Vol.2009-ARC-184 No.23
2009/8/5

TCP による長距離ディスク間データ転送の高速化

谷田直輝[†1]　稲葉真理[†1]　平木敬[†1]

我々は長距離ディスク間データ転送システム ICDC – InterContinental Disk Copy を構築した．ICDC は背景トラフィックの存在する Long Fat-pipe Network（LFN）を経由して大量のデータをディスク間で高速に遠距離に転送することを目的とし，単一 TCP ストリームによるデータ転送を行うことで，一般回線上でも安定して高性能で動作するロバスト性を実現する，小型汎用 PC で構成されるシステムである．ICDC は IPG の調整によってネットワークの高速化を行い，SSD および Direct I/O を組み合わせることでストレージの高速化を行っている．我々は ICDC-1Gbps モデルを用いて日欧間のボトルネック帯域が 1Gbps である LFN 上でデータを転送し，860Mbps のスループットを得た．また，ストレージ性能が 10Gbps ネットワーク世代に対応可能であることを，ICDC-10Gbps モデルを用いて 7Gbps で 180 秒間ネットワーク経由のディスク書き込みを行い実証した．既に我々は 10Gbps LFN におけるメモリ間転送で理論限界に近い速度でのデータ転送を達成済みであり，10Gbps LFN での超高速ディスク間転送を実現する予定である．

Performance Optimization of Disk-to-Disk TCP Data Transfer over Long-Distance Network

Naoki Tanida,[†1] Mary Inaba[†1] and Kei Hiraki[†1]

We developed ICDC – InterContinental Disk Copy, which aims to transfer huge data between disks for long distance via Long Fat-pipe Networks (LFNs) with some background traffic. ICDC consists of a small commodity PC and transfers data with a single TCP stream, which contributes to a stable high performance communication in shared lines. Its high network performance is based on IPG control and its high storage performance is based on the combination of SSDs and Direct I/O. Using ICDC-1Gbps model, we transferred data through a LFN between Japan and France, whose bottleneck bandwidth was 1Gbps. We attained 860Mbps throughput. Using ICDC-10Gbps model, we also demonstrated storage is 10Gbps-network-capable by storing received data into disks for 180 seconds at 7Gbps. We have already attained the throughput of theoretical bound in memory-to-memory data transfer on 10Gbps LFN. We plan to achieve ultrafast speed disk-to-disk data transfer via LFNs.

1. はじめに

コンピュータは初期のころから科学の進歩に重要な役割を果たしてきた．特にスーパーコンピュータは数値流体力学，有限要素法，量子色力学，分子動力学や第一原理といった科学的シミュレーションにおける数値計算に用いられてきた．こういった数値計算に加え，核融合炉，天文台や加速器といった大規模な施設において，大量の実験データ，観測データや設計データを扱う需要が生まれている．これらデータインテンシブ計算の分野では装置が極めて高価になる傾向があり，多くの研究グループが世界的な協力体制を取ることが一般的となっていて，スケーラブルなデータ共有の仕組みが求められている．例えば，国際熱核融合実験炉（ITER）は技術科学的な巨大プロジェクトであり，日中韓印米露欧の国際協力の下に仏カダラッシュに核融合炉を建設している．しかしながら現状では，ネットワーク帯域を十分に生かした設計データの共有を行えておらず，核融合炉の稼働時には 1 日に数十 TB のデータが生成され実験参加国が国際的に共有する必要があるため，プロジェクトを円滑に実施するためには各国を結ぶ高速なデータ転送システムが不可欠である．

一方，長距離高帯域ネットワーク（Long Fat-pipe Network，LFN）は高遅延である，一般回線には背景トラフィックが存在する，経路の途中に存在するスイッチはパケットをバッファリングする，などの理由から大量のデータを高速に長距離転送をするのは困難であり，これを解決するために多くの研究がおこなわれてきた．効率的に速度を回復するための輻輳回避アルゴリズム，送信速度を抑えバースト転送を回避するためのペーシング技術やエンドホストの CPU 負荷を軽減するためのチューニング技術などが提案されている．これらにより徐々に LFN における効率的な転送が達成されてきた．我々は 10Gbps ネットワークにおける理論限界に近い 95%の帯域利用率を達成する過程で，LFN 上でのデータ転送におけるバースト性によるパケットロスを避けるには物理層におけるハードウェアでのペーシングが有効であり，アプリケーション層や TCP 層では解決できないことを示した[13]．過去の実験は主にパケットロスの発生しない専用線を用いて，メモリ間通信という理想的な条件で行っている．

5 章で挙げるように，これまで様々なデータインテンシブ計算のためのファイルサービス

†1 東京大学
the University of Tokyo

が提案されてきたが，科学技術のためのデータ共有には，通信の高速化だけでなく実際にファイルサービスを大域化することが必要である．我々は小型汎用 PC で構成される長距離データ転送装置を構成した．単一 TCP ストリームでのデータ転送により，一般回線上でも安定して高性能で動作するロバスト性を備えることを特徴とする．本稿では，東京大学-ITER 間の 1Gbps ネットワークを用いてディスク間データ転送実験を行い，一般回線上での性能を実証した．また，ストレージ性能が 10Gbps ネットワーク世代に対応可能であることを，ネットワーク経由のディスク書き込み実験を行うことにより実証した．

2. LFN 上のデータ転送における問題点

2.1 理想的な環境における問題点

TCP/IP は信頼性のある通信のための標準的なプロトコルであり，信頼性を実現するために TCP は ACK を用いる．送信側は受信側から ACK が返ってくるまで再送のためにデータを保持し，送信されたが ACK が返ってきていないデータはインフライトデータと呼ばれ，インフライトデータの最大サイズはウィンドウサイズと呼ばれる．ここで RTT は往復遅延時間を意味する．従って最大転送速度は $window\_size/RTT$ で表わされる．多くの輻輳回避アルゴリズムは転送速度を調整するためにウィンドウサイズを用いる．この方法は LFN 上で通信を行う際に次のような問題点を抱えている．

(1) 同じ転送速度を得るために必要なウィンドウサイズは RTT に比例し，同じウィンドウサイズを得るために必要な時間は RTT に比例するため，同じ転送速度を得るために必要な時間は RTT^2 に比例する．

(2) RTT が大きくなるにつれて TCP スタックによって調整可能なデータ転送速度の粒度は荒くなる．RTT が 198ms の環境でデータ転送を行い，1ms 当たりのパケット数を調べてみると，RTT 時間を周期としてバースト状態とアイドル状態を繰り返す（図 1）．例えば，TCP スタックが転送速度をネットワーク帯域の 1/5 に絞ろうとしても，NIC は最大速度で $\frac{1}{5}RTT$ 時間転送し，$\frac{4}{5}RTT$ 時間アイドル状態になることがわかる．

(1) については，High Speed TCP[3]，Fast TCP[11]，BIC TCP[12] や CUBIC TCP[4] といった大きな RTT に対する多くの輻輳回避アルゴリズムが提案されてきた．(2) については，我々はハードウェアによるペーシングが有効であることを実証している[13]．

2.2 実環境における問題点

2.1 章で挙げた問題点は RTT が大きいことに起因する．一部の実験はネットワークに人工的な遅延を発生させる機材を用いて RTT を大きくすることによって行っており，この環境を疑似 LFN と呼んでいる．疑似 LFN ではなく，実際に科学データを LFN 上で転送する際には更に下記のような問題点がある[13]．

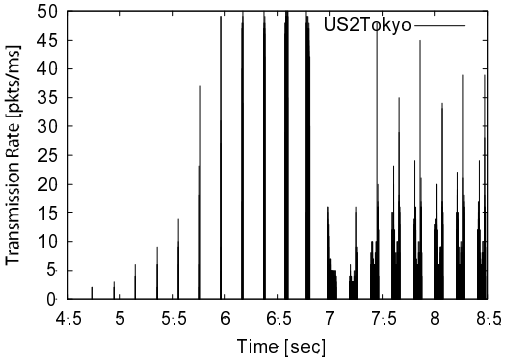

図 1 微視的観点から見たスループット
Fig. 1 Microscopic view of throughput

(1) 中間のスイッチにはバッファが存在する．ACK パケットがバッファリングされ，間隔が抑制されることにより，データの送信にバースト的な挙動が生じる．これは微視的な観点からみたスループットを押し上げる．

(2) 背景トラフィックによりパケットが中間のスイッチでバッファリングされ，バッファ溢れにつながる．また，背景トラフィックの流量は常に変化する．

(3) データを恒久的に用いるためにはストレージに保存する必要がある．通常ストレージの速度はネットワークの速度よりも遅く，バッファ溢れによるパケットロスにつながる．

(1) については，我々は 2.1 章（2）同様ハードウェアによるペーシングが有効であることを実証している[13]．(2) についてもハードウェアによるペーシングが有効であること及び，(3) については従来は大規模なストレージシステムを構築する必要があったが，Solid State Drive（SSD）の爆発的な普及により，ネットワークの速度に匹敵する性能の小型なストレージを構築することが可能になったことを，4 章における実験で実証する．

3. ICDC - 設計と実装

2 章で述べた問題点を考慮に入れ，我々は ICDC – InterContinental Disk Copy を構築

した（図 2）．単一 TCP ストリームによるデータ転送を行い，一般回線上でも安定して高性能で動作するロバスト性を備えた，汎用 PC で構成されるシステムを目指した．

ディスク間転送を実現するためには，ネットワーク上の転送速度以上の速度でストレージにアクセスする必要がある．ネットワークの理論限界速度で転送を行う場合，1Gbps ネットワークでおよそ 125MB/s，10Gbps ネットワークではおよそ 1.25GB/s の性能がストレージに要求されることになる．小型の装置でこの性能要求を満たすため，ストレージにはハードディスクではなく SSD を採用した．また，高速なネットワークとストレージを並列に処理するためには高いメモリ帯域が要求される．例えば 10Gbps で書き込みを行う場合を考えると，NIC からカーネル空間，カーネル空間からユーザ空間，ユーザ空間からカーネル空間，カーネル空間からストレージと 4 回メモリコピーが行われることになり，メモリバスをデータは 6 回通過することになる．そのため単純計算でも最低 1.25GB/s * 6 = 7.5GB/s のメモリ帯域が必要である．そのため，メモリ周りの性能が向上した Intel Core i7 と DDR3 SDRAM の組み合わせを採用した．

ICDC は単体の CPU（Intel Core i7 940），6GB DDR3 SDRAM（1333MHz）及び SSD（Intel X25-E）を MicroATX マザーボード（ASUS Rampage II GENE）に搭載した小型の汎用 PC で構成される．ICDC-1Gbps モデルでは 1 枚の SSD およびネットワークカードに Intel 82562EI を，ICDC-10Gbps モデルでは 8 枚の SSD（Adaptec RAID 5805 による RAID 0）およびネットワークカードに Chelsio S310E を使用する（図 3，図 4）．いずれも各 SSD は SATA 3Gbps で接続される．ソフトウェア環境には CentOS 5.3，linux-2.6.18-128.el5，ext 3 および BIC TCP を使用し，データ転送には iperf-2.0.4-modified を用いて実験を行った．

ICDC の特徴の 1 つはペーシングである．2.1 章で述べたように，TCP による転送はバースト状態とアイドル状態を *RTT* 時間毎に繰り返す傾向がある．IEEE イーサネット標準に従うと，イーサネットアダプタは連続して送出されるパケットの間に遅延を挟む必要があり，これは Inter Packet Gap（IPG）と呼ばれる．Intel 82562EI と Chelsio S310E を含む多くのイーサネットアダプタは IPG のパラメータをソフトウェアで変更可能である．例えば，Intel 82562EI では IPG の値は 4byte から 1027byte まで 1byte 刻みに設定可能であり，Linux のドライバ e1000e のいくつかのパラメータを書き変えることによって IPG の値を変更した．また，Chelsio S310E では，IPG の値は 8byte から 2048byte まで 8byte 刻みに設定可能であり，コマンドラインツールから IPG の値の変更ができる．この IPG を長くすることによってバースト転送の限界速度を下げ，アイドル状態の時間を短くしつつスループットを維持した．一般回線では背景トラフィックとの干渉を避けるため，このペーシングが重要となる．

図 2 ICDC の外観
Fig. 2 Appearance of ICDC

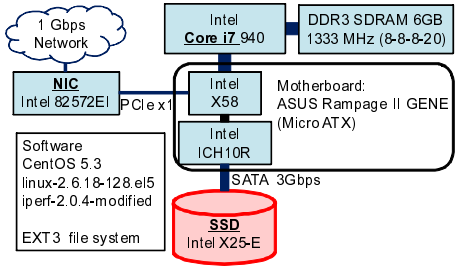

図 3 ICDC-1Gbps モデルのブロック図
Fig. 3 Block Diagram of ICDC-1Gbps Model

次に，ICDC におけるストレージ書き込みについて説明する．iperf-2.0.4 を改造し，クライアントにストレージからの送信機能，サーバに受信時のストレージへの保存機能を追加した．クライアントでは，システムコール sendfile() を用いて zero copy を行い，サーバでは pthread によってストレージへの書き込みスレッドを独立させた．サーバにおいて NIC からデータを受け取ったスレッドはユーザ空間に確保した 4GB のリングバッファに書き込みを続ける．同時に，ストレージへの書き込みスレッドはリングバッファからストレージへ Direct I/O を用いて書き込みを続ける．ここで，Direct I/O によるストレージへの書き込

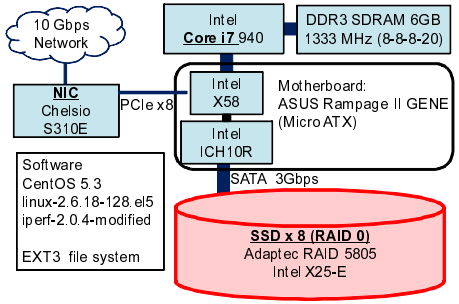

図 4 ICDC-10Gbps モデルのブロック図
Fig. 4 Block Diagram of ICDC-10Gbps Model

みバッファのサイズは 4MB にした．これらにより，平均的な書き込み速度を最大限にしつつ，書き込み速度のぶれをリングバッファで吸収して，NIC がパケットを落とすことを防いだ．

4. 実験および結果

ICDC-1Gbps モデルおよび ICDC-10Gbps モデルの実験を行った．1Gbps モデルの実験は疑似 LFN および LFN でデータ転送を行いスループットを観察した．10Gbps モデルの実験は LAN 内でネットワーク経由のディスク書き込みを行い性能を評価した．

4.1 ICDC-1Gbps モデル

LFN を模擬するために，ネットワークに最大 800ms 程の人工的な遅延を加えることができる Anue H-Series Network Emulator を用いた．図 5 および図 6 は RTT を 302ms としたときのメモリ間データ転送の挙動を示していて，図 5 はペーシングを行わなかったときのスループットであり，図 6 は IPG を 167byte に設定してペーシングを行ったときのスループットである．グラフ中の緑の点と青の点はそれぞれ 1000ms および 1ms でのスループットの移動平均を示しており，両者を比較すると，IPG によるペーシングを行っていない図 5 では 1000ms と 1ms の移動平均が大きく異なる一方で，IPG によるペーシングを行っている図 6 では 1000ms と 1ms の移動平均がほぼ一致していることがわかる．これはバースト的な挙動が抑えられて，スループットが 860Mbps にコントロールされていることを示す．

予備実験の結果を踏まえ，北米大陸を経由し東京から仏カダラッシュに至るネットワーク上でデータ転送を行った（図 7，図 8）．東京からカダラッシュの Cisco C7609-S までは

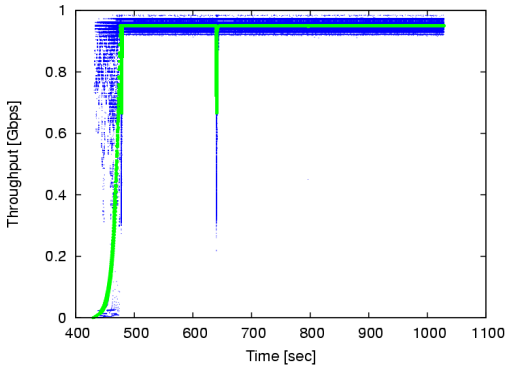

図 5 スループット - メモリ間，IPG 調整無し，疑似 LFN，1Gbps モデル
Fig. 5 Throughput - memory-to-memory, without IPG control, on pseudo LFN, 1Gbps model

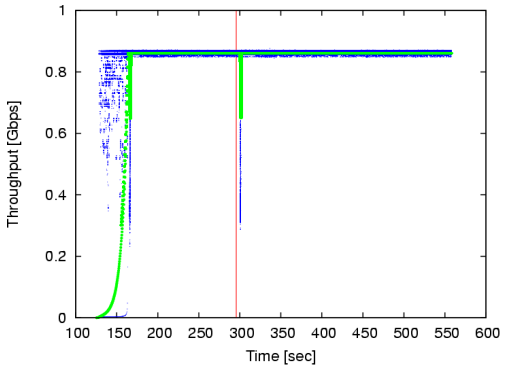

図 6 スループット - メモリ間，IPG 調整有り，疑似 LFN，1Gbps モデル
Fig. 6 Throughput - memory-to-memory, with IPG control, on pseudo LFN, 1Gbps model

10Gbps の回線であったが，カダラッシュ内の Cisco C7609-S から Foundry FESX424 までは 1Gbps の回線であり，このネットワーク経路のボトルネックであった．また，経路の大半は一般回線であり，恒常的に背景トラフィックが存在した．図 9 はペーシングを行っていないメモリ間データ転送のスループットを示す．グラフ中の赤い点は duplicate ACK，すなわちパケットロスが発生したことを示す．1000ms と 1ms におけるスループットの移動

図 7 LFN のネットワーク経路
Fig. 7 Network Path of real LFN

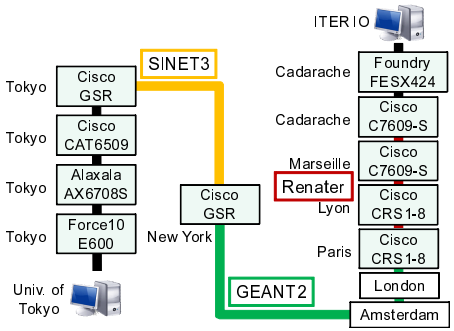

図 8 LFN のネットワークトポロジ
Fig. 8 Network Topology of real LFN

平均が大きく異なり通信が不安定であることがわかるが，実験を繰り返していて，毎回パケットロスのパターンが変わることに気付いた．これは背景トラフィックの違いによる可能性がある．図 10 はペーシングを行ったときのメモリ間転送のスループットを示す．パケットロスが少なくスループットが 1000ms の移動平均 860Mbps で一定に制御されていることがわかる．図 6 と異なって 1ms の移動平均がばらついているのは，受信側で計測行うことにより中間スイッチの影響を反映しているためである．これは LFN におけるデータ転送の難しさを示す例である．最後にディスク間転送の実験結果を示す（図 11）. 図 10 のメモリ間転送と比較して若干性能が低下しているが，ボトルネック帯域の 86%のスループットに達している．

4.2 ICDC-10Gbps モデル

4.2.1 LAN 内での実験

ICDC-1Gbps モデルを用いた LFN での実験の後，LAN 内で ICDC-10Gbps モデルの性能評価を行った．TCP 通信を安定させるためには受信側が処理落ちしないこと，すなわち受信側のディスク書き込み性能が不足してリングバッファを溢れさせないことが重要であるため，送信側のメモリから受信側のディスクへデータを転送し，ネットワーク経由でのディスク書き込み性能の評価を行った．Chelsio S310E に搭載されている粗粒度のスループット調整機能を用いてスループットを 7Gbps に調整し，180 秒間のデータ転送を行った結果，TCP 通信を安定させることに成功した（図 12）．これは 3 分間でおよそ 160GB のデータをネットワーク経由でディスクへと書き込んだことを意味する．

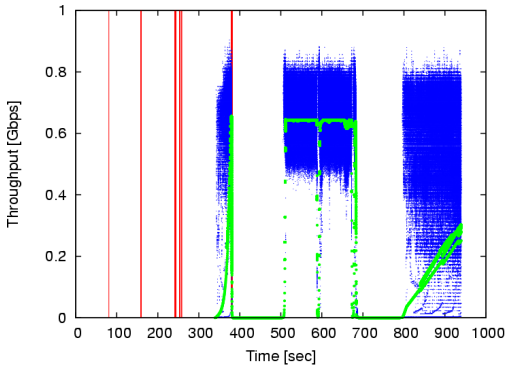

図 9 スループット - メモリ間，IPG 調整無し，LFN，1Gbps モデル
Fig. 9 Throughput - memory-to-memory, without IPG control, on real LFN, 1Gbps model

5. 関 連 研 究

高野らは LFN のためのソフトウェアによる高精度のペーシングメカニズムを設計および評価した.[9]．彼らは Virtual Inter Packet Gap と呼ばれる手法を採用した．彼らはスケジューラを開発するとともに，大きい PAUSE フレームをインターフェース・キューに挿入して IPG として機能するようカーネルに変更を加えた．この手法は正確な IPG 制御を可能

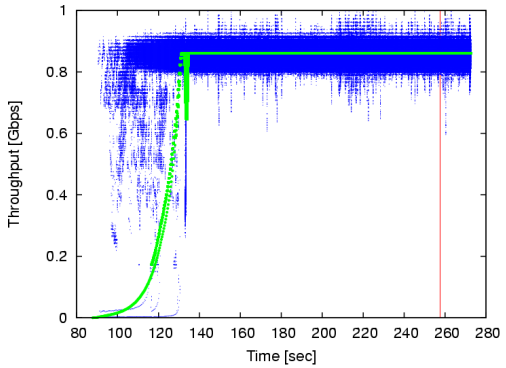

図 10 スループット - メモリ間，IPG 調整有り，LFN，1Gbps モデル
Fig. 10 Throughput - memory-to-memory, with IPG control, on real LFN, 1Gbps model

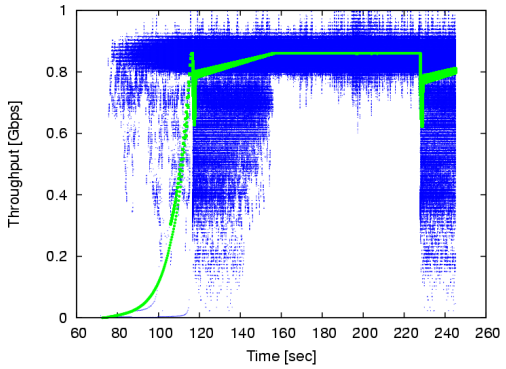

図 11 スループット - ディスク間，IPG 調整有り，LFN，1Gbps モデル
Fig. 11 Throughput - disk-to-disk, with IPG control, on real LFN, 1Gbps model

にするが，バス帯域および CPU を浪費する．一方，ICDC で用いた NIC ベースの IPG 制御はホストに負荷を与えない．

Grid FTP[1] はグリッド環境における一般的なデータ共有システムである．パラレルデータ転送やストライプデータ転送によって性能向上を図る．GPFS[7] はクラスタ環境向けの分散共有ファイルシステムである．ストライピングによって性能向上を図るため，10Gbps

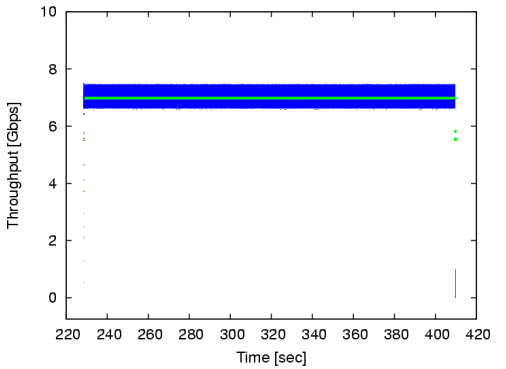

図 12 スループット - ネットワーク経由のディスク書き込み, LAN, 10Gbps model
Fig. 12 Throughput - memory-to-disk, on LAN, 10Gbps model

ネットワークの帯域を使いきるにはノード数を増やす必要がある．PVFS[2] はクラスタ環境における並列ファイルシステムである．Gfarm[10] はメタデータサーバと複数のファイルシステムノードをクラスタ化して構成される仮想的ファイルシステムである．これらを用いて遠隔地にデータを転送する場合，並列 TCP ストリームを扱う必要があるが，一般回線における並列 TCP ストリームはパケットスケジューリングの問題を抱えている．また，背景トラフィックに加えて自らの並列しているストリーム同士で帯域を干渉しあうため，よくチューニングされた単一 TCP ストリームで達成されている理論限界に近いスループットでの転送[13] に匹敵する性能を，並列 TCP ストリームで達成するには困難を伴う．ICDC は単一 TCP ストリームによるデータ転送を行うため，ネットワーク帯域を限界まで利用可能であるという利点がある．

我々の Data Reservoir では iSCSI でのデータ共有を目指してきた[5),6)]．Data Reservoir もまた並列 TCP ストリームの問題を抱えている．Stream Harmonizer[8] はデータ転送時の並列 TCP ストリームの安定化を目的としたハードウェアであり，ファイル転送に有効であると考えられているが，システム構築が必要である．一方，ICDC は小型の汎用 PC1 台で実現している．

6. ま と め

本稿では，汎用ハードウェアおよびソフトウェアを用いた長距離ディスク間データ転送装

置 ICDC を構築した．ICDC-1Gbps モデルを用いた実験では，IPG の調整により ICDC が安定した TCP 通信を米経由日仏間の一般回線で確立し，ボトルネック帯域の 86%の利用率を達成した．これにより一般回線においても IPG の調整が有効であることを実証した．また，ICDC-10Gbps モデルを用いた実験では，7Gbps の速度で NIC からのデータの受信とディスクへの書き込みを同時に 180 秒間維持し続けることに成功した．これは 10Gbps の帯域を利用した高速なディスク間転送に耐える，安価で小型のストレージシステムが構築できたことを意味する．UDP ベースの TCP アクセラレータといった特別な装置及びエンタープライズ向けのハイエンドのストレージ装置は一切使用していない．これはパラメータチューニングおよびペーシングと安価で小型のストレージシステムによって高速な長距離ディスク間データ転送に対応可能なことを示している．我々は 10Gbps LFN におけるメモリ間転送で理論限界に近い速度でのデータ転送を達成済みであり，10Gbps LFN における IPG 調整も行っていることから，コンパクトな汎用装置による 10Gbps LFN 上での高速なディスク間転送に必要な技術が既に揃ったということができ，実証実験の準備を進めているところである．また今後の課題として，背景トラフィックの流量の変化に対応するため，IPG の動的自動チューニングを行うことを考えている．

謝辞 実験に際して助言や支援を頂いた WIDE Project の加藤朗氏と山本成一氏，東京大学の菅原豊氏，玉造潤史氏，吉野剛史氏と小泉賢一氏，APAN の田中仁氏と池田貴俊氏に感謝します．日仏間の実験に協力していただいた NIFS の長山好夫氏，中西秀哉氏と山本孝志氏，ITER 機構の Wilhelm Bjoern 氏と Hans-Werner Bartels 氏に感謝します．実験に際してネットワークを提供してくださった Renater，SURFnet，SINET3，JGN2plus，APAN と Geant2 に感謝します．

参 考 文 献

1) Allcock, B., Bester, J., Bresnahan, J., Chervenak, A.L., Foster, I., Kesselman, C., Meder, S., Nefedova, V., Quesnel, D. and Tuecke, S.: Data management and transfer in high-performance computational grid environments, *Parallel Comput.*, Vol.28, No.5, pp.749–771 (2002).
2) Carns, P.H., Ligon, W.B., III, Ross, R.B. and Thakur, R.: PVFS: A Parallel File System for Linux Clusters, *In Proceedings of the 4th Annual Linux Showcase and Conference*, USENIX Association, pp.317–327 (2000).
3) Floyd, S.: HighSpeed TCP for Large Congestion Windows, *RFC 3649*, Internet Engineering Task Force (2003).
4) Ha, S., Rhee, I. and Xu, L.: CUBIC: a new TCP-friendly high-speed TCP variant, *SIGOPS Oper. Syst. Rev.*, Vol.42, No.5, pp.64–74 (2008).
5) Hiraki, K., Inaba, M., Tamatsukuri, J., Kurusu, R., Ikuta, Y., Koga, H. and Zinzaki, A.: Data Reservoir: Utilization of Multi-Gigabit Backbone Network for Data-Intensive Research, *SC Conference*, p.24 (2002).
6) Kamezawa, H., Nakamura, M., Tamatsukuri, J., Aoshima, N., Inaba, M. and Hiraki, K.: Inter-Layer Coordination for Parallel TCP Streams on Long Fat Pipe Networks, *SC '04: Proceedings of the 2004 ACM/IEEE conference on Supercomputing*, IEEE Computer Society, p.24 (2004).
7) Schmuck, F. and Haskin, R.: GPFS: A Shared-Disk File System for Large Computing Clusters, *FAST '02: Proceedings of the 1st USENIX Conference on File and Storage Technologies*, USENIX Association, p.19 (2002).
8) Sugawara, Y., Inaba, M. and Hiraki, K.: Flow Balancing Hardware for Parallel TCP Streams on Long Fat Pipe Network, *FGCN '07: Proceedings of the Future Generation Communication and Networking*, IEEE Computer Society, pp.391–396 (2007).
9) Takano, R., Kudoh, T., Kodama, Y., Matsuda, M., Tezuka, H. and Ishikawa, Y.: Design and evaluation of precise software pacing mechanisms for fast long-distance networks, *In Proceedings of PFLDNet 2005* (2005).
10) Tatebe, O., Morita, Y., Matsuoka, S., Soda, N. and Sekiguchi, S.: Grid Datafarm Architecture for Petascale Data Intensive Computing, *Cluster Computing and the Grid, IEEE International Symposium on*, p.102 (2002).
11) Wei, D.X., Jin, C., Low, S.H. and Hegde, S.: FAST TCP: motivation, architecture, algorithms, performance, *IEEE/ACM Trans. Netw.*, Vol.14, No.6, pp.1246–1259 (2006).
12) Xu, L., Harfoush, K. and Rhee, I.: Binary increase congestion control (BIC) for fast long-distance networks, *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, Vol.4, pp.2514–2524 vol.4 (2004).
13) Yoshino, T., Sugawara, Y., Inagami, K., Tamatsukuri, J., Inaba, M. and Hiraki, K.: Performance optimization of TCP/IP over 10 gigabit ethernet by precise instrumentation, *SC '08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, IEEE Press, pp.1–12 (2008).