The Management of Heterogeneous Resources in Belle II

Malachi Schram, Vikas Bansal, Antonio Ledesma
Pacific Northwest National Laboratory
August 22nd, 2017
Belle II Overview

- **Goal**: discover new particles and phenomena beyond the Standard Model of particle physics
- Collaboration among **725+ physicists** from 104 institutes in 24 countries
- **50x** the data volume, **40x** rate of collisions relative to previous Belle experiment
- **PNNL led U.S. (DOE)** contribution to Belle II detector construction – now complete
- **Largest** ever U.S. science investment in Japan
  - More Ph.D. physicists (50+) and more institutions (14) than any other country
- **SuperKEKB**: single beam circulation was done **successfully** (phase 1 in 2016)
- Cosmic-ray **data taking is on-going**
- **Physics** run will start in **2018**
Belle II Computing Requirements

- Expected data rates from the Belle II experiment are high
  - Event rate of 6 kHz, corresponds to 11 PetaBytes per year starting in 2022
  - 100 PetaBytes raw data volume by 2024, total data volume ~190 PB
- Processed data samples will be distributed worldwide (Asia, North America, Europe)
- The Belle II Computing Steering Group resource estimations and feedback from Belle Program Advisory Committee and DOE reviews
- Resource requirements depend upon luminosity, event size, replicas, MC samples, etc.

Goal of Belle II/SuperKEKB

Peak luminosity (cm\(^{-2}\)s\(^{-1}\))

Integrated luminosity (fb\(^{-1}\))

9 months/year
20 days/month

Belle II Computing Model

Raw data storage and processing
KEK Data Center
Raw Data Center
PNL Data Center
Raw data duplication process

MC production
Regional Data Center
MC production site

Cloud site
Computer cluster site

Global site

Local resource

End of year 3

Recap

GKD site

9th August 2017
Belle II Monte Carlo Production

- **Accomplishments:**
  - MC production are critical exercises to test the latest software, scalability, and provide physics samples for analysis.
  - Monte Carlo #7 ran from Nov. 1\textsuperscript{st} 2016 to Feb. 7\textsuperscript{th} 2017
    - Approximately 27 billion events simulated
    - Reached 25k concurrent jobs (>200 kHEPSpec)
  - Monte Carlo #8 started on Feb 16\textsuperscript{th} and is done ... MC9 is ramping up
    - New Belle II computing record ~300kHEPSpec as measured by DIRAC pilot jobs

---

![Running jobs by Country](image)
DIRAC (Distributed Infrastructure with Remote Agent Control) INTERWARE

- KIT, CNAF, CESNET, SIGNET, HEPHY, UA-ISMA, ULAKBIM, CYFRONET, .....  
- GRID Middlewares
- Open Science Grid

Distributed infrastructure with Remote Agent Control
(originally developed for LHCb)

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

KEK

A part of DIRAC

VMDIRAC

- CREAM CE
- SiteDirector
- SLURM SiteDirector

Clusters w/o middleware
GE, TORQUE, LSF, ...
Direct submission

BINP, NSU, many universities in Japan

Cloud site

Dynamic Torque

HTCondor Cloud Scheduler

HTCondor VM Manager

cloud site

Academic clouds
- Seen as a traditional CREAM CE site
- Installed in each cloud site

Commercial clouds, Amazon EC2, etc

22th August 2017

ACAT 2017
Belle II Distributed Computing

Diagram showing the system architecture with various components and services:
- Production Manager
- Data Manager
- End Users
- Operations

BelleDIRAC components:
- Production Management
- Fabrication
- gb2 client tools
- Web UI
- DDM
- Monitor
- Transformation System
- VMDIRAC
- WMS
- AMGA
- LFC
- VOMS
- FTS
- CVMFS

DIRAC components:
- Sites
- Cloud services
- Cluster
- DIRAC slave
- Cloud I/F

22th August 2017
U.S. Belle II Computing Requirements

- Resource requirements depend upon luminosity profile & event size (data volume), (re)processing, MC generation, and user analysis.
- MC production accounts for nearly 50% of US Belle II computing resources.
- Using of Leadership Class Facilities (LCFs) for MC production.
- Investigating LCFs for scalable detector and physics studies.
  - Rare physics channels such as $B \rightarrow D^* \tau\nu$ and $B \rightarrow K(*)\tau\tau$.
  - iTOP “ring image” PID performance studies using Deep Learning.
  - Detailed systematic studies when using Deep Learning models.
  - Exhaustive hyper parameter scans.

<table>
<thead>
<tr>
<th></th>
<th>CY17</th>
<th>CY18</th>
<th>CY19</th>
<th>CY20</th>
<th>CY21</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU [kHEPSpecs]</strong></td>
<td>20.11</td>
<td>27.56</td>
<td>58.90</td>
<td>69.71</td>
<td>82.97</td>
</tr>
<tr>
<td><strong>Storage [PB]</strong></td>
<td>0.31</td>
<td>0.81</td>
<td>5.04</td>
<td>6.50</td>
<td>9.28</td>
</tr>
<tr>
<td><strong>Networking In/Out [Gbps]</strong></td>
<td>0.30/0.30</td>
<td>0.49/0.36</td>
<td>1.06/0.26</td>
<td>1.56/0.31</td>
<td>1.89/0.83</td>
</tr>
<tr>
<td><strong>CPU @ PNNL [kHEPSpecs]</strong></td>
<td>0.41</td>
<td>6.35</td>
<td>40.52</td>
<td>29.01</td>
<td>41.58</td>
</tr>
<tr>
<td><strong>CPU @ LCF [kHEPSpecs]</strong></td>
<td>19.70</td>
<td>21.21</td>
<td>18.38</td>
<td>40.70</td>
<td>41.39</td>
</tr>
</tbody>
</table>
PNNL Computing Facilities & Infrastructure

• **Description:**
  – PNNL leverage virtualization technologies using Kubernetes and OpenStack to provide a fault tolerance and flexible infrastructure
  – HEP Computing is funded by a mix of DOE, PNNL HEP project money, and US-Japan funding

• **Major Components:**
  – Gridftp servers w/ bestman2 SRM
  – OpenStack
  – Enterprise Linux-compatible distros
  – Docker
  – DIRAC
  – Condor
  – Kubernetes
  – Ceph object store
  – Lustre shared filesystem

• **Services provided via private cloud:**
  – DIRAC server instances
  – Compute nodes
  – FTS3 to manage file transfers
  – GUMS, VOMS servers
  – CVMFS Stratum 0 & 1
  – Squids
  – Custom services on relational databases (e.g. Belle2db)
Belle II Computing on NERSC

- NERSC has two clusters with grid access:
  - Edison
  - Cori
- NERSC Allocation for 2017
  - Used to develop/test/validate the use of HPC for Belle II
- DIRAC GlobusComputingElement is being modified to provide extra options to enable the use Edison and Cori
- MC9 phase 3 just started and will allow us to monitor the availability of the NERSC resources

<table>
<thead>
<tr>
<th>Project Title</th>
<th>US Belle II HPC Workflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>Pacific Northwest National Laboratory (PNNL)</td>
</tr>
<tr>
<td>DOE Office &amp; Program</td>
<td>HEP - HEP Other Research</td>
</tr>
<tr>
<td>Science Category</td>
<td>High Energy Physics</td>
</tr>
<tr>
<td>Project class</td>
<td>DOE Base Funding</td>
</tr>
</tbody>
</table>

Funded by DOE Office of Science? Y
DOE Manager | Kevin Flood, Glen Crawford
MPP Hours requested in ERCAP | 10,000,000
SRUs requested in ERCAP | 1,000

Cori providing ~15kHEPSpec
Max: 34,146, Average: 9,815, Current: 498

22th August 2017
ACAT 2017
Grid Components for NERSC

**DIRAC**

**Workload Management Agent**
- SiteDirector
  - Using modified `GlobusComputingElement`

**Resource Definition**
- **OSG.CORI.us**
  - Defined as `GlobusComputingElement`
  - Mapped to PNNL StorageElement
- **OSG.EDISON.us**
  - Defined as `GlobusComputingElement`
  - Mapped to PNNL StorageElement

**Belle II HPC Docker**
- pnnlhep/osg-compute
- /cvmfs/belle.cern.ch
  - Sync repo to docker

**NERSC**
- Pull and register into Edison/Cori shifter
- Grid submit with docker image and volume host/docker mount points for input/output/repo
LCFs Status and Plans

For NERSC:
- Request allocation; 10M core hours for CY17
- Setup for NERSC Cori and Edison setup are done
- Docker with software release 00-09-01 used for MC9 is done
  - Develop MPI wrapper to submit “big” jobs and test on NERSC
  - Create multiple DIRAC SiteDirector to submit different size jobs to NERSC

For ORLCF:
- Request allocation; 1M core hours for CY17
  - Develop new DIRAC “SiteDirector”
  - Develop backlog job optimization for scheduling
Belle II Distributed Data Management Computing Software

DIRAC is the main software to orchestrate Belle II computing resources.

MaDDash provides network monitoring. Belle II sites are part of MaDDash mesh.

DIRAC

MaDDash

perfsONAR

WAN Data Challenge Setup

File Transfer Service (FTS3) to schedule transfers between sites

FTS3

BWCTL

BWCTL for initial network test. Setup on perfSONAR boxes

Iperf3, traceroute

GRIDFTP

GridFTP servers render file transfers between sites
Belle II Distributed Data Management Overview

DDM Service Layer
- DataOperationRequest
- StorageElementStatus

DDM Database Layer
- DataOperationDB
- StorageElementStatusDB
- ReplicaAndPopularityDB

DDM Agent Layer
- DataOperationTaskFanout
- DataOperationExecuting
- DataOperationCleaning
- StorageElementStatus
- DataOperationTaskStatusUpdate
- DataOperationRequestStatusUpdate
- DataOperationReplica

DIRAC Base Layer

22th August 2017
ACAT 2017
Belle II Networking

• **Description:**
  – Coordinate with the relevant National Research and Education Network (NREN) providers to organize and evaluate the WAN network status and requirements for the Belle II collaboration
  – Coordinate the development effort to integrate the networking information into the Belle II distributed computing framework
  – Establish a full design and specification of the data transfer and processing/reprocessing workflow between KEK and PNNL

• **Accomplishments:**
  – Full integration into LHCONE
  – Network upgrade by NRENs and major sites: 2x10Gbps dedicated optical path from Seattle, 1x10Gbps dedicated optical path backup path through Boise (tested)
  – Detailed network diagram for each major site (WAN, DTNs, storage backend)
  – Several Network Data Challenges (NDC) before/after SINET upgrade: KEK to PNNL transfer rates based on NDC results is 16Gbps (max 20Gbps)
  – NDC results demonstrates that Belle II networking requirements are satisfied
Latest Belle II Network Data Challenge Results

- A lot of moving part that requires frequent re-evaluation
  - Site reconfiguration: Network, Storage, Data Transfer Nodes, etc.
  - A stable perfSONAR mesh will simplify the network validation process
- New NDC is planned this fall

<table>
<thead>
<tr>
<th>Source ➔ Destination ➖</th>
<th>KEK (Gbps)</th>
<th>PNNL (Gbps)</th>
<th>DESY (Gbps)</th>
<th>KIT (Gbps)</th>
<th>CNAF (Gbps)</th>
<th>NAPOLI (Gbps)</th>
<th>SiGNET (Gbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEK</td>
<td></td>
<td>6.2</td>
<td>11.0</td>
<td>5.0</td>
<td>9.2</td>
<td>15.0</td>
<td>3.0</td>
</tr>
<tr>
<td>PNNL</td>
<td>16.0</td>
<td></td>
<td>10.0</td>
<td>6.0</td>
<td>14.0</td>
<td>10.0</td>
<td>-</td>
</tr>
<tr>
<td>DESY</td>
<td>6.0</td>
<td>6.6</td>
<td></td>
<td>8.0</td>
<td>8.0</td>
<td>8.0</td>
<td>3.0</td>
</tr>
<tr>
<td>KIT</td>
<td>5.6</td>
<td>4.8</td>
<td>8.0</td>
<td></td>
<td>8.0</td>
<td>6.0</td>
<td>3.0</td>
</tr>
<tr>
<td>CNAF</td>
<td>18.0</td>
<td>14.0</td>
<td>10.0</td>
<td>6.0</td>
<td></td>
<td>8.0</td>
<td>3.0</td>
</tr>
<tr>
<td>NAPOLI</td>
<td>16.0</td>
<td>6.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td></td>
<td>3.0</td>
</tr>
<tr>
<td>SiGNET</td>
<td>1.6</td>
<td>0.6</td>
<td>5.0</td>
<td>5.0</td>
<td>5.0</td>
<td>2.0</td>
<td></td>
</tr>
</tbody>
</table>
Integrating Networking in the DDMS

- perfSONAR servers set up at Belle II sites for network monitoring
- Belle II MaDDash mesh is online and provides bandwidth and latency information
- Network information is being part of distributed data management (DDM) inside DIRAC
- Automate notification to sites with network problems and storage status update
DDMS Status and Plans

Status:
- File replication rates reached ~20k files per hour
- File deletion rates reached ~10k files per hour

Short Term Plans:
- Datablock-level replica and popularity catalog implementation
- Performance tuning with multithread agents
- Streamline BelleDIRAC access to FTS3 server
- Job data access lock
- Integrate networking information