論　文

ランダムなネットワークトポロジーのラック配置最適化に関する研究

藤原　一毅[†a)]　　鯉渕　道紘[†b)]

Study on Rack Layout Optimization for Random Network Topology

Ikki FUJIWARA[†a)] and Michihiro KOIBUCHI[†b)]

あらまし　スーパコンピュータの大規模化が進むにつれ，通信遅延が並列アプリケーションの性能に及ぼす影響が無視できなくなっている．このため，高次数スイッチを前提とした低遅延なネットワークトポロジーの活用が注目されている．これまでの研究で我々は，ランダムなショートカットリンクをもつトポロジーが低遅延性の点で優れていることを明らかにした．このようなランダムトポロジーは，一方で，ハイパキューブなどの規則的なトポロジーに比べて配線が非常に長くなるという問題を抱えていた．本研究において我々は，(1) 不均一なランダムショートカットリンクを単純なトポロジーに付加し，(2) ラックレイアウトを施設配置問題として最適化することで，低遅延性を維持したまま総配線長を最大 29%削減できることを示す．

キーワード　相互結合網，ランダムトポロジー，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

1. まえがき

次世代の高性能計算システム上で実行される並列アプリケーションは，数百ナノ秒〜1 マイクロ秒程度の非常に短い MPI 通信遅延を必要とするものが増えると予測されている [1]．これらの並列アプリケーションを効率的に実行するために，超低遅延ネットワークの研究開発が重要な課題となりつつある．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（最も遠い 2 ノード間のホップ数）と平均距離（全ての 2 ノード間の最短ホップ数の平均）の小さいトポロジー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，その次数を生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ー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，ルーチングの容易性，耐故障性などの点でトレードオフをもつ．

これに対し，我々は不規則性をもつトポロジーに着目している．すなわち，リングのような低次数のトポロジーにランダムなショートカットリンクを加えたトポロジーを考える．これを「ランダムトポロジー」と呼ぶ．我々のこれまでの研究 [3] では，ランダムトポロジーが規則的なトポロジーに比べて低遅延であることを定量的に示した．更に，ネットワークの故障やメンテナンスの際，規則的なトポロジーはそのトポロジー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，ランダムトポロジーは故障箇所を迂回するようルーチ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，ランダムトポロジーは規則的なトポロジーに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 2,000 km を大きく超え，京コンピュータも約 1,000 km 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．

† 国立情報学研究所/JST，東京都
National Institute of Informatics/JST, 2–1–2 Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430 Japan
a) E-mail: ikki@nii.ac.jp
b) E-mail: koibuchi@nii.ac.jp

(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．

以上の背景を踏まえ，本研究では，性能（遅延，スループット）は維持したままランダムトポロジー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，ランダムトポロジー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もたせる．次に，ランダムトポロジー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，ラックのフロアへのマッピングを最適化し，総配線長が小さくなるレイアウトを得る．

本研究の貢献は次のとおりである．

（1） ランダムトポロジー生成時，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，直径と平均距離がほとんど増加しないことを示した．(**4.**)

（2）（1）のランダムトポロジーに基づくネットワークは，同じ次数のハイパキューブに比べて遅延が最大 20%小さいことを，サイクルアキュレートシミュレーションにより示した．(**4.**)

（3）（1）のランダムトポロジーについて，ラック間の配線数を最大 15%削減した．(**5.**)

（4）（2）のラックレイアウトについて，総配線長を最大 29%削減した．(**6.**)

2. 関連研究

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

2.1 高次元トポロジー

スーパコンピュータのネットワークトポロジーとして，トーラス，メッシュ，ハイパキューブを含む 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もつ．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得ない．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，トポロジーをラック内とラック外の 2 階層に分け，複数のルータで一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ーを含め，多様なトポロジー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，ランダムトポロジーが同じ次数の規則的なトポロジーに比べて低遅延であることを示した．また，スーパ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

近年，ソーシャルネットワーキングサービスの普及に伴い，大規模グラフの構造から有用な知見を抽出するグラフ分析技術が急速に発達している．こうしたグラフ分析技術は，人間関係ネットワークの分析だけでなく，インターネットや web ページ群といったコンピュータネットワークの分析にも用いられる [5]．

前述のような高次元トポロジーをスーパコンピュータに実装する場合，どのスイッチをどのラックに割り当てるべきか，そのグルーピングの方法が必ずしも明確でない．本研究ではグラフ分析技術を用いてスイッチのグルーピングを行う．

2.3 施設配置問題

複数の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 [6]．これは二次割当問題と呼ばれる NP 困難な組合せ最適化問題に帰着され，厳密解を求めることは極めて難しいが(注2)，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．コンピュータ分野でも，チップ上の配置配線を施設配置問題としてモデル化することは広く行われている．

スーパコンピュータのラックレベルの設計をシステマティックに行うために，上述のグラフ分析技術や施設配置最適化を応用しようとする研究は，我々の知る限り，我々のほかには行われていない [8]．

(注2)：2011 年 5 月現在，128 要素の二次割当問題が，厳密解が知られている最大規模とされている [7]．

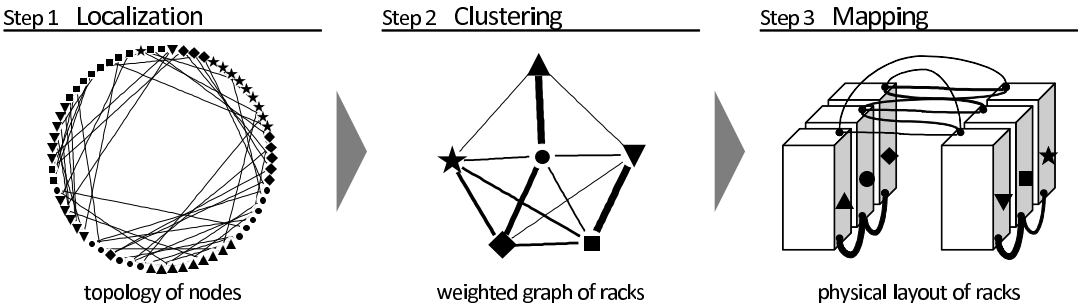


図 1　最適化の方針（記号はノードとラックの対応関係を，線の太さはリンク数を示す）
Fig. 1　Approach for rack layout optimization. The thickness of a line indicates the number of links.

3. 前提と方針

本研究は，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スーパコンピュータを対象とする．本研究の手法は既存のシステムを拡張する場合にも適用可能だが，本論文では扱わない．本論文において「ノード」とは，スイッチと複数台のホストからなる（例えば，48 ポートのスイッチに 16 台のホストが接続され，残り 32 ポートが空いていて他のスイッチに接続できる）サブシステムを意味する．本論文では以後，ノードをネットワーク構成の最小単位とみなし，ノード間の接続関係のみを扱う．そのほか，本論文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．「配線」はリンクの物理的な実体である．「次数」はノードがもつリンク数である（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 ステップからなる最適化の方針を定める．すなわち (1) ランダムトポロジーの局所化，(2) クラスタリング，(3) マッピングである．図 1 にその概略を示す．ステップ 1 では，不均一なランダムトポロジーを生成する（同図左：64 ノードの例）．ステップ 2 では，これをクラスタリングしてラックを頂点とする重み付きグラフを得る（同図中央：6 ラックに格納された例）．ステップ 3 では，これをフロア上にマッピングして物理的なラックレイアウトを得る（同図右：2×3 の格子状配置の例）．

以下 **4.** ～ **6.** においてステップ 1～3 の詳細をそれぞれ述べる．

4. ランダムトポロジーの局所化

本章では，不均一なランダムトポロジーを生成し，ネットワーク性能を劣化させない不均一性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ーは，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ー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ーは一様性が高く特徴的な内部構造をもたないため，本研究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，最適化の手掛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，ランダムトポロジーに局所性をも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，ランダムトポロジーの低遅延性を維持したまま埋め込むことができる局所性の限界を探る．

4.2 手　　法

ランダムトポロジーの生成手法として，局所性をもたない「一様ランダムリング」と，局所性をもつ「近傍ランダムリング」及び「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）トポロジーをベースとし，各ノードに $m-2$ 本のランダ

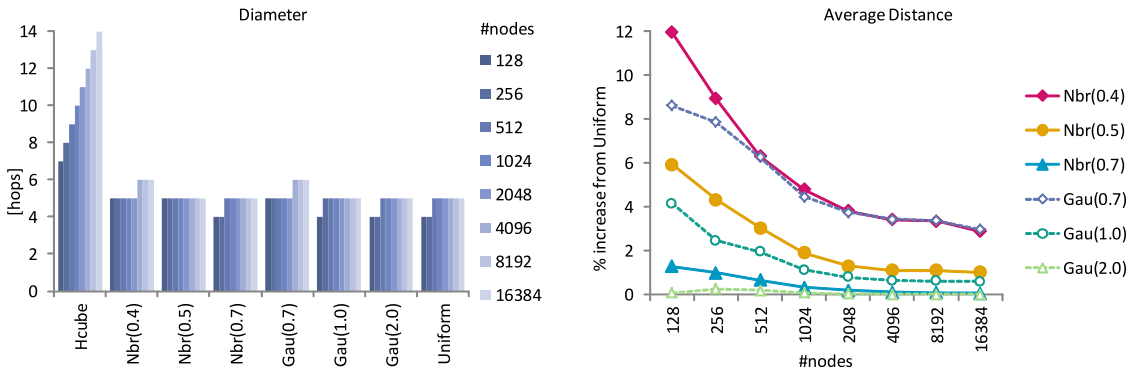


図 2　直径（左）と平均距離（右，一様ランダムリングに対する増加率）
Fig. 2　Diameter (left) and average distance (right).

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ーを生成する．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　全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，$d_{ij} \leq \theta n/2$ であるノードから均等に選ぶ．θ が小さいほど局所性が高く，$\theta = 1.0$ のとき一様ランダムリングと等価である．$\theta n/2$ 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，$x = 6d_{ij}/n$ のときの値 y に比例する選択確率に従って選ぶ．これにより，$\sigma = 1$ のとき，標準正規分布の $\pm 3\sigma$ の範囲の値が選択確率としてリング上の各ノードに当てはめられる．σ が小さいほど局所性が高く，$\sigma \to \infty$ のとき一様ランダムリングに漸近する．局所性をもたせつつ，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，既に接続されているノード対，及び，既に次数 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．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ーについて，その直径と，一様ランダムリングに対する平均距離の増加率を示す．ここでは，乱数種を変えて生成した 20 通りのランダムトポロジーの中から直径が最小のものを選び，その中から更に平均距離が最小のものを選ぶ(注3)．ノード数は $128 \leq n \leq 16384$，次数は $m = \log_2 n$ である．比較のため同一ノード数・同一次数のハイパキューブ（Hcube）も含めた．

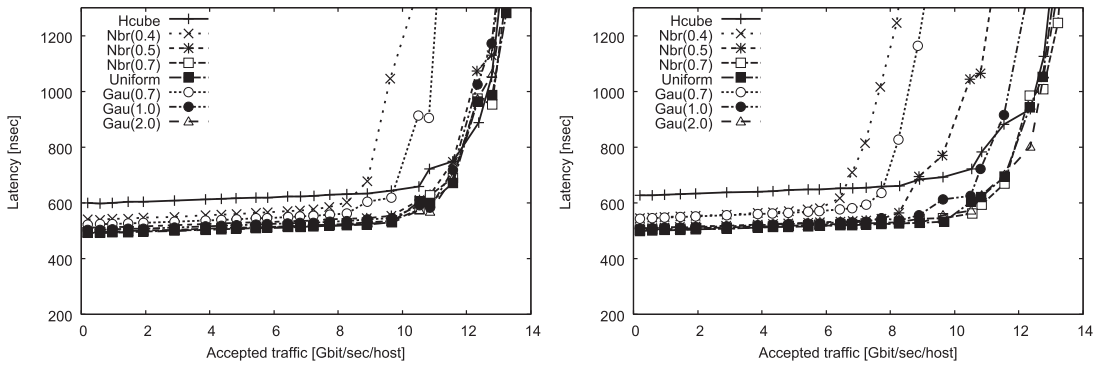
直径について見ると，近傍ランダムリングは $\theta = 0.5$ 以上，単峰ランダムリングは $\sigma = 1.0$ 以上ならば，16384 ノード・14 次のネットワークを一様ランダムリングと同じ直径 5 ホップで連結できた．更に局所性を高めた $\theta = 0.4$ の近傍ランダムリングや $\sigma = 0.7$ の単峰ランダムリングでは，4,096 ノード以上のネットワークで直径 6 ホップとなり，ランダムトポロジーの低遅延性が損なわれることが分かった．平均距離について見ると，4,096 ノードのネットワークにおいて，$\theta = 0.5$ の近傍ランダムリングでは一様ランダムリング比 1.1%増，$\sigma = 1.0$ の単峰ランダムリングは同 0.6%増であるのに対し，$\theta = 0.4$ の近傍ランダムリングと $\sigma = 0.7$ の単峰ランダムリングではいずれも同 3.4%増となった．

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．

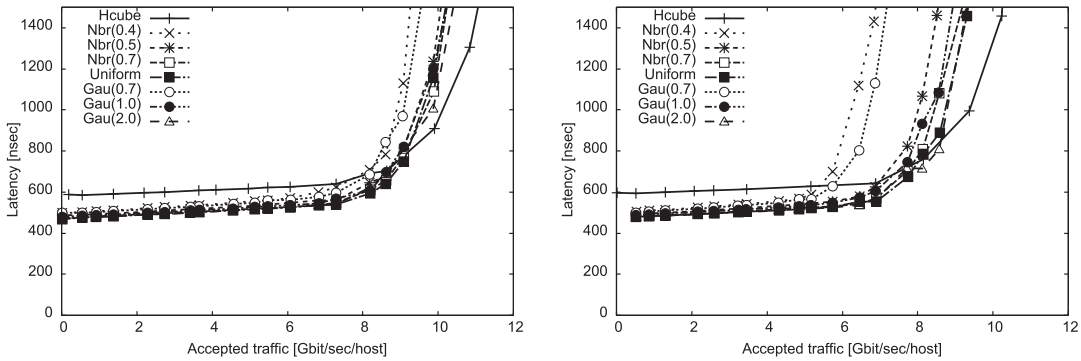
図 3 と図 4 は，ランダムに宛先を選択する Uniform トラヒックと，ソートや高速フーリエ変換を行うア

(注3)：本技術はスーパ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は多数の試行の中から最善の結果を選ぶことができる．



(a) Uniform Traffic　(b) Matrix Transpose Traffic

図 3　スループットと通信遅延（128 ノード）
Fig. 3　Throughput and communication latency (128 nodes).



(a) Uniform Traffic　(b) Matrix Transpose Traffic

図 4　スループットと通信遅延（256 ノード）
Fig. 4　Throughput and communication latency (256 nodes).

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 Matrix Transpose トラヒック [9] を，128 ノード・7 次と 256 ノード・8 次のネットワークにそれぞれ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．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，本論文では以後，任意の乱数種を用いて生成したランダムトポロジーを評価対象とする．比較のためハイパキューブの評価結果も含めた．なお，シミュレータの実行速度の制約から，4,096 ノードのネットワーク性能の評価は省略する．

シミュレーション結果を見ると，近傍ランダムリングは $\theta = 0.5$ 以上，単峰ランダムリングは $\sigma = 1.0$ 以上ならば，トラヒックが飽和しない領域における遅延が，一様ランダムリング比 3%増以内に収まった．他方，$\theta = 0.4$ の近傍ランダムリングや $\sigma = 0.7$ の単峰ランダムリングでは，遅延が増加するだけでなく，トラヒックが飽和する accepted traffic が小さくなり，ネットワーク性能が大きく劣化することが分かった．前節の結果も踏まえると，近傍ランダムリングは分布範囲 $\theta = 0.5$ 以上，単峰ランダムリングは標準偏差 $\sigma = 1.0$ 以上が，ネットワーク性能を維持するための許容範囲と考えられる．

5. クラスタリング

本章では，前章で得たトポロジーに対し，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

トポロジー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，ノード同士の結び付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一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．本論文の評価では，下記 3 種類の手法を用いた(注4)．

Ward 法　[11] 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　[12] 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　[13] 分割型である．辺媒介性（全ての 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，ノードを番号順にラックサイズ r 個ずつクラスタ化していく**等分割法**によるクラスタリングも行った．番号順とは，ランダムリングであれば元のリングに沿った順序，ハイパキューブであれば下位次元から数え上げていく順序である．

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図 5 のような樹状構造を得る．データマイニング用途では，ユーザ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（各クラスタに含まれるノード数）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二つ取り出す．

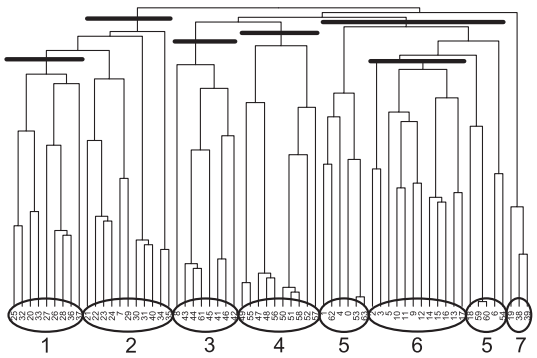


図 5　クラスタリングによって得られる樹状構造の例（Girvan-Newman 法）
Fig. 5　An example of dendrogram by Girvan-Newman clustering.

（2）　二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(注5)に達するまで，手順（1）～（2）を繰り返す．

図 5 において，太い水平線は手順（2）で切り離された枝を示し，楕円は最終的に生成されたクラスタを示す．本手順はクラスタサイズがラックサイズを下回る（ラック内に空きスペースが生じる）ことを許容するため，所要ラック数は必ずしも最小限とならない．なお，等分割法ではクラスタサイズがラックサイズと一致し，所要ラック数は最小限となる．

5.4 リンク数の評価

各クラスタリング手法の効果を測るため，リンク数を評価した．評価には R 言語と igraph ライブラリ [14] を用いた．図 6 は，256 ノード・8 次と 4,096 ノード・12 次のトポロジーを，Ward 法（Ward）・Girvan-Newman 法（Girvan）・Walktrap 法（Walk）・等分割法（Seq）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．一様ランダムリングを等分割法でクラスタリングした場合を削減率の基準とする．比較のためハイパキューブの結果も含める．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

(注4)：我々はこのほか，最短距離法，最長距離法，平均距離法，Newman 法 [10] 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，若しくは削減率が劣る結果となったため，本論文では省略する．
(注5)：切り離された枝を除く全ての頂点が一つのクラスタに含まれる状態．

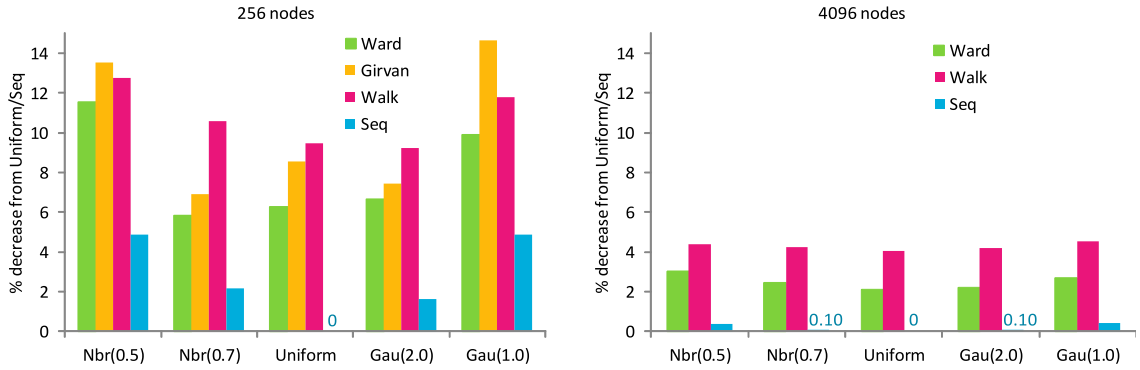


図 6　クラスタリング後のリンク数（一様ランダムリングを等分割した場合に対する削減率）
Fig. 6　Number of inter-rack links after clustering.

では無視する．

256 ノードについて見ると，Girvan-Newman 法でクラスタリングした場合，$\theta = 0.5$ の近傍ランダムリングは 14%，$\sigma = 1.0$ の単峰ランダムリングは同 15%，それぞれリンク数が削減された．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削減率はそれぞれ 4.4%，4.5%だった．局所化された全てのトポロジーにおいて等分割法により正の削減率が得られ，かつ，Ward 法・Girvan-Newman 法・Walktrap 法による削減率は等分割法による削減率を上回った．

以上の結果から，ランダムトポロジー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更に削減することが確かめられた．トポロジー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．

各クラスタリング手法を比較すると，Walktrap 法が Ward 法を削減率で常に上回った．Girvan-Newman 法は削減率で Walktrap 法に匹敵するものの計算量が $O(nl^2)$ と大きく（n は頂点数，l は辺数），数千ノード規模のスーパコンピュータ設計への適用は困難である(注6)．したがって，以後の検討ではクラスタリング手法として Walktrap 法を用いる(注7)．

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付きグラフ（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当問題として定式化される．

今，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．二次割当問題を解くには

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

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，二次割当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）[15] を用いてマッピングを最適化する(注8)．

比較のため，以下に述べるベースライン法によるマッピングも行った．地図上の各地点が格子状に並んでいることを仮定し，グラフの頂点を (1) 全ての列で左から右へ順番に割り当てていく方法と，(2) 1 列目は左から右へ・2 列目は右から左へ・以下同様にジグザグに割り当てていく方法の 2 通りを考える．前者は

(注6)：Girvan-Newman 法の計算時間は，Intel Xeon X5690（3.47 GHz）搭載の計算機を用いて，1,024 頂点・5,120 辺で約 10 分，2,048 頂点・11,264 辺で約 2 時間，4,096 頂点・24,576 辺で約 21 時間だった．
(注7)：Walktrap 法の計算時間は，8,192 頂点・53,248 辺で 22 秒だった．
(注8)：我々はこのほか，Robust Taboo Search [16]，Fant [17]，GRASP [18] 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，本論文では省略する．

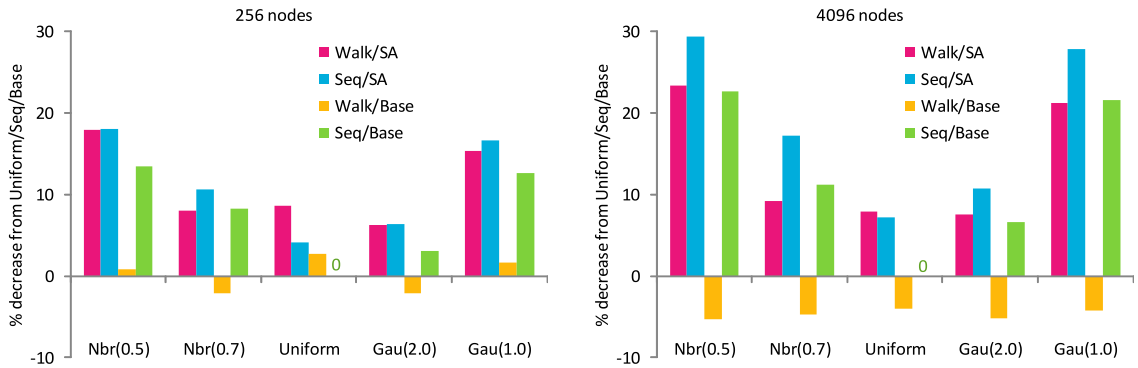


図 7　マッピング後の総配線長（単純実装に対する削減率）
Fig. 7　Total cable length after mapping.

グラフが格子状構造をもつとき地図に合致した解が得られる可能性があり，後者はグラフが環状構造をもつとき妥当な解が得られると考えられる．ベースライン法は両者を試行し，総配線長がより短くなった方を採用する．

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，本技術はユーザが与える任意の地図に対応できる．本評価では，一つのサーバルーム内にラックを縦横同数の格子状に並べると仮定した．具体的には，ラック数が n のとき，列数が $q = \lceil \sqrt{n} \rceil$，1 列当りのラック数が $p = \lceil n/q \rceil$ であるような地図を生成して評価に用いた．

6.4　総配線長の評価

図 7 は，256 ノード・8 次と 4,096 ノード・12 次のトポロジーを，Walktrap 法（Walk）と等分割法（Seq）でクラスタリングした後，SA 法（SA）とベースライン法（Base）でマッピングした結果の総配線長の削減率を示す．一様ランダムリングを等分割法でクラスタリングしてベースライン法でマッピングした場合を「単純実装」と呼び，これを削減率の基準とする．比較のためハイパキューブの結果も含める．ラックの設置間隔は幅 60 cm × 奥行 210 cm（奥行は通路幅を含む）[19]，ラック間にまたがる配線の長さは地図上の両地点間のマンハッタン距離に立上り・立下り各 2 m（計 4 m）を加えた長さ，ラック内で完結する配線の長さは一律 2 m とする(注9)．SA 法の反復数は 1 億回，試行数は 10 回である(注10)．

4,096 ノードについて見ると，$\theta = 0.5$ の近傍ランダムリングを等分割して SA 法でマッピングした場合は 29%（385 km → 272 km），$\sigma = 1.0$ の単峰ランダムリングは同 28%（385 km → 278 km），それぞれ単純実装（385 km）に比べて総配線長が減少した．ハイパキューブ（165 km）と比較すると，総配線長の増加率はそれぞれ 65%，69%（約 1.7 倍）であり，単純実装の増加率 117%（約 2.2 倍）に対して大幅に減少したといえる．

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．例として，図 7 左で最も局所性の低い一様ランダムリングについて見ると，等分割法では 16 ラック・5.89 km だったものが Walktrap 法では 21 ラック・5.61 km となり，ラック数が増加したにもかかわらず，SA 法によるマッピング後の総配線長は減少した．

以上の結果から，ランダムトポロジーの局所化とマッピングの最適化は，共に総配線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総配線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．

7.　議　　論

7.1　スケーラビリティ

本技術はネットワークの規模が大きくなるほど高い

(注9)：本評価はラック内の配線長を加味する点で我々の先行報告 [8] とは異なる．また，ラック設置間隔も先行報告とは異なっている．
(注10)：SA 法の計算時間は，Intel Xeon X5690（3.47 GHz）搭載の計算機を用いて，316 頂点・14,635 辺で約 100 分だった．頂点はラックを，辺はラック間リンクを表す．

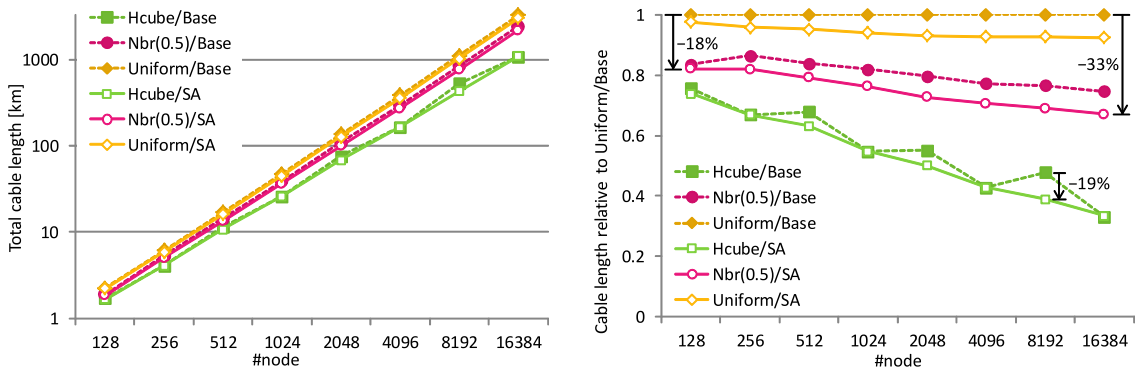


図 8　マッピング後の総配線長（ノード数を変化させたとき）
Fig. 8　Total cable length vs. number of nodes.

配線長削減効果を発揮する．

図 8 は，ノード数を $n = 128$ から $n = 16384$ まで変化させたときの総配線長を示す．同図右は単純実装を 1 として正規化した値である．次数は $m = \log_2 n$，クラスタリング手法は等分割法，ほかの条件は前章と同じである．1 ラックに 16 ノードが格納される仮定から，16,384 ノードは 1,024 ラックであり，これは京コンピュータの 864 ラック [20] を上回る規模である．

$\theta = 0.5$ の近傍ランダムリングについて見ると，128 ノード = 8 ラックで単純実装比 18%減（2.26 km → 1.86 km），16384 ノード = 1,024 ラックで同 33%減（3,257 km → 2,184 km）となった．このことは，単にケーブル敷設コストの削減を意味するだけでなく，省資源化，メンテナンス性の向上，床下空間の占有体積減少による空調効率の向上など，幅広い効用をもたらすと考えられる．逆に言えば，ランダムトポロジーの単純実装は費用対効果が極めて悪い．したがって，実際にランダムなネットワークトポロジーをもつスーパコンピュータを建設するにあたっては，何らかのシステマティックなラック配置最適化手法が必要不可欠であり，本研究はその端緒を開いたものと位置づけられる．

7.2　規則的なトポロジーへの適用

クラスタリングとマッピング最適化は任意のネットワークトポロジーに対して可能であり，特に，そのトポロジーの最適なマッピングが自明でない場合に有用である．

図 8 をハイパキューブについて見ると，8,192 ノード・13 次（512 ラック）のとき，SA 法で最適化された総配線長はベースライン法に比べて 19%減少した．ラックは縦横同数の格子状に並べられる仮定から，512 ラックは 23×23 の格子状配置（最後列の 17 ラック分は空き地）となるが，13 次元のハイパキューブを 23×23 の 2 次元格子に落とし込む方法は自明でなく，ベースライン法のような単純なヒューリスティクスでは最適配置は得られない．同じことが 128 ノード（8 ラック = 3×3），512 ノード（32 ラック = 6×6），2,048 ノード（128 ラック = 12×11）にもいえる．一方，256 ノード（16 ラック = 4×4），1024 ノード（64 ラック = 8×8），4,096 ノード（256 ラック = 16×16），16,384 ノード（1,024 ラック = 32×32）のときは，SA 法とベースライン法の総配線長は同一だった．これらはハイパキューブの次元境界とラック設置場所の端とが合致し，単純なヒューリスティクスで最適配置が得られる「幸運な」場合である．

一般には，ラック設置場所は建物の設計やマシンルームの空きスペースといった外部要因によって決まるものであり，ネットワークトポロジーに合致したラック設置場所が常に確保できるとは考えられない．また，規則的なトポロジーをもつシステムを拡張する場合，ラック配置とトポロジーの最適化は更に困難である [21]．本研究で示したラック配置最適化手法は，マシンルームの形状に依存せず，ラック設置場所相互間の距離さえ定義されれば最適配置を求めることができる．これにより，ネットワークの設計とマシンルームの設計を分離することができ，スーパコンピュータを建設・拡張する際の自由度の向上に寄与するものである．

8.　む　す　び

本研究では，ランダムなネットワークトポロジーの低遅延性を維持したまま総配線長を削減した．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

局所性をもたせ，(2) クラスタリングによってラック間にまたがるリンク数を減らし，(3) ラックレイアウトを最適化することにより総配線長を最小化した．ケーススタディの結果，単純な実装と比べて，4,096 ノードで総配線長を最大 29%，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．

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.S. Hemmert, J.S. Vetter, K. Bergman, C. Das, A. Emami, C. Janssen, D.K. Panda, C. Stunkel, K. Underwood, and S. Yalamanchili, “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] M. Koibuchi, H. Matsutani, H. Amano, D.F. Hsu, and H. Casanova, “A case for random shortcut topologies for HPC interconnects,” Proc. International Symposium on Computer Architecture (ISCA), pp.177–188, 2012.

[4] J. Kim, W.J. Dally, S. Scott, and D. Abts, “Technology-driven, highly-scalable dragonfly topology,” Proc. International Symposium on Computer Architecture (ISCA), pp.77–88, 2008.

[5] C.C. Aggarwal and H. Wang, Managing and Mining Graph Data, Springer, 2010.

[6] Pitu B. Mirchandani and Richard L. Francis, eds., Discrete Location Theory, Wiley-Interscience, 1990.

[7] R.E. Burkard, E. Çela, S.E. Karsch, and F. Rendl, “QAPLIB - A Quadratic Assignment Problem Library”. http://www.seas.upenn.edu/qaplib/

[8] 藤原一毅，鯉渕道紘，“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,” 並列/分散/協調処理に関する『鳥取』サマー・ワークショップ（SWoPP 鳥取 2012），pp.97–102, 2012.

[9] W.D. Dally and B. Towles, Principles and Practices of Interconnection Networks, Morgan Kaufmann, 2003.

[10] A. Clauset, M.E.J. Newman, and C. Moore, “Finding community structure in very large networks,” Phys. Rev. E, vol.70, 066111, Dec. 2004.

[11] Joe H. Ward Jr., “Hierarchical grouping to optimize an objective function,” J. American Statistical Association, vol.58, no.301, pp.236–244, 1963.

[12] P. Pons and M. Latapy, “Computing communities in large networks using random walks,” Computer and Information Sciences - ISCIS 2005, pp.284–293, 2005.

[13] M.E.J. Newman and M. Girvan, “Finding and evaluating community structure in networks,” Phys. Rev. E, vol.69, 026113, Feb. 2004.

[14] “The igraph library”. http://igraph.sourceforge.net/

[15] D.T. Connolly, “An improved annealing scheme for the QAP,” European Journal of Operational Research, vol.46, no.1, pp.93–100, May 1990.

[16] E.D. Taillard, “Robust taboo search for the quadratic assignment problem,” Parallel Computing, vol.17, no.4-5, pp.443–455, July 1991.

[17] E.D. Taillard, “FANT: Fast ant system,” Technical report, Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, Oct. 1998.

[18] Y. Li, P.M. Pardalos, and M.G. Resende, “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.

[19] HP, “Optimizing facility operation in high density data center environments, technoloogy brief,” 2007.

[20] Y. Ajima, S. Sumimoto, and T. Shimizu, “Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers,” Computer, vol.42, pp.36–40, 2009.

[21] A.R. Curtis, T. Carpenter, M. Elsheikh, A. López-Ortiz, and S. Keshav, “Rewire: An optimization-based framework for unstructured data center network design,” Proc. International Conference on Computer Communications (INFOCOM), pp.1116–1124, 2012.

（平成 24 年 10 月 18 日受付，25 年 3 月 10 日再受付）

藤原　一毅　（正員）

2002 東工大・工・制御システム卒．2004 同大大学院理工学研究科機械制御システム専攻了．2004～2008 日立製作所情報制御システム事業部．2012 総合研究大学院大学複合科学研究科情報学専攻了．博士（情報学）．現在，国立情報学研究所特任研究員．IEEE，情報処理学会各会員．

鯉渕　道紘　（正員）

2000 慶大・理工・情報工学卒．2003 同大大学院理工学研究科開放環境科学専攻後期博士課程了．博士（工学）．2002 年度より 2004 年度まで日本学術振興会特別研究員．現在，国立情報学研究所准教授，総合研究大学院大学複合科学研究科情報学専攻准教授（併任）．ハイパフォーマンスコンピューティングとインターコネクトに関する研究に従事．IEEE Computer Society Japan Chapter Young Author Award 2007，2007 年度情報処理学会論文賞受賞．IEEE 会員．