HPC at CSNS

Yakang LI
Institute of High Energy Physics, CAS
HEPiX Fall 2017, KEK
About CSNS—China Spallation Neutron Source

- **Studying neutron characteristics and exploring microstructure of matter**
- **High-level scientific research platform oriented to dimensional academic subjects**
- **The neutron beam was successfully obtained for the first time on 28th August**
- **Scientific research at CSNS needs the support of a high-performance computing environment**
Computing Requirements
CSNS Data Flow

- Data Transfer
- Metadata Catalogue
- Data Process
- Cloud Analysis
Computing requirements

- **Project Construction**
  - Equipment Simulation (accelerator, detector)
  - Neutron Transmission Process in the target and instruments
  - Softwares (ORBIT, Geant 4, PHITS, MCNP) based on MPI

- **Neutron Data Processing**
  - Raw data -> NeXus User data
  - Data Reconstruction based on MPI and GPU

- **Neutron Cross Discipline Computing**
  - Cloud Computing
  - Materials Science, Life Sciences, New Energy and so on.
  - Softwares (vasp)
Computing requirements

- **Cloud computing platform**
  - Mainly for Instrument Users
- **HPC platform**
  - Providing MPI parallel computing environment
  - Support for GPU computing

- **Phase One:** 3 instruments *(will be finished soon)*
  - 500 cores (cloud) + 2000 cores (HPC) + some GPU Cards

- **Phase Two:** 17 instruments
  - The amount of resources currently required * 7
  - As the number of users increases, the demand for resources becomes greater
Design & Implementation of HPC
CSNS platform Based on SLURM

- **Login Farm**
- **CLI**
- **Dashboard**
- **Glusterfs cluster**
- **10Gb Ethernet**
- **InfiniBand FDR 56Gb/s**
- **CPU cores**
- **Queue A**
- **Queue B**
- **Queue C**
- **Backup**
- **10Gb switch**
- **MySQL**
- **GPU Cards**
Login Farm

- **Uniform login entry**
  - Software compiling
  - Job submission

- **Keepalived + LVS**
  - Load Balancer
  - Login nodes

- **LVS working mode**
  - DR (Direct routing)

- **Scheduling algorithm**
  - WRR (Weighted Round-Robin Scheduling)
User Management & Authentication

- **User Management**
  - LDAP

- **Configuration**
  - Nsswitch.conf
  - Nslcd.conf
  - System-auth-ac/passwd-auth-ac
  - Ldap.conf

- **Authentication and Authorization**
  - Integrate IHEP Unified Authentication System

---

<table>
<thead>
<tr>
<th>Ldap account</th>
<th>Role in linux</th>
</tr>
</thead>
<tbody>
<tr>
<td>uid</td>
<td>Linux system user name</td>
</tr>
<tr>
<td>uidNumber</td>
<td>The uid number corresponding to the user name</td>
</tr>
<tr>
<td>gidNumber</td>
<td>The gid number of the group to which the user belongs</td>
</tr>
<tr>
<td>loginShell</td>
<td>Shell used by users</td>
</tr>
<tr>
<td>homeDirectory</td>
<td>User home directory</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Ldap group</th>
<th>Role in linux</th>
</tr>
</thead>
<tbody>
<tr>
<td>cn</td>
<td>Linux group name</td>
</tr>
<tr>
<td>gidNumber</td>
<td>The GID number of the group to which the user belongs</td>
</tr>
<tr>
<td>memberUid</td>
<td>A list of users belonging to this group</td>
</tr>
</tbody>
</table>
User Management & Authentication

- **Authentication**
  - All HPC users are managed and authenticated by the unified authentication system

- **Authorization**
  - The HPC system can only be used by authorized users via IHEP Unified Authentication System
Storage System

- **Glusterfs**
  - Experimental data
  - User home directory

- **Redundancy**
  - RAID 5, replicated volume

- **Mounts to each node via the gluster native protocol**

- **Read-write performance test**
  - 1MB, 640GB, 8 process
  - Read: 1GB/sec
  - Write: 500MB/sec
Software Repository——CVMFS

- **CernVM File System (CernVM-FS)**

- **Stratum0**

- **Stratum1**
  - csns-stratum0
  - ihep-stratum1
Supporting System: HPC web UI

- HPC web UI
- Status Monitor, Feature Monitor, Job Manage, etc.
Supporting System: HPC web UI

- **Monitor Job Features**
### Supporting System: HPC web UI

**Job Management**

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>User</th>
<th>UseTime</th>
<th>Status</th>
<th>Query</th>
<th>Cores</th>
<th>Nodes</th>
<th>RunNode</th>
<th>Operate</th>
</tr>
</thead>
<tbody>
<tr>
<td>6441</td>
<td>p80.sh</td>
<td>zuots(1077)</td>
<td>7:05:03:31</td>
<td>R</td>
<td>public</td>
<td>112</td>
<td>4</td>
<td>node[28-31]</td>
<td></td>
</tr>
<tr>
<td>6505</td>
<td>Zr47-5</td>
<td>chenhuican(1080)</td>
<td>6:03:43:41</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node32</td>
<td></td>
</tr>
<tr>
<td>6856</td>
<td>ev-u-11af</td>
<td>xujuping(1066)</td>
<td>19:34:04</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node17</td>
<td></td>
</tr>
<tr>
<td>6857</td>
<td>ev-u-62af</td>
<td>xujuping(1066)</td>
<td>19:19:40</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node18</td>
<td></td>
</tr>
<tr>
<td>6874</td>
<td>13md300</td>
<td>botao(1064)</td>
<td>07:26:17</td>
<td>R</td>
<td>public</td>
<td>56</td>
<td>2</td>
<td>node[14-15]</td>
<td></td>
</tr>
<tr>
<td>6875</td>
<td>si-o-he-1</td>
<td>yinwen(1078)</td>
<td>07:24:36</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node3</td>
<td></td>
</tr>
<tr>
<td>6876</td>
<td>PC</td>
<td>pflu(1063)</td>
<td>07:04:34</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node7</td>
<td></td>
</tr>
<tr>
<td>6879</td>
<td>zcb-30</td>
<td>zhangcb(1074)</td>
<td>05:00:23</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node19</td>
<td></td>
</tr>
<tr>
<td>6880</td>
<td>zcb-30</td>
<td>zhangcb(1074)</td>
<td>05:00:09</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node20</td>
<td></td>
</tr>
<tr>
<td>6881</td>
<td>zcb-30</td>
<td>zhangcb(1074)</td>
<td>04:59:53</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node21</td>
<td></td>
</tr>
<tr>
<td>6882</td>
<td>zcb-30</td>
<td>zhangcb(1074)</td>
<td>04:59:42</td>
<td>R</td>
<td>public</td>
<td>28</td>
<td>1</td>
<td>node24</td>
<td></td>
</tr>
</tbody>
</table>
Supporting System

- ELK Stack
- Zabbix
- Ganglia
CSNS HPC——Benchmark

- **Computing resource (32 nodes)**
  - 896 Cores, 4TB Memory
  - 94 users

| testing environment | Number of computing nodes: 32
|---------------------|---------------------------------------------------------------|
|                     | cpu: Intel Xeon E5-2680v4(2.4GHz/14c)/9.6GT/35ML3/2400
|                     | Network: infiniband 56Gbps
|                     | OS: CentOS 7.3
| theoretical value  | Computing nodes: 34406.4Gflops
|                     | Optimum test result
|                     | \(N^\ominus\) \(NB^\ominus\) \(P^\ominus\) \(Q^\ominus\) \(Time^\ominus\) \(Gflops^\ominus\)
| 620288\(\ominus\) | 192\(\ominus\) 4\(\ominus\) 8\(\ominus\) 5322.80\(\ominus\) 2.98917e+04\(\ominus\)

- **Linpack efficiency**
  - Continuous computing capability: 2.98917e+04

- **Conclusion**
  - 86.87%
Summary
Summary

- The CSNS HPC Platform has been completed and running well
- The performance of the HPC using the Linpack test meets the construction requirements, but there is also some room for optimization
- The next phase will add GPU computing nodes and large memory computing nodes to HPC platform
- Docker based flexible job HPC system is ongoing
- Would like to learn more from other sites
Thanks for Your Attention!

Yakang Li
CSNS-IHEP, CAS
HEPiX 2017 Fall, KEK