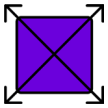
Break the Cryo-EM Storage Performance Barrier with WekaFS

BREAKTHROUGH PERFORMANCE
Shorten image processing and modeling time by 60-90% reducing data transfer time; 10x faster than traditional NAS

EASY MANAGEMENT
Flexibility and ease of deployment whether working on-premises or in the cloud

IMPROVE COLLABORATION
Petascale global namespace, with directory services integration for user authentication and permissions



UNMATCHED SCALIBILITY
Designed to scale to hundreds of petabytes, thousands of compute instances, and billions of files

With the combined power of cryo-electron microscopy (cryo-EM) and high-performance computing (HPC), scientists are opening up new frontiers in biochemistry and genetics. It is helping researchers make the fundamental scientific discoveries needed for the development of life-saving drugs and therapies, like immunotherapy and precision medicine. Cryo-EM is a technique used to construct three-dimensional models of proteins and molecules at near-atomic resolutions. The structure of a protein determines how it interacts with drugs and other substances in the living system. Discovering and modeling a protein structure is therefore essential in accelerating the drug and therapy discovery process.

The technique is now capable of capturing high resolution 2D images of protein molecules frozen in a thin layer of ice under cryogenic conditions using a transmission electron microscope (TEM). With the help of HPC systems and advanced software, thousands of these 2D images are then aligned and combined to form 3D models. Cryo-EM alone doesn't provide all the insights that researchers seek. Those insights come from the combination of cryo-EM and powerful HPC simulations, which process massive amounts of data to yield highly detailed 3D models of biological structures at sub-cellular and molecular scales.

CRYO-EM STRAINS TRADITIONAL IT INFRASTRUCTURE

The surge in cryo-EM is largely a result of better electron detectors and image-processing techniques. Today's TEMs generate image data faster and in greater volumes than typical IT infrastructures can handle. A single detector can produce up to 5 TB of data per day. Throughout the cryo-EM data pipeline, images are iteratively processed and analyzed. The largest raw dataset currently in the Electron Microscopy Public Image Archive (EMPIAR) is 12.9 terabytes. As a result, data storage and computation requirements have increased exponentially presenting the following IT challenges:

- **Data Size & Access Patterns:** High degree of variability between data sizes and access patterns of each process step. While it begins as a high throughput sequential access IO pattern use case, with each step it moves to a smaller size random I/O pattern.
- **Data Management:** Petabytes of research and clinical data are being created which needs to be stored, managed, shared, and analyzed effectively and securely.
- **Compute Limitations:** Traditional CPU-based compute clusters present bottlenecks causing costly delays in the analytics pipeline impacting overall system throughput. New software is optimized to take advantage of GPU-enabled HPC clusters that rapidly process data in parallel to dramatically reduce processing time.
- **Hybrid Cloud:** Local and cloud-based HPC big-data analytics workloads are converging, requiring an ability to seamlessly manage data across both on-premises and cloud environments.

Since this cryo-EM pipeline is composed of multiple processing steps and data access patterns, it takes a modern parallel file system like WekaFS from WekaIO to turbo charge results. With its ability to accelerate GPU workloads, WekaFS further decreases the time required to complete the process, allowing researchers to run additional pipelines for more accurate results. Figure 1 shows how WekaFS can be deployed in a production cryo-EM production environment.

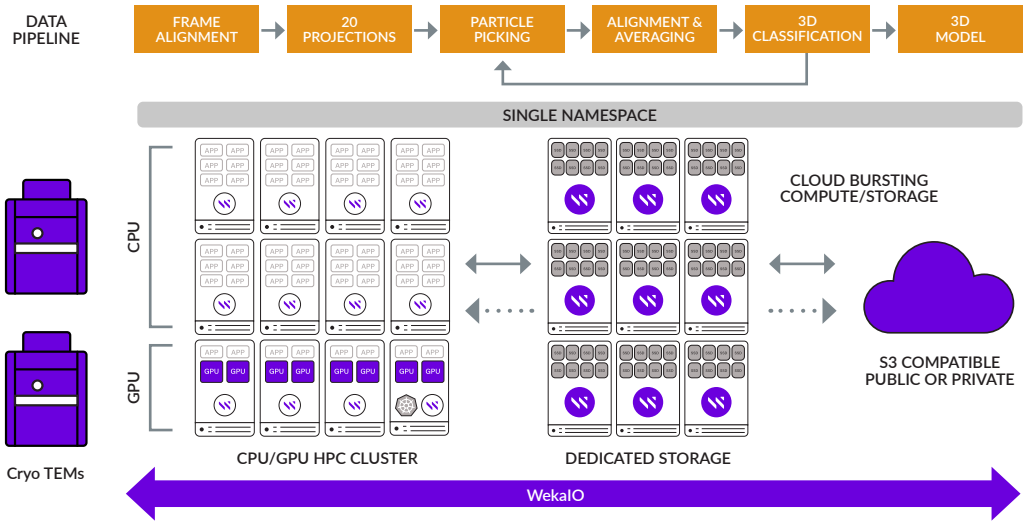


Figure 1 - WekaFS deployed in a cryo-EM production environment

BLAZING PERFORMANCE AT ANY SCALE

WekaFS was built to solve the most demanding and data intensive storage challenges. For large cryo-EM environments running multiple concurrent workloads, ultra-low-latency, high-throughput storage is key to eliminating compute cluster bottlenecks, and reduce valuable processing times. As the world's fastest shared parallel file system, WekaFS is 3x faster than local file systems and 10x faster than traditional NAS. Small file performance (<4k) is exceptional with top rankings on SPEC SFS 2014 and the IO-500 benchmark tests.

WekaFS is a distributed, scale-out and POSIX compliant file system, built on a modern architecture using ultra-low-latency NVMe-over-Fabrics (NVMe-oF). WekaFS does not have separate metadata services that can limit performance; metadata is distributed throughout the cluster via patented mechanisms that prevent hot spots. Performance scales linearly as more nodes are added to the storage cluster. WekaFS delivers the highest-bandwidth, lowest-latency performance to any InfiniBand or Ethernet-enabled GPU-based compute cluster. The system is capable of reaching 10s of millions of IOPS and tens of terabytes/sec of bandwidth at less than 300-microsecond latency.

SCALABLE AND COST-EFFECTIVE

WekaFS scales independently on two dimensions—capacity and performance. Running on standard x86 server platforms with local NVMe SSDs, WekaFS is designed to scale to hundreds of petabytes, thousands of compute instances, and billions of files, but you can start as small as 100TB and scale seamlessly with no added complexity. Current production deployments have anywhere from 8 to hundreds of nodes and tens of petabytes of storage under management. The Weka file system is an easy-to-configure, easy-to-deploy storage solution that adapts to your environment, giving you complete deployment flexibility.

Data is stored in a hybrid tiered storage architecture. Hot data resides on NVMe SSDs within the storage server cluster, and warm data is stored on any Amazon Simple Storage Services (Amazon S3) compatible object storage system with hard disks for cost-optimized capacity. Data movement between the two tiers is managed by WekaFS and is transparent to the application, so no modifications are required to access data on the secondary tier. Patented data protection mechanisms provide the most resilient large cluster deployment with the shortest rebuild time in the industry. WekaFS can easily handle multiple failures and is configured with +2 or +4 levels of failure domain reliability.

RUNS ANYWHERE–SPEED COLLABORATION

WekaFS provides flexibility, ease of deployment, and resiliency whether working on-premises or in the cloud for on-demand scalability. Transitioning datasets between locations is easy taking only a few clicks. By connecting an S3 bucket in Amazon Web Services (AWS) to an on-premises filesystem, researchers can experiment locally to prove functionality prior to moving a dataset in the cloud. WekaFS makes it possible to build filesystems directly from S3 data buckets stored in AWS, removing the hassle of downloading, parsing, and filtering datasets so researchers can focus on making discoveries.

The software supports full encryption all the way from the application clients to the storage system, and supports encryption on the fly and at rest. It is tightly integrated with key management systems that are KMIPS compliant. A petascale global namespace along with directory services integration for user authentication and permissions makes it easy and secure to share data with research collaborators.

Unleash your cryo-EM workloads and break free from current storage constraints with WekaFS.

WEKA

910 E Hamilton Avenue, Suite 430, Campbell, CA 95008 T: 408.335.0085 E: info@weka.io www.weka.io

W25SB202012