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.

ランダムなネットワークトポロジのためのラック配置最適化

藤原　一毅†　　鯉渕　道紘†

† 国立情報学研究所 / JST
E-mail: †{ikki,koibuchi}@nii.ac.jp

あらまし　スーパーコンピュータの大規模化が進むにつれ，通信遅延がメニーコア並列アプリケーションの性能に及ぼす影響が大きくなってきている．そのため，スーパーコンピュータでは高次元スイッチを用いた低遅延トポロジの活用が重要となりつつある．我々のこれまでの研究により，スイッチ間にランダムにリンクを接続したトポロジが低遅延性の点で優れていることが分かったが，ハイパーキューブなどの規則性を持つトポロジと比べてレイアウトの配線が長くなる問題があった．本研究では，(1) ランダムなショートカットリンクの生成時に多少の偏りを持たせ，(2) ラックレイアウトをモデル化して最適化問題として解くことにより，遅延を維持したまま配線長を最大 38%削減できたことを報告する．
キーワード　相互結合網，ランダムトポロジ，配線長，レイアウト，ハイパフォーマンスコンピューティング

Rack Layout Optimization for Random Network Topology

Ikki FUJIWARA† and Michihiro KOIBUCHI†

† National Institute of Informatics / JST
E-mail: †{ikki,koibuchi}@nii.ac.jp

Abstract　As the scale of many-core parallel applications and supercomputer systems increases, the negative impact of communication latencies on performance becomes larger. It is thus necessary to use low-latency topology based on high-radix switches in supercomputer systems. We have previously reported that topologies, which are generated by augmenting classical topologies with random links, make latency shorter, although their wiring length becomes longer than that of non-random topologies, such as hypercube. In this study, to shorten the wiring length of the random topologies, (1) we generate imbalanced random shortcut link set, (2) we solve the optimization problem through their modeling on rack layout. Our analysis results show that wiring length is reduced by up to 38% while communication latency is maintained.
Key words　Interconnection networks, random topology, cable length, cabinet layout, high-performance computing

1. はじめに

次世代の高性能計算システムにおけるメニーコア並列アプリケーションは，強／弱スケーリングを問わず，その多くが数百ナノ秒〜1 マイクロ秒程度の低 MPI 通信遅延を必要とすることが予測されている [1]．したがって，これらの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．ネットワーク遅延を要因別に見ると，スイッチの遅延が支配的であり(注1)，ケーブルの伝送遅延やホストの入出力遅延は相対的に小さい．ネットワーク低遅延化のためには経由スイッチ数を減らすこと，すなわち直径(注2)と平均距離(注3)の小さいトポロジを使うことが有効と考えられる．

現在では数十ポート以上の高次数スイッチ製品が普及し，その次数を活かしたネットワーク設計が可能である．高次数スイッチを前提とした規則的なネットワークトポロジは各種提案されており [2]，それらはネットワークの直径，スイッチの次数，レイアウトの自由度，ルーティングの容易性，耐故障性などの点でトレードオフを持つ．

これに対し，我々は不規則性を持つトポロジに着目している．すなわち，リングのような低次数のトポロジにランダムなショートカットリンクを加えたトポロジを考える．これを「ランダムトポロジ」と呼ぶ．我々のこれまでの研究 [3] では，ランダムトポロジが規則的なトポロジと比べて低遅延であることを定量的に示した．さらに，ネットワークの故障やメンテナンスの際，規則的なトポロジはそのトポロジを維持するために特別な機能や冗長性を必要とするのに対し，ランダムトポロジは故障箇所を迂回するようルーティングを更新することで多くの場合に対応できるという利点がある．

その反面，ランダムトポロジは規則的なトポロジと比べてラックレイアウトの配線長が大幅に増えるという欠点がある．初代地球シミュレータの配線長が 2,000km を大きく超え，京コンピュータも約 1,000km に達していることを考えると，施工性・メンテナンス性・省資源性の観点から，スーパーコンピュータの配線長を抑えることが今後重要となる可能性がある．

以上の背景を踏まえ，本研究では，ランダムトポロジのラックレイアウトの配線長削減に取り組む．具体的には，まず，ランダムトポロジの直径と平均距離を小さく保ったまま，ショートカットリンクに局所性を持たせる．次に，ランダムトポロジをクラスタリングしてラック間の配線数を減らす．最後に，ラックのフロ

(注1)：例えば Infiniband QDR スイッチ 1 台の遅延は約 100 ナノ秒である
(注2)：最も遠い 2 ノード間のホップ数
(注3)：すべての 2 ノード間の最短ホップ数の平均

アへのマッピングを最適化し，ラック間の配線延長が最小となるレイアウトを得る．

本研究の貢献は次の通りである．

（1） ランダムトポロジ生成時，全体の 50%より離れたノード間にはショートカットリンクを張らないようにしても，直径と平均距離はほとんど増加しないことを示した．(4. 章)

（2）（1）のランダムトポロジに基づくネットワークは，同じ次数のハイパーキューブと比べて，遅延が最大 20%小さいことが分かった．(4. 章)

（3）（1）のランダムトポロジに対し，ラック間の配線数を最大 15%削減した．(5. 章)

（4）（2）のラックレイアウトに対し，ラック間の配線延長を最大 38%削減した．(6. 章)

2. 関 連 研 究

ここでは，4 章に関連してスーパーコンピュータのネットワークトポロジを，5 章に関連してグラフ分析を，6 章に関連して施設配置問題を，それぞれ概観する．

2.1 高次元トポロジ

スーパーコンピュータのネットワークトポロジとして，トーラス，メッシュ，ハイパーキューブを含む k-ary n-cubes や，Fat tree が広く利用されてきた．これらは次元数や階層間リンク数を増やすことで高次元ネットワークへ拡張できる．

k-ary n-cubes の他にも各種の規則的な直接網が提案されており，直径と次数の点でトレードオフを持つ．例えば De Bruijn（3,072 ノードにおいて直径 12，次数 4），Kautz（同 11，4），Pradhan（同 12，5），スターグラフ（5,040 ノードにおいて同 7，6），パンケーキグラフなどである [2]．

スーパーコンピュータのネットワークは広帯域を必要とするため，ラック内程度の短いリンクには安価な電気ケーブルが使えるが，ラック間を結ぶ長いリンクには高価な光ケーブルを使わざるをえない．したがって，システムレイアウトがネットワークコストに大きく影響する．ドラゴンフライ網 [4] はこの点に着目し，トポロジをラック内とラック外の 2 階層に分け，複数のルータでひとつの仮想ルータを構成する．ドラゴンフライの各階層には，後述するランダムトポロジを含め，多様なトポロジを埋め込むことができる．

一方で我々は，ランダムなショートカットリンクがネットワークの直径と平均距離を劇的に小さくする現象に着目し，スーパーコンピュータのネットワークへの応用を探究している．これまでの研究 [3] において我々は，ランダムトポロジが同じ次数の規則的なトポロジに比べて低遅延であることを示した．また，スーパーコンピュータの高次元ネットワークの場合，乱数によるネットワーク性能のばらつきが十分小さいことを確かめた．

2.2 グラフ分析

ソーシャルネットワーキングサービスの普及にともない，未知の構造を持つ大規模グラフを分析して有用な知見を得る技術が近年，社会学やマーケティングの分野で急速に発達している．計算機ネットワークの分野でも，インターネットや Web ページ群を対象としたグラフ分析は広く行われている．しかし，こうしたグラフ分析技術を用いてネットワーク機器をラックへ格納する試みは，我々の知る限り行われていない．

2.3 施設配置問題

n 箇所に分散した工場間の物流コストが最小となるように立地を決める問題は施設配置問題と呼ばれ，オペレーションズリサーチの分野で 1960 年代から研究されてきた [5]．これは二次割り当て問題と呼ばれる NP 困難な組合せ最適化問題に帰着され，中規模以上の問題で厳密解を求めることは難しいが，メタヒューリスティクスを用いて実用的な近似解を求める方法が知られている．施設配置問題とその近似解法は集積回路の設計など幅広い産業応用を持つが，これをスーパーコンピュータのラック配置設計に応用しようとする研究は，我々の知る限り行われていない．

3. 前提と方針

本研究が対象とするスーパーコンピュータは，まとまった設置場所（建物内ないしフロア内）に新規に建設されるものとする．既存のシステムを拡張する場合にも本技術は適用可能だが，本報告では扱わない．システムの規模はおおむね 10,000 ノード程度までを想定し，数時間以内にラック配置の解が得られることを要件とする．ここで言う「ノード」は，実際にはスイッチと複数台のホストからなるサブシステムであってよい．本研究ではノードの内部構成を捨象し，ノード間の接続関係のみを扱う．

本報告における用語を次のとおり定義する．「リンク」はノード間の接続である．「次数」はノードが持つリンク数である（ノード内部の接続数を含まないことに注意せよ）．「ラックサイズ」は 1 台のラックに格納できる最大ノード数である．

以上の前提を踏まえ，3 つのステップ ——(1) ランダムトポロジの局所化，(2) クラスタリング，(3) マッピング ——からなる最適化の方針を定めた．この 3 つのステップの関係を図 1 に示す．左は 64 ノードからなる局所化されたランダムトポロジの例である．中央はこの 64 ノードをクラスタリングして 6 台のラックに分散格納した例である．線の太さがリンク数を示す．そして，右はこの 6 ラックをフロア上にマッピングした例である．線の太さが配線数を示す．これら一連のステップにより，ランダムトポロジの低遅延性を維持しつつ，ラック間の配線延長を最小化する．

4. ランダムトポロジの局所化

本章では，偏りを持ったランダムトポロジを生成し，ネットワーク性能を劣化させない偏り具合の許容範囲を定める．

4.1 考 え 方

我々が過去に提案したリングベースのランダムトポロジは，ショートカットリンクを張る際，近いノードも遠いノードも同じ確率で選んでいた．そうして作られたトポロジをネットワークとして実装するには長い物理配線が必要となる [3]．このようなランダムトポロジは一様性が高く特徴的な内部構造を持たないため，本報告を通じて明らかになるように，配線を短くしようとしても最適化の余地がほとんどない．最適化の効果を得て配線を短くするには，最適化の手がかりとなる何らかの内部構造を埋め込む必要がある．

そこで我々は，ランダムトポロジに局所性を持たせるアプローチを探究する．具体的には，ショートカットのリンク先を選ぶ際，リンク元から近いノードは選ばれやすく，遠いノードは選ばれに

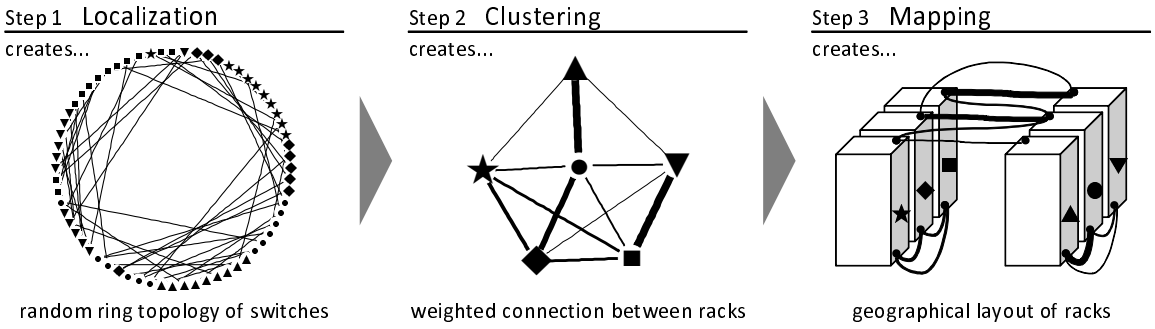

図 1　最適化の方針

くくなるように選択確率を調整する．この操作を本研究では「局所化」と呼ぶ．

局所化によって配線長を最適化する余地が生まれることが期待される反面，直径と平均距離が増加し，ネットワーク性能が劣化することも予想される．そこで，局所化の手法と程度を変えながらシミュレーションを行い，ネットワーク性能とのトレードオフを評価する．これにより，ランダムトポロジの低遅延性を維持したままで埋め込むことができる局所性の限界を探る．

4.2 手　　法

ランダムトポロジの生成手法として，局所性を持たない「一様ランダムリング」と，局所性を持つ「近傍ランダムリング」および「単峰ランダムリング」を定義する．いずれも n ノードからなるリング（環状）トポロジをベースとし，各ノードに $m-2$ 本のランダムショートカットリンクを追加することで m 次のランダムトポロジを生成する．リング上におけるノード i と j の距離を d_{ij} とすると，各生成手法がノード i からのショートカット先 j を選ぶ方法は次のとおりである．

一様ランダムリング　すべてのノードから均等に選ぶ．

近傍ランダムリング　与えられた分布範囲 θ に対し，$d_{ij} \leqq \theta/2n$ であるノードから均等に選ぶ．$\theta = 100\%$ のとき一様ランダムリングと等価である．θ より遠いノードへのショートカットは排除される．

単峰ランダムリング　与えられた標準偏差 σ の正規分布の確率密度関数 $y = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ に対し，$x = \frac{3d_{ij}}{2n}$ のときの値 y に比例する選択確率に従って選ぶ．$\sigma \to \infty$ のとき一様ランダムリングに漸近する．局所性を持たせつつ，遠いノードへのショートカットも排除しないことを意図している．

いずれの生成手法も，すでに i とリンクされているノードと，すでに次数 m に達しているノードは選ばない．

4.3 直径と平均距離

局所化によるグラフとしての性質の変化を知るため，直径と平均距離を調べた．トポロジ生成には我々が Ruby で記述したプログラムを用い，グラフ分析には R 言語と igraph ライブラリを用いた．

図 2 は，一様ランダムリング（Uniform）・近傍ランダムリング（Nbr(θ)）・単峰ランダムリング（Gau(σ)）の各手法で生成したトポロジの直径と平均距離を示す．比較のため同一ノード数・同一次数のハイパーキューブ（Hcube）も含めた．

この結果から，直径・平均距離ともに一様ランダムリングが最も小さく，トポロジの局所性が高まるにつれて大きくなっていくことが分かる．

	256ノード・8次		4,096ノード・12次	
	直径	平均距離	直径	平均距離
Uniform	4	2.89	5	3.65
Nbr(70%)	4	2.92	5	3.66
Nbr(50%)	5	3.03	5	3.69
Nbr(40%)	5	3.16	6	3.78
Gau(2.0)	4	2.89	5	3.65
Gau(1.0)	5	2.97	5	3.68
Gau(0.7)	5	3.11	6	3.78
Hcube	8	4.02	12	6.00

図 2　直径と平均距離

本技術は乱数を用いるため試行ごとに異なる結果を得るが，乱数によるネットワーク性能のばらつきは十分小さいことが分かっている [3] ため，本報告では以後，任意の 1 試行の結果を示すにとどめる(注4)．

4.4 ネットワーク性能の評価

局所化によるネットワーク性能への影響を知るため，遅延を評価した．評価には C++で記述されたフリットレベルシミュレータを用い，スイッチと point-to-point リンクで構成された相互結合網をモデルとした．スイッチの構造とシミュレーションパラメータは [3] に合わせた．

図 3 と図 4 はそれぞれ，ランダムに宛先を選択する Uniform トラフィックと，ソートや高速フーリエ変換を行うアプリケーションが含むシャッフル交換の転置パターンを考慮した合成トラフィックパターンである Matrix Transpose トラフィック [6] を，256 ノード・8 次のネットワークに注入した場合のシミュレーション結果である．縦軸はパケットが生成されてから宛先ホストに到達するまでの end-to-end の遅延を，横軸は各ホストの受信フリットレートである accepted traffic を示す．比較のためハイパーキューブの結果も含めた．なお，4,096 ノードのネットワークについては実行速度の制約からフリットレベルシミュレーションは行っていない．

一様ランダムリングとの比較では，近傍ランダムリングは分布範囲 50%以上，単峰ランダムリングは標準偏差 1.0 以上ならば，トラフィック負荷が飽和する前の遅延の増加を 3%以内に抑えることができた．以後，近傍ランダムリングは分布範囲 50%以上，単峰ランダムリングは標準偏差 1.0 以上を，本検討における許容範囲とする．

(注4)：本技術はスーパーコンピュータの設計段階で利用されるものであり，設計者であるユーザーは多数の試行の中から最善の結果を選ぶことができる．

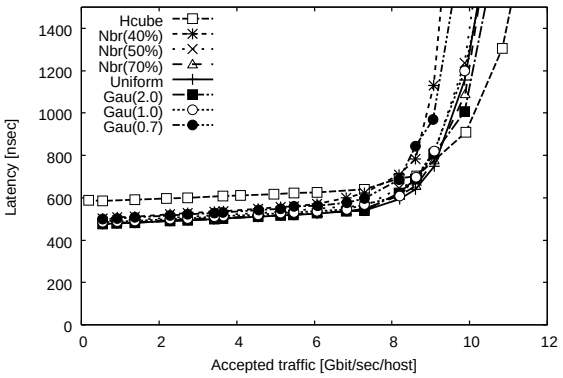

図 3　ネットワーク性能（256 ノード，Uniform トラフィック）

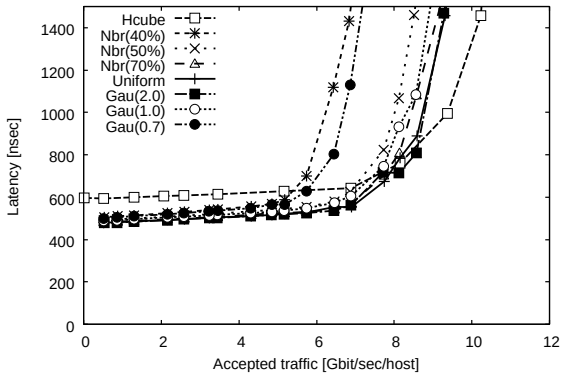

図 4　ネットワーク性能（256 ノード，Matrix Transpose トラフィック）

5. クラスタリング

本章では，前章で得たトポロジに対し，複数のノードを 1 台のラックにまとめることで，ラック間の接続関係を得る．

5.1 モデル化

トポロジはノードを頂点とする重みなし単純無向グラフと等価である．複数ノードを 1 ラックに格納する操作は，このグラフの頂点を縮約して，ラックを頂点とする重み付き単純無向グラフに変換することと等価である（縮約時に生じたループ辺は除去し，多重辺はその本数を辺の重みに変換する）．このとき，ノード同士の結びつきが密な部分をなるべく同じラック内に収める．このような操作はクラスタリングと呼ばれる．

5.2 手　　法

本研究では，データマイニング分野で実績のある階層的クラスタリング手法に基づき，ユーザーから与えられたラックサイズの要件を満たすよう独自に改良した手法を用いてクラスタリングを行う．階層的クラスタリング手法には凝集型と分割型がある．前者は全クラスタが各々ひとつの頂点を含む状態を初期状態とし，クラスタ対を再帰的に併合していく．後者は逆に，クラスタを再帰的に分割していく．本報告の評価では，下記 3 種類の手法を用いた(注5)．

Ward 法　[8] 凝集型である．クラスタ重心からクラスタに含まれる各頂点までの距離の二乗和の増分が最小となるクラスタ対を併合する．

Walktrap 法　[9] 凝集型である．Ward 法と同じ基準でクラスタ対を併合するが，頂点 i と j の「距離」として $d_{ij} = \sqrt{\sum_{k=1}^{n}(P_{ik}^t - P_{jk}^t)^2/\deg(k)}$ を用いる点が異なる．ここで P_{ik}^t は頂点 i から長さ t のランダムウォークを行って頂点 k に到達する確率，$\deg(k)$ は頂点 k の次数である．

Girvan-Newman 法（GN 法）　[10] 分割型である．辺媒介性（すべての 2 頂点間の最短経路のうち，その辺を経由する最短経路の数）の高い辺を順に除去することでクラスタを分割する．

比較のため，直観的手法である**等分割法**によるクラスタリングも行った．等分割法は元のトポロジのベースとなったリングに沿って r ノードずつ順番にクラスタ化していく（r はラックサイズ）．

(注5)：我々はこのほか，最短距離法，最長距離法，平均距離法，Newman 法 [7] も試行したが，上記 3 種類のいずれかと同じ傾向を示すか，もしくは削減率が劣る結果となったため，本報告では省略する．

5.3 クラスタサイズの調整

上述した階層的手法はクラスタリング結果として樹状構造を得る．データマイニング用途では，ユーザーは望ましいクラスタ数を得る高さで樹を水平に切るが，クラスタサイズは不定となる．一方，本研究の用途ではクラスタサイズがラックサイズを超えてはならない．そこで，本研究では次の手順でラックサイズを超えないクラスタを生成した．

（1） 樹状構造の葉であるクラスタのうち，根から最も遠いクラスタを 2 つ取り出す．

（2） 2 つのクラスタサイズの和がラックサイズ以下ならば併合する．ラックサイズを超えるならば併合せず，大きい方のクラスタを樹状構造から切り離して単独のクラスタとする．

（3） 樹状構造の根に達するまで，手順（1）〜（2）を繰り返す．

5.4 リンク数の評価

図 5 と図 6 はそれぞれ，256 ノード・8 次と 4,096 ラック・12 次のトポロジを，Ward 法（ward）・Walktrap 法（walktrap）・GN 法（girvan）・等分割法（naive）の各手法でクラスタリングした結果のリンク数の削減率を示す．比較のためハイパーキューブの結果も含めた．ラックサイズは 16 である．なお，ラック内の配線量はクラスタリングの巧拙にかかわらず一定と考え，本評価では無視する．

256 ノードについて見ると，一様ランダムリングを等分割した場合（直観的手法）を基準として，近傍ランダムリング（分布範囲 50%）を GN 法でクラスタリングした場合に 14%，単峰ランダムリング（標準偏差 1.0）は同 15%，それぞれリンク数が減少した．4,096 ノードについて見ると，Walktrap 法でクラスタリングした場合のリンク数の減少率はそれぞれ 4.4%，4.5%だった．

以上の結果から，ランダムトポロジの局所化がリンク数の削減に寄与することと，適切なクラスタリングがリンク数をさらに削減することが確かめられた．トポロジの局所性が高いほど削減率も高いが，ネットワーク性能とのトレードオフを考慮する必要がある．前章で定めた許容範囲内では，トポロジ生成手法による削減率の差は僅かであった．

クラスタリング手法の比較では，Walktrap 法が Ward 法を削減率で常に上回った．GN 法は Walktrap 法に匹敵するものの，計算量が頂点数の 3 乗オーダーと大きく，4,096 ノードのクラスタリングを数時間以内に終えることができなかったため，要件を満たさないものとして除外する．したがって，以後の検討ではク

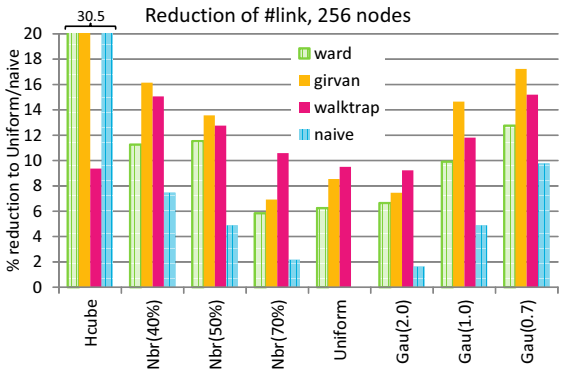

図 5　クラスタリング後のリンク数
（直観的手法に対する削減率，256 ノード）

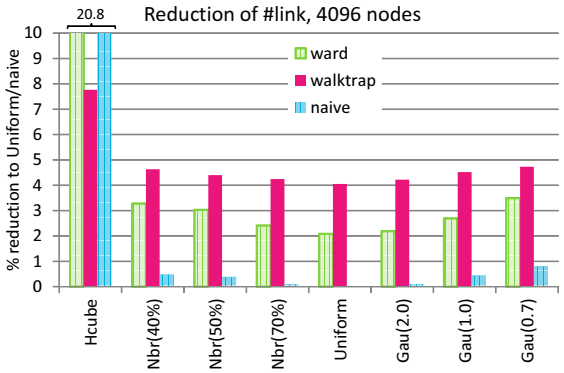

図 6　クラスタリング後のリンク数
（直観的手法に対する削減率，4,096 ノード）

ラスタリング手法として Walktrap 法を用いる．

6. マッピング

本章では，前章で得たラック間の接続関係と，ユーザーから与えられた設置場所の情報に基づき，フロア上のラックレイアウトを得る．

6.1 モデル化

ユーザーから与えられた地図（ラックを設置できる座標のリスト）上の各地点に対し，クラスタリングによって得られた重みつきグラフ（辺の重みがラック間の配線数を表す）の各頂点を割り当てる．このとき，各地点間の距離と辺の重みとの積の総和を最小化することで，ラック間の配線延長が最短となるマッピングを得る．このようなマッピングは二次割り当て問題として定式化される．

いま，ラック数・地点数を n，地点 i と j の距離を d_{ij}，ラック i と j を結ぶ配線数を w_{ij} とする．二次割り当て問題を解くには

$$\text{Minimize } \Sigma_{i=1}^{n}\Sigma_{j=1}^{n} w_{ij} d_{\phi(i)\phi(j)} \qquad (1)$$

なる順列 $\Phi = \phi(1), \ldots, \phi(n)$ を求めればよい．ここで $\phi(i)$ はラック i が割り当てられた地点の番号である．

6.2 解　　法

本研究では，二次割り当て問題に対して適用実績のあるメタヒューリスティクスである Simulated Annealing **法（SA 法）** [11] を用いてマッピングを最適化する(注6)．

比較のため，直観的手法である**ジグザグ法**によるマッピングも行った．ジグザグ法は地図上の各地点が格子状に並んでいることを仮定し，1 列目は左から右へ，2 列目は右から左へ，以下同様にグラフの頂点を定義順にマッピングしていく．グラフが単純なリングである場合に妥当な解が得られる．

6.3 地図の生成

本モデルは各地点相互間の距離さえ定義されれば成立するから，本技術はユーザーが与える任意の地図に対応できる．本報告の評価では，なるべく正方形に近い長方形のフロアにすべてのラックを並べることとし，ラック数に応じた地図を以下の方法で生成した．

いま，ラック数を n，1 ラックの占有寸法を幅 x [cm] ×奥行 y [cm] とする（奥行は通路幅を含む）．ラックを q 列に並べるとき，1 列あたりのラック数は $p = \lceil n/q \rceil$，フロア面積は $a = pxqy$ で表される．ここで，q の候補として $q_1 = \lceil \sqrt{nxy}/y \rceil$，$q_2 = \lfloor \sqrt{nxy}/y \rfloor$ の 2 通りを考え，a がより小さくなる方を q とする（ただし $q > 0$）．そして，フロアの寸法を横 px [cm] ×縦 qy [cm] とし，隅から横方向へ p 台ずつラックを並べる座標リストを，評価に用いる地図とする．

6.4 配線延長の評価

図 7 と図 8 はそれぞれ，256 ノード・8 次と 4,096 ラック・12 次のトポロジを，walktrap 法（walktrap）と等分割法（naive）でクラスタリングした後，SA 法（opt）とジグザグ法（zigzag）でマッピングした結果の配線延長の削減率を示す．比較のためハイパーキューブの結果も含めた．ラックの寸法は幅 60 [cm] ×奥行 180 [cm]，地図上の各地点間の距離はマンハッタン距離である．SA 法の反復数は 1 億回，試行数は 10 回とした．なお，ラック内の配線延長はマッピングの巧拙にかかわらず一定であるため，本評価では無視する．

4,960 ノードについて見ると，一様ランダムリングを等分割してジグザグ法でマッピングした場合（直観的手法）を基準として，近傍ランダムリング（分布範囲 50%）を等分割して SA 法でマッピングした場合に 38%，単峰ランダムリング（標準偏差 1.0）は同 36%，それぞれ配線延長が減少した．また，ハイパーキューブと比べるとランダムリングの配線延長は増加するが，ハイパーキューブを等分割して SA 法でマッピングした場合を基準として，増加率はそれぞれ 94%，99%（約 2 倍）に抑えられた．これは直観的手法の増加率 212%（約 3.1 倍）に対して大幅に減少したと言える．

256 ノードではラック数が 16〜21 と少なく，乱数のばらつきによる結果の揺らぎが無視できないため定量的な議論は難しい．10 通りの乱数を試行したところ，定性的には，局所性が低いときは等分割以外のクラスタリング手法の効果が高い傾向が見られた．例えば図 7 では，一様ランダムリングと単峰ランダムリング（標準偏差 2.0〜1.0）を walktrap 法でクラスタリングすると，等分割と比べてラック数が増加したにもかかわらず，SA 法でマッピングした結果の配線延長は減少した．いずれにせよ，ラック数が少ない場合は多数の試行の中から最善の結果を選ぶことが必要と言える．

(注6)：我々はこのほか，Robust Taboo Search [12]，Fant [13]，GRASP [14] の各手法も試行したが，いずれも SA 法とほぼ同じ最適解が得られたため，本報告では省略する．

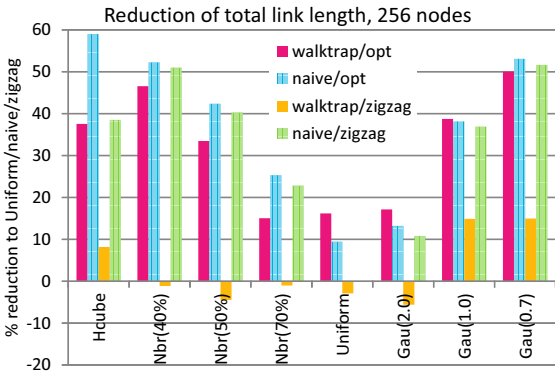

図 7　マッピング後の配線延長
（直観的手法に対する削減率，256 ノード）

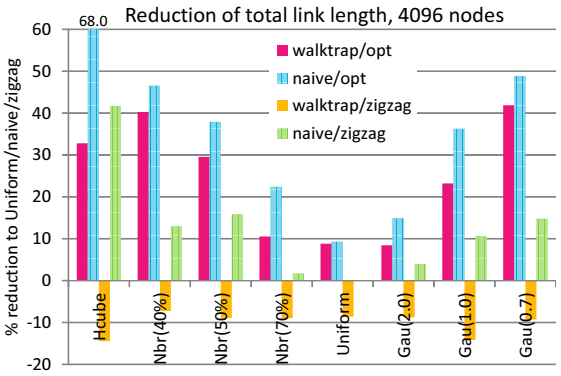

図 8　マッピング後の配線延長
（直観的手法に対する削減率，4,096 ノード）

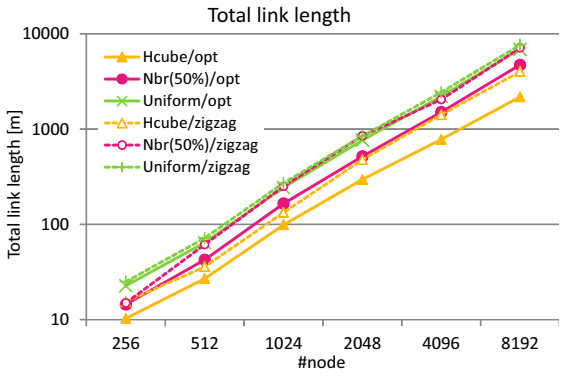

図 9　マッピング後の配線延長（ノード数を変化させたとき）

以上の結果から，ランダムトポロジの局所化とマッピングの最適化は，ともに配線延長の削減に寄与することが確かめられた．また，等分割以外のクラスタリング手法はラック数の増加を伴い，それを相殺してなお配線延長の削減に寄与するか否かはケースバイケースであることが分かった．

6.5　スケーラビリティの評価

図 9 は，ノード数を $n = 256$ から $n = 8192$ まで変化させたときの配線延長を示す（次数は $m = \log_2 n$，ほかの条件は前節と同じ）．この結果から，ジグザグ法でマッピングした場合はノード数が増えると局所化の効果が小さくなるのに対し，SA 法でマッピングを最適化した場合はノード数が増えても局所化の効果が保たれ，近傍ランダムリング（分布範囲 50%）で 38%前後の削減率が安定的に達成できることが分かった．

7.　おわりに

本研究では，ランダムなネットワークトポロジの低遅延性を維持したまま，ラックレイアウトの配線長を削減した．具体的には，(1) ランダムなショートカットリンクの張り方に局所性を持たせ，(2) クラスタリングによりラック間リンク数を減らし，(3) ラックレイアウトを最適化することによりラック間の配線延長を最小化した．ケーススタディの結果，直観的な手法による場合と比べて，4,096 ノードでラック間の配線延長を最大 38%，256 ノードでラック間の配線数を最大 15%，それぞれ削減できることを示した．今後の課題として，ラック数の増加を抑えるクラスタリング手法の開発が考えられる．

謝 辞　本研究の一部は，科学技術振興機構「JST」の戦略的創造研究推進事業「CREST」の支援による．

文　献

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

[2] 天野英晴: 並列コンピュータ, 昭晃堂 (1996).

[3] Koibuchi, M., Matsutani, H., Amano, H., Hsu, D. F. and Casanova, H.: A Case for Random Shortcut Topologies for HPC Interconnects, *Proc. of the International Symposium on Computer Architecture (ISCA)*, pp. 177–188 (2012).

[4] Kim, J., Dally, W. J., Scott, S. and Abts, D.: Technology-Driven, Highly-Scalable Dragonfly Topology, *Proc. of the International Symposium on Computer Architecture (ISCA)*, pp. 77–88 (2008).

[5] Pitu B. Mirchandani and Richard L. Francis(eds.): *Discrete Location Theory*, Wiley-Interscience (1990).

[6] Dally, W. D. and Towles, B.: *Principles and Practices of Interconnection Networks*, Morgan Kaufmann (2003).

[7] Clauset, A., Newman, M. and Moore, C.: Finding community structure in very large networks, *Physical Review E*, Vol. 70, No. 6 (2004).

[8] Joe H. Ward Jr.: Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, Vol. 58, No. 301, pp. 236–244 (1963).

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

[10] Newman, M. and Girvan, M.: Finding and evaluating community structure in networks, *Physical Review E*, Vol. 69, No. 2 (2004).

[11] Connolly, D. T.: An improved annealing scheme for the QAP, *European Journal of Operational Research*, Vol. 46, No. 1, pp. 93–100 (1990).

[12] Taillard, E. D.: Robust taboo search for the quadratic assignment problem, *Parallel Computing*, Vol. 17, No. 4-5, pp. 443–455 (1991).

[13] Taillard, E. D.: FANT: Fast ant system, Technical report, Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale (1998).

[14] Li, Y., Pardalos, P. M. and Resende, M. G.: A greedy randomized adaptive search procedure for the quadratic assignment problem, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Vol. 16, pp. 237–261 (1994).