多次元構造を持つ MIN の提案

朱 笑岩　　亀井 貴之　　塙 敏博　　米田 卓司　　天野 英晴

慶應義塾大学 理工学部
〒 223 横浜市 港北区 日吉 3-14-1
E-mail: cissy,kamei,hanawa,komeda,hunga@aa.cs.keio.ac.jp

本報告では、多次元構造を持つ MIN、$nD-MIN$ を提案する。$nD-MIN$ は多次元化と循環リンクを付加することにより、特定のメモリモジュールに対するステップ数を減らし、アクセスの局所性を利用することができる。

確率モデルによるシミュレーションの結果、アクセス局所性が存在しない場合の性能は通常の MIN と変わらず、アクセスの局所性が存在する場合、アクセスレイテンシィを最大 20%、プロセッサ利用率を最大 2 倍改善することができることがわかった。

並列計算機、MIN、アクセス局所性、性能評価

A proposal of multi-dimensional MIN

Xiaoyan Zhu　　Takayuki Kamei　　Toshihiro Hanawa　　Takuji Komeda
Hideharu Amano

Keio University
223 YOKOHAMA SHI KOUHOKU KU HIYOSHI 3-14-1
E-mail: cissy,kamei,hanawa,komeda,hunga@aa.cs.keio.ac.jp

In this paper, we propose a novel MIN (multistage interconnection network) structure called the multi-dimensional MIN (nD-MIN). Unlike the conventional MINs, the nD-MIN can make the best use of access locality by its multi-dimensional structure and feedback lines.

According to the probabilistic simulation results, the performance of the nD-MIN is almost the same that of the conventional MINs when the access is randomly distributed. Under the traffic with access locality, the access latency and processor utilization ratio are improved up to 20 % and 2.0 times respectively.

Parallel computer, Multistage Interconnection Network, Access Locality, Performance E-valuation

1 まえがき

近年、大規模並列計算機、超並列計算機のプロジェクトが各地で盛んに行われている。こうした大規模な並列計算機の構成を 考える上で、プロセッサ間あるいはプロセッサ-メモリ間の相互結合方式は重要な位置を占めており、並列計算機の規模、応用問題の性質などに応じた最適な結合網を求める研究が積極的に進められている。その中で、多段結合網 **MIN** ($Multistage\ Interconnection\ Network$) はスケーラビリテイが高く、転送容量が大きい利点をもつため、プロセッシングユニット (PU) と共有メモリモジュールを結合する方法として、古くから研究されてきた [1]。

従来の MIN は、PU とメモリモジュールの間で交換されるパケットは 必ず同じ数のステージを通る必要がある。この構造では 、同時に各メモリモジュールから各 PU に大量にデータを転送するには便利であるが、PU がメモリモジュールに対して行なうアクセスに局所性が存在する場合、その利点を活かすことができない。システムが大規模になると、すべての PU がすべてのメモリモジュールに対して一様にアクセスを発生することは考え難いため、MIN の特徴を保持したまま、特定のメモリモジュールに対するアクセスを効率良く行なうことのできる構成が有利になる。

本研究では、超並列計算機 RWC-1 用の直接網として検討された CB^2[2][3] で用いられている結合法を MIN に応用した n 次元 MIN(**nD-MIN**) を提案する。この MIN は、循環構造と多次元構造を持つことで、MIN の特徴を保ったまま、局所性を生かすことができる。さらに、そのレイテンシィについて確率モデルに基づき評価を行なう。

2 多次元 MIN の提案

2.1 nD-MIN の構成原理

nD-MIN は、ステージ間の接続を多次元方向に行なうことにより構成する。ここでは、基本となる MIN に図 1中に示す Banyan 網 (Generalized Cube 網 [4]) を用いる。この網は、ひとつのスイッチングエレメントが、出力と同一番号の端子に対して接続されたリンク (並行リンク) と 2^s (s はステージ番号) 離れた番号の端子に対して接続されたリンク (クロスリンク) を持ち、直接網のハイパーキューブに相当する多段網を構成する。

nD-MIN は、このクロスリンクを多次元方向にも設けることにより構成する。図 1は 3 次元方向への拡張法を示す。

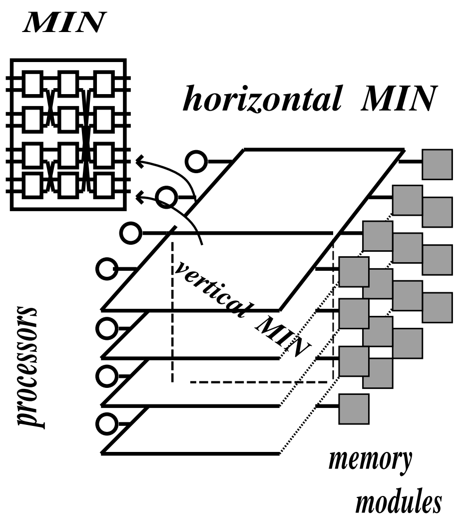

図 1: nD-MIN の基本構成

図 1 の構成 では、自分と同一の水平および垂直平面上のメモリモジュールしかアクセスできない。そこで、図 1の結合網に図 2に示すように循環リンクを付加する。この構造は、直接網の中の循環網 [7] に属する Circular Banyan 網 [2][3] のトポロジを MIN に利用したものと考えることができる。

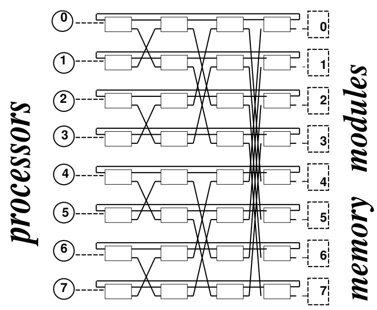

図 2: Circular Banyan 網

Circular Banyan 網を基本とした 3D-MIN を図 3に示す。この構造は直接網である CB^2[7] を MIN に応用したものと考えることもできる。循環リンクを持つことによりハードウェア量は増加するが、任意のメモリモジュールにアクセスすることができ

る。すなわち、3 次元 MIN では水平、垂直のそれぞれの方向で同一平面上のメモリモジュールに対しては直接アクセス可能である。他のメモリモジュールに対しては、最初に垂直方向の MIN を用いて目的地と同一垂直平面上にパケットを送り、一旦 循環リンクを用いて出発地に戻った後 、水平方向の MIN を用いて目的地にパケットを送ることによりアクセスすることができる。

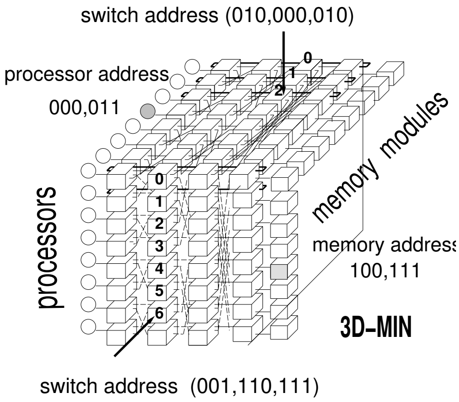

図 3: 3 次元 MIN の構成

2.2 nD-MIN の結合

図 3に示すように 3D-MIN では、各スイッチングエレメントは三次元のアドレス：x_0(ステージアドレス)、x_1(水平平面アドレス)、x_2(垂直平面アドレス) を割り当てる。それぞれのスイッチングエレメントは Banyan 網の接続方法に従い一つの並行リンクとそれぞれの次元方向のクロスリンクにより次のステージのスイッチングエレメントと接続する。接続は以下の式で表すことができる：

- 3D-MIN:

並行リンク
(x_0, x_1, x_2) から
$((x_0+1)modN, x_1, x_2)$

クロスリンク
(x_0, x_1, x_2) から
$(x_0+1, \mathrm{rev}(x_1, x_0), x_2)$
$\ldots$
$(x_0+1, x_1, \mathrm{rev}(x_2, x_0))$

但し、N はスイッチのステージ数で、$rev(a,b)$ は a の b 番目のビットを反転する関数である。$(x_0+1)modN$ は循環リンクに相当する。

上記の結合を一般化することにより、nD-MIN の結合を定義することができる。nD-MIN では、スイッチングエレメントには、n 次元の番号 $(x_0, x_1, \ldots x_{n-1})$ が割り当てられる。3 次元同様に、それぞれの次元の接続を Banyan 網のクロスリンクにより行なうと、nD-MIN の接続は以下の式で定義することができる。

並行リンク
$(x_0, x_1, \ldots, x_{n-1})$ から
$((x_0+1)modN, x_1, \ldots, x_{n-1})$

クロスリンク
$(x_0, x_1, \ldots, x_{n-1})$ から
$(x_0+1, \mathrm{rev}(x_1, x_0), \ldots, x_{n-1})$
$\ldots$
$(x_0+1, x_1, \ldots, \mathrm{rev}(x_{n-1}, x_0))$

2.3 ルーティング

$nD-MIN$ のルーティングアルゴリズムは非常に単純である。まず 3D-MIN を例にとって説明する。

まず、PU とメモリモジュールに 2 次元のアドレス x_0 (水平アドレス) と x_1(垂直アドレス) を割り付ける。まず、パケットを送信した PU のアドレスと、目的地のメモリモジュールのアドレスの、ビットごとの排他論理和を求める。すなわち、プロセッサのアドレスを $P(x_0, x_1)$ とし、メモリモジュールのアドレスを $M(y_0, y_1)$ とすると：

両方のアドレスを排他論理和で構成されるタグ $D(d_0, d_1)$ は

$$
\begin{aligned}
&D(d_0, d_1) \\
&\quad = P(x_0, x_1) \oplus M(y_0, y_1) \\
&\quad = D(x_0 \oplus y_0, x_1 \oplus y_1)
\end{aligned}
$$

ここで、目的地が同一水平平面上になければ $(x_0 \neq y_0)$, 最初に垂直平面上をアクセスのためのパケットを転送する。この時、タグ (d_0) をビット単位でチェックし、1 であれば、垂直方向のクロスリンクから出力し、0 であれば、並行リンクから出力する。次に循環リンクを介してプロセッサ側にパケットを戻して水平平面上で タグ (d_1) を用いて同様にルーティングを行なう。

一般に、$nD-MIN$ のルーティング用のタグは以下の式で示すことができる：

$$
\begin{aligned}
&D(d_0, d_1, \ldots, d_{n-2}) \\
&\quad = P(x_0, x_1, \ldots, x_{n-2}) \oplus M(y_0, y_1, \ldots, y_{n-2}) \\
&\quad = D(x_0 \oplus y_0, x_1 \oplus y_1, \ldots, x_{n-2} \oplus y_{n-2})
\end{aligned}
$$

このタグを d_0から順に適用して、それぞれの次元のリンクにパケットを送ることにより、3 次元同様に、n 次元でもパケットのルーティングが可能になる。

2.4 転送ステップ数

メモリモジュールの数を一定にした場合、 nD-MIN では次元 n に対して最大 n-1 回循環リンクを介して MIN を通らなければならない。このため、最大転送ステップ数は大きくなる。一方、MIN 自体が必要とするステージ数は少なくなることから、同一次元上のメモリモジュールをアクセスする場合のレイテンシィは小さくて済む。

以下、$nD-MIN$ において、パケットの衝突等を考えない場合、転送ステップ数について検討する。あるプロセッサ $P(x_0, x_1, \ldots, x_{n-2})$ に対して、MIN を一回通ることで到着できるメモリモジュールは 0 から $n-2$ までの各桁の中に異なった数字が 1 桁以下のメモリモジュールである。従って、このようなメモリモジュールの数 T_1は下記の式で表される

$$(M_0 - 1) + (M_1 - 1) + \ldots + (M_{n-2} - 1) + 1$$

ただし、M_0から M_{n-2}までの項は 1 桁違いのメモリモジュールの数であり、最後の+1 はアドレスすべて同じであるメモリモジュールである。これはアドレス全く同一であっても、少なくとも一回 MIN を通って (すべて並行リンクで) ルーティングする必要があるためである。

さて、MIN を二回通過する必要があるメモリモジュールはアドレスが 2 桁異なっているモジュールであり、

$$
\begin{aligned}
&((M_0 - 1) \times (M_1 - 1) +, \ldots, \\
&\qquad + (M_0 - 1) \times (M_{n-2} - 1), \\
&+((M_1 - 1) \times (M_2 - 1)) +, \ldots, \\
&\qquad + ((M_1 - 1) \times (M_{n-2} - 1)), \\
&\qquad +, \ldots, \\
&(M_{n-3} - 1) \times (M_{n-2} - 1))
\end{aligned}
$$

で示すことができる。これを T_2とし、同様の式で T_{n-1}までを計算する。

次に、プロセッサ (メモリモジュール) 数を M とすると、ステージ数は、

$$\lceil \frac{\log M}{n-1} + 1 \rceil$$

となる。これより、平均転送ステップ数は、

$$\lceil \frac{\log M}{n-1} + 1 \rceil \times (T_1 + T_2 \times 2 + \ldots + T_{n-1} \times (n-1))$$

となる。この式よりプロセッサ数 64 から 1024 までの平均転送ステップ数を計算した結果をグラフ 4で示す。

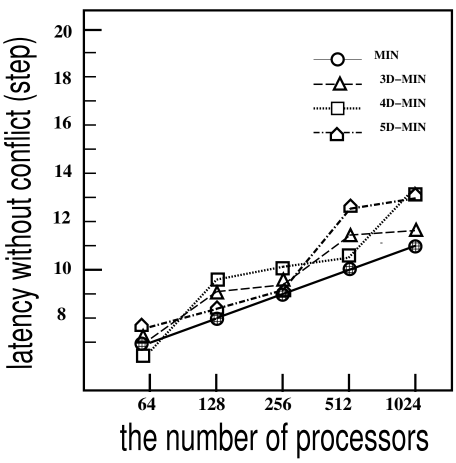

図 4: 平均転送ステップ数

このグラフにより、うまくサイズが合わない場合に、多少変動が見られるものの、平均転送ステージ数は次元数により大きい変化がないことがわかる。

2.5 ハードウェアコスト

先に述べたようにプロセッサ数 (メモリモジュール数) を M、MIN の次元を n とすると、ステージ数は

$$\lceil \frac{\log M}{n-1} + 1 \rceil$$

となる。したがってエレメント数は、

$$M \times \lceil \frac{\log M}{n-1} + 1 \rceil$$

となる。一般の MIN におけるエレメント数は、

$$\frac{M}{2} \times \log M$$

であるので、エレメント数に関しては、nD-MIN は次元数が増える程、一般の MIN に比べて有利であることがわかる。

しかし、nD-MIN はその分入出力が増えるため、クロスポイント数で評価すると、

$$M \times \lceil \frac{\log M}{n-1} + 1 \rceil \times n^2$$

となって、各エレメントのクロスポイント数が常に 4 である MIN に比べて不利であることがわかる。しかし、最近の LSI 技術の発達を考えるとクロスポイント数が多くても LSI のピンネックにならない範囲では一概に不利とはいえず、エレメントの配置、リンクの配線のしやすさ等も含め検討する必要がある。

3　シミュレーションによる性能評価

理論的に求めることのできる転送ステップ数は、ネットワーク内部でのパケットの衝突による遅れが考慮されていないため、実際的ではない。そこで、確率モデルに基づきやや実際的なシミュレーションを行なう。

3.1　評価条件

ここでは、ローカルメモリまたはキャッシュを持つプロセッサが nD-MIN を介してメモリモジュールと接続されているシステムを想定する。命令およびスタックはローカルメモリまたはキャッシュに格納されていると考え、各プロセッサは、以下のアクセスを各クロックでデータに対して発生するとした。

- n_{br}: ブロックアクセス。読み出しまたは同期変数へのアクセス。各プロセッサはデータが到着するまで次のアクセスを発生することができない。

- n_{nb}: ノンブロッキングアクセス。書き込み、プリフェッチ等のアクセスで、プロセッサはアクセスを発生してすぐ次の計算またはアクセスに移ることができる。

ここでは、実際に稼働しているマルチプロセッサ SNAIL[5] の評価に基づき、$n_{br} = 0.05, n_{nb} = 0.025$ に設定した。プロセッサ数は $(2^6, 2^7, 2^8, 2^9, 2^{10})$ とし、100000 ステップをシミュレーションした。なお、ネットワークが安定するまでの 10000 ステップは、シミュレーション結果から除いた。

スイッチのモデルは、エレメント内部に パケット格納用バッファを持つ一般的な MIN であり、パケットは 1 クロックでエレメント間を転送される。読み出しデータ用に、メモリモジュールから PU へ向かう帰還用の MIN を別に持つものとする。スイッチングエレメントは、入力リンクに対して FIFO を、出力リンクに対してパケット 1 個分格納できるバッファを持つ。次のステージのエレメントの FIFO が溢れた場合、パケットは出力リンクのバッファで待ち状態になる。ここではシミュレーション時間とメモリの関係で FIFO で格納できるパケットの最大数は 3 とした。パケットの優先制御は以下のように行なった。

- 並行リンクからのパケットに最高の優先権を与え、順に次元数の低いリンクから来たパケットほど高い優先権を与える。

- 衝突して、一回待たせられたパケットはその都度優先権を上げ、次回の転送に有利になるようにする。

また、メモリのアクセス時間は 3 クロックとし、メモリモジュール内はスイッチングエレメント同様パケット 3 個分の FIFO を持つ。

3.2　衝突率

図 5に、あるパケットが、バッファや FIFO が溢れたため、または行き先がぶつかったため、待たされる割り合い (衝突率) を示す。この結果は、パケットの行き先を均等に分布された場合である。

プロセッサの増加に伴いネットワークの混雑も大きくなり、衝突率はも高くなる傾向にある。通常の MIN は、最も衝突率が低く、n-D MIN はプロセッサ数が増加すれば、衝突率が段々大きくなる。また次元数が高くなるに伴い、衝突率が増加する幅が大きくなる。これは、パケットの行き先を均等に分布された場合、循環リンク を通って戻ってくるパケットによりネットワークが混雑するためである。しかし、このシミュレーションモデルでは最大次元のものでも 20%程度で、さほど大きくなっていない。これは、後に述べるように、ブロッキングアクセスによりプロセッサの待ち時間が長くなり、一定以上にネットワークにパケットが挿入されたいためである。

以上の結果により、アクセスが均等である場合でも循環リンク による衝突率の増加はさほど大きくないことがわかる。

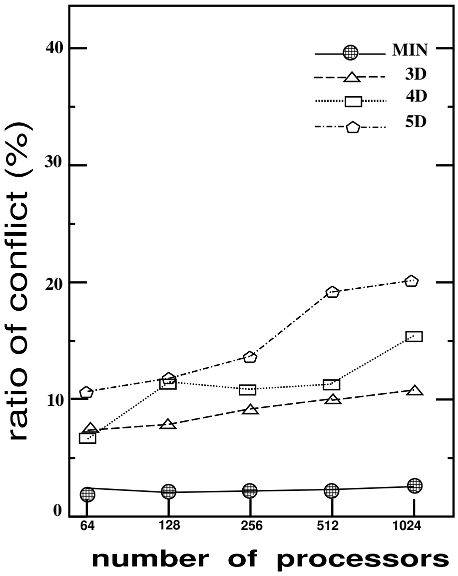

図 5: 衝突率

3.3 アクセスレイテンシィ

ノンブロッキングアクセスについて、プロセッサがアクセスを発生してパケットが戻ってくるまでのクロック数を測定した。まず、アクセスが局所性を持っていない場合に関して、プロセッサ数 64 から 1024 まで変えた場合 のレイテンシィを図 6に示す。

図 6 によると、局所性のない場合でも、nD-MIN は通常の 2 次元 MIN に比べ、サイズが合わない場合を除き、レイテンシィは増加していないことがわかる。また、サイズが合う場合は、2 次元 MIN に比べ低いレイテンシィを実現している。これは、nD-MIN が交換能力の高いスイッチングエレ メントを用いているためと考えられる。

次に、アクセスに局所性のある場合のレイテンシィを測定した。これは、アクセスの 60%がそれぞれ次元数の一致するメモリモジュールに対して行なわれると仮定した。この場合、通常の 2 次元 MIN に対し、アクセスの局所性を与えると、かえってホットスポットを形成して性能が低下するため、通常の MIN に対するアクセスは均一アクセスを用いた。この場合、図 7に示すように、局所性を生かすこと

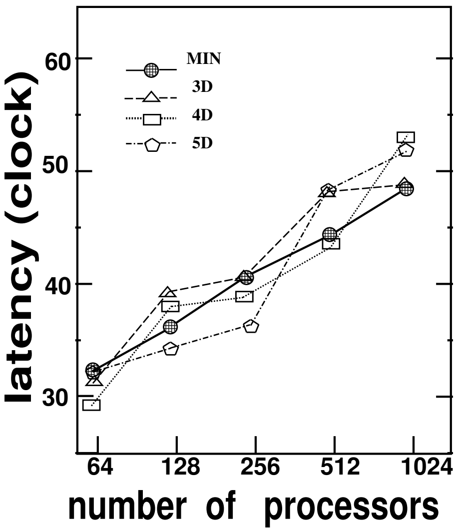

図 6: レイテンシィ

ができる nD-MIN が有利になり、最大 20%改善される。

3.3.1 プロセッサ利用率

ネットワークの最終的な性能は、これを介したアクセスによりプロセッサがどれだけ待たされるかによって決まる。そこで、全実行時間中、アクセスのために待たされる時間の割合の平均を、プロセッサ利用率として測定した。 アクセスが均等に分散している場合の結果を図 8に示す。プロセッサ数が増加すると、レイテンシィが増加するため、プロセッサ利用率が小さくなる。このシミュレーションでは小規模なマルチプロセッサである SNAIL のデータを用いたため、プロセッサの性能をネットワークのレイテンシィに対して高めに設定していることになり、プロセッサ利用率はあまり高くなく、サイズによっては 50%を切っている。このような場合、プロセッサは待ち状態が大きくなることにより、次のアクセスまでの間隔が大きくなり、ネットワークの混雑は緩和される傾向にある。今回のシミュレーションで結合網の衝突率があまり高くならないのはこのためである。

プロセッサ利用率についても、レイテンシィと同様、アクセスが均等な場合は、nD-MIN と通常の MIN の間にはさほど大きな差はない。

これに対し、アクセスに局所性が存在する場合のプロセッサ利用率は図 9に示すように、nD-MIN

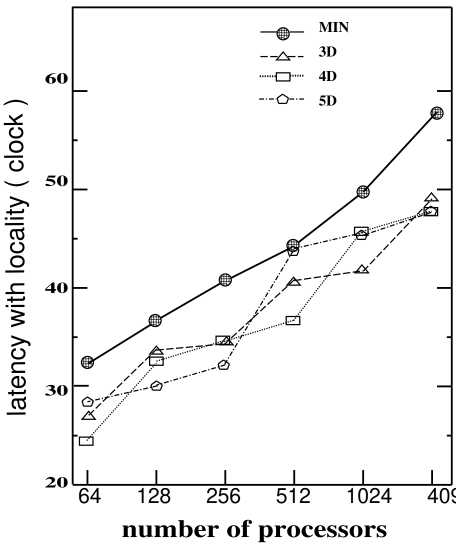

図 7: 局所性アクセス場合のレイテンシィ

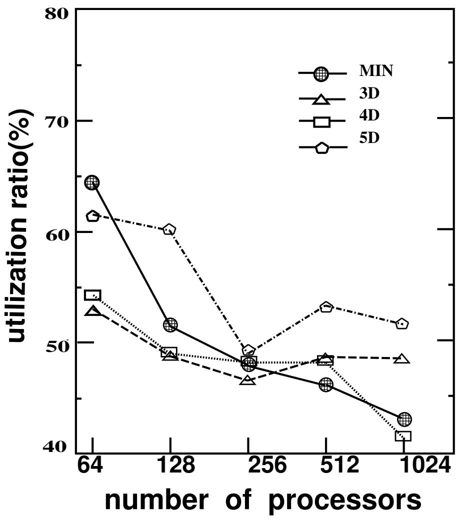

図 8: プロセッサ利用率

が有利になる。サイズが大きい場合、通常の MIN に対して最大 2 倍程度利用率を改善する効果があることがわかる。

4 まとめ

本報告では、局所性を利用することのできる MIN である nD-MIN を提案し、転送レイテンシィ、プロセッサ利用率について簡単な確率モデルによる解析を行なった。その結果、アクセスに局所性がある場合、プロセッサ利用率を 2 倍近く改善できることがわかった。

また、ここで提案した nD-MIN は、以下のような性質がある。

- パーティション: nD-MIN は、各次元内部に 2^m個のリングを含む領域に分割することが可能である。また、次元間も同じく 2^mサイズに分割可能である。さらに、各次元は直交しており、任意の組合せでの分割も可能である。以上により、提案している nD-MIN は基本となる Banyan 網同様、閉パーティションを構成することができる。

- 耐故障性: nD-MIN は、循環リンクを利用することにより、エレメントまたはリンクの故障を迂回することができる。プロセッサ、メモリに直接接続されているエレメントが故障しない限り、プロセッサ-メモリ間の転送機能は維持することができると考えられる。

- 混雑の迂回: 同様に循環リンクを利用することにより、複数のパスを選択して、混雑を迂回する適応型ルーティングを用いることが可能である。

今後、基本的な性能に関してさらに シミュレーションを行ない、評価していく一方、上記の可能性およびパーミュテーション能力などを検討していく予定である。また、同じく多次元構造を持つ、PBSF 網 [4] や MDX 網 [8] との比較等に関しても、検討を行なっていく予定である。

参考文献

[1] Broomell,G. and Heath,J.R. : "Classification Categories and Historical Development of Circuit Switching Topologies," ACM Computing Surveys,Vol.15,No.2,Jun. 1983. pp.93-133

[2] 横田、他「超並列計算機 RWC-1 の相互結合網」情処研報 93-ARC-101

[3] 横田、松岡、岡本、広野、坂井、「超並列向け相互結合網 MDCE の提案と評価」情報処理学会論文誌、Vol.36,No.7, Jul.1995.pp.1600-1609

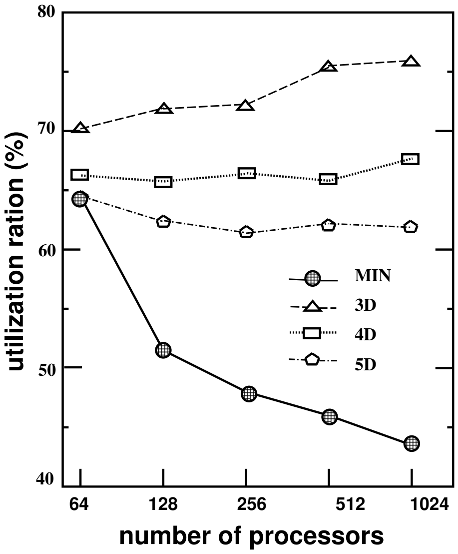

図 9: 局所性アクセス場合のプロセッサ利用率

[4] H.J.Siegel, and S.D.Smith, "Study of multistage SIMD interconnection networks," International Symposium on Computer Architecture 78 ,pp.223-229

[5] 塙敏博, 天野英晴「多重出力可能な MIN の性能評価」情報処理学会論文誌 第 36 巻 第 7 号 1995 年 7 月

[6] 笹原、寺田、大和、塙、天野、「SSS 型 MIN に基づくマルチプロセッサ $SNAIL$ の実装と評価」情報処理学会論文誌 Vol.36,No.7,Jul. 1995. pp.1640-1651

[7] 児玉、坂井、山口「高並列計算機 EM-4 とその並列性能評価」電子情報通信学会論文誌 ,J75-D-I,No.8,1992. pp.607-614

[8] 天野 英晴 「並列コンピュータ」 コロナ社、1996.