Gfarmシンポジウム2023

LustreからGfarmへの透過的なデータアーカイブの検討

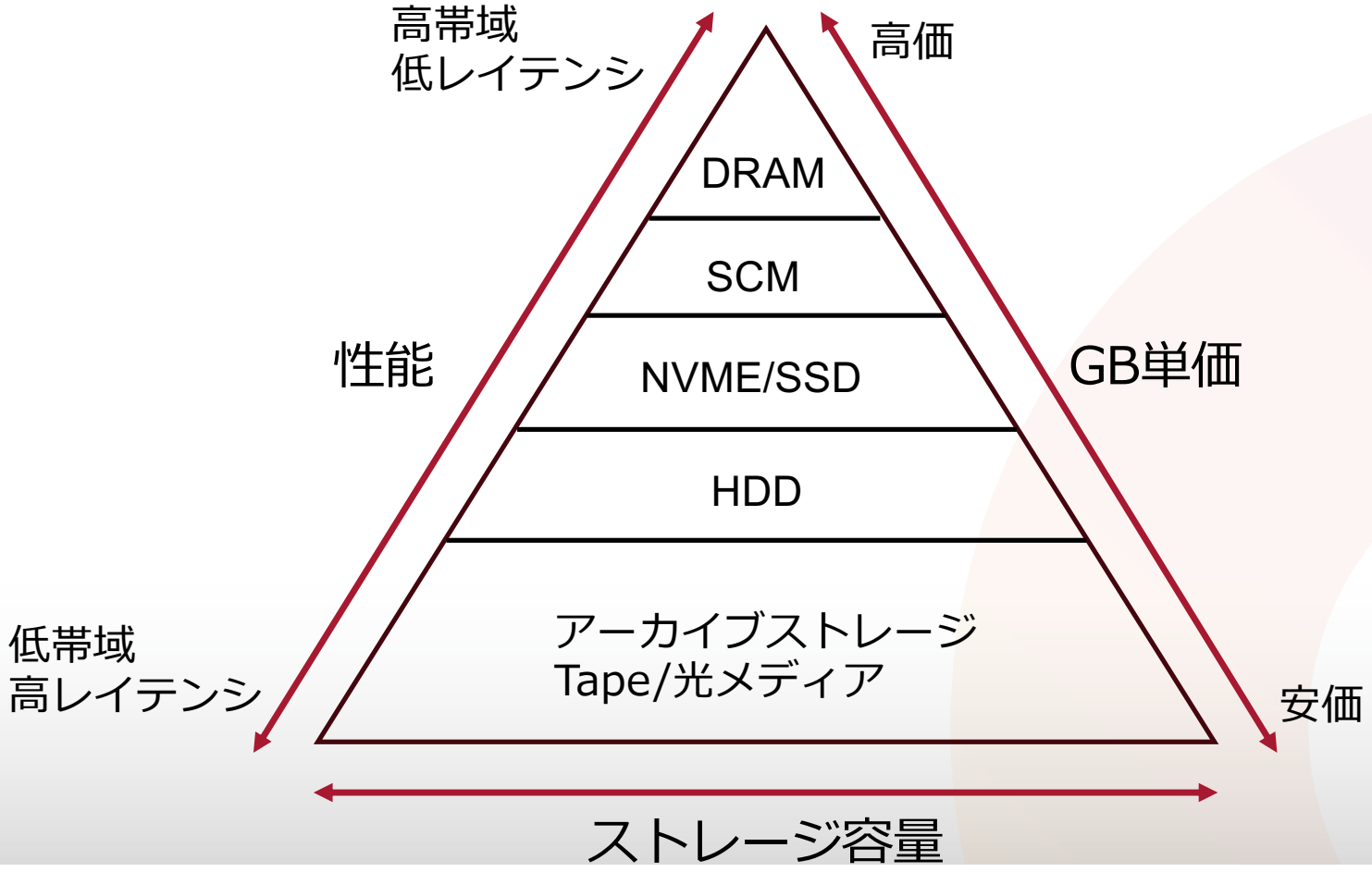
データダイレクト・ネットワークス・ジャパン

井原 修一

本日の内容

- 階層ストレージについて
- Lustre HSM(Hierarchical Storage Management)とは
- Lustre HSMの応用範囲
- Lustre HSM for Gfarmについて
  - Gfarmを使ったLustre HSMの実現方法およびデモ
  - 課題とまとめ

階層ストレージ



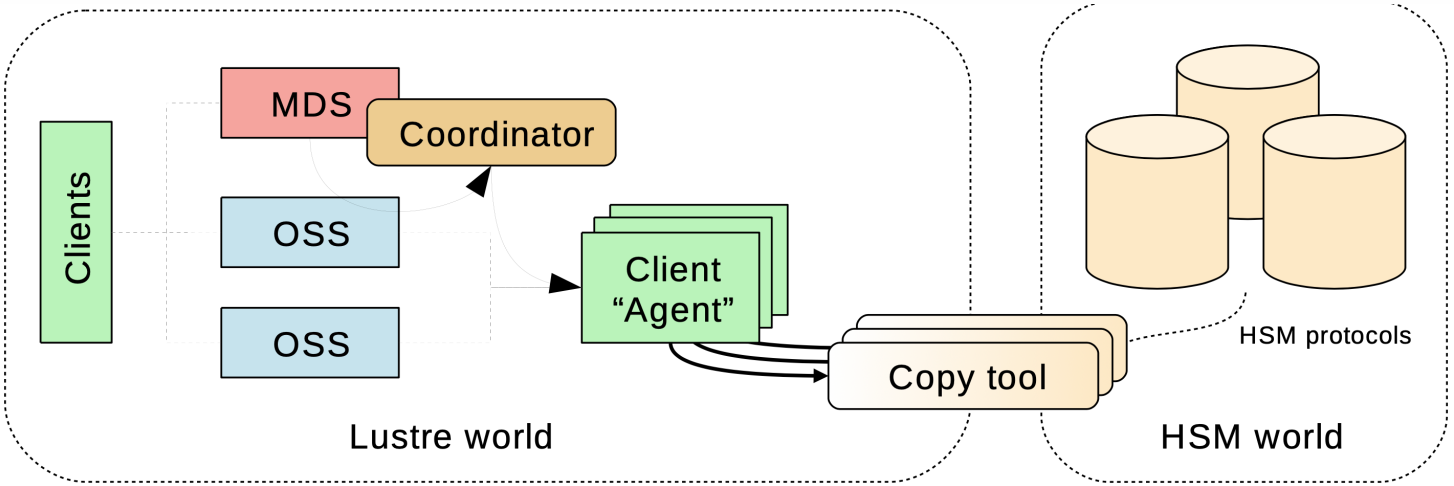
ストレージ階層化を実現するLustreの様々な機能

- **PCC(Persistent Client Cache)**
  - クライアントのローカルストレージを透過的なキャッシュで利用
- **DoM(Data on MDT)**
  - MDTのSSD/NVMeにデータも配置
  - RPCの削減と小さなファイルの高速化
- **OST Read Cache**
  - OSSのRAMにOSTからReadしたものを保持
- **HotPool**
  - 1つのファイルシステム内にてSSDのOSTとHDDのOSTに透過的なデータ管理

本日のテーマ

- **HSM**
  - ネームスペースの異なるストレージシステムをアーカイブストレージとしてLustreの名前空間で管理
- **その他RAMにおけるWrite/Readのバッファキャッシュ**
  - LDLMロックによってデーター貫性の維持
  - ページ単位のキャッシュ管理

Lustre HSM(Hierarchical Storage Management)とは



- Lustre HSMにより異なるネームスペースを接続し単一のネームスペースを構築
  - コピーツールがLustre Client上で動作
  - コピーツールにより様々なアーカイブストレージをサポート
    - Lustre <-> Tape, Other POSIX Filesystem, S3, Google Drive等
- 単一のLustre HSMインターフェース

Lustre HSMの応用

- Lustreからテープライブラリへ
  - HPSS
    - CEA(フランス)がcopytoolの開発およびメンテナンス
  - DMF
    - SGI/HPEによって開発メンテナンス
  - TSM
    - GSI(ドイツ)がcopytoolの開発およびメンテナンス

    https://github.com/tstibor/ltsm
- Lustreからオブジェクトストレージへ
  - AWS FSx for LustreとS3
  - Azure Managed LustreとMicrosoft Azure Blob Storage

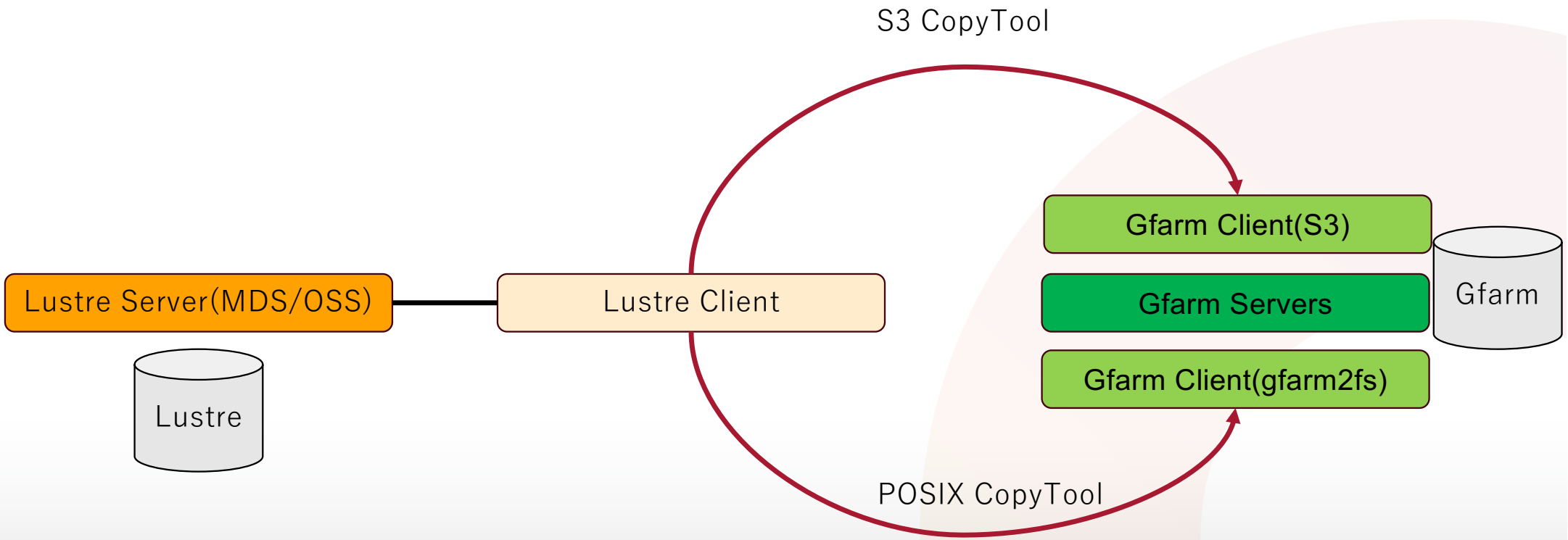
Lustre HSM for Gfarmを検討

- 考えられるワークフロー
  - Lustreで作成したファイルをGfarmにアーカイブ
  - Gfarmで作成したファイルをLustreにインポート
- コピーツール

lustre_hsm_xxx() APIが用意されておりそれぞれのHSMコマンドに対するイベントを定義

  - POSIX Copyツール
    - gfarm2fsでマウントされたPOSIXマウントポイントにアーカイブ
  - S3 Copyツール
    - S3 GWサーバを利用してGfarm上のS3 Bucketディレクトリにアーカイブ
  - Gfarm Copyツール
    - Gfarm APIを利用して新規にGfarm Copyツールを開発(現在存在しない)
    - Gfarmに直接アーカイブ

Lustre HSM for Gfarmのデモ環境



S3を利用したHSM

- S3 Copytool for Lustre HSM
  - Lustre-obj-copytool (Compute Canada)
    - https://github.com/ComputeCanada/lustre-obj-copytool
  - Estuary (ICHEC)
    - https://git.ichec.ie/performance/storage/estuary
    - Lustre-obj-copytoolをフォークして独自にメンテナンス

S3を利用したLustre-HSM初期設定

- HSMの有効化(MDSサーバ)

```
[root@lustre-server1 ~]# lctl set_param -P mdt.*.hsm_control=enabled
[root@lustre-server1 ~]# lctl get_param mdt.*.hsm_control
mdt.lustre-MDT0000.hsm_control=enabled
mdt.lustre-MDT0001.hsm_control=enabled
```

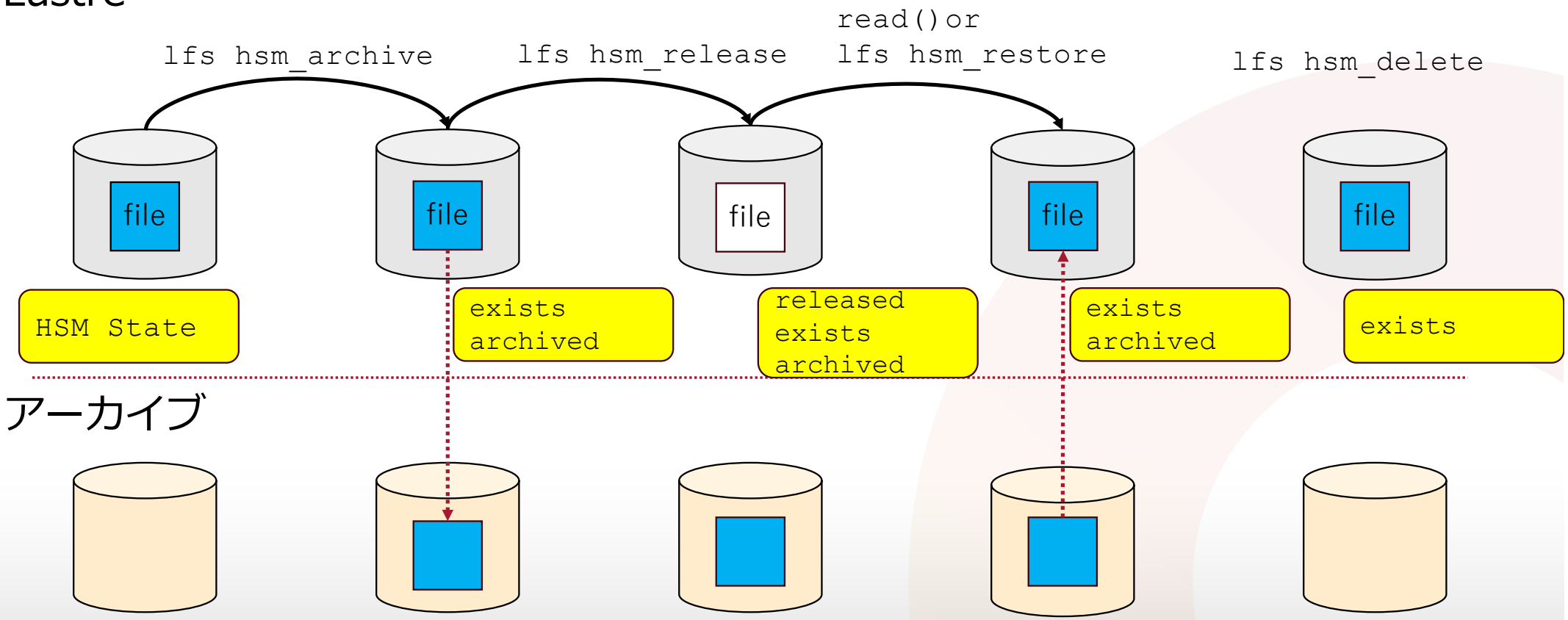
- Gfaram/S3にアーカイブの領域作成(クライアント)

```
[root@lustre-client1 ~]# s3cmd mb s3://mybucket
```

- S3 Copyツールの起動(クライアント)

```
[root@lustre-client1 ~]# s3copytool -c /etc/s3coopytool.conf /lustre
```

デモの前にLustre HSM状態遷移のおさらい



*その他アーカイブデータとLustre上のデータが同期されていない”dirty”の状態が存在

デモ: Lustre HSM for Gfarm/S3 (1)

- LustreからGfarmへアーカイブ

```
[root@lustre-client1 ~]# dd if=/dev/urandom of=/lustre/10mfile bs=1M count=10 oflag=sync
[root@lustre-client1 ~]# df -h -t lustre
Filesystem                 Size  Used Avail Use% Mounted on
10.128.10.233@tcp:/lustre   19G   13M   18G   1% /lustre

[root@lustre-client1 ~]# md5sum /lustre/10mfile
b94494ab06a0f2c8b35e0288bdf9eb2f  /lustre/10mfile

[root@lustre-client1 ~]# lfs hsm_state /lustre/10mfile
/lustre/10mfile: (0x00000000)

[root@lustre-client1 ~]# lfs hsm_archive /lustre/10mfile
[root@lustre-client1 ~]# lfs hsm_state /lustre/10mfile
/lustre/10mfile: (0x00000009) exists archived, archive_id:1
```

サンプルファイル作成

HSM状態の確認

HSMアーカイブ
Lustreに存在かつ
アーカイブされた状態

- Gfarmにてアーカイブを確認

```
user1@client1:~$ gfls -l share/user1/mybucket -h
-rw------- 1 user1    gfarmadm  10.5M Sep  3 09:44 0000000200000bd1_00000004_00000000.0
```

デモ: Lustre HSM for Gfarm/S3 (2)

- Lustreからのリリース

```
[root@lustre-client1 ~]# lfs hsm_release /lustre/10mfile
[root@lustre-client1 ~]# lfs hsm_state /lustre/10mfile
/lustre/10mfile: (0x0000000d) released exists archived, archive_id:1

[root@lustre-client1 ~]# df -h -t lustre
Filesystem                 Size  Used Avail Use% Mounted on
10.128.10.233@tcp:/lustre   19G  2.8M   18G   1% /lustre
```

HSMリリース
Lustreから解放
アーカイブのみの状態

- Lustreへの自動リストア

```
[root@lustre-client1 ~]# echo 3 > /proc/sys/vm/drop_caches
[root@lustre-client1 ~]# md5sum /lustre/10mfile
b94494ab06a0f2c8b35e0288bdf9eb2f  /lustre/10mfile

[root@lustre-client1 ~]# df -h -t lustre
Filesystem                 Size  Used Avail Use% Mounted on
10.128.10.233@tcp:/lustre   19G   13M   18G   1% /lustre

[root@lustre-client1 ~]# lfs hsm_state /lustre/10mfile
/lustre/10mfile: (0x00000009) exists archived, archive_id:1
```

HSMリストア(自動)
Lustreへリストアおよび
アーカイブ状態

デモ: Lustre HSM for POSIX mount point(1)

- LustreからPOSIX/Gfarm(gfarm2fs)へアーカイブ

```
user1@client1:~$ mkdir /tmp/gfarm
user1@client1:~$ gfarm2fs /tmp/gfarm -o allow_root
user1@client1:~$ mkdir /tmp/gfarm/share/user1/archive
root@client1:/# /usr/sbin/lhsmtool_posix --hsm-root /tmp/gfarm/share/user1/archive /lustre
```

POSIX copytoolの仕様上、Lustreのxattrの属性(trusted.hsm, trusted.link, trusted.lov, trusted.lma, lustre.lov)をアーカイブ先のファイルに保存。

Gfarmではuser以外のxattr属性の追記は許可されていないため, アーカイブに失敗。

デモ: Lustre HSM for POSIX mount point(2)

- LustreからPOSIX/EXT4へアーカイブ

```
root@client1:/# lhsmtool_posix --hsm-root /ext4/archive /lustre
```

S3コピーツール同様`lfs hsm_{archive,release,restore,remove}`が利用できる

- アーカイブからLustreへインポート

```
root@client1:~# cd /ext4/archive
root@client1:/ext4/archive# md5sum import/10mfile*
f96841348d47b2727c448628bc11479d  import/10mfile0
7ac66e8e5bed37968958701fc7e3b8b4  import/10mfile1

root@client1:/ext4/archive# lhsmtool_posix --import --hsm-root /ext4/archive import/ /lustre
/lustre
[root@lustre-client1 ~]# lfs hsm_state /lustre/import/*
/lustre/import/10mfile0: (0x0000000d) released exists archived
/lustre/import/10mfile1: (0x0000000d) released exists archived
[root@lustre-client1 ~]# du -h /lustre/import
5.0K /lustre/import

[root@lustre-client1 ~]# md5sum /lustre/import/10mfile*
f96841348d47b2727c448628bc11479d  /lustre/import/10mfile0
7ac66e8e5bed37968958701fc7e3b8b4  /lustre/import/10mfile1
[root@lustre-client1 ~]# du -h /lustre/import
21M /lustre/import
```

インポート後
データはアーカイブのまま

Lustre上はファイルのメタデータが作成されアーカイブ済みのステートになる

Lustre上にてRead実施後
データがリストアされる

課題とまとめ

- Lustre HSM for Gfarmの課題
  - S3やPOSIXのcopytoolをそのまま利用は非効率。要Nativeのcopytoolを実装
  - システムレベル(Lustre/HSM) vs ユーザレベル(Gfarm)
    - Copytoolにおけるユーザ認証
    - Lustreはファイルシステム全体でHSMを有効にする必要がある
- Lustre HSMにおけるGfarmの優位性
  - 高可用性、高性能なアーカイブストレージ
    - 地理的拠点間におけるメタデータおよびデータ多重化
    - ファイルデータチェックサム機能
    - 高帯域のネットワークバックボーン
- 実現可能性
  - クラウド環境にて提供されている”Lustre + オブジェクトストレージ”サービスも実現可能？
  - 多くの日本国内の大学、研究機関ではLustre+Gfarmが既に運用中
  - 将来は”高速ファイルシステム(Lustre) + アーカイブストレージ(Gfarm)”も可能？


ddn