MIN 結合型マルチプロセッサ用トレースドリブンシミュレータ

米田 卓司 †　亀井 貴之 †　安川 英樹 †　塙 敏博 †
寺澤 卓也 ‡　天野 英晴 †

† 慶應義塾大学 理工学部
‡ 東京工科大学

多段結合網 (Multistage Interconnection Network: MIN) は多数のメモリを同時にアクセス可能なことから、数百プロセッサ程度の並列計算機のプロセッサ ·メモリ間結合として将来広く利用されることが期待されている。MIN 結合型マルチプロセッサの評価は、確率モデルを用いて行なう場合が多いが、MIN の転送容量や通過時間はそのアクセスパターンにより大きく変化することが知られており、実際の並列アプリケーションプログラムに基づく評価環境が必要である。

SSS 型 MIN 用トレースドリブンシミュレータ SIPSS(SSS-mIn multiProcessor System Simulator) は、実機あるいは他の命令レベルシミュレータのトレースデータに基づき MIN 結合型マルチプロセッサの性能評価を行なうことのできるシミュレータである。シミュレーションの高速化、シミュレータ自体の簡単化のため、シミュレーション対象は我々の提案している SSS 型 MIN に限定しているが、結合トポロジ、優先度制御、コンバインの有無等を自由に変えることができる。

並列計算機, 多段結合網, トレースドリブンシミュレータ,SIPSS

A Trace Driven Simulator for Multiprocessors with Multistage Interconection Network(MIN)

T. Komeda†　T. Kamei†　H. Yasukawa†　T. Hanawa†
T. Terasawa‡　H. Amano†

† Keio University
‡ Tokyo Engineering University

Multistage Interconnection Network (MIN) has been researched for connecting processors and shared memory modules in a middle scale multiprocessor with hundreds processing units. Although the latency and bandwidth of the MIN is sensitive to the access pattern, most of researches on performance evaluation of the MIN connected multiprocessors use probablistic models. For precise analysis, a realistic simulation environment based on the practical parallel parallel applications.

The trace driven simulator SIPSS (SSS-mIn multiProcessor System Simulator) can simulate the MIN connected multiprocessor using the address trace from real machines or other execution driven simulator. While the control model of the target MIN is limited as the SSS(Simple Serial Synchronized)-MIN for high speed execution, topology of the MIN, priority control, message combining, and other issues of the MIN can be selectable.

Multiprocessor,Multistage Interconnection Network,Trace Driven Simulator, SIPSS

1　はじめに

現在、小規模マルチプロセッサで最も一般的に用いられている共有バスは、接続可能なプロセッサ数の制限が厳しく、大規模化が困難である。これに対し、多段結合網(Multistage Interconnection Network: MIN)は多数のメモリを同時にアクセス可能なことから、数百プロセッサ程度の並列計算機のプロセッサ・メモリ間結合として将来にわたって広く利用されることが期待されている[1][2]。

我々は、高性能低コストなMINの構成法として、パケットを数bitにシリアライズし、同期入力にする事によりスイッチングエレメントを単純化し、ハードウェアコストを減少させると共に高速化するSSS(Simple Serial Synchronized)型MINを提唱している。SSS型MINでは、MIN内での衝突が性能に及ぼす影響が大きいため、高い通過率を持つ結合トポロジが必要になる。そこで我々は複数の出力を持つことにより高い通過率を実現するTBSF(Tandem Banyan Switching Fabrics)の利用、及びPBSF(Piled Banyan Switching Fabrics)の提案を行なってきた。このうちTBSFについてはネットワークチップとともに並列計算機SNAILの実装を行い、評価を行って来た[3]。この結果を踏まえ、さらに通過率の高い結合網やキャッシュ制御機構を持つシステムの設計を開始している。

MINの評価を行なう場合、その転送容量や通過時間はそのアクセスパターンにより大きく変化することが知られている。したがって、実際的な評価を行なうためには、実際の並列アプリケーションプログラムに基づく評価環境が必要である。それにもかかわらず、MIN結合型マルチプロセッサの評価は、確率モデルを用いて行なう場合が多かった。これは、以下の原因によるものと考えられる。

- スイッチングエレメントのシミュレーションを詳細に行なうと、シミュレーション時間が膨大になること
- MIN結合マルチプロセッサのトレースデータがほとんど存在しないこと

SSS型MINにおいては、スイッチングエレメントの制御が簡単であるため、従来型MINに比べシミュレータの構造は簡単にすることができ、実行速度を大きく改善することができる。また、トレースデータに関しては現在稼働中のSNAILから採集するシステムを構築中であり、アクセスアドレスのみならば、現在稼働中のバス結合型システム用命令レベルシミュレータMILL[4]を利用することにより、かなり現実的なシミュレーションが可能である。

そこで、我々はSSS-MIN結合型マルチプロセッサ用トレースドリブンシミュレータ、SIPSS(SSS MIN Multiprocessor System Simulator)を作成し、MINの接続トポロジ、優先制御、コンバインの効果等に関する評価を行っている。本論文では、SIPSSで対象とするSSS型MINに関して解説した後、SIPSSの構造と評価例について報告する。

2　MIN

2.1　SSS型MIN

通常のMINはOmega網を代表とするブロッキング網で構成されており、スイッチングエレメント内で衝突が起きた場合にはエレメント内のバッファに格納され、片方のパケットが通過した後、もう一方のパケットが出力される。このようなMINでは、スイッチングエレメント単位でハンドシェークとバッファリングを行なうため、ハードウェア構成が複雑になると共に、エレメント間の結合リンク数が8~64bitになるため、ピンネックの問題が生じる。また、複数のプロセッサ間で同時に同一メモリへの参照(HotSpot)が集中するとTress Saturationと呼ばれる現象によりネットワーク全体が飽和し、性能が大幅に低下する。これを回避する方法として、Message Combining[6]に代表される様々な解決方法が提案されているが、この機能を実現するためには多くのハードウェア量が必要となる。

そこで、従来型MINの問題点を解決するため、我々は単純なエレメントから構成され、高い通過率を持つ結合網により、1-4bit単位で同期入力されたパケットを交換する方式、SSS(Simple Serial Synchronized)型MINを提案している[5]。この方式では、プロセッサからのアクセスはMINとの間のバッファで1-4bit程度にシリアル化され、B-ISDN等で用いられているATM用交換システム同様、フレームクロックと呼ばれるクロックに同期してネットワークに入力される。さらにスイッチングエレメント内にはパケットバッファは設けられていないため、各エレメントは1bit分の記憶しか行わない。

ネットワーク内ではフレームクロックの立上り数クロック後1クロック毎に一段ずつスイッチの状態が決定され、

段数分のクロック分の遅延の後にネットワーク全スイッチの状態が決定し、ネットワークの入力から出力への論理的なパス (以下トレースと呼ぶ) が設定され、パケットの先頭が出口より出力される。これと同時に、出力側で有効なパケットが到着したかを調べ、ACK(ACKnowledge) もしくは NAK(NotAcKnowledge) をプロセッサ側に返す。プロセッサ側が NAK を受け取った場合、次のフレームでパケットを再送しなくてはならない。また、同一宛先のパケットをエレメント内で結合し、1 つのパケットとして扱う Message Combining により、Hot Spot へのアクセス集中よる Tree Saturation を回避する。

SSS 型 MIN の実際のアプリケーションに基づく評価を行なうために、プロトタイプシステム SNAIL(SSS-Network Architecture ImpLementation) を開発し、現在 16 プロセッサ ·16 メモリの規模で稼働中である [7]。現在この SNAIL 上で SSS 型 MIN の様々な評価を行なっており、SSS 型 MIN が優れた性能を持つ事が示されている。一方、スターベション、転送能力に関しての問題が明らかになっている。これらの問題を解決するため、より転送能力の高い多重出力可能な MIN の研究が行われている。

2.2　多重出力可能な MIN

SSS 型 MIN は、優れた性能を持っているが、ネットワークが混雑すると転送能力が不足するため、従来の MIN よりも高い通過率が必要となる。一方、従来の MIN の通過率を改善するアプローチとして、出力を多重化することにより同じ行き先のパケットを複数通過させるという、多重出力可能な MIN の研究が行われ、以下のものが提案されている。

- **MBSF(Mulit Banyan Switching Fabrics)**
  独立の Banyan 網を複数用意し、トラフィックを分散する方法は、パケットの交換用、サーキットスイッチ用の交換網などに、古くから研究されてきた [8][9][10][11]。MBSF は、図 1のように、各 Banyan 網が全く独立しており、単に各網に対する負荷が網数分に分散される。

- **EBSF(Expanded Banyan Switching Fabrics)**
  Banyan 網のサイズを整数倍に拡張する方法。通過率に関しては、MBSF と同等の性能を持つと考えられる [12]。図 2に、入力数 8、banyan 網のサイズ 3 倍の EBSF を示す。

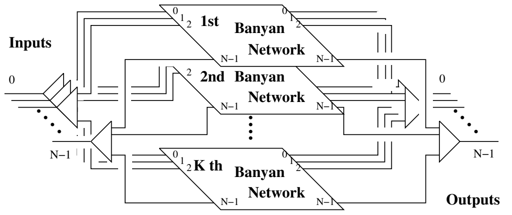

図 1: MBSF(Multi Banyan Switching Fabrics)

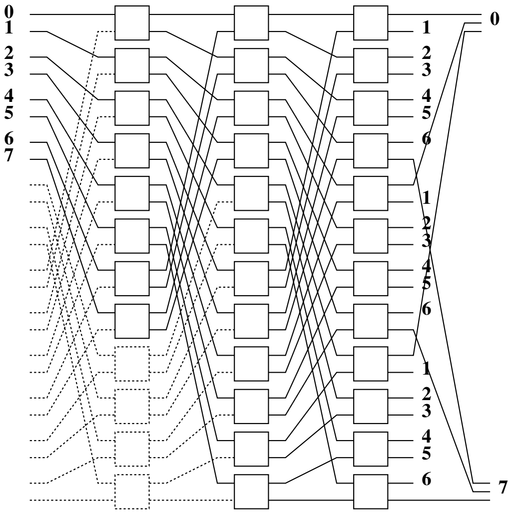

図 2: EBSF(Expanded Banyan Switching Fabrics)

- **TBSF(Tandem Banyan Switching Fabrics)**
  TBSF[13][14] は、図 3に示すように Banyan 網 (omega 網) を直列に接続し、各網の出口にバイパス路を設けた構造を持つ。正しくルーティングされたパケットは、バイパスを出るが、衝突により目的の宛先に到着できなかったパケットは次の段の Banyan 網に入力される。

- **PBSF(Piled Banyan Switching Fabrics)**
  PBSF[15] は、図 4に示すように Banyan(omega) 網を三次元的に接続した構造を持つ。スイッチングエレメントは、最上層と最下層を除いて水平方向の入出力が 2 つずつ、垂直方向を 2 つずつ、計 4 入力 4 出力をもつ。衝突によりルーティングに失敗したパケットは、下の層のネットワークに送られ、ルーティングが継続される。

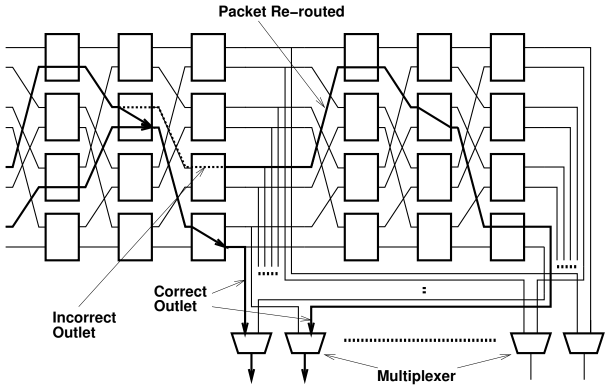

図 3: TBSF(Tandem Banyan Switching Fabrics)

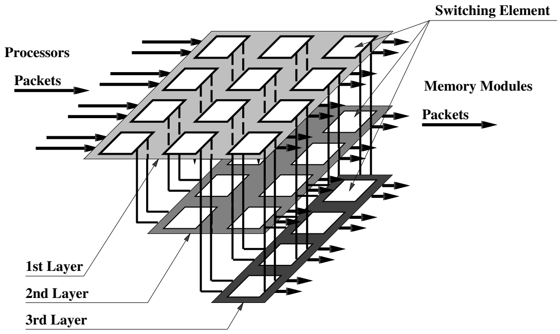

図 4: PBSF(Piled Banyan Switching Fabrics)

SIPSS を用いる事により、これらの様々な結合トポロジについて様々な評価を簡単に行なうことができる。

3 SIPSS

今回作成した SSS 型 MIN 用トレースドリブンシミュレータ SIPSS について説明する。

3.1 SIPSS の構成

SIPSS(SSS-mIn multiProcessor System Simulator) は、SSS 型 MIN で結合されたシステムを前提にしたトレースドリブンシミュレータである。

3.1.1 対象システムのモデル

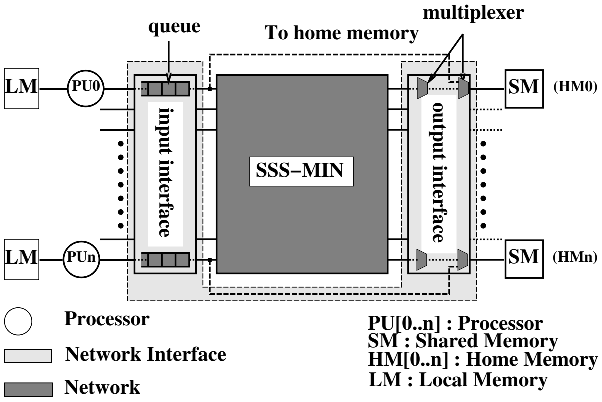

図 5: 対象システムモデル

SIPSS で評価できる対象を図 5に示す。SIPSS では、ホームメモリへのリンクを持たない形と、ホームメモリへのリンクを持った形の 2 種類に関して評価する事が出来る。ここで、ホームメモリと呼んでいるものは、各プロセッサに対応した共有メモリの事である。また、パケットは、ネットワークからの出力でマルチプレクスされ、更にホームメモリへのリンクからの出力とマルチプレクスされる。つまり、パケットは 2 回マルチプレクスされて出力される。

3.1.2 特徴

SIPSS の主な特徴は、

- ホームメモリへのリンクを持ったシステムと持たないシステムの二種類の評価が可能
- ネットワークトポロジの変更が可能
- ネットワーク周辺インターフェイスの設定変更が可能
- コンバイン機能の設定変更が可能
- パケット ·スイッチ ·アクセスの優先度制御が可能
- ネットワークへのアクセス負荷を変える事が可能
- 確率モデルでの評価も可能

である。次に、SIPSS の処理部分について説明する。SIPSS は、基本的にネットワーク部、インターフェイス部、プロセッサ部の 3 つの処理部分から構成されている。

3.1.3 ネットワーク部

ネットワーク部は、SSS 型の MBSF、EBSF、TBSF、PBSF の 4 種類のトポロジ変更をサポートしている。また、各層、各段におけるコンバイン機能の有無、各ステージスイッチの優先度制御に関する設定変更が可能である。

3.1.4 インターフェイス部

インターフェイス部は、入力インターフェイス部と出力インターフェイス部を持つ。入力インターフェイス部では、プロセッサからのメモリアクセスをフレームに同期して、ビットシリアル化して、入力キューに保持し、ネットワーク内に送出する。また、ACK(ACKnowladge) が帰ってこなかったパケットを再送する。入力キューは、初期状態として 3 つのアクセスまで保持が可能であるが、設定ファイルからのキューサイズの変更が可能となっている。

出力インターフェイス部は、ネットワークを通過してきたパケットをマルチプレクスして出力し、通過できたパケットのオーナー (プロセッサ) に対して、ACK を返す。また、マルチプレクサの出力数の変更とマルチプレクサ内でのコンバイン機能の有無に関する設定変更も可能である。

3.1.5 プロセッサ部

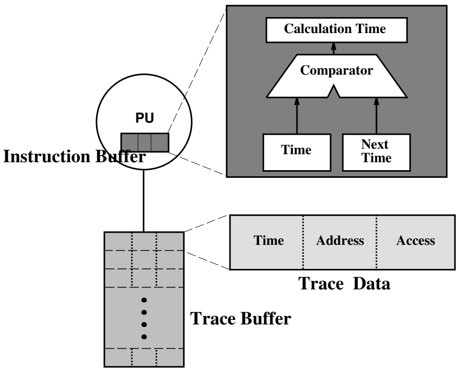

図 6: プロセッサ部

プロセッサ部には、図 6のように、各プロセッサ毎にトレースバッファが用意されており、トレースデータは、そのバッファ内に保持される。各プロセッサは、計算中に、次のトレースデータをトレース用命令バッファにフェッチし、次のアクセスまでの処理時間を計算する。

3.1.6 優先度制御

パケットに優先度を持たせ、スタベーションに関する評価をとることがが可能である。また、スイッチへのパケットの入力位置による優先度、アクセスの種類による優先度の設定変更が可能で、様々な状況下での評価をとることが出来る。

4 評価

今回は、残念ながら SNAIL から、トレースを直接とれなかったため、並列計算機シミュレータ MILL から取り出したトレースを用いて評価を行った。ターゲット・アプリケーションプログラムは、SPLASH[16] のうち MP3D を用い、PU16 台でシミュレートした。

4.1 並列計算機シミュレータ MILL

MILL(Multiprocesser Instruction Level simuLator) は、数十台規模のプロセッサシステム装置を持つバス結合型並列計算機を CPU のインストラクションレベルでシミュレートするシステムである。シミュレートする各 PU にはローカルメモリとキャシュメモリが備えられていることを想定している。MILL の主要な処理は、以下の 3 つである。

プロセッサ部： コンパイル及びリンケージされた、ターゲット・アプリケーションプログラムを読み込み、プログラムカウンタの指す番地からインストラクションの実行をシミュレートする。

キャシュ部： 共有メモリに対するアクセスが生じた時に呼ばれ、ブロックの read/write を行う。hit/miss の判定後、指定されたキャシュプロトコルに従って無効化やブロック転送を行い、共有メモリや他のプロセッサが持つキャッシュとの一貫性を保つ。今回は、キャッシュを持たない SSS 型 MIN システムの評価を行うため、MILL のキャッシュを使用しないように設定してトレースを採取した。

バス・共有メモリ： バスリクエストが生じた時に呼ばれ、IEEE Futurebus のプロトコルに従ったバスオペレーションをシミュレートする。バスリクエストの内容によっては共有メモリの read/write を行い、バス待ち時間や占有時間などの情報採取も行う。

4.2 システムパラメータ

以下に評価に用いたシステムパラメータを定義する。

L_c: クロック長 [sec]

L_f: フレーム長 [sec]

C_F: フレームクロック [clock]

N: ネットワークに入力されるパケット総数 [個]

A_p: 1 フレーム当たりの平均パケット数 [個/frame]

P_A: 通過率

N_A: Acknowladge が帰ってきたパケットの総数 [個]

t_s: シミュレータ上での実行時間 [sec]

t_f: フレーム数 [frame]

mo_1: マルチプレクサ 1 の出力数

mo_2: マルチプレクサ 2 の出力数

これを用いると、フレーム長は、クロック長の整数倍であるので以下のような式で表される。

$$L_f = n \times L_c \tag{1}$$

ただし、n は、フレームサイクル (C_F) と同値である。

$$n = C_F \tag{2}$$

また、フレーム数 t_fは、

$$\begin{aligned} t_f &= \frac{t_s/L_c}{C_F} \\ &= \frac{t_s \times L_f}{L_c^2} \end{aligned} \tag{3}$$

となる。以上の式より、1 フレーム当たりの平均パケット数 (ネットワーク負荷) A_Pは、

$$\begin{aligned} A_p &= N/t_f \\ &= \frac{N \times L_c^2}{t_s \times L_f} \end{aligned} \tag{4}$$

で表される。(4) 式より、クロック長、フレーム長が変化するとネットワーク負荷も変化する事がわかる。以下でクロック長、フレーム長を変化させた時の 4 つの項目に関して評価を行った。

4.3 各トポロジに対する通過率の変化

図 7は、フレーム長を変化させた時の各ネットワークトポロジの通過率の変化を表している。通過率は、次の式を用いて計算した。

$$P_A = \frac{N_A}{N} \tag{5}$$

設定は、PBSF、MBSF に関してはLayer 数を 2、TBSF、EBSF は Banyan 数を 2 とし、全てのトポロジにおいてコンバイン機能を付加させ、多重出力数を 1 としたホームメモリへのリンクを持たないシステムとした。図 7より、PBSF、TBSF、MBSF、EBSF、Omega 網の順番に性能が優れている事が分かる。また、ネットワークの負荷が低い時、全てのトポロジでほぼ同等の性能を示しているが、負荷が高くなると、ネットワーク内でのパケット同士の競合が増加するため、各トポロジ毎の性能差が顕著になっている。

4.4 コンバインによる影響

図 8は、Layer 数 4 の PBSF ネットワークを用いて、クロック長を変化させた時のコンバイン機能の有無による通過率の変化を表している。設定は、ホームメモリへのリンクを持たせ、多重出力数を 1 とした。図 8より、コンバイン機能を付加しない場合、通過率は約 20%位であるが、コンバインを付加する事により、通過率が約 100%近くになることがわかる。また、ホームメモリへのリンクを持ったシステムではホームメモリへ向かうパケットの優先度をネットワークを通過してきたパケットより高くした方が多少の性能向上が伺える。これは、同期アクセスが集中した場合、ホームメモリ内に同期アドレスを持ったプロセッサのアクセスが妨げられるためと考えられる。また、多重出力数を 2 にした場合、約 2 倍ほどの性能の向上がみられた。

4.5 アクセス優先度による影響

図 9は、Banyan 数 2 の TBSF ネットワークでホームメモリへのリンクを持ったシステムを用いて、アクセスの種類によって優先度を持たせた時の通過率の変化を表している。その他の設定は、通過率の差を顕著にするためコンバイン機能は無しとし、マルチプレクスの多重出力数を 1 とした。図より、同期アクセスの優先度を最も高くした方が、性能がよい。これはアクセスのうち同期アクセスが多数を締めているものと考えられる。つま

り、同期アクセスの優先度を下げた場合、ネットワーク内で同期アクセスがデッドパケットとなる確率が大きくなり、再送パケット数が増加し、通過率が下がる。

4.6　ネットワーク負荷の変化

図10は、Layer数4、多重出力数1のPBSFネットワークを用いて、クロック長を変化させた時の1フレーム当たりにネットワークに入る平均パケット数(A_p)の変化を表している。図10より、コンバインを付加した場合、1フレーム当たりネットワーク内に入力されているパケット数は、約3個で飽和している。一方、コンバインを付加していなければ、約8個で飽和している。これは、ネットワーク内で負けたパケットの再送数が大幅に増加するためである。

5　おわりに

シミュレーションは、小規模なシステム(16プロセッサ)で行ない、SS-20ワークステーション上でアクセス10万回当たり約3分の時間を要した。これより、大規模なシステムのシミュレーションに対しても現実的な時間でシミュレートすることができると判断できる。

本稿では、実装されたSSS-MIN型のトレースドリブンシミュレータの詳細について述べた。しかし、今回行った評価は、共有バス結合型の並列計算機シミュレータMILLから収集したトレースを用いた評価であったため、実機SNAILとの比較が行えなかった。今後、稼働中のSNAILからトレースを採集するシステムを構築し、評価を行っていきたい。また、SSS型に限定せず、汎用のMIN型シミュレータの設計を行なっていきたい。

参考文献

[1] J.Konicek, et al,: *The Organization of the Cedar System*, Proc. of 1991 Int. Conf. Parallel Processing, (1991).

[2] G.F.Phister, et al.: *The IBM Research Parallel Processor Prototype (RP3): Introduction and architecture*, Proc. of 1985 Int. Conf. Parallel Processing, (1985).

[3] 周洛, 天野英晴: SSS-MIN用プロトタイプチップの実装, 情報処理学会第44回全国大会論文集,6-103,104 (1992).

[4] T.Terasawa,H.Amano: *Performance Evaluation of the Mixed-protocol Caches with Instruction Level Multiprocessor testbed ATTEMPT-0*, Proc.of IASTED Int'l Conf.on Modeling and Simulation MS'94,pp.1-5,May, (1994)

[5] 天野, 周, 藤川: SSS (Simple Serial Synchronized)型マルチステージネットワーク, 情報処理学会論文誌 Vol.34, No.5, pp. 1134-1143, (1993).

[6] G.F.Phister and V.A.Norton: *Hot Spot Connection and Combining in Multistage Interconnection Networks*, IEEE Trans. on Comput. Vol.C-34, No.10, (1985)

[7] M.Sasahara, et al: *SNAIL:a multiprocessor based on the Simple Serial Synchronized multistage interconnection network architecture*, Proc. of International Conference on Parallel Processing, pp.I117-120, Aug. 1994

[8] Kruskal, C.P. and Snir, M.: *The performance of multistage interconnection networks for multiprocessors*, IEEE Trans. Comput. Vol.C-32, No.12,pp1091-1098,Dec.(1983)

[9] Kumar, M. and Jump, J.R: *Performance of unbuffered shuffle-exchange networks*, IEEE Trans. Comput. VolC-35, No.6,pp573-577,Jun.(1986)

[10] C.T.Lea : *Multi-Log$_2$N networks and their applications in high-speed electronic and photonic switching systems*, IEEE Trans. Comm. 38(10) pp.1740-1749(1990)

[11] C.L.Wu and M.Lee: *Performance Analysis of Multistage Interconnection Network Configurations and Operations*, IEEE Trans. Comput. 41, No.1 pp.18-27(1992)

[12] R.Awdeth and H.Mouftah: *The Expanded Delta Fast Packet Switch*, IEEE Int. Conf. Commun.(ICC),1994.

[13] 坂元, 荒川, 正木, 井上, 天野: 自己ルーティングスイッチの構成とその評価, 信学技報 ISSE88-308,(1988)

[14] Tobagi, F.A. Kwok, T.: *The Tandem Banyan Switching Fablics: a Simple High-Perfomance Fast Packet Switch*, Proc. of IEEE Custom Integrated Circuits Conference, (1991)

[15] 天野英晴, 藤川義文: マルチステージネットワーク PBSF, 情報処理学会計算機アーキテクチャ研究報告 No.94-5, (1992)

[16] Jaswinder Pal Singh, Wolf-Dietrich Weber and Anoop Gupta: *SPLASH : Stanford Parallel Applications for Shared-Memory*, Technical report

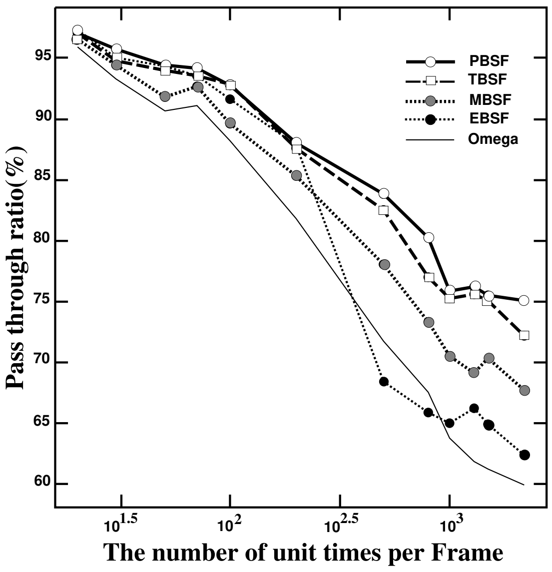

図 7: トポロジによる通過率の変化

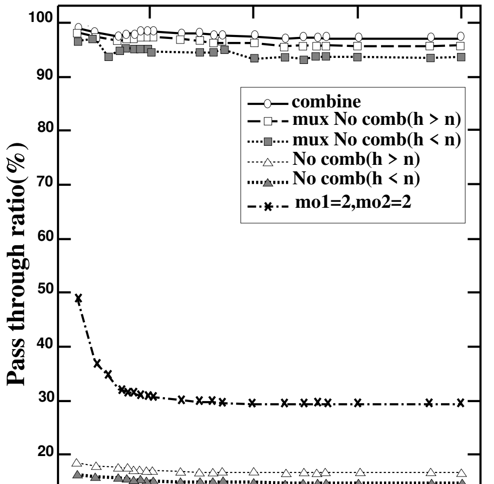

図 8: コンバインによる影響

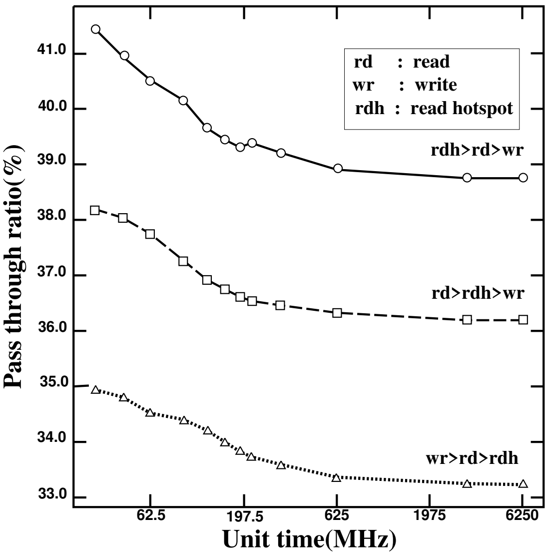

図 9: アクセスの優先度による影響

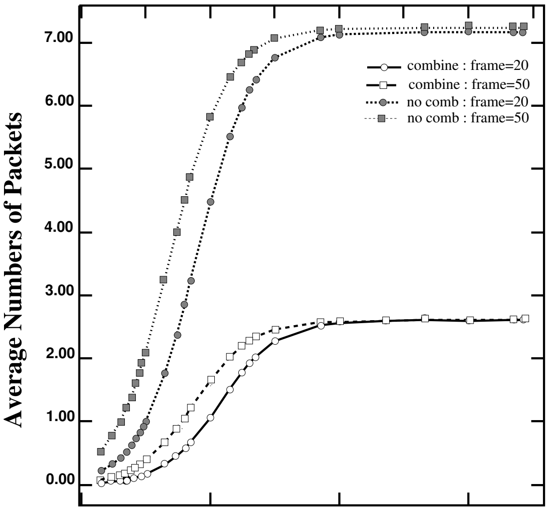

図 10: 負荷の変化