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

2006－HPC－107（42）
2006／8／2

光パケット多重ルータによるテラビット級広域分散計算

太田　昌孝†

†東京工業大学大学院情報理工学研究科　〒152-8552 東京都目黒区大岡山 2-12-1

E-mail:　†mohta@necom830.hpcl.titech.ac.jp

概要 10Gbps100 波長の WDM で利用可能な 1Tbps の全帯域を個々のパケットのエンコードに用いる（500B のパケットは 4ns になる）光パケット多重方式により、高速広帯域光スイッチの光の広帯域性を生かすことができる。ヘッダとペイロードを波長で分け、ヘッダ部分の制御を電気回路で行えば、実用的な数と長さの光ファイバ遅延線により実用的なパケット落ち率の光バッファを有する低消費電力テラビット級光ルータとテラビット級広域網が、既に商用化された技術で構成できる。テラビット級広域網による効率的な広域分散計算のためには、トラフィックの短期的変動を 1Gbps 程度に抑えなるべく長いパケットを使うほうがよい。

Tera-bit Class Wide Area Distributed Computing with Optical Packet Multiplexed Routers

Masataka OHTA†

†Graduate School of Information Science and Engineering, Tokyo Institute of Technology

2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8552 Japan.

E-mail:　†mohta@necom830.hpcl.titech.ac.jp

Abstract By optical packet multiplexing to encode packets at 1Tbps with WDM using 100 wavelengths each at 10Gbps (500B packet becomes 4ns), wide optical bandwidth of high speed wide band optical switches can be fully extracted. By separating a header and a payload by wavelength and processing the header by an electrical circuit, a low-power tera-bit class optical router having optical buffer with practical packet-loss possibility using optical fiber delay lines of practical number and lengths and a tera-bit class wide area network can be constructed with already commercialized technologies. For efficient wide area distributed computing over the tera-bit class wide area network, it is better to limit short-term variation of traffic less than 1Gbps or so and use longer packets.

1. はじめに

光パケット多重の考え方により実用的なテラビット級光ルータ（実用的なパケット落ち率の光バッファー付き）が現在の技術で作成可能であり[1, 2,]、長距離伝送にも対応できる[3]が、これについて紹介し、さらに、これを用いた広域分散計算のありかたについて論じる。

インターネットを用いた Grid などの広域分散計算では、大量のトラフィックが発生するため、高速の長距離通信路と高速のルータが必要である。当面のターゲットはテラビット級であろうか。

高速の長距離通信のためには WDM (Wave-length Division Multiplexing)技術によるテラビット級の信号を 1 本の光ファイバ上で伝送し、長距離伝送に伴う減衰は EDFA (Erbium Doped Fiber Amplifier)等により 1 台のアンプで全波長一括増幅するという、光の広帯域性を存分に生かした解が存在し、既に実用化している。長距離伝送では、分散の問題から、現状では波長あたりの速度は 10Gbps が実用上の限界であり、テラビット伝送のためには 100 波長利用する必要があるが、この程度の波長数なら EDFA で効率的に増幅できる波長域（C バンド）におさまる。

一方、高速ルータについては通常のルータを多数相互接続して並列化によりいくらでも高速化でき[4]、すでにそのようにしてテラビット級の交換能力を持つルータも製品化されている。しかしながらこのようなアプローチでは、並列度の増加とともにルータの大きさ、価格、消費電力[5]などが増加することになる。

そこで、光を利用することで高速ルータをより容易に実現することが考えられる。

これまでのアプローチは、いわゆる「波長ルータ」と言われるものが主体であった。波長ルータは、WDM 伝送されてきた入力の個々の波長をばらばらにし、場合によっては波長変換を施し、別々の出力に送り次段に WDM 伝送するものである。波長ルータの発想は、電気の介在なく光を光のまま交換すれば効率的であろうというものであった。しかしながら、波長ルータは 10Gbps 程度の各波長をばらばらに扱うため、光の広帯域性を全く生かしていないという根本的欠陥がある。

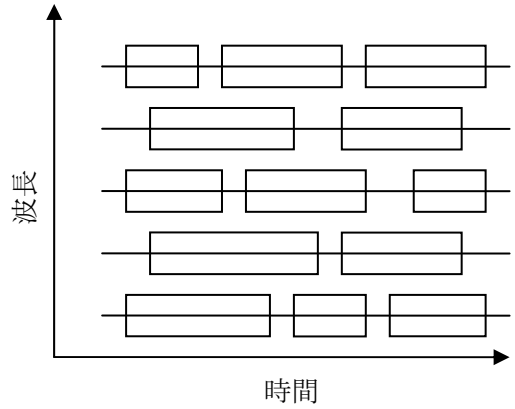


図１ 個別波長での IP over WDM

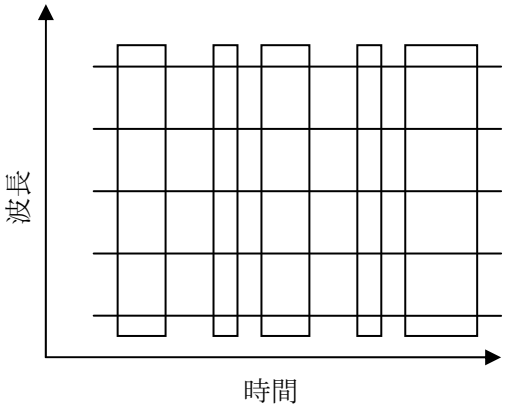


図２ 光パケット多重 IP over WDM

波長ルータの波長利用効率を上げるためにはある程度の波長変換が必要だが、これを電気回路で行ったのでは光のままの交換ではないし、光回路で行ったとしても個々の波長ごとに光回路が必要なので、光回路の広帯域性が生かせず、大きさ、価格、消費電力といった点で、電気より有利にはならない。

波長ルータの別の問題は、名前はルータであるが単なる回線交換装置であり、波長ルータで結合された 2 地点間の通信速度が 10Gbps 単位でしか変えられないことである。今日、あるいは将来の Grid の必要とする帯域に比べて 10Gbps は十分に遅いかもしれないが、民生用のアプリケーションでは 10Gbps どころか 1Gbps でも速過ぎる。また、波長あたりの速度を下げ波長数を増やすと、光の広帯域性がますます意味をなくし、より大規模で高精度な波長ルータが必要となる。かといって、従来からのルータを利用する民生用のインターネットと Grid 専用の波長ルータのネットワークを二重に持つのは、あまりに非効率的である。

問題の根本的な解決のために必要なのは、光の広帯域性を生かしたルータである。

その要素となるのが、近年利用可能になってきた高速広帯域光スイッチ[6, 7]である。[6]のスイッチはすでに製品化されており、例えば 1:8 の（デ）マルチプレクサは C バンド全体にわたる光を 6ns でスイッチできる[8]。スイッチ時間が 3ns のものもある。

このスイッチと光パケット多重の組み合わせにより、光の広帯域性を生かした高速ルータを構成することができるが、その概要を次節に述べる。WDM としては各波長 10Gbps で 100 波長で 1Tbps のものを想定する。

2. 光パケット多重ルータ

2.1 光パケット多重とは

インターネットにおいては、データはパケット単位に区別がつけられるため、パケット多重以外の多重化は不要であり、使われない。例えば SONET の OC-192 の伝送路があるときそれを 64 個の OC-3 の伝送路にわけたりはせず、9.6Gbps の伝送路として利用する。無線 LAN においても、OFDM の個々の周波数ビンに個別にパケットを流すわけではなく、利用可能な全帯域を利用して個々のパケットを伝送する。

この考え方をそのまま WDM にあてはめたのが、光パケット多重の考え方である。

図 1 は IP over WDM の方式としてよく見られるが個々のパケットはどれか一つの波長の中でだけエンコードされ、パケットはまず波長により分離され、波長内では時間軸により分離されている。

これに対して図 2 は光パケット多重による IP over WDM 方式であり、個々のパケットは WDM で利用可能な全体行きを利用して伝送され、時間軸方向にのみ分離される[1]。

このように光パケット多重されたパケット列を、前節で述べた高速広帯域光スイッチで切り替えれば、高速ルータが実現できる。

光パケット多重の考え方自体は古くは[9]にもみられ、また 4 波長を利用した 10G イーサネットでも使われているし、そもそもインターネットの他の媒体ではあたりまえのことである。しかし、テラビット級での例は[10]までなく、後述するようにそれが光バッファの実用化に本質的に重要である。

最近では光パケット多重の考え方をプロセッサ内で利用しようという提案もある[11]が、光バッファにより可変長の非同期パケットの衝突回避も容易に行えるため、光パケット多重をプロセッサ内だけに閉じ込める必要性はない。

現在のインターネット幹線を流れるパケットは平均 500B 程度で最大 1500B だが、これを 1Tbps でエンコ

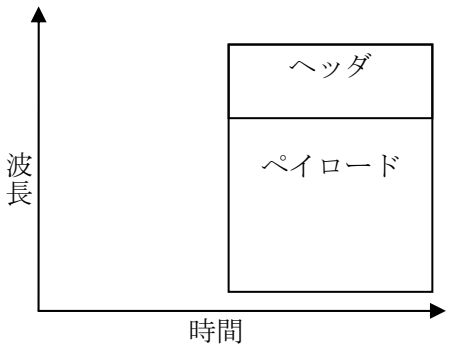


図３　ヘッダとペイロードを時間で分けないパケット形式

ードすると、それぞれ 4ns、12ns となり、高速広帯域光スイッチ等により 4ns の間隔でパケットを並べると、平均実効速度は 500Gbps ということになる。また、この程度の速度ならルータで必要なパケットヘッダの処理は電気回路で十分間に合う。電気回路での処理は一般に多段のパイプラインで行われるが、その間パケットは光ファイバ遅延線（数 m～数十 m）中を通して待たせておく。巨大な経路表の可変長マスクでの検索を一定時間で行うことは不可能ではない[12]しパイプライン化も可能だが、大量の高速メモリを必要とし、それでも速度が追いつかない場合はインターリーブでさらに大量のメモリが必要となる。そこで、IP の規格自体を見直し幹線での経路表を小さく制限しその他の点も考慮して光ルータに都合のいいようにすることも考えられる[13]。

[1]ではさらに、パケット形式を工夫し、ヘッダとペイロードを波長により分け（図３）、光ネットワークの中枢に位置し自ら光パケットを送受する必要のないコアルータではペイロード部分をエンコード・デコードする必要がないようにしている。一方、光パケット多重パケット網と従来の 10Gbps 程度の電気処理網との境界に位置するエッジルータでは、光パケットの送受のために多量の電気回路が必要となる。このような光ルータの構造を図 4 に示す。実際には経路制御などのパケットのやりとりのために少数の波長ではペイロードを入出力すべきである[1]。

2.2 光ファイバ遅延線による光バッファの構成

ルータでは、複数の入力から同時に一つの出力にパケットを送る際(や､光ルータでは考えなくて良いが、より速い入力からより遅い出力にパケットを送る際）には、出力では一時に一パケットしか出力できないため、残りのパケットを一時的にバッファしておく必要がある。バッファはあふれてもよいが、その確率は十分小さい必要がある。

これまで光パケット処理の弱点とされてきたのは、有効なメモリが存在せずパケットを光のままバッファする光バッファの構築ができなかったことである。

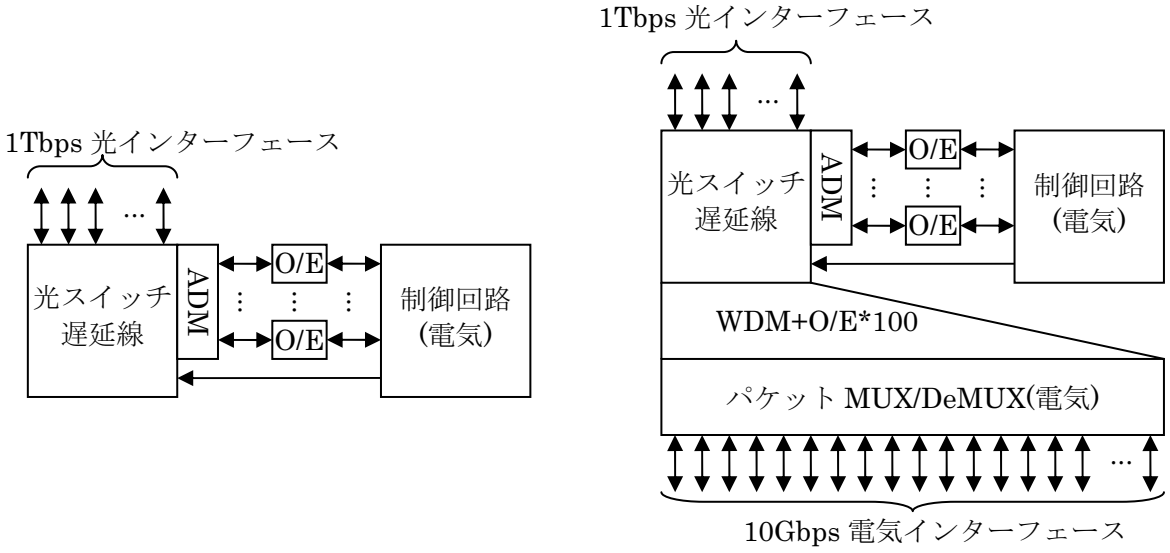
複数の光ファイバを遅延線として用いることで、一応の光バッファは構成できるが、あまり実用的とは考えられていなかった。

その理由は、ひとつにはインターネットのトラフィックには大量のバッファが必要と信じられていたことであり、別の理由は、光ファイバ遅延線の長さと数が非現実的になるというものであった。

しかし近年、バッファが大量に必要なのは回線に少数の TCP トラフィックしか流れていない場合で､極めて多数の TCP の変動が平均化されるインターネット幹線においてはトラフィックはポワソンとみなしてよいことが判明し[14]､帯域利用率が 70~80%でよければ十数～数十パケットのバッファで十分小さなパケット落ち確率とすることができることがわかった。

これにより光バッファへの要求は大幅に緩和されたが、さらに光パケット多重では 1500B のパケットが 12ns で伝送できるため､わずか 3.6m の光ファイバ（石英の屈折率を考慮すると 2.5m だが、以下では無視する）でバッファできることがわかる。

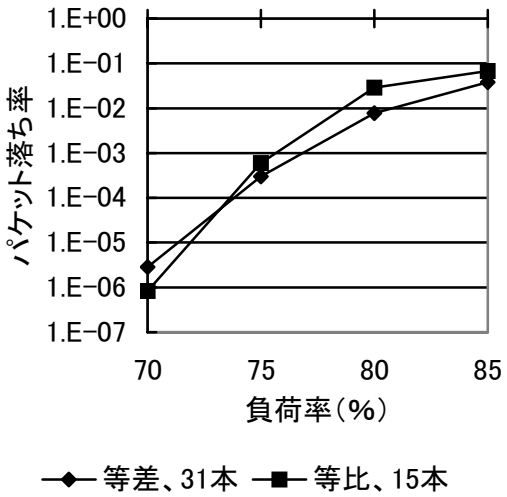
光ファイバ遅延線の構成方法については[9]で論じられており､[2]で光パケット多重の場合の解析を行った。[2]の数値はバッファ方式の比較のためのものなので、[2]と同じ条件でそれとは違う長さと数の光ファイバ遅延線で改めて計算を行った（表 1)。31 本の等差数列的長さ（最長 316m）の光ファイバ遅延線を利用すると、負荷率 75%、80%のとき、それぞれパケット落ち率は 0.03%、0.8%になる。[9]にあるように等比数列的考え方を導入すると光ファイバ遅延線の数を減らすことができ、15 本の等比数列的長さ（公比は 5 本目までは 2、以後 8 本目までは 1.5、以後は 1.333、最長 2063m）負荷率 75%、80%のとき、それぞれパケット落ち率は 0.06%、2.9%になる。



a） コアルータの構成　　b） エッジルータの構成

図４　光パケット多重ルータの構成

表1　光バッファのパケット落ち率



すなわち、1Tbps での光パケット多重により、インターネット幹線のパケットに対し、実用的な数と長さの光ファイバ遅延線により、実用的な負荷に対して実用的なパケット落ち率の光バッファが、[8]などの光スイッチにより構成できる（図５）。図でダミーとあるのは EDFA のサージ対策である[1]。光バッファ内の光スイッチの制御は、パケットヘッダの解析結果に基づき電気回路で行う。

これに対して、1500B のパケットを 100Gbps でエンコードした場合には同じパケット落ち率を達成するのに必要な光ファイバ長は 10 倍となる。これでは物量的にも問題があり、また等比数列的長さの場合は光ファイバ遅延線内での減衰も無視できなくなるため、実用的な光バッファとすることは困難である。わずか 10Gbps でしかエンコードしない場合には必要な光ファイバ長はさらに 10 倍となり、全く実用性はない。

2.3 光パケット多重ルータの消費電力

以上のような考え方で、3 ポートのルータで各出力に 15 本の光ファイバ遅延線を持つ光ルータは、1:8 の光スイッチ 6 個、1:4 の光スイッチ 45 個、1:2 の光スイッチ 3 個で構成できる。光スイッチの消費電力は動作周波数によって変わるが、平均 500B のパケットを 4ns 間隔で伝送する場合、光スイッチは平均 8ns ごとに駆動される。その際全スイッチを切り替える必要はないが、全スイッチを 1/2 の確率で切り替えるとしても動作周波数は 31.25MHz となり、[8]などのデータシートから試算すると、その消費電力は光スイッチだけでは 2.9W、ドライバを含めても 302W でしかない。また、ドライバの消費電力については、まだまだ改善の余地が大きいと思われる。

実際にはこれにヘッダ処理のための光電変換や制御回路の消費電力が加わるが、経路表部分を除けば全体としても光パケット多重コア光ルータの消費電力は小さく抑えられる。これは 1Tbps での光パケット多重により高速広帯域光スイッチの広帯域性を生かした結果

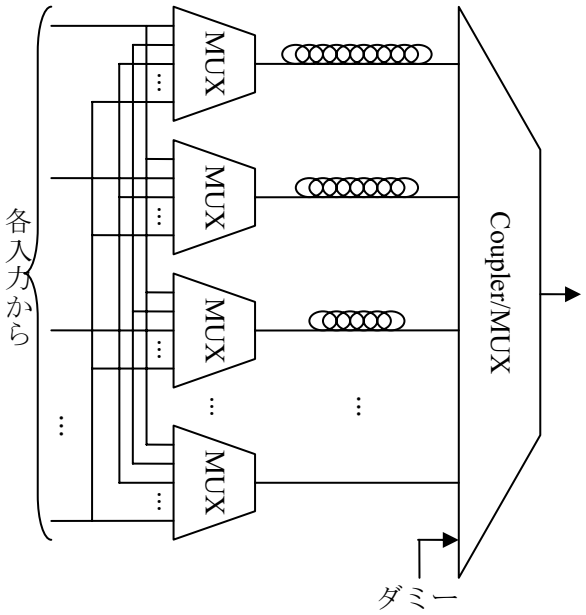


図５　光スイッチと光ファイバ遅延線による光バッファ

である。

一方エッジ光ルータの消費電力はそれほど小さくはならないであろうが、これは仕方がない。なお、光幹線網で 20 ホップ伝送されるとすると両端のエッジルータの消費電力の影響はおよそ 1/10 になる。

3 光パケット多重ルータによるテラビット級広域分散計算

前節で説明した光パケット多重ルータを広域に配置し、WDM 伝送路により接続すれば、現在利用可能な技術を組み合わせるだけでテラビット級光ネットワークが構成できる。長距離多段伝送に伴う各種歪みのパケット単位での補正については、[1, 3]で解決策を提案している。

そこで、テラビット級広域分散計算が可能となる。平均 500B のパケットを 1Tbps で伝送しパケット間を 4ns 空け負荷率を 75%とすると、実効速度は 375Gbps となる。しかし、光パケット多重ルータの特徴から、考慮すべきことがいくつかある。

まず一つは、当面は通信量をあまり気にしなくてもよいということだ。数 Gbps 程度までのトラフィックなら、瞬間的な速度の変動は他のトラフィックと混ざってしまうので、何も気にせず民生用の安価な網を通じて普通にデータをやりとりすることができる。

しかし、数 Gbps 程度を超える通信速度で、通常の TCP のように瞬間的な速度が大きく変動すると、負荷率が%のオーダーで変動するため光バッファでのパケット落ち率も大きく変動し、網全体へ悪影響を与えかねない。そこで、送出速度をほぼ一定にするための何らかの工夫が必要である。TCP の場合、[14]でもある程度の議論はあるが、slow start は 1Gbps 程度までに留めるとか、パケット落ちに対しては送出速度をいき

なり半分にせずに 1Gbps 程度しか低下させない、などの方策が考えられる。後者のような変更を行った TCP は他の TCP より有利であるが、ありもしない fairness にこだわって網全体を不安定にするよりは、ましであろう。

なお、多数の要素 CPU からなる並列プロセッサが個別に比較的低速の TCP を張る場合は、定常状態では全体が平均化されるので問題はあまり起きないが、多数の slow start が同期して発生しないようにはしなければいけない。

光バッファの容量の小さいことには良い面もある。即ち、バッファによる遅れがほとんど発生しないことであり、インターネットの遅延はほとんど伝送遅延だけになることになる。

別の問題として、光パケット多重の効率がある。光パケット多重ではパケット間隔の部分の帯域が無駄になるため、なるべく大きなパケットを送るほうがよい。現在のインターネット幹線ではイーサネットの影響で最大パケット長は 1500B になっているが、全てのパケットが 1500B になれば、それを 1Tbps で伝送しパケット間を 4ns 空け負荷率を 75%とすると、実効速度は 562.5Gbps になる。最大パケット長を 9000B にすれば、より長い光ファイバ遅延線は必要だが、9000B のパケットだけを 1Tbps で伝送しパケット間を 4ns 空け負荷率を 75%とすると、実効速度は 710.5Gbps になる。多数の要素 CPU からなる並列プロセッサが個別に比較的低速の TCP を張る場合はパケット長があまり大きくならない可能性があるので、ある程度の数の CPU をひとまとめにして TCP を張るなどして対処の必要があるであろう。

4 おわりに

1Tbps の WDM での光パケット多重により現在実用化されている要素技術の組み合わせでポートあたりテラビット級の速度の光ルータが作成できることを紹介した。

テラビット級光ルータによる網は広域分散計算に利用できるが、その際、各ルータのバッファが小さいことに配慮しトラフィックの激しい変動を起こさないようにし、また、光パケット多重の効率向上のためなるべく長いパケットを使うべきである。

以上のような点を考慮の上、[11]のようなコンピュータをテラビット級広域（公衆）網に直結し TCP 速度の世界記録を狙うのも一興であろう。

なお、今後も計算機の速度はますます向上し、広域分散計算に必要な通信速度も増大の一方であると予想されるが、光パケット多重ルータの速度もそれに劣らず増大することが期待できる。

高速広帯域光スイッチはサンプル出荷が始まったばかりであり、今後のさらなる高速化と広帯域化によりポートあたり 10Tbps 級のルータも夢ではない。その場合長距離伝送路の容量を超えてしまうことへの広帯域一括波長変換による対処法は[15]に示した。もちろん、並列化による速度向上はいつでも可能である。

文　献

[1] 太田 昌孝、「全光データパスルータの構成要素」、信学技報　PN、2005 年 8 月.

[2] 太田 昌孝、「光パケット多重ルータのファイバー遅延線による光バッファ構成 」、信学技報　PN、2006 年 3 月.

[3] 太田 昌孝、「光パケット多重網間接続と光パケット多重パケットの既存長距離線路上の伝送 」、信学技報　PN、2005 年 12 月.

[4] Masataka Ohta, M. Sola, K. Fujikawa, A. Kojima, H. Fukumori, Y. Muraoka, "Hash Parallel and Label Parallel Routing for High Performance Multicast Router with Fine Grain Qos Control", Proceedings of Internet Workshop '99, pp. 13-16, February 1999.

[5] 小笠原 敦、「情報通信のエネルギー問題　―求められる通信インフラの省電力化―」、http://www.nistep.go.jp/achiev/ftx/jpn/stfc/stt063j/index.html、科学技術動向、2006 年 6 月.

[6] 梨本 恵一他、「PLZT 薄膜光導波路による高速スイッチの開発」、信学技報　PN, 2004 年 10 月.

[7] 池澤 克哉他、「光パケットネットワーク要素技術の開発」、信学技報　PN, 2005 年 4 月.

[8] のぞみフォトニクス 、「AlacerSwitch 0108S: 1x8 high-speed optical switch 」、http://www.nozomiphotonics.com/pdfs/jp/datasheet_alacer_switch_0108S.pdf、2006 年 6 月.

[9] Z. Haas, “The “Staggering Switch”: An Electronically Controlled Optical Packet Switch”, Journal of Lightwave Technology,, Vol. 11, No. 5/6, May/June 1993.

[10]太田 昌孝、「全光ルーティングのインターネットでの利用」、信学 PN 研究会夏合宿、2003 年 9 月.

[11]木村 康則、「次世代スーパコンピュータを実現するインターコネクト技術」、http://www.psi-project.jp/images/dis/yasunori_kimura_20050928.pdf、計算科学技術シンポジウム、2005 年 9 月.

[12]P. Gupta, S. Lin, and N. McKeown, “Routing Lookups in Hardware at Memory Access Speeds”, http://klamath.stanford.edu/~nickm/papers/Infocom98_lookup.pdf, IEEE Infocom’98, March/April 1998.

[13]太田 昌孝、「オプティカル IPv6 」、信学技報　PN、2006 年 5 月.

[14]M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, T. Roughgarden, “Routers with Very Small Buffers”, ACM SIGCOMM Computer Communication Review, Vol 35, No. 2, July 2005.

[15] Masataka Ohta, “IP over Baseband DWDM and SDM”, OECC2006, July 2006.