第４部

特集4 100Gbps級ネットワーク

加藤 朗

第１章　背景および要求点

インターネットの帯域はARPANET時代の56kbpsから飛躍的に増加した。一部の限られた研究者のみだったユーザも、スマートフォンを含む携帯電話からのアクセスまで含めると、数十億人に増加している。これらに対して安定したアクセスを提供するため、ISPのバックボーン回線も、T1/E1(1.5Mbps/2Mbps)から10Gbpsの回線を複数束ねて使うまでに拡張されてきた。

ISPのバックボーンは、非常に多い数のユーザにインターネットアクセスを提供している。つまり、ユーザ一人あたりにしてみれば、インターネットを使っていない時間も多く、webなどをアクセスしている場合でも、実効的な転送レートは100Mbpsに達していないのではないだろうか。

一方、教育・研究ネットワークは事情を異にする。University of AmsterdamのCase de Laat教授によれば、教育・研究ネットワークでは非常に少ない数のユーザが非常に大きな帯域を使っている。研究に必要なデータの発生源は、数は多くないものの、発生するデータ量が膨大であることが多く、地理的な制約がある場合も少なくない。例えば、高エネルギー物理学では、スイスのCERNにあるLHCという加速器が国際共同プロジェクトとして敷設されており、実験中は多量なデータが発生する。効率のためにはCERNに研究者を集めるべきかも知れないが、世界中の高エネルギー物理学の研究者を、多数の学生を含めて集めるのは現実的ではない。データを加工することでデータ量を圧縮することはできるが、最先端の研究者は、可能な限り加工前のデータを分析したいと考えている。

これは、ハワイに設置されている光学望遠鏡からのデータに依存する天文学者でも同じであり、気象条件などの点から望遠鏡は任意の地点に設置できるわけではない。さらに電波天文学では、世界中に分散している電波望遠鏡からのデータを分析する必要があるが、このデータはほとんどがノイズであり、非常にたくさんのデータから必要な情報を抽出している。ここでは情報工学的なデータ圧縮は無力である。

また、計算機科学やその他の分野の、計算機を多用した研究では、常時大きな帯域を必要としていないことも多いが、国際会議などの場でのデモンストレーションでは、研究成果をアピールするとともに、今後の研究資金獲得のためには重要であるため、広帯域のネットワークが必要になる。

電波天文学などのいくつかの例外はあるが、多くの場合、データの損失は研究に深刻な影響を与えることも多い。特にデモンストレーションでのパケット損失の発生は深刻な転送効率の低下を招くため、非常に高い信頼性が要求される。帯域がそれほどではない場合ではForward Error Correctionなどの手法を使うことができるが、10Gbpsを越える世界ではFECのような技法の適用は容易ではない。また、セキュリティ的に十分な対策をすることが容易ではないことも多く、転送地点間を接続する大容量の専用のネットワークが得られればよいが、このような回線の維持・運営には巨額の予算が必要なため、簡単ではない。

この問題を解決するために、教育・研究ネットワークの国際的な連携が重要になってきた。2001年から始まったGLIF - Global Lambda Integrated Facility[*1]- は、それ自身は

＊1　GLIFは2001年にAmsterdamで開催されたGlobal LambdaGrid Workshopに端を発しており、GLIFという名前とhttp://www.glif.is/というURLは2003年にReykjavikで開催されたGlobal LambdaGrid Workshopのときに定められた

何の資産も有していないが、大容量の国際回線を運用している多くの教育・研究ネットワークが参加している。

教育・研究ネットワークの多くは政府からの資金でネットワークが運用されており、したがって、その利用はその国の研究者に限られることが多い。GLIFでは、もちろんその国の研究者の利用を優先するものの、デモンストレーションや実験などの一時的な利用に関しては、使われていない帯域を利用して、通過するトラフィックも容認するという相互の理解のもと、その度に回線を調達する必要を省き、研究活動を進めることができる。

IEEAF - Internet Educational Equal Access Foundation[*2] - は、海底ケーブル事業者であるTyco Telecomに対して、敷設された海底ケーブルの使われていない帯域の寄贈を要請した。太平洋に関しては、2002年に千葉県の江見とオレゴン州のHillsboroの間に敷設されたTGN-Pacificの帯域と、両端のbackhaulの回線を含めて、東品川のTycoのデータセンターとSeattleのWestin Buildingの間にOC12(622Mbps)およびOC192(9.6Gbps)の回線が提供された。2002年に提供が始まった大西洋回線(New YorkとオランダのGronigen)と同様に、5年間の期限付きの寄贈であった。このIEEAF太平洋回線は、University of Washingtonが運用するPacific Northwest Gigapopと、東京側はWIDE Projectがその運用に当たった。Pacific Northwest Gigapoには、既にU.S.の教育研究ネットワークであるInternet2やCanadaのCANARIE、AustraliaのAARNETなどのネットワークが既に10Gbps級の回線で接続されており、Pacific Northwest Gigapopのスイッチ経由で、これらの教育研究ネットワークと相互接続をすることができた。東京端はTokyo Lambda Exchange (T-LEX)として、WIDEやSINET、JGN、APAN-JP等との接続を行った。

当時のネットワークの利用に関しては、Time Domain Multiplexor (TDM)を用いて、帯域保証された複数の回線に帯域を分割し、それを接続地点で相互接続する、いわゆるLambda Network方式と、TDMではなくEthernet VLANによって回線を論理的に分割はするが、帯域保証はしない方式の間で揺れていた。そのためT-LEXでは、OC12回線を収容するL3スイッチ、OC192をOC12 4つに分割することもできるTDM装置の他、OC192を10Gbit Ethernet WANPHYで直接収容できるL3スイッチを準備した。

TDMによる帯域分割は、帯域保証はされているものの、安価なEthernetではなく、OC3 (155Mbps), OC12 (622Mbps)、OC48 (2.4Gbps)といった高価なSONET/SDHインターフェースが必要であり、また、仮に1Gbpsの帯域が必要だとしても、2.4Gbpsの帯域を使わざるを得ないなど、TDMの設定はルータに比べて面倒な点があった。VLANを用いた方式は帯域保証はないものの、他にトラフィックがない時には大きな帯域を使うことができること、対地までVLANを延伸することによって、一般のインターネットとは論理的に独立した接続性が得られるため、セキュリティ装置などを省くこともできた。どのため、いくつかの例外はあったものの、概ね9.6Gbpsの回線は丸ごと10Gbps WANPHYとして扱われることが多かった。

この回線は、定常的なトラフィックは搬送しなかったため、全部の帯域を実験等に用いることができた。そのため、東京大学平木教授らによるData Reservoirプロジェクトでは、この回線を含めGLIFの参加ネットワークの回線を合わせて用いることにより、2006年12月に達成した記録[*3]は、Internet2のSingle Data StreamおよびMultiple Data Streamの各部門で、IPv4およびIPv6の両方でLand Speed Recordの認定を受けるに至った。

このIEEAF太平洋回線は5年間の寄贈の期限が切れた2008年に終了したが、国際的な広帯域のネットワークの協調運用の経験から得られたものは少なくなかった。

第2章　TransPAC/Pacific Wave 100Gbps回線

2000年代は10Gbpsあるいは9.6Gbpsがインターネットの回線として使える上限だったが、2010年に入ると、40Gbps Ethernetおよび100Gbit Ethernetの規格の制定と対応する機器の開発が進んできた。40Gbpsは、当初はデータセンターのサーバへのアクセスとして用いられる

＊2　http://www.ieeaf.org/
＊3　http://data-reservoir.adm.s.u-tokyo.ac.jp/press-20070508.html

ことが想定された。インターネットのバックボーンとしては40Gbpsでは不十分であるということから100Gbpsが注目され、Internet Exchangeでも100Gbpsのサービスを提供するところが現れ始めた。

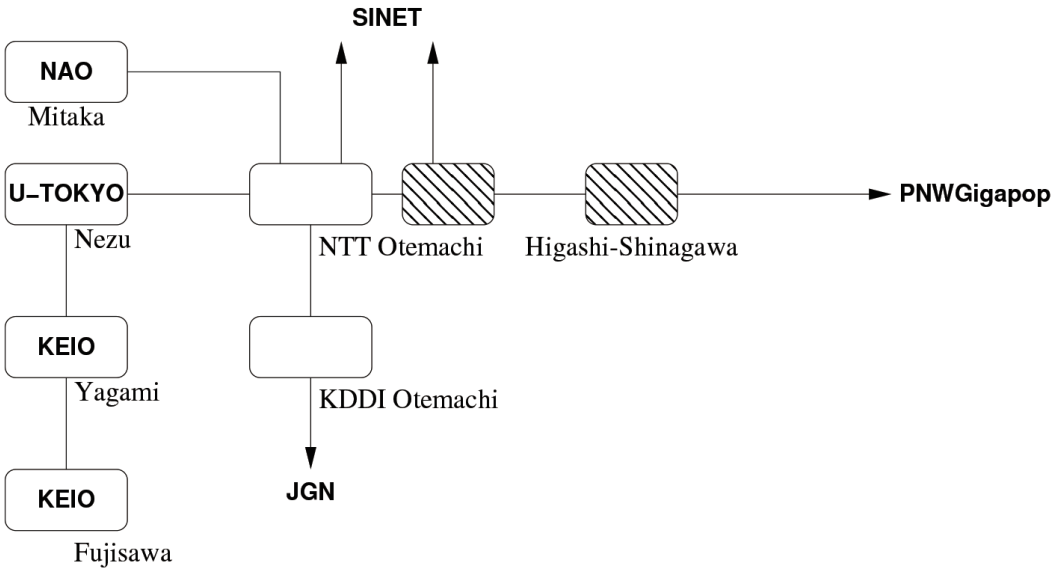
国際的な長距離回線での100Gbpsの利用は必ずしも容易ではない。多くの光ファイバーはOC192を基本帯域として設計されていたため、これを100Gbps Ethernetに対応させるには、ファイバシステムの大規模な改修が必要になる。また、提供される回線のコストや100Gbpsに対応した機器の安定性などの問題もあり、商用ISPで採用するのは必ずしも容易ではなかった。

教育・研究ネットワークでは、単一の研究プロジェクトですら10Gbpsでは足りないデータ転送を必要とするものがあった。2013年7月にオランダのMaastrichtとNew Yorkの間で100Gbpsの回線によるデモンストレーションが行われ、翌年からANA-100Gとしての大西洋回線の運用が始まった。ただし、10Gbpsネットワークとは異なり、100Gbpsネットワークリンクはなかなか安定せず、教育・研究ネットワークといえども一般的に利用できるようになるまでに時間を要した。この回線は2014年11月までには、London・Washington D.C.間の100Gbpsリンクと合わせてリング状のトポロジとなり、単一故障点のないANA-200Gに拡張されている。[*4]

2015年にUniversity of WashingtonのRon Johnson教授は、TGN Pacific回線を運用しているTata Communicationsに要請し、100Gbps回線の寄贈を受けることになった。この回線の運用に関しては、IEEAF時代の9.6Gbpsの回線と同様に、University of WashingtonとWIDE Projectが協調して実施している。また従来から東京とLos AngelesをNSFの予算で9.6Gbpsで結んでいたTransPAC回線は2016年5月末で廃止され、この回線でTransPACの機能も提供することになった。そのため、回線の正式名称としては、

TransPAC/Pacific Wave太平洋回線

と称している[*5]。また、伝送装置や100Gbps回線を収容できるL2/L3スイッチに関してはHuawei Technologiesから、L3スイッチに関してはCisco Systemsから提供を受けている。また東品川と大手町の間の光ファイバーに関



注：ハッチングは CENIC 提供の機材を表す

図2.1 100Gbpsネットワーク

＊4 https://www.nordu.net/content/ana-100g
＊5 http://cenic.org/news/item/pacific-wave-announces-worlds-first-trans-pacific-100-gigabit-re-network

しては、東品川側の伝送装置の設置場所も含めて、国立天文台が提供している。

このTransPAC/Pacific Wave回線はTGN-Pacificを用いたものであるが、2016年4月にはSINETは、Los Angelesへの回線を100Gbpsに増強した。この回線はJapan-USケーブルシステムを用いたものであり、両者が同時に利用できなくなる可能性は非常に低い。そのためSINETのバックアップをTransPAC/Pacific Wave回線で行い、また、TransPACの VLANをSINETでバックアップする運用が行われている。

従来のIEEAFとは異なり、PacificWaveの主宰の一つであるカルフォルニア州の教育・研究ネットワークであるCENICは、100Gbpsスイッチを大手町と東品川に一台づつ設置した。これらの運用・監視はPacific Wave側で実施している。ただし、WIDE Projectではremote handとして、障害発生等で現地作業が必要なときなどには、ベンダとともに協力することになっている。

このTransPAC/Pacific Wave回線に関しては、どのような利用が許されるのか、ということは開通当初は必ずしも明確ではなかった。これに対して、2016年5月にChicagoで開催されたInternet2 Global Summitの際に、University of WashingtonのRon Johnson教授およびJonah Keough氏、APAN-JP、SINET、JGNおよびWIDE Projectの関係者からなる非公式な会合が開かれた。この会合で、Johnson教授は回線の背景について説明し、また回線そのものはTata Communicationsからの寄贈であって特段の制約を受けないものの、Seattleなどの機材がNSFの補助を受けている関係で、任意の目的での利用が許されるわけではないが、学術研究であれば問題がない旨の説明がなされた。

太平洋における100Gbps回線としては、2015年末に開通したLos Angeles — Singapore（Intenet2とSingaRENの共同プロジェクトである）が知られており、また韓国の教育・研究ネットワークであるKREONET（運用母体は、スーパーコンピュータセンタも運用しているKISTIである）は、従来のDaejeon — Seattle 10Gbps回線を更新し、Daejeon — Chicagoに100Gbpsの回線を開設している。また、Seattle — Sydney間のSouthern Cross海底ケーブルを用いているオーストラリアの教育・研究ネットワークAARNETでは、10Gbpsの回線を40Gbpsに拡張しており、100Gbpsへの増強も予定されている。

第３章　運用上の問題点

2000年代では10Gbpsは最も高速な一般的に利用可能なネットワーク技術であったが、1Gbpsのネットワークのつもりではうまく動作しないことも多く見られた。特に光ファイバの品質は重要であり、パケット損失などの問題は光コネクタでの端面の清掃で解決したことも多かった。そのため、1Gbpsのときには気楽に扱うことができた光ファイバも、細心の注意に必要になった。現在は、10Gbpsは特別なものではなく、長距離伝送はできないものの、UTPによる伝送も可能になり、普通に秋葉原で市販されているPC用のマザーボードの中には10Gbpsネットワークのポートがついていたりする位である。

一方、100Gbpsのネットワークは、多くの場合、100GBase-LRが使われている。これは25Gbpsのネットワークを４波波長多重している。基本的な転送レートが2.5倍になっているため、10Gbpsの場合に比べてより慎重な取扱いが必要になる。

前述のANA-100Gの大西洋100Gbpsネットワークでは、しばしば回線がダウンしたり、エラー率が高くなったりするトラブルが発生していた。陸揚げ局などの電気通信事業者の区間を含めて、光ファイバ端面の清掃によってトラブルが解消することが多かったとのことである。

我々のネットワークの例外ではない。例えば、東品川のデータセンターでは、L3スイッチとTata Communicationsの機材間で、数分から数十分経過すると、L3スイッチ側のインターフェースがダウンする現象が発生した。特定のポートやカードに依存した問題ではないことは直ぐに確認できたが、ベンダに状況を報告しても現象の再現が容易ではないためか、数ヶ月が経過した現時点ではこの問題は解決していない。そのため、止むなく、L3スイッチを一時的に外し、大手町への回線を

直接接続することで対応している。

また、ある機器の組み合わせによっては、Symbol Errorが発生する現象が確認されている。これは、100Gbps Ethernetにはありえないbit patternを受信したというエラーで、その瞬間にたまたまパケットが送られていればCRCエラーになり、そちらのカウンタも増えることになる。異なるカード上のポートに接続を変更してもその頻度は多少上下するものの、0にはならなかった。この問題は、接続している機器同士の同期問題とも考えられるが、接続している機器のベンダが異なる場合にはなかなかトラブルシュートができない。また第３のベンダの機器を接続したら問題は発生しなくなるため、どちら側の問題かを識別することはできなかった。なお、この機器の対応するポートを収容するラインカードを、別のトラブルシュートのために抜き差ししたところ、この問題が観測されなくなっている。

100Gbpsは伝送速度が非常に早いので、例えば、カードのバックプレーンへのコネクタに僅かな電気抵抗が発生したとしても、それが信号伝搬の遅延をもたらし、エラーになる可能性もある。そのため、ベンダや接続相手の機器の所有者と綿密に調整の上、トラブルシュートを実施する必要がある。運用の場面でも、10Gbps時代よりも、より協調的な関係が求められている。

第４章　今後の展開

4.1　Global Network Alliance

広帯域回線を国際的に、かつ戦略的に調達し、全体として研究の発展に資するため、GNAという会議が召集されている。選ばれた国の教育・研究ネットワークを運営する母体の長が出席するCEO Forumとその下にworking groupから構成されている。わが国からはSINETを運用する組織である国立情報学研究所(NII)の所長が出席すべきであるが、NIIはSINETを運用しているのは一つの部門に過ぎないため、同部門の安達教授がCEO Forumに出席しており、working groupには漆谷教授と中村教授が出席している。

なお、この問題はSINET単独で解決できる問題ではないため、国内のForumとしてNIIはJGN、APAN-JP、およびWIDE Projectからなる会合を召集し、GNA会議の動向およびわが国としての対応について議論を行っている。

4.2　Guam

100Gbps回線、特にアジア太平洋地域では、それをサポートする海底ケーブルに依存している。古い海底ケーブルシステムでは十分な空き容量がないことも多いが、近年の、主に一般のインターネットに起因する需要の急増により、いくつかのファイバシステムが敷設され、あるいは敷設が計画されている。これらの新しい光ファイバシステムは10Tbps以上の帯域を有していることが多く、100Gbps級ネットワークにとっては重要な基盤になる。

従来のように、アジア地域各国からの回線をUSの西海岸に対して敷設し、そこで相互接続を実施するのは、パケットの伝送遅延が大きく必ずしも最適な方法ではない。そこで太平洋のどこかで相互接続を図ることが検討されるようになった。その一つの案が、Guamに相互接続点にするというものである。Guamは多くの海底ケーブルが陸揚げされており、そのうちのいくつかは100Gbpsの回線をサポート可能である。この可能性に関して、現在University of Hawaiiの学長をしているDavid Lessner氏によって非公式な会議が召集され、議論が行われている。もしGuamでの教育・研究ネットワークの相互接続を行うことになった場合には、以下のような可能性があることが分かっている：

- University of Hawaiiは、Honoluluと Guam間　の100Gbps回線を調達できるように調整を進めている。この回線の一つの目的は、太平洋の島々へのデジタルデバイドの解消であり、海底ケーブルだけではなく、スループットの高い次世代人工衛星も併用することも議論されている。

- U.S.のInternet2とSingapore SingaRENが共同で運営しているLos Angeles — Singapore間の100Gbps回線をGuamに一旦陸揚げする形態にすることが検討されている。
- AustraliaのAARNETは、Guamに対して広帯域回線を調達する方向で調整が行われている。

わが国では、直ちにGuamに対して100Gbps回線を調達できる可能性は高くないが、上記のような計画が実現し、相互接続ができるようになった場合、得られる利益は少なくない。いくつかの大規模電波望遠鏡が設置されているAustraliaからのデータ転送が可能になる他、現在Seattle経由で調整されている国立天文台のHawaii島にあるすばる望遠鏡に対して、より低遅延でのアクセスが可能になる。

しかしながら、以下のような問題も指摘されている：

- Guamで大量の研究データが発生するわけではないので、予算措置は必ずしも容易ではない。

- Guamには、たくさんの海底ケーブルが陸揚げされ、また新たな海底ケーブルの敷設も予定されているが、陸揚げ局は島内にいくつか分散して設置されているために、相互接続点までの島内ファイバの調達可能性が問題になる。

- Guamにおける研究機関としてはUniversity of Guamしかないため、相互接続点をUniversity of Guamに設置するか、あるいは陸揚げ局を含めてた事業者のコロケーション設備にするかは、設置や運用、拡張などの種々の視点での議論が必要である。

第５章　まとめ

以上の様に、教育・研究ネットワークでは、そのバックボーンや国際回線の100Gbps回線への更新が多く行われるようになってきており、特に大容量のデータ転送が必要な大規模科学に対しては力強いツールになってきている。しかしながら、100Gbps Ethernetは、まだまだ接続すれば直ちに利用可能になるほどは安定していない。前述のように、さまざまな問題が影響する。そのため、100Gbps級のネットワークを運用するオペレータは、学術系、商用系を問わず、情報交換を行い、複数のベンダと協力して問題を解決していく態度が重要になる。