ＰＣサーバ向け「ハイブリッドＲＡＩＤ」の開発（２）
～ソフトウエアアーキテクチャとディレイドパリティ生成方式～

2Ｆ－8

兼田 泰典，大枝 高，松並 直人，荒川 敬史，八木沢 育哉
（株）日立製作所　システム開発研究所

1　はじめに

近年のコンピュータ市場では，PCの低価格化／高性能化が急激に進行している。この動向をふまえ，ソフトウエア制御方式をベースとし，パリティ生成等を行うアクセラレータハードを併用して，高性能・高信頼・低価格を実現したPCサーバ向け「ハイブリッドRAID」を開発した。

PCサーバのCPU性能は，２～３年で倍の速度で向上している。これに対し，磁気ディスク装置単体の性能は，アクセス速度を倍にするのに10年かかっており，今後さらにCPUとファイル系の性能差は顕著化する一方である。PCサーバにおけるファイル系のボトルネックを解消するために，磁気ディスク装置をｎ台並列に動作させるディスクアレイがPCサーバに標準で搭載され始めている。ディスクアレイは，磁気ディスク装置をｎ台並列に動作させることでｎ倍の性能向上が見込まれるが，信頼性は磁気ディスク装置単体に比べ１／ｎになる。そこで，ディスクアレイにパリティによる冗長性を持たせたRAID5[1]が多く用いられている。

2　ＲＡＩＤ５の課題

RAID5は冗長構成により信頼性を向上した反面，ライト時の性能が著しく低下するという問題がある[2]。RAID5では，データの書き込みとともに，次に示すパリティの更新処理も行わなければならない。

$$P_{new} = D_{new} \text{ XOR } D_{old} \text{ XOR } P_{old}$$

D_{old}：旧データ，D_{new}：新データ，

P_{old}：旧パリティ，P_{new}：新パリティ

このため１つのライト要求は次の４つのディスクアクセス要求に分けて実行される。

1）D_{old}のリード
2）P_{old}のリード
3）D_{new}のライト
4）P_{new}のライト

このようにライト処理では，従来ディスクを単体使用していた場合に比べ，４倍のディスクアクセスが必要になり，複数ディスクの並列効果が得られない。このライトペナルティを低減することがRAID5高速化の鍵になる。

3　ディレイドパリティ生成方式

ライトペナルティは，ライト時に発生するディスクアクセス回数を削減することで低減できる。この点に関し，

（1）不揮発性キャッシュ＋遅延書き込み
（2）書き込みデータの動的配置［2］
（3）パリティのログ化［3］

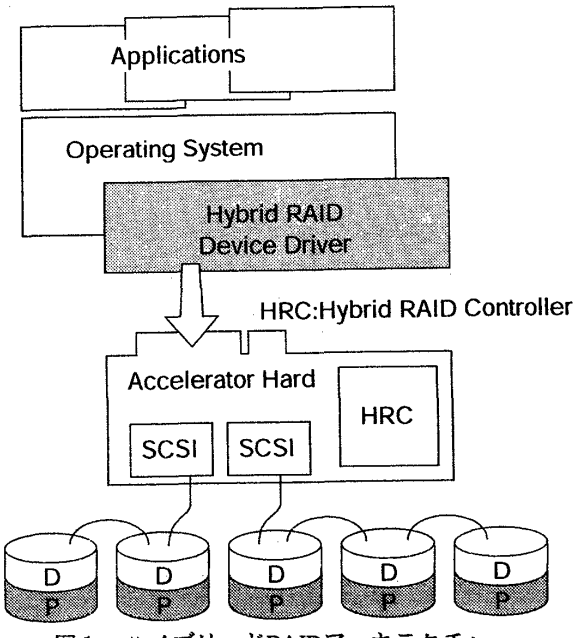

図１　ハイブリッドRAIDアーキテクチャ

Development of "Hybrid RAID" for PC Servers (2)
-Software architecture and delayed parity generation method
Yasunori Kaneda, Takashi Oeda, Naoto Matsunami,
Hiroshi Arakawa, Ikuya Yagisawa
HITACHI, Ltd. Systems Development Laboratory

など種々の方式が提案されている。不揮発性キャッシュはコストが高く低価格なPCサーバには適さない。また，動的配置やパリティのログ化は制御アルゴリズムが複雑でプロセッサ負荷が重く，ソフトウエア制御方式には適さない。

今回，揮発性であるPCサーバの主記憶をキャッシュとして利用し，パリティのみを遅延書き込みする「ディレイドパリティ生成方式」を考案した。ライト要求は以下の手順で処理する。

1）キャッシュを検索。
　旧データ（D_{old}）の有無を調査。
2）旧データ（D_{old}）が無い場合，ディスクから旧データ（D_{old}）を読み出す（1io）。
3）差分データ（D_{diff}）を生成。

$$D_{diff} = D_{new} \text{ XOR } D_{old}$$

4）差分データ（D_{diff}）をキャッシュに保持。
5）新データ（D_{new}）の書き込み（1io）。

以上で，ライト要求に同期した処理は終了する。新データは必ずライト要求に同期して実行するが，旧データの読み込みは，キャッシュでヒットすれば削減可能である。リードモディファイライト処理に着目すれば，旧データのキャッシュヒット率は十分高く期待でき，ライト要求に同期したディスクアクセス回数を１回に削減できる。

パリティの更新処理は，次の手順でライト要求とは非同期にまとめて実行する。

6）旧パリティ（P_{old}）を読み出す（1io）。
7）パリティを生成。$P_{new} = D_{diff} \text{ XOR } P_{old}$
8）新パリティ（P_{new}）の書き込み（1io）。
9）新パリティ（P_{new}）をキャッシュに保持。

ライトの局所性が高ければ，パリティ更新に伴うディスクアクセス回数を削減することができる。さらに，パリティを効率よく更新するために，データ領域とパリティ領域を完全に分離したパリティの配置方式を採用している。

4　ソフトウエアアーキテクチャ

本制御ソフトウエアはOSのデバイスドライバとして実装し，RAID制御機能（データの分散集合），キャッシュ管理機能，ハードウエア制御機能（ディスク制御，アクセラレータ制御）を設けた。デバイスドライバとして実装することで主記憶上のキャッシュを実現できた。

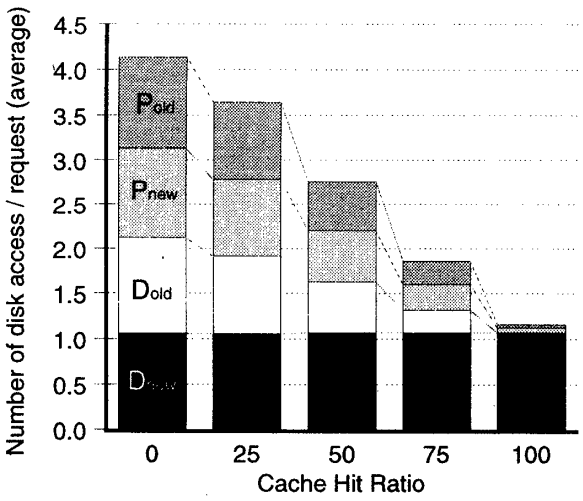

図２　１ライト要求当たりのディスクアクセス回数

5　まとめ

PCサーバの主記憶をキャッシュとして利用し，パリティのみを遅延書き込みすることで，RAID5に特有なライトペナルティを解消するディレイドパリティ生成方式を採用した。本方式の採用により，低価格化の実現と，従来１ライト要求あたり4.15回であったディスクアクセス回数をキャッシュのヒット100%時に最小1.12回にまで削減することができた。

参考文献

[1]David A.Patterson, Garth Gibson, and Randy H.Karz : "A Case for Redundant Arrays of Inexpensive Disks (RAID)", Computer Science Division Department of Electrical Engineering and Computer Science, University of California Berkelay.

[2]茂木和彦, 喜連川優 : "ストライプの動的再構成を伴うRAID5型ディスクアレイにおけるアクセスローカリティが存在する場合の更新処理の性能評価", SWOPP琉球'94電子情報通信学会計算機アーキテクチャ研究会, 電子情報通信学会技術研究報告 ARC 107-25, 1994.7

[3]喜連川優 : "ディスクアレイの技術動向", 日本応用磁気学会誌, Vol. 18, No.4, 1994.