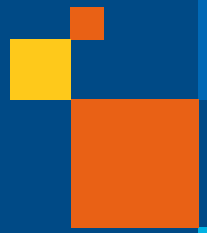
3rd Gen Intel® Xeon® Scalable Processors powers the AI workloads of Alibaba



intel®

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex

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.

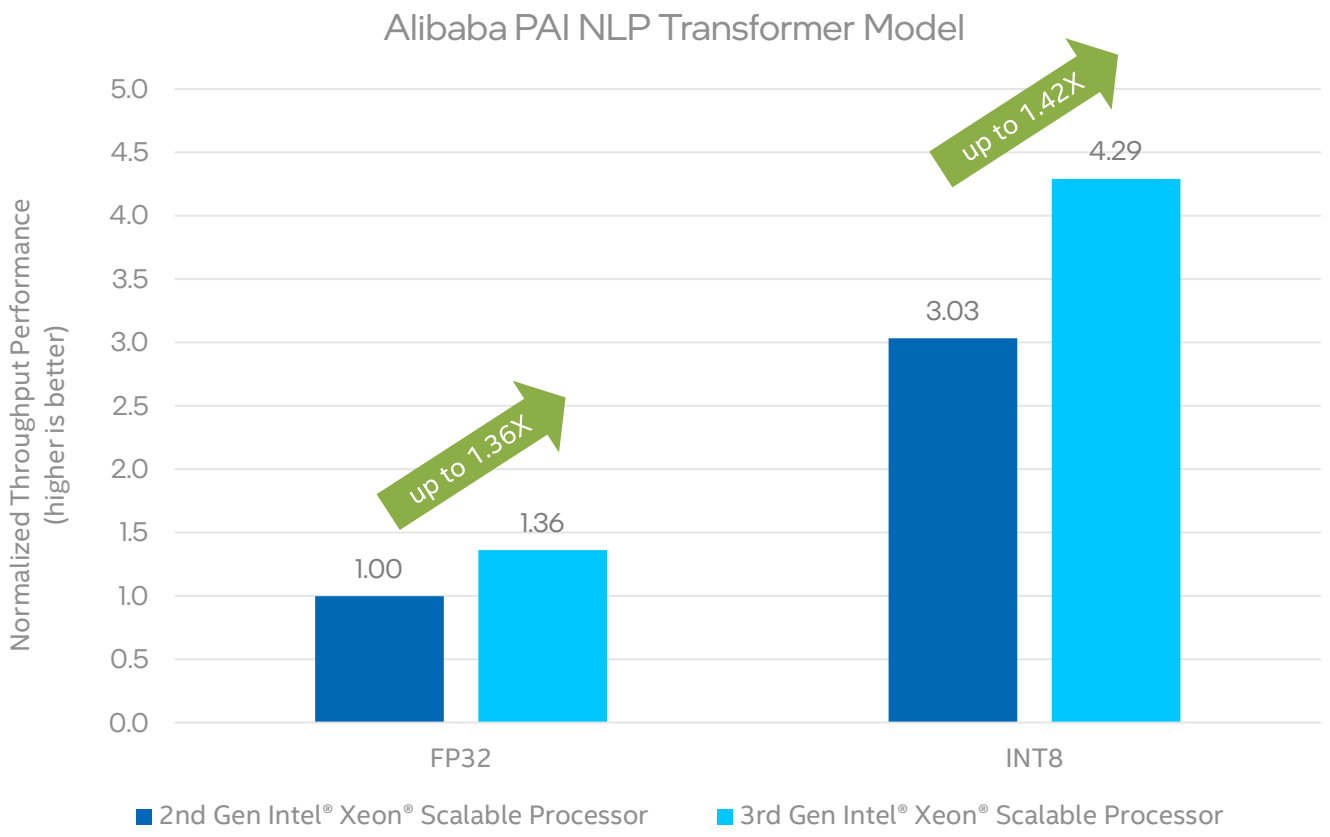
Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

3rd Generation Intel® Xeon® Scalable Processors

Alibaba Cloud PAI Transformer Model

Alibaba Cloud



Application

- Transformer model is a key model in Alibaba Cloud Platform for AI (PAI). It is widely used in natural language processing (NLP) tasks.

Customer Impact

- Unleashes the latest Intel® Xeon® Scalable processor features for Alibaba PAI models, demonstrating Alibaba PAI platform technology leadership.
- It also helps the inference latency on CPU and provide better user experience to Alibaba PAI end-users.

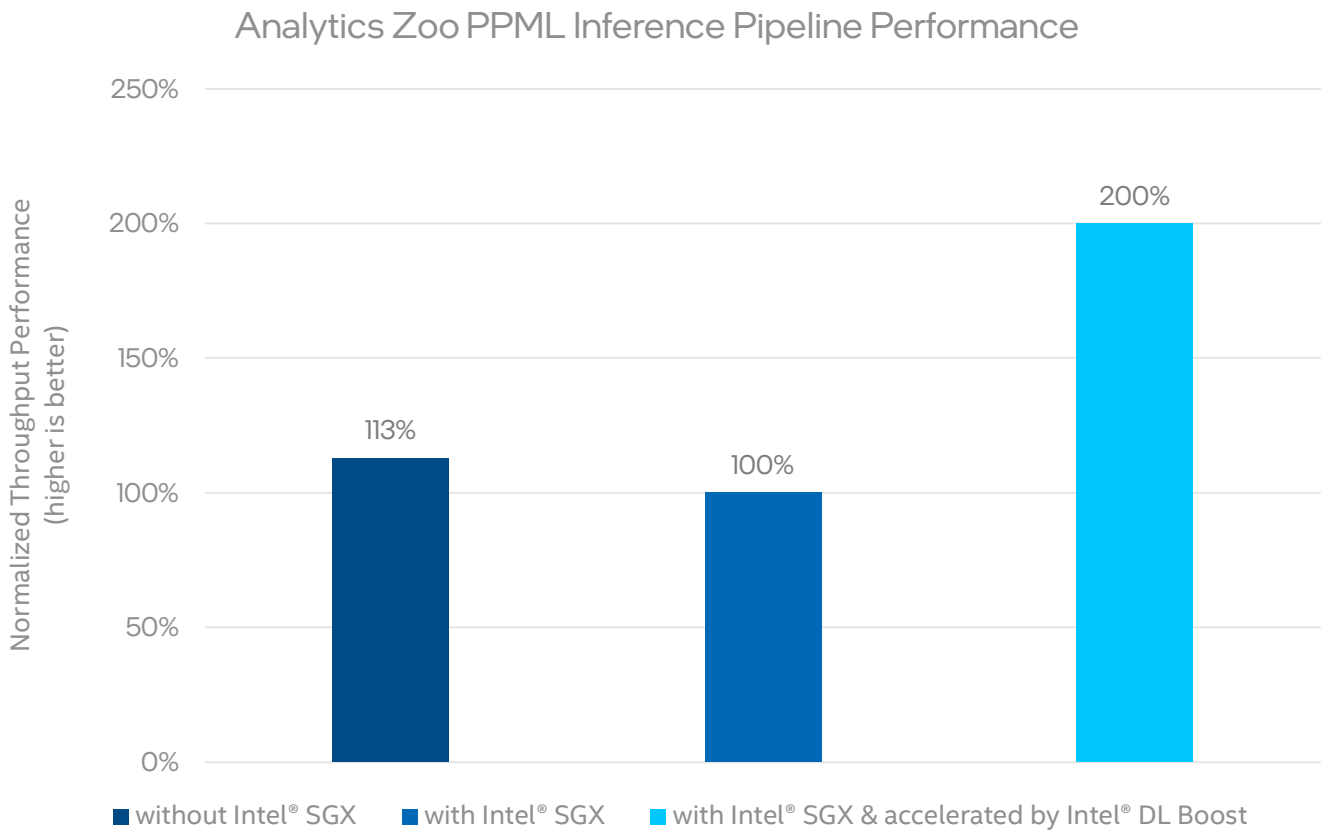
Performance Drivers

- Up to 3X performance using Intel® Deep Learning Boost technology compared to FP32 solution.[1]
- 3rd Gen Intel Xeon Scalable processor **total throughput performance improved up to 36% in FP32 and up to 42%** in INT8 compared to previous generation.[2]

1, 2 – Performance results are based on testing done by Intel March 23, 2021. For complete testing configuration details, see Configurations section. Your costs and results may vary.

3rd Gen Intel® Xeon® Scalable Processors

Ant Group Privacy Preserving Machine Learning



Application

- More secure and distributed inference solution built with Analytics Zoo, better protected by Intel® Software Guard Extensions 2.0 (Intel® SGX) and Occlum* (backed by Ant Group), and accelerated by Intel® Deep Learning Boost (Intel® DL Boost).

Customer Impact

- The end-to-end distributed inference pipeline is more protected by Intel SGX 2.0 and Occlum.
- 2X better inference throughput using Intel DL Boost with Int8 compared to FP32 .[3]

Performance Drivers

- Intel® DL Boost with Int8
- Intel® oneAPI Deep Neural Network Library (oneDNN)

3 – Performance results are based on testing done by Intel March 19, 2021. For complete testing configuration details, see Configurations section.

Configurations

1, 2. Alibaba PAI NLP Transformer Model on PyTorch 1.7.1 Throughput Performance on 3rd Generation Intel® Xeon® Processor Scalable Family

BASELINE: Test by Intel as of 03/19/2021. 2-node, 2x Intel® Xeon® Platinum 8269C Processor, 26 cores HT On Turbo ON Total Memory 192 GB (12 slots/ 16GB/ 2666 MHz), BIOS: SE5C620.86B.02.01.0013.121520200651(ucode: 0x4003003), CentOS 8.3, 4.18.0-240.el8.x86_64, gcc 8.3.1 compiler, Transformer Model, Deep Learning Framework: PyTorch 1.7.1, https://download.pytorch.org/whl/cpu/torch-1.7.1%2Bcpu-cp36-cp36m-linux_x86_64.whl, BS=1, Customer Data, 26 instances/2 sockets, Datatype: FP32/INT8

NEW-1: Test by Intel as of 03/19/2021. 2-node, 2x Intel® Xeon® Platinum 8369B Processor, 32 cores HT On Turbo ON Total Memory 512 GB (16 slots/ 32GB/ 3200 MHz), BIOS: WLYDCRB1.SYS.0020.P92.2103170501 (ucode: 0xd000260), CentOS 8.3, 4.18.0-240.el8.x86_64, gcc 8.3.1 compiler, Transformer Model, Deep Learning Framework: PyTorch 1.7.1, https://download.pytorch.org/whl/cpu/torch-1.7.1%2Bcpu-cp36-cp36m-linux_x86_64.whl, BS=1, Customer Data, 32 instances/2 sockets, Datatype: FP32/INT8

3. Ant Financial Privacy Preserving Machine Learning Performance on 3rd Gen Intel® Xeon® Processor Scalable Family

Test Configuration: Test by Intel Corporation as of 03/20/2021. 2-node, Intel® Xeon® Platinum 8369B Processor, 2 sockets, 32 cores per socket, HT On, Turbo ON, Total Memory 1024 GB (16 slots/ 64GB/ 3200 MHz), EPC 512GB, SGX DCAP Driver 1.36.2, Microcode: 0x8d05a260, Ubuntu 18.04.4 LTS, 4.15.0-112-generic kernel Software Configuration: LibOS Occlum 0.19.1, Flink 1.10.1, Redis 0.6.9, OpenJDK 11.0.10, Python 3.6.9

Workload Configuration: Model: Resnet50, Deep Learning Framework: Analytics Zoo 0.9.0, OpenVINO 2020R2, Dataset: Imagenet, BS=16 per instance, 16 instances/2 socket, Datatype: FP32/INT8

All performance data is tested in lab environment.


intel