PRESIDIO | EVERPURE

High-Performance Storage as the Backbone of AI Infrastructure

A Presidio Perspective with Everpure (Powered by FlashBlade Architecture)

Contents

Intended Audience

This white paper is intended primarily for infrastructure and platform architects responsible for designing, sizing, optimizing, or remediating the data path supporting enterprise AI workloads. The technical sections assume a working knowledge of enterprise infrastructure, data platforms, storage architecture, and AI workload characteristics.

Two secondary audiences are served by specific parts of the document:

- **AI platform owners, and heads of data or ML engineering:** The PATH sections examine the operational model, governance framework, and platform services that connect the underlying infrastructure to AI applications and development workflows.
- **Technology executives sponsoring AI infrastructure investment:** The Executive Summary and The Cost of Poor Storage Performance present the business context, financial considerations, and investment rationale without requiring the reader to engage with the detailed architectural analysis.

Document Structure

The *Executive Summary* and *The Hidden Bottleneck* establish the business and operational context for the discussion. *Why High-Performance Storage Matters* for AI and the Everpure FlashBlade architecture deep dive form the technical core of the paper and are intended primarily for infrastructure and platform architects.

The PATH sections describe the Presidio platform and operational layer that sits above the storage infrastructure, including the mechanisms used to deliver, manage, and govern AI platform services. Readers focused primarily on the business case may conclude their review after *The Cost of Poor Storage Performance*. Readers responsible for architecture, implementation, or platform operations should continue through the technical and PATH sections.

Scope

This white paper evaluates storage performance as one of several factors that can influence the efficiency, reliability, scalability, and economics of enterprise AI infrastructure. It does not characterize storage as the sole or primary determinant of AI program success.

Industry research concerning AI program outcomes is included to establish the broader challenge organizations face in achieving measurable returns from AI investments. It should not be interpreted as evidence that storage limitations are the primary cause of unsuccessful or underperforming AI initiatives. The paper's analysis is limited to identifying the conditions under which storage architecture and data-path performance can materially affect specific AI infrastructure and operational outcomes.

Executive Summary

Artificial Intelligence (AI) is transforming industries, and its success depends on a number of factors. One of the most consistently overlooked is storage performance. Modern AI workloads are intensely data-hungry and highly latency-sensitive. When storage cannot keep pace, GPUs and CPUs—among the most expensive resources in an AI stack—sit idle. The result is slower training and inference, reduced model accuracy iteration speed, and significantly increased operational costs.

The business consequences of getting AI infrastructure wrong are visible on the balance sheet. MIT Project NANDA's The GenAI Divide: State of AI in Business 2025 found that, despite an estimated $30–40 billion in enterprise generative AI investment, the vast majority of integrated AI pilots had yet to produce measurable P&L impact. The study identifies workflow integration, contextual learning, and alignment with day-to-day operations as key factors separating successful deployments from stalled initiatives. These findings establish the broader business imperative for converting AI investment into production value. Within that larger challenge, this paper examines one specific dimension: how the storage architecture and data path can affect the performance, efficiency, and economics of enterprise AI workloads. Boards have stopped asking whether AI budget was approved and started asking what it returned.

A pattern Presidio encounters repeatedly in infrastructure assessments is an environment provisioned generously for compute and thinly for data. Organizations buy accelerators, then discover that the pipeline feeding them cannot sustain the throughput those accelerators require. The visible symptoms are stranded capital in depreciating GPUs, training cycles too slow to support real experimentation, and data preparation stages that run weeks longer than planned. Where the data path is the binding constraint, remediating it is among the highest-leverage changes available. Where it is not, storage investment will not rescue a program failing for other reasons, and the assessment described at the end of this paper is how Presidio establishes which case applies.

This whitepaper explores why high-performance storage is foundational to AI success, the risks of under-provisioned data infrastructure, and how Presidio, in partnership with Everpure, delivers optimized storage architectures designed to unlock the full potential of AI investments.

The AI Infrastructure Challenge

AI workloads differ fundamentally from traditional enterprise applications:

- Massive datasets (terabytes to petabytes)
- High-throughput data pipelines
- Mixed random and sequential I/O patterns, often with small-file metadata storms
- Real-time inference requirements with strict tail-latency SLAs

- **Metadata-bound access patterns:** training corpora of hundreds of millions of small files, where stat, open, and list operations dominate the I/O profile and a metadata tier that cannot scale independently becomes the limiting factor long before raw throughput does

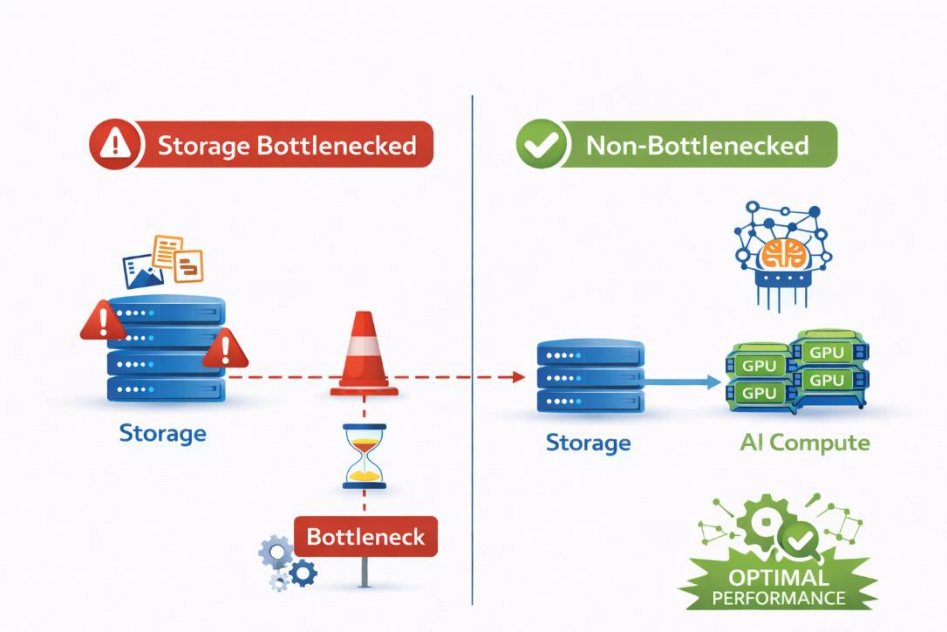
Unlike conventional systems, AI pipelines continuously move data between storage, memory, and compute layers. This creates a critical dependency: **storage performance directly impacts compute efficiency**.

The Hidden Bottleneck

Organizations often invest heavily in GPUs and CPUs but underestimate storage requirements. This leads to a mismatch:

- GPUs waiting on data reads
- CPUs stalled during preprocessing and shuffle phases
- Training pipelines underutilized due to checkpoint write contention

In these scenarios, expensive compute resources are under-utilized. Cast AI's 2026 State of Kubernetes Optimization Report, drawn from roughly 23,000 clusters across the major hyperscalers, put average enterprise GPU utilization at approximately 5%. That figure reflects provisioned capacity against consumed capacity across all clusters, and is driven principally by over-provisioning, idle allocation, and scheduling inefficiency. Academic characterization of production LLM training clusters reports median GPU streaming-multiprocessor activity near 40% during active jobs. Presidio's own infrastructure assessments typically find enterprise training environments operating in a 30–50% band. Storage and pipeline performance is one contributor to that gap, alongside scheduler behavior, batch sizing, network topology, and model implementation. This paper addresses the storage contribution specifically. Where storage is a factor, the effect is amplified by the location of the data relative to the accelerators and by the access patterns the workload generates.



Observed Impact Where Storage Is the Binding Constraint

The ranges below are drawn from Presidio engagements in which storage was measured and identified as the binding constraint before remediation. They are not a forecast for any specific environment. Where the limiting factor is scheduling, model implementation, or network design, storage improvements produce materially smaller gains.

- **GPU utilization:** 30–50% → 60–80%
- **Training time reduction:** 1.5x–3x faster
- **Data pipeline throughput:** 2x or better
- **Infrastructure cost efficiency:** 10–25% reduction in cost per model
- **Inference latency:** 15–40% lower response times

Another key point to consider is the time required to get data ready for consumption, typically clients must account for the hidden cost of data preparation, including:

- **Identify and Discover** – 1-2 weeks, Persona: Data Engineer – Define business requirements and identify relevant data sources.
- **Ingest** – 2-4 weeks, Persona: Data Engineer – load data using LangChain, Fivetran, Airbyte, AWS Glue, Azure Data Factory.
- **Curate** – 3-8 weeks, Persona: Data Scientist/Analyst SME – clean and prepare data using Python, Pandas, Databricks.
- **Transform** – 4-8 weeks, Persona: ML Engineer – experiment with chunking strategies using LangChain, Prefect, Kubeflow.
- **Vectorize** – 5-12 weeks, Persona: ML Engineer – generate vector representations using Hugging Face, OpenAI, Cohere.

- **Index** – 6-14 weeks, Persona: ML Engineer, Data Engineer – load embeddings into vector databases like Pinecone, Weaviate, Milvus.
- **Serve** – 7-16 weeks, Persona: ML Engineer – build LLM application using LangChain, Sagemaker, Google Vertex AI.

These timelines are not fixed costs of doing AI — they are largely a function of the middle-layer tooling and the data path underneath it. We return to this table in the PATH section, under Compressing the Data Preparation Timeline, to show which of these stages compress and why.

Real-World Impact

In environments where Presidio measured the data path to be the constraint, remediation has produced:

- 1.5–3x faster model training cycles
- Measurable reductions in idle accelerator time and the infrastructure spend attached to it
- Improved responsiveness for real-time applications

Presidio's integration expertise combined with Everpure's storage performance enables enterprises to fully realize these benefits.

Why High-Performance Storage Matters for AI

1. Eliminating GPU and CPU Idle Time

AI accelerators require a constant stream of data. If storage throughput or latency is insufficient, training jobs stall, batch processing slows, and inference latency increases. High-performance storage ensures data is delivered at line speed, keeping compute resources fully utilized.

2. Accelerating Model Training

Training large models involves repeated passes over datasets. Faster storage enables reduced epoch times, faster experimentation cycles, and quicker time-to-insight.

3. Enabling Real-Time Inference

For applications like fraud detection, recommendation engines, and autonomous systems, milliseconds matter. Low-latency storage reduces response times, improves user experience, and enables edge and real-time AI use cases.

4. Supporting Scalable Data Pipelines

AI environments must scale seamlessly. High-performance storage supports parallel data access, distributed training, and multi-tenant workloads.

The Cost of Poor Storage Performance

Increased Infrastructure Costs

Idle GPUs waste capital investment. Organizations may compensate by adding more compute instead of fixing the root cause.

Slower Innovation Cycles

Longer training times delay model iteration, reducing competitiveness.

Operational Inefficiencies

- Pipeline bottlenecks
- Resource contention
- Increased management complexity

Energy and Sustainability Impact

Inefficient systems consume more power per unit of useful work, increasing both cost and environmental footprint.

Presidio's Approach to AI Storage Optimization

Presidio brings deep expertise in designing, integrating, and optimizing AI infrastructure. Our approach focuses on aligning storage performance with compute demands.

Key Principles

- **Performance-first architecture:** Storage designed to match GPU throughput requirements
- **End-to-end integration:** Seamless connectivity across compute, networking, and storage layers
- **Workload-aware design:** Tailored solutions for training, inference, and hybrid workloads

Services Offered

- AI infrastructure assessment
- Performance benchmarking
- Architecture design and deployment
- Ongoing optimization and support

Everpure FlashBlade Architecture: A Technical Deep Dive

Before examining the FlashBlade architecture, which is the primary technical focus of this document, it is worth noting that AI data storage can encompass a wide range of platforms and technologies. Everpure's Enterprise Data Cloud is outside the scope of this paper; however, the combination of Purity, Pure1, and Fusion gives organizations a common control plane over enterprise data from the edge to on-premises and public cloud, across a wide range of storage architectures.

The focus of this whitepaper is the Everpure FlashBlade architecture—a disaggregated, all-NVMe, scale-out platform purpose-engineered for unstructured data and modern AI/ML workloads, and part of Everpure's larger Enterprise Data Cloud architecture. Unlike legacy scale-up filers or hybrid arrays retrofitted for AI, Everpure separates compute, capacity, and networking into independently scalable building blocks while presenting a single unified namespace to applications. The sections below detail the architectural pillars that translate directly into measurable AI performance gains.

1. Disaggregated Scale-Out Blade Architecture

The Everpure FlashBlade//S platform is a chassis-based system in which each blade is an independent compute and storage node. Capacity, performance, and metadata processing can all increase as blades are added, though read and write throughput scale at different rates and heterogeneous or partially populated configurations behave differently again. Size configurations against the current Everpure Sizer rather than against a single uniform multiplier. There is no controller pair and no separate metadata server tier to provision.

- **Chassis density:** Up to 10 blades per 5U chassis; up to 10 chassis combined into a single namespace
- **Per-blade capacity:** Each blade hosts up to four DirectFlash Modules (DFMs). R2 greenfield systems support 37.5 TB, 75 TB, 150 TB, and 300 TB DFMs; 24 TB and 48 TB are retained for R1-to-R2 upgrade paths. 150 TB requires Purity//FB 4.6.3 and 300 TB requires 4.8.0, so availability is release- and SKU-dependent. Confirm supported capacities against current FlashBlade//S configuration rules
- **Maximum raw capacity:** Current configuration rules list up to 12 PB raw per chassis on both FlashBlade//S200R2 and FlashBlade//S500R2 with 300 TB DFMs, starting with Purity//FB 4.8.0 and on limited SKUs. Raw capacity is not usable capacity; size usable capacity in Sizer. Multi-petabyte single namespaces are achievable in a single rack
- **Scale-out performance:** Capacity, throughput, IOPS, and metadata performance can all increase as blades are added. Read and write performance do not scale by a fixed multiplier. Actual results depend on blade and DFM population, configuration homogeneity, workload profile, and network topology. Size every configuration in the Everpure Sizer, which is the source of truth
- **Independent fault domains:** Blade-level redundancy, with N+2 within the compute resiliency group, typically 6 to 16 blades. This is the compute resiliency level and not the

storage data-resiliency level, which varies with chassis and DFM population; non-disruptive blade replacement and online expansion

- **Automatic data resiliency:** Purity//FB sets the erasure-coding formation automatically from the system configuration. Formations such as N+2, N+3, and N+4 across the DirectFlash Modules follow from DFM count, blade count, and chassis topology rather than from a user-selected setting.

DirectFlash: Eliminating the SSD Translation Tax

Everpure does not use commodity SSDs. Instead, Purity//FB writes directly to raw NAND through DirectFlash Modules, removing the redundant flash translation layer (FTL) found in off-the-shelf drives. This yields:

- Lower module failure rates than commodity SSD and HDD designs. Comparative annual failure rate figures depend on module type, population size, and measurement period. Source them from current Everpure reliability documentation rather than from this paper
- Lower and more predictable tail latency under sustained AI training I/O patterns
- Global wear-leveling and garbage collection coordinated across the entire array
- Higher rack density and lower watts-per-TB than enterprise SSD designs

2. Purity//FB: The Distributed Operating Environment

Purity//FB is the software layer that makes the disaggregated hardware behave as a single, transactionally consistent system. Every blade contributes to a fully distributed metadata engine and a distributed data plane—there is no master node to recover and no metadata server to size separately.

- **Distributed metadata engine:** Metadata is sharded and replicated across all blades, scaling to billions of files and objects per namespace without the cliff seen in traditional NFS filers
- **Transactional consistency:** File and object writes commit transactionally; failure domains are isolated to the affected blade
- **Distributed metadata structures:** Purity//FB metadata structures are designed to sustain a high rate of concurrent metadata transactions in a scale-out system.
- **Concurrency control:** Purity//FB locking is designed to sustain heavy concurrent metadata load.
- **Always-on data services:** Inline always-on data reduction (deduplication and compression), AES-256 data-at-rest encryption, and fast snapshots
- **Non-disruptive everything:** Blade adds, removes, and software upgrades occur online; no maintenance windows for capacity expansion
- **Replication and DR:** Native asynchronous file and object replication, plus SafeMode immutable snapshots for ransomware resilience

- **Pure1 telemetry:** Cloud-based AI-driven analytics for predictive support and capacity/performance forecasting

3. Unified File and Object: One Namespace for the AI Pipeline

AI pipelines rarely use a single protocol. Ingestion, ETL, training, and inference often span POSIX file access, SMB shares for Windows-based labeling tools, and S3 object access for data lakes and modern frameworks. Everpure presents all of these from a single shared namespace, eliminating the data copies that traditionally fragment AI environments.

Protocol	Versions Supported	Typical AI Pipeline Use
NFS	v3, v4.1 (with pNFS-style parallel mounts)	PyTorch / TensorFlow training datasets, checkpoints, scratch space
SMB	SMB 2.1, SMB 3.x (incl. multichannel)	Windows-based data labeling, annotation, and curation tools
S3	S3-compatible object API (versioning, MPU, bucket policies)	Data lake ingestion, MLflow artifacts, Hugging Face datasets, vector store cold tier
NFS-RDMA	NFS over RDMA / RoCEv2	GPU-server-to-storage bulk reads at line rate

- **Single namespace, multi-protocol:** The same dataset is concurrently readable as files (NFS/SMB) and as objects (S3) without copies or sync jobs
- **RapidFile Toolkit:** Drop-in replacements for ls, cp, chmod, chown, find, and rm that execute in parallel against Purity//FB, accelerating dataset preparation by 10–30x versus stock GNU coreutils
- **Multi-tenancy:** File system quotas, S3 bucket policies, and per-tenant performance shaping support shared AI platforms across business units

4. AI-Specific Integrations: GPUDirect, RDMA, and Reference Architectures

Everpure FlashBlade is engineered for the highest-end NVIDIA GPU compute platforms and is part of NVIDIA's certified storage ecosystem.

- **NVIDIA GPUDirect Storage (GDS):** Certified GDS support enables direct DMA from FlashBlade into GPU memory, bypassing the CPU bounce buffer and reducing latency for large training reads
- **RDMA over Converged Ethernet (RoCEv2):** NFS over RDMA cuts CPU overhead on GPU servers and improves small-IO latency consistency under load
- **Ethernet connectivity:** Each blade fronted by high-speed Ethernet; Fabric I/O Module (FIOM) ports are 100 GbE. 400 GbE is aggregate, XFM-side connectivity available only in specific S500R2 configurations: the 4 x 100G XFM-to-FIOM option consumes all eight

FIOM ports and is supported on greenfield S500R2 multi-chassis systems from Purity//FB 4.6.3, with non-disruptive upgrade support from 4.6.5

- **AIRI//S reference architecture:** Joint Everpure + NVIDIA AI-Ready Infrastructure validated with NVIDIA DGX systems
- **NVIDIA SuperPOD and BasePOD certified:** One of a small number of storage platforms certified for DGX SuperPOD-class deployments
- **MLPerf-aligned:** Sustains the high concurrent-stream read throughput required by MLPerf Storage benchmarks (e.g., ResNet, 3D U-Net, CosmoFlow workloads)

5. Performance and Scale at a Glance

Configuration reference for Everpure FlashBlade//S deployments. Performance figures are deliberately not published here; all figures are configuration-, workload-, and software-version-dependent. Presidio does not publish fixed performance numbers for Everpure platforms. Size every deployment using the current Everpure (Pure Storage) FlashBlade//S datasheet, configuration rules, and Sizer, and record the exact configuration, software version, and units alongside any result.

Dimension	FlashBlade//S200R2	FlashBlade//S500R2
Workload Profile	Capacity-optimized	Performance-optimized
Throughput and IOPS	TBD — size in Sizer	TBD — size in Sizer
Latency (typical small-file metadata)	TBD — size in Sizer	TBD — size in Sizer
Per-blade DFM count	1–4 DFMs	1–4 DFMs
DFM raw capacities	R2 greenfield: 37.5 / 75 / 150 / 300 TB; 24 / 48 TB retained for R1→R2 upgrades (150 TB from Purity//FB 4.6.3, 300 TB from 4.8.0)	R2 greenfield: 37.5 / 75 / 150 / 300 TB; 24 / 48 TB retained for R1→R2 upgrades (150 TB from Purity//FB 4.6.3, 300 TB from 4.8.0)
Networking per blade	50 GbE per blade to each FIOM (100 GbE aggregate across the two FIOM links)	100 GbE per blade to each FIOM (200 GbE aggregate across the two FIOM links)
Maximum raw per chassis	Up to 12 PB raw with 300 TB DFMs (Purity//FB 4.8.0 and later, limited SKUs); raw, not usable	Up to 12 PB raw with 300 TB DFMs (Purity//FB 4.8.0 and later, limited SKUs); raw, not usable
Single namespace scale	Multi-PB across 10 chassis	Multi-PB across 10 chassis
Protocols	NFS / SMB / S3	NFS / SMB / S3 / NFS-RDMA

FlashBlade//S200R2 and FlashBlade//S500R2 are separate system profiles and are not mixed within a single system. Use current Everpure configuration rules and Sizer to establish the exact profile, supported DFM capacities, and expected performance for a given deployment.

6. Mapping Architecture to AI Outcomes

Each architectural attribute of Everpure FlashBlade maps to a specific failure mode in under-provisioned AI environments. The table below makes that mapping explicit so infrastructure architects can size with intent.

AI Pain Point	Everpure / FlashBlade Architectural Answer
GPUs idle waiting for batch reads	GPUDirect Storage + NFS-RDMA + per-blade 100/200 GbE; throughput increases with blade count
Small-file metadata storms (millions of training samples)	Distributed Purity//FB metadata engine; no single metadata server bottleneck
Checkpoint write spikes during training	All-NVMe DirectFlash with consistent write latency through checkpoint spikes; erasure-coding formation selected automatically by Purity//FB from DFM count, blade count, and chassis topology
Multi-protocol data fragmentation	Single namespace exposed concurrently as NFS, SMB, and S3—no copies between data lake and training tier
Capacity growth without performance growth	Adding blades can increase capacity, throughput, and metadata performance. Read and write scaling differ, and heterogeneous or partially populated configurations can be drive-bound. Use current Sizer and scaling guidance for the specific configuration
Slow data prep / ETL phases	RapidFile Toolkit parallelizes ls/cp/chmod/find by 10–30x
Operational risk and ransomware exposure	SafeMode immutable snapshots, AES-256 encryption, replication, and Pure1 predictive analytics

Reference Architecture: AI-Optimized Storage Stack

A modern AI infrastructure built with Presidio and Everpure includes:

- **Compute Layer:** Dell PowerEdge or Cisco UCS with NVIDIA GPU clusters for training and inference
- **High-Speed Networking:** 400/800 Gb RoCEv2 for east-west GPU traffic; 100/200/400 GbE with RoCEv2 for storage traffic
- **Everpure Layer:** FlashBlade//S with GPUDirect Storage and unified file/object access

- **Data Pipeline Layer:** Ingestion, preprocessing, feature stores, and orchestration (e.g., Airflow, Kubeflow, Run:ai)

Best Practices for AI Storage Deployment

1. Align storage throughput to aggregate GPU read bandwidth, not just dataset size
2. Minimize latency across the data path—use NFS-RDMA / RoCEv2 and GPUDirect Storage where supported
3. Design for parallel access, and for capacity and performance that grow together as the system scales out
4. Continuously monitor and optimize using telemetry such as Pure1
5. Avoid retrofitting legacy scale-up filers or hybrid arrays for production AI training
6. Consolidate file and object on a single namespace to eliminate data copies between pipeline stages

Presidio PATH

Now that we have discussed the infrastructure challenges presented by AI, let's take a moment to evaluate the current state of AI in the enterprise. Presidio fundamentally believes AI has crossed a threshold. What began as isolated experiments – chatbots, recommendation engines, productivity copilots – has become foundational infrastructure. AI now powers decisions, operations, and customer experiences at the core of business. The organizations winning with AI are no longer asking whether to adopt it; they are asking how to own it.

PATH is Presidio's answer to that question. A note on naming, because it has changed: PATH is the Presidio portfolio name. PATH Stack is the enterprise AI platform described in this section. PATH Lab is the validation environment described later in this paper. The P.A.T.H. acronym form is deprecated and is not used in current Presidio materials.

PATH Stack is an enterprise AI platform built to give organizations greater control over how AI runs, where it runs, and what it costs.

Internal review note, not for external release: the PATH Stack capability descriptions in this section have been moderated from the previous draft and require sign-off from the Presidio Innovation team before publication. The items needing confirmation are the scope of what PATH Stack manages directly versus integrates, the portability claim across hybrid environments, and the data preparation compression figures in the table below.

PATH Stack reduces the complexity of the AI middle layer and delivers a consistent, secure, and scalable AI foundation with a common operating model across on-premises, colocation, neo-cloud, and hyperscaler environments. The intent is to make enterprise AI substantially simpler to stand up and operate, without trading away governance, compliance, or cost transparency to get there.

The Three-Layer Problem

The AI market is commonly described in three layers: the infrastructure required to run AI (GPUs, ultra-high-speed networks, storage, cybersecurity, and datacenters); the AI applications that deliver business outcomes (agents, copilots, and automation workflows); and, critically, everything in between.

That middle layer is where most enterprises are struggling. The tools required to connect infrastructure to applications – data pipelines, model serving frameworks, GPU virtualization, vector databases, inference orchestration, fine-tuning workflows – change rapidly and require specialized expertise that most organizations cannot build or retain at scale.

The Problem with the Current Model

As AI usage scales within an enterprise, the weaknesses of external API dependency become structural barriers. Consider what happens when AI moves from proof-of-concept to a production workload used by hundreds of employees:

The hidden costs of AI API dependency include:

- **Cost exposure**: token-based consumption pricing becomes unpredictable and unsustainable at enterprise scale.
- **Performance limits**: shared infrastructure means variable latency and no SLA control for mission-critical workloads.
- **Data sovereignty risk**: sensitive enterprise data processed by third party providers creates compliance exposure.
- **Vendor lock-in**: applications built on proprietary APIs cannot be re-platformed without re-engineering.
- **Governance gaps**: no visibility into model behavior, versioning, or audit trails across the organization.

A Different Question

Most enterprises frame AI adoption as a build-vs-buy decision applied to individual applications. That framing misses the deeper question. No organization would rent its ERP system, outsource

its network operating system, or accept that a third party controls the infrastructure its business depends on. AI is becoming that kind of foundational capability – and it deserves that same strategic treatment.

Do you want to rent AI – or own the system it runs on?

Owning the AI operating system does not mean build AI models from scratch. It means having a secure, consistent, infrastructure-agnostic platform that runs whatever models the business needs, in whatever environment makes sense, under full governance. PATH is that platform.

The Enterprise AI Operating System

PATH Stack provides organizations with a stable, consistent AI platform that spans the enterprise hybrid cloud environment. Whether AI workloads run on-premises, in colocation facilities, on neo-cloud providers, or in public cloud environments, PATH Stack is the constant layer above them.

Same Platform One AI operating model regardless of where workloads run.	**Same Security Model** Consistent governance, compliance, and access controls everywhere.
Same Operations A single operational methodology across every deployment.	**Same Developer Exp.** Developers work the same way in every environment.

This consistency is a deliberate architectural decision. As AI applications increasingly need to run close to their data sources, and as physical AI drives more hybrid deployment at the edge, maintaining a single operating model across environments becomes a meaningful operational advantage. Standardizing on PATH Stack is intended to give organizations a consistent basis for AI portability and governance as deployments spread across more environments.

What PATH Removes

PATH Stack takes on the specialized complexity of the AI middle layer. GPU orchestration, data modeling, ML pipeline management, and inference optimization are delivered as managed, governed capabilities rather than as bespoke engineering projects each enterprise staffs and maintains for itself.

PATH Stack is designed to manage the following:

- GPU orchestration and resource scheduling across hybrid environments.
- Data pipeline management and vector database operations.
- Model serving frameworks, versioning, and inference optimization.
- Fine-tuning workflows and model evaluation infrastructure.
- Security policy enforcement, audit logging and compliance reporting.

PATH and Everpure FlashBlade: Control Plane Meets Data Plane

Everything described in the architecture sections above — GPUDirect Storage, a distributed metadata engine, a single multi-protocol namespace — is data plane. PATH is the control plane

that sits above it. The two are designed to be deployed together, and the relationship between them is specific rather than rhetorical.

PATH orchestrates AI workloads; Everpure FlashBlade//S is the low-latency, high-throughput data plane those workloads run against. When PATH schedules a training job, serves a model, or rebuilds a vector index, the operation succeeds or fails on how quickly the underlying storage can move data into GPU memory. PATH abstracts GPU orchestration, pipeline management, and inference optimization away from the enterprise — but abstraction does not repeal the physics underneath. It transfers responsibility for that physics to Presidio, and FlashBlade//S is how Presidio meets it. The table below makes the dependency explicit.

PATH Function	What It Demands of the Data Plane	Everpure FlashBlade//S Answer
GPU orchestration and scheduling across hybrid environments	Sustained line-rate reads to every scheduled GPU, in whichever environment the job lands	GPUDirect Storage and NFS over RDMA (RoCEv2); throughput increases with blade count, so the data path can be sized to keep pace as GPU capacity is added
Data pipeline management and vector database operations	High-concurrency small-file and metadata operations during ingest, chunking, and embedding	Distributed Purity//FB metadata engine sharded across all blades; no separate metadata server tier to become the bottleneck
Model serving, versioning, and inference optimization	Predictable tail latency under mixed read and write load	All-NVMe DirectFlash with consistent small-file metadata latency; erasure-coding formation selected automatically by Purity//FB from DFM count, blade count, and chassis topology
Fine-tuning workflows and model evaluation	Fast checkpoint writes and rapid dataset re-reads across experiment iterations	Consistent write latency through checkpoint spikes; fast snapshots for dataset and checkpoint versioning
Portability across on-premises, colocation, neo-cloud, and hyperscaler	Identical data semantics wherever the workload runs	Unified file and object namespace with native replication; Purity, Pure1, and Fusion provide a common control surface across sites
Governance, audit logging, and compliance reporting	Immutable, encrypted, auditable data at rest	SafeMode immutable snapshots, AES-256 encryption at rest, and asynchronous replication for DR

This pairing is also what makes the PATH Lab commitment to infrastructure-true testing meaningful. A proof of concept validated against a storage tier that cannot match the production data path is not a proof of concept — it is a demonstration. Running Lab workloads against

FlashBlade//S means the throughput, latency, and metadata behavior observed during the engagement are the characteristics the workload will encounter in production, so the cost model built in the Lab still holds when the workload scales.

Compressing the Data Preparation Timeline

Earlier in this paper we set out the hidden cost of data preparation: a sequence of seven stages, each carrying weeks of effort before a single production inference is served. That timeline is the clearest place to see what PATH is actually worth, because most of it is not intellectual work — it is infrastructure work, and waiting on I/O.

Stage	Baseline	Where the Time Goes	How PATH and Everpure Compress It
Identify & Discover	1–2 weeks	Locating and qualifying sources with no common catalogue	PATH governance layer inventories and classifies sources once; the work is reusable across every subsequent use case rather than repeated per project
Ingest	2–4 weeks	Building and maintaining a connector per source, then landing copies	Pre-integrated ingestion tooling, and a single Everpure namespace that is written once and read concurrently as file or object with no copy or sync step
Curate	3–8 weeks	Cleaning at scale; file operations across millions of small objects	RapidFile Toolkit parallelizes ls, cp, chmod, find, and rm by 10–30x; curation runs in place against the primary dataset rather than a staging copy
Transform	4–8 weeks	Chunking experiments, each requiring a full re-pass over the corpus	Managed experimentation environment plus a data plane fast enough to make re-passes cheap, so chunking strategy is tested empirically rather than guessed once and lived with
Vectorize	5–12 weeks	Embedding generation bounded by GPU throughput and data delivery	GPUDirect Storage keeps embedding GPUs fed at line rate; PATH schedules the job across whatever capacity is available, on-premises or in cloud
Index	6–14 weeks	Vector database standup, sizing, tuning, and ongoing operations	Vector database operations are abstracted and managed by PATH; FlashBlade serves as the object tier beneath the index
Serve	7–16 weeks	Serving frameworks, versioning, inference tuning, then a governance retrofit	Model serving, versioning, and inference optimization delivered as managed capabilities, with audit logging and access control present from day one rather than added afterwards

Two honest qualifications belong with that table. First, the compression is not uniform. Stages dominated by human judgment — deciding which data matters, agreeing what "clean" means for a given business domain — are shortened by better tooling but never eliminated by it. Second, the gains are largest precisely where the work is infrastructure-bound: ingest, curation, vectorization, and indexing. Those four stages carry the bulk of the elapsed time in most enterprise programs, and they are the stages most sensitive to both middle-layer tooling and storage performance.

This is the practical argument for evaluating the two layers together rather than sequentially. An enterprise that adopts a managed middle layer but leaves a legacy filer underneath it will find the orchestration is no longer the constraint — the data path is. An enterprise that buys high-performance storage without addressing the middle layer will have a fast data path feeding a pipeline that still takes months to assemble by hand. Presidio's position is that these are one decision, not two.

PATH Lab – From Experimentation to Production

Most enterprise AI initiatives stall between proof-of-concept and production. MIT's Project NANDA research found that only about 5% of custom enterprise AI tools reached production with measurable P&L impact. The authors attribute that gap primarily to integration, workflow design, and organizational approach rather than to model quality or regulation. Infrastructure is one element of integration and not the whole of it. The shortage is rarely of ideas; it is of the governance, cost clarity, and engineering discipline required to move responsibly from experiment to scale.

Four failure points commonly stall the move from POC to production:

- No governance framework: pilots run without compliance and audit infrastructure that production requires.
- Cost confusion: teams cannot accurately model what a use case will cost at scale before committing.
- Infrastructure mismatch: POC environments don't reflect the constraints of production hybrid infrastructure.
- Skills gap: moving from experiment to engineered system requires expertise most enterprises don't have on staff.

What PATH Lab Delivers

PATH Lab addresses each failure point directly, providing the environment, the tooling, the governance frameworks, and the expert guidance to make the transition from experiment to production more reliable and more predictable. Key capabilities of PATH Lab include:

- Use case validation: structured methodology to evaluate AI use cases against real enterprise requirements – technical feasibility, cost model, governance posture and business case.
- Infrastructure-true testing: workloads are tested against the actual hybrid infrastructure topology the production deployment will use – no surprises when you go live.

- Cost transparency: Granular cost modeling for GPU consumption, inference at scale, and operational overhead – enabling confident investment decisions before production commit.
- Governance-first design: compliance frameworks, audit logging, and access controls are built in from day one, not retrofitted after deployment.
- Expert-led delivery: Presidio AI Architects work alongside your team throughout the Lab engagement, transferring knowledge and enabling long-term self-sufficiency.
- **Accelerated time to value:** pre-integrated tooling and established methodologies are intended to shorten the time from POC approval to production deployment. The reduction achieved depends on use case complexity and on the state of the underlying data estate.

Why Presidio – The Differentiated Position

Presidio is a leading provider of AI business outcome solutions, helping organizations modernize IT, optimize performance, and accelerate innovation from edge to core to cloud and across modern networks, modern platforms, cyber security, digital, total experience, managed services and cost optimization practices.

Presidio occupies a unique position in the enterprise AI landscape: infrastructure expertise meets application-layer delivery, bridged by an AI operating system that works across environments. Where most vendors specialize in one layer of the stack, Presidio connects all three – and does so in a way that preserves enterprise choice at every level.

There are four differentiators that should be highlighted:

1. **Full Stack Accountability** – Presidio is accountable from infrastructure to application – not just one layer. We do not hand off when the hard part starts. Our architects span GPU infrastructure, the AI middle layer, and the application workloads that run on top of it.
2. **Infrastructure Agnostic by Design**: PATH works across every environment – on-premises, colocation, neo-cloud, and public cloud hyperscalers. Enterprises are not locked into a single deployment model, and they do not re-architect when their infrastructure strategy evolves.
3. **Governance Built-in, Not Bolted On**: Security, compliance, audit logging, and cost controls are designed into PATH from the ground up. Regulated industries do not need to choose between AI capability and governance posture – they get both.
4. **Proven Enterprise Trust**: Presidio has built deep client relationships across thousands of successful engagements over decades. That track record of accountability, knowledge transfer, and long-term partnership is the foundation PATH is built on.

Conclusion

AI infrastructure success is not determined solely by compute power. Storage performance is a critical enabler that directly impacts efficiency, cost, and innovation speed.

Without high-performance storage, organizations risk underutilizing their most valuable AI investments. GPUs and CPUs sit idle, training and inference slow down, and costs rise.

Presidio, in partnership with Everpure—built on the FlashBlade scale-out architecture, Purity//FB, and certified NVIDIA GPUDirect integrations—provides the expertise and technology required to eliminate these bottlenecks and deliver AI infrastructure that performs at scale.

Next Steps: How to Engage

The only reliable way to establish whether storage is constraining your AI program is to measure it against your own workloads rather than against a datasheet. Presidio offers three defined entry points. They can be run in sequence or taken independently, depending on where you are.

1. AI Infrastructure Assessment. Presidio architects profile your existing AI environment end to end: measured GPU utilization, storage throughput and latency under real load, metadata behavior, and pipeline stage timings. You receive a written findings report identifying where the bottlenecks actually are and a prioritized remediation plan with the expected effect of each change. Start here if you have AI workloads in production and suspect you are not getting what you paid for.

2. Storage Performance Benchmark and Proof of Concept. Presidio runs your actual datasets and training or inference jobs against an Everpure FlashBlade//S configuration sized for your workload profile. You receive before-and-after measurements of throughput, GPU utilization, and epoch or inference times on your own workloads, together with a sizing recommendation and configuration design. Start here if you have a specific workload and need evidence before a capital request.

3. PATH Lab Engagement. Presidio AI architects work alongside your team to validate a specific AI use case end to end against production-true hybrid infrastructure, on the storage platform the production deployment will use. You receive a validated use case, a granular cost model at production scale, a governance and audit framework, and a production deployment plan. Start here if you have a use case stuck between pilot and production.

How to Start

Contact your Presidio account executive, or reach the Presidio AI infrastructure team directly at https://www.presidio.com/contact-us/. To make the first conversation productive, bring whatever of the following you have to hand:

- **Workload profile:** the models and frameworks in use today, and whether the priority is training, fine-tuning, inference, or all three
- **Data profile:** approximate dataset size, file count, and average file size, plus current protocols in use (NFS, SMB, S3)
- **Compute inventory:** current GPU count and generation, and measured utilization if you have it

- **Target environments:** on-premises, colocation, neo-cloud, hyperscaler, or a hybrid of these
- **Constraints:** governance and compliance obligations, data residency requirements, and budget cycle timing

That is enough for Presidio to scope an assessment in a single working session. If you have none of it, the assessment is designed to produce exactly this information — start there.

Sources and References

- MIT Project NANDA, The GenAI Divide: State of AI in Business 2025 (July 2025) — 95% of enterprise GenAI pilots delivered no measurable P&L impact; approximately $30–40 billion in enterprise GenAI investment. The study attributes this outcome to integration, workflow design, and organizational approach. It makes no finding on storage or infrastructure performance, and is cited here only as evidence that an enterprise AI return problem exists.
- Cast AI, 2026 State of Kubernetes Optimization Report — average enterprise GPU utilization of approximately 5% across roughly 23,000 clusters. This is a provisioned-versus-consumed measure across all clusters and reflects over-provisioning, idle allocation, and scheduling inefficiency. It is not a measure of storage-induced stall and is not presented as one in this paper.
- Hu et al., Characterization of Large Language Model Development in the Datacenter (arXiv:2403.07648) — median GPU streaming-multiprocessor activity near 40% in production LLM training clusters.
- Everpure (Pure Storage) published FlashBlade//S datasheets, configuration rules, and AIRI//S reference architecture documentation — capacity and certification figures. Performance figures are configuration-, workload-, and software-version-dependent and must be generated in Sizer for the specific configuration rather than quoted from this paper.
- Presidio infrastructure assessment engagements — observed GPU utilization ranges in enterprise AI environments prior to remediation. Improvement ranges quoted in this paper are drawn from the subset of these engagements in which storage was measured to be the binding constraint.
- Everpure (Pure Storage) FlashBlade//S datasheet, configuration rules, and Sizer — supported DFM capacities and their release dependencies, erasure-coding behavior, FIOM and XFM connectivity, R2 model naming, per-chassis capacity, and read and write performance scaling. All performance and scaling behavior is workload-, configuration-, and software-version-dependent.

About Everpure

Everpure delivers high-performance, scalable storage solutions designed to meet the demands of modern data-intensive workloads, including AI and machine learning. Everpure offerings are built on the FlashBlade scale-out architecture and Purity//FB software, providing a unified file and object namespace, all-NVMe DirectFlash hardware, and certified integrations with the NVIDIA AI compute ecosystem.

For more information, contact Presidio to learn how to optimize your AI infrastructure with Everpure.