光無線によるビッグデータ処理向け相互結合網の研究開発（152103004）

Research and Development of Interconnection Networks for Bigdata Processing using Optical Wireless

研究代表者

鯉渕 道紘　国立情報学研究所

Michihiro Koibuchi　National Institute of Informatics

研究分担者

松谷 宏紀 †　山田 浩史 ††　胡 曜 †††

Hiroki Matsutani†　Hiroshi Yamada††　Yao Hu†††

†慶應義塾大学　††東京農工大学　†††国立情報学研究所

†Keio University　††Tokyo University of Agriculture and Technology　†††National Institute of Informatics

研究期間　平成 27 年度～平成 29 年度

概要

ビッグデータ処理において、並列処理の結果を計算機間でやり取りするための通信待ち時間を短縮するため、光無線を用いて個別に最適化可能な相互結合網を構築し、数千～数万並列で実行するビッグデータアプリケーション性能を飛躍的に向上させる基盤技術を開発した。具体的には、スーパーコンピュータやハイエンドなデータセンターにおいて、フェーズ I では光無線イーサネットによるコンピュータノード間通信、フェーズ II では光無線ストレージネットワークを実現することで、ビッグデータアプリケーションとネットワークのコデザインによる最適化を行い、同じ計算機群を用いて倍以上の実行時間の向上を達成した。そして、ビッグデータ処理に適した数十～数百ラック規模のラック間ケーブルレスデータセンターの設計法を示した。また、2 台のホストで構成されたシステムにて光無線通信稼動率 99%以上を達成した。

１．まえがき

行列計算などの古典的な科学技術並列演算は、スーパーコンピュータ(以後、スパコンと呼ぶ)や汎用プロセッサの持つ計算能力の 60～90%以上の実効性能で計算できる。一方、ビッグデータ並列処理では、京コンピュータ上でのチューニング済み巨大グラフ解析(Graph500)において通信待ち時間が実行時間の 70%を占める報告があるなど、現状のスパコンやデータセンターのネットワーク(以後、相互結合網と呼ぶ)の限界が顕著となっている。

本研究では、光無線(Free Space Optics: FSO)リンク技術の導入により、個別に最適化可能な相互結合網プラットフォームを実現する。データセンター向けの計算機システムであるラックスケールコンピュータでは、メモリ間、ストレージ間すべてが光通信技術により密に結合される歴史的転換点を迎えつつある。本研究では、これら最先端テクノロジに光無線を適用することで、並列アプリケーションに適したプロセッサ、メモリ、ストレージ間の相互接続を動的に確立し、光無線リンク以外同じ構成を取る計算システムに対し大幅な性能向上を達成する。

我々の最終的な狙いは、HP、Intel、Facebook、Microsoft などが設計を進めているデータセンター向け計算機システムであるラックスケールコンピュータへの光無線通信技術の導入である。

以上より、(1)メモリープロセッサ間通信がバスからパケットネットワークに変わり、(2)ストレージを PCI-Express バスに設置することが主流となる相互結合網の歴史的転換点において、最先端の構成要素を柔軟に接続可能とする光無線技術を中心にすえたスパコン、データセンター相互結合網を実現する。

２．研究開発内容及び成果

フェーズ I では 40Gbps 光無線イーサネットを対象にした技術開発に基づき、「光技術と高性能コンピューティング」の招待レビュー論文(電子情報通信学会 ELEX 誌)に今後の展望をまとめた。本論文は 2017 年 7 月の ELEX Top 10 downloads となるなど一定の注目を浴びた。また、フェーズ II ではフェーズ I での知見を元に、ラック単位でプロセッサ、ストレージ、GPU を設置し、それらを光無線で直結する光無線ハイブリッド相互結合網を探求した。

2.1 フェーズ I

1 次元データアクセスを行うビッグデータアプリケーションが多いことから、ネットワークトポロジの直径、平均ホップ数を最小化するようにネットワーク構成を取ることが有望である。その観点で最適化したネットワーク構成を示した。スパコンやデータセンターへの導入が検討されている高次元トーラストポロジと比べて、検討を進めたグラフに基づくネットワークトポロジを用いることで、千プロセス並列以上のグラフ処理やデータ並列ソートベンチマーク性能をイベントドリブンシミュレータ SimGrid により評価し、現状比 250%と劇的に向上させた。

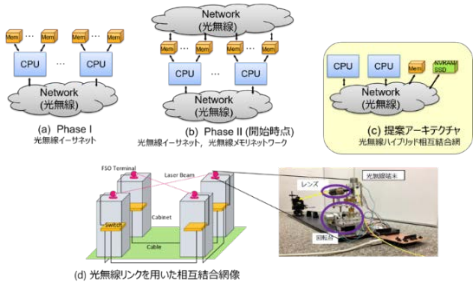

図 1:光無線リンクを用いた将来のデータセンター

プリケーションがデータセンターやスパコンでは同時に動作している。そのため並列アプリケーションを「どの計算ノードに割り当てるか？」というジョブマッピングを光無

線相互結合網向けに改良することが、システム利用効率向上につながる。そこで、マッピングの仕方を工夫することで(未使用の計算ノードが分散することで大きなジョブが実行できなくなるというフラグメンテーションの削減により)システム利用効率を光無線導入前と比べて 20%向上させることに成功した。

2.2 フェーズ II

すべてのコンピュータの構成要素(CPU、GPU、メモリ、NVM/SSD ストレージ)を相互接続する光無線ハイブリッド相互結合網を提唱した(図 1(c))。 (i) 光無線ハイブリッド相互結合網のテストベッドを用いたリモート SSD への読み書き性能とリモート GPU(NVIDIA GTX980Ti)への CUDA カーナル実行の現状比最大 3 倍高速化を達成し、(ii) 光無線を活用したジョブスケジューリング法と VM ライブマイグレーション法を専用に開発した。つまり、光無線ハイブリッド相互結合網という、提案時よりも一歩踏み込んだビジョンのもと、開発目標であるアプリケーションの性能向上とシステム利用率向上を得ることができた。具体的には、単一ラックもしくは複数ラックに各種計算コンポーネント(GPU、ストレージなど)が分散配置されたラックスケールコンピュータを提唱し文書指向データベースの MongoDB の評価などを実機で行い、光無線リンクを直結させることで大幅な処理性能の向上を達成した。

システムソフトウェアに関しては以下の研究開発を行った。現状の 10Gbps イーサネットを用いたデータセンターにおいて、ファイルリード中の VM イメージを移動する場合、空白時間が大きいため差分が生じ、マイグレーションが完了せず非効率なことが多い。そこで、ファイルリードを頻繁に行うビッグデータ処理においてキーコンポーネントとなるデータベース管理システム(DBMS)が稼動する VM の移送技術を開発した。DBMS を稼動する VM は大量のメモリを必要とするため、メモリ転送のみに頼る従来の移送方式では、移送時間が長期化しがちである。そこで、メモリ転送だけでなく、光無線リンクによってストレージノードと動的にリンクを形成し、VM に必要なデータ転送も同時に行う。初期プロトタイプにて最大で 20%の移送時間の削減効果を得られた。

３．今後の研究開発成果の展開及び波及効果創出への取り組み

我々は光無線リンク端末の安定性を示し、これらを用いた大規模計算機システムにおけるジョブマッピング、スケジューリング、仮想マシンのマイグレーション、ネットワーク再構成法などを開発した。これらは光無線データセンターネットワークのパッケージングとして提供し、技術移転できるように準備を進めている。我々は国立情報学研究所オープンハウス 2016、2017 におけるデモ展示、JST 新技術説明会(2016 年 11 月 17 日)(https://shingi.jst.go.jp/list/nii/2016_nii.html)、第 18 回慶應科学技術展 KEIO TECHNO-MALL2016、2017 などで積極的に実用化の可能性を探求した。光無線端末の回転部の設計でコスト問題があるため実用化の目処はたっていないが、光無線技術が提供する「ビッグデータアプリケーション支援のためのネットワーク再構成」と「ケーブルレス」という光無線の特長を訴え、引き続き積極的な活動を続ける予定である。

４．むすび

ビッグデータ処理において、並列処理の結果を計算機間でやり取りするための通信待ち時間を短縮するため、光無線を用いて個別に最適化可能な相互結合網の技術開発を行った。これらは単なるネットワーク構成の研究開発に留まらず、光無線を活用した VM マイグレーション手法にいたるソフトウェア・ハードウェアコデザインを生み、査読付き論文 3 件、査読付き口頭発表論文 11 件、口頭発表 17 件、受賞 5 件、特許申請 1 件と積極的に成果発表を行った。

【誌上発表リスト】

[1]Thao Nguyen Truong, Ikki Fujiwara, Michihiro Koibuchi, Khanh Van Nguyen, “Distributed Shortcut Networks: Low-latency Low-degree Non-random Topologies Targeting the Diameter & Cable Length Trade-off”, IEEE Transactions on Parallel & Distributed Systems, 28(4) 989-1001, (2017 年 4 月)

[2]Yao Hu, Ikki Fujiwara, Michihiro Koibuchi, “Job Mapping and Scheduling on Free-Space Optical Networks”, IEICE Transactions on Information and Systems, E99-D(11) pp.2694-2704, (2016 年 11 月)

[3] Michihiro Koibuchi, Ikki Fujiwara, Kiyo Ishii, Shu Namiki, Fabien Chaix, Hiroki Matsutani, Hideharu Amano ,Tomohiro Kudoh, “Optical Network Technologies for HPC: Computer-Architects Point of View”, IEICE Electronics Express, ELEX,Vol. 13 (2016) No. 6, 15 ページ(招待論文)(2016 年 6 月)

【申請特許リスト】

[1] 中野 浩嗣、藤田 聡、鯉渕 道紘、藤原 一毅、「計算ノードネットワークシステム」特願 2015-224988、日本、2015 年 11 月 17 日

【受賞リスト】

[1] 鯉渕 道紘、情報処理学会 長尾真記念特別賞、“ラックスケールコンピュータ・ネットワークの設計に関する先駆的な研究”、2016 年 6 月 3 日 (https://www.ipsj.or.jp/award/nagao.html)

[2] 松谷 宏紀、情報処理学会 マイクロソフト情報学研究賞、“チップ内からデータセンタ規模に至るマルチスケールな相互結合網の研究”、2018 年 3 月 13 日

[3] Truong Thao Nguyen, Ikki Fujiwara, Michihiro Koibuchi, The ACM Seventh International Symposium on Information and Communication Technology (SoICT 2016) BEST PAPER RUNNER-UP AWARD, “A Diagonal Cabling Approach to Datacenter and HPC Systems” 2016 年 12 月 8 日

【報道掲載リスト】

[1]“新配線法 スパコンなどに提案”、化学工業新聞 7 面、掲載 2015 年 8 月 7 日

[2] 文教ニュース 第 2421 号 P48.NII と JST 共催の新技術説明会 4 名の研究者が新技術説明 NII 本位田真一教授、小野順貴准教授、鯉渕道紘准教授、坂本一憲助教、2016/11/28（月）

【本研究開発課題を掲載したホームページ】

URL:https://www.youtube.com/watch?v=a-kTUxlWZAw&feature=youtu.be&t=8

URL:http://research.nii.ac.jp/~koibuchi/research06.html