

高性能計算機インターコネクต์におけるランダムショートカットトポロジ

鯉 渕 道 紘,^{†1} 松 谷 宏 紀,^{†2} 天 野 英 晴,^{†2}
D. Frank Hsu,^{†3} Henri Casanova^{†4}

メニーコア並列アプリケーションと高性能計算機の大規模化が進むにつれて性能への通信遅延の影響が大きくなってきている。そのため、高性能計算システム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。そこで、本研究では、典型的な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探索する。 N 台の次数 k のスイッチで構成されたトポロジにおいてランダムなショートカットリンクは、直径を理想値である $\log_k N$ に近づけ、平均距離、トポロジの拡張性、耐故障性をスモールワールド効果により改善する。グラフ解析の結果より、ランダムなショートカットリンクは、規則的にショートカットリンクを付加した場合と比べて、直径と平均距離を最大 8 倍改良することが分かった。また、フリットレベルシミュレーションの結果より、ランダムなショートカットリンクは遅延を 35%削減し、ハイパーキューブなどの同じ次数を持つ規則的なトポロジと同程度のスループットを達成した。

Case for Random Shortcut Topologies on High-performance Computer Interconnects

MICHIHIRO KOIBUCHI,^{†1} HIROKI MATSUTANI,^{†2}
HIDEHARU AMANO,^{†2} D. FRANK HSU^{†3} and HENRI CASANOVA^{†4}

As the scale of many-core parallel applications and high-performance computer 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 high-performance computing systems. In this work, we explore to augment classical topologies with random “shortcut” links. Given a topology of N switches with degree k , we content that random shortcut links make it possible to drastically reduce the diameter close to $\log_k N$, reduce the average topological distance, improve topology expandability, and create a small-world effect that improves robustness to faults. Graph analysis results show that adding random shortcut links can improve diameter and average topological distance by up to a factor 8 when compared to adding non-random shortcut links. Flit-level discrete simulation results show that random shortcut links reduce latency by 35% and make it possible to achieve at least the same throughput as existing non-random topologies, including hypercubes.

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は、ストロング/ウィーク・スケーリングを問わず、数百ナノ秒～1 マイクロ秒の低 MPI 通信遅延が必要となることが予測されている¹⁾²⁾。したがって、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 Infiniband QDR 1 台のスイッチ遅延が約 100 ナノ秒など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。したがって、低直径、短い平均距離 (ホップ数) のトポロジを用いることがネットワーク内の低遅延化につながる。

現在、数十ポート以上の高次元スイッチが利用可能であるため、高次元トポロジの採用により低遅延化を探索することが可能である。高次元スイッチを基にした様々な規則的なトポロジがこれまで提案されており³⁾、各々、スイッチの直径、次数、レイアウト (リンク長)、スイッチ数、トポロジに最適化したデッドロックフリールーティング (例: k -ary n -cubes における次元順ルーティング)、耐故障性などの面でトレードオフを持つ。一般的に、これらの規則的なトポロジでは、スイッチの次数が大きくなるにつれて、直径はなだらかに小さくなる。

その他のトポロジの指標としては、拡張性と耐故障性が挙げられる。高性能計算機の規模は、電力、物理面積、コストなどの制限により決定されるが、例えば k -ary n -cubes

^{†1} 国立情報学研究所 / 総合研究大学院大学 / JST
National Institute of Informatics / The Graduate University
for Advanced Studies / JST
^{†2} 慶應義塾大学大学院 理工学研究科
Graduate School of Science and Technology, Keio University
^{†3} フォードダム大学 / Fordham University
^{†4} ハワイ大学マノアキャンパス / University of Hawai'i at Manoa

では k^n 個のノードに限定されるなど既存の規則的なトポロジはシステムサイズを選ぶ。さらに、規則的なトポロジは通常、対称性を利用したタスク割り当て、ルーティングを用いる。そのため、ネットワークの構成要素の故障/メンテナンス時におけるトポロジ維持のための冗長性、あるいは機能の検討が必要となる⁴⁾⁵⁾。

そこで、本研究では、典型的な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探索する。ランダムなショートカットリンクのアイデアは、グラフ理論の研究⁶⁾に端を発する。文献⁶⁾では1) 次数と上限の直径を与えるとランダムグラフは、ランダムではないグラフよりもネットワークサイズが大幅に向上すること、および2) 最適解に近づくためにランダムリンクをトポロジに加える提案が報告されている。

我々は、ランダムなショートカットリンクを規則的な低次元トポロジに加える利点を定量的に示し、高性能計算機のインターコネクトにおけるランダム要素を持つトポロジの有用性を追求する。なお、ランダムショートカットトポロジは拡張性が高く任意のスイッチ数、リンク数で構成することができる。

本研究で得られた知見は以下の通りである。

- グラフの解析結果より、ランダムなショートカットリンクを規則的なトポロジに加えることで、直径が理想値 $\log_k N$ (N : ネットワークサイズ, k : 次数) に近づき、一定間隔で規則的にショートカットリンクを加える場合と比べ、最大 8 倍の効果がある。
- Topology-agnostic デッドロックフリールーティング⁷⁾を用いることで、ランダムなショートカットリンクの付加が経路長の削減に直結する。また、リンク数が増加するにつれて、平均経路長と直径の値が近づく。つまり、経路遅延が均一に近づく。経路長が均一ならば、並列アプリケーションをトポロジの特性に合わせ最適化する負担を抑えることができる。
- フリットレベルシミュレーションの結果より、ランダムなショートカットリンクは低負荷時の遅延を 35% 削減し、ハイパーキューブなどの同じ次数を持つ規則的なトポロジと同程度のスループットを達成した。
- ランダムなショートカットリンクにより、ロバスト性を持つスモールワールド効果を獲得できる。そのため、高い耐故障性、高スループット、低レイテンシのパラメータを高い水準で実現できる。

本論文の構成は以下である。2 章で関連研究を述べ、3 章において、ランダムショートカットトポロジの生成について述べる。4 章では、グラフ解析からランダムショートカットトポロジを評価し、5 章では、フリットレベルシミュレーションによりランダムショートカットトポロジを評価する。最後に、6 章において結論と今後の課題を述べる。

2. 関連研究

2.1 規則的な高次元トポロジ

Fat ツリー、 k -ary n -cubes (トーラス、メッシュ、ハイパーキューブ) がスーパーコンピュータのインターコネクトのトポロジで幅広く利用されてきた。これらは、ツリーの階層間のリンク数増加、あるいは、次元数を増加させることで高次元ネットワークに拡張可能である。

また、規則的な直接網は幅広く提案されており、直径、次数などで様々な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が挙げられる³⁾。

直径の最小値は $\log_k N$ (N : ネットワークサイズ, k : 次数) であり、クロス網、オメガ網やバタフライ網などの間接網 (multistage interconnection network: MIN) は、この値を取ることができる。ただし、一般的に間接網は必要となるスイッチ数が増大する。最近、folded クロス網、flattened バタフライ網が、各々単方向 Clos 網、バタフライ網から拡張されており、文献⁸⁾では、flattened バタフライ網がもっともコスト効率が高いと報告されている。

巨大なスーパーコンピュータの場合、システム・レイアウトがリンク長、ネットワークコストに大きく影響する。これは高バンド幅を実現するためにリンクが長い場合は、高価な光リンクとなる一方、キャビネット内などの短いリンクは電気ケーブルで構成できるためである。ドラゴンフライ・トポロジのフレームワーク⁹⁾では配線レイアウトに焦点をあて、トポロジをキャビネット間ネットワークとグループ内のネットワークの 2 つに分ける。そして、ネットワークの次数を改善するために、複数のルータで 1 つの仮想ルータを構成する。これら 2 種類のサブトポロジには、ランダムショートカットトポロジを含む様々なトポロジが利用可能である。

2.2 分散ループネットワーク (Distributed Loop Network: DLN)

トポロジの直径を削減するために、不規則なショートカットリンクの追加が大変効果的であることが知られている¹⁰⁾。事実、リングトポロジは、直径が $\lfloor N/2 \rfloor$ であるが、例えば、 $N = 36$ の場合、6 本のショートカットリンクを 11 ホップ離れたスイッチにショートカットリンクを加えることで、直径が最適解である 9 になることが報告されている¹⁰⁾。

最低 1 つのリング構造を含むネットワークを、分散ループネットワーク (distributed loop network: DLN) と呼ぶ。DLN において与えられたショートカットリンクの本数における最小の直径を求めることは難しいことが知られている¹⁰⁾。

2.3 複雑ネットワーク

スイッチ間のランダムなショートカットリンクの追加に

よっても、直径と平均距離（ホップ数）を削減することができる．D.J. Watts と S. Strogatz は、スモールワールドモデル (WS Model) を提案した¹¹⁾．彼らはリンクの置き換え確率を決める 1 つのパラメータのみで、1 次元格子トポロジがランダムグラフに移行していくことを示し、少数の長いリンクにより直径が劇的に削減されることを明らかにした．加えて、複雑ネットワークの持つスケールフリー、クラスタ性がネットワークのロバスト性と平均距離を短縮する．そのスモールワールド効果は、ソーシャルネットワーク、インターネット、電力網などで報告されており、チップ内配線に利用する研究も行われている¹²⁾¹³⁾．

これらの特性は高性能計算システムにおいて魅力的であるが、我々の知る限り、複雑ネットワーク効果を高性能計算インターコネクに適用、最適化した研究例はない．そこで、本研究では、スーパーコンピュータのインターコネクで使われきた規則的なトポロジとは一線を画し、ランダムにショートカットリンクを付加するアプローチを探索することでスモールワールド効果を高性能計算機インターコネクの性能向上にもたらすことを目指す．

3. ランダムショートカットトポロジ (RST)

本研究では、 N 台のスイッチを point-to-point リンクで接続したネットワークをトポロジとする．また、スイッチにおいてこれらのリンクを接続するポート数を次数とする．

ランダムショートカットトポロジ (*Random Shortcut Topology: RST*) は、ランダムではないトポロジに対して、ランダムに選択したノード対リンクを追加した構成をとる．これらのリンクをランダムなショートカットリンクと呼ぶ．

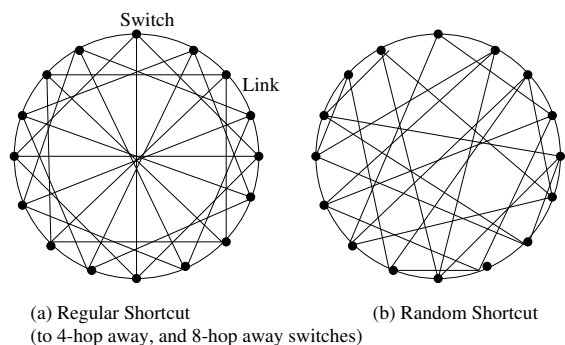


図 1 リングトポロジにおける規則的なショートカットリンクとランダムなショートカットリンクの追加 (次数 5)

Fig. 1 Adding regular shortcut links and random shortcut links to a ring topology (degree is five).

ここではランダムなショートカットリンクの生成の一般化、および文献⁶⁾で述べられた利点を持つリングトポロジ、つまり DLN に焦点をあてる．

リングトポロジの次数は 2 であり、直径は $\lfloor N/2 \rfloor$ である．すでに、DLN の直径を減らすために、ショートカットの

辺を追加する検討が行われている．各頂点に 0 から $N - 1$ までの番号を割り当て、等間隔にショートカットの辺を以下の関係にある頂点 i, j 間に加える．

$$j = i + \lfloor N/2^k \rfloor \bmod N \text{ for } k = 1, \dots, K \quad (1)$$

例えば、図 1(a) では、DLN ($N = 16, K = 2$) であるため次数が 5 となる．一方、同図 (b) では、同じ次数で、リングを基にランダムショートカットトポロジを作成した例である．

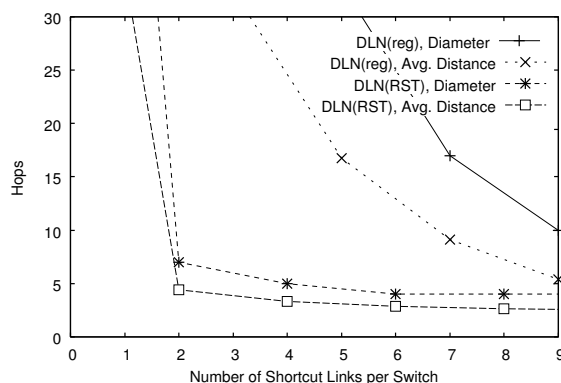


図 2 ショートカットリンク数の増加に伴う規則的な DLN とランダムな DLN トポロジの直径と平均距離 ($N = 256$ スイッチ)

Fig. 2 Diameter and average topological distance for regular DLN and random DLN topologies as the number of shortcut links increases ($N = 256$ switches).

図 2 に、等間隔にショートカットリンクを加えた DLN (図中の DLN(reg)) とランダムな DLN (図中の DLN(RST)) の直径とスイッチ間の平均距離 (ホップ数) を示す．なお、DLN(RST) の生成の詳細は 4 章で述べる．

図 2 より、スイッチあたり 2 本のランダムショートカットリンクを用いることで直径が 8 になるなど、ランダムショートカットリンクにより、極めて大きなホップ数削減効果が得られていることが分かる．さらに、直径と平均距離が接近し、経路長の均一性が増している点も大きな特徴である．これらは、理論的な所見⁶⁾を一層強める結果であり、我々は高性能計算機のトポロジにおいてこの効果を活用する．

4. ランダムショートカットトポロジ (RST) の解析

本章では、ランダムショートカットトポロジにおける直径、スイッチ間の平均距離、耐故障性、基となるトポロジの影響についてグラフ解析を行う．

4.1 方 法

次数 k' のランダムショートカットトポロジを生成するために、 N 台の次数 k のスイッチ (グラフの頂点) で構成された基のトポロジ (以後、基本トポロジと呼ぶ) へのランダムなショートカットリンクを追加する．このために次の 2 種類の追加法を用いた．

1 番目の手法では, $d < k'$ を満たす頂点に対して, $d - k'$ 本の新しいリンクを, k' 未満の次数を持つ異なる頂点に加える. この際, 各頂点を最大 10,000 回ランダムに探す. 失敗した場合 (最後の 2 つの頂点が次数 $k' - 2$ の場合など), 再試行する.

2 番目の手法は, ランダムに選択した $N \times (k' - k)/2$ 組の頂点にリンクを追加する.

そして, 各手法において, 100 個のランダムショートカットトポロジを生成し, 最小の直径のトポロジを選択する. 直径が同じ場合, 総リンク数の少ないトポロジを選択する. また, 耐故障性の評価として, 直径が 3 以上増加しない限り, 生成したランダムショートカットトポロジからランダムにリンクを削除する解析を行った. そして, その最大リンク数 (%) を平均耐故障性の指標とした. これらをグラフ解析プログラムである C++ Snap グラフライブラリ¹⁴⁾ を用いて実装した.

以降では, 両者の手法の比較を行う 4.3 節以外は 1 番目の手法を用いて評価を行う.

4.2 スケーラビリティ

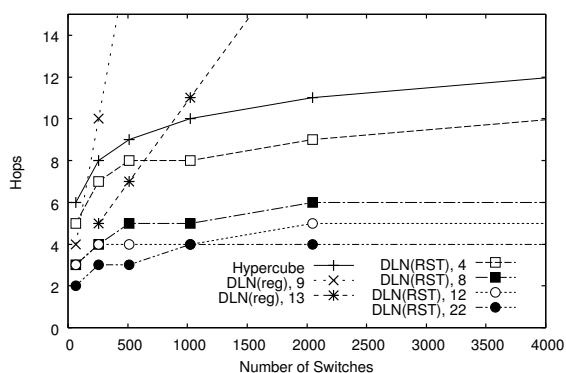


図 3 直径 対 N
Fig. 3 Diameter vs. N .

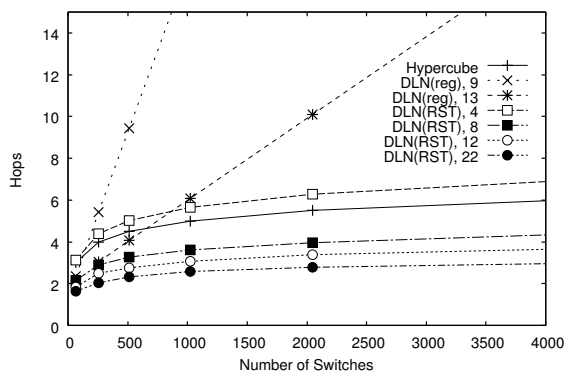


図 4 平均距離 対 N
Fig. 4 Average topological distance vs. N .

ここでは, ネットワークが大きくなるにつれて, 直径と平

均距離がどの位大きくなるのか? というトポロジのスケラビリティに焦点をあてる. 図 3 および 4 は, スイッチ数 N のトポロジにおける, これら 2 つの値を示している. この解析では, 2 つの規則的にショートカットリンクを加えた DLN(reg) と 4 つのランダムショートカットリンクを用いた DLN(RST) とハイパーキューブを比較した. 図中の表記はトポロジとその次数を表している. DLN(reg) では次数 9, 13 の場合 (各々 7, 11 本のショートカットリンクを追加) を評価し, DLN(RST) では次数 4, 8, 12, 22 の 4 つの場合を評価した. なお, DLN(RST) は, 4.1 節において 1 番目のアプローチにより生成されており, ネットワーク内のスイッチの次数は均一である. また, ハイパーキューブの次数はスイッチ数 N に対して $\log_2 N$ となる.

DLN(RST) は DLN と (reg) と比べて, 直径が数倍以上優れており, $\log_k N$ に近づいていることが分かる. また, 4,096 スイッチの場合, 次数 4 のランダムショートカットトポロジは, 次数 12 となるハイパーキューブと比べて直径が小さい. この例からも, ランダムショートカットが直径の削減に極めて効果的であることが分かる.

4.3 スイッチの次数の均一性

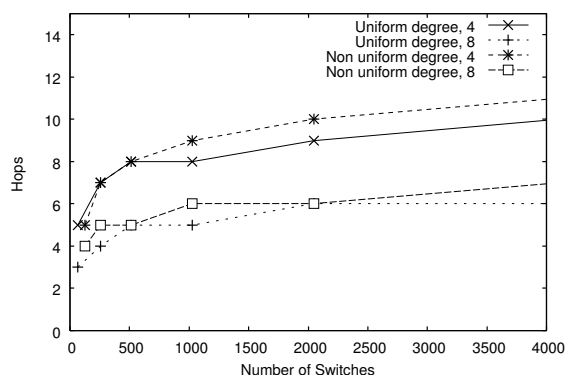


図 5 スイッチの次数が均一な場合と不均一な場合の DLN(RST) の直径 対 N
Fig. 5 Diameter vs. N of DLN(RST)s with uniform and non-uniform switch degrees.

4.1 節において, ランダムショートカットリンクの生成法がすべてのスイッチが同じ次数をもつ場合と, 同数に限定しない場合の 2 通りがあることを述べた. ここでは両手法を比較する.

図 5 および 6 は, 両手法により生成したトポロジの直径と平均距離を示している. “Uniform degree, 4” は, 1 番目の手法により構築された, すべてのスイッチの次数が 4 の DLN(RST) を示し, “Non uniform degree, 4” は 2 番目の手法により構築されている. 両方ともスイッチ次数の平均は 4 である.

両手法ともに直径, 平均距離に大きな差は見られない. 一方, 表 1 は, 同評価における 2 番目の手法におけるスイッ

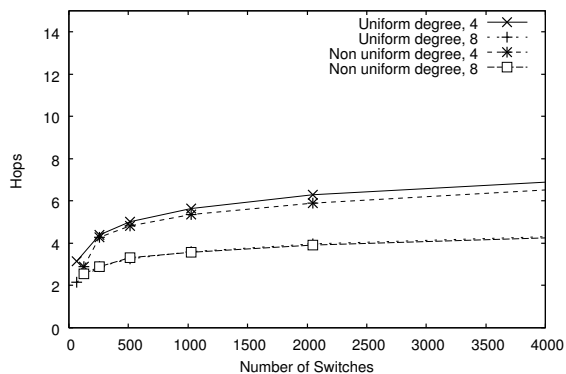


図 6 スイッチの次数が均一な場合と不均一な場合の DLN(RST) の平均距離 対 N

Fig. 6 Average topological distance vs. N of DLN(RST)s with uniform and non-uniform switch degrees.

チの次数の最小値と最大値であり、4 倍から 11 倍ほどの大きな開きがある。本研究では、経路の分散等を考え、1 番目の手法を以後用いるが、2 番目の手法においてスイッチ次数のばらつきがありながら、直径がハイパーキューブよりも低いことは興味深い。

表 1 スイッチの次数が不均一な場合の DLN(RST) の最小と最大のスイッチ次数 (図 5 および 6)

Table 1 Minimum and maximum switch degree for DLN(RST)s with non-uniform switch degrees shown in Figures 5 and 6.

Topology	256		1,024		4,096	
(Avg. Degree)	min	max	min	max	min	max
DLN(RST)(4)	2	8	2	9	2	11
DLN(RST)(8)	2	15	3	16	2	21
DLN(RST)(10)	4	18	3	22	2	22

4.4 耐故障性

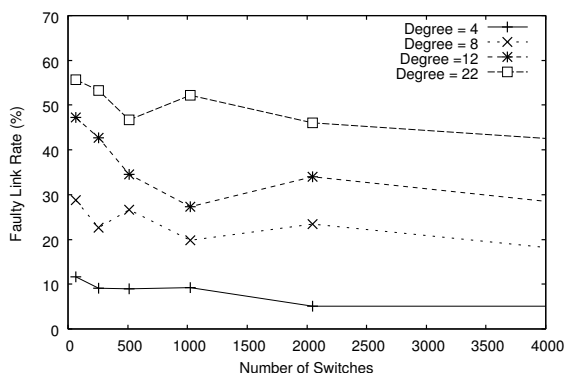


図 7 DLN(RST) における平均耐故障性 対 N

Fig. 7 Average fault tolerance vs. N on DLN(RST)s.

図 7 は、4 つの DLN(RST) における 4.1 節で定義した耐故障性の指標とネットワークサイズとの関係を示している。耐故障性については、いずれの DLN(RST) もネット

ワークサイズが大きくなるにつれて緩やかに減少していることが分かる。また、耐故障性はショートカットリンクの本数が増加するにつれても向上している。例えば、スイッチあたり 10 本のランダムなショートカットリンクを持つ DLN(RST) は、約 30% のリンクが故障しても直径の増加が 2 以内である。

一般的に規則的なトポロジでは独自のルーティングが採用されている。そのため、ネットワークの冗長化あるいは特殊な耐故障ルーティングの実装を耐故障性向上に利用することが多い。一方、ランダムショートカットトポロジは、topology-agnostic ルーティングを用いるため、デッドロックフリーなネットワーク再構成技術により故障箇所を迂回するように経路を更新できる利点もある。

4.5 基本トポロジの影響

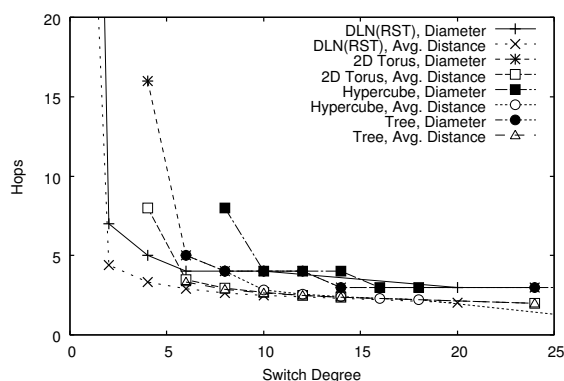


図 8 様々な基本トポロジによるランダムショートカットトポロジの直径 対 次数 ($N = 256$ スイッチ)

Fig. 8 Diameter vs. degree for random shortcut topologies with various baseline topologies ($N = 256$ switches).

ランダムショートカットトポロジにおける基本トポロジの影響について調査した。図 8 はリング、2-D トーラス、ハイパーキューブ、ツリーに各々ランダムなショートカットリンクを追加して構成したランダムショートカットトポロジの直径とスイッチ次数を示している。横軸は、与えた次数であり、基本トポロジにより同次数におけるショートカットリンクの本数は異なることになる。

次数が小さい場合は、ランダムなショートカットリンク数が大きいトポロジが有利であるが、次数が大きくなるにつれて、基本トポロジの選択が直径に与える影響は小さくなる。

5. ランダムショートカットトポロジ (RST) の性能評価

本章では、ランダムショートカットトポロジと比較のためハイパーキューブを含む k -ary n -cubes、参考として間接網である Myrinet Clos を評価する。

5.1 シミュレーション環境

スイッチと point-to-point リンクで構成されたネットワークをモデルとして C++ で記述されたフリットレベルシミュレータを開発した。このモデルでは、スイッチングファブリックはチャネルバッファ、クロスバ、リンクコントローラ、制御回路で構成される。また、各スイッチは、同数の次数を持ち、スイッチング技術としてパーチャルカッスルーを用いた。また、シミュレーション時間の単位を 2.5 ナノ秒と換算してスループット、遅延を計算した、具体的にはヘッダフリットがスイッチを通過する遅延は最低 100 ナノ秒 (40 シミュレーション時間) とした。この中にルーティング計算、仮想チャネルアロケーション、スイッチアロケーション、入力ポートから出力ポートへのクロスバを経由したフリット転送遅延が含まれる。また、フリットの注入遅延、リンク遅延は 5 ナノ秒 (2 シミュレーション時間) とした。

また、フリットサイズは 256 ビットとし、リンクの実効バンド幅は Infiniband QDR と同様に 96Gbps とした。また、低遅延が必要となる通信粒度は細かく (例: 3KB 未満)、さらにその細粒度通信でのスループットが重要であることが報告されているため¹⁾、パケット長は 9 フリット (ヘッダ含む) とした。また、トラフィックパターンはランダムとした。

すべてのトポロジにおいて仮想チャネルは 2 本とし、各スイッチは同数のローカルホストと接続している。ランダムショートカットトポロジは不規則性が強いいため、デッドロックフリーな topology-agnostic ルーティングである Duato のアプローチを採用し、その逃げ道として up*/down* routing を用いた¹⁵⁾。Duato のアプローチでは一度パケットが逃げ道の仮想チャネルに注入された場合、その仮想チャネルを使い続ける必要があるが、多くのパケットが最短経路を取ることができる。メッシュ、ハイパーキューブでは Duato's protocol を用い、逃げ道として次元順ルーティングを用いた。また、参考としてトーラスでは 2 本の仮想チャネルを使う次元順ルーティングを用いた。

5.1.1 トポロジの比較

図 9, 10 および 11 に、64, 256, 512 台のスイッチで構成されたネットワークにおける受信トラフィックとレイテンシの関係を示す。各トポロジにおいて最大受信トラフィック量がスループットとなる。図中の表記はトポロジとその次数を表している。“DLN(RST), 4” は、リングにスイッチあたりランダムショートカットリンク 2 本を追加した次数 4 のトポロジを表している。経由スイッチ数が遅延の主な要因であるため、平均ホップ数の小さい DLN(RST) は、同じ次数を持つトーラスに比べて最大 35 % メッシュと比べて最大 50% の低負荷時の遅延を削減している。

Myrinet Clos は、間接網であるため直接網との公平な比較は難しいが、参考のために 80 台の 16 ポートスイッ

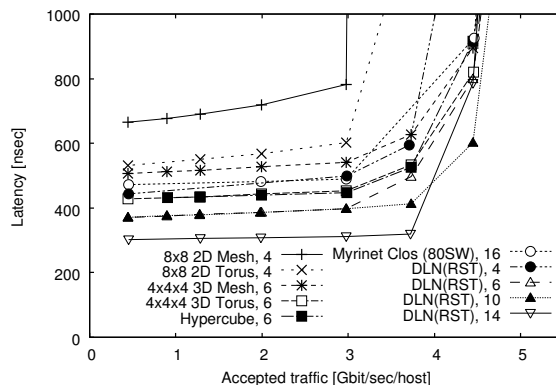


図 9 トポロジの比較 (64 スイッチ, 256 ホスト)
Fig. 9 Topology Comparison (64 switches, 256 hosts).

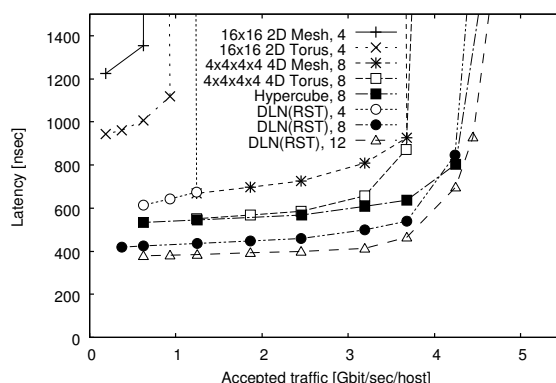


図 10 トポロジの比較 (256 スイッチ, 2,048 ホスト)
Fig. 10 Topology Comparison (256 switches, 2,048 hosts).

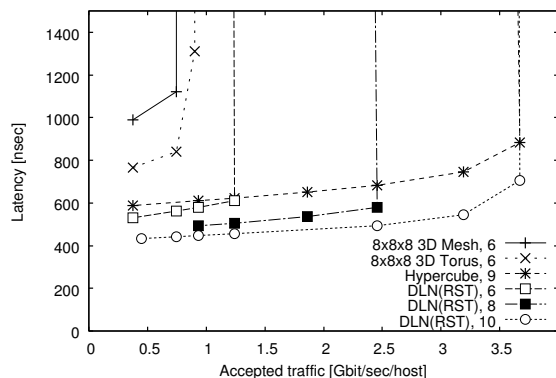


図 11 トポロジの比較 (512 スイッチ, 8,196 ホスト)
Fig. 11 Topology Comparison (512 switches, 8,196 hosts).

チで構成された場合の結果を図 9 に含めている。ただし、Myrinet Clos は、(直接網との比較のために出発地ホストの接続しているスイッチを省いても) 直径が 4 となるため、性能面で次数 14 で直径が 3 である DLN(RST) に性能面で劣ることが分かった。

ハイパーキューブは次数が大きいため、規則的なトポジの中では最も遅延が小さく、スループットが高いため優れているが、同じ次数の DLN(RST) とほぼ同等のスループット

を達成している．しかし、ハイパーキューブは DLN(RST) に比べて遅延の面で劣るためネットワークの低遅延化にはランダムなショートカットが有効であることが改めて明らかになった．なお、2-D および 3-D トーラスとメッシュは、遅延の面でスケラビリティが低く、逆にハイパーキューブと DLN(RST) はネットワークサイズが大きくなるにつれて、遅延が緩やかに大きくなることが分かった．

5.1.2 耐故障性

次に、リンク故障がスループットおよび遅延に与える影響について評価する．本評価では、次数 8 の DLN(RST) から、ショートカットリンク、リングを構成しているリンクを問わず、一定の確率でランダムに削除した上で測定した．

図 12 にリンクの削除割合を 0 % ~ 20 % に変化させた場合のスループットとレイテンシを示している．

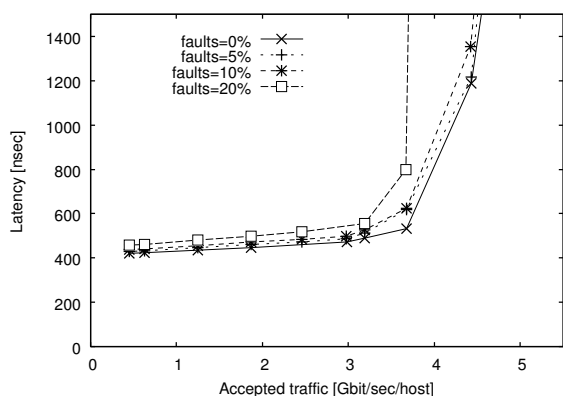


図 12 次数 8 の DLN(RST) の耐故障性 ($N = 256$ スイッチ, 2,048 ホスト)

Fig. 12 Fault Tolerance of DLN(RST)s with eight degrees ($N = 256$ switches, 2,048 hosts).

図 12 より、10%のリンクが故障してもスループットと遅延は緩やかに減少、増加するに留まることが分かる．つまり、DLN(RST) は topology-agnostic ルーティングとの併用により性能面と耐故障性を高い水準で達成していることが分かる．

5.1.3 ランダムショートカットトポロジのばらつき

本節では、はじめに、基本トポロジがスループットと遅延に与える影響を評価する．リング、ツリー、トーラスを基本トポロジとしたランダムショートカットトポロジの結果を図 13 および 14 に示す．“8x8 2D Torus+SC, 6” は、2-D トーラスに、スイッチあたり 2 本のランダムなショートカットリンクを追加した次数 6 のランダムショートカットトポロジを示している．図 13 および 14 より、ネットワークサイズとショートカットリンク数が大きくなると、基本トポロジの影響は小さくなることが分かる．

次に、DLN(RST) におけるランダムショートカットリンクのパターンのばらつきの影響を調査する．DLN(RST) を同一条件において異なる乱数を用いて 15 パターン生成し

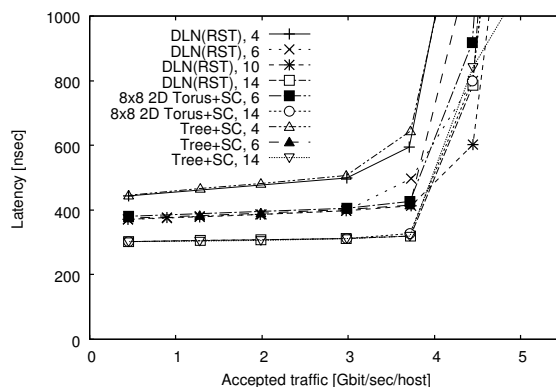


図 13 ランダムショートカットトポロジの比較 ($N = 64$ スイッチ, 256 ホスト)

Fig. 13 Comparison of Random Shortcut Topologies ($N = 64$ switches, 256 hosts).

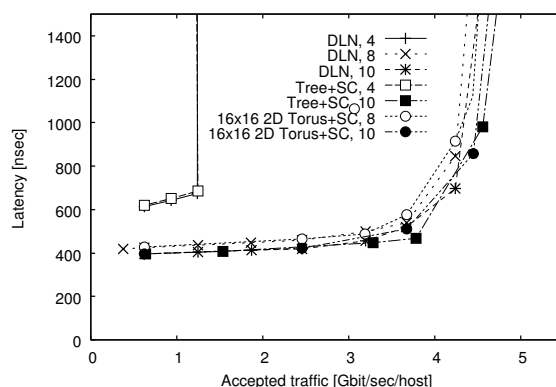


図 14 ランダムショートカットトポロジの比較 ($N = 256$ スイッチ, 2,048 ホスト)

Fig. 14 Comparison of Random Shortcut Topologies ($N = 256$ switches, 2,048 hosts).

た結果を図 15 および 16 に示す．

両図より、“ランダム” に決定されたショートカットリンクのパターンによるスループット、および遅延に関する影響はほとんどないことが分かる．これは、評価した規模のネットワークサイズでは、表 2 に示した通り、各パターンの直径、平均距離がほぼ同じとなるためである．

表 2 図 15 および 16 で用いた DLN(RST) の直径 および平均距離
Table 2 Diameter and average topological distance of DLN(RST)s used in Figures 15 and 16.

Topology (degree)	Diameter		Avg. Top. Distance	
	min	max	min	max
64-sw DLN(RST)(4)	5	6	3.114	3.222
256-sw DLN(RST)(8)	4	5	2.893	2.907

6. まとめと今後の課題

本研究では、高性能計算機インターコネクトの遅延を削減するための手段としてランダムショートカットリンクを

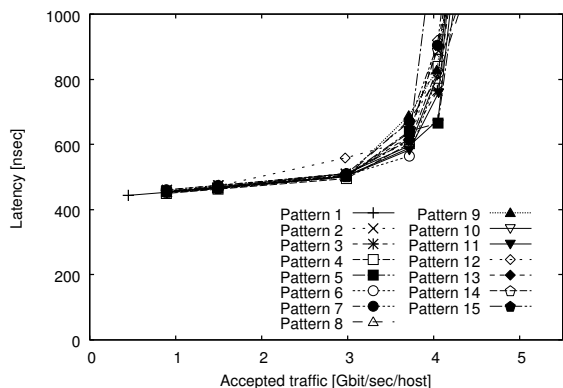


図 15 次数 4 の DLN(RST) におけるランダムショートカットパターンの影響 ($N = 64$)

Fig. 15 Influences of random shortcut patterns on DLN(RST) with four degrees ($N = 64$ switches).

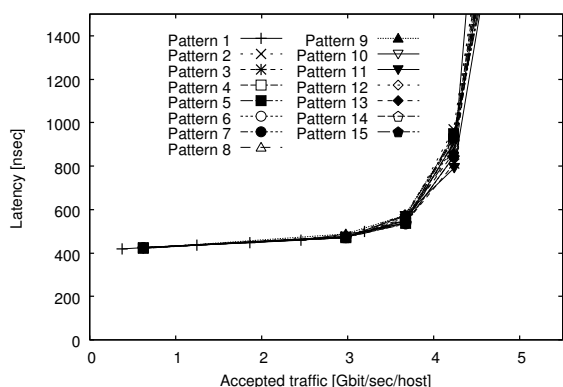


図 16 次数 8 の DLN(RST) のランダムショートカットパターンの影響 ($N = 256$ スイッチ)

Fig. 16 Influences of random shortcut patterns on DLN(RST) with eight degrees ($N = 256$ switches).

規則的なトポロジに追加することを探索し、 N 台の次数 k のスイッチを用いた場合、直径が理想値の $\log_k N$ に近づくことを示した。ランダムなショートカットリンクは、耐故障性を向上させる利点も持つ。例えば、スイッチあたり 10 本のランダムショートカットリンクを持つリングトポロジは、30%以上のリンクを除いても直径が 3 未満の増加に留まった。

また、シミュレーション結果より、ランダムなショートカットリンクは低負荷時の遅延を 35%削減し、ハイパーキューブなどの同じ次数を持つ規則的なトポロジと同程度のスループットを達成した。

今後は、(1) 距離の二乗に応じてショートカットリンクを分散させる¹⁶⁾ などのランダムショートカットトポロジの配線長、レイテンシとスループット面でのチューニングの検討、(2) キャビネットのレイアウトを考慮にいたランダムショートカットリンクの配線長の算出、などの様々な視点においてランダムショートカットトポロジが高性能計算機インターコネクトのトポロジの有力候補であること

をより一層具体的かつ明確にする予定である。

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/pubs-archive/iaa-ic-2008-workshop-report-final.pdf>.
- 2) J. Tomkins: Interconnects: A Buyers Point of View, ACS Workshop, 2007.
- 3) 天野英晴: 並列コンピュータ, 昭晃堂 (1996).
- 4) Coteus, P. and et. al.: Packaging the Blue Gene/L supercomputer, *IBM Journal of Research and Development*, Vol.49, No.2/3, pp.213–248 (2005).
- 5) Ajima, Y., Sumimoto, S. and Shimizu, T.: Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers, *IEEE Computer*, Vol.42, pp.36–40 (2009).
- 6) Bollobás, B. and Chung, F. R.K.: The Diameter of a Cycle Plus a Random Matching, *SIAM J. Discrete Math.*, Vol.1, No.3, pp.328–333 (1988).
- 7) Flich, J., Skeie, T., Mejia, A., Lysne, O., Lopez, P., Robles, A., Duato, J., Koibuchi, M., Rokicki, T. and Sancho, J.C.: A Survey and Evaluation of Topology Agnostic Deterministic Routing Algorithms, *IEEE Trans. on Parallel and Distributed Systems*, Vol.99, No.PrePrints (2011).
- 8) John Kim and William J. Dally and Dennis Abts: Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks, *ISCA*, pp.126–137 (2007).
- 9) Kim, J., Dally, W. J., Scott, S. and Abts, D.: Technology-Driven, Highly-Scalable Dragonfly Topology, *ISCA*, pp.77–88 (2008).
- 10) Bermond, J.-C., Comellas, F. and Hsu, D.F.: Distributed Loop Computer Networks: A Survey, *J. Parallel Distrib. Comput.*, Vol. 24, No. 1, pp. 2–10 (1995).
- 11) Watts, D. J. and Strogatz, S. H.: Collective dynamics of ‘small-world’ networks, *Nature*, Vol.393, No.6684, pp.440–442 (1998).
- 12) Ogras, U.Y. and Marculescu, R.: It’s a small world after all’: NoC performance optimization via long-range link insertion, *IEEE Trans. Very Large Scale Integration Systems*, Vol.14, pp.693–706 (2006).
- 13) Nishioka, Y., Iida, M. and Sueyoshi, T.: Small-World Network to Reduce Delay in FPGA Routing Structures, *International Journal of Innovative Computing, Information and Control (IJICIC)*, Vol.6, No.2, pp.551–566 (2010).
- 14) Leskovec, J.: SNAP Network Analysis Library, <http://snap.stanford.edu/>.
- 15) Silla, F. and Duato, J.: High-Performance Routing in Networks of Workstations with Irregular Topology, *IEEE Trans on parallel and distributed systems*, Vol.11, No.7, pp.699–719 (2000).
- 16) Kleinberg, J.: The Small-World Phenomenon and Decentralized Search, *SIAM News*, Vol.37, No.3, pp. 1–2 (2004).