特集

ALMAを支える相関器制御システム

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

あらまし

自然科学研究機構国立天文台様は，米欧の天文台と共同で，南米チリのアタカマ砂漠近くの高度5000 mの高原に，大型電波望遠鏡の建設を進めている。この計画はALMA（Atacama Large Millimeter/submillimeter Array）計画と呼ばれている。富士通は，（株）FFC（エフ・エフ・シー）と共同で，大型電波望遠鏡で収集したデータを干渉処理する専用計算機である相関器（ACA Correlator），およびLinuxサーバPRIMERGYをベースとする相関器制御システムの開発を担当している。相関器制御システムで用いるPRIMERGY RX300 S3は，チリの高度5000 m，0.5気圧という過酷な環境下で運用する必要があり，安定した運用を実現するために，ディスクレスシステムの採用，同等高度環境下での長時間稼働テストの実施，故障時の速やかな対応のためのリモートメンテナンスシステムの導入を行った。

本稿では，ALMA計画を簡単に紹介し，相関器制御システムの選定から導入において実施した安定運用実現のための施策について述べる。

Abstract

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.

阿部勝己（あべ　かつみ）

計算科学ソリューション統括部
所属
現在，電磁波解析ソフトウェアの開発に従事。

河瀬祥子（かわせ　さちこ）

科学システムソリューション統括部
所属
現在，ALMA計画の相関器用制御システムの開発に従事。

森屋光弘（もりや　みつひろ）

科学システムソリューション統括部
所属
現在，国立天文台業務システムの開発に従事。

ま え が き

高度5000 mのサポート要員が常駐しない地域で，計算機システムを安定運用させなければならない。

自然科学研究機構国立天文台様（以下，国立天文台）では，米欧の天文台と協力し，南米チリのアタカマ砂漠に近い高度5000 mの高原に，大型電波望遠鏡の建設を進めている。この計画はALMA（Atacama Large Millimeter/submillimeter Array）計画[1]と呼ばれている。ALMA計画は，2002年から米欧により，2004年からは日本も参加して，3極体制で進められており，2010年の初期科学運用，2012年の本運用を目指し，現在，各国が担当するハードウェア，ソフトウェアの開発が急ピッチで進められている。

図-1は，建設予定の国際大型電波望遠鏡のイメージである。

国立天文台が担当するのは，全80基のうち16基のアンテナ，七つの周波数バンドのうちの四つの周波数バンド，受信データを超高速に干渉処理する相関器[2]｛ACA（Atacama Compact Array）Correlator｝，および相関器を制御する相関器制御システムなどである。

富士通は，（株）FFC（エフ・エフ・シー）と共同で2004年から上記相関器を，2005年から富士通LinuxサーバPRIMERGYをベースとする相関器制御システムの開発を担当している。**図-2**は，相関器および相関器制御システムの概要である。

相関器制御システムは，スケジュールに組まれた保守日を除き，24時間連続運転される。運用時間中に障害が発生した場合には，できる限り速やかな復旧が要求される。

しかし，高度5000 mの気圧の低い環境下では，当該システムの構成要素であるコンデンサや電源ユニットの故障，冷却効率の低下による装置の過熱，がリスクとなる。このため，相関器制御システムを安定して稼働させるために，以下に示すような対策をとっている。

すなわち，計算機の記憶媒体として使われるハードディスクドライブ（HDD）は，ディスクの回転によって生じる浮力（空力）を利用している[3]。気圧が平地の半分ほどしかない場所では，平地よりも浮力が小さい。その結果，HDDの障害多発のリスクが生じる。リスクを減らすため，ALMA計画の運用システムでは，5000 mの高原でのHDDの使用を禁じている。そのために，HDDを持たないLinuxサーバを，どのようにブートするのかが課題となる。

また，チリは，ほぼ日本の真裏に位置しており，成田からALMA大型電波望遠鏡の設置場所までの移動時間は，約35時間も要する。さらに，天文台の安全基準では，人間が5000 mで作業する時間は制限されており（10時間を上限とする），かつ夜間の作業は禁止，となっている。

このように，簡単に出向くことができない場所に

図-1　ALMAの完成予想図
Fig.1-Rendering of ALMA facilities on plateau near Atacama Desert.

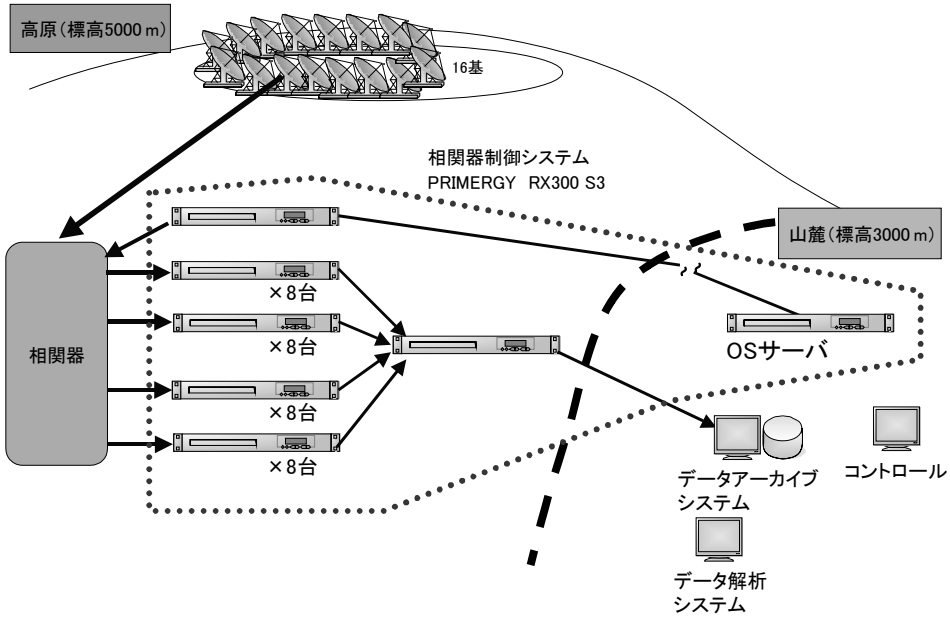

図-2　相関器および相関器制御システム
Fig.2-ACA correlator and its monitor and control system for ACA Correlator for ALMA.

あって作業時間にも制限がある中で，相関器制御システムをメンテナンスするには，どうすればよいのかが課題となる。

本稿では，南米チリの高度5000 mで相関器制御システムを安定稼働させることを目的として，以下の観点から立案，実行した施策について述べる。

・ハードウェア障害のリスクヘッジのための施策
・HDDを持たない計算機をブートするための仕組み
・地理的に離れた場所に設置した相関器制御システムを遠隔でメンテナンスするための，リモートメンテナンスの仕組み

PRIMERGYの稼働性の検証

本システムで採用するLinuxサーバPRIMERGYの稼働条件は，最大高度3000 mとなっている。このため，高度5000 mの高原で安定稼働ができるかを事前検証した。

● PRIMERGY RX300 S2の事前検証

一般的に，計算機を仕様外の物理的環境下で使用すると運用開始後にハードウェア障害が多発するリスクが生じる。このため，計算機を選定する前に，擬似環境で計算機の耐障害性を評価することが重要である。これに関しては，評価結果に基づいて，対策を講じることで，運用開始後の障害多発のリスクを減らすことが可能である。以下に，富士通の工場で行った，PRIMERGY RX300 S2の耐障害性の事前検証の条件と結果を示す。

(1)　動作期間と時間
　2005年10月22日～2005年11月10日
　合計，約480時間
(2)　確認場所
　富士通那須工場・減圧チャンバ施設
(3)　物理環境
　気圧：高度5000 m相当
　温度：27℃
(4)　負荷
　耐久テストツールを利用

負荷は，発熱量が最大となるよう，CPUの使用率が常時100％となるようにした。PRIMERGYには温度センサが搭載されており，温度が閾値を超えると，BIOSに記録される。この情報は，BIOS画面から参照可能である。

試験後，ハードウェアに異常が起こっていないことを確認した。

● システム採用機 PRIMERGY RX300 S3の検証

運用計算機として採用が決まった後にも，現地に納入する前に検証を行うことが重要である。事前検証の時期と，実際に搬入される時期が異なり，計算機に使用されている部材が異なる場合があるからである。

本システムで実際に採用したPRIMERGY RX300 S3（デュアルコアXeon CPU使用）は，RX300 S2（シングルコアXeon CPU使用）の後継機種である。RX300S2と同じように，耐障害性評価試験を行い検証した結果，問題は検出されなかった。

また，本システムでは，ハードウェア障害のリスクをできる限り減らすために，必要最低限のハードウェア構成とすることを念頭に置いた。

高度5000 mのPRIMERGY RX300 S3のHDDレス運用

まえがきで示したように，低圧環境下でのディスク障害のリスクから，ALMA計画の運用システムでは，高度5000 mでHDDを使用することを禁じている。このため，HDDを使用せずにシステムを起動するためにPXE（Preboot eXecution Environment）ブートと呼ぶ手法を採用した。また，HDDを持たないため，サーバ固有情報やファイル情報などをHDD以外の装置で管理しなければならない。以下に，高度5000 mで稼働するLinuxサーバへのブート方式とそのサーバのファイル管理について説明する。

● PXEブート方式の採用

計算機をブートする仕組みは，一般的には，3種類ある。HDDからのブート，CD/DVDからのブート，ネットワーク経由でのブートである。

相関器制御システムではネットワーク経由でブートする方式であるPXEブート方式を採用することとした。高度5000 mで稼働する35台のLinuxサーバへOSを供給するOSサーバを，高度3000 mの拠点（山麓施設）に導入する。OSサーバはHDDを装備する（図-2）。

1台のOSサーバが35台のLinuxサーバにOSを供給することで，管理の手間を省くことが可能となる。

CD/DVDブートを不採用とした理由は，管理の手間の増大である。CD/DVDブートの場合，ブートするLinuxサーバ台数分のCD/DVD媒体を作成する必要がある。また，システムを更新する際には，新たに同じ枚数の媒体を作成し直し，Linuxサーバが稼働する場所に持っていき，媒体を載せ換える作業が必要となる。これらの作業を高度5000 mの35台のLinuxサーバに対し実施するのは非効率的である。

図-3は，PXEブートシーケンスの概略である。PXEブートの手順は以下のとおりである。

(1) Diskless clientを起動すると，BIOSが起動され，POST処理の最後に，PXE boot agentが呼び出される。（PXE boot agentは，NICのROMに格納）

(2) PXE boot agentは，DHCPで，IPアドレスなどをBOOTP経由で取得。

(3) PXE boot agentは，BOOTP 応答パケットに含まれるタグに，“PXEClient”の文字列を調べる。

(4) 決められたboot loader（pxelinux）ファイル

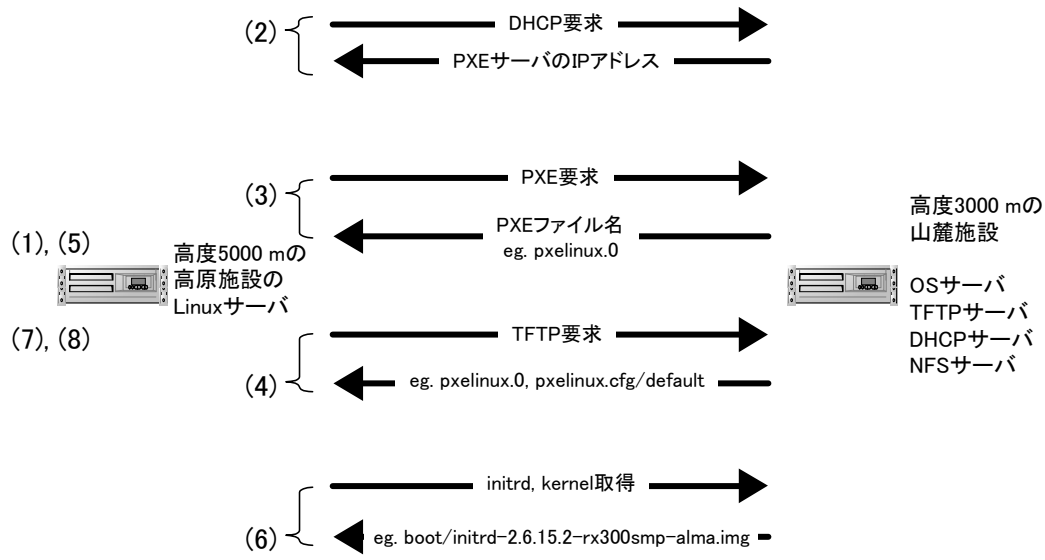

図-3　PXEブートの仕組み
Fig.3-Mechanism of PXE bootstrap.

をTFTP経由でダウンロードする。
(5) PXE boot agentは，ダウンロードしたpxelinuxを利用して起動。
(6) pxelinuxは，カーネルをダウンロード。
(7) カーネルが起動する。
(8) initプロセスを起動し，rcスクリプトによる各種サービス（デーモン）を起動する。

● サーバのファイル管理

Linuxサーバでは，IPアドレスやシステムログ，パスワードファイルといった情報を，サーバ固有のファイルとして保持する必要がある。Linuxサーバが固有のファイルを保持するためには，外部記憶装置やNFS（Network File System）を使用する手段がある。外部記憶装置には，CD/DVD，USBメモリ，ソリッドステートディスクなどがある。サーバ固有の情報を保持するために，CD/DVDを利用することは，管理が非効率的，書込み速度が遅いといった問題がある。USBメモリは廉価であるが，容量が少ない，書込み回数に上限があるなどの問題がある。ソリッドステートディスクは本システムの検討段階では市販されておらず，最近，出回り始めたが，現在でも，価格対容量が高い。NFSは，OS標準であり，1台のNFSサーバによる一元管理が可能である。

以上の評価結果を踏まえて，コスト面，管理の効率化の利点が大きいことを評価して，本システムでは，Linuxサーバ固有のファイルを保持するために，OSサーバをサーバとするNFSを採用した。OSサーバは，高度5000 mにあるLinuxサーバに対し，35台間で共通なファイルシステム，それぞれに固有なファイルシステムを持つ。Linuxサーバは固有ファイルシステム，共通ファイルシステムをNFSマウントしている。

どうやって，メンテナンスするのか

高度5000 mの高原の相関器へは，日本から約35時間，現地山麓施設（以下，山麓施設）からは約1時間を要する。高原の現場で作業できる時間は制限されている。また，夜間の高原滞在は禁止されている。さらに，現地高原の施設（以下，高原施設）では，酸素濃度が薄いことから，思考が緩慢になる。これらの厳しい条件の中で，作業ミスをなくすために，思考を要するメンテナンスなどの作業は，極力，高原施設以外で行う必要がある。

表-1は，日本と高原施設，山麓施設での作業の可能時間を示している。この表から，日本，高原施設，山麓施設で共同して作業を行える時間帯は，日本時間では，22:00～4:00に限られることが分かる。日本と山麓施設で共同して作業を行える時間帯は，日本時間では，22:00～10:00である。このため，時差と作業環境の制約から，相関器制御システムから遠く離れた場所にある山麓施設や国立天文台（東京都三鷹市）から，リモートでメンテナンス作業を行えるための仕組みが必要である。

● リモートメンテナンスの仕組み

リモートメンテナンスの内容は，監視，診断，復旧作業に分類できる。

上記の作業は，筐体（きょうたい）の目視，コンソール，システムへのログインを通して実施可能である。筐体の目視をリモートで行うことは，Webカメラなどのカメラでの監視によって実現可能である。

コンソール作業をリモートで行うためには，リモートで，BIOS画面を含めたコンソール画面の表示，メタキーのリモート操作を含めたキーボード操作，マウス操作が行えることが必要である。

システムへログインして作業するためには，ネットワーク経由でログインできる環境が必要である。

計算機の稼働場所とリモートの両方から，同時に，共同で，メンテナンス作業を行う場面も想定される。システムに矛盾を起こさせないためには，排他制御も必要である。

さらに，リモートメンテナンス作業では，組織のイントラネットを通じた通信のみでなく，インターネットを介在した通信も必要となる。この場合，通信経路の暗号化によるセキュリティの確保が必要である。

上記の要件を満足する製品を調査した結果，

表-1 各拠点（日本，現地高原施設，現地山麓施設）での作業可能時間

日本	日本標準時間	22:00-04:00	04:00-10:00	10:00-16:00	16:00-22:00
	作業	○	○	○	○
現地高原施設	現地標準時間	09:00-15:00	15:00-21:00	21:00-03:00	03:00-09:00
	作業	○	×	×	×
現地山麓施設	作業	○	○	△	△

○：作業可　×：作業不可　△：通常作業時間外

Raritan社のParagon装置を候補とし，この製品に対して次節に示す実証実験を行った。

リモートメンテナンスの概念を**図-4**に示す。東京三鷹市の国立天文台本部からは，高度5000 mの高原施設，高度3000 mの山麓施設それぞれの計算機に直接アクセス可能である。高度3000 mの山麓施設からは，高度5000 mの高原施設に直接アクセス可能である。ただし，高度5000 mへアクセスする際の物理ネットワークは，山麓施設のネットワーク装置を介する。

● 実証実験

実証実験は，以下に示すように2段階に分けて行われた。

(1) 第1段階

国立天文台ハワイ観測所と富士通幕張システムラボラトリ（千葉）間で実験を行った。確認のポイントは，操作性（UI），機能，セキュリティ，性能である。

リモートで操作対象に使用した計算機は，PRIMEPOWER200とSunBlade1000である。ローカルでは，FMV-BIBLO（Windows XP）を使用した。経路は，ハワイ（ヒロ）－［DSL/VPN］→カリフォルニア（サニーベイル）－［富士通WAN］→日本（幕張）である。

実験の結果，以下の事項を確認することができた。

・PRIMERGYでの実験時の画面キャプチャを**図-5**に示す。画面の左側には，接続可能な計算機の一覧が表示され，画面の右側の広い部分は接続した計算機へのログイン画面である。このように，UIについては，計算機に直接ディスプレイを接続して，コンソール画面を見ているのと，ほぼ変わらない見え方である。

・リモートで作業を行う場合，リモートからのキー入力に対する計算機の反応時間が長くなると，計算機がハングアップしているのか，伝送遅延によるものかの判断が難しい。実験時，ハワイ－幕張間のネットワーク遅延は，381 msであった。実験では，コマンドライン操作は若干ストレスを感じる程度であったが，マウス操作は画像表示を伴うため，遅延時間が大きいことが分かった。三鷹－チリ大学間の遅延時間は400 ms程度であるので，国立天文台からリモートメンテナンスを実施する場合においても，実証実験に近い性能を得られると推測できる。

実証実験の第1段階において，UI，機能，セキュリティ，性能の面から，リモートメンテナンスの要求仕様を満足できると判断し，本計算機システムでは，Paragon装置を採用することを決定した。

(2) 第2段階

日本Raritan社から導入予定機器と同一機器を使用し，機能検証を行った。用いた計算機は，運用計算機PRIMERGY RX300 S3である。機能検証では，仮想的にネットワークの遅延・帯域を設定可能なフリーソフトウェアDummyNetを使用した。DummyNetを用いることで，仮想的にインターネット環境を構築することができる。このDummyNetを用いて，ネットワーク遅延を延ばし，帯域幅を狭めることによる性能への影響を評価した。

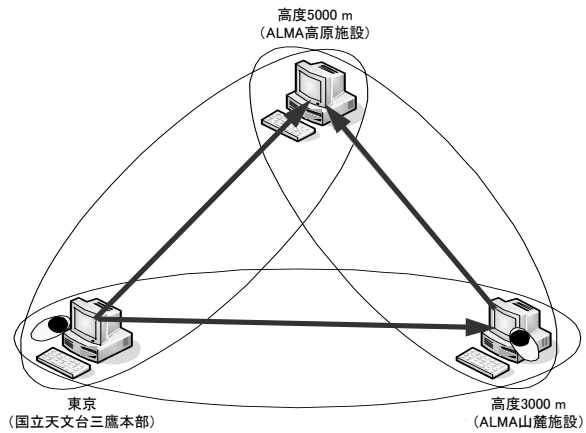

図-4 リモートメンテナンスの概念
Fig.4-Overview of remote maintenance.

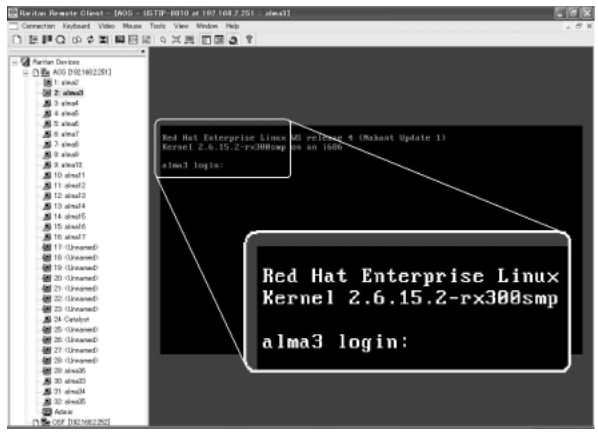

図-5 Paragonを使用したコンソール画面
Fig.5-Console image by Paragon.

その結果，Paragon装置では，リモート操作用PCから，遠隔にあるLinuxサーバに対し，BIOS操作，メタキー入力が可能であることが確認できた。

今後の計画と課題

相関器制御システムは，2008年度に現地での他システムとの結合試験を予定している。試験作業時に問題が発生した場合においても，日本からのサポートに威力を発揮するものと期待している。稼働開始後の効果の測定は2008年度以降となる。

計算機システムでは，事象が発生してから，正常な状態に回復するまでに，四つの段階を経る。事象の発見，事象の切分け，対処，対処した結果の周知，である。相関器制御システムでは，これらの段階ごとに，国立天文台，ALMA側運用支援者，富士通の3者が登場する。3者間での協調がうまくいかない場合，オペレーションの矛盾によりシステムに更なる障害を引き起こすリスクがある。リスクを減らすためには，作業の段取り，周知，状況のリアルタイムでの把握が重要である。そのための仕組みを，お客様を交えて構築することが今後の課題である。

む　す　び

本稿では，ALMA計画において，高度5000 mというサポート常駐者がいない，過酷な環境下で計算機システムを安定運用するための以下の施策について示した。

・減圧チャンバ内での0.5気圧下における運転試験
・HDDレスシステムによる起動の仕組み
・リモートメンテナンスの仕組み

リモートメンテナンスシステムの仕組みにより，強制電源OFF，ONを含めたリモートでの計算機の停止，起動が可能である。また，コンソール画面からの作業が可能である。これにより，Linuxサーバが設置されている場所に極力出動することなく，メンテナンス作業が可能となる。また，問題発生時におけるリカバリ作業開始までの時間を短縮できる。

最後に，本システムの開発をご指導いただいた国立天文台の先生方に厚く感謝申し上げる。

参 考 文 献

(1)　自然科学研究機構 国立天文台：ALMA（Atacama Large Millimeter/submillimeter Array）. http://www.nro.nao.ac.jp/alma/J/
(2) 赤羽賢司ほか：宇宙電波天文学．初版，共立出版，1988，p.291-361.
(3)　有賀敬治：エンタプライズHDDの新たな潮流. *FUJITSU*, Vol.58, No.1, p.2-9（2007）.