計算機アーキテクチャ　90－5
（１９９１．１０．７）

データベース演算処理装置のアーキテクチャ

松田進*　東郷一生*　島川和典**　岩崎孝夫**
（株）東芝　*青梅工場　**情報処理・機器技術研究所

大容量データベースのソート及び関係演算処理を高速に行うデータベース演算処理装置のアーキテクチャについて述べる．入出力チャネルとディスク装置との間に接続し，ディスクを直接アクセスすることによりホストＣＰＵの負荷，主記憶の使用率を最小限にした．データベース演算処理装置は３つのプロセッサモジュール，ソートと関係演算専用のハードウェア及び共有メモリで構成し，機能・負荷分散を図っている．拡張ウィスコンシンベンチマークによる性能はホストのみで行った場合に比べ，関係演算の選択処理で１１倍以上，結合処理で１６倍以上の値が得られた．

An Architecture of a Database Processor

Susumu MATSUDA* Kazuo TOGO* Kazunori SHIMAKAWA** Takao IWASAKI**
TOSHIBA Corp. *OME Works, **Information Systems Engineering Laboratory

9, 2-chome, Suehiro-cho, Ome-shi, Tokyo 198, Japan

An architecture of a database processor, which can process a mass volume of data with high performance, is described. By connecting the processor between an I/O channel and a disk subsystem, a host CPU's load and a rate of memory usage have been minimized. The database processor has three processor modules, special hardware for sorting and relational operations and a shared memory with an optimal balance for both function and load. With the extended Wisconsin benchmark, selection performance is eleven times as fast as that in a host or more, and join performance is sixteen times as well.

1．はじめに

近年のデータ量の増大と情報の複雑化に伴い，大容量データのソート及び関係データベースに対する関係演算の高速な処理要求がますます大きくなってきている．特に関係データベースはデータの操作が容易であることから過去１０年間で急速な普及を果たしてきており，今後はさらに発展していくことが予想される．

このような背景の中，当社は（財）新世代コンピュータ技術開発機構の発足以来，同機構からの再委託研究の一環として“ソート及び関係演算を高速に処理するハードウェア”の開発に従事してきた[1][2][3]．

今回，その成果に基づきＴＰホストコンピュータＴＰ９０／７０モデルの専用付加プロセッサとして高速なソート及び関係演算処理を行うことができるデータベース演算処理装置（以下ＤＢＥ）を製品化したのでそのアーキテクチャの概要を述べる．

2．ハードウェア構成

2．1　ホスト接続方式

図１にＤＢＥのハードウェア構成を示す．ＤＢＥはホストシステムのチャネル部に接続し，ディスクコントローラとの間には独立のパスを持つ．

ホストとのコマンドインタフェースはチャネルを通して行い，演算の対象となるファイルのディスクからの読み出し，ＤＢＥでの処理結果のディスクへの書き込みは，ＤＢＥ／ディスクコントローラ間のパスにより行う．

従って，ホストはＤＢＥに演算を指示すれば，あとはＤＢＥからの演算完了を待つだけでよい．

本方式では，ディスク以外のデバイス上のデータを処理の対象にすることはできないが，通常データベースはディスクに格納されることが多いので，他のデバイスのサポートは重要ではない．

他の方式として，ホストのシステムバスに接続する方式とディスクコントローラに内蔵する方式が考えられるが，それぞれ以下のような特徴がある．

- システムバス接続方式では，システム上の全デバイスを処理の対象とできるが、データアクセス時にホストシステムの介在を必要とするので，ホストシステムの入出力性能の制限を受け，専用ハードウェアの性能が充分に出せないおそれがある．
- コントローラ内蔵方式では，ホストシステムの性能の制限を受けずにアクセスできるが，専用ハードウェアを内蔵するコントローラ下のディスクしかサポートできない．

以上から，ＤＢＥでは機能，性能を重視して，本接続形態を採用した．

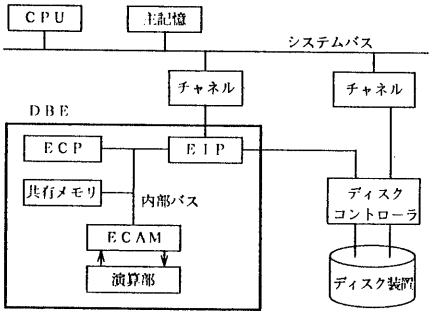


図１　ＤＢＥのハードウェア構成

2．2　構成要素

ＤＢＥは以下のモジュールから構成される．

（1）ＥＩＰ

マイクロプロセッサを内蔵し，ホストとの物理レベルのインタフェースを制御する．同時にディスクコントローラとも接続し，ディスクとの入出力を制御する．また，出力ファイルの再構成処理も行う．

（2）ＥＣＰ

ホストからのコマンドを解析し，ＥＩＰ／ＥＣＡＭを制御して要求された機能を実現する．本モジュールもマイクロプロセッサを内蔵する．

（3）ＥＣＡＭ

マイクロプロセッサを内蔵し，共有メモリ上に入力されたレコードデータからキー部を切り出し，演算部へ送出する．このとき，キーデータの内部データ形式への変換も行う．また，演算部から出力された演算結果を共有メモリ上へ転送する．

また，ＳＥＬＥＣＴ演算はキーを切り出す前に，本モジュールで実行される．

（4）演算部

ＥＣＡＭから送られてくるデータを整列するソータ，整列されたデータに関係演算を施す関係演算部からなる．

（5）共有メモリ

ディスクから入力されたレコードデータ，ＥＣＡＭから出力される演算結果等を格納する大容量の共有メモリである．本装置では最大５１２ＭＢまで実装できる．

２．３　演算処理方式

（1）キー切り出し方式

ＤＢＥでは，演算部へレコード全体を入力するのではなく，キーとそれが切り出されたレコードを対応付ける番号（以下レコード識別子）のペア（以下エレメント）を入力する．

演算結果として，レコード識別子のみ出力されれば，それをもとに元のレコードを取り出すことができる．

一般に，レコード識別子の長さ（ＤＢＥでは４バイト）は，レコードのキー以外の部分の長さよりはるかに短いので，本方式をとることによりソートセルのメモリ量を大幅に削減することができる．

（2）ソータ

複数のソートセルからなるパイプラインソータであり，各ソートセルは２ｗａｙマージソートアルゴリズムによる．本装置では最大１８段までのソートセルが接続可能である．

ＥＣＡＭから送られるエレメントを入力とし，キーをもとにソートする．

初段のソートセルのデータバッファ容量は３２バイトあるので，キー長が１２バイト以下のとき，関係演算部まで含めると５０万件までのエレメントを一度にソートできる．

ここでソートされたエレメントは，関係演算部へ送出される．

（3）関係演算部

２つのリレーション間の関係演算を行うモジュールであり，サポートしている演算はＲＥＳＴＲＩＣＴ，ＪＯＩＮである．また１９段目のソートセルとしても機能する．

ソータから送られるエレメントを入力とし，キーをもとにソート，関係演算等の演算を行う．

演算結果としてレコード識別子のみ，あるいはエレメントを出力する．この出力はＥＣＡＭを通して，共有メモリへ書き込まれる．

２．４　ＤＢＥ内の処理メカニズム

ホストからの演算指示（ＤＢＥコマンド）は，ＥＩＰを通してＥＣＰへ送られる．ＥＣＰはＤＢＥコマンドを解析し，ＥＩＰ／ＥＣＡＭを制御して実行する．

以下，ＳＥＬＥＣＴを例にとってＤＢＥ内の処理メカニズムを次頁の図２に示す．

（1）入力フェーズ

演算の対象となるファイルは，ＥＩＰを通して共有メモリ上の入力ファイル格納領域へ読み込まれる．

対象ファイルは分割して順次読み込まれるので，分割単位が読み込まれる度に，ＥＣＡＭへ渡される．

（2）演算フェーズ

ＳＥＬＥＣＴが指示されているときは，ＥＣＡＭは最初に選択条件の評価を行う．その条件が成立したレコードに対してのみ，キーの切り出しを行い，そのレコードの共有メモリ上の先頭アドレスをレコード識別子としてエレメントを作成し，演算部へ送出する．

ソータは上記エレメントに対し，ソート処理を施し，関係演算部へ送出する．ＳＥＬＥＣＴのとき，関係演算部はエレメント中のレコード識別子のみを取り出し，ＥＣＡＭへ送出する．

ＥＣＡＭは，関係演算部から送られてきたレコード識別子を，あらかじめ指定された共有メモリ上のバッファ領域へ書き込む．

（3）出力フェーズ

レコード識別子の共有メモリへの書き込みが完了すると，ＥＣＰはＥＩＰへ出力処理を指示する．

ＥＩＰは，共有メモリ上のレコード識別子列をもとに，元のレコードを取り出し，ＰＲＯＪＥＣＴＩＯＮが指示されていれば指定されたカラムをならべ直して，出力ファイル形式へ再構成してディスクへ書き出す．

ディスクへの出力が完了すると，ＥＣＰはＥＩＰを通して，ホストへ演算の完了を通知する．

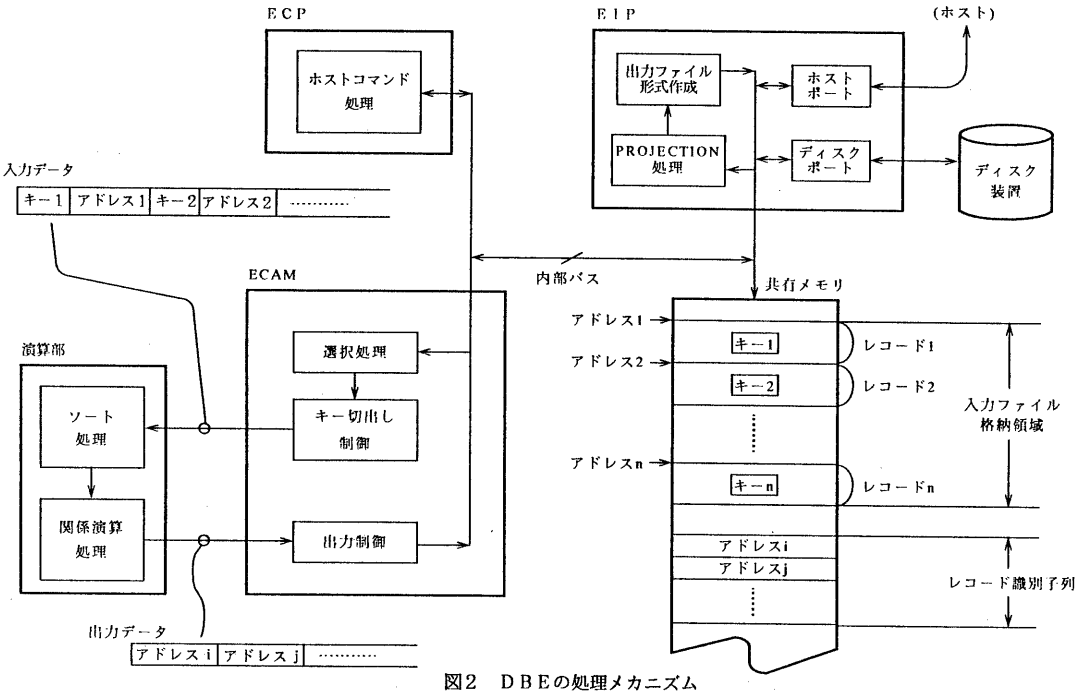


図2　ＤＢＥの処理メカニズム

3．ソフトウェア構成と処理方式

3．1　関係データベース制御システムとＤＢＥ

（1）ソフトウェア構成

図3のようにホスト上の関係データベース制御システム（以下ＲＤＢＣＳ）は，ユーザからのＳＱＬ問合せを解析し関係演算処理を行う．そのときＤＢＥで処理が可能な演算については，ＲＤＢＣＳがその演算の処理単位に対応するＤＢＥコマンド列を生成し，ＤＢＥに対して演算処理を依頼する．ＤＢＥは演算処理を行い結果をＲＤＢＣＳへ返す．

（2）索引の扱い

索引付きのカラムにより演算する場合には，それが一次索引ならばソフト処理となり，二次索引の場合には，ユーザ指定に従いＤＢＥ処理かソフト処理かのどちらかとなる．

（3）同時実行制御

同時実行制御は，

- 他のユーザにデータベースの共用を禁じる排他参照・更新モード
- 他のユーザからの参照のみを許す保護参照モード
- 更新中のデータベースでも共用を許す非整合読込みモード

の3つのモードをサポートしている．

特に非整合読込みは，データベースに対して統計処理を行いたいとき有効となる．一般に統計処理は非定型な問合せとして行われるものが多く，データベース全体を処理しなければならないために負荷が大きい．しかしＤＢＥを非整合読込みモードで使用することによりホストの負荷軽減を図ることができるうえ，高速処理が可能となっている．

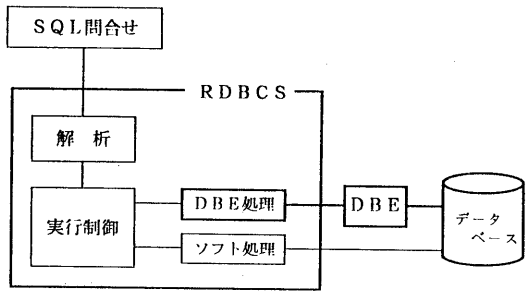


図3　ソフトウェア構成

（4）結果の格納

ＤＢＥが処理した結果は，通常は一時テーブルとして格納される．ＲＤＢＣＳは一時テーブルを入力しつつ処理を続行する．ただし，結果がそのままテーブルとして有効となるような条件が整えば，ＤＢＥからの結果はそのまま最終のテーブルとして格納される．このような条件には，例えばジャーナル属性の付いていないテーブルへの導出等が挙げられる．

（5）応用プログラムでのＤＢＥ使用レベル

DBEを使用する応用プログラムを考えたとき，優先的に問合せ処理の開始から終了までの間専有したい場合もあれば，特定の問合せ処理にのみ使用したい場合もある．このような要求に応えるために，いくつかの使用レベルをサポートしている．レベルには，ジョブ単位，アクセス対象のデータベース単位，ＳＱＬ単位等がある．

3．2　ＤＢＥが処理する演算

ホストのＲＤＢＣＳからのＤＢＥへの処理の依頼は，表1に示すコマンドから成るコマンド列により行われる．ＲＤＢＣＳから送られてきたコマンド列はＥＩＰを経由してＥＣＰに届けられ，ＥＣＰがそれを解釈しＥＩＰとＥＣＡＭを制御して処理を完了した後，ＥＩＰを経由してＲＤＢＣＳに完了状態を報告する．

特にＳＥＬＥＣＴコマンドとＪＯＩＮコマンドは，対象または条件のカラムに対して選択条件が指定されているとき，1つのＤＢＥコマンドとして展開される．この効果は，中間結果を生成することなく選択や結合ができるので，条件が複雑になった場合でもデータの移動を1回におさえることができ，演算の効率を向上することができる．この選択条件には比較述語，ＩＮ述語，ＮＵＬＬ述語，ＬＩＫＥ述語とＡＮＤ／ＯＲの論理式が含まれる．

3．3　ＤＢＥの処理方式

DBEはソートや関係演算処理を高速に行うハードウェアと，密結合された3つのプロセッサモジュール（ＥＩＰ，ＥＣＰ，ＥＣＡＭ）から構成される．各プロセッサモジュールはそれぞれの役割を効率よく処理するため，各々に適した独立の専用モニタにより制御される．

ＤＢＥコマンドの処理の構成要素としてはファイルからの入力，ＥＣＡＭでの処理，ファイルへの出力があるが，1つのコマンドの処理においてはこの3つのうちの必要とされるものが協調して並行に処理を行う．

ＤＢＥの処理方式には，ＤＢＥのハードウェア定数とデータ量により3つのレベルがある．すなわち，1つのコマンドの処理は，そのコマンドの実行時点での共有メモリの使用状況により通常処理と拡張処理に分かれ，さらに通常処理は，ＤＢＥで処理するデータ量が演算部のハードウェア定数を超えないとき内部処理となり，超えるとき外部処理となる．通常処理と拡張処理の違いは，対象及び結果データを共有メモリ内に保持できるかどうかにある．できないとき拡張処理となりＲＤＢＣＳから与えられた作業ファイルを使用することになる．

表1　ＤＢＥが行う関係演算処理

関係演算	ＤＢＥコマンド
テーブルの入力	ＲＥＡＤコマンド
探索条件中の選択	ＡＮＤ／ＯＲ条件付きのＳＥＬＥＣＴコマンド
〃　結合	ＪＯＩＮコマンド（選択条件を同時に指定可能）
副問合せ結果による選択	ＲＥＳＴＲＩＣＴコマンド
並び替え	ＳＯＲＴコマンド
グループ化	ＳＯＲＴコマンド
ユニーク化	ユニーク指定のＳＯＲＴコマンド
カラムの抽出とテーブルの出力	ＰＲＯＪＥＣＴＩＯＮコマンド
カウント	カウント指定の各コマンド
索引の生成	ＳＯＲＴコマンド

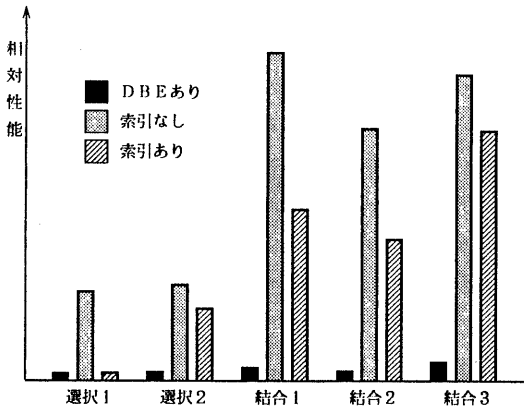
表2　ＤＢＥの処理方式

処理方式	処理方式の説明
通常処理	処理対象及び結果の総データ量が共有メモリを超えない場合の処理
内部処理	総レコード数と総キー長の組み合わせが演算部のハードウェア定数を超えずに1回で処理できる場合の処理
外部処理	ハードウェア定数を超えるために2回以上演算部を使用する場合の処理
拡張処理	処理対象及び結果の総データ量が共有メモリを超える場合の処理（作業ファイルを使用する）

4．性能

ここでは，ＤＢＥを使用した場合の性能とソフトウェアのみによる性能の比を示す．図４は処理対象データがＤＢＥの共有メモリに入る場合の関係演算の性能を示している．関係演算の測定には拡張ウィスコンシンベンチマーク１０万件版を使用した[4][5]．測定時のＤＢＥ構成は共有メモリ１２８ＭＢ，ソータは１８段である．ファイル形式はデータベース特有の統合編成である．

測定結果を見ると，索引のない場合は選択で１１倍以上，結合では１６倍以上の性能が得られた．一方，索引を利用した問い合わせ性能と比較すると，選択率が小さいほどソフトウェアによる方が高速であるが，選択率が大きくなるにつれてＤＢＥを使用したほうがより高速になることが確認できた．



選択１：１０万件→１０００件＝１０００件
選択２：１０万件→１万件＝１万件
結合１：（１０万件→１万件）×１０万件＝１万件
結合２：１０万件×１万件＝１万件
結合３：１万件×（１０万件→１万件）×（１０万件→１万件）＝１万件

注：→は選択，×は結合，＝は結果を示す

図４　関係演算性能比

5．おわりに

ＤＢＥのアーキテクチャとして，ホスト接続方式，ハードウェア構成，ソフトウェア処理方式及び関係演算の性能について述べた．ＤＢＥの特徴は，ディスクの直接アクセスによるホストのＣＰＵと主記憶の負荷軽減，ソートと関係演算を高速に行う専用ハードウェアの装備，３つのプロセッサモジュールによるＤＢＥ内部での機能と負荷分散，大容量共有メモリにある．ＤＢＥのサポートにより，大量データに対する関係演算の高速処理が可能となった．今後は，ホストの持つ機能のなかから負荷の大きいものをＤＢＥで実現していく予定である．

参考文献

[1]Sakai,H.,Iwata,K.,Shibayama,S.,Abe,M and Itoh,H.: Development of Delta as a First Step to a Knowledge Base Machine, in Sood, A.K. and Qureshi, A.H.(eds.), Database Machine Modern Trends and Applications, pp.159-181,Springer-Verlag, Berlin(1986).

[2]岩田，神谷，酒井，柴山，伊藤，村上：関係データベース処理エンジンのソータの試作と評価，情報処理学会論文誌,Vol.28,No.7,pp.748-757(1987).

[3]伊藤，島川，東郷，松田，伊藤，大場：可変長レコード用関係データベース処理エンジンの試作とソート処理性能の評価,情報処理学会論文誌,Vol.30,No.8,pp.1033-1044(1989).

[4]Bitton D.,D.J.Dewitt, and C.Turbyfill, "Benchmarking Database System - A Systematic Approach," Proceedings of the 1983 Very Large Database Conference,October, 1983.

[5]David J.Dewitt, "A Performance Analysis of the Gamma Database Machine",1988 ACM 0-89791-268-3/88/0006/0350