Super-Kamioka Computer System for Analysis

● Hideyuki Okada ● Yukinobu Oyama ● Tomohiro Tsukahara

The Institute for Cosmic Ray Research, the University of Tokyo, is observing supernova, solar, and atmospheric neutrinos, while researching the Grand Unified Theory, using the Super-Kamiokande detector. This large-scale observation detector was constructed 1,000 m below ground level and has a water tank (41.4 m in height, 39.3 m in diameter) capable of storing 50,000 tons of ultrapure water. There are 11,129 photomultiplier tubes (each 50 cm in diameter) mounted on the tank wall. The Super-Kamiokande detector can generate approximately 45 TB of observed data per day, of which approximately 500 GB is saved after noise is removed. The total amount of data stored on the disks is nearly 3 PB, including user information. The Super-Kamiokande detector is used to observe natural phenomena, and the challenge for Fujitsu is to ensure that accurate observed data are captured 24/7 and to store them over a long period of time. In February 2012, Fujitsu installed the Super-Kamioka Computer System for analysis designed to facilitate the storage of a huge volume of data on a large-capacity disk so that the data can be quickly accessed and to provide safe and long-term data storage. This paper explains the system's various innovations that enable accurate capturing of observed data, high-speed processing, and long-term archiving.

1. Introduction

The Kamioka Observatory[1] of the Institute for Cosmic Ray Research, the University of Tokyo, is researching the universe and elementary particles by observing neutrinos and searching for proton decay. It includes the Super-Kamiokande,[2] which began neutrino observations in 1996. Professor Masatoshi Koshiba received the 2002 Nobel Prize in Physics for observing neutrinos from a supernova explosion in February 1987 using the Kamiokande. The Super-Kamiokande is the successor to the Kamiokande. Furthermore, in 1998, it was announced that Professor Takaaki Kajita and his colleagues had confirmed the occurrence of neutrino oscillation, and for that he received the 2015 Nobel Prize in Physics. The neutrino detection experiment at the Kamioka Observatory has thus come to attract worldwide attention.

In 1993, Fujitsu supplied a system for processing the observed data obtained by this experiment. Since then, Fujitsu has replaced this system as needed in line with advances in computer technologies and thereby support the data analysis. Fujitsu supplied, in particular, the Super-Kamioka Computer System for Analysis in February 2012. This system was needed because the data obtained in a single day of neutrino observation comes to approximately 45 TB, and approximately 500 GB of these data are stored after noise is removed and reactions associated with radioactive material are excluded. To enable these data to be saved and to be quickly retrieved for data analysis, Fujitsu upgraded the storage system by adding high-accessibility magnetic storage equipment (FUJITSU Storage ETERNUS DX80 S2). Furthermore, in addition to the analysis of newly observed data, past data are sometimes reanalyzed using new applications or applications designed to vary the parameters. Consequently, to enable all types of analyses to be performed in a relatively short time, the need arose for even faster data access, so it was decided to use Fujitsu Exabyte File System (FEFS), Fujitsu's scalable file system software for achieving high-speed file sharing.

This paper introduces the Super-Kamioka Computer System for Analysis.

2. Overview of observations and data analysis

The Super-Kamiokande detector consists of a cylindrical water tank, 39.3 m in diameter and 41.4 m in height, filled with 50,000 tons of ultrapure water plus 11,129 optical sensors called photomultiplier tubes (PMTs) installed on the tank wall. The detector is installed 1,000 m underground in the Kamioka Mine in Hida-city, Gifu, Japan, to hinder the penetration of cosmic rays, which can affect neutrino observations. The main objectives of the research performed are to observe neutrinos flying in from space and to observe the phenomenon of proton decay. On flying into the Super-Kamiokande detector, neutrinos may react with the water in the tank and generate weak bluish white light (Cherenkov light). The detection of this light by the PMTs enables the energy, reaction location, direction of movement, etc. of a penetrating neutrino to be calculated in a process called "event reconstruction."

Removing background events is particularly important in neutrino observation. Environmental gamma rays entering the tank from the outside and slight amounts of radioactive material such as radon remaining in the tank water can generate Cherenkov light in the same way as neutrino reactions, creating events that can be extremely misleading. At the Super-Kamiokande detector, background events are distinguished from targeted neutrino events by using the characteristics of observed particles such as reaction position and movement direction. The data obtained from neutrino observations are converted into the ROOT format[3] developed by the European Organization for Nuclear Research (CERN)[4] and passed to the Super-Kamioka Computer System for Analysis. This is called the reformat process.

Starting with 45 TB of data obtained in one day of observations, the data to be saved comes to approximately 500 GB per day after removing noise, extracting the data generated by the Cherenkov light, and removing background events. The Super-Kamiokande computer system performs real-time event reconstruction after taking into account parameters related to the individual characteristics of the PMTs and to water quality (such as degree of water transparency). These parameters, however, may change seasonally, and given that applications for performing event reconstruction are constantly evolving, it is not uncommon to reanalyze previously stored raw data. The amount of data targeted for reanalysis can be as much as 500 TB, so very high-speed data access is needed.

In the following sections, we first describe the configuration of the system inside the mine and the system outside the mine in the computer and research buildings. We then describe the technologies used for achieving accurate data collection, high-speed data processing, and long-term data archiving.

3. System configuration

The Super-Kamioka Computer System for Analysis consists of the system inside the mine for collecting the data observed at the Super-Kamiokande detector and performing a format conversion, the system outside the mine for storing and analyzing the format-converted data, routinely used terminals and a backup system, a monitoring system, and a network system for interconnecting the above systems.

Neutrino observation is performed on a 24/7 basis, and data are stored year in and year out, so the media used for storing this observed data must have a huge capacity. The Super-Kamioka Computer System for Analysis has been in operation since 1993 and is now in its 5th generation. System operation period and storage configuration/capacity by generation are listed in **Table 1**.

The 1st to 3rd generations made use of magnetic tape for both data storage and backup purposes. Starting with the 4th generation, it was decided to use a different file system in accordance with purpose and to divide the storage configuration into a large-capacity magnetic disk area requiring high-speed access and a backup area for long-term archiving.

The 5th generation system supporting this experiment consists of the system inside the mine and the system outside the mine in the computer and research buildings. The configuration and features of each are described below (**Figure 1**).

3.1 System inside mine

Super-Kamiokande observes neutrinos on a 24/7 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. This system consists of various types of equipment: front-end data-acquisition servers (FUJITSU

PRIMERGY RX200 S6: 24 units), online data servers (FUJITSU PRIMERGY RX300 S6: 2 units; ETERNUS DX80 S2: 1 unit), reformat servers (PRIMERGY RX200 S6: 10 units), and core network equipment for interconnecting the above (Cisco Systems Catalyst 4506-E switch: 1 unit).

The observation system and front-end data acquisition servers connect to each other via approximately 520 specialized data consolidation boards developed by the Institute for Cosmic Ray Research, the University of Tokyo. The computers use specialized applications to perform the data collection. The front-end data-acquisition servers forward the collected data to the reformat servers, which rearrange the data in

Table 1
System operation period and storage specifications by generation.

Generation	Operation period	Data storage medium	Backup medium
1st	1993–1997	Magnetic tape (CMT) 12.8 TB	
2nd	1997–2002	Magnetic tape (DTF) 200 TB	
3rd	2002–2007	Magnetic tape (LTO1) 440 TB	
4th	2007–2012	Magnetic disk 700 TB	Magnetic tape (LTO3) 547 TB
5th	2012–	Magnetic disk 3 PB	Magnetic tape (LTO5) 1.5 PB

CMT: Cartridge magnetic tape
DTF: Digital tape format
LTO: Linear tape-open

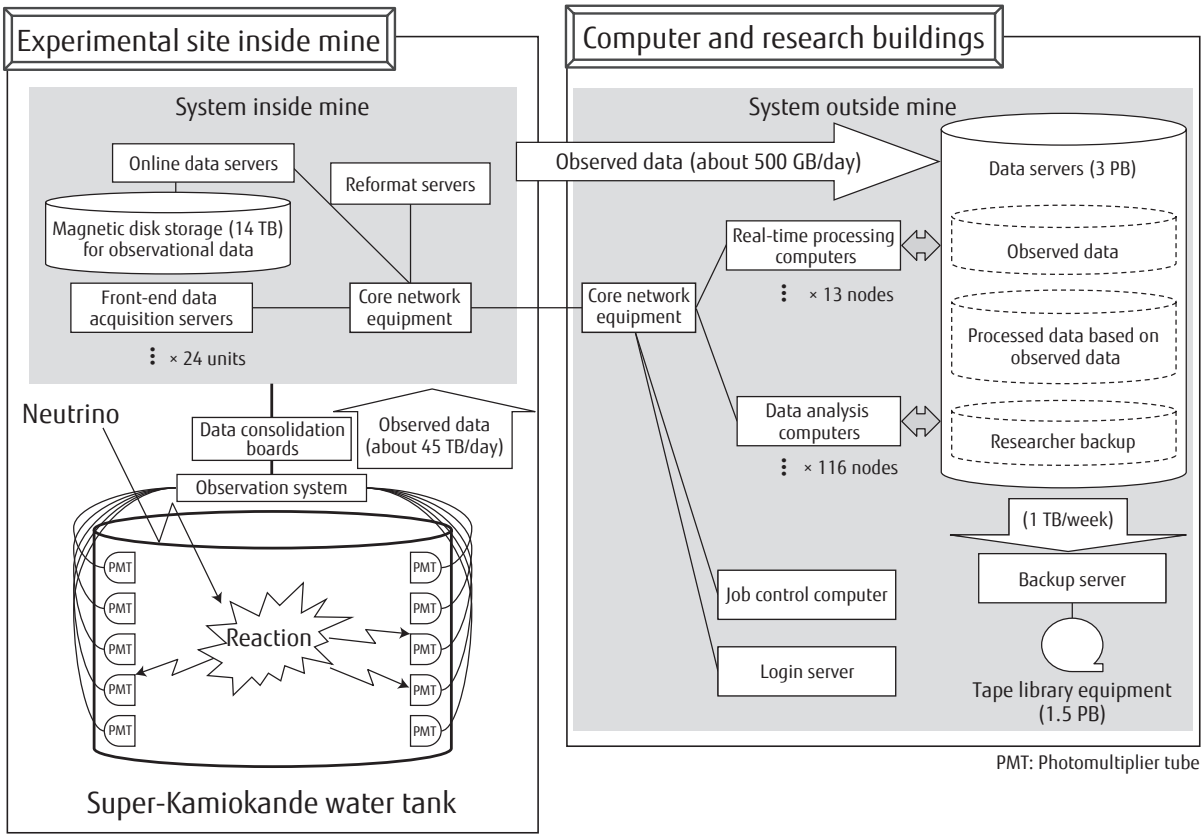

Figure 1
Super-Kamioka Computer System for Analysis.

chronological order, remove noise, and perform event sorting. The resulting data is then forwarded to the online data servers and stored on magnetic disk.

3.2 System outside mine

The system outside the mine is used to accumulate and analyze the observed data sent from the front-end processing system. To support high-speed analysis of this data, the system was designed to be capable of simultaneously executing up to 1,392 jobs (12 CPUs/node × 116 nodes). Newly stored data are constantly being corrected or reanalyzed, and the number of jobs being submitted for execution, while normally about 500, can exceed 3,000, including jobs waiting for execution at busy times. This means that high-speed access to the file system is needed to enable efficient access to the data output from this large number of jobs.

1) Configuration and features of system outside mine

The system outside the mine likewise includes various types of equipment: data servers (PRIMERGY RX300 S6: 8 units; ETERNUS DX80 S2: 16 units), real-time processing computers (PRIMERGY BX922 S2: 13 units), data analysis computers (PRIMERGY BX922 S2: 116 units), and core network equipment for interconnecting the above (Cisco Systems Catalyst 4507R+E switch: 1 unit).

The data servers feature a data storage area with a total capacity of 3 PB. They provide a file-sharing environment with respect to the job control computer and data analysis computer via FEFS. The real-time processing computer and data analysis computer are connected to the data server via InfiniBand (Mellanox Grid Director 4036) to achieve high-speed file access. This scheme ensures stable file access from a large number of jobs and provides a user-friendly data access environment.

2) Program distribution environment and job control

Users develop programs and enter jobs in analysis equipment via the login server. This server provides users with multiple versions of a compiler and access to the FUJITSU Software Technical Computing Suite (TCS). The latter includes software for batch-job operation that enables high-speed scheduling and execution of multiple jobs. In this way, users can perform all the tasks necessary for analysis such as development, execution, and evaluation from a single terminal.

This batch-job system is divided into a real-time system and user-job system. The real-time system automatically executes reformatting, noise removal, and event sorting against the data received from the system inside the mine in real time and saves processing results on the data server. The user-job system runs on the data analysis computer and enables researchers to perform various types of analysis in accordance with the data processed by the real-time system. Dividing the batch-job system in this way minimizes negative effects on operation by limiting the range of impact from a failure, by enabling maintenance to be performed on each system separately, etc.

The TCS can also be used to manage computing resources. With the TCS, a job submitted by the job control computer is automatically executed on an available CPU within the data analysis computer. There is therefore no need for researchers to search for free resources themselves. In addition, memory not being used for job execution can be used for file caching to accelerate file access operations. For this reason, nodes are dispersed as much as possible, and settings are made to optimize job placement so as to minimize memory usage at each node. The above schemes help to speed up the analysis of observed data.

4. Accurate data collection

Given that the target of observation in a neutrino detection experiment is natural phenomena, there is a need for a system that can collect observed data in a stable, round-the-clock manner.

The path from the Super-Kamiokande detector to the tape library equipment in the system outside the mine used for archiving observed data consists of a variety of servers and network devices in a redundant configuration. This results in a system configuration resistant to a simple failure. In particular, the acquisition of observed data will continue as usual even if data transfer to the data server in the system outside the mine (used as the final storage of observed data) cannot be performed due, for example, to maintenance work on the system outside the mine or a temporary network cutoff. As shown in **Figure 2**, the system inside the mine is equipped with magnetic disk storage (14 TB) for saving the observed data received in real time from the online data servers. This scheme enables

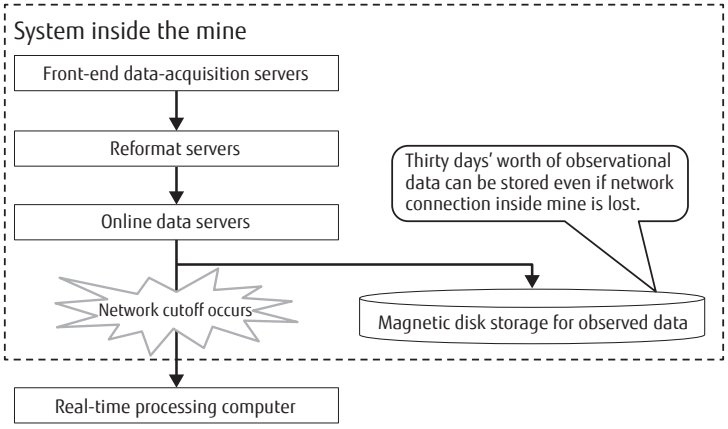

Figure 2
Storage of observational data if network cutoff occurs.

observed data to be stored by the online data servers for up to about 30 days.

5. High-speed and stable data processing

It would be sufficient to provide for a capacity of 500 GB of data per day if the only concern was storing observed data. However, there are times when researchers wish to reprocess previously stored data using the data analysis computer such as when updates are made to an analysis program. There is therefore a need to provide high-speed input/output performance to enable smooth data access by analysis programs. The system must also provide stable operation on a 24/7 basis to enable data-analysis processing to be completed in a relatively short time.

5.1 File system

The data analysis computer and data server transfer data over a high-speed InfiniBand connection providing uniform high-speed file access from any data analysis computer unit. Given the maximum write performance of 2.5 GB/s, the path between the data analysis computer and data server was designed so that operation is not affected even by simultaneous file access by multiple jobs.

The 3 PB capacity of the file system is divided into two completely independent areas, each having a capacity of 1.5 PB. The first area is open to users as a file system to be used in normal operations. The second area is used as a test environment for improving operations and validating revisions and upgrades. This configuration ensures stable system operation.

The servers making up FEFS and the associated communication paths are arranged in a redundant configuration so that a server failure or other problem does not bring down the file system. In addition, the directory quota function in FEFS can be used to limit the volume of directory files allocated to each research group. This function can limit the amount of file usage under specific directories within the same file system. In addition, the limit settings applied to file usage for each research group can be increased or decreased without halting operations, providing for flexible management of file usage.

5.2 Storage

The magnetic disk equipment of the data server is configured so that 14 magnetic disk drives can be used simultaneously in parallel. Each set of 14 magnetic disk drives is called a physical volume. Using 48 physical volumes per file system and distributing load accordingly enables large-capacity and high-speed file access.

5.3 TCS software for batch job support

The command line of the current network queuing system (NQS) is different from that of the previous system. In consideration of the usefulness on an existing user-created script, we prepared a function for executing NQS commands in a TCS environment and realized a smooth transition between the two systems.

The TCS system prepares a job queue for each group, but some data analysis computer units (servers) are included in multiple job queues, thereby minimizing the bias in the number of execution jobs among different groups. Having multiple job queues share a server prevents the utilization rate of that server from dropping even though a particular job queue may become empty.

6. Long-term data archiving

Observed data cannot be regained after the fact and must therefore be saved in a reliable manner. To this end, the system outside the mine saves files in a data server capable of high-speed access while automatically backing up those files to a tape library. The tape library equipment (ETERNUS LT270: 2 units) can mount LTO5 tapes with a total capacity of 1.5 PB uncompressed for backing up observed data. The tape-drive write performance is 1,120 MB/s uncompressed.

The Veritas NetBackup Standard is used for the backup software. Client software is installed in each real-time processing computer unit to back up data onto the tape library equipment. The tape library equipment and Veritas NetBackup Standard are also used for long-term storage of user-created data, which are stored under a specific directory on the backup server and periodically backed up onto the tape library equipment.

7. Conclusion

This paper introduced the Super-Kamioka Computer System for Analysis at the Kamioka Observatory of the Institute for Cosmic Ray Research, the University of Tokyo. The Institute plans to continue and expand its groundbreaking research of neutrinos. Fujitsu intends to make further contributions to this experiment by developing the increasingly advanced systems that will be needed.

The authors would like to extend their deep appreciation to Associate Professor Yoshinari Hayato of the Kamioka Observatory, Institute for Cosmic Ray Research, the University of Tokyo, for his helpful guidance and cooperation over the course of writing this paper.

References

1) Kamioka Observatory, Institute for Cosmic Ray Research, the University of Tokyo.
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) ROOT.
https://root.cern.ch/
4) CERN.
http://home.cern/

Hideyuki Okada
Fujitsu Ltd.
Mr. Okada is currently engaged in negotiation, construction, and operation support for computing systems in scientific fields including the Super-Kamiokande computing system.

Yukinobu Oyama
Fujitsu Research Institute
Mr. Oyama is currently engaged in public-sector consulting work.

Tomohiro Tsukahara
Fujitsu Ltd.
Mr. Tsukahara is currently engaged in the planning and business promotion of computing systems in scientific fields.