MK-4

Data Reservoir

科学技術研究向け超高速ネットワーク基盤

Data Reservoir: new network infrastructure for scientific research projects

平木　敬† 稲葉真理† 玉造潤史§ 酒井英行§ 陣崎　明‡
Kei Hiraki Mary Inaba Junji Tamatsukuri Hideyuki Sakai Akira Jinzaki

1．はじめに

科学技術研究とコンピュータ・ネットワークなどの情報システムとは、コンピュータの誕生時から密接な関係を持ってきた。実際、世界最初のプログラム内蔵コンピュータである EDSAC は、おもに化学・気象・電波天文学における数値計算に用いられ、初期のスーパーコンピュータ達によるシミュレーションで科学は飛躍的な進歩を遂げた。インターネットによる通信により、科学情報やデータの交換が実現し、Web により文献情報の電子化が確立した。

利用可能である最先端の情報システム利用には、情報システム性能の３大指標である、計算速度、記憶容量（メインメモリ容量とディスク容量を分離して考えると、４大指標となる）、ネットワークバンド幅のバランスの取れた有効利用が不可欠である。現在、このバランスは、DWDM や MEMS 技術に代表される光通信技術の急速な発展により新たな転換期を迎えつつある。1997 年に実用期に入ったギガビット・イーサネットから、５年間で 10 ギガビットイーサネット（10GbE）が実用期に入ろうとし、100GbE は既に視野に入りつつある。しかしながら、現存のアプローチでは超高速ネットワークの能力を有効に引き出すことが困難である。このことは、プロセッサチップにおける性能向上よりずっと早い率でネットワークの高速化が達成され、プロセッサの並列化と新しいソフトウェア基盤なしには超高速ネットワークの持つ性能を有効利用することが不可能であることを示している。

「科学技術研究のための新しい超高速ネットワーク利用基盤」プロジェクトは、上記状況を解決するため、10 Gbps から 100 Gbps を超える超高速ネットワークを実験・観測機器からの超大型データを解析する Data Intensive Research の鍵として科学技術研究者が有効に利用し、科学技術研究の対象領域を拡大し、科学技術研究を加速することを目的し、平成１３年度から平成１５年度にわたり科学技術振興調整費、先導的研究等の推進プログラムにより実施中である。

2．プロジェクトの概要

我々が提案する超高速ネットワークの利用基盤では、理学系研究科内の 10 個の実験・観測プロジェクト[1]と、高エネルギー加速器や天体望遠鏡等の実験・観測機器間を超高速ネットワークを介してデータ共有し、ネットワークが持つ能力の高い水準での活用を実現することを目的としている。これらのプロジェクトでは既に多量の巨大ファイルを用いた実験・観測を実施しているため、超高速ネットワーク利用基盤の整備が出来次第、ネットワークバンド幅の高効率利用が可能である。

しかしながら、10 Gbps から 100 Gbps 以上の超高速ネットワークを有効に活用することには多くの困難点が存在する。特に大域／超高速ネットワークは高レイテンシ・高バンド幅であり、通信ウィンドウ制御、エラー制御、バッファー管理上多くの困難点が現出する。また、ネットワークからのデータをユーザが実際に利用するために必要な、ネットワーク・インタフェース・カード内パケット処理速度、I/O バスバンド幅、メモリバンド幅、OS およびライブラリのオーバーヘッド、磁気ディスクドライブとの転送バンド幅などの制限により、一台の PC/WS/サーバからの大域ネットワーク通信は高々600Mbps 程度の通信しか実現しない。すなわち、これら問題点をクリアし、スケーラブルなネットワーク利用基盤を用いない限り多数の ftp や Web 通信を多数束ねた形でしか超高速ネットワークを活用できず、多量な巨大ファイルを用いる実験・観測プロジェクトによる活用は非常に困難であると言わざるを得ない。

我々のプロジェクトは、多量の巨大データを扱う科学技術の研究プロジェクトにおいて超高速ネットワークの持つ能力を活用することを目的として、計算システム研究者（東京大学情報理工学系研究科）、ネットワークシステム研究者（富士通研究所）と理学研究者（東京大学理学系研究科）が共同の体制で、① 遠距離通信と近距離通信を分離し、② ユーザからは通常のファイルとしてアクセス可能であり、③ ネットワークバンド幅とディスク容量に対してスケーラブルである利用基盤を構築し、研究基盤として整備する。

3．データレゼボワール

データレゼボワール（Data Reservoir）は、外部超高速ネットワークからのデータを各研究プロジェクトに分散して置かれるユーザの計算システムから利用可能とする、我々が提案する超高速ネットワーク利用基盤の核となるシステムである。遠距離通信と近距離通信を分離し、実際にデータを利用する計算システムとの接続が近接通信であり高レイテンシ・高バンド幅通信であることから発生する問題点をユーザから隠蔽するため、データレゼボワールは大域ネットワークの端点に設置され、ディスクをキャッシュ層として使用する分散共有ファイルシステムとしてユーザからアクセスされる基本アーキテクチャを持つ（図１）。

† 東京大学情報理工学系研究科
§ 東京大学理学系研究科
‡ 富士通研究所

1. SMART 偏極実験、電波望遠鏡観測実験、スローン・デジタル・スカイ・サーベイ、初期宇宙 X 線・γ線観測実験、地球流体変動シミュレーションデータ解析、LHC・ATLAS 実験、ASTRO-F 赤外線衛星観測、GRAPE-6 恒星系シミュレーション、KEK b-factory 実験、スバル望遠鏡観測。

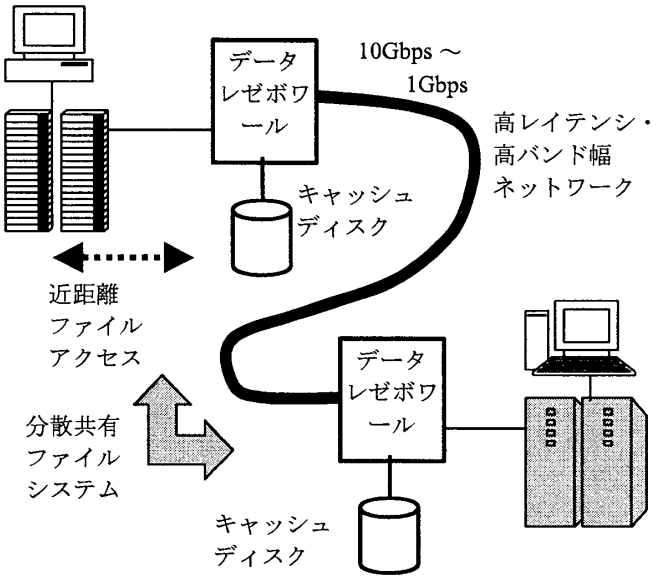

図１．データレゼボワールの概念構成

データレゼボワールでは、ディスクドライブへのアクセスプロトコルとの関連から、SCSI プロトコルを TCP/IP 上に載せた iSCSI プロトコルを採用した。iSCSI プロトコルを用いることにより、高いインターオペラビリティをもつ Storage Area Network を内包することと等価になり、多様なオペレーティングシステムをクライアントとして持つことが可能になるとともに、多段階のストライピングや RAID 機能を透過的に実現することが可能となる。

高レイテンシの超高速ネットワークの持つバンド幅を高効率で利用するため、ストライピングによるデータ転送の並列化に加え、同一 iSCSI デバイスに対して複数ストリームのアクセスを、iSCSI の queue tag 機能を用いて実現する。データレゼボワールは、ユーザからのファイルアクセスを処理し、ストライピングされた iSCSI でのアクセスに変換するファイルサーバ、iSCSI によるアクセスを受理し、更なるストライピングを施し、実際のディスクドライブにアクセスすると共に、遠隔システムとのディスク間データ転送を実現するディスクサーバと、それらを相互接続するネットワーク・スイッチにより構成される。より多数のディスクドライブが必要な場合には、中間ノードにより iSCSI アクセスを更に階層化した iSCSI アクセスに変換する

4. 実証実験

科学技術研究向け超高速ネットワーク基盤であるデータレゼボワールの本格的稼動に先立ち、1 Gbps と 10 Gbps（図２）の２種の実証モデルを構築し、SUPER-SINET を用いた実証実験を実施した。実証実験では、ユーザシステムからは NFS ファイルシステムを介してのデータアクセスを用い、２個のデータレゼボワール間は iSCSI 通信によりデータ同期を取った。実験の結果、1 Gbps モデルでは、SUPER-SINET を用いて 1600Km（遅延時間２６ｍｓ）までの環境において、95%以上の実効バンド幅利用が可能であること（5%は TCP/IP および iSCSI のヘッダ損）、10 Gbps では 10 GEthernet を用いて 7.2 Gbps のデータ転送が実証され、我々の基本アーキテクチャが超高速ネットワークの高効率利用に有効であることが示された。

5. おわりに

「科学技術研究向け超高速ネットワーク基盤整備」プロジェクトでは、1 Gbps および 10 Gbps プロトタイプ実験を通し、遠距離でのデータ共有をネットワークの高効率利用を通して実現することが可能であること、これまで科学技術分野の研究者が築いてきたプログラム基盤を新たなネットワーク基盤上で活用可能なことが実証された。Data Reservoir は新たな汎用的な科学技術研究向き情報基盤として着実に成長している。今後、ファイルシステムおよびユーザインタフェースを一層洗練させるとともに、クラスタ計算機やスーパーコンピュータを含めた系での利用を可能とする予定である。また、遠隔でファイルを共有する基本アーキテクチャから、Disaster Recovery や超高速インターネットにおけるデータ共有というより汎用的な情報基盤への適用も推進する予定である

発表文献

[1] 平木、稲葉、玉造、来栖、陣崎、古賀、酒井、岡村、生田、宮澤：Data Reservoir: 理学研究向け超高速ネットワーク利用基盤、信学会技術報告 CPSY-51、Oct. 2001.

[2] 稲葉、玉造、来栖、陣崎、古賀、生田、平木：Data Reservoir: プロトタイプシステム：アプローチと実験結果、信学会技術報告 CPSY-52 Oct. 2001.

[3] Hiraki, K. Inaba, M., Tamatsukuri, J., Kurusu R., Ikuta, Y., Hisashi, K. and Jinzaki, A., "Data Reservoir: A 4Gbps Long Distance File Sharing Facility for Science Data Processing" Poster, SC2001, Nov. 2001.

[4] Inaba, M., Kurusu, R., Tamatsukuri, J., Koga, H., Jinzaki, A. and Hiraki, K., "Data Reservoir: A very high-speed long distance file sharing facility for scientific data processing," Proc. High-Performace Computing Systems, IPSJ, pp.81-88, Jan 2002.

[5] Hiraki, K. et al, " Data Reservoir: A New Approach to Data-Intensive Scientific Computation," Proc. Int. Symp. on Parallel Architecture, Algorithm and Network, pp.269-274, May 2002.

[6] Hiraki, K. Inaba, M., Tamatsukuri, J., Kurusu R., Ikuta, Y., Hisashi, K. and Jinzaki, A., "Data Reservoir: Utilization of Multi-Gigabit Backbone Network for Data-Intensive Research," to appear Proc. SC2002(CDROM), 2002.

[7]Kurusu, R., Sakamoto, M., Ikuta, Y., Hiraki, K., Inaba, M., Tamatsukuri, J., Koga, H., and Jinzaki, A., "Data Reservoir: Multi-Gigabit Data Transfer Facility, Its Design and Implementation,", Proc. PDCAT, 2002.

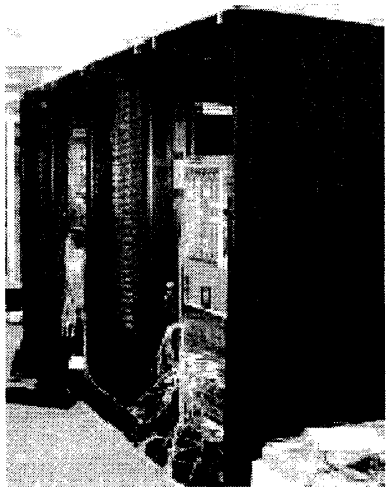

図２．10 Gbps モデル（２セット）