アルマプロジェクトの Ceph ストレージの利用

◯池田恵美[1]、Renaud Jean Christophe Miel[1]、中村光志[2]、芦田川京子[1]、小杉城治[1]
(1 国立天文台アルマプロジェクト、2 国立天文台情報セキュリティ室)

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、解析済みデータは非常に大きく大容量のストレージが求められる。我々はオープンソースでエクサバイト規模にスケールアウト可能な分散オブジェクトストレージ Ceph を導入し、解析サーバのみならず他のサーバも利用可能な大容量ストレージとして利用している。本発表ではアルマプロジェクトの Ceph ストレージの構築、運用について報告する。

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、日米欧チリに構築された解析・パイプライン環境で解析処理を行い、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している。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担っている。解析処理は多くの計算機リソースを必要とし、解析処理結果は数十 GB～数 TB にもなり、今後さらに増大する見込みである。並行して次々と観測データの解析を行う一方で、解析処理結果の削除も行われるため、解析環境には複数台の解析サーバと大容量のストレージが効率よく使える環境が求められる。これまで解析サーバは Lustre という分散ストレージを共有ストレージとして利用してきたが、Lustre サーバ/クライアントは同一の kernel であることなど運用面、利用面で制約が多かった。そのため、解析環境の OS のサポートの End Of Life を機に柔軟性、拡張性のある解析環境を構築することとなった。

2. Ceph とは

Ceph[1]は、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、切り出したストレージプール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(RASODGW)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、オープンソースの分散ストレージソフトウェアである。Ceph はデータを冗長化して Ceph クラスター内のノードに分散配置し、Ceph クライアントは CRUSH アルゴリズムにより自身で格納先を計算してノードに直接アクセスし、効率よくデータを取得する。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 の耐障害性、拡張性、高可用性という特長より解析環境の新たな大容量ストレージに Ceph を採用した。

3. アルマプロジェクトの Ceph ストレージ

3.1. 構成

現在運用しているアルマプロジェクトの Ceph ストレージのクラスター構成を図 1 に示す。Ceph クラスターは 7 台のノードで構成されている。このうち、ノード ceph01～ceph05 では Ceph の基本的なサ

ービスを提供する monitor デーモン(Ceph の構成管理)、manager デーモン(サービスの管理)、osd デーモン(データの格納)を稼働し、ノード ceph-mds, ceph-mds2 では cephFS を提供するための mds デーモン(メタデータの管理)を稼働している。

本発表を行った 2022 年 1 月時点では、Ubuntu20.04LTS 上で Ceph Octopus(ver.15.2.15)を運用している。本ストレージの運用を始めた 2019 年 9 月は、ceph01～ceph03、ceph-mds、ceph-mds2 の 5 台のノードで構成していたが、バージョンアップやノードの追加、ハードウェアの強化、リプレースを重ね、現在に至る(表 1)。また、接続する解析サーバの台数の増加に伴い、L2/L3 スイッチをポート数の多いものにリプレースし、さらにスイッチを多段化した。これらを耐震ラックに格納している。

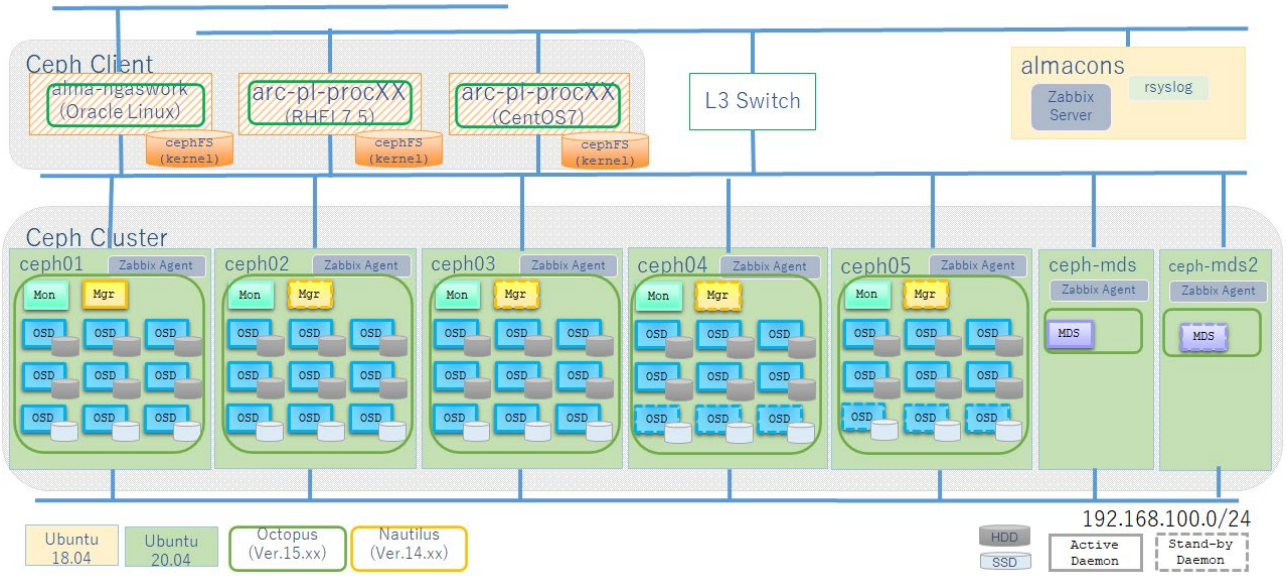

図 1：アルマプロジェクトの Ceph クラスターの構成

表 1：サーバ構成

Node	ceph01	ceph02	ceph03	ceph04	ceph05	ceph-mds	ceph-mds2(*1)
Memory	(32GB->)128GB			128GB		48GB	64GB
CPU	Xeon E5-2609 8C/8T x 2			Xeon Silver 4208 8C/16T x 2			
Disk	HDD 3.5"SAS 6TB x 30, SSD 240GB x 1, 500GB x 11			HDD 3.5"SAS 6TB x 30 SSD 240GB x 1, 500GB x 7			
Network	10GbE 2 ポート x 2 (2x2 ボンディング)					10GbE , 1GbE 各 1 ポート	
OS	(Ub18.04->)Ubuntu20.04LTS (*2)			Ubuntu20.04LTS (*3)		(*2)	(*3)
Ceph version	(14.2.xx->)15.2.15 (*4)			(15.2.xx->)15.2.15 (*5)		(*4)	(*5)
Ceph daemon	mon:1, mgr:1,osd:41			mon:1, mgr:1,osd:30		mds:1	mds:1

注：表のグレーの文字は構築当初の構成。(*1) 故障により 2021 年 12 月にハードウェアをリプレース

3.2. 解析サーバでの利用

アルマプロジェクトでは、約 25 台の解析サーバを保有し、アルマ望遠鏡から得られた観測データの解析処理を行っている。このうち、解析処理方法やソフトウェアにより、約 15 台が Ceph ストレージに接続し、ネットワークファイルシステムとして利用している(図 2 背景が薄い黄色の部分)。アルマ望遠

鏡の観測で得られたデータの解析処理結果は 1 件あたり数十 GB 程度で、これらのデータの解析処理のログも含めると大きいものでは数 TB にもなる。解析サーバは CephFS で同じボリュームを read/write 可能な権限でマウントして共有している。また、解析結果中継サーバ(alma-ngaswork)は CephFS で解析サーバと同じボリュームの一部を read 権限でマウントし、解析処理結果をチリのサーバに転送している(図 3)。その他、環境やデータのバックアップ、移行のために、Ceph ストレージより RBD で領域を切り出して利用、使用後に開放し、一時的なデータ保管場所として柔軟に使用している。

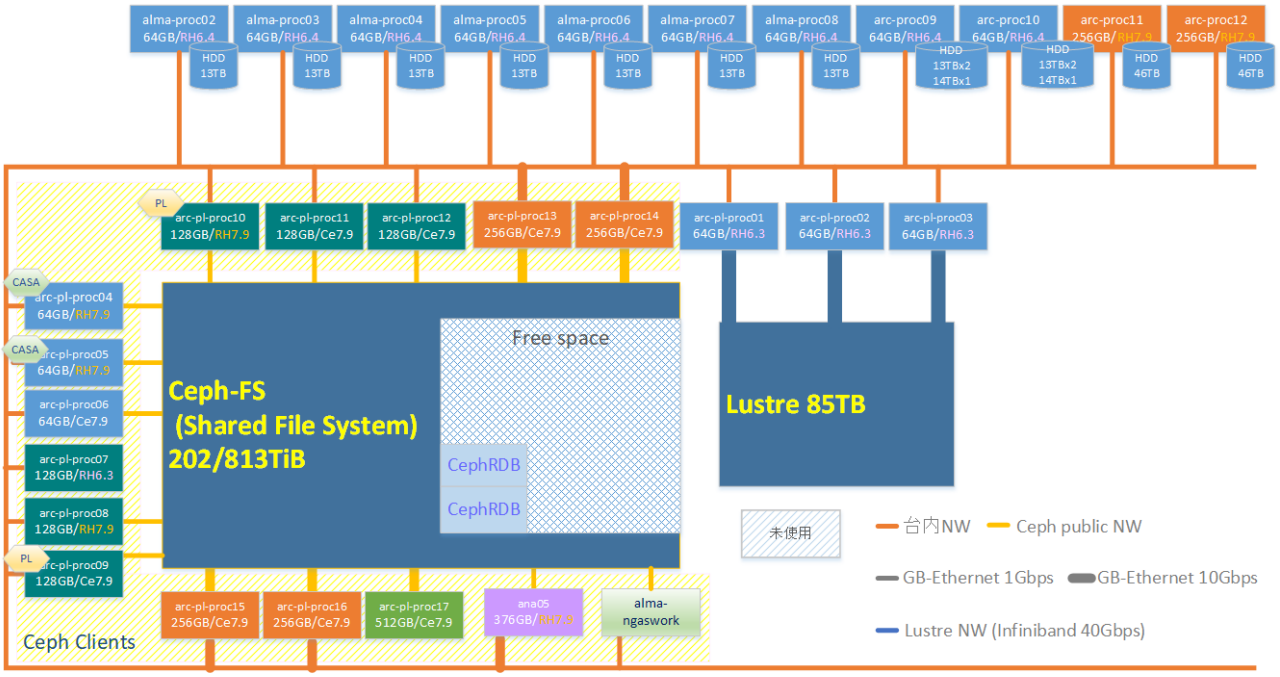

図 2：解析サーバでの Ceph ストレージの利用

```
/mnt/cephfs/alma/ #解析サーバ(rw)
└── qa2/
    ├── PL/
    ├── clai/
    ├── dared_mnt/
    └── results/
        ├── ..../
        ├── products/ #解析結果中継サーバ(r)
        └── ..../
```

図 3：CephFS でのマウント

表 2：CASA 5.4.70 SingleDish 処理時間

OS	Storage	Network	Time(min)	Remarks
RHEL7	LocalSSD	10Gbps	106	Single run
RHEL6	Lustre	40Gbps	129	Single run
RHEL7	RBD	10Gbps	112	Single run
RHEL7	RBD	10Gbps	122	3 parallel run on 1host
RHEL7	CephFS	10Gbps	124	Single run
RHEL7	CephFS	10Gbps	132	4 parallel run on 1host

3.3. 検証

- **I/O 性能および解析処理時間**

解析サーバにおいて、10GiB のファイルの書き込みで RBD は 666MB/s、CephFS で 60MB/s、27GB のファイルの読み込みで CephFS で 220MB/s であった。また、データの格納場所として、これまで使用していた解析環境の LocalSSD, Lustre, 現解析環境の RBD, CephFS でそれぞれ解析処理を行った処理時間を表 2 に示す。処理時間の差は小さく、並列に処理を行った場合の速度の低下も少ない。

- **耐障害性**

現在、データを冗長度 3 で格納している。OSD デーモンを 1 つ停止しても、read/write に影響はなく、速度もほぼ変化しないことを確認した。また、Active/Stand-by で稼働している Ceph MDS デーモンにおいて、Active な MDS デーモンを停止すると、スムーズに Stand-by デーモンに切り替わり、ファイルのメタデータ取得に切り替わりを実感することはなかった。

3.4. 運用・保守

- **監視**

我々は、Ceph の状況を把握するために、Zabbix[2]と Ceph Dashbard を導入し、利用している。Zabbix では、Ceph クラスターの機器の運用状況をブラウザ上で把握している。あらかじめ設定した閾値をトリガーに障害を検知し、メールでその通知を受け取っている。また、Ceph のプラグインの Ceph Dashboard によりブラウザ上で Ceph の状況の把握や操作を行っている(図 4)。

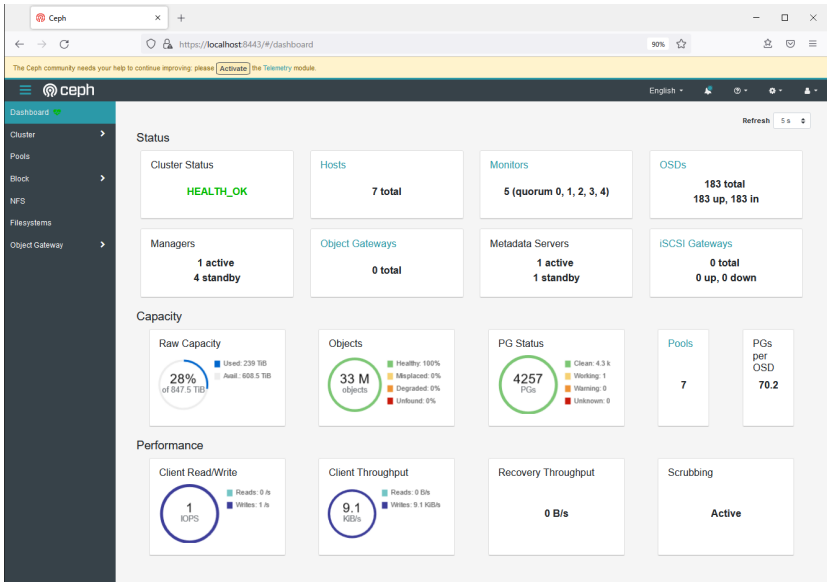

図 4： Ceph Dashboard による Ceph クラスターの監視

- **保守**

「3.3 検証 ・耐障害性」で示したように、冗長構成でサービスの可用性が高いため、多くの場合、サービスを停止することなく、ノードやサービスの保守を行うことができる。事前に仮想環境で構築したテスト環境で行ってから本番環境に実施する。これまでに、OS や Ceph のアップグレード、ハードウェア保守などをローリングアップデートで行ってきた。

- **これまでの苦労・失敗など**

サービス運用開始当初はメモリ不足による動作不安定や SFP+モジュールの不具合による通信障害が度々あり、メモリの追加やモジュールの取り換えなどを行った。また OS をアップグレードした際、起動できなくなり grub の再構成を行った。この時、解決まで 1 週間くらいノードが 1 つ少ない縮退運用していたが、Ceph のシステムが壊れることはなかった。システムとして非常に堅牢である。

4. まとめ、今後の課題

Ceph は拡張性が高く、耐障害性も強い。ただ、柔軟性、拡張性が高い分、パラメータがたくさんあり理解しきれていない。最適化の余地が多分にあると思う。目下、廃止予定の古いストレージからのデータの移行が急がれ、CephFS のプールを利用効率の良い冗長構成に変更する検証を進めている。また複数の用途、ストレージ形式での利用が見込まれているため、ネットワークの拡張が必要となっている。

5. 参考文献

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

[2] Zabbix, https://www.zabbix.com/jp/