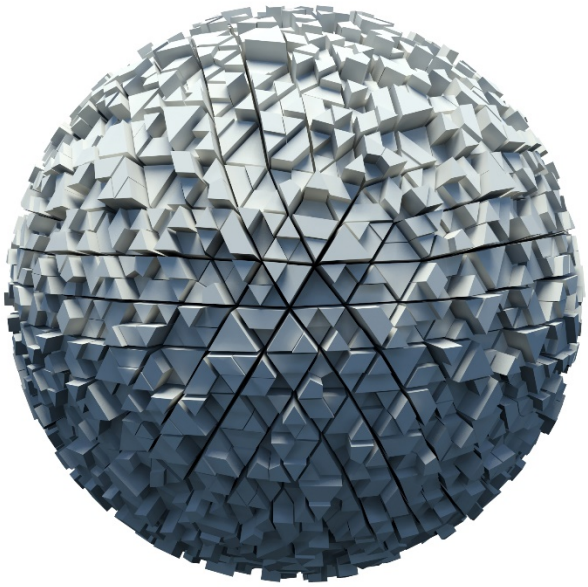
Deloitte.



Considerations for Distributing Artificial Intelligence Workloads to the Intelligent Edge

Introduction

Edge computing sits at the intersection of AI, IoT, and Big Data and is enhancing the next wave of mission capability. It enables organizations to act on location without the need to stream large amounts of information back to a cloud or traditional on-premise computing environment. As an evolution of traditional on-premise or cloud architectures, Edge computing pushes decision making "to the edge of the network" – to locations where it is needed most – where capacity for high-bandwidth network transfers may not be present.

With 41.6 Billion Internet of Things (IoT) devices projected by 2025[i], there is a growing need for platforms and hardware to generate real-time analytics, often without access to centralized data hubs. The convergence of this IoT revolution and Artificial Intelligence maturation will create what is referred to as the "Intelligent Edge" in this paper. At Deloitte, we have positioned ourselves to embrace this Intelligent Edge with a combination of leading technologies from IoT, Cloud, and AI in a solution called "Edge AI". We believe that Edge AI can reduce data streaming bandwidth, operate in a low-connectivity or disconnected state, integrate with drones or vehicles, and produce near-instant mission critical insights.

As seen in Exhibit 1, today's IoT architecture employs a hub-and-spoke model, where information is fed from devices in disparate locations into a centralized hosting infrastructure. Historically, this model was preferred due to the low computational capabilities of Edge devices. As computing resources become cheaper and more

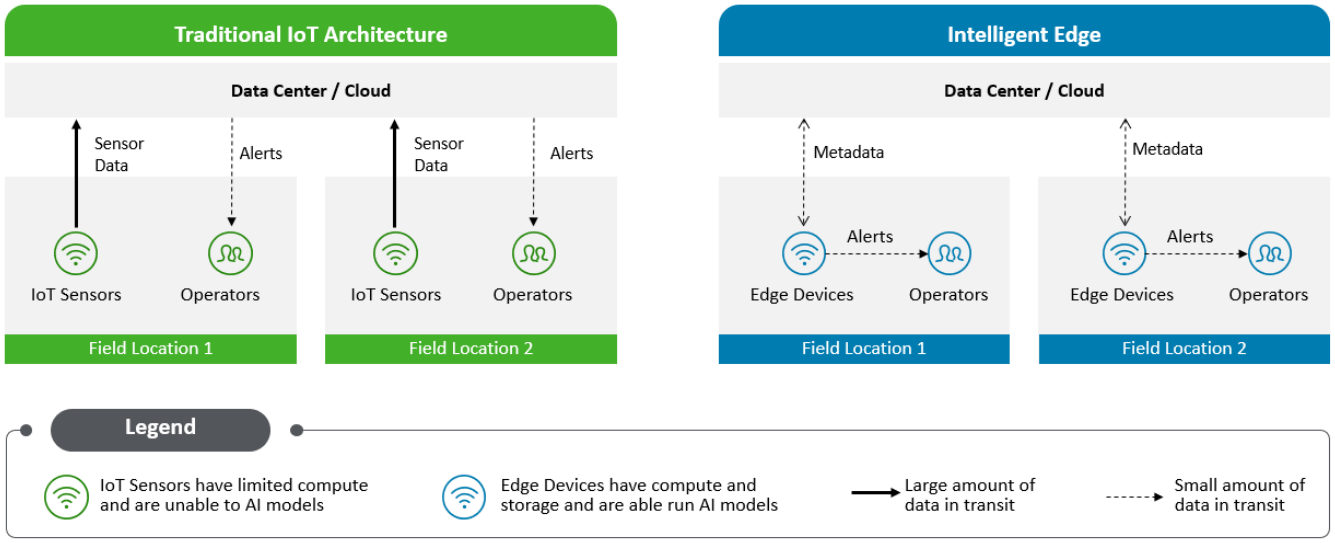
Today's Internet of Things (IoT) architecture employs a hub-and-spoke model with a centralized hosting infrastructure. The Edge model changes that equation, moving computing power and decision-making to AI-enhanced devices located close to the actual events as they occur

i – https://www.idc.com/getdoc.jsp?containerId=prUS45213219

efficient, the Intelligent Edge model can create net-new capabilities for an organization. By moving a higher percentage of the computing power and decision-making responsibility to the devices located close to the actual events, an infrastructure is created where the platform improves the devices, and the devices improve the platform.

Consider, for example, an organization with multiple physical offices, manufacturing facilities, or warehouse locations that wishes to improve their physical security. On-site Edge devices, which we're defining as sensors with additional computing power, can evaluate footage and detect unauthorized entry rapidly at the Intelligent Edge. These devices can alert on-site personnel to unauthorized activity so that they may take quick and effective action. Throughout this paper, this use case will be explored in depth as an example of one of the many possible applications of Edge AI.

Exhibit 1 | Comparison of Intelligent Edge and IoT Architectures



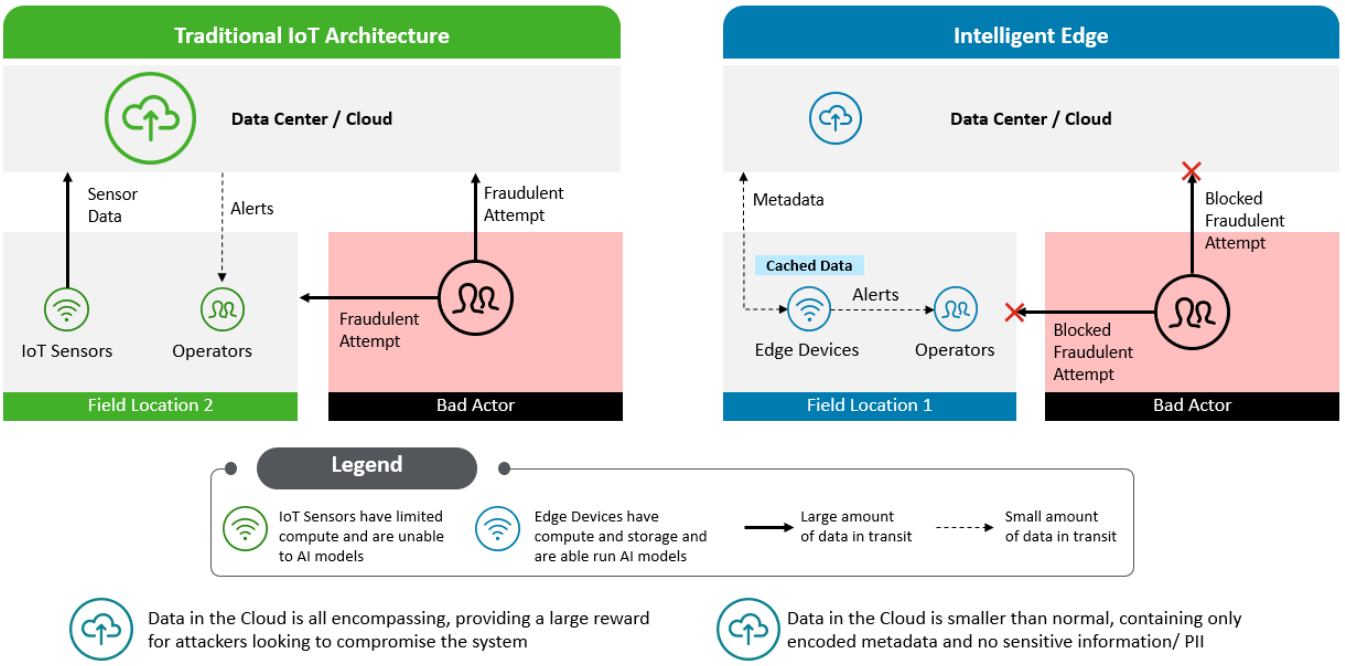
Security

Security at the Intelligent Edge builds on traditional data center security while exhibiting unique considerations and advantages. As shown in Exhibit 2, there are three main areas of differentiation: centralized security configuration updates; prevented connections from fraudulent devices; and decreased size of sensitive data to be uploaded, stored, or lost.

The baseline Intelligent Edge device has a containerized, hardened, edge-specific Kubernetes operating system, featuring both an extensible and lightweight footprint and tamper-resistance. The OS is hardened with containerized communications and code execution subsystems that will only send, accept, and execute digitally signed content. Moreover, the on-board storage of code and search objects are encrypted as well. These security provisions can reduce the risk of loss of data or intellectual property in the event the device is captured or lost. The lightweight design is well-suited to temporary, hand-held, or drone-mounted devices where bandwidth availability is not constant.

The Edge device employs "edge-aware" distributed security mechanisms, without the need for constant reliance on network availability. Employing a customized configuration of industry-standard public-key infrastructure (PKI), this edge-aware design has been rigorously tested and is in production in the military, automotive, IoT, and medical device industries.

Exhibit 2 | Implications on Security



In addition to faster processing and reduced reliance on bandwidth availability, the data privacy enhancements are also noteworthy. Reducing the need to transmit (or store) anything but indicia (i.e., metadata) rather than the full data stream, makes the transmission less subject to hijacking or decoding. If only the indicia are locally stored in the Intelligent Edge device, a compromise of the stored indicia data or streamed updates are far less likely to compromise privacy than a transmission of actual images.

Speed

Quick alerts can be essential for responding to new information effectively. A long chain of networks and computations introduces delay between a sensor detecting an event and an operator receiving a notification. Some alerts are only valuable if they can be reacted to in a timely manner. Taken to the extreme, consider getting alerts that somebody was spotted entering an unauthorized location hours or even days after it occurred. Moving computational power to the Intelligent Edge improves the speed and reliability of the operator alerts.

Intelligent Edge-based computational power increases speed in two key areas: by minimizing the data sent and by removing latency. While these two are generally intertwined, it is possible to evaluate them independently. For example, the Intelligent Edge can provide enough computational power to convert raw video to indicia (the metadata used for pattern matching) inside the imaging device, reducing the amount of data being sent, but still do the matching analysis and long term archival in the cloud. A second option is to capture raw video and analyze the images at the Intelligent Edge – in the AI-enhanced sensor itself. This allows alerts to be sent immediately, while still enabling the raw video to be sent to the cloud for archiving without slowing the alert process.

A second aspect of this redesign may include capabilities outside what is traditionally considered in IoT. For example, image recognition at the Intelligent Edge may no longer need optical cameras and sensors – but might instead turn to LiDAR or depth sensing technologies. These newer sensors also have privacy benefits. The data produced may be disassociated from a specific person or place (compared to traditional imagery) because the images are abstracted into non-identifiable formats.

In the example of an Edge device notifying an operator to a potential unauthorized intrusion, it is clear why speed is important. When analyzing the video at the Intelligent Edge – depending on local network utilization – inferences can be made in as little as 15 milliseconds. The result is a more effective response in the field, which increases the speed at which operators receive the alert and secure the location.

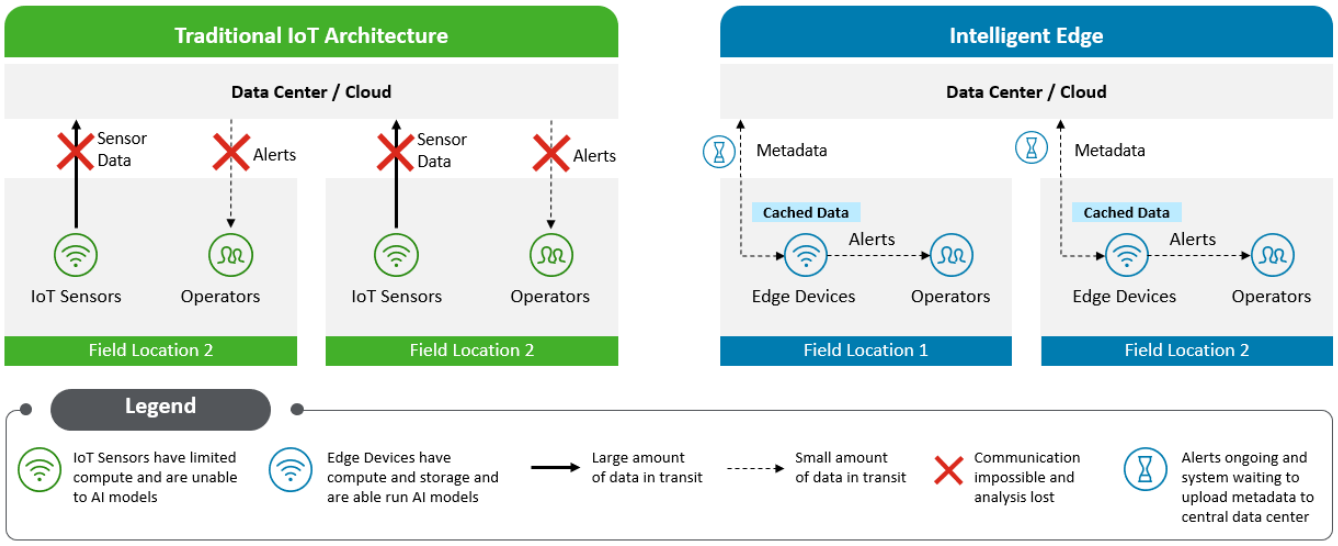
Resiliency

A distinct advantage of Intelligent Edge computing is the ability to operate while disconnected from a central data center. As shown in Exhibit 3, Intelligent Edge devices can communicate directly with the operators at each field location, providing immediate insight and avoiding the latency introduced when processing is done at the data center.

Intelligent Edge devices contain the database of the target parameters; in this case, the devices store the locations of secure and non-secure areas. Operating in a disconnected mode from the central data hub, the recognition training managed via Edge AI can independently report the subject's characteristics, time, and location, thereby accomplishing the complete recognition task in a fraction of the time of a centralized data processing model.

When connections to the data center are lost, the Intelligent Edge devices can cache the data to be transmitted and send it to the data center once the connection has been reestablished, thus removing any gaps in the aggregate data. Alerts to operators do not have to come only from the central data hub – the AI-enabled sensor device can provide faster feedback.

Exhibit 3 | Operations in a disconnected state



Consider, for example, the Security Operators at the field location outlined at the beginning of this paper. If the location were to lose access with the internet, Operators would still be able to receive security alerts on locally networked workstations and respond to incidents as they arise. As shown in Exhibit 3, no such alternative path exists in a traditional IoT architecture. By caching the metadata around each incident, such as time and location, and then sending it when possible to the data center, the organization's data analysis team can still gain every insight possible into security incidents.

Cost

Costs at the Intelligent Edge can shift dramatically, and it is important to plan and architect accordingly. At a high level, costs may shift as network and bandwidth costs decrease and hardware costs increase. Network and bandwidth costs decrease as less data is sent to the cloud. Hardware costs can increase as Intelligent Edge devices are no longer unintelligent sensors but are able to store and process information independently, and this may drive down central data storage costs.

With an Intelligent Edge device, the processing of images to indicia can be completed on-board the device. Network bandwidth is conserved because no raw data has been transmitted. By contrast, in a traditional model, the images captured are transferred and large amounts of network bandwidth are consumed before anything of value is determined. This also means that higher resolutions may not improve the time to recognition in the traditional model; better resolution slows down the process by requiring more time to traverse the network before processing, recognition, and matching can begin.

The hardware costs of shifting computational power to the Intelligent Edge can vary widely depending on many trade-offs. Leveraging available hardware can greatly accelerate time to market and remove “from the ground up” development roadblocks.

When transitioning to an Intelligent Edge architecture, the storage footprint in the cloud goes down, but storage required at the Edge increases. Since cloud storage is generally more expensive than Intelligent Edge storage, cost efficiency increases. However, if inefficient storage distribution among the devices results in the total volume of storage going up, there may be an offsetting storage cost increase. With that in mind, the effect may be that storage cost changes could be much less significant compared to the savings on bandwidth and hardware.

In our example, the organization can drastically cut bandwidth costs by analyzing the raw video on the device rather than transmitting it to the data center for processing. Inexpensive local storage added at the Intelligent Edge supports storage and processing needs as a capital expense, paid for over time by a commensurate amount of storage released from the data center.

Solutions

AI is only as valuable as its ability to be used in production systems and improved over time. In recent years, AI and ML technologies have shown significant improvements in performing complex cognitive tasks such as disease diagnostics and contracts review. Considering the potential to allow human workers to focus on higher level tasks where initial “prep work” is done by AI, many enterprises are scaling up AI/ML initiatives as a necessity to stay competitive in the market, making significant investments in infrastructure, solutions, and talent. Orchestration solutions are ready to scale up the operational and support infrastructure that AI-enabled systems require – managing from hundreds, to millions of AI endpoints without needing to rearchitect the infrastructure.

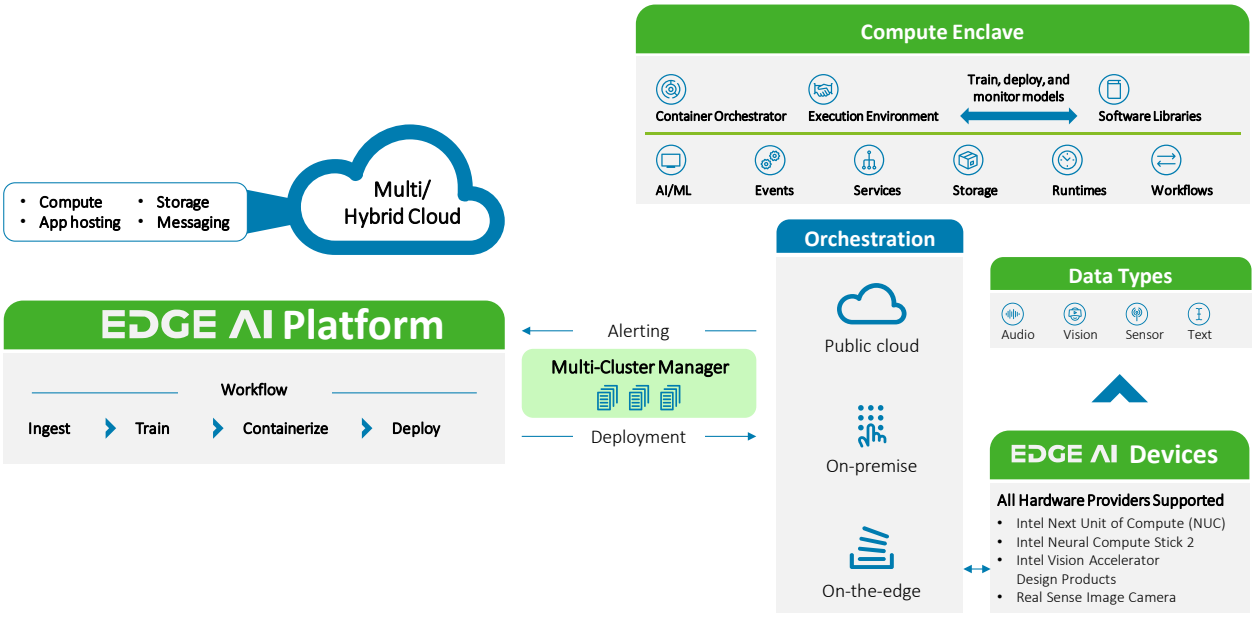
The mature processes around AI implementation must solve the challenge of effectively managing and commanding these intelligent devices. Breakthroughs such as IoT management applications and Google AutoML have enabled this kind of infrastructure. However, for a truly agnostic and generalized platform for managing, controlling, and updating these Intelligent Edge devices remotely, one must look to systems integrators.

Deloitte’s Edge AI shown in Exhibit 4 is a cloud and hardware agnostic accelerator that syncs with Intelligent Edge devices, provides a pipeline for easily & remotely deploying AI models to these devices, and provides a management suite for viewing these Edge AI insights happening across the spectrum of connectivity options, from real time on-device to intermittent. Edge AI is designed to interoperate with any cloud provider, utilize 3rd party vendors, sync Edge devices from any hardware provider, manage data science development of Edge AI models, and view alerts coming directly from Edge devices.

The idea can be best understood through a practical example of the idea-to-development-to-product lifecycle. A new AI model can be trained on the Edge AI platform and converted to Edge model format via Intel’s OpenVINO distribution. This new Edge model can then be containerized, via Docker, with the OpenVINO dependencies, custom Edge AI scripts, and be deployed to Edge devices via k3s, a lightweight Kubernetes distribution installed on the device. This Kubernetes cluster can

then be immediately accessed remotely and securely from Edge AI platform, allowing the user to assess the device's status, set alert thresholds, and control the device's peripherals as well as schedule maintenance operations.

Exhibit 4 | Edge AI Architecture



The potential Edge AI components:

Capability	Product Name	Differentiator
Intelligent Edge Compute	Intel Next Unit of Compute (NUC)	Brings full-scale server compute to a small space, scaling edge computing nodes and enabling the mission directly at the edge without needing to transport the data.
Intelligent Edge Compute	Intel Neural Compute Stick 2	Deep learning development kit built off the Movidius™ VPUs, a full-fledged system-on-chops (SoC), designed specifically for computer vision and analytics. Ideal for small scale pilots and exploring what is possible with edge compute capabilities
Intelligent Edge Compute	Intel Vision Accelerator Design Products	Enterprise-oriented compute devices that enable the hardware acceleration of inferencing at the edge. These accelerators are based on the Intel Movidius VPUs or the Intel Arria Field Programmable Gate Array (FPGA) and can best optimize performance per watt per dollar for some large-scale use cases
Computer Vision Sensor	Intel Real Sense Image Camera	Camera augmented with LIDAR capabilities to improve depth perception for indoor use cases. Ideal for warehouses or storefronts, this camera integrates well with the suite of Intel hardware and software to enable computer vision applications
Machine Learning Software	Intel OpenVINO Toolkit	Open source product that provides a high-performance deep learning interface for machine learning and computer vision. Enables faster development through its integrated development environment and pre-packaged model library and more flexible deployment, enabling teams to write once and deploy anywhere through optimized, heterogeneous, deployment
Machine Learning	OpenVINO	Open Source AI toolkit championed by Intel based in C++ in order to streamline processing time. OpenVINO includes a library of tailorable algorithms to address a number of use cases
Orchestration	Google Kubernetes Engine	The advantages of portability for Edge AI algorithms makes containerization important for the interoperability of models across devices. Kubernetes is the industry standard container orchestration tool, enabling our "build it once, run it anywhere" philosophy
IoT Management	Google IoT Core	Enterprise device management in a centralized location all without needing to manage the underlying hardware

Edge AI Development	Google AutoML	Cloud tool designed to rapidly develop AI algorithms, which can be formatted to run natively in cloud, or on Edge devices

In the earlier example of the Intelligent Edge, the operator at a field location can benefit from timely updates of the models used on the devices at their location by having the device self-train and self-improve. AI models are constantly being refined to eliminate many potential sources of error such as false positives or false negatives. Removing these sources of error for the operator can enable them to more effectively manage the security of their field location without the need for manual updating processes.

Conclusion

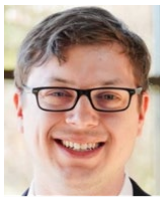
The ability to shift machine learning workloads to the Intelligent Edge through orchestration products such Edge AI provides unique advantages over a traditional data center/cloud hosting model. This new architecture can change the way security at the Intelligent Edge is viewed, increase the speed at which users can respond to complex information, enhance resiliency, and create opportunities to reduce costs.

Meet the Deloitte Analytics Thought Leaders

Key points of contact for any questions regarding the content of this paper.

Doug Bourgeois
Managing Director
Deloitte Consulting LLP
GPS CBO Cloud Engineering

Shailesh Singh
Managing Director
Deloitte Consulting LLP
GPS Analytics and Cognitive



Ryan Luckay
Specialist Master
Deloitte Consulting LLP
GPS Analytics and Cognitive

George Uehling
Senior Consultant
Deloitte Consulting LLP
GPS Core Business Operations

Deloitte.