

AI TRAINING AND INFERENCE AT THE EDGE

ARTIFICIAL INTELLIGENCE: FROM THE PUBLIC CLOUD TO THE DEVICE EDGE

Massive datasets call for entirely new compute and storage solutions

Contents

Introduction	1
Artificial Intelligence Explained	2
The Need to Think of AI from an End-to-End Standpoint	3
Three Critical Questions for AI	5
Why Equinix and NVIDIA?	9
AI as a Service Test Drive from NVIDIA and Equinix	10
References	10

Introduction

Artificial intelligence (AI) has become mainstream and is being leveraged by enterprises in meaningful ways to improve customer service, automate decision-making, predict trends with respect to opportunities and threats, and optimize business processes. It is estimated that 70% of enterprise applications will leverage AI in some form by 2021.¹

AI is making meaningful contributions to businesses in many industry verticals, including:

Robotics

Manufacturing, construction and navigation

Healthcare

Cancer detection, drug discovery and genomics

Internet Services

Image classification, speech recognition and natural language processing (NLP)

Finance

Trading strategy and fraud detection

Media and Entertainment

Digital content creation and game development

Autonomous Vehicles

Pedestrian and traffic sign detection, and lane tracking

Almost every enterprise recognizes the importance of AI in enabling true business transformation. In a recent study, 84% of surveyed executives feared they won't achieve their growth objectives if they don't scale AI. However almost 76% cited their struggle with how to scale AI across their business.² Many businesses are hindered by the complexity and cost of deploying the right infrastructure that can unleash AI-fueled insights from data. This report discusses the need for businesses to look at AI deployment in a holistic end-to-end manner.

Increasingly, AI will not just take place in a centralized manner in a private data center or in a public cloud. Instead, due to data compliance regulations, cost and latency reasons, AI processing will happen in a distributed manner at both the core clouds and at various types of edge locations close to the end devices. In this report we describe how Equinix and NVIDIA are jointly

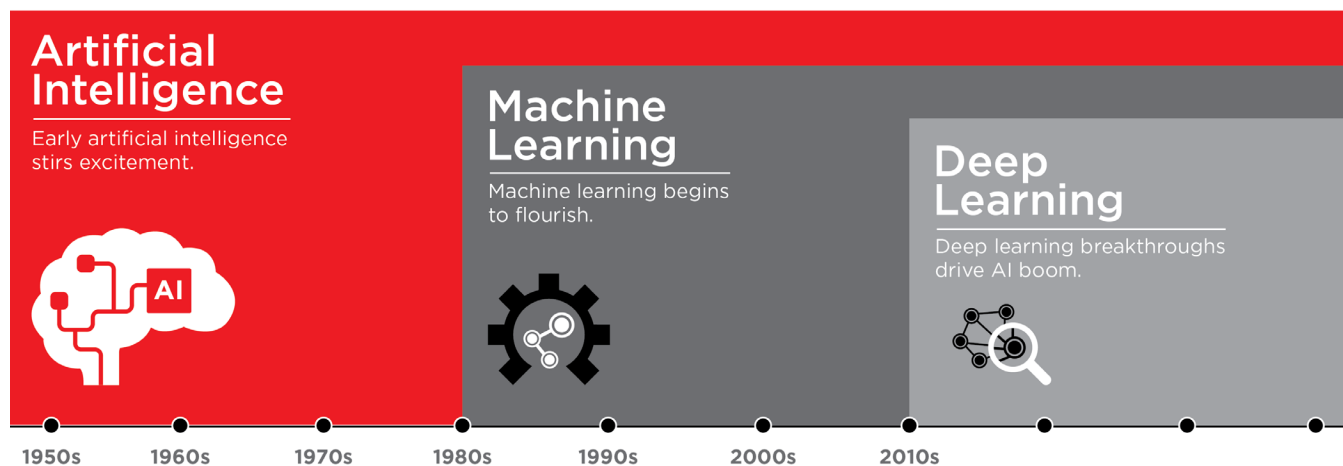


Figure 1: AI landscape

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence—first machine learning, then deep learning, a subset of machine learning—have created ever larger disruptions.

helping with distributed AI applications by providing AI infrastructure that spans across the continuum from edge data centers all the way to the core clouds.

Artificial Intelligence Explained

The confluence of big data, access to massive computing power and innovation in open source-based AI frameworks and algorithms has led to a renaissance in the adoption of AI in enterprise applications. Around 2012, researchers discovered that the modern GPU was well-suited to meet the massive computation demands required for training deep neural networks (deep learning).³

When researchers combined the power of GPUs with very large datasets, traditional neural network algorithms started showing a lot of promise for problems that they had previously not been good at solving (for example, image recognition and speech recognition). Furthermore, problems that had taken weeks or months to solve using traditional CPUs now could be solved in hours or days using GPU-based massively parallel deep learning architectures.

Figure 1 shows the history of developments in the AI field. Machine learning (ML) at its most basic form is the practice of using algorithms to parse data, learn from it, and then use the analysis to decide about or predict something in the world. So rather than programmers hand-coding software routines with a specific set of instructions to accomplish a particular task, as they did with AI algorithms in the 1970s and 1980s, in machine

learning the machine is “trained” using large amounts of data and algorithms that give it the ability to learn how to perform the task.

But this approach to AI is based on supervised learning of engineered features found within the data. This means that ML needs to be supported by developers who have an intimate understanding of the subject domain, the data under analysis and the relationships that exist within that data. ML is exceptionally useful for structured data where an algorithm can be employed to expose known patterns, such as making predictions on time-series data.

Deep learning is a special form of ML and is based on neural networks technology. Neural networks are inspired by our understanding of the biology of our brains—all those interconnections between the neurons. But unlike in a biological brain, where any neuron can connect to any other neuron within a certain physical distance, these artificial neural networks have discrete layers, connections and directions of data propagation.

One might, for example, take an image and chop it up into a bunch of tiles that are input into the first layer of the neural network. Individual neurons in the first layer then pass the data to a second layer. The second layer of neurons does its task and passes the data on to the third layer and so on, until the network produces the final layer and the final output.

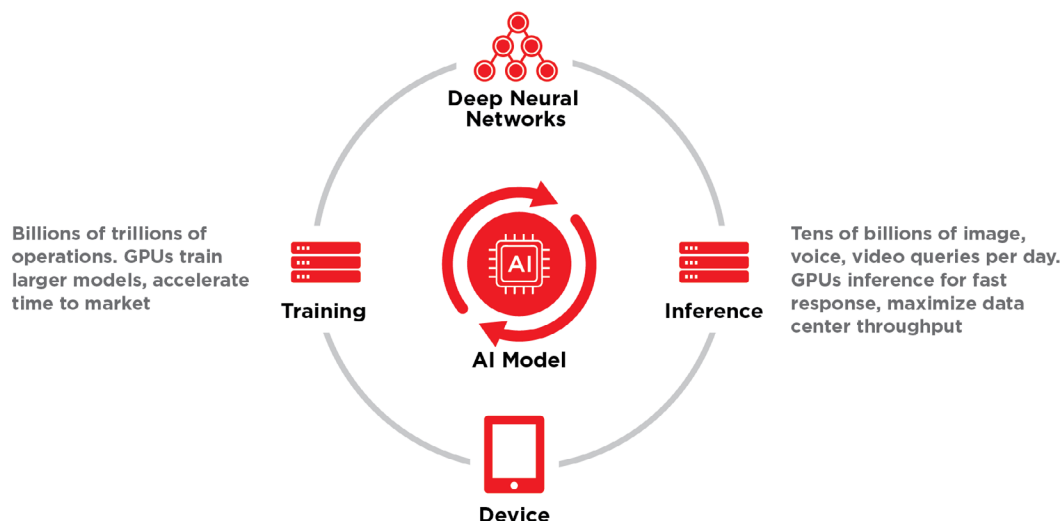


Figure 2: AI model training and inferencing

Each neuron assigns a weight to its input with respect to how correct or incorrect it is relative to the task being performed. The final output is then determined by the total of those weights.

We now describe an example of recognizing a stop sign using deep learning technology. Attributes of a stop sign image are chopped up and “examined” by the neurons: the sign’s octagonal shape, its fire-engine red color, its distinctive letters, its traffic-sign size and its motion or lack thereof. The neural network’s task is to conclude whether it is a stop sign or not. It comes up with a “probability vector,” a highly educated guess, based on the weighting. In our example, the system might be 86% confident the image is a stop sign, 7% confident it’s a speed limit sign, 5% it’s a kite stuck in a tree and so on—and the network architecture then tells the neural network whether it is right or not.

Training vs. inference

As shown in Figure 2, AI consists of training and inferencing components. Deep neural networks are “trained” by running massive datasets repeatedly through a network to optimize its parameters and eventually deliver a model with the highest possible predictive accuracy. In some scenarios, such a model is better than humans, and that ranges from recognizing cats in videos to identifying indicators for cancer in blood and tumors in MRI scans. Google’s AlphaGo learned the game and trained for its Go match—it tuned its neural network—by repeatedly playing against itself. Neural network training is a very compute-intensive

process that most often happens in a data center or in very close proximity to where the data (for training) resides.

A trained neural network is put to work using what it has learned—to recognize images, spoken words or a blood disease, or to suggest the shoes someone is likely to buy next. This speedier and more efficient version of a neural network “infers” things about new data it is presented with based on its training, and this is known as inference. Inference is typically executed in close proximity to where “new” production data is collected, namely data on which a prediction must be made by running it through the trained, optimized network. This can include a very wide set of locations such as the factory floor, an oil rig or a self-driving car.

The Need to Think of AI from an End-to-End Standpoint

Historically, AI model training and inferencing have taken place mostly in a centralized manner at a core data center. The core data center can be either an enterprise’s private data center or a public cloud. However, as shown in Figure 3, this architecture is changing, and AI architectures are becoming more distributed across core and edge locations. This trend is primarily due to data location at the edge, real-time latency, data sovereignty, regulatory compliance and cost reasons.

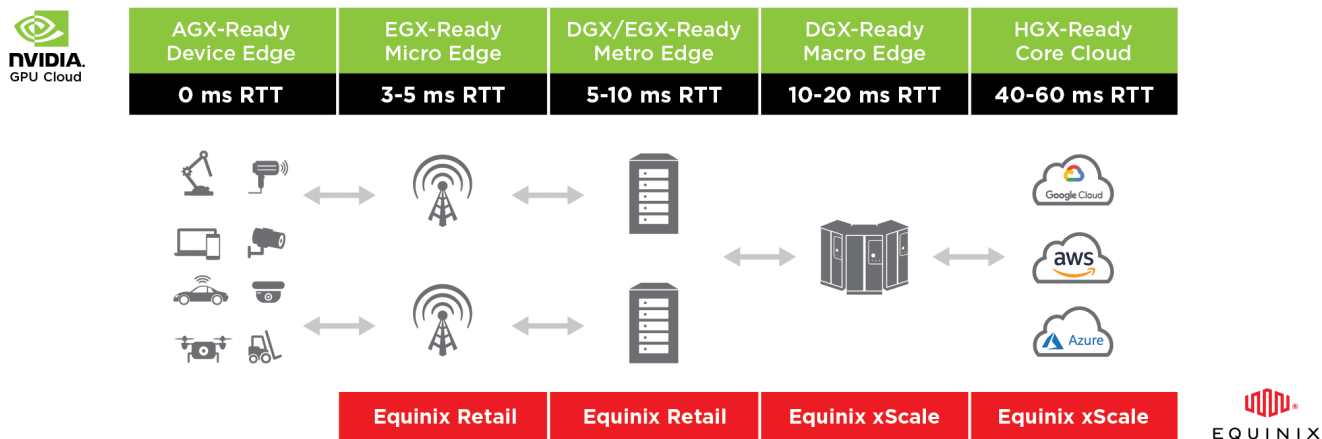


Figure 3: End-to-end AI solution from NVIDIA and Equinix

In this section we first describe the characteristics of the different types of “edges,” and in the subsequent sections we discuss the desired locations for AI training and inferencing operations. In general, enterprises are finding it advantageous to follow the mantra “train where your data lands” in order to minimize costs and ensure fastest time-to-solution on AI model training, and following a similar approach for running inference on production data.

To establish a framework for how AI infrastructure is deployed across a distributed enterprise, we are classifying the different types of edges based on the round-trip network latency (not including the time spent on compute) from the end device or application. A given distributed application can have its microservices running across either a single or multiple of these edge locations in the edge hierarchy (including device edge, micro edge, metro edge, macro edge and cloud).

Device edge (0 ms)

Smartphones, smart cars and internet of things (IoT) devices of various types are all examples of different types of device edges. Here is where the data gets generated and actions (if active devices) take place. Device edges vary in the amount of computational power, whether they are fixed or mobile, how they are connected to the network (for example, low power, wifi, 4G/5G, wired and so on), and whether they can operate without network connectivity. Going forward there will be billions of internet-connected edge devices.⁴ The NVIDIA AGX solution has been designed for device edges.

It brings AI to embedded applications in a compact form factor (power consumption is less than 30 watts) and is being leveraged in nearly every industry (for example, drones, autonomous cars, robots and so on).

Micro edge (3-5 ms)

Closets in shopping malls, stadiums, offices, telecom central offices, parking lots, basements in apartment buildings and cell towers are some examples of micro edges. With the advent of 5G networks, the conversion of the cellular traffic into IP networks will also most likely take place at the micro edges. The primary copy of data is not usually persisted here. Data filtering, stream processing and AI inferencing usually take place here. We quantify locations as micro edges if they can offer a round-trip network latency of 3 to 5 milliseconds (ms). Typical power capacity for micro edge data centers varies from 250 kW to 2 MW.

It is important to note that in many major metro markets current Equinix data centers can satisfy the low latency (3-5 ms) requirements, and thus can act as micro edge for a large segment of use cases. Small-sized micro edges are good places to host AI inferencing infrastructure but typically do not have the capacity (power wise) to support AI training infrastructure. The NVIDIA EGX platform has been designed to operate in micro edges. The EGX portfolio presents an enterprise-grade platform whose compute capability scales from a half trillion operations to 10,000 trillion operations per second. Furthermore, power requirements-wise it can be deployed in a micro edge data center and be used for inferencing purposes.

The EGX platform also supports a cloud-native virtualization execution environment where application programs can run in containers (for example, Docker).

Metro edge (5-10 ms)

These are edges that are in key metro markets globally and can support 5 to 10 ms of round-trip network latency from the end device. Both data caches as well as persistent stores are hosted here. These edges are suitable for handling real-time workloads. Both training and inferencing tasks can be performed at metro data centers. Data flows through here from edge devices en route to the public clouds. There are three types of metro edges:

Interconnection-rich colocation data centers

These multitenant data centers (MTDCs) can handle the power requirements (up to 30 kW per rack) of AI training hardware. Furthermore, they are rich in interconnections—that is, many clouds, networks and enterprises interconnect with each other at these locations over private and secure interconnects. Equinix data centers are interconnection-rich colocation data centers. These data centers are also known as retail data centers, colocating 50+ tenants per data center with 20 to 300 kW average power draw per tenant. The typical power capacity of an Equinix metro edge data center is in the 20 MW range.

Private data centers

Private enterprise data centers typically are not designed to handle the high power requirements of AI training hardware.

Colocation data centers

These are third-party MTDCs that can handle the power requirements of AI training hardware but do not have high connectivity to many clouds or networks, and the scope of most of these providers who are operating these facilities is not global in nature.

Both the NVIDIA EGX and DGX platforms can be hosted at a metro data center (from a power-draw standpoint) for training and inferencing purposes. The DGX platform is the industry-leading AI platform that provides up to 2 petaflops of computing power that can be used to tackle large-scale compute intensive AI training jobs. The DGX platform can also support the hosting of cloud-native container-based applications.

Macro edge (10-20 ms)

These are compute- and storage-dense data centers that typically host fewer but larger tenant deployments. These colocation data centers are also known as “wholesale” data

centers (colocating 1 to 10 tenants per data center with 300 kW to 5+ MW average deployment per tenant), whereas metro edges are known as “retail” data centers. These data centers are not present in every metro and are also not typically interconnection rich.

Macro edge data centers are larger in size (approximate power capacity of around 50 MW) in comparison to the metro edge data centers. These data centers also can host AI training hardware. Equinix xScale™ data centers qualify as macro edges. Some enterprise private data centers also qualify as a macro data center due to their round-trip latency characteristics from the end user device. The NVIDIA DGX platform can be hosted in macro edges for both AI model training and inferencing.

Three Critical Questions for AI

An enterprise must answer the following three questions as it determines how to integrate AI processing (both model training and inferencing) as part of its production-level application architecture:

1. Where should you do AI model training and model building?
2. Where is your edge for AI model inferencing?
3. How do you connect the edge inferencing to the training?

Where Should You Do AI Training and Model Building?

AI model training operations can take place at multiple different types of edges as shown in the edge hierarchy in Figure 3. The following key factors determine the appropriate edge for doing AI model training:

Power Requirements

AI model-training workloads are usually compute and data intensive. Typically, these workloads deal with terabytes of data and require gigabytes of system bandwidth. A fully loaded rack of model-training equipment can consume more than 30 kW per rack. This type of power draw cannot be supported by micro edge data centers and most private enterprise data centers. Thus, model training for most use cases can be performed either in public clouds or in metro edge and macro edge data centers. Equinix provides such data centers globally in more than 50 markets.

Performance Considerations

When performing AI model-training operations, it is important to perform training operations close to where the data is getting generated. Thus, it makes sense to have compute close to where the data resides. The latency associated with moving massive datasets to a far-off remote location for model training impacts the overall time for model training. Secondly, there is performance degradation with the presence of a virtualization layer in a multitenant AI stack. That is, AI training operations running in a container on a bare metal platform perform better than doing AI training on a multitenant virtualized AI platform.

Data Gravity and Data Privacy Requirements

Data that is used as input for AI model training could be generated in clouds, generated by devices or systems at the edge, or procured from external data brokers. Increasingly, the size of the data being generated at the edge is quite large (for example, 4 terabytes [TB] per airplane per day, 3 TB per autonomous car per day, 1 petabyte [PB] per smart factory per day and so on), and moving this data into the core cloud for AI model training is expensive. Thus, enterprises are looking at AI architectures that allow for the processing of data at the edge. Federated AI learning techniques (in certain use cases) allow for the creation of local AI models at the edge. In this approach, one ships only these local AI models—not large amounts of raw data—to the core data center to build a more accurate global AI model. Thus, federated AI learning techniques help reduce the cost of backhauling data from the edge.

Data privacy is another reason why AI model training is moving to the edge. Around 132 countries in the world have data residency or data privacy laws either already in place or are in the process of legislating them.⁵ As a result, enterprises that have operations and data in multiple countries need to process that data locally and are employing distributed AI learning techniques to avoid transferring raw data, in keeping with the privacy and compliance regulations. In many cases, these enterprises want to procure data center space from a single provider that can offer data center services in multiple markets globally, providing the companies a consistent data center experience in all their markets. Equinix has data centers in 26 countries and 56 markets, and thus is attractive to enterprises with a global presence. Public clouds and macro edge vendors also have data centers in multiple markets, but usually they cannot match the global reach of Equinix. Thus, in many cases, they leverage Equinix to increase their global footprint in multiple markets.

Data Aggregation

In many situations an enterprise needs to import data from external sources in order to improve the accuracy of its AI models because it does not have enough data of a certain type or it needs to fuse different variety or types of datasets (for example, weather, traffic and so on). These external datasets can reside in multiple clouds, private data centers or data brokers, or they can be generated at the edge via IoT devices. Furthermore, an AI application does not exist in isolation. It needs to be integrated with various other IT systems. Thus, it is important to host the AI model-training hardware at a location that is well connected via high-speed and secure networks to these different data locations. Equinix is an interconnection hub where many clouds (2,900+), network providers (1,800+) and enterprises (3,000+) come to interconnect with each other. Thus, it is an ideal location for AI data aggregation and analysis.

Simplicity

Public clouds have integrated AI training hardware and open source software, and they have made it easy for data scientists to consume AI services. They also offer AI training as an OPEX service, and this is attractive for enterprises that cannot make upfront CAPEX commitments. Together, NVIDIA and Equinix are now enabling AI as a Service at Equinix data centers. This service integrates NVIDIA® DGX™ Systems, with NVIDIA AI-optimized software in combination with Equinix data centers, making AI in a colocation data center as simple as consuming it in public clouds. Now organizations can benefit from performance and privacy/control benefits of hosting AI in private data centers while enjoying the simplicity and OPEX model benefits of AI in public clouds.

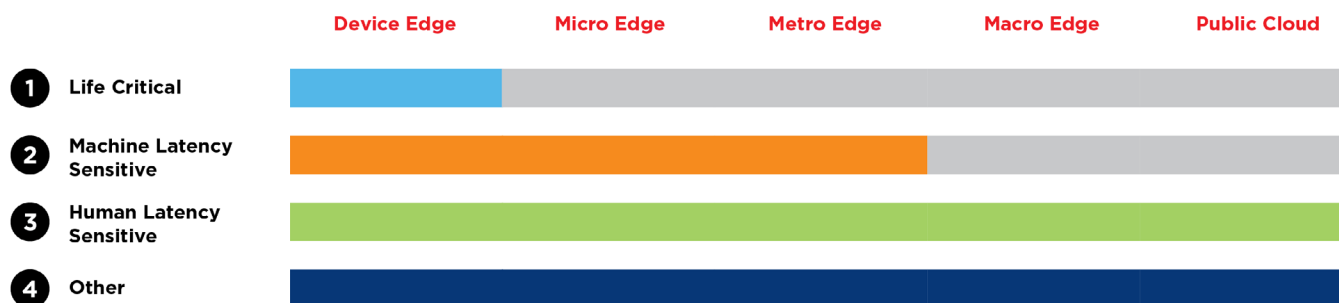


Figure 4: Location of AI inferencing operations for different application archetypes

Where Is Your Edge for Inferencing?

Historically, AI applications did both model training and subsequently model inferencing at the same location. However, increasingly AI inferencing operations are getting decoupled from training locations and are being performed closer to where the data is getting generated. The following are some key factors that enterprises should consider when determining the location for model inferencing.

Latency and Cost Considerations

Different applications have different latency requirements, and thus, AI inferencing can be done at the different types of edges in the edge hierarchy. As enterprises move toward real-time customer engagement models, more applications work in real time. Furthermore, for cost reasons, enterprises will try to do inferencing as far up in the edge hierarchy as possible because that will require them to deploy fewer inferencing nodes at the edge. Figure 4 shows different application archetypes, their inferencing latency requirements and the appropriate edge type that can satisfy those requirements.⁶

Life Critical

Some health, autonomous vehicle and drone applications cannot tolerate network downtime and are extremely latency sensitive. These applications can endanger human health if latency requirements are not satisfied. Inferencing for these applications in most cases will take place at the device edge.

Machine Latency Sensitive

Commodity and stocks trading financial applications, industrial robots, security and surveillance applications, military systems, and content distribution applications need to react and take actions at machine speeds. In

most cases these application latency requirements can only be satisfied by device, micro- or metro-level edges.

Human Latency Sensitive

Some augmented reality (AR) and virtual reality (VR) applications cannot tolerate latencies that are greater than 5 ms. Similarly, voice recognition applications, multiuser gaming applications and smart in-store shopping applications need to run closer to the end devices. Thus, these applications need to run on device, micro- or metro-level edges. Many web applications need to react in real time but can tolerate higher latencies in the 100 to 200 ms range, and thus, the inferencing for these applications can be performed at any of the edges in the edge hierarchy.

Other

Inferencing for other types of applications can take place in any of the nodes in the edge hierarchy. Typically, they will get performed at the same locations as AI training in order to reduce infrastructure deployment costs.

Device Resource Constraints

There is a wide spectrum of edge devices with different network, power and compute capabilities. In some use cases, inferencing will get done on the edge device itself (for example, autonomous car, smartphone and so on), whereas in other use cases, inferencing will get done at a micro data center (for example, for a retail AR/VR application, the AI inferencing will get done in the store wiring closet because the AR/VR goggles don't have enough compute power). With the emergence of high-speed and low-latency 5G networks, increasingly more processing can be offloaded from the edge devices onto the micro edge data centers, and still meet the application's real-time latency requirements while helping to reduce the overall AI inferencing equipment costs.

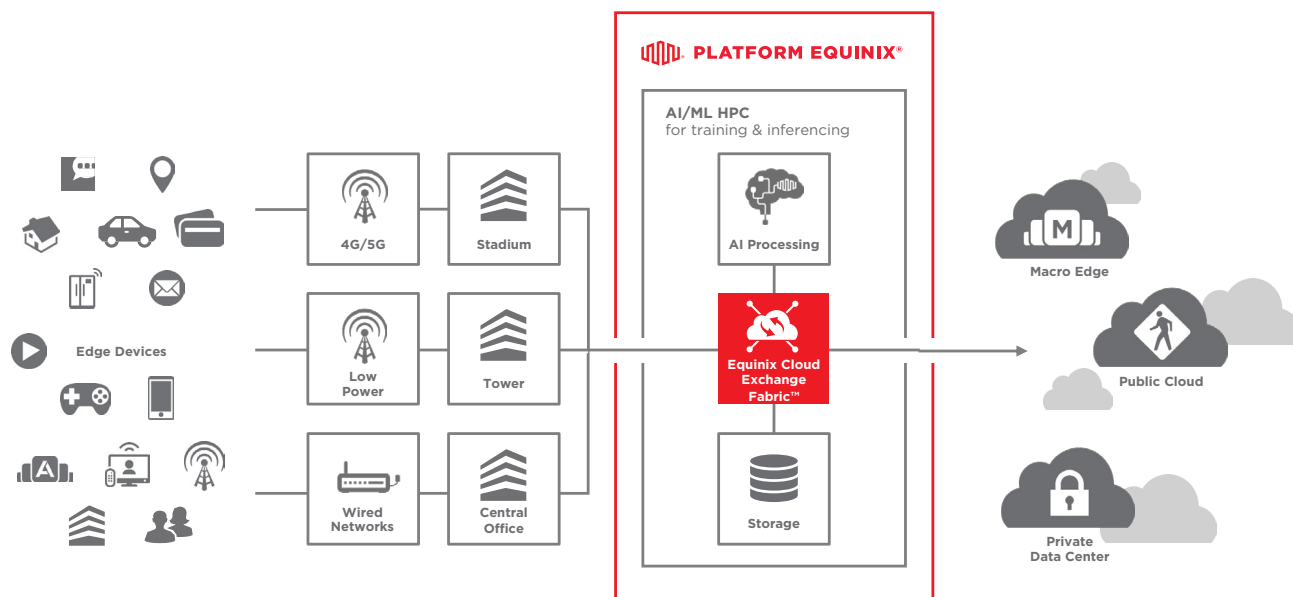


Figure 5: AI processing at an interconnection hub

Data Privacy Requirements

There are use cases where customer or citizen data cannot leave an area or region for data privacy reasons. In those situations, depending on the latency requirements, AI inferencing will most likely get done in the device, micro- or metro-level edges.

Data Aggregation Requirements

In use cases that require fusing data from multiple sources during inferencing, the inferencing tasks will most likely take place at an edge location that is capable of caching or storing the different data sources or data streams and that is well connected via high-speed networks to these different data sources (for example, clouds, data brokers and private data centers). Typically, for use cases with data aggregation requirements, inferencing will most likely get done in an interconnection-rich metro data center.

Service Availability

Depending on the service availability requirements and device connectivity reliability (for example, there might be intermittent connectivity in airplanes, oil rigs in the ocean, mining equipment and so on), AI inferencing can take place either on the device or at the micro edge at the device location.

“In around 80% of the global metro markets, Equinix data centers are located within less than 10 ms round-trip network latency, and in many large metro markets they can support a round-trip latency of even less than 5 ms.”

Thus, Equinix data centers are suitable for hosting most AI inferencing applications except for life-critical applications or machine latency-sensitive applications.

How Do You Connect the Edge Inferencing to the Training?

Figure 3 shows the logical view of how the edges are connected in the edge hierarchy, whereas Figure 5 shows the physical topological view of how the various edge nodes are connected in the edge hierarchy.

The following is a description of how the various edge nodes are connected and the different interconnection options.

Device Edge to Micro Edge

There is an explosion in the number of types of wireless protocols and wireless network providers. Low powered (LoRA), CBRS (free spectrum), wifi, 3G/4G/5G, satellite and so on are some of the different types of wireless network providers that bring traffic from the various types of edge devices to the micro and metro data centers. In addition, wired network providers also bring traffic from wired IoT devices to the micro and metro edge nodes.

Micro Edge to Public Clouds

It is mostly network providers that are helping to move traffic from the edge devices to multiple public clouds and enterprise private data centers for AI model training and inferencing. For connectivity to multiple clouds, network providers usually bring traffic from the micro edges to the public clouds via highly interconnected metro edges (using public internet, MPLS networks or SD-WAN networks). Thus, each network provider needs to interconnect with multiple clouds, and similarly each cloud provider needs to interconnect with numerous network providers. Thus, for cost reasons, network providers and cloud providers peer with each other at interconnection hubs (like Equinix data centers) via exchanges. These exchanges (like Equinix Cloud Exchange Fabric™) make it easy for providers on both A (the network side) and Z (the cloud side) to peer with each other in a cost-effective manner.

Macro Edge to Public Clouds

Macro edges typically only have connectivity to a few network and cloud service providers, and thus, they usually have a network edge presence at metro edge data centers in order to route traffic to multiple network and cloud service providers.

Metro Edge to Public Clouds

Cloud service providers, network service providers (both wired and wireless providers) and enterprises have their network edge infrastructure at an interconnection-rich colocation data center like Equinix. These providers interconnect with each other via cross connects (layer 1), virtual circuits (layer 2) or IP routes (layer 3). Since in most markets Equinix data centers are located within 1 to 2 ms from the public cloud data centers, they are used as the networking on-ramp to the public clouds. That is, the clouds interconnect with network providers at Equinix.

Why Equinix and NVIDIA?

Many organizations want industry-leading AI infrastructure to enable their business transformation, but struggle with the facilities and deployment environment required to support the demands of AI workloads. NVIDIA DGX Systems are an industry-leading solution for high-performance, enterprise-grade AI training and analytics infrastructure. Similarly, NVIDIA EGX is the first platform to deliver enterprise-grade AI inference.

Developed to meet the demands of AI and analytics, NVIDIA Systems are purpose-built for the unique demands of the AI enterprise. Built on a revolutionary architecture, powered by the world's fastest data-centric accelerators, combined with innovative GPU-optimized software and simplified management tools, these fully integrated solutions deliver groundbreaking performance and results. NVIDIA Systems are designed to give data scientists the most powerful tools for AI exploration—from your desk to the data center to the cloud.

Increasingly, with the growth in the size of datasets at the edge, the need to adhere to data residency and privacy regulations, and the need for real-time response, organizations are moving toward distributed AI architectures where model inferencing and even model-training tasks are taking place near the edge. The combination of Equinix and NVIDIA is well-suited to target distributed AI architectures due to the following key reasons.

End-to-End AI Solutions

As customers design end-to-end AI solutions, NVIDIA is capable of providing an end-to-end line of AI training and inferencing systems as shown in Figure 3. The NVIDIA HGX, DGX, EGX and AGX line of solutions span from public clouds all the way to the device edge.

Similarly, Equinix data centers can host NVIDIA systems at the macro, metro and device edges. Together, Equinix and NVIDIA can provide customers with an end-to-end AI solution that spans across the different types of edges. Furthermore, Equinix data centers provide high-speed access to AI systems in the public clouds and allow for the deployment of hybrid AI systems, where compute resides in the public clouds but storage can reside at Equinix. Thus, customers can leverage AI innovation from multiple cloud providers.

Global Presence

Customer-distributed AI solutions can span across multiple regions/countries in order to satisfy government compliance and privacy regulations. More than 130 countries have either already legislated or are in the process of legislating data privacy and residency regulations. In order to satisfy these requirements, enterprises can deploy their AI infrastructure at multiple Equinix global locations (55+ markets in 26 countries). Thus, customers can get the same uniform, consistent service across all of these different regions and countries.

Distributed Fabrics

If the distributed inferencing and training locations are deployed at different Equinix data centers, then they can be connected to each other via high-speed private network fabric (Equinix Cloud Exchange Fabric™). Furthermore, customers can also deploy various federated learning frameworks (for example, Federated TensorFlow, PySyft, NVIDIA Clara, Xain and so on) to do distributed AI across multiple Equinix data centers.

Secure & High-Speed Access to External Datasets

Many AI solutions need to access external data that is spread across private data centers, multiple public clouds, IoT devices at the edge and data brokers. Equinix data centers are located at the interconnection junction between these different data source locations, and thus, are ideally situated for hosting AI training and in many cases also AI inferencing. More than 2,900 cloud companies, 1,800 network companies, 1,250 finance companies, 650 media companies and various other enterprises are already located at Equinix. Thus, an enterprise's AI solution at Equinix can easily tap into these different data sources via high-speed private networks.

Performance

From a performance standpoint, it is desirable to do AI model training and inferencing close to where the data is getting generated. Furthermore, doing AI on bare-metal AI hardware is more performant than doing AI operations on a multitenant virtualized platform in the

cloud. Equinix and NVIDIA together are providing bare metal-based AI solutions at the different types of edges as shown in the edge hierarchy (Figure 3). Equinix data centers are within 10 ms of end devices in most major metro markets (in some markets even less than 5 ms). Equinix has 210+ data centers in 55+ metro markets.

Simplicity

The NVIDIA integrated AI hardware and software appliance ecosystem makes it simple for enterprises to get a cloud-like AI experience. Now, NVIDIA and Equinix together have integrated NVIDIA systems with Equinix edge data centers and networking solutions to provide end-to-end AI as a Service at Equinix.

AI as a Service Test Drive from NVIDIA and Equinix

Equinix and NVIDIA, along with their partners NetApp and Core Scientific, are offering an AI as a Service test-drive setup at Equinix. Customers can use this test-drive setup for both training and inferencing. The goal behind this test drive is to allow customers to assess the capabilities and simplicity of using this joint AI as a Service solution while obtaining the above-mentioned benefits. Customers can access this AI testbed via a high-speed, secure network while persisting data in their respective private cages. This way they keep control over their data while leveraging AI software and hardware innovation from the AI as a Service test drive. After using this test drive, customers can either deploy a private AI stack in their own private cage at Equinix, or they can continue to pay and use AI as a Service at Equinix.

Take the test drive.

[Equinix.com/AI/TestDrive](https://equinix.com/AI/TestDrive)

References

1. Google AI Conference, San Francisco, 2019.

2. "AI: Built to Scale," Accenture Report, 2019.
<https://www.accenture.com/us-en/insights/artificial-intelligence/ai-investments>.

3. A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet Classification with Deep Convolution Neural Networks," Proceedings of the 25th International Conference on Neural Information Processing, 2012.

4. "Global Networking Trends Report," Cisco, 2020.
https://www.cisco.com/c/m/en_us/solutions/enterprise-networks/networking-report.html#.

5. Graham Greenleaf, "Privacy Laws Report," 2019.

6. "Defining Four Edge Archetypes and Their Technology Requirements," Vertiv Technical Report, 2018.
https://www.vertiv.com/globalassets/documents/white-papers/vertiv-edgearchetypes-wp-en-na-sl-11490_229156_1.pdf.