SOLUTION BRIEF

Intel® AI
Intel® Xeon® Scalable Processors

Transforming the Network Edge Enables New Breakthroughs in Near-Real-Time Speech Analytics

Verbio uses network edge servers based on Intel® Xeon® Scalable processors with optimized software to maximize performance and take advantage of reduced latency.

"When moving our speech analytics software to the edge, our customers are looking for a combination of performance and power constraints, and Intel Xeon Scalable processors are providing the right mix, minimizing the latency while delivering the best TCO."

Carlos Puigjaner, CEO, Verbio

The network edge of service providers largely consists of proprietary, fixed-function components with limited compute, memory and storage capacities. As a result, services requiring those IT resources require data to be transmitted across long distances from the customer to data centers, and back again.

The latency penalty introduced by those lengthy roundtrips essentially prevents many high-value services from performing adequately.

Transform the Edge – Unleash New Possibilities

Realizing this, service providers are rapidly transforming the network edge (see Figure 1), replacing proprietary components with powerful, virtualized off-the-shelf servers featuring Intel Xeon Scalable processors and large, high-speed memory and storage capacity to deliver the performance required of today's compute- and data-intensive services.

Intel Xeon Scalable processors are an excellent choice to power servers at the edge because they deliver outstanding performance per watt and per square foot to meet the inherent space and power constraints within edge environments.[1]

These platforms can easily support Network Functions Virtualization (NFV) to bring greater versatility and resource utilization. And they can be seamlessly integrated into a modern, software-defined infrastructure that is self-optimizing, self-healing and adaptive to ever-changing demands.[2,3]

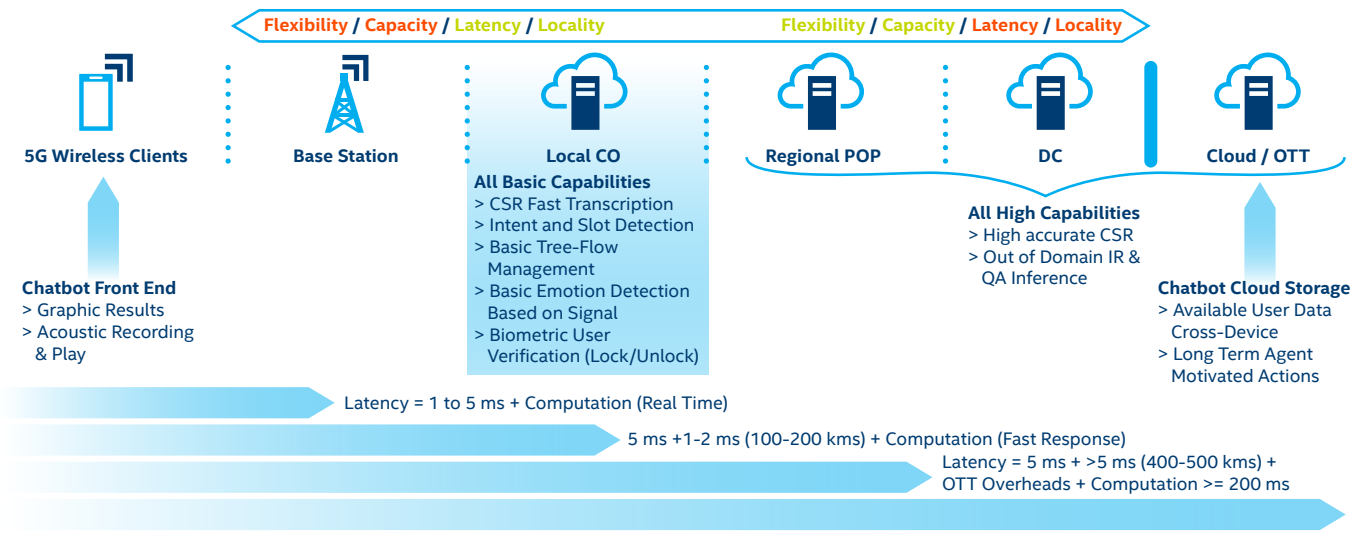


Figure 1. Emerging Network Edge

Network edge locations can become mini data centers, capable of running demanding services, such as analytics, AI, gaming, AR and VR, rich content delivery and more. And by deploying these services at the edge—closer to the customer—latency is greatly reduced, unlocking innovative experiences and possibilities. Network operators win because they can offer new, high-value revenue opportunities. Customers win because they enjoy more immersive experiences and game-changing possibilities.

Verbio and the Power of Advanced Speech Analytics

An example of those game-changing possibilities is an advanced speech analytics deployment from Verbio, a Spanish company that provides deep neural network speech recognition, natural language understanding (NLU) and machine learning technology solutions. There are many potential use cases where Verbio speech analytics software running at the network edge can deliver tremendous benefits, such as next-generation customer call centers, real-time chatbot personal assistants and more. Intel engineering teams from India and Spain worked with the Verbio engineering team in Spain to optimize speech analytics workloads on edge architectures.

Traditionally, speech analytics has been limited to offline, post-conversation processing. This can be helpful for evaluating measures such as call effectiveness after the call, but doesn't permit real-time opportunities.

Recent breakthroughs in AI-based algorithms, including continuous speech recognition (CSR), natural language processing (NLP), speech synthesis or text-to-speech (TTS) and voice biometrics (VB), are now enabling real-time speech analytics. This advancement is made possible through a convergence of hardware performance features, improved algorithms, optimized software and network transformation.

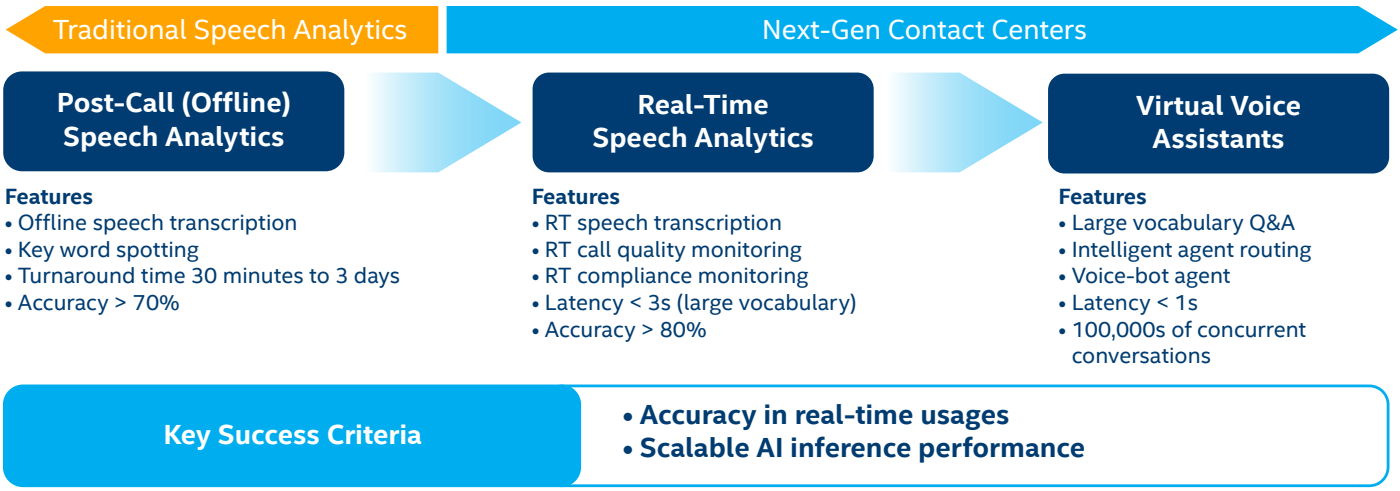


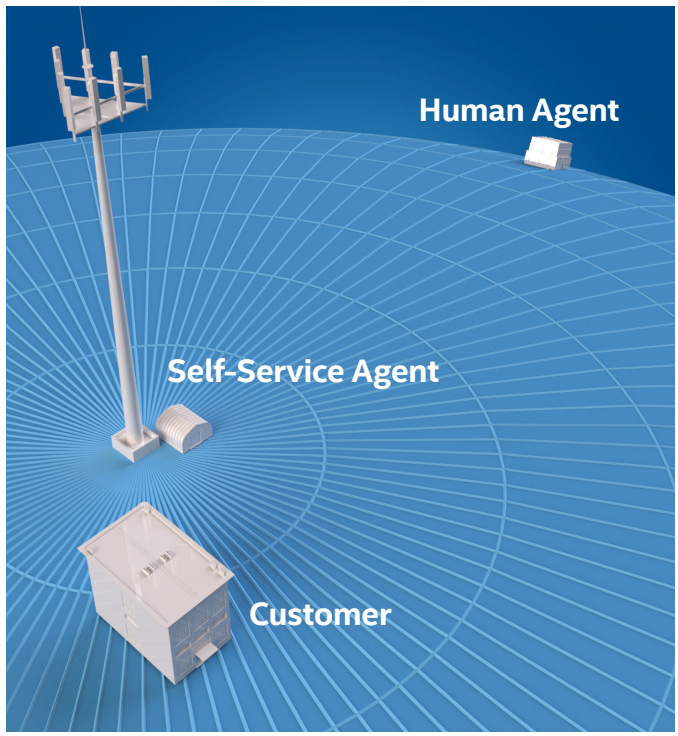
Figure 2. Evolution of contact center speech analytics use cases

Next-Generation Customer Contact Centers

Customer contact and call centers provide important services for countless companies across many industries. Businesses are always trying to deliver better, faster customer service and problem resolution, while improving operational efficiency and reducing costs. Figure 2 illustrates the evolution of contact center speech analytics use cases.

Verbio's speech analytics software employs advanced AI algorithms. In contact centers, the software can be used to determine customers' problems, needs, emotions and other actionable insights. These insights can be used to provide real-time feedback to customer service reps on how to resolve a customer's problem quickly, determine offerings or solutions that might be beneficial to the customer, identify changes in the customer's mood, and take advantage of tips to reduce anger or de-escalate conflict.

With real time virtual voice assistants in next-generation contact centers, businesses can automate more than 60% of total calls previously attended by humans. Calls handled via a virtual assistant can result in significant savings over those same calls handled with live operators. This allows operators to handle more complex, high-value customer service tasks, while also ensuring faster, high-quality service resolutions.



Chatbot Personal Assistants

Chatbot personal assistants go beyond the digital assistants available today in home environments. Today's personal assistants that utilize speech analytics typically require a "wake up" call to the device, often using its product name, followed by a specific question or request.

But what if you had a mobile personal assistant that could help you throughout the day during normal conversations? That's what a chatbot personal assistant can do.

Imagine talking to someone and instantly getting relevant information and links associated with your conversation. Or, imagine your assistant being able to understand the mid-conversation context of your dialogue with a friend or colleague where you discuss meeting at a particular date and time—and the assistant automatically schedules a meeting on your calendar or offers alternate times if you're not available.

Choosing the Right Hardware Architecture

These real-time services require a combination of low latency—made possible by placing those services at the edge, nearer to the customer—and compute performance. The Intel Xeon Scalable processor is a preferred platform for advanced AI and analytics deployments globally, delivering unique performance and scalability benefits to AI-enabled speech analytics workloads. Verbio and Intel collaborated to uncover and implement optimizations to boost performance of these workloads on Intel Xeon Scalable processors.

Optimizing Speech Analytics Performance on Intel Xeon Scalable Processors

Speech analytics includes workloads such as continuous speech recognition (CSR), natural language processing (NLP), text to speech (TTS), voice biometrics (VB) and machine translation (MT). Figure 3 depicts a speech analytics pipeline implementing a virtual voice assistant servicing customer requests via the voice channel.

In these deployments, thousands of concurrent voice channels need to be processed simultaneously without exceeding the 99% latency bound. Speech analytics algorithms need a balance between compute throughput and memory access efficiency and the ability to parallelize the workload across multiple nodes. Our analysis shows that speech analytics workloads stand to gain enormous boosts in performance from systematic analysis of bottlenecks and the implementation of optimizations that leverage the CPU's massive compute and memory resources. Let's take a closer look at these performance drivers.

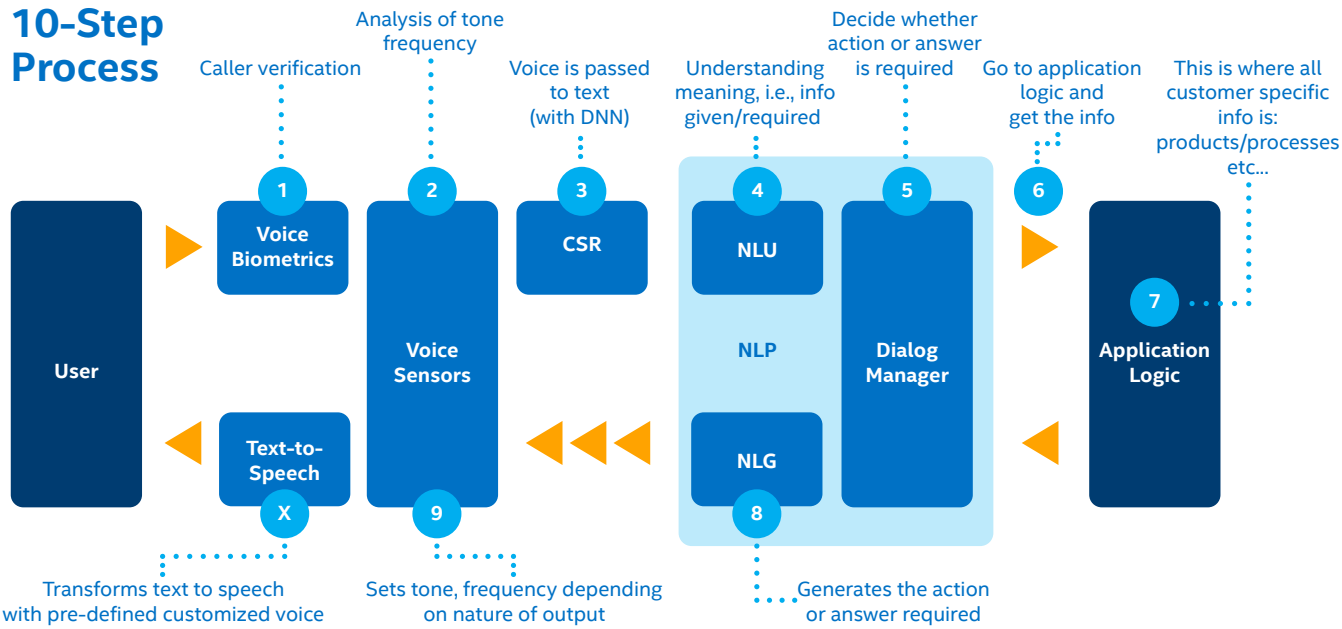


Figure 3. Speech analytics pipeline

Intel-Optimized Software Libraries and Tools

Speech analytics algorithms include deep neural networks (DNNs), signal processing, graph processing, and machine learning functions. Intel offers several optimized software libraries and tools that significantly accelerate these functions, including:

- Intel® Math Kernel Library (Intel® MKL) for BLAS functionality
- Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) for neural network primitives
- Intel® Threading Building Blocks (Intel® TBB) Library for allocation of dynamic memory
- Intel® C++ Compiler tools to speed application performance.

Intel software libraries and tools include a broad range of functions optimized for Intel Xeon Scalable processors. These libraries help improve the overall CPU performance, including memory and compute. Usage of these libraries and tools resulted in a reduction of CPU utilization for CSR workloads by a factor of 4x, from ~90% to ~23%, enabling

a 2.4x improvement in the number of concurrent streams supported (see Figure 4a).

Optimized Topology and Speech Algorithm Building Blocks

Topology and model improvements can deliver a performance boost by optimal configuration of hardware resources for the given operations. Model optimizers help improve the efficient execution of the neural network part of the speech and language solutions. Intel Distribution of OpenVINO Toolkit provides the required model-optimizing compilers, together with an engine to execute the optimized models. Joint investigations have also revealed that batching and data precision choices can offer significant improvement in the arithmetic intensity of these topologies.

Another key component of speech solutions is the decoder. Intel provides optimizations for decoders to boost performance. For example, the Intel® Optimization for TensorFlow* includes an optimization for beam search[4] decoding in CTC-trained neural networks. Intel® Integrated Performance Primitives[5] (Intel® IPP) provide high-performance, platform-aware functions for signal and speech processing. The Intel-optimized decoder solution for the Viterbi algorithm uses weighted finite state transducers (WFST) to combine hidden Markov model (HMM) information and provides a significant performance boost compared to a widely-deployed reference decoder solution.

Further, code vectorization improves in-core parallelism and boosts compute efficiency. Intel® AVX-512 instructions provide a substantial opportunity to boost vectorization of speech analytics workloads. Tools such as Intel® Advisor help developers analyze code to identify optimization opportunities.

Performance Boost Through Workload Scaling

Real-time speech analytics deployments experience large variations in traffic with peak concurrent usages, reaching up to several hundred thousand voice channels. Utilization efficiency of compute cores when scaling the workload is critical to maximize performance (subject to latency bounds). Core and instance scaling are two recommended options (subject to the resource requirement of the workload). The resource requirement of the workload should be managed through a container-based solution that enables deployment across different edge solutions.

A good example of the performance boost due to workload scaling is the Verbio voice biometric use case. Analysis revealed the workload is mostly single threaded, except for the Intel MKL-accelerated portions. As a result, the core identification algorithm was modified to maximize multi-threaded execution. Analysis was performed to understand the resource requirements per instance, which revealed enough headroom on the server to run multiple instances of the biometric engine. When combining these multi-threading and multi-instance opportunities, user identification performance improved by a factor of 10x (see Figure 4b). Usage of thread managers, when coupled with usage of the compute and memory capabilities of the hardware, was primarily responsible for this performance boost. Thread manager tools, such as OpenMP* and Intel® Threading Building Blocks (TBB), can help identify where parallelism can be boosted to increase performance.

The Verbio NLP solution also benefited from a similar approach. Multi-instance helped realize a performance boost of nearly 7x without any code changes (see Figure 4c). The Intel Xeon Scalable processor's large internal caches and high memory bandwidth data access helped realize this multi-threaded and multi-instance approach.

Performance Delivered

Test Configuration

All tests were performed on a single socket Intel Xeon Scalable Platinum 8168 processor. Complete details of the test configuration are provided in Appendix A. Performance improvement achieved by implementing the above recommendations across multiple speech analytics workloads is indicated in Figure 4.

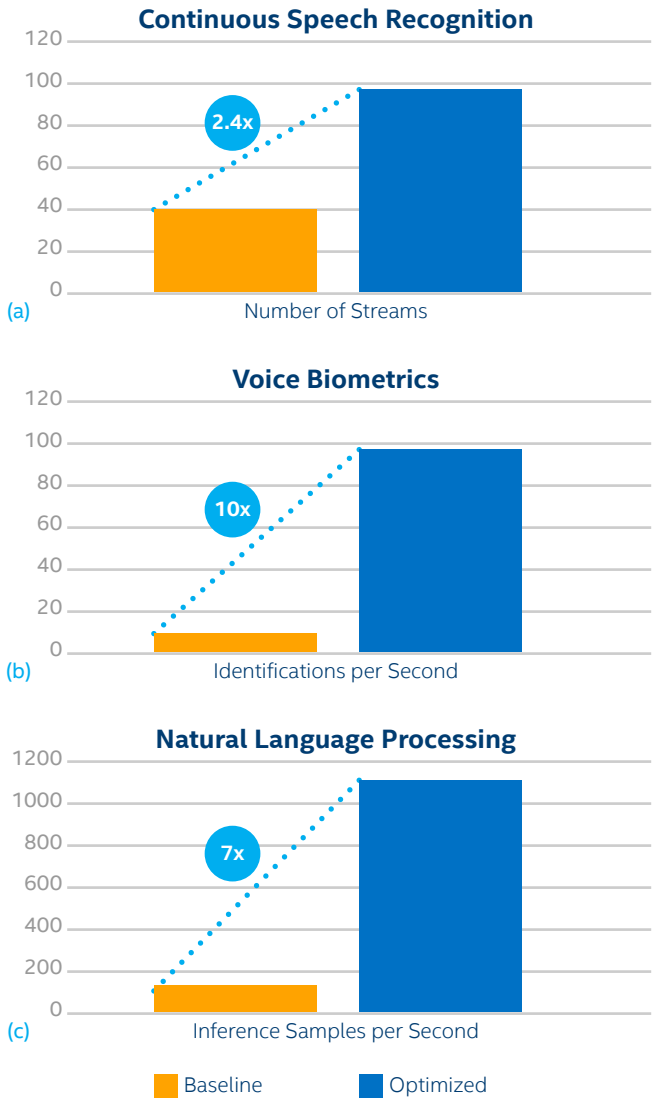


Figure 4. Performance improvement in (a) CSR (b) VB (c) NLP

These results show a clear path to accelerating performance of real-time, AI-enabled speech analytics.

> "The usage of advanced AI algorithms in Speech Recognition, Natural Language Understanding, Text to Speech and Voice Biometrics, combined with the incredible performance of the Xeon Scalable processors, resulted in a winning combination that allows our customers to deploy real time use cases while leveraging their existing hardware architecture."
>
> Carlos Puigjaner, CEO, Verbio

Tap into the Power of Advanced Speech Analytics and Other Opportunities Enabled by AI and the Transformed Network Edge

Real-time performance is non-negotiable in these emerging speech analytics deployments, as delays greatly impair the user experience. Intel Xeon Scalable processors offer unique performance benefits for this class of application, offering incredible performance at peak concurrent usages.

The convergence of AI hardware and software innovation with edge-to-cloud network transformation is driving exciting new possibilities across every industry. Intel is working with developers, solution providers and partners across our vast ecosystem to unleash those possibilities to improve our businesses, our lives and our world.

Learn More:

Intel® Distribution of OpenVINO™ Toolkit

Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN)

Intel® Xeon® Scalable Processors

AI Framework Optimizations from Intel

Speech Recognition Solutions from Verbio

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.

1. https://www.intel.com/content/www/us/en/processors/xeon/scalable/xeon-scalable-platform-brief.html
2. https://builders.intel.com/docs/networkbuilders/creating-the-next-generation-central-office-with-intel-architecture-cpus.pdf
3. https://www.intel.com/content/www/us/en/communications/network-transformation/next-generation-central-office-executive-brief.html
4. https://software.intel.com/en-us/articles/intel-optimization-for-tensorflow-installation-guide
5. https://software.intel.com/en-us/intel-ipp

Appendix A
Configurations: Testing done by Intel Corporation, December 2018.

Application Type	CSR	Biometric	NLP
Benchmark Type	Inferencing	Inferencing	Inferencing
Benchmark Metric	Number of concurrent streams	Number of enrollment/identification per second	Number of Intent detection and slot filling per second
# of Nodes	24+24	24+24	24+24
CPU	Skylake	Skylake	Skylake
Sockets	2S	2S	2S
Processor	8168,24 Cores,205W	8168,24 Cores,205W	8168,24 Cores,205W
Compute Accelerator	NA	NA	NA
CSP Instance	NA	NA	NA
BIO	SE5C620.86B.0D.01.0010.072020182008	SE5C620.86B.0D.01.0010.072020182008	SE5C620.86B.0D.01.0010.072020182008
QDF (Indicate QS / ES2)	NA	NA	NA
Enabled Cores	24 (1 socket)	18 (1 socket)	24 (1 socket)
Platform	Purley	Purley	Purley
Slots	6+6	6+6	6+6
Total Memory	96+96 GB	96+96 GB	96+96 GB
Memory Configuration	2666 MT/s DDR4	2666 MT/s DDR4	2666 MT/s DDR4
Memory Comments	Manufacturer=Micron*	Manufacturer=Micron	Manufacturer=Micron
SSD	INTEL SSDSC2CW240A3, 24S0GB	INTEL SSDSC2CW240A3, 24S0GB	INTEL SSDSC2BA800G3, 800GB
Networking	NA	NA	NA
OS	CentOS* Linux 7, CentOS Linux	CentOS Linux 7, CentOS Linux	Ubuntu* 18.04.1 LTS (GNU/Linux 4.15.0-34-generic x86_64)
OS/Kernel Comments	NA	NA	NA
Other Configurations	NA	NA	NA
HTL	OFF	OFF	OFF
Turbo	ON	ON	ON
Computer Type	Server	Server	Server
Batch Size	24	18	32
Data Version	Verbio dataset	Verbio dataset	Verbio dataset
Data Setup	Data is stored on local storage and read from it	Data is stored on local storage and read from it	Data is stored on local storage and read from it
Storage Abstraction	NA	NA	NA
Object Stores	NA	NA	NA
Data Ingestion	NA	NA	NA
Data Ingestion Software	NA	NA	NA
Data Input Type	Audio	Audio	Text
Complier	18.0.1 20171018	18.0.1 20171018	NA
MKL DNN Library Version	INTEL_MKL_VERSION 20180001	INTEL_MKL_VERSION 20180001	NA
Performance Command	NA	NA	NA
Performance Measurement Knobs	KMP_AFFINITY=granularity=fine, verbose, compact,1,0 and open CPU frequency	KMP_AFFINITY=granularity=fine, verbose, compact,1,0 and open CPU frequency	KMP_AFFINITY=granularity=fine, verbose, compact,1,0 and open CPU frequency
Memory Knobs	numactl -m	numactl -m	numactl -m

Performance results are based on testing as of December 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Optimization Notice: 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.

Notice Revision #20110804

Printed in USA 0119/CVN/ACG/PDF