Data Reservoir — 遠距離超高速ファイル転送システム

来栖竜太郎[1]　坂元眞和[1]　古川裕希[1]　生田祐吉[1]　陣崎明[2]　下見淳一郎[2]
中村誠[3]　玉造潤史[4]　亀澤寛之[5]　稲葉真理[5]
平木敬[5]

1) (株) 富士通コンピュータテクノロジーズ - 長野県長野市三輪田町 1313
2) (株) 富士通研究所 - 神奈川県川崎市上小田中 4-1-1
3) 東京大学情報基盤センター - 東京都文京区弥生 2-11-16
4) 東京大学大学院理学系研究科 - 東京都文京区本郷 7-3-1
5) 東京大学大学院情報理工学系研究科 - 東京都文京区本郷 7-3-1

概 要

データレゼボワールシステムは理学研究の実験・観測施設で生成される巨大データを遠隔施設間で共有することを目的とするネットワーク基盤であり，大域ネットワーク上で信頼性のある高速マルチストリームデータ転送を実現する．本稿では データレゼボワールシステムの概要，および 24,000km 超高速データ転送実験について述べる．

1　はじめに

近年のネットワーク技術の進歩はめざましく，国内の SuperSINET，米国の Abilene に代表される 10〜40Gbps 国内バックボーンネットワークや APAN, SuperSINET, GENKAI といった日米・日韓間の海底光ファイバによる超高速バックボーンネットワークが整備され，科学研究施設はマルチギガビットレベルでの相互接続が可能となってきている．しかしながらネットワーク・インターフェース・カード，I/O バスバンド幅，メモリバンド幅，磁気ディスクドライブ I/O 速度などの制限により，マシン単体でネットワークインフラの能力を十分に活かすことは容易ではない．また 信頼性のある通信として一般に使われている TCP/IP プロトコルは Long Fat Pipe Network と呼ばれる遠距離で通信遅延が大きく広バンド幅ネットワークでは十分な性能を得られないことが知られており，TCP ウィンドウサイズの調整関数を変更することで性能を引き出すための研究が精力的に行なわれている [2, 3].

我々は，理学研究，特に実験・観測プロジェクトが巨大データを遠隔研究施設間で共用するためのネットワーク利用基盤として，データレゼボワールシステムを提案，実装し，性能評価を行ってきた [4, 5].

データレゼボワールシステムは，遠距離通信と近距離通信を分離し，近距離通信は通常のファイルアクセス・インターフェースをもち，遠距離通信はストライプされたデータを並列ストリームで高速に送受信するという特徴を持つ．この遠距離通信用並列ストリームは，ソフトウェアによる通信レートコントロール機構，あるいはハードウェアによる TCP 終端処理により高速化を行ない，ネットワークバンド幅，ディスク容量に対するスケーラビリティーを保持している．

本稿では データレゼボワールシステムの実装，ソフトウェアによる並列ストリーム高速化，および 2003 年 11 月に SC2003(アリゾナ州フェニック) で行なった日米 一往復半 (24,000km) 超高速 (7.01Gbps) データ転送実験について述べる．

2　システム実装

データレゼボワールの基本アーキテクチャーは近距離通信と遠距離通信を分離しアドレスベースでデー

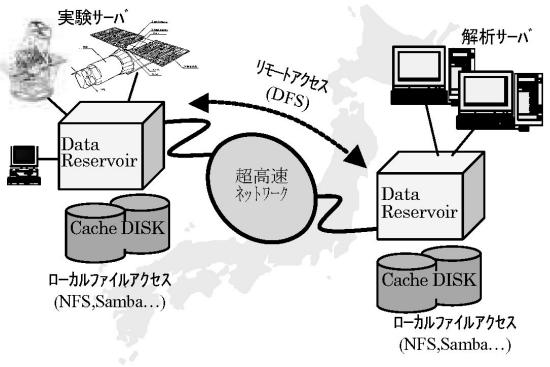

図 1: DSF アーキテクチャ

タにアクセスする Distributed Shared File(DSF) アーキテクチャーであり (図 1)，システムはファイルサーバと複数のディスクサーバから構成される．データアクセスのための通信には TCP/IP 上の iSCSI(internet SCSI) プロトコルを採用している．近距離通信時はディスクサーバが RAW DISK エミュレーションを行うことでファイルサーバがイニシエータ，ディスクサーバがターゲットとなる通信を行なう．一方，遠距離通信時は，転送元・先のディスクサーバがそれぞれイニシエータおよびターゲットとなりブロックレベルで複数ストリームによる並列転送を自立的に行う (図 2)．遠距離通信では，バンド幅を有効に活用するため，2 段階階層的データストライピングを行ないデータの均等分散管理を行なっている．

ファイルサーバおよびディスクサーバのソフトウェアの構成を，それぞれ，図 3，図 4 に示す．近距離

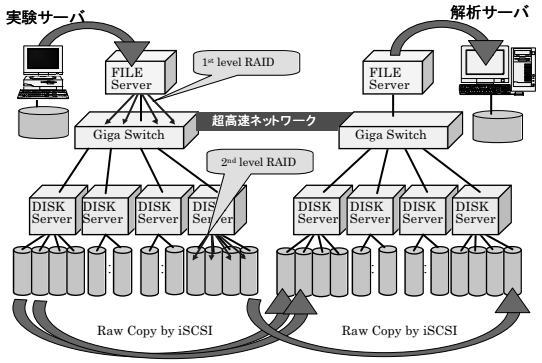

図 2: システム構成図

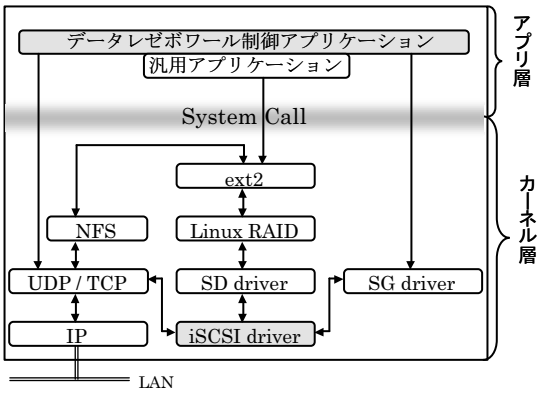

図 3: ファイルサーバレイヤ図

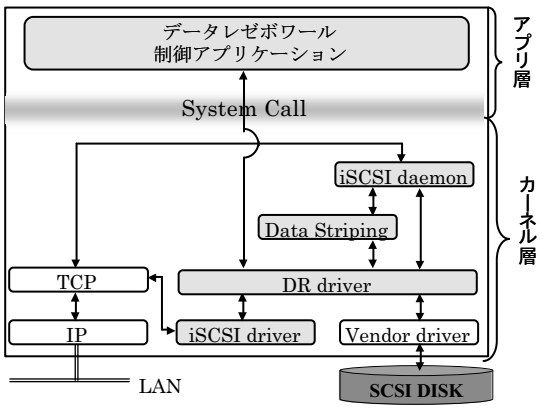

図 4: ディクスサーバレイヤ図

通信元ではファイルサーバの汎用デバイス I/F を通し，遠距離通信元ではディスクサーバの DR デバイス I/F を通しディスク I/O 要求が発行されると，iSCSI ドライバが起動され I/O 要求は iSCSI フレームとしてカプセル化される．この iSCSI フレームは TCP/IP レイヤを通しネットワーク越しに，通信先デバイスであるディスクサーバにパケットとして送られる．当該パケットを受領した通信先ディスクサーバは TCP/IP レイヤを通し iSCSI daemon で iSCSI フレームを受領し，これを SCSI コマンド化し DR ドライバによって自身の物理ディスクに I/O 要求を発行し実際のデータアクセスを行なう．iSCSI ドライバは SCSI の最下層ドライバとして実装されるため，Linux システムの “/dev/sdx” や “/dev/sgx” など標準デバイス I/F を通した iSCSI デバイス利用が可能となっている．遠距離高速転送のため，ディスク

サーバのSCSIディスクへのアクセスとしてはストライプされ分散格納されたデータの高速転送に特化したDRドライバを作成した．また，iSCSI daemonはソフトウェアオーバーヘッドを軽減するため，kernel層で動作するkernel daemonとして実装し高速化を行った．

3 ソフトウェアによる並列ストリームの高速化

TCP/IP は信頼性のある通信プロトコルとして標準的に利用されている．現在一般に使われているNewRenoではネットワークの混雑度は送信パケットに対するACKの欠如およびタイムアウトから推定されるパケット損失によって計られる．この混雑度，すなわちパケット損失情報に基づきTCP ウィンドウサイズの調整による流量制御を行なっている．流量すなわち転送レート (BW) は TCP ウィンドウサイズ (cwnd) と往復遅延時間 RTT で決定され，$BW = cwnd/RTT$ という関係がほぼ成立する．ウィンドウサイズ調整アルゴリズムは，パケット損失に対しては指数的に減少しACK に対しては線形に増加するものでAdditive Increase Multiplicative Decreas(AIMD) と呼ばれる．
遠距離高速ネットワークは Long Fat Pipe Network(以下 LFN と記す) と呼ばれるが，遅延の大きなLFN環境でのACKベースのAIMDアルゴリズムはバンド幅を十分活用できないことが知られている．これは，同じ性能を出すためには遅延時間に比例するサイズのウィンドウサイズが必要となり，またウィンドウサイズの変更速度は，ACKによる推定を利用するため遅延時間に比例するため，ウィンドウサイズ減少からの回復にRTTの2乗に比例するためで，HighSpeed TCP [1] やScalable TCP [2], FAST TCP [3] といったウィンドウサイズ調整の改良が提案されている．
一方，日米間 RTT 200msec, 600Mbps および2.4Gbps帯域ネットワークにおいて，並列ストリームによる高速転送を行なった場合，ストリームごとの速度がばらつきが発生し，時間の経過とともに，この速度差が狭まることは稀で，むしろ差が広がる傾向があることが観測されている．この現象は，Gigabit Ethernet I/F特有のもので Fast Ethernet I/Fでは観測されないためインターフェースによるデータ送出速度と，ウィンドウサイズと RTT で決定される転送レート (BW) の差によって発生するバースト的な振る舞いによって起こされると我々は推測している [6, 7].
我々は，

1. 各ストリームのバースト的振る舞いの抑止
2. 並列ストリームの協調的ウィンドウサイズの調整

を行なうため，以下のようなソフトウェアによるストリームの高速化を行なった．
1. ethernet フレーム間の間隔である Inter Packet Gap(IPG) を延ばすことでインターフェースと転送レート (BW) との差を減じ，各ストリームのバースト的振る舞いの抑止する．具体的には，イーサネットドライバ e1000 に修正を加え，IPG をパラメータ化し設定可能とし LFN 通信においては IPG を最大値 1023 バイトに設定した．
2. 並列ストリームで速度のばらつきをおさえ協調的ウィンドウサイズの調整を行なうため，速い stream を抑制することで速い stream によるネットワークへの負荷を減じ，結果的に遅い stream のバンド幅獲得を容易にすることで全体のバランスをとる方針をとった．具体的には，各コネクションのウィンドウ情報を収集し，ウィンドウサイズに上限を設定するインタフェイスを実装し，外部アプリケーションから各コネクションのウィンドウサイズ調整を行った

4 24,000km データ転送実験

2003年11月にアリゾナ州フェニックスで開催されたSC2003 の バンド幅チャレンジにおいて片側サーバ 33 台ディスク 128 台対向の構成で日米 1 往復半，24,000km のデータ転送実験を行なった．サーバは，IBM x345, Dual Intel Xeon 2.40GHz, 2GB メモリ, Intel 82546EB オンボード NIC, Redhat Linux 7.3, Kernel 2.4.18 USAGI STABLE 20020408 で，各ディスクサーバには，10,000rpm Ultra320 146GB SCSI HDD4 台, 合計 18 ペタバイトのデータディスクを持つ．ネットワークは日米 1 往復半，東京・オ

レゴン州ポートランド間の IEEAF が運用する OC-192(9.6Gbps) を折り返し往復，東京･フェニックスを，NTT コミュニケーションズが運用するネットワーク (4.8Gbps)，APAN が運用する APAN ネットワーク (2.4Gbps)，国立情報学研究所が運営する SUPER-Sinet(1Gbps) の 3 経路で太平洋を渡り，米国 Abilene ネットワークに接続，アリゾナ州フェニックスに到達する経路を取った (図 5)．ネットワークの総長は 24000 km（15000 マイル），遅延時間は，RTT 約 350 ミリ秒，ボトルネックは 3 経路の和による太平洋越えで 8.2Gbps である．

図 5 に，バンド幅チャレンジ時に計測されたスループットと時刻の変化を示す．ここでは，TCP 協調と IPG 調整とを独立に適用しており，500 〜2400sec では 32 台並列 協調 TCP 適用時の，2800sec 〜4200sec では 16 台並列 IPG 調整適用時のデータ転送実験を示している．最大総バンド幅 は協調 TCP 適用時に，7.01 Gbps を記録している．これは総バンド幅の 8.2Gbps の 85% にあたる[1]．ストリーム高速化ではインターフェースのパケット送出レートを下げ高速ストリームの速度の伸びを強制的に落すという，一見，後ろ向きともみえる実装が結果的には，システム全体の性能を著しく向上させた．

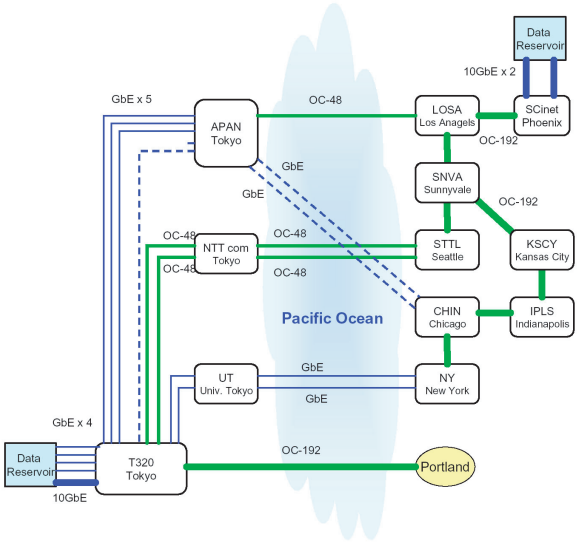

図 5: ネットワーク

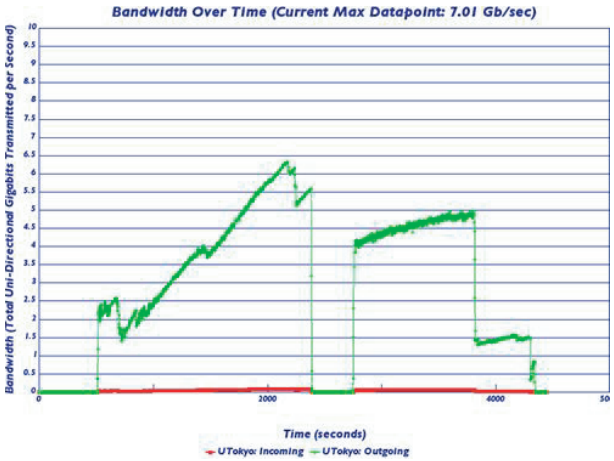

図 6: 実験結果

5 謝辞

本研究は文部科学省科学技術振興調整費先導的研究基盤整備「科学技術研究向け超高速ネットワーク基盤整備」および科学技術振興事業団 CREST による研究領域「情報社会を支える新しい高性能情報処理技術」研究課題「ディペンダブル情報処理基盤」で補助された．日米 24,000km のデータ転送実験は東京大学基盤センター加藤朗助教授，エヌ・ティ・ティ・コミュニケーションズ株式会社，IEEAF，APAN，WIDE プロジェクト，Tyco Telecom，国立情報学研究所，ジュニパーネットワークス株式会社，シスコシステムズ株式会社，物産ネットワークス株式会社，ネットワンシステムズ株式会社，デジタルテクノロジー株式会社の協力により実現された．

参考文献

[1] Sally Floyd, “HighSpeed TCP for Large Congesiton Windows”, Internet Draft, Aug. 2003.
http://www.ietf.org/internet-drafts/draft-ietf-tsvwg-highspeed-01.txt

[2] T. Kelly, “Scalable TCP: Improving Performance in HighSpeed Wide Area Networks”, PFLDnet2003, Feb. 2003.
http://datatag.web.cern.ch/datatag/pfldnet2003/papers/kelly.pdf

[3] C.Jin, et al. “Fast TCP: From Theory to Experiments”, IEEE Communications Magazine, Internet Technology Series, April 1, 2003.
http://netlab.caltech.edu/pub/papers/fast-030401.pdf

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

[5] 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.

[6] M. Nakamura, M. Inaba, K. Hiraki, “Fast Ethernet is sometimes faster than Gigabit Ethernet on LFN — Observation of congestion control of TCP streams”, Proc. PDCS, pp. 854-859, Nov. 2003.

[7] M. Nakamura, M. Inaba, K. Hiraki, ”End-node transmission rate control kind to intermediate routers towards 10Gbps era”, PFLDnet 2004, Argonne, IL, Feb. 2004.

[1]本稿に記載したグラフおよび最大バンド幅は，バンド幅コンテスト中に SCinet(http://scinet.supercomp.org) により計測・記録され公表されたもの