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

データレゼボワールシステムは大域超高速ネットワークを利用し自然科学の実験・観測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る．本稿では, データレゼボワールシステムと核となる大規模データの遠距離超高速転送の実現および 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, 13].

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核となる遠距離データ転送のための並列ストリーム高速化について述べ, 2003 年 11 月に IEEE SC2003(アリゾナ州フェニック) Band Width Challenge において行なった日米 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　遠距離 TCP/IP 通信の問題

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実験的に知られている．これは，同じ性能を出すためには遅延時間に比例する

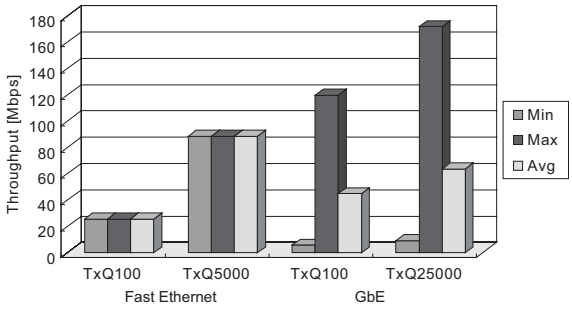

図 1: Throughput using Fast Ethernet and Gigabit Ethernet

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 乗に比例するためと考えられており，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．しかしながら, TCP/IP の LFN での性能の悪さはウィンドウサイズ調整の問題のみに起因するとは考えられない.

図 1 に日米間 RTT 200msec, 600Mbps および 2.4Gbps 帯域ネットワークで Gigabit Ethernet I/F および Fast Ethernet I/F を使い高速転送を行なったときのスループットの比較を示す．Gigabit Ethernet I/F を利用するとストリームごとの速度にばらつきが発生するが, まったく同じ状態, すなわちウィンドウサイズ調整のメカニズムは同一な状態で, I/F を Fast Ethernet に変更し転送を行なうと性能はばらつかず, 最低スループット, 中間値スループットともに Gigabit Ethernet より FastEthernet の方が高速であることが観測される (詳細は [6, 7]). また, 我々は RTT の異なる ストリームについてフロータイムおよびバンド幅利用率に関する AIMD の理論的な解析を行なったが [11, 12], 得られた結果は実験結果と大きく異なっている．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].

3 バースト性の改善とばらつきの抑制

我々は，より性能を高く安定させることを目的とし, (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ことで，TCP/IP の理論的なふるまいと現実のふるまいを近づけるため, それぞれ,

Transmission Rate Controlled TCP (TRC-TCP)
インターフェースによって規定される 速度と TCP の window size の調整によって規定される速度をできるかぎり近づけることで, バースト性の排除を行なう．イーサネットのフレーム間の idle 状態を Inter Packet Gap(IPG) というが, ethernet driver e1000 を変更することで IPG を調整する IPG tuning 方式と, TCP の Slow Start フェーズにおいて 1 つの ACK に対して, 二つの パケットを送出するが, この二つのパケットを, 両方とも ACK 受信直後に送出せず ACK 受信直後に一つパケットを送出したあと次の ACK の想定受信時刻の半分待って残りのパケットを送出する Packet Spacing をそれぞれ実装した．図 2 に IPG を 8 から 1023 に変化させた IPG tuning の効果を示し図 3 に Spacing を行なった結果を示す.

Dulling Edge of Cooperative Parallel Streams (DECP)
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

を提案，実装し実験を行なった.

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

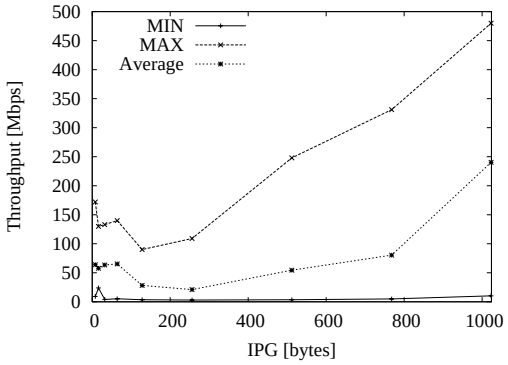

図 2: IPG and transfer rate with GbE RTT=198ms

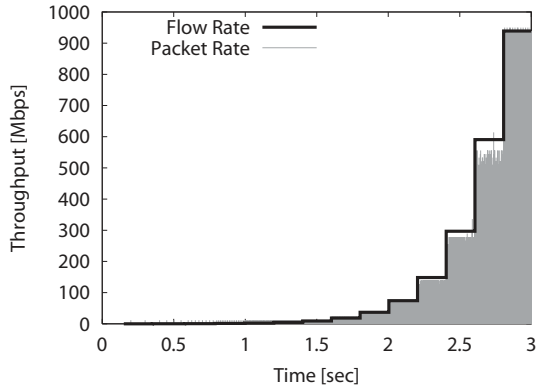

図 3: Flow- and packet-level rate of TRC-TCP during slow start. RTT is 200 ms

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 4)．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 4 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，DECP と TRC-TCP とを独立に適用しており，500 〜2400sec では 32 台並列 DECP 適用時の，2800sec 〜4200sec では 16 台並列 TRC-TCP 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 DECP 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

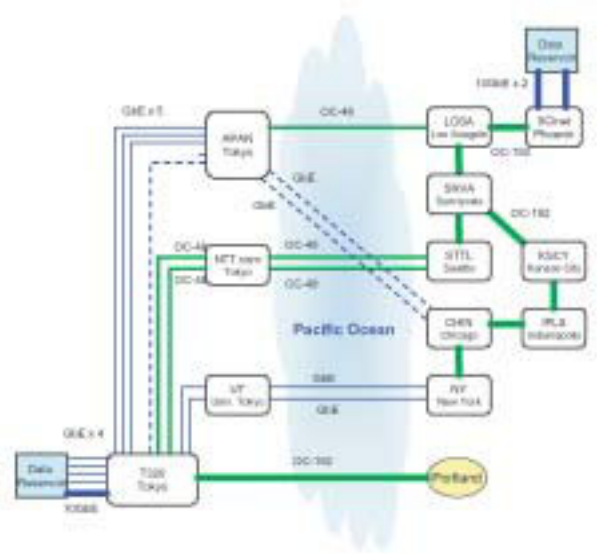

図 4: ネットワーク

図 5: 実験結果

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．

5 まとめ

Data Reservoir システムの中核をなす並列ストリーム高速化についてバースト性の軽減という観点から

[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

シングルストリームの高速化のための TRC-TCP および 並列ストリームの協調的動作のための DECP を提案し, 実験によりその有効性を確認した. 上記 2 手法は, 組み合わせて使うことが可能であり, またウィンドウサイズ調整メカニズムとは独立であるため FastTCP, ScalableTCP, HSTCP などと同時に利用することが可能である. TRC-TCP については現在, ハードウェア化の検討を行なっている [9, 10].

なお, Data Reservoir システムは, バンド幅距離積の世界記録を達成, 更新しており, SC2003 バンド幅チャレンジにおいて「最高バンド幅・距離積・ネットワークテクノロジー賞」を受賞した. これは, バンド幅チャレンジ 2002 における「最高効率賞」に引き続き, 2 年連続の受賞となっている.

6 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.

[8] 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.

[9] 菅原豊　千本潤介　稲葉真理　平木敬, ”10 ギガビットイーサネットを対象とした再構成型ネットワークプロセッサ”, A reconfigurable Network Processor for 10Gigabit Ethernet, 第 1 回リコンフィギャラブルシステム研究会論文集 pp249-256, Sep, 2003.

[10] 千本 潤介 中村 誠 稲葉 真理 平木 敬”FPGA を用いた NIC レベルでのストリームマネージメントアーキテクチャ　ーギガオーダーネットワークにおける通信の最適化”, 第 1 回リコンフィギャラブルシステム研究会論文集　pp263-270, Sep, 2003.

[11] 伊藤剛志, 稲葉真理, ”長距離・短距離通信が混在する環境での TCP/IP のデータ転送速度の理論的解析”, 第 93 回アルゴリズム研究会　Jan 2004

[12] Tsuyoshi Ito and Mary Inaba, ”Theoretical Analysis of Performances of TCP/IP Congestion Control Algorithm with Different Distances”, Networking2004 May 2004 (To appear)

[13] 中村誠, 来栖竜太郎, 坂元眞和, 古川裕希, 生田祐吉, 下國治, 下見淳一郎, 陣崎明, 玉造潤史, 稲葉真理, 平木敬, ”高レイテンシ環境下におけるデータレゼボワールの性能評価”, 情報処理学会研究報告 2003-HPC-93 (HOKKE-2003), pp37-42

[14] 中村誠, 稲葉真理, 平木敬, ”ギガビットイーサネット上での遠距離 TCP 通信における Packet Spacing”, IPSJ SIG Technical Reports, 情報処理学会研究会報告, 2003-QAI-9 高品質インターネット研究報告 No.9, pp13-18