SYSTEM UPGRADE OF KEK CENTRAL COMPUTING SYSTEM

Koichi Murakami (KEK/CRC)

AFAD2016 Kyoto Feb./01/2016
WG6: Network & Computing

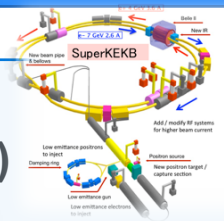
OUTLINE

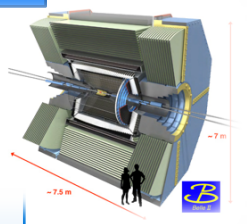
- INTRODUCTION OF KEK
- SYSTEM PURCHASE MODEL
- CURRENT KEKCC SYSTEM (2012-2016)
- NEW KEKCC SYSTEM (2016/AUG - 2020)
- SUMMARY

KEK

MULTIDISCIPLINARY LABORATORIES

Acceralator facilities in two sites





Tsukuba Campus :

Super KEKB (e^+-e^- collider)

Belle II (B physics)

Photon Factory (synchrotron facility)

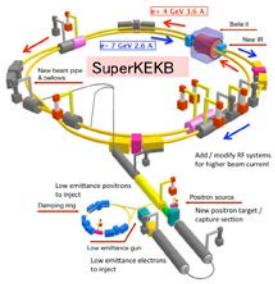
Material science

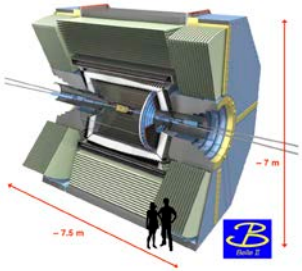
Tokai Capmus:

J-PARC (Proton synchrotron accelerator)

- T2K (Neutrino experiment)
- Hadron experiments
- MLF (Material and Life science Facility

ON-GOING PROJECTS





BELLE, BELLE II EXPERIMENTS

Belle experiment, precise measurements for CP violation.

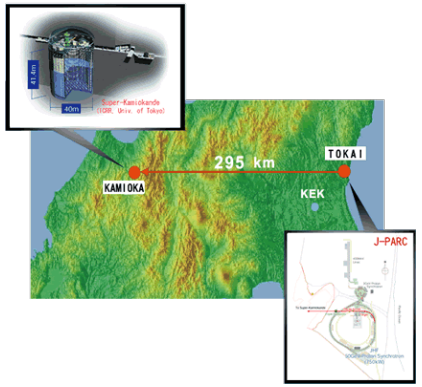
Belle II is the next generation Belle experiment. Aim to discover new physics beyond the SM.

Physics run will start from 2017.

T2K

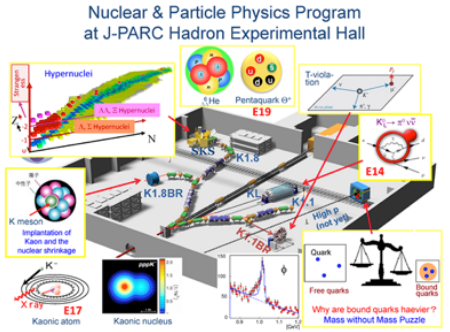
Neutrino experiment for measuring neutrino mass and flavour mixing.

Shoot neutrino from Tokai to the detector at Kamioka mine (300km away)



HADRON EXPERIMENTS AT J-PARC

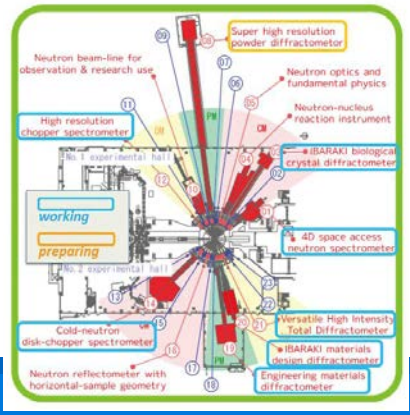
Various experiments for kaon and hadron physics



MATERIAL AND LIFE SCIENCE AT J-PARC

Neutron diffraction, neutron spectroscopy,

nano-structure analysis, neutron instruments, muon spectroscopy



SYSTEM PURRCHASE MODEL

System 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)

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.

KEKCC - DATA ANALYSIS SYSTEM

Central Computing System supporting KEK projects

- Operation started in April 2012.
- System includes IT service such as mail, web (Indico, wiki,…), etc.

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

CPU SERVER

WORK SERVER & BATCH SERVER

- Xeon 5670 (2.93 GHz / 3.33 GHz TB, 6core)
- 282 nodes : 4GB /core
- 58 nodes : 8GB /core
- 2 CPU/node : 4,080 cores

INTERCONNECT

- InfiniBand 4xQDR (32Gbps), RDMA
- Connection to storage system

JOB SCHEDULER

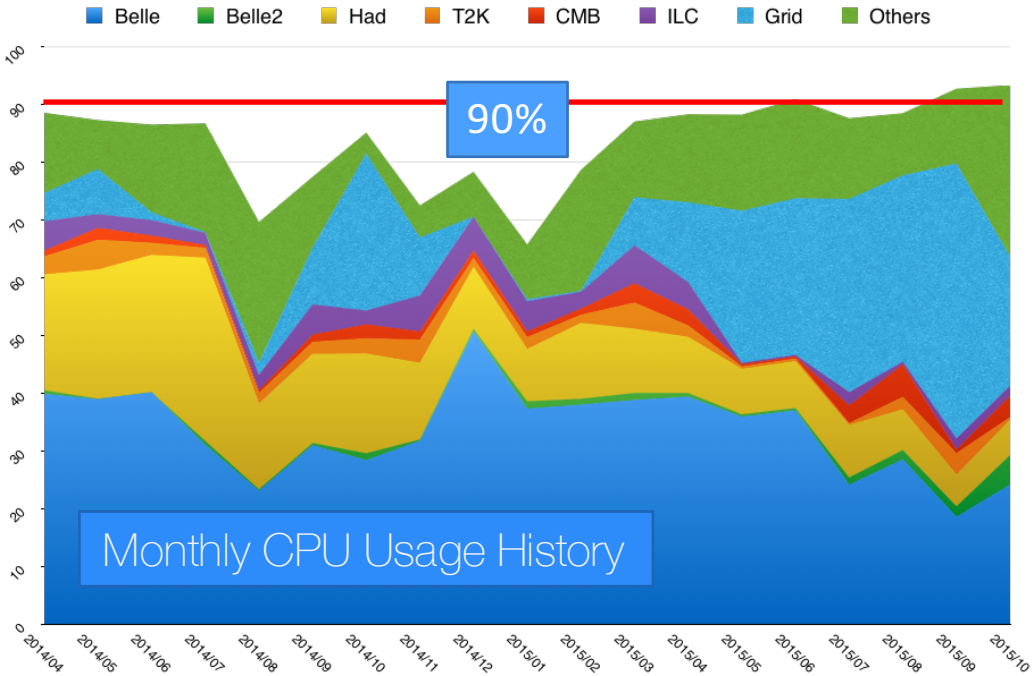
- LSF (ver.9)
- Scalability up to 1M jobs

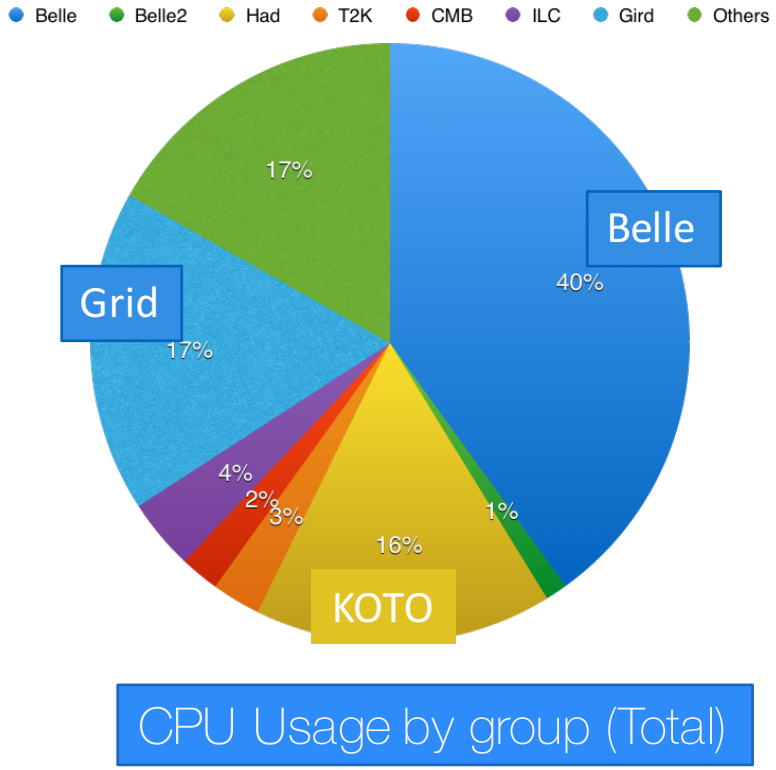
GRID DEPLOYMENT

- EMI
- Work server as Grid-UI, Batch server as Grid-WN

IBM System x iDataPlex

CPU USAGE STATS. (2014.04-2015.10)





CPU Usage by group (Total)

CPU resource is almost full.

Breakdown of group usage:

- Belle / Belle2 (incl. Grid jobs) : 60%
- J-PARC (KOTO, T2K, Hadron) : 20%

DISK STORAGE

DDN SFA 10K X 6

- Capacity : 1152TB x 6 = 6.9 PB (effective)
- Throughput: 12 GB/s x 6
- used for GPFS and GHI

GPFS FILE SYSTEM

- Parallel file system
- Total throughput : > 50 GB/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

- >500MB/s for single file I/O in benchmark test

DDN SFA10000

TAPE SYSTEM

IBM TS3500

TAPE LIBRARY

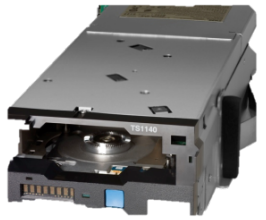
- Max. capacity : 16 PB

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.



IBM TS1140

DATA PROCESSING CYCLE IN HEP EXPERIMENTS

RAW DATA

- Experimental data from detectors, transferred to storage system in real-time.
  - 2GB/s, sustained for Belle II experiment
- x5 the amount of simulation data
- Migrated to tape, processed to DST, then purged
- "Semi-Cold" data (tens to hundreds PB)
  - Reprocessed sometimes

DST (DATA SUMMARY TAPES)

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

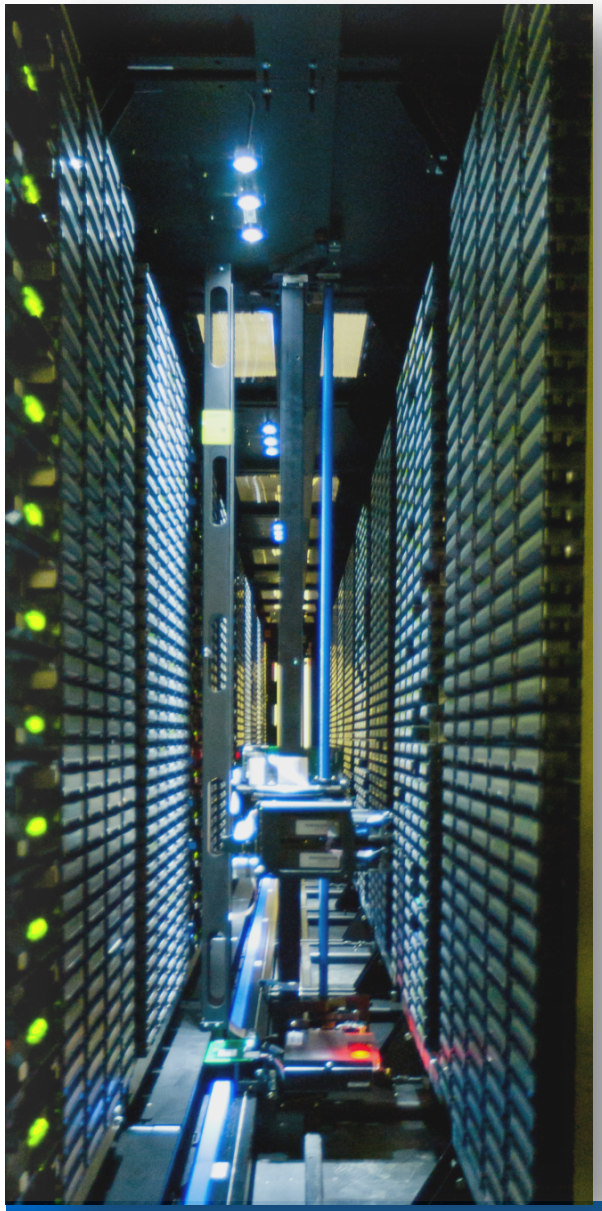
PHYSICS SUMMARY DATA

- 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.

GHI, GPFS + HPSS : THE BEST OF BOTH WORLDS

HPSS

- For exascale storage of DOE labs. Collaboration between DOE labs. and IBM.
- We have used HPSS as HSM system for last 15 years.

GHI, GPFS + HPSS : THE BEST OF BOTH WORLDS

- GPFS parallel file system staging area
- Perfect coherence with GPFS access (POSIX I/O)
- KEKCC is the pioneer of GHI customers (since 2012).
- Data access with high I/O performance and good usability.
  - Same access speed as GPFS, once data staged
  - No HPSS client API, no changes in user codes
  - instead of former VFS/Fuse interface
  - small file aggregation helps tape performance for small data

HPSS
High Performance Storage System

REQUIREMENTS FOR THE NEXT SYSTEM 2016-2020

	CPU (cores)	Disk (PB)	Tape (PB)
Belle	1,000	1.2	3.5
Belle II	7,500	9	29
ILC	400	0.3	1.5
CMB	250	0.5	1
J-PARC	1,650	5.9	27
KOTO	1,000	5	15
T2K	300	0.2	1
MLF	50	0.5	8
Others (J)	300	0.2	3
Total	10,800	17	65
Current Sys.	4,000	7	18
Next Sys.	10,000	13	70

We cannot provide all required resources.

- less improvement on CPU performance, disk density
- resource management in various points of view is needed.
  - resource assignment, priority
  - workload management
  - improvement on software

In future, we have to consider

- space, power supply, cooling、UPS, …

NEXT KEKCC SYSTEM

NEW KEKCC SYSTEM

- Bidding process was ended on the end of Dec.
- 4-years contract
- Service-in : 2016/September -

SYSTEM RESOURCES

- CPU : 10,000 cores (x2.5)
  - Intel Xeon E5-2697v3 (2.6GHz, 14cores) x 2/node, 358 nodes
  - 4GB/core (8,000 cores) + 8GB/core (2,000 cores) (for app.)
- Disk : 10PB: 7PB (GPFS) + 3PB (GHI) (x1.8)
- Interconnect : IB 4xFDR
- Tape : 70 PB (max cap.) (x4.3)
- Total throughput : 100 GB/s (GPFS), 50 GB/s (GHI)

CURRENT VS NEXT

	Current	New	Upgrade Factor
CPU Server	IBM iDataPlex	Lenovo NextScale	
CPU	Xeon 5670 (2.93 GHz ,6core)	Xeon E5-2697v3 (2.6GHz, 14cores)	
CPU cores	4,000	10,000	x2.5
IB	QLogic 4xQDR	Mellanox 4xFDR	
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 x 60	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)	

NEXT SYSTEM COMPONENT

IBM Elastic Storage Server (ESS)

Mellanox TECHNOLOGIES

SX6518
IB 4xFDR

lenovo Next Scale

DataDirect NETWORKS

SFA 12000
for HSM Disk

HSM

HPSS/GHI servers

DataDirect NETWORKS

DDN SFA 12K

IBM

TS3500

SYSTEM SECURITY

JUNIPER NETWORKS

FW SRX3400

Sniper IPS

10GbE

CISCO

Nexsus 7018

Information security is an important matter.

- Operation cost is higher and higher.
- User education
  - management of private keys.
  - monitoring suspicious jobs
- Against system hacking
  - unauthorized accesses
  - patch system security vulnerability
- IPS (Intrusion Prevention System) is installed.
  - monitoring by JSOC (Security Operation Center) for 24h/365d.

DATA EXPLOSION

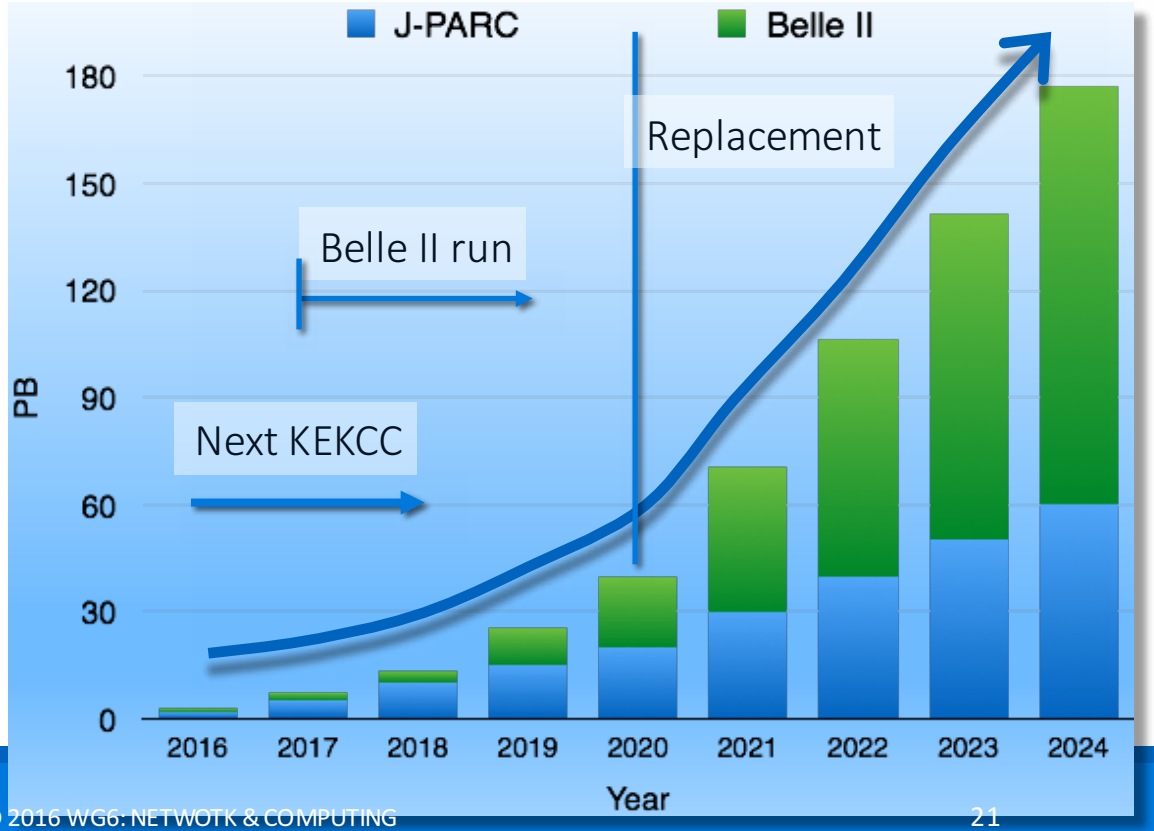
DATA GROWTH EXPECTATION

- J-PARC will constantly produce data.
  - A few – 10 PB /year
- Data explosion is expected for Belle II.
  - Data growth rate after 2020 is very high.

Unexpected factors:

- It depends on economic situation.
- Budget of electricity cost for operating accelerators

Data Explosion



CONCERNS ON DATA MIGRATION

Our System will be replaced every 4-5 years.

- Expected amount of data migration
  - 6PB (2016), Nx10PB (2020), Nx100PB (2024)
- Migration issues will be critical.

Requirements for migration of storage system

- Minimized downtime
- Safe data transfer (checksum)
- Continuous media migration
  - TS1140 : JC/JB -> TS1150 : JD/JC
- Problem with R/O time of small files
  - pining data might be a solution?

Technical issues :

- Users want to manage checksum information for safe data preservation.
- Tape-order recall (RAO) is desired for reading data efficiently.

WORKLOAD MANAGEMENT AND CLOUD SERVICE

Workload management for different groups (DC point of view)

- Requirements on specific system
  - experiments, groups, community
    - e.g. migration to SL6 in Grid service, but Belle I wants to stick to SL5.
  - test for newer OS
- Efficient resource management (servers on demand)

IaaS/PaaS-type of service (internal cloud)

- Middleware choice
  - PCMAE + Platform Dynamic Cluster : coherence with LSF
  - OpenStack-based products
- Provisioning tools
  - KVM (VM), xCAT (baremetal), Docker (future)
- Virtualization technology, not yet enough…
  - Vitual machine (KVM) : CPU virtualization (MC) is ok, but I/O virtualization is not yet enough.
  - Container (Docker, LXC) + Resource management (cgroups)
  - Coherence with JOB scheduler (LSF, UGE)

External cloud service

- Amazon EC2 is tested with Dirac for Belle II MC campaign.

SUMMARY

- Next KEKCC system will start in September 2016.
- Increase computing resources based on requirements of experimantal groups.
  - CPU : 10K cores (x2.5), Disk : 13PB (x1.8), Tape : 70PB (x4.3)
- 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.
  - Belle II 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 migtation as a potential concern