5K－4　スーパーデータベースコンピュータ SDC-II のデータネットワークにおけるバケット平坦化機構の実装と評価

田村孝之　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

我々は，大規模データベースに対する非定型問合せ処理の高速化を目的として，バックエンド型高並列 SQL サーバ SDC-II (Super Database Computer version 2) を開発中である．SDC-II は，数台のプロセッサと SCSI ディスク装置および共有メモリとをバスで結合してデータ処理モジュール (DPM) とし，さらに最大 8 台の DPM をネットワークで結合した構成を採る．SDC-II のデータネットワークは，図1に示すように 2 × 2 のクロスバスイッチから構成される間接多段網 (変形オメガネットワーク) である [2].

SDC-II で用いている結合演算のアルゴリズムは並列ハッシュ分割法であるが，バケットサイズに大きなスキューがある時にも各 DPM 処理負荷の偏りを抑えて高い処理性能を維持するために，バケットを一旦すべての DPM 間に均等に分布するように再分配してから動的に DPM への割り付けを行なう “バケット分散” 手法を採用している．我々はこのバケット再分配機構をネットワークのハードウェアによって実現することを提案しており，それによって処理負荷の偏りと同時にネットワークの閉塞によるスループットの低下も解決することができ，関係演算処理に対して有効であることをシミュレーションにより示してきた．

現在，SDC-II は多モジュールで動作中であるが，今回，バケット平坦化機構を実現するデータネットワークのスイッチ素子を FPGA 上に実装したので，その結果を報告する．

2　バケット平坦化機構の実現

データネットワーク上を流れるタプルには，先頭に経路接続情報を示すヘッダとデータ長とが付加される．宛先指定による接続の際は，ヘッダには宛先 DPM 番号が書かれており，i 段目 (出力側から $i = 0, 1, 2$) のスイッチ素子では，その第 i ビットの値から直ちにスイッチ素子の出力ポートを決定することができる．

これに対してバケット平坦化の際には，ヘッダにはそのタプルの属するバケット番号が入る．スイッチ素子の出力ポートを決定するには，それぞれの入力ポート毎にバケット番号 x_j (j は入力ポートの番号) を添字として配列 M の内容を引き，$M(x_0)$ と $M(x_1)$ を比較した結果の符号ビットを用いる．この配列の各要素 $M(x)$ は，バケット x について 2 つの出力ポートのそれぞれを通過したタプル数の差を累積した値であり，絶対値の小さな正負の整数値を取る．また，タプルの出力先が決定した後は $M(x)$ の値を更新する必要がある．

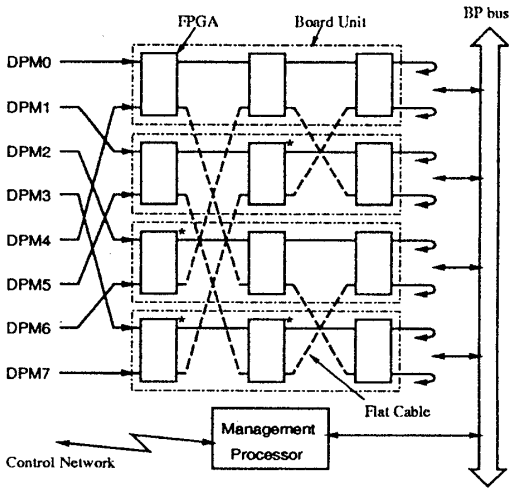

図 1: SDC-II データネットワークの構成

バケット平坦化アルゴリズムはこれまでにいくつかの拡張が施されてきたが [1]，今回実装したスイッチ素子では以下に挙げる動作環境にも対応した．

- 宛先指定モードとバケット平坦化モードの混在
- ポート毎のデータ到着のタイミングは任意
- 任意数の DPM を使用

2 つの入力ポートに宛先指定モードとバケット平坦化モードのデータが同時に到着した場合には，宛先指定モードのタプルを優先して経路を設定する．実際には宛先指定モードの方が経路設定にかかるサイクル数が少ないため，後から到着した宛先モードのタプルが，バケット平坦化モードのタプルを追い越すこともある．

また，一方の入力ポートのみにバケット平坦化モードのデータが到着した時は，他方のポートが接続中でなければ $M(x)$ の符号ビットを用いて出力先を決定し，接続中であれば $M(x)$ とパラメータ T (閾値) との比較結果を用いる．T が小さいと出力先として現在使用中の出力ポートを選択する割合が増え，閉塞が生じやすくなってしまう．大きな T を用いると，一時的に平坦度を犠牲にすることによって空いている出力ポートを有効に利用できるようになる．

使用する DPM 数をネットワークの物理的なサイズと独立にするために，$M(x)$ を更新する時の増分 W_k (k は出力ポートの番号) をパラメータ化し，任意に設定できるように

Implementation and evaluation of bucket flattening mechanism on data network of the Super Database Computer, SDC-II .
T.Tamura, M.Nakamura, M.Kitsuregawa, M.Takagi.
Institute of Industrial Science, University of Tokyo.

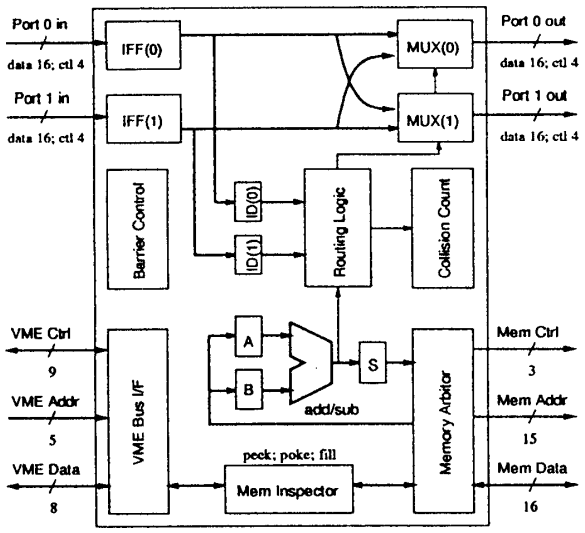

図 2: スイッチ素子のブロック図

している. W_0 と W_1 を出力ポート 0, 1 から到達できる DPM 数の逆比に設定することで, 任意数の DPM 間でのバケット平坦化が可能になる.

3 スイッチ素子の実装と性能

スイッチ素子は, FPGA の一種である Xilinx LCA XC4010-5 (400 CLB, 10000 ゲート相当) 上に実装した. 履歴情報の配列 M はスイッチ素子に外付けの SRAM (32 KWord × 16 bit) に格納する. スイッチ素子の内部は以下のようなブロックに大別される (図 2).

1. 入力レジスタ (IFF), マルチプレクサ (MUX)
   2 × 2 のクロスバスイッチを構成する.
2. ヘッダレジスタ (ID)
3. 経路決定ロジック (Routing Logic)
   バケット平坦化モードでは, ID をアドレスとして $M(x)$ をメモリから読み出し, もう一方のポートが接続中であったら, さらに閾値 T をメモリから読む (アドレスは既定値).
   減算器の出力から出力ポートを決定し, 経路決定後は, メモリから増分 W をフェッチし (アドレスは既定値), $M(x)+W$ を実行した結果をメモリに書き戻す.
4. 演算レジスタ (A,B)
5. 加減算器 (8 ビット 2 の補数表示)
6. VME バス I/F, メモリ監視レジスタ (Mem Inspector)
   VME バスを介して管理プロセッサからスイッチ素子の内部レジスタにアクセスするためのインタフェース. 初期設定やデバッグに使用する. また, スイッチ素子を介してメモリの内容にアクセスできる.

閾値 T や増分 W は, スイッチ素子内部のレジスタに保存することも可能だが, 使用ゲート数を減らし, データパスを簡単にするために, 外付けメモリの中の特定アドレスを割り当てて使っている. ヘッダ中のバケット番号の領域は 13 ビット分なので, メモリの容量 32 KWord の内 8 KWord 分が履歴情報用に使用されることになる.

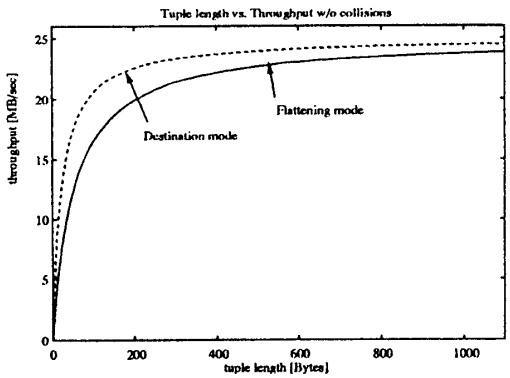

図 3: スイッチ素子の転送性能

また上記の他にも, 閉塞によってタプルがブロックされた時間をカウントするレジスタや, 2 の冪でない数の DPM を使用する時にバリア同期のマスクを設定するレジスタなどが存在する.

今回実装したスイッチ素子においては, 経路設定にかかる最小のオーバーヘッドは宛先指定モードで 3 サイクル, バケット平坦化モードで 8 サイクルであり, クロック周波数 12.5MHz とネットワークの段数 3 を用いると, スループットの最大値 $T_{Max}(l)$ [MB/s] は次式で与えられる.

$$T_{Max}(l) = 25 \times \frac{l}{l+\alpha} \qquad (1)$$

ただし, l は タプル長 [Byte], α は宛先指定モードで 22, バケット平坦化モードで 52 という値になる (図 3).

また, 今回実装したスイッチ素子では, 使用 CLB 数は 338, ゲート数に換算して約 6200 ゲートである. これに対して, 宛先指定モードのみサポートするバージョンでは, 使用 CLB 数 229, ゲート数に換算して約 4100 ゲートである.

4 まとめ

SDC-II のデータネットワークにおけるバケット平坦化機構の実装と性能について述べた. 回路を過度に複雑化させることなく拡張アルゴリズムを採り入れるために, 新たに導入されたパラメータをメモリ上に置くことにした. この結果, オーバーヘッドがやや増加したが, 使用ゲート数が減り, データパスが短くなったのは, 開発を進めていく上で大きな利点だと考える.

今後はさらに詳細な評価を行なう予定である.

参考文献

[1] 田村 ほか. スーパーデータベースコンピュータ (SDC) のバケット平坦化ネットワークにおける縮退動作支援アルゴリズムとその評価. 情処研究会, 1992.

[2] 田村ほか. スーパーデータベースコンピュータ (SDC2) におけるデータネットワーク系の実装. 信学研究会, 1993.