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　元 年　 ６ 月　 ６ 日現在

機関番号：６２６１５
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１２６８２

研究課題名（和文）光無線によるデータセンターの無駄ゼロ運用

研究課題名（英文）Zero Waste Utilization of Optical Wireless Datacenter

研究代表者
胡　曜（Hu, Yao）

国立情報学研究所・アーキテクチャ科学研究系・特任研究員

研究者番号：５０７９１２３２

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：データセンターにおける各コンピュータ要素のハードウェアをラック単位で別々に集約する新たなIRSアーキテクチャを想定し、動的にFSO光無線通信リンクを利用することでアプリケーション毎にリソースを柔軟に分配し、資源利用率の向上を図るタスク実行トポロジの構成法を提案した。ラック間平均通信ホップ数を抑制するとともに、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で、パケットの平均ホップ数や平均通信遅延を最小化することができることを検証した。また、FSO光無線リンク付きIRSアーキテクチャへのジョブスケジューリング性能の評価を行い、アプリケーション実行時間の削減ができることを検証した。

研究成果の学術的意義や社会的意義
本研究で開発した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、将来の光無線環境データセンターに向けた参考とする。
本研究により、低遅延光無線通信時代にあって大規模計算機システムがそのポテンシャルを十分に発揮することで、ニューラルネットワークに代表される最近のビッグデータ処理速度や並列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）：We investigated the effectiveness and efficiency of an Inter-Rackscale (IRS) datacenter architecture which disaggregates hardware components such as CPU, SSD and GPU into different racks according to their own areas. By introducing FSO (free-space optics) channels for wireless connections between racks to make full use of computing resources with a fine-grained granularity during job allocation, we improved the resource utilization and communication latency for datacenter systems. We developed the methods and algorithms of job mapping and job scheduling to fully utilize the interconnection networks of racks connected by the optical wireless links. With a series of event driven simulations, we showed that the FSO links can reduce the hop count and communication latency for user jobs. We also presented the advantage of the FSO-equipped IRS systems in job scheduling performance such as average turnaround time of dispatched jobs for given sets of benchmark workloads.

研究分野：ハイパフォーマンスコンピューティング

キーワード：データセンター　光無線　ネットワークトポロジ

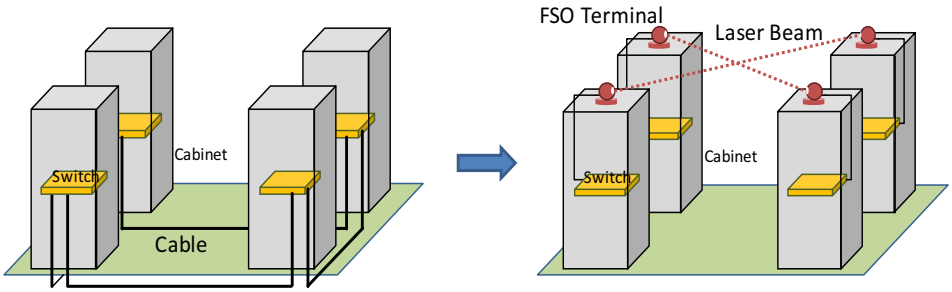
様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

（１）ニューラルネットワークに代表される最近のビッグデータ処理では、入力データやモデルによって定まる不均一な通信パターンを生じる。しかし、アプリケーション毎に異なる通信パターンに適したネットワークトポロジをデータセンターシステムが採用することは難しい。そのため、アプリケーションの通信パターンとデータセンターのネットワーク構成に乖離が生じる。

（２）実際には、トーラス、Fat ツリーなどのネットワークトポロジの中から、直径、スイッチの次数、ルーティングの容易性、耐故障性、レイアウトとコストなどの点でトレードオフを考慮した上で、システム毎に設計者の総合的な判断により異なるトポロジが選択されている。しかし、ケーブルの物理的な制約により、アプリケーションを実行するノード数が足りてもそのトポロジを構成できないことがある。その結果、並列アプリケーションの通信待ち時間が長くなる。

２．研究の目的



現在：ケーブルリンク通信
を用いた相互結合網

将来：FSO光無線リンク
を用いた相互結合網

図１：現在と将来のデータセンター結合網

（１）ニューラルネットワークに代表される最近のビッグデータ処理では、入力データやモデルに依存する不均一な通信パターンが生じる。本研究では、この不均一な通信パターンにあわせたネットワーク構成をとるために、光無線を用いたデータセンターシステムの構成を探求する。具体的には、各並列アプリケーションが必要とするプロセッサ、メモリ、ストレージ、GPU などのアクセラレータ間で FSO（Free Space Optics）光無線リンクを用いて、直接オンデマンドで直結する方式を探求する（図 1）。この方式により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効率的にサポートすることが期待できる。

（２）この再構成法に対して、どのアプリケーションを、いつ、どのように、実行させるのか？というジョブスケジューリング手法を開発することで、本大規模計算機システムの効率的な利用と設計法を確立する。また、FSO 光無線通信リンクを現状のデータセンターシステムに導入することで、システム利用効率向上およびユーザーエクスペリエンス向上ができることを検証する。

３．研究の方法

（１）資源利用率の観点から、実際の大規模計算機システムにおけるスイッチポート数やラック配置といった実装上の制約を考慮に入れ、真のリソース利用率向上をもたらす新たなアーキテクチャを導き出した。その手段として、Intel ラックスケールアーキテクチャの概念をさらに拡張し、ラック間をまたいだリーソスのネットワーク接続を提案した。Python で実装したスパコンスケジュールシミュレータと NetworkX/Pandas で実装したトポロジ構成ツールを活用し、ネットワーク生成・グラフ解析・スケジューリングシミュレーションを行った。また、光無線通信を利用することで光無線リンク数やネットワークのフレグメンテーションの制限を考慮したラック間ネットワークトポロジの動的構成法を開発した。開発したアルゴリズムがアーキテクチャ上の資源や FSO 光無線リンクの利用衝突を避け、有効な資源分配法を実現できることを検証した。その手段として、保有したシミュレータを活用し、スケジューリングアルゴリズムを導入し、タスクを実行するトポロジを動的に計算した性能評価を行った。

（２）システム管理者の観点から、タスクマッピングとタスクスケジューリングの手法を開発した。その手段として、シミュレータ上でタスクのサイズ順や到着時間順などを考慮に入れ、Backfill 付きスケジューラを導入した。また、並列分散アプリケーションを実行し、現在最先端のデータセンターシステムと比べてアプリケーションの待ち時間や実行時間が短縮されることを検証した。この時点では将来の FSO 光無線通信ターミナル付きスパコンが実用化されていないため、遅延の評価にあたっては、慶應義塾大学松谷研の FPGA スイッチを用いた実機評価環境での実験結果に基づいて相対的な比率を調整するなどの対応を行った。

４．研究成果

（１）現状データセンターにおける各コンピュータ要素の CPU、メモリ、ストレージ、GPU などといったハードウェアをラック単位で別々に集約する新たな IRS（Inter-RackScale）アーキテクチャを想定し、動的に FS0 光無線リンクを利用することでアプリケーション毎にリソースを柔軟に分配し、資源利用率の向上を図るタスク実行トポロジの構成法を提案した。

評価実験として、FS0 光無線リンクの動的再構成によるラック間ホップ数削減の効果を調査した。ホップ数が増加するにしたがって実行時間が増加するため、FS0 光無線を用いて送受信間でショートカットリンクを動的に設定することが極めて重要であることを検証した。将来の光無線相互結合網の重要性があらためて示された。

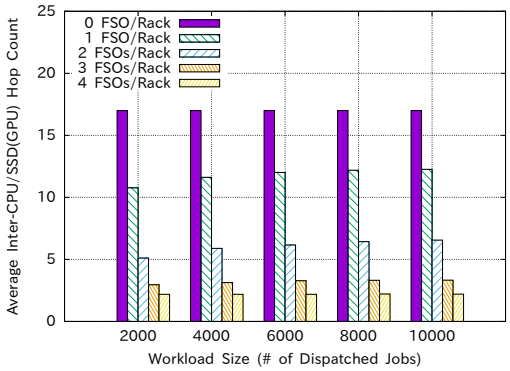


図 2：光無線通信によるラック間平均ホップ数（2-D Torus）

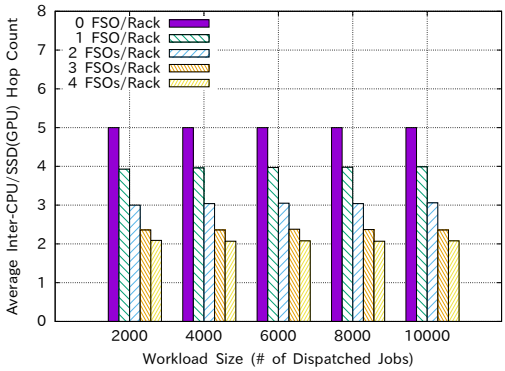


図 3：光無線通信によるラック間平均ホップ数（Fat Tree）

図 2 図 3 は、IRS アーキテクチャの具体的な基本システム構成として 2 次元 Torus トポロジ、Fat Tree トポロジにおける 1152 台のラックを想定し、FS0 光無線リンク数を 0〜4 本/ラックとした場合の評価結果を示している。シミュレーションでは、4 個の CPU で 1 ユニット、64 個の SSD で 1 ユニット、64 個の GPU で 1 ユニットを構成し、各ジョブはランダムに選択した個数のユニット(最大 25 ユニット)の資源を用いることとした。ジョブマッピングにおいて複数リンクを 1 本の辺とみなす(すなわち dilation 値)ことも許容することでジョブを dispatch する機会を増やす工夫を行った。

シミュレーション実行結果として、ラック間で FS0 光無線リンクを利用することで、ラック間平均通信ホップ数を 2 ホップまで（FS0 光無線リンク数を 4 本/ラックとした場合）下げることができることが分かった。

（２）ラック間平均通信ホップ数を抑制するとともに、多数の並列アプリケーションがデータセンターやスパコンで同時に動作するための「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 FS0 光無線相互結合網向けに改良することで、パケットの平均通信遅延を最小化することができることを検証した。この場合、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アプリケーションを実行するためのサブネットワーク構成を取った。

評価実験として、FS0 光無線リンクの動的再構成によるパケットの平均通信遅延の削減の効果を調査した。また、ラック間ホップ数が増加するにしたがって通信遅延も同様に増加するため、光無線通信を用いてラック間でショートカットリンクを動的に設定することが重要であることを示した。

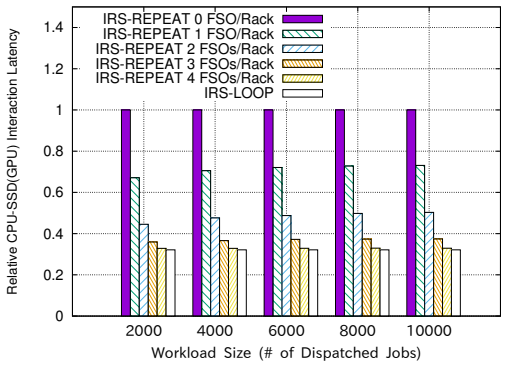


図 4：光無線通信によるラック間通信遅延（2-D Torus）

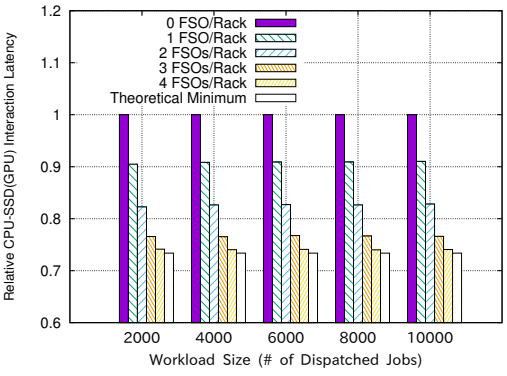


図 5：光無線通信によるラック間通信遅延（Fat Tree）

図 4 図 5 は、2 次元 Torus、Fat Tree トポロジで、タスクマッピングに対して最適化された並

列ジョブを、1 つのスパコンで実行してパケットの平均通信遅延を示している。FSO 光無線リンクの endpoint を交換することでネットワークトポロジを変更し、多くのジョブを収納できるようになっている。それに、シミュレーション実行結果として、2-D Torus の場合、光無線通信によるパケットの平均通信遅延が全ケーブル通信による通信遅延の 32.86%（FSO 光無線リンク数を 4 本/ラックとした場合）へ削減されることができることがわかった。Fat Tree の場合、光無線通信によるパケットの平均通信遅延が全ケーブル通信による通信遅延の 74.01%（FSO 光無線リンク数を 4 本/ラックとした場合）へ削減されることができることがわかった。

（3）FSO 光無線通信リンク付き IRS アーキテクチャへのジョブスケジューリング性能の評価を行った。ワークロードとして NAS Parallel Benchmark (NPB)の実行ログ（並列整数ソートなど大規模データ処理を含む）を用いてジョブマッピングやジョブスケジューリングのシミュレーションを行った。各サブネットワークトポロジでの各並列計算の実行時間は SimGrid イベントシミュレーションを用いることで見積りを行った。各ワークロードのジョブはポアソン分布に基づくジョブ到着間隔とし 2000 個のジョブを実行することと仮定した。IRS アーキテクチャでは、プロセッサノード（CPU ノード）、ストレージノード（SSD ノード）、GPU ノードをそれぞれ異なるラック内に集約配置し、アプリケーション毎に必要となるラック間を、動的に FSO 光無線リンクで直結する。

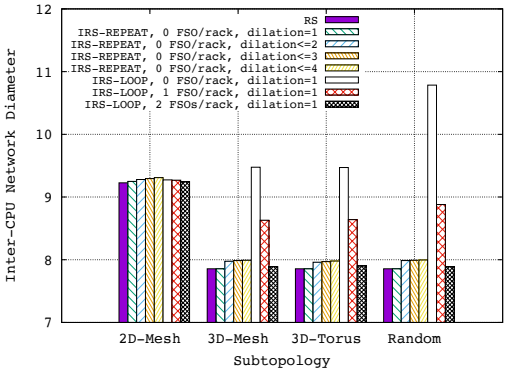


図 6：ジョブマッピングにおける CPU 間の直径

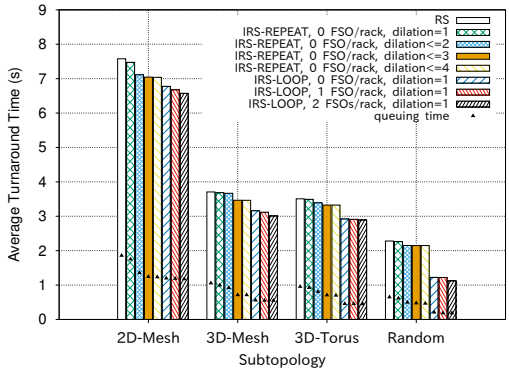


図 7：アプリケーションの平均ターンアラウンド時間

図 6 は、各ジョブ内サブネットワークトポロジと dilation 値に対する CPU ネットワークの直径を示している。図 7 は、アプリケーションの平均ターンアラウンド時間を示している。いずれも値が小さいほど性能が高いことを示している。IRS-LOOP は同種のラックを分散配置した場合、IRS-REPEAT は集中配置した場合である。

評価結果として、CPU ノード、SSD ノード、GPU ノードへのジョブマッピングを効率的に行うことができた場合、ジョブ内の通信のホップ数が小さくなり、並列処理における通信時間の割合を小さくできるため実行時間を小さくする、すなわち平均ターンアラウンド時間を小さくすることができる。また、ラックあたりの FSO 光無線リンク数が多いほど平均ターンアラウンド時間を小さくすることができる。既存の典型的なデータセンターシステムであるラックスケール（RS）アーキテクチャと比べた場合、IRS アーキテクチャは FSO 光無線リンクの endpoint の動的再構成を効率良く用いることでアプリケーションを実行するためのサブネットワークトポロジを構成することにより、平均ターンアラウンド時間、すなわち実行時間を半分以下、つまり、実行性能を倍以上にすることに達成した。なお、FSO 光無線リンクの再構成オーバーヘッド時間を 1 秒～10 秒に変更した場合でも平均レスポンス時間の傾向に大きな変化はなかった。要するに、FSO 光無線通信リンクを現状のデータセンターシステムに導入することで、システム利用効率向上およびユーザーエクスペリエンス向上ができることが望ましい。

（4）前述した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を進め、研究会・国際会議・論文誌などで発表した。

５．主な発表論文等

〔雑誌論文〕（計　1　件）

［1］Yao Hu, Michihiro Koibuchi, "Enhancing Job Scheduling on Inter-rackscale Datacenters with Free-space Optical Links", IEICE TRANSACTIONS on Information and Systems, Vol.E101-D, No. 12, pp. 2922-2932, Dec. 2018. DOI: https://doi.org/10.1587/transinf.2018PAP0010（査読あり）

〔学会発表〕（計　2　件）

［１］Yao Hu, ″Circuit-Switched Interconnects Using Limited Number of Slots″, MS71 Applied Graph Theory in Interconnection Network Design and Operation, SIAM Conference on Parallel Processing for Scientific Computing, Waseda University, Tokyo Japan, March 7-10, 2018.

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者

(2)研究協力者
研究協力者氏名：鯉渕　道紘
ローマ字氏名：Koibuchi Michihiro