宇宙の起源に迫る大型電波望遠鏡「アルマプロジェクト」向けスーパーコンピュータ：ACA相関器システム

Development of Supercomputer System Dedicated for ALMA Project: ACA Correlator System

● 阿部勝己　● 堤　純平　● 檜山貴博

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，チリにおいて大型電波望遠鏡アルマ（正式名称：アタカマ大型ミリ波サブミリ波干渉計）プロジェクトを推進している。富士通グループは大型電波望遠鏡で収集したデータを処理する専用スーパーコンピュータ「ACA相関器システム」を開発した。本システムは，PCサーバ「FUJITSU Server PRIMERGY」35台から成る相関器制御システムと，富士通アドバンストエンジニアリングが開発したACA相関器で構成され，望遠鏡のアンテナが受信する毎秒5120億個の電波信号データを毎秒120兆回（120 TOPS）の計算速度でリアルタイムに処理する性能と，標高5000 m，0.5気圧という過酷な環境での安定動作が求められる。本システムでは，計算性能を実現するためにコストパフォーマンスに優れたFPGA技術を採用し，安定動作を実現するためにディスクレスシステムの採用，低圧環境下での長時間稼働テストの実施やリモートメンテナンスシステムの採用などを行った。2011年から初期科学運用を開始し，試験観測を経て，2013年の開所式を無事に迎えるに至っている。

本稿ではACA相関器システムの全体像と，計算性能の達成および安定動作実現のために行った様々な施策について述べる。

Abstract

The National Astronomical Observatory of Japan (NAOJ) in National Institutes of Natural Sciences (NINS) is promoting the Atacama Large Millimeter/submillimeter Array (ALMA) project on a plateau near the Atacama Desert in Chile jointly with U.S. and European observatories. 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 comprised of 35 PC servers (FUJITSU Server PRIMERGY) and the ACA correlator developed by Fujitsu Advanced Engineering (FAE). The system meets stringent requirements including the need to process in real time huge amounts of radio signal data (512 billion pieces/second) from telescope antenna at a calculation speed of 120 tera operations per second (TOPS) and operating stably in the toughest environment (altitude of 5000 m, atmospheric pressure of 0.5 atm). This system has adopted cost-efficient field-programmable gate array (FPGA) technology to realize the computing performance and a disk-less system for stable operations. The system has been tested for long-time operation in a low pressure environment and it can be maintained remotely. Test operations started in 2011 and, after a pilot observation period, the system proved to be suitable for use in ALMA, which officially began in 2013. This paper describes the entire concept of the ACA correlator system and approaches to realizing its high computing performance and stable operation.

ま え が き

自然科学研究機構国立天文台様（以下，国立天文台）は，米欧の天文台と協力し，南米チリのアタカマ砂漠（高度5000 m）の高原に大型電波望遠鏡を建設した。この計画は，アルマ（アタカマ大型ミリ波サブミリ波干渉計：Atacama Large Millimeter/submillimeter Array）計画[(1)]と呼ばれ，2002年から米欧によって開始され，2004年からは日本も参加した3極体制で進められている。2011年に初期科学運用を開始し，試験観測を経て2013年より本運用を開始している状況である。世界中の天文学者が，アルマを用いた観測に応募しており，既に目覚しい天文学の成果を上げ，学会誌などで報告されている。[(2)]

アルマ望遠鏡の概観を**図-1**に示す。

アルマ望遠鏡は天体からの微弱な電波を受信する66個のアンテナと関連する設備から構成されており，国立天文台が担当するのは16基のアンテナと七つの周波数バンドのうちの四つの周波数バンド用の受信機，受信データを超高速に処理するACA相関器（以下，相関器）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。このうち，相関器とその制御システムをACA（Atacama Compact Array）相関器システムと呼び，その処理速度120 TOPS（tera operations per second）は，スーパーコンピュータに相当するものである。

富士通は，富士通アドバンストエンジニアリング（FAE）と共同で2004年からACA相関器システムの開発に着手した。[(3)] 相関器はアンテナからの大量データをリアルタイムに処理するための非常に高いスループットが求められた。**図-2**は，高度5000 mに設置されたACA相関器システムの外観とシステム構成である。ACA相関器システムは，24時間連続運転され，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。

図-1　アルマ望遠鏡の概観

ACA相関器システムの開発に当たり，以下の三つの大きな課題があった。

（1）大量データをリアルタイム処理する性能の達成

アンテナから相関器へのデータ伝送レートは，1.5 Tbpsである。これらを遅滞なく受信しリアルタイムに処理する必要があった。

（2）極限環境下での安定稼働

高度5000 m，0.5気圧の極限環境であってもシステムを安定に動作させる必要があった。

（3）リモートメンテナンスの実現

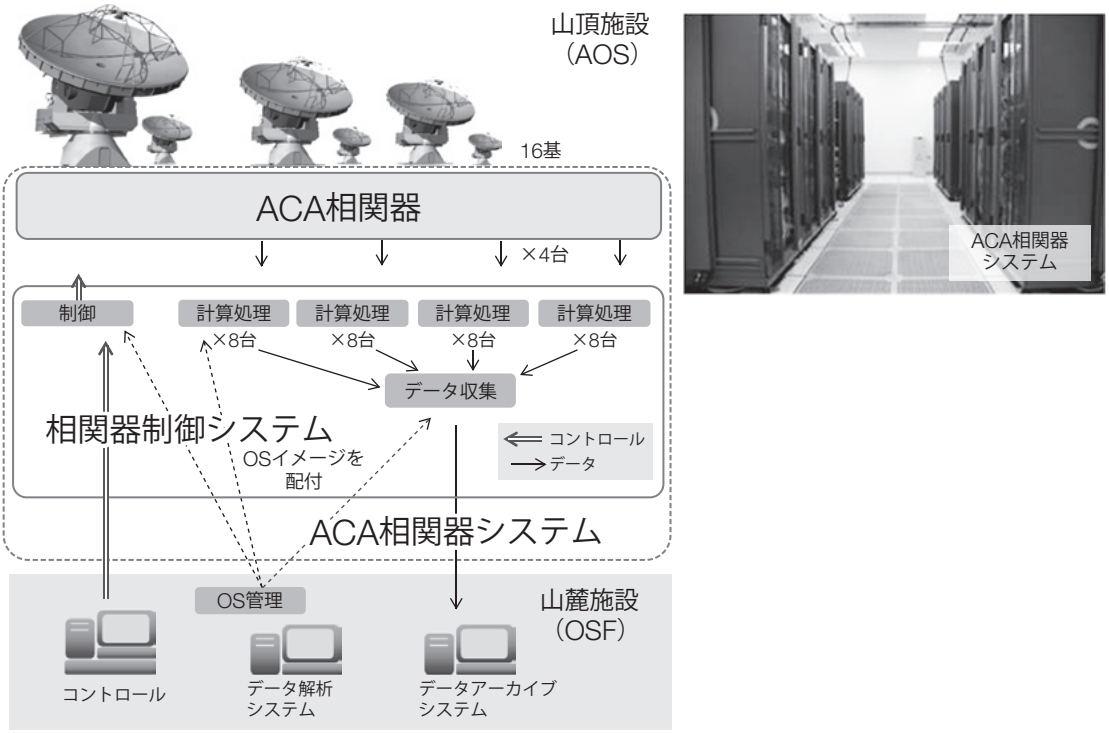
地球の裏側かつ高度5000 mという，人が常駐できない環境に置かれているシステムをリモートで保守し，状態をチェックする必要があった。

本稿では，これらの課題に対し，ACA相関器システムの開発において立案，実行した施策について述べる。

大量データのリアルタイム処理

ACA相関器システムは，米欧の担当する上流のアルマ全体制御システムおよび下流のアーカイブ装置とのインターフェースを持つ。相関器制御システムは，全体制御システムからの指示内容に基づいて観測モードを決定する。観測モードに基づき，相関器はアンテナから送られてきた電波データを相関処理し，処理結果である相関データを相関器制御システムへ送る。相関器制御システム上で，相関データは時間方向や周波数方向の積分や補正処理が行われた後に，データアーカイブシステムへ送出される。

アンテナから相関器へのデータの伝送レートは1.5 Tbpsである。これは一般家庭のインターネット回線2万本分に相当し，これら大量かつ連続出力されるデータを受信し続けることが必要である。そのため1秒間に120兆回の計算を実行する性能が求められた。また，相関器制御システムを構成するPCサーバ FUJITSU Server PRIMERGY（以下，PRIMERGY）には相関器から出力される相関デー



大型電波望遠鏡（左上）の画像提供：国立天文台

図-2　ACA相関器システム外観(右上)とシステム構成

タを受信した後，遅滞なく米欧の担当するアーカイブ装置へ送出するスループットが求められた。

上述の要件を実現するためには，当時の一般的なコンピュータを用いた場合，大規模な計算機が必要となったが，電力，設置場所，コストの面から採用困難であった。また，専用のプロセッサを開発することは巨額なコストが掛かるため実現困難であった。その解決策として，現在一般に広く用いられコストパフォーマンスにも優れたFPGA（Field Programmable Gate Array）を採用し，4096個のFPGAで並列演算方式を開発することで，大量データのリアルタイム処理を実現可能とした。更に一般的なプロセッサは一度作ってしまうと回路の変更ができないのに対し，FPGAは，実現したい機能に応じて回路を柔軟に変更することが可能であるため，運用開始ぎりぎりまで性能を引き出すためのチューニングを行うことが可能となり，研究者であるお客様の飽くなき探究心を満足することに貢献した。

一方，相関器から出力される相関データを相関器制御システム（PRIMERGY）で受信するに当たり，PRIMERGYに搭載可能な通常の通信カードでは相関器から連続して送られてくる大量データを漏れなく受信するための能力を満足しないため，PRIMERGYに搭載する専用受信カードを開発した。更に，受信した大量の相関データをPRIMERGY上で観測モードに応じて適切かつ高速に解析処理するために，計算プログラムに対し，徹底したチューニング（メモリアクセスの高速化，アセンブラレベルでの処理順序を意識したソースコードの記述方法の最適化，コンパイルオプションの組合せの最適化など）を施した。

極限環境の基盤システム

高度5000 m，0.5気圧の環境下では，当該システムの構成要素であるコンデンサー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。また，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を利用しているが，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ため，HDDの障害発生のリスクが高まる。今回，相関器制御システムで採用するPRIMERGYの稼働条件は最大高度3000 mであったため，高度5000 mの高原で安定稼働でき

るかを事前検証した。

● 熱対策～相関器とPRIMERGY～

相関器やPRIMERGYが発する熱を冷却するための工夫が必要であったが，幸い相関器の基板，筐体は専用装置であり，熱対策のための独自設計が可能であった。熱の発生箇所が1か所に集中しないように部品の配置を考え，効率的に熱を放射するように設計した。

PRIMERGYは汎用品であり，計算機を仕様外の物理的環境下で使用するとハードウェア障害が発生するリスクが生じる。このため，計算機を選定する前に，擬似環境で耐障害性を評価し，評価結果に基づいて対策を講じることで障害発生のリスク低減を図った。富士通の工場で行ったPRIMERGYの耐障害性の事前検証では，減圧チャンバ施設で高度5000 m相当の気圧の中での高負荷の厳しい検証とすべく，発熱量が最大となるCPU使用率常時100％で行った。PRIMERGYにはセンサーが搭載されており，結果はBIOSに記録される。この情報からハードウェアに異常が発生しないことを確認した。

また，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ため，実際の導入計算機に対して検証を行うことが重要である。本システムで最終的に採用したPRIMERGY RX300 S3は，事前検証を実施したRX300 S2の後継機種であり，RX300 S2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ため，無事に現地で稼働することが可能となった。

● 高度5000 mのHDDレス運用

アルマ計画では低圧環境下でのディスク障害発生のリスク回避のため，高度5000 mで使用する計算機システムにHDDを使用することを禁止している。HDDを使用しないでブートする仕組みとしてはCDからのブート，ネットワーク経由でのブートなどがある。

相関器制御システムでは，作業ミスを減らしメンテナンス性を高めるために一元管理が必要と考え，ネットワーク経由でブートする方式であるPXE（Preboot eXecution Environment）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PRIMERGYへOSを供給するOSサーバ（HDD搭載）が高度3000 mの拠点（OSF：山麓施設）に設置されている。1台のOSサーバが35台のPRIMERGYにOSを供給することで，メンテナンス性を高めることを可能とした。

CDブート方式も検討したが，CDブートの場合ブートするPRIMERGY台数分のCD媒体を作成する必要があり，システムを更新する際に新たに同じ枚数のCDを作成し直し，PRIMERGYが稼働する高度5000 mに運びCDを載せ換える作業が必要となる。高度5000 mの35台のPRIMERGYに対し，これらの作業実施は作業効率が低い。このため，CDブート方式を採用しなかった。

リモートメンテナンスによる安定稼働

サーバに不具合が発生した場合には，コンソール画面に流れるメッセージを眺める，などの手段を用いてサーバで何が起きているかを知る必要がある。しかし，コンソール画面を見るために現地チリに行く場合，東京から米国とサンチャゴを経由してOSFまで35時間，更に山頂施設（以下，AOS）へは約1時間を要する。また，高度5000 mでの作業は酸素濃度が低いことや夜間の滞在は禁止されていることから作業時間は制限される。

このような作業環境の制約から，チリのACA相関器システムに対して日本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り，これを実現した。現在，AOSでサーバが起動していないなどの状況が発生し，再起動を行った際の起動時メッセージの異常の有無を確認する手段としてリモートでコンソール画面が用いられている。リモートメンテナンスの概念を**図-3**に示す。日本から，高度5000 mのAOSに設置されたPRIMERGYにアクセス可能である。

● リモートメンテナンス

リモートメンテナンスは，監視，診断，復旧作業に分類できる。上記の作業は，筐体の目視，コンソールでの作業，システムへのログインなどを通して実施可能である。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。計算機の稼働場所と遠隔地の両方から，同時にメンテナンス作業を行う状況も想

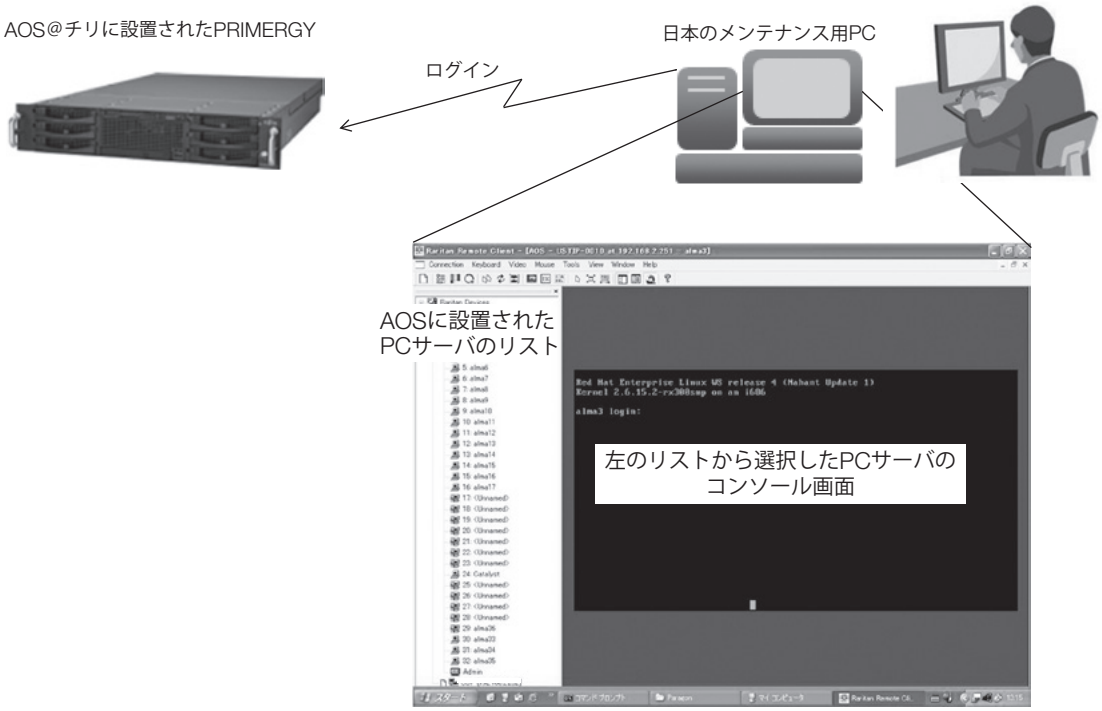


図-3　リモートメンテナンスの概念

定し，システムに矛盾を起こさせないための排他制御も必要であった。

更に，リモートメンテナンス作業では，組織内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であり，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った。

上記の要件を満足する製品を調査した結果，Raritan社のParagon装置を候補とし，次節に示す実証実験を行った。

●　実証実験

実証実験は，2段階に分けて行った。第1段階は擬似環境を用いてリモートメンテナンスの方式の実現性を検証した。第2段階は実際の提案機種を用いて模擬ネットワーク環境下での動作検証を実施した。

（1）擬似環境での実現可能性の検証

2005年，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，機能，セキュリティ，性能である。

リモートで操作対象にした計算機は，PRIMEPOWER200とSunBlade1000である。クライアントPCとして，FMV-BIBLOを使用した。経路は，ハワイ（ヒロ）－［DSL/VPN］－カリフォルニア（サニーベール）－［富士通WAN］－日本（幕張）である。

実験の結果，以下を確認することができた。

・リモートメンテナンスでの画面イメージを図-3の右下に示す。画面の左側には，接続可能な計算機の一覧が表示される。画面の右側には，接続した計算機へのログイン画面が表示される。このように操作性については，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った。

・リモートで作業を行う場合，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－幕張間で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反応遅延が大きいことが分かった。一方，国立天文台のある三鷹－チリ大学間の遅延時間は400 ms程度であったので，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

得られると推測した。

実証実験の第1段階において，リモートメンテナンスの実現可能性を確認でき，Paragon装置を採用することを決定した。

(2) 提案機種での動作検証

導入予定機器と同一の機器を使用し機能検証を行った。用いた計算機はPRIMERGY RX300 S3である。機能検証では，ネットワークの遅延や帯域を模擬的に設定可能なソフトウェアDummyNetを使用した。このDummyNetを用いてネットワーク遅延を発生させ，帯域幅を狭めることで性能への影響を評価した。その結果，リモート操作用PCから遠隔にあるPRIMERGYに対し，BIOS操作やメタキー入力が可能であることを確認することができた。このParagon装置は通信内容を暗号化する機能も有している。また，リモートメンテナンスシステムにより，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。これによりPRIMERGYが設置されている場所に極力出動することがないメンテナンス作業を可能とし，問題発生時における短時間でのリカバリーを実現できた。

実証実験の第2段階において操作性，機能，セキュリティ，性能の面の検証結果から導入予定機器は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た。

むすび

本稿では，アルマ計画において，高度5000 m，0.5気圧という過酷な環境下で，1.5 Tbpsという大量データをリアルタイムに処理し，かつ安定的に稼働するACA相関器システムを安定運用するための以下の技術的背景について示した。

・FPGA採用による高い処理性能の実現
・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

2011年からの初期科学運用を経て，2012年からの本格運用を継続している現在，ACA相関器システムは大きな問題もなく稼働しており，本稿で述べた施策が有効に働いているものと考えている。

アルマ計画は，今後約30年間にわたる運用年数を想定している。その間，科学技術の進歩とともに，アルマ計画も観測技術が向上していくものと思われる。アルマ計画によってもたらされる発見により，人類が宇宙の起源の解明へ一層近づきお客様の夢を形にすることの一助になれたとしたら幸いである。

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ます。

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://alma.mtk.nao.ac.jp/j/
(2) 天文月報 日本天文学会，Vol.106，No.10，(2013).
(3) 阿部勝己ほか：ALMAを支える相関器制御システム. *FUJITSU*，Vol.59，No.5，p.513-519（2008）.

著者紹介

阿部勝己（あべ　かつみ）
テクニカルコンピューティング・ソリューション事業本部科学システムソリューション統括部 所属
現在，国立天文台様観測制御システムの開発に従事。

堤　純平（つつみ　じゅんぺい）
テクニカルコンピューティング・ソリューション事業本部科学システムソリューション統括部 所属
現在，地球観測衛星の地上システムの開発に従事。

檜山貴博（ひやま　たかひろ）
テクニカルコンピューティング・ソリューション事業本部科学システムソリューション統括部 所属
現在，国立天文台様観測制御システムの開発に従事。