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　３ 年　 ４ 月 ２９ 日現在

機関番号：６２６１５
研究種目：若手研究
研究期間：2019～2020
課題番号：１９Ｋ２０２６３

研究課題名（和文）全光無線通信による大規模計算機ネットワーク進化

研究課題名（英文）Large-scale Computer Network Evolution by All-optical Wireless Communication

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）　　3,200,000円

研究成果の概要（和文）：本研究では、ラック間とラック内で多数のレーザービームが相互に干渉しないようなFSO光無線端末レイアウト設計法を確立し、入力した通信パターンに合わせた最適な光無線トポロジの動的構成法を提案した。また、将来の全光無線リンクを用いたデータセンターシステムを想定し、光無線トポロジ動的構成法を活用し、アーキテクチャの最適な物理トポロジへのタスクマッピング手法を導き出した。そして、ラック間とラック内のトポロジとリンク遅延を考慮しアプリケーション毎に計算ノードを柔軟に分配するタスクスケジューリング手法を提案した。

研究成果の学術的意義や社会的意義
本研究で開発したトポロジ動的構成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巨大なデータ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）：We made a comparative study to analyze the impact of application mapping performance on non-random and random network topologies. We investigated the job mapping on FSO-based random topologies so that a newly incoming job can be immediately dispatched as long as there are enough available nodes. Evaluation results show that, for a large compound workload of NAS Parallel Benchmarks (NPB) applications, the random job mapping can reduce up to 64% of makespan and up to 80% of turnaround time when compared with the regular job mapping. Overall, the random topology embedding in random topologies indicates substantial room for improvement of job scheduling performance.
We proposed using topology embedding metrics, i.e., diameter and ASPL, and listed several diameter/ASPL-based application mapping algorithms to compare their job scheduling performances, assuming that the communication pattern of each application is unpredictable to the computing system.

研究分野：高性能計算

キーワード：計算機ネットワーク　光無線通信　ネットワークトポロジ　タスクスケジューリング

１．研究開始当初の背景

（１）ビッグデータ時代のニューラルネットワークにおける巨大なデータ処理では、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発生してしまう。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現状の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、通信パターンと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よくある。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

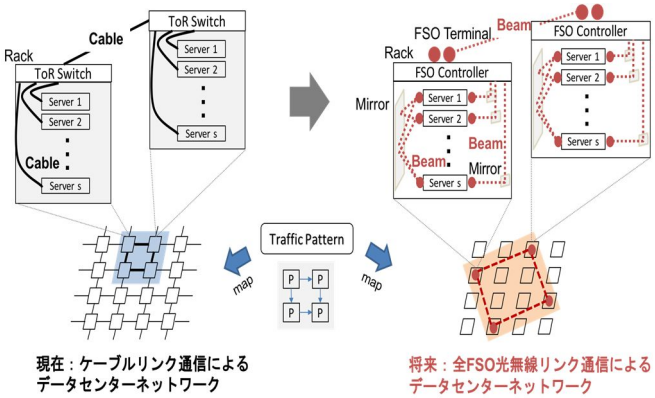

図 １ 現在と将来のデータセンターネットワーク

（１）ビッグデータ時代のニューラルネットワークにおける巨大なデータ処理では、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論理トポロジを検出したうえで、全光無線リンクを用いたデータセンターシステムにおける物理トポロジの動的構成法を探求する（図１）。具体的には、各アプリケーション通信の論理トポロジに基づいて、光無線リンクを用いて計算ノード間を直結することで最適な物理トポロジを構築する方式を探求する。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）光無線リンクを用いてデータセンター内のトポロジを動的に変えることで、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開発する。将来の全光無線通信による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ことで、次世代アプリケーション毎に実行時間が短くなる望みが持てる。

３．研究の方法

（１）ハードウェア的アプローチとして、資源利用率の観点から、実際の計算機システムにおけるラック配置や FSO 光無線レーザービーム干渉といった実装上の制約を考慮に入れ、真の資源利用率向上をもたらす新たな光無線アーキテクチャを導き出した。その手段として、現在最先端の光無線データセンターシステムの概念をさらに拡張し、ラック間とラック内をまたいだ計算ノードの全光無線通信ネットワーク接続を提案した。ソフトウェア的アプローチとして、FSO 通信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、入力した通信パターンに合わせた最適な光無線トポロジの動的構成法を開発した。その手段として、研究代表者である胡が Python で実装したスパコンネットワーク生成シミュレータと NetworkX/Pandas/C++で実装したトポロジ解析ツール群をすでに保有しており、ネットワーク生成・グラフ解析を行った。計算ノード間を、FSO 光無線リンクを利用して直接オンデマンドで直結することで、アプリケーションのホップ数や通信遅延が短縮されることを検証した。そして、データセンター内の通信ケーブルやスイッチが不要となり、低通信遅延 FSO ターミナルの実用化に伴うシステム全体のコストを削減することを検証した。

（２）システム管理者の観点からタスクマッピングとスケジューリングアルゴリズムを開発した。開発したアルゴリズムが光無線アーキテクチャ上の資源や FSO 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か否か正確に評価した。その手段として、研究代表者である胡が保有したスケジューリングシミュレータを活用し、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。提案されたシステム全体を実装してその実現性を検証するとともに、前述したスパコントポロジ生成・解析ツールと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。並列分散アプリケーションを実行し、現状のケーブル計算機システムや開発した単純な光無線システムと比べて、通信待ち時間や総実行時間が短縮されることを明らかにした。

４．研究成果

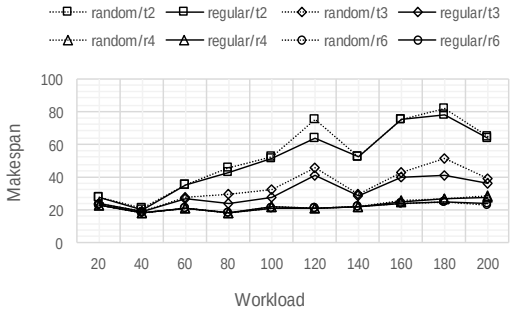

図２ メイクスパン

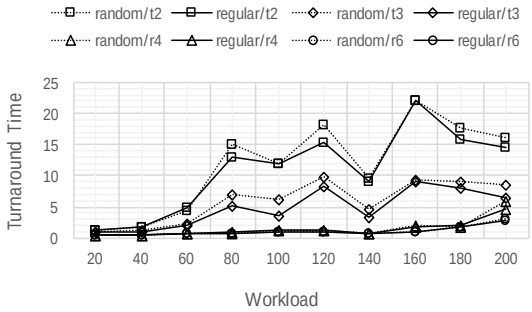

図３ ターンアラウンドタイム

（１）光無線アーキテクチャ上の各種トポロジのタスクマッピングとスケジューリングの性能を評価した。図２は、さまざまなホストトポロジでのジョブスケジューリング全体のメイクスパンを示している。すべてのホストトポロジの中で、2D トーラス（t2）は、メイクスパンが最も長くなるため、パフォーマンスが最も低くなる。 3D トーラス（t3）のホストトポロジは、2D トーラス（t2）よりも短いメイクスパンをもたらす。ホストトポロジは 2D トーラス（t2）と 3D トーラス（t3）の場合、レギュラートポロジでのジョブマッピングがランダムトポロジでのジョブマッピングよりもわずかに優れている。これは、レギュラートポロジでのマッピングにより、連続したグラフ埋め込みが発生し、マップされたジョブの実行時間が短くなるためである。この場合、ランダムトポロジでのマッピングの待機時間は短くなるが、実行時間が長くなることによるパフォーマンスの低下を補うことはできない。ただし、レギュラートポロジでのマッピングとランダムトポロジでのマッピングの違いは実際にはわずかである。単純さとシステム使用率を考慮すると、ホストトポロジが 2-/3-D トーラスの場合、ランダムトポロジでのマッピングにはわずかな罰が許容される。
同様の傾向が図３からもわかる。これは、さまざまなホストトポロジでディスパッチされたすべてのジョブの平均ターンアラウンドタイムを示している。 3D トーラス（t3）のホストトポロジは、2D トーラス（t2）よりも平均ターンアラウンドタイムが短くなる。同様、ホストトポロジが 2-/3-D トーラスの場合、レギュラートポロジでのジョブマッピング性能はランダムトポロジでのジョブマッピング性能をわずかに上回る。
全体として、ランダムホストトポロジは、メイクスパンとターンアラウンドタイムの点で最高のパフォーマンスを示す。ワークロードが 40 より大きい場合、利点が明らかになる。たとえば、ワークロードが 200 の場合、次数 6（r6）のランダムトポロジでのランダムジョブマッピングは、2D トーラスを介したレジュラージョブマッピングと比較すると、メイクスパンが 64%、ターンアラウンドタイムが 80%削減されることと、3D トーラスを介したレジュラージョブマッピングと比較すると、メイクスパンが 36%、ターンアラウンドタイムが 54%短縮されることが分かる。ホストトポロジが次数 4（r4）または次数 6（r6）のランダムトポロジの場合、レジュラートポロジでの埋め込みとランダムトポロジでの埋め込みの間にほとんど違いはなく、後者は前者よりもわずかな利点を示す。また、次数 4（r4）のランダムホストトポロジの性能は次数 6（r6）のランダムホストトポロジのとあまり変わらない。言い換えると、ランダムホストトポロジの次数は、ジョブスケジューリングのパフォーマンスに非常に限定的な影響を及ぼす。

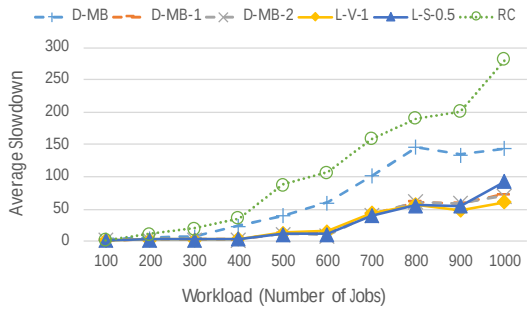

図４　スローダウン（NPB）

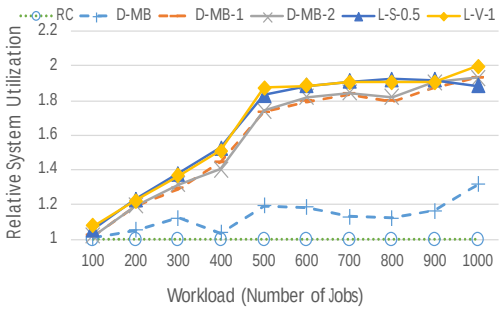

図５ 資源利用率（NPB）

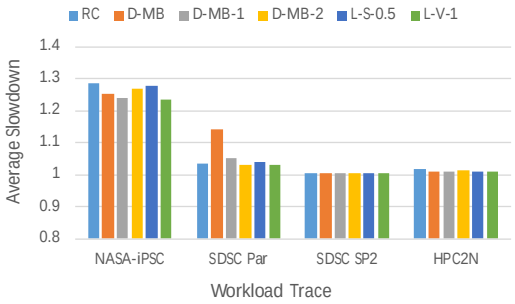

図６　スローダウン（PWA）

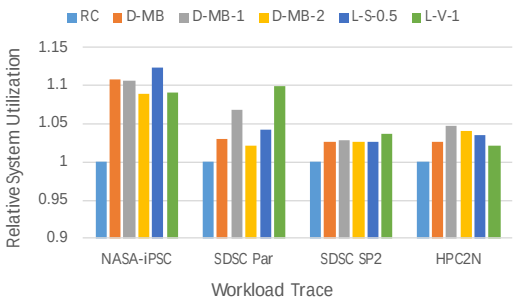

図７　資源利用率（PWA）

（２）光無線アーキテクチャ上の各種タスクマッピングとスケジューリングのアルゴリズムの性能を評価した。図４は、平均スローダウンのパフォーマンスの比較を示している。これは、実行時間に対するターンアラウンドタイムの平均比率を反映している。したがって、値が大きいほど、待機時間の影響が大きいことを示す。比較すると、D-MB は、トポロジマッピングの直径がムーアバウンドに厳密に制限されているため、平均ターンアラウンドタイムと平均スローダウンの点で RC よりも優れている。ただし、ムーアバウンドの厳格な制限により、トポロジマッピングの直径がわずかに大きい値に緩和されて大きな違いがない D-MB-1 と D-MB-2 と比較した場合、トポロジマッピングのパフォーマンスをさらに向上させることもできない。L-S-0.5 と L-V-1 アルゴリズムは、トポロジマッピングの制限として直径の代わりに ASPL を使用し、D-MB-1 と D-MB-2 と同等のパフォーマンスを提供する。ジョブの数が 1,000 になるなど、ワークロードが重くなると、L-V-1 はすべてのトポロジマッピングアルゴリズムの中で最高のパフォーマンスを示す。
同様の傾向が図５からも見られる。これは、メイクスパン全体でのシステム上のノード使用率の比較を反映している。全体として、ASPL ベースのアルゴリズム（L-S-0.5 と L-V-1）は、直径ベースのアルゴリズム（D-MB と D-MB-1 と D-MB-2）よりもわずかに優れている。特に、ジョブ数が 1,000 のようにワークロードが重い場合でも、LV-1 は最高のパフォーマンスを発揮し、ベースライン RC 方式と比較した場合、メイクスパンを最大 48.0%、平均ターンアラウンドタイムを最大 78.1%短縮し、資源利用率を最大 1.9 倍向上させる。
図６、図７に示すように、PWA ワークロードトレースの場合、ほとんどの場合、直径/ ASPL ベースのアプリケーションマッピングは、ジョブスケジューリングパフォーマンスの点でベースライン RC トポロジマッピングよりも優れている。ただし、要求された計算ノードが小さい場合、またはアプリケーションの実行時間がジョブの到着間隔よりも明らかに短い場合（SDSC SP2 や HPC2N の場合など）、異なるトポロジマッピングアルゴリズム間で平均ターンアラウンドタイムと平均スローダウンに大きな違いはない。この場合、ネットワーク全体のほとんどの計算ノードは使用可能な状態を維持するため、任意のトポロジマッピングアルゴリズムを使用して到着ジョブをすぐにディスパッチできる。もう１つの例外は、平均ターンアラウンドタイムと平均スローダウンの点で、SDSC Par の場合、D-MB が他のトポロジマッピングアルゴリズムよりもパフォーマンスが悪い。これは、トポロジマッピングの直径のムーア境界の厳密な制限を伴うゲストトポロジへの待機時間が長いため、ジョブスケジューリングのパフォーマンスに悪影響を与える可能性がある。簡単な最適化は、制限を少し大きい値（たとえば、ムーアバウンドに１つまたは２つを加えたもの）に緩和することである。全体として、ASPL ベースの L-V-1 アルゴリズムは、ジョブスケジューリングのパフォーマンスが安定して向上しているため、どのような場合でも推奨される。
（３）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（うち査読付論文　1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 and Michihiro Koibuchi	４．巻 E103-D(12)
２．論文標題 Application Mapping of Uncertain Communication Patterns onto Non-random and Random Network Topologies	５．発行年 2020年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 2480-2493
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2020PAP0006	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2件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Yao Hu
２．発表標題 Topology Mapping of Parallel Applications onto Random Allocations
３．学会等名 The 21st IEEE International Conference on High Performance Computing and Communications (HPCC-2019)（国際学会）
４．発表年 2019年

１．発表者名 Yao Hu and Michihiro Koibuchi
２．発表標題 Design and Implementation of Circuit Switched Scheduling in Flow-in-Cloud Systems
３．学会等名 The 9th International Workshop on Networking, Computing, Systems, and Software (NCSS-9), First mini Symposium on Computing and Networking (miniCANDAR 2019)（国際学会）
４．発表年 2019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