MSU Research Storage Technical Strategy

Goal

Provide research storage for MSU researchers with the following functionality:

- High performance computational storage (HPCS)
- Dataset Publication Archive (DPA)
  Features:
  - Dataset storage for the Institutional IR
  - "Write Once, Read Rarely"[1] durable storage.
- Research project storage (RPS)
  Features:
  - Network adjacent to HPC (available as HOME)
  - Accessible via SMB, CIFS, NFS, SSH, RSYNC
  - Local granular snapshots
  - Appliance-style management
- Backup Storage (BKS-[0,1,2])
  - Level 0: Human Error Resiliency (local snapshots, ~daily)
  - Level 1: Datacenter Resiliency (nearby snapshots, ~weekly)
  - Level 2: Regional Resiliency (non-local snapshots, ~monthly)

All storage will be accessible via the Science DMZ with the Data Transfer Nodes (DTNs) mediating data ingress and egress. Globus would be the primary method of data transfer, with others available as necessary. The exception would be BKS-3 non-local backups, they would not be available to users, but would be retrievable via administrative action.

Until the specific technological solutions are known for each of the storage functions (DPA, RPS, BKS-[0,1,2]), it is difficult to estimate snapshot granularity, backup granularity, and total storage available to each of the functions. These details will become clear as we implement the solutions. In each of the BKS cases, we would keep as many snapshots as the storage will allow, and back up as often as bandwidth will allow. The periods listed are approximate targets.

Current Setup

We currently have HPCS in the form of the Hyalite Lustre storage. It is available via Globus, scp, sftp, and rsync though only Globus transfer is recommended (the other methods are slow). We are using a variety of methods to simulate the other storage functionality using only HPCS, with limited success. Instead of having custom infrastructure providing the specific storage types,

[1] "Rarely" in this case means read tens or hundreds of times per day, but not interactively edited. It is ok if the file takes several seconds to become available.

they are being emulated using software. This negatively impacts the performance of the storage and is not well suited to the core competency of the infrastructure. Also, it has high admin overhead and is not sustainable as we scale up our research storage services.

HPCS storage is available in a single file-system with four subdirectories:

- /scratch/: no quotas, no backup, files older than 90 days are purged
- /work/: high quotas, no backup
- /store/: low quotas, weekly backups
- /backup/: local backup target for BKS-0

Additionally: HOME is an NFS-mount of the head-node local disk. Each compute node has a LOCAL HDD for non-networked storage (wiped after each job).

Backups

There is no MSU-local storage location large enough to support backups of the entire HPCS storage. HPCS storage is hardware-failure resilient via RAID. BKS-0 is provided via the following strategy:

- HOME is backed up daily to the /backup/ directory of the HPCS.
    - A single tar file per home directory plus a block-level de-duplicated archive of dailies.
- In active development:
    - Weekly tar.gz (max compression) of the daily archives transferred to the Indiana University Scholarly Data Archive (IU-SDA).
    - Weekly snapshot of /store/ into /backup/ de-duplicated archive. Keep N.
    - Monthly tar.gz (max compression) of the STORE archive to the IU-SDA. Keep N.
- BKS-1 is not available.
- BKS-2 is available via the IU Scholarly Data Archive.

Near-term Implementation (2 to 3 months)

The near-term implementation activities will establish BKS-1 and DPA storage. BKS-1 storage will be provided by low-cost storage appliance hardware that will be procured in January 2016. It will be hosted in Renne and isolated to the job of backups. The initial installation will be ~50TB raw, which with compression should be able to meet our current needs. It can be expanded over the course of the next few years up to a maximum size of ~800TB.

The former Research Computing Group storage cluster will be upgraded and reconfigured with the help of BiosIT to provide the DPA functionality. This hardware includes the storage hardware contributed by the library in 2014. The hardware has been upgraded so that each machine is as equivalent as possible, giving us a total of 216TB of raw disk space. This storage will be configured into a Ceph cluster managed by the HPC Bright Cluster Manager. It will be configured with high durability and will have a useable storage space of ~70TB. The datasets stored on the DPA will be backed up to the BKS-1 in Renne.

Next Steps

- Move existing data and services off of the current Gluster servers onto Lustre, wipe and re-provision.
- Purchase and install the BKS-1 backup storage appliance.
- Configure a backup spooler VM to orchestrate backups.
- ETA: February 2016

The cost of this near-term implementation is ~$15,000 for the upgrade of the DPA hardware (already done) and ~$23,000 for the initial BKS-1 hardware. This will come out of cluster upgrade funds.

Long-term Implementation (6 to 18 months)

The Long-term goal is to find and implement a solution for the RPS infrastructure and find sustainable BKS-2 (offsite backups).

Up to this point, the HPCS will have been doing double-duty as HPCS and RPS, a duty that it is not particularly suited to. HOME will still be hosted off of the NFS-mounted head node disk, also non-ideal.

Over the course of Q3 and Q4 2016, we will explore different possible solutions for RPS. The ultimate solution will be available for use as HOME for Hyalite HPC as well as for shared collaborative project storage (SMB/CIFS). The first steps are to define the requirements for the RPS solution, determine a budget, then to pass that information off to an RFP committee. Bids would be reviewed and a solution would be procured and implemented in Fall of 2016.

Ideally the RPS would support local lightweight snapshots, and subsets of the data will enter the existing backup pipeline to BKS-2 and BKS-3. It would take over the "STORE" functionality of the HPCS, so that it would no longer be expected to do any local backing up at all (leaving all of the performance and bandwidth for HPC activity). The target for this would be somewhere around a Petabyte, with each researcher getting 1TB of research storage (with additional available based on need).

We will also be looking into different solutions for scalable off-site backup (BKS-2). The IU SDA will only support up to 50TB of data storage, which we will use up pretty quickly. Past that we will have to pay for a larger allocation, and at that point other options may make more sense. The question of encryption is also important to sort out. When do we encrypt, where do we keep the encryption keys, and what are our policies for encrypted data. The details of backups and will be codified in a technical document which will be reviewed by DISC, ITC Security, and the HPC Advisory Group.