C-016

SSD を活用した HDD ベース分散ストレージの評価
HDD-based Distributed Storage Evaluation using SSD

天野 隆†
Takashi Amano

1. はじめに

画像処理用などの大容量ストレージのアクセス性能高速化が求められている。HDD ベースストレージは安価で大容量であるがアクセス性能が低速である。HDD と比較すると，SSD ベースストレージは高価で小容量であるがアクセス性能が高速であるという特徴がある。これらの特徴などを考慮して大容量でアクセス性能が高速なストレージの実現が課題である。HDD ベース分散ストレージの一部に SSD を活用することでアクセス性能の高速化を図り，その分散ストレージのアクセス性能を評価した。評価の結果，SSD を活用した HDD ベース分散ストレージのアクセス性能高速化を確認した。

2. 画像処理システム

2.1 半導体製造分野のストレージ

半導体製造分野では，半導体に電子回路を作成した後に電子回路が半導体上に正常に作成されているかどうかを検査している。まず，電子顕微鏡装置でスキャンした画像を一旦ストレージに保存する。次に，そのストレージと画像処理サーバから構成される画像処理システムでその画像を読込み，目標とする電子回路の画像とスキャンした画像を比較することで半導体上の電子回路に欠陥がないかを確認する。

電子顕微鏡装置では，年々時間当たりの画像のスキャン可能回数が増加している。画像処理システムでは，電子顕微鏡装置がスキャンした画像を処理するために欠陥検出アルゴリズムや CPU の高速化に加え，ストレージのアクセス性能高速化が求めらている。

2.2 分散ストレージ

本研究では，画像処理システムのストレージとして分散ストレージの Ceph[1]を用いた。Ceph は，ブロックストレージとして使用できる RADOS(Reliable Autonomic Distributed Object Store) Block Device，ファイルシステムとして使用できる CephFS(Ceph File System)，オブジェクトストーレジとして使用できる RADOS Gateway がある。これらは，ネットワークを経由してクライアントから使用することができる。ファイルシステムは，複数のクライアントでマウントして使用可能な共有ファイルシステムとなっている。本研究では，Ceph の構築を HDD をベースにして共有ファイルストレージの CephFS として使用する。

†株式会社 日立製作所
デジタルプラットフォームイノベーションションセンタ
Hitachi, Ltd.
Digital Platform Innovation Center

2.3 分散ストレージの問題

CephFS は，BlueStore[2]と呼ばれる仕組みを使用してデータを保存している。BlueStore は，データを保存するときにデータの付随情報であるメタデータを作成してデータとともに HDD に保存する。このとき，コスト低減のために同一 HDD にデータとメタデータを保存すると，HDD のアクセスが増加し性能劣化の原因となってしまうことが問題である。

2.4 本研究の狙い

上記の性能劣化の問題に対して Ceph ではデータとメタデータを別々の場所に保存できる設定を提供している。

本研究の狙いは，全体コストを抑えつつ CephFS に画像を高速に書込みできるようにすることである。

3. SSD を活用した方式

CephFS を使用して画像にアクセスする性能を確認するにあたり，HDD のみを使用した HDD 方式と SSD を活用した HDD＋SSD 方式を用いる。

3.1 HDD 方式

HDD 方式は，HDD のみを使用してデータとメタデータを同一の HDD に保存するように設定した構成である。(図 1 参照)。

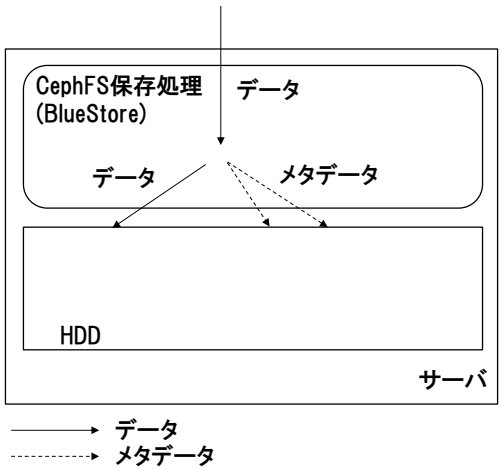

図 1 HDD 方式

3.2 HDD＋SSD 方式

本研究では，HDD 方式を改善した HDD＋SSD 方式を比較対象とする(図 2 参照)。改善の方針としては，コストを抑えつつ性能向上させることとする。

メタデータのサイズは，データのサイズに対して 4%程度である。すべての HDD を SDD にすることで性能向上を

図ると SSD は HDD と比較して高価なため全体コストが高くなってしまう。本研究では，コスト抑えつつ性能向上させるために小容量のメタデータを保存する部分のみを SSD にすることとする。

4TB HDD の 2 台分のメタデータを 480GB の SSD に保存することとした。メタデータは，データ保護用の情報である block.wal と，データの管理情報である block.db の 2 種類ある。block.wal 用に約 2GB と block.db 用に約 238GB のパーティションを SSD に作成してそれぞれのメタデータを保存する。

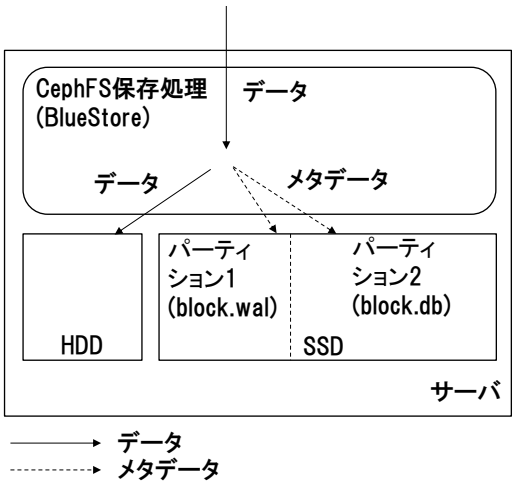

図 2 HDD＋SSD 方式

4. HDD＋SSD 方式の評価結果

4.1 評価環境

CephFS(表 1 と表 2 参照)は，保存領域となる OSD が 84 個で構成される。ノードを画像処理サーバと想定して処理結果で出力される 34MB の画像ファイルの書込み性能を評価する。性能測定ツール fio を使用して，合計 1,680 並列で画像ファイルを書込む性能を評価する(図 3 参照)。

表 1 CephFS のノードの情報

項目	仕様
ネットワーク	10Gbps
HDD	4TB 2 台(SAS)
SSD	480GB 1 台(SAS) 下記メタデータを HDD2 台分保存 block.wal：約 2.1GB x 2 block.db：約 238GB x 2
OS	CentOS 7.6

表 2 CephFS の情報

項目	仕様
OSD 数	合計 84 個
MDS サーバ	アクティブ 1 台
データプール	イレージャーコーディング (データ(k)=4，冗長コード(m)=1)
Ceph バージョン	14.2.1

MDS：Meta Data Server
OSD：Object Storage Device

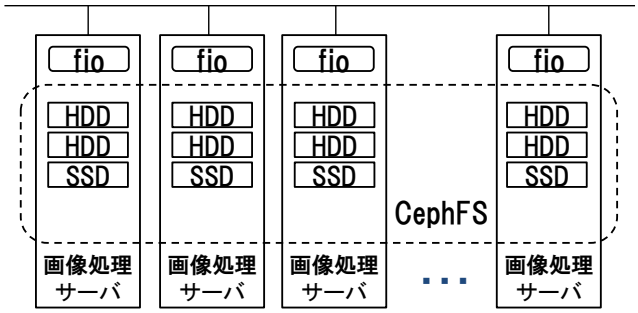

図 3 HDD＋SSD 方式の構成

4.2 評価結果

測定範囲は，画像ファイル(34MB)を書込みしている 5 分間である。評価の結果，HDD 方式は 2.2GB/s，HDD＋SSD 方式は 5.8GB/s の書込み性能であった(表 3 参照)。

表 3 評価結果 [単位：GB/s]

項目	HDD 方式	HDD＋SSD 方式
書込み性能	2.2	5.8

4.3 HDD＋SSD 方式の効果

評価結果から，HDD 方式の 2 倍以上の書込み性能がでることを確認できた。HDD 方式から小容量のメタデータの保存先のみを HDD より高速な SSD にすることで，全体コストを抑えつつ性能向上できることを確認できた。

5. おわりに

CephFS のメタデータ部分のみを SSD に保存した HDD＋SSD 方式を評価し，HDD 方式よりも高速な書込み性能であることを確認した。

本研究の今後の課題を次に示す。

(1) NVMe 対応 SSD などによるさらなる高速化

謝辞

職場の関係者各位には，本研究に関する検討内容について議論頂くとともに，貴重なご意見を頂いた。

参考文献

[1] Ceph，Ceph ホームページ，https://ceph.io/，2022 年 6 月現在
[2] BlueStore, Ceph ドキュメントホームページ, https://docs.ceph.com/en/nautilus/rados/configuration/bluestore-config-ref/，2022 年 6 月現在