



Uncompromising Reliability through Clustered Storage

Delivering Highly Available Clustered Storage Systems

An Isilon Systems Technical Whitepaper

March 2006

Table of Contents

1	Introduction.....	3
2	Storage Challenges with High Availability and Reliability	4
3	Isilon IQ: The Highest Available and Most Advanced Data Protection Solution	6
4	Conclusion	10

Figures

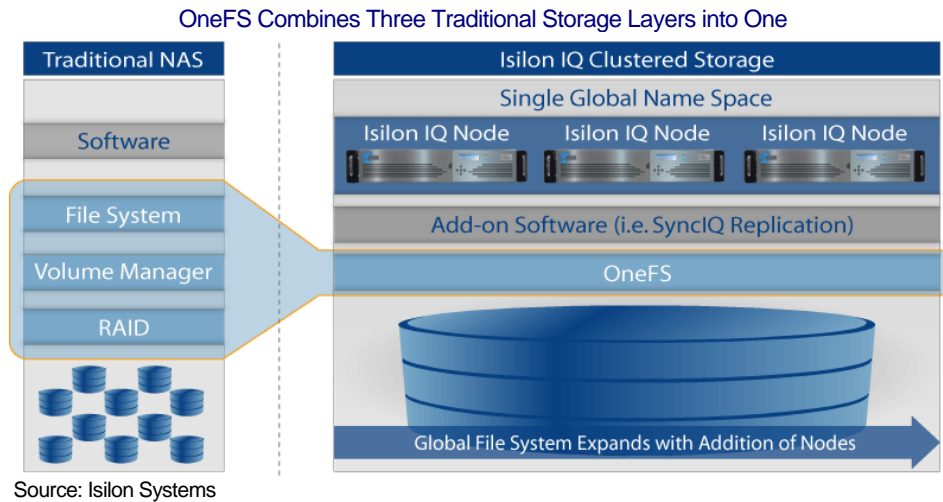
<i>OneFS Combines Three Traditional Storage Layers into One.....</i>	<i>3</i>
<i>Traditional Storage.....</i>	<i>4</i>
<i>Areal Density.....</i>	<i>5</i>
<i>Isilon IQ Network Architecture.....</i>	<i>6</i>
<i>OneFS “n+1” and “n+2” Data Protection Schemes.....</i>	<i>7</i>
<i>Average Disk Drive Rebuild Time (hrs).....</i>	<i>8</i>

1 Introduction

Today, companies of all sizes are dealing with an avalanche of digital content, unstructured data and reference information that is driving an unprecedented increase in storage needs. Corporations dealing with huge amounts of data being produced on a daily basis have a desire to move to clustered architectures based on industry-standard hardware to streamline their storage processes and lower their costs. Clustered network storage enables companies to gain significant advantages in scalability, performance and cost, but one key problem remains: *reliability*.

In designing a clustered storage system that is able to achieve multi-hundred terabyte single file systems spanning thousands of disks, and tens or even hundreds of nodes — all of which are potentially points of failure — new technologies must be developed that will also ensure enterprise-class reliability. This whitepaper describes the limitations of traditional data protection technologies and examines Isilon® IQ's revolutionary new approach to ensuring data integrity and availability in large clustered storage environments.

Isilon's OneFS® is a patent-pending distributed file system software that provides the intelligence behind the award-winning Isilon IQ family of clustered storage systems. It combines the three layers of traditional storage architectures — file system, volume manager and RAID — into one unified software layer, creating a single intelligent symmetric file system that spans all nodes within a cluster. OneFS provides a single point of management for large content stores, faster access to large content files, inherent high availability, and the ability to easily scale a single cluster's capacity — up to 7 Gigabytes per second of total throughput and hundreds of terabytes of capacity, all from a single network file system.



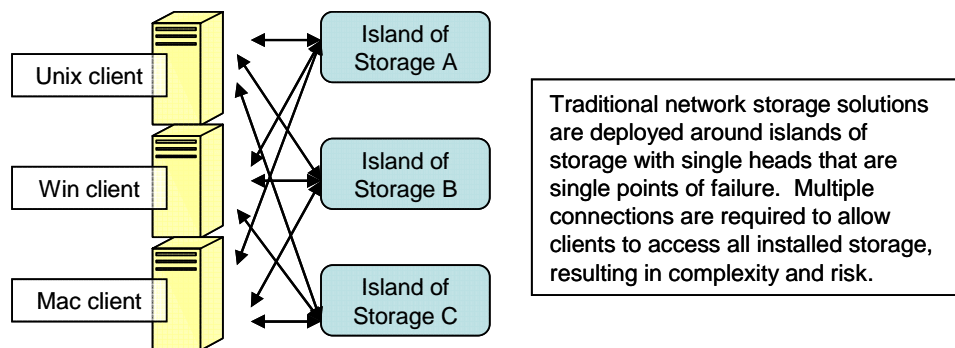
2 Storage Challenges with High Availability and Reliability

Customers today are forced to deal with a rapidly increasing amount of data that must be stored reliably and made available constantly to lots of applications and users. There are fundamental limitations and problems that traditional storage systems present to customers when it comes to high availability and reliability of their mission critical data, and to reliably scale traditional architectures, IT managers are faced with a myriad of technical challenges and complexities, which ultimately leads to higher operating costs.

Single Points of Failure

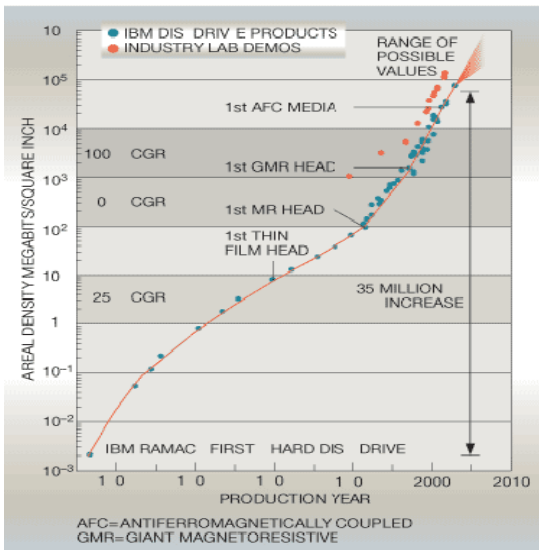
With traditional storage — both NAS and SAN architectures — there are inherent single points of failure. Typically, these systems are deployed with a single server (or head unit) that sits in front of a large amount of disk storage space. If these servers go down for any reason, all access to the disk storage and data behind them is off-line. Over time, these storage systems become even larger “islands of storage” as more capacity is required, exacerbating the risk of data not being available in the event of a server head going down. To address this problem, some vendors have implemented what they term “simple clustering”, which allows for a pair of server heads to be installed for failover. However these solutions do not scale well and are costly, as customers must purchase redundant head units and expensive clustering/failover software add-ons, resulting in a solution that is both expensive and complex to manage.

Traditional Network Storage



Data Protection Schemes with RAID and Single Parity / Drive Rebuild Time

Traditional single parity RAID technology, including RAID4 and RAID5, have long been the norm for storage administrators and offer data protection from a single failed disk drive. These solutions have been acceptable for enterprises using pools of storage with limited amounts of capacity and relatively small volume and file system sizes who are willing to take on the added expense in management and administrative costs to achieve the high availability. The caveat to the protection offered by traditional RAID technology, though, is that no other disk or unacceptable bit error can occur during a read operation while reconstruction of the failed disk is in progress. In today's world of large disk drives, the probability that a secondary failure event will occur has increased dramatically. Areal density (amount of written information on the disk's surface in bits per square inch) is experiencing nearly 100% compound annual growth rate, but the reliability and performance of disk drives is not increasing at the same pace, in part due to the time it takes to rebuild drives. Large-capacity drives, such as the 250GB and 400GB SATA drives, require much longer drive rebuild times, significantly increasing the probability of a multi-failure scenario which almost always results in data loss. Dense SATA disk rebuild times can take up to 24 hours, creating significant windows of risk for multi-failures. Hence the ability of traditional single parity RAID to protect data is being stretched past its limits in the modern data center.



Areal density. Amount of written information on the disks surface in bits per square inch. The diagram shows the areal density improvement for hard disk drives since 1956. Today's CGR (compound growth rate) is essentially 100 percent or doubling every year.

Source: IBM Systems Journal, Volume 42, Number 2, 2003

Detection of Failing Components

Many traditional storage systems today are re-active, not proactive, when it comes to finding and responding to problems with failing hardware components. Without having predictive software intelligence, traditional storage systems are required to put their customer's data at risk because they can not pre-determine when a failure is going to happen. As a result, these systems go through long data reconstruction times after drives fail, during which time there is a heightened risk of data loss if a secondary drive were to fail, instead of taking steps to prevent failures that require data reconstruction before they occur. Because of the increased densities in modern disk drives, the likelihood of error rates has increased, which leads to a critical need to proactively address such errors before they happen in order to ensure a highly available storage system.

A New Solution is Needed

A new solution and architecture is required to deliver maximum data protection and high availability for today's storage trends. This solution must be able to overcome the inherent single points of failures at a cost effective price point while also being easy to operate and manage. In addition, to embrace the new trends of huge single file systems and more dense disk drives, this solution must be able to set a new bar for rebuilding data when these components fail while also having the intelligence to detect failures and provide a "self heal" before such failures occur.

3 Isilon IQ: The Highest Available and Most Advanced Data Protection Solution

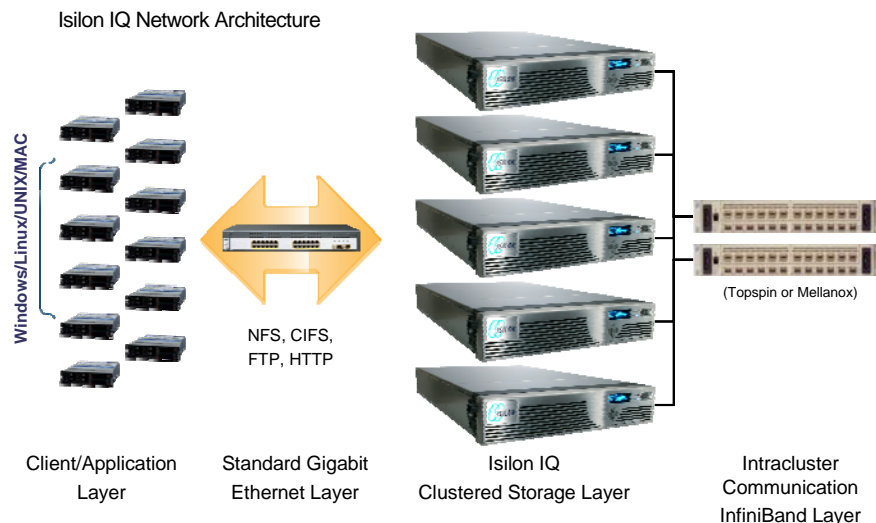
In designing a clustered storage solution, Isilon has focused on solving the key limitations found in traditional storage systems and is setting the bar for reliability by delivering a state-of-the-art clustered storage solution that has:

- ☑ No single point of failure
- ☑ Tolerance for multi-failure scenarios
- ☑ Fastest disk rebuild times in the industry
- ☑ Pro-active failure detection and pre-emptive drive rebuilds
- ☑ Fully journalled file system
- ☑ High transient availability

No Single Point of Failure

Each Isilon IQ cluster consists of anywhere from three to forty-two Isilon IQ nodes. Each modular, self-contained Isilon IQ node contains disk capacity along with a powerful storage server, CPU, memory and network, all in a self-contained, compact, 2U rack-mountable system. As additional Isilon IQ nodes are added to a cluster, all aspects of the cluster scale symmetrically, including capacity, throughput, memory, CPU and network connectivity. Isilon IQ nodes automatically work together, harnessing their collective power into a single unified storage system that is tolerant of the failure of ANY piece of hardware, including disks, switches or even entire nodes.

In a fully distributed architecture, it is critical for each node to stay in sync with all other nodes in the cluster. Isilon IQ storage nodes use either Gigabit Ethernet or high-speed, low-latency Infiniband switching fabric for inter-cluster communication, synchronization and all intracluster operations. This enables each node to share information with every other node on the system, so that each storage node acts as a fully coherent peer with complete understanding of what the other nodes are doing. If any one node were to go down, any other node could fill in, thereby eliminating any single point of failure.



Smart software is at the heart of this reliable architecture. OneFS, Isilon's patent-pending distributed file system, provides the core intelligence that powers Isilon IQ. OneFS keeps nodes synchronized by using a distributed lock manager, coherent caching and a remote block manager that maintains global coherency throughout the entire cluster. It is this global coherency through each node that eliminates any single point of failure for access to the file system. Any node in the cluster can take a write or read

request and each node presents the same unified view of the entire file system. All nodes in the cluster are “peers”, so the system is fully symmetric, eliminating hierarchy and inherent bottlenecks.

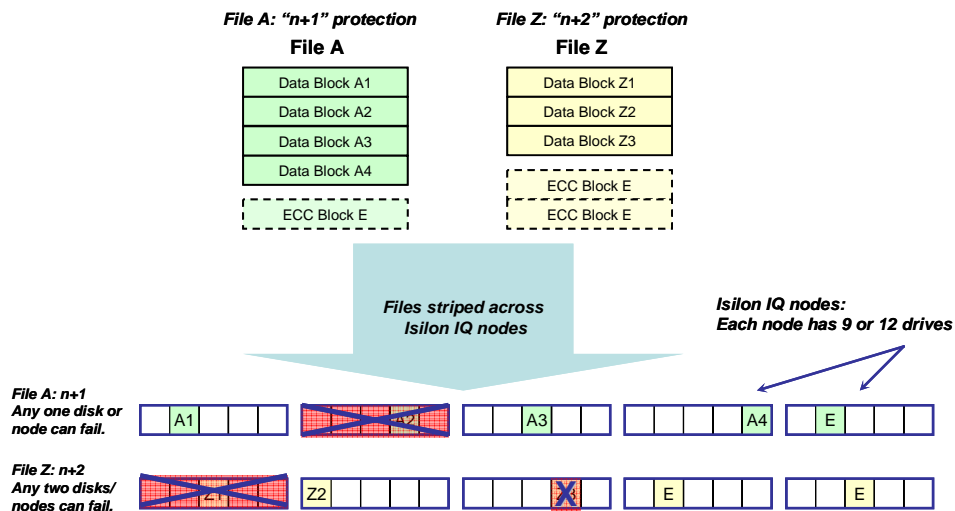
Multi Failure Support

As discussed above, the increase in modern disk densities has created a concomitant increase in the likelihood of multiple failure scenarios and pushed traditional RAID schemes to their limits. However, with Isilon IQ, customers can withstand the loss of multiple disks or nodes without losing access to any content. OneFS’s unique FlexProtect-AP feature utilizes Reed Solomon “n+1 and n+2” ECC (error correction code), parity striping and mirrored file striping (from 2x to 8x) that spans multiple nodes within a cluster. These policies can be set at any level, including cluster, directory, sub-directory, or even individual file level. Additionally, these policies can be changed at any time from a simple WebUI — even while the system is in production and fully available. Because all files are striped across multiple nodes within a cluster, no single node stores 100% of a file; if a node fails, all other nodes in the cluster can deliver 100% of the files within that cluster.

Specifically, “n+2” double ECC error correction allows for multiple failures of disks or even nodes within a single cluster and file system. Each file is striped across multiple nodes within a cluster, with two parity stripes for each data block. Unlike the “n+1” single parity, if a second failure were to occur during this rebuild, all data would still be fully available because the data was originally striped with double ECC error protection. In contrast, the same scenario in a traditional storage system using RAID5 would result in a data loss with no chance of recovery. Isilon engineers estimate that the mean time between failures (MTBF) in n+2 RAID is over 100 times the MTBF in single-parity RAID. Isilon IQ is the only clustered storage solution to offer this level of data protection across a single file system in a clustered architecture.

In the event of a failure, OneFS automatically re-builds files across all of the existing distributed free space in the cluster, eliminating the need to have the dedicated “parity drives” typically required with most traditional storage architectures. OneFS takes advantage of the cluster by leveraging all available free space across all nodes in the cluster to rebuild data. By utilizing this free space while also drawing on the multiple processors and compute power of the cluster, data can be rebuilt much faster when compared to traditional architectures.

OneFS “n+1” and “n+2” Data Protection Schemes

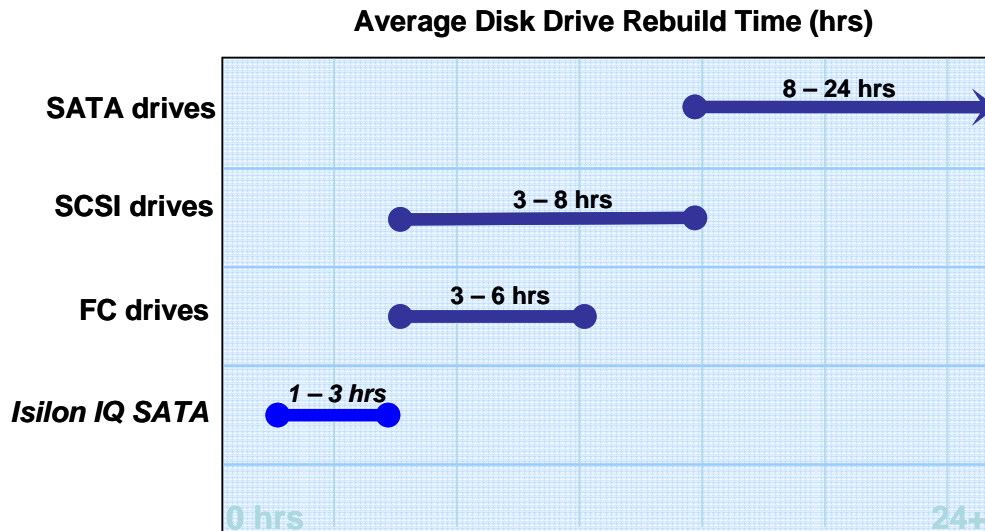


OneFS uniquely stripes data and ECC/parity across entire nodes and not just drives. This allows for advanced protection not achievable with traditional storage architectures. With n+2, Isilon can sustain up to two simultaneous hardware failure scenarios. By leveraging the collective power of the distributed cluster (e.g. CPU, memory, network) Isilon is able to reconstruct data across the available free space of the cluster much faster than traditional storage.

Industry Leading Drive Rebuild Times

The time that it takes a storage system to rebuild data from a failed disk drive is critical to the data reliability of that storage system. With traditional storage systems, the rebuilding process already takes many hours; a trend that is steadily worsening as drive capacities continue to increase. With the advent of ½ terabyte disks, expected within the next year, and the creation of larger and larger single volumes/file systems, traditional storage systems will require up to 24 hours or more to recover from a disk failure. During that time, such traditional storage systems are vulnerable to additional disk failures which will cause data loss and downtime.

Since Isilon IQ is built on a distributed architecture, it leverages all spindles and hardware within the cluster to their maximum capacity in order to reconstruct data from failed disks. Because Isilon IQ is not bound by the speed of any particular disk, Isilon systems are able to recover from disk failures extremely quickly. Disk failures within an Isilon IQ cluster will be rebuilt in one to two hours. When compared to fiber channel and SCSI disks, which can take upwards of 8 hours, or other ATA disk drives, which can take anywhere from 8 to 24 hours to rebuild a single failed disk drive, the advantage of Isilon's architecture is apparent. By delivering industry-leading drive rebuild times, Isilon IQ offers a more reliable storage system that is also more resilient and less susceptible to multi-failure scenarios.



Isilon IQ used Maxtor 250GB Serial ATA drives for the benchmark; disk was 87% full, with file system sizes from 1 – 10MB. Note: Industry average rebuild times are highly variable and depend on many factors. The averages shown above are ranges representing optimal to moderate use cases using drives between 160 – 250GB.

Active Monitoring and Pre-Emptive Data Rebuilds

OneFS constantly monitors the health of all files and disks and maintains records of the smart statistics (e.g. recoverable read errors) available on each drive to anticipate when that drive will fail. When OneFS identifies at risk components, it preemptively migrates the data off of the suspect disk to available free space on the cluster in a manner that is both automatic and transparent to the customer. Once the data is rebuilt, the user is notified to service the suspect drive in advance of actual failure. This feature provides customers with confidence that data written today will be stored 100% reliably, bit-for-bit correct, and available whenever it is needed. No other cluster solution today provides this level of data protection reliability.

Fully Journalled File System

OneFS is a fully-journalled file system with large amounts of battery-backed non-volatile random access memory (NVRAM) within each node, which ensures the integrity of the file system in the event of unexpected failures during any write operation. As each piece of content is written to an Isilon IQ cluster, the content is committed to the journal, protecting the system from a node (or cluster) failure without the need for offline file system checks. The node can then rejoin the cluster more quickly in the event of a node failure because it doesn't require a file system consistency check (FSCK). In addition, OneFS is transactionally safe even in the event of a NVRAM failure with no single point of failure.

Finally, some clustered storage architectures rely on a UPS to provide "non-volatile" memory to their storage nodes. This method is still susceptible to other hardware failure (e.g. node power supply failure) or software failure that creates a single point of failure scenario for NVRAM, putting the file system at risk. Isilon IQ overcomes this problem by locating battery-backed NVRAM in each node and fully distributing the file system across every node in an Isilon IQ cluster.

Transient Availability

A complete clustered storage solution is built on many different industry-standard hardware components. One of the biggest challenges for any vendor or customer looking to put together a clustered storage system is dealing with transient availability of components, or with temporary errors or conditions in one piece of hardware that can randomly arise and impact overall solution availability. In a clustered solution, the potential for this kind of error is exacerbated due to the large total number of industry standard hardware components being added together. For example, servers or nodes may temporarily disconnect from a cluster due to a faulty switch, a power cable may be defective and only work intermittently, power surges may occur, or some other faulty piece of hardware may cause a node to go offline for a period of time.

Therefore, a clustered architecture must be highly resilient and robust to counterbalance such hardware failures and ensure the integrity of the overall solution. In the worst case scenario for multiple failure risk, where the storage is in production with reads and writes occurring, coherency must be maintained within the file system in the event of a very wide range of potential transient conditions. Isilon has focused on this as a major cornerstone of OneFS. Developed with a number of advanced proprietary technologies, OneFS is highly resilient and is able to re-join nodes and maintain coherency even in the event of transient disruptions during production.

4 Conclusion

Today, companies of all sizes are dealing with an avalanche of digital content, unstructured data, and reference information that is driving massive increases in storage needs. Traditional storage architectures are being pushed to their limits and a new technical approach is needed. As a result, there is a broad-scale desire to move to clustered network storage built on standard hardware to gain significant advantages in scalability, performance and costs. However, reliability has remained a challenge for clustered architectures up until now.

Isilon has overcome that challenge and delivered a state-of-the-art reliable clustered storage solution today. Built on industry-standard hardware and powered by Isilon's OneFS distributed file system, Isilon IQ networked storage delivers:

- No single point of failure
- Tolerance for multi-failure scenarios
- Fastest disk rebuild times in the industry
- Pro-active failure detection and pre-emptive drive rebuilds
- Fully journalled file system
- High transient availability

For more information visit us at www.isilon.com.

About Isilon Systems:

Isilon is the premier provider of intelligent clustered storage systems. Isilon's award-winning family of Isilon IQ products are high-performance clustered storage systems that combine an intelligent distributed file system with modular industry standard hardware to deliver unmatched simplicity and scalability. Isilon IQ was designed for digital content and large data-intensive markets, such as media and entertainment, digital imaging, the Federal Government, life sciences, and oil and gas. The Seattle-based company is financed by Lehman Brothers Venture Partners, Sequoia Capital, Atlas Venture and Madrona Venture Group.

© 2001-2006 Isilon Systems, Inc. All rights reserved. Isilon, Isilon Systems and OneFS are registered trademarks, and TrueScale and SyncIQ are trademarks, of Isilon Systems, Inc.

For more information, contact Isilon Systems at:

Isilon Systems, Inc.
3101 Western Avenue
Seattle, WA 98121
Toll-Free: 877-2-ISILON
Phone: 206-315-7602
Fax: 206-315-7501
Email: sales@isilon.com