Ceph is a free-software storage platform, implements object storage on a single distributed computer cluster, and provides interfaces for object-, block- and file-level storage. Ceph aims primarily for completely distributed operation without a single point of failure, scalable to the exabyte level, and freely available.

Ceph replicates data and makes it fault-tolerant, using commodity hardware and requiring no specific hardware support. As a result of its design, the system is both self-healing and self-managing, aiming to minimize administration time and other costs.

Ceph Storage at Lehigh

Research Computing provides a Ceph based storage resource, also called Ceph. In Fall 2018, a 768TB storage cluster was designed, build and deployed to replace the original Ceph cluster, a 1PB storage cluster. In Fall 2020, total storage was increased to 2019TB by the addition of 796TB from Hawk and a further 455TB investment from LTS. Total allocatable storage is 542.87TB with an additional 29.38TB for short term storage on Sol and Hawk.

How is Data Stored in Ceph?

Data is replicated across three disks on three nodes in three racks with distinct power feeds and network paths, secured against simultaneous failure of two full nodes in the primary data center. With current connectivity, the cluster supports an aggregate read/write speed of 3.75GB/s, with capability to increase bandwidth as needed. The Ceph software performs daily and weekly data scrubbing to ensure replicas remain consistent.

NOTE: Ceph does not do backups. If you need backup, one alternative is mount the Ceph project as a network drive and use Crashplan to backup contents in your Ceph project.

System Configuration

- 7 storage nodes
  - One 2.4GHz 16-core AMD EPYC 7351P, 128GB 2666MHz DDR4 RAM, Three Micron 1.9TB SATA 2.5 IN Enterprise SSD
    - Total Raw Storage: 5.7TB for CephFS (Fast Tier)
  - Two Intel 240GB DC S4500 Enterprise SSD (OS only)
  - 13 Seagate 8TB SATA HDD
    - Total Raw Storage: 104TB Ceph (Slow Tier)

- 11 storage nodes
  - One 3.0GHz 16-core AMD EPYC 7302P, 128GB 2666MHz DDR4 RAM, Three 1.9TB SATA SSD
    - Total Raw Storage: 5.7TB for CephFS (Fast Tier)
  - Two Intel 240GB DC S4510 Enterprise SSD (OS only)
  - 9 12TB SATA HDD
    - Total Raw Storage: 108TB Ceph (Slow Tier)

- 10 GbE and 1 GbE network interface
- Debian 10
- Raw Storage: 1916TB (Slow Tier) and 102.6TB (Fast Tier)
- Available Storage: 543TB (Slow Tier) and 29TB (Fast Tier)

Why two tiers of storage?

The original Ceph cluster was designed for archival data storage. In circa 2015, Research Computing decommissioned the storage resource on the then HPC clusters, Corona, Maia, Trits, Capella and Cuda0 and used Ceph as the storage backend instead. This worked fine until Sol, built as a 34 node replacement cluster for Corona, Capella, Cuda0 and Trits, was expanded and upgraded to 56 nodes (81 nodes in Fall 2019) with 66 (120 in Fall 2019) nVIDIA GPUs. The increase in I/O from simulations on Sol caused instability in Ceph. After some research, it was decided that the Ceph replacement should include a fast tier of storage built using SSDs based on the Ceph file system (CephFS) to handle I/O from the ever expanding Sol cluster. The fast tier, CephFS, would provide a distributed global scratch on Sol for writing simulation data from running jobs while the slow tier, Ceph would provide longer term storage of simulation data.

How do I get access to Ceph storage?
To use Ceph as a storage device, Faculty, Staff, Department and Colleges need to purchase a storage project, minimum 1TB for a duration of 5 years or request a storage allocation annually. If purchasing, the cost for 1TB of storage is $375 for 5 year duration. The storage project can be shared with a named group of users including students at no charge. To purchase a Ceph storage project, please contact the Manager of Research Computing with the following:

- Name of the project, default name is the PIs username followed by group for e.g. alp514group
- List of username of users who will have access to the storage. The list can be modified at any time during the 5 year duration.
- Amount of Storage desired (minimum 1TB)
- Banner Index to charge with authorization from Finance Manager

Ceph storage on Sol and Hawk

Ceph is the storage backend for Sol and Hawk. Each Principle Investigator is provided with a 1TB ceph space for his/her research group. If additional space is requested, please include a justification in your compute allocation request or explicitly request a storage allocation. This storage exists as long as your allocation is active and will be deleted (no backups kept) a month after your allocation expires. If you purchase a Ceph project, then your storage will exist for 5 years irrespective of your compute allocation status.

Using Ceph for storage

Ceph storage projects are shared using cifs utilities and can be mounted as a network drive on Windows, Mac OSX and Linux. Ceph projects are mounted on Sol and Hawk. Groups that use Ceph as home directory have access to their projects when they login to Sol. All others can access their Ceph projects at /share/ceph/projectname

Using CephFS on Sol for running jobs

CephFS is available at /share/ceph/scratch/username on the login and compute nodes. Users should use CephFS storage for in flight jobs only and are responsible for transferring simulation data from CephFS to their home directories or Ceph storage projects. The SLURM scheduler automatically creates a folder ${SLURM_JOB_ID} to store data generated from job ${SLURM_JOB_ID}. Users cannot create a folder in CephFS only subfolder in the ${SLURM_JOB_ID}. All data older than 7 days in CephFS will be deleted. It is the responsibility of the user to transfer data from CephFS to their home directories, Ceph project spaces or external storage resource.

Best Storage Practices on Sol

With multiple storage options, home/ceph, cephfs and local scratch, it is the responsibility of users to develop a data management plan for their simulation data.

- home/ceph: permanent storage for the life of your account limited by size of ceph project (default 1TB shared by all members of the research group).
- cephfs: semi-permanent and should moved to permanent storage. This 29TB space is shared by all users and is deleted 7 days after job completion.
- local scratch: temporary - available only for in-flight jobs. This is 500GB space on each compute node that is shared by all users assigned to that node.