7A-02

低直径ネットワークトポロジのラック配置最適化

河野 隆太 †　　松谷 宏紀 †　　鯉渕 道紘 ‡　　天野 英晴 †

† 慶應義塾大学大学院理工学研究科　‡ 国立情報学研究所

1 はじめに

次世代の高性能システムにおける多くのマルチコア並列アプリケーションでは、数百 ns から 1 μs の低 MPI 通信遅延が必要となることが予測されている。ネットワーク内ではスイッチ遅延が支配的である一方、フリットの注入遅延、リンク遅延などは相対的に小さい。従って、低直径、短い平均距離 (ホップ数) のトポロジをスイッチ間ネットワークに適用することがネットワークの低遅延化につながる。

スイッチ間ネットワークを構成するために、各スイッチをノード、リンクをエッジとしたトポロジとしてモデル化し、その直径や平均距離 (ノード間ホップ数の最大値、及び平均値) の小さいトポロジを実際のネットワークに適用することが多い。ラック間の長距離リンク数を大幅に削減しつつ直径 2 を実現可能なトポロジとして、Slim Fly [1] が提案されている。

最近の研究で、Order/Degree Problem と呼ばれる最小直径トポロジを探索するこれまでの取り組み [2] により、規則性を持たない小規模グラフを複製し点対称に接続するトポロジが、広範囲のシステム規模で最適であることが明らかになっている [3]。このトポロジのクラスタ性・対称性を活用し、実システム適用時に総配線延長・総コストを大幅に削減するための拡張手法を提案する。

2 レイアウト最適な点対称トポロジ

本研究で提案するトポロジの基となる点対称性を用いた低直径トポロジ [3] は Order/Degree Problem [2] と呼ばれる大規模・低遅延なネットワークを求めるための最適な解法の一つとされる。低直径なトポロジを求めるため、焼きなまし法 (Simulated Annealing; SA) による解の探索を行っている。

本研究では、この点対称トポロジ生成時に定義される “グループ” に着目し、各グループ内にリンクの局所性を持たせるアプローチを探究する。具体的には、グループ数を必要ラック数と同じ値に設定し、トポロジ全体のリンク数に対する各グループ内に含まれるリンク数の割合の下限値を導入する。焼きなまし法の繰り返しにおいて新たな解を生成する際に、グループ内のリンク数がその割合を下回らないように維持する。グループ内のスイッチを単一ラックに格納することにより、最終的に生成されるトポロジにおいて単一ラックに含まれるリンク数の割合が一定以上であることを保証可能となる。

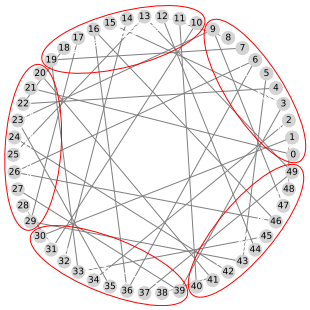


図 1: レイアウト最適な点対称トポロジ

Algorithm 1 新たな対称性トポロジの生成

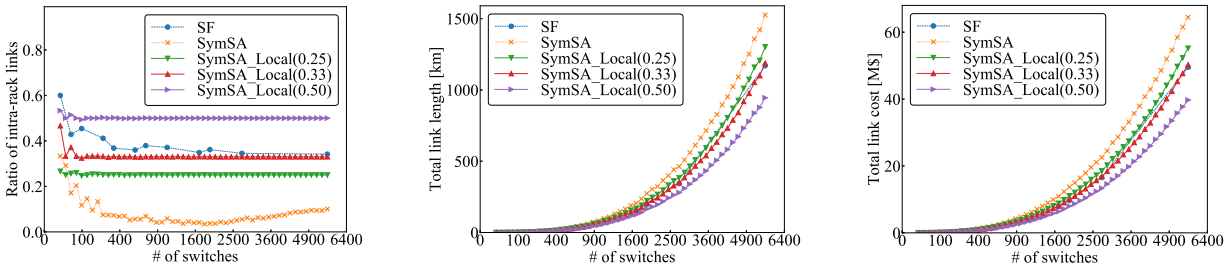
```
1:  function EDGE_EXCHANGE(lines, edge[lines][2], group, min_links)
2:      /* Constant value min_links added to arguments */
3:      do while // Added loop
4:          do while
5:              do while
6:                  do while
7:                      line[0] = Random(lines)
8:                      line[1] = Random(lines)
9:                  end do while line[0] == line[1]
10:             end do while check_duplicated_vertex(line, lines, edge)
11:             if check_symmetric_edge(line, lines, edge, groups) then
12:                 edge_exchange_1g_opt(line[Random(2)], lines, edge, groups)
13:             else
14:                 edge_exchange_2g_opt(line, lines, edge, groups)
15:             end if
16:         end do while check_multigraph(line, lines, edge, groups)
17:     end do while links_in_group(line, lines, edge, groups) < min_links
18: end function
```

点対称トポロジにおいて定義されるグループ数 g をネットワークの構成に必要なラック数と同じ値に設定する。さらに、トポロジの総リンク数に対するグループ内リンク数の割合を示す値 σ を導入する。

提案トポロジの生成例を図 1 に示す。この例においてスイッチ数・次数・グループ数をそれぞれ $n = 50, d = 7, g = 5$ とし、グループ内リンク数の割合を $\sigma = 0.8$ としている。σ を設定しその値を大きくすることにより、グループ内スイッチ間のリンク数が増大し、グループ間のリンク数を削減可能となる。

拡張手法におけるリンク入れ替えによる新たな解を生成するためのアルゴリズムの疑似コードを Alg. 1 に示す。関数の引数として各グループ内に存在するスイッ

Rack Layout Optimization for Low-Diameter Network Topologies
†Ryuta Kawano †Hiroki Matsutani ‡Michihiro Koibuchi †Hideharu Amano
†Graduate School of Science and Technology, Keio University ‡National Institute of Informatics



(a) ラック内リンクの割合　(b) 総配線延長　(c) 総配線費用

図 2: 配線コスト

チ間リンクの下限値 “min_links” を新たに導入している。min_links= $\frac{n \cdot d \cdot \sigma}{2}$ として算出される。17 行目においてリンク入れ替え後のグループ内リンク数が下限値 min_links を下回る場合、入れ替えるリンクの選択をやり直す。

3　配線コスト評価

評価対象のトポロジは Slim Fly (SF) [1] , オリジナルの点対称トポロジ SymSA, 提案のレイアウト最適な点対称トポロジ SymSA_Local(σ) である。ラック数を 3~53, スイッチ数を 18~5,618, 次数を 5~79 とした。提案トポロジのラック内配線の割合を $\sigma \in \{0.25, 0.33, 0.5\}$ とした。焼きなまし法の繰り返し回数は 10,000 回とした。

全体のリンク数に対するラック内リンク数の割合を評価した結果を図 2a に示す。Slim Fly のラック内リンク数の割合はネットワークサイズの増加に伴い値が 1/3 に漸近している。SymSA トポロジはスイッチ数 1,800 においてラック内リンク数の割合が 3.41 %と最小化される一方、提案の SymSA_Local(σ) では、ラック内リンクの割合の下限値を最適化の条件に設定することにより、生成されたトポロジで σ 以上の値を保証可能となっている。

総配線延長・配線コストの評価結果を図 2b, 図 2c にそれぞれ示す。ラックの寸法は幅 60 cm × 奥行 210 cm (通路幅含む) とし、ラック間距離はマンハッタン距離を用いて計算を行った。ラック内配線の長さは 1 m と固定し、ラック外配線には 2 m のオーバーヘッドが生じるものとした。配線コストは既存のコストモデル [1] を基に算出した。

オリジナルの SymSA は最適化の過程でラック外リンクが増加することから、Slim Fly に比べ総配線延長が最大 30.4 %悪化している。一方、本研究の提案である SymSA_Local(σ) は、ラック内リンク数の保証により、5,618 スイッチ, $\sigma = 0.50$ において、総配線延長を Slim Fly, SymSA に比べそれぞれ 19.2 %, 38.0 %削減可能となっている。さらに、総配線費用をそれぞれ 19.8 %, 38.4 %削減可能となっている。

4　まとめ

本研究では、Order/Degree Problem の最適解の一つである点対称トポロジの生成手法を拡張し、ラックレイアウトの配線長を削減可能な新たなトポロジの提案を行った。具体的には、最適化の中での新たな解の生成時にラック内配線数の下限を保証するよう改良し、直観的なクラスタリング・マッピング手法によってラック間配線数・総配線延長を削減した。

ケーススタディの結果、直径 2 の準最適なトポロジとして知られる Slim Fly に比べ、総配線延長・総配線費用をそれぞれ 19.2 %, 19.8 %削減した。

参考文献

[1] Maciej Besta and Torsten Hoefler. Slim Fly: A Cost Effective Low-Diameter Network Topology. In *Proc. of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pp. 348–359, Nov 2014.

[2] Graph golf: The order/degree problem competition. `http://research.nii.ac.jp/graphgolf/`.

[3] Masahiro Nakao, Hitoshi Murai, and Mitsuhisa Sato. A Method for Order/Degree Problem Based on Graph Symmetry and Simulated Annealing with MPI/OpenMP Parallelization. In *Proc. of International Conference on High Performance Computing in Asia-Pacific Region (HPC Asia)*, pp. 128–137, Jan 2019.