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　菅原豊
情報理工学系研究科コンピュータ科学専攻

概要

ネットワークの性能を引き出すためには、時として、階層型モデルによる非効率さが枷となるケースがある。２００４年度、我々はレイヤー間協調方式を提案し、日本・北米大陸・ヨーロッパ大陸にまたがるネットワーク転送実験を行い、internet2 によるバンド幅距離積の世界記録（Land　Speed Record）を更新した。また超高速文字列照合機構を FPGA の上に実装したネットワークカードを作成し、１０Gbps を超える処理速度を達成、レイヤー間協調方式の有用性を示した。

１．はじめに

光スイッチ技術、WDM 等、近年のめざましい技術革新に伴い、ネットワーク環境が整備され、たとえば、日米間、あるいは、欧米間の通信はOC１９２を日常的に利用できるようになってきている。このような環境でネットワークのぎりぎりの性能を引き出すためには、階層型モデルによる非効率さが枷となるケースが多々見られる。特に高速ネットワークに対応するためにハードウェア化を行う際には、ハードウェアとソフトウエアの切り分けをどう行うかの設計が重要であり、たとえば、階層をまたいだハードウェア実装を行うとで、大幅な性能向上が見られることがある。我々は、これをレイヤー間協調方式（inter-layer co-ordination）と呼ぶことにする。我々は、レイヤー間協調方式として、データ転送速度を改善するための、複数ストリーム協調方式、パケット間ギャップ調整方式、TRC-TCP 方式を、また TCP ストリーム用文字列照合ハードウェアを提案してきた。本稿では、２００４年度に行った仕事のうち、ハードウェア TRC-TCP の実装および日本とヨーロッパ大陸、アメリカ大陸をつなぐ遠距離実験、および、書き換え可能である FPGA を持つ２ポート１０Gbps イーサネットカードと、その FPGA 上に実装された TCP ストリームを認識する超高速文字列照合機構について述べる。

２．TRC-TCP　１０Gbps ハードウェア

ネットワーク通信で標準的に使われる TCP は、データの信頼性のため、送信済み受信未確認のデータ（“ inflight data ”）をバッファリングしており、この TCP ウィンドウと呼ばれるバッファサイズの調整により動的にスピードの調整を行っているが、遠距離高速通信では、性能が十分に出ないことが知られている。２００３年度、我々は、TCP のウィンドウサイズによって定まる速度に、ネットワークインターフェースのパケット送出速度を合致させる TRC-TCP 方式で、性能が著しく改善されることを１Gbps のイーサネットを終端とする、約８Gbps の日米間の実験線を用いて実証した。本年度は急速に広まりつつある１０G bps のネットワークに対応するため、すなわち、１０G イーサネットでスタンダードフレームを利用する場合パケット送出間隔が約１．３μ秒、に対応するため、ハードウェア TRC-TCP を既存のネットワークカード上に実装し、日米欧を

またぐ１０Gbps のネットワークの上で実験を行った。

今回、我々は一般的に入手可能な標準的部品を組み合わせたサーバと、インテリジェント１０G bps イーサネットカード Chelsio T110（図１）を組み合わせることでシステムを構成した。この場合のボトルネックは、ネットワークではなく、サーバの I/O バスとなる。ボトルネックスピードは、PCI-X 1.0 64bit-133MHz　で、およそ 8.5Gbps となる。

図１. Chelsio T110

Chelsio　T110 は 512MB パケットバッファを持ちフル TCP オフロード機能を持つため、サーバの CPU 負荷が少ないとう特徴を持ち、データパスは ASIC　制御コードは FPGA という、プログラマブルなインテリジェントイーサネットカードで、近距離通信では７．５G bps を達成する。我々は T110 のバッファのキューイングをシンプルにすることで、シングルストリームによるデータ転送時に再高速が達成できるように最適化したうえで、TCP によって定まる速度とイーサネットパケット送出速度の調整を行った。

３．遠距離データ転送実験

実験は Rioworks HDAMA rev.E のマザーボードに、２．２GHｚ　AMD Opteron248、1GB PC3200 DDR SDRAM を搭載したサーバに Linux 2.6.6 for x86_64　を使用、iperf v1.7.0 を使って、二点間でデータ転送を行った。MTU 1500B のスタンダードと 7832B および 8192B　のジャンボフレームでの実験を行った。バッファサイズは、RTT　250msec のとき、384MB、約 500msec のとき、448MB とした。また複数ストリームを用い ディスク上のファイルデータ転送実験も行った。実験は（１）東京から北米経由 CERN（２）CERN から東京経由ピッツバーグの実際のネットワークを用いて行った。ここでは（２）CERN―ピッツバーグの実験結果を中心に実験結果を述べる。

３．１　東京 – 北米 – CERN

2004 年 10 月 ― 日本、カナダ、米国、オランダと CERN 研究所（スイス・ジュネーブ）の研究者が共同して、日本からスイスまでの世界最長の１０ギガビットイーサネット回線を構築した。　ネットワークの全長は約 18,500Km、17 のタイムゾーンを通過しており、10Gigabit イーサネット WAN PHY 技術を用い光を用いた SONET/SDH 機器の相互運用により、東京大学から CERN 研究所のサーバを恰も同一 LAN 上であるかのように接続を行った。2004 年度最初の遠距離データ転送実験は、この全経路 WANPHY ネットワークを利用して行われた。2004 年度最初の遠距離データ転送実験は、この WANPHY ネットワークを利用して行われ、スタンダードフレームを用いて平均 7.57Gbps のデータ転送を達成した。

３．２　CERN―東京―ピッツバーグ

2004 年 11 月　東京大学と WIDE プロジェクトの研究者が日本、カナダ、米国、オランダと CERN 研究所（スイス・ジュネーブ）の研究者との協力により、米国ピッツバーグから日本を経由してスイス・ジュネーブまでの世界最長の１０ギガビット回線を構築し、ピッツバーグの SC2004 国際会議会場に設置したシステムから CERN 研究所に設置したシステム間

を接続した。ネットワークの全長は約 31,248 Km あり、17 のタイムゾーンを通過している。ピッツバーグの SC2004 国際会議会場に設置された東京大学の計算機から、SCinet を経由し、Abilene, StarLight を経由し APAN/JGN2 ネットワークで東京に到達し、WIDE Project が運用する T-LEX に接続し、T-LEX から米国シアトルまでは TycoTelecommunication が IEEAF に寄贈した回線を用い、Pacific Northwest- Gigapop の機器に接続した。シアトルからは、回線は CA*net4 の専用ラムダによってシカゴの StarLight まで運ばれ SURFnet のシカゴーアムステルダム間のラムダに接続され、アムステルダムの NetherLight に接続された。NetherLight と CERN 研究所の間は、SURFnet のアムステルダムージュネーブのラムダによって接続されている。ネットワークの往復遅延時間(RTT)は約 437 ミリ秒である。(図２)

図２　ネットワーク構成図

実験では、まず一本の TCP ストリームを用いて、データ転送を行い 7.21 ギガビット／秒のデータ転送を実現した。データ転送には 1500 バイトの標準イーサネットフレーム長を用い、データ転送は 20 分間行った。図３に　シングルストリームのデータ転送レートを示す。 定常状態においては、PCI-X の理論的限界値の９５％程度が継続的に達成されており、TRC-TCP が有効であることがわかる。このバンド幅・距離積は 225,298 テラビット メートル／秒となる。

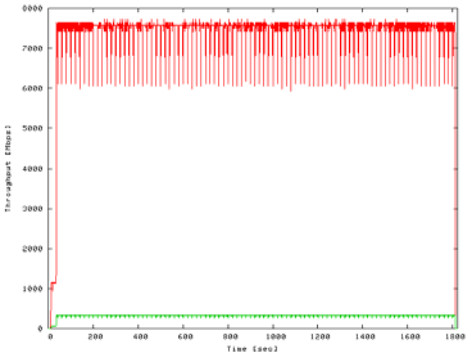

図３　単一ストリーム転送レート

次にマルチストリームによる自己競合のあるシステムにおける、ディスク上のデータファイルの転送実験で、TRC-TCP 方式の有無による比較を示す。

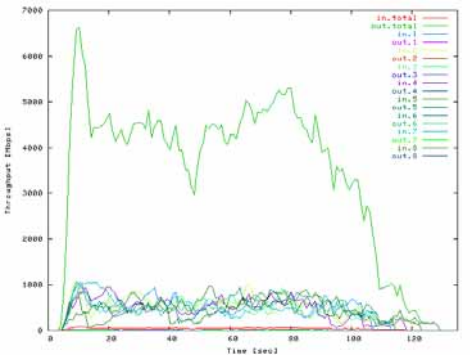

図４　自己競合 TRC-TCP　無し

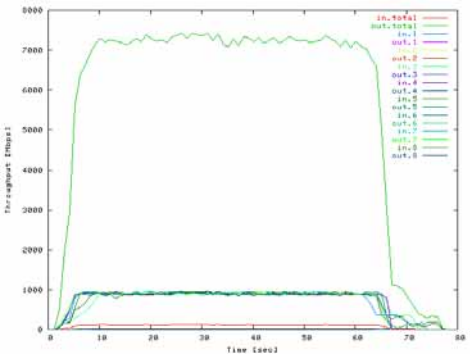

図５　自己競合 TRC-TCP　有り

マルチストリームを用いたシステムにおいては、自己競合による性能低下が問題になることが知られている。図４、図５は、TRC-TCP の無・有による、ストリームごと、そしてシステム全体のパフォーマンスを示している。

TRC-TCP により、バースト性を取り除くことで、著しい安定化およびその結果の性能向上が見られる。ここでは、８台構成のシステムで、1.6Gbps のディスク間データ転送を達成した。

４．プログラマブル１０Gbps　イーサカード

プログラマブルな FPGA を持ち複数の１０Gbps イーサネットポートを持つネットワークカードを開発し（図６）FPGA の上に（１）TCP ストリームを認識する超高速文字列照合機構および(２)10Gps パケットヘッダロガーを実装した。ここでは文字列照合について述べる。

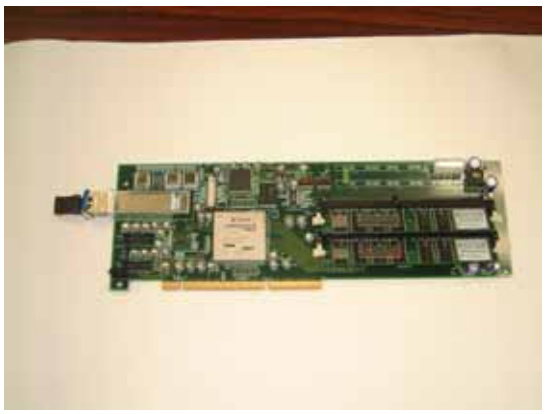

図６．　プログラマブルネットワークカード

近年、Ｗｅｂデータのデータマイニングあるいはネットワークセキュリティーのために、高速文字列照合の要求が高まっている。特にセキュリティーに関しては、照合パターンを必要に応じて変更する必要がある。またエンドノードでの煩雑な処理を無くすためにネットワークの入口でフィルタリングをかければ良いが、そのためには、中間ノードで TCP ストリームごとの文字列照合をワイヤーレートで行う必要がある。我々はＦＰＧＡ向きの文字列照合アルゴリズムＳＢＴ法を提案、実装し、合計 1000 文字のルールが１０Ｇｂｐｓで処理できることを確認した。

５．おわりに

ネットワーク高速化に従い、ハードウェア化の重要性は、日々増している。　一方、技術の進歩に従いユーザ要求は日々変化することも多く、プログラマブルな FPGA と ASIC の組み合わせによるシステムの重要性は今後、ますます増大すると考えられる。

６．謝辞

ネットワーク実験は WIDE プロジェクト、CERN, SURFnet, StarLight, Neitherlight, Tyco Telecommunications, IEEAF, Pacific NorthwestGigapop,CA*net4networks,APAN 富士通コンピュータテクノロジーズ、ＮＴＴコミュニケーションズ、東陽テクニカ、Foundry networks, Cisco Systems, Juniper Networks, Clear Sight 社の協力のもとに行われた。

参考文献

[1] K. Hiraki, M. Inaba, J. Tamatsukuri, R. Kurus, Y. Ikuta, H. Koga,　and A. Jinzaki, “Data Reservoir: Utilization of Multi-Gigabit Backbone Network for Data-Intensive Research,” Proc. Super Computing 2002,　(SC2002) CD-ROM , Nov., 2002.

[2] M. Nakamura, M. Inaba, and K. Hiraki, “Fast Ethernet is sometimes faster than Gigabit Ethernet on LFN--Observation of congestion control of TCP streams” Proc. Int. Conf. on Parallel and Distributed Computing And Systems (PDCS2003) Nov. 2003　pp.854-859

[3] K. Hiraki, M. Nakamura, J. Senbon, Y. Sugawara, T. Itoh and M. Inaba, “End-node transmission rate control kind to intermediate routers - towards 10Gbps era”　(PFLDnet 2004) Feb. 2004

[4] T.Ito, M. Inaba,　“Theoretical Analysis of Performances of TCP/IP Congestion Control Algorithm with Different Distances”, Networking 2004　May 2004

[5] M. Nakamura, H. Kamezawa, J. Tamatsukuri, M. Inaba, K. Hiraki, K.　Mizuguchi, K. Torii, S. Nakano, S. Yoshita, R. Kurusu, M. Sakamoto, Y. Furukawa, T. Yanagisawa, Y. Ikuta, J. Shitami, A. Zinzaki“Long Fat Pipe Congestion Control for Multi-Stream Data Transfer”Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks, ISPAN 2004

[6] Y. Sugawara, M. Inaba, and Kei Hiraki, “Over 10Gbps String Matching Mechanism for Multi-Stream Packet Scanning Systems”, Field-Programmable Logic and Applications, 14th International Conference, FPL 2004

[7] H. Kamezawa, M. Nakamura, J. Tamatsukuri, N. Aoshima, M. Inaba, K. Hiraki, J. Shitami, A. Jinzaki, R. Kurusu, M. Sakamoto, and Y. Ikuta, “Inter-layer coordination for parallel TCP streams on Long Fat pipe Networks”, Super Computing 2004,　High Performance Networking and Computing, SC2004