平成 23 年度 RICC 利用報告書

課題名（タイトル）：

大規模並列アプリケーションの通信性能解析および高速化に関する研究

利用者氏名：南里 豪志
理研での所属研究室名：本所 情報基盤センター

報告内容

１． 本課題の研究の背景、目的、関係するプロジェクトとの関係

計算機が大規模化し、多数の計算ノードで構成されるようになると、プログラムの実行時間におけるノード間通信の占める割合が増加し、高速化の妨げとなる。特に、並列プログラム中のプロセス群が参加して行う、Broadcast や Reduction、Alltoall 等の集団通信は、計算機の規模の増加とともに所要時間が増える。このような集団通信は、多くの科学技術計算で多用されるため、様々な実装アルゴリズムが提案されている。しかしながら、アルゴリズムの優劣はメッセージサイズやプロセス数、さらに計算機のアーキテクチャなどで変動し、常に最速であるものは存在しないため、状況に応じた適切なアルゴリズム選択が重要となっている。

そこで本研究では、計算機の状況に応じて最適な集団通信アルゴリズムを選択する動的アルゴリズム選択技術を開発する。従来、集団通信のアルゴリズム選択は、予め決めておいたメッセージサイズとプロセス数の閾値に従って使用するアルゴリズムを決定する、静的な手法が用いられていた。これは、同じプロセス数とメッセージサイズに対して各アルゴリズムの所要時間が一定であることを前提としている。しかし、プロセスの計算機内での配置やアプリケーションの負荷バランス等によって通信の性能が変動する場合、同じプロセス数やメッセージサイズでも集団通信の所要時間が変わる。例えば、計算機の大規模化に伴って、ノード間のインターコネクトネットワークとして用いられるようになった Fat Tree や多次元トーラスは、ノードの位置関係で通信性能が変化する。特に、複数の独立した通信が同じ経路を同時に使用する通信衝突は、通信速度を大幅に低下させるため、集団通信への影響も大きい。

この問題を解決するため、実行時の状況に応じた動的なアルゴリズム選択技術が提案されている。Faraj らは、同じパラメータで何度も呼び出される集団通信について、最初の数十回を使って用意されている各アルゴリズムを試し、その結果判明した最速のアルゴリズムをそれ以降の呼び出しで利用する Star-MPI ライブラリを提案した[1]。しかし、これは非常に遅いアルゴリズムを試すことによる性能低下が問題である。一方 Nishtala は、集団通信アルゴリズムの性能予測モデルを作成し、それにもとづいて対象アルゴリズムを絞り込んだうえで、最適なアルゴリズムを探索する技術を、通信ライブラリ GASNet 上に実装した[2]。しかし、この性能予測にはプロセスの配置による影響が考慮されておらず、精度に問題がある。また、各アルゴリズムの計測は最初の集団通信呼び出し時にまとめて行うため、オーバヘッドが大きい。

本研究で提案するアルゴリズム選択技術は、プロセスの配置とトポロジの情報に基づいたアルゴリズムの性能予測と、実行時の各アルゴリズムの計測を組み合わせたものである。アルゴリズムの性能予測では、実行開始時に取得したプロセス配置の情報と、予め取得していたトポロジの情報から、集団通信実行時に律速となるリンクの有効バンド幅を算出し、それに基づいて各アルゴリズムの所要時間を見積もる。その結果、他のアルゴリズムよりも大幅に遅いと予測されるアルゴリズムを選択肢から除外した上で、Star-MPI と同様に 1 回の呼び出し時に１つずつアルゴリズムを試して最適なアルゴリズムを選択する。

本年度は、提案手法による集団通信の例として、RICC を対象として動的アルゴリズム選択を行う Alltoall 通信関数を実装した。

２．　具体的な利用内容、計算方法

本研究で提案する動的アルゴリズム選択手法では、各集団通信アルゴリズムについて、トポロジとランク配置に応じた性能予測を行う。この、ランク配置を考慮する重要性を検証するため、予備調査として、ランク配置による集団通信アルゴリズムの性能への影響を計測した。

実験で用いるランク配置のパターンは、RICC のインターコネクトネットワークのトポロジを考慮して決定した。RICC のインターコネクトネットワークは 2 階層のスイッチ群による Fat Tree トポロジで構成されており、下位の 52 台の Leaf Switch が上位の 2 台の Upper Switch と 2 本ずつ、合計 4 本のリンクで接続されている。この Leaf Switch に 20 台ずつの計算ノードが接続され、合計で 1024 ノードの PC クラスタを構成している。各ノードには一意に番号が付けられており、0 番目の Leaf Switch には 0～19、1 番目の Leaf Switch には 20～39 というように、Leaf Switch 内に連続した番号のノードが配置されている。一方、各 Leaf Switch から Upper Switch への 4 本のリンクには、0～3 の番号が付けられている。このトポロジにおける Leaf Switch を跨る通信では、Leaf Switch から Upper Switch へのリンクのうち、その通信の宛先ノードの番号を 4 で割った余りの数字と一致する番号のリンクを経由して、目的のノードにメッセージが送信される。

このようなトポロジでは、プロセスが配置されるノードの位置によって各リンクの使用頻度に偏りが生じる。そこで、この偏りによる影響を検証するため、ランク配置として以下の 5 パターンを用い、プロセス数を 64、128、192、256 と変化させ、各 Alltoall アルゴリズムの所要時間を計測した。

０）4 の倍数のノード番号のみにプロセスを配置

Leaf Switch と Upper Switch の間のリンクを、Leaf Switch あたり 1 本のみ使用する。各 Leaf Switch 内のプロセス数は、192 プロセスまでは 4、256 プロセスでは Leaf Switch 数が不足するので 5 とし、ノード番号の小さいものから順に 4 の倍数の番号を持つノードに配置する。

１）2 の倍数のノード番号のみにプロセスを配置

Leaf Switch と Upper Switch の間のリンクを、Leaf Switch あたり 2 本使用する。各 Leaf Switch 内のプロセス数は、192 プロセスまでは 4、256 プロセスでは 5 とし、ノード番号の小さいものから順に 2 の倍数の番号を持つノードに配置する。

２）各 Leaf Switch の最初の 4 ノードにプロセスを配置

Leaf Switch と Upper Switch の間のリンクを、Leaf Switch あたり 4 本使用する。256 プロセスでは、各 Leaf Switch の最初の 5 ノードにプロセスを配置する。

３）各 Leaf Switch の最初の 8 ノードにプロセスを配置

Leaf Switch と Upper Switch の間のリンクを、Leaf Switch あたり 4 本使用する。パターン２）に対して使用 Leaf Switch 数が半分となり、リンクあたりプロセス数が 2 倍となる。

４）各 Leaf Switch の最初の 16 ノードにプロセスを配置

Leaf Switch と Upper Switch の間のリンクを、Leaf Switch あたり 4 本使用する。パターン２）に対して使用 Leaf Switch 数が 1/4 となり、リンクあたりプロセス数が 4 倍となる。

メッセージサイズを 16KB とし、各プロセス数で各アルゴリズムの所要時間を計測した結果を、図 1～図 4 に示す。各図で横軸はプロセス配置パターンの番号、縦軸は所要時間である。また、それぞれの線が、各アルゴリズムの所要時間である。

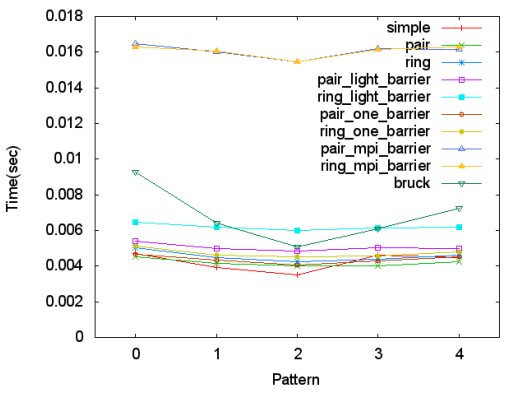

図 1 プロセス配置パターン毎の各アルゴリズムの所要時間

（64 プロセス、16KB）

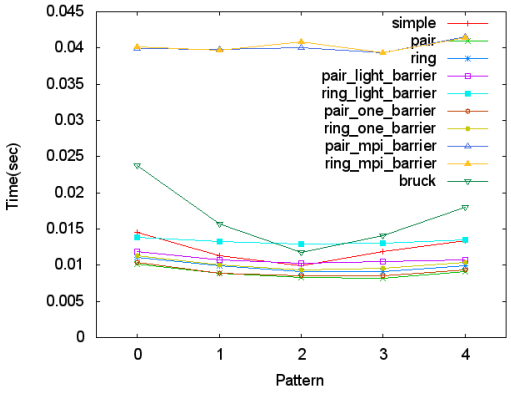

図 2　プロセス配置パターン毎の各アルゴリズムの所要時間（128 プロセス、16KB）

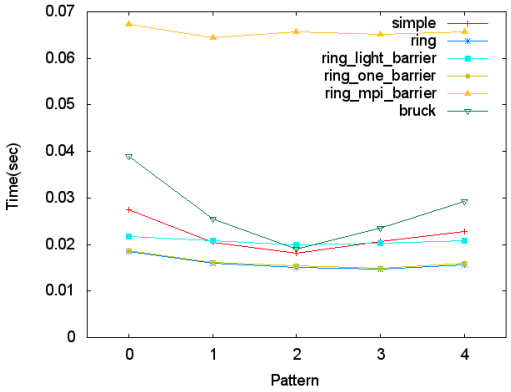

図 3　プロセス配置パターン毎の各アルゴリズムの所要時間（192 プロセス、16KB）

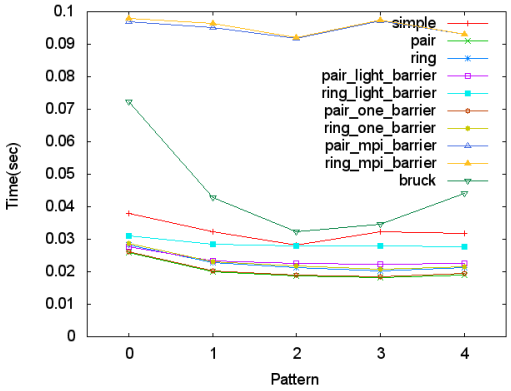

図 4　プロセス配置パターン毎の各アルゴリズムの所要時間（256 プロセス、16KB）

図より、16KB では bruck のパターン毎の性能の変動が大きいことが分かる。パターン 0 やパターン 4 では最速のアルゴリズムに対して 2 倍以上の時間を要しているのに対し、パターン 2 では最速のアルゴリズムに近い時間となっている。これは、プロセス数 P に対して、他のアルゴリズムが 1 回あたり 16KB の一対一通信を P-1 回繰り返して Alltoall 通信を実現しているのに対し bruck は 1 回あたり 16KB*P/2 の一対一通信を $\log_2 P$ 回繰り返しているため、プロセス配置パターンによるバンド幅の変化に影響されたためである。なお、リンクあたりのプロセス数が同じであるパターン 0 と 4 やパターン 1 と 3 で所要時間が異なるのは、パターン 3 と 4 は、パターン 0 と 1 に比べ、同じ Leaf Switch 内に配置されるプロセス数が多く、Leaf Switch をまたぐ通信の数が少ないためである。一方、メッセージサイズが 1MB の場合の 64 プロセスでの各アルゴリズムの所要時間を図 5 に示す。このように、メッセージサイズが大きくなると、他のアルゴリズムでもプロセス配置パターンによって性能の変動が見られるようになる。

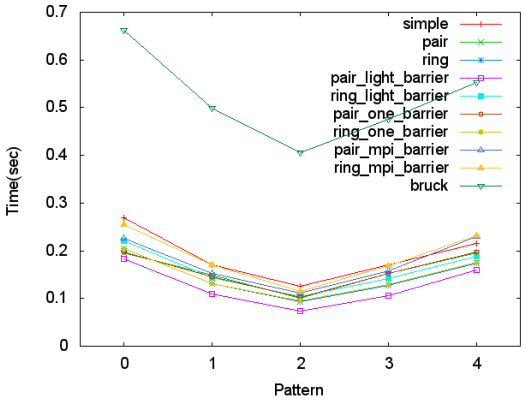

図 5　プロセス配置パターン毎の各アルゴリズムの所要時間（64 プロセス、1MB）

この調査により、プロセスが割り当てられたノードの位置によって有効なバンド幅が変動し、その影響でアルゴリズムの優劣が変化することが確認できた。この結果を受け、プロセスが割り当てられたノードの位置とトポロジを考慮したアルゴリズム選択技術を提案する。

本手法では、まずノードの配置とトポロジの情報に基づいて候補アルゴリズムを絞り込み、その後、各候補アルゴリズムを実際に試して最速のものを選択する。アルゴリズムの絞り込みには、各アルゴリズムの性能モデルを用いる。

本稿では、提案手法の実装例として、RICC のトポロジを対象とした動的アルゴリズム選択による Alltoall 通信関数を実装する。この関数は、実行時のランク配置とトポロジ情報にもとづいて各 Alltoall アルゴリズムの所要時間を予測し、候補アルゴリズムを絞り込む。

各アルゴリズムの性能予測には、性能モデルを用いる。基本的に Alltoall 通信は、参加する全プロセス間で全対全通信を行う。そこで今回作成する Alltoall 通信関数では、全プロセスが同時に通信を行う際の平均帯域幅を計算し、その値を用いて各アルゴリズムの性能を予測する。この平均帯域幅は、あるプロセスと他の全プロセスの間で通信を行う際の個々の通信の帯域幅の平均値とする。ここで個々の通信の帯域幅としては、それぞれの通信時に経由する各リンクの基本帯域幅を、そのリンクを同時に使用する他のプロセスによる通信の数の期待値で割った値を用いる。今回対象とする RICC のトポロジでは、Leaf Switch と Upper Switch の間のリンクにおいて、同時に使用するプロセス数が方向によって異なる。Leaf Switch から Upper Switch への方向は、その Leaf Switch から他の Leaf Switch への通信のうち、宛先ノード番号を 4 で割った値がそのリンク番号と一致するものが使用する。一方 Upper Switch から Leaf Switch への方向は、他の Leaf Switch からこの Leaf Switch への通信のうち、宛先ノード番号を 4 で割った値がそのリンク番号と一致するものが使用する。このうち、今回は、Upper Switch から Leaf Switch への方向の平均帯域幅を用いる。この平均帯域幅は Leaf Switch と Upper Switch の間のリンク毎に変動するため、全てのリンクについて計算した後、全リンクで最も小さい値を、システム全体の通信を律速する平均帯域幅として用いる。

次に、各リンクの平均帯域幅の算出方法を示す。ここで、ノード数を Nnode、ノード内プロセス数を Pn、i 番目の Leaf Switch を LSi、i 番目の Leaf Switch 内と Upper Switch の間のリンクのうち j 番目のものを ULij、プログラムが使用するノードのうち LSi に配置されたものの数を Ni、そのうちノード番号を 4 で割った値が j であるものを NLij とする。また、今回はノード内、Leaf Switch 内、Leaf Switch 間、それぞれ基準となる帯域幅は一定値 B として、平均帯域幅を計算する。

まず、LSi のノードの一つに割り当てられたプロセスが他の全プロセスから受信する場合の各通信の帯域幅を見積もる。ノード内のプロセスから受信する場合の帯域幅は、他の通信に妨げられないと仮定し、B のままとする。一方、Leaf Switch 内のノード間通信は、Leaf Switch からノードへの 1 本のリンクをノード内の全プロセスの受信で共有するため、平均帯域幅は B/Pn とする。また、Leaf Switch を跨ぐ通信の受信では、同じ Leaf Switch 内の各プロセスの通信のうち、同じリンクを経由して他の Leaf Switch から受信するものが 1 本のリンクを共有する。このプロセスが配置されたノードの番号を 4 で割った余りが j である場合 ULij を使って受信するので、同時に同じリンクを使って受信する LSi 内のプロセス数の期待値 RCVij は以下で計算できる。

$$RCVij = 1 + (NULij * Pn - 1) * (Nnode - Ni) * Pn / (Nnode * maxn - 1)$$

これらより、ULij を経由して受信する LSi 内のプロセスについて平均帯域幅 BRij を、以下で計算する。

$$BSij = B/((Pn - 1 + (Ni-1) * Pn * Pn + (Nnode - Ni) * RCVij) / (Nnode * Pn - 1))$$

これを、0≦i<使用スイッチ数、0≦j<4、の各 i, j で計算し、その最小値をシステムの平均帯域幅とする。

各アルゴリズムの性能は、上記で得られた平均帯域幅をそれぞれの性能モデルに適用して予測する。今回の実装に用いた Alltoall の各アルゴリズムの性能モデルを以下に示す。

Algorithm	Model
Simple Spread	(P-1)*L+(P-1)*M*B
Ring	(P-1)*(L+M*B)
Ring with One Barrier	(P-1)*(L+M*B)+(L*$\log_2$P)
Ring with MPI_Barrier	(P-1)*(L+M*B)+2*(L*(P-1)*$\log_2$P)
Ring with Light Barrier	(P-1)*(L+M*B)+L*(P-1)
Pair	(P-1)*(L+M*B)
Pair with One Barrier	(P-1)*(L+M*B)+(L*$\log_2$P)
Pair with MPI_Barrier	(P-1)*(L+M*B)+2*(L*(P-1)*$\log_2$P)
Pair with Light Barrier	(P-1)*(L+M*B)+L*(P-1)
Bruck	L*$\log_2$P+P*M*B*$\log_2$P/2

これらのモデルは、Hockney モデルによる一対一通信の性能モデルをもとにしている。Hockney モデルでは、個々の一対一通信の所要時間を、遅延時間 L とバイト当たりの所要時間 B、すなわち帯域幅の逆数を用いて L + M*B と表す。これを、各アルゴリズム内の一対一通信に適用して、性能モデルを作成した。

このモデル自体は、通信の衝突による影響が考慮されていない。しかし、前述の通り帯域幅をプロセスの配置に応じて調整することにより、衝突の影響を加味した所要時間を見積もることができる。なお、一対一通信の性能モデルとしては、他に LogGP や PLogP{Parameterized Log-P)などが提案されている。特に PLogP は、メッセージサイズの変化に伴う実効帯域幅の変動を表現でき、より高い精度での性能予測が行えると期待できるため、今後適用を検討する。

今回作成した Alltoall 通信関数の実装では、まず事前に基準となる帯域幅と遅延時間を計測した。これは、2 プロセス間の一対一通信を繰り返して計測した結果を用いた。

また、トポロジ情報は事前に調査し、その結果を性能予測手法に反映させた。具体的には、Upper Switch と Leaf Switch の間のリンクの構成、および Leaf Switch にまたがる通信の経路選択ポリシーを調べ、それに基づいて帯域幅を調整する性能予測モデルを作成した。

一方、ランク配置情報の取得は、Alltoall 関数の最初の呼び出し時に行う。これを実行開始時に行うこともできるが、今回は Alltoall 通信関数の試験実装が目的であったため、この関数内で初回呼び出し時にランク配置情報の取得を行っている。

このランク配置情報と、事前に作成していた性能予測を用いて、最初の呼び出し時にアルゴリズムを絞り込む。今回の実装では、用意されているアルゴリズムのうち、最も所要時間が短いと予測されたものに対して、2 倍以上の所要時間がかかると予測されたアルゴリズムを、アルゴリズム選択の候補から除外する。なお、このアルゴリズムを除外する閾値は、今後、システムの性能安定性や性能予測モデルの精度を確認しながら調整する予定である。

候補アルゴリズムを絞り込んだ後の最速アルゴリズムの選択には STAR-MPI を用いる。STAR-MPI の処理は，以下の 2 つのフェーズに分けて行われる．

Learning フェーズ：

集団通信が呼ばれると，選択候補のアルゴリズムのうちの一つを用いて実行し，結果を返す．その際所要時間を計測し，記録する．全ての候補アルゴリズムについて規定回数の計測が終了するまで，各集団通信呼び出しに対してこのフェーズを実行する．全ての候補アルゴリズムの計測が完了すると，それらの所要時間の記録から最も高速なアルゴリズムを選出し，次の Monitoring フェーズで使用するアルゴリズムとする．なお，アルゴリズムの選出に用いる各アルゴリズムの所要時間としては，全プロセスの所要時間の平均値を用いる．

Monitoring フェーズ：

Learning フェーズで選出されたアルゴリズムを用いて集団通信を行う．このフェーズは Learning フェーズが終了してアルゴリズムが選択された後，各集団呼び出しに対して実行する．実際の計算環境では実行中に状況が大きく変化することも考えられるため，定期的に集団通信の所要時間と Learning フェーズで得られた所要時間を比較する．もし，計測した時間が Learning フェーズ時の所要時間から大きく変動した場合，他のアルゴリズムの方が高速となった可能性があるので，再度 Learning フェーズに入り，アルゴリズムを選択する．

今回実装した、動的アルゴリズム選択を行う Alltoall 通信関数の RICC における性能を検証するため，Alltoall 通信を 200 回呼び出すプログラムを用いて所要時間の計測を行った．このプログラ

ムは STAR-MPI のベンチマークプログラムとして提供されているものである.

このプログラムを，プロセス数とメッセージサイズを変えながら RICC のメタジョブスケジューラに投入した．RICC において用いられているメタスケジューラは，出来るだけ空いている計算ノードが少なくなるようにジョブを割り当てて，システムにおけるジョブの充填率を上げることにより，総合的な処理速度の向上を図っている．そのため、同じプロセス数のジョブであっても、割り当てられるノードの位置に規則性が無い。その結果、通信遅延の変動や通信衝突の影響により、通信性能が不安定となる。従来の MPI ライブラリの集団通信関数で用いられている静的なアルゴリズム選択手法では、このような環境では最適なアルゴリズム選択ができない。一方、提案手法や STAR-MPI では、実際に計算機上でアルゴリズムを試すため、状況に応じたアルゴリズム選択を行うことができる。

3． 結果

図 6～図 9に、メッセージサイズが 64KB のときの、1 プロセス x 32 ノード、2 プロセス x 32 ノード、4 プロセス x 32 ノード、8 プロセス x 32 ノードのそれぞれのプロセス数での Alltoall 通信の平均所要時間を示す。X 軸はジョブの番号、Y 軸は所要時間である。示されている所要時間は、アルゴリズムを絞り込んだ動的アルゴリズム選択(DYN GROUP)，アルゴリズムを絞り込まない動的アルゴリズム選択 (DYN NOGROUP)，および，Bruck(Bruck)，Pairwise(Pair)，Pairwise with Light-Barrier(PairLB)，Pairwise with MPI-Barrier(PairMB)，Pairwise with One-Barrier(PairOB)，Ring(Ring)，Ring with Light-Barrier(RingLB)，Ring with MPI-Barrier(RingMB)，Ring with One-Barrier(RingOB)の各アルゴリズムの所要時間である。

これらの結果より、提案手法によってほぼ最適なアルゴリズムが選択され、常に最速に近い性能が得られていることが分かる。また、アルゴリズムを絞り込まない場合との比較より、アルゴリズムを絞り込むことによる性能向上が得られることが分かる。

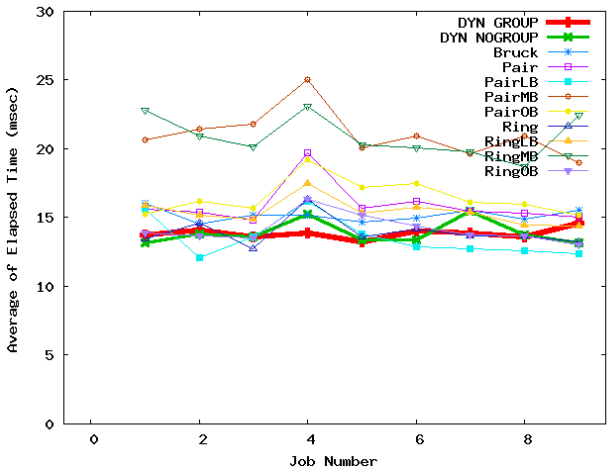

図 6　Alltoall 通信の平均所要時間
64KB、1 プロセス x 32 ノード

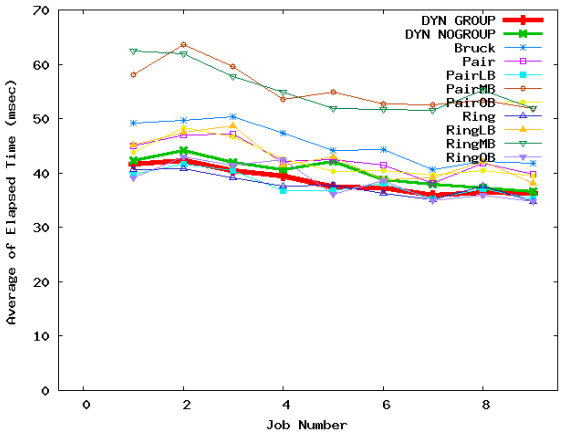

図 7　Alltoall 通信の平均所要時間
64KB、2 プロセス x 32 ノード

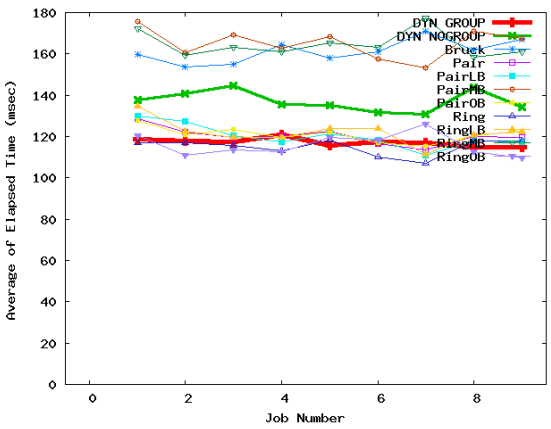

図 8　Alltoall 通信の平均所要時間
64KB、4 プロセス x 32 ノード

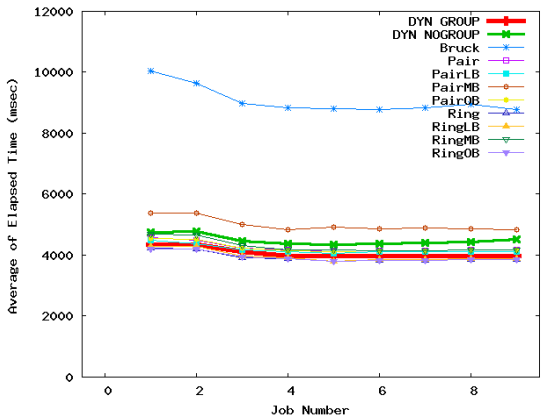

図 9　Alltoall 通信の平均所要時間
64KB、8 プロセス x 32 ノード

さらに、アルゴリズムを絞り込むことによる効果を検証するため、Learning フェーズの所要時間の比率を図 10～図 13 に示す。ほぼすべての場合で、アルゴリズムを絞り込むことによって Learning フェーズの所要時間を短縮できている。

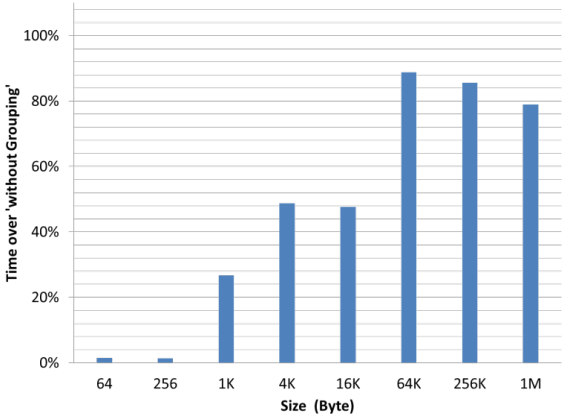

図 10　アルゴリズム絞り込みを行わない場合に対する所要時間の比率（Learning フェーズ）
1 プロセス x 32 ノード

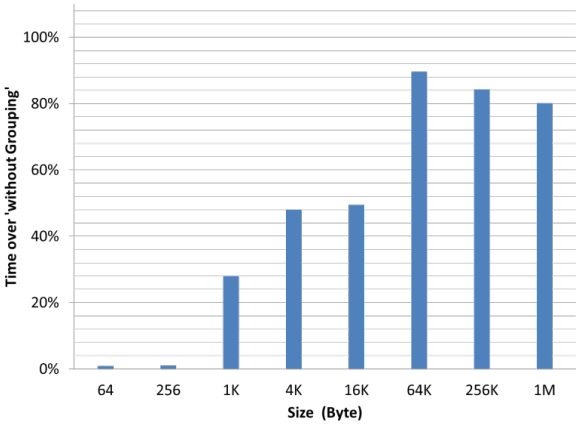

図 11　アルゴリズム絞り込みを行わない場合に対する所要時間の比率（Learning フェーズ）
2 プロセス x 32 ノード

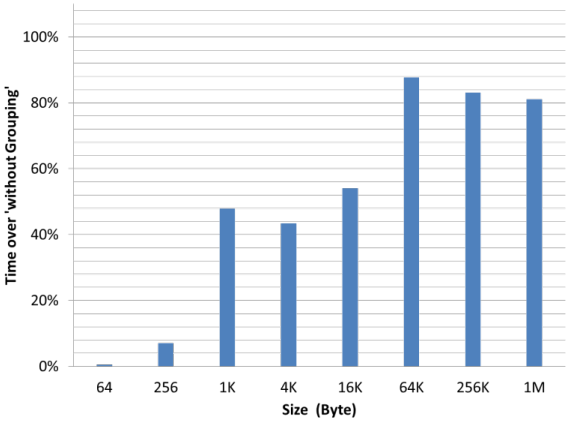

図 12　アルゴリズム絞り込みを行わない場合に対する所要時間の比率（Learning フェーズ）
4 プロセス x 32 ノード

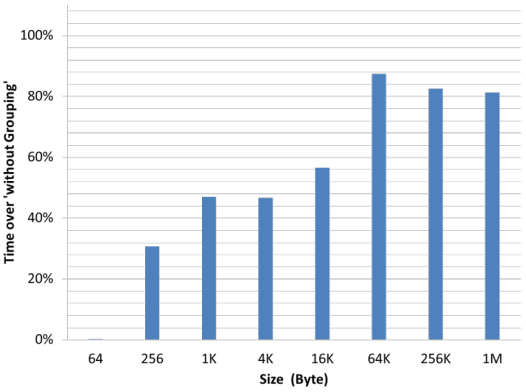

図 13　アルゴリズム絞り込みを行わない場合に対する所要時間の比率（Learning フェーズ）
8 プロセス x 32 ノード

4．　まとめ

プロセスの配置に応じてアルゴリズムを絞りこ

んだうえで、動的にアルゴリズムを選択する集団通信のアルゴリズム選択手法を提案した。また、この手法で Alltoall 通信関数を実装し、RICC で効果を検証した結果、低いオーバヘッドで最適に近いアルゴリズム選択を行えていることが確認できた。

５． 今後の計画・展望

今後は、Alltoall 以外の通信関数についても実装を目指したい。また、実アプリケーションにおいて、負荷バランスが不均衡な場合での提案手法の効果の検証を行いたい。

６． RICC の継続利用を希望の場合は、これまで利用した状況（どの程度研究が進んだか、研究においてどこまで計算出来て、何が出来ていないか）や、継続して利用する際に行う具体的な内容

今年度の研究で、RICC のトポロジを考慮した Alltoall 通信関数を実装できた。来年度も継続して利用し、Alltoall 以外の通信関数の実装を目指したい。

平成 23 年度　RICC 利用研究成果リスト

【国際会議、学会などでの口頭発表】

"Effect of Dynamic Algorithm Selection of All-to-All Communication on Environments with Unstable Network Speed",
Takeshi Nanri and Motoyoshi Kurokawa,
2011 International Conference on High Performance Computing and Simulation, 2011.07