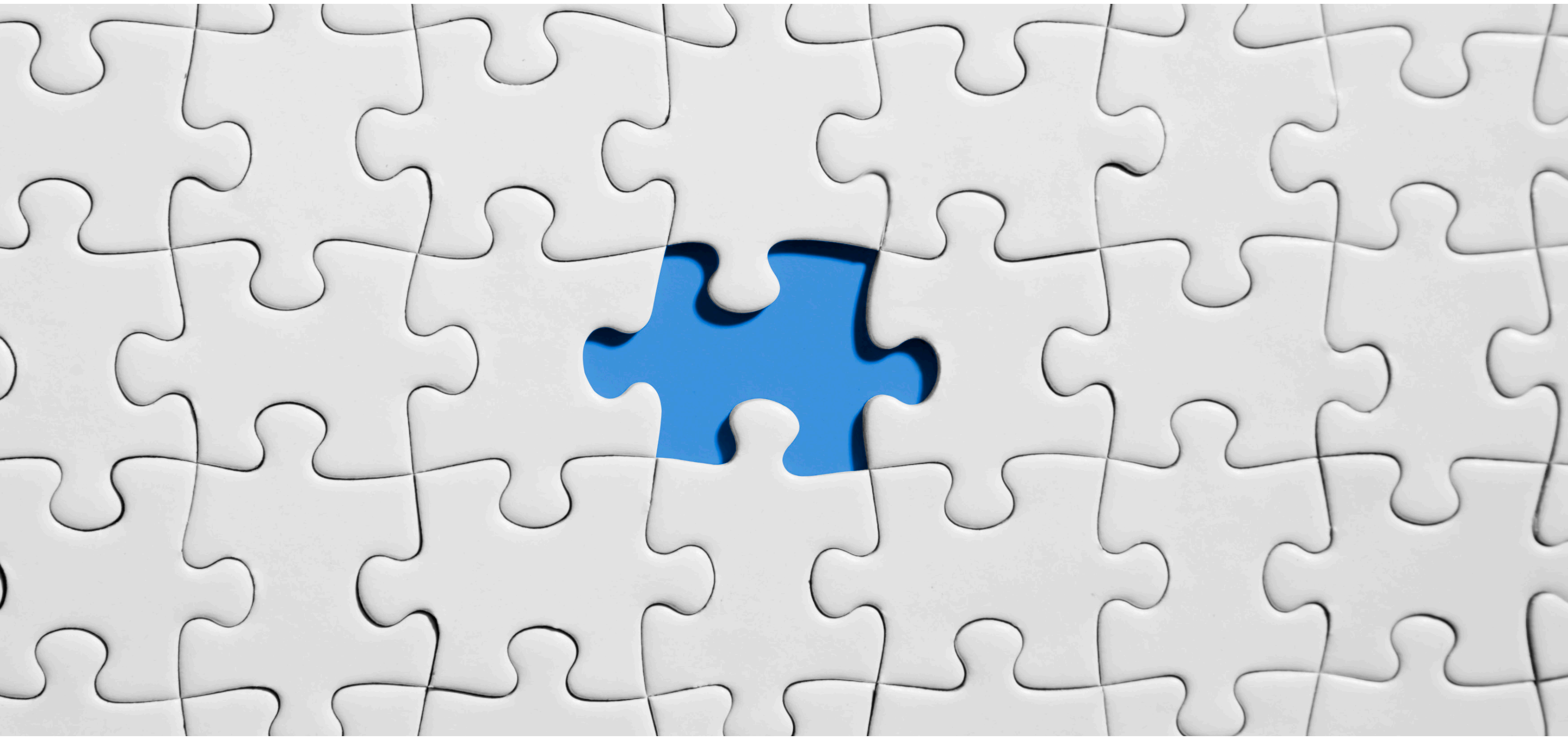
AI AND ML WORKLOADS PLATFORM: CLOUD OR ON EDGE



Harshit Dixit

Senior Sales Engineer Analyst

Dell Technologies

Harshit.dixit@dell.com

Rani Priya S

Associate Sales Engineer Analyst

Dell Technologies

RaniPriya.PriyaS@dell.com

Knowledge Sharing Article

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

Introduction

Once just a figment of the imagination of some of our most popular science fiction writers, Artificial Intelligence (AI) is taking root in our daily lives. Another such science fiction figment coming to reality is Internet of Things (IoT), a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers (UIDs) able to transfer data over a network without requiring human-to-human or human-to-computer interaction. A ‘thing’ in IoT can be a person with a heart monitor implant, a farm animal with a biochip transponder, an automobile that has built-in sensors to alert the driver when tire pressure is low or any other natural or man-made object that can be assigned an IP address and is able to transfer data over a network.

We are entering a new decade, one that will be defined by data. Organizations will succeed or fail largely based on how they collect, use and democratize data analytics throughout their business. At this pivotal point of business transformation, organizations must embrace change and invest in it. Increasingly, organizations in a variety of industries are using IoT to operate more efficiently, better understand customers to deliver enhanced customer service, improve decision-making and increase the value of the business.

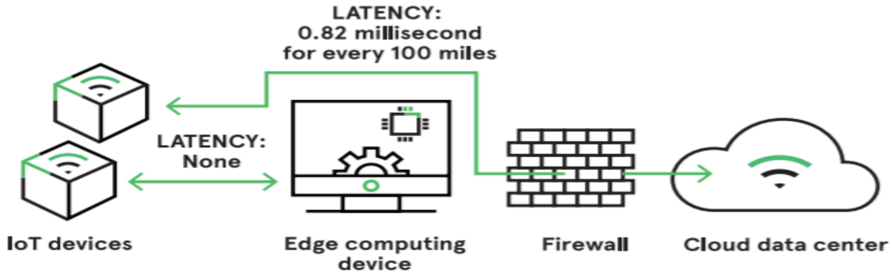
AI is real and is already changing the way businesses used to do things. Developers are trying to bring AI and IOT together as AIoT to help companies in a variety of industries benefit from the data generated by connected devices.

Drawbacks of AI on Cloud

Today, most machine learning (ML) and AI workloads depend on Cloud Data Centers' massive computational, processing, analytics, volume storage, memory and graphic processing capabilities. It is generally agreed that the cloud is the **"least expensive "**platform to host AI development work. But organizations need to decide if cloud is the best platform to host their ML and deep learning (DL) workloads. The simplest way to determine what phase the organization is in their AI journey. Organizations typically turn to cloud service providers (CSP's) for ML and AI workloads because they have a reservoir of development tools and other resources readily available such as pre-trained deep neural networks for voice, text, image, and translation processing. Moreover, for organizations without an expert staff for AI projects, these platforms build DL neural networks automatically saving weeks or sometimes months of labor. But there is a caveat to consider when taking this approach: stickiness. Those applications may only be able to run on the cloud platform on which they were developed. Now platform stickiness (in the case of CSP) in itself is not a bad thing as you have Graphical Processing Units (GPUs) or Field Programmable Gate Arrays (FPGAs) to accelerate training process and you don't have to deal with complex hardware configurations. But the catch here is you will not stop training your neural networks and that eventually will require massive computation. As per market research, this level of compute on cloud can cost you more than 2-3 times than building your own private cloud to train and run neural networks. You can reserve GPU's for a longer period of time in public cloud but building your own private cloud will always remain a cheaper option.

Another major drawback of this type of deployment is latency which occurs when data are collected from sensors and various connected devices and sent back to the cloud. Explosion of data by these devices will grow ever greater. IDC predicts that the digital data created and consumed will grow from around 40 zettabytes of data in year 2019 to 175 zettabytes in year 2025 – more than four times the amount of data produced in 2019. When these edge devices send all the data to the cloud the drawback is obvious; jammed the bandwidth, resulting in latencies.

For every 100 miles data travels, it loses speed of roughly 0.82 millisecond.*



*AVNET-ai-at-the-edge-whitepaper.pdf

Consumer Data Privacy is another area of concern.

"Companies that deal with highly-sensitive consumer data are finding cloud computing to be dangerous due to the high cost of breaches. As a result, many of these companies are using edge computing for consumer data since it affords them more options for security and control. This can complicate the enterprise workflow, but it can bring benefits, especially in medical data companies. - Sean Byrnes, Outlier"

This is where AI Inference at the Edge makes sense. Installing a low power computer with an integrated inference accelerator close to the source of data results in much faster response time. When compared to cloud inference, inference at the edge can potentially reduce the time for a result from a few seconds to a fraction of a second.

AI on Edge

The answer to issues such as latency, bandwidth cost, and network problems is Edge computing. It makes more sense for certain scenarios to perform the processing close to the point from where data is generated.

Driving Factors for AI on Edge

1. **Improved Response Time**

Ruling out the need to transfer data to cloud for processing mitigates response time issues that affect validity of real-time. Because of the explosion of number and types of data generating devices, a massive amount and types of data is being generated are being sensed on the device side. AI is capable of quickly analyzing those huge data sets and giving results that drive high-quality decision making. A component of AI – deep learning – is able to automatically identify patterns and detect anomalies in the data and feed them real-time for predictive decision-making.

2. **Enabling AI with more Applications and Scenarios [2]**

While Deep Learning and Machine Learning highly depend on algorithm and hardware, the role of application and scenarios cannot be overlooked. To make your AI more efficient you need to feed more parameters to your machine learning or deep learning algorithm which illustrates the importance of data in AI. Where the data is being generated is also an important factor to consider. Previously, data was generated in centralized data centers. However, the scenario is changed with the development in IoT. Now sending the zettabytes of data over to cloud data centers requires a massive amount of bandwidth incurring heavy costs for organizations. This problem can be overcome by analyzing data at edge locations where it is generated removing the bandwidth requirement and reducing computing pressures on cloud data centers and bringing this to Edge.

It's not only response time which is improved by placing AI workloads on Edge but we also get a broader range of parameters which can be injected into AI/ML algorithm to build strong and improved neural networks.

3. **AI for Everyone and Everywhere[2] :**

With major organizations having the vision of "AI for everyone and everywhere" it is important that AI should go closer to the devices where the data is being generated. Edge computing is clearly better to achieve this than cloud computing. Why? Because the Servers performing Edge computing are in a closer proximity to end devices and people. Another important factor is that Edge computing is more affordable than Cloud computing. And as discussed earlier Edge computing enables AI to work with a more diverse spectrum of applications than cloud computing.

4. **Security**

CSP’s have improved their security layer but still certain industries have sensitive data which cannot be pushed to public clouds due to regulations. Edge computing keeps the data in the local IT hardware with organizations having full control over data. AI-enabled solutions can also detect anomalies at the edge network and can implement tactics in real time to avoid these cyber-attacks.

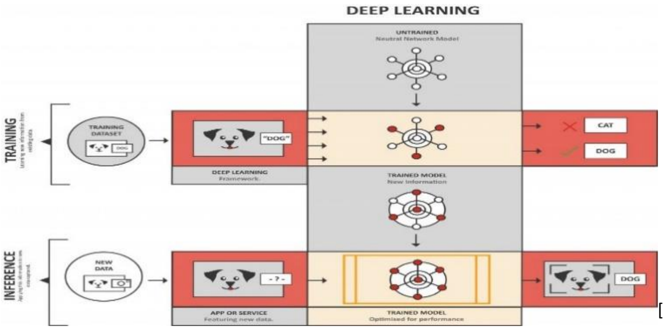
With Edge implementations, organizations identify all possible access points for a hacker and can implement risk mitigation techniques.

Inference

Inference is also a driving force why AI on Edge makes sense. It’s better to first understand what inference is and how it differs from Deep Learning/Machine Learning components of AI.

Deep Learning is the process creating an algorithm to recognize whatever you need it to, such as faces in CCTV footage or defects in product manufacturing.

Inference is the process of taking the above mentioned algorithm (model) and deploying it onto an end user device, which will then process incoming data (usually images or video) to look for and identify whatever it has been trained to do.



[3]

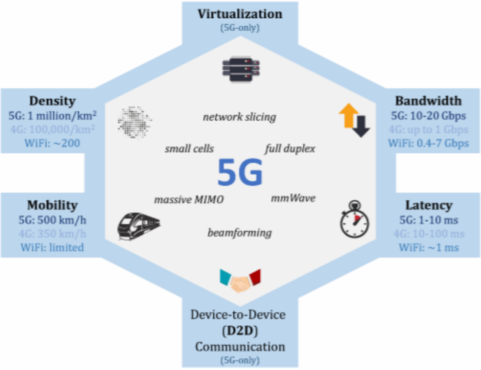
If you are using non-critical workload then having inference on cloud makes sense. However, when working on performance-sensitive or mission-critical workloads, inference should be executed on Edge locations or devices which gives real time results.

Cleary, for real time applications (mission critical and highly sensitive) such as facial recognition or detection of defective products in manufacturing, it is important that the result is generated as quickly as possible, so that a person of interest can be identified and tracked, or the faulty product can be quickly rejected.

Technologies making AI on Edge possible and feasible

1. **5G[4]**

Fifth Generation (5G) mobile communication is here and with its arrival the most important bottleneck for edge computing is removed. At present we agree it is deployed in small areas but its presence is felt in almost all continents but is majorly available in Europe and USA. In future, 5G is predicted to account for almost 15% of the mobile communication network by 2025. Its high bandwidth capabilities of almost 20GBPS, massive device density of almost 1 million devices per square kilometer, low latency of 1 ms and virtualization capabilities is opening new doors for computing. Use cases which were envisioned once like healthcare based on VR, AR, machine-to-machine communication in automotive and smart drones are possible only on 5G. Extensive use of new technologies like NFV, SD WAN with 5G, and basic cell stations have been transformed into mini data centers.



2. AI Accelerators

Chip manufactures have created purpose-built accelerators to enable AI on Edge and bridge the gap between the edge and data centers. Some of the options available are:

- NVIDIA Jetson
- Google Edge TPU
- Intel Movidius and Myriads Chip
- Baidu KUNLUN (Baidu and Samsung Semi-conductor)

3. Powerful Intel Processors[5]

Intel has created a portfolio of processors so that organizations can design a right-sized environment and investment to their application and network. These processors are:

- Intel® Atom® processors which offer up to 40Gb/s packet processing with a low power requirement.
- Intel® Core® processors suited for applications that require high media performance.
- Intel® Xeon® D processors which offer up to 190 Gb/s packet processing and have integrated ethernet and acceleration.
- Intel® Xeon® Scalable processors which offer up to 580 Gb/s packet processing on a dual socket platform.
- 2nd generation Intel® Xeon® Processors also have specially optimized variants for Network function virtualization (NFV) compared to previous generation of Intel Scalable processors. For example, Intel Xeon Scalable 6252N processor operates at 10% percent higher frequency than its predecessor (Intel Xeon Scalable 6252), 2.3Ghz. Using Intel ®Speed ®Select technology can optimize NFV performance and power consumption.

2nd generation Intel Xeon Scalable processor support for Intel® Optane™ DC persistent memory which can be used to store and process data for more applications such as content delivery networks (CDNs), VR, AR and image recognition.

4. HCI, AI Ready and NVMe Solutions

One of the challenges of placing AI workload on edge is reliable compute power. Another challenge with these implementations is that these data centers will be located at remote sites and sending an engineer every time to fix things will not be possible. These infer is that a high level of automation is required, such systems can run themselves with little intervention and, if a certain amount of human intervention is required, a majority of administrative tasks should be able to be performed remotely. This problem is mitigated by HCI solutions and AI ready solutions which are already widely deployed and whose management layer sits atop allowing an administrator to manage these systems remotely regardless of location. These systems have a much smaller footprint making them an optimal choice of mini and micro data centers. These systems also have a measure of redundancy for greater reliability. Feeding data to accelerators is also a key requirement which implies that a high performing storage layer should be there. This can be best delivered through Flash storage, Intel Optane DC SSDs using NVMe these days, as these provide significant I/O performance and reduced latency. Another advantage of amalgamating edge computing with HCI for AI is that it requires less storage space. The best operational feature of HCI is that the technology can function within a smaller hardware design. [5]

Dell EMC offers a broad spectrum of Ready Solutions for AI and HCI including everything you need to accelerate your AI initiatives. Helping make artificial intelligence simpler, these pre-designed, pre-validated solutions are ideal for machine and deep learning so you can get faster, deeper insights into your customers and your business.

Deciding if you are ready to move AI workloads to Edge

The benefits of using Edge Computing are compelling but only if it’s a right choice for you.

According to 2018 research published by MIT Sloan Management Review in partnership with The Boston Consulting Group (http://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/), organizations primarily fall under four different groups:

1. Pioneers (19%) Organizations that both understand and have adopted AI. These organizations are on the leading edge of incorporating AI into their organization's offerings and internal processes.

2. Investigators (32%) Organizations that understand AI but are not deploying it beyond the pilot stage. Their investigation into what AI may offer emphasizes looking before leaping.

3. Experimenters (13%) Organizations that are piloting or adopting AI without deep understanding. These organizations are learning by doing.

4. Passives (36%) Organizations with no adoption or much understanding of AI.

If you fall under Passives, Investigators or Experimenters, it’s better to go with a partner or a CSP as they will help you in different implementation models and also provide a platform to start using AI without having a big AI team in place and investing those dollars into hardware when you don’t have the right skill set to utilize it.

If you fall under Pioneers and Investigators who grasp what AI can unlock for your business, you need to fully understand your operational goal and also if AI on Edge is best for you. You can start by asking the questions below and understand which operational model is best for you.

Cloud-Based Model	**Edge Computing Model**
You don’t need to drive real time decision.	Need collected data to drive real time insights.
Latency in data transmission is accepted.	Near instantaneous data transmission is required.
Can support bandwidth requirement to send large amount of data.	You need local network support.
Network downtime does not affect productivity.	Network downtime affects productivity.
Data is not sensitive and confidential.	Highly confidential and sensitive data.

[1]

Conclusion

With the enormous growth in data generated by connected devices of the world and tools for analyzing this data being hosted on clouds, issues such as latency, high bandwidth costs, less exposure to scenarios for algorithms, and data security have become more prominent.

Introduction of 5G, HCI NvMe-based storage layer, 2nd generation processors, and purpose-built chipsets has enabled Edge computing to host these compute-intensive AI workloads. This implies that cloud is not the only place to host these workloads.

Due to the rise in both AI and IoT there is a pressing need to bring AI workloads in front of the cloud and place them on the Edge, heralding the rise of an Edge Intelligent(EI) world.

But again, it is important to identify the stage of your organization’s AI journey as it makes little sense investing in AI at the edge without having a specific business purpose in mind. Assessment of the costs and benefits of choosing edge over other deployment models is a necessary first stage. Fortunately, tthe Intel AI Developer Program offers resources to help with creation of AI projects from the data center to the edge. As well, Dell EMC AI ready solutions, HCI systems and Dell EMC Cloud platform enable you to deploy the needed hardware for on-premises cloud or Edge data centers without much hassle and save you cost and time, freeing you to seamlessly move your workloads between different cloud platforms when needed.

References

1. AVNET-ai-at-the-edge-whitepaper.pdf
2. Zhi Zhou, Liekang Zeng, Xu Chen, Ke Luo, En Li, Junshun Zhang .: “Edge Intelligence Paving the Last Mile of Artificial Intelligence With Edge Computing”
3. https://www.steatite-embedded.co.uk/what-is-ai-inference-at-the-edge/
4. Dumitrel Loghin, Member, IEEE, Shaofeng Cai, Gang Chen, Member, IEEE, Tien Tuan Anh Dinh, Feiyi Fan, Qian Lin, Janice Ng, Beng Chin Ooi, Fellow, IEEE, Xutao Sun, Quang-Trung Ta, Wei Wang, Xiaokui Xiao, Yang Yang, Meihui Zhang Member, IEEE, Zhonghua Zhang “The Disruptions of 5G on Data-driven Technologies and Applications”
5. https://www.intel.in/content/www/in/en/communications/platform-for-innovation-with-edge-guide.html
6. http://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/

Dell Technologies believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” DELL TECHNOLOGIES MAKES NO RESPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell Technologies software described in this publication requires an applicable software license.