



# MILVUS WITH HPE ALLETRA STORAGE MP X10000



#### About Milvus

Milvus is a leading open-source vector database purpose-built for similarity search and AI-powered applications. It enables fast and accurate search across billions of high-dimensional vectors.

## Advanced vector search architecture for demanding AI workloads

Milvus is a powerful open-source vector database designed to support similarity search and AI-driven applications. It enables organizations to store, index, and query billions of high-dimensional vectors generated by AI models with optimized search performance. These vectors represent complex data types such as unstructured data to support various AI and analytics workloads. Milvus provides a modular, service-oriented architecture that enables scalable, elastic deployment and supports hybrid search across structured and unstructured data. It also offers multitenancy, data isolation, replication, and failover for high availability and secure, production-ready use. Deploying Milvus in a containerized environment with HPE Alletra Storage MP X10000 provides a modern, scalable, and high performance architecture for advanced vector search.

### The challenge

As AI adoption accelerates, organizations are generating and storing vast amounts of vector data derived from deep learning models. These vectors must be indexed and queried in real time to support AI applications. Each vector represents complex data types such as images, audio, video, and natural language, and is essential for powering AI, such as conversational AI recommendation systems, semantic search engines, and anomaly detection platforms.

As AI workloads scale, so does the volume of vector data. Conventional databases and storage architectures are not engineered to handle the high-throughput, low-latency demands of vector-based AI workloads. Their limitations in indexing, parallelism, and real-time inference make them unsuitable for large-scale, performance-critical applications. As AI systems grow in complexity and scale, the underlying infrastructure must evolve to accommodate increasing data volumes and more sophisticated query patterns, demanding architectures that inherently support elasticity and high performance.

## The solution

Deploying Milvus with HPE Alletra Storage MP X10000 addresses these challenges with a cloud-native, high performance architecture. Milvus provides distributed indexing and search capabilities, leveraging GPU acceleration and advanced indexing algorithms to deliver fast and accurate results. Kubernetes orchestrates the deployment, enabling elastic scaling, high availability, and simplified management across hybrid or multicloud environments. HPE Alletra Storage MP X10000 serves as the storage foundation for Milvus, offering ultrafast object storage with high throughput, low latency, and enterprise-grade durability.

As depicted in Figure 1, this integrated solution allows organizations to offload vector data to HPE Alletra Storage MP X10000 object storage, reducing the need for expensive local disks and minimizing infrastructure overhead. Milvus disaggregated compute and storage architecture allows storage to scale independently of compute, giving AI teams the flexibility to handle unpredictable data growth or experimental scaling. HPE Alletra Storage MP X10000 dynamic data placement and disklet-based architecture intelligently distributes data across controllers, helping ensure consistent performance even as demand scales, lowering the need for manual data rebalancing or overprovisioning. This flexibility is critical for AI workloads that experience unpredictable growth or require rapid experimentation.

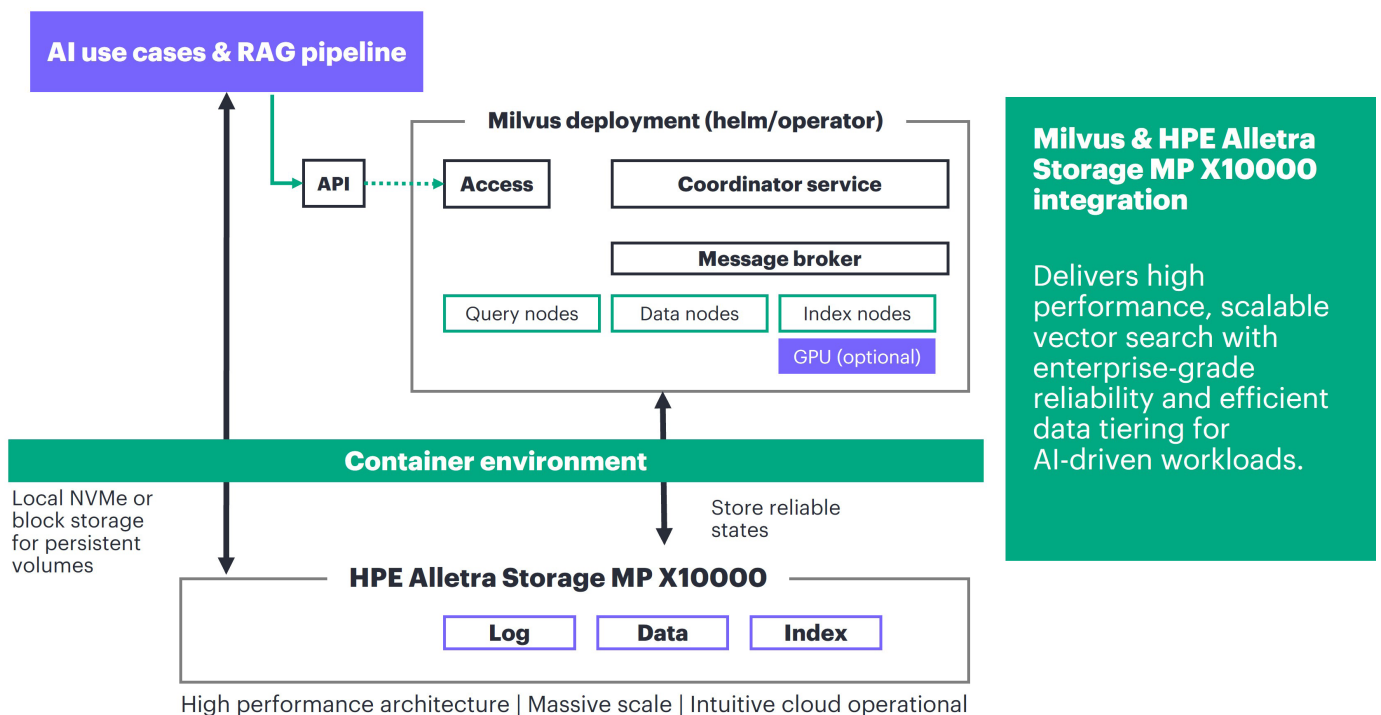


Figure 1. Solution diagram, deploying Milvus with HPE Alletra Storage MP X10000

Performance is a key differentiator in this solution architecture. The HPE Alletra Storage MP X10000 log-structured key-value engine enables high IOPS for small metadata operations and massive throughput for vector segment access, keeping Milvus query latencies low even at scale. This architecture supports the unique I/O profile of vector databases better than traditional object stores. Combined with support for GPU Direct Remote Direct Memory Access (RDMA), it ensures fast, deterministic response times for real-time inference or semantic search workloads.

Milvus can maintain sub-second query latency even when operating on massive datasets. This ensures that AI applications remain responsive and reliable, whether they are serving real-time recommendations or conducting large-scale semantic searches. The cloud-native design of both Milvus and HPE Alletra enables seamless integration with HPE GreenLake, providing unified management, monitoring, and consumption-based pricing. In addition to deployment on standalone Kubernetes, it can also be seamlessly integrated into the HPE Private Cloud AI solution stack for a more comprehensive and streamlined implementation.

Security and compliance are also built into the solution. HPE Alletra offers enterprise-grade data protection, encryption, and lifecycle management, helping organizations meet regulatory requirements and safeguard sensitive AI data. Combined with Kubernetes role-based access controls and Milvus support for multitenancy, the architecture is well-suited for enterprise environments. Seamless integration with HPE GreenLake and HPE Private Cloud AI brings unified management, monitoring, and consumption-based pricing, making the deployment scalable, enterprise ready, and cloud optimized.

## Benefits and value proposition

Milvus, in combination with HPE Alletra Storage MP X10000, provides a future-ready platform for AI applications. This solution empowers organizations to build intelligent systems that are fast, scalable, and cost-effective, helping ensure that vector data remains accessible, actionable, and secure at every stage of the AI lifecycle.

- **Low-latency access:** Milvus can retrieve vector data quickly, even from cold storage tiers, ensuring fast query response times across all data.
- **High throughput:** Supports concurrent access to large volumes of vector data, which is critical for real-time AI inference and analytics.
- **Elastic scalability:** Object storage decouples compute from storage, allowing Milvus to scale storage independently as vector datasets grow.
- **Cost efficiency:** Cold or infrequently accessed vectors can be offloaded to object storage, reducing the need for expensive local SSDs while maintaining accessibility.

- **Cloud-native integration:** HPE Alletra Storage MP X10000 cloud-native design fits seamlessly into Kubernetes environments, aligning with Milvus containerized architecture.

- **Data durability and lifecycle management:** Enterprise-grade features ensure that vector data is protected, compliant, and managed efficiently over time.

Milvus benefits from fast object storage by gaining speed, scalability, and cost-efficiency—all essential for building AI applications that rely on large-scale vector search. Hewlett Packard Enterprise offers a comprehensive full-stack solution—from optimized AI compute with HPE Private Cloud AI to high performance object storage, seamlessly integrated with HPE GreenLake flexible, cloud-native consumption models. This provides a robust, feature-proven architecture that streamlines and accelerates enterprise AI deployment.





Visit [HPE.com](https://www.hpe.com)

## Learn more at

[HPE Alletra Storage MP X10000](#)

[Introducing HPE Alletra Storage MP X10000 video](#)

[Top 10 reasons to choose HPE Alletra Storage MP X10000](#)

[Milvus webpage](#)

[Milvus Sizing Tool](#)

## [Chat now](#)

© Copyright 2025 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

a00149876ENW

HEWLETT PACKARD ENTERPRISE

[hpe.com](https://www.hpe.com)

