paradigm4

flexFS

flexFS™ AI Solutions Brief

Why File System Speed Matters More Than Ever in AI

Today's AI applications are fundamentally different from traditional software. They're **data-hungry**, **parallel by design**, and **performance-sensitive** in ways that expose every weakness in conventional storage systems.

AI workloads—from drug discovery and image analysis to autonomous systems and generative AI—depend on one critical factor that often goes overlooked: **how fast they can access their data.**

Storage is no longer a secondary concern. It's now a core pillar of cluster performance, efficiency, and reliability. When file systems become bottlenecks, AI applications underperform. I/O with higher latency and/or lower throughput increases training duration and degrades inference performance. GPU utilization drops as compute resources wait for data access. This leads to higher infrastructure costs, and lower training and inference throughput—translating to reduced ROI on AI investments.

Enter flexFS, the high performance file system for AI.

GPU Economics Make Storage Speed Critical

AI infrastructure is expensive. High-end GPUs can cost tens of thousands of dollars each, and cloud GPU instances often run $5-25 per hour or more for high-performance configurations. When GPUs wait for data, infrastructure costs increase due to longer training runtimes and slower inferencing. Organizations running large AI clusters quickly discover that storage bottlenecks slow down their applications' output and materially increase their costs—a perilous combination for AI business models.

flexFS Changes the Equation Entirely

flexFS is an elastic cloud file system purpose-built to **optimize the performance** of AI/ML and other high-performance computing (HPC) workloads by eliminating storage bottlenecks and keeping compute elements busy. flexFS delivers exceptional performance, scalability, and minimal operational overhead — **at a lower cost** than WEKA, DDN, and file systems offered by cloud providers.

Traditional cloud storage systems force complex, inter-related tradeoffs among two kinds of performance (throughput and latency) and three kinds of cost (storage, compute, and operational). On AWS, for example, Amazon's Elastic File System (EFS) offers business-computing level latency – but limited throughput and significant cost. Amazon's S3 offers virtually unlimited aggregate throughput at lower cost than EFS – but it also brings poor latency and a need to modify applications to deal with objects instead of files. If properly configured and managed, Amazon FSx for Lustre offers file storage with good latency and throughput – but it also brings higher operational cost and significantly higher storage cost.

flexFS eliminates such entangled tradeoffs, offering excellent latency and throughput performance – *and* lower storage, compute and operational costs.

[1]*This does not include "Intelligent Tiering" storage-class access costs, or operational costs required to configure, monitor and maintain FSx for Lustre volumes.*

[2]*FSx for Lustre recently added elastic capacity thought its "Intelligent Tiering" feature – but throughput must still be pre-provisioned. Moreover, a volume's throughput can only be increased, not decreased.*

[3]*The larger the volume of data, the greater the cost savings.*

flexFS offers:

10x better throughput than Amazon EFS for AI workloads

- Amazon EFS has a hard limit of 60 GiBps. flexFS, by contrast, benefits from S3's throughput scaling and elasticity.

10x more cost-effective than FSx for Lustre

- A 100TB FSx for Lustre volume capable of providing 400 GiBps throughput costs[1] well over 10x the cost of a flexFS volume supporting that level of throughput – or higher.[2]

Zero overprovisioning required—pay only for what you use

- flexFS only charges for capacity used – regardless how much throughput is needed.
- A standard, SSD-based FSx for Lustre volume requires provisioning enough capacity to meet throughput requirements – regardless how much space is used. An "Intelligent Tiering" FSx for Lustre volume charges only for space used but still requires throughput to be provisioned – and adds data access charges.

Instant elasticity — independently scale capacity and throughput – up or down – without manual intervention or downtime.

- You can increase the size of standard FSx for Lustre volumes – manually. They cannot, however, be decreased in size. Achieving a smaller volume requires creating a new smaller volume, moving the data, and deleting the old volume.
- "Intelligent Tiering" volumes have elastic capacity, but a fixed level of provisioned throughput – and with it, a fixed cost – both of which can only be increased. Decreasing provisioned throughput requires creating a new, lower-throughput volume, moving the data, and deleting the old volume.

40-60% cost reduction compared to traditional cloud file systems, while achieving significantly better performance.

- This cost efficiency transforms AI project economics, making it feasible to run larger experiments, iterate more frequently, and deploy more ambitious AI applications.[3]

ROI achieved within 2 months of deployment

- Shorter time-to-value getting started with flexFS, shorter time-to-results, lower storage, compute and operational costs all add up to a rapid return on investment.

Meeting the Challenges of Multi-omics, Biomedical Imaging, Structural Biology, and Clinical Data with AI at Scale

Efficiently managing the ever-growing scale and diversity of data is now fundamental to advancing scientific discovery through AI.

Developing a multimodal AI model to predict drug-target interactions can involve analyzing genomic sequences, protein structures, molecular imaging, and clinical trial data simultaneously. Genomic data alone can exceed 2.5 petabytes across 800 million files—from raw sequencing reads to annotated variants. Add high-resolution cryo-electron microscopy images of protein structures (400 TB of movies and images), molecular simulation trajectories (300TB of time-series data), and clinical imaging data (150TB of DICOM files), and most storage infrastructures can't keep pace.

flexFS Delivers

Protein Structure Prediction

AlphaFold predicts protein folding from sequence data input in FASTA format, using several large sequence and structure databases such as UniRef90, MGnify, BFD, and PDB70, for multiple sequence alignments (MSA) and template searches. These databases involve billions of small files and large datasets. Paradigm4's benchmarks of EFS, Lustre, and flexFS showed all three had comparable performance. Differences in runtimes were less than 2%. Cost, however, is a different story entirely:

flexFS provides 42% cost savings compared with EFS.[4]

flexFS saves 72% over FSx for Lustre.[4]

cryoEM analysis

While AI has streamlined and enhanced cryoEM workflows, I/O access patterns in analysis still present both throughput and latency challenges. flexFS demonstrates high performance on both dimensions at significantly lower costs. Benchmarks run on AWS measuring the performance of automated steps typical of single-particle analysis workflows revealed that:

flexFS provides 17% faster results and costs 72% less than FSx for Lustre.

flexFS provides 23% faster results and costs 43% less than EFS.

flexFS performs faster than both EFS and FSx for Lustre in each step of the cryo-EM workflow.

Moreover, cost relief with flexFS increases as data volume grows.

Looking forward, AI/ ML will be applied throughout the cryoEM analysis workflow to automate, accelerate, and enhance accuracy from particle-picking, denoising, preprocessing, and local resolution determination to model building and validation—all leveraging flexFS.

[4]Pricing for 100TB in the AWS us-east-1 region as of late April 2025.

How flexFS Enables Faster Performance

flexFS was designed from the ground up to solve the most demanding data intensive storage and high throughput challenges. Instead of accepting traditional trade-offs, we built an elastic cloud file system that delivers **exceptional performance**, **seamless scalability**, and **cost efficiency** specifically for demanding AI workloads.

flexFS leverages hyperscale object storage strengths and adds full POSIX file system support; low-latency metadata; tunable, low-latency file data; minimal operational overhead; and outstanding price-performance. The result is truly elastic file storage that outperforms conventional cloud file systems – and costs a whole lot less.

Full POSIX with Cloud-Native Performance

flexFS delivers complete POSIX compatibility without performance compromises. Extended attributes, access control lists, and Linux advisory locking work seamlessly — critical for bioinformatics workflows that depend on complex metadata and concurrent access patterns.

Eliminate Storage Bottlenecks

flexFS delivers the sustained high throughput and low latency that AI applications demand. Our architecture scales linearly with cluster size, supporting hundreds of GB/sec aggregate throughput across thousands of concurrent clients. Whether you're training computer vision models with millions of small image files, processing large video datasets, or running real-time and/or batch inference pipelines, flexFS keeps your AI applications running at full speed.

Seamless Integration with AI Frameworks

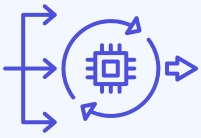
Full POSIX compliance means your existing AI applications work immediately with flexFS—no code changes, no API modifications, no developer retraining required. TensorFlow, PyTorch, Scikit-learn, and other popular frameworks integrate seamlessly, treating flexFS like any standard file system.

True Elastic Performance

Unlike traditional solutions, flexFS separates capacity from performance. Scale throughput independently of storage capacity, paying only for what you use. Need more performance and capacity for a large training run? Scale both up instantly and automatically. Finished training and moved to inference? Scale capacity back down instantly and automatically while maintaining I/O throughput and latency needed to shorten time-to-first-token and inter-token latency—ensuring smooth execution and maximizing overall system efficiency. Need to add 50,000 additional samples (adding 800TB overnight)? Scale it back up instantly and automatically. No manual provisioning, no performance degradation.

Elastic Architecture That Scales Independently

Unlike traditional systems, flexFS separates capacity from performance. Need more throughput for a training sprint? Scale performance up instantly without adding unnecessary storage. Dataset growing? Add capacity without performance penalties.



Massive Parallel Performance

AI training workloads often involve hundreds or thousands of compute nodes accessing data simultaneously. flexFS is architected for exactly this use case, delivering consistent performance even under massive parallel access patterns that bring other file systems to their knees.

Cost Efficiency That Transforms AI Economics

flexFS customers typically see 50% or greater cost reductions compared to traditional cloud file systems, while achieving significantly better performance. This cost efficiency transforms AI project economics, making it feasible to run larger experiments, iterate more frequently, and deploy more ambitious AI applications.

Ready to Accelerate Your AI Research?

Whether you're training multimodal models on genomic data, processing biomedical images at scale, or running complex simulations, flexFS provides the storage foundation and HPC throughput to transform your AI infrastructure from bottleneck to breakthrough.

LEARN MORE

Check out docs.flexfs.io and the flexFS Cryo-EM Benchmark

GET STARTED

Contact info@flexfs.io

MEASURE THE DIFFERENCE

Request a performance benchmark for your specific AI workloads

flexFS™ is a trademark of Paradigm4, Inc. All other trademarks are the property of their respective owners.