老朽化した大容量ストレージから Ceph への移行と統合
(Migration and consolidation from aging mass storage to Ceph)

◯池田恵美、Renaud Miel
(国立天文台アルマプロジェクト)
(◯Emi Ikeda, Renaud Miel
(ALMA Project, NAOJ))

概要(Abstract)

アルマプロジェクトではアルマ望遠鏡の観測で得られたデータの解析処理を行っている。観測データ、解析処理データは非常に大きく、大容量のストレージを必要とする。我々は、これまで使用していたディスクアレイ装置等の機器の老朽化、サポート終了に伴い、分散ストレージシステム Ceph への移行、共有ストレージの統合を行った。本発表ではストレージの移行の準備、ストレージサービス切り替えのプロセスについて報告する。

The ALMA project is analyzing and processing data obtained from ALMA telescope observations. The observation data and analysis processing data are very large and require large storage capacity. We have migrated to a distributed storage system, Ceph, and integrated shared storage due to the aging of the disk array devices and other equipment we had been using and the end of their support. In this presentation, we will report on the preparation for the storage migration and the process of switching storage services.

1. 背景

アルマ望遠鏡で得られる観測データは年間およそ 200TB あり、それらは日米欧チリに構築された解析・パイプライン環境(以下、解析環境)にレプリケーションされて解析処理が行われ、観測後一定期間内に品質保証された解析済みデータとして提供される。アルマプロジェクトは日本側の拠点として、この解析処理を行っている。観測データのセットを並行して解析処理し、その処理結果は一定期間保持した後は削除するため、アルマプロジェクトでは複数台の解析サーバと大容量のストレージを有している。年月を経て、この解析環境に初期に投入されものなどの一部が、サポートの End of Life を迎えたり、機器の老朽化で、そのまま使い続けるのが難しい状況となっていた。

2. アルマプロジェクトの解析環境

2.1. 移行前の解析環境

2022 年 4 月時点では、アルマプロジェクトの解析環境は、約 30 台の解析サーバ、開発サーバと Lustre、NFS、Ceph からなるファイルストレージ、NW 装置等で構成されていた。解析用データ、解析用ソフトウェア、共有ファイルは、用途やクライアントの環境にあったファイルストレージに格納され、解析

サーバ、開発サーバは複数のファイルストレージをマウントしていた(図 1 参照)。しかし、NFS サーバ、Lustre、一部の解析サーバは、OS やデバイスのサポートの End of Life を迎えていた。また、これらのサーバ機やこれらを構成するバックエンドストレージ、NW 装置は老朽化し、安定運用やセキュリティに不安を抱えた状況であった。

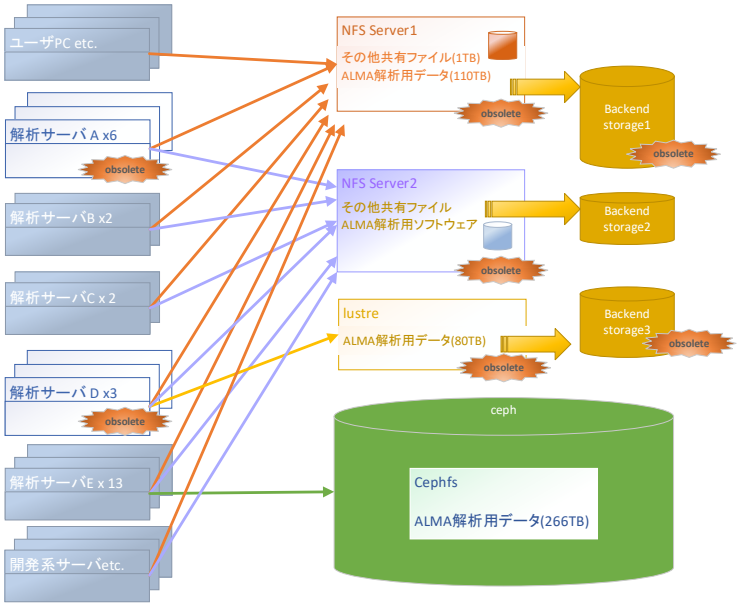

図 1: 移行前の共有ストレージの利用状況

2.2. 解析環境の更新の検討

上記の状況から、解析環境の更新に迫られていたが、解析処理は継続して行われ、データの量は今後ますます増大する見込みであることから、機器の更新ではなく、解析環境の構成を大きく見直して更新を検討することとした。

- **問題点や懸念事項**

  当時の解析環境の問題点や懸念事項は以下のとおりである。

  - ストレージの種類が複数あることで、メンテナンスの作業量が多くなっている。
  - Lustre は、マウントするクライアントも kernel のバージョンが同じである必要があるため、利用できるクライアントが限定される。kernel が変わるようなメンテナンスは、クライアントもすべて一斉に行わなければならず、メンテナンス性に乏しい。
  - NFS サーバは解析サーバ、開発サーバ全てに領域を export しているが、冗長化されていないため、単一障害点となっている。障害が起こった時の影響範囲が大きい。
  - 移行対象のデータのサイズは、少なく見積もっても 150TB 程度ある。大容量のストレージシステムは高額なため、新たに購入するのは予算的に厳しい。

- **格納データの分類**

  更新の対象となるストレージに格納されているデータは、表 1 のように分類できる。

表 1: データの種類、分類

データ			利用者	アクセス
カテゴリ	種類	サイズ		
ALMA データ解析関連	解析ソフト	170GB	解析、開発者	解析環境用 NW
	解析データ	150TB		
その他	解析関連以外の共有データ	400GB	ALMA プロジェクトメンバー	台内 NW
	プロジェクト向けポータルサイトのコンテンツ			
	共有スクリプト			

- **継続して利用可能なストレージ**

  図 1 にある Ceph ストレージは、ALMA の解析環境の共有ストレージの中で最も新しく 2019 年 9 月に導入したストレージ[1]で、解析環境にある複数の OS や異なるバージョンのクライアントが接続できる。2022 年 4 月の時点では総容量 800TB 実効容量 266TB あり、150TB 程度使用していたが、データの冗長形式の変更により実効容量を 480TB 程度にあげることが可能である。また、メンテナンスや障害時に縮退運転で継続してサービスを提供でき、容量の拡張の余地もあるため、当分の間使い続けることが可能である。

以上の検討より、Ceph ストレージのデータの冗長形式を変更して実効容量をあげることで、Lustre, NFS サーバの大部分のデータを、既存の Ceph ストレージに収容し統合することとした。これによりストレージの可用性を高くすることも見込める。

2.3. Ceph とは

我々の解析環境のストレージに導入している Ceph[2]は、オープンソースの分散ストレージソフトウェアである。複数台のコンピュータ(ノード)からなるストレージ領域を 1 つのストレージクラスターとしてまとめ、切り出した領域(pool)ごとにブロック単位(RBD)、ファイル単位(CephFS)、オブジェクト単位(RADOSGW)のインターフェースでアクセス可能なストレージを提供する。データは、冗長化して Ceph クラスター内に分散して格納される。またクライアントは Ceph から取得する CRUSH マップにより自身で動的に格納先を計算してデータノードから直接データを取得する。このように、Ceph はデータのやりとりにおいてノードの障害の影響をうけにくく、単一障害点を作りにくい仕組みをもつ。そして、ノードの冗長構成により、ローリングアップデートや縮退運転ができ、サービスを止めることなく運用できる。また、Ceph クラスターは汎用的な計算機と OS で構築でき、容量をエクサバイト規模まで拡張可能であるため、安価に大容量ストレージを実現できる。我々は、Ceph のこれらの耐障害性、高可用性、柔軟な拡張性という特長から、解析環境の大容量ストレージの一部に Ceph を導入した。2019 年 9 月の運用開始以降、サービスを継続しながら、メモリの増設、容量の拡張、ノードの追加、システムや OS のアップグレードを行ってきた。

3. 移行計画

我々は以下のとおり移行計画をたて、2022 年 1 月より、サービスを継続しながら、A),B)とも並行して移行作業をすすめた。

- **移行先:** 表 1 の ALMA データ解析関連は ceph に移行。その他は他のシステムに移行
- **移行の Deadline:** 三鷹地区全館停電(2022.11.22)でストレージの切り替え。移行完了
- **移行手順**。

  A) Ceph クラスター側の作業: 実効容量をあげ、接続可能なクライアント数を増やす

  1. Ceph クラスターノードの OS(Ubuntu18.04->Ubuntu20.04), Ceph(ver.15.xx->ver.16.xx)のアップグレード
  2. データ格納領域の利用ポリシー制定
  3. Ceph の pool を replication pool から Erasure coded pool へ移行(実効容量 266TB-

>480TB)

4. ceph NFS の構築(ceph public network にアクセスできないクライアントむけ)
5. ceph public network の拡張。クライアントの数の増量(移行後の作業)。

B) データの移行: ユーザ領域はユーザ自身で、それ以外は我々(シスアド)が移行する

表 2: 移行作業手順

作業	データのカテゴリ		作業者	
	解析関連	その他	ユーザ	シスアド
1.ユーザデータ領域の再編成。移行スクリプト作成	○			○
2.ユーザデータ領域のデータ移行	○		○	
3.ポータルサイトの外部立ち上げ。一般向けデータの外部移行		○	○	
4.一般向け共有フォルダの書き込み停止		○		○
5.解析関連のユーザデータ移行	○		○	
6.外部サービスのデータ移行		○		○
7.解析関連のユーザデータ以外のデータを繰り返し移行	○			○
8.NFS サーバで実行していたスクリプトの移植	○			○

4. まとめ

2022 年 1 月より準備を始め、図 1 にあった Lustre, NFS サーバの大容量の解析データを、大きな混乱もなくサービスを継続しながら、ceph に移行し、1 ストレージに集約、統合した。その結果、図 2 のようなマウントの構成となった。Ceph は冗長構成をもち、可用性が高いため、障害に強く、メンテナンスを適宜行える環境で、クライアントの多くの環境に継続して安定したサービスを提供できるようになった。今後は、NFS を利用している解析ソフト関連について内容を整理し、Ceph で提供できるようにし、より可用性を高めていく予定である。長期的には、ますます増大するデータの格納に備え、Ceph のノードを追加する計画もある。

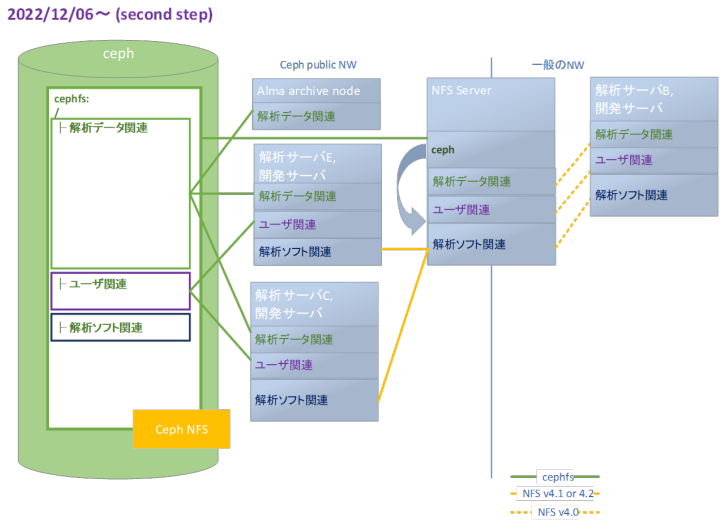

図 2: 移行後のマウント状況

5. 参考文献

[1] 池田恵美 他, アルマプロジェクトの Ceph ストレージの利用, 第 41 回 天文学に関する技術シンポジウム, http://tech.nao.ac.jp/tech-sympo/2021/proceedings/techsympo_2021.html

[2] Ceph, https://ceph.io/