科学技術研究向け超高速大域ネットワーク基盤

平木敬　稲葉真理

情報理工学系研究科コンピュータ科学専攻

概 要

科学者がネットワークの存在を意識することなく，大域超高速ネットワークを使って自然科学の実験・観測データの交換をできるようにするための新しい科学技術研究基盤を構築するため，アドレスベースの大域分散共有ファイル (DSF) アーキテクチャーおよび大域分散データ共有ファシリティー・データレゼボワールシステムを提案，その実現と活用のための技術体系の確立をめざす.

1　はじめに

科学研究とコンピュータ・ネットワークなどの情報システムは，コンピュータの誕生時から密接な関係を持ち続けてきている．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，またこれに続くスーパーコンピュータによるシミュレーションにより 20 世紀後半の科学は 10 の 30 乗オーダーの宇宙論から 10 の -30 乗オーダーの量子論にいたるまで飛躍的な進歩を遂げてきた．また，実験観測装置の微細化およびコンピュータの計算能力およびストレージ容量の増大により巨大データを計算機で取り扱うデータインテンシブリサーチが，脚光をあびつつある．そして近年のネットワーク技術，特にインターネットの発達は，科学情報やデータの交換を実現させ，また Web による文献情報の新しい電子化の形を成立させたのみならず，社会構造を変化するまでの多大な影響を社会に対して及ぼしつつある．日本国内では科学研究基盤ネットワークとして国内の大学・研究機関をつなぐための 10Gbps のバンド幅を持つ SuperSINET が 2002 年より整備されてきている．

最先端の利用可能な情報システムを効率良く活用するためには，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た場合は 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．しかしながら，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 ギがビット・イーサネット (10GbE) が実用期に入ろうとしており，100GbE はすでに視野に入りつつある．これは，プロセッサチップにおける性能向上より，ずっと速い率でネットワークの高速化が達成されることを示しており，現存のアプローチを続けることでは超高速ネットワークの能力を有効に引き出すことは困難であり，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．

我々は，多量の巨大データを扱う実験・観測科学研究プロジェクトが超高速ネットワークの持つ能力を十分に活用してデータを遠隔研究施設間で共用することを目標とし，(1) 遠距離通信と近距離通信を分離し，(2) 近距離通信には通常のファイルアクセス・インターフェイスを提供し (3) 遠距離通信能力はネットワークバンド幅とディスク容量に対しスケーラブルであるネットワーク利用基盤を構築することを目的とし，分散共有ファイル (DSF) アーキテクチャーを提案し分散データファシリティー・データレゼボワール・システムを実装する．

本研究は情報科学の発展への貢献のみならず，実際に運用されることで，実験・観測科学研究の基盤としてデータインテンシブリサーチに利用されることにより理学の発展にも大きく寄与することができる．

2　データレゼボワール・システムの概要

データレゼボワールシステムは大規模実験観測施設と大学等の解析機関をつなぐ超高速かつ高レイテ

ンシな大域ネットワークを前提としている．我々は計算機システムの並列化における分散共有メモリの手法にならいファイルへのアクセスをローカルアクセスとリモートアクセスに分離しディスクをキャッシュ層として使用しアドレスベースでデータにアクセスする分散共有ファイル (Distributed Shared File, DSF) アーキテクチャーを提案した (図 1)．

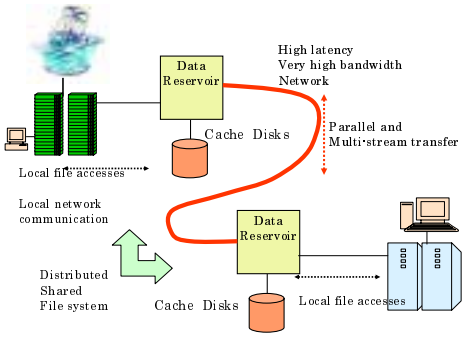

図 1: DSF アーキテクチャー

ネットワーク利用の並列化を図 2 のように各ストレージデバイスの自立的な並列ストリーム転送で実現することにより複数ストリームによる並列バースト転送を実現するという特徴をもつデータレゼボワールシステムを構築した．

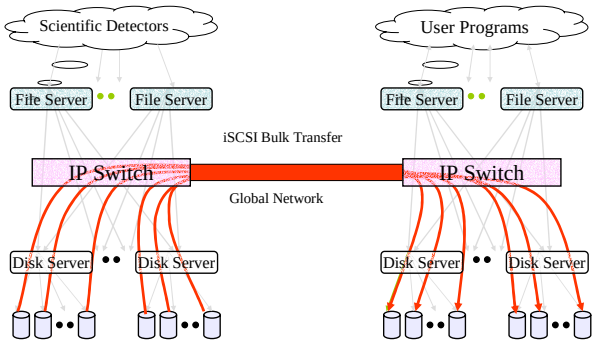

図 2: 自立的並列データ転送

データレゼボワールシステムはファイルサーバとディスクサーバから構成され，階層的なデータのストライピングを行う．すなわちファイルサーバのデータは複数のディスクサーバに，ディスクサーバのデータは複数のローカルストレージにストライプされ分散して格納される (図 3)．データアクセスのための通信には iSCSI(internet SCSI) プロトコルを採用した．

また，データレゼボワールシステムは低位層において並列ストリーム転送の実装が行なわれているため OS，ファイルシステム，そしてユーザプログラムへの透過性を持ち，ユーザはファイルサーバが提供する通常のファイルシステム，あるいは NFS や CIFS

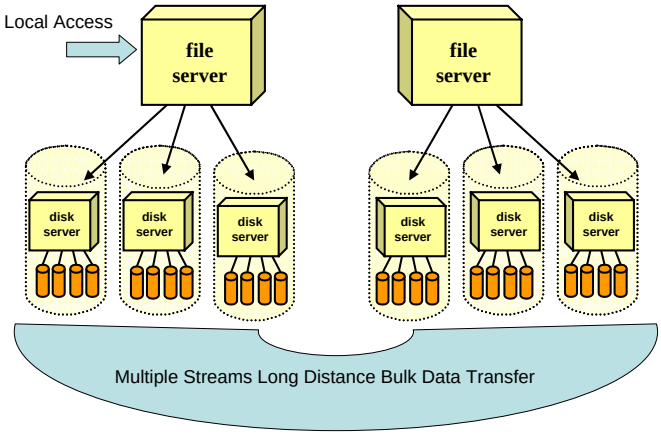

図 3: Data Reservoir システム

といったネットワークファイルシステムを通して本システム上のデータを利用できるという特徴を持つ．

3 実装，実験および性能評価

我々は，DELL Power Edge 1650, (Dual Pentium-III 1.4GHz, 1GB メモリ, 3COM 3C996-SX 1000BASE-SX NIC, 10,000rpm Ultra160 SCSI HDD , Linux 2.4.18 USAGI STABLE 20020408)，Extreme Summit5i 及び 7i, Foundry BigIron8000 ギガビット・イーサネットスイッチでデータレゼボワールシステムの実装を行ない以下の環境でデータ転送実験を行ない，性能評価を行なった．

- 東大・宇宙研 (神奈川県相模原市) 1Gbps 25 マイル 遅延 RTT 3.5msec
- 東大・京大・阪大・東北大・東大 1Gbps 1000 マイル 遅延 RTT 36msec
- 東大・FLA(メリーランド大学)[1] 7500 マイル 遅延 RTT 200msec (SC2002 の予備実験)
- 東大・SC2002 会場 (ボルチモア)[2] 7500 マイル 遅延 RTT 200msec
- 実験室内 遅延，パケットロス装置を使う転送実験 (0〜400msec)
- 実験室内 10Gbps 転送実験

[1] APAN, Abilene, MAX, ボトルネックは 東京シアトル間 TransPAC 回線 OC-12/POS, 約 595Mbps
[2] 日米回線は FLA 実験と同じ

以下，性能の記述は特に断わらない限り，転送データ量を所要時間で割った平均転送レートとし，iSCSI や TCP/IP のヘッダ・オーバヘッドは含まない値を記す．また，ファイルサーバが i 台，ディスクサーバが j 台，各ディスクサーバが k 台のデータディスクを有するシステムを $i \times j \times k$ と略記する．

従来のデータ転送方式では，持続的に高速データ転送を行うことが困難で，たとえば図 4 (GbE で接続された 2 台のホスト間を netcat で 1GB のファイルを転送) に示すように，ピーク性能に対し平均転送バンド幅の低下は不可避であった．一方，データレゼボワール (1x4x2 構成) で 1GB のファイルを転送した時の転送レートを図 5 に示すが，GbE のほぼ限界までコンスタントにパケットを流していることが読みとれる．図 6 に示すようにまたシステム構成にほぼスケールする転送速度を達成している．一方，IPv6 上の転送速度を図 7 に示すが，実験時，IPv6 では経路途中のネットワークが安定していなかったせいと思われるが (図 8 が TCP のパケット再送率)，ストリーム間での転送速度の不均衡が顕著であり，ピーク性能に比して平均性能が悪くなる原因となった．

TCP の輻輳制御の影響を調べるため確認応答待ちパケット数[3] の推移を調べた．図 9 に 1×2×1 構成の 2 本の TCP ストリームに関する送信パケットのうち確認応答パケット未着のパケット数の推移を示す．図 9 のピーク値は，その時点での Congestion Window サイズを表わしており，たとえば，disk2, 9sec でわかるように Fast Retransmit の補償が十分に機能していない．

4 まとめ

我々は，一般に使われている TCP/IP になんら手をいれることなしに，システム構成に応じてスケールし高レイテンシ超高速ネットワーク上でもネットワークバンド幅の 90 〜 95% を持続的に利用できるデータレゼボワールシステムを提案・実装した．本システムは日米回線を使用した実験により，7500 マイルの遠距離通信ではピーク性能 550Mbps，ネットワーク帯域の 91%以上の持続的通信を達成した．通常の TCP によるデータ転送では，通信レイテンシの増加に従ってストリーム当りの性能は低下するが，デバイスプロトコルレベルでの通信の並列化及び最適化により複数ストリームの並列データ転送によって

[3] http://irg.cs.ohiou.edu/software/tcptrace/tcptrace.html

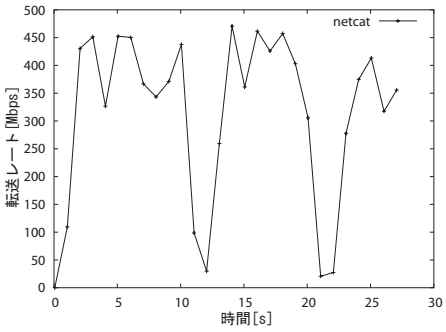

図 4: ファイルシステム，OS を通した従来のデータ転送

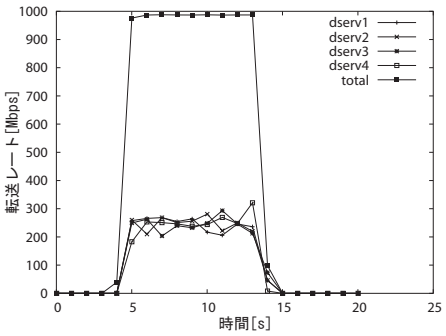

図 5: データレゼボワールによるデータ転送

転送バンド幅の増大と安定化を達成した．

今後の課題としては，TCP の送信 Window サイズの制御がストリームごとに独立して行なわれると高遅延環境の元ではストリーム間での転送レートのばらつきが無視できない事が確認されたため，Window サイズをストリーム間で相関的に制御する方式を開発する事が重要であろう．

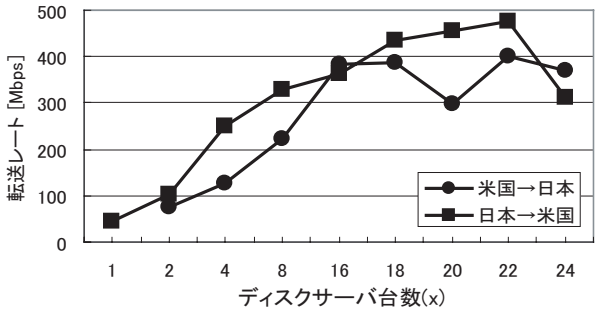

図 6: IPv4，7500 マイル，サーバ数と転送レート

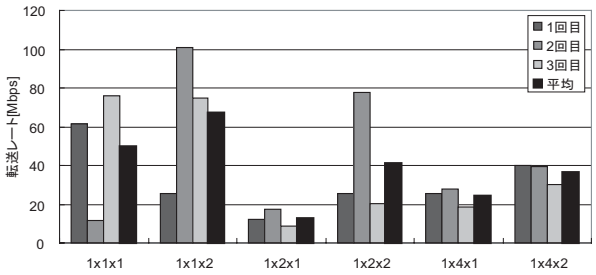

図 7: FLA(7500 マイル) IPv6 実験結果

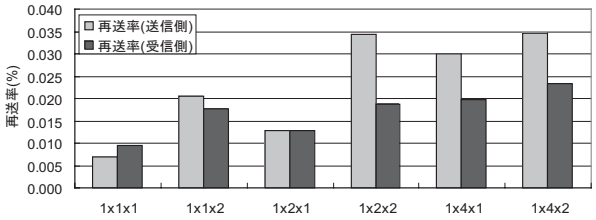

図 8: FLA(7500 マイル) IPv6 再送率

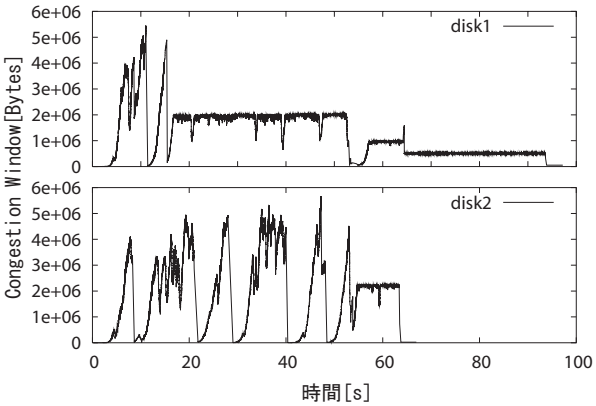

図 9: 7500 マイル，IPv6，1x2x1 構成，outstanding 送信データ量の推移

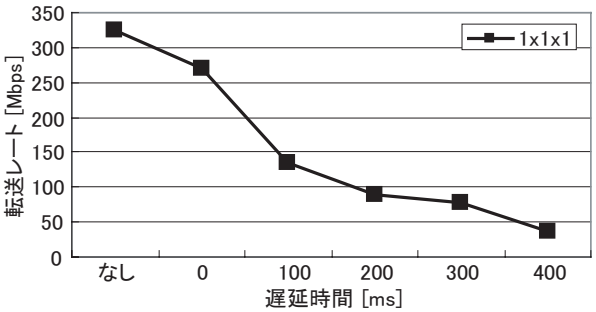

図 10: 実験室，遅延時間による性能低下

図 11: 10G モデル

参考文献

[1] K. Hiraki, M. Inaba, J. Tamatsukuri, R. Kurusu, Y. Ikuta, H. Koga, A. Zinzaki, “Data Reservoir: Utilization of Multi-Gigabit Backbone Network for Data-Intensive Research”, SC2002, Nov. 2002. http://www.sc2002.org/paperpdfs/pap.pap327.pdf

[2] R. Kurusu, M. Sakamoto, Y. Ikuta, K. Hiraki, M. Inaba, J. Tamatsukuri, H. Koga, A. Zinzaki, “Data Reservoir, Multi-Gigabit Data Transfer Facility, Its Design and Implementation”, Proc. PDCAT, pp. 100-108, Sept. 2002.

[3] 平木敬, 稲葉真理, 玉造潤史, 来栖竜太郎, 生田祐吉, 古賀久志, 陣崎明, “超高速ネットワーク用データ共有システム: データレゼボワールの性能評価”, SWoPP, Aug. 2002.

[4] K. Hiraki, M. Inaba, J. Tamatsukuri, R. Kurusu, Y. Ikuta, H. Koga, A. Zinzaki, “Data Reservoir: A New Approach to Data-Intensive Scientific Computation”, Proc. ISPAN, pp. 269-274, May 2002.

[5] 稲葉真理, 来栖竜太郎, 玉造潤史, 古賀久志, 陣崎明, 生田祐吉, 酒井英行, 平木敬, “Data Reservoir: A very high-speed Long distance file sharing facility for Scientific data processing”, Proc. HPCS, IPSJ, pp. 81-88, Jan. 2002.