Computing at Belle II

Silvio Pardi, Guglielmo de Nardo and Guido Russo on behalf of the Belle II computing group

\textsuperscript{a} INFN-Napoli Unit – Campus of M.S. Angelo – Via Cintia 4, Napoli 80126, Italy
\textsuperscript{b} Università degli Studi di Napoli Federico II – Campus of M.S. Angelo – Via Cintia 4, Napoli 80126, Italy

Abstract

The existence of large matter-antimatter asymmetry (CP violation) in the b-quark system as predicted in the Kobayashi-Maskawa theory was established by the B-Factory experiments, Belle and BaBar. However, this cannot explain the magnitude of the matter-antimatter asymmetry of the universe we live in today. This indicates undiscovered new physics exists. The Belle II experiment, the next generation of the B-Factory, is expected to reveal the new physics by accumulating 50 times more data (~50 ab\textsuperscript{-1}) than Belle by 2023. The Belle II computing system has to handle an amount of beam data eventually corresponding to several tens of PetaByte per year under an operation of the SuperKEKB accelerator with a designed instantaneous luminosity. Under this situation, it cannot be expected that one site, KEK, will be able to provide all computing resources for the whole Belle II collaboration including the resources not only for the raw data processing but also for the MC production and physics analysis done by users. In order to solve this problem, Belle II employed the distributed computing system based on DIRAC, which provides us the interoperability of heterogeneous computing systems such as grids with different middleware, clouds and the local computing clusters. Since the last year, we performed the MC mass production campaigns to confirm the feasibility and find out the possible bottleneck of our computing system. In parallel, we also started the data transfer challenge through the transpacific and transatlantic networks. This presentation describes the highlights of the Belle II computing and the current status. We will also present the experience of the latest MC production campaign in 2014.

Keywords: distributed computing, network

1. Introduction

Belle II is an international collaboration that aims at develop a large scientific program on the Physics of flavour through a new e+ e- collider named SuperKEKB, in development at the Japanese laboratory KEK (Tsukuba) [1]. It involves 93 Institutes spread over 23 countries/region for a total of 577 scientists involved in different aspects of the experiment.

SupeKEKB is an upgrade of KEKB; it plans to accumulate an amount of 50 times more data than the previous facility. The scientific program provides a seven-year roadmap to reach the maximum luminosity starting from the first scientific run, currently expected in 2017.

When working, the experiment will produce an amount of data about 20 PB/year. Up to the start of the scientific runs, at least 2 PB/year of Monte Carlo samples must be produced.
To share and process the huge amount of data coming from SuperKEKB, and to produce the required Monte Carlo simulations, a distributed computing infrastructure is needed. The Belle II collaboration is currently working to design the Computing Model that must support the collaboration for the next 10 years.

In this paper we will present the on-going activities including the resources estimation, in term of Storage, CPU and Network, the structure of the framework for job management, and the current experiences in producing Monte Carlo simulations.

The paper is organized as follows: in section 2 we briefly introduce the main characteristics of the Belle II Computing Model. In section 3 we present estimation of needed hardware resources (Storage and CPU). Then in section 4 and 5 we focus our attention respectively on data movement strategies and in the estimation of the future network requirements. In section 6 we introduce the DIRAC framework currently used by the Belle II collaboration for data production, and in section 7 we present the results of the last Monte Carlo Campaign. In section 8 we summarize our work and get the conclusions.

2. Belle II Computing Model

The BELLE II Computing model has to accomplish, in a geographically distributed environment, the following main tasks:

- RAW data processing
- Monte Carlo Production
- Physics analysis
- Data Storage and Data Archiving

The computing infrastructure must guarantee a fast and flexible access to data and to analysis tools by all scientists, distributed in different institutions. To do that, the collaboration is defining the data distribution and data replica model and all the technological components for data processing. The main paradigms for the medium and long-term strategy are: Grid, Cloud and Data Federation[3].

Currently there are several on-going activities finalized to define the main aspects of the computing model, the characteristics, and the growth profile of each subsystems. In the next sections we describe all these activities starting from the estimation of the needed resources.

3. Resource Estimation

One of the crucial aspects in designing a long-term Computing Model is the estimation of the needed resources to store and process all the data, as well as the network requirements in terms of available bandwidth among the sites involved in the distributed computation and analysis.

The main parameter used for all the estimations is the expected luminosity from which we can deduce the event rates. In figure 1 we show the prospect of luminosity expectation of SuperKEKB. The Belle II experiment plans to reach years the target value $8 \times 10^{35}$/cm$^2$/s in seven years.

![Yearly integrated luminosity](image)

Fig. 1. The integrated luminosity in seven years of data taking.

3.1. Storage and Tape

From the storage point of view, we consider two main classes of data that will be produced and stored: the RAW data, produced at KEK and the mDST data, obtained by processing the RAW data saving the reconstruction object. The mDST can be classified in four subclass as follows:

- mDST produced from experimental data
- mDST produced in reprocessing data
- mDST form Monte Carlo simulation
- mDST of reprocessed Monte Carlo

The total amount of data expected every year is a function of the event rate (see table 2) obtained multiplying the luminosity with the cross section. The event size are estimated as follows:

RAW Data event size: 300KB
mDST event size : 40KB
Multiplying the size of RAW data and mDST data for the Event rate, we can easily estimate the needs of Disk space respectively for the whole amount of RAW data produced every year and for the first subclass of mDST. Moreover, considering that we want to permanently store two full copies of the RAW data we can easily estimate the TAPE needs too.

As regarding the rest of mDST data, we have to estimate the storage requirements of the last three scenarios, taking into consideration additional information as the number of Monte Carlo streams and the number of reprocessing. The event size is fixed at 40KB.

The last contribution to take into account in estimating the storage needs is the analysis activity of the final users. This is the most difficult part of the estimation because of the chaotic nature. We estimate using the following assumptions: 500GB/analysis x 200 concurrent analyses x 50 (luminosity factor).

In figure 2 we show the graph describing the total disk space. All the details related the Storage estimation is shown in [2].

3.2. CPU Requirements

The CPU requirements are estimated in kHepSPEC, a common benchmark largely used by High Energy Community.

We base our work using the currently available Belle II software, while the CPU ratio for the MC and RAW data processing is extrapolated from Belle experience.

The input parameters estimated from Belle II software performance are:

- RAW Data processing: 45HepSPEC*s/ev
- Monte Carlo processing: 90HepSPEC*s/ev

We foresee a more frequent data reprocessing activity in the first three years of data taking, and then one reprocessing every 2 years in the second phase (see table 2).

As regarding the Monte Carlo production we consider the number of streams (produced in 11 months for year). All the details relate the CPU estimation for each subclass of data production and skimming are presented in [2].

In figure 3 we show growth of the CPU needs as a function of time.

3.3. Belle II Site Classification

The Belle II Computing Model introduces also a sites classification in order to individuate the role of each contributing data centre the following classes are defined:

- **Raw Data centre**: Stores the RAW data and performs data processing and/or reprocessing as well as participating to Monte Carlo production.

- **Regional Data centre**: Large data centre storing mDST and participating to Monte Carlo production.
MC Production site: Data centre that produces and stores Monte Carlo simulations that include:
- Grid Site
- Cloud Site
- Computing Cluster Site

At the present moment RAW data centre are KEK, for the complete dataset, and a second copy will divided among the sites of PNNL (USA), Germany, Italy, Canada, Korea and India.

All the countries are involved in the Monte Carlo production, and the number of MC samples for each site is defined in portion to the number of Ph.D. researchers.

4. Data Distribution Schema

One of the crucial aspects of the Computing Model, is design the data distribution schema that plays a key role in the estimation process of network requirements.

We have different paths for each kind of data that include: RAW, mDST and Monte Carlo.

The RAW data, are produced at KEK and then replicated in other sites in order to guarantee a second copy.

As regarding the mDSTs produced after data taking, in the current model we plan to produce the 60% of them at KEK and 40% in other RAW Data Centres. While the mDST from data reprocessing will be created in each RAW data Centre but KEK.

We plan to have three full copies of the whole amount of mDST, stored in each Regional Data Centre: One in KEK, one in USA and one in Europe (distribute among Germany, Italy and Slovenia).

Monte Carlo data can be also classified in two categories:
- mDST-MC corresponding to the processed data
- mDST-MC corresponding to the reprocessed data

All the produced mDST-MC are distributed in that way: Each MC production site sends the produced samples to two other sites belonging to two different Regional Area, in order to have three full copies distributed for the user analysis.

Figure 4 shows all the expected data flows that we expect to manage among the different sites.

5. Network requirements

Starting from the data distribution schema is possible to estimate the bandwidth needed to move data among the sites in the target time.

There are multiple issues that affect the data distribution process; one of the most influencing is the high latency environment. In the current transcontinental network infrastructure, we measure the following value of latency:

140ms JAPAN->USA
200ms EU->USA
450ms EU->JAPAN

One of the goal of the network estimation activity, is to identify the current and future bottlenecks in function of the expected traffic growth and then produces recommendations to improve the speed of specific network paths.

Figure 5 shows the expected network occupation in KEK outbound for the RAW data transmission. Under the current design, we plan to move 100% of RAW date to the USA data centre in PNNL in the first three years of data taking. Since the fourth year, the second copy of RAW data will be distributed and stored in different countries. A possible ratio is the follows:

30% in USA, 40% in EU, 30% in Asia and Canada.

In our computation we use 11 months as target time to complete the second copy of all the RAW, using for estimation of the needed bandwidth a tolerance factor of 50%.
The computation shows a peak of 15.5 Gbit/s in outbound at KEK, suggesting an update of the geographic network.

As regarding the mDST movement, we use 11 months as target time to transfer the first data, and 12 months to transfer the mDST from data reprocessing. Table 3 shows the required bandwidth to move the mDST data.

Table 3. Expected network traffic to move the mDST data

<table>
<thead>
<tr>
<th>Event Rate</th>
<th>FY1</th>
<th>FY2</th>
<th>FY3</th>
<th>FY4</th>
<th>FY5</th>
<th>FY6</th>
<th>FY7</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRO - Data (PB)</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>MRO - Reprocessing</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>MC - Stream</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>MC - MCST - Total (PB)</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>MRO - MCST - Reprocessing</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>MRO - MCST - MCST (PB)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>MRO - MCST - Reprocessing</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>MRO - MCST - MCST (PB)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

The last studied item is the Monte Carlo production, using the same target for data transfer (11 months for the MC samples related to data and 12 months in the case of data reprocessing). The table 4 shows that the bandwidth required to copy the total amount of Monte Carlo production reaches a maximum of 2Gbit/s that is one order of magnitude less than the RAW data contribution. Moreover, that traffic must be spread over all the participating sites with a relative low impact for each data centre.

Table 4. Expected network traffic to move the Monte Carlo samples

<table>
<thead>
<tr>
<th>Event Rate</th>
<th>FY1</th>
<th>FY2</th>
<th>FY3</th>
<th>FY4</th>
<th>FY5</th>
<th>FY6</th>
<th>FY7</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRO - Data (PB)</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>MRO - Reprocessing</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>MC - Stream</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>MC - MCST - Total (PB)</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>MRO - MCST - Reprocessing</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>MRO - MCST - MCST (PB)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>MRO - MCST - Reprocessing</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>MRO - MCST - MCST (PB)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Considering all the previous tables and the percentage of data that countries have to deal with, we can calculate the total bandwidth required to move or copy the relative data. The graphs in figure 6 show the expected behaviour of the in and out network traffic for each country. For the KEK outbound trend we recognize a rapid growth as a function of the event rate, in that case the RAW data transfer represents more than the 80% of the total traffic.

At year 4, we can see a rapid decrease of the Outbound traffic at PNNL, due to the slowing-down of data reprocessing rate and to the changed distribution strategy of the RAW data. The behaviour of the other sites is more linear.

The impact of the change of the RAW data distribution strategy after year 4, affects also the inbound graph, as shown in figure 6. More specifically it is evident a rapid rise of the ingress traffic at the European sites of Italy and Germany and in the other countries that work as RAW Data Centre: India, Korea and Canada.

The opposite trend of PNNL and KEK is mainly due to the fact that after year 4 PNNL will store just a 30% of data, while KEK will receive a lesser amount of reprocessed data.
5.1. Data Challenge over the 100Gbit transatlantic link.

The current network infrastructure is based on the following set of Trans-Asia and Trans-Atlantic links, shared with multiple international scientific communities:

- 2x10Gbit/s Tokyo-Los Angeles
- 2.5Gbit/s Madrid-Mumbai
- 3x10Gbit/s Amsterdam-New York
- 3x10Gbit/s Frankfourt-Washington

In prevision of the future LHC update and because the expected exploit of Big Data technologies by other applicative domains, we foresee a large growth of the traffic on these links. These additional data flows can affect the long-term Belle II strategies. To promote the update of the network infrastructure, the Belle II collaboration have participated as pilot application in testing the ANA-100 trans-Atlantic link, a temporary network path that offer 100Gbit/s connection between Amsterdam and NewYork.

The main goals of that activity were:

- Test the ANA-100 Trans-Atlantic routing
- Test/tune/profile the performance of current Belle II data transfer tools

The sites involved in the activity are PNNL, INFN Napoli, KIT and CNAF. The tests have been performed with the following tools:

- Traceroute was used to confirm the routing to each DTN
- Iperf was used to do initial network transfer rate tests
- gridftp and/or srm-copy was used to test sites
- FTS3 server at GridKa was used to schedule data transfers

The tests, performed between May and June 2014 have confirmed the good quality of the ANA-100 link, and the possibility to saturate multiple 10Gbit/s flows on multiple storage systems, among European and USA sites. The results fulfil the requirements for the future data movement strategies.

In figure 7 we show the temporary path created between the two main network providers that have setup the link: GEANT and ESnet.

6. The production tools

To simplify the Monte Carlo production and the future data processing, a central platform able to manage the complete life cycle of pilot jobs is needed. Moreover it should offer all the administration tools and monitoring services.

In order to spread the computation over a large, distributed and heterogeneous environment, the collaboration is adopting a general framework based on DIRAC (Distributed Infrastructure with Remote Agent Control) originally developed for LHCb experiment[4].

DIRAC guarantees the interoperability in heterogeneous computing environment. Thanks to the use of multiple interfaces, it allows to submit jobs to different backend including endpoints based on the European middleware EMI/ARC, the American data centres offering OSG end-point, as well as to local farms and to the new resource provisioning interfaces offered by public and private Clouds.

The current DIRAC infrastructure is largely used for the management of Monte Carlo Challenge production. In figure 8 we show the current setup that is based on a distributed installation with a master server in KEK and 4 slaves in the Belle II centres of PNNL, IHEP, KISTI and Krakow. This implementation guarantees fail over of the subcomponents WMS and DMS and the load balancing on the web interface.

Fig. 7 A representation of the ANA-100 Trans-Atlantic link.

Fig. 8 The distributed DIRAC architecture currently used by the Belle II collaboration, for the MC productions.
7. Monte Carlo production

The Belle II collaboration is currently engaged in the 4th Monte Carlo Campaign.

The past Monte Carlo production, has involved 15 Country and 27 sites, plus 2 non-Belle II computer facilities, using multiple technologies: Grid, Cloud, and local clusters.

During the 3rd Campaign, about 6 billion of Monte Carlo samples has been produced within roughly two months (from 3rd April to 30th May 2014).

![Figure 9: Disk Space in PB including all the contribution](image)

In figure 9 we show the large improvement of the last MC production as well as the previous one. The colours in the graph discriminate the contribution of each site evaluated in terms of HEP SPEC provided.

8. Conclusions

The Belle II collaboration is strongly active in defining the computing model that will support its activities in the next 10 years. Several efforts have been spent to estimate hardware resources requirements in order to dimensioning the economical efforts for the whole collaboration and to understand the current infrastructure limits.

The network estimation activity is just started. We expect that it will give important feedback to improve the trans-continental network and all the links involved to connect the Belle II data centres distributed in world wide area.

At the same time, multiple production tools have been setup, exploiting the DIRAC framework to create a first jobs pilot factory. Thanks to this system we have completed with success three Monte Carlo Campaigns, by growing the number of samples, the number of sites participating and improving the job submission system stability.

The 4th MC Campaign is just started; we aim to produce as twice as the total number of events recorded in the last campaign.

Acknowledgments

We are grateful for the support and the provision of computing resources by CoEPP in Australia, HEPHY in Austria, McGill HPC in Canada, CESNET in Czech Republic, DESY, GridKa, LRZ/RZG in Germany, INFN-CNAF, INFN-LFN, INFN-LNL, INFN Pisa, INFN Torino, ReCaS (Univ. & INFN) Napoli in Italy, KEK-CRC, KMI in Japan, KISTI GSRC in Korea, Cyfronet, CC1 in Poland, NUSC, SSCC in Russia, SiGNET in Slovenia, ULAKBIM in Turkey, UA-ISMA in Ukraine, and OSG, PNNL in USA.

We acknowledge the service provided by CANARIE, Dante, ESnet, GARR, GEANT, and NII.

We thank the DIRAC and AMGA teams for their assistance and CERN for the operation of a CVMFS server for Belle II.

References