Storage options redux

Building on Gluster Alternatives and Cloud Provider Alternatives but with the Whole Tale requirements.

Use cases

Labs Workbench

User has access to a home directory that is mounted in every running container across cluster nodes. The user may have access to other shared volumes with read/write permissions. The user will be running a variety of applications, such as Postgres and MongoDBs, or compiling Cactus for the Einstein Toolkit School. If the Workbench is deployed near an existing cluster (e.g., ROGER), then the user and permissions should be shared.

Whole Tale

Use case: A user creates a new tale based on an existing dataset (readonly). The notebook uses the data to produce new outputs. This is published as a new tale (ideally the notebook is part of the permanent data for the tale, captured in the workspace).

Other requirements

- Shared home directory
  - When capturing a tale, want an exact copy of the home directory at instance in time
  - Relates to provenance, capturing current state, can be published
- Fast
- POSIX?
- Versioning:
  - Conceptually similar to object stores – when you modify a file, you create a new version while potentially maintaining the old one
  - Relates to reproducibility, allowing pointers to immutable versions of data
- Mountable anywhere
  - sshfs
- Notifications: May be implemented in Fuse

Storage options

NCSA

- GPFS shared over NFS
  - Storage condo: Managed by storage team
    - https://wiki.ncsa.illinois.edu/display/NRE/About+NCSA#AboutNCSA-StorageCondo
  - ROGER
  - ADS
    - NCSA/Tech Services/Library
    - $96/TB/year
    - https://www.library.illinois.edu/rds/active-data-storage-overview/
- ZFS shared over NFS
  - Santiago: Managed by ISDA
  - ITS backup: Managed by ITS
  - DXL: Managed by DXL
- OpenStack/Cinder (XFS exported via GlusterFS)
  - Managed by Nebula team

For container-based storage, a common model is to create NFS or GlusterFS file server VMs in Nebula exporting Cinder volumes. GlusterFS has proven non-performant particularly for Docker images (slow untar).

SDSC

- SDSC Cloud/OpenStack
  - Ceph via Cinder
  - Swift
- Comet: Lustre shared over NFS
- Data Oasis?
- Project storage
  - NFS mounted storage
  - Hotel v Condo
  - http://research-it.ucsd.edu/_files/idi_vmware-cloudcompute-storage_posters_part2.pdf

TACC

- Rodeo/OpenStack?
- Wrangler: Lustre shared over NFS

Notes

Ceph/CephFS

- Used by SDSC OpenStack (and 50+% of OpenStack survey respondents)
- CephFS? offers POSIX semantics
- Gluster v Ceph
    - Ceph= object store
    - Gluster = scale-out NAS and object store
    - Both scale out linearly
- More Ceph v Gluster
    - Gluster performs better at higher scales
    - Majority of OpenStack implementations use Ceph
    - Gluster is classic file-serving, second-tier storage
    - Gluster = file storage with object capabilities; Ceph = object storage with block/file capabilities

Rook

- https://github.com/rook/rook
- Distributed storage orchestration for Kubernetes (1.6+) based on Ceph;

minio

- Cloud native+ S3 compatible API
- Used by Deis
- Example built using Gluster...
- Can serve shared NAS
- For Kubernetes

NFS

- Single point of failure
- NFSv2 and NFSv3 have host-based authentication (1). Access control through host and file/directory permissions only.
- NFSv4 has improved security via Kerberos and ACLs
- NFS Ganesha (user level NFS server)

GlusterFS

- Parallel network file storage system
- Good for large static files; immutable files
- Bad for lots of small files; resulting in split brain;
- More complex backup/restore
- Performance degradation under certain load scenarios
- Hard to administer (see Nebula)
- Network authentication, POSIX ACLs
- Version 3.7 supports NFSv4 and pNFS

AWS S3/EBS/EFS

- S3 v EBS v EFS:
    - S3: standalone, durable, storage
    - EBS: For attaching to nodes, single mount; 3 types based on IOPS
    - EFS: Accessible via multiple insteance and other services (shared applications/workloads)
- Comparing Gluster to EFS

Heketi

-

BTRFS

- Copy on write filesystem for Linux
- Used in one example with Minio

BeeGFS

- http://moo.nac.uci.edu/~hjm/fhgfs_vs_gluster.htmlbeegfs

Lustre

- Parallel filesystem used in leadership-class HPC environments

- Used on Comet, for example
- Reviled, but widely used

Flocker:

- Documentation remains unavailable (https://clusterhq.com/flocker/introduction/)

Torus

- https://github.com/coreos/torus
- Development stopped in Feb 2017

Other

- Container Storage Interface
  - https://mesosphere.com/blog/csi-towards-universal-storage-interface-for-containers/
  - Emerging standard?
- https://github.com/cncf/landscap