RAID高速化技術 RAID BOOSTER$_{TM}$

RAID BOOSTER$_{TM}$ RAID Acceleration Technology

関戸　一紀　SEKIDO Kazunori
水野　聡　MIZUNO Satoshi

PCサーバにおけるCPUクロック周波数は，毎年1.5倍のペースで加速度的に速くなりGHzの世界に入ろうとしている。しかし，ディスクの速度（平均アクセス時間）は，こうした現状から取り残されようとしており，アーキテクチャの工夫により両者の性能ギャップを埋めることが，重要な課題となっている。

当社は，この課題の解決を目指して，ランダム書込みをシーケンシャル書込みに変換することで，RAID5書込み性能を2～6倍（当社比）改善するRAID（Redundant Array of Inexpensive（Independent） Disks）高速化技術を開発し，RAID BOOSTER$_{TM}$として製品化した。この技術はOLTPやDWHなどのデータベース処理業務など，膨大な量のデータをランダムに更新する作業の短縮化に大いに効果を発揮する。

This paper provides an overview of RAID BOOSTER$_{TM}$, a new architecture to accelerate write access in PC-based database systems using a RAID (redundant array of inexpensive (independent) disks). A write-mode conversion technique from random accesses into a set of sequential accesses has been realized in a special hardware module, which can achieve a several-fold improvement in the write performance of RAID5. Experimental results have proved that our system is efficient enough to perform database functions in on-line transaction processing (OLTP) and data ware housing (DWH) applications.

1 まえがき

増え続ける一方の情報処理業務にこたえるため，PCサーバのCPU性能は加速度的に向上し，そのペースは機械的動作を伴うハードディスクドライブ（HDD）の性能向上をはるかに上回っている。結果として，このCPUとHDD間の相対的な性能差は，今後も拡大傾向が続くと予想される。この両者の性能ギャップを埋めることが重要な課題である。

HDDの性能向上の手段としてRAID技術が広く利用されている。特に，後述のRAID5はHDDの容量効率が良く，誤り訂正可能なHDD構成により信頼性も高い。また，HDDの台数を増やすことにより性能を向上させることが可能であり，PCサーバに適していると言える。しかし，RAID5はデータ変更時にパリティ（誤り訂正符号）更新を伴うため，書込み時の性能がRAID0に比べると劣るという課題があった。

当社は，RAID5が持つメリットを最大限に生かすRAID高速化技術RAID BOOSTER$_{TM}$を開発した（図1）。この技術は，PCサーバでは世界初であり，課題であったRAID5の書込み性能を大幅に改善できる。以下，その基本原理とシステムの概要，及びその特長について述べる。

2 RAID BOOSTER$_{TM}$の基本原理

RAID BOOSTER$_{TM}$はLSFS（Log-Structured File Systems）技術[1]に基づいている。まず，LSFS技術及びRAID5技術との関係について述べる。

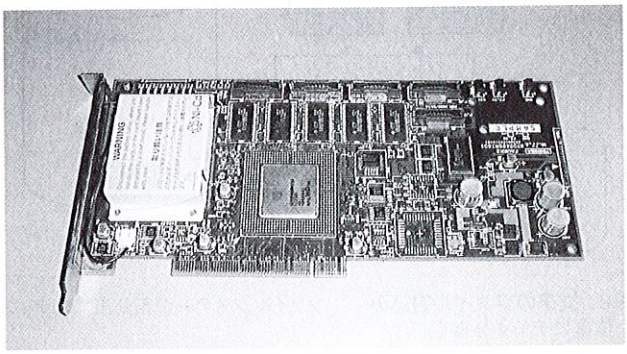

図1. RAID BOOSTER$_{TM}$ボード　PCIカードと制御ソフトウェアというシンプルな構成となっている。
RAID BOOSTER$_{TM}$ board

2.1 LSFS技術

LSFS技術とは「シーケンシャル アクセスは，ランダムアクセスに比較して非常に高速である」というHDDの特性を利用した性能向上の手法である。ランダムアクセス時の性能低下はHDDのシーク時間及び回転待ちによるもので，できればなくしたい部分である。

LSFSでは，データの記憶位置をHDD上で固定しない動的マッピングなどの機構を使って，複数の小さなブロックの書込みデータを，一つの大きいブロックのシーケンシャルな小さなブロックが順番に並んだログへ変換して“まとめ書き”する。ここで，“複数の小さなブロックの書込み”はファイルの論理構造とはまったく無関係である。つまり，異なるファ

イルであろうとも，書込みデータをある程度の大きさになるまでためて，それを大きなシーケンシャルログ(注1)としてHDDに記録する。

従来のファイルの書込みとLSFS技術を使ったRAID BOOSTER$_{TM}$による書込みを図2，図3に示す。アプリケーション プログラム(以下，アプリケーションと略記)A1，A2が，それぞれファイルF1のデータD1とファイルF2のデータD2の書込み要求を発行した場合の，制御とデータの流れを単純化したものである。

従来の書込みでは，ファイルシステムがD1，D2と異なる位置に割り当てたディスク上のそれぞれのブロックに対して書き込む。一方，RAID BOOSTER$_{TM}$が搭載されたシステムでは，ファイルシステムからのD1，D2の書込み要求をRAID BOOSTER$_{TM}$が一括してHDD上へシーケンシャルに書込む。

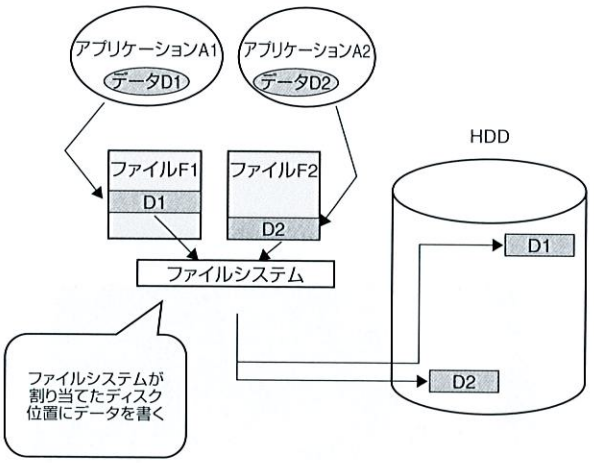

図2．従来のファイル書込み　ファイルシステムが割り当てたディスク位置にデータを書く。
Data flow of file writing without RAID BOOSTER$_{TM}$

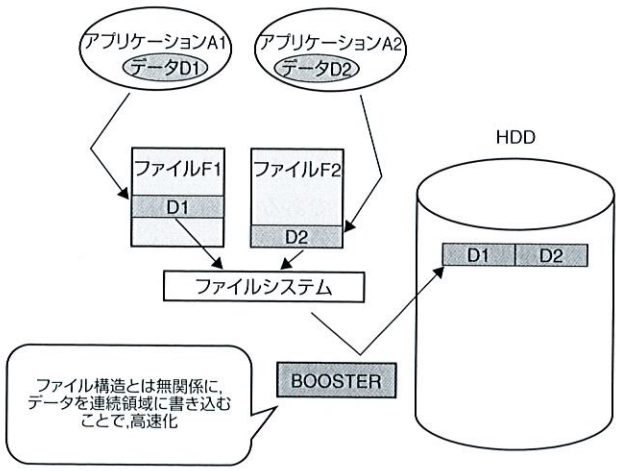

図3．RAID BOOSTER$_{TM}$によるファイル書込み　ファイル構造とは無関係に，データをまとめて連続領域に書き込むことで，高速化する。
Data flow of file writing with RAID BOOSTER$_{TM}$

これによりランダム書込みがシーケンシャル アクセスに変換される。

2.2 RAID5における書込みペナルティの解消

近年，ディスクの高速化と信頼性向上の観点から，RAID技術を搭載したPCサーバが主流となった。RAID技術には0，1，3，5などの種類(2)があるが，その中でも価格容量比の良さと優れた耐障害性から，RAID5が広く利用されている。

しかし，RAID5ではデータを更新する場合，次の処理がRAIDコントローラで行われる。(図4)

(1) 古いデータと古いパリティデータの読出し
(2) 新しいデータを加え，新しいパリティを算出
(3) 新しいデータと新しいパリティデータの書込み

このうち，古いデータの読出し，古いパリティデータの読出し，新しいパリティデータの書込みはHDD障害に備えた処理で，RAID5の書込みペナルティと呼ばれている。

結局RAID5では，信頼性の向上と引替えに書込み性能を犠牲にしている。このため，RAID5を使ったPCサーバでは，書込み性能，特にランダム書込みが悪く，それを意識したシステム構築が必要であった。なお，最新RAIDコントローラでは，パリティデータのキャッシングなど書込みペナルティを軽減する仕組みを持っているが，ランダム書込みに対してはほとんど効果がない。

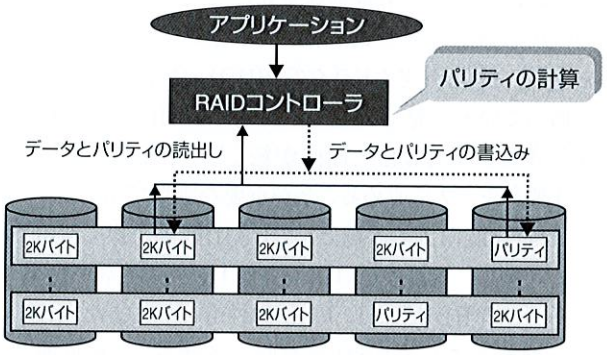

図4．従来のRAID5でのデータ更新　アプリケーションからの1回の書込みに対して，RAIDコントローラが2回の読出しと2回の書込みを行う。
Disk I/Os of RAID controller in RAID5 mode

一方，LSFS技術を使ったRAID BOOSTER$_{TM}$では，アプリケーションからの更新データをためて，大きなブロックのシーケンシャルログとして“まとめ書き”を行う。ここで，“まとめ書き”の単位をパリティグループ一つ分のデータ容量とし，その書出し位置をパリティグループの先頭に合わせることにより，RAIDコントローラでは“まとめ書き”の書込みデータだけからパリティを計算できるようになる。

(注1)　小さな書込みデータが順番に並んだグループ。

よって，先に述べた古いデータの読出し，古いパリティデータの読出し，新しいパリティデータの書込み，すなわちこれら書込みペナルティが発生しなくなり，書込み性能が向上する(図5)。また，書込みペナルティを意識したシステム構築もしなくて済むことから，よりいろいろな応用へ気軽にRAID5が使えるようになる。

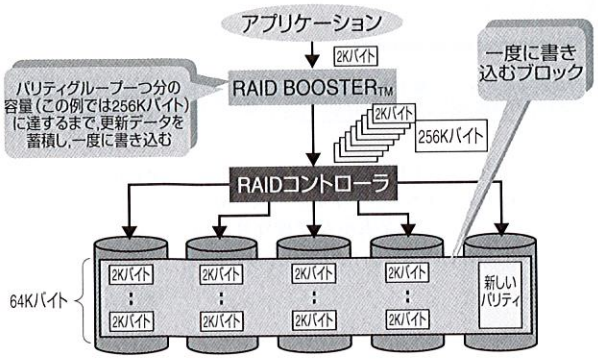

図5. RAID BOOSTER™によるRAID5でのデータ更新　“まとめ書き”の書込みデータだけからパリティを計算でき，書込みペナルティがなくなる。
Disk I/Os of RAID controller in RAID5 mode with RAID BOOSTER™

つまり，RAID5のPCサーバへRAID BOOSTER™を搭載した場合には，これまで説明してきた下記の両方の効果が相乗されて，書込み性能の大幅な改善が可能となる。

(1) “まとめ書き”によりHDDへのI/O(入出力)が減り，書込み性能が向上
(2) 書込みペナルティの解消により，不要なHDDの読出し／書込みがなくなり，書込み性能が向上

3 システムの概要及びその特長

RAID BOOSTER™システムは，図1に示した専用ボードと制御ソフトウェアから構成されており(図6)，RAIDコントローラを搭載したPCサーバMAGNIA™のPCI(Peripheral Component Interconnect)スロットに，このボードを装着するだけで容易に実装できる。また，オペレーティングシステム(OS)はMicrosoft®WindowsNT®(注2)をサポートしている。

RAID BOOSTER™ボードには，書込みデータの蓄積とマッピング情報の保存のために64Mバイトのメモリが実装されている。制御ソフトウェアはLSFS処理を行うフィルタドライバ，専用ボードを制御するカードドライバ，これらドライバの状態やRAIDコントローラの状態を監視するサービス，グラフィカルユーザーインタフェース(GUI)によりRAID BOOSTER™を適用するディスクを設定してその状態を表示するマネージャから構成される。以下，このシステムの特長について述べる。

(注2) Microsoft, WindowsNTは，米国Microsoft Corporationの米国及びその他の国における登録商標。

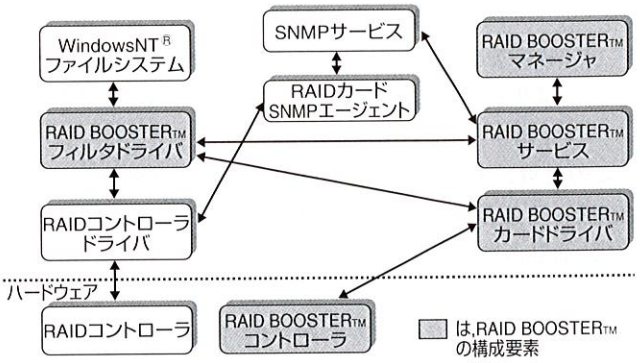

図6. RAID BOOSTER™システムの構成　RAID BOOSTER™は，専用ボード，ファイルドライバ，カードドライバ，サービス，マネージャから構成される。
Configuration of RAID BOOSTER™ system

3.1 書込みデータに対する保全性の確保

2.1節で述べたように，RAID BOOSTER™による性能向上のポイントは，アプリケーションからの書込みデータをためて，一括してHDDへ“まとめ書き”することにある。本来は，HDDに記録されていなければならない書込みデータがメモリ上にしか残っていないので，そのデータに関する保全性には十分な配慮が必要である。そこで，RAID BOOSTER™では以下の対策を実施している。

(1) バッテリによるメモリのバックアップ　バッテリモジュールをボードに内蔵しており，停電などの障害に対してもバッテリバックアップで72時間(通常使用時)メモリ内容をバックアップできる。
(2) ECC(Error Correcting Code)によるデータ保護
ボード上のメモリはECCで保護されており，メモリの1ビットエラーが発生してもそれを修復でき，データを失うことはない。
(3) 書込みデータの二重化　上記の対策だけでは，メモリの2ビットエラーを含むボードそのものが故障した場合，ボードに保存された書込みデータは失われてしまう。そこで，書込みデータに関しては以下の処理により，更にデータの保全性を高めている。
(a) 書込みデータはボード内だけではなく，サーバのメモリにもそのコピーを作成する。
(b) 通常のPCIボードでは，障害が発生した場合にリセット信号を発生し，サーバも止めてしまう。これをフェイルストップと呼ぶ。しかし，RAID BOOSTER™ボードでは障害があった場合でも，それによる影響をボード内だけに閉じ込め，サーバ本体まで波及しないように制御する。これを障害隔離と呼び，サーバ内のソフトウェアはそのまま実行できる。

(c) サーバ側の制御ソフトウェアでは，常にボードの状態をチェックし，障害を検出した場合には，速やかにサーバ上の書込みデータのコピーをディスクへ書き出す。

この対策により，RAID BOOSTER$_{TM}$ボードが故障した場合でも，そのときサーバ側の制御ソフトウェアが動作中であれば，書込みデータを失うことはない。

(4) ディスク障害時のダンプファイル(注3) RAID BOOSTER$_{TM}$では書込みデータをHDDへ“まとめ書き”するため，アプリケーションへ書込み完了を報告する時点と，実際にHDDへの書込みを行う時点が異なる。そのため，HDDへの書込み時にディスク障害が発生した場合，HDDへ書かれているべきデータがメモリ上に残ってしまう。

そこで，HDD障害の場合にはためていた書込みデータをシステムディスクなどへダンプファイルとして保存しておき，ディスク障害が回復した時点で自動的にそのダンプファイルをHDDへ書込むことにより，アプリケーションから見たデータの整合性を確保している。

3.2 完全なオープン性の実現

従来，LSFS技術を採用するには専用仕様のファイルシステムが必要であったが，RAID BOOSTER$_{TM}$ではLSFS処理を行うモジュールをフィルタドライバとして実装しているので，既存のファイルシステム(NTFS)を変更することなく，LSFS技術をPCサーバに適用できる。また，アプリケーション，ミドルウェア，ドライバに対して完全な透過性を持ち，オープンな操作環境を実現している。

3.3 WindowsNT®との親和性の高い設定ツール

RAID BOOSTER$_{TM}$の適用はパーティション単位で設定でき，HDDの容量を気にする必要はない。マネージャとは，RAID BOOSTER$_{TM}$が適用されるパーティションを作成／削除する専用のツールであり，このツールで作成されたパーティションに対してだけRAID BOOSTER$_{TM}$が適用される。

マネージャのGUIは，WindowsNT®のパーティション作成ツールである，ディスクアドミニストレータと同じような表示及び操作性を持っており，WindowsNT®の管理者ならば自然に使えるようにしている(図7)。

また，2.2節で述べたように，RAID5にRAID BOOSTER$_{TM}$を適用する場合，“まとめ書き”の単位をパリティグループ一つ分のデータ容量とし，その書出し位置をパリティグループの先頭に合わせる必要がある。それにはRAIDコントローラが持つRAIDの種類，HDD台数，ストライプサイズなどの情報が必要になる。

そこで，マネージャは，図6に示すように，RAID BOOSTER$_{TM}$サービス，SNMP(Simple Network Management Protocol)

(注3) 緊急対策として，障害時に作成されるファイル。

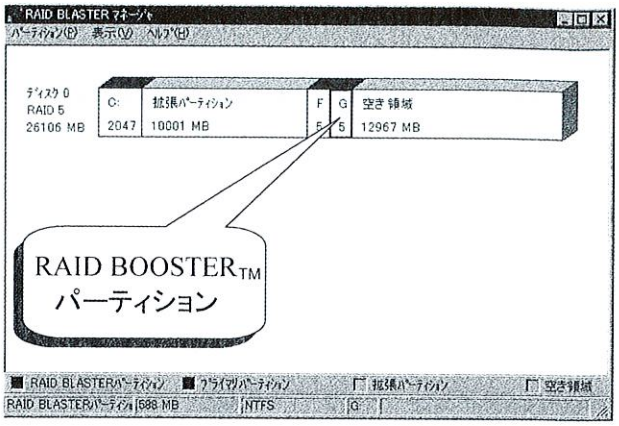

図7. RAID BOOSTER$_{TM}$マネージャ RAID BOOSTER$_{TM}$マネージャは，WindowsNT®のパーティション作成ツールであるディスクアドミニストレータと同じような表示及び操作性を持たせている。
Example of RAID BOOSTER$_{TM}$ Manager display

サービス，RAIDカードのSNMPエージェントを経由して，RAIDコントローラから上記情報を自動的に取得する。よって，管理者はRAIDコントローラの情報を調べる必要はない。

3.4 高性能で小型／低コストなRAID BOOSTER$_{TM}$カード

メモリには高速アクセスできるSDRAM(Synchronous DRAM)を採用し，内部バスをすべて64ビット幅で構成することにより，高いスループット(単位時間内の処理能力)のデータ転送を実現している。

また，主要機能をワンチップ化して部品点数を減らし，PCIハーフサイズ程度の基板サイズ，4層低コスト基板にて機能を実現している。

4 RAID BOOSTER$_{TM}$の性能

実測したRAID BOOSTER$_{TM}$によるRAID5構成で，ランダムアクセスの場合の書込み性能向上率とアクセスサイズとの関係を表1に示す。実用レベルのアクセスサイズで，2～6倍(当社比)の高速書込みができる。

表1. RAID BOOSTER$_{TM}$による書込み性能向上(RAID5，ランダム)
Random write performance improvement ratios achieved by RAID BOOSTER$_{TM}$

アクセスサイズ	2Kバイト	4Kバイト	8Kバイト
性能向上 (倍)	6.7	4.0	2.3

なお，アクセスサイズが大きくなると倍率が下がるのは，ストライプサイズを固定しているため，一回に“まとめ書き”できるブロック数が少なくなりその効果が減少するからであ

る。RAID BOOSTER$_{TM}$ではMicrosoft®SQL Server$_{TM}$[注4]やORACLE[注5]のデフォルト値(初期値)のブロックサイズである2Kバイトで最適となるようにしている。

また,RAID5構成サーバで,RAID BOOSTER$_{TM}$を適用した場合のOLTP(Online Transaction Processing)性能の測定結果を示す(図8)。測定処理には倉庫と在庫管理を模擬したベンチマーク(性能測定用プログラム)を使った。最大スループット(グラフのピーク値)の比較で2.3倍のシステム性能向上を達成できている。RAID BOOSTER$_{TM}$は"書込み"の高速化技術であるが,それがシステム性能の大幅な向上につながっている。

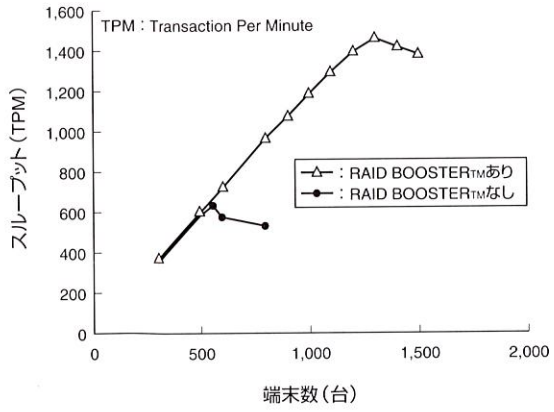

図8. OLTP性能の測定結果(RAID5)　最大スループットの比較で,2.3倍のシステム性能向上が得られる。
Results of OLTP performance measurement

(注4) SQL Serverは,米国Microsoft Corporationの米国及びその他の国における商標。
(注5) ORACLEは,Oracle Corporationの登録商標。

5 あとがき

RAID BOOSTER$_{TM}$はランダム書込みをシーケンシャル書込みに変換することで,RAID5の書込み性能を2～6倍(当社比)改善できる。PCサーバをベースとしたOLTPやDWH(Data Ware Housing)が本格化している。シンプルな構成と少ない投資でRAID5の信頼性とRAID0同程度以上の性能を達成する技術として,これらの応用にこたえていけるものと期待している。

今後は,RAID BOOSTER$_{TM}$の更なる高度化を図るとともに,より進んだディスク高速化技術を追求していく。

文　献

(1) Ousterhout, J., et al. Beating the I/O Bottleneck : A Case for Log-Structured File Systems. ACM Operating Systems Review. 23, 1, 1989, p.11-28.
(2) Patterson, D., et al. A Case for Redundant Arrays of Inexpensive Disks (RAID). Proceedings of ACM Conference on Management of Data. 1988, p.109-116.

関戸　一紀　SEKIDO Kazunori
デジタルメディア機器社 コンピュータ&ネットワーク開発センター 開発第一部主務。計算機アーキテクチャ・性能評価の研究・開発に従事。情報処理学会,IEEE会員。
Computer & Network Development Center

水野　聡　MIZUNO Satoshi
デジタルメディア機器社 コンピュータ&ネットワーク開発センター 開発第一部主務。計算機アーキテクチャ・FT計算機の研究・開発に従事。情報処理学会会員。
Computer & Network Development Center