# 5H-8 スーパーデータベースコンピュータSDCのモジュール間ネットワークにおけるスイッチングユニットの構成

瀬川芳久* 小川泰嗣** 楊維康* 喜連川優* 高木幹雄*
*東京大学生産技術研究所 **リコー中央研究所

## 1 はじめに

SDCは、4台の処理プロセッサと2台のディスクをバス結合した処理モジュールをモジュール間オメガネットワークで結合したアーキテクチャをとる高並列SQLマシンである。[1]

SDCでは処理負荷を効率よく分散させるため、バケット分散並列結合演算法を採用している。[2]この方式では、結合演算を以下のような手順で行なう。まず、被演算リレーションをそれぞれハッシュ分割し、ネットワークを通してすべてのバケットが全モジュール上に平坦に分布するように格納する。次に、バケットの大きさから、処理負荷が均等になるように各モジュールへのバケットの割り当てをスケジュールする。スケジュールに従って、バケットをネットワークを通して転送すると同時に、各モジュールで各バケット内の結合演算を行なう。この際、ネットワーク上で転送の衝突が起こらないようにスケジュールされている。

モジュール間オメガネットワークを使用して、バケットを平坦化するアルゴリズムについては既に発表した。[3]本論文では、このアルゴリズムをハードウェア実装したSDCのモジュール間ネットワークのスイッチングユニットについて、概説する。

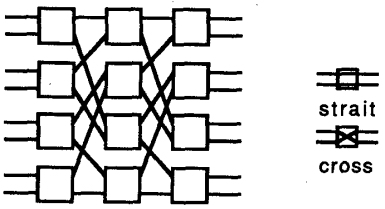


図1: オメガネットワーク

## 2 ネットワークの概要

SDCのデータ転送ネットワークはオメガネットワークの形態をとる。オメガネットワークは、多段結合ネットワークの一種で、[図1]ノード上にある2×2のクロスバスイッチをストレートまたはクロスに接続することにより、入力端から出力端への任意の接続を確立する。

各スイッチは分散制御方式で動作する。分散制御のオメガネットワークでのスイッチングユニットの接続手順は以下の通りである。

スイッチングユニットは入力ポートのデータライン上に示されるヘッダを読み取った後、目的の出力先を解析し、スイッチを接続する。スイッチ接続後はヘッダを次段のスイッチに転送し、次段の接続が開始される。

スイッチ間のデータおよびヘッダの転送はクロックによる同期転送を行なう。

SDCのネットワークの特徴はバケット分散並列結合演算法をハードウェアにより支援する点にある。従って、

- 転送先のIDを指定して転送する(通常転送モード)の他に

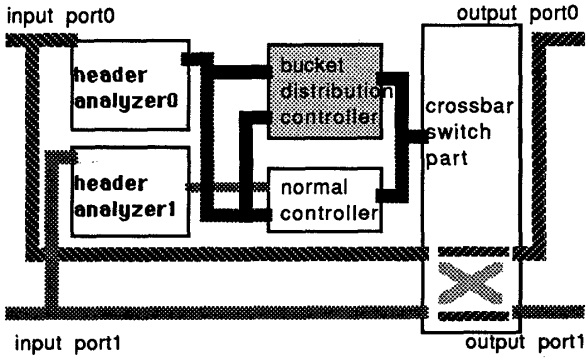


図2: スイッチングユニットのブロック図

| 端子名 | 用途 |
|---|---|
| RVALID | 接続要求と開放要求をユニットに送る。 |
| RACK | 次の接続要求を受け付け可能である事を示す。 |
| DVALID | DATA上のデータが有効である事を示す。 |
| DACK | 出力インタフェースからのデータアクノリッジ信号 |
| DATA< 16 > | 16本のデータライン |

表1: ポート用端子

- ハッシュ分割を行なう際、ハッシュ値に従ってバケット全体が平坦になるように転送先を決定し、転送を行なう機能(バケット平坦化モード)

の機能を有する。

## 3 スイッチングユニットの構成

オメガネットワークにバケット平坦化モードを実装する為、スイッチングユニットを以下の3つの部分に分割した。

- ヘッダ解析部
- 接続制御部
- スイッチ部

バケット平坦化モードは図2に示す通り実現した。接続制御部は2つ用意され、一方は通常転送のためのスイッチ制御を実行し、他方にはバケット平坦化のためのスイッチ制御を実行する。スイッチ部およびヘッダ解析部は通常モード制御部と共用する。接続制御部はヘッダ解析部から発行される制御信号によって接続動作を開始する。制御信号は、ヘッダ解析部がヘッダから転送のモードを解析した後、該当するモードの制御部に、ヘッダとともに送られる。スイッチ部は接続制御部が決定した接続パターンに従ってスイッチの接続を行なう。

スイッチングユニットの外部信号線は大きく分けて、2つの入力ポート用の信号線と2つの出力ポート用の信号線、その他の信号線に分かれる。入出力のポートの有する信号線は[表1]に示す。スイッチングユニットのバケット平坦化モードはハードウェアの多くを通常の転送モードと共有する。その入出力ポート上の信号線も全て通常の転送モードの信号線と共通で特別の信号線は必要としない。

---

[0]Structure of Swiching Unit for Intermojule Network on the Super Database Computer ,SDC

*Y.segawa*, Y.Ogawa**, W.Yang*, M.Kitsuregawa*, M.Takagi**
*The Institute of Industorial Science, University of Tokyo
**Reserch and Development Center,RICHO Co.,Ltd.

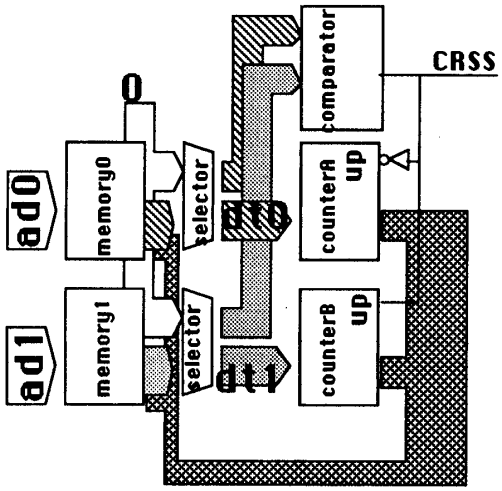


図 3: パケット平坦化接続制御部

## 4 パケット平坦化アルゴリズムの実装

次にパケット平坦化接続制御部での平坦化アルゴリズムの実装を示す。

### 4.1 パケット平坦化アルゴリズム

各スイッチングユニットは、パケット毎に、過去に2つの出力ポートから出力したタプル数の差を逐次記憶しておく。入力ポートに新たなタプルが到着すると、ヘッダからパケット番号を解析し、そのパケットの過去に出力したタプルの数が少ないポートへの接続を行なう。2つの入力ポートに同時にタプルが到着した時には、出力パケット数の差が大きな方を優先する。[3]

### 4.2 接続制御部の構造

パケット平坦化接続制御部の構造を図3に示す。制御部は2個のRAM、U／Dカウンタ、比較器その他の回路からなる。2つのメモリには同じデータが入っており、アドレスX番地のデータはパケット番号Xのタプルの過去の、出力ポート0への出力数と出力ポート1への出力数の差である。メモリは同時に2つのアドレスのデータの読み出しを可能にする為、2つ用意する。実装は主に比較フェーズと計数フェーズに分かれる。

- **比較フェーズ**
  比較フェーズでは過去の履歴からスイッチの接続状態を決定する。新たなヘッダが到着すると、ad 0、1にはそれぞれ、入力ポート0、1に次に入力されるデータのパケット番号が示される。dt 0、1に現れたメモリの内容、即ち出力タプル数の差は比較器で比較されdt 0 >dt 1の時、クロスに、そうでない時ストレートにスイッチングを行なう。一方の入力ポートのみにデータが到着した時、もう一方のdtにはセレクタによって0が出力され、データの到着したポートの優先的な接続を実行する。

- **計数フェーズ**
  計数フェーズはメモリの内容を実際の履歴と整合させる。比較器による比較と同時にdt 0、1の内容はU／Dカウンタにロードする。CRSS信号はカウンタのU／D信号となりカウンタが1進められた後、カウンタAの出力をメモリ0、1にライトし、同様にBの出力もライトする。メモリ内容を正しく保つ為、比較フェーズでデータ未着のため0をロードしたカウンタの内容はライトしない。またad 0、1に現れるパケット番号が同じ時はカウンタの内容は両方ライトしない。

以上の手順でパケット平坦化アルゴリズムが実現できる。

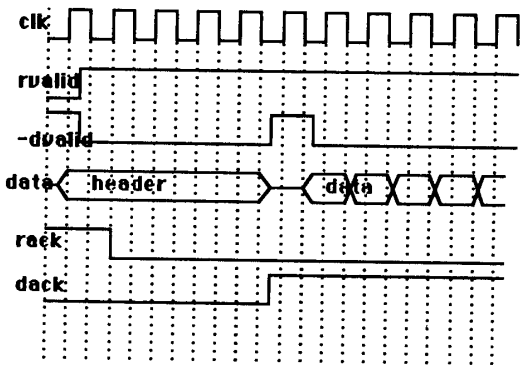


図 4: ユニット間信号のタイミングチャート

## 5 ネットワークを用いてのデータ転送

入出力ポートの信号線を通常の転送モードと共有するため、パケット平坦化モードの接続手順は通常モードと同様に実行する。以下にその手順を述べる。

1. 入力インタフェースは、DATAラインに、転送を行なうモードの情報とパケット番号からなるヘッダをセットし、RVALID、DVALIDをアサートする。（接続開始）
2. ユニットでは、接続処理の実行後、出力ポートのRVALID信号をアサートし、DVALIDとヘッダを出力する。
3. 出力インタフェースまでのスイッチの接続が完了すれば出力インタフェースからDACKが返される。
4. 入力インタフェースでは転送データの同期転送を開始する。データ転送が完了すれば、RACKをネゲートし接続が終了する。

## 6 おわりに

SDCのモジュール間ネットワークのスイッチングユニットの概要についてのべた。現在、ネットワークのスイッチングユニットを実装している。ネットワークインタフェースを実現し、その有効性を確認することが今後の課題である。

## 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] 小川、喜連川「スーパーデータベースコンピュータSDCにおけるパケット分散方式による並列ハッシュ結合演算法」電子情報通信学会データ工学研究会DE89-42、1989
[3] 小川、喜連川「スーパーデータベースコンピュータSDCにおける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11号、1989