2005－ＱＡＩ－16（１）
2005／7／29

擬似ネットワーク環境における TCP/IP の性能評価

中村 誠
東京大学情報基盤センター
〒113-8658 東京都文京区弥生 2-11-16
makoto@is.s.u-tokyo.ac.jp

玉造 潤史
東京大学大学院理学系研究科
〒113-0033 東京都文京区本郷 7-3-1
junji@is.s.u-tokyo.ac.jp

菅原 豊　稲葉 真理　平木 敬
東京大学大学院情報理工学系研究科
〒113-0033 東京都文京区本郷 7-3-1
{sugawara,mary, hiraki}@is.s.u-tokyo.ac.jp

概要　ネットワークに大きな遅延を作り出す遅延装置を用い、Linux PC での RTT の長い高速通信実験を行った。遅延だけが存在する環境では十分なウインドウサイズ、CPU リソースが利用できれば、ネットワークカードの持つ速度性能によりソフトウエアによる TCP/IP でも 7.2Gbps 以上の通信が可能である。

キーワード　高レイテンシ TCP/IP 通信,10 ギガビットイーサネット

Performance evaluation of TCP/IP in pseudo network enviornment

Makoto Nakamura
Information Technology Center,
the University of Tokyo
2-11-16 Yayoi, Bunkyo, Tokyo 113-8658, Japan

Junji Tamatsukuri
Graduate School of Science,
the University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan

Yutaka Sugawara　Mary Inaba　Kei Hiraki
Graduate School of Information Science and Technology,
the University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan

Abstract　We had a long RTT TCP/IP communication experiment on pseudo network by a large latency production equipment. Linux PC with large memory for TCP window and enough CPU resources for packet processing enables over 7Gbit/s communication by current 10Gbit Ethernet Adapters.

Key Words　Long Latency TCP/IP communication, 10Gbit Ethernet

1　はじめに

10 ギガビットイーサネット規格が 2002 年に標準化されて以来、バックボーンネットワークはルータ間の WANPHY 接続に、スイッチ間接続は LAN-PHY にと徐々に統一的な移行が進んで来ており、ルータスイッチともフルワイヤレートでの通信が主流となってきた。

一方、末端の計算機での 10 ギガビットイーサネットアダプタは現在 Intel,S2io(Neterion), Chelsio から入手する事が出来るが、これらのアダプタはいずれも PCI-X 1.0 バスに接続されるため、バンド幅はバス帯域 8.5Gbps 以下に制限される。さらに CPU の処理性能が十分ではないため Intel の 32 ビット CPU Xeon を使用したシステムでは 16KByte のジャンボフレームを使用しても 6Gbps 以上のバンド幅を出せず、1.5KB のフレームでは 3Gbps 以上のバンド幅を出すことも出来ない [2, 3]。

このような 10 ギガビットイーサネットアダプタをもった計算機に求められる能力の一つは今まで以上に高速に地球上に分散したデータをネットワーク上でやり取りすることである。複数の大陸を跨るよ

うな地点間では、数 100ms の往復通信遅延 (RTT) が避けられない。500ms の RTT がある場合、10Gbps で通信するためには TCP では 625MB のウィンドウが必要となる。従来このような大量のバッファ領域を持続的に高速で消費、供給する事は稀であったため、OS を含めたソフトウェアの負荷を十分に考慮しなければならない。

イーサネットは通信速度が向上してもパケットサイズは不変のため、夥しい数のパケットを短時間に処理しなければならない。データパケットが 1500Byte、通信速度が 10Gbps では 1 秒間に 80 万以上のパケットを送受信される。TCP の場合、さらにデータの到着を通知するためにそれと同数から数分の 1 の応答 (ACK) パケット (サイズは 64 Byte 前後) を送受信する。

このような問題に対し、Linux は受信動作で割り込みが大量に発生し CPU 資源を消費し尽す事を避けるために、割り込み発生時に一時的に割り込みを禁止し、アダプタ上のバッファにデータが存在する限りポーリング動作に切り替えて受信処理をする機構 (NAPI) を実装している。また、ギガビットイーサネットの頃から一部のネットワークアダプタはチェックサム計算やデータのセグメント化処理をアダプタ上で行いホスト CPU の負荷を軽減している。さらに、割り込みを静的または適用的に間引いたり、遅延させ回数を削減する (Interrupt Coalescing) 機能を有している。

10 ギガビットイーサネットアダプタでも Intel や S2io も同様の手法を採用し、Intel/S2io では TCP パケットのセグメント処理をアダプタ上で行う TSO(TPC Segment Offloading) の機能を持つ。さらに、Chelsio は TCP/IP 層の処理をすべてアダプタ上のコントローラが行う機能 (TCP Offloaing Engine, TOE) をも有している。実際に我々が行った東京-アムステルダム間往復 (RTT 500ms) の実回線上での通信では、Chelsio T110 のファームウエアに変更を行ったものを用い 1.5KB フレームで 7.21Gbps のバンド幅を得られた [1][4]。ただし Chelsio の TOE が扱うのは IPv4 のみで IPv6 はソフトウェアで処理する必要がある。

従って、ソフトウエアによって RTT の長い TCP/IP 通信を行う場合にはホスト PC の処理能力、OS、ネットワークアダプタを有効に用いる必要があり、未知の問題が存在している。本稿では、長距離の実回線通信実験の予備実験として RTT の長いネットワークで TCP 通信により持続的にデータ転送した際のシステムの振る舞いを解析する。そのため、遅延発生器を用いた擬似広域ネットワーク環境下で、AMD 64 ビット CPU Opteron を使用したシステムで Chelsio/Intel 10 ギガビットイーサネットアダプタのメモリデータのバルク転送性能を評価した。

2 擬似ネットワーク環境での通信実験

遅延装置による擬似ネットワーク環境を用いて、10 ギガビットイーサネットでの TCP/IP によるデータ転送実験を行った。

2.1 実験の構成

遅延発生器として Anue 社製 H シリーズネットワークエミュレータを使用した。これは、10G イーサネットの LANPHY(10.312 Gbps) および WAN-PHY(9.953 Gbps) でのフルワイヤーレートをサポートし、片方向につき 800ms までの遅延を発生できる。

実験に用いた機器は、CPU が Opteron 248 2.2GHz の Dual 構成、マザーボードが Rioworks HDAMA、メモリが DDR3200 CL2 2GB(512MB モジュールを 4 枚でメイン CPU のメモリスロットに実装した) という構成の PC である。

ソフトウェアは OS に Linux 2.6.6 を、ドライバは T110 に cxgbtoe ver.2.1.1 を (ただし TOE 使用時は chtoe-t1 ver.1.1.4)、N110 に cxgb ver.2.1.1、Intel PRO/10GbE に ixgb ver.1.0.110 を用いた。アプリケーションは Iperf 2.0.2 を使用した。

図 1 のように富士通 XG1200、Foundry MG8 を介し、遅延装置を経由して通信するよう端末を接続した。ネットワーク機器は遅延装置を挟んで異なる VLAN を構成し、フォワーディングは L2 で動作させた。そのため、すべての端末は同一の L2 VLAN 上に存在している。また、ワイヤー上での転送速度を測定するために MG8 の SNMP 統計情報を用いた。

通信は スタンダードフレーム (1500Byte) とジャンボフレーム (9198Byte,Foundry MG8 の最大 MTU) それぞれで行い転送性能、CPU リソース、メモリリソースを計測する。

本実験でターゲットとしたネットワークアダプタは、以下の 3 種類である。

1. Chelsio T110 Protocol Engine
   - TCP/IP 通信を TOE を用いた場合とソフトウエアによる場合の両方を測定した。TOE の場合、最大性能を測定するため Jumbo Frame は 7832Byte とした。
2. Chelsio N110 Server Adapter

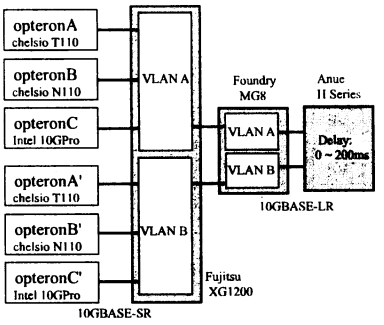

図 1　実験構成図

- T110,Intel PRO/10GbE のような TCP/IP を支援するハードウエアを持たない。

3. Intel PRO/10GbE SR Server Adapter(以降 IXGB と示す)
   - TCP/IP 通信の CPU 負荷を削減する TSO の機能、割り込み負荷を削減する NAPI を利用した。

これらのネットワークアダプタを持った PC を対向で通信を行い、ネットワーク設定のパラメータ (Window Size,TX Queue,Buffer Size) により最適な通信速度を計測した。

高レイテンシ通信の比較元となるのは、Chelsio T110 による TOE を用いた高速通信である。(表 1)

MTU/RTT(ms)	0	100	200	300	400
1500B	7.53	7.21	7.21	7.20	7.20
7832B	7.53	7.52	7.54	7.52	7.54
Window Size(MB)	1	94	190	285	375

表 1　Chelsio T110 ToE 時の性能 (Gbps)

この結果は実際のネットワークで計測した場合の通信速度と同等であり、擬似ネットワーク環境が高レイテンシ通信実験に用いることができると言える。

2.2　ハードウエアによる TCP/IP 通信

T110 の TOE は、TCP 層のコントロールを全てアダプタのハードウエア上で行う。そのため、システムはウインドウに十分なデータを書くことができれば TCP のパケット処理に CPU リソースを使わない。TOE 時には性能が制限されることをさけるため Nagle アルゴリズムを無効にした。

MTU が 1500Byte の場合、転送開始数秒後から、性能が 1-2Gbps で停滞する期間がある。(図 2) この時、XG1200 のカウンタを観測すると送信側のアダプタから PAUSE フレームが継続的に送出されおり、ACK パケットの受信動作に問題があると考えられる。20 秒後から、この現象が解消し転送速度が上昇する。図 3,4 は各々送信側、受信側での TOE 使用時の CPU 使用率と割り込み回数のグラフだが、7.2Gbps のバンド幅が出る期間の受信側の割り込み回数が毎秒約 16,000 回なのに対し、1-2Gbps で停滞する期間の割り込み回数が毎秒約 20,000 回と多い。CPU 使用率は 7.2Gbps のバンド幅が出る時には送信側で約 85%(idle が約 15%)、受信側で約 60%(idle が約 40%) でソフトウェアによる TCP の場合と同様送信側の負荷が大きい。

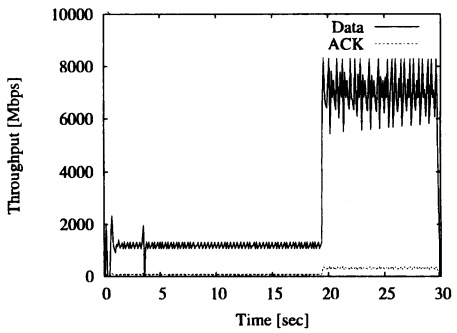

図 2　TOE 使用時の転送速度 (MTU=1500B, RTT=400ms)

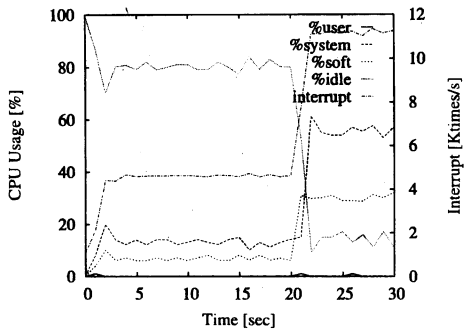

図 3　TOE 使用時の CPU 使用率と割り込み回数 (送信側, MTU=1500B, RTT=400ms)

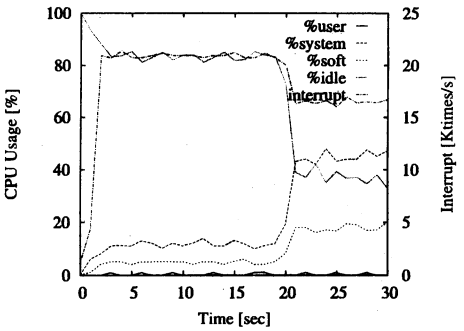

図 4 TOE 使用時の CPU 使用率と割り込み回数 (受信側, MTU=1500B, RTT=400ms)

2.3 ソフトウエアによる TCP/IP

図 5, 表 2 が遅延がない状態での各ネットワークアダプタの通信性能である。スタンダードフレームで 3Gbps、ジャンボフレームで 6 から 7.2Gbps で通信した。

図 6, 表 3, 図 7 が遅延がある場合の性能である。ジャンボフレームでは RTT=200ms,RTT=400ms ともに遅延がない場合との性能差は小さい。共通してハードウエアの場合と比べ大きなウインドウサイズ (MEM) を必要としており、理論値との比率が約 3 倍以上で最大性能が得られている。

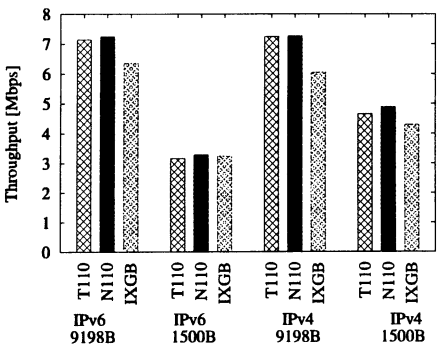

図 5 ピークバンド幅 (RTT=0ms)

2.3.1 Chelsio T110/N110

これら 2 種類のネットワークアダプタは非常に似た振る舞いをする。どちらもソフトウエア TCP/IP 通信で遅延がある場合にスタンダードフレームは安定した動作をしなかった。ジャンボフレームでは

IP	MTU (B)	Card	Rate (Gbps)	MEM (MB)
V6	1500	T110	3.16	1
V6	1500	N110	3.28	1
V6	1500	IXGB	3.23	10
V4	9198	T110	7.24	1
V4	9198	N110	7.26	1
V4	9198	IXGB	6.05	10
V6	9198	T110	7.14	1
V6	9198	N110	7.23	1
V6	9198	IXGB	6.35	10

表 2 転送性能 (RTT=0ms)

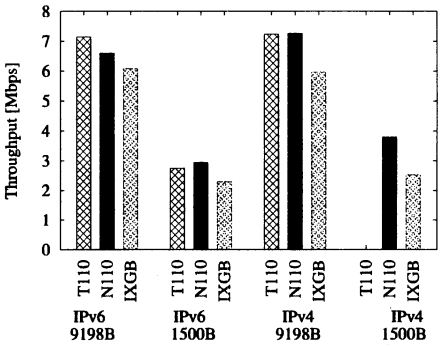

図 6 ピークバンド幅 (RTT=200ms)

IP	MTU (B)	Card	Rate (Gbps)	MEM (MB)	MEM 理論値	MEM 比率
V6	1500	T110	1.80	960	90	10.67
V6	1500	N110	1.80	896	90	9.96
V6	1500	IXGB	1.98	896	99	9.05
V4	9198	T110	7.20	960	360	2.67
V4	9198	N110	7.08	896	354	2.53
V4	9198	IXGB	5.98	896	299	3.00
V6	9198	T110	7.07	960	354	2.72
V6	9198	N110	6.60	896	330	2.72
V6	9198	IXGB	6.30	896	315	2.84

表 3 転送性能 (RTT=400ms)

安定して TOE 時と同等の性能を示した。カードの応答性能がよく、割り込みがシステムを過負荷にしやすいためネットワークアダプタのドライバで割り込みは 50us ごとに設定 (最大 20000 回) してあり、転送時には約 17000 から 20000 回の割り込みが発生する。割り込み処理の軽減のため待ち合わせ時間を小さくしても性能には貢献せず、受信時の coalescing である adaptive-rx に効果はなかった 。

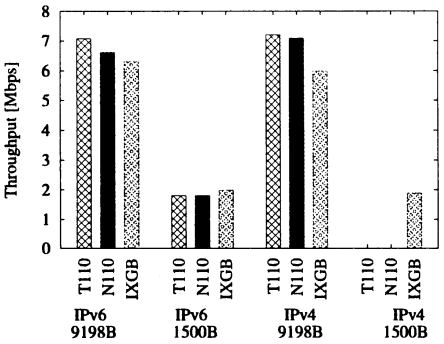

図 7　ピークバンド幅 (RTT=400ms)

RTT	V4 0	V4 400	V6 0	V6 400
T110	16837	16517	17320	16510
N110	17042	15561	17128	15363
IXGB	5236	3710	9169	16068

表 4　各ネットワークアダプタの割り込み平均発生回数 (回数/秒)

N110 では SMP 環境では irqbalance 機能 (割り込みを複数の CPU に振り分ける機能) が有効な場合、再送処理が発生してしまう場合がある。ドライバの実装中で並列処理の制御に問題があると考えられる。

2.3.2　Intel PRO/10GbE

Chelsio よりも最大性能は低く転送能力は 6.4Gbps 程度である。NAPI の効果により、Chelsio よりも少ない割り込みしか発生しない。(表 4) Chelsio と同条件の実験であるため、最大性能を決める要因は PCI-X バスではなくアダプタ本体にあると推測される。TSO を使用すると約 10%CPU 使用率が軽減するが、転送速度は向上しなかった。また、TSO を使用し、遅延がある場合には再送処理が発生し、さらに再送処理が 1 時間以上続き強制終了せさるを得ない場合が多い。そのため、遅延環境では TSO なしで計測を行った。

2.3.3　性能評価

遅延環境で Linux の TCP スタックを使用する場合には、輻輳ウインドウを非常に大きくする必要がある。輻輳制御には、*RTT* × 通信速度 のウインドウが必要である。しかし、表 3 あるようにピーク性能で持続的にデータ転送するためには、理論値であるバンド幅遅延積の少なくとも 2.5 倍以上の大きさのウインドウサイズ指定する必要がある。400ms の遅延環境の場合、通信の開始からの 5 秒間に TCP の slow start アルゴリズムにより急激にウインドウサイズが増加する。アプリケーションプログラムはこの間用意されているウインドウを埋め尽くすようにデータの送出を行う。この時、増加するウインドウサイズに合わせて、ネットワークへのデータ送出も増加し、メモリから TCP/IP スタックに渡される。小さなウインドウサイズでは、このデータ供給が間に合わず、ネットワークへの送出データがウインドウサイズを満たすことができないと推測される。このため、ネットワークアダプタへの十分な送出データをウインドウメモリ上に持つことで増加するネットワークへのデータ送出量を確保し、ネットワークアダプタの十分な通信能力を発揮することができる。(図 8)

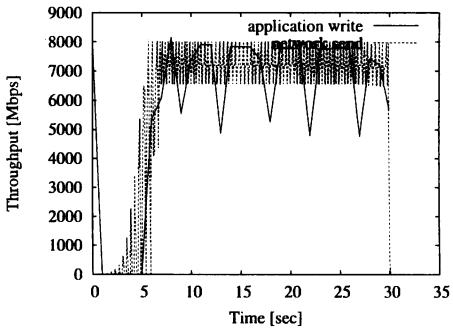

図 8　アプリケーションのデータ送出とウインドウサイズの推移 (RTT=400ms)

ホスト PC のメモリとバスの通信性能とアーキテクチャが通信性能に影響する。Opteron の AMD チップセットの SMP 構成ではプロセッサのそれぞれが 2 本ずつのメモリバスを持っており、今回の構成では 1GB ずつ異なるメモリバスに接続されている。メモリへのアクセスが複数のメモリバスにまたがる場合アクセス時間増加することが考えられる。RTT が 400ms の時にはウインドウサイズが 1GB を超えるため、RTT=0,200ms の場合と比較して性能が低下している。

また、高速なデータ転送を行うため、システムの TCP/IP スタックとアプリケーションがバランスよく動作することが必要である。今回の転送実験では、転送性能にアプリケーションが送出に用いるユーザバッファサイズが性能に大きく影響していることが分かった。通信時には、送信側の方が、データ生成とパケット生成のために受信側よりも負荷が

$MSS \times n$	9158	18316	36632	73264
$n=$	1	2	4	8
BW/CPU	0.224	0.214	0.210	0.185

表 5　送信側のバッファサイズと性能比率 (MTU=9198)

$MSS \times n$	36632	54948	73264	146528
$n=$	4	6	8	16
BW/CPU	0.183	0.204	0.160	0.192

表 6　受信側のバッファサイズと性能比率 (MTU=9198)

高い。CPU リソースを測定した際、ユーザバッファサイズを大きくすると特に送信側の CPU 負荷が上昇する。データ転送中は、送信側の CPU 使用率が 100% で、受信側は 10% 以上の空き使用時間がある場合が最も転送性能がよく安定して通信した。ユーザバッファサイズが実行性能に与える影響を求めるため 通信速度/*CPU* 使用率 を比較したのが表 5,6 である。ユーザバッファはネットワーク上でのデータサイズ ($MSS = MTU -$ ヘッダサイズ) の整数倍でとるのが効率よい。$MSS \times n$ で変化させた場合、大きすぎるユーザバッファは CPU 負荷を増加させ通信性能を悪化させる。これらの表から、送信側ユーザバッファは $MSS \times 1$, 受信側ユーザバッファは $MSS \times 6$ が最も良く、遅延が増加しても、これらの場合が最も良い性能であった。

高速通信時は割り込みが大幅に増加する。実験中最大 22000inter/s の割り込みがある状況で転送を行ったが、遅延によって割り込み数は変化しなかった。また、アダプタからの割り込み回数はアダプタによる発生間隔が制限されていたり、NAPI による割り込み禁止期間があるため、割り込みによる CPU への負荷は通信性能に影響しないことが分かった。

3　まとめ

RTT の長い擬似ネットワーク環境を構築し、TCP/IP による通信実験を行い、ネットワークカードの性能を引き出す方法を求めた。適切なサイズのウインドウサイズ、ユーザバッファを用いて通信することで RTT=400ms の場合に 9KByte ジャンボフレームを使用すれば Chelsio T110/N110 で 7.2Gbps、Intel PRO/10GbE で 6.4Gbps のバンド幅を得た。一方、1.5KByte スタンダードフレームを使用した場合に Chelsio T110/N110 で 1.8Gbps、Intel PRO/10GbE で 1.98Gbps に留まっている。この原因を明らかにすることは今後の課題である。Intel PRO/10GbE の TSO(TCP Segment Offload) を使用しても CPU 負荷は軽減されるものの転送性能が向上しなかった事から、PCI-X バス、CPU のパケット処理能力だけでなくネットワークアダプタの I/O 処理効率が重要だと言える。

今後、今回の結果を元に実際の長距離回線での通信実験を行う予定である。現実の回線ではルータなどが存在するため、パケットのロスやリオーダリングが発生する。この場合には TCP/IP のウインドウコントロールが発生し、擬似ネットワーク環境のような遅延だけが存在するネットワークとは異なる状況での通信となり、通信特性もシステムの振る舞いも大きく変化すると考えられる。その比較を行う予定である。

謝辞

本研究は, 文部科学技術省 科学技術振興調整費「重要課題解決型研究等の推進－分散共有型研究データ利用基盤の整備」, 基盤研究 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機構」, および２１世紀 COE「情報科学技術戦略コア — 大域ディペンダブル情報基盤で補助された。

参考文献

[1] M. Nakamura, R. Kurusu, F. Marti, M. Sakamoto, Y. Ikuta, J. Tamatsukuri, Y. Sugawara, N. Aoshima, M. Inaba, K. Hiraki, "Experimental Results of inter-layer cooperative hardware for TRC-TCP on 10Gbps Ethernet WANPHY 18,500km Network", PFLDnet 2005, Feb. 2005.

[2] G. Hurwitz, W. Feng, "Initial End-to-End Performance Evaluation of 10-Gigabit Ethernet", In Proceedings of Hot Interconnects 11 (HotI' 03), Aug. 2003.

[3] Richard Hughes-Jones. "PCI-X Activity and UDP measurements using the Intel 10 Gigabit Ethernet NIC", PFLDnet 2004, Feb. 2004.

[4] Data Reservoir Project., "Internet2 Land Speed Record in single and multiple TCP stream", http://data-reservoir.adm.s.u-tokyo.ac.jp/lsr-20041225/index.html, Dec. 2004.