Accelerate Your AI Today

on 3rd Generation Intel® Xeon® Scalable Processors

up to **10-100x**
faster with Intel-optimized versions over default TensorFlow (image recognition) / Scikit-Learn (SVC & 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 perf 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]

How Do We Do It?

Continued hardware innovation and software optimizations drive AI performance gains on Intel® Xeon® Scalable Processors.

Intel® Xeon® Scalable Processors continue to be the foundation for artificial intelligence, machine learning, and deep learning, and Intel continues to focus on speeding up the ENTIRE data pipeline, not just accelerating small chunks of it.

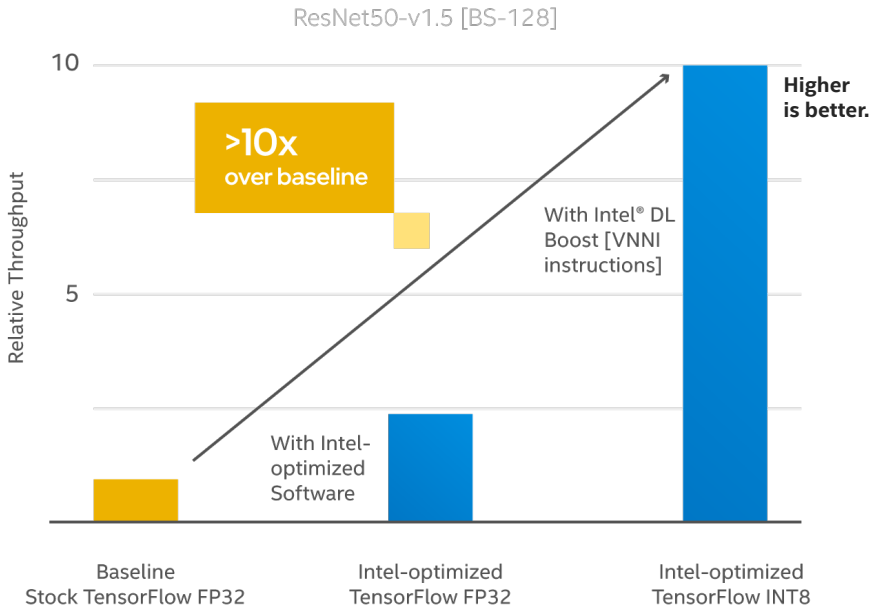
There are many startups creating AI hardware accelerators promising huge performance gains, but did you know that you can get "software AI accelerator" on Intel® Xeon® Scalable Processors that can deliver up to 100x performance gains **on machine learning (ML) workloads** by simply installing our ***free*** Intel® Distribution for Python.[6]

For **deep learning (DL) workloads,** 3rd Generation Intel® Xeon® Scalable Processors are showing a greater than 10x improvement with Intel® DL Boost and optimized software!

But how does that compare against the competition? Since most data scientists don't run a single AI workload, we looked at a recent Kaggle survey and chose a broad range of popular machine and deep learning models, including training and inference.

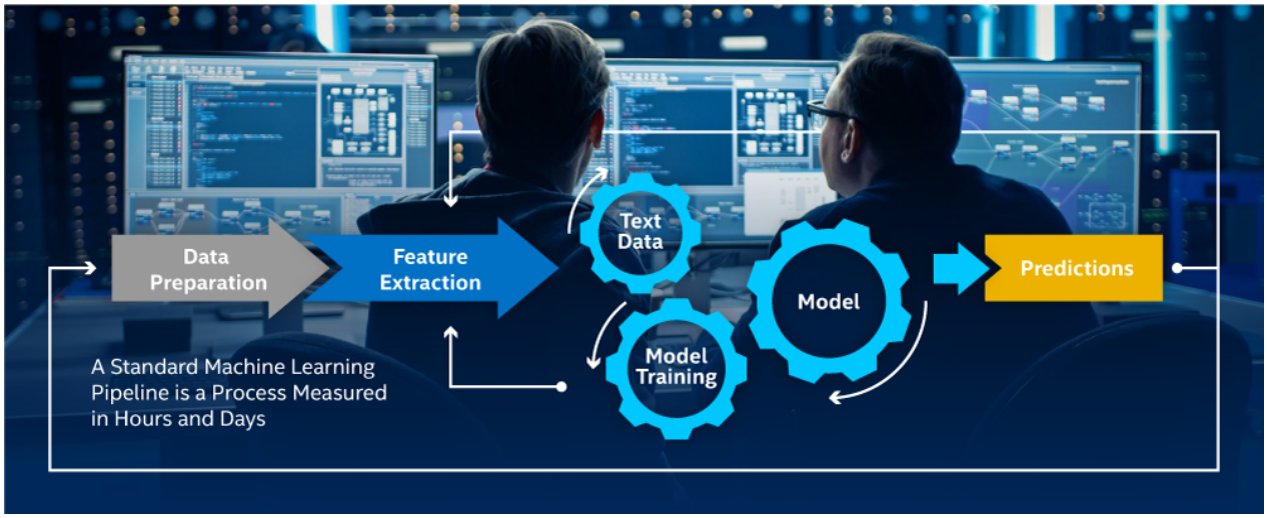
Not only did we **outperform both AMD EPYC 7763 (codenamed Milan) and Nvidia A100 GPUs** on a majority of workloads, but we also outperformed them across the geomean of these key customer workloads. Long story short, customers choose Intel® Xeon® Scalable Processors for their performance and TCO benefits, and—as you'll read later—it delivers the performance they ***need!***

Inferencing Performance with Software Optimizations[7]



A Day in the Life of a Data Scientist

How much extra money should you spend to save a little time? The answer: It all depends how much time we're talking about. It's an unrealistic representation of a data scientist's day to say that they run a program and then just sit on their hands waiting for it to resolve. They ingest, process, and experiment with data to create accurate models and strategies. This process takes a lot of experimentation and time … time measured in hours and days, not microseconds. Intel looks at the entire pipeline, not just one aspect of it. The graphic below shows a standard machine learning (ML) pipeline and how data scientists spend their time. There are misconceptions in the industry that GPUs are required to handle this workflow, which is not based on what a data scientist does on a daily basis.



So how do you compare the performance of different solutions for an end-to-end (E2E) ML pipeline? We've tested many real E2E workflows that read data from a large dataset (**Readcsv** in the chart below), iterate on the data multiple times to create a model (**ETL** and **Model Training**), and then run predictions on the model (**ML Time**).

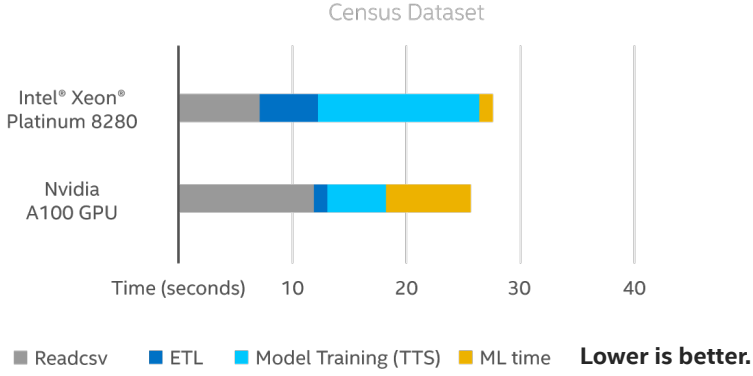
3rd Generation Intel® Xeon® Scalable Processors deliver 25% faster E2E data science at all phases of the pipeline comparable to 2nd Generation Intel® Xeon® Scalable Processors.[8]

But what about GPUs? Will we be waiting hours or days longer to get results? When you look at the whole picture, 3rd Generation Intel® Xeon® Scalable Processors deliver competitive performance as GPUs for this representative E2E workload. In fact, the difference in completion time is less than the average time between eye blinks[9]—a far cry from what some may want you to believe!

Consider This:

3rd Generation Intel® Xeon® Scalable Processors deliver competitive performance without the likely added cost and complexity of switching to a GPU platform

End-to-End Machine Learning Performance[10]



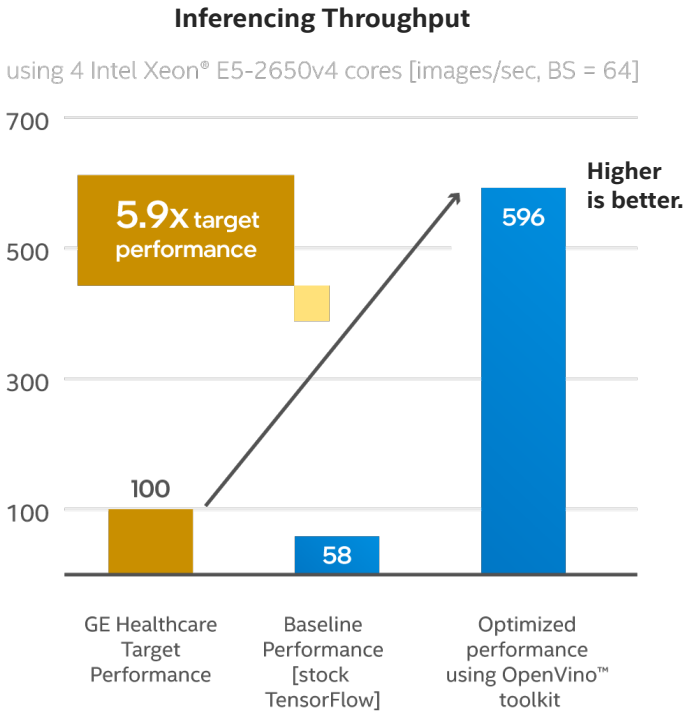
Do I Always Need the Highest Performance?

Get the performance you NEED by optimizing on the Intel® Xeon® Scalable Processor hardware you already use and trust.

GE Healthcare

GE engineers needed an inferencing solution that could keep pace with their imaging pipeline and make it flexible enough to deploy on different CT scanner models or even in the data center or cloud ... all without increasing their costs. They had four unused Intel® Xeon® Processor cores in their machines and needed to hit a goal of at least 100 images per second in order to keep up with their imaging pipeline. In collaboration with the Intel team and utilizing the OpenVINO™ toolkit, GE was able to realize high performance and low TCO on Intel® Xeon® Scalable Processors, resulting in a 14x speed increase compared to their baseline solution and 5.9x above their inferencing targets![11]

Check out the white paper here.



intel.

Performance varies by use, configuration and other factors. Learn more at www.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 technologies may require enabled hardware, software or service activation.

Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

Performance Results:

1 See [117] at www.intel.com/3gen-xeon-config. Results may vary.

2 See [123] at www.intel.com/3gen-xeon-config. Results may vary.

3 See [45] at www.intel.com/3gen-xeon-config. Results may vary.

4 See [43] at www.intel.com/3gen-xeon-config. Results may vary.

5 See [44] at www.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/2.3GHz, 270W TDP) processor on Intel® Software Development Platform with 512 GB (16 slots/ 32GB/ 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/32GB/2933MHz) 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, Pre-processing Modin 0.8.3, Omniscidbe v5.4.1, Intel Optimized Scikit-Learn 0.24.1, OneDAL Daal4py 2021.2, XGBoost 1.3.3, Dataset source: IPUMS USA: https://usa.ipums.org/usa/, Dataset (size, shape): (21721922, 45), Datatypes int64 and float64, Dataset size on disk 362.07 MB, Dataset format .csv.gz, Accuracy metric MSE: mean squared error; COD: coefficient of determination, tested by Intel, and results as of March 2021.

9 Source: https://www.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/2.3GHz, 270W TDP) processor on Intel® Software Development Platform with 512 GB (16 slots/ 32GB/ 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/ 32GB/ 3200) total DDR4 memory, ucode 0x8301034, HT on, Turbo on, Ubuntu 18.04.5 LTS, 5.4.0-42-generic, NVIDIA A100 (DGX-A100), 1.92TB 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, Omniscidbe 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 Sklearn 0.24, OneDAL CuML 0.17, XGBoost 1.3.0dev.rapidsai0.17, Nvidia RAPIDS 0.17, CUDA Toolkit CUDA 11.0.221. Dataset source: IPUMS USA: https://usa.ipums.org/usa/, Dataset (size, shape): (21721922, 45), Datatypes int64 and float64, Dataset size on disk 362.07 MB, Dataset 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 16GB / 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 15th 2018 and may not reflect all publicly available security updates. No product can be absolutely secure.