氏　　名　平澤 将一

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2328 号

学位授与の日付　2022 年 3 月 24 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　The Application of Routing Cache for High-Bandwidth Low-Latency Switch on Interconnection Networks

論 文 審 査 委 員　主　査　合田 憲人
情報学専攻 教授
五島 正裕
情報学専攻 教授
竹房 あつ子
情報学専攻 教授
鯉渕 道紘
情報学専攻 准教授
平木 敬
株式会社 Preferred Networks シニアリサーチャー
／東京大学 名誉教授

（様式３）

博士論文の要旨

氏　　名　平澤　将一

論文題目 The Application of Routing Cache for High-Bandwidth Low-Latency Switch on Interconnection Networks

Parallel applications become sensitive to communication latencies and bandwidth of interconnection networks between compute nodes on parallel computers. Interconnection networks have been studied for parallel computers, including supercomputers and high-end datacenters. Switch delay dominates communication latencies in interconnection networks, especially for short messages because switch delays are massive relative to the link and packet injection delays. At a conventional switch, routing decision is based on CAM (Content Addressable Memory) table lookup, and it imposes a significant delay. A main problem of the packet forwarding processing is the significant operation latency to the CAM at a switch. Reducing the access latency to CAM is crucial for the upcoming low-delay switch in parallel computers. Besides the CAM latency problem, the packet forwarding rate is not proportional to the switching capacity on cutting-edge commodity switches of interconnection networks. A switch will not be able to forward incoming packets at the maximum line rate. It is also difficult to provide the proportional packet forwarding rate to a high line rate on a future switch even for long packets. The key design to resolve the problem of the packet forwarding performance is a packet forwarding cache architecture explored in this dissertation. More precisely, we should find out “address patterns” of interconnection networks, and the packet forwarding cache architecture should be optimized for enjoying the address pattern.

To resolve the latency and throughput problems, an on-chip packet forwarding cache to a switch is explored. An incoming packet avoids large-latency accessing a CAM forwarding table if the cache hits. Firstly, the cache for up to 2K-node jobs is optimized because a large number of workloads are smaller sizes than 2K compute nodes. The conventional cache design supports an almost 100% hit rate (no capacity miss nor conflict miss) for packets generated in up to 2K-node jobs on arbitrary network topologies, which affect the access pattern of a switch. An incoming packet avoids large-latency accessing a CAM forwarding table if the cache hits. Only an exclusive layer-1 (L1) cache at an input port contributes to achieving a high line rate, e.g., 800Gbps for the incoming short packets. Zero-load communication latency with the packet forwarding cache in a large scale interconnection network is evaluated. From the evaluation results, the reduction percentage of zero-load communication latency gradually decreases from 19% to 13% with the effects of capacity misses of the packet

forwarding cache when used in a 9K computation node large scale interconnection network. Additionally, with additional entries in the packet forwarding cache, the reduction percentage of zero-load communication latency gradually increases from 9% to 19%. The adoption of the packet forwarding cache clearly decrease the communication latency of interconnection network and increase both the line rate and the performances of parallel applications Consequently, the packet forwarding cache is strongly recommended to be adopted in HPC switches. However, larger jobs make the cache hit rate almost “zero” on any network topologies, and the cache effect becomes almost “zero.”

Secondly, a switchable node reduction function to refer to a packet forwarding table on a switch is presented for 100% hit rate on larger jobs. The main idea is that a large number of packet destinations share a same index tag, resulting in the same required number of cache entries as the number of output ports. This design can be enabled by the path regularity of the above network topologies. A general node reduction function, which obtains two addresses and their indices and returns a cache tag, is defined to achieve the path regularity. The switchable node reduction function is then optimized to typical network topologies, i.e., k-ary n-cubes, fat trees, and Dragonfly. Evaluation results show that the reasonable packet forwarding cache supports a 933Gbps line rate even for incoming shortest packets on the above network topologies. In addition, they illustrate that parallel applications obtain the performance gain of 5.07x speed up using the cache switches since the impact of the switch delay and link bandwidth is significant on the end-to-end communication performance.

Through this dissertation, it is concluded that a commodity switch should have a packet forwarding cache with switchable node reduction functions. The packet forwarding cache is efficient for forwarding a large number of shortest packets, and the switchable node reduction function is necessary for large scale parallel computers.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　平澤 将一

Title
論文題目　The Application of Routing Cache for High-Bandwidth Low-Latency Switch on Interconnection Networks

本学位論文は、「The Application of Routing Cache for High-Bandwidth Low-Latency Switch on Interconnection Networks（相互結合網における広帯域低遅延スイッチのためのルーティングキャッシュの応用)」と題し英文で記述され、全 7 章から構成されている。

第 1 章「Introduction」では、HPC（ハイパフォーマンスコンピューティング）システムの相互結合網の研究分野の概況と本研究の目的を述べている。本章では相互結合網で要求されるスイッチのスループットとルーティング処理について述べ、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スイッチアーキテクチャの提案であると述べている。

第 2 章「Background Information」では、相互結合網のネットワークトポロジと、そのスイッチのマイクロアーキテクチャについて解説するとともに、各々に関してスループットと通信遅延の向上を実現する研究動向をまとめている。

第 3 章「Problem Statement」では、最近のスイッチのスループットがパケットのルーティング処理性能に律速されつつある点を指摘し、現状の汎用スイッチのマイクロアーキテクチャでは、スループットの向上が困難であることを示している。

第 4 章「Packet Forwarding Cache」では、スイッチ内のルーティング処理のためにキャッシュ機構を提案している。本キャッシュはスイッチング ASIC（Application Specific Integrated Circuit）の中に配置され、キャッシュに格納された目的地宛てのパケットは、CAM（Content Addressable Memory）にアクセスすることなくルーティング処理が完了するため、既存のスイッチと比べてスループットの向上が期待できる。本キャッシュの基本構成はセットアソシアティブ方式である。出願者は、NAS 並列ベンチマークなどの並列計算の実行時間の大幅な短縮効果をシミュレーションにより示している。また、キャッシュのハードウェア量は、十分に小さいことをシミュレーションにより示している。

第 5 章「Switchable Node Reduction Function」では、第 4 章にて提案したキャッシュの基本構成では、任意のネットワークトポロジに対応できる一方、2,000 台以上の計算ノードで構成された大規模 HPC システムではスループット向上の効果が薄いことを指摘している。そして、大規模化に対応するため、一部のネットワークトポロジのサポートに限定するハードウェアノード削減関数を提案し、キャッシュ機構に追加している。出願者は、Node Reduction Function のハードウェア量は、十分に小さいことをシミュレーションにより示し、大規模 HPC システムにおいてもスイッチのスループット向上が達成できることを明らかにしている。

第 6 章「Discussions」では、第 4 章、第 5 章で提案したキャッシュの他の利用方法とサポートする HPC システム規模について検討を行っている。
第 7 章「Conclusions」では、本研究を総括し、得られた成果および今後の課題について述べている。

公開発表会では博士論文の章立てに従って発表が行われ、その後に行われた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究が相互結合網のスイッチのルーティング処理性能とスループットの向上に貢献することが評価された。以上を要するに本学位論文は、HPC システムにおいて必要となる広帯域低遅延通信を実現するために、相互結合網のスイッチにルーティング処理を行うキャッシュアーキテクチャを提案し、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。また、本学位論文の成果は、学術雑誌論文 1 件、フルペーパー査読付き国際会議論文 1 件として発表され、学術的な貢献も認められる。以上の理由により、審査委員会は、本学位論文が学位の授与に値すると判断した。