Intersect360 Research White Paper: MANAGING DATA AT SCALE: DELL EMC STORAGE FOR HPC

EXECUTIVE SUMMARY

It is hard to find an enterprise-class user that has not been touched by major industry trends related to the continued influx of data. Correspondingly, there has been a proliferation of technologies and solutions for managing data at scale. At the enterprise level, successful implementations evaluate a range of these technologies and implement them in proportion to specific workloads or needs. Any company may find itself influenced by high-performance analytics, edge computing, machine learning, or cloud. Managing data at scale has become a preponderant theme of enterprise computing that will continue through the 2020s.

The most important aspect of these trends is that they are converging. AI, analytics, and High Performance Computing (HPC) are not independent markets at all, but rather interdependent approaches that coexist across an organization. This presents a tremendous challenge in data management. Companies have access to more data than ever before, and it has the potential to be extremely valuable, with multiple approaches to unlocking that value. Point products certainly exist, but across the broader market, a larger opportunity exists for a storage vendor that can provide a full range of scalable, high-performance solutions, backed by enterprise features and reliability.

Dell EMC is such a company. With a broad, trusted portfolio spanning HPC, analytics, and AI, Dell EMC has become a leader in storage, computing, and complete solutions for high-performance workloads, including storage arrays like Dell EMC Isilon, ECS, and PowerVault. Since Dell's acquisition of EMC in 2016, the company's leadership across high-performance data management has continued to evolve through trends in HPC, analytics, and AI, with solutions in hardware, software, and services to optimize environments for maximum value.

This white paper explores ways in which industry and technology trends are driving demands for storage and data management, and Dell EMC offerings for meeting these demands. Intersect360 Research expects more organizations to recognize the critical role of storage architecture in extracting value from data, as HPC, analytics, and AI continue to consolidate. Leading storage vendors will be those who can not only deliver a wide range of optimized products, but more importantly implement intelligent data management solutions that migrate data automatically and on-demand between higher-performance storage elements and economical long-term archive, while maintaining accessibility enterprise-wide for users and applications. This will become the ultimate test in turning data into value.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

INDUSTRY TRENDS: ANALYTICS, SCIENTIFIC COMPUTING, AND AI

All Kinds of Workloads; All Kinds of Data

Across all fields of computing, much has been made about managing the ongoing onslaught of data. Scientific instrumentation that has always generated data now does it at higher resolution. Other types of scientific data that used to be analog are now digitized. Personal devices, such as phones and tablets, create new types of consumer data. And the Internet of Things (IoT) has begun generating data from previously uninteresting devices. Purchases get scanned, photos get posted, and notifications chime. What do we get from all that data?

In some cases, the answer is obvious. The scientific data is generated specifically because it is valuable. Oil and gas companies capture high-resolution seismic images in order to locate and curate new oil fields. Manufacturers create digital models of their products to improve quality and to reduce time to market. Pharmaceutical companies run simulations to predict the efficacy of new drug treatments. Medical research, weather simulation, models of the Big Bang: with increases in data, any of these examples, or countless others, can be done at increased levels of fidelity, further enhancing their scientific value.

This escalating data dynamic is one of the key drivers of the High Performance Computing (HPC) market, which is projected to exceed $50 billion worldwide by 2023. New innovations in engineering and business are continuously fueled by available data and the ability to drive insights from it. But the data-driven use cases—and the resultant need for reliable performance at scale—don't stop there.

Earlier this decade, a new type of workload was so entrenched with this concept it was tagged with the accompanying moniker: "Big Data." What big data (now lower-case) represented was the desire to drive value from all the data being generated, particularly at the enterprise level. In the era of data warehousing, the question had once been, "How can we add structure to our data, so that we might learn from it?" Big data analytics advanced the game: "How can we learn from our data, regardless of whether it has any structure?"

Learning from All That Data

If big data changed the game by asking how to learn from data without giving it structure, another technology area more recently upped the stakes further with a new inquiry: "How can we learn from data in non-traditional formats?" Increasingly, data has been accumulating in rich-media formats, such as images, video, and audio? Humans have traditionally been good at sorting through this data, not computers. Now, thanks to advancements in the availability of data, the scale of HPC, and investment by hyperscale web companies, there has been a revolution in using artificial intelligence (AI)—or more specifically, *machine learning*—as a complement to other application approaches.

Historically, most scientific applications have been *deterministic*; based on a set of inputs, a program runs a bunch of calculations and comes up with an output, which is the answer.

Machine learning represents a new category of application that is *experiential*. Based on patterns seen previously, a machine learning algorithm makes inferences about current or future situations. This approach is called "artificial intelligence" because it mimics how humans learn.

Machine learning can be deployed any time there is a wealth of data to draw on, coupled with a reward from making more intelligent inferences based on that data. As described above, this limitation is not much of a limitation at all, as there are ever more organizations wondering how they can take advantage of available data, much of which is not in a format that is easy to compute on.

And in fact, this data is everywhere. We have "all kinds of data," not only in the colloquial, quantitative sense, but also in the literal sense of all types. Data is at the point of sale, and it is in the R&D center, and it is at corporate headquarters. Data is on-premise (which is often multiple locations), and it is in the cloud (also usually multiple locations). It is in smart devices at the edge, in live video streams, in instrumentation, and in satellite capture. Data is internal, or it is accessible online. It is as old as public archives and as new as social media streams. The question is, how to get value from it.

Bringing It All Together

Data isn't the only thing that the converging trends in HPC, analytics, and AI have in common. The other is that they tend to coexist, often sharing the same budgets, same infrastructure, or same personnel. On the whole, budgets are increasing—faster in organizations that are pursuing machine learning that at those that are not—but in general, these approaches work together, not separately.[1]

This presents a tremendous organizational challenge in data management. Companies have access to more data than ever before. It has the potential to be extremely valuable, and there are multiple approaches to unlocking that value. Where there is data consolidation, many organizations are aggregating it into "data lakes," expanding collections of data in its raw or unaltered form. For years, the notion of a data lake had a negative connotation, in that it implied that the data was unorganized and therefore not in a format for analysis or even search. Advancements in AI and analytics have the potential to change that, gleaning valuable information from consolidated data regardless of its format or organization. Now the great question seems to be, "How can I manage my data so that I can take advantage of it, with HPC, analytics, or AI?"

Machine learning can be deployed any time there is a wealth of data to draw on, coupled with a reward from making more intelligent inferences based on that data.

Organizations are seeking to take advantage of their available data, much of which is not in a format that is easy to compute on.

[1] Intersect360 Research, HPC User Budget Map Special Report: Machine Learning Impact on HPC Environments, 2019.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

TECHNOLOGY TRENDS: HIGH-PERFORMANCE STORAGE

With so much interest in finding value in ever-growing amounts of data, including data lakes, there has been a corresponding proliferation of technologies and solutions for managing data at scale. At the enterprise level, successful implementations evaluate a range of these technologies and implement them in proportion to specific workloads or needs.

No list of storage technologies would ever be complete. This section is intended to be an overview of the most relevant trends in high-performance enterprise segments across HPC, analytics, and machine learning, and the ways in which they drive the need for storage technologies capable of high performance at extreme scale.

Parallel and Scalable File Systems

Scientific computing applications have long demanded access to large data sets, and naturally in such cased the bandwidth for moving data from storage to server is of great importance. At the single server level, this is easy enough to determine and to monitor, but as systems scale, the problem becomes more complicated.

In conventional file system implementations—such as with NFS, the standard for UNIX or Linux—when multiple servers or nodes submit requests simultaneously, they may be routed through a common I/O server or port, sometimes referred to as a file server or filer head. As systems scale, the I/O server becomes a potential bottleneck to data access. Putting in multiple I/O servers can carry substantial cost overhead.

Many HPC users combat this by moving to a *parallel file system*. Parallel file systems allow equal, direct access to storage systems from any node in a scalable cluster, without routing traffic through an intervening I/O server. I/O bandwidth therefore scales along with the cluster.

Parallel file systems have become more common over time, although almost exclusively among scientific computing users. Carrying the benefits of parallel file systems into enterprise environments requires commercial support and reliability from a trusted enterprise vendor.

The most commonly implemented parallel file systems are Lustre, an open-source option supported professionally by high-performance solution vendors, and GPFS, developed by IBM and now available across other vendors' storage systems. In recent years, Intersect360 Research has been tracking growing adoption of BeeGFS, another open-source solution with roots at the Fraunhofer Institute in Germany; so far, BeeGFS is most commonly found in Europe, though it has the potential to grow to other geographies.[2] OneFS, a distributed, clustered file system[3] from Dell EMC, also shows up in Intersect360 Research studies, particularly in data-rich commercial environments such as entertainment and life sciences.

---

[2] Intersect360 Research, *HPC User Site Census: Middleware and Developer Tools*, 2019.

[3] Whether OneFS qualifies as a parallel file system is a matter of definition. Dell EMC does not market OneFS as a parallel file system; however, Intersect360 Research considers it to be a parallel file system and tracks it as such.

Solid-State Storage and Data Tiering

The past decade has seen the rise of solid-state storage devices, including flash drives and other types of non-volatile storage. These components got their start in consumer technologies, particularly smartphones and tablets. As enterprise-class reliability improved, the potential applicability widened.

The primary advantage of flash storage is speed. In particular, it can have higher data throughput than disk-based storage. On the other hand, it has been more expensive to implement than disk, for the same capacity. Nevertheless, it grew in popularity as analytics workloads drove a greater need for high I/O throughput. Today many organizations choose all-flash arrays for part or all of their storage environments. Some storage systems implement flash components as a burst or metadata tier.

Having storage components with different costs and performance characteristics drives the notion of *tiered storage*. For decades, organizations have maintained storage archives, usually with tape libraries. Concepts like data migration, information lifecycle management, and hierarchical storage all leveraged this idea, that active data should be kept on faster, costlier storage, whereas inactive data could be pushed to lower-performance, lower-cost archives.

Beyond tape, the evolution of flash has introduced the potential for new, faster tiers in front of disk. There are also more options for archiving than ever before. Some companies offer "warm archive" object-based libraries, which sit between disk and tape. Deep archives and remote archives take data farther away for security purposes. And all of these are only the on-premise solutions. Many organizations also have data with public cloud resource providers. This data in the cloud could be in any stage from active to archive.

In today's tiered storage implementations, an organization might have not just two tiers, but perhaps five, six, or more. In these cases, managing high-performance storage is only partly about optimizing performance to any one tier. Perhaps even more important is having data management tools for migrating data to the right tier at the right time, reducing latency in data access while optimizing for cost. Advanced solutions for enterprises are policy-based, allowing automated data tiering between flash, disk, and archives, including cloud integration capabilities.

Data Management (… and Security. And Governance. And Stewardship. And Sovereignty. And …)

Each of the above trends individually implies a greater need for data management. Placing data, migrating data, accessing data: these are all aspects of managing enterprise-wide data at scale. Parallel file systems and data migration tools are some examples of data management tools.

Moving data isn't always straightforward. There is the obvious challenge of "data gravity." It takes a lot of bandwidth and time to move substantial amounts of data, and therefore there can be a strong preference for leaving data in place when possible. "Data sovereignty"

becomes an issue in hybrid cloud environments, monitoring the control of data: its ownership, possession, and stewardship.

Some implementations, particularly those which are object-oriented, have storage operating systems that treat data with a level of abstraction; the implied overhead of this virtualization of data management can pay off if it helps move data more fluidly through a heterogeneous storage environment. Imagine a valet parking system. This doesn't make sense when nearby parking is ample, but when it is hard to find or remote, the valet can improve my experience, particularly if he can predict at all when I might return and have my car ready.

Beyond the movement or migration of data, modern storage management carries a wealth of issues in managing data through its lifecycle. This starts at inception, as organizations consider what data should be created or gathered to begin with. Will the data have value, or is it storage based simply on fear? Then there is the question of the sensitivity of the data. Who has access to it? Are there compliance or regulatory implications, as with GDPR or HIPAA? How is the data audited? Many organizations monitor the stewardship of data, archiving or deleting it as it hits certain milestones, depending on whether the organization prefers it to be kept or forgotten. Once kept, does the data need to be curated or prepared for further learning? Details like these form the crux of data implementation issues for forward-looking businesses as they strive to combine enterprise storage with performance and scalability.

The role of data management itself continues to evolve. This is an area in which the applications themselves might improve how they are run. Simulations of data flows and predictive analytics may soon be applied in order to better place and migrate data. The future is clear to bring more advances.

Successful implementations will come from vendors that can span a wide range of technologies, maintain enterprise levels of support and reliability, and keep an eye to future innovations in data management. Perhaps most importantly, with such a wide range of potential solutions to bring to bear against a wide and evolving range of challenges, the provider should be able to recommend which combination of solutions is best suited to a particular set of workloads.

Will this data have value? Who has access to it? Are there regulatory implications? How is the data audited? Once kept, does the data need to be curated for further learning?

Details like these form the crux of data implementation issues for forward-looking businesses as they strive to combine enterprise storage with performance and scalability.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

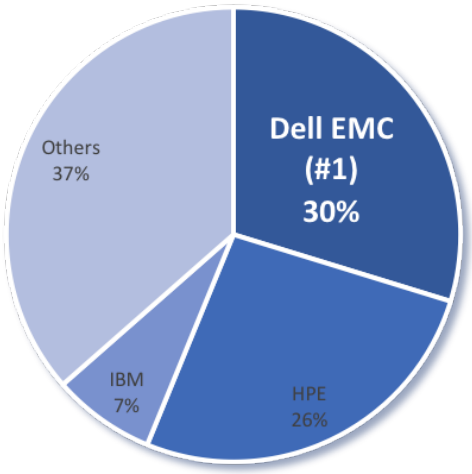
HIGH-PERFORMANCE STORAGE SOLUTIONS FROM DELL EMC

With trusted products across both computation and data management, Dell EMC is the industry leader in both storage for HPC environments and total HPC solution revenue.[4] (See figure below.) Dell EMC leverages this breadth of offerings with converged solutions that incorporate HPC, data analytics, and AI.[5]

Dell EMC's leadership across high-performance data management comes from the company's long heritage in enterprise storage, its breadth of solutions in both hardware and software, and its services in helping to optimize environments for maximum value. When Dell completed its acquisition of EMC in 2016, it immediately became the storage leader in technical environments, and as the market continued to evolve, Dell EMC grew its position to lead in total solutions as well.

Share of Combined Worldwide HPC Server and Storage Revenue, 2018

Intersect360 Research, 2019



Dell's growth has not simply been organic. As described above, the HPC industry has gone through significant changes over the past decade, specifically in the incorporation of new types of applications, and the need for technologies to address these expanded workflows in mixed enterprise environments. This section gives an overview of some of the products and services Dell EMC offers for HPC, analytics, and AI.

Dell EMC PowerVault ME4 Series

When a business is ready to deploy HPC technologies to help streamline data workflows, it would help to understand that the storage infrastructure objectives may be different—with some expected overlap—than for general-purpose, transactional workloads. For example,

[4] Intersect360 Research, "Vendor Overview and Outlook: Dell EMC in HPC," 2019.

[5] https://www.dellemc.com/en-us/collaterals/unauth/brochures/solutions/hpc_ai_convergence_brochure.pdf.

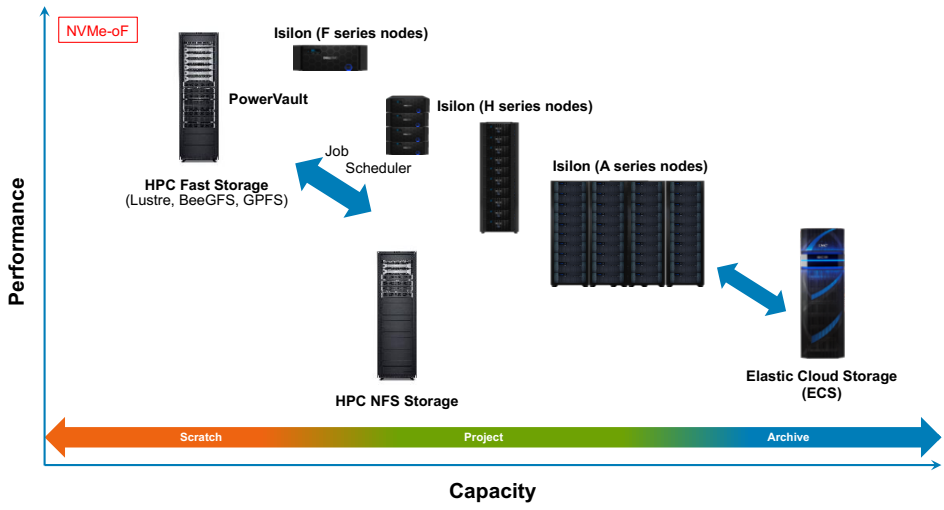
instead of an enterprise focus on I/O throughput, smaller file sizes, random file access and mixed read/write concentrations, HPC batch-oriented workloads may be oriented toward high bandwidth, larger files, sequential data access, and far more writes than reads, all within a cluster environment.

For these high-performance enterprise implementations, the workhorse of Dell EMC portfolio is the PowerVault ME4 series, optimized for either direct-attach storage (DAS) or storage area network (SAN) deployments. PowerVault is positioned to deliver many of the HPC infrastructure needs described above; it has a 12 Gigabit/second SAS back end, and it is capable of 7.0 GB/sec for reads and 5.5 GB/sec for writes. PowerVault can be flexibly configured with or without a parallel file system, including optimizations of Lustre in conjunctions with Dell PowerEdge servers.

For data protection, the ME4 series incorporates ADAPT, Autonomic Distributed Allocation Protection Technology, Dell EMC's unique erasure coding solution that reduces drive rebuild times from traditional RAID deployments.[6] ADAPT provides another example of how Dell EMC aims to combine greater performance with enterprise reliability in a tiered-storage architecture.

Dell EMC High-Performance Computing Storage Portfolio

Source: Dell EMC



[6] https://www.dellemc.com/en-us/collaterals/unauth/white-papers/products/storage/powervault-me4-series-adapt-software-white-paper.pdf

Dell EMC Isilon

Isilon has been a provider of storage solutions to technical users since well before its acquisitions by EMC and then Dell. With the high-throughput OneFS file system, Isilon storage has long sold well in high-performance, data-intensive areas, such as entertainment and life sciences.

The Dell EMC Isilon Scale-Out NAS solution scales to tens of petabytes in a single file system and namespace, with hundreds of GB/sec aggregate throughput for data reads and writes, extending its appeal in an era of data explosion. Moreover, the policy-based auto-tiering capabilities of OneFS enable automated data migration between hot SSD tiers and SATA archives, including compatibility with public cloud providers. Natively implemented front-end client access—through NFS (for UNIX or Linux), CIFS (for Windows), HDFS (for Hadoop), or object-based, public cloud tools—serves any common application environment.

Addressing the sudden rise in both machine learning and analytics, Dell EMC Isilon offers all-flash array (AFA) options, optimizing for maximum data throughput. The Isilon all-flash solution can support hundreds of thousands of concurrent connections for parallel applications at scale.

Under Dell EMC, the Isilon Scale-Out NAS product has added enterprise features to OneFS for data management and compliance. Data governance and compliance, for regulations such as GDPR or HIPAA, are supported by features such as WORM (write once, read many), access zones for multi-tenancy, and role-based administration and auditing. Isilon is architected to run analytics on data in-place, while conforming to standard enterprise protocols like NFS and HDFS. Isilon also includes tools for organizations designing their own analytics processes and enables the creation of analytics as a service. With this combination of enterprise features and performance at scale, Isilon targets workloads such as analytics and long-term storage of sensitive personal data, such as in healthcare or financial services.

ECS Scale-Out Object Store

For users who are ready to migrate away from tape, the ECS Scale-Out Object Store offers a higher-performance, disk-based alternative, while leaving existing applications and workflows intact, with multi-protocol support. Unlike some tape implementations, ECS is always online, matching well to workflows with streaming or continuous data flows, such as real-time analytics.

ECS deploys scalability, from petabytes to exabytes, delivering cloud-like scalability with on-premise efficiency. Cloud-native applications can be run on ECS with accelerated performance, removing internet latency. Furthermore, ECS can be deployed globally, with geo-distributed data that is architected to be accessible locally for performance, yet isolated for data protection, and administrated and accessed through a single namespace.

Across the broader market, an opportunity exists for a storage vendor that can provide a full range of scalable, high-performance solutions, backed by enterprise features and reliability. Dell EMC is such a company.

Dell EMC's storage portfolio is positioned to serve the intersection of HPC, analytics, and AI, combining high-performance scalability with enterprise data protection and regulatory compliance. This will become the ultimate test in turning data into value.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

INTERSECT360 RESEARCH ANALYSIS

It is hard to find an enterprise-class user that has not been touched by major industry trends related to the continued influx of data. Even if a company does not identify as "HPC," it may find it is influenced by high-performance analytics, edge computing, machine learning, or cloud. Managing data at scale has become a preponderant theme of enterprise computing that will continue through the 2020s.

The most important aspect of these trends is that they are converging. AI, analytics, and HPC are not independent markets at all, but rather interdependent approaches that coexist across an organization. For most users, it is impractical to consider buying a scalable solution for scientific computing, separate from a high-throughput solution for analytics, separate from an object store for archiving, separate from an appliance for machine learning. As the market continues to evolve, Intersect360 Research expects trends to continue toward consolidation. With an increased emphasis on tiering, we predict that object-based storage will continue to expand in enterprise environments.

Point solutions certainly exist for all of these specialties, but across the broader market, a larger opportunity exists for a storage vendor that can provide a full range of scalable, high-performance solutions, backed by enterprise features and reliability. Dell EMC is such a company, with a broad portfolio of product across HPC, analytics, and AI. Dell EMC is a leader in storage, computing, and total solutions for high-performance workloads. Its offerings across Dell EMC Isilon, ECS, and PowerVault all play in HPC directly, and even other enterprise products such as Dell EMC Unity XT and PowerMax find their way into supported end-to-end solutions.

Dell EMC's storage portfolio is positioned to serve the intersection of HPC, analytics, and AI, combining high-performance scalability with enterprise data protection and regulatory compliance. Dell EMC not only delivers a wide range of optimized products, but more importantly implement intelligent data management solutions that migrate data automatically and on-demand between higher-performance storage elements and economical long-term archive, while maintaining accessibility enterprise-wide for users and applications.

This will become the ultimate test in turning data into value.

For more information about Dell EMC solutions for HPC, analytics, and AI, visit:

- Isilon: https://www.dellemc.com/en-us/storage/isilon/index.htm.
- PowerVault ME4: https://www.dellemc.com/en-us/storage/powervaultme4.htm.
- ECS: https://www.dellemc.com/en-us/storage/ecs/index.htm.

P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com