2モジュール構成のスーパーデータベースコンピュータ (SDC) の試作と評価

3L−7

平野聡 原田昌信 中村稔 相場雄一 鈴木和宏 喜連川優 高木幹雄 楊 維康
東京大学 生産技術研究所

1 概要

我々は高並列リレーショナル・データベース・サーバーである、「スーパーデータベース・コンピュータSDC」を開発している。モジュール1台からなる試作機は既に稼働しており、商用データベース・マシン Teradata (プロセッサ20台、ディスク40台) と比較して20倍から50倍と高い性能を有することを示した [1]。

本論文では、多モジュールSDCの実現に向けて、ネットワーク・インターフェースの試作、及び、多モジュールの制御方式を確立する事を目的とし、更に1台のモジュールを試作して2モジュールでの動作確認を行なったので、その結果について報告する。

2 SDCについて

SDCのハードウェアは、プロセッサ4台と磁気ディスク装置2台を密に結合し「処理モジュール」とし、それらを高機能オメガ・ネットワークで疎に結合したハイブリッド・アーキテクチャをとる。この構成では、密結合の利点である軽い通信コストによる高速性と、モジュール数の増減によるスケーラビリティが同時に得られる [1]。

SDC のソフトウェアは、並列関係演算アルゴリズム、多モジュール制御方式、プロセス・モデルの各レベルで高並列化を図る。並列関係演算アルゴリズムのレベルでは並列 GRACE Hash アルゴリズム [4] などを用い、データ処理の並列性を引き出す。その際、多モジュールに渡った制御が必要となる [4章]。

プロセス・モデルのレベルでは、ディスクやネットワークといったI/O処理とプロセッサ群による処理をオーバーラップしI/Oからのデータを「On The Fly」に処理する事、及び、複数のプロセッサ間の負荷を均等に分散する事が必要になるが、I/Oバウンドのデータベース処理に「プロセッサがI/Oを制御する」という立場をとる従来のオペレーティング・システムのプロセス・モデルを導入しても性能の向上は見込めない。そこでSDCでは、「I/Oがプロセッサを駆動する」モデルを取る。SDCのプロセス・モデルはハードウェアのハイブリッド・アーキテクチャと密接に関連しており、モジュール内のプロセス間通信は共有メモリ・アクセス、モジュール間のプロセス間通信はメッセージ・パッシングという形態を取る。また、プロセスからはディスクならびにネットワークからのデータを同一の手順で扱うことが可能であるという大きな特徴がある [2]。

Implemantaion and Evaluation of two module SDC, The Super Database Computer
S.Hirano, M.Harada, M.Nakamura, Y.Aiba, K.Suzuki, M.Kitsuregawa, M.Takagi
Institute of Industrial Science, University of Tokyo

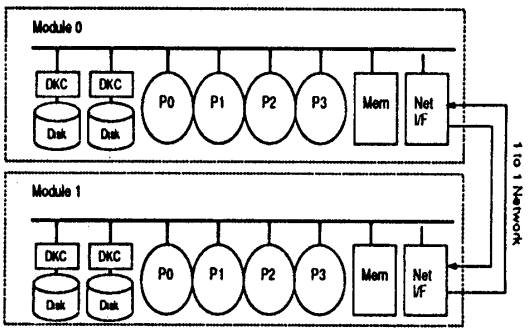

図1: 2モジュール版SDCの構成

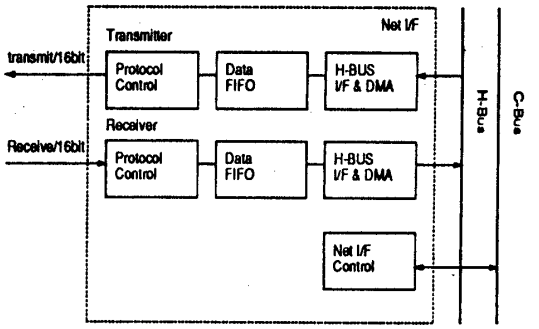

図2: ネットワーク・インターフェースの構成

3 ネットワーク・インターフェースの構成と動作

試作したハードウェアの構成を図1に示す。1モジュール内にプロセッサ MC68020 (20MIIz) を4台、ステージング・バッファ・メモリを8 MB、8インチ・ディスクを2台備える。今回、新たにネットワーク・インターフェース (図1中 Net I/F) を付加した。Net I/F の内部構成を図2に示す。構成要素は以下の通りである。

H-BUS IF − データ転送用共有バス・インターフェース
H-BUS DMA − SDC のプロセス・モデルを支援するタスク生成機能を有する DMA コントローラ
Data FIFO − データ 用 FIFO 8K バイト
Protocol Control − ネットワークの回線接続、ハンドシェークなどを行なう
Net I/F Control − Net I/F の設定、タスク生成のイベント通知を行なう

ネットワークはデータ幅16ビットの回線接続であり入力チャネル、出力チャネル2系統を有する。モジュール間は本来、高機能オメガ・ネットワークで結合されるのであるが、今回は2モジュールであるので1対1に直接接続した。

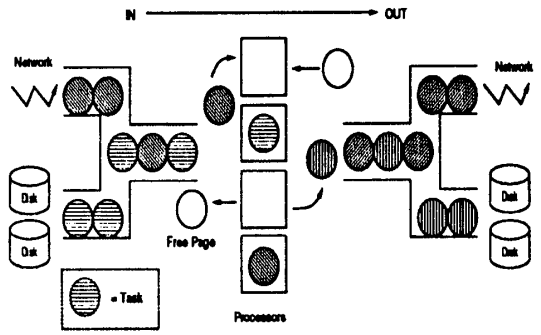

図3: ネットワーク上のデータの流れ

Net I/F は以下のように動作する。プロセッサ群は並列関係演算アルゴリズムに従って宛先タグを付加したタプルを数個まとめてネットワークへのタスクとし、共有メモリ上に構成される仮想的なパイプに送り出す(図3)。Net I/F はパイプからタスクを拾得すると、順に Data FIFO に格納する。Protocol Control 部は Data FIFO に1タプル分のデータが揃うと起動され、FIFO から取り出したタプルに付いている宛先タグに従って相手側モジュールに回線の接続を試みる。その際、相手側 Net I/F の Data FIFO に1タプル分以上の空きができるまで、回線の開設は保留される。回線接続が完了した後、両側の Net I/F はハンドシェイクを行ないながらデータ転送を行なう。回線の切断は送信側のデータ転送終了によって行なわれる。

Net I/F の受信側の動作もほぼ同様である。

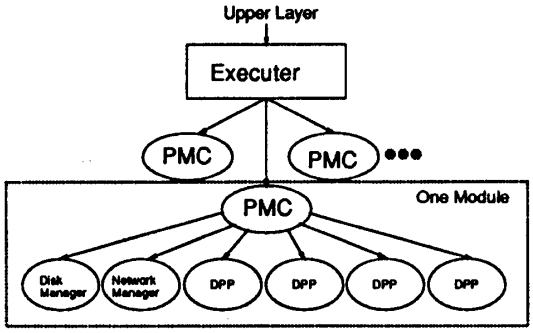

図4: ソフトウェアの構成

4 2モジュール SDC の制御方式

図4にソフトウェアの構成を示す。図中、点線で囲まれた部分が1モジュールを表している。構成は全モジュールで共通しており、試作機では2モジュール分が存在する。モジュール群の制御方式は以下のようになる。ここでは例として GRACE HASH アルゴリズムを用いた結合演算を取り上げる。GRACE HASH アルゴリズムでは、全モジュールの全ディスクから入力リレーションを読み、各タプルをハッシュ値によってネットワークを介してモジュール間に分配し(スプリット・フェーズ)、その後各モジュール内で独立に結合(ジョイン・フェーズ)を行なう。

1. 結合演算命令が Executer に送られる。
2. Executer は各モジュールの PMC (モジュール管理プロセス) にスプリット・フェーズの開始を指示する。
3. 各 PMC はそれぞれのモジュール内の Disk Manager, Network Manager, データ処理プロセス群 DPP (1プロセッサに1 DPP) にスプリット・フェーズの開始を指示する。
4. ディスク、ネットワーク、プロセッサは上記のプロセス・モデルに則り、タスクのやりとり、即ちデータ処理を行なう。
5. PMC は入力リレーションのディスク入力とネットワーク送出が終了すると Executer にその旨報告する。
6. Executer は PMC からの報告を受け、モジュール群の同期を取る。
7. PMC はネットワークから来るタスクを処理し終った後、Executer に処理終了を報告する。
8. Executer は全 PMC から処理終了の報告を受けるとスプリット・フェーズを終了とし、ジョイン・フェーズを開始する。
9. ジョイン・フェーズを実行の後、終了。

このように、モジュール群の大局的な同期は Executer によって明示的に統括され、データ処理そのものに伴う同期や通信はプロセス・モデルによって暗黙的に行なわれる。

5 2モジュール SDC の性能評価

試作機上で拡張ウィスコンシン・ベンチマーク・テストを行なった。条件は、タプル長208バイト、タプル数100万、選択率10%の結合演算である。表1に結果を示す。

これから2モジュール SDC の構成、制御方式が妥当であることが確認できた。

1モジュール SDC	127.7
2モジュール SDC	64.0
20ノード teradata	3535

表1 (単位: 秒)

参考文献

[1] 平野、原田、中村、小川、楊、喜連川、高木「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」並列処理シンポジウム, 1990

[2] 平野、原田、中村、楊、喜連川、高木「スーパーデータベースコンピュータ SDC のソフトウエア」電子情報通信学会技術研究報告 Vol.90 No.144, 1990

[3] 喜連川、小川「パケット平坦化機能を有するオメガネットワーク」, 情報処理学会論文誌第30巻第11号 p1494,1989

[4] Kitsuregawa M., Ogawa Y., "A New Parallel Hash Join Method with Robustness for Data Skew in Super Databas Computer (SDC)", Proc. of International Conf. on Very Large Data Bases, 1990