Lustre File System

〜Lustre File System〜

山口 尚平
Shohei YAMAGUCHI

Abstract: The number of nodes which cluster have is increasing sharply in recent years.Distributed file system such as NFS and ASF do not satisfy the requirements of today's high-performance computing environments.Lustre was developed in order to solve the problem.Lustre provides high I/O throughput in clusters and shared-data environments and also protection from single points of failure.Therefore Lustre satisfy the requirements of large clusters today.

1　はじめに

近年，並列計算に用いられるクラスタが大規模化し，通常クラスタで用いられるファイルシステムである NFS や AFS では．マウントできるノードの制限やネットワークの負荷の面で対応しきれない．

これらの問題を解決するため，クラスタ用に最適化されたファイルシステムである Lustre が開発されている．本稿ではこの Lustre の概要を述べる．

2　Lustre とは

Lustre[1] とは Clustre File Systems,Inc. によって開発されているオープンソースのファイルシステムである．これは既存のファイルシステムでは対応できないような大規模なクラスタなどのシステムで利用することを目的として開発されている．

2.1　Lustre の構成

Lustre の基本的な構成要素として Client，Meta Data Server(MDS)，Object Storage Targets(OST) があげられる．これらを用いた Lustre の構成図を Fig. 1 に示す．

ここでは Client，MDS．OST について述べる．

- Client： Client はユーザからファイルの読み書きの要求を受け，実際にファイルの読み書きが行われるまでの作業，つまり MDS や OST に対する問い合わせをユーザの変わりに行うものである．
- MDS: MDS はファイルシステムのメタデータ (ファイルの位置情報を持つデータ) を保持している．つまり実際のデータはそこには無く，Client に要求されたデータがどこにあるかをという情報を Client に送信する機能を持っている．
- OST: OST はストレージを管理しているサーバであり，ここでは実際のデータを扱う．ここでは Client からの要求に従ってファイルをストレージに書き込んだり，データを Client に送信する．

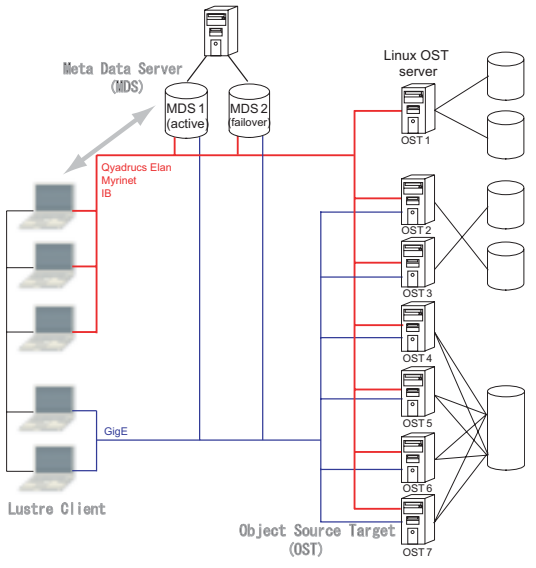

Fig. 1　Lustre の構成

2.2　Lustre におけるファイル参照手順

Lustre のサブシステムの相互関係を Fig. 2 に示す．

以下に Client がファイルを参照する際に Lustre が行う処理を述べる．

1. Client が MDS に対して参照したいファイルのメタデータを要求する．
2. MDS は求められたファイルに対応するメタデータを Client に送信する．
3. Client はそのメタデータを基に対応する OST へファイルを参照させるよう要求する．
4. OST はメタデータに対応するファイルをストレージから読み込み Client へ送信する．

Lustre では複数のストレージがあるが，以上の操作を Lustre が自動的に行ってくれるため，Lustre を使用している人はどのストレージにほしいファイルがあるかを考える必要がない．

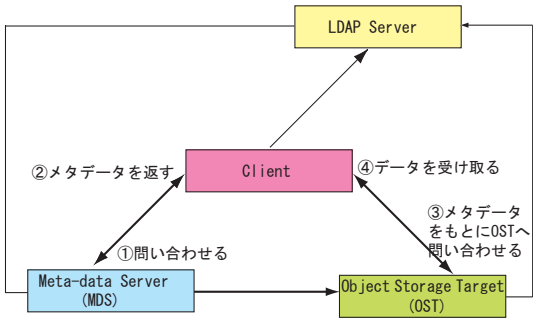

Fig. 2　Lustre のサブシステムの相互関係

2.3　ロック機能

Lustre ではファイルやディレクトリをロックする機構がある．これを用いることによってネットワークのボトルネックを減らし，全体的なデータ処理能力を向上させることができる．

例えば，2 つの Client から同時に同じ名前を持つファイルの作成処理が発生した場合，Lustre ではディレクトリをロックし排他制御を行うため，同一のファイル名が複数存在することはない．しかし，この機構を持たない他のファイルシステムではファイル名が重複する可能性がある．

2.4　障害時の対応

稼動中の MDS に障害が発生した場合，代理サーバへの引き継ぎ（フェイルオーバー）の仕組みを Fig. 3 に示す．

Client の要求に対して MDS からの応答が無いとき，代理サーバの情報を持っている LDAP サーバに問い合わせを行う．その後すぐに要求を直接代理サーバに行う．

OST が故障した場合はその故障した OST は使用できないため，次に新しいファイルを作成する要求が来た場合は自動的に故障した OST を避ける．

3　NFS と Lustre の比較

NFS と Lustre を比較した表を Table 1 に示す．

Table 1　NFS と Lustre の比較

	NFS	Lustre
ノード数	△	◎
容量	△	◎
耐故障性	×	○
I/O の処理能力	△	◎

NFS はマウントするノード数が多ければ多いほど NFS サーバにかける負担は大きくなる．それに比べ，Lustre はメタデータと実際のデータを分けているので，付加を分散させることができる．このようなことから Lustre は Client の数を 10000 まで構築することができ，I/O の処理能力も高い．

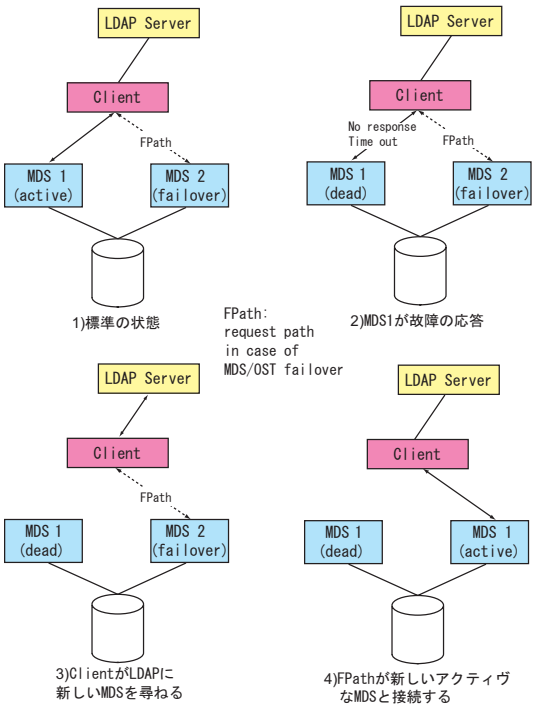

Fig. 3　Lustre におけるフェイルオーバーのメカニズム

また，NFS はサーバに障害が起きれば使用できなくなるが，Lustre の場合 MDS や OST に障害が起きてもそれを回避して動作し続けることができる．

よってノード数の多いクラスタで用いるファイルシステムとして Lustre が良いことが分かる．

4　まとめ

本稿では Lustre の概要を述べた．Lustre は大規模なクラスタで用いることができるファイルシステムであり，フェイルオーバーを行うことで耐故障性を向上させている．

また，メタデータと実際のファイルを分けられているのでシステムにかかる負荷を分散させ大規模なストレージ環境と高速な I/O の処理能力を実現している．

5　今後の課題

本稿では Lustre に関する調査を行った．今後は Lustre の構築を行い，性能に関する調査を行う予定である．

参考文献

1) Lustre, *http://www.lustre.org/*.