スーパーデータベースコンピュータ SDC のシステム・ソフトウエアの概要

4N-5

平野 聡　楊 維康　喜連川 優　高木 幹雄
東京大学 生産技術研究所

1 概要

本論文ではスーパー・データベース・コンピュータ SDC のシステム・ソフトウエアの構成について概観する。 SDC[1] は、プロセッサ4台、磁気ディスク装置2台、ハードウエア・ソータをバス共有型の密結合としてクラスタ化し、クラスタ間をオメガ・ネットワークで結合する構成をとる。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、クラスタ数の増減によるスケーラビリティが同時に得られる。

以下、SDC のシステム・ソフトウエアの全体構成、クラスタ内の構成、プロセス間通信について述べる。

2 SDC のシステム・ソフトウエアの全体構成

[図1] に SDC のシステム・ソフトウエアの全体構成を示す。ホスト計算機上には DBMS 及びユーザ・インターフェース・プロセスが常駐する。DBMS はユーザ・インターフェース・プロセスから与えられた adhoc query や、アプリケーション・プログラムからのトランザクション要求を受け付けると、排他制御、最適化を行ない、基本演算列に変換する。基本演算は結合演算の分割フェーズと演算フェーズのようにサブコマンドで構成される。DBMS はサブコマンドをメッセージとしてコントロール・ネットワークを介して各モジュールに与える。クラスタ間の同期制御はサブコマンドの発行とそのコマンドに対する終了報告で実現される。

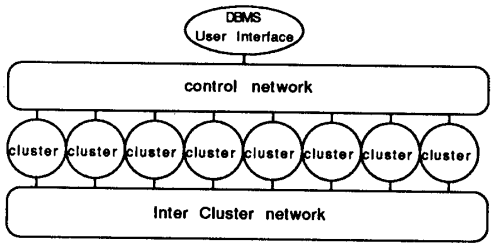

図 1: SDC のソフトウエアの全体構成

3 クラスタ内のソフトウエアの構成

[図2] にクラスタ内のソフトウエアの構成を示す。図中において、細い線はコントロール・バスを通るコントロール・メッセージの流れを表し、太い線は データ・バスを通るデータの流れを表している。また、破線に囲まれた部分はひとつのプロセッサ上で実行される。クラスタ内のプロセスには以下のものがある。

- Cluster Control Process (CCP): クラスタ内の他のプロセスを管理する。
- Disk Manager (DiskMan): ファイルの論理的な構造をディスク上の物理的な構造にマッピングする。
- Memory & DMA Manager(MemMan): ディスクからのデータをバッファ・メモリ上に展開する。バッファ・メモリは4キロバイトから、10キロバイトほどの大きさの固定長のページとして管理されている。[図3] ページはディスクリプタ部とデータ部を持ち、ディスクリプタ部はコントロール・バス上のメモリへ、データ部はデータ・バス上のメモリへ置かれる。
- SCSI Driver: SCSI コントローラを制御する。転送速度 2.5MB/sec の同期バスが2本あり、2台のディスクが同時に動作する。
- Inter Cluster Network Manager(NetMan): インター・クラスタ・ネットワークを介したデータ・メッセージ・パッシング（後述）を管理する。
- Sort Manager(SortMan): ハードウエア・ソータの初期化、DMA 転送の管理をする。
- Data Processing Process(DPP): 選択、ハッシング等の実際のデータ処理を行なう。
- Massage Passing Manager(MsgMan): プロセス間、クラスタ間のコントロール・メッセージ・パッシングを管理する。このプロセスは全てのプロセッサ上に存在する。
- Network Activation and Network Notification Observer (Nanno): リモート・プロシジャ・コールのサーバである。このプロセスは全てのプロセッサ上に存在する。

クラスタは Cluster Control Process(CCP) によって管理される。CCP はホスト計算機上の DBMS からのメッセージを解釈し、処理内容に従って各プロセスにコマンドを送る。

4 制御の流れ

結合演算を例に取り、制御の流れについて解説する。結合演算の処理方式は Grace Hash[2] を SDC のアーキテクチャに最適化して用いる。分割フェーズでは、各クラスタのディスクに水平に分割して格納されているリレーションを読み出しながら、ハッシュをかけ、バケットをインター・クラスタ・ネットワーク介して、クラスタ間に分散する。演算フェーズでは、バケットを収集しながら突き合わせを行なう。

分割フェーズ制御の流れを以下に示す。

1. DBMS が各クラスタの CCP にジョイン要求メッセージを送る。
2. CCP が4つの DPP にジョイン要求メッセージを送る。
3. CCP が DiskMan にリード・ライト要求メッセージを送る。
4. CCP が NetMan に入出力要求メッセージを送る。
5. DiskMan がディスク・リードと DMA をスタートする。
6. ディスクからデータが来ると、MemMan がリード・バッファに格納する。MemMan はディスクから読み出されたタプルをフリーリストから持ってきた空きページに詰め、FIFO を構成しているリードバッファ・リストの入力端へつないでゆく。([図3] 中の In)
7. DPP はリード・バッファ・リストの出力端を監視している。ページがつながれるとそのページを獲得して、選択処理、ハッシングを行ない、インター・クラスタ・ネットワークに送出する。（他のシステムでは、数メガバイトのディスク・バッファを2面持ち、交互に使うことが多いが、SDC ではこのようなメモリ管理方式によりわずか 50 キロバイトで済んでいる。）

---

[0] SDC, The Super Database Computer, Overview of the system software

S.Hirano, W.Yang, M.Kitsuregawa, M.Takagi
The Institute of Industrial Science, University of Tokyo

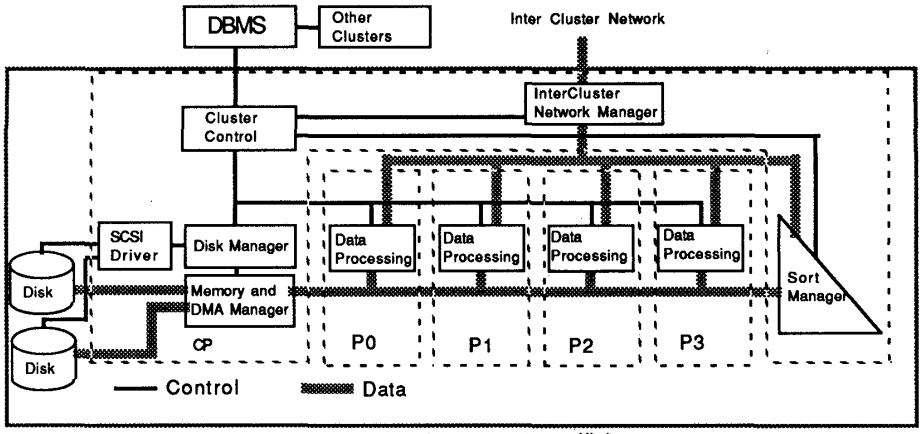

図 2: クラスタ内のソフトウエア構成

8. インター・クラスタ・ネットワークからのデータはディスクのライトバッファに溜められる。バッファがニア・フルになるとDiskManはディスク・リードを中断し、バッファをディスクにライトする。
9. リレーションを全て読みだし、リードが終了すると、DiskMan、DPPはCCPに終了を報告する。
10. CPPはDBMSにディスク・リードの終了を報告する。
11. DBMSは全てのCCPからの終了報告を受領後、各CCPにライト・バッファのフラッシュを要求する。

このようにCCPはデータベースの基本演算についてのシナリオを持ち、ディスク2台の転送速度に追従して処理が行なわれるよう、クラスタ内の計算資源をスケジューリングする。

5 プロセス間通信

SDCのシステム・ソフトウエアのインフラ・ストラクチャであるプロセス間通信には、

- コントロール・メッセージ・パッシング
- データ・メッセージ・パッシング
- 共有メモリ・アクセス

の3種類がある。

コントロール・メッセージ・パッシングによるプロセス間通信は、メッセージ・パッシング・マネージャMsgManとサーバ・プロセスNannoによって実現されており、クラスタ内、クラスタ間、ホスト/クラスタ間で、プロセスの生成、コマンド送付、終了通知といった実行制御に使用される。プロセス間の多対多のメッセージの送受、リモート・ファイル・アクセス、リモート・システム・コール、仮想端末等の豊富な機能があり、rsh、rcp等のユーティリティと、プログラミング言語ライブラリから利用する。

共有メモリ・アクセスによるプロセス間通信はMemManとDPPとの間でディスク・データの処理を行なうために用いられる。ここでは、非常に高速な処理が必要とされるので、メモリ上のデータを共有し、スピン・ロックで互いに同期ををとりながら協調して動作する。ところが、スピン・ロックを多用すると、バスに大きな負担をかけ、全体の性能が著しく低下することが知られている。そこでSDCでは、クラスタ間のネットワークと、クラスタ内のバスについて、コントロールとデータの物理的なパスを完全に分離し、さらにデータ用バスについてはバス・ボトル・ネックを解消するため、データ流に沿って複数のバスを配置した。

ネットワークは、

- コントロール・ネットワーク: コントロール・メッセージ・パッシング用。
- インター・クラスタ・ネットワーク: データ・メッセージ・パッシング用。バケット平坦化等の機能を持つ。[3],[4]

の2組。バスは、

- VME-BUS: コントロール・メッセージ・パッシング、ページによるメモリ管理用。
- SCSI-BUS × 2: ディスク・データ転送用。
- H-BUS: ディスク・コントローラ、プロセッサ間のデータ転送用。
- S-BUS: プロセッサ、ソータ/インター・クラスタ・ネットワーク・アダプタ間のデータ転送用。

の4組である。

6 終りに

以上、SDCのシステム・ソフトウエアの構成について概観した。SDCは現在1クラスタのプロトタイプ上で基本性能の評価と、ネットワークの実装を行なっている。

参考文献

[1] 楊、平野、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」情報処理学会第39回全国大会、1989

[2] M.Kitsuregawa,H.Tanaka,T.Moto-oka, Architecture and Performance of Relational Algebra Machine GRACE, Proc. of Int.Conf. on Parallel Processing,1984

[3] 小川、喜連川、「スーパーデータベースコンピュータにおけるバケット分散並列結合演算法とその性能予測」情報処理学会第39回全国大会、1989

[4] 喜連川、小川、「バ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に投稿中

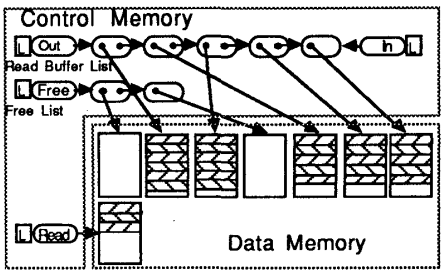

図 3: ページによるメモリ管理（リード・バッファ）