AI-Ready Infrastructure: A large organization selected hyperscale AIRI™ (AI-Ready Infrastructure) — jointly developed by Pure Storage® and NVIDIA® — to accelerate and simplify the processing of massive volumes of high-resolution images and video to dramatically improve the productivity of data scientists from disparate departments who now have simultaneous access to a blindingly fast shared pool of storage.

BUSINESS TRANSFORMATION

IT staff spends far less time configuring and managing its hyperscale AIRI infrastructure and more time pursuing high-value strategic objectives. Data scientists are more productive at any scale with AIRI Software stack and focus on rapidly building and deploying models.

SOLUTION BENEFITS

- Hyperscale AIRI combines the best-in-class options for compute, storage and networking — multi-racks of DGX-2 servers and FlashBlades interconnected with Mellanox InfiniBand or Ethernet fabrics.
- The on-premise hyperscale AIRI solution is far more cost-effective than cloud-based approaches.
- Hyperscale AIRI is a proven reference architecture designed in partnership with the leaders of AI supercomputing systems — NVIDIA and Mellanox — and validated by a growing number of customers.

APPLYING ARTIFICIAL INTELLIGENCE TO OBJECT IDENTIFICATION AND CLASSIFICATION

A large organization involved with security is tasked with analyzing increasingly large volumes of high-resolution still images and videos. It is now applying AI to these critically important workloads in order to increase the volume of images that can be analyzed, reduce the time required to perform analysis, and improve accuracy in mission critical applications.

This customer faces a number of challenges to successfully apply AI to object identification and classification, including:

- *Massive volumes of data* — Each super-high-resolution image or video is hundreds of megabytes in size, meaning a single data set for analysis can routinely run into the petabytes.
- *Greater accuracy requires ever more data* — Even a small improvement in training accuracy can require a 10x increase in data-set size.
- *A complex AI pipeline* — The process of preparing data of this volume and complexity for analysis involves many steps, and a bottleneck at any one of them can seriously impede the flow of data required for timely analysis.
- *A multi-tenant environment* — The customer wanted its image-processing capabilities to be available to many different users from cross-disciplinary teams across the organization.
- *Keeping data teams productive at scale* — Like most organizations pursuing AI or machine-learning initiatives, this organization's most valuable resource are its data scientists. Keeping them fully engaged and productive by eliminating idle cycles is a top priority, so anything that holds up the continuous flow of data for analysis is a roadblock that must be removed.
- *Building a high-performance, manageable IT infrastructure at reasonable cost* — Various "point" solutions that solve particular problems exist in the abstract, claiming to handle the compute, network or storage requirements of AI. But the reality of delivering a proven infrastructure — one that is reliable, scalable, cost-effective, easy to manage, and keeps data scientists productive — must be validated with reference architectures that are fully operational in real-world production environments.

CHALLENGES:

- A single data set can routinely run into the petabytes. Even a small improvement in training accuracy can require a 10x increase in data-set size.
- Preparing data involves many steps, and a bottleneck at any one of them can seriously impede the flow of data required for timely analysis.
- Image-processing capabilities must be available to many different users from cross-disciplinary teams across the organization.
- Keep data scientists fully engaged and productive by eliminating idle cycles.
- The organization needed a high-performance, manageable IT infrastructure at reasonable cost.

IT TRANSFORMATION:

- Analytic tasks that used to require 8, 12 or even 16 hours can now be accomplished in minutes.
- The solution radically simplifies installation and accelerates time-to-deployment.
- The costs of third-party design, integration, and support services are eliminated, and the small footprint and reduced power requirements slash operating costs.
- Data scientists are more productive and the IT staff spends far less time configuring and managing its IT infrastructure and more time pursuing high-value strategic objectives.

This customer faced all these challenges, and more. It had five key criteria for a solution: Performance, scalability, manageability, cost-effectiveness and resource optimization. All five were met by AIRI™, the AI-Ready Infrastructure from Pure Storage and NVIDIA.

THE AIRI SOLUTION FROM PURE STORAGE AND NVIDIA

AIRI is a converged infrastructure stack that is purpose-built for demanding large-scale learning environments such as artificial intelligence, machine learning, genomics and predictive healthcare. An AIRI solution has two key elements: the NVIDIA DGX™ servers and the Pure Storage FlashBlade™ data hub.

AIRI's flexible, hyperscale architecture enables enterprises to add DGX servers or storage blades independently based on the unique requirements of an organization. The stack is configured and tested as a complete end-to-end solution, eliminating the need for customers to perform their own configuration and tuning.

The hyperscale AIRI solution deployed by this customer consists of six DGX-2 servers, providing a total processing capacity of 12 petaflops; and two FlashBlades with multi-chassis, and storage capacity scalable to petabytes with zero downtime. The DGX servers and FlashBlades are scaled-out across three racks and interconnected with a Mellanox Infiniband™ network.

This hyperscale AIRI configuration solves all five of the customer's key requirements:

PERFORMANCE

To effectively employ the 12 petaflops processing power of the NVIDIA DGX-2 servers — without consuming multiple racks of space in a data center — storage must be exceptionally fast to feed data to the NVIDIA GPUs. The scale-out FlashBlade data hub can deliver data at rates up to 75Gb/s bandwidth and up to 7.5 million IOPS. Processing power on this scale means that analytic tasks that used to require 8, 12 or even 16 hours can now be accomplished in minutes.

SCALABILITY

Hyperscale AIRI is based on a highly extensible architecture that offers almost limitless expansion possibilities. NVIDIA DGX-2 servers and FlashBlade blades can be added at any time and in any combination, to meet the growing needs of applications and users. Hyperscale AIRI is architected to scale-out to 64 racks of DGXs and FlashBlades interconnected with Mellanox Fabrics.

MANAGEABILITY

While performance is the number-one priority for the customer, ease of deployment and administration is a close second. They needed a solution that could start on a small scale, then grow without disruption as the scale and complexity of their workloads grow.

Hyperscale AIRI meets this requirement in several ways. First, hyperscale AIRI is a complete solution that comes fully configured out of the box, based on a reference architecture proven in numerous deployments with government, academic and commercial customers. Customers do not need to be their own system integrator. This radically simplifies installation and accelerates time-to-deployment. And ongoing maintenance is dramatically simplified through the software included, intuitive management tools, and non-disruptive software upgrades and capacity expansions.

COST-EFFECTIVENESS

Hyperscale AIRI saves on both capital and operating costs in several ways. The DGX-2 servers and FlashBlade systems deliver performance levels that alternative, unproven, do-it-yourself solutions could approach only at much higher cost and significantly greater management complexity.

Since ordering, maintenance, support and expansion are handled through a single vendor, the costs of third-party design, integration and support services are eliminated. The small footprint and reduced power requirements of the AIRI system slash data-center operating costs and/or co-location fees. In addition, the need for specialized IT staff to manage historically complex storage environments is sharply reduced or even eliminated.

RESOURCE OPTIMIZATION

Data scientists are more productive in an AIRI environment, for three reasons.

First, the AIRI software stack, with pre-optimized and dockerized AI frameworks, means data scientists do not have to spend weeks compiling and tweaking all the dependent libraries, and instead can focus on building models.

Second, the FlashBlade platform provides a shared pool of storage that multiple analysts can use simultaneously. Data scientists do not have to wait for their turn at accessing data or concern themselves with the tedious and time-consuming process of preparing data for analysis.

Third, the high performance of the DGX-2 servers and FlashBlade means more work gets done in less time. More data means more accurate training; more accurate training means less time correcting errors; less time correcting errors improves productivity and accelerates time-to-insight. In addition to improved productivity for data scientists, the customer's IT staff spends far less time configuring and managing its IT infrastructure and more time pursuing high-value strategic objectives.

HYPERSCALE AIRI: DELIVERING SUPERCOMPUTING CAPABILITIES TO ENTERPRISES PIONEERING REAL-WORLD AI

Artificial intelligence is still in its early stages as a mainstream technology, so enterprises looking to tap the power of AI confront numerous alternatives for compute, storage and systems management. Hyperscale AIRI is the superior option for several reasons:

- Hyperscale AIRI combines the best-in-class options for compute, storage and networking. The hyperscale AIRI is architected by Pure Storage in partnership with the leaders of AI and supercomputing infrastructure NVIDIA and Mellanox.
- An on-premise hyperscale AIRI solution is far more cost-effective than cloud-based approaches for enterprises supercharging AI initiatives. The volumes of data involved in their real-world AI applications make it prohibitively expensive to move data to and from a cloud service. Plus, keeping data on-premise means analysts have near-instant access to it.
- Hyperscale AIRI is a proven reference architecture designed to offer a modular, streamlined approach to building an AI infrastructure that can scale to 64 racks of DGX servers and FlashBlades interconnected with Mellanox High Performance Fabrics.

PURESTORAGE®

info@purestorage.com
www.purestorage.com/customers