研究課題別評価

1. 研究課題名：超高速 I/O 指向オペレーティングシステム

2. 研究者氏名：河合　栄治

3. 研究の狙い

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。特に、現在のオペレーティングシステムアーキテクチャがその I/O 処理機構において、同期的、すなわちイベント駆動的な構造を用いているために高いオーバヘッドが生じている点に着目し、非同期性を導入することによりこれからの超高速ネットワークに対応したサーバプラットホームを創出することを目的とした。

4. 研究結果

本研究では、高負荷のかかるネットワークサーバにおいて、クライアントからの多数のコネクションを管理する多重化 I/O と呼ばれる機構の高速化を中心に、開発を行った。主要な成果は下記の 2 点である。

(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。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。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。

⑵　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。

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。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。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。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。

5. 自己評価

3年間のさきがけ研究を振り返って、ネットワークサーバの I/O 機構に非同期性を導入によることによる性能向上という目標は達成されたと考えている。さらに私が提案した手法はサーバソフトウエアおよびオペレーティングシステムへの変更が少なくて済み、導入コストが小さく現実的であるという利点も持っている。また、サーバアーキテクチャの検討という基礎研究的な側面を持ちつつ、実際の商用サービスを含む数多くの運用現場での実証実験を通じ、使える技術の開発ができたことは、本研究の大きな特長である。

一方で、当初掲げていた目標の中で達成できなかったものに、マルチスレッド環境における I/O 処理量に着目したスケジューリングによるサーバ性能向上手法がある。本目標については、実際にスケジューラを構築し、性能評価を行ったが、最終性能的に 5%程度の向上しか観測されなかった。商用サービスに見られる大規模アプリケーションサーバのように、多種多様なスレッドが多数動作する環境で評価ができれば、大きな効果が見られたのではないかと予想している。その意味では、近年の大規模化したサービスプラットホームを対象とした研究において、実験環境構築の難しさを痛感した。

本研究では、その成果物として、Chamomile と名付けた Web Accelerator を開発した。本ソフトウェアは ISP や通信機器ベンダー等からも問い合わせを受けた。今後も開発を進め、ドキュメント等を整備し、フリーソフトウエアとしての公開を目指していきたいと考えている。

ネットワークの高速化の流れは現在も加速中であり、今後はホスト内におけるハードウエア的な分散処理を視野に入れた基盤の構築が必要となると考えている。特にプロセッサ技術では従来の SMP

技術に加え、SMT 技術が広く利用可能になってきている。また、ネットワークプロセッサのような ネットワークインタフェースに近い場所に処理能力を確保する技術も登場してきている。そのため、これらの多様な処理要素の容易かつ効率的な利用を可能にするような機構を構築していく必要があるだろう。これらの技術と、現在盛んに開発が進められているサーバクラスタリング技術を融合し、統一的な真の分散サービスプラットホームを構築していきたい。

6. 研究総括の見解

超高速ネットワークを利用したインターネットによる情報配信サービスが社会インフラとなっている現在、その効率化が大きな問題となっている。河合研究者は、サーバに搭載されているオペレーティングシステムの I/O 処理機構が同期的であることにより高いオーバヘッドが生じている点に着目し、非同期性の導入によりこの問題を解決する方法に挑戦した。その結果、CPU 処理能力の枯渇防止のための多重化 I/O 実行間隔制御、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という、2つの技術を開発することにより目標を達成することが出来た。また、具体的な成果物として、Chamomile と名付けた Web Accelerator を開発したが、これは ISP や通信機器ベンダー等からも高い評価を受けており、河合研究者は目標とした研究課題を現実的な形でも解決した。

7. 主な論文等

招待講演

1.高速ネットワークサービスを実現するオペレーティングシステム, NETWORLD + INTERNET Tokyo 2001, 2001 年 6 月.

2.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, 2003 年 6 月

論文

1.Eiji Kawai, Youki Kadobayashi, and Suguru Yamaguchi. Alleviation of Processor Usage on Heavily-Loaded Network Servers with POSIX Real-time Scheduling Control. (Submitted to IEICE Transactions)

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, Vol.45, No.2, 2004 年 2 月.

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, Vol.45, No. SIG1 (ACS4), 2004 年 1 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 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, 2003 年 3 月.

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, 2002 年 11 月.

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. International Conference on Parallel and

Distributed Computing and Networks (PDCN 2004), Innsbruck, Austria, February, 2004.

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 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）