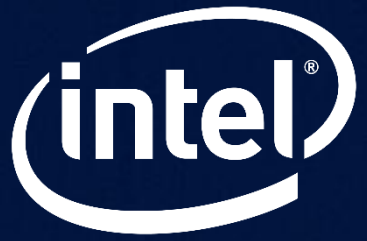


3rd Gen Intel® Xeon® Scalable Processor: Alibaba Proof Points

Notices & Disclaimers

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

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.

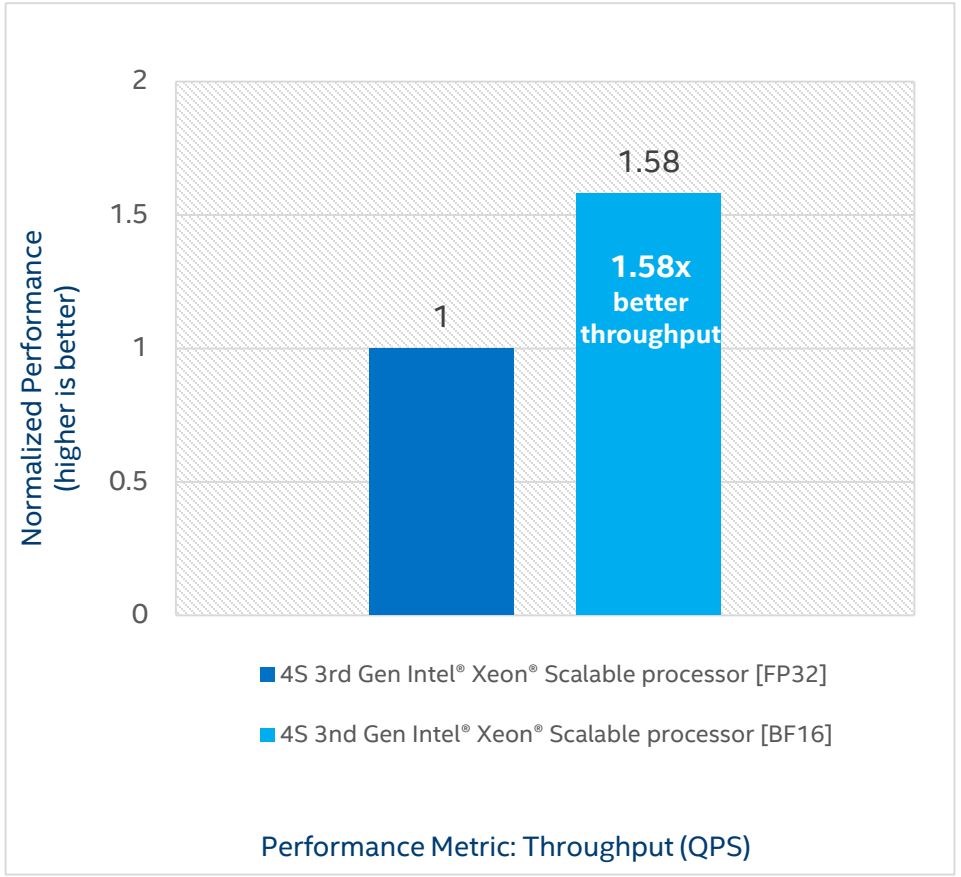
Your costs and results may vary.

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.

3rd Gen Intel® Xeon® Scalable Processor

Alibaba Cloud* PAI (TextCNN)



Application

- TextCNN is one of the key models in Alibaba Cloud Platform for AI (PAI). It is widely used in natural language processing (NLP) applications for different services

Benefit

- Up to **1.58x better throughput** using Intel® Deep Learning Boost (Intel® DL Boost) with bfloat16 compared with FP32 with minimal accuracy loss on Fused TextCNN solution in Alibaba Cloud PAI[1]
- Improved total cost of ownership (TCO) with bfloat16 which delivers better throughput per server

Performance Drivers

- Intel® DL Boost with bfloat16
- oneAPI Deep Neural Network Library (oneDNN) 1.3

At a Glance

Intel® architecture + Adjacencies:
- 3rd Gen Intel® Xeon® Scalable processor

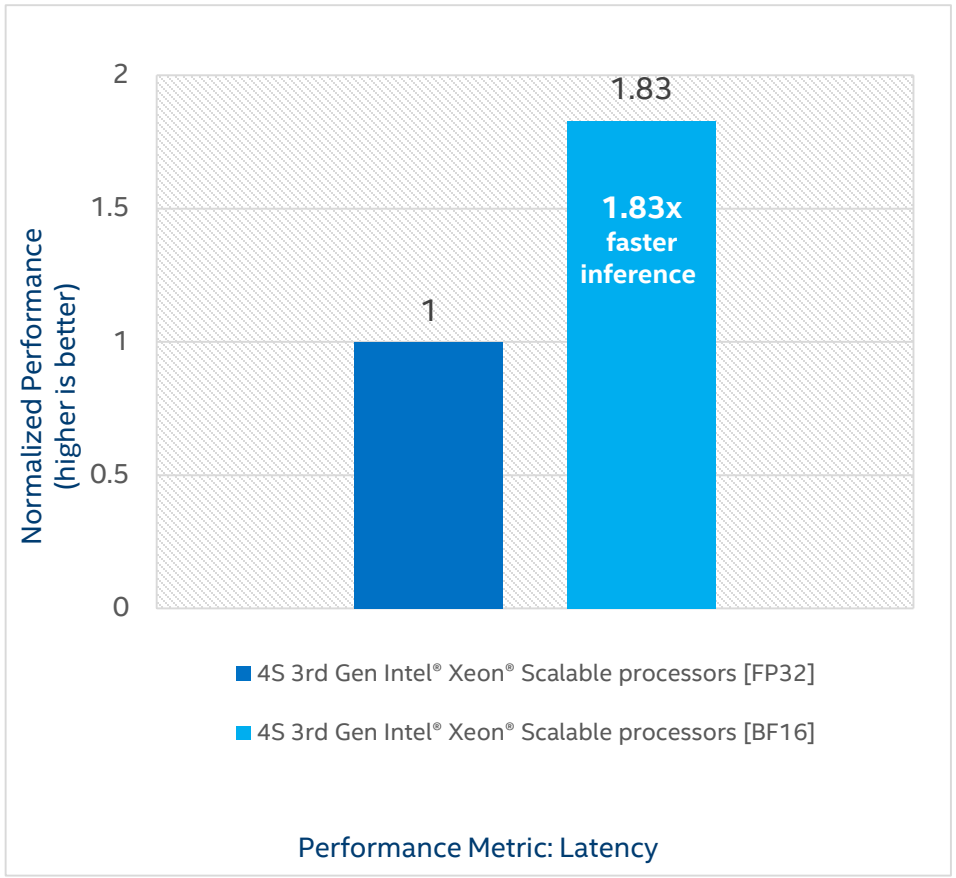
Feature Enabling
- Intel® DL Boost with bfloat16

Intel Software Tools/Libraries
- oneDNN 1.3

1 – Performance results are based on testing done by Intel on April 23, 2020 in a lab environment. Actual deployment plan was being developed by Alibaba Cloud at the time of testing. For complete testing configuration details, see Configuration section.

3rd Gen Intel® Xeon® Scalable Processor

Alibaba Cloud* PAI (BERT)



Application

- BERT is one of key models of the Alibaba Cloud Platform for AI (PAI). It is widely used in natural language processing (NLP) tasks for different artificial intelligence (AI) related business

Benefit

- On Fused BERT BigOP solution, 3rd Gen Intel® Xeon® Scalable processors with Intel® Deep learning Boost (Intel® DL Boost) with bfloat16 helped **improve the customized BERT latency by 1.83x** with similar accuracy compared to FP32 models[1]
- This improved performance enables Alibaba Cloud customers to have a better user experience

Performance Drivers

- Fused 12 layers BERT into one bigOP for both FP32 & BF16
- Over 80% hot function is FP32 SGEMM
- Intel® DL Boost with bfloat16: Replaced SGEMM by BF16 GEMM with oneAPI Deep Neural Network Library (oneDNN) 1.3

At a Glance

Intel® architecture + Adjacencies:
- 3rd Gen Intel® Xeon® Scalable processor

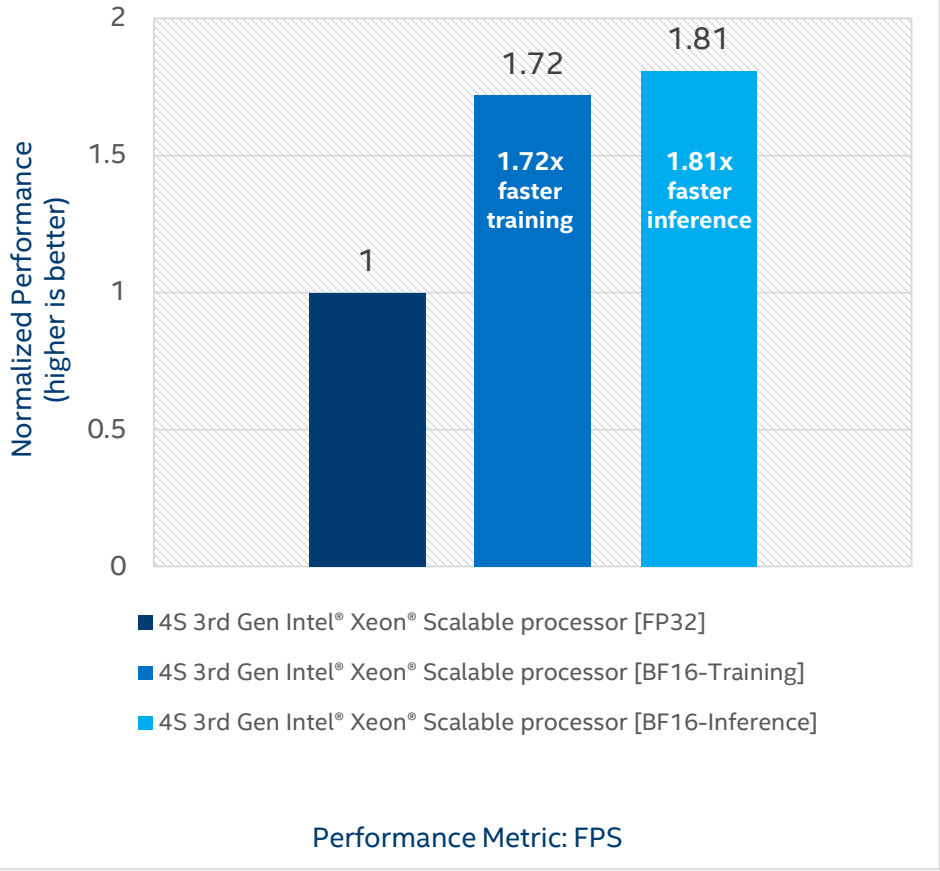
Feature Enabling
- Intel® DL Boost with bfloat16

Intel Software Tools/Libraries
- oneDNN 1.3

1 – Performance results are based on testing done by Intel on April 23, 2020 in a lab environment. Actual deployment plan was being developed by Alibaba Cloud at the time of testing. For complete testing configuration details, see Configuration section.

3rd Gen Intel® Xeon® Scalable Processor

Alibaba Ant Financial (3D-CNN I3D Video)



Application

- 3D CNN model, I3D, is used to analyze video and classify the content into different categories

Benefit

- Up to **1.72x faster training** with bfloat16 (vs FP32) with improvements in FPS and wall-clock time, without any changes to hyper-parameters[1]
- Up to **1.8x faster inference** with blfloat16 than FP32 inference[2]
- Provide better SLA for end-users

Performance Drivers

- No hyper-parameter changes with bfloat16 Training vs. FP32
- 16-bit data format helps with memory bandwidth
- oneAPI Deep Neural Network Library (oneDNN) provides a solid bfloat16 performance foundation

At a Glance

Intel® architecture + Adjacencies:
- 3rd Gen Intel® Xeon® Scalable processor (pre-production)

Feature Enabling
- Intel® DL Boost with bfloat16

Intel Software Tools/Libraries
- oneDNN 1.3

1,2 – Performance results are based on testing done by Intel April 20, 2020. For complete testing configuration details, see Configuration section.

Configurations

AliCloud PAI Customized TextCNN on TF1.14 Throughput Performance on 3rd Gen Intel® Xeon® Scalable Processor:

New: Tested by Intel as of 4/23/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family(Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel Ethernet Controller 10G X550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized TextCNN(Confidential), BS=32, Dummy data, 4 instances/4 socket, Datatype: BF16

Baseline: Tested by Intel as of 4/23/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family(Ali Customized SKU) Processor, using Intel Reference Platform 24 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel Ethernet Controller 10G X550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, MKL version: 2020.1.217, Customized TextCNN(Confidential), BS=32, Dummy data, 4 instances/4 socket, Datatype: FP32

AliCloud PAI Customized BERT on TF1.14 Latency Performance on 3rd Gen Intel® Xeon® Scalable Processor:
New: Tested by Intel as of 4/23/2020. 4 socket Intel® Xeon® Platinum 83xx(Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel ethernet Controller 10G x550T, OS: CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized BERT(Confidential), BS=1, MRPC data, 12 instance/4 socket, Datatype: BF16
Baseline: Tested by Intel as of 4/23/2020. 4 socket Intel® Xeon® Platinum 83xx(Ali Customized SKU) Processor using Intel Reference Platform, 24 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x7000017), Storage: Intel SSDPE2KX010T7, NIC: 2x Intel ethernet Controller 10G x550T, OS:CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: TF1.14 https://pypi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb6be6148ec4f971/intel_tensorflow-1.14.0-cp36-cp36m-manylinux1_x86_64.whl, Compiler: gcc 8.3.1, MKL version: 2020.1.217, Customized BERT(Confidential), BS=1, MRPC data, 12 instance/4 socket, Datatype: FP32

Alibaba Ant Financial Inference and Training on 3rd Gen Intel® Xeon® Scalable Processor:
Tested by Intel as of 4/20/2020. 4 socket 3rd Gen Intel® Xeon® Scalable processor (18-core, 170W, pre-production) Processor using Intel Reference Platform, 18 cores HT OFF, Turbo ON Total Memory 768 GB (24 slots / 32GB / 2666), BIOS Version: 166.08 (6BC51780-BFDE-1000-03E6-000000000000) Microcode: 0x8600000b, CentOS 7.7.1908, 3.10.0-957.el7.x86_64, Deep Learning Framework: Pytorch Intel optimized Pytorch-1.0.0a0+3ca7205 https://gitlab.devtools.intel.com/cce-ai/pytorch, dnnl (mkldnn) commit id:7b53785 https://github.com/oneapi-src/oneDNN, Model: 3d CNN I3D, Compiler: gcc 7.3.1, Libraries: dnnl (mk-dnn), Dataset: UCF101 (size: 13320 shape: 3x64x224x224, Baseline Training: BS=24*4, FP32, New Training: BS=24*4, BF16; Baseline Inference: BS=32, 4 instances/4sockets, FP32, New Inference: BS=32, 4 instances/ 4 sockets, BF16.


intel®