5-2　光パケットスイッチ構成とその制御技術に関する研究

5-2 Studies on Architecture and Control Technology for Optical Packet Switches

原井洋明
HARAI Hiroaki

要旨

インターネットへの適用、スループット 10 Tbps 超をターゲットにした、光パケットスイッチの概略とその構成技術、実用化への要求条件を述べる。また、情報通信研究機構（NICT）における研究を中心に、スイッチアーキテクチャ、全体システムの開発動向、電子制御システムの開発動向を報告する。

In this paper, we first describe overview and advanced technology of optical packet switches (OPS) and requirement for practical use of OPS of which target is the Internet and 10 Tbps throughput. Then, we report switch architecture, recent activities of integrated technology and electronic control systems for OPS.

[キーワード]
光パケット交換，スイッチ構成，占有出力バッファ，バッファ管理
Optical packet switching, Switch architecture, Dedicated output buffer, Buffer management

1　はじめに

1990 年代を中心に光 ATM（Asynchronous Transfer Mode）スイッチとして研究がなされていた光パケットスイッチの研究は、現在もそのターゲットをインターネットに向けて研究が継続している。光パケットスイッチの適用領域は、徐々に方向性が固まりつつある。MPLS（Multi-Protocol Label Switching）のように、IP（Internet Protocol）ネットワークの下位に閉域した光ラベルスイッチネットワークを形成し、そこで光パケットを高速転送するものである。これ自体は IP over ATM において光 ATM スイッチを使うことと大差ない。しかし、ATM は 53 バイトの「同期固定長パケット」を扱うのに対し、今のターゲットは MPLS や IP など、インターネットへの適用をにらんだ 40〜1,500 バイトの「非同期可変長パケット」である。また、スループットのターゲットも、当時の 10 Gbps 超レベルから、現在は 10 Tbps 超に変わっている。

本稿では、上述のインターネットへの適用、スループット 10 Tbps 超をターゲットにした、光パケットスイッチの概略とその構成技術を述べる。また、構成技術に関して、要求性能を述べる。制御技術の動向も述べる。なお、本論の一部は、文献[1]により詳しく紹介している。また、光技術については、[2]にも述べている。

2　光パケットスイッチと要求性能

2.1　光パケットスイッチの機能

最初にパケットスイッチの機能を説明する。図 1 に示すように、パケットスイッチは五つの機能、すなわち、スイッチング（交換；ON-OFF ではなく、方路の切替え）、バッファリング、フォワーディング（方路検索）、バッファ管理（衝突回避処理）、ルーティング（経路制御）に分けられる。本稿の対象は可変長非同期パケット処理であり、同期機能は基本的に不要である。

スイッチング（**2.2**に詳細）とバッファリング

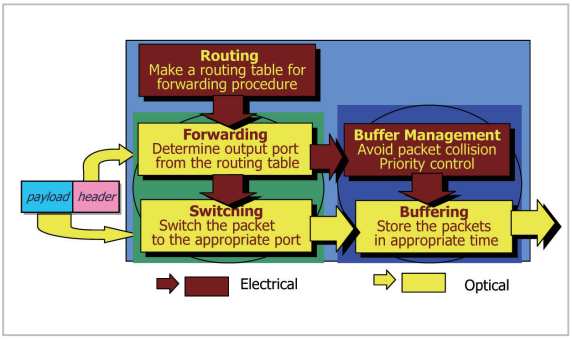

図1 光パケットスイッチの内部機能。矢印は占有出力バッファ構成における制御情報及びデータの動きを示す。

(**2.3** に詳細)は、以前から、光システムによる実装技術が研究されている。パケット交換において、中継ノードでペイロードにアクセスする必要はなく、ペイロードを高速にするほど光システムの利点を享受できるからである。一方、オール電子処理のパケットスイッチでは、ペイロードを速度にするほど内部バスの高速化やメモリアクセス速度の向上などが必要なためにコストがかさむ。

方路検索には、ラベルを、ラベルテーブルを含むメモリと照合する従来型の電子処理による方法が多く取り入れられている[3]-[5]。ただし、回線速度 40〜160 Gbps では、複数プロセッサを用いた処理が必須となる。パケット到着数は回線速度に依存するので、ラベル速度が低速かどうかは関係ない。そこで、方路検索を瞬時にするために、光システムにより実現する方法が考えられている[6]-[8]。これらは光信号が進行しながら方路検索が行えるので、その点で並列処理は不要である。しかし、現状で大量のラベルを処理するには、複数の光システムを用いるので装置規模の点で解決が必要である。

バッファ管理には論理演算を、ルーティングには論理演算とメモリを要するので、光化は困難である。それぞれに要求される時間オーダ(バッファ管理数〜数 10 ナノ秒、ルーティング数秒以上)に適した電子処理をするのが現実的である。ただし、光システムの特性を考慮した処理を行わねばならない。

2.2 スイッチングシステムと要求性能

代表的なスイッチングシステム構成は以下の 3 種類である。

(1) 光空間スイッチにより方路を切り替える方法[5]。

(2) 光カプラ(スプリッタ)と光ゲートによる構成[4][7]。各入力ポートの信号を N 分岐し、それぞれに光ゲートを配置する。それぞれの出力に対しゲートが高々一つ開くように制御することで、完全非閉塞なスイッチを構成する。

(3) 波長変換と AWGR(Arrayed Waveguide Grating Router)による構成[3]。例えば、N をスイッチにつながるファイバ数、W をファイバの波長多重数として、N 個の DEMUX/MUX、NK 個の可変波長変換器(TWC)と固定波長変換器(FWC)、$WK \times WK$ の AWGR により構成される。WK 種類の波長に変換できる TWC を用いると、本スイッチは完全非閉塞な(Strict sense Non-blocking)スイッチになる。

これらの混在型も考えられる。NICT では、(1)又は(2)の構成により開発実証を行っている[7]。

スイッチを大規模にするには、波長変換の切替幅増加、AWGR の分離波長数増加、光ゲート及び光スイッチのパワーロス削減、光スイッチの規模増加が必要となる。例えば、回線(ポート)速度 160 Gbps で 64 回線をスイッチに接続するスループット 10 Tbps の完全非閉塞スイッチの構成に必要な光スイッチの規模は 1×64 であり、波長変換幅は 12.8 THz(102.4 nm)となる。なお、波長変換幅は、160 Gbps(200 GHz とする)の波長で一ポートを構成する時、64 波長が AWGR の入出力に必要として算出した。

ガード時間(最小パケット間隔)を小さくし回線使用効率を高めるために、波長変換やスイッチの切替時間を小さくすることも大切である。例えば、46 バイトの IP パケットを 64 B/66 B 変調した 10 ギガビットイーサネット(物理速度 10.3125 Gbps)で送る場合の回線使用効率は約 53 %である(プリアンブル 8 バイト、イーサネットフレーム 64 バイト、インターフレームギャップ 12 バイトとして算出)。1500 バイトのパケットだと約 94 %であり、計測[9]より求めた平均長付近の 250 バイトに対しては、約 84 %となる。これと同等の回線使用効率を回線速度 40 Gbps の光パケットスイッチで得るには、ガード時間をそれぞれ

7.6 ナノ秒、16.7 ナノ秒、8.9 ナノ秒に抑えないといけない（図 2）。したがって、回線速度の増加を享受できるようにガード時間、すなわち、波長変換の切替速度や光スイッチの切替時間をナノ秒オーダ以下に縮める必要がある。

2.3 光ファイバ遅延線バッファ及び衝突回避

従来の電子制御のパケットスイッチと光パケットスイッチとの本質的な違いは、ストアアンドフォワード型か進行型かである。電子パケットスイッチは、いったんメモリ（RAM）にデータを蓄えてから衝突回避処理やヘッダ処理が行われる。現状、光メモリがないので、光バッファは進行型の光ファイバ遅延線（FDL；Fiber Delay Line）を用いて構成する。光 FDL バッファ（以降、「光バッファ」という。）は、

（1）光スイッチと FDL[5] [7] [8]
（2）光カプラと FDL、光ゲート
（3）波長変換と AWGR、FDL

の 3 通りの方法や、類似した方法で構成できる。

図 3 左に、光スイッチと長さの異なる（単位長 D に比例する）FDL から構成される 4 入力 1 出力、4 通りの遅延を選択できる光バッファの例を示す。光スイッチは同期固定長の場合には再構成可能な非閉塞（Rearrangeable Non-blocking）スイッチでよい。一方、非同期可変長パケットの場合には、完全非閉塞スイッチが必要となる。

バッファ規模に関する考察を以下に述べる。半導体メモリを用いると、コストを考えなければ、今や 10,000 程度のパケットを保持するバッファを作ることは簡単である。一方、光ファイバ遅延線に頼る光パケットスイッチでは、しばらくその規模は望めない。その実現の妨げが光バッファ実現及びその研究に対する障害になっていた。しかし、近年、ネットワーク上位層の助けを必要とするものの、20 程度のパケットを保持できれば十分という見解[10]がある。また、単純に ITU-T の指標や統計データを参考に負荷と平均パケット棄却率だけで考察すると、数十の遅延線があれば十分という見方もある[11]。したがって、まずは、

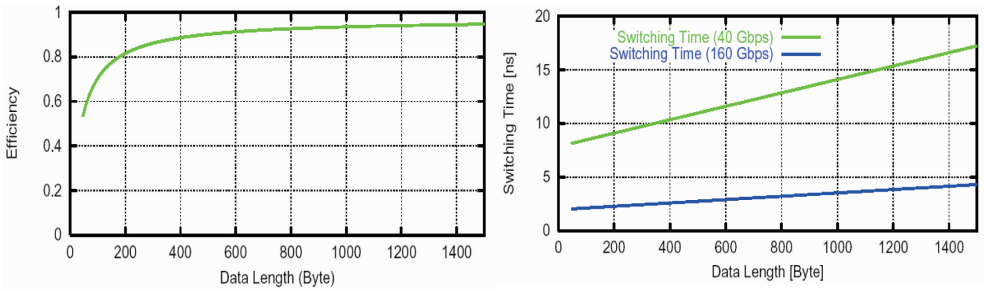

図2 （左）IP パケットをイーサネットフレームにマッピングするときの効率。（右）パケット長が同じ条件で、左図と同等の効率を得るために必要なスイッチング時間。

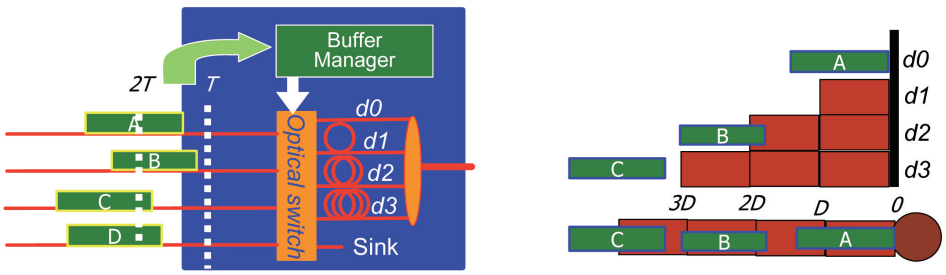

図3 （左）光ファイバ遅延線バッファ。（右）左のパケット到着に対して内部及び出力時にパケット衝突を起こさない遅延線割当例。

その程度の光バッファがあれば実用レベルになる。

なお、光バッファを構成する際には、論理レベルで二つの問題がある。(1) 離散値の遅延時間 (0, D, $2D$, ...) しか与えられない (粒度が粗い) ために起る回線使用効率やパケット棄却性能の劣化(2) HOL (Head-of-Line Priority Queueing) のような割込処理の難しさである。

以下に問題 (1) の具体例を述べる。複数の方路から到着したパケット (図 3 左) は、バッファ管理装置により、スイッチ及び出力時に衝突しないように適切な遅延線が選択される。図 3 右上下は、パケット A、B、C が、それぞれ、遅延線 d_0、d_2、d_3 に転送されたときの出力ポートから見たパケットの相対位置を示す。光バッファが与える遅延は離散値のみなので、連続する二つのパケット間に空きができてしまう。この空きが回線利用率やパケット棄却性能の劣化につながる。

また、メモリがないので、連続して複数のポートから同時にパケットが到着する場合に対処して、最小パケット長に相当する時間内で回線数分のパケットを処理する機能が必要となる。異なる見方をすれば、その処理速度がパケットスイッチの回線速度や回線数を決めることになる。逆から見て、最小パケット長を 64 バイトとし、回線速度 160 Gbps、64 ポートの 10 Tbps のパケットスイッチを実現するためには、3.2ナノ秒で 64 個のパケット処理 (遅延線の選択) が必要になる。

なお、衝突回避のために波長変換が有効で、波長数を増やすほど波長変換によるパケット棄却性能は改善することが今までの研究で明らかになっている。ただし、使用できる波長数は隣接ノードの有する波長数に依存する。多くの報告にあるように、波長変換は遅延線バッファと組み合わせると有効になる。

3 光パケットスイッチの技術動向

本章では、まず、光パケットスイッチのアーキテクチャ技術とプロトタイプ開発例を述べ、電子処理技術の動向を述べる。要素別の光技術については、[1] や本特集の [2] などが詳しい。

3.1 スイッチアーキテクチャ

NICT での研究開発は、図 4 左に示した占有出力バッファ型 $N \times N$ パケットスイッチを対象としている。図は $N=4$ の場合である。図中、ラベルスイッチには図 1 のフォワーディングとスイッチング機能が含まれ、バッファには図 1 のバッファ管理とバッファリング機能が含まれる。図 1 に示した経路制御は本構成の選択に影響を与えないため省略する。ほかに、入力バッファ構成、共有周回バッファ構成 (図 4 右)、共有出力バッファ構成などがある。以下では、NICT において占有出力バッファ構成を用いた理由を述べる。

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及びスループット特性を持つ。これは、HOL (Head of Line) ブロッキングが起こらないためである。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 4 左に示

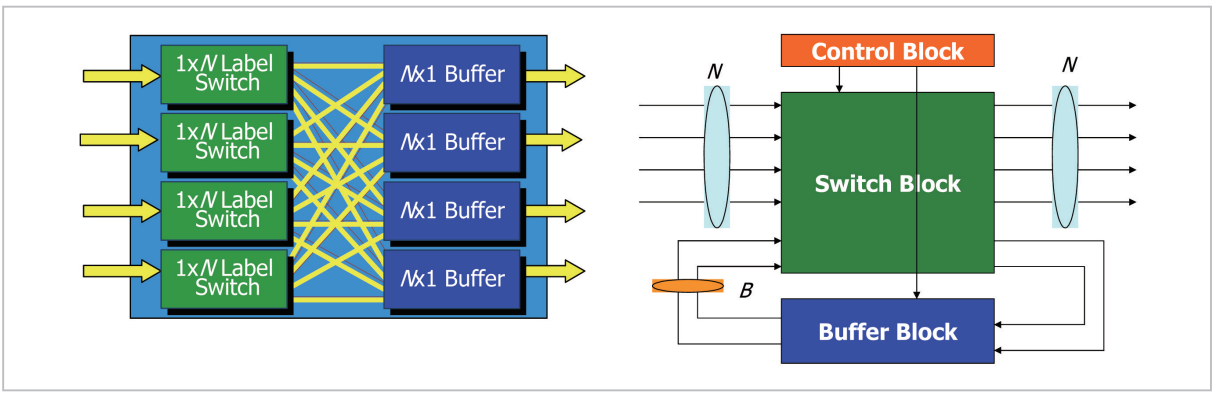

図4 (左) 占有出力バッファ型光パケットスイッチ構成 (制御ブロックは独立にラベルスイッチ、バッファ内に持つ)。(右) 共有周回バッファ型光パケットスイッチ構成。

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（出力調停）機能を必要とする。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。

占有出力バッファ構成は、共有出力バッファ構成や共有周回バッファ構成より、以下の理由から有利である。占有出力バッファ構成は、スイッチ部とバッファ部を完全に切り分け制御できる。具体的には、出力ポートを決める処理はスイッチ部で一度のみでよく、高速の光ラベル処理を生かせる。また、いったんバッファにデータを入れれば出力ポートは一意に定まり、バッファから複数のポートへスイッチする可能性を排除する。したがって、バッファ制御を一出力ポートずつ独立に制御できるので、高速なバッファ制御が可能となる。制御が完全に無視できる状態（制御時間0）では、共有バッファ構成の方が少ないバッファ資源で同等の性能を得られる。しかし、現状、ポート数をNとし、バッファの遅延線数をBとすると、占有出力バッファ構成の方が、NB倍高速なバッファ管理方式[12]を有している。また、占有出力バッファ構成では、スイッチは小規模（$1\times N$や$N\times B$、$1\times B$など）のものを複数使えばよい。しかし、図4右に示すように、共有バッファ構成では、比較上、大きなスイッチ（$(N+B)\times(N+B)$など）を要する。

以上の特長から判断して、先述のとおり、NICTで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。従来、2.4 Gbpsや10 Gbpsベースの光バッファを有する光パケットスイッチの開発例はあった[3][4]。最近では、40 Gbpsベースの開発例もある[5]。NICTでは、2002年に回線速度40 Gbpsベースの光ラベル処理及び光バッファを有する光パケットスイッチの開発に成功した[7]。さらに、2005年に、光ラベル処理及び光バッファ処理を工夫して、回線速度160 Gbpsベースの光パケットスイッチ開発に成功した[8]。

3.2 バッファ制御システムの技術動向

電子ラベル処理に関しては、従来のIPやATMでのアドレス／ラベル処理手法がそのまま応用できると考えられる。また、筆者の知る限り、光パケット処理独特の方式が提案されたという報告は見られない。そこで、以降は、バッファ管理（スケジューリング）の動向を紹介する。

一般に、バッファ管理の研究は、大きく三つに分類できる。

(1) バッファ利用の効率化。さらに三つに分類できる。
　ア　Voidを減らしてリンク使用効率を高める（例えば[13]）。
　イ　波長変換を導入してファイバ使用効率を高める（例えば[14]。[15]はア、イ両方）。
　ウ　光バッファの粒度最適化（例えば[16]）。
(2) 処理の高速化
(3) 優先制御

以降では、処理の高速化及び優先制御について述べる。

・処理の高速化

電子ルータでも指摘されているように、衝突回避処理の速度向上は、パケットスイッチのスループット性能の向上に必須である。筆者らは複数プロセッサを規則的に構成した並列パイプラインアーキテクチャ（図5）をバッファ管理に導入し、同時に全Nポートからパケットが到着しても、プロセッサ当たり1ステップ（計算量$O(1)$）で処理する方式を示した[12][17]。パイプライン処理は有限の$(\log_2 N+1)$ステップ（Nパケットに対するアルゴリズムの計算量$O(\log N)$）である。図5において、丸印がプロセッサを示し、四角印がレジスタを示す。キューの更新を最短パケット長に相当する時間（単位処理時間）当たり一度に抑えるので、単純なラウンドロビン処理（単位処理時間当たり$O(N)$回）よりもN倍のスループットを提供できる。例えば、Void Fillingの計算量がVoid数Vの増加関数（計算量$O(V)$）で与えられること及び共有周回バッファ構成でのバッファ制御アルゴリズムがバッファ内の遅延線数Bの増加関数（計算量$O(B)$）で与えられること等を考えると本方式は有効である。一方で、本方式は、複数のプロセッサを用いるため、ハードウェア面、特に回路規模での検証が必須となる。本方式は、0.22 μm

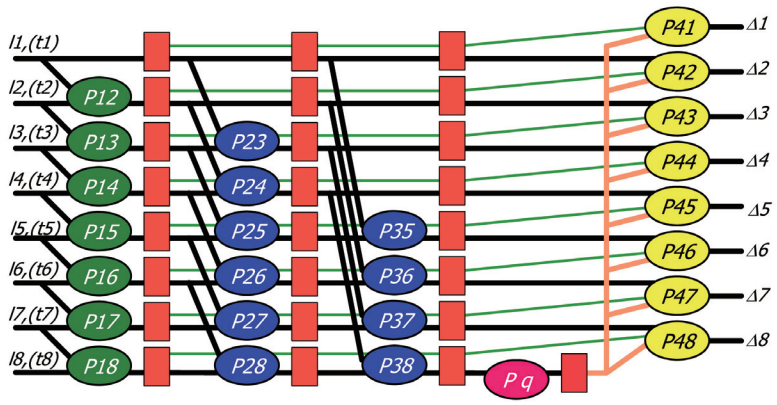

図5 高速バッファ管理実現のための並列パイプライン処理機構の概略

FPGA (Field Programmable Gate Array) のゲートレベルシミュレーションにより、最小 64 バイトの光パケットを処理する、回線速度 40 Gbps、8 ポートの光パケットスイッチのバッファ管理をサポートできること及び 0.13 μm FPGA の論理セルの見積りで、回線速度 40 Gbps、128 ポート (5.12 Tbps) のパケットスイッチのバッファ管理をサポートできることを確認している。また、パケット長が異なり、同期固定長を対象とするものの、回線速度 160 Gbps ベースの NICT の光パケットスイッチプロトタイプ[8] に導入できるハードウェアを開発している。本方式は、拡張により優先制御[18] や公平制御[12] も提供できる。

・優先制御

例えば光バッファの遅延線数に制約があって、十分なパケット棄却性能が提供できない時には、優先制御 (品質差別化) によって、あるクラスのパケットに対する性能を引き上げられる。[19] では、波長変換と使用できるファイバ遅延線数を閾値にし、TD (Threshold Dropping) 方式 (PBS；Partial Buffer Sharing は同意) を光バッファに適用している。NICT では、PBS を拡張し、バッファのみで優先制御を実現する PBSO (PBS with Overwriting) 方式により、PBS 方式よりも性能が向上することを確認した[20] [21]。上述の並列パイプライン方式を拡張した Parallel PBFS 方式は高速制御と優先制御が両立できる[18]。

4 まとめ

本稿では、インターネットへの適用、スループット 10 Tbps 超をターゲットにした、光パケットスイッチの概略とその構成技術について述べた。また、各構成に関して要求性能を述べ、制御技術の動向を記述した。

今後、光パケットスイッチをネットワークに展開するためには、各構成技術の規模の進展、微小化、低消費電力化が必要になってくる。さらに、ネットワークアーキテクチャに関する以下の検討も大切である。

(1) 光パケットと上位層パケット (IP パケットなど) とのインタフェースの検討。例えば、10 Gbps の IP パケットから先述の 160 Gbps の光パケットをいかに構成するかである。[22] では、DWDM 技術を用いて回線速度の和が 160 Gbps の光パケットを構成する方法を提案している。

(2) ルーティング。MPLS のように独自のネットワークのみで光パケット交換を使うのはさほど難しくないが展開が難しい。光パケット交換を広範に展開するには、光パケットスイッチが他種交換ノードと接続しても使えるような方式の開発が大切である。例えば、光パケットスイッチを交換ノードの代表である IP ルータと接続しても使えるルーティング[1] など制御技術の開発である。

参考文献

1 原井洋明，“フォトニックパケットスイッチ開発の最新動向とネットワークへの展開”，電子情報通信学会技術研究報告（PN2004-67），pp.35-40，Dec．2004．（招待講演）．

2 和田尚也，“160 Gbit/s/port 光パケットスイッチプロトタイプ及び関連技術の研究開発”，本特集.

3 S.J.B.Yoo, F.Xue, Y.Bansal, J.Taylor, Z.Pan, J.Cao, M.Jeon, T.Nady, G.Goncher, K.Boyer, K.Okamoto, S.Kamei, and V.Akella, "High-performance optical-label switching packet routers and smart edge routers for the next-generation Internet", IEEE Journal on Selected Areas in Communications, Vol.21, No.9, pp.1041-1051, Sep. 2003.

4 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 Transactions on Communications, Vol.E83-B, pp.2304-2311, Oct. 2000.

5 池澤克哉ほか，“光パケットネットワーク要素技術の開発”，電子情報通信学会技術研究報告（PN2005-5），pp.25-30，Apr．2005.

6 K.Kitayama and N.Wada, "Photonic IP routing", IEEE Photonic Technology Letters, Vol.11, No.12, pp.1689-1691, Dec. 1999.

7 N.Wada, H.Harai, and F.Kubota, "Optical Packet Switching Network Based on Ultra-Fast Optical Code Label Processing", IEICE Transactions on Electronics, Vol.E87-C, No.7, pp.1090-1096, Jul. 2004. (Invited)

8 H.Furukawa, N.Wada, and T.Miyazaki, "Demonstration of 160 Gbit/s Optical Packet Switching and Buffering Based on All-optical Code Label Processing", IEEE LEOS 18th Annual Meeting, No.MG5, pp.89-90, Oct. 2005.

9 "WAN packet size distribution", available from "http://www.nlanr.net/NA/Learn/packetsizes.html".

10 D.Wischik and N.McKeown, "Part I: Buffer Sizes for Core Routers", ACM/SIGCOMM Computer Communication Review, Vol.35, No.3, Jul. 2005.

11 原井洋明，“光パケットスイッチング技術”，平成 17 年度チュートリアル講演会（電子情報通信学会フォトニックネットワーク研究会／フォトニックネットワーク協議会），Feb．2006.

12 H.Harai and M.Murata, "High-Speed Buffer Management for 40Gb/s-Based Photonic Packet Switches", IEEE/ACM Transactions on Networking, Vol.14, No.1, pp.191-204, Feb. 2006.

13 L.Tancevski, S.Yegnanarayanan, G.Castanon, L.Tamil, F.Masetti, and T.McDermott, "Optical routing of asynchronous, variable length packets", IEEE Journal on Selected Areas in Communications, Vol.18, pp.2084-2093, Oct. 2000.

14 A.Ge,L.Tancevski, G.Castanon, and L.S.Tamil, "WDM fiber delay line buffer control for optical packet switching", in Proceedings of SPIE Vol.4233 (OptiComm 2000), pp.247-256, Oct. 2000.

15 T.Yamaguchi, K.Baba, M.Murata, and K.Kitayama, "Scheduling algorithm with consideration to void space reduction in photonic packet switch", IEICE Transactions on Communications, Vol.E86-B, pp.2350-2357, Aug. 2003.

16 F.Callegati, "Optical buffers for variable length packets", IEEE Communications Letters, Vol. 4, pp.292-294, Sep. 2000.

17 原井洋明，村田正幸，“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”，電子情報通信学会技術研究報告（PN2003-7），pp.31-36，Sep．2003.

18 H.Harai, "Parallel and Pipeline Processing for Prioritized Buffer Management in Photonic Packet Switches", in Proceedings of HPSR 2004 (The 2004 IEEE Workshop on High-Performance Switching and Routing), pp.156-161, Apr. 2004.

19 F.Callegati, G.Corazza, and C.Raffaelli, "Exploitation of DWDM for optical packet switching with quality of service guarantees", IEEE Journal on Selected Areas in Communications, Vol.20, pp.190-201, Jan. 2002.

20 H.Harai and M.Murata, "Photonic Buffer Architecture to Support Prioritized Buffer Management for Asynchronously Arriving Variable-Length packets", in Proceedings of ONDM 2003 (The 7th IFIP Working Conference on Optical Network Design and Modelling), pp.1103-1118, Feb. 2003.

21 原井洋明，村田正幸，“DiffServ を実現するための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法”，電子情報通信学会技術研究報告（PS2001-59），pp.139-144，Dec．2001．

22 H.Harai, "Skew Compensation for Multi-Wavelength Optical Packets in Photonic Packet-Switched Networks", in OECC/COIN 2004 Technical Digest (9th Optoelectronics and Communications Conference), pp.588-589, Jul. 2004.

原井洋明（はらい ひろあき）
新世代ネットワーク研究センターネットワークアーキテクチャグループ研究マネージャー（旧情報通信部門超高速フォトニックネットワークグループ主任研究員） 博士（工学）
ネットワークアーキテクチャ、光パスネットワーク、光パケットスイッチの研究開発