平成 25 年度 RICC 利用報告書

課題名（タイトル）：

大規模並列計算機の効率的な利用・運用に関する研究

利用者氏名： ○南里 豪志*, 稲富 雄一*

所属： *本所 情報基盤センター

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

並列計算機をバッチ型で利用する場合、利用者から申請されたジョブがジョブスケジューラによって計算ノードに割り当てられる。この時ジョブスケジューラは、予め設定されたスケジューリングポリシーに従って、プロセスの配置を決定する。従来、多くのジョブスケジューラは、計算ノードをジョブクラスと呼ばれる大小の連続ブロックに分割し、それらのジョブクラス単位でプロセスを割り当てていた。これに対し、理化学研究所の RICC(RIKEN Integrated Cluster of Clusters) のメタスケジューラでは、出来るだけ空いている計算ノードが少なくなるように、計算ノードの位置に関係なくプロセスを配置する。その結果ジョブの充填率が向上するため、並列計算機全体のスループットも向上することが期待できる。一方、個々のジョブにおけるプログラム中の通信の性能は、そのジョブに割り当てられた計算ノードの相対的な位置関係によって大きく影響を受ける。例えば計算ノード同士の距離によって通信遅延時間が変動する。さらに、距離が遠い場合、他の通信と経路が競合する通信衝突が発生し、通信帯域幅が低下する可能性が高くなる。

特に、並列プログラムにおける全プロセスによる総和の計算や全対全のコピー等の、集団通信と呼ばれる通信では、この転送性能の低下が大きな問題となる。集団通信には複数の実装アルゴリズムが用意されており、状況に応じて最適なアルゴリズムを選択して使用する。従来は、事前に十分な調査を行うことにより、使用するプロセス数と転送するメッセージサイズから最適なアルゴリズムを選択している。しかし通信性能が安定しない環境では、同じプロセス数、メッセージサイズでも事前の調査時とは最適なアルゴリズムが変化する可能性が高い。そのため、実行時の状況に応じて適応的にアルゴリズムを選択する仕組みが必要となる。

そこで本課題では、集団通信の実装アルゴリズムを実行中に試しながら、最適なものを選択する動的最適化技術の研究開発を行っている。この最適化技術は、プログラム中で繰り返し呼ばれる集団通信について、その最初のほうの繰り返し時に実装アルゴリズムを変えながら実行することにより、実行時の状況における各アルゴリズムの性能を取得し、その情報に基づいて最適なアルゴリズムを選択するものである。この手法の問題として、他のアルゴリズムに対して非常に遅いアルゴリズムも試すことによるオーバヘッドの増大がある。そこで、アルゴリズムの性能予測モデルを作成し、それを用いて選択対象のアルゴリズムを絞り込む技術を開発した。本年度は、この技術を用いた集団通信関数を実装し、効果を計測した。

２． 具体的な利用内容、計算方法

集団通信のアルゴリズムを動的に選択するツール STAR-MPI に、我々が開発したアルゴリズムの絞り込み技術を実装し、評価した。絞り込みによって期待される効果は、遅いアルゴリズムを除外することによる動的選択のオーバヘッド低減である。しかし、予測モデルによる各アルゴリズムの予測時間には誤差が含まれるため、確実に遅いプログラムを特定することが出来ない。そのため、絞り込み方によっては、間違って実際は速いアルゴリズムを除外する可能性がある。そこで、絞り込み手法として以下の３通りを考え、比較した。

・Limit Number: 予測時間の順位で上位のものを残す。

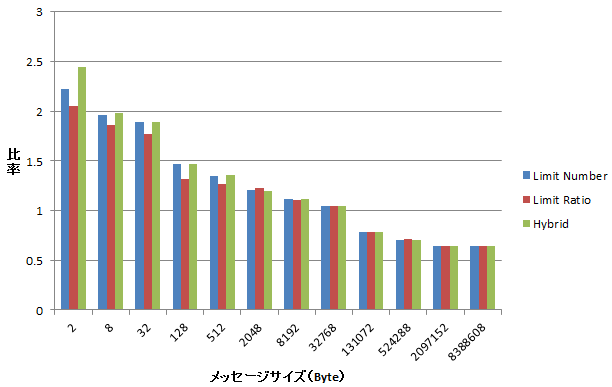
・Limit Ratio: 予測時間が最短のものに対する比率が一定の値以下のものを残す。

・Hybrid: 上記の両方のアルゴリズムを残す。

これらの方法を、Allgather 通信を 100 回呼び出すプログラムで試し、アルゴリズム絞込みの効果を検証した。実験には京コンピュータの互換機である Fujitsu PRIMEHPC FX10 を使用した。

3. 結果

以下に、96 ノードでの実験結果を示す。



これは、ジョブ投入時に、利用するノード群の仮想的な形状として 8x3x4 を指定した場合のものである。縦軸はアルゴリズムの絞り込みを行わなかった場合に対する所要時間の比率である。メッセージサイズが 32KB 以下の場合、アルゴリズムの絞り込みを行わないほうが高速であった。これは、性能を予測するためのオーバヘッドが、アルゴリズムの絞り込みによる効果を上回ったためである。一方、メッセージサイズが 128KB 以上では所要時間を大幅に短縮できて切ることが分かる。また、アルゴリズムの絞り込み手法による違いは若干あるものの、全体的な効果は、どの手法でも大差ないことが分かった。

4. まとめ

集団通信アルゴリズムを実行時に選択する手法として、既存の実測時間のみを用いる手法に対し、事前に各アルゴリズムの性能予測を行って遅いアルゴリズムを除外する手法を提案し、実装した。実験により、提案手法の効果を確認した。

5. 今後の計画・展望

提案手法の有効性を実アプリケーションで確認するとともに、Allgather 以外の集団通信についても実装する。

6. 利用がなかった場合の理由

本年度は、主に Tofu ネットワークでの実装と計測に時間を取られてしまい、RICC を利用できなかった。