Lightbits Data Platform for AI and Machine Learning

Build Your AI Cloud with High-performance Software-defined Storage

PERFORMANCE AT SCALE

- Up to 75 million IOPS with sub millisecond tail latency
- Perfect for storage-intensive AI-oriented workloads such as large-scale vector and streaming databases
- Built on the power and simplicity of NVMe/TCP

COST EFFICIENCY

- Up to 80% lower TCO than DAS or SAN storage
- Scale compute and storage resources independently to maximize resource utilization
- Thin provisioning and compression to increase effective storage capacity

HIGH AVAILABILITY & RESILIENCY

- Clustered architecture provides high availability and dynamic resource scaling
- Synchronous replication within or across zones
- Built-in fast, efficient snapshots and database clones
- Multi-tenancy with Quality of Service assurances

Storage can either be an accelerator or a bottleneck for the AI pipeline. With the rapidly accelerating adoption of AI and machine learning, newly emerging workflows are demanding more performance at a greater scale than ever before. Many companies leveraging AI have discovered that file and object storage aren't enough—they need block storage too. They need high-performance block storage to support ever-growing vector, streaming and other databases, and to store and manage the operating system images for their private cloud and cloud services customers.

The Lightbits cloud data platform enables enterprises to build AI clouds with extreme performance at scale to capitalize on rapidly expanding AI opportunities. To accelerate AI pipelines, Lightbits' high-performance, scalable, cost-efficient block storage can streamline data pre-processing, boost model training, speed up real-time inference, and optimize retrieval-augmented generation. With Lightbits, you gain a flexible platform that scales with your AI cloud.

SCALE AI SERVICES WITH MAXIMUM PERFORMANCE

AI models and datasets are constantly growing and becoming more complex, so the underlying storage infrastructure must scale as well. The Lightbits data platform scales beyond the petabyte level and delivers performance of up to 75 million IOPS and consistent sub-millisecond tail latency even under a heavy load. This exceptional performance profile makes it the ideal solution for vector and other AI-oriented databases whether they manage real-time AI application data or store training parameters and tags.

Built by the inventors of the NVMe®/TCP protocol, Lightbits delivers this exceptional performance using standard TCP/IP networks and Ethernet NICs without requiring any configuration changes. No proprietary software is installed on client systems.

MAXIMIZE RESOURCES TO LOWER YOUR TOTAL COST OF OWNERSHIP

Unlike AI architectures built with local NVMe, the Lightbits clustered approach gives you the flexibility to provision any capacity you need to any server and to share data across servers. Where AI servers with DAS are often 15-25% utilized, Lightbits disaggregated storage allows you to scale performance and capacity independently and dynamically to maximize utilization.

To maximize your storage resources further, Lightbits optimizes SSD media through smart data placement, thin provisioning, and compression for up to 4:1 total data reduction. Lightbits also offers efficient thin snapshots and clones for data protection, so there's no need to copy data from local NVMe to external storage which can further reduce storage capacity requirements.

Lightbits' clustered architecture also avoids disruptions to AI pipelines or training if nodes or drives fail or become inaccessible. So, there's no need to handle storage failures by resuming from checkpoints.

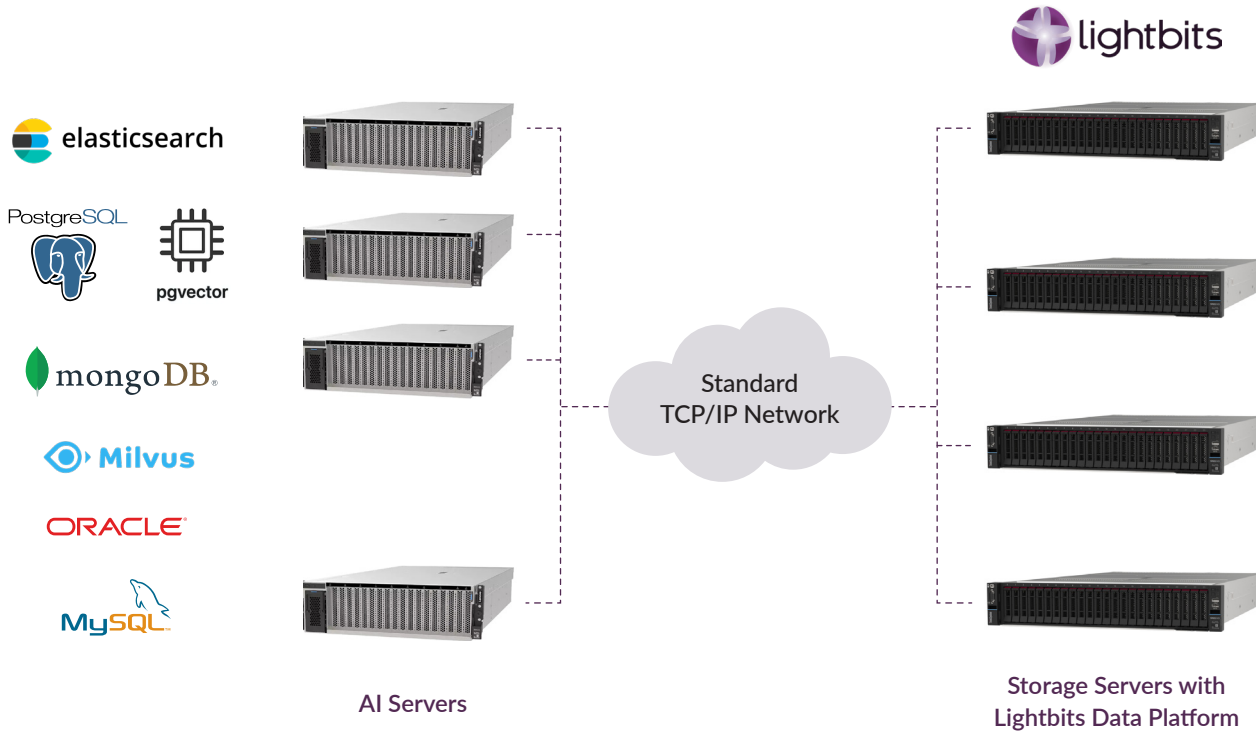

Figure 1: Typical AI deployment showing AI servers running databases and storage servers running Lightbits

OPTIMIZE A RANGE OF AI WORKFLOW PROCESSES

Streamline Data Preprocessing
Shorten the time required to clean and transform raw data into a format that can be more effectively utilized in machine learning models.

Accelerate Model Training
During the model training phase, data can be accessed faster, and model parameters can be continuously adjusted more frequently in a given time period.

Efficient Real-time Inference
Real-time AI requires high speed storage to make instant predictions for fraud detection, eCommerce personalization's, or autonomous vehicles reactions.

Retrieval-augmented Generation (RAG)
The vector databases used in LLMs require high-performance to return RAG-customized results for chatbots and other applications.

Fast, Secure Checkpointing
With Lightbits, your checkpoint data is persistent and protected without requiring a copy to a second storage tier. Lightbits' high write throughput ensures no GPU idle time.

1830 The Alameda, San Jose, CA 95126 E: info@lightbitslabs.com www.lightbitslabs.com