ＰＣサーバ向け「ハイブリッドＲＡＩＤ」の開発（4）
2Ｆ－１０　～CPU負荷率低減の方法～

松並 直人，大枝 高，兼田 泰典，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、CPUの高性能化や本格的なマルチタスクOSの登場により、PCサーバの業務への適用が進行し、そのディスクサブシステムとしてトランザクション処理に適したRAID5[1]が一般的に用いられている。

ホストCPUですべてのRAID5の制御を行う、従来の「ソフトウェアRAID」は専用ハードが不要なので低価格であり、さらに、高速なホストCPUを利用するので限界性能が高いという利点がある。一方、(a)CPU負荷率が上昇する、(b)RAIDからのOSのブートアップができない、(c)ディスク活線挿抜ができない等、の欠点がある。

そこで我々は、ソフトウェアRAIDの長所を生かしながら上記の課題を解決すべく、ホストCPUでRAIDのデータ分散・集合制御等を行うソフトウエア(デバイスドライバ)と、パリティ生成・OSブートアップ・ディスク活線挿抜等を行うアクセラレータハードを統合して、高性能・高信頼・低価格を実現するPCサーバ向け「ハイブリッドRAID」を開発した。

本報告ではアクセラレータハードでパリティ生成を行いCPU負荷率を低減する方法について述べる。

2. パリティ生成方式

(1) ソフトウェアパリティ生成方式

ソフトウェアRAIDは、CPUによる排他的論理和(XOR)演算でパリティを生成する。以下「ソフトウェアパリティ生成方式」と称する。ソフトウェアパリティ生成方式には次の２点の課題がある。

(a) CPU負荷率の上昇

パリティ生成処理は、ライト時のRAID制御時間全体の約40%以上を占め、CPU負荷率を上昇させる主原因である。これは、CPUは100MIPS超の性能を有し、その最大転送速度は533MB/s(64Bit 66MHz)と高速であるが、パリティ生成速度は低速な主記憶DRAMのデータ転送速度で決定されるためである。主記憶の転送速度は約120MB/sであり(64Bit 高速ページモードDRAM)、CPUは潜在的な転送能力の約1/4しか発揮できない。

(b) CPUのキャッシュヒット率の低下

パリティ生成に必要なデータは、一度CPUの主記憶1次/2次キャッシュに転送された後、CPUに転送される。このデータはパリティ生成時以外には使用されないのでキャッシュヒットは期待できない無効データである。よって、キャッシュ上の有効データはこれらの無効なデータで上書きされてしまい、CPUのキャッシュヒット率が低下する。

(2) DMAパリティ生成方式

ハイブリッドRAIDは、ソフトウェアパリティ生成

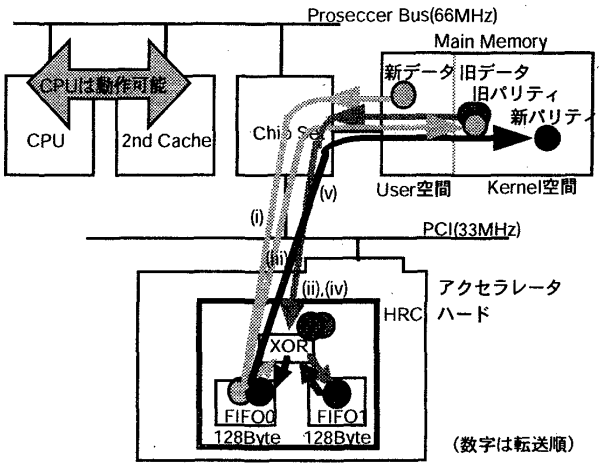

図1　DMAパリティ生成方式

Development of "Hybrid RAID" for PC Servers (4)
～Method of Reducing CPU Load ～
Naoto Matsunami, Takashi Oeda, Yasunori Kaneda,
Hiroshi Arakawa, Ikuya Yagisawa
HITACHI, Ltd. Systems Development Laboratory

RAID: Redundant Arrays of Inexpensive Disks
DMA: Direct Memory Access
PCI: Peripheral Components Interconnect
IOPS: Input/Output Per Second

方式の課題を解決するため、PCI直結アクセラレータボード上に、133MB/sの高速DMA転送でパリティ生成を行う「DMAパリティ生成方式」を実装したLSI "HRC (Hybrid RAID Controller)"を搭載する。ハイブリッドRAIDは、主記憶にディスクキャッシュ(以下単にキャッシュ)を構築し、ディレイドパリティ生成方式と称するパリティ遅延一括更新制御でライト処理の高速化を行っている[2]。HRCは、同方式に対応する以下3つのDMAモードを有する。

[DMA Mode]

(a) Mode1: メモリ間コピーモード
ユーザ空間のデータをカーネル空間のディレイドパリティキャッシュにコピー(又はその逆)するモード。リード時に使用する。

(b) Mode2: XOR演算モード
主記憶上の最大4項のデータのXOR演算を実行するモード。ライト時にパリティグループの全データが揃った時に使用する。

(c) Mode3: XOR演算+メモリ間コピーモード
(a)メモリ間コピーと(b)XOR演算を同時に実行するモード。ライト時に旧データ、新データ、旧パリティから新パリティを生成し、同時に新データをキャッシュにコピーする時に使用する。

図1はMode3 DMAパリティ生成方式の動作図である。処理の手順は次の通りである。

(i)主記憶上の新データをHRCのFIFO0にリード。
(ii)主記憶上の旧データとFIFO0上の新データのXOR演算を実行し差分データをFIFO1に格納。
(iii)HRCのFIFO0上の新データを主記憶にライト(新データのメモリ間コピー)。
(iv)主記憶上の旧パリティとFIFO1上の差分データのXOR演算を実行し結果をFIFO0に格納(新パリティの生成)。
(v)FIFO0上の新パリティを主記憶にライト。

以上のパリティ生成処理の間、ホストCPUは一切関与する必要がないので、本方式はCPU負荷率を低減できる。また、DMA制御専用のMPUや、パリティ生成作業用バッファメモリも不要なので、本方式は低コストで実現できる。

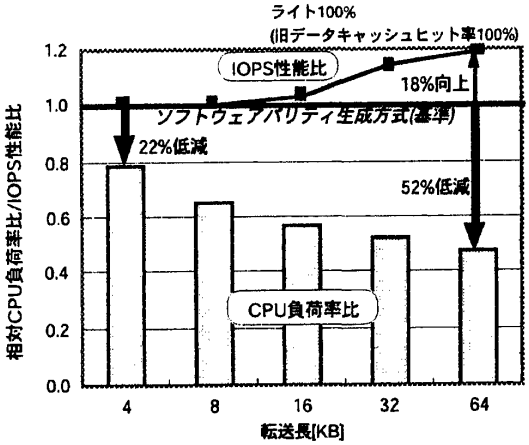

図2 DMAパリティ生成方式の相対CPU負荷率比/IOPS性能比

3. 性能評価

図2にDMAパリティ生成方式の1ライトI/O当たりのCPU負荷率およびIOPS性能を、ソフトウェアパリティ生成方式の結果を1とした時の相対比で示す。

DMAパリティ生成方式により、CPU負荷率はソフトウェアパリティ生成方式に比較し、4KB時には22%、64KB時には52%低減できた。また、IOPS性能は、32KB以上で性能が向上し、64KB時にDMAパリティ生成方式が18%高速であった。

4. おわりに

CPUの介在無く、また専用のMPUやバッファメモリを必要とせずに主記憶上で高速にパリティ生成を行う「DMAパリティ生成方式」をアクセラレータハードに実装した。本方式により、ソフトウェアパリティ生成方式に対し、1ライトI/O当たりのCPU負荷率を最大52%低減できることを実機で確認した。

参考文献

[1]D.A.Patterson, et al. : "A Case for Redundant Arrays of Inexpensive Disks (RAID)", Computer Science Division Department of Electrical Engineering and Computer Science, University of California Berkeley.

[2]兼田他:"PCサーバ向け「ハイブリッドRAID」の開発(2)～ソフトウェアアーキテクチャとディレイドパリティ生成方式～", 情報処理学会第57回全国大会, 1996.9.