Belle II computing

Ikuo UEDA (KEK IPNS)

The 40th Anniversary Symposium of the US-Japan Science and Technology Cooperation Program in High Energy Physics
April 15, 2019 @ University of Hawaii
JFY 2012 - 2015  Japan:KEK + US:PNNL  Project Completed
Establishment of a remote data center for acceleration of Belle II data center

Belle IIデータ再プロセス高速化のためのリモート・データセンターの設立

Development of a scalable and automatized production system for the Belle II experiment (US side research title : 2016)

Automatized Production System for Belle II (research title was renamed :
2017 or later : funded by the Japan side only)

Belle II実験における拡張性を考慮した自動化プロダクション・システムの開発

Hiding Data Access Times in HEP Distributed Workflow
Establishment of a remote data center for acceleration of Belle II data center

**JFY 2012 - 2015**  Japan:KEK + US:PNNL  Project Completed

*Belle II*データ再プロセス高速化のためのリモート・データセンターの設立

Development of a scalable and automatized production system for the Belle II experiment


(US side research title : 2016)

*Automatized Production System for Belle II* (research title was renamed : 2017 or later : funded by the Japan side only)

*Belle II*実験における拡張性を考慮した自動化プロダクション・システムの開発

Hiding Data Access Times in HEP Distributed Workflow

Purposes

JFY 2012 - 2015

Establishment of a remote data center for acceleration of Belle II data center

Goal: Acceleration of the speed of the Belle II data reprocessing
by establishing the remote data center in U.S.A.

- to trigger the Belle II computing activity in U.S.A.
- to let the KEK computing resource concentrate on RAW data process
- to reduce the risk of data loss in unexpected contingency
- to develop human resources for computing and middleware

JFY 2016 - 2018

Automatized Production System for Belle II

Goal: Integration of the scalable and automatized production system to the Belle II experiment

- to reduce the burden on expert time and chance of human errors
- to control complicated and different types of jobs smoothly and effectively
- to deliver physics data to users as soon as data-taking finishes
**Belle II Computing Model**

- **GRID sites**
- **Raw data storage and (re)process**
- **mdst storage**
- **MC production and Physics analysis skim**
- **User analysis (Ntuple level)**

**End of year 3**

- **RAW data + produced mDST**
- **MC mDST**
- **Skim uDST**

**KEK Data Center**

**Data Center in US**

**Regional Data Center**

**MC production site**

**Detector**

**HPC sites**

**Cloud sites**

**Computer cluster sites**

**Local resource**

**CPU**

**Disk**

**Tape**

**Storage for original + copy**

**Storage for copy**

**Temporary storage**

**Raw data**

**mdst Data**

**mdst MC**

**dashed inputs for uDST**

**Ntuple**
Belle II Distributed Computing Structure

**Human**

**Software interface**
- BelleDIRAC v4r6p0
  - Interware extension
  - Analysis user interface

**Interware**
- Management system

**Cyberinfrastructure**
- GRID services for Belle II

**Platform**
- GRID Middleware
- OS
- Hardware
- Network

**Sites**

**Production Manager**
- Client Tools
- Production Management System
- Fabrication System

**Data Manager**
- Web Portal
- Monitoring
- Distributed Data Management System
- RMS
- DMS

**End Users**
- CE
- SE
- Cluster
- Cloud I/F

Cloud site

VCYCLE

Cloud I/F

DIRAC slave

CE

SE

Cloud slave

Cluster

Cluster

SE

CE

2017-Dec-13. Computing in HEP - Ueda I.

CE: grid computing element
SE: grid storage element
Automatized Production System

Different types of production
- MC production (w/ or w/o BG)
- Skim production
- RAW data process

Huge variety of modes
- BB, udsc, signal, background
- many physics skims

Complicated data management
- over world-distributed sites

Reduce human error and
- perform effective operation
Research Highlight: One page summary

Proto-Production system
(Automatic job submission
Automatic Issue detection [monitor])

Manual job submission
(no automated Production system)

Normalized CPU power

Proto-Production system
(Automatic Data distributed)

Full-Production system

Coninuous operation
running various types of productions

US is increasing resources gradually

US joined since 2013
Establish the data center in US

US-Japan project (JYF 2012-2015)

US-Japan project (JYF 2016-2018)
Establishment of a remote data center for acceleration of Belle II data center

Belle IIデータ再プロセス高速化のためのリモート・データセンターの設立

JFY 2012 - 2015  Japan:KEK + US:PNNL  Project Completed

Development of a scalable and automatized production system for the Belle II experiment

(US side research title : 2016)

Automatized Production System for Belle II

(research title was renamed : 2017 or later : funded by the Japan side only)

Belle II実験における拡張性を考慮した自動化プロダクション・システムの開発


Hiding Data Access Times in HEP Distributed Workflow

MC production jobs

“Jobs go to Data”

requires storage at each site

Issue: Inefficient use of compute resources without local storage
MC production jobs

“Jobs go to Data” ———> “Remote Data Access”

enables contribution of compute resources
w/o local GRID storage

computing site with Storage

MC job w/ BG

Storage-server-less computing site

necessary BG files from other sites

BG files

computing site with Storage

Issue: Time consumed in Remote Accesses
Belle II computing sites

GRID sites
KEK, BNL, DESY, GridKA, KISTI, CNAF, many European sites
~30 sites: ~75%

Cloud sites
Univ. of Victoria, Univ. of Melbourne
several sites: ~15%

Computer cluster sites
Many Universities in Japan, Korea, India, China, Russia, Mexico,
~25 sites: ~10%

Large contribution from
Compute Resources w/o local GRID storage

Normalized CPU usage by Site
46 Weeks from Week 52 of 2017 to Week 46 of 2018

Max: 339, Average: 183, Current: 61.6

Legend:
- LCG.KEK.jp
- DIRAC.Uvic.ca
- ARC.CERN.ch
- ARC.KIT.de
- LCG.KDE2.jp
- OSG.BNL.usr
- LCG.Napoli.it
- LCG.ESY.de
- LCG.KML.jp
- ARC.SIgnET.it
- ARC.MPI/MU.de
- DIRAC.RCNF.jp
- DIRAC.Uvic-local.ca
- LCG.NTU.tw
- LCG.CESNET.cz
- LCG.MPIR.mn
- LCG.Frascati.it
- LCG.HEPFI.at
- DIRAC.NDUnu.jp
- LCG.Pisa.at
- LCG.KISTI.kr
- DIRAC.IITG.in
- DIRAC.PNNL.us
- LCG.Torino.it
- LCG.BUC.fr
- DIRAC.CIENVESTAV.mx

Comments:
- 12 sites for 25%
- 25 sites for 10%
Remote Data Access

- **Download**
  - copying whole files unnecessarily
  - CPU idle during download

- **Direct I/O** (e.g. xrootd)
  - chaotic remote accesses can be inefficient

- **Organized Streaming (TAZeR: Transparent Asynchronous Zero-copy Remote I/O)**
  - I/O optimization with pre-fetching to memory
  - Intelligent job scheduling

<table>
<thead>
<tr>
<th>Execution time (min)</th>
<th>Download</th>
<th>Direct I/O</th>
<th>TAZeR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core (ordered by exec time)</td>
<td>500min</td>
<td>200min</td>
<td>30min</td>
</tr>
<tr>
<td>Network read time</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Existing remote I/O technique

Support from DOE as a part of “Integrated End-to-end Performance Prediction and Diagnosis”
Applying TAZeR to Belle II

TAZeR
✓ Hiding network and I/O latencies with
  - I/O optimization
  - Intelligent job scheduling

Belle II
✓ Efficient use of compute resources without local GRID storage

Proposed project: “Hiding Data Access Times in HEP Distributed Workflow”

To increase throughput of Belle II Monte Carlo simulations
To identify the conditions under which TAZeR improves HEP workflow