Overview and Highlights of the Belle II Computing and Software

MALACHI SCHRAM
PACIFIC NORTHWEST NATIONAL LABORATORY
On Behalf of Belle II collaboration
Computing and Software Workflow

**Belle II Detector**

Raw data center outside KEK

**Data**

processed @ "raw data center"

**Online disk**

(Sequential root)

**Offline storage**

(Sequential root)

raw data in root: merge

300kB / event

20kB / event

1-5 sec/event @ KEKCC

Event generation

**MC** produced on the world-wide distributed computing

Detector simulation

BG effect (data)

Digitization

Reconstruction

No intermediate stage is saved

mDST distributed to each region

"coordinated" group skimming

**group official skim**: mDST, μDST, index

mDST: reconstruction level info,

μDST: mDST + particle level info,

index: collection of pointers to events

First 2-3 years ("baseline")

User distributed analysis

investigate another possibility

User-job → Ntuple

mDST

more disk requirements

index ← User-job → Ntuple

mDST

less disk requirement

less network requirement access locally or streaming with xrootd/http more chaotic network access

August 4th 2016

ICHEP 2016
Belle II Analysis Software Framework (BASF2)

Data Object:
- C++ class that encapsulates the variables of a specific object
- Various types: StoreObject, StoreArray, and Relational Objects

Data Store:
- Implementation in the Belle II framework to store/retrieve (share) data objects
- Data objects can be stored per event and/or per job

Module:
- Algorithm that performs operations on data object
Belle II is able to run using multiple cores
fork()-based:
- reduces memory requirements
- few thread-safety issues
Multi-core jobs would be very beneficial for distributed computing
- Fewer jobs reduces the strain on the distributed system
- Shorter jobs would allow more efficient back fill on HPC systems and parasitic use of clouds
- Smaller memory footprint per core

<table>
<thead>
<tr>
<th>Input process</th>
<th>Worker processes</th>
<th>Output process</th>
</tr>
</thead>
<tbody>
<tr>
<td>input mod.</td>
<td>Rx</td>
<td>Tx</td>
</tr>
<tr>
<td>Ring Buffer</td>
<td>Rx</td>
<td>Tx</td>
</tr>
<tr>
<td></td>
<td>Rx</td>
<td>Tx</td>
</tr>
<tr>
<td></td>
<td>Rx</td>
<td>Tx</td>
</tr>
<tr>
<td></td>
<td>Rx</td>
<td>output mod.</td>
</tr>
</tbody>
</table>

August 4th 2016
Parallel Processing Memory Reduction

Proportional memory

- 4 independent single-core jobs
- Parallel processing (4 cores) – Total
- Fraction of memory for one sub-process

Memory / MB

- 4557.44 MB
- 2086.12 MB

Time / s

August 4th 2016

ICHEP 2016
Belle II is using Geant4.10.1.2
- Default BASF2 physics list is FTFP_BERT
- Geometry parameters are being migrated to the conditions database
- Three geometries:
  - Designed geometry is available for all sub-detectors
  - Displacement geometry is the difference between designed and actual
  - Mis-alignment geometry for subtle shifts outside of Geant4

Magnetic Field:
- In MC 5, the B field is constant at 1.5T
- In MC 6, a 3D B field was made available

Digitization is available and managed by each sub-detector group

Validation
- Simulation histograms are added to the basf2 validation package
- Variable monitored include: particle multiplicity, energy distributions, etc
The beam background is simulation using **Strategic Accelerator Design (SAD)**, provided by the accelerator group.

- Simulates transport of particles through the accelerator.
- If a particle leaves a nominal beam trajectory and collides with the beam pipe or collimator, its 4-momentum is saved to a file.
- The file corresponds to one $\mu$s of running at the nominal SuperKEKB luminosity.
- The file is passed to Geant4 for simulation and deposited energy is stored (SimHits).
- Backgrounds include: Radiative Bhabha, Touschek scattering, Beam-gas interactions, two-photon QED.
- The backgrounds are mixed with the signal SimHits.

Currently working on background overlay using random trigger events.
For more details please see Peter Lewis’s talk titled “Status of SuperKEKB commissioning and Belle II detector construction”

http://indico.cern.ch/event/432527/contributions/1071301/
Modular architecture to support re-usability and distributed development

Real data reconstruction: VXD Test beam and CDC Cosmic Rays
Electromagnetic Calorimeter Reconstruction

- Belle reconstruction worked well for low background but not with high background
- Novel hypothesis based clustering is being investigated
- Different hypothesis will have different number of crystals within a “cluster”

![Graph showing energy resolution for true energy versus measured energy for old and new clustering methods.](image)

- e.g. 30 MeV
Charged hadron identification primarily from combined information from TOP, ARICH and dE/dx measurements in the CDC and SVD

Each detector component provides log likelihoods of 6 particle hypotheses: electron, muon, pion, kaon, proton and deuteron

dE/dx only
kaon vs. pion selection
$\log \mathcal{L}_K > \log \mathcal{L}_\pi$

Top only
kaon vs. pion selection
$\log \mathcal{L}_K > \log \mathcal{L}_\pi$

ARICH only
kaon vs. pion selection
$\log \mathcal{L}_K > \log \mathcal{L}_\pi$

Combined
kaon vs. pion selection
$\log \mathcal{L}_K > \log \mathcal{L}_\pi$
Provide read/write access to calibration, alignment, geometry, environmental parameters

- Adopted REST API
  - *Apache* server
  - Application Interface (*GlassFish*)
  - *PostgreSQL* database
  - *Lustre* file system

- A global tag is a unique identifier to define a collection of payloads with defined IOVs

- Belle II will use global tags to define processing runs, etc
Scaling the Conditions Database

Prototype developed using **OpenStack**

Use **Docker** to develop modular components which allows for quick scaling as needed

Use **Swagger** for automated API documentation

Use **Payara** and **Hazelcast cluster** for flexible and scalable in-memory data grid solution

Testing **OpenStack Trove** provide redundancy and high availability
SuperKEKB luminosity projection

Goal of Belle II/SuperKEKB

- Commissioning started in early 2016
- Peak luminosity (cm$^{-2}$s$^{-1}$)
- Integrated luminosity (ab$^{-1}$)

Calendar Year

~50x the integrated luminosity of Belle

~40x the peak instantaneous luminosity of Belle

Preliminary

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Storage (PB)</td>
<td>3</td>
<td>8</td>
<td>10</td>
<td>22</td>
<td>60</td>
<td>116</td>
<td>190</td>
<td>265</td>
<td>341</td>
</tr>
<tr>
<td>Total CPU (kHEPSpec)</td>
<td>240</td>
<td>340</td>
<td>475</td>
<td>610</td>
<td>710</td>
<td>880</td>
<td>1060</td>
<td>1420</td>
<td>1640</td>
</tr>
</tbody>
</table>
DIRAC (Distributed Infrastructure with Remote Agent Control) INTERWARE

KIT, CNAF, CESNET, SIGNET, HEPPHY, ULAKBIM ...

- Provided as a DIRAC plugin
- Need additional installation
- Multiple cloud sites allowed
- Handle each cloud as a site
- No modification in cloud site

Clusters w/o middleware
GE, TORQUE, LSF, ...
Direct submission
BINP, NSU, many universities in Japan
Belle II Distributed Computing Layers

Production Manager  Data Manager  End Users  Operations

Production Management
Fabrication
Transformation System

BelleDIRAC

gb2 client tools
Web UI
Monitor

DIRAC

VMDIRAC

WMS

VMDIRAC

Transformation System

Sites

cloud site

Cloud I/F

Cluster

SE

AMGA
LFC
VOMS
FTS
CVMFS

Grid Services

August 4th 2016  ICHEP 2016
“Production Manager” (human)
- to define a production: sw release, event type, number of events to produce, …
- and put the parameters into Production Management System

Production Management System
- to create Fabrication instance(s): simulation, reprocess, skim, merge, …
- and chain them as needed

Fabrication System
- to create and look after jobs, replace failed jobs, verify output of successful jobs
  - simulation jobs can run everywhere, merge jobs can run at the sites with “major” SEs.
- to feed output files to Distributed Data Management System

Distributed Data Management System
- to gather output files to “major” SEs in blocks (important for merge)
- to distribute “final” products over the grid (for further processing / analysis)
Road to full production system

**MC3** (2014 May): no Production System

- 2.5 M jobs: *manual submission* by the *shifters* (with a script)

**MC4** (2014 Nov): with “prototype” Fabrication System

- ~5 M jobs: *automatic submission* by the system, controlled by a single system manager
- Output files uploaded directly from WNs to “major SEs”

**MC5** (2015 Aug — 2016 Feb): with improved “prototype” Fabrication System

- ~10 M jobs: automatic submission by the system, controlled by a single system manager
- Output files transferred by FTS from production sites to “major SEs”
- Monitoring and automatic issue detection for shift tasks and expert investigation

**Cosmic ray data replication** (2016 Jun ~):

- An agent to register raw data files as they get available on a storage, and make transfer requests
- Automatic replication of raw data
Road to full production system (2)

**MC6** (2016 Jun): with Fabrication System and DDM

- O(1) M jobs: automatic submission by the system, controlled by a single system manager
- Output file transfers by FTS **controlled by DDM**
- Monitoring and automatic issue detection for shift tasks and expert investigation

**MC7** (2016 Nov ~ 5 months): with the Production Management System

- **The software should be well tested by the software group, validated by a production manager**
- Automatic job submission by the system, **controlled by the system**
- Automatic file transfers by the system, **controlled by the system**
- Monitoring of **production progress**
- **Automatised operations** (eg. ggus tickets, downtime, storage health)
- **Tune and improve the system** during the production and afterwards
- **The system will be ready for workflows** other than MC production (raw dta processing, reprocessing, skim, …)
Scale Tests (to be repeated as needed in between “production” activities)

- To see if we have any limitation in job or file transfer throughputs beyond the current level
- And address the issues found

Cosmic ray data processing (2017)

- Raw data format and reconstruction software need to be ready and well tested by this time
- The production system should be ready, it has been developed not only for MC production, but with other use cases (incl. raw data processing)
- Still, this will be the first real use case to try raw data processing workflow

System dress rehearsal (2017): before Phase-2 runs

- To try out the full chain workflow from raw data to skim
- Find issues and address them before phase-2 runs

System full dress rehearsal (2018): before Physics runs
Belle II Networking Requirements

► In 2014, Belle II was added to the LHCONE network

► International network requirements are based on (production only):
  ▪ RAW data
  ▪ mDST from data and Monte Carlo - after data taking and reprocessing

► KEK estimated outgoing (incoming) traffic of 20Gbps (5Gbps) in 2024.

► PNNL estimated maximal incoming (outgoing) traffic 8Gbps (4Gbps)

► European estimated maximal incoming (outgoing) traffic 10Gbps (4Gbps)

► India, Canada and Slovenia will require ~2Gbps traffic

► Other site (Australia) will be interested by a traffic less than 1Gbps
Latest Network Data Challenges

DC 1 (2015):
- KEK ↔ PNNL
- KEK → EU

DC 2 (2015/2016): LHCONE sites
- PNNL ↔ EU
- EU ↔ EU

DC 3 (2016): SINET5 (phase 1)
- KEK ↔ PNNL
- KEK → EU

Software stack used:
- perfSONAR suite
- SRM storage elements + gridftp servers
- FTS3 with manual channel configuration
Network Data Challenge Results

Results of key sites based on the latest data challenges

<table>
<thead>
<tr>
<th>Source</th>
<th>KEK</th>
<th>PNNL</th>
<th>DESY</th>
<th>KIT</th>
<th>CNAF</th>
<th>NAPOLI</th>
<th>SIGNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEK</td>
<td>N/A</td>
<td>6.3 Gbps</td>
<td>8 Gbps</td>
<td>7 Gbps</td>
<td>7 Gbps</td>
<td>6.6 Gbps</td>
<td>2.5 Gbps</td>
</tr>
<tr>
<td>PNNL</td>
<td>3.9 Gbps</td>
<td>N/A</td>
<td>6.6 Gbps</td>
<td>6 Gbps</td>
<td>9 Gbps</td>
<td>8 Gbps</td>
<td>-</td>
</tr>
<tr>
<td>DESY</td>
<td>3 Gbps</td>
<td>6.6 Gbps</td>
<td>N/A</td>
<td>8 Gbps</td>
<td>8 Gbps</td>
<td>8 Gbps*</td>
<td>3 Gbps</td>
</tr>
<tr>
<td>KIT</td>
<td>3.2 Gbps</td>
<td>4 Gbps*</td>
<td>8 Gbps</td>
<td>N/A</td>
<td>8 Gbps</td>
<td>6 Gbps</td>
<td>3 Gbps</td>
</tr>
<tr>
<td>CNAF</td>
<td>3.8 Gbps</td>
<td>8 Gbps</td>
<td>10 Gbps</td>
<td>6 Gbps</td>
<td>N/A</td>
<td>8 Gbps*</td>
<td>3 Gbps</td>
</tr>
<tr>
<td>NAPOLI</td>
<td>3 Gbps</td>
<td>3 Gbps*</td>
<td>3 Gbps*</td>
<td>3 Gbps</td>
<td>3 Gbps*</td>
<td>N/A</td>
<td>3 Gbps*</td>
</tr>
<tr>
<td>SIGNET</td>
<td>0.8 Gbps</td>
<td>0.6 Gbps*</td>
<td>5 Gbps</td>
<td>2 Gbps</td>
<td>5 Gbps</td>
<td>2 Gbps*</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Networking Status and Plans

► **Current Status**
  - Network data challenge verify that:
    - KEK→PNNL requirements are satisfied until 2019
    - KEK outgoing traffic appears to be limited to ~3-4Gbps
    - PNNL→EU requirements are satisfied
    - EU ↔ EU requirements are satisfied
  - Next data challenge will test new KEKCC and full LHCONE setup (Fall 2016)

► **Moving Forward (FY16-17):**
  - Integrate networking transfer tools into the Distributed Data Management System
  - Integrate KEK in LHCONE and perform new Data Challenge
  - Update network requirements to include estimation on non-production traffic
Summary

► **Software**

- Analysis tools are in place
- Physics feasibility studies are being performed using Analysis tools and the latest Monte Carlo samples
- Software chain is being exercised with the latest RAW data from: VXD, CDC, TOP, etc.

► **Computing**

- Prototype production system is available
- Working on scalability test of each components and overall system
- Three separate systems: Develop/Validate/Production
- Developing automated unit testing