Cometインテリジェント NICの応用（第1版）

稲葉真理[†] 陣崎明[††] 平木敬[†]

本稿では2002〜2004年度に行われた高速遠距離データ通信に関する富士通研究所Cometグループと東京大学Data-Reservoirグループとの共同研究について述べる．本共同研究は巨大データを取り扱う実験科学のためデータ共有基盤Data Reservoirシステム実現のために，ネットワークストリーム処理Comet技術を応用し，遠距離データ転送のバンド幅距離積世界記録を達成した．また共同研究を進めるうちに当初の想定を上回る多様な応用への展開がなされた．

Application of Comet Intelligent NIC (Version 1)

Mary Inaba,[†] Akira Jinzaki[††] and Kei Hiraki[†]

This paper describes joint research of Data Reservoir Group of University of Tokyo and Comet Group of Fujitsu Laboratory. This joint work aims to develop information infrastructure for scientific researchers, named Data Reservoir System, using Comet Network Processor, and we successfully attained the world record of the distance bandwidth products. In addition, for pre-experiments and data analysis, we apply Comet technology to several other network facilities such as packet analyzer and pseudo Long Fat Pipe Network (LFN) emulator.

1. はじめに

近年のネットワーク技術の進歩は急速であり，年々プロセッサ速度向上を上回る速度向上が実現している．また，高速ネットワークが新たに開拓したアプリケーション分野も急速に拡大し，相乗効果によりネットワークシステムは最も急速に進化しつつある分野の1つと考えられる．特に高速ネットワークに係るハイエンド技術では研究から実用化までのサイクルが短く，厳しい国際競争の中で研究開発分野を自ら確立していくことが求められている．

高速ネットワーク処理に用いるチップを独自に開発する場合，研究開発には莫大な金銭的投資が必要となり，これを大学の一研究室レベルで行うことは一般には困難である．また競争的研究資金を獲得し研究を行う場合，多くのケースでは資金獲得時に提案書で示した成果をあげ，論文としてまとめて世に問うという義務が発生するため，論文になりにくい実装・開発，あるいは完成品とするための作り込みの部分が軽視されやすい傾向がある．一方，企業では採算の見込みがたつ開発に対してはかなりの量の物的・人的リソース投入が可能であるが，採算の見込みが不明確な冒険的な開発や，ある要素技術の特長を活かしてはいるが短期的収益性の悪い応用開発に対するリソース投入は厳しいことが多い．また企業規模が大きく分業体制が確立している大企業においては，要素技術開発と，それを利用する応用開発は異なる部署が担当することが一般的であるが，その完全な相互理解はしばしば困難をともなう．企画部門がその役割を担うことも多いが，いずれの場合も，要素技術開発者が把握していた潜在的可能性を応用にあたり十分に活かしきれないことが多い．企業からみた産学協同研究のメリットは開発すべき先端技術の選択と方向性が「学」の知見によって明確化されること，短い期間で検討実装検証を繰り返すことで正しい方向への軌道修正が行え，結果的にリスクが軽減されることである．

本稿では2002年度から2004年度にかけて行われた富士通研究所Cometグループと東京大学Data-Reservoirグループとの高速遠距離データ転送に関する共同研究について述べる．この共同研究は，巨大データを取り扱う実験科学のためデータ転送を効率良く高速に行うための研究基盤Data Reservoirシステム[1)〜3)]実現のために，富士通研究所が開発したネットワークストリーム処理のためのComet（COMmunication Enterprising Technology）技術[4),5)]を活用し超高速TCP通信を実現することを目的として開始さ

† 東京大学
The University of Tokyo
†† 富士通株式会社
FUJITSU JAPAN

れた．共同研究開始時にはComet Network Processor（Comet-NP）が完成しており，共同研究期間中に，この Comet-NP を搭載したプログラマブル intelligent NIC（Network Interface Card）Comet i-NIC および DVD でよく用いられる IEEE1394 規格（以下 i.Link）を UDP にエンキャプシュレートして転送するための映像転送に特化したネットワークカード Comet-DVIP が商品化された．共同研究開始時は，Comet を搭載した NIC を利用したネットワークストレージシステムの開発が共同研究の目標であったが，共同研究実施の過程において検討を進めていくうちにデータ通信の高速化のほかにも応用範囲が広がり，Comet i-NIC と汎用 PC を組み合わせて実験に必要なテスト環境生成装置やパケット解析装置の開発を行った．また一般の遠隔会議システムでは黒板の文字が読めないため Comet DVIP を利用した黒板の文字が読める高精細遠隔講義システムの開発も行った．Comet 技術の特長を活かした応用開発という点から考えると,「共同研究」という枠組みがあったため，企画段階でさまざま検討が行え，短いサイクルでポジティブフィードバックが可能であるという利点が大きかった．

本稿では，2 章でまず Comet 技術の概説を行う．3 章では Comet i-NIC を利用した実験解析機材，パケットログ収集装置および擬似遠距離環境生成装置について述べる．4 章では遠距離高速転送を実現するための Comet-TCP について述べ，擬似遠距離環境および日米 1 往復半の実ネットワークで行った遠距離ディスクデータ転送の実験結果について記す．5 章で黒板に書いた文字が読める遠隔講義システムを紹介し，6 章でまとめる．

2. Comet 技術

2.1 背　　景

近年のネットワークの速度向上はめざましいものがあり，プロセッサスピードの速度向上を上回っている．たとえば，1983 年に 10 Mbps の 10 BASE-5（IEEE802.3 CSMA/CD）が標準化されてから 1995 年に IEEE 802.3u Fast Ethernet が標準化されるまで 12 年の歳月が流れたが，ギガビットイーサネット（以下 GbE）IEEE802.3z 1000BASE-SX が標準化されたのは 1998 年，そして 10 ギガビットイーサネット（以下 10 GbE）IEEE802.3ae 10 Gbit/s Ethernet over fiber が標準化されたのは 2003 年であり，8 年間で約 100 倍の通信速度の向上が見られる．一方，プロセッサ速度向上は 1997 年当時，約 100 MHz が，現在約 4〜6 GHz，とほぼ 50 倍であり，標準的なベンチマークを用いた計算速度においても，ほぼムーアの法則に則った速度向上が実現しており，ネットワークのデータ転送速度向上は，プロセッサチップの性能向上を上回っている．その結果，たとえば IEEE802.3i 10BASE-T は当時の Workstation でフィルタリング処理が十分可能であったが，10 GbE のフィルタリング処理を現在一般的な PC，Workstation で行うことは不可能であり，従来メインプロセッサで行えたプロトコル処理に対し何らかの高速化のアプローチが必要になってきている．高速なプロトコル処理機構としては，専用ハードウェアによるもの（主としてスイッチ・ルータ）とマイクロプロセッサによるもの（主として NIC）に大別される．専用ハードウェアは高速処理が可能である反面，複雑な処理を組みにくくプロトコル変更に柔軟に対応できない．一方，マイクロプロセッサは複雑な処理を実現できプロトコル変更にも柔軟に対応できる反面，性能的限界が厳しい．これらの問題を解決するためには柔軟性を持つ高速パケット処理が可能なハードウェアアーキテクチャが必要である．

2.2 Comet ハードウェアアーキテクチャ

Comet プロトコル処理エンジンは，ストリーム処理に特化したプログラム可能なネットワークプロセッサであり，プロトコル高速処理およびホストの負荷軽減を目標としている．Comet プロトコルエンジンは，プログラム可能なストリームプロセッサと，コントロール用の汎用プロセッサから構成される．ストリームプロセッサは有限状態機械（Finite State Machine，以下 FSM）であり，FSM エンジンと，状態遷移表（State Transition Table，以下 STT）メモリで構成される（**図 1**）．ストリームプロセッサはチェックサム計算やテーブル検索等パケット処理に必要な機能ユニットを備え，On-the-fly で水平型マイクロプログラムによる並列処理を行うことができ，ストリームの実時間処理が可能となっている．コントロールプロセッサはネットワークインタフェース，ストリームプロセッサの初期化や制御，ホスト計算機との通信を行う．

2.3 Comet Network Processor（Comet NP）

Comet Network Processor（Comet NP）は 0.35 μm CMOS テクノロジのプロトコル処理用 FSM を担当するネットワークプロセッサであり汎用プロセッサとともに Comet アーキテクチャを実現する．Comet NP は 2 個のストリームプロセッサ，2 本の PCI-64 bit/66 MHz 外部バスを持ち，80 MHz で動作し，DES/3DES，Checksum，Table Lookup 機能ユニットを持つ（**図 2**）．この機構により，パケットに対す

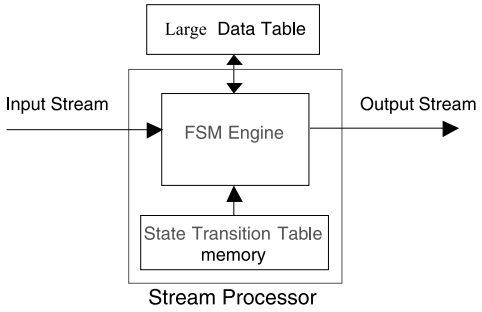

図 1 ストリームプロセッサ構成
Fig. 1 Stream processor.

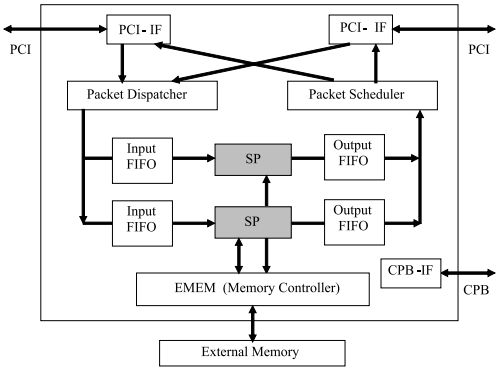

図 2 Comet NP 構成
Fig. 2 Comet NP.

るヘッダ処理，Checksum 処理，暗号化・復号化処理と，付属するバッファメモリを用いた再送処理，遅延時間挿入が可能となる．たとえば IPchecksum，DES 処理であれば 80 MHz 動作時に 320 MB/sec 程度の処理速度を出すことが可能である．Comet NP 処理方式の詳細は文献 4) に詳述されている．

2.4 Comet NIC

富士通研究所は Comet NP を搭載した複数の NIC を開発した．ここでは，本共同研究で活用した Comet i-NIC および Comet DVIP について述べる．

Comet i-NIC はフルサイズ PCI カードフォームをしたプログラム可能なインテリジェント NIC であり，外部用・ホスト用それぞれに 64/32 bit，66/33 MHz PCI bus を持ち，外部用 PCI bus には 2 port オンボード 10/100/1000Base-T が搭載されている（**図 3**）．Comet i-NIC は，プロトコル処理用の Comet NP およびコントロール用汎用プロセッサとして Strong Arm 1110 233 MHz，メモリ 256 MB SDRAM，16 MB Flash を搭載しており，IPsec，Single DES，3DES プロトコルオフロード機能が用意され，automatic forwarding 機能や Interrupt Coalescing 機能を持つ．たとえばパケットフォワーディングを利用し，片方のポートから入ったストリームを on-the-fly で処理し，これを他方のポートから出す，すなわちブリッジすることで，ホストに対してストリームを送らずに処理が行えるため，ホスト側の通信負荷ほぼゼロで高速にストリーム処理を行うことができる（**図 4**）．コントロールプロセッサは Comet NP の STT の書き換えや，外部とのネットワークインタフェースのコントロールを行う．我々は，Comet i-NIC のパケットフォワーディング機能を用いて，パケット解析装置 DR Giga-Analyzer と擬似遠距離環境 Comet-Delay，Comet-Drop を開発，また遠距離ネットワークで安定な TCP 通信を実現する TCP トンネルである Comet-TCP を実現した．

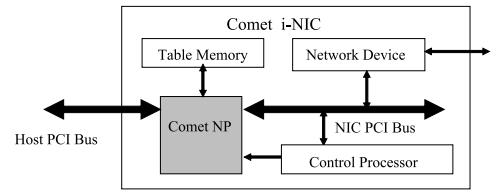

図 3 Comet i-NIC ブロック図
Fig. 3 Block diagram of Comet i-NIC.

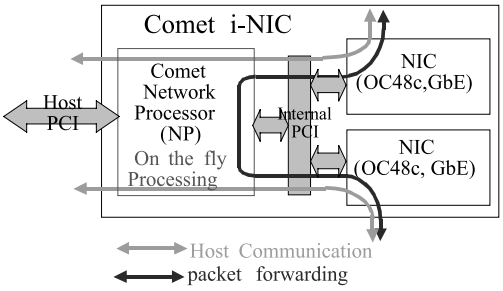

図 4 Comet i-NIC データフロー構成
Fig. 4 Dataflow of Comet i-NIC.

Comet DVIP は DV 映像を UDP にエンキャプシュレートしてネットワークごしに転送することを目的として開発された i.link DVIP 21 cm PCI カードであり，IEEE1394 を 2 port，100B-TX を 1 port 搭載している．Comet NP を利用した IPsec ESP 処理により，映像のセキュアな 3DES つき高速転送が可能になるほか，使用可能帯域に応じて，データ量削減が望ましい場合，音声帯域は削らずに画像帯域のみ削減する機能や，100 μsec 粒度でのデータの平滑化する機能を持っている．我々は Comet DVIP を利用して黒板の字が読める高精細度遠隔講義システムを実現した．

3. Comet を応用したネットワーク実験用機材

3.1 遠距離通信の難しさ

TCP/IP はインターネットで標準的に利用されているプロトコルであるが，高速で遅延の大きなネットワーク（Long Fat pipe Network，以下 LFN）においては十分な性能を出すのが容易でないことが知られている．流量・再送制御は送信データに対する応答（ACK）をもとに行われ，流量制御はウィンドウサイズの調整により実現されている．転送レート（BW）は TCP ウィンドウサイズ（cwnd）と往復遅延時間 Round Trip Time（RTT）で決定され，BW = cwnd /RTT いう関係が成り立つ．すなわち転送レートはウィンドウサイズに比例し，遅延時間に反比例する．このため日米間（RTT 約 200 ms）で 1 Gbps の性能で通信をする場合，理論的にはウィンドウバッファは 25 MB（MTU 1500 バイトで約 16,600 パケット）必要となる．ウィンドウサイズの調整を Additive Increase Multiplicative Decrease（AIMD）方式で行う場合，パケット損失による転送レート減少からの回復にかかる時間は RTT の 2 乗に比例する．このように標準 TCP は LFN ではバンド幅を十分活用できないことが知られており，HighSpeed TCP [6] や Scalable TCP [7]，FAST TCP [8] といったウィンドウサイズ制御の改良アルゴリズムが提案されてきている．しかしながら LFN 環境で高速データ転送を行う際に問題となるのはウィンドウサイズ調整の問題だけではない．

2002 年に行われた Supercomputing 2002 バンド幅チャレンジにおいて，Data Reservoir システムを用いて 12,000 km，RTT 200 msec，ボトルネック太平洋をわたる OC-12（622 Mbps）の日米サーキットでディスクデータ転送実験を行った．**図 5** にバンド幅チャレンジの公式記録のグラフを示す．

実験では Intel Architecture（以下 IA）サーバ 26 台対向でデータ転送を行いネットワーク経路ボトルネックの東京シアトル間 OC-12 の 92%平均利用効率を達成した．しかしながら各ストリームの性能は予測値に比べ悪く，またストリームごとの性能差も著しいことが分かった．原因を探るうち，GbE NIC を Fast Ethernet NIC に交換し他は同じ条件で実験比較すると，GbE NIC 使用は性能がばらつき，Fast Ethernet NIC 使用の方が平均性能が良いことが観察されたが，この差異は TCP ウィンドウ制御アルゴリズムからは説明がつかない[9]．このような性能低下を引き起こす原因を探るため，パケットレベルでの通信の解析，すなわち通信の両端でパケットログを収集しつき合わせて解析を行うことを目標として 1 Gbps ワイヤレートでのパケット記録と解析を行うパケット解析装置 DR Giga-Analyzer を開発した[10]．また実験室内で予備実験を行うための擬似 LFN 生成装置 Comet Delay，Comet Drop を開発した．

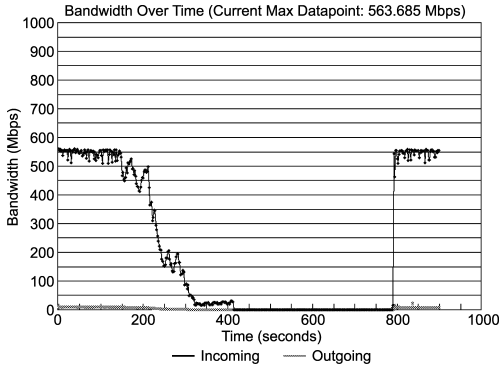

図 5 SC2002 バンド幅チャレンジ公式記録
Fig. 5 SCinet official record of Data Reservoir in SC02 BWC.

3.2 パケットアナライザ DR Giga-Analyzer

DR Giga-Analyzer は 1 Gbps の通信に対し高精度タイムスタンプを付加したパケットログを 2 時間にわたり収集可能なパケットアナライザであり，通信の両端点のログをつきあわせることで遠距離 2 点間の TCP/IP 通信の振舞いを解析することを目的としている．我々は GPS 標準時刻同期タイムスタンプを付加するパケットアナライザ DR Giga-Analyzer を Comet i-NIC 上に実装した．対象とする GbE（1000B-SX）はフルワイヤレート双方向で 2 Gbps で，最短パケット（64 Byte 長）では 2,976,000 pps を超えるキャプチャ能力が必要となる．

DR Giga-Analyzer は Comet i-NIC 上にパケット振り分け機能を実装した Scatter Comet 2 枚を挿した振り分けサーバ 1 台と通常の NIC を挿したデータ格納のための収集サーバ 8 台，光タップおよび GbE スイッチから構成される（**図 6**）．被観測ネットワーク上の双方向（上り/下り）のパケットは，光タップにより分岐され，それぞれ，振り分け装置の ScatterComet に受信される．Scatter Comet 内部にはあらかじめ各収集装置宛ての UDP カプセリングヘッダを用意しておき，GPS 時計によるタイムスタンプを付加したのち，ラウンドロビン方式で受信パケットを 8 台の収集用サーバにフォワーディングする（**図 7**）．収集サーバは，受信したフォワーディングパケットからカプセリングヘッダを取り除き，ディスク装置に書き込みを

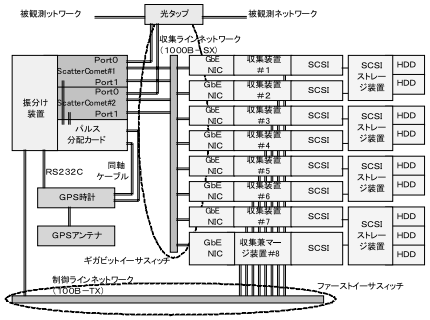

図 6 DR Giga-Analyzer の構成
Fig. 6 DR Giga-Analyzer.

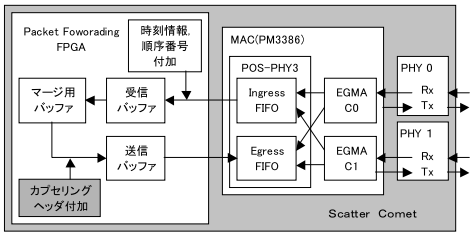

図 7 Scatter Comet ブロック図
Fig. 7 Block diagram of Scatter Comet.

行う．

通常の NIC は複数パケットの送受信を 1 回の割込みで通知する Interrupt Coalescing 機能を有しているが，Coalescing パラメタをチューニングしても，ショートパケット性能は 1 パケット 64 Byte 時で 60 Mbps, 122,700 pps 程度である．この限界を超えるため，Scatter Comet から収集装置にフォワーディングする際，ショートパケットは複数パケットをマージしてフォワーディングし収集装置側の受信パケット数を減らしている．

またギガビットイーサネット環境では，1,500 Byte 長パケットは 12 μsec 間隔で，64 Byte 長パケットであれば 0.5 μsec 間隔で送受信される．一方，異なる 2 点間のネットワークトラフィック解析を行う場合，2 点間で同期した時刻関係を把握することが必要である．このため，DR Giga-Analyzer では GPS 時計が出力する標準時刻と同期する高精度のタイムスタンプを付加した．精度は標準時刻に対する実精度 100 nsec，計時解析度 25 nsec であり，GPS 電波が得られない実験状況においても，1 ppm の時間確度を実現している．

安定した収集性能を確保するため，長時間占有するプロセスの平滑化等のチューニングも実施し，長時間のキャプチャリングに対しては，汎用のサーバとディスク装置を用いることにより，スケーラブルな最大キャプチャ時間の展開が可能な構成としている．また，ヘッダトレースモードを用意しており，パケットのフルキャプチャのほか，受信したパケットの先頭から 64 Byte，128 Byte，256 Byte をキャプチャすることも可能である．収集サーバ側の分散ストレージには Data Reservoir システムで開発した Distributed Shared File Architecture（DSFA）を利用している．各収集装置で分散して記録されたパケット情報は，時刻情報および別途付加されている受信順序番号により整順され，1 つのストリームデータとしてマージされ出力される．キャプチャされたパケットデータは上述の高精度時刻情報とともに，tcpdump フォーマットで出力される．これにより，さまざまな tcpdump 出力形式の恩恵を得ることができる．

完成した 2003 年当時，標準時刻に対する実精度 100 nsec，ディスク容量 2 TB，フルパケットトレースで約 2.3 時間，64 Byte ヘッダトレースでは最大 40 時間超のフルワイヤレートキャプチャが可能であり，双方向 2 Gbps ネットワーク上のパケットを長時間フルトレース可能な解析機器は，DR Giga-Analyzer のみであった．

3.2.1 DR Giga-Analyzer 解析例

米国富士通研究所（以下 FLA）と東京大学の間で 12,000 km，RTT 200 msec，ボトルネック OC-12（622 Mbps）の回線上で通信を行い DR Giga-Analyzer による観察を行った．NIC 以外は同じ環境で，`iperf`，スタンダードフレーム（約 1,500 Bytes）を用いたメモリ間通信実験を行った比較実験結果を示す．**図 8** に GbE NIC を，**図 9** は Fast Ethernet NIC を用いた場合の受信側における 1 msec あたりのパケット到着数を示す．

Fast Ethernet NIC の場合，1 msec ごとに約 8 パケット NIC の速度限界の 9 割程度が均等に届くのに対し，GbE NIC の場合は，RTT 200 msec ごとにサーキットのボトルネックの 9 割程度約約 50 パケットが一瞬届き，その前後にもある程度のパケットが到着するが，全体の 8 割以上の間はパケットが到着せずバースト的な振舞いが観察される．GbE NIC は Fast Ethernet NIC を使用した場合に比べ，たとえ TCP スタックの Congestion Window サイズが同じであっても，パケット送出がバースト的になり，マクロスコーピックには起こる必要のないパケットロスが起き，その結果，性能低下を招いていると推測される．

3.3 擬似遠距離ネットワーク環境生成

帯域を大幅に使用する実験は回線を共有する他のトラフィックに対し多大な影響を与えるため，一般回線上での実験は行えない．しかしながら遠距離回線を排

他的に利用するのは容易ではなく，また多くの人々に協力を仰ぐ必要があるため，実際の回線による実験を行う前に可能なかぎり予備実験により性能予測・改善を行うことが不可欠である．我々は，Comet i-NIC で L2 ブリッジを行うことで，実験室内のローカルなネットワーク環境で，広域ネットワークと同様な実験環境を構築する擬似遠距離ネットワーク環境生成装置を開発した．遅延発生装置 Comet Delay は Comet i-NIC のデュアルポート GbE をネットワークインタフェースとして使用する．あらかじめホストインタフェース経由で遅延時間を 0 秒から 127 秒まで 100 μs きざみで指定する．Comet Delay は一方のポートから入ってきたパケットをいったん Comet i-NIC 上のバッファメモリ上に格納し，指定された遅延時間後に，他方のポートからパケットを送出することで遅延を発生させる．ここでも，パケットフォワーディング機能を用いており，ホスト側の OS オーバヘッドの影響を受けずパケットごとの遅延時間のゆらぎを非常に小さくおさえることができる．

TCP の性能を解析するためには，実際のパケットロスの影響が TCP ウィンドウサイズに与える影響を調べる必要がある．擬似遅延環境エミュレータ Comet Drop は，遅延発生装置 Comet-Delay に，指定した確率でパケットを廃棄する機能を付加したもの，すなわち，Comet i-NIC バッファメモリ上のパケットを（1− 廃棄率）の確率で他方のポートに送るものである．パケット廃棄率は 1/65536 単位で任意の値に設定が行える．

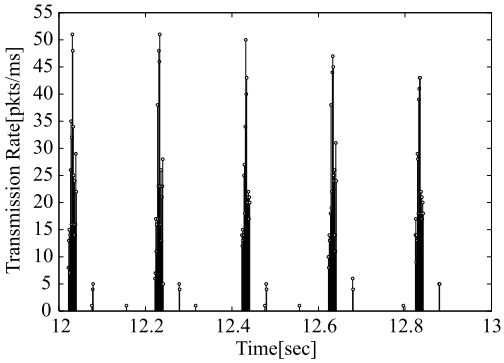

図 8 1 msec あたりのパケット数（GbE NIC）
Fig. 8 Number of received packets per 1 msec (GbE NIC).

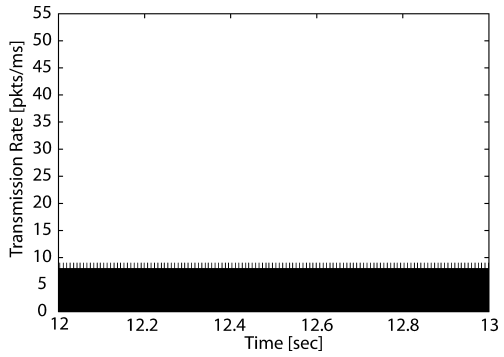

図 9 1 msec あたりのパケット数（Fast Ethernet NIC）
Fig. 9 Number of received packets per 1 msec (Fast Ethernet NIC).

4. Comet を応用した高速通信

4.1 Comet TCP

3.2.1 項で述べたようなバースト的な振舞いをなくし，本来起こらなくてもよいパケットロスによる性能低下を防ぐために，2 点間の TCP 通信を Comet i-NIC を用いて透過的に遠隔中継することで，TCP 通信を高速化する Comet TCP を開発した[12]．Comet TCP は，独自の LFT（Long Fat Tunnel）プロトコルを用い Comet i-NIC は TCP に対するプロキシとして動作する（**図 10**）．

アプリケーションと Comet の間は通常の TCP を用いて通信を行い，遠距離ネットワーク上の Comet 間 LFT 通信は TCP のデータをカプセル化しトンネルする方式をとる．両端の Comet 上でカプセル化・データ解析・再構成の処理を行いデータの送出はハードウェアにより平滑化される．

Comet はアプリケーションのすぐ近くに位置するため見かけの RTT が小さくなり遅延の影響を受けにくく，高スループットを達成できる．Comet TCP は OS のプロトコルスタックや TCP を利用する既存のアプリケーションにいっさいの変更を必要とせず，通常の TCP としてそのまま利用可能であるところに特長がある．

4.2 Comet TCP 実験

ネットワーク遅延がスループットに与える影響について比較するため，3.2 節で述べた Comet Delay を利用し擬似 LFN 環境を作り Comet TCP，Scalable TCP，High Speed TCP，スタンダード TCP のシン

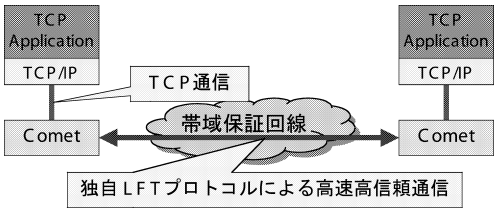

図 10 Comet TCP
Fig. 10 Comet TCP.

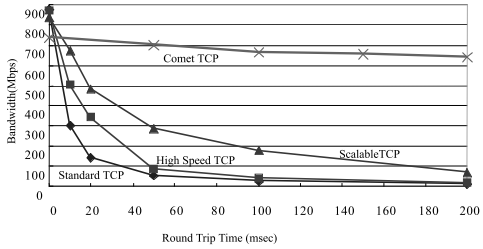

図 11 TCP 変種に対する遅延の影響
Fig. 11 Effects of latency for TCP variations.

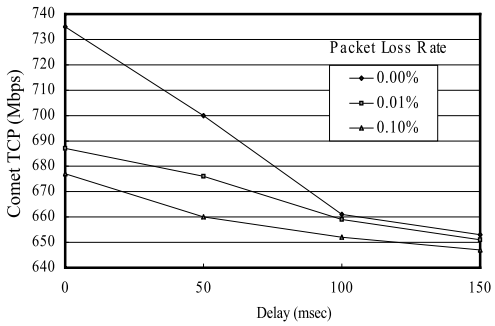

図 12 Comet TCP パケットロス耐性
Fig. 12 Comet TCP packet-loss tolerance.

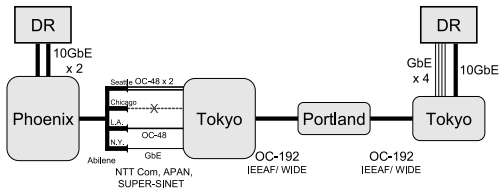

図 13 24,000 km ネットワーク，ボトルネックは 3 経路合計 8.2 Gbps
Fig. 13 24,000 km network, aggregated 8.2 Gbps bottleneck.

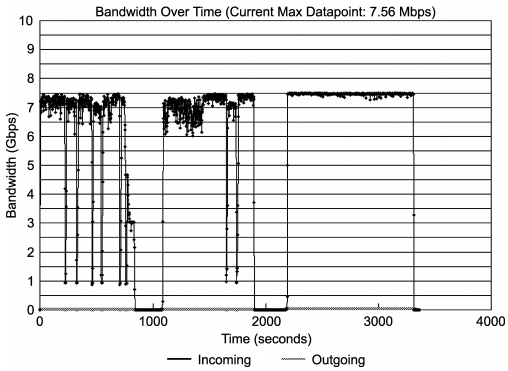

図 14 SCinet によって計測された公式記録（Comet-TCP）
Fig. 14 SCinet official record of Data Reservoir with Comet-TCP.

グルストリーム通信の比較を行った．

図 11 に TCP 変種ごとに遅延を 0 msec から 200 msec まで変化させたときのスループットの変化を示す．TCP および Comet TCP 以外の TCP 変種は 20 msec 遅延で明らかな性能低下が見られ，50 msec では 0 msec の 1〜3 割程度まで性能低下を起こすのに対し，Comet TCP は遅延 200 msec でも 9 割程度の性能が保持できていることが分かる．

TCP はパケットロス等で ACK が戻らない場合急激にスループットを低下させる．Comet TCP はパケットロスの後，見かけの RTT が小さいために，近距離通信と同様に元の通信帯域に復帰させることができる．Comet TCP のパケット落ちに対する耐性を調べるために Comet Drop を用いた実験を行った．

図 12 に，パケットロスレート 0%（パケットロス無し），0.01%，0.1%それぞれについて遅延のスループットに与える影響を示す．Comet TCP では 0.1%パケットロスがあっても，RTT 50 msec ではパケットロス無しに比較し 1 割程度しか性能が落ちず，また遅延が長くなるほどパケットロスによる性能低下の割合は少なくなることが分かる．

擬似遠距離ネットワーク環境下で Comet TCP を利用すると性能低下が十分少なかったため，Supercomputing 03 バンド幅チャレンジで日米 1 往復半，24,000 km，ボトルネック 8.2 Gbps の実回線でディスク間データ転送実験を行った．**図 13** に実験に用いたネットワークの構成図を示す．東京から，Internet Educational Equal Access Foundation（以下 IEEAF）の OC-192 の回線を用いてポートランド折り返し設定を行い，東京ポートランド間を往復，東京から 2 本の OC-48 と GbE4 本で再び太平洋を渡り米国内は Abilene 経由で SC の会場となった Phoenix に到達する．

ディスクデータ転送には，16 台の Data Reservoir システムを用い，スループット 7.56 Gbps 帯域の 92%利用を達成した（**図 14**）．計測時，ネットワークが安定せず，900 秒近辺，および 1,900 秒近辺でネットワークが 1 分程度止まり，パケットロスが発生したが即座に元の通信帯域に復帰している．これはマルチストリーム協調実験のために SC03 で我々が別途行った 32 台対向のパケットロス後のスループット回復に比べても格段に早かった．

5. 応用 — 遠隔講義システム

ここでは Comet DVIP を用いた応用例として高精細度遠隔講義システムを紹介する．遠距離会議システムは普及し，会議だけでなく，インターネット上で

聴講ができる講義も多数存在している．しかしながらそのプレゼンテーションマテリアルはプロジェクタに映すことが前提とされており，会議中あるいは講義中に黒板あるいはホワイトボードに書かれた文字を普通に認識できる実用的かつ安価なシステムは 2002 年当時，存在しなかった．Comet DVIP は映像をターゲットとした IEEE1394 を 2 port，100B-TX 1 port の NIC で，Comet NP を利用した IPsec ESP 処理により，映像のセキュアな 3DES つき高速転送が可能であり，データ量削減が望ましい場合，音声帯域は削らずに，画像帯域のみ削減するといった調整が可能である．我々は Comet DVIP を利用して黒板の字が読める遠隔講義システム Comet Blackboard を実現した．Comet Blackboard システムは，ネットワーク帯域が 10 Mbps から 100 Mbps のレンジを想定し，黒板の文字を認識するための高解像度静止カメラと，講師（話者）映像を撮影するモーションカメラ，マイク，ホストマシンおよび Comet DVIP で構成される．使える帯域が少ない場合，静止カメラによって撮影される黒板画像は，時間方向でデータをまびくことで，レスポンスは悪くても解像度を落とさない一方，講師画像は解像度は落としてもレスポンスはあまり落とさないという方針をとることで，黒板に書かれた文字が読めるシステムを実現した．本システムは実際に 2005 年度に東京大学の柏キャンパスと本郷キャンパスの間で運用された[11]．

6. おわりに

本実験の成果は，2003 年 11 月，東京ポートランド折り返しの予備実験で世界記録を達成し，その後，フェニックスで開催された Supercomputing 2003 High Performance Bandwidth Challenge において，“Distance x Bandwidth Product & Network Technology Award（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．またこの成果を受け，翌年の日本・アメリカ・ヨーロッパをつなぐ WANPHY ネットワーク開通時のテスタとして選ばれ，これが後に Internet2 Land Speed Record の全カテゴリ制覇へとつながった．この記録達成は，商品化のため，チップや NIC の開発にリソースが投入できる企業と，単発的で先の見えない実験的試みに対してある程度の人的リソースが投入可能な大学が協力し，共同研究という枠組みの中で「Comet i-NIC をしゃぶりつくす」ということを一番の目標として掲げ,「採算」および「論文」についてはこの枠組みの中では重要項目とはしなかったことで，初めて可能であったと思われる．実際，データ解析装置や遅延発生装置の開発に際しては,「どこにオリジナリティがあるのか」等はいっさい考えることなく自分たちが使いたいものを追求し,「性能」と「使い勝手」を一番に考え，良い意味で楽しく開発を行うことができた．その一方で，競争の激しい分野でツールを自作できることの価値，すなわち必要な物・機能を欲しいときに比較的安価に入手できるということの大切さが再認識された．意外なことに楽しく行ったツールの開発が，10 Gbps，そして 40 Gbps，100 Gbps の時代を迎えるにあたりハードウェアアプローチを行うためのフレームワークを考える次の研究につながっている[14),15)]．

Comet の特長を活かした開発という点から考えると,「販売元」対「ユーザ」という関係に比べ，共同研究というより枠組みは，連絡が密であり非常に短いサイクルでのポジティブフィードバックが可能であるという利点があったと思われる．隔週 1 度のミーティングによる意見交換は，たとえば適切なメモリサイズの考察等製品開発という側面から考えても，また次の製品企画という点からも有用だったと考えられる．実際，富士通研究所は本共同研究の成果をふまえて，ビジネス適用を目指した ASIC の開発を行い，低消費電力なネットワークプロセッサを実現した[13)]．このネットワークプロセッサを用いて共同研究で開発した Comet TCP を NIC サイズ（7.5 cm × 15 cm）のボードで実現することに成功した．これは類似製品と比較して容積で 1/50 以下，消費電力で 1/20 以下であり，インテリジェント NIC 技術がネットワーク処理装置における大きなブレークスルーであることを実証したといえる．

先端技術の開発には技術的，ビジネス的に大きなリスクがある．企業からみた産学協同研究のメリットは開発すべき先端技術の選択と方向性が「学」の知見によって明確化されること，短い期間で検討実装検証を繰り返すことで正しい方向への軌道修正が行え，結果的にリスクが軽減されることである．特にネットワークシステム分野においては要素技術だけでなく，システム構築が大きな比重を占めるため，技術開発目標があいまいになりやすい．このように多面的なシステム把握が必要な分野において産学共同研究は大きな役割を果たすと考える．反省点としては，商品化を考えたとき，開発およびデシジョンメーキングに時間がかかりすぎることがあることがあげられると思う．たとえば黒板の文字が読める遠隔講義システムについては，企画を始めた 2002 年時点では各人の机上に設置可能な普及版の開発を考えていたが，プロトタイピング作

成に時間がかかったこともあり，2005 年の時点で商品化を断念せざるをえなかった．

謝辞 共同研究全般にあたってネットワークの構成および議論していただいた東京大学の加藤朗氏に感謝します．東京大学の玉造潤史氏，亀沢寛之氏，富士通研究所の下見淳一郎氏，古賀久志氏に感謝します．DR Giga-Analyzer を実際に開発した富士通コンピュータテクノロジーズの中野理氏，鳥居健一氏，吉田昇一氏，柳沢敏孝氏，水口健二氏，生田祐吉氏に感謝します．また富士通研究所の Comet グループ河合純氏，下国治氏，長沼征典氏，的場宏純氏，都築俊秀氏，米国富士通研究所の益岡竜介氏，富士通コンピュータテクノロジーズの来栖竜太郎氏，坂元眞和氏，古川裕希氏，東京大学の西村亮氏，中村誠氏，青嶋奈緒氏に感謝します．

遠距離実験を行うにあたってサポートを行っていただいた WIDE，IEEAF，APAN，JGN，SCinet，SuperSINET 諸機関に感謝します．特に米国接続について Don Reilly 氏および福田健平氏，村上満雄氏，根尾美由紀氏に感謝します．ネットワーク接続について山本成一氏，長谷部克幸氏，田中仁氏，小林克志氏，関谷勇司氏に感謝します．

Foundry，Cisco，Juniper 各社に実験用機材貸し出しおよび現場サポートを感謝します．

データレゼボワールプロジェクトは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」の一環として研究開発が行われ，基板研究 B（2）15300014「アプリケーショントランスペアレントな大域データインテンシブ計算機構」によって補助され，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」の補助により実験を実施しました．

参 考 文 献

1) Hiraki, K., Inaba, M., Tamatsukuri, J., Kurusu, R., Ikuta, Y., Koga, H. and Zinzaki, A.: Data Reservoir: A New Approach to Data-Intensive Scientific Computation, *Proc. ISPAN*, pp.269–274 (2002).
2) Kurusu, R., Sakamoto, M., Ikuta, Y., Hiraki, K., Inaba, M., Ta matsukuri, J., Koga, H. and Zinzaki, A.: Data Reservoir, Multi-Gigabit Data Transfer Facility, Its Design and Implementation, *Proc. PDCAT*, pp.100–108 (2002).
3) 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, *SC2002* (2002). http://www.sc-2002.org/paperpdfs/pap.pap327.pdf
4) 陣崎 明：Stream Processor Comet，並列処理シンポジウム JSPP2000，IPSJ Symposium Series, Vol.2000, No.6, pp.205–212 (2000).
5) 下國 治，河合 純，陣崎 明，山澤昌夫，中村 修，村井 純：Security Network Processor による低消費電力 IPSec ESP の実装と評価，インターネットコンファレンス 2003 論文集，pp.51–58 (2003).
6) Floyd, S.: HighSpeed TCP for Large Congestion Windows, RFC 3649 (2003).
7) Kelly, T.: Scalable TCP: Improving Performance in Highspeed Wide Area Network, *PFLDnet03* (2003).
8) Jin, C., Wei, D. and Low, S.: FAST TCP: Motivation, Architecture, Algorithms, Performance, *IEEE Infocom 2004* (2004).
9) Nakamura, M., Inaba, M. and Hiraki, K.: Fast Ethernet is sometimes faster than Gigabit Ethernet on LFN — Observation of congestion control of TCP streams, *Proc. PDCS*, pp.854–859 (2003).
10) 中野 理，鳥居健一，吉田昇一，柳沢敏孝，水口健二，生田祐吉，陣崎 明，下見淳一郎，玉造潤史，中村 誠，稲葉真理，平木 敬：Data Reservoir—遠距離超高速ファイル転送システム，大域ディペンダブルシンポジウム，pp.199–202 (2004).
11) DVIP 遠隔講義システムスタートアップガイド. http://www.i.u-tokyo.ac.jp/edu/training/ss/enkaku/
12) 下見淳一郎，河合 純，下國 治，陣崎 明，中村 誠，稲葉真理，平木 敬：長距離 TCP 高速化機構の開発，インターネットコンファレンス 2004 論文集，pp.83–91 (2004).
13) 陣崎 明，都筑俊秀，鈴木英好：Gigabit ネットワーク向け低消費電力セキュリティ LSI の開発，組込技術とネットワークに関するワークショップ ETNET2005 (2005).
14) 菅原 豊，稲葉真理，平木 敬：動的再構成を用いたアプリケーションレイヤ処理エンジンの設計，デザインガイア 2005 電子情報通信学会技術研究報告 RECONF2005-59〜71，pp.7–12 (2005)
15) 吉野剛史，玉造潤史，稲上克史，菅原 豊，稲葉真理，平木 敬：ハードウェア・エンジンを用いた 10GbE 上の TCP 通信解析，2006 年並列／分散／協調処理に関する『高知』サマー・ワークショップ，SWoPP 高知 (2006).

(平成 18 年 5 月 15 日受付)
(平成 18 年 9 月 14 日採録)

稲葉　真理（正会員）

東京大学工学部建築学科卒業．武市コンサルティングオフィス，株式会社リコーに勤務した後，東京大学大学院理学系研究科修士課程修了，博士課程中退．理学系研究科助手，講師，情報理工学系研究科特任助教授．博士（理学）．アルゴリズム，ネットワークの研究に従事．

陣崎　　明（正会員）

1954 年 9 月 1 日生まれ．1978 年広島大学工学部電気工学科卒業，1980 年同大学院修士課程修了，同年株式会社富士通研究所入社．以来，一貫して，コンピュータネットワークの高速処理技術およびネットワークプロセッサの研究開発に従事．2006 年より，富士通株式会社先端ソリューション事業部に所属，富士通研究所システム LSI 開発研究所主管研究員を兼務．現在，電子情報通信学会会員．

平木　　敬（正会員）

1976 年東京大学理学部卒業．1982 年東京大学理学系研究科物理学専門課程退学．1986 年理学博士（東京大学）．1995 年より東京大学教授．この間電子技術総合研究所研究員，IBM 社 T.J. Watson 研究センター客員研究員，東京大学助教授等．計算機・ネットワークの高速化の研究に従事．GRAPE-DR project，Data Reservoir project において超高速計算システム，超高速ネットワークに関する研究開発を実施中．