iSCSI インタフェースを持つ分散ファイルシステムを用いた大容量写真共有アプリケーション

大西 祥代
Sachiyo ONISHI

1　はじめに

近年，研究で扱うデータだけでなく，音楽や写真といったマルチメディアファイルなどの，日常の生活で使用するデータ量も増加の傾向にある．そういった背景より，大容量のデータを記録・保存可能なストレージが必要となっている．そこで本研究では，複数のディスクを仮想的に統合することで，大容量のストレージを構築する．構築したストレージを iSCSI[1] インタフェースを用いてネットワークに接続することで，複数のサーバからの容易なアクセスを可能とする．さらに，このシステム上で動くアプリケーションとして，大容量の保存領域を活かした写真共有アプリケーションを作成する．

2　システムの概要

1 つのディスクでは容量に限界がある．そのため，数 TB から数 PB オーダーの容量を持つ巨大なストレージを作るには，複数のディスクを仮想的に 1 つに統合することが必要である．そこで，本システムでは，Lustre[2] や GPFS[3] などの分散ファイルシステムを用いて大容量ストレージを実現する．そして，このストレージ上のファイルに多数のサーバからネットワークを介したアクセスを行うため，本システムでは，ネットワークとストレージ間のインタフェースとして，iSCSI を用いる．iSCSI は，既存の SCSI 技術を IP ネットワーク上で使用可能にしたもので，Ethernet を介して遠隔のストレージへネットワーク接続を実現するためのものである．そして，このシステム上では，大容量ストレージを活かした写真の保存・共有を行う ISDLphoto が動く．ISDLphoto では，大量の写真の検索や管理を容易にするために，写真にタグを付与し，写真同士を関連付ける．本システムの全体像を Fig. 1 に示す．

3　分散ファイルシステム

分散ファイルシステムとは，ネットワーク上の複数のコンピュータのストレージを仮想統合し，ストレージ上のファイルをローカルマシンと同様に利用することを可能とする技術である．複数のストレージを用いることで，大容量のストレージを実現する．また，複数のストレージにデータを分割して保存するため，高速なアクセスが可能で障害に強いシステムとなる．

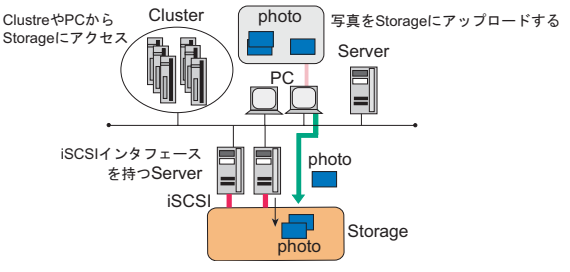

Fig. 1　システムの全体像

本研究では，Lustre と GPFS を用いて分散ファイルシステムを構築し，その性能比較を行う．

3.1　Lustre

Lustre は，Cluster File Systems，Inc. によって開発された大規模な分散ファイルシステム構築ツールである．Lustre の基本的な構成は，MDS(Metadata Server)，OST(Object Storage Target)，Client からなっており，認証サーバとして LDAP を用いている．以下にそれぞれの機能を示す．

- MDS
  ファイルの位置情報のみを持つメタデータを管理する．
- OST
  実データのみを管理しており，Client の要求に従って，実データの読み書きを行う．
- Client
  MDS や OST に対して問い合わせを自動的に行う．ユーザは Client とのみ通信を行う．

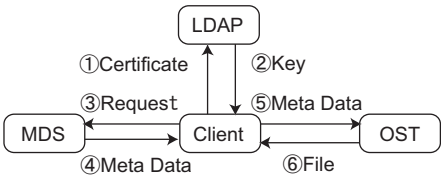

Fig. 2　Lustre の動作

Lustre の動作を Fig. 2 に示す．まず，Client は LDAP サーバへ接続し認証が行われると，MDS へ接続するためのセッションキーを受け取る．そして MDS から参照したいファイルの位置情報を持つメタデータを受け取り，OST はそのメタデータから，実データを取り出し Client に返す．以上の処理によりユーザは求めるデータを取得することができる．

以下に Luetre の特徴を挙げる．

- 高速な I/O 性能
  Lustre は，1 つのデータを複数のハードディスクに分けて同時に書き込むストライピングを用いている．データを保存する OST を複数台設置し，データを分割して各 OST に格納するので，OST の増加に伴い，I/O 性能が上昇する．
- 耐故障性の向上
  OST・MDS に待機系サーバを用意しフェールオーバーを行うことで，障害に強いシステムとなる．

3.2 GPFS

GPFS(General Parallel File System) は米 IBM 社により開発された分散ファイルシステムである．メタデータと実データを同じノードで管理する．GPFS では SAN(Storage Area Network) を用いてディスクの共有を行う．SAN とはサーバとストレージ間を Fiber Channel を用いて接続するものであり，高速なデータ転送を可能とする．以下に GPFS の特徴を挙げる．

- 高いシステム・パフォーマンス
  複数のノードから，同一ファイルへの同時アクセスができるため，高速アクセスが可能となる．
- システムの柔軟性
  GPFS がシステムに構成された後でも，GPFS を再構成することができる．

4 iSCSI

iSCSI とは，記憶装置とコンピュータ間の通信に利用する SCSI コマンドを IP ネットワーク経由で送受信するためのプロトコルである．

iSCSI では，SCSI コマンドやレスポンスなどのメッセージとデータを，TCP/IP パケットにカプセル化することで，パラレル・バスを利用した SCSI トランスポートを既存の IP ネットワークによるトランスポートに置き換え，データ送受信を行う．

4.1 特徴

以下に iSCSI の特徴を挙げる．

- IP ネットワーク上に簡単にストレージを接続可能
- Ethernet を使用できるため導入コストが安価
- ルーティング可能な TCP/IP ベースのため，通信距離が無制限
- ブロック単位でディスクに読み書きするため高速

4.2 iSCSI の実装形態

iSCSI を実装するためには，以下の 2 つの方法がある．

- ソフトウェアによる利用
  既存の NIC を使用できるため，コストが安いというメリットがある．しかし TCP/IP 処理や iSCSI のプロトコル処理を CPU で行うためサーバへの負担が大きい．
- ハードウェアによる利用
  専用ハードウェアが高価なため，コストが高い．しかしサーバの CPU 負担を軽減することができる．

5 ISDLphoto

本システム上で動作するアプリケーションとして，写真共有アプリケーションである ISDLphoto を作成する．ISDLphoto は以下の 3 つの機能を持つ．

- 大容量保存
  大容量のストレージがあるため，容量を気にすることなく大量の写真保存が可能である．
- 共有機能
  ストレージがネットワーク上にあるため，保存された写真の閲覧・ダウンロードが誰でも可能である．
- 検索・管理機能
  Fig. 3 に示すように写真にタグを付与することで，写真同士が関連を持つ．Fig. 4 のようにタグで写真を分類することで検索・管理が容易になる．また，写真を辿っていくことにより，新たな興味の発見や，他者とのコミュニケーションが生まれるなどのおもしろさがある．

分散ファイルシステムを利用することで，以上の要件を満たすことができる．

Fig. 3　写真へのタグ付け

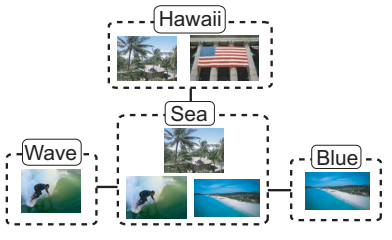

Fig. 4　タグによる写真の関連

6 まとめ

本研究では，Lustre や GPFS を用いて大容量ストレージを持つ，大規模な分散ファイルシステムを構築し，iSCSI を用いることで，既存のネットワークを介したアクセスを可能にする．そして，本システム上で動作するアプリケーションとして写真共有や，タグによる写真検索・管理を行う ISDLphoto を作成する．今後は分散ファイルシステムを Lustre と GPFS で構築し，性能検証を行う．

参考文献

1) IETF IPS
http://www.ietf.org/rfc/rfc3720.txt
2) Cluster File Systems, Lustre
http://www.lustre.org/docs/whitepaper.pdf
3) IBM, General Parallel File System
http://www-03.ibm.com/servers/eserver/clusters/software/gpfs.html