White Paper

Everpure

AMD

Everpure AI Factory with AMD

AMD Instinct MI350 series
and FlashBlade//SR2

Contents

Introduction

This white paper describes the Everpure™ AI factory with AMD, a turnkey, enterprise-grade AI infrastructure platform purpose-built for large-scale model training, inference, and high-performance computing (HPC) workloads. By pairing AMD Instinct™ MI350 series accelerators with Everpure FlashBlade//SR2 storage, organizations can build sovereign, high-performance AI factories that scale from hundreds to thousands of GPUs—all without forklift upgrades, planned downtime, or storage bottlenecks. This solution is channel-delivered and backed by the economics of Everpure Evergreen®.

Solution overview

Everpure AI factory with AMD, the solution described in this white paper, integrates AMD Instinct MI350 series GPU accelerators with Everpure FlashBlade//SR2, an industry-leading unified fast file and object (UFFO) storage platform. Together, they form a multi-user, scalable AI compute and storage fabric designed for simultaneous enterprise workloads with enterprise-grade access control, scheduling, and data management.

This solution addresses three compounding challenges enterprises face when building AI factories:

- **Storage bottlenecks**: AI training and inference demand sustained, multi-terabyte-per-second throughput. Inadequate storage stalls GPU utilization and inflates training time.
- **Operational complexity**: Disconnected storage, networking, and compute layers require deep expertise to deploy and tune, delaying time to value.
- **Cost and scalability risk**: Forklift upgrades and proprietary lock-in make AI infrastructure expensive to evolve as model sizes and cluster requirements grow.

Key solution benefits

- **Scalability**: Scales from 512 to 4,096+ AMD GPUs and from single-chassis to multi-chassis FlashBlade® configurations; ideal for large-scale large language model (LLM) training and inference
- **Performance**: Greater than 50% performance improvement for FlashBlade//SR2 compared to the first generation FlashBlade//S™; up to 4,096GB/s of aggregate read throughput matched to AMD MI350X workload I/O profiles
- **Software integration**: Ships with the AMD ROCm™ 7 open source software stack, including AI workflow management tools compatible with PyTorch, JAX, and vLLM
- **Nondisruptive upgrades**: Zero planned downtime for storage processor, capacity, and blade upgrades—within or across technology generations—with Evergreen Architecture.
- **Enterprise support**: Dedicated Everpure and AMD expertise with end-to-end services for deployment, tuning, and ongoing management
- **Agile AI platform**: Built for future innovations as AI requirements evolve; modular, flexible, and continuously improved through the Evergreen//One™ subscription model

Note: This solution has been tested on the Everpure AI factory with AMD internal research infrastructure and is designed to solve the most demanding computational problems of today.

Solution components

AMD Instinct MI350X accelerator

The AMD Instinct MI350X is a high-performance AI accelerator built on the AMD CDNA™ 4 architecture. It is engineered for the most demanding LLM training, fine-tuning, and inference workloads at enterprise scale.

- **Memory**: 288GB of HBM3E (approximately 1.5 times the capacity of the NVIDIA B200), enabling 500-billion-parameter models to run on fewer GPUs and reducing multi-GPU memory communication overhead
- **Precision support**: Native FP4 and FP6 support for a 35 times generational leap in inference throughput over the MI300 series (critical for high-throughput serving of large foundation models)
- **Software ecosystem**: Backed by the open source ROCm 7 stack with immediate compatibility with PyTorch, JAX, and vLLM; no proprietary lock-in on the software layer
- **Interconnect**: High-bandwidth, GPU-to-GPU interconnect fabric supporting large-model parallelism across multi-rack configurations

Note: High-resolution computer vision datasets can exceed 30TB and may require up to 4GB/s of read performance per GPU. Contact your Everpure representative for directional, workload-specific sizing guidance.

Everpure FlashBlade//SR2

FlashBlade is an industry-leading UFFO storage platform, engineered from the ground up for flash to deliver cost-effective reliability, simplicity, and multidimensional performance for exabyte-level modern datasets. The SR2 blade generation delivers greater than 50% performance improvement over the first generation without forklift upgrades.

Enterprise features include:

- **Multidimensional performance**: Very high throughput across any file size, sequential or random I/O, batched or real-time workloads, and tens of millions of file operations per second for sustained GPU utilization across all training phases
- **DirectFlash® architecture**: Eliminates translation layers with proprietary flash module design, delivering lower latency, higher density, and simpler management without constant tuning
- **Nondisruptive upgrades**: Hardware and software upgrades and capacity expansions without planned downtime, eliminating operational risk as AI infrastructure grows
- **Always available**: 99.9999% uptime SLA backed by Evergreen//One
- **Dynamic scalability**: Independent scaling of capacity, performance, metadata, and file/object counts without overprovisioning or data migration
- **Zero Move Tiering**: Intelligent data access across heterogeneous all-flash clusters with no manual data movement
- **Multi-protocol support**: Tuned for Network File System (NFS) over Remote Direct Memory Access over Converged Ethernet (RoCE); natively supports S3 and Server Message Block (SMB), enabling unified access across AI training, inference, and data preparation workloads

Architecture and integration

Storage fabric network integration

FlashBlade//SR2 connects to the Everpure AI factory with AMD environment via 16 400GbE ports, enabling seamless scaling from a minimal configuration into multiple chassis, blade, and DirectFlash Module configurations in a simple, nondisruptive manner.

Centralized storage enables seamless data access and sharing across distributed GPU compute environments, whether for healthcare AI imaging, financial risk modeling, or media content creation. RoCE further minimizes latency and offloads CPU workloads for demanding AI applications.

FlashBlade integrates with the Everpure AI factory with AMD cluster via a pair of Arista 7060DX4-32F switches—a standard process proven across thousands of enterprise deployments worldwide.

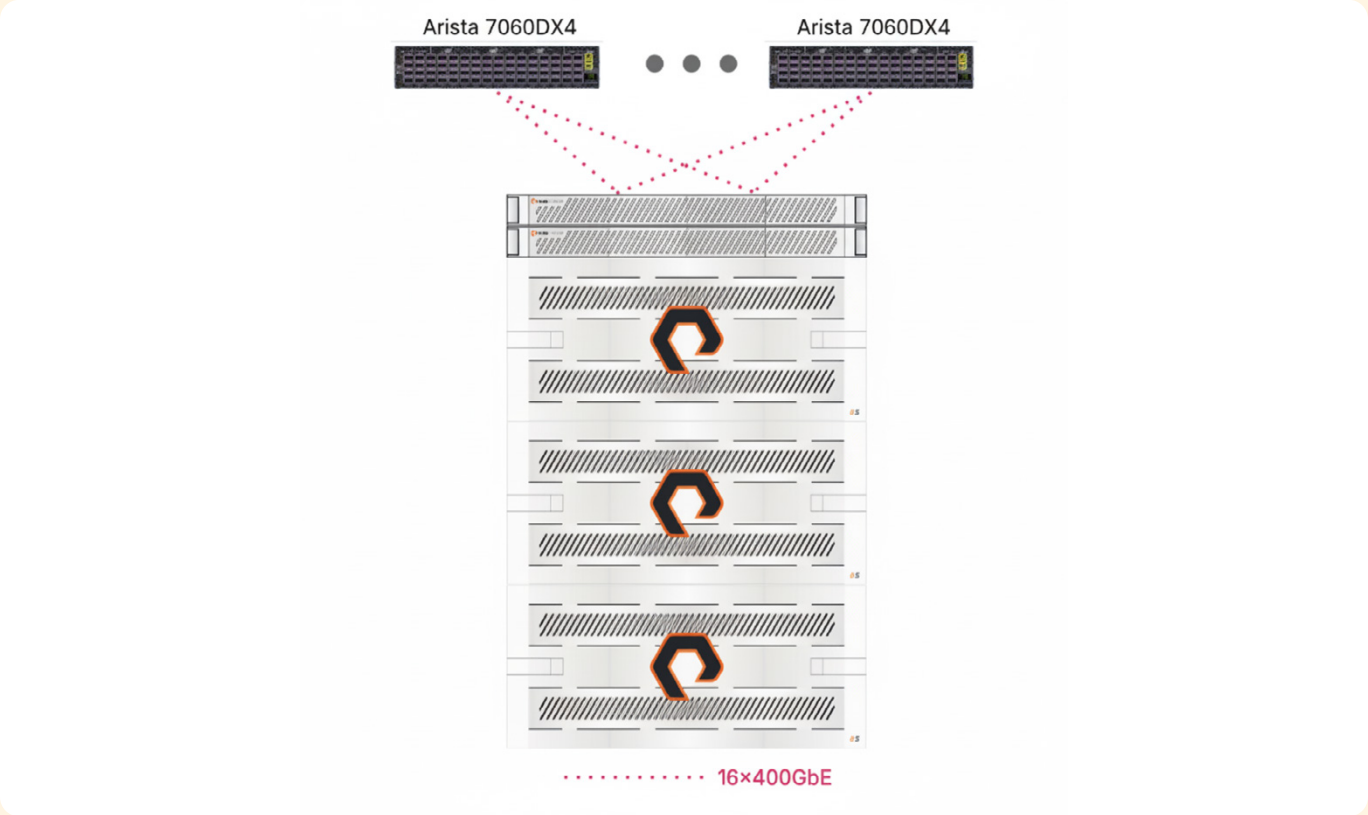

FIGURE 1 Everpure AI factory with AMD storage fabric network integration. FlashBlade//SR2 connects to AMD compute nodes via 16 400GbE uplinks through Arista 7060DX4-32F switches.

FlashBlade//SR2 physical architecture

A FlashBlade array comprises four key components working together to deliver multidimensional performance and nondisruptive upgrade capabilities:

- **External fabric modules (XFMs)**: Aggregate all traffic from the Arista 7060DX4-32F switches to the FlashBlade chassis blades
- **Chassis**: Houses the blades and provides the mechanical and power infrastructure for the storage nodes
- **Blades**: Individual storage compute nodes that run the Purity//FB operating environment and serve data over NFS, S3, and SMB
- **Direct flash modules**: High-density DirectFlashModules that attach to blades, providing raw flash capacity independent of blade compute resources

XFM8400 connectivity

Dual XFM8400 switches each uplink to the Arista 7060DX4-32F switches with eight connections for 16 400GbE connections total. This dual-path architecture provides full redundancy and maximum aggregate bandwidth for AI workloads.

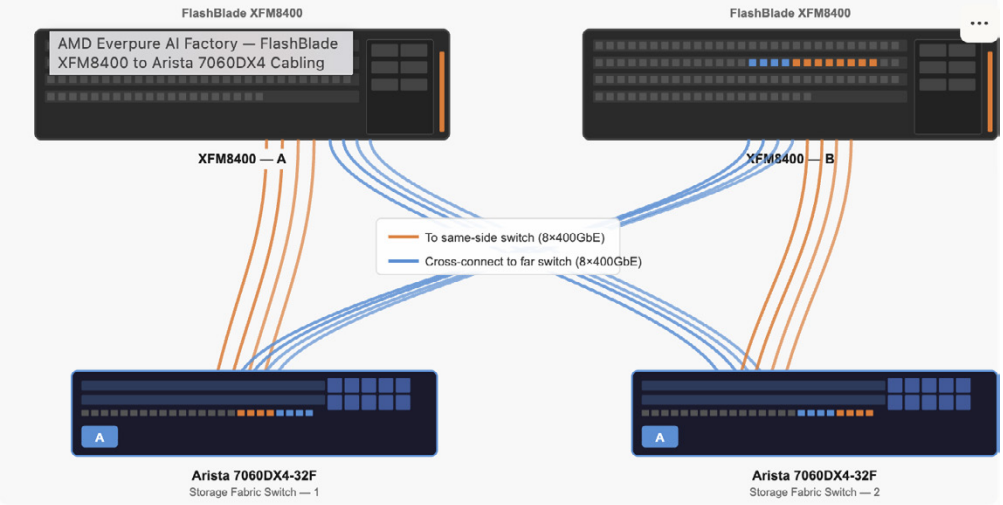

FIGURE 2 XFM8400 aggregate uplink connections to the Arista 7060DX4-32F storage fabric switch pair. Each XFM8400 provides eight 400GbE connections for 16 total uplinks.

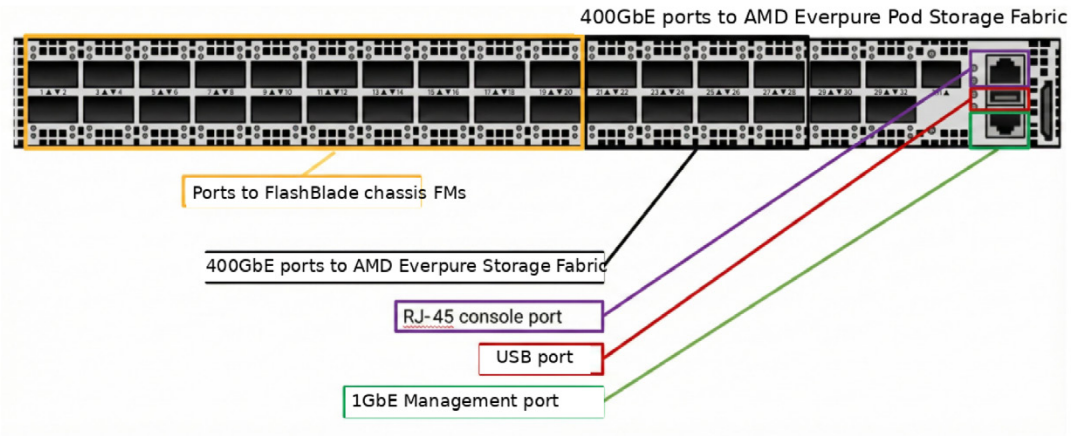

FIGURE 3 XFM8400 connection detail showing individual port assignments and labeled port groups for chassis blade connections.

XFM8400 cabling guide

Connect cables from the XFM8400 modules to each chassis in the order specified in Figure 4. Follow the connection order exactly to ensure symmetric traffic distribution and full redundancy across both XFM8400 modules.

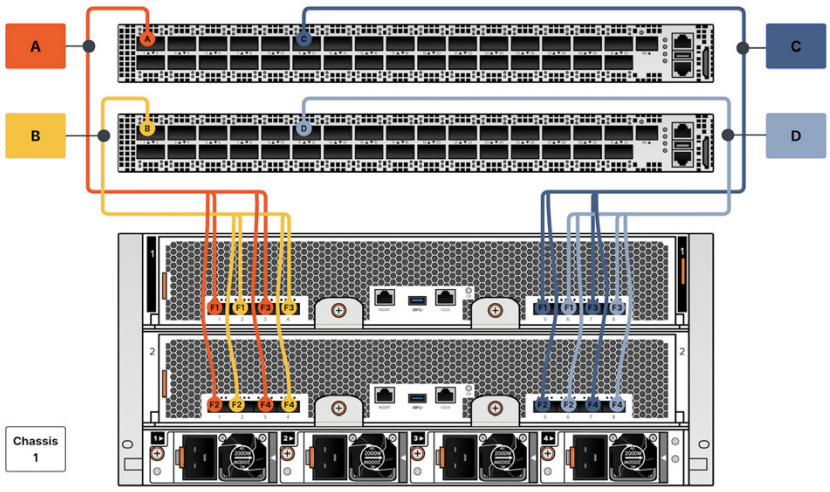

FIGURE 4 XFM8400-to-chassis cabling connection order. Refer to the FlashBlade multi-chassis installation guide for complete cabling specifications.

Note: Always refer to the current FlashBlade multi-chassis installation guide (see Additional resources) for the authoritative cabling order. Connection sequences vary by chassis count and XFM model revision.

Scalability and directional sizing guidance

Everpure AI factory with AMD scales from single-rack starter configurations to multi-rack, multi-petabyte deployments. FlashBlade capacity and performance scale independently as GPU cluster size grows. Tables 1 and 2 provide directional sizing guidance based on sample deployments for the AMD Instinct MI350 series across two performance tiers.

Note: All sizing data is directional guidance based on sample configurations. Contact your Everpure representative for a complete sizing exercise based on your specific workload profiles, redundancy requirements, and data protection policies.

FlashBlade//SR2 blade sizing guidance

Good performance tier

GPUs	Namespaces	Chassis/ namespace	Capacity (raw) TB	RU	Racks	Power (kW)	FlashBlade configuration
512	1	4	4,500	22	1	11.7	1x40x3x37.5
1,024	1	10	11,250	52	2	28.3	1x100x3x37.5
2,048	2	10	22,500	104	3	56.7	2x100x3x37.5
4,096	4	10	45,000	208	5	113.3	4x100x3x37.5

TABLE 1 Everpure AI factory with AMD—FlashBlade//SR2 directional sizing guidance, good performance tier. Configurations are scaled to AMD Instinct MI350X cluster sizes from 512 to 4,096 GPUs.

Better performance tier

GPUs	Namespaces	Chassis/ namespace	Capacity (raw) TB	RU	Racks	Power (kW)	FlashBlade configuration
512	1	10	11,250	52	2	28.3	1x100x3x37.5
1,024	2	10	22,500	104	3	56.7	2x100x3x37.5
2,048	4	10	45,000	208	5	113.3	4x100x3x37.5
4,096	8	10	90,000	416	10	226.7	8x100x3x37.5

TABLE 2 Everpure AI factory with AMD—FlashBlade//SR2 directional sizing guidance, better performance tier. Higher chassis density delivers two times the read throughput of the good performance tier at equivalent GPU counts.

Note: Power (kW) values are indicative. Contact your Everpure representative for final power budgeting and cooling requirements specific to your data center.

Storage performance guidance

AMD provides workload I/O guidelines for Everpure AI factory with AMD. The numbers in Table 3 represent directional best-case guidance and help establish the I/O load a single rack generates under full GPU utilization. Always benchmark and size performance and capacity to your organization's specific workloads.

Number of GPUs	Good—Read (GB/s)	Good—Write (GB/s)	Better—Read (GB/s)	Better—Write (GB/s)
512	256	128	512	256
1,024	512	256	1,024	512
2,048	1,024	512	2,048	1,024
4,096	2,048	1,024	4,096	2,048

TABLE 3 AMD Instinct MI350X storage performance guidance per GPU cluster size. Read and write throughput requirements for the good and better performance tiers are provided for 512–4,096 GPU configurations.

Performance design guidance

- **LLM training and checkpointing**: LLM training demands sustained, peak read and write throughput simultaneously. Allocate at least 50% of the recommended read performance as write performance for LLM or large-model workloads. FlashBlade//SR2 delivers symmetric read/write performance without performance degradation under mixed I/O profiles.
- **Computer vision datasets**: High-resolution computer vision datasets can exceed 30TB. These workloads require up to 4GB/s per GPU of sustained read performance. The better performance tier configuration is recommended for large-scale computer vision training.
- **Inference serving**: AMD Instinct MI350X native FP4/FP6 support drives 35 times inference throughput gains over the MI300 series. Storage I/O for inference is typically read-heavy, so the good performance tier is sufficient for most inference-serving deployments.
- **Checkpointing strategy**: For models above 100 billion parameters, implement frequent checkpointing to FlashBlade//SR2 to minimize training recovery time. FlashBlade symmetric throughput ensures checkpoint writes do not stall training forward passes.

Note: Benchmark your specific workloads before finalizing configuration. AI workload I/O profiles vary significantly by model architecture, batch size, and parallelism strategy.

Summary

The Everpure AI factory with AMD solution delivers an enterprise-grade AI infrastructure platform that pairs the raw compute power of AMD Instinct MI350 series accelerators with the industry-leading storage performance and simplicity of Everpure FlashBlade//SR2.

FlashBlade was purpose-built for flash and optimized to sustain the high-throughput, low-latency I/O demands of large-scale AI initiatives. Its modular architecture scales capacity and performance independently, and it continuously improves through Evergreen nondisruptive upgrades without planned downtime or forklift replacements.

Together, Everpure and AMD deliver an AI factory that is:

- **Performant**: Up to 4,096GB/s aggregate read throughput matched to AMD MI350X workload profiles across 4,096 GPU clusters
- **Scalable**: From 512 to 4,096+ GPUs and from single-chassis to multi-petabyte storage without architectural changes
- **Simple**: Single integrated solution for all AI pipeline stages—ingest, training, checkpointing, and inference—that is easy to deploy, manage, and scale
- **Energy-efficient**: High-density DirectFlash architecture with unmatched data center efficiency in an environmental, social, and governance–friendly AI infrastructure
- **Future-proof**: Built for the next generation of AMD and Everpure innovations as AI model sizes and computational requirements continue to evolve

Additional resources

To learn more about Everpure AI factory with AMD and related solutions, explore:

- FlashBlade//S
- Everpure Platform for AI
- AMD Instinct MI350 series
- AMD ROCm open software platform

Visit Our Website | **800.379.PURE**

ID26973-01-en-06/26