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

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，異なる通信パターンを持つ並列アプリケーションを 1 台のハイパフォーマンスコンピューティング (HPC) システム内で効率良く動作させることを目指している．一般的に，並列アプリケーションは特定のトポロジを想定して最適化されるため，アプリケーションが想定する論理トポロジと実際の物理トポロジが異なると性能を十分に発揮できない．本研究では，通常の電気スイッチネットワークに加え，光サーキットスイッチの補助的利用により，様々なスイッチ間トポロジを内包可能な低遅延結合網を提案する．評価結果より，提案トポロジは従来のトポロジに比べ，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

HPC Interconnect for High Topological Embeddability 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 parallel applications that have different communication patterns in a single High-performance computing (HPC) system. As parallel applications can be optimized for specific topologies, they cannot show full performance when a given physical topology is far from the logical topology the application assumed. This paper proposes to supplementally use optical circuit switches (OCSes) to patch the electrically-switched baseline network so that the proposed low-latency interconnection network can embed or emulate various network topologies. Experimental results show that the proposed topology achieves both high topology embeddability and the overall low latency compared to the conventional topologies.
Key words　High-performance computing (HPC), optical circut switching, network topology, interconnection networks, datacenter networks

1. はじめに

専用超並列計算機，PC クラスタなどの並列計算機，データセンターネットワークでは，システム設計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．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課題の 1 つとなる．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].

理想的には，対象とした並列アプリケーションの通信パターンに適したトポロジを採用することが望ましい．しかし，異なる通信パターンを持つ並列アプリケーションを実行する既存の HPC システムでは，そのようなトポロジの選択は難しい．したがって，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

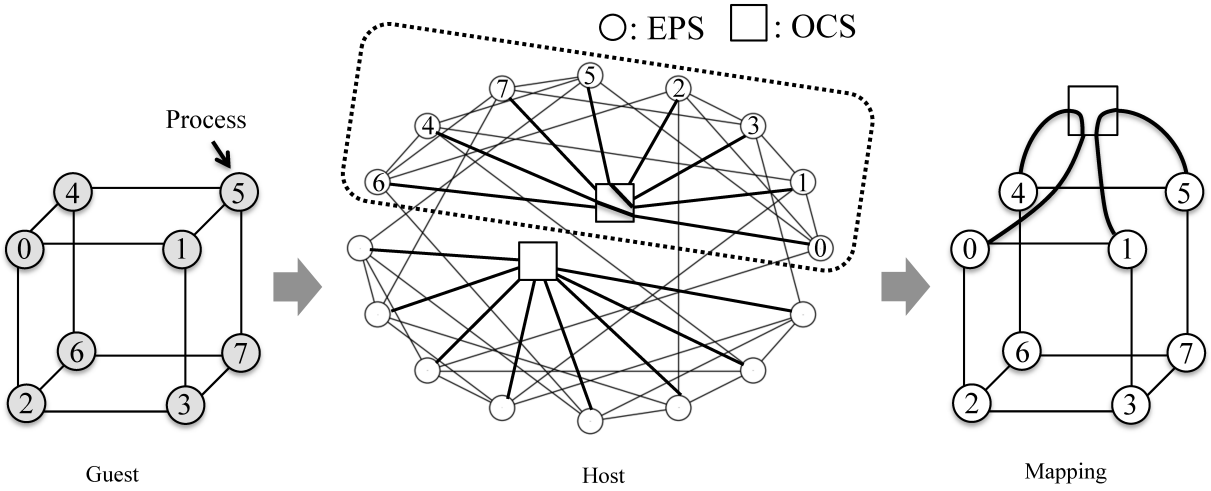


図 1　光サーキットスイッチを用いた論理トポロジのマッピング例

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，本報告では，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付加することにより，様々なトポロジを内包可能な相互結合網 (図 1 中央) を提案する．本相互結合網は，並列アプリケーション毎に適したトポロジを提供する．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．そこで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する．

具体的には，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．

本報告の貢献は以下である．

- 電気パケットスイッチ間に適用すべきトポロジを提案する．これは図 1 の Host トポロジから光サーキットスイッチを除いた構造である．グラフ解析の結果から，このトポロジは，TORUS や FAT TREE のようなトポロジの局所性を保ちつつ，トポロジ全体でのノード間ホップ数が小さい性質を持つ．
- 前述のトポロジを電気パケットスイッチ間に適用し，光サーキットスイッチを複数挿入することにより，従来の規則的なトポロジを複数内包可能な相互結合網を提案する．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ることにより，マッピング対象である Guest トポロジを多数探索し，配置できることが分かった．

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提案トポロジについて述べる．さらに，4. 章において光サーキットスイッチの挿入手法及びトポロジの探索手法について述べ，トポロジ内包性の評価を行う．最後に 5. 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 [2] 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 [3]〜[5] 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 [6] 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 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 [7]．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 [7] 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 [8] 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 [9]．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．

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 [10] 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，局所化ランダムトポロジ (L-RANDOM) を提案する．

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．

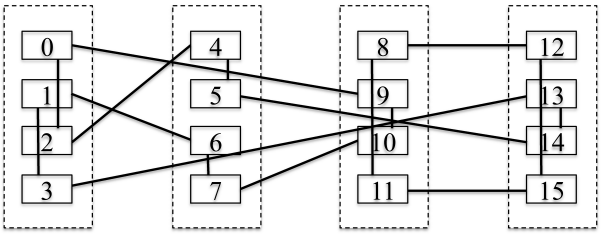


図 2 局所化ランダムトポロジの構成例 ($n = 2$, $N = 4$, $d = 1$)

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．

3.2 トポロジ解析

本章では，従来の規則的トポロジ及び我々が従来提案したランダムトポロジ [5], [11] との比較を行う．

本章での評価対象のトポロジとして，以下の 4 つを採用する．

- 局所化ランダムトポロジ (L-RANDOM) (本章では，N 及び d は全ての次元で全て等しいとする)
- HYPERCUBE

- 一様ランダムリング (DLN) [5]: Ring トポロジに対して Random Shortcut リンクを付加することにより構成されたトポロジ
- 単峰ランダムリング, $\sigma = 0.7$ (GAU(0.7)) [11]: Ring トポロジに対して，ショートカットリンクを以下の生成規則により生成，付加することにより構成されたトポロジ：標準偏差 0.7 の正規分布に比例する確率に従いリンクの目的地をランダムに選択し，ショートカットリンクを付加する．

ここで，各トポロジにおけるノード番号を以下のように定義する．まず，局所化ランダムトポロジについては，3.1 章で定義した通りとする．また，DLN, GAU(0.7) については，基となる Ring トポロジに沿って時計回りに番号を付けることとする．また，HYPERCUBE については，リンク接続の規則に用いられる一般的な番号付けの通りとする．このノード番号の定義は，以降の章でも同様に用いることとする．

3.2.1 Partitioning 性能評価

まず，提案トポロジのトポロジ分割性能の評価・比較を行う．トポロジ分割性能の指標として，あるトポロジを同サイズの複数トポロジに分割した際の，各ノードの平均次数を用いる．この次数が高いほど，分割後のトポロジがより多くのリンクを内包するため，Topology Embedding を考慮する際に特定のゲストトポロジを内包する可能性が高まる．

トポロジの分割方法については，ノード番号順に均等分割することとし，分割後の全トポロジの平均ノード次数の平均を測定した．

なお，ノード次数を 10, ノード数を 1024 に統一した．また，局所化ランダムトポロジについて，$n = 10$ とし，全ての次元について $N = 2, d = 1$ とした．

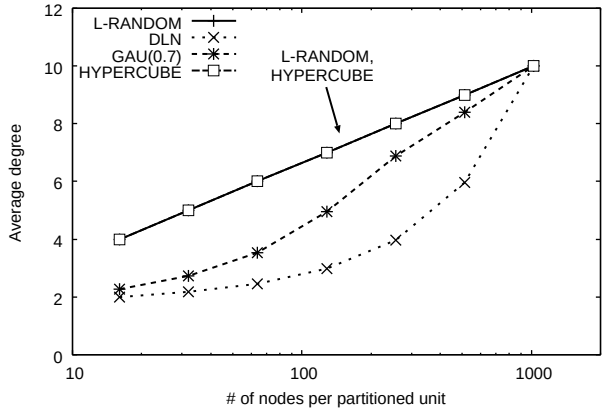


図 3　トポロジ分割後の平均次数

トポロジ分割後の平均次数の評価結果を図 3 に示す．DLN トポロジでは分割サイズが 512 ノードである場合まで，平均次数が HYPERCUBE に比べ著しく小さいことが示されている．これは，DLN トポロジが一様に接続を持つトポロジであるため，トポロジを分割した際に分割後のトポロジ間にリンクが多く存在し，分割後のトポロジ内でのリンクが少なくなるためである．

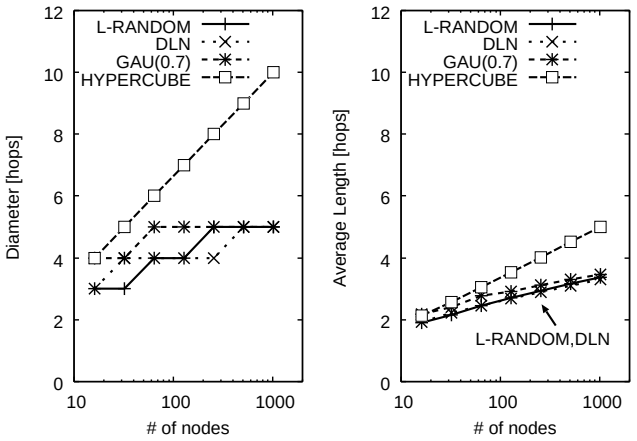
また，GAU(0.7) トポロジでは，局所的にリンクが接続されているため，DLN より平均次数が高いものの，リンクの接続先が確率的に局所化されていることから，HYPERCUBE に比べ平均次数は小さくなっている．

一方，局所化ランダムトポロジでは，HYPERCUBE と同様にノード番号に沿って規則的にリンクを接続しているため，分割後の平均ノード次数は HYPERCUBE と全く同じとなっており，DLN トポロジに比べ最大で 2.44 倍の平均次数を達成している．

3.2.2 全体性能評価

本章では，ネットワークサイズを変化させた場合の全体性能の比較を行う．ここでの全体性能とは，システム全体でのノード間の直径 (最小ホップ数の最大) 及び平均距離 (最小ホップ数の平均) を示す．

本章では，局所化ランダムトポロジについて全ての次元で $N = 2, d = 1$ として，n を 4 から 10 まで (ノード数を 2^4 から 2^{10} まで，次数を 4 から 10 まで) 変化させた．さらに，比較対象である他トポロジのノード数及びノード次数を局所化ランダムトポロジと合わせた．



(a) 直径　　(b) 平均距離

図 4　分割後トポロジのホップ数

評価結果を図 4 に示す．この図より，1024 ノードにおける局所化ランダムトポロジの直径は DLN と同じであり，平均距離は DLN から 2%の悪化に留まっている．さらに，HYPERCUBE に比べ 50%の直径の削減，33%の平均距離の削減を達成している．すなわち，本章で提案した局所化ランダムトポロジは，従来提案されたランダムトポロジである DLN や GAU(0.7) と同様に，ノード間のランダム接続によるスモールワールド性によりホップ数を削減出来ていることが分かる．

3.2.1 章の結果と合わせ，局所化ランダムトポロジが高いトポロジ分割性能と全体性能の高さを併せ持つトポロジであることが示されている．

4. 光サーキットスイッチの挿入による Embedding

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することにより，トポロジの Embedding 性能 (内包性) を向上させるシステムを提案する．また，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 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し，その手法を適用したトポロジ内包性の評価を行う．

4.1 光サーキットスイッチの挿入方法

本章では，光サーキットスイッチの挿入方法の提案を行う．

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．

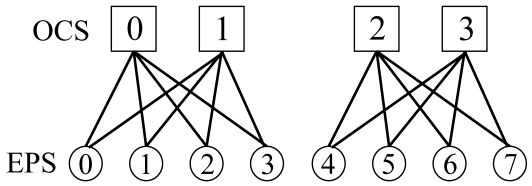


図 5　光サーキットの Regular 挿入 (電気パケットスイッチ数: 8, $p = 4, c = 2$)

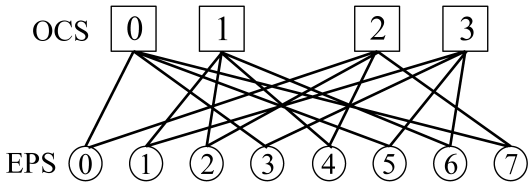


図 6　光サーキットの Random 挿入 (電気パケットスイッチ数: 8, $p = 4, c = 2$)

- Regular 挿入 (REG): 図 5 のように，3.2 章で定義した各ノード (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 6 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．

4.2　2 次元 TORUS の Embedding

本章では，ゲスト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 2 次元 TORUS をマッピング対象のトポロジとした．

TORUS 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．

4.2.1　GA(遺伝的アルゴリズム) を用いた TORUS の探索

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 TORUS 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 [12] 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．

GA の各個体は探索対象トポロジである TORUS のノード数と同じ長さのベクトルから構成され，各個体の i 番目の要素の値は TORUS の i 番目のノードがマッピングされるホストトポロジ内のノード番号である．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 50%, 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ノード及びリンク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探索対象の TORUS のノード数で分割する．各分割トポロジに対する GA の試行数を最大で 3 回とし，全ての分割トポロジについて探索を行う．
- 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．

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

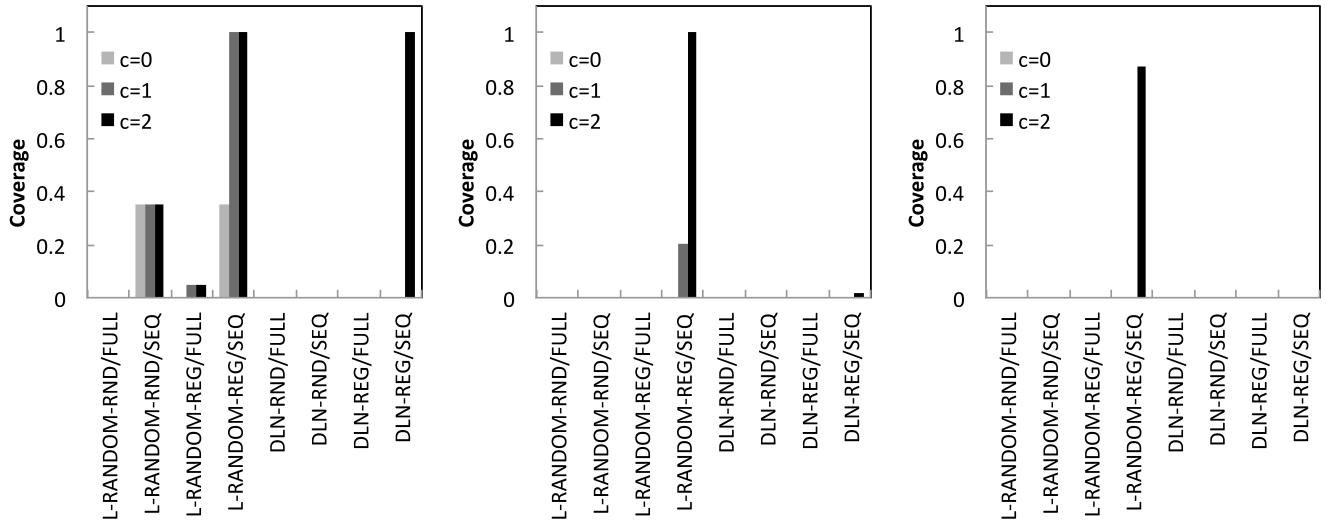
本章では，電気パケットスイッチ間トポロジに光サーキットスイッチを挿入した場合のトポロジ内包性の評価を行う．

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．

4.1 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 4.2 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 2×4 TORUS, 4×4 TORUS, 4×8 TORUS とした．

評価結果を図 7 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 TORUS の完全マッピングに用いられた電気パケットスイッチ数を内包率 (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，3.2.1 章で示した通り，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



(a) 対象トポロジ: 2×4 TORUS　(b) 対象トポロジ: 4×4 TORUS　(c) 対象トポロジ: 4×8 TORUS

図 7　2 次元 TORUS の内包率

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，2×4 TORUS の内包率が 100%となり，2 個挿入することにより，4×4 TORUS, 4×8 TORUS の内包率がそれぞれ 100%, 87.5%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. 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を提案し，さらに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．

- グラフ解析より，提案トポロジである局所化ランダムトポロジ (L-RANDOM) は，HYPERCUBE と同等の局所性 (トポロジ分割性能) と一様ランダムリングトポロジ (DLN) と同等の低遅延性を達成した．
- 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，2 次元 TORUS の内包率が大幅に向上した．

今後の課題として，(1) ランダム，FAT TREE などの TORUS 以外の論理トポロジを対象とした内包性の評価，(2) アプリケーション毎のネットワークのエネルギー最適化，ケーブリングの見積りが挙げられる．

謝 辞　本研究の一部は科学研究費 (#25280018,#25730068), 戦略的情報通信研究開発推進制度（SCOPE）若手 ICT 研究者等育成型研究開発, 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] J. Kim, W. J. Dally, S. Scott and D. Abts: “Technology-driven, highly-scalable dragonfly topology”, ISCA, pp. 77–88 (2008).

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

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

[5] 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).

[6] 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).

[7] 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).

[8] 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).

[9] 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).

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

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

[12] 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).