様式 C-19

科学研究費補助金研究成果報告書

平成　21 年　4 月　13 日現在

研究種目：若手研究（B）
研究期間：　2006～2008
課題番号：18700065
研究課題名（和文）IPv6 と Myrinet による階層型クラスタ上の OpenMP

研究課題名（英文）　OpenMP on Hierarchical Cluster with IPv6 and Myrinet

研究代表者
南里　豪志　（NANRI TAKESHI）
九州大学情報基盤研究開発センター・准教授
研究者番号：70284578

研究成果の概要：IPv6 対応インターネットと Myrinet で構成された階層型クラスタ向けに RMA 機構を開発するとともに、通信最適化技術を開発した。また、ソフトウェアで制御されたキャッシュメモリシステムを RMA 機構上に構築して、有効性を確認した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	1,300,000	0	1,300,000
2007 年度	700,000	0	700,000
2008 年度	600,000	180,000	780,000
年度			
年度			
総　計	2,600,000	180,000	2,780,000

研究分野：総合領域
科研費の分科・細目：情報学・ソフトウェア
キーワード：ハイパフォーマンスコンピューティング、分散共有メモリ、OpenMP、プログラム最適化、ソフトウェアキャッシュ

１．研究開始当初の背景

近年、クラスタの普及により、複数のクラスタを結合した複合的なクラスタが、大規模計算機システムのプラットフォームとして注目されており、その性能の有効活用が望まれていた。しかしながら、このような計算機システムはクラスタ内部の高速ネットワークとクラスタ間の汎用ネットワークによる階層的な構造をしているため、以下の問題があった。

- プログラミングモデルが複雑：
  このような階層型クラスタ上で動作する並列プログラムの記述手段としては、低レベルなメッセージパッシングモデルを用いるものしか提供されていないため、プログラム作成やチューニングが困難である。
- クラスタ間の通信コストが大きい：
  クラスタ間の通信に汎用ネットワークを用いるため、通信遅延時間が大きい。さらに既存の並列計算環境では、IPv4 のアドレス枯渇問題を回避するためにクラスタ内ではプライベートアドレスが用いられており、他のクラスタのノードと直接通信できない。

階層型クラスタ上の並列プログラム開発環境としては、MPICH-G2 や PACX-MPI などがあるが、これらはメッセージパッシングモデルであり、しかも IPv4 にしか対応していないため、クラスタ間通信を中継するノード

がボトルネックとなって性能が低下する。

２．研究の目的

本研究では、IPv6 対応の OpenMP 環境を階層型クラスタ上に構築するために必要な技術について研究、開発を行う。この技術を用いることにより、階層型クラスタにおけるプログラミングモデルとして、メッセージパッシングモデルよりもはるかにプログラム開発が容易な OpenMP を利用することができるようになる。また、IPv6 に対応することにより、プライベートアドレスを用いる必要がなくなり、クラスタ内の各ノードが直接他クラスタのノードと通信できるようになるため、クラスタ間通信コストを軽減することができる。

具体的には、まず共有メモリモデルである OpenMP による並列プログラムを階層型クラスタ上で動作させるために必要となる分散共有メモリシステムを研究対象とする。さらに、この分散共有メモリシステムにおける通信コストをさらに軽減させるための技術として、通信最適化技術、及びソフトウェアキャッシュ技術について研究する。これらにより、階層型クラスタシステムの性能を容易に活用できる環境の構築を可能とする。

３．研究の方法

本研究では、まず研究代表者が平成 15～17 年度の若手研究｛B｝で開発した遠隔メモリアクセス機構について、クラスタ間通信に用いたソケットを IPv6 に対応させる。これにより、ネットワークアドレスの枯渇問題が解消されるため、各計算ノードにグローバルアドレスを持たせ、クラスタをまたがった任意のノード間で、ゲートウェイノードを介さずに直接通信を行うことができる。さらに、遠隔メモリアクセス発生時に、通信相手が自分のクラスタ内にいるか否かを判断して、自動的に適切な通信手段を選択するために、ノード情報の取得と管理を行うための機能を追加する。

次に、通信最適化手段として、特に計算機の規模が大きくなってくると問題となる集団通信と呼ばれるものに着目し、その高速化を図る。集団通信は、並列計算に参加しているプロセスのグループ全体が参加して行う通信の総称であり、具体的には、全ノードでの総和計算(Allreduce)や、あるノードが所有するデータの全ノードへのコピー(Broadcast)等である。これらについて、実行時の状況に応じた実装手段を選択する機能に関する研究を行い、通信の効率化を図る。さらに、並列プログラム中の通信パターンから、プロセスのノードへの配置を最適化する技術について、研究を行うことにより、クラスタ間通信における通信衝突の回避を図る。

さらに、遠隔メモリアクセス時に、アクセス対象のデータを含むブロック単位でコピーを取得し、ローカルメモリに蓄えておくソフトウェア制御のキャッシュメモリシステム技術について研究を行い、プログラム中の参照の局所性を活かした高速化が行えるようにする。

４．研究成果

まず、階層型クラスタにおいて、他の計算ノードへの遠隔メモリアクセスを可能とする RMA 機構の IPv6 への対応を行った。、また、相手先のノードがクラスタ内か否かに応じて自動的に通信方式を切り換え、常に最も効率の良い方式で通信を行う気候についても実装した。その結果、クラスタ間通信によるボトルネックの影響が無くなり、通信に要するコストが大幅に低減できることを確認した。

次に、通信最適化技術として、集団通信の高速化と、プロセス配置の最適化について研究を行った。

このうち集団通信の高速化では、各プロセスの負荷状況に応じて集団通信内部で行う通信の順序を調整することにより、負荷の不均衡の影響を軽減し、システム全体の利用効率を向上させる技術について開発して評価した。この技術は、集団通信における各プロセスの待ち時間の違いから負荷状況を検出し、負荷が低くて待ち時間が長いプロセスから順に通信を行うことができるように、通信の順序を調節するものである（図 1)。

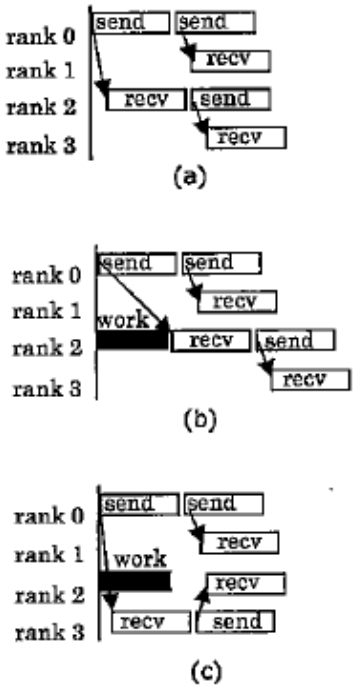

図 1　通信順序の調整

この技術に関する予備実験として、疎行列同士の積を行うプログラムに対して適用した

結果を表１に示す。ここで Total はプログラム全体の時間、Broadcast は集団通信 Broadcast に要した時間であり、Orig が最適化前、Opt が最適化後の時間を示す。これにより、順序調整が通信時間削減に効果があることを示した。

表 1 集団通信内の通信順序調整の効果

Procs	Total(sec)		Broadcast(sec)	
	Orig	Opt	Orig	Opt
4	74.5	75.7	8.49	8.92
8	89.3	87.8	21.5	17.1
16	120	114	50.2	42.5

また、集団通信高速化の技術として、集団通信アルゴリズムの自動選択技術についても研究を行った。集団通信は内部で一対一通信を繰り返すことによって実現されるが、その組み合わせ方、すなわち実装アルゴリズムによって性能が大きく変化する。また、それぞれのアルゴリズムについて、メッセージサイズやプロセス数によって適性が変化するため、状況に応じて最適なアルゴリズムを選択する必要がある。さらにシステムの規模が大きくなると、同じメッセージサイズ、同じプロセス数でもプログラム内の負荷バランスや、そのシステムで実行されている他のジョブの影響、通信衝突の発生頻度の違い等により、最適なアルゴリズムの予測が難しくなる。そこで本研究では、プログラム中で集団通信が呼び出されるたびに異なるアルゴリズムを試すことにより、実行中にアルゴリズムを選択する技術に関する研究を行った（図２）。

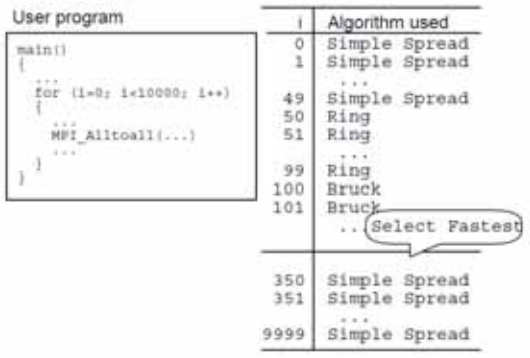

図 2 実行時アルゴリズム選択技術

同様の技術として、米国の STAR-MPI があるが、本研究の手法は、各アルゴリズムの性能予測モデルを用いて対象アルゴリズム数を絞り込むことにより、アルゴリズム選択のコストを低減している。この技術の予備実験として、Alltoall 通信を繰り返し呼び出すプログラムを用いて性能評価を行ったところ、図３の Dynamic が示す通り、メッセージサイズの変化に対して、常に最も効率の良いアルゴリズムを選択することにより、性能向上が図れていることが分かる。

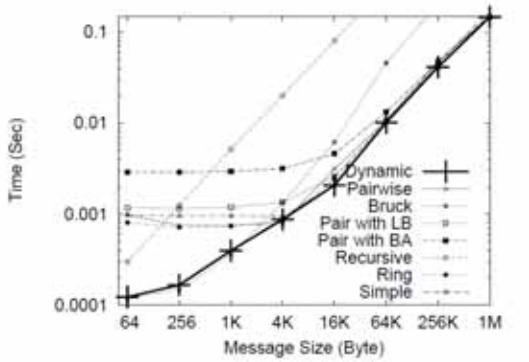

図 3 アルゴリズム選択の効果

もう一つの通信最適化技術として、通信パターンを考慮したプロセス配置最適化技術について研究を行った。これは、プログラム中の通信パターンを解析し、それをもとに通信頻度の高いプロセス同士を近くに、頻度の低いプロセス同士を遠くに配置することにより、通信路の衝突を軽減し、通信コストの削減を図るものである。この技術は、プログラム中のデータ依存関係から同時に発行される可能性の高い通信グループを解析し、それぞれの通信グループでの通信コストが最低となるようにプロセス配置の組み合わせを選択することにより、実現する。この、組み合わせの選択は、組み合わせ最適化問題となるため、simulated annealing による発見的手法を用いた。NPB(NAS Parallel Benchmarks)の CG を用いた予備実験の結果、本手法によって通信のボトルネックとなる経路における通信衝突を軽減でき、その結果、通信コストを削減できることを示した。

最後に、ソフトウェアキャッシュメモリ技術について、紹介する。これは、分散共有メモリシステムにおいて、あるノードのプロセスが他のノードが所有するデータに対してアクセスする際、そのデータを含むブロック単位でローカルメモリにコピーすることにより、それ以降、そのブロック内のデータへのアクセスを高速におこなうことを可能とする技術である。この技術の開発においては、複数のノードが同じメモリブロックをコピ

ーした場合にシステム全体の一貫性を保持するためのコストが重要な問題となる。本研究では、この一貫性保持に必要な情報をクラスタ毎に管理し、クラスタ内のコピーを一括して管理することにより、少ない通信コストでの実現を行った。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① 森江 善之，末安 直樹，松本 透，南里豪志，石畑 宏明，井上 弘士，村上 和彰，通信タイミングを考慮した衝突削減のための MPI ランク配置最適化技術，情報処理学会論文誌(トランザクション) コンピューティングシステム，Vol.48, No. SIG13, pp. 192-202, 2007，査読有.
② 曽我 武史，栗原 康志，南里 豪志，黒川 原佳，村上 和彰,負荷バランスの動的最適化による MPI ブロードキャスト性能改善，情報処理学会論文誌（トランザクション）コンピューティングシステム，Vol. 1, No. 3, pp. 67-82, 2008，査読有.
③ Nzigou Mamadou, H., Nanri, T. and Murakami, K., Performance Models for MPI Collective Communications with Network Contention, IEICE Transactions on Communications}, Vol. E91-B, No. 4, pp.1015-1024, 2008，査読有.

〔学会発表〕（計６件）

① Gu, FL., Nanri, T., Murakami, K., Implementation of GAMESS on Parallel Computers: TCP/IP versus MPI, Intl. Conf. of Comp. Methods in Sci. and Eng., pp. 1517-1519, 2006，査読有.
② Domingues, G., Morie, Y., Gu, FL., Nanri, T. and Murakami, K., SMMH - A Parallel Heuristic for Combinatorial Optimization Problems, International Conference of Computational Methods in Sciences and Engineering, 2007，査読有.
③ Gu, FL. Nzigou Mamadou, H., Domingues, G., Nanri, T. and Murakami, K., Investigating the Performance of Collective Communications on SMP Clusters: a Case for MPI_Allgather, International Conference of Computational Methods in Sciences and Engineering, 2007，査読有.
④ Soga, T., Kurihara, K., Nanri, T., Kurokawa, M. and Murakami, K. Dynamic Optimization of Internal Communications in MPI Broadcast, Recent Advances in PVM and MPI. LNCS 4757, pp. 387-388, poster, Oct. 2007，査読無.
⑤ Baba, S., Onoue, Y., Nanri, T. and Fujino, S. Dependence on loop distribution of performance in hybrid-parallel IDR(s) method, Proceedings of HPC Asia, pp.46-53, 2009，査読有.
⑥ Soga, T., Nanri, T., Kurokawa, M. and Murakami, K., Profiling Technique for Dynamic Optimization According to Waiting Time, Proceedings of HPC Asia, pp.270-276. 2009，査読有.

６．研究組織
(1)研究代表者

南里　豪志（NANRI TAKESHI）
九州大学・情報基盤研究開発センター・准教授
研究者番号：70284578