IDC PERSPECTIVE

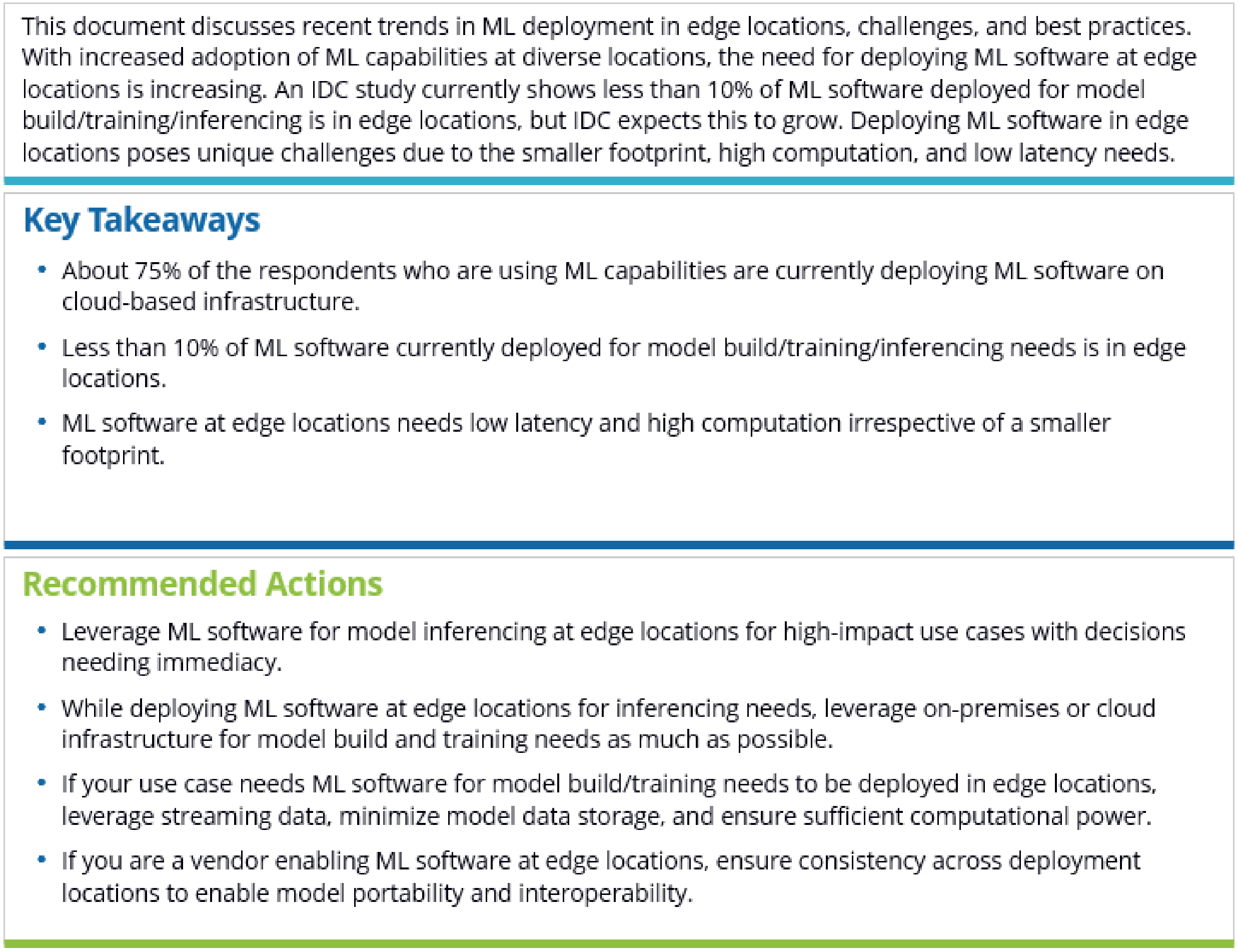
Machine Learning at the Edge

Sriram Subramanian Jennifer Cooke

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: Machine Learning at the Edge



This document discusses recent trends in ML deployment in edge locations, challenges, and best practices. With increased adoption of ML capabilities at diverse locations, the need for deploying ML software at edge locations is increasing. An IDC study currently shows less than 10% of ML software deployed for model build/training/inferencing is in edge locations, but IDC expects this to grow. Deploying ML software in edge locations poses unique challenges due to the smaller footprint, high computation, and low latency needs.

Key Takeaways

- About 75% of the respondents who are using ML capabilities are currently deploying ML software on cloud-based infrastructure.
- Less than 10% of ML software currently deployed for model build/training/inferencing needs is in edge locations.
- ML software at edge locations needs low latency and high computation irrespective of a smaller footprint.

Recommended Actions

- Leverage ML software for model inferencing at edge locations for high-impact use cases with decisions needing immediacy.
- While deploying ML software at edge locations for inferencing needs, leverage on-premises or cloud infrastructure for model build and training needs as much as possible.
- If your use case needs ML software for model build/training needs to be deployed in edge locations, leverage streaming data, minimize model data storage, and ensure sufficient computational power.
- If you are a vendor enabling ML software at edge locations, ensure consistency across deployment locations to enable model portability and interoperability.

Source: IDC, 2021

SITUATION OVERVIEW

Many industries are progressing on the journey to digital-first operations and are quickly recognizing the potential for machine learning (ML) workloads. These workloads extract value from data — unlocking insights that can improve safety of people and communities, increase operational efficiency, and drive greater innovation. Today, most machine learning and big data analytics workloads are happening in core or cloud datacenters. In August 2021, IDC surveyed enterprise organizations to understand where AI and ML workloads are run. More than 75% of organizations said that these workloads are run in core datacenters today.

Based on the positive outcomes of early projects and the greater availability of industry-specific tools, IDC believes that ML adoption will accelerate. As this happens, ML workloads will migrate to where the data is generated. By placing compute resources adjacent to where data is created, organizations can achieve better performance, comply with data sovereignty/regulatory requirements, and save on data transport costs. ML needs data gathered from a myriad of devices or "things" that are geographically dispersed. Video cameras, sensors, and connected devices are generating massive volumes of data. Organizations are generating and collecting massive amounts of data, and the key to shifting to data-driven decisions is the ability to extract insights from this data. Having the right infrastructure and platform is key to making this shift. Industries are quicky recognizing that more needs to be done to help them become not just data rich but data driven. The ability to analyze and extract insights from data is driving more investments in resources outside of the core datacenter — or at the edge. These edge locations will comprise manufacturing floors, hospitals, stadiums, airports, warehouses, and windmills, to name just a few examples. These same industries will move more workloads to new edge locations as they seek to modernize their operations.

Why Machine Learning at the Edge?

A massive opportunity exists to transform and improve operations by extracting insights from data. The reality is that all of this data is generated outside of the four walls of a datacenter. As industries seek to further improve processes and operational efficiency, more of the machine learning will happen outside of their core datacenters and closer to where the data is generated. As latency and performance needs rise and bandwidth-intensive applications proliferate, a distributed approach to compute is required. Machines, video cameras, sensors, and all sorts of devices and things are creating and collecting massive amounts of data. Machine learning can extract insights from this data. But the costs of transporting data, the regulatory and compliance associated with where data lives and travels, and latency/performance needs are requiring that more analytics and processing for machine learning be done very close to where the data is created. For this reason, edge IT resources are needed.

Today, many industries are at the very early stages of their journey toward improving operational efficiency and safety with machine learning applications. Manufacturing, transportation, retail, and healthcare organizations are just learning and exploring ways to improve operations by relying more on data-driven insights. One of the limiting factors as organizations begin harnessing the value of data is the challenge of moving, managing, and securing that data. Digital-first business is soon confronted with the cost of transporting data and data sovereignty regulations. These two factors are driving more ML workloads to run outside of the organization's core datacenter and more frequently in the field, at a remote site, in a hospital, or on the manufacturing floor. For many organizations, leveraging a services partner — such as a cloud or colocation provider — will speed the process of gaining insights and advantage from data.

In manufacturing settings, edge resources support ML and AI for visual inspection of assets (preventative maintenance), quality control for product consistency and asset inspection, warehouse space optimization, using drones and sensors for perimeter protection and intrusion detection, and the use of digital twins to understand and optimize operations.

In stadium or public event venues and airports, access control, facial recognition, crowd monitoring, abandoned bag tracking, and parking occupancy and availability are powered by ML platforms.

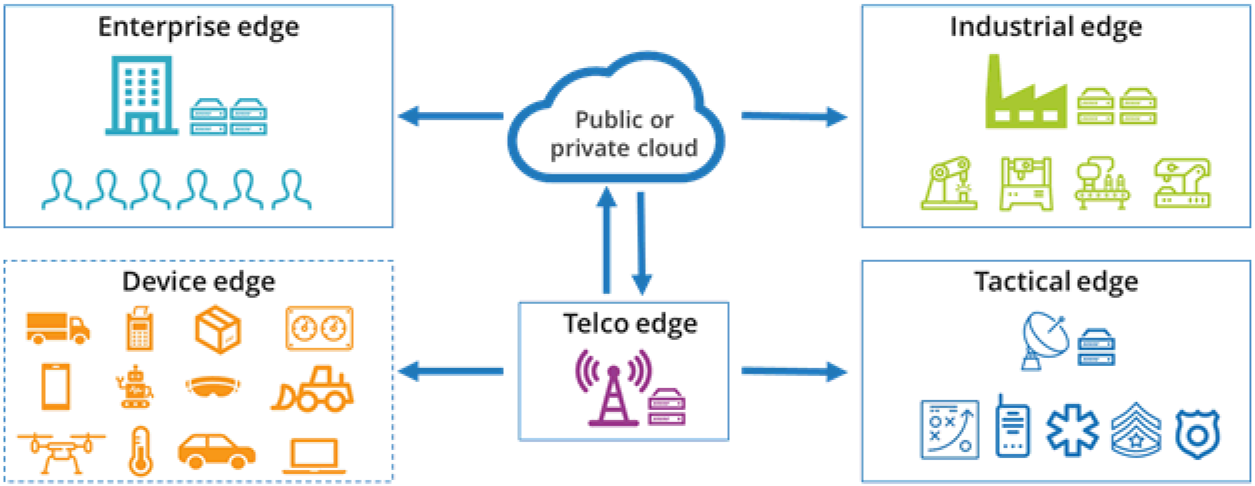
In the energy and utilities industries, ML is streamlining pipeline inspections that impact personal and community safety. ML is also used for oil field and refinery optimization, asset monitoring, and predictive maintenance on equipment to ensure resilient and safe operations.

In the transportation industry, machine vision is used to identify any anomalies and speed the repair of train cars before they break down. Autonomous vehicles of all kinds – ships, cars, and freight trucks – will rely on edge IT to operate, and ML will optimize their routes, schedules, and maintenance.

The aforementioned examples are just a few of the uses for ML workloads at the edge. As service providers and large organizations increase adoption of these capabilities, they will explore the IT platforms that best support their digital-first journeys. Edge IT will exist across many different locations, including cloud service provider and colocation providers' datacenters, within an operational setting, in the field, or at a remote office (see Figure 2). Most organizations will leverage infrastructure from multiple sources, not just one or two.

FIGURE 2

Types of Edge Deployments



Source: IDC, 2021

Challenges Deploying ML Software at the Edge

End-to-end machine learning workflow consists of various steps including data ingestion, data preparation, model build/train, model evaluation, deployment, inferencing, and monitoring. Each step has unique requirements and challenges.

Data ingestion/data preparation stages employ training and testing data and are storage intensive and need high QoS for data access. Hence they need to be supported by a scalable, high-performing storage infrastructure. Model training is a highly compute-intensive task, with periodic bursts of computational activity. It also needs a scalable, high-performing storage infrastructure. Model inferencing requires low latencies since model predictions need to be immediate.

As discussed in the previous section, edge deployments are of four types – enterprise, device, industrial, and tactical edge. Each type of deployment has its own limitations. For example, device edge deployments may not have large compute and storage resources due to their capacity/size limitations. Tactical edge deployments may have limited connectivity. Enterprise and industrial edge deployments may introduce heterogeneity of infrastructure. Given these limitations and the distributed nature of edge locations, deploying ML software at the edge poses the following challenges:

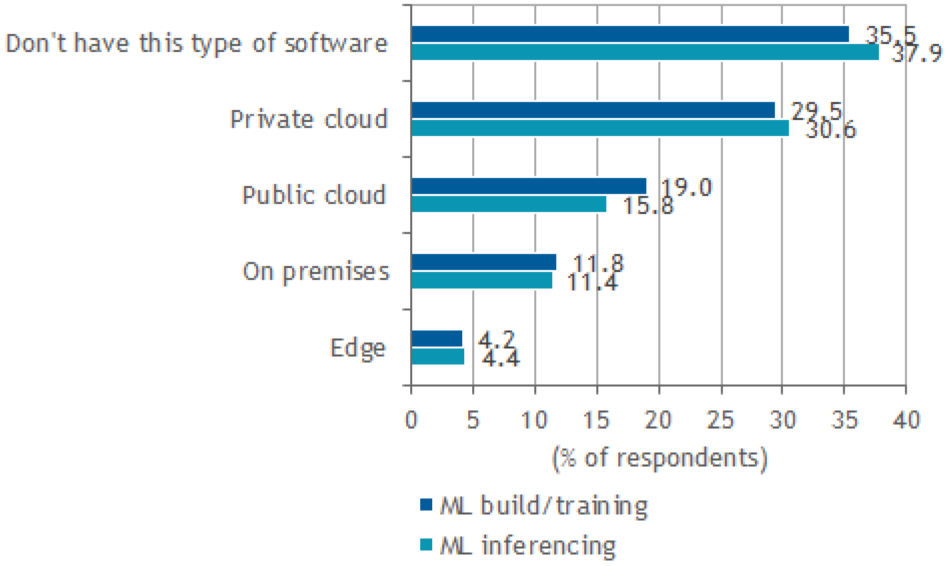
- **Heterogeneity:** The difference in edge deployment types influences the type of infrastructure deployed, which in turn influences the type of ML software, compute/storage management platforms, and infrastructure provisioning software. This introduces heterogeneity in the ML software and accompanying platforms/tools, which makes machine learning operations (MLOps) more complex. MLOps engineers now must manage heterogenous platforms and tools at different locations.
- **Capacity limitations:** Some of the edge deployments such as device edge deployments may be limited by space constraints, and hence infrastructure deployed may have capacity limitations. This in turns prevents running compute-intensive tasks such as ML training on these deployments. In such cases, the devices may need to stream/send the data to a different location, such as core or cloud, to perform such tasks.
- **Privacy and security concerns:** Edge deployments also introduce privacy and security concerns to data collected at the deployment location. ML software deployed at the location needs to be able to handle such concerns, either all by itself or in tandem with ML software deployed in core or cloud locations.

These challenges inhibit ML software deployment at edge locations. A recent IDC study shows that less than 10% of ML software for model build/train/inferencing is currently deployed in edge locations. However, with growing adoption of ML capabilities, expanding edge-based use cases, and increasing investments in edge deployments, IDC expects this share to grow rapidly (see Figure 3).

FIGURE 3

Machine Learning Software Deployment by Location

Q. Where are customers deploying ML software?



n = 920

Source: IDC's *Future Enterprise Resiliency and Spending Survey, Wave 7,* August 2021

Challenges Deploying and Supporting Hardware at the Edge

IT challenges that are difficult at core datacenters become insurmountable at the edge. IT infrastructure deployed in edge locations introduces several additional challenges, including:

- Limited or no IT staff
- Unsecured spaces
- Environmental conditions including temperature, humidity, and dust

As organizations increase their reliance on ML workloads at the edge, downtime for IT becomes downtime for all operations. The impact to the business ranges from being costly to potentially introducing a safety risk. Because of this, hardware that is easily managed and updated, secure, standardized, and open or cloud native is best suited for edge deployments. Edge infrastructure will exist in many different locations, and the challenges related to physical proximity aren't always fully appreciated. The lack of trained IT staff onsite or the ability to rapidly transport someone to the site increases the need for equipment that can operate autonomously. In situations where equipment does need to be physically altered or replaced, it should be simplified and clear to the point where people who are not trained IT staff can easily understand directions. Color-coded cables and language-agnostic instructions are key.

IDC's research has shown that when planning an edge strategy, enterprises often turn to their hardware provider first. But assembling an edge solution is far more than deploying a server in a new location. Connectivity, remote management, data and physical security, and a degree of autonomous operation are table stakes for successful edge projects. Edge IT often requires coordination across many different partners, such as telcos, service providers, systems integrators, managed service providers, and hardware/software providers. Standardized platforms and trusted partners with industry-specific knowledge of unique challenges will be preferred. Having a provider that is able to anticipate problems that organizations didn't know they'd be facing is a great advantage.

For many organizations, edge deployments are their first forays into high-performance workloads. Practical considerations of performance, cost of moving data, and data sovereignty regulations are necessitating AI and ML processing closer to where data is created. The combined challenge of deploying and managing edge IT, as well as supporting demanding new applications and platforms, will be a major hurdle for these organizations. A technology provider's ability to simplify ML at the edge will be the key to success. In IDC's October 2021 *Future Enterprise Resiliency and Spending Survey,* 72% of organizations said that the complexity of deploying edge computing solutions is causing them to make greater use of managed service provider or value-add integrators and partners. Organizations need help, and technology providers that can provide guidance and streamline edge decisions will be preferred.

ADVICE FOR THE TECHNOLOGY BUYER

- In some edge implementations, place infrastructure where people shop, work, and reside. Edge compute needs to be a good neighbor and not be intrusive or disruptive to the surrounding environment. Seek providers that offer purpose-built infrastructure to ensure that the equipment is protected from the environment and unauthorized access and operates quietly and unobtrusively.
- Seek out providers that have a deep understanding of the infrastructure needs to support high-performance workloads such as AI and ML workloads. By leaning on a trusted, experienced provider, your organization can avoid some of the early pitfalls of early edge adopters.
- Leverage ML software for model inferencing at edge locations for high-impact use cases with decisions needing immediacy.
- While deploying ML software at edge locations for inferencing needs, leverage on-premises or cloud infrastructure for model build and training needs as much as possible.
- If your use case needs ML software for model build/training needs to be deployed in edge locations, leverage streaming data, minimize model data storage, and ensure sufficient computational power.

IDC'S POINT OF VIEW

Expanding Use Cases

IDC observes that ML use cases that require edge deployments are increasing rapidly. Various OT use cases that have high impact and a sense of immediacy in their predictions are well served by edge location-based deployments. Recent IDC studies show that across all industries, slightly more than half have started to use AI/ML in analyzing operational data. Specifically, 20% of respondents from the oil and gas industry and 22.1% from the discrete manufacturing industry have indicated that they have realized rapid paybacks from executing AI/ML projects. IDC expects adoption of edge-based ML deployments across industries to increase.

Growth

IDC observes that the vendor ecosystem to support edge-based deployments of ML software is evolving rapidly. Some of the ML platform vendors and cloud-based ML services vendors are increasing their capabilities to support edge-based ML inferencing on heterogenous infrastructure. Open-source software such as lightweight Kubernetes (K3s) is also providing end users with more options for enabling edge-based ML use cases.

IDC also observes that investments on edge deployments are increasing much more rapidly than investments in core. Edge infrastructure spending is forecast to increase at a compound annual growth rate (CAGR) of 18.7% through 2025, whereas core infrastructure spending is only growing at a CAGR of 3.3%. . With increasing adoption of ML techniques, expanding use cases, and increased investments in edge deployments, edge-based deployment of ML software will also increase rapidly.

Recommendations to Vendors

Machine learning models go through various stages in the end-to-end ML pipeline. They also go through various infrastructure environments from experimentation to production. With the introduction of heterogenous environments through edge-based deployments, migrating models is only getting more complex.

Consistency across such environments enables MLOps engineers to move models easier. While containerizing ML models enables portability across these environments, maintain consistency to ensure smoother operations. MLOps engineers may also need a common interface across these environments to ensure model monitoring.

IDC recommends that vendors enabling ML software at edge locations ensure consistency across deployment locations to enable model portability and interoperability. Enabling consistent interfaces also increases MLOps productivity.

Organizations struggle with the complexity of edge deployments. They are rapidly recognizing that these edge resources should be modernized, cloud-native, and high-performance edge resources. Faced with this complexity, they are turning to providers to help navigate the journey. Technology providers' ability to meet customers where they are and prescribe an affordable, scalable solution to their needs will be well accepted in the market.

LEARN MORE

Related Research

- *IDC FutureScape: Worldwide Artificial Intelligence and Automation 2022 Predictions* (IDC #US48298421, October 2021)
- *Market Analysis Perspective: Worldwide Edge Strategies, 2021* (IDC #US48264221, September 2021)
- *Market Analysis Perspective: Worldwide AI Life-Cycle Software, 2021* (IDC #US47712321, September 2021)
- *Enterprise AI: An Architectural Shift Emerges* (IDC #US47602221, April 2021)

Synopsis

This IDC Perspective provides an overview of machine learning (ML) at the edge and the challenges associated with ML software and infrastructure deployment at the edge. This document also provides recommendations to the end user to overcome such challenges and leverage ML at the edge successfully.

"Edge-based deployments pose unique challenges to deploying ML software at the edge. A recent IDC study shows that less than 10% of ML software currently deployed is at edge locations," said Sriram Subramanian, research director, AI and Automation Software research group at IDC. "With increasing adoption of ML techniques, expanding use cases, and increased investments in edge deployments, edge-based deployment of ML software is expected to increase."

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com