Monitoring system for the Belle II distributed computing

Kiyoshi Hayasaka for the Belle II computing group

Kobayashi-Maskawa Institute for the Origin of Particles and the Universe, Nagoya University
Chikusa-ku Furo-cho, Nagoya, Japan

E-mail: hayasaka@hepl.phys.nagoya-u.ac.jp

Abstract. Belle II experiment is a next-generation B-factory at KEK in Japan, which will collect 50 ab^{-1} data sample for 10 years, that corresponds to about 5×10^{10} $B\bar{B}$-pair events. To handle such a huge data sample, Belle II has adopted the distributing computing. A monitoring system is necessary to operate the computing system stably and we have been developing the monitoring system for Belle II computing based on our experience we have gained through the mass production test of the Monte Carlo simulation events. In this paper, we introduce our monitoring system, especially, the one we call "active-way" monitoring.

1. Introduction

Belle II experiment is a next-generation B-factory at KEK in Japan, which will start for physics run without (with) vertex detector in 2017 (2018). For 10 years, Belle II will collect 50 ab^{-1} data sample, which corresponds to about 5×10^{10} $B\bar{B}$-pair events. To analyze this huge number of events, we roughly need to handle 1 MHS06 CPU resources, 100 PB storage for one set of raw data and 100 PB one for Monte Carlo (MC) simulation and analysis-data events, finally. Thus, we adopt distributed computing technique. As a distributing computing software framework, DIRAC (Distributed Infrastructure with Remote Agent Control) has been adopted, which is originally developed for LHCb and can handle grid, cloud and local traditional cluster resources. [1] At the present, around 40 sites participate in Belle II computing, where 25,000 concurrent jobs are handled at peak, as shown Fig. 1. To understand the system well, for example, to find a bottle-neck, to check the scalability and the stability of the central servers, we have perform the mass production test of the MC simulation events, called MC campaign, where the MC events same as that Belle II collaborators actually analyze are generated using Belle II software. On our distributed computing system, Belle II software is provided through Cern Virtual Machine File System (CVMFS). [2]

2. Experiences through the operation

We have performed MC campaign four times so far and experienced many things through the operation. In particular, we have faced several kinds of typical troubles such as problems on Computing Element (CE), on Worker Node (WN), on the central servers which execute DIRAC components and so on. When some problem happens on CE, the pilot job, which makes the Belle II job-executable environment beforehand, cannot be accepted by CE, fails to transfer payload files or keeps idle for a long period on CE. In the first case, CE is down. In the second case,

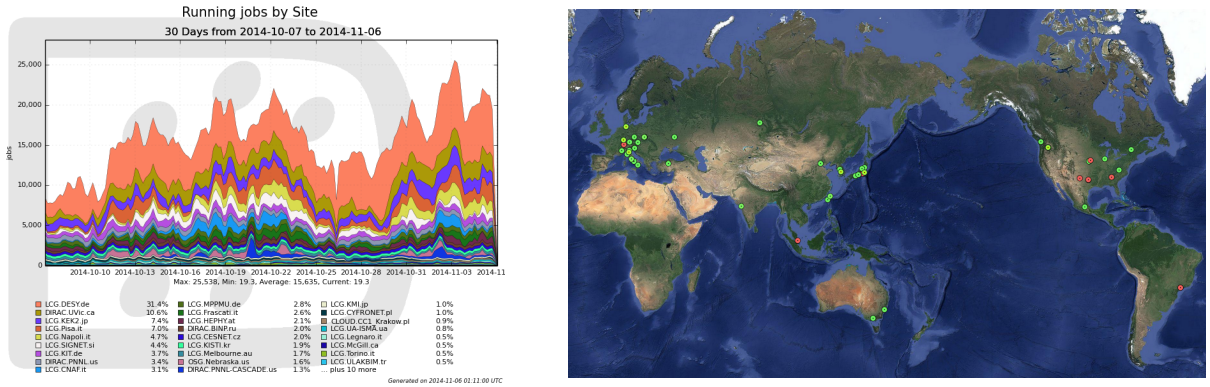


Figure 1. (Left) History for the number of the concurrent jobs on Belle II computing system from 7th Oct. to 6th Nov. in 2014. (Right) Location of sites joining Belle II computing. Green and red indicate enabled and disabled to submit a job, respectively.

CE's certificate or proxy on the pilot job is expired. In the last case, CE is alive but batch job system is down. Of course, they are the most probable causes based on our experience through the operation and there are many kinds of exception. For example, merely when the network is so busy, the payload transfer is failed.

Similarly, when a WN is problematic, the pilot job can start, but, it cannot make the environment for the Belle II jobs and finishes soon. In many cases, the WN has hardware trouble such as HDD failure. Sometimes, bad-behavior jobs remain many files and they make less disk space. Newly joined site or newly added WN tends to forget to install some required rpm packages. Occasionally, clock adjustment is out of work. Since Belle II software is provided via CVMFS on a WN, it should work well, too. Here we list up the problems relating CVMFS, that we have faced: (i) CVMFS is not mounted. (ii) CVMFS is mounted but some files on CVMFS can not be read with I/O error. (iii) Files on CVMFS can be read but they are not up-to-date. In the last case, the Belle II job can run but it may make a wrong result.

Since these problems make the efficiency of our resources low, we like to detect and fix them as quickly as possible.

3. Belle II monitoring system

However DIRAC has already had the monitoring system, we consider that it is not sufficient to detect the problems mentioned in the previous section. Therefore, we have been developing our own monitoring system. We classify the monitoring ways into two categories. The former utilizes information collected by DIRAC, makes statistics and visualizes them. While this is discussed in [3], the latter is introduced here. The active way monitors the system using several tools outside DIRAC, which probe problems on WNs, CEs and so on, and we newly develop them.

3.1. SiteCrawler

The SiteCrawler collect information on a WN such as CPU type, no. of cores, amount of the equipped memory, disk free space, OS type and version, mounted CVMFS revision, and diagnoses the WN concerning installed rpm packages, open ports, http proxy status, clock adjustment and Belle II software environment, via a DIRAC job. The job is submitted via DIRAC to each site once a hour periodically. The result is stored in DB and summarized in the web page as shown Fig. 2. SiteCrawler can detect the change of the settings such as opening ports and so on, for example, after downtime. In addition, when a new site jobs or some WNs

are added, SiteCrawler is helpful.

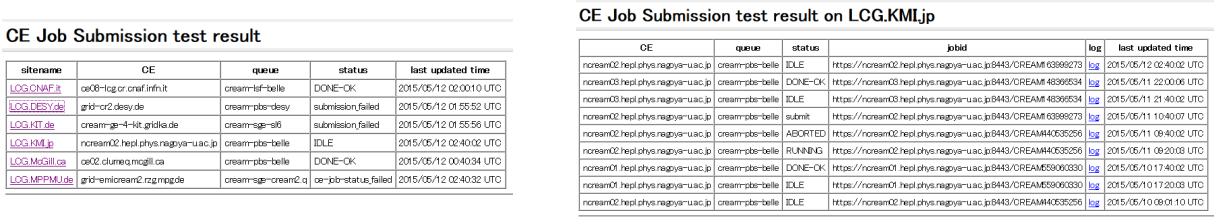
Site status summary

site	worker node	CPU	#core	memory	OS	Kernel	rpm	cvmfs	releases	CPU Norm.	last updated
CLOUD.AWS.Singapore.sg	ip-172-31-4-30	Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz	x1	3768MB	None	3.14.35-28.38.amzn1.x86_64	7 problems found	Rev. 52	OK (build-2015-02-09)	6.5 HS06	2015/04/20 20:00:48
CLOUD.AWS.Sydney.au	ip-172-31-25-11	Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz	x1	3768MB	None	3.14.35-28.38.amzn1.x86_64	7 problems found	Rev. 54	one problems found (build-2015-02-09)	6.2 HS06	2015/05/12 07:21:58
CLOUD.AWS.Tokyo.jp	ip-172-31-22-158	Intel(R) Xeon(R) CPU E5-2670 v2 @ 2.50GHz	x1	3768MB	None	3.14.34-27.48.amzn1.x86_64	7 problems found	Rev. 54	one problems found (build-2015-02-09)	6.4 HS06	2015/05/11 20:35:42
DIRAC.Beiheng.cn	hep108.buaa.edu.cn	Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz	x32	1004MB/cores	Scientific Linux release 5.8 (Boron)	2.6.18-308.1.1.el5	OK	Rev. 54	OK (build-2015-02-09)	13.0 HS06	2015/05/12 07:35:53
DIRAC.BNP.ru	0-bondar.inp.nsk.su	Intel(R) Core(TM) i5-4690 CPU @ 3.50GHz	x4	3961MB/cores	Scientific Linux release 6.6 (Carbon)	2.6.32-504.3.3.el6.x86_64	6 problems found	Rev. 54	OK (build-2015-02-09)	23.2 HS06	2015/05/12 08:46:44
DIRAC.CINVESTAV.mx	jaguar.fis.cinvestav.mx	Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz	x48	1337MB/cores	Scientific Linux CERN SLC release 6.6 (Carbon)	2.6.32-504.12.2.el6.x86_64	OK	Rev. 54	OK (build-2015-02-09)	16.2 HS06	2015/05/12 08:46:05
DIRAC.Hiroshima-HU.jp	cs12	Intel(R) Xeon(R) CPU X5550 @ 2.67GHz	x16	7277MB/cores	CentOS release 5.6 (Final)	2.6.18-238.9.1.el5xen	one problem found	Rev. 54	OK (build-2015-02-09)	11.2 HS06	2015/05/12 09:15:31
DIRAC.Niigata.jp	ngbe5.sc.niigata-u.ac.jp	Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz	x32	201.9MB/cores	Scientific Linux SL release 5.5 (Boron)	2.6.18-371.6.1.el5	one problem found	Rev. 54	OK (build-2015-02-09)	6.9 HS06	2015/05/12 08:53:10
DIRAC.Osaka-CU.jp	cpu7	Intel(R) Core(TM) i7-4790 CPU @ 3.60GHz	x8	3998MB/cores	Scientific Linux CERN SLC release 6.6 (Carbon)	2.6.32-504.8.1.el6.x86_64	OK	Rev. 54	OK (build-2015-02-09)	12.0 HS06	2015/05/12 08:58:49
DIRAC.PNNL.us	cwn40	AMD Opteron 62xx class CPU	x32	1950MB/cores	Scientific Linux release 6.5 (Carbon)	2.6.32-431.17.1.el6.x86_64	OK	Rev. 54	OK (build-2015-02-09)	3.4 HS06	2015/05/12 09:00:23
DIRAC.TIFR.in	belle.net.phop.tifr.res.in	Intel(R) Xeon(R) CPU E5-2640 0 @ 2.50GHz	x24	1336MB/cores	Scientific Linux release 6.6 (Carbon)	2.6.32-504.1.3.el6.x86_64	OK	Rev. 54	OK (build-2015-02-09)	12.1 HS06	2015/05/12 08:49:48
DIRAC.Tokyo.jp	belle2	Intel(R) Celeron(R) CPU G1620 @ 2.70GHz	x2	1965MB/cores	Scientific Linux CERN SLC release 6.6 (Carbon)	2.6.32-431.el6.x86_64	OK	Rev. 54	OK (build-2015-02-09)	11.9 HS06	2015/05/12 09:02:03
DIRAC.UVic.ca	cc-east-67070993-8636-43a9-b3d8-5cdc830ad41d3.openstacklocal	Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz	x8	3770MB/cores	Scientific Linux release 6.5 (Carbon)	3.10.64-85.cernvm.x86_64	OK	Rev. 54	24 problems found (build-2015-02-09)	11.8 HS06	2015/05/12 09:06:36
DIRAC.Yamagata.jp	serv001	Intel(R) Core(TM) i7 CPU 860 @ 2.80GHz	x8	983MB/cores	Scientific Linux release 6.6 (Carbon)	2.6.32-431.20.3.el6.x86_64	7 problems found	Rev. 54	OK (build-2015-02-09)	10.1 HS06	2015/05/12 09:55:20
DIRAC.Yonsei.kr	compute-0-1.local	AMD Opteron(tm) Processor 6380	x32	4096MB/cores	Scientific Linux release 6.5 (Carbon)	2.6.32-431.29.2.el6.x86_64	OK	Rev. 54	OK (build-2015-02-09)	8.7 HS06	2015/05/12 09:15:50
LCG.CESNET.cz	skurut21.grid.cesnet.cz	Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz	x32	201.2MB/cores	Scientific Linux release 6.2 (Carbon)	2.6.32-504.3.3.el6.x86_64	OK	Rev. 54	OK (build-2015-02-09)	8.6 HS06	2015/05/12 08:56:47
LCG.CNAF.it	wn-200-03-37-01-a.cr.cnaf.infn.it	AMD Opteron(tm) Processor 6320	x16	4025MB/cores	Scientific Linux release 6.4 (Carbon)	2.6.32-504.8.1.el6.x86_64	OK	Rev. 54	OK (build-2015-02-09)	5.3 HS06	2015/05/12 08:17:27
LCG.Cosenza.it	recas-wn-04	AMD Opteron(tm) Processor 6376	x64	4036MB/cores	Scientific Linux release 6.6 (Carbon)	2.6.32-504.12.2.el6.x86_64	one problem found	Rev. 54	OK (build-2015-02-09)	5.7 HS06	2015/05/09 19:46:51
LCG.CYFRONET.pl	ri217-g7e.zeus	Intel(R) Xeon(R) CPU E5645 @ 2.40GHz	x12	2001MB/cores	Scientific Linux release 6.5 (Carbon)	2.6.32-431.37.1.cyf1.el6.x86_64	OK	N1	OK (build-2015-02-09)	9.5 HS06	2015/05/12 08:57:56

Figure 2. Summary of the result obtained by SiteCrawler. First row shows the site name defined inside DIRAC and the second shows the latest investigated worker node's name. From the third row, the collected information and the diagnosed results on the worker node. Shaded columns indicate the site on which SiteCrawler does not execute more than 2 days. By clicking the site name, the previously investigated result on other worker nodes can be seen.

3.2. CE test job submitter

In order to check CE's health, we develop CE test job submitter. When a failure rate for DIRAC's pilot job submission on some CE is higher than 70% or SiteCrawler result is not updated more than 6 hours, this tool submits a test job to the CE and stores the submission result and the process after the submission if the submission is succeeded. In Fig. 3, the web interface of the CE test job submitter is shown. When the submission is failed, by the failure reason from the output of the submission command, we can know what happens on the CE. In the case where the job submission is succeeded but the job status can not be running, CE works but something is wrong inside the site, for example, its batch queue system is down, CRL or CA is not up-to-date. We can guess it from the status history of the job or the reason why the job is aborted. If everything goes well, it turns out that the problem is outside the CE anyway. At the present, only LCG sites are supported and the job submission test is performed by "glite-ce-job-submit". In the future, OSG and traditional cluster sites will be supported.



CE Job Submission test result

sitename	CE	queue	status	last updated time
LCG.CNAF.it	ce06-lcg.cr.cnaf.infn.it	cream-lsf-belle	DONE-OK	2015/05/12 02:00:10 UTC
LCG.DESY.de	grid-cr2.desy.de	cream-pbs-desy	submission_failed	2015/05/12 01:55:52 UTC
LCG.KIT.de	cream-ge-4-kit.gridka.de	cream-sge-sl6	submission_failed	2015/05/12 01:55:56 UTC
LCG.KMI.jp	ncream02.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	IDLE	2015/05/12 02:40:02 UTC
LCG.McGill.ca	ce02.clumeq.mcgill.ca	cream-pbs-belle	DONE-OK	2015/05/12 00:40:34 UTC
LCG.MPPMU.de	grid-emicream2.rzg.mpg.de	cream-sge-cream2.q	ce-job-status_failed	2015/05/12 02:40:32 UTC

CE Job Submission test result on LCG.KMI.jp

CE	queue	status	jobid	log	last updated time
ncream02.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	IDLE	https://ncream02.hepl.phys.nagoya-u.ac.jp:8443/CREAM163998273	log	2015/05/12 02:40:02 UTC
ncream03.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	DONE-OK	https://ncream03.hepl.phys.nagoya-u.ac.jp:8443/CREAM148366534	log	2015/05/11 22:00:06 UTC
ncream03.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	IDLE	https://ncream03.hepl.phys.nagoya-u.ac.jp:8443/CREAM148366534	log	2015/05/11 21:40:02 UTC
ncream02.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	submit	https://ncream02.hepl.phys.nagoya-u.ac.jp:8443/CREAM163998273	log	2015/05/11 10:40:07 UTC
ncream02.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	ABORTED	https://ncream02.hepl.phys.nagoya-u.ac.jp:8443/CREAM140535256	log	2015/05/11 09:40:02 UTC
ncream02.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	RUNNING	https://ncream02.hepl.phys.nagoya-u.ac.jp:8443/CREAM140535256	log	2015/05/11 09:20:03 UTC
ncream01.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	DONE-OK	https://ncream01.hepl.phys.nagoya-u.ac.jp:8443/CREAM159060330	log	2015/05/10 17:40:02 UTC
ncream01.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	IDLE	https://ncream01.hepl.phys.nagoya-u.ac.jp:8443/CREAM159060330	log	2015/05/10 17:20:03 UTC
ncream02.hepl.phys.nagoya-u.ac.jp	cream-pbs-belle	IDLE	https://ncream02.hepl.phys.nagoya-u.ac.jp:8443/CREAM140535256	log	2015/05/10 09:01:10 UTC

Figure 3. (Left) Sites on which some job submission recently is failed. (Right) History of the job submission test and its result. By clicking "log", the output from "glite-ce-job-submission" or "glite-ce-job-status" can be checked.

3.3. DIRAC port checker

When a central DIRAC server receives more requests for some DIRAC service such as providing the configuration than it can manage, the server will be down and the unexpected down of the

server may make a damage to the server. Besides, when some service is not supplied for the running pilot jobs, they stops. To avoid this kind of trouble, we make a load distribution of the crowded DIRAC services and monitor of ports used by the services on the DIRAC servers. When a DIRAC service is crowded, the corresponding port gets busy. Therefore, we check if the port can be open or not using Berkeley sockets. When some port can not be open, the operators are reminded by the e-mail message. The history is visualized by MUNIN [4] as shown in Fig. 4. Also, this helps to check if the load distributing works well or not.

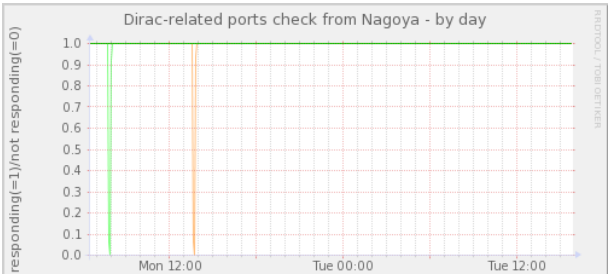


Figure 4. History of a response of ports used by central DIRAC servers, where 1 (0) means that the port can (not) response. This figure shows, around 8:00 on Monday, the port 9135 on dirac3.cc.kek.jp was busy while around 14:00 on Monday, the port 9135 on dirac4.cc.kek.jp was busy within a very short period.

4. Summary

When a huge computing resources is handled, it is important to monitor the whole system for the effective operation. Belle II computing utilizes 1 Million HS06 CPU resources and $\mathcal{O}(100)$ PB storage and adapts DIRAC as a distributed computing framework. For our operation, DIRAC does not have suitable monitoring system. Therefore, we need to develop our monitoring system based on the experience through the several MC campaigns. In this paper, we have introduced our monitoring tools. which are called the active way. They can detect several kinds of the problems on the WNs, CEs and the central servers. Near the future, we make them more automated and develop the automatic notification system to the site maintainers to operate Belle II computing more efficiently.

Acknowledgements

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 GSDC 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

[1] http://diracgrid.org/, A. Tsaregorodtsev *et al.*, J. Phys.: Conf. Ser. **119** (2008) 062048
[2] http://cernvm.cern.ch/portal/filesystem
[3] Kato-san's proceedings in CHEP2015.
[4] http://munin-monitoring.org/