3rd Gen Intel® Xeon® Scalable Processor: Tencent AI 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.

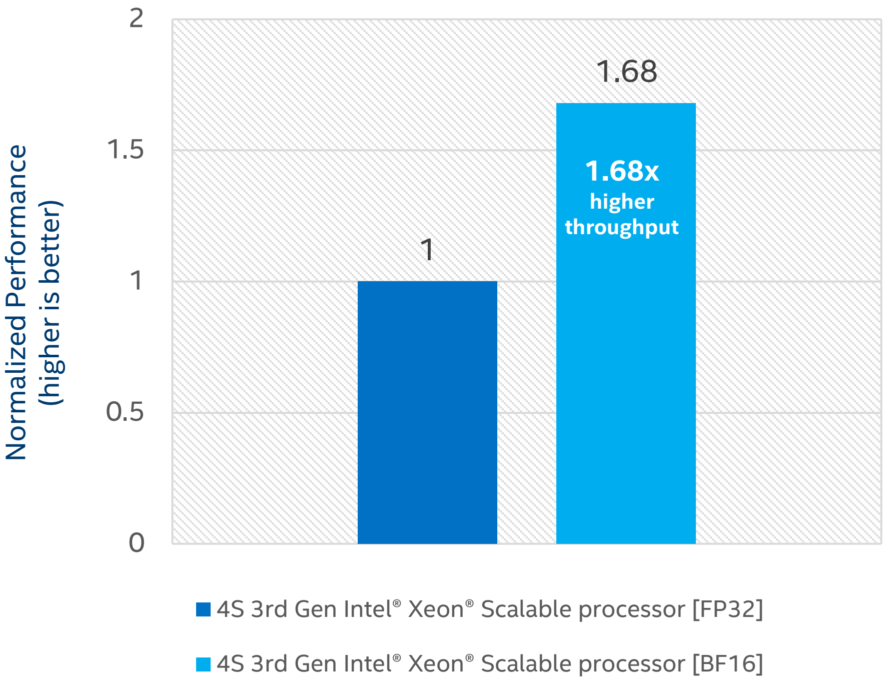
Your costs and results may vary.

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

Tencent* Search Engine



Performance Metric: Throughput (QPS)

Application

- Tencent Search Engine uses a customized Natural Language Processing (NLP) Model with sub 30ms latency requirement for inference

Benefit

- Intel® Deep Learning Boost (Intel® DL Boost) with bfloat16 with oneAPI Deep Neural Network Library (oneDNN) improved NLP throughput by 1.68x with similar accuracy[1]
- Enabled Tencent to increase NLP throughput on CPU, resulting in improved total cost of ownership (TCO) for Tencent search business

Performance Drivers

- Intel® DL Boost with bfloat16
- oneDNN 1.3

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

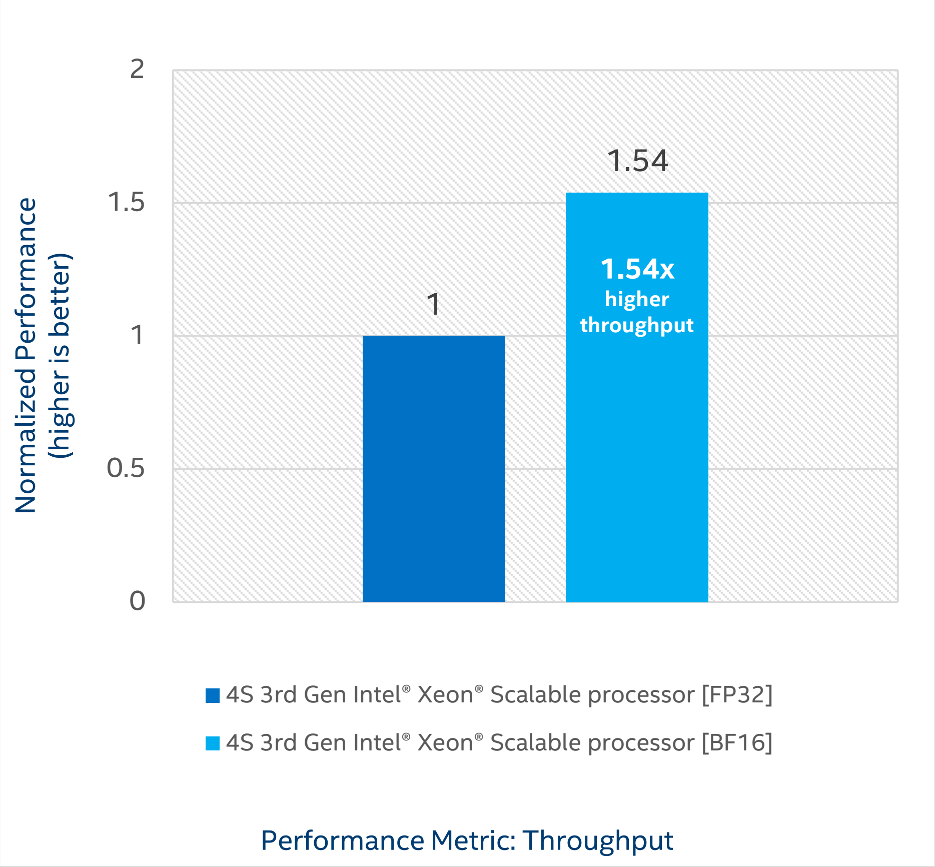
[1] – Performance results are based on testing done by Intel April 28, 2020. No product or component can be absolutely secure. For complete testing configuration details, see Configuration section.

*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

Tencent Cloud

Tencent Cloud Xiaowei* (TTS WaveRNN)



Application

- Tencent Cloud Xiaowei provides key end-to-end AI solution for smart devices (speaker, car, robot, TV, etc.)
- Speech synthesis is one of the key services with pWavenet, WaveRNN, CBHG being top models for text to speech (TTS)

Benefits

- Increased inference throughput on CPU and delivered better experience to Tencent's end users.
- With Operation (OP) fusion, GEMV and cache utilization optimization, Intel® Deep Learning Boost (Intel® DL Boost) with bfloat16 helped improve custom waveRNN throughput by 1.54x at similar accuracy[1]

Performance Drivers

- Intel® DL Boost with bfloat16 to optimize WaveRNN Inference
- Operation (OP) fusion to decrease memory access
- Load balance for sparse GEMV and dense GEMV optimization

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

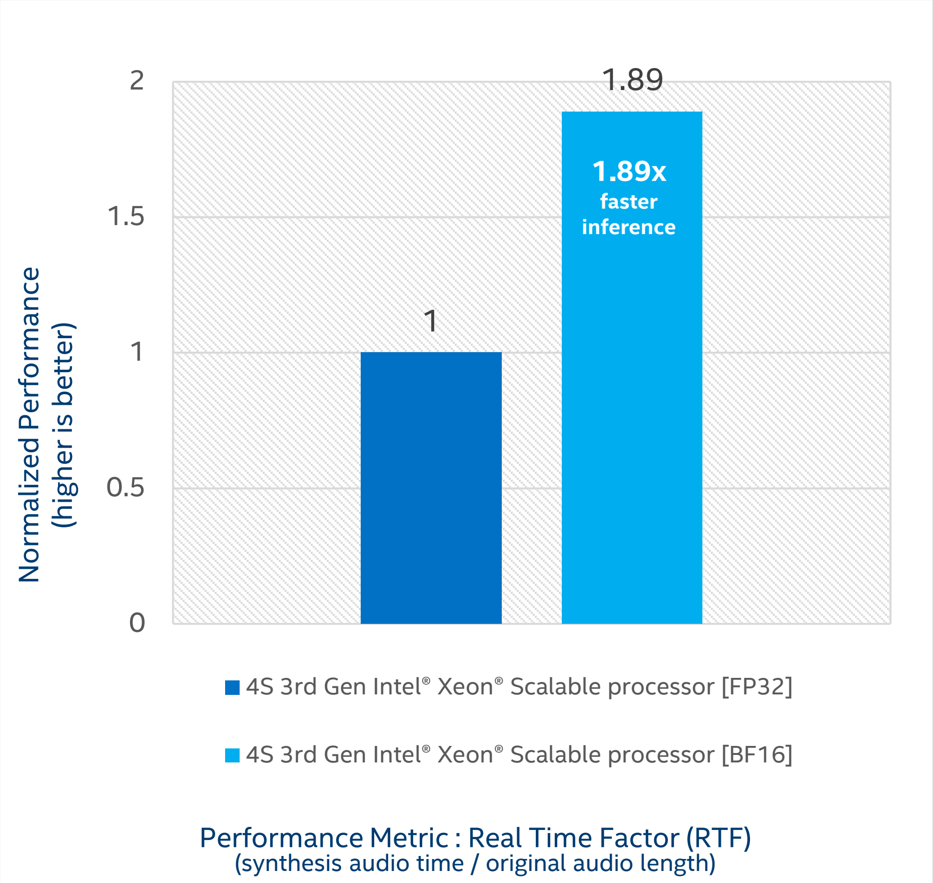
[1] – Performance results are based on testing done by Intel April 28, 2020. No product or component can be absolutely secure. For complete testing configuration details, see Configuration section.

*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

Tencent Cloud

Tencent Cloud Xiaowei* (TTS Parallel WaveNet)



Performance Metric : Real Time Factor (RTF)
(synthesis audio time / original audio length)

Workload Description

- Tencent Cloud Xiaowei provides key end-to-end AI solution to smart devices (speaker, Car, Robot, TV, etc.)
- Speech synthesis is one of the key services with pWavenet, WaveRNN and CBHG being top models for text to speech (TTS)

Potential Customer Impact

- Reduced inference latency on the processor to provide better service experience to Tencent Cloud Xiaowei's end users
- With operation (OP) fusion, pWavenet big OP solution using Intel® Deep Learning Boost (Intel® DL Boost) with bfloat16 helped speed-up custom pWavenet by 1.89x with similar accuracy[1]

Performance Drivers

- Intel® DL Boost with bfloat16
- oneDNN 1.3

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] – Performance results are based on testing done by Intel May 11, 2020. No product or component can be absolutely secure. For complete testing configuration details, see Configuration section.

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 http://www.intel.com/performance.

Configurations

Tencent Search Engine Customized NLP model on TF1.14 Throughput Performance on 3rd Generation Intel® Xeon® Processor Scalable Family:

New: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x86000017), 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 NLP model(Confidential), BS=1, MRPC data, 8 instances/4 socket, Datatype: BF16

Baseline: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x86000017),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 NLP model(Confidential), BS=1, MRPC data, 8 instances/4 socket, Datatype: FP32

Tencent Cloud Xiaowei Customized WaveRNN on MXNetv1.7 Throughput Performance on 3rd Generation Intel® Xeon® Processor Scalable Family:

Opt. BF16 Solution: Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x86000017), CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: MXNet1.7 https://github.com/apache/incubator-mxnet/tree/v1.7.x, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized WaveRNN(Confidential), BS=1 , Customer Provided data, 104 Instances/4 socket, Datatype: BF16

BASELINE(Opt. FP32 Solution): Tested by Intel as of 4/28/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x86000017),CentOS 8.1, 4.18.0-147.5.1.el8_1.x86_64, Deep Learning Framework: MXNet1.7 https://github.com/apache/incubator-mxnet/tree/v1.7.x, Compiler: gcc 8.3.1, oneDNN version: DNNLv1.3, Customized WaveRNN(Confidential), BS=1, Customer Provided data, 104 Instances/4 socket, Datatype: FP32

Tencent Cloud Xiaowei TTS P_Wavenet on TF1.14 Latency Performance on 3rd Generation Intel® Xeon® Processor Scalable Family:

New: Tested by Intel as of 5/11/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots/ 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x86000017), 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 TTS Pwavenet(Confidential), BS=1, Customer Provided data, 4 instances/4 Socket, Datatype: BF16

Baseline: Tested by Intel as of 5/11/2020. 4 socket 3rd Generation Intel® Xeon® Processor Scalable Family (CPX pre-production SKU) Processor, 26 cores HT On Turbo ON Total Memory 384 GB (24 slots / 16GB/ 2933 MHz), BIOS: WCCCPX6.RPB.0018.2020.0410.1316 (ucode:0x86000017),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 TTS Pwavenet(Confidential), BS=1, Customer Provided data, 4 instances/4 Socket, Datatype: Datatype: FP32 intel®