様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月　 ７ 日現在

機関番号：３２６１２
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１８５９

研究課題名（和文）広帯域光通信によるFPGA主導型相互結合網

研究課題名（英文）FPGA-driven Broadband Optical Communication Interconnect Network

研究代表者
胡　曜（HU, Yao）

慶應義塾大学・デジタルメディア・コンテンツ統合研究センター（日吉）・特任助教

研究者番号：５０７９１２３２

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では、入力した通信パターンに合わせた最適なFPGA主導型の光通信トポロジの動的構成法を提案し、ポストムーアにも対応できるデータセンターシステムの設計法を検討した。また、光通信トポロジ動的構成法を開発・活用し、アプリケーション毎に計算ノードを柔軟に分配するスケジューリング手法を提案した。そして、高性能計算ネットワーク上での高い移植性を実現するために、近似通信のアプリケーションレベルの技術も検討した。これにより、並列アプリケーションの通信待ち時間や総実行時間を最小化するとともにシステム全体のスケーラビリティを最大化することが期待できる。

研究成果の学術的意義や社会的意義
本研究で開発したトポロジ動的構成法、データ圧縮法とスパコンスケジューラのプログラムをオープンソースソフトウェアとして公開した。研究過程で得られた知見については、産業界・学術界の技術者・研究者らと幅広い議論を交えながら、研究会・国際会議・論文誌などで発表し、将来のスパコンネットワーク設計や異種資源環境データセンター構築に向けた参考とする。
本研究により、異種資源環境システムにおける大規模計算機ネットワークがそのポテンシャルを十分に発揮することで、ビッグデータ時代のニューラルネットワークにおける多様なワークロード処理の速度や次世代アプリケーション実行性能をより一層向上させることが期待できる。

研究成果の概要（英文）：In this study, we proposed a dynamic configuration method for FPGA-driven optical communication topologies tailored to the input communication patterns, and investigated a design approach for datacenter systems that can also accommodate post-Moore architectures. This enables the reduction of overall power consumption and communication volume in datacenters, as well as the number of hops (communication delay) between computing nodes as the computer systems scale up. In addition, we developed and utilized dynamic configuration methods for optical communication topologies, and proposed a flexible job scheduling method for parallel applications. We also examined application-level techniques for approximate communication to enable high portability on high-performance interconnection networks. This is expected to minimize waiting time and total execution time for parallel applications while maximizing overall system scalability.

研究分野：高性能計算

キーワード：高性能計算　計算機ネットワーク　データ圧縮　タスクスケジューリング

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

人工知能(AI)や 5G 通信による巨大かつ高度なデータ処理では、計算ノード間の通信に膨大な時間を要する。機械学習(ML)やディープラーニング(DL)をはじめとする新世代アプリケーションの非常に速い進化速度に対し、ムーアの法則に従った CPU や GPU などの汎用プロセッサの性能向上が限界に達しつつある。インターネットやクラウドコンピューティングを支えるデータセンターには、より高度な処理と高い性能を求めるために大量の機器がぎっしり詰まっている。その結果、電力消費量が爆発的に増加するとともに、システム全体の利用も非効率な状態になっている。

２．研究の目的

本研究は、将来の広帯域光通信技術を用いた大規模計算機システム(図 1)においてより高効率的な資源利用を実現する。そこで、様々な並列アプリケーションを１つのデータセンターで低遅延かつ効率的にサポートすることが期待できる。

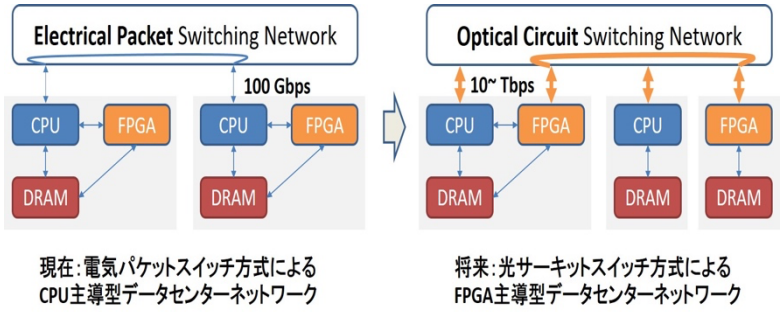

図　1　現在と将来のデータセンターネットワーク

３．研究の方法

(1) 光通信による異種ハードウェアトポロジの動的構成法:パケット競合が起きないように従来データセンターのパケットスイッチと異なる時分割多重(TDM)サーキットスイッチ方式を採用することにより、通信帯域や遅延を保証することができる。そして、入力した通信パターンに合わせた最適な FPGA 主導型の光通信トポロジの動的構成法を提案し、ポストムーアにも対応できるデータセンターシステムの設計法を検討する。これにより、計算機システム大規模化に伴うデータセンター全体の電力消費量や通信量がかさむ計算ノード間のホップ数(通信遅延)を削減することが可能である。

(2) 異種ハードウェア混在環境の大規模計算機システムにおけるジョブマッピングやスケジューリング手法:将来広帯域光通信技術を相互接続ネットワークに用いたデータセンターシステムを想定し、(1) の光通信トポロジ動的構成法を活用し、異種ハードウェアの物理トポロジへの最適なジョブマッピング手法を導き出す。そして、アプリケーション毎に計算ノードを柔軟に分配するスケジューリング手法を提案する。これにより、アプリケーションの通信待ち時間や総実行時間を最小化するとともにシステム全体のスケーラビリティを最大化することが期待できる。

４．研究成果

(1) 対象となる細粒度サーキットスイッチング(FGCS)ネットワークにおいて、シンプルで最小の混雑を把握した MiniCAR 経路手法を提案した。MiniCAR では、データパケットを、送信元から宛先まで混雑しているリンクを避けて適応的に経路を選択する。MiniCAR は、FGCS ネットワークの経路の全体的な情報に依存し、最低限必要なスロット数が FGCS スイッチに増加しないように混雑を回避するように設計されている。評価を通じて、FGCS ネットワークのサイズが増大するにつれて、最低限必要なスロット数が増加することを示した。伝統的な経路アルゴリズムと比較して、MiniCAR は、大規模な 2 次元トーラス FGCS ネットワークにおいて多様な通信パターンを使用して、最低限必要な時間スロット数を最大 49.2%削減することができる。そのため、MiniCAR は並列 FGCS コンピューティングシステムにおける通信遅延を大幅に減少させるのに役立つ。

(2) Hamorder という軽量のリオーダリングアルゴリズムを提案した。これは、ランダムネ

ットワークトポロジー上で、インタータスクとイントラタスクのノードの並べ替えを通じてトポロジー埋め込みの局所性を向上させるものである。従来のグラフの並べ替え手法とは異なり、Hamorder は並列アプリケーションの実行パフォーマンスを向上させることを重視する。結果は、Hamorder が現行の最先端の手法である Gorder を上回り、最大 27.3%の速度向上をもたらすことを示している。さらに、Hamorder を OpenTuner によるオートチューニングフレームワークで活用し、ランダムトポロジー上での並列アプリケーションの実行を最適化した。フレームワークは、最小限の検索試行回数で最大 2.29 倍のスピードアップを達成した。

(3) 近似通信のためのデータ圧縮を最適化することで、並列コンピュータの相互接続ネットワークにおける有効なネットワーク帯域幅を向上させることを可能にする。ネットワークインターフェースでのハードウェア圧縮を使用する代わりに、浮動小数点通信用のアプリケーションレベルの精度保証を前提としたビット単位の圧縮アルゴリズムを開発した。圧縮された浮動小数点値はビットストリームに結合され、バイト配列にカプセル化される。これは MPI の unsigned char 型に対応しており、高いポータビリティを確保している。さらに、伝送中に発生する可能性のあるビットフリップに対する圧縮データのエラーチェックと訂正技術を探求した。評価結果によれば、提案したビット単位の圧縮アルゴリズムは、高性能な相互接続ネットワーク上で特定の誤差限界を保持しつつ、並列アプリケーションの実行性能を向上させる。

(4) ニューラルネットワークに代表される先進アプリケーションに使われるノードコミュニティ検出やランダムウォーク並列化を設計・開発した。単純なネットワーク、複雑なネットワーク、および相互接続されたマルチプレックスネットワークを含むさまざまなネットワークタイプでランダムウォークを実行するために設計された汎用ネットワークモデルを提案した。複数のサーバーにまたがる分散グラフでランダムウォークを実行した結果、提案されたネットワークモデルの効率性を強調するだけでなく、さまざまなシナリオでの並列ランダムウォークの効果を示している。これにより、異種ハードウェア環境でのアプリケーションのマッピングや実行を容易にすることが可能となる。

(5) 本研究で開発したトポロジ動的構成法、データ圧縮法とスパコンスケジューラのプログラムをオープンソースソフトウェアとして公開した。これにより、研究成果を社会に広く還元し、将来の光無線環境データセンターに向けた参考とした。また、研究過程で得られた知見については、産業界・学術界の技術者・研究者らと幅広い議論を交えながら、研究会・国際会議・論文誌などで発表し、将来のスパコンネットワーク設計や異種ハードウェア環境データセンター構築に向けた参考とする。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Yao Hu	４．巻 11
２．論文標題 Accelerating Parallel Applications Based on Graph Reordering for Random Network Topologies	５．発行年 2023年
３．雑誌名 IEEE Access	６．最初と最後の頁 40373～40383
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/access.2023.3269793	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yao Hu	４．巻 11
２．論文標題 Exploring Approximate Communication Using Lossy Bitwise Compression on Interconnection Networks	５．発行年 2023年
３．雑誌名 IEEE Access	６．最初と最後の頁 59238～59249
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/access.2023.3281834	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 胡曜
２．発表標題 並列計算機における自動チューニングを用いた通信最適化
３．学会等名 ETNET2022
４．発表年 2022年

１．発表者名 胡曜
２．発表標題 非可逆圧縮を用いたMPI通信の性能評価
３．学会等名 HotSPA2022
４．発表年 2022年

１．発表者名
Yao Hu

２．発表標題
MiniCAR: Minimal Congestion-Aware Routing Method in Fine-Grained Circuit-Switched Networks for Parallel Computing Systems

３．学会等名
26th IEEE Symposium on Computers and Communications (ISCC 2021)（国際学会）

４．発表年
2021年

１．発表者名
Yao Hu, Qian Huang

２．発表標題
Optimizing Link Prediction by Community Detection on Dynamic Networks

３．学会等名
IEEE 11th International Conference on Information, Communication and Networks (ICICN 2023)（国際学会）

４．発表年
2023年

１．発表者名
Yao Hu

２．発表標題
Parallel Computation of Random Walks on Distributed Graphs

３．学会等名
26th IEEE/ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD2023-Winter)（国際学会）

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	鯉渕　道紘 (Koibuchi Michihiro)		

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関