様式 C-19

科学研究費補助金研究成果報告書

平成２２年　６月１０日現在

研究種目：若手研究（スタートアップ）
研究期間：2008～2009
課題番号：20800083
研究課題名（和文）　ソフトウェアによる高精度パケットスケジューリング機構の開発

研究課題名（英文）　Development of a precise software packet scheduling mechanism

研究代表者
高野　了成（TAKANO RYOUSEI）
独立行政法人産業技術総合研究所・情報技術研究部門・研究員
研究者番号：10509516

研究成果の概要（和文）：
本研究は、大容量ネットワークにおける帯域保証を実現するために、ソフトウェアによる高精度かつ柔軟性のある新しいパケットスケジューリング機構を開発することを目的とする。本機構はパケットの送信予定時刻を送信済みバイト数を基に決定することで、ハードウェアタイマに依存しない精密なスケジューリングを可能にする。ギャップパケットと高解像度タイマを用いた 2 種類の方式を実装してその得失を明らかにした他、商用音楽データ配信サービスでの安定運用も達成できた。

研究成果の概要（英文）：
The aim of this study is development of a new software-based precise and flexible scheduling mechanism to guarantee the data transmission performance on high-speed networks. The proposed mechanism determines transmission timing of packets by the number of bytes transferred to enable a precise packet scheduling without dependence on a hardware timer. We show the pros and cons of two types of implementations: a gap packet based and a high-resolution timer based. We also report it has been in stable operation on a commercial music data delivery service.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,330,000	399,000	1,729,000
2009 年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	2,530,000	759,000	3,289,000

研究分野：オペレーティングシステム
科研費の分科・細目：計算機システム・ネットワーク
キーワード：インターネット高度化、計算機システム

１．研究開始当初の背景

近年の広域ネットワーク通信技術の急速な発展により、IP 電話、映像コンテンツ配信、大容量ファイル転送など、実時間性やスループットなどさまざまな要求を持つアプリケーションが大容量ネットワークを介して実行可能になる。アプリケーションのサービス品質およびネットワーク利用効率の向上を

達成するには、各アプリケーションの要求帯域を正確に保証し、競合する要求に対しては優先度に基づいた調停機能を有するパケットスケジューリング機構が必要である。

しかし、一般的な帯域制御で用いられているトークンバケット方式では、平均レートを制御することはできるが、短時間に指定したレートを超えてデータが送出されるバーストトラフィックの発生を回避できないという根本的な問題がある。ネットワークの大容量化が進展するほど、バーストトラフィックがネットワークに与える影響は増大し、輻輳によるアプリケーション通信性能の低下やネットワーク利用効率の低下が問題となる。

この問題に対して、既定の送信レートに基づいてパケット送信間隔を均一に制御することでバーストトラフィックを平滑化し、安定した帯域制御を実現する技術としてパケットペーシングが知られている。しかし、タイマ割込み処理の精度や負荷増大などの問題から、マイクロ秒精度の制御が要求されるギガビットクラスのネットワークにおける実現は困難とされ、実環境では利用されてこなかった。これに対して、我々の先行研究ではパケットの送信予定時刻を送信済みバイト数を基に決定することで、端末のハードウェアタイマに依存することなく精密に制御できるバイトクロックスケジューリング方式を提案した。しかし、さらなる高速ネットワークへの対応や柔軟かつ簡便な利用といった課題が存在していた。

２．研究の目的

本研究は、大容量ネットワークにおける帯域保証を実現するために、ソフトウェアによる高精度かつ柔軟性のある新しいパケットスケジューリング機構を開発することを目的とする。具体的な目的を次に述べる。

（１）アプリケーションごとの要求に合わせた Kbps から 10Gbps の範囲に渡る高精度な帯域制御を実現する。10 ギガビットイーサネットさらに 40、100 ギガビットイーサネットへの対応も考慮に入れて、先行研究で提案したギャップパケット方式に加えて、近年オペレーティングシステムへの対応が進んでいる高解像度タイマを用いた方式も検討する。

（２）トラフィック量の増減に合わせて、クラス間で帯域の貸し借りを許す柔軟なパケット優先度制御を実現する。

（３）上記のパケットスケジューリング機構が、実アプリケーションから容易に利用できるようにトラフィック制御設定を自動化するソフトウェアを開発する。一般的に採用されている帯域制御機構はクラスベース制御を前提としているが、ここでは IP フロー単位での帯域制御を対象とする。

３．研究の方法

提案するパケットスケジューリング機構を Linux オペレーティングシステムのネットワークスタック上に実装し、大容量通信アプリケーション実行における効果を実証する。

実験環境として、10 ギガビットイーサネットを有する 2 台の計算機を用意し、後述する GtrcNET-10 を介して接続する。各計算機は CPU に Quad-core Xeon E5430/2.66GHz dual、8GB のメモリ、そして 10 ギガビットイーサネット NIC（Network Interface Card）として PCI-Express x8 バスに接続した Myricom Myri-10G を搭載する。OS として Ubuntu server 9.04 を用いた。

評価に際しては、アプリケーションの実効通信性能とは別にパケットスケジューリングの精度を比較解析するために、10 ギガビットイーサネットに対応したハードウェアネットワークテストベッド GtrcNET-10 を使用する。GtrcNET-10 は通信性能に影響を与えることなくワイヤレートでパケットをキャプチャすることが可能であり、取得したパケットキャプチャ結果をオフラインで解析し、パケット送信間隔やバースト性を評価する。

４．研究成果

（１）バイトクロック方式に基づき、ギャップパケットまたは高解像度タイマを用いた高精度なパケットスケジューリング機構を 10 ギガビットイーサネット上で実現した。特に高解像度タイマ方式はスケジューリングの精度を維持しつつ、汎用性を高めることでその適用範囲を広げることを可能にした。具体的にはワイヤレートで通信できないシステムやイーサネット以外の通信媒体でも利用可能になった。

まずバイトクロック方式の概要について述べる。本方式では、パケット送信時刻を正確にスケジューリングするために、実時間の代わりに通信開始からの送信バイト数を用いる。その根拠は、1 バイトのデータ送信に要する時間は、例えばギガビットイーサネットであれば 8 ナノ秒と一定であり、パケットを隙間なく連続して送信できれば、送信時刻は正確に制御できる点にある。その動作の概要を図 1 に示す。入力パケットは宛先アドレスやプロトコルクラスに基づいてクラス分けされ、クラスごとのキューにキューイングされる。各クラスは目標送信レートを持つ。キューの先頭パケットの送信予定時刻をクラスクロック、現在時刻はインタフェースごとのグローバルクロックと呼ぶ。スケジューラはグローバルクロックと各クラスクロックを比較し、送信予定時刻に達したパケット

から順に送信する。送信時には、クラスクロックにパケット送信に要する時間と次のパケット送信までの待ち時間を加算する。上記のスケジューリングの結果、図1右のようにパケット送信が不要な通信アイドル時間が発生する。この時間を正確に制御し、バースト送信を抑えることが、正確な帯域制御の要となる。

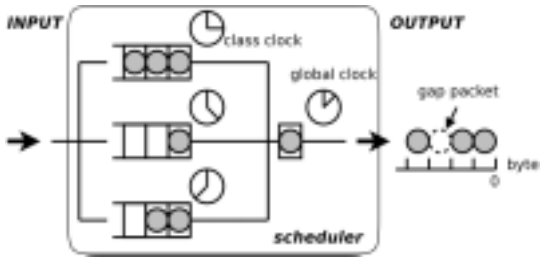

図1　バイトクロック方式

本研究では通信アイドル時間を制御するために、ギャップパケットと高解像度タイマの二つの実装方式を採用した。前者の方式は、通信アイドル時間に相当する長さのダミーパケットを計算機から送信することで後続するパケット送信時刻を制御する。このダミーパケットをギャップパケットと呼び、イーサネットではフロー制御に用いられる PAUSE フレームを利用することで実現する。後者の方式では、Linux が提供するマイクロ秒精度の高解像度タイマ hrtimer をワンショットタイマとして利用して送信時刻を制御する。

帯域制御の正確さを評価するため、目標送信レートを 100Mbp から 10Gbps まで、100Mbps 刻みに設定した場合のグッドプットを測定した。比較対象として、PSPacer（ギャップパケット方式)、PSPacer/TB（高解像度タイマ方式）の他に、Linux 標準の HTB（Hierarchical Token Bucket）を比較した。目標送信レートと実測して得られたグッドプットの差分を図2に示す。PSPacer では目標レートの増加に比例して、わずかだが差分が大きくなり、実測値は最大で 0.14Gbps 下回っている。これは本実験環境では、

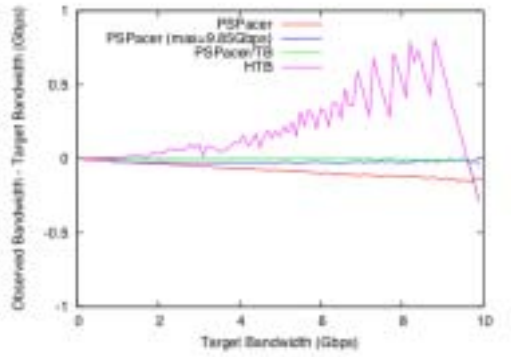

図2　目標送信レートと実測値との差分

Myri-10G がワイヤレートをわずかに下回る速度でしか送信できていないことに原因がある。そこで最大送信レートを 10Gbps の代わりに 9.85Gbps へと調整した結果（ラベル PSPacer (max=9.85Gbsp))、差分は無視できるまで小さくなった。PSPacer/TB ではさらに差分が縮まり、一番精度が高いという結果になった。一方、HTB では目標送信レートが高くなるにした正確な帯域制御に失敗している。特に、HTB は差分が最大 0.81Gbp に達するなどと大きく、さらに差分の増加が一様ではない。HTB も内部的には hrtimer を使っているが、PSPacer/TB と比較しても正確さに劣る理由はスケジューリング方式の違いにある。

この結果より、PSPacer は 10 ギガビットイーサネット環境においても既存方式と比較して、非常に高いスケジューリング精度を実現しており、大容量アプリケーションの通信性能向上に大きく貢献できると考える。またデータセンタなどクラスタ内部での TCP/IP 通信において、少数の計算機に通信が一時的に集中することによって極端に通信性能が劣化する TCP incast 問題が注目を集めている。提案機構はこのような問題に対しても有効な解決策の一つになると考える。

今後さらに NIC 性能が向上すると、ギャップパケット方式ではワイヤレート性能を前提とする点が足かせになるので、高解像度タイマ方式が有効になると考える。ただし高解像度タイマ方式には CPU 負荷の点で課題が残っている。制御対象クラス数が少ない場合はギャップパケット方式よりも軽量であるが、クラス数が増加すると負荷が逆転することを確認している。スケジューリングの精度をできるだけ維持しつつタイマ割込み回数を削減できる手法を開発する必要があると考える。

（2）目標送信レートの合計が物理リンク帯域を超える over subscribed 設定に対応するため、バイトクロック方式のアルゴリズムを改善した。この改善によりアプリケーションの多様な要求に答える柔軟な帯域制御が可能となった。

over subscribed 設定ではクラスクロックがグローバルクロックより遅れるが、その際に送信可能状態でキューイングされているパケットがバースト送信される可能性がある。そこで over subscribed 状態を検出し、クラスクロックを調整する仕組みを追加した。クラスクロックを調整する際にどのクラスを優先して送信するか考慮する必要がある。ギャップパケット方式ではクラス間の公平性を重視し、目標送信レートに傾斜比例した帯域になるように制御する。一方、高解像度タイマ方式では Linux 既存の QoS 機能と組み合わせることをより重視して開発し、優先

度制御や階層的な帯域制御なども実現した。したがって、その組み合わせ次第で優先度制御できるようにした。したがって今後新たな QoS ポリシが提案されたとしても容易に対応が可能と考える。

ギャップパケット方式において、目標送信レートの設定を 4、4、4、2、2Gbps に設定し、合計帯域を 16Gbp に設定した場合の実験結果を図3に示す。各通信は10秒おきに開始され、20秒から60秒の区間では10Gbps のリンク帯域を 2:2:2:1:1 の割合で正確に共有できている。これは HTB と比較して誤差が小さいこともわかった。

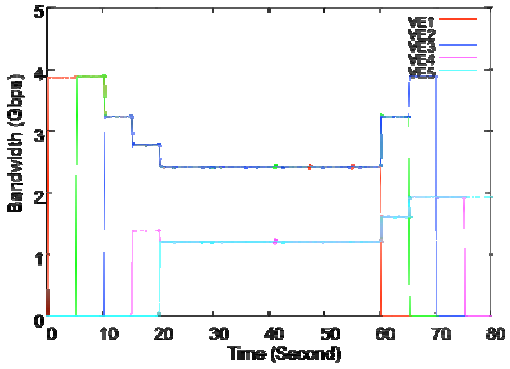

図3　over subscribed 設定での帯域共有

（3）提案機構をデータ配信サービスに適用するための手法を検討し、IP フロー単位のトラフィック制御設定を自動化するソフトウェアシステムを開発した。さらに商用音楽データ配信サービスでの安定運用を実現した。

Linux のトラフィック制御機能は基本的にクラスベース QoS を前提としており、フローベース QoS を実現するには、フロー数の増加に伴うオペレーションコストの増大を解決する必要がある。本システムでは、制御対象アプリケーションのシステムコール呼出しを監視して IP フローを識別することで、あらかじめシステム管理者が記述した平易なルールに従って、自動的にフローキューを作成し設定する。

本システムおよび PSPacer は、2009 年 7 月より IP ネットワーク経由で音楽データを配信する商用サービスにおいて運用が開始され、複数台のサーバから 2000 以上クライアント端末へデータ配信が行われている。その結果、帯域制御専用ハードウェア装置と同様の簡便さで帯域制御を実現することが可能となり、システム構築費用が削減できたこと、オペレーションコストが削減できたことを確認した。今後は性能保証型分散ファイルシステム Papio との連携など、多様な応用への展開を考えていきたい。

（4）上記すべての開発成果物は GNU GPL ライセンスに基づくオープンソースソフトウェアとしてプロジェクトのホームページにて公開済みである。高速ネットワーク上で大容量通信を効率的に行いたいという要求を持つ国内外のユーザに利用され始めており、商用利用にも耐え得るソフトウェアとして国内外の研究者・開発者に普及させることができた。今後は Linux カーネルへのマージも含めてさらなる普及化を探っていきたい。

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３件）

①高野了成、工藤知宏、児玉祐悦、岡崎史裕、精密な帯域共有とトラフィック隔離を実現するパケットスケシューリンク方式、情報処理学会研究会報　2009-HPC-119、査読無、2009、67-72

②高野了成、工藤知宏、児玉祐悦、岡崎史裕、10 ギガビットイーサネットを用いた精密なパケットスケジューリング機構の開発、電子情報通信学会技術研究報告 109-188、査読無、2009、31-36

③高野了成、工藤知宏、児玉祐悦、小竹賢、丹羽証、IP フロー単位のトラフィック制御設定の自動化機構、電子情報通信学会技術研究報告 109-273、査読無、2009、59-64

〔学会発表〕（計３件）

①高野了成、IP フロー単位のトラフィック制御設定の自動化機構、電子情報通信学会 NS 研究会、2009 年 11 月 13 日、金沢工業大学

②高野了成、10 ギガビットイーサネットを用いた精密なパケットスケジューリング機構の開発、電子情報通信学会 NS 研究会、2009 年 9 月 10 日、東北大学

③高野了成、精密な帯域共有とトラフィック隔離を実現するパケットスケシューリンク方式、情報処理学会 HPC 研究会（HOKKE 2009）、2009 年 2 月 27 日、北海道大学

〔その他〕

ホームページ等

https://www.gridmpi.org/pspacer.jsp

６．研究組織

(1)研究代表者

高野　了成（TAKANO RYOUSEI）
独立行政法人産業技術総合研究所・情報技術研究部門・研究員
研究者番号：10509516

(2)研究分担者

(3)連携研究者