マルチポート・ページメモリをキャッシュとして用いた高性能ディスク・サブシステムのアーキテクチャと処理性能†

岡田義広†† 田中 譲††

マルチポート・ページメモリ (MPPM) をディスク・キャッシュとして用いることにより，スケーラブルな性能が得られるディスク・サブシステム (DIMP) を提案する．従来のキャッシュメモリ付きディスク・サブシステムでは，キャッシュのヒット率が低い場合には，頻繁にディスク・ドライブへアクセスしなければならない．ディスク・ドライブとディスク・キャッシュ間のデータ転送がボトルネックとなる．また，RPS (Rotational Positioning Sensing) ミスによる回転待ちが頻繁に生じ処理性能を低下させる．DIMP では，MPPM をキャッシュメモリとして用いることにより，上記のボトルネックを解消できる．MPPM を用いたことにより，データ転送時のオーバヘッドが小さい．高トランザクションに対して実用的な応答時間で処理でき，チャネルの転送幅に応じた高いスループットが得られる．本論文では，このディスク・サブシステム DIMP の構成と動作機構について解説する．シミュレーションによって得た処理性能を挙げて，DIMP の有効性を示す．

1. はじめに

近年の計算機システムでは，大量のデータを処理可能となった．これは，中央処理装置の処理速度の高速化と記憶装置の大容量化による．二次記憶装置であるディスク・サブシステムの大容量化では，ディスク媒体自体を高密度化して大容量化を図るほかに，ディスク・ドライブ装置を複数用いて大容量化を図る傾向にある．

ディスク装置は，機械的な動作が必要である．中央処理装置や一次記憶装置に比べて，処理速度が遅い．そのために，計算機システム全体の処理性能が悪くなる．そこで，処理装置間（中央処理装置とディスク装置間）での処理速度の差を緩和するために，緩衝メモリとしてディスク・キャッシュ[1)-3)]が用いられる．

本論文では，高スループットを達成するために，ディスク・キャッシュとして，マルチポート・ページメモリ (MPPM)[4)]を用いたディスク・サブシステム DIMP (DIsk subsystem using MPpm) を提案する．MPPM をディスク・キャッシュとして用いることにより，ディスク・ドライブとディスク・キャッシュ間のパス幅を増やすことができ，高スループットが期待できる．

従来のキャッシュメモリ付きディスク・サブシステムでは，スループットおよび応答時間がキャッシュのヒット率に依存する．キャッシュのヒット率が低い場合には，ディスク・ドライブ装置とディスク・キャッシュ間のデータ転送が頻繁に起こり，これがオーバヘッドとなりスループットを悪化させる．また，キャッシュのヒット率が低い場合には，RPS (Rotational Positioning Sensing) ミスによる回転待ちが頻繁に生じ処理性能を低下させる．RPS ミスによる回転待ちをなくし応答時間を向上させる研究として，B-DISK[5),6)]がすでに提案されている．応答時間を向上させる他の研究として，集合ディスク[7),8)]がある．ディスク・ドライブ装置に格納するデータを分割して，それぞれ別々のディスク・ドライブ装置に格納することにより，並列性を上げてデータの転送時間を小さくし，応答時間を向上しようとするものである．マルチポート・ページメモリをディスク・システムに適用した研究としては，マルチポート・ページメモリをバッファ・メモリとして用いた知識ベース・マシンの研究[4),9)]がある．マルチポート・ページメモリを用いたことにより，ディスク・ドライブ装置とのデータ転送におけるオーバヘッドが小さく，高性能が得られる．

中央処理装置の処理速度の高速化と記憶デバイスの高密度化は，今後ますます図られると思われる．ディスク装置の記憶容量も，ますます増加することになる．記憶容量の増加に伴いディスク装置に格納されているデータの共有性が増す．したがって，多数の上位処理装置から，同時にアクセス可能である必要がある．そこで，高スループットが得られるディスク・サブシステムのアーキテクチャについて研究を行った．

† A High-Performance Disk Subsystem Architecture Using MPPM (Multi-Port Page Memory) as Its Cache Memory and Its Performance Evaluation by YOSHIHIRO OKADA and YUZURU TANAKA (Electrical Engineering Department, Faculty of Engineering, Hokkaido University).

†† 北海道大学工学部電気工学科

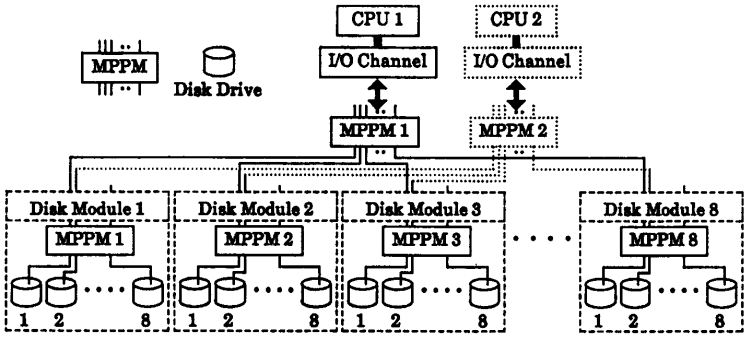

図 1 MPPM を階層的に用いたディスク・サブシステムの構成
Fig. 1 The configuration of a disk subsystem using MPPMs.

図 1 に示すのが DIMP の構成例である．MPPM をディスク・キャッシュとして用いることにより，十分な本数のパス（ディスク・ドライブとディスク・キャッシュ間パスと，ディスク・キャッシュとチャネル間パス）を得ることができる．チャネルの転送幅によって，あるいは，ディスク・ドライブの処理速度によって制限されるだけの高い処理性能（スループット）が得られる．図 1 は，二階層の例であるが，種々の階層化が可能である．多数の入出力ポートをもつ MPPM を 1 個用いた一階層構成も考えられる．本論文では，DIMP の構成と動作機構について述べる．また，シミュレーションにより得られた処理性能を示し DIMP の有効性を示す．

以下では，第 2 章で，まず従来のキャッシュメモリ付きディスク・サブシステムのボトルネックについて簡単に説明する．また，MPPM をディスク・キャッシュとして用いる効果について述べる．第 3 章では，DIMP のアーキテクチャについて解説する．ここで，DIMP の基本的な構成例とキャッシュ動作を示す．第 4 章では，従来型ディスク・サブシステムの処理性能を示すとともに，シミュレーションにより得られた DIMP の処理性能を示し，DIMP の有効性を述べる．第 5 章でまとめを述べる．

2. 従来型ディスク・サブシステムのボトルネック

本章では，従来のキャッシュメモリ付きディスク・サブシステムの構成と基本的なキャッシュ動作図を示し，ディスク・ドライブとディスク・キャッシュ間のパス（以下 D-C パスと呼ぶことにする）がボトルネックであることを指摘する．また，MPPM をキャッシュメモリとして用いることによる効果について解説する．

2.1 従来型ディスク・サブシステムの構成

図 2 に示すのが従来のキャッシュメモリ付きディスク・サブシステムの構成である．複数のディスク・ドライブ装置は，数本（4 本程度）のパスでディスク・コントローラ中のディスク・キャッシュと接続され，このパス（D-C パス）を介してデータ転送を行う．I/O チャネルも，数本（4 本程度）のパスでディスク・コントローラ中のディスク・キャッシュと接続され，このパス（C-C パスと呼ぶことにする）を介してデータ転送を行う．

2.2 D-C パスのオーバヘッド

基本的なキャッシュ動作として，ライトスルー方式が一般的である．**図 3** に，その動作を示す．キャッシュでリードミスが起きた場合には，ディスク・ドライブから当該レコードを転送する．一般に，読み込まれたレコードの前後のデータは，この後すぐに読み込まれる確率が高い．したがって，図 3 で示されるように，キャッシュでリードミスが起きた場合には，通常ディスク・ドライブからキャッシュメモリへ当該レコードを含む 1 トラック分（あるいは数レコード分）のデータを転送する．ところが，キャッシュのヒット

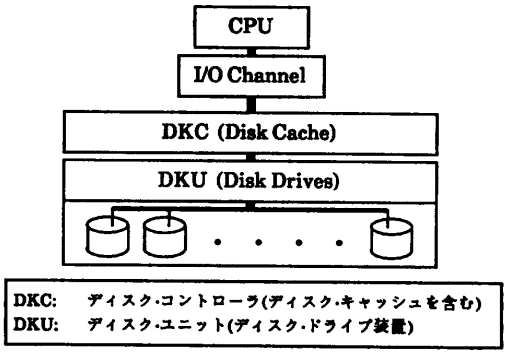

図 2 キャッシュメモリ付きディスク・サブシステム
Fig. 2 The disk subsystem with a disk cache.

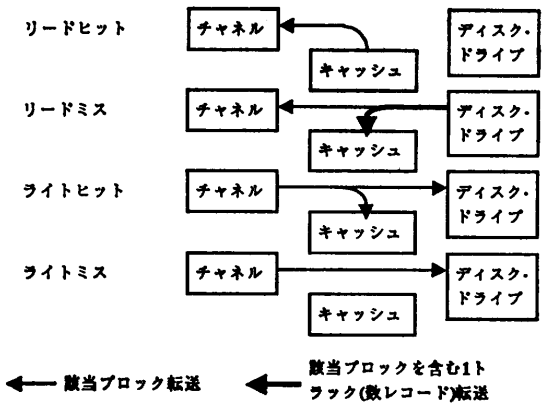

図 3 ディスク・サブシステムのキャッシュ動作
Fig. 3 The disk cache behavior of conventional disk subsystems.

率が低い場合には，転送された1トラック分のデータ中，当該レコード以外は，この後読み込まれる割合が低く，無駄なデータである．この無駄なデータ転送が頻繁に行われるため，これがオーバヘッドとなり，スループットを悪化させる．したがって，キャッシュのヒット率が低い場合にも，高スループットを得るためには，上記のオーバヘッドに勝るだけ十分大きなD-Cパス数が必要である．

2.3 RPS ミスによる一回転待ち

ディスク・ドライブ装置がシーク動作と回転待ちをし，データ転送が可能になったにも関わらず，D-Cパスが他のディスク・ドライブのデータ転送中で，使用可能なパスがない場合には，このディスク・ドライブはデータ転送が行えない．これを RPS ミスという．RPS ミスが起きると，ディスク・ドライブは，さらに一回転するのを待たなければ，データ転送を行えない．ヒット率が低い場合には，ディスク・ドライブ装置へのアクセス回数が増え，D-Cパスの負荷が重くなる．これに伴い RPS ミスが頻繁に生じ，処理性能を低下させる．この点に関しては，ディスク・ドライブ装置内に小容量のバッファを設けた B-DISK により解決されている．ディスク・ドライブからデータを読み込む時，RPS ミスが起きた場合には，このバッファへデータをいったん転送しておく．D-Cパスが空いた時点で，このバッファからデータの読み込みが行われる．ディスク・ドライブへデータを書き込む時も，D-Cパスが空いた時点で，このバッファへいったんデータを転送しておく．ディスク・ドライブ装置が書き込み可能になった時点で，このバッファからディスク・ドライブにデータの書き込みが行われる．これにより，RPS ミスが起きても，一回転するのを待つ必要がなく，待ち時間を小さくでき，処理速度の向上が望める．

上記のように，RPS ミスは，多数（64 台程度）のディスク・ドライブ装置が数本（4本程度）の D-C パスを共用するために起こる．したがって，ディスク・ドライブ装置の台数と同数の専用の D-C パスをもつことで，RPS ミスを回避でき，性能の低下を抑えることができる．

以上述べたように，従来型ディスク・サブシステムでは，ディスクキャッシュとディスク・ドライブ間のデータ転送がボトルネックとなっている．したがって，この間のデータ転送幅を増やさなければ，著しいスループットの向上は望めない．

2.4 MPPM を用いる効果

図4に示すように，ディスク・サブシステムは，ディスク・ドライブ-(D-C パス)-ディスク・キャッシュ-(I/O チャネル)-主記憶装置というメモリの階層構造を成す．ディスク・ドライブ数は，将来，ますます増加すると予想される．ディスク・ドライブ台数が増えることにより，記憶容量が増す．それに伴って，スループットも向上するようなアーキテクチャが望ましい．

上述したように，スループットを向上させるためには，D-C パス数を十分に増やす必要がある．D-C パス数を増やす場合，従来型のディスク・サブシステムでは，ディスク・キャッシュがネックになる．**図5**に示されるのが，従来のディスク・キャッシュの構成である．複数のパスと接続されるため多重ポート化される．(入出力ポートの転送速度)×(入出力ポート数) のアクセス速度をもつ高速なメモリデバイスを用い，速度差を吸収するバッファと $n \times 1$ スイッチにより時分割制御することで，仮想的に多重ポート化している（n はポート数）．この構成では，各ポートのバッファの状態を監視し，効率良くスイッチの切り替えを行わ

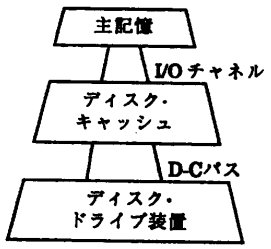

図 4 ディスク・サブシステムのメモリ階層
Fig. 4 The memory hierarchy of disk subsystems.

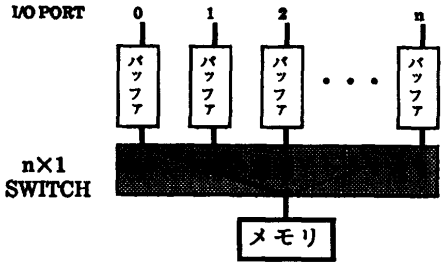

図 5 従来のキャッシュメモリの構成例
Fig. 5 The conventional cache-memory architecture.

なければならない．ポート数を増やす場合も，アクセス速度の速いメモリデバイスを用い，高速に制御を行わなければならない．特に，すべてのディスク・ドライブが専用のパスを介してディスク・キャッシュを独立にアクセスする構成では，ディスク・ドライブ数以上の入出力ポートが必要となる．この場合，上述のメモリ構成では実現不可能といえる．多数の入出力ポートをもち，すべての入出力ポートから独立にデータのアクセスが行えるメモリ構成が必要である．そこで，われわれは，MPPM をディスク・キャッシュとして用いることにした．

[MPPM の構成]

図 6 に示されるように，MPPM は，複数の入出力ポート(図の例では，8 ポート)をもち，ブロック単位でデータの入出力が行えるメモリデバイスである．各ポートは，スイッチング・ネットワークを介して，各メモリバンクと接続されている．各ポートは，常に，異なるメモリバンクをアクセスする．

スイッチング・ネットワークに与える制御コード(図の例では，3 ビット) は 0 から 7，0 から 7 と常に変化する．このとき，各ポートは，メモリバンクを(例えば，ポート 0 は，メモリバンクの 0 から 7，0 から 7 と) 順にアクセスする．図では，制御コードの値が 3 (011_2) の時の，入出力ポートとメモリバンクの接続を太線で示している．

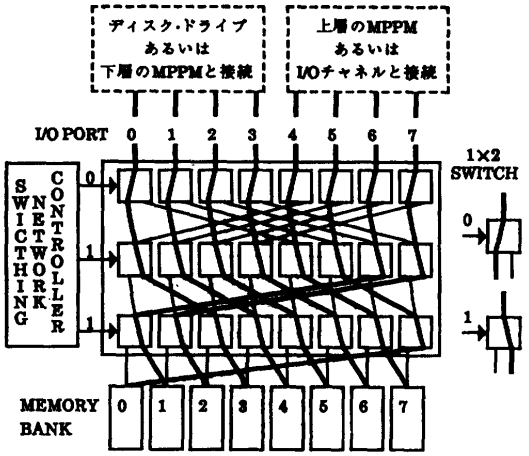

図 6 MPPM の構成例
Fig. 6 The MPPM architecture.

[MPPM の動作]

ポート 0 を例に，データの入出力を説明する．ポート 0 がメモリバンク 0 と接続された時点で，データの転送が開始される．この時の制御コードは，0 である．制御コードが 0 から 7 へ一巡する間 (これを 1 スライスと呼ぶ) に，ポート 0 は，順にメモリバンクの 0 から 7 へ接続される．1 スライスで，8 個のデータをアクセスする．この間，すべてのメモリバンクへ，各々アドレス m が与えられる．次の 1 スライスでは，すべてのメモリバンクへ，アドレス $m+1$ を与え，同様に 8 個のデータをアクセスする．ブロックサイズは，メモリバンク数の倍数でなければならない．仮に，ブロックサイズを 64 とすると 8 スライスでブロックのすべてのデータのアクセスが行える．8 スライス目で，メモリバンクへ与えられるアドレスは，$m+7$ となる．このようにして，1 ブロックのデータのアクセスが行われる．

同様に，ポート 1 は，制御コードが 7 の時に，メモリバンク 0 と接続され，データ転送が開始される．ポート 1 の場合の 1 スライスとは，制御コードが 701～6 と一巡する期間をいう．各ポートは，常に，異なるメモリバンクをアクセスするので，アクセス衝突はない．同一のブロックであっても，異なるポートから同時に (厳密には，数クロックずれて) アクセス可能である．このように，各ポートは，アクセス衝突なくブロック単位でデータの入出力が行える．

MPPM のデータの入出力では，ポート切り替えによる待ち時間を必要とする．通常これは，入出力ポート数に比例する．しかし，転送ブロックサイズに比べて，入出力ポート数が小さな場合には，著しいオーバヘッドではない．特に，ディスク・キャッシュとして用いる場合には，転送ブロックサイズが大きいので問題とならない．

先の図 1 に示したように，MPPM をディスク・キャッシュとして用い，複数あるディスク・ドライブ装置のそれぞれを MPPM の入出力ポートのそれぞれに接続することで，ディスク・ドライブ数と同数の D-C パスが得られる．各ディスク・ドライブ装置は，ディスク・キャッシュと専用のパスで接続されることにな

り，RPS ミスの発生を抑えることもできる．

3. DIMP のアーキテクチャ

この章では，DIMP の基本的な構成とキャッシュ動作について述べる．

3.1 DIMP の基本構成

先の図1に示すように，16の入出力ポートをもつMPPM と8台のディスク・ドライブ装置からディスク・モジュールを構成できる，16の入出力ポートのうち，8ポートがディスク・ドライブ装置との接続に使用される．残りの8ポートは，上層の MPPM との接続，あるいはチャネルとの接続に使用される．上位装置から見た場合，8台分のディスク・ドライブ容量があり，8本の入出力パスをもつディスク・ドライブ装置と同様に扱うことができる．さらに，16の入出力ポートをもつ MPPM と先のディスク・モジュール8個から新たにディスク・モジュール（64台分のディスク・ドライブ容量があり，8本の入出力パスをもつ）を構成できる．16の入出力ポートのうち，8ポートがディスク・モジュールとの接続にそれぞれ使用される．残りの8ポートは，上層の MPPM との接続，あるいはチャネルとの接続に使用される．MPPM は，複数の入出力ポートをもち，アクセス衝突なく，ブロック単位でデータの入出力が行えるメモリデバイスである．MPPM をディスクキャッシュとして用いることにより，各ディスク・ドライブは，衝突なくキャッシュメモリをアクセスできる．同様に，各ディスク・モジュールは，上位のキャッシュメモリを衝突なくアクセスできる．

また，図1の破線で示したように，上層の MPPM をさらに1個用いることで，チャネルと接続されるパス数を増やすことができる（これをデュアル・フレーム構成[3]と呼ぶ）．このように，比較的少ない数の入出力ポートをもつ MPPM を用いて，これを階層的に構成したディスク・サブシステムが DIMP である．MPPM を階層的に構成することで，システム拡張が容易に行える．また，デュアル・フレームのような構成上の工夫を適用しやすい．

図1に示した DIMP の構成は，二階層の構成であるが，種々の構成が考えられる．128 の入出力ポートをもつ MPPM を1個用い，64 の入出力ポートを 64 台のディスク・ドライブ装置のそれぞれと接続し，残りの 64 の入出力ポートをチャネルとの接続パスとして用いた一階層の構成も考えられる．ただ，システム拡張が容易に行え，構成上の工夫を適用しやすいという点では，図1に示した構成が良いと，著者らは考えている．

いずれの階層構成の場合にも，ディスク・ドライブ装置と接続される一番下層の MPPM のメモリ容量を大きくし，これにディスク・キャッシュの機能をもたせる．従来のディスク・キャッシュと同程度のキャッシュ容量を仮定すると，数 GB の容量のディスク・ドライブ装置を用いる場合，図1の構成の最下層の MPPM（8台のディスク・ドライブ装置が接続される）のメモリ容量は，数十 MB となる．また，上層の MPPM もキャッシュとして機能するが，これを挟む上・下層の MPPM とのデータ転送のためのバッファとしての機能を期待しており，比較的小さな容量で良い．データ転送単位についても，キャッシュ・ミスが起きて，ディスク・ドライブ装置から最下層の MPPM へデータを読み込む場合のみ，トラック単位のデータ転送とし，他のデータ転送は，すべてブロック単位とする．

3.2 ポート切り替えにより起こる待ち時間の解消

DIMP は，MPPM による階層構造を成す．先述したように，MPPM のデータの入出力は，バンク番号が0のメモリバンク[4]を基準に行われる．よって，MPPM 同士の入出力ポートの接続では，間にバッファを入れる必要がある．しかし，MPPM 間の入出力ポートの接続関係を考慮して，同期した制御コードをそれぞれの MPPM のスイッチング・ネットワークに与えることで，バッファを入れなくてもよい．また，バッファでの待ち時間を解消できる．**図7**に示した，3つの場合について説明する．

（1） 上層の MPPM 1 のポート0に下層の MPPM 2 のポート4を，ポート1に別の下層の MPPM 3 のポート4を接続した場合を考える．上層の MPPM 1 の制御コードを 012…と与えた場合，ポート0が接続するメモリバンクは，012… となり，ポート1が接続するメモリバンクは，123… となる．したがって，下層の MPPM 2 のポート4が接続するメモリバンクが 012… となるためには，制御コードを 456… とすればよい．同様に，下層の MPPM 3 のポート4が接続するメモリバンクが 123… となるためには，制御コードを 567… とすればよい．

（2） これは，下層の MPPM に対する制御コードを統一した場合である．上層の MPPM 1 のポート0は，下層の MPPM 2 のポート4に，ポート1は，下

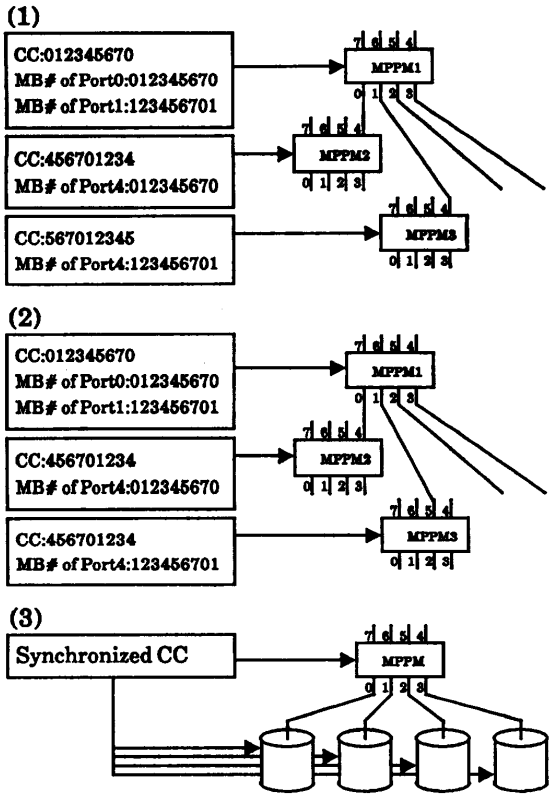

図 7 MPPM のポート接続と制御コード
Fig. 7 The relationship of the connection between MPPM's I/O ports and MPPM's control code.

層の別の MPPM 3 のポート 5 に接続すればよい.

(3) ディスク・ドライブからその上層の MPPM へのデータ転送は，1 トラックごとに行われるものとする．よって，各ディスク・ドライブを同期して回転させ，これと MPPM へ与える制御コードを同期させることにより，バッファ（待ち時間）なしで，データ転送を行うことが可能であると思われる．この点については，ディスク・ドライブ台数が多い場合には，制御が難しくなると思われ，検討が必要だと考えている．

3.3 DIMP のキャッシュ動作

DIMP では，MPPM をキャッシュメモリとして用いる．キャッシュの動作方式は，一般のキャッシュメモリと同様に，ライトスルー方式とライトバック方式が考えられる．ここでは，それぞれの方式について簡単に説明する．それぞれの場合の評価性能については，後の章で解説する．

3.3.1 ライトスルー（ライトアフタ）方式

図 8 に示すのが，ライトスルー方式の基本動作である．リードヒットの場合には，即座にデータの読み込みが行える．リードミスの場合には，下層のドライブ

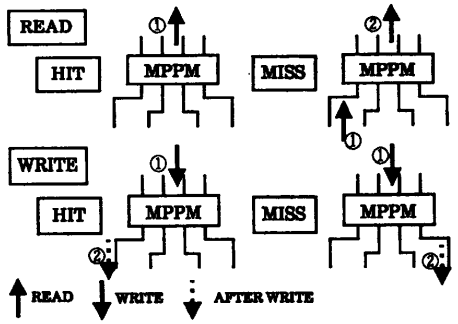

図 8 DIMP のキャッシュ動作（ライトスルー方式）
Fig. 8 The cache behavior of a DIMP (Write-Through protocol).

あるいは MPPM から当該データを読み込み，いったん MPPM にデータを書き込む．これと同時にデータの読み込みが行える．ドライブから読み込む場合には，当該データを含む 1 トラック分を読み込む．ライトの場合には，必ずディスク・ドライブへ当該データを書き込む．まず，いったん MPPM にデータを書き込む．下層の MPPM に対するポートが使用中の場合には，ポートが空くのを待った後，下層のドライブあるいは MPPM にデータを書き込む（ヒット・ミスに関係なく，書き込みデータは，必ずいったん MPPM に書き込まれる．よって，ライトアフタと呼ばれる方式に近い）．この操作が繰り返されて，書き込みデータは，次第に下層に移動し，最後にディスク・ドライブに書き込まれる．ディスク・ドライブに対する書き込みは，当該ブロックのみである．書き込みデータが移動した通り道となる MPPM には，LRU (Least Recently Used) 規則により捨てられるまで，このデータが残り続ける．したがって，書き込みデータが下層へ移動中に，このデータに対する読み込み要求が来た場合でも，上層の MPPM から即座に読み込みが行える．

3.3.2 ライトバック方式

図 9 に示すのが，ライトバック方式の基本動作である．リードヒットおよびライトヒットの場合には，即座に読み込みおよび書き込みが行える．リードミスの場合には，下層のドライブあるいは MPPM から当該データを読み込まなければならない．そのためには，要らないブロックを MPPM から追い出す必要がある．したがって，まず，LRU 規則にしたがいブロックを追い出す．このとき，それらが更新されたブロックならば，下層のドライブあるいは MPPM に書き戻す．ライト・ミスの場合も同様に，追い出されるブロックが更新されたブロックならば，書き戻しの処理

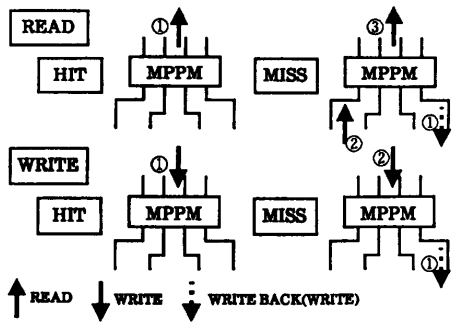

図 9 DIMP のキャッシュ動作（ライトバック方式）
Fig. 9 The cache behavior of a DIMP (Write-Back protocol).

を行う．これらの処理は，各層の MPPM で独立に行われる．ディスク・ドライブから上層の MPPM へのデータ転送は，当該ブロックを含む1トラックごとに行われる．ディスク・ドライブに対する書き込みは，更新された当該ブロックごとに行われる．

3.3.3 固定長レコードと可変長レコード

ワークステーションやパーソナルコンピュータ等では，ディスク装置等の I/O 機器との接続のために SCSI (Small Computer System Interface) バスが用いられる．一般に，SCSI バスに接続されたディスク装置では，固定長レコードを扱う．固定長レコードの場合，ディスク装置に格納される領域が固定されており，データの書き込みに対して，そのデータのディスク上での領域が存在しないということはない．一方，汎用大型機に，I/O チャネルを介して接続されるディスク・サブシステムでは，CKD フォーマット[3]と呼ばれる可変長レコードを扱うものがある．この場合には，ディスク・ドライブ装置に格納される領域が不定であり，ディスク上に存在しない新しいデータの書き込みに対しては，まず，ディスク上に領域を確保する必要がある．よって，常にキャッシュにデータを書き込むという動作が行えない．したがって，CKD フォーマットと呼ばれる可変長レコードの場合には，上述したキャッシュ動作をそのまま用いることができない．

本論文では，固定長レコードを扱うディスク装置を対象として考えており，CKD フォーマットと呼ばれる可変長レコードの扱いについては，今後検討するつもりである．

4. シミュレーションによる評価と解析

DIMP の性能を見るためにシミュレーションによる性能評価を行った．また，従来型ディスク・サブシステムの性能評価を併せて行った．本章では，これらの評価結果を挙げて，DIMP の有効性を示す．

4.1 従来型ディスク・サブシステムの処理性能

4.1.1 ディスク・ドライブの性能

表1に，シミュレーション対象とするディスク・ドライブの性能を示す．これらの値は，文献3)中に示されているもので，実際に使用されているディスク・ドライブの性能である．したがって，得られるシミュレーション結果は，現実的な値である．ディスク・ドライブ装置にアクセスする場合，常にシーク動作が行われるわけではない．同一トラック上のデータであれば，シーク動作は行われない．表中のランダム度とは，シーク動作が行われるか否かを確率として与えたものである．

4.1.2 シミュレーション条件と結果

表2に，シミュレーション条件とシミュレーション結果を示す．実際に使用されているディスク・サブシステムの入出力リクエストのトレース・データを用いることで，より現実的に性能評価が行える．しかし，そのようなデータを得ることができないため，入出力リクエストは確率分布にしたがうものとした．待ち行列網モデルを基本とした性能評価シミュレータ CAB[10] を用いて，シミュレーションを行った．単位時間当たりの I/O 数と応答時間（TAT）の関係を求めた．

表 1 ディスクドライブの性能
Table 1 The specification of a disk-drive model.

[回転速度]	
一回転時間	: 16.7 msec
回転待ち時間	: 平均 8.3 msec
シーク時間	: 平均 12.5 msec
ランダム度	: 1/3
転送速度	: 3,000 KB/sec
[転送時間]	
リードヒット，ライト時	: 3 msec/BLK
リードミス時	: 18.4 msec/TRK
(1 BLK＝4 KB 固定/1 TRK＝1 トラック)	

表 2 シミュレーション条件
Table 2 The simulation conditions.

(1)	I/O の到着間隔はランダム（指数分布）である．
(2)	リード：ライト比を 4：1 固定とする．
(3)	転送ブロック長を 4 KB 固定とする．
(4)	リードミス時には，当該データを含む1トラックを読み込むものとする．
(5)	キャッシュとチャネル間のデータ転送時間も 3 msec/ブロックとする．
(6)	I/O の発行件数は，20,000 件とする．

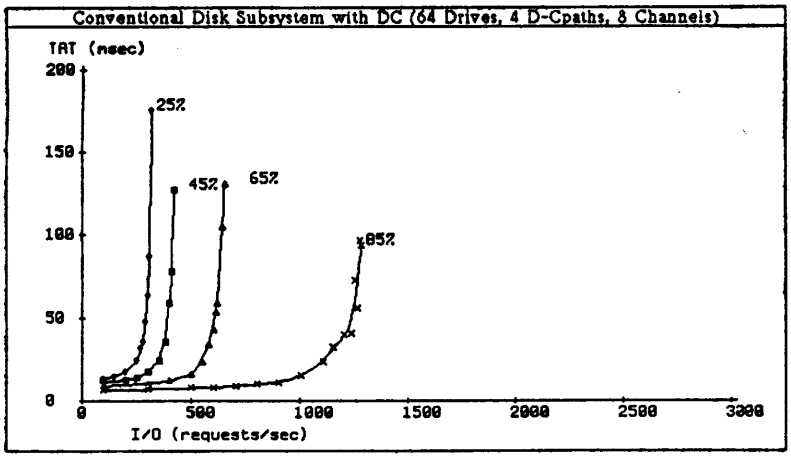

図 10 DC のみのディスク・サブシステムの性能
Fig. 10 The performance of a disk subsystem with DC.

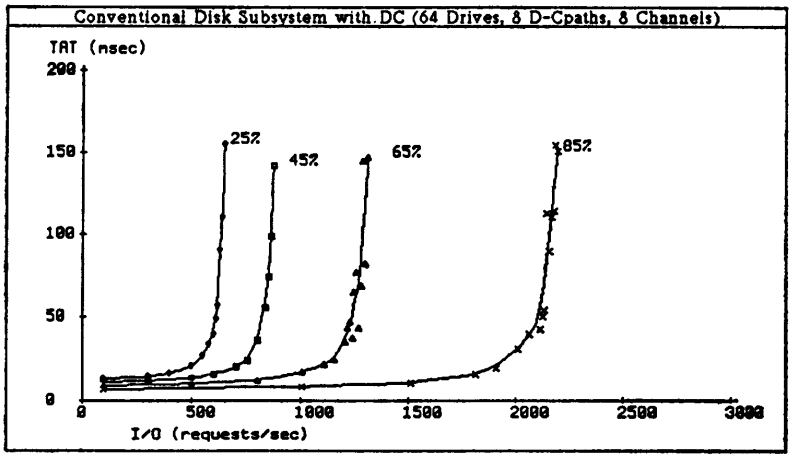

図 11 DC のみのディスク・サブシステムの性能
Fig. 11 The performance of a disk subsystem with DC.

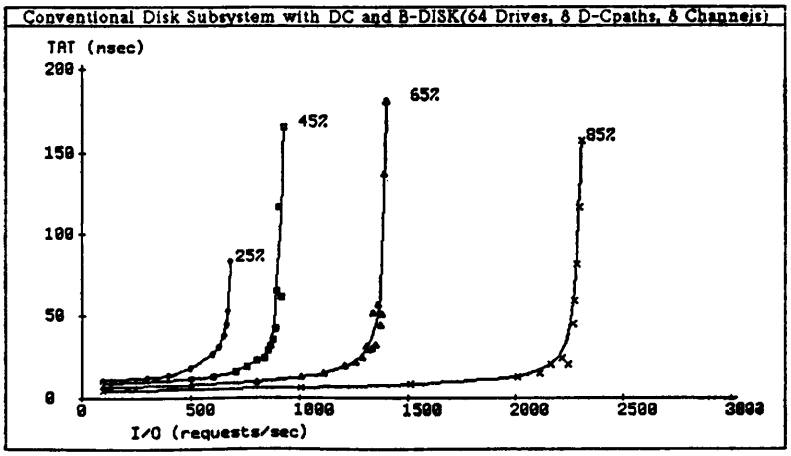

図 12 DC と B-DISK のディスク・サブシステムの性能
Fig. 12 The performance of a disk subsystem with DC and B-DISK.

DC のみをもち，ディスク・ドライブ数64台，D-C パス数4，チャネルパス数8のディスク・サブシステムを構成1とする．構成1に対して，D-Cパス数を8にしたディスク・サブシステムを構成2とする．DCと B-DISK の両方を備え，ディスク・ドライブ数64台，D-Cパス数8，チャネルパス数8のディスク・サブシステムを構成3とする．**図10**，**図11**，**図12**は，それぞれ構成1，2，3について，キャッシュのヒット率を25，45，65，85%と変えた結果のグラフである．

4.1.3 考　　察

図10，図11，図12より，どの構成においても，キャッシュのヒット率が低い場合には，スループットが著しく悪いことが分かる．

表3は，応答時間（TAT）が50 msec の時のスループット（単位時間当たりの I/O 数）を，図10，図11，図12のグラフから比較したものである．どのヒット率においても，D-Cパス数が4（表中[1]）の場合に比べて D-C パス数が8（表中[2]）の場合は，ほぼ2倍のスループットが得られている．したがって，ディスク・ドライブとディスク・キャッシュ間のデータ転送がボトルネックであると分かる．また，D-Cパス数が8で B-DISK を用いない（表中[2]）場合と，D-Cパス数が8で B-DISK を用いた（表中[3]）場合を比べてみると，スループットは，10%程度しか向上していない．以上から分かるように，スループットを向上させるためには，ディスク・ドライブとディスク・キャッシュ間のデータ転送幅を増やす必要がある．

表4は，各構成について，低い

表3 TATが50 msecの時のI/O件/秒
Table 3 I/O requests/sec (TAT=50 msec).

ヒット率 (%)	25	45	65	85
[1] 構成1 (件/秒)	300	400	620	1230
[2] 構成2 (件/秒)	620	840	1260	2120
[3] 構成3 (件/秒)	690	910	1410	2310
割合 ([2]/[1])	2.06	2.10	2.03	1.72
割合 ([3]/[2])	1.11	1.08	1.11	1.09

表4 I/Oが100件/秒の時のTAT
Table 4 TAT (msec) (I/O=100 (requests/sec)).

ヒット率 (%)	25	45	65	85
[1] 構成1 (msec)	13.86	11.53	9.30	7.08
[2] 構成2 (msec)	13.47	11.38	9.30	7.08
[3] 構成3 (msec)	11.07	8.98	6.76	4.66

トランザクション域（スループットが100件/秒）での応答時間を示したものである．応答時間は，キャッシュのヒット率が大きくなるにしたがって，値が小さくなっている．B-DISKを用いた場合には，ディスク装置に対するデータ転送をディスクの回転とは非同期に行うことができ，B-DISKを用いない場合に比べて，応答時間が良くなっているのが分かる．

4.2 DIMPの処理性能

ドライブの性能および転送速度は，4.1節と同様である．MPPM間のデータ転送もすべて，データ長を4 KB/ブロック（転送時間3 msec）固定とした．シミュレーションに用いたDIMPの構成は，先の図1に示したものである．ディスク・ドライブ数が64台，チャネル数が8で，先に示した従来型のディスク・サブシステムのシミュレーションに用いたモデルに対応するものである．I/Oの発行件数を50,000件としてシミュレーションを行った．図13, 14に示すのが，キャッシュ動作をそれぞれライトスルー，ライトバックにした場合の結果のグラフである．ライトバックの場合，キャッシュミス時に，更新されたデータはディスクに書き戻す必要がある．更新されたデータである割合は，I/Oのライト比と同じ0.2とした．キャッシュのヒット率（個々のMPPMにおけるヒット率）をそれぞれ10%，30%，50%（1個の等価なMPPMに置き換えた場合には，19%，51%，75%）とした場合の結果である．

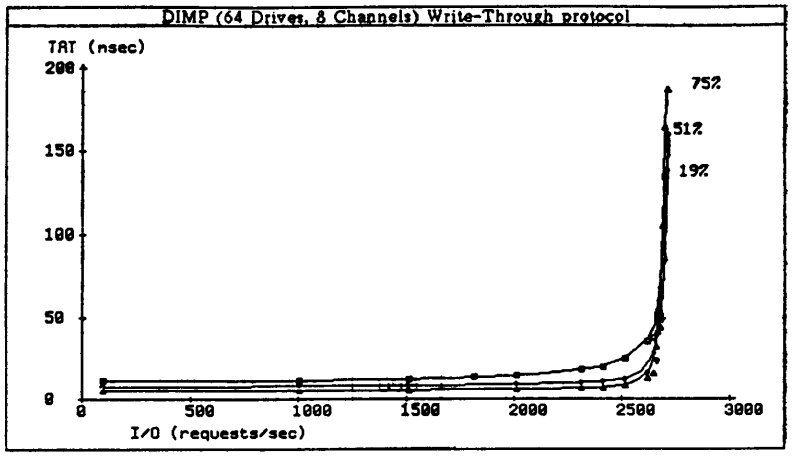

図13 DIMPの性能（ライトスルー方式）
Fig. 13 The performance of a DIMP (Write-Through protocol).

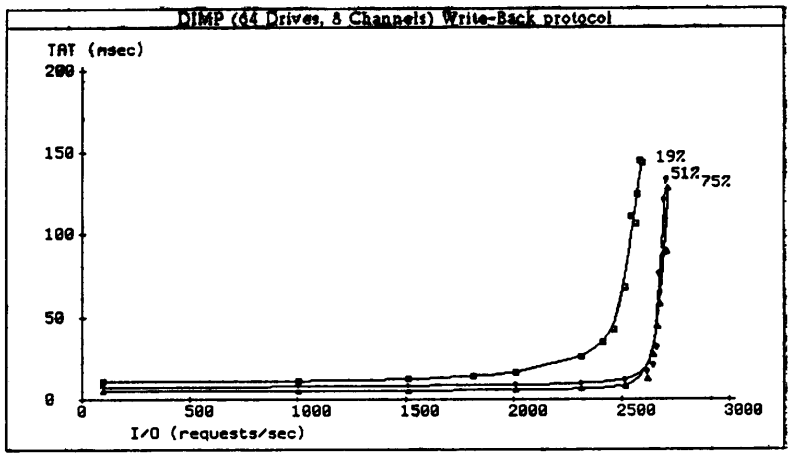

図14 DIMPの性能（ライトバック方式）
Fig. 14 The performance of a DIMP (Write-Back protocol).

4.2.1 考　　察

チャネルのデータ転送時間は，3.0 msec/ブロックである．チャネル・パス数が8の場合のチャネルの転送幅は，$8/(3.0\times10^{-3})=2,666$ 件/秒である．これがスループットの限界性能である．図13，図14の性能グラフより，キャッシュのヒット率が低い場合にも，限界性能にほぼ近いスループットが得られることが分か

る．したがって，DIMP は，キャッシュのヒット率に依存せず，チャネルの転送幅に応じた高いスループットが得られると言える．キャッシュのヒット率が 19% では，ライトバックのほうが，ライトスルーより性能が悪い．これは，I/O のライト比が 0.2 と低いために，ライトバックでは，書き戻しのためのオーバヘッドがあるためと考えられる．

表 5 に示されるように，高トランザクション域（スループットが 1,000 件/秒）でも，低トランザクション域（スループットが 100 件/秒）とほとんど同じ応答時間である．応答時間についても，従来型のディスク・サブシステムに比べて，良い性能が得られると言える．

表 5 I/O が 100 件/秒と 1,000 件/秒の時の TAT
Table 5 TAT (msec) (I/O=100 and 1,000 (requests/sec)).

ヒット率 (%)	19	51	75
(1) I/O=100 (requests/sec)			
・ライトスルー (msec)	11.11	7.90	5.49
・ライトバック (msec)	11.14	7.92	5.48
(2) I/O=1,000 (requests/sec)			
・ライトスルー (msec)	11.59	8.18	5.71
・ライトバック (msec)	11.72	8.15	5.65

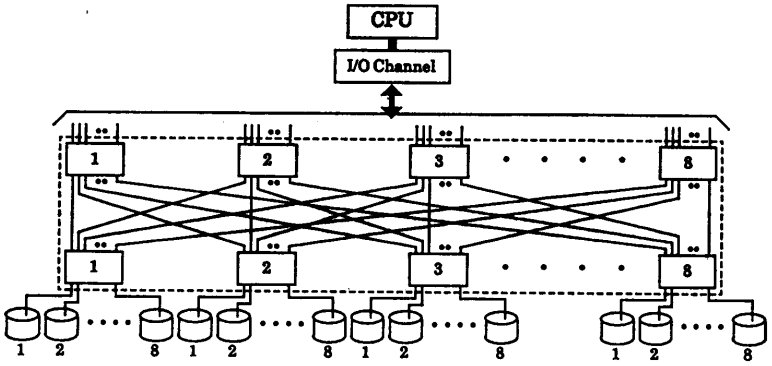

図 15 MPPM を用いたディスク・サブシステムのアーキテクチャ（二層全接続型構成）
Fig. 15 The architecture of a disk subsystem using an MPPM (Dual-Layer Complete-Connection).

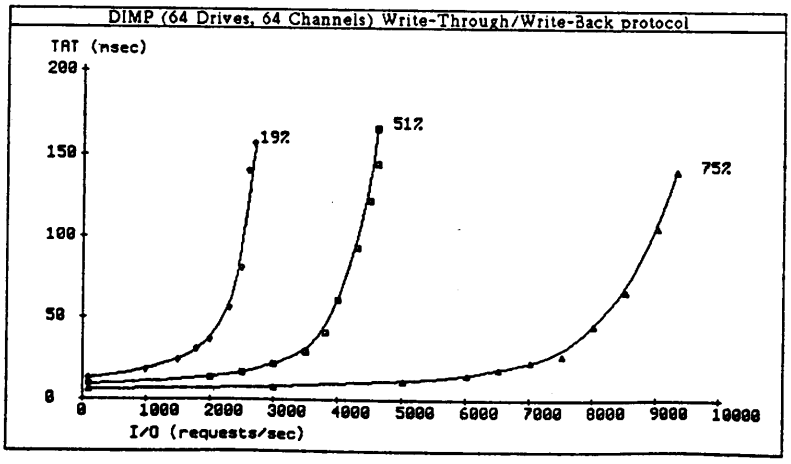

図 16 DIMP の性能（二層全接続型構成）
Fig. 16 The performance of a DIMP (Dual-Layer Complete-Connection).

4.3 DIMP の二層全接続型構成

図 15 は，先の図 1 に示した DIMP に対して，チャネルに接続されている側の MPPM の数を 1 個から 8 個に増やしたものである．この構成（これを二層全接続型構成と呼ぶ）では，チャネルと接続されるパスが 64 本もあり，高スループットが期待できる．以下では，この構成に関して，処理性能を示す．

4.3.1 キャッシュ動作

二層全接続型構成は，チャネルと接続されている側の MPPM が複数あり，それらがディスク・ドライブ装置を共有している．したがって，密結合型マルチプロセッサ・システムに用いられるパラレルキャッシュと同様に，キャッシュ・データのコヒーレンシの問題がある．一般に，ライトスループロトコルでは，無効化方式が一番簡単な方法である．ライトバックプロトコルでは，無効化方式は使えず，ブロードキャストやスヌーピングといった複雑な方法が行われる．MPPM では，データのアクセスは，スイッチングネットワークに与える制御コードに同期して行われる．よって，MPPM をキャッシュとして用いる場合には，無効化は容易に行えるが，ブロードキャストやスヌーピングといった方式は難しい．また，ライトバックプロトコルとライトスループロトコルを比較すると，一般に，ライトの割合が大きい場合には，ライトバックプロトコルのほうが有利である．

二層全接続型構成では，上層の（チャネルと接続されている）MPPM については，キャッシュ・データのコヒーレンシの問題があるが，下層の（ディスク・ドライブ装置と接続されている）MPPM については，キャッシュ・データのコヒーレンシを考慮する必要が

ない．以上の点から，キャッシュ動作については，上層のMPPMはライトスループロトコル（無効化）で動作し，下層のMPPMはライトバックプロトコルで動作する，ライトスルー/ライトバック方式とした．

4.3.2 処理性能

以下の図16に示すのは，図15に示した構成を持つDIMPの性能グラフである．I/Oの発行件数を50,000件としてシミュレーションした結果である．64本ものチャネル・パスがあり，チャネルの転送幅は，64/(3.0×10^{-3})=21,333件/秒である．性能グラフを見ると，この値までのスループットは得られていない．この場合は，ディスク・ドライブがネックになっている．ディスク・ドライブ台数を増やす，あるいは，ディスク・ドライブ装置の処理速度を向上する（シーク時間を小さくし，回転速度を上げる）ことで，さらに性能の向上が期待できる．

4.4 検　討

将来，ディスク・サブシステムの記憶容量は，ますます大きくなり，しかも高いスループットが望まれるようになる．データの転送速度を上げることで，スループットを上げることは可能である．けれども，処理速度の速いメモリデバイスを用い，短時間で複雑な制御をしなければならない．しかも，キャッシュのヒット率が低い場合には，ディスク・ドライブに頻繁にアクセスしなければならず，機械的に動作するディスク・ドライブ装置の処理速度を上げることは難しい．したがって，データの転送速度を上げることで，スループットを向上させるには限界がある．このように考えると，MPPMをキャッシュメモリとして用いたDIMPは，スケーラブルな性能が得られ，将来のディスク・サブシステムとして有効であると思われる．特に，DIMPの二層全接続型構成は，多数のチャネルと接続されるパスがあり，ディスク・ドライブがネックになるだけの高いスループットが得られ有望であると思われる．

また，データ・ベース・システムなどで，データの検索を，高速に行うためには，ポインタでつながれた関連あるデータ群を一度に読み込める必要がある．これらデータ群は，通常複数のディスク・ドライブ装置にわたって格納されている．したがって，複数のディスク・ドライブ装置から並列にデータのアクセスが行える場合，高速なデータ検索が可能である．このような点から，複数のディスク・ドライブ装置から並列にデータのアクセスが行えるDIMPは，データベース・マシンのディスク・サブシステムとして有用であると思われる．

5. おわりに

本論文では，MPPMをディスク・キャッシュとして用いたディスク・サブシステムDIMPを提案した．DIMPの構成と基本的なキャッシュ動作について解説した．従来のディスク・サブシステムでボトルネックとなっていたディスク・ドライブ-ディスク・キャッシュ間のパス数を，MPPMを用いることにより，大きくすることができる．

DIMPは，以下の特徴があることを述べた．

（1） キャッシュのヒット率が低い場合にも，チャネルの転送幅あるいは，ディスク・ドライブの処理速度によって制限される高いスループットが得られる．

（2） RPSミスの発生を抑えることができ，待ち時間を小さくできる．

（3） 構成上の自由度が高く，システム拡張が容易である．

また，従来のキャッシュメモリ付きディスク・サブシステムとDIMPのシミュレーションによる性能評価を行い，シミュレーション結果を挙げて，DIMPの有効性を示した．

本論文では，DIMPを提案すること，上記の3つの特徴が得られること，およびその将来的な有効性について述べた．実際に，DIMPを実現する場合には，構成・動作に関して，より詳細な検討が必要だと思われる．また，コストについての検討も必要であると思われる．DIMPは，複数のMPPMを用いており，これをバッファとして用いた場合，集合ディスク（ディスク・アレイ）へも即座に対応できると考えられる．今後は，この点について検討するつもりである．

謝辞 本研究の遂行において，多くのご助言をいただいた日立製作所小田原工場，宮崎道生氏に深謝いたします．

さらに，貴重なご意見をいただいた査読者の方に感謝の意を表します．

参考文献

1) 平野正信：アクセス・ギャップを埋めるディスク・キャッシュの機能を見る，日経コンピュータ，1982年3月22日号，pp. 71-85 (1982).

2) 小畑征二郎，松沢　茂，宮崎正俊，神山　典，表　俊夫：ディスク・キャッシュの効果に関する一考察，情報処理学会論文誌，Vol. 26, No. 6,

pp. 1009-1016 (1985).
3) 日立製作所：HITAC (H-8538-C 3 ディスク制御装置，H-6585 ディスク駆動装置，H-8598 ディスク駆動装置) 解説書，資料番号 8080-2-094-10.
4) Tanaka, Y.: A Multiport Page-Memory Architecture and a Multiport Disk-Cache System, *New Generation Computing*, Vol. 2, pp. 241-260, OHMSHA, Tokyo (Feb. 1984).
5) 宮地泰造，三石彰純，溝口徹夫：バッファ内蔵型ディスク装置の性能評価，電子通信学会論文誌，Vol. J 67-D, No. 11, pp. 1301-1308 (1984).
6) 宮地泰造，三石彰純，溝口徹夫：階層型ディスク・キャッシュ・サブシステムの性能評価，電子通信学会論文誌，Vol. J 68-D, No. 9, pp. 1609-1616 (1985).
7) Salem, K. and Garcia-Molina, H.: DISK STRIPING, *Int. Conf. on Data Engineering*, IEEE, pp. 336-342 (1986).
8) Kim, M. Y.: Synchronized Disk Interleaving, *IEEE Trans. Comput.*, Vol. C-35, No. 11, pp. 978-988 (1986).
9) 物井秀俊，森田幸伯，伊藤英則，岩田和秀，酒井浩，柴山茂樹：マルチポートページメモリを用いた知識ベースマシンの並列制御方式と処理性能，情報処理学会論文誌，Vol. 29, No. 5, pp. 513-520 (1988).
10) 岡田義広，田中 譲：視覚的シミュレータの開発支援システム：FES，情報処理学会論文誌，Vol. 32, No. 6, pp. 766-776 (1991).

(平成3年10月2日受付)
(平成4年10月8日採録)

岡田 義広 (正会員)

昭和39年生．昭和63年北海道大学工学部電気工学科卒業．平成2年同大学院修士課程修了．現在北海道大学大学院工学研究科博士後期課程電気工学専攻在学中．コンピュータアーキテクチャ，データベースマシンの研究に従事．電子情報通信学会会員．

田中 譲 (正会員)

昭和25年生．昭和47年京都大学電気工学科卒業．昭和49年京都大学電子工学専攻修士課程修了．工学博士．現在，北海道大学電気工学科教授．データベースマシン，データベース理論，メディア・ベース，論理型プログラミング等の研究に従事．主たる著書，「コンピュータ・アーキテクチャ」(オーム社，共著)．IEEE，ソフトウェア科学会，人工知能学会各会員．