

Monitor and Control System for ACA Correlator Based on PRIMERGY for ALMA Project

● Katsumi Abe ● Sachiko Kawase ● Mitsuhiro Moriya

(Manuscript received March 19, 2008)

The National Astronomical Observatory of Japan is constructing a large interferometer called the Atacama Large Millimeter/submillimeter Array (ALMA) on a plateau near the Atacama Desert in Chile at an elevation of about 5000 m, in cooperation with the National Radio Astronomy Observatory of the USA and the European Southern Observatory. In cooperation with FFC Limited, Fujitsu is taking charge of development of the Atacama Compact Array (ACA) correlator used in correlation processing of data collected with the interferometer. Fujitsu is also charged with developing the monitor and control system based on Fujitsu's Linux PRIMERGY servers. The PRIMERGY RX300 S3 used for this system must operate stably in a severe environment at an elevation of about 5000 m and under atmospheric pressure of about 0.5 atm. We employed a diskless system for stable operation, thus making the system reliable even at low atmospheric pressure. We also conducted long-time running tests in an environment similar to the Atacama high-altitude environment, and adopted a remote maintenance system in order to make error handling and recovery much easier. This paper briefly introduces the ALMA project, and then describes the means of stable monitor and control system operation.

1. Introduction

We are developing a computer system that can stably operate on-site at an elevation of 5000 m without requiring resident support engineers.

The National Astronomical Observatory of Japan (NAOJ), a member of National Institutes of Natural Sciences (NINS), is constructing a large interferometer on a plateau near Chile's Atacama Desert at an elevation of about 5000 m, in cooperation with the National Radio Astronomy Observatory of the USA and the European Southern Observatory. This project is called the Atacama Large Millimeter/Submillimeter Array (ALMA) project.¹⁾ The ALMA project was launched in 2002 by the USA and Europe, and with Japan's participation from 2004, three major groups are now advancing the project. The respective countries in charge are rapidly devel-

oping the hardware and software necessary for preliminary scientific testing slated to begin in 2010, and for full-scale system operation planned for 2012.

Figure 1 shows a rendering of ALMA to be constructed.

NAOJ takes charge of 16 antennas (among 80 total antennas), four frequency bands (out of seven frequency bands), the Atacama Compact Array (ACA) correlator (for ultrahigh-speed correlation processing of received data),²⁾ and the correlator's monitor and control system.

In cooperation with FFC Limited, Fujitsu began developing the correlator described above in 2004, and then in 2005 launched development of the monitor and control system based on Fujitsu's Linux PRIMERGY servers. **Figure 2** shows an outline of the correlator and its monitor and control system.

The monitor and control system must run continuously 24 hours a day, except on scheduled days of maintenance. Should any trouble occur during operation, the system must be recovered as soon as possible.

Since the monitor and control system will be installed at a site located 5000 m above sea level and under very low atmospheric pressure, however, the capacitors and power supply units of this system are liable to break down or overheat due



Figure 1
Rendering of ALMA facilities on plateau near Atacama Desert.

to largely reduced cooling efficiency. Therefore, we needed to take various measures to operate the monitor and control system stably under such conditions.

A hard disk drive (HDD) as storage media of computers utilizes the buoyancy (air pressure between a head and disk) caused by disk rotation.³⁾ However, the air pressure at the installation site is lower than that at ground level because atmospheric pressure at the installation site is about half that at ground level. Such insufficient buoyancy is liable to result in frequent HDD faults. To reduce such risk, the ALMA project prohibits the use of HDDs in the on-site system at an elevation of 5000 m. This restriction, however, poses a problem: How to boot Linux servers with no HDDs.

In addition, Chile and Japan are antipodes (at opposite points) on the earth. It therefore takes about 35 hours to travel from Narita in Japan to the ALMA installation site in Chile. Moreover, the astronomical observatory's safety standards restrict humans from working more

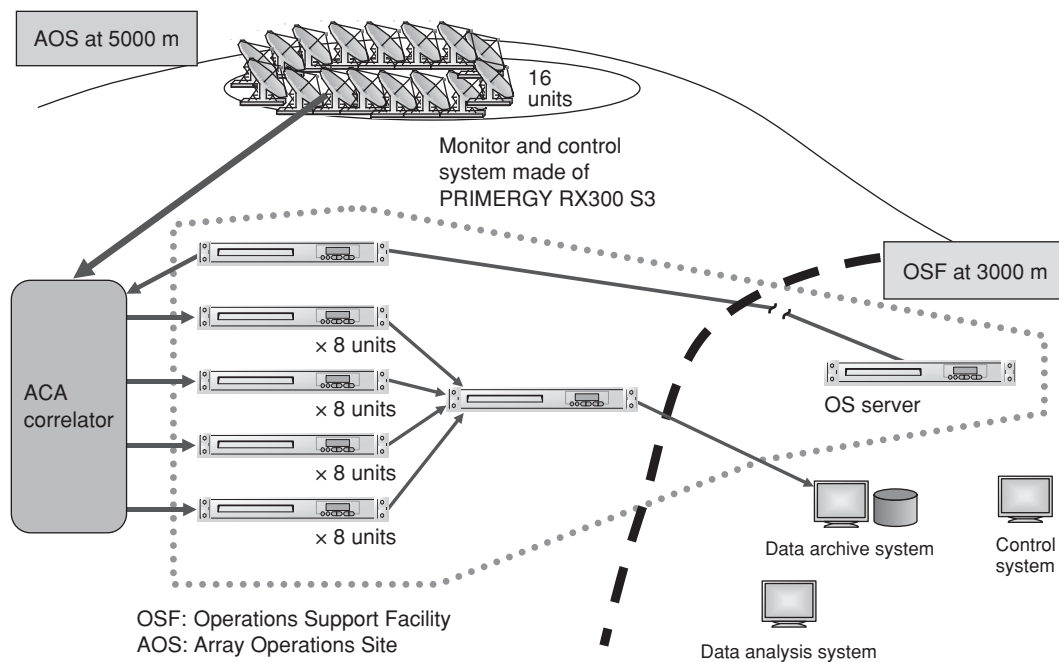


Figure 2
Monitor and control system of ACA correlator for ALMA.

than 10 hours a day at an elevation of 5000 m, and also prohibit humans from working at night.

Therefore, our mission in terms of addressing the problem was how to maintain the monitor and control system installed at a remotely located site where work hours are restricted.

This paper describes the measures we planned and executed to achieve stable operation of the monitor and control system installed at a site in Chile at an elevation of 5000 m as follows:

- Hedging the risk of hardware errors
- Booting computers that have no HDDs
- Conducting remote maintenance of the monitor and control system installed at the remote site

2. Verification of PRIMERGY operation

The Linux PRIMERGY servers used in this system can be operated normally at a maximum elevation of 3000 m. Consequently, we had to verify beforehand that PRIMERGY servers could also operate stably in a plateau-like environment at 5000 m.

2.1 Prior verification of PRIMERGY RX300 S2

Generally speaking, using a computer in a physical environment not compliant with its specifications may frequently result in hardware errors after beginning a production run. To prevent such trouble, it is important to verify the computer's fault tolerance in a similar environment beforehand. Then taking measures based on the verification results can reduce the risk of frequent hardware errors after starting a production run. We conducted such prior verification of the fault tolerance of the PRIMERGY RX300 S2 at Fujitsu's plant under the following conditions:

- 1) Operation term and time
October 22, 2005 to November 10, 2005
About 480 hours total
- 2) Verification site
Pressure-reducing chamber facilities at

Fujitsu's Nasu Plant

3) Physical environment

Atmospheric pressure: Equivalent to that at an elevation of 5000 m

Temperature: 27°C

4) Load

Use of endurance test tool

Load was applied so that the CPU use rate was always 100% in order to maximize heat output. PRIMERGY has a thermal sensor to detect any temperature exceeding the threshold value for recording in BIOS. The recorded temperatures can be referenced from the BIOS screen. After completing the test, we confirmed that no hardware errors had occurred.

2.2 Verification of PRIMERGY RX300 S3 used in the system

After determining which machines for actual operation, we must, in certain cases, verify the machines again before delivery to the site. This is because, if the actual delivery time is much later than the verification time, parts used in the tested machines may differ from those used in the delivered machines.

We delivered the PRIMERGY RX300 S3 equipped with a dual-core Xeon CPU for use as a production-run machine in the system. The PRIMERGY RX300 S3 is the successor model to the PRIMERGY RX300 S2 with a single-core Xeon CPU. Therefore, we also verified the fault tolerance of the RX300 S3 in the same way as for the RX300 S2, and confirmed that no hardware errors had occurred.

Also, to minimize hardware errors, we tried to minimize the hardware configuration.

3. PRIMERGY RX300 S3 operation without a HDD at 5000 m

As described in Section 1, the ALMA project prohibits HDDs from being used in facilities located at an elevation of 5000 m in order to reduce the risk of disk errors in an environment under

low atmospheric pressure. To start the system without using HDDs, we adopted a technique called the Preboot Execution Environment (PXE) bootstrap. Server-specific information and file information must be managed using devices other than HDDs. The following introduces the bootstrap system of Linux servers that can operate normally at 5000 m, and describes the file management system of these servers.

3.1 Use of the PXE bootstrap system

There are generally three methods of booting a computer: Bootstrap with a HDD, bootstrap with a CD/DVD drive, and bootstrap via a network.

We employed the PXE bootstrap system (bootstrap via a network) in the monitor and control system. This system uses an OS server to supply the operating system to 35 Linux servers installed in the plateau facilities (AOS) at an elevation of 5000 m. This OS server is installed in the foothill facilities (OSF) located at about 3000 m and is equipped with a HDD (Figure 2).

Supplying the operating system to 35 Linux servers from one OS server reduces the management load.

We did not adopt the bootstrap system using the CD/DVD drives because it would entail a high management load. In other words, we would have to create CD/DVD disks for the 35 Linux servers. Moreover, when updating the system in the future, we would have to create 35 new CD/DVD disks, carry these disks to the site, and then replace the old disks. This would certainly prove inefficient working on 35 Linux servers located at an elevation of 5000 m.

Figure 3 outlines the flow of PXE bootstrap, which is executed as follows:

- 1) The diskless client starts up, and then BIOS starts. At the end of POST processing, the PXE boot agent is called. (The PXE boot agent is stored on the NIC ROM.)
- 2) By using dynamic host configuration protocol (DHCP), the PXE boot agent obtains

an IP address and related information via BOOTP.

- 3) The PXE boot agent checks the PXEClient character string in the tag of the BOOTP response packet.
- 4) A specified boot loader (pxelinux) file is then downloaded via TFTP.
- 5) By using the downloaded pxelinux, the PXE boot agent starts bootstrap operation.
- 6) The pxelinux downloads the kernel.
- 7) The kernel starts.
- 8) The init process starts. Various services (daemons) are started with the rc scripts.

3.2 Server file management

The Linux server must hold IP addresses, system logs, password files, and other necessary information as server-specific files. To hold such server-specific files, the Linux server can use external storage devices or the network file system (NFS). The usable external storage devices are CD/DVD drives, USB memory, and solid-state disks. The CD/DVD drives used to hold server-specific data are inefficient in terms of management and have slow write speeds. USB memory is inexpensive, but offers small storage capacity and limits the maximum number of write operations. Solid-state disks were not

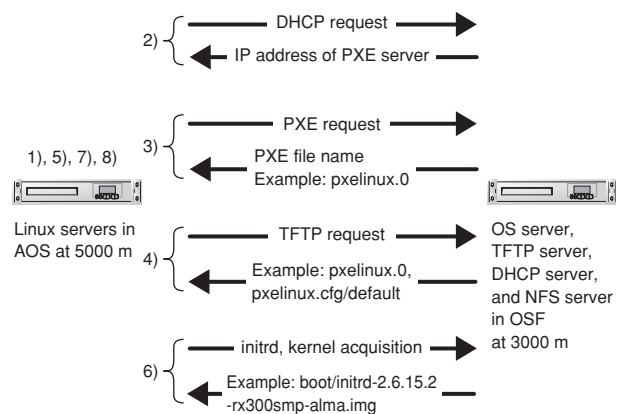


Figure 3 Mechanism of PXE bootstrap.

commercially available when this system was examined. Although solid-state disks are now being sold, the price-to-capacity ratio is rather high. On the other hand, NFS is supported as a standard function of the Linux operating system, and one NFS server can centrally control server-specific files.

In considering the evaluation results above, we decided to adopt the OS server and NFS system to store files specific to the Linux servers. The OS server has both file systems that are specific and common to the 35 Linux servers installed at 5000 m. Each Linux server gets a specific file system and a common file system by using the NFS mount.

4. How to maintain the system

It takes about 35 hours for an engineer to travel from Japan to reach the correlator equipment at the Array Operations Site (AOS) in Chile at an elevation of 5000 m, and about one hour to go from the Operations Support Facility (OSF) to AOS. Moreover, the maximum work time at AOS is restricted and visiting AOS on the plateau at night is prohibited. Given the lower oxygen level at AOS, an engineer’s thinking may be impaired. To minimize work errors under such severe conditions, maintenance work requiring high-level concentration must be performed at other than AOS as far as possible.

Table 1 lists the maximum work times permitted at the facilities in Japan, and at AOS and OSF in Chile. As listed in this table, the only time zone in which engineers at the facilities in

Japan and engineers at AOS and OSF in Chile can cooperate is 22:00 to 04:00 (Japan Standard Time). Similarly, the only time zone in which engineers at the facilities in Japan and engineers at OSF in Chile can cooperate is 22:00 to 10:00 (Japan Standard Time). Expanding such limited work time requires remote maintenance to be performed from remote facilities; that is, OSF in Chile or from facilities at NAOJ in Mitaka (Tokyo).

4.1 Mechanism of remote maintenance

Remote maintenance work is divided into supervision, diagnosis, and recovery.

Remote maintenance work can be performed by visually inspecting the cabinets and operating the console and system after logging in. Web cameras and other means can be used to remotely inspect the cabinets.

Remote console operation requires using the remote display of console screens (including BIOS screens), operating the keyboard (including remote operations of meta keys), and using the mouse.

Logging into the system requires that an environment be prepared for login via a network.

Cooperative maintenance from both local and remote sites may occur at the same time. Such simultaneous maintenance necessitates exclusive control to prevent conflicts in the system.

Remote maintenance also requires communication via the Internet and intranets. Security with communication path encryption must be

Table 1
Maximum work times permitted at facilities in Japan, and at AOS and OSF in Chile.

	Japan Standard Time	22:00–04:00	04:00–10:00	10:00–16:00	16:00–22:00
Japan	Work	○	○	○	○
AOS in Chile	Field standard time	09:00–15:00	15:00–21:00	21:00–03:00	03:00–09:00
	Work	○	×	×	×
OSF in Chile	Work	○	○	△	△

○: Work permitted ×: Work prohibited △: Other than standard work time

maintained for Internet communications.

We researched products that satisfy the remote maintenance conditions above. We then selected the Paragon unit (manufactured by Raritan Computer Inc.) as a candidate, and conducted demonstration experiments on this unit.

Figure 4 shows the concept of remote maintenance. The head office of NAOJ in Mitaka (Tokyo) can directly access each computer at AOS (located at 5000 m) and OSF (located at 3000 m). OSF at 3000 m above sea level can directly access AOS at 5000 m. Accessing AOS also requires using the physical network facilities of OSF.

4.2 Demonstration experiments

The demonstration experiments were conducted in two stages:

1) First stage

The first stage of demonstration experiments was conducted between NAOJ's

observatory in Hawaii and Fujitsu's Makuhari System Laboratory in Chiba, Japan. Through these experiments we confirmed the operability (user interface, UI), functions, security, and performance of the Paragon unit.

The remote computers used were the PRIMEPOWER200 and SunBlade1000. The local computers used were FMV-BIBLO units (running Windows XP). The communication path used was "Hawaii (Hilo) - [DSL/VPN] -> California (Sunnyvale) - [Fujitsu WAN] -> Japan (Makuhari)".

We confirmed the following results from these experiments.

Figure 5 shows the screen captured in the experiments using PRIMERGY. The left side shows a list of connectable computers. The right side shows the login window to the connected computer. As shown in this figure, the UI on the screen approximates that on the console screen of the monitor that is directly connected to the

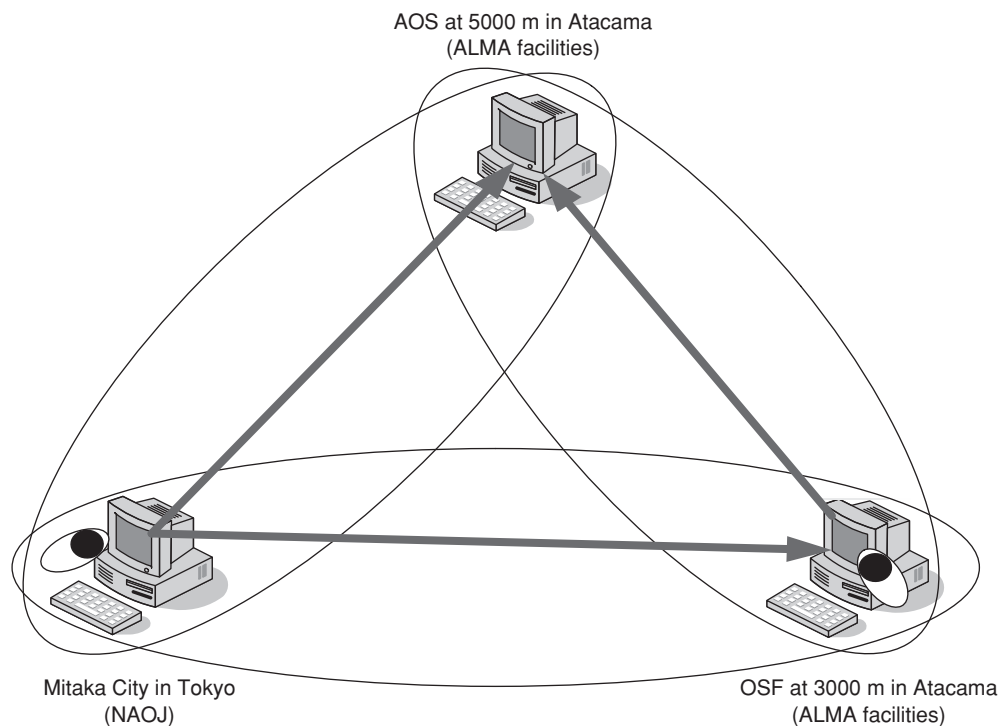


Figure 4
Overview of remote maintenance.

computer.

If the computer's response to remote key input is considerably delayed, it is difficult to determine the current status (whether computer hang-up or network latency). In our experiments, the network latency between Hawaii and Makuhari was 381 ms. We felt somewhat stressed over the command line operation. At the same time, we were more concerned about mouse operation because image display caused longer latency. The latency between Mitaka and the University of Chile was about 400 ms. We expected that remote maintenance conducted from NAOJ would provide latency similar to that in the demonstration experiments described above.

In the first stage of the demonstration experiments, we found that the UI used, functions, security, and performance all satisfied the necessary conditions of remote maintenance. We then decided to use the Paragon unit in the computer system.

2) Second stage

Raritan Computer Inc. provided us with the same Paragon unit as that to be installed, and we verified the unit's functions by using the PRIMERGY RX300 S3 (which would also be used in a production run). To verify the functions,

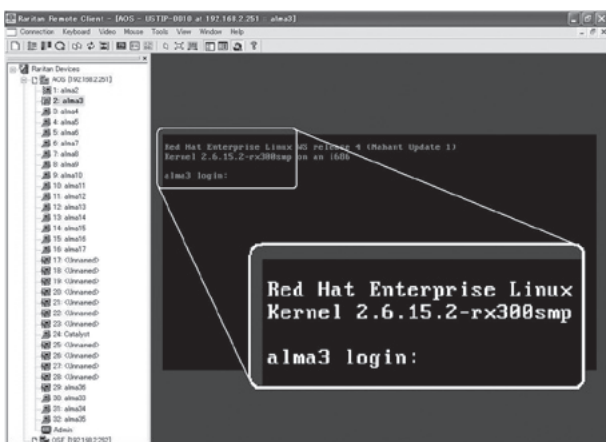


Figure 5 Console image displayed by Paragon.

we used free “DummyNet” software that could set a network latency and bandwidth virtually. DummyNet can be used to create a virtual Internet environment. We increased the network latency and narrowed the bandwidth by using DummyNet, and then evaluated the effects on performance. We consequently confirmed that both BIOS and meta key operations could be performed normally from a local PC connected to a remote Linux server via the Paragon unit.

5. Future plan and problems

We will continue testing of the monitor and control system connected to other on-site systems in 2009. We expect that remote support from Japan will be effectively provided in case problems occur in the tests. The operation results after the production run will be evaluated in 2009 or later.

Computer systems have four operation steps from the occurrence of an error to recovery. These steps are error detection, error identification, handling, and handling result notification. In the monitor and control system, three parties (NAOJ, the ALMA system administrator, and Fujitsu) must act in each of these steps. A lack of successful cooperation among these three parties may result in further errors due to conflicting operations. To reduce such risk, work arrangements, notification, and real-time confirmation of status must be made. Our future problem is to construct a mechanism for reducing said risk, in cooperation with the customer.

6. Conclusion

This paper described the following measures for achieving stable operation of the ALMA-project computer systems without requiring resident support engineers in a severe environment at an elevation of 5000 m:

- Operation test in pressure-reducing chamber (0.5 atm)
- Mechanism of system startup without HDDs
- Mechanism of remote maintenance

The mechanism of remote maintenance enables the startup/shutdown (including forcible power-on/power-off) of remote computers. Because the remote computers can be operated from a local console screen, maintenance can be performed without having to visit remote Linux server installation sites. Moreover, the remote maintenance function reduces the time from when trouble occurs to when recovery work can begin.

Acknowledgement

We would like to express our gratitude to the persons concerned at NAOJ for their helpful suggestions and advice on the development of this system.



Katsumi Abe
Fujitsu Ltd.

Mr. Abe received the Ph.D. in Engineering from Shizuoka University, Hamamatsu, Japan in 1996. He joined Fujitsu Ltd., Chiba, Japan in 1996, where he has been engaged in development of electromagnetic field analysis software.

E-mail: abe.katsumi@jp.fujitsu.com



Mitsuhiro Moriya
Fujitsu Ltd.

Mr. Moriya received the B.S. degree in Applied Chemistry from University of Yamanashi, Kofu, Japan in 1977. He joined Fujitsu Ltd., Tokyo, Japan in 1977, where he has been engaged in development of operating systems and support of computing systems for researchers and developers.

E-mail: mmoriya@jp.fujitsu.com



Sachiko Kawase
Fujitsu Ltd.

Ms. Kawase received the M.S. degree in Information and Computer Sciences from Nara Women's University, Nara, Japan in 2003. She joined Fujitsu Ltd., Chiba, Japan in 2003, where she has been engaged in development of astronomical systems.

E-mail: kawase.sachiko@jp.fujitsu.com

References

- 1) National Institutes of Natural Sciences (NINS), National Astronomical Observatory of Japan (NAOJ): ALMA (Atacama Large Millimeter/submillimeter Array). <http://www.nro.nao.ac.jp/alma/E/>
- 2) A. Richard Thompson, et al.: Interferometry and Synthesis in Radio Astronomy. A WILEY-INTERSCIENCE PUBLICATION, 1986.
- 3) Y. Goto, et al.: Head Disk Interface Technologies for High Recording Density and Reliability. *FUJITSU Sci. Tech. J.*, Vol.42, No.1, pp.113-121 (2006).