# グリッド技術を駆使して日米拠点間での超大規模データ処理に成功

## 1万km離れた日米間で記録的な741Mbpsのデータ転送を実現

産総研グリッド研究センターは、グリッド技術「グリッドデータファーム」による大規模データ解析の実証実験に世界で初めて成功した。日米の7拠点をネットワークで接続して計190台のPCからなるシステムを構築し、シミュレーションにより高エネルギー物理学の大規模実験データを生成、処理した。その過程におけるデータ複製作成において、米国内で2.286Gbpsのデータ転送速度を達成、1万km離れた日米間では世界で初めて741Mbpsのデータ転送速度を達成した。この結果により世界規模の超大規模データセンターの実現や、国際的な共同実験による超大規模データ解析にめどをつけることができた。

### グリッドデータファーム

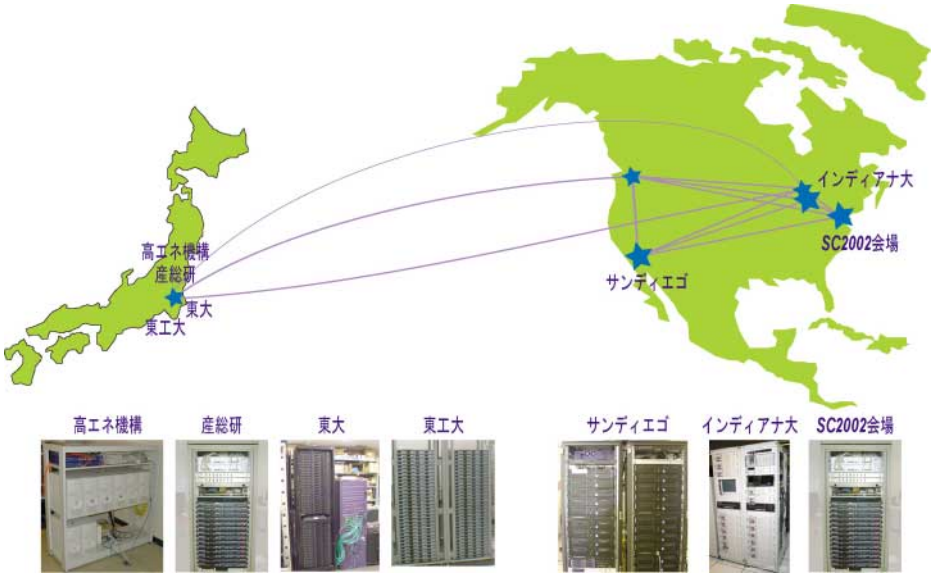
グリッド研究センターでは、超大規模データを複数拠点で協調して解析するグリッド技術「グリッドデータファーム」を研究開発している。これは、広域に分散設置されたPCのハードディスクを利用して大規模並列ファイルシステムを構築するもので、PB（ペタバイト：1PBは1,000兆文字、CD170万枚相当）級の大規模データ計算の基盤システムを目指している。グリッドデータファームの主な特徴は、ファイルアクセスの局所性を利用した大規模データの高速アクセスと、ファイルの複製によるハードディスクやネットワークの耐故障性の実現である。

グリッドデータファームは、年間数PBの実験データの解析が必要な素粒子物理学や、天文学における全天多波長の観測データの解析、生命情報学の遺伝子解析など大規模データ解析、大規模データシミュレーションを必要とする理論・実験科学だけではなく、電子政府・電子商取引などビジネス分野における大量のデータ処理や、地理的に離れた拠点間の高速データ複製による冗長性の確保と負荷分散を行うことができる。世界規模のデータベースなど大規模データの高速処理を、多くの人々が安全に共有するための基礎技術として非常に有効であり、幅広い産業応用が考えられる。

### 日米拠点による実証実験

2002年11月16日から22日まで米国ボルチモアで開催された国際会議SC2002において、日米の7拠点を高速ネットワークで接続し、本グリッドデータファームによる大規模データ解析の実証実験に世界で初めて成功した。

日米の6研究機関（産総研、高エネルギー加速器研究機




●図1　各拠点のネットワーク地図とPCクラスタ

構、東京工業大学、東京大学、米国インディアナ大学、米国サンディエゴ·スーパーコンピュータ·センター[SDSC])とSC2002会場の7拠点に分散配置された計190台のパソコンからなるPCクラスタ7システムをグリッドデータファームにより統合した(図1)。ネットワークには、つくばWAN、APAN/TransPAC、MAFFINのサポートを得た。

構成したシステムは、産総研先端情報計算センターに設置されているスーパーコンピュータSR8000の2倍近い962 GFlopsのピーク計算性能をもち、6600 MB/s(CD1枚を0.1秒で読み書きする速度)の高速アクセス性能をもつ18TB(テラバイト)の大容量ファイルシステムをもつことになる。

本実証実験では、主に東京工業大学の大規模PCクラスタで素粒子実験を模擬する大規模データを生成し、他拠点のPCクラスタに数百GB(ギガバイト)規模の複製を作成した。

## 記録的なファイル転送性能達成

今回の実験では、日本国内はつくばWANとSuperSINET、日米間はAPAN/TransPACとNII-ESnet HEP PVC、米国内はAbileneとESnetといった複数の高速広域ネットワークを利用し、SC2002会場内ではSCinetを利用した。インディアナ大、SDSCとSC2002会場間のネットワーク性能はそれぞれ622 Mbps、日米間は2本のネットワークを利用して893 Mbpsであり、SC2002会場とその他の6拠点を結ぶネットワークの理論性能は片方向あたり2.137 Gbpsとなる(図2)。

実験の結果、外向き1.691 Gbps、内向き0.595 Gbpsの計2.286 Gbps(毎秒約2億9千万文字、CD1枚分のデータを2.3秒で転送できる速度)のデータ転送速度を達成した。この際には、SC2002会場では12台のPCを用いた。また、産総研つくばセンターに設置された4台のPCとSC2002会場内の4台のPCを用いて1万km離れた日米間で、ネットワーク理論性能893 Mbpsの83%に相当する741 Mbps(毎秒約9千万文字、CD1枚分のデータを7秒で転送できる速度)の実効データ転送速度を達成した(写真1)。

特に今回の実験ではAPAN/TransPACの2本の日米ネットワークを効率的に利用することにより高い性能を達

●写真1　大規模データ転送中の模様

成している。日米間で一つのアプリケーションによって741 Mbpsの転送速度を達成したことはこれまでに例がない。

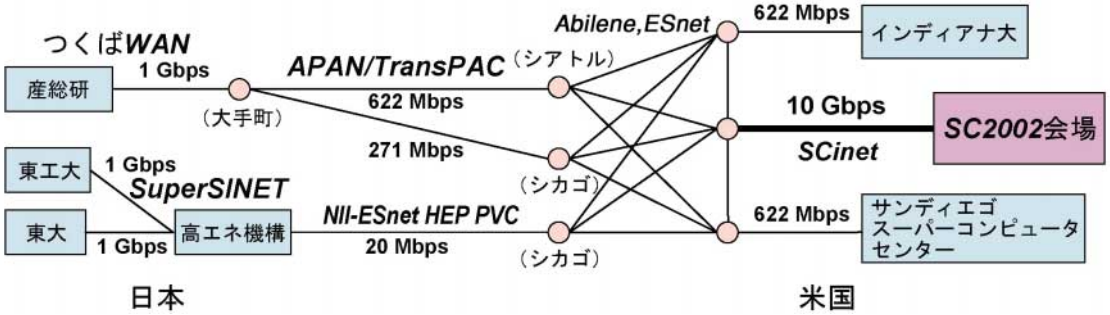
## 長距離高速ファイル転送実現のポイント

グリッドデータファームでは、ネットワーク転送性能、ディスク入出力性能の両方を向上させ、さらに同時に複数のPC間で並列にデータ転送することにより、長距離でのデータ転送速度を大幅に改善することを可能とした。

長距離通信に用いられる光ファイバー上を信号が伝達する場合、1 mあたり5 ns程度の遅延がある。実験を行った日米拠点間は直線距離で約10,800 km離れている。実際のファイバーの長さはこれより長く、途中の通信機器での遅延も加わるため、今回の実験環境では往復で0.2秒程度の通信遅延が観測された。

通常インターネットで利用されているTCPと呼ばれる通信方式では、0.2秒の遅延があると、途中のネットワーク性能に拘らずデータ転送速度は2 Mbps程度に落ちてしまう。これは、もともとTCPが長距離の高速データ転送を考慮していないためである。本実験では現在仕様策定が進められているHigh Speed TCPを利用して性能改善を図った。さらに、High Speed TCPでも解決できないネットワークの状況に合わせた性能改善のため、通信ストリームの流量制限、並列ストリームの数など細かい通信方式の設定を行った。

また、ディスク入出力性能を向上させるために、それぞ




●図2　本実験で構築したグリッド環境のネットワーク論理図

れのPCではハードディスクを同時に4台利用し、ネットワーク性能（1 Gbps）とほぼ同程度のディスクアクセス性能を実現した。これらはすべて1UのPC（高さ約4.5 cmのラックマウント型サーバ）に収まる高密度な実装となっており、省スペースで高性能を得ることができる（写真２）。

## 長距離高速ファイル転送の本当の難しさ

広域ネットワークは専用ネットワークではなく、多くの人々により利用される。そのため、通信方式を変更し本来は性能が向上するはずであったとしても、ネットワークが混雑していると性能向上を実測することができず、その結果だけからでは通信方式変更の効果が分からない。さらに、長距離転送では通信性能が安定するまで時間がかかることもあり、高性能を達成するための通信方式の決定には時間が必要となる。今回の実証実験では、SC2002会場に持ち込んだPCクラスタと、会期中だけ設置されるSCinetの利用が主であったため、通信方式の決定に費やすことができる時間は実質二日しかなく非常に大変な作業となった。

また、理論性能に近い性能を達成するには、利用するネットワークの状態が良好に保たれていなくてはならない。これには実験に用いるネットワークの運用者との間で十分な調整、調査が必要となる。さらに、今回の実験では日本、米国、日米間と複数の広域ネットワークを利用したため、ネットワーク不調の際の切り分けも複雑となる。実際、ネットワークの不調と思われる症状があり、その切り分けおよび原因の解明に苦労した。結局、米国内ネットワークの不調と判明し、修正前は35 Mbpsしか測定できなかったが、修正後は500 Mbpsを超える性能を出すことができた。このような問題は「ネットワークのバンド幅を使いきろう」というような機会でもない限り発見されることがないため、図らずも米国内ネットワークのバグ取りにも貢献することとなった。


●写真２　グリッドデータファームPCクラスタ

●写真３　日米拠点での超大規模データ処理に成功したグリッドデータファームバンド幅チャレンジチームの産総研、高エネ機構メンバー。前列左がグリッド研究センター建部 研究員

## 国際研究協力の成果

本実験にあたり、産総研はグリッドデータファームのソフトウェア開発および実証実験環境の調整を、高エネルギー加速器研究機構は素粒子実験シミュレーションのプログラム開発を行った。また、東京工業大学は大規模PCクラスタにより計算資源を必要とする実験シミュレーションデータを生成した。インディアナ大学およびSDSCはPRAGMA（環太平洋におけるグリッドアプリケーションに関するコラボレーション）による共同研究として議論に参加すると共に、計算資源、ネットワーク資源、ディスク資源の提供および環境構築、性能評価に対する協力を行った（写真３）。

## 今後の予定

グリッドデータファームは、TBあるいはPB規模の超大規模データに対する高速処理を、多くの人々が安全に共有することを目指したグリッド技術である。より高速なネットワークを用い、装置の規模を拡大することにより、今後必要とされる今回の実験の10倍の高速データ転送、100倍の大規模データに対する高速処理にも対応できる。今後は、欧州も含めた世界規模のグリッド環境により、さらに大規模な実証実験を進めていく予定である。

※通常、1 Mbps＝1,000,000 bpsとして計算されるが、プレス発表時はファイル転送に限り1 Mbps＝1,024×1,024 bpsとして計算していたため、転送速度は707 Mbpsとした。本記事では通常の換算に従い741 Mbpsとしている。

●問い合わせ
〒305-8568
茨城県つくば市梅園 1-1-1　中央第２
独立行政法人 産業技術総合研究所 グリッド研究センター
大規模データ応用チーム　建部 修見
E-mail o-tatebe@aist.go.jp