PURESTORAGE®
Uncomplicate Data Storage, Forever

BUSINESS WHITE PAPER

The Pure Storage Platform for AI

Pure Storage® accelerates and simplifies AI deployments, enhancing their value to the enterprise.

With readily available generative artificial intelligence (GenAI), AI has become a sine qua non of information technology (IT) operations. Enterprises in finance, medicine, manufacturing, transportation, security, and others all realize that AI is now a survival issue for them. Those that use AI to identify trends, make accurate predictions, serve clients faster with less effort, and so forth have distinct competitive advantages over those that don't.

The increasing importance of AI-based solutions makes reliable, easy-to-use IT services a must for production deployments. This brief surveys AI needs throughout the project lifecycle (primarily from a storage perspective) and shows how the Pure Storage® portfolio of storage system, data services, and workflow management products for Kubernetes promote efficiency both for AI and IT infrastructure teams as well as developers and MLOps engineers who design, implement, and run AI applications.

> *"The playing field is poised to become a lot more competitive, and businesses that don't deploy AI and data to help them innovate in everything they do will be at a disadvantage."*
>
> **PAUL DAUGHERTY**, *CHIEF TECHNOLOGY AND INNOVATION OFFICER, ACCENTURE QUOTED IN: HTTPS://WWW.SALESFORCE.COM/BLOG/AI-QUOTES/*

The AI Project Lifecycle

Organizations undertake AI projects to support mission objectives such as,

- More accurate medical diagnoses
- Acceleration of genomic research
- More predictable market fluctuations
- Bank card fraud detection
- Rapid identification of security threats
- etc.

Whatever their objectives, in-house AI projects tend to follow a trajectory similar to that shown in Figure 1, from conception, development, production, to evolution.

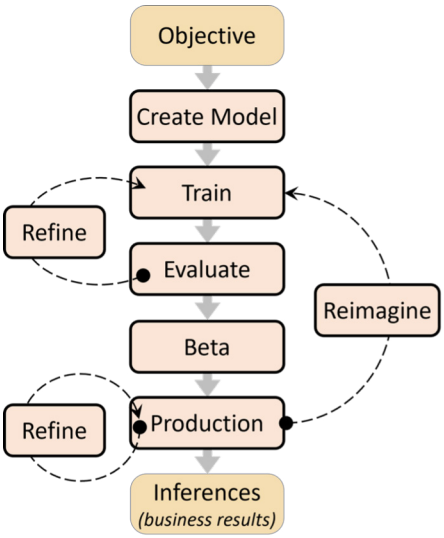

FIGURE 1 AI Project Life Cycle

AI projects generally start with a proposed model (algorithm) and train (refine) it iteratively using steadily increasing amounts of available or easily acquired input data for which outcomes are known until it reliably produces inferences (outcomes) that support the mission objective. For example, a medical diagnosis model might be trained using thousands of MRI scans with known diagnoses. The finished model would then take live scans as input and suggest diagnoses to medical practitioners.

Models are typically envisioned by small groups of data scientists, often in functional or business organizations rather than in IT teams. Data scientists start with modest IT resources (e.g., public cloud virtual machines and storage) to experiment with model variations.

> *As projects progress, the intellectual property embodied in models and the curated training data become increasingly valuable. Along with performance and reliability, protection against data loss and theft of intellectual property become important considerations.*

As model refinement progresses, development is typically taken over by Machine Learning Operations (MLOps) engineering organizations which:

- Assume on-going model training responsibilities from data scientists.
- Manage the transition of project resources and collateral to IT environments with the scalability, performance, and reliability needed for late-stage training and production.
- Introduce automated workflows that enable self-service training f or data scientists and other developers, and ultimately support robust production.

When models produce reliable inferences, they move into production. Production models use live data to produce inferences that assist with business decisions. As with any mission-critical IT application, both AI models and the environments in which they run must be stable. Reliable access to a model and the data it needs to function is key to stability.

Models in production are usually monitored for "data drift" to ensure that results continue to meet mission objectives. With GenAI, open source or proprietary *large language models* (LLMs) are often combined with *retrieval augmented generation* (RAG) using context-specific vector databases that are updated frequently to include new data and remove what has become less relevant. In some cases, AI models may be completely reimagined as business needs change and/or if new types of training data become available.

Information Technology in AI Projects

While AI projects typically begin using modest in-house or public cloud IT resources, most are destined for eventual production. Planning for production computing, storage, and software needs, and designing production workflows early in the development process minimizes mid-stream infrastructure and procedure changes and accelerates return on investment.

Planning for production is particularly important with storage. Ideally, an AI storage infrastructure should provide:

- Non-disruptive expansion to meet rapidly growing data and changing I/O needs.
- Seamless data sharing among developers, training jobs, and production.
- Decade-long "24 365" duty cycles.
- Security to protect intellectual property from intrusion and theft.

How AI Projects Utilize IT

From an information technology perspective, AI projects can be thought of as consisting of two general types of tasks:

Data Curation

The best models are trained using input data from multiple disparate sources—event records, documents, images, sensor readings, etc. Data curation is a blanket term for the acquisition, storage, and pre-processing of data for use in model training, and later in production.

Projects usually preserve raw input data to avoid the time and cost of recreating or re-acquiring it. In most cases, it must be curated (preprocessed) for model training. Data must be anonymized; timelines, measurement units, graphics resolutions, and so forth must be reconciled; and items must be transformed into the file or object formats that training and production tools require. Most projects preserve both raw and curated data.

What Data Curation Means for Storage Infrastructure

A large project might use petabytes of curated training data. Its storage infrastructure must be able to scale from a few hundred gigabytes of files to petabytes contained in billions of files and/or objects.

Raw data is used as input to curation but is then largely idle until a model is reimagined. Storage used for it should be low-cost, but highly scalable.

Data is typically curated by processing large batches of raw data sequentially, but items are accessed randomly during training and production. Flexible systems that perform well with both sequential, random, and mixed access I/O workloads are key.

Training and Production

The fundamental assumption of AI is that iterative training with large amounts of input data with known outcomes converges on a model that reliably makes useful inferences when presented with input[1] for which outcomes are unknown. In a medical imaging scenario for example, thousands of images with known diagnoses might be used to train a model that, when presented with new images, would propose diagnoses used to advise physicians. Subject matter-specific applications often combine application-specific data sets with readily available general-purpose *large language models* (LLMs) and use *retrieval-augmented generation* (RAG) to satisfy natural language queries in their subject matter area.

As training progresses, the amount of curated data and the intensity with which it is used increase rapidly. In addition, the data, files, and other collateral that comprise the evolving model become increasingly valuable.

When a completed model moves into production, its I/O needs change from the very intensive demands of a relatively small number of training jobs to the agility required to service thousands of concurrent client transactions making many unrelated I/O requests.

Finally, it is common to preserve production inputs and the corresponding output inferences for retraining and other techniques that adapt models to changing conditions.

What Training and Production Mean for Storage Infrastructure

Storage for curated data must be scalable, both in capacity and performance, and must be easily sharable by many concurrent training jobs. As the number of concurrent training jobs grows, automated scheduling and data sharing are musts.

The business value of a production model requires storage with production-grade reliability, performance, and administrative simplicity. The value of the intellectual capital embodied in it and its training and production data requires robust security for "data at rest."

In Summary: The Ideal Storage for an AI Infrastructure

Perhaps the most important attribute for AI project storage is *agility*—the ability to grow from a few hundred gigabytes to petabytes, to perform well with rapidly changing mixed workloads, to serve data to training and production clients simultaneously throughout a project's life, and to support the data models used by project tools. The attributes of an ideal AI storage solution are:

Performance Agility

- I/O performance that scales with capacity.
- Rapid manipulation of billions of items, e.g., for randomization during training.

Capacity Flexibility

- Wide range (100s of gigabytes to petabytes) with easy, non-disruptive expansion.
- High performance with billions of data items.
- Range of cost points optimized for active and seldom-accessed data.

Availability & Data Durability

- Continuous operation over decade-long project lifetimes.
- Protection of data against loss due to hardware, software, and operational faults.
- Non-disruptive hardware and software upgrade and replacement.
- Seamless data sharing by development, training, and production.

Space and Power Efficiency

- Low space and power requirements that free data center resources for power-hungry computation.

Data Models

- Support for block, file, and object data models and common network protocols.

Security

- Strong administrative authentication.
- "Data at rest" encryption.
- Protection against malware (especially ransomware) attacks.

OperationAL Simplicity

- Non-disruptive modernization for continuous long-term productivity.
- Support for AI projects' most-used interconnects and protocols.
- Autonomous configuration (e.g. device groups, data placement, protection, etc.).
- Self-tuning to adjust to rapidly changing mixed random/sequential I/O loads.

Appendix A lists potential pitfalls to be avoided when designing an AI storage infrastructure.

Pure's AI Product Portfolio

Storage Systems

The Pure Storage® portfolio of all-flash storage systems, illustrated in Figure 2, includes three FlashArray™ scale-up *Unified Block and File* (UBF) servers, a FlashBlade® scale-out *Unified Fast File and Object* (UFFO) server, and two *Unified Data Repository* (UDR) servers, one based on FlashArray and the other on FlashBlade, for low-cost large-scale storage. All systems support broad ranges of capacity that is easily expandable online. Each is optimized for specific capacity ranges, data type(s), and cost/performance targets. With these servers, Pure Storage can satisfy virtually any AI storage requirement from project conception through model training, and on into production.

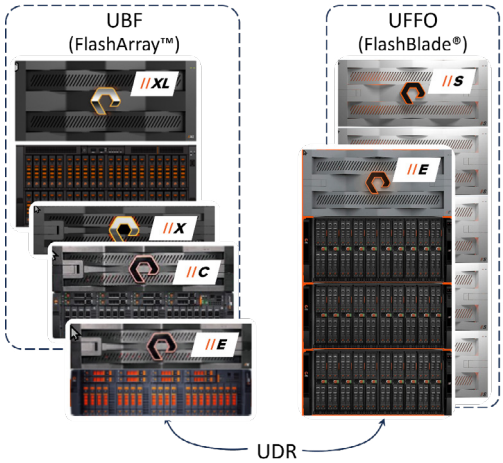

FIGURE 2 Pure's AI Storage System Portfolio

Features Common to All Systems

All Pure Storage systems share key properties:

Reliability and Availability

Systems are designed to continue operating when any internal component fails. For example, they survive at least two (in most cases more) rare *DirectFlash® Module*[2] (DFM) failures that overlap in time without loss of data or client access to it.

Efficiency

Systems optimize capacity utilization by *thin provisioning*—deferring space allocation until clients write data. They allocate space autonomously to balance utilization and load. DFMs' high density[3] minimizes system "footprint," power consumption, and ultimately, e-waste.

Simplicity

Systems minimize administrative tasks to the greatest extent possible—there are no "device groups" to manage, and no data placement or protection decisions to make. Systems report status and events to the Pure1® Cloud frequently; Pure1 analyzes behavior and proactively initiates any necessary service operations.

Continuity and Longevity

Systems are designed for lifetimes of a decade or more of continuous operation with no planned downtime or service outages, even during software and hardware upgrades and modernizations.

Evergreen®

Evergreen subscriptions that include regular software and hardware updates and periodic modernizations are perhaps Pure's most important benefit for AI projects of long duration. The company offers subscriptions both for purchased systems and for storage delivered as a managed service (Evergreen// One™). Technical briefs TB-230601f and TB-230601o, available at https://support.purestorage.com or from a Pure Storage representative, describe the company's Evergreen offerings in more detail.

Product-Specific Features

Spectrum of Performance and Cost Options

From latency-optimized FlashArray//XL™ for rapid response in production to throughput-optimized FlashBlade//S for training with very large data sets, to cost-optimized FlashArray//E™ and FlashBlade//E™ for less-active data (e.g., raw data, feature stores, etc.), Pure's product line offers cost/performance options that span AI project requirements.

Pure systems' very high-performing metadata operations on large numbers of files and objects make them particularly suitable for AI project training.

Capacity Flexibility

With maximum capacities ranging from FlashArray//C50's 1.6PB (effective[4]) to FlashBlade//S's nearly 20PB (physical), Pure's systems can support many in-house AI projects from concept through production with a single project-wide *data hub*. Systems can "start small" in a single chassis with minimal physical capacity and be expanded up to a model's maximum supported capacity without interrupting service to applications.

Where multiple systems *are* required, for capacity, performance, cost, or data model reasons, Pure1 centralizes storage management and supports AI-based "what if" capacity planning for users' entire "fleets" of Pure systems.

Data Reduction

Pure's systems optimize flash utilization by compressing data prior to storing it. In addition, FlashArray systems, achieve further efficiency by *deduplicating* blocks of data, replacing duplicate blocks with links to a single stored instance. Deduplication works well with *structured* data (e.g., databases, tables, etc.).

NVIDIA Collaboration

With approximately 80% market share,[5] NVIDIA Corporation is the acknowledged leader in AI computation. Since 2017, Pure Storage has collaborated with NVIDIA to develop solutions for AI. The collaboration has resulted in the jointly-developed *AIRI® Pure Storage NVIDIA DGX BasePOD Reference Architecture*[6] for AI, based on NVIDIA's DGX servers and network fabric coupled with Pure's FlashBlade//S™ storage systems. As an NVIDIA BasePOD certified reference architecture, AIRI eliminates the design, deployment, and management complexity inherent in custom-crafted AI infrastructures.

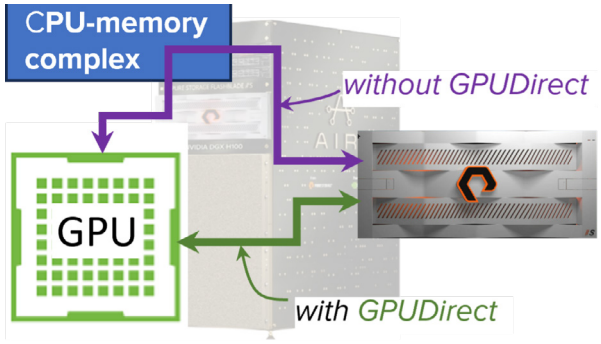

FIGURE 3 Pure-NVIDIA Collaboration

More recently, the Pure Storage-NVIDIA collaboration has resulted in:

- Pure's implementation of the NVIDIA GPUDirect[7] storage protocol to transfer data directly between FlashBlade//S storage and NVIDIA's GPUs, bypassing control CPU memory.
- Storage partner validation for FlashBlade//S in NVIDIA-certified OVX L40S reference architectures offered by major server vendors. When combined with FlashBlade//S storage, OVX-certified servers are complete AI platforms that accelerate small model training and fine-tuning, as well as GenAI RAG and production inference workloads.

Finally, when used in conjunction with Portworx®, NVIDIA's *device plugin for Kubernetes*[8] provides comprehensive management and scheduling of both GPU and storage resources at all AI project stages.

Portworx: Pure's Secret Weapon

Portworx by Pure Storage is the company's Kubernetes data services platform that provides persistent storage, data sharing and protection, workflow automation, and (optional) disaster recovery for containerized applications.

Portworx accelerates development of IT environments for the containerized applications used in most AI projects with a *software-defined storage* model that enables infrastructure-neutral access to data. Portworx supports any type of block storage, whether located on-premises or in a public or private cloud.

Portworx presents standardized virtual block or file storage devices to applications, regardless of the on-premises or cloud technology used to instantiate it. It does this by making architect-defined *storage classes* available to developers. Storage classes standardize storage properties, simplifying self-service job creation and promoting stable, reliable development and production environments. In addition, Portworx includes templates that assist with setup for applications like Apache Kafka, Zookeeper, Elasticsearch, as well as for popular databases including SQL Server, MongoDB, Postgres, and Cassandra, all of which are commonly used in AI projects. It provides consistent development environments while enabling self-service job creation by data scientists and other developers.

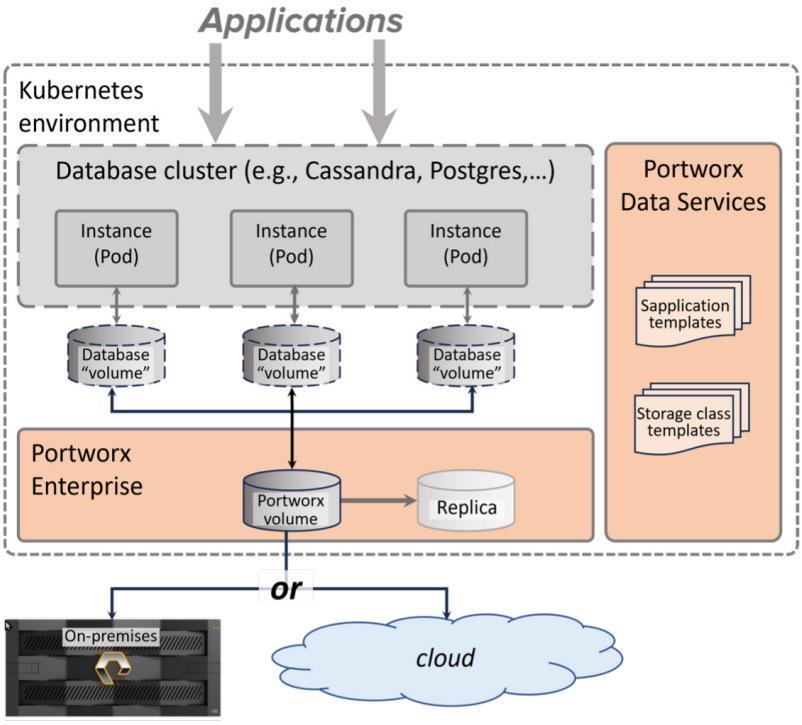

FIGURE 4 Using Portworx with a Database App

The software-defined storage model enables Portworx to share data among multiple Kubernetes pods running separate jobs. This makes it particularly useful for model training, where running many concurrent jobs that share the same input data is key to rapid implementation. As an example, Figure 4 shows how Portworx can simplify deployment of a training application that utilizes data from a database.

Finally, Portworx provides fault tolerance by replicating its virtual storage devices to either on-premises resources or to a public cloud. It can also protect entire project environments by copying them to S3 objects in a public or on-premises private cloud.

Table 1 suggests ways in which Portworx can be used to advantage in in-house AI projects.

	Portworx Feature	Application to AI
Portworx Data Sharing & Resource / Workflow Management	Up to 640PB per cluster	Adequate storage capacity for any size AI project.
	Software-defined ("cloud-native") storage	Presents on-premises, private cloud, and public cloud storage to applications as feature-consistent block volumes or NFS file systems.
	Public and hybrid cloud storage	Supports mix of AWS or Microsoft Azure and on-premises storage for a Kubernetes cluster.
	Dynamic provisioning	Enables convenient self-provisioning of virtual storage resources for greater data scientist and developer agility.
	Easy data migration	Minimizes public cloud access charges by copying data to on-premises storage for use by applications in multiple clusters.
	Vector database support	Facilitates *Retrieval Augmented Generation* (RAG) for training customized GenAI models managed by Kubernetes.
	Python libraries for Portworx services	Simplifies common operations in data scientists' Jupyter notebooks for self-service resource instantiation and management.
	Disaggregated architecture	Enables independent scaling of computing and storage eliminates need to provision unnecessary resources.
	Storage class, resource, and application templates	Standardizes Kubernetes resources for consistency while eliminating needs for developer awareness of implementation details.
	Data replication and backup	Transportable snapshots and backup of Kubernetes environments protect data protection transparently to developers.
	Application checkpoints	Enables easy-to-use preservation of progress in long-running jobs.

TABLE 1 Portworx Features That Optimize AI Projects

While Portworx is designed to utilize any storage platform, combining it with Pure's systems creates a reliable, scalable, high-performing storage, data protection, and resource provisioning and workflow management environment for containerized AI projects.

At the End of the Day...

The Pure Storage system portfolio includes storage for all phases of AI projects, large and small. Pure's systems relieve IT, data scientists and MLOps teams from most common storage management tasks. With them, data scientists can concentrate on modeling and MLOps teams can provide reliable, scalable, high-performing environments for project data with sharable storage that expands to meet both training and production needs without disruption.

Available in performance-optimized and capacity-optimized models that scale on demand, The Pure Storage platform accelerates and enhances AI projects for healthcare, genomics, exploration, financial and many other fields. It can do so while sharing storage capacity and I/O bandwidth with other data-intensive applications such as analytics, database backup and restore, software development, media and entertainment post-production, electronic design automation (EDA) and more. .

Portworx takes the guesswork out of creating robust, scalable Kubernetes environments for containerized AI training and production applications. Its templates simplify configuring and implementing applications and databases commonly used in AI projects; its built-in storage services enable data sharing and backup of both project data and entire project environments.

For Further information...

Additional information on topics discussed in this brief is available at:

Pure Storage in Artificial Intelligence

https://www.purestorage.com/solutions/analytics-and-ai/artificial-intelligence.html

Overview of Pure Storage Portworx

https://portworx.com/

Appendix A: AI Storage Pitfalls

Data storage can have a profound impact on in-house AI project cost, development speed, and ultimately, project success. This appendix lists pitfalls to be avoided when planning and developing project storage infrastructures.

Configuration Rigidity

It is virtually impossible to predict lifetime storage needs accurately at the start of an AI project. Storage systems that are awkward or impossible to reconfigure, expand, or upgrade should therefore be avoided. Even upgradable systems can be problematic if upgrading entails service outages, especially in the production phase when inferences are driving business decisions.

Fragility

If an AI project is important enough to invest in, it's important enough to keep available. Data scientist productivity is important, but as projects move to intensive training and production, availability is vital. Storage that can't survive component failures, be repaired or upgraded online, or recover from user and administrative errors isn't up to the AI job.

Data Isolation

Data that can't be easily shared among developers and between containerized training apps and production tasks must be copied from where it is to where it's needed. As data sets expand, copying large data sets becomes disruptive. In addition to being time and resource-consuming, data set copying initiated by humans can be error-prone.

Inflexibility

I/O requirements vary throughout the life of an AI project. Storage that is highly optimized for narrow use cases (e.g., by data set location, block sizes, file types, etc.) can make it difficult to respond as a project evolves. Manual storage "tuning" interrupts development and production and requires expensive expertise. And even experts don't always get it right.

Data Model Rigidity

At the outset, AI projects typically utilize files and/or structured databases, hosted either on block storage or in file servers. As models develop and training data scales, many tools use data in object form to simplify organization and processing and construct vector databases that maximize the quality of search results. Storage systems that support only a single data model (e.g., file servers, block storage systems) may make converting input data into forms usable by training jobs and in production awkward and time-consuming.

Appendix B: Using Pure's Products for AI

The Pure Storage portfolio of all-flash systems meets AI capacity, I/O performance, and orchestration needs from project conception through model development, on into production. Table 2 suggests the optimal applications for each of Pure's systems in AI projects.

	Product	Training	Production	Infrastructure
UFFO	FlashBlade//S™ 0.35PB—19.6PB (physical)	very large objects (images, video,...)	I/O-intensive production (e.g., GenAI)	• Vector databases • Retrieval Augmented Generation (RAG)
UBF	FlashArray//XL™ 0.96PB—5.5PB (effective)	large data sets	Latency-sensitive production	• Ecosystem streaming (Kafka, etc.) • Kubernetes volumes & VMs
UBF	FlashArray//X™ 0.31PB—3.3PB (effective)	medium-large data sets		• Vector databases • Retrieval Augmented Generation (RAG) • Kubernetes volumes & VMs
UBF	FlashArray//C™ 1.6PB—8.9PB (effective)		• Feature stores • Inference input & output	
UDR	FlashArray//E™ 1.0PB—4.0PB (effective) FlashBlade//E™ 4.0PB—8.0PB (physical)	Raw training data seldom-accessed after initial use	Feature stores Inference input & output	Exception logging Project archives

TABLE 2 Pure Storage Systems in AI Projects

Storage for Every AI Project

As Table 2 suggests, Pure's systems offer a breadth of capacity, cost/performance, and protocol support that are a fit for all aspects of an AI project. In terms of the desirable AI storage properties listed on page 5:

Performance Agility

The portfolio includes a range of cost/performance options to meet lifetime project needs:

- All systems utilize reliable, high-performing, power-efficient flash-based DFMs exclusively.
- Latency-optimized UBF systems support *quality-of-service* (QoS) limits for prioritizing performance among data sets.
- Throughput-optimized UFFO systems scale out—adding blades to a system increases both storage capacity, network performance, and processing capability.
- Cost-optimized UDR systems make retaining large amounts of infrequently-accessed data affordable.
- Rapid File Toolkit for UFFO systems accelerates bulk operations on millions of files.

Capacity Flexibility

All Pure systems are easily expandable from minimum to maximum capacity by non-disruptive addition of DFMs (and blades in scale-out UFFO systems). AI project architects can specify systems based on production expectations with initial configurations sized to meet early-stage capacity needs and add capacity incrementally as projects progress:

Availability & Durability

Pure's systems provide reliable "24×7" storage for AI model training and production:

- Systems provide continuous availability over lifetimes that exceed a decade.
- Systems are highly available—internal redundancy allows them to continue operating if components fail. Purity software is designed to sustain a minimum of two concurrent DFM failures (in many cases, more) without loss of data availability.
- All updates and modernizations are *non-disruptive*, performed while systems are serving client I/O requests and with minimal performance impact.
- DFM service lifetimes are over 10 years—far longer than typical SSD specifications. Data durability (i.e., statistical *mean time to data loss*, or MTTDL) due to DFM failure is on the order of millions of years.

Space and Power Efficiency

Pure's high-density all-flash system architectures:

- Hold up to 2 petabytes of physical storage in five rack units.
- Consume up to 90% less power per terabyte than competitive storage systems.
- Result in much less "e-waste" due to longer lifetimes of devices and storage systems.

Data Models

AI development and production tools utilize block storage, file servers, or object stores. Pure's systems support block, file, and object protocols for easy integration:

- UBF systems support client-side file systems and object stores with virtual block devices as well as providing file services for hundreds of file systems containing as many as a half-billion files.
- UFFO systems support petabyte-scale data sets containing billions of files or S3 objects in thousands of file systems or object buckets.
- UDR systems store less-frequently accessed data, whether blocks, files, or objects, economically.

Security

Many AI projects utilize proprietary data. As models mature, both curated training data and the intellectual property in models' digital representations become increasingly valuable. Pure's systems provide a secure environment for digital information:

- They use strong administrator authentication to control access.
- They automatically encrypt all stored data and metadata.
- They can be configured to take periodic snapshots of selected data to protect against application and administrative errors.
- Users can enable *SafeMode*™ to protect against malware (e.g., ransomware).

Operational Simplicity

As AI projects evolve from concept through training and into production, a smooth operating environment becomes increasingly important. Pure storage systems' ease of operation can simplify a project's operating environment in several ways:

- They are typically installed and ready for use in a few hours rather than days.
- With support for all common I/O protocols (NVMe, NFS, SMB, S3, iSCSI, GPUDirect), they integrate easily into data centers. UBF systems can provide block and file services simultaneously; UFFO systems can simultaneously serve files and objects.
- They eliminate many common management operations—specifying protection, placing data sets, reserving spare capacity, tuning for performance, etc. are all autonomous.
- Both software/hardware upgrades and system modernizations are *non-disruptive* (i.e., are performed while systems are operating) and have negligible impact on performance. Data migration is never required.
- They are self-monitoring; they upload status and event logs[9] every few seconds to the Pure1 monitoring and consolidated management service (available to all customers with active Evergreen subscriptions). Pure1's AI-based behavioral models identify and report potential issues and provide experience-based capacity planning advice.

Accelerate AI Adoption With The Pure Storage Platform

1 For GenAI, "input data" usually takes the form of natural language queries.
cf: https://www.techtarget.com/searchdatamanagement/feature/Vector-search-now-a-critical-component-of-GenAI-development
"LLMs...are trained with extensive vocabularies and can determine the meaning of a question even if it isn't phrased in the business-specific language..."
2 DFMs are Pure Storage-designed flash memory modules that provide the persistent storage in all Pure systems. They fulfill the role played by SSDs in conventional flash-based storage systems.
3 Up to 75 terabytes per module at publication time.
4 Effective capacity is that available for user data, net of erasure code overhead and system metadata. Users typically experience up to 2:1 data reduction with FlashBlade compression and as much as 5:1 with FlashArray deduplication and compression. Data reduction depends strongly on the nature of data. For example, structured text and tabular data usually reduces by at least 2:1, whereas images, streams, and encrypted data are essentially uncompressible.
5 https://www.britannica.com/topic/NVIDIA-Corporation
6 https://www.purestorage.com/docs.html?item=/type/pdf/subtype/doc/path/content/dam/pdf/en/reference-architectures/ra-airi-nvidia-dgx-basepod-architecture-config-guide.pdf
7 GPUDirect is a registered trademark of NVIDIA Corporation.
8 https://catalog.ngc.nvidia.com/orgs/nvidia/containers/k8s-device-plugin
9 Except where customer policies or regulations prohibit external connections.

PS2571-02-en 05/24