Transforming Drug Discovery Through Data-Centric Approach

Life Sciences
North America

RECURSION

- Artificial Intelligence and Deep Learning
- Train machine learning models for inference
- GPU Accelerated
- Large Amounts of Small Files

Solution

A 2PB high-performance, multi-tier data management infrastructure maximizes GPU compute resources for accelerated AI workflows

Kris Howard
Principal Systems Engineer

Recursion Pharmaceuticals

AI-driven solution exceeds performance of platforms used by largest pharmaceutical companies, allowing identification of drug types in weeks, not months or years.

Recursion Pharmaceutical is reinventing drug discovery through artificial intelligence (AI) with a solution that rivals, and exceeds, the processing platforms used by the largest pharmaceutical companies. The six-year old startup, based in Salt Lake City, executes about 350,000 experiments weekly and screens thousands of compounds against hundreds of disease models—at a fraction of the cost and time of traditional drug discovery methods.

According to Kris Howard, Principal Systems Engineer at Recursion Pharmaceutical, AI and machine learning play an important role in understanding human biology. "We leverage machine learning to ask complex questions about biology, and get faster answers," he explains. "This enables us to quantify and analyze hundreds of different features at the single cellular level to accelerate drug discovery efforts."

The Challenge

- Small files drag down analysis performance
- Traditional drug discovery process is long, expensive and prone to failure
- AI and ML can accelerate drug discovery, but require high-performance compute and storage resources
- Fully optimized storage infrastructure was needed to accelerate AI applications and support fast-growing pipeline

When new drugs are discovered, it can take up to 10 years of clinical trials before that drug ever makes it to the market—if it succeeds, since around nine out of 10 drug candidates fail to win FDA approval. With mounting drug-discovery costs and time-to-market challenges, Recursion Pharmaceutical devised an innovative data-first approach that dramatically decreases the costs and increases the efficiency of biological research.

"We do things a little differently," says Howard. "This means designing drug discovery workflows with data science in mind." The result is a high-throughput platform designed to unlock the maximum value of data from the world's largest repository of biological images.

In addition to ingesting massive amounts of cell image data, Recursion required high-performance drug discovery processing that was fully optimized for AI and machine learning. While another obstacle was large amounts of small file sizes, which created massive overhead for the company's metadata server and performance bottlenecks for critical workloads.

"Our data is our company, so we needed a robust storage architecture to support our AI-driven models," Howard adds. "Managing our at-scale data needs required faster ingest, optimized processing and reduced application run times."

The Solution

- High-performance hybrid data management which allows fast access to millions of small files and extended attributes while automatically tiering data to more efficient spinning disk

Since traditional storage architectures wouldn't meet Recursion's stringent performance demands, Howard leveraged his 15 years in HPC to select the optimal processing and storage solution. He quickly dismissed other seemingly scalable storage solutions as inadequate for meeting the high-performance file processing demands.

"I've worked on lots of different solutions, and DDN storage is the most mature," Howard recalls. "Their tooling is better, and their engineers really know their products. If I need a two-node, 200-node or 200,000-node cluster, I know DDN is going to perform."

Performance

Reduced access times to files by two milliseconds

Scale

Architecture that supports AI-driven models

Flexibility

Faster than internal flash, economics of shared storage, and a path to cloud adoption

Experience

A career-long partnership built on expertise and trust

"Feeding the GPUs is our number one priority. Machine learning necessitates getting the data to the GPUs as fast as possible, so you can process it, get it out and train your models. I was most comfortable working with DDN."

Kris Howard

Principal Systems Engineer

DDN's reputation as a storage leader is reinforced by its increasing focus on AI data storage infrastructure. "Feeding the GPUs is our number one priority," says Howard. "Machine learning necessitates getting the data to the GPUs as fast as possible, so you can process it, get it out and train your models."

In collaboration with DDN's engineers, Howard initially created a POC, encompassing DDN's ES400NV® and ES7990X® storage appliances running Lustre with 2PBs of capacity for staging machine learning models. An all-flash layer was deployed as a front-end to the file system supported by ample spinning disk. The first 64K of each file is stubbed to this layer, which then accelerates access to the first part of the data before streaming the rest to spinning disk. Not only did this approach deliver extremely fast performance for Recursion's demanding workloads, it alleviated file-access bottlenecks while enabling efficient streaming to the GPUs.

"Our DDN storage is wicked fast," says Howard. "The Flash layer reduced our access times to the files, and we can get our GPUs to 100% utilization, and keep them pegged there. It's highly unusual to train data off a PFS, but it's a perfect solution for our use case."

DDN's flexibility in sizing Recursion's configuration to meet specific workloads has resulted in robust storage that seamlessly supports 18 nodes and 136 GPUs. "My infrastructure is bulletproof," adds Howard. "This lets me focus on other things, like maximizing how the cluster runs and helping users optimize their jobs. DDN hardware just works, which frees up so much room in my day."

The Benefits

- High-performance, reliable storage supports broader, faster, more efficient drug discovery pipeline
- "Wicked fast" DDN storage feeds world's largest dataset of cellular image data while feeding computational database of tens of millions of biological perturbations
- Lifesaving drug treatments now can be identified in weeks—instead of months, or years

Thanks to DDN's flexible, high-performance storage, Recursion is creating a broader, faster, more efficient drug discovery pipeline. The company's data-centric approach is a prime example of how intersecting biological research and machine learning can yield impressive results. The sophisticated platform addresses multiple areas of biology, leveraging SiRNA for sequencing genetic diseases, secreted factors for immunology and inflammation, along with CRISPR for sequencing infectious diseases. The high-performance system trains machine learning models required for drawing inferences from Recursion's massive cadre of cellular images. The models then facilitate development of computational "fingerprints," or phenotypes, of different biological perturbations.

"Our pipeline of more than 30 programs was developed in less than five years," says Howard. "Thanks to our AI-driven platform, we can identify potentially lifesaving molecules and drug types in weeks—instead of months, or even years, it can take to discover new treatments."

Future Challenges

Looking ahead, DDN will play an increasingly pivotal role in helping Recursion streamline AI-driven computational workflows while reducing time to discovery. Recursion is looking to synchronize DDN with cloud storage for handling raw images as well as supporting metadata extractions and data tagging.

Other ideas will be driven by the strong partnership between the two organizations. "I need people behind me I can trust," concludes Howard. "I know if I reach out to the DDN team, they'll do everything in their power to make sure issues are resolved and Recursion continues its mission to change the drug-discovery paradigm."

About DDN

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. DDN has designed, developed, deployed, and optimized systems, software, and solutions that enable enterprises, service providers, research facilities, and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud.

v1 (4/20)