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.

© Intel Corporation. Intel, Xeon, Optane, DL Boost, AVX, the Intel logo, Optane logo, and Xeon logos and other marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.
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
• 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

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.

*Other names and brands may be claimed as the property of others. Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products. Results have been estimated or simulated.
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
• 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

Performance Metric: Latency

Normalized Performance (Higher is better)

- 4S 3rd Gen Intel® Xeon® Scalable processors [FP32]
- 4S 3rd Gen Intel® Xeon® Scalable processors [BF16]

1.83x faster inference

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.

*Other names and brands may be claimed as the property of others. Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products. Results have been estimated or simulated.
### 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
- Up to 1.8x faster **inference** with bfloat16 than FP32 inference
- 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

---

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

*Other names and brands may be claimed as the property of others. Performance results are based on testing as of dates in configuration and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks. Refer to http://software.intel.com/en-us/articles/optimization-notice for more information regarding performance and optimization choices in Intel software products. Results have been estimated or simulated.
Configurations

AliCloud PAI Customized TextCNN on TF1.14 Throughput Performance on 3rd Gen Intel® Xeon® Scalable Processor:
https://pyipi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb66e16148ecf4971/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

https://pyipi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb66e16148ecf4971/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:
https://pyipi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb66e16148ecf4971/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

https://pyipi.tuna.tsinghua.edu.cn/packages/4a/f4/e70311ed73205b12793660641e878810f94fca7d1a9dbb66e16148ecf4971/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: