KEKCC – KEK Central Computer System

Go Iwai

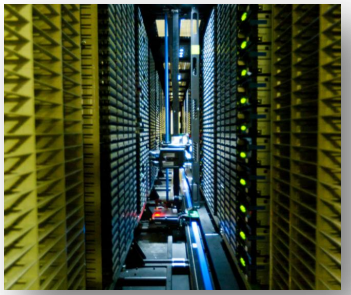
High Energy Accelerator Research Organization (KEK)
Computing Research Center (CRC)

Two Large-scale Computer Systems in KEK

Central Computer System (KEKCC)

Linux Cluster + GPFS/HPSS

4,000 cores (Xeon 5670)



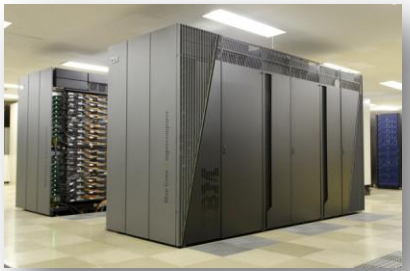
7 PB disk storage and tape library (up to 16 PB)

Grid instance is running in the KEKCC

Supercomputer System

System-A Hitachi SR16000 model M1

Total peak performance: 54.9 TFplops

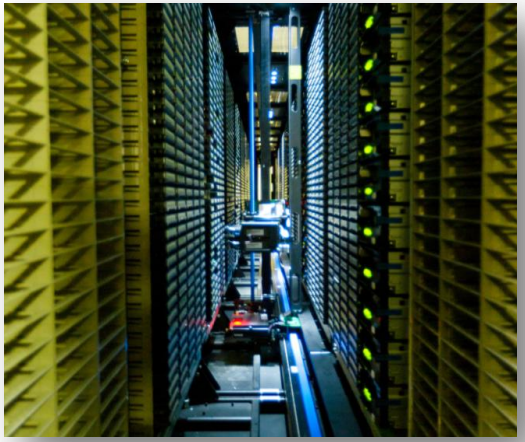


System-B IBM Blue Gene/Q

Total peak performance: 1.26 PFlops

KEKCC Overview

- Central Computer System supporting KEK projects, e.g. Belle/Belle2, ILC, J-PARC, and so on.
  - Current KEKCC has started in April 2012 and will be ended in August 2016.
- Data Analysis System
  - Login servers, batch servers
    - IBM iDataPlex, Intel Xeon X5670, 4,080 cores (12cores x 340nodes)
    - Linux Cluster (SL5) + LSF (job scheduler)
  - Storage system
    - DDN SFA10K 1.1 PB x 6 sets
    - IBM TS3500 tape library (16 PB max)
    - TS1140 60 drives
    - GPFS (4PB)+ HPSS/GHI (HSM,3PB)
    - Storage interconnect : IB 4xQDR (Qlogic)
    - Grid (EGI) SE, iRODS access to GHI
    - Total throughput : > 50 GB/s
- Grid computing system: EMI and iRODS
- System includes common IT services, mail, web (Indico, wiki,...) as well.



Who Are Working on KEKCC

- **Belle**
  - Belle experiment, precise measurements for CP violation.
- T2K
  - Neutrino experiment for measuring neutrino mass and flavour mixing. Shoot neutrino from Tokai to the detector at Kamioka mine (300km away)
- **Hadron experiments at J-PARC**
  - Various experiments for kaon and hadron physics
- Material and Life science at J-PARC
  - Neutron diffraction, neutron spectroscopy,
  - nano-structure analysis, neutron instruments, muon spectroscopy

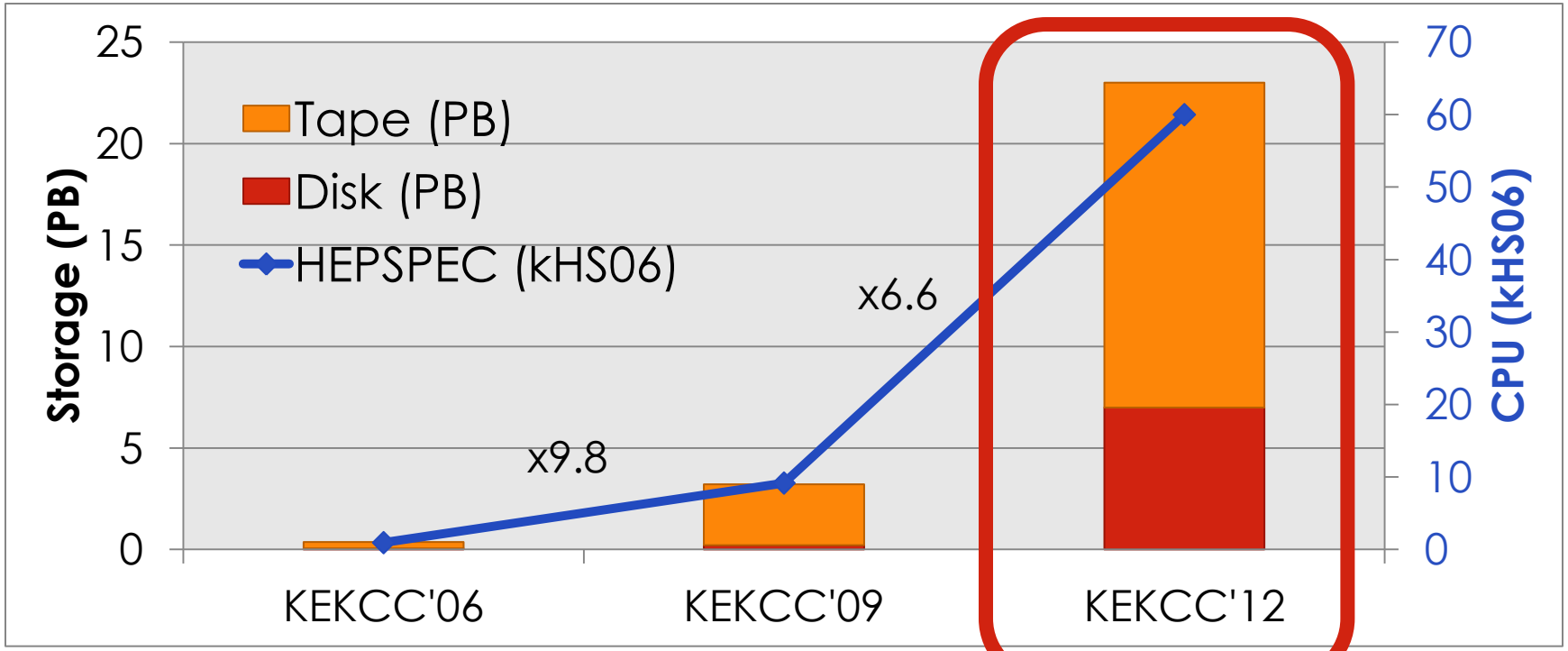
- **Belle2**
  - Belle II is the next generation Belle experiment. Aim to discover new physics beyond the SM. Physics run will start from 2017.
- Kagra
  - Gravitational wave (GW) detection experiment at Kamioka
  - Expected to start operation in 2018
  - Just started to utilize Grid computing resources for sharing data with LIGO and VIRGO.
- ILC
  - Linear collider experiment
  - Japan is a candidate site
  - Two active VOs: ILC and calice

These 3 experiments are big consumers last years.

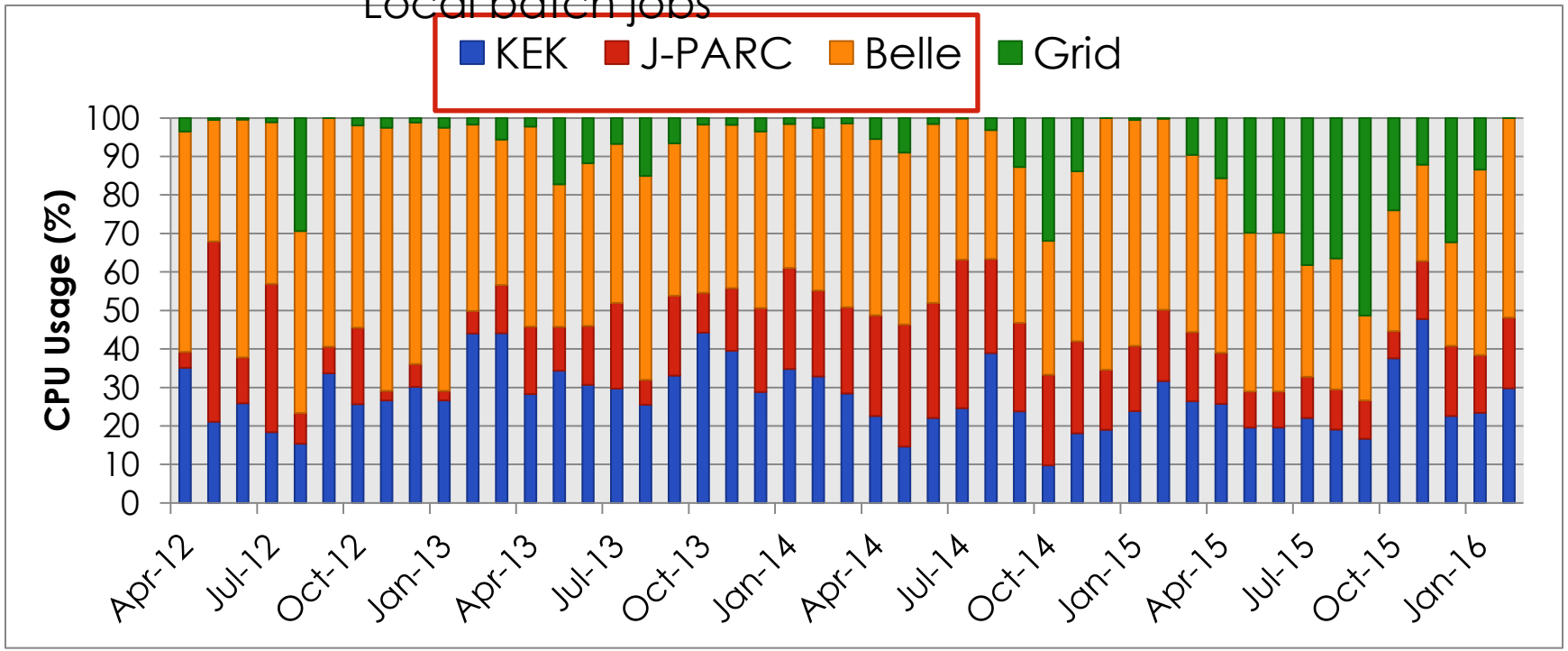
Experiments			Start Year	How work on KEKCC
Ongoing or Completed	Belle		1999	Mainly work on local batch servers
	J-PARC	T2K	2009	Mainly work on local batch servers Partly using Grid for data delivery
		Hadron Exp.	2010	Mainly work on local batch servers
		MLF (Material&Life Science)	2010	Mainly work on local batch servers Partly using Grid (iRODS) for data transfer
Future	Belle2		2017	Actively work on Grid
	Kagra		2018	Just started preparation work for sharing data over the Grid
	ILC		202X	Working on Grid, and local batch servers

Resource History

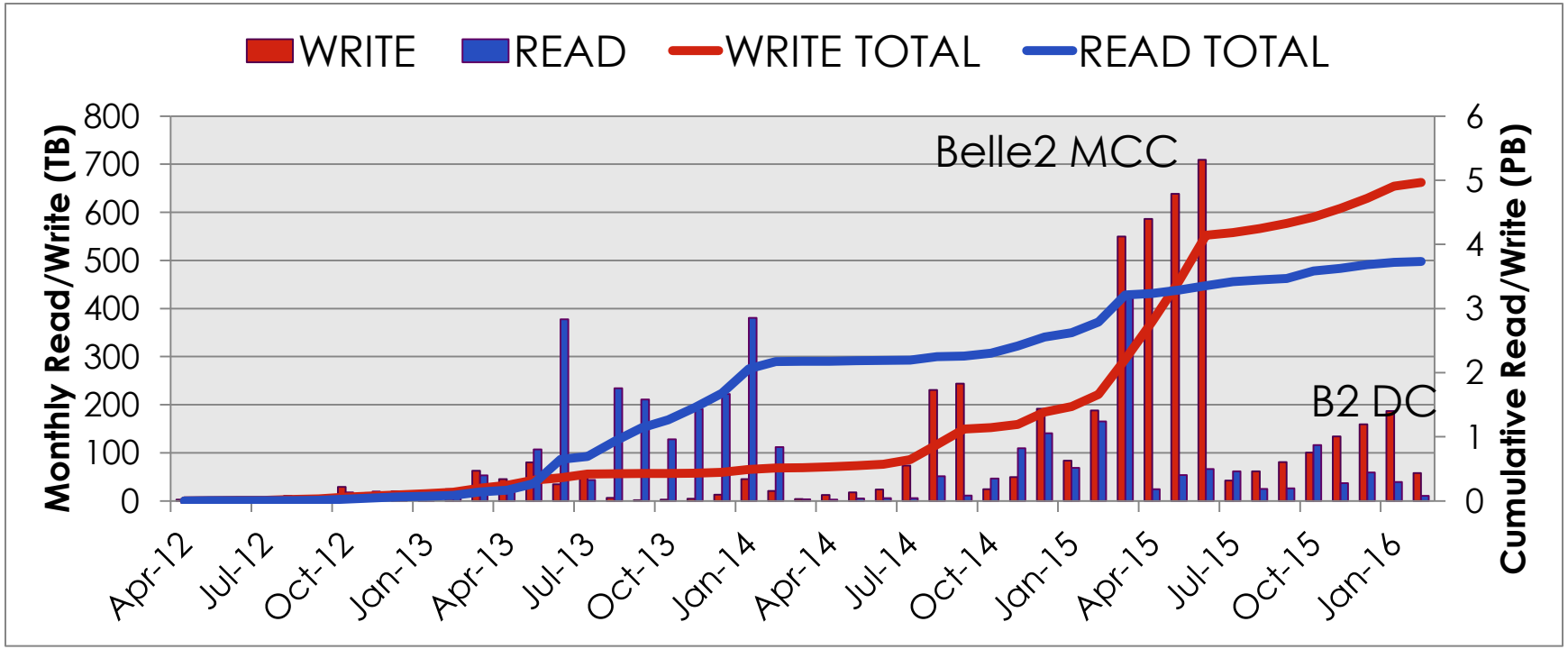
60 kHS06 of CPU
7 PB of disk
Max 16 PB of tape capacity



Grid is Now Using Nearly Half of KEKCC



Yearly 1PB of Read/Write to SRM



4 PB of readout and 5 PB of writing to the SRM has been achieved

System Procurement

- KEKCC is totally replaced every 4-5 years, according to Japanese government procurement rule for computer system.
  - International bidding according to GPA (Agreement on Government Procurement by WTO)
  - Bidding is processed for 1 year.
- Purchase and operation model
  - NOT in-house scale-out model, BUT rental system
  - Completely different purchase/operation model from US/EU sites
  - Much less human resource in computer center
    - 25 staffs (KEK/CRC) vs 270 staffs (CERN-IT)

Hardware purchase by lease
+
Service (implementation/operation staffs)

System Replacement Cycle

- Bidding is processed for 1 year.
  - Committee was launched in Feb/2015.
  - RFx (Request for Information/Proposal/Quotation)
  - RFC (Request for comments)
  - Bidding
    - Score for price + benchmark
    - Bid-opening was done on the end of Dec/2015.
- System implementation (Jan – Aug / 2016)
  - Facility updates (power supply, cooling)
  - Hardware installation
  - System design / implementation / testing

Service-in of the new system is scheduled on Sep/2016.

CPU Server

- Work server & Batch server
  - Xeon 5670 (2.93 GHz / 3.33 GHz TB, 6 cores/chip)
  - 282 nodes : 4 GB/core
  - 58 nodes : 8 GB/core
  - 2 CPU/node : **4,080 cores**, **60 kHS06**
- Interconnect
  - **InfiniBand 4xQDR** (32Gbps), RDMA
  - Connection to storage system
- Job scheduler
  - **LSF** (ver.9)
  - Scalability up to 1M jobs
- Grid deployment
  - EMI & iRODS
  - Work server as Grid-UI, Batch server as Grid-WN

Disk Storage

- DDN SFA10K x6
  - Capacity : 1,152 TB x 6 = **6.9 PB** (effective)
  - Throughput: 12 GB/s x 6
  - Used for GPFS (4PB) and GHI (3PB)
- GPFS File System
  - Parallel file system
  - Total throughput : **> 50GB/s**
  - Optimized for massive access
    - IB connection : non-blocking / RDMA
    - Number of file servers
  - Separation of meta-data area
  - Support for larger block size
- Performance
  - >500 MB/s for single file I/O in benchmark test

Tape System

- Tape Library
  - TS3500
  - Max. capacity : 16 PB
- Tape Drive
  - TS1140: 60 drives
  - latest enterprise drive
  - We do not use LTO because of less reliability.
    - LTO is open standard. Could be different quality of tape drive/media for a specification.
- Tape Media
  - JC: 4TB, 250 MB/s
  - JB: 1.6TB (repack), 200 MB/s
  - Users (experiment groups) pay tape media they use.
  - 7PB is stored so far.

Data Processing Cycle in HEP Experiments

~10-1,000PB

- Raw data
  - Experimental data from detectors, transferred to storage system in real-time.
    - 2GB/s, sustained for Belle II experiment
  - Migrated to tape, processed to DST, then purged
  - "Semi-Cold" data (tens to hundreds PB)
    - Reprocessed sometimes DST (Data Summary Tapes)

~10-100PB

- DST (Data Summary Tapes)
  - "Hot data" (~tens PB)
  - Data processing to make physics data
  - Data shared with various ways (GRID access)

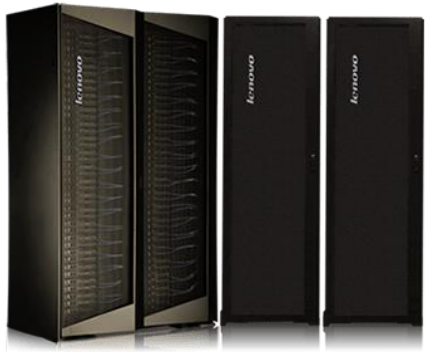
- Physics summary data
  - Handy data set for reducing physics results (N-tuple data)

- Requirements for storage system
  - High availability (considering electricity cost for operating acc.)
  - Scalability up to hundreds PB
  - Data-intensive processing w/ high I/O performance
    - Hundreds MB/s I/O for many concurrent accesses (Nx10k) from jobs
    - Local jobs and GRID jobs (distributed analysis)
  - Data portability to GRID services (POSIX access)

High Performance Tape Technology is the Key

- Hundreds PB of data is expected for new HEP experiments.
  - Cost-efficient on capacity
  - Less electricity cost
- Not only the cost/capacity issue,...
  - Performance, Usability and Long-term Preservation are also very important.
  - Hardware as well as middleware (HSM) are keys.

Next System Component



Lenovo

NextScale

Mellanox TECHNOLOGIES

SX6518

IBM

Elastic Storage Server

DataDirect NETWORKS

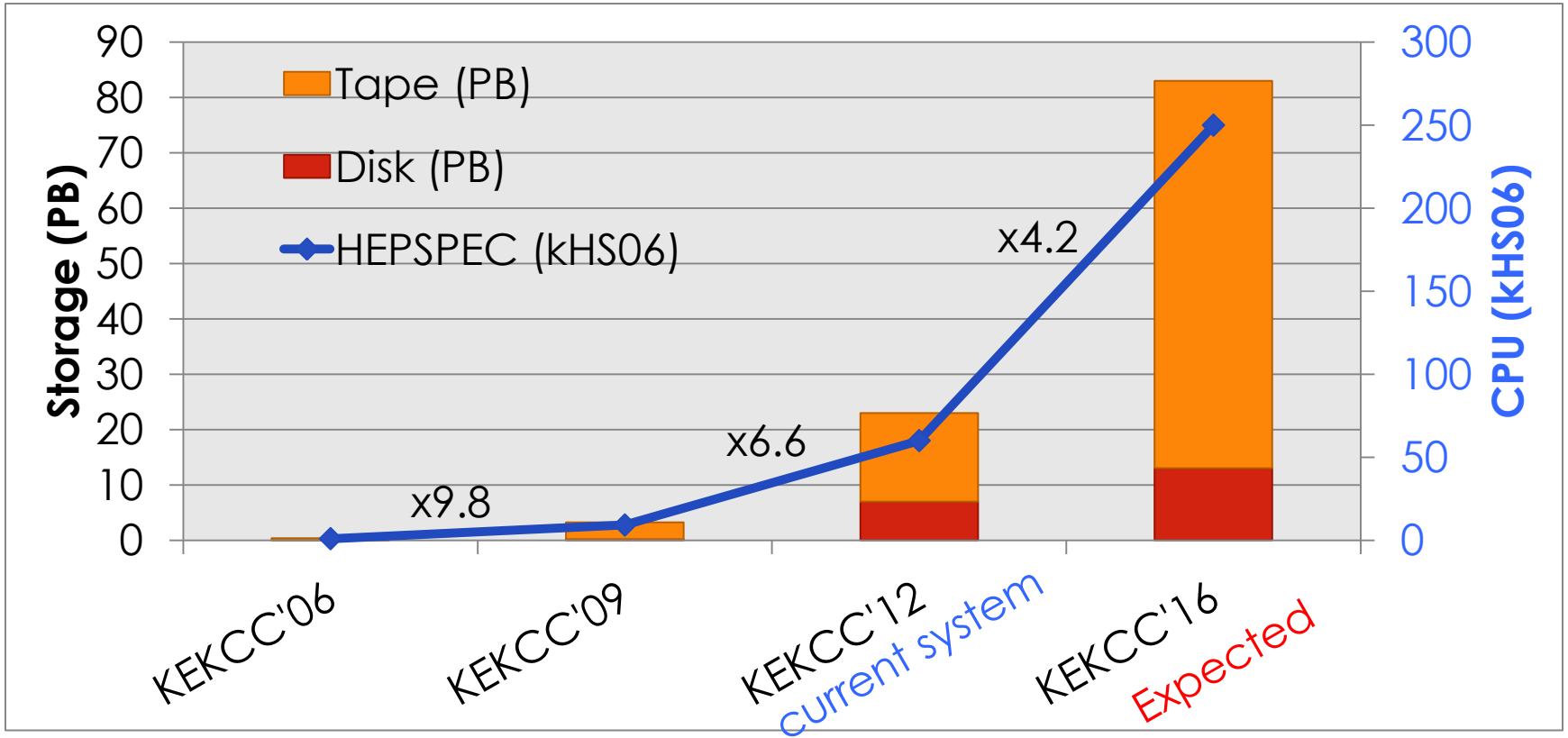
SFA 12K

Current VS Next

	Current	New	Upgrade Factor
CPU Server	IBM iDataPlex	Lenovo NextScale	
CPU	Xeon 5670 (2.93 GHz, 6 cores/chip)	Xeon E5-2697v3 (2.60 GHz, 14 cores/chip)	
CPU cores	4,000	**10,000**	**x2.5**
HEPSPEC (kHS06)	60	**250**	**x4.1**
IB	QLogic 4xQDR	Mellanox 4xQDR	
Disk Storage	DDN SFA10K	IBM Elastic Storage System (ESS)	
HSM Disk Storage	DDN SFA10K	DDN SFA12K	
Disk Capacity	7 PB	**13 PB**	**x1.8**
Tape Drive	IBM TS1140 x60	IBM TS1150 x54	
Tape Speed	250 MB/s	350 MB/s	
Tape max capacity	16 PB	**70 PB**	**x4.3**
Power Consumption	200 kW (actual monitored value)	< 400 kW (max estimation)	

Expected Resource in KEKCC'16

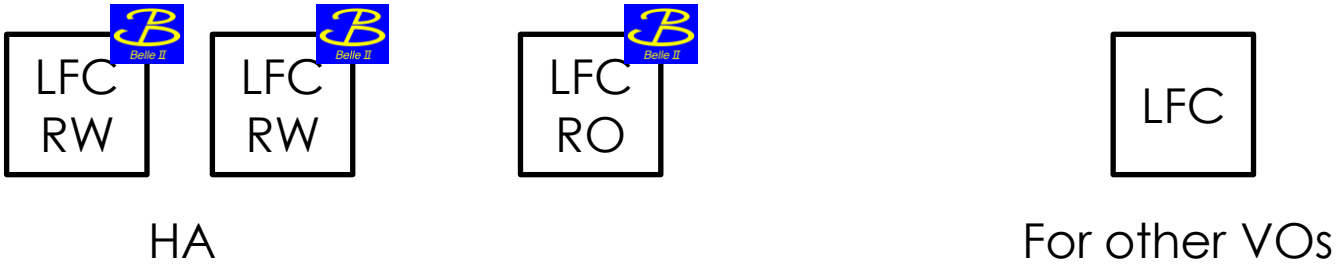
250 kHS06 of CPU
13 PB of disk
Max 70 PB of tape capacity



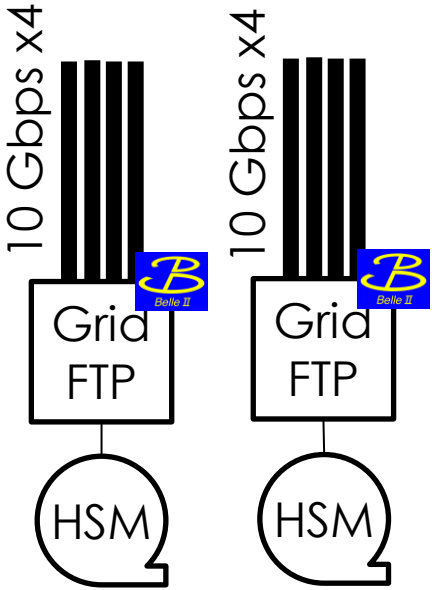
Belle2 Dedicated Services

- The big change on Grid is many Belle2-critical services, e.g. LFC, SRM, AMGA, FTS, CVMFS S0, are isolated to the other VOs for more stable operation with no downtime.

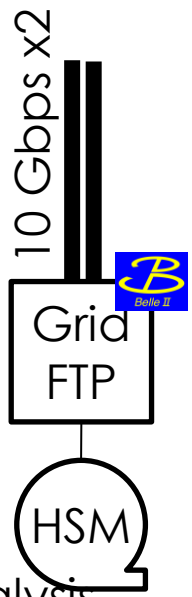
Example 1



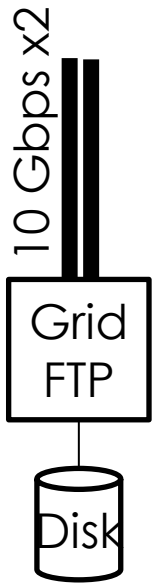
Example 2

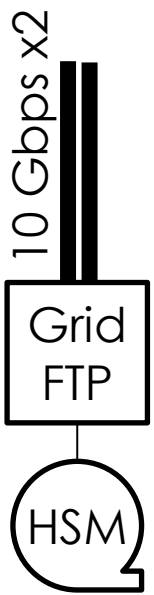


RAW data transfer to US



Analysis
Every Activities other than raw data transfer





For other VOs

Situation on Security

- We have had no serious security incident during the current contract period of KEKCC.
- Nevertheless the security cost is increasing in recent years.

- Background:
  - Many identity fraud: many personal identifiable information has been stolen from governmental institutes some times.
  - "My Number" (Individual Number) has started in 2015.

- Japanese people is getting nervous, government is getting more nervous.
- Government suddenly gives us an order to provide the investigative reports for all machines, which connects to the Internet.
  - Nearly 1,000
- We have no full-time staff for these kind of works.
  - DUTY-STOP during this works

Summary

- Grid job is being the biggest consumer in the KEKCC.
  - 30-50% of CPU time
- Next KEKCC system will start in September 2016.
  - CPU : 10K cores (x2.5), 250 kHS06 (x4.1), Disk : 13PB (x1.8), Tape : 70PB (x4.3)
  - Some Grid components (LFC, StoRM, etc) will be deployed as Belle2-dedicated services on Belle2-dedicated servers with HA.
  - CVMFS stratum-0 and -1 will start.
- Tape system is still important technology for us, not only hardware but software (HSM) points of view.
  - We have been a HPSS user for long years. We adopt GHI since 2012.
  - GHI is a promising solution for HSM for large scale of data processing.
- Scalable data management is a challenge for next 10 years.
  - Belle2 experiment will start in 2017.
  - Data processing cycle (data taking, archive, processing, preservation...)
  - Workload management w/ cloud technology: Job scheduler (LSF) + Virtualization (KVM, Docker)
  - Data migration as a potential concern

Thank You!