パブリッククラウドを利用したALMA観測データの品質保証実証実験

小杉 城治*1，森田 英輔*1，中里 剛*1，林 洋平*1，ミエル ルノー*1，合田 憲人*2，吉田 浩*2

Demonstration experiment of data quality assurance infrastructure for ALMA astronomical data using public cloud

KOSUGI George*1, MORITA Eisuke*1, NAKAZATO Takeshi*1, HAYASHI Yohei*1, RENAUD Miel*1, AIDA Kento*2, YOSHIDA Hiroshi*2

ABSTRACT

We experimentally built a data archive and data analysis environment for the ALMA telescope on the public cloud. In order to optimize the usability of the data and operation cost, the older data, which was opened to the public for more than 2 years, and relatively new data are stored and managed separately. By doing so, the operation cost of the archive system is not very different from that of on-premise case. In the cloud reduction/pipeline environment, it was found that the generation of the instance greatly affects the processing capacity and the cost, and that the prediction of the computer resources required for the processing is important for the cost optimization.

Keywords: ALMA, public cloud, Amazon Web Services (AWS), Data archive, Data reduction, Pipeline, data quality assurance.

概　要

パブリッククラウド上に ALMA 望遠鏡のデータアーカイブと解析環境を実験的に構築した。データの使い勝手と運用コストを最適化するために、公開から２年以上経過してアクセス頻度が低くなった古いデータと、それより新しくアクセス頻度が比較的高いデータを別々に保管することで、既存のオンプレミスアーカイブに近いコストでクラウド上にアーカイブが構築できる可能性が見えてきた。クラウド解析環境では、インスタンスの世代が処理能力やコストに大きく影響すること、また、コストの最適化には処理に必要な計算機リソースの予測が重要であることがわかった。

* 2020 年 12 月 4 日受付（Received December 4, 2020）
*1 国立天文台（National Astronomical Observatory of Japan: NAOJ）
*2 国立情報学研究所（National Institute for Informatics: NII）

1 はじめに

天文観測データはある時点における宇宙の歴史的な記録である。また、実験物理学と異なり、測定条件を主体的に制御することはできない。その意味で天文観測データは取り直しが利かない。そのため、国立天文台のデータポリシー[1)] 第２項には、「国立天文台は観測データを利用可能なデジタル形式で永続的に保管する」と記述されているのだが、それを実現するのは容易ではない。望遠鏡や観測装置が運用を続ける限り保管データは今後も増加し続ける。国立天文台では、これまで台内に構築された計算機環境で観測データの保管をおこなってきた。計算機環境は経年劣化を考慮して何年か毎に更新する必要がある。計算機や周辺機器の性能は向上しているものの、計算機更新時のデータ移行は更新前の計算機の I/O 速度に律速されるため、運用中に増加し続けた観測データも含めて全てのデータを決められた時間内に移行することは困難になってきている。

観測データを永続的に保管して公開する目的は、それらを現在、及び、後世のより広い範囲のユーザーの使用に供することにある。なぜなら、全ての観測データを現時点の限られた知識だけで完全に理解することは不可能だからである。望遠鏡や観測装置を熟知していない広い範囲のユーザーがデータを「正しく」利用して科学的成果を出すためには、データが適切に較正処理されている必要があり、また、品質が担保されていることが重要である。国立天文台のデータポリシー第３項に「国立天文台は、観測データを利用しやすい形式で公開する」とあり、その具体的内容として、「データは特定のソフトウェアを用いなくても解析できる水準まで較正処理を進め、できる限りそのまま物理量として扱えるようにした後に公開する」と記述されている。

望遠鏡や観測装置の運用を通じて、その特性の理解が進むにつれ、要求される較正処理はより複雑で高度なものとなる。データの滞留を防ぐためには、観測時間と同程度以下の時間で較正処理をおこなう必要があり、データ量やデータ処理内容に対応できる計算機資源を確保しなければならない。観測効率の向上や較正処理の高度化に対応するためには、計算機資源の継続的な増強・更新が求められる。地上からの天文観測は天候や季節（太陽や月と天体との位置関係など）の影響を受けるため、較正処理するデータ量や処理内容は必ずしも一定ではない。従来は計算機処理の繁忙期に対応できる十分な計算機資源を確保してきたため、逆に処理の閑散期には計算機資源には比較的余裕がある状態となっていた。

安全に効率的に天文観測データを「利用しやすい形式」で「恒久的」に保管と公開を続けていくためには、これらの課題に目を向け、各時代のテクノロジーに即した対応策を模索していく必要がある。パブリッククラウドを利用した今回の実証実験は、上記課題を解決する際のヒントとなる。

2 ALMA 観測データとその品質保証

2.1 ALMA 望遠鏡による観測

ALMA 望遠鏡はミリ波・サブミリ波干渉計として世界最高の感度と解像度を備えている。そのため、世界中の研究者が観測時間を求めて利用申請をおこなう。通常は年に１回、観測プロポーザルを提出することで利用申請をおこなう。観測プロポーザルには、観測の目的や期待される成果、望遠鏡をどのようにどの程度の時間使用するか、観測天体情報、観測する周波数帯域や周波数分解能、必要とする空間分解能や感度など、観測に必要なパラメータを全て記述する。

集まった観測プロポーザルは天文学の分野毎に分類され、それぞれの分野の複数の研究者による審査を通じて採点される。また、プロポーザルに記述された観測方法が実現可能かどうか、必要とされる感度や分解能と観測時間や望遠鏡配列などに矛盾がないか等、採点とは別に技術的な審査も実施される。審査を経て実際に ALMA 望遠鏡で観測されるプロポーザルは、観測時間換算でプロポーザル全体の 20～25%程度になっている [2)]。

2.2　観測データの品質保証

国際 ALMA 観測所は、4～5 倍という狭き門を通った優れたプロポーザルが科学成果に繋がるように、最終的なデータ品質（感度や分解能など）がプロポーザルの要求要件を満たしているかどうかデータ解析を通じて確認し、観測スケジュールにフィードバックをかけている。要求された品質が満たされるまで観測が繰り返され、満たされたことが確認され次第、観測提案者に観測データ、及び、解析済みデータが配布される。この一連の活動を観測データの品質保証と呼んでいる。

2.3　観測データの保管と公開

ALMA 望遠鏡のデータレートは、最大 60M Bytes/s、平均 6M Bytes/s、年間データ量の目安は 200T Bytes 程度である。データレートは、観測天体の明るさ、周波数分解能、周波数帯域幅、周波数帯域数、空間解像度、観測に使われる望遠鏡数、積分時間など、観測プロポーザルの要求に沿って変化しうる。年間 200T Bytes という上限は、2010 年以前に ALMA 望遠鏡の運用計画を策定する中で運用コストの面から規定された。計算機能力や観測システムの進化、観測装置（ALMA 望遠鏡の場合は受信機システムや相関器など）の改良などにより、数年後には最大データレートが引き上げられ、年間データ量も今の数倍にまで達する予定である。アーカイブに保管された観測データ・解析済みデータ量の推移を図 1 に示す。2017 年以降には年間データ量が計画通り 200T Bytes 程度となっており、アーカイブ保管されているデータ総量は現在 1PB に迫っている。

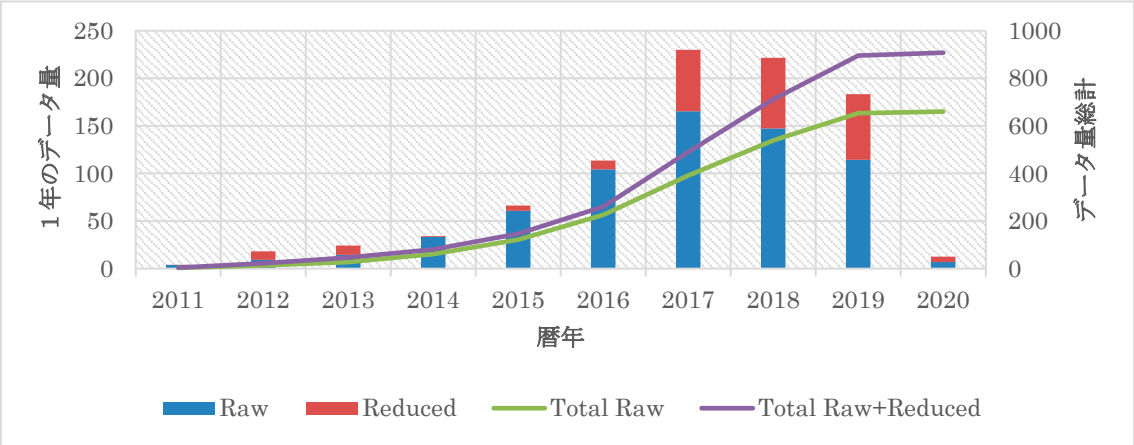

図 1　ALMA アーカイブデータ量の推移

ALMA 望遠鏡の観測データは全てアーカイブに保管される。解析済みデータもまた同じアーカイブに保管される。観測データの品質保証が完了して観測データや解析済みデータが観測提案者に配布されてから 1 年間（執筆時点では COVID-19 の影響を踏まえて 1 年 3 ヶ月間に延長中）は、観測データへのア

クセス権は観測提案者だけが有する（データ占有期間）。データ占有期間が終わるとデータは全世界に公開され、誰もが自由に利用できるようになる。ALMA 望遠鏡の観測データは、データ保全、及び、ローカルコミュニティへのサービスのために、日米欧、及び、チリでそれぞれ保管されている。また、品質保証活動も日米欧智で分担して進められているが、本論文は日本の活動部分にフォーカスしたものである。

3 運用環境とパブリッククラウド実証実験環境

国立天文台アルマプロジェクトで実際に運用しているデータアーカイブ、及び、データ解析機能を模擬する環境をパブリッククラウド上に構築し、性能やコストの比較をおこなった[3)4)5)]。利用可能なパブリッククラウドには幾つか候補があるが、今回の実証実験では、国立情報学研究所 (NII) が構築・運用している SINET[6)] (Science Information NETwork)から直接乗り入れが可能なクラウド事業の一つである Amazon Web Services (AWS)を利用した。

3.1 データのアクセス頻度と対応するクラウドサービス

クラウドのストレージサービスには複数種類があり、データのアクセス頻度などストレージの用途に適したものを選ぶことが重要である。コールドストレージは、通常のストレージサービスよりも 1 桁低いコストでアクセス頻度の低いデータ（コールドデータ）を格納する。頻繁にアクセスされないことが前提なので、データへのアクセスコストは高く設定されている。

AWS の主なストレージサービスには AWS S3 という標準オブジェクトストレージサービス、AWS S3 IA (Infrequent Access)という標準オブジェクトサービスのオプション、並びに、AWS Glacier というコールドストレージサービスがある。AWS Glacier は容量当りの保管コストが低い代わりに、データアクセス時に長時間の復元処理が必要であったり、データの最低保持期間が指定されていたりと、幾つかの制約が課せられている。データをクラウドから外に取り出す場合、AWS S3 ではデータ保管料金、データアクセスリクエスト料金、及び、対 SINET データ転送料金が発生する。AWS S3 IA ではデータ保管料金は低いものの、更にデータ取り出し料金が追加される。AWS Glacier ではデータ保管料金はより低いものの、更にデータリストアリクエスト料金とデータリストア用スペース料金が加算される。

3.2 データアーカイブ機能

国立天文台で 2018 年に更新され稼働している現システム（以降、オンプレミスアーカイブと呼ぶ）は、2023 年まで運用される予定である。観測データや解析済みデータを保管するストレージの総容量は 2P Bytes（ペタバイトはテラバイトの 1000 倍）、また、管理データベース領域として 10T Bytes の高速ディスクが割り当てられている。これらのストレージと 3 台のデータ管理サーバー、及び、1 台のデータベースサーバーは 32G bps の高速ファイバーチャネルスイッチで接続されている。

これを模擬する環境を AWS 上に作成し、アーカイブやデータベースソフトウェアのインストールの後、既に公開された ALMA の観測データや解析済みデータを全て登録してクラウドアーカイブを構築し、オンプレミスアーカイブとデータダウンロードの性能比較をおこなった（図２）。なお、本測定では、コールドストレージサービスとして、S3 IA を使用した。

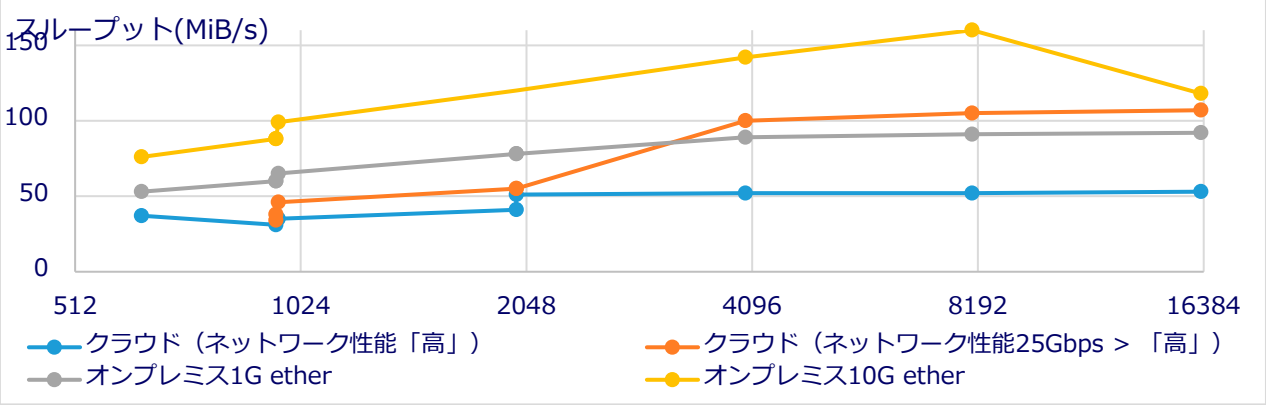

図２　オンプレミス、及び、クラウドアーカイブによるデータのダウンロード性能比較。横軸の単位は MiB。

全体的な傾向として、データサイズが大きくなるほど、スループットが上がっている。クラウド上の ALMA アーカイブは、オンプレミスと比較すると、同程度のネットワーク帯域のものでも性能が低下している。これは、ストレージの性能差、オーバーヘッド、及び、データベースの性能差に起因するものと考えられる。ただし、アーカイブから取り出したデータをインターネット経由で転送する場合には、インターネットの帯域でリミットされるため、性能差はほとんど目立たなくなる。

ALMA アーカイブには、現時点でおよそ 900T Bytes のデータが保管されている（図１）。国立天文台のオンプレミスアーカイブから最近１年でダウンロードされたデータ量は約 550T Bytes である。この実測値を使って、クラウドアーカイブの運用コストをシミュレートした（図３）。AWS のストレージサービスでは、データの保管費に加えて、データの読み出しやインターネットを経由したダウンロードに対しても課金される。550T Bytes ものダウンロード量であっても、保管費が運用費の大部分を占めるため、AWS Glacier を利用した方がかなり安価に運用できる。AWS Glacier の運用費であれば、オンプレミスアーカイブの運用費と比べて大きな遜色はない。但し、AWS Glacier のようなコールドストレージはデータにアクセスしてダウンロードする前に、直接アクセスできるストレージ領域にデータをリストアする必要があり、これに 200 分程度の時間を要する。クラウドアーカイブユーザーにとって、ダウンロード開始までに常に長い待ち時間が発生するのは、我慢できることではないであろう。

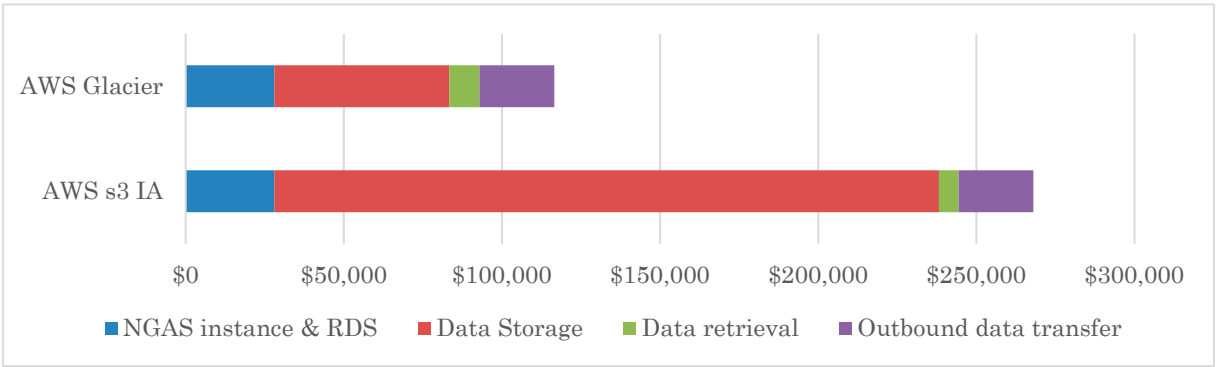

図３　クラウドアーカイブの年間運用費試算。AWS s3 IA で約$270K、AWS Glacier では約$120K。NGAS: Next Generation Archive System は ALMA アーカイブのソフトウェア。RDS: Relational Database Service。

3.3 データ解析機能

観測提案者に品質保証された解析済みデータを観測後一定期間内に提供するために、日米欧の各ALMA Regional Center (ARC)とチリの Joint ALMA Observatory (JAO)ではそれぞれ解析・パイプライン環境を構築し、観測データを逐次解析処理している [7)]。標準観測モードで取得された観測データは、基本的に自動解析パイプラインを通り、最終プロダクトのチェックのみARC或いはJAOの天文学者が担う。解析パイプラインを通らない、或いは、処理の途中で失敗する観測データは、データ解析者によるマニュアル解析に回される。観測提案者に解析済みデータを提供する目標期限は、自動解析パイプラインが使える場合には観測完了後 30 日以内となっている。解析処理にかかる時間は、観測提案者が決めた観測設定（望遠鏡配置、周波数分解能、周波数帯域幅、帯域数、観測空間範囲、積分時間など）で大いに左右される。観測時間に対する解析処理時間は一定ではなく、観測設定によっては観測時間と解析処理時間が同程度のものから、解析処理時間が数倍や十数倍以上かかる場合もある。処理の重い観測データが短時間の内に多数取得された場合でも解析済みデータの目標配布期限を満たせるよう、ARC や JAO では十分な計算機資源をデータ解析システムに割り当てている。

本論文執筆時点において国立天文台アルマプロジェクトで構築・運用しているデータ解析環境は、表１の通りである。なお、解析処理の高精度化や新しい観測モードの導入に伴い、要求される計算機資源が増大するため、現在も順次解析計算機環境を拡充しつつある。

表１　国立天文台のオンプレミス解析計算機諸元

Category	台数	CPU	Memory	Storage
解析パイプライン計算機	12	1 CPU with 4 - 6 cores	64 – 256 GB	高速共有ファイル 200 TB 程度
マニュアル解析用計算機	12	1 CPU with 4 - 6 cores	64 – 128 GB	ローカル高速ファイル 12 – 48 TB

クラウド解析環境も AWS 上に構築した。AWS 上の仮想マシンインスタンスで観測データの解析・パイプライン処理をおこない、解析処理時間などの性能と計算コストを評価した。また、観測データを AWS 上に構築したクラウドアーカイブ（3.2 参照）に格納し、そこから直接クラウド解析環境に読み出す仕組みも構築し、実証実験に供した。

3.3.1 オンプレミス解析環境とクラウド解析環境の比較

クラウド解析環境の性能評価や運用コストの算出のために、搭載メモリサイズやインスタンス世代を変えたクラウド解析環境を複数用意し、オンプレミス解析環境と全く同じデータセットを処理させた。準備したデータセットの概要を表２に、また、実験に使用したインスタンスの概要を表３に示す。尚、表３のコストは実験当時の概算であり、契約形態や契約時期等によって変わりうるものである。ALMA のデータセットは表２のものよりも更に大きなものがあり、データサイズが 100 GB 程度、処理時間は１週間程度かかる。今回の結果には含めていないが、クラウド解析環境上での処理自体は問題なく完了している。また、オンプレミスでの処理時間と比較しても、特に大きな違いは出なかった。

表２　ALMA 解析処理用データセットの概要

データセット	データサイズ	典型的な処理時間
小データ	500 – 700 MB	１時間程度
中データ	3 – 5 GB	５時間程度
大データ	10 – 30 GB	１日程度

表３　クラウド解析環境のインスタンス概要

インスタンス	CPU core	Memory	Generation	Cost (approx. US$/month)
m4.2xlarge	4 cores	32 GB	4-th generation	$385 -
m5.2xlarge	4 cores	32 GB	5-th generation	$370 -
r4.2xlarge	4 cores	61 GB	4-th generation	$480 -
r5.2xlarge	4 cores	64 GB	5-th generation	$455 -
x1e.2xlarge	4 cores	244 GB	-	$1,800 -

図４に実験結果を示す。左から順に小データ（データセット#1）、中データ（データセット#4）、大データ（データセット#9）を処理した時間、課金額を示している。処理時間については、どれもオンプレミス解析環境と大きな差異はないが、メモリを多く積んだクラウド解析環境ほど解析処理に時間がかかっているのがわかる。メモリを多く積むとインスタンスのコストが上がるため、相乗効果で処理コストがかさんでいる。この解釈については考察で詳しく述べる。

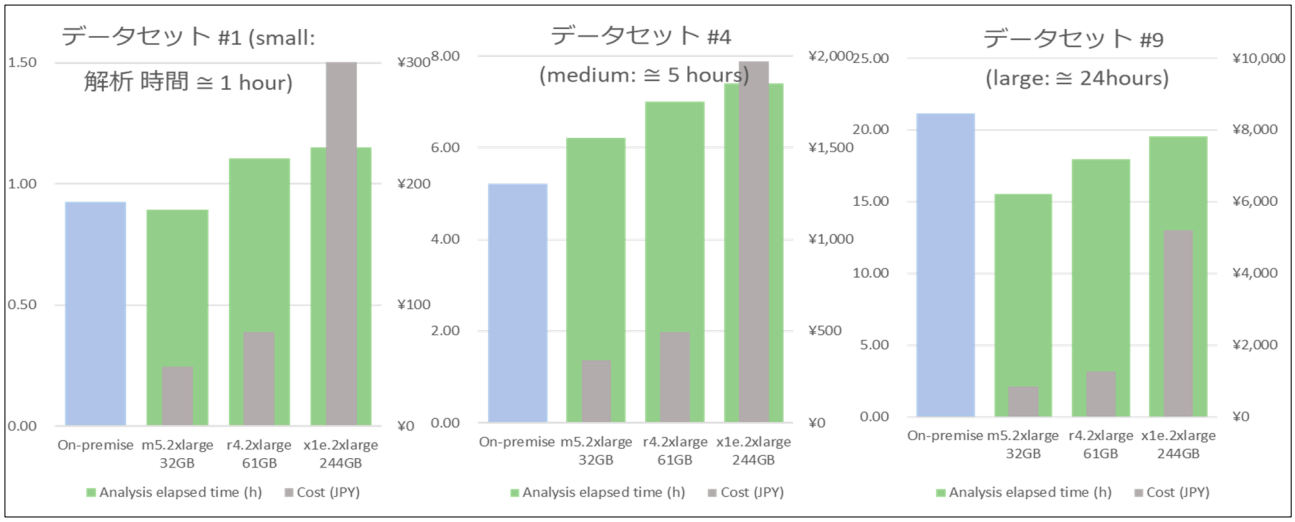

図４　オンプレミス環境とクラウド解析環境の処理時間の比較と課金額。太い縦棒は左側縦軸に対応した処理時間（単位 hour）を示し、細い縦棒は右側縦軸に対応する課金額を示す。1USD＝¥110 で換算。

4 考察

4.1 クラウドアーカイブの最適化

3.1 では全てのデータを AWS s3 IA、或いは、AWS Glacier に保管した場合の運用費を試算したが、運用費を抑えようとするとデータの復元処理のためにダウンロードまでの待ち時間が極端に長くなる。頻度を減らすためには、アクセスの少ないデータを選択的に AWS Glacier に保管し、ある程度以上アクセスが期待されるデータは AWS s3 IA やオンプレミスアーカイブに保管すれば良い。アーカイブデータは新しいほど、或いは、公開されて間がないものほどたくさんアクセスされることが期待できる。そこで、我々は観測提案者にデータが配布された後の経過時間によって、データのアクセス数がどのように推移するかを調べた（図５）。その結果、ダウンロードされる大部分のデータは、観測提案者にデータを配布した後２年以内の比較的新しいデータであることがわかった。

そこで、観測提案者へのデータ配布後２年以内のデータを、頻繁にアクセス可能なオンプレミス、或いは、AWS s3 IA に保管し、それ以外の古いデータを AWS Glacier に格納する。このようなクラウド同士、或いは、クラウドとオンプレミスのハイブリッドをおこなうことで、アクセスの多いデータは待ち時間を少なく、アクセスの少ないデータは待ち時間が長くても我慢してもらい、しかも、トータルの運用費をある程度抑制することが可能となる。またデータ配布後２年以内のデータ量は今後数年間あまり変化しないのに対し、２年以上のデータは定常的に増加する。定常的に増加するデータ部分を AWS Glacier に保存することは、将来の運用コストの増加を押さえる方向に働く。

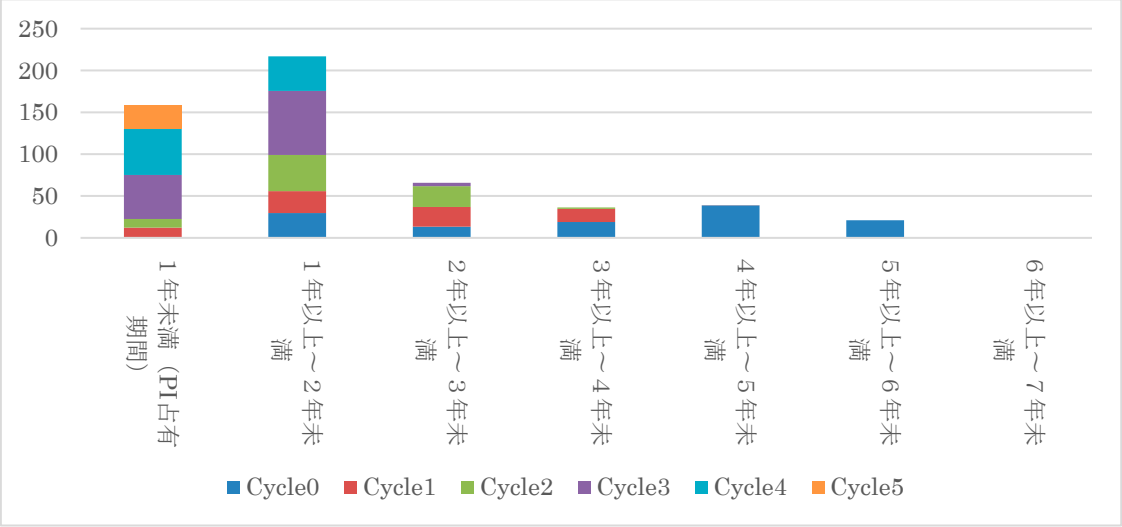

図５　観測提案者にデータ配布後の経過年数別ダウンロードサイズ。単位はテラバイト(TB)。

4.2 クラウド解析の最適化

3.3.1 で示したように、今回の結果はメモリを多く積んだインスタンスほど処理が遅くなることを示唆しているようにも見えるが、実は、インスタンスの世代による処理能力の違いが見えていることがわかった。図６は同じデータセットを同じサイズのメモリを積んだ世代が異なるインスタンスで処理した結果である。表３にそれぞれのインスタンスの世代情報を載せた。m5.2xlarge は第５世代インスタンス、m4.2xlarge は第４世代インスタンスである。測定時点で最新だった m5 のインスタンスは、旧世代の m4 インスタンスと比較して、仕様上も CPU や I/O 性能が向上している。しかも表３のように、新しいイン

スタンスの方が価格も若干下がっている。処理時間が短縮され、しかも価格が下がることにより、処理コストの差が大きく出る。すなわち、インスタンスの価格設定が新旧であまり差がなければ、新しいインスタンスを選択すべきである。オンプレミス解析環境の場合、一旦 CPU やマザーボードを購入すると、性能が陳腐化するまで使い続けることがよくあるが、クラウド解析環境の場合は、積極的に新しいインスタンスに乗り換えていくことが可能となる。それにより、処理速度が向上し、コストも抑えられるという利点がある。

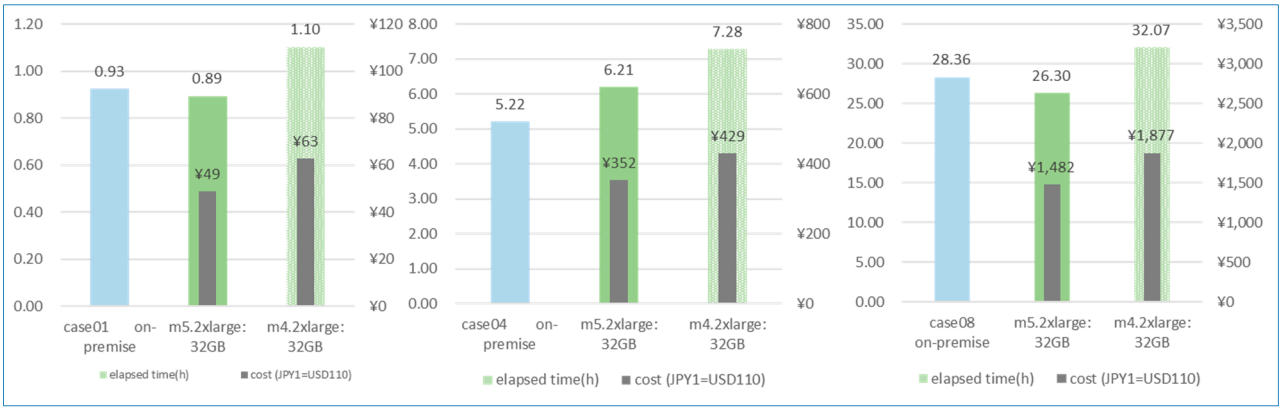

図６　インスタンス世代の違いによる処理時間とコストの比較

図４では同じデータをメモリ量の異なるインスタンスで処理しても、処理が正しく完了することが示されている。しかし、メモリ量の効果はほとんど見えていない、或いは、インスタンス世代の違いに隠される程度しか変化していない。一方、メモリを無駄に多く積んだインスタンスは高価になる。図７は大きなサイズのデータセットを処理した際の CPU の利用率、図８はメモリのプロファイルを示している。少なくとも処理の前半部分は、CPU コアを１つしか使わず、また、メモリも一部しか使われていないことがわかる。すなわち、解析パイプライン処理を前半、後半に分け、前半部分はコア数もメモリ数も少ないインスタンスで処理し、後半部分のみ必要数のコアとメモリを積んだインスタンスに載せ替えることで、クラウド解析環境の効率的な利用ができる。実際、我々のチームでは、インスタンスに対するブロックストレージの動的な取り付け／取り外しが可能というクラウドの特性を活用して、前半処理を終えたストレージを、後半処理のために CPU コアやメモリが多いインスタンスに付け替えて、続きの処理を進めることに成功している。

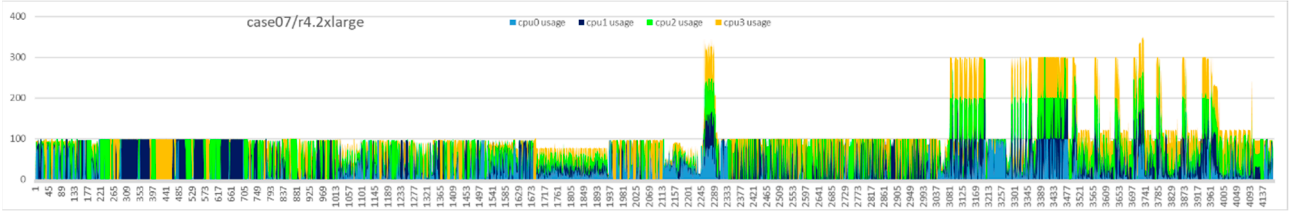

図７　大きめのデータを解析処理したときの CPU 利用率の推移

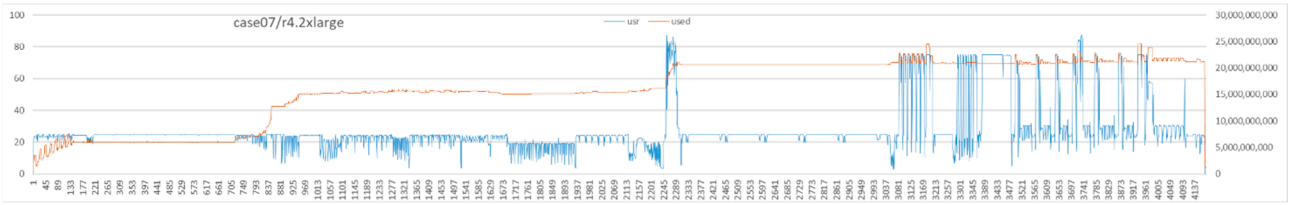

図８　大きめのデータを解析処理したときの使用メモリ量の推移

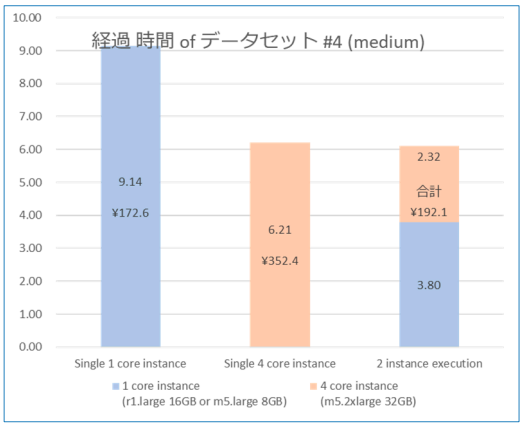

図９　インスタンスを組み合わせて処理したときの時間、コスト例

図９はその実例であり、中程度のサイズのデータセットを、最初から最後まで１コア 16G Bytes メモリのインスタンスで処理した場合（左）、４コア 32G Bytes メモリのインスタンスで処理した場合（中央）、前半を１コアで処理し、ストレージ付け替えをおこなった後、後半を４コアで処理した場合（右）の処理時間とインスタンスのトータルコストを示している。前半をコストの低い小さなインスタンスで処理し、後半を大きなインスタンスで処理することによって、処理時間は全てを大きなインスタンスで処理した場合と同等、コストは小さなインスタンスだけで処理した場合と近い値まで抑えられた。

5　今後の課題

各解析処理に必要なリソースが正確に予測できれば、必要最小限のリソースを持ったインスタンスを解析処理に使うことが可能となり、ストレージの付け替えと組み合わせることで、効率的なクラウド利用が可能となる。現在、解析処理に必要とされるメモリ量やディスク容量を予測するために、様々な観測パラメータと使用リソースの関係を調査中である。

参考文献

1) 国立天文台のデータポリシー, https://www.adc.nao.ac.jp/J/cc/public/center/ADC_Data_policy_20140522.pdf （2021 年 1 月 21 日閲覧）
2) ALMA Reports, https://almascience.eso.org/documents-and-tools/alma-reports （2021 年 1 月 21 日閲覧）
3) 吉田 浩, 合田 憲人, 上田 郁夫, 原 隆宣, 小杉 城治, 森田 英輔, 中村 光志, 「クラウドコールドストレージに対する大量データ格納の試行と評価」, 情報処理学会研究報告 2017-HPC-160 No.25
4) 吉田 浩, 合田 憲人, 上田 郁夫, 原 隆宣, 小杉 城治, 森田 英輔, 中村 光志, 「クラウドコールドストレージに対する大規模実験データ格納のケーススタディ」, 情報処理学会研究報告 Vol.2018-HPC-165 No.8
5) Hiroshi Yoshida,“Experiments in Storing Scientific Research Data in Cloud Cold Storage Services”, Storage Networking Industry Association (SNIA)主催 Storage Developers Conference 2018,

https://www.snia.org/sites/default/files/SDC/2018/presentations/Cloud_Storage/Yoshida_Hiroshi_Experiments_in_Storing_Scientific_Research_Data_in_Cloud_Cold_Storage_Services.pdf （2021 年 1 月 21 日閲覧）

6) SINET, https://www.sinet.ad.jp/ （2021 年 1 月 21 日閲覧）

7) George KOSUGI, “Extensive data handling for ALMA (Atacama Large Millimeter/Submillimeter Array)”, 宇宙科学情報解析論文誌, 第 1 号, 2012, 77