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.

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

— 同期到着固定長パケットの場合 —

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制御アルゴリズムを提案する．ポート数 N に依存しない計算量 $O(1)$ の提案アルゴリズムにより，従来の N 倍の高速化を図れる．回線速度 40Gbps の 8×8 光パケットスイッチにおいて，64 バイトの固定長パケットを処理できる機能を 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 512×512 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，同期固定長パケット，FPGA

A Parallel and Pipeline Algorithm for Output-Buffer Management in Photonic Packet Switches

— For a case of Synchronous, Fixed-Size Packets —

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 an $O(1)$ algorithm based on parallel and pipeline processing for this purpose. The proposed 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 synchronously arriving 64byte fixed-size packets.

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

1. はじめに

大量のトラヒックを転送する基幹ネットワークを構築するには，リンク容量の増加のみならず，ノードにおけるパケット転送能力 (ノードスループット) を改善する必要がある．現在，パケット転送には電子処理技術が用いられている．電子処理技術を用いると，ムーアの法則に従った LSI 技術の進展や大規模分散／並列システムの構築によりパケット転送能力を向上することはできる．しかし，リンク容量の増加は光ファイバを束ねることで簡単にできるのに比べ，集積化や並列処理によるノードスループットの増加は限界を生じやすい．ノードスループットを増加させるには，ネットワークレイヤの下位に光技術を用いた新たなレイヤを導入する方法がある．現在では，光パスネットワーク [1] を GMPLS (Generalized Multi-Protocol Label Switching) [2] で制御する手法が有望である．しかし，その実現には，複雑なトラヒックエンジニアリングが必要であり，かつ，帯域の粒度が粗いという問題がある．したがって，我々は，パケット交換への光技術の導入を目指す．

高スループットの光パケットスイッチを実現するには 2 つの方法が考えられる．現状の回線速度のポート数を増やしたスイッチを構築するか，または，回線速度を増したスイッチを構築するかである．

最近では 10Gbps の回線を 32 ポート備えた IP ルータ [3] 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，現状の光技術ですでに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 [4], [5] を対象としたバッファ管理方式の検討を行なう．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 (アドレス検索，フォワーディング)，交換 (スイッチング)，バッファ管理 (キュー管理，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，ラベル処理 (多波長ラベル処理 [6]，光位相符号ラベル処理 [7])，交換，バッファ [4], [8] 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 [4]．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．電子処理性能がこのまま向上しつづけるとは限らず，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．

筆者らは，出力バッファ型 $N \times N$ パケットスイッチを対象としている．出力バッファ方式は，入力バッファ方式と比べて，より良好な遅延特性およ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 $1 \times N$ スイッチを束ねて構成しているので，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 [9] 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している．

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 (逐次処理方式) 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．今後の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 $O(N)$ よりも高速なバッファ管理アルゴリズムは検討されていない．

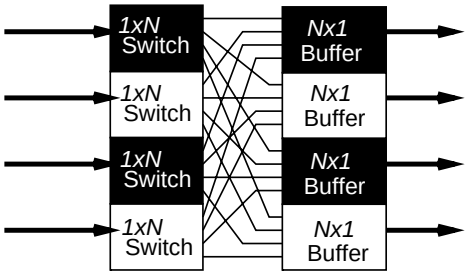

図 1　$N \times N$ 光パケットスイッチ ($N = 4$)

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．まず，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 [9], [10]．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のが一般的になっている．しかし，先述のように，複数入力バッファ方式を光パケットスイッチに適用することは困難である．

文献 [11] 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 [12], [13] を行ない，計算量 $O(\log_2 N)$ の高速化を達成している．対象とするパケットは固定長／同期である．

本稿では，出力バッファ型光パケットスイッチにおけ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に基づいた計算量 $O(1)$ のアルゴリズムを提案する．そのために，プロセッサ数 (装置規模) $O(N \log_2 N)$ のマルチプロセッシング構成を提案する．ただし，N はパケットスイッチのポート数である．本構成は二つの部分，プレフィクス演算部および遅延決定部から成る．プレフィクス演算部では，「到着時にバッファにパケットはなく，ポート ID 順のラウンドロビンスケジューリングにしたがって，すべてのパケットがバッファに格納される」と想定して，到着したパケットの相対的な遅延時間を求める．ここで，$O(N \log_2 N)$ 個のプロセッサを用い，並列プレフィクス演算 [12], [13] をパイプライン方式で実行させ

ることで，計算量 $O(1)$ の高速化を達成する．遅延決定部では，現在のバッファ占有量とプレフィクス演算部から送られた相対的な遅延時間から，全 N ポートに到着するパケットの遅延を同時に決定する．また，バッファ占有量も同時に更新する．提案するアルゴリズムの計算量は $O(1)$ であり，従来の光パケットスイッチの管理方式よりも N 倍高速である．また，従来のパケットスイッチの管理方式 [11] よりも $(\log_2 N)^2$ 倍高速である．

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ゲートレベルのハードウェアシミュレーションにより，固定長 64 バイトのパケットが同期して到着する，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可能であることを確認した．さらに，電子技術による最新の IP ルータの少なくとも 64 倍の性能となる，回線速度 40Gbps の512 × 512 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す．

本稿の構成を以下にのべる． 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．最後に 5.において，まとめと今後の課題を述べる．

2. 出力バッファ型光パケットスイッチ

2.1 光パケットスイッチの概略

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$N \times 1$バッファからなる．$1 \times N$ パケットスイッチと$N \times 1$バッファはメッシュ状に光接続される．$1 \times N$スイッチでは，光宛先検索機能 [6], [7] により高速宛先検索が可能である．その結果，本スイッチは高ノードスループットが期待できる．ただし，パケットの衝突を避け棄却率を改善するためにはバッファを必要とする．そこで，$N \times N$ パケットスイッチの各出力ポートに光バッファが接続される．我々が$N \times 1$ バッファに用いるのは，光メモリではなく，光遅延線バッファである．図 2(a) に4×1 光バッファ構成を示す．各バッファは，B 本の光ファイバ遅延線 (図は$B = 4$)，および，$N \times (B + 1)$ 光空間スイッチ，光カプラ，バッファ管理装置からなる．B 本のファイバ遅延線 $\{d_0, d_1, \ldots, d_{B-1}\}$ の長さは単位長 D の倍数 $(0, D, \ldots, (B-1)D)$ である．本光バッファは，0 から $(B-1)D$ までの離散時間の遅延を生じる．

本稿では，すべてのパケット長が等しく，パケットは同期してバッファに到着することを想定している．パケットスイッチ内部でパケットを固定長パケットに分解する場合には，入力部にその機能を加える．また，同期を行なうために，その後に光同期システム [14] 〜 [16] を配置する．

2.2 バッファ管理装置の振舞い

IP ルータなどの電子技術によるノードシステムに実装されている RAM と異なり，遅延線バッファを用いると，光の直進性のため古典的な蓄積交換型の手法を用いることが難しい．遅延線を光バッファとして用いるには，各光パケットが光バッファに到着する前に，光パケットが適切な遅延線に進むように選択せねばならない．$N \times N$出力バッファ型光パケットスイッチには，同じ周期に最大 N パケットが同一の出力ポートを目指して到着する．したがって，そのパケット長を L とすると，L に相当する時間以内に N パケットの遅延を求めるバッファ管理装置が必要になる．本稿では，逐次処理 (ラウンドロビン処理) によってポート順に光パケットを処理する方法を基本とする．

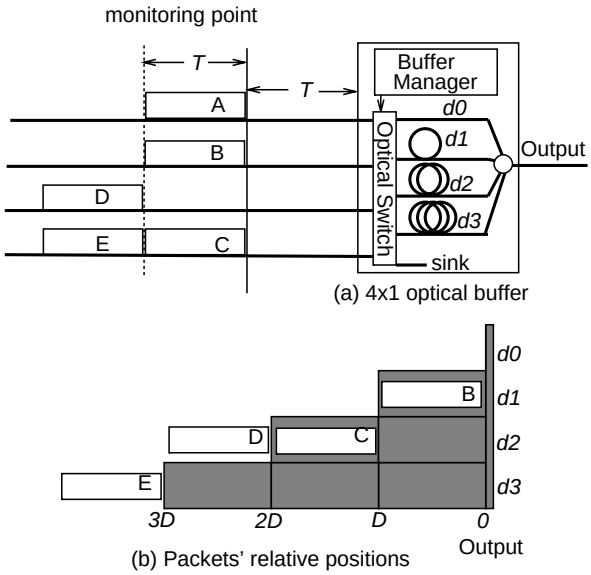

図 2 光バッファに到着したパケットに対するバッファ管理装置の振舞い．(a) 光遅延線バッファ ($N = 4, B = 4$) とパケットの到着例． (b) バッファ内におけるパケットの相対位置

光バッファに到着した光パケットは，バッファ管理装置が制御する光スイッチによって適切な遅延線を通過した後，光バッファから出力される．ただし，$N \times (B + 1)$ 光スイッチの最下部のポートに送られた光パケットは廃棄される．

次章における並列パイプラインバッファ管理方式を理解するために，ここでは，その方式の基になる逐次処理方式を用いたバッファ管理方式のパケットへの振舞いを述べる．先述の図 2 には，4×4 光パケットスイッチの 1 バッファに到着するパケット A から E を示している．バッファ管理装置は逐次処理方式の周期にあたる内部クロック (周波数 $1/T$) を持つ．各周期 kT $(k = 1, 2, \ldots)$ において，バッファ管理装置は，次周期 (時刻 $(k + 1)T$) に光バッファ内の光スイッチに到着する光パケットの到着情報を受け取る．バッファ管理装置は，時間 T の間に最大$N = 4$パケットのそれぞれに対して遅延時間を求める．連続して到着するパケットを扱うためには，周期 T はパケット長 L に相当する時間以下でなければならない ($T \leqq L$)．以降では，$L = T = D$とする．

以降では，バッファ管理装置の振舞いを詳述する．l_n をポート n におけるパケットの到着の有無 (到着時 $l_n = 1$，非到着時 $l_n = 0$) とする．バッファ管理装置は変数 q を管理する．これは，バッファに存在するパケットがすべて出力される時刻と現時刻との差，すなわち，バッファ占有量である．同期固定長パケットの場合，q はパケット長

```
for n := 1 to N do
begin
  if (l_n = 1) then begin
    if q < B then begin
      Packet n is given delay qD;
      q := q + 1; end
    else Packet n is discarded;
  end
end
q := max(q − 1, 0);
```

図 3　逐次処理を実現するための擬似コード

で正規化された値，すなわち，パケット数である．バッファ管理装置は，各周期において到着するすべてのパケットの遅延を，ポート $1, 2, \ldots, N$ の順に計算する．バッファ占有量 q が遅延線数 B より小さければ ($q < B$)，ポート n のパケットには，遅延 $q \times D$ が与えられる．一方で遅延線数以上であれば，そのパケットは棄却される．パケットがバッファに格納される場合には，次ポート以降のパケットを適切に処理するために，バッファ占有量を $q \leftarrow q + 1$ と更新する．すべてのポートのパケットの遅延を計算した後に，次周期のパケットの遅延を衝突なく最小限にするために，$q \leftarrow \max(q - 1, 0)$ と更新する．図 3に同期固定長パケットを処理するための擬似コードを示す．先述のように，パケット長および，処理の周期は遅延線の単位長 D に等しい．また，バッファ占有量 q も D で正規化している．“Packet n is given delay qD” とは，パケットを遅延線 d_q に送ることを表わす．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．

図 2 (a) に示すように，3 パケット A，B，C が同一周期に到着し，次周期に 2 パケット D，E が到着する場合を想定する．バッファ管理装置は，それぞれの最初の 3 パケットをそれぞれ遅延線 d_0, d_1, d_2 に送るよう決定する．また，次周期の 2 パケットは，それぞれ遅延線 d_2, d_3 に送られる．図 2 (b) には，パケット A が光バッファから出力された直後の，光バッファ内における残りの 4 パケットの相対的な位置を示す．図に示すように，すべての光パケットが衝突なく光バッファから出力されることがわかる．

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，固定長光パケットが同期して到着する光パケットスイッチのバッファ管理に用いる計算量 $O(1)$ の並列パイプラインアルゴリズムを提案する．同期を行なうには，光同期システム [14] 〜 [16] がスイッチの入力部に必要になり，光システムの規模が増大する．また，可変長パケットを固定長に直す処理がネットワークの入口，または，スイッチの入力部に必要になる．しかし，バッファ管理においては，入力パケットを 1 または 0 で扱えるため，可変長パケットや非同期到着の場合よりも遅延を求めるための電子処理が簡素になる．

3.1 マルチプロセッサシステムとその機能分担

我々の提案するアルゴリズムは，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，$\left(s_n = \sum_{i=1}^{n} a_i \right)$ 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 [12], [13]．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．

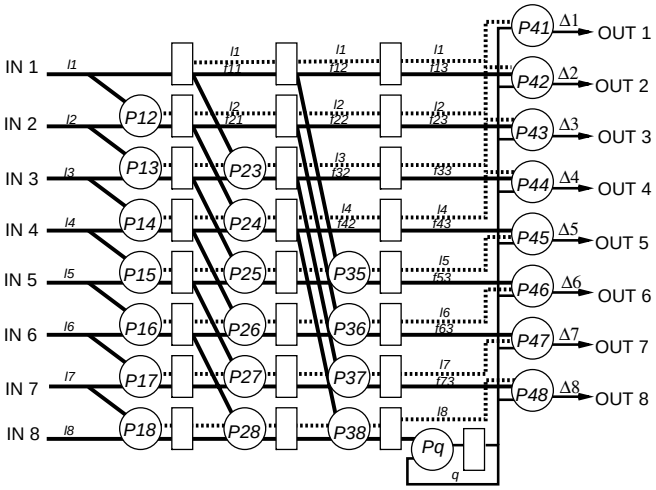

図 4　並列パイプライン処理アーキテクチャ ($N = 8$)

図 4に，$N = 8$ におけるアルゴリズムを実現するための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，「パケット到着時にパケットがバッファになく，到着したパケットは逐次処理によってポート順に処理される」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．固定長パケットが同期到着する本稿の場合では，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数に等しい．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てきた相対遅延から，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 は，バッファ占有量を更新するために用いる．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ために用いるバスを示し，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ためのバスを示す．

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ためには，$(\log_2 N+1)$ 周期を要する．バッファ管理装置では，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せねばならない．

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，$(\log_2 N+1)T$ 時間後にポート n へパケットが到着する場合には “1” 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あたえられる．

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，すべての処理では反復処理を行なわないので，計算量 $O(1)$ が実現できる．

3.2 各プロセッサの内部処理

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，$(\log_2 N+1)$周期後におけるポート n，$(n-1)$へのパケット到着の有無を示す値 l_n と l_{n-1} が入力される．それらの値はパケットが到着するなら 1 で，到着しなければ 0 である．プロセッサ P_{1n} では，下記の処理を行ない，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数の和を示す値 $f_{n,1}$ を出力する．値 $f_{n,1}$ は直後のレジスタに格納される．

for each processor P_{1n}, in parallel ($n := 2$ to N)
  $f_{n,1} := l_n + l_{n-1}$;

次に，第 2 段のプロセッサの処理を述べる．$(N-2)$ 個のプロセッサでは，並列プレフィクス演算の第 2 回目の処理を行なう．プロセッサ P_{2n} $(3 \leqq n \leqq N)$ において，その入力部は第 1 段のプロセッサ P_{1n} に接続するレジスタと $P_{1,n-2}$ に接続するレジスタとに接続されている．プロセッサ $P_{2,n}$ は，値 $f_{n,1}$ と $f_{n-2,1}$ を受取り，以下の処理にしたがって，ポート $\max(n-3,1)$ から n に到着するパケットの数を表わす値 $f_{n,2}$ を出力する．値 $f_{n,2}$ は直後のレジスタに格納される．

for each processor P_{2n}, in parallel ($n := 3$ to N)
  $f_{n,2} := f_{n,1} + f_{n-2,1}$;

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されている．プロセッサ P_{kn} は，値 $f_{n,k-1}$ および $f_{n-2^{k-1},k-1}$ を受けとり，以下の処理にしたがってポート $\max(n-2^k+1,1)$ から n に到着するパケットの数を表わす値 $f_{n,k}$ を出力する．値 $f_{n,k}$ は，直後のレジスタに格納される．

for each processor P_{kn}, in parallel ($n := 2^{k-1}+1$ to N)
  $f_{n,k} := f_{n,k-1} + f_{n-2^{k-1},k-1}$;

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数，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) 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，$(\log_2 N+1)$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されており，プロセッサ $P_{(\log_2 N+1),n}$ は値 $f_{n-1,\log_2 N}$ を受けとる．同時にプロセッサは到着情報 l_n とバッファ占有度 q も受け取る．プロセッサ $P_{(\log_2 N+1),n}$ は以下の処理にしたがって，ポート n におけるパケットの遅延を表わす値 Δ_n を出力する．

for each processor n, in parallel ($n := 1$ to N)
begin
  if ($l_n = 0$) **then** exit;
  $\Delta_n := q + f_{n-1,\log_2 N}$;
  if ($\Delta_n < B$) **then** Packet n is given delay $\Delta_n D$;
  else Packet n is discarded;
end

最終段では，プロセッサ P_q においてバッファ占有度を同時に更新する．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ている．プロセッサは値 $f_{N,\log_2 N}$ を受け取り，以下の式 (1) にしたがって バッファ占有量 q を更新する．

$$q := \max\left(\min(q + f_{N,\log_2 N}, B) - 1, 0\right). \qquad (1)$$

図 3の最終行において，次の周期の処理のためにバッファ占有量を更新する際に用いられる関係 $(q-1)$ は，式 (1) では $\min(q+f_{N,\log_2 N},B)-1$ となっている．この変更は，以下に述べるように，提案する処理と逐次処理とにおけるバッファ占有量の更新結果を一致させるために必要な関係である．2.に述べたように，逐次処理においてパケットが棄却される場合には，バッファ占有量は更新されない．一方，提案する並列パイプライン方式では，値 $f_{N,\log_2 N}$ は到着したパケット数を示しており，一部のパケットは棄却されるかもしれない．そこでバッファ占有量を見積りすぎないように，上記の関係を導入している．

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アルゴリズムに基づいた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．ここでは同期到着する固定長光パケットを処理するものを扱う．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

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
パケット長	(L)	64 byte
遅延線数	(B)	15
遅延線単位長	(D)	3.125m (64 バイト相当)

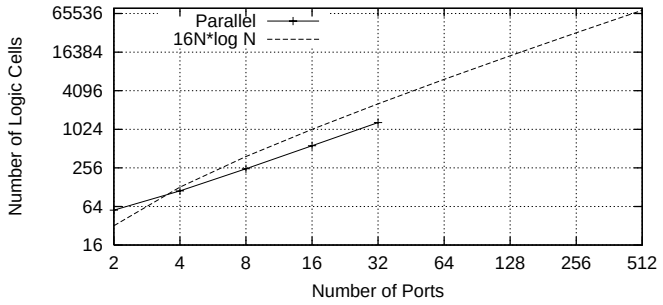

図 5　ポート数ごとの実現可能性の検証

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．なお，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) とパケット長 L (バイト) を基にして，以下の式 (2) で与えた．

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

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 5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 $(16N \log_2 N)$ の曲線もプロットした．

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$N = 512$においても，対象とした FPGA における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 512 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，本結果は固定長パケットと可変長パケットの違いや，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 512×512 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 48 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. 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アルゴリズムを提案した．提案アルゴリズムの計算量は，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，同期到着する 64 バイト固定長のパケットを処理できる機能を FPGA に実装できることを確認した．また，回線速度 40Gbps の 512×512 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．今後，光同期システムや固定長パケットへの分解処理を取り除くために，長さの異なるパケットの処理を実現することも重要である．

文　献

[1] I. Chlamtac, A. Ganz, and G. Karmi, “Lightpath communications: An approach to high bandwidth optical WAN’s,” *IEEE Trans. Communications*, vol. 40, pp. 1171–1182, July 1992.

[2] E. Mannie *et al.*, “Generalized multi-protocol label switching architecture (draft-ietf-ccamp-gmpls-architecture-05.txt),” *IETF Internet Draft (Work in Progress)*, Mar. 2003.

[3] Juniper Networks *available from ”http://www.juniper.net/”*.

[4] N. Wada, H. Harai, and F. Kubota, “40Gbit/s interface, optical code based photonic packet switch prototype,” *OFC 2003 Tech. Digest*, pp. 801–802, Mar. 2003.

[5] M. Duelk, J. Gripp, J. Simsarian, A. Bhardwaj, P. Bernasconi, M. Zirngibl, and O. Laznicka, “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.

[6] N. Wada, H. Harai, W. Chujo, and F. Kubota, “Photonic packet routing based on multi-wavelength label switching using fiber Bragg gratings,” *ECOC 2000 (No.10.4.6)*, pp. 71–72, Sep 2000.

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

[8] K. Habara, H. Sanjo, H. Nishizawa, Y. Yamada, S. Hino, I. Ogawa, and Y. Suzaki, “Large-capacity photonic packet switch prototype using wavelength routing techniques,” *IEICE Trans. Commun.*, vol. E83-B, pp. 2304–2311, Oct 2000.

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

[10] R. Sivaram, C. B. Stunkel, and D. K. Panda, “HIPIQS: A high-performance switch architecture using input queueing,” *IEEE Trans. Parallel and Distributed Systems*, vol. 13, pp. 275–289, Mar. 2002.

[11] A. Prakash, S. Sharif, and A. Aziz, “An $O(\log^2 N)$ parallel algorithm for output queueing,” *Proc. IEEE INFOCOM 2002*, pp. 1623–1629, June 2002.

[12] T. H. Cormen, C. E. Leiserson, and R. H. Rivest, “Algorithms for parallel computers,” *Introduction to Algorithms*, ch. 30, MIT Press, 1989.

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

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

[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] T. Sakamoto, A. Okada, M. Hirayama, Y. Sakai, O. Moriwaki, I. Ogawa, R. Sato, K. Noguchi, and M. Matsuoka, “Demonstration of an optical packet synchronizer for an optical packet switch,” *OFC 2002 Tech. Digest*, pp. 762–763, Mar. 2002.