計算機アーキテクチャ　62－3
（１９８６．７．２１）

機能ディスクシステムとその評価

喜連川　優，原田リリアン，中野美由紀，高木　幹雄

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

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、メモリイ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IBM3330，3350，3380 をみてみると、平均シークタイムが各々30 msec、25msec、16msec、又転送レートは0.8MB/sec，1.2MB/SEC，3.0MB/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出来ず、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。

又，我々は，これまで，データベースマシン，データベース管理システムの研究を行ってきた。その研究過程において，汎用計算機にデータベースシステムの構築するばあい，現状のオペレーティングシステムの入出力ルーチンではその非効率性から極めて低い性能しか得られないことを認識している。入出力はスーパバイザ配下にあり，通常のユーザからは制御困難であるため，既存システムとは独立に，我々はその根底から問題の解決を試みることとした。

機能ディスクシスムはこの旧態依然たる二次記憶系改革の第一歩である。通常のＳＭＤディスクドライヴと数台のＭＣ６８０００からなる簡単な試作システムは，ウィスコンシン関係データベースベンチマークに関し現在の商用ＤＢＭＳに比べ，一桁以上の極めて高い性能を達成した。本稿では，機能ディスクシステムの構成ならびにその評価結果について報告する。

Functional Disk System & Its Evaluation

Masaru KITUREGAWA,　Lilian HARADA,　Miyuki NAKANO,　Mikio TAKAGI

Institute of Industrial Science
The University of Tokyo

The major performance bottle neck in the current computer system is in the low-performance secondary system. The performance of the CPU has increased dramatically so far, about several orders of magnitude improvement has been achieved. On the other hand, that of the disk system has shown little advance since nineteen sixties. The von Neumann bottle neck between the CPU and the secondary storage subsystem has been much more enlarged.

In this paper we propose a new secondary system, 'Functional Disk System'. Functional disk system incorporates both a large staging buffer and parallel processing mechanism, where the intelligent data management as well as efficient data processing is performed. Dynamic clustering mechanism is employed in the disk controller. Dedicated i/o environment eliminates the large overheads usually incurred in the i/o driver of the operating system kernel. Its capability is much enhanced than the associative disk system proposed so far which is characterized by the on-the-fly filter processing.

The simple pilot system was constructed to examine the feasibility of our approach and its performance was evaluated with the Wisconsin Benchmark. The functional disk system showed much higher performance than the current software DBMS on VAX 750.

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、メモリイ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IBM3330，3350，3380 をみてみると、平均シークタイムが各々30msec、25msec、16msec、又転送レートは0.8MB/sec，1.2MB/SEC，3.0MB/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出来ず、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。

又，我々は，これまで，データベースマシン，データベース管理システムの研究を行ってきた。（７，８，９）その研究過程において，汎用計算機にデータベースシステムの構築するばあい，現状のオペレーティングシステムの入出力ルーチンではその非効率性から極めて低い性能しか得られないことを認識している。入出力はスーパバイザ配下にあり，通常のユーザからは制御困難であるため，既存システムとは独立に，我々はその根底から問題の解決を試みることとした。

機能ディスクシスムはこの旧態依然たる二次記憶系改革の第一歩である。通常のＳＭＤディスクドライヴと数台のＭＣ６８０００からなる簡単な試作システムは，ウィスコンシン関係データベースベンチマークに関し現在の商用ＤＢＭＳに比べ，一桁以上の極めて高い性能を達成した。本稿では，機能ディスクシステムの構成ならびにその評価結果について報告する。

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。即ち，ＣＰＵは二次記憶上のデータをアクセスしようとする際，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め、という大変低レベルの指令を出す必要がある。この命令はＣＰＵから，チャネル、ディスクコントローラを経てディスクに長い道のりを経てたどりつくのである。実際には計算機上にはオペレーティングシステムがあり、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出来ることになる。この様な状況を考えると、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事が一般的である。

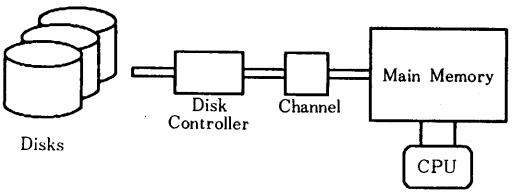

図１　従来の二次記憶システム

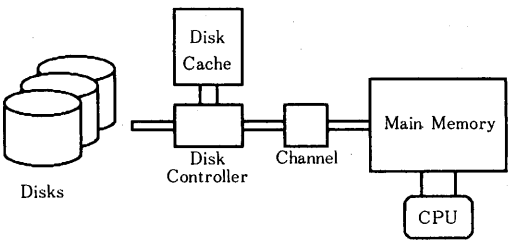

図２　ディスクキャッシュシステム

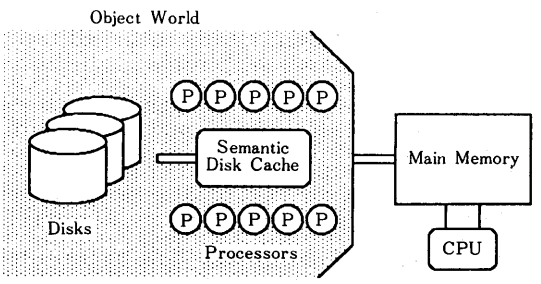

図３　機能ディスクシステム

さて、ほとんど変化のない二次記憶系の世界にも1980年代に入って2つのアーキテクチャ的工夫がなされている。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、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、主記憶を超える莫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だけでしかないというのはいかにももったいない次第である。

もう一つの改良はＸＡ（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

この様な背景を鑑み、今日のコンピュータシステムに於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。これは従来の如く、シリンダ、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イ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利用し、高度な並列処理メカニズムを導入する事により極めて高い性能を実現する事を目的としている。（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出来る。即ち、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．試作ハードウェアシステムの構成

以上の考えに基き，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、３台のＭＣ６８０００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台がマスタプロセッサ，他の２台がスレーブプロセッサである。各々，ローカルメモリを１ＭＢ，５１２ＫＢ持つ。マスタプロ

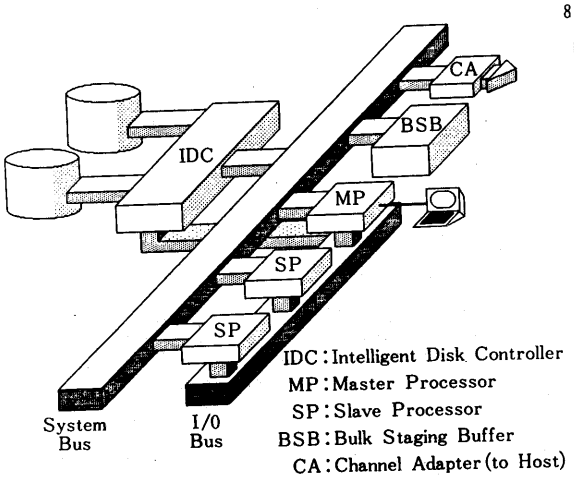

図4　機能ディスクシステムの構成

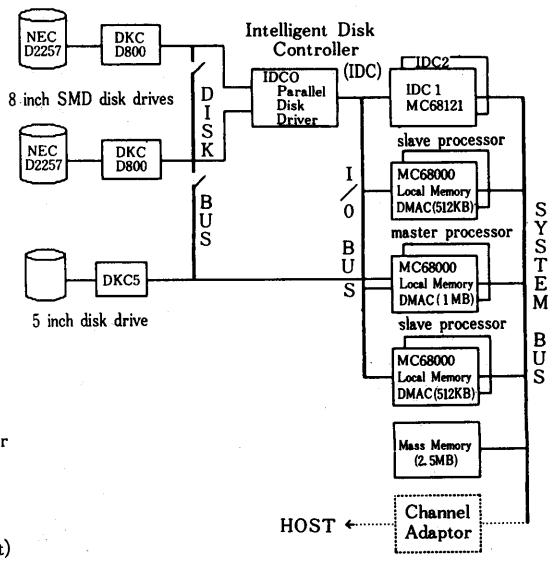

図5　機能ディスクシステムのハードウェアアーキテクチャ

セッサはプログラム開発の為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事から，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ＤＭＡコントローラはチェイ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ト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１．２ＭＢ／ｓｅｃ，記憶容量１６８ＭＢ）のドライブ（日本電気製　Ｄ２２５７）を用いており，性能向上の為、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：Ｉｎｔｅｌｌｉｇｅｎｔ　Ｄｉｓｋ　Ｃｏｎｔｒｏｌｌｅｒ）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ＭＣ６８１２１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ＭＣ６８１２１内の通信用レジスタを介して，ＩＤＣにコマンドを与え，ＭＣ６８１２１がこれを解釈実行する。ＩＤＣ内では，１．２ＭＢ／ＳＥＣのデータ流れ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，セクタ／レコードアセンブリ機構，フィルタ機構，不要フィールド除去機構，動的クラスタリング機構等が登載されている。

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，及び，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勿論異なる。

５．機能ディスクシステムのシステムソフトウエア

５．１　全体構成

機能ディスクシステムは，試作機の段階であり，実装を短期間で実現するために，既存のツールを出来る限り有効利用する方針に基づき，開発を行ってきた。しかし，バッファ（共有メモリ）資源の有効管理，マルチプロセッサ等に関しては，既存のシステムで補えない部分が存在し，機能ディスクシステムの目的に合わせた制御方式を導入した。

開発環境としては，ＲＯＭベースのＯＳ－９(V1.2)を利用している。ＯＳ－９は，本体が小さいが，かなり強力なマルチプロセス機能と整備されたI/O マネジャ，ドライバ群があり，今回のシステムでは，十分利用出来ると考えられるからである。また，ソフト開発は，マスタプロセッサ上でＣ言語を用いて行った。

機能ディスクシステムのソフトウェアは，図２に示す構成をしている。ＯＳ－９の説明については省略する。まず，最も基本的な機能を提供しているのが，ＦＤＳコミュニケーション・システム（ＦＤＳ：機能ディスクシステム）である。ＦＤＳコミュニケーション・システムの機能は，複数台のプロセッサの同期，ソフトウェア資源（プログラム・モジュール，データ）のプロセッサ間の転送，共有メモリ上の通信バッファの管理等である。各アプリケーションは，このＦＤＳコミュニケーション・システムを通じて，全システム

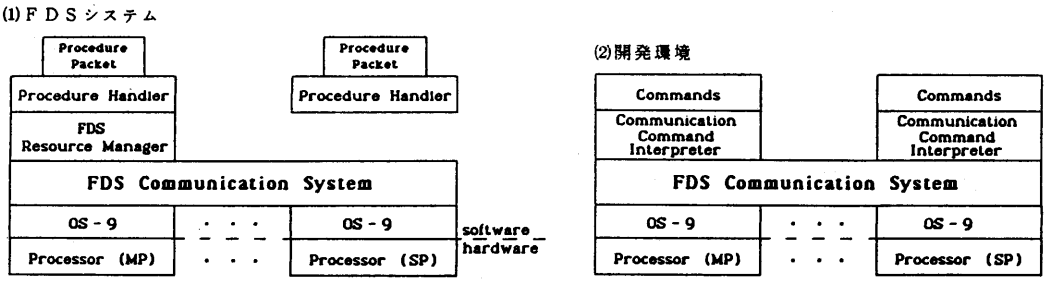

図６．ＦＤＳソフトウェアシステムの構成

を制御することになる。

ＦＤＳコミュニケーション・システムの上で実際にシステムを統合しているのが，ＦＤＳリソース・マネジャである。ＦＤＳリソース・マネジャは，マスタプロセッサ上のプロセスであり，ＩＤＣへの資源情報の設定，ＩＤＣの制御，各プロセッサ上のプロシジャ・ハンドラの制御等を行っている。また，ＦＤＳシステムテーブルを始めとするシステム資源の管理も行っている。

プロシジャ・ハンドラは，ＦＤＳリソース・マネジャによって与えられたプロシジャ・パケットを並列に処理する。プロシジャ・パケットには，ＩＤＣによって共有メモリに展開されたデータブロックに対する処理手続きが，記述されている。一般に，プロシジャ・パケットは，データブロックのローカルメモリへのＤＭＡ転送部，データ処理部，処理結果の書き戻し部から構成される。

ＦＤＳコミュニケーション・システム上に，開発支援ツールとして，ＦＤＳコミュニケーションコマンド・インタプリタを作成した。これは，ソフトウェア開発時に各プロセッサの制御をマスタプロセッサを通じて行うものである。主な機能は，各プロセッサ上でのコマンド実行，プログラム・モジュール，データ・ファイルのプロセッサ間の転送などである。

その他，関連ユーティリティとして，ＩＤＣシュミレータ，ハードウェアテストツール，性能評価用ツール等が用意されている。

５．２　ＦＤＳコミュニケーション・システム

ＦＤＳシステム・ソフトウェアの中核をなすのが，ＦＤＳコミュニケーション・システムである。ＯＳ－９は単一プロセッサ用のＯＳであり，マルチプロセッサ制御機能及び共有メモリ管理機能を支援していない。これらの機能はＦＤＳコミュニケーション・システムが実現している。ただし，単一プロセッサ内のプロセス間通信については，ＯＳ－９の機能をそのまま利用している。

第一に，マルチプロセッサ制御として，ＦＤＳコミュニケーション・システムでは，複数台のプロセッサ間でのプロセス間通信機能を共有メモリを介して行なっている。これは，大きく，メッセージ交換を行うTRANSFERプロセスとコミュニケーション・ライブラリに分けられる。TRANSFERプロセスでは，コミュニケーション・システム情報の初期化，退避，復元，及びプロセッサ間のメッセージの受け渡しを行う。コミュニケーションライブラリには，表１に示されるルーチンが提供されており，これらは，プロセス間の制御を行う上位のアプリケーションによって，使用される。

第二に，共有メモリ管理機能として，ＦＤＳコミュニケーション・システムでは，モードによって異なる二種類の共有メモリ管理を実現している。ＦＤＳの本来の機能は，二次記憶から，共有メモリ上に，データを取り出すことである。そこで，効率良く共有メモリを利用するために，ＦＤＳモードとコミュニケーション・モードの二つのモードにより，メモリの使用形式を変更する方式を採用した。コミュニケーション・モードでは，共有メモリは，コミュニケーションシステムによって管理され，メッセージバッファとシステム情報に利用される。ＦＤＳモードになると，コミュニケーションに関する情報は，最小限必要な領域を除いて退避され，バッファは，開放される。共有メモリは，ＦＤＳリソース・マネジャによって，管理される。再び，コミュニケーション・モードに戻ると，退避された情報が復元され，プロセス通信が再開される。

６．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，巨大データベース，大規模画像処理などを考えている。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(1)を用いて機能ディスクシステムを評価したので、その結果を示す。

関数名	動作概要
comminit	システム情報の初期化，再設定。バッファリンクの構築
commsFDS commrFDS commwFDS	ＦＤＳモードの設定。コミュニケーション情報の退避。 ＦＤＳモードの解除。コミュニケーション情報の復元。 (MP 上で使用) ＦＤＳモード解除を待つ。(SP 上で使用)
commregp commextp commendp	通信対象プロセスの登録。 通信対象プロセスが，登録されていることの確認。 プロセスの登録抹消。
commbid	プロセッサIDを得る。
receive send	受信データを取り込む。（バイト単位） 指定されたプロセスへデータを送る。（バイト単位）

表１．ＦＤＳコミュニケーション・ライブラリ

表２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、ウイスコンシン大学により，ＶＡＸ１１／７５０上でかなり平等な環境で測定された値である。問合せは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１８２Ｂ、全体で１．８２ＭＢ）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ヶ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ヶ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合せであり，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，その各々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なわれるものである。a1,a2は２バイト整数フィールドであり，１００００ヶのレコードからなるファイルがa2によって１００ヶに分類される。

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臓するデータベース専用ディスクコントローラ，更

(1) 選択演算

```
range of t is tenKtuple
retrieve into TEMP(t.all)
    where a0<1000;
```

	インデックス無（秒）
U-INGRES	64.4
C-INGRES	53.9
ORACLE	230.6
IDM noDAC	33.4
IDM DAC	23.6
DIRECT	46.0
FUNCTIONAL	
DISK	３．１９７＋α＋β

(2) 射影（重複除去）演算

```
range of t is oneKtuple
retrieve unique into (t.all)
```

U-INGRES	236.8
C-INGRES	132.0
ORACLE	199.8
IDM noDAC	122.2
IDM DAC	68.1
DIRECT	58.0
FUNCTIONAL	
DISK	０．９７５＋α＋β

(3) ファイル結合演算

```
range of t is tenKtuple1
range of w is tenKtuple2
retrieve into TEMP(t.all,w.all)
where (t.a0 = w.a0) and w.ao < 1000
```

	（分）	
U-INGRES	10.2	
C-INGRES	1.8	
ORACLE	> 300	
IDM noDAC	> 300	
IDM DAC	> 300	(108sec)
DIRECT	10.2	
FUNCTIONAL	（秒）	
DISK	６．１８＋α＋β	

(4) 集計演算

```
range of t is tenKtuple
retrieve into sum(t.a1 by t.a2)
```

	100 partitions
U-INGRES	174.2
C-INGRES	484.8
ORACLE	1487.5
IDM noDAC	67.5
IDM DAC	38.2
DIRECT	229.5
FUNCTIONAL	（秒）
DISK	３．２４３＋α＋β

表２．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）

に，ＭＣ６８０００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。

但し，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。ＱＵＥＬのコンパイラは現在実装中であり，プロシジャは全てＣで直接記述されている。つまり，α：問合せ言語の翻訳等システムオーバーヘッドは含まれていない。又，β：結果データのディスクへの書出し時間も含まれていない。問合せは簡単であり，αはそれ程大きくないと考えられるし，又，β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略等しい為無視できないが，(1)(3)では出力は入力の1/10，(4)では更に少なく，第１項に比べて充分小さいと考えられる。

７．おわりに

今日の計算機システムに於ける最大のボトルネックは二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７，図８）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。

リレーションサイズ，クラスタサイズ，タプル長，プロセッサ台数，パーティション数等に対する機能ディスクシステムの性能依存性についても，既に詳細な評価をすすめすており（５，６），稿を改めて発表する予定である。

尚，先の評価ではディスクを１台駆動した場合について示したが，現在ディスクを２台並列に駆動する実験を進めており，例えば先のベンチマークの集計演算に関してがディスク２台，プロセッサ３台で１．８０秒という値を得ている。

今後，ディスク，プロセッサの並列度を更にあげると共に，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ゆく予定である。又，ＱＵＥＬの実装も進めている。

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。

［参考文献］

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

2） 喜連川優，高木幹雄： 機能ディスクシステム，生産研究，37巻， 2号(1986)

3） 喜連川優，原田リリアン，高木幹雄：機能ディスクシステム，第31回情報処理全国大会5D-8 (1985)

4） 喜連川優，中野美由紀，高木幹雄：機能ディスクシステムに於けるシステムソフトウェアの設計，第32回情報処理全国大会5S-3 (1986)

5） 高木幹雄，喜連川優，原田リリアン：関係データベース処理に於ける機能ディスクシステムの性能評価，第32回情報処理全国大会5S-4 (1986)

6） M.Kitsuregawa, M.Takagi, 'Performance Evaluation of Functional Disk System' to appear in Int. Conf. Computer Design (1986)

7） M.Kitsuregawa, H.Tanaka, T.Moto-oka, 'Application of Hash to Database Machine and Its Architecture' New Generation Computing Vol.1 No.1 (1983)

8） M.Kitsuregawa, H.Tanaka, T.Moto-oka, 'Architecture and Performance of Relational Algebra Machine GRACE' Proc. of Int. Conf on Prallel Processing (1984)

9） S.Fushimi, M.Kitsuregawa, H.Tanaka, 'An Overview of the System Software of a Prallel Relational Database Machine GRACE' to appear in 12th Int. Conf. on VLDB 86 (1986)

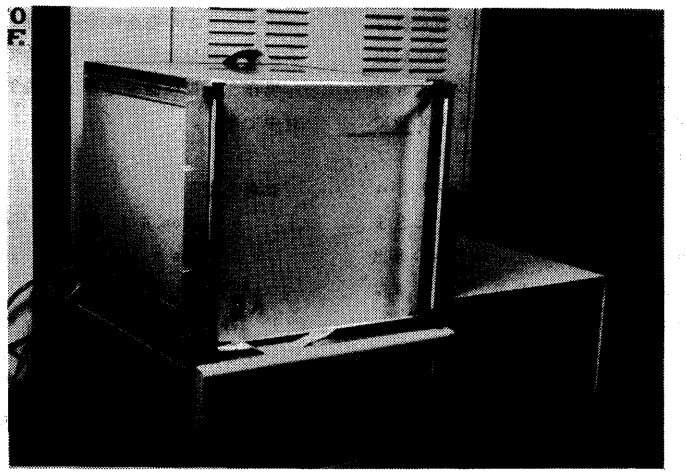

図７．　機能ディスクシステムの概観

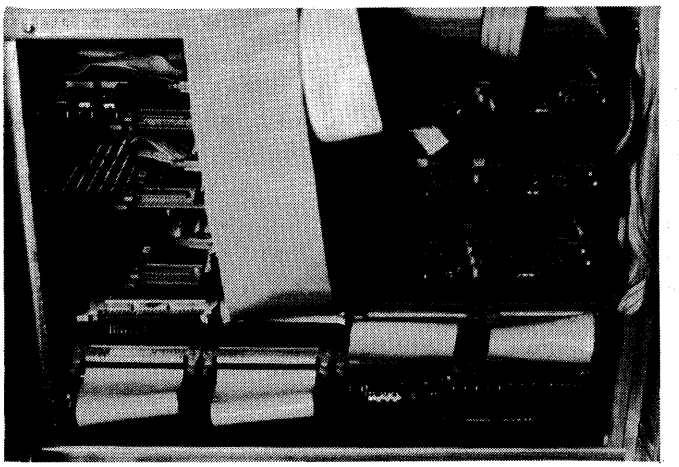

図８．　機能ディスクシステムの内部構成