2A-01

オフチップ相互結合網向け不規則トポロジのための容易に実装可能なルーティングアルゴリズム

河野 隆太† 中原 浩† 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 ‡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，1μ秒以下の低MPI通信遅延が必要となることが予測されている[1]．従って，これらの高性能計算システムに向けた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フリットの注入遅延，リンク遅延に比べ支配的である[2]．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やデータセンターネットワーク(DCN)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[3]．また，この応用として，2次元座標上でノード間の配線長を制限しつつランダムにノード間を接続するトポロジが，低遅延性とレイアウト容易性を両立することが可能であると示されている[4]．このようなネットワークでは，ノード間のランダム接続によるスモールワールド性と，近隣ノード同士が密に繋がる局所性の2つの性質が同時に表れている．

ランダムなネットワークの問題点として，各ルータ間の最短経路を単純なロジックを用いて求められない点が挙げられる．これまでランダムネットワークにおいては，up* / down* ルーティングなどに代表される，各ノードがほかの全ノードへの次ホップの経路情報を持つルーティング手法が用いられてきた．しかし，テーブルサイズがノード数Nに対して各ルータ内で$N \log N$となりスケーラビリティが乏しくなる問題があった．

そこで，本研究では，先述の2次元空間上に配置されたノードに対して配線長制限を課すランダムトポロジを対象とした，低ホップ数と高スケーラビリティを両立する新たなルーティング手法を提案する．

Easily implementable routing algorithms for irregular topologies in off-chip interconnects
†Ryuta Kawano †Hiroshi Nakahara ‡Ikki Fujiwara †Hiroki Matsutani †Hideharu Amano ‡Michihiro Koibuchi
†Graduate School of Science and Technology, Keio University ‡National Institute of Informatics

2 ルーティングテーブルの構築

各ノードのルーティングテーブルをE_nとして，ノード番号n_tと対応する次ホップの組を1つのエントリとする．また，各ノードにおけるテーブルエントリ数の上限を$e_{\max}$と設定し，すべてのノードで同じ上限値を持つものとする．本提案では，ルーティング情報として，各ノードから2次元座標上で上下左右に隣接するノードへ到達可能となるような情報が必要となる．そこで，以下の手順で各ノードのルーティングテーブルE_nへエントリを追加する．(1) トポロジ上の各ノードi $(0 \le i < N)$について，2次元座標上でマンハッタン距離1で隣接する各ノードへの最短経路$P'_i = \{P_{i,0}, ..., P_{i,(|P'_i|-1)}\}$を計算する．(2) 各隣接ノード$j$への経路$\{i, m_0, m_1, ..., j\}$について，隣接ノード$j$を除く経路上の全ノードのルーティングテーブルへ，宛先ノードjと，各ノードでの次ホップの組をエントリとして追加する．

先述のエントリ追加を全ノード，全経路について行った後，余っているエントリー領域に対して，座標上で隣接しない遠方ノードへの最短経路情報を以下の手順で追加する．(1) トポロジ上のすべてのノード対$(i, j)(0 \le i, j < N; i \ne j)$について，最短経路$S_{i,j}$を計算する．(2) 各最短経路を (A) ホップ数，(B) ノード番号, の優先順位で順に取り出す．取り出した最短経路について，宛先ノードjを除く経路上の全ノードのルーティングテーブルに空きエントリがある場合，それらのノードに宛先ノードjと，各ノードの次ホップの組をエントリとして追加する．最後に，各ノードのルーティングテーブルE_nについて，格納されているエントリー順を，そのノードから宛先ノードn_tへのホップ数が小さい順にソートする．

3 ルーティング方法

本章では，先述のルーティングテーブルを用いたルーティング方法について記述する．パケットの宛先ノード$n_{\mathrm{p_dst}}$とパケットの存在する現在地のノードn_{tmp}を入力として，現在地のノードが保持するルーティングテーブルに記されている経路情報から，次の条件で次ホップを選択する．ノードi, j間の座標距離を$C(i, j)$としたとき，パケットが存在するノードのエントリテー

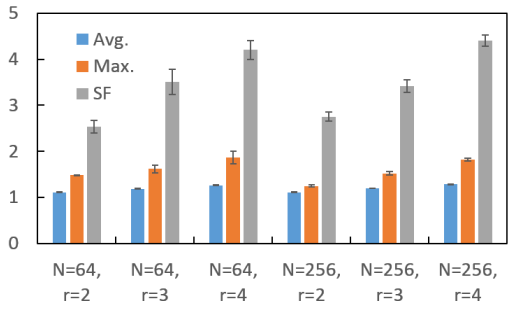
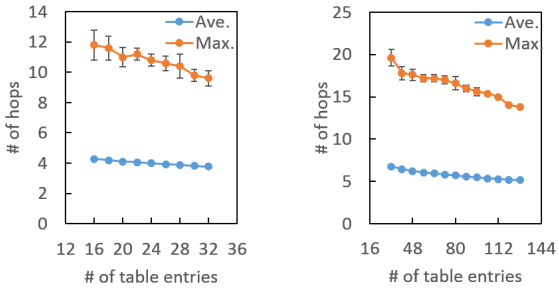


図 1: 平均/最大ホップ数の悪化率，Stretch Factor



(a) $N = 64, r = 4$ (b) $N = 256, r = 8$

図 2: テーブルサイズと平均ホップ数

ブルの中で，$C(n_{\text{p\_dst}}, n_{\text{t}})$ が最も小さくなるエントリ n_{t} を選択し，対応するポート番号を次ホップとする．その際，$C(n_{\text{p\_dst}}, n_{\text{t}})$ が最小となるエントリが複数存在する場合は，パケットの存在するノード n_{tmp} からエントリのノード n_{t} までのホップ数 $h(n_{\text{tmp}}, n_{\text{t}})$ が最も小さいエントリを選択する．各エントリはホップ数の小さい順にソートされているため，候補となるエントリーから順番が最も若いものを選択する．

2 章の記述の通り，テーブル構築時に最短経路上のすべてのノードに次ホップ情報を追加しているため，あるエントリによって転送されたパケットは，次ホップの中間ノードでもその最短経路の情報を利用可能である．また，2 次元座標上での隣接ノードへの最短経路を保持することにより，選択するエントリノードを単調に宛先ノードへ近づけることができる．以上により，ライブロックを回避可能である．

4 評価

図 1 に $N = 64, 256$ におけるホップ数評価を示す．本図における評価結果は，提案手法を適用した場合のノード間ホップ数を，最短経路を取った場合のホップ数で正規化した値となっている．各凡例はそれぞれ，“Ave.”, “Max.” が平均, 最大ホップ数を表し，“SF” は Stretch Factor を表す．SF は全ノード対のうち最短経路距離に対する悪化率の最大値をとったものである．最大配線長の制限については，$r = 2, 3, 4$ とした．また，次数は $d = 4$ とした．各ノードのルーティングテーブルのエントリ数は $N = 64$ の時 $e_{\max} = 16$, $N = 256$ の時 $e_{\max} = 25$ とした．$N = 256, d = 4, e_{\max} = 25$ の場合，1 ノード当たりのテーブルサイズは 250 bit となり，up* / down* ルーティング等で全ノードへの宛先を持つ場合と比べてテーブルサイズを 51%削減できる．さらに，同ノード，同次数で最大配線長 $r = 2$ の時，最短経路に対する平均ホップ数の悪化率を 11%に抑制可能である．

また，図 2a，図 2b に，次数 $d = 4$ の下でエントリ数を変化させた場合の平均・最大ホップ数を示す．エントリ数の増加により，特に最大ホップ数を削減可能であることが分かった．$(N, r) = (256, 8)$ においては，エントリ数を 4 倍にすることにより，最大ホップ数を 30%削減可能である．よって，局所性の少ないトラフィックを持つアプリケーションに対しては，エントリ数を増やすことで効率的に実行性能を向上させることが分かる．さらに，最大エントリ数をノード数 N の半分とした場合，$N = 64, 256$ のそれぞれにおいて，同ノード数，同次数の 2D Torus に比べ，平均ホップ数をそれぞれ 7%, 36% 削減可能であることが分かった．

5 おわりに

本研究では HPC 用不規則網ネットワークに対する実装容易性の高いルーティング手法を提案した．遠方ノードと隣接ノードへの経路情報により，256 ノードにのネットワークおいて，ホップ数の悪化率を 11%に抑えつつ，テーブルサイズを最大 51%削減できることが分かった．さらに，同ノード・次数の 2D Torus に比べ，平均ホップ数を最大 36%削減可能であることを示せた．

参考文献

[1] K. Scott Hemmert et al. Report on Institute for Advanced Architectures and Algorithms, Interconnection Networks Workshop 2008. http://ft.ornl.gov/doku/_media/iaaicw/iaa-ic-2008-workshop-report-v09.pdf.

[2] J. Tomkins. Interconnects: A Buyers Point of View. ACS Workshop, 2007.

[3] Michihiro Koibuchi et al. A Case for Random Shortcut Topologies for HPC Interconnects. In *Proc. of the International Symposium on Computer Architecture (ISCA)*, pp. 177–188, 2012.

[4] Michihiro Koibuchi et al. 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.