氏　　名	TRUONG THAO NGUYEN
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2002 号
学位授与の日付	平成 30 年 3 月 23 日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Cable-Geometric and Moderate Error-Proof Approach for Low-Latency Interconnection Networks
論 文 審 査 委 員	主　査　准教授　鯉渕　道紘 教授　計　宇生 准教授　福田　健介 教授　米田　友洋 教授　五島　正裕 准教授　松谷　宏紀　慶應義塾大学　理工学部

(別紙様式 2)
(Separate Form 2)

論文の要旨
Summary (Abstract) of doctoral thesis contents

For decades, computational science and computer industry pursue a low latency interconnection network. In particular, as numerous endpoints of highly parallel computing systems (such as data centers or supercomputers) demand high computing and large storage, an interconnection network becomes a critical component of the highly parallel computing systems. The interconnection network is expected to provide a low latency and a high communication bandwidth. Indeed, the switch delay to forward a message becomes dozens or hundreds of nanoseconds. The receiving and sending overhead at a host could be hundreds of nanoseconds by enabling intelligent network interfaces. As device technology and its corresponding software overhead continue to improve, an expected Message Passing Interface (MPI) level communication naturally becomes latency sensitive, such as dozens of nanoseconds for custom supercomputers or hundreds for commodity clusters.

In this dissertation, we propose a new type of low latency interconnection network working towards Exascale and Big-Data computing challenges. Our approach is a combination of three following technologies; a short-diameter network topology with its low-cost physical layout, a custom routing avoiding latency overheads for accessing routing tables, and a low-latency error control mechanism for the usage of high-bandwidth optical links.

We first focus on designing short diameter network topologies and their physical layouts for achieving both low total switch latency and low total cable latency. We propose a new non-random approach named Distributed Shortcut Networks (DSNs), for designing a cost-effective and layout-conscious interconnect topology. The DSN approach is based on a common technique seen in traditional regular topology designs, e.g., the use of “virtual supernodes”, combined with a different design philosophy learned from observing small-world networks. We discussed both the ring-based DSN and the 2-D grid-based DSN topologies, where the ring-based DSN provides insight and theoretical analysis to the approach while the 2-D grid-based DSN provides competitive and practical use. By these techniques, DSN achieves logarithmic diameters which are in proportional to the total switch latency. DSN also has advantages of structured topological properties for shorter cable lengths when it is implemented into a machine room, i.e., leading to a low total cable latency. In addition, we illustrate that DSN is fully flexible to a required network size and to an incremental expandability after its installment on a machine room. When incrementally adding nodes and cabinets to the proposed topology, its diameter and average shortest path length increase modestly. Its network cost and power consumption modestly increase when compared to the counterpart non-random topologies.

Secondly, we study the ideal performance of a proposed interconnection network when a routing algorithm is implemented. Numerous endpoints require large CAM (Content Addressable Memory)-based routing tables at each switch consuming electric power drastically. A large CAM table implemented outside the router chip becomes a latency bottleneck especially when the switch delay becomes extremely small, e.g. 40ns. In this regard, avoiding routing table is a goal for a lower switch delay. We propose an effective non-minimal custom routing algorithm on DSNs without routing tables by computing directly of synthesis hardware at each switch. These latency analyses show that our custom routing algorithm achieves a good result in the low switch-latency era.

Thirdly, we address the high switch-delay problem by reducing the latency overhead of error correction codes in high-bandwidth optical communications, e.g., 100 Gb/s or 400 Gb/s. Recent high-bandwidth optical communication has error correction codes (ECC), such as forward error correction components (FEC) at each switch for maintaining the same bit error rate (BER) as that in traditional low-bandwidth interconnection networks. However, the FEC operation latency overhead becomes higher than the sum of all the other switch operation overhead including routing computation and switch allocation. The FEC operation overhead significantly degrades the performance of parallel applications in highly parallel computing systems especially when the BER is high. In this study, we design low-latency unreliable networks using a Hamming code. Although it does not provide rigid error-free communication, some parallel applications obtain the acceptable quality of computation results with short execution time.

Finally, we illustrate our best interconnection networks integrated with DSN topologies with its custom routing and the high-speed cables with our moderate error-proof approach. We show that our low-degree, cable-geometric moderate error-proof approach achieved a better performance compared with the same-degree counter-part network such as Torus or Random.

(別紙様式 3)
(Separate Form 3)

博士論文審査結果の要旨

Summary of the results of the doctoral thesis screening

本学位論文は、「Cable-Geometric and Moderate Error-Proof Approach for Low-Latency Interconnection Networks (低遅延相互結合網のためのケーブルジオメトリックおよび節度あるエラー防止のアプローチ)」と題し、全 7 章から構成されている。第 1 章「Introduction」では、相互結合網の研究分野の概況と目的を述べている。本章では 10 万計算ノード規模のスーパーコンピュータに代表される大規模計算機の相互結合網で要求される低遅延性、高帯域性、信頼性などについて述べ、現状の相互結合網の問題点を指摘し、本論文の目的がこれらの問題点を解決するための相互結合網のアーキテクチャおよび性能向上手法の提案であると述べている。第 2 章「Background」では、相互結合網の主要な構成要素であるネットワークトポロジ、ルーティング、通信リンクについて解説するとともに、各々に関して低遅延化を実現する最新の研究動向をまとめている。第 3 章「Network Topology」では、ネットワークトポロジに関して、遠くの計算ノードへのショートカットリンクの活用およびパケットの移動距離を抑制することで通信遅延に占めるケーブル伝搬遅延の削減の 2 点を特徴とする手法を提案している。提案したネットワークトポロジは、同じリンク数、スイッチ数を用いて構成された 3 次元トーラスと比べて、通信遅延が削減できることをネットワークシミュレーションにより示した。さらに、ネットワークトポロジの直径と平均最短経路長を解析した結果、提案手法により設計されたネットワークトポロジは 3 次元トーラスなどの実用化されているネットワークトポロジと比べてホップ数の面で優れていることを示している。また、実際の大規模計算機は、運用開始後に計算ノード数を増加させることが多いため、相互結合網の拡張性が必要となる。提案したネットワークトポロジは、計算ノードを追加した場合においてホップ数の増加を抑制する効果があり、有利であることを示している。第 4 章「Custom Routing」では、相互結合網の大規模化にともない、スイッチの処理時間に占めるフォワーディングテーブルのパケット参照時間の増大、および、複雑なネットワークトポロジの登場によりルーティング処理の複雑化の 2 つの問題点を指摘し、これらの問題点を解決するための手法として、計算ノードの位置情報を用いた計算アルゴリズムにより実装可能なカスタムルーティングを提案し、計算ノード間の通信時間が削減できることをネットワークシミュレーションにより示している。第 5 章「Moderate Error-Proof Approach」では、今後、相互結合網の通信リンクの帯域が大きくなるにつれて、通信リンクのビット誤り率が劇的に悪化し、その結果、エラー検出訂正処理に係る遅延がパケットの通信時間の支配的要因になる問題点を指摘している。この問題点を解決するために、アプリケーション毎に必要となる解の精度を最低限満たすように、エラー検出訂正処理を簡略化することでネットワークの低遅延化を実現する手法を提案している。提案手法は大規模計算機における並列計算の高速化が達成されることをネットワークシミュレーションにより示している。第 6 章「Integrated Interconnection Networks」では、第 3 章、第 4 章、第 5 章にて提案したネットワークトポロジ、カスタムルーティング、リンクに関する技術を統合した相互結合網について、大規模計算機における並列計算の更なる高速化が達成されることをネットワークシミュレーションにより示している。第 7 章「Conclusion」では、本研究を総括し、得られた成果お

(別紙様式 3)
(Separate Form 3)

よび今後の課題について述べている。本学位論文の成果として、出願者は主著で IEEE Transactions on Parallel and Distributed Systems を含む査読付き学術雑誌論文 2 件、フルペーパーの査読付き国際会議 1 件他として発表し、研究内容が国内外で認められている。以上を要するに本学位論文は、大規模計算機において必要となる低遅延通信を実現する相互結合網のアーキテクチャおよびネットワークトポロジ、カスタムルーティング、通信リンクという 3 つの構成要素に関して並列計算の実行性能を向上させるための手法を提案し、ネットワークシミュレーションの性能評価によりその有効性を示したものである。これらの成果は、今後さらなる大規模化が見込まれる計算基盤を構築するための技術的な問題を解決した点で、学術上貢献が大きい。よって、本学位論文は博士（情報学）の学位論文として価値あるものと認め、合格とする。