Rucio/BigData Express/SENSE (ROBIN)
a Next Generation High Performance Data Service Platform

Dr. Wenji Wu (wenji@fnal.gov)
Wednesday, October 7, 2020
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 challenge!

Need a comprehensive solution that incorporates:

- Service designed for scientists
- Scientific workflows
- Data management
- Science DMZs
- High-performance 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
- Peer-to-peer, Scalable, Extensible
- Software Defined-Networking (SDN) for End-to-End Virtual Guaranteed Network
- A Highly Intuitive “Intent”-based Interface

BigData Express High-performance Data Transfer Service
- Storage Support
- AA Support
- End-to-end Performance Optimization
- High Performance Data Transfer Engine
- CILogon-base Security

SENSE Smart Network Service
- Consistency
- Proven Track Record
- Consistency
- Proven Track Record
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.
1. A Rucio client sends a replication request to the Rucio server.
2. 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*.
3. 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.
4. 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.
5. 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.
6. 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.
7. The Rucio server updates the internal states and notifies the client upon completion.
On-demand Provisioning of End-to-end Path 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
Results (I) – Site Registration

RSE Name: STARLIGHT-SITE

$ rucio-cmd rucio-admin rse info STARLIGHT-SITE

Settings:
-------------------
third_party_copy_protocol: 1
rse_type: DISK
domain: [u'last', u'write']
availability_delete: True
delete_protocol: 1
rse: STARLIGHT-SITE
deterministic: False
write_protocol: 1
read_protocol: 1
availability_read: True
staging_area: False
credentials: None
availability_write: True
lifepath_algorithm_hash
sign_url: None
volatile: False
verify_checksum: True
id: belle5b690ee4aeb16c6bc1d8b36f
Attributes:
-------------------
bdo: https://165.x.x.157:5000
name_convention: BDE
STARLIGHT-SITE: True
Protocols:
-------------------
bdo
extended_attributes: {u'v.storage_portal': {u'v.storage_url': u'/api/v1/object-store?list&format=json&prefix=/a/b/c&domain=[u'last', u'write'], u'v.last modifies': 1, u'v.last deletes': 1}, u'v.last': u'/api/v1/object-store?list&format=json&prefix=/a/b/c&domain=[u'last', u'write'], u'v.last modifies': 1, u'v.last deletes': 1}, u'v.last': u'/api/v1/object-store?list&format=json&prefix=/a/b/c&domain=[u'last', u'write'], u'v.last modifies': 1, u'v.last deletes': 1}
schema: bdo
port: 5000
impl: rucio.rse.protocols.bdo.Default
Usage:
-------------------
rucio
files: 0
used: 0
rse: STARLIGHT-SITE
updated_at: 2020-08-26 20:12:58
free: None
source: rucio
total: 0
rse_id: belle5b690ee4aeb16c6bc1d8b36f

RSE Name: CERN-SITE

$ rucio-cmd rucio-admin rse info CERN-SITE

Settings:
-------------------
third_party_copy_protocol: 1
rse_type: DISK
domain: [u'last', u'write']
availability_delete: True
delete_protocol: 1
rse: CERN-SITE
deterministic: False
write_protocol: 1
read_protocol: 1
availability_read: True
staging_area: False
credentials: None
availability_write: True
lifepath_algorithm_hash
sign_url: None
volatile: False
verify_checksum: True
id: e1786435ee6d40e3eac7e683171a70
Attributes:
-------------------
bdo: https://cisp-surfnet-ds.cern.ch:5000
CERN-SITE: True
name_convention: BDE
Protocols:
-------------------
bdo
extended_attributes: {u'v.storage_portal': {u'v.storage_url': u'/api/v1/object-store?list&format=json&prefix=/a/b/c&domain=[u'last', u'write'], u'v.last modifies': 1, u'v.last deletes': 1}, u'v.last': u'/api/v1/object-store?list&format=json&prefix=/a/b/c&domain=[u'last', u'write'], u'v.last modifies': 1, u'v.last deletes': 1}, u'v.last': u'/api/v1/object-store?list&format=json&prefix=/a/b/c&domain=[u'last', u'write'], u'v.last modifies': 1, u'v.last deletes': 1}
schema: bdo
port: 5000
impl: rucio.rse.protocols.bdo.Default
Usage:
-------------------
rucio
files: 0
used: 0
rse: CERN-SITE
updated_at: 2020-08-26 20:12:57
free: None
source: rucio
total: 0
rse_id: e1786435ee6d40e3eac7e683171a70
Results (II) – Rucio Rules

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

The replication rules created after the Rucio data replication request
Results (III) – BigData Express Data Transfer Process

- **FROM:** StarLight
  - Source: /dev/zero/zero/zero/dev/mm/0.nl
  - /disk1/35g-1.blk, and etc.

- **TO:** CERN
  - Source: /dev/mapper/centos_dtn01-home
  - /disk0/

- **Total Size:** 25.000 GB
  - Number of files: 1
  - 26843545600 bytes

- **Job Status**
  1. Preprocessing
  2. Setup network path
  3. Data transfer
  4. Resource cleanup

- **Log Messages**
  - [2020/03/03 2:31:12 PM] - DSN matching
  - [2020/03/03 2:31:11 PM] - Setup DSN path
  - [2020/03/03 2:32:03 PM] - Setup DSN path
  - [2020/03/03 2:32:03 PM] - Path verification
  - [2020/03/03 2:32:13 PM] - Launching
  - [2020/03/03 2:34:57 PM] - Transferring
  - [2020/03/03 2:45:31 PM] - Teardown DSN path

- **Source DTN:** 165.123.1.12
- **Destination DTN:** 162.96.240.29
- **Type:** bucket
- **Current Rate:** 123.60 MB/s
- **Average Rate:** 128.64 MB/s
- **Transferred:** 29660.00 MB
## Results (IV) – 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>Service Negotiation</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>Service Reservation</td>
<td>9s</td>
<td>Reserve with RMs</td>
<td>7s</td>
</tr>
<tr>
<td>Service Allocation</td>
<td>94s</td>
<td>Commit with RMs</td>
<td>34s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Verify Service Model</td>
<td>51s</td>
</tr>
<tr>
<td>Service Deallocation</td>
<td>60s</td>
<td>Release with RMs</td>
<td>4s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Commit with RMs</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