超高速I/O指向オペレーティングシステム

「機能と構成」領域　　河合　栄治

要　　旨

インターネットの超高速化にともない、ネットワークサービスのボトルネックが、従来の低速で高価なネットワークから、サーバに移行してきている。そのため、現在サーバのクラスタ化などの分散技術が多く開発されているが、ネットワークの性能向上の速度は、プロセッサの性能向上の速度を大きく上回っており、このままではシステムは大規模化する一方である。

本研究では、個々のサーバ自体の性能を向上させることを目的として、オペレーティングシステム技術に着目して今後の超高速ネットワーク時代を支えるシステム技術の開発を行った。具体的には、サーバにおいて処理される多数のソケットに関するイベントを一括管理する多重化I/Oがボトルネックになることに着目し、その機能的なフレームワークを変更することなく、性能のスケーラビリティを改善する技術を開発した。また、実時間スケジューリングを応用することにより、システム負荷に応じたプロセッサ利用率の実現も達成した。これらの手法は、従来のオペレーティングシステムへの変更も不要であり、適応が非常に容易であるという利点も持つ。

1．研究のねらい

インターネットにおける従来の情報配信サービスでは、ボトルネックは低速で高価なネットワークにあるとされ、そのような状況を改善するために配信の最適化技術が数多く開発されてきた。その代表として挙げられるのがキャッシュ技術であり、現在広く普及しつつあるCDN（Contents Delivery Network）サービスもその一例である。

しかしながら、ネットワーク技術の発展は当初の予想を遥かに越えて進んでおり、バックボーンネットワークにおける性能（スループット）は6ヶ月に2倍向上しているという報告があるほどである。そのため、これまでのネットワークサービスインフラストラクチャの構造に多くの歪みが発生してきている。例えば、先に挙げたCDNなどは、当初の目的であった分散化によるネットワーク負荷の軽減という意義が薄れる一方で、一時的なリクエスト集中によるシステムダウンの回避や、経路切断など故障からのサービスの保護、さらにはDoS

攻撃などからのサービスの防御など、目的の多様化が進みつつある。

本研究では、そうした「歪み」が最も顕著に現れる場所としてエンドノードすなわちサーバに着目し、超高速ネットワークを支えるサーバのシステム技術に焦点を当てて研究を行った。

2. 研究方法と成果

2.1 多重化I/Oの実行間隔制御

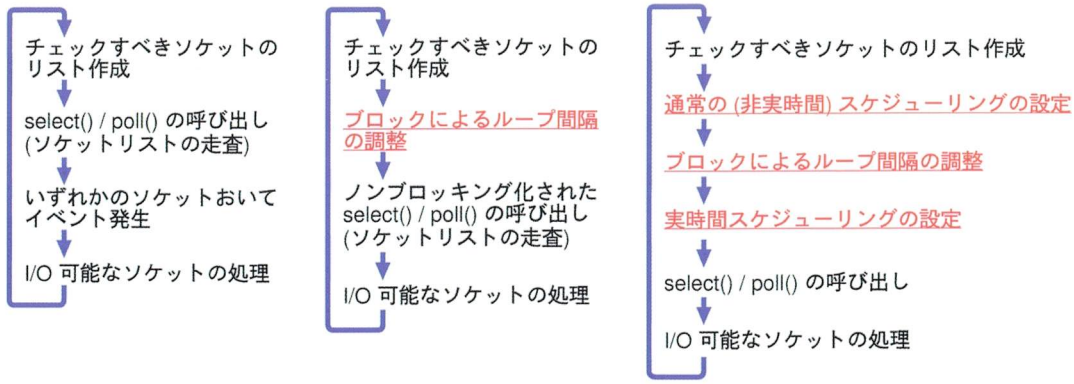
高速ネットワークサーバでは、数千から数万ものソケットを同時に扱うことができなければならない。特に現在代表的なネットワークサービスの一つであるWebにおいては、HTTP/1.1永続コネクションが導入されたため、サーバにおける同時ソケット数が増加する傾向にある。

Unix上のサーバプログラムなどで、このような同時ソケットにおけるI/O処理を多重化するのによく用いられるselect()やpoll()（多重化I/O）には、サーバ負荷の増加に対する性能のスケーラビリティに欠けるという問題がある。この問題の原因は、select()やpoll()におけるソケットテーブルの走査の処理コストが大きいことにあると広く認識されており、それゆえこれまでに提案されてきた解決手法は、こうしたソケットテーブルの走査を廃止し、特別なイベント通知機構を設けるものが多い。しかし、これらの手法は、オペレーティングシステムの改造が必要であったり、プログラミングモデルの変更を要したりするため、導入コストが高いという別の問題がある。多重化I/Oにおいて真に問題なのは、ソケットテーブルの走査そのものではなく、多重化I/Oがそのイベント駆動的な処理構造により必要以上に頻繁に呼び出されてしまうことにある。

そこで本研究では、多重化I/Oの呼び出し間隔を制御し、サーバの性能を向上させる手法を提案した（図1b）。本手法により、高頻度の多重化I/O呼び出しによって引き起こされていたCPU処理能力の枯渇が防止され、サーバの処理能力が向上する（図2）。また、本手法は従来のselect()やpoll()を用いたプログラミングモデルを踏襲するため、適用コストが非常に小さいという特長も併せ持つ。

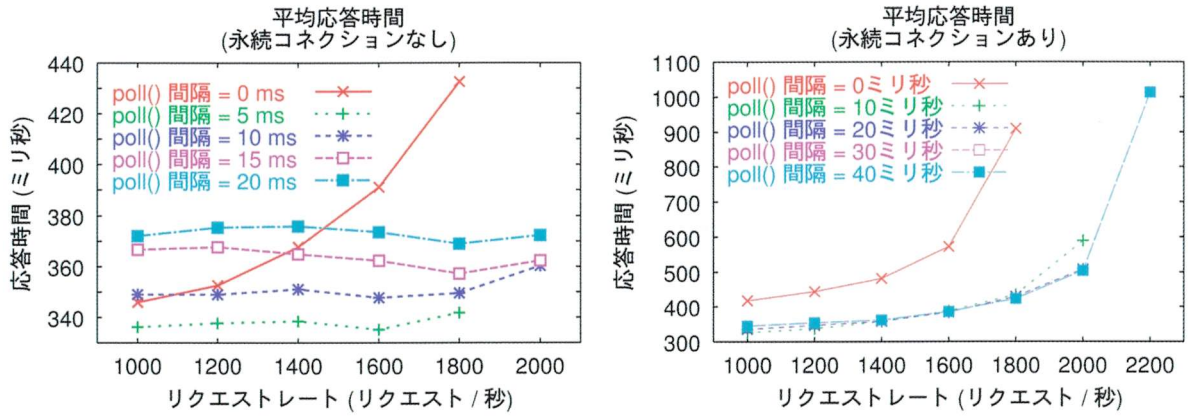
2.2 実時間スケジューリングによる実行間隔制御における確定的なプロセッサ利用の実現

本研究で開発した多重化I/Oにおける実行間隔制御機構において、同時ソケット数が非常に大きい場合、サービス遅延時間の低減などの効果は確認できるものの、いくつかの特異な



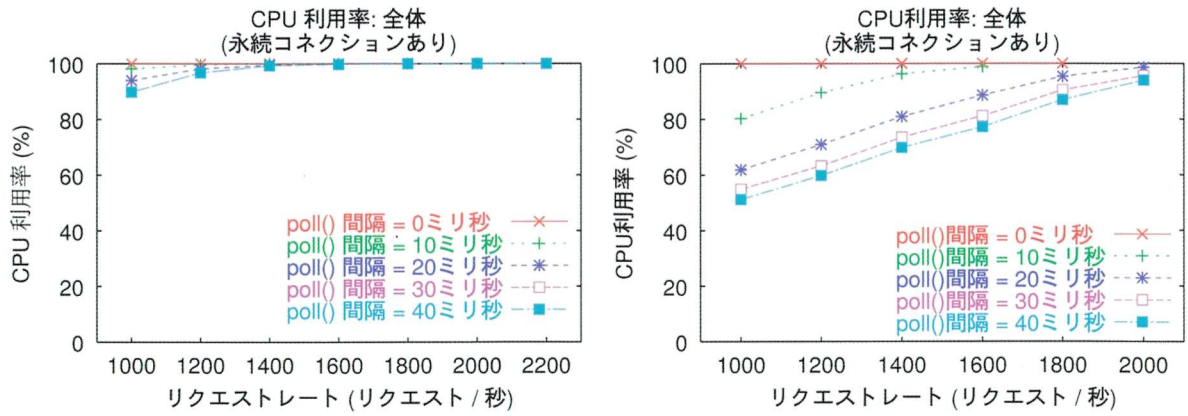
(a) 従来方式　　(b) 提案方式　　(c) 実時間スケジューリングの応用

図１：多重化I/Oにおける実行間隔制御



(a) 永続コネクション無効時　　(b) 永続コネクション有効時

図２：Webアクセラレータに実装した場合の性能
(poll()間隔が0の場合は実行間隔制御をまったく行わない)



(a) 永続コネクション無効時　　(b) 永続コネクション有効時

図３：Webアクセラレータに実装した場合のプロセッサ利用率
(poll()間隔が0の場合は実行間隔制御を行わない)

現象が二点見られた。

一つは、サービス遅延時間の増加傾向である。実行間隔制御を行う場合、一定間隔でソケットのチェックが行われるため、リクエストレートの増加に対してサービス遅延時間はほぼ一定で推移すると考えるのが妥当である。しかし、実験では同時ソケット数が大きい場合は、提案方式を組み込まない場合と比較して大幅にサービス遅延時間の低減を実現しているものの、サービス遅延時間に増加の傾向が観測された（図２b)。

もう一つは、プロセッサの利用率がほぼ100%となってしまう点である（図３a)。実行間隔制御では、リクエストレートに応じてスレッドをブロックするため、プロセッサの利用率はリクエストレートの増加に対して線形に増加すると考えるのが妥当である。

これらの点を考察した結果、制御を行う間隔がオペレーティングシステムのスケジューリングにおけるタイムスライスを越えてしまい、予期しないコンテキストスイッチ等が含まれてしまうことが判明した。そこで本研究では、実時間スケジューリングを用いることでこれらのコンテキストスイッチを防止する手法を提案した（図１c)。本方式により、ソケット数が非常に多い場合でも、サービス遅延におけるスケーラビリティが向上した。また、プロセッサ利用率もリクエストレートに対して線形に推移するようになった（図３b)。後者の利点は、特に近年プロセッサの消費電力が増加しているため、データセンターなどにおける大規模PCサーバクラスタ等で大きな利点になると考える。

２．３　その他の研究

本さきがけ研究は、超高速ネットワークI/Oオペレーティングシステムと題してのプロジェクトであったが、その他にもいくつかネットワークサービスに関連するいくつかの研究を行った。本節ではそれらについて簡単に紹介程度にまとめる。

１．ネットワーク経由のデバイス制御手法の研究

ネットワーク環境の遍在化により、ユビキタスコンピューティング技術に代表されるような、いつでもどこでも計算機を利用可能にする技術が着目を集めている。本研究では、特にUSBで接続されるデバイスに着目し、ネットワーク経由で制御するための基本的なフレームワークを開発した。現在のインターネットは高速化してきているが、デバイスを遠隔から制御するのに十分な品質等が確保されているとは言えない。そこで、デバイス制御に関するネットワークにおける要求分析を行い、最適な通信方式、それらの自動的な選択方式に関する研究を行った。

２．ネットワークにおける新しい情報発信・共有モデルの研究

現在、ネットワークを用いた情報システムが広く利用されているが、それらの中で、特にコミュニケーションを主眼として開発されたシステムを用いて情報等の蓄積等が行われている。

例えば、メーリングリストやWeb掲示板等がそうしたシステムにあたるが、これまでのシステムは情報発信者に強いコミュニティ的意識を強いるものがほとんどであった。そのため、多くの場合で、ごく一部の者だけが情報発信し、その他の大勢は情報を受信するだけという状況が形成される。

本研究では、より個人的な視点に立った、情報発信・共有モデルを提案し、システム開発を行った。そこでは、基本的には個人的なメモ蓄積アプリケーションという形をとり、完全に個人の制御下で共有知識として流通させることで、情報発信者の心理的負荷の軽減を実現している。

3．WWWサーバクラスタを用いたリアルタイム情報発信技術の研究

インターネットが情報流通インフラとしての地位を獲得していくにつれ、スポーツのスコア情報や金融情報など、リアルタイム性の高い情報を大規模に配信する必要性が高まっている。本研究では、大規模サーバクラスタにおけるこうした秒単位で更新されていく情報を同期して配信するシステムを開発した。具体的には、更新情報を管理するマスターサーバと、実際にクライアントへ情報を配信するスレーブサーバの間で、更新・確認・公開という三つのフェーズをもつプロトコルを開発した。本システムは、実際に商用サービスにおいて実証実験を行い、良好な結果を得ている。

4．WWWサービスのIPv6移行技術の研究

IPv6は日本が主導して普及に取り組んでいる技術の一つであるが、IPv4からの移行の際のコストや、移行後のサービスの欠如などの問題がある。特に、サービス品質、費用対効果、セキュリティなどに非常に敏感である商用サービスなどでは、こうした問題が重要視され、なかなか移行が進んでいないのも事実である。

本研究では、商用Webサービスに焦点を当て、そのIPv6移行に必要な要件を分析し、リバースプロキシサーバを用いた低コストなIPv6サービスフレームワークを提案した。本システムで主に開発したのは、IPv6－IPv4中継機能、高い処理能力を達成するメモリキャッシュ、IPv6に未対応なサービスを一元的に管理するサービスフィルタである。また、本システムは、既存のIPv4ネットワークセキュリティのフレームワークへの組み込みも容易であるという特長も持つ。また、本システムも実際の商用サービスにおいて実証実験を行っている。

5．P2P型情報配信技術の研究

今後のネットワークの高速化を考えると、サーバシステムのクラスタ化による負荷分散方式から、P2Pに代表されるような、ユーザが利用する計算機を含めたサービスフレームワークが必要になると考えられる。本研究では、P2P型の情報配信網を用いたWebキャッシュシステムを提案した。

従来より、Webキャッシュシステムはキャッシュヒット率でその性能が議論されていた。そのため、Webキャッシュの目標としては、多くアクセスされる、すなわちより高頻度でキャッシュヒットするコンテンツをいかに効率よく保持しておくかが焦点となっていた。しかしながら、近年のネットワーク高速化により、多くの場合キャッシュがエンドユーザにおけるサービス遅延の解消にあまり貢献していないばかりか、逆にキャッシュが性能のボトルネックになってしまう現象まで観測されるようになっている。そこで本研究では、発想の転換を行い、ダウンロードに非常に長い時間を要するコンテンツのみをP2P接続された分散型キャッシュクラスタからダウンロードする機構を開発した。本システムは、従来のシステムではヒットしにくかったサイズの大きなコンテンツや、アクセス頻度の低いコンテンツなどへのアクセスを改善した。

3．今後の展開

本研究は、さきがけ研究21の枠組みの中では一旦終了するが、これからも高速ネットワークをサービス支えるシステム技術を継続して行きたいと考えている。今後は、以下に掲げる二つの方向性で研究を発展させる予定である。

1．高速イベント処理機構の再構築

これまで、高速ネットワークにおけるI/Oならびにイベント処理の効率化を実現するために、その原因の究明と解決法の提案、技術開発を行ってきた。今後はこうした要素技術を体系化し、高速ネットワークサーバプラットホームとしてのオペレーティングシステムの機構的再構築を行っていきたいと考えている。特にTCP/IPの高速ネットワークへの適応化に関する研究や、ソケットインタフェースなどのプログラミング的な側面など、総合的な検討を行う。

2．サーバプラットホームにおけるネットワークプロセッサの応用

近年、ネットワーク処理専用のハードウエアとして、ネットワークプロセッサ（NP）技術が注目を集めている。今後、個々のサーバの性能を大幅に向上させるには、オペレーティングシステムにおける構造的な改善だけではなく、このNPを用いた高速ネッ

トワークサービスの実現が必須であると考えている。そこで、本研究で得られた知見をこうしたNPを用いたサーバ技術に応用し成果展開していくことを考えている。

4．成果リスト

招待講演

2001年6月　大規模Webサーバの設計・実装・運用，NETWORLD + INTERNET Tokyo 2001

2003年6月　オペレーティングシステムからみた高速ネットワークサービスの実現，NETWORLD + INTERNET (N + I) Tokyo 2003

論　　文

1．河合栄治，門林雄基．山口　英．POSIX実時間スケジューリングを用いた高負荷サーバのスケーラビリティ改善手法（投稿中）．

2．河合栄治，門林雄基，山口　英．ネットワークサーバにおける多重化I/Oの実行間隔制御による性能向上手法．情報処理学会論文誌，情報処理学会（2004年2月掲載予定）．

3．河合栄治，門林雄基，山口　英．ネットワークプロセッサ技術の研究開発動向．情報処理学会論文誌 コンピューティングシステム，情報処理学会（2003年10月掲載予定）．

4．河合栄治，白波瀬章，塚田清志，山口　英．商用WWWサービスのIPv6への現実的な移行手法．情報処理学会論文誌，情報処理学会，Vol.44, No. 3 , Mar. 2003.

5．西馬一郎，河合栄治，知念賢一，山口　英，山本平一．通知によるコンテンツ一斉公開機構を用いたWWWクラスタシステム．情報処理学会論文誌，情報処理学会，Vol.43, No.11, pp.3439-3447, Nov. 2002.

口頭発表（国際会議）

1．Eiji Kawai, Youki Kadobayashi, and Suguru Yamaguchi. Improving Scalability of Processor Utilization on Heavily-Loaded Servers with Real-Time Scheduling.（投稿中）

2．Eiji Kawai, Youki Kadobayashi, and Suguru Yamaguchi. Efficient Network I/O Polling with Fine-Grained Interval Control. International Conference on Communication, Internet, and Information Technology (CIIT 2003), Scottsdale, AZ, USA, November, 2003.

3．Eiji Kawai, Akira Shirahase, Kiyoshi Tsukada, and Suguru Yamaguchi. Practical Migration Strategy to IPv6 for Enterprise Web Services. The 11th International World Wide Web Conference, May, 2002.

その他3件（計6件）

口頭発表（国内研究会等）

1．河合栄治，門林雄基，山口　英．ネットワークプロセッサ技術に関するサーベイ．電子情報通信学会IA研究会，2003年5月．

2．河合栄治，門林雄基，山口　英．多重化I/Oの実行間隔制御におけるスケジュール操作による確

定的なプロセッサ利用の実現．情報処理学会 システムソフトウエアとオペレーティングシステム研究会，2003年５月．

3．河合栄治，門林雄基，山口　英．多重化I/Oの実行間隔制御による効率化手法．日本ソフトウエア科学会，SPA2003, 2003年３月．

4．河合栄治，白波瀬章，塚田清志，山口　英．商用WWWサービスの IPv6環境移行技術の研究．情報処理学会 マルチメディア通信と分散処理研究会，2002年３月．

その他８件（計12件）