情報処理学会論文誌　コンピューティングシステム　Vol.14　No.3　1–13 (Dec. 2021)

二重同型Hypercubeネットワーク

細見 岳生[1,2,a)]　安戸 僚汰[3,b)]　鯉渕 道紘[4,c)]　下條 真司[2,d)]

受付日 2021年3月31日, 採録日 2021年9月30日

概要：HPC システムにおいて，ノード間を接続するネットワークの高性能化の重要性は GPU などのアクセラレータの採用により増している．近年その高性能化の要望に応えるために，多数のシステムが複数のネットワーク・プレーンを持つ Multi-Plane 方式を採用している．本論文では，この Multi-plane 方式においてネットワークの高性能化を実現する二重同型ネットワークを提案する．この二重同型ネットワークは，各プレーンをスイッチ間の接続が異なるグラフ同型のネットワークで接続することを特徴とする．Hypercube および Foleded-Hypercube を題材にして二重同型ネットワークの構成法を示し，またそれらの性能評価を実施した．その結果，適切な同型ネットワークの選択により，同一のネットワークを二重化した場合と比較して．ネットワークの経済的コストを増加させることなく遅延およびスループットを改善できることが明らかとなった．

キーワード：ネットワーク，ハイパーキューブ，多重 Plane，遅延，スループット

Dual-plane Isomorphic Hypercube Networks

Takeo Hosomi[1,2,a)]　Ryouta Yasudo[3,b)]　Michihiro Koibuchi[4,c)]　Shinji Shimojo[2,d)]

Received: March 31, 2021, Accepted: September 30, 2021

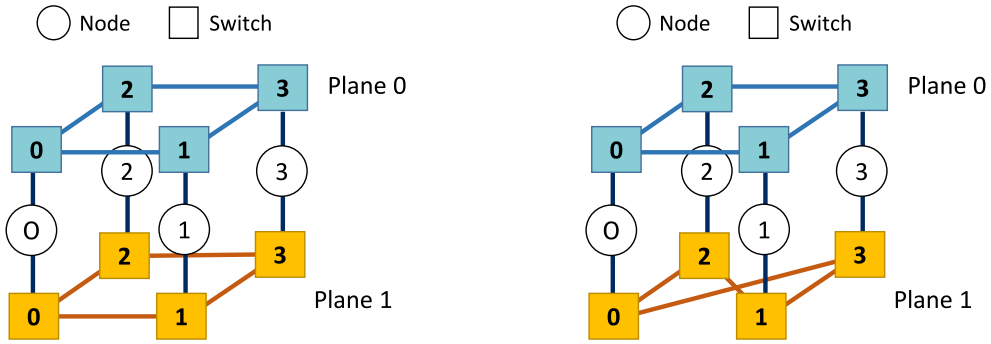
Abstract: The importance of improving the performance of interconnection networks in HPC systems is increasing due to the adoption of high performance accelerators such as GPUs. In recent years, in order to meet the demand for higher performance networks, the systems have adopted a multi-plane network. In this paper, we propose a dual isomorphic network that improves the network performance in the multi-plane system. This dual isomorphic network is characterized by connecting each plane with a graph isomorphic network that has different connections between switches. We discuss the dual-plane isomorphic Hypercube and the dual-plane isomorphic Foleded-Hypercube. The evaluation results show that this scheme can improve latency and throughput without increasing economic costs of network by choosing the appropriate isomorphic network.

Keywords: network, hypercube, multi-rail, multi-plane, isomorphic, latency, throughput

1. はじめに

HPC システムにおいて，プロセッサ間を接続するネットワークの高性能化は重要な課題である．プロセッサ性能やメモリバンド幅は向上を続けており，それに対応したネットワークの低遅延化と高スループット化が求められている．また近年 GPU などのアクセラレータの活用により HPC システムの 1 ノードあたりの性能が飛躍的に向上している．そのため，1 ノードが複数のネットワーク・ポートを持つ Multi-Rail 方式を採用し，処理性能向上に見合ったネットワークの高スループット化を実現するシステムも増加している [24], [25], [26].

[1] 日本電気株式会社
NEC Corporation, Kawasaki, Kanagawa 211–8666, Japan
[2] 大阪大学
Osaka University, Ibaraki, Osaka 567–0047, Japan
[3] 広島大学
Hiroshima University, Higashihiroshima, Hiroshima 739–8511, Japan
[4] 国立情報学研究所
National Institutes of Informatics, Chiyoda, Tokyo 101–8430, Japan
[a)] takeo.hosomi@nec.com
[b)] yasudo@cs.hiroshima-u.ac.jp
[c)] koibuchi@nii.ac.jp
[d)] shimojo@cmc.osaka-u.ac.jp



(a) 二重 2 次元 Hypcube ネットワーク
(a) Dual-plane 2-Hypercube Network.

(b) 二重同型 2 次元 Hypcube ネットワーク
(b) Dual-plane Isomorphic 2-Hypcube Network.

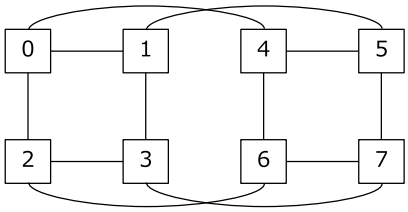
図 1　二重 Hypercube ネットワークと二重同型 Hypcube ネットワークの例
Fig. 1　Example of dual-plane hypcube network and dual-plane isomorphic hypercube network.

従来，Multi-Rail を持つシステムにおいては，2 つの方式が存在した．1 つは，ネットワークが 1 つのプレーンからなり，ノードが持つ複数のポートを 1 つのネットワーク・プレーンに接続してスループットを向上させる Port-Aggregate 方式である．この方式は，スループット向上が図れるものの，ネットワーク・プレーンのサイズがポート数にノード数をかけた大きさが必要となりトポロジや接続するノード数によっては平均最短距離が悪化する問題があった．もう 1 つは，ネットワークが複数のプレーンからなり，各ポートをそれぞれ異なるプレーンに接続する Multi-plane 方式である．この方式では，各プレーンは同一であり，均等にプレーンを利用すればスループットをプレーン数倍にすることができる．また，どのプレーンを利用してもノード間の距離は同一であるため，平均最短距離は単一プレーンと同じとなる．

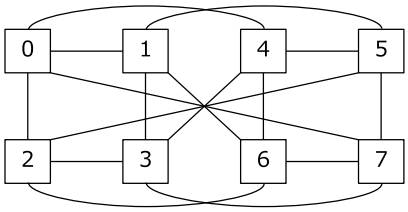
本論文では，この Multi-plane 方式においてネットワークをさらに高性能化する二重同型ネットワークを提案する．二重同型ネットワークは，各プレーンに同一ではなく同型のネットワークを採用することで，遅延を削減しスループットを向上させるものである．ここで同型とは，グラフ同型のトポロジでありスイッチ間接続の組合せが異なるものである．**図 1** に従来の Multi-plane 方式での二重 2 次元 Hypercube ネットワークと，提案している二重同型 2 次元 Hypercube ネットワークを例として示す．図に示すように，二重 2 次元 Hypercube ネットワークでは．プレーン 0 とプレーン 1 ともに，ノードとスイッチ間．およびスイッチのスイッチ間の接続が同一の構成をとる．一方，二重同型 2 次元 Hypercube ネットワークでは，プレーン 0 とプレーン 1 ともに同じ Hypercube トポロジではあるが，異なるスイッチ間の接続をとる．同一のネットワークを二重化する従来の Multi-plane 方式と比較して，このような二重同型ネットワークを採用することで，あるノードへの距離がそれぞれのプレーンで異なるものとなるため，パケットを送出する際に距離が短い方のプレーンを選択することで，遅延を改善することが期待できる．また，そのような選択を行うことで，パケットを送信する場合に利用するスイッチ間リンク数を削減することができるため，スループットの向上も同時に期待できる．

本論文では，Hypercube ネットワーク [10] と Folded-Hypercube ネットワーク [11] の 2 つのよく知られたトポロジに対して，提案する二重同型ネットワークを適用し議論する．Hypercube ネットワークは HPC システムにおいて重要なトポロジの 1 つであり，現在の HPC システムでも広く用いられている [24], [27]．またこのようなシステムでの標準的なインターコネクトである InfiniBand [22] でもサポートされている．我々の先行研究 [29] では，これらのネットワークに関する初期検討を行い，グラフ解析で平均最短距離を削減しスループットを向上させる効果を持つことを確認した．ただし，サイクルレベルのシミュレーションでは，スループットは同様の効果を示すものの，システム規模が大きくなると平均最短距離が小さくなるにもかかわらずスイッチ間のケーブル長が伸びケーブル遅延が増大しノード間の遅延が悪化するという知見も得た．本論文では先行研究 [29] を発展させて，二重同型 Hypercube ネットワークのスイッチ間接続を決める際にケーブル遅延を考慮した最適化方式を新たに加え，総合的な評価を通して有効性を明らかにする．

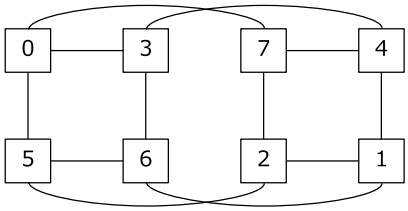
本論文の構成は以下である．2 章において二重同型 Hypercube ネットワークについてグラフ解析を行い，平均最短距離やスループットについて評価した結果を示す．また，3 章において実システムに適用した場合の評価として，ケーブル遅延を考慮した二重同型 Hypercube ネットワークの最適化方式を提案し，ネットワークの経済的コストの解析，サイクルレベルシミュレーションによる性能評価を行う．4 章で関連研究を述べ，最後に 5 章において結論を述べる．



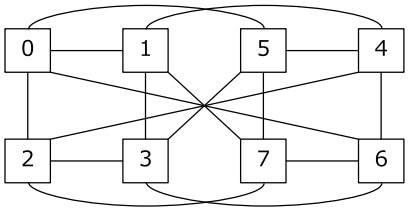
(a) 3-Hypercube $H(0,3) = \{1,2,4\}$



(c) 3-Folded-Hypercube $FH(0,3) = \{1,2,4,7\}$



(b) Isomorphic 3-Hypercube $H(1,3) = \{3,5,7\}$



(d) Isomorphic 3-Folded-Hypercube $FH(1,3) = \{1,2,5,6\}$

図 2　二重同型 Hypercube（DI-HC）および Folded-Hypercube（DI-FHC）の接続例（$n = 3$）

Fig. 2　Example of dual-plane isomorphic hypercube (DI-HC) and folded-hypercube (DI-FHC) ($n = 3$).

2. グラフ解析

この節では，二重 Hypercube（D-HC），二重同型 Hypercube（DI-HC），二重 Folded-Hypercube（D-FHC），二重同型 Folded-Hypercube（DI-FHC）の 4 つのネットワークを比較する．まずは同型ネットワークの生成方法と二重同型ネットワークでのルーティングについて述べる．次に，多数の選択肢が存在する同型ネットワークの中から最適なものを選択するために 2 つの性能指標を導入する．1 つは，平均最短距離であり遅延を表すものとして用いる．もう 1 つはスループットを表す指標として全対全最大トラフィックを新たに定義しそれを用いる．最後に，様々なネットワーク・サイズでの解析結果を示す．なお，本章では，簡単のためにスイッチあたりのノード数を 1 として検討を行った．

2.1 二重同型 Hypercube および二重同型 Folded-Hypercube の生成

まず，DI-HC の生成について述べ，次に DI-FHC の生成について述べる．二重同型では，両プレーンにおいて異なるスイッチ間接続をとることを特徴とする．ここで，各ノードには $0 \sim 2^n - 1$ の ID が割り当てられ，両プレーンのスイッチにも $0 \sim 2^n - 1$ の ID を割り当てられる．ノードと接続する両プレーンのスイッチの関係はどの同型においても同一で，同じ ID を持つ者どうしが接続するものとする．DI-HC のプレーン 0 には一般的な Hypercube 接続を用い，スイッチ ID のハミング距離が 1 のスイッチどうしを接続することによって構成する．一方プレーン 1 には，前記プレーン 0 とは異なるスイッチどうしを接続して同型 Hypercube を構築する．

ここで，Hypercube のスイッチ間の接続関係に着目して，それぞれのプレーンの接続を表す $H(0,n)$ と $H(1,n)$ を定義する．プレーン 0 の n-Hypercube の接続は集合 $H(0,n) = \{1, 2, \ldots, 2^{n-1}\}$ で定義できる．各スイッチは自 ID と $H(0,n)$ の各要素とビットごとの排他的論理和をとって得られる ID を持つ n 個のスイッチと接続する．すなわち，ID が x のスイッチは，以下の式で求められるスイッチ集合 Y と接続する．$Y = \{x \oplus h \,|\, h \in H(0,n)\}$ たとえば，**図 2** (a) に示す 3 次元 Hypercube は $H(0,3) = \{1,2,4\}$ で定義され，ID が 1 のスイッチは，ID が $\{0,3,5\}$ のスイッチと接続する．$H(1,n)$ も要素数 n の集合で，各要素は 1 以上 2^n 未満の自然数となる．この $H(1,n)$ を用いて，プレーン 1 の ID が x のスイッチは，以下の式で求められるスイッチ集合 Y と接続する．$Y = \{x \oplus h \,|\, h \in H(1,n)\}$ ここで，構成される Network が 1 つの連結グラフとなる $H(1,n)$ のみが同型 Hypercube となる．たとえば，図 2 (b) に同型 3 次元 Hypercube $H(1,3) = \{3,5,7\}$ の例を示す．この同型 Hypercube ではたとえば ID が 1 のスイッチは ID が $\{2,4,6\}$ のスイッチと接続する．このように $H(0,n)$ と $H(1,n)$ で定義されるグラフ同型の Hypercube を利用することで，DI-HC を構成する．

DI-FHC も同様に構成できる．Folded-Hypercube は，Hypercube の各頂点に対して最も遠い距離の頂点と接続するリンクを追加するものである．Hypercube と同様にプレーン 0 の Folded-Hypercube は頂点間の接続関係から以下の $FH(0,n) = \{1, 2, \ldots, 2^{n-1}, 2^n - 1\}$ で定義できる．またこの Folded-Hypercube において，ID が x のスイッチは，以下の式で求められるスイッチ集合 Y と接続する．$Y = \{x \oplus h \,|\, h \in FH(0,n)\}$ 図 2 (c) に 3 次元 Folded-Hypercube の接続を示す．この例では，

表 1　ルーティング用のテーブル（図 2 (a)，(b) に示す二重同型 3 次元 Hypercube の例）

Table 1　Routing table in the nodes for dual-plane isomorphic 3-hypercube shown in Fig. 2 (a), (b).

Src.ID ⊕ Dst.ID	Dist.@ Plane 0	Dist.@ Plane 1	Routing Value @ Plane 1
0	0	0	0
1	1	3	7
2	1	2	6
3	2	1	1
4	1	2	5
5	2	1	2
6	2	2	3
7	3	1	4

$FH(0,3) = \{1,2,4,7\}$ であり，ID が 1 のスイッチは ID が $\{0,3,5,6\}$ のスイッチと接続する．DI-FHC の $FH(1,n)$ も，最終要素を除いた n 個の要素について DI-HC の構築と同様の規則で生成する．Folded で追加されるリンクについては，求めた n 個の値のビットごとの排他的論理和で求められる値となる．図 2 (d) に 3 次元同型 Folded-Hypercube の接続を示す．この例では，$FH(0,3) = \{1,2,5,6\}$ であり，ID が 1 のスイッチは ID が $\{0,3,4,7\}$ のスイッチと接続する．

ここで，$H(1,n)$ や $FH(1,n)$ には多数の選択肢が存在するが，以降で述べる評価指標を用いて効果の大きい集合を選択する．

2.2　ルーティング

DI-HC および DI-FHC の最短パスルーティングは，以下の 2 つのステップからなる．

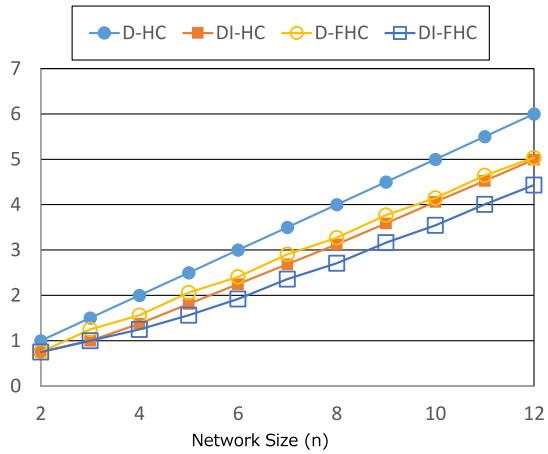
(1) ノードは距離が短いプレーンを選択しそのプレーンにパケットを送出する．同一の場合はどちらかを選択する．
(2) 選択したプレーン内で各スイッチは最短パスルーティングを行う．

ここで，各ノードは**表 1** に示すルーティング用のテーブルを保有する．表はノード数分のエントリからなり，ソースとあて先ノード ID のビットごとの排他的論理和をとって得られる値をインデックスとして，プレーン 0 での距離，プレーン 1 での距離，プレーン 1 でのルーティング用の値を保持する．以降表 1 に示した DI-HC を例に説明するが，DI-FHC においても同様の考え方でルーティングが可能である．

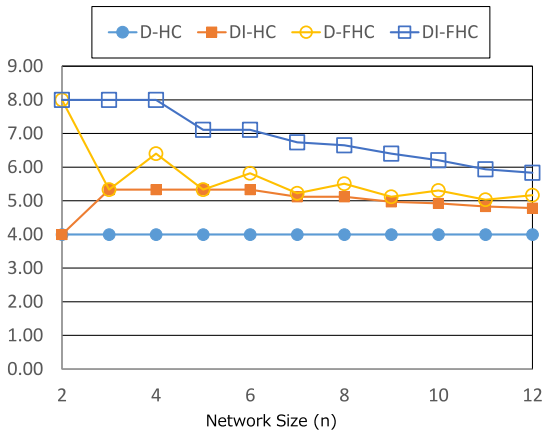
ステップ (1) においては，表 1 を用いて，あて先ノードへの距離をそれぞれのプレーンについて読出し，値を比較することによって決定する．ステップ (2) におけるプレーン 0 のルーティングは，既存の次元オーダの最短パスルーティングが利用可能である．次元オーダのルーティングでは，ソースとあて先ノード ID のビットごとの排他的論理和をとって得られる値を用い，下位から 1 が立っているビットを順に選択し対応するリンクをたどることで実現される．一方プレーン 1 では，前節で示したようにプレーン 0 とは異なる接続となるため上記のプレーン 0 と同じ方法でルーティングすることはできない．そのため，両プレーンがグラフ同型であることを利用し，ソースとあて先ノード ID のビットごとの排他的論理和をとって得られる値ではなく，それを変換した値を用いることでプレーン 0 と同様の次元オーダのルーティングを可能となる．この変換は，プレーン 1 でのあるスイッチ ID が，プレーン 0 のどのスイッチ ID と同じ位置にあるかを求めるものである．これにより，プレーン 1 でのあるスイッチ ID へのルーティングパスが，プレーン 0 ではどのスイッチ ID へのルーティングパスと同一であるかを求めることができ，次元オーダの最短パスルーティングが実現できる．表 1 のプレーン 1 でのルーティング用の値は，図 2 (b) に示した 3 次元同型 Hypercube の例での変換となる．プレーン 1 におけるスイッチ ID 0，1，2，3，4，5，6，7 は，それぞれプレーン 0 におけるスイッチ ID 0，7，6，1，5，2，3，4 と同じ位置にあるため，前記プレーン 0 におけるスイッチ ID の値が表に保持される．

上記より，以下のようにノードからノードへルーティングが行われる．ノード 0 からノード 6 へのルーティングは，ステップ (1) で表 1 よりプレーン 0 とプレーン 1 の距離が同じなため，どちらかのプレーンが選択される．プレーン 0 が選択された場合，ステップ (2) において次元オーダの最短パスルーティングが行われ，下位から 2 および 3 ビット目が 1 であることから $H(0,3) = \{1,2,4\}$ の 2，3 番目の接続が順に利用され，スイッチ ID 0-2-6 とルーティングされる．プレーン 1 が選択された場合，ステップ (2) において表 1 より 6 から 3 に変換され，下位から 1 および 2 ビット目が 1 であることから $H(1,3) = \{3,5,7\}$ の 1，2 番目の接続が順に利用され，スイッチ ID 0-3-6 とルーティングされる．

表 1 において，プレーン 0 での距離やプレーン 1 での距離は，それぞれ src⊕dst の値および表 1 の Routing Value@plane 1 から popcount により算出可能であり，テーブルで保有するのではなくルーティング時に算出してもよい．また，各パケットは 1 つのプレーンのみを用いて転送され，プレーン内では既存の次元オーダの最短経路を用いる．そのため本ルーティングはデッドロックフリーかつライブロックフリーであることは自明である．なお，1 つのパケットを他ノードを経由して 2 つのプレーンを用いて転送することで，単一のプレーン内でルーティングするよりも距離を短くできる可能性はあるが，本論文では扱わないものとする．



(a) 平均最短距離
(a) Average Shortest Path Length



(b) 全対全最大トラフィック
(b) All-to-All Maximum Traffic

図 3　二重同型 Hypercube（DI-HC）および Folded-Hypercube（DI-FHC）のグラフ解析結果

Fig. 3　Graph analysis results of dual-plane isomorphic hypercube (DI-HC) and folded-hypercube (DI-FHC).

2.3　評価指標

二重同型ネットワークの評価指標として，遅延に相当する平均最短距離に加えて，スループットに相当する全対全最大トラフィックを導入する．この全対全最大トラフィックは，各スイッチに 1 ノードが無限のスループットを持ったリンクで接続して Uniform Random 通信を行ったときに，スイッチ間リンクのスループットを 1 として得られるノードあたりの最大スループットを求めるものである．以下に，その算出方法を示す．

(1) 全ノードが全ノードに対してそれぞれ 2（=プレーン数）パケットを送付する．
(2) 前節のルーティングのステップ (1) において，距離が短いプレーンがあれば，両パケットをそのプレーンに流す．一方，距離が同一の場合はそれぞれのプレーンに 1 パケットずつ流す．
(3) 選択したプレーンで固定の最短パスルーティングを行う．
(4) 各スイッチのリンクにおいて通過するパケット数をカウントする（リンクは有向グラフとして方向別にカウントする）．
(5) ノードとスイッチ間のリンクを除く，すべてのスイッチ間リンクにおいて最大の通過パケット数を求める．
(6) 1 ノードが送出するパケット数（ノード数 × 2）を上記最大値で割った値を全対全最大トラフィックと定義する．

たとえば，単一の Hypercube はこの値が 2 となり，D-HC では 2 倍の 4 となる．この指標がシミュレーション結果と一致するかどうかは 3.5 節で検証される．

2.4　解析結果

図 3 に，D-HC，DI-HC，D-FHC，DI-FHC の 4 つの構成について，ネットワークサイズとして n が 2 から 12 までについて，平均最短距離と全対全最大トラフィックを求めた結果を示す．これらは，それぞれ多数の選択肢の中から全対全最大トラフィックと平均最短距離の改善率が最大となる $H(1,n)$ と $FH(1,n)$ を選択したものである．なお，n が 8 以下については，全解探索によって最適な値を求め，9 以上については部分探索結果による準最適解となる．

図より，二重同型化によって平均最短距離を削減し，全対全最大トラフィックを向上させる効果があることが確認できる．$n = 8$ のとき，平均最短距離は二重同型 Hypercube（DI-HC）で 22%（D-HC で 4，DI-HC で 3.13），二重同型 Folded-Hypercube（D-FHC）で 17%（D-FHC で 3.27，DI-FHC で 2.71）削減されている．平均最短距離の削減量はほぼ一定であり，n の増加に従って削減率は徐々に減少している．$n = 12$ の場合，DI-HC で 17%，DI-FHC で 12%の削減となっている．全対全最大トラフィックも $n = 8$ のときに，DI-HC で 28%（D-HC で 4，DI-HC で 5.12），DI-FHC で 21%（D-FHC で 5.51，DI-FHC で 6.65）改善している．こちらも平均最短距離同様に n が大きくなるにつれて向上率は減少し，$n = 12$ のときに DI-HC で 20%，DI-FHC で 13%の改善となっている．

このように，二重同型化が Hypercube においても Folded-Hypercube において，平均最短距離および全対全最大トラフィックを改善する効果を持つことがグラフ解析結果から確認できる．

3. 実システムを想定した最適化

実システム，特に大規模なシステムでは，ネットワークの遅延としてスイッチの遅延とケーブルの遅延の両方を考慮することが重要になってくると考えられる．2 章で同型ネットワークの評価指標として用いた平均最短距離は，これら 2 つのうちスイッチ遅延のみを反映したものであり，ケーブルの遅延は考慮していないものとなる．そこで，この節では平均最短距離に替わる新たな指標として，ケーブル遅延も考慮した平均最短遅延を導入し，その指標をもとに最適な二重同型ネットワークの接続を求める．

本章では，まず想定するシステム構成について示したあと，平均最短遅延を指標とした最適な二重同型ネットワークのグラフ解析結果を示す．次に，ネットワークの経済的コストを試算し，最後にサイクルレベルのシミュレータによるネットワーク性能評価結果を示す．

3.1 実システムの想定

3.1.1 システム構成

HPC システムで標準的に用いられている 19 インチラックを複数利用し，それらがなるべく正方形に近い形で並べられるシステム構成を想定する．1 ラックのサイズは幅 80 cm，奥行き 150 cm，高さ 200 cm からなるとし，幅方向には隣接して並べ，奥行方向には保守スペースとして 100 cm のスペースをとり並べることとした．また，正方形に近い形となるラック配置として，幅および奥行のラックの比率を 2 : 1 あるいは 4 : 1 のどちらかを選択して並べた．

1 ラックには 16 ノードとそれらのノードが接続するスイッチを収容するものとした．また，スイッチあたりに接続するノード数は 4 とした．この 4 という数値は，ノードとスイッチ間．およびスイッチとスイッチ間にはすべて同じ Link 速度で接続する想定で，ノード–スイッチ間リンクではなく，スイッチ間のリンクがボトルネックとなるように選択した値である．なお，二重同型ネットワークを構成するうえで，接続を変更するのはスイッチ間のケーブルのみとし，ノードとスイッチ間の接続についてはどちらのプレーンも同一とした．

3.1.2 ケーブル長

市販されているケーブルを利用して配線を行うことを想定してケーブル長を算出した．現在 Mellanox 社が販売している InfiniBand EDR 用のケーブルの種類を**表 2** に示す [28]．InfiniBand のケーブルには銅ケーブルと光ケーブルが存在し，銅ケーブルは 2 m から 5 m，光ケーブルは 10 m から 100 m が提供されている．

表 2　InfiniBand EDR ケーブルの種類
Table 2　Type of InfiniBand EDR cables.

Cable types	Length（m）
Copper	2.0，2.5，3.0，4.0，5.0
Optical	10，15，20，30，50，100

ラック内およびラック間の配線長は以下のようにして求めた．ラック内の，ノードとスイッチ間の配線長はノードの位置によって異なり，最短でほぼ 0，最長でラック高の 200 cm となる．ここでは平均をとり 100 cm とし，配線猶予の 100 cm を加えた 200 cm とした．ラック内の，スイッチとスイッチ間の配線長は，スイッチが隣接していることから，100 cm とした．最後に，ラック間配線長は，マンハッタン距離でカウントし，200 cm のケーブル余裕を加えることとした [13]．上記により算出した配線長に対して，表 2 に示されているケーブル候補から満たすものを選択し，今回試算するケーブル長とした．

図 4 に 8 ラック構成時（$n = 5$）のラック間配線の配線長と選択されたケーブル長の例を示す．8 ラック構成時は幅方向に 4 ラック，奥行方向に 2 ラック並べる構成となる．ラック内の数値はそのラックに配置されるスイッチの ID を示す．たとえば，スイッチ ID8 と 12 を接続する場合，幅方向で隣接するラック間の配線となり，280 cm が配線長となる．この配線長を満たすケーブルは 3.0 m のものとなるためそれがケーブル長となる．また，スイッチ ID0 と 31 を接続する場合，最も遠いラック間の接続となり，690 cm が配線長でケーブル長は 10 m となる．

3.2 実システムを想定した同型ネットワークの選択

3.2.1 遅延指標

実システムを反映した遅延指標として新たに下記の式で表される平均最短遅延を導入する．下記の平均最短遅延は，スイッチ遅延とケーブル遅延の両方を考慮したモデルであり，Uniform Random 通信を行ったときの平均遅延の最小値を表すものである．

平均最短遅延
= (平均ケーブル長 × ケーブル遅延 + スイッチ遅延)
×平均最短距離 + α

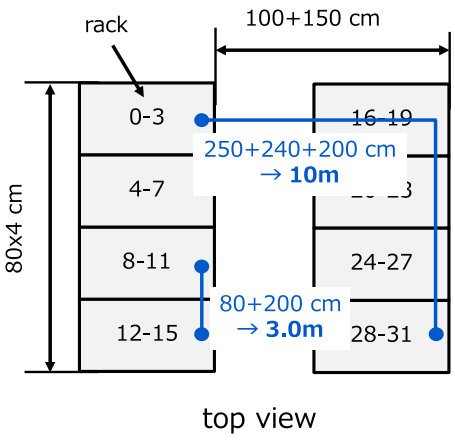
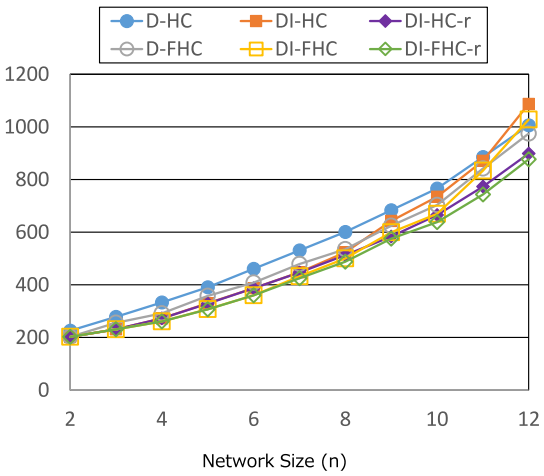
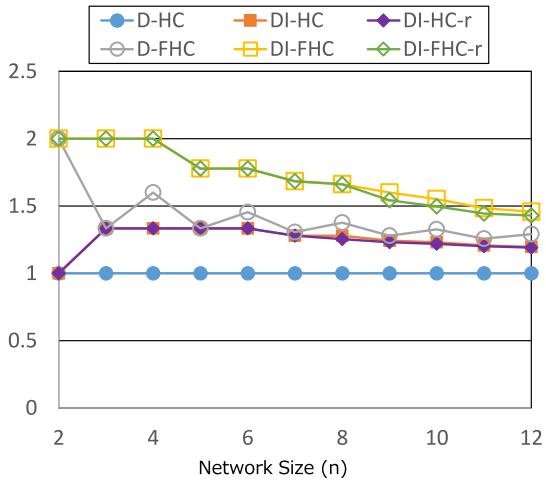


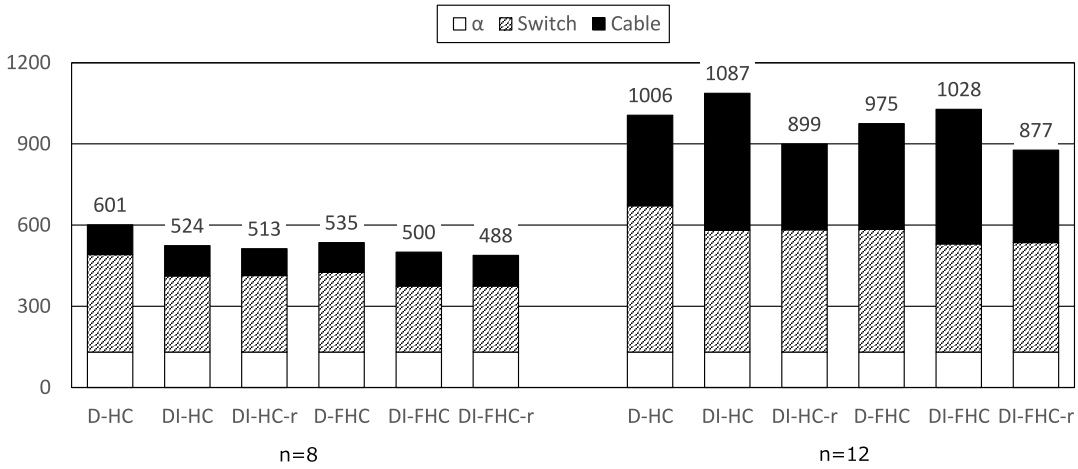
図 4　ケーブル配線長の計算
Fig. 4　Cable length calculation.



(a) 平均最短遅延（ns）
(a) Average Shortest Path Latency (ns)



(b) 全対全最大トラフィック
(b) All-to-All Maximum Traffic



(c) 平均最短遅延（ns）の内訳
(c) Breakdown of Average Shortest Path Latency (ns)

図 5　実システムを想定した二重同型 Hypercube の解析結果
Fig. 5　Graph analysis results of dual-plane isomophic hypercube on the real system.

このモデルでは，スイッチ間を 1 ホップするのに要する遅延を平均ケーブル長とスイッチの遅延から計算し，それに平均最短距離をかけることでスイッチ間の伝送遅延を求めている．α は定数分で，ノードからスイッチ，スイッチからノードに要する時間などに相当する．本節では後述するシミュレーション評価と同一のパラメータとして，スイッチ遅延を 90 ns，ケーブル遅延を 5 ns/m，α は 131 ns とした．なお，この指標がシミュレーション結果と一致するかどうかは 3.5 節で検証される．

3.2.2　二重同型ネットワークの効果

平均最短距離を指標としたものと区別するために，平均最短遅延を指標とした二重同型ネットワークをそれぞれ DI-HC-r，DI-FHC-r と記す．

図 5 (a)，(b) に，D-HC，DI-HC，DI-HC-r，D-FHC，DI-FHC，DI-FHC-r の 6 つのネットワークについて，平均最短遅延，全対全最大トラフィックを求めた結果を示す．DI-HC と DI-FHC は，それぞれ多数の選択肢の中から全対全最大トラフィックと平均最短距離の改善率が最大となる $H(1,n)$ と $FH(1,n)$ を選択したものである．DI-HC-r と DI-FHC-r は，それぞれ多数の選択肢の中から全対全最大トラフィックと平均最短遅延の改善率が最大となる $H(1,n)$ と $FH(1,n)$ を選択したものである．なお，n が 8 以下については，全解探索によって最適な値を求め，9 以上については部分探索結果による準最適解となる．全対全トラフィックについては，図 3 (b) に示した値の 1/4 になっている．これは，図 3 (b) においては 1 スイッチに接続するノード数を 1 としていたが，この評価では 1 スイッチに接続するノード数を 4 としているためである．

図より，$n \leqq 8$ においては，平均最短距離で最適化した DI-HC，DI-FHC，および平均最短遅延で最適化した DI-HC-r，DI-FHC-r ともにほぼ同様の傾向を示していることが確認できる．$n = 8$ の場合に，DI-HC，DI-HC-r の平均最短遅延は，D-HC の 601 ns からそれぞれ 524 ns（13%）と 513 ns（15%）に削減できており，DI-FHC，DI-FHC-r

も，D-FHC の 535 ns からそれぞれ 500 ns（6.6%）と 488 ns（8.9%）に削減できている．全対全最大トラフィックも，D-HC の 1.0 から DI-HC で 1.28（28%），DI-HC-r で 1.25（25%），D-FHC の 1.38 から DI-FHC，DI-FHC-r とも 1.66（21%）に向上している．

また，$n > 8$ の場合には違う傾向を示していることも確認できる．平均最短遅延において，平均最短距離で最適化した DI-HC，DI-FHC とも n が増加するにしたがって改善効果が減少し，$n = 12$ においては D-HC や D-FHC よりも悪化する結果となっている．一方，平均最短遅延で最適化した DI-HC-r と DI-FHC-r は，悪化することなく改善を維持できていることが分かる．$n = 12$ の場合に，平均最短遅延は D-HC の 1,006 ns から，DI-HC は 1,087 ns と 8.1%の悪化となっているが，DI-HC-r は 899 ns と 11%の改善となっている．D-FHC の場合も，975 ns から，DI-FHC は 1028 ns と 5.5%の悪化となっているが，DI-FHC-r は 877 ns と 10%の改善となっている．全体全トラフィックについては，$n > 8$ の場合も両者で大きな差は生じていない．$n = 12$ の場合に，D-HC の 1.0 から DI-HC で 1.20（20%），DI-HC-r で 1.19（19%），D-FHC の 1.29 から DI-FHC で 1.46（13%），DI-FHC-r で 1.43（11%）に向上している．

平均最短遅延の傾向の違いを確認するために，その内訳を図 5 (c) に示す．図は，3.2.1 項の式で求められる値で，スイッチ遅延，ケーブル長遅延，およびその他遅延に分類して示したものである．

この図より，$n = 8$ のケースにおいては，内訳に占めるスイッチ遅延の割合が高く，ケーブル遅延の割合が低いことが分かる．また，遅延の改善はおおむねスイッチ遅延の改善によることも分かる．一方，$n = 12$ のケースにおいては，異なる様相を示し，内訳に占めるケーブル遅延の割合が増大している．$n = 8$ のケースと同様に，スイッチ遅延についてはどの同型 Network も同様に改善している．しかし，ケーブル遅延については，平均最短距離で最適化した DI-HC および DI-FHC はケーブル遅延が大幅に増大し，それがスイッチ遅延の削減を打ち消し全体として平均最短遅延の悪化につながっている．それに対して，平均最短遅延で最適化した DI-HC-r および DI-FHC-r はケーブル遅延をわずかながらも改善しており，スイッチ遅延の削減効果とあわせて平均最短遅延の削減につながっている．DI-HC での数値を見ると，スイッチ遅延は D-HC と比較して 90 ns 減少しているが，ケーブル遅延は 172 ns 増加し，全体として 81 ns の悪化となっている．DI-FHC においても，D-FHC と比較してスイッチ遅延は 54 ns 減少しているがケーブル遅延は 107 ns 増加し，全体としては 53 ns の悪化となっている．一方，DI-HC-r の数値を見ると，スイッチとケーブルの遅延は D-HC と比較してそれぞれ 89 ns と 18 ns 改善し，全体として 107 ns 改善している．D-FHC-r においても，スイッチとケーブルの遅延は D-FHC と比較してそれぞれ 48 ns と 49 ns 改善し，全体として 98 ns 改善している．

上記の結果より，ケーブルによる遅延を考慮した平均最短遅延で最適化した二重同型ネットワークが，平均最短遅延を削減し全対全最大トラフィックを向上させるという観点で有効であることが確認できる．

表 3　InfiniBand EDR の市販価格
Table 3　Price of InfiniBand EDR.

(a) NIC

Ports	Price（US$）
1	399
2	485

(b) Switch

ports	Price（US$）
36	11,170
216	50,500
324	61,995
648	93,000

(c) Copper cable

Length（m）	Price（US$）
2.0	134
2.5	140
3.0	148
4.0	242
5.0	278

(d) Optical cable

Length（m）	Price（US$）
10	545
15	575
20	610
30	685
50	975
100	1,665

3.3　経済的コストの計算

経済的コストは，InfiniBand の市販価格をベースに試算する．今回のコスト算出の範囲は，NIC とスイッチ，およびノードとスイッチ間のケーブル，スイッチとスイッチ間のケーブルとし，ノードあたりのコストを様々なネットワークサイズで評価した．以降，コストを計算するにあたっての想定と算出方法を述べ，試算結果を示す．

3.3.1　コスト試算方法

Mellanox 社が販売している NIC とスイッチのポート数と価格の一覧を**表 3** (a)，(b) に示す [28]．NIC のコストは各ノードが 2 ポートの NIC を 1 個利用するものとした．スイッチのコストについては，文献 [14], [17], [19], [20] を参考に，以下の方式で算出した．今回評価する構成のスイッチのポート数は 36 以下となる．スイッチあたりのノード数は 4 としているので，n が 8 のときに Hypercube で 12 ポート，Folded-Hypercube で 13 ポート．n が 12 のときに Hypercube で 16 ポート，Folded-Hypercube で 17 ポートとなる．36 ポートを大きく下回るポート数からなるため，スイッチコストがポート数に比例するとして，36 ポートのスイッチコストから以下の方式で推定することとした．

$$\text{スイッチコスト} = 310.28 \times \text{ポート数} \quad (US\$)$$

ケーブル長の算出は 3.1.2 項で示した方法で行い，選択したケーブルのコストを表 3 (c)，(d) に示す市販価格 [28]

を元にコストを算出した．

3.3.2　コスト試算結果

前節の試算方法に従って算出した，ノードあたりのネットワーク・コストを**図 6** に示す．図から，ネットワークサイズの増加によってコストが悪化していることが確認できる．これは，ネットワークサイズが増えることによって，ノードあたりのスイッチポート数やケーブル数が増加すること，同時にラック数が増えてシステム規模が大きくなることよりケーブル長が伸びることによる．また，平均最短距離で最適化した二重同型ネットワーク（DI-HC, DI-FHC）はネットワークサイズの増加によって二重ネットワーク（D-HC, D-FHC）よりもコストが悪化するが，平均最短遅延で最適化した二重同型ネットワーク（DI-HC-r, DI-FHC-r）ではそれが抑えられていることも確認できる．以降，二重同型化によるコストの変化に注目して議論する．

二重同型化によって変化するのはスイッチ間のケーブル長とそのコストである．**図 7** にそれぞれ，平均ケーブル長とノードあたりのケーブルコストを示す．図 7 (a) に示した平均ケーブル長の変化を見ると，平均最短距離で最適化した場合は同型化によって大幅に悪化するが，ケーブル長を考慮した平均最短遅延で最適化した場合はその悪化を抑えることができていることが分かる．また，n が大きいほど平均最短遅延で最適化した場合にケーブル長の悪化を抑えられているのは，ケーブル長が遅延に与える影響が大きいためと考えられる．$n = 8$ のときに，DI-HC は D-HC の 5.5 m から 7.1 m（30%）と悪化しているが，DI-HC-r は 6.3 m（15%）に抑えられている．DI-FHC においても，D-FHC の 6.7 m から DI-FHC は 9.3 m（38%）に悪化しているが，DI-FHC-r は 8.4 m（24%）に抑えられている．さらに，$n = 12$ のときを見るとその差は顕著で，DI-HC は D-HC の 11.2 m から 20.3 m（82%）と大幅に悪化しているが，DI-HC-r では 12.7 m（13%）程度に抑えられている．DI-FHC の場合も同様に，D-FHC の 15.5 m から 22.5 m（45%）と悪化しているが，DI-FHC-r では 15.2 m（−2.2%）と逆に減少している．図 7 (b) に示したケーブルコストからも，ケーブル長と同様の傾向が見てとれる．平均最短距離で最適化した場合は同型化によって大幅に悪化するが，ケーブル長を考慮した平均最短遅延で最適化した場合は，特に n が大きくなると，その悪化を抑えることができていることが分かる．$n = 8$ のときに，DI-HC は D-HC の 0.63 k$から 0.73 k$（15%）と悪化しているが，DI-HC-r は 0.67 k$（5.6%）に抑えられている．DI-FHC においては，D-FHC の 0.78 k$から DI-FHC は 0.93 k$（20%）に悪化しており，DI-FHC-r では 0.90 k$（20%）と若干改善している．さらに，$n = 12$ のときを見るとその差は顕著で，DI-HC は D-HC の 1.26 k$から 1.66 k$（32%）と大幅に悪化しているが，DI-HC-r では 1.28 k$（1.7%）に抑えられている．DI-FHC の場合も同様に，D-FHC の 1.57 k$から 1.90 k$（22%）と悪化している

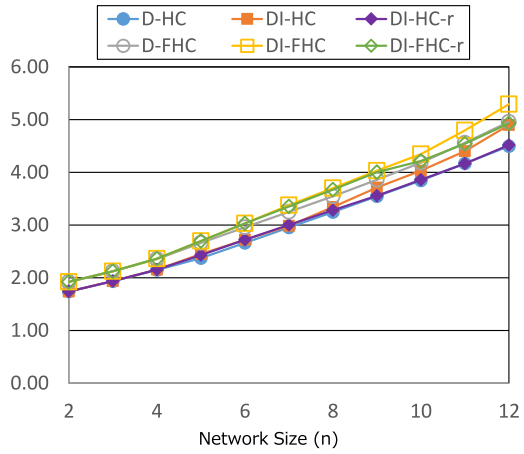
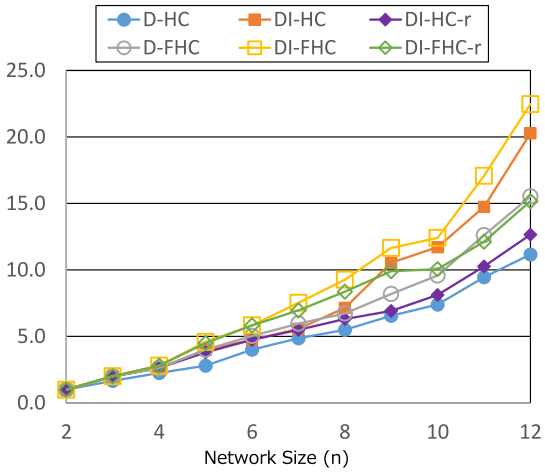
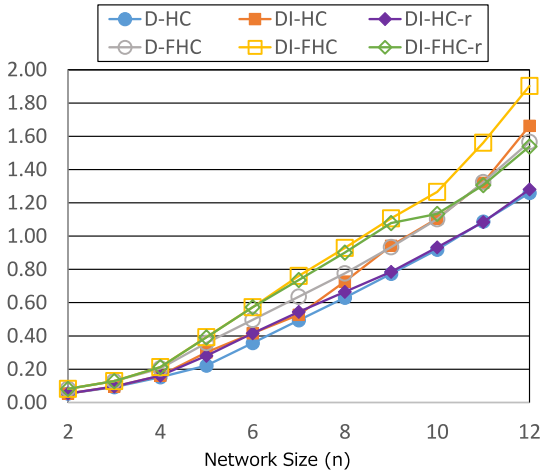


図 6　経済的ネットワークコスト評価結果（k US$）
Fig. 6　Economic network costs evaluation results (k US$).



(a) Average Cable Length (m)



(b) Average Cable Costs per Node (k US$)

図 7　ケーブル長とケーブルコスト
Fig. 7　Cable length and costs.

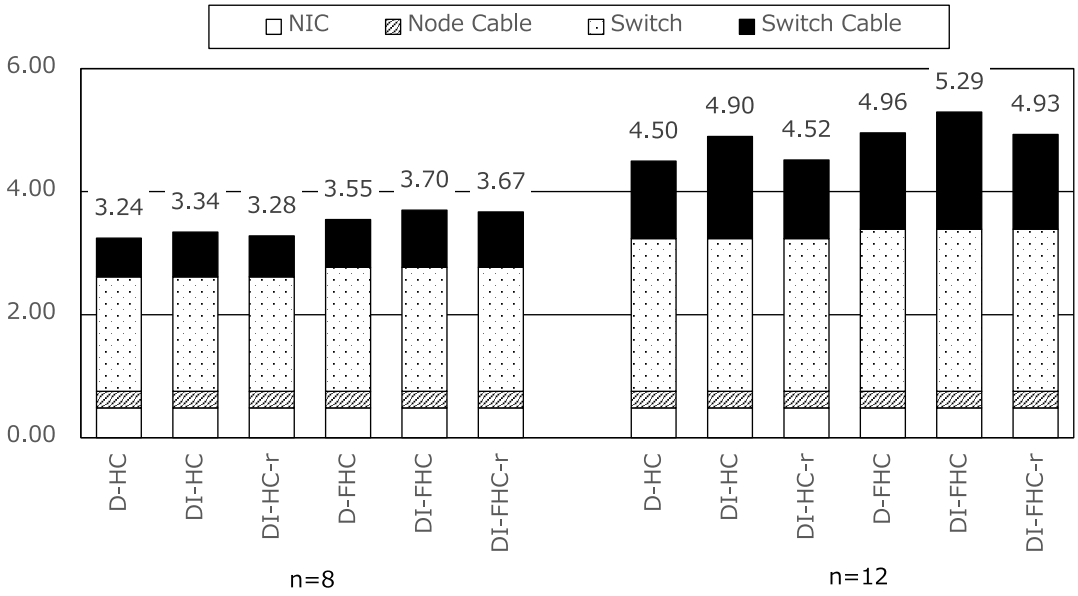


図 8　ネットワークコストの内訳（k US$）
Fig. 8　Breakdown of network costs per node (k US$).

が，DI-FHC-r では 1.54 k$（−1.7%）と逆に減少している．ケーブル長の変化よりもケーブルコストの変化が小さくなっているが，これは表 3 (c)，(d) より m あたりのコストはケーブル長が伸びるほど低下するためと考えられる．

次に，**図 8** に示したコストの内訳を見ると，n が 8 のときはケーブルコストが全体に占める割合が比較的低いことから，ケーブルコストの悪化がネットワークコストの悪化に与える影響が低くなっていることが分かる．また，n が 12 のときはケーブルコストの影響が大きくなり，平均最短距離で最適化した場合はケーブルコストの悪化によって全体コストも悪化しているものの，平均最短遅延で最適化することでケーブルコストの悪化を抑え全体コストの悪化も抑えられていることが分かる．$n = 8$ のときを見ると，スイッチケーブルのコスト割合は DI-HC の場合で 19%，DI-FHC の場合で 22%となっている．そのため，二重同型ネットワークの採用によるコストの増加割合は，最も大きい DI-FHC の場合でも D-FHC からの 4.3%となっている．$n = 12$ のときでは，スイッチケーブルのコスト割合が増加して DI-HC に占める割合は 28%，DI-FHC では 32%となっている．そのため平均最短遅延で最適化してケーブル長の増加を抑えた影響が見てとれ，DI-HC の場合では D-HC の 4.50 k$から 4.90 k$（9.0%）に増加しているが DI-HC-r では 4.52 k$（0.47%）に抑えられている．また，DI-FHC の場合でも，D-FHC の 4.96 k$から 5.29 k$（%）に増加しているが，DI-FHC-r では 4.93 k$と逆に 0.55%改善している．

上記の結果より，ケーブル遅延を考慮した平均最短遅延で最適化した同型ネットワークを選択することが，ネットワークコストの観点でも有効であることが確認できる．

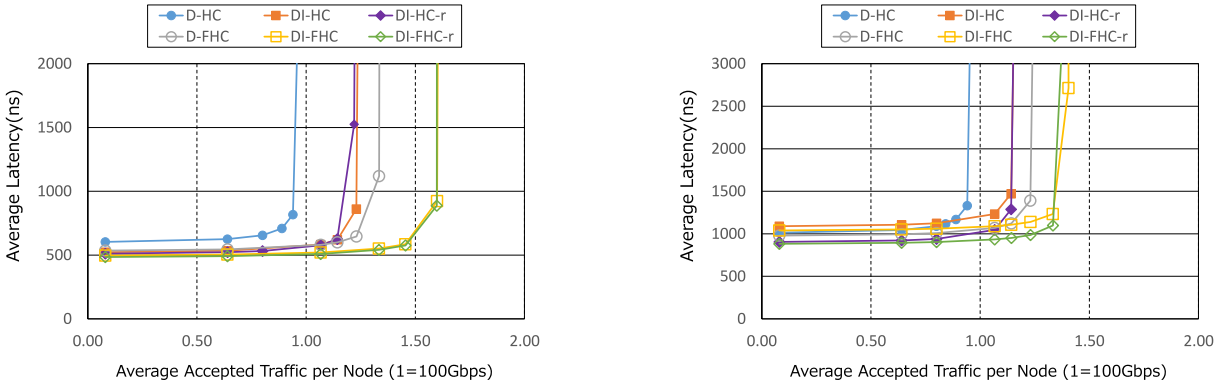
3.4　サイクルレベルシミュレーションによる評価

3.4.1　性能評価方法

サイクルレベルのシミュレータを用いネットワークの遅延およびスループットの評価を行った．

ネットワークのサイズとしては，$n = 8$ および $n = 12$ の 2 構成で評価を行った．なお，スイッチあたりのノード数は 4 としているため，総ノード数は 1,024 および 16,386 の構成となる．スイッチの構成は，入力バッファ，クロスバ，および小容量の出力バッファからなるものとした．ランダム通信においても内部構成がボトルネックにならずリンクと同等に近い性能が出せるように，スイッチ内のクロスバは倍速動作の 1 ポートあたり 200 Gbps で動作するものとした．スイッチのポート to ポートの最短遅延は文献 [23] を参考に 90 ns とした．スイッチ間およびノード・スイッチ間のリンク速度は，InfiniBand EDR のリンク速度である 100 Gbps とした．リンクの遅延は，システム構成で求めたケーブル長に 5 ns をかけた値とした．ネットワーク内でのルーティングについては，静的な次元オーダの最短ルーティングで評価を行った．

ノード間の通信パターンとしては，Uniform Random で評価を行った．各ノードは，上記通信パターンのパケットをある一定間隔で Network に送出し，シミュレーションを走らせて安定状態での受け取ったパケット量のノードあたりの平均（Average Accepted Traffic）と，パケットの送出から受信までの平均遅延（Average Latency）を測定した．ノードのパケット送出間隔を変更して複数回シミュレーションを行い，それぞれのシミュレーションで得られた複数の Average Accepted Traffic と Average Latency の組みでネットワークの特性を明らかにした．



(a) $n = 8$　(b) $n = 12$

図 9　性能評価結果

Fig. 9　Performance evaluation results.

3.4.2　性能評価結果

性能評価を行った結果を，**図 9** に示す．以降の議論では，得られる最大の Average Accepted Traffic をネットワークのスループットとし，また最小の Average Latency をそのネットワークの遅延とした．これらは，文献 [5] でも述べられている一般的な定義である．

図より，$n = 8$ および $n = 12$ の両方のケースで，二重同型化（DI-HC，DI-HC-r，DI-FHC，DI-FHC-r）によってスループットが向上していることが確認できる．$n = 8$ のケースでは，DI-HC と DI-HC-r のスループットはそれぞれ D-HC から 27%と 25%向上している．また，DI-FHC と DI-FHC-r は，それぞれ D-FHC から 20%と 19%向上している．$n = 12$ のケースでは，DI-HC と DI-HC-r のスループットはそれぞれ D-HC から 19%と 18%向上している．また，DI-FHC と DI-FHC-r は，それぞれ D-FHC から 12%と 10%向上している．

遅延を見ると，$n = 8$ のケースで二重同型化によって遅延が削減されていることが確認できる．平均最短遅延と同様に，$n = 12$ のケースでは平均最短距離で最適化した DI-HC および DI-FHC は元の D-HC，D-FHC よりも悪化している．一方平均最短遅延で最適化した DI-HC-r および DI-FHC-r であれば遅延を削減できていることも確認できる．$n = 8$ のケースでは，DI-HC と DI-HC-r の遅延はそれぞれ D-HC から 13%と 15%減少し，DI-FHC と DI-FHC-r はそれぞれ D-FHC から 7.3%と 9.5%減少している．$n = 12$ のケースでは，DI-HC の遅延は D-HC から逆に 7.4%増加し，DI-FHC の遅延は D-FHC から 6.2%増加している．一方，DI-HC-r，DI-FHC-r であればそれぞれ 11%と 9.9%減少している．

上記の結果より，これまでの解析結果と同様に，ケーブルによる遅延を考慮した平均最短遅延で最適化した同型ネットワークを選択することが，遅延を削減しスループットを向上させるという観点で有効であることが確認できる．

3.5　評価指標の検証

全対全最大トラフィックと平均最短遅延は，Uniform Random 通信時の最大スループットと最小遅延を表すものとして導入した．図 5 のグラフ解析結果と，図 9 のサイクルレベルシミュレーションによる評価結果比較すると，全対全最大トラフィックの誤差は平均して 2.75%（最小 1.32%から最大 3.82%）となっており，平均最短遅延は 0.48%（最小 0.14%から 1.1%）となっている．両者とも大きな差にはなっておらず高い精度で一致しているといえる．

4.　関連研究

ネットワークの遅延削減やスループット向上に関する研究として，ネットワークトポロジの研究が多く行われている．近年の HPC システムに利用されているトポロジとしては，今回評価を行った Hypercube に加えて，Fat-tree [16]，Dragonfly [14]，Torus などが複数のシステムで用いられている．また，直径が小さいトポロジとして，Random Topology [15]，SlimFly [17]，Flattened Butterfly [13] などが提案されている．いずれも，1 つのネットワーク・プレーンにおける接続関係に着目したネットワークの低遅延化および高スループット化に関連する提案である．一方，本論文の提案は，多重プレーンのネットワークを有効活用することにより低遅延化や高スループット化を実現するものである．ノード間の距離がすべて 1 の完全網を除いた他のトポロジにも適用可能な技術である．

Multi-plane の性能についていくつかの論文が存在する．性能向上に関する論文としては文献 [1], [2], [3], [4] などがある．文献 [3], [4] は複数プレーンの Fat-tree におけるスループット向上の性能評価を行っている．文献 [2] は複数プレーンのネットワークにおいて各プレーンへのパケット割当手法の提案と評価を行っている．文献 [1] では，InfiniBand を用いた MPI 通信において，複数プレーンへのパケット割当てを実現する MPI の設計に関する議論を行っている．これらは，各プレーンに同一のトポロジを多重化する方式

に関する提案である．一方，本論文の提案は，各プレーンに同型のトポロジを採用することを新規に提案し，同一のトポロジを採用した方式よりも低遅延化や高スループット化を実現するものである．

2つの独立した，異なるネットワーク構成を用いる研究もある．代表例としては，ハイエンドなデータセンターやスーパコンピュータを対象にして，大きなバルクデータ転送は光サーキットスイッチネットワーク，それ以外のデータ転送は電気スイッチで構成されるネットワークを用いる研究 [9] や，ステンシル通信用に3次元トーラス，バリアなどの集合通信用にツリートポロジの2つのネットワークを有するスーパコンピュータである BlueGene/L があげられる．

また，ケーブル長を課題として議論している論文として Random Topology でのケーブル長とそのコスト削減手法を提案した文献 [18] が存在する．本論文では，二重同型 Hypercube ネットワークを対象に，ケーブル長やそのコスト削減に加えて，ケーブル長が遅延に与える影響を考慮した Network の選定方法やその影響の評価を行っている点に違いがある．

5. おわりに

本論文では，二重同型 Hypercube ネットワークおよび二重同型 Folded-Hypercube ネットワークの提案を行い，その遅延とスループットの評価を行った．これらの二重同型ネットワークは，2つのネットワーク・プレーンを保有するシステムにおいて，各プレーンをグラフ同型のネットワークで接続することを特徴とする．

同型ネットワークの候補は複数存在するため，2つのネットワーク選定法を示した．1つは，ネットワークの評価指標としてよく用いられる平均最短距離と，スループットに相当する全対全最大トラフィックの両指標を改善する同型ネットワークの選定法である．もう1つは，遅延指標として平均最短距離だけでなくスイッチ間のケーブル遅延も考慮した平均最短遅延と，全体全最大トラフィックの両指標を改善する同型ネットワークの選定法である．

グラフ解析やサイクルレベルのシミュレーションの評価結果から，後者の選定法が遅延・スループット・経済的コストの観点から有効な選定法であることが確認された．両手法ともスループットについては改善効果を示し，両者に大きな差がないことが確認された．遅延についても8次元程度までの Hypercube であればどちらの方法でも同程度に改善できることも確認された．しかし，ネットワーク・サイズが大きくなりシステム規模が大きくなると，平均最短距離で最適化する手法ではケーブル長が伸びる影響を受けてケーブル遅延が増加し，遅延が逆に悪化することも明らかとなった．一方，平均最短遅延で最適化する手法であればケーブル遅延を指標に組み込んだことからそれを改善できることも確認された．経済的コストにおいても，遅延同様に，ネットワーク・サイズが大きくなると平均最短距離で最適化した場合はケーブル長が伸びることからコストが増加するが，平均最短遅延で最適化するとコストの増加を防ぐことができることも確認できた．

参考文献

[1] Liu, J., Vishnu, A. and Panda, D.K.: Building Multirail InfiniBand Clusters: MPI-Level Design and Performance Evaluation, *Proc. 2004 ACM/IEEE Conference on Supercomputing*, pp.33–44 (2004).

[2] Coll, S., Frachtenberg, E., Petrini, F., Hoisie, A. and Gurvits, L.: Using multirail networks in high-performance clusters, *Proc. 2001 IEEE International Conference on Cluster Computing*, pp.15–24 (2001).

[3] Wolfe, N., Mubarak, M., Jain, N., Domke, J., Bhatele, A., Carothers, C.D. and Ross, R.B.: Preliminary Performance Analysis of Multi-rail Fat-tree Networks, *Proc. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pp.258–261 (2017).

[4] Jain, N., Bhatele, A., Howell, L.H., Böhme, D., Karlin, I., León, E.A., Mubarak, M., Wolfe, N., Gamblin, T. and Leininger, M.L.: Predicting the Performance Impact of Different Fat-tree Configurations, *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*, pp.1–13 (2017).

[5] Duato, J., Yalamanchili, S. and Ni, L.: Interconnection Networks: An engineering approach, Morgan Kaufmann (2002).

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

[7] Dally, W.J.: Performance Analysis of k-ary n-cube Interconnection Networks, *IEEE Trans. Computers*, Vol.39, pp.775–785 (1990).

[8] Duato, J.: A Necessary and Sufficient Condition for Deadlock-Free Adaptive Routing in Wormhole Networks, *IEEE Trans. Parallel and Distributed Systems*, Vol.6, No.10, pp.1055–1067 (1995).

[9] Barker, K.J., Benner, A., Hoare, R., Hoisie, A., Jones, A.K., Kerbyson, D.K., Li, D., Melhem, R., Rajamony, R., Schenfeld, E., Shao, S., Stunkel, C. and Walker, P.: On the Feasibility of Optical Circuit Switching for High Performance Computing Systems, *Proc. 2005 ACM/IEEE Conference on Supercomputing*, p.16 (2005).

[10] Bhuyan, L.N. and Agrawal, D.P.: Generalized Hypercube and Hyperbus Structures for a Computer Network, *IEEE Trans. Computers*, Vol.33, No.4, pp.323–333 (1984).

[11] El-Amawy, A. and Latif, S.: Properties and performance of folded hypercubes, *IEEE Trans. Parallel and Distributed Systems*, Vol.2, No.1, pp.31–42 (1991).

[12] Miller, M. and Siran, J.: Moore graphs and beyond: A survey of the degree/diameter problem, *Electronic Journal of Combinatorics*, Vol.61, pp.1–63 (2005).

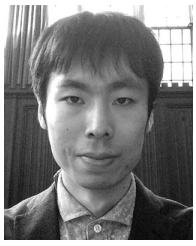
[13] Kim, J., Dally, W.J. and Abts, D.: Flattened Butterfly: A Cost-efficient Topology for High-radix Networks, *The 34th Annual International Symposium on Computer Architecture* (*ISCA'07*), pp.126–137 (2007).

[14] Kim, J., Dally, W.J., Scott, S. and Abts, D.: Technology-Driven, Highly-Scalable Dragonfly Topology, *The 35th Annual International Symposium on Computer Archi-*

tecture (*ISCA'08*), pp.77–88 (2008).
[15] Koibuchi, M., Matsutani, H., Amano, H., Hsu, D.F. and Casanova, H.: A case for random shortcut topologies for HPC interconnects, *The 35th Annual International Symposium on Computer Architecture* (*ISCA'12*), pp.177–188 (2012).
[16] Leiserson, C.E.: Fat-trees: Universal networks for hardware-efficient supercomputing, *IEEE Trans. Computers*, Vol.34, No.10, pp.892–901 (1985).
[17] Besta, M. and Hoefler, T.: Slim fly: A cost effective low-diameter network topology, *The International Conference for High Performance Computing, Networking, Storage and Analysis* (*SC'14*), pp.348–359 (2014).
[18] Koibuchi, M., Fujiwara, I., Matsutani, H. and Casanova, H.: Layout-conscious Random Topologies for HPC Off-chip Interconnects, *The 19th International Symposium on High-Performance Computer Architecture* (*HPCA'13*), pp.484–495 (2013).
[19] Curtis, A.R., Carpenter, T., Elsheikh, M., Lopez-Ortiz, A. and Keshav, S.: REWIRE: An optimization-based framework for unstructured data center network design, *The International Conference on Computer Communications* (*INFOCOM*), pp.1116–1124 (2012).
[20] Mudigonda, J., Yalagandula, P. and Mogul, J.C.: Taming the flying cable monster: A topology design and optimization framework for data-center networks, *The USENIX Conference on USENIX Annual Technical Conference*, p.8 (2011).
[21] Top 500 Supercomputer Sites, available from ⟨http://www.top500.org/⟩.
[22] Deploying HPC Cluster with Mellanox InfiniBand Interconnect Solutions, available from ⟨http://www.mellanox.com/related-docs/solutions/deploying-hpc-cluster-with-mellanox-infiniband-interconnect-solutions-archive.pdf⟩.
[23] SB7700 InfiniBand EDR 100Gb/s Switch System, available from ⟨http://www.mellanox.com/related-docs/prod_ib_switch_systems/pb_sb7700.pdf⟩.
[24] NASA Peiades Supercomputer, available from ⟨https://www.nas.nasa.gov/hecc/resources/pleiades.html⟩.
[25] Papatheodore, T.: Summit System Overview, available from ⟨https://www.olcf.ornl.gov/wp-content/uploads/2018/05/Intro_Summit_System_Overview.pdf⟩.
[26] Matsuoka, S.: TSUBAME3 and ABCI: Supercomputer Architectures for HPC and AI/BD Convergence, available from ⟨http://on-demand.gputechconf.com/gtc/2017/presentation/S7813-Matsuoka-scalable.pdf.pdf⟩.
[27] HPE SGI 8600 System, available from ⟨https://www.hpe.com/jp/ja/product-catalog/detail/pip.hpe-sgi-8600-system.1010032504.html⟩.
[28] COLFAX DIRECT, available from ⟨https://colfaxdirect.com/store/pc/home.asp⟩.
[29] Hosomi, T., Yasudo, R., Koibuchi, M. and Shimojo, S.: Dual-plane Isomorphic Hypercube Network, *Proc. International Conference on High Performance Computing in Asia-Pacific Region* (*HPCAsia2020*), pp.73–80 (2020).

細見 岳生

平成 6 年京都大学大学院情報学研究科修士課程卒業．同年 NEC 入社．コンピューティングシステムの研究開発に従事．



安戸 僚汰　（正会員）

平成 31 年慶應義塾大学大学院理工学研究科後期博士課程修了．博士（工学）．同年より広島大学情報科学部特任助教．現在，相互結合網，GPU や FPGA を用いた高速計算に関する研究に従事．

鯉渕 道紘　（正会員）

平成 15 年慶應義塾大学理工学研究科博士課程修了．博士（工学）．平成 17 年国立情報学研究所助手．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准教授（兼任）．相互結合網と計算機システムに関する研究に従事．

下條 真司　（正会員）

昭和 61 年大阪大学基礎工学部大学院後期課程修了．同年大阪大学・助手．平成元年同大型計算機センター・講師．平成 3 年同助教授，平成 10 年同教授，平成 12 年同大学サイバーメディアセンター副センター長，平成 17 年同大阪大学センター長，平成 19 年同副センター長，平成 20 年から 3 年間情報通信研究機構大手町ネットワーク研究統括センターセンター長/上席研究員．平成 23 年サイバーメディアセンター教授．平成 27 年よりセンター長．現在に至る．