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

遠距離データ通信においてネットワークの性能を限界まで引き出すためには経路上の帯域幅の変化、特に帯域幅が減少するポイントでのパケットロスをいかに起こさせないかが一つの大きなキーポイントとなっている。

2005 年度，我々はハードウェアプログラマブルなネットワークツール ミドルハードウェアを提案，10G bps イーサネット２ポートを持ちブリッジとして利用できる MH-BOX TGNLE-1 を作成した。TGNLE-1 により我々が 2004 年度に提案したレイヤー間協調方式の適応範囲がホストマシン単体ではなくクラスタに拡がった. またアドレスマッチングの問題で，IPv4 限定だった実装を IPv6 対応とすることも可能となった．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 IPv4, IPv6, Single stream, Multiple streams の全種目について世界記録（Land Speed Record）を更新した.

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．10Gbps Ethernet の普及とあわせ遠距離通信において回線速度と エンドノードの PCI-X・X2, PCI-express 等の内部バスの速度がほぼ同じオーダーとなってきている. 我々は遠距離通信に際しての階層化によるオーバーヘッドを軽減するためのレイヤー間協調（inter-layer co-ordination）を提案し，複数ストリーム協調，パケット間ギャップ調整，TRC-TCP を提案・実装し実ネットワークでの実験を行い遠距離高速データ転送を実現してきた。特にハードウェア TRC-TCP は エンドノードの CPU では不可能なパケットレベルでの細粒度のタイミングの調整を実現し超高速データ転送におけるハードウェアサポートの重要性を示唆した．ハードウェア TRC-TCP は，市販の インテリジェント NIC Chelsio T110 を採用したがこの T110 の FPGA はバッファ管理やクロック管理は行えるが, データパスはチップが制御しているため可能なパケット処理には限界があった．2005 年度，我々はネットワークストリームをワイヤーレートで取り扱うためネットワークファンクションを FPGA のファームウェア作成により実現するためのプログラマブルハードウェアの枠組、ミドルハードウエアアプローチを提案し 2 ポートの 10Gbps イーサネットインターフェースと FPGA を持つミドルハードウェアボックス(MH-Box) TGNLE を実装し，FPGA 用のファームウェアを開発することで TRC-TCP，ストリームハーモナイザ，暗号化複合化，パケットフラッダ，パケットロガーを実現した．また 2004 年度に引き続き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を行い，Chelsio T110 の PCI-X2 対応である T210 お

よび TGNLE に実装したハードウェア TRC-TCP を用いることで Internet2 Land Speed Record を全種目制覇した．本稿では TGNLE の実装について述べ， TGNLE-1 上に実装したハードウェア TRC-TCP の遠距離実験を TGNLE-2 上に実装されたパケットロガーを利用して解析した結果について記す．

２．MH-Box　TGNLE

ミドルハードウェアボックス(MH-Box)はネットワークストリームをワイヤーレートで取り扱うためのプログラマブルな枠組みであり，FPGA のファームウェア開発によりをヘッダ採取，速度調整，フィルタリングなどの機能を実現する．MH-Box はエンドノードとしてだけでなく，Proxy として，ネットワークブリッジとして TCP ストリームを中間地点で処理することを想定している．たとえば 10Gbps ワイヤーレートはスタンダードフレームを使った場合，ほぼ　1.25μ秒に一個パケットが送出されるため，中間地点で TCP ストリームを処理する場合，この 1.25μ秒間隔で TCP コンテクストを切り替え可能である必要があるためコンテクスト情報は FPGA のメモリにのせることが必要である．またヘッダ情報収集やパケットコントロールを行うためのデータバッファが必要となる。

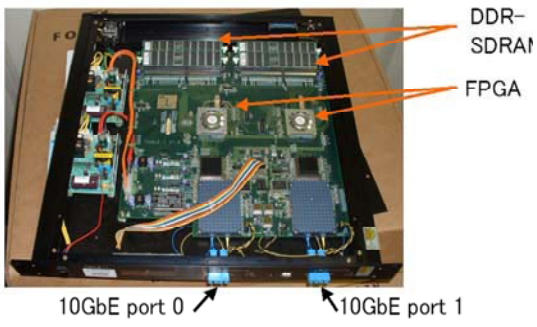

図１　TGNLE 写真

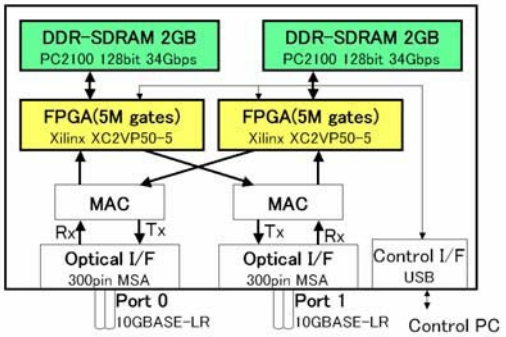

図２　TGNLE ブロック図

図１に今回実装した MH-Box TGNLE の写真を、図２にブロック図を示す．TGNLE は 10Gbps イーサネット LR 光インターフェースを 2 ポート（port0, port1）持ち port0 RX から port1 TX と　port1RX から port0 TX へという二つのデータパスを持つ．

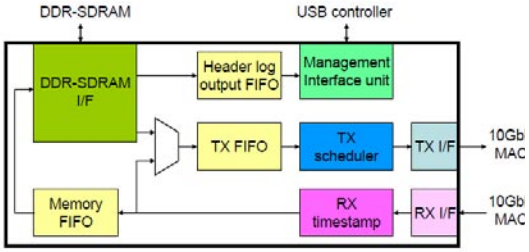

図３　データパス

図３にデータパスを示す．それぞれのパスはコントロール用の５００万ゲートの FPGA (Xilinx XC2VP50-5) および ２ギガバイト DDR-SDRAM メモリ（PC2100,128bits, 34Gbps)を持つ．Rx と FPGA の間にはメモリバッファがないため Rx から入った信号が FPGA に届くまでは一定時間となる．また　コントロール用に双方の FPGA に接続する USB を全体で 1 ポートを持つ．

TGNLE 用ファームウェアとしては　(1)パケットペーサー，(2)ストリームハーモナイザー，(3) 擬似遅延環境，(4) ネットワークプッ

シャー(5)パケットロガー, (6)ストリングマッチャー, (7)エンクリプションデクリプションを提案・実装したが紙面の制限によりここでの説明は割愛する.

３．データ転送実験

３．１　予備実験

予備実験を、Opteron サーバ dual Opteron 248 2.2GHz Rioworks HDAMA マザーボード PCI-X, DDR3200 CL2 2GB(メインＣＰＵのメモリスロットに 512MB メモリ４枚) Linux 2.6.12 ネットワークカード Chelsio T110 cxgbtoe ver 2.1.1　環境で行った. RTT を 400ms に固定した 10Gbps イーサネット擬似遠隔環境で Iperf 2.0.2 を使ったメモリからメモリへのデータ転送を行った. メモリメモリ転送の場合は PCI-X が通信ボトルネックになるため, ネットワークバッファを受信側にいれると効果的であることが確認された.

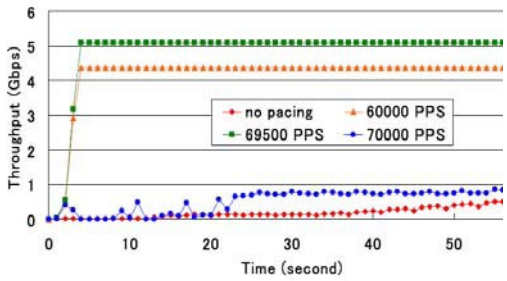

図４ 受信側ペーシング効果

図４にハードウエア TRC-TCP による秒あたりのパケット数(PPS)設定によるスループットの変化を示す. 適正なパケット数 69500 は２分探索法で求めた. パケット数が最適値を１％下回ると若干の性能低下が見られるのに対しパケット数が最適値を１％上回ると性能は大幅に低下することがわかる.

３．２　30,000km データ転送実験

2005 年 10 月から 2 月にかけて　東京大学と WIDE プロジェクトの研究者が日本、カナダ、米国、オランダの研究者との協力により、東京・シカゴ・アムステルダム・シアトル・東京の全長約 32,372 Km 世界一周線を構築した.（図５）

図5　ネットワーク構成図

このサーキットの上で様々な実験を行い, 2005 年１０月から 2006 年２月にかけて, 下記代表的な３件を含む計６回の世界記録更新を行った.

(A) 11 月 10 日 IBM X366 サーバ Interl Xeon MP 4 個 メモリ PC3200 DDR2(送信側 8GB, 受信側 32GB), PCI-X2 Netrion Xframe II ネットワークインターフェース, Windows Server 2003 SP1 x64(version 5.2 build 1830) 環境で 9014 バイトジャンボフレームを用い NTttcp を用いたメモリメモリ転送を行い 7.99Gbps (TCP payload) を達成 239,820 Terabit meter/sec を達成した. これは IPv4 の Single and Multiple stream の記録更新となったが、OS に Windows を使いまたネットワークインターフェースに非インテリジェントカードを使ったところにこの実験の特徴がある。

(B) Opteron サーバ dual Opteron 248 2.2GHz Rioworks HDAMA マザーボード PCI-X, メモリ 1 GB Corsair Twinx Linux 2.6.14 ネットワークカード Chelsio N110(10GBase-SR) PCI-X/133 MHz サーバ.

9130Byte ジャンボフレーム、IPv6 packets 6.96 Gbps (TCP payload), 208,800Terabit meter / secを達成した．また アドレス処理で不利となるといわれている IPv6 でも充分実用的な効率よいデータ転送が可能であることが示しされ.

(C)2 月２０日 IBM X366 サーバ Interl Xeon MP 4 個 メモリ PC3200 DDR2 32GB, PCI-X2 Chelsio T310-X インテリジェントネットワークインターフェース,iperfメモリメモリ転送を行い 8.8049Gbps (TCP payload)を達成 239,820 Terabit meter/sec を達成した(zu６)。この記録は、internet2 の LSR のレギュレーション 10% 記録更新が必要とてらしあわせると 10Gbps WANPHY では更新不能な値であるため、おそらく 40Gbps ネットワークを利用しないと記録更新ができないと予測されている.

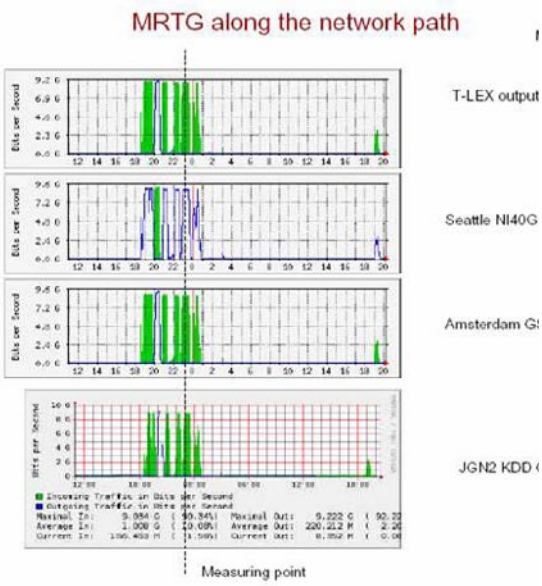

図６ ＭＲＴＧグラフ

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

東京大学の加藤朗氏、三精システムの藤代康之氏に特別に感謝する．Chelsio 社および米国 Microsoft 社の協力に感謝する．ネットワーク実験は WIDE プロジェクト，Pacific Northwest Gigapop, SURFnet, StarLight, Neitherlight, Tyco Telecommunicati4 networks, 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, Clear Sight 社の協力のもとに行われた。

参考文献

[1] 中村誠,菅原豊,玉造潤史,稲葉真理,平木敬, “擬似ネットワーク環境における TCP/IP の性能と評価”, IA 研究会, 信学技報, vol. 105, no. 219, IA2005-7

[2] 菅原 豊, 稲葉 真理, 平木 敬, “細粒度パケット間隔制御の実装と評価”, 2005 年並列／分散／協調処理に関する『武雄』サマー・ワークショップ（SWoPP 武雄 2005）

[3] 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]

[4] Makoto Nakamura, Ryutaro Kurus, Felix Marti, Masakazu Sakamoto, yukichi Ikuta, Junji Tamatsukuri, Yutaka Sugawara, Nao Aoshima, Mary Inaba, and KEI HIRAKI, “Experimental Results of inter-layer cooperative hardware for TRC-TCP on 10Gbps Ethernet WANPHY 18,000km Network”, Third International Workshop on Protocols for Fast Long-Distance Networks PFLDnet 2005

[5] Yutaka Sugawara, Mary Inaba, and Kei Hiraki,, “High-speed and Memory Efficient TCP Stream Scanning using FPGA”, 15th International Conference on Field Programmable Logic and Applications, FPL2005

[6] Junji Tamatsukuri, Katsushi Inagai, Mary Inaba, and Kei Hiraki, “Experimental Results of TCP/IP data transfer on 10Gbps IPv6 Network”, Fourth International Workshop on Protocols for Fast Long-Distance Networks, PFLDnet 2006

[7] インターネット２ランドスピードレコード http://lsr.internet2.edu