Data preservation of BESIII/IHEP

Lu WANG, Gang CHEN, IHEP/CAS

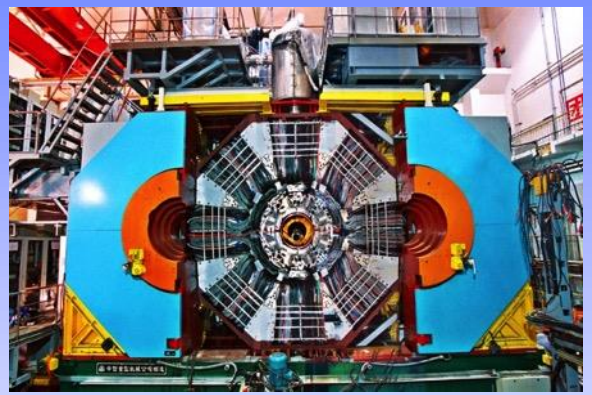
DPHEP Collaboration Workshop

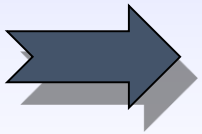
June 09, 2015

Introduction of BESIII Computing Environment

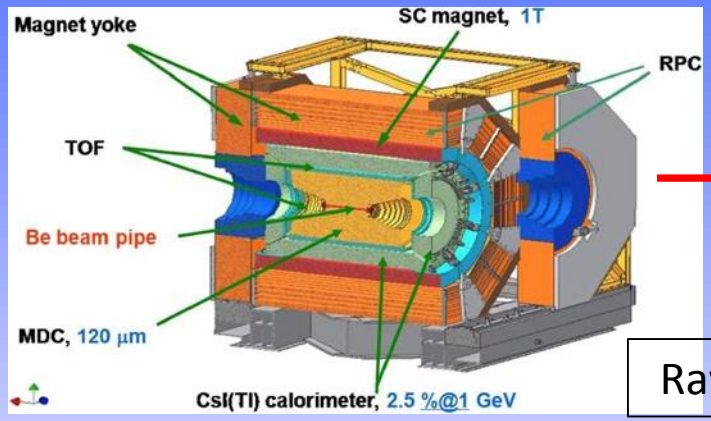
BECPII/BESIII

- BEPC: Beijing Electron Positron Collider
  - Started in 1989, upgraded to BEPCII since 2004
  - Dual-Ring, 2~5GeV/C
  - Luminosity (3~10)$\times 10^{32}$ $cm^{-2}s^{-1}$
- BESIII: Beijing Spectrometer
  - Upgraded to BESIII with BEPCII
  - Started to collect data in May 2009
  - End of data acquisition: 2022 (likely to extend)





Data Analysis Model



Raw data on disk
(All, Bhabha, Dimu., Random trigger ……) .raw

Raw data on tapes

Detector Calibration

Detector Simulation

.rtraw, ROOT

Offline Reconstruction (DST Production)

.dst, ROOT

Background Mixing
(with random trigger events)

.raw

Physics Skimming (nprong, DTag, …)

.dst, ROOT

Reconstruction

.dst, ROOT

User Analysis

Data Types

- Raw Data: delivered by DAQ for reconstruction
  - 15KB/event, in byte stream format
- MC-Raw Data: Data generated by simulation
  - 8 KB/event, ROOT format
- Reconstructed Data (DST):
  - 5 KB/event, ROOT format
- Reconstructed. MC Data (DST):
  - 13 KB/event, ROOT format
- User Analytic Data (DST): tag data, skimmed data…
  - ROOT format

Data Volumes

	Current (2009-2014)	Increase Per Year	2020
Raw	600	100	700
MC-Raw	600	100	700
Rec.	600	100	700
Rec. MC	600	100	700
User Analytic Data	350	50	650
Total (TB)	2750 (TB)	450 (TB)	3450 (TB)

Software Framework: BOSS

- BESIII Offline Software System (BOSS), is an offline data processing software system, developed based on GAUDI framework
- External Libs: CERNLIB, ROOT, CLHEP, Geant4,...
- Developing language: C++, some Fortran, and Java for web applications
- Database: MySQL
- Configuration management tool: CMT
- Operation system : SLC6/64bit
- Complier: GCC 4.3.2

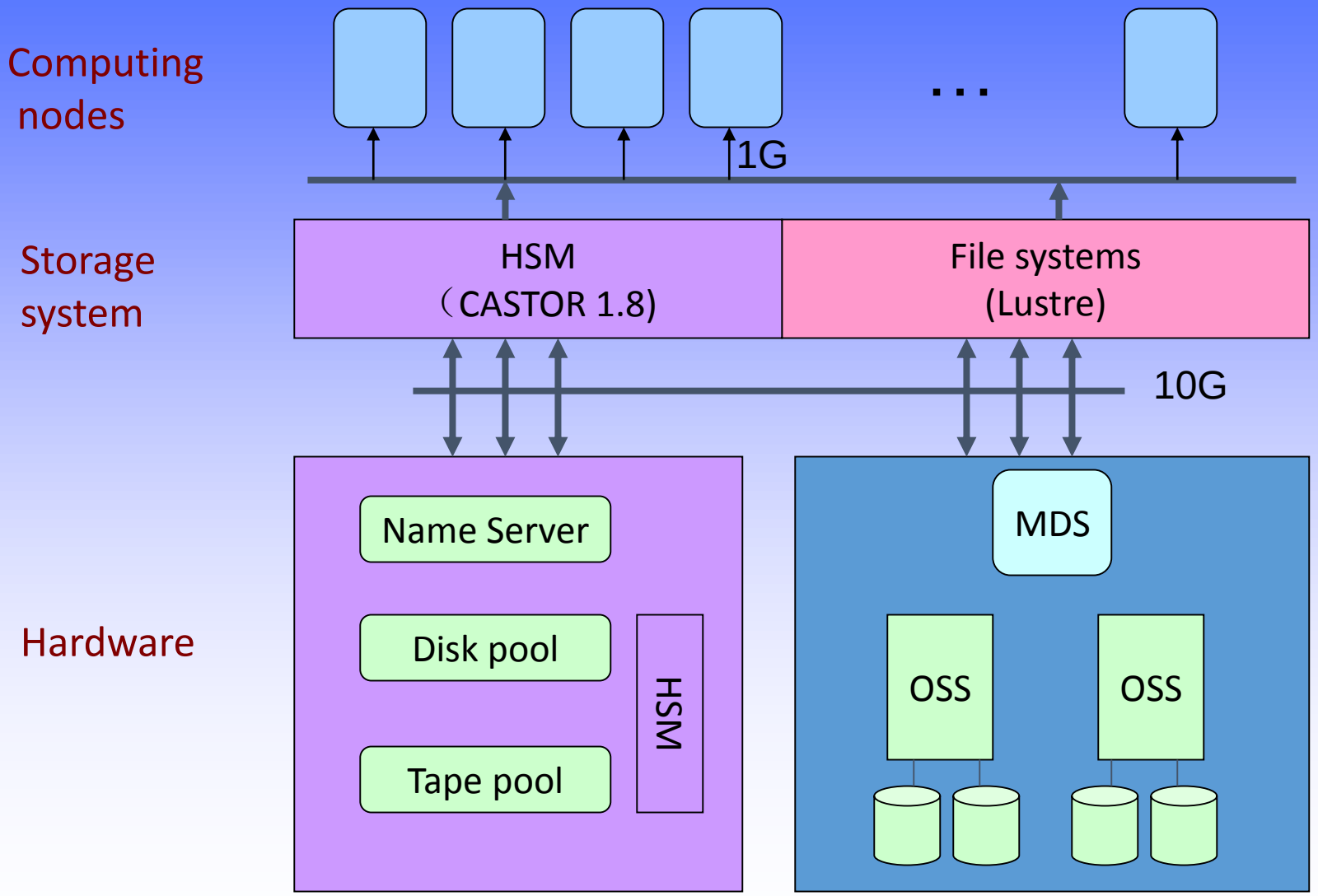
Computing Resources

- PC cluster: for reconstruction and physics analyses
  - 4260 cores, Xeon/3.0GHz, E5140, E5430, E5520,E5540
  - 1GB-2GB memory/core
  - IBM/HP/Dell blade systems with 10Gb uplink to Core Switch.
- GPGPU cluster: for partial wave analyses
  - 35 nodes (4 GPUs/nodes, 2 AMD HD 4870x2 cards)
  - 100 nodes (2 GPUs/nodes, 2 AMD HD 7950 cards)
  - Developed with OpenCL (firstly with Brook+) and C++, mainly for computing intensive job. ~100 times faster than CPU
- Special budget has been secured to upgrade hardware in the next years.

Computing Resources

- Distributed computing: integrate various resources from collaborations for MC production and analysis
- ~ 2000 cores (1GB-2GB memory/core), ~ 400 TB
- Resource types include cluster, cloud, grid
- The experiment software in distributed environment is stored and published through CVMFS
- The experiment data replicas in sites is managed by File metadata and replicas catalog with query functions provided
- The sqLitedb to get calibration data in sites is stored and published through CVMFS, synchronizing with central mysql offline database regularly

Storage System



HSM Storage

- Based on CASTOR v1.8
  - IBM3584 tape library, LTO 4
  - Stage system re-written
  - File reservation function added
  - Read performance
    - 60MB/s ->90MB/s per driver , 2.3 GB/s aggregated
  - Current capacity for BESIII
    - 2.7 PB, 2.2 PB used , 0.5 PB available
  - A remote replication of important raw data
    - ~ 900 cartridges, 700 TB

Disk Storage

- Based on Lustre distributed file system(2.5.3)
  - Infortrend/Dell disk array
  - HP G6/G7/G8 2U disk server
  - Automatic tuning of client cache
  - Process level I/O behavior traced
  - Fine-grained usage statistics provided
  - Read Performance for BESIII
    - 800MB/s per disk server, 20 GB/s aggregated
  - Current capacity for BESIII
    - 2.7 PB, 2.3 PB used , 0.4 PB available

Status of BESIII data preservation

Targets

- Follows the Model 4 of DPHEP recommendation：
  - Data about the experimental conditions and various parameters like calibration constants, detector geometry data etc.
  - Raw data and DST data should be conserved when the experiment system becomes stable,
  - Every stable version of BOSS,
  - Documents…
  - MC-Raw data will be deleted after reconstruction
- The experiment is expected to stop data taking at 2022, and Lifespan of preserved data is expected to be about 15 years after then.

Bit Preservation

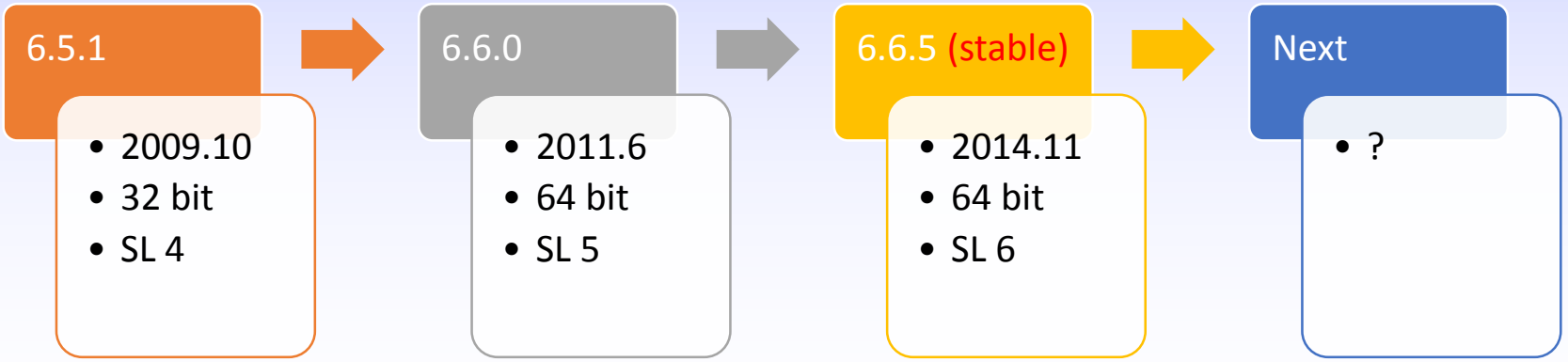
- Media: Tape
- Raw data/Random trigger data:
  - Flows from on-line farm to tape directly,
  - A copy on disk after reconstruction.
- Rec/DST/skimmed DST/tag data :
  - Replicated from disk to tape when a certain software version is stable.
- Condition/document databases:
  - Snapshots are copied to tape routinely

Bit Preservation

- Integrity check
  - A MD5 integrity check is done when data is copied from disk to tape
- Annual examination of tape library and LTO4 tapes
  - 200+ damaged tapes found during 2013 examination
  - Very few damaged tapes reported during 2014 examination
  - Considering biannual examination of tapes since the operation itself may introduce damages

Software Preservation

- We are lucky that BOSS is a integrated software package which includes all the blocks required in BESIII data processing.
  - Simulation, reconstruction, calibration, analysis…
  - functions could be preserved all together in a package
- However, it evolves quickly according to the upgrade of hardware, OS and requirements of physicists.

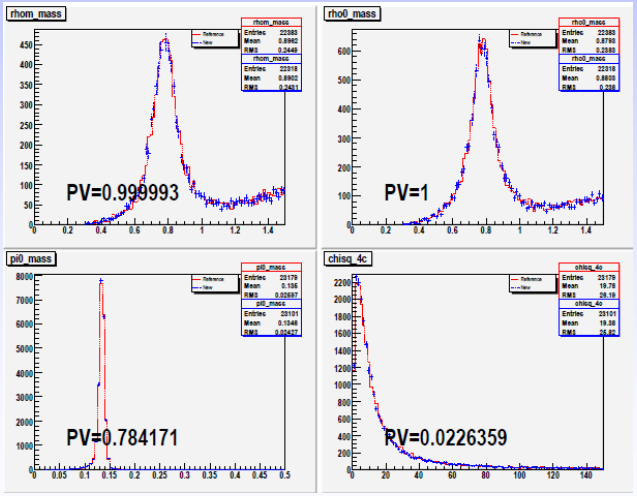


Software Preservation

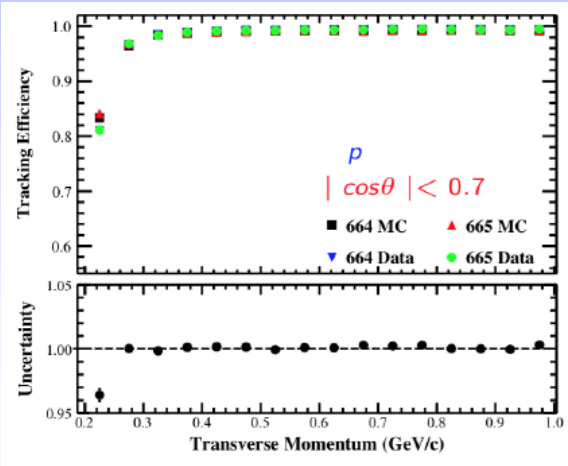
- For an old but stable version of BOSS, we preserve following items:
  - A complete package of software,
  - A runnable virtual machine image
  - The puppet template and RPM repository from which a runnable OS is created,
  - Release documents, book keeping parameters…
- A functional validation is done according to the standard process of software release.

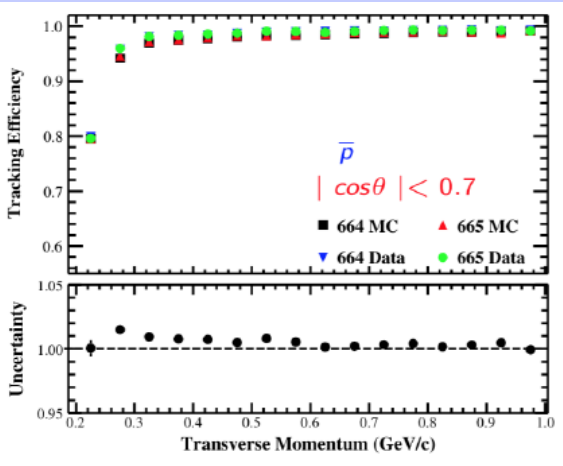
Validation Process

- MC data samples are generated for MC validation
- Related data samples are reconstructed to check the consistency between MC and real data



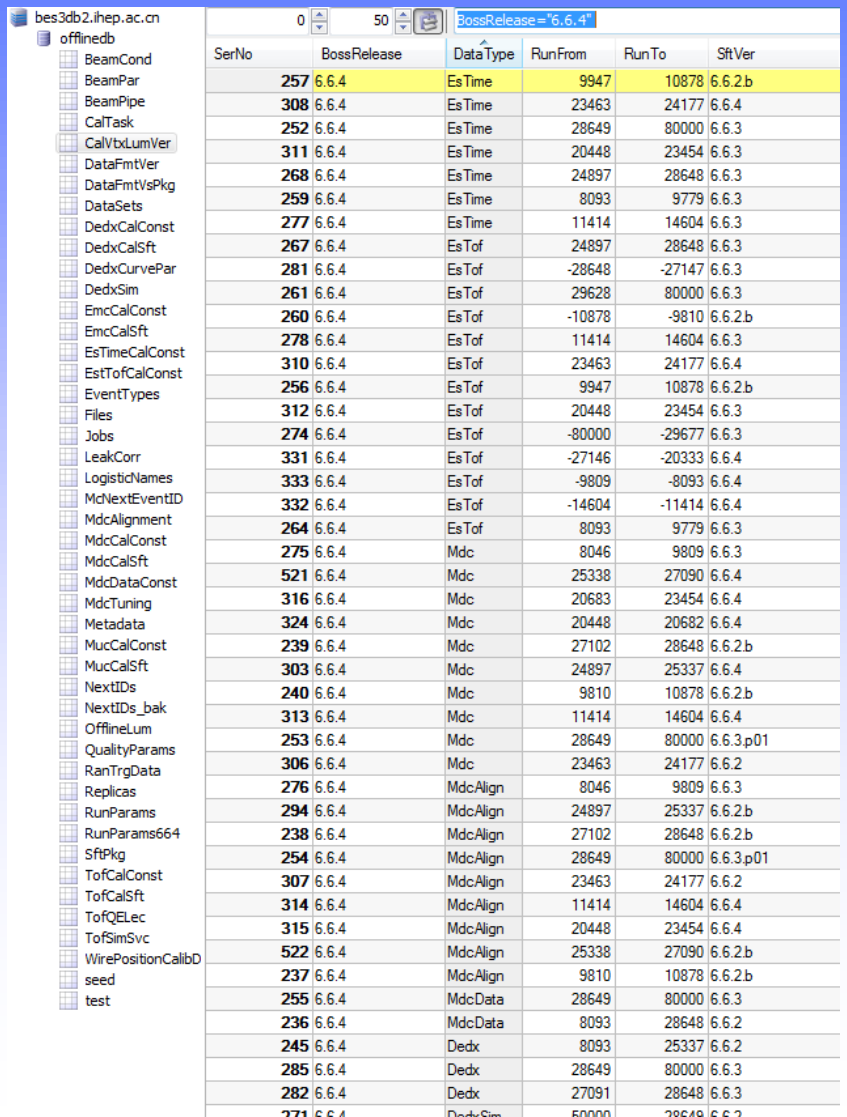
Jpsi to Rhopi





Tracking efficiency between BOSS 6.6.5 and 6.6.4

Bookkeeping for Calibration Constants



SerNo	BossRelease	DataType	RunFrom	RunTo	SftVer
257	6.6.4	EsTime	9947	10878	6.6.2.b
308	6.6.4	EsTime	23463	24177	6.6.4
252	6.6.4	EsTime	28649	80000	6.6.3
311	6.6.4	EsTime	20448	23454	6.6.3
268	6.6.4	EsTime	24897	28648	6.6.3
259	6.6.4	EsTime	8093	9779	6.6.3
277	6.6.4	EsTime	11414	14604	6.6.3
267	6.6.4	EsTof	24897	28648	6.6.3
281	6.6.4	EsTof	-28648	-27147	6.6.3
261	6.6.4	EsTof	29628	80000	6.6.3
260	6.6.4	EsTof	-10878	-9810	6.6.2.b
278	6.6.4	EsTof	11414	14604	6.6.3
310	6.6.4	EsTof	23463	24177	6.6.4
256	6.6.4	EsTof	9947	10878	6.6.2.b
312	6.6.4	EsTof	20448	23454	6.6.3
274	6.6.4	EsTof	-80000	-29677	6.6.3
331	6.6.4	EsTof	-27146	-20333	6.6.4
333	6.6.4	EsTof	-9809	-8093	6.6.4
332	6.6.4	EsTof	-14604	-11414	6.6.4
264	6.6.4	EsTof	8093	9779	6.6.3
275	6.6.4	Mdc	8046	9809	6.6.3
521	6.6.4	Mdc	25338	27090	6.6.4
316	6.6.4	Mdc	20683	23454	6.6.4
324	6.6.4	Mdc	20448	20682	6.6.4
239	6.6.4	Mdc	27102	28648	6.6.2.b
303	6.6.4	Mdc	24897	25337	6.6.4
240	6.6.4	Mdc	9810	10878	6.6.2.b
313	6.6.4	Mdc	11414	14604	6.6.4
253	6.6.4	Mdc	28649	80000	6.6.3.p01
306	6.6.4	Mdc	23463	24177	6.6.2
276	6.6.4	MdcAlign	8046	9809	6.6.3
294	6.6.4	MdcAlign	24897	25337	6.6.2.b
238	6.6.4	MdcAlign	27102	28648	6.6.2.b
254	6.6.4	MdcAlign	28649	80000	6.6.3.p01
307	6.6.4	MdcAlign	23463	24177	6.6.2
314	6.6.4	MdcAlign	11414	14604	6.6.4
315	6.6.4	MdcAlign	20448	23454	6.6.4
522	6.6.4	MdcAlign	25338	27090	6.6.2.b
237	6.6.4	MdcAlign	9810	10878	6.6.2.b
255	6.6.4	MdcData	28649	80000	6.6.3
236	6.6.4	MdcData	8093	28648	6.6.2
245	6.6.4	Dedx	8093	25337	6.6.2
285	6.6.4	Dedx	28649	80000	6.6.3
282	6.6.4	Dedx	27091	28648	6.6.3

- Calibration constants and tuning parameters of each sub-detector are recorded according to run numbers and for each BOSS Release, to make sure the production of simulated data and reconstruction are reproducible

Document Preservation

- A non-trivial work which involves different systems of the experiment.
- A serials of community shared software have been leveraged.
  - DocDB: paper, technical notes, minutes…
  - Hypernews: notifications of software release, paper publishment …
  - Indico: Conference slides,
  - Inspire: published paper

Budget and FTEs

- Since the experiment are still working, budget and FTEs are shared with the operation of computing center.
- Not just meaningful to HEP, but also attracts interests from other communities of CAS
- No funding scheme was defined, but support from CAS will be pursued.

Other Experiments

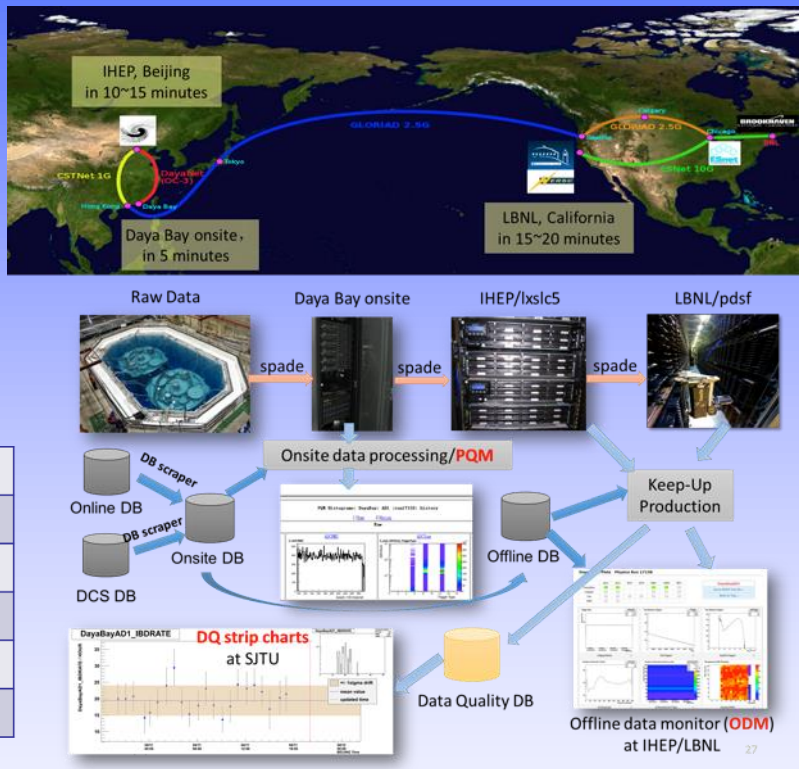
Daya Bay Reactor Neutrino Experiment

- To measure the mixing angle θ_{13}
  provided the most precision measurement of θ_{13}
- 6 reactor cores, 17.4 GW
- Relative measurement
  - 2 near sites, 1 far site
- Multiple detector modules
- Good cosmic shielding
- Produces ~200TB data/year
- will continue taking data

till the end of 2017

Daya Bay Reactor Neutrino Experiment

- Distributed computing system
  - Daya Bay, IHEP, LBNL
- Computing and storage
  - Roughly 1/3 of BESIII capacity
  - Same architecture at IHEP

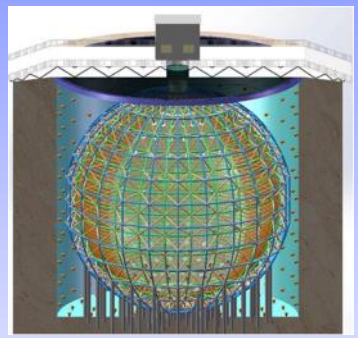


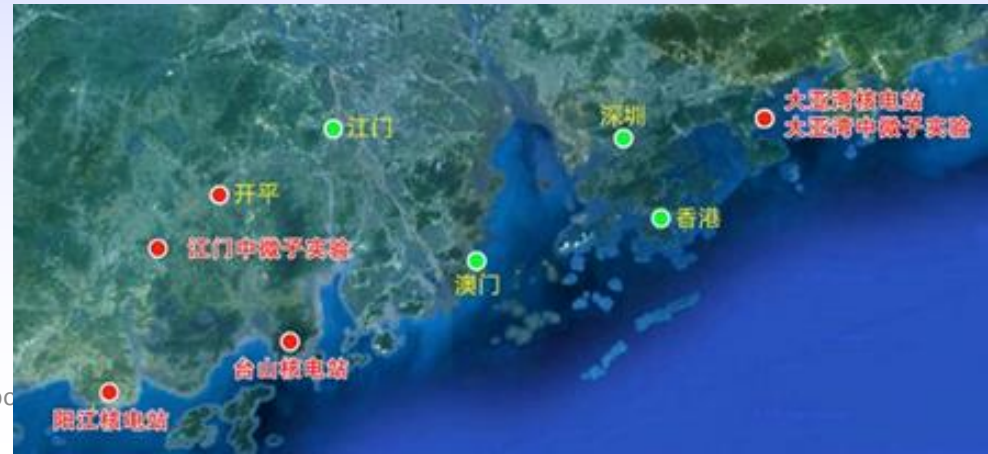
	IHEP	LBNL
Disk space	750TB	963TB
Harddisk file system	Lustre	GPFS
Tape system	CASTOR	HPSS
Accumulated raw data	320TB (disk), 640TB (tape)	320TB (disk), 640TB (tape)
CPU (cores)	~800	371 dedicated + additional shared

- Software framework (NuWa): Neutrino at Daya Wan (Bay)
  - Adoption of LHCb/ATLAS Gaudi framework
  - The system for simulation, reconstruction and analysis

JUNO

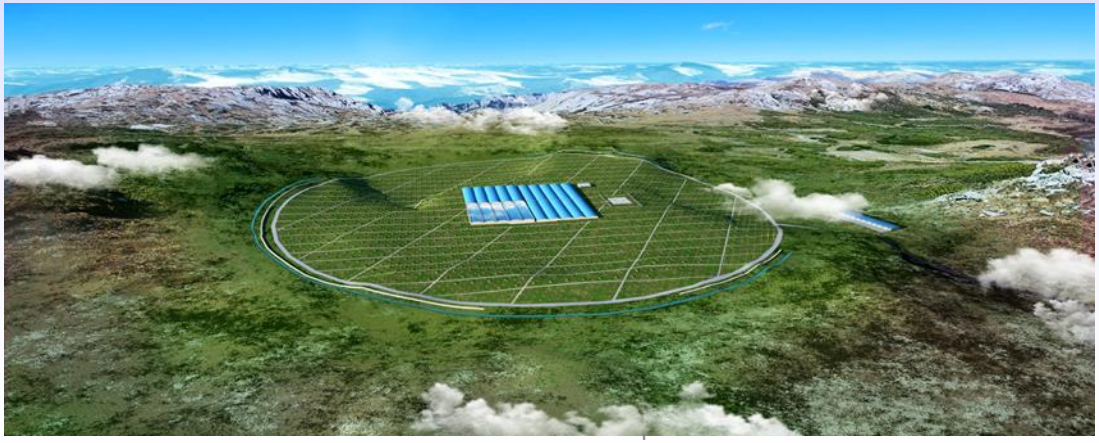
- Jiangmen Underground Neutrino Observatory
- Started to build in Jan. of 2015, operational in 2019
- 20 kt LS detector
- 2-3% energy resolution
- Rich physics opportunities
- Estimated to produce 2PB data/year for 10 years
  - Larger than BESIII





LHAASO

- Large High Altitude Air Shower Observatory, located on the border of Sichuan and Yunnan Province
- multipurpose project with a complex detector array for high energy gamma ray and cosmic ray detection
- Expected to be operational in 2019
  - ~1.2PB data/year * 10 Years, larger than BESIII



DP of other experiments

- Expected to generate larger data volumes than BESIII, and that means more intensive task of DPHEP.
- Experience of BESIII data preservation could be reused into these experiments
  - Similar software framework, computing architecture
  - Shared software developers and system operators
- Actually, DP of Daya Bay data has started this year following the same procedure with BESIII
- For those future experiments, it is always better to consider DP earlier!

Conclusion

- Considerations and Works on BESIII data (documents, data, software and execution environment) preservation has been initiated.
- As a running experiment, the perseveration work is synchronized with usages of users, upgrades of software and operations of the cluster.
- We have learnt a lot from peer experiments in DPHEP group.
  - Technical level and strategy level

Thank you!