Overview of the Lustre File System at IHEP

Lu Wang
Lu.Wang@ihep.ac.cn
Computing Center, Institute of High Energy Physics,
Chinese Academy of Sciences

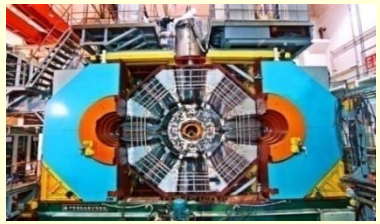
Outline

- Introduction of HEP computing in IHEP
- Lustre deployment, performance and management
- Expected Lustre features from IHEP
- News of China Lustre workshop 2011

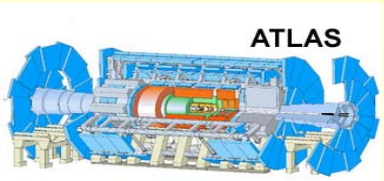
IHEP at A Glance

- The largest(~1000 stuffs) fundamental research center in China with research fields includes:
  - Particle Physics Experiments
  - Cosmic Ray/Astrophysics experiments
  - Theoretical Physics
  - …
- Scientific projects:
  - BESIII experiment running on BEPC
  - ARGO-YBJ experiment
  - Daya Bay reactor neutrino experiment
  - ATLAS, CMS experiment on LHC
  - AMS, HMXT …

Beijing Spectrum



ARGO-YBJ Detector



Computing Center at IHEP

More than 7000 CPU cores

Outside data farms:
LHC, YBJ,DAYABAY

WAN

Login, monitoring,
scheduling,AFS,backup…

10Gbit Ethernet

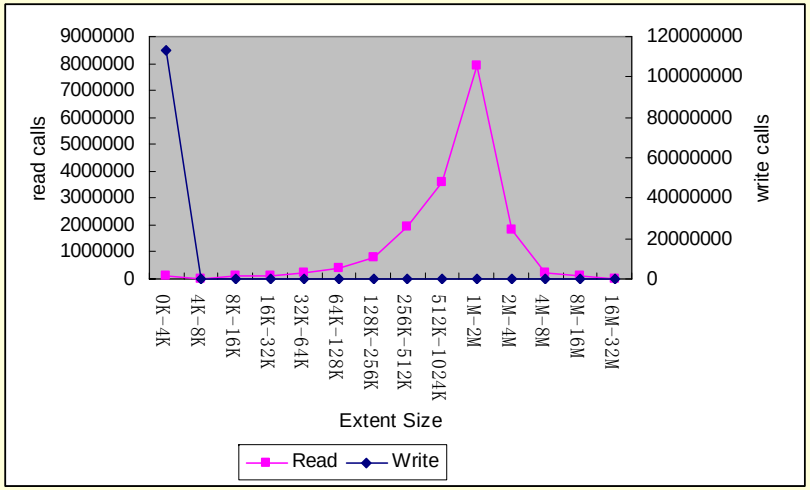
Nearly 2 PB disk storage (Lustre)

Local data
farm

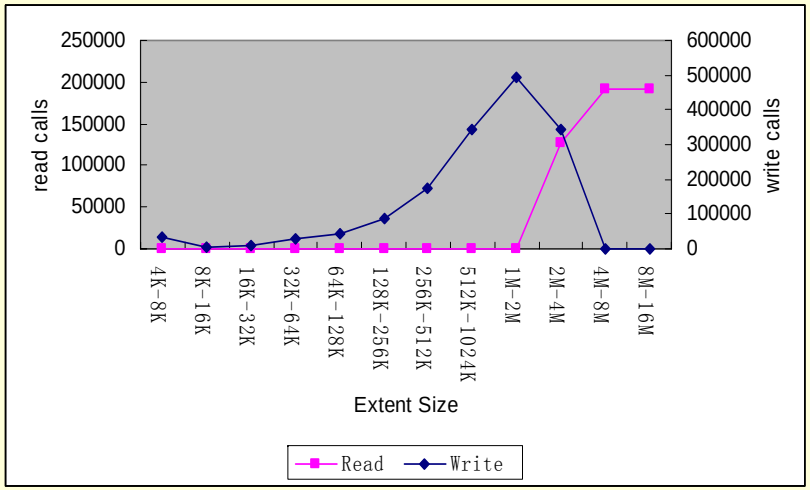
5PB tape storage

Requirements of HEP Computing

- Most HEP(high energy physics) computing are *data intensive, high throughput* computing
  - Read per job: [0.5,6] MB/s
  - Write per job: 0.1 MB/s
  - Big files (>1 GB), Write once, Read Many
  - I/O extent size



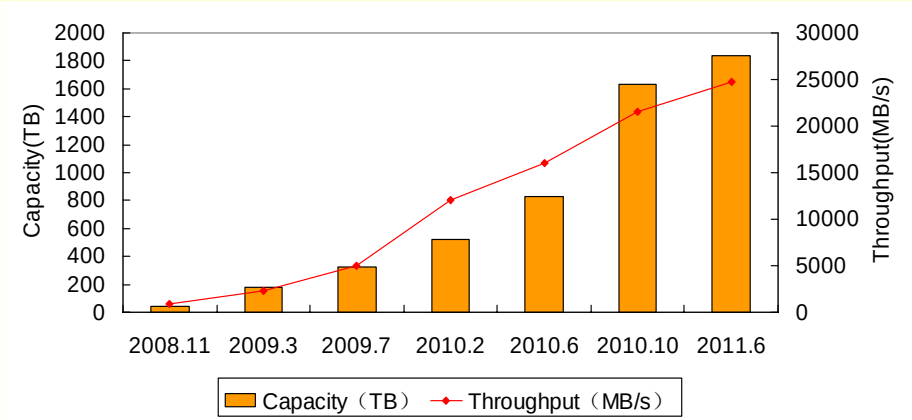
Distribution of I/O extent size for analysis jobs



Distribution of I/O extent size for reconstruction jobs

Brief History of Lustre at IHEP

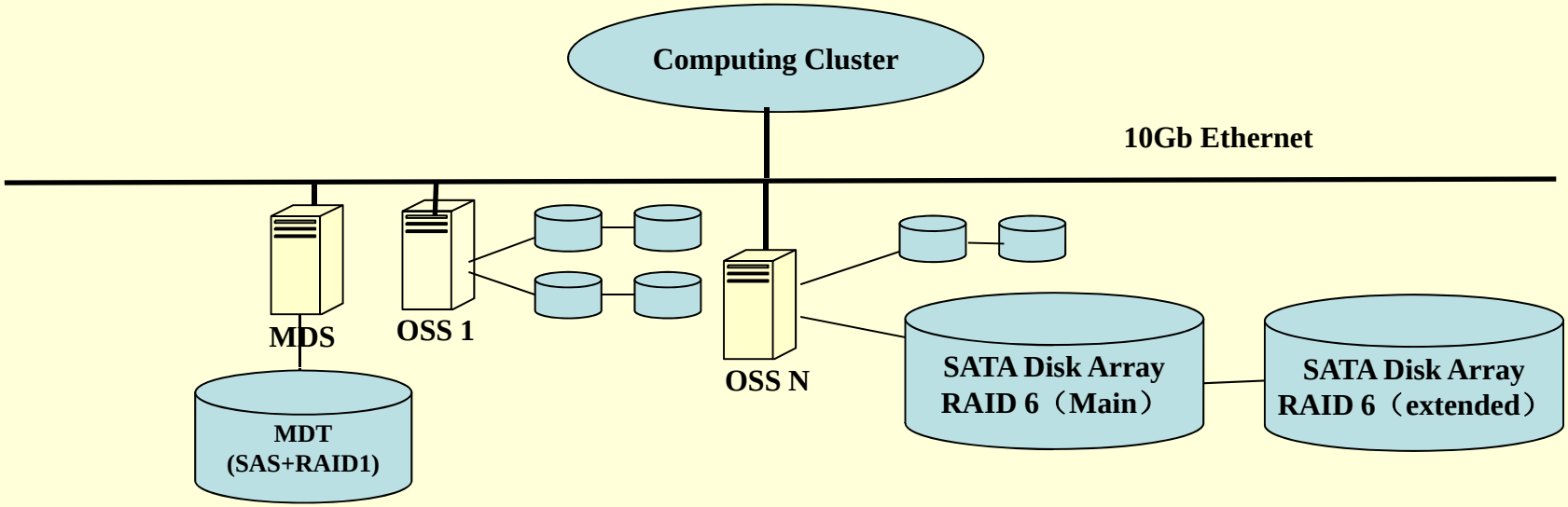
- 2008.8 A Lustre file system (1.6.5) was firstly deployed
- 2010.1 upgraded to *1.8.1.1*, 2011.7 upgraded to *1.8.5*



- Current:
  - ***3*** *MDSs*, ***31*** OSSs, *300+* OSTs, *800* client nodes, *100* million files
- At the end of 2011:
  - *+ 7 OSSs*，*+84 OSTs*, *+ 500 TB*, *+125 clients*

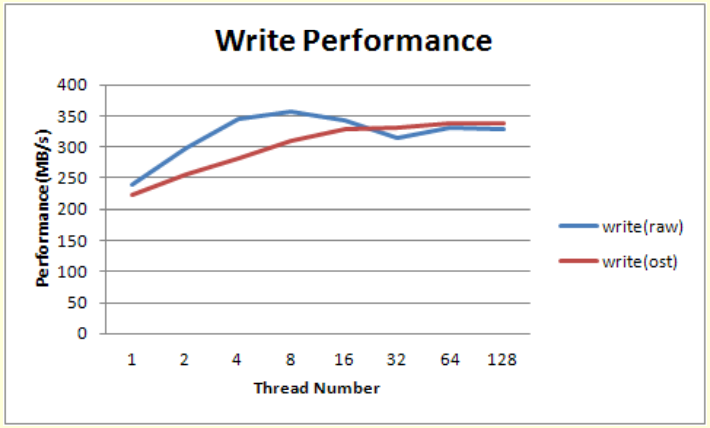
Current Lustre Deployment

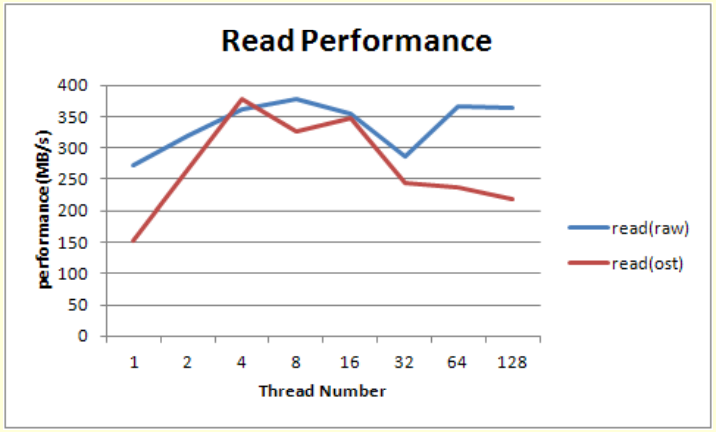
- MDS attached with 1 disk array
  - RAID 10,SAS
- OSS attached with 4 disk arrays through two 4 Gb HBA
  - RAID 6+1, SATA,1TB/2TB



Performance Test

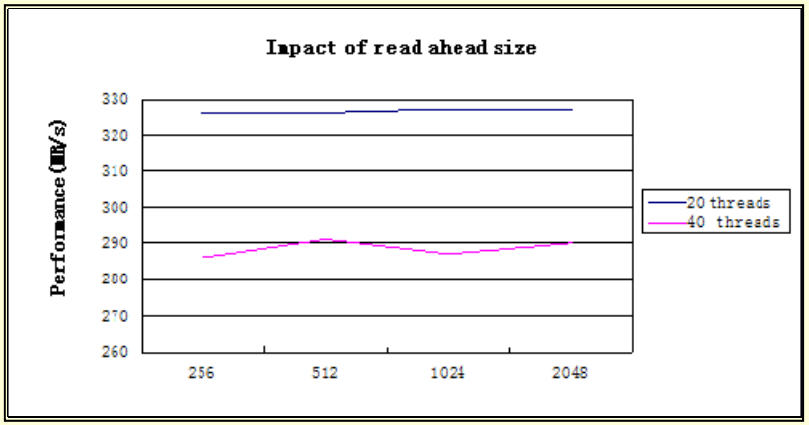
- Performance of raw device
  - Raid 6+1, 12, 1TB, SATA,write through
  - with *spgdd-survey*, multi-thread, each thread R/W single region
  - Write [239-357]MB/s, Read [273-377]MB/s
- Performance of each OST
  - With *obdfilter-survey*,multi-thread, each thread R/W single object
  - Write[223-337] MB/s,Read [219-379]MB/s
  - When thread Number >32, read performance lost more

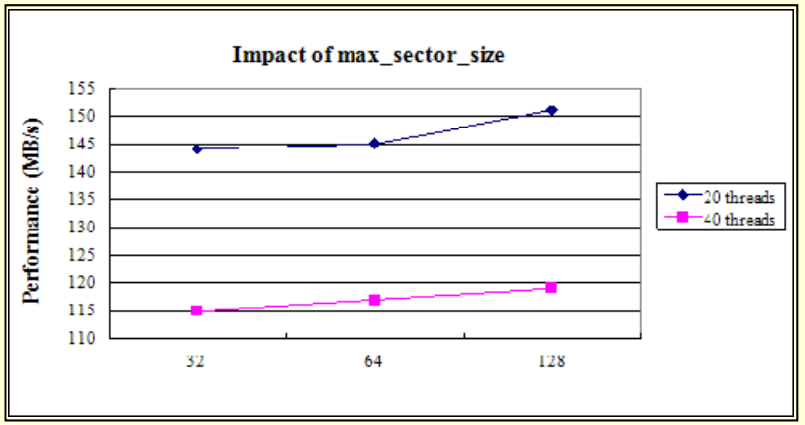


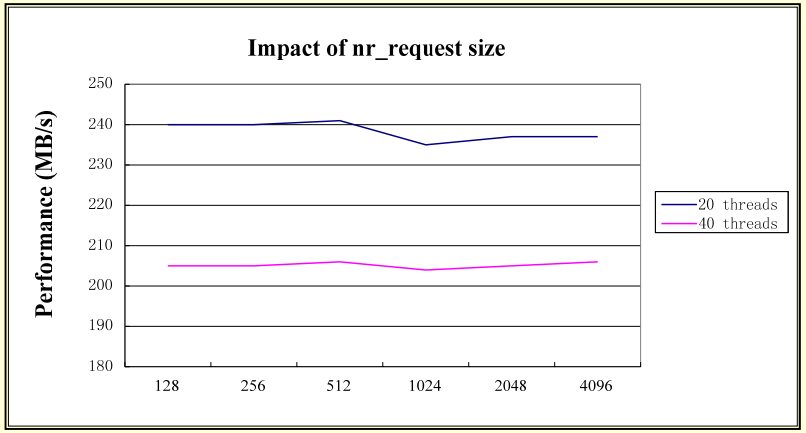


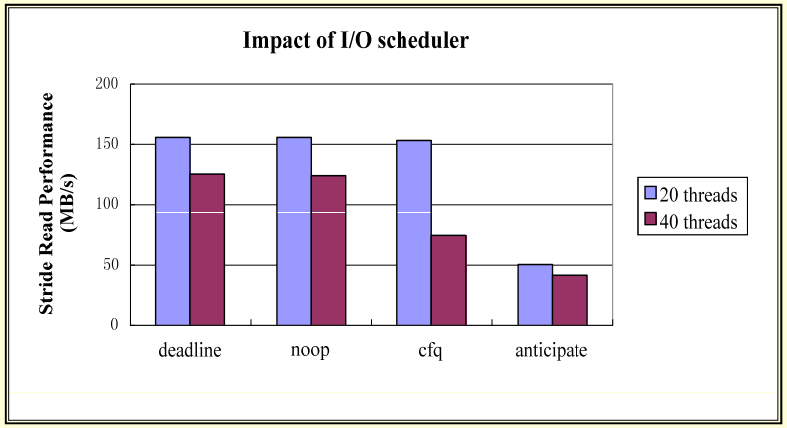
Performance Test

- Multi-thread stride read performance,stride=2MB



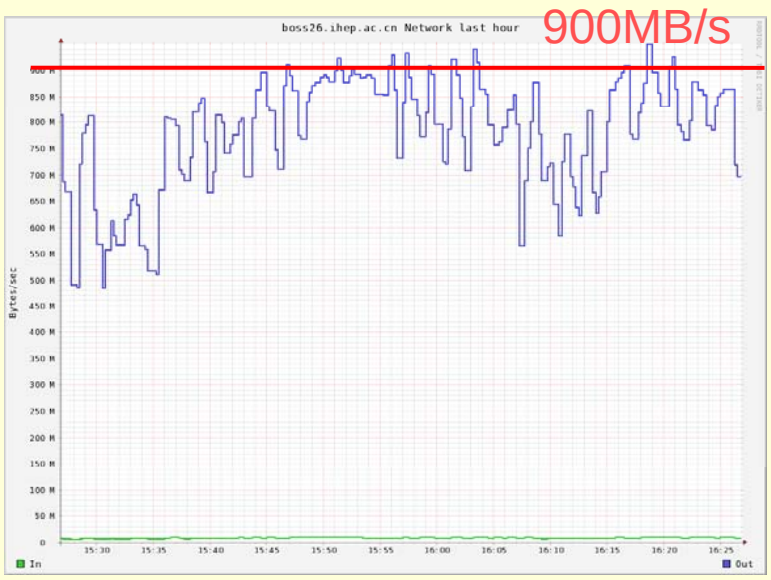




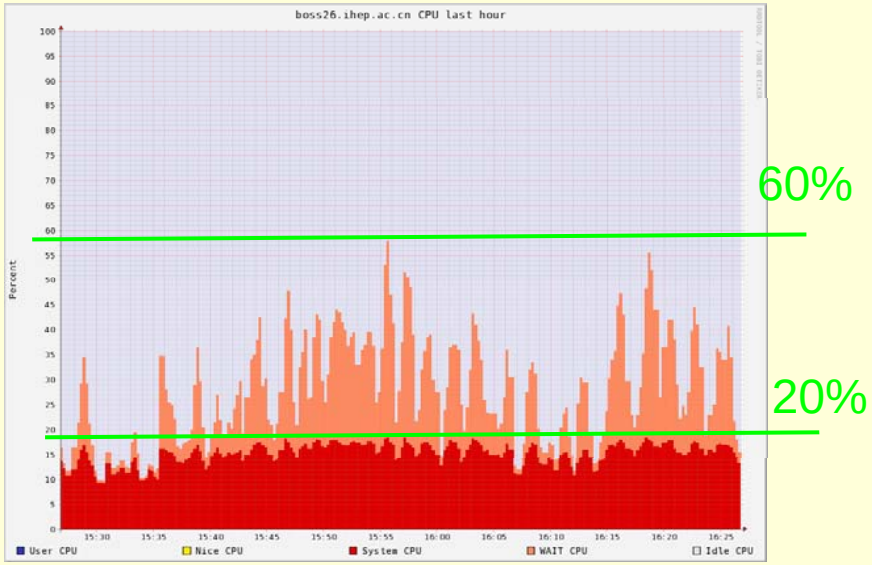


Real Performance

- Performance for one OSS with 4 disk arrays, 8 OSTs
  - each disk array serving ~50 client processes
  - Read performance, 225 MB/s/disk array,similar with test preference
  - The peak read performance of the whole system is expected to be 24 GB/s



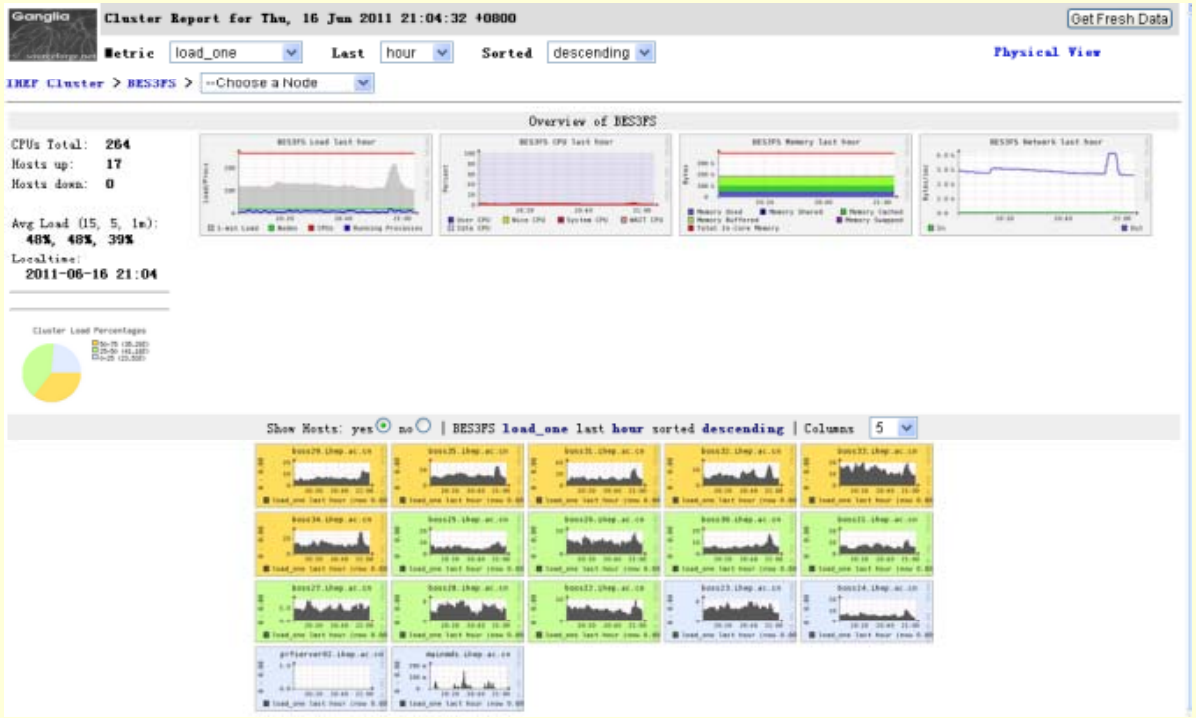
Read throughput of single OSS



IOwait on single OSS

Management: Monitor

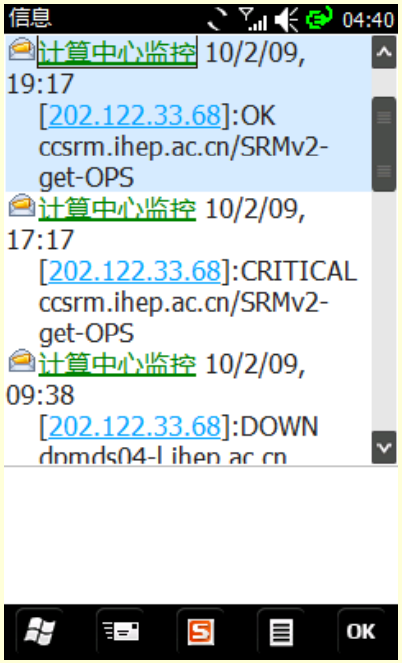
- OST/MDT : ***llstat, llobdstat, iostat***
- Common OS Metric: ***Ganglia***



- Client: ***rpc_stats, extent_stats,offset_stats***

Management :Problem Detection

- Problem detection integrated with Nagios
  - Nagios plugin on all clients report availability of file system
  - Probe script
  - Error alarms sent to administrators through Web, E-mail and Mobile



Management: Problem Diagnostic

- Tools
  - lctl debug_kernel, lctl debug_file..
  - crash dump
  - debugfs
  - /var/log/message integrated with Syslog-ng
- Reference
  - Lustre Manual
  - Discussion List

Problems and Solutions

Problem	solution
Frequent crashes of 32 bit OSS	Change all OSS to 64 bit
Instability of clients with 2 NICs	specify the NIC used by Lustre client with *options lnet networks=tcpx(ethx)*
Instability caused by low default timeout (100s)on Lustre 1.6.X	1. Specify timeout through lctl set_param 2. Update to 1.8.x, which support AT
1.User synchronization with LDAP DB 2.Quota on directories 3.parralel data migration between Disk and Tape	Home made scripts
Memory usage control of OSS	turn off *read_ahead_cache_enable* *Writebehind_enable* on OSS

Problems and Solutions

Problem	solution
After the expansion of Lustre, the 32bit Clients face serious low memory shortage, crashes frequently	have not find solution through Mail List
communication with a dead client will stuck one CPU core on OSS(1.8.1.1)	After upgrade to 1.8.5, the problem have not happend again
certain ptlrpcd-recov,ldlm process process take 100% CPU, unable to be killed	After upgrade to 1.8.5, the problem have not happned again

Expected Features of Lustre

1. OST auto balance
   - After new devices added in, it is difficult to balance data on application level
   - Writing data more frequently on new devices will cause unbalance of future read
2. File level replication
3. Scalability of MDS
4. HSM
   - combine Lustre with Tape storage under managed by CASTOR*
5. Better memory usage control
6. On line backup of metadata
   - LVM or DRBD may caused performance penalty
   - it is not possible to make offline backup everyday

* a HSM software developed by CERN

China Lustre Workshop 2011

- It is the first Lustre Workshop in China.
- Co-organized by the Institute of High Energy Physics and Whamcloud, Inc.
- Aims at providing an communication channel for Chinese Lustre users and developers
- Received *44* registrations from different industries in 4 weeks!
- Hope to build a Chinese users' community to:
  - exchange experience
  - forge and sustain a bond with the development team
  - feed back problems and collect common requirements
- More details:
  - http://lustrechina.csp.escience.cn/dct/page/65595

Thank you!