B-008

RAIDシステム内蔵型NAS(1) -高速I/Oコマンドインタフェース-

Embedded NAS for RAID System (1) - High Performance I/O Command Interface -

山崎 康雄† 須藤 敦之† 坂口 明彦†
Yasuo Yamasaki Atsushi Sutoh Akihiko Sakaguchi

1．はじめに

情報システムの利用拡大により、システムが保持するデータ容量は膨大なものとなっている。その膨大なデータを管理するコストを削減するため、ストレージ統合が求められている。これまで、SAN(Storage Area Network)接続による管理の統合が進められてきたが、近年、LAN(Local Area Network)上に配置することが可能なNAS(Network Attached Storage)の利用も増大しつつある。より大規模なストレージ統合実現のため、NASのデータアクセス性能に対する要求はますます高まっている。

そこで、RAIDシステムにNASブレードを内蔵するRAIDシステム内蔵型NASにおいて、NASサーバOSとRAIDコントローラとが協調して動作するデータアクセス性能向上方法を検討した。とくにNASサーバOSとRAIDコントローラ間の高速I/Oインタフェースについての実装および評価を行った。

2．システム構成

RAIDシステム内蔵型NASは、RAIDシステム内にNASブレードと呼ぶ専用ボードを実装する。これにより、1台のRAIDシステムでSAN/NASの2つの機能を提供する。システム構成を図1に示す。

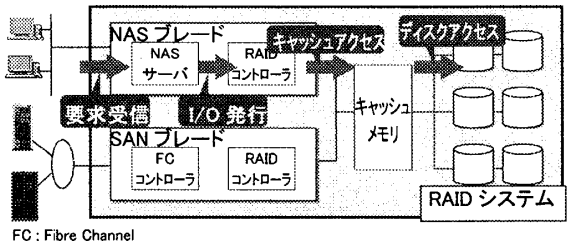

図1 RAIDシステム内蔵型NASシステム構成

RAIDシステムは、大容量のディスクドライブ、ディスクドライブ上のデータを一時的に配置するキャッシュメモリ、およびこれらの制御を行うRAIDコントローラを装備している。

また、NASブレードには、ネットワークと接続しI/O要求を処理するNASサーバとRAIDコントローラを搭載する。NASサーバはLinuxの改良版によって実装しており、NFSはカーネルスレッドにより実現した。

NFS要求のREAD処理は以下のようになる。

1) NASサーバがネットワークからのNFS要求を受信

2) NFSカーネルスレッドはNFS要求をファイルアクセス要求に変換してファイルシステムに渡し、その後ブロックI/Oドライバ、SCSI(Small Computer System Interface)ドライバを経てRAIDコントローラドライバにディスクI/O要求を発行する

3) RAIDコントローラドライバはRAIDコントローラにSCSI I/Oコマンドを発行し、その結果を発行元に返す

4) RAIDコントローラは受け取ったSCSI I/Oコマンドを解析し、キャッシュヒットミス判定を行い、必要であればディスクからデータを読み込みキャッシュを充填し、データをDMA(Direct Memory Access)転送しコマンドの結果(成功・エラー)をNASサーバに通知する

3．従来のI/Oコマンドインタフェース方式と課題

RAIDシステムとLinuxを用いたシステムに従来のI/Oコマンドインタフェース方式を適用する場合以下の2つの課題がある。

1) Linuxの割り込み処理オーバヘッドの削減

一般のSCSIディスクを用いたLinuxにおいてI/Oコマンドを発行した場合、コマンド完了の検出に割り込みを用いる。この割り込み処理のオーバヘッドおよび割り込み処理からの復帰処理のオーバヘッドの大きさは、低速なSCSIディスク装置の場合は問題とならないが、高速なRAIDシステムを接続した場合は問題となる。

2) RAIDコントローラのコマンド解析オーバヘッドおよびキャッシュ判定オーバヘッドの削減

RAIDコントローラはSCSIコマンドを受け取るとコマンドを解析し、キャッシュヒットミス判定を行う。これらの処理はコマンド数に比例して時間がかかるため大量のI/Oコマンドを生成する高性能NASにおいては処理時間を削減することが課題である。また、後者のキャッシュヒットミス判定処理はRAIDシステム全体で共有するメモリに格納されたキャッシュ管理情報をアクセスする必要があるため不必要なアクセスは極力抑える必要がある。

4．高速I/Oコマンドインタフェース方式

RAIDシステム内蔵型NASでは独自インタフェースが使えることに着目しRAIDコントローラとNASサーバOS間の高速I/Oコマンドインタフェースを設計した。

† (株) 日立製作所 中央研究所

1) マルチ LBA コマンド形式

I/O 対象のディスク上のブロック位置(LBA:Logical Block Address)とメモリ上のアドレスを複数指定できる専用の独自コマンド形式であるマルチ LBA コマンド形式を開発した(図 2)。複数のコマンドを一括して処理して RAID コントローラのコマンド解析処理のオーバヘッドを削減する。

また、ひとつのマルチ LBA コマンドに含まれる I/O 対象が RAID システム内のキャッシュメモリ単位に含まれると想定できる場合には、Linux がコマンドにフラグを付与するようにした。RAID コントローラはこのフラグを見てキャッシュヒットミス判定を一括化することができる。

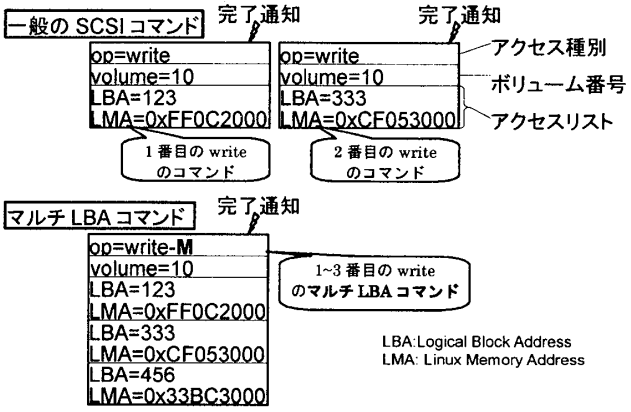

図 2　マルチ LBA コマンド形式

2) SCSI コマンドポーリング

Linux における割り込み処理オーバヘッドを削減するために、定期的にコマンドの完了をチェックする SCSI コマンド完了ポーリングを開発した。Linux は NFS 処理がきりのよいところでコマンドの完了をまとめてチェックすることで、割り込み処理のオーバヘッドをなくした(図 3)。また、次のコマンドの発行も同じタイミングで行うため、マルチ LBA コマンドの一括化効果を高めることができる。

ポーリングは次の 3 箇所で行う

(1)スケジュール時点でのポーリング

NFS 高負荷のケースでは、数 10 マイクロ秒単位で Linux 処理が切り替わる。これを適度に間引いてポーリングする。

(2)定期タイマでのポーリング

NFS 処理に直接関係ないプロセスがプロセッサを使用するケースでは、スケジュールの機会が減る。必ず 10 ミリ秒毎に発生するタイマ割り込みでポーリングする。

(3)アイドルスレッドでのポーリング

NFS 低負荷のケースでは、プロセッサはアイドル状態となる。このとき eager にポーリングを行って応答性を向上する。

上記の 3 箇所でポーリングすることでオーバヘッドを大幅に削減し、かつアクセスレイテンシが悪化しないように歯止めをかけた。

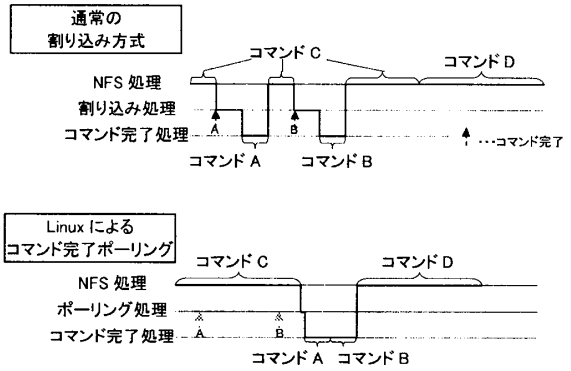

図 3　SCSI コマンドポーリング

5. 評価

コマンド数削減効果を調べるため、NAS サーバ OS の要求したディスクアクセスがいくつのコマンドとして発行されたかを調査した(図 4)。NAS サーバ OS のディスクアクセス要求数に対して，実際に発行したコマンドの個数が約 1／4 のみだった。これはコマンド発行ポーリングとマルチ LBA コマンドの効果である。また、ディスクアクセス回数すべてについて割り込みが削減された。これはコマンド完了ポーリングの効果である。

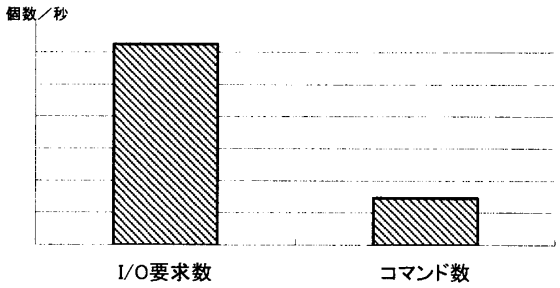

図 4　コマンド数削減効果

6. おわりに

専用の高速 I/O コマンドインタフェース方式により、コマンド数を 1 ／ 4 に削減し、割り込みオーバヘッドを排除した。

参考文献

[1] 坂口明彦，山崎康雄，須藤敦之：RAID システム内蔵型 NAS(4) –高信頼内部通信機能–，情報処理学会第 66 回全国大会，5D-6(2004)

[2] 須藤敦之，坂口明彦，山崎康雄：RAID システム内蔵型 NAS(6) –キャッシュメモリ制御–，情報処理学会第 66 回全国大会，5D-6(2004)