NN Models [BFloat16]: 185.67s. Geomean of performance speedup for Aible LightGBM models on 4th Gen Xeon Scalable Processor over 3rd Gen Xeon Scalable Processor is 1.56.

Customer Solutions
Katana Graph configuration details



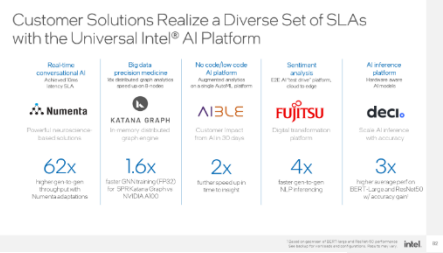
Distributed GNN Training:

- 8-node each with: 2x 4th Gen Intel Xeon Scalable processor (pre-production Sapphire Rapids >40cores) on Intel pre-production platform and software with 512 GB DDR5 memory, microcode 0x90000c0, HT on, Turbo off, Rocky Linux 8.6, 4.18.0-372.26.1.el8_6.crt1.x86_64, 931 GB SSD, 455 TB Luster filesystem with HDR fabric, Katana Graph 0.4.1 vs. DGL 0.9, test by Intel Corporation on 09/19/2022.
- Single node Graph Partitioning:
- 1-node, 2x 4th Gen Intel Xeon Scalable processor (pre-production Sapphire Rapids >40cores) on Intel pre-production platform and software with 1024 GB DDR5 memory, microcode 0x90000c0, HT on, Turbo off, Rocky Linux 8.6, 4.18.0-372.26.1.el8_6.crt1.x86_64, 894 GB SSD, 105 TB Luster filesystem with OPA fabric, DGL 0.9.0 random graph partition on single node, test by Intel Corporation on 08/17/2022.

Distributed GNN Training with GPU:

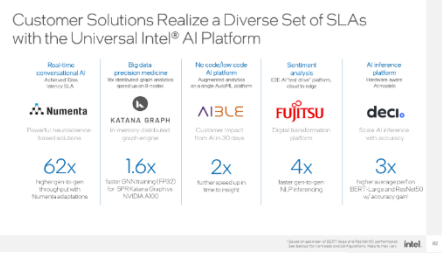
- 8-node, 2x 3rd Gen Intel Xeon Scalable processor with 256 GB DDR4 memory, microcode 0xd000270, HT on, Turbo on, Rocky Linux4.18.0-372.26.1.el8_6.crt1.x86_64, 931 GB SSD, 455 TB Luster filesystem with HDR fabric, 2 A100-PCIE-40GB per node, DGL 0.9, test by Intel Corporation on 09/19/2022.

Customer Solutions
Fujitsu configuration details



- BASELINE(ICX): Tested by Intel as of October 2022. 2 socket Intel(R) Xeon(R) Platinum Ice Lake 8380 CPU @ 2.30GHz Processor(ICX), 40 cores/socket, HT On, Turbo ON, Total Memory 384GB (12slots/32GB/3200 MT/s DDR4), BIOS: SE5C6200.86B.0022.D64.2105220049, ucode 0xd000375, Ubuntu 20.04.5 LTS, 5.4.0-126-generic ,GCC 9.4.0 compiler, Inference Framework: Pytroch 1.12.0, Sentiment analysis in NLP eCommerce Recommender, Topology: HuggingFace :German-Sentiment-Bert model, Multiple streams, Datatype: FP32.
- Config1(ICX): Tested by Intel as of October 2022. 2 socket Intel(R) Xeon(R) Platinum Ice Lake 8380 CPU @ 2.30GHz Processor(ICX), 40 cores/socket, HT On, Turbo ON, Total Memory 384GB (12slots/32GB/3200 MT/s DDR4), BIOS: SE5C6200.86B.0022.D64.2105220049, ucode 0xd000375, Ubuntu 20.04.5 LTS, 5.4.0-126-generic ,GCC 9.4.0 compiler, Inference Framework: OpenVINO 2022.2.0, Sentiment analysis in NLP eCommerce Recommender, Topology: HuggingFace :German-Sentiment-Bert model, Multiple streams, Datatype: FP32.
- BASELINE(SPR): Tested by Intel as of October 2022. 2 socket Intel(R) Xeon(R) Platinum 8480+(SPR), 56 cores/socket, HT On, Turbo ON, Total Memory 512GB(16slots/32GB/4800 MT/s DDR4), BIOS: SE5C6200.86B.0022.D64.2105220049, ucode 0x2b000041, Ubuntu 22.04.1 LTS, 5.15.0-48-generic , GCC 9.4.0 compiler, Inference Framework: Pytorch 1.12.0, Sentiment analysis in NLP eCommerce Recommender, Topology: HuggingFace :German-Sentiment-Bert model, 1 instance/2 socket, Multiple stream, Datatype: FP32.
- Config1(SPR): Tested by Intel as of October 2022. 2 socket Intel(R) Xeon(R) Platinum 8480+(SPR), 56 cores/socket, HT On, Turbo ON, Total Memory 512GB(16slots/32GB/4800 MT/s DDR4), BIOS: SE5C6200.86B.0022.D64.2105220049, ucode 0x2b000041, Ubuntu 22.04.1 LTS, 5.15.0-48-generic , GCC 9.4.0 compiler, Inference Framework: OpenVINO 2022.2.0, Sentiment analysis in NLP eCommerce Recommender, Topology: HuggingFace :German-Sentiment-Bert model, 1 instance/2 socket, Multiple stream, Datatype: FP32.
- OPTIMIZED(Config2:SPR): Tested by Intel as of October 2022. 2 socket Intel(R) Xeon(R) Platinum 8480+(SPR), 56 cores/socket, HT On, Turbo ON, Total Memory 512GB(16slots/32GB/4800 MT/s DDR4), BIOS: SE5C6200.86B.0022.D64.2105220049, ucode 0x2b000041, Ubuntu 22.04.1 LTS, 5.15.0-48-generic , GCC 9.4.0 compiler, Inference Framework: Intel OpenVINO toolkit 2022.2.0, Sentiment analysis in NLP eCommerce Recommender, Topology: HuggingFace :German-Sentiment-Bert model, 1 instance/2 socket, Multiple stream, Datatype: AMX_BF16.

Customer Solutions
Deci.ai configuration details



- ResNet50: Test by Intel as of 11/29/22. 1-node, 2x Intel® Xeon® Platinum 8480+, 56 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 4800 MT/s),BIOS 3A03, ucode 0x2b000021, OS Ubuntu 20.04.5 LTS, kernel 5.15.0-52-generic, ImageNet Benchmark, IPEX==1.13.0, Resnet50, pytorch==1.13.0, intel-openmp==2022.2.1, score 9838 ips @ BS1, 13310 ips @ BS116
- DeciNet: Test by Intel as of 11/29/22. 1-node, 2x Intel® Xeon® Platinum 8480+, 56 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 4800 MT/s),BIOS 3A03, ucode 0x2b000021, OS Ubuntu 20.04.5 LTS, kernel 5.15.0-52-generic, ImageNet Benchmark, IPEX==1.13.0, DeciNet, pytorch==1.13.0, intel-openmp==2022.2.1, score 28998 ips @ BS1, 46288 ips @ BS116
- BERT-Large: Test by Intel as of 11/29/22. 1-node, 2x Intel® Xeon® Platinum 8480+, 56 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 4800 MT/s),BIOS 3A03, ucode 0x2b000021, OS Ubuntu 20.04.5 LTS, kernel 5.15.0-52-generic, SQuADv1.1 Benchmark, IPEX==1.13.0, BERT-Large, sequence length 384, pytorch==1.13.0, intel-openmp==2022.2.1, score 322 ips @ BS1, 380 ips @ BS56
- DeciBERT: Test by Intel as of 11/29/22. 1-node, 2x Intel® Xeon® Platinum 8480+, 56 cores, HT On, Turbo On, Total Memory 512 GB (16 slots/ 32 GB/ 4800 MT/s),BIOS 3A03, ucode 0x2b000021, OS Ubuntu 20.04.5 LTS, kernel 5.15.0-52-generic, SQuADv1.1 Benchmark, IPEX==1.13.0, DeciBERT, sequence length 384, pytorch==1.13.0, intel-openmp==2022.2.1, score 1052 ips @ BS1, 1296 @ BS56


intel