社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

河野　隆太[†,††]　藤原　一毅[†††,††]　松谷　宏紀[†,†††]　天野　英晴[†,†††]　鯉渕　道紘[†††,††]

† 慶應義塾大学大学院 理工学研究科　223–8522 神奈川県横浜市港北区日吉 3–14–1
†† 科学技術振興機構
††† 国立情報学研究所　101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †{kawano,matutani,hunga}@am.ics.keio.ac.jp, ††{ikki,koibuchi}@nii.ac.jp

あらまし　我々は，1 台のハイパフォーマンスコンピューティング (HPC) システムにおいて，複数の小規模並列アプリケーションを効率良く動作させることを目指している．古典的な並列アプリケーションは様々なトポロジを想定して最適化されてきたため，想定する論理トポロジと実際の物理トポロジが異なる場合は，性能を十分に発揮できない．そこで，我々は，これまでに電気スイッチネットワークに対し，光サーキット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てきた．本報告では元になる電気スイッチネットワークのトポロジ，および，トポロジの内包方法について詳細な評価を行う．評価結果より，提案トポロジは従来のトポロジに比べ，より低い導入コストで，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を示した．
キーワード　高性能コンピューティング, 光サーキットスイッチ, ネットワークトポロジ，相互結合網, データセンターネットワーク

Interconnect Design for Low Latency, High Topological Embeddability and Partitioning Capability by Supplementary Optical Circuit Switches

Ryuta KAWANO[†,††], Ikki FUJIWARA[†††,††], Hiroki MATSUTANI[†,†††], Hideharu AMANO[†,†††], and Michihiro KOIBUCHI[†††,††]

† Graduate School of Science and Technology, Keio University　Hiyoshi 3–14–1, Kohoku-ku, Yokohama, Kanagawa, 223–8522 Japan
†† Japan Science and Technology Agency
††† National Institute of Informatics　Hitotsubashi 2–1–2, Chiyoda-ku, Tokyo, 101–8430 Japan
E-mail: †{kawano,matutani,hunga}@am.ics.keio.ac.jp, ††{ikki,koibuchi}@nii.ac.jp

Abstract　This paper focuses on how to efficiently run multiple small parallel applications in a single High-performance computing (HPC) system. As parallel applications can be optimized for several specific topologies, some of them cannot show full performance when a given physical topology is different from the logical topologies they assumed. Previously we proposed to supplementally use optical circuit switches (OCSes) to patch the electrically-switched network that achieves low latency and supports high topology partitionability, so that the electrically-switched topology efficiently embed or emulate k-ary n-cubes, fat trees and random connections. In this report we evaluate the baseline electrically-switched network topology and the embedded method in detail. Our emperical results show that the proposed topology achieves high topology embeddability, the overall low latency and a low-cost implementation compared to the conventional topologies.
Key words　High-performance computing (HPC), optical circuit switching, network topology, interconnection networks, datacenter networks

1. はじめに

通常，データセンターネットワークや，並列計算機などの大規模計算システムでは，システム導入時に電気パケットスイッチ間のネットワークトポロジが決定されている．しかし，並列アプリケーション毎にアプリケーション内で生じる通信アクセスパターンは異なる．さらに，システム導入時に想定していない通信アクセスパターンを持つ並列アプリケーションが将来的に登場する可能性もある．そのため，並列アプリケーションのプロセス間通信パターンを示す論理トポロジと，システム導入時に決定された物理ネットワークトポロジとの乖離を抑えることが，アプリケーション性能向上のための 1 つの課題となる．例えば，並列数値シミュレーションを 2048 コア規模のスーパーコンピュータで実行させる場合，アプリケーションの問題空

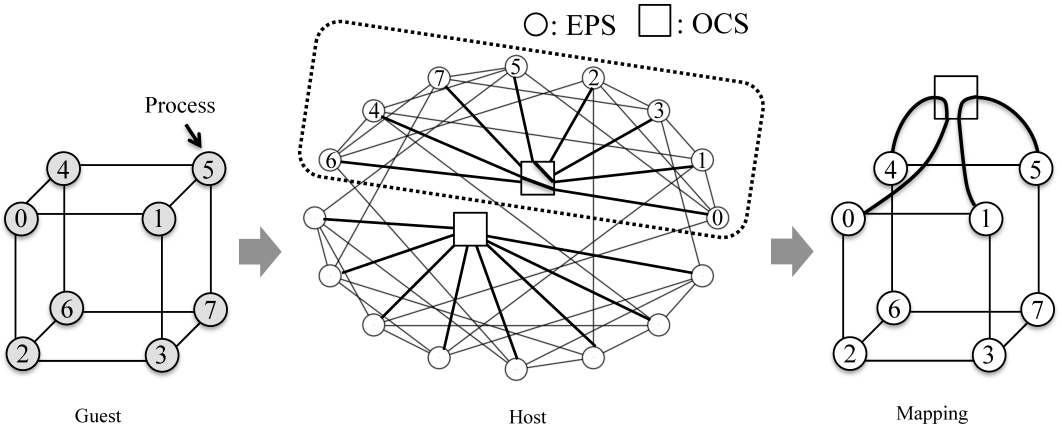


図 1　光サーキットスイッチを用いた論理トポロジのマッピング例

間 (論理トポロジ) の実トポロジへのマッピング最適化により，15%の性能向上が達成されることが報告されている [1].

現状では，TORUS，FAT TREE などのネットワークトポロジの中から，直径，スイッチの次数，ルーティングの容易性，耐故障性，レイアウトとコストなどの点でトレードオフを考慮した上で HPC システム毎に設計者の総合的な判断により（異なる）トポロジが選択されている．例として，京コンピュータでは 6 次元 TORUS が採用されているほか，TSUBAME 2.0 では FAT TREE が用いられている．したがって，現状ではシステムが採用したトポロジ毎にユーザが並列アプリケーションの最適化を行うことが必要となる．

そこで，このユーザの最適化の負担を軽減するために，我々は，電気パケットスイッチ間トポロジに光サーキットスイッチを付加することにより，様々な (電気パケットスイッチ間の) トポロジを内包可能な相互結合網 (図 1 中央) の提案を行ってきた [2]．本相互結合網は，光サーキットスイッチの構成を並列アプリケーション毎に変更することにより，各アプリケーションに適したトポロジを提供することが可能である．光サーキットスイッチは，通常，10〜100 ミリ秒のサーキット構成オーバヘッドがかかるため，本提案ではサーキットの構成は対象とするアプリケーション実行前に 1 度のみ実行することを前提とする．また，各光サーキットスイッチは，波長分割や時分割による多重化を行わない，1 リンク 1 チャネルの単純なサーキットスイッチとして利用することとしている．

本相互結合網の具体的な目的及び手法を説明する．本提案では，アプリケーション実行前に，光サーキットスイッチのサーキットを確立し，そのコネクションを 1 つのリンクとみなす．そして，そのリンクの endpoint を切り換えることで，様々な (電気スイッチ間) トポロジをエミュレートする．例を図 1 に示す．この図において，内包させたい論理トポロジを “Guest” とし，提案相互結合網の物理トポロジを “Host” とし，論理トポロジの物理トポロジへのマッピング結果を “Mapping” としている．本図左の Guest トポロジはアプリケーションのプロセス番号及びプロセス間通信を表す．また，Host トポロジ，Mapping の図において，白丸は電気パケットスイッチ，四角は光サーキットスイッチ，リンクは物理リンクを表し，番号はマッピングされた論理トポロジの各プロセスを示す．本図に示すように，Host トポロジの中から Guest トポロジをマッピング可能な箇所 (点線で囲んだ箇所) を見つけ，電気パケットスイッチ間の実リンクと，光サーキットスイッチにより確立された電気パケットスイッチ間リンクを組み合わせて，Guset トポロジの完全なマッピングを実現している．

本報告は，我々の以前の提案トポロジ [2] の有用性を詳細に検証するための細部の設計と評価を行う．主な貢献は以下の通りである．

- 本相互結合網のより詳細な Guest トポロジの探索方法を提案する．以前の提案 [2] では，Guest トポロジが直接網である場合に限定した探索手法の提案及び評価を行ったが，本研究では，Guest トポロジが間接網である場合に適用可能なトポロジ探索手法を提案する．グラフ解析の結果から，この相互結合網において，遺伝的アルゴリズムを用いた探索手法により，一般的な並列アプリケーションに広く使われる FAT TREE を Guest トポロジとした際のトポロジ内包性が大幅に向上することが分かった．
- 本相互結合網の実用化に向けた，導入コストの評価を行う．クラスタリング及びキャビネット配置の最適化手法を用いることにより，従来の規則的な物理トポロジと同程度の総配線延長・消費電力で本提案物理トポロジを実装可能であることが分かった．

以下，2. 章において関連研究を述べ，3. 章において電気パケットスイッチ間に適用するトポロジ，及び光サーキットスイッチの挿入手法について述べる．4. 章ではトポロジの探索手法について述べ，FAT TREE トポロジの内包性の評価を行う．5. 章では本提案に関するコスト評価を行い，最後に 6. 章で結論を述べる．

2. 関 連 研 究

2.1 典型的なネットワークトポロジ

現在，HPC 向け相互結合網として様々なものが提案されており，それらはネットワークのスループットを向上させる，あるいは，経由電気スイッチ数を小さくすることに主眼がおかれている．さらに，高次元のスイッチを用いたネットワークをキャビネット内とキャビネット外の 2 階層ネットワークに分け，階層ごとに多様なトポロジを埋め込むことのできる Dragonfly [3] といったトポロジも提案されている．

近年のスーパーコンピュータでは，TORUS や FAT TREE などの規則的なトポロジが用いられる場合が多い．これらのトポロジはマルチユーザ環境において比較的小規模なノード数を利用するアプリケーションを実行する際に，トポロジを分割して部分的に利用出来る点でも優れたトポロジである．

また，HPC 向けの低遅延なトポロジとして，ノード間を不規則に接続するランダムトポロジが提案されている．このト

ポロジはスモールワールド性と呼ばれるノード間ホップ数が $\log N$(N: ノード数) に比例して小さくなる性質を持ち．電気パケットスイッチ間トポロジに適用した場合，帯域や拡張性，耐故障性などに優れることから，データセンター向けに活用する提案 [4]〜[6] が報告されている．

2.2 Topology Embedding

並列アプリケーションの実行性能向上のために，プロセス間通信の論理トポロジを実ネットワークトポロジの一部に効率的にマッピングすることが必要となる．一般的なグラフ理論において，大きなホストトポロジ H の頂点及び辺の一部に小さなゲストトポロジ G の頂点及び辺を効率的にマッピングする問題は Topology Embedding と呼ばれる問題である．

Topology Embedding において，マッピングの効率を表す指標がいくつか存在し，その中に *dilation* と *congestion* がある．トポロジ G 内の各頂点をトポロジ H の各頂点の一部にマッピングし，さらにトポロジ G 内の各辺について，マッピング先の 2 頂点間の複数経路のうち 1 つにマッピング先を定めることとする．このとき，トポロジ G 内のある辺の dilaton とは，その辺のマッピング先として定めたトポロジ H 内の経路の距離 (ホップ数) を表す．トポロジ G 内のある辺で dilaton が 1 を超えているとき，ゲストトポロジ G のマッピング先での辺の長さが 1 より長くなっていることを示す．また，ホストトポロジ H 内のある辺での congestion とは，トポロジ G の全ての辺のマッピング先経路のうち，その辺を通る本数を表す．トポロジ H 内のある辺で congestion が 1 を超えるとき，ゲストトポロジ G の複数の辺がマッピング先で辺を共有していることを示す．

並列アプリケーションのマッピングでは，マッピング先の遅延や帯域の悪化を防ぐため，dilaton 及び congestion の最大値を共に 1 とするような完全なマッピングを行うことが理想である．TORUS や HYPERCUBE といった規則的トポロジをホスト及びゲストトポロジとした Topology Embedding 問題については，dilaton 及び congestion の最大値を小さくするようなマッピング最適化を行う提案がなされている [7] が，dilation 及び congestion の最大値を 1 とするような完全なマッピングは困難とされており，さらに一般的に完全なマッピングの探索は NP 困難であるとされている．

本研究では TORUS および FAT TREE トポロジをゲストトポロジとし，dilation 及び congestion の最大値を共に 1 とするような完全マッピングを目的としたホストトポロジ及び探索手法を探求する．

2.3 光サーキットスイッチと電気パケットスイッチのハイブリッドネットワーク

データセンターや並列計算機の相互結合網では，光サーキットスイッチと電気パケットスイッチの両方を持つハイブリッド型が提案されてきた [8]．電気スイッチのみ，光サーキットスイッチのみで構成された 2 系統のネットワークを持つ典型的なハイブリッドネットワーク [8] では，大規模なバルク転送は光サーキットネットワーク，小さいデータ転送は電気パケットネットワークで行う．また，この方針は滝澤らの研究 [9] でも踏襲されている．

一方，ToR(Top-of-Rack) スイッチと (同じラック内の) ノード間は電気パケットスイッチを用いて通信を行い，ラック間 (ToR スイッチ) 通信は光サーキットスイッチで構成するデータセンターネットワークが提案されている [10]．本ネットワークでは最新の光サーキットスイッチのサーキット切り替えが 11.5ns で実現可能なことから，サーキットの再構成を数十〜数百 ns 秒単位で行うことを想定している．これらの研究では，光サーキットスイッチの長所である高バンド幅通信を生かすことに注力されている．一方，我々は光サーキットスイッチのサーキットをリンクとみなし，その “リンク” の endpoint (電気スイッチのポート) を更新する．つまりサーキットの再構成により，トポロジの内包性を向上させることに尽力する．

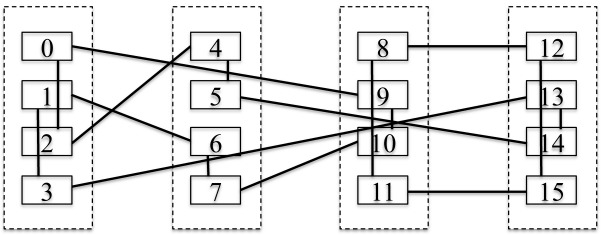


図 2　局所化ランダムトポロジの構成例 ($n = 2$, $N = 4$, $d = 1$)

なお，1 つのインターネット回線の帯域の一部を GMPLS 技術により，150Mbps 単位で end-to-end の専用パスとして利用する技術 [11] などがインターネットワーキングでは使われているが，現時点では本研究が対象とする 40Gbps 以上の光パスを多数リンクに集約したネットワーク技術を安価に HPC で利用することは難しい．そのため，本研究ではこのような WDM を用いた光通信技術の利用は想定しない．

3. 局所化ランダムトポロジ L-RANDOM

本章では，我々がこれまでに提案した局所化ランダムトポロジ (L-RANDOM) [2] 及び，L-RANDOM トポロジにおける光サーキットスイッチの挿入手法について記述する．

3.1 局所化ランダムトポロジ (L-RANDOM) の構成方法

本トポロジでは，TORUS や HYPERCUBE と同様に，次元数を n 次元と定義する．また，$i = 1, ..., n$ の下で，N_i を各次元の頂点数，d_i を各次元の頂点の次数と定義する．さらに，頂点数を $\prod_{i=1}^{n} N_i$，各頂点の次数 (各頂点からの辺の数) を $\sum_{i=1}^{n} d_i$ とする．また，各頂点に 0〜$(\prod_{i=1}^{n} N_i) - 1$ まで番号をつけることとする．

トポロジの構成方法は以下の通りである．各頂点の次数を増やすため，次に示す動作を n 回繰り返すこととし，以下を $j(= 1, ..., n)$ 回目の試行とする．

（1） 頂点群を番号順に $\prod_{i=1}^{j} N_i$ のサイズで均等に分割する．

（2） 分割後の各ユニット内で，頂点の次数を d_j だけ増やすようランダムマッチングを行う．ただし，$j > 1$ である時，$j - 1$ 回目の試行で同じユニットに属していた 2 頂点間を接続箇所として選ばないこととする．

トポロジの構成例を図 2 に示す．この構成例において，次元数は $n = 2$ であり，頂点数及び頂点の次数は全ての次元で等しく，$N = 4$，$d = 1$ としている．また，実線で囲んだ四角は各頂点を表し，各頂点内の数字は頂点番号を示している．この図において，点線で囲んだ各領域内で 1 次元目の頂点間接続を行い，点線の領域外で 2 次元目の頂点間接続を行っている．

頂点の次数が全ての次元で等しく $d = 2$ であるとき，本トポロジは n 次元 TORUS の各辺についてランダム入れ替えを行った方式と言える．また，頂点数及び次数が全ての次元で等しく $N = 2$, $d = 1$ であるとき，本トポロジは HYPERCUBE の各辺についてランダム入れ替えを行った方式と言える．

本トポロジを電気パケットスイッチ間に適用した場合，従来の規則的トポロジ及び従来のランダムトポロジ [6], [12] に比べ，

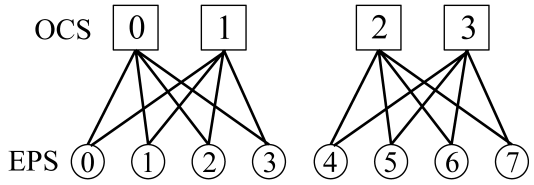


図 3　光サーキットの Regular 挿入 (EPS 数: 8, $p = 4, c = 2$)

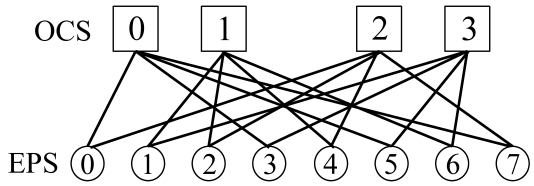


図 4　光サーキットの Random 挿入 (EPS 数: 8, $p = 4, c = 2$)

低遅延性能とトポロジ分割性能の両立性が大幅に向上することが分かっている [2].

3.2　光サーキットスイッチの挿入方法

本章では，トポロジの Embedding 性能 (内包性) を向上させることを目的とした，電気パケットスイッチ間トポロジに対する光サーキットスイッチの挿入方法について記述する．

本研究では，次の通りの仮定の下で，光サーキットスイッチの挿入を行う．光サーキットスイッチのポート数は全て等しく $p = 32$ とする．電気パケットスイッチ数は 1024 個とし，電気パケットスイッチ間トポロジの次数は 10 とする．各電気パケットスイッチは光サーキットスイッチとの接続に c 個のポートを用いる．

光サーキットスイッチの挿入手法として，以下の 2 通りを検討する．

- Regular 挿入 (REG): 図 3 のように，トポロジ上の各ノード (電気パケットスイッチ) の番号順に，ツリー状に光サーキットスイッチへ接続する．
- Random 挿入 (RND): 図 4 のように，各電気パケットスイッチに対し，接続先の光サーキットスイッチをランダムに c 個選択する．

4. 光サーキットスイッチの挿入による Embedding

本章では，GA(遺伝的アルゴリズム) を利用した手法により，電気パケットスイッチと光サーキットスイッチを混在させたトポロジにおいて，間接網である Embedding 対象トポロジの完全なマッピングを探索する手法を提案する．さらに，その手法を適用した FAT TREE トポロジの内包性の評価を行う．

4.1　FAT TREE の Embedding

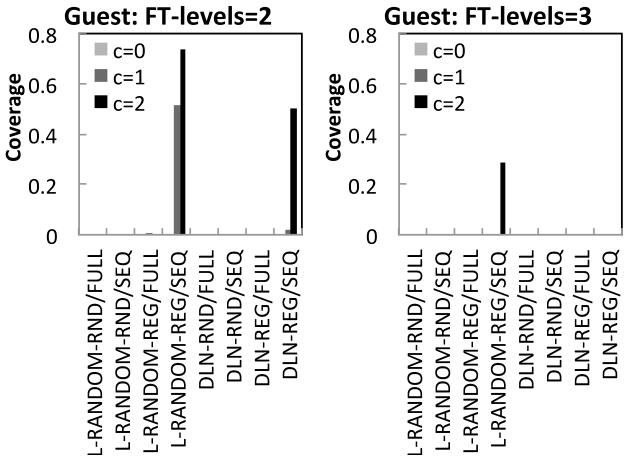
本章では，間接網の Guest トポロジのマッピング手法について記述する．本研究では FAT TREE をマッピング対象のトポロジとした．

FAT TREE の完全なマッピング (dilation 及び congestion の最大値を 1 とするマッピング) を行うため，以下の手順を用いる．

（1） congestion の最大値を 1 に保ちつつ，dilation が 1 を超える辺が少なくなるようなマッピングを探索する．

（2） 光サーキットスイッチによるサーキットを用いて，dilation が 1 を超える辺を補完し，全ての辺で dilation が 1 となるようにする．

2.2 章で述べた通り，dilation が 1 を超える辺の数が最小となるような FAT TREE の探索は NP 困難である．一方で，Topology Embedding の問題には GA(Genetic Algorithm; 遺伝的アルゴリズム) が有効であることが示されている [13] ため，GA をマッピング手順 (1) の探索手法として採用した．我々は他のメタヒューリスティックな探索手法であるグローバル/ローカルなランダムサーチや SA(Simulated Annealing; 焼きなまし法) も適用したが，GA に比べ結果が劣った．これは，本章で扱う探索問題の解となる完全なマッピングは，取りうる全マッピング数に対し少ない傾向にあるためである．



(a) G: FAT TREE (level = 2)　(b) G: FAT TREE (level = 3)

図 5　FAT TREE の内包率

本研究では，マッピング対象のゲストトポロジである FAT TREE について，以下の通り想定する．ある構成の FAT TREE トポロジを FAT TREE(l, x, y) と定義する．この定義において，l はツリーの階数を表す．x, y はそれぞれ各スイッチからの上位，及び下位方向への接続リンク数を表す．すなわち，FAT TREE(l, x, y) は $2^{l-2}y$ 個の “末端” スイッチと，$(2^{l-1} - 1)x$ 個の “中間” スイッチからなる．末端スイッチにはゲストトポロジのマッピング先となる計算ノードが直接接続されているものと仮定し，中間スイッチは FAT TREE 内の上位階層に位置するハブスイッチとしての役割を果たす．

GA の各個体は探索対象トポロジである FAT TREE のスイッチ数と同じ長さのベクトルから構成される，各個体の i $(0 \leqq i < 2^{l-2}y)$ 番目の要素の値は FAT TREE の i 番目の末端スイッチがマッピングされるホストトポロジ内のスイッチ番号であり，j $(j \geqq 2^{l-2}y)$ 番目の要素の値は FAT TREE の $(j - 2^{l-2}y)$ 番目の中間スイッチがマッピングされるホストトポロジ内のスイッチ番号である．各個体の評価値は dilation が 1 を超える辺の数であり，評価値を小さくするよう探索を行う．交叉操作では，2 個体をランダムな 1 点で連結し，重複する値の入った要素について，別のランダムな値に置き換える．突然変異操作は，個体内の 2 要素の値を入れ替える方法と，個体内の 1 要素の値を別のランダムな値に置き換える方法の 2 手法を用意する．選択操作はトーナメント法を適用する．個体数を 100，交叉確率を 10%, 突然変異確率を 2 手法に対しそれぞれ 20%とし，トーナメントサイズを 3 とする．世代数は最大で 20000 とし，5000 世代の間で最も良い個体の評価値が一定の場合は GA を終了する．

GA 実行後，得られた全ての個体について光サーキットスイッチのサーキットを用いた完全なマッピングを力任せ探索を用いて探索する．完全なマッピングが見つかった場合，そのマッピングで使われたホストトポロジ内のリンクと，末端スイッチとして使われたノードをホストトポロジから削除する．

ホストトポロジの探索範囲及び GA の試行数として次の 2 通りを検討する．

- Sequential 探索 (SEQ): ホストトポロジを番号順にスイッチ数 32 で分割する．各分割トポロジ内で GA の試行を繰り返す．ただし，まだ末端スイッチとしてマッピングされていないホストトポロジ内のスイッチ数がゲストトポロジの末端スイッチ数を下回るか，連続して 3 回の試行で完全なマッピングが見つからない場合，その分割トポロジ内での探索を終了する．このような探索を全ての分割トポロジについて行う．
- Full 探索 (FULL): ホストトポロジの全ノードを探索対象とする．完全なマッピングが見つかるまで GA の試行を繰り返すが，連続して 10 回の試行で見つからない場合，全体の探索を終了する．

4.2 光サーキットスイッチを挿入したシステムの FAT TREE 内包性評価

本章では，電気パケットスイッチ間トポロジに光サーキットスイッチを挿入した場合のトポロジ内包性の評価を行う．

電気パケットスイッチ間トポロジには 3. 章で提案した局所化ランダムトポロジ (L-RANDOM) と，比較対象として一様ランダムリングトポロジ (DLN) を採用した．電気パケットスイッチ間トポロジのノード数は 1024 とし，次数は 10 とした．また，局所化ランダムトポロジについて，$n = 10$ とし，全ての次数で $N = 2, d = 1$ とした．

3. 2 章で示した 2 通りの光サーキットスイッチの挿入手法を適用し，さらに 4. 1 章で示した 2 通りの GA による探索手法を適用して探索を行う．電気パケットスイッチ 1 個当たりの光サーキットスイッチ数を $c = 0, 1, 2$ とする．マッピング対象トポロジは FAT TREE(2, 2, 4), FAT TREE(3, 2, 4) の 2 種類とした．

評価結果を図 5 に示す. 本章では, 評価対象を “TOPOLOGY-OCS_SPEC/SEARCH_RANGE” と表記する．この表記は，“TOPOLOGY” が電気パケットスイッチ間トポロジ，“OCS_SPEC” が光サーキットスイッチの挿入手法，“SEARCH_RANGE” が GA による探索範囲を示す．また，システム全体の電気パケットスイッチ数に対するマッピングに用いられた FAT TREE の末端スイッチの数を内包率 (Coverage) と定義する．

本図より，電気パケットスイッチ間トポロジが L-RANDOM である場合，DLN に比べトポロジの内包率が高い．これは，DLN が密に接続された場所が少ないトポロジである一方で，L-RANDOM が局所化されたトポロジで，密に接続された場所が多いためである．また，挿入手法，探索範囲が (REG, SEQ) の場合，光サーキットスイッチを電気パケットスイッチ当たり 1 個挿入することにより，FAT TREE(2, 2, 4) の内包率が 51%となり，2 個挿入することにより，FAT TREE(3, 2, 4) の内包率が 28%となる．これは，L-RANDOM が番号順に局所化されたトポロジであるため，番号順に分割したトポロジ内で探索する方がマッピング成功率が高まり，さらに光サーキットスイッチを番号順に挿入にすることにより，電気パケットスイッチ間が密に接続された箇所に多くのサーキットを確立することができるためである．

5. コスト評価

本章では，EPS 間局所化ランダムトポロジに対し規則的な OCS 挿入を行う提案ネットワークの総配線延長・消費電力の最適化を行う．また，従来のネットワークとの比較評価も提示する．

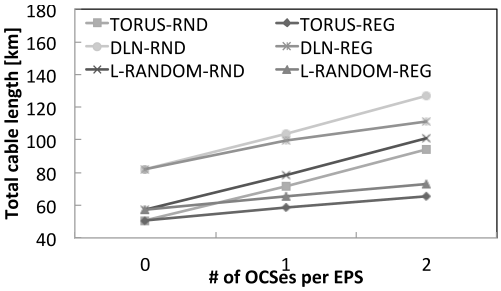


図 6　総配線延長

5.1 総配線延長評価

5.1.1 総配線延長の削減手法

本章では，トポロジの物理レイアウトを最適化し，フロア上でのキャビネット配置において，キャビネット内・キャビネット間の総配線延長を減らすことを目指す．

最適化手法は我々の先行研究 [14] に示されたものを用いる．本手法は，(1) クラスタリングによる各スイッチの格納キャビネットの最適化，及び (2) 2 次割当問題の適用によるキャビネット配置の最適化，の 2 つの手順より構成されている．各手順の目的は，手順 (1) はキャビネット間の配線数を最小化することであり，手順 (2) はキャビネット間の総配線延長を最小化することである．

本研究では，手順 (1) において Ward 法 [15] をクラスタリング手法として採用した．さらに手順 (2) において Simulated Annealing 法 (SA 法) [16] を配置手法として採用した．本研究では SA 法の反復回数は 1 億回，試行数は 5 回とした．

キャビネット配置について，通路幅を含めたラックの占有面積は幅 0.6[m] × 奥行 2.1[m] とした．配線のオーバーヘッドについては先行研究 [17] に基づき，キャビネット内配線を 2[m]，キャビネット間の配線オーバーヘッドをキャビネット当たり 2[m] とした．

5.1.2 比較評価

5. 1. 1 章で提示した配線長削減手法を本提案システム及び従来トポロジに光サーキットスイッチを挿入したシステムに適用した場合の総配線延長の評価を行う．

電気パケットスイッチについて，個数は 1024 個，スイッチの次数は 10 とした．キャビネットの格納する最大のスイッチ数を 2 個とした．光サーキットスイッチの挿入手法は 3. 2 章で記述した 2 通りの手法を用いた．電気パケットスイッチ間に適用するトポロジとして，5 次元 TORUS, 一様ランダムリングトポロジ (DLN), そして提案手法の L-RANDOM の 3 種類を用いた．局所化ランダムトポロジのパラメータ設定は $(n = 10, N = 2, d = 1)$ とした．

評価結果を図 6 に示す．この図において，横軸を電気パケットスイッチ 1 個当たりの光サーキットスイッチ数としている．

一様ランダムリングトポロジを電気パケットスイッチ間に用いる場合や，光サーキットスイッチをランダムに挿入する場合は，多数の長距離配線が出現するため，総配線延長が大きくなる傾向にある．局所化ランダムトポロジを用いる場合，最適化手法の適用により局所的に接続された短距離の配線が多くなるため，総配線延長が小さくなる．また，光サーキットスイッチを番号順に挿入する場合も同様の傾向が見られる．

これまで示した通り，局所化ランダムトポロジを電気パケッ

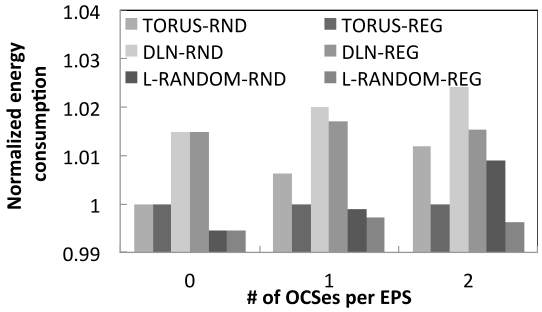


図 7 消 費 電 力

トスイッチ間に適用し，さらに光サーキットスイッチを番号順に挿入する手法は，低遅延性とトポロジ内包性を両立する．さらに，本章の評価により，提案システムは規則的な電気パケットスイッチ間トポロジに光サーキットスイッチを規則的に挿入した場合に対して，配線オーバーヘッドを 11%に抑えることが可能であることが示せた．

5.2 消費電力評価

本章では，5.1 章で評価した各システムの総消費電力の比較を行う．HPC システムの総消費電力は各スイッチに接続される配線の長さ等に左右される．本研究では商用のスイッチ及び配線の持つ特性 [18] [19] に基づき，配線当たりの消費電力を 1) 5[m] 未満: 4[W/link], 2) 5[m] 以上: 6[W/link] として評価を行った．対象のシステムの各配線長は 5.1 章での最適化手法適用後の値を用いた．

評価結果を図 7 に示す．本図において，横軸を電気パケットスイッチ当たりの光サーキットスイッチ数としている．縦軸はシステム全体の消費電力を示し，各値は，電気パケットスイッチ間 TORUS トポロジに光サーキットスイッチを番号順挿入したシステムの全体消費電力の値で正規化している．

評価結果は 5.1.2 章における総配線延長の結果と同様の傾向を示しており，局所化ランダムトポロジに光サーキットスイッチを番号順挿入するシステムが従来の規則的トポロジを用いたシステムと同程度以下の全体消費電力を達成することを示せた．

6. ま と め

本研究では，HPC 向け電気パケットスイッチ間トポロジの高いトポロジ内包性及び高い全体性能の両立を目指して，電気パケットスイッチと光サーキットスイッチを混在させたシステムを提案した．より具体的には，スモールワールド性と局所性を併せ持つ電気パケットスイッチ間トポロジに対する，複数の光サーキットスイッチの挿入手法，及び遺伝的アルゴリズムを用いたマッピング対象トポロジの探索手法を提案した．得られた知見を以下にまとめる．

- 局所化ランダムトポロジ (L-RANDOM) に対し，光サーキットスイッチを適切な場所に複数挿入し，さらに遺伝的アルゴリズムによる探索範囲を限定することにより，FAT TREE の内包率が大幅に向上した．
- 提案システムに対しクラスタリング及びキャビネット配置の最適化を施すことにより，従来の規則的トポロジを用いたシステムと同程度の総配線延長，総消費電力を達成した．

今後の課題として，マッピングに用いていない余剰リンクを用いたアプリケーション性能の向上等が挙げられる．

謝 辞 本研究の一部は科学研究費 (#25280018,#25730068), および JST CREST の助成を受けたものである．

文 献

[1] H. Subramoni, S. Potluri, K. C. Kandalla, B. Barth, J. Vienne, J. Keasler, K. A. Tomko, K. W. Schulz, A. Moody and D. K. Panda: “Design of a scalable infiniband topology service to enable network-topology-aware placement of processes”, SC, p. 70 (2012).

[2] 河野, 藤原, 松谷, 天野, 鯉渕：“光サーキットの補助的利用による高いトポロジ内包性を持つ HPC インターコネクト”, 電子情報通信学会技術研究報告 CPSY2013-111 (2014).

[3] J. Kim, W. J. Dally, S. Scott and D. Abts: “Technology-driven, highly-scalable dragonfly topology”, ISCA, pp. 77–88 (2008).

[4] J.-Y. Shin, B. Wong and E. G. Sirer: “Small-world datacenters”, SoCC, p. 2 (2011).

[5] A. Singla, C.-Y. Hong, L. Popa and P. B. Godfrey: “Jellyfish: Networking data centers randomly”, NSDI, p. 17 (2012).

[6] M. Koibuchi, H. Matsutani, H. Amano, D. F. Hsu and H. Casanova: “A case for random shortcut topologies for hpc interconnects”, ISCA, pp. 177–188 (2012).

[7] J. A. Ellis, S. Chow and D. Manke: “Many to one embeddings from grids into cylinders, tori, and hypercubes”, SIAM J. Comput., **32**, 2, pp. 386–407 (2003).

[8] K. J. Barker, A. F. Benner, R. R. Hoare, A. Hoisie, A. K. Jones, D. J. Kerbyson, D. Li, R. G. Melhem, R. Rajamony, E. Schenfeld, S. Shao, C. B. Stunkel and P. Walker: “On the feasibility of optical circuit switching for high performance computing systems”, SC, p. 16 (2005).

[9] 滝澤, 遠藤, 松岡：“光サーキットネットワークの補助的利用による HPC アプリケーション性能向上”, 情報処理学会論文誌コンピューティングシステム, **2**, pp. 110–121 (2009).

[10] G. Porter, R. D. Strong, N. Farrington, A. Forencich, P.-C. Sun, T. Rosing, Y. Fainman, G. Papen and A. Vahdat: “Integrating microsecond circuit switching into the data center”, SIGCOMM, pp. 447–458 (2013).

[11] “Sinet: Science information network”, `http://www.sinet.ad.jp/`.

[12] M. Koibuchi, I. Fujiwara, H. Matsutani and H. Casanova: “Layout-conscious random topologies for hpc off-chip interconnects”, HPCA, pp. 484–495 (2013).

[13] R. Chandrasekharam, V. V. Vinod and S. Subramanian: “Genetic algorithm for embedding a complete graph in a hypercube with a vlsi application”, Microprocessing and Microprogramming, **40**, 8, pp. 537–552 (1994).

[14] I. Fujiwara, M. Koibuchi and H. Casanova: “Cabinet Layout Optimization of Supercomputer Topologies for Shorter Cable Length”, Proc. of International Conference on Parallel and Distributed Computing, Applications and Technologies (2012).

[15] P. Pons and M. Latapy: “Computing communities in large networks using random walks”, Computer and Information Sciences ISCIS, pp. 284–293 (2005).

[16] D. T. Connolly: “An improved annealing scheme for the QAP”, European Journal of Operational Research, **46**, 1, pp. 93–100 (1990).

[17] J. Kim, W. J. Dally and D. Abts: “Flattened butterfly: a cost-efficient topology for high-radix networks”, ISCA, pp. 126–137 (2007).

[18] “7050 series 10/40g data center switches”, `https://www.cfa.harvard.edu/twpub/SMAwideband/NetworkSwitch/7050S_Datasheet_11_10_11.pdf`.

[19] “Cables - mellanox technologies”, `http://www.mellanox.com/page/cables`.