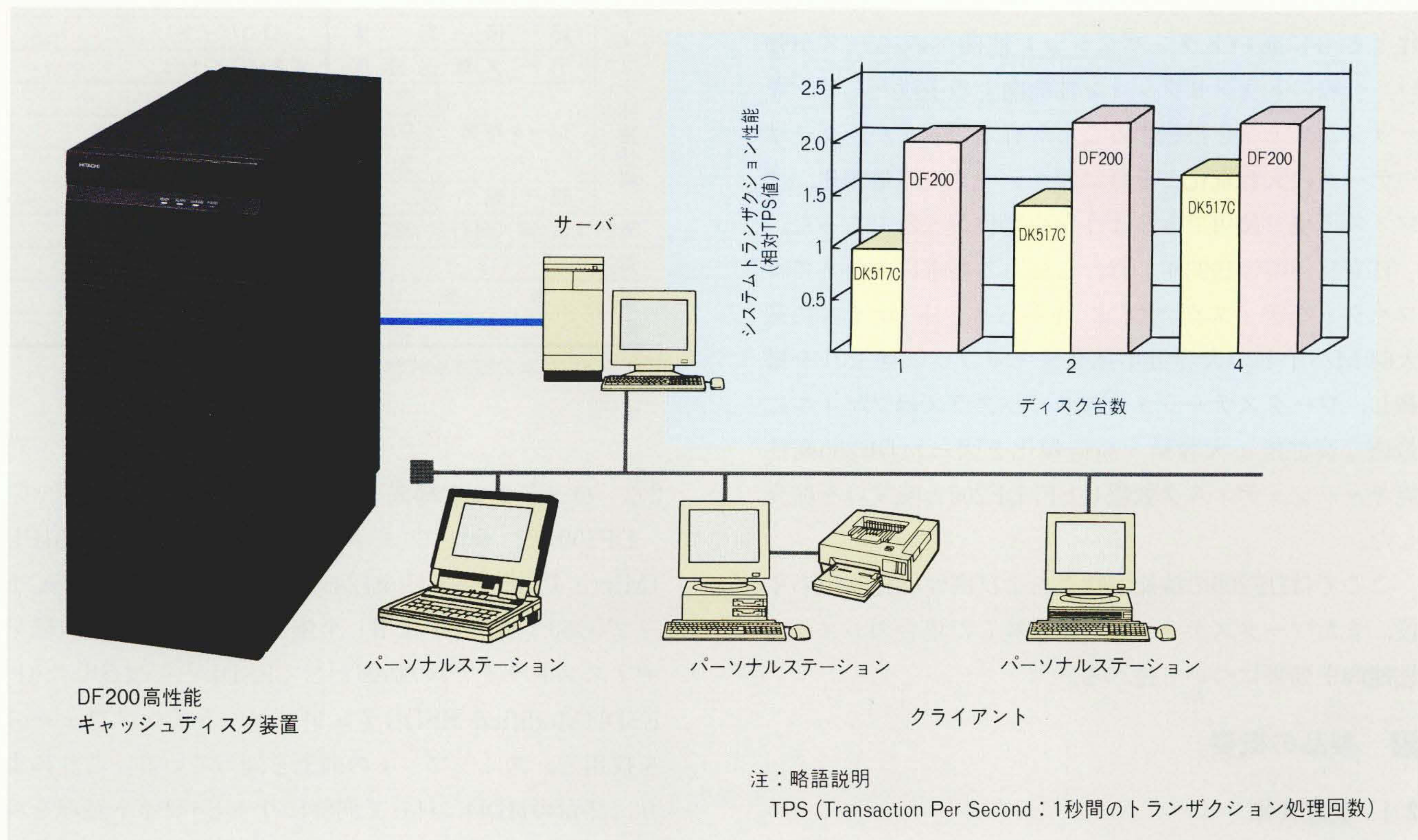


データ処理能力を大幅に向上させた 高性能キャッシュディスク装置

High Performance Cache Disk Subsystem

吉田 稔* Minoru Yoshida
春名利之* Toshiyuki Haruna
永井英男* Hideo Nagai



高性能キャッシュディスク装置の構成例とシステム性能評価 DF200高性能キャッシュディスク装置は最大64 Mバイトの大容量不揮発キャッシュメモリを搭載し、高いトランザクション性能によってシステム性能向上に威力を発揮する。右上の性能グラフはRDB(リレーショナルデータベース)管理ソフトを使用し、銀行の預金取引業務をモデルとした性能ベンチマークテスト結果を示す。

ダウンサイジング、クライアントサーバシステムといった中・小型コンピュータを利用したネットワーク・分散処理が急速に進展している。その中心となるワークステーションなどのサーバに使用される磁気ディスクファイルには、大容量化とともに高い性能と高信頼性が要求されてきている。

今回これらの要求にこたえ、最新の半導体デバイスを使用し、最大64 Mバイトの大容量不揮発キャッシュメモリを搭載した最大記憶容量11.4 Gバイトの

DF200高性能キャッシュディスク装置を開発した。

これまで日立製作所は、メインフレームコンピュータシステムに使われる磁気ディスクファイルに、大容量の不揮発キャッシュメモリと高度の制御論理を駆使した、高性能・高信頼性のディスク装置を提供してきた。この技術を継承・活用し、ワークステーションのサーバクラスのファイルに対するディスク装置の提供にも積極的に取り組んでいる。

* 日立製作所 ストレージシステム事業部

1 はじめに

オープンシステムや分散処理の拡大により、ワークステーションのビジネス分野での進出が著しい。なかでもRDBMS(Relational Database Management System)を使用したOLTP(On-line Transaction Processing)分野の拡大が進んでいる。このようなシステムのサーバ装置に使用される磁気ディスクファイルには、容量の増大化とともに高いトランザクション性能へのニーズが強い。一般にトランザクション性能向上の手段として、キャッシュメモリを搭載することが有効であるが、ユーザーデータの大容量化とそれに伴うデータの高信頼化、オフィス環境で使えるような小型化が課題であった。

日立製作所は1993年3月、これらの要求にこたえてコンパクトなディスク サイド キャビネット サイズに最大64 Mバイトの大容量不揮発性キャッシュメモリを搭載し、ワークステーションのサーバクラスのファイルに最適な高性能・大容量・高信頼化を図ったDF200高性能キャッシュディスク装置(以下、DF200と略す。)を開発した。

ここではDF200の機能、特長および高性能化実現の手段、またワークステーションに接続した場合のシステム性能向上効果について述べる。

2 製品の概要

2.1 装置仕様

DF200の主な仕様について表1に示す。ディスク サイド キャビネット サイズの筐(きょう)体に最大4台の大容量磁気ディスク装置を搭載し、最大11.4 Gバイトの記憶容量を実現した。また最大64 Mバイトの不揮発性キャッシュメモリ、高性能RISC(Reduced Instruction Set Computer)プロセッサの採用と高度なキャッシュ制御論理により、日立製作所従来装置(DK516C)と比較して約6倍のI/O性能向上を実現した。また上位ホストとのインタフェースにはSCSI-2^{※1)}を採用し、キャッシュはホストから意識する必要はなく、従来の単体ディスク装置との上位互換性を保っている。

表1 DF200の主な仕様 ホストインタフェースにSCSI-2を採用し、ディスク サイド キャビネット サイズの筐(きょう)体で最大11.4 Gバイトの高性能ファイルを提供する。

項 目		DF200-5・DF200-11
記 憶 容 量		5.7 Gバイト・11.4 Gバイト
イ ン タ フ ェ ー ス		SCSI-2
最 大 デ ー タ 転 送 速 度		10 Mバイト/s
デ ー タ キャ ッ シ ュ 容 量		16・32・48・64 Mバイト
キャッシュデータ保持時間		7 日間
デ ィ ス ク 装 置 仕 様	内 蔵 デ ィ ス ク 装 置	5 インチディスク
	記 憶 容 量	2.87 Gバイト
	デ ー タ 転 送 速 度	4.8 Mバイト/s
	シーク時間	最 大 27.0 ms
		平 均 12.8 ms
		最 小 2.0 ms
	回 転 速 度	5,400 r/min
外形寸法(幅×奥行き×高さ)		250×600×600(mm)
質 量		最大60 kg
電 源 条 件		AC100 V・AC220 V
周 波 数		50/60 Hz

注：SCSI-2[本文脚注※1)参照]

2.2 ハードウェア構成

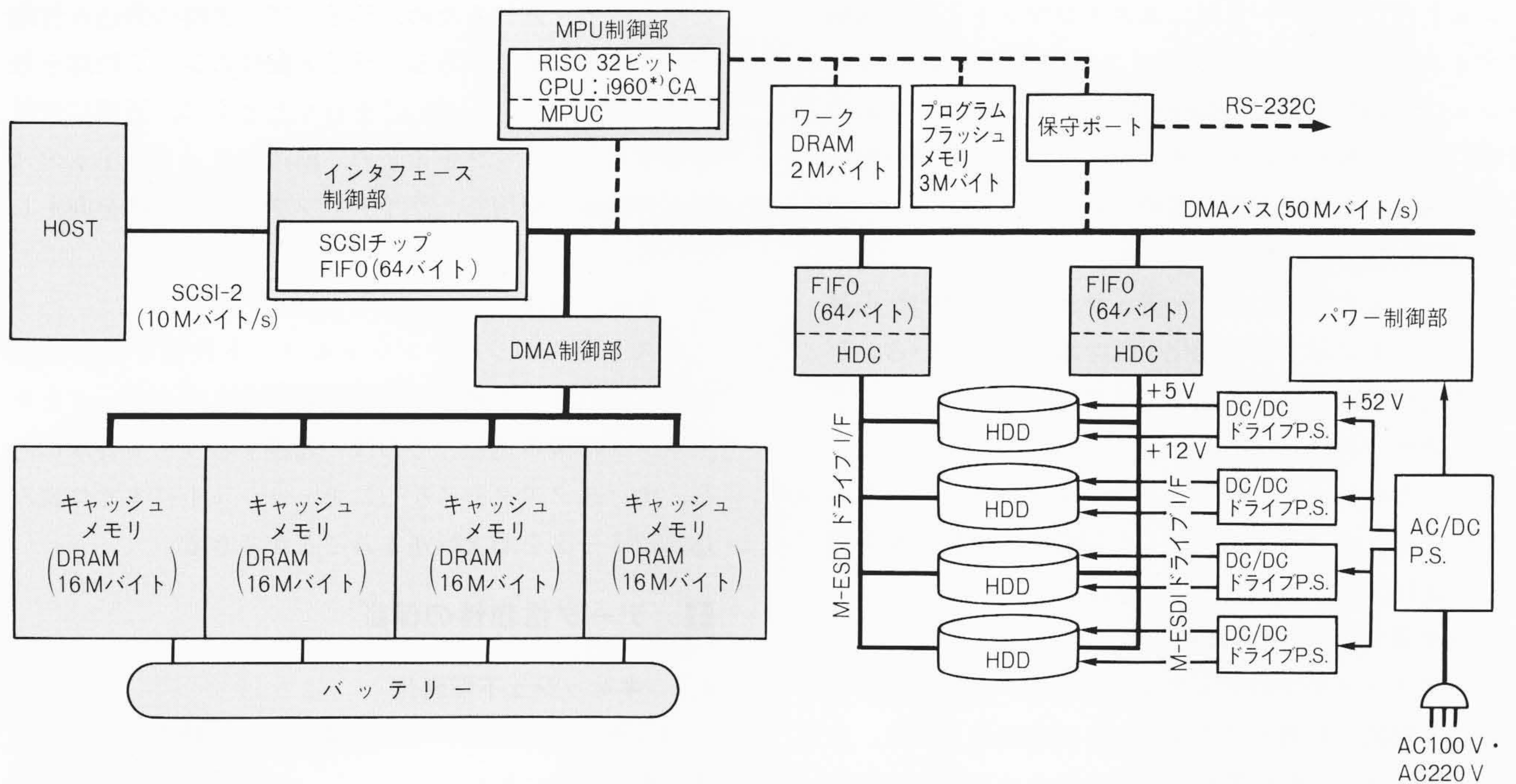
DF200のハードウェア構成を図1に示す。MPU(Micro Processor Unit)制御部には32ビットのRISCチップ(i960 CA^{※2)})を採用し全体を制御する。HDD(磁気ディスクドライブ)の制御としてESDI^{※3)}を改良したM-ESDI(Modified-ESDI)デュアルポートインタフェースを採用し、スループットの向上を図っている。これにより、2台のHDDに対して同時にリード・ライト処理を可能にする。またキャッシュメモリには、16 Mビット省電力タイプ・セルフリフレッシュ機能内蔵の最新DRAMを採用し、基本構成16 Mバイト(メモリカード1枚実装)、最大64 Mバイト(メモリカード4枚実装)までのキャッシュ記憶容量を持つ。このキャッシュメモリは、バッテリーでバックアップされて不揮発化を図っている。

内部DMA(Direct Memory Access)バスは32ビット構成で最大50 Mバイト/sの高速データ転送能力を持ち、上位ホスト装置および2台のHDDに対する同時データ転送処理を可能にしている。また上位ホストインタフェースのワイドSCSI(20 Mバイト/s)への拡張に対しても

※1) SCSI-2は、Small Computer System Interface-2の略語で、ANSI(米国国家規格協会)で規格標準化された小型システム向け周辺装置の標準インタフェースである。

※2) i960は、米国Intel Corp.の商標であり、またi960 CAは、Intel Corp.の製品名称である。

※3) ESDIは、Enhanced Small Disk Interfaceの略語で、米国ANSIで規格標準化された小型ディスク装置用インタフェースである。



注：略語説明はか

FIFO (First In First Out Buffer), P.S. (Power Supply), HDC (Hard Disk Controller), RISC (Reduced Instruction Set Computer), MPUC (Micro Processor Unit Controller), M-ESDI (Modified ESDI), I/F (Interface), HDD (Hard Disk Drive), DMA (Direct Memory Access), SCSI (脚注※1参照)

※) i960は、米国Intel Corp.の商標である。

図1 DF200ハード構成のブロック図 高速DMA(Direct Memory Access)バスを持ち、ホストおよび2台のHDDに対する同時データ転送処理が可能である。

対応可能としている。

入力電源仕様としては、全世界規模の対応を可能とし、また、HDDへの電源給電については、独立したDC/DCコンバータで各HDDへ供給する方式を採用して信頼性の向上を図っている。また、これらの電源電圧値およびキャッシュバッテリーの出力電圧値について、パワー制御部によって常時監視することにより、電源異常が発生した場合、MPUに指示を出して速やかにデータの保護機構を稼働させる。万一の障害発生時の装置状態は、フロントパネル内の8ビットLED(Light Emitting Diode)でエラー状態を表示できるようにし、保守を容易化している。

3 高性能処理の実現

3.1 キャッシュ機能

最大64 Mバイトの大容量不揮発キャッシュメモリは、16 kバイト単位のセグメントに分割しリード・ライトデータ共用に使用する。各セグメントはさらに512バイトのブロック単位に分割され、ヒットミスの管理が行われる。

ホストからのキャッシュを使ったリード・ライト動作

を図2に示す。リードヒット時はディスクアクセスなしに直ちにデータがホストへ転送されてホストコマンドが終了し、高速応答を実現する。リードミス時にはいったんディスクアクセスを伴うため、通常のディスクとほぼ同等の応答時間となる。ライトでは、ライトヒット、ライトミスいずれもキャッシュにデータを書き込みライトコマンドは直ちに終了し、高速応答が実現される。キャ

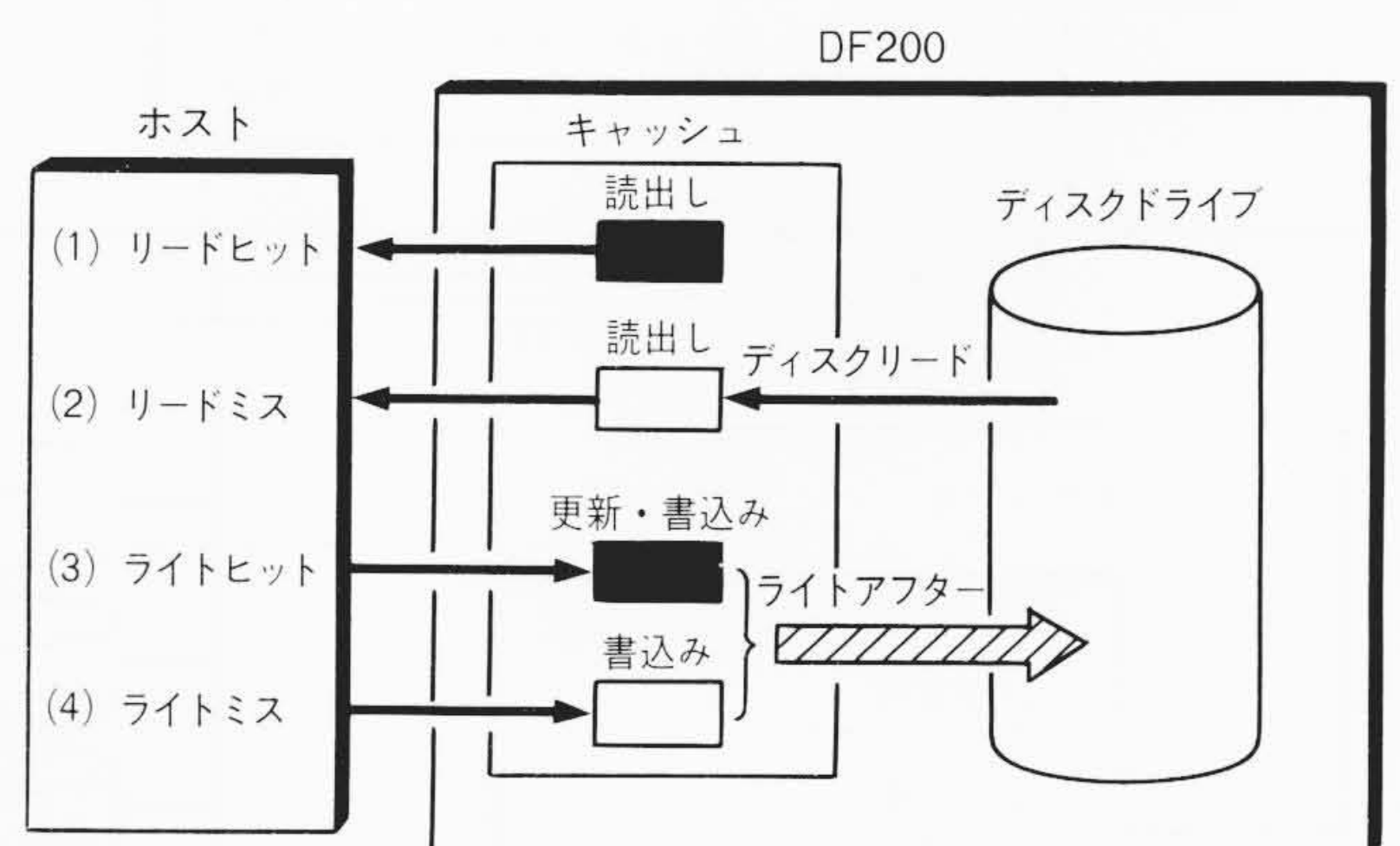


図2 DF200キャッシュリード・ライト動作 (1), (3), (4)動作は、キャッシュアクセスで高速応答する。(2)動作はディスクリード後、ホスト転送し、単体ディスクとほぼ同性能である。

ッシュ上のライトデータは、ホストコマンドとは非同期にディスクに書き込まれる(ライトアフター機能)。キャッシュは、LRU(Least Recently Used)アルゴリズムの管理に従って最も古いデータをキャッシュから追い出して再使用され、ヒット率を高めている。

3.2 高速化アルゴリズム

キャッシュを使用したリード・ライトデータの転送制御には、以下に示す高性能化手法を取り入れている。その高性能化手法の概要を図3に示す。

(1) 同時データ転送機能

ホストとキャッシュ間、キャッシュと両ディスクポート間の合計3チャンネルの独立した同時データ転送を可能とし、高I/O多重処理性能を実現している。

(2) 高多重ディスク制御機能

ディスクドライブのデュアルポートインタフェースを使用し、動的に処理パスを選択する制御機能を持つ。また、ディスクに回転位置検出機構を持たせ不要な回転待ちを行わない。これらによってディスクドライブ使用率を高め、スループットを向上している。

(3) リードライト・スキップ機能

キャッシュとディスク間のリードライト時、不要なブロックデータの転送をスキップするハードウェア機構を持たせ、1回のディスクアクセスで不連続なブロックを余分な回転待ちなく効率よくリードライトする。

(4) シークソート、まとめ書き機能

キャッシュがライトデータでいっぱいになり性能低下

となることを避けるため、ライトアフタ時の書込み性能を向上させる必要がある。ライト動作のシーク距離を最短化するソート(並べ替え)を行うとともに、近傍に位置するライトブロックをまとめ、前述のリードライト・スキップ機能を使用し、ライトアフタの処理性能を向上している。

(5) 先読み機能

ホストからのシーケンシャルリード性能を高めるため、1回のリード動作で要求範囲よりも先のデータをキャッシュに読み込む。その後の連続するリード要求に対し、ディスクアクセスなしにキャッシュ上にすでに読み込んだデータを順次転送することができる。

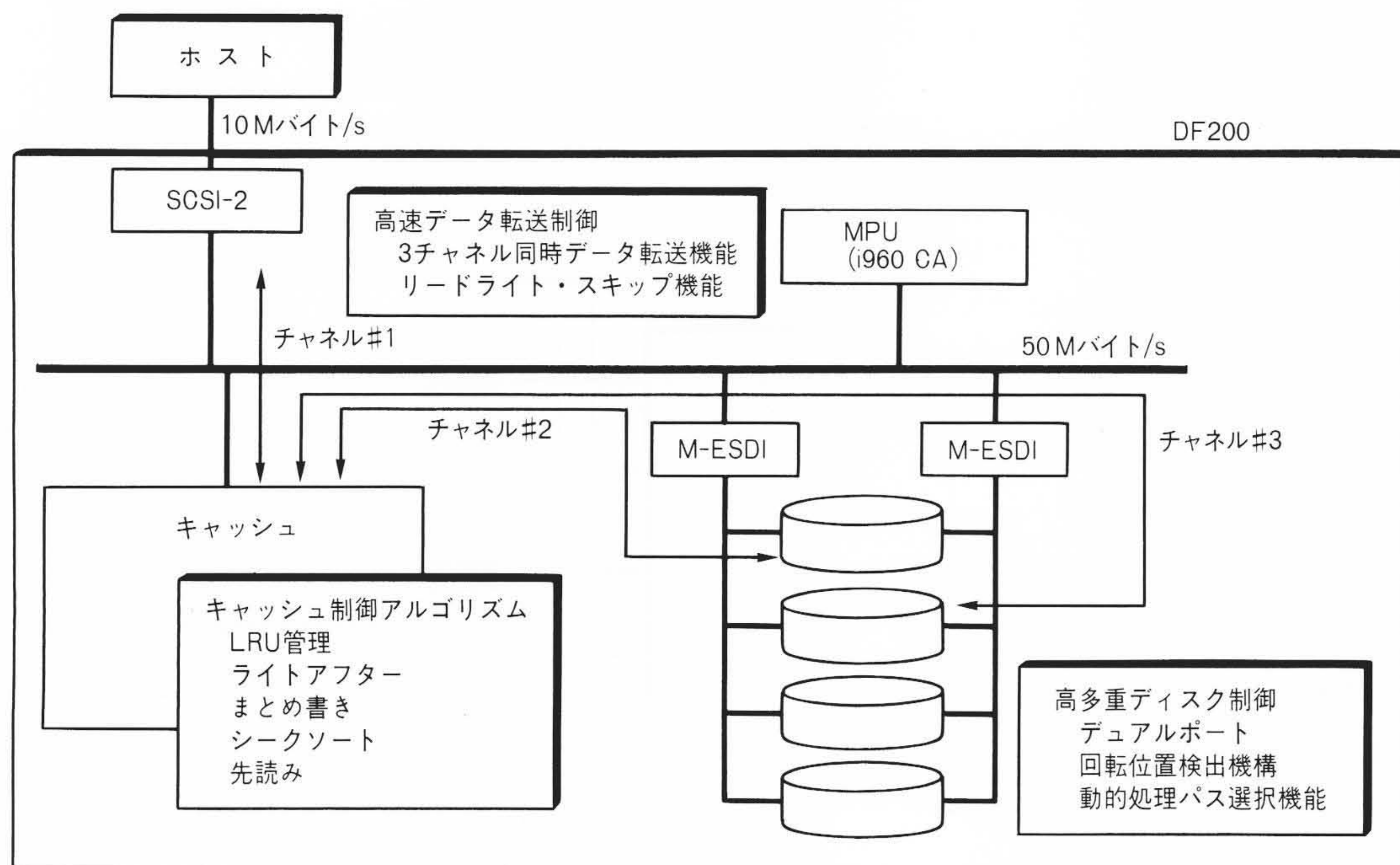
4 データ信頼性の保証

4.1 キャッシュ不揮発化

バッテリーバックアップ機構により、動作中の電源異常、電源切断に対し、キャッシュ内容を最低1週間保持する。その後電源回復、再度パワーオン時に、自動的に状態回復および未書込みデータのディスクへの強制書込みを行い、ユーザーデータを保証している。

4.2 データ保証チェックコード

高度で複雑なキャッシュ制御では、ユーザーデータの破壊に対する十分な保護が必要である。ユーザーデータはパリティ、ECC(Error Correction Code)のチェックコードで保護されると同時に、各ブロックごとにLRC(Longitudinal Redundant Code)および内部の論理的ア



注：略語説明
MPU (Micro Processor Unit)

図3 高性能処理を実現するための各機能
高スループット性能を実現するためのハード・ソフト機能を示す。

ドレス情報を付加したチェックコードを付加する機能を持たせた。このチェックコードでホストからのライトデータをキャッシュに書き込んでから、ディスクを通し再びキャッシュからホストに出るまでのすべての転送経路を保護し、データの信頼性を高めている。

5 性能の評価

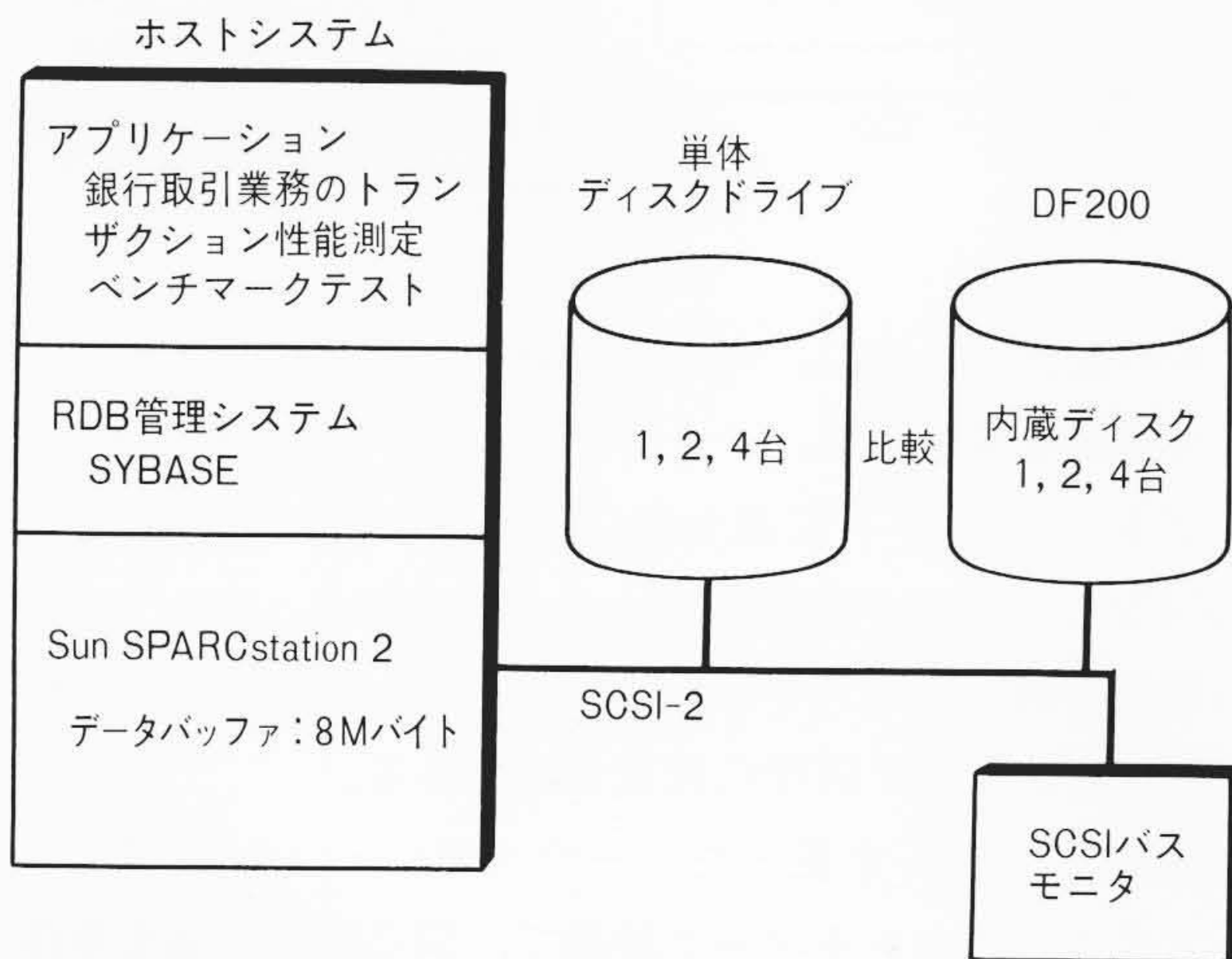
5.1 ベンチマークテストによるシステム性能評価

(1) 評価システムの構成

DF200を使用する代表的アプリケーションの一つに、RDB(リレーショナルデータベース)を使用したオンラインランザクション処理がある。その典型として、銀行の預金取引業務をモデルとしたベンチマークテストを利用し、システム性能全体としての評価を行った。

テスト環境を図4に示す。ホストにはSun Microsystems社のSPARCstation 2^{※4)}を使用し、RDB管理ソフトにSYBASE^{※5)}を使用した。ベンチマークテストプログラムは、銀行での預金データベースの作成・更新・残高照会業務をシミュレートしたものを使用し、ネットワーク通信処理を除いたデータベースシステム自身の処理性能を評価するようにした。

性能比較の対象として、DF200の内蔵ディスクとほぼ同性能でSCSIインタフェースを持つ日立製作所製の単



注：略語説明 RDB (リレーショナルデータベース)

図4 リレーショナルデータベースシステムでの性能評価ベンチマークテストの構成 単体ディスクドライブとDF200の比較を行った。ディスクは1台、2台、4台と変えて評価した。

※4) SPARCstation 2は、米国Sun Microsystems社の製品名称である。SPARCは米国SPARC International, Inc.の商標である。

※5) SYBASEは、米国Sybase Inc.の商標である。

体5インチ磁気ディスク、DK517Cを使用した。データベース件数は160万件、ホストワークステーション上のデータバッファに8Mバイトを割り当てた。またディスクへのI/O動作をトレースする目的で、SCSIバス上にI/Oモニタツールを接続した。

(2) ベンチマークテスト結果

このベンチマークテストの結果を図5に示す。データベースシステムの処理性能であるTPS(Transaction Per Second: 1秒間のランザクション処理回数)値で、単体ディスク比で2.1倍(ディスク1台構成)から1.3倍(ディスク4台構成)のTPS性能値の向上が得られる結果を得た。

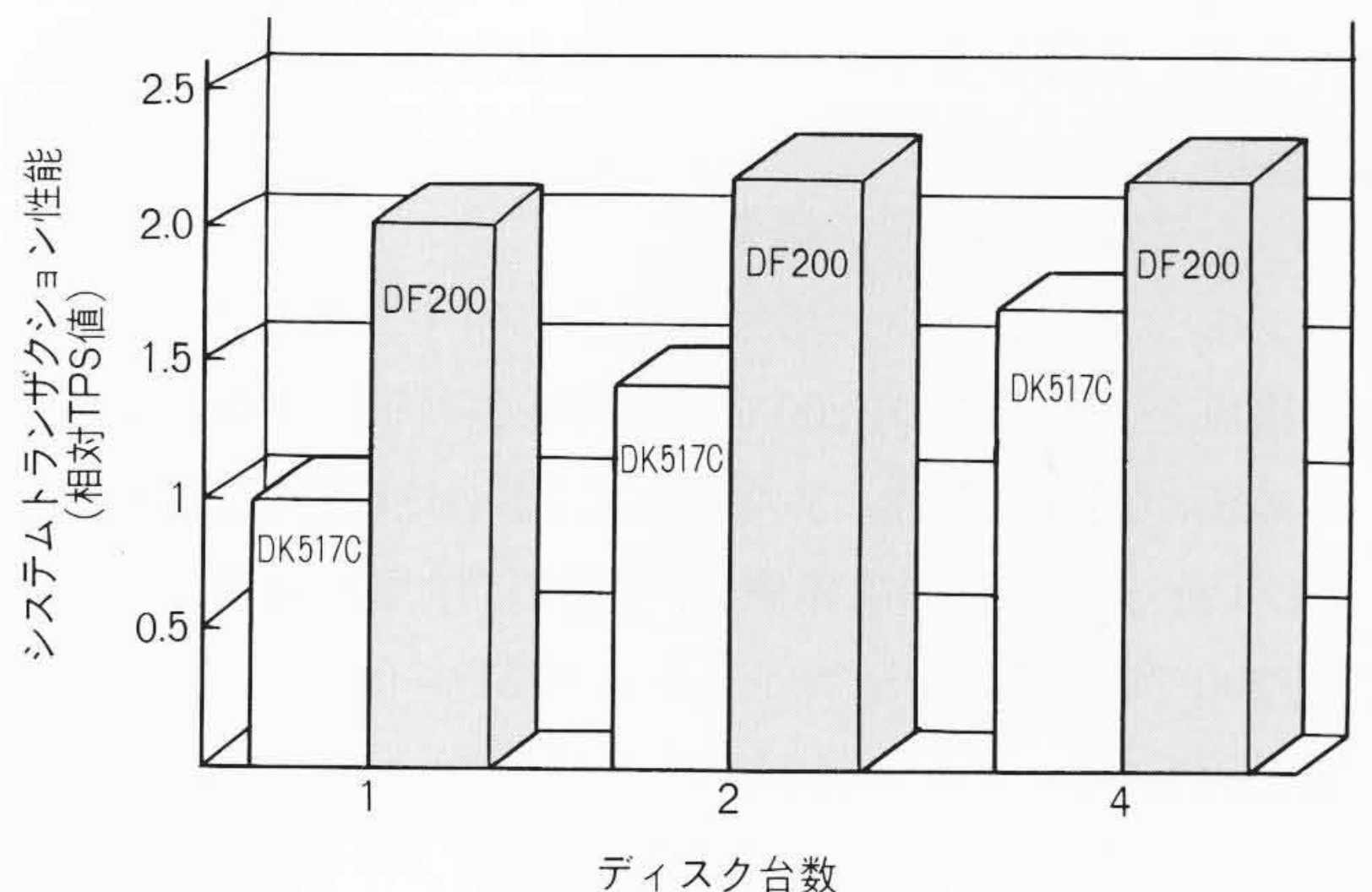
ベンチマークテスト時に接続したディスクI/Oモニタの採取データから、データベース・管理ソフトが行う障害時の復旧を可能とするためのログファイルの書き込み処理時間が、DF200のキャッシュによって大きく短縮され、性能向上に寄与しているとの結果を得た。

5.2 シミュレーションによる性能評価

(1) シミュレーションシステム構成

DF200自身の限界性能を、シミュレーションで評価した。使用したシミュレーションシステムの構成を図6に示す。

シミュレーションはGPSS(General Purpose Simulation System)言語を使用した待ち行列解析型のモデルに、キャッシュ機能(ヒット率、まとめ書きなどの高速化アルゴリズム)のモデルを付加したものである。また、実際のシステム環境に近づけるため、前述のベンチマーク

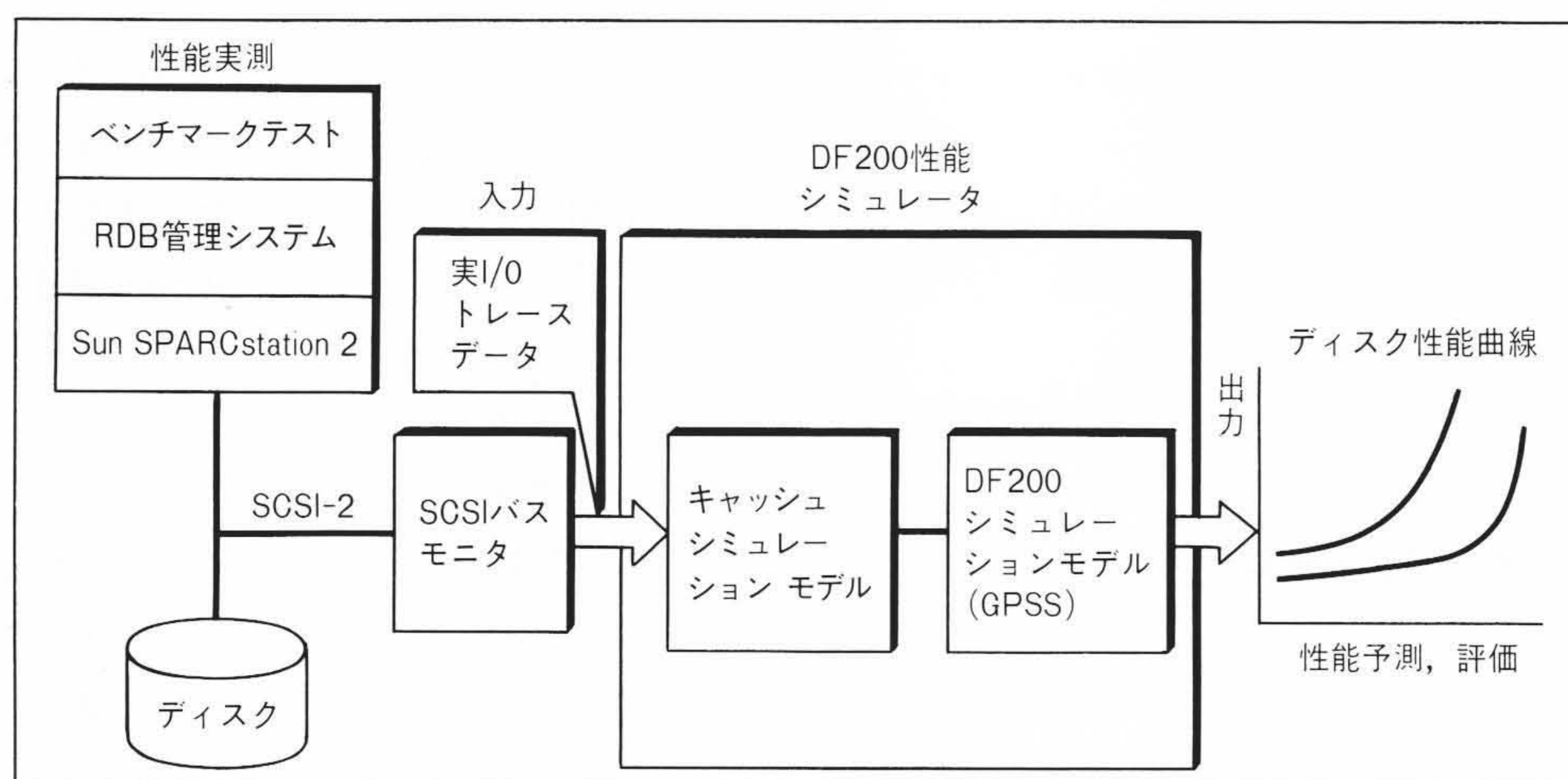


注：略語説明

TPS (Transaction Per Second: 1秒間のランザクション処理回数)

図5 ベンチマークテストによるシステム性能評価の結果

DF200は単体磁気ディスク構成との比較でTPS値が2.1倍(ディスク1台構成)から1.3倍(ディスク4台構成)の範囲でそれぞれ向上した。



注：略語説明

GPSS (General Purpose Simulation System)

図6 シミュレーションによる性能評価システムの構成 実I/Oのトレースデータを使用し、キャッシュを含む性能を評価する。

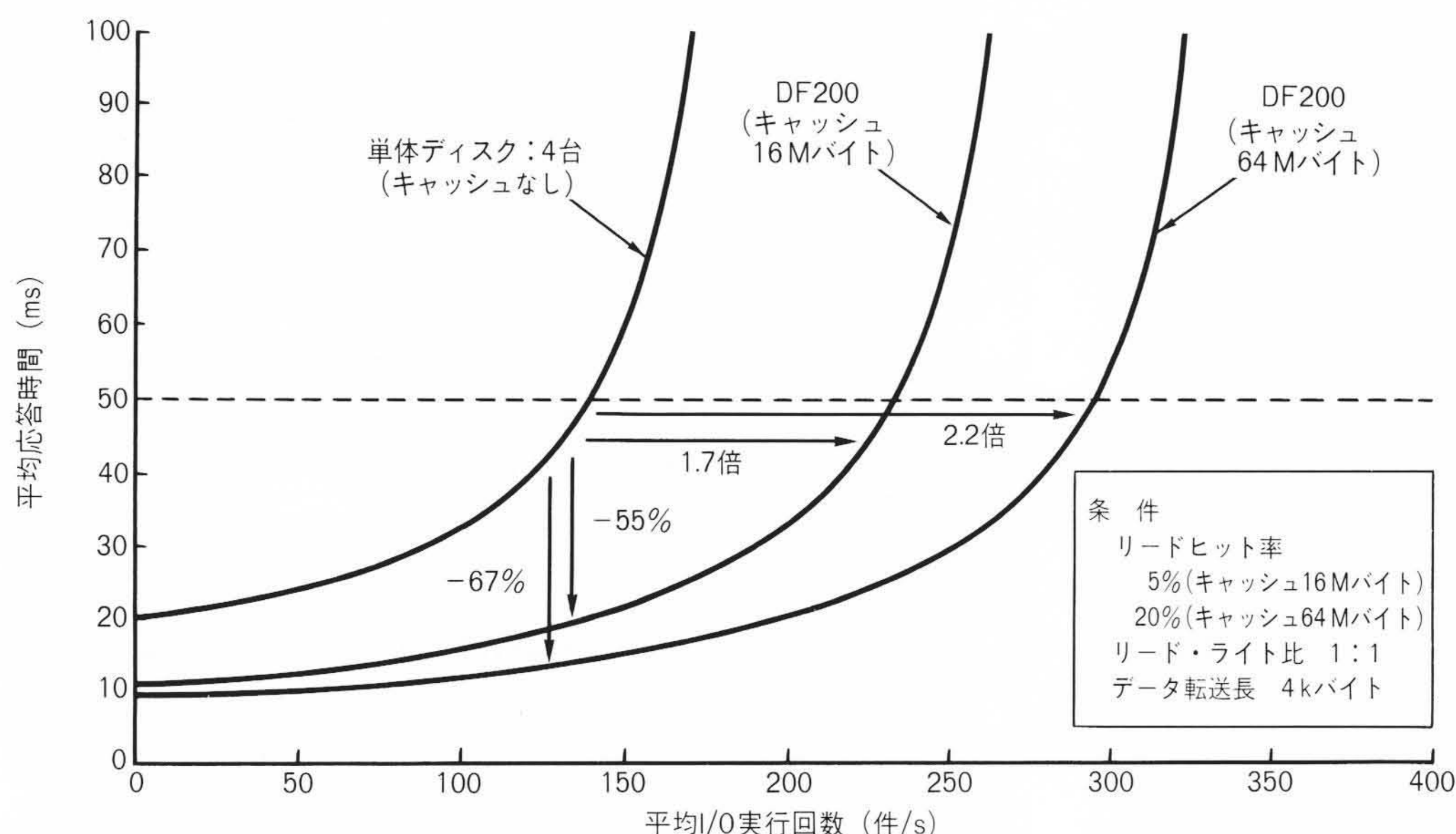


図7 シミュレーションによるDF200の性能評価結果 I/O実行回数の限界性能(平均応答時間50 msのとき)は、単体ディスク構成に比べて1.7倍から2.2倍に向上する。また、平均応答時間でも50%以上の改善が期待できる。

テストでI/Oモニタで採取したディスクのアクセスパターンデータをシミュレーションの入力データとして使用した。

(2) シミュレーション結果

シミュレーションの結果を図7に示す。単体ディスクの構成と比較してDF200では平均応答時間で約50%以上の短縮が期待できる。またディスクI/O処理性能(IOPS: I/O Per Second)の限界値(平均応答時間50 msのとき)はDF200で約230 IOPS(キャッシュ16 Mバイトのとき)～約300 IOPS(キャッシュ64 Mバイトのとき)となり、単体ディスクによる構成時に対し約1.7倍から2.2倍の向上が期待できる。

6 おわりに

ワークステーションなどのデータベースサーバで代表される分野で必要とする高トランザクション性能のファ

イルの提供を目的に、今回DF200高性能キャッシュディスク装置を開発した。コンパクトなディスク サイド キャビネット サイズに最大11.4 Gバイトの記憶容量と最大64 Mバイトの不揮発性キャッシュを持ち、高いデータ信頼性を持つ。ホストからはキャッシュを意識せず、従来の単体ディスク同様に接続使用できる。

RDBシステムを使ったワークステーション環境でのシステム性能ベンチマーク評価で、同じ記憶容量を単体ディスクで構成した場合に比較し、システムの処理性能のTPS値が1.4倍から2.2倍に向上する結果を得て、この製品の有効性が検証できた。

今後、種々のシステムで、ユーザーのより実環境に近い状態での性能評価を行い、いっそうの機能および性能面での強化ならびに拡充を図っていく考えである。