WHITE PAPER

Performance and Scale for Modern Database Workloads

Accelerate OLTP, OLAP, and AI workloads with FlashArray//XL to drive workload modernization.

Contents

Executive Summary

Enterprises are under intense pressure to deliver near-instant insights while lowering cost and complexity. Slow analytics at quarter-close can delay revenue recognition; an unplanned failover that stretches beyond minutes can cost millions in lost transactions. To keep these high-stakes moments from becoming business-critical fire drills, the underlying storage platform must do more than perform. It must simplify operations, protect data, and scale cleanly as workloads expand.

FlashArray//XL™ answers these needs. By collapsing OLTP, OLAP, and AI workloads onto one high-performance storage tier, it shrinks hardware footprints, cuts operational overhead, and keeps latency within microseconds even at peak concurrency. Built-in replication, snapshots, and encryption align storage with business KPIs, delivering faster time to insight and a measurable return on investment.

- Run more workloads on fewer systems without performance tuning
- Accelerate batch jobs, reporting pipelines, and AI queries
- Reduce administrative effort through a single storage pool for OLTP, OLAP, and vector search
- Future-proof capacity and performance for emerging workload demands
- Lower TCO by consolidating diverse workloads

The Platform Era: Converging Storage and Workloads

Modern data strategy is shifting from siloed systems toward converged architectures where transactional, analytical, and AI workloads share the same infrastructure. This shift is driven by the need for faster insight, simplified governance, and tighter cost control.

Workload convergence is inevitable: A single tier of high-performance storage removes the delays caused by data movement and duplicated copies. As regulations tighten and budgets contract, enterprises see convergence as the fastest route to compliance, efficiency, and agility.

A chorus of technology leaders is signaling that convergence is not optional:

- Thomas Kurian (CEO, Google Cloud) emphasized in an investor call in 2023 the need for hybrid storage architectures that unify structured and unstructured data for analytics and AI. **“We offer a single system to analyze structured and unstructured data... a data warehouse and a data lake in one.”**
- Larry Ellison (CTO, Oracle) highlighted database engine convergence at Oracle CloudWorld in 2023, integrating vector search for AI directly into databases. **“The best way to specialize AI models is to put supplemental training data in an Oracle vector database. And that’s what we built.”**
- Satya Nadella (CEO, Microsoft) stressed at Build 2023 workload convergence, where SQL analytics and AI processing run on a unified infrastructure. **“It [Microsoft Fabric] unifies all analytics workloads—SQL, machine learning—on the same compute infrastructure… fueling the next generation of AI applications.”**
- Frank Slootman (CEO, Snowflake) described functional convergence at Snowflake Summit in 2023, where modern database platforms support transactional, analytical, and AI workloads in a single system. **“Warehousing doesn’t define us anymore. It is just a workload type now.”**

These statements underscore an industry consensus: the future belongs to unified platforms that collapse silos and deliver speed at scale.

Rising Pressure on Storage Infrastructure

The 2025 State of the Database Landscape Report by Redgate reveals a pivotal shift in database strategy. Organizations are rethinking cloud adoption, consolidating database platforms, and prioritizing performance, cost-efficiency, and data security—all of which are directly impacted by storage decisions.

Trend	Key Data Point	Implications for Administrators
Hybrid Storage Strategies	30% fully in the cloud; 46% favor a hybrid approach	**Storage admins:** Manage on-prem arrays that can burst capacity or replicate to the cloud. **DBAs:** May need to move mission-critical workloads back on-prem to maintain consistent performance. With Pure1® and Pure Fusion™ automation, capacity planning and deployment become simpler by eliminating pre-planning.
Rising On-premises Retention	On-prem hosting up from 31% (2023) to 34% (2024)	**DBAs:** Must optimize local systems for performance workloads. **Storage admins:** Ensure consistent SLAs with reliable on-prem infrastructure. FlashArray//XL™ AI-driven monitoring and smart recommendations for workload placement reduces overhead.
Cloud Challenges	63% cite cost management, 40% performance issues, 37% inefficiencies	By reducing cloud-related tuning and troubleshooting and native support for hybrid deployment engineers can spend more time innovating. Automation, monitoring, and management tools like Pure1 proactively handle bottlenecks and capacity expansions, assist in troubleshooting, and can ensure compliance across large deployments with reduced risk for human error.
Database Platform Consolidation	74% reducing to three or fewer platforms	**DBAs:** Reduced complexity but expect higher performance demands. **Storage admins:** Manage fewer arrays handling mixed workloads. Consolidation on FlashArray//XL lowers TCO by minimizing the total hardware footprint.

A Storage-first Design Principle for Modern Data Architectures

Modern data architecture for databases is evolving rapidly. Structured workloads like OLTP, OLAP, and vector similarity search are increasingly expected to coexist, interact in real time, and operate across hybrid environments.

Adopting a storage-first design mindset doesn't mean buying infrastructure before software—it means recognizing that the success of the software stack depends on a foundation that can deliver both performance and business-centric benefits, including:

- Consistent microsecond latency for mixed workloads
- Fewer arrays, less sprawl, and lower power and space costs
- Real-time access to live data without duplication
- Simplified operations across on-prem and cloud with integrated replication and snapshots

Storage sits at the base of the modern data architecture—underneath ingest pipelines, processing engines, metadata services, and user-facing analytics. When that foundation is fast, unified, resilient, and cost-efficient, it removes downstream bottlenecks, mitigates the risks associated with multiple siloed systems, and unlocks the full value of the platform above it.

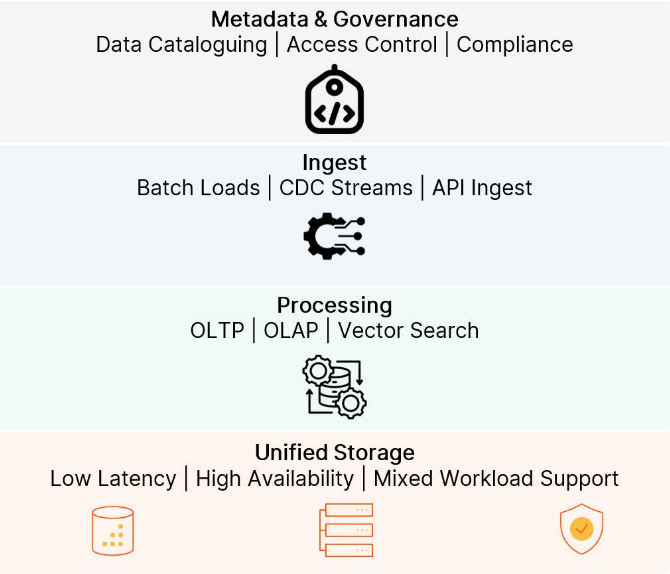
Metadata & Governance
Data Cataloguing | Access Control | Compliance
Ingest
Batch Loads | CDC Streams | API Ingest
Processing
OLTP | OLAP | Vector Search
Unified Storage
Low Latency | High Availability | Mixed Workload Support

FIGURE 1 The storage-first modern data architecture

For database teams, this approach results in fewer systems to manage, more predictable performance, and a smoother path to scaling structured workloads in any direction—all while optimizing resource usage and containing operational costs.

Evolving Workloads: Convergence and Consolidation for the Modern Database

Modern enterprises increasingly run transactional, analytical, and AI workloads side by side—and often against the same data sets. Each workload stresses storage differently, yet business leaders expect a single service-level envelope with no compromises. This section explains the unique demands of each workload and shows how FlashArray//XL satisfies them simultaneously.

Workload Characteristics and Requirements

Workload	Characteristics	Requirements
OLTP (Online Transaction Processing)	High concurrency; small, random reads/writes	Microsecond latency; transactional integrity
OLAP (Online Analytical Processing)	Large-scale scans; parallel queries	High sustained throughput; wide parallelism
Vector Search & AI Workloads	Data footprint, similarity look-ups; index updates	Rapid random I/O; concurrency at scale

Why It Matters

- OLTP latency spikes lead directly to abandoned shopping carts, failed trades, or delayed reservations.
- Slow OLAP jobs push reporting past executive deadlines, eroding confidence in the data.
- Vector search underpins real-time product recommendations and anomaly detection—both intolerant of I/O stalls.

Building an On-premises Platform Your Data Teams Will Love

Cloud services have taught developers to expect self-service speed and instant scale, yet many data-intensive workloads still run best in the data center. A converged, cloud-like stack on premises lets you give developers the agility they want while you keep control of governance, cost, and performance. Pure Storage FlashArray provides the foundation, delivering 99.9999% measured availability so your most critical data stays online even during upgrades.

Why Convergence Matters Now

Ignoring convergence risks the following:

- **Regulatory exposure** when sensitive data leaves approved environments
- **Unpredictable latency** that slows transactions, analytics, and AI training
- **Runaway cloud spend** from variable egress and burst fees
- **Fragile integrations** between legacy systems and cloud services
- **Talent drain** as engineers fight siloed tooling instead of building new features

Competitors that converge their stack remove these roadblocks, compress project timelines, and reach insight faster. Falling behind is a real possibility.

The Four Layers of Convergence in Structured-Data Architecture

Convergence does not mean forcing every workload into one monolith. It means reducing fragmentation at each layer so you can simplify operations, cut cost, and adapt quickly.

Layer	What to Converge	Key Benefits
1. Storage	Consolidate OLTP, analytics, and AI data on a single high-performance FlashArray tier.	Fewer arrays to buy and manage; uniform snapshots, replication, and encryption; lower TCO.
2. Database Engines	Choose engines that support multiple models (row, column, vector) instead of a separate product for every workload.	Less patching and licensing; minimal data copies; predictable tuning across OLTP, OLAP, and AI.
3. Workloads	Run OLTP, analytics, and AI on common storage and, where practical, within the same engine family.	Real-time insight from live transactional data; fewer brittle ETL jobs; unified backup and security.
4. Data Integration	Offer SQL, NoSQL, vector, and streaming APIs through a standard gateway or services layer.	Faster app launches using approved interfaces; consistent governance; simple integration for new tools.

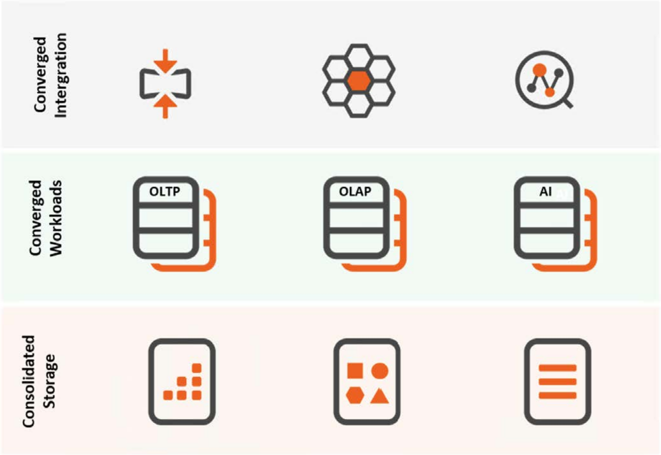

FIGURE 2 Data management convergence

FlashArray//XL—the Unified Platform for Modern Workloads

Modern application landscapes combine high-volume transactions, real-time analytics, and AI techniques such as vector similarity search on the same data sets. Running these workloads on separate arrays multiplies data copies, introduces performance variability, and increases operating expense. FlashArray//XL aligns with the four-layer convergence model by providing a single, highly available storage foundation capable of serving all workload types at enterprise scale.

High-performance in an efficient footprint

- Millions of transactions per minutes
- < 150 microsecond latency
- Up to 7.4 PB effective capacity in a 5U chassis

Simplified, cost-efficient data protection

- Built-in snapshots & always-on encryption
- Active-Active replication
- Average 5:1 Data reduction ratio

Always-on Data Services

- 99.9999% measured availability
- No disruption during controller upgrades
- Seamless Software refreshes

Fast AI vector insights without GPU silos

- Early lab validation complete
- Sub-second vector search performance
- Tested on multi-terabyte datasets

FIGURE 3 FlashArray//XL is the unified platform for modern workloads

FlashArray//XL Capability	Impact on Modern Workloads
Up to 7.4PB effective capacity in a 5U chassis, throughput to 45 GB/s, and latency as low as 150 µs	Large structured databases, column stores, and multi-terabyte embedding sets remain in one place without performance isolation issues
99.9999% measured availability, maintained during controller upgrades and software refreshes	Transactional, analytical, and inference services remain online without scheduled downtime
Built-in snapshots, active-active replication, encryption, and an average 5:1 data-reduction ratio	Uniform data-protection controls reduce administrative overhead while lowering effective cost per terabyte
Early lab validation of sub-second vector search on multi-terabyte data sets	AI-driven search and recommendation engines operate beside core databases without dedicated GPU tiers

Performance Characterization and Benchmark Results

This performance characterization quantifies the benefits of modern workloads using benchmark results for OLTP, OLAP, and vector similarity search workloads on FlashArray//XL170. Comparisons are compiled using the same test harness against a previous generation model (FlashArray//XLR1, launched in 2021) against a next-generation model (FlashArray//XLR5, launched in 2025). Characterization and benchmarking is run in an on-premises data center to mimic the conditions of production deployments.

Note: Real-world results vary with schema design, data volume, and concurrency. Customers should run a proof of concept that mirrors their own workload mix.

HammerDB Benchmark Suite

HammerDB is an industry-standard, open source database benchmarking tool designed to simulate realistic database workloads, measuring the performance of transactional (OLTP) and analytical (OLAP) databases. It provides standardized benchmarks such as TPROC-C and TPROC-H, widely recognized for evaluating database system performance and scalability.

VectorDBBench

VectorDBBench is an open source benchmarking tool designed to compare the performance and cost-effectiveness of leading vector databases and cloud services. It provides an intuitive interface for both experts and non-professionals to run benchmarks, reproduce results, and evaluate systems with ease.

With real-world testing scenarios like data insertion, vector search, and filtered search, VectorDBBench reflects production-like workloads using public data sets such as SIFT, GIST, Cohere, and OpenAI-sourced data. It also includes cost-effectiveness reports for cloud services, helping users make practical and informed decisions.

Workloads, Methodology, and Results

Transaction Processing Workload (TPROC-C)

The transaction processing characterization testing used the following environment:

Component	Detail
Compute	**Specification:** 8 x Cisco UCSC-C220-M5SX servers with Intel® Xeon(R) Platinum 8160 CPU @ 2.10GHz and 512GB memory **Operating system:** Windows Server 2025 Datacenter **Networking:** 2 × 25GB ethernet ports in 802.3ad (LACP) **Fiber channel:** Emulex LightPulse LPe31000/LPe32000 (2 × 32GB FC ports)
Connectivity	**Fiber channel:** Cisco MDS 9396T 32-Gbps 96-Port Fibre Channel Switch **Ethernet:** N9K-C9336C-FX2
Storage	**Previous generation model:** FlashArray//XL170 R1 with 16 FC ports (8 per controller) **Next-gen model:** FlashArray//XL170 R5 with 16 FC ports (8 per controller) **Storage configuration:** 4 data packs – 40 × 4.56T Direct Flash Modules **Purity OS version:** Purity//FA 6.8.7
Microsoft SQL Server	SQL Server 2022 Enterprise Developer Edition **TPCROC-C database:** 4 data files (128GB each) and 1 log file, each within its own volume **Max degree of parallelism:** 8 **Cost threshold for parallelism:** 50 **Memory limits (TPROC-C):** 64GB **Memory limits (TPROC-H):** 100GB
HammerDB	Version 4.12

To evaluate storage bottlenecks under intense transactional workloads, eight separate SQL Server database instances were deployed, each running on its own physical host with a dedicated operating system, SQL Server instance, and near-identical database configuration. Each database was initialized with a unique TPROC-C data set, followed by index rebuilds to ensure optimal baseline performance. Testing was then repeated on the different storage platforms.

All workloads were executed in parallel to simulate a highly consolidated, high-intensity OLTP environment.
This approach aimed to stress the storage system and highlight performance limits under concurrent pressure.

The primary performance metric was New Orders per Minute (NOPM)—a key indicator for OLTP workloads. Databases were first built and backed up, then restored to the specific storage configuration being tested.

The following configuration was used for the transaction processing workload:

- Number of warehouses: 5,000
- Virtual users: 200
- User delay (ms): 500
- Repeat delay (ms): 500
- Iterations: 1
- TPROC-C driver script: Timed Driver Script
- Total transactions per user: 10,000,000
- Checkpoint when complete
- Minutes of ramp-up time: 5
- Minutes for test duration: 30
- Use all warehouses
- Time profile

Results and Analysis

In the TPROC-C transaction processing benchmark, the FlashArray//XL170-R5 achieved a peak throughput of 2,128,331 NOPM, delivering a **100%** improvement over the FlashArray//XL170-R1, which reached 1,062,841 NOPM under the same workload and configuration.

This result highlights the architectural enhancements in the XL170-R5, which provides significantly greater throughput capacity and I/O handling compared to the previous generation. Both systems were tested using eight parallel SQL Server database instances with identical TPROC-C configurations, allowing for a fair comparison focused solely on storage performance.

The XL170-R5's ability to maintain high transactional throughput under a heavily consolidated OLTP workload demonstrates its value for enterprise customers looking to scale database performance without redesigning applications or increasing server resources.

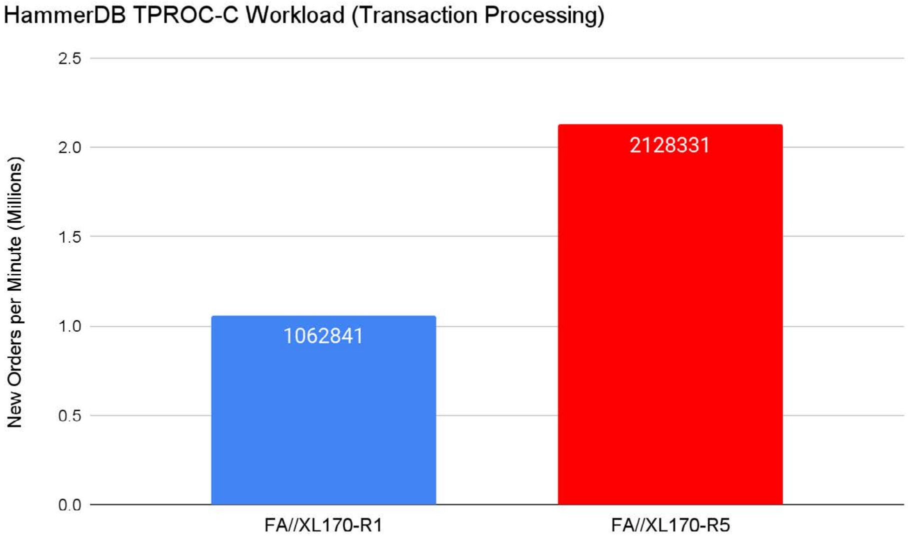

FIGURE 4 TPROC-C performance results comparing FlashArray// XL170-R1 and XL170-R5

Importantly, transaction processing is an IOPs intensive workload and while both the R1 and R5 models occupy the same physical rack space, the XL170-R5 delivers double the transactional performance without increasing data center footprint. The result is a two times increase in performance density, enabling customers to consolidate more workloads per rack unit and improve overall infrastructure efficiency.

Note: When running transaction processing characterization it was observed that write IOP size was 9Kb and read IOP size was 8Kb.

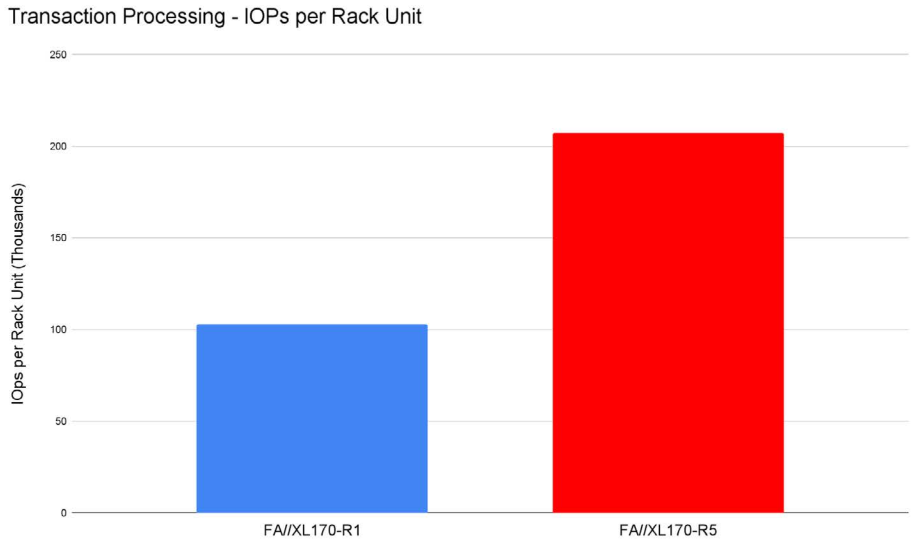

FIGURE 5 IOPs per rack unit achieved while running the TPROC-C benchmark

Decision Support Workload (TPROC-H)

The decision support characterization testing used the following environment:

Component	Detail
Compute	**Specification:** Cisco UCS C480-M5 running 4.3(2.240053) with 4 x Intel® Xeon® Platinum 8276M CPU @ 2.20GHz and 512GB memory **Operating system:** Windows Server 2025 Datacenter **Networking:** 2 × 100GB ethernet ports in 802.3ad (LACP) **Fiber channel:** 2 x Emulex LightPulse LPe31000/LPe32000 (8 × 32GB FC ports)
Connectivity	**Fiber channel:** Cisco MDS 9396T 32-Gbps 96-Port Fibre Channel Switch **Ethernet:** N9K-C9336C-FX2
Storage	**Previous generation model:** FlashArray//XL170 R1 with 16 FC ports (8 per controller) **Next-gen model:** FlashArray//XL170 R5 with 16 FC ports (8 per controller) **Storage configuration:** 4 data packs – 40 × 4.56T Direct Flash Modules **Purity OS version:** Purity//FA 6.8.7
Microsoft SQL Server	SQL Server 2022 Enterprise Developer Edition **TPCROC-H database:** 4 data files (256GB each) and 1 log file, each within its own volume **Max degree of parallelism:** 8 **Cost threshold for parallelism:** 50 **Memory limits (TPROC-C):** 64GB **Memory limits (TPROC-H):** 100GB
HammerDB	Version 4.12

This workload simulates a complex analytical and decision-support environment, representative of OLAP workloads common in data warehousing scenarios. The comparative metric used is Queries per Minute (QPM). The database schema is generated and loaded at a defined scale factor, backed up, and subsequently restored onto the evaluated storage solutions before executing the benchmark queries.

The following configuration was used for the analytical processing workload:

- Scale factor: 1000
- MAXDOP: 8
- Clustered ColumnStore
- Virtual users: 1
- Total query sets per user: 1
- Iterations: 1 full run of the query set
- Execution profile: Sequential execution of queries

Results and Analysis

In the TPC-H analytical processing benchmark, the FlashArray//XL170-R5 achieved a throughput of 23.01 QPM, significantly surpassing the FlashArray//XL170-R1, which delivered 18.71 QPM. This result represents approximately a 23% improvement in analytical query performance with the R5 model.

The results demonstrate the FlashArray//XL170-R5's ability to handle intensive analytical workloads with greater efficiency. Organizations running SQL Server decision-support workloads can expect improved query responsiveness and reduced processing times compared to previous-generation platforms.

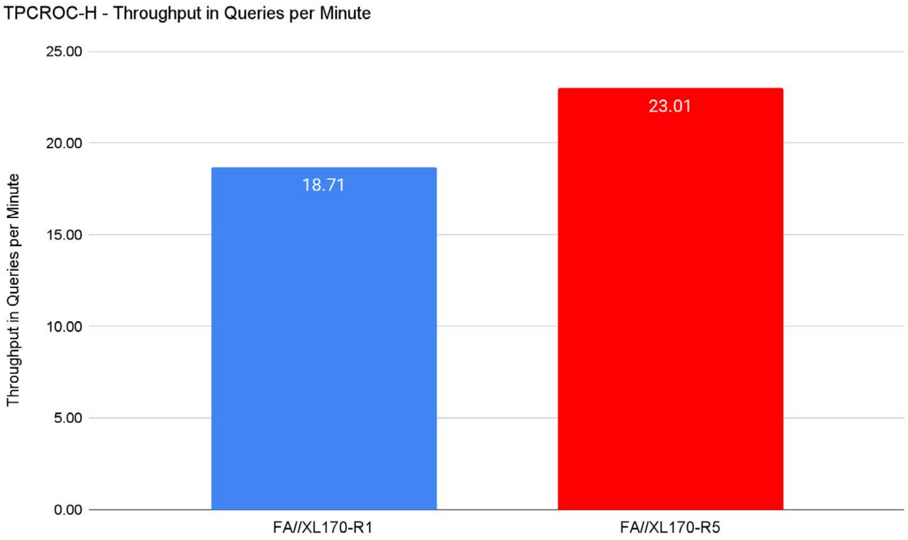

FIGURE 6 TPROC-H performance results comparing FlashArray//XL170-R1 and XL170-R5

Vector Similarity Search

Vector search characterization testing used the following environment:

Component	Detail
Compute	**Specification:** Cisco UCS C480-M5 running 4.3(2.240053) with 4 x Intel® Xeon® Platinum 8276M CPU @ 2.20GHz and 512GB memory **Operating system:** Red Hat Enterprise Linux release 9.5 **Networking:** 2 × 100GB ethernet ports in 802.3ad (LACP) **Fiber channel:** 2 x Emulex LightPulse LPe31000/LPe32000 (8 × 32GB FC ports)
Connectivity	**Fiber channel:** Cisco MDS 9396T 32-Gbps 96-Port Fibre Channel Switch **Ethernet:** N9K-C9336C-FX2
Storage	**Next-gen model:** FlashArray//XL 170R5 with 16 FC ports (8 per controller) **Storage configuration:** 4 data packs – 40 × 4.56T Direct Flash Modules **Purity OS version:** Purity//FA 6.8.7
PostgreSQL	PostgreSQL Server Version 17.4 PGVectorScale extension version 0.7.1 PGVector extension version 0.8.0
VectorDBBench	Version: 0.0.25
Data Set	The Cohere/wikipedia-2023-11-embed-multilingual-v3 data set on Hugging Face contains approximately 250 million paragraph embeddings from the 2023-11-01 Wikimedia Wikipedia dump, covering 300+ languages. Each embedding is generated by the Cohere Embed V3 multilingual model and has 1024 dimensions. The dataset is stored in compressed Parquet format, with a raw size of 536GB.

Results and Analysis

High-dimensional vector embeddings can occupy substantial storage space. Applying data reduction techniques helps manage capacity requirements and maintain operational efficiency. The table below summarizes the impact of FlashArray//XL170-R5 data reduction on a 536GB vector data set.

Metric	Footprint
Logical Data Footprint Post Indexing	1.5TB
Physical Data Footprint	670GB
Data Reduction Ratio	2.4:1
Footprint Reduction	56.77%

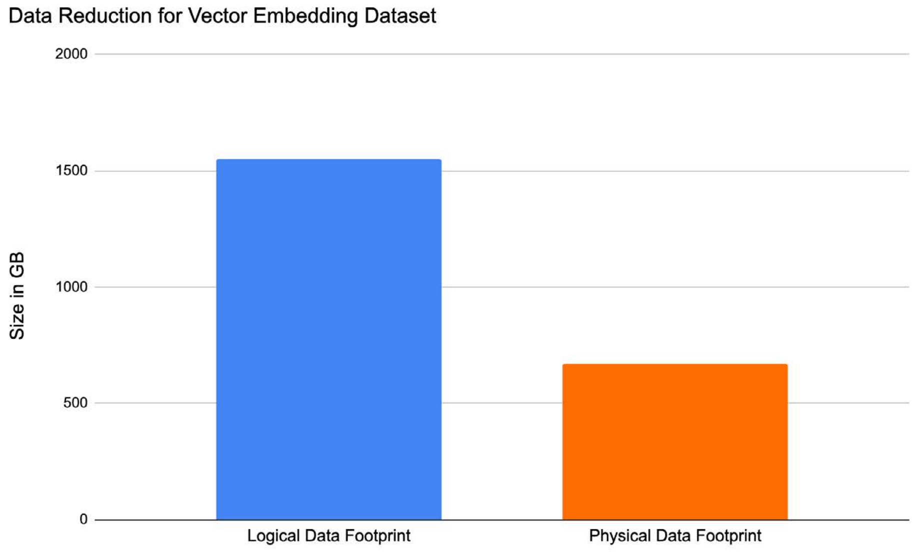

FIGURE 7 Vector embedding storage footprint before and after data reduction

FlashArray//XL170-R5 applies inline compression and deduplication to reduce the data set's physical footprint by over half. A 2.4 to 1 data reduction ratio means that 2.4GB of logical data requires only 1GB of physical storage. This reduction results in a 56.77% decrease in capacity consumption.

Key considerations:

- **Logical data footprint** reflects the total size of raw vector embeddings, indexes, and metadata before reduction.
- **Physical data footprint** reflects the actual on-disk usage after data reduction is applied by the array.
- **Reduction impact**: Less physical capacity is needed to support large-scale vector workloads, which can help reduce hardware requirements, improve rack density, and simplify data protection operations like snapshots and replication.

These results highlight how FlashArray//XL170-R5 can improve storage efficiency in environments that rely on vector search and indexing.

Customer Story: IAA Puts Its Vehicle Marketplace in the Fast Lane

IAA facilitates the sale of total-loss, damaged, and low-value vehicles.

Its 24×7 digital marketplace makes auto auctions fast, reliable, and simple for buyers and sellers around the world. Every year, buyers and sellers trust IAA's automotive marketplace to facilitate and process millions of sales. Storage uptime is crucial to support mission-critical processes like vehicle shipments, document creation, payments, and more. IAA's acquisition by heavy equipment auctioneer Ritchie Bros., and the subsequent creation of their new parent company—leading, omnichannel marketplace RB Global, Inc.—also meant its infrastructure needed to fuel sustainable growth on a whole new scale. Additionally, slow processes and response times with their existing storage system had begun to affect their digital customer platform. Pure Storage FlashArray//XL provided the performance and headroom IAA needed to scale its operation at pace, especially with the continuously upgraded, non-disruptive Evergreen® architecture. IAA now can scale at pace, with non-disruptive upgrades and reliable storage that delivers 24X7 services to every customer. The results have been transformative: Data processing power has increased 33% while their data center footprint has shrunk 97%. The simplicity of managing Pure Storage gives the IT team more bandwidth to focus on improving the customer experience.

Conclusion

The needs of modern database workloads have exposed the limits of legacy storage. FlashArray//XL is built for what current database environments demand and well architected to support future growth and innovation. Designed for consolidation, it combines unmatched performance density, simplified management, and enterprise-grade resilience to handle the scale and complexity of OLTP, OLAP, and AI-driven workloads.

Explore how FlashArray//XL can future-proof your organization's database storage needs.

Schedule a demo to see it in action and discuss your business goals and technology needs.

purestorage.com

800.379.PURE

PURESTORAGE®