様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ９ 日現在

機関番号：６２６１５
研究種目：基盤研究(B)（一般）
研究期間：2013～2015
課題番号：２５２８００１８

研究課題名（和文）ランダムショートカットと光通信技術による超低遅延グリーンインターコネクト

研究課題名（英文）Low-Latency Green Interconnects using Random Shortcuts and Optical Communication Technology

研究代表者

鯉渕　道紘（Koibuchi, Michihiro）

国立情報学研究所・アーキテクチャ科学研究系・准教授

研究者番号：４０４１３９２６

交付決定額（研究期間全体）：（直接経費）　　14,000,000円

研究成果の概要（和文）：本研究では，エクサスケール規模以上の高性能計算機システムにおいて，ランダムショートカットリンク接続を光波長多重スイッチ技術により実現することで（１）最長通信遅延1μ秒，（２）現状の電気スイッチのみを用いたHPCインターコネクトと比べて電力性能比数倍を実現するインターコネクト技術を提案し，有効性を示した．本光多重技術の利用により配線を隣接キャビネット間に抑えることが可能となるため，総配線長をほぼ最小にできる．さらに，スモールワールド性を利用してランダムトポロジや通信パターンにあわせて最適なネットワークトポロジを構成することが可能である．

研究成果の概要（英文）：We have proposed and evaluated interconnects technology for exascale HPC computers in order to provide (1) lus communication latency to a farthest node and (2) multiple-times improvement of power efficiency when compared to existing electric-switched networks. Our technology relies on optical wavelength division multiplexing switches using random shortcut links.
Using the optical switches efficiently provides inter-rack cabling that minimizes the total length of wires. We take a suitable network topology for a given traffic patterns by using the small-world phenomenon on our proposed interconnection networks.

研究分野：相互結合網、計算機システム・ネットワーク

キーワード：ハイパフォーマンス・コンピューティング　フォトニックネットワーク　計算機システム　相互結合網

１．研究開始当初の背景

(1) 2020 年頃登場のエクサスケールのスーパーコンピュータ（以後スパコンと呼ぶ）では，１つのアプリケーションに 100 万～数億並列処理が必要となる．これを効率よく実現するためにノード間最小通信遅延 300ns，最長(最も離れたノード間)通信遅延 1us が要求されているが，ムーアの法則にしたがった向上では現状の数倍の改良に留まり，最長遅延を数 us 以下にすることが難しい．データセンターにおいても，メニーコア化による細粒度並列処理，HPC に近い解析ワークロードなどの高速化において同様の現象が予測される．加えて，１台のスパコンの配線長が 2,000km を超えるなど物理資源量，配線密度とレイアウト，ケーブルの張替などの高メンテナンスコストの点で，規則的なトポロジで結合するネットワーク構築は限界が近づいてきている．

(2) 今後，半導体集積が進むため，配線遅延（5ns/m）すら，大規模インターコネクトの遅延ボトルネックになる．したがって，今後はマシンルーム内のスイッチとキャビネットレイアウトの工夫および，波長多重化の積極的な利用により通信遅延を改善することが必要となる．この点で，ROADM(Reconfigurable Optical Add Drop Multiplexer)技術に基づき光パスを事前に設定することにより長距離のランダムショートカットリンクをマンハッタン距離で実現することが有望である．

２．研究の目的

本研究では，エクサスケール規模以上の高性能計算機システムにおいて，ランダムショートカットリンク接続を光波長多重スイッチ技術により実現することで（１）最長通信遅延 1μ秒，（２）現状の電気スイッチのみを用いた HPC インターコネクトと比べて電力性能比数倍向上（３）通信パターンに応じたネットワークトポロジを実現するインターコネクトを探求する．本光多重技術の利用により配線を隣接キャビネット間に抑えることが可能となるため，総配線長をほぼ最小にできる．さらに，スモールワールド性を利用してランダムトポロジや通信パターンにあわせて最適なトポロジを再構成可能とする．

３．研究の方法

(1) ベースライン相互結合網の構成

我々は低遅延，低消費電力の面から多角的な構成，可能性の検討を進める．ベースラインとして，ROADM 技術を用いた光スイッチを各キャビネットの Top-of-Rack(ToR)に据え，有線の配線を隣接キャビネット間に限定する．光スイッチの接続はランダム，あるいは通信パターンに合わせて静的に設定するが，波長の多重度には限界があるため，電気スイッチを複数経由してパケットは転送される．そして並列アプリケーションを実行する毎に光スイッチの設定を更新することで可変トポロジを実現する．

(2) ランダム性を含むネットワークトポロジ

ベースラインとなるキャビネット内ネットワークとキャビネット間ネットワークの構成をふまえた上で，光多重リンクの静的な更新により実現可能な可変トポロジと，その実現に必要となる光多重スイッチの次数とレイアウトについて定量的に解析し，一部の構成について我々が保有するサイクルアキュレートシミュレーションにより有効性を示し，ポストエクサスケール HPC システムまでスケーリングすることを示す．

(3) 光多重の静的再構成(WDM スイッチと光パス)

コストと本光スイッチの次数との議論を進め，定量化した上で，システムサイズに対する本光スイッチ数の配置とコストのトレードオフを定量化し，システム実現性を明瞭にする．

(4) スパコンネットワーク設計への応用

本成果技術をすぐに実機において利用できるように Cray 社 Black Widow スパコンや ANSI/TIA/EIA-942 などの規格に沿った本インターコネクトの設計を示し，大規模化は結合網解析，また小規模な構成についてはサイクルアキュレートシミュレーションにより有効性を提示する．

(5)データセンタネットワーク設計への応用

Top-of-Rack スイッチを用いたツリートポロジを採用していることが多い点が HPC システムと異なる．このトポロジを更新した上で本ランダムショートカットリンクの追加をした評価（グラフ解析と MapReduce などのトラヒックパターンに対するサイクルアキュレート・ネットワークシミュレーション）を進めることで本成果の適用可能な対象であることを明確にする．

以上より，本研究は電気スイッチを規則トポロジで構築した既存のスパコン，データセンター・ネットワークが今後直面する遅延，電力面での大規模化の限界突破を行う．具体的には,ランダムショートカットのアイデアを本光スイッチにより実装することで，革新的な低遅延スパコン，データセンター・グリーン（省電力，省資源の意味）ネットワークを実現する．

４．研究成果

(1) エクサスケール規模以上の高性能計算機システムにおいて，ランダムショートカットリンク接続を光波長多重スイッチ技術により実現するベースライン相互結合網を提案した．具体的にはスパコンのインターコネクトに，低次元スイッチを用いた場合を想定した．その結果，階層的にリンクを追加することで直径と次数がランダムトポロジの特性に近づきつつ，レイアウト時にトーラスよりも短い配線長に抑えること成功した．この

見積により，スイッチ遅延とケーブル遅延の和で定まる最長通信遅延を従来の規則網に比べて大幅に削減可能なことが分かった．さらに，光波長多重スイッチ技術にトラヒックをオフロードすることにより大規模計算機システムにおいてジョブ毎のパーティショニングが極めて上手くできることが分かった．

(2) 光スイッチを用いたHPCインターコネクトのトポロジについて，光スイッチ技術，ネットワーク構成，要求パフォーマンス等の観点から最適化技術を確立した．また，メモリ間通信をCPUを介さず行うダイレクトメモリコピーアーキテクチャについても検討を行った．試算の一つとして，10万計算ノードの相互接続に必要となる光スイッチのハードウェア規模をラック数で換算すると，シリコンフォトニクス技術を用いた高密度集積により，計算ノードを収容するラック数の数%程度に収まるとの結果を得た．また，現在実現されている光スイッチは切替ガードタイムが比較的大きいため，転送サイズが数～数十メガバイト以上という領域において効率的なバーストスイッチングが可能であるとの試算を得た．

(3) 電気スイッチネットワークに対し，光波長多重スイッチを補助的に利用することにより，k-ary n-cube，Fat ツリー，ランダムという3種類の電気スイッチ間トポロジを効率良く内包可能な低遅延結合網を提案した．提案手法は，既存のトポロジに比べ低い導入コストで，かつ，遺伝的アルゴリズムによる探索範囲を限定することでトポロジ内包性とシステム全体での低遅延性を両立できることが分かった．そして多くの場合，並列アプリケーションのプロセス空間に適したトポロジを全体の計算システムから割り当てることが可能となることが分かった．

(4) 本相互結合網を導入する場合のフロアプランについて，拡張可能なように光パッチパネルを導入する方法，低遅延低次元トポロジの提案とそのトポロジにおける任意の電気，光波長多重スイッチ数の拡張が可能なフロアプランを開発した．そして，最終的に本相互結合網における数千並列で実行する並列科学技術アプリケーションの評価を行うことが可能なイベントドリブンシミュレーションであるSimGrid環境の構築を行い，通信プロファイルの見える化を実現する拡張を行った．

(5)インターコネクト網に光スイッチ技術を導入する際に，光スイッチの切替を分散制御的に実行可能とする新たな手法を提案した．これにより，従来提案されている集中制御手法に比べ，既存インターコネクト網制御技術との高親和性の達成，及び，切替にかかるオーバーヘッドタイムの削減が可能となる．両者より，ランダムショートカットリンクを光スイッチ上で効果的に利用することが可能となり，効率的に低遅延性と高バンド幅を要求するアプリケーションをサポートすることが可能となった．

1～5の結果より，光波長多重スイッチと電気スイッチの混在する相互結合網における(電気スイッチの)ネットワークトポロジの構築方法を確立した．その結果，並列アプリケーションの通信パターンに合わせた構成が可能となり，数百ラック規模の計算機システムにおいて最長ゼロ負荷通信遅延 1μ秒を達成できる場合が多いことを示すことができた．現状のHPCネットワークは汎用のアクティブ光ケーブルと InfiniBand などの電気スイッチで構成されている．しかし，ムーアの法則の終焉が近づいていることを考えると，これらの既存のネットワークアーキテクチャでは，将来的に大規模化が進むHPC/データセンターを構成するネットワークに十分な低遅延性，電力性能比，帯域を提供することは難しいと考えられる．一方で，本研究で開発したランダムショートカットリンク技術を波長多重スイッチ技術により実現することで，最長システム内通信遅延 1μ秒，電力性能比数倍，1Tbps 級の高帯域リンクを達成可能である．よって，これらの技術は，今後のネットワーク設計において極めて有望である．

５．主な発表論文等

（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 12 件）

(1) Michihiro Koibuchi, Ikki Fujiwara, Kiyo Ishii,Shu Namiki, Fabien Chaix, Hiroki Matsutani, Hideharu Amano and Tomohiro Kudoh, Optical Network Technologies for HPC: Computer-Architects Point of View,IEICE Electronics Express (ELEX), Vol. 13 2016 No. 6, 査読無(招待レビュー論文)

(2) Nguyen T. Truong, Van K. Nguyen, Ikki Fujiwara, Michihiro Koibuchi, Layout-conscious Expandable Topology for Low-degree Interconnection Networks, IEICE TRANSACTIONS on Information and Systems, E99-D, No.5, pp.1275-1284, 2016年5月, 査読有

(3) Ahmed Shalaby, Ikki Fujiwara and Michihiro Koibuchi, The Case for Network Coding for Collective Communication on HPC Interconnection Networks, IEICE TRANSACTIONS on Information and Systems, Vol.E98-D,No.3,pp.661-670,Mar. 2015.査読有

(4) Nguyen T. Truong, Van K. Nguyen, Nhat T. X. Le, Ikki Fujiwara, Fabian Chaix and Michihiro Koibuchi, Layout-aware Expandable Low-degree Topology, The 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), pp.462--470, Dec. 2014,DOI:

10.1109/PADSW.2014.7097842, 査読有
(5) Van K. Nguyen, Nhat T. X. Le, Ikki Fujiwara, Michihiro Koibuchi, Distributed Shortcut Networks: Layout-aware Low-degree Topologies Exploiting Small-world Effect, the International Conference on Parallel Processing(ICPP), pp.572-581, Oct 2013, 10.1109/ICPP.2013.71, 査読有

〔学会発表〕(計 17 件)
(1)Michihiro Koibuchi, Singularity of Future Computer-System Networks, the sixth ACM international symposium on information and communication technology (SoICT), Dec. 2015,インペリカルフエフエホテル(ベトナム・フエ)(招待講演)
(2)Michihiro Koibuchi, Future Low-latency Networks for High Performance Computing, The First International Symposium on Computing and Networking Across Practical Development and Theoretical Research (CANDAR), Japan, Dec, 4-6, 2013,ひめぎんホール(愛媛県松山市), (招待講演)
(3)鯉渕 道紘, 超低遅延 HPC インターコネクトのためのランダムトポロジ,電子情報通信学会通信方式研究会のワークショップ, 第 26 回情報伝送と信号処理ワークショップ pp.29-33,2013 年 11 月, 第一滝本館(北海道登別市)(招待講演)

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況(計 0 件)

○取得状況(計 0 件)

〔その他〕
ホームページ等
高性能並列分散システムの超省電力・高信頼・低遅延インターコネクト
http://research.nii.ac.jp/~koibuchi/research02.html

６．研究組織
(1)研究代表者
鯉渕 道紘 (KOIBUCHI Michihiro)
国立情報学研究所・アーキテクチャ科学研究系・准教授
研究者番号：40413926

(2)研究分担者
石井 紀代 (ISHII Kiyo)
産業技術総合研究所・電子光技術研究部門・主任研究員
研究者番号： 90612177

(3) 研究分担者
天野 英晴 (AMANO Hideharu)
慶應義塾大学・理工学部・教授
研究者番号： 60175932

(4) 連携研究者
工藤 知宏 (KUDOH Tomohiro)
東京大学・情報基盤センター・教授
研究者番号：00234451

(5) 連携研究者
並木 周 (NAMIKI Shu)
産業技術総合研究所・電子光技術研究部門・副研究部門長
研究者番号:30415723

(6) 連携研究者
藤原 一毅 (FUJIWARA Ikki)
国立情報学研究所・アーキテクチャ科学研究系・特任准教授
研究者番号:90648023

(7) 連携研究者
松谷 宏紀 (MATSUTANI Hiroki)
慶應義塾大学・理工学部・講師
研究者番号:70611135