2008－HPC－116 (12)
2008／8／5

光ネットワークの補助的利用によるHPC性能向上

滝澤 真一朗[†1] 遠藤 敏夫[†1] 松岡 聡[†1,†2]

将来のペタスケール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コストや電力消費の問題で実現困難である．この問題を解決するために，低バイセクションバンド幅電気パケット（EPS）ネットワークと共に高バンド幅光サーキット（OCS）ネットワークを補助的に使用するネットワーク，およびその上でのEPS-OCS間で中継転送を行うMPIアプリケーション用通信手法を提案する．OCSネットワークに接続された一部のノードが，同一EPSスイッチ下の他のノードからの他EPSスイッチ下ノードへの転送を中継する．シミュレーション評価の結果，EPSネットワークのみの場合より最大で30%の実行時間短縮を確認できた．

HPC Performance Improvement by Supplementing a Small Optical Network

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

For future peta-scale HPC systems, it will be unfeasible to construct a fully-connected network with high bisection bandwidth due to cost and power consumption, etc. We propose a network which is composed of an electronic packet switching (EPS) network with low bisection bandwidth and a high bandwidth supplemental optical circuit switching (OCS) network, and a communication methodology where messages are relayed from EPS to OCS and vice versa for MPI applications. In this network, partial nodes connected to the OCS network relay messages from other nodes under the same EPS switch to nodes under other EPS switches. From results of simulations, we confirmed that our approach reduces execution time against EPS only network by 30 % at maximum.

1. はじめに

マルチコアプロセッサを多数搭載するペタスケール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，コストや性能面で実現困難となる．現状でも，例えばBlue Gene/Lでは65536機のプロセッサを接続数の少ない3Dトーラスネットワークに接続している．[1]．また，東京工業大学のTSUBAME Grid Cluster[2]では全対全接続を維持しているものの，上流のバンド幅が低い構成になっている．TACC Ranger[3]やT2K筑波[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．

この問題の解決策として，我々はバイセクションバンド幅の低い電気パケット（EPS: Electronic Packet Switching）ネットワークと高バンド幅かつ低消費電力の光サーキット（OCS: Optical Circuit Switching）ネットワークのハイブリッドネットワークを提案し，さらに，その環境上での光回線を用いたフォワーディング方式によるMPI通信手法を提案した．[5] 従来のようにEPSネットワークのバイセクションバンド幅を増加させるのではなく，EPSスイッチ間の大容量通信を高バンド幅のOCSネットワークにさせることで，コストや消費電力を押さえつつ，アプリケーション実行性能を上げることを目的とした．その結果，ノード数に対して十分少ない光回線数でフルバイセクションバンド幅のEPSネットワークの性能を上回ることを確認した．このとき，半数以上のノードでOCSネットワークは使用されていない状況にあった．

本研究では，バイセクションバンド幅の低いEPSネットワークを使用したHPCシステムでOCSネットワークを補助的に利用すること，および，その環境でのEPS-OCS間で中継を行うMPIアプリケーション用通信手法を提案する．OCSネットワークの補助的利用とは，各EPSスイッチ下のOCSネットワークに

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

接続されたノード数が等しくなるように，一部のノードを OCS ネットワークに接続することである．MPI プロセスを実行するノードを EPS スイッチ単位でグルーピングし，EPS スイッチをまたぐ長距離通信では光回線で接続されたノードが他ノードのメッセージを EPS-OCS 間を中継して転送する．シミュレーションによる評価より，EPS ネットワークのみの場合より最大で 30%の実行時間短縮を確認できた．また，バイセクションバンド幅による影響が少ないことも確認した．

2. OCS ネットワークの補助的利用

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．上半分の 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．この EPS ネットワークに対し，我々は，末端の EPS スイッチ以下の OCS ネットワークに接続されたノード（OCS ノード）数が等しくなるように，ノードを単一の OCS ネットワークに接続する．そして，頻繁に大容量通信を行うノード間に高バンド幅光回線を割り当て，EPS ネットワークの迂回路として使用する．そのために OCS ネットワークは任意のノード間で回線が確立できるよう，full CLOS で構成する．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．EPS スイッチ下の OCS ノード数をバランスすることで，EPS スイッチ下の多くのノードが大容量通信を行う最悪ケースでも，安定した通信性能を保つことが可能である．OCS ネットワークではノードが一度に通信できる宛先ノードは 1 つに限られる．ノード間で通信を行う前に回線を確立しなければならず，また，通信終了後に回線を解放しない限り，他のノードと通信を行うことができない．さらに，回線確立／解放には機械操作が必要となるため，数ミリから数十ミリ秒の時間を要する．[6] そのため，ネットワークを広域に渡り使用する情報サービス，ストレージ等の通信には EPS ネットワークを用いる．

3. MPI 通信手法

本章では，OCS ネットワークを補助的に利用した HPC システム上で EPS，OCS の両方を活用した MPI アプリケーション用通信手法を提案する．各計算ノードは 1 プロセスのみを実行する仮定を設けているが，以下の手法はノード毎の通信パターンを計測，あるいはノード上の全プロセスの通信パターンを集約することで容易に複数プロセスの場合に適応可能である．

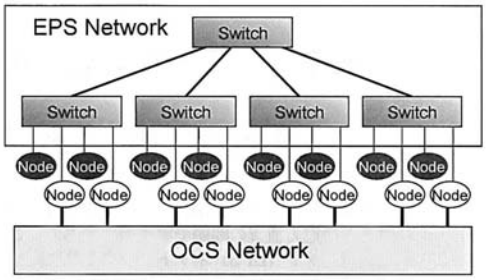

図 1 OCS ネットワークの補助的利用例

3.1 OCS を補助的利用した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が，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利用できるが，一度に通信できる宛先ノードは 1 つであること，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のため，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 EPS-OCS 間中継通信

我々は，OCS ノードが同一 EPS スイッチ下のノードからの他 EPS スイッチ下のノードへの大容量メッセージを OCS ネットワークへ中継転送する通信手法を提案する．中継を行うことで EPS 上流ネットワークを使用せず，帯域圧迫せずに済む．一方で OCS ノードの EPS リンクが混雑するが，末端リンクであるため，混雑を緩和するためのバンド幅増設は容易に行える．大容量メッセージと判断するメッセージサイズの閾値として OCS ネットワークの帯域遅延積を用いる．この閾値より小さいサイズのメッセージは EPS ネットワークのみを用いて転送される．帯域遅延積以下のサイズのメッセージを OCS ネットワークに送信すると，受信完了までに無転送時間が生じ，バンド幅の浪費となるためである．本通信手法による EPS，OCS の使い分けを以下にまとめる．

- EPS スイッチ内通信で済む場合は EPS ネットワークのみ用いる
- 小容量通信は EPS ネットワークのみ用いる
- EPS スイッチをまたぐ大容量通信には両ネットワークを用い，EPS-OCS 間の中継転送を行う

この手法で通信を行うために，あらかじめ MPI アプリケーションの各プロセスの位置情報，通信パターンを取得し，OCS ノード間で可能な限り通信パターンを満たすように光回線を確立し，EPS-OCS 間の通

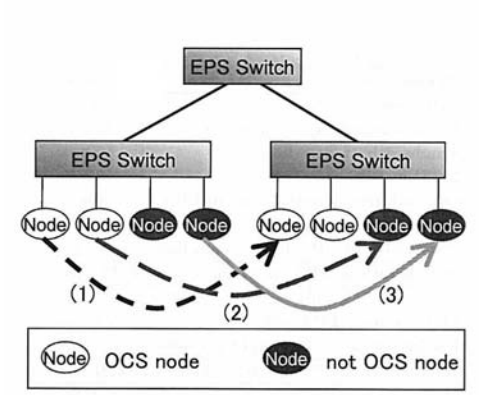

図 2　3 種類の通信パターン

信経路を計算する．この事前準備は以降の 3.2.1 から 3.2.3 の順で行う．

3.2.1　プロセス配置，通信パターンの取得

プロセス配置として，各プロセスを実行しているノードが接続している EPS スイッチの識別子（名前など）を得る．そして，EPS スイッチ単位でプロセスをグルーピングする．また，OCS ノード上で実行されているプロセスのランク（MPI プロセス ID）も取得する．どのプロセスが EPS-OCS 間中継可能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一対一送信の受信相手と，その相手への平均送信メッセージサイズを取得する．

3.2.2　光回線の割り当て

3.2.1 で取得した通信パターンと OCS ノードの配置状況を基に，OCS ノード間で光回線を確立する．ここでは，ノード間通信のうち，EPS スイッチをまたぐ通信にのみ光回線を割り当てる．総ノード数に対し OCS ノード数は少ないため，光回線を有効利用するためである．

アプリケーションの通信パターンと OCS ノードの配置より，EPS スイッチ間通信には図 2 に示す 3 パターンがある．

(1)　通信を行う 2 プロセスがともに OCS ノード上で実行されている場合．OCS ノード間に光回線を割り当てることで，OCS ネットワークを用いた直接通信が可能．

(2)　通信を行う 2 プロセスがそれぞれ OCS ノード，OCS ノードでないノード上で実行されている場合．

(3)　通信を行う 2 プロセスがともに OCS ノードでないノード上で実行されている場合．

パターン (2)，(3) ではノード間に光回線を割り当てることができない．

上記 3 種類の通信パターンを考慮して，次の 3 ステップの光回線割り当てを行う．

1. パターン (1) の通信に光回線を割り当てる．特定の EPS スイッチ間通信を行う複数の通信がある場合には，通信量の多い順に割り当てる．通信量の等しい複数の通信がある場合には，現状では，参加するプロセスのランクの小さいものを優先している．特定の EPS スイッチ間に多くの光回線を割り当ててしまわないよう，ラウンドロビン方式で EPS スイッチ間に割り当てる．また，ステップ 2 のために，各 EPS スイッチ間で確立した光回線の数を記録する．
   パターン (1) の通信すべてに光回線を割り当てた後でも，未使用の OCS ノードがあればステップ 2 へと進む．
2. パターン (2)，(3) の通信を，OCS ノード間の EPS-OCS 中継でまかなう為に EPS スイッチ間に光回線を割り当てる．あらかじめ通信パターンから各 EPS スイッチ間の通信数を計算しておく．このステップでは，光回線が割り当てられていない EPS スイッチ間に，スイッチ間通信数の多い順に光回線を割り当てる．通信数が等しい複数の EPS スイッチペアがある場合には，現状では EPS スイッチ識別子の小さいものを優先している．このステップにおいても各 EPS スイッチ間で確立した光回線数を記録する．
   すべての EPS スイッチ間通信に光回線を割り当てた後でも，未使用の OCS ノードがあればステップ 3 へと進む．
3. 通信数の多い EPS スイッチ間に光回線を増設割り当てする．各 EPS スイッチ間で，通信数とスイッチ間の割り当て済み光回線数の差分を計算する．この差分値の大きい EPS スイッチ間に順に光回線を割り当てる．ステップ 2 同様，差分値が等しい複数の EPS スイッチペアがある場合には，現状では EPS スイッチ識別子の小さいものを優先している．

ステップ 3 まで実行しても使用されない OCS ノードがあるが，それらは，1)EPS スイッチ間通信を行うものの，相手 OCS ノードが既に他の OCS ノードと回線確立済みで光回線を確立できない，2) そもそも EPS スイッチ間通信を行わない，という状況であるため使用しない．

3.2.3　通信経路の作成

3.2.2 で割り当てた光回線を用いて，EPS-OCS 両ネットワークに渡る大容量メッセージの転送経路を作成する．リンクバンド幅（の逆数）を基準とした距離ベクトル型アルゴリズムを用いてフォワーディングテーブルを作成し，それに従った経路を用いて転送を行う．光回線を用いた経路と，EPS ネットワークだけを用いた経路の距離が等しい場合は，前者を優先使用する．結果として，図 3 に示す 3 パターンの経路が出来上がる．

(1)　光回線を用いた OCS ノード間の直接通信経路

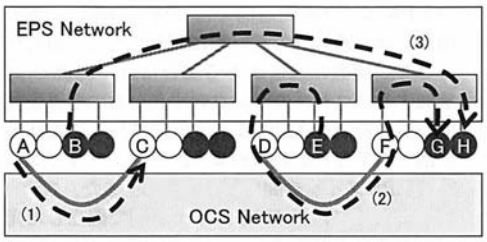

図 3　3 種類の大容量メッセージ通信経路

である．宛先ノードと直に光回線で接続されている場合に使用される．

(2)　OCS ノードが他のノードからのメッセージを中継転送する経路である．光回線に接続されていないノード，あるいは宛先ノードとは異なるノードに接続されている OCS ノードは，同一 EPS スイッチ以下の OCS ノードのうち，宛先ノードと同じ EPS スイッチ下の OCS ノードと接続されているノードに中継依頼する．ここで中継されるのは MPI メッセージであり，OCS ノードで MPI メッセージを復元した後にソフトウェアレイヤで中継することで，コモディティ NIC を用いた 2 ネットワーク間中継を可能にする．図 3 では，ノード E とノード G 間の通信経路として，$E \Rightarrow D \Rightarrow F \Rightarrow G$ と中継経路が使用されている．この経路では光回線は 1 本しか用いられていないが，複数の光回線を用いて複数の EPS スイッチを経由する長距離の中継経路も存在しうる．

(3)　EPS ネットワークのみを用いた経路である．互いに OCS ノードで接続されていない EPS スイッチ間，また，経路 (2) で示した中継経路を持ってしても到達できない EPS スイッチ間通信経路として使用する．OCS ノードの数が足りず，十分な数の光回線を確立できない場合にこのような経路が多用される．

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行う．さらに，バイセクションバンド幅が極端に低い EPS ネットワークを用いた評価も行った．

4.1 実 験 設 定

256 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 スイッチ構成として，8×32 と 16×16 の 2 パターンを採用した．それぞれ，32 ポートスイッチ 8 機，16 ポートスイッチ 16 機に相当する．OCS ネットワークの規模として，OCS ノード 64，128 の 2 通りを用いた．これより，EPS スイッチあたりの OCS ノード数は 8×32 の場合は 8 と 16，16×16 の場合は 4 と 8 となる．シミュレーションのパラメータを表 1 に示す．EPS と OCS を用いた提案環境（Ours）と EPS ネットワークのみを用いた環境（EPS only）では，上流リンクと下流リンクのバンド幅比率を 1:5 とした．またバイセクションバンド幅が極端に低い EPS ネットワークとして，上記比率が 1:20 の環境を用意した．

表 1　シミュレーションパラメータ

Parameter	Value
CPU speed of node	2.0GHz
Propagation delay in the OCS	120ns
Bandwidth of OCS link	4Gbps
MTU in EPS	4096B
One link propagation delay in EPS	40ns
Switching delay in EPS switch	420ns
Bandwidth of EPS downstream link	1Gbps
Bandwidth of EPS upstream link in **8x32 case**	
Ours and EPS only	6.4Gbps
Full bisection EPS only	32Gbps
low bisection in Ours and EPS only	1.6Gbps
Bandwidth of EPS upstream link in **16x16 case**	
Ours and EPS only	3.2Gbps
Full bisection EPS only	16Gbps
low bisection in Ours and EPS only	0.8Gbps

隣接通信（後述）と NAS Parallel Benchmarks (NPB) の CG，LU，MG[7] の 4 アプリケーションを用いた．ただし，NPB の各アプリケーションは 5 イテレーションまでの実行とした．隣接通信とは，16×16 の格子上に配置した各プロセスが隣接 4 プロセスとそれぞれ 4MB のメッセージを交換し合うアプリケーションである．格子の第 1 行第 1 列から行方向に順にランクを割り振る．各アプリケーションのプロセスは EPS スイッチ以下に連続になるように，かつ，ランクの小さいプロセスが OCS ノード上で実行されるように配置する．すなわち，EPS8 × 32，OCS64 ノードの構成では，ランク 0 ～ 31 が同一 EPS スイッチ下，ランク 32 ～ 63 が別の同一 EPS スイッチ下となり，ランク 0 ～ 7，32 ～ 39 のプロセスが OCS ノード上で実行される．また，大容量メッセージと判定する閾値は 8192 と設定した．

シミュレーションでは，計算処理時間と通信時間を足し合わせ実行時間を求める．計算処理時間として，実計算機上での実行時間を用いた．通信時間には「遅延 + メッセージサイズ / バンド幅」を用いた．ただし，複数のメッセージが同一リンクに集中する場合の混雑を模倣するため，過去のメッセージの送信時刻分，後続メッセージの送信を遅らせた．また，EPS ネットワークではメッセージは MTU サイズに分割し，EPS スイッチは Store-and-Forward 方式で動作する．

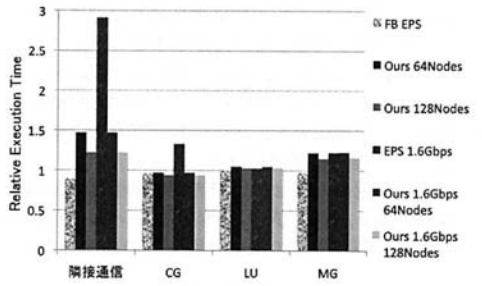

図 4　8 × 32 構成の結果

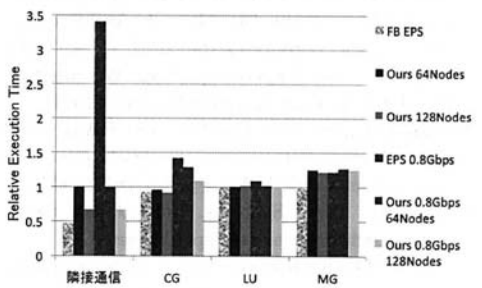

図 5　16 × 16 構成の結果

4.2　結果と考察

図 4 に 8 × 32 構成の，図 5 に 16 × 16 構成の結果を示す．縦軸は EPS only に対する相対実行時間を表す．凡例 FB EPS はフルバイセクション EPS only を，Ours 64Nodes，Ours 128Nodes はそれぞれ EPS only に 64，128 ノード分の OCS ネットワークを接続した提案を表す．EPS 1.6Gbps，Ours 0.8Gbps 64Nodes などバンド幅の記述があるものは低バイセクションバンド幅 EPS ネットワークを用いた結果である．

図 4 において，提案手法は隣接通信において OCS ノード数が 64 の場合に最大 47%実行時間が増加している．これは 4.1 で示したプロセス配置の場合，図 2 のパターン（1）の通信が存在せず，すべての EPS スイッチ間通信が中継転送になるため，中継ノードへの負荷が原因となっている．OCS ノード数を 64 から 128 に増やすことでこの負荷は減り，増加率も 22%に減り，性能が向上している．一方図 5 では実行時間短縮が見られる．OCS ノードが 128 の場合には EPS only に対し，33%の短縮となった．これは，16 × 16 構成ではパターン（1）の通信も存在し，さらに，EPS スイッチ下ノード数が減ったことによる中継負荷の減少が理由である．EPS ネットワークの構成，プロセス配置が大きく影響する結果となった．また，図 4，図 5 の両方において，低バイセクションバンド幅 EPS only は大幅な実行時間増加が見られるが，提案手法には影響が無い．この理由は，アプリケーション通信パターンを満たすように OCS ノード間に光回線が割り当てられ，EPS スイッチ間通信すべてが OCS ネットワークに流れ，EPS 上流リンクを使用しないためである．

表 2　光回線を割り当てられた OCS ノード数

Topology	隣接通信	CG	LU	MG
8x32 64nodes	56	64	56	64
8x32 128node	112	128	112	128
16x16 64nodes	60	60	60	64
16x16 128node	120	126	120	128

CG では，図 5 から確認できるように，低バイセクションバンド幅 EPS ネットワークとともに用いた場合に提案手法で実行時間が増加している．OCS ノードが 64 の場合に最大で 30%の増加であった．このとき EPS 上流スイッチは全 16 本のリンクから計 2GByte のデータを受信していた．一方で 8 × 32 の構成の場合，82KByte しか受信していない．この理由は，特定の EPS スイッチ間を光回線で接続できず，低バンド幅 EPS 上流リンクを使用した通信（図 3 経路（3））を行わざるを得なかったためである．3.2.2 で示した現状の光回線割り当てアルゴリズムでは，条件を満たす複数の回線があった場合に，他の通信状況を考慮せず，単純な方法で 1 つを選び出している．この場合，重要な通信に光回線が割り当てられず，EPS 上流リンクを使用せざるを得なくなり，性能低下することがある．今後の課題として対処する．

LU はバイセクションバンド幅は性能に影響を与えず，全パターンで同程度の実行時間であった．MG では提案手法は EPS only に対し，どの構成でも 20%ほど実行時間が増加した．MG では表 2 に示すように，すべての場合で全 OCS ノードを使用している．EPS スイッチをまたいだ長距離通信が他のアプリケーションより多く，中継負荷が原因である．

総じて，性能向上した場合もあれば，低下した場合も多い．特に OCS ノード数を全ノード数の 1/4 の 64 にした場合には性能低下の方が多かった．隣接通信の場合に，EPS ネットワーク構成，プロセス配置が問題となっていたため，性能低下したものに関しては，あらかじめ通信を行うプロセスの組を意識して OCS ノードに配置するなど，プロセス配置を工夫することでさらなる性能向上の余地があると考えている．

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．Barker らは，各ノードが単一の低バンド幅 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．[8] 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．[9] Kamilらは低バンド幅EPSネッ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のハイブリッドネットワークHFAST（Hybrid Flexibly Assignable Switch Topology）を提案している．[10] 従来のネットワークを用いた場合には局所性のあるノード同士の通信を最適化するためにはタスクマイグレーションが用いられていたが，HFASTでは光回線の切り替えで任意のノードを近づけることが可能であり，軽量な通信最適化が行える．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．これらは複数のEPS，OCSネットワークを使用するため，必然的にネットワーク規模が大きくなる問題がある．Kodiらはボード（ノード集合）内通信にEPSネットワークを用い，ボード間通信にOCSネットワークを用いるE-RAPIDを提案している．[11] E-RAPIDは通信頻度の高いボード間に光回線を多く割り当てる，通信頻度の低いボード間には転送レートを下げ消費電力を下げる，ことを行う．しかしながら，ボード数に応じて光送出ユニットの数を増やさなければならず，スケーラビリティに欠ける．

6. まとめ

将来のペタスケール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高バンド幅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EPS-OCS間で中継を行う通信手法を提案した．OCSネットワークの補助的利用とは，各EPSスイッチ下のOCSネットワークに接続されたノード数が等しくなるように，一部のノードをOCSネットワークに接続することである．提案通信手法では，MPIプロセスを実行するノードをEPSスイッチ単位でグルーピングし，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EPSスイッチ間通信に光回線を割り当て，光回線に接続されたノードが他ノードからのEPSスイッチをまたぐ通信メッセージを中継転送する．評価の結果，EPSネットワークのみを用いた場合より最大で30%の実行時間短縮と，バイセクションバンド幅による影響が少ないことを確認した．また，EPSネットワークの構成とプロセス配置により性能が大きく変わることも確認した．

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．現状では，光回線を割り当てる際に条件を満たす複数の通信がある場合，単純な方法で1つを選び出している．各OCSノードは1つの光回線しか持てないため，選び出された通信に光回線を割り当てることで，他の重要な通信に光回線を割り当てることができず，性能低下につながる可能性がある．そのため，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する．また，OCSノードのEPSリンクバンド幅を増加した環境での評価を行う．これにより，OCSノードの中継負荷が減り，性能向上が見込める．さらに数千プロセス規模のより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., et al.: 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) Takizawa, S., et al.: Locality Aware MPI Communication on a Commodity Opto-Electronic Hybrid Network, *Workshop on Large-Scale Parallel Processing* (2008).
6) Dobbelaere, P.D., et al.: Degital MEMS for Optical Switching, *Communications Magazine, IEEE*, Vol.40, pp.88–95 (2002).
7) der Wijngaart, R. F.V.: NAS Parallel Benchmarks Version 2.4, Technical Report NAS Technical Report NAS-02-007, NASA Ames Research Center (2002).
8) Barker, K. J., et al.: On the Feasibility of Optical Circuit Switching for High Performance Computing Systems, *Proceedings of the 2005 ACM/IEEE conference on Supercomputing* (2005).
9) Barker, K. J., et al.: Performance Analysis of an Optical Circuit Switched Network for Peta-Scale Systems, *Euro-Par 2007*, pp.858–867 (2007).
10) Kamil, S., et al.: Reconfigurable Hybrid Interconnection for Static and Dynamic Scientific Applications, *ACM International Conference on Computing Frontiers* (2007).
11) Kodi, A. et al.: Performance Adaptive Power-Aware Reconfigurable Optical Interconnects for High-Performance Computing (HPC) Systems, *International Conference for High-Performance Computing, Networking, Storage and Analysis (SC'07)* (2007).