Quantum®

TECHNOLOGY BRIEF

StorNext for HPC Storage

ABSTRACT
Nearly all HPC environments share five important requirements for their data storage infrastructure. Quantum's StorNext® appliances provide the flexibility and performance needed to meet these requirements.

AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

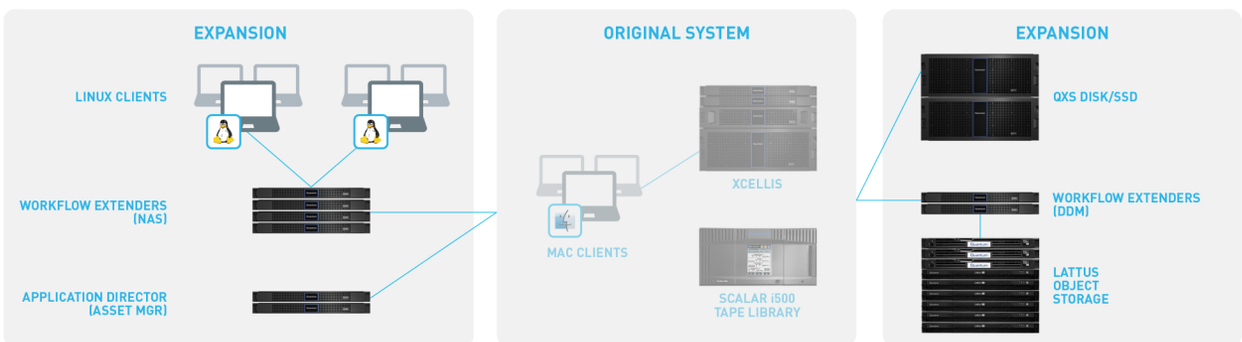
Capacity Expansion

StorNext's File System Expansion feature enables additional primary tier capacity to be added dynamically, without service interruption or downtime. StorNext clients are notified of the new capacity and it is immediately available for use. Similarly, tiers behind StorNext Storage Manager may also be expanded at any time, also without disruption. It is just as easy to add a whole new tier of storage, for example adding a cloud tier to an existing tape-centric system.

Functional Expansion

Other functions may need to scale besides storage. An influx of new users may require more NAS capacity on the front end. An increase in the amount of data moving to and from archive tiers may demand distributed data movers (DDM) on the back. These functions and more are scaled up simply by adding Xcellis Workflow Extender or Application Director nodes to share the load.

Figure 6 - Capacity and Functional Expansion



Storage Migration

Just as it is easy to add capacity and tiers, StorNext also makes it easy to decommission storage that is no longer needed. The Migrate Data tool enables data migration off of an existing stripe group onto other compatible stripe groups. Data migration is run in the background while the file system remains online. When all stripe groups on a storage array are migrated, that array may be decommissioned. SNSM also contains flexible migration tools for the tiers of storage it manages. Whether you wish to migrate older media to the latest generation of LTO tapes in a new library, move data off tape into the cloud, or even migrate from one cloud provider to another, StorNext's migration tools make it possible. Data has a lifespan independent of the life and limits of any individual storage device.

Open Architecture

The StorNext software was designed from the beginning to be hardware-independent. Xcellis systems allow customers to use a variety of disk and flash arrays from Quantum and a wide range of third-party vendors. In addition to supporting Quantum targets, SNSM in Xcellis supports tape libraries from other name-brand suppliers, as well as all of the most popular object storage systems and public cloud storage providers. This openness makes it possible to re-use existing assets within a StorNext architecture, and preserves choice over time.

The Center for Remote Sensing of Ice Sheets (CReSIS) tracks changes in the ice sheets in Greenland and Antarctica, and took advantage of StorNext's storage flexibility at the time of their deployment.

> We did not have to rip and replace anything – we could protect our existing investment. At the same time, we could easily add a tape archive tier into the storage solution. With the StorNext platform, we now have a single file system that manages our entire multi-tier storage solution, from tape to primary disk to the HPC cluster.

Riley Epperson
IT Engineer, CReSIS

AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

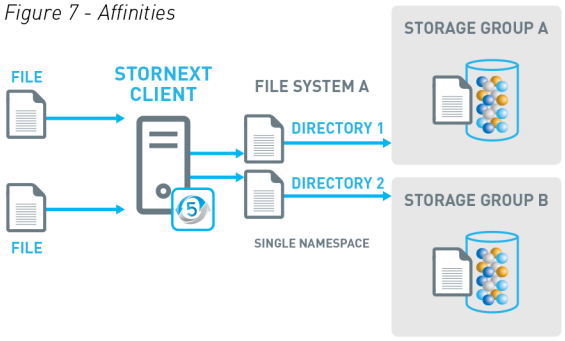
Affinities and Storage Policies

HPC and research environments frequently leverage shared computing and data storage infrastructure. The challenge is to ensure that the unique requirements of different groups of stakeholders or customers can be met, while minimizing the overhead and administration needed to manage the shared system.

An affinity in StorNext is a logical connection between a directory in the file system and specific stripe groups in the primary storage. This simple but powerful concept enables "steering" files to the storage that is most appropriate. If one part of a project or workflow generates lots of small files and read/write activity, that directory can be tied to storage capable of the high IOPS needed. If another researcher or project needs to manage large files that are read and written sequentially, they can use another directory in the same file system that is tied to an array tuned for high streaming performance.

Storage Policies within SNSM are also customizable with directory-level granularity, but this doesn't mean you have to manage every directory independently. Each policy may be associated with multiple directories, and typically most needs are met with only a few common policies. Where needed however, data management behavior may be highly customized to control which data lives where, when, how many copies and versions are maintained, in what formats, and for how long.

Figure 7 - Affinities



The Scripps Research Institute, one of the largest non-profit research institutes in the world, relies on StorNext's policy-driven tiering to define per-group data migration policies to meet the workflow needs of different groups.

> One group might want to archive data to tape immediately, while another might want to keep data on primary disk for a longer period. With StorNext, we can define policies to best suit the specific workflow needs of each group.
>
> **Brant Kelley**
> Director of IT Services, Scripps Research

CONCLUSION

High-performance computing is not limited to rarefied academic or government facilities with unlimited budgets. Research institutions and commercial enterprises alike use HPC techniques and tools to further their goals. As a result, HPC systems - including the data storage systems that support them - must provide more than raw speed. Key requirements in addition to high performance are large scale, shared access, storage tiering, and future flexibility. Customers confirm that Quantum's StorNext appliances meet all of these requirements and are perfectly suited to solve the data management challenges of the vast majority of HPC environments.

AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

TABLE OF CONTENTS

AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

INTRODUCTION

High-Performance Computing, or "HPC," is a term for which there is no universally accepted definition. What is "high performance" for one application might be considered impossibly slow for another. To complicate things, the concept of high performance changes over time as technology marches forward. Even Wikipedia punts, redirecting "High-Performance Computing" to the article on supercomputers without venturing a definition.

What is clear is that architectures and techniques pioneered in the supercomputing world have trickled down. Institutions and businesses in a variety of industries rely on these technologies to do real work, both academic and commercial. Use cases range from climate modeling to genome sequencing, mapping to design automation, and petroleum exploration to animation. You no longer need one of the top 100 supercomputer installations to get things done.

HPC requires high-performance storage. But as with the computational platform, how high is "high enough" depends on your goals. For the upper-end subset of HPC users, speed is the only thing that matters. It trumps cost, ease of use, everything. They will deploy file systems like Lustre or GPFS because they are the fastest, even though they are complex to deploy and difficult to manage.

For the rest of the HPC user population – the vast majority – priorities aren't so singular. Performance is still top on the list, but it's balanced with other things. Time is money in more ways than one, and time needed to integrate, maintain, and manage products from multiple vendors or a complex pile of open-source technology is extremely expensive.

DATA STORAGE REQUIREMENTS FOR HPC

Storage environments associated with nearly all HPC applications share five key characteristics and requirements, many of which are interrelated.

Requirement	Definition
Large Scale	"Lots" of data - so much that it cannot be managed and protected by traditional methods. Single file systems that must grow without restrictive spindle or LUN limits, for example. Enough data that batch backup becomes impractical or physically impossible as a protection method. And yet this data is important and must be protected somehow.
High Performance	Very fast access is needed to at least some of the data, some of the time. Fast enough that traditional NAS architectures are not sufficient. Often a large fraction of the data does not require fast access, or any access at all—but must be retained.
Shared Access	Some type of sharing and/or workflow requirement. This can be parallel—many nodes or users simultaneously accessing the same data set—serial, with multiple processes that transform or act on a data set in sequence, or a combination of the two. Multiplatform heterogeneity is another form of sharing, where machines with different operating systems must access the same data without getting in each other's way. Maintaining access controls in a cross-platform multi-user environment is a particular challenge.
Storage Tiering	The ability to place data on different classes of storage automatically, based on policies. A necessity when there is enough data that it can't economically spend its life on the most expensive storage. Very beneficial where a substantial amount of data requires long retention—years to decades or more.
Future Flexibility	HPC environments with petabyte or exabyte-scale data repositories must be flexible and expandable in situ. Disruptive forklift upgrades aren't an option. It must be possible to expand and upgrade capacity, performance, software, storage and interface technology as the state of the art advances – with minimal disruption. An open architecture allows mixing best-of-breed products from different vendors, and preserving existing investment.

Clearly, meeting all of these requirements simultaneously is a tall order, and solution designs can be complex. Ease of use is a critical factor, as it translates directly to ongoing operational expense. A system that is easy to use during initial configuration, through day-to-day administration, changes and upgrades will cost dramatically less to run than one that isn't as well constructed.

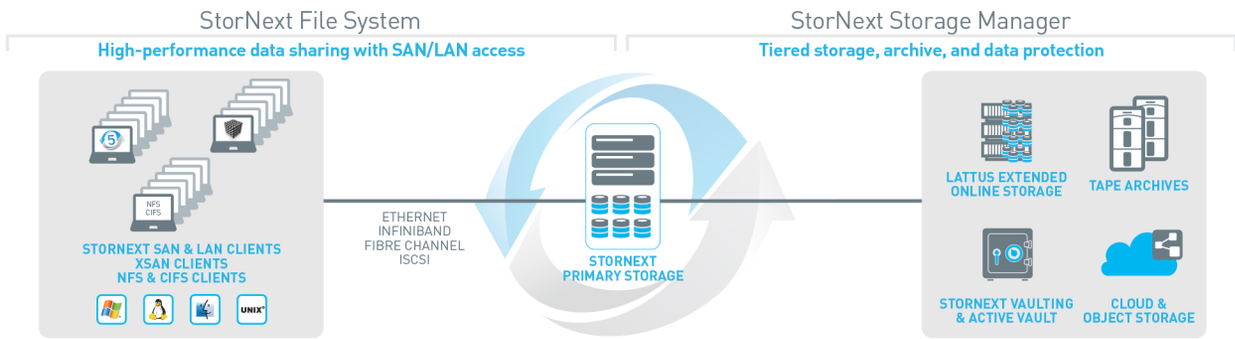
AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

QUANTUM STORNEXT

StorNext® is the name of the software engine within Quantum's Artico™, Xcellis™, and related family of appliances. Encapsulating the software into modular appliances makes the system easy to deploy and support, with well-defined performance characteristics. Time from design to deployment and use is much shorter with appliances vs. integrating a system from scratch with open-source software components. "Appliantized" does not equal "limited" or "locked in." The StorNext architecture is open and highly customizable.

StorNext has two main components, the StorNext File System (SNFS) and the StorNext Storage Manager (SNSM). SNFS is a high-performance, heterogeneous, shared file system. This file system may be accessed a number of ways, including via NAS protocols (SMB and NFS), an S3-style object interface, or for the highest performance, dedicated LAN and SAN clients. SNSM is a policy and tiering engine that transparently extends the file system to other tiers of storage such as tape, object storage or cloud, while managing capacity, data protection, archiving, versioning, migration, and related tasks. All of the SNSM functions are automatic and behind the scenes from the point of view of applications and users.

Figure 1 - StorNext Architecture



STORNEXT MEETS HPC REQUIREMENTS

StorNext offers a unique combination of features that make it ideal for HPC environments with the demanding requirements outlined above. Let's examine how StorNext is meeting HPC requirements in real-world applications today.

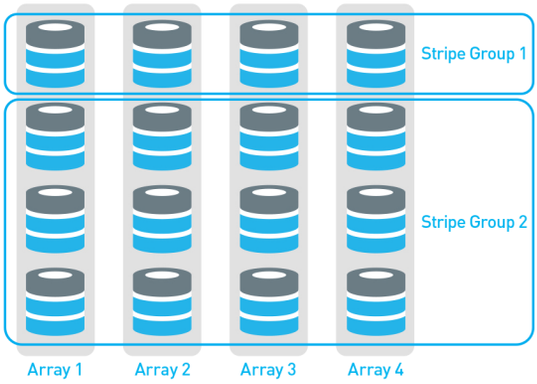
This HPC Requirement	Is Enabled by These StorNext Features
Large Scale	Stripe Groups, Storage Virtualization
High Performance	Direct Access, Metadata Separation, Parallelism, Tunability, Storage Agnosticism
Shared Access	Simultaneous Access, Multiple Access Methods, OS heterogeneity
Storage Tiering	Caching, Copies and Versions, Proactive Data Integrity Checking
Future Flexibility	Capacity Expansion, Functional Expansion, Storage Migration, Open Architecture, Affinities and Storage Policies

Large Scale

With deployed single file systems with over 550PB of capacity and 500 million files, and having a limit of 18EB and at least 1.4 billion files, it's clear there are few practical limits on the scale of a StorNext file system. To achieve this, SNFS uses a construct known as 'stripe groups', along with the capability to extend a file system across multiple storage tiers.

AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

Figure 2 - Stripe Groups



Stripe Groups

A stripe group is simply a collection of LUNs. A StorNext file system is composed of multiple stripe groups, which may be hosted on different storage arrays. This enables a single file system to scale beyond the bounds of any single storage array, enabling scalability in capacity as well as performance.

Storage Virtualization

In conjunction with SNFS, SNSM may be used to extend the bounds of a file system onto other tiers of storage, including cloud. No matter how many flash or disk arrays or other tiers of storage are included, clients always see a single homogeneous file system with a single namespace. If one file system isn't sufficient it's easy to deploy multiple file systems in a single, simple to manage environment.

GDWG, the computing center shared by Germany's Göttingen University and the Max Planck Society provides computing facilities to support scientific research. They have built a large-scale storage environment to store, protect, and provide access to millions of files and large data volumes for tens of thousands of users.

> "In our experience, StorNext is the only platform that can handle the number of files and the amount of data we manage, while also delivering high performance."
>
> **Stefan Teusch**
> Deputy Head of IT Infrastructure, GDWG

High Performance

The StorNext file system was designed to handle digital video nearly two decades ago, when that was a new and difficult computing task. SNFS was designed to "stay out of the way" and make it possible to wring every drop of speed out of the storage hardware. As a result, StorNext has been able to grow in capability and performance without alterations to its fundamental architecture, even as storage hardware has advanced in performance by orders of magnitude. Several aspects of the file system design contribute to this high performance capability.

Direct Access

SNFS is a shared file system, where multiple clients can see and access the same files at the same time. Shared file systems require an arbiter to coordinate access and avoid chaos. In StorNext this function is provided by a cluster of Metadata Controllers, or MDCs.

If all data access had to funnel through the MDCs they would quickly become a bottleneck. In StorNext this problem does not occur because the MDCs are not in the data path. Data access from SAN and LAN clients is direct from the client to the storage, traditionally over a fast network such as Fibre-Channel or Infiniband. When a StorNext client needs to perform a write, it makes a request to the MDCs for a block allocation on the storage. The client then writes directly to the storage with no intermediary in the way. For NAS access there is one additional hop through the NAS presentation, but if needed a cluster of NAS gateways may be deployed for maximum scalability, and each NAS gateway has direct access to storage without funneling through the MDCs.

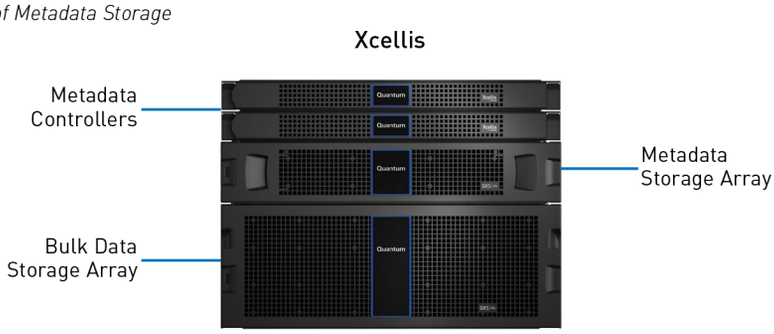
AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

Metadata Separation

In addition to metadata communication being out-of-band, metadata storage may also be separate from bulk data storage. This enables performance optimization in a number of ways. For environments with a high rate of file activity, metadata storage performance is critical to overall throughput. Because the storage used for metadata is relatively small and can be separate, it's easy and budget-friendly to amp-up metadata performance. Simply deploy some flash or a few SSDs for metadata instead of HDDs. You don't have to deploy a huge, unaffordable all-flash array for all of your data just because you need more metadata performance.

Metadata transactions are by their nature small relative to most reads and writes. In general, any block storage can be optimized for small or large I/Os, but not both simultaneously. StorNext's ability to separate metadata from data storage enables each to be tuned for their role, further maximizing overall system performance.

Figure 3 - Separation of Metadata Storage



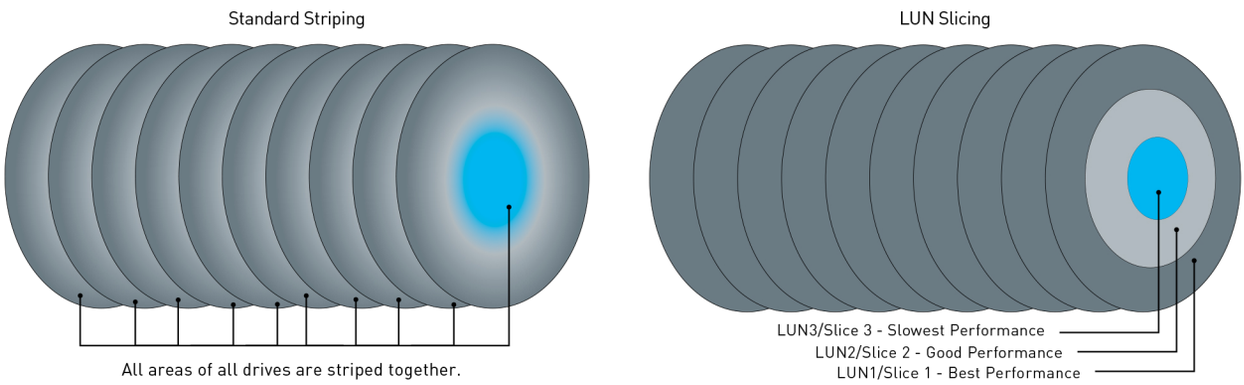
Parallelism

Because a StorNext file system may be built on an arbitrary set of LUNs, its performance is not limited by a single array controller. By spreading I/O across LUNs governed by multiple controllers, performance is aggregated.

Tunability

Achieving the highest performance for a particular workload requires understanding both the data characteristics and how the files are accessed. Is there a lot of random I/O with small files, or sequential streaming of large files? What about read vs. write? Given that information, the storage configuration may be optimized. With SNFS, the level of tuning that is possible is unmatched. From LUN sizes to LUN slicing, stripe groups to stripe breadth, affinities to allocation session reservation, client mount options and more. But just because there are a lot of knobs and dials doesn't mean they all have to be used. In most environments extreme customization is not required. A comprehensive tuning guide is provided with the product, and optional professional services are available.

Figure 4 - LUN Striping vs. LUN Slicing



AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

CERN—the European organization for nuclear research based in Geneva Switzerland—is the world's leading laboratory for particle physics. CERN uses StorNext for data collection for the ALICE experiment running on their Large Hadron Collider. Performance testing on the system prior to deployment in 2010 demonstrated sustained write performance of 4.5GB/s, with simultaneous sustained read performance of 2.5GB/s on a single file system. That's 7GB/s of aggregate I/O.

> Data is CERN's most precious commodity. Quantum StorNext is instrumental in collecting that data quickly and reliably.
>
> **Pierre Vade Vyvre**
> Project Leader, CERN

And that was in 2010. 7GB/s isn't bad even today, but it's far from the limit. Standard StorNext reference architectures for 4K video production today range up to 12GB/s of aggregate throughput.

Storage Agnostic

SNFS has always been storage agnostic. As long as the storage can be presented as LUNs, SNFS can use it. This has enabled StorNext performance to grow along with data storage technology, from HDD to SSD to whatever comes next.

Shared Access

Sharing of data takes many forms, from colleagues collaborating via NAS to Linux cluster nodes accessing data in parallel via Infiniband. StorNext enables many types of sharing through a number of architectural features.

Simultaneous Access

Unless more restrictive access rights are in place, multiple StorNext clients may open, read, and write to the same files at the same time. This is true whether access is via SAN client, LAN client, or NAS. This open access enables serial workflows to be parallelized. A raw data set may be processed by several applications at once, instead of a serial pipeline approach. Cluster computing methods may also be applied, where many machines process different segments of a large input file in parallel. Simply enabling multiple researchers to share a data set without having to make local copies can enhance collaboration, speed results and reduce costs.

Multiple Access Methods

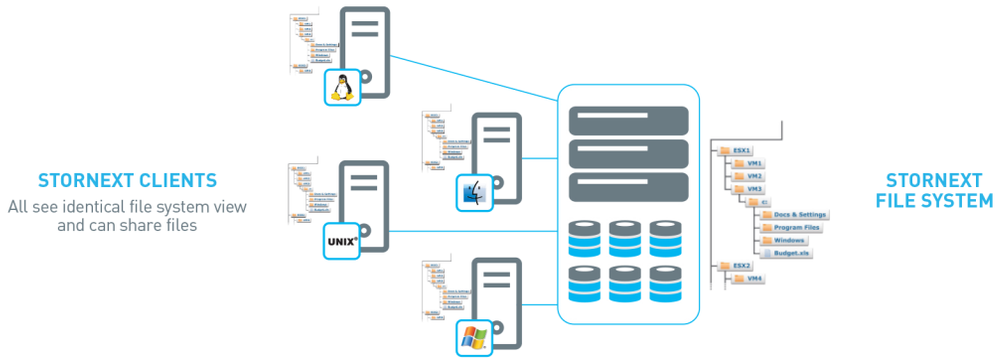
Users and systems may connect to a StorNext system via a mix of SAN, LAN, or NAS. The right choice depends on the number of clients, their location and connectivity, and individual performance requirements. Aside from performance, all clients are equal and may share access to all files.

OS Heterogeneity

The computing world has coalesced around three main operating system families; Linux, Mac OS, and Microsoft Windows. HPC environments commonly include all three. Data acquisition and cluster processing may happen on Linux, while researchers view results or perform further processing on Mac and Windows systems. All of these platforms have some ability to share resources with the others, but with compromises in performance and access control. By providing native support for all common OS families including cross-platform access control, StorNext simplifies heterogeneous sharing of data without restricting performance.

AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

Figure 5 - OS Heterogeneity and Simultaneous Access



The National Institutes of Health (NIH) department of Radiology and Imaging Sciences maintains one of the world's most important libraries of medical images on StorNext, managed by contractor MedData Research, Inc. The project heavily leverages StorNext's sharing capability.

> StorNext works seamlessly with all of the applications, even the roll-your-own applications that the research teams develop, as well as many different platforms like Windows, Mac OS, Linux, and UNIX. ... Scientists can analyze data in the lab over high-speed Fibre Channel, then go back to their offices and download the same files to their laptops to use in presentations.
>
> **Jeff Plum**
> Vice President, MedData Research, Inc.

Storage Tiering

The ability to incorporate different types of storage at a range of cost and performance points – transparently - is key to the value of StorNext in HPC environments. Due to the quantity of data these projects generate and use, it isn't simply a matter of saving a few dollars per year. Often it's the difference between being able to save data vs. discarding it to make room for another set. SNSM drives all policy actions, including managing the disk cache and tiering. Based on the policy configuration for the directory, a file that lands in the file system may have multiple copies created immediately to different storage tiers.

Caching

One of the functions of the primary storage tier in a StorNext environment is as a cache. Capacity on this tier is managed via a system of policies and watermarks. When the primary tier is near full, SNSM selectively removes files, leaving the copies on other tiers. If a file is accessed that is no longer on primary, SNSM automatically retrieves it back to the primary tier and makes it available to the requestor. This cache-and-tier scheme ensures that the most recently used files remain on primary storage for quick access. The details of this behavior are highly customizable – selected files may be pinned to the primary tier, others may pass quickly through the cache on the way to the archive, and other file groups may age off in a least-recently-used fashion. No matter how the system is configured, all files remain visible in the file system and accessible, no matter where they are located.

Copies and Versions

Unlike old-fashioned HSM, SNSM makes file copies essentially immediately – by default after five minutes of inactivity. The tiered copies serve as an important data protection mechanism. Because changed files are being continuously protected, it is unnecessary to "back up" a StorNext system in the traditional sense. This is ideal for data that is "too big to back up," but still has to be protected. Each file may have up to four copies, and each copy may reside on a different storage destination.

AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com

Copies are essential for disaster recovery and to protect against equipment failure, but they are less useful for recovering from everyday errors or mistakes. SNSM employs versioning for this purpose. Again based on policy configuration, up to 45 versions of each file may be maintained, with many options for how those versions are managed. When a researcher needs to roll back a file to the version from last Tuesday, it's easy to do exactly that.

At the forefront of genomics and proteomics research, the SIB Swiss Institute of Bioinformatics generates a lot of data. As 'omics and genomics in particular move toward the point of patient care, StorNext gives SIB a proactive strategy for protecting genomics data for decades.

> StorNext not only helps us make sure we capture data fast – it also makes archiving an automated, cost-effective process to help us fulfill our role as data steward. We always make two copies of the files on tape, keeping one available in the archive and the other vaulted to provide an additional layer of protection against any kind of hardware failure or damage to a site. We are dealing with some of the most valuable data sets on earth. StorNext gives us a multi-petabyte archive capability, long-term data protection, and the ability to easily roll back file versions.
>
> **Roberto Fabbretti**
> IT Manager, Vital-IT Group, SIB

Proactive Data Integrity Checking

StorNext helps ensure data integrity, even if files aren't accessed for long periods of time. To monitor for corruption, checksums may be calculated when files are stored and checked upon retrieval. If a corrupt file is detected, SNSM will fulfill the retrieve request from another copy, if available. Even better, if SNSM is paired with a Quantum Scalar i6 or i6000 tape library, it integrates with the Enterprise Data Lifecycle Management (EDLM) capability of the library. The library will scan all media on a user-defined schedule, and if degraded media is detected, StorNext will automatically copy the affected files to a new tape, before data is lost. This process is performed entirely in the background.

Airbus Defence and Space archives critical data streaming from Earth observation satellites managed by the European Space Agency. The Airbus team relies on StorNext, including integration with EDLM on a Scalar tape library to ensure data integrity is maintained over time.

> Archived data is not retrieved frequently, but when it is, we need to make sure that nothing has been lost. The policy-based EDLM features of the StorNext tape archive help proactively maintain the integrity of data on the tapes over the long term. If a problem develops with a tape, StorNext automatically migrates data to a better tape, without our administrators having to lift a finger. EDLM provides a huge advantage.
>
> **Mark Curtis**
> Technical Design Authority, Airbus Defence and Space

Future Flexibility

In traditional IT, it's common to replace data storage hardware every three years via the so-called "forklift upgrade", or even more frequently if existing storage runs out of capacity. For the large data sets common to HPC environments, this model is simply too disruptive. StorNext makes it possible to expand the capacity of any tier of storage on the system. Requirements always change over time, and new projects utilize storage in new ways. StorNext policies are granular and flexible, to adjust and evolve instead of becoming obsolete. Even the MDCs, the core of the system, may be replaced with zero downtime thanks to their clustered design.

AUTONOMOUSTUFF
Tel. 309.291.0966 | www.AutonomouStuff.com | info@AutonomouStuff.com