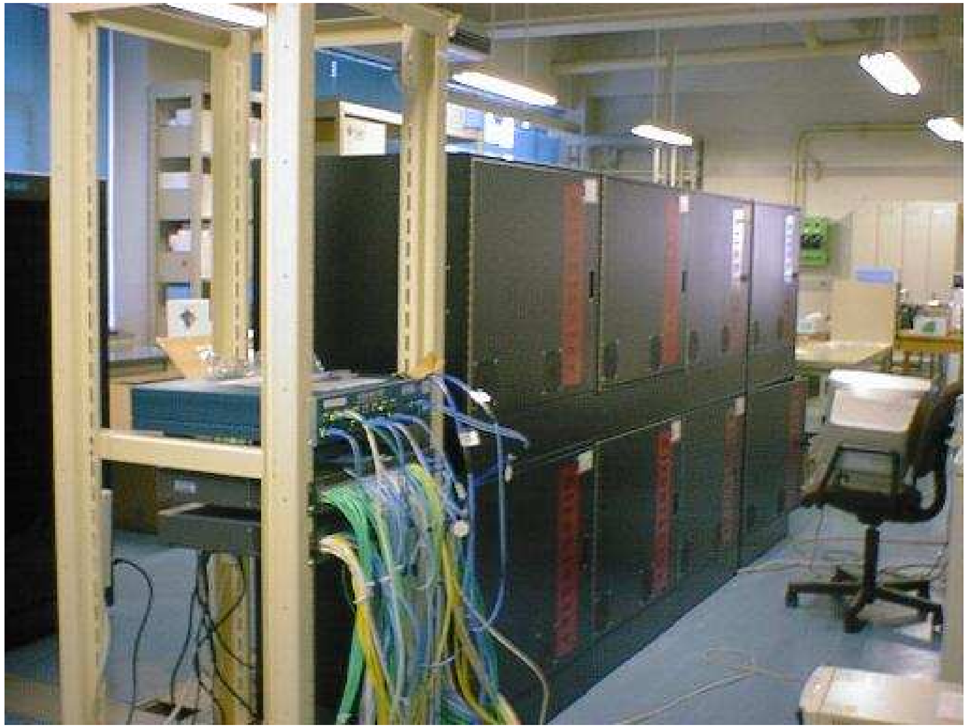
The RIKEN CCJ (computing center in Japan) for RHIC Physics

S. Yokkaichi, T. Ichihara, Y. Watanabe,
N. Hayashi, S. Sawada(1), Y. Goto(2), O. Jinnouchi,
H. Hamagaki(3), H. En'yo
RIKEN, KEK(1), RBRC(2), CNS(3)



CPU farm (Alta Cluster)

Tape Robot (STK PowderHorn)

What is CCJ

- Three purposes for PHENIX experiment
    - ...required 11470 SPECint95 (~ P III 1GHz x 244) in plan
  - Detector Simulation (30TB/year, 8000SPECint95)
  - 'Regional' analysis (calibration, DST−>microDST, etc.)
    - analysis by local users (Japanese and Asian)
  - Spin Physics analysis
- History of Construction and Operation
  - 1999/4 : construction starts at RIKEN (Wako, Japan)
    - grow to full size in 3 years
  - 2000/6 : open for collaboration
    - simulation projects for PHENIX
    - 2000/12 : DST production for PHENIX year−1 data
    - analyses and simulations by 'regional' users

Data Transfer Japan <–> BNL

- Data amount from PHENIX in plan
  - Raw data : 200TB/year (20MB/sec x 120 days)
    - year–1 data: 2.9TB / year–2 data taking was started 2001/8
  - DST (reconstructed data) : 100TB/year ?
- Airborne data transfer was started for year–2 data
  - SD–3 tape(50GB/tape) / RedWood drive (11.25MB/sec)
  - 120 tapes/week = 10MB/sec
- Wide Area Network between RIKEN and BNL/RCF

  RIKEN–IMnet (50Mbps) –APAN(70Mbps)–ESnet(155Mbps)–RCF
  - 2.2 MB/sec using 'scp' (25sessions/25processes)
  - 2.8 MB/sec using 'bbftp' (4sessions/44processes)
  - 1TB of year–1 raw data were transferred using network

CPU farm and work disks

- Linux cpu farm (Red Hat 6.1 / Alta Cluster)
  - 96node 192cpu (PIII 1GHz x 64 / 850MHz x 96 /700MHz x 32)
    - ~8100 SPECint 95 in total
    - memory 512MB + swap 1GB /cpu
  - more 96cpu will come in this year
- Login server and NFS server (Solaris2.6/SUN E450 x 3)
  - 9TB RAID disks in total and more 9TB will come
  - Fibre Channel RAID : ~40MB/sec I/O
- Gigabit Ethernet (alteon ACE Switch 180)
  - Jumbo frame (MTU 9KB): reduce cpu usage for transfer

HPSS (High Performance Storage System) 4.2

- hierarchical storage system supported by IBM
  - compatible with RCF (:RHIC Computing Facility)
- hardware at CCJ
  - IBM RS/6000 SP x 5, RedWood tape drive x 4
  - 4TB disk cache (FC RAID) / 100TB tape volume
  - tape robot : STK PowderHorn 9310
  - 50MB/sec achieved in write stress test from the Linux farm
- Tape duplication facility at BNL
  - IBM RS/6000 F50 x 1 + RedWood drive x 2
    - pftp from HPSS to local disk / tar from local to SD3 tape in RW drive
  - tape duplication with ~10MB/sec is achieved

    ...ready for PHENIX year–2 data

Software for operation

- AFS daily 'mirroring' for stable operation
  - PHENIX software is shared using AFS
  - copy (rsync) the AFS region to local disk
    - AFS connection between Japan and BNL is unstable
    - AFS client on Linux was unstable
  - also 'mirror' server for Hiroshima/CNS/Tsukuba/TIT
- LSF 4.1 : batch queue system
- VxFS (journaling file system) : free from fsck for ~TB disks
- RH6.1 (compatible with RCF) installation using 'KickStart'
- Objectivity/DB 6.1 for PHENIX calibration DB
- 'cftp' and 'data carousel' for HPSS access
- MRTG for network/cpu monitoring

Objectivity/DB access

- calibration DB in PHENIX
  - current size is ~600MB... to be grown in near future
- direct access from Japan to BNL via AMS
  - too slow to use with the PHENIX reconstruction software
- making a 'snap – shot ' copy of PHENIX –FDB
  - use '*oobackup*' to gather DB files, compress and transfer to CCJ the files and '*ooinstallfd*'
  - periodically transferred and already used for DST production at CCJ
- AMS server at CCJ : dual 1GHz Pentium III /Gigabit Ether
- Next Task : 'incremental' copy method for large size DB

Near Future

- PHENIX year–2 data handling
  - 100TB of tape transfer / 100K of files management
  - I/O between HPSS–disk–Linux farm in analysis
    - 10–100 times larger than the year–1 data experience
- planned size at 2002/3 (end of the first 3–years plan)
  - CPU : 288 cpu / ~13000 SPECint95
  - DISK: 26TB
- next phase
  - HPSS: 400TB tape volume
some more tape drives (9940 series)
240MB/sec tape I/O
  - second tape robot