大阪大学大型計算機センターにおける次世代インターネット・QoS技術への取り組み

大阪大学大型計算機センター
門林　雄基

1　はじめに

次世代インターネットは現在のインターネットと比べて 1000 倍以上の帯域が利用可能であるため、今までとは異なる利用方法が考えられている。例えば、次世代インターネットを用いてテレビの画像を配信することや、従来はシステム内にとどまっていた並列機のプロセッサ・インタコネクトを地理的に離れた場所まで拡張することが可能である。

筆者らのグループでは Grid Toolkit と呼ばれる、次世代インターネット上のアプリケーションのための汎用ミドルウェアについて他の研究機関との協調実験をおこなっている。Grid Toolkit を用いることでインターネット規模での動作時におけるセキュリティ問題、CPU やネットワークといった資源の割り当て問題からアプリケーションを解放することができる。また、そのような次世代のアプリケーションではネットワーク通信品質のばらつきが問題となる。このための基盤技術である Diffserv についても開発および実験をおこなっているので、これについても紹介する。

これら Grid Toolkit, Diffserv といった基盤技術を用いたアプリケーションの一例として Telemicroscopy (電子顕微鏡の遠隔観察と三次元標本情報の復元) を紹介する。

2　Grid Toolkit – 次世代インターネットとアプリケーションの架け橋

次世代インターネットでは、現在のインターネットと比べて 1000 倍以上の帯域が利用可能である。これは、ルータに並列計算機などのスイッチ技術が積極的に採り入れられていることと、光波長多重 (WDM; Wavelength Division Multiplexing) によってファイバあたりの容量が飛躍的に増大することにより、現実のものとなりつつある。また、光インターコネクト技術と MPLS 技術の融合などにより、さらなる飛躍的な成長が期待できる。

インターネットの速度が 1000 倍になれば、並列計算のような高性能・低遅延な通信路を必要とするアプリケーションもインターネット上で動作させることができるかもしれない。次世代のインターネットでは、電子メールのような人対人のトラフィックだけではなく、あるタスクをこなすために複数の計算機が有機的に連結されて動作し、機械対機械のトラフィックをやりとりすることが一般的になるだろう。

このような、次世代インターネット上に様々な機能をもった計算機が分散している環境で、それらを有機的に組み合わせ、大きなタスクをなしとげることは簡単ではない。かりに専属のオペレータがいたとしても、たとえば彼は科学計測機器からのデータの取り出し、データ形式の変換、スーパーコンピュータへのジョブ投入、結果のワークステーションへの転送、可視化といった繁雑な手続きを踏まねばならない。また、ファイル転送のプロセス等を簡単にするため、セキュリティを犠牲にしてしまう場合も多い (例えば、特定ホストからだけパスワードなしでログインできるよう設定するなど)。また、複数のスーパーコンピュータで同時にシミュレーションを走らせ、より大規模な問題を解くといったことは現在のシステムでは難しい。

これら、機能分散と利便性のトレードオフ、分散システムの利便性とセキュリティのトレードオフ、複数の計算機にまたがる CPU 資源の同時割り当て、といった問題は次世代インターネットが現実のものとなり、組織の壁を超えるような広範囲にわたる機能分散が必要不可欠になった現在、避けて通れない問題となりつつある。

Grid Toolkit はこれらの問題を包括的に取り扱うミドルウェア (アプリケーションと基本ソフトウェアの間に位置するソフトウェア) である。特に、筆者のグループが他の研究機関と協調実験をおこなっている Globus (Grid Toolkit のひとつ) は、機能分散されたサブシステム間の連結機能、公証局と個人証明書にもとづく強力な認証システム、利便性とセキュリティを両立する Single Sign-On, 複数の計算機にまたがる資源割り当て機構などを備える。

このような機能を実現するミドルウェアは Globus に限らないが、Globus は特に High Performance Computing を対象として設計されており、当センターのソフトウェア・プラットフォーム、センターの目的とも合致するため実験的に運用・評価をおこなっている。

評価環境として、まず Globus を HP Exemplar (HP-UX 11.0) に移植し、APAN と vBNS のネットワークを用いて動作確認を行った。この環境を用いて SC'98 の iGrid Demonstration の一貫としてデモンストレーションを行った。これは、米国 Argonne National Lab と ISI にて作成されたマイクロトモグラフィ実験のためのアプリケーション [1] の解析部分を当センターの HP Exemplar で実行し、可視化部分を SC'98 会場の Onyx で実行するというものである。筆者が移植に用いたコードは Globus v1.0 以降のリリースに含まれている。

また、大阪大学医学部田村研究室 水野 (松本) 由子博士 (現在 Johns Hopkins Medical Inst.) らの研究グループとともに、脳磁計 (MEG) からのリアルタイムでのデータ取得、解析、可視化に取り組んでいる。これは脳磁計の 64 部位に設置されたセンサーから得られる時系列データを、119 のセンサーの組み合わせについて Wavelet に基づく手法で相関を解析するというものである。この処理をシングルプロセッサで行うと膨大な時間を要する (例えば 8 秒分のデータで 12 時間) が、これを Globus を用いて細粒度の並列処理 (MPI) と粗粒度の並列処理 (job co-allocation) へ分割することで、ネットワーク上の複数の並列計算機を利用することができ、32CPU を用いた場合で解析時間を 16 分の 1 まで短

縮することができた。また、Globus を用いることで解析しながらの可視化も可能となった [2, 3]。

Grid Toolkit を用いることで、次世代インターネットを活用したアプリケーションを比較的容易に作成することができる。しかしながら、Grid Toolkit をより効率的に動作させるためには、大容量のデータ転送を高速に行えることと、プロセッサ間通信において遅延を最小限に抑えることが必要となる。このための技術 (帯域保証やパケット優先制御の仕組み) は QoS 技術と呼ばれている (QoS: Quality of Service)。これを次世代インターネット上に実現することは必要不可欠である。

また、次世代インターネットにおけるその他の有望なアプリケーションとして、電話網のインターネットへの融合、放送網のインターネットへの融合が挙げられる。しかし、インターネット電話では遅延が問題となり、また、放送網への応用では遅延のゆらぎやパケット廃棄率が問題となる。これらのアプリケーションは、遅延や帯域といったトラフィック特性の面で従来のデータ・トラフィックとは異なった要求を持っているのである。このような新たなアプリケーションと従来のアプリケーションがうまく共存していくためにも、QoS 技術はなくてはならないものである。

3 Diffserv – 次世代インターネットの核となる QoS 技術

このような、従来のデータ転送と音声・ビデオの配送などを異なるサービス要求としてとらえ、それぞれについて一定のサービス品質 (Quality of Service) を達成するという考えは、最近になって出て来たものではない。サービス品質の絶対的な保証 (hard guarantee) はインターネットのような大規模なネットワークでは現実的ではないため、長い間インターネットで採用することを見合わせていた。しかし近年、QoS 研究の分野においてインターネットの特性を考慮した「サービス品質の統計的な差別化」という考え方が提唱され、これを実現するための仕組みについてめざましい進展がみられた。IETF (Internet Engineering Task Force, インターネットの技術標準化団体) では、これらの知見をもとに Diffserv (Differentiated Services) という方式の標準化がすすめられている。

表 1: 従来方式と Diffserv の転送速度の測定結果

	従来方式	Diffserv
ビデオ (DV/UDP)	11.6 Mbps	32.6 Mbps
データ (netperf)	31.5 Mbps	9.38 Mbps

現在のインターネットでは、インターネット上をながれるデータはすべてパケットに分割され、ルータにおいてパケット単位で経路選択がおこなわれ、到着した順にパケットが出ていく。また、ルータ内のキューが一杯になり、パケットを廃棄しなければならないときは、キュー内のパケットから廃棄すべきパケットを公平に選ぶ。これに対し、Diffserv

方式にもとづくルータでは、パケットに付けられた優先制御コード (diffserv codepoint) をもとに、パケットの送出順序、廃棄順序を決める。この codepoint をアプリケーションの要求にしたがって使い分けることで、パケットの遅延をできるだけ抑えることや、データ転送速度があまり遅くならないよう制御することが可能となる。

codepoint をパケットに書き込むのはユーザの端末機器と、プロバイダのユーザ収容点に置かれたルータ (DS 境界ルータ) である。DS 境界ルータではユーザとプロバイダの契約にもとづいて codepoint を設定し、プロバイダ内部のルータ (DS 内部ルータ) では codepoint にもとづいて契約に見合った優先制御をおこなう。

細かい説明は省略させていただくが、パケットの送出順序、廃棄順序を制御するためのさまざまな仕組みを組み合わせることで、インターネットが提供しているサービスの「幅」が広くなると御理解いただきたい。

一例として、筆者らの研究グループが取り組んでいる Diffserv ルータの実装を紹介する。これは FreeBSD におけるトラフィック優先制御の仕組み (ALTQ と呼ばれるフリーソフトウェア) に Diffserv 対応機能を付加したもので、DS 境界ルータとしても DS 内部ルータとしても動作する。

これをルータとして用いた小規模な実験ネットワークを構築し、データトラフィックとビデオトラフィックを同時に流し、データとビデオの望ましくない干渉を避けることができるかどうか、実験を行った。ビデオトラフィックとしては民生用の DV 機器から発信された映像信号 (IEEE-1394 パケット) を UDP パケットに変換したものを用いた [4]。この方式で、ビデオをブロック欠損なく完全に再生するためには 32.6Mbps が必要である。また、データトラフィックとしては netperf というテスト・トラフィック発生ソフトウェアを UDP_STREAM モードで用いた。

このときの測定結果を表 1 に示す。従来方式を用いた場合はビデオの転送速度がデータトラフィックの干渉をうけ、著しく低下したのに対し、Diffserv を用いた場合は codepoint の設定によって優先制御をおこなうことができ、データトラフィックを抑えてビデオの品質を保つことができた。

上で示したものは Expedited Forwarding と呼ばれる Diffserv のパケット配送方式のひとつで、単なる優先配送に過ぎないが、このほかに、最低帯域の統計的保証と余剰帯域の分配をおこなうための Assured Forwarding と呼ばれる配送方式についても実験をおこない、動作を確認している。

4 Telescience – 次世代インターネットの科学への応用

筆者らの研究グループでは、Grid Toolkit と Diffserv という次世代インターネットの二つのキーテクノロジの、遠隔地からの科学観察 (Telescience) への応用に取り組んでいる。

これまでに、大阪大学超高圧電子顕微鏡センターの森教授、米国 UCSD の Mark Ellisman

教授らによって、同センターが有する超高圧電子顕微鏡を次世代インターネットを経由して米国 UCSD で操作し、遠隔地から科学観察を行うという実験がなされている。筆者らの研究グループはこのうち、科学観察のための画像伝送方式としてインターネット上での DV 伝送を用いることを提案し、DV 伝送システムの運用、APAN, vBNS などとの技術的調整を担当した [5]。

DV を用いることで MPEG-2 などの画像伝送方式と比べて飛躍的に低遅延かつ高画質の画像伝送を行うことが可能であるが、APAN と vBNS の接続ポイント (StarTAP) において従来のデータトラフィックとの競合が起きており、動画像においてブロック欠損がみられるなど科学観察にとって望ましくない結果となっている。Diffserv を用いることで DV トラフィックを優先制御することができ、この問題を解決することができる。筆者らの研究グループでは Diffserv に関する技術開発ならびに運用推進に寄与することで、動画像とデータが混在するような Telescience プロジェクトにひろく寄与することができると考えている。

また、超高圧電子顕微鏡などからの科学観察で得られた、撮像角度の異なる二次元の高精細画像を複数枚重ね合わせ、三次元画像を復元するといった手法が確立されているが、この処理には多大な計算時間を要する。このプロセスに Grid Toolkit を応用し、地理的に分散した複数台のスーパーコンピュータで処理し、解析時間を劇的に短縮するといったことも考えられている。筆者らの研究グループでは Grid Toolkit の移植、改良、Globus ディレクトリサービスの日本での運用、アプリケーションの Grid Toolkit への対応 (部分書き換え) といった作業を通じて、Telescience プロジェクトにおける次世代インターネットの活用、Telescience プロジェクトにおける計算時間の短縮をはかることができると考えている。

5 おわりに

本稿では次世代インターネットにおける二つの重要な技術である Grid Toolkit と Diffserv について簡単に紹介した。また、Grid Toolkit の応用として、脳磁計 (MEG) からのリアルタイムでのデータ取得、解析、可視化への取り組みについて述べ、Diffserv の応用として大阪大学超高圧電子顕微鏡センターが有する超高圧電子顕微鏡の遠隔観察について述べた。

さまざまな科学計測機器が計算機に接続され、さらにインターネットに接続されるようになった現在、計算機とインターネットが先端科学研究にたいして果たしうる役割はかつてないほど大きい。また、情報技術が高度に進化しつづけているため、最先端の情報技術を用いた場合と、そうでない場合とを比べると科学研究のプロセス全体のスピードに大きな差が生じる場合すらある。

筆者らの研究グループでは、先端科学研究に対して最先端の情報技術を適用することと、そのための技術開発の双方が重要なミッションであると考えている。しかしながら、大阪大学大型計算機センターは運用センターとして位置づけられており、これらのミッション

を遂行するための人手はまったくないというのが実情である。科学にとっての情報技術の重要性と、情報技術のますますの高度化という長期的な傾向を鑑みれば、情報技術の応用(Applied IT)にたいし一層の研究開発人員を割り当てるべきであると考えられる。

参考文献

[1] Gregor von Laszewski, Mei-Hui Su, Joseph A. Insley, Ian Foster, John Bresnahan, Carl Kesselman, Marcus Thiebaux, Mark L. Rivers, Steve Wang, Brian Tieman, and Ian McNulty. Real-time analysis, visualization, and steering of microtomography experiments at photon sources. In *Ninth SIAM Conference on Parallel Processing for Scientific Computing*, April 1999.

[2] Y Mizuno-Matsumoto, S Date, Y Tabuti, R.A. Zoroofi, S Shimojo, Y Kadobayashi, H Tatsumi, H Nogawa, K Shinosaki, M Takeda, and T Inoue. Integration of signal processing and medical image for evaluation of brain function on globus. In *Proceedings of IWS'99*, pages 297–302, February 1999.

[3] 伊達進, 水野（松本）由子, 田村進一, 佐藤嘉伸, アガイザデ ゾルフィー レザ, 田渕裕士, 下條真司, 門林雄基, 辰巳治之, 野川裕記, 篠崎和弘, 武田雅俊, 井上健, and 宮原秀夫. 脳磁計（ＭＥＧ）用広域分散コンピューティング環境. *Medical Imaging Technology*, 1999. to appear.

[4] Katsushi Kobayashi. Design and implementation of firewire device driver on freebsd. In *1999 USENIX Annual Technical Conference*, June 1999.

[5] Seeing submicron structures from 10,000 kilometers: Trans-pacific microscopy enabled by international advanced networks. http://www.npaci.edu/online/v3.10/, May 1999. NPACI Online.