社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

回線速度 40Gbps の 128×128 光パケットスイッチをサポートする並列パイプライン制御によるバッファ管理方式

原井　洋明†　　村田　正幸††

† 独立行政法人通信総合研究所 情報通信部門　〒 184-8795 小金井市貫井北町 4-2-1
†† 大阪大学 サイバーメディアセンター　〒 560-0043 豊中市待兼山町 1-30
E-mail: †harai@crl.go.jp, ††murata@cmc.osaka-u.ac.jp

あらまし　出力バッファ型光パケットスイッチにおける高速バッファ管理手法について検討する．並列処理とパイプライン処理を組みあわせた複数プロセッサ構成による制御方式を提案する．各プロセッサにおいて，ポート数 N に依存しない計算量 $O(1)$ の処理を行ない，従来の N 倍の高速化を図る．回線速度 40Gbps の 8×8 光パケットスイッチにおいて，最小 64 バイトの長さの異なるパケットを処理する機能を FPGA に実装できることをシミュレーションによって確認する．また，提案方式と FPGA 技術を用いて回線速度 40Gbps の 128×128 光パケットスイッチをサポートできることを示す．

キーワード　光パケットスイッチ，出力バッファ，並列パイプライン処理，非同期可変長パケット， FPGA

Buffer Management based on a Parallel and Pipeline Mechanism to Support 128×128 Photonic Packet Switches with 40Gbps Ports

Hiroaki HARAI† and Masayuki MURATA††

† Communications Research Laboratory　Koganei-shi, Tokyo 184-8795, Japan
†† Osaka University　Toyonaka-shi, Osaka 560-0043, Japan
E-mail: †harai@crl.go.jp, ††murata@cmc.osaka-u.ac.jp

Abstract　We investigate a high-speed buffer management mechanism for output-buffered photonic packet switches. We propose a parallel and pipeline mechanism on multi-processing architecture for this purpose. The machanism provides N times faster processing than an existing $O(N)$ mechanism does, where N is the number of ports. Through hardware simulation after place and route operation, we confirm feasibility of an FPGA-based buffer management hardware for 8×8 photonic packet switches with 40Gbps ports, which is capable of asynchronously arriving variable-size packets, of which minimum is 64byte. A support of 128×128 packet switch with 40Gbps ports is also feasible by using our mechanism and a latest FPGA technology.

Key words　Photonic packet switch, Output buffer, Parallel and pipeline processing, Asynchronous variable-size packets, Field programmable gate array

1. はじめに

最近では 10Gbps 回線を 32 ポート備えた IP ルータ [1] があるが，電子処理は限界に近付きつつあると言われて久しい．本稿では基幹ネットワークへの適用を目指して，現状の光技術で実証されている回線速度 40Gbps のポートを備えた光パケットスイッチ [2], [3] を対象としたバッファ管理方式を検討する．ポート数のサポートがどこまで実現できるかを明らかにし，それによって，電子処理のみのパケットスイッチに対する光パケットスイッチの優位性を示す．

光パケットスイッチの機能は大きくラベル検索 (フォワーディング)，交換，バッファ管理 (スケジューリング)，バッファリング，経路制御 (ルーティング) の 5 機能にわけられる． 40Gbps や 160Gbps といった高速の光パケットを O/E/O 変換なく転送するには，交換とバッファリングでは，パケットを光信号のまま扱わねばならない．さらに，短時間で大量のパケットを転送するためには，ラベル検索におけるメモリへのアクセス速度やバッファ管理における処理速度がボトルネックになる．今後の光パケット交換技術の進展のためには，ラベル処理をメモリアクセスを伴なわない光技術で行なうことが望ましい．実際，ラベル検索 (多波長ラベル処理 [4]，光位相符号ラベル処理 [5])，交換，バッファ [6] などは光技術による実現性が確認されている．これらの機能を備えたプロトタイプも開発されている [2]．光バッファは光ファイバ遅延線 (FDL; Fiber Delay Line) を用いて構成できる一方，実用的な光論理や光メモリ (RAM) はまだなく，バッファ管理は光パケットの遅延時間を決めるための電子処理が必要である．電子処理性能が今後も向上し続けるとは限らず，計算量が小さなバッファ管理アルゴリズムを開発する必要がある．

我々は出力バッファ型 $N \times N$ パケットスイッチを対象とする．入力バッファ方式と比べ，出力バッファ方式は良好な遅延特性およ

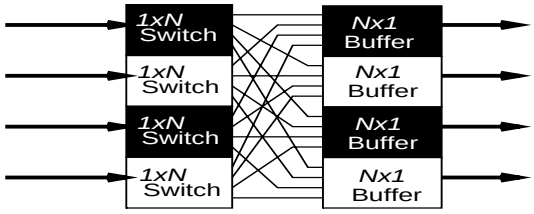

図 1　$N \times N$ 光パケットスイッチ ($N = 4$)

びスループット特性を持つ．これは， HOL (Head of Line) ブロッキングが起らないためである．一方，入力バッファ方式よりも N 倍バス速度の大きなスイッチを必要とするので，実装が難しい．そこで，HOL ブロッキングを回避し，出力バッファ方式と同等の論理性能を得る複数入力バッファ方式 (MIQ; Multiple Input Queue) が考えられている．しかし，図 1に示すように，我々が対象とする光パケットスイッチは，N 個の $1 \times N$ スイッチを束ねた構成なので，単一出力ポートに N 本の回線を備える．これにより N 倍バス速度が大きなスイッチを用いるのと同等の性能を得られる．さらに，出力ポートにおける衝突回避のために， MIQ はアービトレーション機能 [7] を必要とする．このアービトレーションは入力側にてメモリバッファの使用を前提としたものであり，光ファイバ遅延線バッファを使用しての実現は困難である．それゆえ，我々は出力バッファ方式をそのまま実現するアーキテクチャに着目する．

光パケットの衝突回避には光ファイバ遅延線バッファを用いる．光パケットを光バッファに保持するには，光パケットがパケットスイッチに到着後，出力バッファに到着するまでの固定時間で，光パケットの遅延時間を求めねばならない．したがって，連続して到着する光パケットをすべて処理するには，パケット長に相当する時間以内に最大 N パケットを処理するバッファ管理が必要になる．単純なラウンドロビンスケジューリング方式を用いる場合の計算量は $O(N)$ である．ポート数が大きくなると，処理が追いつかなくなることがじゅうぶんに考えられる．高スループットの光パケットスイッチ実現には，電子処理によるボトルネックの回避が不可欠にも関わらず，計算量 $O(N)$ より高速なバッファ管理方式は検討されていない．

一方，既存の電子処理を駆使した，ルータや ATM 交換機を代表としたパケットスイッチでは，段階を追って高速化が検討されてきた．最初は伝統的な単一サーバ方式からの移行である．近年の基幹に用いられている装置では，複数プロセッサを用いたパイプライン処理が一般的である．すなわち，入力回線ごとに宛先検索処理を行ない，出線毎に用意したバッファにパケットを格納し，その後，別プロセッサを用いて出線のアービトレーションを行なう [7], [8]．本方式では，プロセッサ (LSI 回路規模) は増加するが，単位時間により多くのパケットを処理でき，高スループット化が図れる． FPGA (Field Programmable Gate Array) や ASIC (Application Specific Integrated Circuit) など大規模 LSI の進展により，このように複数のプロセッサを用いて並列／パイプライン処理を行ない，高スループットを実現する．しかし，先述のように，複数入力バッファ方式の光パケットスイッチへの適用は困難である．

文献 [9] では，出力バッファ型パケットスイッチにおいて，メモリバッファにパケットを格納するための計算量が $O((\log_2 N)^2)$ のバッファ管理方式を提案している．本方式では，時系列を周期にわけ，各周期に到着するパケットの出力時刻を求め，同一メモリから同時に複数のパケットを出力する処理を避けるようにバッファにパケットを蓄積する処理を行なう．出力時刻を求めるために N プロセッサによる並列プレフィクス演算 (Parallel Prefix Operation) という並列処理 [10], [11] を行ない，計算量 $O(\log_2 N)$ の高速化を達成している．

文献 [9] では対象とするパケットは固定長であり，同期到着する．しかし，インターネットを流れるパケットの長さは様々である [12]．固定長同期パケットのためのバッファ管理方式を適用するためには，(1) 対象とするネットワークの送信エッジノードにおいて可変長パケットを複数の固定長パケットに分割し，さらに受信エッジノードでもとのパケットに戻す処理が必要である．または， (2) ネットワークの各ノードにおいて，入力時に可変長パケットを固定長に分割し，出力時にもとのパケットに戻す処理が必要である．いずれもノードにおける光同期処理 [13] 〜 [15] が必要で，光システムがより複雑になる．中継ノードの処理の負担を減らすには，同期処理を用いず長さの異なるパケットを処理する機能が有用である．近年，非同期到着する可変長パケットを扱う光パケットスイッチのためのバッファ管理方式も提案されている [15] 〜 [17]．しかし，計算量は $O(N)$ と同等かそれより複雑である．

本稿では，出力バッファ型光パケットスイッチにおいて非同期到着し長さの異なるパケットを処理する高速バッファ管理方式を検討し，並列処理とパイプライン処理を組み合わせた複数プロセッサ構成による制御方式を提案する．各プロセッサにおいて，パケットスイッチのポート数 N に依存しない計算量 $O(1)$ の処理を行なう．複数プロセッサ構成の装置規模は，$O(N \log_2 N)$ であるが，提案方式は，従来の光パケットスイッチの管理方式よりも N 倍のスループットを達成する．従来のパケットスイッチの管理方式 [9] のスループットよりも $(\log_2 N)^2$ 倍大きい．

我々が提案する方式の装置規模は $O(N \log_2 N)$ なので，集積性について実現可能性を検討する必要がある．光パケットスイッチでは，大規模なメモリを使わず，光遅延線バッファを用いるので，メモリバッファを用いた構成よりも簡単に実現できる．我々は， FPGA への配置配線処理を施した後のハードウェアシミュレーションにより，最小 64 バイトの長さの異なるパケットが非同期に到着する，回線速度 40Gbps の 8×8 光パケットスイッチのバッファ管理装置を実現できることを確認した．さらに，電子技術による最新の IP ルータの少なくとも 16 倍の性能となる，回線速度 40Gbps の128×128光パケットスイッチにおけるバッファ管理装置のサポートが提案方式と既存の FPGA 技術を用いてできる．

本稿の構成を以下に述べる． 2.において，対象とする光パケットスイッチ構成を示し，基本バッファ管理方式を述べる． 3.において，並列パイプライン処理による高速バッファ管理方式を提案する． 4.ではハードウェアによる実現可能性を検証する． 5.では光パケットスイッチのスケジューリング特性を示す． 6.ではまとめを述べる．

2. 光パケットスイッチ

2.1 出力バッファ構成

現在，入力バッファ型 [13]，周回型 [13]，出力バッファ型 [2], [6], [15] の光パケットスイッチが提案されている．我々は出力バッファ構成のスイッチを用い，波長変換を用いない．本稿におけるパケットスイッチ構成は，文献 [2] のバッファ管理装置を非同期可変長パケット対応に置きかえた構成に近い．

先述の図 1に，バッファ管理手法を適用する光パケットスイッチ構成を示す．$N \times N$ パケットスイッチは，N 個の$1 \times N$ バッファレスパケットスイッチと N 個の $N \times 1$ 光バッフの光学的フルメッシュ接続から構成される．$1 \times N$ パケットスイッチでは， [4], [5] 等を用い超高速のラベル検索を行なう．可変長パケットは同一ラベルをペイロードの前後につけることでサポートできる [18]．$N \times 1$ 光ファイバ遅延線バッファは，光スイッチと長さの異なる B 個のファイバ遅延線 $d_0, d_1, \ldots, d_{B-1}$ からなる．遅延線の単位長は D であり，遅延線 d_k の長さは kD である．光バッファでは，0 から $(B-1)D$ までの離散時間の遅延を与える．図 2 (a) は，$N = 4, B = 4$ の光ファイバ遅延線の構成を示す．

2.2 逐次制御によるバッファ管理装置の振舞い

ルータなど電子処理によるノードで用いられる RAM バッファと異なり，光遅延線バッファでは，光の直進性のために光パケットを蓄積できない．光遅延線バッファの実現は，到着するパケットを異

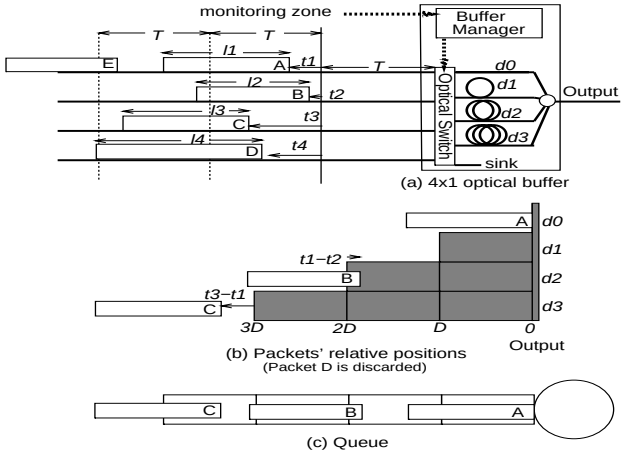

図 2　(a) 光ファイバ遅延線バッファ ($N = 4, B = 4$) とパケットの到着 (b) 割当てられる遅延線とパケットの相対位置 (c) 論理的なパケットの位置

なるファイバ遅延線に割当て，パケットの衝突を回避することで可能になる．選択される遅延線は各パケットがバッファに到着する前に与えねばならず，バッファ管理装置がその処理を行なう．出力バッファ型$N \times N$光パケットスイッチでは，最大 N パケットが同時にある$N \times 1$バッファに到着する．遅延線バッファを用いる場合，最小パケット長 $l_{\min}$ に相当する時間以内に N パケットに対して遅延を与えるバッファ管理装置が不可欠である．

次章で述べる並列パイプライン方式の理解のため，ここでは，その基となるラウンドロビンスケジューリング (逐次処理) によるバッファ管理装置の振舞いを述べる．4×4 光パケットスイッチの一つのバッファに到着するパケットを書いた図 2に再度注目する．非同期到着する可変長パケットを処理するバッファ管理装置は，逐次処理の周期時間を示す内部クロック (周波数 $1/T$) を持つ．バッファ管理装置は，実際にパケットが光スイッチに到着する T 前に，T 時間後から $2T$ 時間後までに到着するパケットの情報を基にして，時間 T 以内にすべての到着パケットの遅延時間を求める．パケットの連続到着を許す場合，その周期 T は最小パケット長 $l_{\min}$ 以下でなければならない ($T \leqq l_{\min}$)．この制約は，パケットが光バッファの光スイッチに到着する前に遅延を求めるために必要である．

バッファ管理装置の処理を述べる．対象とする周期にポート n $(n = 1, 2, \ldots, N)$ に到着するパケットの長さを l_n，周期開始からパケット到着までの時間 (到着ギャップと呼ぶ) を t_n とする (図2(a)参照)．バッファ管理装置はパケットの到着情報として，パケット長と到着ギャップを受取る．バッファ管理装置は，周期の開始時刻を 0 として，バッファ内にある全パケットが出力される時刻を表わす変数 q を管理する．q をバッファ占有度と呼ぶ．また，バッファ管理装置は，周期時間 T でポート $1, 2, \ldots, N$ の順に，対象とするパケットが進む遅延線を求め，実際にパケットが到着した時に適切に光スイッチを切替える．

パケットの衝突回避に十分な遅延時間は $q - t_n$ である．しかし，光遅延線バッファの遅延の離散時間特性により，パケットの遅延は $\Delta_n D$，ただし，$\Delta_n = \lceil \frac{q-t_n}{D} \rceil$ となる．$\Delta_n < B$ であれば，パケットは遅延線 d_{Δ_n} に転送され，$\Delta_n \geqq B$であれば，パケットは棄却される．パケットが遅延線に転送される場合，バッファ占有度 q は次のパケットを適切に処理するために更新され，$q \leftarrow t_n + l_n + \Delta_n D$となる．全ポートのパケットの遅延を求めた後は，次周期に到着するパケットを適切に処理するために，$q \leftarrow \max(q - T, 0)$ と更新する．図 3に逐次処理を実現する擬似コードを示す．図中 "packet n is given delay $\Delta_n D$" とは，パケットを遅延線 d_{Δ_n} に送ることを意味する．N ポートに到着する最大 N パケットを一周期で順に処理するので，逐次処理方式の計算量は $O(N)$ となる．

4 パケット A ～ D が同一周期に， E がその次周期に光バッファ

```
for n := 1 to N do
begin
  if (l_n ≠ 0) then begin
    Δ_n := ⌈(q − t_n)/D⌉;
    if Δ_n < B then begin
      q := t_n + l_n + Δ_n D;
      Packet n is given delay Δ_n D; end
    else Packet n is discarded;
  end
end
q := max(q − T, 0);
```

図 3　N ポートパケットスイッチにおいて逐次処理を実現する擬似コード

に到着する状態を想定する (図 2(a))．バッファ管理装置は，パケット A,B,C をそれぞれ遅延線 d_0, d_2, d_3 に転送する．その時点でバッファあふれとなり，パケット D は棄却される．バッファリングと廃棄処理は，光スイッチを駆動することで実現する．図 2(b) には，パケット A が遅延線 d_0 に格納された直後における出力ポートから見たパケットの相対位置を示す．図に示すように， 3 パケットが衝突なくバッファから出力される．図 2(c) には，パケットが図 2(b) の状態の光バッファにおけるパケットの論理的な位置を示す．連続するパケットの間には空き (Void) が存在する．光遅延線バッファの離散特性のため，光バッファは新たなパケットとその直前に格納されたパケットの間に$\Delta_n D - q + t_n$の空きを持つ．この空きを減らしバッファ利用率を改善するために， Void Filling という手法がある [16]．しかし， Void Filling の計算量は逐次処理の計算量よりも大きく，回線速度を制限しがちである．さらに，パケットの到着順と出力順が異なるかもしれない．逐次処理では，ポートごとに注目すると，棄却されるパケットを除いて順序の一貫性が保たれる．例えば，パケット E はパケット A より後に到着し，出力する．順序を維持し高い回線速度を維持しながらパケット棄却率を小さくするには，逐次処理を用いると共に遅延線の単位長 D の設定が重要である．例えば，文献 [15], [19] では，負荷 0.8 において D を平均パケット長の 0.3 倍程度にすると棄却率が最小になることが示されている．

3. 並列／パイプライン処理による高速バッファ管理

本章では，長さの異なるパケットが非同期に到着する光パケットスイッチのバッファ管理に用いる，並列処理とパイプライン処理を組合せた複数プロセッサ構成による制御方式を提案する．文献 [20] に述べた固定長パケットが同期して到着する場合の管理よりも処理が複雑になるが，光同期システムやパケットの分解／再構成という処理が不要になる．

3.1 マルチプロセッサシステムとその機能分担

我々の提案する制御は，並列プレフィクス演算をパイプライン化する機能を用いることで実現する．並列プレフィクス演算とは，N 個の要素$< a_1, a_2, \ldots, a_N >$が与えられた時に，$N$ 個のプロセッサを用いて，$(s_n = \sum_{i=1}^{n} a_i)$ で定義される N 個のプレフィクス和 $< s_1, s_2, \ldots, s_N >$を求める演算である [10], [11]．本稿ではパイプライン化した演算を並列パイプラインプレフィクス演算と呼ぶ．

図 4に，$N = 8$ における制御を実現するプロセッサの配置を示す．本構成は，$(\log_2 N + 1)$ 段のパイプラインステージからなり，それぞれ複数のプロセッサ (図中丸) と複数のレジスタ (図中四角) から構成される．第 k 段 $(k = 1, 2, \ldots, \log_2 N)$ には，$(N - 2^{k-1})$ 個のプロセッサ $P_{k,n}$ $(n = 2^{k-1} + 1, \ldots, N)$ が配置され，最終段には，$(N + 1)$ 個のプロセッサが配置される．

前 $(\log_2 N)$ 段では，「到着時にバッファにパケットはなく，逐次処理に従って，すべてのパケットがバッファに格納される」と仮定して，到着パケットの相対遅延時間を求めるプレフィクス演算部を構成する．例えば，ポート 1 へ到着するパケットの相対遅延は常に 0 となる．ポート 2 へ到着するパケットの相対遅延は，同じ周期にポート 1 にパケットが到着するなら，そのパケット長であり，到着

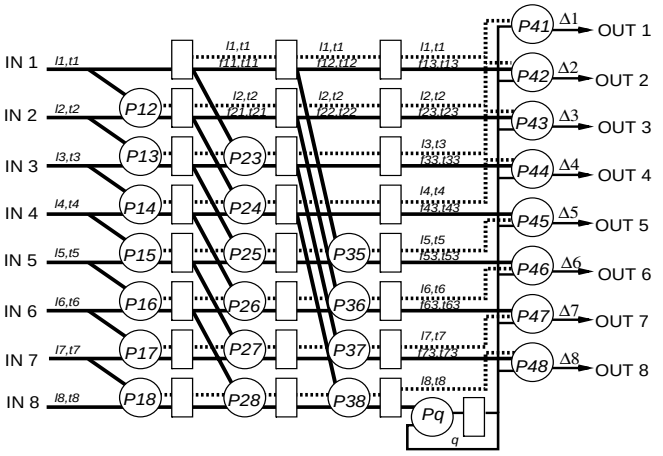

図 4　並列パイプライン処理アーキテクチャ ($N = 8$)

しなければ 0 となる．第 k 段のプロセッサは，並列プレフィクス演算の第 k 段の演算として用いられる．第 $(\log_2 N)$ 段のプロセッサに後方にある第 n 番目のレジスタに格納された値は，ポート$(n+1)$に到着するパケットに与えられる相対遅延を示す．ポート $(n+1)$ に到着するパケットの相対遅延は，ポート 1 から n に到着するパケットの連結パケット (3.2参照) の長さと到着ギャップから求める．最後方の第$(\log_2 N+1)$段に配置された$(N+1)$ 個のプロセッサが，遅延決定部を構成する．そのうち，N 個のプロセッサ$P_{\log_2 N+1,n}$ $(n = 1, 2, \ldots, N)$ は，現在のバッファ占有量と，プレフィクス演算部より送られた相対遅延から，全 N ポートに到着するパケットの遅延を並列的に求める．残りのプロセッサ P_q では，バッファ占有量を更新する．図 4において，実線は各プロセッサで計算した値を転送するバスを示し，破線はマルチプロセッサシステムの入力ポートに到着した値を転送するバスを示す．

このマルチプロセッサシステムがバッファ管理装置となる．ポートに到着した各パケットの遅延は，$(\log_2 N+1)$個のプロセッサを経由して求められる．したがって，本構成において到着するパケットの遅延を求めるには，$(\log_2 N+1)$ 周期を要する．バッファ管理装置は，パケットが光バッファ内の光スイッチに到着する $(\log_2 N+1)T$ 時間前に，そのパケットの情報を得るようにする．

図 4において INn で示された入力ポート n への到着情報は，$(\log_2 N+1)$ 周期後にポート n へパケットが到着する場合にはその長さ l_n と到着ギャップ t_n であり，到着しない場合には “0” である．OUT n からは，ポート n へ到着する光パケットの遅延が出力される．その遅延を基に光バッファ内の光スイッチが適切に駆動され，光パケットに遅延が与えられる．

3.2　連結パケットの導入

非同期到着かつ可変長パケットを処理するときのポイントは，パケットの相対遅延をいかに求めるかである．光同期を用いれば，単に，パケット長を各要素としてプレフィクス和を求めればよい．しかし，光同期システムはなく，遅延線バッファの離散特性を考慮すると，パケット長のみならずパケットの到着ギャップを考慮して相対遅延を求めねばならない．2 パケット間の間隔を適切に保つために，我々は同一周期に異なるポートに到着した複数のパケットが仮想的に接続された仮想連結パケットを導入する．2 つのパケットからなる連結パケットの長さ l' を以下に定義する．

$$l' = \left\lceil \frac{t_1 + l_1 - t_2}{D} \right\rceil D + t_2 + l_2 - t_1 \qquad (1)$$

ここで t_1, l_1 はそれぞれ，前方のパケットの長さと到着ギャップを表わし，t_2, l_2 は後方のパケットのそれを表わす．到着ギャップ t' は，$l_1 \neq 0$，すなわち，着目した 2 ポートのうち，前方のポートにパケットの到着がない時には $t' = t_2$ とし，到着があれば$t' = t_2$ とする．図 5に連結パケットの例を示す．本章での並列パイプラインプレフィクス演算では，この定義を用いる．図 6 には，プレフィクス演算部における各ステージで発生する連結パケットの長さを示している．各記号は 3.3で定義する．2 パケットの間隔を適切に保持することで，光パケット内および出力での衝突はない．この連結パケットを用い，「到着時にバッファにパケットはなく，逐次処理に従って，すべてのパケットがバッファに格納される」と仮定して到着パケットの相対遅延を計算する．

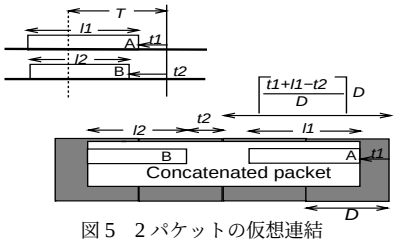

図 5　2 パケットの仮想連結

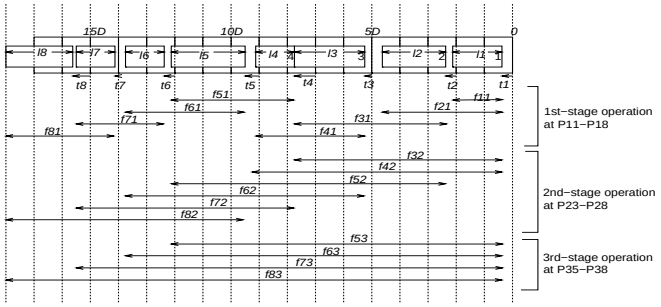

図 6　前 $\log_2 N$ 段で発生した連結パケットの長さ ($N = 8$)

for each processor n, in parallel ($n := 2$ to N)
begin
  if ($l_{n-1} = 0$) **then** $t_{n,1} := t_n$
  else $t_{n,1} := t_{n-1}$;
  if ($l_n = 0$ **then** $f_{n,1} := l_{n-1}$
  else begin
    $\Delta_n := \left\lceil \dfrac{t_{n-1} + l_{n-1} - t_n}{D} \right\rceil$;
    $f_{n,1} := t_n + l_n + \Delta_n D - t_{n-1}$;
  end ;
end

図 7　並列パイプライン方式の第 1 ステージにおける処理

3.3　各プロセッサの内部処理

以下に本構成の各プロセッサが周期ごとに行なう処理を述べる．なお，いずれの処理においても反復処理を行なわないので，計算量 $O(1)$ が実現できる．

マルチプロセッサシステムの前 $(\log_2 N)$ 段のプレフィクス演算部では，そのパイプラインステージを用いて並列パイプラインプレフィクス演算を行なう．以降では，各プロセッサの処理を述べる．第 1 段は並列プレフィクス演算の第 1 回目の処理に用いられる．$(N-1)$ 個のプロセッサ P_{1n} $(2 \leqq n \leqq N)$において，$(\log_2 N+1)$周期にポート n，$(n-1)$へ到着するパケットの情報が入力される．それらの情報を以下に示す．

- l_n, t_n: ポート n に到着するパケットの長さと到着ギャップ (2.参照)
- l_{n-1}, t_{n-1}: ポート $(n-1)$ に到着するパケットの長さと到着ギャップ

ポート n にパケットが到着しなければ，$l_n = t_n = 0$ とする．プロセッサ P_{1n} では，図 7に示す処理を行ない，着目した周期において 2 つのポートに到着するパケットの連結パケットの長さと到着ギャップを示す値 $f_{n,1}$，$t_{n,1}$ を出力する．それらの値は直後のレジスタに格納される．

プレフィクス演算部の第 2 段以降は一般化できる．これを第 k 段 $(2 \leqq k \leqq \log_2 N)$ としてその処理を述べる．第 k 段の $(N - 2^{k-1})$ 個のプロセッサでは，並列プレフィクス演算の

for each processor n, in parallel ($n := 2^{k-1}+1$ to N)
begin
  if ($f_{n-2^{k-1},k-1} = 0$) **then** $t_{n,k} := t_{n,k-1}$
  else $t_{n,k} := t_{n-2^{k-1},k-1}$;
  if ($f_{n,k-1} = 0$) **then** $f_{n,k} := f_{n-2^{k-1},k-1}$
  else begin
    $\Delta_n := \left\lceil \dfrac{t_{n-2^{k-1},k-1} + f_{n-2^{k-1},k-1} - t_{n,k-1}}{D} \right\rceil$;
    $f_{n,k} := t_{n,k-1} + f_{n,k-1} + \Delta_n D - t_{n-2^{k-1},k-1}$;
  end ;
end

図 8　並列パイプライン方式の第 k ステージにおける処理

for each processor n, in parallel ($n := 1$ to N)
begin
  if ($l_n = 0$) **then** exit;
  if ($f_{n-1,k} = 0$) **then** $q' := q$;
  else $q' := \left\lceil \dfrac{q - t_{n-1,k}}{D} \right\rceil + t_{n-1,k} + f_{n-1,k}$;
  $\Delta_n := \left\lceil \dfrac{q' - t_n}{D} \right\rceil$;
  if ($\Delta_n < B$) **then** Packet n is given delay $\Delta_n D$;
  else Packet n is discarded;
end

図 9　並列パイプライン方式における遅延計算処理

第 k 回目の処理を行なう．プロセッサ P_{kn} ($2 \leqq k \leqq \log_2 N$, $2^{k-1}+1 \leqq n \leqq N$) において，その入力部は，$(k-1)$ 段目のプロセッサ $P_{k-1,n}$ に接続するレジスタと $P_{k-1,n-2^{k-1}}$ に接続するレジスタとに接続している．プロセッサ P_{kn} は，値 $f_{n,k-1}$ および $t_{n,k-1}, f_{n-2^{k-1},k-1}, t_{n-2^{k-1},k-1}$ を受取り，図 8に示す処理に従ってポート$\max(n-2^k+1, 1)$ から n に到着するパケットの連結パケットの長さと到着ギャップを表わす値 $f_{n,k}, t_{n,k}$ を出力する．それらの値は直後のレジスタに格納される．

前 $(\log_2 N)$ 段において，上記の処理をパイプライン方式で行なうことで，第 $(\log_2 N)$ 段では，ポート 1 から n $(n = 1, 2, \ldots, N)$ までのプレフィクス和を出力する．ここでのプレフィクス和は，ポート 1 から n に到着したパケットの連結パケットの長さと到着ギャップ，言いかえれば，ポート $(n+1)$ へ到着するパケットへ与える相対遅延である．もともと入力されている情報 (l_n, t_n) も遅延決定部で用いるために，並行して別のレジスタに格納される．

次に，$(\log_2 N + 1)$ 段における遅延決定部の処理を述べる．N 個のプロセッサが遅延を求めるために使われる．プロセッサ $P_{(\log_2 N+1),n}$ ($1 \leqq n \leqq N$) の入力部は $(\log_2 N)$ 段のプロセッサ $P_{\log_2 N,n-1}$ の直後のレジスタに接続しており，プロセッサ $P_{(\log_2 N+1),n}$ は値 $f_{n-1,\log_2 N}$ および $t_{n-1,\log_2 N}$ を受取る．同時にプロセッサは到着情報 l_n, t_n とバッファ占有度 q も受取る．着目したポートにパケットが到着する場合，すなわち，$l_n \neq 0$の場合，プロセッサ $P_{(\log_2 N+1),n}$ は，バッファ占有度と相対遅延を用いて，ポート 1 から$(n-1)$に到着した全パケットがバッファから出力する時刻を表わす q' を求める．その後，ポート n へのパケットの遅延時間 $\Delta_n D$ を求める．最後に，$\Delta_n < B$であればパケットに対して遅延時間 $\Delta_n D$ を与え，そうでなければパケットを棄却する．それらの処理を図 9に示す．

遅延決定部では，同時にプロセッサ P_q においてバッファ占有度を更新する．プロセッサの入力は，プロセッサ $P_{\log_2 N,N}$ の直後のレジスタに接続される．プロセッサは全ポートに到着するパケットの連結パケットの長さと到着ギャップを示す値 $f_{N,\log_2 N}, t_{N,\log_2 N}$ を受取り，図 10に示す処理に従ってバッファ占有量 q を更新する．

並列パイプライン処理を用いると，図 3に示した逐次処理と比べて N 倍高速なバッファ管理を行えるが，一方，バッファ占有量を安全側に見積る可能性がある．プロセッサ集合 $\{P_{(\log_2 N+1),n}\}$ がバッファあふれによりパケットの一部を棄却する場合でも，プロセッサ P_q はすべてのパケットがバッファに格納されると仮定してバッファ

if ($f_{N,k} = 0$) **then** $q := \max(q - T, 0)$
else begin
  $\Delta := \left\lceil \dfrac{q - t_{N,k}}{D} \right\rceil$;
  if ($\Delta < B$) **then**
    $q := \max\left(t_{N,k} + f_{N,k} + \Delta D - T, 0\right)$;
  else $q := q - T$;
end

図 10　並列パイプライン方式のバッファ占有度の更新処理

表 1　バッファ管理装置の諸元

動作周波数	$(1/T)$	78.2 MHz
回線速度	(C)	40.0 Gbps
入力回線数	(N)	8
最小パケット長	$(l_{\min})$	バイト
最大パケット長	(MTU)	2,047 バイト
遅延線数	(B)	31
遅延線単位長	(D)	3.125m (64 バイト相当)

占有量を更新するからである．例えば，図 6において，最初の 2 つのパケットが格納された後にバッファが一杯になった時でも，残りの 6 パケット分の長さがバッファ占有量として追加される．しかし，上記の見積りによる性能の劣化は無視できる程度である [21].

4. ハードウェア実現性の検証

本章では，提案した並列パイプライン処理に基づきバッファ管理装置回路を設計し，ハードウェア規模と速度の実現性を検証する．実装のかわりに，設計したバッファ管理装置を 0.22μm FPGA デバイスに配置配線処理したイメージを用いてゲートレベルシミュレーションを行なう．そのために提案アルゴリズムを，ハードウェア記述言語 (HDL) 用に書き直している．表 1にバッファ管理装置回路の諸元を示す．表に示すとおり，バッファ管理装置回路は動作周波数 78.2MHz で動作する．これを回線速度 (実行帯域) に換算すると 40Gbps になる．回線あたりの実効帯域 C (Gbps) は，本装置の動作周波数 $f_{\max}$ (MHz) と最小パケット長 $l_{\min}$ (バイト) を基にして，以下の式 (2) で与えた．

$$C = 8 l_{\min} \times f_{\max} \times 10^{-3} \qquad (2)$$

次に，光パケットスイッチのポート数をどこまで増やせるかを，バッファ管理装置回路実現性の面から検討する．ここでは 79,040 個の論理セルが集積された最新の 0.13μm FPGA デバイスを対象とする．論理合成ソフトウェアによってバッファ管理装置回路に必要な論理セル数を見積り，その結果を図 11に示す．図中 “Parallel” で示した特性が提案方式によるバッファ管理装置回路に必要な論理セル数である．マルチプロセッサシステムは装置規模 $O(N \log_2 N)$ なので，参考のために関数 $(64N \log_2 N)$ もプロットした．

図より，N の増加に対する論理セル数の増加の割合は，参考とした関数の増加割合よりも小さく，また，その関数は$N = 128$においても，対象 FPGA の論理セル数より小さいことがわかる．したがって，我々が提案するバッファ管理装置回路は 128 ポートのパケットスイッチをサポートできる．0.13μm FPGA は 0.22μm FPGA よりも高速動作するので，本バッファ管理装置も回線速度 40Gbps での動作が期待できる．したがって，本結果はバッファ管理装置以外の機能や動作速度を無視した結果ではあるが，回線速度 40Gbps の 128×128 パケットスイッチのサポートができる．本バッファ管理装置を用いることによって，光パケットスイッチは，最新の回線速度 10Gbps の 32×32 IP ルータの 16 倍のスループットが得られる．クリティカルパスの最適化や ASIC の導入によってさらに高速な回線速度のポートや高スループットの光パケットスイッチをサポートするバッファ管理装置の実現も期待できる．

5. スケジューリング特性

本章では，シミュレーションによって提案する並列パイプライン

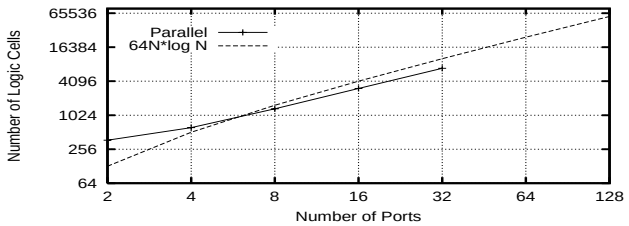

図 11　ポート数ごとの実現可能性の検証

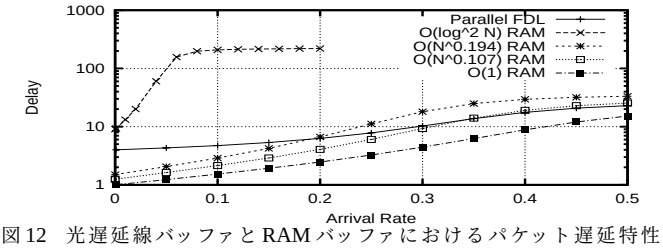

図 12　光遅延線バッファと RAM バッファにおけるパケット遅延特性 ($B = 25, D = 64$ バイト)

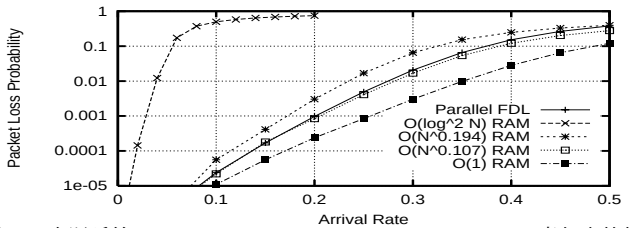

図 13　光遅延線バッファと RAM バッファにおけるパケット棄却率特性 ($B = 25, D = 64$ バイト)

処理方式の有効性を示す．比較の対象は，RAM バッファを有する電子処理によるパケットスイッチのバッファ管理方式である．光遅延線バッファを用いると Void 空間があるためにバッファの利用効率が悪くなる．しかし，この効率劣化は高速バッファ管理によって無視できることを示す．

シミュレーションでは，光パケットスイッチの一出力ポートに着目する．そこに対して 10^6 パケットを発生させた．パケットの到着はポアソン過程に従う．パケット長は 64 バイトから 1500 バイトまでであり，平均長は 141.6 バイトである．パケット長の分布はほぼ指数的であり，下記のように発生させた．まず，平均 128 バイトの指数分布に従って変数を発生させ，最小 (最大) パケット長よりも小さな (大きな) 値は 64 バイト (1500 バイト) に変更した．提案方式では，到着率 0.45 で負荷がほぼ 1 になる．

RAM バッファを用いた実行可能な方法は文献 [9] における計算量 $O((\log_2 N)^2)$ のスケジューリングであるが，それは，提案方式と比べてかなり低速である．したがって，RAM バッファには架空のスケジューリングである $O(N^{0.194})$，$O(N^{0.107})$，$O(1)$ の 3 方式も参照のために用いた．ポート数は$N = 8$で評価する．

図 12，13にそれぞれパケットの遅延と棄却率を示す．遅延時間はスケジューリングに要する時間とバッファ滞在時間の和とし，最小パケット長 64 バイトで正規化した．回線速度 40Gbps における単位時間は 12.8nsec に相当する．図の横軸には提案方式における最小パケット長相当の時間を単位としたパケットの到着率を示す．図中 "$O(x)$" とラベルされた方式によるパケットスイッチのポート速度は，処理速度を考慮して提案する並列パイプライン方式の $1/x$ 倍にしている．図からわかるように，光遅延線バッファであっても提案方式によって，電子処理のパケットスイッチに既存の $O((\log_2 N)^2)$ スケジューリング方式を用いた場合の性能を上回る．電子処理のパケットスイッチが光パケットスイッチとほぼ同等の性能を示すのは，$O(N^{0.107})$ スケジューリングを用いる場合である (低負荷における遅延特性を除く)．以上より，光遅延線バッファでは，Void 空間を生じるが，提案方式によるスケジューリングを用いることで，電子処理によるパケットスイッチよりも良好な性能を得ることがわかる．

6. ま と め

本稿では，出力バッファ型光パケットスイッチにおけるバッファ管理を高速化するために，並列処理とパイプライン処理を組みあわせた制御方式を提案した．本方式は複数プロセッサ構成で，各プロセッサの計算量は，ポート数に依存しない $O(1)$ なので，従来の逐次処理方式の N 倍の高速化を図れる．シミュレーションにより，回線速度 40Gbps の 8×8 光パケットスイッチにおいて，非同期到着する最小 64 バイトの長さの異なるパケットを処理する機能を FPGA に実装できることを確認した．また，最新ルータの 16 倍のスループットとなる回線速度 40Gbps の 128×128 光パケットスイッチにおけるバッファ管理装置のサポートが提案方式と既存の FPGA 技術を用いて可能であることを示した．

文　献

[1] Juniper Networks *available from "http://www.juniper.net/"*.

[2] N. Wada *et al.*, "40Gbit/s interface, optical code based photonic packet switch prototype," *OFC 2003*, pp. 801–802, Mar. 2003.

[3] M. Duelk *et al.*, "Fast packet routing in a 2.5 Tb/s optical switch fabric with 40 Gb/s duobinary signals at 0.8 b/s/Hz spectral efficiency," *OFC 2003 Post Deadline* (PD8), Mar. 2003.

[4] N. Wada *et al.*, "Photonic packet routing based on multi-wavelength label switching using fiber Bragg gratings," *ECOC 2000*, pp. 71–72, Sep. 2000.

[5] K. Kitayama and N. Wada, "Photonic IP routing," *IEEE Photonic Tech. Letters*, vol. 11, pp. 1689–1691, Dec. 1999.

[6] K. Habara *et al.*, "Large-capacity photonic packet switch prototype using wavelength routing techniques," *IEICE Trans. Commun.*, vol. E83-B, pp. 2304–2311, Oct. 2000.

[7] N. McKeown, "The *i*SLIP scheduling algorithm for input-queued switches," *IEEE/ACM Trans. Networking*, vol. 7, pp. 188–201, Apr. 1999.

[8] R. Sivaram *et al.*, "HIPIQS: A high-performance switch architecture using input queueing," *IEEE Trans. Parallel and Distributed Systems*, vol. 13, pp. 275–289, Mar. 2002.

[9] A. Prakash *et al.*, "An $O(\log^2 N)$ parallel algorithm for output queueing," *Proc. IEEE INFOCOM 2002*, pp. 1623–1629, June 2002.

[10] T. H. Cormen *et al.*, "Algorithms for parallel computers," *Introduction to Algorithms*, ch. 30, MIT Press, 1989.

[11] J. Jájá, *An Introduction to Parallel Algorithms*. Addison Wesley, 1992.

[12] "WAN packet size distribution," *available from "http://www.nlanr.net/NA/Learn/packetsizes.html"*.

[13] D. Hunter and I. Andonovic, "Approaches to optical Internet packet switching," *IEEE Commun. Mag.*, vol. 38, pp. 116–122, Sep. 2000.

[14] T. Sakamoto *et al.*, "Demonstration of an optical packet synchronizer for an optical packet switch," *OFC 2002*, pp. 762–763, Mar. 2002.

[15] M. Murata and K. Kitayama, "Ultrafast photonic label switch for asynchronous packets of variable length," *Proc. IEEE INFOCOM 2002*, pp. 371–380, June 2002.

[16] L. Tancevski *et al.*, "Optical routing of asynchronous, variable length packets," *IEEE J. Select. Areas in Commun.*, vol. 18, pp. 2084–2093, Oct. 2000.

[17] A. Ge *et al.*, "WDM fiber delay line buffer control for optical packet switching," *SPIE Vol. 4233 (OptiComm 2000)*, pp. 247–256, Oct. 2000.

[18] N. Wada *et al.*, "Photonic variable length packet routing based on multi-wavelength label switch using multi-section fiber Bragg gratings and supercontinuum light source," *OAA/BGPP 2001*, vol. JW4, July 2001.

[19] F. Callegati *et al.*, "Exploitation of DWDM for optical packet switching with quality of service guarantees," *IEEE J. Select. Areas in Commun.*, vol. 20, pp. 190–201, Jan. 2002.

[20] 原井, 村田, "出力バッファ型光パケットスイッチにおける並列バッファ管理方式," 信学技報 (IN2003-42), pp. 1–6, July 2003.

[21] H. Harai and M. Murata, "An $O(1)$ parallel and pipeline algorithm for output-buffer management in photonic packet switches," *submitted to IEEE INFOCOM 2004*, 2003.