M-016

階層型 RAID を用いた大規模仮想ディスクの構築

Constructing a Large-Scale Virtual Disk using Hierarchical RAID

チャイ エリアント† 上原 稔† 森 秀樹†
Erianto Chai Minoru Uehara Hideki Mori

1 はじめに

PC 教室などの教育環境は数百台の PC から構成され、大規模ストレージを必要とする。しかし、実際の教育環境ではクライアント PC の HDD はほとんど利用されない。教育環境では信頼性や管理の容易さを重視するため、数 TB の HDD 容量を持つ高価なファイルサーバを導入することが多い。しかし、このようなファイルサーバはより安価なシステムを求める教育現場のニーズと乖離している。ここで、一例をあげる。60TB のファイルサーバを定価で見積もると 2.5 億円になるが、120GB の HDD を持つ 500 台の PC からなるシステムは HDD 単価 1 万として 500 万円で済む。そのコスト比は 1:50 になる。信頼性等は重要であるが、これほどコストが違うと異なる選択肢も考えられる。

本研究の目的は、遊休資源を活用するため、数百台 PC のハードディスクの空き容量を集めて、1 台の信頼性のある大規模ストレージを実現することである。ストレージのファイルサーバは 64 ビット Linux で構築され、Linux クライアントからは NFS (Network File System)、Windows クライアントからは CIFS (Common Internet File System)でアクセスされる。

2 関連研究あ

2.1 RAID

RAID[1][2]（Redundant Arrays of Inexpensive Disks、レイドと読む）とは、記憶すべきデータと障害回復のための冗長データを複数のハードディスクドライブに分散して格納することで、パフォーマンス（性能）とフォルトトレラント（耐障害）性を同時に確保するための技法である。

現在のエンタープライズ環境では従来よりも大容量化が進んだディスクドライブを多数使用するようになっているほか、複数のサーバによるストレージの共有も進んでいるため、1 つのディスク障害が多数のシステムに影響を与える危険性が高まっている。RAID 5 システムでは、オペレーターは時々間違った駆動中ドライブを引いて、2 つドライブが同時に失敗するようになってしまった。そこで、RAID 5 より信頼性が高い RAID 6 が採用される。

2.2 NBD

NBD[3]（Network Block Device）は、Linux カーネルの拡張の一つである。リモート・サーバが提供するブロックデバイスを、ネットワークを経由してローカルのブロックデバイスのように扱うことができる。NBD は NFS に良く似ている。NFS では、リモート・サーバが提供するファイルシステムを ネットワークを経由してローカルのファイルシステムのように扱うことができる。一方 NBD ではファイルシステムではなくブロックデバイスとして扱うことを可能にする。ブロックデバイスの上にファイルシステムを構築することはもちろん可能なので、その点では NFS よりも柔軟性があるものということができる。

2.3 ファイルシステム

SMB(Server Message Block)/CIFS は Windows で標準のネットワーク共有プロトコルである。Linux は Samba を用いて SMB/CIFS サーバとなることができる。

NFS[4]は UNIX でデファクトスタンダードな分散ファイルシステムである。Linux でも標準でサポートされている。また、Windows でも Service for UNIX(SFU)を用いて NFS を利用することができる。ただし、SFU を用いた運用は Samba を用いた運用ほど容易ではない。そのため、クライアント OS によって適切なファイルシステムのプロトコルを選択しなければならない。

XFS[5]は業界最先端の高性能ファイルシステムとして広く認識されて、システムクラッシュからの迅速な復旧や非常に大規模なディスクシステムのサポートを提供する。

3 システム構成

既存のファイルシステムの本質的な問題点はディスク容量を 100%利用できないことである。例えば、120GB ずつ集めて 60TB のストレージを構築しても、それぞれのファイルは 120GB を超えることはできない。また、1 つのディレクトリの中でファイルの合計サイズが 120GB を越えることもできない。また、仮想ファイルシステムは下位層のファイルシステムに依存するためファイルサイズの上限が 2GB に制約されることもある。

このように NFS をはじめとするファイルシステムレベルの分散ストレージでは空き容量を連結して 1 つのストレージにすることができない。そこで、本研究ではディスクレベルの分散ストレージを採用する。ディスクレベル分散ストレージでは仮想的なディスクを構築することでファイルシステムに依存しない運用が可能となる。

システム構成は図 1 に示す。仮想ディスクのディスク・サーバは必要とする領域を用意し、ファイルシステムが書き込みや読み込みができるように機能を提供する。ファイルシステムは複数のディスク・サーバにアクセスし、大容量ディスクの実現を可能にする。ディスクの性能を上げるために RAID6 を取り入れる。そして、ファイルシステムの上に NFS と Samba を使うことによって、実際にストレージとして利用できるようになる。

† 東洋大学大学院工学研究科情報システム専攻

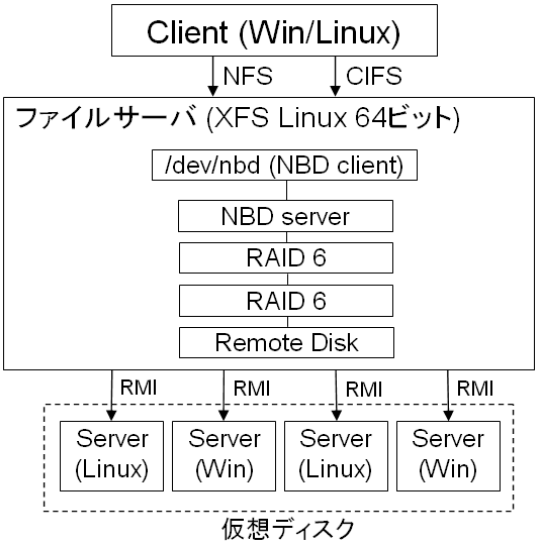

図 1　システム構成図

4　実装

85TB(512*170GB)仮想システムを作って、仮想ディスクプログラムを起動する。そして、512 ディスクを 32 グループ（1 グループ=16 ディスク）にして RAID66 にする。ファイルシステム側プログラムを起動し、ディスクを XFS でフォーマットする。Windows からアクセス出来るように Samba を起動する。図 2 はディスクドライブ・プロパティの図である。

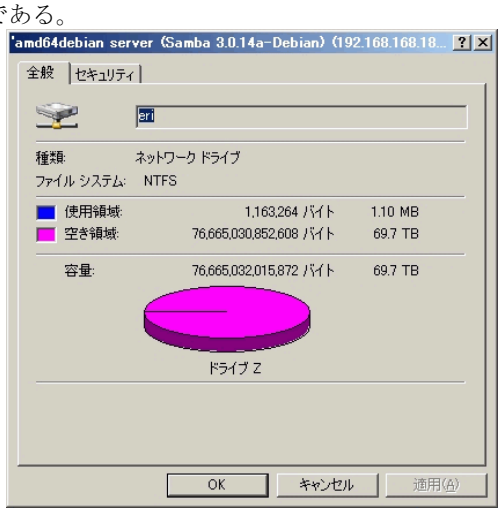

図 2　ディスクドライブのプロパティ

5　評価

実験では固定長ディスクと可変長ディスクを XFS でフォーマットし、フォーマット速度を比較する。可変長ディスクは必要とする分だけイメージファイルに書き込む。

表 1 は格容量に対して、固定長ディスクと可変長ディスクのフォーマット平均時間である。

表 1　フォーマット平均速度

容量	フォーマット平均時間（s）	
	固定長ディスク	可変長ディスク
1GB	12.33	1.33
2GB	20.67	1.33
5GB	51	1.33
10GB	103	1.33
20GB	202	1.33
50GB	561.33	3
100GB	1046.67	5
200GB	1669.33	10

図 3 は XFS で固定長ディスクと可変長ディスクのフォーマット平均時間を比較する。グラフから見ると、可変長ディスクの方がフォーマットする時間が速い。理由は可変長ディスクではディスクのシークする時間が速いためと考えられる。

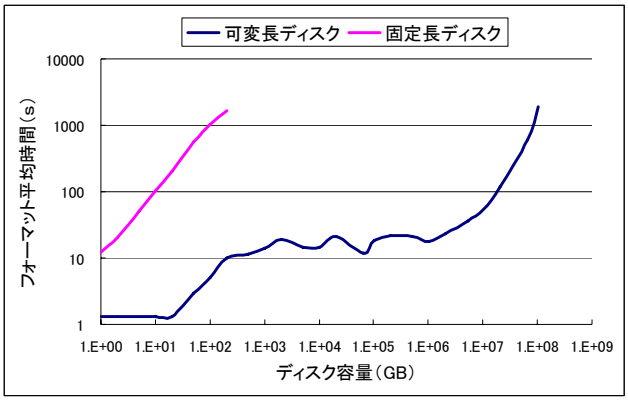

図 3　XFS で固定長ディスクと可変長ディスクの比較

6　まとめ

RAID66 で 69.7TB の大規模仮想ディスクを構築した。今後は、故障の種類、故障の数、故障ディスクなどにより信頼性を評価したり、故障時の性能を評価したりする。また、サーバの遠隔管理やボトルネックを解消する。

参考文献

[1] 宇野俊夫: 「ディスクアレイテクノロジ　RAID」, エーアイ出版株式会社, pp. 69-70 (July 2000)
[2] Intelligent RAID 6 Theory Overview and Implementation, http://download.intel.com/design/storage/papers/30812202.pdf
[3] IBM Linux Hint & Tips, http://www-06.ibm.com/jp/linux/developers/techinfo/nbd.pdf
[4] S. Shepler, et.al.: NFS version 4 Protocol, http://www.ietf.org/rfc/rfc3010.txt
[5] XFS 高性能ジャーナルファイルシステム, http://www.sgi.co.jp/projects/xfs/