3B－2　スーパーデータベースコンピュータ SDC-II のバケット平坦化ネットワークにおける負荷特性の評価

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

SDC-II (Super Database Computer version 2) は，大規模データベースに対する非定型問い合わせ処理の高速化を目的として開発された高並列関係データベースサーバである [1]. SDC-II は複数の二次記憶装置と複数のプロセッサからなる共有バス方式のクラスタをデータ処理モジュール (DPM) とし，8 台の DPM 間を間接多段網 (変形オメガネットワーク) で結合したアーキテクチャを採用している．

SDC-II では結合演算のアルゴリズムとして並列ハッシュ分割法を用いているが，データベースが大規模化した際には避けられないバケットサイズのスキューや DPM 間に発生するランダムな通信による性能低下を解消するため，各スイッチ素子に “バケット平坦化機構”[2] を持たせ，ネットワークのハードウェアによって DPM 間に均等にバケットを分布させることを可能にしている．この機構は同時に，各スイッチにおける輻輳を避けるように宛先を決定するため，実効スループットを向上させる効果を持っている．

本論文では，SDC-II のバケット平坦化ネットワークのさまざまな入力負荷に対する特性を評価し，本機構の有効性を示す．

2　バケット平坦化機構

ハッシュ結合法は，大きなリレーションをハッシュを用いて主記憶サイズより小さなバケットに分割し，各バケットを順次主記憶上で独立に処理する方法であり，各バケットを異なる DPM に割り当てることで容易に並列化可能である．並列化の際に，各バケットを処理する DPM をハッシュ値から静的に決定する “バケット集中” 型の方法をとると，バケットの大きさにスキューがある場合には DPM 間の負荷が不均等になってしまい，期待される台数効果を達成できない．また，DPM 間の通信がランダムになってしまうため，ネットワーク上で輻輳が頻発し，実効的なスループットが低下してしまう．

そこで，“バケット分散” 型の方法では，バケットの分割数を増やして大きなバケットが生成されるのを防ぎ，分割の完了後にバケットサイズの統計に基づいて各 DPM への割り当てを決定する [3]. 分割を行なっている間は，1) 特定の DPM へのデータの集中を避け，2) 統計情報の取得を容易にし，3) バケットを割り当て DPM に収集する際のネットワーク上での輻輳を避けるために，各バケットは全ての DPM 間に均等に分散して配置する．

このような，バケットの平坦な分布を実現するには，各 DPM がそれぞれのバケットに含まれるタプルを全ての DPM に対して等量ずつ送信すればよいが，ソフトウェア的にバケット毎の宛先 DPM アドレスを循環的に変化させることでも実現は可能である．しかし，この方法では DPM 間の通信パターンがランダムになるため，ネットワークがボトルネックとなることによる性能低下は避けられない．個々のタプルの宛先は便宜的に決定されたものに過ぎないが，輻輳を生じないような通信パターンになるように宛先を決定するには全 DPM 間でのグローバルな通信・同期が必要になり，却ってオーバヘッドが大きくなってしまう．

これに対してネットワークのバケット平坦化機能を利用すると，各タプルがスイッチに到着した時点で宛先を適応的に決定できるため，ブロックの発生を抑えることができる．ただし，宛先をランダムに決定するとバケットの平坦分布が実現できなくなるので，スイッチを通過したタプル数の累積値を各バケット毎に保持しておき，その値に基づいて最もコストの小さな接続状態を決定している．また，現実の環境では各 DPM からのネットワークに対する入力は非同期に与えられるため，ブロックの発生が全くなくなるわけではない．そこで，パラメータとして閾値を設定し，バケット分布の平坦度とブロック率の間のトレードオフを調整可能にしている．

3　閾値による性能の変化

実験では，プロセッサがあらかじめ生成しておいたデータをネットワークに送信した際の，ポート当たりのブロック率および受信したデータのバケット分布の平坦度を測定した．送信データは，Zipf(1.0) 分布にしたがってランダムに分布するバケット ID を持つ．使用 DPM 数 4 台とバケット数 32 は固定し，タプル長を 100 ～ 2000 Byte の範囲で，閾値の値を 0 から 10 の範囲で変化させた．また，DPM 当たりの転送タプル数は全データ長が約 8 MB になるように決定した．

ポートあたりのブロック率は，各スイッチにおけるブロック時間の総和をポート数で割り，全通信時間に占める百分率で表す．また，バケット分布の平坦度は，バケット毎の DPM 間分布の標準偏差の全バケットに関する平均 MSD (Mean Standard Deviation) で表す．MSD の値が大きいほど平坦度が悪いことを示す．

図 1および図 2に，閾値を変化させた時のブロック率と平坦度の変化のそれぞれを，ソフトウェアによるバケット平坦化の結果と共に示す．図 1ではタプル長による影響はあまり

Performance evaluation of bucket flattening network of the SDC-II under various workloads .
T.Tamura, M.Nakamura, M.Kitsuregawa, M.Takagi.
Institute of Industrial Science, University of Tokyo.

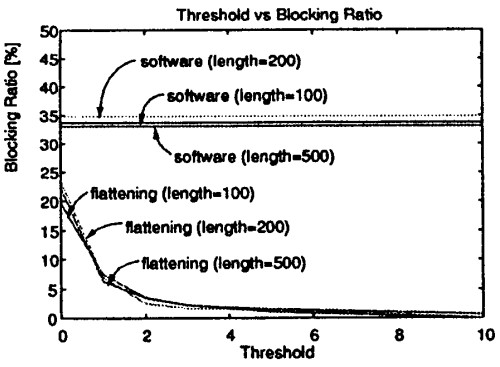

図 1: 閾値に対するブロック率の変化

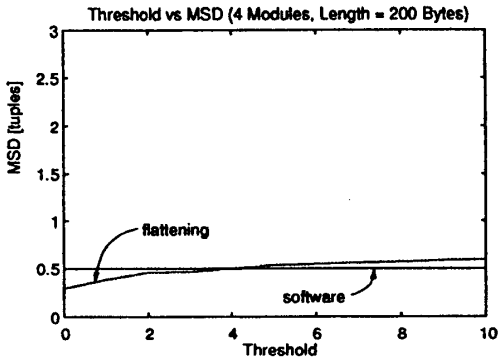

図 2: 閾値に対する平坦度の変化

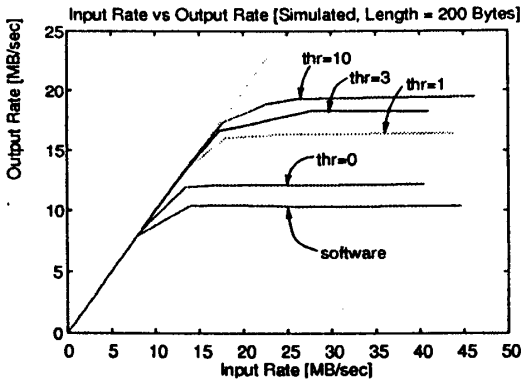

図 3: 入力速度に対する応答性能の変化

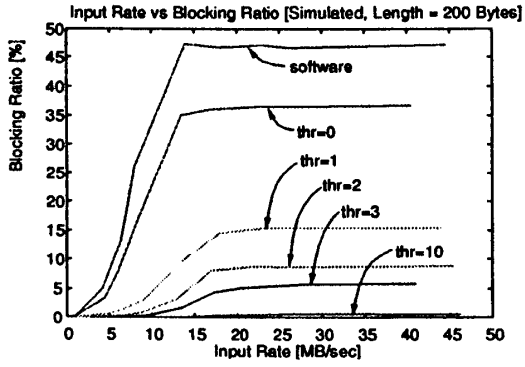

図 4: 入力速度に対するブロック率の変化

見られず,閾値を大きくするにつれてブロック率が急速に小さくなることが分かる. 閾値が 0 の時でもブロック率はソフトウェアによる平坦化よりも小さくなっているが, 1 以上の値に設定することで大幅にブロックを解消できることが分かる. 特に, 3 以上の値にするとブロック率はほとんど 0 に近くなる.

閾値を大きくすることのペナルティは, 結果のバケット分布の平坦度が失われることであるが, 図2によるとバケットサイズのばらつきは 1 タプル以内に収まっており, 平坦度の悪化は問題にならないと言える.

4 入力速度に対する性能の変化

次に, ネットワーク上の転送とプロセッサからのデータ生成がオーバラップして行なわれる場合について, データの生成速度が変化した時の特性をシミュレーションにより調べた. データは 4KB のページ単位で発生し, その発生間隔は指数分布に従うものとした.

図3および図 4は, データの生成速度に対するスループットおよび平坦度の変化のシミュレーション結果である. 図3においては, ソフトウェアによる平坦化では生成速度が 10 MB/s 程度に達するとブロックの発生によってスループットが飽和してしまうのに対し, 閾値を設定したハードウェアによる平坦化では理想値である 25 MB/s に近い 20 MB/s 程度まで入力に追随できることが分かる. 同様に, ブロック率が増加しはじめる生成速度も両者の間で差があることが図 4から読み取れる. さらに, 閾値として 3 以上の値を設定しておけばブロックの増加は問題にならないことも分かる.

5 まとめ

本論文では, SDC-II のバケット平坦化ネットワークの負荷特性を述べた. 並列ハッシュ結合演算におけるバケットサイズのスキューは, ソフトウェア的な方法によっても解決することができるが, ネットワーク上での輻輳が頻発するため, 本来のバンド幅の半分以下しか利用できないことが分かった. ハードウェアの機能を利用することで, データスキューの問題と同時に輻輳の発生を防ぐことが可能であり, より大きな並列度を達成することができる.

参考文献

[1] 中村ほか. スーパーデータベースコンピュータ SDC2 における多重結合演算の実装と評価. 信学研究会, 1994.

[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.

[3] M. Kitsuregawa and Y. Ogawa. Bucket spreading parallel hash: A new, robust, parallel hash join method for data skew in the super database computer(SDC). *16th VLDB*, pp. 210–221, 1990.