Data Grid Deployment for High Energy Physics in Japan

Hiroshi Sakamoto *

International Center for Elementary Particle Physics (ICEPP), the University of Tokyo, Tokyo, Japan

Abstract

Recent activities in Japan to deploy a data grid for high energy physics will be reported. Japanese collaboration of LHC-ATLAS experiment is now preparing a Tier-2 analysis center at Tokyo. Several institutes/universities have grid facilities and they are connected to the Tokyo Tier-2 center. Collaborators of Phenix of RHIC at BNL and CDF of Tevatron at FNAL are also using grid tools for data transmission. KEK-CRC is also hosting the Belle experiment VO and is running a site for it as a part of LCG. A general purpose virtual organization for Japanese high energy physics society is to be launched.

Key words: Data Grid, High Energy Physics, Distributed Data Analysis, Data Transmission, TCP Protocol

1. Introduction

As a general trend, recent high energy physics experiments, especially collider experiments require very high data transfer bandwidth, very high data processing speed and very large storage capacity. Recently, thanks to the growth of international network connectivity, distributed data analysis scheme is going to be employed in such experiments. Now the collaboration is truly world-wide, and the regional effort of distant countries is very crucial, in order to establish a stable and efficient data analysis infrastructure spread around the globe. In this article, Japanese activities will be reported from this point of view.

The Atlas experiment of LHC at CERN is fully utilizing data grids for its data analysis infrastructure. The experiment will start in 2007, and the commissioning of the computing system is now going on. Institute for Elementary Particle Physics (ICEPP) of the University of Tokyo is running a Tier-2 anal-

ysis facility as a member of the WLCG (Worldwide LHC Computing Grid) collaboration. As this facility locates far apart from the experimental site, special treatment is necessary for data transfer due to the long distance network connection, which gives very large round trip time (RTT) and many hops in between. Intensive study to achieve high throughput data transmission has been started. Some preliminary result will be shown later.

Running collider experiments, the Phenix experiment of BNL RHIC, and the CDF experiment of FNAL Tevatron accelerator are also partly utilizing data grid technologies. Japanese collaboration institutes of these experiments are operating local analysis facilities. Grid tools are used mainly for intensive data transfer. Their experiences are also shown.

Study of grid application to the Belle experiment of KEK-B has been just started. There are many collaboration institutes which are also members of LHC experiment collaborations, and they have computing facilities connected to some of existing grids. They want to transfer and analyze Belle data through their grids. Plan to deploy a grid infrastructure for Belle will be discussed.

Finally, a more generic virtual organization (VO),

---

* Corresponding author.
  Email address: sakamoto@icepp.s.u-tokyo.ac.jp
(Hiroshi Sakamoto).

called as "HEPnet-J grid" is going to be launched. This local VO is the one for common R&D for groups who want to experience how a grid works.

2. Atlas of LHC at CERN

The world largest collider accelerator LHC is now under construction at CERN near Geneva. At this highest energy proton-proton colliding experiments, world-wide data grids will be fully utilized for the first time. Because of the energy, the beam luminosity and also the complexity of the experimental apparatus, unprecedentedly large size of data must be processed. The total size of the data accumulated in one year becomes order 10PB (Peta-Bytes, 10^{15} bytes), and in order to process these data, more than 10 Million SI2000 processing capacity is necessary. These sizes are far beyond the possible amounts which a single institute can provide. Introduction of data grids are therefore quite natural. The LHC Computing Grid (LCG) project was established in 2002, and since then, substantial efforts have been made to deploy a usable grid infrastructure in truly world-wide scale.

2.1. The fabric

International Center for Elementary Particle Physics (ICEPP) of the University of Tokyo has been participating the LCG project from the beginning, and now it is running a Tier-2 analysis center for the ATLAS experiment. The definition of terms "Tier-0", "Tier-1" and "Tier-2" are given in the Memory of Understanding (MoU) of the world-wide LHC computing grid [1]. As a national analysis facility, ICEPP provides services for 15 collaborating institutes in Japan.

Study on the grid fabric, i.e. PC farms, disk storage, tape drives and network devices have been continuing since 2003. Specification of the current computing resources which is pledged to LCG is listed in the Table 1. In order to construct a resource center efficiently, several critical issues must be taken care of. Under a given financial conditions, we want to introduce as much resources as possible, so as to be able to contribute to the project. As the capacity of available rooms for computing resource is limited, the size of components is a big concern. Several kinds of blade type PC servers have been tested. Blade servers can be installed at higher density in a limited space. However, at the same time, the power consumption is also concentrated locally as the size of the equipment becomes small, and therefore, cooling of the system is also crucial and needs special care. Specification of our new production system has been completed based on the result of the study.

CPU	Xeon 2.8GHz
Memory	2GB/node
Disks	36GB × 2
Network	GbE NIC × 3
Enclosure	8 Brades / 6U
System	108 Brads / 3 racks

Table 1. Specification of ICEPP PC server pledged for WLCG in 2006. The value listed in WLCG MoU [1] is, 200 kSI2000 of CPU, 40TB of disk capacity and 1Gbps of WAN connection for 2006.

The new production system will be delivered in this December, which will fulfill the required amount of computing resources to be pledged in 2007 and thereafter. Its international tendering is in progress now.

EGEE gLite3 middleware [2] is installed for the present system. From the grid operation point of view, this resource center is taking care of by the regional operation center. In the Asia-Pacific region, ASGC is playing this role.

2.2. Network Connectivity

Network connectivity with other grid sites is a very important issue. Inside Japan, almost all collaborating institutes are connected via SINET, [3] Japanese national research education network operated by National Institute of Informatics. The bandwidth of the domestic connections is continuously increasing and now major collaboration institutes are connected with 1 Gbps or more. The maximum round trip time (RTT) among these institutes is about 20 ms and therefore the Atlas-Japan collaboration employed a deployment model of establishing a single analysis center (Tier2) at Tokyo and each institute has UI nodes to be connected to Tier2.

As for the international connection, we have to consider several issues arising from geographical conditions.

– Very large round trip time (RTT). Between Europe and Japan, RTT is around 300ms.
– Bandwidth to US/Europe is large. Whereas, the mutual connection inside the Asia-Pacific region is fairly thin.

According to the ATLAS computing model [4], A Tier-2 center should be associated to a Tier-1. However, from these observations, we decided to have two associated Tier-1 sites, i.e., IN2P3-CC Lyon and ASGC Taipei.

We have started connectivity study with these two sites. In both case, we have to use general purpose networks and therefore the quality of traffic is not predictable.

– ASGC Taipei: RTT between Tokyo and Taipei is around 40 ms. The bandwidth is limited to 600 Mbps.
– IN2P3CC Lyon: RTT between Tokyo and Lyon is around 300 ms. NII operates a connection between Tokyo and New York, and at New York, it is directly routed to the European academic network, Geant. The bandwidth is 10 Gbps.

Practically, the experimental data transfer is managed by a software framework named as FTS [5]. So, the performance of the FTS file transfer is our primary concern. It is ideal if the wire speed is achieved without any tunings.

Preliminary result of Taipei-Tokyo test is rather satisfactory. Figure 1 shows a snap shot of network traffic monitor taken during the FTS test. Transmission is quite stable and sustaining 40MB/sec transfer has been achieved.
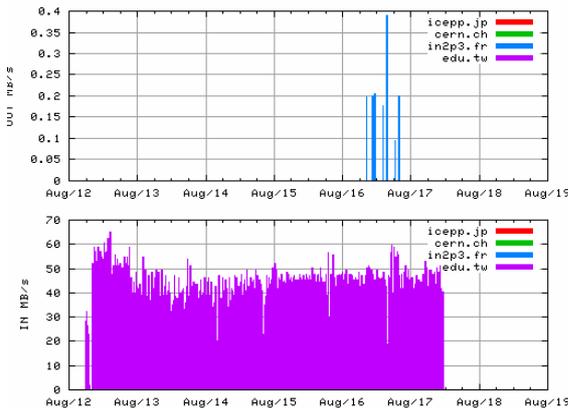

Fig. 1. A snap shot of Taipei-Tokyo data transfer tests. Network traffic monitored at ICEPP is shown. The lower figure shows incoming traffics whose source is from the edu.tw domain. Sustained transmission of more than 40 MB/s has been achieved.

In the case of Lyon-Tokyo, it is not so easy as the Taipei case. As a nature of the TCP protocol implementation, quality of network traffic is very crucial in the long distance transmission. Once traffic intervention happens, due to the large RTT, the transmission is suspended so long time. Even after the resumption, the TCP window size is reduced according to the congestion control mechanism and it takes much time to recover the maximum transfer speed. In the very preliminary test, about 10 MByte/sec of transfer speed was observed.

In order to understand the quality of data transmission between Lyon and Tokyo, several 'iperf' tests were performed. In the series of the tests,

– Two kinds of TCP protocol implementations were tested.
  · Conventional Reno TCP.
  · BIC TCP [6] introduced in recent Linux distribution.
– Effects of 'pacing' were tested. A software implementation of TCP spacing, PSPacer [7] was employed.

Network traffic has been monitored during the tests as shown in Figure 2. Preferable combination is BIC TCP and PSPacer to achieve the wire speed. In these tests, the both end-points are PCs with GbE NIC. So the absolute maximum is 1Gbps or nearly equals to 110MB/sec.

A more sophisticated congestion avoiding algorithm is employed in BIC TCP, and therefore more efficient transmission is expected for BIC TCP compared to Reno. PSPacer was turned on/off during the tests. By comparing the traffic where it was on and off, it is seen that this method is more effective when used with BIC TCP. Further study is continuing in the collaboration with IN2P3-CC people.
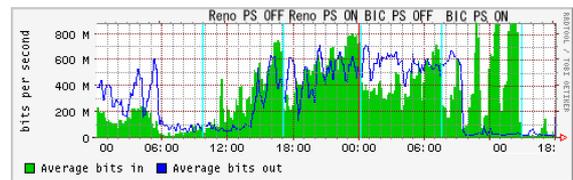

Fig. 2. Network traffic monitored at ICEPP during Lyon Tokyo iperf tests. The test was done for 4 cases, namely, standard Reno TCP with and without PSPacer, and BIC TCP with and without PSPacer. Combination of BIC TCP and PSPacer shows the highest performance in the iperf transmission. At the peak, incoming data rate reached to 1Gbps.

3. Phenix of RHIC at BNL

In Japan, there are also collaborating institutes of the Phenix experiment of RHIC in BNL. RIKEN is now running Japanese analysis center (CCJ) for Phenix. Until a few years ago, they used to transfer data in forms of tape media shipped to Japan. After the JP-US connection was upgraded, they started to try to transfer their data by using grid-FTP.

Figure 3 shows a snapshot of network traffics monitored at RIKEN. Since 2005, Phenix data are all transferred via the international network. This is a remarkable achievement of an application using grid tools [8].
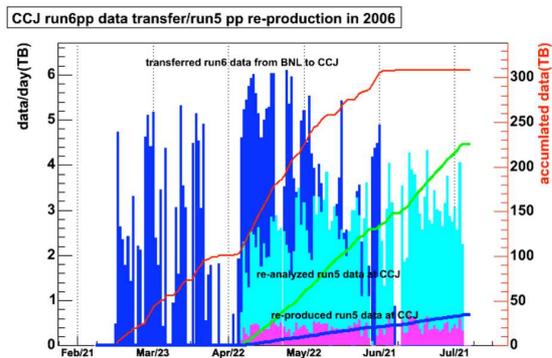

Fig. 3. Daily statistics of transfer and re-processing of Phenix data at CCJ RIKEN. They achieved 6 TB/day data transfer. This plot is taken from T. Ichihara and Y. Watanabe [9]

4. CDF of Tevatron at FNAL

As for the CDF experiment of Tevatron at FNAL, the University of Tsukuba is running a regional analysis facility (DCAF: Decentralized analysis facility), named as JPCAF. Data transfer between FNAL and Tsukuba is also managed using grid tools. They offer about 400GHz of CPU and 10TB of disks and connected to SINET via 1Gbps bandwidth of MPLS-VPN.

They are preparing to introduce LCG middleware, as they are also participating to the Atlas experiment and have intention to utilize their facility for Atlas data analysis.

5. Belle of KEK-B at KEK

The Belle experiment of KEK-B at KEK is very characteristic experiment and it has been running since 1999 gathering very huge amount of data. It's luminosity is increasing gradually, and also they have a plan to upgrade the accelerator to 10 times intense one. At such a high luminosity, consequently at such high rate, data analysis scheme should be reconsidered.

Distributed data management is not new for them. MC production has been done by many collaborating institutes and the results are all sent to KEK for common use. They had deployed SRB [10] among collaborating institutes around the world.

However, the total amount of real and simulation data will increase rapidly in near future, and therefore, they start considering to employ a grid technology for their production system.

The Belle experiment body is of course a large international collaboration. Among the collaborating institutes, many are also committed to LHC experiments. Therefore, some such sites have already deployed data grids for LHC. This is also a strong pressure for Belle to start the program.

Computing Research Center of KEK started research and development on the grid usage. They provided three grid sites for the R&D.
- testbed
- preproduction system
- production system

Composition of the production system is shown schematically in Fig. 4. This production system and the preproduction system are registered at EGEE-GOC.

As mentioned before, they have a rich experience of deploying SRB federations [11]. In their R&D program, study on interoperability between the LCG middleware and SRB is an interesting subject.

Recently they held a meeting dedicated to coordinate the introduction of a data grid for them. Virtual organization of Belle has been already established and its VOMS is running at KEK. As the next step, the implementation of Belle VO support at each site is ongoing.

6. Summary

There are many high energy physics programs and international collaborations are running in Japan. In this report, experiences at LHC-Atlas, RHIC-Phenix, Tevatron-CDF and KEKB-Belle are shown, where fully or partly the grid technology has been introduced and being utilized actively.
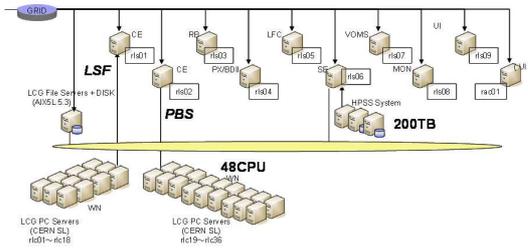
On the other hand, there are also many groups

Fig. 4. Schematic view of KEK production system JP-KEK-CRC-02. This system has been running since 2006. This diagram is provided by G. Iwai et al.[12].

who have interest on this technology, but have not yet taken actions. A discussion to establish a general purpose VO for training and exercise was triggered. A VO named as HEPnet-J will start soon as an incubator grid testbed for newcomers. A nation-wide meeting was held at Tokyo in July to discuss how to launch the HEPnet-J VO. People from various high energy physics programs, Atlas, Belle, CDF, Phenix, ILC, K2K etc, attended the meeting.

The "Cyber-Science Infrastructure (CSI)" project has started in Japan, which includes upgrade of academic network to SINET3 [3] and Japanese national grid activity NAREGI [13]. The HEPnet-J VO will be operated as a part of the CSI activity. KEK is now participating to the NAREGI project, in which high energy physics is also taken as a leading application. Interoperability of NAREGI with other running grids is their strong concern and a collaborative work with EGEE has been started.

Acknowledgements

References

[1] "Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid", David Jacobs, CERN-C-RRB-2005-01/Rev., 2005.

[2] "EGEE gLite-3 Middleware", http://glite.web.cern.ch/glite/packages/R3.0/.

[3] "Science Information Network", http://www.sinet.ad.jp/english/index.html.

[4] "Atlas Computing Technical Design Report", Atlas Collaboration, CERN-LHCC-2005-022, 2005.

[5] "File Transfer Service", https://twiki.cern.ch/twiki/bin/view/EGEE/FTS.

[6] "BIC TCP - a TCP variant for high-speed Long Distance Networks", http://www.csc.ncsu.edu/faculty/rhee/export/bitcp/.

[7] "PSPacer version 1.2", http://www.gridmpi.org/pspacer-1.0/index.en.jsp.

[8] "PHENIX experiment uses Grid to transfer 270TB of data to Japan", CERN Courier Volume 45 Number 7 September 2005, p15.

[9] T. Ichihara and Y. Watanabe, private communication.

[10] "SRB - SDSC Storage Resource Broker", http://www.sdsc.edu/srb/.

[11] "SRB SYSTEM AT BELLE/KEK", Y. Iida et al., Computing in High Energy Physics 2004, Interlaken, 29 Sep. 2004. ID=216.

[12] G. Iwai, Y. Iida, S. Kawabata, T. Sasaki, Y. Watase, private communication.

[13] "NAREGI - National Research Grid Initiative", http://www.naregi.org/index_e.html.