4N-4

スーパーデータベースコンピュータ SDCのアーキテクチャ

楊維康　平野聡　瀬川芳久　喜連川優　高木幹雄
東京大学生産技術研究所

1．はじめに

我々は，データベースマシンに関するこれまでの研究成果を基に，現在スーパーデータベースコンピュータSDCを構築中である．SDCは複数台の密結合マルチプロセッサで構成される処理モジュールをネットワークで疎に結合したハイブリッド並列アーキテクチャを採用している．マルチプロセッサで構成された各処理モジュールは，高速ディスクインターフェスとハードウェアソータを有しており，選択，ソート等の関係演算に対し，ディスクのデータ転送速度に追従できる処理速度を有し，結合等負荷の重い関係演算に対しては，ハッシュとソートによる新しい結合アルゴリズムを採用している．本稿では，SDCの全体アーキテクチャについて報告する．

2．SDCの開発背景

データベースマシン（DBM）の研究は既に20年の歴史があり，これまでに様々なDBMが提案され，試作されてきた．従来のDBMのアーキテクチャに関する研究は，大容量のデータ記憶系と高速処理系の間を繋ぐデータ転送路に於けるボトルネックの解消が，重要な研究課題であると言えよう．

又，従来のメインフレームでのデータベース処理環境では，ディスクに対するI/O 処理は，そのOSのI/O ハンドラの管理下で行われ，データの転送はページ単位で行われる．データベースの処理に於いては，ディスクに対するI/O が頻繁に行われ，その為のCPU オーバーヘッド及びディスクのアクセス特性による転送遅延が大きく，又，バッファ管理もデータベースに適した方式が採用されておらず，二次記憶アクセスに係わるもう一つの側面でのI/O ボトルネックが存在する．

一方，関係データベースの基本演算に対する高速処理アルゴリズムが種々提案されているが，それを効率的に実装するには，アーキテクチャ上のサポートが不可欠である．

以上の認識に基づき，高性能のDBMの研究に於いては，I/O の並列化と処理の並列化を有機的に結合したシステムを実現することが重要だと考えれる．我々は従来データベースマシンGRACE [1] 及び機能ディスクシステム[2] の研究を通して，アルゴリズムの提案からシステムの実装まで様々な実験を通して，有意義な経験が得られた．又，専用ハードウェアとしてLSI ソートチップ[3,4] を開発し，その成果はデータベースマシンGREOとして商用化された．今までの研究成果に基づき，我々はSDCのアーキテクチャを提案し，現在その構築を進めている．

3．SDCのアーキテクチャ

SDCのアーキテクチャを図1に示す．本章では，SDCの特徴を挙げ，そのアーキテクチャについて説明する．

Architecture of the Super Database Computer-SDC

W. Ynag, S. Hirano, Y. Segawa, M. Kitsuregawa, M. Takagi
Institute of Industial Sience,
University of Tokyo

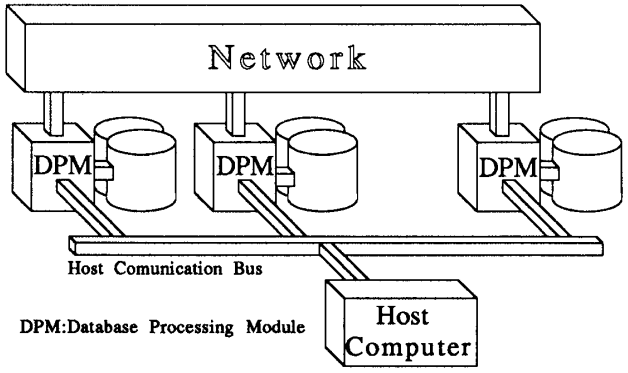


図1．SDCのアーキテクチャ

①ハイブリッド並列アーキテクチャ．SDCに於いてデータ処理系は，複数の密結合マルチプロセッサから構成されるデータベース処理モジュールをネットワークで疎に結合して構成される．この構成では，処理の並列性は複数台の処理モジュール，とモジュール内のマルチプロセッサの２つのレベルで実現されている．

②強化されたI/O システム．SDCの大容量２次記憶系は，高速ディスクを複数台用いて構成される．ディスクとのインターフェスには高速高機能ディスクコントローラを開発し，ページ単位のメモリ管理機構をデータ流に追従しつつ実現する．関係データベースのデータを水平に分割して，各ディスクに均等に格納し，ディスクが並列に動作することによってバンド幅の広いI/O スループットを得る．

③処理系と２次記憶の結合．SDCでは，二次記憶系と処理系は物理的に離れた構成ではなく，幾つかのディスクをグループ化し，処理モジュールと密結合して，複数台のディスクと処理モジュールを併せて一つの処理クラスターを形成している．このような処理クラスター構成は，従来のシステムに存在するI/O チャネルのボトルネックの問題を無くし，クラスター自体が２次記憶系にデータ処理系を持たせた機能化した記憶モジュールと見ることができる．このような構成では，ディスクからのデータストリームに対するFiltering,Hashing 等の前処理が，物理的にディスクに近い所で行い，不必要なデータ移動を避けることによって，ネットワークの負荷を軽減し，そしてシステム全体で高い性能を得ることができる．

④高機能相互結合網．各クラスター間のデータ交換がネットワークを通して行われる．ネットワークは多機能オメガネットワークを採用し，ハッシュ法による結合アルゴリズムをサポートする．

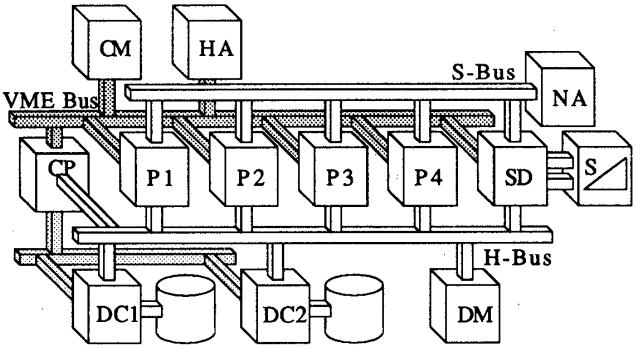
⑤Bimodal Sorting Memory[4]．SDCの処理モジュールには，19個のLSI ソートチップ[3] を用いて構成された大容量ハードウェアソータを有する．ソータの大容量メモリはソータに従属しているのではなく，モジュール内のプロセッサからも通常のメモリとしてアクセスできるようになっており，ソーティング機能を有するバイモードメモリと見なすことができる．

⑥ダイナミックGRACE ハッシュアルゴリズム[5]．我々が

開発したGRACE ハッシュは米国に於いて改良されHybridハッシュとして多用されているが，それを更に改良した動的GRACE ハッシュをアーキテクチャレベルで支援する．

4．ＳＤＣの処理モジュールの構成

ＳＤＣの処理モジュールは，複数のマイクロプロセッサ(CP,P1～P4,MC68020,20MHz) とメモリ(DM,CM),ディスクコントローラ(DC1,DC2),ソータ(S) 及び専用ハードウェアインターフェス(SD,HA,NA)を用いて構成される（図２）．



CP: Control Processor　　CM: Control Memory
DC1,DC2: Disk Controller　　DM: Data Memory
HA: Host Adapter　　NA: Nework Adapter
P1,..., P4: Processors　　S: Hardware Sorter
SD: Sorter Driver

図２．ＳＤＣの処理モジュールの内部構成

CPは制御プロセッサである．その機能は，モジュール内の資源の管理，ディスク制御及びデータ管理，各プロセッサの動作協調，各専用ハードウェアインターフェスの初期化及び制御等である．

P1～P4はデータ処理用のプロセッサである．関係データベースに対する総ての処理はP1～P4の並列動作によって行われる．

DC1,DC2 は高速データ転送インターフェスを有するディスクコントローラである．DCはSCSIインターフェスでディスクとつなぎ，ディスクの同期転送モードでディスク，DM間のデータのDMA 転送を行う．DCには，DMA 転送が一時中止しても，ディスクからの同期データ転送を中断しないように，大容量FIFOバッファメモリを有し，又，ディスクにページ単位の書き込みをサポートするハードウェア機構を有する．DCのインテリジェントな制御機構によって，ソフトウェアの負荷を軽減し，ディスクへのREAD/WRITE切替えの高速化を実現する．

DMは高速データメモリで，ディスクR/W 時のバッファ，プロセッサの中間処理結果の格納等に使用される．

CMはモジュール内の共用資源を管理する為のデータ構造，制御情報及び各プロセッサの動作を同期させる為の情報を格納するのに使用されるメモリである．

S は大容量ソートメモリであり,19 個のLSI ソートチップを用いて構成され，512Kレコードのソート能力を有し，最大８MBのデータを2.6MB/Sec.の転送速度でソートすることができる．ソータの大容量メモリはDMとオーバラップし，通常の大容量RAM としてH-BUS からアクセスできるようになっている．

SDはソータ駆動系であり，ソータとのインターフェスである．SDはソータに入力するデータに応じて制御フラグを自動的に生成する機能，ソータの出力をDMに転送するDMA機能を有する．

HAとNAはそれぞれホスト，ネットワークとのインターフェスである．

データベースに対する問い合わせ処理を行う時，大量のデータがディスク，メモリ，プロセッサ，ソータの間を移動する．処理モジュール内では，バスの負荷を分散させ，データ移動が渋滞すること無くスムーズに行われる為に，３つのバスを設けた.H-BUSは高速データバスで，データベース処理を行う為に，DMとDC1,DC2,P1～P4，ソータとの間のデータ移動に使用される．S-BUS はソータへのデータ入力，ネットワークを介してモジュール間のデータ転送に使用される．VME-BUS はシステム制御バスで，モジュール内の資源の制御情報へのアクセス，CPによる各プロセッサの制御，プロセッサ間の同期制御を実現する為のロック機構，割り込み等に使用されるバスである．

5．ＳＤＣに於ける処理方式

ＳＤＣでは処理単位はページ等の小さいgranule ではなく，オペランドリレーション全体が構成する巨大なデータストリームを処理単位とするデータストリーム指向の処理方式を用いる．高速ディスクインターフェスを介して，２次記憶からのデータを高速連続的に読み出す．総ての関係処理はディスクのデータ転送と重畳して，そのデータ流に沿って行われる．モジュール内のバスのバンド幅，マルチプロセッサの処理能力等が総てディスクからのデータ流に追従できるように設計されている．結合等従来リレーションサイズの自乗オーダーの演算はハッシュ・ソート・マージアルゴリズムで線型化し，ハッシュによる動的クラスタリングはネットワークを介して行われる．

6．むすび

本稿では，スーパーデータベースマシンＳＤＣのアーキテクチャについての概要を述べた．単一モジュールは既に一部稼働しており，マルチプロセッサによって，ディスクの最大速度で転送されたデータ流に沿って関係処理が行われることが既に確認され，高い性能が得られている．システム全体を完成させ，精確な性能評価を行うことが今後の研究課題である．

〔参考文献〕

[1] Kitsuregawa,M. et.al."Architecture and Performance of Relational Algebra Machine GRACE",Int. Conf. on Parallel Processing 84, 1985.

[2] Kitsuregawa,M. et.al."Functional Disk System for Relational Database", Proc. of the 3rd Int. Conf. on Data Engeneering, pp.88-95, 1987.

[3] 楊，他『LSI ソートチップの試作』
情報処理学会第37回全国大会 7Q-4, 1988.

[4] Kitsuregawa,M. et.al."Implementation of LSI Sort Chip for Bimodal Sort Memory",the Int. Conf.on Very Large Scale Integration, Aug.1989,Munich,West Germany

[5] Kitsuregawa,M. et.al."The Effect of Bucket Size Tuning in the Dynamic Hybrid GRACE Hash Join Method" 15th Int.Conf. on Very Large Data Bases, Aug.1989, Amsterdam,The Netherlands