ESG SHOWCASE

Supporting Federal Agencies with AI Infrastructure-as-a-Service from ViON and NVIDIA

Date: September 2021 **Author:** Mike Leone, Senior Analyst

ABSTRACT: As the use of AI permeates throughout federal agencies, the struggle to support the diversity of AI use cases and their workload requirements is forcing infrastructure transformation. To overcome operational burdens, performance challenges, and rising costs from force-fitting the use of legacy and/or commodity hardware to satisfy diverse AI workload requirements, agencies are looking for help in consolidating and modernizing their infrastructures to deliver a next-generation platform that accelerates their AI journey. Agencies are looking for solutions that deliver the simplicity, elasticity, and resource availability of the public cloud, but with the deterministic performance of a powerful, on-premises AI platform made available through an operationally efficient, as-a-Service delivery model.

Federal AI Infrastructure Challenges

As federal agencies look to transform through the accelerated adoption of AI, infrastructure change is becoming inevitable. Existing legacy infrastructure is unable to satisfy the unique demands of AI development. Components like traditional CPUs and even commodity GPUs are not optimized to deliver the performance required for a diverse set of AI workloads. ESG research shows that nearly 1 in 3 organizations state that one of their greatest barriers to AI success is the need for better IT infrastructure capabilities.[1] For some agencies, assembling an AI-centric infrastructure has led to wasted capital expenses, as IT leaders new to AI think they can scale complex AI workloads the same way as they do their traditional workloads by buying more servers packed with PCIe-connected GPUs and scaling out the same storage that supports their mainstream workloads. They build the infrastructure for a specific use case or AI workload, and it leads to overburdened IT, delays from moving massive amounts of data to separate environments, unique onboarding experiences depending on the environment, and wasted capital through duplicate deployment and management of infrastructure and tools.

A great example of this can be found in the Department of Defense, where they had several disparate AI environments consisting of different infrastructure components, tools, and data across the agency. These AI silos created a barrier for the Department to effectively use the right infrastructure to support their hundreds of unique AI use cases. This was all discovered after the formation of the Joint Artificial Intelligence Center (JAIC) evaluated the existing, department-wide AI infrastructure that needed consolidation to help lower the barrier for AI developers to have access to the right tools across a common infrastructure.

Transforming IT to Enable AI Success

As agencies search for help in assembling a modern and complete AI infrastructure stack to support their diverse AI workloads, the status quo leads agencies down the path of utilizing the public cloud due to promises of fast onramp, availability of tools/services/resources, and limitless scalability. The as-a-Service model is appealing to overburdened IT

[1] Source: ESG Master Survey Results, *Artificial Intelligence and Machine Learning: Gauging the Value of Infrastructure*, March 2019.

staffs that can now deliver controlled environments that enable AI developers to effectively experiment and learn about the best ways to leverage data to support an AI use case. But as model complexity grows, the need for increased compute power and speed and more storage forces a cost tradeoff that would sacrifice time or improved accuracy due to rising costs. On top of that, data gravity is forcing agencies to repatriate their AI workloads to on-premises environments to offset the increase in time and money it takes to push large data sets to where the compute resides. Between development roadblocks, escalating costs, and data gravity, agencies are hitting an inflection point: they need the best of both public and private cloud worlds. They need a hybrid approach to AI infrastructure.

Agencies need dense computational power to support the massively parallel architecture of neural networks and high-performance storage with ultra-low latency networking to keep compute clusters fed with the right data. They realize there is a benefit to an on-premises platform that gives developers the deterministic performance they need at a reasonable cost and still enables the rapid iteration of AI development at scale. Agencies are looking for a new IT infrastructure standard through a viable AI platform that is delivered in an as-a-Service model that enables embracing data gravity while supporting both temporal needs as AI development gets underway and the diversity of AI workloads across all use cases.

Accelerate AI Adoption with ViON and NVIDIA

ViON has helped federal agencies for years acquire new technology to support their existing and increasingly next-generation workloads through a flexible and affordable as-a-Service operating model. Through key partnerships with infrastructure and technology providers, ViON delivers on-premises infrastructure that can be selected, configured, and utilized based on customer workload requirements and demands. Not only does this help agencies accelerate their IT modernization efforts through the delivery of cloud-like infrastructure, but ViON's as-a-Service model introduces flexibility in how infrastructure resources are delivered, managed, and expensed. Agencies treat the infrastructure as an operational expense (versus traditional capital expenditure) and the pay-as-you-go model ensures budgets remain in check. The result is a consolidated data center using modern technology that accelerates deployment times of right-sized resources to support diverse data-centric workloads.

How It Works

Infrastructure is deployed by ViON at a data center or co-location facility based on customer requirements. Using the OpEx financial model, ViON acquires and retains ownership of the infrastructure, while customers (if they choose) operate the infrastructure and own the processes for configuration, control, and management. Once deployed, IT can procure and manage IT infrastructure delivered as-a-Service. This gives IT greater control and oversight into requirements, procurement, workflows, contracts, and key performance indicators to best support end-users' requirements based on a workload's resource demand. And together with a pay-per-use model, end-users pay only for the capacity allocated.

AI Infrastructure-as-a-Service with NVIDIA

ViON's as-a-Service offering extends to AI infrastructure. In partnership with AI technology provider NVIDIA, agencies gain the simplicity and resource elasticity of the cloud with the deterministic performance of a robust AI platform. Understanding that agencies want effortless access to resources that can be spun up quickly, ViON's AI Infrastructure-as-a-Service is lowering the barrier to AI by enabling AI developers to get started quickly with access to the right resources during the development phase of AI through productive experimentation. As experimentation leads to more complex training and production, AI Infrastructure-as-a-Service delivers the right amount of optimized computational power in combination with high-performance storage to efficiently feed data sets. And as AI models increase in complexity, ViON and NVIDIA ensure linearly predictable performance at scale.

NVIDIA DGX

The foundational infrastructure of ViON's AI Infrastructure-as-a-Service is the NVIDIA DGX™ A100 system featuring eight NVIDIA GPUs and two second generation AMD EPYC™ processors. Optimized to support the end-to-end AI lifecycle, the DGX is delivered as a universal building block for the AI data center. Through embedded technologies, DGX allows IT to right-size computational power for the workload at hand. In other words, agencies can consolidate their disparate AI environments into a single platform offering homogenous infrastructure that supports heterogeneous workloads. Whether the workload is heavy on analytics for data being engineered and prepared to feed a model training run, taking a viable prototype and training it at scale, or taking a fully tuned model to production for inference, the same system and infrastructure can easily pivot to support three different and common AI workloads over time by dynamically adjusting the resource and computational horsepower based on demand. As workloads get increasingly more complex, the DGX system offers multi-nodes scaling so data scientists can iterate faster on training runs and improve their accuracy using growing data sets. NVIDIA DGX brings together an optimal balance of compute, storage, and networking to enable the fastest time to solution on the most complex AI workloads.

The Bigger Truth

With AI adoption permeating throughout the public sector, IT staffs across federal agencies recognize the barriers to faster AI adoption. Legacy infrastructure cannot effectively deliver performance at scale. The public cloud, while great for onramp and resource availability through an as-a-Service delivery model, can quickly create scale and cost challenges. And data gravity is all but forcing IT to consider on-premises AI deployments to complement their clouds.

ViON with NVIDIA recognize these AI challenges and together are rising to the occasion by delivering an AI Infrastructure-as-a-Service solution that can meet lofty agency IT demands and AI workload requirements. Agencies gain a cloud-like infrastructure on-premises that delivers right-sized resources optimized for all AI workloads. Resources are delivered to AI developers as-a-Service, minimizing resource waste and maximizing resource utilization. Wrapped with a pay-as-you-go cost model, budgets remain in check. And for IT, they gain unprecedented peace of mind knowing ViON and NVIDIA experts ensure infrastructure flexibility and agility across the AI lifecycle from development and experimentation to training and inference.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides market intelligence and actionable insight to the global IT community.

www.esg-global.com contact@esg-global.com 508.482.0188