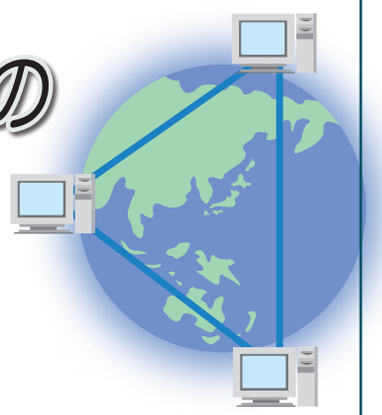
5

Lambda Networking：広帯域ネットワーク利用の一形態



加藤　朗
東京大学情報基盤センター
kato@wide.ad.jp

山本 成一
東京大学大学院情報理工学系研究科
yama@wide.ad.jp

関谷 勇司
東京大学情報基盤センター
sekiya@wide.ad.jp

Lambda Networking

インターネットの普及およびアプリケーションの高度化に伴い，インターネットで交換されている情報量は飛躍的に増大してきた．各々のサーバやホスト，アクセス回線の帯域の改善やその低廉化がこれに大きく貢献しているが，一方これらのトラフィックをさばくための，ネットワークリンクの広帯域化やルータの高性能化によるところも大きい．媒体が無線や銅線から光ファイバーになったことによって，ネットワークの帯域は大きく改善された．単一の光信号で伝達できる帯域も飛躍的に向上し，また複数の光信号を同一光ファイバーを波長多重する WDM によって，1 芯あたりの伝達容量は 1Tbps に達しようとしている．

通信回線の帯域が飛躍的に増加したとはいっても，単一リンクでも帯域は現在は約 10Gbps が上限であり，一部に OC-768c（約 38Gbps）をサポートしたものがあるにすぎない．また回線容量が増加しても，遅延時間は変わらないため，特に長距離大容量のデータ転送では，帯域遅延積は非常に大きくなってしまう．たとえば，RTT が約 250ms の日本とヨーロッパの間で 7Gbps の通信を想定すると帯域遅延積は 220MB にも達するが，関連する各ルータやスイッチにこれに匹敵するバッファを確保するのは容易ではない．また，遅延が非常に大きいため，何らかの原因で受信側でのフロー制御が必要になったとしても，それが有効になるには RTT 分の時間が必要であり，それまで 220MB のデータが到達することになる．

TCP を改良し，このような帯域遅延積の大きい環境にも適用できるようにする研究も盛んに行われている．通常の TCP は slow start のため，ウィンドウが帯域遅延積に匹敵するまで成長するのに非常に時間がかかるという問題のほか，パケット損失が発生すると輻輳ウィンドウを半分にするため，転送性能が大きく低下してしまい，帯域遅延積の大きな通信で性能を確保するのは容易ではない．これに対して，輻輳ウィンドウが一定以上の大きさがあるとき，パケット損失発生時の輻輳ウィンドウの増減について工夫をすることでパフォーマンスの劣化を抑える HighSpeed TCP[1] や Scalable TCP[2]，また RTT の計測値に基づいて輻輳ウィンドウを制御する FAST TCP[3] などが提案されている．また TCP ではなく UDP に独自の制御プロトコルを搭載して高速伝送を実現しようとしている提案もある[4), 5)]．いずれもインターネット上で実行することを前提にしているため，公平性を考慮する必要がある点が 1 つのネックになっている．

帯域遅延積の大きな通信に対する別の解決法は，アプリケーションに対して直接 Layer-1 の通信路を，あるいは途中に Layer-2/Layer-3 デバイスが介在するとしても回

線の全帯域を専用に割り当てることである．数 Gbps という帯域要求が現実的になってきているため，光通信技術を用いて単一の波長を用いた通信路をそのままアプリケーションに割り当てることも可能になってきている．専用の通信路が得られれば，スイッチ等の中間に介在する機材のバッファも専用に使用することができ，低い割合で発生するビットエラーまで妨げることはできないが，中間でのパケット損失は発生せず，また公平性の枠に縛られることなく，任意のプロトコルを実行することができる．このような通信形態は Lambda Networking，得られた通信路は光パスと呼ばれている．

このような Lambda Networking は，一般的なインターネットトラフィックを搬送する回線として利用することも可能であるが，非常に大きな帯域遅延積を持つ通信では Lambda Networking が期待されている．特に最近の巨大科学では，観測データの発生源である望遠鏡や加速器等を任意の場所に設置することは不可能であるため，巨大な帯域遅延積を持つ通信は避けられない傾向にある．

GLIF

各国の学術ネットワークのほとんどは，学術情報の交換を目的として Layer-3 のサービスを提供しているが，その中には，大容量のデータ転送や各種デモンストレーション，あるいは将来のネットワークを研究する目的で，Layer-2 あるいは Layer-1 のサービスを提供しているものもある．もし国際的に広帯域で Layer-2 あるいは Layer-1 のサービスを提供しているネットワークを相互に接続することができれば，単一のネットワークでは実現が難しかった長距離大容量のデータ転送を伴う各種実験が可能になる．

2001 年よりオランダの学術ネットワークである SURFnet の Kees Neggers 氏により，毎年小規模な Lambda Workshop という会合が開催され，光技術を用いた学術ネットワークの国際的な接続に関して議論されてきた．2003 年 8 月にアイスランドの Rejkyavik で開催された Lambda Workshop では，その会合の名称が GLIF[☆1] に変更された．

多くの学術ネットワークでは Layer-3 のサービスも必須であるため，実際に国際的な接続を Layer-2 あるいは Layer-1 で実現するためには，複数の広帯域国際回線を運用していることが必須になる．国際回線は非常に高価であるため，この実現は容易ではなかったが，近年の非常に広帯域な海底ケーブルの敷設による回線価格の低下や，回線事業者からの回線の寄贈などによってこれが可能になってきた．2002 年 9 月から運用が始まった IEEAF[☆2] の大西洋回線は，冗長性はないものの 9.6Gbps の回線が Tyco Telecom によって寄贈されたものである．

ネットワーク	運用開始	区間	運用
IEEAF	2003年1月	東京ーSeattle	Layer-2
JGN2	2004年8月	東京ーChicago	Layer-2 化予定（現在はLayer-3）
SINET	2005年4月	東京ーNew York	Layer-3
TransPAC	2005年4月	東京ーLos Angeles	Layer-3

表-1　我が国の9.6Gbps国際リンク

我が国では，従来は商用インターネットや学術インターネットでは，海外，特にアメリカ合衆国に対して広帯域な接続性を獲得することに主眼が置かれてきたが，最近では帯域のみならず Layer-1/Layer-2 の接続性の提供も考慮されているものが少なくない．**表 -1** に我が国に関連する学術系ネットワークで使用されている国際リンクのうち，現在 9.6Gbps のものを示す．ここで，Lambda Networking に対する期待が高まっているとはいえ，通常の接続性を安定に確保することが目的である SINET のような安定運用を指向する Layer-3 サービスを提供するネットワークの重要性が失われたわけではないことを明記しておきたい．

Data Reservoir

「Lambda Networking」に述べたように，最近の科学プロジェクトでは，帯域遅延積の非常に大きなデータ転送を要求するものが少なくない．しかし，これらのプロジェクトに従事する研究者は計算機ネットワークの専門家ではなく，それぞれの研究分野での活動に専念したい．この問題を解決するため，遠隔地で発生したデータを研究者の近傍に効率的に転送することによって，研究者に対して大量のデータに対するより容易なアクセスを提供することを目的とした研究プロジェクトが Data Reservoir[6)～8)]（以下 DR と記す）である．

まず観測データを，近傍に設置した DR 装置に送る．DR 装置はいったんデータをディスクに格納した後，データの利用者の近傍に設置した DR 装置に対して，広帯

☆1　GLIF ― Global Lambda Integrated Facility: http://www.glif.is/

☆2　Internet Educational Equal Access Foundation: http://www.ieeaf.org/

域ネットワーク上を iSCSI/TCP を用いてデータ転送を行う．この転送の帯域遅延積は非常に大きくなるため，転送効率が十分に得られるように TCP のアルゴリズムやパラメータなどの調整を行っておく．その結果，観測データの利用者は近傍の DR からデータを取り出し，高性能計算機で解析を行うことができる．

DR プロジェクトは 2002 年に Baltimore で開催された SC2002 に出展し，「Most Efficient Use of Available Bandwidth Award」を受賞した．翌 2003 年に Phoenix で開催された SC2003 では，現地に 30 台の Xeon サーバを持ち込み，東京に設置されていた対抗側のシステムとのデータ転送を行った．

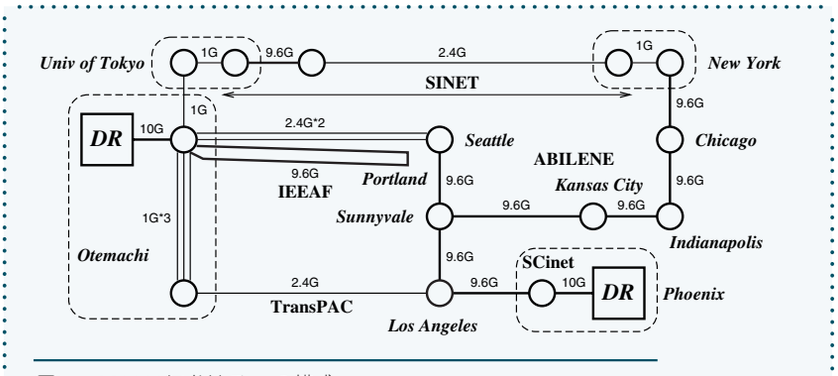


図-1　SC2003におけるDRの構成

図 -1 に示すように，東京まで（1）会場から Seattle までは主催者が準備した SCinet および UCAID が運用するネットワークである Abilene を経由し，Seattle から東京までは 2.4Gbps の回線を 2 回線を用いた経路，（2）Abilene によって Los Angeles を経由して，TransPAC によって東京に至る別の 2.4Gbps による経路，（3）Abilene によって New York を経由し，SINET により東京に至る 1Gbps の経路を確保し，合計 8.2Gbps の帯域を得た．さらに東京でポリシルーティングにより，Portland までの 9.6Gbps の折り返し回線に送出することによって，実質的に 24,000km を越える距離の通信路を確保し，7.56Gbps のディスク間転送速度を達成した．

SC2003 に出展したシステムでは，それぞれ 16 台のサーバによるシステムによって上記の性能を達成し，「Distance × Bandwidth Product & Network Technology Award」を受賞したが，よりコンパクトな実装が好ましいことは言うまでもない．そのため，一般的になりつつあった 10Gbit Ethernet を用いて，単一のサーバ間での伝送性能の向上を目指すことにし，2004 年当初から開発が進められてきた．10Gbps を目標とした場合，PCI-X バスの容量が 8.5Gbps 程度しかないため，TCP/IP の処理は NIC で行い，PCI-X バスの利用を最小限にするとともに，メモリへの帯域やアクセス遅延で有利な AMD 社の MPU Opteron を用いたシステムが開発されてきた．

日本〜スイス間の光パス

2004 年 7 月に Cairns で開催された APAN Meeting において，CANARIE の René Hatem 氏によって，IEEAF Pacific，CA*NET4，IEEAF Atlantic，SURFnet が運用する回線を相互に接続し，東京と Geneva の間を 9.6Gbps で接続する実験が提案された．その後，同 9 月に Nottingham で開催された GLIF Meeting で関係者の打ち合わせが行われた．この実験の主な目的は，日本とヨーロッパを結んだ 9.6Gbps の回線を正しく設定し，それを確認することにあったが，この回線を利用した実験も併せて実施することが確認された．そして前章で述べた DR 装置を Geneva の CERN と東京大学にそれぞれ設置し，伝送実験を行うことになった．

本実験で得られる光パスは 9.6Gbps の SONET OC-192 であり，Packet over SONET（POS）に対応したハードウェアで終端するのが一般的であったが，非常に高価であるという欠点があった．そのため，製品の出荷が始まったばかりであるものの，10Gigabit Ethernet を SONET 上で稼働させる，いわゆる WANPHY を用いて接続を行うことにした．

当初の予定からは若干の変更はあったものの，図 -2 に示すような回線が構成された．光パスの設定はほぼ 1 日で完了したが，東京側での 10Gigabit Ethernet の WANPHY が当初予定していた機材では動作せず，代替機が必要になったこと，また中間の多重化装置においてハードウェア障害が発生しており，保守部材の輸送を含めて時間がかかったが，それでも約 4 日間で全体の設定が完了した．得られた光パスは少なくとも☆3 18,600km にわたり，RTT は 263ms であった．

この光パスを用いて，DR の転送実験を行ったところ，Opteron PC 間の単一 TCP によるメモリ間転送で 7.57Gbps の転送性能が得られた．また，9 台の Xeon PC からなるシステムを用いたディスク間転送で 9Gbps の転送性能が得られた．特にメモリ間転送の性能は，Chelsio T110 という 10Gigabit Ethernet NIC に TCP/IP の

☆3　上記ネットワーク機材の設置場所間の距離の総和．実際の光ファイバーは最短距離で敷設されているとは限らない．

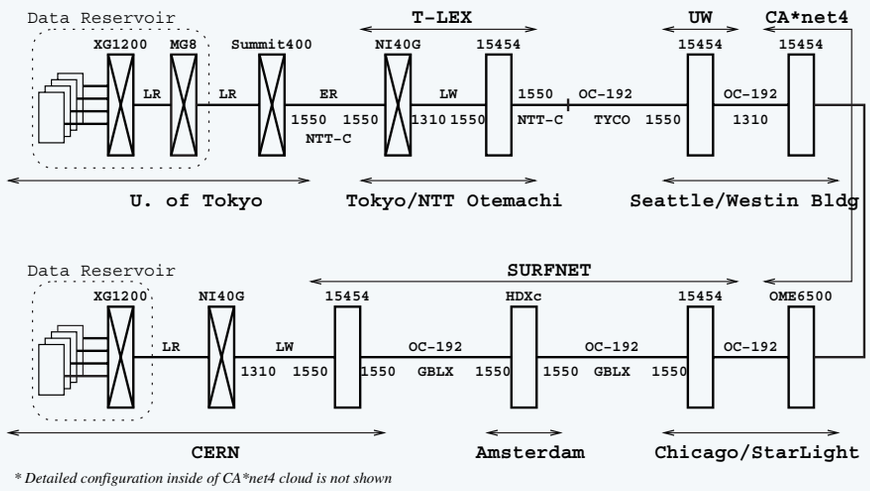


図-2　日本～スイス間の光パスの構成

機能を担当させ，PCI-X バスを介する転送を最小にとどめたこと，および送信側で送信レートを調整し，受信側での取りこぼしが発生しないようにしたことによるところが大である．

この実験の目的は，短時間で光パスを生成することではなく，光パスが正しく設定できること，およびそれを有効に利用できることを示すことである．これに関しては，UCLP[☆4] や GMPLS[9] などの制御プロトコルを用いることにより，光パスの設定時間を非常に小さくすることが期待されている．

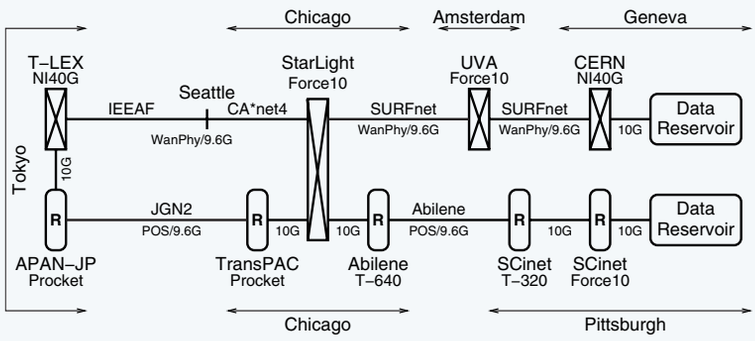


図-3　SC2004の際の光パス構成

この実験の結果，得られた光パス上にエラーが発生している場合，その発生場所を特定するのはそれほど容易ではないことが分かった．上記ハードウェア障害に対しては，中間地点で回線折り返しを依頼し，エラー発生区間の切り分けをしながら，対象区間を絞っていくという原始的な方法に頼らざるを得なかった．したがって本格的な光パスの利用を考える場合，パス制御プロトコルだけではなく，ユーザに回線折り返しテストを提供する枠組みや，途中で光分岐器を用いた回線モニタの設置などの対応が必要であることが示唆される．

LSR受賞まで

DR の実験グループは 2004 年 11 月に Pittsburgh で開催された SC2004 に参加し，前章の転送実験の経験を元に，より長い距離での転送実験を行った．図 -3 に示すように，Amsterdam に設置した DR と会場に設置した DR の間の転送を，いったん Chicago を経て東京を経由し，JGN2 国際回線によって Chicago を経由する経路上で行った．経路長は少なくとも 31,200km あり，1500byte の通常の Ethernet フレームを用いた単一 TCP による転送では 7.21Gbps を記録した．また，単一サーバによるディスク間転送では，1.6Gbps を記録している．この SC2004 における転送実験では，DR は「Single Stream, Longest Path, Standard MTU TCP Throughput Award」を受賞した．

Internet2 における Land Speed Record（LSR）[☆5] は，

- 少なくとも一区間は Abilene などの運用ネットワークを経由すること
- 経路長は，ルータ間の距離の合計で表現し，最大長 30,000km とする

[☆4] User Controlled LightPath Provisioning: http://phi.badlab.crc.ca/uclp/

[☆5] Internet2 Land Speed Record: http://lsr.internet2.edu/

などの規定が設けられている．このうち，上の規定は，Chicago—Pittsburgh 間は Abilene によっているため問題ないが，下の規定によると，Geneva—東京間は，途中にルータが存在していないため，実際にはアメリカ大陸経由のところを，直線距離で評価されてしまい，認定された経路長は 20,645km となった．そのため，IPv4 Single Stream では 148.85 Pbm/s の記録が認定されたが，IPv4 Multiple Stream での記録の更新には至らなかった．

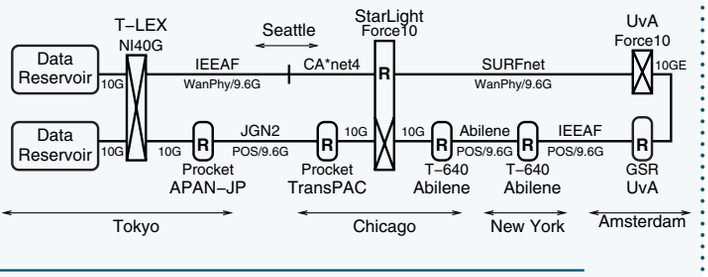


図-4　LSR受賞の際の光パス構成

このため DR グループでは，東京から Amsterdam を経て東京に戻る経路を想定し，

- New York － Chicago 間は Abilene 経由とするが，これがボトルネックになる可能性があるため，最もトラフィックが少ない時間帯を選ぶ
- 東京－ Amsterdam 間は，途中の Chicago の Starlight Switch をルータとして動作させ，距離計算上東京－ Chicago，Chicago － Amsterdam の距離の和で評価されるようにする

という工夫をすることによって，Internet2 の LSR の規則を満足した上で再度記録に挑むことにした．

実際の構成は図 -4 に示す通りであり，2004 年 12 月 24 日に実施された．なおこの際，Chicago － Seattle 間の光パスは CA*net-4 の 9.6Gbps 回線を用いたが，同時期に実施されている他の実験との時分割共有のため，UCLP を用いたユーザによる光パスの設定が行われた．総経路長は約 34,000km であるが，LSR の規則によって最大値の 30,000km と評価された．この環境でも DR は単一 TCP ストリームによって 7.21Gbps を記録し，帯域距離積では 216.3Pbm/s が LSR の「Single and Multiple Stream, Longest Path, Standard MTU TCP Throughput Internet2 Land Speed Record」として認定された．

まとめ

帯域遅延積の大きな通信に対して，光パスを設定することによって，帯域を確保する方法が有効であることが DR の実験によって示された．しかし，光パスは回線交換であるため，設定されている間は他の通信と帯域を共有することはできない欠点もあり，UCLP や GMPLS などの制御システムの導入によって，必要な時間だけ光パスを設定することも必要になる．

一方，これらの制御プロトコルの導入によって短時間で光パスの設定は可能になるが，得られた光パスの品質の保証や，問題の検出や測定，切り分けなど解決するべき問題もあり，これらについて今後検討していきたい．

謝辞　光パス上でのデータ転送実験を担当していただいた DR プロジェクトの東京大学平木教授，稲葉助教授，玉造講師，中村助手，富士通コンピュータテクノロジーズの来栖氏，坂元氏，生田氏に深く感謝します．本実験に関して回線をお貸しいただいた WIDE Project，IEEAF，Pacific Northwest Gigapop/University of Washington，CANARIE，StarLight，SURFnet，CERN の関係各位，T-LEX の運用にご協力いただいている NTT Communications の長谷部氏，また機材等に関してご協力いただいた Cisco Systems，Net One Systems，物産ネットワークス，NTT Communications の関係各位に感謝します．

参考文献

1) Floyd, S.: HighSpeed TCP for Large Congestion Windows, RFC3649 (Dec. 2003).
2) Kelly, T.: Scalable TCP: Improving Performance in HighSpeed Wide Area Networks, In Proceedings of the First International Workshop on Protocols for Fast Long-Distance Networks (Feb. 2003).
3) Jin, C., Wei, D. X. and Low, S. H.: FAST TCP: Motivation, Architecture, Algorithm, Performance, In Proceedings of IEEE INFOCOM (Mar. 2004).
4) He, E., Leigh, J., Yu, O. and DeFanti, T.: Reliable Blast UDP: Predictable High Performance Bulk Data Transfer, In IEEE Cluster Computing (2002).
5) Sivakumar, H., Grossman, R. L., Mazzucco, M., Pan, Y. and Zhang, Q.: Simple Available Bandwidth Unitization Library for High-Speed Wide Area Networks, In Journal of Supercomputing (2003).
6) Hiraki, K., Inaba, M., Tamatsukuri, J., Ikuta, Y., Hisashi, K. and Jinzaki, A.: Data Reservoir: Utilization of Multi-Gigabit Backbone Network for Data-Intensive Research, In Proceedings of Supercomputing (2002).
7) Nakamura, M., Inaba, M. and Hiraki, K.: Fast Ethenret is Sometimes Faster than Gigabit Ethernet on LFN － Observation of Congestion Control of TCP Streams, In Proceedings of International Conference on Parallel and Distributed Computing and Systems (PDCS2003), pp.854-859 (2003).
8) Nakamura, M., Senbon, J., Sugawara, Y., Itoh, T., Inaba, M. and Hiraki, K.: End-node Transmission Rate Control Kind to Intermediate Routers － Towards 10Gbps Era, In Proceedings of the Second International Workshop on Protocols for Fast Long-Distance Networks (2004).
9) Berger, L.: Generalized Multi-Protocol Label Switching (GMPLS) Singaling Functional Description, RFC3471 (Jan. 2003).

（平成17年7月5日受付）