Experience of BESIII data production with local cluster and distributed computing model

Z.Y. DENG[1], W.D. LI[1], L. LIN[2], H.M. LIU[1], C. NICHOLSON[3], Y.Z. SUN[1], X.M. ZHANG[1], A. ZHEMCHUGOV[4]

[1] Institute of High Energy Physics, China

[2] Soochow University, China

[3] Graduate University of Chinese Academy of Sciences

[4] Joint Institute for Nuclear Research, Russia

Introduction

BESIII Experiment

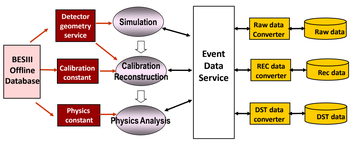
- **BESIII, the Beijing Electron Spectrometer**
  - Collider experiment to study tau-charm physics
  - Energy region: from 2 to 4.6GeV
  - L_{peak}=5*10^{32}cm^{-2}s^{-1}

States	Energy (GeV)	Peak luminosity (10^{33} cm^{-2}s^{-1})	Physics crosssection (nb)	Events/year
J/ψ	3.097	0.6	3,400	1×10^{10}
$\psi(2S)$	3.686	1.0	640	3×10^{9}
$\tau^+\tau^-$	3.670	1.0	2.4	1.2×10^{7}
$D^0\bar{D}^0$	3.770	1.0	3.6	1.8×10^{6}
D^+D^-	3.770	1.0	2.8	1.4×10^{6}
D_sD_s	4.030	0.6	0.32	1×10^{6}

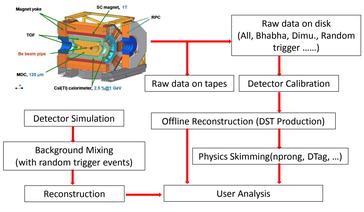
Expected BESIII data samples in a calendar year

BESIII Offline Software System (BOSS)

- BESIII Offline Software System (BOSS), is a new offline data processing software system which is developed based on GAUDI framework
- External Libs: Geant4, ROOT, GDML, MySQL……
- OS: Scientific Linux Cern 5.5, GCC 4.3.2
- Simulation, calibration, reconstruction, and analysis algorithms are core software for data processing and physics analysis, software framework provides these algorithms event data service and constants data service



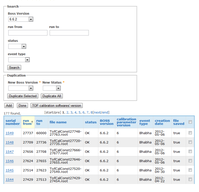
BESIII dataflow



Local Data Processing

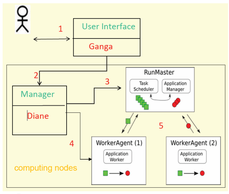
Local data management system

- **Management of raw data**
  - Import information of raw data files from online database
  - Create/search dataset
- **Management of calibration constants**
  - Save calibration constants for specific sub-detector, software version, run range
  - Interface for users to search specific constants
  - Permission control for different users



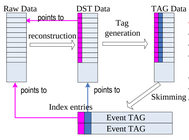
Locally distributed computing: DistBoss

- DistBoss is a distributed system developed for fast calibration, reconstruction, software validation, etc.
- Ganga is used as User Interface
- Diane is used to control and manage the running of master and workers
- The data processing is paralleled at event level



Run cmd:
distboss jobOptions.txt [queue] [worknumber]

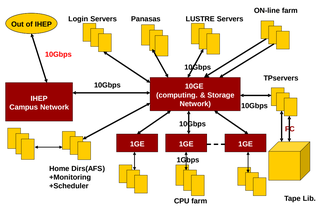
Event Navigation



- **The event tag information could be:**
  - EF type, Good run type;
  - Offline type: nProngs, Dtag result, …
  - Back navigation: run#, event#, physical locations
  - vertex info
  - net charge, evis, …
  - n_goodtrack, n_goodphoton, n_piKp, n_piO, n_Ks…

- To provide information sufficient for primary event selection, quick analysis
- Enable event navigation to help to access information in upstream data by using event index, quick reconstruction of selected events
- Save disk space

BESIII Computing Environment Architecture



BESIII data processing (Storage)

- **BESIII data collected by the end of 2011:**
  - J/ψ: 225M, ψ': 106M
  - ψ'':2.9fb^{-1}(3.5xCLEO-c), ψ(4040): 0.5fb^{-1}
- **Data size**
  - MC raw event size: 5-7kB/event, MC DST event size: 18-25kB/event
  - 225M jpsi inclusive mc(raw+DST): 5TB
  - Total MC data(raw+DST): 16TB (for one software version)
  - Total real raw data: 250TB, Total random trigger data: 9TB
  - Total real DST data: 85TB (for one software version)
- **Plans: more J/ψ, ψ', ψ'', and data at higher energies**
  - The BESIII will take about 10 billion J/psi data and the data collected in other energy points such as Psi', Psi(3770), Psi(4040), etc. will be of equivalent size.
  - The total amount of raw data is estimated to be about 3.6 PB.
  - It is supposed the data reconstruction is repeated at least twice a year, the total size of the DST will be about 1.8 PB.
  - For all the Monte Carlo events, the total storage capacity should be 1.0 PB.

BESIII data processing (CPU)

- **Computing Resources in IHEP**
  - CPU cores
    - 3398 cores, 1536 cores to be available in the near future
  - Tape space (Castor)
    - 4PB, 3PB available
  - Local file system(Lustre)
    - 2464TB, ~1000TB available
- **CPU time of production jobs (with 2000 cores)**
  - Produce 1 billion jpsi inclusive mc DST events: 8 days
  - Reconstruct 1 billion jpsi raw data: 7 days
  - Reconstruct 0.1 billion psip raw data: 1 days
  - Reconstruct 2.9fb^{-1} psipp raw data: 13 days
- With more data accumulated year by year, it's more difficult for IHEP to provide all the computing resources for both raw data processing and MC production
- Hope to use the computing resources from other institutes or universities in BESIII collaboration, distributed computing is necessary

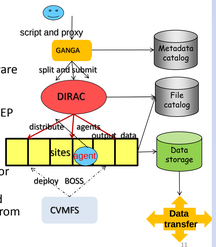
Distributed Computing

Distributed computing

- **Status of remote sites**
  - Most sites are in small scale, and lack experts on grid computing
  - Limitation of network speed between IHEP and other sites
- **Design principles**
  - Make it as simple as possible for sites to join and for users to use
  - Use existing software wherever possible
- **Computing model**
  - IHEP is responsible for processing and storage of all the real raw data
  - MC production and analysis jobs are distributed among a number of sites with enough computing resources
- **Basic requirements for sites**
  - 100 cores, 100TB storage

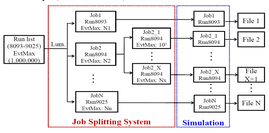
Main Components

- BES VO and VOMS
- GANGA for job submission and management
  - GangaBoss plug-in has been written to support BES software
- DIRAC for running distributed computing jobs
  - DIRAC server is running at IHEP with clients at remote sites
- DFC
  - for data management
- CVMFS (CERN VM File System) for deploying BOSS on target sites
  - Clients running at distributed sites can load BOSS version from server at IHEP



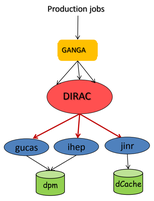
MC production job splitting

- Random seed got from database, and reset new seed to database after job splitting
- Database configure for each site (mysql or sqlite)
- Random trigger data for background mixing
  - There is a copy of all the random trigger files in the SE at each site. Each reconstruction job then downloads the respective random trigger data files to do background mixing.
  - For each job, 2~3 data files, 800MB~3GB



Test Prototype

- Current sites
  - PBS sites
    - GUCAS: 80 cores
    - IHEP-PBS: 96 cores
  - LCG sites
    - JINR: 220 cores
    - IHEP-LCG: 8 cores
- IHEP SE
  - DPM, 200TB
  - for storage of all the MC data
  - Buffer for transferring data between local Lustre system and remote SE
- JINR SE
  - dCashe, 3.5TB served for BESIII test jobs



Performance test

- BOSS 6.6.0 is successfully deployed to 4 sites with CVMFS
- About 3000 BESIII production jobs have been split and submitted to DIRAC, 50M MC events produced
- Output files registered in DIRAC File Catalogue automatically
  - /bes/File/jpsi/660/mc/rhopi/exp1/stream001
  - /bes/File/jpsi/660/mc/rhopi/exp1/stream002
  - /bes/File/psip/660/mc/incl/exp1/stream001
  - ……
  - File level metadata registered after job finished
- Validation between local and distributed computing finished, results are consistent
- Small number of jobs failed with
  - software libraries not found, after reschedule, jobs can finished successfully
  - Job stalled, pilot not running
  - resource temporarily unavailable

Plan

- Analysis of failed production jobs, try to find the reason
- Test physics analysis jobs at distributed sites
- Local Lustre file system + SRM server
  - to enable files registered in the DIRAC File Catalog to be accessed by jobs
- Data transfer tool
  - IHEP already has bbftp-based tool for data transfer between sites
    - Local file system to local file system
  - Selection and integration of data transfer tool
    - FTS, RTS, …
- Continued cooperation with DIRAC developers to make DIRAC more suitable for BESIII experiment
- Integrate more sites to BESIII grid
- Volunteer computing BOINC+CernVM

Summary

Large scale data samples from BESIII detector have been successfully processed.
DIRAC based distributed computing system is set up.
The performance test based on the prototype system shows it works.
More work needed before it comes into use.

CHEP2012, New York University