# Data oriented job submission scheme for the PHENIX user analysis in CCJ

Tomoaki Nakamura, Hideto En'yo,
Takashi Ichihara, Yasushi Watanabe
and Satoshi Yokkaichi

*RIKEN Nishina Center for Accelerator-Based Science*

# PHENIX Computing Center in Japan (CCJ)

**CCJ history:**

- Developed since 1998 as regional analysis center for PHENIX and other RIKEN related experiments.
- Construction was started at 1999.
- Operation was started from 2000.
- 23 publications and ~30 doctoral dissertations have been completed by using CCJ resources.

**PHENIX:**

- DST production (polarized $p+p$).
- User level data analysis ($p+p$, A+A).
- Detector analysis, calibrations.
- MC simulations.
- Detector R&D.

**Other experiments:**

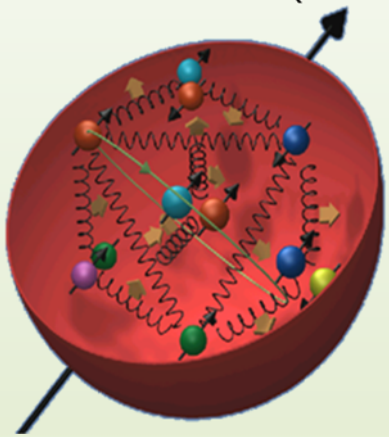
- KEK-PS E325, E559, Belle.
- J-PARC, E16, E26, g-2.

# PHENIX experiment at RHIC

## Spin physics by polarized proton beam

Cross Section and Parity Violating Spin Asymmetries of $W^{\pm}$ Boson Production in Polarized $p+p$ Collisions at $\sqrt{s} = 500$ GeV

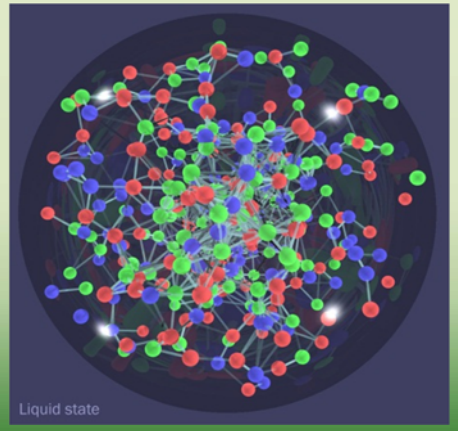
arXiv:1009.0505 (2010)

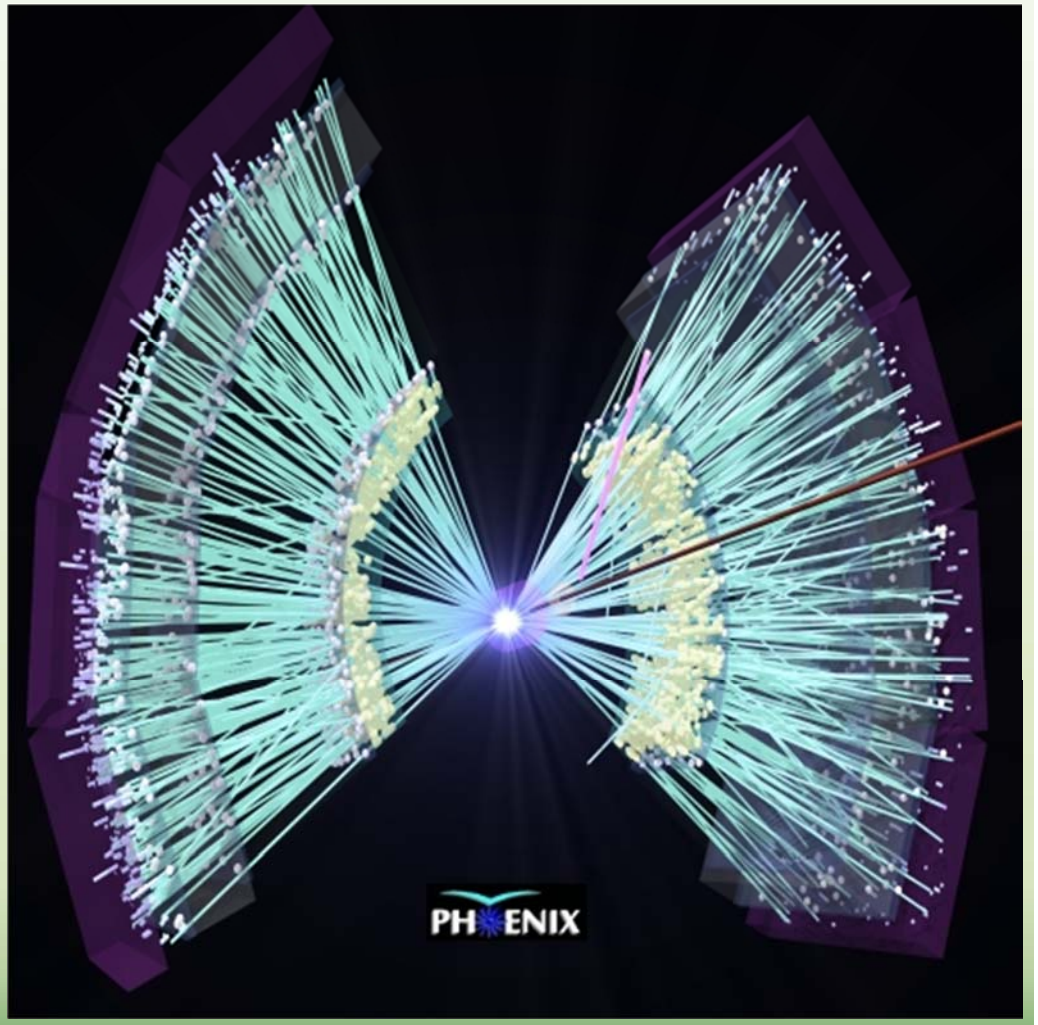


## QGP physics by heavy-ion beam

Enhanced Production of Direct Photons in Au + Au Collisions at $\sqrt{s_{NN}} = 200$ GeV and Implications for the Initial Temperature

Phys. Rev. Lett. 104, 132301 (2010)



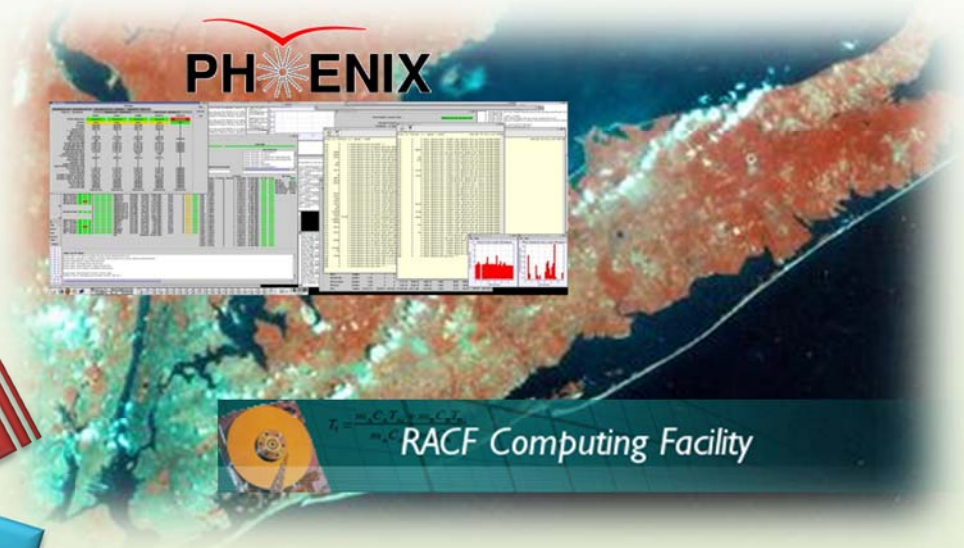




http://www.bnl.gov/bnlweb/pubaf/pr/newsroom.asp

# CCJ specification

## For the data production:

- Network:
  - ✓ 10Gbps from BNL to RIKEN (SINET3)
- GridFTP servers:
  - ✓ 4+2 servers (directory connected with buffer box in PHENIX counting room)
- Tape storage:
  - ✓ HPSS 7.1
  - ✓ TS3500 library 2.5PB (Maxim 8PB)




Raw or Reconstructed

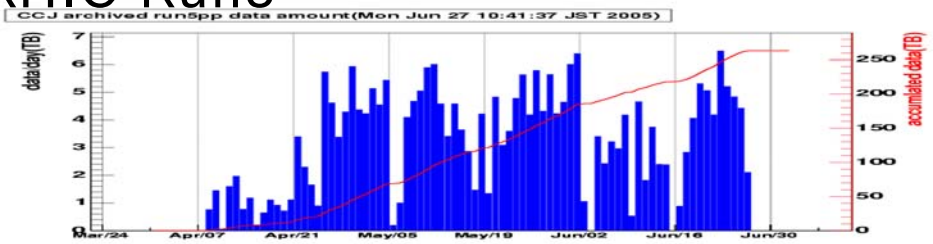
Reconstructed


## For the user analysis:

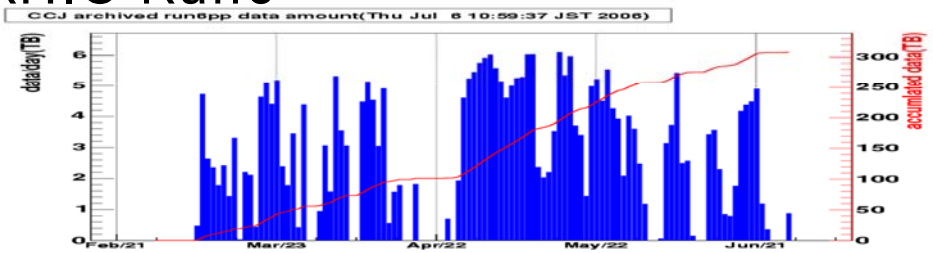
- PC cluster:
  - ✓ ~252core
  - ✓ 160core (Joint operation with RIKEN Integrated Cluster of Clusters)
    http://accc.riken.jp/ricc_e.html
- Disk storage:
  - ✓ ~320TB
- Analysis environments:
  - ✓ Offline library
  - ✓ Data bases
- Maintaining equivalent resources since the start of operation (2000).

# Data transfer record by GridFTP

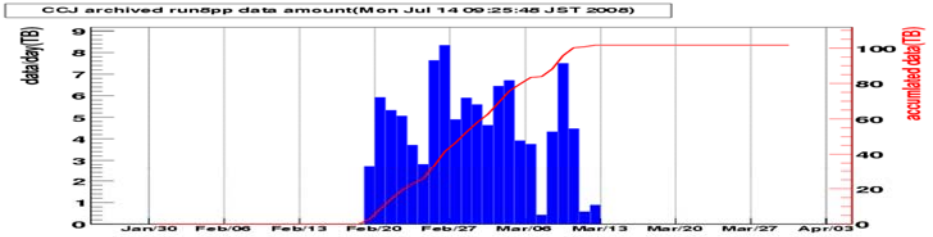
RHIC-Run5



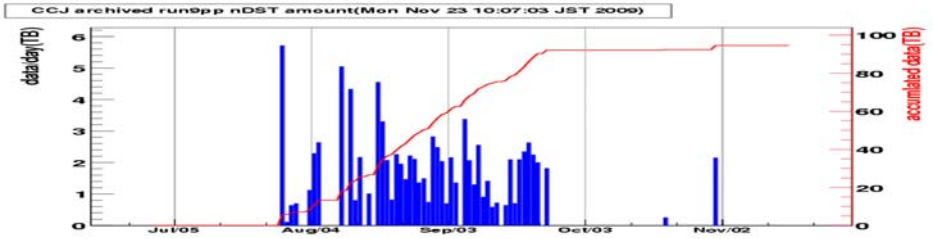

RHIC-Run6




RHIC-Run8




RHIC-Run9




**FY2005**
- Polarized *p*+*p* 200GeV
- Raw data
- Total 263TB

**FY2006**
- Polarized *p*+*p* 200GeV
- Raw data
- Total 308TB

**FY2008**
- Polarized *p*+*p* 200GeV
- Raw data
- Total 100TB

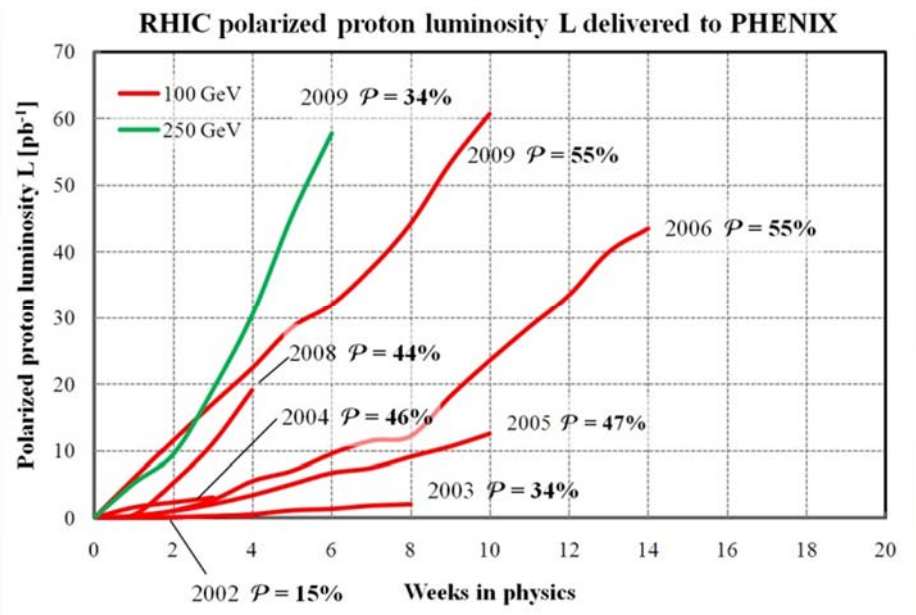
**FY2009**
- Polarized *p*+*p* 200/500GeV
- **RECONSTRUCTED DATA**
- Total 95TB

1.5PB of raw and reconstructed data was stored as of 2010.

# Effective utilization for the user analysis

## Growing data size

- Improvement of beam at RHIC
  - ✓ Luminosity
  - ✓ Polarization
- Detector upgrade:
  - ✓ ~400Kch + Silicon VTX 4.4Mch.
- DAQ upgrade:
  - ✓ 500 ~ 800MB/sec at RHIC-Run10.
  - ✓ Will be upgraded factor 2 or more.




http://www.bnl.gov/cad/

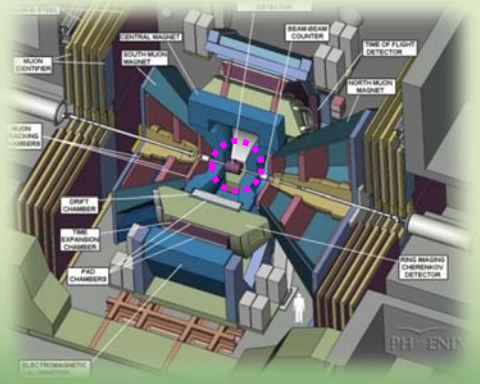
## For the user data analysis

- Data distribution, storage to many nodes:
  - ✓ Not easy to avoid the I/O bound problem by the conventional methods.
- NFS export:
  - ✓ Too slow due to the multiple access.
- Disk cash:
  - ✓ No way, when you scan all of the data set.

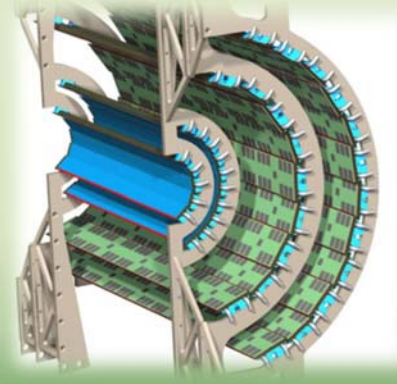
## MINIMIZATION OF I/O

Preparing special nodes with large capacity of local disk to store the read only data in advance.

PHENIX



Silicon VTX



http://www.rarf.riken.go.jp/lab/radiation/

# Disk configuration and I/O performance

## HP ProLiant DL180 G5

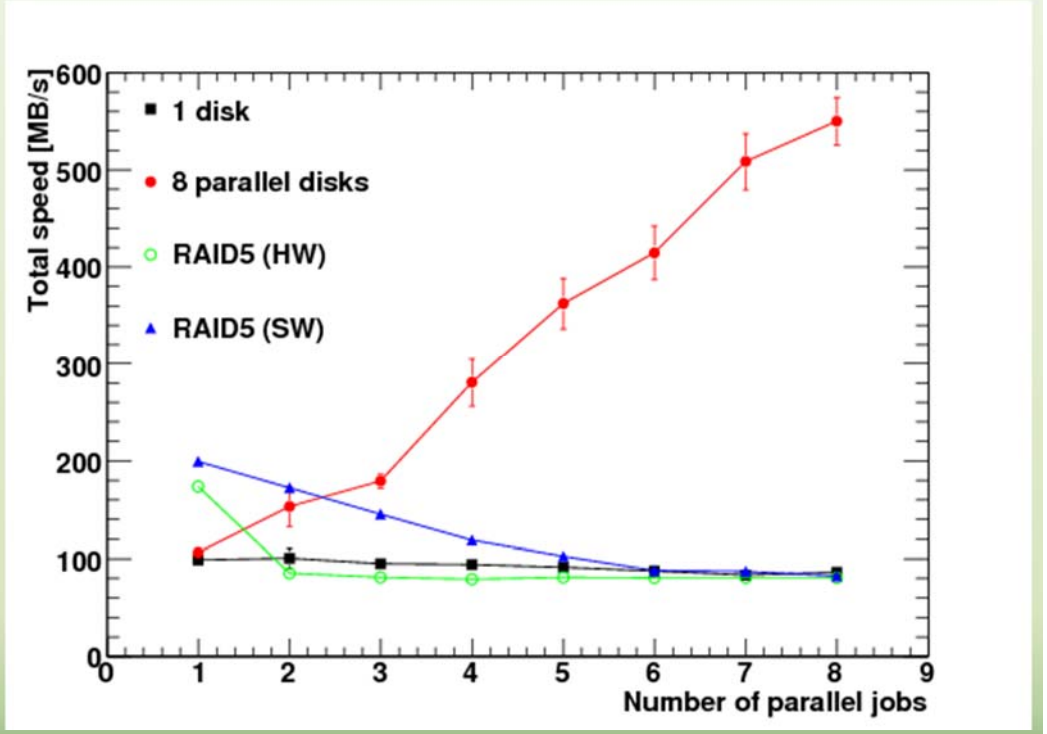
- Chassis: 2U, 12HDD-bay
- CPU: Quad-core × 2
  2.66GHz Xeon E5430
- Memory: 16GB (RDIM)
- HDD:
  146GB SAS × 2 for the system (RAID1)
  1TB SATA × 10 for the data storage
- NIC: 1Gbps
  Uplinked by 10Gbps switch in rack.

## System

- OS: Scientific Linux 5.3 x86_64
- File system: ext3
  XFS has no gain for the read only usage.

## HDD configuration

- RAID5 configuration:
  Advantage is only for 1 process.
- Good performance without RAID:
  Total ~550MB/sec at 8 parallel jobs.



# Data allocation

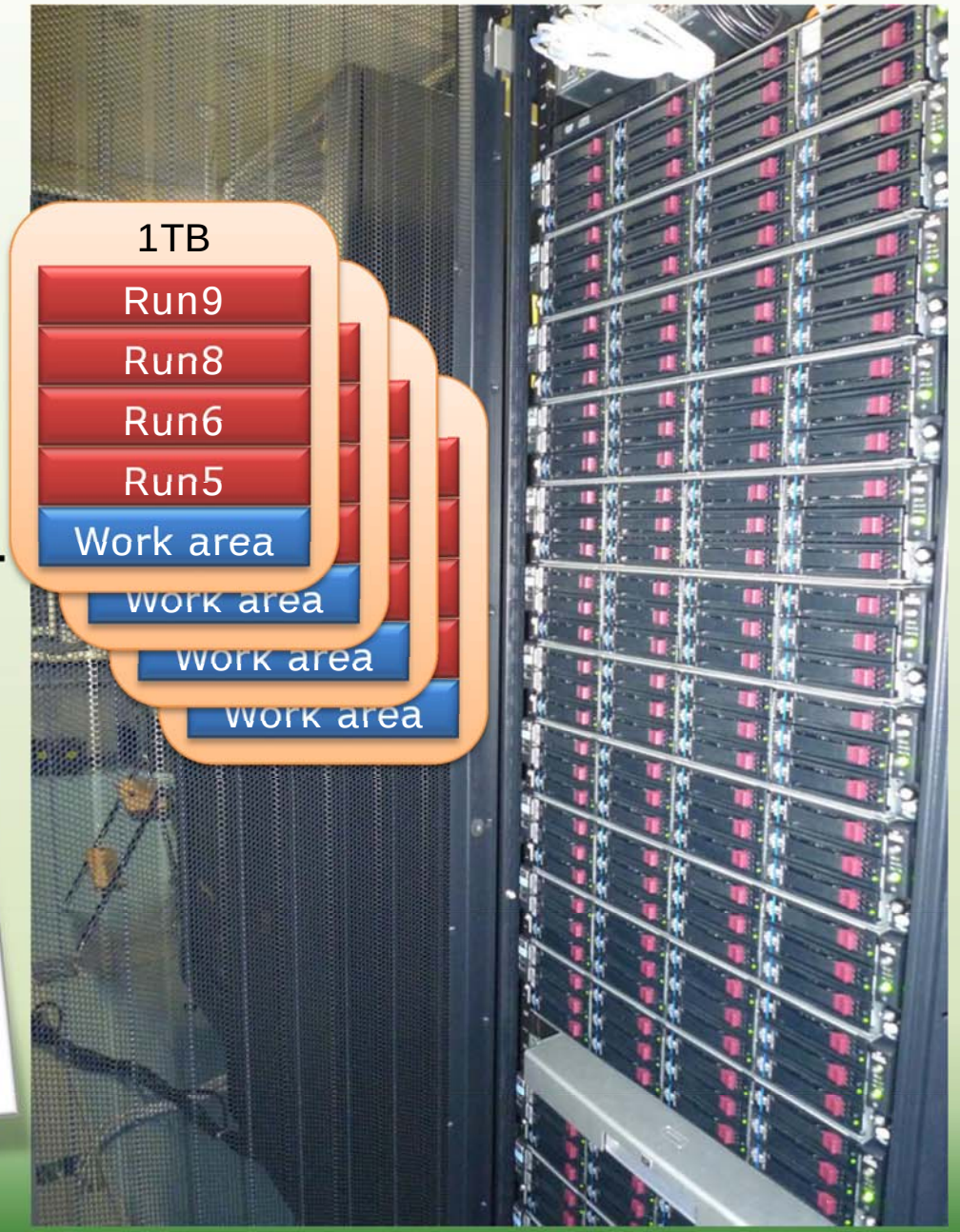
## PC Cluster

- Total 18 nodes
- Total 144 CPU cores
- Total 180TB for the data storage

## Data sets in HDD

- Every HDD has a part of each data set segmented by run number.
- User jobs can use all of CPU cores without time loss at data distribution.
- ~40GB work area is allocated in each HDD for the non data analysis jobs e.g. simulation.

| Dataset | nDST type | Data amount |
|---|---|---|
| Run 9 $p+p$ 200 GeV | All type | 65.4 TB |
| Run 9 $p+p$ 500 GeV | All type | 31.2 TB |
| Run 8 $p+p$ 200 GeV | All type | 21.2 TB |
| Run 6 $p+p$ 200 GeV | w/o detector | 14.6 TB |
| Run 5 $p+p$ 200 GeV | w/o detector | 9.9 TB |



# Job submission scheme

**Job submission script**

- Submit the user job to the appropriate node automatically, which has the required data.

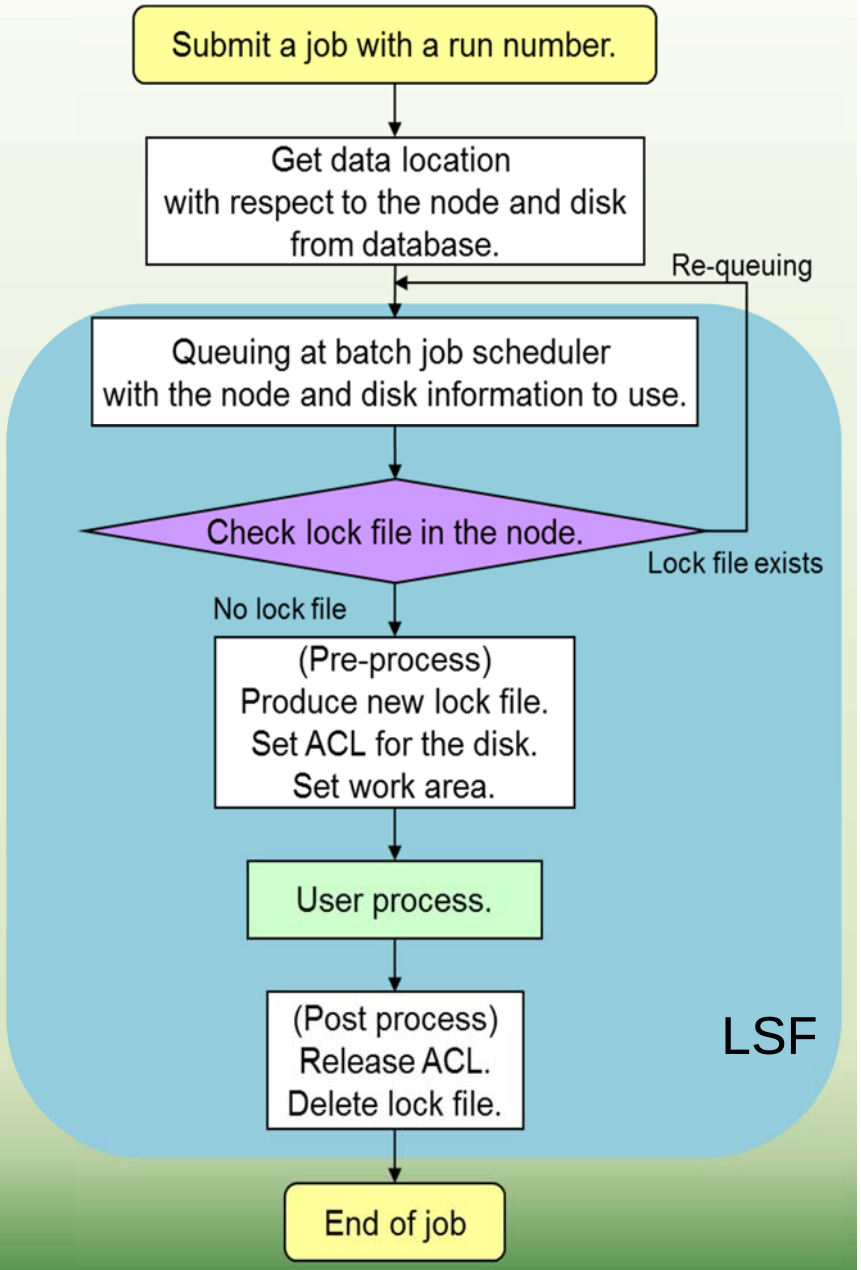
**Batch Job scheduler (LSF 7.0)**

- Fair-share among users.
- Set the path to the data directory and work area for environmental variables in the job.

**Lock file and directory ACL**

- Create at the pre-process for the exclusive use of a physical disk.
- Second job, which will use the data in occupied disk by the first job, stay in queue until the end of the first job.

User got to be able to analyze all of the existing data in local HDD (150TB) within 9 hours.



# Summary

- 18 calculating nodes with 180 TB local disks were introduced for effectively analyzing huge amounts of PHENIX data.

- A data-oriented batch queuing system was developed as a wrapper of the LSF system to increase the total computing throughput. Indeed, the total throughput was improved by roughly 10 times as compared to that in the existing clusters; CPU power and I/O performance are increased threefold and tenfold, respectively. Thus, users can analyze data of 150TB within 9 hours.

- We have experienced just 2 times disk failure out of 180 disks in 1.5 years operation. Therefore, we could conclude that this method is highly effective for the I/O bound data analysis.

# Next step

Preparing the same set of cluster with same data set,

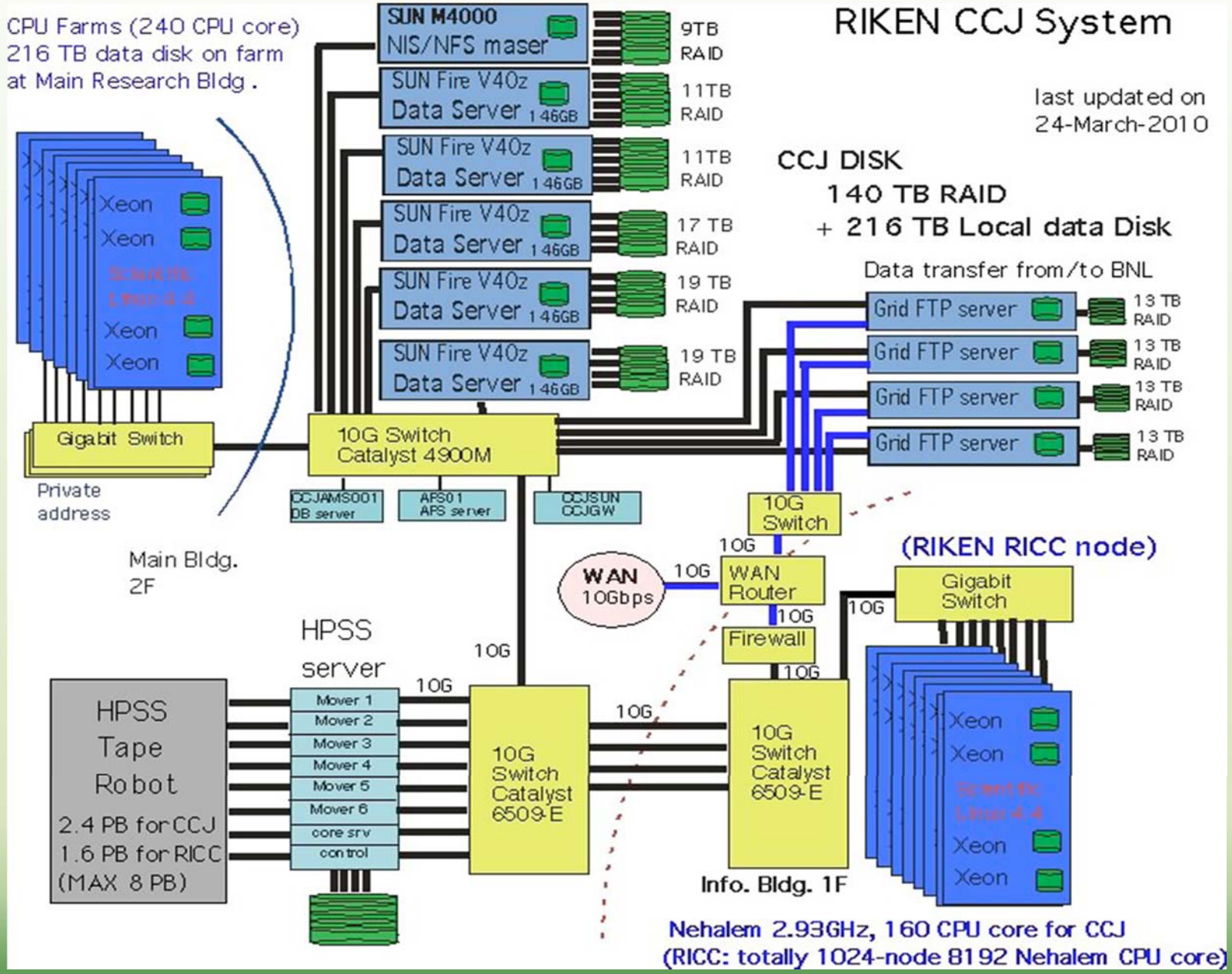
**High availability:**
- Reducing down time by the redundant data.
- Balance of job distribution with respect to any kind of jobs.

**Easy to expand for the growing data size:**
- Without any other special software.
- No special servers.

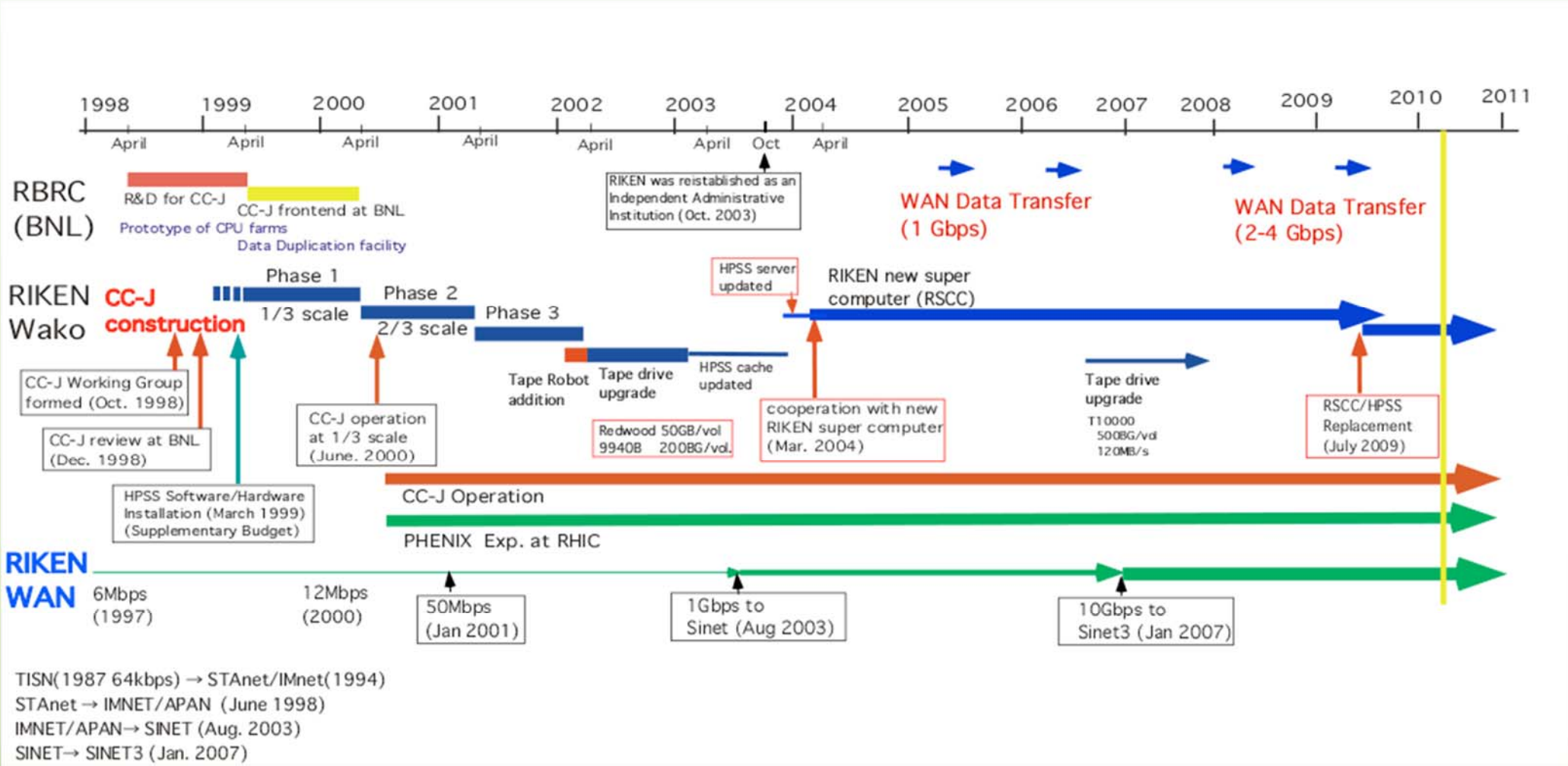
**By low cost:**
- 144CPU core, 2GB memory/core, 180TB disk for the local data storage.
- ~$150K as of 2009.
- The capacity of a HDD is growing and its price is going down steadily.

# Backup

# CCJ specification

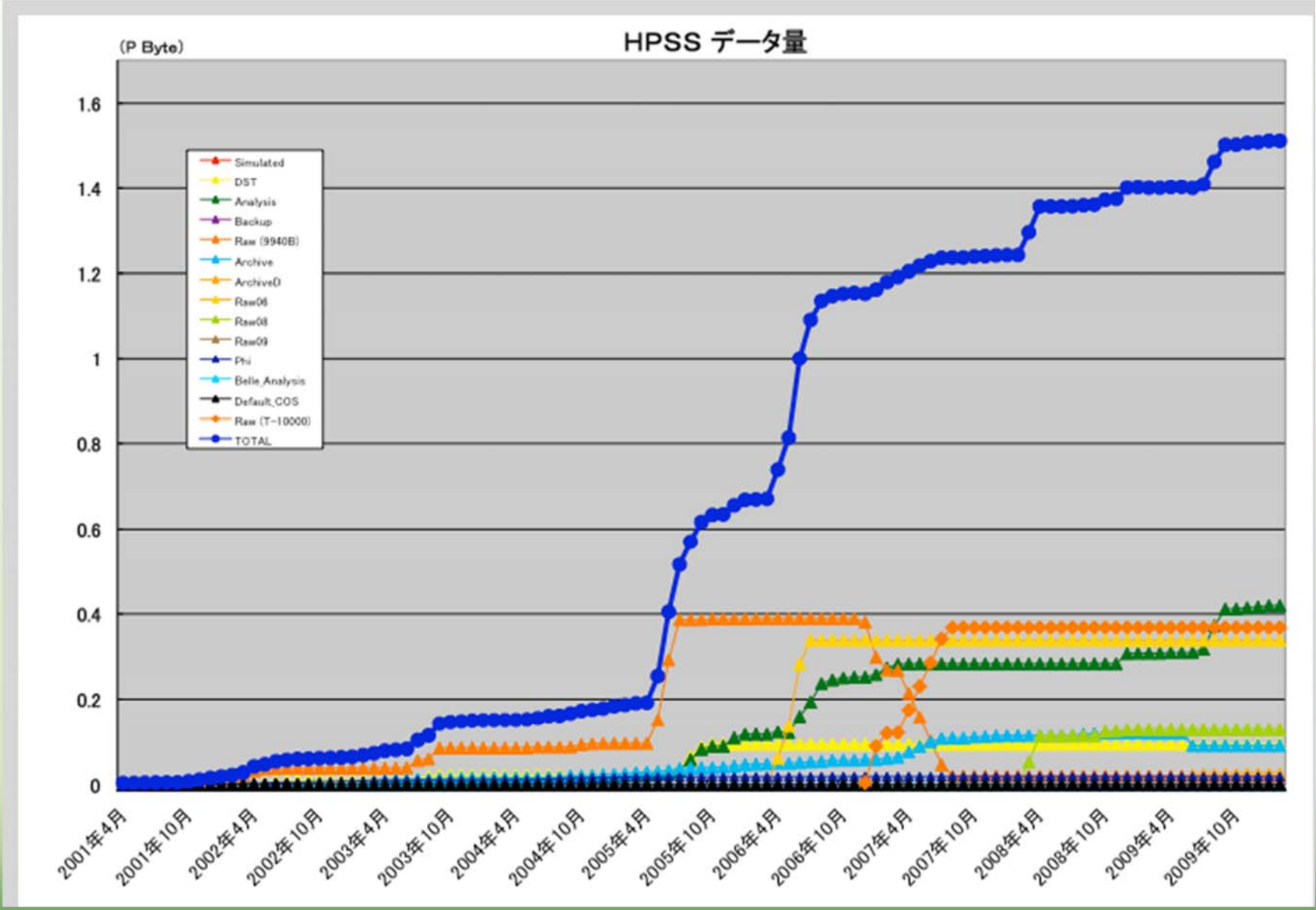


# CCJ history

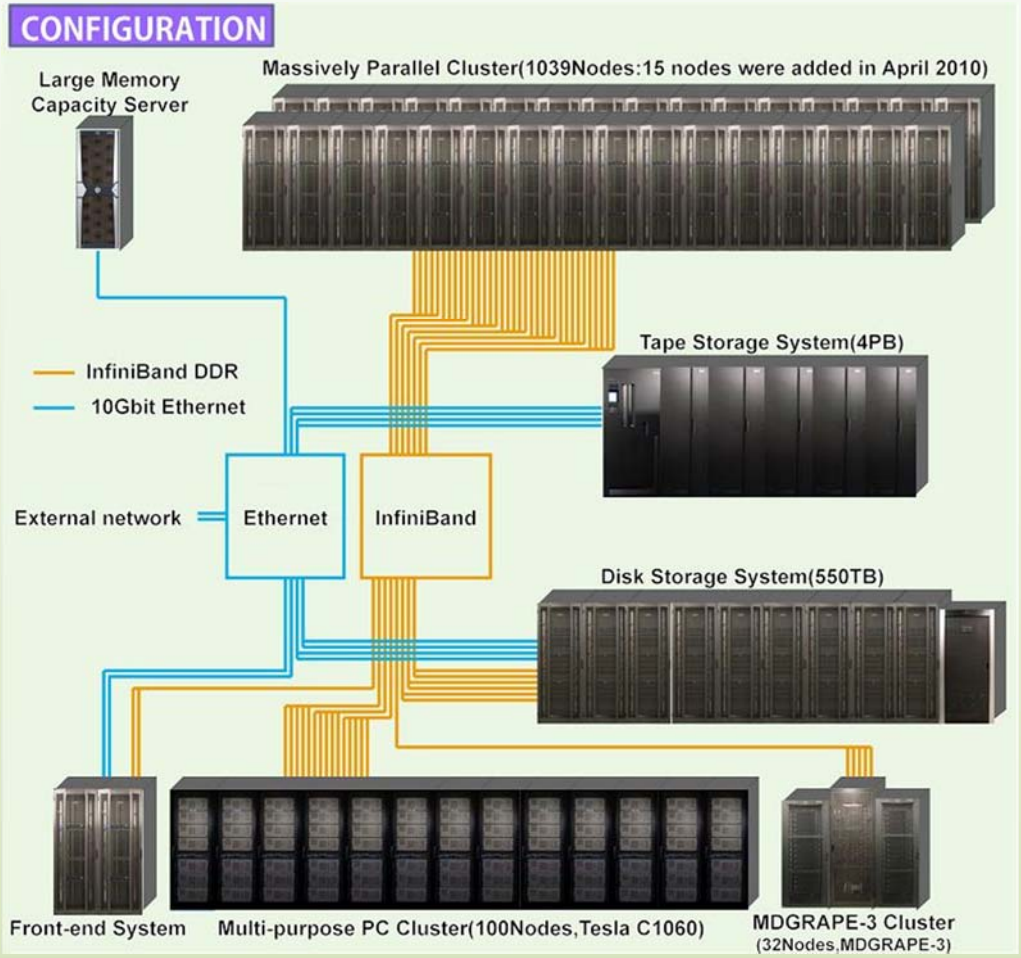



TISN(1987 64kbps) → STAnet/IMnet(1994)
STAnet → IMNET/APAN (June 1998)
IMNET/APAN→ SINET (Aug. 2003)
SINET→ SINET3 (Jan. 2007)

# HPSS data amount



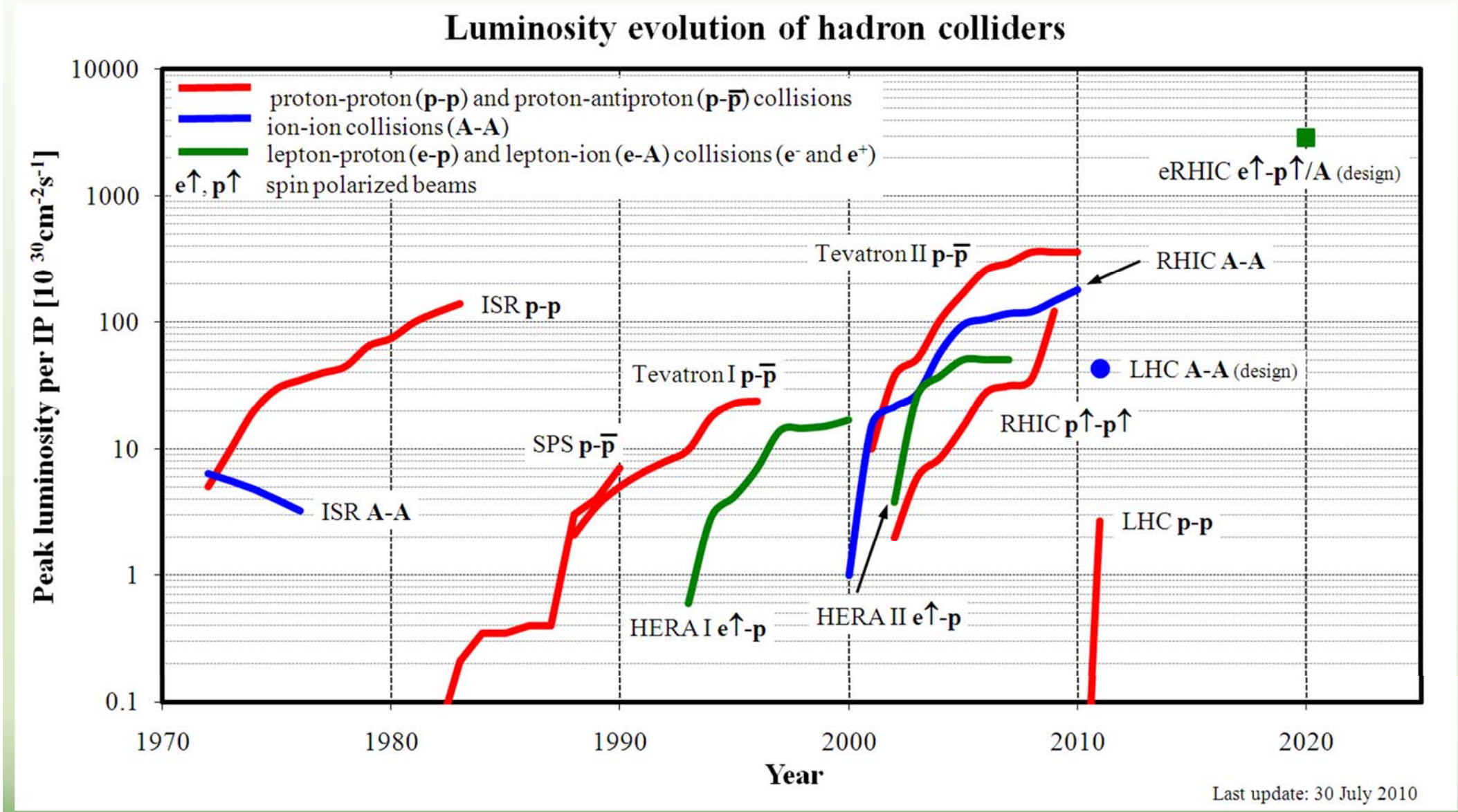
# RIKEN Integrated Cluster of Clusters




http://accc.riken.jp/ricc_e.html

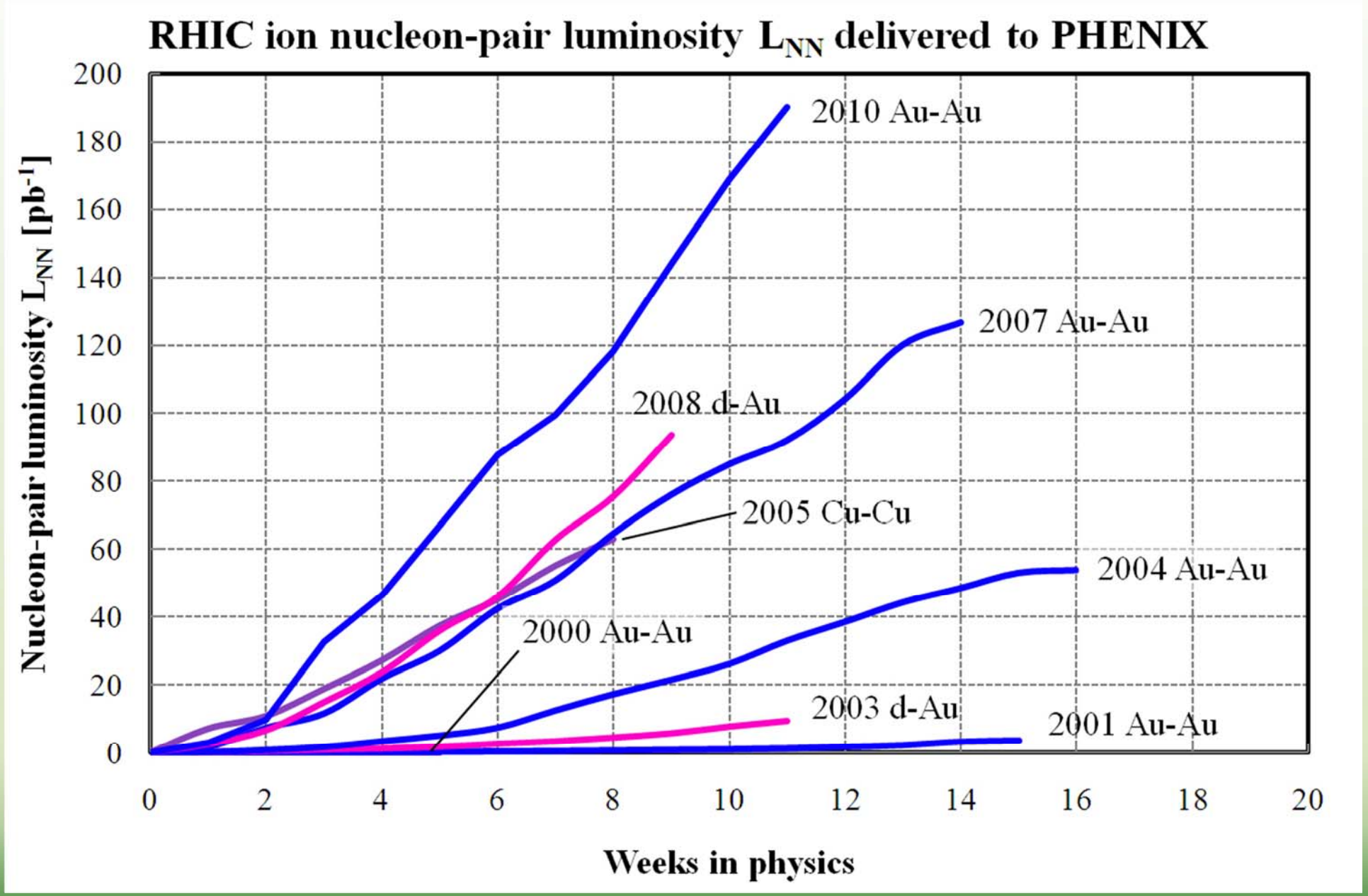
| System | Model | Specifications |
|---|---|---|
| Massively Parallel PC Cluster (97.4TFLOPS) | Fujitsu PRIMERGY RX200S5 (1024nodes) | CPU: Intel Xeon 5570 (2.93GHz) X 2<br>Memory: 12GB<br>HDD: 500GB (RAID0, SAS) |
| Multi-purpose Parallel Cluster (9.4+93.3TFLOPS(SP)) | NEC Express 5800/56Xg (100nodes) | CPU: Intel Xeon 5570 (2.93GHz) X 2<br>Memory: 24GB<br>HDD: 250GB<br>Accelerator: NVIDIA Tesla C1060 |
| MDGRAPE-3 Cluster (3.1+64TFLOPS) | SGI Altix XE250 (32nodes) | CPU: Intel 5472 (3.0GHz) X 2<br>Memory: 32GB<br>HDD: 750GB<br>Accelerator: MDGRAPE-3 |
| Large Memory Capacity Server (239GFLOPS) | SGI Altix 450 (1node) | CPU: Intel Itanium 9140M (1.66GHz) X 18<br>Memory: 512GB<br>HDD: 12TB<br>I/O: PCI-X (MDGRAPE-3 available) |
| Disk Storage System(550TB) | Filesystem: QFS+SRFS<br>File Server: SPARC Enterprise M9000<br>RAID: Eternas 2000 Model200 X 24 | |
| Tape Storage System(4~PB) | HSM: High Performance Storage System<br>Core Server: System p570 X 1<br>Mover: System p570 X 6<br>Cache: DS4800 X 6 (20TB)<br>Tape: TS1040 X 12 (LTO Ultrium4)<br>Library: TS3500(L53+D53+S54) | |
| Network | InfiniBand: X4 DDR | Switch: Qlogic SilverStorm 9024 X 60 and 9120 X 2<br>Topology: Fat-tree (bisection bandwidth: 240GB/s) |
| | Ethernet: 10GbE, GbE | Switch: Cisco 6509-E X 2 |

# Luminosity at hadron collider



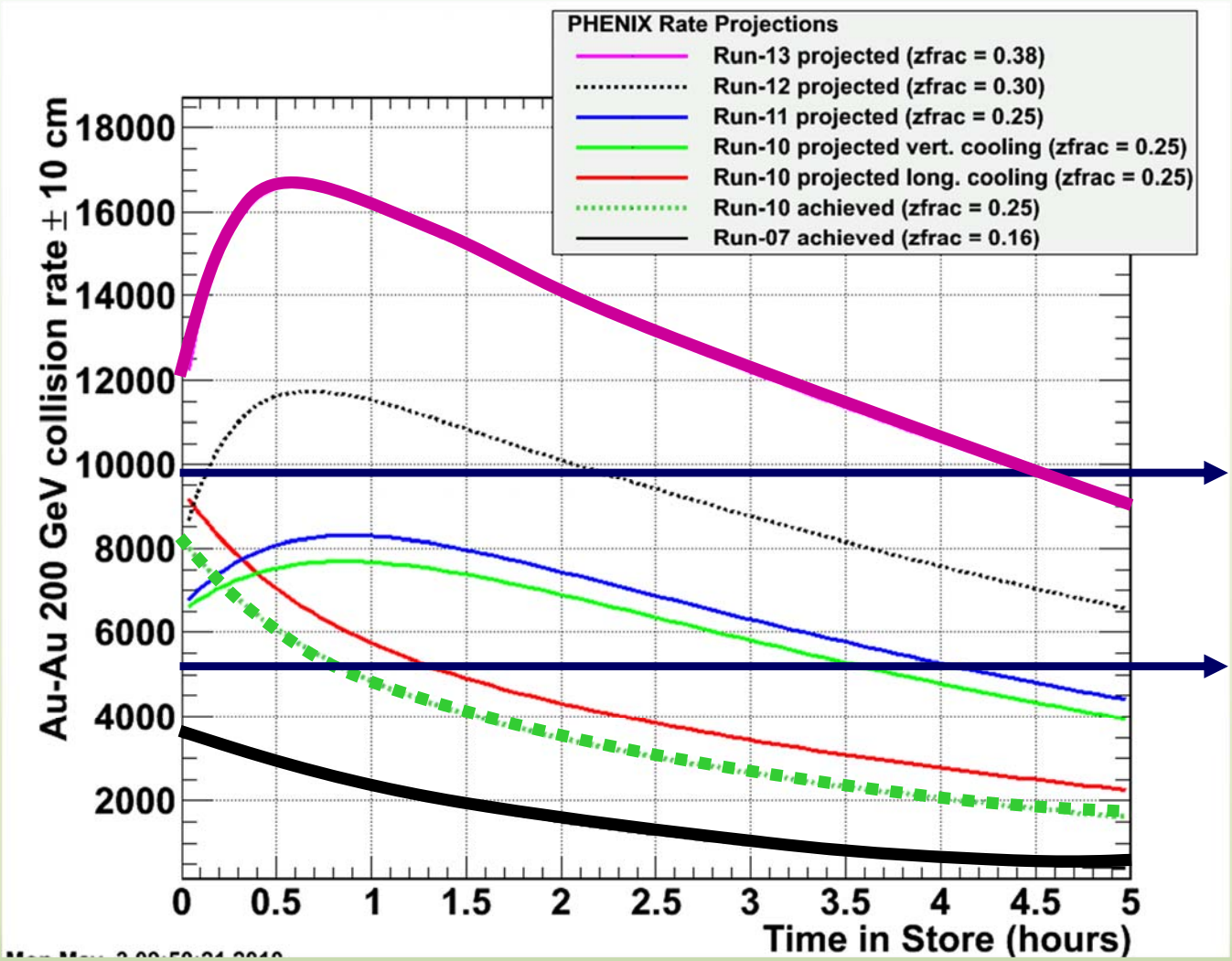

http://www.bnl.gov/cad/

# Integrated luminosity A+A at RHIC




http://www.bnl.gov/cad/

# PHENIX DAQ upgrade




SuperDAQ
(i.e. sDAQ)

DAQ2010

James Nagle: Next Decade Plans for the PHENIX Experiment
Workshop on Saturation, the Color Glass Condensate and Glasma