High Performance Compute and Storage Solutions for High-End Microscopy

Real-time analysis, scaling capacity and seamless data sharing to accelerate discovery

DDN provides fully-integrated and optimized data platforms for Gatan microscopes. Designed in close collaboration with Gatan, DDN's innovative technology solutions enable high-end microscopy workflows at any scale for life sciences, materials research, and other industry applications. DDN vastly accelerates the discovery process compared to current storage with architectures that simplify and streamline all aspects of data storage and processing.

DDN enables data capture and subsequent analysis from Gatan K3, K3 IS, Rio, OneView, and other advanced microscopes at dozens of top research centers globally. This paper presents an overview of workflow designs from DDN for high-end microscopy and recommended architectures for high-performance data storage and compute infrastructure. Several customer specific success stories are also available from DDN.

High-End Microscopy Imaging and Processing Workflows

High-end microscopy workflows are data intensive and typically require a high-performance centralized data platform to ensure the best utilization of instruments, processing facilities, and personnel time. A properly integrated and optimized data architecture is an enabler to the scientific process and catalyst to collaboration and breakthrough discovery. High-end microscopy workflows are dynamic and multiple types of imaging and processing are engaged depending on experiment being conducted.

Solving the Data Challenges of High-End Microscopy

Real-time analysis and continuous production is a challenge for most research facilities. Sensors used in the microscope are often deployed with a local data storage platform. Typically, local data storage platforms have sufficient capacity to store only a small volume of data compared to the actual capabilities of an instrument. This puts crippling limitations on the type and duration of imaging, often limiting experiments and preventing operators from taking full advantage of the capabilities of high-resolution instruments. Local data storage also introduces a very high data management overhead. When using local data storage systems, data must be continuously copied between different storage locations. This requires operational diligence and coordinated effort among all collaborators using the instrument. It also introduces noticeable delays in data availability for analysis and can require lengthy instrument downtime. Local data storage platforms also have limited redundancy and protection which puts the data at risk of complete loss in case of operational error or hardware component failure.

To eliminate these limitations, overhead and risks, DDN recommends a centralized shared storage and compute architecture built using high-performance, scalable, and reliable data platforms. The DDN architecture eliminates bottlenecks and reduces data management overhead by enabling data captured from the microscope to be written directly to a central storage system during acquisition. From there, with no additional transfer steps, the captured

data is immediately available to the compute cluster for use in processing applications and other analysis. The DDN Shared architecture enables a significant leap in efficiency and significantly accelerates high-end microscopy workflows.

The DDN architecture is very flexible and scales simply to meet evolving workflow capacity, performance, and capability needs—all independently. It is capable of supporting continuous acquisition from multiple microscopes running at the same time, and instantly sharing the data with collaborators across multiple sites. Data availability and security are well-recognized cornerstone of DDN products. Data is protected through multiple mechanisms, and the systems are architected with multiple layers of redundancy. This ensures maximum infrastructure uptime to support microscopy workflows at production centers of any scale, while ensuring that data governance requirements are met.

DDN Solutions Enrich High-End Microscopy Imaging and Accelerate Time to Results

High-end microscopy imaging involves multiple iterations between sample optimization, sample preparation and screening the grid on the microscope to see if the sample is good enough to do high-end data collection. During these iterations, a lot of data is collected and analyzed, with the direct electron detection camera in integrating mode. This valuable data is oftentimes stored and retained long term for research. By eliminating data transfer and enabling realtime analysis, the DDN shared storage architecture enables faster turn-around of sample iterations, leading more quickly to high-end data collection and increased imaging bandwidth.

High-end data collection also involves several iterations with the direct electron detector in electron counting mode which generates an incredible amount of data that needs to be stored and processed. For example, with its high frame rate, the Gatan K3 direct detection sensors produce up to 3.6 GB/s. High-end microscopy facilities routinely generate tens of terabytes (TB) per day from a single microscope. Processing this data requires additional space, as much as 3X the size of the raw data. It can be useful to have an offloading server to preprocess the raw movie stream coming off the acquisition system to realign and average the movie frames, as well as to extract particles from the images and remove all the blank background. This strategy reduces the size of the data set as part of the compute process before it is written to persistent storage.

Total data storage requirements depend greatly on data management strategies and data retention policies. For example, organizations often consider whether to keep all of the data produced as part of their activities, or only published data. At one center, where long-term data storage is growing at a rate of about 100TB per month, they keep movie frames for a couple of months and then automatically delete them, with a two-week warning to users so that they can transfer them elsewhere if they wish to keep that data longer. Most will keep some selected raw images or frame average images for as long as they can, and processed data is similarly kept.

While it may be desirable to keep some data forever, that does not mean it will be actively used all that time. DDN solutions facilitate a structured, efficient and effective data management and governance. For instance, DDN data storage systems can be deployed with multiple tiers of drives, maintaining seamless and immediate access to data for users, while enabling technology-efficient and cost-optimized data placement. Microscopy workflows at centers supported by this tiered storage system, leverage a pool of high-performance drives for active data, and a pool of higher-capacity, slower drives to store data that does not get accessed as frequently. The data placement across tiers is done automatically based on customizable policies and the data remains seamlessly available at all times to the users. DDN also offers several tools that provide advanced data management capabilities, such as real-time storage analytics and usage reporting, and for short- or long-term data archiving. These tools can make powerful data management solutions accessible to data experts and non-experts alike.

Data set sizes vary a lot from project to project, spanning several orders of magnitude from the smallest to the largest. For example, high-resolution maps can run for over a week on the microscope. A typical project data set size is perhaps 5TB in size, including raw images and processed data, but some can be quite a bit larger. This is particularly true in tomography which runs at a much higher frame rate than single particle and can generate well over 30TB of raw data in a single day. A typical project will then take several days of computation requiring a cluster of GPU servers. DDN solutions ensure compute and storage can process the data in real-time, enabling maximum utilization of the microscope, avoiding unnecessary interruptions, and preventing the accumulation of backlogged jobs waiting to be processed.

The data processing typically goes through various steps: initial model building, model refinement, visualization, and validation. Specific applications used in those steps were traditionally both memory and CPU intensive. For instance, alignments are iterative processes so need to hold all the data in memory. Processing jobs typically require 160-200 GB of memory, and might run for several weeks on hundreds of CPU cores. Recently many processing applications have been optimized for GPU computing to speed up analysis, enabling faster time to results. These analysis workflows are much more data IO intensive. In order to fully-benefit from GPU accelerated applications, the data storage infrastructure supporting the compute systems must be capable of handling the augmented IO load, reading and writing data at a rate that keeps pace with the GPU. Within the cluster, supporting the data IO intensive GPU and storage traffic is a high throughput and low latency InfiniBand™ (IB) network fabric.

A common approach is to deploy expensive solid state drives (SSDs) in every GPU compute node as fast local data workspace. This introduces a workflow requirement to load and offload data to the compute node as part of the processing pipeline, which results in lower productive utilization of the node and heavy data management overhead. Also, this approach restricts heavily the use of distributed parallel compute and hinders multi-node collaborative job processing. Rather than using local disk storage, DDN recommends deploying a centralized storage system capable of meeting the real-time IO requirements of all CPU and GPU compute nodes used within a processing cluster. This approach eliminates unnecessary data migration and management operations, and enables maximum utilization of compute resources for fastest time to results. Also, the DDN architecture delivers the performance and agility required for effective use of advanced distributed compute techniques such as machine learning.

DDN solutions are uniquely capable of delivering data storage and compute performance to meet the demands of a high-resolution microscopy workflow. The DDN storage systems can deliver sustained performance to ensure that IO needs of all operations—acquisition, processing, visualization, archiving—are met in realtime and never hindered waiting for data from an underperforming infrastructure. Beyond delivering high-performance, DDN solutions are also extremely agile. While the overall size of the project directory might be several terabytes (TB), individual file sizes accessed by applications can range from a couple of kilobytes (KB) to a couple of gigabytes (GB). Unlike enterprise data platforms commonly deployed in IT infrastructure, DDN solutions are designed to perform well for a wide variety of data sizes and access patterns, maximizing the value that users can get out of their microscopes, as well as potentially supporting a breadth of users with a variety of research workloads other than microscopy.

An additional strength of the DDN solution is its ability to scale from small capacity to an extremely large capacity seamlessly with small modular building blocks. This provides flexibility for DDN to deliver solutions tailored to meet a wide-range of customer needs. It also provides confidence for customers that their systems can be expanded easily and reliably as their capacity and performance needs evolve over time. DDN supports sites that require an extra-large system to meet the needs of a consortium of users and independent labs for which a smaller system meets all of the current needs equally well. Several High-end microscopy sites initially deployed a small solution, and regularly added capacity to become a large system over time.

With DDN, High-end microscopy workflows run faster, better and more effectively.

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.

v3 (4/21-KK)