INFN-Roma1 & QoS

Alessandro De Salvo

07-02-2020

Storage Infrastructure @ INFN-ROMA1 [1]

- **Roma is hosting 2 Tier2 sites in the same Computing Center**
  - Belonging to ATLAS and CMS
  - Two different storage technologies: DPM (ATLAS) and dCache (CMS)
  - Most of the R&D is on the ATLAS side, although we might decide to expand the testbeds to CMS (and dCache) as well
  - All what we describe in the following slides is manly ATLAS-driven, many details are still only plans

- **DPM (1.13.2)**
  - 1.9 PB of storage space on FC storage systems
  - 10 Gbps connectivity (IPv4 + IPv6)
  - Spinning disks only (NLSAS)

- **dCache**
  - 1 PB of storage space on FC storage systems
  - 10 Gbps connectivity (IPv4)
  - Spinning disks only (NLSAS)

Storage Infrastructure @ INFN-ROMA1 [2]

- **Glusterfs (>= 6)**
  - 3 clusters, mainly used for oVirt/Openstack facilities
  - ~100 TB of available space in replica 2 + arbiter mode or replica 3
- **Ceph (>= luminous – 12.2.X)**
  - 2 clusters, with different QoS targets
  - «Production» cluster
    - Dedicated FC disks, on SAN (Direct Attach) systems, shared with glusterfs (multi-protocol servers)
    - 10 Gbps connectivity
    - ~0.5 PB of raw disk space [bluestore]
    - RDB (cinder), CephFS, RGW (s3) are supported
    - SSD caches are available
  - «Scratch» cluster
    - Using the unused disk space of the WNs or dedicated spare disks installed
    - 1 Gbps collectivity
    - ~150 TB of raw disk space
- **Combined NFS exporter (HA mode)**
  - Both glusterfs and ceph (CephFS, RGW) volumes are exported

QoS [1]

- **The main target is to make an efficient use of the different technologies, available at the site level, and integrate them with the main WLCG Storage Systems**
  - **The current plans are to make use of Ceph, in different ways, integrating with our DPM storage system**
  - **No plans to use Glusterfs for other purposes than oVirt/Openstack VMs**

- **Several different ways of integration are possible, with different levels of QoS**
  - **Virtual DPM hosts on Ceph RBD volumes**
  - **CephFS exports**
  - **RGW/S3 buckets**

QoS [2]

- **Virtual DPM**
  - Fully transparent from the applications and infrastructure point of view
  - VMs running on Ceph RBDs (either replicated or in erasure coding) via oVirt or Openstack
  - Currently available as testbed for the ESCAPE project, running on oVirt VMs and replicated RBDs on the Ceph production cluster
  - Very powerful and easy to maintain solution
    - May be limited by the VMs' architecture, like the network or memory/CPU overhead

- **CephFS**
  - Can provide a single namespace across multiple servers
  - DPM is not directly taking advantage of the single namespace, but certainly the configuration could be simpler this way
  - Limited by the number of MDS enabled/active and their latencies in serving the metadata
  - Can be easily configured with erasure coding, to save space while keeping the data safe
  - Plans to test that in the ESCAPE testbed

QoS [3]

- **RGW/S3**
  - Very interesting option, as it's totally independent from the server the DPM services are running
  - Two options:
    - Posix mount the S3 buckets and use them from the DPM pool disk node
    - Directly use a DPM plugin to consider an S3 endpoint as disk pool node
  - The first option is not the preferred one, as it can be suffering from limitations on the nodes themselves
  - The second option would allow us to mount remote endpoints as well as local ones, without any special configuration other than in the DPM head node
    - Allows bandwith and resource optimization
    - …but requires a specific plugin that at the moment is not fully maintained
    - We are willing to try to help (as users, in general) bringing the plugin back to life
  - RGW/S3 are only limited by the number of RGW daemons that we may have
    - We currently load balance the traffic via haproxy on several services, some performance tests are needed before we may claim this is working well with all payloads

QoS tuning

- **Disk space optimization**
  - Depending on the QoS type we want to support we may choose different parameters, e.g.
    - Replicated storage space (>= 3 replicas) for frequently accessed data
    - Erasure coded data for near-archival data, with different options, depending for example on the level of high availability we want to have on the specific datasets
      - Easy examples are scratch data or local-only data on the Storage Elements
  - We are already experimenting different parametrizations on the erasure code pools, trying to find the optimal values for our configuration
    - More tests needed, and validations with real payloads are needed

- **Tiering and SSD caching**
  - Can use SSD to accelerate in many different ways
    - Pool Tiering
    - Bluestore DB optimization
    - All-Flash pools
  - Each different technique is acting on one or more Ceph facility
  - Our experience with Tiering was not so positive
    - Data corruption, no big performance gain
  - We're now concentrating on the other categories of SSD caching, but we also plan to have new tests of the Tiering with more recent versions of Ceph

Conclusions

- **All the R&D work on QoS is involving Ceph and its integration with with the existing systems (DPM)**
  - **Several years of experience running Ceph, integrating it with different systems and facilities**
  - **Different future improvements include disk space optimization and cache acceleration (tiering/SSD journaling/all-flash pools)**
  - **S3 backend very promising**
- **Testbed with Virtual DPM already available for the ESCAPE project**
- **Different QoS tests are enabled via different underlying configurations on Ceph**
- **R&D activities only limited by funding and manpower!**