RAIDシステム内蔵型NAS(4) －高信頼内部通信機能－

(High-Reliability Internal Communication for Embedded NAS)

(株)日立製作所 中央研究所[1] 坂口 明彦 山崎 康雄 須藤 敦之[2]

1 はじめに

近年企業において大容量ストレージに対する要求が高まっており、その集中管理による運用管理コストの削減が重要な課題となってきた。この問題を解決するために、複数のストレージを FC(Fiber Channel)ネットワークで接続したSAN(Storage Area Network)によるストレージのコンソリデーションが進められてきた。しかしながら、全てのサーバがFCネットワークを介してSAN接続されているとは限らないため、LAN(Local Area Network)を介して利用できる NAS(Network Attached Storage)によるストレージコンソリデーションを導入する企業も増えてきた。エンタープライズ用途でのNASにおいては、従来の RAID システムと同等の高信頼性、特に単一障害でのシステム無停止が重要な課題となっている。

本研究では RAID システム内蔵型 NAS ブレードにおいて、NAS ブレード間の相互監視を二重化された内部ネットワークを用いて行う高信頼内部通信機能の開発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の構成を図1に示す。通常のRAIDシステムにFiber Channel で接続するSANブレードのスロットに IP ネットワークで接続するNAS ブレードを実装する。これにより、1 台のサブシステムでSAN/NASの両方の機能を提供することが可能となる。NAS ブレードからは内部ネットワークを介して、共有メモリ・キャッシュメモリ・ディスク装置と接続されており、RAIDコントローラがそれらの制御を行っている。IPネットワークで外部と接続されており、外部からのI/O要求はNASサーバの処理の後、RAIDコントローラを経由してディスクアクセスが行われる。

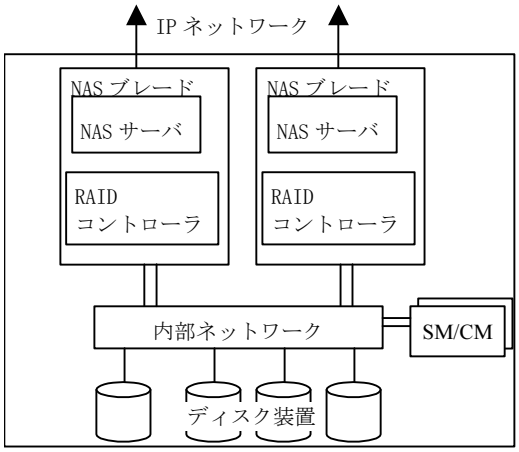

図1 RAIDシステム内蔵型NASの構成

内部ネットワーク及び共有メモリ(SM)、キャッシュメモリ(CM)はそれぞれ二重化されており、単一障害時には代替パスを介して正常動作が保証される。

3. 内部通信仕様

本研究における内部通信の主な使用目的はNASブレード間のハートビート通信である。NAS ブレードは相互監視対象のNASブレードと定期的にハートビート通信を行い、互いに生死確認を行う。ハートビート通信の遮断により、他方のNASブレードのダウンをいち早く検出し、検出した場合にはサービスの引継ぎを行う。このため、ハートビート通信の特徴は以下の通りである。

(1)通信の高信頼性が重要

まず第一に、高信頼性である。NAS ブレード間のハートビート通信では、通信が途絶えた場合に相手のNASブレードがダウンしたと判定し、サービスの引継ぎを行う。ハートビート通信の遮断を誤検出した場合、相互監視を行うNASブレード同士が互いに相手のサービスを引き継ぎ二重サービスとなるのを防ぐため、引き継ぐ場合には他方のNASブレード

1 Hitachi, Ltd., Central Research Laboratory
2 Akihiko Sakaguchi, Yasuo Yamasaki, Atsushi Sutou

を確実に停止させる。無駄なサービスの引継ぎ処理を防ぐためハートビート通信の高信頼化が重要となる。

(2)定期的に通信が発生

第二にハートビート通信はNASブレード間の相互監視のための通信であり、定期的に発行される。発行間隔は1秒から数秒程度の比較的頻度の少ない場合が一般的である。

(3)通信量は一定

第三にハートビート通信は通信が継続されていることを確認することで互いのNASブレードが動作していることを確認することを目的とする。したがって、一回の通信における通信量は一定であり、更に通信量は比較的少ない。

4. 高信頼内部通信の実装方式

前述のハートビート通信の特徴を踏まえて、内部ネットワークを介した共有メモリおよびキャッシュメモリを用いた高信頼内部通信を開発した。

共有メモリ/キャッシュメモリとその通信路は二重化されており、単一障害時には代替パスを介して正常動作が保証される。そのため本内部ネットワークを用いることでハートビート通信の高信頼性が確保される。また、通信頻度および通信量が少ないハートビート通信に用いる場合には、通信負荷によるI/O性能への影響は問題とならない。

高信頼内部通信の基本構成は図2の通りである。キャッシュメモリ上にNASブレード間でデータ転送を行うためのメッセージキューを用意し、送信側NASブレードが書き込んだデータを受信側NASブレードが読み込むことで通信可能となる。メッセージキューの制御は、共有メモリ上にキュー制御情報を配置することで可能となる。共有メモリアクセスはレイテンシが短いがスループットは低く、逆にキャッシュメモリアクセスはレイテンシは長いが、スループットは高いという特徴がある。この特徴を生かし、データ量の少ない制御情報は共有メモリ上に配し、メッセージキューはキャッシュメモリ上に配する。送信NASブレードがメッセージキューにデータを書き込んだ後、キュー制御情報を更新する。受信NASブレードはそれを検知するとメッセージキューからデータを読み出す。

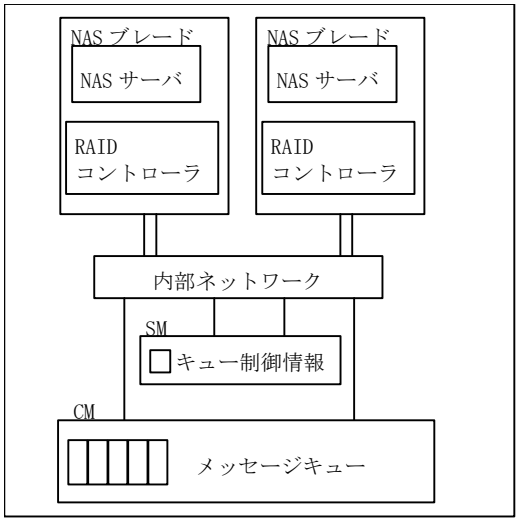

図2 高信頼内部通信基本構成

NASサーバ上でNASブレードの相互監視を行うHAソフトを稼動させ、ハートビート通信を行う。一般のHAソフトが利用できるように、内部ネットワーク上に通常のTCP/IPプロトコルを実装し、上位ソフトには通常のLAN経由で通信を行っているように見せかける。共有メモリ/キャッシュメモリへのデータ書き込み、およびキュー制御情報の更新の検知はRAIDコントローラが行う。

5. 評価

本内部通信を実装したプロトタイプを用いて本方式の評価を行い、要求仕様を満たしていることを確認した。評価項目は二点である。

・内部ネットワーク/共有メモリ/キャッシュメモリの単一障害時にハートビート通信が継続すること。
・ハートビート通信時にI/O性能の劣化が発生しないこと。

6. おわりに

二重化された内部ネットワークで接続されたキャッシュメモリを通信バッファとすることで、障害に強い高信頼通信の開発を行った。NASブレード間のハートビート通信に適用することで、ネットワークの障害や高負荷によるNASブレードの生死確認の誤検出を防ぐことができた。