Build an Edge to Cloud AI Infrastructure

Break down data siloes to deliver insights whenever and wherever they are needed most

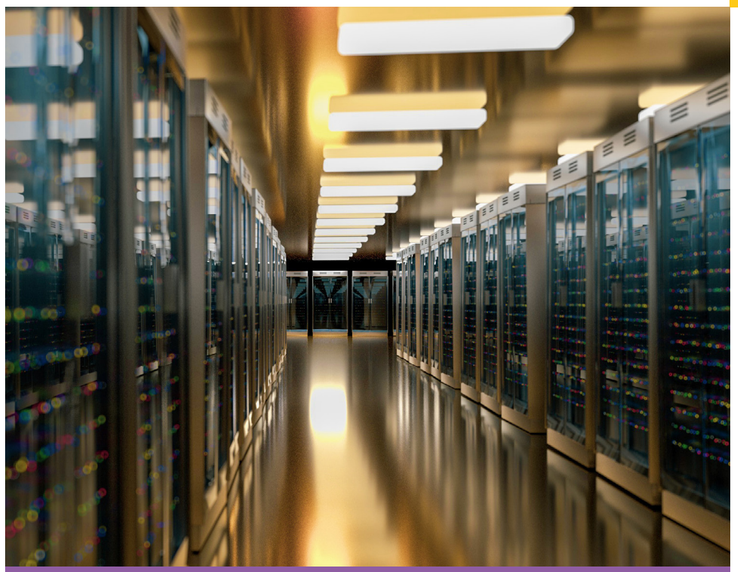
Contents

Introduction

The increasing speed of business operations paired with constantly rising customer expectations means that for many organizations, decision making must be increasingly devolved away from the head office. Whether they're dealing with customers on the shop floor, or working on a busy production line, individual employees need the insight and information to make critical decisions in real time. In some cases, these decisions may need to be entirely automated.

Increasingly, decisions are being made based on data generated at the edge. In fact, so much data is being captured at the edge that Gartner predicts that by 2025, up to 75 percent of all enterprise data will be generated outside traditional data centers[1]. Putting compute capabilities closer to this data at its point of origin enables new real-time use cases and derives fresh revenue streams from even the most sensitive data.

Achieving this relies on the effective combination of three technologies: edge computing, the cloud and artificial intelligence (AI). While all three already add value individually, enterprises that ensure they can integrate them across their infrastructure and enable full edge-to-cloud intelligence, will be best placed to succeed. We can think of them existing in a virtuous cycle. More devices and compute power at the edge create more data, which feeds more complex AI use cases. This in turn creates more insight that can be disseminated through the cloud across the organization and used to further optimize data collection and analysis in the future.



Edge-to-Cloud AI: Definitions

- **Edge:** Edge computing is the placement of resources that move, store and process data closer to the data's source or the point of service delivery. As a result, businesses can reduce latency, improve quality of experience, optimize total cost of ownership, comply with data locality requirements, and enable actionable insights.
- **Cloud:** Cloud computing gives you remote access to computing, storage, and networking resources within your data center or through a public cloud service provider.
- **AI:** Encapsulates a broad set of computer sciences designed to replicate human-like abilities, such as perception, logic, and learning. An emerging trend is the use of different AI techniques, like deep learning and reinforcement learning, in combination to make progress toward generalized intelligence.

The edge consists of all the data gathering, processing, storage and communications beyond the service's remote core (in the data center or cloud). In this paper, we will split the broader concept of the edge into two:

- Edge devices: single-user assets that generate or consume data, such as drones, wearables, smartphones, smart speakers, industrial sensors, smart cameras.
- Edge infrastructure: devices that pull multiple data streams from different sources, such as network video recorders, gateways, local servers, hyperconverged edge infrastructure (or "data center in a box").

This paper will explore the edge-to-cloud AI journey within the organization, looking at how and where AI is used, and which technologies enable it.

Use Cases for AI in the Cloud

If AI at the edge provides micro-level insights, running AI in the cloud enables organizations to create deeper intelligence at a macro level. Using the cloud, they can tap into larger datasets, often pulling data in from multiple sites across the infrastructure (or from external sources) to build a comprehensive and integrated view of trends and patterns over time.

AI in the Cloud Use Case Example: Language Recognition

Speech and text recognition (known as natural language processing) can support a wide variety of use cases, such as automated chatbots for customer service, or real-time support for contact center staff. Voice recognition software provider iFLYTEK offers voice-based cloud solutions for a range of industries.

Use Cases for AI at the Edge

AI at the edge is a key enabler of many of today's most innovative and transformational use cases across multiple industries.

Edge Device AI Use Case Example: Retail
By building in AI capabilities to devices across the store, retailers can enhance the customer experience, make more efficient use of space, and enhance inventory management. With its ability to read codes, text, and numbers, machine vision can help manage, track, and analyze inventory levels, ensuring that critical materials are in the hands of those who need them most. Connected, intelligent and responsive digital signs can suggest offers or products to customers based on their behavior and interests, and help the retailer know when their message is truly effective. Self-service kiosks or even entirely autonomous stores can provide customers with a range of services to help personalize their experience while keeping it contactless and streamlined. Meanwhile, machine learning can analyze footage of shoppers streamed to on-premises edge gateways, helping identify potentially criminal behavior in real time.

Edge Infrastructure AI Use Case Example: Healthcare
There are many potential uses for AI in healthcare, but one of the most popular is medical imaging. Thousands of medical images—such as CT scans, X-rays and MRIs—are produced daily, each requiring careful analysis to identify anomalies and achieve an accurate diagnosis. By implementing deep learning capabilities at the point of capture, Philips has been able to accelerate CT scan imaging by 188x[2], helping clinicians diagnose and treat patients more quickly.

Essentials for an Edge-to-Cloud AI Infrastructure

Many organizations today already make use of AI, although often in a relatively piecemeal way. Individual initiatives can deliver valuable results and add significant value for the business. However, taking the next step to enable AI from end-to-end across your organization can significantly accelerate and expand these advantages. It can also open up a range of new opportunities.

When implementing AI on a broader scale across the business, it is important to ensure that each of the three main elements of your infrastructure—edge devices, edge infrastructure and cloud—are equipped with the capabilities to support it. These are:

- **High performance:** AI workloads tend to be computationally intensive, so it is essential to have strong compute performance wherever AI training or inference takes place.
- **Low latency:** One of the beauties of AI is its ability to support real-time decision making. Moving some AI workloads to the edge can reduce latency across compute, storage and networking.
- **High capacity:** AI depends on large volumes of data, so any infrastructure running AI can avoid bottlenecks by ensuring compute, storage and memory capacity are up to the task.
- **Robust security:** Not only do AI workloads need large volumes of data, this data is often sensitive in nature (for example, in the healthcare or public safety industries). Whatever the AI workload, the devices and software that it runs on must be reliably secure.

Intel offers technologies and solutions that enable organizations to support AI workloads from edge to cloud while meeting these overarching requirements (see Figure 1). Intel® Edge Technology Solutions for AI enable high-performing inference on everything from on-premises servers to PCs to cameras, robots, and drones. Our portfolio of CPUs, VPUs, and FPGAs are tuned for low-latency inference to help remove data bottlenecks alongside the latest in networking, memory, and storage.

CPU *only*
For mainstream AI use cases

CPU + GPU
When compute is dominated by AI, HPC, graphics, and/or real-time media

CPU + DEDICATED
When compute is dominated by deep learning (DL)

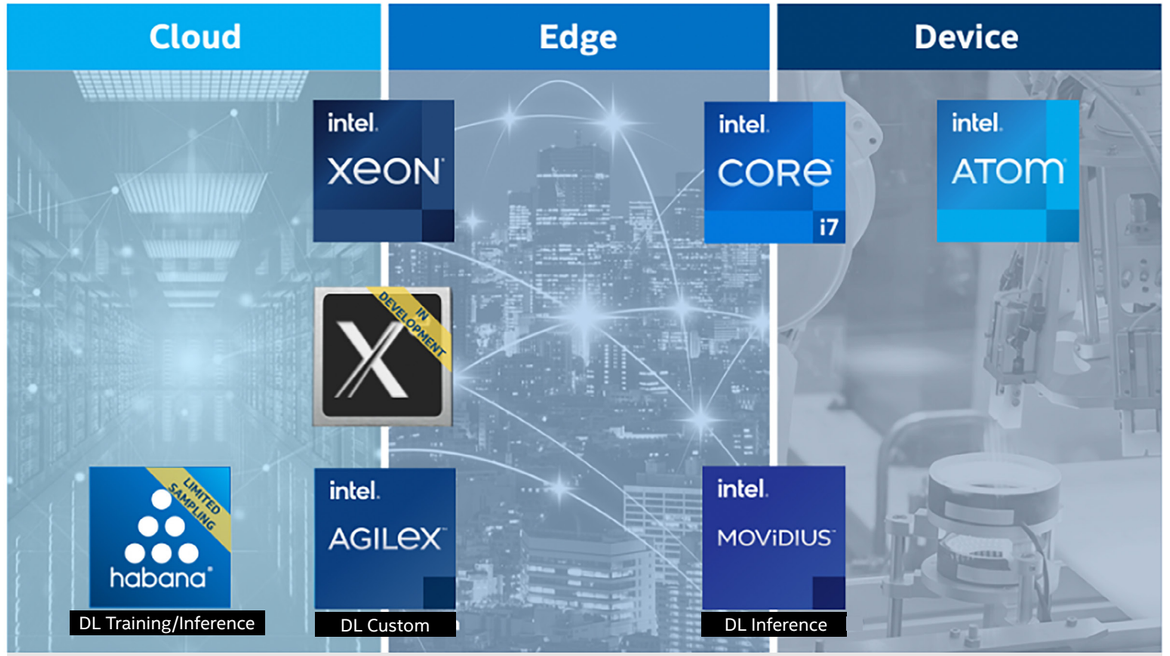


Figure 1. Intel® AI technologies support edge-to-cloud AI.

Edge Device

Edge devices are typically small (for example a smart watch or a smart camera), with little or no space for hefty components. They often need to run on a limited power supply as well, meaning any hardware must be space and power efficient. In order to support AI workloads on these devices though, they must also deliver high performance.

While edge devices can—and often do—operate entirely independently and primarily carry out AI inference workloads, in some use cases it can be beneficial to connect multiple edge devices to enable federated learning for AI training. This enables edge devices to collaboratively learn a shared prediction model while keeping all training data on the device and removing the need to store all learning data in the cloud. It also allows devices to play a role in training AI models, as the device downloads the latest model, learns from the data on the devices and then sends back its changes in a small, focused update. This update can then be encrypted and sent to the cloud where it can contribute to improving the shared model. With this federated approach, all data is kept on the device, helping maintain data security.

Use case example: Robot prototypes help fight against Covid-19

Robots are being used in the US to help sterilize surfaces in hospitals using bursts of UltraViolet (UV) light. The light is highly effective at killing off viruses but can also be harmful for humans. The robot uses AI to navigate around the hospital, and to check that a space is clear of people before it emits the UV light to sanitize large spaces quickly and thoroughly. This helps ensure safety is maintained across the hospital while keeping busy areas open for use as much as possible.

The robots use the Intel® Movidius™ VPU, a purpose-built AI accelerator for computer vision and deep learning inference. It combines computer vision, camera image processing, and AI deep learning inference into a stand-alone system on chip (SoC). This means it can be deployed independently as an edge device that enables powerful on-device recognition and computer vision analysis in real time. When deployed in combination with a host CPU at the edge or in the cloud, the Intel Movidius VPU also offers strong deep learning acceleration. This can help speed up media processing and image analysis for applications such as network video recorders that can view, process and catalog video information in real time.

Use case example: Computer vision enables automated license plate recognition

Smart cameras can be invaluable in helping to automate repetitive routine tasks in order to free up employees to focus on more complex challenges. One such example is the AI-driven license plate recognition solution developed by Intel AI Builders member, Wahtari. The cameras run on Intel Movidius VPUs and Intel® Atom® processors. Their low power consumption and high performance mean they are energy efficient while providing the compute capabilities needed to run complex AI workloads at the edge. Support for high-resolution Ultra HD 4K, high frames per second (FPS) multimedia streaming means Intel Atom processors can also offer low-power, cost-efficient support for deep learning workloads as well, when combined with an accelerator such as the Intel Movidius VPU. As a result, the Wahtari nLab AI training platform delivers AI inference at 45 frames per second and detects over 7,000 license plates an hour.

The Wahtari solution was built using the Intel® Distribution of OpenVINO™ toolkit. This convolutional neural network (CNN)-based toolkit enables the development and acceleration of applications that emulate human vision and run across Intel® hardware.

Edge Infrastructure

Hardware that supports more holistic or complex edge computing in the form of edge clusters or network servers, for example, tend to rely on components that offer higher performance than stand-alone edge devices. They may also make use of security or connectivity features as required to support their designated use case. Intel Edge Technology Solutions offer the flexibility to seamlessly add support for AI at the edge in both brownfield and greenfield environments.

Use case example: Computer vision helps ensure public safety

Computer vision and AI can help large organizations like governments and transportation companies ensure public safety and convenience. Engineering and design company Klas Telecom, for example, has developed a computer vision-based solution for the railway industry, which enables human and vehicle detection at crossroads, onboard empty seat detection, and intruder detection. This helps maintain public safety and security while enabling staff to focus on issues or risks that most need their attention.

The Klas Telecom solution uses 10th Generation Intel® Core™ i7 processors. These offer high performance to run complex edge AI workloads, while remaining power efficient. Its deep learning algorithms were also built using the Intel Distribution of OpenVINO toolkit, helping accelerate its AI workloads from edge to cloud across a heterogenous Intel® architecture, using a common API.

Use case example: Machine learning at the edge helps enhance product quality

There is huge potential for edge computing in manufacturing and industrial settings. Audi's Neckarsulm factory assembles up to 1,000 cars each day, with around 5,000 welds per car. To manually inspect millions of welds in a single day is costly, labor-intensive, and nearly impossible. Audi set a goal to inspect 100 percent of welds to an exceptional degree of accuracy. It did this using machine learning algorithms, Intel's Industrial Edge Insights software and the Nebbiolo edge platform for streaming analytics. The resulting solution automates inspection based on data from welding-gun controllers. It was able to reduce labor costs by 30 to 50 percent, freeing employees for other valuable opportunities within the company. Ultimately, the factory boosted weld inspections 100x, with 18ms latency for each weld inspection.

The solution is powered by the latest 3rd Generation Xeon Scalable processors, with built-in Intel® Deep Learning Boost, which offer embedded performance acceleration for AI workloads. They enable up to 30x performance improvement for inference workloads compared to the previous generation[3], creating an agile, robust and scalable edge foundation.

Use case example: Computer vision supports delicate conservation efforts

Smart cameras and video analytics provide a great opportunity to help monitor and protect endangered habitats, where the physical presence of conservation workers can be difficult. For instance, monitoring coral reefs typically involves human divers directly collecting data underwater or manually capturing video or images of the reef for later analysis. While reliable, these methods create the risk of divers interfering with wildlife behavior and inadvertently affecting research results. They also mean the opportunity for data capture is limited, as divers are only able to safely spend around 30 minutes underwater at a time. Project: CoRaiL in the Philippines overcomes these issues by analyzing coral reef resiliency using smart cameras and video analytics.

Data from the cameras is analyzed using Accenture's Video Analytics Services Platform, which is powered by Intel Xeon Scalable processors and Intel Movidius VPUs, using algorithms developed with the Intel Distribution of OpenVINO toolkit. The solution also uses Intel® Field Programmable Gate Arrays (Intel® FPGAs) to further accelerate their AI workloads. As blank, modifiable canvases, these components can be adapted multiple times to serve different purposes. They can be particularly valuable in high-throughput, low-latency applications like AI at the edge.

Use case example: Semantic data lake and AI enhance proactive treatment

AI thrives on large data sets. The healthcare industry creates a lot of these, generating huge volumes of data daily through things like medical images and genomic tests. Montefiore Health System, based in the Bronx in New York City, built a semantic data lake architecture to house all its data. This structure combines Montefiore's data stores with various ontological databases that define over 2.5 million terms and their relationships to one another. In this way, data from disparate sources and in a variety of formats can be considered when running analyses.

The solution, called the Patient-centered Analytical Machine Learning (PALM) platform runs on Intel Xeon Scalable processors and Intel® Optane™ Solid State Drives (Intel® Optane™ SSDs). Designed to help break through memory and storage bottlenecks, these devices can ingest, organize and distribute edge data to AI pipelines in real time. They can be configured for fast storage, or as a method to extend system memory, which means that this system has the potential to be a semantic data lake powerhouse.

The machine learning model that Montefiore first ran on the PALM platform helped identify patients at high risk for respiratory failure or death in the hospital. As a next step, the organization plans to apply more predictive AI use cases to the platform, including efforts to more effectively route people to the most appropriate care, predict appointment no-shows that waste precious care resources, and forecast and allocate hospital beds to more efficiently house patients and reduce their length of stay.

Edge-ready AI toolkits for developers

Distributed edge solutions are among the most complex to develop. Intel helps developers streamline their workflows and speed the deployment of distributed edge solutions, with a focus on open standards and supporting containerized and cloud-native development.

Our full-stack, optimized software is built on open standards and interfaces. For example, the Intel® Distribution of OpenVINO™ toolkit supports the development of deep learning applications essential for computer vision use cases at the edge, while the Intel® Distribution for Python helps deliver fast machine learning application performance across Intel® platforms.

Support for common libraries and frameworks like TensorFlow, Keras, PyTorch, oneDNN and BigDL enable quick application development for a range of AI workloads. For example, Intel® AI Builders member, Taboola, built its custom inferencing solutions using the TensorFlow Serving (TFS)[1] framework. It collaborated with Intel to optimize and significantly accelerate its custom TensorFlow Serving application, using the Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) on Intel Xeon Scalable processors.

Even as AI is used more frequently at the edge, the cloud continues to play an important role. For example, AI models may be trained in the cloud, taking advantage of access to more and/or larger datasets and the ability to draw on a larger pool of compute power. These models can then be rolled out to edge servers or individual devices to run inference on new data as they generate or collect it.

Meanwhile, priority data from the edge can be returned to the cloud as well. By collating data from multiple edge locations into a centralized point in the cloud, organizations can build up a rich data resource that provides visibility into operations across the infrastructure. Running AI on this data can help identify and predict larger trends or provide deeper business-level insights.

Use case example: AI helps scale metadata management and search

Enterprises in multiple industries are challenged with keeping track of all their data to make it searchable, manageable and scalable. Global IT services provider phoenixNAP offers a service that uses AI to help its customers more effectively store, search and analyze their data and metadata, across multi-cloud environments.

phoenixNAP's customers can now store their data in scale-out object storage instead of memory and keep a cache of the hottest data in Intel® Optane™ persistent memory (Intel Optane PMem, delivered with Intel Xeon Scalable processors) to accelerate performance. Intel Optane PMem provides edge servers with up to triple the maximum storage per node, while greatly reducing data latency[4]. In phoenixNAP's case, the technology has helped cut latency by 80 percent and accelerate indexing by 3x compared to hosting the solution in a hyperscale cloud environment[5].

Intel® technology and AI: Evolving together

The new 3rd Gen Intel Xeon Scalable Processors evolve Intel's 4- to 8-socket processor foundation to provide additional support for AI-infused, data-intensive cloud services like deep learning. The evolution of Intel Deep Learning Boost in the 3rd Gen Intel Xeon Scalable Processor makes it the first general-purpose server CPU to offer built-in bfloat16 instructions. This will make mainstream AI training more widely deployable for applications like image classification, speech recognition, and language modeling. In addition, Intel Optane PMem 200 Series supports up to 36TB of memory in an 8-socket configuration, helping accelerate AI inference on large and high-value data structures like medical or seismic imaging.

Approaches to Integrate AI Environments

As with any AI initiative, it's important to approach your edge-to-cloud development strategically. When working with our customers, we recommend following four broad steps:

- **Build a strong foundation** of aligned, committed stakeholders with a shared goal that addresses a real business need. Ensure there's a strong business case for your proposed AI use case, wherever you plan to deploy it.
- **Map your data pipeline** to ensure data strategy and AI needs are aligned. Be clear on what data you need and where it will come from. Have a plan in place to ensure all data is ingested, stored, processed and analyzed in the right place, at the right time.
- **Develop your AI model.** You can simplify this process by using the wealth of readily available toolkits, libraries, frameworks and software optimizations.
- **Deploy your AI workloads** using the right mix of optimized hardware, networking and applications

Learn more about these steps and find tips on tackling each stage in this eGuide:
Ease Your Organization Into AI.

Security

As IT environments grow and become more complex, protecting them is an ever-present and increasingly critical priority. Sensitive data and workloads must be secured wherever they reside, move, or are used across the infrastructure. This means ensuring you have strong security built into your data center and edge devices for on-premise workloads, and that the cloud service providers (CSPs) represented in your hybrid cloud environment can match those measures. At the same time, these protections must be as inobtrusive as possible to the end user. Individual employees should be able to access and use the data they need, on their device of choice, without their productivity being impacted.

Intel works closely with the industry's leading security companies, CSPs and independent software vendors (ISVs) to develop customized security solutions that begin at the silicon level and supports the entire edge-to-cloud infrastructure. Our security technologies align with current standards, like those set out by the National Institute of Standards and Technology (NIST), and benchmarks.

Key Intel® security technologies are:

- **Intel® Software Guard Extensions (Intel® SGX):** Intel SGX is a set of instructions that increase the protection of application code and data while in-use. Application developers can use the instructions to create trusted execution enclaves (TEEs) within the CPU memory. Data and code are only unencrypted when inside the TEE, where they are processed using cryptographically permitted applications. They are then re-encrypted before being released. The TEEs remain isolated from the rest of the environment, including the operating system, hypervisor and BIOS server, helping prevent anyone—the application developer, system administrator, server owner, or CSP—from being able to access the data and code within the enclave.
- **Intel® Control-flow Enforcement Technology (Intel® CET):** Intel CET provides CPU-level security capabilities to help protect against common malware attack methods that have been a challenge to mitigate with software alone. It targets the misuse of legitimate code through control-flow hijacking attacks, which are widely used techniques in large classes of malware. Intel CET offers software developers two key capabilities to help defend against control-flow hijacking malware. Indirect branch tracking delivers indirect branch protection to help defend against jump/call-oriented programming (JOP/COP) attack methods. Shadow stack delivers return address protection to help defend against return-oriented programming (ROP) attack methods.
- **Intel® Total Memory Encryption (Intel® TME):** Intel TME encrypts full system memory (DRAM) for improved protection against physical attacks from interposers, freeze spray, DIMM removal etc. It is enabled directly in the system BIOS with a single CPU-supplied key, and requires only a small overhead on memory performance. Intel TME encrypts the entire memory using AES-XTS, a NIST-standard 'storage class' algorithm. By encrypting data before writing it to the platform memory, and decrypting on read, it remains transparent to software. The technology is easy to implement, requiring no operating system (OS) or application enabling.

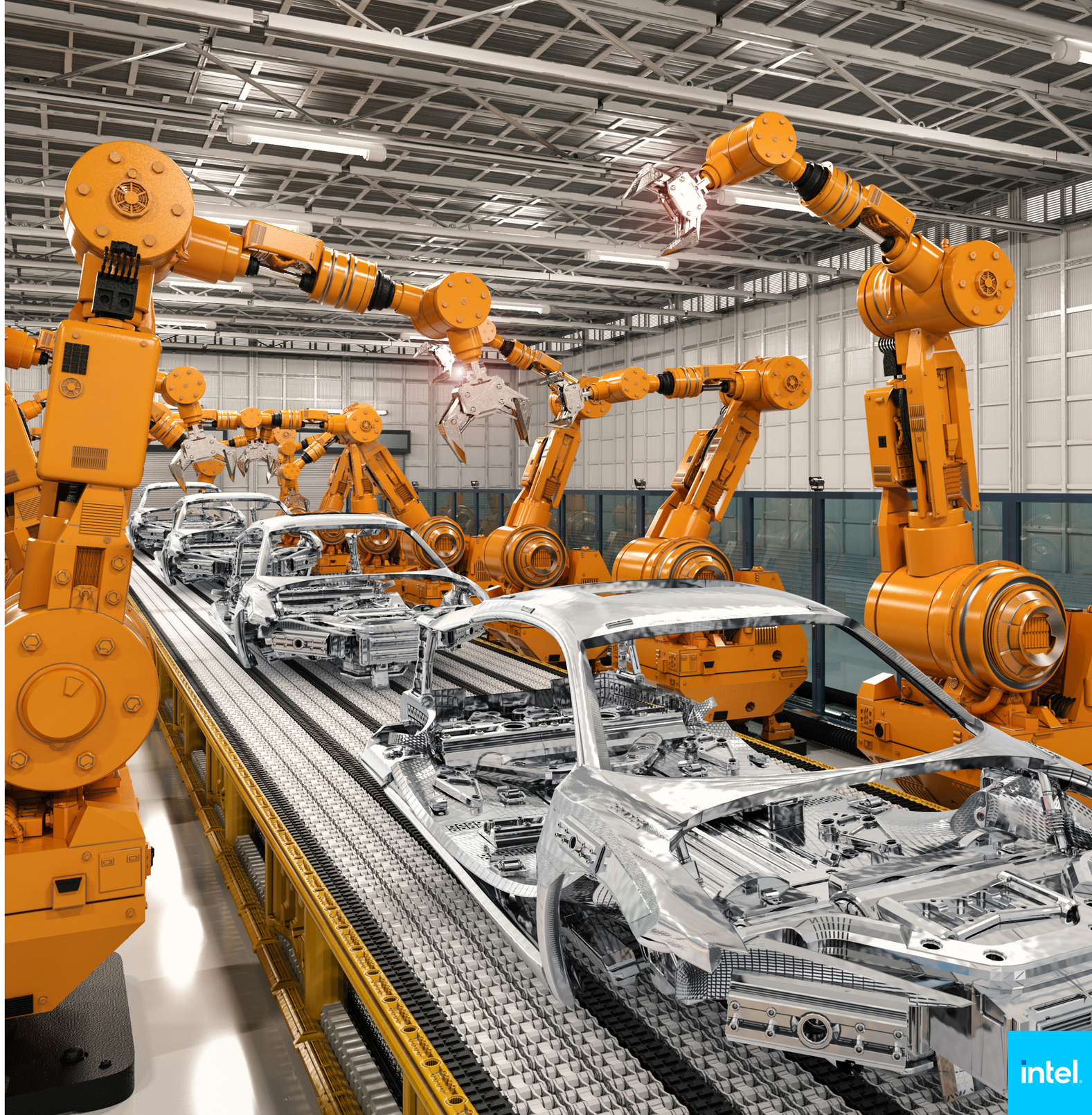
Accelerate your path to edge-to-cloud AI

Intel has decades of experience working across the entire edge value chain—from builder to integrator to cloud and network provider to developer. We have aligned use cases and fixed common integration headaches, resulting in hundreds of preconfigured packages backed by a mature developer ecosystem that is constantly optimizing and innovating. Take advantage of this ecosystem to speed development time and accelerate your time to results:

- **Ready-to-deploy enterprise AI solutions:** Intel® AI Builders offers access to over 300 leading global AI software, hardware, and service providers with more than 150 solutions across diverse use cases and markets, enabling any business to quickly harness AI.
- **Ensure optimal AI deployments:** Intel® Select Solutions for AI help you simplify and accelerate infrastructure deployment with rigorously benchmark-tested and verified solutions optimized on Intel® Xeon® processors.
- **Reduce development & collaboration challenges:** Intel® AI: in Production helps accelerate path to production with Intel® technologies, software tools, development kits, code samples, and solutions from our ecosystem.

Learn More

- Solution Brief: Intel® Select Solutions for BigDL on Apache Spark
- Solution Brief: Intel® Select Solutions for AI Inferencing
- Webpage: www.intel.com/ai
- Webpage: AI success stories
- Webpage: www.intel.com/cloud



[1] https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders/

[2] **Configuration details: Hardware:** Model: Intel® Xeon® Platinum 8168 processor at 2.70 GHz, Intel® Hyper-Threading Technology disabled. BIOS version: SE5C620.86B.0D.01.0010.072020182008. System Memory: 192 GB, 2,666 MHz. Intel® Turbo Boost Technology: Enabled. Solid State Drives: ATA device with non-removeable media. Model number: INTEL SSDSC2CW240A3. **Software:** Ubuntu 18.04.1 LTS (GNU/Linux 4.15.0-29-genericx86_64). Keras 2.1.1. TensorFlow 1.2.1. OpenVINO Toolkit: 2018 R2. Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) v0.14. Dataset: Bone age prediction model: 299x299x3 .png images. Testing carried out by Philips, August 2018. https://newsroom.intel.com/news/intel-philips-accelerate-deep-learning-inference-cpus-key-medical-imaging-uses/#gs.o36v3z

[3] 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 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".

[4] 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".

[5] Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Configurations: Up to 3x indexing and 80% cache latency decrease – based on phoenixNAP and Panzura testing as of March 2019 on Elasticsearch: Intel® Xeon® Gold 6230 processor, Total Memory 256 GB RAM, 1.5TB of Intel® Optane™ DC persistent memory, HyperThreading: Enabled, Turbo: Enabled, ucode: 0x043, OS: ('centos-release-7-5.1804.el7.centos.x86_64'), Kernel: (3.10.0-862) vs. AWS i3xlarge (Intel) Instance, Elasticsearch, Memory: 30.5GB, Hypervisor: KVM, Storage Type: EBS Optimized, Disk Volume: 160GB, Total Storage: 960GB, Elasticsearch version: 6.3.

Performance varies by use, configuration and other factors. Learn more at HYPERLINK "file:///C:/Users/abhewitt/AppData/Local/Microsoft/Windows/INetCache/Content.Outlook/M92GXFTD/www.Intel.com/PerformanceIndex" 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.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

0821/JL/CAT/PDF 344959-001EN