様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月１２日現在

機関番号：１２６０１
研究種目：若手研究（B）
研究期間：2009 ～2011
課題番号：21700068
研究課題名（和文）　イーサネットに適したＴＣＰ／ＩＰ通信方式の研究

研究課題名（英文）　Research of effective TCP/IP communication on Ethernet

研究代表者
玉造　潤史（TAMATSUKURI JUNJI）
東京大学・大学院理学系研究科・准教授
研究者番号：90322049

研究成果の概要（和文）：
10 ギガビットネットワークが普及し、高速な TCP/IP 通信が行われるようになった。TCP/IP プロトコルではパケット損失を検出し、途中継路上で利用できる最大帯域を推定してロスの無い通信を実現している。ネットワークメディアとして多く利用されているイーサネットのインターフェースは、特性として通信データを受け取るとメディアの最大速度で送出する。そのため、平均通信速度での最大性能を引き出す TCP/IP プロトコルで利用すると、瞬間的な送出速度の問題でパケット損失を生じる。この影響を抑えるため TCP/IP スタックがネットワークインターフェースに渡すデータを送出速度に合わせて調整することで、安定的な通信を行うことを Linux 上に実装しテストを行った。

研究成果の概要（英文）：
According to spread the 10 Gigabit network, servers could communicate in high speed by TCP/IP protocol. TCP/IP protocol detects packet loss and assumes available maximum bandwidth in the communication path. An Ethernet interface, generally used in network connection, transmits data putting from TCP/IP protocol stack at its maximum media rate. When TCP/IP protocol used for exploiting the average communication rate, an Ethernet generates packet loss at the maximum media rate. To avoid these packet losses, we tried to adjust the transmission speed from TCP/IP stack to Ethernet interface. We examine its effect in Linux servers.

交付決定額

（金額単位：円）

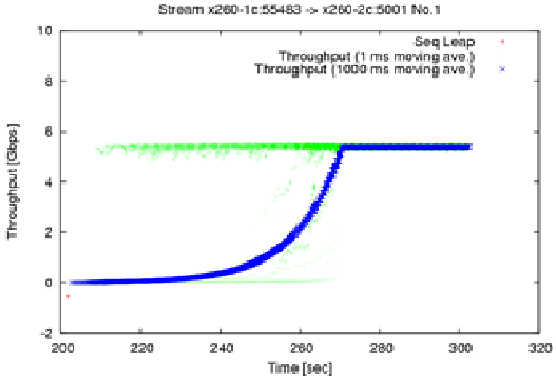
	直接経費	間接経費	合　計
２００９年度	2,000,000	600,000	2,600,000
２０１０年度	800,000	240,000	1,040,000
２０１１年度	500,000	150,000	650,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学、計算機システム・ネットワーク
キーワード：イーサネット、TCP/IP

１．研究開始当初の背景

本研究では、インターネット通信で通常使われている TCP/IP プロトコルによる通信を、ネットワーク物理レイヤとして一般的に用いられているイーサネット上で行う際の問

題点を明らかにより効率的な通信を可能とするネットワークドライバと TCP/IP スタックを構築することを目的とした。
既に、世界中をつなぐネットワークが 10Gbps クラスの大容量接続であることが一般的となり、これまで不可能であった超高速通信によるデータ交換が可能となりつつある。ネットワーク回線としては SONET/SDH による回線容量の 9.26Gbps OC-192 回線が海外回線としても利用されている。これらは WAN 公衆ネットワークでの物理層として用いられてきたが、WANPHY 技術により 10Gbps イーサネットに変換され利用可能となっている。WANPHY では広域系のネットワークで用いられる SONET/SDH フレーミング規格の回線上にイーサネットフレームをそのまま乗せることを可能としており、WANPHY によって、世界中の長距離ネットワーク回線がすべてイーサネットによって接続されることとなった。結果として以前はローカルネットワークに限定的に用いられてきたイーサネットが世界中を接続する通信に利用されるようになっている。
このような長距離イーサネットにおける通信において問題として当初考えられたのはこれまでのネットワーク同様、接続しているネットワーク機器でのバッファあふれ、他の通信との通信回線への入力の輻輳であった。しかし、実際に通信を行うと経路上の通信速度に問題がない場合であっても安定的な通信ができないという問題が明らかになってきた。



これはイーサネットが持つ本質的な通信上の振る舞いに問題があると考えられる。右のグラフの太い実線が 1 秒平均での速度であり通常ユーザが通信速度として用いる通信速度である。ところが、イーサネットのインターフェースは送出時にプロセッサからインターフェースが受け取るデータをハードウエア割り込みなどによって一括して受け取り、一気に送り出す。そのためインターフェースからネットワークへの送出の速度はマクロな時間ではプロセッサが送っているデータの供給速度(TCP/IP プロトコルスタックがコントロールしての速度)となるが、ミクロな時間では物理層の最大速度で送出してしまうという振舞いである。その結果、グラフ中の実線外に分布している無数の点が 1 ミリ秒あたりの平均のパケット送出速度であり、この送出速度はほぼインターフェースの限界速度で送出していることが判明した。(本グラフでは約 5.5Gbps)このように送出可能なデータを規格上の最大速度での送出することは通信性能を向上させるように思えるが、実際には、経路上の機器や受信側のサーバは常に物理層最大速度での受信をしなければならないということを意味し、通信経路上に常に最大速度で影響を与えていることになる。また、ホストでは受信動作を効率的に行うために、現在の通常のネットワークドライバはパケットの到着ごとにプロセッサへのハードウエア割り込みを発生させる。これが ACK パケットの場合では、非常に短いパケットが非常に短時間に到着し、非常に細かい粒度での割り込みを発生させる。そのため、実際に利用可能な大容量のネットワークであってもこのような通信をしている限りより高速な通信を実現することは難しい。安定なネットワークへの送出速度と平均通信速度とのギャップを埋ることが安定的に通信を実現するためには必要となる。

２．研究の目的

本研究ではイーサネットにおける TCP/IP 通信を安定化し、高性能を発揮する協調して動作するネットワークドライバおよび TCP/IP スタックを開発し、ネットワークで安定な性能を得ることにある。
初めに、イーサネットによるネットワーク通信速度のコントロールを実現する。既に現在のイーサネットにおける通信の状況を測定する方法は TAPEE(Tengigabit hardware Aided Precise Ethernet Engine)により確立しており、この方法での計測結果を活用することで、イーサネットドライバにおいてミクロなネットワークへの送出速度のコントロールを実現することを目的とする。既存のいくつかのネットワークインターフェースには送出時のパケット間隔(IPG,Inter Packet Gap)をコントロールすることができるものがある。この機能を活用し通信時にダイナミックに IPG をコントロールしながら通信を行うことを実現する。さらに IPG の設定機能のないインターフェースにおいて同様の機能を実現するため、通信に影響を与えないパケット間隔を生成するよう、速度をコントロールしてネットワークインターフェースにデータ送出することで通信の状況をコント

ロールする方法を検討することとした。
さらに、このデータ送出コントロールにより通常の TCP/IP の輻輳制御による輻輳ウインドウコントロールとネットワークドライバの協調を実現することを検討することとした。これは既存の TCP/IP スタックに最小限の変更を加えるだけで実現し通常利用可能な TCP/IP プロトコルスタックとして開発する。高速で高負荷の TCP/IP 通信では輻輳ウインドウ制御部分の負荷がプロセッサ負荷急激に上昇させる。その状況をホストの計測から明らかにし、急激に上昇するプロセッサ負荷の上昇を TCP/IP スタックが追従できるよう、ウインドウ成長を先取りした負荷上昇と同様に、スタックの実行時間スケジュールを行う TCP スタックとして作成することを目指す。このスタックでは低負荷状態からの TCP のウインドウサイズのスケーリングに合わせた擬似的な負荷上昇をスケジューリングによってエミュレートする。このスタックの実行状況にあわせた輻輳ウインドウ成長をコントロールすることによって、プロセッサ資源が一時的に枯渇し、輻輳が起こった場合と同様に振舞う状況を改善する。このようなネットワークドライバと TCP/IP スタックが協調し安定した通信性能を得ることが本研究の目的である。

３．研究の方法

本研究で行う具体的なシステム実装と実験内容は以下のとおりとした。
(1)TCP/IP 通信実験を行うサーバシステムの構築
(2)イーサネットでの TCP/IP 通信の精密な解析とホスト状態の解析から通信不安定原因要素の確定
(3)Linux におけるイーサネットドライバにおいて IPG 調節および破棄可能なパケット挿入によるパケット送出レートコントロール機能の付加
(4)Linux カーネル上での送出速度コントロール機能を付加したイーサネットドライバと強調する TCP/IP スタックの設計と実装
(5)TCP/IP スタックのテストとファイル転送プロトコルを用いた詳細な通信状況の解析
(6)実際にイーサネットによる長距離高遅延ネットワーク上と近距離ネットワーク上での実行環境の構築と性能測定

これらの実装を行い、測定結果を基に本研究での提案方式の効果をまとめることとした。

初年度である平成 21 年度は主として実験環境の構築とイーサネット通信の詳細な計測およびイーサネットドライバにおける送出速度調整の実現を主として行った。

１）近接対向環境の構築
購入したサーバをローカルに設置し 2 台ずつを直接イーサネットで接続した実験環境を構築する。構築されたサーバ環境においてイーサネットの影響を計測できるようにし、サーバ単体での通信性能を確認した。
２）イーサネットパケット送出速度コントロール機能のカーネルへの実装の調査
実装開始時の最新 stable カーネルのソースコードを調べ、現在利用可能なネットワークドライバで利用可能なパケット送出速度調整機能について精査した。
３）イーサネットパケット送出速度コントロール機能のカーネルへの実装の設計
２の実装調査に基づき、stable カーネルにに対して、パケット送出速度コントロール機能付きドライバとなる機能付加について設計を行う。IPG を活用して実現可能なネットワークカードでは、可能な限り IPG により実現を行うが、多くのドライバでの利用を考慮しできるだけ特定のネットワークカードの機能に依存しない方法で設計を行った。
４）イーサネットパケット送出速度コントロール機能のカーネルドライバへの実装
３での設計によるパケット送出速度をコントロールできるネットワークドライバの実装を行う。この機能付加は可能な限り CPU および I/O の性能低下を起こさず、ローカルネットワーク上で破棄するか、もしくはネットワークカードのコントローラ内でパケットが挿入されていることと同じ状態を作り出す方法で実装を行った。

平成 22 年は引き続き、10G イーサネットインターフェースの整備を行い、10Gbps ネットワークでの詳細な計測を行うとともに、ネットワークドライバに協調する TCP/IP スタックの開発を行った
１）遠距離対向環境の増強
追加購入した 10 ギガビットイーサネットインターフェースを用いて対外ネットワークと接続し、長距離ネットワーク上での通信テストが行える実験環境を構築した。特に、外部接続については、本拠地である東京大学から海外拠点を回り、東京大学に戻る世界一周回線が利用可能となるように構築した。
２）カーネルドライバを用いての実ネットワーク上でのパフォーマンステスト
実装した送出速度コントロール機構付きカーネルドライバを用いて通信テストを行った。アプリケーションとしてはホスト性能に依存するがメモリからメモリへの通信を行う iperf による通信システムをサーバホスト上に構築して利用した。
３）ファイル転送プロトコルの通信状況の精査

パフォーマンステストにおいて、実際のネットワーク上では起こる通信パケット間隔がつまる現象や、実際の輻輳などと、ホスト内で起こりうる輻輳がどのように起こっているかを、開発したカーネルドライバと通常の Linux TCP スタックの両方で精密に計測を行う。通信状況の精密な計測にはデータレゼボワールプロジェクトで開発した TAPEE (Tengigabit hardware Aided Precise Ethernet Engine)を用いて行った。両方のスタックでの通信状況と Acknowledge packet の状況を調べることで、実装の実ネットワーク上での有効性を確認した。

４）送出速度コントロールドライバと協調する TCP/IP スタックについての考察と調整

上記の長距離、高遅延ネットワーク上での TCP 通信の測定結果から実装した送出速度コントロール機構と協調する TCP/IP スタックの評価を行い、その有効性に関する考察を行う。明確に問題が分かった場合には再調整を行い効果の有効性を再検討した。

平成 23 年度にさらにいくつかのネットワークインターフェースを整備することにより、より詳細な通信状況の計測および解析を行う。その結果を報告書としてまとめ、総括を行った。

１）送出速度コントロール機能の汎用化とその有効性の検証とイーサネットを活用するコンピュータシステムでの有効性の検証

前年度までに開発を行った機能について汎用化を行い複数のネットワークインターフェースドライバにおいても利用可能とするため、実験環境に仮想化環境を導入して送出コントロールの実験を行った。仮想環境としては Xen および kvmの両方を用い、速度測定を行った。

２）研究成果のとりまとめ

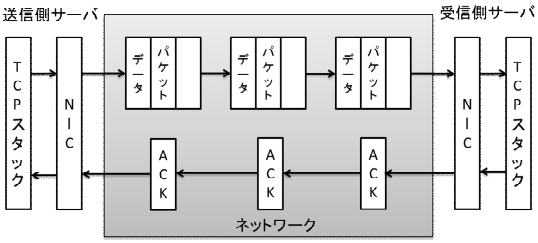
本研究において得られた通信計測の結果をとりまとめ、イーサネットと TCP/IP スタックの協調に関する評価を行った。

４．研究成果

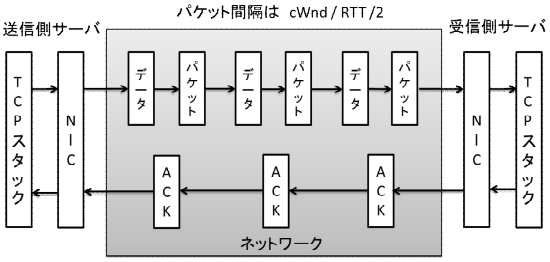
本研究では、高性能サーバ 2 台を 10Gbps イーサネットスイッチで繋ぎ仮想化クラスタ化して TCP/IP 通信を行った。

１）イーサネットインターフェースを活用する TCP/IP スタックの実装について

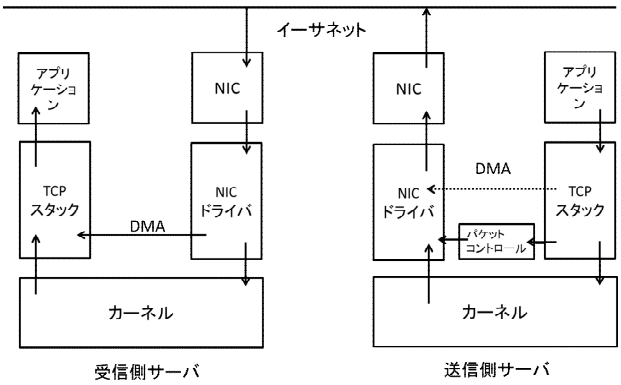
TCP/IP スタックからのデータ送出コントロールをカーネルを介し、DMA を用いて行っているデータのやりとりを、TCP/IP 通信での RTT と cwnd のサイズに合わせて行う実装を行った。TCP/IP 通信は、通常下図のようにデータを送出速度に合わせて渡している。



高速な通信の場合 ACK パケットは複数のデータパケットの到着をまとめて送出するため間隔を空けて通信しているが、データパケットは、TCP/IP スタックからネットワークインターフェース（NIC）に送出される際に DMA 転送を用いてデータパケットがまとめて送出される。これを下図のように低速な通信の場合にはデータパケット側を cWnd / RTT /2(データパケットの送出は片道であるため)の間隔で送出する実装を行った。



実際に通信する部分としては下図のように TCP/IP スタックからの送出部分を、データサイズで行うのではなく、cWnd / RTT /2 を基にしたパケットサイズ単位で送出することを行った。



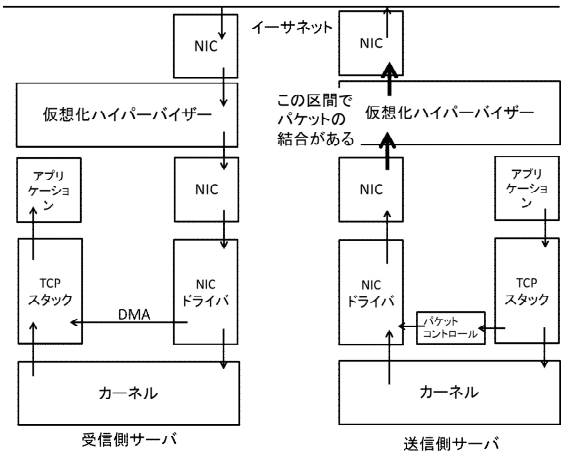
現時点では、TCP パケット生成のオフロード機能は使えていない。しかし、パケットサイズ単位のデータ転送を実施しているため、TCP パケット生成のオフロード機能である TSO(TCP/IP Segment Offload)機能は有効に

することができると考えている。また、オフロード機能の有効性は TCP/IP スタックの負荷としてパケットのチェックサム生成にあり、TSO が機能できれば求められている CPU 負荷低減は可能であると考える。

２）単体性能について
行った実装について、対向環境での通信速度測定を行った。イーサネットのパケット送出の振る舞いは想定したように、パケット間隔を空けて送出することができた。この状況は通信速度が低い状況でないと確認できなかったため、通信速度が低くなるように TCP/IP スタックのウインドウサイズを制限し、低速状態での確認となった。
今回用いたサーバはパケット間調整を行って通信を行う負荷があっても十分に早い環境であったため、低速状況においては特に影響はなく、イーサネットでのパケット損出が発生するような状況を創出しての確認は困難であることが分かった。多数のサーバが通信し複数の TCP/IP 通信が一つのイーサネットセグメント上で輻輳を生じるような状況での通信状況測定を行うことが本方式の効果の最終確認ではあるが、そこまでの確認はできなかった。

３）仮想化環境での性能について
現実に近い環境として今回の実験環境に仮想化環境を構築しての実験を行った。
仮想化ハイパーバイザーとして Linux 環境で標準的に用いられている Xen と Kvm の環境を 2 台のサーバに導入して実験を行った。どちらの環境でもネットワークはブリッジモードで利用し、内部的には仮想ネットワークインターフェースに対して通信するように設定した。
今回開発した TCP/IP スタックを用いたが、通信自体は残念ながら想定したような振る舞いとならずサーバ外からの通信としては通常の通信と同じくパケットがまとまったような通信形態となった。

この原因は、サーバのハイパーバイザー内でのパケットの振る舞いによる影響が出ていると考えられた。下記のように、仮想サーバから出たパケットはハイパーバイザー内で DMA 転送として送出しており、そのため、送出間隔を調整したパケットが複数合わさって転送されていることに起因していると想定される。



この問題を解決するにはハイパーバイザー内でのネットワークインターフェースへの通信を DMA 方式による方法を止め、パケット単位のフォワーディングを行うような改変を行うか、もしくは TCP/IP スタックでのコントロールによる方法ではなく、ネットワークインターフェースでの速度調整の方法でのパケット間調整をより高度に実施する方法が考えられる。しかし、後者においては現時点ではあまり大きな IPG 調整ができるネットワークインターフェースは存在しないため、IPG が大きく可変（少なくとも転送性能を 1/10 程度まで低減できるよう IPG が設定できる）ようなインターフェースを開発する必要がイーサネット通信のパケットバースト現象を抑止するためには必要である。

４）まとめ
本研究では、イーサネットのパケット送出速度を調整し、輻輳が起こらない TCP/IP 通信をする TCP/IP スタックの開発行った。低速ではあるが、パケット間隔を空けた通信が可能であることを確認した。開発した TCP/IP スタックの安定性の確認には多数のサーバが必要であり、この検証が今後必要となると考える。
現実に近い適用として仮想化環境での通信計測も行った。仮想化ハイパーバイザーを介した状況下では、実サーバとの通信の振る舞いが異なり、ハイパーバイザーのネットワークパケットの転送による影響があることが分かった。この場合、ハイパーバイザー上の仮想サーバだけのパケット間コントロールだけでなくハイパーバイザーのイーサネットインターフェースにおけるパケット間隔のコントロールが必要であることが分かった。

５）今後の展開について
イーサネットのパケットバースト送出により輻輳が生じる状況はこれまでネットワークシミュレータ上では確認されているが、本

研究のパケット間調整を行った通信の効果を確認するには、多数のサーバが同様にパケット間調整を行った状況での安定性を確認する必要がある。最低限で6台程度のサーバを用いて通信する必要があるため、そのような状況下での検証を今後行っていきたいと考えている。
実環境での効果として仮想化クラスタ上での通信について適応を行ったが、現時点では仮想化ハイパーバイザーを経由しての通信としてはうまくパケット間調整ができていない。パケット単位のフォワーディングを行うようなハイパーバイザーの利用はできると思われるためその確認をして、仮想化環境での速度計測を実施したいと考えている。これが可能となれば前述したパケット輻輳状況下の再現も台数を減少でき、通信安定性の効果測定には必要であると考えている。
また、性能計測において問題があったため、これまでの成果ついて、まとめきれていないが今後成果発表を行っていきたいと考えている。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計０件）
なし

〔学会発表〕（計０件）
なし

〔図書〕（計０件）
なし

〔産業財産権〕
なし

〔その他〕
ホームページ

６．研究組織
(1)研究代表者
玉造　潤史（TAMATSUKURI JUNJI）
東京大学・大学院理学系研究科・准教授
研究者番号：90322049

(2)研究分担者
なし

(3)連携研究者
なし