社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2006－ＡＶＭ－55（11）
2006／12／15

リニアスケーラブルなスイッチ実現に向けた ロードバランス型スイッチ方式の提案

西崎　秀樹†　山田　憲晋†

†日本電気株式会社システムプラットフォーム研究所　〒211-8666　神奈川県川崎市中原区下沼部 1753
E-mail:　†{nisizaki@da, kenshin@ap}.jp.nec.com

あらまし　年々増加し続けるアクセス及びバックボーンネットワークのトラヒックに対応するため、コアルータに代表される通信装置ではスイッチ容量の大容量化が必須である。本稿では、リニアスケーラブルなスイッチとしてロードバランス型スイッチに着目し、入力 Interface と出力 Interface 双方の低遅延を両立する方式として NWES(No Wait Exact Spreading)を提案する。さらに、NWES 方式の特性シミュレーションにより、スイッチ Interface 数及びトラヒック負荷に依存せず、低遅延を実現することを示す。

キーワード　ルータ、ロードバランス型スイッチ、クロスバスイッチ

Proposal of a new distributing algorithm for Load Balanced Linear Switch

Hideki NISHIZAKI† and Kenshin YAMADA†

† System Platforms Research Laboratories, NEC Corporation
1753, Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa 211-8666 Japan
E-mail:　†{nisizaki@da, kenshin@ap}.jp.nec.com

Abstract　To accommodate increasing traffic of access and backbone networks year by year, communications equipment such as routers must require very high capacity switches. We propose a novel distributing algorithm for Load Balanced Switch, NWES (No Wait Exact Spreading), achieving low delays in both input and output ports. Simulation results show that NWES can achieve low latency regardless of the number of switch ports and traffic load.

Keyword　Router、Load Balanced Switch、Crossbar Switch

1. はじめに

近年、インターネットの普及、アクセス回線の高速化によりデータトラヒックが急激に増加している。このトラヒックの急激な増加に対応するためには、通信路の高速化やルータを始めとする通信装置の高速、大容量化が必要になってくる。

ここでは、大規模ネットワークを構築するルータなどに適用されるスイッチ技術の一方式であるロードバランス型スイッチに着目して、高速 Interface 収容、多 Interface 収容の観点から現方式の課題とその課題を解決する新方式の提案を行う。

2. ロードバランス型スイッチ[1]~[5]

ロードバランス型スイッチは、入力 Interface と出力 Interface 間に分散配備された中間段 Buffer を有する。各入力 Interface は全中間段 Buffer にセル(受信パケットを固定長サイズに区切ったもの等)を均等に分散し、全中間段 Buffer から均等に各出力 Interface にセルが配送される。このため、入力 Interface－中間段 Buffer、中間段 Buffer－出力 Interface 間は速度 R/N(R:収容 Interface 速度、N:中間段 Buffer 数)の信号のメッシュ接続となる。このため、ロードバランス型スイッチでは、従来のスイッチ構成である入力バッファ型スイッチのように入力出力 Interface 間の接続を決定するためのスケジューラを必要とせず、出力バッファ型スイッチのようにトラヒックの偏りに対応するためのスイッチの速度アップを必要としない。

2.1. ロードバランス型スイッチ実現例

ロードバランス型スイッチでは構成上、入力 Interface－中間段 Buffer 間と中間段 Buffer－出力 Interface 間の接続はメッシュ接続となる。ただし、この形態では収容 Interface 容量の増大や多 Interface 収容を考慮すると実現上非常に困難になってしまう。この

ため、ロードバランス型スイッチを実現する場合には図 2 に示すようにメッシュ接続をそれぞれ前段 Crossbar Switch、後段 Crossbar Switch に置き換える。前段 Crossbar Switch、後段 Crossbar Switch 共に図 3 に示すように周期的な固定設定となる。

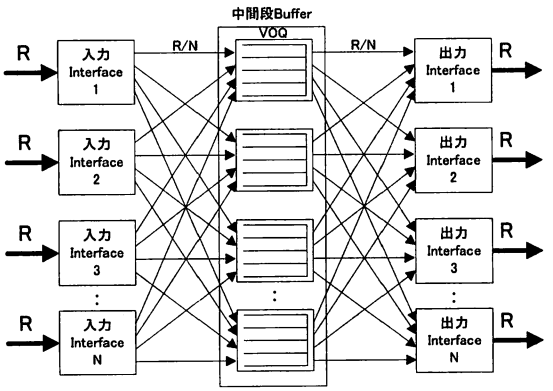

図 1　ロードバランス型スイッチ構成例

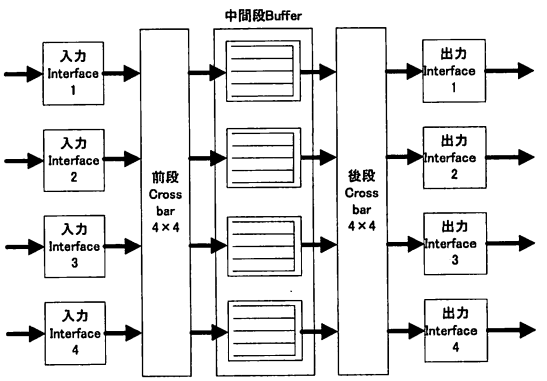

図 2　ロードバランス型スイッチ実現例
(収容 Interface 数=4 の場合)

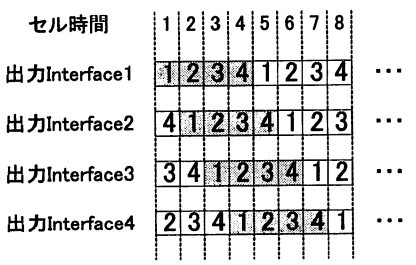

図 3　Crossbar Switch 設定(前段、後段共通)

2.2. ロードバランス型スイッチ基本動作

図 4 をもとにロードバランス型スイッチの基本動作を説明する。図 4 は入力 Interface1 から出力 Interface1 に対して N 個のセルを送出する場合である。入力 Interface1 では、N 個のセルを N 個の中間段 Buffer に 1 セルずつ出力する。各中間段 Buffer は宛先となる出力 Interface 毎にセルを格納する VOQ(Virtual Output Queue)構成のバッファを配備しており、受信したセルは該当宛先の Queue に蓄積される。各中間段 Buffer に格納されたセルはセル本来の宛先である出力 Interface1 に出力される。

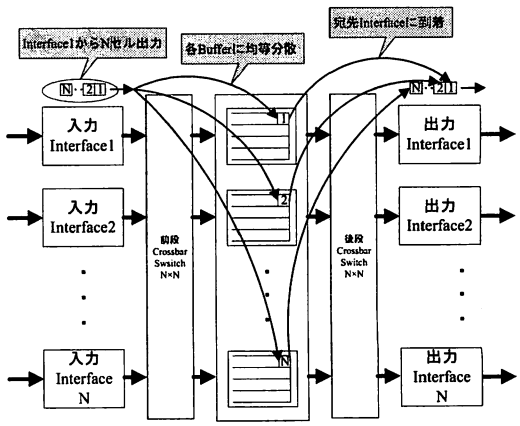

図 4　ロードバランス型スイッチ基本動作

2.3. セル順序入れ替え

ロードバランス型スイッチは、各中間段 Buffer のセル蓄積量に差分が生じるとセル順序が逆転して出力 Interface に到着する場合がある。図 5 にセル順序逆転の一例を示す。

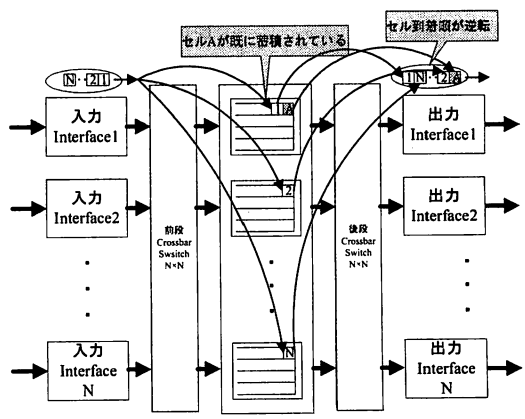

図 5　セル順序入れ替え例

入力 Interface1 から N 個の出力 Interface1 行きのセルを送出する場合、各中間段 Buffer に 1 セルずつ(合計 N セル)送出することになるが、1 つ目の中間段 Buffer に既に 1 セル(セル A)が蓄積された状態では 1 つ目の中間段 Buffer に送出されたセル 1 のみが Queue の 2 セル目に格納されることになる。この後、各中間段から 1 セルずつセル送出が行われると出力 Interface 1 へはセル A→セル 2→・・・→セル N→セル

1 の順で到着することになりセル順序逆転が起こる。このため、出力 Interface では、セル順序の並び替え処理(Reordering 処理)が必要となる。

2.4. ロードバランス型スイッチの遅延

ロードバランス型スイッチでは、各中間段 Buffer に蓄積されるセル数のばらつきにより、2.3.にて述べたセル順序の並び替え処理(Reordering 処理)が必要となる。Reordering 処理による遅延時間を削減するためには、入力 Interface において宛先となる出力 Interface 毎に各中間段 Buffer に対して厳密にセルを均等分散する必要があるが、アルゴリズムによっては入力 Interface から各中間段 Buffer へセルを均一分散する処理のために大きな遅延が発生する。よって、入力 Interface と出力 Interface 遅延時間は一般にトレードオフの関係にある。

3. ロードバランス型スイッチ従来方式

ロードバランス型スイッチの主な従来方式として、Base 方式、UFS(Uniform Frames Spreading)方式と FOFF(Full Ordered Frames First)方式の 3 方式がある。いずれの方式も入力 Interface から中間段 Buffer へのセル振り分け方法に独自性を持たせる方式となっている。

3.1. Base 方式

Base 方式は、入力 Interface へのセル到着順に中間段 Buffer に送出していく方式である。中間段 Buffer におけるセル蓄積状態を考慮せずに入力 Interface からセルを送出するために入力 Interface でのセル送出待ち時間は論理的には 0 となるが、中間段 Buffer 間でのセル蓄積数差分によるセル順序の逆転範囲が規定できないため、条件によっては出力 Interface での Reordering 処理遅延が増大してしまう。

3.2. UFS(Uniform Frames Spreading)方式[4]

UFS 方式は、入力 Interface から中間段 Buffer へセルを出力するときには同一宛先セルを常に N セル(N:収容 Interface 数)単位で送出する方式である。出力セル数が N セルに満たない場合は、空きセルを挿入して N セル単位に揃える。常に N セル単位で送出することにより、中間段 Buffer 間でセルの蓄積差分がなくなるために、出力 Interface での Reordering 処理の必要がないというメリットがある。反面、N セル揃うまでの待ち時間や空きセル挿入判定処理などの入力 Interface での処理待ち遅延に課題がある。

3.3. FOFF(Full Ordered Frames First)方式[4]

FOFF 方式は、同一宛先セルに関して入力 Interface から各中間段 Buffer に順番に出力していく方式である。例えばある出力 Interface 行きのセルを 3 セル出力する場合、中間段 Buffer1～3 に 1 セルずつ出力し、その後に同一出力 Interface 行きのセルが来た場合は、中間段 Buffer4 から出力していく。これにより、各中間段 Buffer のセル蓄積差分を 1 入力 Interface あたり 1 セル以内に抑えることができる。この場合、出力 Interface でのセル順序逆転が起こる場合があるが、セル順序の逆転範囲が規定できるため、出力 Interface での Reordering 処理遅延を一定時間内に抑えることが可能となる。反面、宛先となる出力 Interface 毎に次に出力する中間段 Buffer が 1 つに決まってしまうため、入力 Interface での処理待ち遅延が大きくなってしまう課題がある。

4. NWES(No Wait Exact Spreading)方式

本稿にて提案する NWES 方式は入力 Interface から中間段 Buffer へのセル振り分け方法に独自性を持たせる方式となっている。NWES 方式では、同一宛先毎に入力 Interface から各中間段 Buffer へ出力したセル数をカウントしている。入力 Interface からのセル出力先となる中間段 Buffer は同一宛先セルに関して各中間段 Buffer へのセル出力数差分がある一定値(M とする)以内に収まるように決定される。この場合、入力 Interface からのセル送出となる中間段 Buffer は常に一つに限定されるということはなくなるために入力 Interface での処理待ち遅延を小さく抑えることが可能となる。また、中間段 Buffer 間のセル蓄積数差分も 2M セル以内に抑えることが可能となるため、出力 Interface での Reordering 処理遅延を一定時間内に抑えることが可能となる。

4.1. 入力 Interface 処理

入力 Interface では、セルの宛先となる出力 Interface 毎に各中間段 Buffer にセルの出力数をカウントしている。図 6 に入力 Interface 処理の一例を示す。

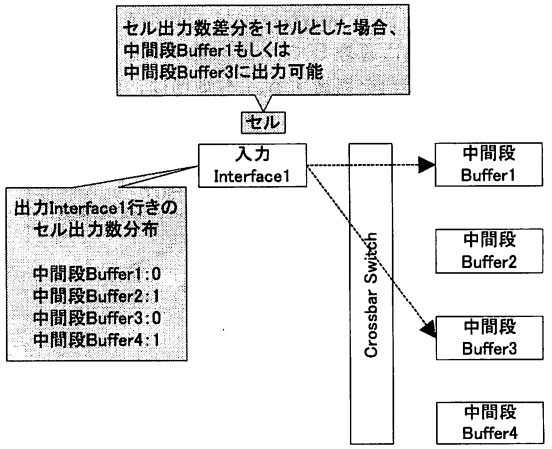

図 6　入力 Interface 処理

入力 Interface1 では、出力 Interface1 行きのセルを中間段 Buffer1 と 3 には 0 セル、中間段 Buffer2 と 4 には 1 セル出力している。各中間段 Buffer へのセル出力数差分を 1 とした場合、中間段 Buffer2 または 4 にセル

を出力すると差分が 2 となってしまうために、新たに出力 Interface1 行きのセルを出力できるのは中間段 Buffer1 と 3 になる。中間段 Buffer1 と 3 のどちらにセルを出力するかは現在のセル時刻(図 3 の Crossbar Switch 設定参照)から最も待ち時間が少なく出力できる中間段 Buffer を選択する。

4.2. 中間段 Buffer 処理

中間段 Buffer では、内部に VOQ(Virtual Output Queue)を配備している。入力されたセルは宛先 Interface 毎に該当する Queue に格納される。VOQ を構成する各 Queue からは後段の Crossbar Switch の設定に従って周期的に 1 セルずつ読み出しが行われる。本処理に関しては、ロードバランス型スイッチの各方式に関して共通である。

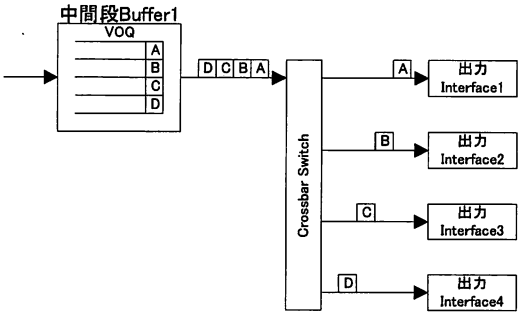

図 7　中間段 Buffer 処理

4.3. 出力 Interface 処理

出力 Interface では、同一出力 Interface 行きのセルに関して中間段 Buffer 間にセル蓄積数に差分が生じると、セル順序が逆転して出力 Interface に到着する場合がある。このため、出力 Interface では、セルを送出元の入力 Interface 毎に管理して、セル順序の並び替え処理(Reordering 処理)を実施する。本処理は、UFS 方式では不要であるが、Base 方式、FOFF 方式では NWES 処理と同様に必要である。

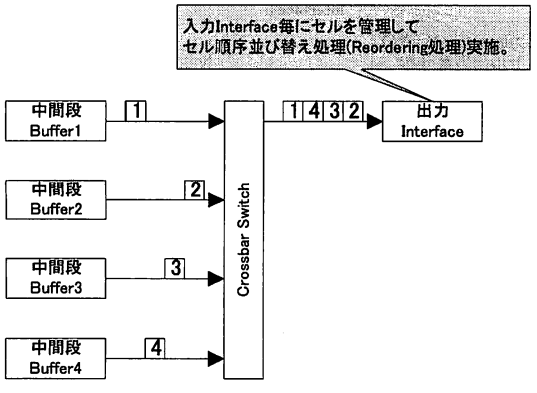

図 8　出力 Interface 処理

5. ロードバランス型スイッチの遅延分析

ロードバランス型スイッチでは特に入力 Interface から中間段 Buffer への送出待ち時間と出力 Interface でのセル順序並び替え処理(Reordering 処理)が主な遅延となる。本遅延に関して以下に分析する。

5.1. 入力 Interface での遅延

5.1.1. Base 方式

Base 方式では、中間段 Buffer でのセル蓄積状態を考慮せずに入力 Interface から中間段 Buffer にセル送出するために論理的には遅延 0 となる。

5.1.2. UFS 方式

UFS 方式では、入力 Interface にて N セル(N:収容 Interface 数)揃うまでの時間と前段 Crossbar Switch の設定が所定の中間段 Buffer へ送出できるようになるまでの時間が待ち時間となる。特にセル送出を開始する中間段 Buffer は常に 1 つ(例えば中間段 Buffer1)に固定されているために、セルが出力可能になってから実際に出力されるまでの待ち時間が大きくなる場合がある。

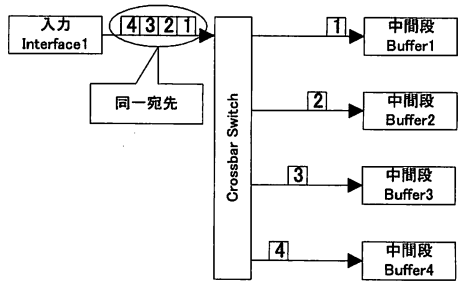

図 9　UFS 方式での入力 Interface 処理

5.1.3. FOFF 方式

FOFF 方式では、前段 Crossbar Switch の設定が所定の中間段 Buffer へ送出可能となるまでの時間が待ち時間となる。セル送出先の中間段 Buffer は、前回送出した中間段 Buffer の次の中間段 Buffer と決められている。このためにセル送出を開始する中間段 Buffer は常に 1 つになるために、セルが出力可能になってから実際に出力されるまでの待ち時間が大きくなる場合がある。

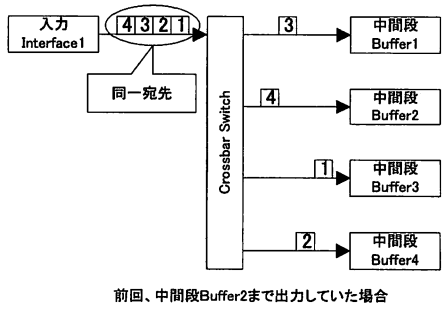

図 10　FOFF 方式での入力 Interface 処理

5.1.4. NWES 方式

NWES 方式では同一出力 Interface 行きのセルに関して、各中間段 Buffer への送出セル数差分が一定値(例えば1セル)以内であればどこの中間段 Buffer でも送出できるため、最も待ち時間の少ないセル送出可能な中間段 Buffer を選択することができる(図 6 参照)。

5.2. 出力 Interface での遅延

5.2.1. Base 方式

Base 方式では、中間段 Buffer 間のセル蓄積状態が不均衡になりやすく、かつ最大差分値も規定できないために状況によっては Reordering 処理遅延が増大してしまう場合がある。

5.2.2. UFS 方式

UFS 方式では、各中間段 Buffer でのセル蓄積状態に差分が出ないために Reordering 処理の必要がない。

5.2.3. FOFF 方式

FOFF 方式では、1 入力 Interface あたり中間段 Buffer 間のセル蓄積数差分が最大 1 セルなので N 個の Interface 収容時には最大 N セルの差分となる。このため出力 Interface では、並び替えが必要なセル到着時間は最大 N^2 セル時間(N 中間段 Buffer×N(セル差分))となる。

5.2.4. NWES 方式

NWES 方式では、同一出力 Interface 行きのセルに関して中間段 Buffer 間への送出セル数差分を M セルとした場合、中間段 Buffer 間のセル蓄積数差分が最大 2M セルとなるので N 個の Interface 収容時には最大 2MN セルの差分となる。このため出力 Interface では、並び替えが必要なセル到着時間は最大 $2MN^2$ セル時間(N 中間段 Buffer×2MN(セル差分))となる。

6. シミュレーション結果

6.1. NWES 方式におけるセル送出数差分適正値

NWES 方式における入力 Interface から中間段 Buffer 間への送出セル数差分の最適値を求めるためにシミュレーションを実施した。

6.1.1. シミュレーション条件

シミュレーション条件を表 1 に示す。

表 1　シミュレーション条件

項目	条件
取得特性	トラヒック負荷 vs. 平均遅延
トラヒックモデル	ランダム
	バースト(On-Off Model) バースト長：16、32、64、128 セル
Interface 数	N=4,16,32,64,128
負荷	Load=1,10,30,50,70,90,99%
計測対象	10,000,000 セル
対象方式	NWES
セル数差分	1,2,4,8

6.1.2. シミュレーション結果

表 1 の条件によるシミュレーション結果を図 11 に示す。

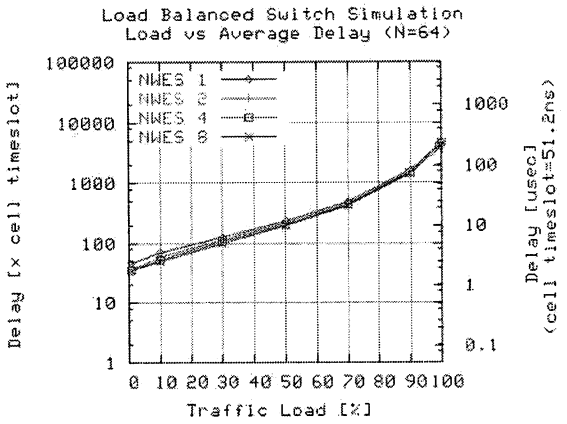

図 11　トラヒック負荷 vs. 遅延特性
(収容 Interface 数 N=64、平均バースト長=128 セル時)

6.1.3. シミュレーション結果考察

本シミュレーション結果から、Interface 数やトラヒック条件を変化させても入力 Interface から各中間段 Buffer に出力する同一宛先セルのセル数の差分 1,2,4,8 の各場合における特性に大きな違いは見られない。厳密に言えば、各場合において高負荷時において差分が小さい方の特性が良くなる傾向にある。

6.2. 方式間比較

6.2.1. シミュレーション条件

表 2　シミュレーション条件

項目	条件
取得特性	トラヒック負荷 vs. 平均遅延
トラヒックモデル	ランダム
	バースト (On-Off Model) バースト長：16、32、64、128 セル
Interface 数	N=4,16,32,64,128
負荷	Load=1,10,30,50,70,90,99%
計測対象	■Interface 数 128 の場合 測定セル数 100,000,000 セルで平均遅延の算出は後半の 50,000,000 セル ■Interface 数 64 の場合 測定セル数 25,000,000 セルで平均遅延の算出は後半の 12,500,000 セル ■Interface 数 32,16,4 の場合 測定セル数 10,000,000 セルで平均遅延の算出は後半の 5,000,000 セル ■Load=1%の場合は測定セル数を上記の 1/10 とし、平均遅延の算出は後半分のセルとする。
対象方式	NWES、Base、UFS、FOFF

※NWES 方式におけるセル送出最大差分値は 1 セル。

6.2.2. シミュレーション結果

表 2 の条件によるシミュレーション結果を図 12、図 13 に示す。

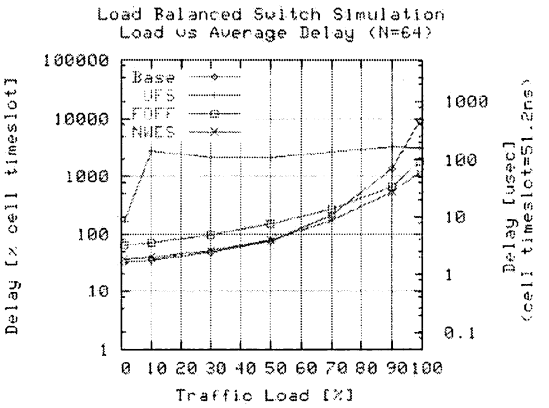

図 12　トラヒック負荷 vs. 遅延特性
(収容 Interface 数 N=64、ランダムトラヒック時)

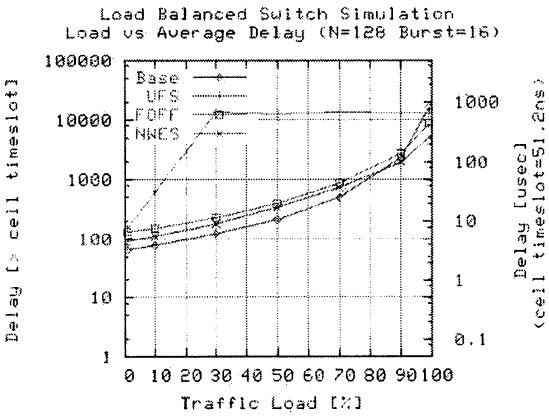

図 13　トラヒック負荷 vs. 遅延特性
(収容 Interface 数 N=128、平均バースト長=32 セル時)

6.2.3. シミュレーション結果考察

以下にシミュレーション結果について考察する。

- 低負荷では、どの方式であっても中間段 Buffer 間のセル蓄積状態の不均衡があまり発生しないために、入力 Interface にてできるだけ遅延を発生させない方式である Base 方式、NWES 方式が他方式に比べて特性が良い。
- 高負荷になると、各中間段 Buffer へのセル均等分配処理を行わない Base 方式では中間段 Buffer 間のセル蓄積状態に不均衡が生じ、これにより Reordering 処理遅延が大きくなるため、遅延特性が他方式に比べて劣化する。
- 従来方式である Base 方式、UFS 方式、FOFF 方式の各方式と NWES 方式の遅延特性を比較すると、Load やトラヒック(ランダム、バースト)の各条件にて常に低遅延になっているのは NWES 方式のみである。これは、NWES 方式が入力 Interface における各中間段 Buffer に対するセルの分散処理遅延と出力 Interface での Reordering 処理遅延の両方の低遅延化を実現しているためである。

7. まとめ

高速、大容量化を実現する上で有力なスイッチ構成であるロードバランス型スイッチに対する処理方式として NWES 方式を提案し、従来方式よりも低遅延が実現できることをシミュレーションにより確認した。今後は、マルチキャスト、優先処理の適用などスイッチ実用化に向けた検討を行っていく予定である。

8. おわりに

本内容の一部は、総務省委託研究「次世代バックボーンに関する研究開発」の成果である。

文　献

[1] C. S. Chang, D. S. Lee, Y. S. Jou, "Load balanced Birkhoff-von Neumann
switches, Part I: one-stage buffering," Computer Communications,
vol. 25, pp. 611-622, 2002.

[2] C. S. Chang, D. S. Lee, C. M. Lien, "Load balanced Birkhoff-von
Neumann switches, Part II: multi-stage buffering," Computer Communications,
vol. 25, pp. 623-634, 2002.

[3] I. Keslassy, S. T. Chuang, K. Yu, D. Miller, M. Horowitz, O. Solgaard,
and N. McKeown, "Scaling Internet routers using optics," ACM SIGCOMM,
Karlsruhe, Germany, 2003.

[4] I. Keslassy, "The Load-Balanced Router," Ph.D. Thesis, Stanford University,
2004.

[5] C. S. Chang, D. S. Lee, Y. J. Shih, "Mailbox switch: a scalable two-stage
switch architecture for conflict resolution of ordered packets," IEEE
INFOCOM, Miami, FL, 2004.