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 ２７ 年　 ６ 月　 ３ 日現在

機関番号：６２６１５
研究種目：若手研究(B)
研究期間：2013～2014
課題番号：２５７３００６８

研究課題名（和文）超低遅延時代へ向けた大規模計算機の相互結合網のトポロジとルーティング

研究課題名（英文）HPC Interconnects toward Ultra-low-delay Era

研究代表者
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

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）：この研究は、次世代スーパーコンピュータなどの大規模並列計算機システムの中で、数万台のコンピュータが非常に短い時間（1マイクロ秒以下）で相互に通信するためのネットワーク設計法を明らかにした。将来出現する超低遅延ネットワーク機器を想定し、従来のネットワーク設計法では無視されていたケーブル内の信号伝搬時間をきちんと考慮したうえで、最適な配線と通信経路を決める方法を開発した。これにより、研究開始前に最先端だったネットワーク設計法に比べて、ケーブルの長さを65%も短くしながら通信遅延がわずか6%しか増加しないという、省資源性と低遅延性を高いレベルで両立する次世代のネットワーク設計が可能になった。

研究成果の概要（英文）：This research clarified how to design an ultra-low-latency network that connects tens of thousands of computers within a large-scale parallel computing systems such as next-generation supercomputers. We developed a sophisticated method to design a network topology and a routing algorithm that minimizes the end-to-end latency. Existng design method for supercomputer networks did not take a cable delay into account, of which our proposed method explicitly take care. When comparing to a former state-of-the-art network design, our design achieves merely 6% higher communication latency while consuming 65% smaller amount of cable. As such, our proposed design method provides a desirable trade-off between cost and performance.

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：ネットワーク

１．研究開始当初の背景

スーパーコンピュータの大規模化が進むにつれ、通信遅延がアプリケーション性能向上の足かせとなっている。2019 年ごろのエクサスケール計算機システム上で実行される並列アプリケーションは、300 ナノ秒～1 マイクロ秒程度の非常に小さい通信遅延を要求することが予測されている。通信遅延はスイッチ遅延とケーブル遅延からなり、現在はスイッチ遅延（100～200 ナノ秒以上）が支配的である。そのため、少ない経由スイッチ数で多くのノードを結合できる、いわゆるスモールワールド性を持つネットワークを大規模計算機システムに応用しようとする研究が、2012 年に入ってから注目を集めている。すなわち、現在最先端の研究でもノード間結合網におけるケーブル遅延は実質的に無視されている。ケーブル遅延はケーブル内の信号伝搬速度に束縛され、電気ケーブル・光ケーブルとも、およそ 5 ナノ秒／メートルを下回ることができない。他方、スイッチ遅延は半導体技術の進歩にともなって継続的に短縮されつつあり、近い将来 60 ナノ秒を下回る製品が実用化されると予測されている。したがって、数十メートルに及ぶケーブルを用いる大規模計算機システムにおいては、将来、ケーブル遅延がスイッチ遅延に対して相対的に大きくなり、これを無視できなくなることが確実である。

図 1 は、横軸にスイッチ遅延をとり、縦軸にアプリケーションレベルの最長通信遅延を試算したものである。実線が従来型の典型的なトポロジ（ハイパーキューブ）、点線がスモールワールド性を持つトポロジ（ランダムリング）、二重線が本研究で提案するべき将来の低遅延トポロジを示す。この図から明らかなように、60 ナノ秒を下回る超低遅延スイッチが登場すると、現在最先端の研究成果が通用しなくなり、むしろ従来型のトポロジが低遅延性で有利となるようなパラダイムシフトが生じる。このように、将来実用化される超低遅延スイッチの恩恵をアプリケーションが十全に享受するためには新しいネットワーク設計の方法論が必要不可欠であり、その実現に向けた研究にただちに取り組まなければならない。

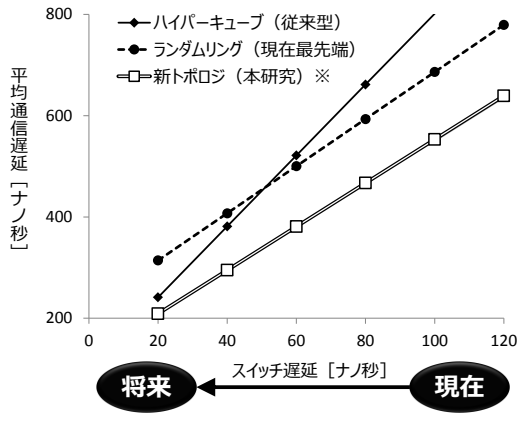

図 1　現在と将来の通信遅延(試算値)

２．研究の目的

将来の超低遅延スイッチを用いた大規模計算機システムのノード間結合網において真に効率的なネットワーク設計を実現するため、以下の 2 点を明らかにする。

(1)　トポロジ構成法: ハードウェア面からのアプローチとして、ラック配置と物理ケーブル長を陽に考慮し、各ノード間における経由ケーブル長と経由スイッチ数をともに小さくするネットワークトポロジの構成法を提案する。

(2)　ルーティング手法: ソフトウェア面からのアプローチとして、ケーブル遅延を陽に考慮し、端末ノード間の総通信遅延を最小化するパケットルーティング手法を提案する。

両者について、その効果を数理解析とシミュレーションによって定量的に評価する。

３．研究の方法

(1) ネットワークトポロジ構成法の開発・評価

本研究の目的を達成するには、純粋にグラフ理論的な観点からホップ数を削減することに加え、実際の計算機システムにおけるスイッチ遅延やケーブル遅延、さらには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、真の低遅延性をもたらすトポロジを導き出さなければならない。そのために、上述の諸条件をグラフに反映した独自のネットワークシステムモデルを構築する。その手段として、トポロジ生成ツールとグラフ解析ツールをそれぞれ開発し、1 万ノード程度までのトポロジ生成・グラフ解析を可能とする

(2) ルーティングアルゴリズムの開発・評価

将来ケーブル遅延が支配的となった状況を想定し、単に目的地までのホップ数に基づいてパケット転送経路を選ぶ従来型のアルゴリズムとは対照的に、スイッチ遅延やケーブル遅延といったシステム全体の物理的パラメータを考慮した経路選択アルゴリズムを開発する。また、開発したアルゴリズムが(1)のトポロジ上で極端な輻輳やデッドロックを起こさず、十分な帯域を保ちつつ低遅延な通信を実現できるか否かを正確に評価しなければならない。その手段として、研究協力者が C++ で実装したフリットレベル・ネットワークシミュレータを活用し、任意のルーティングアルゴリズムを用いた回路レベルでのシミュレーションを可能とする。これにより、グラフ解析では予測できないパケットの輻輳などを含めた精密な性能評価を行う。

４．研究成果

マシンルームのフロア上に数百台のラックが格子状に配置される大規模並列計算機を想定する。各ラック内には複数のスイッチがあり、各スイッチには複数の計算ノードが

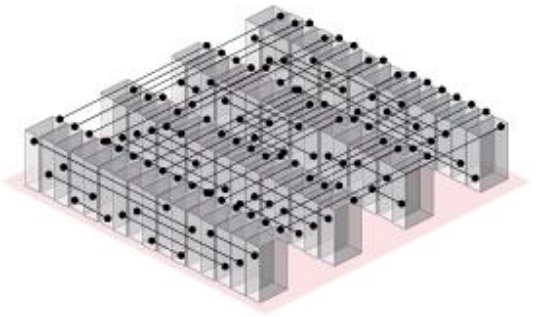

図 2　超低遅延時代に適合するスパコン向けネットワークトポロジ “Skywalk”

接続されているものとする。本研究で開発したネットワークトポロジ構成法（以下 Skywalk と称する）は、フロア上のラック配置を所与として、ラック間を結ぶ通信の信号伝搬経路が地理的に遠回りとならないようにケーブルを敷設するものである。このとき、各ラックを構成単位としてスモールワールド性を持つトポロジを構成することにより、幾何学的方法論とグラフ理論的方法論の「いいとこ取り」をして、低遅延と低コストを高いレベルで両立することを目指す。また、本研究のパケットルーティング手法（以下 Fastest Routing と称する）は、各ケーブルの信号伝搬時間をあらかじめ計算しておき、ネットワークを重み付きグラフとしてモデル化することにより、遅延が最小となる経路を選択するものである。

図 2 は、64 台（16 台×4 列）のラックがフロア上に並ぶ大規模並列計算機の俯瞰図であり、一例として単純な Skywalk トポロジのリンクを黒線で表す。Skywalk トポロジ構成法は、まず、各ラック内のスイッチ群を完全結合する。次に、幅方向に並んでいるラック群（図 2 では 16 台）をランダムに結合する。最後に、通路を挟んで奥行方向に並んでいるラック群（図 2 では 4 台）をランダムに結合する。ラック同士を結合する場合、実際には両端のラック内のスイッチ同士を結合することになるが、このとき各スイッチの使用ポート数をできるだけ均等にする。各方向の結合に用いるスイッチポート数はパラメータとして設計者が決める。つまり、Skywalk には 3 つの整数パラメータがあり、その合計値がトポロジの次数となる。設計者

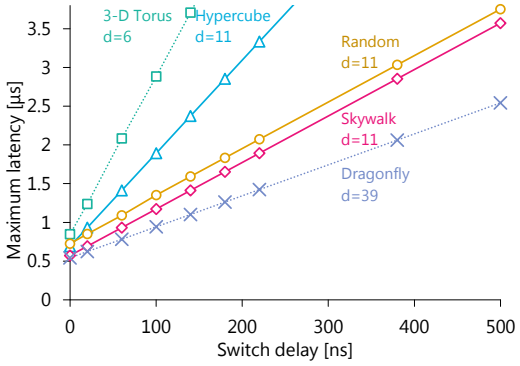

図 3　トポロジ別のグラフ解析結果（256 ラック）

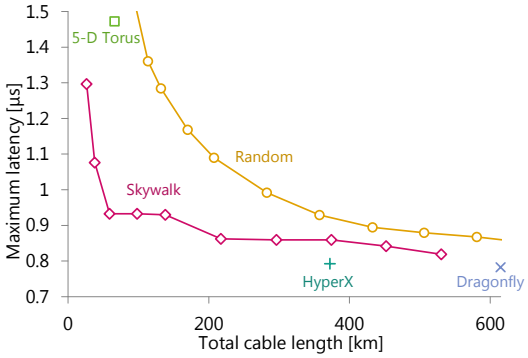

図 4　トポロジ別の配線延長と通信遅延（256 ラック）

は 3 つのパラメータを動かすことでポート数（＝コスト）と性能のトレードオフを探ることができる。

図 3 は、グラフ解析によって得られたスイッチ遅延（横軸：小さいほど未来）と最大通信遅延（縦軸：小さいほど良い）の関係を示す。11 ポートを用いる Skywalk トポロジは、同じポート数のランダムトポロジよりも低遅延である。また、39 ポートを用いる Dragonfly トポロジ（注 1）は Skywalk トポロジよりも低遅延だが、将来スイッチ遅延が 60ns となった時点では、両者の差はわずか 16%まで縮まる。したがって、将来の超低遅延ネットワークを設計する場合、Skywalk はきわめてコストパフォーマンスの高いトポロジ構成法となる。

図 4 は、256 台のラックをフロア上に配置したときの配線延長（横軸：短いほど良い）と通信遅延（縦軸：小さいほど良い）をトポロジ別に示す。Skywalk とランダムはパラメータを動かすことで配線延長と通信遅延のトレードオフが得られる。両トポロジで同じ通信遅延を達成するために必要な配線延長を比較すると、Skywalk はランダムよりも配線延長を大幅に削減できることがわかる。

図 5 は、フリットレベル・ネットワークシミュレーションによって得られたランダムトラフィックに対する各トポロジのスループット（横軸：大きいほど良い）と通信遅延（縦軸：小さいほど良い）の関係を示す。Skywalk トポロジと Fastest Routing 手法を組み合わせることで、他のトポロジと比べて

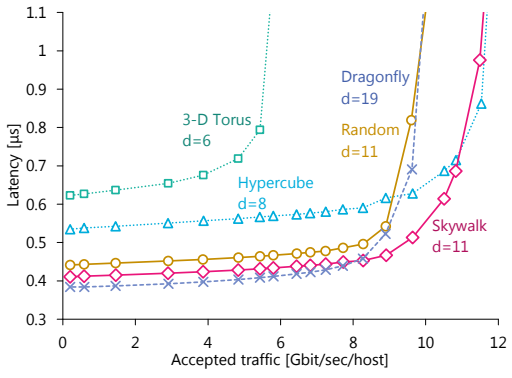

図 5　トポロジ別の性能評価結果（64 ラック）
d はトポロジの使用ポート数

少ないポート数で低遅延・高スループットを達成できることがわかる。

上述のように、本研究で開発した Skywalk トポロジは、本研究開始時点で最先端の研究成果だったランダムトポロジと比べ、ポート数が同じなら通信遅延を小さく、通信遅延が同じなら配線延長を短くできる。また、現在普及しつつある Dragonfly トポロジと比べ、同等以上の低遅延・高スループット性能を低コストで達成できる。

以上が平成 25 年度に得られた研究成果である。当初の想定を上回る成果が得られたことから、平成 26 年度は本研究において開発・改良したソフトウェア群（トポロジ生成ツール・グラフ解析ツール・ラック配置最適化ツール）を他の研究にも適用し、アイディアを融合することによる相乗効果を追究した。具体的には、(1)リンク交換による既設ネットワークの低遅延化技術との融合、(2) 敷設後に拡張可能な低次数トポロジ構成法との融合、(3)空間光無線ネットワーク技術との融合、(4)低遅延ネットワーク・オン・チップ設計法との融合である。本研究で開発したトポロジ生成・グラフ解析・ラック配置最適化等のソフトウェア群を適用することで、大はデータセンターから小はチップ内ネットワークまで、あらゆるインターコネクトの設計法を高度化・最適化できることを示した。

注 1　Dragonfly トポロジ: 2008 年に提案され、現在スパコン用として徐々に採用されつつあるネットワークトポロジ。ポート数の多いスイッチを使って低遅延を実現する。

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 3 件）

[1] Ikki Fujiwara, Michihiro Koibuchi, Hiroki Matsutani, Henri Casanova: "Swap-and-randomize: A Method for Building Low-latency HPC Interconnects", IEEE Transactions on Parallel and Distributed Systems (TPDS), 掲載確定, 2014 年 7 月, 査読有, DOI 10.1109/TPDS.2014.2340863

[2] 藤原一毅, 鯉渕道紘: “ランダムなネットワークトポロジのラック配置最適化に関する研究”, 電子情報通信学会論文誌, vol.J96-D, no.8, pp.1903-1912, 2013 年 8 月, 査読有, DOI なし

〔学会発表〕（計 10 件）

[1] Ryuta Kawano, Seiichi Tade, Ikki Fujiwara, Hiroki Matsutani, Hideharu Amano, Michihiro Koibuchi: “Optimized Core-links for Low-latency NoCs”, The 23rd Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP 2015), pp.172-176, 2015 年 3 月 4 日,トゥルク（フィンランド）, 査読有, DOI 10.1109/PDP.2015.15

[2] Ikki Fujiwara, Michihiro Koibuchi, Tomoya Ozaki, Hiroki Matsutani, Henri Casanova: “Augmenting Low-latency HPC Network with Free-space Optical Links”, The 21st IEEE International Symposium on High Performance Computer Architecture (HPCA 2015), pp.390-401, 2015 年 2 月 7 日, サンフランシスコ・ベイエリア（アメリカ）, 査読有, DOI 10.1109/HPCA.2015.7056049

[3] Nguyen T. Truong, Van K. Nguyen, Nhat T. X. Le, Ikki Fujiwara, Fabian Chaix, Michihiro Koibuchi: “Layout-aware Expandable Low-degree Topology”, The 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS 2014), pp.462-470, 2014 年 12 月 16 日, 新竹（台湾）, 査読有, DOI 10.1109/PADSW.2014.7097842

[4] 藤原一毅, 鯉渕道紘: “Pursuit of Low-latency Networks for Supercomputers”, 第 10 回情報科学ワークショップ, 2014 年 9 月 17 日, ツネイシしまなみビレッジ（広島県福山市）

[5] 河野隆太, 藤原一毅, 松谷宏紀, 天野英晴, 鯉渕道紘: “光サーキットの補助による低遅延性及びトポロジ内包性・分割性をもつネットワーク”, 2014 年並列／分散／協調処理に関する『新潟』サマー・ワークショップ (SWoPP 新潟 2014), 2014 年 7 月 29 日, 朱鷺メッセ（新潟県新潟市）

[6] Ikki Fujiwara, Michihiro Koibuchi, Hiroki Matsutani, Henri Casanova: “Skywalk: a Topology for HPC Networks with Low-delay Switches”, The 28th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2014), pp.263-272, 2014 年 5 月 19 日, フェニックス（アメリカ）, 査読有, DOI 10.1109/IPDPS.2014.37

[7] Van K. Nguyen, Nhat T. X. Le, Ikki Fujiwara, Michihiro Koibuchi: “Distributed Shortcut Networks: Layout-aware Low-degree Topologies Exploiting Small-world Effect”, The 42nd International Conference on Parallel Processing (ICPP 2013), pp.572-581, 2013 年 10 月 1 日, リヨン（フランス）, 査読有, DOI 10.1109/ICPP.2013.71

[8] Ikki Fujiwara, Michihiro Koibuchi: “Mapping Non-trivial Network Topologies onto Chips”, The

International Symposium on Embedded Multicore/Many-core System-on-Chip (IEEE MCSoC-13), pp.73-78, 2013 年 9 月 26 日, 一橋講堂（東京都千代田区）, 査読有, DOI 10.1109/MCSoC.2013.10
[9] 藤原一毅, 鯉渕道紘: “高次元トポロジ NoC の配線長最小化手法”, 2013 年並列／分散／協調処理に関する 『北九州』 サマー・ワークショップ (SWoPP 北九州 2013), 2013 年 7 月 31 日, 北九州国際会議場（福岡県北九州市）【情報処理学会 山下記念研究賞】
[10] Ikki Fujiwara: “Does light speed affect topologies?”, NII 湘南会議 031, 2013 年 9 月 23 日, 湘南国際村センター（神奈川県葉山町）

〔その他〕
鯉渕研究室ホームページ
http://research.nii.ac.jp/~koibuchi/

2013 年度国立情報学研究所オープンハウス
http://www.nii.ac.jp/userimg/openhouse/2013/leaf_op25pre.pdf

６．研究組織
(1) 研究代表者
藤原 一毅（Ikki Fujiwara）
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