Joint solution brief
Artificial intelligence

AI DATA SCIENTISTS TEST DRIVE NEW SOLUTION

Equinix, NVIDIA, NetApp and Core Scientific pilot an Artificial Intelligence (AI) as a Service test drive at the metro edge

Overview

Equinix, in partnership with NVIDIA, NetApp and Core Scientific, has announced an Artificial Intelligence (AI) as a Service test drive setup at Equinix data centers. This solution brings together industry-leading AI hardware technologies from NVIDIA and NetApp, the global interconnection-rich data center platform from Equinix and best-in-class Cloud for Data Scientists™ software technology from Core Scientific.

With the Test Drive program, enterprises can explore AI as a Service at Equinix and decide whether they want to deploy at Equinix.

Challenge

There is no doubt that groundbreaking machine learning algorithms and the rise of deep learning have enabled businesses to attack some of their toughest challenges with the power of AI. Demand for AI compute power will only accelerate in the future. To empower data scientists to prototype, test and iterate on their AI models more rapidly, companies require direct access to advanced computing power and fast local storage. Proximity of datasets to compute resources becomes paramount, and in many use cases, data scientists need access to data coming from external sources such as sensors, cameras and equipment at the edge of the enterprise. To access that data efficiently, they need to host their AI infrastructure at a well-connected location with high speed, secure access to multiple clouds, private data centers and data brokers. It has become apparent that traditional CPU-based cloud infrastructure constrains data science workflow and does not scale to meet the demands of training increasingly large AI models and datasets.

Data scientists need unfettered access to powerful tools and a platform that enables rapid iteration to deliver the best models with the highest predictive accuracy, in the shortest time possible, if businesses are to reap returns on their AI investments. Data science teams expect “ready now,” IT-approved AI platforms that can be provisioned quickly and easily from a web UI. Successful AI development requires teamwork with orchestration of a myriad of tools, including GPU-accelerated data analytics and computing in tandem with high-performance storage, data management and high availability. Furthermore, concerns about data center readiness, data locality, escalating

About NVIDIA

NVIDIA pioneered accelerated computing to solve problems that normal computers cannot solve. The company innovates at the intersection of graphics, HPC and AI.

nvidia.com

About NetApp

NetApp is the leader in cloud data services, empowering global organizations to change their world with data. Together with our partners, we are the only ones who can help you build your unique data fabric.

netapp.com

About Core Scientific

Core Scientific’s mission is to be the premier AI and blockchain provider, delivering best-in-class infrastructure and software solutions for a rapidly evolving market.

corescientific.com

cost, scalability and the rapidly evolving nature of AI innovation and infrastructure can easily delay time to insights. These challenges necessitate a full-service solution that can empower data science teams.

Due to data residency, compliance, performance requirements (of moving large datasets to far-off core clouds) and cost reasons (for backhauling large datasets to core clouds), it becomes critical for businesses to place their AI compute infrastructure co-resident with their data, following the mantra of "train where your data lands."

Furthermore, AI platforms cannot exist in isolation. They need to be integrated via secure and high-speed networks to an enterprise's corporate IT systems, which can exist in private data centers and public IaaS and SaaS clouds. Thus, there is a need to connect AI systems with the rest of an enterprise's IT infrastructure.

Finally, AI platforms need to be situated where they can ingest data from multiple sources in order to fuel model prototyping and improve accuracy. In many instances these datasets reside in multiple public clouds, data brokers and private data centers, and they are also getting generated at the edge. Thus, it is desirable to host the AI infrastructure at an interconnection hub that has high speed and secure access to these different data sources.

Solution

Equinix, NVIDIA, NetApp and Core Scientific together offer a fully integrated AI as a Service platform, available in the Test Drive program, to help businesses meet the rapidly growing demands of AI development in every industry. The key highlights of this solution are:

- AI as a Managed Service: A cloud-native (container-based) AI as a Service offering that makes it easy for data scientists to consume AI services.
- Industry-leading AI technology stack: Consisting of high-performing compute, network and storage technologies that are optimized for running all the major AI software frameworks.
- AI stack at global interconnected metro edge: A global and highly interconnected data center platform that provides high-speed and secure interconnection to IT systems and data sources that are spread across public clouds, private data centers and edge locations.

AI as a Service

Core Scientific's Cloud for Data Scientists™ brings together innovation in cloud-connected all-flash storage systems with NVIDIA's leadership in AI and GPU computing to provide an on-demand AI solution. It includes Plexus™, which makes available GPU-optimized software containers and advanced orchestration and scheduling. Cloud for Data Scientists is available to customers to deliver the ease of the public cloud with the cost benefits of colocation.

Plexus™ by Core Scientific brings a complete solution for deep learning into a single web-based AI fabric that simplifies the way data scientists access advanced compute.

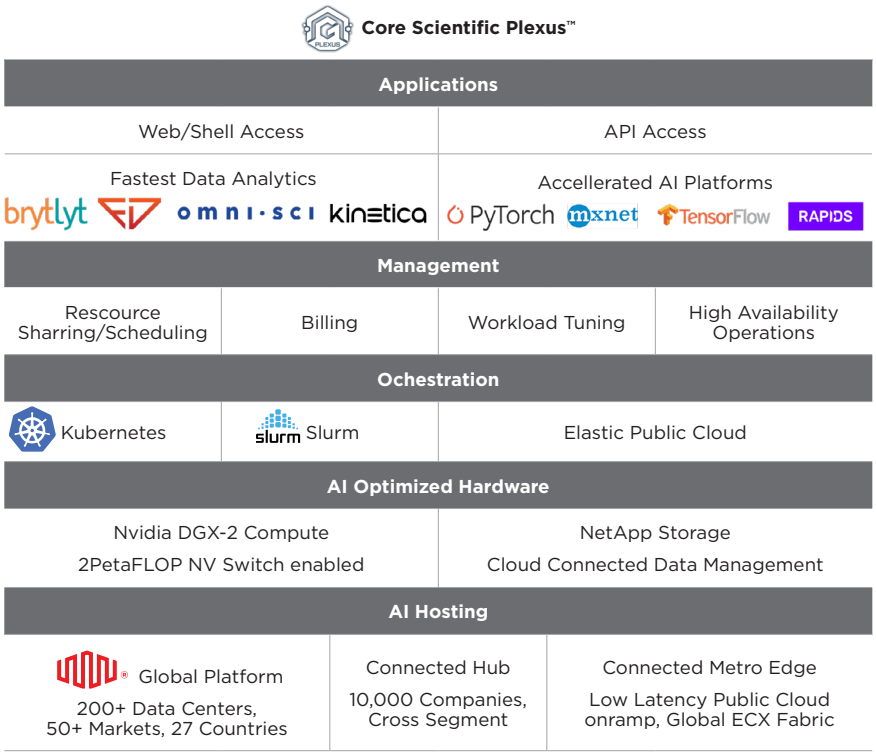
Features

Trusted partners
Rely on the solution that combines key strengths of Equinix, NVIDIA, NetApp and Core Scientific.

AI at the metro edge
Perform AI at 200+ securely connected Equinix International Business Exchange™ (IBX®) data centers worldwide in 55 global metros.

Simplified AI as a Service
Build the data science you need with high-performing tooling, workflows and resource management.

Industry-leading AI technology stack
Run containerized AI frameworks and ML algorithms on industry-leading GPU and storage technologies.



Ease of use with Core Scientific Plexus:

- Walk-up tooling, workflows and resource management for data scientists.
- Highly available Kubernetes and Slurm cluster management.
- GPU application portal including NGC containers and leading GPU analytics software applications OmniSci, BrytLyt, FastData.IO and SQream.
- Accelerated AI platforms: TensorFlow, Rapids, PyTorch and MXNET.
- OPEX model with on-demand scaling, both colocated and bursting to the public cloud. (In the Test Drive setup, there is no bursting of processing to public clouds.)
- Unlike in the public cloud, there are no ingress or egress fees for moving data in or out of the AI setup at Equinix.

Industry-Leading AI Technology Stack

The technology stack consists of industry-leading compute, network and storage hardware, along with an AI deep-learning software stack consisting of many open source AI frameworks that have been optimized to run on the NVIDIA® DGX-2 platform. This technology stack provides industry-leading performance on the leading AI benchmarks for both training and inferencing.[1]

- NVIDIA DGX-2: The world's first AI system to deliver 2 petaflops of AI performance in a single node, powered by the integration of 16 NVIDIA

Benefits

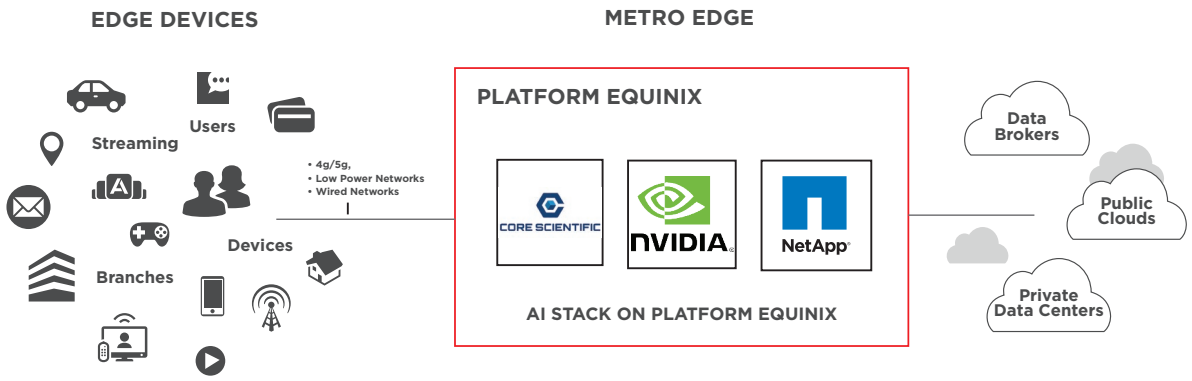
Keep control over your data
Avoid cloud lock-in and keep control over sensitive persistent data.

Control costs, cut latency and satisfy compliance regulations
Process data at the edge in order to cut data transfer costs, perform real-time inferencing and satisfy data residency requirements.

Interconnect with global ecosystems
Get secure and high-speed access to data sources to financial services, media networks, clouds and enterprise ecosystems.

Scale on demand
Build capacity to handle surges in data, and step it down to save when business slows.

1 Shar Narasimhan, "NVIDIA Clocks World's Fastest BERT Training Time and Largest Transformer Based Model, Paving Path for Advanced Conversational AI" NVIDIA Developer, August 13, 2019. https://devblogs.nvidia.com/training-bert-with-gpus/



V100 Tensor Core GPUs. With this technology you can train some of the world's most complex AI models on a single system. Its architecture is built on NVIDIA NVSwitch technology, which interconnects all 16 GPUs with an ultrahigh-bandwidth, low-latency AI fabric that enables the highest levels of AI model and data parallelism.

- NVIDIA GPU Cloud: A repository of containerized AI frameworks and ML algorithms that have been optimized for the NVIDIA DGX Systems and are available as part of this software stack.
- NetApp AFF A800 Storage System: Cloud-connected all-flash storage that spans from edge to cloud, can scale up to 20 petabytes and can support up to 400 billion files. Test Drive will be a scaled-down version of this storage capacity.
- Mellanox Networking: Providing a high-performance AI network fabric using 8x100 Gbps bandwidth to each DGX system, with ultralow latency supporting both EDR Infiniband and 100 Gigabit Ethernet connectivity.

AI Stack at Global Interconnected Metro Edge

In addition to having an industry-leading fully managed AI stack, it is important to host this stack at an interconnection hub that is close to the edge and public clouds for performance, cost and compliance reasons. Hosting an AI stack at an Equinix colocation data center provides the following benefits:

- Global presence and consistent service across 55 markets in 26 countries. This helps to satisfy enterprise requirements to deploy in multiple locations in order to comply with government regulations for data residency.
- Global high-speed and secure interconnection fabric connecting these data centers so that you can securely do AI processing across multiple regions.
- Proximity to the edge in most metros (less than 10 ms from the end devices) allows for AI inferencing and training at the metro edge. This, in turn, provides cost and latency benefits.
- High-speed and low-latency connectivity to the public cloud (between 1 and 2 ms to most major clouds in most markets). This architecture allows customers to have hybrid multicloud AI architectures.

- Ecosystem of 9,800+ companies consisting of clouds, network providers, financial companies, media companies and other enterprises at Equinix. Thus, this AI service allows companies to easily perform AI on their own data at Equinix, and get high-speed and secure access to data from other partners. In addition, new companies that join this ecosystem at Equinix can leverage the benefits of AI as a Service.
- Retail colocation data centers that can support from small (10 kW) to larger AI infrastructure deployments (~300 kW).

Use Cases for This Solution

AI at Equinix is specifically designed for use cases where:

- Enterprises want to do AI training or inference in their private data centers instead of in public clouds for control, cost, performance and privacy reasons. But many are finding it difficult to host AI in private data centers due to their inability to handle high power-density requirements, and the complexity of managing AI hardware and software infrastructure. Thus, enterprises want to access AI as a Service at a colocation data center.
- Datasets need to get processed at the edge instead of hauling the data back to a remote core data center, for cost, latency, and privacy and compliance reasons.
- AI applications need to integrate with enterprise IT systems or need to access external data from multiple sources such as clouds, private data centers, data brokers and edge locations. In these hybrid multicloud use cases, it's best to host the AI infrastructure at an interconnection hub where businesses can integrate with distributed IT systems via high-speed and secure networks.
- More than 9,800 enterprises and providers already have their infrastructure at Equinix and are interconnected to each other. We want to make it easy for both existing customers and new customers to do AI processing while leveraging data from this ecosystem.

More Information

Equinix, along with its partners NVIDIA, NetApp and Core Scientific, is piloting the AI as a Service Test Drive at Equinix. By using Test Drive for both training and inferencing, you can assess the capabilities and simplicity of this joint AI as a Service solution and its benefits. After using Test Drive, you can deploy this AI platform at Equinix.

About Equinix

Equinix, Inc. (Nasdaq: EQIX) connects the world's leading businesses to their customers, employees and partners inside the most-interconnected data centers. On this global platform for digital business, companies come together across more than 50 markets on five continents to reach everywhere, interconnect everyone and integrate everything they need to create their digital futures.

Equinix.com

If you are interested in trying out Test Drive at Equinix, you can find more information at:

Equinix.com/LandingPage url://www.equinix.com/AI/testDrive

The global interconnection platform for a cloud-first world

Globally deploy your infrastructure and services wherever opportunity leads. Directly and privately interconnect to your most important clouds, services and networks. Activate edge services on-demand to scale for success. On Platform Equinix®, you'll reach everywhere, interconnect everyone and integrate everything you need to create your best future. Get digital ready with Equinix.

NVIDIA & NetApp & CoreScientific-Partner Solution Brief_d1_US-EN | 10881 | v043020 | 0420 | Q419