第 XXXI 部

超広帯域オプティカルネットワークの設計と運用

第 31 部
超広帯域オプティカルネットワークの設計と運用

[1]

第 1 章　Lambda Networking

近年、9.6 Gbps などの広帯域ネットワーク回線を 1 Gbps や 2.4 Gbps に分割し、あるいは分割せずにそのまま 9.6 Gbps の回線として用い、それを単一のアプリケーションである時間専有するいわゆる Lambda Networking が注目されている。一般的なインターネットは、非常に多数のユーザが少しずつ帯域を分け合っており、多くの場合、あるユーザは特定の宛先との通信を行うのではなく、しばしばそのユーザには気が付かないうちに、非常に多くの通信相手とデータのやり取りをしている。しかし、特に大規模な科学技術研究では、非常に大量のデータを発生する研究施設が国際的に共有されていることが珍しくなく、これらのデータ転送は必然的に大容量長距離に渡って行う必要が発生する。

これらの長距離の広帯域伝送に関して、eVLBI のように天体からの雑音情報を離れた地点で観測し、それらのデータの相関を取るため、多少のパケット損失は許されるようなアプリケーションも存在するが、多くの場合には信頼性のあるデータ転送が必要になる。特に遅延が 100〜300 ms に及ぶような距離を 1 Gbps を越える転送性能のもと TCP でデータを送信するには、多くの問題がある。特にパケット損失は転送性能を確保する上で致命的なものであるため、QoS よりもより固い帯域確保が必要になる。このような通信は一般には限られた地点間に発生するため、性能劣化問題やセキュリティ問題を回避するため、しばしば Layer-3 より低い通信路が用いられる。

その極端な例では end to end に Layer-1 の通信路を確保することになるが、この場合には、途中にはルータやブリッジなどのパケット単位でのデータ転送をおこなう装置は介在せず、パケット損失は残存 bit エラーのみになり、パケット損失の発生を最小限にすることができる。そこまで行かなくても帯域が確保されているため、インターネット上のトランスポート層プロトコルに要求される公平性は全く考える必要がなくなり、より積極的に再送を行うプロトコルを用いることも可能である。そのため、このような通信形態である Lambda Networking が大規模科学技術をサポートするためには有望視されており、設定される通信路はしばしば Light Path とも呼ばれている。

第 2 章　GLIF

定常的に広帯域回線を使用するためには、もちろん二地点間に専用回線を調達することが必要になるが、それを準備するためには多額の予算が必要になるほか、長期契約を求められたり、発注してから利用できるまでに長時間掛かるなど気軽に実験することはできない。しかし、多くの研究・教育ネットワークは利益を目的にしていない。教育・研究ネットワークの多くは国の予算によって運営されているため、当然その国の研究者に優先権があるが、空いている帯域を融通することによって複数の国を跨ぐ Light Path を設定することが可能になった。

このような Light Path を提供しうる研究・教育ネットワークは GLIF — Global Lambda Integrated Facility —[2] という集団を形成しており、WIDE プロジェクトが運営に協力している IEEAF の太平洋リン

1　本ロゴは奈良先端科学技術大学院大学の寺田直美博士による。
2　http://www.glif.is/

表 2.1. GOLE 一覧表

GOLE	Location	Operator
AMPATH	Miami (FI)	Florida International University
CANARIE-StarLight	Chicago (IL)	CANARIE
CANARIE-PNWGP	Seattle (WA)	CANARIE
CENIC	Los Angeles, CA	CENIC
CERN	Geneva, CH	CERN
CzechLight	Prague, CZ	CESNET
HKOEP	Hong Kong, HK	CSTNET
KRLight	Daejeon, KR	KISTI
MAN LAN	New York (NY)	Internet2 and NYSERNET
NetherLight	Amsterdam, NL	SurfNET and SARA
NorthernLight	Stockholm, SE	NORDUNET
Pacific Northwest GigaPoP	Seattle (WA)	PNWGigapop and Univ. of Washington
StarLight	Chicago (IL)	Northwestern University and UIC
T-LEX	Tokyo	WIDE Project
UKLight	London, UK	UKERNA

クの存在から、WIDE プロジェクトも GLIF に参加している。GLIF は 2003 年 8 月に、それまで SurfNET の Kees Neggers 氏が開催していた invitation only な小さな会合である Lambda Workshop を発展させたもので、それ以降は毎年会合が開かれている。また、Tech ワーキンググループと Control Plane ワーキンググループはその中間にも会合を持ち、議論を行っている。2004 年 9 月の英国の Nottingham までは invitation only で開催されたが、2005 年 9 月は San Diego で iGRID2005 に併設され、誰でも参加できるようになった。2006 年 9 月には第 5 章に示すように WIDE プロジェクトは NICT・NII と共にホスト役を努めた。

GLIF は現在 47 のネットワークや国際広帯域ネットワークの運用に関連深い大学などの組織が参加している。また、GLIF で用いることができる広帯域ネットワークリンクは、トラフィック交換の目的で幾つかの地点に集まっており、これらの集合地点での運用が Light Path の設定管理上重要であることが認識されてきた。これらの地点は GOLE — GLIF Open Lambda Exchange — と呼ばれており、表 2.1 の 15 地点が登録されている。Light Path の設定に関して、幾つかの GOLE を経由することが必要になるため、これらの運用は GLIF では極めて重要になる。そのため、GOLE の運用担当者は定期的に電話会議を行って、相互の調整に努めている。

第 3 章　T-LEX

T-LEX は IEEAF[3]太平洋回線を終端し、WIDE プロジェクトを含む国内の研究・教育ネットワークとの相互接続を行うため、2004 年から運用を開始した相互接続プロジェクトである。既存のネットワークのみならず、IEEAF のアジアへの回線の延長にも考慮した設計になっており、光プラットホームである Cisco ONS-15454 を中心に、Layer-2/3 スイッチであり、IEEAF OC-12c 回線の収容も行っている Foundry BigIron-15000、10GbE 対応スイッチである Foundry NetIron 40G、さらに OC-48c などの回線にも対応できる Catalyst-6500/SUP720 から構成されており、図 3.1 のように接続されている。

開設当初、IEEAF OC-192c 回線は、その対抗側である Seattle の Pacific Northwest Gigapop との関係から、主に STS-24 に帯域を分割し、Gigabit Ethernet を 8 回線通すことが想定されていた。これは OC-192c を Packet over SONET で収容できる機材が非常に高価で対抗で準備できなかったためである。

その後、10GbE WANPHY が一般化し、また、ちょっと高性能な PC でも Gigabit Ethernet を埋めることができるようになってくると、OC-192c の帯域を一つのアプリケーションで占有するような使

3　http://www.ieeaf.org/

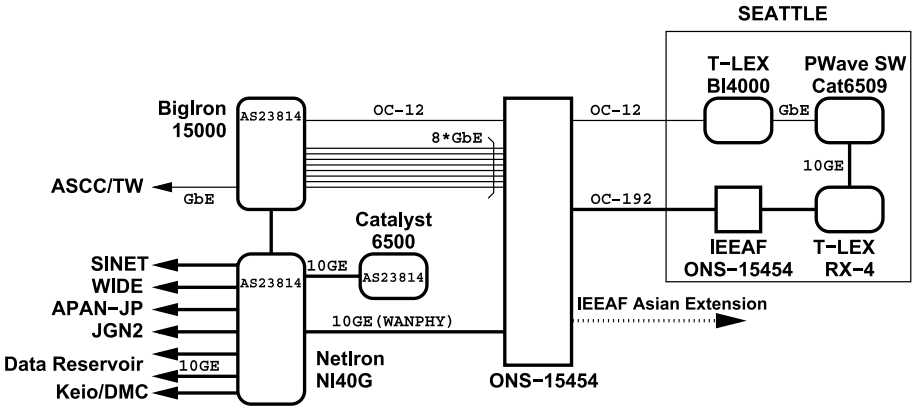

図 3.1. T-LEX の構成

い方が広まってきた。Gigabit Ethernet を対象に設計された BigIron-15000 では 10GbE WANPHY サポートや Jumbo Frame、転送性能の点で満足できないものであるため、NetIron-40G を追加した。ただし、IEEAF OC-12c 回線の終端は NetIron-40G ではできないため、引き続き BigIron-15000 が運転されている。

2006 年度は、ハードウェアの変更はなかったものの、SINET への 10GbE のリンクの敷設を行った。2007 年 2 月になってからであるが、従来の APAN-JP 経由だった WIDE と SINET の間の通信を、T-LEX を Layer-2 で経由の直接の接続に変更することができた。

第 4 章 Data Reservoir LSR

東京大学平木教授が率いる研究グループ Data Reservoir Project では、かねてから Internet2 Land Speed Record（LSR）[4] に挑戦してきた。2006 年度当初までに、IPv4 および IPv6 の両方のプロトコルにおいて、Single TCP ストリーム部門および Multiple TCP ストリーム部門の両方で LSR の記録を樹立し、Internet2 LSR 委員会によってその記録が認定されている。

このうち、2006 年 2 月 20 日に樹立した IPv4 の LSR は、LSR で認定される最大距離である 30,000 km を上回る伝送距離に対して、Single TCP で 8.80 Gbps を記録している。この記録を打破するためには、LSR の規則では 10%以上の性能向上が必要であるため、少なくとも 9.68 Gbps を達成しなければならない。これは単一の STM-64c/OC-192c 回線では決して得られない速度であり、10GbE LANPHY で 30,000 km を越える経路を設定するか、STM-256c/OC-768c を使えるようになるまで待たなければならない。

一方、IPv6 は TCP Offload Engine などのハードウェアのサポートをなかなか受けられないため、Data Reservoir の LSR は 6.96 Gbps に留まっていた。これを打破する試みが 2006 年の 12 月下旬に計画された。これは、クリスマス休暇の週は多くの研究者は休暇で活動をしていないため、GLIF の環境で必要な OC-192c の回線を東京〜United States〜Amsterdam の区間で二重に設定するためには、最適の季節である。単一の OC-192c 回線では、データストリームとその反対方向の ACK ストリームが帯域を奪い合うため、好ましい結果を得にくいと考えられている。

2006 年 12 月の GOLE 電話会議では、この LSR への挑戦に関する Light Path の設定に関して、その半分以上の時間を裂いて議論が行われた。その結果、大西洋回線に関しては SurfNET の Amsterdam〜New York 回線および Amsterdam〜Chicago 回線を使うのが他の活動との競合が少なく最適で、New York〜Chicago 間は CANARIE が新たに設定した OC-192c 回線、Chicago〜Seattle は National Lambda Rail の一波長を用いた TransLight（10GbE LANPHY）を使用することになった。つまり、図 4.1 のような構成となった。この構成は、T-LEX の NetIron-40G

4 http://www.internet2.edu/lsr/

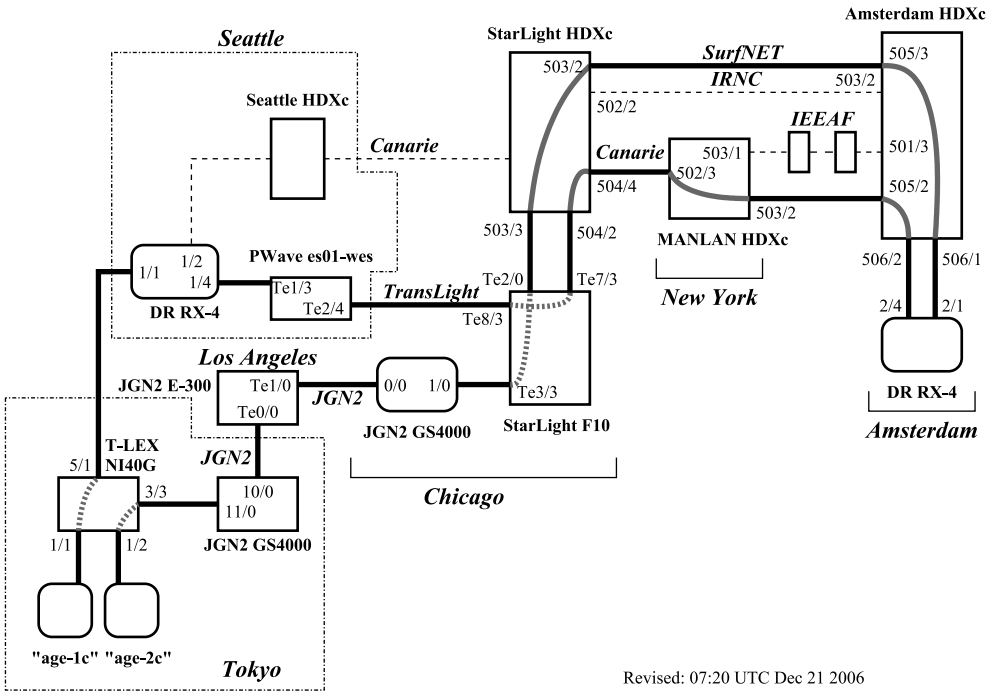

図 4.1.　DR LSR 実験のネットワーク構成

と StarLight の Force10 E-1200 以外は往復のストリームが競合する Layer-2 以上のデバイスが存在せず、LSR への挑戦に適した構成であるということができる。

Data Reservoir は、2006 年 12 月 30 日に IPv6 の Single TCP で 7.67 Gbps、翌 12 月 31 日には 7.67 Gbps を 10%以上上回る 9.08 Gbps を記録し、それぞれ Internet2 LSR 委員会へ申請を行った。

12 月 30 日の記録は、PCI-Express x8 を用いた Chelsio S310E-SR を使用したもので、メモリ間の転送を `iperf` によって実施したものである。IPv6 では TOE は利用できないが、S310E-SR の持っている出力パケットに関する pacing の機能を活用したものである。12 月 31 日の記録は、同じハードウェアを使用したものであるが、`iperf` を `mmap()` を使うように改造して、コピーオーバヘッドを削減したものを使用している。

この二つの LSR 候補は本報告書執筆時点で Internet2 LSR 委員会の判断はなされていない。前者は 2006 年 2 月の IPv4 LSR とほぼ同じ条件であるため、問題なく認定されると推測される。後者に関しては、プログラムを変更しているため、この点を LSR 委員会がどう判断するかに依存している。後者の記録が認定されると、それを打破するためには 9.988 Gbps の記録が必要になり、10GbE LANPHY でも不可能であると考えられる。

第 5 章　GLIF2006 関係イベント

5.1 GLIF2006

GLIF の会合を東京で行うことは、2004 年の Nottingham ミーティングの際、東京大学大学院情報理工学系研究科（当時）の青山教授と WIDE プロジェクトの加藤・篠田から Kees Neggers 氏らに提案を行い、それが承認されたものである。その結果、2006 年 9 月 11 日–12 日で秋葉原のダイビルで開催されることになった。

一方、U.S. の政府関連の研究ネットワークに関して光を用いたテストベッド構築に関する議論を行い、政府に報告書を提出することを目的としたワークショップ ONT — Optical Network Testbeds — は、2004 年に U.S. のネットワークの関係者のみによって開催されたが、国際的な展開が不可欠であることから、翌 2005 年に開催された ONT-2 は諸外国のネットワーク関係者も講演に招待された。我が国

からはJGNIIを代表し、NICTの加藤理事が講演を行っている。ONTとGLIFはその性格は異なるものの、講演をする人の集合は概ね一致したため、東京の同じ会場でGLIF2006の前、9月7日8日と開催され、120人が出席した。

GLIF2006はONT3と同数の約120人であったが、多少参加者の入れ替わりがあり、日本人の比率が少なくなり、国際的な色彩を帯びた会合になった。GLIF2006の講演資料等はGLIFのWeb Page[5]から入手することができる。

GLIF2007はCESNETのホストのもと、9月17–18日にCzechのPragueで開催されることになっている。

5.2 Global Lambda Networking Symposium

Lambda Networkのリーダが GLIF2006のために東京に集まるため、この絶好の機会を活かし、主に日本のネットワークコミュニティにLambda Networkingや GLIF を紹介する Global Lambda Networking Symposium を GLIF2006 の翌日 9 月 12 日に同じダイビルの 2 階ホールで開催し、200 名の参加を得た。午前中は主にGLIF2006の参加者からGLIFやLambda Networkingの関連技術の現状や問題点、将来像について同時通訳付きで紹介した。午後は我が国でのGLIFに関連する技術や活動や研究・教育ネットワークの状況などの紹介が行われた[6]。

第6章　GLIFの今後

帯域を占有することは、広帯域大容量データ転送を効率よく実施するためには効果的であり、特にCERNのLHC加速器やITERなどの巨大科学に関連するデータ転送には必須の技術であると考えられる。ただし、layer-3で経路制御を行うIPと異なり、end-to-endに直接通信路を確保した場合、途中のノードで回線品質のチェックを実施することは、SONET/SDHのoverhead byteを用いるしか方法がなくなる。また、わずかなジッタが受信側のバッファを非常に短い時間ではあるが溢れさせ、パケット損失に繋がり、転送性能を劣化させてしまう。このため、GLIFユーザをサポートする技術として、10 Gbpsの帯域で使い物になり遠隔操作が可能なパケット監視などの開発が必要になると考えられている。

5　http://www.glif.is/meetings/2006/
6　http://www.e-side.co.jp/glifsymposium2006/