Super-Kamioka Computer System for Analysis

● Akira Mantani ● Yoshiaki Matsuzaki ● Yasushi Yamaguchi
● Kouki Kambayashi

(Manuscript received April 7, 2008)

The Institute for Cosmic Ray Research (ICRR) of the University of Tokyo newly established the Kamioka Observatory in 1996, and has continued to observe the elementary particles known as neutrinos by using the Super-Kamiokande Neutrino Detection Equipment. This equipment contains a 50 000-ton ultrapure water tank measuring 39.3 meters in diameter, 41.4 meters in height, and located 1000 meters underground. A total of 11 129 photomultiplier tubes (PMTs, 50 cm in diameter) are mounted on the inner wall of the tank. The Kamioka Observatory continues observation 24 hours a day, 365 days a year in order to detect neutrinos observable only for 10 seconds from a supernova explosion which may occur once every dozens of years. The current size of total accumulated data is nearly 350 TB. In February 2007, Fujitsu installed a computer system (known as the "Super-Kamioka Computer System for Analysis") using mass storage disk drives for saving and rapidly accessing the observed data. This paper describes the configuration of the Super-Kamioka Computer System for Analysis, explains how data is managed and rapidly accessed, and how throughput performance is improved.

1. Introduction

The Kamioka Observatory of the Institute for Cosmic Ray Research (ICRR) of the University of Tokyo[1)] has been conducting research on elementary particle physics through neutrino detection and nucleon decay searches. Super-Kamiokande[2)] (**Figure 1**) was constructed as neutrino observation equipment in 1996, and enabled the discovery of the phenomenon of neutrino oscillations in 1998. Super-Kamiokande subsequently achieved other impressive results such as proving that neutrinos have mass, and is now being used to solve such cosmic mysteries as the birth of black holes and stars. Neutrino observation entails the storage of about 50 GB of raw data a day, and the total size of currently accumulated data is almost 350 TB (including the size of processed data). In February 2007, we replaced the conventional, hierarchical tape-drive storage management system with a new storage system offering faster access based on Fujitsu's ETERNUS4000 model 500 magnetic disk drive system. The Kamioka Observatory must analyze past data using newly changed parameters. Given the need for faster data access in such analysis, Fujitsu's Parallelnavi Shared Rapid File System (SRFS) was installed.

This paper outlines the observation of data, describes the analysis system, and explains how high-speed data access is acquired.

2. Outline of data observation

The Super-Kamiokande detector consists of a cylindrical stainless-steel water tank (39.3 meters in diameter, 41.4 meters in height, and designed to hold about 50 000 tons of ultrapure water) and 11 129 photomultiplier tubes (used as photosensors) mounted on the inner wall of the tank.

Super-Kamiokande is located 1000 meters underground in the Kamioka Mine in Gifu Prefecture in order to prevent cosmic rays from influencing observation. Super-Kamiokande is used for such research purposes as observing neutrinos coming from space and proton decay events. Neutrinos from space can be roughly classified into three types: neutrinos from the sun (solar neutrinos), neutrinos generated by the reaction of cosmic rays with the earth's atmosphere (atmospheric neutrinos), and neutrinos generated by a supernova explosion at the end of a star's life (supernova neutrinos). Neutrinos entering the tank of Super-Kamiokande may react with pure water in the tank and generate a glimmering bluish-white light (known as "Cherenkov light"). The energy, reaction position, and travel direction of the incoming neutrinos are calculated by sensing Cherenkov light with the photomultiplier tubes. This operation is called event reconstruction.

One particularly important point of neutrino observation is to remove background events. For example, the background events that may occur during the observation of solar neutrinos involve environmental gamma rays coming from outside the tank and minute residual amounts of such radioactive materials as Radon in the water inside the tank. Such background events generate Cherenkov light in the same way as a neutrino reaction and thus are a source of confusion. However, the Super-Kamiokande system can distinguish neutrino events from background events by analyzing the observed particle generation position and travel direction. The data observed on background events clearly identified by using the reaction position and other factors is discarded immediately after being acquired.

After the data of background events is discarded, the remaining data is converted into a world standard format (known as "ZEBRA"[3]) established by the European Organization for Nuclear Research (CERN).[4] The converted data is then sent to the Super-Kamioka Computer System for Analysis. This format conversion operation is called reformat operation. One record consists of about 5 KB and about 11 million events are recorded each day.

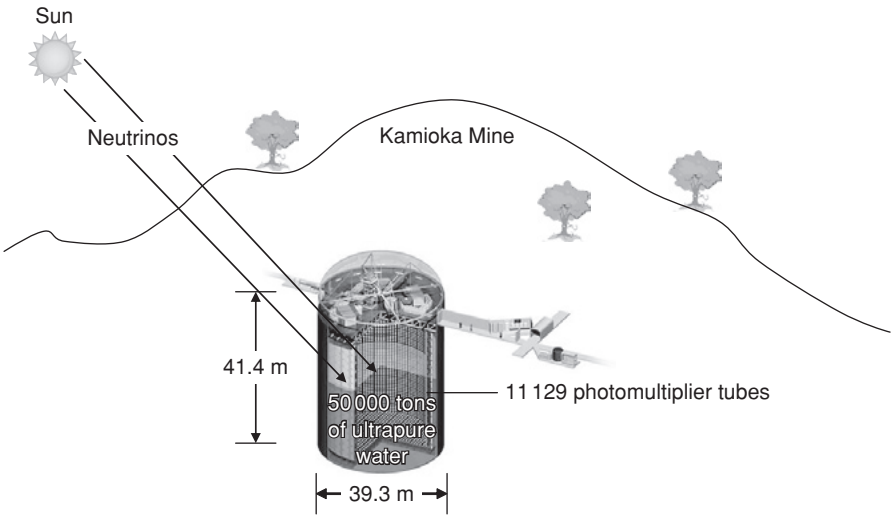

Figure 1
Super-Kamiokande Neutrino Detection Equipment.

Therefore, about 50 GB are saved per day. The Super-Kamioka Computer System compensates the data of all remaining events by using the parameters for respective characteristics of the photomultiplier tubes and the parameters for water quality (e.g., transparency of water), and reconstructs these events in real time. Since these parameters may be subject to seasonal variations and the applications used to reconstruct events may be constantly improved and advanced, the events are often reanalyzed by using all the accumulated raw data. Given the huge amount (about 110 TB) of raw data needed for reanalysis, a system capable of superfast data access must be used.

Figure 2 shows the configuration of the Super-Kamioka Computer System for Analysis. The next section describes the configuration of system located at the experimental site in the mine, the system outside the mine (located in the computation and research buildings), and the technology necessary to acquire high-speed access.

3. System configuration of Super-Kamioka Computer System for Analysis

The Super-Kamioka Computer System for Analysis consists of the system inside the mine (for collecting data observed by Super-Kamiokande and converting the data format), the system outside the mine (for accumulating the format-converted data and analyzing the accumulated data), the terminals used daily and data backup system, the monitoring system, and the Gigabit Ethernet for connecting these systems.

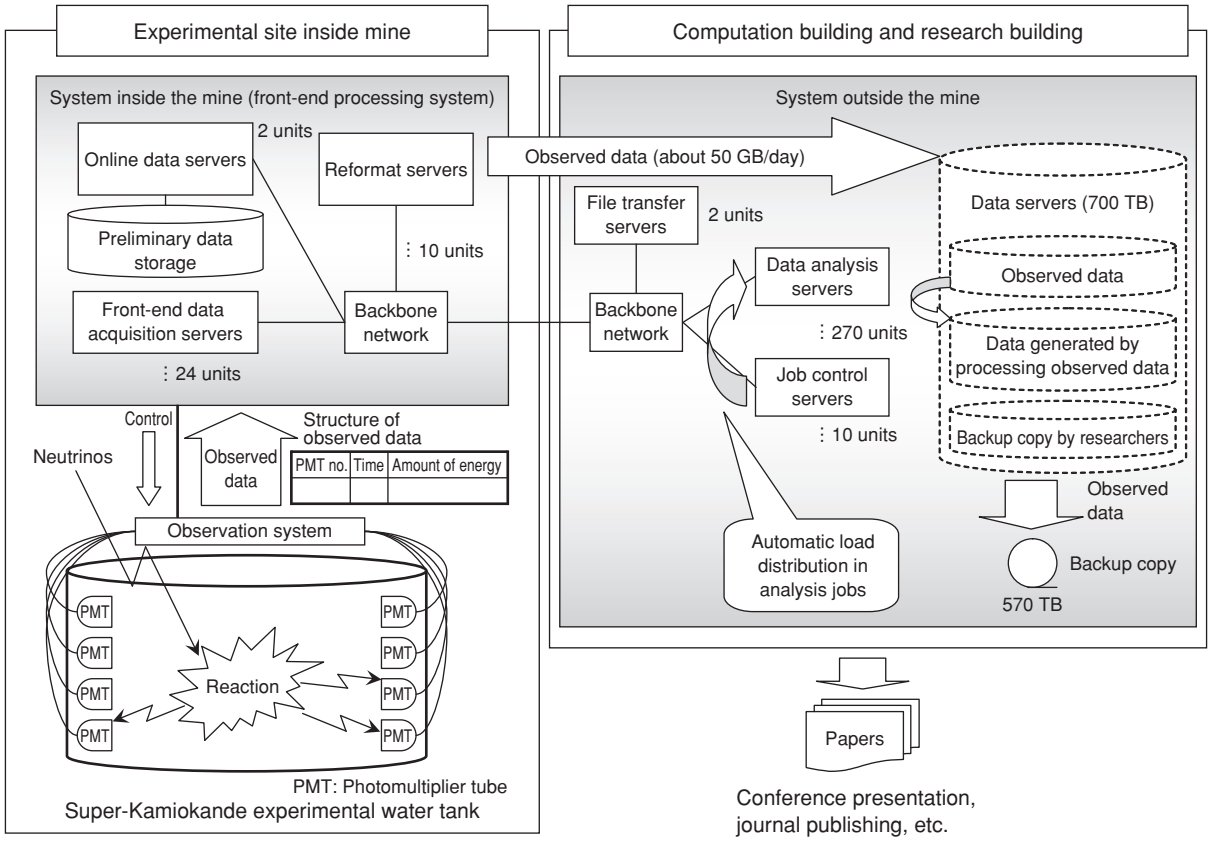

Figure 2
Super-Kamioka Computer System for Analysis.

The following describes the main constitutive systems inside and outside the mine.

3.1 System inside the mine

Super-Kamiokande observes neutrinos on a 24-hour basis in order to steadily detect important events involving cosmic rays. The front-end processing system for this detection is a real-time system that requires high availability.

The front-end processing system consists of front-end data acquisition servers (24 PRIMERGY RX200 S3 units), online data servers (two PRIMERGY RX300 S3 units and one ETERNUS4000 model 100), reformat servers (ten PRIMERGY RX200 S3 units), and an online network (four Catalyst4948 units and two Catalyst2960G units) for connecting these components.

The observation system is connected to the front-end data acquisition servers via a dedicated interface developed by ICRR of the University of Tokyo. Dedicated applications collect the observed data. The front-end data acquisition servers send the data collected to the preliminary data storage connected to the online data servers. Thus, the sent data is accumulated in this preliminary data storage.

The online data servers communicate with data servers (used as the final storage of observed data) in the system outside the mine. Should this data-server communication be interrupted, the observed data sent to the online data servers in real time must not be lost. To reduce the risk of losing such real-time observed data, the storage capacity of online data servers must be as large as possible. The ETERNUS4000 model 100 has limited storage capacity of up to 2 TB in one RAID group. However, we constructed a 5-TB file system by connecting multiple RAID groups to one file system by using the Logical Volume Manager (LVM) of the Linux system. This enables the online data servers to accumulate observed data for up to about 100 days.

The front-end data acquisition servers execute the event reconstruction described above. Then, the reformat servers reformat the data and transfer necessary data to the data servers at a rate of about 50 GB per day.

3.2 System outside the mine

The system outside the mine is used to accumulate and analyze the observed data sent from the front-end processing system. This system can simultaneously process up to 1080 jobs for high-speed parametric data analysis (using four CPUs/node × 270 nodes). Calibration for newly accumulated observation data and reanalysis jobs are always executed, and typically about 500 jobs and sometimes more than 1080 jobs (including jobs awaiting execution) are input. The disk drives and file systems must therefore be carefully designed to ensure efficient data access in these jobs. The next section describes the concrete points for such design.

1) System configuration and main functions

The system outside the mine consists of data servers (three PRIMEQUEST 520 units and six ETERNUS4000 model 500 units), job control servers (ten PRIMERGY BX620 S3 units), file transfer servers (two PRIMERGY BX620 S3 units), and data analysis servers (270 PRIMERGY BX620 S3 units), and a backbone network (Catalyst6509E) for connecting these components. The servers are connected via the backbone network and multiple Gigabit Ethernet lines for high-speed network access.

The data servers have a total data storage capacity of 700 TB. SRFS provides the file sharing environment for the job control servers and data analysis servers.

2) Program development environment and job control

Users can use the job control servers to develop programs and enter jobs into the data analysis servers. The job control servers have a development environment such as Intel compilers and the VTune performance analyzer as part of CPU performance analysis applications. The

Parallelnavi Network Queuing System (NQS) environment—batched-job operation support software—is also provided to immediately execute the developed and debugged programs. Thus, all operations required for analysis operations in development, execution, and evaluation can be executed on one terminal.

The jobs input from a job control server are assigned by NQS to idle CPUs selected from among the data analysis servers, and then executed by the CPUs. Therefore, users need not search for idle resources. In each cabinet of the blade servers that constitute the data analysis servers, two 1000BASE-T interface lines are shared by ten blades. If too many jobs are concentrated in one blade-server cabinet, network overhead may increase excessively. Therefore, the NQS environment was set so that jobs are as evenly distributed to the cabinets as possible.

4. High-speed data access

Only accumulating the observed data would require a function for storing about 50 GB in a 24-hour period. Since the data analysis servers must process all accumulated data, however, data must be input from and output to the servers as quickly as possible so that analysis processing can always continue.

We took the following measures for this purpose:

1) File system

The data analysis servers and file transfer servers constitute a file system mounted over a network so that data files can be utilized evenly from any data analysis server. The Network File System (NFS) is generally used as such a file system mounted over a network. However, our experience shows that NFS does not offer high processing speed and encounters difficulty in simultaneously processing the input/output requests concurrently issued from all the data analysis servers. To solve these problems, we employed SRFS instead of NFS.

We then created input/output model jobs for the analysis programs, and used SRFS to measure job performance. Based on the measurement results, we decided on an input/output data length of 8 MB and designed the conditions necessary to achieve maximum performance when using this input/output data length.

2) Storage

The file transfer servers are comprised of sets of seven hard disk drives (HDDs) for enabling concurrent use of the drives in parallel. A set of seven HDDs is called a physical volume. When this configuration is used without modification, the maximum physical volume performance restricts maximum input/output performance. To ease this restriction, we constructed a logical volume by combining multiple physical volumes so that the physical volumes could be used when using a logical volume.

The important points for increasing maximum logical volume performance are determining how many physical volumes to use for constituting a logical volume (number of stripe columns), and how many volumes of data to input to and output from one physical volume at a time per input/output operation on the logical volume (stripe width). We determined these values according to the performance measurement results. Specifically, we decided that the stripe width should be 128 or 256 KB using 16 or 8 stripe columns in order to achieve optimum performance. When using 16 stripe columns, however, we found that the system cannot be comprised of a number of HDDs exceeding the limit. We therefore employed 8 stripe columns and a stripe width of 256 KB.

3) Network

SRFS is a network file system in which real data is accessed via the Gigabit Ethernet. However, traffic other than input/output operations can be expected to reduce input/output performance, and the broadcast packets issued by SRFS itself may disturb other communications. As a result, we constructed an input/output-dedicated network.

4) Input/output library

We developed a dedicated input/output library in which the input/output data length was 8 MB. This provides high-speed data input/output from user-developed analysis programs.

5. Throughput performance

Since up to four analysis programs can be executed on one data analysis server of this system, up to 1080 input/output requests may be simultaneously issued by all data analysis servers. For smooth analysis, high-speed throughput performance that can process this many input/output requests without delay must be supported.

The following describes the measures taken for acquiring high throughput and cites the throughput measurement results.

1) Suppression of network slowdown

Ideally, the network band should be fully used in order to efficiently use a network. The network band can be fully used by employing one of two methods: (1) dividing a request into multiple operations to be executed in parallel on one network, or (2) simultaneously executing multiple requests on one network. Since there are too many physical interfaces on the data analysis system side in this system as compared with the number of physical interfaces on the data server side; however, the band at the data server is too narrow and the network may slow down. To prevent such network slowdown, we restricted the number of data analysis servers assigned to the physical interface of each data server.

More specifically, each data server has seven physical interfaces connected to the input/output network, but up to 270 data analysis servers must be processed. Therefore, we assigned each physical interface to 38 or 39 data analysis servers.

When this setting is applied, an error in any physical interface of a data server may render the 38 or 39 connected data analysis servers unusable. However, we employed this design by considering high throughput performance.

2) Suppression of communication time-out

The time-out value and retry count in network communications are important design points. An excessively large time-out value delays error detection and recovery. Conversely, an inadequately small time-out value increases the retry communication count and reduces communication performance.

Since throughput performance was considered more important than error detection performance, we set a time-out value for continued processing even under a high load. Given the difficulty in acquiring the appropriate time-out value by calculation, however, tuning operation to obtain a time-out value was eventually performed according to the actual measurement results.

In the actual measurement, the 1080 model jobs described in the previous section were executed concurrently to simultaneously issue input/output requests to one file system. Measurement was repeated by incrementing and decrementing the time-out value. Moreover, the occurrence of a time-out invokes retry processing and results in varying job execution time. We therefore determined the optimal value by considering that the job execution time was roughly the same as when no time-out occurred.

3) Measurement results of throughput performance

In measuring throughput performance, the 1080 jobs used for the time-out value tuning described above were concurrently executed on 270 data analysis servers, in order to measure the input/output speed in input/output operations on three data servers. In this case, it takes a certain amount of time from starting execution of the first job to starting the concurrent parallel operation of 1080 jobs, and the number of jobs operating in parallel gradually decreases because shorter jobs terminate earlier. We therefore executed one job several times continuously, then discarded the first and last execution results so as to exclude the measurement results for less

than 1080 concurrent jobs.

We executed jobs on the data analysis servers to read/write data on the data servers and acquired throughput performance of 960 MB/s as the average read/write value.

6. Conclusion

This paper outlined the observation of data at the Kamioka Observatory of the Institute of Cosmic Ray Research, University of Tokyo, introduced the computer system used for on-site data analysis, and then described the means and background of acquiring high-speed data access. Because the actual data observation system has different site data characteristics and a different system configuration, our solution cannot be considered optimal in all cases. However, we hope that our computer system design concepts and problem-solving approaches prove helpful to you.

Acknowledgement

The authors wish to thank Research Associate Yusuke Koshio of the Kamioka Observatory, Institute for Cosmic Ray Research, The University of Tokyo, for his kind guidance and cooperation.

References

1) The Institute for Cosmic Ray Research (ICRR) of the University of Tokyo, Kamioka Observatory. *http://www-sk.icrr.u-tokyo.ac.jp/index_e.html*
2) Super Kamiokande. *http://www-sk.icrr.u-tokyo.ac.jp/sk/index-e.html*
3) CERN: The ZEBRA System. *http://wwwasdoc.web.cern.ch/wwwasdoc/zebra_html3/zebramain.html*
4) CERN. *http://public.web.cern.ch/Public/Welcome.html*

Akira Mantani
Fujitsu Ltd.
Mr. Mantani joined Fujitsu Ltd., Japan in 1985 and has been engaged in the installation support of computer systems for national science institutions. He worked in London from 1996 to 2003 and was engaged in the installation and support of major European supercomputer systems.

Yasushi Yamaguchi
Fujitsu Ltd.
Mr. Yamaguchi joined Fujitsu Ltd., Japan in 1980 and was engaged in the installation support of supercomputer systems up to 2001. Then he was engaged in the development of applications and the construction of social computer systems until 2006. Since 2007, he has mainly supported critical computer systems in various projects.

Yoshiaki Matsuzaki
Fujitsu Ltd.
Mr. Matsuzaki joined Fujitsu Ltd., Japan in 1988 and has been engaged in the development of applications and the support of computer systems for research and development institutes in the Tsukuba area. Since 2007, he has also been engaged in developing the scientific computer business.

Kouki Kambayashi
Fujitsu Ltd.
Mr. Kambayashi joined Fujitsu Ltd., Japan in 1989 and has been engaged in the installation support of computers and network systems for research and development institutes. He has been supporting the Super-Kamioka Computer System ever since its installation in 2007.