Status of the ESCAPE CERN XCache

Riccardo Di Maria

CERN

October 8th, 2019

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

Overview

- CMS XCache - INFN Experience
- CERN Vanilla XCache
- Disk Caching Proxy (DCP) Cluster
- Open Discussion/Questions
  (with respect to possible issues faced within DCP implementation)
- Next Steps - To Do List

Funded by the European Union's Horizon 2020 - Grant N° 824064

ESCAPE Data Infrastructure for Open Science (DIOS)

- Data Lake Infrastructure and Federation Services - Xavier Espinal, CERN
- Data Lake Orchestration Service - Patrick Fuhrmann, DESY
- Integration with Compute Services - Yan Grange, ASTRON-NWO
- Networking - Rosie Bolton, SKAO
- Authentication and Authorization - Andrea Ceccanti, INFN

Simone Campana, CERN as WP leader and Rosie Bolton, SKAO as deputy

Funded by the European Union's Horizon 2020 - Grant N° 824064

CMS XCache - Caching On-Demand Effort

- The goal was to acquire knowledge on XCache by reproducing what INFN is currently deploying for CMS (thanks to Diego and Daniele).
- Useful documentation:
  - https://cloud-pg.github.io/CachingOnDemand/BARE/
  - https://buildmedia.readthedocs.org/media/pdf/xcache/latest/xcache.pdf
- CERN setup:
  - VM on CERN-OpenStack under ESCAPE WP2 CERN project;
  - m2.small flavour (1 VCPUs, 1.8 GB RAM, 10 GB);
  - no external storage mounted.
- Requirements on Cache-side:
  - valid CMS /etc/vomses to be copied from e.g. lxplus;
  - user:group xrootd:xrootd for /data/xrd and /etc/grid-certificates/xrd;
  - (personal certificate used as) host certificate (without passwd).

Funded by the European Union's Horizon 2020 - Grant N° 824064

CMS XCache - Caching On-Demand Effort

- The PoC was quite smooth for a non-expert.
- The XCache origin point was set as xrootd-cms.infn.it, i.e. the CMS Italian global redirector.
- The server's xrootd/xcache configuration had “Direct Mode Proxies” implemented:
  - xrdcp -f -v xroot://testxcache.cern.ch ← XCache
    no need to explicitly pass the origin → //store/data/Run2017C/MET/MINIAOD/29Jun2019_UL2017validation-v1/270000/E200B10F-41DF-454C-8EBA-DC1BBB5ADA3B.root /dev/null
- A similar PoC could be performed also for ATLAS with the only difference (at this stage) of using a “Forwarding Mode Proxies”.
- “Combination Mode Proxies” can be used to allow a client to connect to a particular destination or to forward a connection via a URL type of path.
- Both CMS and ATLAS rely on Docker, K8s, Slate, Singularity Containers, ecc…; however, for now, a Vanilla implementation would perfectly serve the purpose of this work.

Funded by the European Union’s Horizon 2020 - Grant N° 824064

CERN Vanilla XCache - without Auth-method

- The goal is to eventually deploy a caching layer that could serve both LHC-based experiments and Astronomy.
- At this stage, the only requirement on Cache-side is to install xrootd* xrootd-server* xrootd-client*.
- The XCache origin point is set as eulake.cern.ch, thus no Auth-method needed.
- The server's xrootd/xcache configuration has “Combination Mode Proxies” implemented.

- Implementation quite simple → moving towards a Disk Caching Proxy (DCP) cluster.
- Redirector escape-wp2-xcache-01.cern.ch points to testxcache.cern.ch, which is aware of data location.
- However, the redirector expects a reasonable-in-size cache as a host:
  ==> /var/log/xrootd/xcache/cmsd.log ⇐
  190902 15:49:06 26650 Meter: Insufficient space; 7GB available < 11GB high watermark
- The use of a Cluster Management Service directive (cms.space 1g 0.5g) solved the issue.

Funded by the European Union’s Horizon 2020 - Grant N° 824064

CERN Vanilla XCache - DCP Cluster

- CERN setup:
  - redirector: escape-wp2-xcache-01.cern.ch;
  - m2.small flavour (1 VCPUs, 1.8 GB RAM, 10 GB);
  - 2 m2.large flavour (4 VCPUs, 7.3 GB RAM, 40 GB): escape-wp2-xcache-0*.cern.ch;
  - no external storage mounted.
- The first host that tries to connect to the redirector is established as the primary server.
- If caches do not have the file requested in cache, the primary server is contacted by the redirector, and future calls are redirected always to the same cache.
  - This could have pro (global national primary server - redundancy needed) and cons (work would be always carried out by the primary server if files are not yet cached).
  - A test to check the behaviour for overloading should be performed.
- If the primary server is disabled/stopped/..., a restart of xrootd and cmsd deamons is necessary for the redirector and for at least one cache in order to choose another primary server.

Funded by the European Union's Horizon 2020 - Grant N° 824064

CERN Vanilla XCache - DCP Cluster

- If a second cache is forced to cache the same file, the redirector will route the request still to the primary server (50 calls).
  - A restart of xrootd and cmsd deamons is always necessary.
- If all caches are forced to cache the same file and the xrootd and cmsd deamons are restarted in the whole cluster, the request is assigned to each cache in fairshare-mode (50 calls).
- To test:
  - different origin points for different files;
  - a file cached in one storage having different origin points.

Funded by the European Union's Horizon 2020 - Grant N° 824064

Open Discussion/Questions

- Is the establishment of the primary server a modus operandi correct by default?
- Is a dynamical allocation possible for the primary server?
- How could reliability be ensured for the primary server? Has someone tried to have redundancy of the redirector service (for reliability)?
- Why couldn't the redirector be a XCache itself?
- If a file is cached to a second cache, how can the redirector be instructed to route the request in fairshare-mode?
- Is a restart of xrootd and cmsd deamons always necessary for changes in the infrastructure?

Funded by the European Union's Horizon 2020 - Grant N° 824064

CERN Vanilla XCache - To Do List

- Implement monitoring:
  - Ilija ad-hoc solution (python script) and/or BHAM (to follow-up).
- Integration with other storages besides eulake:
  - investigate horizontal scaling with several stages, load balancing, ecc… .
- Investigate other protocols such as http (follow-up with Wei after ATLAS week).
- XCache stress test using HammerCloud mainly to investigate data corruption:
  - setup HC jobs from existing analysis functional tests;
  - create specific HC test to stress the storage;
- CMS implement model using global redirector:
  - caching solution on national level;
  - possibly federated solution with multiple caching layers.
- ATLAS doesn't have a global redirector but multiple ones (relying on Rucio):
  - investigate the integration with Rucio as a first step;
  - they are also interested in a common monitoring solution.

Funded by the European Union's Horizon 2020 - Grant N° 824064

Backup

Funded by the European Union's Horizon 2020 - Grant N° 824064

ESCAPE Goals

- Implementing Science Analysis Platforms for EOSC researchers to stage data collections, analyse them, access ESFRIs' software tools, bring their own custom workflows.
- Contributing to the EOSC global resources federation through a Data-Lake concept implementation to manage extremely large data volumes at the multi-Exabyte level.
- Supporting "scientific software" as a major component of ESFRI data to be preserved and exposed in EOSC through dedicated catalogues.
- Implementing a community foundation approach for continuous software shared development and training new generation researchers.
- Extending the Virtual Observatory standards and methods according to FAIR principles to a larger scientific context; demonstrating EOSC capacity to include existing frameworks.
- Further involving SMEs and society in knowledge discovery.

Funded by the European Union's Horizon 2020 - Grant N° 824064