バケット平坦化機能を有するオメガネットワーク†

喜連川 優†† 小川泰嗣†††

関係データベースシステムにおいて処理負荷の大きい結合演算等に対しハッシュ分割法は有効である．ハッシュ分割法は並列処理向きであるため，結合演算等を高速実行する並列データベースマシンではハッシュ分割法を用いるものが多い．各バケットを複数の処理モジュールにステージングするバケット分散方式の高速化のためには，複数の処理モジュール上にバケットを均一に分配する必要がある．この機能をバケット平坦化機能と呼ぶ．本論文では処理モジュール間結合にオメガネットワークを用いる並列データベースマシンを想定し，そのためのバケット平坦化機能を持つオメガネットワークを提案する．バケット平坦化の制御方式に分散制御を採用することで，バケットの分布を集中管理する必要がなく，ネットワークの大きさに依存せず制御時間を一定にできる．シミュレーションによる評価から本オメガネットワークがバケット平坦化に有効であることが確認された．2×2スイッチング装置を用いた場合平均標準偏差を0.7タプル，ゆらぎを2.7タプル以下，4×4スイッチング装置を用いた場合平均標準偏差を0.5タプル，ゆらぎを2タプル以下にできた．実装には，バケット数をBとしたとき，B個のカウンタと比較器を付加するだけで良い2×2スイッチング装置を用いたオメガネットワークが望ましい．

1. はじめに

関係データベースシステムは簡潔なユーザインタフェース，高度なデータ独立性，明確な論理的基盤等から標準的データベースシステムとなりつつあり，それに伴い処理の高速化に対する要求が高まっている．結合演算は重複除去演算や Group-by 句を含む集計演算等と同じく処理負荷が大きいため，ソフトウェアおよびハードウェアの両面から研究が進んでいる[1]．

結合演算のアルゴリズムとして様々なものが提案されているが，GRACE ハッシュ，ハイブリッドハッシュ等のハッシュ分割法は対象リレーションを結合属性値に基づいて複数のバケットと呼ばれるクラスタに分割する方式で，大規模リレーションに対しても有効である[2)~4)]．また，バケットの処理を1つのプロセッサに割り当てることで簡単に並列処理が実現できるため，ハッシュ分割法を複数のプロセッサで並列実行するデータベースマシンが提案されている[5)~7)]．これらマシンはプロセッサ，ディスクおよび大容量のステージングバッファ等から構成された処理モジュールを複数持つ．結合演算は，ディスクから読み出したリレーションをバケットに分割し，複数の処理モジュールが独立にバケットの結合演算を行うことで並列処理される．その際，各バケットを複数の処理モジュールに分散格納するバケット分散方式では，バケットの処理順序の調整によりバケットの大きさが不均一な場合でもバケットの大きさのゆらぎを吸収し効率的処理を実現できる[2),5)]．

バケット分散方式では，複数の処理モジュールに分散格納されているバケットをその処理に割り当てられた処理モジュールに収集する必要があるため，ボトルネックとなる可能性がある．このバケット収集を巡回的に行うことで効率化できるが，バケットが処理モジュールに不均一に格納された場合，パイプラインの擾乱により処理速度が低下する[8)]．したがって，バケット分散方式の効率的実行には1つのバケットを複数の処理モジュール上に均一に分配する必要がある．バケットを複数の処理モジュールに均一に分配する機能をバケット平坦化機能と呼び，その実現法をわれわれはすでに提案している[9),10)]．しかし，リングバスによる方式ではバス容量の点からプロセッサ数が制限され，間接nキューブネットワークによる方式では集中制御方式を採用したため，転送速度を低下させないためにネットワークの並置やスイッチング装置へのバッファの付加等が必要となった．

本論文ではバケット平坦化機能を持つオメガネットワークを提案する．本ネットワークでは，バケット平坦化の制御にネットワークの各スイッチング装置がバケットの分布に関する局所的な情報に基づいて状態を自律的に決定する分散制御方式を採用する．通常の2入力2出力のスイッチング装置だけでなく，オメガネットワークの構成要素として4入力4出力のスイッ

---

† An Omega Network with Bucket Flat Distribution Mechanism for a Parallel Database Machine by MASARU KITSUREGAWA (Institute of Industrial Science, University of Tokyo) and YASUSHI OGAWA (Research and Development Center, Ricoh Co., Ltd.).

†† 東京大学生産技術研究所
††† (株)リコー中央研究所

チング装置を用いた場合の制御方式も提案する．さらに，本オメガネットワークのバケット平坦化の能力をシミュレーションにより評価する．

本論文の構成は以下のとおりである．まず，2章でバケット平坦化機能の説明を行う．ここでは，結合演算の効率的方式であるハッシュ分割法を解説し，その効率的並列実行法を検討する．3章でバケット分配に採用したオメガネットワークを紹介する．本論文で提案するバケット平坦化機能を有するオメガネットワークは4章で説明される．分散制御方式による2入力2出力および4入力4出力のスイッチング装置の状態決定方式を詳細に説明する．5章で本論文で提案したオメガネットワークのシミュレーションに基づく評価と解析を行い，最後の6章で本論文をまとめる．

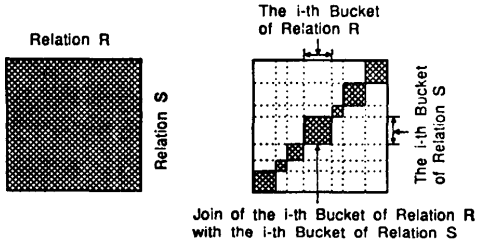
2. ハッシュ分割法による結合演算とその並列実行

この章では，結合演算の効率的方式であるハッシュ分割法を解説し，その並列実行法を検討する．さらに，バケット分散式の処理高速化のために必要となるバケット平坦化機能を説明する．

2.1 ハッシュ分割法による結合演算

結合演算法として従来から用いられているネストループ法は2つのリレーションに属するタプルの全組合せについて結合属性値を比較するため，処理時間は2つのリレーションの大きさの積に比例する．一方，ハッシュ分割法は結合属性値にハッシュ操作を施しリレーションをバケットと呼ばれるクラスタに分割する．同じハッシュ値を持つタプルが1つのバケットを構成するので，異なるバケットに属するタプルは結合しない．したがって，結合属性値を比較すべきタプルの組合せを斜線部で表すと**図1**のように組合せが減少し，処理時間が大幅に短縮される[2]．

ハッシュ分割法は分割フェーズと演算フェーズの2つのフェーズに分けられる．分割フェーズでは，対象リレーションの各タプルの結合属性値にハッシュ操作を施し，リレーションを複数のバケットに分割する．このとき，2つのリレーションに同じハッシュ関数を用いる．分割フェーズに続いて行われる演算フェーズでは，各バケットの結合演算が実行される．2つの対象リレーションのバケットのなかから対応するもの（同じハッシュ値を持つもの）が順に選択され，その結合演算が行われる．ハッシュ分割法は結合演算のほか，重複除去演算やGroup-by句を含む集計演算等負荷の重い関係代数演算にも有効である．



(a) Ordinary processing. (b) Hash-partitioned processing.

図1 ハッシュ分割法による処理負荷の低減
Fig. 1 Load reduction effect by hashing.

2.2 ハッシュ分割法のバケット分散方式による並列実行

ハッシュ分割法では各バケットの処理を1つのプロセッサに割り当てると，異なるバケットに属するタプルは結合しないので，各プロセッサは他のプロセッサと通信なしに結合演算を効率的に並列実行できる[2),5),6)]．いま想定しているデータベースマシンのアーキテクチャは**図2**のように複数の処理モジュールから構成される．処理モジュールにはプロセッサ，ディスクおよび大容量のステージングバッファ（以下略してバッファと呼ぶ）があり，各リレーションは水平分割され複数の処理モジュールのディスクに格納されているものとする．

ハッシュ分割法は以下のように並列実行される．分割フェーズでは，ハッシュ操作を各処理モジュールが独立に行い，ディスクから読み出したリレーションをバケットに分割する．バケットに分割されたリレーションは，各バケットが全処理モジュールに分散するバケット分散方式でいったん格納される．すなわち，タプルは適切な処理モジュールに転送され，ディスクに格納される．演算フェーズでは，各バケットの処理が1つの処理モジュールに割り当てられ，処理モジュールは各々並列に結合演算を行う．演算フェーズでは，各処理モジュールは割り当てられたバケットを複数の

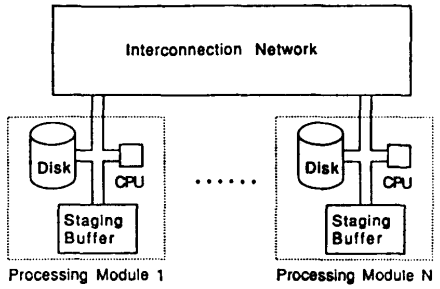


図2 想定するデータベースマシンのアーキテクチャ
Fig. 2 Architecture of parallel database machine.

処理モジュール内のディスクから自分のバッファに順次読み込み，結合演算を行う．バケットの大きさが異なると処理速度は若干低下するが，バケットを小さい順に処理することで影響を小さくできる[2),5)]．

2.3 バケット分散方式におけるバケット平坦化機能

バケット分散方式では各バケットが複数の処理モジュールに分散格納されているため，演算フェーズで各処理モジュールは割り当てられたバケットを収集する必要がある．このとき，同じディスクに格納されている異なるサブバケットに対し複数の処理モジュールから読み出しが発生し，アクセス競合となる可能性がある．ここで，サブバケットとは各処理モジュールに格納されているバケットの一部分を指す．アクセス競合はサブバケットを巡回的に読み出すことで回避できる．**図3**は処理モジュールが4個の場合を示しており，各ステップでは太線で囲まれたサブバケットが読み出される．

バケットが処理モジュール上に不均一に分布している場合，すなわち，あるバケットのサブバケットの大きさが処理モジュールで異なる場合，パイプラインの擾乱が起こり効率が低下する．これに対し，**図4**のように各バケットが処理モジュール上に均一に分布すれば，パイプラインの乱れは著しく減少する．サブバケットの大きさがお互いに等しいようなバケットの処理モジュール上の分布をバケット平坦分布，その実現のための機能をバケット平坦化機能と呼ぶ．

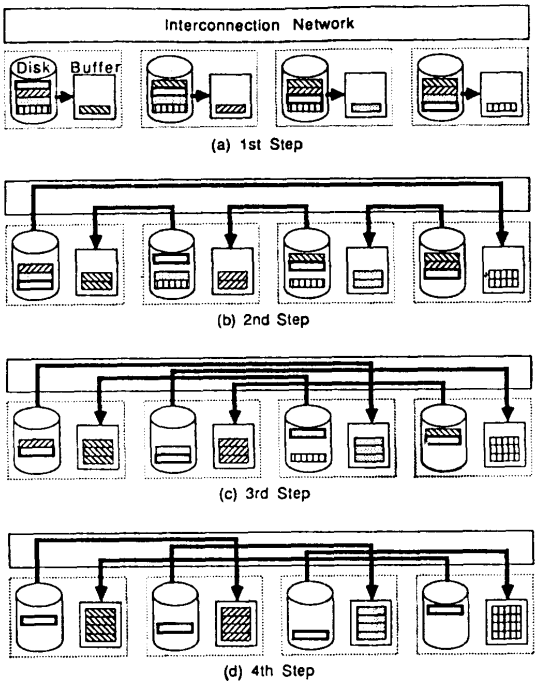


図3 巡回的に行われるバケット収集処理の様子
Fig. 3 Example of bucket collection cycle (N=4).

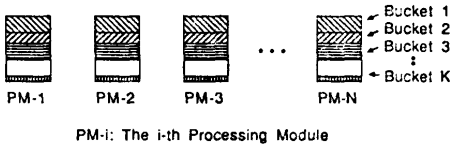


図4 バケット平坦分布
Fig. 4 Flat bucket distribution.

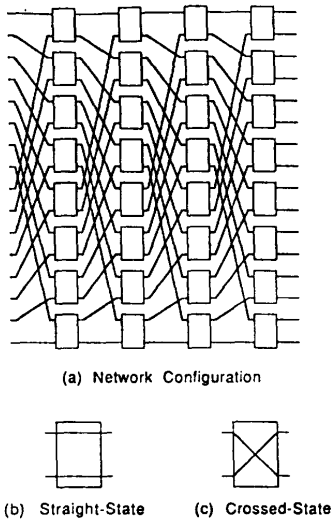


図5 2×2 スイッチング装置で構成されるオメガネットワーク
Fig. 5 Omega network configured by 2×2 switching units (N=16).

3. オメガネットワーク

以下の章で，処理モジュール間結合に多段ネットワークの1つであるオメガネットワーク[11)]を用いた場合のバケット平坦化機能の実現法を考察する．オメガネットワークを採用したのはバス型のネットワークと比べ閉塞が起こり難く，また，クロスバスイッチ等と比較してハードウェアのコストが低いからである[12)]．

オメガネットワークは多数の2入力2出力（以下では 2×2 と書く）スイッチング装置から構成される．**図5**(a)にネットワークの大きさ（入出力ポート数）が16 の場合のオメガネットワークを示す．ある段の出力ポートとつぎの段の入力ポートは Shuffle-exchange で結合される．ネットワークの大きさが N の場合，ネットワークの段数は $\log_2 N$，各段のスイッチング装置数は $N/2$ である．各スイッチング装置は 2×2 のクロスバスイッチで，図5(b)(c)に示す Straight

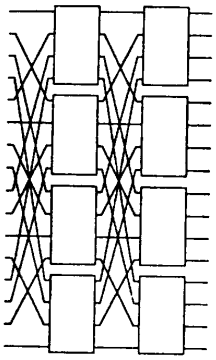


図 6 4×4 スイッチング装置で構成されるオメガネットワーク
Fig. 6 Omega network configured by 4×4 switching units (N=16).

と Crossed の2つの状態がある．また，4×4 のクロスバスイッチをスイッチング装置としてオメガネットワークを構成することもできる（**図 6**）．ある段の出力ポートとつぎの段の入力ポートは2段直列結合した Shuffle-exchange で結合される．ネットワークの段数は $\log_4 N$，各段のスイッチング装置数は $N/4$ である．各スイッチング装置は 4×4 のクロスバスイッチなので，入出力ポートを1対1対応させる状態数は $4!=24$ である．

オメガネットワークは閉塞型ネットワークであり，任意の入出力ポートを結合することはできるが，複数の結合を同時には実現できない場合がある．2×2 スイッチング装置でオメガネットワークを構成した場合，実現可能な結合数は $2^{N/2\times\log_2 N}$ である．一方 4×4 スイッチング装置の場合，結合数は $24^{N/4\times\log_4 N}$ であり閉塞が減る．

4. バケット平坦化機能を有するオメガネットワーク

この章ではバケット平坦化の制御方式を考察し，分散制御方式による実現法を詳細に説明する．

4.1 分散制御方式と集中制御方式

バケットを平坦化するには，各タプルについてそれが属するバケットの処理モジュールのバッファ上の分布を調べ，そのバケットのサブバケットのなかから大きさ（サブバケット内のタプル数）が最小の処理モジュールに転送すれば良い．その際オメガネットワークは閉塞型ネットワークの1種であるため，閉塞が起こらないようにタプルの行き先を決定する必要がある．

バケット平坦化の制御方式には集中制御方式と分散制御方式がある．集中制御方式では中央管理装置がタプルの行き先を決定する．各タプルには行き先タグが付けられ，ネットワークのスイッチング装置はそのタグに従って状態を決定する．これに対し，分散制御方式では処理モジュールからネットワークへの出力時にタプルの行き先を決定するのではなく，ネットワークのスイッチング装置がバケットの分布に関する局所的な情報のみからバケット平坦化するよう自律的に状態を決定する．集中制御方式では，ネットワークで閉塞を避けるためタプルの行き先を逐次的に決定しなければならず，バケットの分布からタプル数が最小である処理モジュールを探索する必要がある．したがって，ネットワークの大きさ N に対し行き先決定に $O(N^2)$ の時間がかかり，ネットワークが大きい場合不適切である．これに対し，分散制御方式はタプルの行き先の決定が逐次的ではなく，ネットワークが大きい場合でも適用できる．したがって，集中制御方式より分散制御方式の方が適切と考えられる．さらに，本論文で提案する分散制御方式では処理モジュール上のバケットの分布を知らなくて良いため，スイッチング装置間あるいはスイッチング装置と処理モジュール間の通信の必要がなく実装が簡単になる．

4.2 分散制御による2×2スイッチング装置の状態決定法

2×2 スイッチング装置の場合，上下2つのポートから出力されたタプル数をバケットごとに記録するため，合計 $2B$ 個のカウンタが必要である．ここで，B はバケット数である．分割フェーズのはじめにカウンタはすべて0に初期化され，以後各ポートからタプルを出力するごとにそのタプルが属するバケットのカウンタがインクリメントされる．$C_i(X)$ が出力ポート i ($i=upper, lower$) のバケット X に対するカウンタ値，X_j が入力ポート j($j=upper, lower$) に到着したタプルのバケット，$f(i)$ が入力ポート i と結合される出力ポートを表すとき

$$S(f)=C_{f(upper)}(X_{upper})+C_{f(lower)}(X_{lower})$$

は状態 f が与えられたときのそのスイッチング装置から出力されたタプルの累積数を表すもので，この S を最小にする f をスイッチング装置の状態とすればバケットは平坦化される．2×2 スイッチング装置の場合，取り得る状態の Crossed と Straight の S を求め，値が小さい方の状態とすれば良い．状態が Crossed ならば $f(upper)=lower$, $f(lower)=upper$ であり，Straight ならば $f(upper)=upper$, $f(lower)=lower$ であるから，S の差 Dif(=S(Straight)−S(Crossed))

は

$$Dif=(C_{upper}(X_{upper})+C_{lower}(X_{lower}))-(C_{lower}(X_{upper})+C_{upper}(X_{lower}))$$

と求められる．Dif はスイッチング装置に入力された2つのタプルが属するバケットの出力の上下のポートによる偏りを表すので，スイッチの状態はこの値が正ならば Crossed，負ならば Straight とすれば良い．

以下の方法で，状態決定に必要なカウンタ数をB個に減らすことができる．上下のカウンタの値の差

$$D(X)=C_{upper}(X)-C_{lower}(X)$$

から Dif をつぎのように計算できるからである．

$$Dif=D(X_{upper})-D(X_{lower})$$

スイッチング装置の状態は先ほどと同様，この値が正ならば Crossed，負ならば Straight とする．分割フェーズのはじめにカウンタはすべて0に初期化され，以後各ポートからタプルを出力するごとに上のポートから出力されたタプルの属するバケット用カウンタはインクリメント，下のポートから出力されたタプルの属するバケット用カウンタはデクリメントされる．

図7にスイッチング装置の状態が決定される例を示す．ここでは，上下ポートに各々バケットmとバケットnに属するタプルが入力される．現在のカウンタ値から $Dif=5-(-2)=7>0$ なので，このスイッチング装置の状態は Crossed に設定される．その結果，バケットmのカウンタはデクリメントされ，バケットnのカウンタはインクリメントされる．

4.3 分散制御による4×4スイッチング装置の状態決定法

4×4 スイッチング装置の場合，4つの出力ポートごとに出力したタプル数をバケットごとに記録するため，$4B$ 個のカウンタが必要である．$C_i(X)$ が出力ポート $i(i=1,2,3,4)$ のバケットXに対するカウンタ値，X_j が入力ポート $j(j=1,2,3,4)$ に到着したタプルのバケット，$f(i)$ が入力ポート i と結合される出力ポートを表すとき

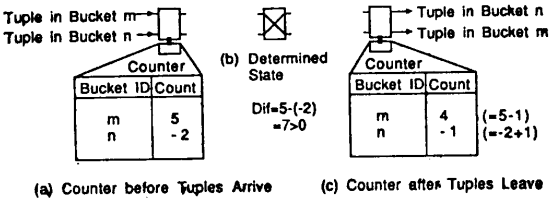


図7 2×2 スイッチング装置における状態決定の様子
Fig. 7 State determination in 2×2 switching unit.

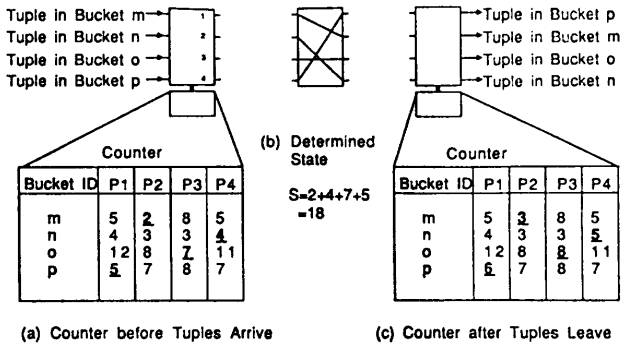


図8 4×4 スイッチング装置における状態決定の様子
Fig. 8 State determination in 4×4 switching unit.

$$S(f)=\sum_{i=1}^{4} C_{f(i)}(X_i)$$

を最小にする f をスイッチング装置が取り得る24の状態のなかから探し，そのスイッチの状態とする．さらに，各ポートから出力されたタプルが属するバケットのカウンタをインクリメントする．

図8にスイッチング装置の状態が決定される例を示す．ここでは，各ポートにバケット m, n, o, p に属するタプルが入力される．現在のカウンタ値から $f(1)=2$, $f(2)=4$, $f(3)=3$, $f(4)=1$ のとき $S=2+4+7+5=18$ で最小となる．タプルを送り出した後のカウンタは，各ポートから出力されたタプルの属するバケットの値がインクリメントされ，図8(c)のようになる．

5. シミュレーションによる評価

バケット平坦化のシミュレーションを行い本オメガネットワークを評価する．

5.1 シミュレーションモデル

今回のシミュレーションでは処理モジュール数 N をネットワークの大きさと呼ぶ．その他，バケットの種類 B と処理モジュールのディスクに格納されているタプル数 T をパラメータとし，シミュレーションを行う．バケットの処理モジュール上の分布の平坦度はつぎのように求める．まず，あるバケットの処理モジュールの違いによるサブバケットの大きさの標準偏差を求め，つぎに，その値のバケットによる平均をとる．この値 $\bar{\sigma}$ を平均標準偏差と呼び，バケット平坦度の評価に用いる．D_{ij} をバケット i の処理モジュール j に格納されているサブバケットの大きさとすると，平均標準偏差はつぎのように計算される．

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

ここで，E{ } は期待値を表す．$\sigma=0$ がバケット平坦分布に相当し，大きいほど分布が不均一である．また，あるバケットについてサブバケットの大きさが最大のものと最小のものとの差をバケットについて平均をとった値を処理モジュール間のサブバケットの大きさのゆらぎと定義し，この値も平坦度の評価に用いた．

タプルの属するバケットはつぎの2つのいずれかの分布に従うものとする．第1の分布では，タプルがどの処理モジュールに格納されているかによらずタプルの属するバケットを全バケットからランダムに選択される．処理モジュール i ($i=1,\cdots,N$) に格納されていたタプルがバケット b ($b=1,\cdots,B$) である確率 $p_i(b)$ は

$$p_i(b)=\frac{1}{B} \quad \text{for all } b$$

である．これをバケットの均一分布と呼ぶ．第2の分布では，処理モジュールによって格納されているタプルの属するバケットの範囲が限定され，確率 $p_i(b)$ はつぎのように与えられる．

$$p_i(b)=\begin{cases}\dfrac{N}{B} & : (i-1)\times\dfrac{B}{N}<b\leq i\times\dfrac{B}{N}\\ 0 & : \text{otherwise}\end{cases}$$

この分布をたんざく分布と呼ぶ．

5.2 シミュレーション結果

まず，3つのパラメータの影響を調べる．各シミュレーションで本オメガネットワークのバケット平坦化機能の効果を見るため，処理モジュールのディスク上での分布に対する平均標準偏差を求めた（平均標準偏差は付録に示したように解析的に計算できる）．

ネットワークの大きさを変化させた場合：処理モジュール当たりのタプル数とバケット数を一定にし，ネットワークの大きさを変化させる．**図9**に $T=1$k ($=1,024$)，$B=128$，均一分布の場合の平均標準偏差を示す．横軸はネットワークの大きさ，縦軸は平均標準偏差をそれぞれ対数表示したものである．ネットワークが大きくなるほどディスク上でのバケットの平坦度も大きくなり，$N=64$ で平均標準偏差は 3.0 程度である．この結果は解析的に求めた値と良く一致している．2×2スイッチング装置で構成した本オメガネットワークにより，$N=64$ の場合でも平均標準偏差は 0.7 でバケット平坦化機能の有効性が確認された．4×4 スイッチング装置では，平均標準偏差は 0.5 であり効果はさらに大きい．特に図に示していないが，

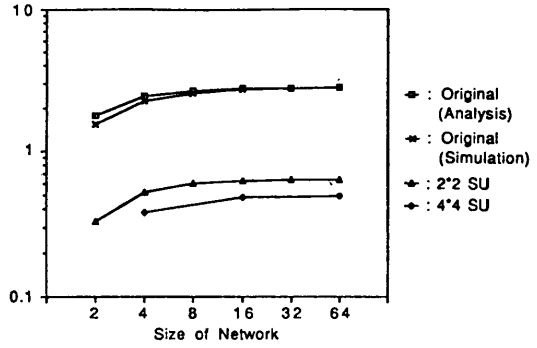


図 9 ネットワークの大きさと平均標準偏差
($\beta=128$, $T=1$k, 均一分布)

Fig. 9 Size of network vs. mean standard deviation ($B=128$, $T=1$k, uniform distribution).

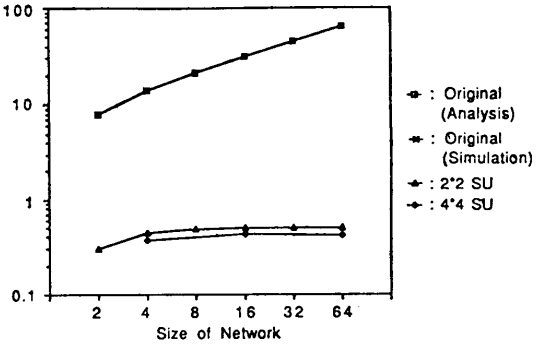


図 10 ネットワークの大きさと平均標準偏差
($B=128$, $T=1$k, たんざく分布)

Fig. 10 Size of network vs. mean standard deviation ($B=128$, $T=1$k, rectangular distribution).

ゆらぎも平均標準偏差と同じ傾向が見られ，ネットワークが大きくなるに従いゆらぎも大きくなる．$N=64$ の場合，ゆらぎは 2.7 (2×2 スイッチング装置)，1.9 (4×4 スイッチング装置) であった．たんざく分布の場合は**図 10** のとおりであり，均一出力分布と同じ傾向が見られるが，平坦化に対する効果は大きい．$N=64$ の場合のゆらぎは 2.0 (2×2 スイッチング装置)，1.6 (4×4 スイッチング装置) であった．

処理モジュール当たりのタプル数を変化させた場合：ネットワークの大きさとバケット数を一定にし，タプル数を変化させる．**図 11** に $N=16$，$B=128$，均一分布の場合の結果を示す．横軸はタプル数，縦軸は平均標準偏差をそれぞれ対数表示したものである．先ほどと同様バケット平坦化機能が有効であり，特に 4×4 スイッチング装置の効果が大きい．ディスク上でのバケット分布では，タプル数が大きくなると（解析による結果と一致して）平均標準偏差も大きくなるのに対し，本ネットワークではタプル数に関係なくほ

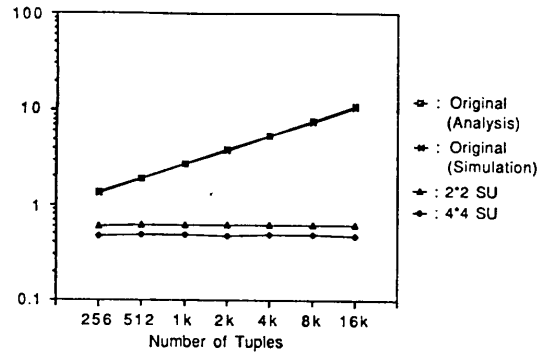


図 11 タプル数と平均標準偏差
(N=16, B=128, 均一分布)

Fig. 11 Number of tuples vs. mean standard deviation (N=16, B=128, uniform distribution).

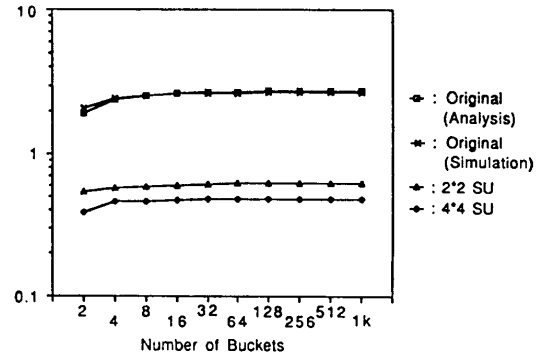


図 13 バケット数と平均標準偏差
(N=16, T=8×B, 均一分布)

Fig. 13 Number of tuples vs. mean standard deviation (N=16, T=8B, uniform distribution).

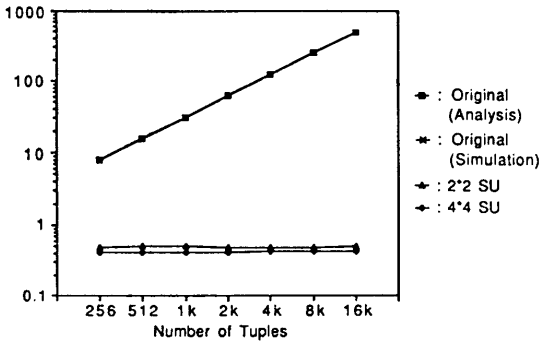


図 12 タプル数と平均標準偏差
(N=16, B=128, たんざく分布)

Fig. 12 Number of tuples vs. mean standard deviation (N=16, B=128, rectangular distribution).

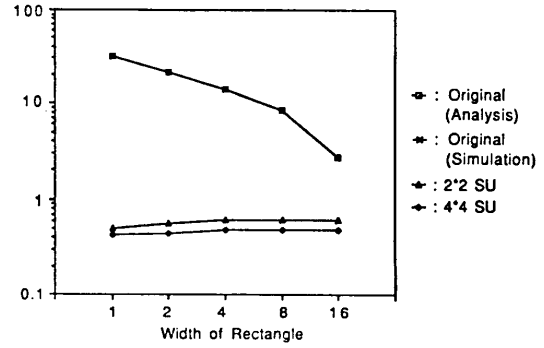


図 14 たんざくの幅と平均標準偏差
(B=128, T=1k, たんざく分布)

Fig. 14 Width of rectangle vs. mean standard deviation (N=16, B=128, T=1k, rectangular distribution).

ぼ一定である．ゆらぎもタプル数に関係なく 2.0 程度に収まっていた．**図 12** はたんざく分布の場合である．

バケット数を変化させた場合：ネットワークの大きさを一定にし，バケット数を変化させる．この場合，前の結果に示されたようにバケットの平坦度は処理モジュール当たりのタプル数に影響されるため，タプル数をバケット数に比例させてシミュレーションを行った．**図 13** は N=16, T=8×B (バケット数の 8 倍)，均一分布の場合の結果を示す．横軸はバケット数，縦軸は平均標準偏差をそれぞれ対数表示したものである．バケット数の違いはほとんど平均標準偏差に影響を与えないことがわかる．ゆらぎについても同様の結果が得られた．特に結果を図示していないが，たんざく分布でもバケット数の違いはほとんど平均標準偏差およびゆらぎに影響を与えなかった．

つぎに，バケットの分布の違いが平坦度に与える影響をさらに検討する．たんざく分布において 1 つの処理モジュールに格納されているバケットの範囲 X をたんざくの幅と呼び，それと平坦度の関係を調べた．先に定義したたんざく分布におけるたんざくの幅を 1 とすれば，X=N が均一出力分布に相当する．N=16, B=128, T=1k としてシミュレーションを行った結果が**図 14** である．横軸はたんざくの幅，縦軸は平均標準偏差をそれぞれ対数表示したものである．

ディスク上でのバケット分布ではたんざくの幅が広くなると平均標準偏差は小さくなるのに対し，本ネットワークでは大きくなっている．これは，たんざくの幅が広くなるほど異なる処理モジュールに格納されていたタプルが干渉し合うためと考えられる．

以上の結果をまとめると，平坦度が比較的悪くなる場合でも平均標準偏差で 1 以下，ゆらぎは 3 以下になり，分散制御によるバケット平坦化機能は有効であった．4×4 スイッチング装置を用いればバケット分布はより平坦になるが，2×2 スイッチング装置でも十分である．処理モジュール上での分布に偏りがあるたんざく分布に対する方が平坦度を改善する効果は大き

い.

5.3 考 察

分散制御方式はスイッチング装置が独立に動作するので，大域的に最適な行き先が選択されない．その影響を調べるため，集中制御方式との比較を行う．

シミュレーションでは，集中制御方式として以下のアルゴリズムを用いた．ネットワークで閉塞が起こらないように中央管理装置がタプルの行き先を逐次的に決定する．タプルはそのタプルの属するバケットのタプル数が最小の処理モジュール（行き先候補と呼ぶ）に転送されれば良いが，行き先候補数が少ないものほど閉塞を起こさない処理モジュールを見つけられる可能性が低いため，はじめに全タプルの行き先候補を捜し，その候補数の少ない順に行き先を決定する．タプルの行き先はつぎのように決定される．中央制御装置は現在行き先決定順にある処理モジュールのタプルの行き先候補のなかから，既に決定されている結合によってネットワークに閉塞が生じない処理モジュールを捜す．見つかった場合，その処理モジュールを行き先とし，行き先決定権をつぎの処理モジュールに移動する．見つからなかった場合，そのタプルに属するバケットのタプル数が2番目に少ないサブバケットを捜し，そのタプルに対する新たな行き先候補とする．（2番目に少ない行き先候補から閉塞を起こさない処理モジュールが見つけられなかった場合3番目に少ないものを行き先候補とする.）ここで，まだ行き先の決定されていない処理モジュールから行き先候補数が最小のものを選択し，その処理モジュールを新たな行き先決定の対象とする．以上の操作をすべての処理モジュール間の結合が決定するまで繰り返す．このアルゴリズムはバケット分散の最適な方法ではないが，かなり高い平坦度を実現できると考えられる．

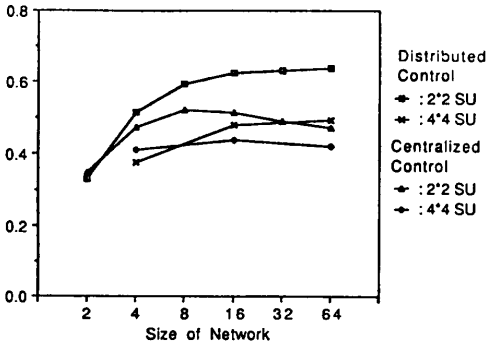


図 15 集中制御方式と分散制御方式の比較
(B=128, T=1k, 均一分布)

Fig. 15 Comparison of distributed control with centralized control (B=128, T=1k, uniform distribution).

均一分布として，集中制御方式と分散制御方式によるシミュレーションの結果を示したのが図15である．図15は処理モジュール当たりのタプル数とバケット数が一定（T=1k, B=128），均一分布の場合のネットワークの大きさと平均標準偏差の関係を示している．ネットワークが小さい場合，分散制御と集中制御の差は小さいが，ネットワークが大きいと集中制御の結果の方が良い．これはネットワークが大きくなると各スイッチング装置内の情報の局所性が増大し，有効な情報量が減少するためと考えられる．しかし，分散制御でもバケットを十分均一に分配しており，制御時間の少ない分散制御方式が適切であることが確認できた．

6. おわりに

本論文でバケット平坦化機能を持つオメガネットワークを提案した．バケット平坦化の制御方式として分散制御を採用した．分散制御方式では，ネットワークの各スイッチング装置はバケットの分布に関する局所的な情報に基づいて状態を自律的に決定し，バケットを処理モジュール上に均一に分配する．分散制御方式にはバケットの分布を集中的に管理する必要がない，バケット分配に必要な時間がネットワークの大きさに依存しない等の利点がある．また，ネットワークの構成要素として通常の2入力2出力のスイッチング装置を用いた場合だけでなく，4入力4出力のスイッチング装置を用いた場合の状態決定方式も提案した．

本オメガネットワークのバケット平坦化機能の評価をシミュレーションで行った．バケットの平坦度を平均標準偏差およびゆらぎで測定したが，本オメガネットワークはバケット平坦化に有効であることが確かめられた．バケットの処理モジュール上での分布が均一分布の場合でも 2×2 スイッチング装置で構成した本オメガネットワークにより平均標準偏差を 0.7 タプル，ゆらぎを 2.7 タプル以下にできた．たんざく分布の場合改善の効果は非常に大きかった．4×4 スイッチング装置は平坦化の効果がさらに大きく，平均標準偏差を 0.5 タプル，ゆらぎを2タプル以下に改善できた．バケット数 B の場合，2×2 スイッチング装置では B 個のカウンタと2つのカウンタ値を比較する機能があれば良いが，4×4 スイッチング装置では $4B$ 個のカウンタと 24 の組合せから評価値を最小にするものを選択する機能が必要となる．したがって，実装上

は 4×4 スイッチング装置では付加すべきハードウェアが複雑なため，2×2 スイッチング装置が望ましいと考えられる．

今後は本オメガネットワークの詳細な設計を進めてゆきたい．また，本ネットワークにより結合演算等の負荷の重い関係代数演算がどの程度高速化されるかについては稿を改めて述べることにしたい．

参考文献

1) 喜連川優，伏見信也：データベースマシン，情報処理，Vol. 28, No. 1, pp. 223-234 (1987).
2) Kitsuregawa, M., Tanaka, H. and Moto-oka, T.: Application of Hash to Data Base Machine and Its Architecture, *New Generation Computing*, Vol. 1, No. 1, pp. 66-74 (1983).
3) DeWitt, D. J., Katz, R. H., Olken, F., Shapiro, L. D., Stonebraker, M. R. and Wood, D.: Implementation Techniques for Main Memory Database Systems, *ACM SIGMOD '84*, pp. 1-8 (1984).
4) Yamane, Y.: A Hash Join Technique for Relational Database Systems, *Proc. on Foundations of Data Organization*, pp. 388-398 (1985).
5) Kitsuregawa, M., Tanaka, H. and Moto-oka, T.: Architecture and Performance of Parallel Relational Database Machine GRACE, *Proc. of the 14th Int. Conf. on Parallel Processing*, pp. 241-250 (1984).
6) DeWitt, D. J. and Gerber, R. H.: Multiprocessor Hash-Based Join Algorithms, *Proc. of the 11th Int. Conf. on VLDB*, pp. 151-164 (1985).
7) DeWitt, D. J., Gerber, R. H., Graefe, G., Heytens, M. L., Kumar, K. B. and Muralikrishna, M.: GAMMA, A High Performance Dataflow Database Machine, *Proc. of the 12th Int. Conf. on VLDB*, pp. 228-237 (1986).
8) 坂井修一，喜連川優，田中英彦，元岡　達：関係代数マシン GRACE におけるバケット収集網，電子通信学会論文誌，Vol. J 86-D, No. 1, pp. 9-16 (1985).
9) 坂井修一，喜連川優，田中英彦，元岡　達：関係代数マシン GRACE におけるバケット分配網，電子通信学会論文誌，Vol. J 86-D, No. 6, pp. 1272-1279 (1985).
10) Fushimi, S., Kitsuregawa, M. and Tanaka, H.: An Overview of the System Software of a Parallel Relational Database Machine GRACE, *Proc. of the 12th Int. Conf. on VLDB*, pp. 209-219 (1986).
11) Lawrie, D. H.: Access and Alignment of Data in an Array Processor, *IEEE Trans. Comput.*, Vol. C-24, No. 12, pp. 1145-1155 (1977).
12) 黒川恭一，相磯秀夫：結合方式，情報処理，Vol. 27, No. 9, pp. 1005-1021 (1986).

付録　ディスク上のバケット分布の平均標準偏差

バケットのディスク上での分布の σ を解析的に求める．ここで，N はネットワークの大きさ，T は処理モジュール当たりのタプル数，B はバケット数である．まず，均一分布の場合を扱う．平均標準偏差はつぎのように計算される．

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

タプルが属するバケットはタプルが格納されている処理モジュールに依存しないので σ は i とは独立に求めることができ，$\tilde{D}_i=\frac{1}{N}\sum_{j=1}^{N}D_{ij}$ とすれば

$$\sigma=\sqrt{\mathrm{E}\left\{\frac{1}{N}\sum_{j=1}^{N}D_{ij}^2-\left(\frac{1}{N}\sum_{j=1}^{N}D_{ij}\right)^2\right\}}$$
$$=\sqrt{\mathrm{E}\left\{\frac{1}{N}\sum_{j=1}^{N}(D_{ij}^2-\tilde{D}_i^2)\right\}}$$

となる．タプルが属するバケットはランダムに決定されるので，$\bar{D}=\mathrm{E}\left\{\frac{1}{BN}\sum_{i=1}^{B}\sum_{j=1}^{N}D_{ij}\right\}$ とすれば

$$\mathrm{E}\left\{\frac{1}{N}\sum_{j=1}^{N}(D_{ij}^2-\tilde{D}_i^2)\right\}$$
$$=\mathrm{E}\left\{\frac{1}{N}\sum_{j=1}^{N}(D_{ij}-\bar{D})^2\right\}$$
$$+2\bar{D}\,\mathrm{E}\left\{\frac{1}{N}\sum_{j=1}^{N}(D_{ij}-\tilde{D}_i)\right\}$$
$$-\mathrm{E}\left\{\frac{1}{N}\sum_{j=1}^{N}(\tilde{D}_i-\bar{D})^2\right\}$$
$$=\mathrm{E}\{(D_{ij}-\bar{D})^2\}-\mathrm{E}\{(\tilde{D}_i-\bar{D})^2\}$$

一方，

$$\mathrm{E}\{(\tilde{D}_i-\bar{D})^2\}=\mathrm{E}\left\{\left[\frac{1}{N}\left(\sum_{j=1}^{N}D_{ij}\right)-\bar{D}\right]^2\right\}$$
$$=\frac{1}{N^2}\mathrm{E}\left\{\left[\sum_{j=1}^{N}(D_{ij}-\bar{D})\right]^2\right\}$$
$$=\frac{1}{N^2}\mathrm{E}\left\{\sum_{j=1}^{N}(D_{ij}-\bar{D})^2\right\}$$
$$=\frac{1}{N}[\mathrm{E}\{(D_{ij}-\bar{D})^2\}]$$

ここでは，$\mathrm{E}\{\sum_{k\neq j}(D_{ij}-\bar{D})(D_{ik}-\bar{D})\}=0$ の関係を用いている．したがって，

$$\mathrm{E}\left\{\frac{1}{N}\sum_{j=1}^{N}(D_{ij}^2-\tilde{D}_i^2)\right\}=\left(1-\frac{1}{N}\right)\mathrm{E}\{(D_{ij}-\bar{D})^2\}$$

一方，タプルがディスクにある確率は $\rho=1/B$ なので

$$E\{(D_{ij}-\bar{D})^2\}=T\rho(1-\rho)=\frac{T}{B}\left(1-\frac{1}{B}\right)$$

以上から，均一分布の場合の平均標準偏差は

$$\bar{\sigma}=\sqrt{\frac{T}{B}\left(1-\frac{1}{B}\right)\left(1-\frac{1}{N}\right)}$$

つぎに，たんざく分布の場合を扱う．X はたんざくの幅である．いま

$$\frac{1}{N}\sum_{j=1}^{N}D_{ij}^2-\left(\frac{1}{N}\sum_{j=1}^{N}D_{ij}\right)^2$$
$$=\frac{X}{N}\left\{\frac{1}{X}\sum_{j=1}^{N}D_{ij}^2-\left(\frac{1}{X}\sum_{j=1}^{N}D_{ij}\right)^2\right\}$$
$$+\frac{1}{N}\left(\frac{1}{X}-\frac{1}{N}\right)\left(\sum_{j=1}^{N}D_{ij}\right)^2$$

である．たんざく分布では1つのディスクにあるバケット数は $X\times B/N$ であり，均一分布の式を用いると

$$E\left\{\frac{1}{X}\sum_{j=1}^{N}D_{ij}^2-\left(\frac{1}{X}\sum_{j=1}^{N}D_{ij}\right)^2\right\}$$
$$=E\left\{\frac{1}{X}\sum_{j=1}^{X}D_{ij}^2-\left(\frac{1}{X}\sum_{j=1}^{X}D_{ij}\right)^2\right\}$$
$$=\frac{TN}{XB}\left(1-\frac{N}{XB}\right)\left(1-\frac{1}{N}\right)$$

さらに，$E\{\sum_j D_{ij}\}=TN/B$ だから

$$E\left\{\frac{1}{N}\sum_{j=1}^{N}D_{ij}^2-\left(\frac{1}{N}\sum_{j=1}^{N}D_{ij}\right)^2\right\}$$
$$=\frac{X}{N}\frac{TN}{BX}\left(1-\frac{N}{BX}\right)\left(1-\frac{1}{X}\right)+\frac{1}{N}\left(\frac{1}{X}-\frac{1}{N}\right)\left(\frac{TN}{B}\right)^2$$
$$=\frac{T}{B}\left(1-\frac{N}{BX}\right)\left(1-\frac{1}{X}\right)+\frac{NT^2}{B^2}\left(\frac{1}{X}-\frac{1}{N}\right)$$

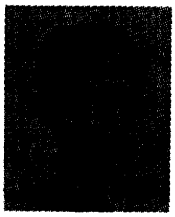
したがって

$$\bar{\sigma}=\sqrt{\frac{T}{B}\left(1-\frac{N}{BX}\right)\left(1-\frac{1}{X}\right)+\frac{NT^2}{B^2}\left(\frac{1}{X}-\frac{1}{N}\right)}$$

均一分布 ($X=N$) の場合，均一分布の式と一致することは容易に確かめられる．また，たんざく分布 ($X=1$) の場合，つぎのようになる．

$$\bar{\sigma}=\frac{T}{B}\sqrt{N-1}$$

(平成元年4月6日受付)
(平成元年7月18日採録)



喜連川 優 (正会員)

昭和30年生．昭和53年東京大学工学部電子工学科卒業．昭和58年同大学院情報工学専門課程博士課程修了．工学博士．同年，東京大学生産技術研究所講師．現在，同研究所助教授．並列コンピュータアーキテクチャ，データベースマシン，データ工学等の研究に従事．電子情報通信学会，電気学会，IEEE 各会員．

小川 泰嗣 (正会員)

1962年生．1985年東京大学工学部計数工学科卒業．1987年同大学院工学系研究科情報工学専攻修士課程修了．同年，(株)リコー入社．中央研究所にて情報検索およびデータベースシステムの研究開発に従事．IEEE, ACM 各会員．