RAID システム内蔵型 NAS(3) —障害処理機能—

Embedded NAS for RAID System (3) – Design of Processing against Failures –

宮田賢一† Kenichi Miyata
橋本顕義† Akiyoshi Hashimoto
薗田浩二† Koji Sonoda
中谷洋司† Yoji Nakatani
中野隆裕† Takahiro Nakano
原純一† Junichi Hara

1. はじめに

本研究では大規模 RAID システムに NAS を内蔵したシステムにおける障害処理機能について検討する。

大規模 NAS では、24 時間 365 日無停止である高可用性が求められ、システムで障害が発生した場合でも利用者に対してファイル共有サービスを透過的に継続するとともに、システム内部では発生した障害を確実に検知し、障害の原因を解明してより高可用なシステムを構築するための情報を取得できるような手段が必要となる。

本研究では内蔵型 NAS という特徴を活かし、障害の検出・対処・情報取得という処理を、装置内の各部が密に連携して行えるようなシステムの検討を行った。

2. 障害処理の課題

本節では大規模なファイルサーバが満たすべき機能の課題を、障害検出と障害への対処という二つの観点から示す。

2.1 障害検出

障害検出に関して次の二つの課題がある。

(1) 障害の誤検出防止
(2) 障害ノードの確実な停止

クラスタを管理するソフトウェアが障害を検出した場合、後述するようにファイル共有サービスをクラスタ内で引き継ぐためにサービスが一時的に保留状態になる。したがって障害を誤検出すると不当にサービスを保留にすることになり、ユーザへのサービス提供の質が低下することになる。

さらに、それ以上動作させるとハードウェアやファイル共有のためのファイルシステムを破壊する可能性のある緊急度の高いハードウェア障害やソフトウェアの致命的な障害の場合は、迅速かつ確実にノードをダウンさせる必要がある。

2.2 障害への対処

障害への対処として次の二つの課題がある。

(1) サービス継続
(2) 障害情報取得

ファイル共有サービスにとっての 24 時間 365 日無停止である条件は、ファイル共有を利用するクライアントやアプリケーションが障害発生に気づかないことである。そのためには、障害が発生しても、ユーザへのサービスを継続できるような手段が必要である。

障害は構成部品の自然劣化により発生する場合のほかに、ハードウェアやソフトウェアのバグにより特定の条件を満たしたときに発生する場合がある。そこで障害情報を取得できる機能を持つことで、障害情報をシステム保守員が取得して開発元にフィードバックすることが可能になり、より高可用なシステムとして再構成できる。

3. 解決手段

2 節で述べた課題を解決するためのシステム構成を設計し、その構成の上で解決手段を検討・実装した。

3.1 システム構成

本 RAID システム内蔵型 NAS の障害処理という観点によるシステム構成を図 1に示す。

一つのノードにはファイル共有サービスを提供する NAS プロセッサと、RAID 装置とデータ入出力を行うための RAID コントローラが搭載される。このノードを 2 個束ねてクラスタとして構成する。NAS サービスを提供するノードが RAID 装置に内蔵されることにより、ノードは RAID 装置内の資源を活用することができる。

図 1のノード内監視・ノード間監視・RAID 装置内通信路については以下の節で詳細を示す。

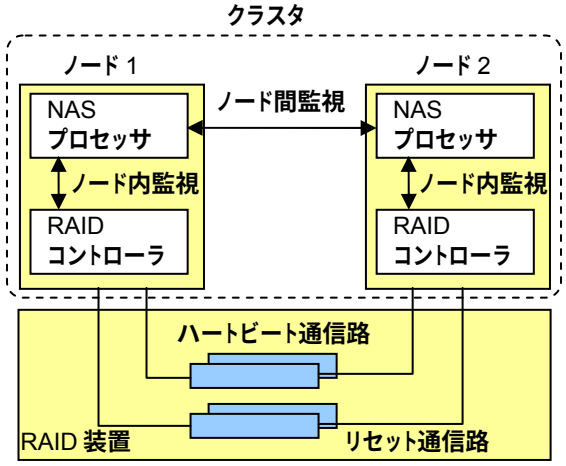

図1 監視系統と冗長通信路

3.2 障害検出

障害検出のために、ノード内およびクラスタ内ノード間で相互に監視を行うことで、ノードで障害が発生

† (株)日立製作所システム開発研究所
Systems Development Laboratory, Hitachi, Ltd.

したことを確実に検出する。

3.2.1 ノード内監視

ノード内の NAS プロセッサと RAID コントローラとの間で相互に監視を行う。これにより以下のような確実かつ速やかな検出が可能になる。

- NAS プロセッサがダウンしている場合
  ノード間監視でノードダウンが検出されるのを待たずに NAS プロセッサがダウンしていることを直接検出できる。
- RAID コントローラがダウンしている場合
  一般的に数分の時間が必要な RAID 装置への I/O 要求のタイムアウトを待たずに、速やかに RAID コントローラがダウンしていることを検出できる。

3.2.2 ノード間監視

クラスタを構成している他のノードの死活を監視するために、ハートビートと呼ぶパケットをノード間で交換することが一般的に行われている。あるノードが相手ノードからのハートビートの受信に失敗した場合、相手ノードがダウンしていると判定し、その結果ノード間のフェイルオーバが発生する(3.3.1節参照)。ハートビート断を誤検出することは、不必要なフェイルオーバを発生させ、ユーザへのサービスが一時的に保留状態となりサービスの質が低下することになる。

そこでハートビートの通信路には確実にハートビートの交換ができるような高信頼性が求められる。RAID システム内蔵型 NAS システムでは、ハートビートの通信路として RAID 装置内の専用通信路を利用し、これをソケットインターフェースとして Linux に見せるようにした(図 2)。この通信路は二重化されており、片方の通信路が利用不可になっても他の通信路に自動的に切り替わる。また専用通信路へのパスそのものが障害を起こした場合に備えて、ファイル共有用の LAN ポートをハートビートの代替通信路として用いることでさらに可用性を高める。

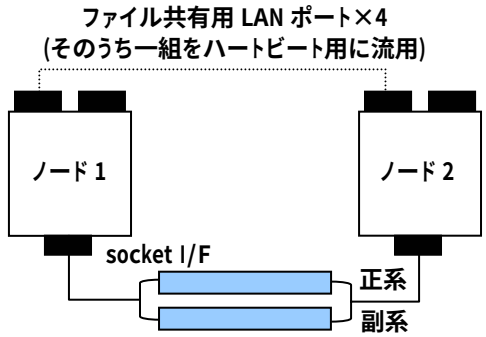

図2 冗長化された RAID 装置内専用通信路

3.3 障害への対処

3.3.1 ノード間フェイルオーバ

障害が発生してもユーザへのファイル共有サービスを継続して行うために、ノード間フェイルオーバを行う。ノード間のフェイルオーバにより、障害が発生したノードで提供していたサービスを、他のノードで再開することができる。

フェイルオーバを行う際、ディスクへの I/O 要求が両ノードから発行されてデータ破壊を起こすのを防止するために、フェイルオーバ元のノードを確実にリセットしなければならない[1]。

このリセット要求をノードをまたがって発行するために、ハートビート用通信路と同じように RAID 装置内の冗長化通信路を用いる。これによりリセット要求を確実に相手ノードに届けることができる。

3.3.2 メモリダンプ

3.3.1節で述べたリセット処理の一連の流れの中で、NAS プロセッサが管理する全メモリ空間を RAID ディスク上にダンプする処理を行う(図 3)。

ハートビート断を検出したノードが相手ノードのリセットを行った場合、NAS プロセッサには割り込みが上げられ、NAS プロセッサが再起動した直後に NAS プロセッサが管理している全メモリ空間をダンプする。

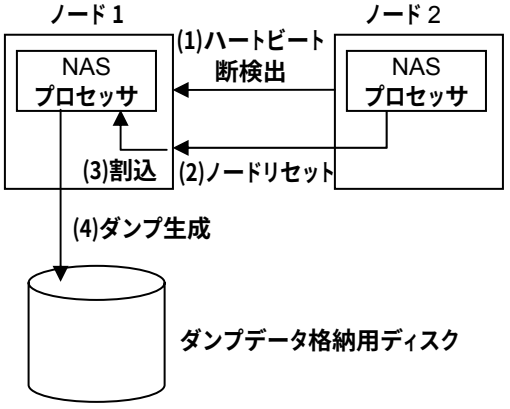

図3ダンプ生成の手順

4. まとめ

内蔵型 NAS という特長を活かして、システム内部で発生する障害の検出･対処･情報取得を確実に行うための設計を行った。

RAID 内蔵型 NAS システムの中で、RAID 装置が持つ冗長なノード間通信路を利用することで、確実なノード間監視と障害検出を行い、また確実にノードリセット処理を行うことを実現できた。

---

[1] ノード 1 上にディスクに未反映のクライアントからの I/O 要求が待っているときにノード 2 へのフェイルオーバが発生したとする。クライアントはノード 1 からの応答が返ってこないため再度ノード 2 に I/O 要求を発行する。この後両ノードから同じディスク反映要求が発生するとデータを破壊する可能性がある。