Vol.2014-HPC-145 No.12
2014/7/28

スイッチ台数の削減と高いAll-to-all通信性能を両立する多層Fullmeshトポロジーの提案

中島 耕太[1,a)]　三輪 真弘[1]

概要：本稿では，スイッチ台数の削減と高い All-to-all 通信性能を両立する多層 Fullmesh トポロジーを提案する．クラスタシステムの大規模化に伴い結合のために必要なスイッチ台数も増加している．スイッチ台数が増加すると電力コストが増加するだけでなく部品点数が増加するため故障率が増加し保守コストも増加する．そこで，スイッチ台数を削減する多層 Fullmesh トポロジーを提案する．数千ノードクラスのシステムにおいて接続のために必要なスイッチ数を 3 段 Fat Tree と比較して約 4 割削減しつつ，All-to-all 通信時のデータ交換順序をスケジュールすることで経路競合を回避し，Fat Tree と同等の性能を維持できる．通信フローにおけるスループット評価では，多層 Fullmesh において Fat Tree と同様に経路競合が回避できることを確認し，ランダム通信においても大規模構成においては Fat Tree とほぼ同等の性能となることを確認した。

1. はじめに

HPC システムにおけるクラスタシステムの利用が普及している．クラスタシステムは IA サーバに代表されるコモディティ部品から構成されるサーバをノードとし，数百〜数千のノードを高速なネットワークで接続したシステムであり，コストパフォーマンスに優れる特徴を持つ．

特に近年，GPGPU や Xeon Phi といったアクセラレータの出現により，ノードあたりの演算性能が劇的に向上している．幅広いアプリケーションにおいて高い演算性能を活かして高い性能を実現するには演算性能の向上に見合うだけのネットワーク性能を実現する必要がある．

高いネットワーク性能を実現するために，多くのシステムにおいてノード間を接続するネットワークとして InfiniBand[1] が採用されている．2014 年 6 月の Top 500[2] に登録されているシステムのうち 44.6%が InfiniBand を採用している．また，高い性能を実現するために Fat Tree トポロジーが利用されている．Fat Tree は All-to-all 通信のような全ノードが全ノードに対して通信する場合においても経路競合を避けることができ，高い性能を実現できる．一方で，数千ノードを接続するためには 3 段 Fat Tree を構成する必要があり，スイッチ台数が大幅に増加する問題がある．

そこで，スイッチ台数の削減とネットワーク性能を両立する多層 Fullmesh トポロジーを提案する．多層 Fullmesh は，Fullmesh の各辺にスイッチを接続し，複数の Fullmesh を多層化して接続したトポロジーである．この多層 Fullmesh は FBB(Fully Bisectional Bandwidth) 構成の 3 段 Fat Tree と比較して約 4 割スイッチ数を削減することができる．また，All-to-all 通信において，各通信フェーズにおける通信相手の選択順序をスケジュールしデータ交換手順を最適化することで通信競合を回避し，Fat Tree と同等の性能を維持する．

FBB 構成の 3 段 Fat Tree と多層 Fullmesh の通信フローによる性能比較を行ったところ，All-to-all 通信においては Fat Tree と同様に各フェーズにおける経路競合が回避できるため FBB 構成 3 段 Fat Tree と同等の性能を達成した．またランダムな通信パターンにおいても，ネットワーク規模が大きくなるに連れて FBB 構成 3 段 Fat Tree に匹敵する性能を達成できることがわかった．

本稿では，スイッチ台数の削減と高い All-to-all 通信性能を両立する多層 Fullmesh トポロジーを提案する．以降，2 章では課題について述べ，3 章で広く用いられている Fat Tree と All-to-all 通信について説明する．4 章では多層 Fullmesh を提案する．5 章では，多層 Fullmesh と Fat Tree の特徴を比較し，6 章では，性能評価について述べる．7 章で関連研究について述べ，8 章でまとめと今後の課題について述べる．

[1] （株）富士通研究所
Fujitsu Laboratories Ltd.
[a)] nakashima.kouta@jp.fujitsu.com

2. 課題

多くのクラスタシステムでは InfiniBand による Fat Tree トポロジーが採用されている．Fat Tree は，All-to-all 通信のような通信負荷の高い通信パターンにおいても高い性能を達成できることが特徴である．例えば，並列化された FFT では，All-to-all 通信性能を高めることがアプリケーション性能を向上させるにあたって非常に重要である．FFT は最も顕著な例の一つであるが，この他にも様々な数値計算において負荷の高い通信処理は利用されており，様々なアプリケーションにおいて演算性能を引き出すためには，クラスタシステムにおけるネットワーク性能を十分に高める必要がある．

一方で，クラスタシステムの大規模化に伴い，接続するノード数が増加している．このため，接続のために必要となるスイッチの台数は増加している．スイッチの台数が増加すると，部材費用，電力，設置面積のコストが増加する．

Fat Tree では，数千ノード規模のクラスタシステムを構成する場合，十分な帯域を確保するためには 3 段の Fat Tree を用いる必要がある．3 段 Fat Tree は，表 1 に示すようにノードと接続されるポート数の合計をスイッチのポート数の合計で割ったポート利用効率が 1/5 と低い．このため，接続ノード数が増加すると，必要となるスイッチ数も大幅に増加し，例えば 5,832 台のノードを接続するためには 810 台のスイッチが必要となる．

さらに，ノードあたりの演算性能の向上が進んでいるため，通信性能も増強する必要がある．例えば，TSUBAME 2.5 や HA-PACS では，ネットワークを 2 重化し，ネットワークバンド幅を増強している．このような多重化構成では，さらに必要となるスイッチ数が増加してしまう．

したがって，ネットワークを構成するスイッチの台数を削減しつつ，Fat Tree と同じような高い性能を実現することが課題である．

3. Fat Tree と All-to-all 通信

Fat Tree の構成を図 1 や図 2 に示す。図 1 や図 2 のように，上段に向かうリンクの本数が複数存在するトポロジーであり，各段のスイッチにおいて，下段側と上段側のポート数の比率が 1:1 であるとき，下段側からの通信をすべて異なる上段側のポートへ転送することが出来れば経路競合が発生しないため，高い通信性能を実現できるトポロジーとして知られており，クラスタシステムで広く用いられている．

一定の段数で接続できるノード数の上限はスイッチのポート数で決定される．例えば，現在主流となっている InfiniBand スイッチは 36 ポートであるが，この場合，2 段 Fat Tree では最大 648 ノードまで，3 段 Fat Tree だと

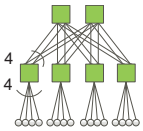

図 1 2 段 Fat Tree

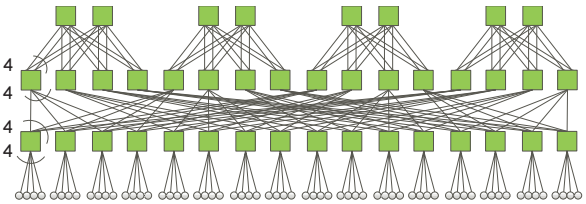

図 2 3 段 Fat Tree

11,664 ノードまで接続できる．Fat Tree の段数が増加してしまうと，表 1 に示すようにポート利用効率が下がってしまい，必要となるスイッチ数がさらに増えてしまう特徴がある．

All-to-all 通信は，数多くの HPC のアプリケーションで利用されている通信パターンであり，各ノードが持つデータを全ノードに対して配信する通信である．メッセージサイズが大きい場合は，実際に各ノードがすべてのノードに対してデータを転送する．すなわち n ノード間で All-to-all 通信を行う場合には，n フェーズの通信が実行される．

Fat Tree では，シフト通信パターン [3] と呼ばれる転送順序を用いることで All-to-all 通信時の経路競合を回避している．シフト通信パターンは，自身のノード番号を S とした場合に，i 番目のフェーズにおいて (1) 式で算出されるノード D に対して転送する．

$$D = (S + i)\%n \tag{1}$$

このように通信すると図 3，4 に示すようにすべてのフェーズにおいてあるスイッチの配下にあるすべてのノードからの通信がそれぞれ別々の上段スイッチを経由するように転送されるため経路競合を回避できる．このように Fat Tree では All-to-all の経路競合をこのようにして回避している．

4. 多層 Fullmesh の提案

Fullmesh は図 5 に示すような全スイッチが全スイッチと直接接続されるトポロジーである．文献 [4], [5] によると Fullmesh において，All-to-all 通信において競合を発生させることなく転送順番をスケジュールできることが知られている．一方で単一の Fullmesh だけだと接続ノード数に限界があり，大規模なクラスタシステムを構築することはできない．

そこで，新たに多層 Fullmesh というトポロジーを提案する．多層 Fullmesh の例を図 6 に示す．多層 Fullmesh は，Fullmesh の各辺にスイッチを配置し，複数の Fullmesh を

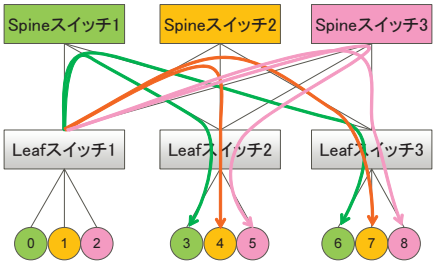

図 3 Fat Tree における All-to-all 通信

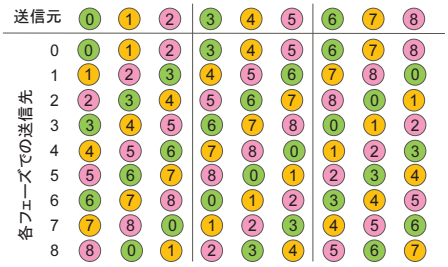

図 4 シフト通信パターンの順序

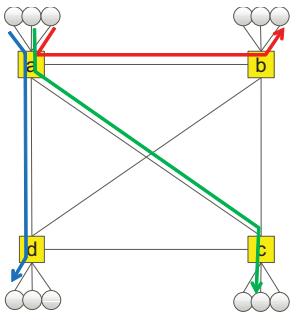

図 5 Fullmesh

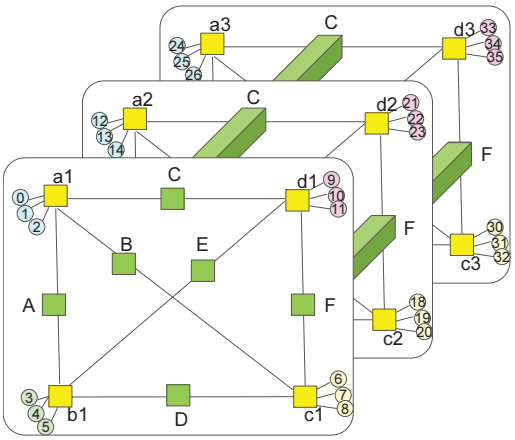

図 6 多層 Fullmesh(6 ポートスイッチ)

接続した構成である．

多層 Fullmesh では，各 Fullmesh の層の内部では Fullmesh と同様に通信競合を避けることができる．通信競合を避けるための宛先の決定式を式 (2)，式 (3) に示す．

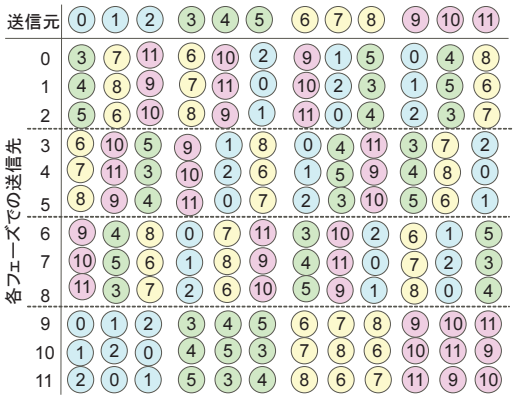

図 7 層内の通信順序

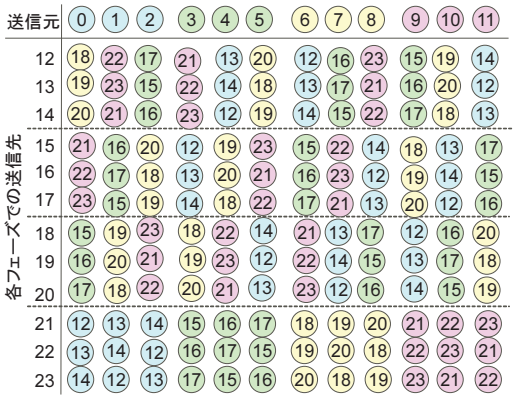

図 8 1 層目から 2 層目への通信順序

$$D = [\{S/d + (o + i/d)\%d + 1\}\%(d+1)] \cdot d + (o+i)\%d + [i/\{d(d+1)\}] \cdot d(d+1) \quad (2)$$

$$D = (S/d \cdot d) + (o+i)\%d + [i/\{d(d+1)\}] \cdot d(d+1) \quad (3)$$

数式中の各変数は以下の通りである．D:転送先ノード番号，S:転送元ノード番号，i:フェーズ，d:次数 (ポート数の 1/2)，o:送信元ノード番号のグループ内オフセット ($o = S\%d$)．また，フェーズが $d(d+1)k$ ～ $d(d+1)k+d^2-1$ の場合は式 (2) を，フェーズが $d(d+1)k+d^2$ ～ $d(d+1)(k+1)-1$ の場合は式 (3) を用いて算出する．なお，k は 0～$d-1$ の整数である．

例として，図 6 における 1 番目の層の中の通信順序を図 7 に示す．Fullmesh の頂点となるスイッチ (以降，頂点スイッチ) に接続されている各ノードが他の頂点に接続されているノードに対して通信する場合には，各ノードが別々の頂点スイッチに向かって通信するようにスケジュールする（フェーズ 0～8）．各頂点スイッチの中に閉じる通信は，各頂点に所属するノード同士で交換する（フェーズ 9～11）．

さらに，j 番目の層から $j+1$ 番目の層へ通信する場合は，j 番目の層のある頂点スイッチに接続されている各ノードは，それぞれ $j+1$ 番目の別々の頂点スイッチに向かって通信するようにスケジュールする（フェーズ 12～20）．例えば，フェーズ 12 では，1 番目の層にある頂点スイッチ

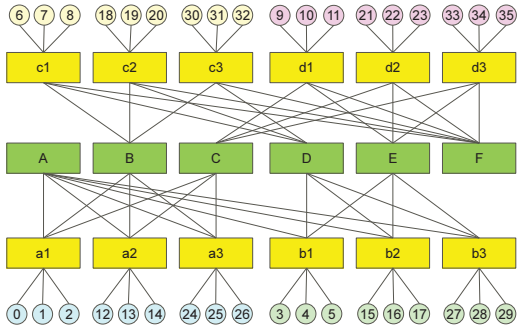

図 9 多層 Fullmesh(図 6 と同一)

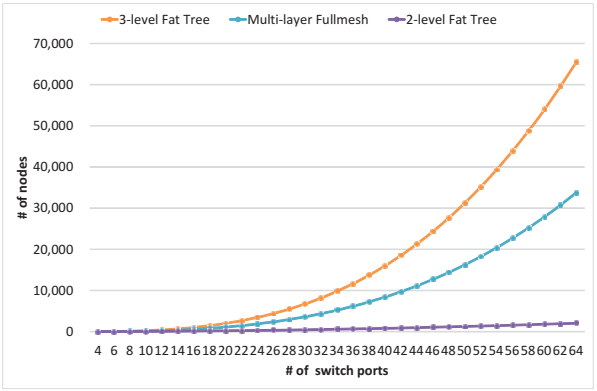

図 10 最大接続ノード数

a1 に接続されるノード 0，1，2 は，2 番目の層にある頂点スイッチ c2，d2，b2 に接続されるノード 18，22，17 へ転送する．また，異なる層における同一の頂点に位置するスイッチに接続される各ノードへは，頂点スイッチが隣接するすべての辺に配置されるスイッチを経由する（フェーズ 21〜23）．例えば，頂点スイッチ a1 に接続されるノード 0，1，2 は，A，B，C を経由して a2 に接続されるノード 12，13，14 へ転送する．

図 7，8 では，1 番目の層のノードからの転送についてのみ記載しているが，各層について同様の通信が同時に行われる．すなわち，1 番目で層内の通信を実施しているときには同時に 2 番目と 3 番目でも層内で通信を行う．また，1 番目の層が 2 番目の層に送信しているしているときは，2 番目の層は 3 番目に，3 番目の層は 1 番目に送信している．また，j 番目の層から $j+1$ 番目の送信が完了すると，j 番目の層から $j+2$ 番目の層へと層の間隔を 1 層ずつ広げていき，最終的にはすべての層間での通信を行う．

このように通信順序をスケジュールすることで，All-to-all 通信のすべてのフェーズにおいて経路競合を回避できる．したがって，Fat Tree と同様に All-to-all 通信において高い性能を実現することができる．

また，図 6 を書き直すと図 9 のように表すことができる．このように同一頂点に相当する頂点スイッチ同士は 2 段 Fat Tree を構成している特徴を持つ．また，スイッチのポート利用効率は 2 段 Fat Tree と同一である．

5. 多層 Fullmesh と Fat Tree の比較

5.1 スイッチ台数と接続規模

スイッチのポート数を p とする場合の最大接続ノード数，スイッチ数，ポート利用効率を 2 段 Fat Tree，3 段 Fat Tree，多層 Fullmesh について比較した．ポート利用効率とは，ノードと接続されるポート数の合計をスイッチのポート数の合計で割ったものである．比較を表 1 に示す．また，スイッチのポート数を 4 から 64 まで変化させた時の最大接続ノード数を図 10 に示す．

多層 Fullmesh は 2 段 Fat Tree と同じポート利用効率で 3 段 Fat Tree の約 1/2 の規模まで接続可能である．2014 年現在広く用いられている InfiniBand スイッチは 36 ポートであるので，多層 Fullmesh を利用して最大 6,156 ノードまで接続できる．同じ利用効率である 2 段 Fat Tree では 648 ノードまでしか接続できないので同じ利用効率で接続ノード数を大幅に増加させることができている．

さらに将来，スイッチのポート数が 48 あるいは 64 まで増加すると接続可能ノード数は 14,400 あるいは 33,792 ノードまで増加する．このように将来的には 1 万ノードクラスの接続も可能である．

表 2 2 分割バンド幅

トポロジー	Fat Tree	Fullmesh
ノード当たり 2 分割バンド幅	1	$\frac{d+1}{2d} \simeq \frac{1}{2}$

(d:次数，p:ポート数の場合，$d = p/2$)

5.2 2 分割バンド幅

Fat Tree と Fullmesh のノードあたりの 2 分割バンド幅を表 2 に示す．多層 Fullmesh は Fullmesh を多層化したトポロジーである．4 で議論したように同一頂点に相当するスイッチ間は Fat Tree 群をなしている．したがって，2 分割バンド幅を考える場合には Fullmesh 層における 2 分割バンド幅を考えれば良い．表 2 から，次数が十分に大きい場合は，Fat Tree の 1/2 の帯域であると言える．

6. 性能評価

6.1 評価方法

多層 Fullmesh の性能を評価するために，同程度のノード数の FBB 構成の 3 段 Fat Tree と帯域を 1/2 にした 3 段 Fat Tree との性能を比較する．評価に用いるネットワークを表 3 に示す．

多層 Fullmesh 多層 Fullmesh は，図 11 に示すように，Fullmesh の頂点となるスイッチにはそれぞれ d 台のノードが接続され，$(d+1)$ 角形の Fullmesh を構成する．Fullmesh の各辺に層間を接続するスイッチが配置される．層の数は d 枚である．

FBB 3 段 Fat Tree 図 2 に示すように，1 段目のスイッ

表 1 トポロジーによる接続規模とポート利用効率

トポロジー	2 段 Fat Tree	3 段 Fat Tree	多層 Fullmesh
最大接続ノード数	$\frac{1}{2}p^2$	$\frac{1}{4}p^3$	$\frac{1}{8}p^3+\frac{1}{4}p^2 \simeq \frac{1}{8}p^3$
スイッチ数	$\frac{3}{2}p$	$\frac{5}{4}p^2$	$3(\frac{1}{8}p^2+\frac{1}{4}p)$
ポート利用効率	$\frac{1}{3}$	$\frac{1}{5}$	$\frac{1}{3}$

表 3 評価に用いるネットワーク構成

	次数	多層 Fullmesh	FBB Fat Tree	1/2 帯域 Fat Tree
スイッチ数	4	30	40	28
	8	108	160	112
	18	513	810	567
ノード数	4	80	64	64
	8	576	512	512
	18	6,156	5,832	5,832

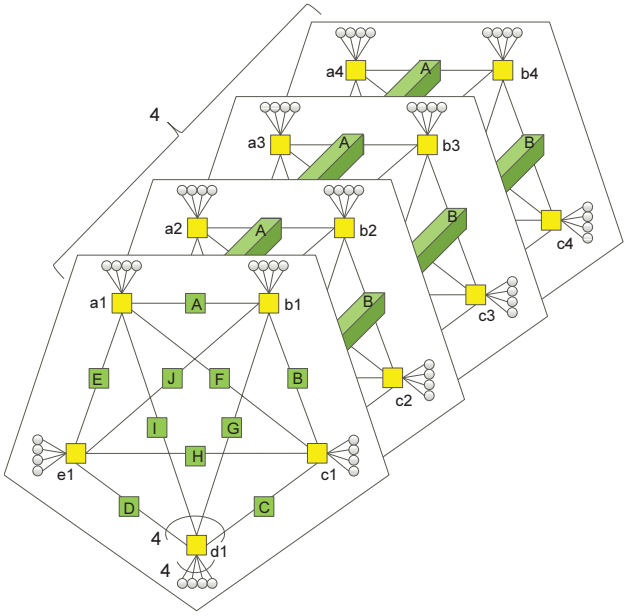

図 11 多層 Fullmesh(8 ポートスイッチ，次数 $d=4$)

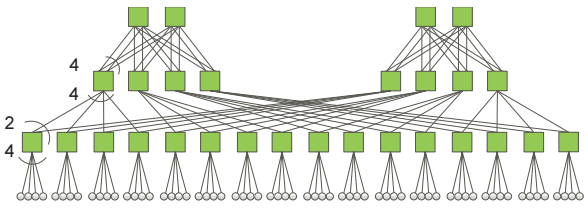

図 12 1/2 帯域 Fat Tree(8 ポートスイッチ，次数 $d=4$)

チにはそれぞれ d 台のノードが接続される．1 段目のスイッチからは 2 段目のスイッチに対して合計 d 本接続され，2 段目のスイッチ 1 つとそれぞれ 1 本ずつ接続される．2 段目のスイッチからは 3 段目のスイッチに対して合計 d 本接続され，3 段目のスイッチ 1 つとそれぞれ 2 本ずつ接続される．

帯域 1/2 Fat Tree 図 12 に示すように，1 段目のスイッチにはそれぞれ d 台のノードが接続される．1 段目のスイッチからは 2 段目のスイッチに対して合計 $d/2$ 本接続され，2 段目のスイッチ 1 つとそれぞれ 1 本ずつ接続される．2 段目のスイッチからは 3 段目のスイッチに対して合計 d 本接続され，3 段目のスイッチ 1 つとそれぞれ 2 本ずつ接続される．

それぞれのネットワークにおいて All-to-all 通信時の性能とランダム通信時の性能を評価する．評価指標には経路競合数から算出するスループットを用いる．経路競合数とは，複数の送信ノードがそれぞれ対応する複数の受信ノードに対して同時に転送した場合における同一リンクを経由する通信フローの数である．経路競合数が 1 である場合は，ある送信ノードから受信ノードへの通信フローにおいて，競合が発生しなかったことを意味する．経路競合数が n である場合は，ある送信ノードから受信ノードへの通信フローにおいて，経由するリンクにおいて最大 n 本の通信フローが同一リンクを経由し競合したことを意味する．

All-to-all 通信時の評価は，All-to-all 通信における各フェーズにおける最大経路競合数の平均値を比較する．All-to-all 通信では各フェーズにおいて全ノードから全ノードに対して同時に転送を行う．各フェーズにおける各通信フローの最大の経路競合数の逆数の平均をスループットとする．すなわち，全フェーズにおいて経路競合数が 1 である場合が性能の最大値であり，スループットは 1.0 となる．

ランダム通信における評価では，全ノードがそれぞれ異なる全ノードに対して同時に転送する場合の各フローの経路競合数を算出する．この経路競合数の逆数の平均をスループットとする．この際，送信ノードと受信ノードの対応はランダムであり，一対一である．このランダムの組み合わせ 100 通りの平均値を算出する．また，経路競合数の分布を確認するために，ヒストグラムを求める．

6.2 All-to-all 性能

All-to-all 性能を図 13 に示す．凡例における **ftree** は

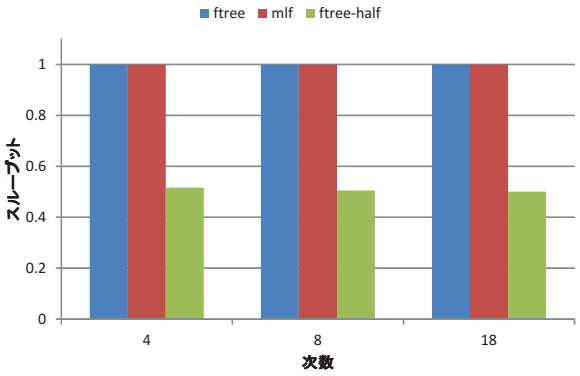

図 13　All-to-all 通信のスループット

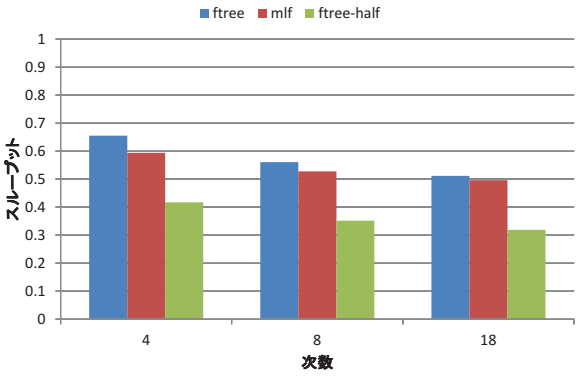

図 14　ランダム通信のスループット

FBB Fat Tree，**mlf** は多層 Fullmesh，**ftree-half** は 1/2 帯域 Fat Tree を示す．多層 Fullmesh と FBB Fat Tree のスループット値はいずれの次数の場合も 1.0 であり，すべてのフェーズにおいて経路競合が回避できていることがわかる．多層 Fullmesh は 2 分割バンド幅が約 1/2 であるにも関わらず FBB Fat Tree と同等の性能を達成できている．1/2 帯域 Fat Tree のスループット値は 0.52〜0.50 であり，ほぼ 1/2 である．帯域を 1/2 に絞っているためスループットは約半分である．このようにトポロジーと転送手順の組み合わせにより少ないハードウェア資源で高い性能が達成できていることが確認できる．

6.3　ランダム通信性能

ランダム通信におけるスループットを図 14 に示す．多層 Fullmesh の性能は FBB Fat Tree に対して 3.0%〜9.4%程度の性能劣化にとどまっている．特に次数が大きくなればなるほど，性能差は小さくなる．このようにスイッチ数を約 4 割削減しているにも関わらず，FBB Fat Tree に迫るランダム通信性能を達成している．

経路競合数の分布を図 15，16，17 に示す．図 15-17 の縦軸は，各経路競合数の全体に占める割合を示している．多層 Fullmesh は，いずれの次数においても同程度のスイッチ数で構成する 1/2 帯域 Fat Tree と比較して経路競合数が少ないことが確認できる．また次数が大きくなると FBB Fat Tree と多層 Fullmesh の経路競合数の分布は近づき，次数が 18 の場合は，ほぼ同等の分布になる．このように，大規模な多層 Fullmesh は FBB Fat Tree に近い性能特性を持つ可能性が高いと言える．

7.　関連研究

文献 [4], [5] では，単一の Fullmesh における All-to-all 通信の経路競合回避手法を提示している．パケットレベルのシミュレーションにおいて効果を確認している．一方で，単一の Fullmesh に対する適用にしか言及していない．単一の Fullmesh だけだと 36 ポートスイッチ構成の場合でも

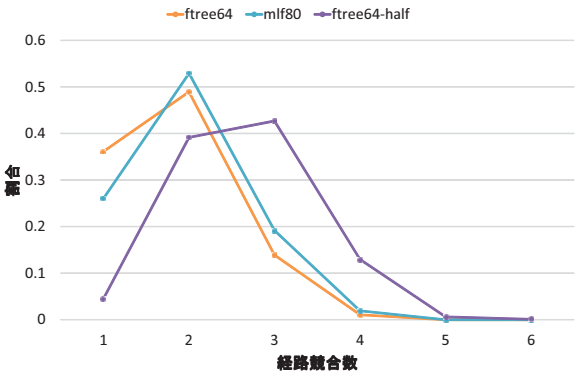

図 15　ランダム通信時の経路競合の比較 (次数 4)

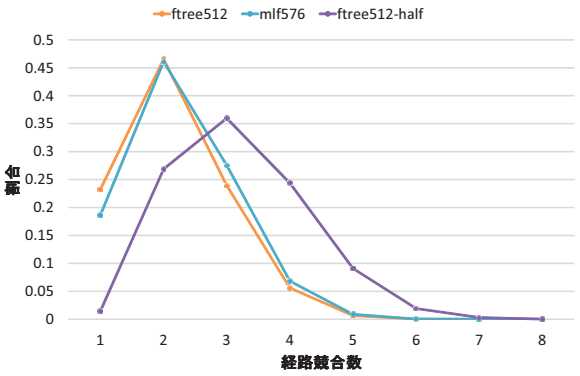

図 16　ランダム通信時の経路競合の比較 (次数 8)

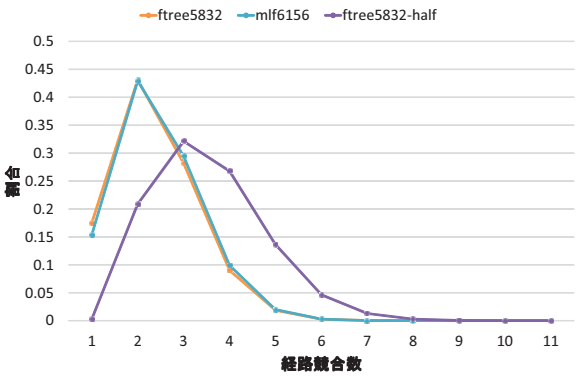

図 17　ランダム通信時の経路競合の比較 (次数 18)

高々 342 ノードまでしか接続できず，数千ノードクラスの構成を実現することは困難である．

8. おわりに

本稿では，スイッチ台数の削減と All-to-all 性能を両立する多層 Fullmesh トポロジーを提案した．クラスタシステムで広く採用されている 3 段 Fat Tree と比較して，スイッチ数を約 4 割削減できることを示した．また，All-to-all 通信時に経路競合が発生しないことを示した．さらに，ランダム通信パターンでも Fat Tree と互角であることを示した．

多層 Fullmesh は，2 段 Fat Tree と同じポート利用効率で 36 ポートスイッチ使用時に最大 6,156 ノードまで接続できる．全ノードにおける All-to-all 通信では FBB Fat Tree と同様に全フェーズで競合回避でき，しかもランダム通信でも同程度の性能を実現できる可能性が高いことを示した．

今後の課題として，より厳密なパケットレベルのシミュレーションによる評価，一部のノードを切り出して使用する場合のジョブ割り当てと経路競合回避の方法の確立，数十万台規模に拡張する手法の検討，実機における実アプリ評価がある．

参考文献

[1] InfiniBand Architecture Specification Release 1.2, InfiniBand Trade Association, http://www.infinibandta.org.
[2] Top 500 http://www.top500.org.
[3] E. Zahavi, G. Johnson, D.J. Kerbyson, and M. Lang : “Optimized Infiniband Fat-tree routing for shift all-to-all communication patterns,” In Proceedings of the International Supercomputing Conference 2007 (ISC07), 2007.
[4] E. Totoni and L. V. Kale: “ACM SRC poster: optimizing all-to-all algorithm for PERCS network using simulation,” Proceedings of the 2011 companion on High Performance Computing Networking, Storage and Analysis Companion, 2011.
[5] E. Totoni, A. Bhatele, E. J. Bohm, N. Jain, C. L. Mendes, R. M. Mokos, G. Zheng, and and L. V. Kale: “Simulation-based Performance Analysis and Tuning for a Two-level Directly Connected System,” Proceedings of IEEE 17th International Conferenceon Parallel and Distributed Systems, 2011.