ACA Correlator System: Supercomputer System Developed for ALMA Project

● Katsumi Abe ● Junpei Tsutsumi ● Takahiro Hiyama

The National Astronomical Observatory of Japan (NAOJ) in the National Institutes of Natural Sciences (NINS) is jointly promoting with U.S. and European observatories the Atacama Large Millimeter/submillimeter Array (ALMA) Project. The Fujitsu Group developed a supercomputer, the Atacama Compact Array (ACA) correlator system, dedicated to processing data from large radio telescopes. This system consists of a correlator control system comprising 35 FUJITSU PRIMERGY PC servers and the ACA correlator developed by Fujitsu Advanced Engineering. The system meets stringent requirements including the need to process in real time huge amounts of radio signal data (512 billion pieces/s) collected by telescope antennas at a calculation speed of 120 tera-operations per second and to operate stably in an extreme environment (a plateau near the Atacama Desert in Chile at an altitude of 5000 m with an atmospheric pressure of 0.5 atm). It uses cost-efficient field-programmable gate array (FPGA) technology and a disk-less system to achieve reliable operation. Lengthy testing in a low pressure environment demonstrated that the system can operate in such an environment for a long period of time and that it can be maintained remotely. Test operation began in 2011 and, after this pilot operation proved that the system is suitable for use in the ALMA Project, it was put into regular use in 2013. This paper describes the entire concept of the ACA correlator system and the approaches taken to realizing its high computing performance and reliable operation.

1. Introduction

The National Astronomical Observatory of Japan (NAOJ) of the National Institutes of Natural Sciences (NINS) has joined up with U.S. and European observatories to construct a large radio telescope on a plateau in Chile's Atacama Desert at an altitude of 5000 m. This project, which is known as the Atacama Large Millimeter/submillimeter Array (ALMA) Project,[1] was initiated in 2002 under U.S. and European supervision and became a trilateral system in 2004 with the addition of Japan. Test operation began in 2011, and regular operation began in 2013 after a pilot observation period. Astronomers from around the world are eager to make observations using ALMA, and amazing results have already been reported in academic journals and elsewhere.[2]

The ALMA telescope consists of 66 antennas and related facilities for receiving and processing very weak signals from celestial objects. A view of ALMA's antennas and buildings is shown in **Figure 1**. NAOJ is in charge of 16 antennas, the receivers used for receiving signals in four of seven frequency bands, the Atacama Compact Array (ACA) correlator for ultra-high-speed processing of received data, and the correlator control system. The correlator and its control system combined are called the ACA correlator system; its processing speed of 120 tera-operations per second (TOPS) is equivalent to that of a supercomputer.

Fujitsu began developing the ACA correlator system in 2004 in collaboration with Fujitsu Advanced Engineering (FAE).[3] The correlator had to have exceptionally high throughput for processing in real time the huge amounts of data received from the antennas. A view of the ACA correlator system installed at an altitude of 5000 m and the system configuration are shown in **Figure 2**. This system had to be capable of 24-hour continuous operation and of prompt restoration in the event of a system fault during operations.

Specifically, three requirements had to be met when setting out to develop the ACA correlator system:

1) High-performance processing of huge amounts of data in real time

The data transfer rate from antennas to correlator must be 1.5 Tb/s. These data must be received without any delays and processed in real time.

Figure 1
View of ALMA telescope.

2) Reliable operation in an extreme environment

The system must be able to run stably under severe environmental conditions, including an altitude of 5000 m and an atmospheric pressure of 0.5 atm.

3) Remote maintenance

Given a system situated in a high-altitude environment on the other side of the world in a largely uninhabitable area, it must be possible to perform remote maintenance and remote checking of facility status.

In this paper, we describe the plans made and measures taken with respect to these requirements in the development of the ACA correlator system.

2. Real-time processing of huge amounts of data

The ACA correlator system includes interfaces to an upstream ALMA total control system and to the downstream data archive equipment, both under U.S. and European management. On the basis of instructions received from the total control system, the correlator control system determines which observation mode

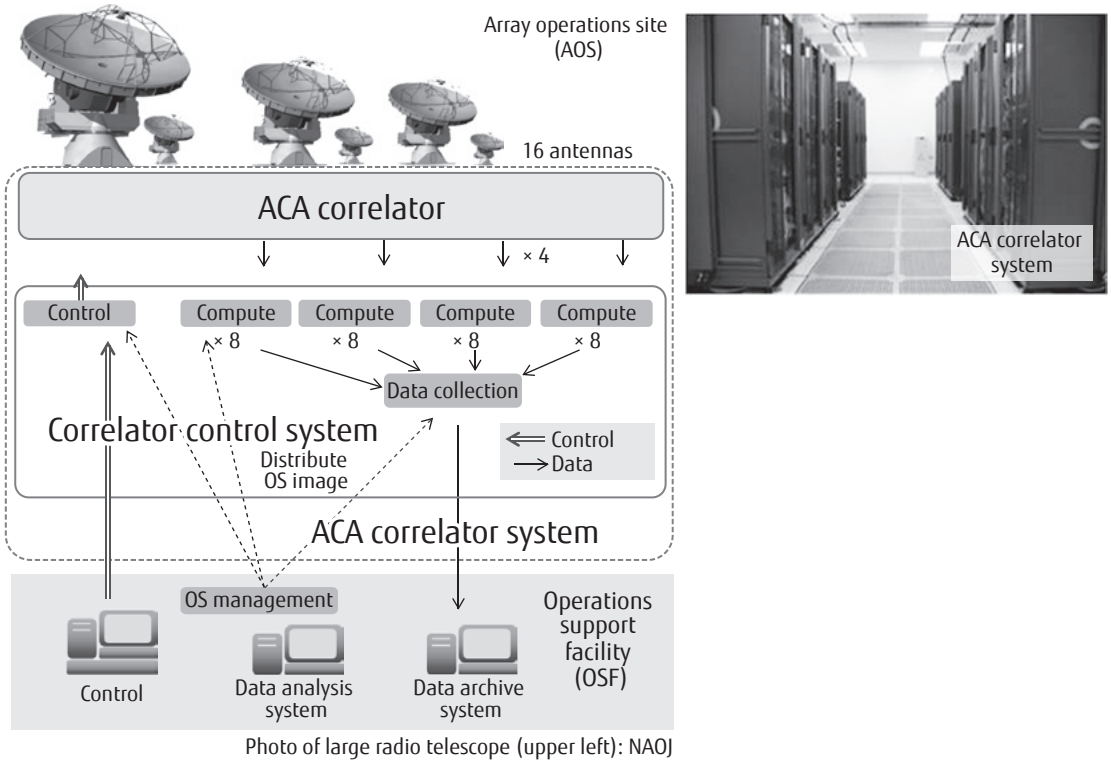

Figure 2
View of ACA correlator system (upper right) and system configuration.

to use. Then, in accordance with the mode selected, the correlator performs correlation processing of radio signal data received by the antennas and forwards the processing results (correlation data) to the correlator control system for integration and correction processing in both the time and frequency domains. On completing this processing, the correlator control system sends the resulting data to the data archive system.

The bit rate for transferring data from the antennas to the correlator is 1.5 Tb/s, which is equivalent to the capacity of 20 000 home Internet lines. This huge amount of data that is continuously being output from the antennas must be received in a continuous, uninterrupted manner, which requires a level of performance that can execute 120 trillion calculations per second. To this end, the PC servers[note)] that make up the correlator control system must have enough throughput to receive the data output from the correlator and to forward the processing results without delay to the data archive equipment.

Meeting these requirements would require a large-scale computer, but at the time of system development, the computers that were available could not be used without difficulty given the power required, installation-site requirements, and cost required. Moreover, developing a specialized processor for this purpose would be extremely costly. To overcome these problems, we used field-programmable gate array (FPGA) technology, which is known for its superior cost performance, and developed a parallel computing system consisting of 4096 FPGA chips. This approach enabled us to achieve real-time processing of huge amounts of data. Furthermore, in contrast to an ordinary processor with circuits that cannot be changed, the FPGA chips enable flexible alteration of circuits to support desired functions. This means that circuit tuning to extract maximum performance can be performed immediately before beginning actual operations.

Since mounting standard communication cards in the correlator control system (PRIMERGY servers) to handle the data output from the correlator would not provide enough performance to receive that huge amount of data without error, we developed a specialized receiver card for use in the servers. We also performed extensive tuning of the calculation program (speeding up access to memory, optimizing the method for describing source code while keeping processing order at the assembler level in mind, optimizing the combination of compiler options, etc.) so that the huge amount of received correlation data could be analyzed at high speed in a manner appropriate to the observation mode.

note) The PC server name is "FUJITSU Server PRIMERGY."

3. Base system for extreme environment

Severe environmental conditions that include an altitude of 5000 m and an atmospheric pressure of 0.5 atm can cause the elements comprising the ACA correlator system such as capacitors and power-supply units to fail and equipment to overheat due to reduced cooling efficiency. Hard disk drives (HDDs), moreover, operate on the basis of buoyancy generated by disk rotation, and using them in a location in which air pressure is only about half that at ground level means that buoyancy would be relatively low, which could lead to HDD failure. Since the PRIMERGY servers used in the correlator control system were designed for use at a maximum altitude of 3000 m, we conducted functional tests to see whether they could operate stably at 5000 m, as described below.

3.1 Anti-heat measures for correlator and servers

Measures had to be taken to reduce the heat generated by the correlator and servers, and since the boards and cabinets of the correlator were custom made, the designs could be tailored to counteract heat. Specifically, they were designed so that the components were arranged in such a way that prevented the generated heat from accumulating in one location, meaning that the heat was efficiently radiated.

The PRIMERGY PC server is a general-purpose product, so using one under physical conditions outside its operating specifications increases the risk of hardware failure. We therefore evaluated its fault tolerance under severe test conditions: a CPU usage rate of 100% in order to generate maximum heat and the atmospheric pressure found at 5000 m (generated using a low-pressure chamber). We placed a sensor inside the server and recorded its readings using a basic input/output system (BIOS). The recorded data showed that no anomalies had occurred. On the basis of these results, we implemented measures to reduce the risk of

failure.

The delay between this testing and the delivery of the servers to be used meant that there had been changes in the product. It was therefore important to also test the delivered servers. The delivered PRIMERGY RX300 S3 servers were the successor to the RX300 S2 model that had been tested. We subjected them to the same fault tolerance testing. No problems were detected, enabling us to deploy them on site without problem.

3.2 Disk-less system at 5000 m

The ALMA Project forbids the use of HDDs to avoid the risk of a disk failure in the low-atmospheric-pressure environment. There are various mechanisms for booting the system without a HDD such as booting from a CD and booting from the network. Considering that unified control would be necessary in the correlator control system to reduce operational mistakes and facilitate maintenance, we selected the second mechanism and adopted the Preboot eXecution Environment (PXE) as the booting system. An OS server (with an HDD) in the project's operations support facility (OSF) located at 3000 m is used to supply an OS to the 35 PRIMERGY servers operating at 5000 m. Having one OS server supply an OS to 35 servers makes for more efficient maintenance.

We decided not to use a CD boot system because doing so would require that we prepare 35 CDs, one for each server, and that we prepare the same number of new CDs whenever the system was updated. It would also require transporting those CDs to an altitude of 5000 m, the altitude at which the servers operate, and loading them onto the servers. This would not be an effective approach.

4. Remote maintenance for reliable operation

In the event that a problem occurs on a server, the nature of the problem needs to be determined by some means such as by examining messages appearing on a console screen. However, traveling to Chile to observe a console screen would involve a 35-hour trip from Tokyo to the OSF via the United States and Santiago and about another hour to reach the array operations site (AOS). In addition, work time at 5000 m is limited given the low oxygen level and the prohibition on residing at the AOS overnight.

Taking these work-environment restrictions into account, we recognized the need for a mechanism that would enable maintenance work on the ACA correlator system in Chile to be performed remotely from Japan, and we implemented such a mechanism. At present, we use a remote control screen as a means of examining the messages appearing after rebooting a down server at the AOS and checking for anomalies. The concept of remote maintenance is shown in **Figure 3**. This system enables us to access the servers located at the AOS.

4.1 Remote maintenance

The work of remote maintenance comprises monitoring, diagnosing, and restoring and can be accomplished by visually inspecting cabinets, working at a console, logging into the system, etc. To perform console operations remotely, it had to be possible to remotely display a console screen including a BIOS screen and to perform keyboard operations, including meta-key remote operations and mouse operations. Furthermore, given the possibility that maintenance operations might need to be performed simultaneously on-site and from a remote site, it was essential that an exclusive-control mechanism be incorporated so that no operations could generate inconsistencies in the system.

Remote maintenance, moreover, requires communications via the Internet in addition to communications via an intranet within the organization, which means that we also had to implement an appropriate level of security by encrypting data on the communication path.

After surveying products that could satisfy these requirements, we selected Raritan's Paragon system as a candidate and performed a validation experiment.

4.2 Validation experiment

This experiment was conducted in two stages. In stage 1, we tested the feasibility of achieving a remote maintenance system using a simulated environment, and in stage 2, we tested operations in a mock network environment using actual models of the proposed equipment.

1) Feasibility test in simulated environment

This test was conducted in 2005 between the NAOJ

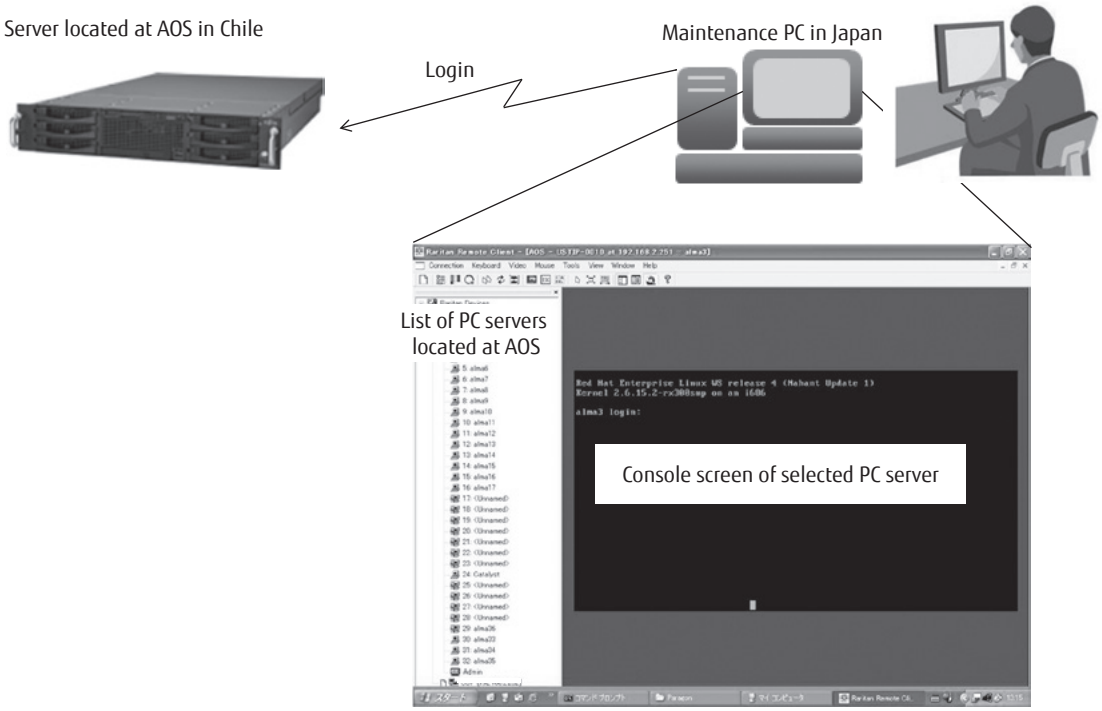

Figure 3
Concept of remote maintenance.

Hawaii observing station and the Fujitsu Makuhari Systems Laboratory (Chiba prefecture, Japan). The test examined operability, function operation, security, and performance.

The computers targeted for remote operations in this test were the PRIMEPOWER 200 and Sun Blade 1000, and the client was Fujitsu's FMV-BIBLO notebook PC. The communication path ran from Hilo, Hawaii, to Sunnyvale, California (DSL/VPN) and from Sunnyvale to Makuhari, Japan (Fujitsu WAN).

A screen shot of the remote maintenance achieved is shown in the lower right of Figure 3. A list of accessible computers is shown on the left side of the screen, and the login screen of the selected computer is shown on the right. The display is virtually the same as that on the screen of the on-site console connected directly to the computer.

When remote operations were performed, it was difficult to determine whether a long response time was caused by a hang-up on the computer side or simply a transmission delay. At the time of this experiment, the network transmission delay between Hawaii and Makuhari was 381 ms. Command line operations had a relatively short delay while mouse operations, which involve the display of graphics, had a relatively long delay. The transmission delay between NAOJ in Mitaka, Tokyo, and the University of Chile, is 400 ms, meaning that performing remote maintenance from NAOJ in Tokyo would result in nearly the same performance as that revealed by this validation experiment.

In summary, stage 1 of the validation experiment demonstrated that remote maintenance was feasible. We thus decided to use the Paragon system for this purpose.

2) Operations test using proposed equipment

In this test, we examined the operation of functions using the same equipment as that slated for system implementation, that is, the PRIMERGY RX300 S3 server. We also used Dummynet software for emulating network delay and frequency bandwidth to evaluate the effect of network delay and narrow bandwidth on performance. This test revealed that BIOS operations and meta-key input for a remotely located PRIMERGY server could be performed on a local PC.

The Paragon system includes a function for encrypting communication data and the remote maintenance system is capable of halting and booting a remotely located computer through functions that

include forced OFF/ON of the power supply. This means that maintenance operations could be performed without having to dispatch personnel to the server site. This capability enables recovery to be quickly achieved after problem occurrence.

In summary, the results of testing operability, function operation, security, and performance in stage 2 of the validation experiment showed that the equipment slated for implementation satisfied the requirements for a remote maintenance system.

5. Conclusion

The ACA correlator system used in the ALMA Project needs to process huge amounts of data in real time at 1.5 Tb/s under extreme environmental conditions. Three technical solutions were used to achieve reliable operation of this system: FPGA technology for high-performance processing, a disk-less system for stable operation, and a remote maintenance capability.

The ACA correlator system has been running continuously since 2012 after an initial test period that began in 2011. No major problems have been encountered, which suggests that the measures described in this paper have been effective.

The ALMA Project is expected to run for about 30 years. During this time, we can expect the observation technology used in this project to improve along with advances in science and technology. It would give us great pleasure if discoveries made as part of the ALMA Project were to help realize our customer's dream of bringing humanity closer to understanding the origins of space.

Finally, we would like to extend our deep appreciation to the staff at NAOJ for their valuable suggestions and comments during the development of this system.

References

1) National Astronomical Observatory of Japan, National Institutes of Natural Sciences: Atacama Large Millimeter/submillimeter Array (ALMA). *http://alma.mtk.nao.ac.jp/e/*
2) The Astronomical Herald, *the Astronomical Society of Japan*, Vol. 106, No. 10, (2013) (in Japanese).
3) K. Abe et al.: Monitor and Control System for ACA Correlator Based on PRIMERGY for ALMA Project. *Fujitsu Sci. Tech. J.*, Vol. 44, No. 4, pp. 418–425 (2008).

Katsumi Abe
Fujitsu Ltd.
Mr. Abe is engaged in the development of observation control systems for the National Astronomical Observatory of Japan (NAOJ).

Takahiro Hiyama
Fujitsu Ltd.
Mr. Hiyama is engaged in the development of observation control systems for the National Astronomical Observatory of Japan (NAOJ).

Junpei Tsutsumi
Fujitsu Ltd.
Mr. Tsutsumi is engaged in the development of ground systems for earth observation satellites.