情報処理学会論文誌　コンピューティングシステム　Vol.13　No.2　1–12 (Sep. 2020)

HPCスイッチにおけるルーティングテーブルキャッシュの研究

平澤 将一[1,2,a)]　八巻 隼人[3]　鯉渕 道紘[1,2,4]

受付日 2020年1月14日, 採録日 2020年5月21日

概要：HPC（High-Performance Computing）システムにおける並列アプリケーションの性能は，計算ノード間の相互結合網の通信遅延により影響を受ける．相互結合網においてパケットは複数のスイッチを経由して転送されるため，特にメッセージサイズが小さい場合に，スイッチ遅延が通信遅延の支配的要因である．典型的なスイッチでは，ルーティング処理がオフチップCAM（Content Addressable Memory）に基づくテーブルルックアップで実装されるため，大きな遅延が生じる．そこで，本研究ではHPCスイッチにルーティングテーブルキャッシュを適用した場合のスイッチ遅延の削減効果を示す．本HPCスイッチでは，各入力ポートにルーティングテーブルキャッシュを配置する．本キャッシュがヒットした場合，CAMアクセスを避けることができるため，スイッチ遅延を削減することが期待できる．キャッシュのシミュレーション結果より，4ウェイの連想数で2,048エントリのキャッシュを有するスイッチで構成された相互結合網において，512台の計算ノード間の通信に対する競合性ミスの発生は0.1%以下となることが分かった．また，SimGridシミュレーションの結果，256台のスイッチを用いた相互結合網において，ルーティングテーブルキャッシュの導入により，NAS並列ベンチマークの性能を平均6.9%向上させることに成功した．さらに，大規模な相互結合網の解析結果より，ジョブサイズが十分に大きい場合に生じる容量性キャッシュミスが与える通信遅延への影響は限定的であり，本キャッシュを導入することにより，無負荷通信遅延を13%から19%と大幅に削減できることが分かった．

キーワード：スイッチアーキテクチャ，相互結合網，キャッシュ

A Study of Routing-table Cache on HPC Switches

Shoichi Hirasawa[1,2,a)]　Hayato Yamaki[3]　Michihiro Koibuchi[1,2,4]

Received: January 14, 2020, Accepted: May 21, 2020

Abstract: Parallel applications become sensitive to communication latencies of interconnection networks between compute nodes on HPC (High-Performance Computing) systems. Switch delay dominates communication latencies in interconnection networks especially for short messages, because a packet is transferred to a destination via multiple intermediate switches. At a conventional switch, routing decision is based on off-chip CAM (Content Addressable Memory)-based table lookup, and it imposes a significant delay. In this study, we exploit the application of on-chip routing-table cache for HPC switches. We place routing-table cache at each input port on the switch. The routing-table cache can bypass the CAM table lookup when it hits, then significantly reducing the switch delay. Our cache simulation results show that a 4-way set associative cache with 2,048 entries has less than 0.1% of the conflict miss rate on 256-nodes interconnection networks. Our SimGrid simulation results show that the introduction of routing-table cache on each switch improves 6.9%, in average, of performance of NAS Parallel Benchmarks on 256-node interconnection networks. Our analysis results show that the impact of the capacity cache miss on the communication latency is negligible even if a job size becomes large. The routing-table cache efficiently reduces the zero-load communication latency by 13% to 19%.

Keywords: switch architecture, interconnection networks, cache

1. はじめに

高性能計算（High-Performance Computing：HPC）システムにおけるマルチコア並列アプリケーションは，数百ナノ秒〜1 マイクロ秒の低 MPI（Message Passing Interface）通信遅延の達成が必要である [1]．したがって，低遅延相互結合網の研究が，これからの高性能計算システムの開発に重要となる．MPI 通信の遅延は，計算ノードの送受信遅延，複数本のリンク遅延，そして複数台のスイッチ遅延の和であるが，Mellanox 社 SB7790 InfiniBand EDR 100 Gbps スイッチの遅延が約 90 ナノ秒であるなど，スイッチ遅延が支配的である．したがって，スイッチの遅延を削減することが重要である．

スイッチにおけるパケット処理は，ルーティング，出力ポートのバッファの割当て，クロスバーの割当て，および入出力ポート間のパケット転送の 4 つに分けることができる．このなかで，本研究では，スイッチにおけるルーティング処理の遅延削減に焦点を当てる．

一般的に，スイッチのルーティングテーブル[*1]は，パケットの目的地情報を入力とし，パケットが転送されるスイッチの出力ポート情報を出力するため，エントリ数は大きい．たとえば，Mellanox 社 MTS3600（36 ポート，40 Gbps InfiniBand Switch）では 48K エントリ，同社 SN3000 シリーズのデータセンター用スイッチ（32〜48 ポート，最大 400GigabitEthernet（GbE））では 512K エントリを有する．よって，ルーティングテーブルは CAM（Content Address Memory）を用いて実装することが有力である．

CAM 自体は，近年の製品では 200 MHz 前後の動作周波数，すなわち 4〜5 ナノ秒程度のサイクル遅延でアクセスが可能である [2]．これに加えて，スイッチチップの Network-on-Chip（NoC）とスイッチチップ–CAM チップ間の往復通信に 20 ナノ秒程度の遅延を要する．つまり，ルーティング処理に 25 ナノ秒程度のアクセス遅延が生じる．そのため，パケットのスイッチ遅延において，CAM アクセスが占める割合は大きい．

そこで，本研究では，スイッチ遅延を削減するために，ルーティングテーブルキャッシュを各入力ポートに有するスイッチ構成を探求し，本キャッシュが与えるシステム性能への影響を評価する．我々は先行研究において，HPC スイッチに本キャッシュを導入した場合の小規模相互結合網のスループットと通信遅延について，サイクルレベルのシミュレーションにより評価済み [3], [4] である．しかし，詳細なキャッシュヒット率の解析，大規模相互結合網に対する本キャッシュの通信性能への影響，および並列アプリケーションに対する本キャッシュの実行性能への影響は示されていない．

[1] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[2] 総合研究大学院大学
SOKENDAI, Hayama, Kanagawa 240–0193, Japan
[3] 電気通信大学大学院
The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[4] 国立研究開発法人科学技術振興機構，さきがけ
JST, PRESTO, Chiyoda, Tokyo 101–8430, Japan
[a)] hirasawa@nii.ac.jp

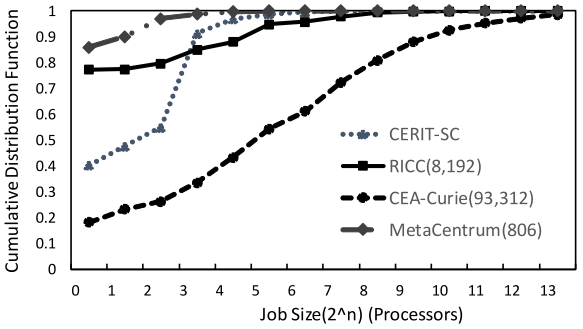

図 1　並列ジョブサイズの累積分布
Fig. 1　The cumulative distribution of parallel-job sizes.

ルーティング処理において，本キャッシュがヒットした場合，CAM アクセスが発生しないため，通信遅延を削減することができる．一方，キャッシュがヒットしなかった場合，キャッシュへのアクセス遅延に加えて，CAM アクセスが生じるため，通信遅延が大きくなる．ただし，最近の HPC システムでは，1 つの並列プログラムをシステム全系で動かすことはまれであり，一般的には同時に多数の小規模な並列プログラムを，異なる隣接計算ノード群に割り当て，異なるジョブで生じた通信が相互に干渉しないように実行される．したがって通常の HPC システムの運用において，ジョブサイズは数千程度が一般的な最大値となることから，本キャッシュのヒット率はきわめて高くなる場合が多いと考えられる．

図 1 は，CERIT-SC，RICC，CEA-Curie，MetaCEntrum スパコンにおけるユーザジョブのサイズを累積分布で示している．これは，Parallel workloads archive（http://www.cs.huji.ac.il/labs/parallel/workload/）および文献 [5] で用いられているトレースから解析した．

図 1 において，凡例中の括弧内の数字は，そのスパコンが有するプロセッサ総数である．ここで横軸は，計算ノード数ではなくプロセッサ数である．たとえば，RICC では 1 台の計算ノードあたり 8 プロセッサ数を有し，CEA-Curie では CPU と GPU を併用し，計算ノードごとに異なる多数のプロセッサ数を有する．したがって，各ジョブが利用する計算ノード数は，図 1 で示したプロセッサ数と比べて大幅に小さくなる．

図 1 より，実際の HPC システムでは，RICC において 98%のジョブが 128 個のプロセッサ（32 台の計算ノード）数以下で実行されているなど，多くの場合ジョブサイズ

[*1] InfiniBand ではフォワーディングテーブル，イーサネットでは MAC アドレステーブルと呼ばれる．

は小さい．1 つのジョブの実行中に生成されるパケットの目的地の総数は，そのジョブを実行する計算ノード数となる．よって，この目的地数に対して十分な容量のキャッシュが実装可能である場合，キャッシュヒット率が限りなく 100%に近づくことが期待できる．

本論文の貢献は以下である．

(1) キャッシュのシミュレーション結果より，4 ウェイの連想数で 2,048 エントリのキャッシュを有するスイッチで構成された相互結合網において，512 台の計算ノード間の通信に対する競合性ミスの発生は 0.1%以下となることが分かった（4 章）．

(2) SimGrid v3.12 を用いてアプリケーション性能を評価した結果，256 台のスイッチを用いた相互結合網において NAS 並列ベンチマークの性能を平均 6.9%向上させることが分かった（5 章）．

(3) 相互結合網の解析結果より，本キャッシュは，ネットワークサイズが十分に大きい場合に容量性ミスを引き起こす．そのため，実行するジョブサイズが数百から数万ノード規模へと大きくなると，無負荷通信遅延の削減効果が 19%から 13%へと緩やかに低下することが分かった．また，予想どおり，本キャッシュのエントリ数が大きいほど，同遅延の削減効果が 9%から 19%へと緩やかに大きくなることが分かった（6 章）．

本論文の構成は以下である．2 章では，本研究の背景となる技術を紹介し，3 章では，HPC スイッチにおけるルーティングテーブルキャッシュの構成について述べる．4 章では，ルーティングテーブルキャッシュの競合性ミス率を解析する．5 章では，たかだか，数百計算ノードを用いた並列計算アプリケーションの実行時間の面でルーティングテーブルキャッシュの効果を評価する．6 章では，数万計算ノード間通信の無負荷通信遅延の面でルーティングテーブルキャッシュの効果を解析する．7 章では，ルーティングテーブルキャッシュの導入による電力面とスループット面の影響に関して議論を行う．8 章では，まとめと今後の課題を述べる．

2. 背景

2.1 Content Address Memory（CAM）

スイッチにおけるルーティングの実装は CAM の利用が有力である．Anton-2 や IBM BlueGene/Q のように，専用の規則（例：トーラストポロジにおける次元順ルーティング）に従ってルーティングする専用の相互結合網では，（CAM を用いずに）ハードウェア合成によるルーティング実装が可能であり，スイッチ遅延は 40 ナノ秒程度ときわめて小さい．しかし，そのような専用相互結合網の利用は HPC 分野において限定的である．

一方，CAM は，HPC 分野において幅広く使われている InfiniBand のように，任意のネットワークトポロジをサポートするスイッチのルーティングテーブルの実装に適している．しかし，ルーティングテーブルはネットワークアドレス（InfiniBand では 16 bit の Local ID）すべてに対して経路情報を持つことから，テーブルサイズが大きくなる．よって，CAM をスイッチチップとは別のチップに実装することが有力である．

ルータやスイッチに搭載される数十 Mbit CAM へのアクセス遅延は過去 10 年にわたり，製品レベルでは大きく改良されていない．たとえば，ルネサスエレクトロニクス社の CAM では，2009 年に出荷した R8A20410BG のアクセス遅延は 6 ナノ秒程度であるのに対し，最新版である R8A20611BG でもアクセス遅延は 4 ナノ秒程度となる．スイッチチップから別チップの CAM へのアクセス遅延は，CAM 自体のアクセス遅延，ルータチップの Network-on-Chip，SerDes 変換を含めたチップ間通信の往復遅延の和となるため，25 ナノ秒程度となる．この遅延は，プロセッサチップにおいても，オフチップ L3 キャッシュのメモリ自体は 3 ナノ秒程度で動作可能だが，プロセッサからのアクセスには 20 数ナノ秒 [6] を要するという報告とも近い値で一貫性がとれている．

CAM は高消費電力であることが問題視されており，論文 [7] によると TCAM（Ternary CAM）の消費電力は同容量の SRAM（Static Random Access Memory）の 15 倍と報告されている．そのため，静的電力を削減するパワーゲーティング技術を導入する研究 [8] や，動的電力を削減する DVFS を導入する研究 [9] がなされているが，いまだ大幅な電力の削減にはいたっていない．

2.2 スイッチマイクロアーキテクチャ

図 2 に，結合網の教科書 [10] で述べられている典型的なスイッチ構成に基づくブロック図を示す．**図 3** に，そのスイッチにおける 4 フリット長のパケットに対するパイプライン処理の流れを示す．このスイッチのパイプラインは，ルーティング（Routing Computation：RC），出力ポート

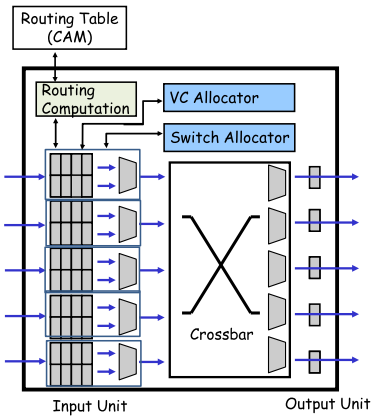

図 2　典型的な 5 × 5 スイッチのブロック図
Fig. 2　Block diagram of conventional switch.

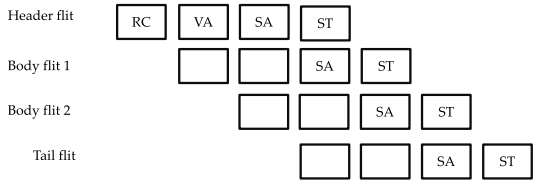

図 3　スイッチにおけるパケットのパイプライン処理
Fig. 3　Pipeline processing of a packet.

のバッファの割当て（Virtual-channel Allocation：VA），クロスバーの割当て（Switch Allocation：SA），および入出力ポート間のパケット転送（Switch Transfer：ST）の 4 つのステージに大別される．

図 3 は，スイッチのパイプラインステージごとの処理を示している．1 ステージ目では，パケットヘッダが到着後にルーティング処理が開始され，出力ポートの選択を行う．2 ステージ目では，ルーティング処理結果に基づき，出力ポートのチャネルバッファの確保を行う．3 ステージ目では，そのパケットが出力ポートに転送できるようにクロスバーの割当てを行い，4 ステージ目では，パケットヘッダが入力ポートから出力ポートへ転送される．これら 4 ステージの処理は，実際のスイッチでは各々複数サイクルに細分化されて実装される [10]．

商用スイッチの詳細なパイプライン構成とサイクル数はブラックボックスのことが多い．ただし，IBM BlueGene/Q ルータでは 40 ナノ秒（500 MHz，20 cycles）[11]，Cray 社 YARC ルータでは 31.25 ナノ秒（800 MHz，25 cycles）[12]，富士通社製 10 GbE スイッチでは 450 ナノ秒（312.5 MHz，140 cycles）[13]，RHiNET-2/SW では 160 ナノ秒（125 MHz，20 cycles），RHiNET-3/SW では 240 ナノ秒（100 MHz，24 cycles）[14] については公開されている．

本研究ではスイッチのマイクロアーキテクチャに関して，ルーティング処理以外のステージには改良を施さない．よって，ルーティング処理以外のステージのサイクルレベルの挙動については，本研究ではこれ以上言及しない．

2.3　ルーティングテーブルキャッシュ

ルーティングテーブルキャッシュは，電気インターネットルータのパケット転送のスループット向上のために提案された [15]．前述のように，CAM によるルーティングテーブル検索は遅延が大きく，インターネットのバックボーン電気ルータにおけるスループット上のボトルネックとなる [16]．そこで，ルーティングテーブルキャッシュでは，CAM アクセスの前段に高速なキャッシュを導入し，キャッシュヒットした場合には CAM アクセスをバイパスすることでルータのスループットを向上させる．ルーティングテーブルキャッシュと同様のキャッシュ機構を持つ最新の研究 [17] では，64 バイトの最短パケット長を仮定したパケット転送においても 1 Tbps のスループットをラインレートで実現可能であることを示した．

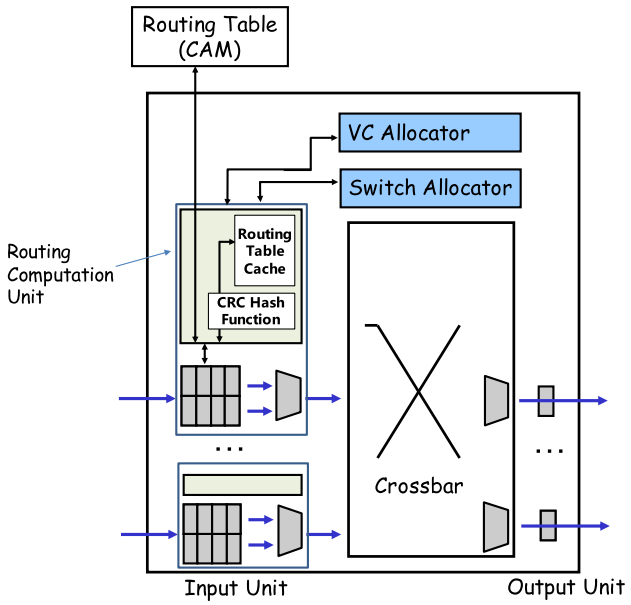

図 4　ルーティングテーブルキャッシュを有するスイッチのブロック図
Fig. 4　Block diagram of switch with routing table caches.

しかし，我々の知る限り，HPC 分野においてスイッチにルーティングテーブルキャッシュを用いる研究は我々の先行研究 [3], [4] のみであり，詳細な通信遅延削減効果と実際の HPC アプリケーション性能への影響については分かっていない．

3.　HPC スイッチにおけるルーティングテーブルキャッシュ

3.1　ルーティングテーブルキャッシュの動作

ルーティングテーブルキャッシュを有するスイッチにおいて，ルーティング処理以外の挙動は図 2 に示した典型的なスイッチの場合と同様である．そこで，ここでは，スイッチのルーティング処理について詳細に述べる．

図 4 に，本研究対象であるルーティングテーブルキャッシュを有するスイッチのマイクロアーキテクチャを示す．ルーティングテーブルキャッシュを有するスイッチは，典型的なスイッチ（図 2）のルーティング部に対して，新たにルーティングテーブルキャッシュ機構を導入した構成をとる．ルーティングテーブルキャッシュは，スイッチの入力ポートごとに備えられる．これは，同じ宛先であっても一般には出力ポートが異なるルーティングを行う可能性があるためである．

図 2 で示した典型的なスイッチでは，スイッチの入力ポートに到着したパケットが入力ポートのチャネルバッファに蓄えられた後，CAM アクセスを通して，そのパケットの出力ポート情報を得る．一方，本スイッチでは，スイッチの入力ポートに到着したパケットが入力ポートのチャネルバッファに蓄えられた後，自ポートのルーティングテー

ブルキャッシュを参照する．この際，参照するキャッシュラインは，パケットヘッダに含まれる目的計算ノード情報から計算されるハッシュ値をもとに決定される．参照したキャッシュラインのタグ情報とパケットの目的計算ノードが一致，すなわちキャッシュヒットした場合には，キャッシュから当該パケットの出力ポート情報を得る．一方で，パケットがキャッシュミスした場合には，スイッチファブリックを介して CAM にアクセスすることで出力ポート情報を得る．したがって，本スイッチではキャッシュヒットした場合に CAM アクセスが省略される．なお，CAM はすべての目的地アドレスを格納できる十分な容量を有していることとする．

本スイッチのルーティング処理は，(1) ハッシュ値計算，(2) キャッシュアクセス，(3) CAM アクセスの 3 ステージのパイプラインで構成される．各ステージに要するサイクルは以下のとおりである．ハッシュ値計算は CRC（Cyclic Redundancy Check）を用いて 1 サイクルで実行される．キャッシュアクセスは，オンチップキャッシュの場合，キャッシュ容量により 1 サイクルから数サイクル程度を要する．本論文では，4.1 節で後述するように 24 KiB の小容量なキャッシュを想定しており，そのアクセス遅延は CACTI6.5 [18] による評価では 32 nm プロセステクノロジで 0.41 ナノ秒である．これは，1 GHz 以下で動作するスイッチにおいてヒット/ミス判定と合わせても 1 サイクルに収まる遅延である．CAM アクセスは，前段のキャッシュアクセスにおいてパケットがキャッシュミスした場合にのみ実行され，前述したように 25 ナノ秒（1 GHz 動作のスイッチで 25 サイクル）程度を要する．

3.2　キャッシュ構成

3.2.1　データ構造

ルーティングテーブルキャッシュのエントリは，キャッシュタグとして目的計算ノードの識別子 8 バイト，キャッシュデータとして出力ポートの識別子 2 バイトおよび QoS などで用いる予約フィールド 2 バイトの計 12 バイトで構成される．

3.2.2　階層

各入力ポートに設置するキャッシュの容量は，多くの場合で実行中のジョブサイズよりも十分に大きいことが想定される．したがって，多くのジョブ実行時に生じる通信に関しては，キャッシュの容量性ミスは生じない．そこで，本研究では各入力ポートで共有する大容量でオンチップの 2 次ルーティングテーブルキャッシュは用いず，各入力ポートに 1 つだけキャッシュがあり，その入力ポートを排他的に使うこととする．

3.2.3　一貫性保持プロトコル

プロセッサのキャッシュと異なり，（ルーティング情報の更新がない限り）キャッシュデータの更新がない．そのため，更新が必要な場合はキャッシュを一度空にする単純な制御で十分と考えられ，WAW（Write After Write），RAW，WAR のデータハザードに対する一貫性処理は検討しない．

4.　ルーティングテーブルキャッシュの競合性ミス率の解析

一般的に，キャッシュミスは，競合性ミス，初期参照ミス，および容量性ミスの 3 つが原因となる．キャッシュエントリ数に対して十分に小さい計算ノード数で構成されるジョブ内で生じる通信では，容量性ミスはほとんど生じない．また，初期参照ミスはプログラム実行前に計算ノード間でメッセージ交換を行うことで防ぐことができる．そこで，本章では，ルーティングテーブルキャッシュの競合性ミス率について評価を行う．競合性ミス率は，連想数に大きく影響を受けるため，連想数とキャッシュの競合の発生率について解析する．

4.1　条件

キャッシュ容量は，容量性ミスとスイッチチップの面積増加との兼ね合いから，一般的なプロセッサが備える L1 キャッシュと同程度の 24 KiB（=2,048 エントリ）とした．24 KiB とすることで，ジョブサイズが 2,048 以下の場合，容量性ミスは発生しない．ルーティングテーブルキャッシュによるスイッチチップの面積増加は，CACTI6.5 [18] を使用し，高性能トランジスタモデル，32 nm と 65 nm プロセステクノロジの 2 つのケースについて見積もったところ，各々，ポートあたり 0.18 mm^2 と 0.73 mm^2 にとどまることが分かった．京コンピュータのインターコネクトに用いられていた ICC チップ（10 ポート，富士通セミコンダクタ 65 nm プロセステクノロジ）の面積が 329 mm^2 [19] であることを参考にすると，ルーティングキャッシュ機構は 10 ポート分で 7.3 mm^2，すなわち約 2%の面積増加にとどまる．なお，CRC ハッシュの計算ハードウェアは数百 μmm^2 程度 [20] であり，スイッチチップの面積にはほぼ影響を与えない．

トラヒックパターンは，目的地アドレスの分布が均一かつ，ハッシュ値を用いることで一様分布になるユニフォームトラヒックと，MPI 版 NAS 並列ベンチマーク 3.3.1 の通信トレースを用いた．NAS 並列ベンチマークの問題サイズとランク数は，それぞれ Class A，256 とした．NAS 並列ベンチマークのキャッシュヒット率の評価において，ネットワーク構成は，3 次元トーラス（$4 \times 4 \times 4$）とし，各スイッチに計算ノードが接続されていることとした．

4.2　評価結果（ユニフォームトラヒック）

図 5 は，連想数と競合により追い出されたエントリの割合をソフトウェアシミュレーションにより解析した結果を示している．

評価結果より，スイッチへの挿入データ数 512 の場合は

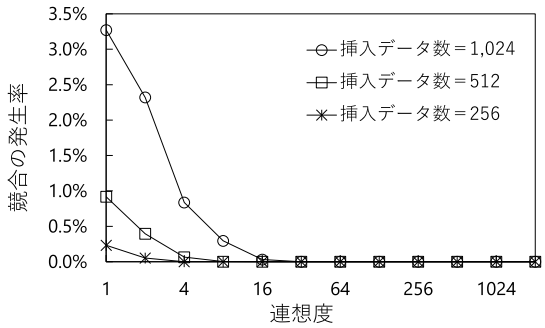

図 5　連想数と競合により追い出されたエントリの割合（ユニフォームトラヒック）

Fig. 5　Associative number and ratio of entries replaced by conflicts (uniform traffic).

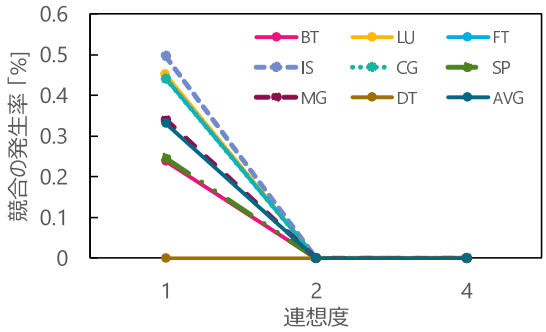

図 6　連想数と競合により追い出されたエントリの割合（NAS 並列ベンチマーク）

Fig. 6　Associative number and ratio of entries replaced by conflicts (NAS parallel benchmarks).

4 ウェイで，また挿入データ数 1,024 の場合は 16 ウェイで競合率 0.1%以下を達成している．つまり，1,024 台の計算ノードを用いるジョブまでは 16 ウェイの連想数のキャッシュを実装することにより，競合性ミスをほぼ 0 とすることが期待できる．図 1 に示したように，多くの HPC システムでは 90%以上のジョブがジョブサイズ 1,024 以下であるため，本キャッシュ構成によって大多数のジョブで競合性ミスの影響はきわめて小さいことが分かる．

4.3　評価結果（NAS 並列ベンチマーク）

図 6 は，NAS 並列ベンチマーク実行時のルーティングテーブルキャッシュの連想数と競合により追い出されたエントリの割合を示している．

いずれのアプリケーションにおいても，連想度が 2 と 4 の場合，競合ミスは発生しなかった．つまり，初期参照ミスを除くと連想度 2，4 において 100.0%のキャッシュヒット率が得られることが分かった．なお，図 6 には提示していないが，DT を除くすべてのアプリケーションにおいて，連想度 4 では初期参照ミスを含めても 99.7〜100.0%ときわめて高いヒット率[*2]に達した．

[*2]　総通信回数の少ない DT では 95.6%であった．

表 1　SimGrid シミュレーションのパラメータ

Table 1　Parameters of SimGrid simulation.

ケーススタディ 1	
キャッシュ無スイッチのルーティング遅延	25 ナノ秒
キャッシュ有スイッチのルーティング遅延（ヒット時）	2 ナノ秒
同（ミス時）	27 ナノ秒
ルーティング以外の処理遅延	75 ナノ秒
計算ノードの性能	5 TFlops
ネットワーク構成	3 次元トーラス
ケーススタディ 2	
キャッシュ無スイッチのルーティング遅延	20 ナノ秒
キャッシュ有スイッチのルーティング遅延（ヒット時）	2 ナノ秒
同（ミス時）	22 ナノ秒
ルーティング以外の処理遅延	40 ナノ秒
計算ノードの性能	50 TFlops
ネットワーク構成	ランダムトポロジ

5.　並列アプリケーションの評価

本章では，ルーティングテーブルキャッシュを有するスイッチを用いた相互結合網と，ベースラインとなる同キャッシュを持たない相互結合網について，科学技術並列計算のベンチマークの実行時間から比較評価する．

5.1　シミュレーション条件

本章では，SimGrid v3.12 [21] を用いてアプリケーション性能を評価する．2 つの並列計算システム構成を対象とし，各々の評価環境とパラメータを**表 1** に示す．ケーススタディ 1 は保守的なシステム構成，ケーススタディ 2 はアクセラレータを搭載した計算ノードを用いたハイエンドなシステム構成を想定している．

ケーススタディ 1 では，ベースラインとするキャッシュ無スイッチにおいて，ルーティング遅延を 25 ナノ秒とした．ルーティング遅延（25 ナノ秒）は，2.1 節で述べたとおり，CAM チップへの往復アクセス遅延（Network-on-Chip とチップ間通信）と，CAM 自体の処理遅延の和である．それ以外のスイッチ処理遅延を 75 ナノ秒とした．これは，出力ポート（と仮想チャネル）のバッファの割当て，クロスバーの割当て，入出力ポート間のパケット転送遅延の和である．すなわち，パケット転送に必要となるスイッチの最小遅延は 100 ナノ秒となる．これは Mellanox 社 SB7790 InfiniBand EDR 100 Gbps スイッチの遅延（約 90 ナノ秒），HPCI ロードマップ白書 [22] で 2018 年頃を想定したスイッチ遅延（100 ナノ秒/hop），相互結合網の教科書 2 章 [10] におけるベースラインのスイッチ遅延（100 ナノ秒）と同様の水準である．

一方，キャッシュ有スイッチのルーティング遅延は，キャッシュヒット時 2 ナノ秒とした．これは，3.1 節で述べたように，ハッシュ値の計算 1 サイクルとキャッシュアクセス遅延 1 サイクルの和であり，1 GHz 動作のスイッチチップを想定した場合の 2 サイクルに相当する．

キャッシュミス時のルーティング遅延は，ミスペナルティとして，キャッシュ無スイッチのルーティング遅延と同一（25 ナノ秒）とした．つまり，キャッシュ有スイッチの遅延は，キャッシュがヒットした場合 77 ナノ秒，キャッシュが外れた場合は，102 ナノ秒となる．

ケーススタディ 2 では，キャッシュ無スイッチにおけるルーティング処理遅延を 20 ナノ秒とし，それ以外のスイッチ処理遅延を 40 ナノ秒とした．よって，このスイッチ遅延は 60 ナノ秒となる．一方，キャッシュ有スイッチでは，キャッシュヒット時 42 ナノ秒，キャッシュミス時 62 ナノ秒となる．このスイッチ処理遅延は 2.2 節で述べた実スイッチのなかで，最も遅延の小さいスイッチに準じる値である．

いずれのケーススタディともに，各スイッチは 6 台の隣接スイッチと相互接続し，別途，1 台のプロセッサノードと接続している．リンクバンド幅は 200 Gbps とする．アプリケーションは，MPI 版 NAS 並列ベンチマーク 3.3.1 を用いる．問題サイズは，シミュレーション実行時間の制約から IS，FT のみ Class A，他は Class B とする．前章の結果より，連想度 4 における NAS 並列ベンチマーク実行時の容量性ミス率，競合性ミス率はいずれも 0%である．初期参照ミスは，ジョブの実行前に，利用する計算ノード間で最小バイト数の全対全通信を行うことで回避することができる．そこで，本評価では単純にキャッシュミスが生じないことと仮定する．

5.2　評価結果（ケーススタディ 1）

NAS 並列ベンチマークの評価結果を**図 7**，**図 8** に示す．縦軸は Millions of Operations Per Second（MOPS 値）の相対値であり，ベースラインとなるキャッシュのないスイッチ（Conv. Switch）を利用した場合を 1.0 としている．相対 MOPS 値は大きいほどアプリケーションの実行性能が高いことを示している．評価結果より，ルーティングテーブルにキャッシュを有するスイッチ（Swtich w/ Routing Cache）を用いることで NAS 並列ベンチマークの MOPS 値を向上できていることが分かる．64 台のスイッチを用いた場合，平均 5.2%，256 台のスイッチを用いた場合，平均 6.9%の向上を達成している．NAS 並列ベンチマークにおいて，プロセス数が増加することによりプロセスあたりの計算量が減る．そのため，プロセス数が増えた場合，すなわち，ネットワークサイズが大きい場合，通信遅延がプログラムの実行時間に与える影響が相対的に大きくなると考えられる．よって，ネットワークのサイズが大きいほど，ルーティングテーブルキャッシュを有するスイッチを用いた性能向上の効果が大きくなったと考えられる．

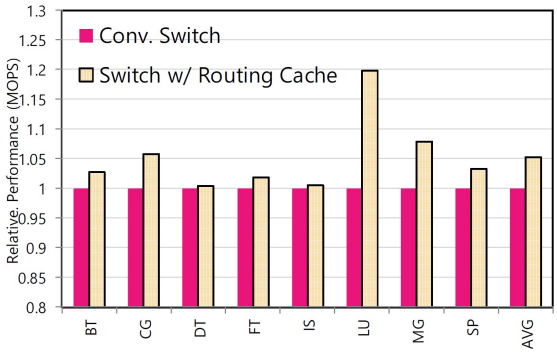

図 7　NAS 並列ベンチマークの評価結果（64 台のスイッチ，ケーススタディ 1）

Fig. 7　Evaluation results of NAS parallel benchmarks (64 switches, case study 1).

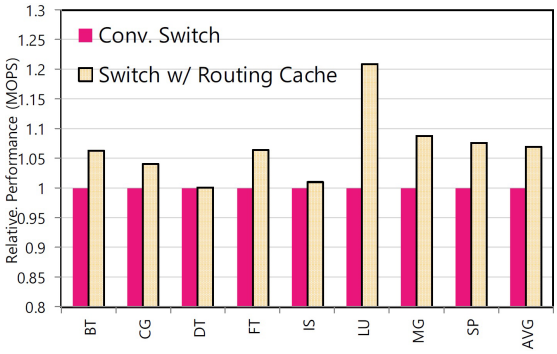

図 8　NAS 並列ベンチマークの評価結果（256 台のスイッチ，ケーススタディ 1）

Fig. 8　Evaluation results of NAS parallel benchmarks (256 switches, case study 1).

NAS 並列ベンチマークでは，IS や FT のように全対全あるいはネットワーク全体にわたるアクセスが発生する場合や，SP，BT，および LU のように隣接プロセス間通信が多く発生する場合など，様々な通信パターンが生じる [23]．本評価結果は，すべての並列ベンチマークの実行において，ルーティングテーブルキャッシュを有するスイッチを用いることで性能向上を達成したことが特筆される．つまり，ネットワーク遅延を削減する本キャッシュの効果はきわめて大きいといえる．

5.3　評価結果（ケーススタディ 2）

ケーススタディ 2 における NAS 並列ベンチマークの評価結果を**図 9** に示す．ルーティングテーブルキャッシュをスイッチに導入することで，平均 5.5%の性能向上を達成している．ケーススタディ 1 と比較して，計算ノードの性能が大幅に向上したため通信間隔が短くなる．つまり，通信性能の影響が与えるアプリケーション性能の影響が大きくなりやすいといえる．一方，ネットワーク構成の面では，ケーススタディ 1 の場合と比較して，(1) ベースライン

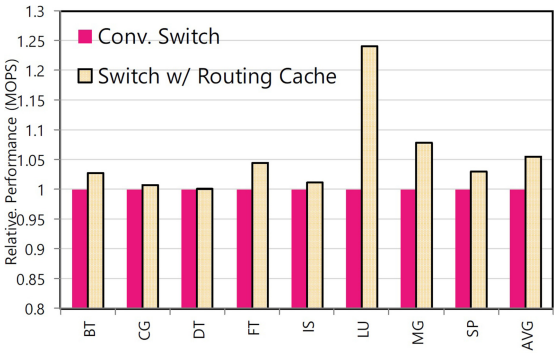

図 9　NAS 並列ベンチマークの評価結果（64 台のスイッチ，ケーススタディ 2）

Fig. 9　Evaluation results of NAS parallel benchmarks (64 switches, case study 2).

スイッチの遅延が 100 ナノ秒から 60 ナノ秒と小さくなり，かつ，(2) ネットワークトポロジがホップ数の小さいランダムトポロジとなったため，通信性能がアプリケーション性能に与える影響が小さい環境ともいえる．

両者のトレードオフの結果，全体のシミュレーション結果としては，ケーススタディ 1 の場合と比べてほぼ同等の平均性能向上を達成したと考えられる．

6. 無負荷通信遅延の解析

本章では，大きなジョブを実行したときに生成される通信の評価を解析する．前章の SimGrid シミュレーションは，我々の評価環境において計算時間の面で 256 台の計算ノードを超える規模の評価を行うことができない．そこで，我々は数千〜30K 計算ノード間の通信を対象にしたルーティングテーブルキャッシュの効果を無負荷通信遅延の解析を通して評価する．なお，先行研究 [24] で本章で行うような結合網の性能解析は行われており，結合網のシミュレーションの評価結果と一貫性のある結果がすでに報告されている．本解析評価は，システム開発の初期検討における性能指標として有益なものであると考えられる．

6.1 条件

前章のケーススタディ 1 ではネットワーク構成で 3 次元トーラスを用いたが，本章ではより一般化し，ジョブを K-ary N-cube トーラス（K^N 台の計算ノード，K は奇数）の矩形領域に割り当て，実行することとする．1 台のスイッチにつき 1 台の計算ノードが接続されており，次元順ルーティングを想定する．なお，1 台のスイッチにつき複数台の計算ノードが接続された場合，あるいは，K が偶数の場合も以下，同じ考え方で求めることができるが，簡単化のためここでは省略する．

キャッシュサイズは入力ポートごとに M エントリとする．

計算ノードは，独立にポアソン分布に従った間隔でパケットを生成することとする．そして，各計算ノードにおいて，パケットは生成順にネットワークに注入される．パケットの目的地は，パケット生成時に，すべての計算ノードのなかからランダムに 1 つ選択することで決定される．

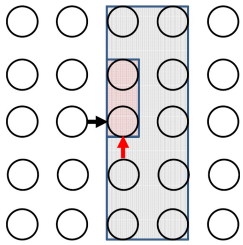

図 10　5-ary 2-cube トーラスの例

Fig. 10　Example of 5-ary 2-cube torus.

4 章と 5 章の評価環境と異なり，本章では大きなジョブを想定するため，容量性ミスが生じ，通信遅延に影響する．一方，ランダムトラヒックを想定し，十分な連想数を有することで競合性ミスは発生せず，かつ，ネットワーク始動後十分に時間が経過してキャッシュが十分に利用されている，すなわち，初期参照ミスは生じないこととする．

6.2 キャッシュヒット率

図 10 は，5-ary 2-cube トーラスにおいて，x 次元および y 次元入力ポートを通過するパケットの目的地となる頂点群を示している．x 次元の入力ポート（図中，黒矢印）のパケットは，次元順ルーティングに従って，黒点線の枠内の $K\lfloor K/2 \rfloor$ 個の目的地をとりうる．一方，y 次元の入力ポート（図中，赤矢印）のパケットは，赤斜線の枠内の $\lfloor K/2 \rfloor$ 個の目的地をとりうる．なお，ローカル計算ノードからの注入入力ポートのパケットは，自身を除く $K^N - 1$ 個の目的地をとりうる．

ランダムトラヒックの場合，K-ary N-cube トーラスにおける i 次元の入力ポートのキャッシュヒット率 P_i $(0 \leq P_i \leq 1)$ は入力ポートを通過するパケットの目的地の総数から，式 (1) により求まる．なお，ローカル計算ノードからの注入入力ポートは 0 次元とする．

$$P_i = \begin{cases} t_i & (t_i < 1) \\ 1 & (\text{otherwise}) \end{cases} \tag{1}$$

$$t_i = \begin{cases} \frac{M}{K^N - 1} & (i == 0) \\ \frac{M}{K^{N-i}\lfloor K/2 \rfloor} & (0 < i) \end{cases} \tag{2}$$

6.3 無負荷通信遅延

計算ノード〜スイッチ間リンク遅延，スイッチ間リンク遅延ともに C，ルーティングテーブルキャッシュがヒットした場合のスイッチ遅延を S_{base}，ミスした場合のペナルティ遅延を S_{pnl} とする．

あるスイッチの i 次元の入力チャネルにあるパケットが，

i 次元にそって，j ホップ離れた座標のスイッチの入力チャネルへ到着するまでの無負荷通信遅延 $L_{i,j}$ は，以下である．

$$L_{i,j} = j(S_{base} + S_{pnl}(1 - P_i) + C) \quad (3)$$

無負荷通信遅延の最大 L_{max} は，K-ary N-cube トーラスの各次元において $\lfloor K/2 \rfloor$ ホップ離れた目的地への無負荷通信遅延となる．キャッシュヒット率は，スイッチの入力ポートの次元 i に依存するため，L_{max} は以下となる．

$$L_{max} = L_{0,1} + \sum_{i=1}^{N} L_{i,\lfloor K/2 \rfloor} + C \quad (4)$$

式 (4) の右項において，0 次元であるローカル計算ノードからの注入を 1 ホップとする点，目的地の計算ノードまでに経由するリンク数は経由スイッチ数よりも 1 つ大きくなる点に注意が必要である．

6.4　最大無負荷通信遅延の評価

キャッシュサイズ，ジョブの実行サイズ，ベースラインとなるスイッチ遅延をパラメータとして変更した場合のジョブ内全対全のランダムトラヒックに対する最大無負荷通信遅延を**図 11**，**図 12** に示す．また，スイッチ遅延，キャッシュのミスペナルティなどの評価パラメータは前章ケーススタディ 1 と同じとする．キャッシュのエントリ数は 2,048 とする．

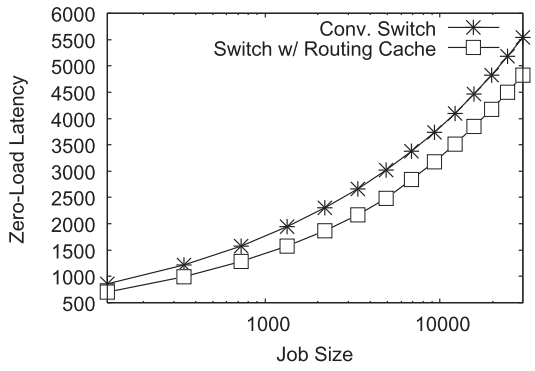

図 11　ジョブサイズに対する無負荷通信遅延

Fig. 11　Zero-Load communication latency for various job sizes.

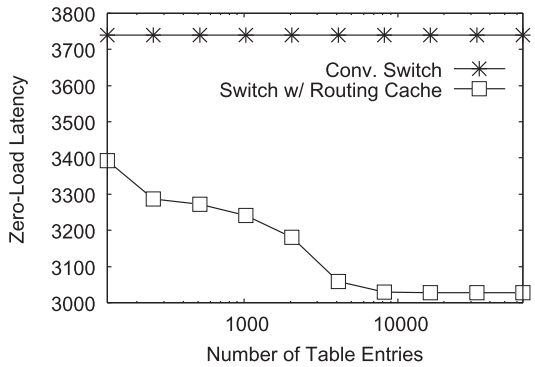

図 12　入力ポートあたりのキャッシュエントリ数に対する無負荷通信遅延

Fig. 12　Zero-Load communication latency vs. number of cache entries per input port.

6.4.1　ジョブサイズ

図 11 は，ジョブサイズを大きくした場合の最大無負荷通信遅延を示している．ジョブサイズが大きくなるに従って，容量性のキャッシュミスが生じる．しかし，ルーティングテーブルキャッシュによる最大無負荷通信遅延の削減効果は 13〜19%と依然，大きいことが分かる．よって，ルーティングテーブルキャッシュは，ジョブサイズが大きい場合でも通信遅延の削減に有効であるといえる．

6.4.2　キャッシュエントリ数

図 12 は，大規模な相互結合網，具体的には $21 \times 21 \times 21$ 3 次元トーラスにおいて，入力ポートあたりのルーティングテーブルキャッシュのエントリ数を大きくした場合の最大無負荷通信遅延を示している．比較のため，キャッシュを有しないベースラインとなるスイッチ（Conv. Switch）の最大無負荷通信遅延をプロットしている．キャッシュのエントリ数が 128 の場合において，最大無負荷通信遅延の 9%削減に成功している．本構成では，9,261 ($21 \times 21 \times 21$) 台の計算ノードで構成されているため，9,261 以上のキャッシュのエントリ数を有する場合は，最大無負荷通信遅延の削減効果は 19%で一定となる．

7.　ルーティングテーブルキャッシュ導入による他の側面の影響

7.1　消費電力

ルーティングテーブルキャッシュを有するスイッチでは，キャッシュヒット時に CAM アクセスが省略されるため，CAM の読み出しにかかる動的エネルギーが削減される．特に，ルーティングテーブルキャッシュは前述したように 100%に近いヒット率を獲得できることから，CAM の動的エネルギーの大幅な削減が期待できる．本節では，ルーティングテーブルキャッシュを有するスイッチによるテーブル検索の消費電力削減効果を，モデル式を用いて評価する．

本スイッチのテーブル検索消費電力 S は，論文 [17] で検討された単一のルーティングテーブルキャッシュを備えるインターネットルータのテーブル検索消費電力モデル式をもとに，以下で求められる．

$$S = n(E_{cache} + E_{cam} r_{miss}) + (dP_{cache} + P_{cam}) \quad (5)$$

ここで E_{cache} と E_{cam}，P_{cache} と P_{cam} はそれぞれキャッシュおよび CAM の動的エネルギー，静的電力を表す．n，d，r_{miss} はそれぞれ毎秒処理パケット数，スイッチ次数，キャッシュミス率を表す．

なお，従来スイッチのテーブル検索消費電力は式 (5) の E_{cache} および P_{cache} を 0 とし，r_{miss} を 1 とすることで求

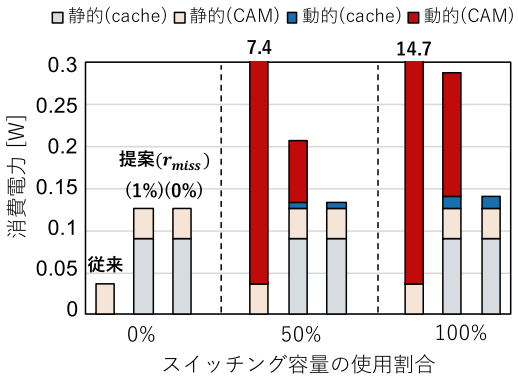

図 13　スイッチング容量の使用割合に対するテーブル検索消費電力と内訳

Fig. 13　Power consumption of table lookups in three cases of switching capacity utilization.

められる．

E_{cache} および P_{cache} は CACTI6.5 [18] を用いて見積もるとキャッシュメモリあたり 0.039 nJ，13 mW となる．一方で，E_{cam} および P_{cam} は近年の CAM の電力に関する論文 [25], [26] を参考に 42 nJ，36 mW とした．d は 3 次元トーラスを想定して，7 ポート（ローカルポート込）とした．

以上のパラメータを用いて各スイッチのテーブル検索消費電力 S を求めた結果を**図 13** に示す．なお，n は流動的であるため，図 13 ではスイッチング容量 1,400 Gbps（200 Gbps × 7）に対して，スイッチング容量の 0%，50%（700 Gbps），100%（1,400 Gbps）が使用される 3 ケースについて n を求めている．また，r_{miss} は 5.2 節の結果をもとにキャッシュミス率 1%と 0%の 2 ケースを想定し，テーブル検索消費電力を求めた．

図 13 より，従来スイッチでは処理パケット数に応じて増加する CAM の動的エネルギーが支配的となる．一方，ルーティングテーブルキャッシュを有するスイッチでは CAM アクセスを減らすことで消費電力の大幅な削減が可能となることが分かる．ポートごとにキャッシュを用意するスイッチではキャッシュの静的電力が増えるが，スイッチング容量の 50%使用の場合においては従来スイッチの 1%程度に過ぎず，CAM の動的エネルギー削減に対する寄与のほうが大きい．

以上より，ルーティングテーブルを導入することで，スイッチの消費電力を削減することが可能である．

7.2　スイッチング容量

従来スイッチにおけるルーティング処理の最大スループット T_{base} は式 (6) で求まる．

$$T_{base} = \frac{1}{L_{cam}} \tag{6}$$

ここで L_{cam} は CAM のアクセス遅延を表す．

1 パケットあたり 1 回のルーティング処理が必要となる．2.1 節で言及したようにオフチップ CAM へのアクセスには 25 ナノ秒程度を要するため，単一の CAM を有する従来スイッチのルーティング処理のスループット T_{base} は 40 Mpps（packet per second）程度が限界となる．

一方，ルーティングテーブルキャッシュを有するスイッチでは，CAM アクセスはキャッシュミス時にのみ生じるため，ルーティング処理の最大スループット T_{rcache} は式 (7) で求まる．

$$T_{rcache} = \min\left(\frac{1}{L_{cache}}, \frac{1}{L_{cam} r_{miss}}\right) \tag{7}$$

ここで L_{cache} はキャッシュのアクセス遅延を表す．5.1 節で評価したように，キャッシュミス時のルーティング処理遅延はハッシュ計算とキャッシュアクセス遅延，CAM アクセス遅延の和となる．したがって，キャッシュミス時には従来スイッチと比べてルーティング遅延が悪化する．一方で，スループットは従来スイッチと同等あるいは高くなる．これは，各処理がパイプラインで動作しており，全パケットがキャッシュミスするワーストケースであっても，CAM は 40 Mpps のスループットを達成できるためである．

本論文で想定する 24 KiB キャッシュのアクセス遅延は 3.1 節で言及したように 1 ナノ秒程度であり，同スループットはポートあたり最大で 1 Gpps となる．よって，本ルーティングテーブルキャッシュを有するスイッチの T_{rcache} は式 (7) より，100%に近いキャッシュヒット率が得られる場合 1 Gpps となり，従来スイッチの T_{base} と比べて 25 倍高くなる．キャッシュミス率が 4%以上になると，CAM アクセスがボトルネックとなりスループットは低下していく．しかしながら，たとえば 10%のミス率であっても T_{rcache} は 400 Mpps であり，従来スイッチの T_{base} と比べ 10 倍のスループットが得られる．

以上より，ルーティングテーブルキャッシュを導入することで，ルーティング処理の大幅なスループット向上が可能である．

8.　まとめと今後の課題

HPC（High-Performance Computing）システムにおける並列アプリケーションの性能は計算ノード間の相互結合網の通信遅延により影響を受ける．そこで，本研究では，スイッチ遅延を削減するために，ルーティングテーブルキャッシュを各入力ポートに有するスイッチ構成を探求し，本キャッシュが与えるシステム性能への影響を評価した．

まず，キャッシュのシミュレーション結果より，4 ウェイの連想数で 2,048 エントリのキャッシュを有するスイッチで構成された相互結合網において，512 台の計算ノードを用いて実行した場合の通信に対する競合性ミスの発生は 0.1%以下となることが分かった．次に，SimGrid v3.12 を用いてアプリケーション性能を評価した結果，256 台のスイッチを用いた相互結合網において NAS 並列ベンチマー

クの性能を平均 6.9%向上させることが分かった．最後に，大規模な相互結合網におけるルーティングテーブルキャッシュが与える無負荷通信遅延への影響を解析した．解析結果より，ジョブサイズが十分に大きい場合（例：9k 台の計算ノードの利用），本キャッシュは，容量性ミスが生じるため，最大無負荷通信遅延の削減効果が 19%から 13%へと緩やかに低下することが分かった．一方，本キャッシュのエントリ数が大きいほど，同遅延の削減効果が 9%から 19%へと緩やかに大きくなることが分かった．

以上より，相互結合網の通信遅延の削減，かつ，並列ベンチマークの性能向上への貢献が明瞭であることから，ルーティングテーブルキャッシュを HPC スイッチに導入することを，強く推奨する．

今後の課題としては，異なるスイッチマイクロアーキテクチャにおけるルーテイングテーブルキャッシュ導入の有効性に関する検討があげられる．具体的には，小規模システムを対象としたスイッチでは，ルーティングテーブルがスイッチチップに内蔵されることが想定される．このようなスイッチではルーティングテーブルへのアクセス時間を小さくすることが期待できる．この場合におけるルーティングテーブルキャッシュの有効性に関する定量的な評価を進める予定である．

謝辞　本研究は JST さきがけ JPMJPR19M1，JSPS 科研費 19H01106 の助成を受けたものである．図 1 の作成に協力いただいた国立情報学研究所アーキテクチャ科学研究系胡曜博士に感謝する．

参考文献

[1] Hemmert, K.S. et al.: Report on Institute for Advanced Architectures and Algorithms, *Interconnection Networks Workshop 2008*, available from ⟨http://ft.ornl.gov/pubs-archive/iaa-ic-2008-workshop-report-final.pdf⟩.

[2] RENESAS Electronics Corp.: Network Search Engine, available from ⟨https://www.renesas.com/jp/ja/products/memory/application-specific-memory/network-search-engine.html⟩.

[3] Koibuchi, M., Ishida, S. and Nishi, H.: The impact of routing cache on high-performance switches, *Proc. 10th International Conference on Optical Internet* (*COIN*), Tuj2 (2 pages) (May 2002).

[4] Ishida, S., Koibuchi, M. and Nishi, H.: A case for routing cache on HPC switches, *IEICE Communications Express*, Vol.1, No.1, pp.49–53 (2012).

[5] Klusacek, D. and Chlumsky, V.: Evaluating the Impact of Soft Walltimes on Job Scheduling Performance, *Job Scheduling Strategies for Parallel Processing*, pp.15–38 (May 2018).

[6] Molka, D. Hackenberg, D., Schone, R. and Nagel, W.E.: Cache coherence protocol and memory performance of the intel haswell-EP architecture, *Proc. 44th International Conference on Parallel Processing* (*ICPP 2015*), pp.739–748 (Sep. 2015).

[7] Agrawal, B. and Sherwood, T.: Ternary CAM power and delay model: Extentions and uses, *2008 IEEE Trans. Very Large Scale Integration* (*VLSI*) *Systems*, Vol.15, No.5, pp.554–564 (2008).

[8] Matsunaga, S. et al.: Fabrication of a 99%-energy-less nonvolatile multi-functional CAM chip using hierarchical power gating for a massively-parallel full-text-search engine, *Symposium on VLSI Circuits*, pp.C106–C107 (2013).

[9] Soyata, T. and Liobe, J.: pbcam: Probabilistically-banked content addressable memory, *2012 IEEE International SOC Conference*, pp.27–32 (2012).

[10] Dally, W.J. and Towles, B.: *Principles and Practices of Interconnection Networks*, Morgan Kaufmann (2003).

[11] Chen, D. et al.: The IBM Blue Gene/Q Interconnection Network and Message Unit, *SC*, pp.1–10 (Nov. 2011).

[12] Scott, S., Abts, D., Kim, J. and Dally, W.J.: The blackwidow high-radix clos network, *International Symposium on Computer Architecture* (*ISCA*), pp.16–28 (2006).

[13] 堀江健志，清水　剛，服部　彰：1 チップ 10 ギガイーサネットスイッチ LSI, FUJITSU, Vol.55, No.6, pp.553–558 (Nov. 2004).

[14] Nishimura, S., Kudoh, T., Nishi, H., Yamamoto, J., Harasawa, K., Matsudaira, N., Akutsu, S., Tasho, K. and Amano, H.: RHiNET-3/SW: An 80-Gbit/s high-speed network switch for distributed parallel computing, *Hot Interconnects*, Vol.9, pp.119–123 (2001).

[15] 奥野通貴，西村信治，石田慎一，西　宏章：ネットワークトラフィックの時間的局所性を利用したブロードバンドネットワーク向けキャッシュ型パケット処理技術，情報処理学会論文誌，Vol.47, No.2, pp.346–354 (2006).

[16] Yang, T., Xie, G., Li, Y., Fu, Q., Liu, A.X., Li, Q. and Mathy, L.: Guarantee IP Lookup Performance with FIB Explosion, *Proc. 2014 ACM Conference on SIGCOMM*, pp.39–50 (2014).

[17] Tanaka, K., Yamaki, H., Miwa, S. and Honda, H.: Multi-level packet processing caches, *2019 IEEE Symposium in Low-Power and High-Speed Chips* (*COOL CHIPS*), pp.1–3 (Apr. 2019).

[18] Muralimanohar, N., Balasubramonian, R. and Jouppi, N.P.: Cacti 6.0: A tool to model large caches, *HP laboratories*, Vol.27, p.28 (2009).

[19] 吉田利雄, 池田吉朗, 安島雄一郎ほか：スーパーコンピュータ「京」：3. ハードウェア—ラック，冷却，プロセッサ，インターコネクト，情報処理，Vol.53, No.8, pp.767–773 (2012).

[20] Yamaki, H., Nishi, H., Miwa, S. and Honda, H.: Data prediction for response flows in packet processing cache, *Proc. 55th Annual Design Automation Conference* (*DAC 2018*), No.110, pp.1–6 (2018).

[21] SimGrid: Simulation of Distributed Computer Systems, available from ⟨https://simgrid.org/⟩.

[22] HPCI 技術ロードマップ白書，入手先 ⟨http://www.open-supercomputer.org/workshop/sdhpc/⟩.

[23] Chaix, F., Fujiwara, I. and Koibuchi, M.: Suitability of the random topology for HPC applications, *Euromicro International Conference on Parallel, Distributed, and Network-Based Processing* (*PDP*), pp.301–304 (2016).

[24] Ould-Khaoua, M.: A performance model for Duato's fully adaptive routing algorithm in k-ary n-cubes, *IEEE Trans. Computers*, Vol.48, No.12, pp.1297–1304 (1999).

[25] Mahendra, T.V., Mishra, S. and Dandapat, A.: Self-controlled high-performance precharge-free content-addressable memory, *IEEE Trans. Very Large Scale Integration* (*VLSI*) *Systems*, Vol.25, No.8, pp.2388–2392 (2017).

[26] Mishra, S. Mahendra, T.V. and Dandapat, A.: A 9-T 833-MHz 1.72-fj/bit/search quasi-static ternary fully associative cache tag with selective matchline evaluation for wire speed applications, *IEEE Trans. Circuits and Systems I: Regular Papers*, Vol.63, No.11, pp.1910–1920 (2016).

平澤 将一　（正会員）

平成 12 年東京大学理学部情報科学科卒業．平成 17 年同大学大学院情報理工学系研究科コンピュータ科学専攻博士課程単位取得退学．平成 18 年電気通信大学大学院情報システム学研究科助手．平成 24 年東北大学大学院情報科学研究科産学官連携研究員．平成 29 年より国立情報学研究所特任研究員．高性能計算システムとプログラミング言語処理に関する研究に従事．

八巻 隼人　（正会員）

平成 28 年慶應義塾大学大学院理工学研究科開放環境科学専攻博士課程修了．博士（工学）．現在，電気通信大学大学院情報理工学研究科助教．インターネットルータ等のハードウェアアーキテクチャに関する研究に従事．

鯉渕 道紘　（正会員）

平成 15 年慶應義塾大学理工学研究科博士課程修了．博士（工学）．平成 17 年国立情報学研究所助手．現在，国立情報学研究所准教授，総合研究大学院大学複合科学研究科情報学専攻准教授（兼任）．相互結合網と計算機システムに関する研究に従事．