ATM 結合 PC クラスタ 上での並列関係データベースサーバの構築

1R－2

田村孝之　小口正人　喜連川 優
東京大学 生産技術研究所

1 はじめに

近年，マイクロプロセッサや標準バスの高速化によってパーソナルコンピュータの性能は著しい向上を続けており，また 10 Mbps Ethernet に代わる次世代標準 LAN の候補として幾つかの方式が提案され，急速に技術開発が進められている．このような状況から，安価な PC を標準的なネットワークで結合した PC クラスタは，商用超並列計算機はもちろん，ワークステーションクラスタと比べても価格対性能比で圧倒的に有利であり，研究が活発化しつつある．

しかし，これまでに行われた PC クラスタに関する研究ではノード数が十数台以下の小規模なシステムが対象であり，アプリケーションも数値演算に限られていた．一方，社会的なニーズとしては，企業を始めとする様々な組織において電算化が浸透した結果蓄積された膨大な量のデータを多角的に分析することへの需要が高まっており，高いデータ処理性能を持つシステムを安価に構築することは極めて重要な課題であると考えられる．

我々はこれまで関係データベースにおける大規模問合せ処理の高速化を目指し，高並列関係データベースサーバ SDC-II の開発を行ってきた [1]．SDC-II においては，共有バス，磁気ディスク I/F，相互結合網などをボードレベルから設計し，専用のシステムソフトウェアと高性能な並列関係演算アルゴリズムを実装することで，マイクロプロセッサによる並列関係データベース処理の有効性を実証することができた．今回，この研究で得られた成果を更に大規模な問題に対して適用し，PC クラスタによる超並列関係データベースサーバ実現の可能性を示すため，100 台規模の PC を ATM スイッチを介して結合した大規模 PC クラスタを構築し，並列関係データベース処理系を作成したのでその概要を報告する．

2 ATM 結合 PC クラスタの構成

今回我々が構築した PC クラスタは，図 1に示すように 200 MHz Pentium Pro を搭載するノード 100 台を 155 Mbps の ATM および 10 Mbps Ethernet の 2 系統のネットワークで相互結合したシステムである．各ノードは 64 MB の主記憶を持ち，OS 用の IDE ディスクとデータベース格納用の Ultra Wide SCSI ディスク (4GB, 約 10 MB/s) を内蔵している．ATM スイッチには 128 ポート の UTP-5 インタフェースを持つ日立製 AN1000 を用い，ATM NIC (Network Interface Card) には米 Interphase 社製 5515 PCI ATM Adapter を用いた．

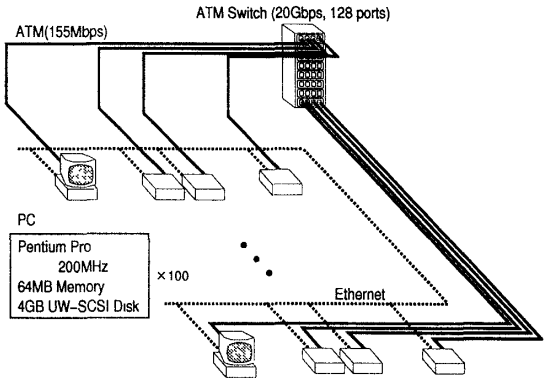

図 1: ATM 結合 PC クラスタの構成

各ノードでは，Solaris 2.5.1 オペレーティングシステムが稼働しており，ATM NIC に対しては PVC 及び SVC 上の TCP/IP プロトコルによる通信が可能である．TCP を用いたときのポイントツーポイントの最大実効スループットは 115 Mbps (8 KB/message) 程度である [2].

3 PC クラスタにおける並列関係データベースサーバの実装方式

大規模データベースに対する関係演算処理においては，大量のデータの移動に伴って処理が進むため，ディスクやネットワークに対する入出力操作には高い効率が要求される．SDC-II においては，ディスクおよびネットワークに専用の管理 CPU を用いてメインの CPU を割り込み処理やバッファ割り当てなどの低レベルの処理から解放し，さらにプログラムを常にカーネルレベルで動作させることで，コピーオーバヘッドの排除や関係演算アルゴリズムに適したバッファ管理方式の採用を可能にしていた．しかし，PC クラスタにおいては，最終的にはカーネルレベルでの実装を目指すものの，可搬性と開発期間の短縮が重要であると考え，ユーザレベルプロセスによる実装を行った．その際，OS のレイヤをハードウェアと見做すことでプログラムの構造をできるだけ同一に保ち，また OS 依存の部分を最小限に留めるようにした．

サーバプロセス自体は，Solaris のスレッドライブラリ (POSIX pthread とほぼ互換) を用いてマルチスレッ

Implementation of a parallel relational database server on an ATM connected PC cluster
T. Tamura, M. Oguchi, and M. Kitsuregawa
Institute of Industrial Science, The University of Tokyo
7-22-1, Roppongi, Minato-ku, Tokyo 106, Japan.

ドにより実現している．以下の 6 つのスレッドが常駐している．

1. File flush daemon
2. Disk I/O daemon
3. Network transmitter daemon
4. Network receiver daemon
5. Query executor daemon
6. User request server

この内，ディスクとネットワークに関する I/O daemon はデバイスドライバの動作をシミュレートするためのものである．SDC-II においては，ファイルを読む際にページ毎に I/O リクエストを発行するのではなく，最初にファイル全体に対するリードリクエストを発行し，順次バッファを割り当てて格納することで完全な非同期 I/O を実現していた．PC クラスタではこの方法は取れないため，Disk I/O daemon がファイル全体に対するリードリクエストをページ毎の I/O に変換するようにしている．ネットワークに関しては，エラー回復のためのプロトコルが未実装のため，各ノードに対して TCP ソケットを開いて通信を行っている．一方，ファイルの書き込みにおいては delayed write により可能な限りメモリ上のバッファに残すようにしているため，フリーメモリの量がスレッショルドを下回った時に，オープンされているファイルのリストを走査してバッファ消費量が最も大きなものからディスクにフラッシュする処理を File flush daemon が受け持っている．

問合せ実行時の流れは以下のようになる．ユーザアプリケーションから発行された SQL 問合せ文は，コンパイル・最適化の後に実行可能オブジェクトとして出力される．そして，各ノード上の User request server に対して実行要求メッセージが送られ，その結果 User request server は指定されたオブジェクトを動的にリンクし新たなスレッドを生成して実行を開始させる．各ノード上のスレッドからの出力は User request server を経由し実行結果としてユーザアプリケーションに返される．

問合せ毎に新たに生成されるスレッドは，入出力をオープンし，入力データを処理するコールバックルーチンを設定した後は休止状態となる．実際にデータを処理するのは Query executor daemon であり，あらゆる入力イベントをこのスレッドが扱う．これは，スレッドの切り替えに伴うオーバヘッドを回避し，各ルーチンの実行順序の制御を容易にするためである．

4 ベンチマークによる予備評価

本システム上で，アドホックな問合せ処理の標準ベンチマークである TPC-D [3] の問合せを実行した時の性能を，公表されている商用並列 DBMS の値と共に表 1に示す．

ここでは，30 ノード上に 10GB のデータベースを作成したが，SCSI ディスクに関しては最終仕様と異なり，2 GB，約 5.5 MB/s のものを用いている．最終的には，100 ノード上に 100GB のデータベースを作成することを目標としているので，1 台当たりの処理負荷は約 3 倍になる．Query 3 (3-way 結合演算) や Query 5 (6-way 結合演算) で商用機の性能が相対的に高くなっているが，これはインデックスによりアクセスするデータ量が減少しているからである．我々は，ワーストケースでの性能を評価するために，インデックスの使用は避けた．これに対して，Query 1 (集計演算) ではインデックスの効果が無いため条件は等しい．また，Query 9 (6-way 結合演算) においてもインデックスの効果が薄い．これらの結果から，我々のシステムは最先端の商用機と比較しても十分高い処理性能を有していることが分かる．

表 1: TPC-D 問合せの実行時間

System	PC Cluster	NCR 5100M
CPU	Pentium Pro 200 MHz × 30	Pentium 133 MHz × 160
Memory size	64MB × 30	1GB × 20
Disk capacity	2GB × 30	2GB × 400
Database size	10GB	100GB
Query 1	63.1 [s]	630.5 [s]
Query 3	82.6 [s]	440.1 [s]
Query 5	88.9 [s]	342.1 [s]
Query 9	105.4 [s]	953.3 [s]

5 まとめ

本論文では ATM で相互結合された 100 台の PC からなる PC クラスタについてその概要を述べた．TPC-D ベンチマークによる予備的な評価により PC クラスタが十分な性能を持っていることを確認できた．

参考文献

[1] 中村，平野，田村，喜連川，高木．スーパーデータベースコンピュータ SDC-II におけるシステムソフトウエアの設計と実装．信学論 Vol.J78-D-I, No.2, pp.129–141, 1995.

[2] M. Oguchi, T. Shintani, T. Tamura, and M. Kitsuregawa. Preliminary Experimental Results of a Parallel Association Rule Mining on ATM connected PC Clusters. Intl. Symp. on CODAS, pp. 278–281, 1996.

[3] Transaction Processing Performatce Council. TPC Benchmark^{TM}D (Decision Support) Standard Spec. Rev. 1.1. 1995.