SCS Distributed File System Service Proposal

Project Charter:

To cost effectively build a Distributed networked File Service (DFS) that can grow to Petabyte scale, customized to the size and performance that the School of Computer Science desires.

The deliverable system motivated and outlined in this proposal will provide the School of Computer Science starting with 400TB of usable, secure and highly available file storage to support the current and future services of the CSCF at acceptable performance levels. This system will be designed to be incrementally expandable without service interruption and replace the NetApp hardware within a evergreen generation.

Introduction:

Open-source software-defined storage systems are quickly evolving to match the feature set, performance, availability and capacity heretofore only available with proprietary storage solutions. This revolution encompasses the length and breadth of storage ecosystems, from performance (eg. Lustre) to cloud (eg. Ceph) to scale-out distributed storage (eg. Gluster) and in object, block and file storage. With these open-source tools, a capable IT department can reliably field, manage and support a highly-available and redundant storage system on commodity hardware with features and performance available only on specialized, proprietary hardware only a few years ago.

Many of these open-source software-defined storage systems leverage the modern high-performance networking options that are available to federate cost-effective single storage servers into a single distributed service. The bandwidth available in a modern Ethernet network solution (>40Gbps) in conjunction with load-reducing features such as Remote Direct Memory Access (RDMA), allows for the efficient integration of single server performance. Such Distributed File Systems (DFS) offer many attractive features:

1. Distribution of storage workload across several servers to increase performance.
2. Replication of stored data across several servers to provide data security.
3. Ability to provide large capacity single-namespace storage volumes in excess of tens of petabytes through the aggregation of individual server storage resources.
4. Physical separation of system component servers to provide high-availability against power outages, floods and other disruptive events.
5. A natural, non-disruptive physical mechanism for system expansion: add a server to the network.

This document explores the possibilities and benefits of providing an order-of-magnitude replacement of our existing NAS service that can be scaled easily by simply adding more of the building blocks used to create it. It also provides a 4-year transition plan for implementing the new service for all School of Computer Science DFS needs.

Requirements:

Any viable DFS storage solution must meet the following availability, performance, service and feature requirements:

- The solution must be available (functional) and redundant (with no data loss) against the loss of a single server room due to any disruption.
- The solution must provide data security against on-disk degradation (bit-rot) and failure.
- The solution must provide performance levels adequate to at least meet those currently provided by various CSCF network storage services.
- In the event of the loss of a server room, the service must provide usable performance for dependent CSCF services throughout the duration of the disruption.
- The DFS will natively provide block storage for CSCF virtual and bare-metal infrastructure.
- The DFS will natively provide several filesystem options to dependent CSCF services, including:
    - NFSv4 including Kerberized NFS
    - pNFS
    - POSIX-compliant distributed filesystem (one client to many servers) eg. glusterfs, CephFS
    - SMB/CIFS (possibly distributed using CTDB).
    - encrypted file volumes (desired but not essential feature)
- The solution will be sufficiently flexible to meet new requirements as they become apparent, with minimal effort.
- DFS storage nodes would be distributed evenly between the MC 3015, M3 3101 and DC 3558 server rooms to provide a redundantly dispersed data storage service that will function even if an entire server room (building) goes off-line.

Services we'd like included natively are:

- dedicated block storage (iSCSI)
- nfsv4 (krb/nfs and/or pnfs)
- gfs2
- cifs
- encrypted file volumes

Deployment Requirements:

- Nodes to have full remote lights-out-management capability
- Base OS distribution auto-deployed via CSCF PXE-boot service
- OS and Application software and configuration to be fully managed via CSCF SaltSTack using information contained in SCS Inventory database
- Full health, resource and response monitoring via current established CSCF monitoring services.(see http://watcher.cscf.uwaterloo.ca/)

Services This Proposal Would Enhance/Enable:

1. OwnCloud:
   A self-hosted file sync and share server. It provides access to personal data through a web interface, sync clients or WebDAV while providing a platform to view, sync and share across devices easily—all under the clients' control. The ownCloud open architecture is extensible via a simple but powerful API for applications and plugins and it works with any file storage system. The anticipated storage requirement for this service is 1TB per School academic faculty and staff member and 500GB per School graduate student for a total of 280TB.
   - File synchronization would help with backup and evergreening of School workstations as evergreened hardware simply sync with our cloud and then be ready for deployment. The staff hands-on interaction time for the new hardware would be greatly reduced if not eliminated.The ability to share files with colleagues without being concerned with "FIPPA" nor "Patriot Act".
2. Depot.cs.uwaterloo.ca
   - Distribution of our Debian meta-packages that enables others to easily duplicate school provided services on their own hardware. Total storage requirement: 10TB
3. Mirror.cs.uwaterloo.ca
   - Redundancy for "mirror.csclub.uwaterloo.ca" which is an on site mirror of the open source software we use to provide the School's IT services. Total storage requirement: 40TB
4. Central location for system logs (rsyslog). Total storage requirement: 10TB
5. Time Machine file space for SCS Staff, Faculty and Grad Mac systems.
6. Backup (encrypted) Storage for DataBase services with 3-year history. Total storage requirement: 10TB
7. Replace and augment the Direct Attached Storage on the backup system so the tape juke box is only used for "Archive" storage.
8. SCS private cloud to support more automated and agile deployment of IAAS/PAAS virtual hosts. This also includes distributed management of our virtual host environment. Total storage requirement: 50TB
9. Home Directory Service:
   - Preliminary experimentation with existing storage hardware indicates that this service could provide a more cost effective home directory service than our current 40TB NetApp service. Storage requirement: 50TB
10. Swing space for research data:
    - Research computer installations in the School are now commonly host in excess of 100TB of storage. There is a demonstrated need for CSCF to be able to house data on this scale, at least temporarily, in the event of disruption or installation reconfiguration. Being able to provide large blocks of data storage so a research group can rebuild/expand their raid setups. CSCF has had requests for 20TB to 100TB of temp disk space from researchers.
11. Development and deployment of such a DFS on the scale discussed here (grand total usable storage for an initial installation ~400TB) would greatly increase the CSCF institutional competence for dealing with systems of this type, on this scale. This knowledge will be useful in the (near) future for supporting research system requests from users in "big-data" and other

fields with large storage capacity needs. Ultimately, this class of support capability increases the attractiveness of the School to prospective Faculty.

Existing CSCF Deployed Network File Services:

1. NetApp service
   size:
   cost:
   features:
2. Six 12x3.5" disk-bay Data Servers - Dell 515s
   Purpose: setup for backing up grad PCs, but now just Tier 2 storage
- What's on them now, mirror.csclub (20TB), depot.cs, xhier data tree, syslogs
- network-mount storage not in home directories
  size:
  cost:
  features:
3. Three Direct Attached Storage Jetstore Shelves -
   size:
   cost:
   features:

Transition Plans

Build a (Number of Server Rooms) * N data storage server nodes where

N = ceil(
(Minimum Size of desire DFS system)
/ ((Number of Server Rooms-1)*Size of Storage Node)
)

Purchase Storage nodes in groups of (Number of Server Rooms) every other year such that the oldest nodes get replaced in their sixth year of service.

Hardware Configuration

There are a number of possible hardware configurations. We estimate that approximately $120,000 would be required to setup an initial system to satisfy the above requirements.

Dependencies

The 40 Gb network infrastructure currently being put into place for the MySQL cluster is a requirement for this service to operate in a production environment.

Costs: Staff Time

Staff time to investigate DFS distribution we choose.

- 30 hours (~20 so far) - Lori / Dave

Staff time to build the initial system.

- 80 hours - Lori / Guoxiang
- 10 hours - create documentation for adding file services / maintenance

Staff time to configure file services

- 10 hours - Guoxiang / Lori (based on documentation)

Ongoing staff time to monitor and replace failed drives and other hardware.

- estimate : 1 hour per week (1 unit) over the lifetime of the system

Evergreen - and next set of evergreen nodes (expected to be multiples of three)

-

Costs: How the Yearly Fiscal Budget Would be Affected

It is expected that this service will replace the NetApp and Backup.cs JBOD shelves in regards to the budget line item, so new net annual request is expected.

External References

Linux 4.0 Hard Drive Comparison With EXT4 / Btrfs / XFS / NTFS / NILFS2 / ReiserFS

A PERFORMANCE COMPARISON OF ZFS AND BTRFS ON LINUX

Comparison of ZFS and BTRFS on LINUX