Belle II Distributed Data Management and Networking

VIKAS BANSAL, MALACHI SCHRAM, ANTONIO LEDESMA

Pacific Northwest National Laboratory, Richland, WA

DPF ‘17, Fermilab, August 3, 2017
Belle II Collaboration

PNNL is US DOE lead institution for Belle II

US: 87 members from 14 institutes

More US institutions and Ph.D. physicists on Belle II than any other country including Japan (excluding KEK)

Countries: 24
Institutes: 104
Members: 722

c.f. CERN Greybook July 2017
ATLAS: 39 countries, 217 inst., 7783 members
CMS: 49 countries, 208 inst., 6217 members
ALICE: 41 countries, 167 inst., 2799 members
LHCb: 17 countries, 74 inst., 1494 members
The Belle II experiment is for Super B factory at KEK in Japan.

- Complementary physics to the LHC based on precision measurements from high-intensity beams.
- Total integrated luminosity: 50 ab\(^{-1}\).
- Collisions start in Early 2018.
- Similar data rate as from LHC Run I.

Raw Data: 100 KB / event
Detector mDST: 5 KB / event
MC mDST: 6 KB / event
Reconstruction: 20 HEPSPEC *s / event
MC: 60 HEPSPEC *s / event

<table>
<thead>
<tr>
<th></th>
<th>2017</th>
<th>2018</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk [TB]</td>
<td>3000</td>
<td>4500</td>
<td>11000</td>
</tr>
<tr>
<td>Tape [TB]</td>
<td>2000</td>
<td>2200</td>
<td>6500</td>
</tr>
<tr>
<td>CPU [KHEPSPEC06]</td>
<td>210</td>
<td>400</td>
<td>480</td>
</tr>
</tbody>
</table>
Belle II computing sites across the globe

Running Jobs by Sites

24 Hours from 2017-01-18 19:00 to 2017-01-19 19:00 UTC

Sites employ
EMI
OSG
Cloud
Local cluster

Univ. of Victoria
PNNL
Belle II computing model until Year 3
Belle II distributed computing software

- DIRAC (Distributed Infrastructure with Remote Agent Control) as the solution of choice
- Successfully used by LHCb. Extended for Belle II.
Distributed data management system 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**
Purpose:
- Provide near real time storage elements information

Current implementations:
① StorageElementStatus provide API layer for internal and external users.
② StorageElementStatusAgentDB provide persistified information
③ StorageElementStatusAgent:
  - Available space
  - Access rights (w/r) at file and director level
  - ADLER32 Checksum

Extra features:
- Include full ACL
DataOperation: Workflow Overview

User → DataOperation Service → TaskFan out Agent → Requests

Requests → Tasks

Tasks → Task Exec Agent

Task Exec Agent → Replicate, Delete
DataOperation: Replication Workflow

User

DataOperation("Replication")

DataOperationTaskFanoutAgent

DataOperationDB

DataOperationTaskStatusUpdateAgent

DDM Layer

DataOperationCleaningAgent

DataOperationExecutingAgent

StorageElementStatus()

DataOperationRequestStatusUpdateAgent

DDM Layer

AccountingDB

DIRAC Base Layer

Request Management System DB

RequestExecutingAgent

FTS DB

FTS Agent
DDM by design does not delegate deletion of files to base DIRAC to avoid race conditions on data files.
- In addition, Belle II utilizes AMGA (metadata service) that is not available in base DIRAC.
- Data operation executing agent only acts on one operation on physical file at any given time.
Distributed data management system in production

- Distributed Data Management System (DDMS) is successfully used in Monte Carlo (MC) samples production
- Production system hand overs data files to DDMS
  - DDMS decides to transfer the files/data to specific sites as per policy.

![Data blocks distribution over storage elements](image-url)
Distributed data management system in production

- Production system hand overs data files to DDMS
  - DDMS check for site health before starting any transfers. Replicates through DIRAC using File Transfer Server and reports back finished transfers

![Bar chart showing files transfer request assigned to DDMS](image1.png)

- 3100 files/hour

![Bar chart showing files transfer request finished by DDMS](image2.png)

- 4100 files/hour
  - Maintenance time hence quiet
Belle II wide area networking

- Powered by National Research and Education Networks (NREN), e.g. ESnet
- Dedicated servers set up at Belle II sites for network monitoring
- PNNL leads the effort in setting up Belle II Monitoring Dashboard (MaDDash) mesh to measure sites-pair wise bandwidth and latency
- Network info is part of DDMS inside DIRAC
- Automate notification to sites with network problems
WAN data challenge KEK ↔ PNNL

- Estimated network bandwidth for peak outgoing traffic from KEK: 9 Gbps
- KEK outgoing traffic measured at 16 Gbps
- DDM can be made to assess various network paths for optimal data transfers
Belle II computing needs on par with LHC Run I
Belle II distributed computing software is written in DIRAC framework
Distributed data management system handles data transfers and deletion and is currently deployed in production mode
We are actively working on next version of DDM that will directly schedule transfers to FTS to avoid DIRAC’s RMS layer
We foresee to add network health and monitoring in distributed data management system
Detailed Deletion Workflow

1. Initialize deletion task list ("Deleter" or "Deleter2")
2. Do list of tasks associated with above task

   For each task in task list

   If
   
   - True: Update status in DBM
   - False: Continue

   For each LAIR in list

   If
   
   - True: Remove LAIR from DB
   - False: Continue

   Modify LAIR's status to "Deleter"