論文審査の要旨及び担当者

No.1

報告番号	甲 乙 第 号	氏 名	松谷 健史

論文審査担当者	主 査	政策・メディア研究科委員	兼環境情報学部教授 村井 純
	副 査	政策・メディア研究科委員	兼環境情報学部教授 中村 修
		政策・メディア研究科委員	兼環境情報学部准教授 バンミーター, ロドニー
		理工学部教授	天野 英晴
学力確認担当者：			

（論文審査の要旨）

松谷健史君提出の学位請求論文は「IP (Internet Protocol) を用いた低遅延分散アーキテクチャ」と題し、7 章からなる。

本研究は、現在のインターネットサービスにおいて非常に重要な拠点であるデータセンタ内での通信性能を飛躍的に向上するための研究開発である。本研究では、一般のネットワークインターフェイスカード（以下 NIC と略す）やスーパーコンピュータで用いられる RDMA 方式の NIC の仕組みを詳細に解析し、特に遅延に関する通信性能に大きな影響を及ぼす構造の分析を行った。そして、その結果に基づき、標準的な遅延時間をはるかに超える低遅延な通信を実現するための手法をソフトウェアとハードウェアの両面において研究開発を行い、実証した。

本研究で提案した IP-NUMA 方式を採用した NIC を、コモディティ（市販の）FPGA ボードを用いて実装した。そのためのソフトウェアは Linux OS 上に構築し、10Gbps クラスの実ネットワーク上で実験をおこない、その有用性を示した。また、実際のアプリケーションへの適応例として memcached を用いた場合のシミュレーションをおこないその実効性を明らかにした。

現在のデータセンタにおけるネットワークへの要求は転送速度（スループット）の向上だけでなく、サーバの処理能力の向上（処理時間の削減）が要求されている。Hard Disk の衰退と Flash Disk や Persistent memory の躍進によりアプリケーションの処理時間は大幅に削減できているが、相対的にデータセンタ内における内部的な通信処理時間が多くを占めるようになってきている。プロトコルスタックを含め通信処理の多くは本体コンピュータのソフトウェアによって処理されるが、本研究ではこれらの通信処理を NIC 上のハードウェアで実現することで、ソフトウェアの処理時間を大幅に削減した、IP プロトコルを使った低遅延通信“IP-NUMA 方式”を提案し、これを開発した。“IP-NUMA 方式“を用いることにより、一般の NIC を用いた通信に対して通信遅延時間を 90%以上削減した。また、低遅延通信で用いられる、RDMA 方式と比べても 11%以上の削減を可能とした。

本論文の構成は、まず、第 1 章で研究背景と研究全体について概観したのち、第 2 章で、関連技術とその問題点について論じた。第 3 章では、IP-NUMA アーキテクチャについて述べ、特に，既存の NIC のハードウェアによって生じる処理遅延を調べるために、PCI Express のコマンドを分析し、PCI Express コマンドの実行時間を計測するために、FPGA 上で動作する PCI Express バス測定回路とソフトウェアを実装し、遅延の要因が DMA リードにあることを明らかにした。この解析結果をもとに、”IP-NUMA 方式“では、DMA リードの代わりに CPU ライトを用いる手法を提案した。

第 4 章では、実際に IP-NUMA 方式の実装と評価について論じている。ハードウェアの実装は PCI Express とイーサネットインターフェイスを持つ FPGA を用い、ソフトウェアは Linux カーネル上で動作するドライバとライブラリを開発することで実装している。
評価には、市販 NIC と IP-NUMA NIC を用い、２つの PC 間で pinpong プログラムを走らせることで RTT(ラウンドトリップタイム) を計測した。
10 ギガビットイーサネット環境では、RTT が市販 NIC において約 22μs なのに対し、IP-NUMA NIC では約 2μs となり、通信遅延を一桁削減することに成功した。
RDMA 方式においてユーザ回路を含まない理論上の最小 RTT に対しても約 10%を超える遅延時間の削減ができることを示した。

論文の前半の第 4 章までは、ホスト端末における IP プロトコルを用いた低遅延通信の為の方式について論じたが、後半では、データセンタ内のネットワーク環境における低遅延に関する研究について論じている。

低遅延通信を考慮した IP ネットワークを設計する場合、L3 スイッチにおける IP パケット転送遅延時間を把握しておく必要がある。第 5 章では、転送遅延に着目し、L3 スイッチが IP パケットを転送する場合に必要な工程を詳細に分析し、各工程の処理と最小転送遅延時間を明らかにした。第 5 章では、この分析に基づいた L3 スイッチのハードウェアを FPGA を用いて実装した(FIBNIC スイッチ)。
4 ポートギガビットイーサネットのためのインターフェースを持つ FPGA を用いて FIBNIC を実装し、評価したところ、市販の L3 スイッチで 64 バイトパケットの転送遅延時間が約 4ms だったのに対して、今回実装した FIBNIC スイッチでは、パケットサイズに関わらず、転送遅延時間が 1ms 以下で転送できることが確認され、本実装が有効であることを示した。

第 6 章では、第 1 章から第 4 章で説明したホスト端末内での低遅延手法である IP-NUMA 方式と第 5 章で説明した L3 スイッチの低遅延手法である FIBNIC を組み合わせて、データセンタ内ネットワーク上でアプリケーションを利用した場合を想定し、シミュレーションによって評価をおこなっている。
　評価アプリケーションには, Facebook や Twitter などでも利用されている分散型メモリキャッシュの memcached を用いた。シミュレーションには、まず memcached のベンチマークツールを用いてパケットをキャプチャしたものを分析し、第 4 章の計測によってあきらかにした IP-NUM 方式による転送遅延時間と、第 5 章で計測した市販 L3 スイッチと FIBCNIC スイッチの IP 転送遅延時間を用いた。その結果、ラック内の接続においては、市販 NIC と市販スイッチの組み合わせでは 16,981 回/秒のトランザクション数だったのに対して、本手法である IP-NUMA 方式と FIBNIC スイッチの組み合わせでは 73,384 回/秒となり、約 4 倍の性能向上となった。
また複数の L3 スイッチによって接続されるラック間で接続においては、IP-NUMA 方式と FIBNIC スイッチの組み合わせではトランザクション数として約 4 倍の性能向となった。ラック間接続においては、市販 NIC を使いながらスイッチのみを本手法である FIBNIC にするだけで、1.41 倍の性能向上となり、L3 スイッチによる IP 転送遅延が大きいラック間接続では L3 スイッチの低遅延化が有効であることが示された。

第 7 章で議論をまとめている。

現在、クラウドサービスの普及により、アプリケーションの処理はデータセンタ内の膨大なサーバに集中するようになり、データセンタ内に設置されているサーバ間を接続するネットワークではスループットだけでなく、低遅延通信が求められている。そのためデータセンタによっては、主にスーパーコンピュータ用に開発された RDMA NIC を利用している例もある。

本論文では、NIC の仕組みと PCI Express の処理を詳細に分析することにより、DMA リードを用いないことで RDMA 方式より低遅延通信が可能な IP-NUMA 方式を提案した。
本研究では、市販の FPGA 上にこの IP-NUMA 方式を実装し、評価をおこなった。その結果、IP-NUMA 方式は、通常の NIC を用いた通信に対しては、一桁早い低遅延通信を実現し、RDMA に比べても約 10%の低遅延通信を可能とした。
また、本論文では、ホスト端末における遅延時間削減だけではなく、L3 スイッチのパケット転送における遅延時間に関する詳細な考察をおこない、FPGA を用いた L3 スイッチとして FIBNIC の開発をおこない、データセンタ内の高効率なネットワーク環境を提案した。実際に Facebook や Twitter などでも利用されている memcached を用いて、本論文で提案した IP-NUMA 方式および FIBNIC を用いた場合の性能をシミュレートした結果、処理可能なトランザクション数を約 4 倍にすることが可能な性能の向上ができることを明らかにした。

上記の成果と、それを記述した本論文を通して、著者の先端研究成果は次世代の同様分野の研究者に対しての大きな示唆と勇気を与えた。工学分野の研究を社会貢献に結びつける能力を有することを示したものといえる。よって、本委員会は、本論文の著者は、博士 (政策・メディア)の学位を受ける資格のあるものと認める。