5H−9

スーパーデータベースコンピュータ(SDC)における データネットワークの論理設計

田村孝之 原田昌信 平野聡 中村稔 喜連川優 高木幹雄
東京大学 生産技術研究所

1 はじめに

スーパーデータベースコンピュータ(SDC)は，SQLを高速に実行する高並列リレーショナルデータベースサーバである．SDCは，数台のプロセッサと磁気ディスク装置とを共有バスで密に結合した処理モジュール(PM)を相互結合網で疎結合したハイブリッドアーキテクチャをとる[1][2]．さらに，共有バスと相互結合網はそれぞれがデータ用およびコントロール情報用に2重化されている．

SDCでは基本的に，ディスク装置から読み出されるデータの流れを止めずに関係演算処理を行なうことにより高速化を図っている．したがって，SDCのデータネットワークに対しては高速性が強く要求される．また，特定の関係演算アルゴリズムの一部を処理することや，モジュール間の同期を簡単にする手段を与えることも望まれる．

我々は，SDCにおけるデータネットワークとして，バケット分散ハッシュ結合アルゴリズムを支援するバケット平坦化機能付きのオメガネットワークを提案し[3]，その有効性をシミュレーションにより示した[4]．

本稿では，提案されたデータネットワークの構成要素である，スイッチングユニットおよびネットワークとPM間のインターフェース(以下ネットワークアダプタと呼ぶ)の設計について述べる．

2 データネットワーク系の構成

SDCは現在，1つのPMがプロセッサ(M68040)5〜7台と磁気ディスク装置4台とから構成される第2版の製作が進行中である．今回，このモジュールを用いたデータネットワークとして，図1に示すような16×16オメガネットワークを設計の対象とした．

一つ一つのスイッチングユニットは2×2クロスバースイッチから成り，クロック毎に各段を入力側から出力側へとパイプライン式にデータが送られていく．その際に，送り側からは現在データライン上にあるデータの有効/無効を示す制御信号を出力し，受け側ではビジー状態かどうかを示す信号を前段に伝える．

また，ネットワークに出力するタプルの先頭には，接続のモードとIDを示す2バイトのヘッダと2バイトのタプル長を付ける．各スイッチでは，これらの情報により経路を自律的に決定する．

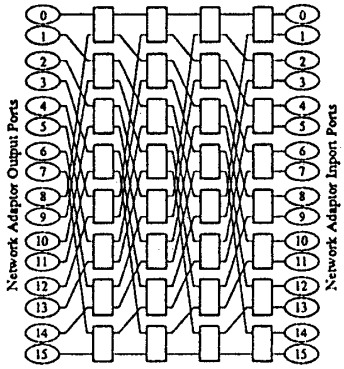

図1: データネットワーク系の構成図

さらに，ネットワーク上にマネージメントプロセッサを置き，コントロールネットワークを通じての処理モジュールとの通信，ネットワークの初期設定，処理フェーズの切替え，エラー処理などの大域的な処理を行なう．

3 ネットワークアダプタの構成

ネットワークアダプタの基本動作は，II-BUS(データ用高速バス)上でDMAを行ない，共有メモリ(DM)上のデータをネットワークに出力することと，ネットワークから入力したデータをDMに書き込むことの2つであり，内部構成は図2のようになっている．

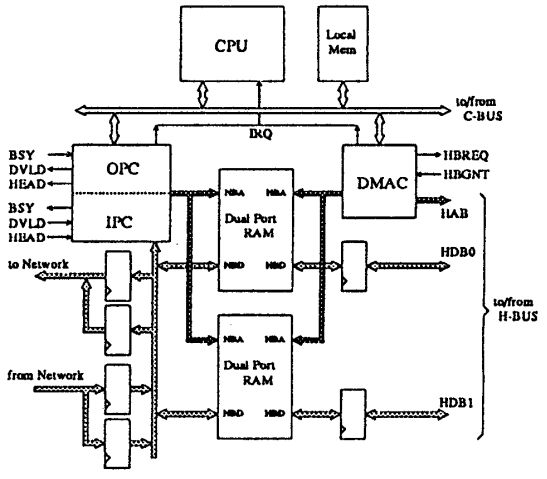

図2: ネットワークアダプタのブロック図

モジュール内の処理はページを単位として行なわれるが，ネットワークとはタプルを認識しながらやりとりする必要があるので間にバッファを置いて処理を分離し，制御を簡単化する．さ

Logic Design of the Data Network for The Super Database Computer SDC
T.Tamura, M.Harada, S.Hirano, M.Nakamura, M.Kitsuregawa, M.Takagi
Institute of Industrial Science, University of Tokyo

らに，デバイスとしてデュアルポートRAMを用いることにより，タイミング系もモジュールの内側とネットワーク側で切り離すことが出来る．各モジュールは独立したクロックで動作しているが，ネットワーク側の制御回路をネットワークから供給されるクロックに基づいて動作させることにより，モジュール間の非同期性の吸収を図る．

バッファとDMとの間のデータ転送を受け持つのがDMAC (H-BUS DMA Controller) である．2重にインターリーブされたデータバス，連続ワード転送のサポート，バッファRAMの2バンク構成などにより非常に高速なDMAを行なうことが可能である．バスが飽和していなければ，DMからの16バイトの読み出しとDMへの16バイトの書き込みがそれぞれ360ns，280nsで完了する．

OPC (Output Port Controller) はネットワークへの出力を制御する．バッファに蓄えられた1ページ分のデータを，各タプルの先頭に付加されたデータ長を読むことによってタプル毎に切り分けて出力する．また，ヘッダを出力する時には初段のスイッチに対して接続動作を行なうように伝える．

一般に，ページ内のデータは末尾に空白を含むため，これらを有効なデータと区別しなければならない．このためにページ内の有効タプル数を共有メモリ上から得て，1タプル出力する毎にデクリメントすることによって，ページの終りを検出する．

IPC (Input Port Controller) はネットワークからの入力を行なう部分であるが，OPCに比べて複雑になっている．モジュール内では，コンテナモデルに基づきページ単位の処理が行なわれるために，ネットワークから入力されたタプルをまとめてページにし，プロセッサの待行列に入れなければならない．この際に，異なるページは異なるプロセッサによって処理される可能性があるのでタプルがページ間にまたがることは許されない．そこで，一般には次に来るタプルのデータ長が未知のため，タプルの先頭を入力してデータ長を読んでからページがあふれるかどうかを判断し，あふれるようだったらそのタプルは次のページの先頭から書き込むという制御を行なう．

DMAC, OPC, IPCのそれぞれはページの終了を検出すると制御用のCPUに対して割り込みを上げ，CPUが新たなページに対する入出力を再起動するようになっている．このCPUはモジュール内の他のプロセッサとC-BUS (制御用バス) を通して接続されており，共有メモリ上からページの先頭アドレスを取って来たり，空になったページを解放したりする働きを持っている．

4 スイッチングユニット

スイッチングユニットの接続動作には，通常の行先ノード指定モードの他にバケット平坦化モードがあり，これはさらにタプル数に基づく平坦化とタプル長に基づく平坦化とに分けられる．どちらの平坦化も経路の決定法は同じであるが，履歴情報として用いるものが異なっている．これらのモードの切替えは，タプルのヘッダの上位2ビットに基づいて行なわれる．

スイッチングユニットのブロック図を図3に示す．

通常モードでは，ヘッダの下位ビットをノードアドレスと解釈して接続を行なう．スイッチに新たなタプルが到着した時のクロック毎の状態遷移は，

1. ヘッダをレジスタにラッチ
2. ポート間での行先の比較

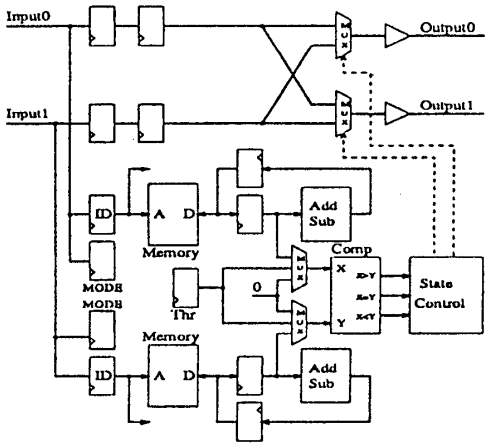

図3: スイッチングユニットのブロック図

3. スイッチの切替え
4. 接続

となり，ブロックが発生しなければ3クロックで経路が確定する．

バケット平坦化モードでは，ヘッダの下位ビットをバケットIDとしてレジスタにラッチし，カウンタ用メモリへのアドレスとする．スイッチの状態遷移は，

1. ヘッダをレジスタにラッチ
2. カウンタの読み出し
3. ポート間でのカウンタ値の比較
4. スイッチの切替え
5. 接続
6. カウンタ更新
7. カウンタの書き込み

となり，ブロックの発生がない場合で経路確定までに4クロックを要する．スイッチの切替えに引き続くカウンタ更新サイクルにおいては，タプル数による平坦化の時はカウンタの値と1との加減算が行なわれ，タプル長による平坦化ではカウンタの値とデータ長との加減算が行なわれる．

5 おわりに

我々が今までに提案し，シミュレーションによりその有効性を示した，バケット平坦化機能を有するデータネットワークの設計について述べた．

今後は，実際に製作した装置についての評価と機能の拡張についての検討を行ないたい．

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報，90(144)，1990.

[3] 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，30(11) pp.1494，1989.

[4] 相場，平野，喜連川，高木: “スーパーデータベースコンピュータ(SDC)におけるバケット平坦化オメガネットワークの動作特性”，電子情報通信学会技報，4(33)，1991.