

## 機 能 デ ィ ス ク シ ス テ ム

## ——設計思想とベンチマーク評価——

## Functional Disk System

喜連川 優\*・高 木 幹 雄\*

Masaru KITSUREGAWA and Mikio TAKAGI

## 1. は じ め に

CPU の処理能力は一昔前に比べると 100 倍以上も高速化されてきた。これには命令の先取り、パイプライン演算処理、キャッシュ、メモリアンタリーブなど数多くのアーキテクチャ的改革、素子技術の進歩が大きく寄与している。一方、二次記憶システムはどうであろう。IBM のディスクの製品系列 IBM 3330, 3350, 3380 をみていると、平均シークタイムがおおの 30 msec, 25 msec, 16 msec, また転送レートは 0.8 MB/sec, 1.2 MB/sec, 3.0 MB/sec となっており、記憶メディアとしての性能は高々一桁向上した程度にとどまっていることがわかる。つまり、CPU の演算速度の急速な改善に二次記憶システムの性能は追いつくことができず、その性能のギャップはますます大きくなる一方である。ディスクキャッシュ、電子ディスクはその一つの解決策と考えられるが、主記憶容量にも匹敵する大容量記憶空間上で単なるディスク操作のエミュレーションしか行っておらず、本質的な解決とはいえない。今日のコンピュータシステムの抱える最も大きな性能のボトルネックは処理装置と二次記憶系との間にあるといえる。

## 2. 二次記憶システムの歴史

ここでディスクシステムの歴史をながめてみる。図 1 に示されるシステム構成は、昔からほとんど変化していない。すなわち、CPU は二次記憶上のデータをアクセスする際、どのシリンダのどのトラックのどのセクタのデータを読み、という極めて低レベルの指令を発行する必要がある。この命令は CPU から、チャンネル、ディスクコントローラそしてディスクへと長い道のりを経てたどっていくのである。実際には計算機上にはオペレーティングシステムが作動しており、ディスクの駆動にはカーネルの I/O ドライバ、データ管理ルーチン、そしてアプリケーションプログラムへとソフトウェア的にも長い道のり

を経てやっとデータをアクセスすることができることになる。このような状況を考えると、二次記憶系へのアクセスは先の物理的性能からさらに一桁以上低下することが一般的である。

従来、ほとんど変革のない二次記憶系に対し 1980 年代に入り 2 つのアーキテクチャ的工夫がなされた。1 つは図 2 に示されるディスクキャッシュと呼ばれるもので、

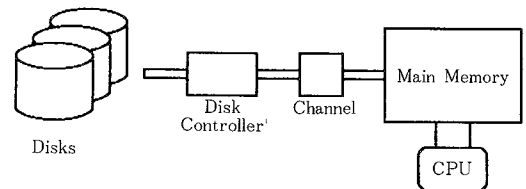


図 1 従来の二次記憶システム

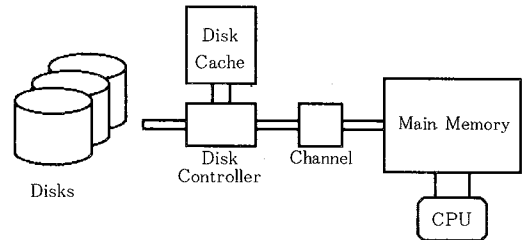
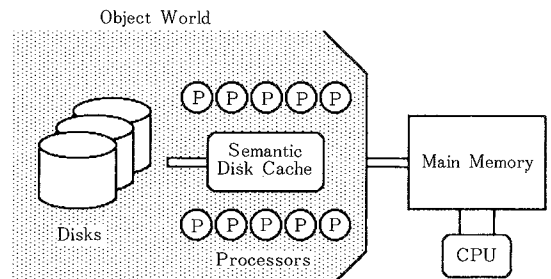


図 2 ディスクキャッシュシステム



Functional Disk System

図 3 機能ディスクシステム

\* 東京大学生産技術研究所  
機能エレクトロニクス研究センター

## 研 究 速 報

半導体メモリが安価になったことからこれを大量に利用して実効的にディスクのアクセス時間を短縮化させようとするものである。80年代から商品化が顕著となり、現在では16MB～512MBにも及ぶ大容量ディスクキャッシュ、電子ディスク装置が提供されている。拡張アーキテクチャを採用しないメインフレームでは、主記憶は16MB止まりであり、主記憶を超える膨大なメモリ空間が存在することになるが、この大きなメモリ空間に対して実質的に何ら処理能力(プロセッサ)を有しておらず、単なるディスクのエミュレーションを行っているにすぎない。

もう一つの改良は拡張アーキテクチャで採用されたダイナミックサブチャネルという方式であり、二次記憶装置とチャネルシステムの間でのパスを動的に柔軟に制御することにより実効的な二次記憶へのバンド幅を向上させることをねらったものである。いずれにしても、二次記憶系に対する変革は僅かであり、しかも抜本的な解決になっているとは言い難い。

### 3. 機能ディスクシステム

このような背景に鑑み、今日のコンピュータシステムにおける最大のボトルネックを解消すべく新しい超高性能二次記憶システム：機能ディスクシステムの開発を行っている。これは従来のごとく、シリンドラ、トラック、セクタという極めて低レベルのアドレスアクセスではなく、オブジェクト指向概念を取り入れることにより、より高位のアクセスインタフェースを提供するとともに、現在のディスクキャッシュシステムにみられる大容量半導体ステージングバッファに加え、安価なマイクロプロセッサを多数利用し、高度な並列処理メカニズムを導入することにより極めて高い性能を実現することを目的としている。(図3)

ソフトウェア的には、従来のオペレーティングシステム上ではI/Oはユーザからは制御困難であり、大きなオーバーヘッドに甘んじねばならなかったが、専用高速I/Oドライバによりアプリケーションに適した入出力環境を提供することができる。すなわち、ディスクを単なる記憶メディアとして使うのではなく、そこに“機能”を実現することにより、ディスク上でのデータ管理、データ処理を可能とし、極めて高い性能が期待されるわけである。

### 4. 試作システムの構成

以上の考えに基づき、まず簡単な実験システムを構築した。図4はその概観、図5にそのハードウェアアーキテクチャを示すが、極めて単純かつ簡素であり、16ビットマイクロプロセッサMC68000 3台からなる共有メモ

リ型マルチプロセッサシステムとなっている。1台がマスタープロセッサ、他の2台がスレーブプロセッサである。おのおの、ローカルメモリを1MB、512KB持つ。マスタープロセッサはプログラム開発のためメモリ容量を大きくしてある。大容量データの操作ではメモリ間転送が頻繁に生ずると考えられ、プログラムモードでは大きな性能の低下が予想されることから、各プロセッサはMC68450なるDMAコントローラを登載している。このDMACはチェーンモードを支援しており、リスト構造などの非連続データの効率の良い操作が可能となる。

プロセッサ群は16ビットのシステムバス(モトローラ

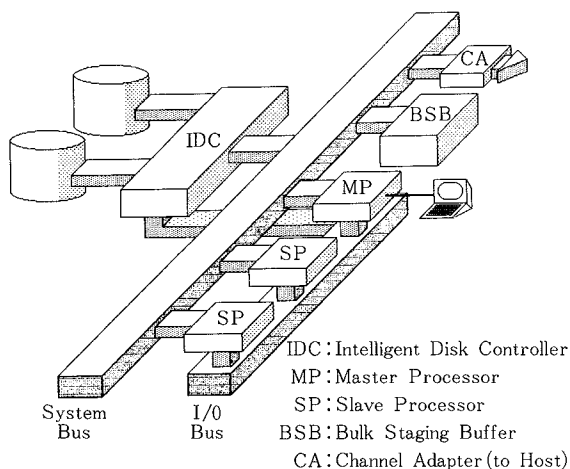


図4 機能ディスクシステムの構成

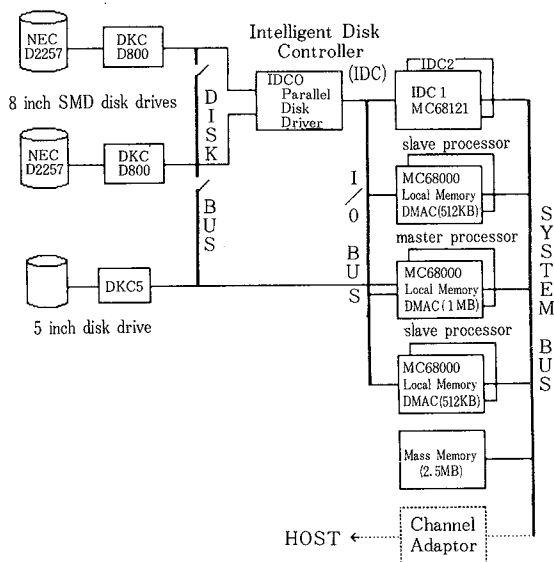


図5 機能ディスクシステムのハードウェアアーキテクチャ

研 究 速 報

標準バス Versa Bus) を介して結合されており, 互いの通信は共有メモリ上でなされる. 本マシンではデータラヒックの高いアプリケーションを想定しており, バスの能力を上げるため, プロセッサからディスクへのデータ出力にはシステムバスとは別の I/O バスを設けてある.

ディスクは 8 インチウィンチェスタ SMD タイプ (転送レート 1.2 MB/SEC, 記憶容量 168 MB) のドライブ (日本電気製 D 2257) を用いており, 性能向上のため, 2 台のディスクを並列に駆動する専用のディスクコントローラ (IDC) が用意されている. IDC とディスクの間は SCSI インタフェースである. IDC は MC 68121 なる 8 ビットのマイクロプロセッサを用いており, これによりディスクの低レベル制御を司る. すなわち, ディスクに対する種々のコマンドの発行, IDC 内の各論理ユニットの初期化, 駆動, 監視を行う. また, マスタプロセッサは MC 68121 内の通信用レジスタを介して, IDC にコマンドを与え, MC 68121 がこれを解釈実行する. IDC 内では, 1.2 MB/sec のデータ流に沿って動作する機能部位が複数存在するが, これらはすべて高速マイクロプログラム制御によって実現されている. IDC は Versa ボード

3 枚から構成されている.

ソフトウェアはマスタプロセッサ上の Microware 社 OS-9 で開発されている. 開発言語は C およびアセンブラである. 先の 8 インチディスクとは別に, システムディスクとして 5 インチディスク, および, フロッピーディスクがマスタプロセッサに直接結合されている.

このシステムはあくまでも低コストで早期に実現することを目的とした実験システムであり, 機能ディスクシステムの理想像とはもちろん異なる.

## 5. 性 能 評 価

機能ディスクは汎用の二次記憶システムを目指しているが, 当面のアプリケーションとしては, 大規模画像処理, 巨大データベースなどを考えている. これは, これらのアプリケーションの二次記憶系に対する負荷が極めて重く, 機能ディスクシステムの効果が最も大きいと考えるからである.

米国ウィスコンシン大学で 1983~1984 年にかけて開発された関係データベースシステム評価用ベンチマーク<sup>1)</sup>を用いて機能ディスクシステムを評価したので, そ

表 1 ウィスコンシンベンチマークによる機能ディスクの性能評価  
(商用データベース管理システムとの性能比較)

(1) 選択演算		(3) ファイル結合演算	
range of t is tenKtuple		range of t is tenKtuple1	
retrieve into TEMP (t, all)		range of w is tenKtuple2	
where a 0 < 1000;		retrieve into TEMP (t, all, w, all)	
インデックス無 (秒)		where (t, a 0 = w, a 0) and w, a 0 < 1000	
U-INGRES	64.4		(分)
C-INGRES	53.9	U-INGRES	10.2
ORACLE	230.6	C-INGRES	1.8
IDM no DAC	33.4	ORACLE	> 300
IDM DAC	23.6	IDM no DAC	> 300
DIRECT	46.0	IDM DAC	> 300 (108 sec)
FUNCTIONAL		DIRECT	10.2
DISK	$3.197 + \alpha + \beta$	FUNCTIONAL	(秒)
		DISK	$6.18 + \alpha + \beta$
(2) 射影 (重複除去) 演算		(4) 集計演算	
range of t is oneKtuple		range of t is tenKtuple	
retrieve unique into TEMP (t, all)		retrieve into sum (t, a 1 by t, a 2)	
U-INGRES	236.8	100 partitions	
C-INGRES	132.0	U-INGRES	174.2
ORACLE	199.8	C-INGRES	484.8
IDM no DAC	122.2	ORACLE	1487.5
IDM DAC	68.1	IDM no DAC	67.5
DIRECT	58.0	IDM DAC	38.2
FUNCTIONAL		DIRECT	229.5
DISK	$0.975 + \alpha + \beta$	FUNCTIONAL	(秒)
		DISK	$3.243 + \alpha + \beta$

の結果を示す。

表 1 に示される INGRES (U は大学版, C は商用版), IDM, ORACLE などの関係データベースプロダクトの性能値は, ウィスコンシン大学により, VAX 11/750 上でかなり平等な環境で測定された値である。問い合わせは QUEL 言語で記述してある。第一のベンチマークは単純な検索で 10000 ケのレコードからなるファイル (1 レコード 182 B, 全体で 1.82 MB) から 1000 レコードを内容検索するもので, インデックスが無い場合について比較した。a0 は 2 バイト整数フィールドである。第二は射影演算で, 1000 ケのレコードからなるファイルに対して, 重複除去を行っている。第三のファイル結合演算では, 10000 ケのレコードからなる 2 つのファイルを a0 フィールドで結合する問い合わせであり, 片側のファイルは条件節により 1/10 に絞られている。第四は集計演算であり, a2 のフィールドの値によってファイルを分類し, そのおのおののレコード群に対して a1 のフィールドの総計をとっている。この類の処理はオフィスデータベースでは頻繁に行われるものである。a1, a2 は 2 バイト整数フィールドであり, 10000 ケのレコードからなるファイルが a2 によって 100 ケに分類される。

いずれの場合にも機能ディスクは高い性能を示していることがわかる。データベース処理への I/O のチューニング, そして, フィルタ, クラスタリング機構を内蔵するデータベース専用ディスクコントローラ, さらに, MC 68000 マイクロプロセッサ群による高度の並列処理によりこれらの値が達成されている。

ただし, 実験システムは未完成であり, この測定値はディスク 1 台プロセッサ 2 台での値である。仮想記憶メカニズムもないため, 取り扱えるファイルの大きさはメモリ容量で制限される。インデックスも未実装である。また,  $\alpha$ : 問い合わせ言語の翻訳等システムオーバーヘッド,  $\beta$ : 結果データのディスクへの書き出し時間は含まれていない。問い合わせは簡単であり,  $\alpha$  はそれほど大きくないと考えられるし, また,  $\beta$  についてはベンチマーク (2) では出力データ量は入力データ量とほぼ等しいため無視できないが, (1) (3) では出力は入力の 1/10, (4) ではさらに少なく, 第 1 項に比べて小さいと考えられる。

## 6. お わ り に

今日の計算システムにおける最大のボトルネックは二

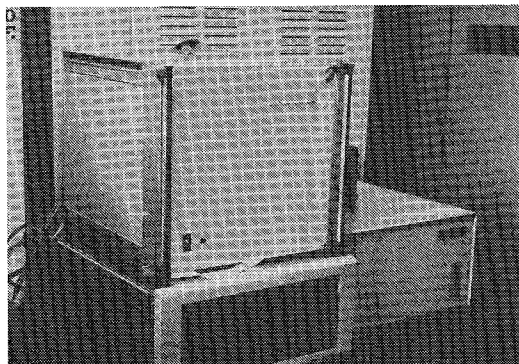


図 6 機能ディスクシステムの概観

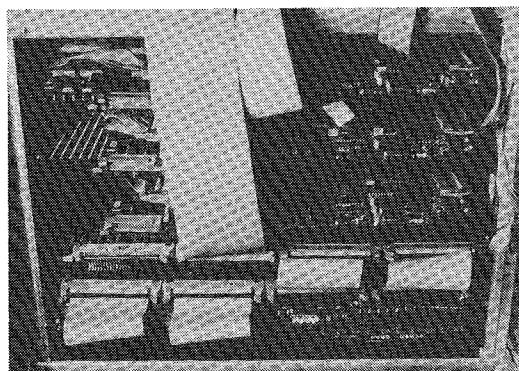


図 7 機能ディスクシステムの内部構成

次記憶システムにあるとの考えから, 超高性能二次記憶系の開発を目的とし, 機能ディスクシステム第 1 版の試作を行った (図 6, 図 7)。試作システムは数台のマイクロプロセッサ, 数 MB の DRAM, 専用ディスクコントローラからなる簡単なものであるが, 従来のソフトウェア DBMS に比べ, 極めて高い性能を確認した。

今後, ディスク, プロセッサの並列度をさらにあげると共に, 汎用計算機システムへの組み込みに関して, ホスト OS との整合性等の検討をすすめていく予定である。

現在, 大規模画像処理の機能ディスクへの実装を検討中である。

(1985 年 10 月 7 日受理)

## 参 考 文 献

- 1) D. Bitton, D. J. DeWitt, 'Benchmarking Database Systems A Systematic Approach' Proc. of VLDB 83