

小特集 ITER に向けた遠隔実験の現状

3. 高速データ通信技術と遠隔実験データ解析開発の現状と展望

3. Technologies for High-Speed Data Transfer and Remote Experiments' Data Analyses

山本孝志, 江本雅彦, 中西秀哉

YAMAMOTO Takashi, EMOTO Masahiko and NAKANISHI Hideya

核融合科学研究所

(原稿受付: 2013年5月23日)

ITER など次世代装置においては遠隔実験という研究形態が大幅に普及すると考えられるが、これを成立させるには、遠距離高速通信を基盤とした大量データの参照・共有と、サイト内外の差を感じさせないデータ解析等研究環境の開発、提供が不可欠である。現在の国際回線帯域 10 Gbps をフル活用する TCP データ通信技術はほぼ実証されて来ており、将来構築される ITER 装置～遠隔実験センター間の回線は 100 Gbps 超になると想定され、それに向けた研究開発が進められている。また日欧間の回線経路や VPN の種別についても複数の可能性が検討されている。クラウド・コンピューティング技術により、インターネット上にデータ解析環境をすべて実現し、軽量端末でどこからでも同一環境にアクセスできる状況が実現しつつある。カギとなるのが各種の仮想化技術であり、精力的に研究開発が行われている。目下、演算能力の伸びが鈍化してきた CPU に代わって画像処理プロセッサを計算に応用する GPGPU 技術が急速に伸びており、利用環境の整備が進むことで今後の大規模並列計算を支える基盤技術になると予想される。

Keywords:

long fat network (LFN), packet pacing, SINET, GEANT, 100 Gbps, L2VPN, cloud computing, SaaS, DaaS, Virtualization Technology (VT), GPGPU

3.1 はじめに

遠隔実験を快適に行うために何よりもまず必要なのが、遠隔地間の高速な通信を可能にするデータ伝送路である。20年前においてはアナログ・モデム装置を用いてコンピュータ同士をシリアル通信で接続する方式が主流であり、Ethernet に代表されるローカル・ネットワーク (LAN) の通信速度よりもはるかに低速な毎秒 1 k～数 10 k ビット (2400 baud などと呼ばれた) しか利用できなかった。これらはアナログ電話回線網を用いていた。その後、ISDN と呼ばれるデジタル電話網が普及、1 回線あたり 1.5 Mbps の高品質な回線を利用可能になり、核融合研究の分野でも FECnet のような共同研究ネットワークの構築が行われた[1]。

これに対し、現在の学術情報ネットワーク SINET 4 の東京～大阪間の幹線帯域は 40 Gbps × 2 で、これに大学サイト等が 10 Gbps で接続、末端 PC の Ethernet ポートに至っては 1 Gbps と、広域ネットワーク (WAN) の幹線帯域の方が LAN や末端のコンピュータの通信速度を上回っている。日米間の海外回線、欧米間の大西洋回線についても国内幹線と同様である。

このように通信速度の観点からみれば、LAN - WAN の境目は地球規模でなくなりつつあるといえる。ITER 遠隔実験センターが技術的に成立する可能性がある

大きな理由の一つとなっている。フランスにある ITER 装置と日本の遠隔実験センターとの間でも、途中の通信回線品質や物理的帯域幅による制限はほとんどないといえる。残された問題は、いかに遠距離回線の伝送能力やコンピュータの計算リソースをフルに活用できるかの、利用技術面に集約されている。その端的な一例が、遅延の大きい遠距離間でインターネット標準の TCP/IP プロトコルを利用した際、実効データ転送速度が大幅低下する問題である。

遠隔実験を円滑かつ快適に進めるには、距離の違いを感じさせない高速なデータ伝送と臨場感のあるデータ解析環境とが高い次元で融合し、効率的に通信回線やコンピュータの物理性能を活用できる技術が求められる。次節以降にそれら技術開発の現状と課題、ITER 時代に向けた展望を述べる。

(中西秀哉)

3.2 高速データ通信技術

実験や演算結果のデータ転送を始め、Web、遠隔ログイン等に用いられ、インターネットに流れるデータの 90% は Transmission Control Protocol (TCP) プロトコルによる通信である。しかし、TCP による遠距離通信においては確保された帯域を十分に有効利用できない Long Fat Network (LFN) 問題と呼ばれる現象が発生する[2]。

National Institute for Fusion Science, Toki, Gifu 509-5292, Japan

Corresponding author's e-mail: nakanisi@nifs.ac.jp

TCPでは確実にデータを誤りなく送信するために、送信側は先に送信したデータが問題なく受信側にて受領したという確認応答を待って、次のデータを送信する(図1(a))。このままでは転送効率が低いために、TCPの接続時に受信ウィンドウサイズを規定する。送信側は受信ウィンドウサイズ内のデータ量であれば確認応答を待たずに先行して送信を行う(図1(b))。通常の設定値ではキャンパスLANなど近距離通信では問題はないが、遠距離通信になると、パケットが往復する時間(Round-trip time; RTT)が大きくなり、転送効率が低下する(図1(c))。一般的には、

$$\text{受信ウィンドウサイズ} \geq \text{RTT} \times \text{ネットワーク回線帯域}$$

と設定する必要がある[3]。

他のLFN問題の原因として、経路の途中でデータが失われるパケットロスがある。TCPはパケットロスが発生すると回線が他の通信により混雑していると判断し、これ以上混雑しないよう、自らの送信速度を低下させる機能を有する。その後、回線の余白を利用するために徐々に送信速度を上げていくが、RTTが大きいと元の速度に回復するまで著しく時間がかかり効率的な転送が望めなくなる。パケットロスは混雑の他、ネットワーク機器が生ずる揺らぎによるものもあり、これらの対策として、予め送信速度の最大値を抑制する手法がある。

一例として、日本(核融合研)とフランス(ITER IO)間におけるデータ転送速度のグラフを図2に示す。利用可能な回線帯域は両端まで1 Gbpsであったにもかかわらず、標準の設定では図2(a)に示すよう3 Mbps程度の転送速度しか得られない。図2(b)は、受信ウィンドウサイズを上記の条件に合わせて設定した場合である。瞬間的に800 Mbpsを得るが、安定しない。図2(c)は、さらに、バーストによるパケットロスを避けるためにネットワークカードのドライバ内部でパケットの送信間隔を調整した結果である[4]。非常に安定した転送速度が得られている。

続いて10 GbE回線を用いた日本(核融合研)とフランス(ITER-IO)間のデータ転送試験結果を図3に示す。この転送試験では、共用回線(SINET, GEANT, Renater)とL

2 VPNによる専用回線(APAN/JGN, SURFnet, Renater)の両方を用いた。両者の途中経路に4 Gbpsの帯域制限が課せられたため、最大転送速度を3.5 Gbpsと指定した。何れも遜色のない安定した転送結果を得られた[5, 6]。

LFN問題に対して、TCPストリームを複数(多数)同時に並列使用する方法もある。欧州CERNの素粒子実験LHC等にて利用されているグリッド環境GlobusのGridFTP[7]が著名であり、国立情報学研究所(NII)においては次世代の100 Gbps帯域フル活用をめざしてMMCFTPの開発も進められている[8]。また、LFN問題の原因となるTCPを使わずUser Datagram Protocol(UDP)プロトコルにてデータ送信を行う方法もある[9, 10]。UDPにはパケットロスと補完するなど信頼性を維持する機能がないため、最低限必

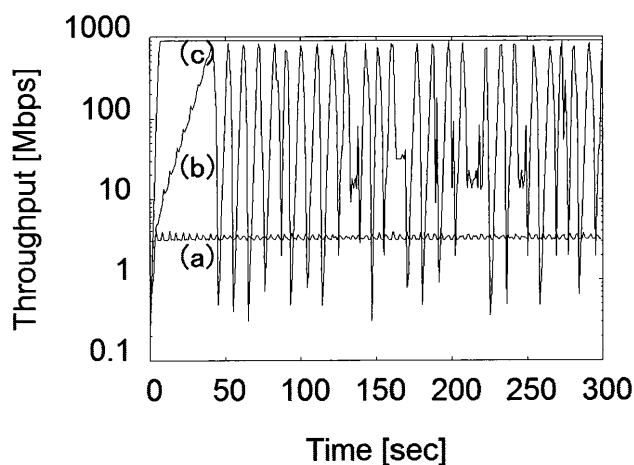
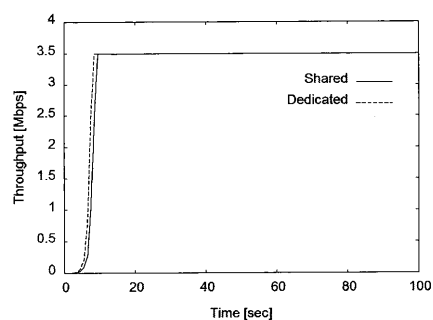


図2 日仏間のメモリ間転送速度 (a)デフォルトの状態。(b)受信ウィンドウサイズを拡大。(c)パケット間ギャップを調整し、最大送信速度を900 Mbpsに制限。



(a) 経路図



(b) 転送速度

図3 10 GbE回線における経路とデータ転送試験結果。途中経路に4 Gbpsの制約があるため、最大3.5 Gbpsの指定を行った。共用回線(SINET-GEANT-Renater)と専用回線(APAN/JGN-SURFnet-Renater)の何れも安定した転送を行うことができた。

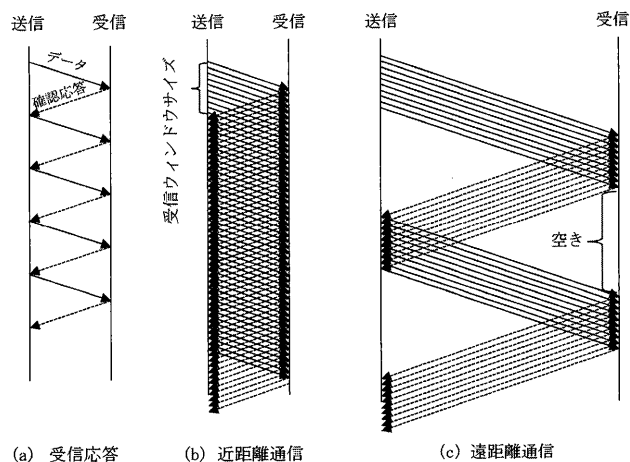


図1 TCPのデータ転送方法。実線がデータを、破線が確認応答を示す。受信ウィンドウサイズが適切でないと、遠距離回線ではデータが流れない隙間が生じる。

要な接続制御用に TCP 接続（必要となる帯域は小さい）を併用する方法が、リアルタイムビデオ配信などでよく使われている。

青森県六ヶ所村にある国際核融合エネルギー研究センター（IFERC）は、国立情報学研究所が運用する SINET 4 の最寄りのデータセンター（弘前）に2012年に接続され、ITER との接続は SINET の国際回線を利用することになると思われる。現在、SINET の国際回線帯域は米国向けで 10 Gbps×2 であるが、2016年に計画されている次期 SINET では、この10倍の 100 Gbps への拡張が期待される。100 Gbps の IEEE 規格は2010年7月に制定され、すでに米国の研究ネットワーク Internet2 や欧州の広域ネットワークである GEANT の幹線にて利用されている。100 Gbps を超える伝送規格については、400 Gbps と 1 Tbps のいずれかと予想されていたが、2013年3月に実現可能性から 400 Gbps の制定を行うことが決まった[11]。400 Gbps の規格制定は2017年頃を目標としているため、ITER 運転開始時（2020年）において利用可能となるか微妙なタイミングであり、仕様確定時期には間に合わない可能性も高く、現状で想定可能な ITER と遠隔実験センター間の回線帯域は 100 Gbps と考えられる。

また、遠隔実験の臨場感を大きく左右する遅延時間の減少を図るためには、物理的な距離を短くすることが有効である。現在、日本からフランスなど欧州への経路は北米経由（約 26000 km）であるが、六ヶ所村を中心とした正距方位図で示すよう最短経路はロシア経由となる（図 4）。2013 年時点ではまだ存在しないが、ロシア北極海経由の回線（約 16000 km）が2014年に完成する計画もある[12, 13]。しかし、回線品質や国際政治情勢、冬に海底ケーブルが切断された場合の保守方法など、不確定な要素が多いため、利用する場合でも、信頼性の高い北米経由を併用した冗長構成が望しい。ロシア経由以外にも、すでにインド洋経由の日欧回線（約 22000 km）が存在するが、現状、運用が安定しない等の問題が残っている。なお、衛星通信に関しては、通信用衛星は赤道上空約 36000 km に位置し伝送路長はその 2 倍以上となるほか、回線帯域が 1 Gbps 程度と狭く伝送

路エラーも大きい等、大容量データ伝送には不適である。例えば JAXA が運用する超高速インターネット衛星きずな（WINDS）においても通常提供する最大帯域は 1.2 Gbps である[14]。

データ転送を行う際には、セキュリティにも十分な注意を払う必要がある。データ保護のために暗号化する手段もあるが、一般的に十分安全といわれる水準の暗号化はかなり重い演算処理が必要なため、大容量データの転送前後でこれを行おうとすると転送速度の大幅低下を招き実用的ではない。そのため近年、ITER のようなプロジェクトベースの大規模データ転送には、当該ネットワークを論理的に他の通信と分離してしまう仮想閉域網（VPN）技術を用いる流れが大きくなっている。この広域 VPN サービスには、仮想化するネットワーク階層（L1：物理層，L2：データリンク層，L3：ネットワーク層）に応じて 3 種類ある。

L1VPN は 1 対 1 接続の専用線環境を提供する。L1VPN はパケットロスの原因となる揺れがきわめて小さくデータ転送にとっては理想的な VPN であるが、複数拠点の場合はメッシュ状に構成する必要があるが、また、SINET～GEANT など複数国際回線をまたがって実現するかは定かではない。L2VPN は複数拠点があたかも単一のハブ（L2 スイッチ）に接続されているかのような環境を提供する。IP ネットワーク的には同一サブネットであるため、すべての端末の MAC アドレスが識別可能であり、全拠点へブロードキャストが行き渡る。また、IP 以外の独自プロトコルも利用可能である。L3VPN は複数のルータを経由して拠点毎のサブネット群が相互接続される環境である。複数のサブネットが作成されその間の経路情報の交換が行われるため、ブロードキャストは各サブネット内に留まる。いずれが遠隔データ転送にふさわしいか今後の検討課題となるが、簡素に構成できる点と実績面から L2VPN が有力と思われる。（山本孝志）

3.3 遠隔実験データ解析

ITER で収集される膨大なデータを転送するためには高速のネットワークが必要であるが、既存のネットワーク技術では長距離のデータ通信には問題があり、そのための技術導入が必要であることは前節で述べた。しかしながら、大小さまざまな研究サイトすべてに、膨大なデータを一律に転送し各々の環境で解析させることは非効率でもあるため、データは拠点となる研究所に置き、遠隔から解析を行うという形態も必要とされる。

かつては、データ解析と言えば、FORTRAN で開発された解析コードをコンパイル・実行して行うというスタイルが一般的であったが、PC の性能の向上とともに、IDL や MATLAB といった汎用解析用アプリケーションを用いて、対話的にデータ解析を行うというスタイルが普及していった。しかしながら、現在では、スマートフォンやタブレット型 PC の普及が目覚しく、今までデスクトップ PC で行っていた電子メールや WEB ブラウジング等の簡単な仕事の多くはこれらの軽量端末上で行うようになり、計算能力や膨大なストレージを必要とする作業は、ネットワー

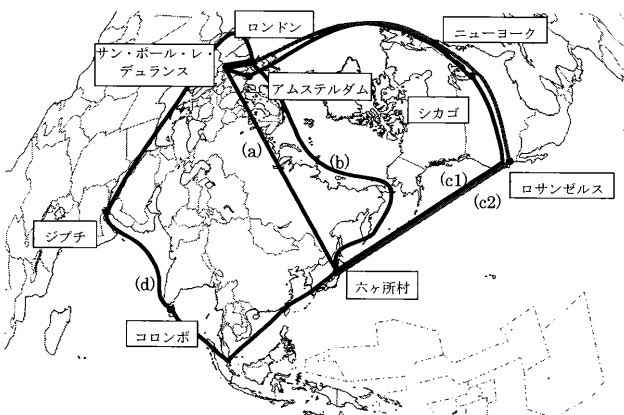


図 4 六ヶ所村を中心とした正距方位図法。(a)最短経路。(b)ロシア北極海経由。(c1)北米シカゴ経由。(c2)北米ニューヨーク経由。(d)インド洋経由の概略を示す。(c1),(c2)はそれぞれ図2の専用回線と共用回線に該当する。

ク上にあるリソースを必要に応じて使うという、いわゆる、クラウド・コンピューティングが盛んに行われるようになってきた。この潮流は遠隔実験においても例外ではなく、軽量端末での利用は増加するはずである。このため、これからの遠隔実験では、ネットワーク上の多様なコンピュータリソースを利用するというスタイルが中心になっていくと思われる。

クラウド・コンピューティングの広がりとともに、Everything as a Service という考えが浸透してきた。これはコンピュータ処理に必要なものをコンポーネントとして提供し必要に応じて利用するというので、以下のような物がある(図5)。

- SaaS (Software as a Service)
- DaaS (Desktop as a Service)
- PaaS (Platform as a Service)
- STaaS (Storage as a Service)
- ……等

この内、SaaSは最も普及しているものの一つで、必要なソフトウェアのみをサービスとして提供するものである。SaaSを利用することにより、クライアント毎にソフトウェアをインストールすることなく、サーバ上で管理することができるため、メンテナンス性が向上し、デマンドの急激な変化に対応しやすい。このような形態のサービスは早くから遠隔実験参加で利用されており、各研究所ではデータ表示や解析を行うプログラム等をCGI等から呼び出すことにより、WEB上で実行できるようなサービスを提供してきた。例えば、核融合科学研究所では、LHD実験の進捗状況を遠隔地からリアルタイムで把握するためにポータルページを提供しており、このページから最新ショットのサマリーグラフの表示や、WEB上から生の収集データに対して簡単な演算を行って可視化できるツールの利用が可能と

なっている[15]。

一般にこのようなサービスはブラウザ上から実行するようにつくられているため、融通が利かないことが多いが、これらのサービスをSOAP[16]やRPC-XML[17]を利用しコンポーネント化することにより、自身で開発したプログラムの中から容易に呼び出すことができ、より柔軟な利用が可能となる。

遠隔実験参加では、DaaSもまた重要な役割を果たす。これはクライアントにデスクトップ環境を提供するもので、あらかじめ解析を行うためのプログラムをセットアップしておくことにより、新規に参加する研究者でもすぐに解析環境を利用でき、また、実行環境を固定することにより保守性が高まる等の効果がある。DaaSの形態としては、旧来からあるWindowsのターミナルサービスのようなセッションベースの他に、最近では、サーバ上で仮想マシンを動かし、仮想マシンのデスクトップを提供するという方法がある(デスクトップ仮想化)。この方法の利点は、ユーザが完全に仮想マシンを操作することができるため、管理者権限の必要な作業を行うことができる他、必要に応じてCPUやメモリ等の資源の割り当てを動的に変更することができる点にある。

近年では1コアあたりの演算能力の向上速度は鈍化しており、各メーカーは1CPUに多くのコアを詰め込むことにより、総合的な計算能力や電力辺りの計算効率を競うようになってきた。この結果、マルチコアのCPUを載せた少数のサーバ上で多数の仮想マシンを同時に動かすことが可能となった。一台の仮想PCでCPUの能力を活かすためには、複数のコアを同時に動作させるようにプログラムを開発する必要がある。このためには、マルチタスクやマルチスレッドを利用したプログラムが必要であり、科学技術計算の分野ではOpenMPやMPIといった並列コンピューティング環境を利用した開発が行われている。

一方、CPUとは別にGPUを利用した計算(GPGPU)が注目を集めるようになってきた。GPUは本来画像処理を行うための装置であるが、3Dの描画性能の急速な発展とともに、GPUの性能が飛躍的に高まった。画像処理で必要とされる計算は簡単な演算の繰り返しが多いため、ハードウェア化しやすく、高速化がCPUに比べて容易である。単純な比較ではハイエンドデスクトップPCで使われるCPUであるCore i7 3770Kが100 GFLOPS程度なのに対し、1枚で数TFLOPSの演算能力を持つグラフィックカードが登場している(表1参照)。このGPUの能力を汎用の計算に利用しようというのが、GPGPU[18]である。

GPUの処理は汎用CPUとは大きく異なる。例えば、GPUには数百~数千ものの演算プロセッサが搭載されている

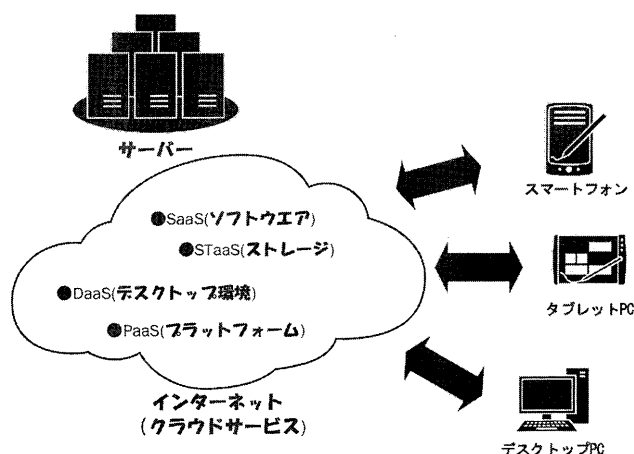


図5 クラウド・コンピューティング概念図。

表1 代表的なCPUおよびGPU。

名称	タイプ	プロセッサ数	演算速度(GFLOPS)
Intel Core i7 3770K	汎用CPU	4	121.6
NVIDIA GeForce GTX 690	GPU	3072 (CUDA コア)	5622
AMD Radeon HD 6990	GPU	3072 (ストリームプロセッサ)	5099
Intel Xeon Phi 5110P	コプロセッサ	60	1011

が、個々のプロセッサが独立して動作しているわけではなく、高速に実行するためには、複数のプロセッサが異なるデータに対し同じ命令を同時に実行する必要がある、条件分岐等が起こると著しく演算速度が低下してしまう。また、メインメモリからGPUの内部メモリへのデータ転送がボトルネックとなるため、極力内部メモリ内で計算を行う必要がある。このようなGPUの特性に適合したものとして、分子運動や天文シミュレーションがあり、CPUに比べ高い演算能力が得られている。

GPUの開発環境として、NVIDIA社は自社のGPU向けにCUDA[19]を提供しており、C言語に類似した言語やFORTRAN 2003の拡張言語を用いてGPUを利用したプログラムを開発することができる。一方、AMD社はATI Stream[20]を提供しているが、両者には互換性がない。そこで、最近ではOpenCL[21]という標準規格が登場し、この規格に沿った開発環境を利用することにより、異なるCPUやGPUの混在する環境でも透過的に並列演算を行うプログラムが開発できる。また、コンパイラやMATLAB等のデータ処理アプリケーションのGPUの対応が始まっており、今後はGPUの存在を意識せずに、その演算能力を活用した解析プログラムが開発できるようになると思われる。

また、NVIDIA社はNVIDIA VGXというGPU仮想化技術を提供しており、この技術を使えば、CPUの仮想化と同様にGPUの仮想化が可能となり、非力な軽量端末でクラウドを利用しながら、必要に応じて仮想CPUや仮想GPUを割り当て、データ解析に利用することが可能になる。

これからのデータ解析ではデスクトップPCだけでなく、スマートフォンやタブレット等の多様な端末が混在することになり、ネットワークを介したコンピュータの利用がより重要になってくる。このような環境下では、様々な仮想化技術が重要な役割を果たし、CPUやGPU等の計算資源やディスク等のストレージが仮想化され、必要に応じてそれらが割り当てられる。マルチコアのCPUやGPUを活用するためには、専用のプログラミングが必要であり、利用のための敷居が高いが、コンパイラやアプリケーションの対応により、容易に活用できるようになってくるはずである。(江本雅彦)

3.4 まとめと今後に向けた展望

ITERなどの次世代実験プロジェクトの一つの大きな特徴は、国際連携で行われる点である。ITERは世界7極による多極間協定で進められており、2020年ごろの稼働予定である。こうした世界規模での連携研究プロジェクトは、すでに素粒子分野のCERN LHCや天文学のALMA電波望遠鏡など、多分野で共通した潮流となっている。

こうした状況で、実際に共同研究を支えるのがInformation and Communication技術(ICT)である。核融合研究分野の遠隔実験プロジェクトでは、最新のICTを駆使して、ローカルでリアルな制御室(実験室)とほとんど同じ臨場感をもった「仮想制御室」を地球の裏側にある共同研究サイトに実現することが究極の目的といえる。

前節でも述べられたとおり、GPUの処理能力はCPUの

演算能力を超えてしまったここ数年、三次元グラフィックスの処理能力は格段に進化しており、核融合プラズマの3-D像再構成・リアルタイム表示も視野に入ってくるようになった。ITERなど次世代の遠隔実験センターでそれが実現した暁には、一般社会から大きな注目を浴びるであろうことは間違いない。核融合分野でICT応用に携わる者として、近い将来そうした先駆的な遠隔実験室を実現できることを期待してやまない。

末筆ながら、SNETの運用に大変お世話になっている国立情報学研究所の漆谷重雄先生を初めとするSINET利用推進室の皆様、日仏データ転送試験を指導して下さった東京大学平木敬先生と同研究室の皆様に深く感謝の意を表します。なお本研究は、NIFS共同研究(NIFS12ULHH006, NIFS11UMSS001)、平成18-21年度国立情報学研究所CSI委託事業などの支援を受けています。

参考文献

- [1] 居田克巳, 中西秀哉: プラズマ・核融合学会誌 73, 377 (1997).
- [2] T. Yamamoto, Fusion Eng. Des. 83, 516 (2008).
- [3] 性能改善のためのTCP受信窓サイズとその設定方法, http://www.sinet.ad.jp/document/sinet_doc/rwin_size_v5.pdf
- [4] T. Yoshino *et al.*, Proc. 2008 ACM/IEEE conf. Supercomputing, article No.11 (2008).
- [5] T. Yamamoto *et al.*, Proc. 12th Int. Conf. Accelerator and Large Experimental Physics Control Systems (ICALEPCS 2009), 12-16 October 2009, Kobe, Japan, WED003, 618 (2009).
- [6] H. Nakanishi *et al.*, Nucl. Fusion 51, 113014 (2011).
- [7] GridFTP, <http://www.globus.org/toolkit/docs/latest-stable/gridftp/>
- [8] K. Yamanaka *et al.*, Proc. 9th IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research, O8-1, 6-10 May 2013, Hefei, China (2013).
- [9] Tsunami UDP Protocol, <http://tsunami-udp.sourceforge.net/>
- [10] 富士通プレスリリース, <http://pr.fujitsu.com/jp/news/2013/01/29-1.html>
- [11] IEEE 802.3 400GbE Study Group, <http://www.ieee802.org/3/400GSG/>
- [12] WIRED 日本語版, <http://wired.jp/2012/03/27/europe-moving-60-ms-closer-to-japan-with-new-undersea-cables-ars/>
- [13] POLARNET, <http://polarnetproject.ru/>
- [14] きずな(WINDS), <http://www.satnavi.jaxa.jp/project/winds/>
- [15] Y. Nagayama *et al.*, Fusion Eng. Des. 87, 2218 (2012).
- [16] <http://www.w3.org/TR/soap12-part1/>
- [17] <http://www.xmlrpc.com/>
- [18] R. Fernando, GPU Gems: Programing Techniques, Tips and Tricks for Real-Time Graphics (Addison-Wesley, Reading, MA, 2004).
- [19] http://www.nvidia.com/object/cuda_home_new.html
- [20] <http://www.amd.com/jp/products/technologies/stream-technology/Pages/stream-technology.aspx>
- [21] <http://www.khronos.org/opencl/>