PROCEEDINGS OF SCIENCE

Introduction to GSDC project and activities

Beob Kyun Kim[1]
Korea Institute of Science and Technology
245 Daehangno, Yuseong-gu, Daejeon, 305-806, Republic of Korea
E-mail: trugens@gmail.com

Sul-Ah Ahn
Korea Institute of Science and Technology
245 Daehangno, Yuseong-gu, Daejeon, 305-806, Republic of Korea
E-mail: snowy@kisti.re.kr

Tatyana Khan
Korea Institute of Science and Technology
245 Daehangno, Yuseong-gu, Daejeon, 305-806, Republic of Korea
E-mail: tanya@kisti.re.kr

Haengjin Jang
Korea Institute of Science and Technology
245 Daehangno, Yuseong-gu, Daejeon, 305-806, Republic of Korea
E-mail: hjjang@kisti.re.kr

The GSDC project founded by the Korean government promotes global collaboration in data intensive researches by providing computing and storage infrastructure, and technical support. In this project, we support ALICE, CDF, Belle, LIGO, STAR, Neuroscience, and a few small experiments. The activities of GSDC project on ALICE Tier1 testbed, ALICE Tier2, OSG integration to support CDF, Belle and Belle2 support are presented in this paper. Moreover, the future plan and work is outlined.

The International Symposium on Grids and Clouds and the Open Grid Forum
Academia Sinica, Taipei, Taiwan
March 19 - 25, 2011

[1] Speaker

1. Introduction

Since 2009, the Korean government has established a new master plan on science and technology development. The promotion of data intensive research in cyber-environment became one of the important issues. By this plan, the Global Science Data Center (GSDC) project is assigned to the Korea Institute of Science and Technology Information (KISTI). The activities of this project include support of data intensive research with large-scale computing and storage facilities, technical support, and pilot research on these infrastructures.

In this paper, an introduction to the GSDC project and the technical details of its facilities will be provided. In addition, the future plan to extend capacity and support will be presented.

2. Activities of GSDC

GSDC's mission, given by the government, is the promotion of data intensive researches by providing a cyber-environment. The cyber-environment includes computing and storage resources, as well as technical support. The GSDC project has started in 2009 with ALICE[1], Belle[2], and CDF[3].

For ALICE, KISTI established the ALICE Tier-2 center in 2008 and since Feb. 2009, the site availability is around 98%. At the moment, GSDC is trying to extend its activity to it's Tier-1 center. As a preliminary Tier-1 center, a testbed was setup and successfully put in production in 2010. In addition to this, in order to maximise Tier-1's effect, the ALICE Analysis Facility (AAF)[4] was set up and tested. It was named KISTI Analysis Facility (KiAF), and it will provide extra research opportunities to ALICE researchers around the Asian area. In 2010, KISTI became an ALICE Associate Member to establish stable communication channels with CERN and ALICE researchers. 128 cores and 50TB, 144 cores and 200TB, and 48 cores and 20TB are allocated for Tier-2, Tier-1, and KiAF, respectively.

Belle data from KEK and computing environment serve Korean researchers. Since 2009, the grid site that enabled Belle VO has been setup and Belle MC data production has been done. Belle data at GSDC are available on local access basis and grids simultaneously. For the support of Belle2 as an extension of Belle support, grid resources are used as a test infrastructure. Since the Belle2 project is planned to run on grids - unlike Belle, that uses basic local computing - there is a great need in development of new software. Therefore, GSDC's infrastructure on grids is utilised for Belle2 software development including large-scale data handling with AMGA[5] and distributed computing. 104 cores and 200TB are allocated to support both Belle and Belle2 experiments.

Since 2009, GSDC enabled specially devoted cluster to support CDF experiment on OSG[6]. GSDC resources became the first x86_64 platform used outside of the US for CDF Software tests. Recently, from the second half of 2010, CDF decided to reprocess data to improve tagging efficiency. For this work, KISTI (GSDC) and CNAF (INFN) work together as the reprocessing offsite. 404 cores and 200TB are allocated for CDF and CDF data reprocessing.

There are more experiments supported by GSDC: the LDG (LIGO Data Grid) testbed was setup and run for LIGO; for neuro science, local cluster is allocated and is being used for brain image analysis to detect Alzheimers. Hanyang University in Korea is working on this project with McGill University in Canada as an extension of CBRAIN[7].

While clients are being asked to use grid interfaces with GSDC resources, there are communities that still use local access. In order to keep resources secure, GSDC is working on the development of SPHINX that provides token-less one time password generation via a graphical interface. This solution will be officially integrated in 2011.

3.Infrastructures

Most of GSDC's first members worked for ALICE Tier-2 center at KISTI. Both ALICE and Belle experiments require to run on LCG, therefore, GSDC resources are operated basically with LCG. The running grid middleware services from LCG are gLite-UI, WMS, LB, Apel, BDII, VOBOX, LFC, lcg-CE, CREAMCE, SEDPM, VOMS, MyProxy, and pure xrootd. For the monitoring of resources, Nagios and SMURF are used. For the management automation, puppet is used with operating system and middleware repository. Basically, all hosts are connected to each other with a dedicated 1Gbps network except storage services that include disk pool and xrootd servers and use a dedicated 10Gbps network. The main switch is connected to the Global Ring Network for Advanced Application Development(GLORIAD). GLORIAD provides a bandwidth of 10Gbps to Europe, the US, Russia, China, as well as Korea. While operating with GLORIAD, the Korea Research Environment Open Network(KREONET) is attached to these resources in 10Gbps to most Academic and Research institutes in Korea.

To support CDF, GSDC has a number of OSG grid services including osg-CE and GUMS. Unfortunately, since GSDC resources mostly use LCG, the GSDC technical staff has had to procure compatibility on LCG enabled resources.

GSDC has around 900 cores and 850TB disks at the moment and will increase two times within this year. IBRIX from HP is used as a parallel filesystem to serve smooth data access. All data access is provided by this filesystem, including grid services with SEDPM and xrootd, data access via metadata service with SAM. For light-weighted services, like MyProxy, VOMS, LFC, LDG master, and GUMS, virtualization technology is used. Because of the very expensive license for virtualization, open-source cloud solution is being tested.

4. Future plan and ongoing projects

4.1 Experiments

As a preliminary Tier-1 for ALICE, GSDC will do collaborative processing of ESD data with Gangneungwonju National University in Korea and INFN in Italy. Currently used resources for KiAF have small amounts of main memory and disk space, therefore, it cannot support users' jobs sufficiently. New resources purposely specified for Analysis Facility will be setup and work for ALICE researchers in Asia. KiAF with new resources will be tested by Korean researchers first and will become available to all ALICE users at a later time.

Full-recon data analysis from Belle will be performed by collaborative research with Yonsei University and KEK. For this work, data is reconstructed at KEK and transferred to GSDC, where analyses work can be done much faster. Belle2 software development will be supported continuously and GSDC will work as a regional center for Belle2 once it starts.

Offsite CDF data reprocessing will run at GSDC continuously and GSDC will run CDF analysis jobs later on. Moreover, software development on cloud computing and grid information services will be studied in collaboration with FNAL.

Markov-chain Monte Carlo analysis of gravitational-wave signals from LIGO will be done under GSDC resources by collaboration of Korea Gravitational Wave Group(KGWG) and LIGO. The GSDC will run as an Asian hub for STAR and extend its current support for neuro science into grids. For Reactor Experiment for Neutrino Oscilliation at Yonggwang(RENO), GSDC will serve as a kind of Tier-0 with 10Gbps network to detector facility near nuclear power plant.

4.2 Resources

The procurement policy of GSDC is different from other big computing centers. Every year, new resources are setup and extended. In a few years, the plan promotes to have more than 5,000 cores and 5PBs of disk space. In 2010, GSDC started to operate 2,000 cores and 2PBs. However, these computing resources are planned to be extended every year. Virtualization technology is expected to provide more flexibility and fail over capability. OpenNebula is being tested as an open source virtulization solution and Nimbus will be tested in the near future.

References

[1] ALICE, http://aliceinfo.cern.ch

[2] Belle, http://belle.kek.jp

[3] CDF, http://www-cdf.fnal.gov

[4] ALICE Analysis Facility, http://aaf.cern.ch

[5] S. Ahn et al, *Design of the Advanced Metadata Service System with AMGA for the Belle II Experiment*, *JKPS 57 (4)*

[6] Open Science Grid, http://www.opensciencegrid.org

[7] CBRAIN, http://cbrain.mcgill.ca