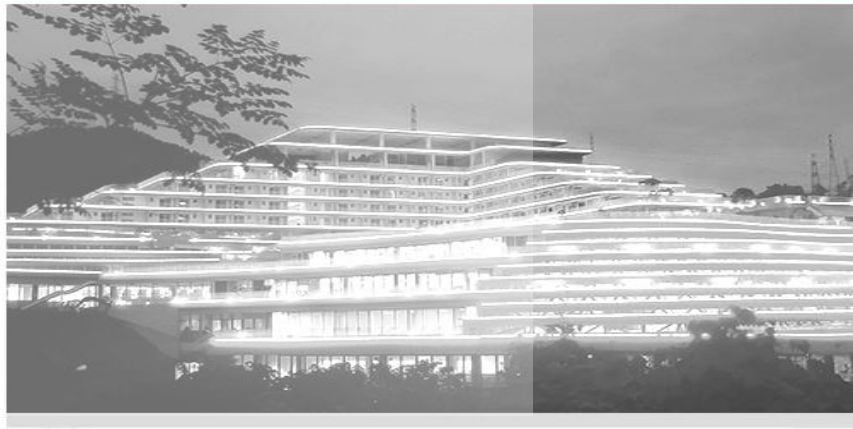
Lustre at China National GeneBank

Liping Yang氏



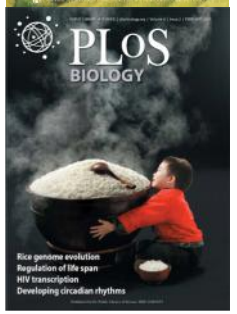
China National GeneBank 国家基因库

Part One

Introduction to China National GeneBank (CNGB)

China National Genebank (CNGB)

- Non-Profit Institute
- Funding by:
  - China NDRC (National Development and Reform Commission)
  - Shenzhen local government
  - BGI self-funding
- Operated by BGI
- Three Banks & Two Platforms



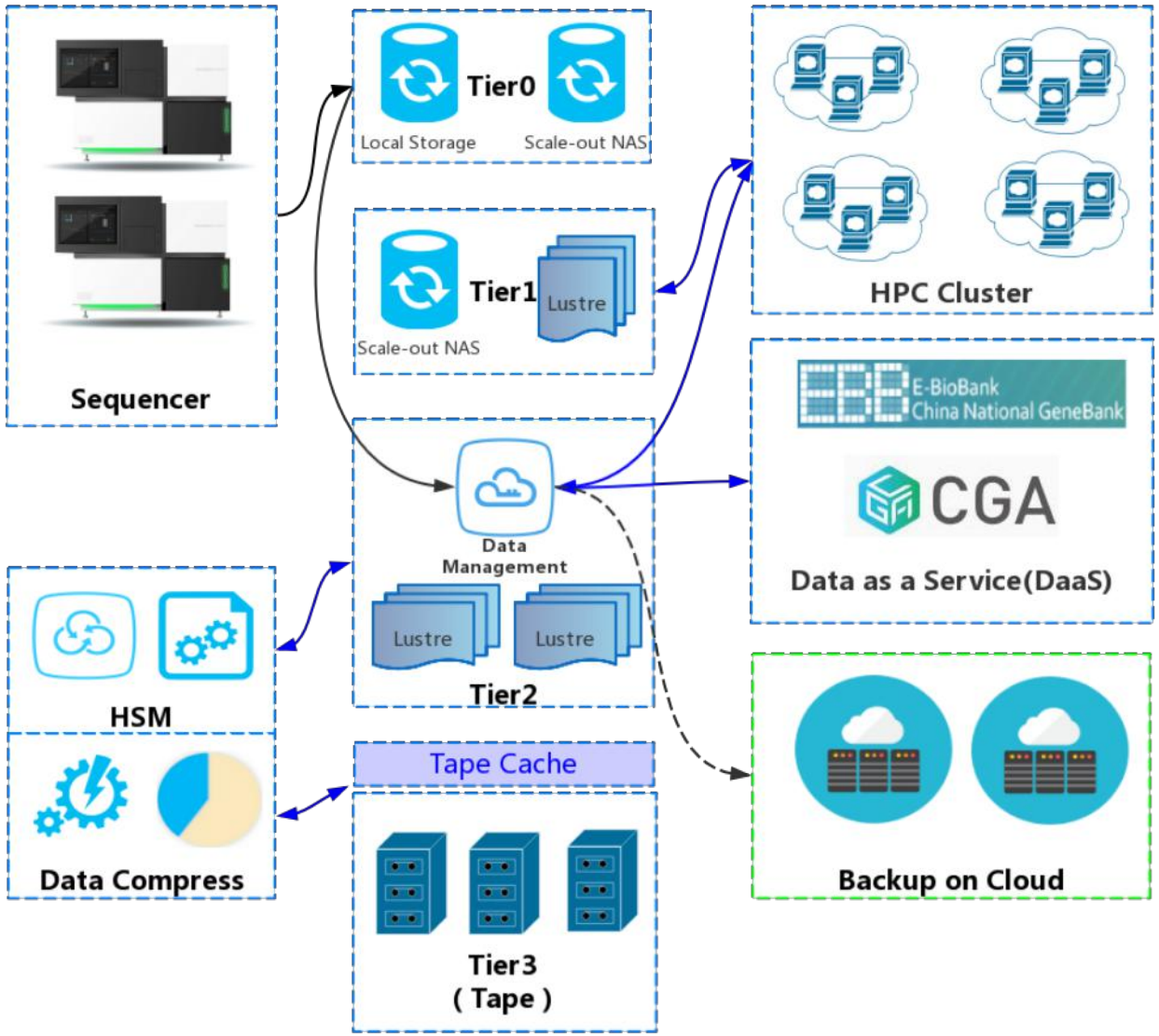
CNGB

Read
Write
3 Banks & 2 Platforms
Living Biobank
Bio-informatics Data Center
Biorepository

CNGB Bio-Informatics Data Center

- 65,072 CPU Cores
- 258.36 TB Memory
- 119.41 PB Storage
- 917.23 T Flops

Data Flow in CNGB



Part Two

The Road to Large Scale Lustre File System

First Attempt

- Configuration:
  - Sun Fire x4500 (500GB hard drive * 48) * 8
  - Soft RAID. No HA.
  - Lustre 1.6
- Problems:
  - The reseller failed to deploy the system.
  - Lustre engineer from Sun help us deploy it, and the benchmark result is pretty good.
- Resolution:
  - We break the system into NAS
- Lessons learned:
  - Proof of concept is not optional.

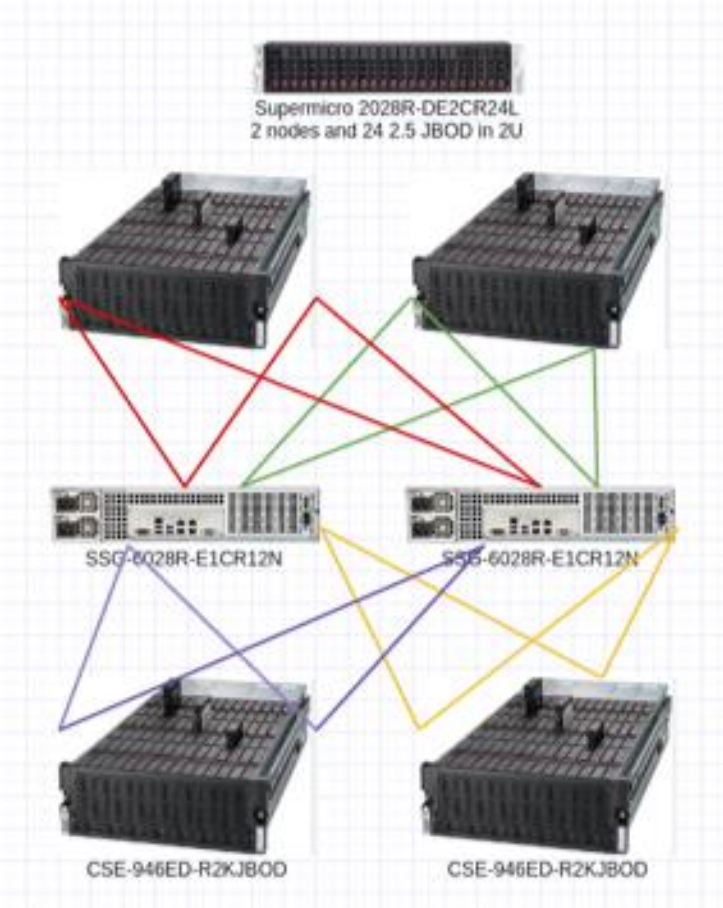
Source: https://www.youtube.com/watch?v=lwT3Hrk4BSo

Looking for solutions

- We buy commercial parallel storage systems like Isilon and Panasas.
- We tried many commercial and open source parallel file systems like ParaStor, BWFS（Blue Whale File System），glusterfs and Lustre.
- We use Glusterfs and Lustre is our testing environment

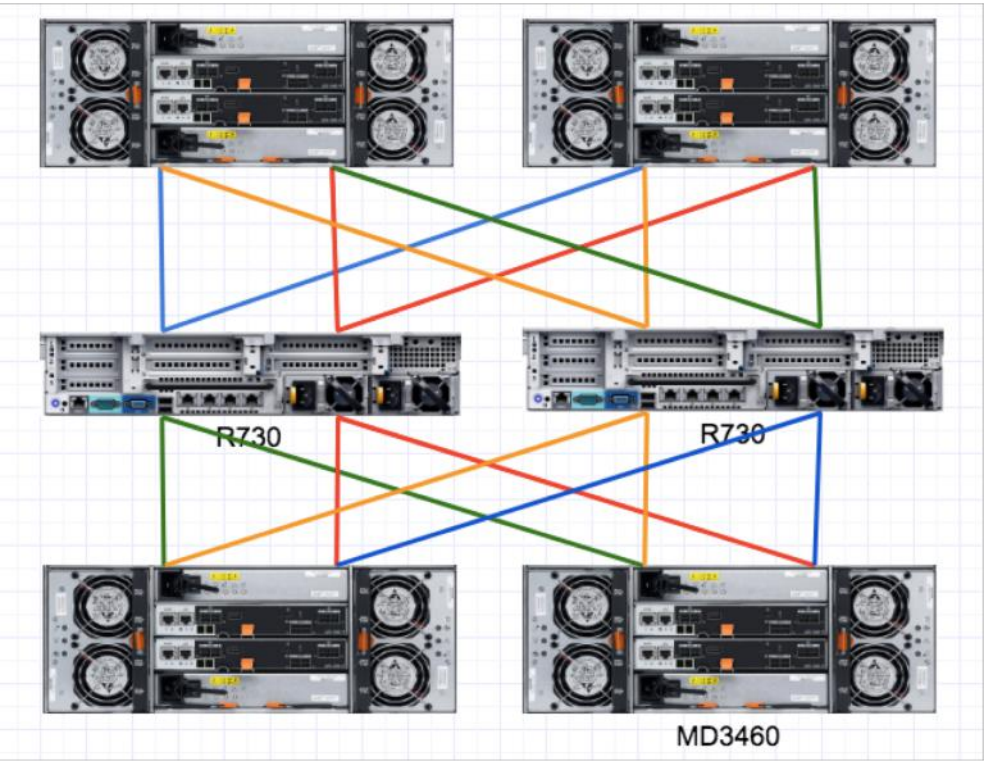
Lustre as Archive Storage

- Background:
  - Two copies. One of them in Lustre File System.
- Configuration:
  - Supermicro 4U 90Bay JBOD + 1U OSS Server
  - IEEL 2.x ~ IEEL 3.x
- Results:
  - No vendor lock-in
  - Stronger negotiating position
  - High density space saving
  - Gain Lustre production system experience



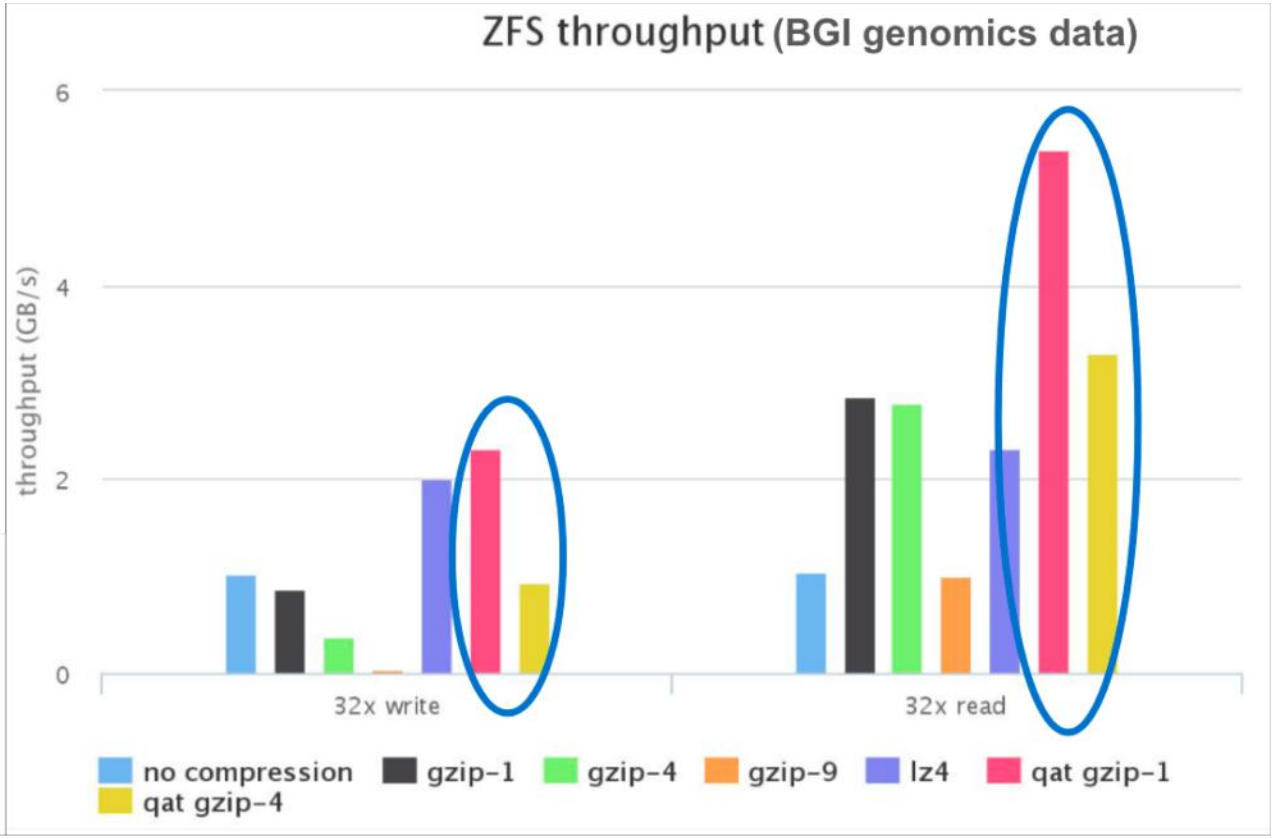
Lustre for gene data analysis

- Configuration:
  - Dell MD3460 + R630、Supermicro 4U 90Bay JBOD + 1U OSS Server
  - IEEL 2.x ~ IEEL 3.x
- Compare to other commercial solutions
  - No vendor lock-in
  - 10x faster
  - Half-price



Extended Lustre Functionality

- Ali OSS copytool
- Intel QAT Compression for ZFS on Linux
- Cooperate with DDN to use Quip
- DDN LIPE



LIPE

Lustre Integrated Policy Engine

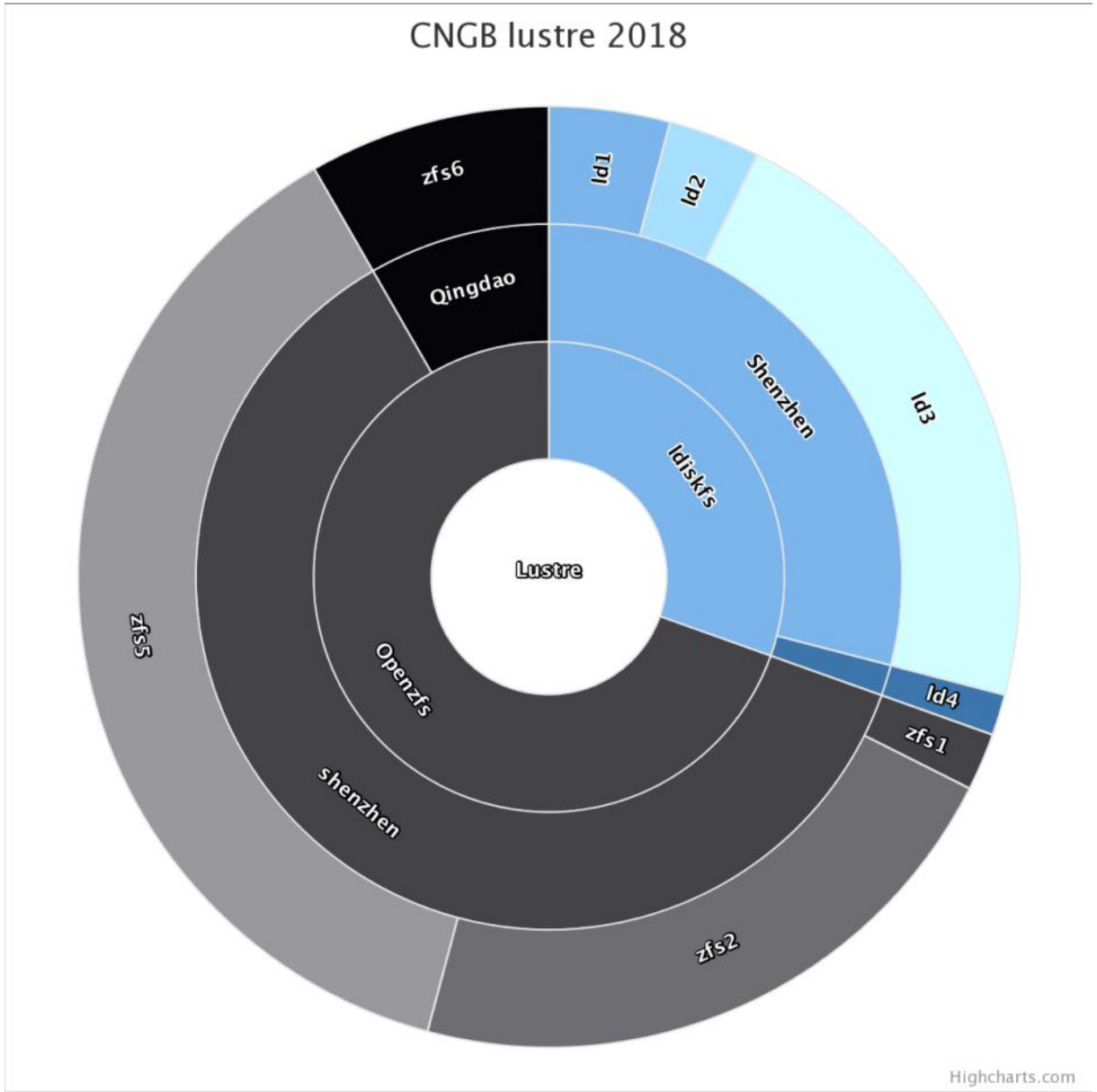
Source: http://wiki.lustre.org

Part Three

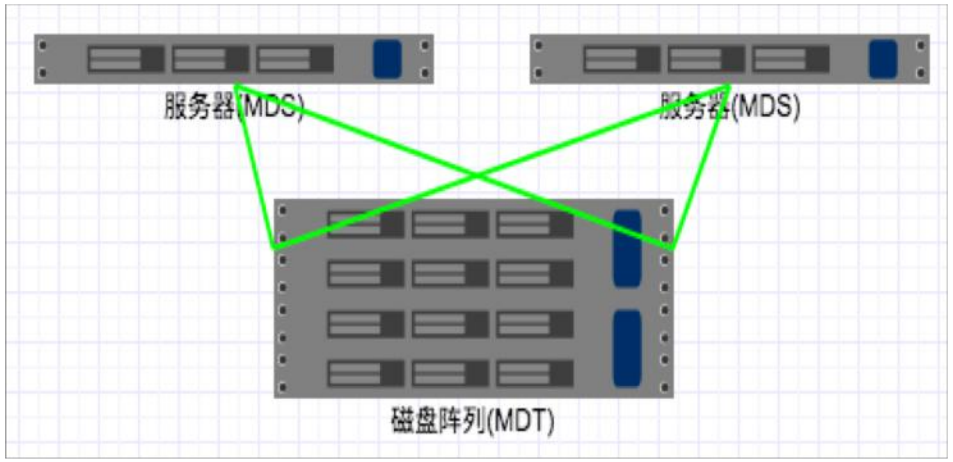
Run and Manage Lustre

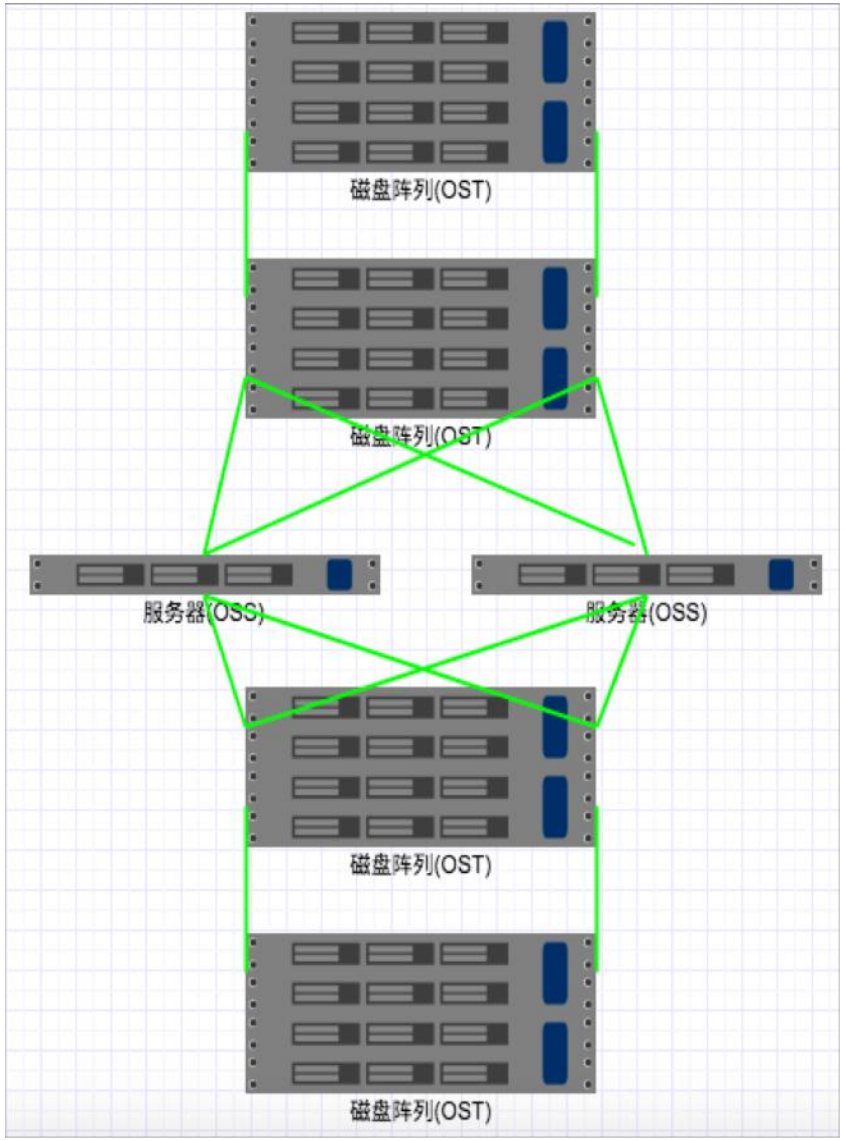
Overview of CNGB Lustre Systems

- Raw capacity: 75.4 PB
  - ldiskfs: 24.1 PB
  - OpenZFS: 51.3 PB
- Largest system: 26 PB
- Peak Performance
  - Single system: 85 GB/s
  - Single OSS: 6 GB/s



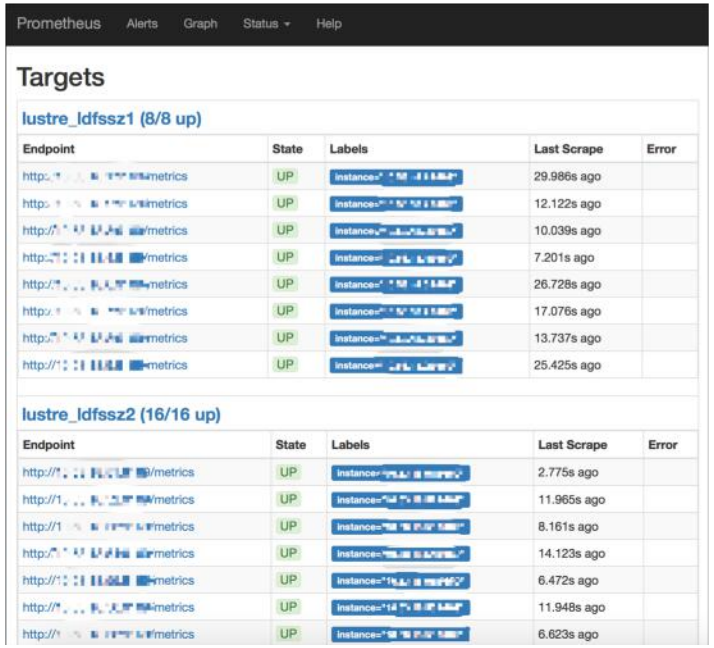
- Storage hardware vendors
  - Supermicro, DELL, HPE, HGST
- Level 3 software support from DDN/Whamcloud
- Manual Failover





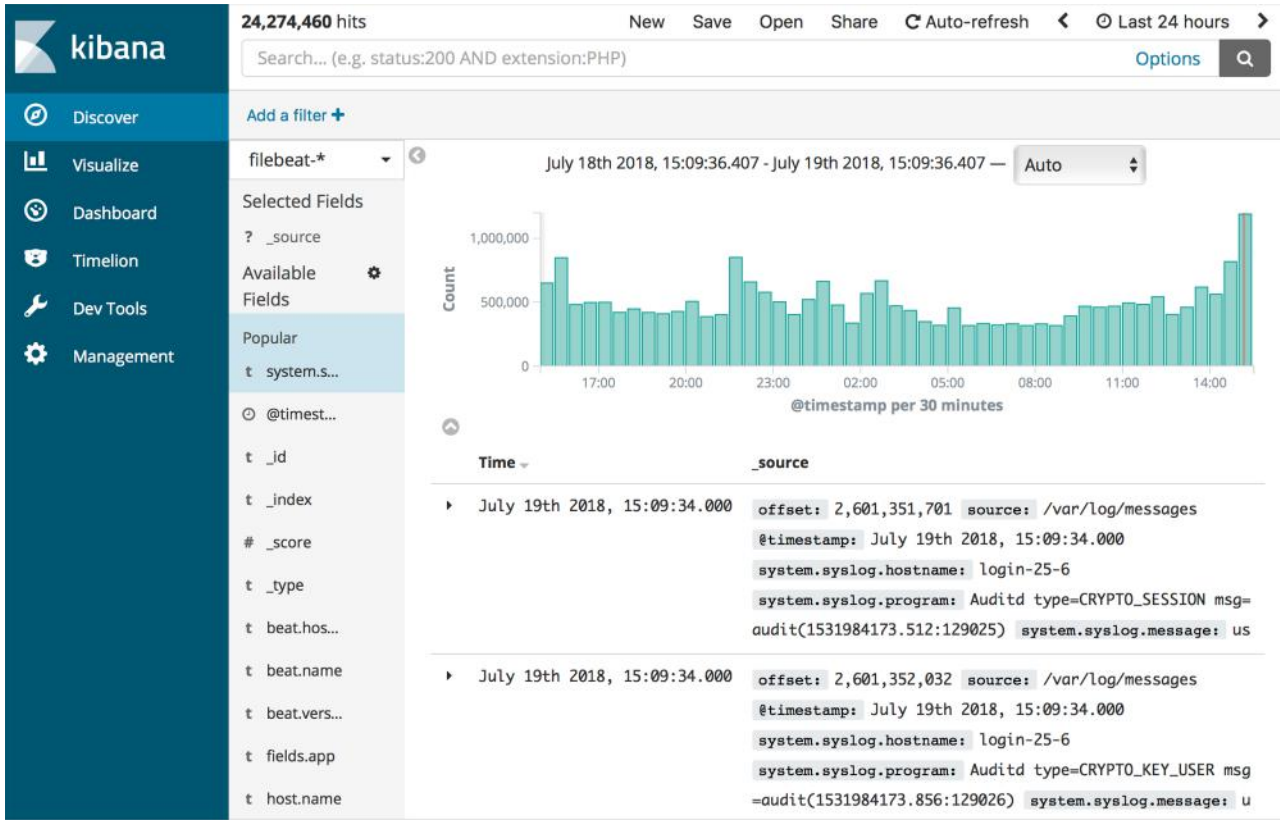
Monitoring

- Nodes metric collection: Prometheus, node_exporter
- Network metric collection: Prometheus, node_exporter and scripts
- Lustre metric: Prometheus and lustre_exporter
- Dashboard: Grafana



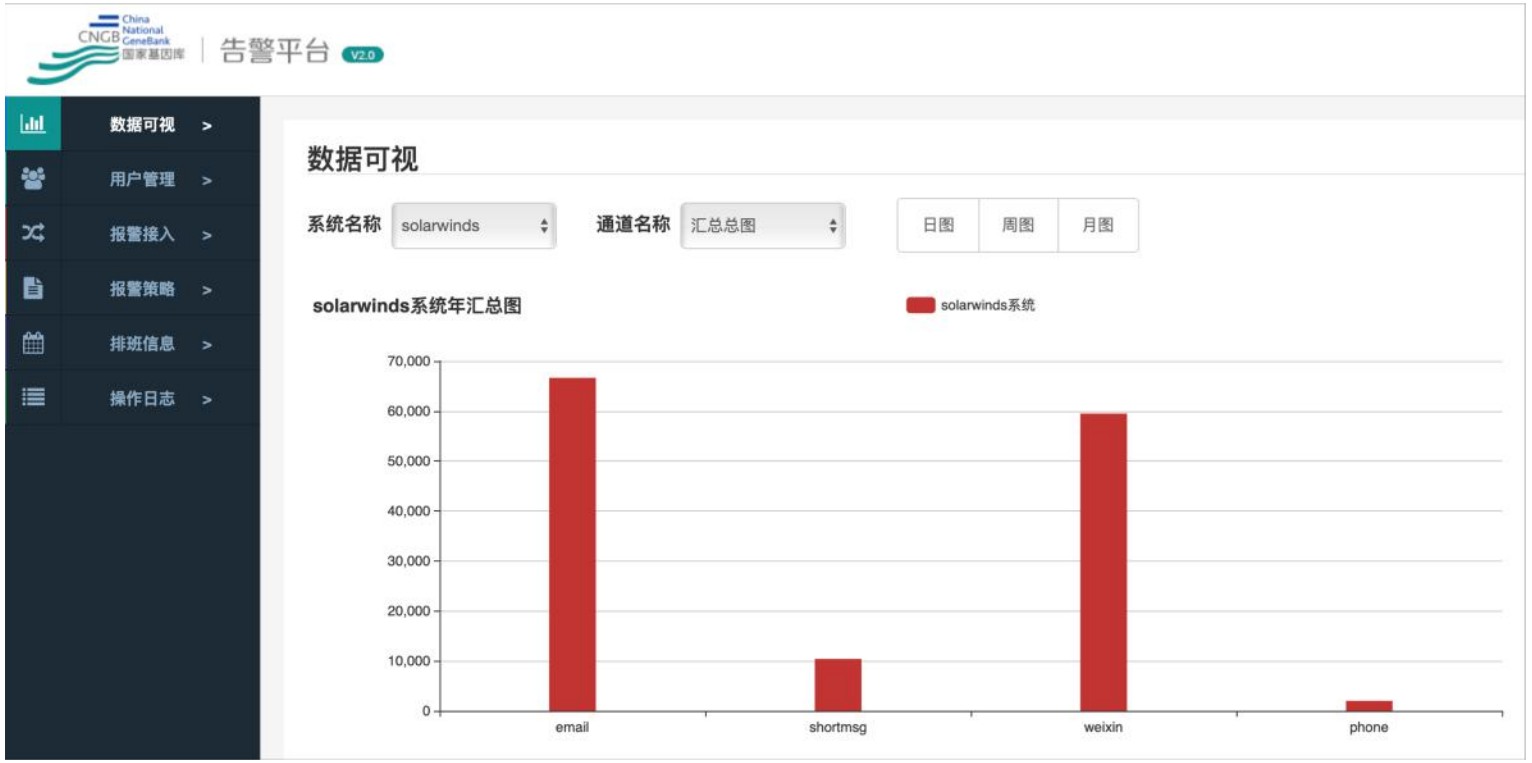
Logs

- Log collection: Filebeat
- Extract, Transform, Load (ETL): Elastic Search Grok Processor
- Storage: Elasticsearch and graphite
- Dashboard: Kibana and Grafana



Notifications

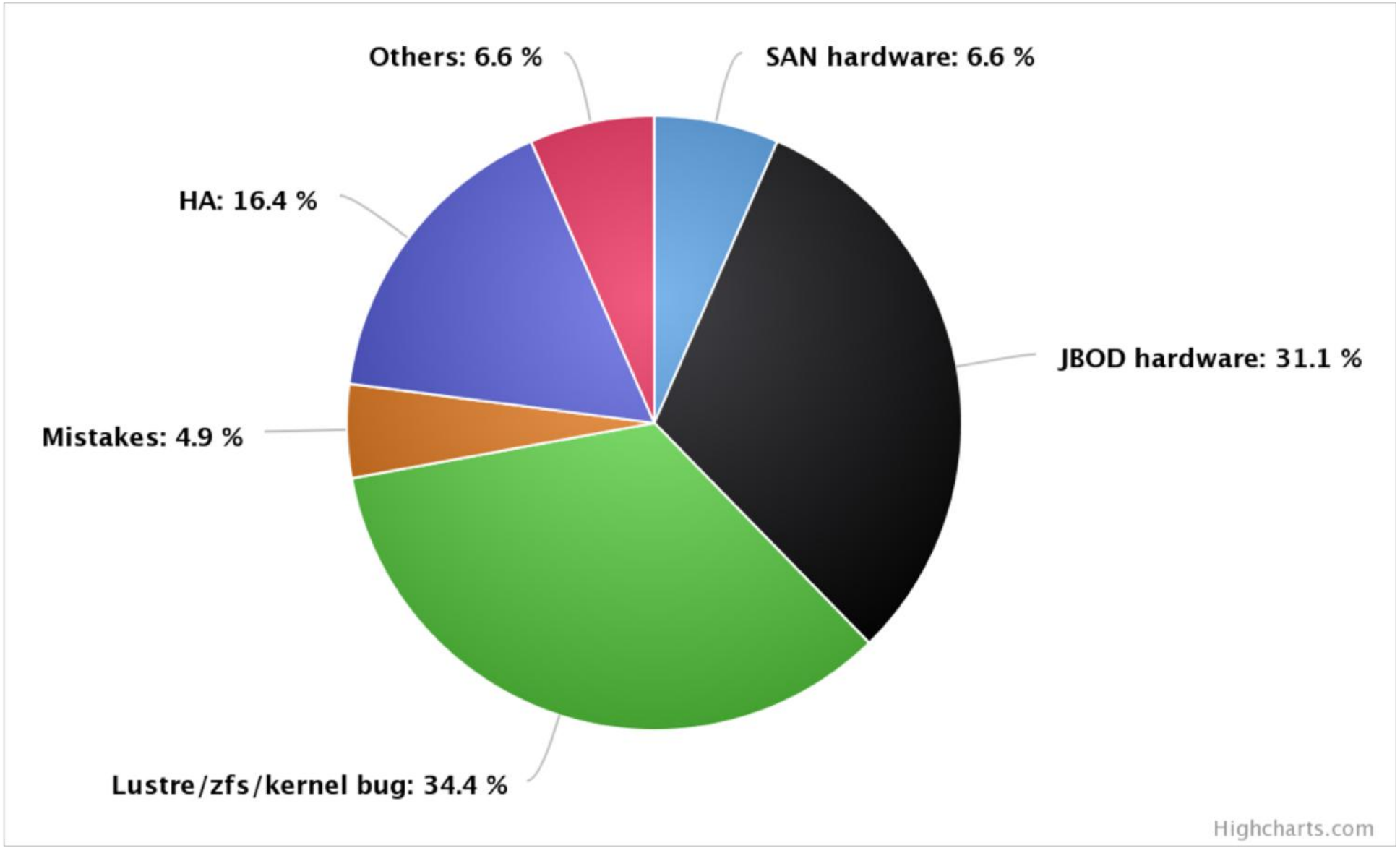
- Nagios
- In-house notification gateway
  - SMS text, WeChat, email and phone call



Part Four

Issues and Lessons learned

CNGB Lustre Failures



Hardware problems

- Problems
  - Firmware (HBA, Hard Drive etc.)
  - Defective modules or cables
  - Network adapter and network switch cooperation issue
- Resolutions
  - POC and benchmarking
  - End-to-end quality control process
  - Working on in-house automated test system

- Pros
  - Open Source
  - Unified RAID administration for different hardware (no 3rd party RAID tools)
  - End-to-end data integrity
  - Scalable, online filesystem check/scrub/repair
  - Lower hardware costs
- Cons
  - Slightly lower metadata performance, especially for MDT
- Considerations
  - zfs 0.7.0 or above for MMP. Keep your eyes on bugfix and new feature in Github.
- Future works
  - QAT hardware-assisted checksums, compression and encryption

- Timeout
- Adaptive Timeouts
  - at_min
  - at_max
- LDLM Timeouts
  - ldlm_enqueue_min
  - ldlm_timeout
- Debug - dump_on_timeout, dump_on_eviction
- ARP issue in our environment

Reference: Lustre Resiliency: Understanding Lustre Message Loss and Tuning for Resiliency

Case study: A post-disaster survival experience

- Background:
  - Tight Deadline
  - POSIX compliant file system storage
  - A fixed budget, with prioritization on storage capacity
- Solution:
  - Intel Enterprise Edition Lustre
  - Intel Lustre reference architecture hardware without POC

Case study: A post-disaster survival experience cont.

- Issue 1 – Network sluggish
  - *LustreError: 51342:0:(import.c:372:ptlrpc_invalidate_import()) zfs4-OST0005_UUID: RPCs in "Unregistering" phase found (1). Network is sluggish? Waiting them to error out.*
  - Network interface is flapping up and down a few times per day
- Root Cause
  - Continuous Time Linear Equalizer (CTLE) setting incompatible between network adapter and network switch
- Resolution
  - Apply a patch to the network switch

Case study: A post-disaster survival experience cont.

- Issue 2 – Some compute nodes network interface dropping packets
  - *cngb-compute-e01-4: err: 109 drop: 1438345 frame_error: 109 nfs_err_num:8*
  - *cngb-compute-e03-5: err: 0 drop: 65514579 frame_error: 0 nfs_err_num: 0*
  - */opt/gridengine/default/spool/cngb-compute-f22-5/job_scripts/5054771: line 2: echo: write error: Input/output error ln: accessing `/zfs2/ntdb/group5_20/result/process/CNV_CNVnator/101608698/chr14/101608698.bam': Cannot send after transport endpoint shutdown*
- Root Cause
  - MTU size mismatch
- Resolution
  - Set and verify MTU setting for all network adapter and network switch

Case study: A post-disaster survival experience cont.

- Issue 3 – 25G NICs bonding showing only half the throughput
- Root Cause
  - Network adapter was put in a low speed PCI Express x4 slot
- Resolution
  - Relocate the network adapter to a PCI Express x8 slot

Case study: A post-disaster survival experience cont.

- Issue 4 – Cross data center network interruption
  - *00000800:00000100:57.0:1528296092.423675:0:9079:0:(socklnd_cb.c:2276:ksocknal_find_timed_out_conn()) A connection with 12345-10.xxx.xxx.xx@tcp (10.xxx.xxx.xxx:988) timed out; the network or node may be down.*
  - *00000100:00000400:31.0:1528296182.884581:0:108584:0:(client.c:2114:ptlrpc_expire_one_request()) @@@ Request sent has timed out for slow reply: [sent 1528296065/real 1528296065] req@ffff8859b92123c0 x1596679597925056/t0(0) o41->lustre-MDT0000-mdc-ffff8840501ca800@10.xxx.xxx.xxx@tcp:12/10 lens 224/368 e 0 to 1 dl 1528296182 ref 2 fl Rpc:X/0/ffffffff rc 0/-1*
  - *00000100:02000400:31.0:1528296182.884615:0:108584:0:(import.c:178:ptlrpc_set_import_discon()) szebra-MDT0000-mdc-ffff8840501ca800: Connection to lustre-MDT0000 (at 10.xxx.xxx.xxx@tcp) was lost; in progress operations using this service will wait for recovery to complete*
- Root Cause
  - ARP proxy occasionally not responding
- Resolution
  - Use router instead of ARP proxy

Case study: A post-disaster survival experience cont.

- Issue 5 – HBA reset cause ZFS hang
  - *May 31 02:43:11 cngb-oss-a2-2 kernel: mpt3sas_cm3: fault_state(0x5853)!*
  - *May 31 02:43:11 cngb-oss-a2-2 kernel: mpt3sas_cm3: sending diag reset !!*
  - *May 31 02:43:12 cngb-oss-a2-2 kernel: mpt3sas_cm3: diag reset: SUCCESS*
- Root Cause
  - Firmware version phase15IT/22.00.00.00 of HBA is not compatible with RHEL 7.4
- Resolution
  - Upgrade HBA firmware

Case study: A post-disaster survival experience cont.

- Other issues
  - Fiber optic cable defects. *cngb-compute-e03-8: err: 1000601 drop: 913804 frame_error: 1000601 nfs_err_num: 1377*
  - CRC error increasing in 25G NIC
  - Hybrid port VLAN tagging disordered
- Resolution
  - Replace fiber optical cables
  - Configure FEC in Optical module, and replace some faulty optical module
  - Apply a patch to switch
- Lessons learned:
  - Try before you buy
  - Software version do matters
  - Burn-In Test is not nice to have

Part Five

Summary

Summary

- Lustre is super fast and stable enough, but it's still complex and fragile, IMHO.
  - Technical support is very important.
- POC and benchmarking for systems you want to buy.
  - Commercial product like DDN is good
  - If you want to build your own Lustre system, do extensive benchmarking.
- It's worth taking a hard look at ZFS.
- Ethernet is not the best option for Lustre, but it works.
  - Carefully tune the timeout parameters.
- You will encounter Lustre Bug one day.
  - Prepare for it. Set up Core Dump. Monitor any metrics related. Collect logs.