Data storage services at KEK/CRC
-- status and plan

KEK/CRC

Hiroyuki Matsunaga

Most of the slides are prepared by Koichi Murakami and Go Iwai

KEKCC System Overview

KEKCC (Central Computing System)

- The only large cluster system that supports various projects in KEK
  - We also operate supercomputers to support different user communities
- In operation since April 2012
  - 3.5-year lease system (until Aug. 2015) in the original plan
  - Decided to extend the lease period by one year (until Aug. 2016)
  - Such a large system is fully replaced every 4-5 years by bidding, by following the Japanese government procurement system
  - Migrated to GHI (GPFS-HPSS interface) from VFS/client-API/PFTP
    - Big challenge for HPSS / HSM

IBM

IBM DataDirect NETWORKS QLOGIC intel Xeon inside

CPU

- Work (interactive) server & Batch server
  - Xeon 5670 (2.93 GHz / 3.33 GHz TB, 6core)
  - 282 nodes : 4GB memory/core
  - 58 nodes : 8GB memory/core
  - 2 CPU/node : 4080 cores
  - Scientific Linux 5
- Interconnect
  - InfiniBand 4xQDR (4GB/s), RDMA
  - Connection to storage system
- Job scheduler
  - LSF (ver. 8 -> ver.9 in last Aug.)
  - Scalability up to 1M jobs
- Grid deployment
  - EMI CREAM CE
  - Work server also acts as UI, Batch server as WN

IBM System x iDataPlex

Disk

- DDN SFA10K x 6 racks
  - Capacity : 1152TB x 6 = 6.9 PB (effective), 3TB HDD
  - Throughput: 12 GB/s x 6
  - Used for GPFS and GHI staging area (3PB each)
- GPFS file system
  - Parallel file system
  - Total throughput : > 50 GB/s
  - Optimized for massive access
    - number of file servers
    - no bottle-neck interconnect, RDMA-enabled
    - Separation of meta-data area (on SSD)
    - larger block size
- Performance
  - >500MB/s for single file I/O in benchmark test

DDN SFA10000

DataDirect NETWORKS

Tape

IBM TS3500

- Tape Library
  - Max. capacity : 16 PB
- Tape Drive
  - TS1140 : 60 drives
    - Latest enterprise drive
  - Did not choose LTO because of less reliability
- Tape Media
  - JC : 4TB, 250 MB/s
  - JB : 1.6TB (repack) , 200 MB/s
  - Tapes are provided by each user/group

IBM TS1140

HPSS
High Performance Storage System

HSM (Hierarchical Storage Management)

- HPSS
  - Disk (first layer) + Tape (second layer)
  - Experience in the previous KEKCC

- Improvements from the previous system
  - More tape drives
  - Faster I/O speed for tape drive
  - Faster interconnect (10GbE, IB)
  - Performance improvement on staging area (capacity, access speed)
  - Integration with GPFS file system (GHI)

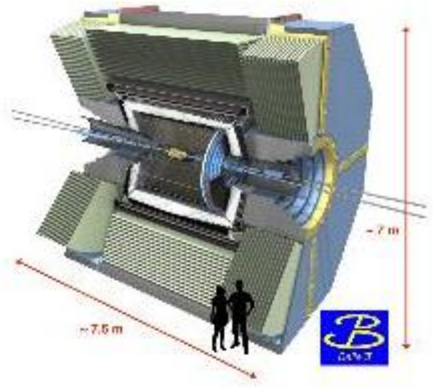
- GHI (GPFS-HPSS interface)
  - GPFS as staging area
  - Perfect coherence with GPFS access (POSIX I/O)
    - No use of HPSS client API and VFS interface
    - Taking advantage of high performance I/O of GPFS
  - Grid SE and iRODS access to GHI

Storage Management

Data Management Cycle

Different use cases

HEP



Big Data, different kinds of data sets

Raw data : Experimental data from detectors

- Written to HSM in real-time (through several buffer stages)
- High availability is required
- Cold data (PB to hundreds PB)
- Reprocessed from time to time
  - Reprocessing frequencies depend on stage of the experiment

DST (summary data) : For data analysis

- Hot data (1 – tens of PB)
- Access from jobs

Data-intensive processing

- High I/O performance required.
- Hundreds MB/s of I/O, many concurrent accesses from jobs

Material Science

Large amount of Image data taken by detectors

- A few PB/year expected
- Service as data archive
- Easy accessibility rather than I/O performance

Tape Storage Technology is Important

We are facing a challenge of Big Data.

Hundreds of PB of data is expected in new HEP experiments.

- Belle II expects 2-3 hundred PB/year at peak luminosity of the SuperKEKB accelerator.
- Cannot afford a disk-only storage solution.
- Much less electricity cost for tape storage.

On the other hand:

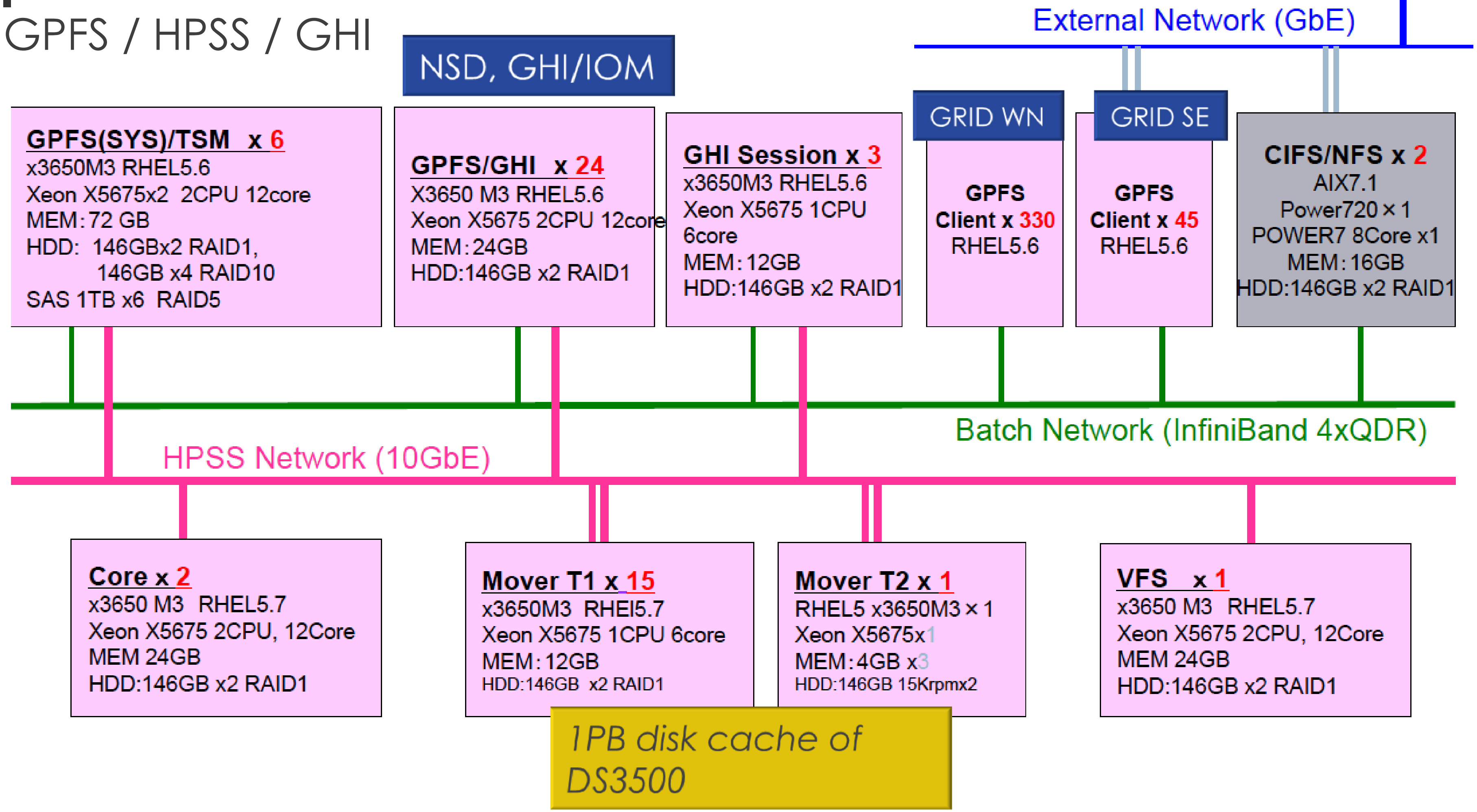
- *Performance*, *Usability* and *Long-term Preservation* are also very important.
- Middleware (HSM) in addition to hardware is a key point.

GHI, GPFS + HPSS : ***A Best solution at present***

- Data access with high I/O performance and good usability
  - Same access speed as GPFS once data are staged
  - No need for HPSS client API nor changes in user codes
  - Aggregation of small files helps tape performance a lot

KEKCC Storage Configuration

GPFS / HPSS / GHI



External Network (GbE)

NSD, GHI/IOM

GPFS(SYS)/TSM x 6
x3650M3 RHEL5.6
Xeon X5675x2 2CPU 12core
MEM:72 GB
HDD: 146GBx2 RAID1,
146GB x4 RAID10
SAS 1TB x6 RAID5

GPFS/GHI x 24
X3650 M3 RHEL5.6
Xeon X5675 2CPU 12core
MEM:24GB
HDD:146GB x2 RAID1

GHI Session x 3
x3650M3 RHEL5.6
Xeon X5675 1CPU
6core
MEM:12GB
HDD:146GB x2 RAID1

GRID WN

GPFS
Client x 330
RHEL5.6

GRID SE

GPFS
Client x 45
RHEL5.6

CIFS/NFS x 2
AIX7.1
Power720 × 1
POWER7 8Core x1
MEM:16GB
HDD:146GB x2 RAID1

Batch Network (InfiniBand 4xQDR)

HPSS Network (10GbE)

Core x 2
x3650 M3 RHEL5.7
Xeon X5675 2CPU, 12Core
MEM 24GB
HDD:146GB x2 RAID1

Mover T1 x 15
x3650M3 RHEl5.7
Xeon X5675 1CPU 6core
MEM:12GB
HDD:146GB x2 RAID1

Mover T2 x 1
RHEL5 x3650M3 × 1
Xeon X5675x1
MEM:4GB x3
HDD:146GB 15Krpmx2

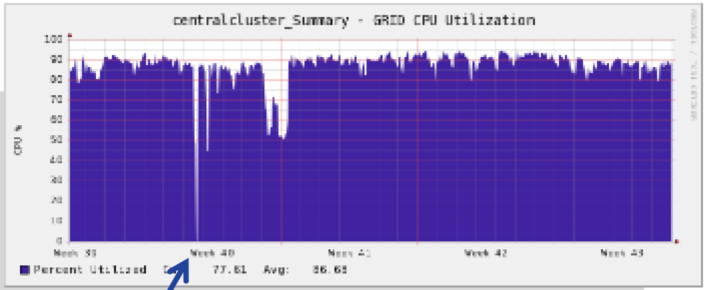
VFS x 1
x3650 M3 RHEL5.7
Xeon X5675 2CPU, 12Core
MEM 24GB
HDD:146GB x2 RAID1

1PB disk cache of
DS3500

GPFS : 3.5.0.18 / HPSS : 7.3.3.9.1a
GHI : 2.3.1.1

System Operations

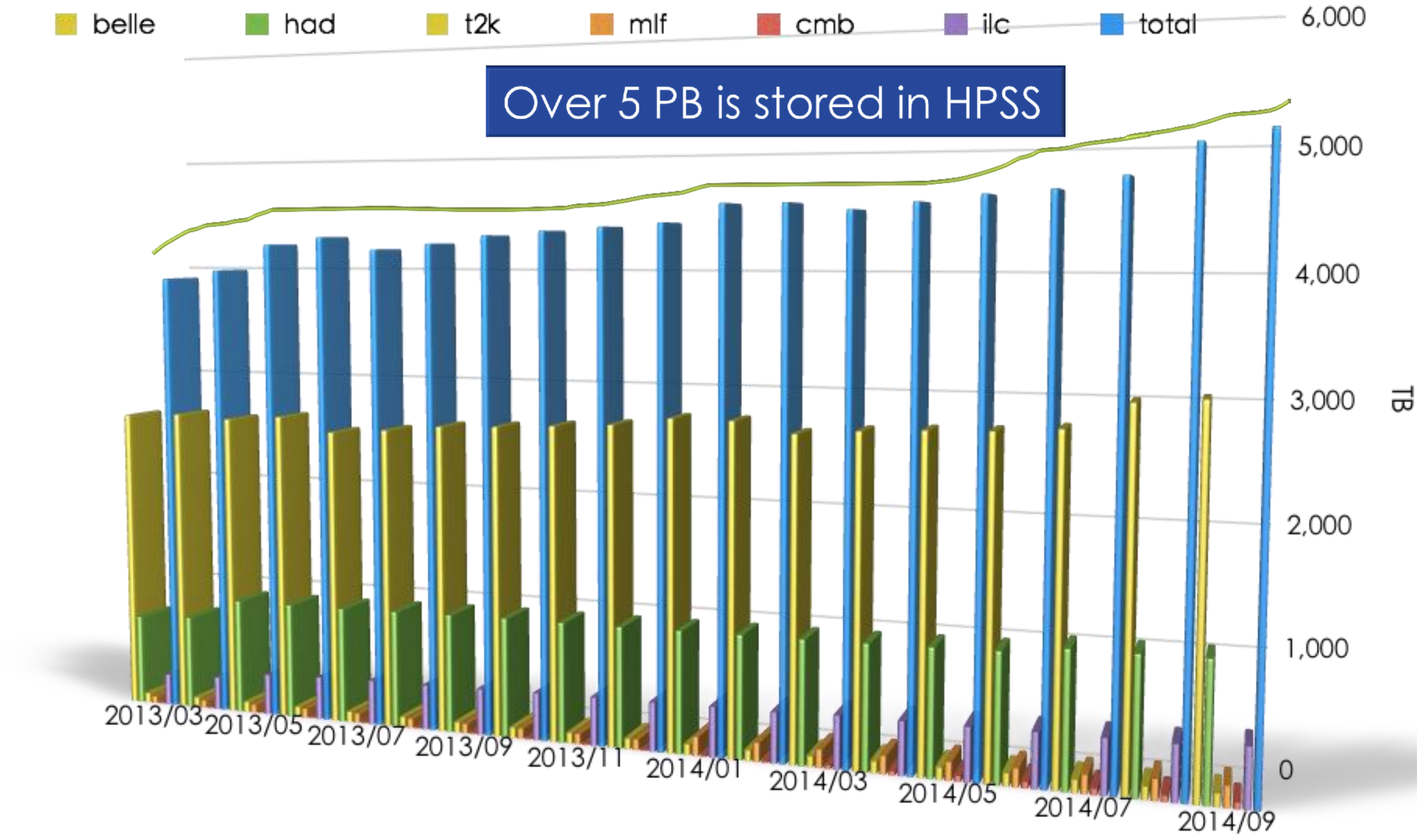
Storage system affects system performance and availability



GPFS	*Nearly stable* Main focus on performance issues (massive access, load balance, scalability). File system gets unmounted unexpectedly. Many nodes are affected while recovery. *Critical* in case of total outage of GPFS.
HPSS	*Almost stable* But occasional troubles with staging and data consistency
GHI	*Less stable, or problematic* Unclear behavior between GPFS and GHI might cause GPFS instability Staging outage, data loss

History of Data Stored in HPSS

For the past 1.5 year, over 5PB

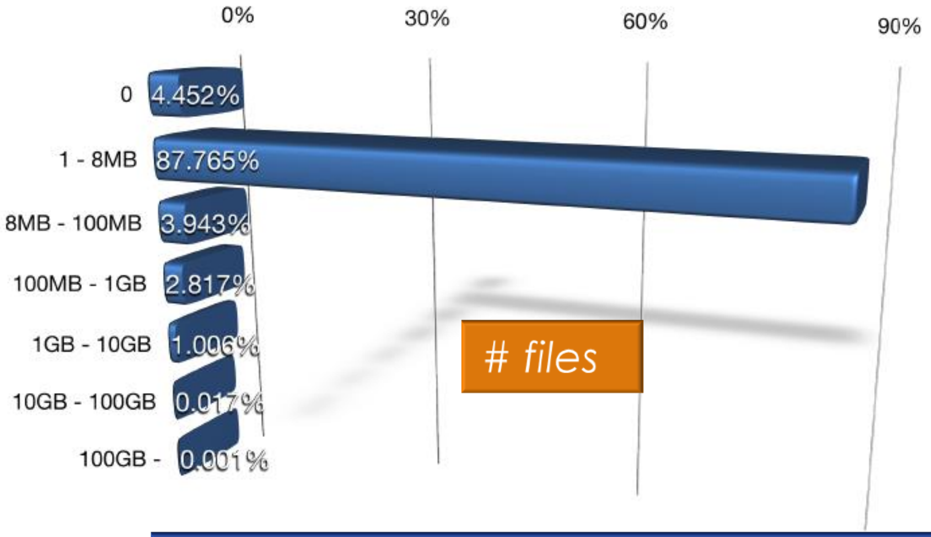


File Size Distribution

HPSS stats.

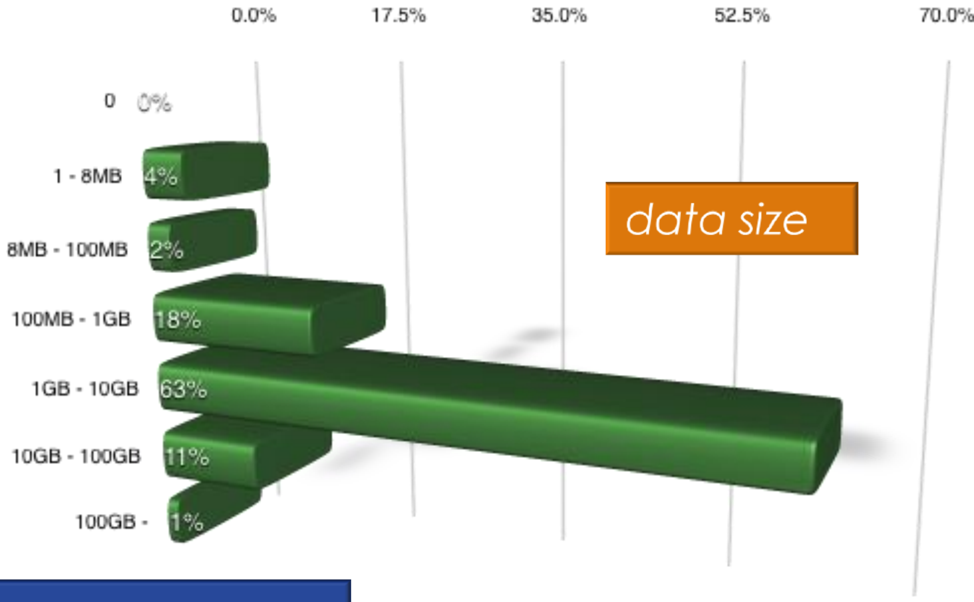
Number of files stored in HPSS

- 142 Million files in HPSS
- Files under 8MB are aggregated in HTAR.
  - To facilitate tape operation



files

File Size Distribution



data size

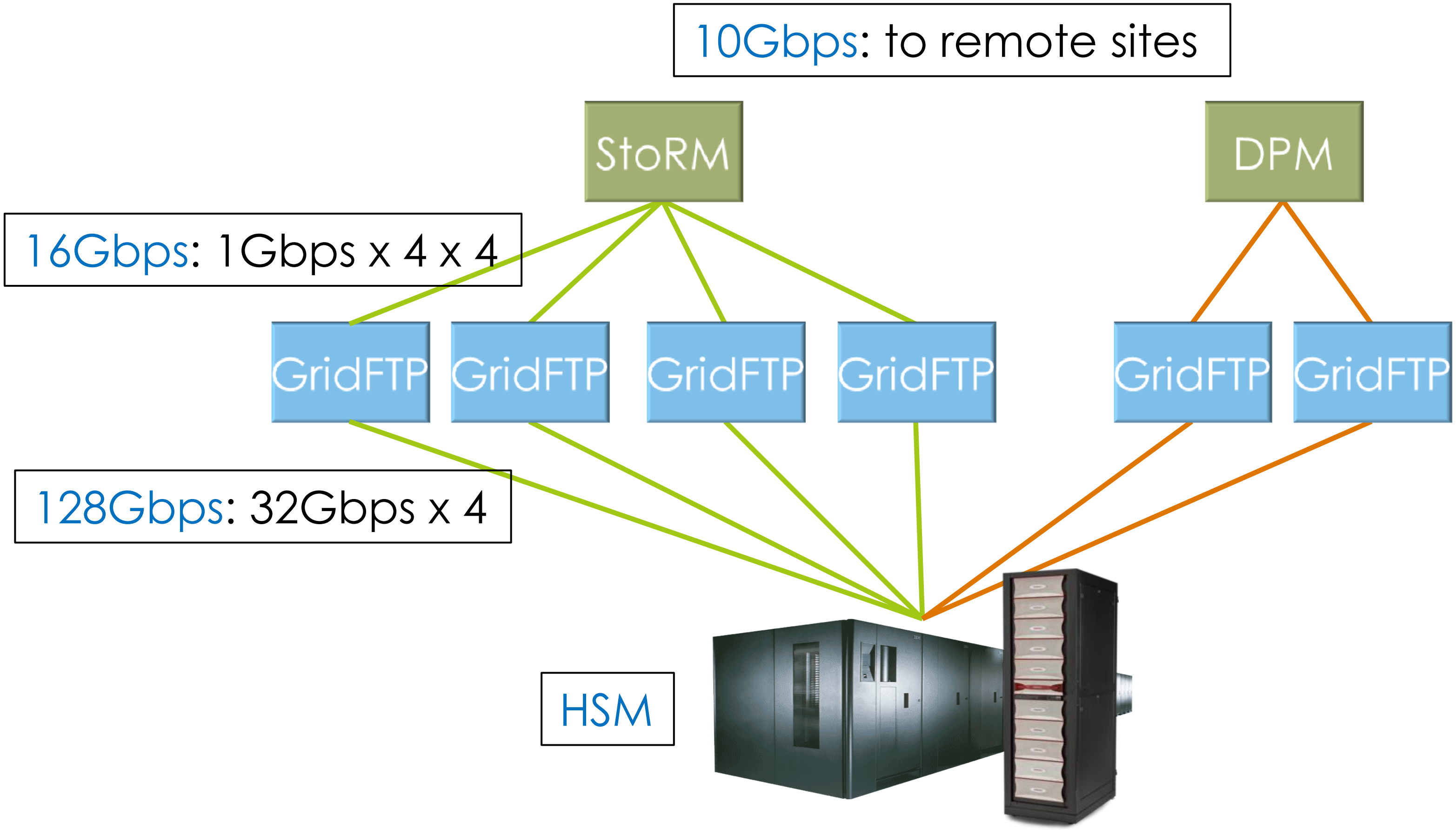
Average file size :

- depends on COS (Class of Service)
- data type : 10MB – 2GB

GRID Setup

EMI middleware

- StoRM is the main SE
- DPM will retire in the near future.



10Gbps: to remote sites

StoRM

DPM

16Gbps: 1Gbps x 4 x 4

GridFTP GridFTP GridFTP GridFTP

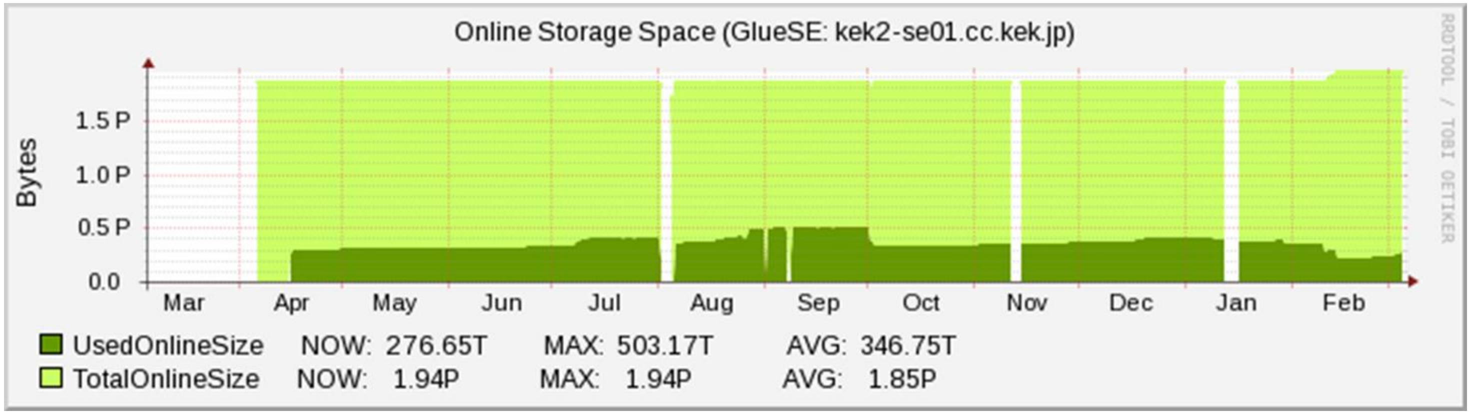
GridFTP GridFTP

128Gbps: 32Gbps x 4

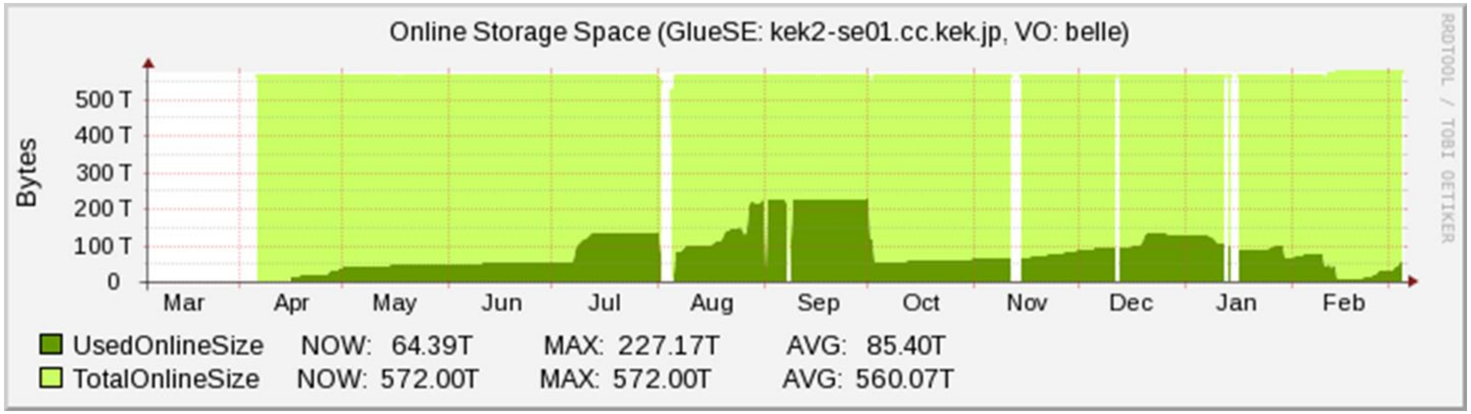
HSM

Storage Capacity for Grid

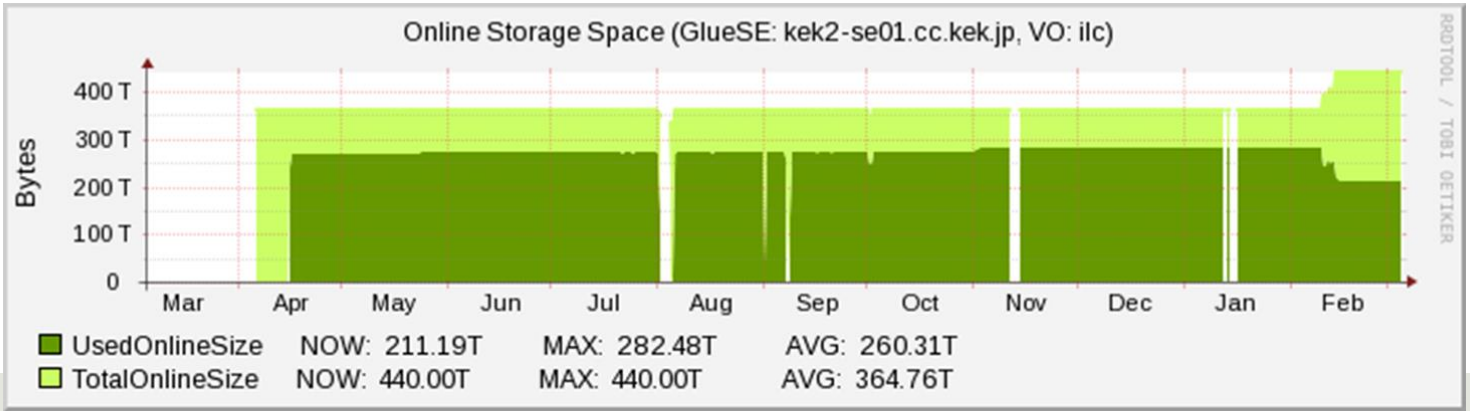
ALL VOs



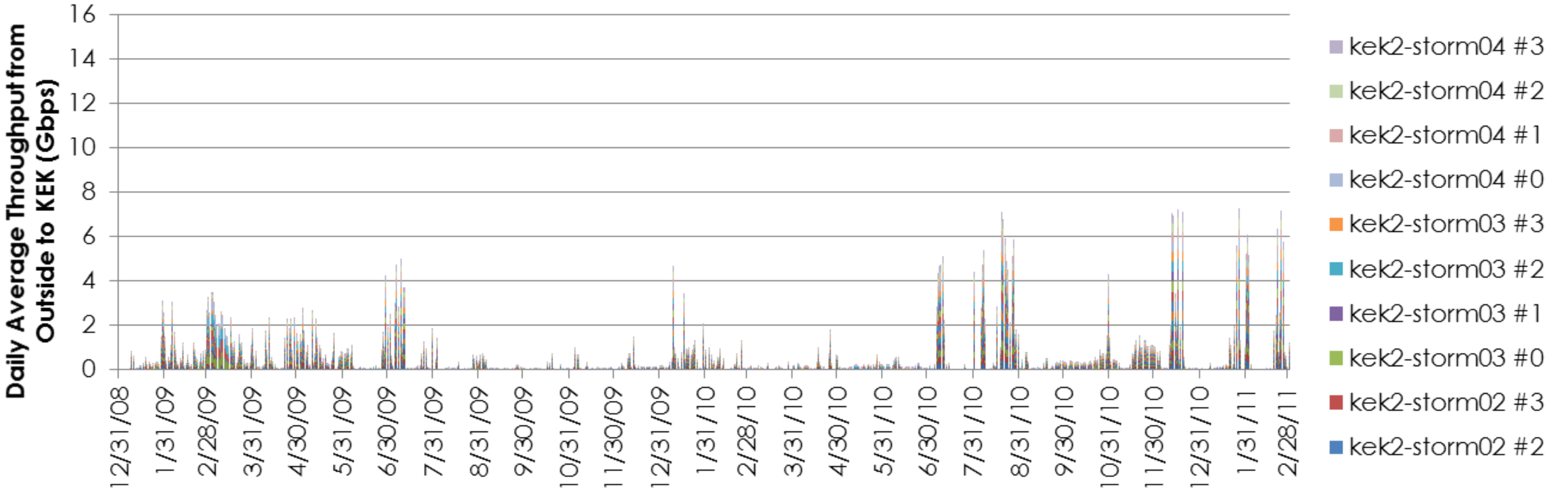
Belle VO (Belle & Belle II)

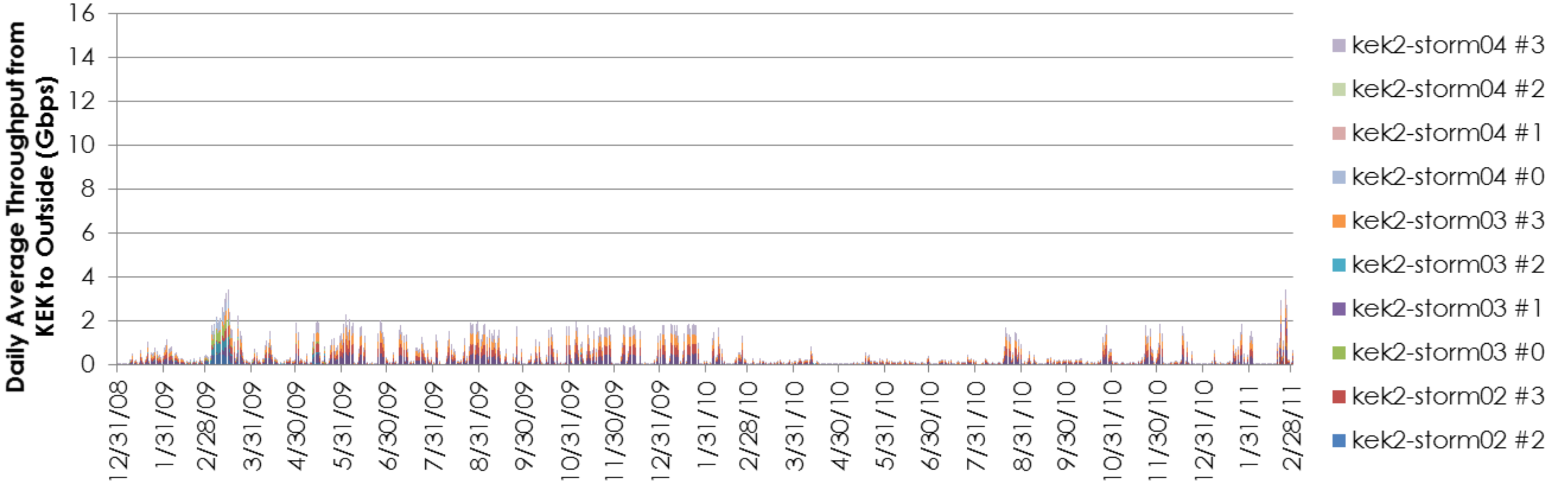


ILC VO

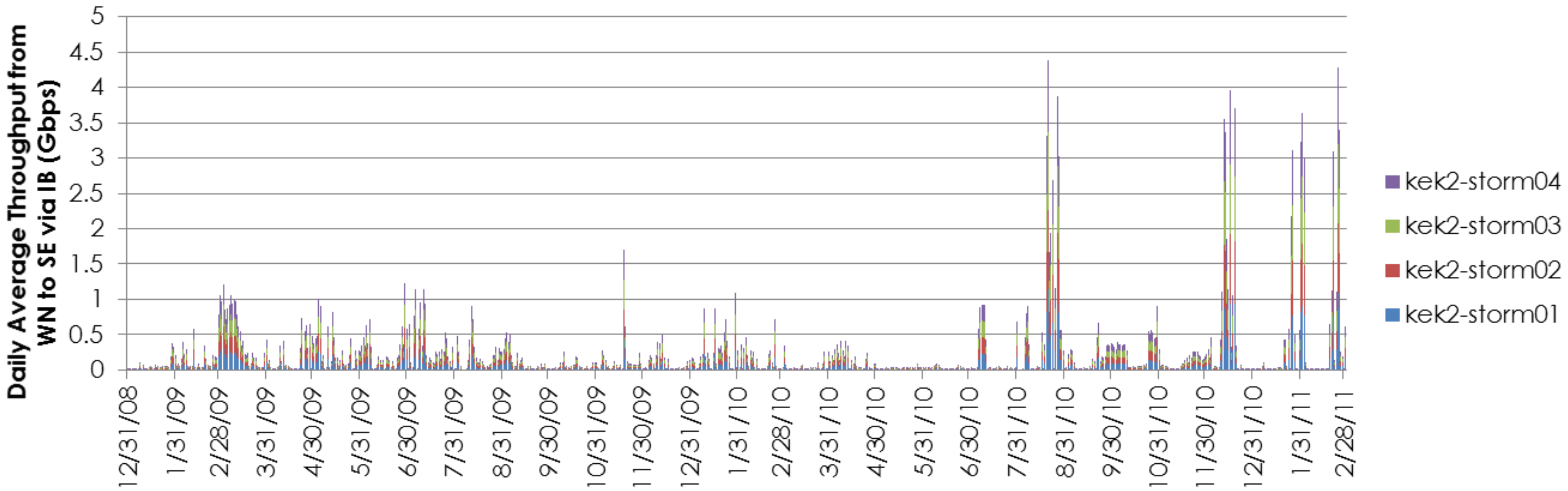


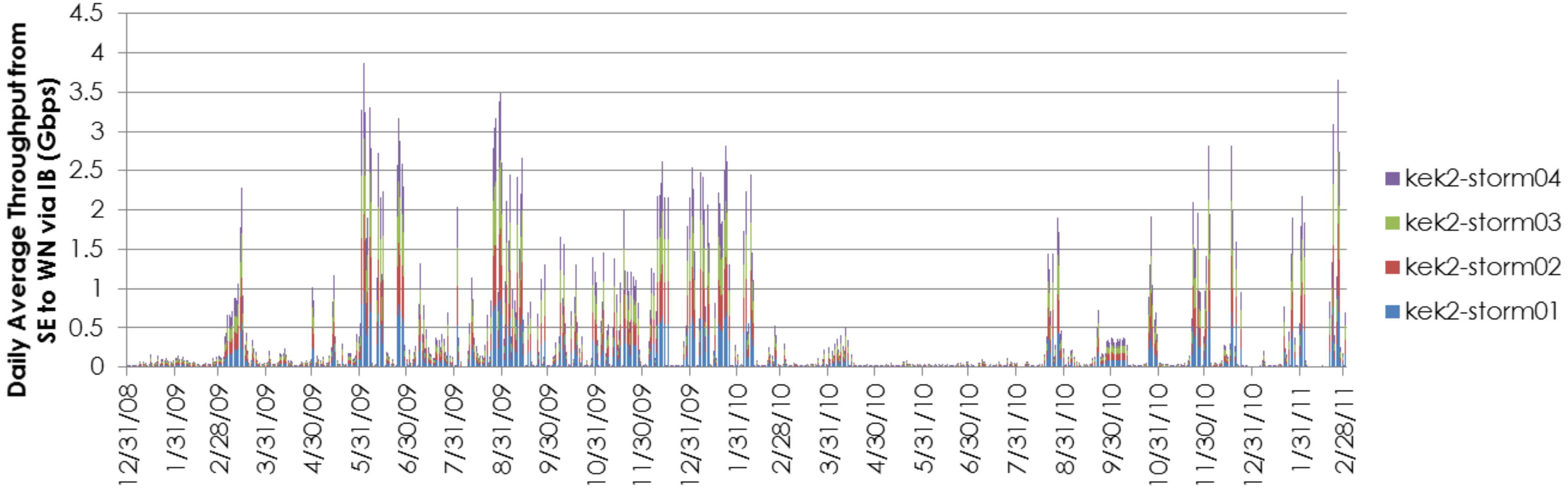
Throughputs for external traffic





Throughputs for internal traffic





Reduce Risk of Tape-Data Loss

- **Some data loss incidents occurred due to different reasons :** *software bugs, media errors, human errors*
  - Many users think that tape is cold device, safer device to preserve data for a long period.

- **How to reduce the data-loss risk?**
  - Hot media might be safer, since hardware health check seems to work well so far.
    - If a media degradation detected, then the media is copied out to another media soon.
  - Cold media might be potentially in danger.
    - Media could be damaged without notice.
      - Due to media degradation, dust intrusion, etc
  - Important data should be stored redundantly.
    - Write data in multiple media, via parallel writing or re-migration
    - User should pay extra media cost for preventing unexpected data loss.
  - Data integrity in HPSS could help.
    - Checksum, end-to-end integrity
    - Use trash can against human errors?

Data Explosion and Migration

- Hundreds of PB of data is anticipated in the next several years.
  - System is replaced every ~4 years.
    - HSM system (or tape library) may change by future bidding.
    - If that happens, data migration is a big issue.
      - How to do it efficiently and safely?
    - We expect 6-12 months for migration, but want to minimize the migration period for operational and lease cost reasons.
      - Two data storage systems should be running in parallel during the data migration.
- Technical issues
  - Checksum should be managed in HPSS for data integrity.
  - Optimization of reading out tape data in the recorded order.

Summary

- We have been a HPSS user for many years.
- GHI in operation since 2012.
- Tape system is important technology for us.
  - In terms of both hardware and software.
- GHI is a promising solution for HSM for a large scale of data processing.
- The peformance of HPSS / GHI is good, but the stability of GHI should be improved.
- Scalable data management is a challenge for next several years.
- Will upgrade the system two years later. Design and specification of the new system will start soon.