PURESTORAGE®

CORE SCIENTIFIC

WHITE PAPER

Run AI Workloads on Premium Infrastructure

Tap into the power of AI-as-a-service by Core Scientific, built with hardware from Pure Storage®.

Contents

Introduction 3

AI Infrastructure Is Challenging 3

- Costs and Complexity 3
- Limitations of Most Cloud Providers 3
- Lack of Adequate Management Tools 4

Gain an AI Infrastructure Advantage with AIRI™ Infrastructure 4

- Highest Performance 5
- Reduced Complexity 5
- Lower Total Cost of Ownership (TCO) 6

AIRI Architecture for the AI Lifecycle 7

- The AI Data Lifecycle 7
- Saturate the GPU 8
- AIRI System Architecture 9
- Advantages of the FlashBlade Product in AIRI Infrastructure 9

Core Scientific and Pure Storage Make AI Accessible 10

Additional Resources 10

Introduction

Businesses across a wide range of industries understand that artificial intelligence (AI)-driven insights are crucial to staying competitive. At the same time, the difficulty of deploying the right infrastructure to drive those discoveries is thwarting many of those same organizations from even getting started in AI. The stakes are clearly high. A recent survey revealed that most C-suite executives believe that their businesses risk failure if they don't scale AI in the next five years.[1] Still, as many as 76% report they are struggling with how to achieve that goal.[1] Now, Core Scientific and Pure Storage have partnered to make AI-optimized infrastructure widely and quickly available.

AI Infrastructure Is Challenging

Behind the aim to scale AI is the need for top-level performance, which is essential not only to power better and faster insights, but also to attract the best AI talent. And building world-class AI infrastructure to support this objective begins with the top-performing, latest-generation graphics processing units (GPUs). However, the storage throughput of a system must be equally high performing to feed data to the GPUs as quickly as they can process it. Without extremely high throughput to saturate the fastest GPUs, a bottleneck will form that will constrain the system's performance. The fastest GPUs, together with all-flash storage, form the foundation of the world's best AI infrastructure.

Costs and Complexity

But even if businesses understand the components of cutting-edge AI infrastructure, acquiring and assembling that infrastructure is easier said than done. First, building world-class AI infrastructure in a do-it-yourself (DIY) fashion requires expertise beyond the capabilities of many organizations. Second, cost is a huge obstacle, because GPU-powered servers tend to be much more expensive than CPU-powered ones. And though some organizations might be able to splurge on the latest AI infrastructure, the GPU power advantage of the latest system lasts only until the next generation is released. And even if an organization is able to purchase a new system year after year to preserve an AI performance edge, embracing a capital-expenditure (CAPEX) model for AI infrastructure might be a wildly inefficient and cost-prohibitive way to stay ahead of the competition.

Limitations of Most Cloud Providers

To reduce costs as they pursue AI infrastructure, many organizations choose to embrace an operating-expense (OPEX) pricing model by purchasing AI services from a cloud provider. But this option comes with a catch—performance limitations:

- Cloud providers are typically a generation or two behind the latest GPUs in offering AI compute power.
- Cloud providers' fabrics typically rely on virtualization and are not built for bare-metal computing.

- Customers of cloud services typically are not offered the opportunity to store their enormous datasets on media with the highest industry throughput rates, such as Pure FlashBlade® products.
- Customers often cannot ensure that their cloud-based data will be co-located in the same rack—or data center—as the GPU-powered servers assigned to run AI algorithms on that data.

As a result of all these limitations, relatively high latencies are common in cloud-based AI, which rules out the performance edge that many organizations are seeking in the first place.

Lack of Adequate Management Tools

Another important reason why many AI projects fail is the lack of good management software to support the compute clusters used for AI. Data scientists, understandably, do not often have a DevOps background, and many are accustomed to the "walk-up" services available through cloud providers that help reduce friction in setting up desired computations (such as training machine learning [ML] models on datasets). Without adequate software, users might end up simply staring at a terminal for a million-dollar AI infrastructure, unsure how to start. To help ensure success, businesses need software that makes it easy for users to configure the infrastructure to run AI workloads, with options available for key features such as bursting from the data center to a public cloud when extra compute is needed.

Gain an AI Infrastructure Advantage with AIRI® Infrastructure

Pure Storage and Core Scientific, a leading managed service provider (MSP) for AI, have partnered to solve these challenges in an AI-based offering, Cloud for Data Scientists™. Offered through Core Scientific, Cloud for Data Scientists delivers the industry's first data center services for cutting-edge AI. It is a managed solution that combines Core Scientific's expertise with Pure Storage AI-ready infrastructure (AIRI), which itself is built from GPU-powered NVIDIA DGX servers and ultrafast FlashBlade storage. This AI infrastructure offering is available as an optimized AI-as-a-service solution based in partner data centers, as a co-location option, or as a managed solution on customer premises. With each option, Cloud for Data Scientists helps optimize performance by ensuring that customer data is always co-located with compute. Finally, Cloud for Data Scientists comes bundled with Core Scientific® Plexus™ software tooling, which makes it easy for data scientists to get started with running their GPU-powered workloads.

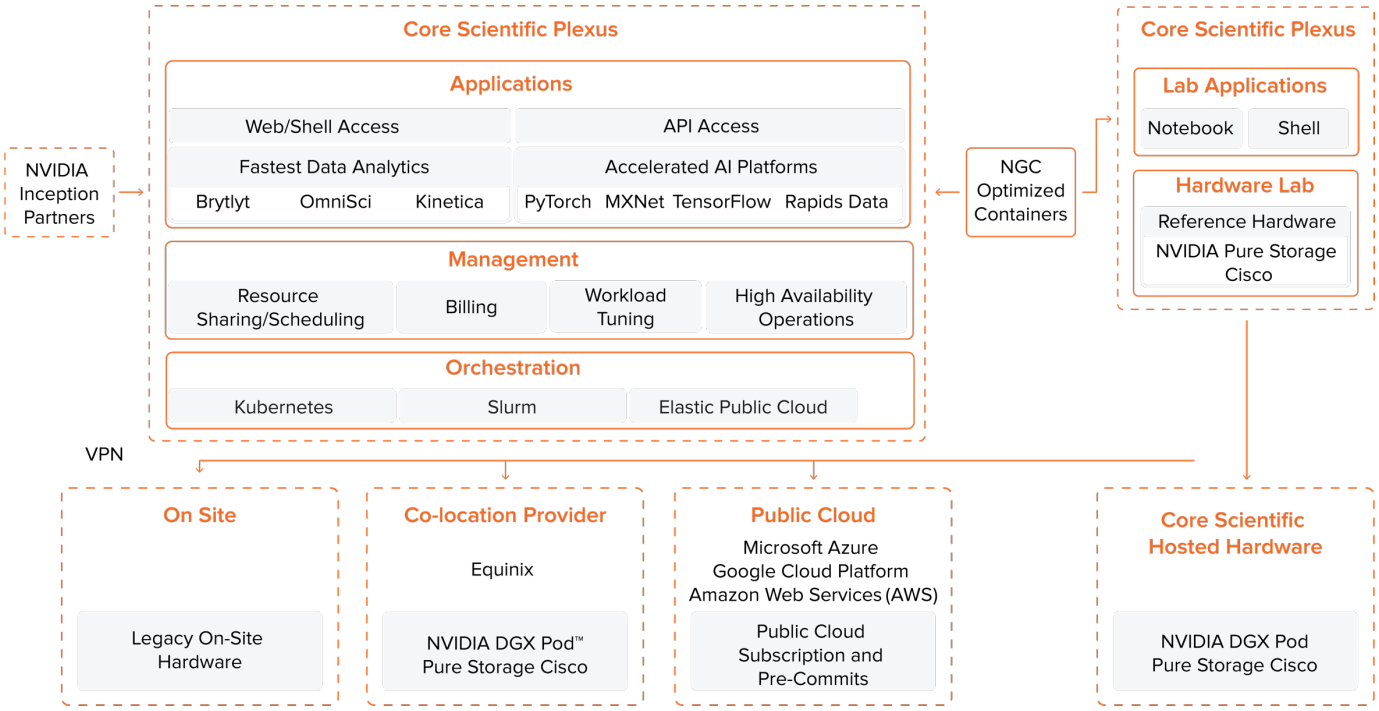


Figure 1. Core Scientific offers a managed AI infrastructure solution accelerated by Pure Storage.

Highest Performance

As the primary hardware component of the solution, the AIRI infrastructure with the FlashBlade® product offers extremely high throughput, fast enough to saturate the most powerful GPUs and get the most performance out of the NVIDIA DGX servers. The Cloud for Data Scientists service also ensures that hardware is kept updated on an ongoing basis, so its AI infrastructure can always deliver the best available GPU hardware and software, far ahead of the public cloud cycle. The service also ensures high performance for cloud-based AI through data locality, offering co-location of data and compute at the data center level or even the rack level.

Reduced Complexity

Cloud for Data Scientists, as a managed service, completely eliminates the complexity of setting up and maintaining AI infrastructure. Storage is complex, but Pure's infrastructure epitomizes simplicity. Simplicity is a foundational attribute of Pure's product and support strategy. The Pure Evergreen Storage™ subscription model delivers a fundamentally different storage-consumption experience. MSPs receive a "subscription to innovation" that includes non-disruptive software and hardware upgrades, free controller upgrades every three years (with an eligible Evergreen™ subscription), and trade-in opportunities to make use of advances in flash technology.

Effortless support experience: The Pure1® platform is a cloud-data-management solution powered by Pure1 Meta™, an AI/ML engine that processes more than a trillion telemetry data points per day.[2] It provides 24x7 end-to-end visibility, and it enables predictive analytics and planning across your entire storage fleet.

Easy installation and maintenance: It's easy to install, manage, and scale Pure Storage arrays with nothing to configure or tune. This simplicity not only leads to higher operational efficiency, but also to faster customer onboarding and accelerated book-to-bill cycles. Another advantage is higher asset utilization: You do not have to procure assets well ahead of anticipated demand to allow for lengthy site preparation, installation, setup, configuration, and tuning. These higher utilization rates can help substantially reduce cost structure.

Automation: Pure follows an API-first approach, and Pure arrays also seamlessly integrate with several platforms and ecosystems, including VMware (such as VMware Cloud Director), containers, Microsoft solutions, and OpenStack. This means software development, testing, environment maintenance, migration, and expansion become easy to manage and automate.

Core Scientific contributes ongoing expert guidance in helping customers meet their infrastructure needs as they evolve, in addition to ensuring that the technology stack and data models are tuned for peak performance.

The AIRI infrastructure–based solution is also bundled with Core Scientific Plexus management software, which offers the following features and benefits:

- A single pane of glass to manage all AI and deep learning (DL) projects with walk-up tooling
- Seamless access to a GPU-accelerated software app portal, including Kinetica, SQream DB, Deepgram, and other NVIDIA Inception partners
- Ease-of-use through intelligent orchestration, management, and scheduling of data-science workloads
- A simple option to burst to the cloud in hybrid and CPU-bound scenarios, in addition to shared co-located clusters
- High-availability cluster management

Lower Total Cost of Ownership (TCO)

The Core Scientific and Pure Storage solution helps reduce the cost of acquiring AI infrastructure for customers.

Core Scientific, through Cloud for Data Scientists, helps lower the TCO of AI infrastructure by making an OPEX model available specifically for AI workloads, whether co-located in regional data centers or on customer premises. Paying for access to AI-optimized hardware only as needed—with no ingress or egress fees—spares businesses from having to invest in expensive hardware through CAPEX. The service also allows customers to manage these costs while avoiding being locked into old hardware, in addition to avoiding the price and performance inefficiencies of the public cloud.

Pure Storage helps reduce TCO with the Evergreen Storage business model and industry leading data reduction. The Evergreen Storage model from Pure addresses the expensive upgrade cycles, disruptive downtime, and rebuys of terabytes that are typical of legacy storage providers. The Evergreen Storage subscription model offers seamless, rapid upgrades and expansion, without disruption. These upgrades can result in dramatic feature improvements for no extra cost, such as capacity expansion, increased throughput rates, lower power consumption, and increased density and consolidation. Pure Storage also helps reduce costs by enabling smaller footprints through five data-reduction

technologies: data deduplication, compression, pattern removal, deep reduction, and copy reduction. These are always-on, global, and designed for mixed workloads. The data reduction from Pure is highly granular, employing variable 512-byte addressing and multiple data-reduction and compression technologies. The combination of these factors, together with thin provisioning, results in data-reduction levels of up to 10:1.[3]

AIRI Architecture for the AI Lifecycle

The AI Data Lifecycle

Data is the heart of modern AI and DL algorithms. Before training can begin, the first problem is collecting the labeled data that is crucial for training an accurate AI model. A full-scale AI deployment must continuously collect, clean, transform, label, and store large amounts of data. Adding additional high-quality data points directly translates to more accurate models and better insights.

Data samples typically undergo a series of processing steps:

1. Ingest the data from an external source into the training system.
2. Clean and transform the data, and save it in a format convenient for training.
3. Explore parameters and models, quickly test with a smaller dataset, and iterate to converge on the most promising models to push into the production cluster.
4. Train by feeding random batches of input data into production GPU servers for computation to update model parameters.

It is important to have fast central storage nearby during these processing steps to maximize data throughput and saturate the GPU cycles.

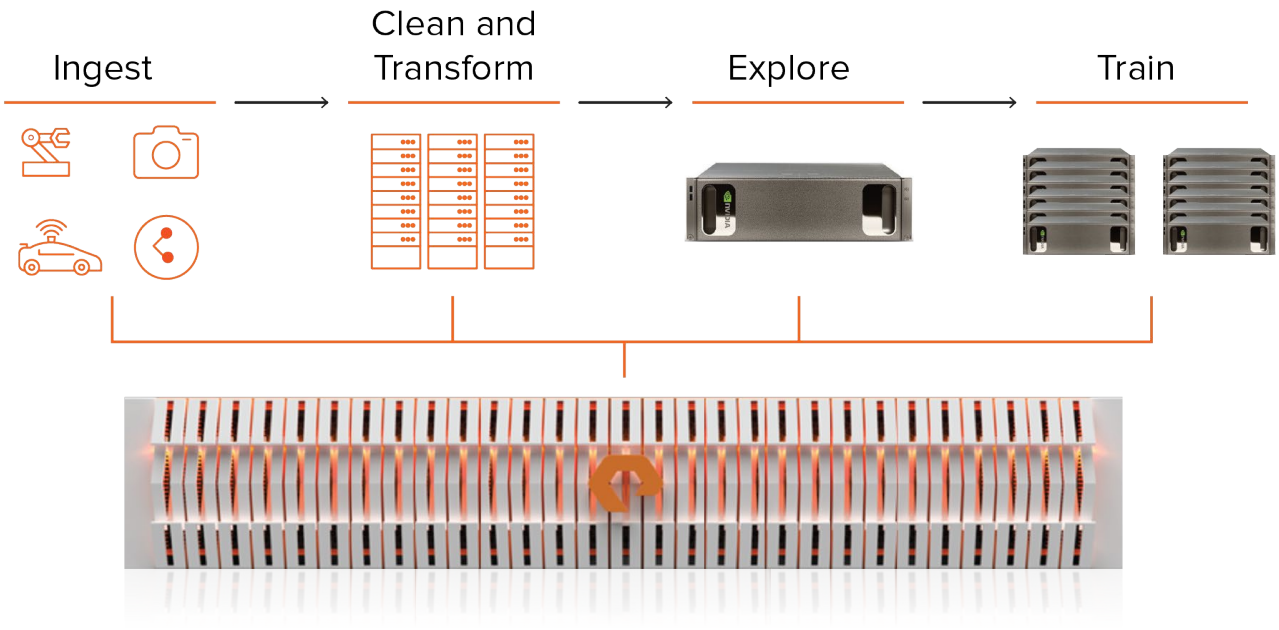


Figure 2. The FlashBlade product serves as fast central storage through all phases of AI data handling.

Saturate the GPU

Training software needs to fill the data pipeline from the storage system to the GPUs to ensure that the GPUs always have the next batch of training data available. For an end-to-end model training session using AIRI infrastructure, the data flows as follows:

- **Decode and augment:** Input files from the FlashBlade product are loaded and converted to a form appropriate for training. The decode step can be pre-computed or performed on the host CPU. Augmentation uses the host CPU to introduce dynamic changes in the input.
- **I/O queues:** Multiple threads read random batches of input records from storage and populate an internal queue. These threads run on the NVIDIA DGX-2 host CPUs and are responsible for pre-fetching batches and ensuring training records are available in DRAM.
- **Training:** A second set of threads pulls data from internal queues to feed to the GPUs for computations needed for training and updating model parameters.

Each training batch requires retrieving the data from persistent storage, then decoding and augmenting that data. If the GPUs are idle—waiting for I/O and the host CPU to fetch and decode the inputs—then the compute power of the GPUs will not be efficiently utilized, as shown in Figure 3.

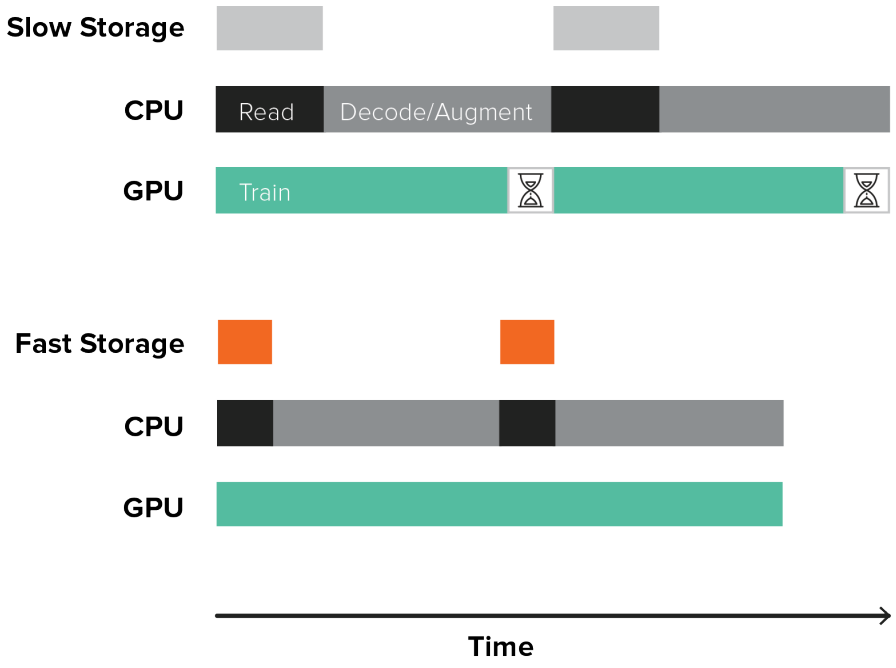
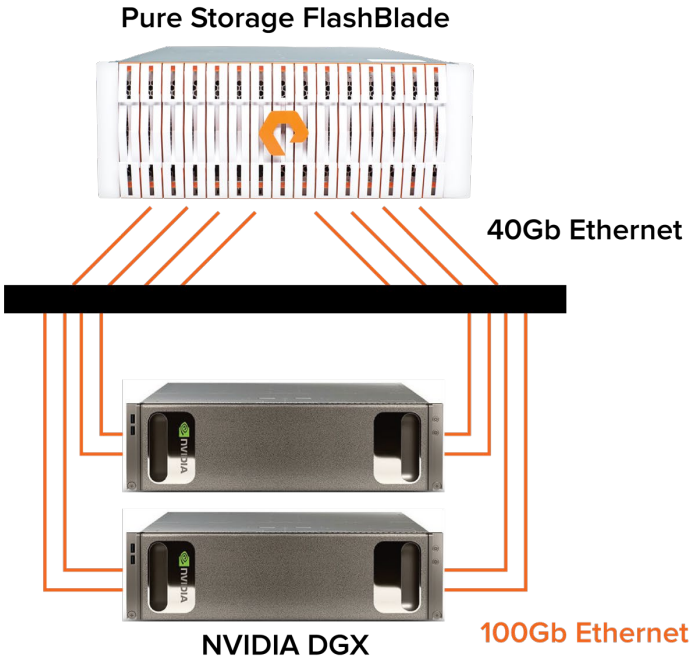


Figure 3. Fast storage helps keep the GPU pipeline saturated.

The pipeline training flow in AIRI infrastructure helps ensure that the training devices always have the next input set available upon completion of a training batch. As long as the I/O and host CPU portion is faster than the training computations, then GPUs will operate at maximum utilization.

AIRI System Architecture

The AIRI architecture is depicted in Figure 4, along with a description of core components of the system.



FlashBlade: 15x17TB blades:
179TB usable, before data reduction
8x 40 Gb/s uplinks

2x NVIDIA DGX-2, each with:
16x NVIDIA Tesla V100 GPUs with 512GB memory
2x Intel Xeon Platinum 8168 processor at 2.7GHz
1.5TB DDR4 system memory

Figure 4. AIRI system architecture.

The AIRI system is designed for scale-out DL workloads, and it is not restricted to this ratio of storage to compute. As datasets and workload requirements scale, additional NVIDIA DGX-2 servers can be provisioned and can instantly access all available data. Similarly, as storage capacity or performance demands grow, additional blades can be added to the FlashBlade system with zero downtime or reconfiguration.

Advantages of the FlashBlade Product in AIRI Infrastructure

The centralized data hub in the AIRI DL architecture can help increase the productivity of data scientists and makes scaling and operating simpler. FlashBlade products specifically make building, operating, and growing an AI systems easier for the following reasons:

Performance: With more than 15GB/s of large-file read bandwidth per chassis, and up to 75GB/s total, FlashBlade products can support the concurrent requirements of an end-to-end AI workflow. Overall, FlashBlade products can deliver a 10x to 20x improvement in analytics performance compared to other infrastructures.[4]

Small-file handling: The ability to randomly read small files (50KB) at 10GB/s from a single FlashBlade chassis (50GB/s with 75 blades) means that no extra effort is required to aggregate individual data points to make larger, storage-friendly files.

Native object support: Input data can be stored as either files or objects.

Non-disruptive upgrade: Software upgrades and hardware expansion can happen anytime, even during production model training.

Built for the future: Purpose-built for flash to take advantage of new generations of NAND technology for density, cost, and speed.

Core Scientific and Pure Storage Make AI Accessible

Core Scientific and Pure Storage have partnered to offer the industry's first AI-as-a-service, based on AIRI hardware, for companies looking to quickly and cost-effectively tap the power of AI with extremely fast infrastructure. Core Scientific's Cloud for Data Scientists combines the power of NVIDIA DGX systems and Pure FlashBlade storage to deliver AI on-demand that is easier, faster, and less expensive. Core Scientific software and expertise, combined with Pure Storage technology, can result in lower TCO and zero storage maintenance for end customers, so they can focus on their business outcomes and not have to worry about AI infrastructure.

Additional Resources

Equipping and enabling Data Scientists to take on the world's most advanced AI challenges

AIRI: Finally, AI at Scale

[1] "AI: Built to Scale," Accenture, November 2019.
[2] "AI-Driven Management, Full-Stack Analytics, and Predictive Support," Pure Storage, 2018.
[3] "Confused about Efficiency Claims? Apples to Apples Just Got Easier," Pure Storage.
[4] "Man AHL Accelerates Time-to-market with Pure Storage Data Hub," Pure Storage, 2018.

Pure Storage, Inc.
650 Castro Street, #400
Mountain View, CA 94041