光サーキットネットワークの補助的利用による HPC アプリケーション性能向上

滝澤 真一朗[†1] 遠藤 敏夫[†1] 松岡 聡[†1,†2]

多数のノードからなる大規模 HPC システムでは，全ノードを高バンド幅で全対全接続するネットワークは金銭コストや電力消費の問題で実現困難である．我々は低バイセクションバンド幅電気パケット（EPS）ネットワークと高バンド幅光サーキット（OCS）ネットワークからなるネットワーク環境と，その環境での通信手法を提案する．この環境では，各ノードは単一リンクで EPS ネットワークに接続され，一部のノードは OCS ネットワークへも単一リンクで接続される．アプリケーションの通信パターンを考慮して，異なる EPS スイッチに属する OCS ネットワークに接続されたノード間に光回線を割り当て，さらに他ノードからのメッセージを中継させることで，EPS ネットワーク上流で起こりうる混雑を回避する．シミュレーションによる評価の結果，全ノードの半数だけを OCS ネットワークに接続することで，フルバイセクションバンド幅 EPS ネットワークと同程度の性能を示し，高いバイセクションバンド幅を要求するアプリケーションに対して有効であることが確認できた．

HPC Application Performance Improvement by a Supplemental Optical Circuit Switching Network

Shin'ichiro Takizawa,[†1] Toshio Endo[†1] and Satoshi Matsuoka[†1,†2]

For large scale HPC systems which consist of many nodes, it will be unfeasible to construct a fully-connected network with high bisection bandwidth due to cost and power consumption, etc. We propose a hybrid network that is composed of an electronic packet switching (EPS) network with low bisection bandwidth and a high bandwidth supplemental optical circuit switching (OCS) network, and communication method on the network. In this network, each node connects to the EPS network with one link and partial nodes also do to the OCS network with another one link. We assign optical pathways to node pairs that are connected to the OCS network and are not in the same EPS switch by considering application's communication pattern. We avoid contentions on the EPS upstream network by letting these nodes relay messages from other nodes. By conducting simulations, we confirmed that our approach can achieve almost the same performance as that of full bisection EPS only network by connecting only half of nodes to the OCS network. We also confirmed that our approach can achieve high performance with applications that require high bisection bandwidth.

1. はじめに

マルチコアプロセッサを多数搭載する大規模 HPC システムでは，従来使われていた電気パケット交換方式を採用した高バイセクションバンド幅のクロスバーや Fat Tree などのノード間全対全接続ネットワークは，金銭コストや性能面で実現困難である．そのため，現状では Blue Gene で用いられている 3D トーラスネットワーク[1] のようなノード間接続数の少ないネットワークや，東京工業大学の TSUBAME Grid Cluster[2] で用いられているバイセクションバンド幅が低い Tree ネットワークなどが採用されている．TACC Ranger[3] や T2K システム[4] のような高バイセクションバンド幅ネットワークを持つシステムもあるが，将来は並列度のさらなる増加が予想され，そのようなネットワークの規模の維持は困難になると考えられる．

一方で HPC システム上で実行される MPI アプリケーションの多くには通信に局所性があり，各プロセスは一部の特定のプロセスとのみ通信を行う特徴がある[5]．通信局所性を持つアプリケーションに対しては，高いバイセクションバンド幅を提供せずとも，通信パターンに着目した通信最適化手法を用いることで実行性能向上を実現できる．その手法の 1 つとして，電気パケット（EPS: Electronic Packet Switching）

†1 東京工業大学
Tokyo Institute of Technology
†2 国立情報学研究所
National Institute of Informatics

ネットワークと光サーキット（OCS: Optical Circit Switching）ネットワークを組み合わせたネットワーク環境が提案されている[6)–8)]．EPS ネットワークとは，HPC システムで広く用いられている Ethernet や InfiniBand からなるネットワークである．OCS ネットワークは以下の特徴を持つネットワーク環境である．

- エンドツーエンドを光信号で通信を行う．
- 回線交換型ネットワーク．通信前後に 2 ノード間で光回線の確立・解放が必要であり，それには機械処理を要するため，ミリ秒オーダーの時間がかかる[9)]．
- 広帯域，低遅延，低消費電力．

これら既存研究ではアプリケーションの通信に合わせて光回線を割り当てることで，EPS ネットワーク上では遠く離れたノード間の通信を OCS ネットワーク上で高速に行うことが可能になる．しかしながら，複数の EPS，OCS ネットワークを必要とするため，ネットワーク規模が大きく構築が容易でない．

大規模 HPC システム用ネットワークとして，我々は各ノードが単一の低バイセクションバンド幅 EPS ネットワークと，単一 OCS ネットワークに接続された EPS-OCS ハイブリッドネットワーク環境を提案し，その環境での MPI 通信手法を提案した[10)]．アプリケーション通信パターンに従い EPS スイッチ間通信を行うノード間に光回線を割当て, 大容量の EPS スイッチ間通信を OCS ネットワーク上で中継転送する．これにより, 単一の OCS ネットワークであっても，アプリケーションの実行性能がフルバイセクションバンド幅の EPS ネットワークに匹敵することを確認した．この性能は，わずか 1/4 のノードのみ光回線を用いた状況で達成された．残りの 3/4 を使用することでさらなる性能向上を確認したが，微々たるものだった．OCS ネットワークの規模に無駄があり，さらに縮小してもこの性能を維持できると考えられる．

本研究では上記の研究で得た知見を基に，単一の小規模 OCS ネットワークを補助的に使用する EPS-OCS ハイブリッドネットワーク環境を提案する．OCS ネットワークの補助的利用とは，EPS ネットワークに接続されたノードの一部のみを OCS ネットワークに接続することである．さらに, 上記研究で提案した, OCS ネットワーク上で中継転送を行う通信手法に対する以下の 3 点の改良を提案する．一部のノードのみ OCS ネットワークに接続されている制約の下で, アプリケーションの通信パターンを可能な限り満たすように光回線を割り当てる．EPS スイッチ間の通信量に応じて, 通信の多い EPS スイッチ間の中継バンド幅を増強すべく, それらスイッチ下のノード間に優先的に光回線を割り当てる．中継ノードの負荷を減らすために, 中継ノードの EPS リンクバンド幅を増強する．シミュレーションによる評価の結果，EPS ネットワークのバイセクションバンド幅が低い場合でも，全ノードの半数だけを OCS ネットワークに接続することで，フルバイセクションバンド幅 EPS ネットワークと同程度の性能を示し，また，高いバイセクションバンド幅を要求するアプリケーションに対して有効であることが確認できた．

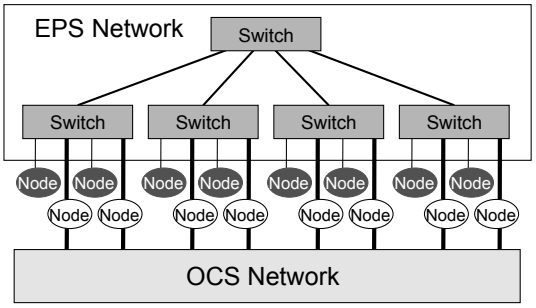

図 1 OCS ネットワークの補助的利用環境

2. OCS ネットワークの補助的利用の提案

図 1 に OCS ネットワークを補助的に接続したネットワーク環境を示す．EPS ネットワークは既存の HPC システムで用いられいている，InfiniBand 等を利用したパケット交換型のノード間相互通信網を表す．図中では 2 階層の Tree トポロジで描かれているが，全ノードが全対全で接続されていればどのようなトポロジでも構わなく，低バイセクションバンド幅であっても構わない．システム運営に必要となるアカウントなどの情報サービス，ストレージ通信はこの EPS ネットワークを用いて行われるとする．我々は，この EPS ネットワークに属するノードの一部を単一の高バンド幅 OCS ネットワークに接続する．図 1 では，各末端 EPS スイッチ下の 2 ノードが OCS ネットワークに接続されている．以降，OCS ネットワークに接続されたノードを OCS ノードと呼ぶ．

各 OCS ノードは単一リンクで OCS ネットワークに接続されるため，OCS ネットワーク上では一度に通信できる宛先ノードは 1 つに限られる．接続に制限のある OCS ネットワークを最大限に活用するため，頻繁に大容量通信を行うノード間に光回線を割り当て，EPS ネットワークのショートカット経路として使用する．OCS ネットワークは任意の OCS ノード間で回線が確立できるよう構成する．また，3 章で述べるように，OCS ノードは同一 EPS スイッチ下の他ノードからのメッセージを他 EPS スイッチ下ノードへ中継転送するため，OCS ノードの EPS リンクバンド幅は他のノードの EPS リンクバンド幅よりも高く構成する．

3. 提案 MPI 通信手法

OCS ネットワークを補助的に用いた環境での通信の特徴をまとめ，アプリケーション通信パターンを考

慮して光回線割当てを行う，通信手法を提案する．

3.1 OCS 補助的利用環境での通信の課題

EPS ネットワークは全ノードに渡り提供されているものの，バイセクションバンド幅が低い場合にはネットワーク上流で混雑が起こりうる．OCS ネットワークでは高バンド幅光回線を用いて任意の OCS ノード間の通信をショートカットできるが，一度に通信できる宛先ノードは 1 つであり接続に制限があること，回線確立／解放に長時間要すことより，複数の通信相手との頻繁な通信には不向きである．さらに，OCS ノードは全ノードの一部であるため，OCS ノード上で大容量通信を行うプロセスが実行されていない限り光回線は有効利用できず，さもなくば，EPS 上流ネットワークに大容量メッセージを送出することになり，帯域圧迫，性能低下へとつながる．EPS，OCS のそれぞれに利点，欠点があるため，互いの欠点を補うよう通信を行う必要がある．

3.2 通信パターンに応じたメッセージ経路

アプリケーションで通信されるメッセージのサイズ，宛先プロセスの位置に応じて異なる通信経路をとる．使用される全経路パターンを図 2 に示す．なお，この図ではすでに OCS node である A と D の間，および F と H の間に光回線が割り当てられているとする．

メッセージサイズが OCS ネットワークの帯域遅延積よりも小さい場合，メッセージは EPS ネットワークのみを用いて通信される（図中 $B \Rightarrow C$，$E \Rightarrow J$）．帯域遅延積以下のサイズの場合，光回線のバンド幅を十分に活かしきれないためである．また，宛先プロセスが同一 EPS スイッチ下のノード上で動作している場合には，メッセージサイズによらず EPS ネットワークのみを用いた通信を行う（図中 $B \Rightarrow C$）．これには 2 つ理由がある．1 つは単一 EPS スイッチ内の通信であれば，高いスイッチ内バンド幅を活かした高速通信が可能だからである．もう 1 つは，接続に制限のある OCS ネットワークを，すでに十分ノード間距離の近い EPS スイッチ内の通信に用いるには贅沢だからである．

メッセージサイズが OCS ネットワークの帯域遅延積より大きく，宛先プロセスが他 EPS スイッチ下ノード上で動作している場合にのみ，OCS ネットワークを用いる．大容量通信を行う OCS ノード間を光回線で接続し，その通信を OCS ネットワーク上で行う（図中 $A \Rightarrow D$）．さらに，OCS ノードでないノード間の大容量通信を OCS ノードが中継する（図中 $G \Rightarrow F \Rightarrow H \Rightarrow I$）．このように，従来なら EPS ネットワーク上流に流れていたメッセージを OCS ネットワーク側に追いやることで，EPS ネットワーク上流の帯域圧迫を避け，混雑による遅延を回避する．さらに EPS スイッチ間に複数の光回線を割り当てることで，スイッチ間通信のバンド幅増強も行える．一方で OCS ノードの EPS リンクが混雑しうるが，バンド幅を増強することで緩和する．末端リンクであるため，EPS 上流リンクのバンド幅を増強することより容易であると考えている．

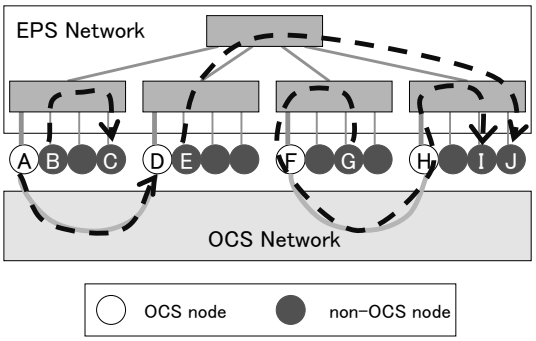

図 2 EPS-OCS ハイブリッドネットワーク上の全通信経路

3.3 大容量通信のための経路作成方法

以上の通信手法を実現するためには，大容量通信用の経路を作成する必要がある．経路作成には，アプリケーションの通信パターン，プロセスの配置情報を基に，OCS ノード間で光回線を確立し，ノード間のルーティングテーブルの計算が必要となる．以降 3.3.1 から 3.3.3 にかけて経路作成方法を説明する．

3.3.1 プロセス配置情報，通信パターンの取得

プロセス配置情報として，各プロセスを実行しているノードの ID（EPS ネットワーク上の IP アドレスなど）を取得し，ノード，およびプロセスを EPS スイッチ単位でグルーピングする．また，OCS ノードの ID も取得する．前者はノード間通信を行う際に EPS スイッチ内で閉じた通信か，EPS スイッチ間通信かを判断するために，後者はどのノードが光回線を用いた通信を行えるのか判断するためである．

通信パターンとして，プロセス毎に OCS ネットワークの帯域遅延積以上のサイズのメッセージ送信の受信相手と，その相手への平均送信メッセージサイズを取得する．さらに 1 ノード上で複数プロセスを実行している場合，プロセス間の通信パターンからプロセス配置情報を基に，ノード間の通信パターンを抽出する．

3.3.2 光回線の割り当て

3.3.1 で取得したプロセス配置と通信パターンを基に，可能な限り通信パターンを満たすように OCS ノード間で光回線を確立する．光回線割当てアルゴリズムの概要を図 3 に示す．

Onodes は MPI プロセスを実行する OCS ノードの集合である．CommPaths はこれら OCS ノード間で行われる通信のうち，EPS スイッチをまたぐ通信の集合である．各要素は，通信を行う 2 ノードの ID，通信量からなる．

OCS ノード間に光回線を割り当てる際には，まず OCS ネットワーク側で接続する EPS スイッチのペアを定める（2 行目）．続いて，実際に光回線を割り当

```
入力
Onodes     := OCS ノードの集合
CommPaths  := EPS スイッチをまたぐ OCS ノード間の
              通信の集合
出力
OnPairs    := 光回線で接続する OCS ノードペアの集合

1: while (Onodes または CommPaths が空でない)
2:    SwPair := 光回線で接続する EPS スイッチのペア
                を選択
3:    OnPair := 光回線を割り当てる OCS ノードのペア
                を選択 (Onodes, CommPaths,
                        SwPair)
4:    OnPairs   << OnPair
5:    Onodes    -= OnPair
6:    CommPaths -= OnPair が関係する通信
7: end
```

図 3 回線割当てアルゴリズム

てる OCS ノードのペアを，先ほど求めた EPS スイッチのペア，使用可能な OCS ノードの集合，考慮すべき OCS ノード間通信パターンを基に決定する（3 行目）．2 行目，3 行目の処理の詳細は後述する．選ばれた OCS ノードペアを記録（4 行目）後，光回線で接続された OCS ノードにはこれ以上光回線を割り当てることができないので，OCS ノードの集合から除外する（5 行目）．同様に，OCS ノードが数多くの EPS スイッチ間通信を行うとしても，接続された相手以外とは直に OCS ネットワーク上で通信を行えないため，すべての関係する EPS スイッチ間通信を集合から除外する（6 行目）．

以上の処理を終了条件となるまで繰り返す（1 行目）．Onodes が空となる状況は，すべての OCS ノード間で光回線が割り当てられた場合である．CommPaths が空となる状況は，アプリケーションの通信を満たすように十分な数の光回線が割り当てられた場合，あるいは，宛先 OCS ノードが既に他の OCS ノードと接続済みで回線を確立できない場合である．後者の状況となった場合には光回線が割り当てられない OCS ノードが生じるが，これ以上通信パターンを満たす必要がない，また満たせないため，使用しない．

以下に 2 行目，3 行目の詳細を述べる．

EPS スイッチペアの選択　あらかじめノード間通信パターンから各 EPS スイッチ間の通信数を求めておく．通信数とは，EPS スイッチ間で行われるノード間の通信リンクの数である．

まず，通信を行う EPS スイッチ間に 1 本ずつ光回線が割り当てられるように，通信数の少ない順に EPS スイッチペアを選択する．通信を行うすべての EPS スイッチ間に最低 1 本の光回線が割り当てられた後は，「EPS スイッチ間通信数 − 割当て済み光回線数」を計算し，通信の多い EPS スイッチ間バンド幅を増強すべく，この値の大きい EPS スイッチペアを選択する．

OCS ノードペアの選択　選択された EPS スイッチ間通信を行う OCS ノードペアを，ノード間通信量（メッセージサイズ）の多い順に選択する．過去の OCS ノードペア選択状況によって，選択された EPS スイッチ間通信を行うノードペアを選択できない場合もある．そのような場合は任意のノードペアを選択する．

3.3.3 通信経路の作成

以上で割り当てた光回線を用いた，EPS-OCS 両ネットワークに渡る大容量メッセージの通信経路を作成する．そのためのルーティングテーブルを以下の手続きより生成する．

(1) 各ノードは 1 ホップで到達可能な宛先ノードへの経路を作成する．すなわち，同一 EPS スイッチ下のノードへの経路と，OCS ノードの場合，光回線で接続された相手ノードへの経路である．
(2) OCS ノードは光回線で接続されたノードと経路情報を交換し，宛先ノードへのホップ数が最小になるように，フォワーディングテーブルを更新する．
(3) 各ノードは，同一 EPS スイッチ下の OCS ノードの経路情報を取得し，手順（2）同様にフォワーディングテーブルを更新する．
(4) 手順（2），(3）を最大 EPS スイッチ数回繰り返す．

手順（4）で EPS スイッチ数分繰り返すのは，複数の光回線を使用したホップ数の長い通信経路を考慮するためである．しかしながら，そのような経路は遅延が大きくなるために除外し，繰り返し回数を削減することも可能である．この場合や，OCS ネットワークの規模が小さく十分な光回線を割り当てることができない場合には，ノード間の大容量通信を OCS ネットワークを用いて中継できないことがある．このときは大容量通信であっても，図 2 中の経路 $E \Rightarrow J$ のような，EPS ネットワーク上流を用いた通信を行う．

4. 評　　価

OCS ネットワークを補助的利用した環境と提案通信手法の組み合わせと，EPS ネットワークのみを用いた場合，フルバイセクションバンド幅 EPS ネットワークとの実行時間比較をシミュレーションにて行う．本提案の評価を行う際には，通信パターンを基に光回線を割当て，フォワーディングテーブルを作成済みとした．

4.1 実 験 設 定

4.1.1 アプリケーション

2 次元格子上の隣接通信と，NAS Parallel Benchmarks（NPB）の CG，IS，LU[11] の 4 アプリケーショ

ンを用いた．256 プロセスを使用し，NPB の 4 アプリケーションでは問題サイズを C とした．2 次元格子上の隣接通信では，16 × 16 の格子に配置した各プロセスが隣接する最大 4 プロセスとそれぞれ 4MB のメッセージを交換し合う．格子の第 1 行第 1 列から行方向に順にランク（MPI プロセス ID）を割り振った．各アプリケーションは同じ計算の繰り返し処理からなっており，本実験では 5 イテレーションまで実行した．IS は全対全通信を繰り返し実行する，局所性のないアプリケーションである．

4.1.2 シミュレーション環境パラメータ

64 ノードからなる環境を想定してシミュレーションを行った．EPS ネットワークのトポロジとしては図 1 と同様な 2 階層の Tree トポロジを用いた．末端 EPS スイッチ構成は，16 ポートスイッチ 4 機とした．OCS ネットワークの規模として，OCS ノード数 4，8，16，32，64 の 5 通りを用いた．EPS スイッチあたりの OCS ノード数をバランスし，各スイッチ以下 1，2，4，8，16 と配置した．

詳細なシミュレーションパラメータを表 1 に示す．CPU core speed は後述するシミュレータで用いるためのアプリケーション MPI 関数トレースを取得した環境の CPU 速度である．各ノード 4CPU core からなるとし，1 ノード上で 4 プロセスを実行した．EPS ネットワークの上流リンクのバンド幅として，提案ネットワークと EPS ネットワークのみの場合（表中 Ours & EPS only）は 20Gbps とした．上流リンクはストレージなど他の通信でも用いられているとし，バンド幅を小さくすることで仮想的な混雑状況を表現した．OCS ノードの EPS リンクのバンド幅は，OCS リンクと同じ 20Gbps とした．OCS ネットワークの帯域遅延積は 10,000bit となる．

4.1.3 プロセス配置

次の 2 パターンの配置を用いた．

Sequential 各 EPS スイッチ下，各ノード上にプロセスを連続配置する．すなわち，プロセス 0 から 63 を 1 つの EPS スイッチ下に配置し，プロセス 0 から 3 を 1 ノードに，プロセス 4 から 7 を別の 1 ノード上で実行する．実行する MPI プロセスのランクの合計が小さいノードを OCS ノードとする．

CommPattern アプリケーションの通信パターンからプロセス間通信グラフを作成し，4 分割し，グループごとに EPS スイッチ下に配置する．さらに，グループごとにプロセスを 16 分割し，1 ノードに 4 プロセス配置する．グラフ分割ライブラリ metis[12] を用い，グループ間の通信量が最小となるように分割した．EPS スイッチをまたぐ通信量が多いノードを OCS ノードとする．

評価環境上でそれぞれの配置を行った場合の，各アプリケーションの EPS スイッチ間通信量を表 2 にまとめる．CG では CommPattern 配置の方が通信量が増えている．metis はヒューリスティックに基づいてグラフ分割を行うために，最適に分割できるとは限らず，精度に問題があるためである．

表 1 シミュレーション環境パラメータ

Parameter	Value
Node Parameter	
CPU core speed	2.0GHz
Number of cores	4
Propagation delay of intra-node comm.	100ns
Bandwidth of intra-node comm.	68Gbps
EPS Network Parameter	
One link propagation delay	500ns
Bandwidth of upstream link (Ours & EPS only)	20Gbps
Bandwidth of upstream link (Full Bisection EPS only)	160Gbps
Bandwidth of downstream link	10Gbps
Bandwidth of OCS node's downstream link	20Gbps
MTU	4096Bytes
OCS Network Parameter	
Propagation delay	500ns
Bandwidth	20Gbps
MTU	4096Bytes

表 2 各アプリケーションの EPS スイッチ間通信量

Application	Sequential	CommPattern
隣接通信	1920MB	1480MB
CG	2142MB	2544MB
IS	3461MB	3461MB
LU	108MB	79MB

4.1.4 シミュレータ

評価のためにシミュレータを作成した．このシミュレータは，MPI アプリケーションを実機で実行したときに取得した MPI 関数トレースを入力として，環境パラメータを当てはめ再生し，実行時間を求める．MPI 関数トレースとして，各関数へ渡された引数と，呼び出し時刻，終了時刻を記録する．

シミュレータは CPU 処理，MPI 通信処理をトレースファイル中のすべての関数トレースを処理しきるまで繰り返す．CPU 処理時間として，MPI 関数呼び出し時刻から直前の MPI 関数の終了時刻を差し引いた値を用いた．MPI 通信処理時間は次のように計算した．一対一通信の場合，リンクごとの「遅延 + メッセージサイズ / バンド幅」の和を通信時間とした．OCS ネットワーク上での通信は，光回線を確立すればスイッチ内では待ちは発生しないため，単一リンクとして処理した．EPS ネットワークでは，各 EPS スイッチは Store-and-Forward 方式とし，EPS スイッチ内通信であれば 2 リンク，EPS スイッチをまたぐ通信であれば 4 リンクとして処理した．また，ノードおよび EPS スイッチでメッセージを受信する際に，同一宛先にメッセージが集中したときの混雑をシミュレー

トするため，先に受信したメッセージの時刻を考慮し，後続するメッセージの受送信を遅らせた．MTU サイズを超えるメッセージは MTU サイズに収まるように分割し，複数回送信処理を行う．EPS，OCS ネットワーク間の中継を行う際には，分割されたメッセージをすべて受信した後に中継する．集団通信の場合，MPICH2 で用いられているアルゴリズム通りに一対一通信の組合せとして通信時間を計算した[13]．

4.2 結　　果

図 4 に Sequential 配置の，図 5 に CommPattern 配置での各アプリケーションの結果を示す．横軸はアプリケーション，縦軸は EPS ネットワークのみを用いた場合の実行時間に対する，各ネットワークでの実行時間の短縮率を表す．値が大きいほど短い時間で実行を終えたことを表し，処理速度の向上を意味する．凡例「FB EPS」はフルバイセクションバンド幅の EPS のみのネットワークを，「Ours 数字」は提案ネットワークにおいて OCS ノード数を指定の数にした場合を表す．

隣接通信の結果をみると，OCS ノード数が全ノード数の半数の Ours 32 の場合，Sequential 配置では EPS に対して 26%の実行時間短縮が確認できた．しかしながら，FB EPS の方が 31%の短縮と，効果が高い．一方，CommPattern 配置の場合は，FB EPS は 3%の短縮しか得られていないが，Ours 32 では 9%の短縮が確認できた．このフルバイセクションバンド幅 EPS ネットワークの短縮率の劇的な減少の理由は，EPS 上流リンクを用いた通信量が減り，上流での混雑が減ったため，EPS のみのネットワークの実行時間が短縮されたことによる．提案手法では EPS スイッチ間通信を行うノード同士を光回線で直接接続し，低遅延，高バンド幅通信を行うことで，さらなる性能向上を実現している．このように，通信を考慮してプロセス配置をすることで性能向上することは知られていたが，提案手法を用いることでさらなる性能向上が実現できることが確認できた．

隣接通信の結果の特徴的な振る舞いとして，Ours 8 より，回線数の少ない Ours 4 の方が性能が優れている点が挙げられる．これは以下の理由による．OCS ノード数が 4 の場合，使用できる光回線数は 2 本のため，すべての EPS スイッチ間通信を OCS ネットワーク上で行うことができず，一部の通信は EPS 上流リンクを用いて行われる．一方，OCS ノード数が 8 の場合，4 本の光回線でより多くの EPS スイッチ間を接続でき，より多くの EPS スイッチ間通信を OCS ネットワーク上で行える．特に Sequential 配置の場合にはすべての EPS スイッチ間通信が OCS ネットワークに移譲された．しかしながら，OCS ノードでの中継混雑や，宛先ノードに到着するまでのホップ数の増加により性能低下が起こる．CommPattern 配置の Ours 8 で 46%以上性能低下しているが，このときの最長経路は 3 つの EPS スイッチをまたぐものであり，遅延の大きい経路を選択したことによる．グラフ分割の精度の問題で，逆に Sequential 配置よりも通信を行う EPS スイッチペアが増えてしまったためである．

この振る舞いは，以降で説明する NPB アプリケーションでも確認された．上記の問題は OCS ノード数をさらに増やすことで解決できる．EPS スイッチ間の OCS ネットワーク上でのリンクの本数，バンド幅の増強となるためである．なお，EPS 上流リンクのバンド幅がさらに小さい場合には，このような性能の逆転現象は起こらないと考えられる．

NPB の結果を見ると，IS のみ大幅な性能向上が確認でき，その他では効果が少ないと確認できる．IS において Ours 32 のとき，Sequential 配置では 16%の短縮であり，FB EPS と同程度の向上であった．CommPattern 配置では Ours 32 は 17%の短縮であり，FB EPS の 14%を上回った．IS は全対全通信を行うため，高いバイセクションバンド幅が要求される．OCS ノード数を増やすことで，提案手法は複数の光回線で EPS スイッチ間を接続することになり，スイッチ間通信のバンド幅増強，混雑の削減につながり，フルバイセクションバンド幅 EPS ネットワークの性能を上回ることが可能となる．CG は IS 同様に EPS スイッチ間通信量は多い（表 2）が，どちらの配置でも FB EPS の短縮率が小さいことから，IS ほど性能がバイセクションバンド幅依存ではないとわかる．LU はそもそも EPS スイッチ間通信量が少ないため（表 2），CG 同様，バイセクションバンド幅依存でない．

以上より，提案手法は高いバイセクションバンド幅を要求するアプリケーションに対して有効であると言える．そうでないアプリケーションであっても，EPS 上流リンクがストレージや他のノード，他のアプリケーションによる通信で圧迫されている場合にも有効である．また，隣接通信，IS の結果より，全ノードの半数だけを OCS ネットワークに接続することで，フルバイセクションバンド幅 EPS ネットワークと同程度の性能を発揮することが確認できた．

5. 関 連 研 究

EPS ネットワークと OCS ネットワークを利用するネットワークは他にも数多く提案されている[6)–8)]．Barker らは，各ノードが低バンド幅 EPS ネットワークと，複数の OCS ネットワークに接続するネットワークを提案している[6)]．EPS ネットワークは小規模メッセージ通信，集団通信に使用し，OCS ネットワークは一対一の大規模メッセージ通信に使用される．また，同ネットワーク上で複数 OCS ネットワークに渡り各ノードが大規模一対一メッセージをフォワードする手法も提案している[7)]．Kamil らは低バンド幅 EPS ネッ

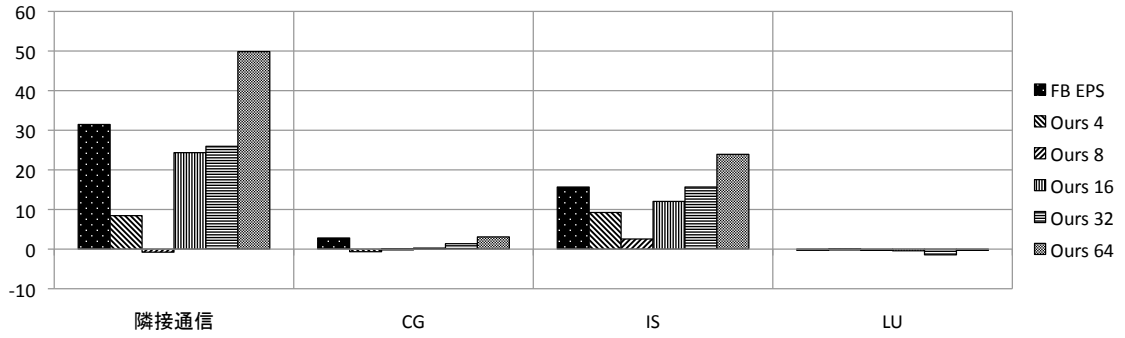

図 4 Sequential 配置での EPS のみのネットワークに対する性能向上率

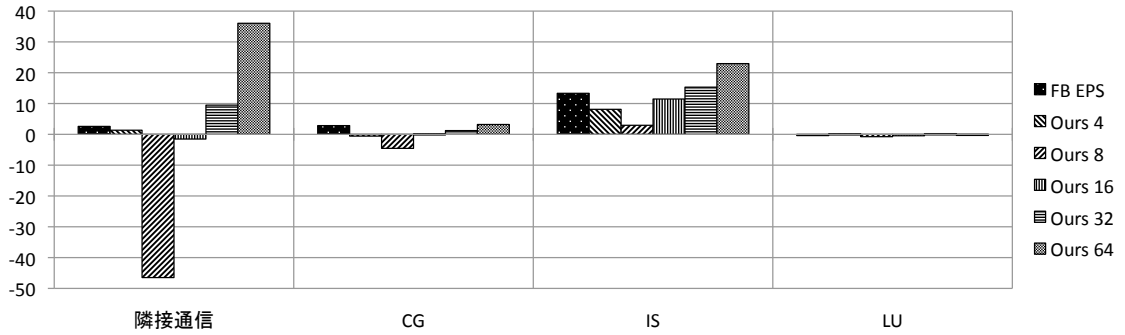

図 5 CommPattern 配置での EPS のみのネットワークに対する性能向上率

トワークとノードとの間に OCS ネットワークを挿入した構成を持つハイブリッドネットワーク HFAST を提案している[8]．従来ならプロセスマイグレーションが用いられていたところを，光回線を割当て直すことで，通信を行う任意の 2 ノードを同一 EPS スイッチ下に配置し，通信最適化を行う．HFAST は一対一の大規模メッセージ通信に使用され，小規模メッセージ通信，集団通信には別の EPS ネットワークが使用される．これら既存研究では複数の EPS，OCS ネットワークを使用する．特に全ノードを OCS ネットワークに接続する必要がある．必然的にネットワーク規模が大きくなり，金銭コスト，消費電力が大きくなる問題がある．また，複数のネットワークにメッセージが分散されるため，個々のネットワークの利用効率が我々の提案よりも低くなる．

過去の研究で我々は，上記既存研究のネットワーク規模の問題を解決するべく，単一 EPS ネットワークと，それと同規模の単一 OCS ネットワークからなるネットワーク環境，およびその環境での通信手法を提案した[10]．本研究では，OCS ネットワークの規模を縮小し，さらに OCS ノードの EPS リンクバンド幅を増やすことで中継負荷削減を行った．通信手法については，一部のノードのみ OCS ネットワークに接続されている制約を考慮して光回線割当てを行った点，EPS スイッチ間通信数を考慮して通信数の多いスイッチ間に光回線を多く割り当てた点が異なる．小規模 OCS ネットワークを補助的に使用することで，過去の研究，および，フルバイセクションバンド幅 EPS ネットワークと同程度の性能を示すことを実証した．

アプリケーションの通信パターンを活かした通信最適化手法には，プロセスマイグレーションを行うもの[14]，プロセスをネットワーク上に最適配置するもの[15),16)] がある．しかしながら，プロセスマイグレーション手法では移動先プロセスの決定にかかるコスト以外にも，メモリイメージ転送コストがかかる．今後システムがより大規模化されるにつれ，移動するプロセス数や，使用メモリ量も増加していくと考えられるので，転送コストは大きくなり，効果が薄れる．Dixit-Radiya ら[15)] や，Bhanot ら[16)] はトポロジの固定されたネットワーク上で，アプリケーション通信パターンとネットワーク通信コストからプロセスの最適配置を求めている．しかしながら，プロセス最適配置手法だけではネットワークトポロジに制約されるため，通信パターンを活かしきれない．トポロジの再構成が可能なネットワークと組み合わせることで，より効果を発揮する手法である．

6. ま と め

大規模 HPC システム用のネットワークとして，EPS

ネットワークを使用するシステムで OCS ネットワークを補助的に使用すること，および，その環境でのアプリケーション通信パターンを考慮した MPI 通信手法を提案した．OCS ネットワークの補助的利用とは，EPS ネットワークに接続されたノードの一部を単一小規模 OCS ネットワークへも接続することである．提案通信手法では，アプリケーションの通信パターンを可能な限り満たすように，OCS ネットワークに接続されたノード間に光回線を割当て，それらノードが EPS スイッチをまたぐ大容量メッセージを中継転送する．評価の結果，EPS ネットワークのバイセクションバンド幅が低い場合でも，全ノードの半数だけを OCS ネットワークに接続することで，フルバイセクションバンド幅 EPS ネットワークと同程度の性能を示すことを確認した．高いバイセクションバンド幅を要求するアプリケーションに対して特に有効であることも確認できた．

今後の課題として，光回線割り当てアルゴリズムの改良を考えている．今回の提案手法では，光回線を割り当てる際に条件を満たす複数の OCS ノード間通信がある場合，単純に通信量により選択している．この場合，各 OCS ノードは 1 つの光回線しか使用できないため，他の重要な通信に光回線を割り当てることができず，性能低下となる可能性がある．アプリケーション全体の通信状況を考慮して光回線を割り当てるように変更したいと考えている．さらに，より多くのアプリケーションでの評価，数千プロセス規模の大規模環境での評価を行う．

謝辞 本研究の一部は科学研究費補助金特定領域研究 (18049028)，および，JSPS グローバル COE プログラム「計算世界観の深化と展開」の補助による．

参 考 文 献

1) Davis, K., Hoisie, A., Johnson, G., Kerbyson, D.J., Lang, M., Pakin, S. and Petrini, F.: A Performance and Scalability Analysis of the BlueGene/L Architecture, *SC '04: Proceedings of the 2004 ACM/IEEE conference on Supercomputing*, Washington, DC, USA, IEEE Computer Society (2004).
2) Matsuoka, S.: The Road to TSUBAME and Beyond, *High Performance Computing on Vector Systems 2007*, Vol.6, pp.265–267 (2007).
3) : TACC HPC Systems, `http://www.tacc.utexas.edu/resources/hpcsystems/`.
4) : T2K Open Supercomputer Alliance, `http://www.open-supercomputer.org/`.
5) Shao, S., Jones, A. K. and Melhem, R.: A Compiler-based Communication Analysis Approach for Multiprocessor Systems, *Proceedings of the 20th IEEE International Parallel and Distributed Processing Symposium* (2006).
6) Barker, K. J., Benner, A., Hoare, R., Hoisie, A., Jones, A.K., Kerbyson, D.J., 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, *Proceedings of the 2005 ACM/IEEE conference on Supercomputing* (2005).
7) Barker, K.J. and Kerbyson, D.J.: Performance Analysis of an Optical Circuit Switched Network for Peta-Scale Systems, *Euro-Par 2007*, pp.858–867 (2007).
8) Kamil, S., Pinar, A., Gunter, D., Lijewski, M., Oliker, L. and Shalf, J.: Reconfigurable Hybrid Interconnection for Static and Dynamic Scientific Applications, *ACM International Conference on Computing Frontiers* (2007).
9) Dobbelaere, P.D., Falta, K., Fan, L., Gloeckner, S. and Patra, S.: Degital MEMS for Optical Switching, *Communications Magazine, IEEE*, Vol.40, pp.88–95 (2002).
10) 滝澤真一朗，遠藤敏夫，　松岡聡：次世代光インターコネクトでの MPI 通信に関する研究，コンピュータソフトウェア (to appear) (2008).
11) der Wijngaart, R. F.V.: NAS Parallel Benchmarks Version 2.4, Technical Report NAS Technical Report NAS-02-007, NASA Ames Research Center (2002).
12) : METIS - Family of Multilevel Partitioning Algorithms, `http://glaros.dtc.umn.edu/gkhome/views/metis/`.
13) Thakur, R., Rabenseifner, R. and Gropp, W.: Optimization of Collective Communication Operations in MPICH, *International Journal of High Performance Computer Applications*, Vol.19, No.1, pp.49–66 (2005).
14) Maghraoui, K.E., Desell, T., Szymanski, B.K., Teresco, J.D. and Varela, C.: Towards a Middleware Framework for Dynamically Reconfigurable Scientific Computing, *Grid Computing and New Frontiers of High Performance Processing* (Grandinetti, L., ed.), Advances in Parallel Computing, Vol.14, Elsevier, pp.275–301 (2005).
15) Dixit-Radiya, V. A. and Panda, D. K.: Task Assignment on Distributed-Memory Systems with Adaptive Wormhole Routing, *the Fifth IEEE Symposium on Parallel and Distributed Processing*, pp.674–681 (1993).
16) Bhanot, G., Gara, A., Heidelberger, P., Lawless, E., Sexton, J.C. and Walkup, R.: Optimizing task layout on the Blue Gene/L supercomputer, *IBM Journal of Research and Development*, Vol.49, No.2/3, pp.489–500 (2005).