The Beijing Tier-2 Site:

current status and plans

Lu Wang, Computing Center
Institute of High Energy Physics,
Beijing
3/15/10
Outline

- Grid activities in 2009
- Grid Resource plan for 2010
- Computing system for local experiments
### Growth of Grid Fabric

<table>
<thead>
<tr>
<th>CPU Cores</th>
<th>Storage Capacity</th>
<th>Install&amp;Conf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1100</td>
<td>200TB DPM</td>
<td>200TB d-Cache</td>
</tr>
<tr>
<td></td>
<td>Quattor</td>
<td></td>
</tr>
</tbody>
</table>

3/15/10  The Beijing Tier-2 Site  3/29
Network Status

- **TEIN3** Link to Europe: 1Gbps
  - Timeout <170ms
- **GLORIAD** Link to America: 622Mbps
- Data I/O per day: ~3TB
Monitoring System--DIGMON
The Reliability of the site is from 98%-100% through the whole year.
Improvement of data analysis ability through using FroNTier/Squid:

<table>
<thead>
<tr>
<th>Site</th>
<th>BEIJING</th>
<th>IRFU</th>
<th>LAL</th>
<th>LPNHE</th>
<th>LAPP</th>
<th>TOKYO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>16</td>
<td>1.3</td>
<td>1.4</td>
<td>1.5</td>
<td>1.5</td>
<td>13</td>
</tr>
</tbody>
</table>

3/15/10

The Beijing Tier-2 Site
Job Management on different Platforms

- **Supported backend:**
  - PBS, gLite, GOS

- **User interface:**
  - Command Line
  - Web Portal

- **Finished:**
  - MC & Rec Job split
  - Bulk Job submit
  - Job Accounting
Job Management on different Platforms

- Provide two user interfaces
  - Users who have afs account can use them
Outline

- Grid activities in 2009
- Grid Resource Plan for 2010
- Computing system for local experiments
## Resource Plan

<table>
<thead>
<tr>
<th></th>
<th>China, IHEP, Beijing</th>
<th>2009</th>
<th>2010</th>
<th>Split 2010</th>
<th>ALICE</th>
<th>ATLAS</th>
<th>CMS</th>
<th>LHCb</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (HEP-SPEC06)</td>
<td></td>
<td>5600</td>
<td>8000</td>
<td>Offered</td>
<td>4000</td>
<td>4000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>% of Total</td>
<td>50%</td>
<td>50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk (Tbytes)</td>
<td></td>
<td>400</td>
<td>600</td>
<td>Offered</td>
<td>300</td>
<td>300</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>% of Total</td>
<td>50%</td>
<td>50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nominal WAN (Mbits/sec)</td>
<td></td>
<td>1000</td>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Grid activities in 2009
- Grid Resource Plan for 2010
- Computing system for local experiments
Computing cluster for local experiments

- Support experiment: BES, YBJ, DayaBay neutrino...
- Operating System: SLC 4.5
- Computing resource management
  - Resource Manager: Torque
  - Job Scheduler: Maui
  - Monitoring: Ganglia
- Automated installation & configuration: Quattor
- Storage management
  - Home dir.: openAFS
  - Data dir.: Lustre, NFS
  - Mass storage system: Customized CASTOR 1.7
Status of Job Management

- Computing Resource
  - CPU core: 4044
  - Job queue: 23

- Features
  - Bulk Job Submit for MC and Rec Job
  - Job error detection and resubmit
  - Tools for bulk data copy
  - Integrated with dataset bookkeeping
  - Job accounting and statistic interface
Job Accounting

IHEP Cluster Running Statistic: 225 Completed Job(s).

2010-03-01 — 2010-03-01

<table>
<thead>
<tr>
<th>JobID</th>
<th>Queue</th>
<th>User</th>
<th>Group</th>
<th>JobName</th>
<th>WallTime (h)</th>
<th>CPUTime (h)</th>
<th>Efficiency</th>
<th>ExecHost</th>
<th>SubmitHost</th>
<th>Real Mem (Mb)</th>
<th>Virtual Mem (Mb)</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>198776</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002050</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198878</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002060</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198879</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002070</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198880</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002080</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198881</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002090</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198882</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002100</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198883</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002110</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198884</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002120</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198885</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002130</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198886</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002140</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198887</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002150</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198888</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002160</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198798</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002050</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198799</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002060</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198800</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002070</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198801</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002080</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198802</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002090</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198803</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002100</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198804</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002110</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198805</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002120</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198806</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002130</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198807</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002140</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198808</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002150</td>
<td>1.070</td>
<td>1.500</td>
<td>0.836</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>382.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
<tr>
<td>198809</td>
<td>torquev</td>
<td>lijiang</td>
<td>physics</td>
<td>Data1002160</td>
<td>2.075</td>
<td>2.680</td>
<td>0.936</td>
<td>pvs0012</td>
<td>pvs0012</td>
<td>383.640</td>
<td>120.640</td>
<td>2010-01-26</td>
<td>01:37:26</td>
</tr>
</tbody>
</table>
Cluster Statistic

Jobs/Queue

Job Success Rates

CPU Efficiency
Storage Architecture

Computing nodes

Storage system

Hardware

- HSM (CASTOR)
- File systems (Lustre, NFS)

- Name Server
- Disk pool
- Tape pool

- MDS
- OSS
- OSS

Hardware:
- 10G
- 1G
CASTOR Deployment

- **Hardware**
  - 2 IBM 3584 tape libraries
  - ~5350 slots, extend > 4PB tape capacity
  - 20 tape drivers (4 LTO3, 16 LTO4)
  - ~2400 tapes (2000 of them are LTO4)
  - >800TB of data is stored in tapes for the moment
  - 10 tape servers and 8 disk servers with 120TB disk pool

- **Software**
  - Modified version based on CASTOR 1.7.1.5
  - Support the new types of hardware, such as LTO4 tape
  - Optimize the performance of tape read and write operation
  - Reduce the database limitation of stager in CASTOR 1
Performance Optimizing

- **Write**
  - Raise the data migration threshold to improve writing efficiency, > 100GB
  - Increase size of data file, 2GB for raw data, 5GB for rec. data
  - Store one data set on to more than one tape in order to stage in parallel later

- **Read**
  - Read tape files in bulk, and sort them in ascending order
  - Copy data from CASTOR to the LUSTRE file system directly and skip the disk servers in CASTOR
  - Stage in files from different tapes in parallel
  - Setup dedicated batch system for data migration. Distributed the data copy task to several nodes for higher aggregated speed

- **Result**
  - Write: 330MB/sec for 8 tape drivers
  - Read: 342MB/sec for 8 tape drivers, 40MB+/driver/sec
Performance of the Castor System

8 tape driver: >700MB/sec
Deployment of Lustre File System

Version: 1.8.1.1
I/O servers: 10
Storage Capacity: 326 TB

Diagram showing the structure of the Lustre File System with various components and connections, including MDS (Main and sub), OSS 1 and OSS N, SATA Disk Array RAID 6 (Main and extended), and 10Gb Ethernet connections.
Performance of the Lustre File System

- Throughput of Data analysis: $\sim 4GB/s$
- WIO% on computing nodes: $< 10\%$
- We added 350TB storage space, 10 I/O servers to the system a few weeks ago, the throughput is estimated to be $\sim 8GB/s$!
Real time Monitoring of Castor

- Based on Adobe Flex 3 and Castor 1.7 API
- Shows the system real time status with animation, color, and user friendly graphics
- Integrated Information from Ganglia

![Diagram of real time monitoring system]

- Web Browser
  - Action Script, Flex, Cairngorm Events
- Cairngorm data Model
  - Map
  - HTTP Protocol
  - Java data Model
  - Socket
- Cmonitor

Adobe LiveCycle Data Service On Tomcat
Real time Monitoring of Castor
The File Reservation component is a add-on component for Castor 1.7. It is developed to prevent the reserved files from migrating to tape when disk usage is over certain level.

The component provides a command line Interface and a web Interface. Through these two Interfaces, data administrators can:

- Browse mass storage name space with a directory tree
- Make file-based, dataset-based and tape-based reservation
- Browse, modify and delete reservation.

According to test results, current system is stable under 400 to 500 users concurrent access.
### File Reservation for Castor

#### Castor 文件预留系统

**Loading data...**
- castor
  - hep.ac.cn
    - bes
      - bes2
    - raw
    - test
    - offline
    - run
    - userTest
  - public
  - user
  - ybj

**你当前的位置：/castor/hep.ac.cn/bes/best/raw 文件列表**

**当前列表的位置：/castor/hep.ac.cn/bes/best/raw 文件列表**

<table>
<thead>
<tr>
<th>名称</th>
<th>类型</th>
<th>文件名</th>
<th>操作数</th>
<th>文件大小</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rw</td>
<td>test001.rw</td>
<td>0</td>
<td>100MB</td>
</tr>
<tr>
<td>2</td>
<td>rw</td>
<td>test002.rw</td>
<td>0</td>
<td>200MB</td>
</tr>
<tr>
<td>3</td>
<td>rw</td>
<td>test003.rw</td>
<td>0</td>
<td>300MB</td>
</tr>
<tr>
<td>4</td>
<td>rw</td>
<td>test004.rw</td>
<td>0</td>
<td>400MB</td>
</tr>
<tr>
<td>5</td>
<td>rw</td>
<td>test005.rw</td>
<td>0</td>
<td>500MB</td>
</tr>
<tr>
<td>6</td>
<td>rw</td>
<td>test006.rw</td>
<td>0</td>
<td>600MB</td>
</tr>
<tr>
<td>7</td>
<td>rw</td>
<td>test007.rw</td>
<td>0</td>
<td>700MB</td>
</tr>
<tr>
<td>8</td>
<td>rw</td>
<td>test008.rw</td>
<td>0</td>
<td>800MB</td>
</tr>
</tbody>
</table>

**需要删除的文件列表如下：**

<table>
<thead>
<tr>
<th>名称</th>
<th>类型</th>
<th>文件名</th>
<th>操作数</th>
<th>文件大小</th>
</tr>
</thead>
<tbody>
<tr>
<td>/castor/hep.ac.cn/bes/best/raw/test001.rw</td>
<td>10</td>
<td>100MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/castor/hep.ac.cn/bes/best/raw/test002.rw</td>
<td>10</td>
<td>200MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/castor/hep.ac.cn/bes/best/raw/test003.rw</td>
<td>10</td>
<td>300MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/castor/hep.ac.cn/bes/best/raw/test004.rw</td>
<td>10</td>
<td>400MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/castor/hep.ac.cn/bes/best/raw/test005.rw</td>
<td>10</td>
<td>500MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/castor/hep.ac.cn/bes/best/raw/test006.rw</td>
<td>10</td>
<td>600MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/castor/hep.ac.cn/bes/best/raw/test007.rw</td>
<td>10</td>
<td>700MB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/castor/hep.ac.cn/bes/best/raw/test008.rw</td>
<td>10</td>
<td>800MB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summery

- The Beijing Tier-2 Site
  - Resource and plan
  - Reliability and Efficiency
  - Monitoring and cooperating tools

- Computing System for local experiments
  - Job Management
    - Features, accounting, statistics
  - Customized Castor 1.7 as HSM
    - Performance optimization and result
  - Distributed disk storage using Lustre
    - Deployment and current scale
  - Realtime monitoring for Castor
    - Animation based on Adobe Flex
  - File reservation for Castor
Thank you!

Lu.Wang@ihep.ac.cn