White Paper

EMC ISILON ONEFS OPERATING SYSTEM

Powering the Isilon scale-out storage platform

Abstract

This white paper provides an introduction to the EMC® Isilon® OneFS® operating system, the foundation of the Isilon scale-out storage platform. You'll gain an overview of the architecture of OneFS as well as the benefits of a scale-out storage platform.

April 2011

Part Number h8202

Table of Contents

Executive summary

Communications bandwidth and computing power are at an all-time high, and following Moore's law, these capabilities will continue to double every 12 to 18 months. Technology is truly evolving and many constructs that were relevant 20 years ago are museum history pieces today. In turn, today's cutting edge technology will be relegated to the same status 20 years from now. For those of us who have participated in the growth and development of technology over the last several years, it's easy to forget the scale and speed at which it is evolving.

There is no area where this acceleration is more apparent than with enterprise storage and it mirrors the changes in computing and bandwidth. Spinning media and basic containers of storage have existed almost as long as the microprocessor, but things are quickly changing. The advent of Flash-based memory technology, high-speed, low-latency networking, ubiquitous industry standards, and open-source software give us the tools to create a fresh paradigm. The ever-increasing demands of businesses and consumers provide the need and motivation.

Current enterprise storage platforms have served their purpose; network-attached storage systems (NAS) and storage area networks (SAN) have been a crucial, simplifying step over direct-attached storage (DAS). However, in essence SAN and NAS platforms are nothing more than an evolutionary step from DAS and, ultimately, have only extended the lifespan of DAS storage techniques. A decade ago, SAN and NAS began to replace DAS as the standard for enterprise storage—but both of these technologies were designed nearly 20 years ago.

The basic storage container (as well as the SAN and NAS technologies that extend it) espouses a fundamental architectural flaw; it is static and wasn't designed for scalability. Traditional data protection techniques, such as mirroring and RAID, can't continue to scale. The storage model of moving to larger and more powerful equipment can't meet the growing capacity and performance requirements without incurring large costs and/or increasingly putting data at risk. Finally, the organizational and staffing reality of managing ever-increasing storage complexity is not sustainable.

The most basic of technologies, volume, is the clearest sign that a new storage paradigm is needed. For purely organizational purposes, there are very few reasons to constrain data to a specific location. Assuming proper security and reliability, individuals and applications should be able to access files and folders from any geography as easily as if the data was local. As data sets grow, so do users' needs. Administrators can no longer afford to spend precious time moving data between containers, reorganizing workflows and shares, and dealing with more and more complexity due to the technical constraints of traditional storage.

Another significant challenge in the SAN or NAS storage environment is waste; many storage vendors have determined that up to 50 percent of their storage is going unused. While this helps the bottom line for storage vendors, it results in wasted

power, cooling, and management. When volumes are limited in size due to reliability concerns, technical constraints, or performance capabilities, they ultimately cause inefficiencies to occur. These fixed sets of resources cannot be maximized to their full potential without constant tuning, oversight, and workflow changes—taking staff time away from more strategic and critical tasks in the data center.

Data protection techniques such as mirroring and RAID are designed to reduce complexity in storage media and provide reliability through redundancy. This key component of traditional storage is now faced with a huge scalability hurdle in that the amount of media and file-based data is expanding far more rapidly than the speed at which it can be accessed. RAID can be effective only if data can be reconstructed before another failure occurs and at best today's technologies can absorb two simultaneous failures. As the length of time increases to repair these failures, so does the probability that another failure will occur. With 3 terabyte (TB) drives on the market, 4 TB drives just around the corner, and 6 TB drives on the horizon, RAID for data protection is not sustainable and a new approach is necessary.

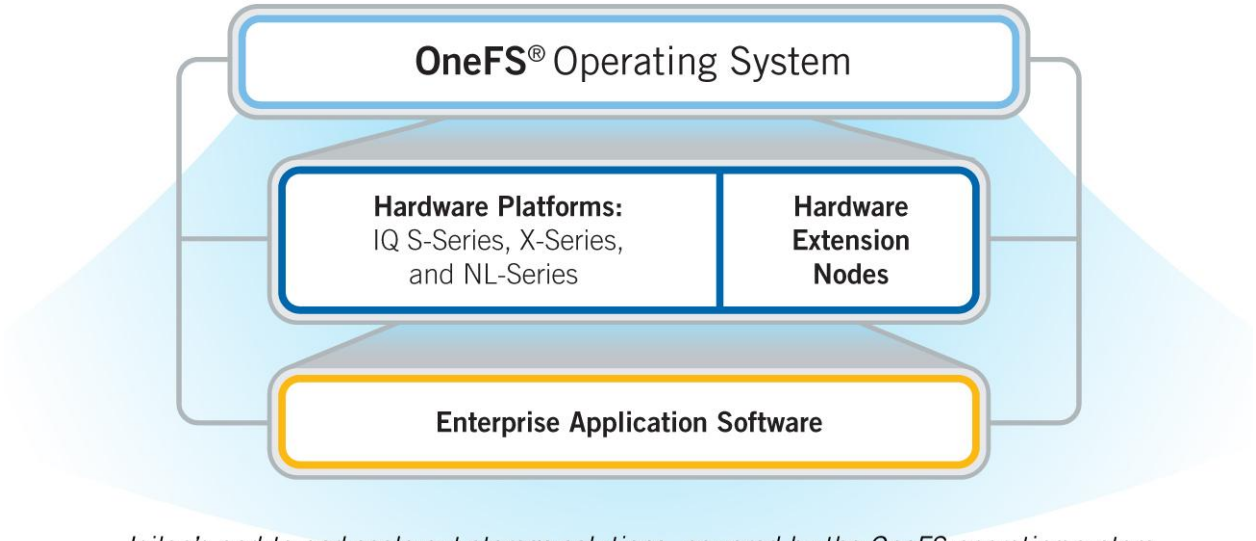
In most enterprise storage environments, RAID is supplemented by mirroring technologies that are in-system, in the data center, or at another location. This simple, brute-force technique of duplicating data solves one availability challenge but isn't designed to scale. Replicating data within systems or within a single data center quickly becomes cost-prohibitive in terms of the additional capacity consumed, processing overhead, network bandwidth, and operational costs. In addition, the data placed into storage systems is inherently dynamic—what is relevant or mission-critical today, won't be tomorrow. Organizations must have the ability to align protection mechanisms to the value of the data and to easily adjust protection as priorities and needs dictate.

In the world of Moore's law, achieving high performance in traditional storage systems is also a significant challenge. Increasing performance by replacing full systems is an extremely expensive and disruptive proposition. Purchasing more equipment than necessary, consuming more power than needed, and taking up valuable data center floor space merely to avoid significant upgrade steps in the future is not a sustainable model. Micro-managing performance by creating specific volumes and constantly tweaking and adjusting those volumes may be manageable in low numbers, but at scale it is an impossible task. With these storage challenges administrators simply can't meet required service levels due to the dynamic nature of modern workloads.

The 20-year-old technology and design decisions that SAN and NAS storage systems are built on cannot scale to meet the challenges of today's enterprises. A new storage platform is required—one that is designed specifically to scale when needed. Scaling a storage platform must occur in a cost-effective, nondisruptive, predictable fashion—and it must continue to be simple to manage and administer as it grows in capacity and performance. This scalability design has become known as "scale-out." It provides the ability to seamlessly grow an environment without bounds and without complexity—and this is the age of the scale-out imperative.

The Isilon OneFS operating system

In 2000, seeing the challenges with traditional storage architectures and the pace at which file-based data was increasing, the founders of Isilon began work on a revolutionary new storage architecture, the OneFS® operating system. The most important design choice and fundamental difference of EMC® Isilon® storage is that with OneFS the storage system does not rely on hardware as a critical part of the storage architecture. Rather OneFS (see Figure 1) combines the three layers of traditional storage architectures—file system, volume manager, and RAID—into one unified software layer, creating a single intelligent file system that spans all nodes within a storage system. Isilon scale-out storage provides the appliance hardware base on which OneFS executes. While the hardware is commodity, enterprise-quality components produced by manufacturers, such as Intel, Hitachi, and Super Micro, nearly all aspects of the storage system are provided in software, by OneFS. The hardware includes a high-speed battery-backed NVRAM journal and a 20 Gb/s point-to-point microsecond-latency interconnect, Infiniband. On this commodity hardware base, the OneFS operating system enables data protection and automated data balancing and migration, as well as the ability to seamlessly add storage and performance capabilities without system downtime.



Isilon's end-to-end scale-out storage solutions, powered by the OneFS operating system, give users a broad range of options to meet their specific storage needs.

Figure 1. OneFS operating system: Powering the Isilon scale-out storage platform

OneFS works exclusively with the Isilon scale-out storage system, referred to as a "cluster." A single Isilon cluster consists of multiple storage "nodes," which are constructed as rackmountable enterprise appliances containing memory, CPU, networking, NVRAM, Infiniband, and storage media. An Isilon cluster starts with as few as three nodes and can scale out as high as 144 nodes. At the time of publication, an Isilon scale-out storage system's total single file system capacity ranges from a minimum of 6 TB to a maximum of 10 petabytes (PB). Each node

added to a cluster increases aggregate disk, cache, CPU, and network capacity. As a result of this aggregate increase, a 144-node cluster can access as much as 13.8 TB of globally coherent, shared cache. With capacity and performance delivered in a single storage system, a single file system, and a single volume, the complexity of the system and management time for the storage administrator does not increase as the system scales.

OneFS stripes data across all storage nodes in a cluster. As data is sent from client machines to the cluster (using industry-standard protocols, such as NFS and CIFS), OneFS automatically divides the content and allocates it to different storages nodes in parallel. This occurs on the private Infiniband network, which eliminates unnecessary network traffic. The Isilon cluster is managed as a single file system and the coordination and data distribution are completely transparent to end-user clients. When a client wishes to read a file, OneFS will retrieve the appropriate blocks from multiple storage nodes in parallel, automatically recombining the file, and the initiating client sees exactly what was originally written. This ability to automatically distribute data across multiple nodes in a transparent manner is fundamental for the ability of OneFS to enable growth, next-generation data protection, and extreme performance.

Scalability

In contrast to traditional storage systems that must "scale up" when additional performance or capacity is needed, OneFS enables an Isilon storage system to "scale out," seamlessly increasing the existing file system or volume into petabytes of capacity. In addition, with the flexibility of OneFS, different node types can be mixed in a single cluster or "pool," through the addition of the SmartPools™ application software, providing investment protection and eliminating the need for "forklift" upgrades when different capacity or performance levels are needed. SmartPools (see Figure 2) enables businesses and storage administrators to easily deploy a single file system to span multiple tiers of performance and capacity. This single file system automatically adapts to business data and application workflows over time.

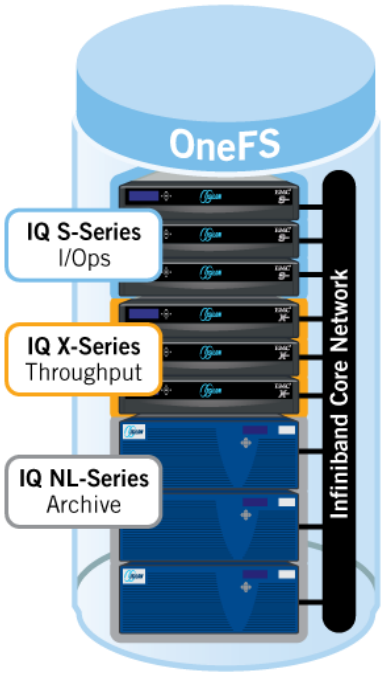


Figure 2. SmartPools single file system for multiple tiers with automated, transparent data movement

In addition to tiering data automatically across different nodes, SmartPools can also use solid state drives (SSDs) as a tier for metadata and file-based storage workflows. SSDs as a tier can be used within a pool to improve metadata or data access performance, or the SSDs in one tier can be leveraged to hold the metadata of files on other tiers—accelerating the performance even of nodes that have no SSDs.

Adding capacity and performance capabilities to an Isilon cluster is significantly easier than with other storage

systems—requiring only three simple steps for the storage administrator: adding another node into the rack, attaching the node to the Infiniband network, and instructing the cluster to add the additional node. The new node provides additional capacity and performance since each node includes CPU, memory, and network. The Autobalance™ feature of OneFS will automatically move data across the Infiniband network in an automatic, coherent manner so existing data that resides on the cluster moves onto this new storage node. This automatic rebalancing ensures the new node will not become a hot spot for new data and that existing data is able to gain the benefits of a more powerful storage system. The Autobalance feature of OneFS is also completely transparent to the end user and can be adjusted to minimize impact on high-performance workloads. This capability alone allows OneFS to scale transparently, on-the-fly, from 18 TB up to 10 PB with no added management time for the administrator, nor increased complexity within the storage system.

Allocating data with a single, scalable pool of storage is an often understated benefit and added efficiency found in a single file system. Forcing administrators to coerce users into choosing the volumes that have the requisite amount of free space or manually moving data is time-consuming and inefficient. If the users choose incorrectly, the performance demands of a particular workflow cannot be satisfied by a particular volume. Also, if the organization cannot address a particular volume, or if the storage administrator cannot move data transparently and quickly, then storage efficiency will be sub-optimal. Industry analysis of storage deployments suggests that on average 43 percent of storage capacity is wasted due to these inefficiencies. An Isilon scale-out storage system has no such constraints—it operates with storage efficiencies typically in excess of 80 percent.

Performance

A large-scale storage system must provide the performance required for a variety of workflows, whether they be sequential, concurrent, or random. Different workflows will exist between applications and within individual applications. OneFS provides for all of these needs simultaneously with intelligent software. More importantly, with OneFS (see Figure 3), throughput and IOPS scale linearly with the number of nodes present in a single system. Due to balanced data distribution, automatic rebalancing, and distributed processing, OneFS is able to leverage additional CPUs, network ports, and memory as the system scales.

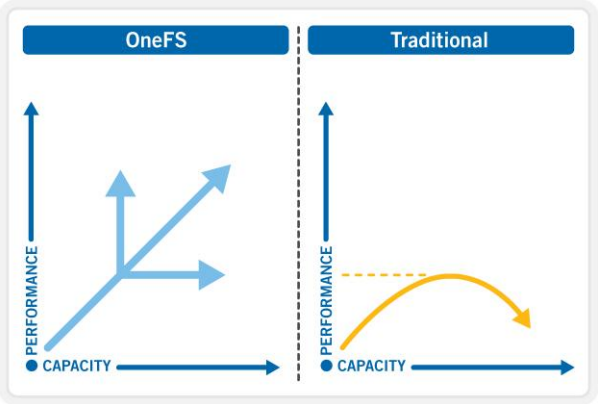


Figure 3. OneFS linear scalability

In order to fully exploit locality and meet the needs of various workflows, OneFS provides a globally accessible and coherent cache across all nodes. Each storage node contains standard DRAM (between 6 and 96 GB) and this memory is primarily used to cache data that has been placed on that particular storage node and is actively being accessed. This cache grows as more nodes are added to a cluster, allowing an increasing working set to continually remain in cache and up to 13.8 TB can be cached in a single system. In addition, OneFS allows the storage system administrator to specify the type of workload on a per-file or per-directory basis, indicating whether the access pattern to a particular file/directory is random, concurrent, or sequential. This unique capability allows OneFS to tailor on-disk layout decisions, cache-retention policies, and data pre-fetch policies in order to maximize performance of individual workflows.

Management

As organizations face more data and more management complexity, they are offered a wider variety of potential solutions. The emphasis for the next-generation data center is meeting customer requirements in a sustainable, scalable, and efficient fashion and the key to success is reducing management complexity. Human capital, traditionally measured by “Operating Expense” (or “OpEx”), must be leveraged to focus on the activities that enable a business to do more to improve its productivity, resourcefulness, and ultimately, bottom line.

Traditional storage systems require lengthy planning, upgrade, and maintenance activities. Trivial tasks, such as increasing capacity, scaling performance, and adding additional users, often require horizontal scaling and reconfiguring applications, and result in a disruption of user activities and ultimately lost productivity and revenue.

OneFS has been designed to simplify administration activities and maintain this simplicity as the overall system scales, as shown in Figure 4. The ability to add performance and/or capacity in 60 seconds with an Isilon node, avoid manual data and connection rebalancing with SmartConnect™ and Autobalance, and mitigate hardware and software upgrades is uniquely enabled by OneFS.

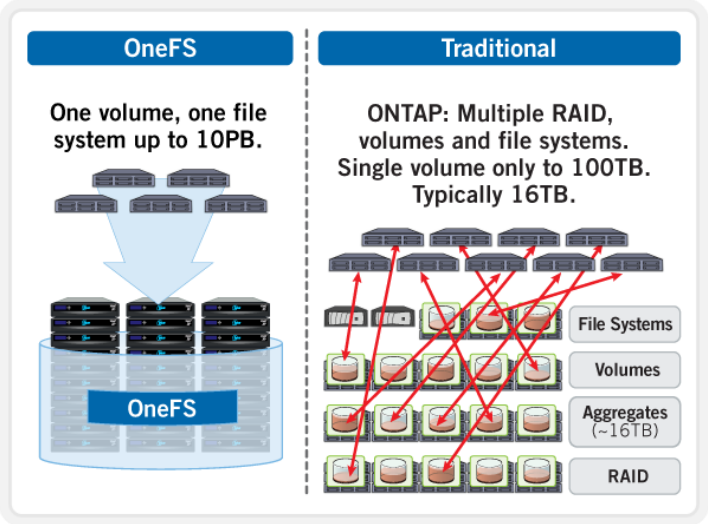


Figure 4. OneFS simplicity vs. complexity of traditional NAS

Data protection

As traditional storage systems scale, techniques that were appropriate at a small size become inadequate at a larger size, and there is no better example of this than RAID. RAID can be effective only if the data can be reconstructed before another failure can occur. However, as the amount of data increases, the speed to access that data does not and the probability of additional failures continues to increase. OneFS does not depend on hardware-based RAID technologies to provide data protection. Instead, OneFS includes a core technology, FlexProtect™, which is built on solid mathematical constructs and utilizes Reed-Solomon encodings to provide redundancy and availability. FlexProtect provides protection for up to four simultaneous failures of either full nodes or individual drives and as the cluster scales in size, FlexProtect delivers on the need to ensure minimal reconstruction time for an individual failure.

FlexProtect is a key innovation in OneFS and takes a file-specific approach toward data protection, storing protection information for each file independently. This independent protection allows protection data to be dispersed throughout the cluster (see Figure 5) along with the file data—dramatically increasing the potential parallelism for access and reconstruction when required. When there is a failure of a node or drive in an Isilon storage system, FlexProtect is able to identify which portions of files are affected by the failure and employs multiple nodes to participate in the reconstruction of only the affected files. Since the Autobalance feature in OneFS spreads files out across the cluster, the number of spindles and CPUs available for reconstruction far exceeds what would be found in a typical hardware RAID implementation. In addition, FlexProtect doesn't need to reconstruct data back to a single spare drive (which with RAID creates an unavoidable bottleneck); instead the file data is reconstructed in available space, providing a virtual "hot spare."

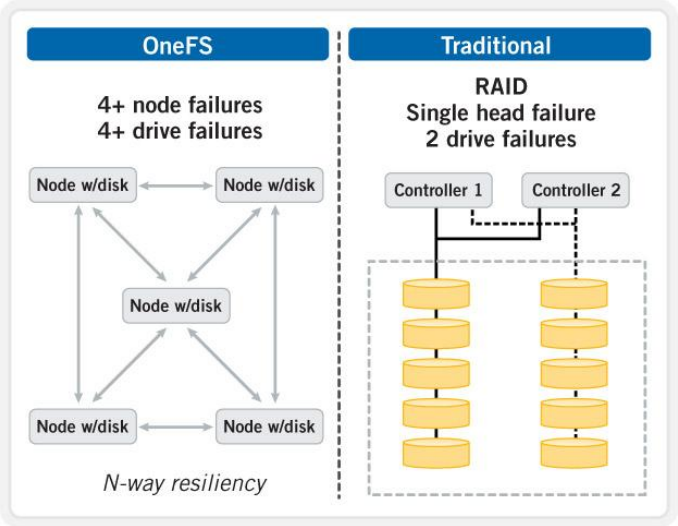


Figure 5. OneFS N+4 data protection

OneFS constantly monitors the health of all files and disks within the cluster and if components are at risk, the file system automatically flags the problem components for replacement, transparently reallocating those files to healthy components. OneFS also ensures data integrity if the file system has an unexpected failure during a write operation. Each write operation is transactionally committed to the NVRAM journal to protect against node or cluster failure. In the case of a write failure, the journal enables a node to rejoin the cluster quickly, without the need for a file system consistency check. With no single point of failure, the file system is also transactionally safe in the event of an NVRAM failure.

Since the FlexProtect feature in OneFS is file-aware, it also provides file-specific protection capabilities. An individual file (or more typically, a directory) can be given a specific protection level and different portions of the file system to be protected at levels aligned to the importance of the data or workflow. Critical data can be protected at a higher level whereas less critical data can be protected at a lower level. This provides storage administrators with a very granular protection/capacity trade-off that can be adjusted dynamically as a cluster scales and a workflow ages.

Conclusion

Scalability, performance, ease of management, and data protection are critical in a storage system that can meet user needs and the ongoing challenges of the data center in the world of Moore's law. With OneFS organizations and administrators can scale from as little as 18 TB to as high as 10 PB within a single file system, single volume, with a single point of administration. OneFS delivers high performance, high throughput, or both, without adding management complexity.

Next-generation data centers must be built for sustainable scalability. They will harness the power of automation, leverage the commoditization of hardware, ensure the full consumption of the network fabric, and provide maximum flexibility for organizations intent on satisfying an ever-changing set of requirements.

OneFS is the next-generation file system designed to meet these challenges.

Figure 6. The OneFS operating system and suite of scale-out storage enterprise software applications