PROWESS

How Intel® CPUs Can Accelerate Your AI Pipeline

Thinking outside of the graphics processing unit (GPU) box can reward you with faster time to solution for artificial intelligence (AI) and a lower total cost of ownership (TCO).

Executive Summary

Intel reports that Intel® Xeon® Scalable processors account for 70 percent of artificial intelligence (AI) data center inferencing.[1] There appear to be several reasons behind this overwhelming popularity. Many organizations are finding that x86 data-center processors with built-in accelerators can boost performance without specialized software workarounds or expensive hardware add-ons. And in many cases, these AI-enhanced CPUs enable deployment of an open, unified platform that can deliver better performance for AI and non-AI workloads, and that can also help lower total cost of ownership (TCO).

A Faster AI Time to Solution

We see the benefits of AI in our everyday lives. AI is the "smart" behind voice-assisted speakers, curated online content, home security systems, autonomous vehicles, phone apps, suggested audio and video playlists, and many more technologies. Organizations across all industries are discovering how data-driven decision making holds the key to business and operational success. Businesses of all sizes, from local independents to global enterprises, are using AI to turn an endless sea of data into actionable insights. They use these insights to make their services and products stand out, to stay responsive and agile, and to get ready for future growth. If you are not using AI by now, you should realize that your competitors are. If you are already using AI, you might be wondering how you can make it work even better.

Even with the apparent growth of AI-driven business strategies, Prowess Consulting believes that many organizations are barely scratching the surface of the potential benefits they could gain. This is because the AI pipeline—data, model, and deploy—can require significant expertise to successfully build and deploy. We suggest that the solution to extracting more actionable insights from your AI pipeline begins with examining your hardware, the bones of your data-centric infrastructure. Today, the majority of workloads passing through the data and deploy stages of the AI pipeline are handled by CPUs, especially data-intensive preprocessing and inferencing. While graphics processing units (GPUs) dedicate the most cycles for DL training during the model stage, CPUs also contribute to machine learning (ML) algorithm processing and some deep learning (DL) workloads.

Technical business decision makers understand that CPUs, such as Intel Xeon Scalable processors, provide a more economical solution for running AI workloads than GPUs and other dedicated hardware. Your organization might be among those who choose these x86 data center CPUs to run inferencing workloads seven times out of 10.[1] However, you might not be aware that you can improve AI performance and TCO even further by using the Intel Xeon Scalable processor's built-in AI accelerators. These hardened AI accelerators are engineered to deliver highly performant processing for DL inferencing workloads during the deploy stage of the AI pipeline.

If you already have Intel Xeon Scalable processors installed, you can use any number of open standards–based tools to tune your AI pipeline. Optimizing your existing infrastructure with software tools can upgrade AI performance more quickly and cost-effectively than buying, installing, and administering new hardware. Speaking of software, we recommend maintaining an open software environment, whether you are tuning existing hardware or installing new hardware. Open AI platforms, such as cnvrg.io®, can help unify your AI pipeline, so it performs better and is simpler to administer. Open-source frameworks, such as the popular TensorFlow™ and PyTorch® frameworks, allow your infrastructure engineers, app developers, and other IT staff to work with software, operating systems, and infrastructures they know how to build and support. Running ISV-certified, AI-optimized enterprise database applications, such as SAP HANA®, Microsoft® SQL Server®, Oracle® Database, or Oracle® Exadata® also helps ensure platform interoperability and reliability. An ISV-certified and open standards–based ecosystem can help prevent IT staff from having to integrate new stacks that require specialized knowledge to administer.

In short, if lower TCO and accelerated time to solution are important benefits to you, then CPUs with advanced AI capabilities, like Intel Xeon Scalable processors, could be an effective solution to your AI deployment challenges.

A Unified, Scalable AI Pipeline

An AI pipeline encompasses five data-processing workloads:

- Data ingress
- Data exploration/preprocessing
- Model selection
- Training
- Inferencing

So, it makes sense to examine the whole AI pipeline when seeking ways to optimize performance. Figure 1 shows an often-overlooked aspect of the AI pipeline, that the bulk of AI pipeline activity is devoted to data ingress, data exploration, preprocessing, and inferencing workloads. A less obvious point is that AI-accelerated CPUs are well-suited for DL training during the model stage.

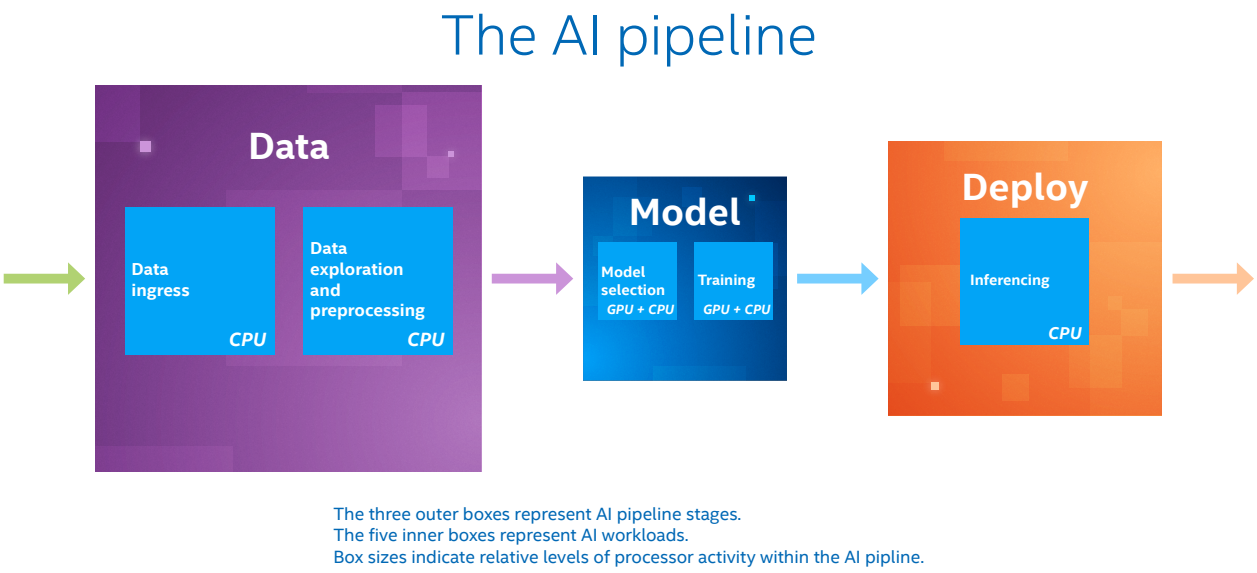

Figure 1. Instead of focusing on one or two AI workloads, zoom out and investigate how you can improve performance across the entire AI pipeline

The importance of GPU-based processing for using large datasets when creating models cannot be dismissed—it is the preferred approach for parallel-processed model selection and training workloads. But as you can see, these two workloads represent just one stage in an AI pipeline, and not always the most active one. Workloads are constantly moving through the pipeline as data is ingested, models get trained, and processing shifts over to deployment testing and production. And workloads can cycle back to ingress, exploration, or preprocessing as new data is ingested.

Figure 1 shows how upgrading the GPU to improve model-stage performance might not be as beneficial as optimizing the CPUs to improve performance across the entire AI pipeline. Optimizing existing Intel Xeon Scalable processors, for example, is probably more cost-efficient than buying new GPU hardware. In addition to costing less, you could achieve greater overall performance because optimizing the CPUs improves processing across all workloads in the AI pipeline.[2] If you decide it is time for a hardware upgrade, investing in AI-enhanced CPUs is probably a better price-performance choice than upgrading dedicated AI accelerator hardware, such as GPUs.

Unlike dedicated AI accelerator hardware, CPUs can be configured to automatically switch to processing non-AI data-analytics workloads as needed, keeping them from sitting idle. This scalable processing lets you extract greater operational efficiency across your data-centric infrastructure, which helps lower TCO. CPU scale-outs also help reduce processing time for certain workloads. For example, it might take eight Intel Xeon Scalable processors approximately 16 hours to train an image recognition model during normal working hours. Automated scaling that recruits as many as 64 Intel Xeon Scalable processors can reduce this time by up to 4x during weekends and evenings.[3] Not only does this improved processing efficiency deliver shorter time to solution, but it also means more workloads can get processed within a given period.

Accelerated Preprocessing

The AI pipeline begins in the data stage with data ingestion and preprocessing. Data preprocessing constitutes the heaviest lift in today's modern data pipeline, whether it's for AI, data mining, high-performance computing (HPC), or data science.[4,5,6] In a 2020 survey conducted by Anaconda, respondents reported that 45 to 66 percent of their data-science pipelines were occupied by preprocessing.[7] This percentage means that regardless of model selection and training performance, a slow data stage can delay your AI pipeline's time to solution.

Data preprocessing is critical for achieving accurate ML-driven medical diagnoses, such as differentiating malignant cancer from a chronic health condition like diabetes.[8] Preprocessing helps ensure that image classification and predictive models are generated from accurate, complete, and relevant data.

Preprocessing workloads are almost exclusively handled by CPUs, which are ideally suited for compute-intensive, single-node processing. GPUs typically remain idle during the data stage and kick in during the model stage. An AI-accelerated CPU, such as an Intel Xeon Scalable processor, can help deliver high-throughput, low-latency processing for massive datasets,[9] such as medical imaging files.

Pandas is a popular open-source library used for data preprocessing. The Intel® Distribution of Modin library enables Pandas to be scaled using one line of code. This scale-up implementation can boost 3rd Generation Intel Xeon Scalable processors' preprocessing performance by up to 90x.[3,10] Accelerated preprocessing enables 3rd Generation Intel Xeon Platinum 8380 processors to outperform AMD EPYC™ 7763 processors, which cannot be accelerated, on end-to-end workloads by up to 20 percent.[11]

Faster Inferencing

Healthcare organizations are using Intel Xeon Scalable processors to improve AI-driven analyses of high- and super-resolution medical imaging and structured reports. The faster a DL model can accurately analyze medical imaging, the earlier medical staff can make appropriate diagnoses and prescribe treatments. In an optimization test of the Huiyi Huiying Medical Technology Company's AI-driven diagnostic platform, the Dr. Turing™ AI platform, Intel Xeon Scalable processors were optimized with the Intel® Distribution of OpenVINO™ toolkit. The accelerated platform showed improved inferencing performance in a breast cancer–detection model by as much as 8.24x, with a less than 0.17 percent decrease in accuracy compared to the original AI deployment.[12] In computed tomography (CT) image analysis for detecting signs of COVID-19 in patients, optimized Intel Xeon Scalable processors helped lower inference times as much as 35 percent from the baseline deployment.[13] In both cases, these performance boosts were achieved by using software to optimize processors already in place, rather than buying new AI-dedicated hardware.

Product manufacturers are using AI solutions built on Intel Xeon Scalable processors to speed up production time, improve product quality, boost operational efficiency, and lower system downtime. According to byteLAKE, an industrial simulations and monitoring solutions provider, optimized Intel Xeon Scalable processors helped reduce computational fluid dynamics (CFD) simulation run times for liquid chemical production from a few hours to 10–20 minutes, with more than 93 percent prediction accuracy—and without having to upgrade or overhaul the hardware infrastructure.[14]

International users of Alibaba's online services expect real-time responsiveness when interacting with AI-powered translation services such as natural language processing (NLP) and neural machine translation (NMT). DL workloads can use 32-bit floating-point precision (FP32), 16-bit brain floating point (BF16), or 8-bit integer low-precision (INT8). The lower the precision number, the faster the workload gets processed. Intel® Deep Learning Boost (Intel® DL Boost) enables 3rd Generation Intel Xeon processors to handle INT8-based inferencing for NLP, which not only delivers up to 3.1x better AI performance but also maintains higher accuracy than FP32-based inferencing running on previous-generation Intel Xeon Scalable processors.[15]

An Open Software Ecosystem

> "OSS enables and increases AI adoption by reducing the level of mathematical and technical knowledge necessary to use AI."
>
> *– The Brookings Institution*[16]

According to a report from The Brookings Institution, "How open-source software shapes AI policy," open-source software (OSS) plays a major role in AI adoption and standardization.[16] And we agree. Deploying an open software ecosystem for AI can help you reap the benefits of easier implementation, standards-based interoperability, best-practices testing, and de facto future-proofing.[16]

It is generally acknowledged that proprietary solutions are not as economical to own and operate as open solutions. Proprietary platforms are also notorious for creating vendor lock-in, which usually means that add-ons and upgrades to improve performance will keep getting more expensive. Non-standard tools often require IT staff with specialized knowledge to develop, install, administer, and troubleshoot. A non-standard ecosystem is more likely to develop workloads that are siloed by application. Improving performance for a siloed AI pipeline can require complex workarounds or expensive infrastructure overhauls.

The Goldwind SE wind farm relies on an open software ecosystem for developing a highly responsive, multi-model solution that helps predict optimal electricity generation under ever-changing weather conditions. Its AI platform, which runs on Intel Xeon Scalable processors, integrates Apache Spark™, TensorFlow, Keras®, Apache® MXNet, and other applications, frameworks, and libraries into a single AI pipeline.[17] The open platform integrates different data sources and software, which work together to rapidly develop, prototype, and deploy AI models. In power-prediction field trials, testers fed 30,000 records into an LSTNet DL model (Apache MXNet framework) to generate 5,000 iterative optimizations. The improved AI pipeline delivered a wind-power prediction for a two-hour window within 50 ms of data ingress, 4x faster than the old solution, and with 79 percent accuracy, a 20 percent improvement.[18]

Open Tools, Heterogeneous AI Hardware

Open standards toolkits, such as oneAPI, SigOpt, Intel® oneAPI Deep Neural Network (Intel® oneDNN) library, and the Intel Distribution of OpenVINO toolkit, offer familiar, standardized interfaces that allow you to write once and deploy everywhere across your AI hardware. Free resources, forums, and support are provided through a highly knowledgeable, international open-source community of AI architects, developers, and administrators.

Pre-optimized AI frameworks and libraries can help you get ML and DL deployments up and running at peak performance sooner. They can help you eliminate the time-consuming chore of building frameworks and foundational libraries from scratch, and then testing those frameworks and libraries. Intel Xeon Scalable processors natively support open AI frameworks and libraries like TensorFlow, PyTorch, scikit-learn®, and XGBoost to deliver as much as 10- to 100-fold performance improvements over unoptimized hardware infrastructures.[19]

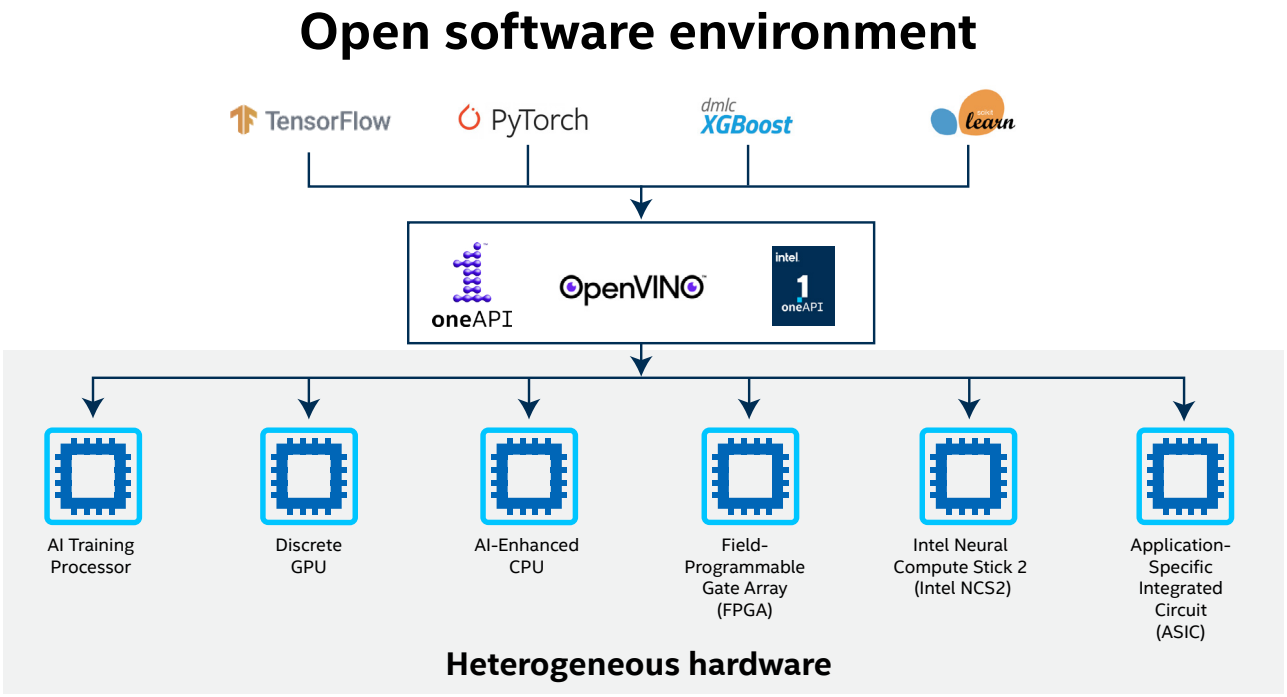

Figure 2. We envision the future of cost-efficient, high-performing AI deployments as an open software ecosystem running on heterogeneous hardware

As AI becomes an increasingly prevalent data-analysis tool for all types of industries, the future of AI deployment appears to be heading toward an open software ecosystem running on heterogeneous hardware. Heterogeneous computing directs traffic to the processor best suited for a particular AI workload. An open AI platform lets you plug in the high-performing GPU, CPU, or other AI accelerator hardware of your choice for a winning combination of price and performance.

Optimize Your Entire AI Pipeline

With operational efficiency, interoperability, scalability, performance, and TCO in mind, we recommend optimizing your installed CPUs before spending money on new hardware. CPUs with built-in AI accelerators are easily fine-tuned to deliver highly performant inferencing, while also saving you time and money. If you need to upgrade your AI hardware, we suggest the best price-performance strategy is installing x86 server CPUs with built-in AI accelerators. Intel Xeon Scalable processors fall into this category; they contribute processing to all workloads and help extend your AI pipeline seamlessly from the workstation to the data center or cloud, and to the edge. And because they are not dedicated accelerators, these CPUs can shift over to other compute-intensive workloads when AI demands are low.

We suggest deploying AI on an open ecosystem, using open software, frameworks, and toolkits for optimizing, and debugging AI deployments. Not only will you be rewarded with a more streamlined AI pipeline that is simpler to administer, you will more likely be able to update and optimize your data infrastructure without resorting to expensive overhauls.

You can learn more about improving AI performance, lowering your TCO, and future-proofing data centers, by reading Intel's "Critical Considerations for AI Deployments."

[1] Based on Intel market modeling of the worldwide installed base of data center servers running AI inference workloads as of December 2021. Information was provided through direction communication between Prowess and Intel as of May 2, 2022.

[2] Source: Claim 100 at Intel. "Performance Index—3rd Generation Intel® Xeon® Scalable Processors." www.Intel.com/3gen-xeon-config.

[3] Intel. "Critical Considerations for AI Deployments." www.intel.com/content/www/us/en/products/performance/nvidia-ai-facts.html.

[4] Finance Train. "Data Preprocessing in Data Science and Machine Learning." May 2020. https://financetrain.com/data-preprocessing-in-data-science-and-machine-learning.

[5] Towards Data Science. "Understanding Data Preprocessing." https://towardsdatascience.com/data-preprocessing-e2b0bed4c7fb.

[6] CEOWORLD Magazine. "Data Preprocessing: what is it and why is important." December 2019. https://ceoworld.biz/2019/12/13/data-preprocessing-what-is-it-and-why-is-important/.

[7] Anaconda. "The State of Data Science 2020: Moving from hype toward maturity." www.anaconda.com/state-of-data-science-2020.

[8] Krishna Kant Singh, Mohamed Elhoseny, Akansha Singh, Ahmed A. Elngar (editors). "Chapter 5: Diagnosing of disease using machine learning." *Machine Learning and the Internet of Medical Things in Healthcare.* 2021. Pages 89–111. ISBN 9780128212295. https://doi.org/10.1016/B978-0-12-821229-5.00003-3.

[9] Claim 43 at Intel. "Performance Index—3rd Generation Intel® Xeon® Scalable Processors." www.Intel.com/3gen-xeon-config.

[10] Intel. "Performance Optimizations for End-to-End AI Pipelines—Optimized Frameworks and Libraries for Intel® Processors." Workloads and configurations at https://techdecoded.intel.io/resources/one-line-code-changes-to-boost-pandas-scikit-learn-and-tensorflow-performance/#gs.bzkn2n.

[11] Source: Claim 56 at Intel. "Performance Index—3rd Generation Intel® Xeon® Scalable Processors." www.Intel.com/3gen-xeon-config.

[12] Tested by Intel as of 9/18/2019. **Processor:** 2-socket Intel® Xeon® Platinum 8268 processor, 28 cores, Intel® Hyper-Threading Technology (Intel® HT Technology) on, Intel® Turbo Boost Technology on, 384 GB total memory (12 slots, 32 GB, 2,933 MHz); **BIOS:** SE5C620.86B.0X.02.0001.051420190324 (ucode: 0x4000024), CentOS® Linux® 7 (core); **Intel software:** Intel® Distribution of OpenVINO™ toolkit version 2019R2, Intel optimizations for TensorFlow™, Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN), Vector Neural Network Instructions (VNNI) to accelerate inferencing; **topology:** RetinaNet, https://github.com/fizyr/keras-retinanet; **compiler:** gcc 4.8.5, Intel MKL-DNN version v0.17, BS=1, eight asynchronous requests, customer data, 1 instance/2 socket; **datatype:** INT8; **dataset:** 366 X-ray mammogram images with an image resolution of 1280x640 provided by Huiyi Huiying. Source: Intel. "Huiying Medical Technology Optimizes Breast Cancer Early Screening and Diagnosis with Intel® AI Technologies." November 2019. https://builders.intel.com/docs/aibuilders/huiying-medical-technology-optimizes-breast-cancer-early-screening-and-diagnosis-with-intel-ai-technologies.pdf.

[13] Data is based on Huiyi Huiying internal testing results using the following configuration: **Processor:** Dual-socket Intel® Xeon® Gold 6252N processor, 2.30 GHz; **cores/threads:** 24/48; **operating system:** Ubuntu® 18.04.4 LTS; **DL framework:** PyTorch® version 1.5.1; **Intel® software:** Intel® Distribution of OpenVINO™ toolkit version R2020.3.194; **network model:** nested U-Net, HR-Net. Source: Intel. "HYHY: Full-Cycle AI-Based Medical Imaging Solution." www.intel.com/content/www/us/en/customer-spotlight/stories/hyhy-customer-story.html.

[14] IDC. "How byteLAKE Creates AI-Driven Industrial Solutions Using Intel Xeon Scalable Processors." Sponsored by Intel. April 2022. www.intel.com/content/www/us/en/data-center/idc-bytelake-case-study.html.

[15] Configuration details Alibaba® PAI NLP transformer model on PyTorch® 1.7.1 throughput performance on 3rd Generation Intel® Xeon® Scalable processors. **Baseline configuration:** Tested by Intel as of 03/19/2021. 2-node, 2 x Intel Xeon Platinum 8269C processor, 26 cores, Intel® Hyper-Threading Technology (Intel® HT Technology) on, Intel® Turbo Boost Technology on, 192 GB total memory (12 slots, 16 GB, 2,933 MHz), BIOS: SE5C620.86B.02.01.0013.121520200651 (ucode: 0x4003003), CentOS® 8.3, 4.18.0-240.1.1.el8_3.x86_64, gcc 8.3.1 compiler, transformer model, DL framework: PyTorch® 1.7.1, https://download.pytorch.org/whl/cpu/torch-1.7.1%2Bcpu-cp36-cp36m-linux_x86_64.whl, BS=1, customer data, 26 instances/2 sockets, datatype: FP32/INT8. **New configuration:** Tested by Intel as of 03/19/2021. 2-node, 2 x Intel Xeon Platinum 8369B processor, 32 cores, Intel HT Technology on, Intel Turbo Boost Technology on, 512 GB total memory (16 slots, 32 GB, 3,200 MHz), BIOS: WLYDCRB1.SYS.0020.P92.2103170501 (ucode: 0xd000260), CentOS 8.3, 4.18.0-240.1.1.el8_3.x86_64, gcc 8.3.1 compiler, transformer model, DL framework: PyTorch 1.7.1, https://download.pytorch.org/whl/cpu/torch-1.7.1%2Bcpu-cp36-cp36m-linux_x86_64.whl, BS=1, customer data, 32 instances/2 sockets, datatype: FP32/INT8. All performance data is tested in lab environment. Source: Intel. "Accelerating Alibaba Transformer model performance with 3rd Gen Intel® Xeon® Scalable Processors (Ice Lake) and Intel® Deep Learning Boost." www.intel.com/content/dam/www/central-libraries/us/en/documents/alibaba-lpot-blog-download.pdf.

[16] The Brookings Institution. "Report: How open-source software shapes AI policy." August 2021. www.brookings.edu/research/how-open-source-software-shapes-ai-policy/.

[17] Testing configuration: processor: Intel® Xeon® Gold 6130 processors; memory: 192 GB DDR4 2,666 MHz; operating system: CentOS® 7.6; Apache Spark™ version: 2.4.3. Source: Intel. "Goldwind SE: Intelligent Power Prediction Solution." www.intel.com/content/www/us/en/customer-spotlight/stories/goldwind-customer-story.html.

[18] Tech Field Day. "Building Big Data AI Applications on Intel Analytics Zoo." November 2020. https://techfieldday.com/video/building-big-data-ai-applications-on-intel-analytics-zoo/.

[19] Source: Sessions, AI002, Rachel Oberman, #9 at Intel. "Performance Index—Innovation Event Claims." https://edc.intel.com/content/www/us/en/products/performance/benchmarks/innovation-event-claims/.