

大容量で高信頼な分散共用ディスク

明石孝祐[†]，細坪久賢[‡]，竹内理恵[‡](株)富士通[†]，(株)富士通北陸システムズ[‡]

akashi@rp.open.cs.fujitsu.co.jp, {hoso, rie}@yah.so.fjh.se.fujitsu.co.jp

1 D-4

1 はじめに

Grandpower7000 クラスタシステム[1]は従来より SCSI マルチイニシエータによるノード間のディスク共用(以下 SCSI 共用)を提供してきた。SCSI 共用では、1 チャンネルあたりの接続台数が少ないこと、また伝送速度が低いことから今日のような大規模なクラスタシステムを構築する上でその要求に十分に対応できなくなっている。

そこで今回、シェアドナッシングのアーキテクチャ上でも、ソフトウェアによりディスク共用を可能とする技術(分散共用ディスク、以下 DSD)を開発した。ノード間インタコネクトとして AP-Net を採用し、ノード間メモリ転送機能を使った効率的な実装を行なうとともに、通信路およびノード、ディスクなどのハードウェア障害に対し二重化構成/ミラーリングを組み合わせることでデータの可用性を高めた。

2 DSD の概要

DSD は特定ノード上のディスク装置に 1:1 に対応付けられた仮想ディスクデバイスを提供する。通常のデバイス特殊ファイルと同様に扱うことができ、全ノードで同一のファイル名(/dev/rdsdsk/dsdXX)を持つのでディスクがどのノードに属するか意識せずアプリケーションを作成できる。

また、論理ボリューム機能(以下 LVCF)を DSD ディスクに適用することで、ノードをまたがったディスクのミラーリングやストライピングを可能にしている。

ハードウェア上は AP-Net により、ローカルにディスク装置を備えたノードが相互接続されていることだけを前提としている。SCSI 共用の場合とは異なり、ア

クセス経路上の AP-Net 通信チャネルおよびサーバノード、またディスク装置が故障すればリモートアクセスはできなくなる。そこで AP-Net 故障に対しては DSD 自体が自動的にカレントチャネルを切り替える機能を備えた。ノードおよびディスク装置に対しては LVCF によりミラーリングを行うことでデータ可用性を保証する。以上の構成により DSD 環境下でも論理ボリュームでアクセスする限りアプリケーションから単一故障を隠蔽することができる。

3 DSD の実現技法

3.1 基本構成

リモートアクセスを実現するため、DSD では IO シッピングと呼ばれている手法[3]を採用した。すなわち DSD は UNIX のデバイスドライバとして実装されており、ドライバがノード内で提供する open や strategy サービスをクラスタ内全ノードへ拡張するものであるといえる(図 1)。この手法はドライバがドライバを呼び出す方式であり既存ドライバの改造なく DSD を実現できる。

DSD ドライバ…DSD プロトコルを実行しドライバセマンティクスを提供する部と AP-Net 通信制御部とから成る。

DSD デーモン…サーバ側で HDSK ドライバのエントリを呼ぶ出すときのコンテキストとなる。

DSD モニタ…DSD ドライバのドライバの初期化およびハード障害に対して資源管理機構と連携[2]し構成変更をドライバに指示するデーモン。

HDSK ドライバ…ハードディスクドライバ。

LVCF ドライバ…仮想ボリュームドライバ。ディスクをミラー、ストライプ、コンカチネートする機能を提供する。

^{*}Scalable and High-Available Distributed Shared Disks

[†]Takahiro Akashi, Fujitsu Limited

[‡]Hisayoshi Hosotsubo, Rie Takeuchi,
Fujitsu Hokuriku Systems Limited

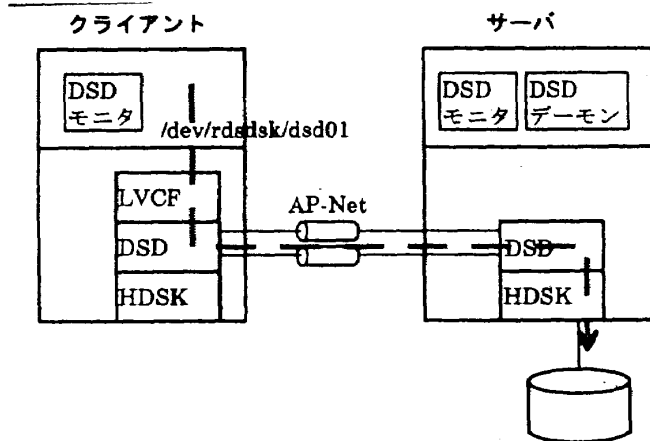


図1 DSD の基本構成

3.2 DSD プロトコルと通信制御

DSDはソケットやTLIなど汎用APIは使わずAP-Netのメッセージ送信(SEND)とノード間メモリ転送(PUT)を直接利用する。

a) ユーザ空間とのDMA

メッセージベースの実装ではデータサイズが大きくなるに従い、データコピーによるCPUオーバヘッドが常に問題になる。PUT機能はノード間のメモリ転送をハードウェアが実行するので、これをプロトコルに組み込むことでコピーの回数を減らすことができる。

read処理の場合のプロトコルを図2に示す。readデータはディスクドライバによりサーバのバッファに一度コピーされた後、ユーザ空間へ直接コピーされる。

b) メッセージの保証

ハードの提供するSENDではメッセージの送達保証がないため、メッセージ脱送を考慮したプロトコルが必要でリカバリ処理が複雑になる。そこでメッセージに番号を付加しメッセージの保証機能を実装した。更にAP-Net論理チャンネルを二本持っている場合、カレントチャンネルで故障が発生した際、受信されていないメッセージを他チャンネルに転送する機構を持たせている。

また、ウィンドウ数によるフロー制御も組み込んでいるが、メッセージを固定長(128B以下)とすることでCPUオーバヘッドの増加を軽減している。

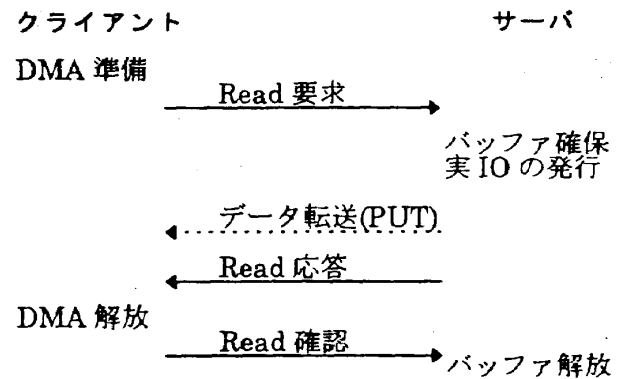


図2 read プロトコル

3.3 資源管理機構との連携

AP-Netの故障時に論理チャンネル切り替えを制御するのはDSDモニタであるが、DFSDモニタは、AP-Netばかりでなく他ノードや他DSDモニタの状態にも依存した制御を行っている。そこで資源管理機構の枠組みに従いDSDモニタをAP-Net資源、ノード資源、DSDモニタ資源のイベントレシーバとして実装した。統一されたインターフェースで処理することができ実装を簡単にする事ができた。

4 おわりに

本稿は高速インターコネクトAP-Net上に実現されたディスク共用技術DSDの概要を紹介し、実装上の問題をいくつか議論した。今後は性能/スケーラビリティの検証を行うとともに、ノード間にまたがるソフトRAID構成への応用や、ハードディスク以外のデバイスへの拡張なども検討していきたい。

参考文献

- [1] 富川ほか, "Granpower7000 クラスシステムの設計思想と新技術", 第56回情報処理全国大会論文集, 1D-01, 1998
- [2] 阿部ほか, "高信頼性を実現する資源管理機構とイベント管理機構", 第56回情報処理全国大会論文集, 1D-02, 1998
- [3] G.F. Pfister, "In search of Clusters: The Coming Battle in Lowly Parallel Computing", Prentice Hall, 1995