Status report from Tokyo Tier2 at ICEPP

Tomoe Kishimoto
ICEPP, The University of Tokyo

Jan. 29 2017
Main projects in ICEPP

- ATLAS experiment at LHC
- MEG experiment at PSI ($\mu \rightarrow e \gamma$ rare decay)
- R & D for ILC

ATLAS–Japan group
- 17 institutes and ~150 members
- Tokyo Tier2 is the only WLCG site in ATLAS–Japan
ICEPP regional analysis center

✓ Resource overview
  - Support only ATLAS VO in WLCG (Tier2) and provide ATLAS-Japan dedicated resources (local use)
  - Hardwares are leased, and are replaced in every three years
  - ~10000 CPU cores including service instances and ~10 PB disk storage (T2 + local use)
    ‣ 18.11HS06/core (Intel Xenon E5–2680 v3)

Single VO and uniform architecture

✓ Operation team
  - H.Sakamoto (will retire in next Mar.),
    J.Tanaka, T.Mashimo, N.Tomoaki,
    T.Kishimoto, N.Matsui
## WLCG pledge

<table>
<thead>
<tr>
<th></th>
<th>CPU [HS06]</th>
<th>DISK [TB]</th>
<th>(*)LOCALGROUPDISK [TB]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2017</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pledge</td>
<td>34,000</td>
<td>4,000</td>
<td>-</td>
</tr>
<tr>
<td>Deployed</td>
<td>111,268</td>
<td>4,000</td>
<td>1,000</td>
</tr>
<tr>
<td><strong>2018</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pledge</td>
<td>40,000</td>
<td>4,800</td>
<td>-</td>
</tr>
<tr>
<td>Deployed</td>
<td>111,268</td>
<td>4,800</td>
<td>1,000</td>
</tr>
</tbody>
</table>

(*) Grid disks for ATLAS–Japan group

- Tier2 resources
  - The current system (4th system) satisfies 2018 WLCG pledge
  - New system will be provided for 2019–2021
    - (Need to migrate 5.8 PB data to the new system...)

✓ Tier2 resources

- The current system (4th system) satisfies 2018 WLCG pledge
- New system will be provided for 2019–2021
  - (Need to migrate 5.8 PB data to the new system...)

Jan. 29 2018
Site status in ATLAS

✓ Fraction of # of completed jobs for the last year:
  - Production: 4.0% (Tier2) – 2.2% (All)
  - Analysis: 6.3% (Tier2) – 4.1% (All)
  ← Good contributions

✓ > 99% site availability has been achieved using the 4th system (for 2 years)

#' of completions & CPU consumption

# of ATLAS-J authors ~ 150
# of ATLAS authors ~ 3000
CE and batch system update

✓ Migration from “CREAM+Torque/Maui” to “ARC+HTCondor” has been completed

HTCondor pool occupancy

✓ Introduced dynamic partitioning for single- and multi-core jobs
  - Improvement of CPU utilization was observed
  - (Reported at AFAD2017, see backup)
Disk storage is managed by DPM, and its database is MySQL.

Previous configuration of SE:

- No redundancy in MySQL database..., risk of producing dark data.
MySQL replication

✓ Semi-synchronous replication in MySQL has been implemented
  - Master server is replicated to slave server automatically
    → Can use slave server as new master server when a trouble occurs in master server
  - Daily backup from slave server (takes ~10 mins)
    → No impact on master server performances

lcg-se01.icepp.jp

- Fusion–IO ioDrive has been attached for database spaces to reduces time for maintenances
- Binary log increases by 8GB per day
ATLAS data management monitor

Transfer efficiency: source is Tokyo

File deletion efficiency

Downtime for the database upgrade

✓ No issues have been observed after the database upgrade
International network status

✓ SINET5 is a NREN in Japan
- 2016 Mar.: 20 Gbps for London and 100 Gbps for LA become available
- 2016 Apr.: LHCONE peering for EU sites
  › ICEPP←CERN latency improved by 30%
- 2016 Sep.: LHCONE peering for US sites

- 2017 Sep.: LHCONE peering for ASGC, KREONET2 and TEIN via JGN-X VRF in HongKong (100 Gbps for Tokyo ⇔ HongKong)
Data transfer with other site

Total transfer volumes last year

Europe: 4.2 PB (67 %)
North America: 2.0 PB (32 %)
Asia: 94 TB (2%)

Incoming data
~800MB/s in one day average

European: 4.2 PB (67 %)
North America: 2.0 PB (32 %)
Asia: 94 TB (2%)

Outgoing data
Status of IPv6 migration

✓ Long pause due to problems of main switch firmware…
   - The firmware was fixed last year, and our procedure/experience for IPv6 filtering have been matured

✓ IPv6 migration plan:
1. Enable the dual stack mode of perfSONARs (done)
2. Enable LHCONE peering via IPv6, need to discuss with SINET and University network team (by end of Aug. 2018)
3. Enable the dual stack mode of storage system (by end of Dec. 2018)

perfSONAR is a key tool to measure IPv6 performances
PerfSONAR measurements

 ✓ Data measured by PerfSONARs are also stored to ELK stack for good visualization
 ✓ Latency tests with ATLAS Tier1s:

 - IPv6 tests are stable so far, but differences of performance are expected since LHCONE peering for IPV6 is not ready yet
PerfSONAR measurements

IPv4

Packet loss

Perfsonar line packet destination

Bandwidth

Perfsonar line throughput-1g destination

IPv6

Perfsonar line packet destination ipv6

Perfsonar line throughput-1g destination ipv6

IPv4

IPv6
Summary

✓ Tokyo Tier2 with the 4th system is running
  - Providing enough computing resources for ATLAS
  - > 99% site availability is achieved

✓ Migration from Torque/Maui to HTCondor has been completed

✓ Redundancy in MySQL database has been implemented
  - Reduced the risk of producing dark data

✓ International network connectivity has been improved thanks to Japanese NRENs (SINET and JGN)

✓ IPv6 migration is ongoing
  - PerfSONARs are IPv6 ready, tests are working well
CPU utilization

- Improvement of CPU utilization has been observed thanks to the dynamic partitioning.
- HTCondor is stable so far.

There was a pbs_server crush

Test jobs (e.g. ops job) are overcommitted in HTCondor system

<table>
<thead>
<tr>
<th></th>
<th>November 2016</th>
<th>December 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>week1</td>
<td>week2</td>
</tr>
<tr>
<td>Static partitioning</td>
<td>-</td>
<td>98.8%</td>
</tr>
<tr>
<td>(Torque/Maui)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic partitioning</td>
<td>-</td>
<td>99.4%</td>
</tr>
<tr>
<td>(HTCondor)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tier2 configuration

- **Network**
  - Brocade MLXe-32 x 2
  - Non-blocking 10Gbps
  - Inter link 16 x 10Gbps
  - 10GE (SFP+) 176 ports

- **Tier2**
  - DPM file servers
  - LCG service nodes
  - LCG worker nodes

- **Non-grid**
  - GPFS/NFS file servers
  - Tape servers
  - Non-grid service nodes
  - Non-grid computing nodes

Main switches: continued use from previous (3rd) system