ARTIFICIAL INTELLIGENCE: A GUIDE FOR CLOUD SERVICE PROVIDERS

How to grow your cloud services business with AI and machine learning

intel XEON® PLATINUM inside™

CONTENTS

IN BRIEF

- AI is beginning to add real value to businesses and organizations of all types
- CSPs have the opportunity to grow their business by helping their customers to take advantage of the benefits AI can bring
- To create the foundation for building a compelling AI proposition, CSPs need to focus on:

  a. improving the scalability of their infrastructure;

  b. selecting the right frameworks for their AI projects;

  c. ensuring their storage systems are architected correctly to support ML and DL workloads;

  d. improving the performance and speed of the platforms that will be running ML or DL algorithms;

AI AND MACHINE LEARNING: THROUGH THE HYPE

It has been said several times that AI is about to change the world. But now, the prediction may be about to come true. Three things are coming together for the very first time: data, compute and evidence.

1. We have more data than ever before, driven by exponentially increasing numbers of devices.
2. Computational power now allows us to start processing vast quantities of data at high speed.
3. We are starting to see evidence of AI's transformative power in the use cases and business models it is starting to make possible.

The combined maturity of data and compute technologies has accelerated development of machine learning and deep learning models and neural nets. The processing power and volume of data storage required to train and run these models is huge but advances in technology and modern computing models, such as cloud computing, have finally put it within our grasp. In other words, the affordability and efficiency of compute, storage and data technologies are key factors driving the current surge in artificial intelligence. As these factors drive early successes with AI use cases, the achievements and lessons learned inspire new projects. The impact of AI on the automotive industry as the enabler of autonomous driving is one example of this: continuous investment, prototyping and iteration have moved the industry forward in leaps. Another example is the clear power of advanced analytics for all types of organizations. Both of these use cases are helping to drive home the potential that AI technologies can offer.

TERMINOLOGY

Artificial Intelligence
Umbrella term to represent any program that can sense, reason, act and adapt.

Machine learning
Learning algorithms build a model from data, which they can improve on as they are exposed to more data over time.

Deep learning
A subset of machine learning in which multi-layered neural networks learn from vast amounts of data.

Find out more

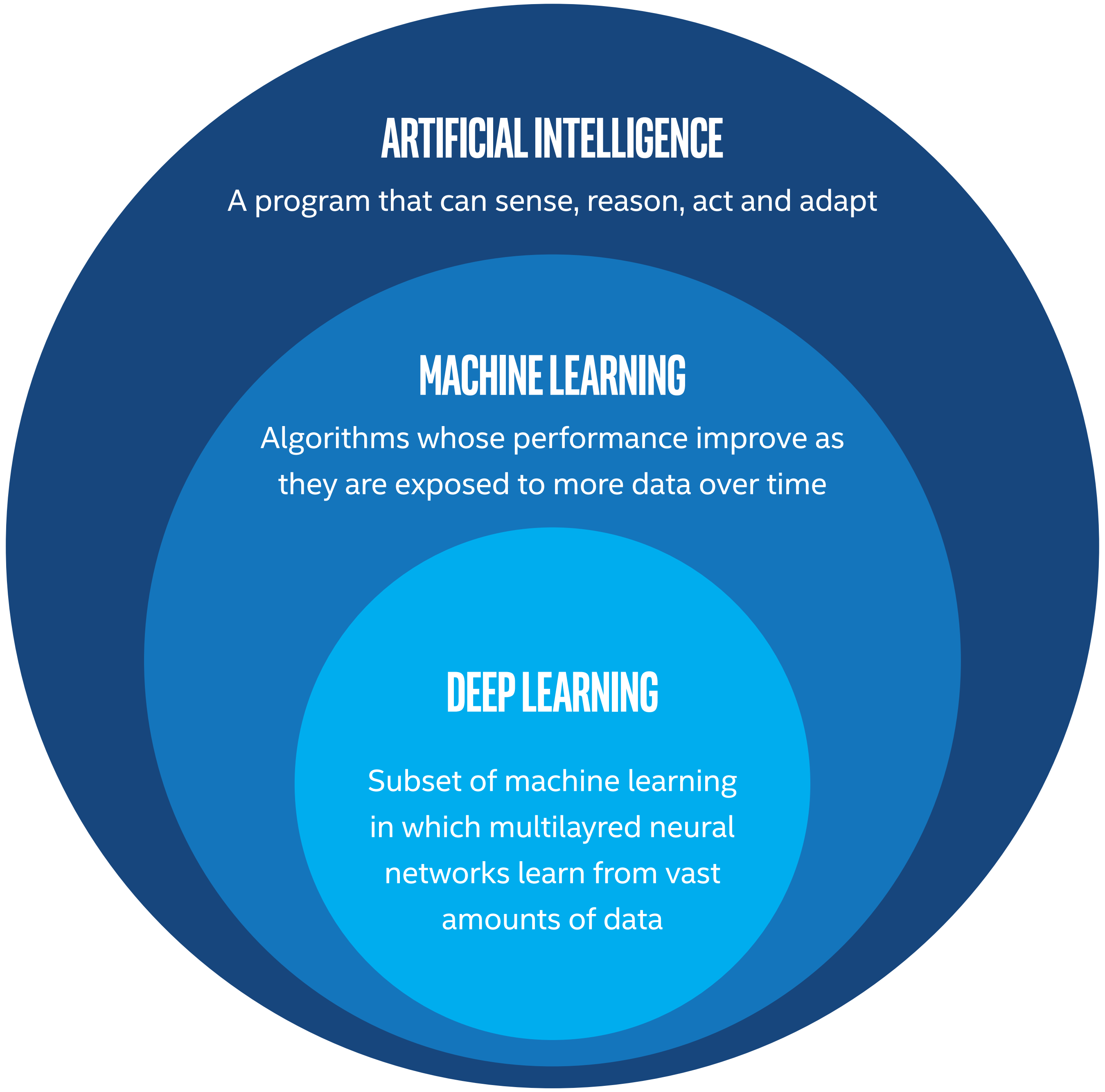

ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayred neural networks learn from vast amounts of data

Figure 1: Understanding artificial intelligence, machine learning and deep learning.

CURRENT LANDSCAPE: WHERE ARE WE NOW?

While we are beginning to understand the transformative power that AI brings, it is primarily only the largest organizations that have heavily invested in AI to date[1]. Most medium-sized and smaller organizations are still at the very early stages of implementing machine learning or deep learning technologies, if they have begun at all[2].

The enthusiasm for AI is well founded. As well as enabling us to automate processes at scale and without human intervention, AI is now helping us to achieve things that were impossible before. For example, computers can determine a healthy medical scan from one that contains a tumor faster than a human, dealing with more images per minute than a medical expert could. And AI systems are able to observe data at scale that no human or group of humans ever could, such as evaluating health data across the entire population to predict and prevent life-threatening illnesses. In this latter example AI is adding something genuinely new, not taking the job of a human or humans but performing a task that lies outside of human capabilities. We are likely to see many more of these use cases emerge over the coming years. AI is not a single technology event – our use of it will evolve and develop over time.

The current emphasis for most enterprises and organizations developing AI systems is on machine learning (ML), where machines are trained to develop the skills required to facilitate decision making, such as learning how to recognize patterns in images or other data. The primary growth area, however, is a subset of machine learning, known as deep learning (DL). DL allows developers to use sensing data to create services that add a layer of intelligence to applications, for example recognizing audio input to allow users of an application to use voice commands (e.g. Amazon Alexa*). Autonomous vehicles are another area in which DL is making strides. The process that allows a machine to control a car is extremely complex, requiring a repeating loop of:

1. Sensing the world state, including the position and speed of other cars, rules on this part of the road, pedestrians and obstacles, the direction required to continue the journey and so on
2. Planning and calculating, making a decision about the optimum next move
3. Taking the action required to proceed. Once the action has been taken, the world state has changed as a result and so the machine's sensing process must immediately begin again (see Figure 2).

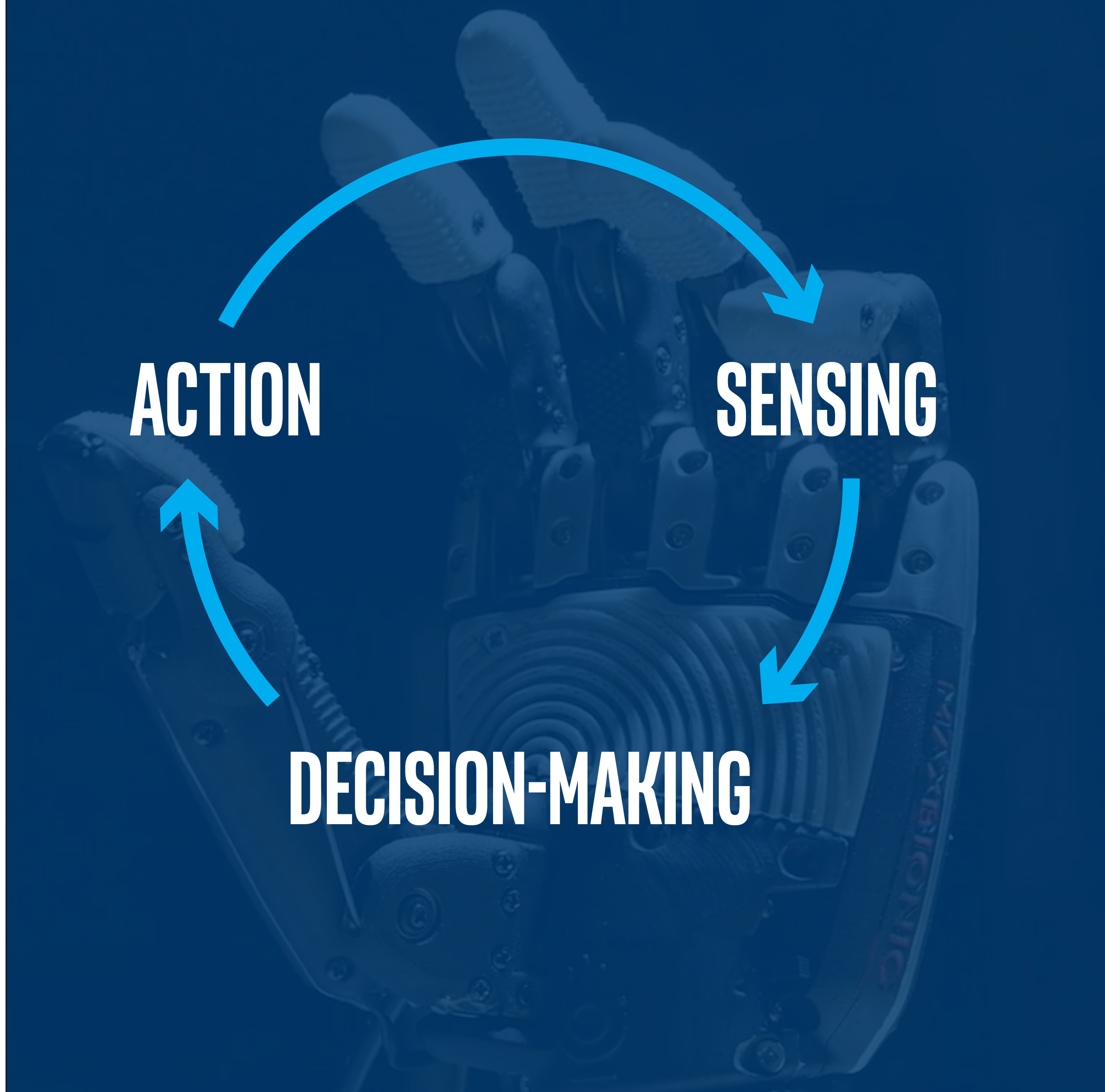


Figure 2: The process that allows a machine to control a car requires a loop of sense perception, decision-making and action.

WHAT ARE WE MISSING?

AI systems have developed rapidly over the past few years and some areas such as autonomous vehicles have seen early success. In other places, however, we have some distance to go before AI technologies go mainstream. More sophisticated sensing technologies are still being perfected to deal with complex and fast-changing environments, for example, and we do not yet have all the data in place to train the AI systems we plan to use. There are also issues to resolve outside of the technology itself, for example the fact that we cannot always explain the decisions made and actions taken by AI. This may be a feature required in legal circumstances, such as if an automated vehicle is involved in a traffic collision.

WHAT'S NEXT?

The possibilities for AI are huge and could potentially touch every area of our lives. In many ways, defining the scope for AI is like trying to define the scope for traditional software – its uses are potentially limitless. In macro terms, the current focus for AI developers is on pattern recognition – for example, image recognition or audio recognition. Perfecting these technologies is enabling use cases from audio assistants such as Amazon Alexa, Microsoft Cortana*, and Apple Siri* to visual search capabilities such as looking for criminal suspects in security footage.

A newer opportunity for AI is in some respects the opposite of pattern recognition: anomaly detection. The potential for AI-based anomaly detection is far greater and could have much more impact: for example, detecting an anomalous heartbeat that could provide early warning of cardiac arrest or other health problems or modelling the normal network behaviors in order to detect and flag malicious network behaviour of criminals. Detecting anomalous behavior in people or events is challenging because by definition, there is no or very little data on the anomaly to begin with. Observing what data there is at scale is outside of human capabilities, so it gives us a hint of the real power and opportunity for AI to bring net new benefits to society.

WHERE CAN CSPS ADD VALUE?

AI presents a number of opportunities for cloud service providers, both for improving services and data center processes, and in terms of new AI-based services to offer customers, including:

Help customers navigate their AI challenges

70 percent of enterprises expect to implement AI in 2018[3] but 91 percent anticipate significant barriers to adoption[4]. In other words, your customers' commitment to investing in AI is clear but many need more clarity on what they can achieve and how to navigate some of the technological and cultural barriers it presents. The good news is that the most commonly encountered barriers to AI adoption are things that CSPs can help their customers to overcome: a lack of the right IT infrastructure and limited access to the personnel and understanding required to develop AI solutions.

Support customers with organizing their data

Often enterprises require technical support in their development of AI, especially around cleaning and organizing the data that their machine learning and deep learning algorithms use. CSPs able to help their customers with this piece of the puzzle could find this a profitable way to build out their AI services, that also helps emphasize their credentials as a trusted advisor.

Aggregate customers' edge device and sensor data where required

Many use cases for AI, such as autonomous driving, involve data collected by sensors and in some cases processed by edge devices: edge computing rather than cloud computing. Without the latency associated with transferring all of the sensor data for processing in a central cloud, the AI system can make decisions faster. This is critical in applications such as autonomous driving, where the speed of decisions is paramount. However, retaining the data on that decision making is still important for the development of future systems. By aggregating this sensor and decision-making data, CSPs can create a foundation from which to offer value-added services, such as blending it with other data sources and applying big data analytics to deliver deep insights back to the customer.

Store the data that customers use to train their AI

Massive quantities of data may be required to train some machine learning and deep learning algorithms. Many organizations are not able to store this amount of data in a cost-effective way, so it can take them a long time to train the algorithms that they want to develop. Infrastructure-as-a-Service (IaaS) providers are perfectly positioned to occupy this space, offering data storage solutions specifically tailored to AI training needs.

Work with customers to create the algorithms they need

It's possible that your customers want to make strides in AI and know roughly what they want to do but don't yet have the capacity or expertise to build their own AI capabilities. Offering AI development solutions, such as algorithm creation, could be a profitable route for cloud service providers. Larger CSPs including Amazon and Google already have algorithms and tools available to customers and developers, so this type of offering may not be unique. It is, however, a compelling value-added service for your existing and potential customers who might require a more tailored solution. The opportunity here is to create a model that more closely matches the needs of your customer base.

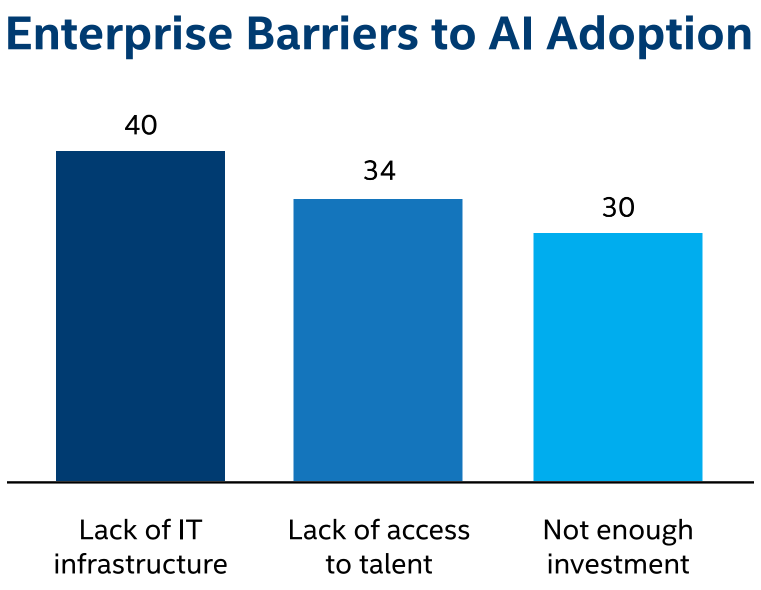
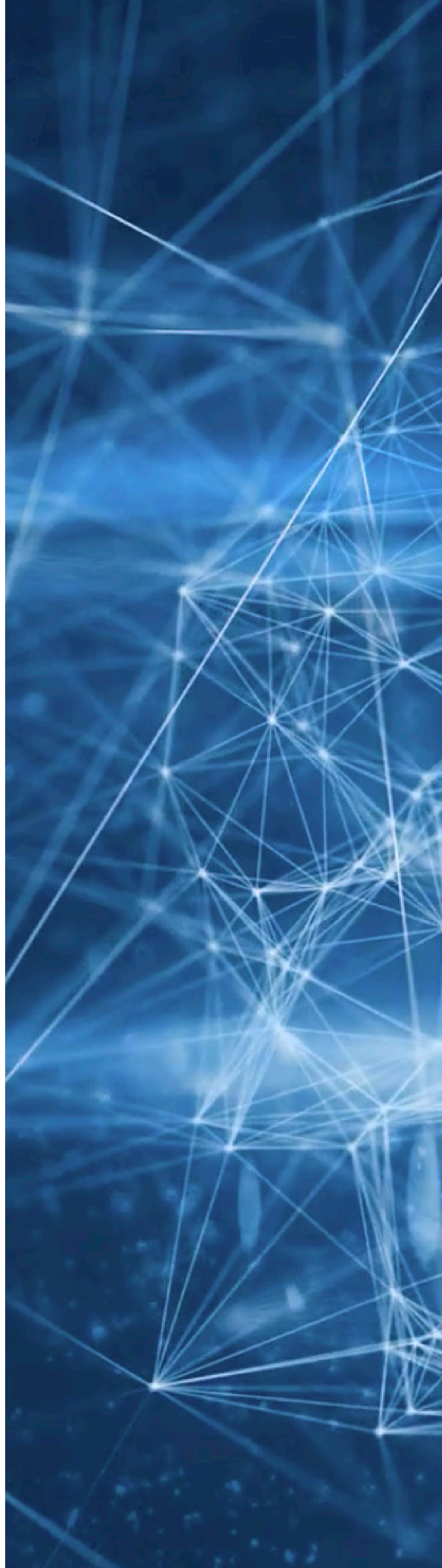


Figure 3. Top barriers to enterprise adoption of AI technologies (Source: Teradata)[5]



WHERE CAN CSPS ADD VALUE?

Offer continually evolving AIaaS services

Hyperscale CSPs, including Google, Amazon, Microsoft and IBM were quick to realize the opportunity to monetize their investments in AI and machine learning. Over the past few years, they have transformed the technologies they developed for improving their services into new offerings for their customer base.

Each of these market leaders has taken a different approach to Artificial Intelligence-as-a-Service (AIaaS), however. While Google has wrapped and resold its ML technologies behind Google Image Search (visual recognition) and Google Translate (audio recognition), IBM has taken a sector-specific approach, targeting industries such as healthcare and retail to apply its Watson-branded cognitive computing technologies to particular industry pain points[6]. The diversity of approaches among the world's largest cloud providers hints at the different opportunities available for other CSPs to offer a unique brand of AIaaS to their customers.

The potential of AI to help solve some of the biggest challenges in society and business means that there are many possible areas of focus for CSPs. To help identify an area to begin with AIaaS, some initial questions to ask include:

- Which machine learning, deep learning or other AI deployments in your business could be modified to provide the foundation of an AIaaS offering for customers?
- What type of AI solutions might your customer base find most useful? For example, does your business serve developers who might benefit from an AI test environment? Do you have a large proportion of healthcare organizations that might find a straightforward image classification system most useful? How many of your customers have call centers and therefore might be interested in an AI speech recognition system?
- Does your organization have capabilities in a broadly applicable area of AI, such as predictive analytics? Could you look to offer this as a service across industry sectors?
- How are organizations similar to your customers using AI right now? Learning how organizations similar to your customers are using AI and machine learning technologies can help you to establish a path forward for your own customers. Whether you serve the healthcare, industrial, financial services, agriculture, automotive, app development, research or something else entirely, there are plenty of use cases that can help inform your decision about where to start.

FIND INSPIRATION

See how Microsoft is using Intel® Xeon® Scalable processors with Intel® FPGAs to solve the challenges associated with AI, including accelerating deep neural networks.

Watch the video

TECHNOLOGY CONSIDERATIONS

Once you've identified your business's path to AI, the next step is to establish the changes your infrastructure will require to support AI workloads. Machine learning and deep learning processes often make intensive demands on processing, storage and power consumption but most use cases do not require a completely new infrastructure. Establishing how to blend new technology and your existing systems to achieve the best possible results with AI at the lowest possible total cost of ownership (TCO) is the first step to success. But which elements of infrastructure performance do you need to pay particular attention to? We've outlined the priority considerations below.

1. SCALABILITY

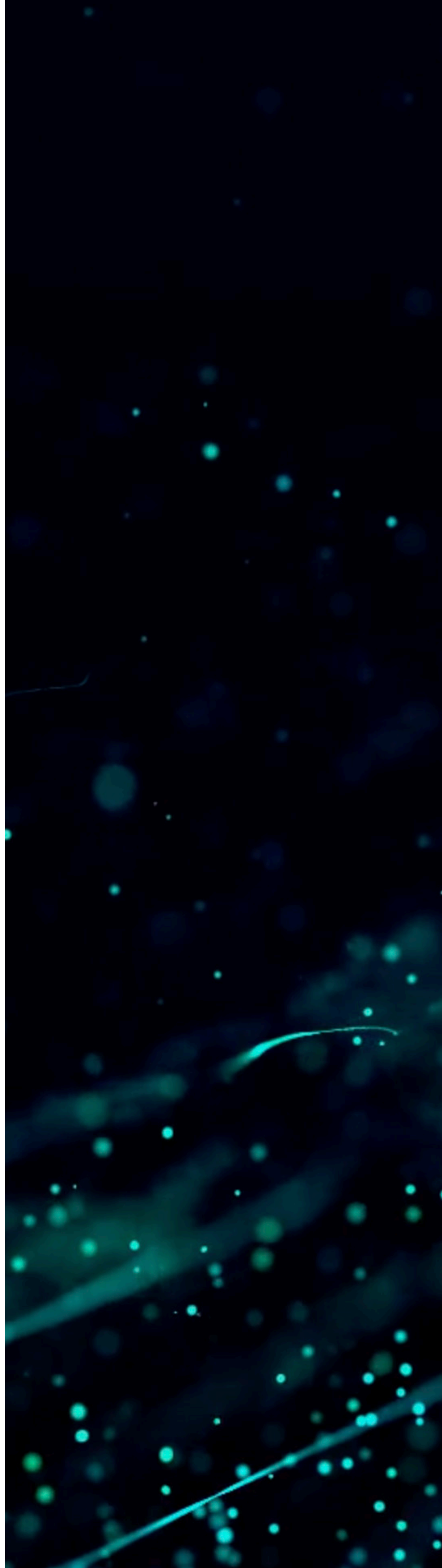
However innovative your cloud AI services, you will need to ensure that your prices are competitive if you are to convince your customers to invest in the AI solutions you are offering. This means that you will need to offer them as cost effectively as possible, utilizing as much of your existing infrastructure as you possibly can. Deploying new systems specifically to handle AI workloads can be expensive in terms of capital expenditure and power consumption.

The good news is that most existing Intel architecture-based virtualized infrastructures can support common machine learning and artificial intelligence processes. The trick is in maximizing the scalability of your existing infrastructure to ensure it can scale out to support the machine learning algorithms you want to run. This is where the Intel® Xeon® Scalable processors can help, offering up to 2.2x faster AI/deep learning training versus previous generation technologies[7] and up to 65 percent lower total cost of ownership[8]. Updating your infrastructure to maximize scalability while reducing total cost of ownership will provide the foundation you need for delivering efficient and cost-effective AI workloads.

2. FRAMEWORKS

Before starting to build AI solutions, it is important to select the best framework for your use case, purposes and in-house skillset (see Table 1). Whereas some frameworks are more suited to machine learning tasks such as visual classification, others are deep learning-specific. Some have plenty of existing pre-trained models, making them easier to get started with, whereas others may have a steeper learning curve. It may even be that your development team has an existing preference for a particular tool based on the language or platform that they are most comfortable with.

Crucially, however, Intel has optimized each of the frameworks in Table 1 for Intel® Xeon® technology, resulting in performance improvements in orders of magnitude in some cases[9]. Using Intel Xeon processors can provide over 100x performance increase with the Intel MKL-DNN library[10,11]. For example, inference across all available CPU cores on AlexNet*, GoogleNet* v1, ResNet-50*, and GoogleNet v3 with Apache MXNet on the Intel Xeon processor E5-2666 v3 (c4.8xlarge AWS* EC2* instance) can provide 111x, 109x, 66x, and 92x, respectively, higher throughput[10,11] . Figures 1 and 2 compare the training and inference throughput, respectively, of the Intel Xeon processor E5-2699 v4 and the Intel Scalable Platinum 8180 processor with the Intel MKL-DNN library for TensorFlow and Intel Optimized Caffe. The performance of these and other frameworks is expected to improve with further optimizations.



TECHNOLOGY CONSIDERATIONS

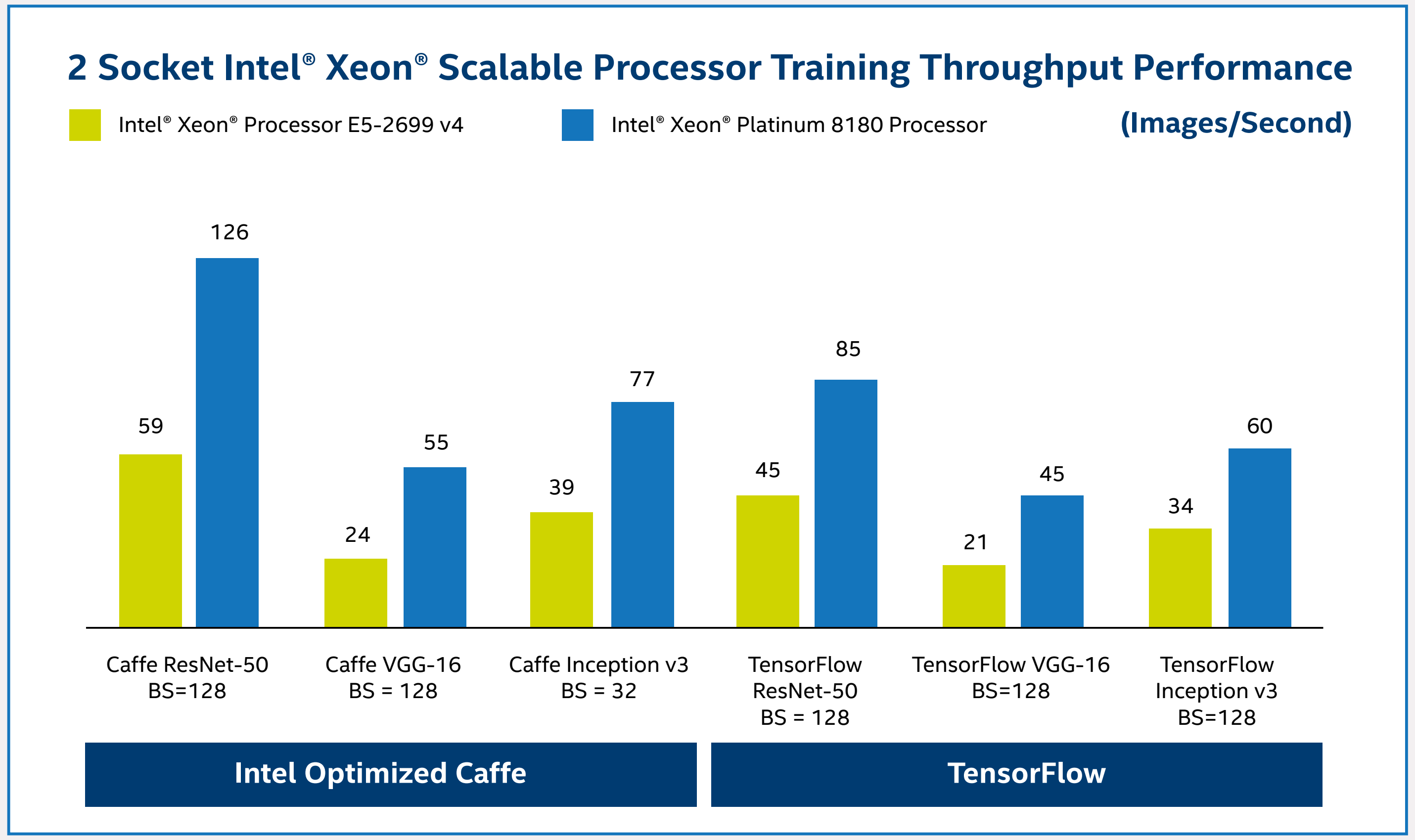


Figure 4. Training throughput of Intel Optimized Caffe and TensorFlow across Intel Xeon processor v4 (formerly codename Broadwell)h (light blue) and Intel Xeon Scalable processor (formerly codename Skylake)j (dark blue) with ResNet-50, VGG-16 and Inception-v3 with various mini-batch sizes (BS). Intel® MKL-DNN provides significant performance gains starting with the Intel Xeon processors v4 with AVX-2 instructions and a significant jump with the Intel Xeon Scalable processors when AVX-512 instructions are introduced.

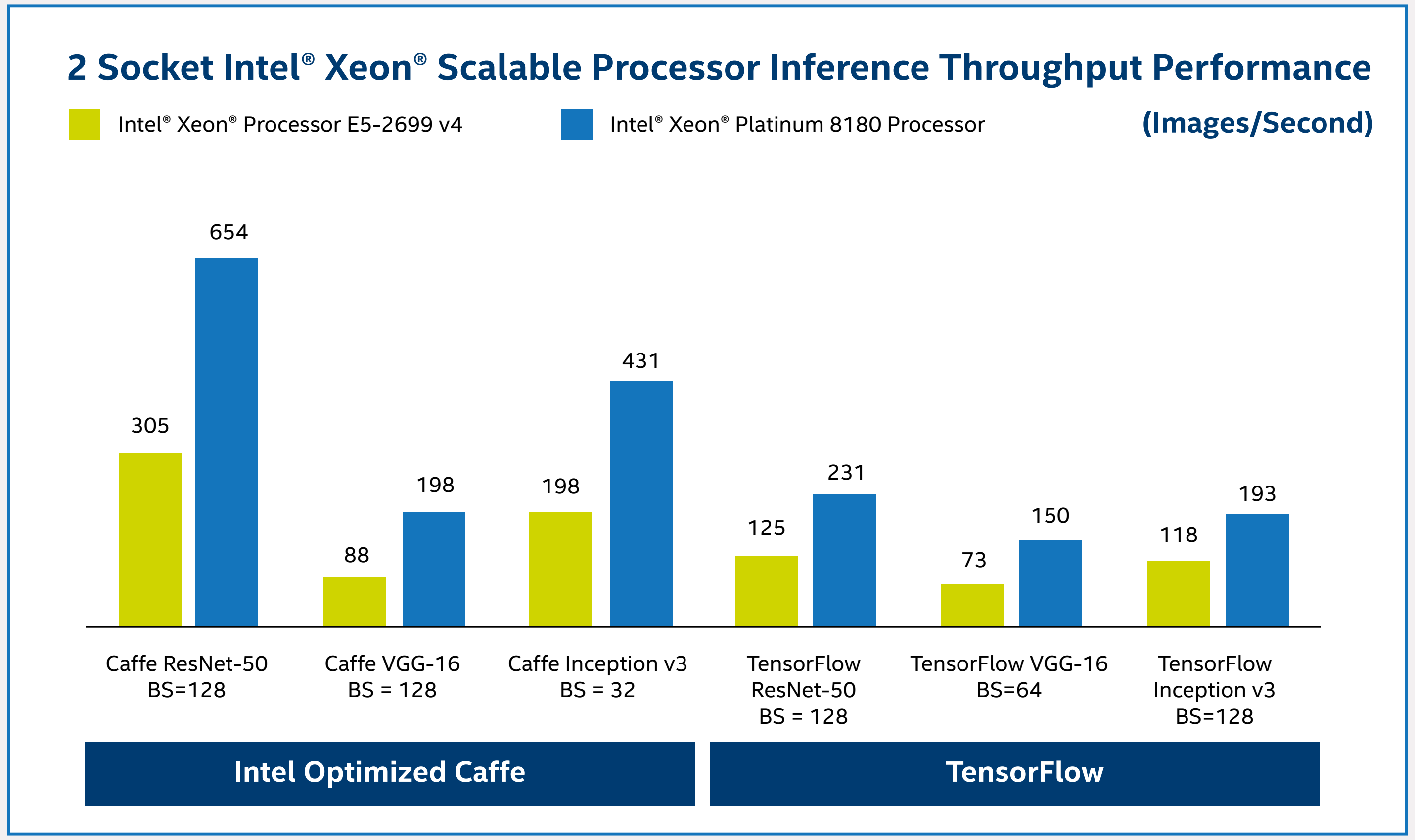


Figure 5. Inference throughput of Intel Optimized Caffe and TensorFlow across Intel Xeon processor v4 (formerly codename Broadwell)h (light blue) and Intel Xeon Scalable processor (formerly codename Skylake)j (dark blue) with ResNet-50, VGG-16 and Inception-v3 with various mini-batch sizes (BS). Inference was computed with fp32 precision. Lower precision can improve performance.

By using optimized frameworks on Intel® Xeon® technology, you can benefit from Intel's deep investment in performance improvements to achieve better results on your AI projects from the outset. As you continue to modernize your infrastructure by upgrading to newer Intel® architecture, you increasingly benefit from the performance improvements that the optimizations for newer processor technology deliver.

Framework	Use Case	Programming Language	Pros	Cons
TensorFlow*	Computation using data flow graphs for scalable machine learning. Supports convolutional neural networks and recurrent neural networks	Python C++	• Well documented • Uses computational graph abstraction • Visualization capabilities via TensorBoard make it easy to understand, debug and optimize programs	• Can be slow • Few pre-trained models • Not fully open source
Theano*	Numerical computation library. Allows users to define, optimize and evaluate mathematical expressions on arrays and tensors.	Python	• Optimized for CPU and GPU • Efficient with numerical tasks • Higher-level spin-off frameworks • Lots of sample code and tutorials	• Limited compared to other libraries • Considered difficult to use – cryptic error messages • No major developments after November 2017
Caffe*	Fast, open framework for deep learning. Effective for building convolutional neural networks (CNNs) for image classification.	C++	• Train models without writing code • High performance • Available bindings for Python and MATLAB	• Ineffective for recurrent networks • Lower performance with new architectures • Generalist
Caffe2*	Open source deep learning framework created by Facebook.	Python C++	• Built for expression, speed and modularity • Train large machine learning models and deliver AI on mobile devices	• Less support and fewer services and functions than some other frameworks
Torch*	Scientific and numerical operations. Deep learning research. Recurrent neural network and convolutional neural network modelling.	C	• Easy-to-use modular front end • Fast and efficient • Large ecosystem of community-driven packages and many pre-trained models available	• Lack of quality documentation • Limited plug-and-play code for immediate use • Based on Lua
Microsoft CNTK*	Computation networks, learning algorithms and model descriptions	C++	• Flexible and relatively fast • Allows for distributed training • Supports C++, C#, Python and Java • GPU support	• Implemented in Network Description Language (NDL) – a new language • Lack of visualizations
Apache Spark* MLlib	Scalable machine learning library. Classification, regression, clustering. Processing large data volumes.	Scala	• Usable in Java, Scala, Python and R • Fast over large data volumes	• Challenging to learn at the beginning • Plug-and-play only available for Hadoop
MXnet*	Modern deep learning framework with auto-parallelization features. Allows non-experts to design, train and re-use deep neural networks.	Python	• Easy to deploy large-scale deep learning on AWS cloud • Pre-defined CloudFormation templates • Possible to migrate from Tensor Flow • Supports many programming languages	• Smaller user base • Potentially more challenging to debug
Neon*	Deep learning framework designed for modern deep neural networks, including recurrent and convolutional neural networks, as well as Long Short-Term Memory Models and Deep Autoencoders		• Easy to use • Extensible on AlexNet, Visual Geometry Group (VGG) and GoogLeNet • Tightly integrated with Intel GPU kernel library • Fast	• Cannot be configured to use multiple CPU/GPU threads

Table 1. Considerations for selecting an AI framework[12]

TECHNOLOGY CONSIDERATIONS

3. STORAGE AND BIG DATA INGESTION

Machine learning and deep learning algorithms require a lot of capacity to store the data they need to train on, so your storage systems will need to be an area of particular focus prior to embarking on your AI journey. Whether you are undertaking machine learning as part of your business processes or running the algorithms for your customers, you will almost certainly need to work with huge pools of data from different sources: structured and unstructured data will need to be brought together and processed, and streamed data must be pooled with existing data.

Depending on the AI use cases your business is focusing on, you may need to retain exponentially increasing amounts of data – after all, it will not be clear which is the crucial piece of information until analysis is complete. To establish the right storage environment for your particular AI use case:

- Take time to determine the storage demands your AI use case will create. Work with all the AI project stakeholders to understand: What data will your machine learning algorithms need to train on? Where will this data come from? How will you pool it? Do you need to create a data lake? Does all of this storage need to be equally quickly accessed?
- Identify the storage resources in your existing infrastructure that can form part of the solution, as well as the type and volume of new capacity you will require for effectively delivering your AI workloads.
- Investigate whether and how your storage resources will need to be tiered to run machine learning algorithms most effectively at the lowest possible TCO. If you do need to tier your storage resources, you may want to investigate methods of automatic tiering to reduce the storage management burden that machine learning can present.

4. PERFORMANCE AND SPEED

Fast, efficient processing is extremely important for all AI instances but there are several ways to approach meeting these demands, depending on the specific activity. China-based CSP, JD.com, was able to speed up its image detection and extraction solution more than three times by switching to BigDL for distributed deep learning using Apache Spark* on Intel® Xeon® clusters[13].

A key element in achieving the speed and performance required for AI is ensuring that your processing engine has fast access to the storage and networking interfaces. I/O can be a critical bottleneck inside the data center, particularly in AI instances where the flow of information is so continuous. Make sure your I/O hardware can keep up with the demands of the AI workloads you expect to run.

JD.COM SPEEDS UP IMAGE DETECTION WITH INTEL® XEON® PROCESSORS

JD.com, a major Chinese e-commerce platform, upgraded its previous GPU-based image detection and extraction solution to BigDL for distributed deep learning using Apache Spark on Intel® Xeon® clusters. The switch brought ~3.83X speedup vs. GPU.[1] **Find out more**

[1] **Platform: 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).

MEITUAN LAUNCHES AIAAS WITH HELP FROM INTEL

Like many cloud service providers, Meituan was using GPUs for its image classification and AI workloads. By switching to servers based on the Intel® Xeon Phi™ processor for its data center expansion, the company was able to increase its performance and lower its hardware costs. Intel has worked closely with Meituan for several years, and guided the company through the migration, including co-developing software and offering support with optimizing its performance. **Find out more**

CHINA UNIONPAY IMPROVES RISK CONTROL WITH NEURAL NETWORKS

China UnionPay, an international financial institution specializing in banking services and payment systems, handles up to 20 billion payments every year across emerging channels such as mobile, online and social media. It has used AI to improve the accuracy of its risk control by moving from a rules-based model to a neural-network model based on Apache Spark* and powered by Intel technology. **Find out more**

TECHNOLOGY CONSIDERATIONS

5. HARDWARE HOMOGENEITY

AI and machine learning workloads may benefit from accelerated computing, for example offloading a portion of the processing to Field Programmable Gate Arrays (FPGAs) or GPUs[14]. FPGAs are often used for implementing neural networks as they can handle different algorithms in computing, logic, and memory resources in the same device. Faster performance is achieved by hardcoding operations into the hardware[15].

Acceleration technologies can, however, make compute far less homogenous which may create additional work for CSPs in terms of management, maintenance and security of the total systems. Ensuring you have homogeneity of hardware, as opposed to having to maintain separate clusters for accelerators, will make AI workloads far easier to run with benefits including simplified management, easier patch, repair and replacement, and more straightforward security controls.

6. NEW AI TECHNOLOGIES

Technologies in the field of AI are developing at a rapid rate and hardware and software designed to deal with the specific challenges that machine learning and deep learning technologies present is emerging regularly. Last year Intel announced that it will be launching the Intel® Nervana™ Neural Network processor (Intel® Nervana™ NNP), designed to meet the needs of speed, numerical parallelism and scalability that AI use cases demand. The architecture has been built for neural networks from the ground up, supporting all deep learning primitives while making the core components as efficient as possible. Technologies such as the Intel Nervana NNP will allow those investing in AI to improve the power and performance of their AI systems while simplifying their deployment. Staying alert to AI-specific technologies such as this will ensure that you can quickly ramp up your AI projects and products, accelerating the time to market for deploying new services for your customers, helping you to stand out from your competitors.

CONCLUSION & NEXT STEPS

There is a wealth of opportunity for CSPs in AI but taking advantage of it will depend on having the right infrastructure in place to support the new types of workloads. It may be important for CSPs to build out their AI capabilities and services in line with their existing expertise, to build trust with their customers and facilitate early success. The next steps for CSPs looking to grow their business with AI are:

Research

Gather inspiration and best practice what your customers and competitors are doing in AI – you can start with the case studies and further reading on Intel in AI and the Intel® AI Academy for Developers.

Build your team

Ensure that all of your key stakeholders across business functions participate in your AI project. Product/service development, line of business heads, IT, dev, operations and leadership will all need to input for a project to be successful.

Evaluate your existing IT systems

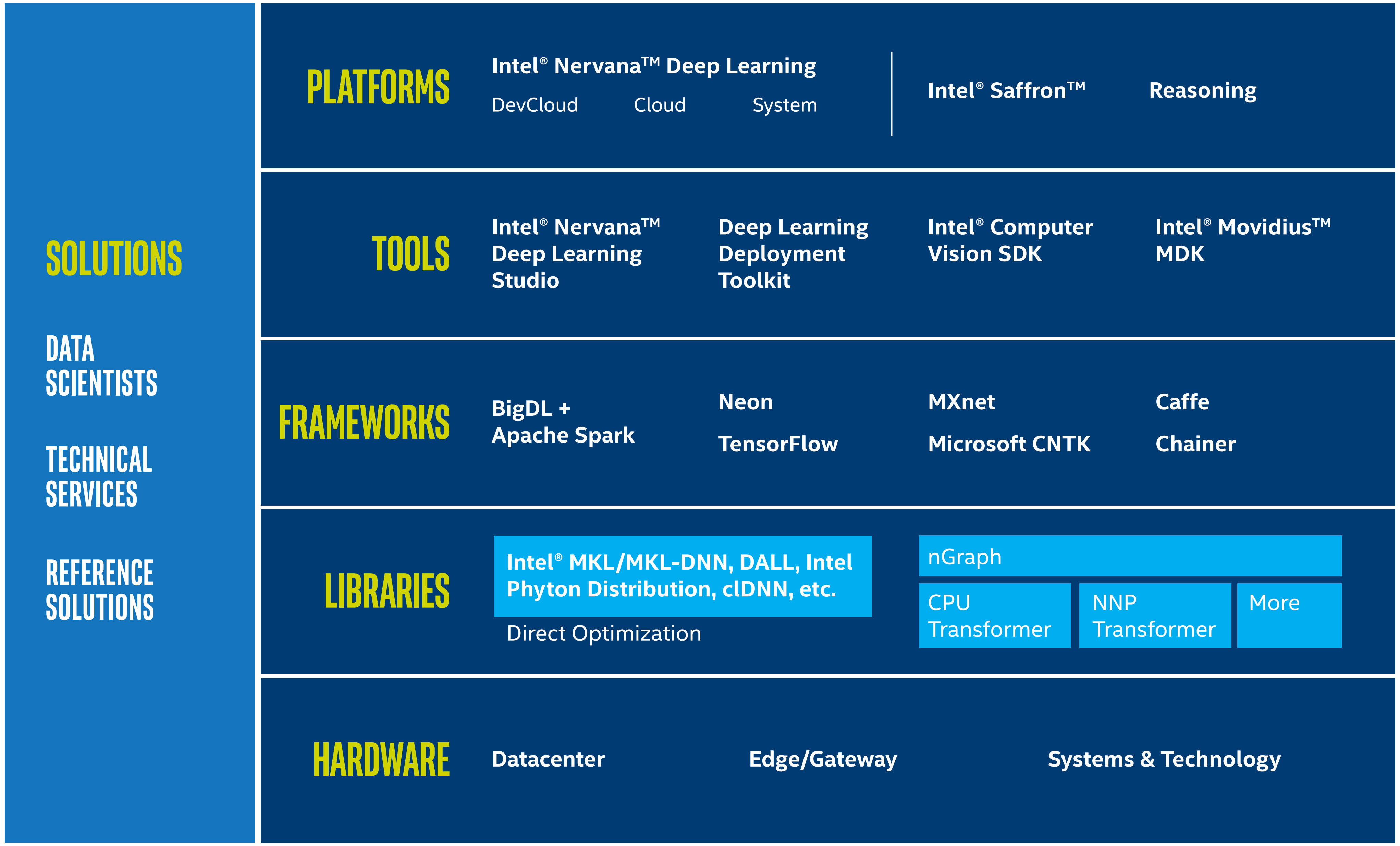
Use the Technology Considerations section above as a starting point for establishing the suitability of your infrastructure and processes, and identifying candidates for upgrade or improvement.

Discover more resources for CSPs

Visit intel.com/csp

INTEL IN AI

The Intel® AI portfolio features broadly applicable and flexible solutions for every business need.



SOLUTIONS

DATA SCIENTISTS

TECHNICAL SERVICES

REFERENCE SOLUTIONS

PLATFORMS

Intel® Nervana™ Deep Learning

DevCloud

Cloud

System

Intel® Saffron™

Reasoning

TOOLS

Intel® Nervana™ Deep Learning Studio

Deep Learning Deployment Toolkit

Intel® Computer Vision SDK

Intel® Movidius™ MDK

FRAMEWORKS

BigDL + Apache Spark

Neon

TensorFlow

MXnet

Microsoft CNTK

Caffe

Chainer

LIBRARIES

Intel® MKL/MKL-DNN, DALL, Intel Phyton Distribution, clDNN, etc.

Direct Optimization

nGraph

CPU Transformer

NNP Transformer

More

HARDWARE

Datacenter

Edge/Gateway

Systems & Technology

Figure 6. the Intel® AI portfolio

FURTHER READING

- Meituan builds AI on IA: from enhanced customer experience to differentiated cloud services
- Intel® AI Academy for developers building AI solutions
- Intel Resources for AI professionals
- Intel Resources for Cloud Service Providers

1 McKinsey (2017) Artificial Intelligence: The Next Digital Frontier? https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx

2 McKinsey (2017) Artificial Intelligence: The Next Digital Frontier? https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx

3 Forbes (2017) '10 Predictions for AI, Big Data and Analytics in 2018'

4 Teradata (2017) 'State of Artificial Intelligence for Enterprises'

5 Teradata (2017) State of Artificial Intelligence for Enterprises, http://assets.teradata.com/resourceCenter/downloads/AnalystReports/Teradata_Report_AI.pdf

6 Fast Company (2017) 'How Amazon, Google, Microsoft, And IBM Sell AI As A Service, https://www.fastcompany.com/40474593/how-amazon-google-microsoft-and-ibm-sell-ai-as-a-service

7 Intel® Xeon® Platinum 8180 processor compared to Intel® Xeon® processor E5-2699 v4 NOTE: 113x gain in last 2 years, using optimized frameworks & optimized Intel® MKL Libraries compared to Intel® Xeon® processor E5-2699 v3 with BVLC-Caffe Platform: 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

Deep Learning Frameworks: 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, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs fromhttps://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), andhttps://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

Platform: 2S Intel® Xeon® CPU E5-2697 v2 @ 2.70GHz (12 cores), HT enabled, turbo enabled, scaling governor set to "performance" via intel_pstate driver, 256GB DDR3-1600 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.21.1.el7.x86_64. SSD: Intel® SSD 520 Series 240GB, 2.5in SATA 6Gb/s, 25nm, MLC.

Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact,1,0', OMP_NUM_THREADS=24, CPU Freq set with cpupower frequency-set -d 2.7G -u 3.5G -g performance

Deep Learning Frameworks: Caffe*: (http://github.com/intel/caffe/), revision b0ef3236528a2c7d2988f249d347d5fdae831236. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs fromhttps://github.com/intel/caffe/tree/master/models/intel_optimized_models (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), andhttps://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). GCC 4.8.5, Intel MKL small libraries version 2017.0.2.20170110.

8 Up to 65% lower 4-year TCO estimate example based on equivalent rack performance using VMware ESXi* virtualized consolidation workload comparing 20 installed 2-socket servers with Intel® Xeon® processor E5-2690 (formerly "Sandy Bridge-EP") running VMware ESXi* 6.0 GA using Guest OS RHEL* 6.4 compared at a total cost of $919,362 to 5 new Intel® Xeon® Platinum 8180 (Skylake) running VMware ESXi* 6.0 U3 GA using Guest OS RHEL* 6 64 bit at a total cost of $320,879 including basic acquisition. Server pricing assumptions based on current OEM retail published pricing for 2-socket server with Intel® Xeon® processor E5-2690 v4 and 2 CPUs in 4–socket server using E7-8890 v4 – subject to change based on actual pricing of systems offered.

9 Intel (2017) TensorFlow* Optimizations on Modern Intel® Architecture https://software.intel.com/en-us/articles/tensorflow-optimizations-on-modern-intel-architecture

10 Platform: 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, dummy 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 (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl -l".

TensorFlow: (https://github.com/tensorflow/tensorflow), commit id 207203253b6f8ea5e938a512798429f91d5b4e7e. Performance numbers were obtained for three convnet benchmarks: alexnet, googlenetv1, vgg (https://github.com/soumith/convnet-benchmarks/tree/master/tensorflow) using dummy data. GCC 4.8.5, Intel MKL small libraries version 2018.0.20170425, interop parallelism threads set to 1 for alexnet, vgg benchmarks, 2 for googlenet benchmarks, intra op parallelism threads set to 56, data format used is NCHW, KMP_BLOCKTIME set to 1 for googlenet and vgg benchmarks, 30 for the alexnet benchmark. Inference measured with --caffe time -forward_only -engine MKL2017option, training measured with --forward_backward_only option.

MxNet: (https://github.com/dmlc/mxnet/), revision 5efd91a71f36fea483e882b0358c8d46b5a7aa20. Dummy data was used. Inference was measured with "benchmark_score.py", training was measured with a modified version of benchmark_score.py which also runs backward propagation. Topology specs from https://github.com/dmlc/mxnet/tree/master/example/image-classification/symbols. GCC 4.8.5, Intel MKL small libraries version 2018.0.20170425.

Neon: Internal-version. Dummy data was used. The main.py script was used for benchmarking in mkl mode. ICC version used : 17.0.3 20170404, Intel MKL small libraries version 2018.0.20170425. Latest version neon results: https://www.intelnervana.com/neon-v2-3-0-significant-performance-boost-for-deep-speech-2-and-vgg-models/

11 Platform: 2S Intel Xeon CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel_pstate driver, 256GB DDR4-2133 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.el7.x86_64. OS drive: Seagate* Enterprise ST2000NX0253 2 TB 2.5" Internal Hard Drive.

Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact,1,0', OMP_NUM_THREADS=36, CPU Freq set with cpupower frequency-set -d 2.3G -u 2.3G -g performance

Intel Caffe: (http://github.com/intel/caffe/), revision b0ef3236528a2c7d2988f249d347d5fdae831236. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy 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 (GoogLeNet, AlexNet, and ResNet-50), https://github.com/intel/caffe/tree/master/models/default_vgg_19 (VGG-19), and https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). GCC 4.8.5, MKLML version 2017.0.2.20170110.

BVLC-Caffe: https://github.com/BVLC/caffe, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training BVLC Caffe (http://github.com/BVLC/caffe), revision 91b09280f5233cafc62954c98ce8bc4c204e7475 (commit date 5/14/2017). BLAS: atlas ver. 3.10.1.

12 Sources: DZone (2018) 10 Best Frameworks and Libraries for AI https://dzone.com/articles/progressive-tools10-best-frameworks-and-libraries, Towards Data Science (2017) A Survey of Deep Learning Frameworks https://towardsdatascience.com/a-survey-of-deep-learning-frameworks-43b88b11af34, Bosch Research and Technology Center (2016) Comparative Study of Deep Learning Software Frameworks https://arxiv.org/pdf/1511.06435.pdf, Microway (2016) Deep Learning Frameworks: A Survey of TensorFlow, Torch, Theano, Caffe, Neon and IBM< Machine Learning Stack https://www.microway.com/hpc-tech-tips/deep-learning-frameworks-survey-tensorflow-torch-theano-caffe-neon-ibm-machine-learning-stack/, Liu & Zang (2017) Caffe2 vs TensorFlow: Which is the better deep learning framework? http://cs242.stanford.edu/assets/projects/2017/liubaige-xzang.pdf

13 GPU: 20 * NVIDIA Tesla* K40. CPU: Intel® Xeon® processor E5-2650 v4 @ 2.20GHz, 1200 logical cores (each server has 24 physical cores with Intel® Hyper-Threading Technology (Intel® HT Technology) enabled, and is configured to support 50 logical cores in Yet Another Resource Negotiator (YARN). https://software.intel.com/en-us/articles/building-large-scale-image-feature-extraction-with-bigdl-at-jdcom

14 IDC (2017) Accelerated Computing Is Quickly Gaining Traction as Enterprises Seek to Manage Cognitive Workloads, According to IDC, https://www.idc.com/getdoc.jsp?containerId=prUS42909117

15 Intel FPGA, https://www.altera.com/solutions/industry/computer-and-storage/applications/machine-learning/machine-learning.html

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

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.

Benchmarks were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system. Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer, or learn more at www.intel.com.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer, or learn more at www.intel.com.

0318/CAT/XX/PDF 337374-001EN