Report on Large Scale DH Test

J.H Kim[1] & S. I Ahn[2] & K. Cho[1]

[1]High Energy Physics Team
[2]e-Science Grid IT Team
KISTI, Daejeon, Korea

Belle II General Meeting, 2010.04.01

Overview

1. Paper status
2. Estimating the meta-data of Belle II
3. Large Data Handling test
4. Summary and next plan

Paper Status

- 2009.12.19 : Presented a talk at CCP2009.
- 2010.03.10 : Submitted a paper of meta-system,The advanced data searching system with AMGA at the Belle II experiment, which is presented at CCP2009.
- 2020.03.11 : Got the confirmation of submisstion from CPC.
- 2020.03.11 ∼ : On processing of review.

Elsevier Editorial System(tm) for Computer Physics Communications
Manuscript Draft

Manuscript Number:

Title: The advanced data searching system with AMGA at the Belle II Experiment

Article Type: Special issue: CCP 2009

Keywords: Keywords: High energy physics, Belle II, Large data handling, Metadata service, AMGA

PACS: 07.05.-t; 07.05.Tp; 29.85.-c

Corresponding Author: Prof. Kihyeon Cho, Ph.D.

Corresponding Author's Institution: KISTI

First Author: Junghyun Kim, Ph.D.

Order of Authors: Junghyun Kim, Ph.D.; SunIl Ahn, Ph.D.; Kihyeon Cho, Ph.D.; M. Bracko, Ph.D.; Z. Drasal, Ph.D.; T. Fifield; R. Fruhwirth, Ph.D.; R. Grzymkowski, Ph.D.; T. Hara, Ph.D.; M. Heck, Ph.D.; S. Hwang, Ph.D.; Y. Iida, Ph.D.; R. Itoh, Ph.D.; G. Iwai; H. Jang, Ph.D.; N. Katayama, Ph.D.; Y. Kawai, Ph.D.; C. Kiesling, Ph.D.; B. K Kim, Ph.D.; T. Kuhr, Ph.D.; S. Lee; W. Mitaroff, Ph.D.; A. Moll, Ph.D.; H. Nakazawa, Ph.D.; S. Nishida, Ph.D.; H. Palka, Ph.D.; K. Prothmann, Ph.D.; M. Rohrken, Ph.D.; T. Sasaki, Ph.D.; M. E Sevior, Ph.D.; M. Sitarz, Ph.D.; S. Stanic, Ph.D.; Y. Watase, Ph.D.; H. Yoon; J. Yu; M. Zdybal, Ph.D.

Abstract: We use a metadata service at the Belle experiment which provides a mechanism to locate files using descriptive information. However, at the Belle II experiment, we will have 50 ∼ 60 times more data than that of the Belle experiment. Therefore, it is expected that the existing metadata service has problems with performance, scalability, and durability, in particular, if it is extended to an event-level for searching metadata. To deal with this issue, we have designed a new metadata schema for Belle II which significantly reduces disk space for metadata, and proposed a new metadata service system which provides good performance and scalability based on Arda Metadata catalog for Grid Application. The control of the event-level metadata provides an efficient scheme of processing such as events with many tracks.

Estimating the size of meta-data at Belle II

- We estimated the size of meta-data using that of Belle.
- Partially, we have minor corrections to estimate the meta-data size.
  - The Belle II will take the other method to store the data.
  - We need to consider the definition of computing part of TDR (suggested by Kuhr.
- We need to change the Table 14.2 of TDR based on the suggestion.

- Last year, we performed to realize the meta data based on exp07.
  - Data type is on_resonance, uds and stream 0.
  - We processed 2013 files
  - The files are around 12 million events.
  - The space occupation in DB is $\sim$ 145GB for the event-level.
  - The meta-data was evaluated as 12 Bytes/event and 600 Bytes/file.

Table: Reference

Space Occupation per file in DB	600Bytes
Average number of events in a file	111,190
Space Occupation per event in DB	12Bytes
Multiples in Belle II	60

- We refer from the TDR (page 434-435) for the realistical estimation of the meta-data size.
- We suppose that it be 4GB/file.
- We suppose that skims be around 30 kinds.
- We suppose that skim ratio be 10%.

Table: Summary of data of Belle II (TDR page 435)

data type	size
Raw data[PB]	386.4
mDST[PB]	17.2
MC[PB]	51.5

Table: Estimatation of the meta-data size for Belle II

	# of files (Belle)	# of files	Size for file level	Size for event-level
Raw data		100.0M	60.0GB	
mDST	24k	4.3M	2.6GB	5.6TB
skims	720k	12.9M	7.8GB	
MC	240k	12.5M	7.5GB	17.0TB
skims	7,200k	37.5M	22.5GB	

- We constructed the large meta-data ($\sim$140GB) based on the event-level last year.
- We expect to handle the size of file-level since total size of meta-data for the file level is less than 140GB.
- We will have a lot of size for the meta-data of the event-level.
  Therefore, we need to try to reduce the meta-data size with new method for event-level.
- We need to evaluate the patterns of end-user for searching the interesting data more and more.
  - To make the rule or limitation of searching.
  - To improve the performance.

The large Data Handling test

- Object: discussion (2010.03.05)
  - The first phase will be a local stress test of the AMGA system at KISTI. Scalability of read and write accesses to AMGA will be tested.
  - The second phase will involve meta-data replication and data transfers to other sites. This tests the administrator component of the DH system and the performance of the components involved in the transfer (network, disks/tape, LFC).

Currect status

- Transfer:
  When will the data transfer to KISTI be completed?
  $\rightarrow$ around 80% (100TB/120TB)
  How much data is needed for a meaningful test?
  $\rightarrow$ all of data (100TB) in KISTI
- Extraction:
  Extract meta-data from the Belle Data.
  $\rightarrow$ on the 1st processing(3.13 $\sim$ 3.23)
  $\rightarrow$ on the 2nd processing(3.24 $\sim$ current)
  $\rightarrow$ will be done by 4.08
- The scalability test:
  need a week (4.15)
- Replication: Which sites will participate in the Data Handling test?
  $\rightarrow$ to be discussed(KISTI, Melbourne, ?)
- Plan:
  Time-line and personnel resources to complete the Data Handling test.
  $\rightarrow$ Trying to finish the middle of April.

Resources and used data

- The data is based on case A (old data).
- Used data;
  experiment:exp07-exp65
  data type :on_resonance
  type :uds, charm, charged, mixed
  stream :0,1,2
- The data is in NSDC (KISTI).
- The data transfer have finished around 80%.
- Belle library is b20090127_0910.
- KISTI data was synchronized with that of KEK. $\rightarrow$
  The directory structure and the file location are same between KEK and KISTI.
- NSDC of KISTI allocates the resources;
  CPUs : maximum 300 nodes.
  disk space : 1TB for the extraction area.
  MC storage : around 120TB

Issues

- We have two issues on the extraction.
- 1^{st}: We need to process whole event-level even if we need the information in file-level only.
- 2^{nd}: We faced a problem in processing itself.

-> Due to the issues, we delayed the time-line.

- 1^{st} issue: To process whole data

- To extract the information of file-level.
  - We try to extract the information with command of OS without processing.
    E.g.
    Directory structure,file names and so on
  - We can’t obtain some information.
    E.g.
    total number of events
    highest event number
    lowest event number
    file creation date

  -> We needed to process the full data.

- 2^{nd} issue: Processing itself

- Processing;
  - We take 4-8hr to process the data for each experiment in KEK.
  - We take 1-3days to processing the data for each experiment in KISTI cluster.
- The occupation of CPUs;
  - We use the 20 nodes in KEK.
  - We use 24 $\sim$ 84 nodes in KISTI.
  - NSDC of KISTI reports that clusters use NAS and GPFS for the storage and data I/O. However, we suspect to have the data I/O problem when we process the data in KISTI.
  - In the future, we can't use maximum 300 nodes if many processing jobs come into cluster.
- Currently, we use maximum 24 node for extraction.
- We notice that the problem is common sense.

To solve the problems

- We need to optimize the work node and SE systems.
- We develop dCache system (working on).

Summary and Next plan

1. We submitted the paper to CPC (a special section of CCP2009).
2. We re-estimate the meta-data size realistically with TDR and generate the meta-data of Belle.
3. We expect the meta-system have enough capacity to handle the Belle II data.
4. We start to test the large data handling.
5. We will finish the large data handling test until the middle of April.