

HPCネットワーク用ランダムトポロジ向けの 低遅延な分散ルーティング手法

2015年12月18日

河野 隆太（慶應大）
中原 浩 （慶應大）
藤原 一毅（国情研）
松谷 宏紀（慶應大）
天野 英晴（慶應大）
鯉渕 道紘（国情研）

目次

- スパコン向けネットワーク
 - 低ホップ数なネットワークトポロジの探求
- ランダムネットワーク
 - “近道”を利用したホップ数の削減
 - 配線長制限による実用性向上
 - ルーティングに関する問題
- 分散情報を用いたルーティング手法
 - 近傍・遠方ノードの情報追加によるルーティング
- 評価
 - 通信遅延削減効果
 - テーブルサイズ
- 結論

目次

- スパコン向けネットワーク
 - 低ホップ数なネットワークトポロジの探求
- ランダムネットワーク
 - “近道”を利用したホップ数の削減
 - 配線長制限による実用性向上
 - ルーティングに関する問題
- 分散情報を用いたルーティング手法
 - 近傍・遠方ノードの情報追加によるルーティング
- 評価
 - 通信遅延削減効果
 - テーブルサイズ
- 結論

スパコン向け相互結合網への要求



System Interconnects

	2011		2015		2019	
System Size	32,768		32,768		32,768	
Sockets	32		200		800	
Peak PF	1.0		6.1		25.0	
TF/Socket						
	Expect	Want	Expect	Want	Expect	Want
NIC B/W (B/F)	0.01 - 0.1	1.0	0.005 - 0.03	1.0	0.025 - 0.25	1.0
Link B/W (B/F)	0.01 - 0.1	1.0	0.005 - 0.03	1.0	0.025 - 0.25	1.0
MPI Latency (ns)	750 - 1500	500	500 - 1000	400	400 - 750	300
MPI Throughput (M Msg/s)	20	50	80	300	300	1200
Load/Store (M Msg/s)	75	400	150	1,600	300	6400
Load/Store Latency (ns)	300	100	300	100	300	100

システム内1μ秒 [1]

スイッチ経由遅延: 100n秒/sw

リンク通過遅延: 5n秒/m

[2]

エンド間**経由スイッチ数**
の削減

低ホップ数な
スイッチ間NW

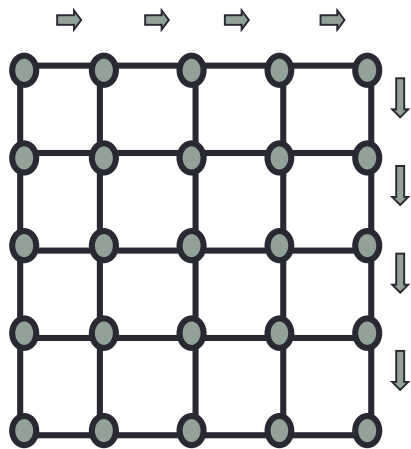


[1] K. Scott Hemmert et al: "Report on Institute for Advanced Architectures and Algorithms, Interconnection Networks Workshop 2008", <http://ft.ornl.gov/doku/media/iaaicw/iaa-ic-2008-workshop-report-v09.pdf>.

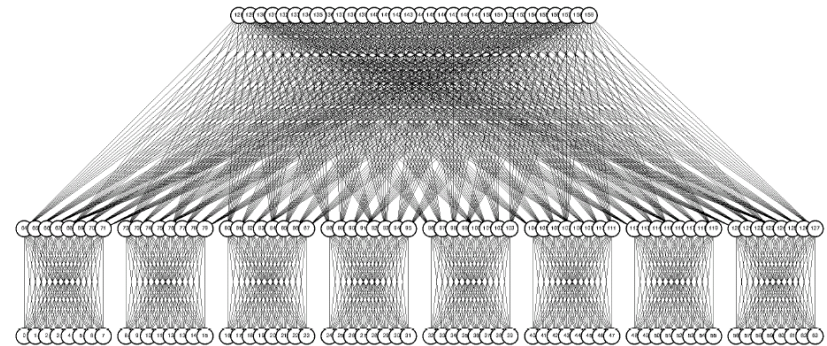
[2] J. Tomkins: "Interconnects: A Buyers Point of View", ACS Workshop, 2007.

スイッチ間ネットワークのトポロジとしてのモデル化

- 「スイッチ $\Rightarrow$ ノード, リンク $\Rightarrow$ 辺」とした
トポロジグラフにモデル化
- ポート数 ($\Rightarrow$ 次数) の大きなスイッチを用いて
ホップ数を削減
 - エンドツーエンドの最大ホップ数 ($\Rightarrow$ 直径) を削減



次数 4 : 直径 2 ($n - 1$)



The 512 hosts connect to 8 ports on each of these 64 "leaf" switches

次数 16, 直径 5 : 512ホスト

実例：スパコンで用いられるNWトポロジ

- Top500, Jun 2014 [3]

順位	名称 (国)	NWトポロジ
1	天河2号 (中国)	Fat tree
2	タイタン (アメリカ)	Torus
3	セコイア (アメリカ)	Torus
4	京 (日本)	Torus
5	ミラ (アメリカ)	Torus
6	Piz Daint (スイス)	Dragonfly
7	Stampede (アメリカ)	Fat tree
8	JuQUEEN (ドイツ)	Torus
9	Vulcan (アメリカ)	Torus
10	(非公表) (アメリカ)	Dragonfly

- Dragonfly網 [4]

- スイッチ間網を
キャビネット内・キャビネット外
に分け,
各区域でトポロジ形成し低遅延化
- 高基数スイッチを利用

[3] top500.org, <http://www.top500.org/lists/2014/06/>.

[4] J. Kim, W. J. Dally, S. Scott and D. Abts: "Technology-Driven, Highly-Scalable Dragonfly Topology", ISCA'08.

平均ホップ数の小さいトポロジの探索

- 与えられたノード数・次数に対する最小ホップ数トポロジの探索 [5]

平均ホップ数の小さいトポロジの探索

- 与えられたノード数・次数に対する最小ホップ数トポロジの探索 [5]

[Home](#)
[Problem](#)
[Solutions](#)
[Submit](#)
[Event](#)
[Q&A](#)
[About](#)

Find a graph that has smallest diameter & average shortest path length given an order and a degree.

News

- 2015-12-10: Announced the final rankings for the 2015 competition!
- 2015-10-16: Submission closed! Finalists will be notified by 2015-10-20
- 2015-10-06: Added 2 new solutions! This is the final update before the deadline
- 2015-09-28: Added 3 new solutions!
- 2015-09-21: Added 5 new solutions!
- 2015-09-14: Added 4 new solutions!
- 2015-09-07: Added 6 new solutions!
- 2015-08-31: Added 8 new solutions!
- 2015-08-24: Added 7 new solutions!
- 2015-08-17: Added 17 new solutions!
- 2015-08-10: Added 12 new solutions!

Best solutions

Update 2015-12-10

Degree d	Order n				
	16	64	256	4096	10000
3	3 / 2.200 0.000%	5 / 3.770 0.211%	8 / 5.636 0.861%	13 / 9.787 2.928%	15 / 11.111 3.122%
4	3 / 1.750 0.962% ²	4 / 2.869 0.417%	6 / 4.134 1.065%	9 / 6.753 4.373%	10 / 7.601 3.480%
16	N/A	2 / 1.746 0.000%	3 / 2.093 8.020% ²	4 / 3.253 8.716%	5 / 3.625 1.060%
23	N/A	2 / 1.635 0.000% ¹	2 / 1.910 0.000%	4 / 2.886 0.731%	4 / 3.200 8.675%
60	N/A	2 / 1.048 0.000% ¹	2 / 1.765 0.000% ¹	3 / 2.295 8.975%	3 / 2.650 0.615%
64	N/A	N/A	2 / 1.749 0.000% ¹	3 / 2.242 12.994% ²	3 / 2.610 1.005%

Legend: Diameter / Average shortest path length (ASPL)
Gap from the lower bound of ASPL (%)

Notes: 1. A random graph is optimal. Submissions with this size will not be awarded.
2. There are no graphs with this size that satisfy the lower bound of diameter and ASPL [Erdős 1980].

Show complete list

Motivation

平均ホップ数の小さいトポロジの探索

- 与えられたノード数・次数に対する最小ホップ数トポロジの探索 [5]

[Home](#)
[Problem](#)
[Solutions](#)
[Submit](#)
[Event](#)
[Q&A](#)
[About](#)

Find a graph that has smallest diameter & average shortest path length given an order and a degree.

News

- 2015-12-10: Announced the final rankings for the 2015 competition!
- 2015-10-16: Submission Finalists will be notified 2015-10-20
- 2015-10-06: Added 2 new solutions! This is the final before the deadline
- 2015-09-28: Added 3 new solutions!
- 2015-09-21: Added 5 new solutions!
- 2015-09-14: Added 4 new solutions!
- 2015-09-07: Added 6 new solutions!
- 2015-08-31: Added 8 new solutions!
- 2015-08-24: Added 7 new solutions!
- 2015-08-17: Added 17 new solutions!
- 2015-08-10: Added 12 new solutions!

Best solutions

ノード数 (スイッチ数)

Degree d	16	64	256	4096	10000
3	3 / 2.200 0.000%	5 / 3.770 0.211%	8 / 5.636 0.861%	13 / 9.787 2.928%	15 / 11.111 3.122%
4	3 / 1.750 0.962% ²	4 / 2.869 0.417%	6 / 4.134 1.065%	9 / 6.753 4.373%	10 / 7.601 3.480%
16	N/A	2 / 1.746 0.000%	3 / 2.093 8.020% ²	4 / 3.253 8.716%	5 / 3.625 1.060%
23	N/A	2 / 1.635 0.000% ¹	2 / 1.910 0.000%	4 / 2.886 0.731%	4 / 3.200 8.675%
60	N/A	2 / 1.048 0.000% ¹	2 / 1.765 0.000% ¹	3 / 2.295 8.975%	3 / 2.650 0.615%
64	N/A	N/A	2 / 1.749 0.000% ¹	3 / 2.242 12.994% ²	3 / 2.610 1.005%

Legend: Diameter / Average shortest path length (ASPL)
Gap from the lower bound of ASPL (%)

Notes: 1. A random graph is optimal. Submissions with this size will not be awarded.
2. There are no graphs with this size that satisfy the lower bound of diameter and ASPL [Erdős 1980].

Show complete list

Motivation

平均ホップ数の小さいトポロジの探索

- 与えられたノード数・次数に対する最小ホップ数トポロジの探索 [5]

Home Problem Solutions Submit Event Q&A About

Find a graph that has smallest diameter & average shortest path length given an order and a degree.

News

- 2015-12-10: Announced the final rankings for the 2015 competition!
- 2015-10-16: Submission Finalists will be notified 2015-10-20
- 2015-10-06: Added 2 new solutions! This is the final before the deadline
- 2015-09-28: Added 3 new solutions!
- 2015-09-21: Added 5 new solutions!
- 2015-09-14: Added 4 new solutions!
- 2015-09-07: Added 6 new solutions!
- 2015-08-31: Added 8 new solutions!
- 2015-08-24: Added 7 new solutions!
- 2015-08-17: Added 17 new solutions!
- 2015-08-10: Added 12 new solutions!

Best solution

ノード数 (スイッチ数)

次数 (ポート数)

Degree d	16	64	256	4096	10000
3	3 / 2.200 0.000%	5 / 3.770 0.211%	8 / 5.636 0.861%	13 / 9.787 2.928%	15 / 11.111 3.122%
4	3 / 1.750 0.962% ²	4 / 2.869 0.417%	6 / 4.134 1.065%	9 / 6.753 4.373%	10 / 7.601 3.480%
16	N/A	2 / 1.746 0.000%	3 / 2.093 8.020% ²	4 / 3.253 8.716%	5 / 3.625 1.060%
23	N/A	2 / 1.635 0.000% ¹	2 / 1.910 0.000%	4 / 2.886 0.731%	4 / 3.200 8.675%
60	N/A	2 / 1.048 0.000% ¹	2 / 1.765 0.000% ¹	3 / 2.295 8.975%	3 / 2.650 0.615%
64	N/A	N/A	2 / 1.749 0.000% ¹	3 / 2.242 12.994% ²	3 / 2.610 1.005%

Legend: Diameter / Average shortest path length (ASPL)
Gap from the lower bound of ASPL (%)

Notes: 1. A random graph is optimal. Submissions with this size will not be awarded.
2. There are no graphs with this size that satisfy the lower bound of diameter and ASPL [Erdős 1980].

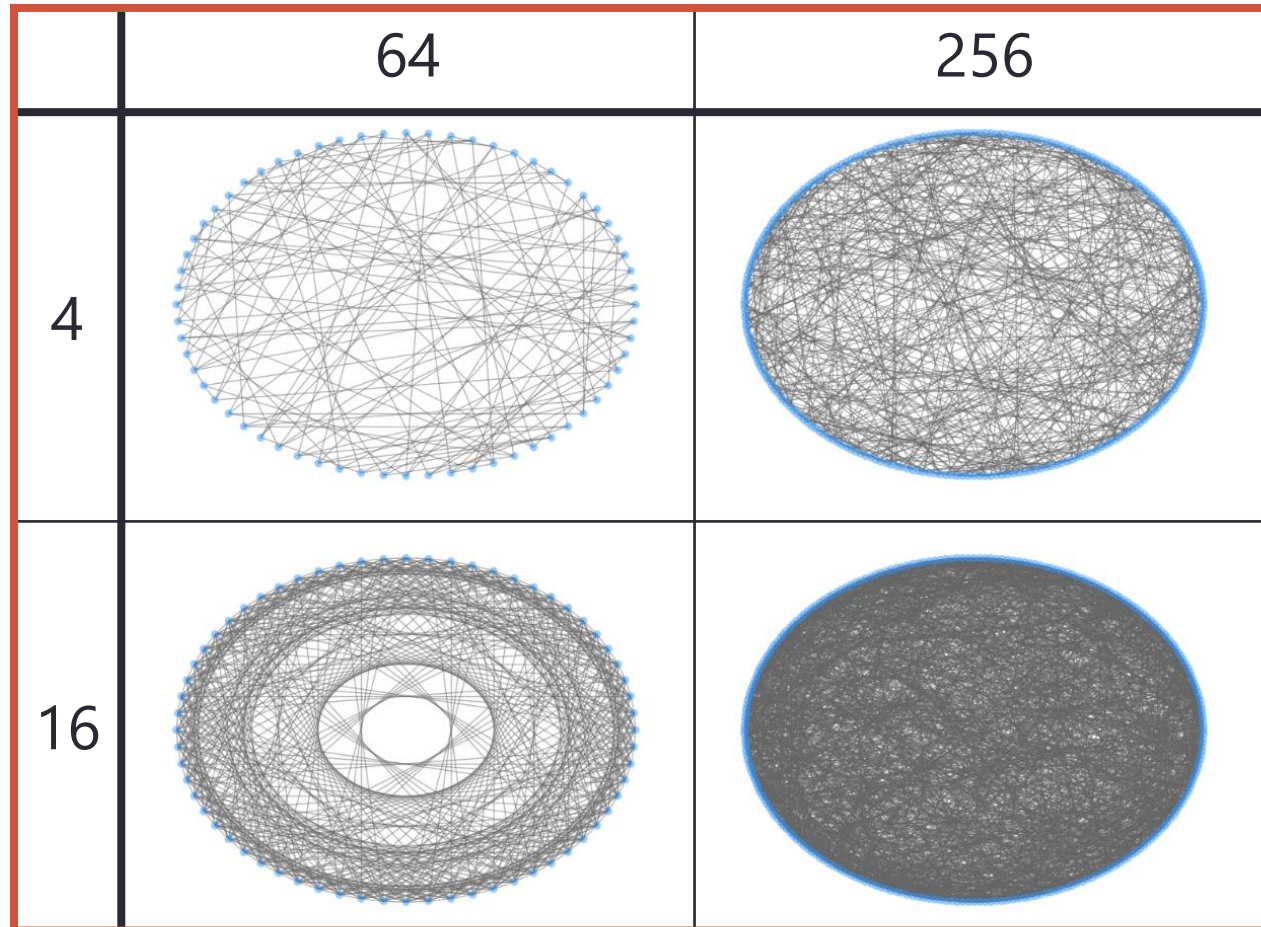
Show complete list

Motivation

探索により得られた低ホップ数トポロジ [5]

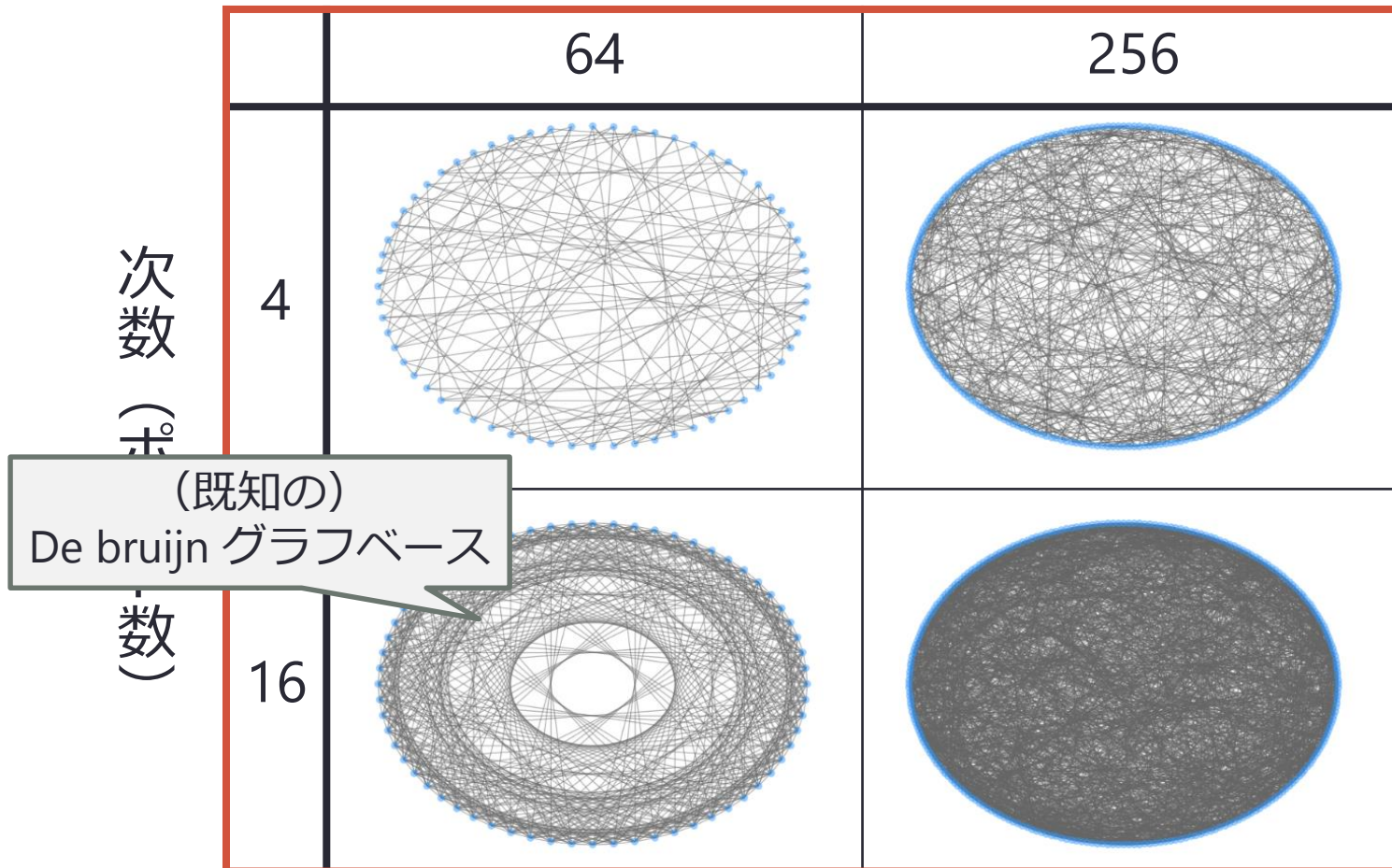
ノード数（スイッチ数）

次数（ポート数）



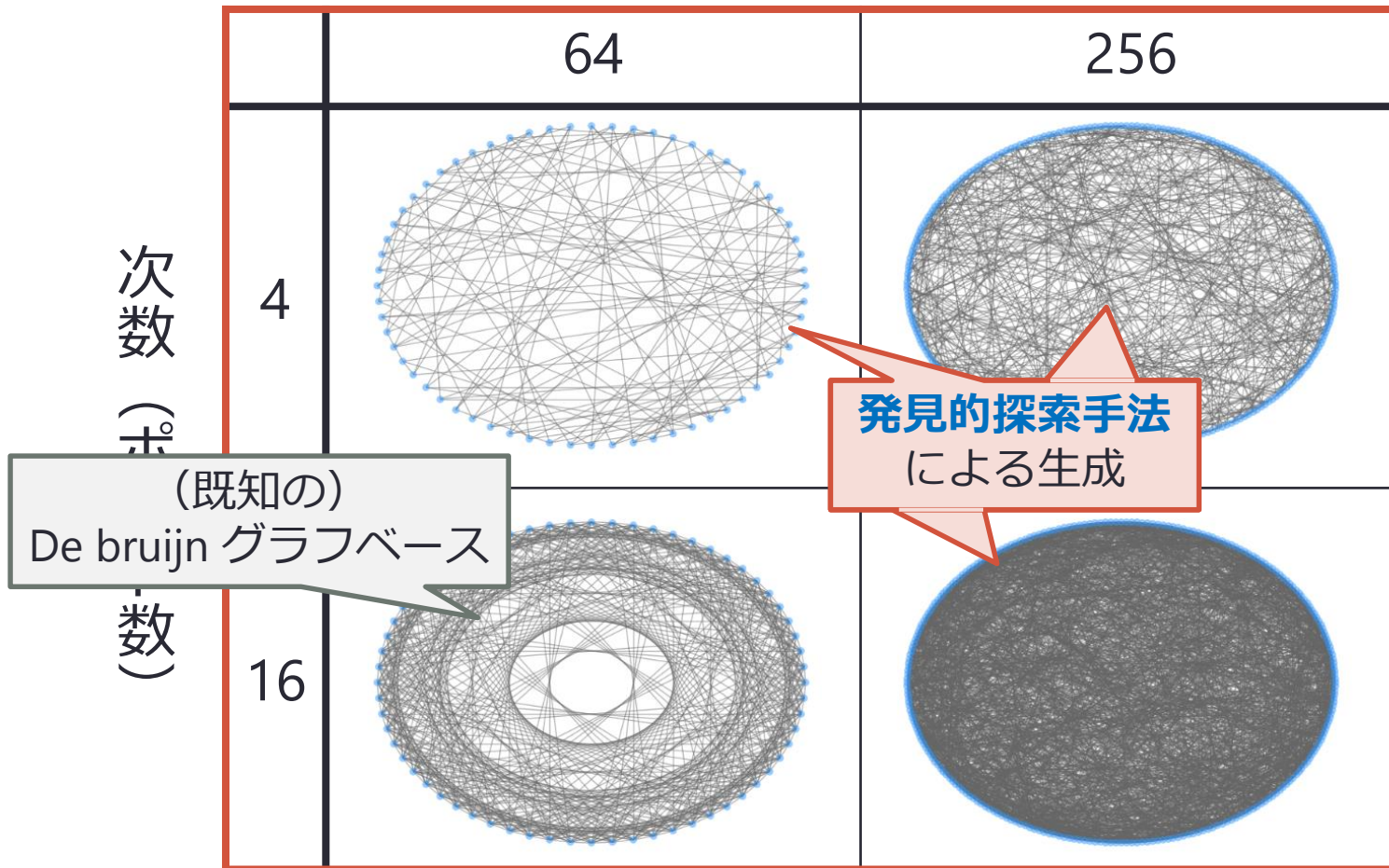
探索により得られた低ホップ数トポロジ [5]

ノード数 (スイッチ数)



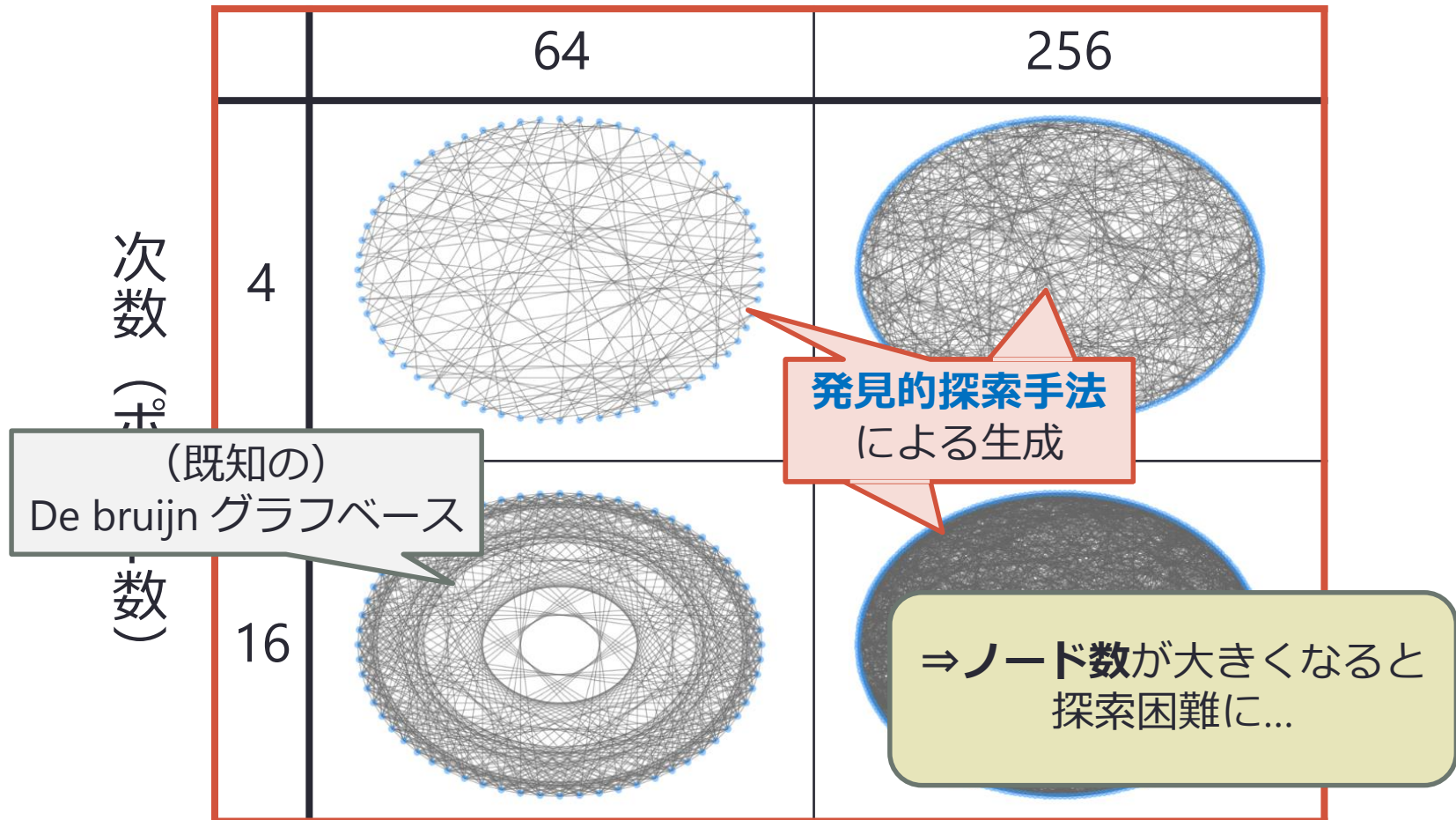
探索により得られた低ホップ数トポロジ [5]

ノード数（スイッチ数）



探索により得られた低ホップ数トポロジ [5]

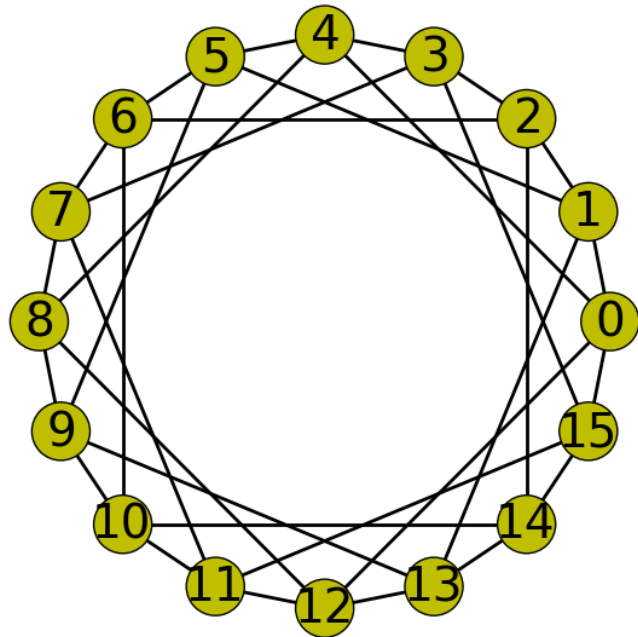
ノード数 (スイッチ数)



目次

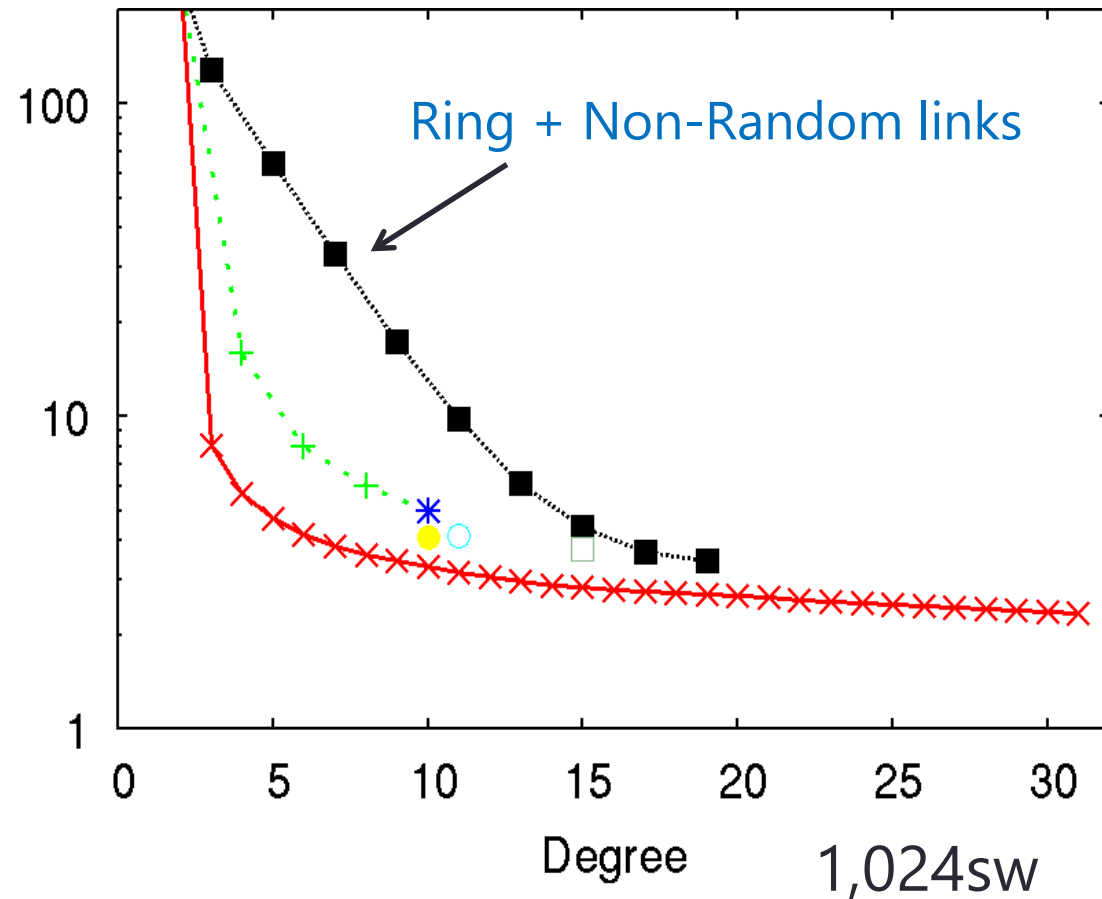
- スパコン向けネットワーク
 - 低ホップ数なネットワークトポロジの探求
- ランダムネットワーク
 - “近道”を利用したホップ数の削減
 - 配線長制限による実用性向上
 - ルーティングに関する問題
- 分散情報を用いたルーティング手法
 - 近傍・遠方ノードの情報追加によるルーティング
- 評価
 - 通信遅延削減効果
 - テーブルサイズ
- 結論

ランダムトポロジによる低遅延化 [6]

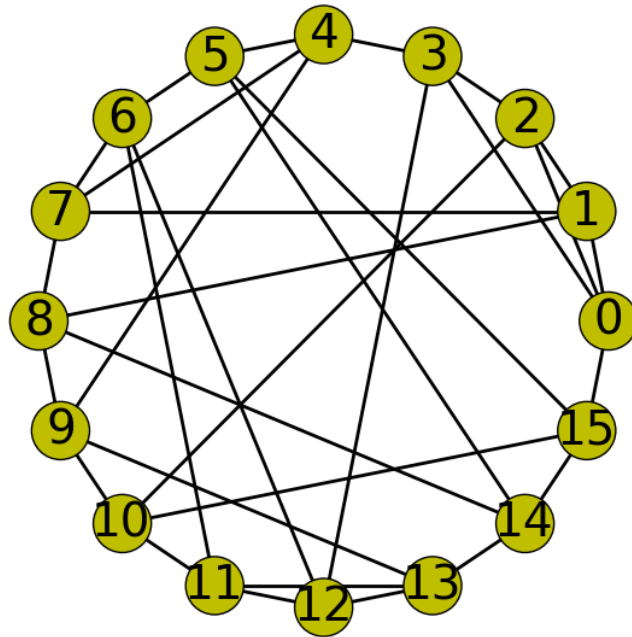


Ring + Non-Random Links
ルータ間トポロジ

Average Shortest Path Hops

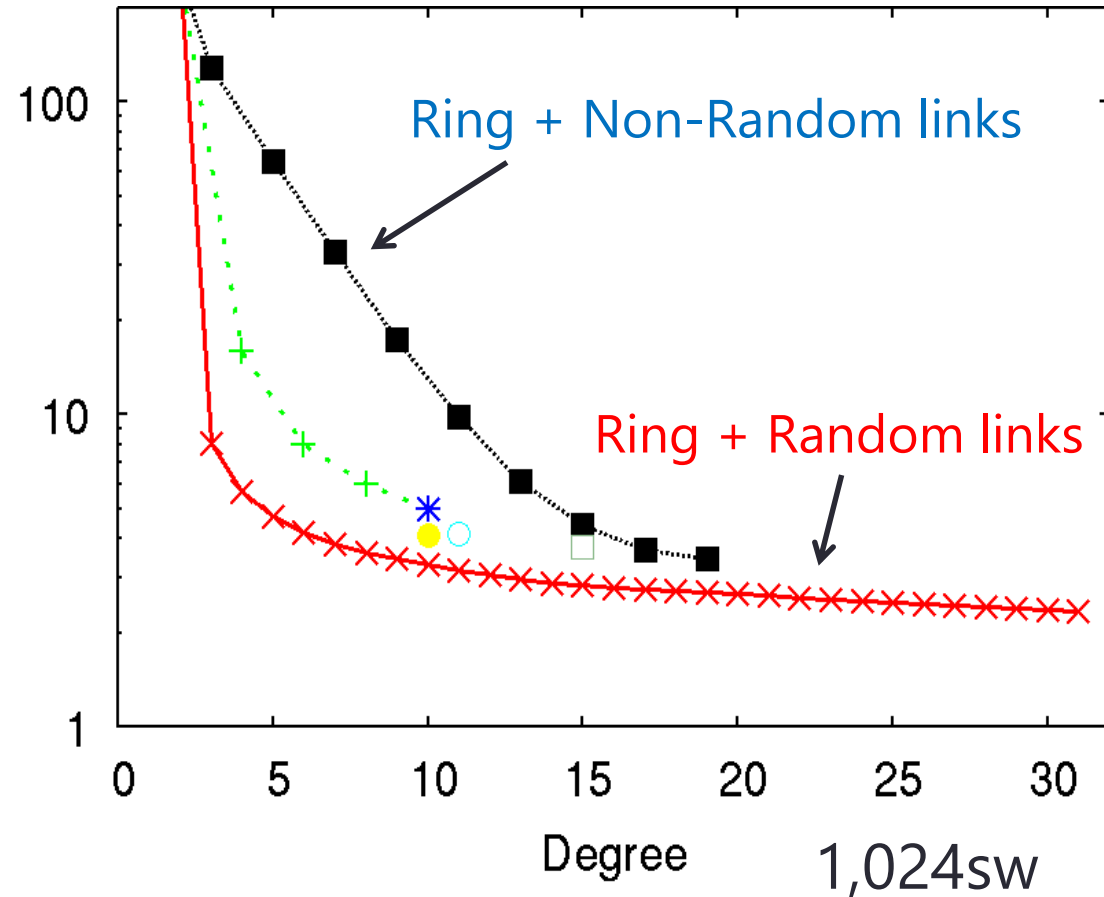


ランダムトポロジによる低遅延化 [6]

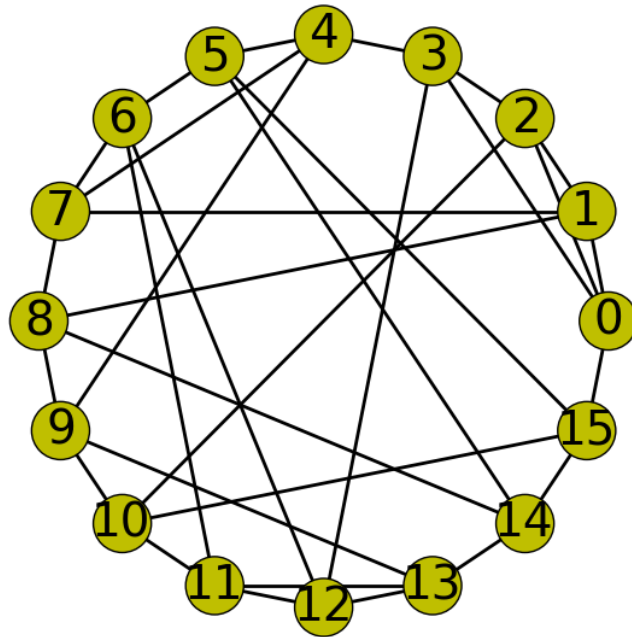


Ring + Random Links
ルータ間トポロジ

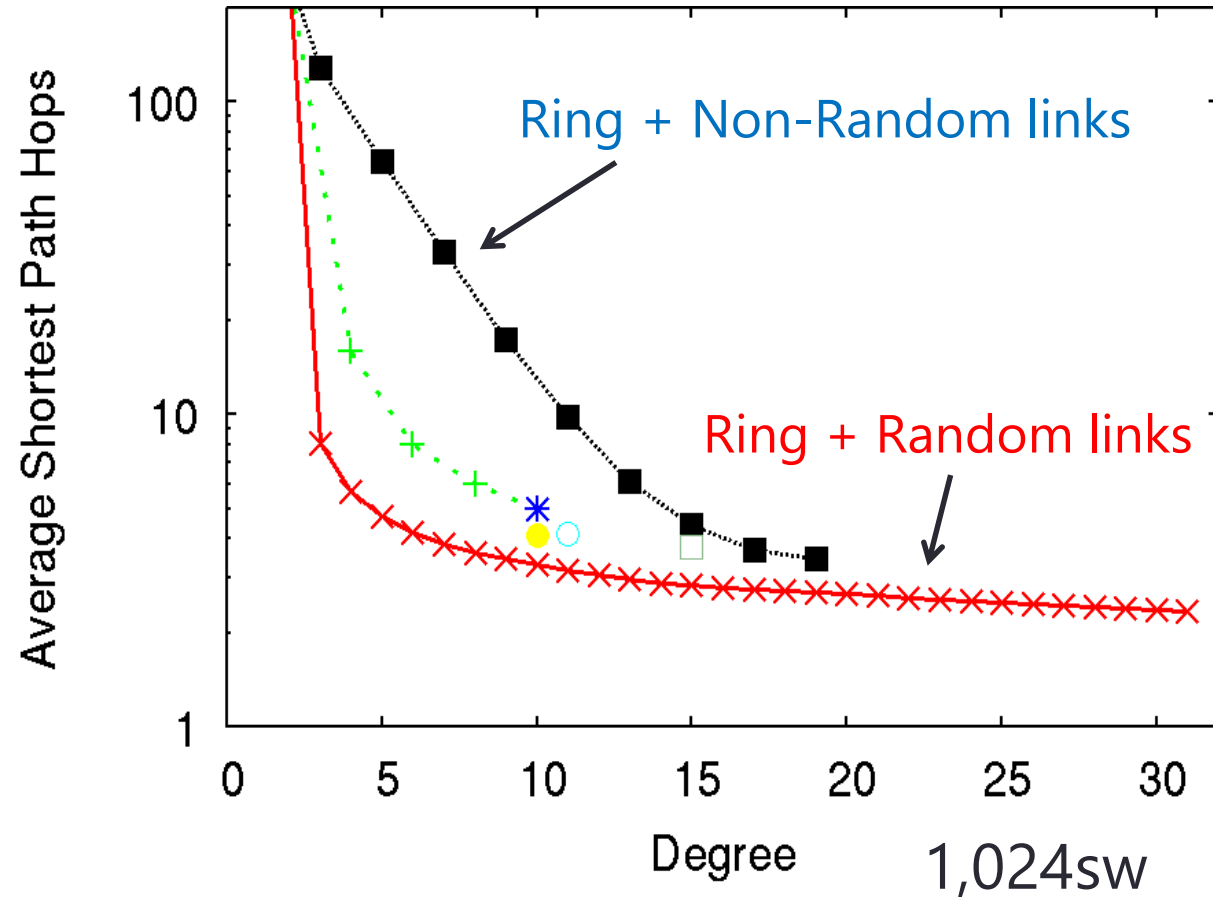
Average Shortest Path Hops



ランダムトポロジによる低遅延化 [6]



Ring + Random Links
ルータ間トポロジ



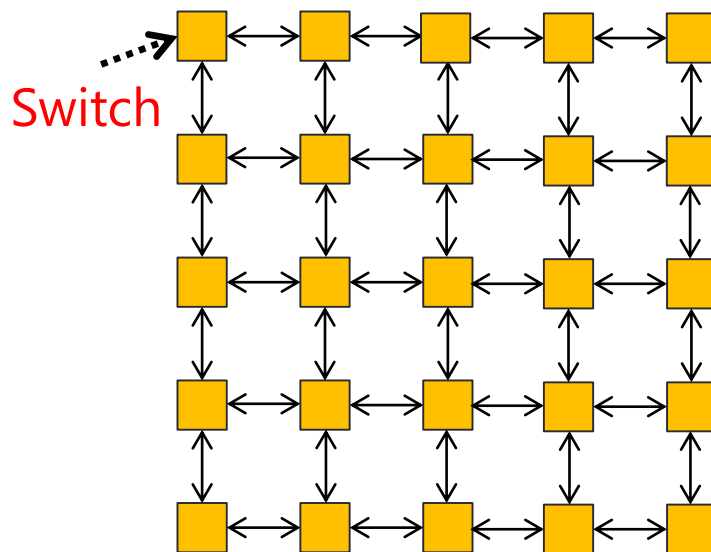
最適化することなくホップ数削減可能

配線長制限ランダムトポロジによるホップ数削減

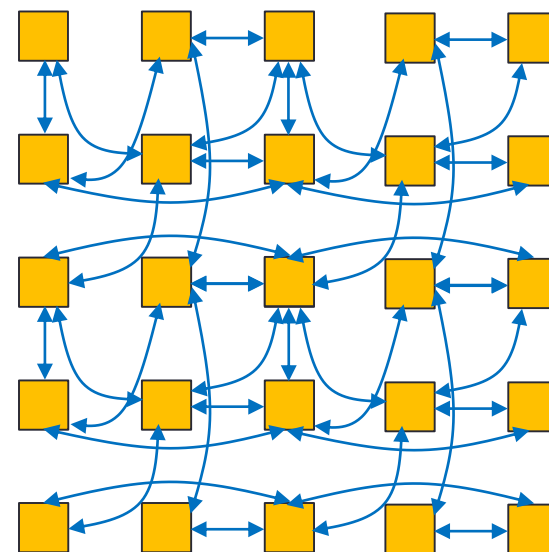
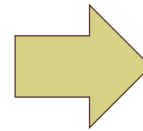
配線長を制限したランダムリンク

- ・ スイッチ間接続に**局所性**が存在
- ・ 配線コスト削減による実用性向上

[7]



Regular Topology



Random Topology

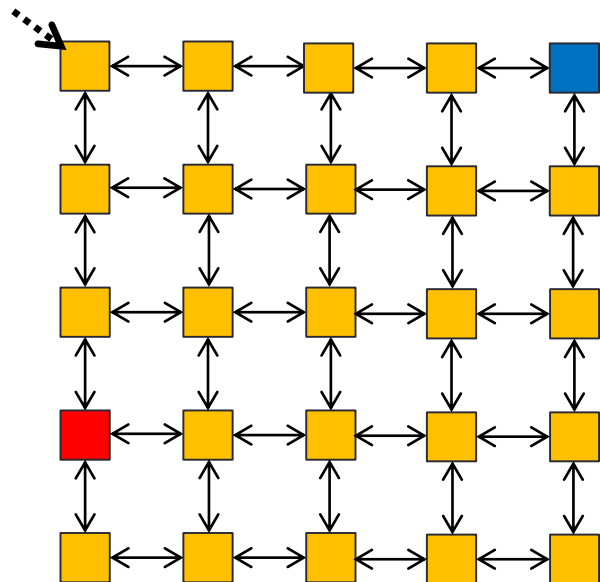
[7] M. Koibuchi, I. Fujiwara, H. Matsutani and H. Casanova: "Layout-conscious Random Topologies for HPC Off-chip In-terconnects", HPCA'13.

配線長制限ランダムトポロジによるホップ数削減

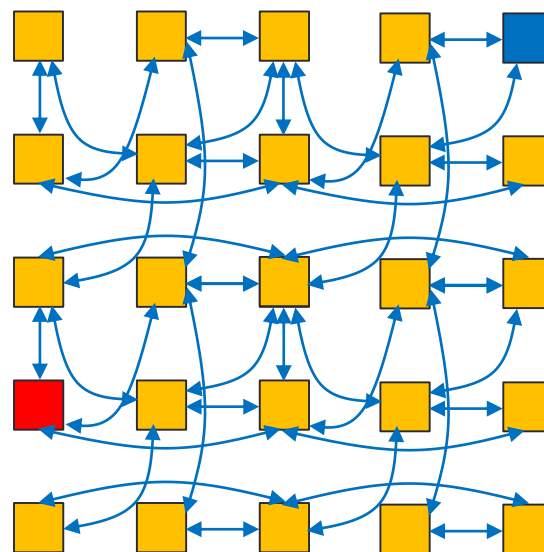
与えられた 送信ノード & 受信ノードに対し,

[7]

Switch



Regular Topology



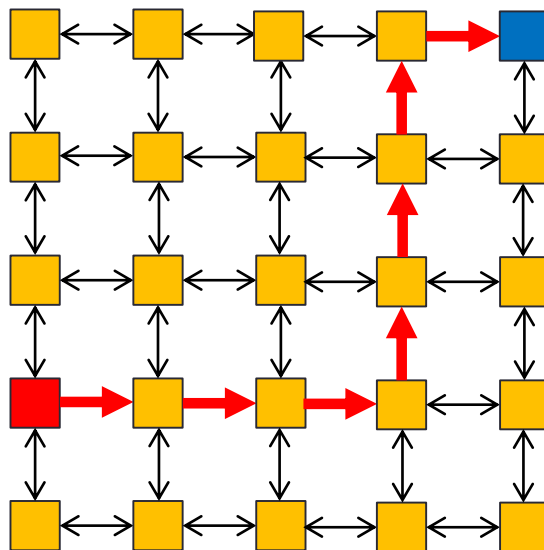
Random Topology

[7] M. Koibuchi, I. Fujiwara, H. Matsutani and H. Casanova: "Layout-conscious Random Topologies for HPC Off-chip In-terconnects", HPCA'13.

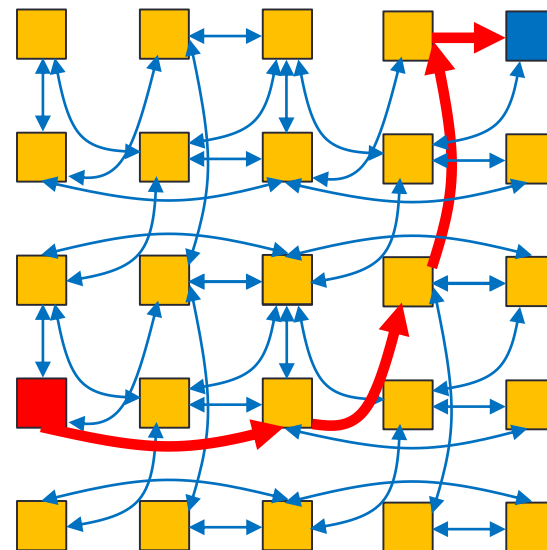
配線長制限ランダムトポロジによるホップ数削減

与えられた 送信ノード & 受信ノードに対し、
低ホップ数経路により遅延削減

[7]



Regular Topology



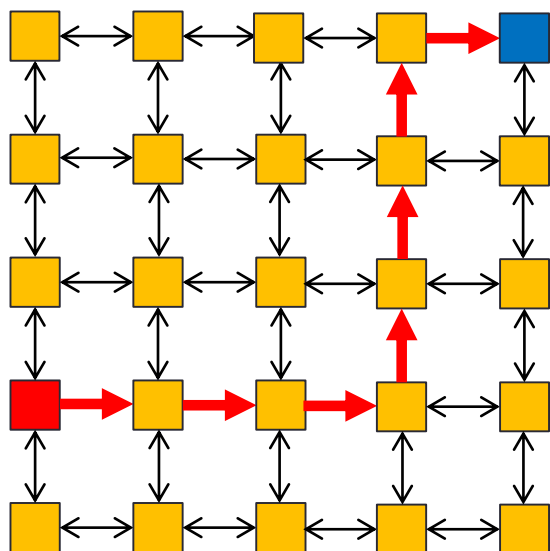
Random Topology

[7] M. Koibuchi, I. Fujiwara, H. Matsutani and H. Casanova: "Layout-conscious Random Topologies for HPC Off-chip In-terconnects", HPCA'13.

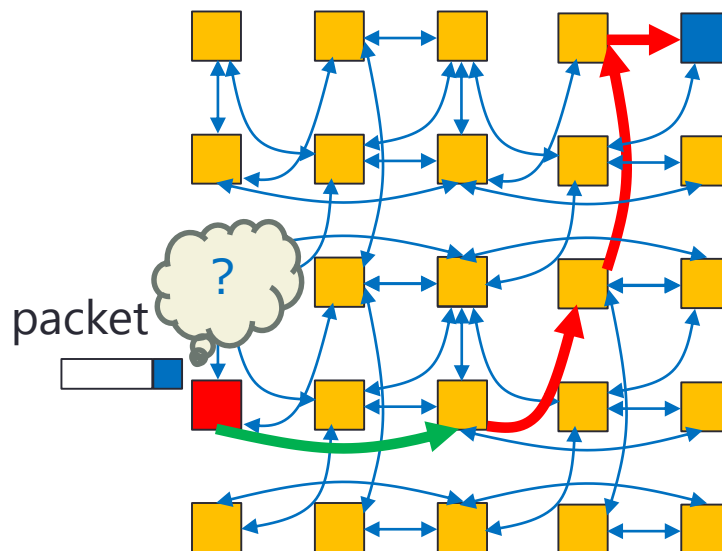
配線長制限ランダムトポロジによるホップ数削減

課題: 宛先ノードへの次ホップをどう知るか？

[7]



Regular Topology

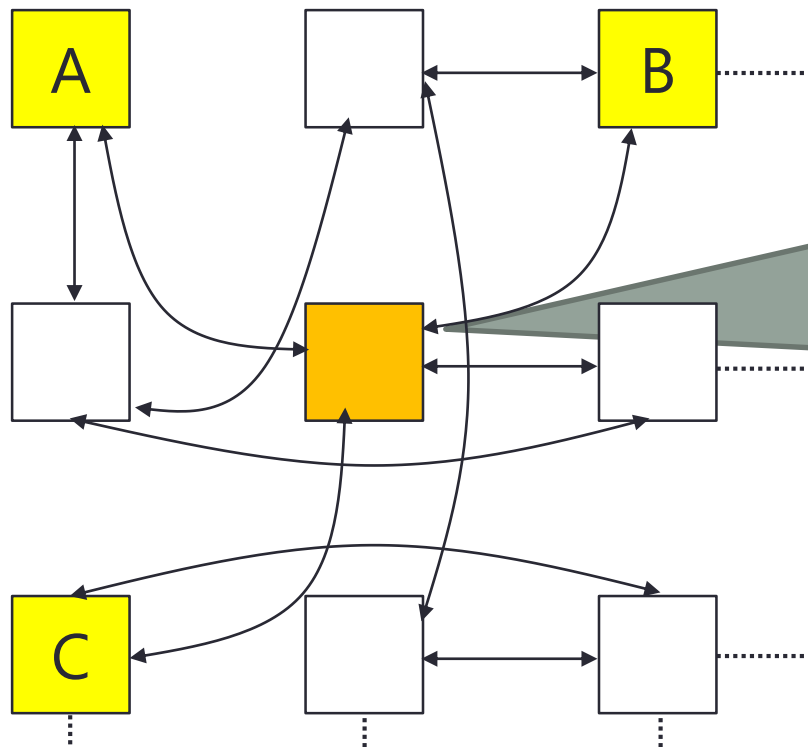


Random Topology

[7] M. Koibuchi, I. Fujiwara, H. Matsutani and H. Casanova: "Layout-conscious Random Topologies for HPC Off-chip In-terconnects", HPCA'13.

従来の不規則網向けルーティング

- 他の全スイッチへの経路の次ホップを保持
- 問題点：テーブルサイズの増大
 - スイッチ当たり $N \log N$ の記憶域



T	
Dst	Next
1	A
2	B
...	..
...	..
N	A

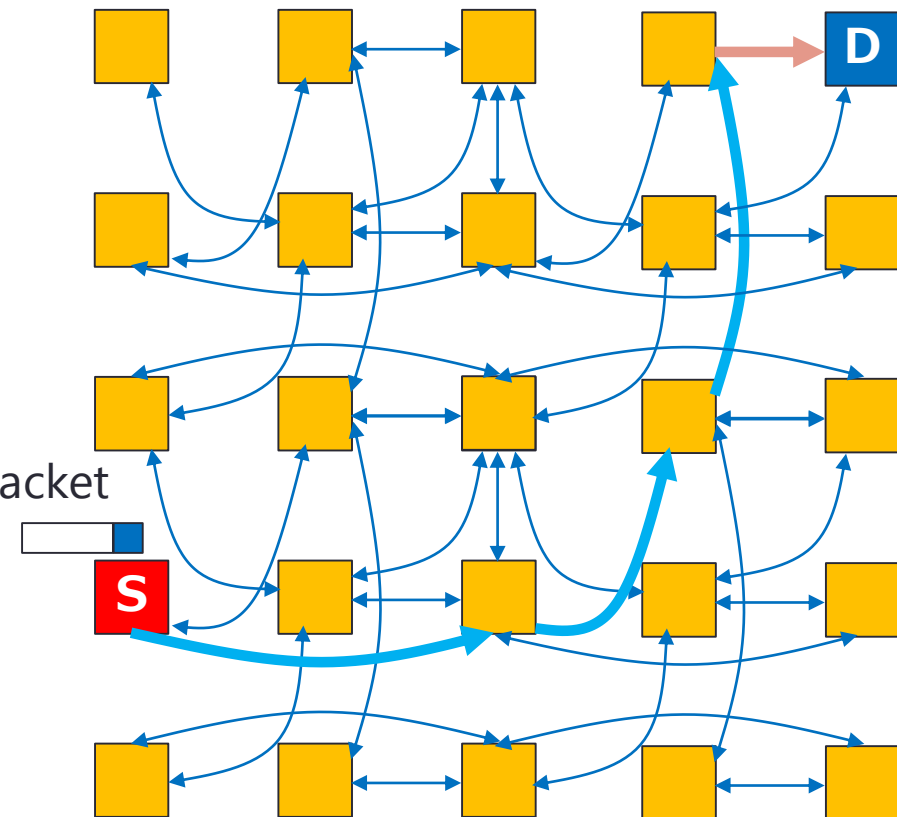
コンパクトルーティング

- ルーティングテーブルサイズとホップ数にトレードオフ
⇒ **コンパクトルーティング**と呼ばれる問題
- 従来の解決手法：ホップ数の最大悪化率を3倍に抑えつつ,
 $O(N^{2/3} \log^{4/3} N)$ を達成するアルゴリズムが存在 [8]
 - 少数の**目印ノード**を設定
 - 各ノードが「**目印ノード**」と「**近傍の複数ノード**」への次ホップを保持
 - 宛先が遠方の場合, 最寄りの**目印ノード**に向けてルーティング
- 問題点：目印ノードの**ホットスポット化**

目次

- スパコン向けネットワーク
 - 低ホップ数なネットワークトポロジの探求
- ランダムネットワーク
 - “近道”を利用したホップ数の削減
 - 配線長制限による実用性向上
 - ルーティングに関する問題
- 分散情報を用いたルーティング手法
 - 近傍・遠方ノードの情報追加によるルーティング
- 評価
 - 通信遅延削減効果
 - テーブルサイズ
- 結論

基本アイディア：近傍・遠方情報の利用



各ノードが、

- ・ 隣接する**近傍ノード**への最短経路
- ・ 一定ホップ先までの**遠方ノード**への最短経路

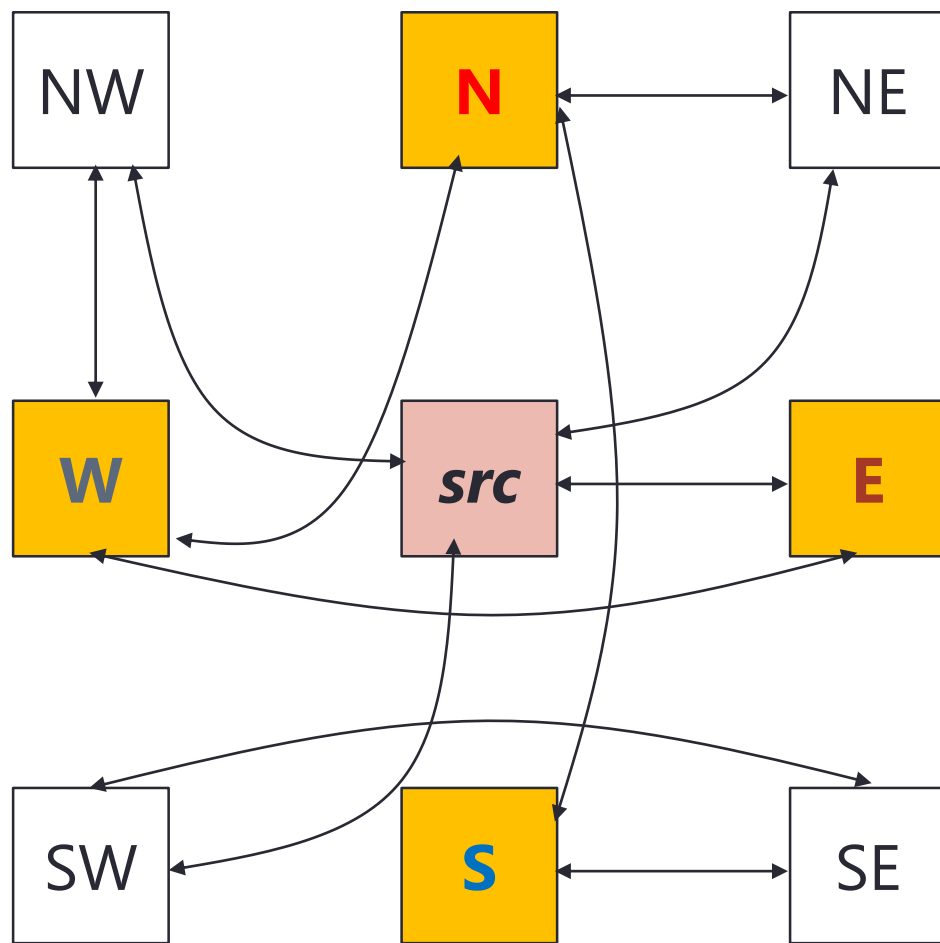
の次ホップを

ルーティングテーブルに保持

目的地が遠い：**遠方ノード**へのエントリー

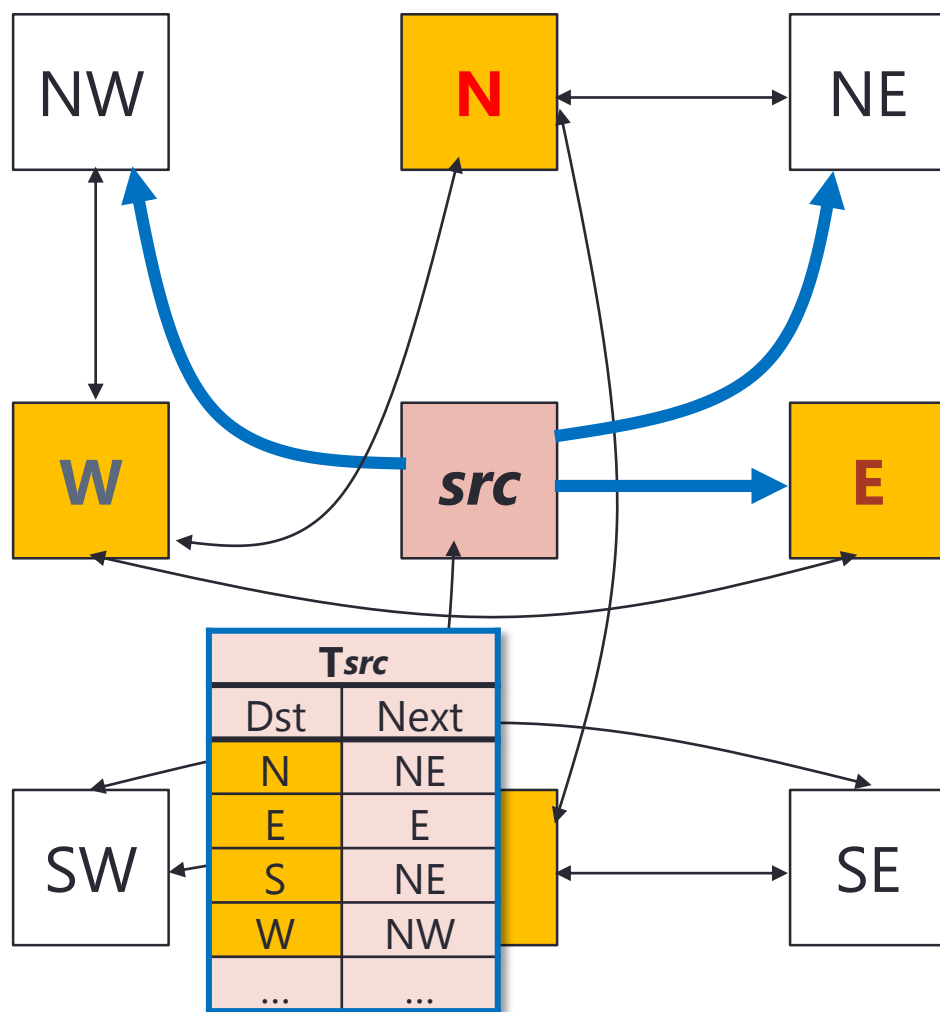
目的地が近い：**隣接ノード**へのエントリー

提案：近傍ノードへの次ホップ情報の追加



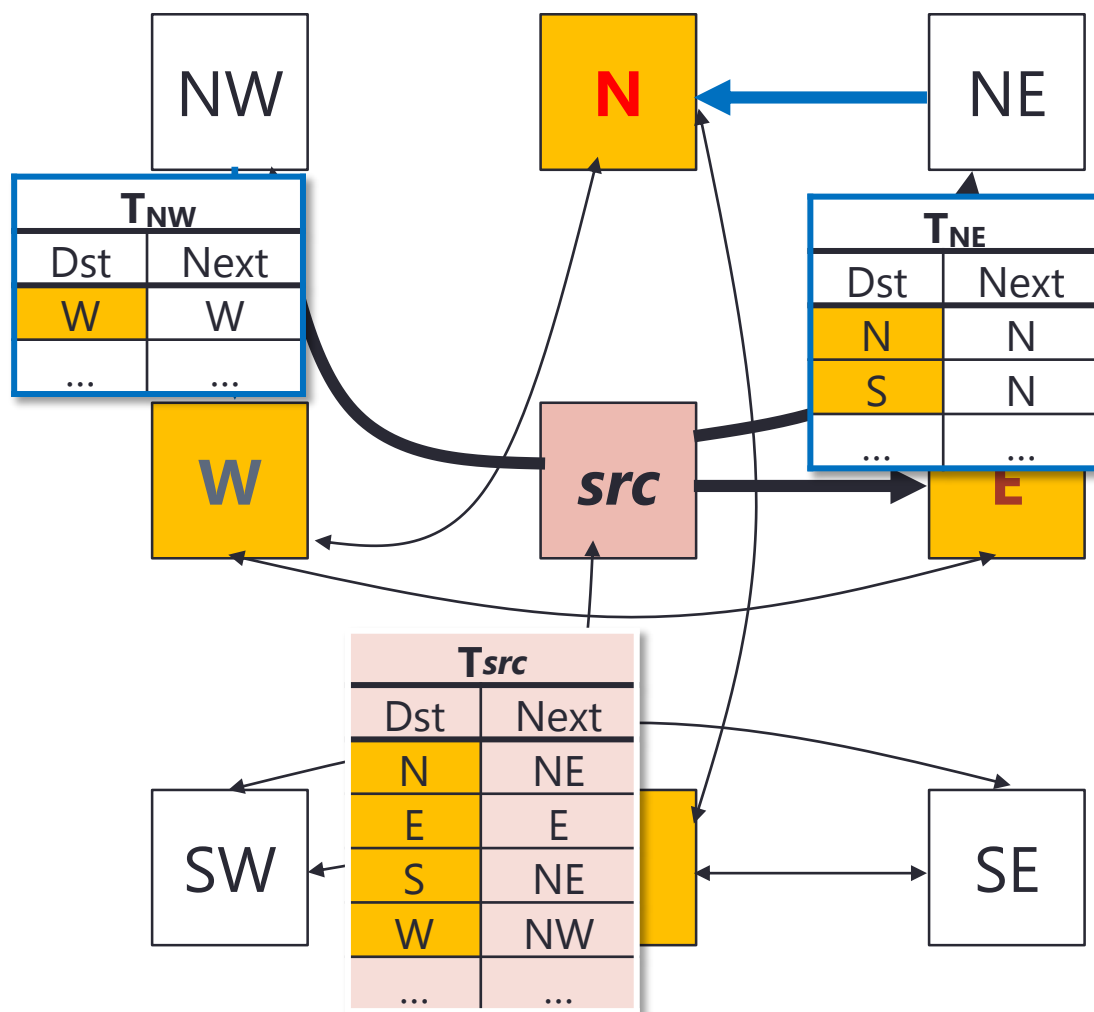
- 各ノードから、座標上で**上下左右に隣接**するノードへの経路情報を保持

ルーティングテーブルへの近傍情報の追加



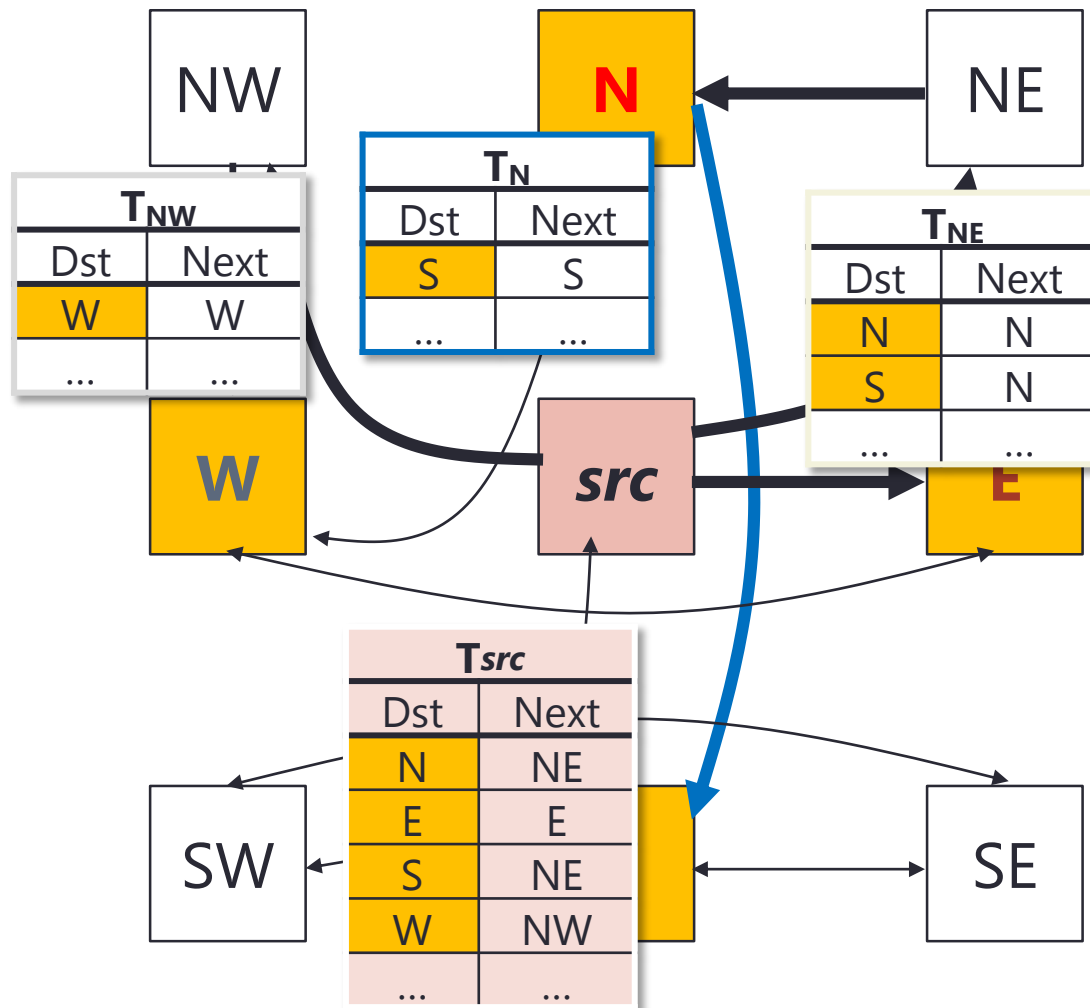
- 各ノードから,
座標上で**上下左右に隣接**する
ノードへの経路情報を保持
- 隣接ノードへの**次ホップ**を
ルーティングテーブルへ追加

ルーティングテーブルへの近傍情報の追加



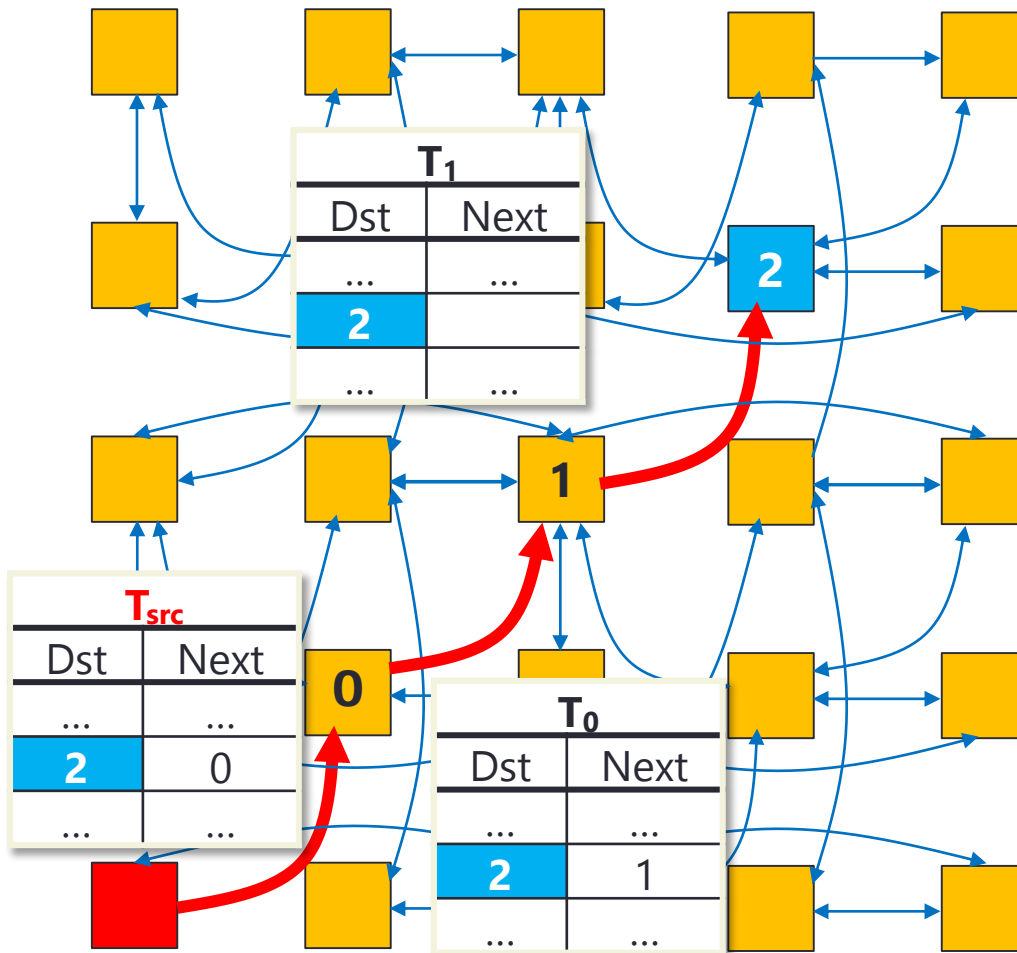
- 各ノードから,
座標上で**上下左右に隣接**する
ノードへの経路情報を保持
- 隣接ノードへの**次ホップ**を
ルーティングテーブルへ追加
- 経路上の**中間ノード**にも
エントリーを追加

ルーティングテーブルへの近傍情報の追加



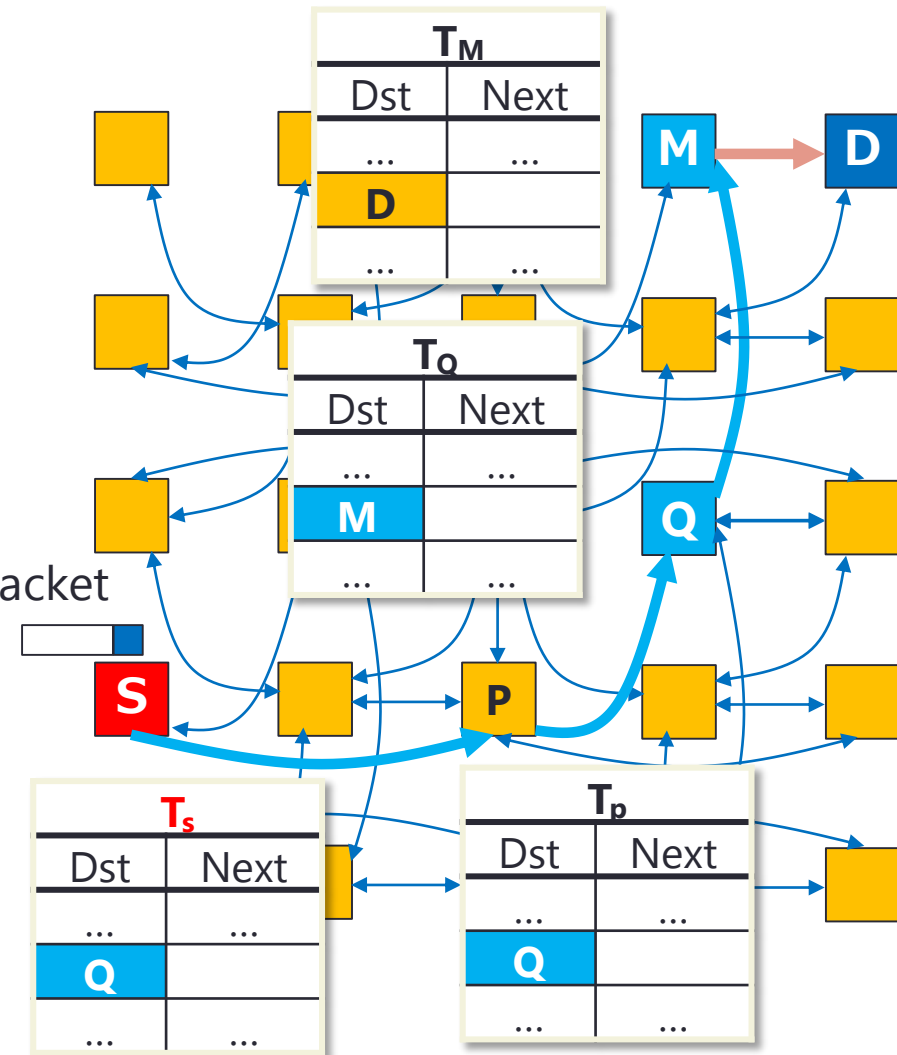
- 各ノードから、座標上で**上下左右に隣接**するノードへの経路情報を保持
- 隣接ノードへの**次ホップ**をルーティングテーブルへ追加
- 経路上の**中間ノード**にもエントリーを追加

提案：ルーティングテーブルへの遠方情報の追加



- ・ 各スイッチから
遠方スイッチへの
経路情報を同様に追加
- ・ 各スイッチから
一定ホップ数内に限定
⇒ 負荷が特定ノードに
集中せず**分散**可能

実際のルーティング例



パケットの宛先スイッチの座標に対し、
最も近い座標へのエントリーを利用

目的地が遠い：遠方ノードへの
エントリー

目的地が近い：隣接ノードへの
エントリー

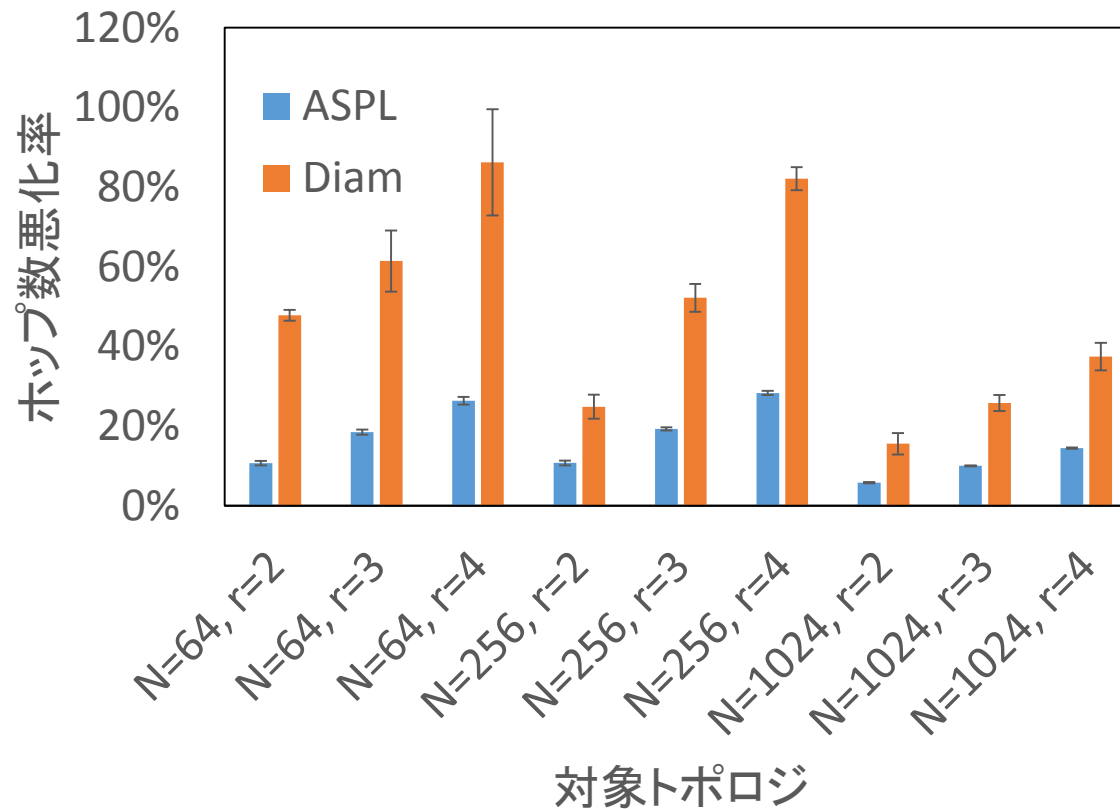
目次

- スパコン向けネットワーク
 - 低ホップ数なネットワークトポロジの探求
- ランダムネットワーク
 - “近道”を利用したホップ数の削減
 - 配線長制限による実用性向上
 - ルーティングに関する問題
- 分散情報を用いたルーティング手法
 - 近傍・遠方ノードの情報追加によるルーティング
- 評価
 - 通信遅延削減効果
 - テーブルサイズ
- 結論

評価1: ホップ数

- ノード数 $N = 64, 256, 1024$
- 配線長制限 $r = 2, 3, 4$
- 次数 $d = 4$
- 遠方ノードの上限ホップ数 = 10
- 上限エントリー数
 - $N = 64$: # of entries = 16; 128 bit/router ($N \log d = 128$)
 - $N = 256$: # of entries = 25; 250 bit/router ($N \log d = 512$)
 - $N = 1024$: # of entries = 85; 850 bit/router ($N \log d = 2048$)

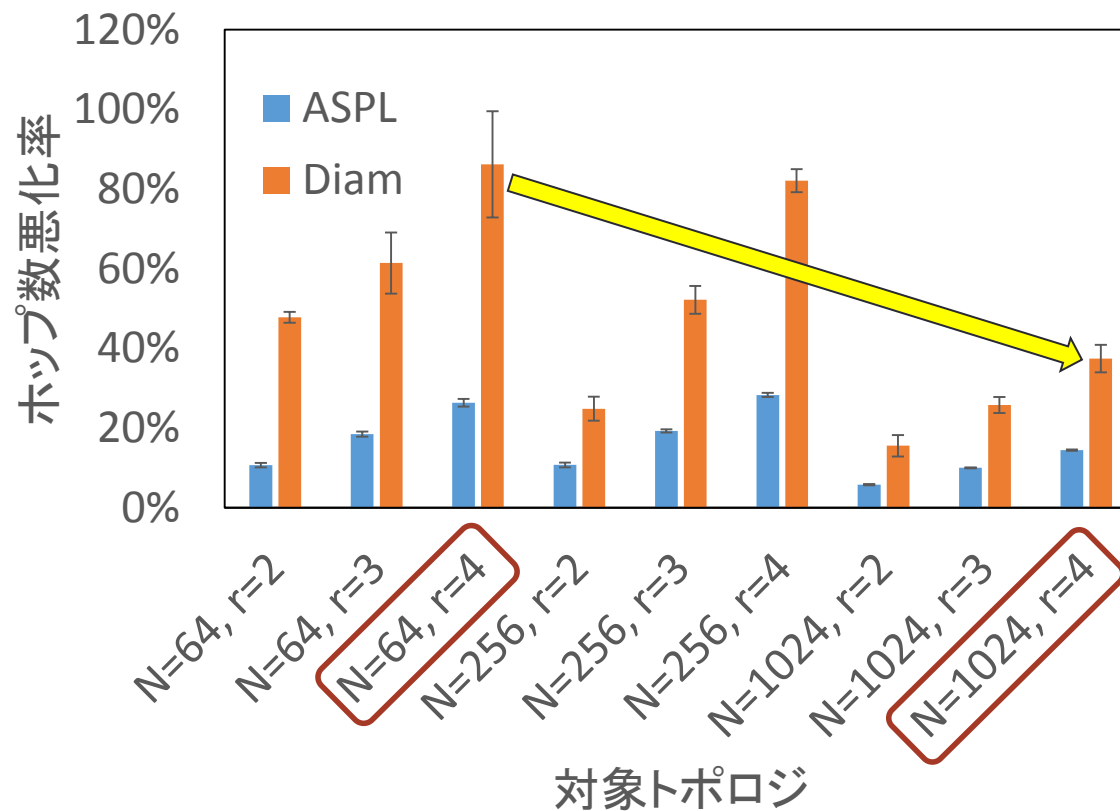
ASPLの悪化率、Diameterの悪化率、 stretch factor (各経路でのホップ数悪化率の最大)



- 再掲: 上限エントリー数

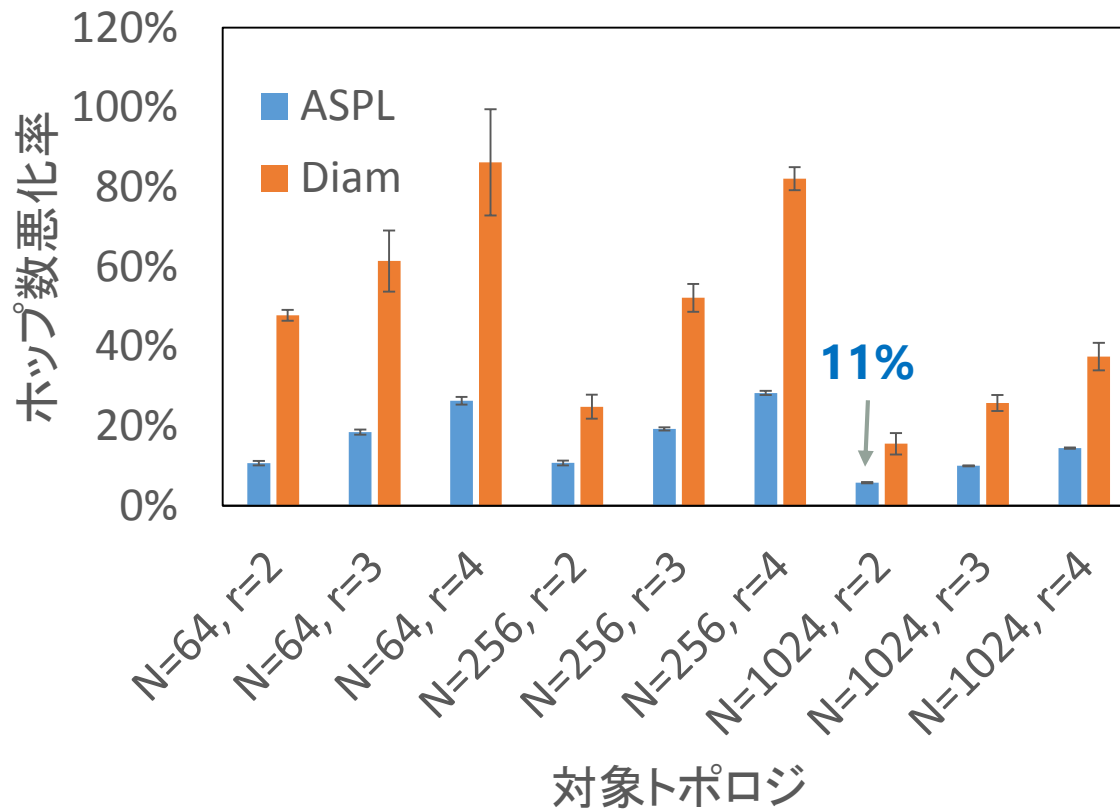
- $N = 64$: 128 bit/router ($N \log d = 128$)
- $N = 256$: 250 bit/router ($N \log d = 512$)
- $N = 1024$: 850 bit/router ($N \log d = 2048$)

ASPLの悪化率、Diameterの悪化率、 stretch factor (各経路でのホップ数悪化率の最大)



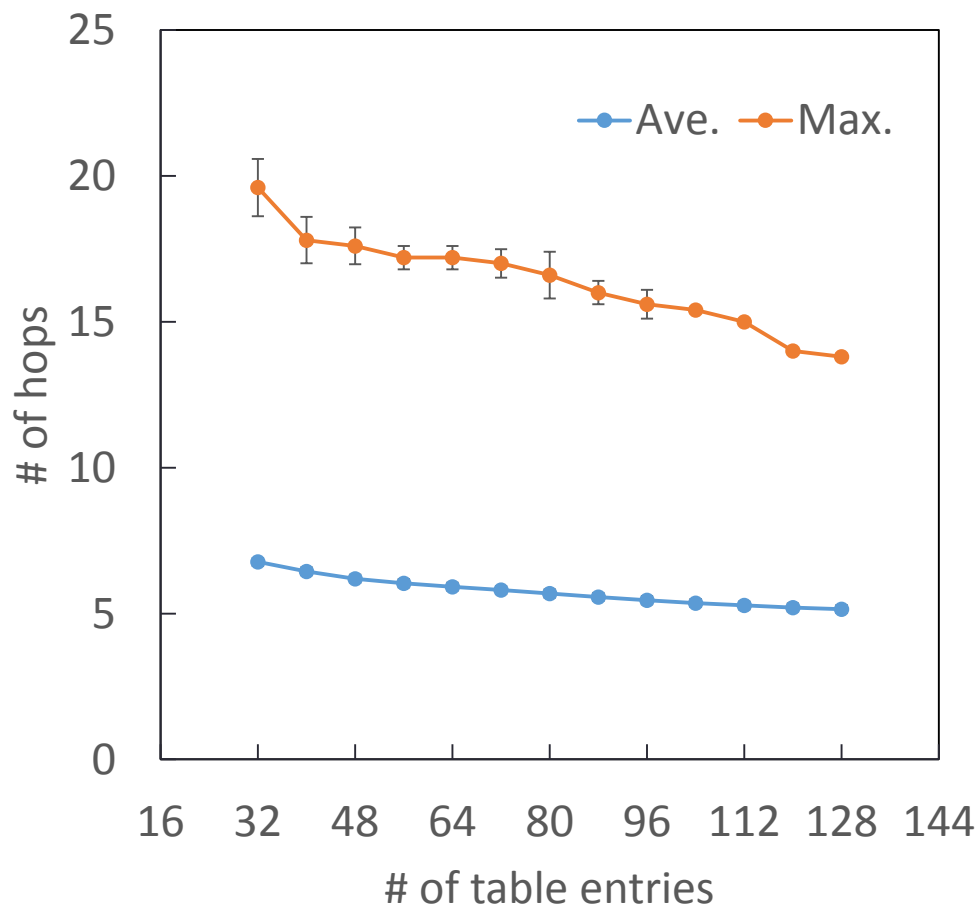
- 再掲: 上限エントリー数
 - $N = 64$: 128 bit/router ($N \log d = 128$)
 - $N = 256$: 250 bit/router ($N \log d = 512$)
 - $N = 1024$: 850 bit/router ($N \log d = 2048$)
- 全経路保持の場合
($N \log d$) より小さいス
ケール増加で悪化率改善

ASPLの悪化率、Diameterの悪化率、 stretch factor (各経路でのホップ数悪化率の最大)



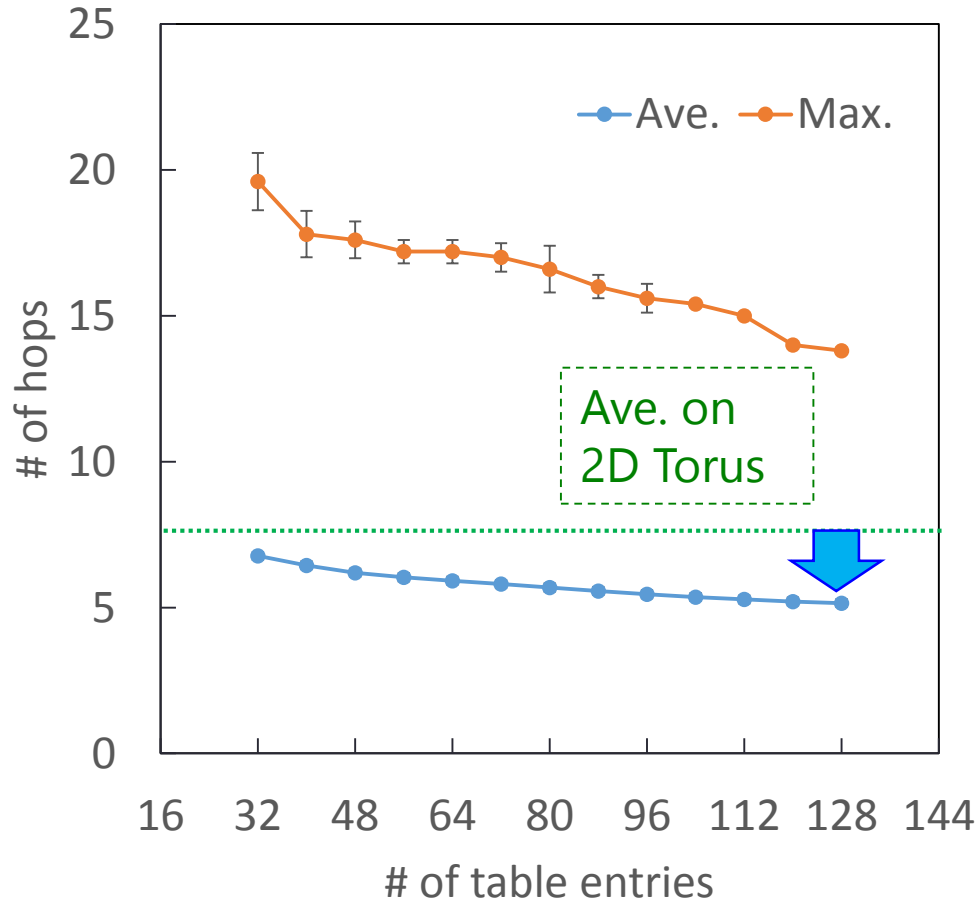
- 再掲: 上限エントリー数
 - $N = 64$: 128 bit/router ($N \log d = 128$)
 - $N = 256$: 250 bit/router ($N \log d = 512$)
 - $N = 1024$: 850 bit/router ($N \log d = 2048$)
- 全経路保持の場合
($N \log d$) より小さいスケール増加で悪化率改善
- 1024ノードにおいて
 - ホップ数悪化を**11%**に抑制
 - テーブルサイズを**51%**削減

評価2: テーブルサイズと達成遅延



- $N = 256$ nodes
- max. link length = 8
- テーブルサイズ緩和による影響
 - 最大遅延への削減効果大
 - 平均遅延への削減効果小

評価2: テーブルサイズと達成遅延



- $N = 256$ nodes
- max. link length = 8
- テーブルサイズ緩和による影響
 - 最大遅延への削減効果大
 - 平均遅延への削減効果小
- 同ノード数・同次数の 2D Torus に比べ、平均遅延を最大**36%** 改善

結論

- 対象：スパコン向けNW
- 目的：idealのホップ数に近い
低遅延ルーティング
- 解決法：配線長制限ランダムトポロジに対する分散ルーティング
 - “近道”を用いたホップ数削減
 - 近傍と数ホップ先までの遠方情報により
効率的なルーティングを実現
- 成果：遅延削減＋スケーラビリティ
 - Torus網に比べ遅延を **36%** 改善
 - ルーティングテーブルを **51%** 削減

