Data Processing for 'SUBARU' Telescope using GRID

Y. Shirasaki M. Tanaka S. Kawanomoto S. Honda M. Ohishi Y. Mizumoto

National Astronomical Observatory of Japan, 2-21-1, Osawa, Mitaka, Tokyo 181-8588, Japan

N. Yasuda

University of Tokyo, 5-1-5 Kashiwa-no-Ha, Kashiwa Chiba, 277-8582 Japan

Y. Masunaga

Ochanomizu Univerisity, 2-1-1 Otsuka Bunkyo-ku, Tokyo, 112-8610 Japan

Y. Ishihara J. Tsutsumi

Fujitsu Ltd., 4-1-1 Kamikodanaka Nakahara-ku, Kawasaki, 211-8588 Japan

H. Nakamoto Y. Kobayashi M. Sakamoto

Systems Engineering Consultants Co. Ltd., 22-4 Sakuraoka-cho Shibuya-ku, Tokyo, 150-0031 Japan

Abstract

The amount of astronomical data is rapidly growing as the progress of telescope technology and the detection technique of recent years. As a result, the way of traditional analysis is becoming insufficient for utilizing the large amount of data effectively and efficiently. The SuprimeCam, which is one of the instruments equipped with the Subaru telescope, has been generating 6 TB of public data since its start of operation. It is almost impossible to transfer all the data to the local machine to analyze them. It is, therefore, desirable to have an environment where the data is analyzed where it is stored. In addition, it is not easy for an researcher who is not familiar with the Subaru data to reduce the Subaru's raw data. To overcome these difficulties, we have applied grid technology to construct a system that carry out multiple jobs on multiple servers. Integrating this system to the Japanese Virtual Observatory, a user can easily utilize the grid system through the web browser interface.

Preprint submitted to Elsevier *30 May 2007*

Key words:
GRID, Astronomy, Virtual Observatory, Web

1 Introduction

Thanks to the progress of telescope technology and the detection technique of recent years, it is expected that we will meet with a situation where a large scale of high-quality data is continuously generated. The way of traditional analysis, however, is becoming insufficient for using the large amount of data effectively and efficiently, and for getting the maximum scientific results. Although many astronomers recognize the importance of research that uses the multi-wavelength data, such research actually needs considerable effort. One reason is that, for each data set, one needs to learn how to reduce and analyze the data, and even needs to know where the analysis tools are available (see top panel of Figure 1). To overcome such situation and maximize the scientific return from a big project like Subaru and ALMA, it is important to construct an environment where user can access to the science-ready data with very few effort (see bottom panel of Figure 1). National Astronomical Observatory of Japan (NAOJ) started its VO project (Japanese Virtual Observatory – JVO [1]) in 2002. The objectives of the JVO project are to provide a seamless access to the distributed data service of the world, and to provide user-friendly analysis environment. To realize the interoperability among the astronomical databases of the world, it is crucial to define a world standard of the access protocol for the databases. The International Astrnomical Ob-

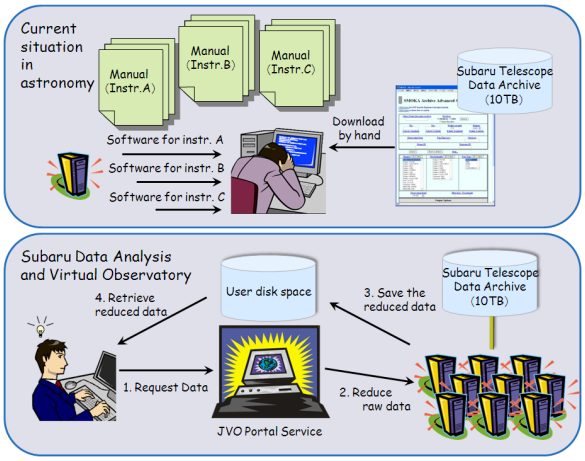

Fig. 1. Current situation of astronomical study (top) vs astronomy of the future (bottom).

[1] http://jvo.nao.ac.jp/

servatory Aliance (IVOA [2]) was formed in 2002 for that purpose. The IVOA now comprises 16 VO projects from Armenia, Australia, Canada, China, Europe, France, Germany, Hungary, India, Italy, Japan, Korea, Russia, Spain, the United Kingdom, and the United States. This paper describes our recent progress on the data analysis environment on JVO.

2 Subaru Data Archive

Subaru [3] is an optical-infrared 8.2 m telescope operated by National Astronomical Observatory of Japan (NAOJ) at Mt. Mauna Kea, Hawai. Subaru has seven open use instruments: CIAO, COMICS, FOCUS, IRCS, SuprimeCam, HDS and MOIRCS. Using these instruments, observation can be made for wavelengths from optical (300 nm) to infrared (20 μm) with spectrum resolution up to 10^5 (HDS).

As of October 2006, 8 TB of data is archived in the public area. More than 70% of the data are from the SuprimeCam, which is a mosaic of ten 2048 $\times$ 4096 CCDs and covers a 34' x 27' field of view with a pixel scale of 0.20" (Figure 3). More than 90% of all the data requests are for the SuprimeCam, so our current priority issue is how to improve the usability of the SuprimeCam data. There is a plan to upgrade the SuprimeCam to the Hyper SuprimeCam (HSC) in 2011. The HSC will have ten times larger detection area than the SuprimeCam, and is expected to generate data in ten times hihger rate.

The data are registered in the Subaru Telescope Archive System (STARTS) in Hawai as soon as the data are acquired by the instruments, so an observer retrieves his data from STARS during and/or after the observation. The data

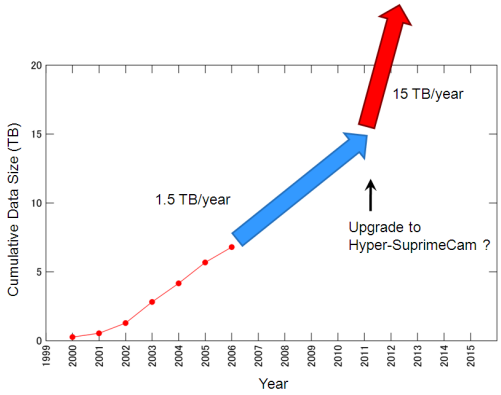

Fig. 2. Cumulative totla data size of the Subaru SuprimeCam.

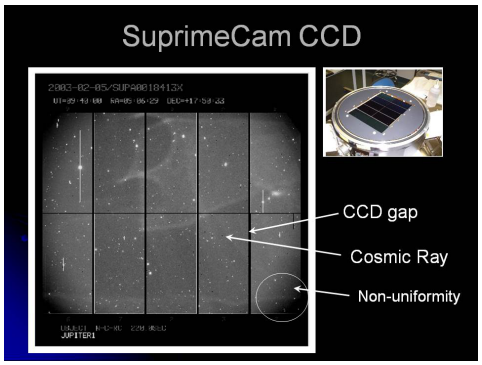

Fig. 3. Subaru SuprimeCam CCDs and an image taken by the SuprimeCam.

[2] http://www.ivoa.net/
[3] http://subarutelescope.org/

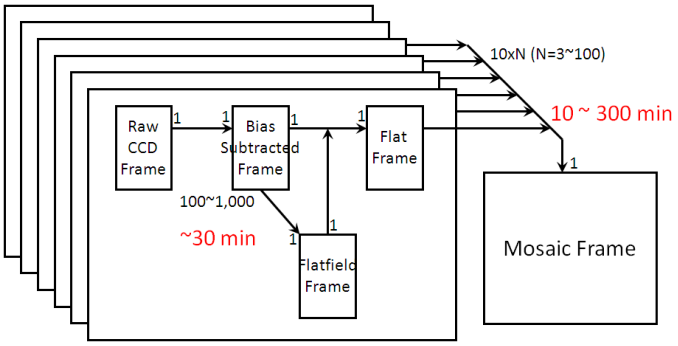

Fig. 4.

of STARS is mirrored to the Mitaka Advanced STARS (MASTARS) in Japan, so the observer can retrieve the data also from the MASTARS when he returns to Japan. STARTS and MASTARS are not public data archive. To use the system, you need to get an account on the Subaru computing system for STARS or an account on the Mitaka computing system for MASTARS. The data that passed 18 months of a proprietary period becomes publicly available through the SMOKA [4] and JVO. system. The SMOKA system provides various query modes for the Subaru archive. The JVO system provides IVOA standard access interface to the Subaru archive.

3 Data Reduction of SuprimeCam

The amount of data, especially of SuprimeCam, is very large, so it is important to provide a way to analyze the data without moving the data to a users' machine. One of the ways to do so is to login to the Subaru or Mitaka computing system and analyze the data on the machine. It is, however, not practical to use the visualization tool from a remote machine, especially when accessing through a slow network, and is also not easy for a novice user to analyze the Subaru data with command line user interface. Another solution is to provide a web service, through which one can access to the data analysis software and visualize the data in a compact graphical format such as GIF, JPEG and PNG. Recently a lot of open source framework are available for making such a service, and it has been realized that interactivity of the web based service can be improved by using Ajax technique as demonstrated by the google map service.

To reduce the data of Subaru SuprimeCam, a lot of computing resources are required. To make one final mosaic image from raw data, flat-field frames must be prepared for each CCD (Figure 4). The flat-field frame is derived from hundreds of observation frames by calculating their median values for each

[4] http://smoka.nao.ac.jp/

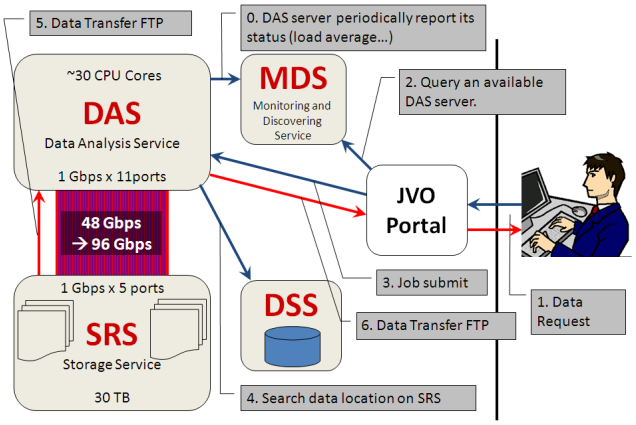

Fig. 5. Subaru data reduction pipeline architecture

2k×4k pixels. It typically takes two hours for creating one flat-field frame, and at lease ten flat-field fames are required to make the final mosaic image. Thus it takes 20 hours for prepareing the flat-field frames, if the calculation is perfromed on just one CPU. After all the raw data are corrected by using the flat-field frames, we need to coadd the frames to remove the gaps between the CCDs, to remove cosmic rays event, and to enhance the signal to noise ratio. It takes one hour for typical observations.

Another important point is to provide the data in a form that detector and environment dependencies are removed. Those dependencies are usually not known by an archive user, so it should be properly reduced by a data provider. The quality of the data reduction is improved as the experience is accumulated and the reduction software is also evolves continuously, so it is adequate to reduce the data on demand with the most developed algorithm.

4 Grid Computing System

It is expected that several tens user will submit data reduction jobs, and the number of users may increase unexectedly. So the computing resources need to be integrated in a scalable manner. We have developed a Web service based grid computing system. This system is composed of four services: Monitoring and Discovering Service (MDS), Data Analysis Service (DAS), Data Search Service (DSS) and Storage Service (SRS). Figure 5 shows a diagram of the Subaru SuprimeCam analysis system. The MDS manages the status of each DAS server. Each DAS server periodically reports its status, such as load average, number of running job and job status, and the status is stored on an MDS database. A GRID client queries to the MDS to ask which server is free for a job submission. The MDS returns a service endpoint URL that is appropriate for the job submission. The client submit the job, and wait it to finish. While waiting, the client periodically polling to the MDS for querying

the job status. The DAS sends a query to the DSS to get Subaru RAW data, the DSS returns the URLs for the data, then the DAS retrieves the data and starts analysis. When a job status is changed to "finish", the client queries the DAS for an URL to retrieve the result. The URL is passed to the SRS, the result is stored on the storage of the SRS, and the metadata of the result is registered on the DSS.

We have integrated the GRID computing system to the JVO portal [5] so that user can easily access to the computing resources through a well-familiar interface, Web browser. The details of the JVO system were described in elsewhere (Shirasaki et al. 2006, Shirasaki et al. 2006b, Tanaka et al. 2006, Ohishi et al. 2006).

5 Summary

The construction of the GRID-based Subaru analysis service has just started last year (2006). Currently the data of SuprimeCam is a primary target of the development, but data of other instruments will be available. The grid computing system are constructed to obtain enough computing resource to analyze all the SuprimeCam data in reasonably short time. Using the system, we have demonstrated that all the SuprimeCam data can be reduced in less than one week. If the same thing is done without this sytem, it will take more than one year to accomplish the work. The JVO portal service is available for everyone at `http://jvo.nao.ac.jp/portal`, although currently only the limited functionality is publicly available. To use all the JVO functionality, user registration will be required.

References

Tanaka, M. et al. , 2006, in ASP Conf. Ser., Vol. 351, Astronomical Data Analysis Software and Systems XV, ed. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique (San Francisco: ASP)

Shirasaki, Y. et al., 2006, Advanced Software and Control for Astronomy. ed. Lewis, Hilton; Bridger, Alan. Proceedings of the SPIE, 6274, p42

Shirasaki, Y. et al. , 2006, in ASP Conf. Ser., Vol. 351, Astronomical Data Analysis Software and Systems XV, ed. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique (San Francisco: ASP)

Ohishi, M. et al. , 2006, in ASP Conf. Ser., Vol. 351, Astronomical Data Analysis Software and Systems XV, ed. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique (San Francisco: ASP)

[5] http://jvo.nao.ac.jp/portal