3L−8 スーパーデータベースコンピュータ用バケット平坦化オメガネットワークの非同期動作特性

相場 雄一　平野 聡　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は複数のCPUが密結合した処理モジュール（PM）を相互結合網により結合したハイブリッドアーキテクチャを採用した並列データベースマシンである[1]。SDCではバケット分散並列結合演算法を採用する[2]。バケット分散方式では、各バケットを全PMに均一に分配する必要があるが、この方式をバケット平坦化機能と呼ぶ。本発表では、SDCの相互結合網に用いるバケット平坦化オメガネットワーク[3]の非同期動作について説明し、さらにシミュレーションによる性能評価を報告する。

2 SDCにおける平坦化ネットワーク

SDCで採用されるバケット分散方式では、バケットをそれぞれのPMにデータ分布に基づいて動的に割り当てる。そのため、PM間で負荷を一定に調整できデータ分布が不均一な場合でも効率的処理が実現される。この際各バケットは複数のPMにサブバケットとして分散格納される。分散格納されたバケットはその結合演算時に割り当てられたPMに収集されるが、このサブバケットの大きさが不均一であると収集に要する時間が大きくなってしまう。逆に均一であると効率良く処理が行なわれる。すなわち、各バケットについてサブバケットの大きさが等しいことが必要となる。このような分布をバケット平坦分布と呼ぶ。

SDCではオメガネットワークにバケット平坦化機能を付加することを我々はしている。オメガネットワークは多数の2入力2出力（以下2×2）スイッチング装置SUから構成される。各段はシャッフル交換によって結合される。ネットワークの大きさが$N \times N$の場合、段数は$log_2 N$、各段のSUの数は$N/2$である。各SUは2×2のクロスバスイッチで、StraightとCrossedの2つの状態をとる。

3 スイッチング装置における非同期動作制御方式

ネットワークの各SUはバケットの分布に関する局所的な情報から自律的に状態を決定する。このSUの状態決定法では、次のようなカウンタ$D(X)$を用意する。XはバケットID、$D_{left}(X)$、$D_{right}(X)$はそれぞれ左右の出力ポートから出力されたバケットXに属するタプルの数である。

$$D(X) = D_{left}(X) - D_{right}(X)$$

非同期動作においてはSUへタプルが到着する状況には次の3つの場合が考えられる。

- 2つの入力ポートに同時にタプルが到着する場合
- 一方の入力ポートにのみタプルが到着する場合
- 一方の入力ポートの入力によってSUの状態が決定されている時に他方の入力ポートにタプルが到着する場合

以下、X_{left}とX_{right}を左右の入力ポートに到着したタプルの属するバケットとしてそれぞれの場合についてSUの状態決定法を説明する。

2つの入力ポートに同時にタプルが到着する場合

- $D(X_{left}) \geq D(X_{right})$ ⟶ Crossed
- $D(X_{left}) < D(X_{right})$ ⟶ Straight

一方の入力ポートにのみタプルが到着する場合

- タプルが左側のポートに到着した場合:
  $D(X_{left}) \geq 0$ →Crossed, $D(X_{left}) < 0$ →Straight
- タプルが右側のポートに到着した場合：
  $D(X_{right}) \geq 0$ →Crossed, $D(X_{right}) < 0$ →Straight

一方のポートに入力されている時他方にタプルが到着した場合

この場合はすでにSUの状態は決定されているわけである。新たに到着したタプルについて$D(X)$の絶対値が小さくなるように出力すればバケットの偏りが少なくなるが、すでにSUの状態がこれを大きくするように決定されていることも考えられる。この時そのタプルをブロックするかどうかが問題となる。そこで、しきい値（以下Thrと書く）を導入し、$D(X)$とThrの大小によってブロックするか否かを決定する。新たなタプルが到着した時点でのSUの状態がCrossedかStraightによって2つの場合に分けられる。

1. SUの状態がCrossedの場合：
   - タプルが左側のポートに到着した場合：
     $D(X_{left}) \geq -Thr$ ⟶ ブロックしない
     $D(X_{left}) < -Thr$ ⟶ ブロック
   - タプルが右側のポートに到着した場合：
     $D(X_{right}) \leq Thr$ ⟶ ブロックしない
     $D(X_{right}) > Thr$ ⟶ ブロック
2. SUの状態がStraightの場合：
   - タプルが左側のポートに到着した場合：
     $D(X_{left}) \leq Thr$ ⟶ ブロックしない
     $D(X_{left}) > Thr$ ⟶ ブロック
   - タプルが右側のポートに到着した場合：
     $D(X_{right}) \geq -Thr$ ⟶ ブロックしない
     $D(X_{right}) < -Thr$ ⟶ ブロック

4 シミュレーションによる性能評価

4.1 シミュレーションモデル

シミュレーションで用いられたパラメータは、処理モジュール数：N、バケット数B、処理モジュールあたりのタプル数T、タプル発生確率L、しきい値Thr、タプル長TLである。タプル発生確率とは、タプルが入力されていないある単位時間に新たなタプルがネットワークに入力される確率である。タプル長は1単位時間に転送される長さを1としている。

Asynchronous Operation in an Omega Network with the Flat Bucket Distribution Mechanism for the Super Database Computer
Y.Aiba, S.Hirano, M.Kitsuregawa, M.Takagi
Institute of Industrial Science, University of Tokyo

パケットのバッファ上の分布の平坦度を次に定義する平均標準偏差$\overline{\sigma}$で評価する。D_{ij}をパケットiのPMjに格納されているサブパケットの大きさとする。$\overline{\sigma}=0$がパケット平坦分布に相当、大きいほど分布は不均一となる。

$$\overline{\sigma}=E\left\{\frac{1}{B}\sum_{i=1}^{B}\sqrt{\frac{1}{N}\sum_{j=1}^{N}{D_{ij}}^2-(\frac{1}{N}\sum_{j=1}^{N}D_{ij})^2}\right\}$$

4.2 シミュレーション結果

タプル長TLを変化させた場合： $T=1k(=1024)$、$N=8$、$B=128$、$L=0.1$と固定し、タプル長$TL=10,30,100$とした場合各々Thrを変化させた場合の結果を図1、2に示す。図1、2は横軸にThrをとり、それぞれ平均標準偏差、全処理時間との関係を示している。Thrが0に近いほど平坦化能力が高いが、ブロックが起きやすいため処理時間が長くなっていることがわかる。処理時間については、$L=0.1$の場合あるタプルと次のタプルの間の空き時間の期待値は9単位時間となり、タプル長100の場合を考えると1024タプルを転送するのに$(100+9)\times1024\simeq112\times10^3$単位時間かかる。このことを考えると$Thr=15$のところでほとんどブロックがないことがわかる。$Thr=0$の近くでもこれと比較して1.3倍程度の処理時間で済む。また、平均標準偏差においても$Thr=0$では0.5以下と非常に小さな値となる。Thrを大きくした場合2.0以上になることもあるが、タプル数を増やしてもこの値はあまり変化せず相対的に減ることが確認され、有効であることがわかった。

$TL\times L=1$としてTL、Lを変化させた場合： ここでいう発生確率からはタプルとタプルの間の空き時間の期待値が算出できる。この空き時間とタプルの長さの比を一定に保つことによってSUの負荷を一定にできる。$TL\times L=$一定とすることによってほぼこれを満たすことができる。$TL=10,30,50$とした場合の結果を図3に示す。予想された通りほぼ同じ値を示している。さらに処理時間については縦軸に（単位時間）／TLをとることによって、図4のようにほぼ同じ値を示すことがわかる。

5 おわりに

本発表では、SDCのPM間結合網に採用するパケット平坦化オメガネットワークの非同期動作について報告した。シミュレーションによる性能評価の結果その有効性が確認できた。

参考文献

[1] 楊維康他、“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”、第39回情処全国大会

[2] 小川泰嗣、喜連川優“スーパーデータベースコンピュータにおけるパケット分散並列結合演算方とその性能予測”、第39回情処全国大会

[3] 喜連川優、小川泰嗣“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、Vol. 30.

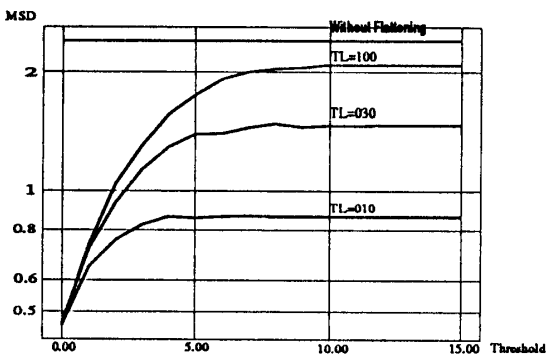

図 1: タプル長を変化させた場合の平均標準偏差

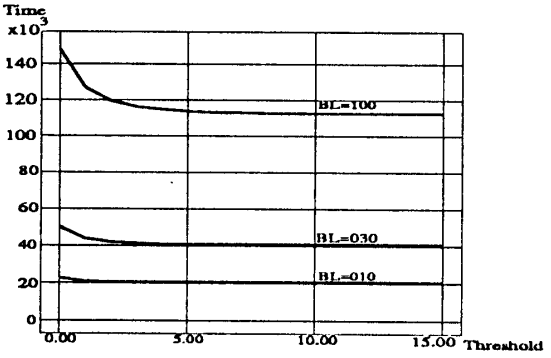

図 2: タプル長を変化させた場合の全処理時間

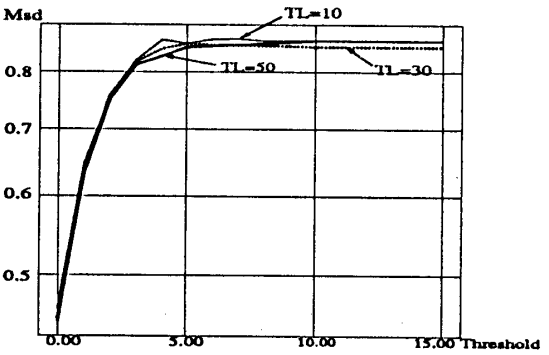

図 3: $TL\times L=$一定とした場合の平均標準偏差

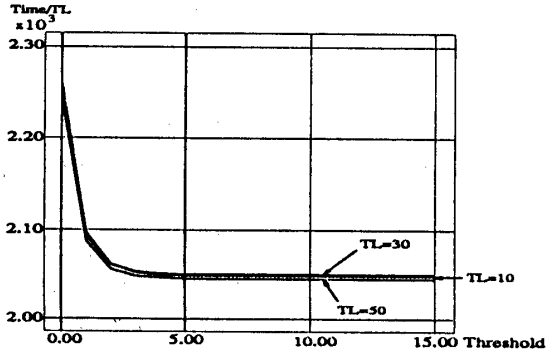

図 4: $TL\times L=$一定とした場合の（全処理時間）／TL