Solution Brief

Intel® AI

Neusoft and Intel Deliver Swift Results for Pathology Analysis

Support for bfloat16 technology in 3rd Generation Intel® Xeon® Scalable processors yields impressive performance improvements for AI-driven diagnoses.

Neusoft

Chinese solution provider Neusoft joined forces with Intel in early 2020 to improve the speed of artificial intelligence (AI) for medical diagnoses with the help of 3rd Generation Intel Xeon Scalable processors. These advanced processors succeeded in accelerating machine learning (ML) inferencing through their support for the Brain Floating Point 16-bit (bfloat16) numeric format. Thanks to bfloat16, Neusoft is now able to process images used for diagnoses much more quickly, with testing revealing 1.91x faster throughput and 91.8 percent lower latency for a key pathology-analysis application.[1]

Neusoft's CareVault is an intelligent cloud platform for medical research

CareVault Intelligent Medical Research Cloud Platform provides a trusted medical-research platform for AI-based research and development (R&D). Created by Neusoft, it supports collaborative innovation between medical and AI technology professionals. The platform consolidates medical data from many hospitals and public datasets, and it provides an R&D environment with complete lifecycle management for code development, testing, and deployment. CareVault provides tools such as a knowledge-service platform, a data-science platform, a medical data–structured platform, and AI tools. It also offers libraries such as a medical corpus, a medicine knowledge base, diagnosis- and health-knowledge maps, and medical-knowledge maps.

Based in a trusted cloud environment, the CareVault platform can be used to support applications that analyze and diagnose pathologies.

Pathology Analysis application

One CareVault application, named Pathology Analysis, uses digital-photography slides taken of patients to diagnose cancers and other diseases. Pathology Analysis enables medical professionals and hospital researchers to quickly diagnose diseases with the help of CareVault's software tools and capabilities.

Processing pathology slides as fast as possible is important for Neusoft customers, and for this reason Neusoft has been committed to making ongoing improvements in the technology that supports the Pathology Analysis application.

About Neusoft

Neusoft is a global solutions and services provider committed to promoting social development and change through IT innovation. Founded in 1991, Neusoft was China's first listed software company, and it currently employs 20,000 people across 60 cities through its many subsidiaries throughout the world.

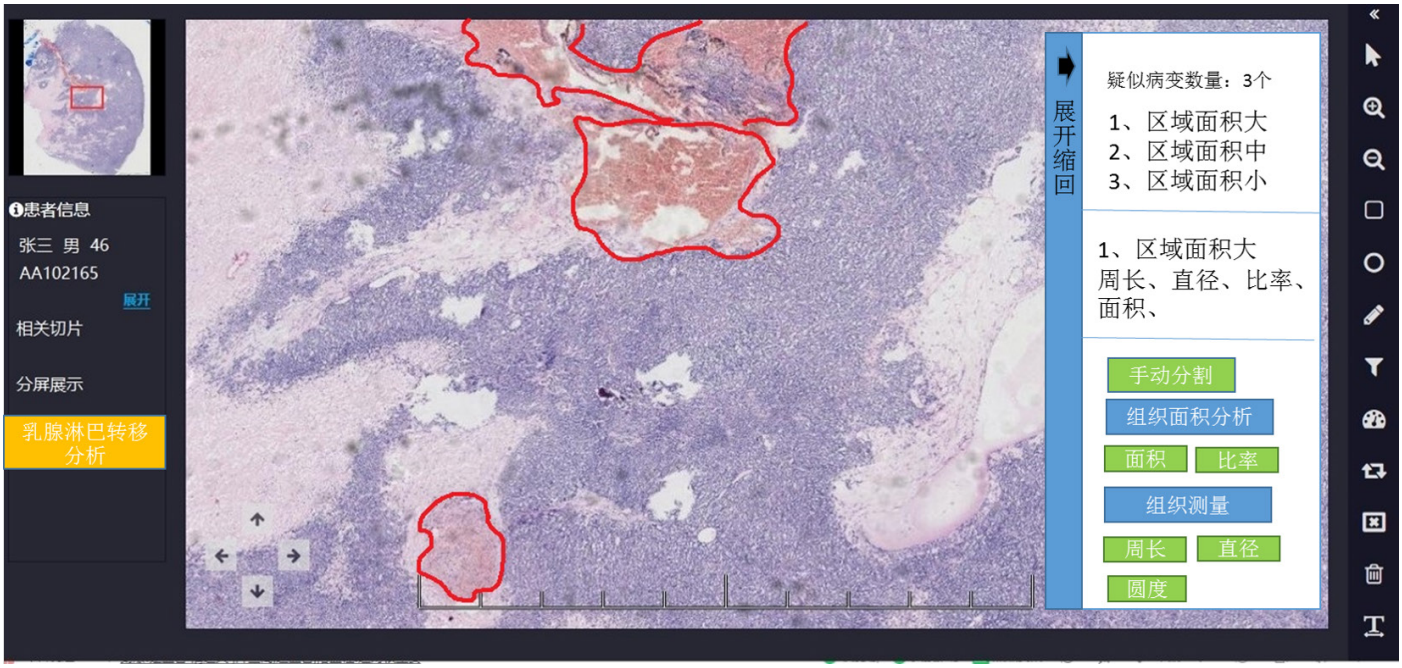

Figure 1. Neusoft's Pathology Analysis application uses digital-photography slides to diagnose cancers and other pathologies on the CareVault platform

3rd Generation Intel Xeon Scalable processors accelerate AI

Neusoft partnered with Intel to upgrade the CPUs used to support the Pathology Analysis application to 3rd Generation Intel Xeon Scalable processors. These advanced processors are built specifically to accelerate AI workloads on the same hardware that's used for non-AI workloads, which spares organizations the need for additional AI infrastructure. A key feature behind the AI acceleration in these newer processors is an update to the Intel Advanced Vector Extensions 512 (Intel AVX-512) instruction set, part of Intel Deep Learning Boost (Intel DL Boost). Called AVX-512_BF16, this update delivers the industry's first x86 support for the bfloat16 numeric format.

The key benefit of bfloat16 is faster performance for workloads, such as ML, that are heavy in floating-point arithmetical calculations. This performance improvement is possible thanks to bfloat16's use of numbers that are much more efficient and lightweight.

Similar structure allows easy transition

Another advantage of bfloat16 is that, despite its use of more compact numbers, its format at the bit level is similar to that of the default numeric format used for floating-point arithmetic, FP32. This overall structural similarity makes it easy to update existing code in support of the newer, faster format. In fact, when Intel worked with Neusoft to build software support for bfloat16 into the CareVault platform and the Pathology Analysis application, the transition was completed seamlessly in a matter of days because so few code changes were needed.

Bfloat16 speeds machine learning

The speed at which a CPU completes mathematical calculations can be heavily affected by the numerical format used in those calculations. Take floating-point arithmetic, for example. Efficiency in performing floating-point computations is especially important in ML because these workloads rely heavily on floating-point math. The native format in x86 processors for floating-point arithmetic is IEEE-754 32-bit (FP32), which specifies more digits for each number than is optimal for deep learning (DL). More digits require more processing, and this extra weight can act as a drag on even the most powerful servers.

Enter bfloat16. Bfloat16 is a 16-bit standard that trades this unneeded precision for large gains in performance. Bfloat16 reserves only seven bits for numerical precision (as opposed to 23 bits for FP32), which helps improve performance in two main ways. First, it helps improve compute resources by allowing data to move faster through the memory hierarchy, which reduces memory-bandwidth bottlenecks. Second, bfloat16 can enable higher operations per second (OPS) because its lower degree of precision takes up less space in the silicon.

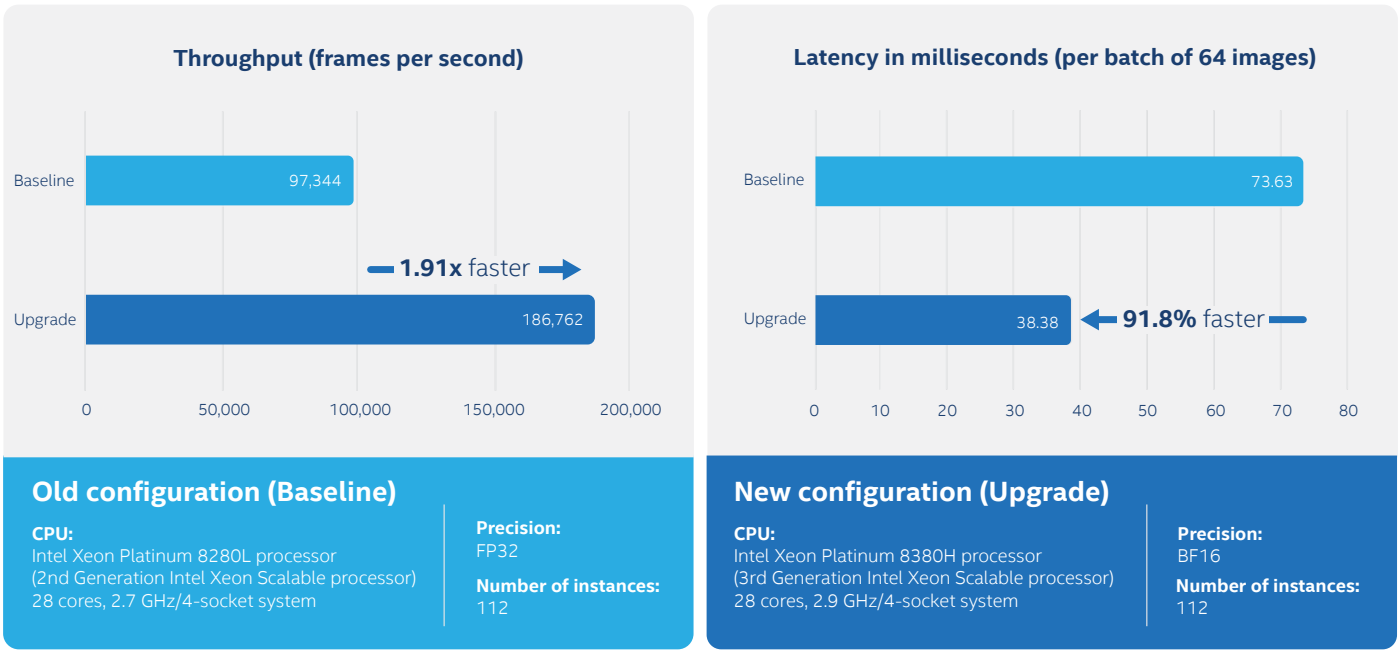

Figure 2. The upgrade to 3rd Generation Intel Xeon Scalable processors allowed the Pathology Analysis application to process almost twice as many frames per second and process slides with significantly less latency

Testing confirms performance improvements

Testing conducted by Intel in a laboratory environment confirmed impressive performance improvements after the upgrade to 3rd Generation Intel Xeon Scalable processors.[1] The testing results and configurations are shown in Figure 2. When the Pathology Analysis application used the bfloat16 numeric format, it was able to process frames with 1.91x faster throughput and 91.8 percent lower latency.

Bottom line: 3rd Generation Intel Xeon Scalable processors speed ML and enable real-world improvements for health services

As the industry's first mainstream server CPUs with built-in bfloat16 support, 3rd Generation Intel Xeon Scalable processors are able to significantly improve ML performance through lighter-weight floating-point calculations. Neusoft's Pathology Analysis application was able to benefit from this feature to quickly improve the speed of its medical diagnoses with only simple changes required to the infrastructure and codebase. Bfloat16 can deliver similar improvements for many types of ML workloads, including those for image classification, speech recognition, and language modeling.[2]

Learn More

Accelerate your AI journey with Intel: **https://intel.com/ai**

Intel DL Boost: **intel.com/content/www/us/en/artificial-intelligence/deep-learning-boost.html**

Neusoft website: **neusoft.com/**

Neusoft Medical Systems website: **neusoftmedical.com/en/**

[1] Configuration details for testing TensorFlow on Neusoft pathology inference throughput performance on Intel Xeon Platinum 8380 processors with bfloat16:

New: Tested by Intel as of May 15, 2020. 4-socket Intel Xeon Platinum 8380 processor, 28 cores, Intel Hyper-Threading Technology (Intel HT Technology) on, Intel Turbo Boost Technology on, 768 GB total memory (24 slots/32 GB/2,933 MHz), BIOS: WLYDCRB1.SYS.0015.D19.2002140555 (ucode: 0x87000016), CentOS 8.1, 4.18.0-147.8.1.el8_1.x86_64, deep learning framework: TensorFlow ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git -b utb 4c711446a4d42fa1ef8759602345fb75f50154ee, compiler: GCC 8.3.1, Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) version: v1.2.0, BS=64, customer's real data, 112 instances/4 sockets, datatype: bfloat16.

Baseline: Tested by Intel as of May 16, 2020. 4-socket Intel Xeon Platinum 8280 processor, 28 cores, Intel HT Technology on, Intel Turbo Boost Technology on, 768 GB total memory (24 slots/32 GB/2,933 MHz), BIOS: 4.1.10 (ucode: 0x4000024), CentOS 8.1, 4.18.0-147.8.1.el8_1.x86_64, deep learning framework: TensorFlow ssh://git@gitlab.devtools.intel.com:29418/TensorFlow/Direct-Optimization/private-tensorflow.git -b utb 4c711446a4d42fa1ef8759602345fb75f50154ee, compiler: GCC 8.3.1, Intel MKL-DNN version: v1.2.0, BS=64, customer's real data, 112 instances/4 sockets, datatype: FP32.

[2] Intel. "3rd Generation Intel® Xeon® Scalable Processors - Claims and Benchmark Library." https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/.

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.

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.

Printed in USA 1020/RG/PRW/PDF Please Recycle 343949-001US