

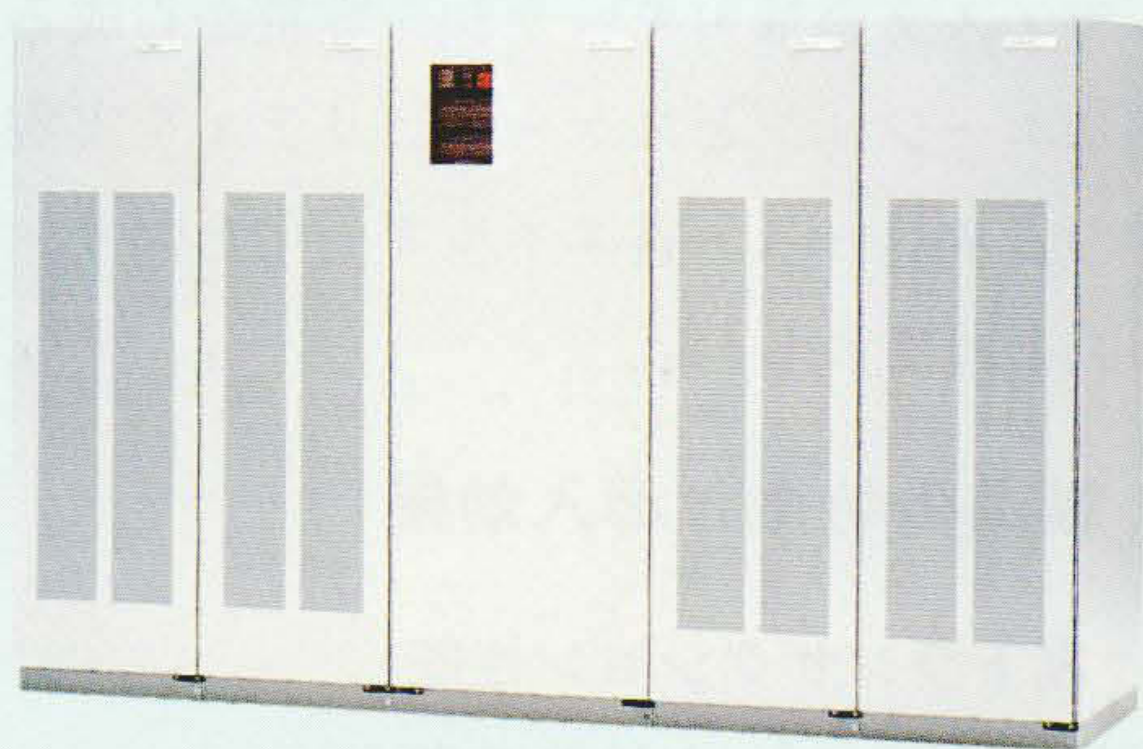
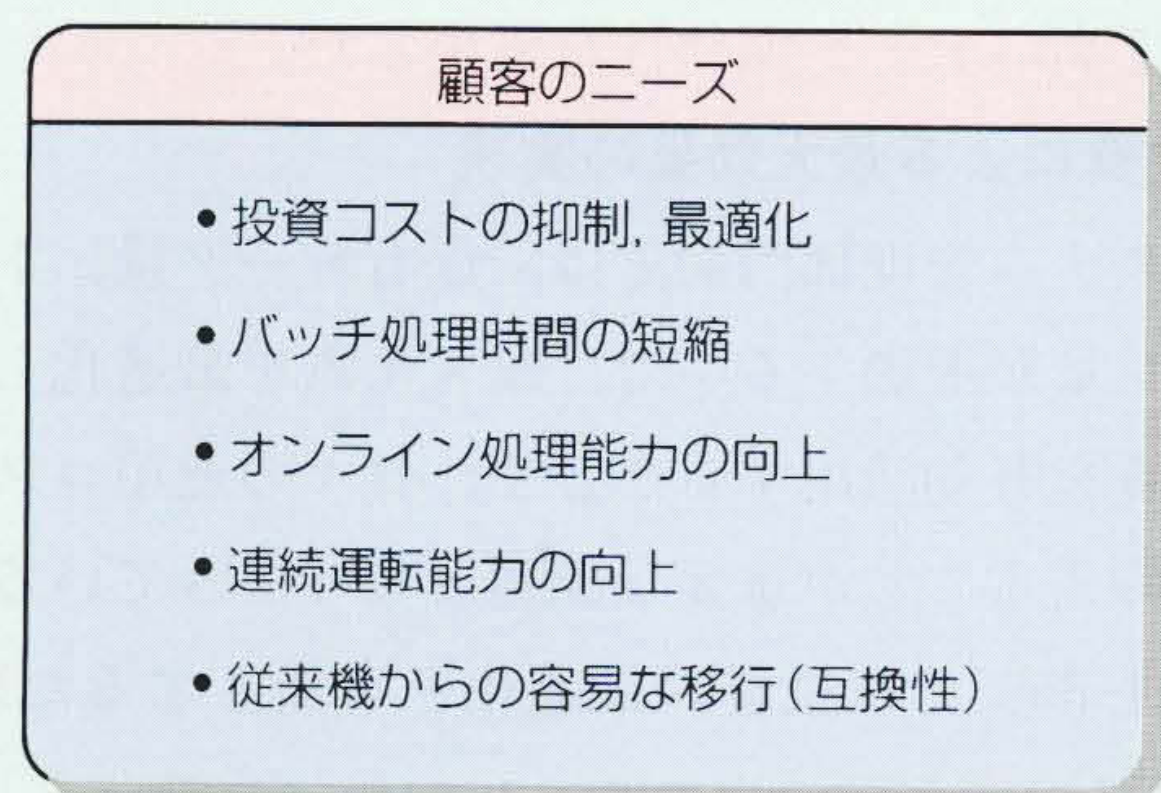
24時間・365日無停止運転を実現する ディスクアレー装置

Disk Array Subsystems Realizing Non-Stop 24-Hour 365-Day Operation

野沢正史* Masafumi Nozawa

高松久司* Hisashi Takamatsu

島田朗伸* Akinobu Shimada



H-6591/H-6595ディスクアレーサブシステム

製品コンセプト

連続運転性とデータ保全性の追求

- ハードウェア障害によるシステムダウン回避
ハードウェアコンポーネントの二重化, 冗長構成化
- 計画停止機会減少への対応
無停止保守, 無停止ハードウェア増設機能
- 重厚な障害防止策

利用環境に応じた最適性能の提供

- バッチ性能の向上
(RAID5データストライピング, キャッシュプリロード機能)
- 高トランザクション環境でのレスポンス安定化
- マルチプロセッササブシステム

投資の最適化

- 最適モデル選択可能
(RAID5/0, H-6588-3/H-6586-K, 通常モデル・一体形モデル)
- 移行コストの低減(H-6588-3/H-6586-Kエミュレーション)
- 運用コスト削減(省スペース, 省エネルギーの追求)

ディスクアレーサブシステムの製品コンセプト

多様化するニーズに対して最適なソリューションを提供するため、連続運転性とデータ保全性の追求、利用環境に応じた最適性能の提供、投資の最適化の三つの基本コンセプトに基づいてディスクアレーサブシステムを開発した。

企業の国際化の進展、業務の多様化に伴って基幹情報を支えるデータセンタに対しては、24時間切れ目の無いサービスの提供や大量データの高速処理への要求がますます強まっている。

これらのニーズにこたえる「Mパラレルシリーズ」の中核ストレージ製品として、ディスクアレーサブシステム“H-6591/H-6595”を開発した。H-6591/H-6595は、日立製作所の大型コンピュータ向けディスクサブシステムとして初めてRAID5技術を採用するとともに、制御装置の二重化または冗長構造によって高信頼性を実現する設計とし、また無停止保

守機能によって無停止運転を実現した。Mパラレルシリーズプロセッサの持つ無停止運転機能との連携により、Mパラレルシリーズトータルとして24時間・365日無停止運転に対応するものである。大容量キャッシュメモリのインテリジェント制御技術によって高速入出力処理も可能とする。

H-6591/H-6595は、従来実現されていなかった高次元の可用性や信頼性あるシステムを最小限のコストで構築するものである。この製品を使用することによってユーザーは、無停止運転や運用負担の軽減など数多くのメリットを得ることができる。

* 日立製作所 ストレージシステム事業部

1 はじめに

近年、企業の国際化の進展に伴ってコンピュータシステムの運用の切れ目が無くなってきている。また、各企業が取り扱う業務は多種多様化しており、深夜や休日もオンラインシステムを稼動させるユーザーが増加している。このため、特に企業の基幹情報を扱う大型システムに対しては、障害によるシステムダウンが許されないことはもちろんのこと、ハードウェア増設や定期保守のための計画停止もほとんどできない状況となってきている。

一方、大型コンピュータでは、大量の全社レベルデータを一括管理するデータサーバとしての役割に加え、インターネット・イントラネット環境をはじめとするオープン環境でのスーパーサーバとしての新たな役割への期待も大きい。こうした中で大型ストレージシステムには、多数のユーザーの要求に応じてデータの高速検索・参照・更新を可能とする入出力性能向上への要求が強まっている。

こうした背景の中で日立製作所は、今後いっそう厳しくなる可用性、データ処理性能向上へのニーズにこたえるため、新世代大型コンピュータシステム「Mパラレルシリーズ ストレージ サブシステム」の中核製品としてディスクアレーサブシステム“H-6591/H-6595”を開発した。大型システム向けディスクサブシステムとして初めてRAID(Redundant Array of Independent Disks)技術を採用し、24時間・365日無停止運転を可能とするものである。

ここでは、H-6591/H-6595の製品コンセプトと特徴、および導入時の効果について述べる。

2 製品コンセプト

“H-6591/H-6595”は、新世代大型システムに求められる多様なニーズに対して最適なソリューションを提供するため、以下に示す三つの基本コンセプトに基づいて開発した。

2.1 利用環境に応じた最適性能の提供

大型システムが処理する業務は、オンライントランザクション処理とバッチ処理の2形態に大別される。オンライントランザクション処理に対しては、トランザクション量の増大に対応するスループット性能と安定したレスポンス性能の提供が求められている。一方、バッチ処理に対しては、オンライン時間延長に対応して処理時間の短縮が求められている。これらに対応するため、ランダム入出力環境でのスループットの向上と、シーケンシャル入出力性能の向上を図った。

2.2 連続運転性とデータ保全性の追求

前章で述べたように大型システムに対しては24時間・365日無停止運転へのニーズが強い。これに対応するため、1点故障がサブシステム全面ダウンにつながらないハードウェアアーキテクチャ、無停止ハードウェア増設・保守機能を実現した。また、大型システムの厳しい稼動環境の下で安定稼動を実現するため、新たに開発した高信頼ディスクの採用や多種のデータ保護・チェック機能によって高いデータ保全性を実現した。

2.3 最小投資による最大効果の実現

企業のシステム管理部門間では、保有データ量、トランザクション量が増加する中で、システムを最適化して限られた投資を有効活用することや、日々の運用コストを最小限に抑えることが重要な課題となってきている。

H-6591/H-6595では投資の最適化を可能とするため、二つのRAIDレベル、二つのエミュレーションモード、幅広いディスク容量・キャッシュ容量から成る豊富なモデルバリエーションとスケラビリティを実現した。また、装置の省スペース・省エネルギー化を図り、運用コストの削減を可能とした。

3 製品の特徴と導入効果

3.1 高トランザクション性能

3.1.1 スループット性能の向上

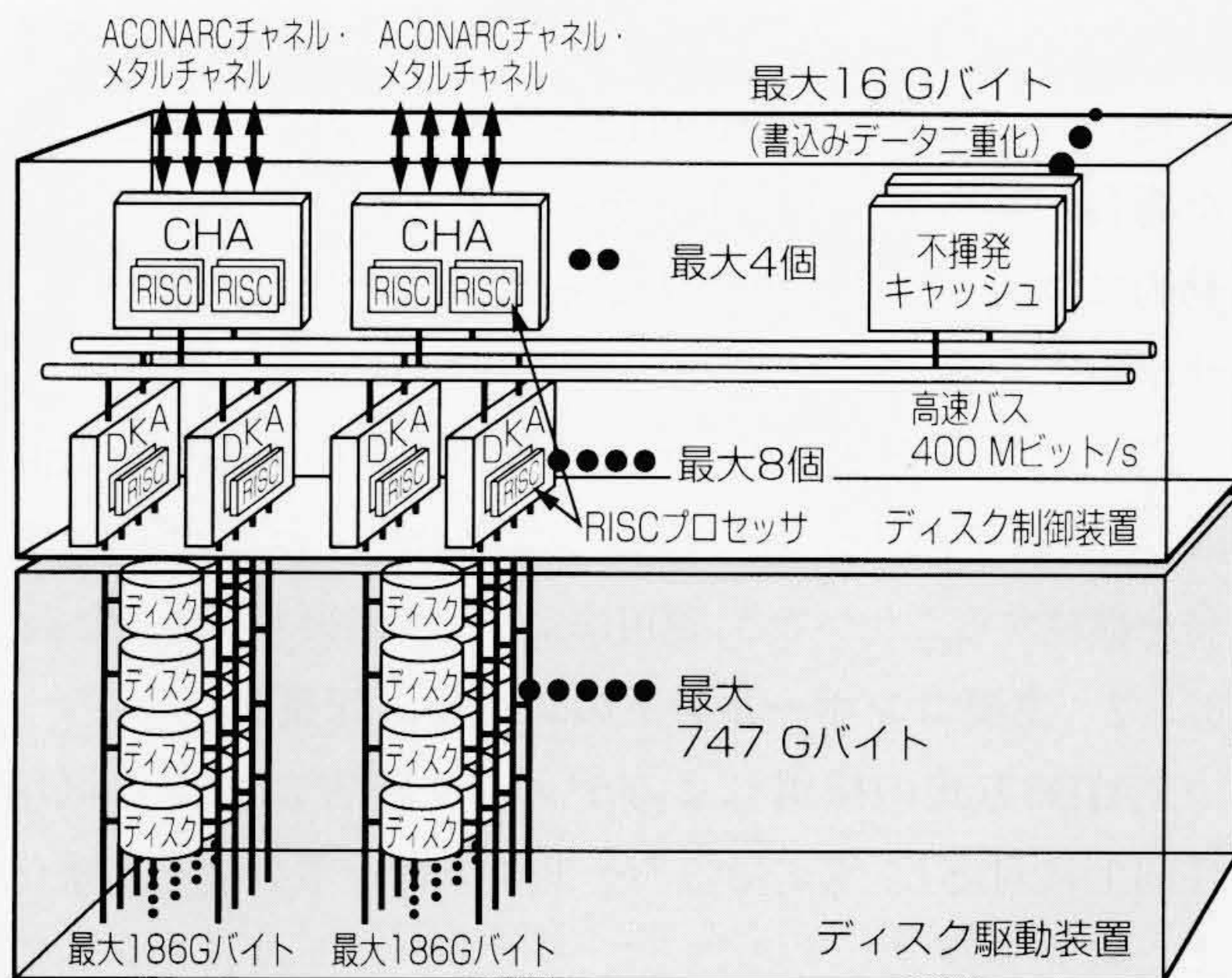
(1) RAID5およびRAID0方式の採用

トランザクション量の増大が続いているユーザーにこたえて安定した入出力性能を提供するためには、スループット性能を向上させる必要がある。そこで、複数個の入出力要求を同時に多重処理することが可能なRAID5およびRAID0レベルを採用した。

(2) マルチプロセッサ構成

ディスクを効率的に多重動作させてスループット性能を向上させるため、ホストーキャッシュ間のデータ転送を制御するCHA(Channel Adapter)側、およびキャッシューディスク間のデータ転送を制御するDKA(Disk Adapter)側にそれぞれ8プロセッサずつの最大16マルチプロセッサ構成を採用した。高スループットを実現するハードウェアアーキテクチャを図1に示す。各プロセッサは同時に独立して動作可能であり、サブシステム当たり最大40多重の同時転送が可能な能力を持つ。

以上のように、RAID5およびRAID0方式の採用と、マルチプロセッサによるスループット向上によって高トランザクション性能を実現した。



注：略語説明

ACONARC (Advance Connection Architecture)

CHA (Channel Adapter ; プロセッサ2個搭載)

DCA (Disk Adapter ; プロセッサ1個搭載)

RISC (Reduced Instruction Set Computer)

図1 高スループットと高信頼を実現するハードウェアアーキテクチャ

このディスクアレーサブシステムは、ホストとキャッシュ間のデータ転送を制御するCHA、キャッシュとディスク間のデータ転送を制御するDCA、不揮発キャッシュ、CHA—キャッシュ—DCAを結ぶ高速バス、およびディスクドライブで構成する。

3.1.2 レスポンス時間の短縮

(1) 大容量キャッシュ

オンラインレスポンス時間を短縮するためにはキャッシュ上にアクセスデータが存在する確率(以下、ヒット率と言う。)を高め、ディスクへのアクセス回数を削減する必要がある。そこで、ディスク容量に対するキャッシュ容量の比率を従来製品よりも高めることでヒット率の向上を図った。H-6591/H-6595では、従来製品の16倍に当たる最大16 Gバイトのキャッシュが搭載可能であり、高いキャッシュ比率の構成とすることができる。

(2) 拡張高速書き込み機能

一般にデータの更新処理では、更新前のデータの参照処理が行われるために高いヒット率が期待できる。さらに、固定長形式で記録されているデータに関しては、キャッシュ上にデータが存在しない場合もキャッシュ上にデータを書き込み、この時点でホスト側には書き込み終了の通知を行う拡張高速書き込み機能をサポートした。これにより、データ更新時のヒット率をほぼ100%とすることができ、メモリ性能でのデータ更新処理が可能となる。

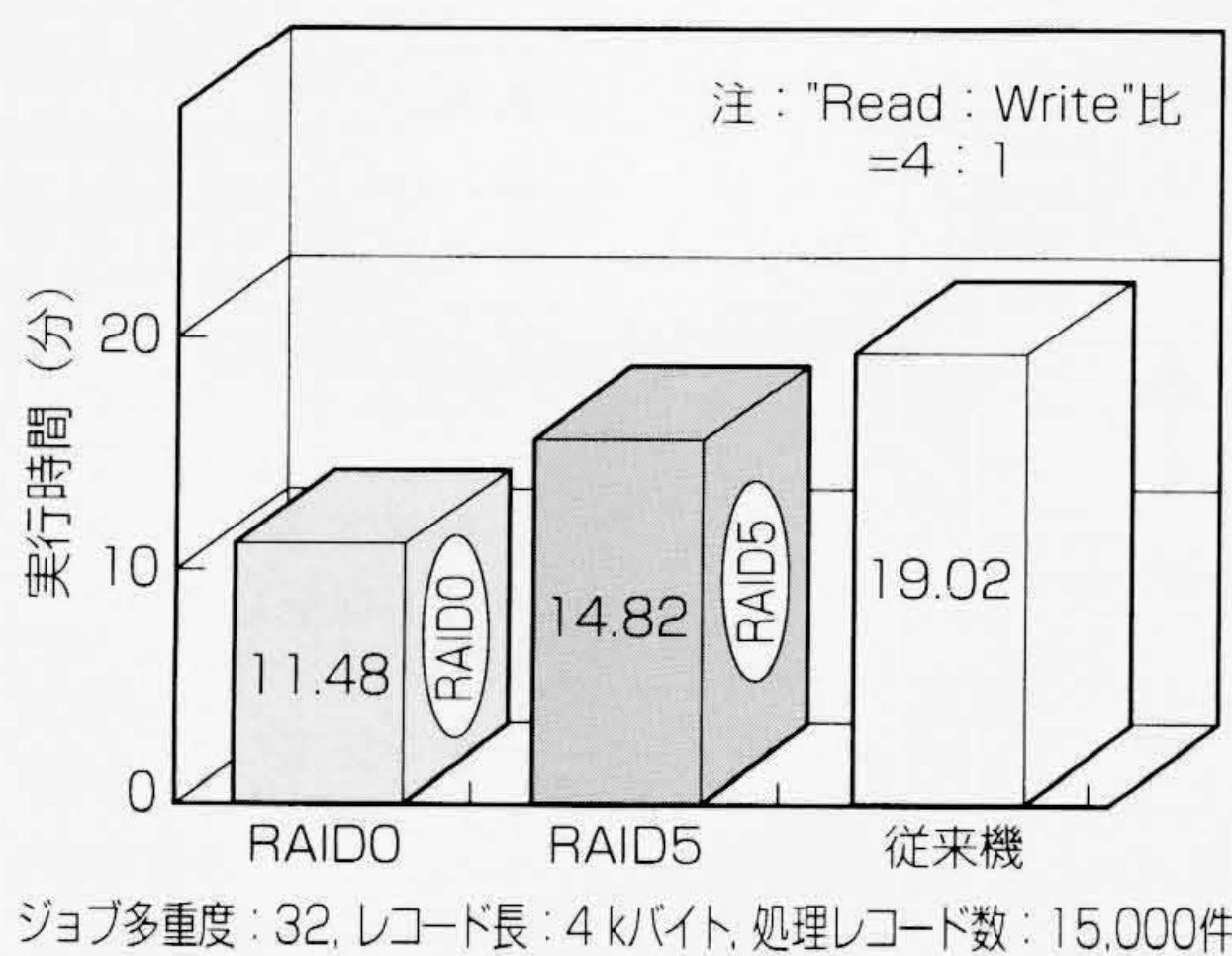
(3) ライトペナルティの克服

RAID5方式では、データ更新の際にディスクへの更新データ書き込み以外に、パリティデータを更新するための処理が必要となる。この処理を「ライトペナルティ」と呼ぶ。前述のように、更新処理のキャッシュヒット率はほぼ100%となるため、このライトペナルティが直接ホストへのレスポンス時間に影響を与えることはない。ただし、高負荷環境ではサブシステム全体のスループットを低下させる原因となる。そこで、キャッシュからディスクへのデータのデステージング処理^{※1)}の際、ディスク上の格納位置が隣接するデータをまとめて書きする方式を採用した。この方式により、ディスクへのアクセス回数が削減でき、ライトペナルティ自体の発生を低減した。したがって、ユーザーのRAID5を意識した性能上の考慮は不要であり、従来製品と同様のアプリケーションに適用することができる。

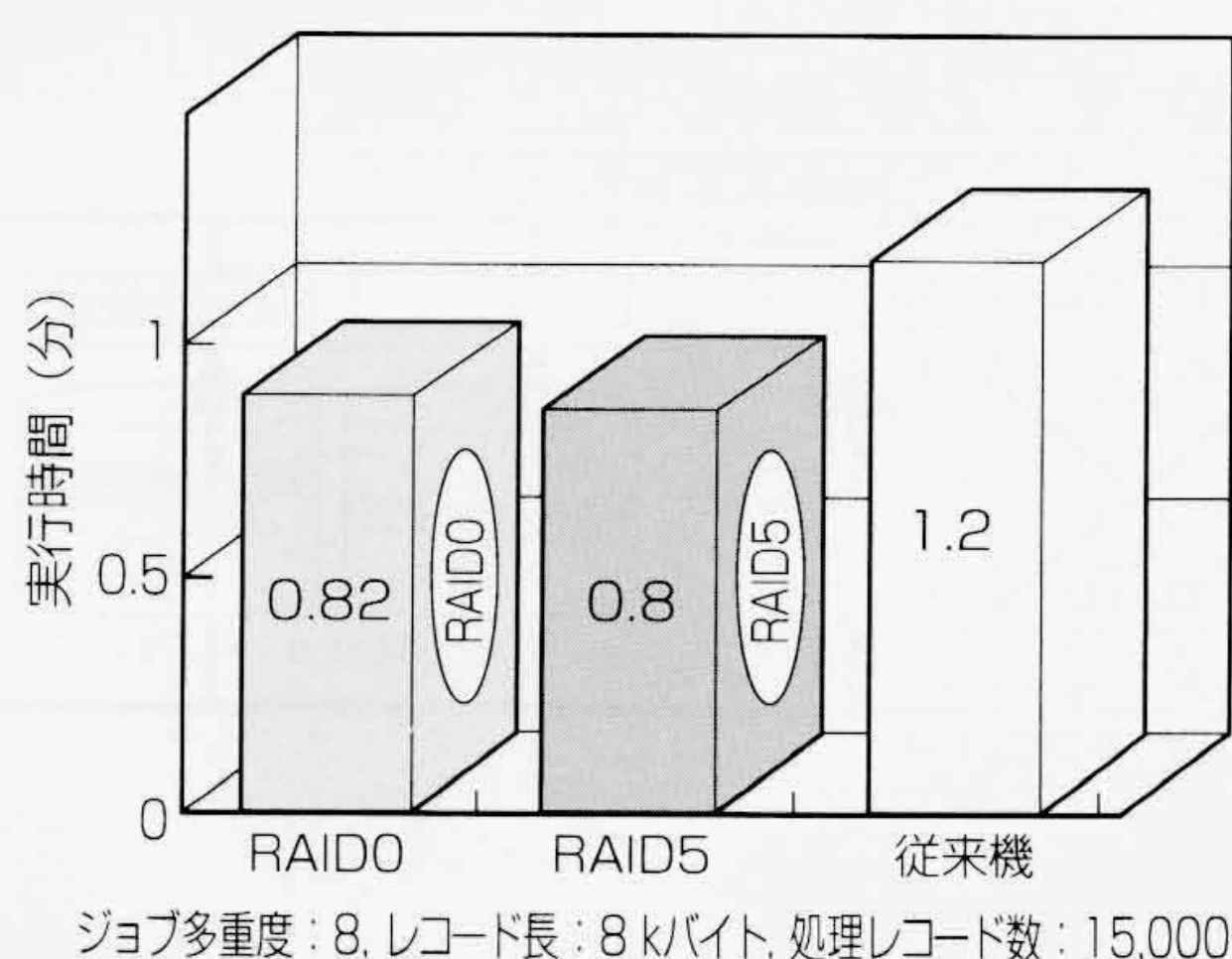
3.2 バッチ処理の高速化

大型システムでのもう一つの代表的なアプリケーションとしてバッチ処理がある。バッチ処理でのディスクサ

※1) デステージング処理：キャッシュ上の記録されたライトデータをディスク上に反映する処理のこと



(a) ランダムアップデート



(b) シーケンシャルリード

図2 性能向上効果(ベンチマークテスト結果)

オンラインランザクション処理の代表的なアクセスパターンであるランダムアップデートでは従来機に比べてRAID5は22%, RAID0は40%, またバッチ処理のアクセスパターンであるシーケンシャルリードではRAID5およびRAID0で30%以上、それぞれ性能が向上した。

ブシステムへの入出力形態は、シーケンシャル リード・ライトが主体である。H-6591/H-6595では、シーケンシャルアクセス時にキャッシュへデータの先読み処理を行うプリロード機能をサポートした。RAID5モードでは、複数のディスクを並列動作させて高速転送で先読み処理が実行できる。これにより、大量データのシーケンシャルリード・ライト処理をすべてキャッシュ・ホスト間の転送性能である17 Mバイト/s(ACONARCチャネル接続時)で処理でき、ジョブ実行時間を大幅に短縮することが可能である。

以上のスループットおよびレスポンス性能向上の効果、プリロード機能によるバッチジョブ実行時間短縮の効果の一例を図2に示す。この例では、トランザクション処理に代表されるランダムアップデート処理で、従来機(H-6588-3)に比べてRAID5で22%、RAID0で40%性能が向上した。またシーケンシャル リード ジョブでは処理時間を従来機に比べてRAID5で35%、RAID0で30%短縮した。

3.3 24時間・365日無停止運転対応

3.3.1 RAID5によるデータ回復機能

RAID5モードでのディスク障害時のデータ回復処理の流れを図3に示す。ホストからの要求データが格納されているディスクに障害が万一発生した場合でも、ほかのディスクのデータとパリティデータを使って要求データを回復し、ホストへ転送することが可能である。これを「コレクションリード機能」と呼ぶ。またスペアディスクを搭載した場合、障害が発生したディスク上のデータをスペアディスク上に再構築する。これを「コレクシ

ョンコピー機能」と呼ぶ。さらに、通常のリード・ライト処理時に発生する修正可能エラー回数が一定値を越えた場合、該当ディスク上のデータをスペアディスクに自動的にコピーする「ダイナミックスペアリング機能」をサポートする。

以上の3機能により、ディスク障害発生の確率を低減し、ディスク障害が万一起った場合でもユーザーは業務を継続することができ、運用による障害回復も不要となる。

3.3.2 主要コンポーネントの二重化、冗長化

RAID5方式の採用によるディスク障害に対する可用性向上に対応し、すべてのハードコンポーネントを二重化または冗長構成とすることで、サブシステムトータルで高い信頼性を実現する設計とした。二重化、冗長化したハードウェアシステム構成を図4に示す。これにより、どの部位に障害が発生した場合でもサブシステムダウンは発生せず、別コンポーネントを使用して処理が継続できる。

3.3.3 無停止保守・増設機能の装備

ディスク、キャッシュ、電源、ファン、制御パッケージなどの主要コンポーネントはすべて、サブシステム稼動中に交換可能な構造としている。また、制御装置内のマイクロコードの保守、バージョンアップをサブシステム稼動中に行うことも可能である。これにより、障害時の部品交換はもちろんのこと、容量拡張やオプション増設もサブシステムへのアクセスを継続したまま行うことが可能であり、保守・増設のためのシステムの計画停止を極少化できる。

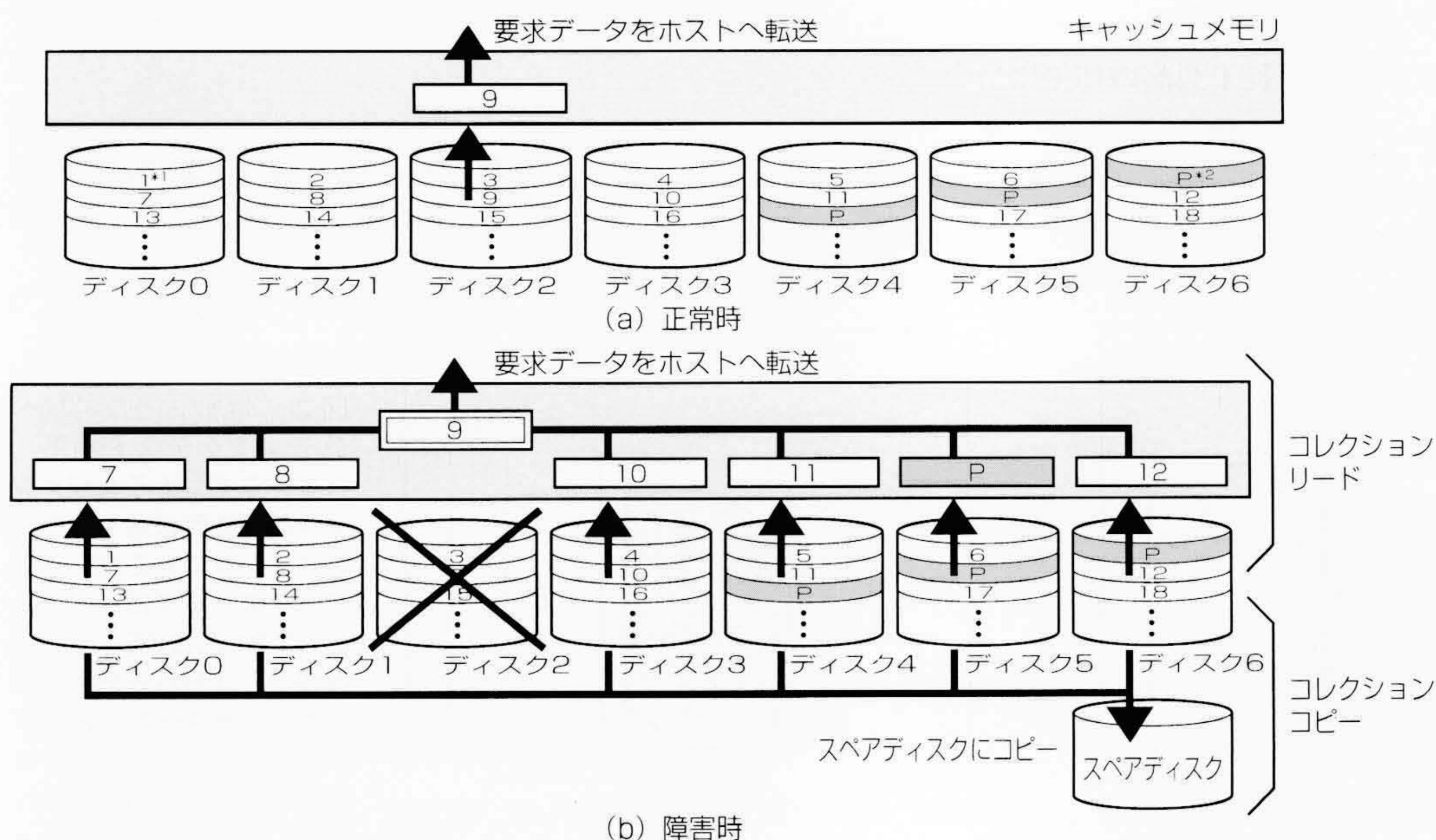


図3 RAID5のデータ回復機能

障害が発生したディスク2に格納されているデータ9の読み出し要求があった場合、正常なディスクからデータ7, 8, 10, 11, 12, およびパリティデータPをキャッシュメモリ上に読み出し、データ9を復元してホストへ転送する(コレクションリード機能)。これと並行して、ディスク2のすべてのデータを正常なディスク0, 1, 3~6のデータから復元し、スペアディスク上にディスク2を再構築する(コレクションコピー機能)。

注：*1 1, 2, 3... ; 各ディスクに格納されているユーザーデータ *2 P ; データ回復用パリティデータ

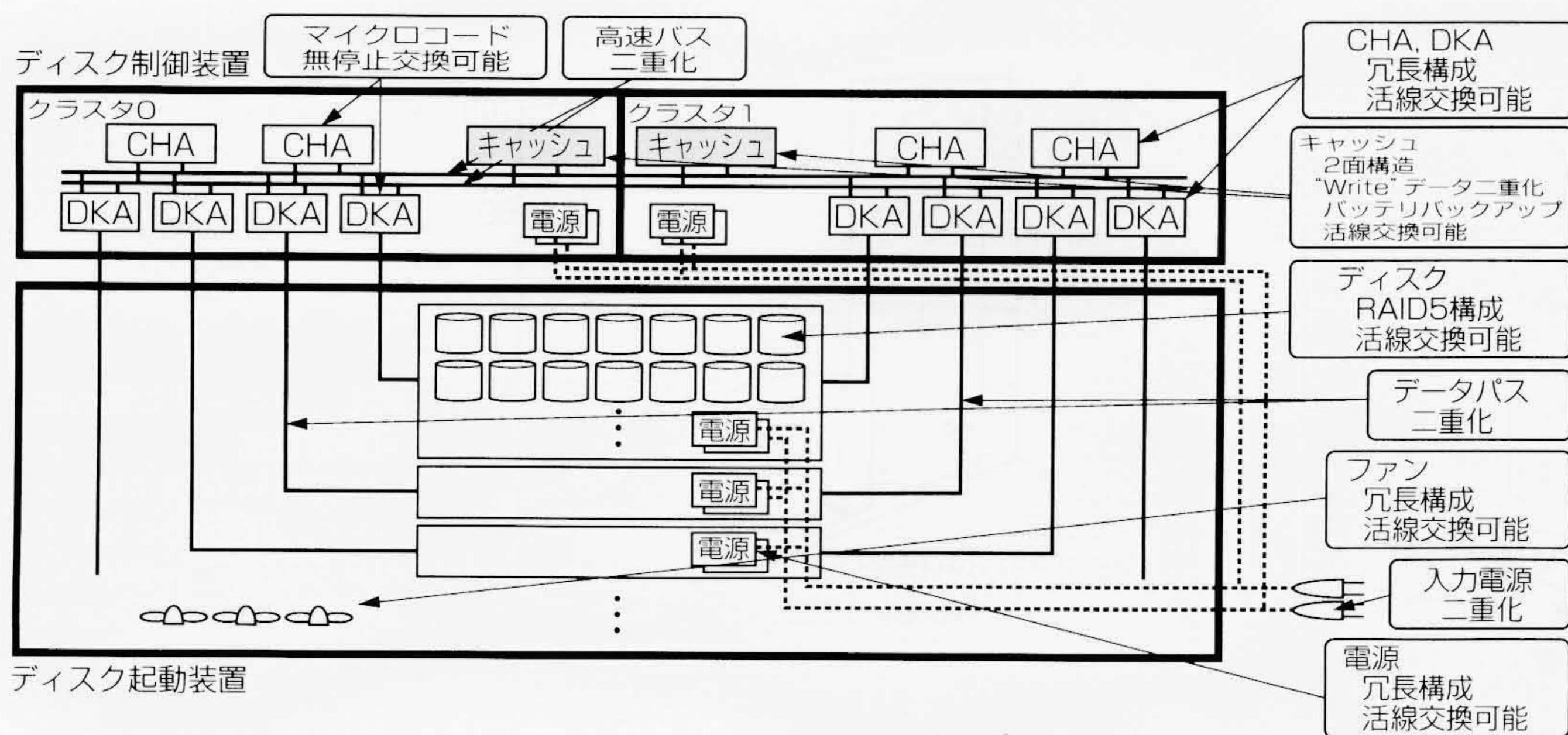


図4 二重化、冗長化ハードウェアのシステム構成
主要ハードウェアコンポーネントはすべて二重化または冗長構成とし、1点障害によるサブシステムダウンが発生しないシステム構成としている。また、無停止保守機能をサポートしている。

3.3.4 ディスク二重書き機能のサポート

RAID0モードでは、ホスト側のプログラムとの組合せによるサブシステム単位のデータ二重化機能であるディスク二重書き機能をサポートした。これにより、制御装置を含めた完全二重化システムが構築でき、金融系オンライン元帳ファイルなどの最重要データに対して最高レベルの信頼性を提供する。

3.4 高いデータ保全性の提供

高度なデータ保証と長時間連続運転が要求される大型システムの環境の中で、ダイナミックスペアリング、コレクションリード、コレクションコピーの過度の発生を防ぎ、ユーザーに対して安定稼動を提供するため、このサブシステム専用に開発した高信頼3.5インチディスクを採用した。

このディスクは従来の大型ディスクの設計思想・技術を踏襲したものであり、ディスク単体レベルでも従来機と同等の信頼性を実現した。また、制御装置内の各プロセッサ間での相互チェックやデータ保証コードを多数付加することにより、高いデータ保全性を実現した。

3.5 豊富な製品ラインアップ

H-6591/H-6595はディスク容量が23 Gバイトから747 Gバイト、キャッシュ容量が256 Mバイトから16 Gバイトまでの幅広い容量帯をサポートするスケーラブルモデルと、ディスク容量が最大93 Gバイト、キャッシュ容量が最大2.5 Gバイトのシングルキャビネットモデルの2モデルをラインアップする。どちらのモデルもRAID5およびRAID0の二つのRAIDレベル、およびH-6588-3・H-6586-Kの二つのエミュレーションタイプをサポートしている。このため、ユーザーのシステム規模、適用業務、移行前機器タイプに応じて最適なモデル・構成・RAIDレベル・エミュレーションタイプを自由に組み合わせるこ

とができ、ユーザーは必要最小限の導入コストで最適なシステムを構築することができる。各RAIDレベルおよびディスク二重書き機能の実業務での使い分けを図5に示す。

3.6 運用コストの削減

3.5インチディスクと高密度実装技術の採用により、設置スペースと消費電力を大幅に削減した。設置スペースと消費電力の従来機との比較を図6に示す。単位容量当たりの設置面積を比較した場合、H-6588-3の64%減、H-6587-3の84%減である。同じく消費電力ではH-6588-3の58%減、H-6587-3の87%減である。これらにより、フロア賃借料や空調費、電力料金などのシステムのランニングコストを大幅に削減することが可能である。

3.7 ソフトウェアの透過性と移行の容易性

H-6591/H-6595では、従来機H-6588-3およびH-6586-

	RAID5	RAID0+	RAID0+WDCP
適用業務	●高可用性、高シーケンシャル性能が要求されるデータセット ●オンラインファイル ●バッチファイル ●システムファイル	●低コスト重視 ●ワークファイル ●従来と同等レベルの信頼性が要求されるデータセット	●最高レベルの信頼性、高ランダム性能が要求されるデータセット ●オンライン重要ファイル ●システムファイル
特徴	経済性 1.6 1.3 1.5 10⁵ 可用性の高さ	経済性 1.7 1.7 1.1 1.5 可用性の高さ	経済性 0.8 1.7 1.1 10⁷ 可用性の高さ

注：略語説明

WDCP (Double Disk Volume Control Program ; ディスク二重書き機能制御用プログラムプロダクト)

図5 各RAIDレベルの特徴と適用分野

従来機を1とした場合のH-6591/H-6595の経済性、性能、可用性は各RAIDレベルによって異なる。これら各RAIDレベルの特徴に応じて、RAID5はオンライン・バッチ・システムファイルなど全般に、RAID0はワークファイルに、RAID0とディスク二重書き機能の組合せはオンライン重要ファイルにそれぞれ適用するのが効果的である。

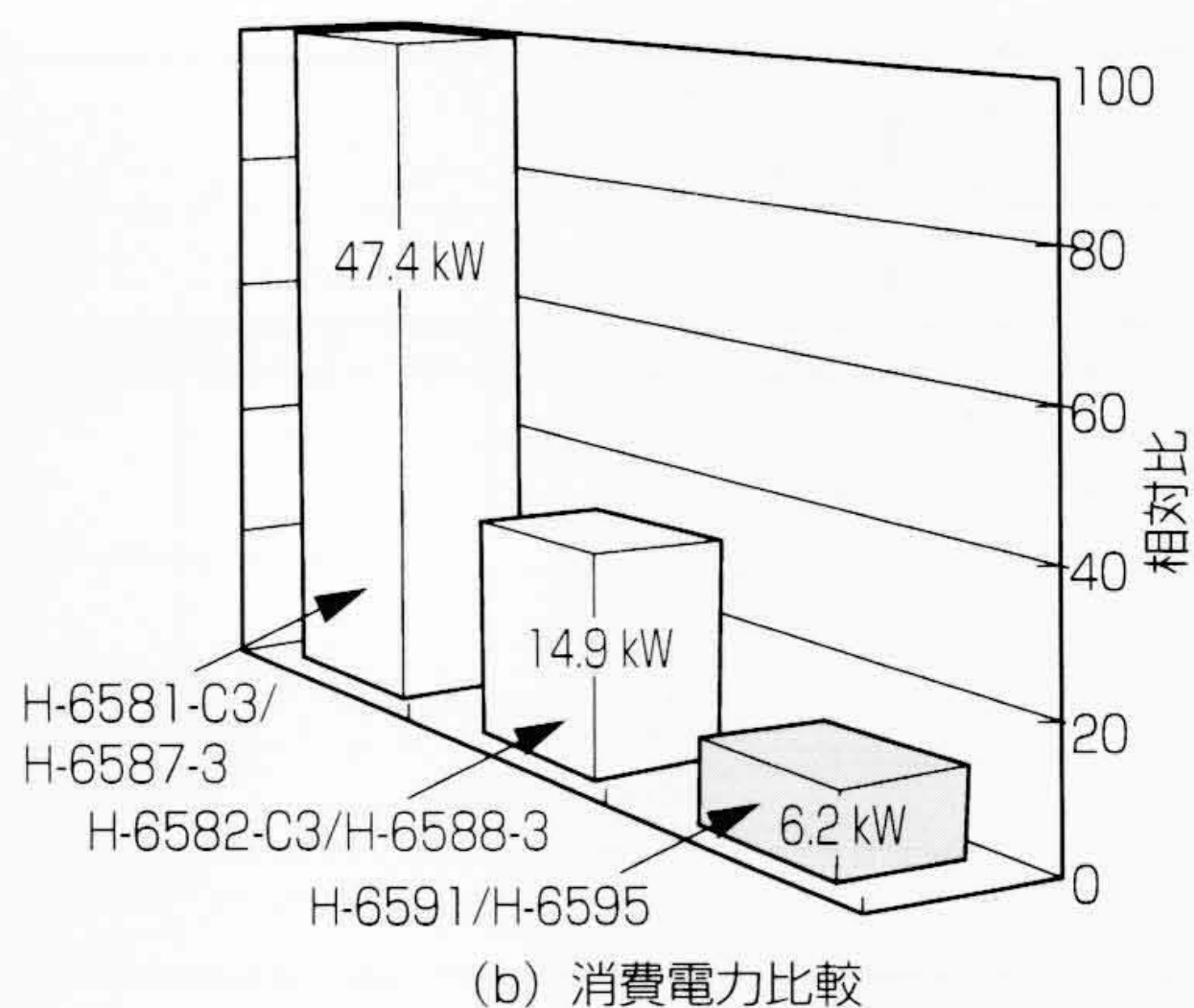
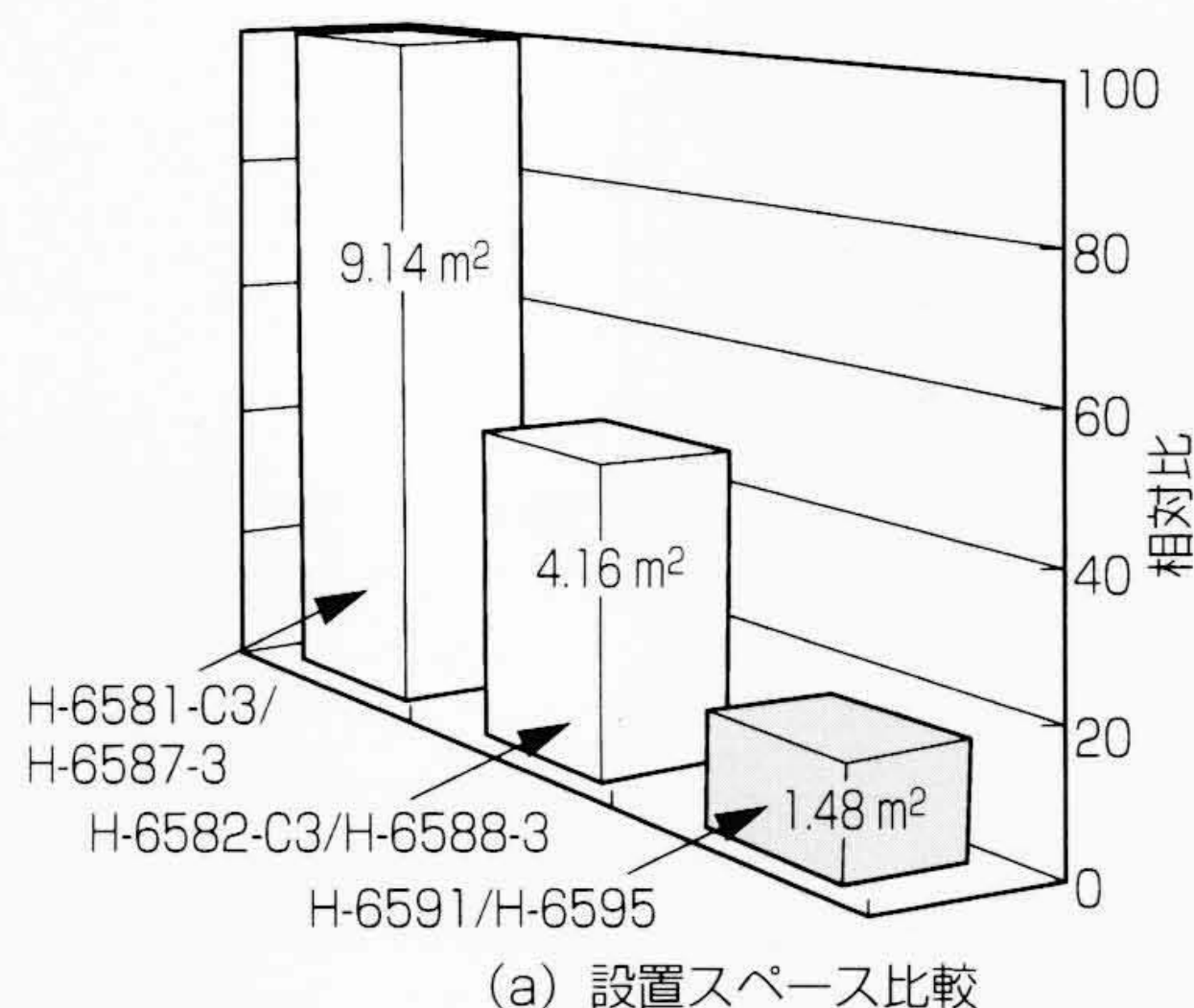


図6 スペースと電力削減効果比較

H-6591/H-6595では、スペース、消費電力ともにH-6588-3の約 $\frac{1}{3}$ 、H-6587-3の約 $\frac{1}{6}$ にそれぞれ大幅に削減した。

Kのトラックフォーマット、ボリュームフォーマットをエミュレートする方式を採用している。RAID5はディスク7台をRAIDグループとし、H-6588-3での8ボリュームまたはH-6586-Kでの12ボリュームをエミュレートする。RAID0ではディスク1台でH-6588-3の1ボリュームをエミュレートする。これにより、ユーザーおよびアプリケーションからはディスクアレーの構造を意識することなくH-6588-3、H-6586-Kと同様に使用できるので、プログラムや運用の変更は不要である。また、従来機からのデータ移行時もデータセットの再配置などは必要なく、ボリューム対ボリュームのコピーによる容易な移行が可能である。

4 将来展開

4.1 大容量ディスク

ユーザーの保有データ量の増大への対応や、いっそうの低価格化への要求にこたえるため、使用ディスクの大容量化を図り、サブシステム容量の拡大とコストパフォーマンスの大幅向上を目指す。

4.2 マルチプラットフォーム

大型システムとオープンシステムとの融合・連携は今後ますます進展していくものと考ええる。この中で、UNIX^{※2)}、パソコンなどの異なるプラットフォームとの

接続性を強化していく。これにより、オープンシステムにも大型システム並みのレベルの高い信頼性、性能を提供するとともに、ユーザーのストレージへの投資コストの最適化を可能とする。また、ストレージを一元化することによって管理コストを削減することも可能となる。最終的にはこれをさらに発展させて、大型システム、オープンシステム間でのデータ共有の実現を目指す。

4.3 高機能コントローラ

制御装置レベルでのデータの遠隔二重化、アクセスを継続しながらの自動データマイグレーション、データのアクセス形態に応じた最適RAIDレベルへの自動配置などの機能をサポートし、ストレージ運用の自動化を推進する。

5 おわりに

ここでは、連続運転性の向上やデータ処理性能向上のニーズにこたえて開発したディスクアレーサブシステム“H-6591/H-6595”の特徴、導入時の効果について述べた。

これにより、システムの24時間・365日無停止運転の実現、運用負担・コストの軽減、入出力性能の向上を図ることができる。

今後も、この装置の基本プラットフォームをベースとして、大容量ディスクの採用、異なるプラットフォーム間のデータ連携機能の強化、コントローラの高機能化などを実現していき、この装置からより多くのユーザーメリットが生じるように力を注いでいく考えである。

※2) UNIXは、X/Open Company Limitedがライセンスしている米国ならびに他の国における登録商標である。

参考文献

1) 斉藤, 外: 高信頼・小型化を実現した大容量磁気ディスク装置, 日立評論, 76, 2, 135~140(平6-2)

2) 安部, 外: 新世代大型コンピュータ Mパラレルシリーズの開発, 日立評論, 77, 5, 371~374(平7-5)