次世代光インターコネクトでの MPI 通信性能の評価

Evaluation of MPI Communication Performance on Next Generation Optical Interconnect

滝澤 真一朗 † 遠藤 敏夫 †
Shin'ichiro TAKIZAWA Toshio ENDO
松岡 聡 †,††
Satoshi MATSUOKA

† 東京工業大学
Tokyo Institute of Technology
takizawa@matsulab.is.titech.ac.jp endo@gsic.titech.ac.jp
matsu@is.titech.ac.jp

†† 国立情報学研究所
National Institute of Informatics

将来の数万プロセッサ規模のシステムでは，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構築が難しい．そこで安価に構築可能な，上流リンクバンド幅が低い電気パケットネットワークと光サーキットネットワークの双方を活用するネットワークを提案する．光ネットワークは電気ネットワークの上流リンクでの混雑を避けるために，パケットスイッチをまたぐノード間通信にのみ，ショートカットとして使用する．このネットワーク上での MPI 通信手法として，通信パターンを解析し，あらかじめ必要な光回線を割り当て，プロセス間でメッセージをフォワードする手法を提案する．MPI 全対全通信をシミュレーションした結果，提案ネットワークは，パケットネットワークの上流リンクバンド幅を 2 倍に増強した環境と同程度の性能を達成できることを確認した．

1 はじめに

将来のペタスケール時代のスーパーコンピュータは，シングルプロセッサコアのクロック上昇率の頭打ちのため，マルチコアプロセッサを数千から数十万搭載した高度に並列化されたシステムとなりうる．そのようなシステムのノード間インターコネクトとして，過去のスーパーコンピュータやクラスタシステムで用いられていた，パケット交換方式を採用するクロスバーや Fat Tree などの全体全接続ネットワークは，コスト面・性能面において現実的ではない．この問題は現状のシステムでも確認できる．例えば，Blue Gene/L は 65536 個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，各プロセッサがメッセージをフォワーディングすることにより離れたプロセスとの通信を実現している [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，1:5 と上流のバンド幅が下流バンド幅に比べ低い構成になっている．

次世代インターコネクトとして，各ノードを安価な低バンド幅電気パケットネットワークと，高バンド幅光サーキットネットワークの双方に接続するネットワークが提案されている [3, 4]．これらネットワークではサイズの小さいメッセージは電気ネットワークで送信され，サイズの大きいメッセージは光ネットワークを用いて，ノード間で回線を確立した上で送信される．光回線確立には数ミリ秒時間を要するが，これら手法では通信パターンの解析，予測を行い，通信が起こる前にあらかじめ回線を確立する方法が提案されている．高価な高バンド幅パケットネットワークを使用せず，高価で高消費電力な OEO（Optical-Electrical-Optical）変換機が必要ない光サーキットネットワークを用いるため，低コストで実装できるメリットがある．一方，デメリットとして大容量メッセージを全対全で交換する場合に性能低下が見込まれる．しかし MPI 並列アプリケーションの多くは通信に局所性があり，各プロセスは総プロセス数に対し

てはるかに少ない数の相手としか通信をしない．さらに，集団通信の多くは小さいメッセージを交換することが報告されている [5]．そのため，通信の局所性を満たすようにプロセス配置，あるいは光回線確立の管理・スケジューリングを行なえば，性能を維持したままコストを削減することができる．しかし，これら既存研究では光ネットワーク部の規模が大きくなるため，構築コスト，消費電力，必要面積を考慮すると大規模環境の構築は現実的ではない．

本研究では，安価な上流低バンド幅電気パケットネットワークと光サーキットネットワークの双方を活用したノード間インターコネクトを提案する．提案ネットワークでは，計算ノードは電気ネットワークと光ネットワークにそれぞれ 1 つずつ NIC を持ち，光ネットワークは，電気ネットワークにおいてスイッチをまたぐ大容量通信が起こる場合にのみ，上流リンクでの混雑を避けるためにサプリメンタルに使用する．MPI 通信を行なう際には，あらかじめアプリケーションの通信パターンに合致するように光回線を確立し，光回線を優先利用するノード間のメッセージフォワーディングテーブルを作成し，それに従いメッセージをフォワードする．全対全通信をシミュレーション実行した結果，使用光回線数を増やすほど性能向上が確認できた．また，電気ネットワークの上流リンクのバンド幅を 2 倍に増強した場合と同程度の性能が得られた．

2 関連研究

光ネットワークを用いたグリッド・クラスタ環境での MPI アプリケーションの評価を行なっている既存研究があり，本章ではそれらで用いられているネットワークトポロジとその利用法について述べる．

Barker らは，各ノードが低バンド幅電気パケットネットワークに 1 つの NIC を，光サーキットネットワークに多数の NIC を持つネットワークを提案している [3]．この環境での一対一通信はまず電気ネットワークを用いて開始される．トラヒックを監視しつつ，通信データ量が増加した場合には，光回線を確立し光ネットワーク上での通信に切り替える．通信 2 ノードのいずれかでも全ての光 NIC を使いきっている場合には，LRU ルール等に従い古い回線を解放し，新規に回線を確立する．一方，集団通信は電気ネットワーク上でのみ行われる．本研究とは，各ノードに多数光 NIC を持たせる点と，集団通信を電気ネットワーク上だけで行なっている点が異なる．

Shalf らは低バンド幅電気パケットネットワークと光サーキットネットワークを組み合わせたハイブリッドネットワーク HFAST を提案している [5]．HFAST は電気ネットワークと計算ノードの間にコネクションプールとして光ネットワークを挿入する構成になっている．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 を用いれば光回線の割り当て問題となり，軽量な通信最適化が可能である．この研究では帯域遅延積以上のデータ転送にのみ HFAST ネットワークを用い，小容量データ転送や集団通信には別の低バンド幅電気ネットワークを用いる．2 つの分断された電気ネットワークを用いている点と，集団通信は電気ネットワーク上だけで行う点が本研究とは異なる．

光ネットワーク環境での MPI アプリケーション性能評価として，Kim らはサイト間を光ネットワークで接続した光グリッド環境と，シングルクラスタ環境での実行性能の比較評価を行なっている [6]．光ネットワークを用いると，IP 通信に比べ通信遅延のばらつきが小さくなるため，MPI_Barrier など同期を取る際に process skew が起こりにくくなると述べている．MPI アプリケーションは，光グリッドを用いると最大で倍性能が向上すると述べているが，光グリッド環境のノード数とシングルクラスタ環境のノード数が異なること，アプリケーションの通信量を一切考慮していないことより，フェアな評価とは言えず，更なる詳細な評価が必要である．また，実験に用いられた光グリッドも 2 サイトを接続したものと小規模な環境である．

井本らは計算ノードが直接光ネットワークに接続された環境での MPI ライブラリの実装を行なっている [7]．このライブラリは，リングトポロジ型光ネットワーク上に構成された共有メモリインターフェースを介して通信を行なう．しかし，リングネットワーク上の通信は単一トークンパッシング方式であり，1 度に 1 プロセスしか通信を行えないため，Ethernet を用いた場合より実行性能が落ちている．この性能低下は MPI_Alltoall のように，全体全でメッセージを交換し合う場合に特に顕著に現れると考えられる．光ネットワークしか利用しない点で本研究とは異なる．

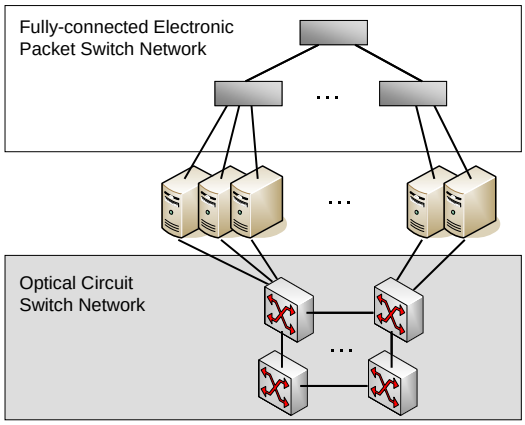


図 1: 提案ネットワーク

3 サプリメンタル光ネットワークの提案

3.1 サプリメンタル光ネットワーク

将来の高度に並列化された並列計算機のノード間インターコネクトとして，図 1 に示すネットワーク環境を提案する．計算ノードは電気パケットネットワークと光サーキットネットワークにそれぞれ 1 つずつ NIC を持つ．電気ネットワークは Tree ネットワークのような全対全接続である必要があるが，上流リンクのバンド幅は低くてよい．光サーキットネットワークには光バーストスイッチング技術を用い，各スイッチはメッシュ上に接続されており，数ミリ秒単位での回線切り替えが可能とする．ノードは光ネットワークへは 1 つしか NIC を持たないので，回線交換方式の性質上，同時に接続できる通信相手は 1 ノードに限られる．しかし，4 章で具体的に述べるように，光ネットワークは電気ネットワークのショートカット経路として使用するため，接続数が 1 でも通信可能である．また構築コストの都合から，回線確立ミス（呼損）が起こりうる波長数の少ない安価なネットワークを構築しても機能する．

提案ネットワークは安価な機材で構築できるので，数万ノードからなる大規模環境にも適応できる．また既存環境を，光ネットワークを追加するだけで，用意に拡張可能である．

3.2 他ネットワークとの比較

提案手法と 2 章で紹介した関連研究で用いられているネットワークとの比較を，全対全接続性，電気ネットワークのポート数，光ネットワークのポート数，光ネットワークにおける呼損の可否の 4 項目について行い，結果を表 1 にまとめた．表中の N はノード数を，n は電気パケットスイッチ数を，k はノードあるいはスイッチ 1 つが持つ光 NIC 数を表す．ここで，ネットワークのポート数はネットワークとノード，あるいは別のネットワークを接続可能なスイッチにおけるポート数の合計を意味する．また,「光バックエンド」は Kim らの光グリッド環境や，電気ネットワークのバックエンドに光ネットワークを使用する手法を一般化したネットワークを表し,「全光ネットワーク」は井本らのリング型光ネットワークのなどを一般化したネットワークを表す．ここでは回線交換方式の光ネットワークを用いたとして比較を行なった．

本提案や Barker ら，Shalf らの提案ネットワークでは，全対全接続の電気パケットネットワークを用いるため，全対全接続性がある．光バックエンドや全光ネットワークでは，電気スイッチやノードに搭載する光 NIC 数に応じて，全対全接続性は決まる．しかし，金銭的コストやノード構成により搭載できる NIC 数に制限があるため，全対全接続を実現するのは難しい．

本提案，Barker ら提案においては単一電気ネットワークを用いるので，電気ネットワークのポート数はノード数 N に等しい．光バックエンドネットワークではノードの接続以外にも，複数の電気ネットワーク間を接続する必要があるため，ポート数は $N + n$ となる．一方，Shalf ら提案のネットワークは，2 つの電気ネットワークを用いるため，そのポート数は最小で $2N$ になる．

提案ネットワークでは各ノードは 1 つの光 NIC を持つので，光ネットワークのポート数はノード数 N に等しくなる．一方，Barker ら提案では各ノードが複数光 NIC を持つこと許可しているため，ポート数は kN になる．Shalf ら提案の HFAST では，光ネットワークは N 個のノードを N 個の電気ポートに接続し，さらに電気スイッチ間をも接続する場合があるため，最小 $2N$ となる．全光ネットワークにおいては，搭載する NIC 数に応じてポート数が決まる．

最後に呼損について，提案ネットワークにおいては光回線が確立できない場合でも電気ネットワークで通信が行えるのでアプリケーション実行には問題無い．同様なことは Barker ら，Shalf ら提案にも当てはまる．しかし，光バックエンド，全光ネットワークにおいては主な通信を光ネットワーク上で行なうため，呼損の発生は通信不能，大幅な性能低下につ

表 1: 提案ネットワークと関連研究で提案されているネットワークの比較

ネットワーク	全対全接続	電気ネットワークポート数	光ネットワークポート数	呼損の可否
提案ネットワーク	あり	N	N	可
Barker ら提案	あり	N	kN	可
Shalf ら提案	あり	最小 $2N$	最小 $2N$	可
光バックエンド	構成依存	$N+n$	n	否
全光ネットワーク	構成依存	0	kN	否

ながる．特に光バックエンドの場合には，特定スイッチ以下の全ノードが通信できなくなる状況も起こりうる．

以上より，ネットワークの規模が小さく，光ネットワークの呼損の影響が少ない提案手法は，他手法に対し安価に構築可能である．

4 提案ネットワークにおける MPI プロセス通信

本章では提案ネットワーク上での MPI アプリケーションのプロセス間通信手法について述べる．1 計算ノード上で 1MPI プロセスを実行する設定でアルゴリズムを考案した．

4.1 MPI 通信の要件

提案ネットワークは特性の異なる 2 つのネットワークから構成されるため，それらのネットワークの特性に着目した MPI 通信手段が必要となる．電気パケットネットワークに関しては，上流リンクのバンド幅が低いことを想定しているので，複数の大容量データ通信を行なうと混雑発生が想定される．また，光サーキットネットワークでは，回線確立・解放に数ミリから数十ミリ秒ほどの時間を要すため，頻繁に確立・解放を行なうのでは，通信性能に影響される MPI アプリケーションでは大幅な性能低下につながる．また，小容量データ通信のために光回線を用いるとバンド幅の浪費，回線確立・解放コストの増大といった問題が生じる．

これらの課題より，電気パケットネットワークでの混雑を避け，バンド幅を有効活用し，光回線確立・解放コストを隠蔽できる通信手法が必要となる．

4.2 MPI 通信アルゴリズム

提案ネットワーク上で MPI アプリケーションを実行する際には，以下の条件にマッチする通信のみ光ネットワークを使用する．

- 一定サイズ以上のメッセージ交換を行なう場合
- 電気パケットネットワークでスイッチをまたいだ通信が起こる場合

1 つ目の条件は，高バンド幅の光ネットワークを有効活用するために，サイズの小さいメッセージは従来どおりの電気ネットワークで転送するという方針である．メッセージサイズの閾値は，光ネットワークの帯域遅延積以上のサイズとする．2 つ目の条件は，1 つ目の条件を満たす通信において，バンド幅の低い電気ネットワークの上流リンクは用いずに，通信ペア間で光回線を確立し，光ネットワーク上で通信を行なうという方針である．この 2 つの条件から，提案ネットワークにおける MPI 通信は以下に要約される．

- 小さいサイズの一対一通信，集団通信はパケット通信のみで行なう
- 大きいサイズの一対一通信，集団通信には光回線通信も用いる
- 電気パケットスイッチ内通信で済む場合には光ネットワークは使用しない

このように提案手法では，バンド幅の低い電気ネットワークの上流リンクのショートカットとして，サプリメンタルに光ネットワークを使用する．

理想としては，電気パケットスイッチをまたいだ全てのプロセス間通信は光ネットワークを用いて行ないたいが，光ネットワークの規模や他アプリケーションの光回線利用状況により，十分な光回線を用

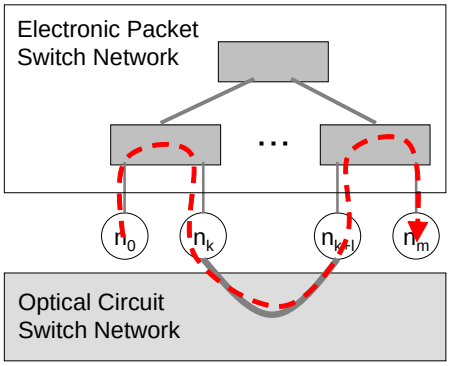


図 2: 提案ネットワークにおけるフォワーディング例

意できない場合もある．そこで，アプリケーションの通信パターンを基に，利用可能な光回線だけを用い，あらかじめプロセス間で通信トポロジを構築し，その上でのフォワーディングテーブルを作成し，各プロセスがメッセージをフォワードする手法を用いる．例えば，図 2 において，ノード n_k とノード n_{k+l} だけが光回線で接続されている場合に，ノード n_0 がノード n_m に大容量メッセージを送信する場合には，図中の点線で表される矢印に従いメッセージをフォワードする．光回線数が足りず，他のスイッチ下ノードと光回線で接続されない孤立スイッチができてしまった場合にのみ，電気スイッチの上流リンクで通信を行なう．このようにあらかじめ光パスを確立し，フォワーディングテーブルを作ることにより，アプリケーション実行中に時間のかかる光回線切り替えを行なわずに済む．しかし，アプリケーションの通信パターンを取得する必要があるので，実現するためには事前実行をするか，繰り返し処理を行なうアプリケーションである必要がある．

提案アルゴリズムは以下の 3 ステップの処理を順に実行し通信準備を行なう．以降 NAS Parallel Benchmarks[8] の MG，クラス C，ノード数 16 を図 3 に示す環境で実行した場合の例を交えて各ステップの詳細を述べる．

1. ネットワークトポロジとアプリケーション通信パターンの取得
2. プロセスのグルーピングと光回線の割り当て
3. フォワーディングテーブルの作成

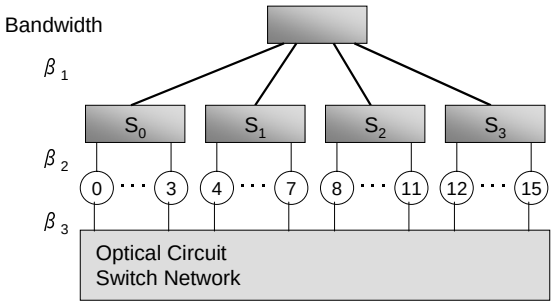


図 3: NPB MG:C:16 の実行環境

4.2.1 ネットワークトポロジとアプリケーション通信パターンの取得

ネットワークのトポロジ情報として取得する情報は電気ネットワークの上流・下流リンクの論理バンド幅と光ネットワークの論理バンド幅，電気パケットスイッチ下の MPI プロセスの配置情報である．論理バンド幅は設定ファイルから参照できるとする．MPI プロセスの配置情報は，例えば各プロセスが IP アドレスを MPI_Alltoall で交換し合い，IP アドレスのレンジから推測して取得する．図 3 の例では，電気上流バンド幅が β_1，下流バンド幅が β_2，光ネットワークのバンド幅が β_3 と取得でき，プロセス配置に関しては MPI ランク 0〜3 が同一スイッチ下，4〜7 が同一スイッチ下，という具合に取得できる．

アプリケーションの通信パターンとして取得する情報は，一対一通信の通信相手と送信データサイズとする．このとき，通信データサイズが光ネットワークの帯域遅延積以上の通信情報のみを取得する．アプリケーションによっては集団通信しか行わないものもあり，その場合には集団通信の通信パターンを取得する．NPB MG の場合，メッセージ閾値を 8KB に設定すると各プロセスはそれぞれ他の 4 プロセスとのみ通信することになる．例えば，ランク 0 のプロセスはランク $1, 2, 4, 12$ のプロセスと通信し，それぞれに対し合計 46MB ほどのメッセージを送受信することになる．

4.2.2 プロセスのグルーピングと光回線の割り当て

本手法では，電気ネットワークの上流リンクを使用しないように，電気ネットワークを用いた通信は可能な限り単一スイッチ内に収める．電気パケットスイッチをまたぐ通信には，電気上流リンクの迂回路として光回線を用いた通信を行なう．そのため，プロセスをスイッチ単位でグルーピングし，グループを

表 2: SP 方式における NPB MG:C:16 のスイッチ間通信数

Switch ID	S_0	S_1	S_2	S_3
S_0	−	16	0	16
S_1	16	−	16	0
S_2	0	16	−	16
S_3	16	0	16	−

超えた通信に光回線を割り当てる 2 手法を考案した．

1 つは，プロセスの電気スイッチ配置そのままにグルーピングを行なう Switch Partitioning（SP）方式である．先のステップで取得した通信パターンを基に，電気スイッチをまたぐ通信を行なうプロセス間に光回線を割り当てる．少ない光回線数でもスイッチ間を接続できるように，ラウンドロビン方式でスイッチ間に回線を割り当てる．このとき，同じスイッチ間通信を行なうプロセスペアに対しては，ランク値の小さいプロセスから順に回線を割り当てるルールとした．

NPB MG の場合，通信パターンから光回線数の増加につれプロセス間を図 4 に示すように接続していく．図 4(e) において全スイッチ間が光回線で全対全で接続される前に再度スイッチ S_0 とスイッチ S_1 下プロセスが接続されている．この理由は表 2 に示すように，各スイッチはそれぞれ他の 2 つのスイッチと接続すればアプリケーションの通信パターンを満たすため，十分な回線を確立し終えたので，次のラウンド処理が開始されたためである．

しかし，この方法では電気スイッチ配置に拘束されるので，通信パターンによっては離れたスイッチ下のプロセスとの通信が多くなり，光回線使用数が増える傾向がある．

そこで 2 つ目として，アプリケーションの通信パターンによりグルーピングを行なう Communication Partitioning（CP）方式を提案する．この方式では，通信量の多いプロセス同士をグルーピングし，同一電気パケットスイッチ下に再配置をする．通信量の少ないプロセス間リンクを切断することで，スイッチ数分のプロセスグループを作成し，切断されたリンクに光回線を割り当てる．光回線割り当てルールは SP 方式の場合と同じである．

NPB MG の場合，図 5 に示すようにプロセスのランク値が再配置され，光回線数の増加につれ各スイッチ下プロセスが接続される．図 3 のトポロジでは最大で 6 回線確立すれば通信パターンに合致したトポロジが構築される．

この方式を用いると，通信パターンに従った回線割り当てが可能なので，SP 方式より光回線数は少なく済むが，プロセスマイグレーションや MPI ランク値の再割り当てが必要となる．

4.2.3 フォワーディングテーブルの作成

電気上流リンクでの混雑を避けるため，高バンド幅の光回線に接続されたプロセスが他のプロセスのスイッチをまたぐ通信を中継する手法を取る．この手法では中継プロセスの電気ネットワーク側で混雑が起こりうるが，将来の高性能計算機にはエンドノードにも十分なバンド幅が供給されることと，電気上流リンクで混雑を起こすよりはシステム全体への影響は少ないと判断し，このように設計した．

メッセージ中継のためのフォワーディングテーブルは，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作成する．また，電気上流リンクより光回線を優先使用するルールとした．回線数が足りず，孤立してしまったスイッチ下プロセスへの通信には電気ネットワークを用いる．例えば図 4(b) において，プロセス 5 からプロセス 11 に通信する場合には，スイッチ S_1-スイッチ S_2 間に光回線が無いので電気上流リンクを用いて通信を行なう．

5 評価

提案ネットワーク上での全対全通信の性能をシミュレーションにより評価する．シミュレーションネットワーク環境を図 6 に示す．64 ノードからなる環境で，電気ネットワークの上流・下流リンクのバンド幅比率は 1:5 とした．電気パケットネットワークの詳細なパラメータは図中に示すとおりであり，これは東工大の TSUBAME で使用されている Voltaire 社製 InfiniBand スイッチ ISR9288 のスペックシートから取得した．光サーキットネットワークの遅延は電気ネットワークと等しいとし，1 回線のバンド幅はプロセスが実行されるノードのバス速度に律速されるとし，PCI Express x16 のバンド幅 32Gbps とした．

提案 2 手法（SP，CP）と電気ネットワークだけを用いた場合を比較する．SP 方式と，電気ネットワークだけの場合には，MPI ランク値は図中の左のプロセスから順に割り振る．CP 手法におけるプロセスグ

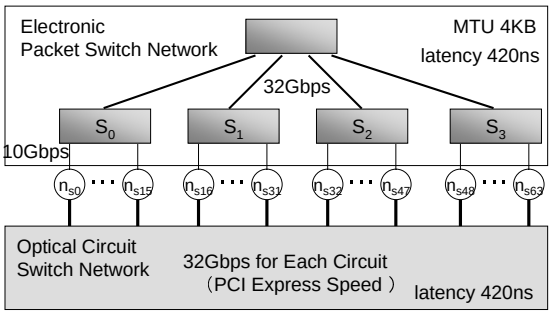


図 6: 想定環境

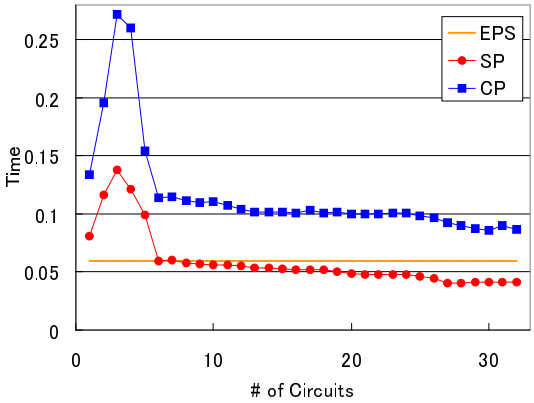


図 7: MPI_Alltoallv の性能

ループ分割にはグラフ分割ライブラリ Metis[9] を使用した．

全対全通信パターンとしては NAS Parallel Benchmarks の IS，クラス C，プロセス数 64 で実行される MPI_Alltoallv を用いた．今回は単一通信だけの評価であるが，IS では計算処理の 1 イテレーション内で通信は MPI_Allreduce と MPI_Alltoall，MPI_Alltoallv の 3 種類しか実行せず，MPI_Allreduce と MPI_Alltoall での通信量は少ないため，MPI_Alltoallv が支配的な通信となりアプリケーション性能を如実に表す．具体的に IS の MPI_Alltoallv では，各プロセスは全プロセスに対し 132KB ほどのメッセージを送信する．1 プロセスあたり，合計 8.5MB 送信する通信である．通信アルゴリズムは MPICH2[10] のアルゴリズムと同様に以下の式で計算されるランクを持つプロセスに対して順に送受信する．始めに全プロセスに対し送信し，その後全プロセスから受信するとした．

$$(Rank + i) \bmod 64 \quad (i = 1, 2, ...63) \qquad (1)$$

評価用にシミュレータを実装した．シミュレータでは通信時間のみを考慮し，遅延（α），バンド幅（β），メッセージサイズ（n）より $\alpha + n/\beta$ と計算し求めた．電気ネットワークに関しては，必要に応じてメッセージを MTU サイズに分割し複数回送信を行なった．また，電気スイッチは Store-and-Forward 方式とし，リンクごとに通信時間を求め，足し合わせた結果をプロセス間 End-to-End の通信時間とした．さらにパケットスイッチにおいて同一出力ポートを宛先とする通信の混雑をシミュレートするため，過去のメッセージの送信タイムスタンプを保存し，後続するメッセージのタイムスタンプ計算ではその値を加味して行なった．光ネットワークでは，回線を確立すればスイッチにおける待ちは発生しないため単一リンクと仮定し通信時間を計算した．電気ネットワークからの MPI メッセージを光ネットワークにフォワードする場合には，MTU サイズで分割されたメッセージを全て受信した後に，光ネットワークに送信するとした．

シミュレーション結果を図 7 に示す．図中凡例の EPS は電気パケットスイッチネットワークだけの結果を表す．X 軸に光回線数を取り，Y 軸に全対全通信実行時間の全プロセスでの平均値を取った．全体的に CP 方式の性能が悪い．CP 方式では通信パターンによりプロセスを再配置しなおすのだが，Alltoallv 通信は全対全通信であることと，プロセス間通信量に大きな違いがないことより再配置によるメリットが少なく，また，式（1）による通信順では離れたノードとの通信・中継回数が増え性能が低下する．SP の場合においても回線数が少ない場合には EPS に比べ性能が低下している．回線数が少ないことによる中継ノードへの混雑と，遠回りの経路の利用による．一方，6 回線以上では EPS の性能を上回る．これは，6 回線を超えると 4 機の電気パケットスイッチがノードを介して光ネットワーク側に全対全で接続されるためである．回線数が増えるほど，スイッチ間の経路が増え，中継ノードの電気ネットワーク側の混雑が減る．

最後に，電気ネットワークの上流リンクのバンド幅を増強した場合との比較を行なった．比較には SP 手法を用い，SP 手法の場合の電気上流リンクバンド幅は先と同じ 32Gbps とした．結果を図 8 に示す．SP 手法は 27 回線以上使用して電気上流リンクバンド幅 64Gbps と同程度の性能に達している．しかし，提案手法では電気の上流リンクは用いないので，他アプ

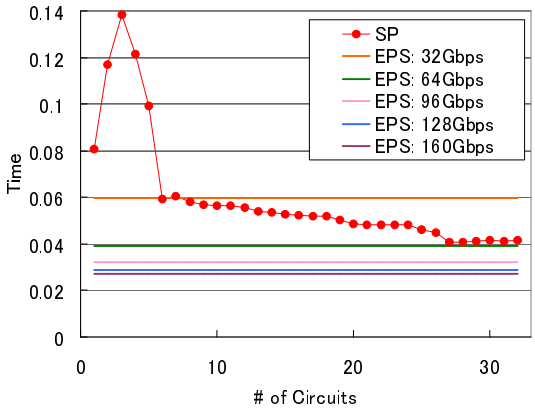


図 8: 電気ネットワークの上流リンクバンド幅を増強した場合と SP 手法の比較

リケーションやストレージなど，他の通信も考慮すると異なる結果になると考えられる．今後，他の通信も考慮した評価を行なう．

6 おわりに

計算機間インターコネクトとして，電気パケットネットワークと光サーキットネットワークの双方を活用するネットワークを提案した．光ネットワークは電気ネットワークの上流リンクのショートカットとして使用するため，電気上流リンクのバンド幅を増強する必要がなく，ネットワークの規模も小さくて済むので，安価に構築できる．提案ネットワーク上での MPI アプリケーションの通信の局所性を利用した通信アルゴリズムを提案した．アプリケーションの通信パターンと電気パケットスイッチ下のノード配置によりプロセスをグルーピングし，グループ間をまたぐ通信に光回線を割り当て，プロセス間でメッセージをフォワードする方式である．グルーピング手法として電気スイッチによる配置を利用した Switch Partitioning 方式，アプリケーションの通信パターンにより分割する Communication Partitioning 方式を提案した．

提案ネットワーク上でこのアルゴリズムを用いて全対全通信を行なった結果，波長数の増加に伴い電気ネットワークだけを用いた場合より性能向上が確認できた．また，電気ネットワークの上流リンクにコストをかけバンド幅を 2 倍にした場合と同程度の性能が得られた．提案手法では電気ネットワークの上流リンクは使用しないため，他の通信を考慮すると更なる性能向上が見込める．

今後の課題として以下の項目を考えている．

- アプリケーション全体のシミュレーションによる評価
- 複数アプリケーション存在下でのシステムとしての評価
- 1 ノード複数プロセスの場合の通信アルゴリズムの考案
- 大規模環境を想定した性能評価
- Barker らや，Shalf らの提案するネットワークやその他光ネットワークを利用した環境上でのアプリケーション実行性能との比較

謝辞

本研究の一部は科学研究費補助金特定領域研究 (18049028) 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, p. 41 (2004).

[2] 松岡 聡: TSUBAME の飛翔：ペタスケールへ向けた「みんなのスパコン」の構想, 情報処理学会研究報告 2006-HPC-107 (pp37-42 July 31) (2006).

[3] 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, *SC '05: Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, Washington, DC, USA, IEEE Computer Society, p. 16 (2005).

[4] 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).

[5] Shalf, J., Kamil, S., Oliker, L. and Skinner, D.: Analyzing Ultra-Scale Application Communication Requirements for a Reconfigurable Hybrid Interconnect, *Proceedings of the ACM/IEEE SC 2005 Conference* (2005).

[6] Kim, D., Jin, H.-W., Jeong, K., Lee, J. and Noh, M.: Performance Measurement and Analysis of High-Performance Parallel Applications over

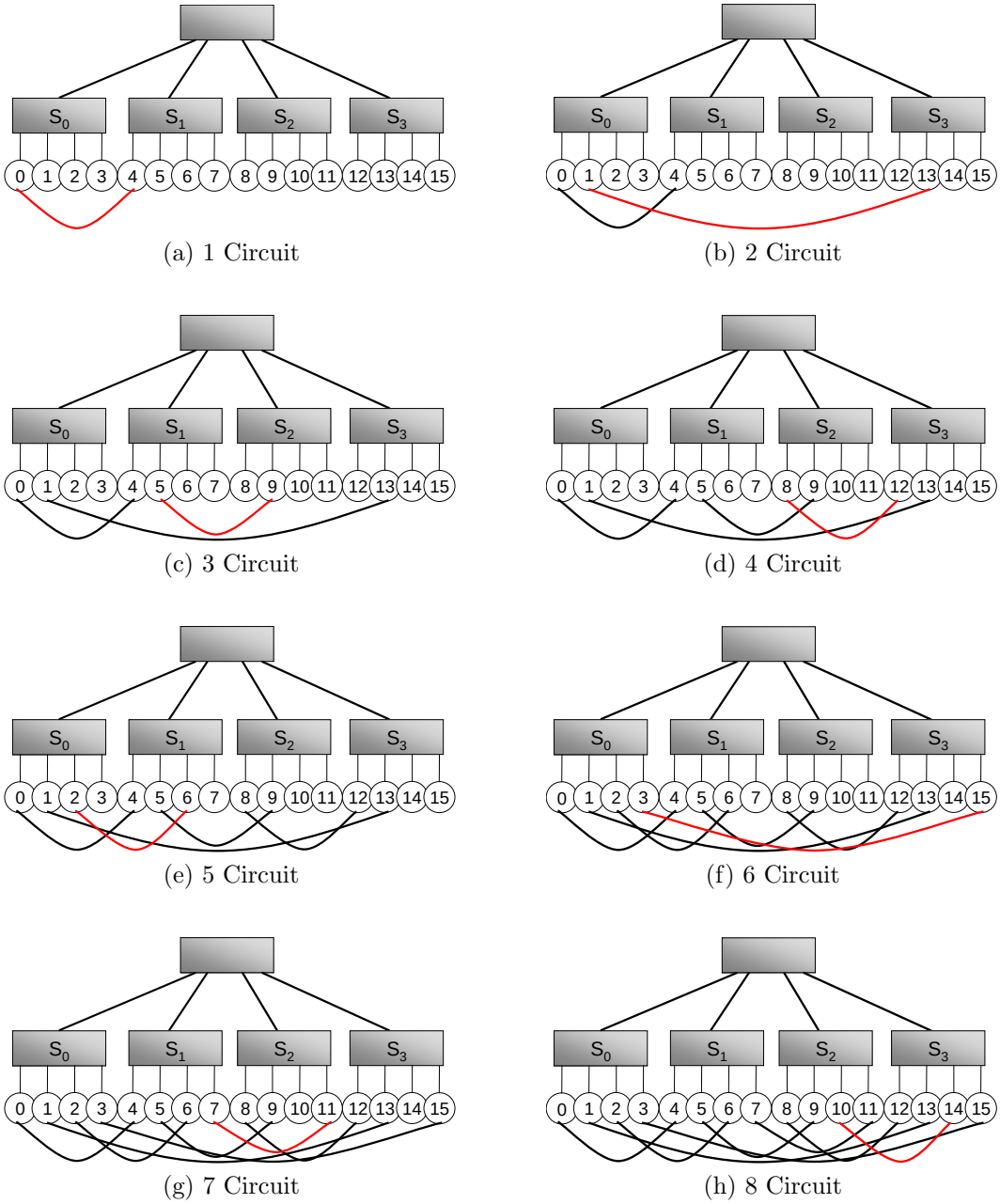
Lambda Grid, *The 9th International Conference on Advanced Communication Technology*, Vol. 1, pp. 792–796 (2007).

[7] Imoto, M., Taniguchi, E., Baba, K. and Murata, M.: Implementation and evaluation of MPI library with Globus toolkit for establishing λ computing environment, *Proceedings of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies*, pp. 421–426 (2005).

[8] der Wijngaart, R. F. V.: NAS Parallel Benchmarks Version 2.4, Technical Report NAS Technical Report NAS-02-007, NASA Ames Research Center (2002).

[9] : METIS - Family of Multilevel Partitioning Algorithms, `http://glaros.dtc.umn.edu/gkhome/views/metis/`.

[10] : MPICH2 home page, `http://www-unix.mcs.anl.gov/mpi/mpich2/`.



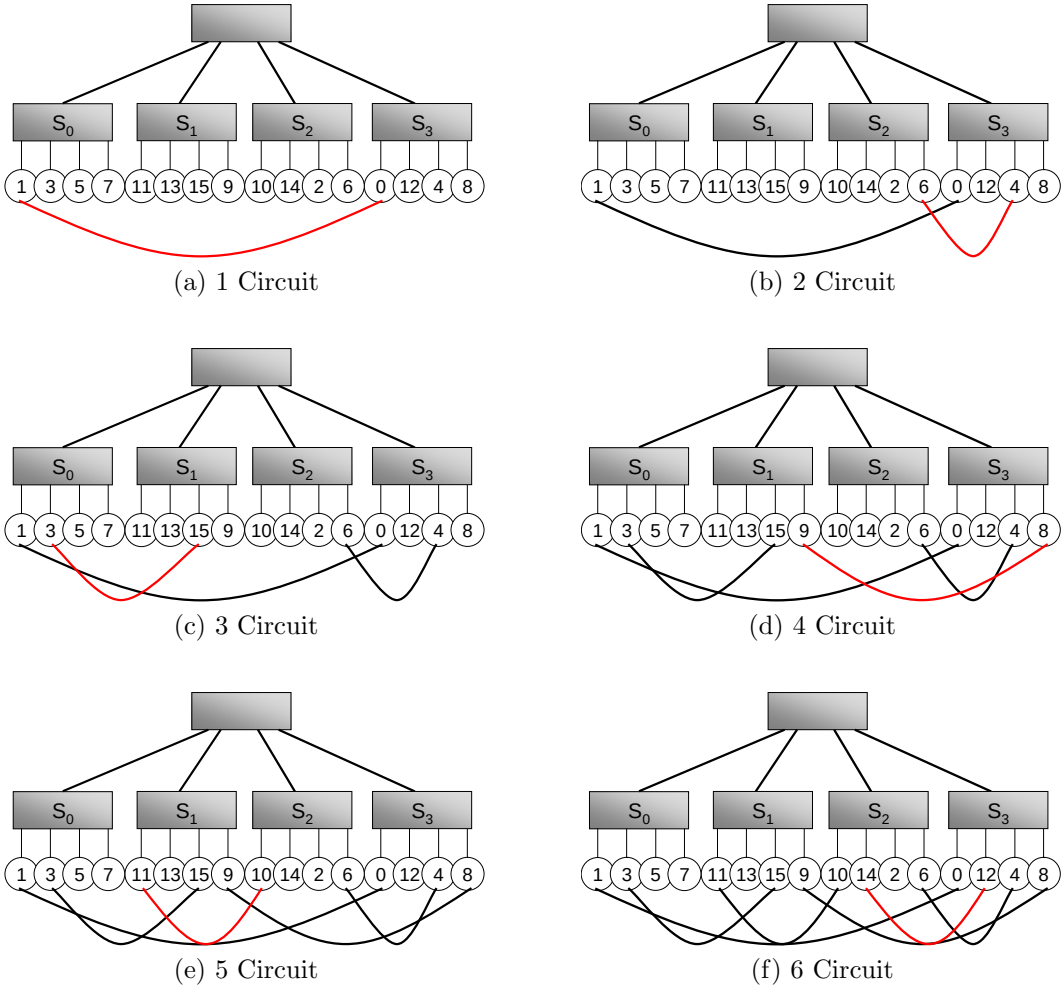
(a) 1 Circuit (b) 2 Circuit

(c) 3 Circuit (d) 4 Circuit

(e) 5 Circuit (f) 6 Circuit

(g) 7 Circuit (h) 8 Circuit

図 4: NPB MG:C:16 での Switch Partitioning 方式による光回線割り当て



(a) 1 Circuit (b) 2 Circuit

(c) 3 Circuit (d) 4 Circuit

(e) 5 Circuit (f) 6 Circuit

図 5: NPB MG:C:16 での Communication Partitioning 方式による光回線割り当て