C-011

画像保存クライアントを想定した分散ストレージ評価
Distributed Filesystem Evaluation with Client to Store Image File

天野 隆†
Takashi Amano

1. はじめに

画像処理用ストレージのアクセス高速化が求められている。画像保存クライアントの OS によっては，分散ストレージを直接マウントできず高速にデータを保存できない場合がある。そのような画像保存クライアントの場合でも分散ストレージに高速にデータを保存することが課題である。画像保存クライアントから Samba[1]サーバ経由で分散ストレージにデータを保存する環境を構築して評価を行った。評価の結果，画像保存クライアントから分散ストレージに高速にデータを保存できることを確認した。

2. 画像処理システムの課題

2.1 画像像保存クライアント

半導体製造の分野では，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，検査処理サーバ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。

画像保存クライアントは，このような電子顕微鏡装置を想定している。一般的なクライアントとの違いは，データの読込みがなく書込みだけという点である。

2.2 分散ストレージ

本研究では，分散ストレージとしてオープンソースの Ceph[2]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。

2.3 画像処理システム

画像処理システムは，前述の画像保存クライアントと分散ストレージ Ceph の共有ファイルシステム CephFS を組み合わせて構築することを想定している。

画像保存クライアントの OS は Windows‡である。Ceph の OS は CentOS である。

†株式会社日立製作所
デジタル PF イノベーションショセンタ
Hitachi, Ltd.
Center for Technology Innovation - Digital Platform
‡Windows は Microsoft 社の登録商標である。

2.4 画像処理システムの問題

前述した画像処理システムでは，画像保存クライアントと CephFS の接続に問題がある。CephFS は Windows でのマウントがサポートされていないため，画像保存クライアントから直接 CephFS に対して画像を保存することができず，性能のボトルネックとなっている。

2.5 本研究の狙い

上記の問題に対して，Windows から間接的に CephFS にアクセスする方式として Samba を使用する方式が提供されている。

本研究の狙いは，Samba を使用した方式で画像保存クライアントから CephFS に高速に画像を保存できるようにすることである。

3. Samba を使用した方式の提案

Samba を使用した方式で画像保存クライアントから CephFS に画像を保存する性能を確認するにあたって，従来方式と本研究での提案方式を用いる。

3.1 従来方式

従来の方式としては，CephFS 向けに提供されている Samba の vfs_ceph[3]モジュールを使用した vfs_ceph 方式を用いる(図 1 参照)。

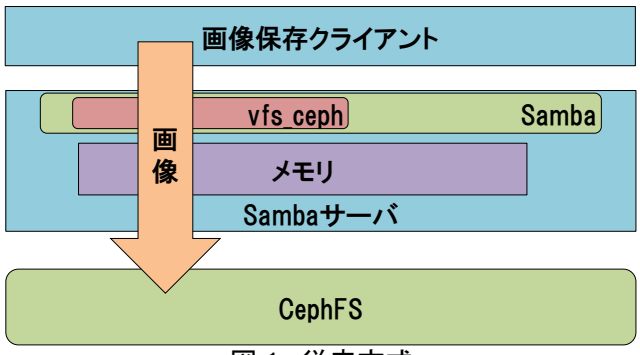

図 1　従来方式

3.2 本研究での提案方式

本研究では，Samba を使用した 2 段階書込み方式を提案する(図 2 参照)。

2 段階書込み方式は，まず画像保存クライアントから Samba サーバのメモリに画像ファイルを書込む(第 1 段階書込み)。次に，メモリから CephFS に画像ファイルを書込む(第 2 段階書込み)。

第 1 段階書込みでは，下記を事前に実施しておく。

(1-1)メモリに tmpfs ファイルシステムで一時保存領域を事前に作成

(1-2)Samba で(1)の一時保存領域を画像保存クライアントでマウントできるように共有設定
(1-3)画像保存クライアントで(2)で共有された一時保存領域をマウント

第 2 段階書込みでは，一時保存領域から CephFS への画像の書込みに Lsyncd を使用する。Lsyncd は，指定ディレクトリのファイル書込みを検知できる inotify イベントを監視し，ファイル書込みを検知した後に別ディレクトリへファイルを移動するように設定することができるツールである。この Lsyncd を使用することにより，一時保存領域の画像ファイルの書込みを検知して，画像ファイルを CephFS に移動することができる。

第 2 段階書込みでは，事前に下記を実施しておく。
(2-1)CephFS をマウント
(2-2)Lsyncd に画像ファイルの移動を設定

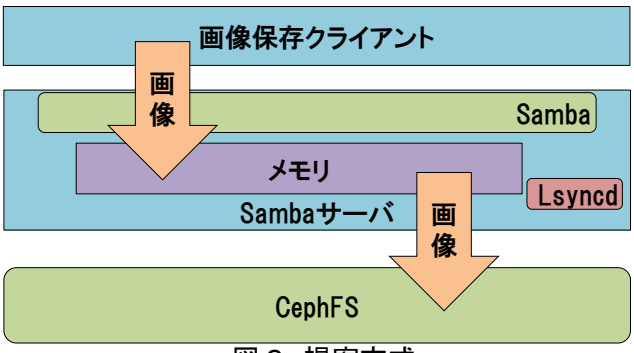

図 2　提案方式

4. 提案方式の評価結果

4.1 評価環境

画像保存クライアント(表 1 参照)は 6 台，Samba サーバ(表 2 参照)は 1 台の構成である。CephFS(表 3 と表 4 参照)はノード 42 台から構成される。

表 1　画像保存クライアントの情報

項目	仕様
CPU	Intel Xeon Gold 6138T 2GHz
CPU ソケット	2 個使用
メモリ	256GB
ネットワーク	10Gbps
画像ファイル	34MB
書込み並列度	1
OS	Windows 10

表 2　Samba サーバの情報

項目	仕様
CPU	Intel Xeon Gold 6138T 2GHz
CPU ソケット	1 個使用
メモリ	128GB
ネットワーク	10Gbps
Samba バージョン	4.9.1
OS	CentOS 7.6

表 3　CephFS のノードの情報

項目	仕様
CPU	Intel Xeon Gold 6138T 2GHz
CPU ソケット	2 個使用
メモリ	256GB
ネットワーク	10Gbps
HDD	4TB　2 台
OS	CentOS 7.6

表 4　CephFS の情報

項目	仕様
ノード数	42 台
OSD 数	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

4.2 評価結果

測定範囲は，書込みを開始して 10 分経過後の 5 分間である。評価の結果，従来方式は 303MB/s，提案方式は 934MB/s の書込み性能であった(表 5 参照)。

表 5 評価結果　[単位：MB/s]

項目	従来方式	提案方式
書込み性能	303	934

4.3 提案方式の効果

評価結果から，従来方式の 3 倍以上の書込み性能がでることを確認できた。従来方式では，vfs_ceph モジュールがボトルネックになっていると考える。

5. おわりに

Samba を使用した方式で画像保存クライアントから CephFS に高速に画像を保存する 2 段階書込み方式を提案し，従来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) Samba サーバの障害時の対応
(2) CephFS の性能低下時の対応

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] Samba，Samba ホームページ，https://www.samba.org/，2021 年 6 月現在
[2] Ceph，Ceph ホームページ，https://ceph.io/，2021 年 6 月現在
[3] vfs_ceph, Samba ホームページ, http://www.samba.gr.jp/project/translation/current/htmldocs/manpages/vfs_ceph.8.html，2021 年 6 月現在