The Dynafed Data Federator As Grid Site Storage Element

F Berghaus\textsuperscript{1}, K Casteels\textsuperscript{1}, C Driemel\textsuperscript{1}, M Ebert\textsuperscript{1}, F Fernandez Galindo\textsuperscript{3}, F Furano\textsuperscript{2}, O Keeble\textsuperscript{2}, C Leavett-Brown\textsuperscript{1}, M Paterson\textsuperscript{1}, R Seuster\textsuperscript{1}, R Sobie\textsuperscript{1}, R Tafirout\textsuperscript{3}

1. University of Victoria [CA]
2. CERN
3. TRIUMF [CA]
The dynamic federator [Dynafed] redirects HTTP

- Dynafed [1] redirects to nearby storage
- Operating three configurations:
  - Belle-II at UVic:
    - R/O access to cloud storage
    - R/W access to grid storage
  - ATLAS at CERN and TRIUMF:
    - R/W to cloud storage
- Instances operated by others:
  - data-bridge at CERN for *@home
  - Belle-II Dynafed at INFN
  - RAL ECHO
- Part of a WLCG demonstrator
Distributed site with Cloudscheduler

- **Distributed cloud system**
  - Cloudscheduler [2]
  - In production for >8 years

- **User:**
  - DIRAC (Belle-II) or PanDA (ATLAS)

- **Cloudscheduler at UVic and CERN**

- **Cloud Resources:**
  - In Canada, US, UK, Germany, Austria and at CERN
  - \(O(10^3)\) cores - easy to add more

- **CE:** HTCondor & Cloudscheduler
- **SE:** dCache (UVic), EOS (CERN)

- **Goal:** operate as production SE for ATLAS and Belle-II
Belle-II storage element at UVic

- **DIRAC SE: UVIC-SE**
  - HTTP/WebDAV -> Dynafed
  - SRM -> UVic dCache (232TB -> 400TB)
- Authentication with X.509 using VOMS roles
  - Configured for ATLAS and Belle-II
- Back-end storage
  - Grid site near cloud if available (read-only)
  - Object storage over S3
    - MinIO [3] (100GB/instance)
    - Manual replication of inputs
    - CephS3 (20TB)
- UVic dCache accessible for read & write
  - Via Dynafed for HTTP/WebDAV
ATLAS storage element at TRIUMF

- Site: CA-TRIUMF-DYNAFED
  - DATADISK & SCRATCHDISK (30TB)
  - Analysis & production queue
- Gateway to TRIUMF Ceph S3
- HTTP/WebDAV only
- X.509 authentication with authorization using VOMS roles
  - Macaroons supported
- Production scaling issues:
  Apache [4] hangs/dies every few days
  - HAproxy balance “first”: fail over to second Dynafed
  - Shared Memcached [5]
ATLAS storage element at CERN

- Site: CERN-EXTENSION
  - DATADISK & SCRATCHDISK (50TB)
  - Production queue
- Gateway to CERN Ceph S3
- HTTP/WebDAV only
- X.509 authentication with authorization using VOMS roles
  - Macaroons supported
- Queue and storage set to Test
Issue

- Mechanism:
  - Grid: User is responsible, Want-Digest [RFC3230]
  - Cloud: Provider is responsible, Content-MD5 [RFC1544]
- Algorithm:
  - Grid: ADLER32 [RFC1950], for many reasons
  - Cloud: MD5 [RFC1321]

Solution

- Dynafed handles Want-Digest requests:
  - Use native support of grid storage
  - Call out to user executable if Want-Digest is not supported
## Checksum implementation

<table>
<thead>
<tr>
<th></th>
<th>Checksum Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CERN</td>
<td>Calculated by load balanced, external web service</td>
</tr>
<tr>
<td>TRIUMF</td>
<td>Calculated on Ceph Rados gateways</td>
</tr>
<tr>
<td>UVic</td>
<td>Calculated on Dynafed</td>
</tr>
</tbody>
</table>

- Once calculated store checksum digest as object metadata
  - Future requests use metadata value
  - Implemented in `dynafed_storagestats [6]`
3\textsuperscript{rd} party COPY [TPC]

- Object storage does not implement TPC requests
- Dynafed handles copy requests:
  - Forward to storage implementations that support TPC
  - Call out to user executable if TPC is not supported
- Challenge: must report performance markers

<table>
<thead>
<tr>
<th>Dynafed</th>
<th>COPY Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CERN</td>
<td>SSH call to load balanced set of server</td>
</tr>
<tr>
<td>TRIUMF</td>
<td>SSH call to Ceph Rados gateways</td>
</tr>
<tr>
<td>UVic</td>
<td>Executed locally on Dynafed</td>
</tr>
</tbody>
</table>
Accounting and reporting

- Provide storage space accounting using WLCG JSON [7]
  - Used by Rucio for ATLAS and DIRAC for Belle-II
- Produce content dumps to allow dark data checks
- Add free space information to memcached for Dynafed
  - Dynafed configured to only redirect WRITE requests to storage with sufficient free space
- Implemented in reports and stats feature of dynafed_storagestats
Conclusion

- Dynafed allows access to object storage as grid storage element

- Implemented workarounds for differences in cloud and grid storage
  - Checksum implemented by call out and object metadata
  - 3rd party copy implemented by call out
  - Reporting and accounting implemented using dynafed_storagestats

- Operating Dynafed as production SE in Belle-II and ATLAS
Dynafed storage stats

- https://pypi.org/project/dynafed-storagestats
  
  pip3 install dynafed-storagestats

- Features:
  - Checksums
    - get: retrieved checksum from object metadata
    - put: store checksum digest in object metadata
  - Reports
    - filelist: dump all files in a dynafed path
    - storage: report free space and quota information
  - Stats: add free space and quota information to memcached for Dynafed

- Currently running with:
  - **Azure** Storage Blob, **AWS S3**, **Ceph S3**, **Minio S3**, **DPM** (via WebDAV), **dCache** (via WebDAV)
Bibliography


2) UVic HEP research computing, Cloucscheduler [software], available from https://github.com/hep-gc/cloucscheduler

3) MinIO project, MinIO [software], available from https://min.io

4) Apache software foundation, Apache HTTP Server [software], version 2.4, available from https://httpd.apache.org

5) Memcached project, memcached [software], version 1.5.19, available from https://memcached.org

6) HEP research computing, dynafed_storagestats [software], version 1.0.28, available from https://pypi.org/project/dynafed-storagestats

7) Worldwide LHC computing grid, Storage Space Accounting [standard], available from https://twiki.cern.ch/twiki/bin/view/LCG/StorageSpaceAccounting