Commodity Storage Solutions at the Center for High Performance Computing

by Sam Liston (University of Utah)

The goal of this report is to describe in light detail the history of storage at the Center for High Performance Computing (CHPC), the current group storage solution provided by CHPC, and that solution's architecture. It is not intended to be an exhaustive description or a deeply technical dive or cookbook, but a high-level overview highlighting the positive characteristics as well as the limitations of this storage solution. Currently, group storage is standard NFS or SMB/CIFS mountable disk space, often with a RAID6 configuration, for contiguous storage in the 1-1920 TB size range that can be purchased at cost by researchers on campus.

CHPC has been providing storage for many years and continues to learn and gain knowledge, experience that started from very dedicated and specialized systems and advanced into expertise in multi-tenant, shared, and commodity solutions; they have spanned SAN (storage area network) and NAS (network attached storage) systems, to parallel file systems. Through these iterations the center has worked to find solutions that address the limitations of previous solutions. Improvements from faster hardware, better software, and new technologies shape those solutions. The goal is to provide storage which is as accessible and usable as possible, with a reasonable level of robustness, reliability, and manageability, and with good performance and scalability at a price point that is approachable. As new solutions are explored, they are thoroughly vetted for stability and resiliency. The positives and negatives are weighed. We have also learned the importance of facilitation, communication, and interaction with CHPCs diverse user community to help educate the users regarding the pros/cons and limitations of the solution (performance, handling of large numbers of small files, back-ups, downtimes if specific components fail, etc) so that they have clear service level agreements and understanding of our storage solutions.

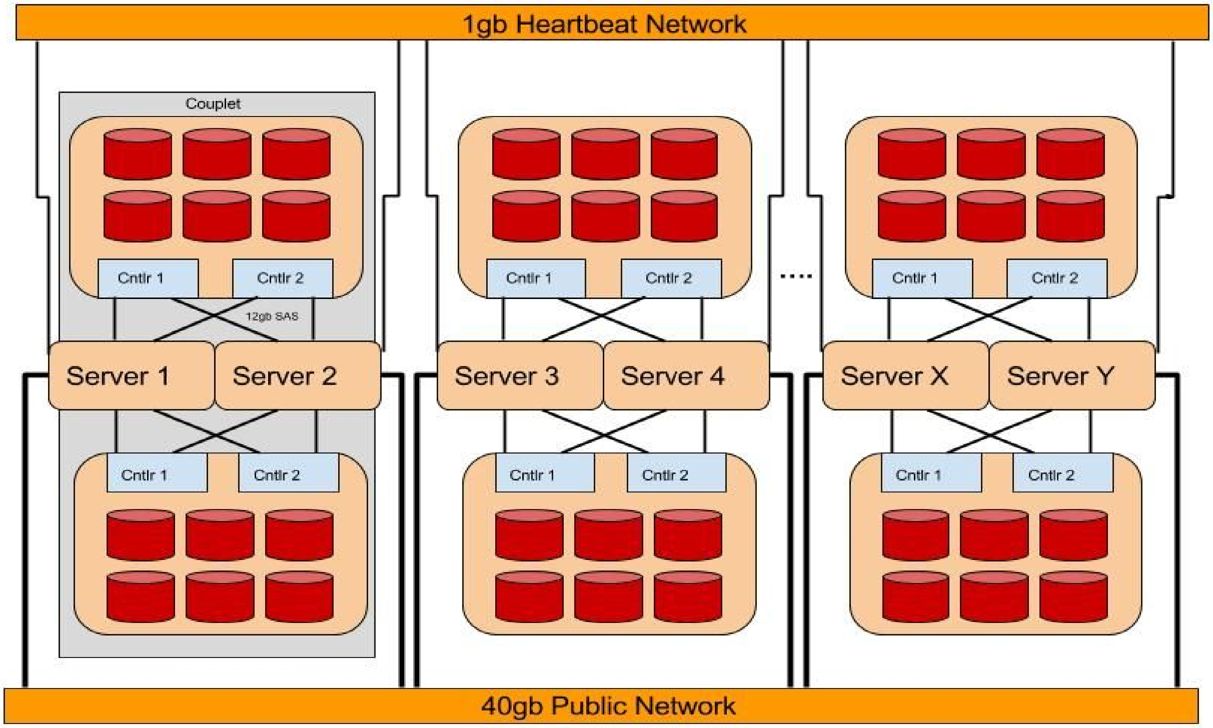
To help achieve the above stated goal of being conscious of the cost, CHPC tries to engineer and deploy solutions that can run on a variety of hardware. This minimizes vendor lock-in and allows CHPC to seek aggressive pricing. A solution that can be built and expanded as needed in a "condo" model is also a key feature, thereby allowing researchers to purchase space in large or small chunks as funds allow and their needs require. Many new researchers come to campus with the assumption that they will need to create their own storage solution for their new group, and they will do so if a compelling solution is not presented to sway them from that course of action. Thus CHPC strives to create solutions that are better than any reasonable solution a researcher and their group can set up on their own.

The solution described in this article is currently used on both the group and home directory offerings. This solution will continue to be used for the group space offering. However, as we look to refresh current home directory solution, we are moving away from this storage architecture. Details of this move are given at the end of this report.

CHPC Storage Architecture

In searching for a storage solution that was robust and minimized cost and complexity, CHPC iterated through several architectures. Both FC and iSCSI SANs were deployed in different generations of solutions. In 2011 CHPC purchased a turn-key distributed file system solution for “scratch” storage. Through this purchase CHPC was introduced to the “couplet” architecture. The “couplet” is the current architecture of choice.

Shown below is a schematic of the current couplet architecture used for CHPC’s storage offering.



A couplet consists of two servers each directly attached via 6 Gbps or 12 Gbps SAS (serial attached SCSI) to a controller tray. SAS architecture is such that each connection consists of four serial lanes at either 6 Gbps or 12 Gbps, so in reality a single 6 Gbps or 12 Gbps SAS connection can transfer data at 24 Gbps or 48 Gbps. Each tray contains redundant controller modules and the two servers each contain two dual port SAS cards with each controller module connected to one of the two SAS cards in each of the servers. This SAS topology allows for loss of a server, SAS card, cable, or controller module, without losing data access. By directly attaching servers to storage, the cost and complexity of the storage fabric is eliminated.

Initial implementations of the couplet model consisted of 12 drive trays filled with 2 TB drives. The current generation is built with 60 drive trays, with 8 TB drives. In selecting our current 60 drive unit, CHPC tested several trays from different vendors. The tray that we selected for both price and performance is the Dell MD3460. CHPC also tested, and could use if necessary, similar trays from HP, Nexsan, DotHill and Quantum.

Disk arrays within these 60 drive trays are defined in four 15 drive sets, and configured as either a 14 drive RAID6 plus one hot spare or a 15 drive Dynamic Disk Pool (DDP) with a single drive worth of reserve capacity. In this arrangement, each storage controller module has ownership of two arrays allowing for a more equally distributed workload on each controller module. At the OS level these arrays are stripped together in pairs using Logical Volume Management (LVM). Configuring the arrays in this fashion fully utilizes the horsepower of both controllers. Having dual controller storage trays enables continued functionality during a failure of one of the controllers.

The couplet model also allows for expansion without additional infrastructure through the use of JBOD (Just a Bunch Of Disks) trays. Each controller tray can have two additional JBOD trays connected behind it. At current drive capacities in our current RAID/DDP configuration, a fully expanded couplet can provide up to 1,920 TB of storage. In this configuration a single pair of controllers would collectively manage 360 disks. While this is very efficient for cost, expanding to this density could impact performance. It is possible in this configuration for the aggregate I/O from the disks to exceed the performance capabilities of the pair of controllers, causing them to be a bottleneck in the I/O path.

Each storage server is connected to the network via two 40 Gbps ethernet connections. These connections are setup in an active/passive bond, to increase resiliency in the case of a failure. The storage servers are also dual connected in an active/passive manner to a 1 Gbps network. This internal 1 Gbps network serves as a communication channel for heartbeating and quorum disk checks among the servers in the storage cluster. A final additional single 1 Gbps connection is dedicated to the baseboard management controller, which provides lights-out management and remote access to the server.

By default we purchase all storage with 5 years of warranty. This can usually be extended to seven years without issue, at modest cost. For the majority of the systems this covers support calls and next-business-day parts. As budgets allow and the criticality of the data warrants, higher levels of warranty can and are purchased, particularly as larger research groups purchase dedicated storage hardware.

In the current model RAID6 or DDP arrays are configured on storage trays filled with enterprise-grade 7200 rpm SAS drives. Other RAID configurations, (i.e. RAID5 or RAID10) and faster drive types (i.e. 10K, 15K, or SSD) could be easily used in this model, if different I/O or performance characteristics are needed. Directly connecting storage to servers using SAS was a choice to keep costs down and keep the configuration simple, but FC or iSCSI could be used as an alternative.

The software stack for the CHPC group storage solution starts with a standard CentOS operating system. Currently this is CentOS7. The high-availability (HA) suite, one of the optional sets of packages in CentOS, RHEL or other Redhat derivations, is used to form and manage the storage cluster. The HA suite orchestrates the location and motion of services among machines that are members of the cluster. These services can be configured to serve many purposes. Almost any application, or daemonized process can be started, stopped or migrated between cluster members manually, or according to prescribed failover domains. A service can also be configured to start multiple daemons in a particular order. For the purposes of bringing up a complete storage service CHPC configures file system mount, virtual IP address, Samba/CIFS export, and NFS export as a complete service.

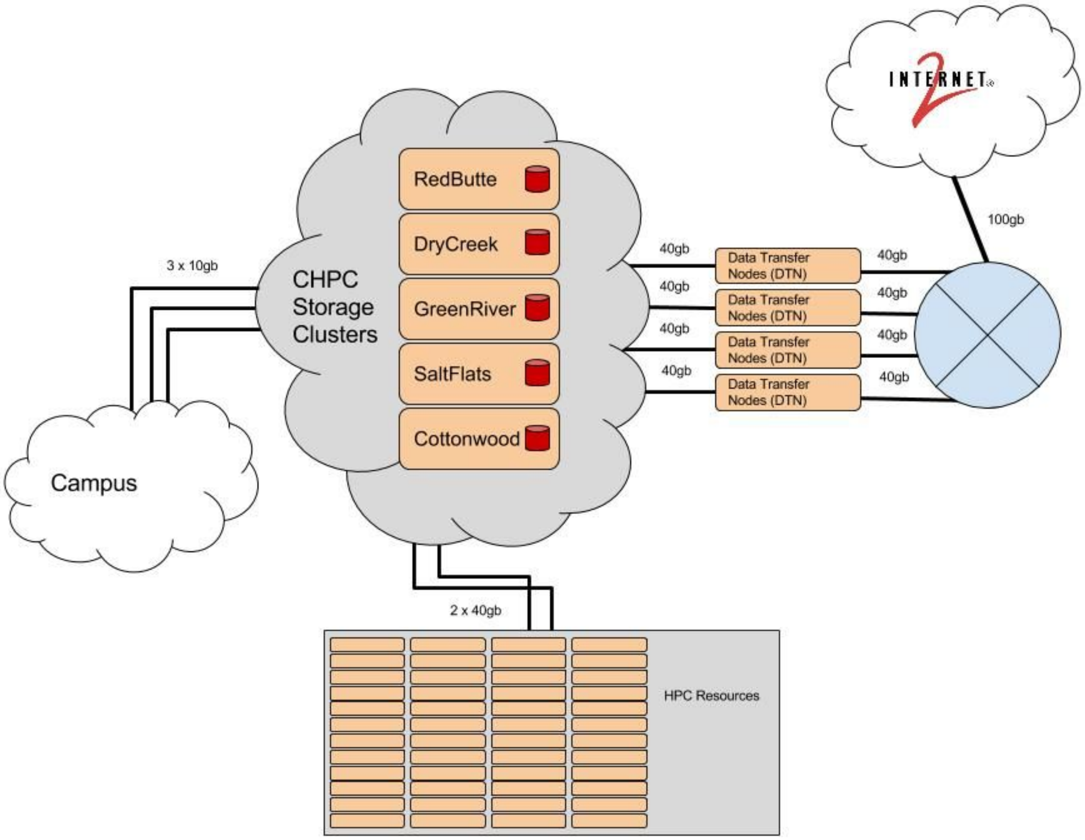
The two main components of the HA suite are corosync and pacemaker. Corosync acts as a messaging interface used between members of a cluster to keep track of how is in and who is out, what services are running, or not running. It handles the management of the cluster of servers themselves. This includes managing the heartbeating functions and fencing of members as needed. Pacemaker orchestrates the services themselves. The functions of starting, stopping, migrating, recovering and relocating of services is handled by pacemaker. It also understands failover domains and restricts the location of services in accordance with those domains. In the CHPC model a couplet is the failover domain. One server in the couplet is defined as the primary for half the file system services and the other the secondary for those services within the failover domain. This is limited to the couplet in the CHPC design due to the choice of directly attached storage trays. If iSCSI or fiber channel were used as the interconnect to the storage tray, a failover domain for a particular file system service could potentially contain all the servers within the cluster, enabling a service to be freely moved between all members of the cluster. Management software for the HA suite is available in both a WebUI and a full CLI.

The file system of choice for the current storage cluster generation is XFS. This is another flexible aspect of the CHPC storage solution. There are many file systems with a variety of features. XFS offered a good array of features along with a stable and reliable history. ZFS and Btrfs are being explored for future use. Both offer many new features and capabilities, but are still a bit experimental. XFS allows for creation of very large file systems. These XFS file systems are configured with 64-bit inodes to facilitate file counts into the billions. Quotas are configured on each file system. Three type of quotas are allowed: user, group and project quotas. Project quotas are used on the CHPC storage clusters to subdivide a file system into smaller pieces which are sized based upon the amount of storage purchased by a researcher. By provisioning storage space in this way researchers can purchase as little a 1 TB. Whole couplets can also be purchased by a particular group and added to the storage cluster, if their needs require.

All storage servers, their associated disks, and their networks are monitored for health and tracked for performance using Nagios and Cacti. Nagios uses defined heuristics to ensure hardware and software availability and functionality. Cacti is used to track individual system metric, such as CPU, memory or network utilization. However, as both of these applications are antiquated, CHPC is beginning to explore alternatives. In the event of an issue, or hardware failure alert emails, from Nagios, the servers, and/or the disks trays, are sent to the appropriate admins, to ensure necessary actions are taken as soon as possible. Additional linux tools created for diagnostic, trending and troubleshooting purposes are in place and allow admins to determine and diagnose issues, bottlenecks and user behaviors. As new and better tools are discovered that enable admins to better understand the functions of the system, they are deployed into the CHPC suite of tools.

Provisioning and management of systems is handled through a Spacewalk server. Spacewalk provides a single pane of management for linux systems, handling OS installation, configuration and customization for a single machine or across an entire cluster. It manages software updates and pushes out security patches and bug fixes. As an open-source tool, Spacewalk, enables CHPC to efficiently deploy and manage large sets of systems and maintain OS homogeneity and consistent security without software costs.

Network Topology of CHPC storage



Shown above is a high-level diagram of the bandwidth characteristics and network topology and associated with CHPCs storage clusters. Accessibility is a crucial component in CHPCs storage deployment. All storage: home, group and scratch are accessible on the HPC resources. They are also accessible on the Data transfer Nodes (DTN) which are connected to the science DMZ architecture that provides a fast path, bypassing the campus firewall, for file transfer to remote locations such as peer institutions and national resources. Shares are also available on campus or through the campus VPN via SAMBA/CIFS.

Limitations of the current solution

At present, backup resources at CHPC are limited and not scalable with data growth. Group space which comprises the majority of the storage managed by CHPC is backed up to tape as requested by the owner. That researcher, through CHPC, purchases tapes and CHPC archives the requested data to those tapes. Due to the sheer scale of the group space storage and the limited nature of the backup systems at CHPC an archive run can be performed, at most, once a quarter, and even this is not sustainable due to the growth in data. The fact that a good portion of the total data housed at CHPC--while being resilient against failure--is not backed up is a known deficiency in this storage solution. However, CHPC has neither the budget or man-power required to build and maintain a traditional backup solution equal in scale to primary storage.

An additional limitation is one that is intrinsic of all RAID based solutions, specifically the difficulties surrounding recovery from failure. Catastrophic failure of a RAID system is always a possibility. Measures are in place to defend against small failures, such as disks, power supplies or even controllers, but the possibility of large-scale failure, such as errors in the logic of the RAID structure itself, is real -- thus the saying: “RAID is not a backup.” Recovery from a large-scale failure in the CHPC RAID-based storage solution can be complex. If the failure is in the hardware or firmware, the problem must be diagnosed and either deemed corrected or new hardware to correct the problem must be shipped before data can be restored. In addition, as most backup solutions are tuned to efficiently backup data, as opposed to restoring it, if a backup or archive of the data is available on tape, the restoration process takes significantly longer than the original backup process.

This deficiency can be most impactful to home directory space -- space that users access continually. To address this CHPC is in the process of deploying a new home directory system. This Dell Compellent system will be configured in two hemisphere. One hemisphere will be synchronous mirror of the other. In the case of a catastrophic failure of one of the hemispheres the other hemisphere will take over serving the data to users without interruption, allowing us time to triage and repair the failed hemisphere while the other side is live, thus greatly decreasing the chances of an outage of data access.

In an additional effort to mitigate another of these deficiencies CHPC has implementing an object-storage based archive solution. For a detailed description of this solution see (https://www.chpc.utah.edu/documentation/white_papers/ArchiveSolutionattheCenterforHighPer formanceComputing.pdf). This solution, called Pando, is based on Ceph and is configured as a self-service private cloud that researchers can purchase portions of to serve as a near-line backup copy. This archive solution removes much of the management overhead related to our traditional tape-based backup processes. It can be used by researchers as a place to store a secondary, or tertiary copy of important data, or a place to stash data that is at rest to free up space on particular group shares for data is being more actively being used.