M U L T I - T I E R S T O R A G E Joe Little, Electrical Engineering, Stanford University (revision 1.0)

ABSTRACT

The Electrical Engineering department at Stanford University has often spearheaded the use of new technology via its collaboration with the industry or through its own research. Often, though, major changes in technology may drive an entire shift in the infrastructure required for the research and education activities of the department. The explosion of cheap, readily available storage has required us to adopt new strategies and technologies to manage data, whether its the primary storage or backup and recovery services. Foremost, as storage itself gets commoditized, we need to avoid expensive solutions and creatively adopt the next wave of solutions at minimal costs to keep abreast of what appears to be constant storage growth. Such standbys as tape-based backups may not be dead, but as will be shown, the realities of data growth require us to move beyond previous data management mechanisms. This white paper covers our use of new growth-oriented filesystems with snapshot capabilities, iSCSI for network layer independence, and a collection of other technologies provided by various vendors and open source projects to create a multi-tiered storage solution with self-service data restoration, long term growth, and disaster recovery.

In the not so distant past, research groups would tend to acquire their own infrastructure as funding sources, grants, and donations would warrant. This hardware generally included a localized file server for the group, or enough combined storage among the various hardwire networked workstations to provide for the group. Those drives were large for that time, upwards of 4.3 or 9GB at most, and most were directly used without any form of RAID or redundancy of storage. Backups generally could be run centrally, collecting from each file server the relatively trivial megabytes to potentially a gigabyte of delta change in data per day. Even with multiple research groups, incremental backups combined would average to less than 20-50GB per day at most, fitting easily in both the time window to perform the backup, as well as the average tape capacity of DLT or other media.

Within the past few years, storage use, driven by email, multi-media, and burgeoning application and file format sizes, has dramatically grown. Aiding this is the constant march to greater yet cheaper storage form factors. Where as 5 years ago, UNIX servers got by with 2-9GB, we now find workstations with a minimal of 160GB drives, and servers utilizing RAID arrays of multiple 73GB (SCSI/FC) or now 750GB (SATA) drives. Storage has also taken flight, with much data now on mobile computing platforms. Now, each group has local systems with over 200GB and servers holding terabytes of data. They still want those nightly backups, and even worse, they want further granularity. The task of backing up such data within the provided windows is arduous enough. The thought of actually recovering files, directories, or systems from multiple tapes within an acceptable time frame is almost impossible to reckon.

The solutions have already been presented by many vendors in vertical silos. We have centralized solutions of some flavor or another, including SANs, NAS, virtual tape libraries, NDMP, etc. We have even adopted some of the best in class of these, pooling money as best we can to adopt such things as Network Appliance's Filer. The advent of self-restore through regular file system snapshots has greatly aided in our perceived abilities to keep up. However, storage needs continue to grow and the cost of maintaining such products as the NetApp over many years gets to the point where we spend money feeding a low-terabyte count beast every few years for the cost of buying 10 times that storage outright in a new chassis. There is also the need to manage multiple silos of storage, and add or take away hardware without requiring clients to realign their mounts. A strategy to embrace the economics of storage pricing over time, shift storage to generic commodity units, provide advanced features such as snapshots and single logi-

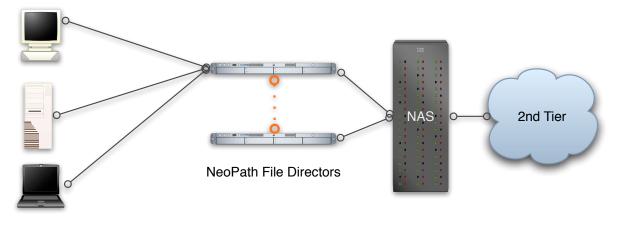
cal pools of storage, and encourage the use of open standards for longevity of data access and growth is required, and we feel we've figured out the pieces.

What has become evident is that a multi-tier approach to storage, utilizing arrays of disks for all backup requirements to match primary storage, and utilizing ultra efficient filesystems, allow us to both provide backups and restores to online storage in a timely manner, as well as scale our backup in line with our primary storage as it grows. Tapes have not been obsoleted, as they now fit nicely as a data archive mechanism, allowing easy migration to the safety of offsite locations for disaster recovery scenarios. If we can maintain stable, read-only online backups that persist over many days, we are now afforded the opportunity to write out to tape over multiple days a complete archive of our data. We now move from daily tape windows to monthly or as we'll see, quarterly windows.

Tier 1 Storage

We start by considering the first tier of storage. Today this is supplied by the likes of a Network Appliance FAS3050 and historically the 840 series. In the long run, we envision technologies that mirror the capabilities of Data OnTap, the NetApp operating system, to supplant the NetApps, but today, they do provide the best turn key file server solution, offering multiple point in time snapshots at the granularity of inodes for efficiency, NVRAM-backed fast NFS write capability, and large redundant arrays of disk that are easily serviceable. It also helps that their NFS implementation is considered one of the best in the market.

As previously mentioned, customers regularly need to expand to multiple NetApp filers or upgrade/ replace them. We have been using the NeoPath File Director to abstract the back end NAS file server from the data representation that clients see. The NeoPath product provides some special NFS features we make use of, such as virtual NFS servers that combine multiple backend exported file systems into a single tree. We use this to maintain a single logical tree for users even when the backend data comes from both old and new NAS systems, or different organizational volumes that need combined for ease of management. Even Linux and older Solaris file servers can be added into these synthetic trees. We also have access to live migration capabilities, allowing entire portions of the tree to move from one NAS to another while users actively access the data. An important property of the migration feature the File Director offers is its transactional nature - all file and directory updates are fully committed on both the old and new NAS systems, and the transition to the new NAS system is atomic. This product can also be used to best present both first tier and second tier snapshots to the end user in read-only backup virtual directories to aid in complete self-service restoration of files of all online backups. The 2nd tier solution will be covered at length in the following pages.



(Diagram showing Tier 1 Layout)

The File Director does have an ideal configuration that we recommend. For our purposes, the high availability setup provided by two clustered File Directors is a must. The backend NAS systems export their volumes only to the NeoPath, which in turn controls all client access through ACLs against its virtual servers and synthetic directories. Configuring the interfaces into 802.3ad aggregates (6 total) is also a must for both reliability and performance, and its expected that the backend NAS systems also use aggregation of multiple gigabit ethernet ports where possible.

Tier 2 Storage

What differentiates the next tier of storage is its fundamental purpose: to make readily available copies of the first tier covering multiple months of data. It is a complete copy of the first tier, also with snapshots of the filesystems it contains, usually daily and monthly in scope. The first tier, where snapshots exist, tend to be granular to hours and a few days. Ideally, this second tier should have daily snapshots that span multiple months covering between six months to a year.

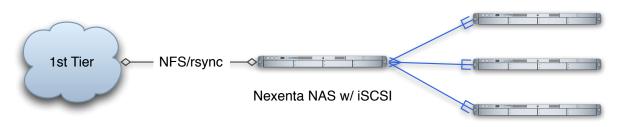
The foundations of this layer of storage need to be commodity in nature from the outset, as one needs to more than match the growth of the primary storage systems. So far, we've found that 1.15 to 1.5 times the storage of the first tier of storage is required given the efficiencies of inode-level snapshots. How does one possibly hope to scale as the first tier grows? There are a few assumptions that can be made. The make up of first tier technologies tend to be faster disks with their overall reliability tested and measured over time, with low latency connections and high aggregate bandwidth switches. You'll find that fibre channel is king here. That's a great part of what we buy in a Network Appliance or similar. The second tier can out scale the primary storage by adopting larger storage medium, which tends to come out initially in what has historically been slower interfaces using more commoditized interconnects like ethernet between large disk arrays. Thus, the ideal second tier should strive to embrace the most amount of storage with the least amount of per-chassis cost as its grows in size.

We have found two solutions that combined together meet this need. The ability to scale a filesystem over multiple commodity PC chassis is provided by iSCSI targets. The field of iSCSI, especially targets, takes some time to fully understand. Early on, it became apparent that software targets that handled direct I/O to large arrays of SATA storage would be the best fit. This led us to the SBEi iSCSI software target solution. Using commodity hardware with controllers from 3Ware or LSI, we've been able to acquire multiple units from such vendors as PogoLinux and Rack-Stanford University Electrical Engineering Multi-Tier Storage

able Systems that allow us to combine many disks of high quality (400GB to 750GB RAID edition) into virtualized iSCSI luns of around 2TB in size, usually of over 6TB per chassis. Each of these luns is backed by RAID5 redundancy to provide the best reliability and capacity mix. On the second tier, we tend to limit the number of hot spares used as the online requirements of such arrays is reduced compared to the requirements of first tier users.

The iSCSI target provided by SBEi is a best of breed, easily beating most software targets for performance, but more importantly, offering multiport I/O and reliability up to and including Error Recovery Level 2. ERL2 is not often fully implemented in the industry, and it shows the completeness of this software stack. Tying the iSCSI target to redundant initiators would allow us continuous growth of storage, bound only by the cost of disk, base chassis, and gigabit ethernet ports. All three of these tend to always drop in cost over time, allowing one to stay ahead of the curve with regards to cost of first tier storage. The only weakness of iSCSI over commodity components is the built-in latency that such a solution generally incurs. However, that latency is well within the requirements of a mostly read-only backup solution that the second tier represents.

All of these targets only provide half the solution. The other half of the solution requires an appliance iSCSI initiator that is itself predominantly a NAS. By this I mean to say it requires similar functionality we find in a Ne-tApp, but at a commodity cost. This past year has seen the release of what appears to be the ideal solution from Sun Microsystems in some of their new technologies included in OpenSolaris, namely the collection of their persistent iSCSI initiator, ZFS 128-bit filesystem with unlimited snapshots and growth, as well as their ownership of NFS that is included in the operating system. Put this together as an appliance on x86 hardware and you start to approach the level of a full fledged NetApp Filer.



SBEi iSCSI Targets

(Diagram of the Tier 2 Storage)

A variant of OpenSolaris, namely Nexenta, with the targeted functionality of a NAS could itself become a first tier candidate. We have actively been working with a Nexenta-based NAS solution. OpenSolaris itself is currently highly volatile, so using the base distribution itself is best geared for second tier adoption at this time, but I consider any well architected NAS solution, such as Nexenta's, built on that foundation with controlled product evolution to some day be sufficient for placement alongside NetApps in the first tier.

The general purpose of the ZFS-based second tier appliance is to pull via the rsync protocol a copy daily of all first tier data (specifically a given snapshot) onto the iSCSI backed volumes created on this system. That version of data persists itself over multiple daily snapshots here lasting for at least six months. This system exports via NFS these mirrored volumes and their snapshots back to the users, preferably via the same NeoPath File Directory within the virtual file server and directory structure it presents.

The use of rsync is essential since it compares both tiers of storage but only sends the differences over the wire. It is hoped that done in parallel, multiple rsync tasks can effectively tier very large data pools within a few

hours, returning us to reasonable backup windows. The combination of NFS/ZFS/iSCSI has one current drawback in that writes of many small files over NFS is not performant due to NFS syncs and high latency for iSCSI. However, by reading the source data via NFS or rsync and writing it to the local iSCSI-backed filesystem avoids this problem. There is no issue in the NFS read scenario from this filesystem, as NFS doesn't add the synchronization penalty for reads. For the given requirements, a ZFS-based NAS using iSCSI storage pools appears to be ideally suited to the second tier, providing both the mostly limitless growth and cost curve we desire. It has been found that the online recovery time provided by this second tier approach is very effective for random restoration requests of files, directories and such at the granularity of a given day, and that in the case of entire volume recovery completely wins out over older tape based recovery scenarios. The Nexenta-based NAS solution using direct-attached-storage is also being used in a first-tier capacity, and with support for tiering via rsync or native ZFS snapshot mirroring, we can have a commodity oriented solution at both tiers.

Tier 3 Storage

The final component in this storage solution returns us to those just mentioned tape technologies. Even with the relatively high performance recovery offered by disk to disk storage solutions, we still need to handle disaster recovery. First, the iSCSI-based solution also affords us some great benefits here at some cost. By tying ZFS together with mirrored iSCSI luns, we can conceivably place our targets at multiple locations around a campus as long as the interconnected bandwidth is reasonable (at least gigabit ethernet). Beyond this, one still needs to resort to tape based technologies.

Even here, iSCSI was adopted to allow for location independence. A 12TB LTO-2 tape library configured as an iSCSI target allows for multiple second tier head nodes, or replacement hardware, to access a multi-tape restores regardless of location. Even here, open source based technologies have arisen to the point of rivaling their proprietary counterparts, as both Bacula and Amanda (2.5) allow management of such tape arrays with tape spanning. In our case, we have adopted Amanda to effect the best transition from previous Amanda usage and move us away from our historical use of Legato for large volume backup at usually high costs. Adopting such a solution requires us only to backup the trailing edge of the snapshots of the second tier (within a week of expiring), giving us ample time to have automated backup of 12TB of archival data at a time. As data grows, these off site backup times can be staggered, still protecting administrators from most manual effort.

The restoration picture from such archival media is still not as ideal as one would prefer, but by limiting statistically most restorations to online disk within the first and second tier, the scenario of full tape recovery seems acceptably remote. It helps that with the solutions described here, most recovery efforts entail no involvement of the storage administrator at all.

Summary

This white paper detailed the general reasoning behind the mult-tier storage strategy of the Electrical Engineering department at Stanford University, with the goal of handling large storage growth and backup in a manageable and cost effective way for the long term. It is believed that utilizing our current first tier storage, as well as the evolving commodity NAS solutions of the second tier, we can eventually move to an end-to-end commodity solution with relatively limitless scalability in the design itself. Key to this are the NeoPath File Director, SBEi iSCSI targets, and OpenSolaris-based ZFS NAS system.

Stanford University Electrical Engineering