Workload Analysis of GSDC Cluster Using PBS Batch System

Ilyeon Yeo, Sang Un Ahn, and Heejun Yoon

Abstract— GSDC (Global Science experimental Data hub Center) has limited computing resources, and GSDC is supporting several kinds of scientific experiments, such as ALICE (A Large Ion Collider Experiment), Belle II, Genome and so on. We have to share computing resources with community users for data analysis. Therefore, there is an issue of how to allocate resources effectively. Currently GSDC allocates resources based on MoU. For this reason, other experiment resources may not be used even if the resources for the experiment are fully utilized and new tasks must wait for a long time in the queue. In this paper, I analyzed workload of GSDC cluster that are using PBS (Portable Batch System) batch system.

Research Keywords—Workload, PBS

1 INTRODUCTION

It is very important to analyze the workload of cluster to allocate resources more effectively [1]. GSDC is supporting High Energy Physics such as ALICE, CMS, Belle II and other experiments by providing computing resources. And we are using two kinds of batch system, HTCondor and PBS [2] for job scheduling. ALICE, Belle II and Genome experiments are run using PBS batch system. ALICE and Belle II cluster system are grid cluster system and Genome cluster system is local cluster system. That means ALICE and Belle II cluster are opened to every user who is involved in corresponding community and Genome cluster is only available to about 50 domestic users approved by our system. In case of ALICE, there are about 1700 researcher from 160 institutions in 41 countries around the world. And about 700 researchers from 100 institutes in 23 countries around the world are participating in Belle II collaboration. In this paper, I focused on the workload of cluster system which is using PBS system.

2 ENVIRONMENT

GSDC has about 8,000 cores CPU, 6 PB disk and 2.5 PB tape storage in total. ALICE experiment is using about 3,400 cores, Belle II experiment is using about 300 cores and Genome experiment is using about 700 cores. As I commented, ALICE and Belle II cluster are grid system, so they can be used by all overseas users. These experiments are run under PBS batch system and Maui Scheduler [3]. Maui is a scheduling policy engine that is used with PBS batch system. PBS manages the receipt of jobs in queues and execution on cluster nodes. MAUI is a priority based scheduler but it is not event driven, it regularly polls the PBS queues to decide which jobs to run. MAUI allows to add a priority property for each queue.

3 WOLKLOAD ANALYSYS

The user's behavioral model can be obtained through an analysis of the workload. The model is indispensable for understanding how various parameters change the use of the resource center.

In this paper I analyzed job count, mean wall time, mean CPU time and mean queue time for each experiment. To get a specific data, I used PBS accounting data that was stored on the system automatically. I analyzed accounting data for 2016 for each experiment. PBS accounting data provides information about total wall time, total CPU time, the time job was queued and the time job was started. Wall time is the actual amount of time taken to perform a job. It is the sum of three terms: CPU time, I/O time, and the communication channel delay (e.g. if data are scattered on multiple machines). In contrast to CPU time, which measures only the time during which the pro-

- *Ilyeon Yeo is with the Korea Institute of Science and Technology Information, Daejeon, 34141. E-mail: ilyeon9@kisti.re.kr.*
- *Sang Un Ahn is with the Korea Institute of Science and Technology Information, Daejeon, 34141. E-mail: sahn@kisti.re.kr.*
- *Heejun Yoon (corresponding author) is with the Korea Institute of Science and Technology Information, Daejeon, 34141. E-mail: k2@kisti.re.kr.*

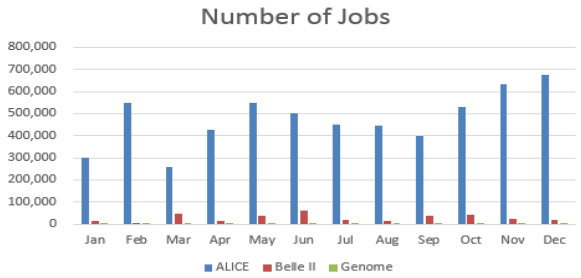

Fig. 1. Number of Jobs per month in 2016.

cessor is actively working on a certain task, wall time measures the total time for the process to complete.

Fig.1 shows the total number of jobs completed for each experiment during 2016. We allocate about 3,400 core to ALICE cluster and it is running on grid cluster system, so many jobs are performed a month. The Genome experiment uses more core than the Belle II experiment, but the total number of job is small.

Fig.2 shows the mean wall time of each experiment job for 2016. The total wall time is the sum of the wall time of each job and the mean wall time is the total wall time divided by the number of competed jobs. Genome experiment shows that the time it takes to complete each job is much longer than in other experiments.

Fig.3 shows the mean CPU time of each experiment job for 2016.The total CPU time is the sum of the CPU time of each job and the mean CPU time is the total CPU time divided by the number of competed jobs. Generally, CPU time is measured to be less than wall time, but in genome experiments, multithreading jobs are performed, so CPU time is longer than wall time.

Fig.4 shows the mean queue time of each experiment job for 2016. The queue time indicates how long the job remains in the queue before the actual job starts. We can get the queue time by subtracting the

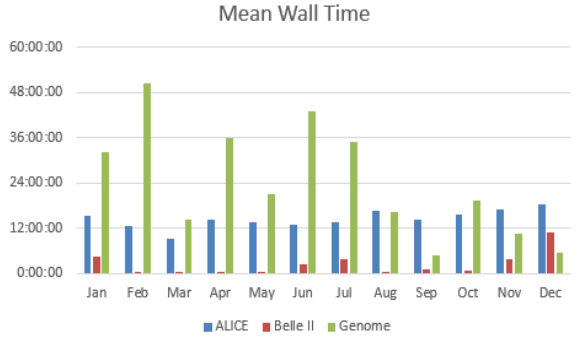

Fig. 2. Mean Wall Time per month in 2016.

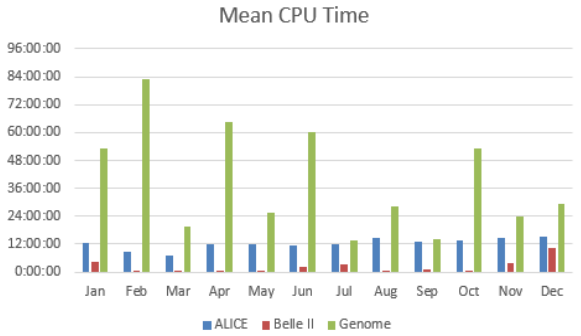

Fig. 3. Mean CPU Time per month in 2016.

time the job has been queued from the time the job is started. The total queue time is the sum of queue time of each task and the mean queue time is the total queue time divided by the number of competed jobs. As can be seen from this figure, the mean queue time for ALICE experiments is longer than for other experiments. That means, cluster for ALICE experiment is fully used.

GSDC is attempting to build an integrated batch system to jointly utilize resources for experiments running as HTCondor-based local cluster system. This allows you to dynamically allocate resources to an entire resource by adjusting the priority of the resource. However, it is not easy to change an experiment running in the PBS batch system environment to an HTCondor environment. Furthermore, experiments such as the ALICE and Belle II experiments run on a grid cluster system are impossible without the help of the global community.

4 Conclusions

So far, we have analyzed the workload of each experiment based on PBS batch system. In the case of ALICE experiment, cluster utilization rate was high and queue time was longer than Belle II and Genome ex-

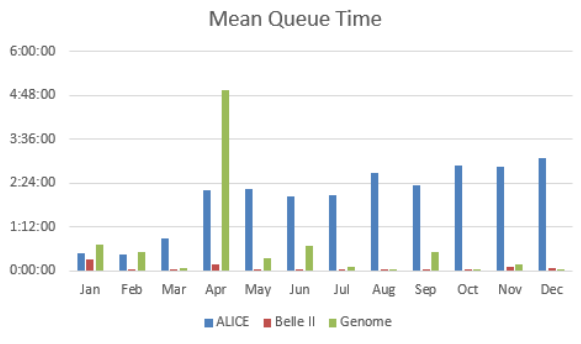

Fig. 4. Mean Queue Time per month in 2016.

periments. To solve this problem, it is necessary to study the dynamic allocation of resources based on fair-share like the HTCondor-based integrated resource management system studied in GSDC.

ACKNOWLEDGMENT

This work was supported by the Program of Construction and Operation for Large-scale Science Data Center funded by KISTI and by the Program of the Global hub for Experiment Data of Basic Science funded by the NRF.

REFERENCES

[1] Dror G. Feitelson, "Workload modeling for performance evaluation," LNCS, vol. 2459, pp. 114-141, Sep. 2002.

[2] J.P. Jones, "PPB-Portable Batch System," Beowulf cluster computing with Linux, MIT Press, pp. 356-367, 2001.

[3] B. Bode, D.M. Halstead, R. Kendall, Z. Lei, W. Hall, and D. Jackson, "The Portable Batch Scheduler and the Maui Scheduler on Linux Cluster," Proc. Fourth Ann. Linux Showcase & Conference, 2000.

Ilyeon Yeo received the B.S. and M.S. degrees from the School of Electronic Engineering at Kyungpook National University, in 2000 and 2002, respectively. He has been serving as a Senior Researcher of Global Science experimental Data hub Center at Korea Institute of Science and Technology Information (KISTI) since 2012. He served as a Senior Researcher of Knowledge Information Center and National Science and Technology Information Service at KISTI since 2002. His research interests include Parallel Computing, Information Retrieval, Database, Grid Computing, and Security.

Sang Un Ahn received a Ph.D. in Subatomic Physics from Blaise Pascal University in 2011 and in Nuclear and Particle Physics from Konkun University in 2012. He is currently a Senior Researcher in the department of Global Science experimental Data hub Center at KISTI (Korea Institute of Science and Technology Information). His research interests are workload scheduling and processing, data management and policy, and quantum computing.

Heejun Yoon received the M.S from the Computer Engineering at Chungnam National University, KR in 1997. He has been serving as a Senior Researcher at Korea Institute of Science and Technology Information since 2000. His research interests include Large data processing, Parallel Computing, Cyber-Infra system, Computer Education.