2022-23 DCIG TOP 5

RISING VENDORS IN STORAGE FOR LIFE SCIENCES SOLUTIONS

By Mike Matchett

Table of Contents

Rising Vendors in Storage for Life Sciences Solutions

SOLUTIONS EVALUATED:

- DDN EXAScaler
- Dell PowerScale F900
- Hitachi Vantara HCP
- HPE ClusterStor E1000
- Huawei OceanStor 9000
- IBM Spectrum Scale
- Nasuni
- Netapp HPSS (E-series)
- Panasas ActiveStor Ultra
- Pavilion HyperParallel Data Platform
- Pure Storage FlashBlade
- Quantum StorNext
- Qumulo File Data Platform
- Quobyte
- VAST Data Universal Storage
- WekaIO WekaFS

SOLUTION FEATURES EVALUATED:

- *Deployment Capabilities*
- *Data Protection*
- *Product and Performance Management*
- *Documentation Support*
- *Technical Support*
- *Licensing and Pricing*

Rising Vendors

DCIG recently conducted research into Storage for Life Sciences solutions. This 2022-23 DCIG TOP 5 Rising Vendors in Storage for Life Sciences report focuses on the top solutions by vendors of size less than $200M whom we consider to be "rising" in the market. The use cases are the same, but rising companies tend to be younger companies bringing newer technologies and disruptive solutions to market. Those organizations looking for modern approaches and potential competitive advantage may want to consider working with a rising vendor.

Storage for Life Sciences Challenges

Life Sciences organizations depend on some of the most compute and data-intensive applications in the world for primary research, on-line analysis, global collaboration and product development. Many of these applications are High Performance Computing (HPC) quality workloads that include genomics sequencing, molecular simulations, protein folding, AI/ML optimization, and intensive media processing. These applications can push even cutting edge IT data storage implementations to their performance and capacity limits.

And increasingly, upstream life sciences workloads feed critical data into downstream workflows that might for example manage real-time medical-grade production lines or oversee global distribution.

The stakes are high for IT in life sciences with competitive global pressures, the search for life-saving solutions, expensive data "source" equipment, elite research staff and ever-increasing data volumes and performance demands. For example, voracious genomics sequencing equipment can easily generate TB's of raw data in a few hours, overwhelming local legacy file storage often configured with the default operator workstation. This data must then be offloaded into downstream research-feeding storage while that equipment is idling, wasting opportunity, time and resources.

Other workloads present challenges too—molecular simulations running on scale-out HPC clusters impose HPC data consumption patterns, which require local high-speed parallel file IO of very large or very many files to many client nodes at once. Many life sciences workloads need to strongly leverage critical GPU resources, and maximizing the utilization of large numbers of GPUs may require specific GPU-related storage features. And on-line data capacity demands mount with each type of research conducted, each form of analysis required, and every AI/ML model built.

Traditionally, data requirements for these highly demanding applications have been served with parallel file system solutions. Parallel file systems can deliver massive IO volumes to data hungry applications (like those running in large HPC or HPC-like compute clusters), but historically implementing, operating and tuning massively parallel file systems for top performance at scale has been the province of dedicated PhD level academics.

Today, huge volumes of data are not only created and processed by primary life science applications, but that data then must be shared with multiple consumers and global collaborators. It is obviously expensive (and often prohibitively slow) to make and maintain many multiple copies of very large data. It can also be quite expensive to maintain large histories of data online for downstream research activities. It is traditional to archive HPC data sets

into object storage for downstream use. However, this secondary storage dataflow means that data lives in multiple locations and in different forms, adding friction and delay to sharing, complicating data access and preventing optimal data value extraction.

In addition, "academic" storage solutions have tended to be light on enterprise storage management features like data protection, security and backup/disaster recovery functionality. While primary HPC storage must first meet the HPC application requirements, it is increasingly the case that research data stays on-line for longer time periods to be recalled and leveraged on-demand and accessed by a wider set of applications. Maintaining this key data over time properly then becomes more critical to the ongoing overall success of the organization.

An effective storage solution will ingest and store data faster than source equipment can produce it and then deliver it as fast as the sum of consuming workloads calls for it.

Benefits of an Effective Storage for Life Sciences Solution

At the top level, scalable storage performance is critical. A life science organization should try to fully utilize their high-end research and laboratory equipment, HPC clusters and GPU-intensive analytical servers. The objective of life science storage then is to store and flood massive amounts of scientific data into all the expensive data pipelines serving the organization's goals. An effective and equally scaled storage solution will ingest and store data faster than source equipment can produce it and then deliver it as fast as the sum of consuming workloads calls for it.

Scalable storage will scale-out almost indefinitely, as needs and requirements grow or expand. The best solutions can offer a single "namespace" for all files even as the data storage under management grows into exabytes, eliminating multiple arrays, needless replication and extraneous data copies. Storage solutions that really support massive scale-out to thousands of devices also offer resilient designs and non-disruptive upgrade/repair features to avoid single points of failure and downtime.

Perhaps even more important than maximizing resource utilization is fully empowering scientific staff. The top storage solutions work transparently to keep data flowing to all users, in real-time on-demand, driving their applications and analytical workflows without lag or downtime. The best scalable storage solutions also present simple (and automated) data storage interfaces, freeing staff from onerous storage management concerns and enabling them to focus more on their productive research.

There is plenty of competition in the global race to life science insight and discovery, with massive opportunity for organizations that can most efficiently leverage resources, empower researchers and minimize IT risk and distraction. Life sciences storage can make a significant difference in organizational outcomes by delivering world-class performance, scaling readily to handle the largest of online data requirements and significantly increasing organizational efficiency.

Finally, we are seeing top end storage solutions increasingly supporting downstream and collaborative workflows through wider multiprotocol support, native storage tiering, inherent data protection, multi-tenancy and data security features. Overall, top life sciences storage solutions, even though often accelerated on high-end appliances or custom hardware, are becoming more cloud-like in utility, economics and management.

Distinguishing Features of Storage for Life Sciences

In addition to the broad capabilities mentioned above, all of the life sciences storage solutions evaluated in this report share some features that help distinguish them from the broader IT storage market.

File performance. First, these are not simply scaled up NAS solutions, but designed from the start for high-end file storage performance and capacity.

Large capacities. Scalable architectures are the norm, with the ability to add or expand on-line and tiered capacity without grossly affecting operations or performance.

Broadening application support. Life sciences research encompasses a wide variety of applications, usage, access, cost, risk and data management concerns. The storage solutions evaluated for this report have demonstrated significant utility in some slice of life sciences and threaten to increasingly consolidate storage with broadening application support.

Resiliency at scale. All of the life sciences storage we evaluated have features that address the resiliency and resulting availability of the solution at large scales of deployment.

Similarities Among the DCIG TOP 5 Rising Vendors in Storage for Life Sciences

In addition to the distinguishing features above that all of the evaluated storage solutions share, the selected DCIG Rising Vendors in Storage for Life Sciences solutions have the following traits in common:

File performance first. The top solutions need to be highly performant file storage solutions designed to serve HPC class performance requirements for high-volume data reads and writes.

Boundless scale-out. The top solutions all are capable of scaling-out to many PB's (or even EB's) of storage without impacting performance, often to hundreds or thousands of storage/server nodes.

Large and small files. High-speed ingest and performant read of large files to feed HPC-class workloads are key life science storage design points, but also increasingly these systems support billions (or trillions) of small files, random IO access patterns and low latency applications.

The DCIG TOP 5 rising vendor solutions also deliver the following product features:

Files integrated with objects. All of the top solutions provide unified object services or internally tier to cloud and/or cloud-like object storage.

Integration simplicity. They all provide modern storage consoles for management of storage at large, but also REST APIs for management integration at scale.

Aim to ease management burdens. These solutions are all designed to ease storage management burdens at scale especially those caused by siloed NAS deployments.

Differences Between the DCIG TOP 5 Rising Vendors in Storage for Life Sciences

Not all storage solutions are the same. Many have evolved from differing original design implementations with different original design goals. Our evaluated storage solutions each may have a "sweet" spot within the broader life sciences market, and most life sciences organizations today deploy more than one storage solution.

The 2022-23 DCIG TOP 5 Rising Vendors in Storage for Life Sciences solutions can be seen to differ from one another in the following ways:

Deployment architecture. Most storage solutions these days are inherently described as software (i.e. "software defined"), but many are often and perhaps best acquired pre-integrated on certified appliances. With wider cloud adoption and global workflow initiatives, cloud deployment or extension of scalable storage can be increasingly important to larger data and storage architectural designs. Also note that some storage described in this report can support either a different storage OS or lay on type of widely varying infrastructure.

GPU acceleration. It is important to many life sciences applications (and to recoup resource investment) to be able to drive available GPU's as hard as possible. Different storage solutions may provide native support for GPU Direct operations, converged storage/GPU systems (e.g. NVIDIA DGX), or other features designed to drive GPUs as hard as possible.

Parallel perspectives. Some form of parallelism in design does seem to be required to deliver on demanding life sciences performance requirements, but it no longer strictly means just a classic parallel file system. Some of our top solutions offer parallel operations due to network switching, storage node distribution, edge caching or back-end storage parallelism.

Qumulo is especially suited for the capacity storage of petabytes of media files (images, video) while simultaneously servicing demanding video processing or research workloads with cloud-elastic or bursty IO patterns.

DCIG TOP 5 Rising Vendors in Storage for Life Sciences Solution Profiles

Each of the solution profile highlights three notable features that make the solution attractive for this market.

Qumulo

Qumulo is a cloud-native, "single-tier" global file system designed for efficient, extreme-scale unstructured data capacities and HPC-class file performance. Qumulo is especially suited for the capacity storage of petabytess of media files (images, video) while simultaneously servicing demanding video processing or research workloads with cloud-elastic or bursty IO patterns.

Qumulo applies native intelligent file-level analysis, predictive pre-fetch and caching best utilize assigned NVMe resources to maximize performance, while delivering linear scalability through auto-tiering across diverse hybrid cloud and active-archive storage capacities. Qumulo clusters can be deployed across all major public cloud providers, HPE, Dell, Pure, Fujitsu and other underlying storage arrays or are available through Qumulo pre-integrated hardware.

Three of the key features that earned Qumulo a spot among DCIG TOP 5 Rising Vendors in Storage for Life Sciences solutions include

Simple global hybrid cloud. Qumulo delivers a single namespace with multiprotocol access (SMB, NFS, FTP, REST) across all assigned storage. No application migration or transformation is required to access files in any environment, including in the cloud, easing both storage consolidation and data collaboration tasks.

Consistent workflow integration. Rich API's from Qumulo enable the tight workflow integration of storage with complex data pipelines and critical application workflows found in key use cases like genomics, molecular simulations, image processing and PACS.

Real-time analytics. Qumulo provides instant data and usage visibility across billions of files. The resulting intelligence enables the "full utilization" of storage resources and powers the Qumulo platform's automatic predictive file-level pre-fetch and caching for high performance.

Pavilion HyperParallel Data Platform

The Pavilion HyperParallel Data Platform provides massive IO parallelism through its unique architectural approach that basically hyper-converges a high-speed network switch with a unified flash storage array.

The Pavilion HyperParallel Data Platform provides massive IO parallelism through its unique architectural approach that basically hyper-converges a high-speed network switch with a unified flash storage array. The base 4RU system can hold over 2PB and up to 20 controllers, while serving high-performance block, file and object I/O as every controller can talk directly to any disk. Pavilion can cluster an unlimited number of these base arrays with linear (or even increasing) performance.

While the Pavilion HyperOS and its inherent Pavilion HyperParallel File System provide a combined NFS/S3 global namespace across an unlimited number of array units and assure both performance and availability at scale, Pavilion's flash arrays can also host other high-performance file systems (e.g. IBM Spectrum Scale) if desired.

Three of the key features that earned Pavilion HyperParallel Data Platform a spot among DCIG TOP 5 Rising Vendors in Storage for Life Sciences solutions include

Unlimited linear performance (and capacity) scaling. A remarkable storage density enables Pavilion HyperParallel platform users to start fairly small but grow both scaling up (to 2PB per unit) and scaling out (unlimited) cost-efficiently as necessary. Because of the integrated network switch design, performance across clusters can be linear with growth.

Built for serious workflows. The Pavilion HyperOS and HyperParallel File System provide a native multi-protocol access to enable a single, unified storage repository to power both very high-speed data ingest and parallel client consumption of fast data in disparate forms. Protocols include NVMe-of(RDMA, IB or TCP), iSCSI, NFS and S3 with native support on any combination of controllers across any number of arrays.

Integrated enterprise data services. Tiering, replication, snapshots, clones, security, encryption and application plugins (popular S3-based apps, ML, Big Data, Spark, TensorFlow, Kafka, Splunk, Teradata and more) are all included in the core platform. Hardware upgrades and cluster expansion can be made non-disruptively, and the platform design eliminates single points of failure common to other arrays.

Quobyte

Quobyte readily combines flash and HDD into a single namespace with inherent transparent migration and tiering. Uniquely, files (especially large files) can be internally tiered across HDD and flash.

Quobyte is a software storage system based on a parallel distributed POSIX file system built for high performance computing with a distinctly hyperscaler-oriented design. Quobyte is deployable on-prem, in clouds and in K8 environments. It readily combines flash and HDD into a single namespace with inherent transparent migration and tiering. Uniquely, files (especially large files) can be internally tiered across HDD and flash. The scale-out Quobyte cluster provides linear performance scaling of throughput, IOPs and metadata with full "mesh" communication and "quorum voting" to eliminate bottlenecks and avoid critical failures. Quobyte works well for both serving large numbers of small files as well as high speed large file access. As true software storage, Quobyte can be installed on a node with a one line install (no server configuration required).

Three of the key features that earned Quobyte a spot among DCIG TOP 5 Rising Vendors in Storage for Life Sciences solutions include:

Automatic internal file tiering. Without wasting resources on cache capacity or specific storage tuning, and maximizing flash investments, large file access receives the best of both worlds - low latency and high throughput as the first "blocks" come are served from flash while parallel access across multiple nodes will then deliver a flood of streaming data from more cost-efficient disk.

Full multiprotocol access. Quobyte provides a massive single namespace and unified ACL across all interfaces including Linux and Windows (NFS and SMB), S3/Object, HDFS, TensorFlow and more. Files are essentially objects in the Quobyte file system, and benefit from the backend object storage principles without impacting file performance.

Security built-in. Quobyte leverages advanced erasure coding for cluster protection. It provides encryption at the client-side so storage admins can't access data. And storage management operations including updates, expansion/replacement, data migration, and policy configurations are all non-disruptive to the larger cluster. Broken hardware and other resource failures are automatically factored out as they occur, supporting the hyperscaler "replace rather than repair" model which is key to easily maintaining cloud-scale storage.

VAST Data Universal Storage

VAST leverages of large amounts of Storage Class Memory to buffer writes at memory speed. Flash-optimized wide-striping across cost-efficient flash storage ensures maximum parallel flash read performance for all data.

VAST Data's all-flash scale-out storage is designed to take advantage of several newer technologies—NVMe-oF, a "hyperscale" flash architecture, and storage class memory. The VAST Data Universal Storage solution attempts to converge primary and secondary storage into one tier for all data, enabling deep capacities of data to reside essentially inside a single Exabyte-scalable globally name-spaced system.

VAST front-end servers are essentially fungible, stateless points of access that can be scaled up to 10,000 nodes (VAST appliances or containers). Each server can access every storage node over an NVME fabric. Every one of up to 1000 storage nodes can support over 2PB/2RU (VAST's own storage appliances, nodes can also be run on other certified hardware).

Three of the key features that earned VAST a spot among DCIG TOP 5 Rising Vendors in Storage for Life Sciences solutions include:

Performance in and out. VAST's leverage of large amounts of Storage Class Memory buffers writes at memory speed. This memory staging then enables background deduplication and destaging out to flash-optimized wide stripes across cost-efficiently driven flash storage distributed across the cluster. Keeping all data in flash (and cached in memory) ensures a minimum of parallel flash read performance for all data—great for random access, metadata intensive operations, and AI/ML applications.

Storage simplicity. As a single Exabyte-scalable tier with multiprotocol access to all data (NFS, NFS over RDMA, S3, SMB and containers), management is simplified and usage is completely "democratized". Storage is protected with a proprietary Global Erasure Code that declusters error correction to provide near instant recovery of data due to failed devices.

Capacity optimization. Incoming data is cleverly deduplicated using a proprietary algorithm that leverages a measure of similarity to existing blocks to create a "local" and more performant encoding/decoding capability within the larger scale-out cluster. Decompression is fast (out of flash). Overall, this capability makes the investment in their cost-efficient flash architecture go even farther.

WekaIO

Weka can integrate and internally auto-tier with a third-party object store to create a "single-tier", single-namespace solution with combined high-performance and scalable capacity storage.

Weka is a POSIX-compliant high-performance clustered parallel file system designed to run natively over NVMe drives. It can integrate and internally auto-tier with a third-party object store to create a "single-tier", single-namespace solution with combined high-performance and scalable capacity storage.

WekaFS is software that runs on commodity hardware and is deployable across a range of environments from virtual servers, containers and on-premise x86 Linux servers to cloud instances as a dedicated appliance, converged, or cloud-native solution.

Three of the key features that earned WekaIO a spot among DCIG TOP 5 Rising Vendors in Storage for Life Sciences solutions include:

Parallel flash performance. WekaFS optimizes the use of local NVMe drives on each storage node and converts local (Linux) node caching into "adaptive caching", a type of cluster-coherent shared caching. Weka can be leveraged simply to accelerate the performance of slower object storage implementations.

Resiliency at scale. WekaFS provides enterprise NAS management features like storage native snaps, clones and a "snap to object" which can be subsequently rehydrated on non-mirror systems for a wide variety of agility-enhancing use cases. Weka is resilient to failures with a unique node striping approach and advanced erasure coding that enables rapid drive rebuilds.

Cloud-like economics. In addition to essentially pooling industry-standard server and storage for performance, WekaFS also integrates tiering to offload colder data to cloud and/or on-premise object storage through S3 and Swift protocols. File-based apps can run on Weka in the cloud without modification, and data can be served out of archive object stores with high performance.

Inclusion and Evaluation Criteria for Storage for Life Sciences Solutions

In this report, DCIG specifically focused on Storage for Life Sciences solutions possessing the following characteristics. DCIG identified fifteen different solutions meeting these inclusion criteria:

- Commercially available on December 1, 2021.
- Sufficient, publicly available information available for DCIG to make an informed decision.
- Clear vendor intention to support life sciences and healthcare use cases as evidenced by publicly-available, solution-focused content.
- Demonstrated business in these industries.
- Capable of supporting global initiatives, meeting demanding application file performance requirements, and scalable to petabytes.

DCIG evaluated each of these solutions in the following areas:

1. ***Deployment capabilities.*** Evaluate the capabilities concerning on-premise deployment options, cloud provider deployment options, cloud provider targets supported, storage protocols supported, virtual environments supported, and certifications with equipment, operating systems, and applications.
2. ***Data protection capabilities.*** Evaluate solution capabilities supporting availability, encryption, replication, and snapshot features.
3. ***Product and performance management features.*** Evaluate options to manage the underlying hardware and optimize it for performance. Examples include dashboard views, predictive analytics, storage optimization, quality of service features, auto-tiering capabilities, and directory service integration.
4. ***Documentation support.*** Evaluate the breadth and depth of documentation the provider makes available to customers. Examples include whitepapers, knowledge bases, online manuals, as well as community forums.
5. ***Technical support.*** Evaluate the availability and technical support options of the solution provider. Examples include support availability, response time commitments, options to open cases, escalation support, and proactive problem resolution.
6. ***Licensing and pricing.*** Evaluate the relative ease of doing business through flexibility and simplicity in contract lengths, pricing elements, and bundled pricing options.

DCIG Disclosures

Vendors of some of the solutions covered in this DCIG report are or have been DCIG clients. There are some important facts to keep in mind when considering the information contained in this report and its merit.

- No vendor paid DCIG any fee to research this topic or arrive at predetermined conclusions.
- DCIG did not guarantee any vendor that its solution would be included in this DCIG TOP 5 report.
- DCIG did not imply or guarantee that a specific solution would receive a DCIG TOP 5 designation.
- All research is based upon publicly available information, information provided by the vendor, and the expertise of those evaluating the information.
- DCIG conducted no hands-on testing to validate how or if the features worked as described.
- No negative inferences should be drawn against any vendor or solution not covered in this report.
- It is a misuse of this DCIG TOP 5 report to compare solutions included in this report against solutions not included in it.

DCIG wants to emphasize that no vendor was privy to how DCIG weighted individual features. In every case the vendor only found out the rankings of its solution after the analysis was complete. To arrive at the solutions included in this report, DCIG went through a seven-step process to come to the most objective conclusions possible.

1. DCIG established which features would be evaluated.
2. The features were grouped into six general categories.
3. Solution providers were given an opportunity to complete a product survey. A DCIG analyst examined the feature data for each solution or completed a survey based upon the analyst's own knowledge of the solution and publicly available information.
4. DCIG identified solutions that met DCIG's definition for a Storage for Life Sciences solution.
5. DCIG weighted each feature to establish a scoring rubric.
6. DCIG evaluated each solution based on information gathered in its survey.
7. Solutions were ranked using standard scoring techniques. ■

About DCIG

The Data Center Intelligence Group (DCIG) empowers the IT industry with actionable analysis. DCIG analysts provide informed third-party analysis of various cloud, data protection, and data storage technologies. DCIG independently develops licensed content in the form of DCIG TOP 5 Reports and Solution Profiles. Please visit **www.dcig.com.**