情報爆発時代の光インターコネクト上でのMPI通信アルゴリズム

MPI Communication Algorithm over an Optical Interconnect for the Information Explosion Era

滝澤 真一朗[†] 遠藤 敏夫[†] 松岡 聡[†,††]

Shin'ichiro Takizawa Toshio Endo Satoshi Matsuoka

[†] 東京工業大学 [††] 国立情報学研究所

Tokyo Institute of Technology National Institute of Informatics

1 背景

情報爆発時代の大容量データを処理するためのHPC(High Performance Computing)システムには、数千〜数万のプロセッサを高速に相互接続するインターコネクトネットワークが求められる。このネットワーク設計にあたり、従来より広く使われてきたInfiniband等電気パケットネットワークには高消費電力、ネットワーク上流での混雑の問題があり、近年では低消費電力、高バンド幅の光サーキットネットワークの利用が注目されている。サーキット通信では通信相手が限定されるため、既存研究[1, 2, 3]では、光サーキットスイッチング(OCS)ネットワークと電気パケットスイッチング(EPS)ネットワークのハイブリッドインターコネクトを提案している。しかしこれら研究では、ネットワークの規模が大きい、特殊なハードウェアを必要とする、といった構築コストの問題がある。我々はコモディティパーツを用いて安価に構築可能なハイブリッドネットワークを提案し、さらにその上でのMPI通信手法を提案する。

2 OCS/EPSハイブリッドネットワークとMPI通信手法

提案するハイブリッドネットワークの構成例を図1に示す。低上流リンクバンド幅かつ全対全接続のEPSネットワークと、ノンブロッキングで再割り当て可能なEPSネットワークからなり、各計算ノードはそれぞれのネットワークに対し1つづつリンクを持つ。ノードに搭載するNICとEPSネットワークはコモディティパーツより構成でき、OCSネットワークは $Ninput \times Noutput$ の光サーキットスイッチの組み合わせで容易に構築できるため、既存ネットワークの拡張としても構成できる。

提案ネットワークには、1)EPSネットワークの上流リンクバンド幅が低い、2)各ノードはOCSネットワークへのリンクを1つしか持たなく、同時に複数のノードと通信ができない、2つの制約がある。そのため、OCSネットワークはEPSネットワークの上流リンクでの混雑を避けるためのショートカットとして、EPSネットワークにおいてスイッチ間で通信がおこる場合にのみ使用する。この要件をみたすMPIアプリケーション通信手法を、以下のステップから構成した。

1. **トポロジと通信パターンの取得**： アプリケーションの第1イテレーションの間に（あるいは事前実行して）、EPSスイッチ下のプロセスの配置と、各プロセスでのMPI通信相手とデータ量を取得する。

2. **プロセスのグルーピング**： プロセス配置と通信パターンに応じてプロセスを分割する2手法を提案する。1つはスケジューラによる初期配置をそのまま用いた **Switch Partitioning (SP)** である。スイッチを単位として、同一スイッチ下のプロセスを1グループとする方法である。もう1つは通信パターンにより通信頻度の高いプロセス同士を1EPSスイッチ下に移動しグループとする **Communication Partitioning (CP)** である。プロセスのグループ分割には通信パターンを入力としたedge cutアルゴリズムを用いる。CP手法を実装するにはプロセスマイグレーションやプロセスIDの再割り当てと言った手法が必要となる。

3. **サーキットの割り当て**： アプリケーションの通信パターンを満たすようにグループ間通信用サーキットを確立する。このとき、IDの小さいプロセスから順に、少ない回線数でグループ間を接続するグループ間ラウンドロビン方式で確立していく。

4. **メッセージフォワード通信**： 生成したOCS/EPSハイブリッドネットワーク上でのノード間MPI通信用フォワーディングテーブルを作成し、それに従い以降の通信を行う。フォワーディングテーブル作

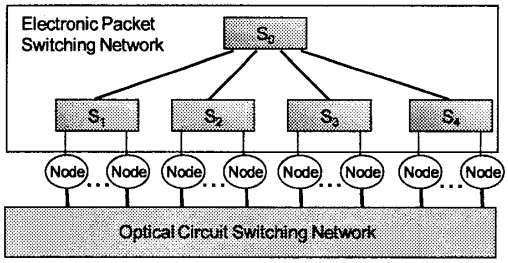

図 1: 光・電気ハイブリッドネットワーク

成には、理論バンド幅を基準とした距離ベクトルアルゴリズムを用いる。

3 評価

提案ネットワークと通信アルゴリズムの組み合わせと、EPS ネットワークのみ (以下 EPS only) を用いた on demand 通信の場合との比較評価をシミュレーションにより行った。図 1 のトポロジに表 1 で示すパラメータを設定した環境を用いた。使用したアプリケーションは、プロセスを 8×16 の格子上に行方向に配置し、各プロセスは隣接 4 プロセスと 40MB のメッセージを交換し合う通信を 5 回行うものである。SP 手法と EPS only ではプロセス ID は図中の左から順に割り当て、CP 手法ではあらかじめ通信パターンを取得し、グルーピング結果に従い割り当てた。

結果を図 2 に示す。横軸は使用した光サーキット数を表し、縦軸は上流リンクバンド幅が 64Gbps の EPS only に対する相対実行時間を表す。凡例中の $U \times X$ は上流リンクバンド幅を X 倍した場合を表し、$U \times 5$ の場合には EPS ネットワークはフルバイセクションバンド幅を提供する。光サーキット数が少ない場合は提案手法は実行時間が増加していることが確認できる。これは、メッセージが複数の中継プロセスを経由して宛先プロセスへ到着することによる遅延に起因する。しかしながら、サーキット数の増加につれホップ数が減るため、実行時間は減少していき、フルバイセクションバンド幅 EPS only ネットワーク以上の性能を示す。さらに EPS only の場合には上流リンクバンド幅が低い場合には大きく実行時間が増加しているが、提案手法ではほとんど変化がない。このことより、提案手法はストレージや他アプリケーションの通信による影響が小さいとわかる。

また、我々は NAS Parallel Benchmarks による評価も行っており、そこでも通信の局所性が高いアプリケーションでは同様の傾向が確認できた。しかし、局所性の低いアプリケーションでは逆に実行時間の増加が確認され、さらなるアルゴリズムの改良を考えている。

表 1: シミュレーションパラメータ

Parameter	Value
Node count	128
Node count under each packet switch	32
CPU speed of node	1.6GHz
Link speed in the OCS	32Gbps
Link speed of upstream links in the EPS	64Gbps
Link speed of downstream links in the EPS	10Gbps
Propagation delay in the OCS	80ns
One link propagation delay in the EPS	20ns
Switching delay in packet switches	420ns
MTU in the EPS	4096B

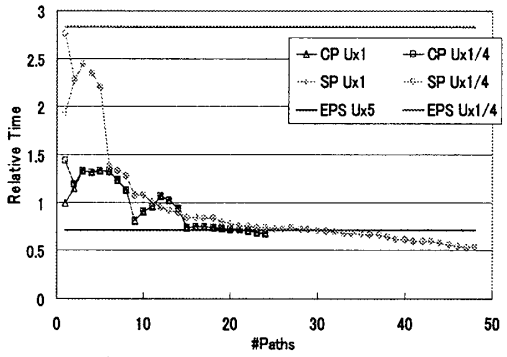

図 2: 相対実行時間

4 まとめ

情報爆発時代の HPC システム用 OCS/EPS ハイブリッドネットワークとその上での MPI 通信手法を提案し、有効性を示した。今後の課題として、より大規模環境を想定した評価、ノードのネットワーク性能をヘテロにした場合の評価を考えている。

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] Kevin J. Barker et al. On the Feasibility of Optical Circuit Switching for High Performance Computing Systems. In *SC '05: Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, page 16, Washington, DC, USA, November 2005. IEEE Computer Society.

[2] Shoaib Kamil et al. Reconfigurable Hybrid Interconnection for Static and Dynamic Scientific Applications. In *ACM International Conference on Computing Frontiers*, 2007.

[3] Avinash Kodi et al. Performance Adaptive Power-Aware Reconfigurable Optical Interconnects for High-Performance Computing (HPC) Systems. In *International Conference for High-Performance Computing, Networking, Storage and Analysis (SC'07)*, November 2007.