Rucio/BigData Express/SENSE (ROBIN)

a Next Generation High Performance Data Service Platform

Dr. Wenji Wu (wenji@fnal.gov)

Thursday, March 11, 2021
Many people’s hard work

FNAL: Wenji Wu, Liang Zhang, Qiming Lu, Amy Jin, Phil DeMar, Robert Illingworth

iCAIR/StarLight: Joe Mambretti, Se-young Yu, Fei Yeh, Jim-Hao Chen

ESnet: Inder Monga, Xi Yang, Tom Lehman, Chin Guok, John Macauley
Agenda

• Motivation

• ROBIN: a next generation high performance data service platform
  • Architecture
  • Key Mechanisms

• Initial Evaluation
  • An international testbed
  • Experiments
Motivation (I)

- Big data has emerged as a driving force for scientific discoveries.

- Large scientific instruments generates large amount of data.

- Science data must be collected, indexed, archived, shared, and analyzed, typically in a widely distributed, highly collaborative manner.
Motivation (II)

Managing and moving extremely large volumes of science data worldwide is a special multidimensional challenges!

Need a comprehensive service solution that incorporates:

- Services designed for scientists
- Scientific workflows
- Data management
- Science DMZs
- Large volume data transfer services
- Orchestration
- High-performance networking

Our Solution

ROBIN (Rucio/BigData Express/SENSE):
A Next Generation High-performance Data Service Platform
ROBIN (Rucio/BigData Express/SENSE)
A Next Generation High-performance Data Service Platform

Scientific Applications (e.g., LSST, CMS)

Rucio Data Management Service
- Smart Namespace
- Easy Integration

BigData Express High-performance Data Transfer Service
- Storage Support
- AA Support
- Peer-to-peer, Scalable, Extensible
- High Performance Data Transfer Engine

SENSE Smart Network Service
- Consistency
- Proven Track Record
- End-to-end Performance Optimization
- CILogon-base Security
- Software Defined-Networking (SDN) for End-to-End Virtual Guaranteed Network
- A Highly Intuitive “Intent”-based Interface
ROBIN (Rucio/BigData Express/SENSE)
A Next Generation High-performance Data Service Platform
ROBIN Key Mechanisms

• Site Registration

• Rucio/BigData Express (BDE) job launching mechanism

• On-demand provisioning of end-to-end network path with guaranteed Qos

• Security
Site Registration

- Register a BigData Express site as an RSE with the Rucio server
  - The RSE name
  - The information necessary to access the new RSE
    - Hostname, port, protocol, and local file system path
  - The distance metric between the new RSE and other RSEs

A new protocol “bde” is defined to support BDE-based data transfer.
Rucio/BDE Job Launching Mechanism

1. A Rucio client uses Rucio CLI to request replication on a destination RSE.
2. The client sends the replication request to the Rucio server.
3. The Rucio server creates a replication rule for the request and generates the data transfer tasks. The tasks are temporarily kept in a task queue.
4. The Rucio server regularly pulls tasks from the queue. It ranks the sources for each task, selects the protocol “bde” for src/dst RSEs, submits the tasks in groups to BDE.
5. BDE schedules and assigns resources (DTNs, network) to execute the data transfer tasks.
   - BDE calls SENSE to provision WAN paths with guaranteed QoS between sites.
6. After the DTNs and the paths have been successfully reserved, BDE launches the data transfer tasks, monitors the progress of the tasks, retries in case of errors, and notifies the Rucio server upon completion.
7. The Rucio server closely monitors the status of the transfers. A failed data transfer will be resubmitted in the task queue for retries until the maximum retry limit is reached.
8. The Rucio server updates the internal states and notifies the client upon completion.
On-demand Provisioning of End-to-end Path with Guaranteed Qos

A End-to-End Data Transfer Loop with Guaranteed QoS
Security

• Keep each system’s security intact

• Execute a logic mapping between them to enforce security at all levels
  • Direct mapping between Rucio and BDE accounts with X509 certificate delegation
  • Each BDE site, acting as a SENSE client, with pre-configured client credential.

<table>
<thead>
<tr>
<th>Systems</th>
<th>Authentication/Authorization methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rucio</td>
<td>Username/password, X509 certificates, Kerberos tickets, SSH-RSA public key</td>
</tr>
<tr>
<td>BigData Express</td>
<td>Username/password, X509 certificates</td>
</tr>
<tr>
<td>SENSE</td>
<td>Username/password, OIDC</td>
</tr>
</tbody>
</table>
ROBIN Cross-Atlantic Testbed

StarLight site:
- DTN: dtn110.sl.startap.net, with several Intel NVMe drives for data storage, a 100GE Mellanox NIC for data transfer, and a 1G NIC for control.
- Head node: 165.x.x.157, with a 1G NIC for control.

CERN site:
- DTN: dtn01.cern.ch, with a rotational disk for data storage, a 10GE Mellanox NIC for data transfer, and 1G NIC for control.
- Head node: cixp-urfnet.cern.ch, with a 1G NIC for control.
Experiments

1. Register each BDE site in the testbed as an RSE with the Rucio server

2. Create an experiment file named 25g-1.bin and register the file with the Rucio server

3. Use the Rucio client to submit a request to the Rucio server to replicate the registered file from StarLight to CERN
$rucio-cmd rucio list-file-replicas test:25g-1.bin

+---------+----------+----------+---------+---------+-----------------+
| SCOPE   | NAME     | FILESIZE | ADLER32 | RSE: REPLICA  |
+---------+----------+----------+---------+---------+-----------------+
| test    | 25g-1.bin| 26.844 GB| 49576448| STARLIGHT-SITE: bde://165.124.33.157:5000/165.124.33.142/disk0/25g-1.bin |
+---------+----------+----------+---------+---------+-----------------+

$rucio-cmd rucio list-rules test:25g-1.bin

<table>
<thead>
<tr>
<th>ID</th>
<th>ACCOUNT</th>
<th>SCOPE:NAME</th>
<th>STATE [OK/REPL/STUCK]</th>
<th>RSE_EXPRESSION</th>
<th>COPIES</th>
<th>EXPIRES (UTC)</th>
<th>CREATED (UTC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>554acd9b7ddb4a319a37308a1285753f</td>
<td>root</td>
<td>test:25g-1.bin</td>
<td>OK[0/0/0]</td>
<td>STARLIGHT-SITE</td>
<td>1</td>
<td></td>
<td>2020-08-31 04:04:32</td>
</tr>
</tbody>
</table>

The replica and the replication rule for 25g-1.bin

$rucio-cmd rucio list-rules test:25g-1.bin

<table>
<thead>
<tr>
<th>ID</th>
<th>ACCOUNT</th>
<th>SCOPE:NAME</th>
<th>STATE [OK/REPL/STUCK]</th>
<th>RSE_EXPRESSION</th>
<th>COPIES</th>
<th>EXPIRES (UTC)</th>
<th>CREATED (UTC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c510e4cd53b44b77b6c6db2c367036766</td>
<td>root</td>
<td>test:25g-1.bin</td>
<td>REPLICATING[0/1/0]</td>
<td>CERN-SITE</td>
<td>1</td>
<td></td>
<td>2020-08-31 04:09:34</td>
</tr>
<tr>
<td>554acd9b7ddb4a319a37308a1285753f</td>
<td>root</td>
<td>test:25g-1.bin</td>
<td>OK[1/0/0]</td>
<td>STARLIGHT-SITE</td>
<td>1</td>
<td></td>
<td>2020-08-31 04:04:32</td>
</tr>
</tbody>
</table>

The replication rule created for the Rucio data replication request
Results (II) – BigData Data Transfer Process

Subtask - 5f4d62bb856a310d86557449

FROM: StarLight
source: /home/sereser/esederei/dem/mmed/bdl
/disk7/23q-1.bl, and etc.

TO: CERN
/home/sereser/esederei/dem/mmed/bdl
/home/cern/home
/disk7/

Total Size: 25.000 GB
number of files: 1
(26443545600 bytes)

Job Status
1. Preprocessing
   - query reservations
   - DWM matching
2. Setup network path
   - planning
   - SENSE path
   - path verification
3. Data transfer
   - launch
   - transfer
4. Resource cleanup
   - SENSE path

Log Messages
[3/31/2020, 2:32:09 PM] - Path verification

Source DTN | Destination DTN | Type | Current Rate | Average Rate | Transferred
--- | --- | --- | --- | --- | ---
165.124.33.142 | 160.91.249.29 | boot_offset | 133.60 MB/s | 128.64 MB/s | 25960.00 MB

Flow Status: Waiting to be closed
## Results (III) – BDE/SENSE Interactions

<table>
<thead>
<tr>
<th>BDE Control Event</th>
<th>Transaction Time</th>
<th>SENSE Control Event</th>
<th>Transaction Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Service Negotiation</strong></td>
<td>5s</td>
<td>Compute Service Intent (initial)</td>
<td>3s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Re-Compute Service (negotiate)</td>
<td>2s</td>
</tr>
<tr>
<td><strong>Service Reservation</strong></td>
<td>9s</td>
<td>Reserve with RMrs</td>
<td>7s</td>
</tr>
<tr>
<td><strong>Service Allocation</strong></td>
<td>94s</td>
<td>Commit with RMrs</td>
<td>34s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Verify Service Model</td>
<td>51s</td>
</tr>
<tr>
<td><strong>Service Deallocation</strong></td>
<td>60s</td>
<td>Release with RMrs</td>
<td>4s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Commit with RMrs</td>
<td>33s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Verify Service Model</td>
<td>12s</td>
</tr>
</tbody>
</table>
Future Plans (Work in Progress)

- Continue to evaluate/test ROBIN
  - 100Gbps international WAN paths
  - High-end DTNs
  - Multiple site deployment
  - Increased automation
  - Enhanced parameter analytics

- Compare ROBIN with Rucio/FTS
Conclusion

• **ROBIN (Rucio/BigData Express/SENSE)**
  A Next Generation High-performance Data Service Platform

• A unique comprehensive set of integrated services designed specifically for managing and moving extremely large amounts of data over long distances
Questions?

Additional Information