Alibaba PAI NLP Transformer Model on 3rd Generation Intel® Xeon® Scalable processors
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.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

© 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.
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 inference latency on CPU and provide better user experience to Alibaba PAI end-users.

Performance Drivers

- ~2X performance speedup 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.
Configurations

**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:** 1-node, 2x Intel Xeon Platinum 8369B processor on Coyote Pass with 512 GB (16 slots/ 32GB/ 3200) total DDR4 memory, ucode 0xd000260, HT on, Turbo on, CentOS Linux 8.3, 4.18.0-240.11.el8_3.x86_64, 1x INTEL 450G SSD, gcc 8.3.1, Transformer Model, 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, test by Intel on 3/19/2021.
