SOLUTION BRIEF

AI Performance

Intel® Xeon® Scalable Processor

Company Logo Goes Here

STRONG AI PERFORMANCE IN THE CLOUD

Build strong AI performance in the cloud with 3rd Generation Intel® Xeon® Scalable Processors, the foundation for artificial intelligence, machine learning, and deep learning. Intel continues to focus on accelerating the entire data pipeline to optimize and scale the cloud journey.

OPTIMIZED FOR AI

Hardware innovation and software optimizations drive AI performance gains on Intel® Xeon® Scalable processors.

up to **10–100x**
faster with Intel-optimized versions over default TensorFlow (image recognition)/scikit-learn (SVC and kNN predict)[1]

up to **74%**
faster from gen on gen (natural language processing)[2]

up to **25x**
faster than AMD EPYC 7763 (object detection)[3]

up to **1.5x**
higher performance than AMD EPYC 7763 (Milan) across 20 key customer AI workloads[4]

up to **1.3x**
higher performance than NVIDIA A100 across 20 key customer AI workloads[5]

Our Offering

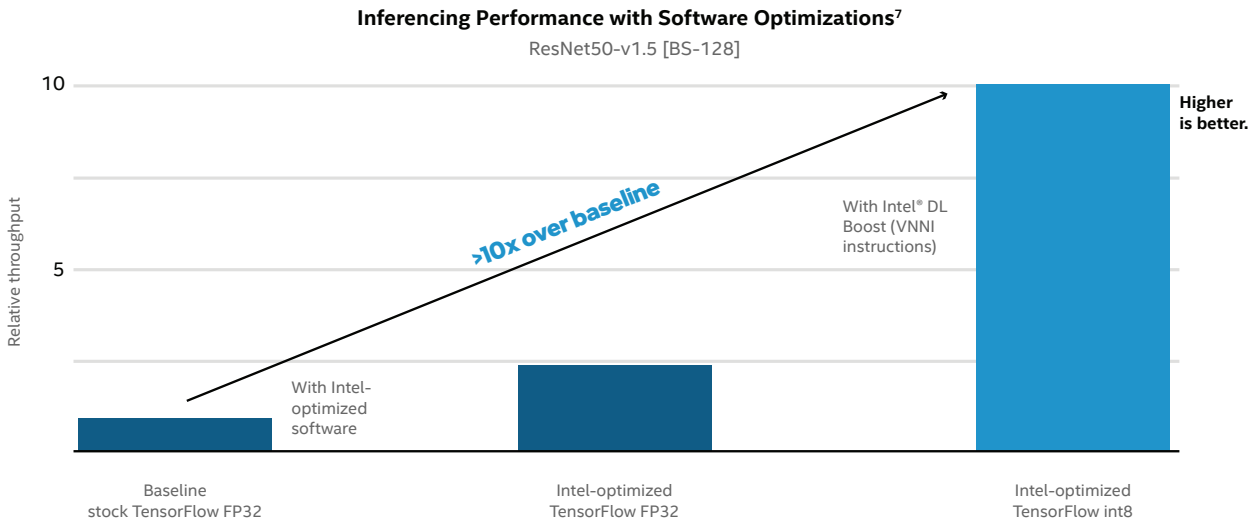
At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus. Ulluptaspit harias et voluptaquam aspedic iendercit mos raesequia nem int eicienia vent ut labo. Tia qui qui dolore non escit volor sa esciisqui doloreptur, nonsendi quatur sunt unt, ommoles etur mo omnimincidBus atet vidunt eate consequi net fugia quostius, si culparum doloressequi te seque velit hariam voluptati aut deris simi, qui ipitem dolorum dia ipsum et aut hitiuntectur alicia quibus maio min con perrovid eosandae sed eicipid ut.

To unlock the full power of your cloud platform, [company or product name] AI solution developers most often start with software-defined AI accelerators on Intel® Xeon® Scalable processors. Other providers may promise huge performance gains with hardware accelerators, but we choose software-based accelerators that can deliver up to 100x performance gains on machine learning (ML) workloads with Intel® Distribution for Python[6] to enhance the power of your cloud platform at little or no additional cost.

For deep learning (DL) workloads, 3rd Generation Intel Xeon Scalable processors can achieve a 10x or greater improvement with Intel® DL Boost and optimized software.

Comparing the performance of competitive offerings can be complicated because most data scientists run a variety of AI workloads. Intel tested its platforms on a range of machine and deep learning models, including training and inference, that were deemed popular by more than 2,500 data scientists who participated in a recent **Kaggle survey**.

The tests showed that Intel Xeon Scalable processor-based platforms outperformed both AMD EPYC 7763 (code-named Milan) and NVIDIA A100 GPUs on most workloads. The Intel platforms also outperformed those competitors across the geo mean of those key customer workloads. Those results persuaded us to choose Intel® Xeon Scalable processors for our customers, who reap the benefits of high performance and lower total cost of ownership (TCO). We also recommend and specify Intel technology-based instances from our cloud services providers and yours, to unlock AI performance and scalability in the cloud.



A day in the life of a data scientist

Data scientists run programs, but they don't sit around waiting for those programs to resolve. They ingest, analyze, and test their data to improve the accuracy of their models and strategies. These iterative processes take a lot of time, which is measured in hours and days. In that context, a few microseconds of runtime are hardly relevant.

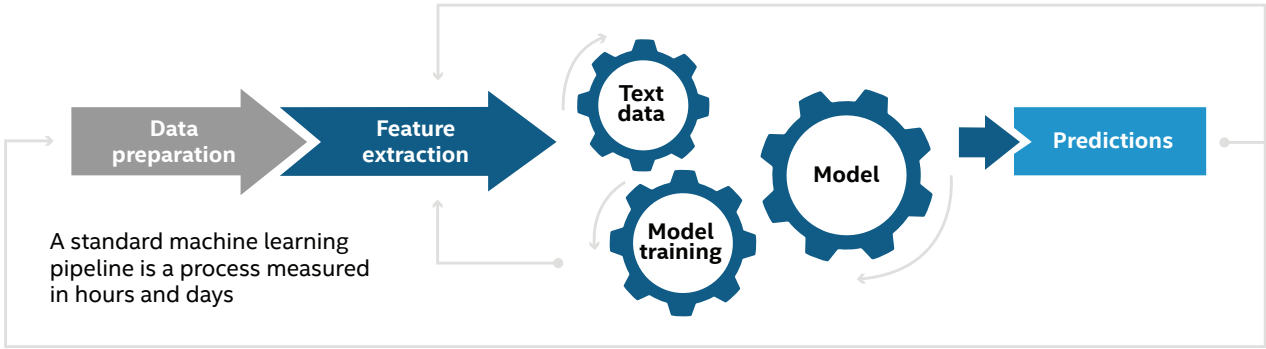
When our developers [at company name] design solutions for data scientists, they consider the entire data pipeline, not just the processing speed for one aspect of the job. A common ML pipeline reveals that data scientists spend most of their time in iterative processes. While some integrators insist that GPUs are required to handle machine learning and similar workflows, the additional hardware may not confer significant benefits to data scientists performing real-world tasks.

Continuous improvement

To compare the performance of different solutions in an end-to-end (E2E) ML pipeline, Intel tested real E2E workflows that perform the following tasks:

- Read data from a large data set (example: Readcsv in the chart below)
- Iterate on the data multiple times to create a model (ETL and model training)
- Run predictions on the model (ML time)

The results: 3rd Generation Intel Xeon Scalable processors deliver 25 percent faster E2E data science at all phases of the pipeline compared to 2nd Generation Intel® Xeon® Scalable processors.[8] That is a significant improvement that gives us tremendous confidence in designing and recommending solutions based on Intel technology, on-premises or in the cloud.

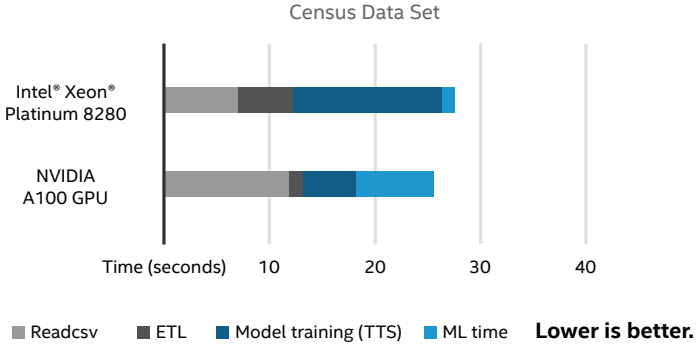


Testing also revealed that 3rd Generation Intel Xeon Scalable processors deliver competitive performance when compared to popular GPUs for representative E2E workloads. Despite the additional hardware cost, the difference in completion time was insignificant: less than the average time between eyeblinks.[9]

CONSIDER THIS

3rd Generation Intel® Xeon® Scalable processors deliver competitive performance without the additional cost and complexity of switching to a GPU platform.

End-to-End Machine Learning Performance[10]



High performance where it counts

GET THE PERFORMANCE YOU NEED BY OPTIMIZING ON THE INTEL® XEON® SCALABLE PROCESSOR HARDWARE YOU KNOW AND TRUST.

GE Healthcare

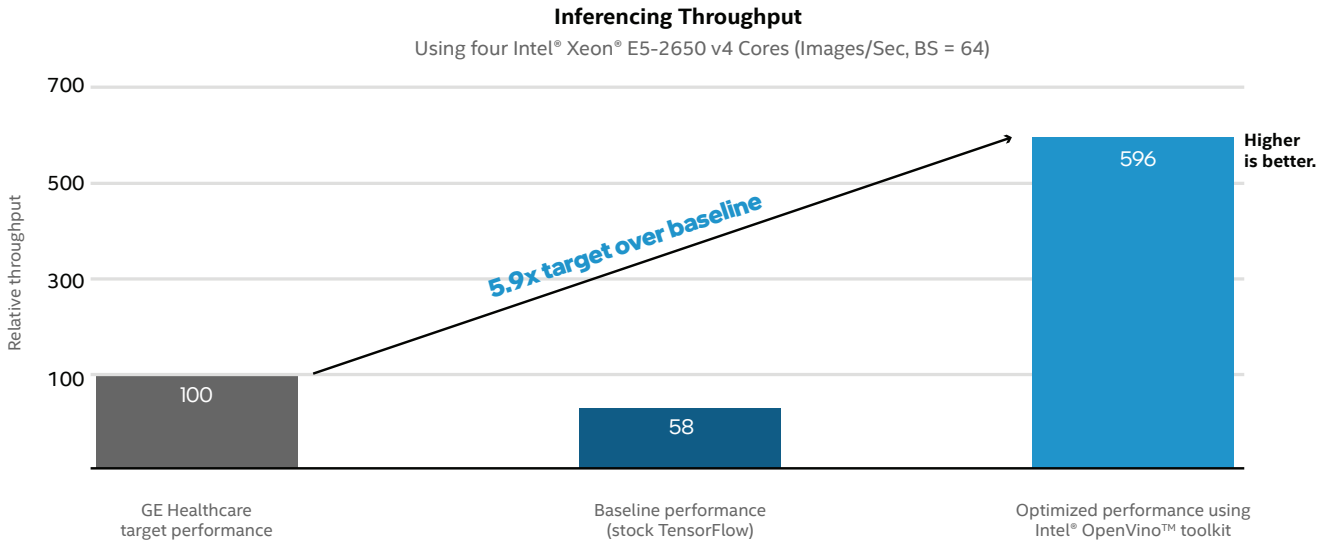
GE Healthcare

GE Healthcare designs and manufactures advanced medical equipment, including computerized tomography (CT) scanners. GE engineers needed an inferencing solution that was fast and powerful enough to keep pace with the needs of a busy hospital's radiology department. A single CT scanner can produce 100 images or more per second, and the inferencing solution must analyze those images rapidly to help radiologists identify indicators of cancer or other diseases.

GE also wanted the new inferencing solution to be deployable on a variety of the company's CT scanner models and to run in a customer's data center or in the cloud, to accommodate IT and OT infrastructure and requirements in hospitals throughout the world.

Luckily, GE's CT machines already had four underutilized or unused Intel® Xeon® processor cores. In collaboration with the Intel team, and using Intel® Distribution of OpenVINO™ toolkit, GE developed an optimized, software-defined solution that ran on their existing hardware. The new, Intel® Xeon® Scalable processor-based solution achieved a 14x speed increase compared to GE's baseline and a 5.9x acceleration compared to the company's inferencing targets.[11]

Details of the GE solution are presented in a **white paper** about medical imaging optimizations assisted by Intel software tools.



SOLUTION BRIEF

AI Performance

Intel® Xeon® Scalable Processor

BUILD SMARTER, FASTER, STRONGER CLOUDS FOR A BETTER FUTURE.

About [Company Name]

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus.

Contact us to get started.

Company Logo Goes Here

intel®

COMPANYNAME.COM

Contact info goes here

1. See [117] at **intel.com/3gen-xeon-config**. Results may vary.
2. See [123] at **intel.com/3gen-xeon-config**. Results may vary.
3. See [45] at **intel.com/3gen-xeon-config**. Results may vary.
4. See [43] at **intel.com/3gen-xeon-config**. Results may vary.
5. See [44] at **intel.com/3gen-xeon-config**. Results may vary.
6. Intel® Distribution for Python is available to optimize performance for all Intel data center CPUs.
7. See [118] at www.intel.com/3gen-xeon-config. Results may vary.
8. Hardware configuration for Intel® Xeon® Platinum 8380: 1-node, 2x Intel® Xeon® Platinum 8380 (40C/2.3 GHz, 270W TDP) processor on Intel® Software Development Platform with 512 GB (16 slots/32 GB/3200) total DDR4 memory, ucode X55260, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 2x Intel® SSD D3-S4610 Series. Hardware configuration for Intel® Xeon® Platinum 8280: 1-node, 2x Intel® Xeon® Platinum 8280L processor on Intel® Software Development Platform (28C) with 384GB (12 slots/32 GB/2933 MHz) total DDR4 memory, ucode 0x4003003, HT on, turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 2x Intel® SSD DC S3520 Series. Software: Python 3.7.9, preprocessing Modin 0.8.3, OmniSciDB v5.4.1, Intel-optimized scikit-learn 0.24.1, OneDAL Daal4py 2021.2, XGBoost 1.3.3, data set source: IPUMS USA: **usa.ipums.org/usa/**, data set (size, shape): (21721922, 45), data types int64 and float64, data set size on disk 362.07 MB, data set format .csv.gz, accuracy metric MSE: mean squared error; COD: coefficient of determination, tested by Intel, and results as of March 2021.
9. Source: **ncbi.nlm.nih.gov/pmc/articles/PMC4043155/**
10. NVIDIA A100 is 1.9 seconds faster than 3rd Gen Intel® Xeon® Scalable processor supporting Intel® DL Boost on Census end-to-end machine learning performance. Hardware configuration for Intel® Xeon® Platinum 8380: 1-node, 2x Intel Xeon Platinum 8380 (40C/2.3 GHz, 270W TDP) processor on Intel® Software Development Platform with 512 GB (16 slots/32 GB/3200) total DDR4 memory, ucode X55260, HT on, Turbo on, Ubuntu 20.04 LTS, 5.4.0-65-generic, 4x Intel® SSD D3-S4610 Series, tested by Intel, and results as of March 2021. Hardware configuration for NVIDIA A100: 1-node, 2-socket AMD EPYC 7742 (64C) with 512 GB (16 slots/32 GB/3200) total DDR4 memory, ucode 0x8301034, HT on, turbo on, Ubuntu 18.04.5 LTS, 5.4.0-42-generic, NVIDIA A100 (DGX-A100), 1.92 TB M.2 NVMe, 1.92TB M.2 NVMe RAID. Software configuration for Intel® Xeon® Platinum 8380: Python 3.7.9, pre-processing Modin 0.8.3, OmniSciDB v5.4.1, Intel-optimized scikit-learn 0.24.1, OneDAL Daal4py 2021.2, XGBoost 1.3.3. Software configuration for NVIDIA A100: Python 3.7.9, pre-processing CuDF 0.17, Intel Optimized scikit-learn 0.24, OneDAL CuML 0.17, XGBoost 1.3.0dev.rapidsai0.17, NVIDIA RAPIDS 0.17, CUDA toolkit CUDA 11.0.221. data set source: IPUMS USA: **usa.ipums.org/usa/**, data set (size, shape): (21721922, 45), data types int64 and float64, data set size on disk 362.07 MB, data set format .csv.gz, accuracy metric MSE: mean squared error; COD: coefficient of determination, tested by Intel, and results as of March 2021.
11. Configuration: 2-socket Intel® Xeon® E5-2650 v4 processor 24 cores HT OFF, Total Memory 256 GB (16x 16 GB/2133 MHz), Linux-3.10.0-693.21.1.el7.x86_64-x86_64-with-redhat-7.5-Maipo, BIOS: SE5C610.86B.01.01.0024.021320181901, Intel® Deep Learning Deployment Toolkit version 2018.1.249, Intel® MKL-DNN version 0.14. Patch disclaimer: Performance results are based on testing as of June 15, 2018 and may not reflect all publicly available security updates. No product can be absolutely secure.

Performance varies by use, configuration, and other factors. Learn more at **intel.com/PerformanceIndex**. 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 is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's **Global Human Rights Principles**. Intel® products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

Intel does not control or audit third-party data. You should review this content, consult other sources, and confirm whether referenced data is accurate.

Intel technologies may require enabled hardware, software, or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.

1021/ADS/CMD/PDF