BUSINESS BRIEF

Enterprise & Government
Artificial Intelligence

Accelerating AI Adoption

Deliver enterprise AI with minimal cost and complexity, using familiar Intel® infrastructure and the only CPU with AI acceleration built in

IDC predicts that **75%** of enterprise applications will integrate AI by 2021[1]

Artificial intelligence (AI) holds great potential to drive digital transformation in the enterprise, with IDC predicting that 75 percent of enterprise applications will integrate AI by 2021[1]. However, many organizations are grappling with how to begin even as they accumulate an overwhelming variety and volume of data. While more than half of the world's data was created in the last two years, less than 2 percent has been analyzed[2]. Each organization must figure out which data can drive competitiveness and boost business performance. Then they must solve the challenge of capturing, analyzing and using that data.

Align AI to the data lifecycle

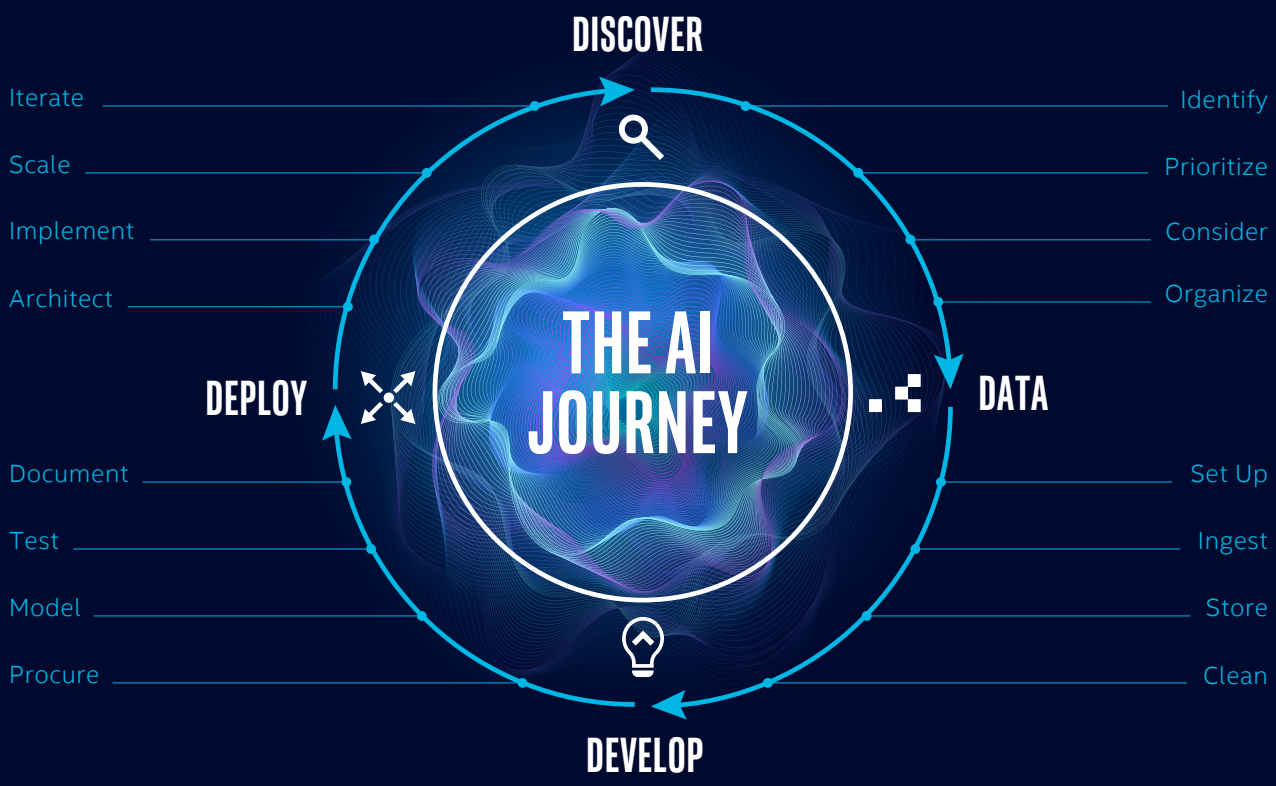
Creating value from data—whether using basic analytics and business intelligence (BI) or true AI—demands a robust strategy across the data lifecycle.

The Intel® Xeon® platform offers an ideal foundation for organizations to get started with machine learning and AI using the existing data, analytics program, and infrastructure resources in place today. Through its work with the broader AI solution provider ecosystem, Intel is able to support enterprises in delivering on their AI strategy across every step of the data lifecycle, which can be broken down into four main stages:

Discover:

Before any AI initiative begins, it is important to ensure the groundwork has been laid. Organizations should begin by identifying possible AI use cases, pulling in perspectives from across the business and prioritize those likely to realize the greatest return in the quickest timeframe. Following that, the next step is to set up cross-business unit teams, including all the business, IT and engineering skills required, that can work together to drive an AI program.

For inspiration on where to start, consult resources like Intel® AI Builders, which offers a portfolio of over 100 ready-to-use AI solutions for a range of industries and use cases.



Data:

Data enters the enterprise diverse in source, volume, format, structure, and frequency. A key first step in is creating a system to capture, store and process it all as it enters, from edge to cloud. A unified corporate data framework can bring order to this torrent. Compression technologies like Intel® QuickAssist can help further reduce the size of data to be transferred. Also make sure that across your environment, strong, reliable connectivity is in place to support the transfer of high volumes of data. Increase bandwidth as your data demands grow, using Intel® Ethernet 700 series offerings or, for higher speed connectivity over greater distances, Intel® Silicon Photonics products.

Data scientists today spend around 40 percent of their time simply gathering and cleaning data[3]. By breaking down data siloes and creating a unified view across all corporate data, it's possible to help reduce this burden. Having a strong compute foundation like the 2nd generation Intel® Xeon® Scalable processor is also important at this stage to help accelerate these processes, while Intel® Optane persistent memory can enable more real-time processing, holding more data closer to the CPU. Once the data has been prepared, make high-capacity, cost-effective data storage available using Intel® Optane™ SSDs.

Develop:

With the data ready to use, organizations must next plan to architect for and integrate their chosen AI workloads. Conveniently, most machine learning in the enterprise is already possible on Intel® architecture, making it simple to implement new AI workloads in existing, familiar, general-purpose environments. Tools like the Intel® Math Kernel Library help further accelerate machine learning algorithms. This same infrastructure can support more complex deep learning workloads, with deep learning acceleration built-in. For example, Intel® Deep Learning Boost, delivered with 2nd and 3rd generation Intel Xeon Scalable processors, provides a performance boost for embedded AI workloads. The 3rd generation Intel Xeon Scalable processor, to be launched later this year, will include additional support for model training and inference with the bfloat16 floating point format.

Deploy:

Once an AI use case has been proven valuable, the last step is to operationalize new projects so they can run at scale. Making sure application developers have the tools they need to build, launch and maintain AI applications in line with business need is critical. For example, the Intel® Distribution of OpenVINO Toolkit is designed to help developers quickly build new inferencing applications, whether vision, text, speech or other type of unstructured data.

Intel® technologies help simplify and accelerate AI innovation

Many enterprise organizations see implementing AI in their environment as a significant investment that they can't afford to get wrong.

The good news is that in most cases, that large investment is not needed. It's possible to make a start using familiar Intel® platforms in the data center today, and build capabilities over time, as value is proven.

The only server CPU with AI acceleration built in

Building on the Intel Xeon platform's suitability for getting started with machine learning using existing general-purpose hardware, the latest, 2nd generation Intel® Xeon® Scalable processors offer the general utility and high performance needed to drive a wide range of compute-heavy workloads. It also includes Deep Learning Boost, built in AI acceleration, that can help power complex AI workloads while supporting future growth and unanticipated application needs.

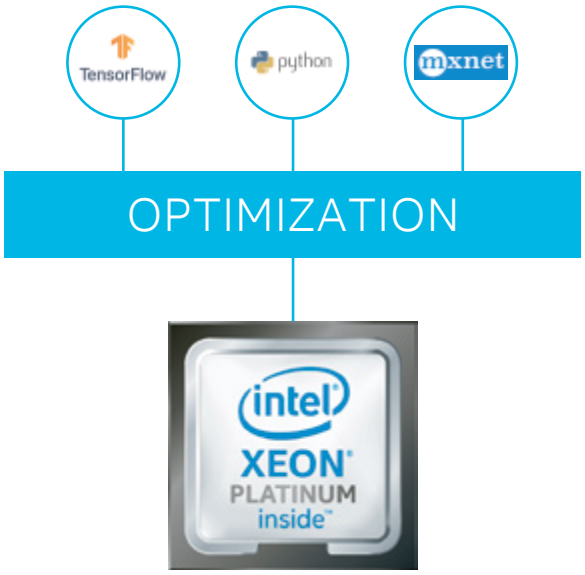
By relying on Intel Xeon Scalable platform as you build your AI environment, you can also benefit from significant cost advantages. Investment in code and software development will continue to deliver return thanks to the platform's broad compatibility with common software, frameworks and toolkits. Also, by staying on a known programming model and trusted infrastructure, you can avoid technical debt and unnecessary complexity in the data center. This combination of proven performance and reliability today, along with focus on continued development for the future, makes the Intel Xeon Scalable processor the go-to foundation for AI.

Software optimized for AI

Software is critical to AI. Recreating from scratch the math and algorithms required for each application is a complex and expensive undertaking. Using frameworks, libraries, tools, and pre-trained models, teams can benefit from the hard work already done in a given knowledge domain, such as vision or speech.

To meet this need, Intel has developed optimizations of the most widely used AI software frameworks for Intel architecture, which help drive significantly higher performance. For example, image classification workloads running on ResNet-50 can achieve up to 14X improvement in inference performance on Intel® Xeon® Platinum 8280 processor with Intel® DL Boost for ResNet-50 using Intel® Optimizations for Caffe, compared to running the same workload on Intel® Xeon® Platinum 8180 processor at launch[4]. Other optimizations include those for TensorFlow, MXNet, and Python, ensuring enterprises can achieve the performance they need on the data center servers they know.

Intel has optimized the leading frameworks to make it possible to develop highly effective AI models in existing environments using well known Intel® Xeon® platforms.



Simplify AI deployment with Intel's ecosystem

To simplify AI deployments further, Intel® Select Solutions for AI are pre-configured and workload-optimized to deliver high performance, cost efficiency, and security, delivered through Intel's solution provider partners. These solutions include:

- Intel® Select Solution for AI Inferencing, which uses industry-standard Intel technology along with the Intel Distribution of OpenVINO toolkit, a deep learning reference software stack, and Kubeflow and Seldon Core to help simplify model deployment. Providing low-latency, high-throughput inference and built-in acceleration, the solution suits a variety of budgetary and performance requirements.
- Intel® Select Solution for BigDL on Apache Spark, which offers a workload-optimized combination of deep learning software and Intel hardware that eliminates the need for organizations to research and manually optimize infrastructure to support their AI initiatives. This offering is particularly useful for those looking to build out their AI capabilities at speed and scale.

Major cloud service providers have also developed pre-configured, Intel® architecture-optimized AI environments to support on-demand AI deployments that take advantage of features like Intel DL Boost:

- AWS: Optimized EC2 instances for machine learning or compute-intensive applications.
- Microsoft Azure: Intel-optimized Data Science Virtual Machine (DSVM)
- Google Cloud Platform: TensorFlow optimizations for Intel platforms and the latest Intel machine instances

Leading OEMs—including Dell, HPE and Lenovo—offer AI capabilities as a service, optimized to run on Intel Xeon platforms to deliver strong price performance and technical agility to flex in line with user demands.

In addition, Intel works closely with a range of AI software providers, such as Data Robot and H20.ai to help optimize their applications for Intel® technology-based environments and ensure quick and seamless deployment. The Intel AI Builders community is also dedicated to accelerating AI deployments on Intel® architecture, and offers a range of market-ready solutions.

As your organization grows

As organizations' data resources continue to grow, and the questions they ask become more complex, it's likely their AI applications will also evolve over time. As use cases mature, especially in enterprise environments generating data with economic value, many will need to optimize and invest.

Intel provides customers a cost-effective path to AI today and a commitment to their evolving journey. It continues to invest in the future to meet the ever-expanding hunger for processing and data represented by today's—and tomorrow's—workloads. It is committed to the ongoing development of the industry's broadest AI compute portfolio, based on the Intel® Xeon® Scalable platform and a range of accelerators that allow enterprises to grow their AI applications in line with their own business needs and IT environment.

Intel® Technology Differentiators

Intel® Xeon® processor's enhanced AI performance: 2nd generation Intel® Xeon® Scalable processor delivers up to 30x AI performance improvement with Intel® Deep Learning Boost, compared to previous generation Intel® Xeon® processors without Intel DL Boost[5]. New 3rd generation Intel® Xeon® Scalable processors continue Intel's leadership in built-in AI acceleration, and are the first general-purpose server CPUs with built-in bfloat16 support, which accelerates both AI training and inference performance

Software optimized for AI: 3.75x AI inferencing performance increase[6] using OpenVINO or Tensorflow with Intel DL Boost

Market-ready AI solutions: A rich ecosystem of solution providers, combined with workload-optimized and verified Intel® Select Solutions, enable you to get up and running quickly

Business Brief | Accelerating AI Adoption 4

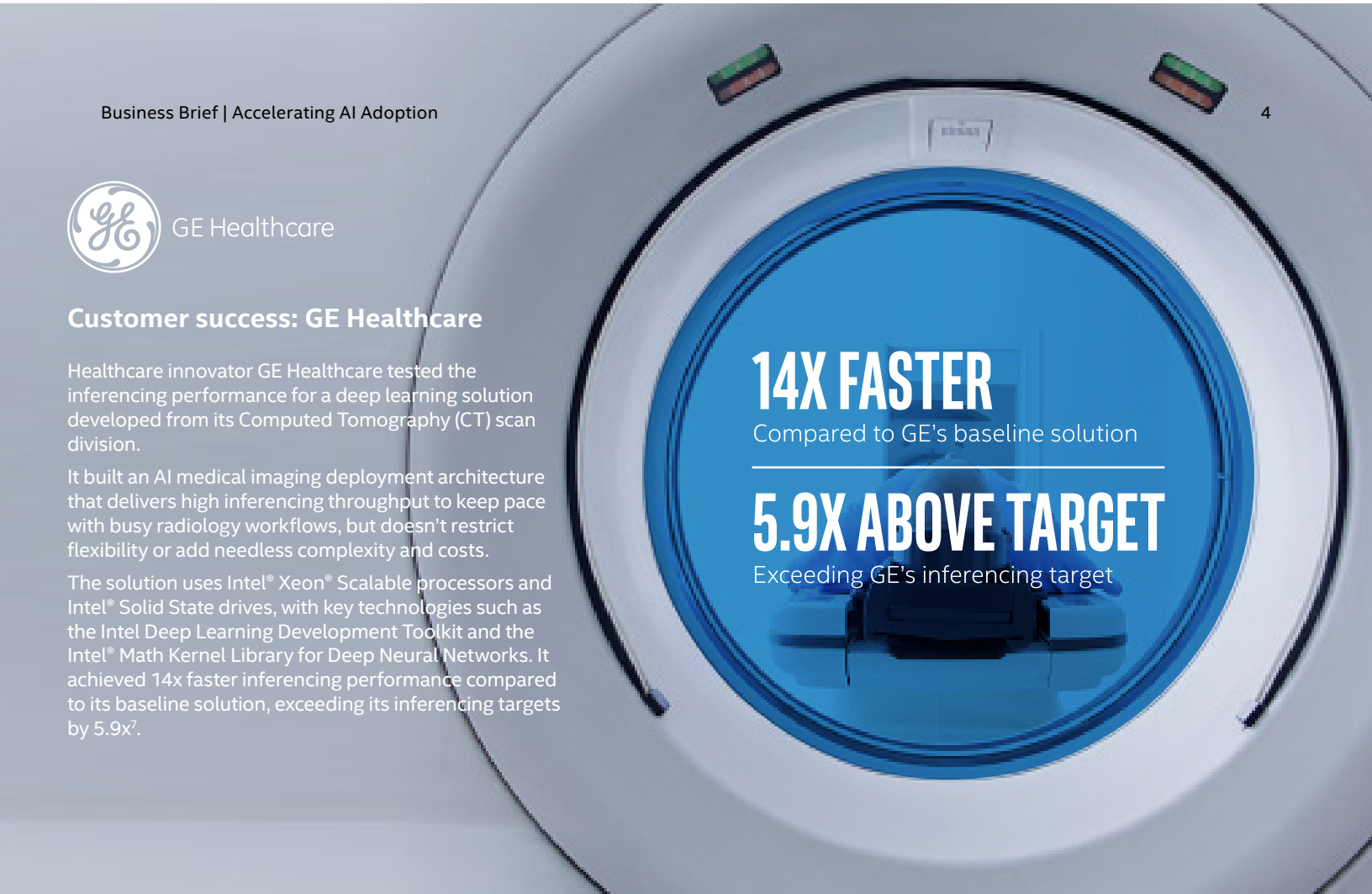
GE Healthcare

Customer success: GE Healthcare

Healthcare innovator GE Healthcare tested the inferencing performance for a deep learning solution developed from its Computed Tomography (CT) scan division.

It built an AI medical imaging deployment architecture that delivers high inferencing throughput to keep pace with busy radiology workflows, but doesn't restrict flexibility or add needless complexity and costs.

The solution uses Intel® Xeon® Scalable processors and Intel® Solid State drives, with key technologies such as the Intel Deep Learning Development Toolkit and the Intel® Math Kernel Library for Deep Neural Networks. It achieved 14x faster inferencing performance compared to its baseline solution, exceeding its inferencing targets by 5.9x[7].



UP TO 3.5X INCREASE

In **training** performance on 2nd Gen Intel Xeon Scalable processor using Intel-optimized Tensorflow

UP TO 2.4X IMPROVEMENT

in inference on 2nd Gen Intel Xeon Scalable processor using Intel-optimized Tensorflow

IFDAQ

Customer success: IFDAQ

IFDAQ began as a scientific research project in 2008, when a consortium of AI pioneers and big data leaders mutually explored new technologies to refine the world's largest fashion data asset.

The fashion and luxury industry is highly complex, and insight into fundamental KPIs gives companies valuable information which they can use to improve their competitive edge. IFDAQ needs compute performance which can deliver these results to its customers in a timely and cost-efficient manner.

Its AI solution helps in visualizing, predicting and monitoring the market's dynamics and impacts in real-time and with meaningful transparency. It enabled the team to predict careers and performances of professional fashion models by calculating a final benchmark value from decisive factors under complex conditions. This guarantees an accurate and precise rating of the talent and shows 3.5x improvement in training performance and a 2.4x improvement in inference performance , using Intel optimized Tensorflow on 2nd Gen Intel Xeon Scalable processor.

Learn more

- **eGuide:** Ease Your Organization into AI
- **Solution Brief:** Intel® Select Solutions for BigDL on Apache Spark
- **Solution Brief:** Intel® Select Solutions for AI Inferencing
- **Webpage:** CSP and OEM Deployment Solutions

intel

[1] IDC Market Forecast, Worldwide Storage for Cognitive/AI Workloads Forecast, 2018–2022, https://www.forbes.com/sites/louiscolumbus/2018/11/04/idc-top-10-predictions-for-worldwide-it-2019/#10c156c97b96

[2] Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018, https://www.seagate.com/gb/en/our-story/data-age-2025/

[3] https://businessoverbroadway.com/2019/02/19/how-do-data-professionals-spend-their-time-on-data-science-projects/

[4] 14x inference throughput improvement on Intel® Xeon® Platinum 8280 processor with Intel® DL Boost: Tested by Intel as of 2/20/2019. 2 socket Intel® Xeon® Platinum 8280 Processor, 28 cores HT On Turbo ON Total Memory 384 GB (12 slots/ 32GB/ 2933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x200004d), Ubuntu 18.04.1 LTS, kernel 4.15.0-45-generic, SSD 1x sda INTEL SSDSC2BA80 SSD 745.2GB, nvme1n1 INTEL SSDPE2KX040T7 SSD 3.7TB, Deep Learning Framework: Intel® Optimization for Caffe version: 1.1.3 (commit hash: 7010334f159da247db3fe3a9d96a3116ca06b09a) , ICC version 18.0.1, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a, model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt,
BS=64, synthetic Data, 4 instance/2 socket, Datatype: INT8 vs. Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

[5] 30x inference throughput improvement on Intel® Xeon® Platinum 9282 processor with Intel® DL Boost : Tested by Intel as of 2/26/2019. Platform: Dragon rock 2 socket Intel® Xeon® Platinum 9282(56 cores per socket), HT ON, turbo ON, Total Memory 768 GB (24 slots/ 32 GB/ 2933 MHz), BIOS:SE5C620.86B.0D.01.0241.112020180249, Centos 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, No datalayer syntheticData:3x224x224, 56 instance/2 socket, Datatype: INT8 vs Tested by Intel as of July 11th 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50GHz (28 cores), HT disabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 384GB DDR4-2666 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800GB, 2.5in SATA 6Gb/s, 25nm, MLC).Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, synthetic dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50),. Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

[6] 3.75x improvement with AI Inferencing Intel Select Solution. The solution was tested with KPI Targets: OpenVINO/ ResNet50 on INT8 on 02-26-2019 with the following hardware and software configuration:

Base configuration: 1 Node, 2x Intel® Xeon® Gold 6248; 1x Intel® Server Board S2600WFT; Total Memory 192 GB, 12 slots/16 GB/2666 MT/s DDR4 RDIMM; HyperThreading: Enable; Turbo: Enable; Storage(boot): Intel® SSD DC P4101; Storage(capacity): At least 2 TB Intel® SSD DC P4610 PCIe NVMe; OS/Software: CentOS Linux release 7.6.1810 (Core) with Kernel 3.10.0-957.el7.x86_64; Framework version: OpenVINO 2018 R5 445; Dataset:sample image from benchmark tool; Model topology: ResNet 50 v1; Batch Size: 4; nireq: 20. 0

[7] Configuration: 2-socket Intel(R) Xeon(R) 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 ver: 2018.1.249,Intel® MKL-DNN ver:0.14 https://simplecore.intel.com/nervana/wp-content/uploads/sites/53/2018/04/Intel-Software-Development-Tools-Optimize-Deep-Learning-Performance-for-Healthcare-Imaging-paper-ForDistribution.pdf Patch disclaimer: Performance results are based on testing as of June 15th 2018 and may not reflect all publicly available security updates.

[8] Configuration: IFDAQ Configuration: NEW: Tested by Intel as of 08/06/2019. 2 socket Intel® Xeon® Gold 6248 Processor, 20 cores per socket, OS Ubuntu 18.04.2 LTS, Deep Learning Framework: Intel Optimized TensorFlow 1.12.0, custom test data, for Feed Forward (single layer), Batch size : 200, python 3.6 BASELINE: Tested by Intel as of 08/06/2019. 2 socket Intel® Xeon® Gold 6248 Processor, 20 cores per socket, OS Ubuntu 18.04.2 LTS, Deep Learning Framework: TensorFlow 1.12.0, custom test data, for Feed Forward (single layer),Batch size : 200

[9] Configuration: IFDAQ Configuration: NEW: Tested by Intel as of 08/06/2019. 2 socket Intel® Xeon® Gold 6248 Processor, 20 cores per socket, OS Ubuntu 18.04.2 LTS, Deep Learning Framework: Intel Optimized TensorFlow 1.12.0, custom test data, for Feed Forward (single layer), Batch size : 200, python 3.6 BASELINE: Tested by Intel as of 08/06/2019. 2 socket Intel® Xeon® Gold 6248 Processor, 20 cores per socket, OS Ubuntu 18.04.2 LTS, Deep Learning Framework: TensorFlow 1.12.0, custom test data, for Feed Forward (single layer),Batch size : 200

Performance results are based on testing as of the date set forth in the configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

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

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation.

Your costs and results may vary.

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.

All product plans and roadmaps are subject to change without notice.

0420/JL/CAT/PDF 343148-001EN