C-045

配線長制限ランダムトポロジ向けのスケーラブルなルーティング手法

河野 隆太 †　中原 浩 †　藤原 一毅 ‡　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

次世代の高性能計算システムにおける多くの並列アプリケーションは 1μ 秒以下の通信遅延が必要と予測されている．よって，こうした高性能計算システム向けの低遅延ネットワークの研究開発が今後，重要となる．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．

最近の研究で，従来の規則網とは異なり，ノード間をランダムに接続したネットワークトポロジがホップ数を劇的に削減でき，それらが HPC やデータセンター用のネットワークに適用可能であることが示されている [1]．このような不規則網は低ホップ数のために総配線延長の悪化を引き起こすことから，配線長に制限を設けた上で不規則にノード間を接続するという解決策が提案されている [2].

不規則網のスケーラビリティを向上させる上でさらに問題となるのは，従来の最短経路ルーティングではすべての宛先ノードへの経路情報を各ノードが蓄える必要がある点である．本研究ではこの問題を解決するため，先述の配線長制限を課したランダムネットワークの局所性とスモールワールド性に着目し，ルーティングテーブルの削減と低ホップ数の維持の両方を達成する新たなルーティング手法を探求する．

2 問題定義

2.1 グラフに関するパラメータ

本論文では，計算ノードを集約する Top-of-Rack (ToR) スイッチが格納されたキャビネットが，平面フロア上に配置されることを想定する．さらにスイッチ間ネットワークを，各スイッチをノード，スイッチ間リンクをエッジとする無向グラフとしてモデル化する．トポロジが展開される平面空間を $n \times n$ の二次元座標上とし，各格子点上に 1 つのノードが存在することとする．すなわち，ノード集合 N と二次元座標の一辺 n は $|N| = n^2$ の関係を満たす．エッジ集合 E に含まれるノード i, j 間のエッジを $e_{i,j}$ とし，その配線長 $l_{i,j}$ を，2 ノード間のマンハッタン距離に等しいこととする．すなわち，ノード i, j の x, y 座標をそれぞれ $(x_i, y_i), (x_j, y_j)$ として，$l_{i,j} = |x_i - x_j| + |y_i - y_j|$ を満たすものとする．

2.2 テーブルエントリの定義

本論文では，ルーティング情報である単一のテーブルエントリを，宛先ノード v_{dst} とそのノードへの最短経路上の次ホップのノード v_{next} の組として定義し，$< v_{\mathrm{dst}}, v_{\mathrm{next}} >$ と表記する．本研究では，グラフと，各ノードが蓄えることが可能な最大テーブルエントリ数 $t_{\max}$ を入力として，最短経路ルーティングに対するホップ数の悪化率がなるべく小さいルーティングテーブルエントリを構築することを目標とする．

3 提案手法

3.1 テーブルエントリの構成法

本節では，本提案で各ノードが蓄えるルーティング情報の構成法について記述する．以下に本提案で構成されるルーティングテーブルによって形成される 3 種類の最短経路及びそれに対応するテーブルの構成アルゴリズムを記述する．

1. エッジで接続された 2 ノード間の経路: グラフ上で隣接する 2 ノード間について経路情報を与える．すなわち，全ての $e_{i,j} \in E$ に対し，エントリ $< i, i >$ をノード j に，エントリ $< j, j >$ をノード i に，それぞれ加える．

2. 座標上で隣接する 2 ノード間の経路: 2.1 節で定義した $n \times n$ の 2 次元座標上で，マンハッタン距離が 1 だけ離れた 2 ノード間 i, j について，最短経路を取るための情報を以下の通り追加する．ノード i からノード j への最短経路を

$$P_{i,j} := \{m_z \mid 0 \le z \le |P_{i,j}|-1, m_0 := i, m_{|P_{i,j}|-1} := j\}$$

とし，$0 \le z \le |P_{i,j}|-2$ に対して，エントリ $< j, m_{z+1} >$ をノード m_z に追加する．さらに，ノード j からノード i への最短経路についても同様に，ノード i へのエントリーを追加する．

A Scalable Routing Method for Random Topologies with the Link Length Limited
†Ryuta Kawano †Hiroshi Nakahara ‡Ikki Fujiwara †Hiroki Matsutani ‡Michihiro Koibuchi †Hideharu Amano
†Graduate School of Science and Technology, Keio University ‡National Institute of Informatics
AcknowledgmentA part of this work was supported by JSPS KAKENHI Grant Number 15J03374.

3. 上記以外の 2 ノード間: 上記 1, 2 に示した 2 ノード間の最短経路についてエントリを構築した後，それ以外の最短経路について，ホップ数の小さいものからテーブルエントリが空きがある限りエントリを追加していく．これを実現するため，グラフ内の全ノードについて，各宛先ノードをルートとした最短経路を示すスパニングツリーを生成し，これらのツリーについて同時に幅優先探索を行う．探索中に訪れたノード u のテーブルに空きがある場合，ノード u の親ノードを w_{succ} としてエントリ $< v, w_{\mathrm{succ}} >$ をノード u に追加する．ノード u のテーブルに空きがありエントリを追加した，もしくはすでに同一エントリがノード u のテーブル内に存在していた場合は，さらにノード u の子ノード $w_{\mathrm{prev}} \in W_{\mathrm{prevs}}$ について，幅優先探索を続行する．それ以外の場合は探索を打ち切る．

3.2 ルーティング手法

宛先ノード v のパケットに対するノード u 上でのルーティングは以下のように行う．ノード u に蓄えられる全てのエントリのうち，パケットの宛先ノード v とエントリの宛先ノード v_{dst} のマンハッタン距離 $l_{v,v_{\mathrm{dst}}}$ を最小とするようなエントリ $< v_{\mathrm{dst}}, v_{\mathrm{next}} >$ を選択し (そのようなエントリが複数ある場合は，宛先ノードの番号順が最も若いものを選択する．)，v_{next} を次ホップとする．

3.1 節で示したエントリ情報のうち，1, 2 番目に列挙した経路情報により，任意のノード間での到達性を保証可能となる．さらに，3 番目に示した遠方ノード間の経路情報により，効率的にホップ数削減が可能となる．

4 最大エントリ数と平均ホップ数の評価

本章では，提案手法を配線長制限を課したランダムグラフに適用し，平均ホップ数の評価を行う．評価対象とするグラフを，最大配線長 r, 次数 d の条件の下で，すべてのノードについて次数が等しくなるよう生成した正則ランダムグラフとする．本評価では，$|N| = 256, r = 4, d = 4$ とし，最大テーブル数 t_{max} を変化させた際に達成される平均ホップ数を評価した．また比較対象を Cowen により提案されているコンパクトルーティング手法 [3] とした．この従来手法では，グラフ中の一部のノードを “landmark” ノードとし，各ノードが landmark ノードへのグローバルなルーティング情報を保持している．また, 従来手法ではパラメータの設定により各ノードが保持するエントリ数・平均ホップ数が変化するため，パラメータにより取りうる全ての最大テーブル数と，その最大テーブル数により達成された平均ホップ数を評価結果とした．

図 1 にその結果を示す．この図において，各プロットは乱数の異なる 10 個のグラフについて評価結果の平均を取ったものである．従来手法が取りうる t_{max} の最小値 42 において，提案手法は 6.5%平均ホップ数が悪化している．一方，テーブルサイズが大きい場合，提案手法は従来手法に比べ平均ホップ数を最大 1.8%改善している．これらは．配線長制限ランダムグラフのスモールワールド性が，十分大きなローカル情報によってホップ数削減に効果的となることを示している．

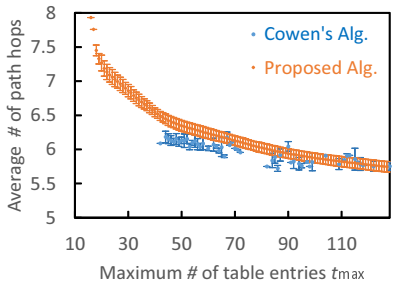

図 1: $|N| = 256, r = 4, d = 4$ における最大エントリ数 t_{max} に対する平均ホップ数の変化.

さらに，提案手法は従来手法で取りえないテーブルエントリサイズを取ることが可能となっている．取りうる最小のテーブルエントリサイズは従来手法に比べ 62%削減可能となっており，より実装面での柔軟性が高いことが示されている．

5 おわりに

本稿では，配線長制限を課した不規則網向けの新たなルーティング手法を提案し，従来のコンパクトルーティング手法に比べほぼ同等のホップ数を達成しつつテーブルサイズをより柔軟に設定可能であることを示した．

謝辞 本研究の一部は JSPS 科研費 15J03374 の助成を受けたものである．

参考文献

[1] Michihiro Koibuchi, Hiroki Matsutani, Hideharu Amano, D. Frank Hsu, and Henri Casanova. A Case for Random Shortcut Topologies for HPC Interconnects. In *Proc. of the International Symposium on Computer Architecture (ISCA)*, pp. 177–188, 2012.

[2] Michihiro Koibuchi, Ikki Fujiwara, Hiroki Matsutani, and Henri Casanova. Layout-conscious Random Topologies for HPC Off-chip Interconnects. In *Proc. of the International Symposium on High Performance Computer Architecture (HPCA)*, pp. 484–495, 2013.

[3] Lenore J. Cowen. Compact Routing with Minimum Stretch. In *Proc. of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 255–260, 1999.