超並列計算機用多段結合網における転送性能の解析

三島 健[†1] 朴 泰祐[†2]
中村 宏[†3] 中澤 喜三郎[†4]

本論文では，並列計算機用相互結合網である多段結合網に対する，確率モデルに基づく理論解析手法を新たに提案する．これまでの同網に対する理論解析は，通信の分野におけるパケット網あるいは回線交換を主として行われており，現在の並列計算機の主流となっているwormhole転送方式をモデルとしたものは数少ない．また，それらは解析の都合上，いくつかの制約を持っており，そのまま超並列計算機における結合網解析手法として用いるには難しいものも多い．本論文で提案する手法は，結合網上の各種状態値の確率を厳密に定め，それに基づく平均遅延およびバンド幅を，マルコフ連鎖等のモデルに比べはるかに低コストで求めることが可能である．超並列計算機を想定した大規模多段結合網を対象とし，各種パラメータの下で本手法による解析値と計算機シミュレーションによる値を比較した結果，両者が非常に高い精度で一致することを確認した．

Performance Analysis of Multistage Interconnection Network for Massively Parallel Processors

TAKESHI MISHIMA,[†1] TAISUKE BOKU,[†2] HIROSHI NAKAMURA[†3] and KISABURO NAKAZAWA[†4]

In this paper, we propose a new method for theoretical performance analysis of multi-stage interconnection network, based on a probability model. In several researches on that class of network so far, they mainly focused on packet switching and circuit switching networks for telecommunication fields, while there were few works focussing wormhole routing algorithm which is generally used in parallel processing systems. Moreover, these research models have several restrictions due to easy analysis, which is difficult to apply directly for massively parallel processing systems. In our proposing method, we exactly define and estimate all status values in the network, and can estimate the average latency and bandwidth of the network with very low cost compared with other models like Marcov Chain. Comparing with computer simulation results for various conditions and parameters, we confirmed that the results in our method match with them in very high accuracy.

1. はじめに

超並列計算機を効率良く動作させるためには，PU（Processing Unit）間の通信を高速に行うことが重要であり，これは相互結合網のトポロジやルーティング・アルゴリズムに大きく依存する．これまでに多くの結合網が提案されてきたが，基本特性や拡張性を考えた場合，数千台規模の超並列計算機に採用できるものは限られている．

その中でも，多段結合網（Multistage Interconnection Network，以下MINと省略）は，高速通信用ネットワークとして広く研究されてきた．MINは，比較的小規模なクロスバ・スイッチを組み合わせた多段の間接網であり，通信チャネル数が多いため，大規模なシステムにおける複雑な転送パターンに対しても比較的高い通信性能を保つことができる．また，並列計算機システムのネットワークばかりでなく，高速通信転送におけるスイッチング・ノードなどにも広く使われている．さらに最近では，VLSI技術およびパッケージング技術の発展にともない，中規模（8×8あるいは16×16程度）で高バンド幅を持つクロスバスイッ

†1 NTTネットワークサービスシステム研究所
NTT Network Service Systems Laboratories
†2 筑波大学電子・情報工学系
Institute of Information Sciences and Electronics, University of Tsukuba
†3 東京大学先端科学技術研究センター
Research Center for Advanced Science and Technology, University of Tokyo
†4 明星大学情報学部電子情報学科
Department of Electronics and Computer Science, Meisei University

チも実現されており，超並列計算機への適用が注目されている．

MINに対する過去の研究では，計算機シミュレーションおよび理論解析によってその転送性能が評価されてきたが，MINに対する多くの理論解析では，主として回線交換やパケット交換など通信技術における転送方式のみが対象とされてきた[1)~7)]．これに対し，分散メモリ型の並列計算機における転送方式では一般にWormhole方式を採用しているため，これらの解析結果を適用することは難しい．そればかりでなく，それらの多くは，そのネットワークが並列計算機に応用された場合に一般に想定される仮定に合わないモデルとなっている．たとえば，文献1)では，クロスバにバッファがないモデルを想定していて，競合に負けたメッセージは捨てるという仮定で解析を行っている．文献2)および3)では，クロスバが2×2の場合に限定しているため，中規模のクロスバを使ったネットワークの解析には適用できない．文献4)は，メッセージのブロッキングによる時間依存性を考慮していないため，現実的ではない．文献5)および6)は，ネットワーク中で生じる時間依存性は考慮しているが，ネットワークの入口での時間依存性を解決していない．また，クロスバが2×2の場合に限定されているため，汎用性がない．文献7)も，クロスバが2×2の場合に限定されている．さらに1つのバッファに複数のメッセージが同時に入ることができると仮定しているが，このような仮定はネットワークの高速性を考えると実現は困難である．

一方，Wormhole方式におけるネットワークの理論解析はk-ary n-cube Networkなど直接網のみが対象とされてきた[8),9)]．これは，これまでの超並列計算機用ネットワークが主として直接網を中心に研究されてきたという背景によるものであるが，間接網におけるWormhole方式における理論解析はおろそかにされていた．

つまり，任意のクロスバ・サイズで解析可能であること，バッファ・サイズが有限であること，メッセージ長は1以上の有限であること，メッセージのブロッキングによる時間依存性を考慮に入れること，という現実的なMINの解析を行うための4つの条件をすべて満たし，かつWormhole方式を仮定した理論解析はこれまで行われていない．したがって，並列計算機に適した，より現実的な評価手法が望まれる．本研究では，より現実的な並列計算機の解析手法を提案する．

以下では，まず2章で本研究で新たに提案する2つの理論解析手法を示し，3章ではその理論解析によって得られた値とシミュレーションによって得られた値を比較し，提案した解析手法の有効性を評価する．最後に結論を述べる．

2. MINの解析

2.1 解析の基本方針

本論文で提案する手法は，確率モデルを基本とするものである．すなわち，システム中の各種状態変数の値を確率的に求め，それらの組合せにより，ネットワーク全体の転送性能を求める手法を用いる．

ここでは，MINのネットワーク中での待ち時間および転送スループットを理論的に求める手法について述べる．解析を行うために次のような仮定を用いた．

- メッセージ長は固定長で，flit[10)]単位で扱う．
- メッセージの転送先PUは一様ランダムに分布し，メッセージごとに独立に与えられる．
- メッセージ間には，依存関係がない．
- 各PUのメッセージ発生確率は同一とする（後述のλは各PUで一定）．
- ルーティングはWormhole方式である．
- 衝突したメッセージの調停方法は，FCFS（First Come First Service）priorityに基づく．
- クロスバ・スイッチの入力には1 flit分のバッファがあり，ブロックされたメッセージの一部はこのバッファに蓄えられる．

図1に解析の対象となるクロスバ・スイッチとPUのモデルを示す．もし，PUがメッセージを発生したとき，それ以前に発生したメッセージがPUから出ていなかったら，PU中のバッファにメッセージは蓄えられる．

基本的には，解析が複雑で，計算コストの大きいマルコフ連鎖などは用いず，きわめて小さな計算コストで済む解析手法を用いる．それは，前の時間における

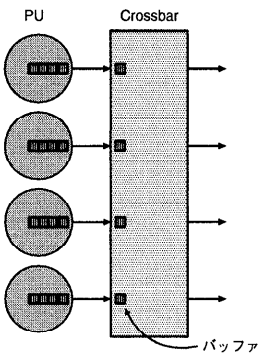


図1 クロスバ・スイッチ
Fig. 1 Crossbar switch.

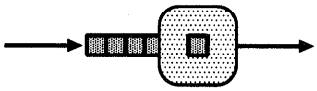


図2 簡単化されたモデル
Fig. 2 Simple model.

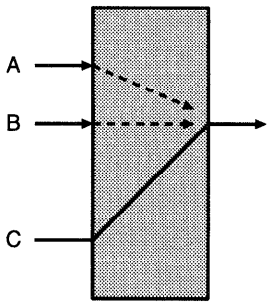


図3 メッセージの同時到着
Fig. 3 Message arriving at the same time.

状態確率から現在の状態確率を求める理論式を繰り返し計算して定常確率を求める手法である．各PUのメッセージ発生確率が同じで，かつメッセージの転送先PUを一様ランダムに決めるので，各入力，出力チャネルを平均化して考えることができる．つまり，同じステージ中の入力チャネルどうし，出力チャネルどうしは区別する必要がない．このようにすることによって解析を簡単化することができる．つまり，**図2**のようにサーバと待ち行列といった簡単化されたモデルで解析を行えばよい．

解析を行うにあたって，メッセージの同時到着を考慮に入れる解析手法とメッセージの同時到着を考慮に入れない解析手法が考えられる．メッセージの同時到着とは，**図3**のメッセージAとメッセージBのようにクロスバの入力に同時に到着して，かつ同じ出力へ同時に向かう場合を指す．このことを考慮するかしないかによって解析手法が変わる．PUのメッセージ発生確率が小さい場合はメッセージの同時到着の確率も小さいので，そのような条件下では同時到着の効果を無視することができる．しかし，現実的な解析を行うためにはメッセージの同時到着を考慮しなければならない．メッセージの同時到着を考慮しない場合は，考慮する場合よりも解析が簡単になるので，まずその手法について述べ，さらにそれを同時到着を考慮した手法に拡張する．

2.2 メッセージの同時到着を考慮しない解析

ここでは，メッセージの同時到着はない，という制約の下での理論解析について述べる．同時到着でないのであれば，複数のメッセージが同じ出力へ向かってもよい．この場合，後から到着したメッセージは先に到着したメッセージによってブロックされる．

MINは複数のクロスバ・スイッチから構成されているため，MINの解析を行う前に単体クロスバ・スイッチから成るCrossbar Network（入出力数 n）の転送性能を考える．

2.2.1 Crossbar Networkの転送性能

システムに固定的に与えられるパラメータは，以下のとおりである．

- n ··· クロスバ・スイッチのサイズ $(n \times n)$
- l ··· メッセージ長
- λ ··· PUのメッセージ発生率

また，解析で用いる中間変数は以下のとおりである．

- μ ··· メッセージがバッファから出ていく確率
- ρ ··· クロスバの入力バッファの利用率
- i ··· 注目しているメッセージと衝突するメッセージの数

解析の結果，転送性能として求められる変数は，以下のとおりである．

- w ··· ネットワーク中でのメッセージの待ち時間の期待値
- t ··· スループット

以下に手順を述べる．1つのメッセージがバッファを通過するために必要な時間の平均値は $l+w$ なので，ある時刻において，メッセージがバッファから出ていく確率 μ は

$$\mu = \frac{1}{l+w}$$

で求められる．これは，待ち行列理論におけるサービス率に相当する．

ネットワーク中のバッファは，待ち行列理論における $M/D/1$ モデル[11]に相当するので，バッファの利用率 ρ（バッファのビジー率）は

$$\rho = \frac{\lambda}{\mu}$$

で求められる．以上より，Crossbar Networkにおけるメッセージの待ち時間，スループットは以下のようにして求めることができる．

Crossbar Networkのバッファの利用率は

$$\rho = \begin{cases} \lambda \times (l+w) & \textbf{if } \lambda \times (l+w) \leq 1 \\ 1 & \textbf{if } \lambda \times (l+w) > 1 \end{cases} \quad (1)$$

である．$\lambda(l+w)$ を計算して1以上になる場合がある．これは，バッファにメッセージが入ることができずにPU側に溜まっている状態である．このようなときにはバッファにはつねにメッセージが存在する状態なので，$\rho = 1$ と補正する．

メッセージのネットワーク中での待ち時間の期待値 w は，次のように考えて求める．ある注目しているメッセージがクロスバの入力バッファに到着したときに，クロスバの同じ出力へ向かおうとしているすでに到着しているメッセージが i 個あるとする．クロスバの入力は n 本あるから，i の最大値は $n-1$，すなわち $0 \leq i \leq n-1$ である．i 個のうち1つは現在転送中である．現在転送中のメッセージによって待たされる時間の期待値は $(l+1)/2$ である．なぜならば，転送されずに残っているメッセージ長は，1 flit または 2 flit，…，l flit のいずれかであってかつどの場合が起こる確率も等しいので，その平均値が待ち時間の期待値となる．また，メッセージの衝突は FCFS priority で調停されるため，現在転送中でないメッセージによって待たされる時間は，$i-1$ 個のメッセージが通過に必要な時間 $l(i-1)$ となる．したがって，i 個のメッセージによって待たされる時間は

$$l(i-1)+\frac{(l+1)}{2}$$

となる．以上より，i 個のメッセージが衝突する確率と i 個のメッセージによって待たされる時間の期待値を掛け合わせたものを，すべての i について足し合わせることによって w を求めることができる．

注目しているメッセージが他の1つのメッセージと衝突する確率は ρ/n なので，それが i 個のメッセージと衝突する確率は

$$\binom{n-1}{i}\left(\frac{\rho}{n}\right)^i\left(1-\frac{\rho}{n}\right)^{n-1-i} \tag{2}$$

である．

したがって，

$$w=\sum_{i=1}^{n-1}\binom{n-1}{i}\left(\frac{\rho}{n}\right)^i\left(1-\frac{\rho}{n}\right)^{n-1-i} \times\left\{(i-1)l+\frac{1}{2}(l+1)\right\} \tag{3}$$

となる．i が0の場合は，待ち時間がゼロになるので，$i=0$ の場合の待ち時間を足し合わせる必要がない．よって，式(3)は $i \geq 1$ の場合のみを考慮している．

式(3)はさらに簡単化でき，式(4)のようになる（証明は付録の証明1に示す）．

$$w=l\left[\frac{n-1}{n}\rho-\left\{1-\left(1-\frac{\rho}{n}\right)^{n-1}\right\}\right] +\frac{l+1}{2}\left\{1-\left(1-\frac{\rho}{n}\right)^{n-1}\right\} \tag{4}$$

以上のように w が求まるが，この式には確率変数 ρ が含まれている．ρ は式(1)において，w より求まるため（それ以外の変数はすべて入力パラメータとして与えられている），両者の定常値を以下のステップ2からステップ3を繰り返して求める．

ステップ 1 w に適当な初期値を代入する．

ステップ 2 その w を使って ρ を計算する．

ステップ 3 その ρ を使って新たな w を計算する．

そして求められた w を使って，スループットを計算する．スループットは，クロスバの入力バッファにメッセージが存在する確率 ρ に，実際に出力から出てくる割合を掛け合わせたものである．実際に出力から出てくる割合は，無衝突で出てくるという理想的な場合の転送時間 l を，実際の転送時間 $l+w$ で割った値である．よって，

$$t=\frac{l}{l+w}\times\rho \tag{5}$$

となる．

2.2.2 MIN の転送性能

次に Crossbar Network の解析方法を MIN に拡張する．例として3 stage MIN の解析を示す．以下に新たに必要となる変数を示す．

- n_s … 第 s stage のクロスバ・サイズ
- ρ_s … 第 s stage のクロスバの入力バッファの利用率
- w_s … 第 s stage のクロスバでのメッセージの待ち時間の期待値

第1 stage クロスバの入力バッファの利用率は，Crossbar Network の解析と同様に考えて

$$\rho_1=\begin{cases}\lambda\times(l+w_1+w_2+w_3) & \textbf{if}\ \ \lambda\times(l+w_1+w_2+w_3)\leq 1\\ 1 & \textbf{if}\ \ \lambda\times(l+w_1+w_2+w_3)>1\end{cases} \tag{6}$$

となる．第1 stage での待ち時間も単一クロスバの解析とほぼ同様だが，第2 stage 以降で待たされる時間を考慮する必要がある．つまり，図4のメッセージ A が第1 stage で待たされる時間は，メッセージ B が第1 stage を通過するのに必要な l clock とメッセージ B が第2，3 stage で待たされる時間の和である．よって，見かけのメッセージ長を L_1 とすると，

$$L_1=l+w_2+w_3$$

である．したがって，第1 stage での待ち時間 w_1 は，式(4)を使って，

$$w_1=L_1\left[\frac{n_1-1}{n_1}\rho_1-\left\{1-\left(1-\frac{\rho_1}{n_1}\right)^{n_1-1}\right\}\right]$$

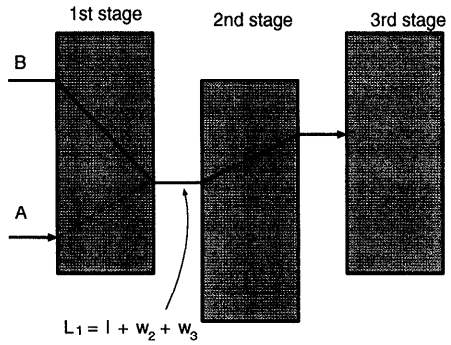


図4 見かけのメッセージ長
Fig. 4 Seeming message length.

$$+\frac{L_1+1}{2}\left\{1-\left(1-\frac{\rho_1}{n_1}\right)^{n_1-1}\right\} \quad (7)$$

となる．

次に第2 stageについて考える．1つのメッセージが第1 stageのバッファを通過するために必要な時間は，$l+w_1+w_2+w_3$ である．バッファにメッセージが存在する確率を，バッファを通過するために必要な時間で割ることによって，第2 stageへのメッセージ到着率を求めることができる．これは，$\rho_1/(l+w_1+w_2+w_3)$ となる．この値を使って，単一クロスバの ρ を求める式(1)と同様にして以下のように ρ_2 を求めることができる．

$$\rho_2 = \frac{\rho_1}{l+w_1+w_2+w_3} \times (l+w_2+w_3) \quad (8)$$

w_2 についても式(7)と同様に求めることができる．見かけのメッセージ長 L_2 は

$$L_2 = l + w_3$$

なので，第2 stageでの待ち時間 w_2 は

$$w_2 = L_2\left[\frac{n_2-1}{n_2}\rho_2 - \left\{1-\left(1-\frac{\rho_2}{n_2}\right)^{n_2-1}\right\}\right] + \frac{L_2+1}{2}\left\{1-\left(1-\frac{\rho_2}{n_2}\right)^{n_2-1}\right\} \quad (9)$$

となる．以下，同様にして ρ_3, w_3 も計算できる．

w_1，w_2，w_3 を求めるためには，以下のステップ2からステップ7を定常解が得られるまで繰り返し，計算する．

ステップ1 w_1，w_2，w_3 に適当な初期値を代入する．
ステップ2 その w_1 を使って ρ_1 を計算する．
ステップ3 その ρ_1 を使って新たな w_1 を計算する．
ステップ4 その w_2 を使って ρ_2 を計算する．
ステップ5 その ρ_2 を使って新たな w_2 を計算する．
ステップ6 その w_3 を使って ρ_3 を計算する．
ステップ7 その ρ_3 を使って新たな w_3 を計算する．

最終的に，3 stage MINの総合的なスループット t は，式(5)と同様にして

$$t = \frac{l}{l+w_3} \times \rho_3 \quad (10)$$

として求められる．

2.3 メッセージの同時到着を考慮した解析

2.2節で示した手法ではメッセージの同時到着が考慮されていない．そこで，さらに厳密な解析のために，その効果を反映させるよう，2.2節の手法を拡張する．メッセージの同時到着の効果を反映するために，以下のようにメッセージの調停方法を改良した．

- 同時到着でないメッセージどうしは，FCFS priorityで調停される．
- 同時到着であるメッセージどうしは，random priorityで調停される．

2.2節と同様に，まずCrossbar Networkの理論解析について述べ，次にMINについて述べる．

2.3.1 Crossbar Networkの転送性能

新たに必要となる中間変数は，以下のとおりである．

- a ··· クロスバ・スイッチの入力バッファへのメッセージ到着率
- j ··· 今，注目しているメッセージと衝突する i 個のメッセージのうち，注目しているメッセージと同時刻に到着したメッセージの数

バッファの利用率 ρ は，同時到着を考えない場合とまったく同じようにして

$$\rho = \begin{cases} \lambda \times (l+w) & \textbf{if } \lambda \times (l+w) \le 1 \\ 1 & \textbf{if } \lambda \times (l+w) > 1 \end{cases} \quad (11)$$

で求められる．

メッセージのネットワーク中での待ち時間の期待値 w は，次のように考えて求める．ある注目しているメッセージがクロスバの入力バッファに到着したときに，同じクロスバの出力へ向かおうとしているメッセージが i 個あるとする．i 個のうち j 個は同時に到着したメッセージで，$i-j$ 個はすでに到着しているメッセージである．さらに $i-j$ 個のうち1つは現在転送中である．同時に到着したメッセージはrandom priorityで調停されるため，それらにより待たされる時間の期待値は $jl/2$ である．また，すでに到着しているメッセージはFCFS priorityで調停されるため，それらにより待たされる時間は $l(i-j-1)+(l+1)/2$ である．よって i 個のメッセージが衝突する確率に i 個のメッセージによって待たされる時間の期待値を掛け合わせたものを足し合わせることによって，w を求めることができる．

注目しているメッセージが i 個のメッセージと衝突する確率は，式 (2) より

$$\binom{n-1}{i}\left(\frac{\rho}{n}\right)^{i}\left(1-\frac{\rho}{n}\right)^{n-1-i}$$

である．衝突するメッセージのうち注目しているメッセージと同時到着である確率は a/ρ である．なぜならば，これは，バッファにメッセージが存在するという条件（ρ）において，メッセージが同時到着するという，条件付確率として定めなければならないからである．したがって，i 個のメッセージのうち j 個のメッセージが同時に到着する確率は

$$\binom{i}{j}\left(\frac{a}{\rho}\right)^{j}\left(1-\frac{a}{\rho}\right)^{i-j} \tag{12}$$

である．ここでクロスバ・スイッチの入力バッファへのメッセージ到着率 a は，バッファにメッセージが存在する確率を通過に必要な時間で割れば得ることができる．すなわち

$$a=\frac{\rho}{l+w}$$

で計算できる．以上より

$$\begin{aligned} w=\sum_{i=1}^{n-1}\Bigg[&\binom{n-1}{i}\left(\frac{\rho}{n}\right)^{i}\left(1-\frac{\rho}{n}\right)^{n-1-i}\\ &\times\Bigg\{\sum_{j=0}^{i}\binom{i}{j}\left(\frac{a}{\rho}\right)^{j}\left(1-\frac{a}{\rho}\right)^{i-j}\\ &\times\left(f(i,j,l)+\frac{1}{2}jl\right)\Bigg\}\Bigg] \end{aligned}$$

$$f(i,j,l)=\begin{cases}0 & \text{for } i=j\\ (i-j-1)l+\frac{1}{2}(l+1) & \text{for } i\neq j\end{cases} \tag{13}$$

となる．式 (13) は，さらに簡単化でき，式 (14) のようになる（証明の基本方針は同時到着を考慮しない場合とほぼ同様なので省略する）．

$$\begin{aligned} w=l\Bigg[&\frac{n-1}{n}(\rho-a)\\ &-\left\{1-\left(1-\frac{\rho-a}{n}\right)^{n-1}\right\}\Bigg]\\ &+\frac{l+1}{2}\left\{1-\left(1-\frac{\rho-a}{n}\right)^{n-1}\right\}\\ &+\frac{n-1}{n}a\times\frac{l}{2} \end{aligned} \tag{14}$$

w を求めるためには，以下のステップ 2 からステップ 3 を定常解が得られるまで繰り返し，計算する．

ステップ 1 w に適当な初期値を代入する．

ステップ 2 その w を使って ρ，a を計算する．

ステップ 3 その ρ，a を使って新たな w を計算する．

そして求められた w を使って，スループットを計算する．式 (5) と同様にして

$$t=\frac{l}{l+w}\times\rho \tag{15}$$

となる．

2.3.2 MIN の転送性能

次に Crossbar Network の解析方法を MIN に拡張する．例として 3 stage MIN の解析を示す．新たに必要となる中間変数は，以下のとおりである．

- a_s … 第 s stage のクロスバの入力バッファへのメッセージ到着率

第 1 stage のクロスバの入力バッファの利用率は，同時到着を考慮しない解析とまったく同じで

$$\rho_1=\begin{cases}\lambda\times(l+w_1+w_2+w_3)\\ \quad\textbf{if}\ \ \lambda\times(l+w_1+w_2+w_3)\leq 1\\ 1\\ \quad\textbf{if}\ \ \lambda\times(l+w_1+w_2+w_3)>1\end{cases} \tag{16}$$

となる．

また，見かけのメッセージ長 L_1 は

$$L_1=l+w_2+w_3$$

だから，第 1 stage での待ち時間 w_1 は，式 (14) を使って，

$$\begin{aligned} w_1=L_1\Bigg[&\frac{n_1-1}{n_1}(\rho_1-a_1)\\ &-\left\{1-\left(1-\frac{\rho_1-a_1}{n_1}\right)^{n_1-1}\right\}\Bigg]\\ &+\frac{L_1+1}{2}\left\{1-\left(1-\frac{\rho_1-a_1}{n_1}\right)^{n_1-1}\right\}\\ &+\frac{n_1-1}{n_1}a_1\times\frac{L_1}{2} \end{aligned} \tag{17}$$

となる．第 2 stage についても，同時到着を考慮しない解析手法を同様にして拡張できる．

$$\rho_2=\frac{\rho_1}{l+w_1+w_2+w_3}\times(l+w_2+w_3) \tag{18}$$

$$\begin{aligned} L_2&=l+w_3\\ w_2&=L_2\Bigg[\frac{n_2-1}{n_2}(\rho_2-a_2)\\ &\quad-\left\{1-\left(1-\frac{\rho_2-a_2}{n_2}\right)^{n_2-1}\right\}\Bigg]\\ &\quad+\frac{L_2+1}{2}\left\{1-\left(1-\frac{\rho_2-a_2}{n_2}\right)^{n_2-1}\right\}\\ &\quad+\frac{n_2-1}{n_2}a_2\times\frac{L_2}{2} \end{aligned} \tag{19}$$

となる．以下，同様にして ρ_3，w_3 も計算できる．

w_1，w_2，w_3 を求めるためには，以下のステップ 2 からステップ 7 を定常解が得られるまで繰り返し，計算する．

ステップ 1 w_1，w_2，w_3 に適当な初期値を代入する．
ステップ 2 その w_1 を使って ρ_1，a_1 を計算する．
ステップ 3 その ρ_1，a_1 を使って新たな w_1 を計算する．
ステップ 4 その w_2 を使って ρ_2，a_2 を計算する．
ステップ 5 その ρ_2，a_2 を使って新たな w_2 を計算する．
ステップ 6 その w_3 を使って ρ_3，a_3 を計算する．
ステップ 7 その ρ_3，a_3 を使って新たな w_3 を計算する．

最終的に，3 stage MIN のスループット t は

$$t = \frac{l}{l + w_3} \times \rho_3 \tag{20}$$

となる．

3. 結果と考察

本章では，2 章に示した解析手法の有効性を検証するため，いくつかのネットワークの例に対し，計算機シミュレーションによる結果と比較し検討する．

シミュレーションは，乱数のシードを変えて，10 回行った．グラフ中のシミュレーションの値は，その 10 個のデータの平均値を表している．エラーバーは，信頼区間 95%を表す．凡例の analysis 1 は同時到着を考慮しない解析によって得られた値を，analysis 2 はこれを考慮した解析によって得られた値を，それぞれ表す．横軸は PU のメッセージ発生確率 λ で縦軸はネットワーク中での待ち時間 w を表す．グラフは特徴の現れているもののみとした．本来，待ち時間とスループットの両者のグラフを載せるべきであるが，飽和点等の特性は両者で非常に似ているため（たとえば図 8 と図 11），紙面の都合上，基本的に待ち時間のグラフを示し，参考のためスループットに関するグラフは**図 11**（クロスバ・サイズ = 16 × 16，メッセージ長 = 10 flit）のみを示す．

図 5 から**図 8** までは，Crossbar Network における待ち時間を，クロスバ・サイズが小さい場合（2 × 2）と大きい場合（16 × 16）と，メッセージ長が短い場合（1 flit）と長い場合（10 flit）について，計 4 通りの組合せの結果をそれぞれ示す．

3 stage MIN については，メッセージ長が 10 flit の場合について，各段を構成するクロスバ・サイズが 2 × 2 の場合を**図 9** に，クロスバ・サイズが 16 × 16 の場合を**図 10** に，それぞれ示す．図 9 の総 PU 数は 8 であるが，図 10 の総 PU 数は 4096 という大規模ネットワークを想定している．

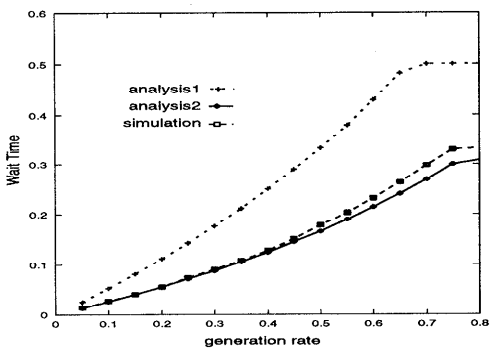


図 5 Crossbar Network における待ち時間（size = 2, length = 1 flit）
Fig. 5 waiting time on Crossbar Network (size = 2, length = 1 flit).

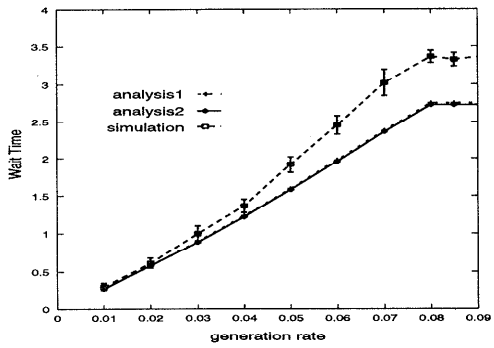


図 6 Crossbar Network における待ち時間（size = 2, length = 10 flit）
Fig. 6 waiting time on Crossbar Network (size = 2, length = 10 flit).

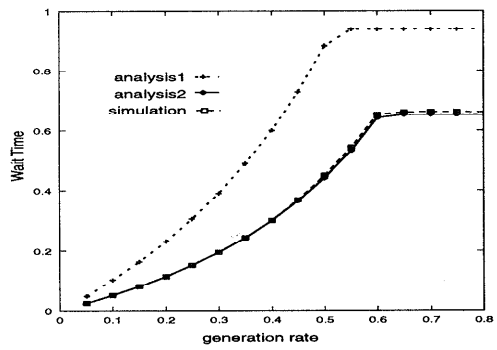


図 7 Crossbar Network における待ち時間（size = 16, length = 1 flit）
Fig. 7 waiting time on Crossbar Network (size = 16, length = 1 flit).

3.1 全体の傾向

analysis 1 は，メッセージ長が極端に短い場合

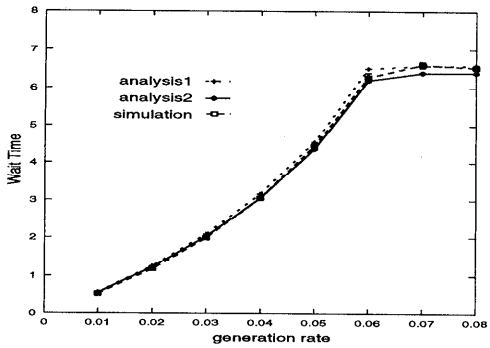


図8 Crossbar Networkにおける待ち時間（size = 16, length = 10 flit）

Fig. 8 waiting time on Crossbar Network(size = 16, length = 10 flit).

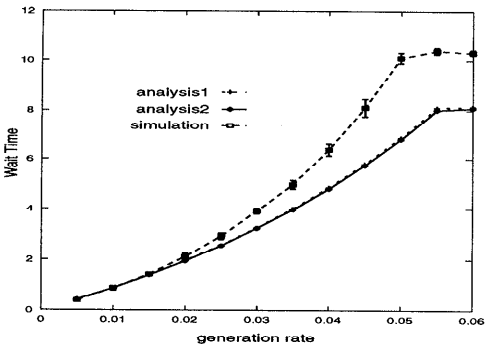


図9 3-stage MINにおける待ち時間（size = 2, length = 10 flit）

Fig. 9 waiting time on 3-stage MIN (size = 2, length = 10 flit).

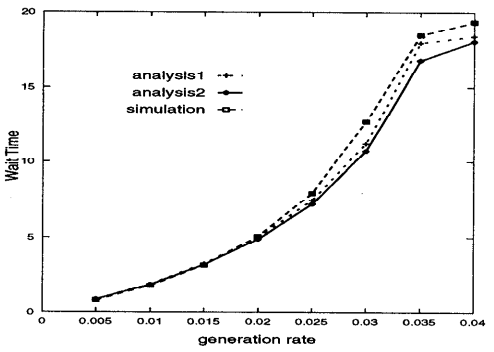


図10 3-stage MINにおける待ち時間（size = 16, length = 10 flit）

Fig. 10 waiting time on 3-stage MIN (size = 16, length = 10 flit).

（1 flit）では誤差が大きくなるが，analysis 2はそのような場合でも非常によく近似している．つまり，メッセージ長が非常に短い場合は，同時到着の影響が出ており，これを反映しているanalysis 2の方が正確であるということがいえる．analysis 2においても，クロスバ・サイズが小さい場合（size = 2）は，20% 前後の誤差が生じるが，クロスバ・サイズが比較的大きい場合（size = 16）は，最大でも5%前後の誤差である．さらに，結合するステージ数が少ないほど誤差は少ない．

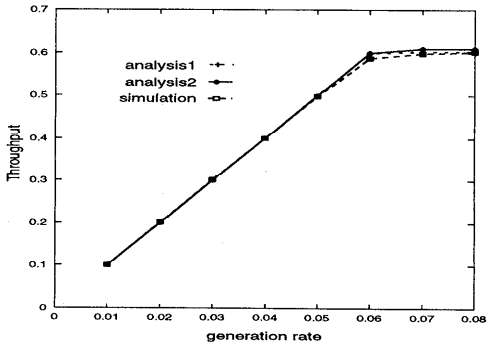


図11 Crossbar Networkにおけるスループット（size = 16, length = 10 flit）

Fig. 11 throughput on Crossbar Network (size = 16, length = 10 flit).

メッセージ長が長いほど，analysis 1とanalysis 2の差は小さくなる．この原因は以下のように考えられる．メッセージ長が長い場合，メッセージ発生確率が小さい領域で，ネットワークは飽和状態になってしまう．よって，グラフにはその領域のみが示されているが，メッセージ発生確率が小さいため，メッセージの同時到着確率もまた小さくなる．したがって，同時到着を考慮しているか否かの差がなくなるため，両解析手法の結果の差が小さくなっていると考えられる．

3.2 ネットワーク中での待ち時間とスループットの関係

一般に，PUのメッセージ発生確率が高くなりネットワークの負荷が大きくなれば待ち時間は大きくなる．よって，右上がりのグラフになる．スループットは，実際にネットワークから出てきたデータ量である．したがって，ネットワークが飽和していないならばネットワークの混み具合にかかわらず，スループットはネットワークに入ったデータ量に比例する．つまり，右上がりの原点を通る直線となる．

本研究で対象としている待ち時間は，PU内の待ち時間を含んでおらず，ネットワーク中のみの待ち時間なので，ネットワークが飽和に達すると待ち時間は一定になる．また，ネットワークが飽和に達すると，PUがいくらメッセージを発生してもネットワークに入ることができるメッセージ量は一定なので，スループットも一定になる．ネットワーク中での待ち時間の飽和点とスループットの飽和点は，理論解析値もシミュレー

ション値も一致していることがグラフから分かる．

3.3 有効性

Crossbar Networkでメッセージ長が1 flitの場合，analysis 1の誤差は無視できない．それ以外では，メッセージ長が長いほど，analysis 1とanalysis 2の差は少ない．したがって，メッセージ長が長く，メッセージの同時到着の確率が小さい場合では，同時到着を考慮に入れたより現実的な手法を用いる必要はないが，メッセージ長が短い場合は，同時到着を考慮に入れたより現実的な手法を用いることが有効である．

本論文で提案した手法は，並列計算機の解析に必要な4つの条件（クロスバ・サイズは任意であること，バッファ・サイズは有限であること，メッセージ長は1以上の有限であること，メッセージのブロッキングによる，時間依存性を考慮にいれること）を満たしている．他の解析手法では，これらの条件の一部しか満たしていない．したがって，本研究で提案した手法は，他の理論解析手法に比較すると，現実的な並列計算機の理論解析に有効である．

本手法と対照的な解析方法としては，マルコフ連鎖を用いる方法が考えられる．本手法では全PUにおけるパケット発生確率が同一であるとしているのに対し，マルコフ連鎖を用いた解析では，この条件を緩めることが，理論的には可能である．しかし，マルコフ連鎖を用いる解析ではシステム規模が大きくなった場合，状態数が爆発的に増加するため，近年の高速ワークステーションを使ったとしても，その適用範囲には限界がある．また，実際に各PUごとのパケット発生確率を変化させようとすると，その分だけ状態数が増え計算量がさらに膨大になり，現実的には難しいと考えられる．これに対し，本論文で提案している手法は，使用されているクロスバのサイズが大きくなっても計算コストはまったく増えず，またさらに多段のネットワークを想定した場合でも，基本的に3段の場合の手法の延長で容易に実現可能である．また，演算量はマルコフ連鎖に比べはるかに少なくて済み，数万以上のノードを想定した超並列計算機への適用も非常に容易であると考えられる．

また，一般に超並列計算機のネットワークを構成するクロスバ・スイッチは中規模のスイッチが使用され，かつオーバヘッドを少なくするために，ある程度の長さのメッセージが用いられる．さらに，高い転送性能を実現するためにステージをなるべく少なくする．このことから，これらのネットワークの評価を行うためには，本研究で提案した手法は，非常に有効な手段であるといえる．

さらに，本論文で提案している，クロスバ・スイッチにおけるメッセージのWormhole的振舞いの解析方法は，MINのような等距離間接網だけでなく，ある程度の規模のクロスバ・スイッチを規則的に用いている不当距離間接網，たとえばハイパクロスバ網[12]や，MDX網[13]にも適用可能であると考えられる．

クロスバ・サイズが小さい場合には，理論解析値とシミュレーション値には小さな誤差が見られる．これは，理論解析またはシミュレーションのどちらかに厳密性に欠けている点があるためと思われる．誤差の原因が何処にあるのかははっきりと分かっていない．しかし，原因の1つとして，シミュレーションは離散値を対象としており，理論解析は連続値を対象としていることにあると考えられる．これらの原因の究明は，今後の課題の1つである．

4. おわりに

本研究では，より現実的なモデルにおけるMINの転送性能の理論解析手法として，確率モデルに基づく手法を提案した．提案した手法には，メッセージの同時到着を考慮しない解析手法と同時到着を考慮する解析手法の2つがあり，同時到着を考慮した手法によって現実的な並列計算機の解析を行うことができた．さらに，理論解析によって求められた値と同時到着を考慮した理論解析と同じ仮定の下でのシミュレーションによって求められた値を比較することによって，次のような結果を得た．

- ネットワークを構成するクロスバ・サイズがある程度の大きさであれば，理論解析値はシミュレーション値に大変よく近似する．
- メッセージ長が長いほど，同時到着の影響は少なくなる．メッセージ長が短い場合は，より現実的な解析である同時到着を考慮した手法はより有効である．
- 非常に小規模のクロスバ・スイッチを用いた場合については，本手法による解析結果と実験値の誤差はいまだ十分小さいものとはいえない．しかし，一般に超並列計算機にはある程度の大きさのスイッチを使用し，かつ，なるべくステージ数を少なくするため，これらのネットワークの解析には十分有効である．

本研究の今後の課題としては以下のようなことがあげられる．

- クロスバ・サイズが小さい場合に現れる誤差を究明し，さらに誤差の少ない手法を提案する．
- クロスバ・スイッチから構成される，MIN以外の

ネットワークの解析に拡張する．

これらの課題を解決することによって，ネットワークの転送性能に対し，より完全な理論解析評価を与えることが，今後の超並列計算機のネットワークの実現に大きく役立つことを期待する．

謝辞 本研究を進めるにあたり貴重なご意見をいただいた筑波大学西川博昭助教授ならびにアーキテクチャ研究室諸氏に深く感謝します．なお，本研究の一部は文部省科学研究費補助（奨励（A）09780234）によるものである．

参考文献

1) Patel, J.H.: Performance of processor-memory interconnections for multiprocessors, *IEEE Trans. Comput.*, Vol.C-30, No.10, pp.771–780 (1981).
2) Wu, C., et al.: Performance Analysis of Multistage Interconnection Network Configurations and Operations, *IEEE Trans. Comput.*, Vol.41, No.1, pp.18–27 (1992).
3) Hsiao, S., et al.: Performance Evaluation of Circuit Switched Multistage Interconnection Networks Using a Hold Strategy, *IEEE Trans. Comput.*, Vol.3, No.5, pp.632–640 (1992).
4) Yoon, H., et al.: Performance analysis of multibuffered packet-switching networks in multiprocessor systems, *IEEE Trans. Comput.*, Vol.39, No.3, pp.319–327 (1990).
5) Mun, Y., et al.: Performance analysis of finite buffered multistage interconnection networks, *IEEE Trans. Comput.*, Vol.43, No.2, pp.153–162 (1994).
6) Youn, H.Y., et al.: On Multistage Interconnection Networks with Small Clock Cycles, *IEEE Trans. Parallel and Disributed Systems*, Vol.6, No.1, pp.86–93 (1995).
7) Lin, T., et al.: Performance Analysis of Finite-Buffered Multistage Interconnection Networks with a General Traffic Pattern, *ACM SIGMETRICS*, Vol.19, No.1, pp.68–78 (1991).
8) Kim, J., et al.: Hypercube Communication Delay with Wormhole Routing, *IEEE Trans. Comput.*, Vol.43, No.7, pp.806–814 (1994).
9) Dally, W.J.: Performance Analysis of k-ary n-cube Interconnection Networks, *IEEE Trans. Comput.*, Vol.39, No.6, pp.775–785 (1990).
10) Dally, W.J.: Virtual-Channel Flow Control, *IEEE Trans. Parallel and Distributed Systems*, Vol.3, No.2, pp.194–204 (1992).
11) Kleinrock, L.: *Queueing Systems*, A Wiley-Interscience (1976).
12) 朴 泰祐ほか：ハイパクロスバ網における適応ルーチングの導入とその評価，電子情報通信学会論文誌，Vol.J78-D-I, No.2 (1995).
13) Murata, A., Boku, T. and Amano, H.: The MDX (Multi-Dimensional X'bar): A Class of Networks for Large Scale Multiprocessors, *IEICE Trans. Information and Systems*, Vol.E79-D, No.8 (1996).

付　録

A.1 証　明　1

まず，式(21)と式(22)に分けて，さらに式(21)を式(21a)と式(21b)に分けて考える．

$$w_a = \sum_{i=1}^{n-1} \binom{n-1}{i} \left(\frac{\rho}{n}\right)^i \left(1-\frac{\rho}{n}\right)^{n-1-i} \times (i-1)l \tag{21}$$

$$w_b = \sum_{i=1}^{n-1} \binom{n-1}{i} \left(\frac{\rho}{n}\right)^i \left(1-\frac{\rho}{n}\right)^{n-1-i} \times \frac{1}{2}(l+1) \tag{22}$$

$$w_{a1} = \sum_{i=1}^{n-1} \binom{n-1}{i} \left(\frac{\rho}{n}\right)^i \left(1-\frac{\rho}{n}\right)^{n-1-i} \times il \tag{21a}$$

$$w_{a2} = \sum_{i=1}^{n-1} \binom{n-1}{i} \left(\frac{\rho}{n}\right)^i \left(1-\frac{\rho}{n}\right)^{n-1-i} \times (-l) \tag{21b}$$

式(21a)において，$m = n-1$ とおくと

$$\begin{aligned} w_{a1} &= l \times \sum_{i=1}^{m} \binom{m}{i} \left(\frac{\rho}{n}\right)^i \left(1-\frac{\rho}{n}\right)^{m-i} \times i \\ &= l \times \sum_{i=1}^{m} \frac{m!}{i!(m-i)!} \left(\frac{\rho}{n}\right)^i \left(1-\frac{\rho}{n}\right)^{m-i} \times i \\ &= l \times m \times \left(\frac{\rho}{n}\right) \\ &\quad \times \sum_{i=1}^{m} \frac{(m-1)!}{(i-1)!(m-i)!} \left(\frac{\rho}{n}\right)^{i-1} \left(1-\frac{\rho}{n}\right)^{m-i} \\ &= l \times m \times \left(\frac{\rho}{n}\right) \\ &\quad \times \sum_{i=1}^{m} \binom{m-1}{i-1} \left(\frac{\rho}{n}\right)^{i-1} \left(1-\frac{\rho}{n}\right)^{m-i} \end{aligned}$$

となる．ここで，

$$\sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i}=1$$

であることから

$$w_{a1}=l\times m\times\left(\frac{\rho}{n}\right)=\frac{n-1}{n}l\rho \tag{23}$$

となる．

次に，式(21b)について説明する．

$$\sum_{i=0}^{n-1}\binom{n-1}{i}\left(\frac{\rho}{n}\right)^i\left(1-\frac{\rho}{n}\right)^{n-1-i}=1$$

だから，

$$w_{a2}=-l\times\left\{1-\left(1-\frac{\rho}{n}\right)^{n-1}\right\} \tag{24}$$

となる．

最後に，式(22)について説明する．やはり，

$$\sum_{i=0}^{n-1}\binom{n-1}{i}\left(\frac{\rho}{n}\right)^i\left(1-\frac{\rho}{n}\right)^{n-1-i}=1$$

という関係式を使って，

$$w_b=\frac{1}{2}(l+1)\left\{1-\left(1-\frac{\rho}{n}\right)^{n-1}\right\} \tag{25}$$

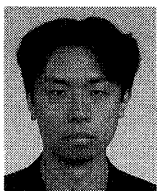
となる．

以上より，

$$\begin{aligned}w&=w_{a1}+w_{a2}+w_b\\&=l\left[\frac{n-1}{n}\rho-\left\{1-\left(1-\frac{\rho}{n}\right)^{n-1}\right\}\right]\\&\quad+\frac{l+1}{2}\left\{1-\left(1-\frac{\rho}{n}\right)^{n-1}\right\}\end{aligned} \tag{26}$$

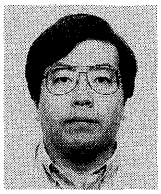
となる． □

(平成10年9月1日受付)
(平成11年1月8日採録)



三島 健（正会員）

平成6年筑波大学第三学群情報学類卒業．平成8年同大学院工学研究科電子・情報工学専攻修士課程修了．同年NTTネットワークサービスシステム研究所入社．マルチプロセッサ構成による次世代交換ノードシステムの研究，開発に従事．

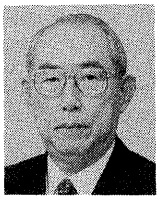


朴 泰祐（正会員）

昭和59年慶應義塾大学工学部電気工学科卒業．平成2年同大学院理工学研究科電気工学専攻後期博士課程修了．工学博士．昭和63年慶應義塾大学理工学部物理学科助手．平成4年筑波大学電子・情報工学系講師，平成7年同助教授，現在に至る．超並列処理ネットワーク，超並列計算機アーキテクチャ，ハイパフォーマンスコンピューティング，並列処理システム性能評価の研究に従事．電子情報通信学会，IEEE各会員．

中村 宏（正会員）

昭和60年東京大学工学部電子工学科卒業．平成2年同大学院工学系研究科電気工学専攻博士課程修了．工学博士．同年筑波大学電子・情報工学系助手．同講師，同助教授を経て，平成8年より東京大学先端科学技術研究センター助教授．計算機アーキテクチャ，ハイパフォーマンスコンピューティング，計算機の上位レベル設計支援，非同期式計算システムの研究に従事．本会平成5年度論文賞，平成6年度山下記念研究賞各受賞．電子情報通信学会，IEEE，ACM各会員．



中澤喜三郎（正会員）

昭和30年東京大学工学部応用物理卒業．昭和35年同大学院数物系博士課程応用物理修了．同年日立製作所入社．TAC，HITAC 5020，E/F，8800/8700，M-200H/280H，680H，S-810等，超大型コンピュータ・スーパコンピュータの開発に従事．平成元年より筑波大学電子・情報工学系教授，計算物理学センター向きの超並列処理システムCP-PACSの研究に従事．平成8年より電気通信大学情報工学科教授，平成10年より明星大学情報学部教授．工学博士．電子情報処理学会，IEEE，ACM各会員．平成5年度本学会論文賞，平成8年度本学会功績賞各受賞．