Vol.2013-HPC-138 No.13
2013/2/21

遠隔 Atomic 通信を用いた省メモリ性実現のための方式検討

住元 真司[1]　安島 雄一郎[1]　秋元 秀行[1]　安達 知也[2]　岡本 高幸[1]　三浦 健一[1]

概要：ポストペタスケール規模での通信ライブラリでは、省メモリ性の実現が必須となる。省メモリ性を実現するためには、少ないバッファでいかに効果的かつ安定した通信を行うかが課題となる。本論文では、ポストペタスケール規模での省メモリ性実現のために、何が問題となりえるのかを明らかにし、遠隔 Atomic 通信を用いた実現方式の設計について検討する。

Methods for Memory Saving Communication Library Using Remote Atomic Memory Communication

Abstract: It is important to reduce memory consumption of communication library for post peta-scale systems. To realize the saving memory of communication library, the issue is to implement high performance communication with less memory and stable communicationx effectively. This paper presents issues and design for realizing momory saving communication library for post peta-scale systems, and some methods to solve the issues using remote atomic communication.

1. はじめに

我々は JST-CREST において、ポストペタスケールシステムにおいて省メモリを実現する通信ライブラリの研究開発を行っている。ポストペタスケールシステムにおいては、演算性能の増加に比例したメモリ量の増加は期待できない上、並列プロセス数は増加すると想定され、通信ライブラリに必要なメモリ量も増加する。このような環境においては、大規模アプリケーション実行時にメモリが逼迫した状況になりえるため、プログラムの実行安定性と性能安定性が問題となる [1]。このため、ポストペタスケール規模における通信ライブラリでは、少ないバッファでいかに効果的かつ安定した通信を行うかが課題である。

本論文では、ポストペタスケール規模での省メモリ性実現のために、何が問題となりえるのかを明らかにし、その課題を解決するため、遠隔 Atomic 通信を用いた実現方式について検討した結果を報告する。　本論文の構成は、第 2 章で、省メモリ低遅延通信プロトコルの実現方針について述べ、第 3 章で省メモリ化の課題を述べる。第 4 章で省メモリ通信のための通信方式についての検討内容、そして、第 5 章で関連研究、第 6 章でまとめを述べる。

[1] 富士通株式会社　次世代テクニカルコンピューティング開発本部/(独) 科学技術振興機構　戦略的創造研究推進事業
Fujitsu Limited., Next Generation Technical Computing Unit Japan Science and Technology Agency (JST), Core Research for Evolutional Science and Technology (CREST)
[2] 富士通株式会社　次世代テクニカルコンピューティング開発本部/ Fujitsu Limited., Next Generation Technical Computing Unit

2. 省メモリ・低遅延通信プロトコルの実現方針

我々は，JST CREST の研究領域「ポストペタスケール高性能計算に資するシステムソフトウェア技術の創出」の一つのテーマである「省メモリ技術と動的最適化技術によるスケーラブル通信ライブラリの開発」の研究課題プロジェクトに参画している。

本プロジェクトでは数千万 数億プロセスに耐える省メモリの通信レイヤを実現するために、通信資源の動的管理と低遅延通信を両立する技術開発を行っている。 既存の通信プロトコルは通信性能や品質を確保するために通信相手数に応じたメモリ確保が必要な構造となっており、エクサスケール向けの通信プロトコルでは省メモリ化技術が必須となっている。

しかし、一般に資源 (メモリ使用量) と性能・品質はトレードオフの関係にあるため、性能・品質の劣化をいかに最小にした省メモリ化技術を実現するかが課題である。 一

ⓒ 2013 Information Processing Society of Japan

般に通信を行うために必要なメモリには、送受信に必要な通信バッファと相手先の情報を格納する制御情報から構成されるが、エクサスケールシステムにおいては、この両者ともに極限までの省メモリ化の取り組みが必要になる。

このため、省メモリ・低遅延通信プロトコルの実現方針としては、以下の 3 つを基本とする。

(1) 通信時に必要な最小限の制御情報のみを格納し、必要時に動的に制御情報を取得する通信制御方式の採用
(2) 信頼性確保に必要な制御通信を含めたデータ通信処理の最適化
(3) これらを、RDMA と遠隔 Atomic 操作を活用することにより実現

3. ポストペタスケール規模での省メモリ化実現の課題

ポストペタスケール規模において、プログラムの実行安定性と性能安定性を実現するために省メモリ化の実現が必須である。本章では、省メモリ化実現のための課題について述べる。このために、既存 MPI[2] 実装においてメモリ消費に大きく関係がある通信方式について整理し、課題を明らかにする。

3.1 ポストペタスケール規模での省メモリ通信の目標

ポストペタスケール規模での省メモリ通信の目標は、高い通信性能（低遅延高バンド幅）を確保しながら，プログラムの実行安定性と性能安定性を確保することである [1]。

3.2 既存の MPI 実装

既存の一般的な MPI 通信においては、次の 2 つの通信方式が利用される。

Eager 方式: 通信遅延を抑えるため、通信相手との同期なしにメッセージを送信する方式である。送信側が主体となって通信するため送信側主導型 (Sender Initiated) 通信とみなされる。本方式は、受信時にメモリコピーが発生するため、通信遅延に比べメモリコピー時間が小さい場合に有効である。通信性能は、一般に受信側のバッファメモリ量に比例するため、通信性能をあげるためにはより多くのメモリが必要になる。MPI 実装においては、一つの送信要求 (MPI_Send) に対して受信 (MPI_Recv) が一致した場合は、指定された受信バッファに直接コピーされるが、一致する受信がない場合は Unexpected Message として一時的な受信バッファを割り当てコピーした上で格納し、一致する受信が発行されるまで保持される。

Rendezvous 方式: 使用するメモリ量を抑制するため、制御通信を利用し受信側のバッファと送信側のバッファの準備が完了後にデータ転送を行う方式である。受信側が主体となって通信するため受信側主導型 (Receiver Initiated) 通信とみなされる。転送要求を受信側で制御できる他、RDMA 通信と組み合わせることにより、プロセッサによるメモリコピーが不要な通信を提供することが可能である (Zero-Copy 通信)。Rendezvous 方式の制御に必要な通信は数十バイトの短いメッセージであり、かつ、制御通信はメッセージ受信時に直ちに処理されるため、既存 MPI のの実装では制御通信は Eager 方式で制御メッセージを送信している。

MPICH[3] や Open MPI[4] では、これらの通信方式について、通信のメッセージ長が比較的に短い場合には Eager 方式を採用し、それ以上のメッセージ長の場合には Rendezvous 方式による実装となっている。これは、Eager 通信が多くの受信バッファを必要とするために、バッファが確保可能なメッセージ長については Eager 方式による通信性能の向上を狙っているからである。

3.3 ポストペタスケール規模通信での局所集中問題

第 3.2 節で述べた Eager 方式による通信は、高性能である半面、受信先の状態に関係なくメッセージを送る。このため、特定ノードに局所的にメッセージ送信が集中した場合にメッセージ受信が滞り指数関数的に性能劣化が発生する場合がある。これは、全体のシステム規模の増大と共に問題は拡大し、ハードウェアの受信バッファがオーバフローする場合ある。このオーバフローによる挙動はインタコネクトの種類によりハードウェアエラーになる場合がある。

ここで、インタコネクトハードウェアの受信バッファオーバフロー時の挙動を次にまとめる。

- InfiniBand: 基本的にオーバフローしないように受信バッファを割り当てるのが基本、Unreliable Datagram の場合パケットロス、Reliable Connection の場合は、一定の時間経過後に QP エラーが発生する。
- Tofu の場合： ユーザレベルでは RDMA のみのため、パケットロスは発生しない。半面，通信が集中した場合はネットワーク上にパケットが溜まり、パケットが一掃されるまでネットワークが通信不能となる。

このように、特定ノードに局所的にメッセージが集中した場合に、ネットワークが一時的に通信不能、もしくは、プログラム実行が停止する可能性がある。

さらには、本問題と Eager 方式による Unexpected Message による使用メモリ量増加が重なると、メモリ枯渇によるプログラム実行性能の劣化ならびにプログラムがエラー終了する場合がある。

これら 2 つの問題はノード規模が現在より一桁以上増えると想定されるポストペタスケール規模においては、さら

ⓒ 2013 Information Processing Society of Japan

に状態は悪化し、論文 [5] で述べたように、通信メッセージ長が短い場合にも、受信バッファのオーバフローが発生する。これは、制御メッセージの通信においても局所集中問題が発生しえることを意味している。それ故、これら 2 つの問題はポストペタスケール規模では根本的に解決する必要がある。

3.4 ポストペタスケール規模通信における省メモリ化実現の課題

第 3.3 節で述べた問題は、Eager 型（送信側主導型）通信が受信側の状態を把握すること無くメッセージを送信することに問題があるため、Eager 型通信ではなく Rendezvous 型方式 (受信側主導型) を採用すれば Unexpected Message によるメモリ枯渇問題はなくなるはずである。

しかし、Rendezvous 方式を採用したとしても、ポストペタスケール規模通信では、依然、制御通信のパケットによる局所集中問題が発生する。したがって、制御パケットの集中問題に対し，いかに少ない通信バッファ量で回避するかがポストペタスケール規模での省メモリ化実現の課題である。

4. 遠隔 Atomic 通信を用いた省メモリ通信のための通信方式検討

本章では，第 3.3 節で述べた，ポストペタスケール規模での省メモリ化実現の課題を解決する手法として、遠隔 Atomic 通信を用いた省メモリ通信のための通信方式について検討を行う。

4.1 通信集中問題緩和のためのアプローチ

通信集中問題の緩和を実現するためには単一ノードの受信処理自体の高速化に加え、システム全体で考える必要がある。単一ノードの受信処理の問題点は、既存の Rendezvous 型方式が要求受信、要求処理、応答パケットの配送処理をすべてホスト CPU で実行していることが問題である。したがって、これらの処理の簡略化が課題である。また、システム全体での対策としては、局所的な処理集中の検出と単位時間あたりの総要求数を一定以下に抑える仕組みの導入が重要となる。このため、次のアプローチをとることにする。

単一ノード処理の向上： 受信ノードの処理能力の向上として遠隔 Atomic 通信の活用により、受信ノード処理のホスト CPU 処理のオーバヘッド削減を考える。

システム全体での対策： 処理集中するパケット数を分散、緩和するために、処理ノードの分散配置による対策を考える。

4.2 遠隔 Atomic 通信

遠隔 Atomic 通信とは，分散メモリシステムにおいて、Atomic 処理を遠隔に実行するものである。RDMA Read/Write と組み合わせることで、相手のホスト CPU の介在することなく、メモリ参照と更新が可能になる。遠隔 Atomic 通信処理としては，例えば，InfiniBand に置いては、”Fetch and Add 処理” や”Compare and Swap 処理” が利用可能である [6]。

4.3 遠隔 Atomic 通信を用いた制御通信の実現方式検討

Rendezvous 型方式の目的はひとつの送信に対する相手先の受信バッファを固定することである。しかし、既存の Rendezvous 型方式はホストプロセッサにより処理されるため、ホストプロセッサの処理状態により処理が滞る場合がある。したがって，複数のノードから一つの受信バッファキューを排他的に選択するか、受信バッファ数分だけの送信に絞ることができればよい。

この 2 つの処理を実現するために、遠隔 Atomic 通信を用いた Redezvous 型方式を提案する。本方式は，受信バッファの確定を実現する受信バッファ獲得方式と既存の通信方式上にメッセージ送信の制限を付加する送信メッセージ数制限方式がある。これらの方式の詳細を述べる。

受信バッファ獲得方式： 特定ノードの制御通信の受信バッファを遠隔ノードから獲得することにより、制御通信の受信バッファオーバフローを避ける方式である。具体的には、制御通信受信バッファとしてバッファ配列を準備、Producer-Consumer カウンタを受信バッファのあるノードに確保し、送信ノードから遠隔 Fetch and Add 処理でインデックス番号を排他的に獲得、獲得したバッファに対して RDMA Write することにより、バッファ配列数だけを送信可能にする。制御通信の受信バッファを持つノードは，制御メッセージの処理後に Consumer 側のカウンタを+1 することにより、受信バッファを開放する。バッファ配列は Producer-Consumer カウンタによりサイクリックに再利用する。

送信メッセージ数制限方式： 特定ノードへの送信メッセージ数を受信バッファ数を制限することにより、制御通信の受信バッファオーバフローを一定数以下に制限する方式である。具体的には、Producer-Consumer カウンタを準備して遠隔 Fetch and Add 処理で Producer カウンタを+1 してメッセージ転送、受信側はメッセージ受信毎に Consumer カウンタを+1 する。受信メッセージキュー数分だけを送信可能にする。カウンタはサイクリックに再利用する。

具体的な動作の違いを、既存の方式との比較を行うことで明らかにする。

Eager 方式による Message 通信

図 1 に Eager 方式による Message 通信モデルを示す。

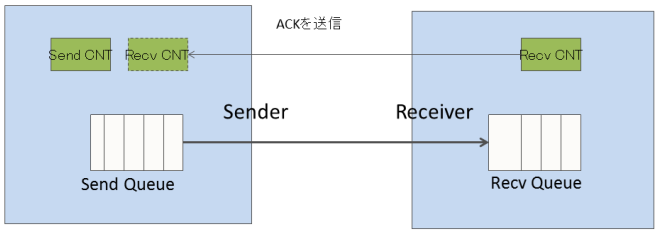

図 1 Eager 方式による Message 通信

Eager 方式は受信側の受信バッファの空き状況を把握すること無くメッセージを送信するために、例えば，受信側に複数のノードからの送信が集中しても、送信側ではその状態を把握することができない。受信バッファ数以上のメッセージが到着した場合、インタコネクトにより、メッセージの廃棄、受信待ちのために、メッセージが溜まる。

Rendzvous 方式による RDMA 通信

図 2 に Rendzvous 方式による RDMA 通信モデルを示す。

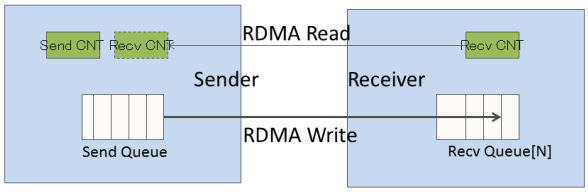

図 2 Rendzvous 方式による RDMA 通信

本方式においては、受信側との同期をとることにより通信を行うため、受信側の状況を把握することができる。しかし、メッセージが局所的に集中した場合に、Rendzvous 方式による同期のためのホストプロセッサにより制御メッセージ処理を受信側で処理するため、ホストプロセッサによる処理がボトルネックになる他、制御メッセージの集中によるネットワーク網の通信遅延が課題である。

遠隔 Atomic 通信を用いた受信バッファ固定方式

図 3 に遠隔 Atomic 通信を用いた受信バッファ固定方式のモデルを示す。

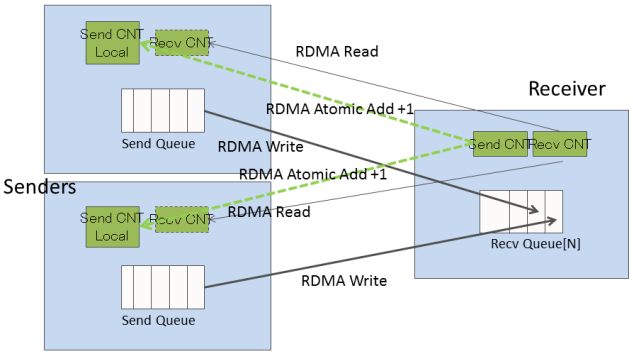

図 3 遠隔 Atomic 通信を用いた受信バッファ固定方式

遠隔 Atomic 通信を用いた受信バッファ固定方式のモデルでは、図 3 に示すとおり、受信側に配列の要素数 N(N は整数) の受信バッファ配列 RecvQueue[N]、送信カウント SendCNT と処理済みを示す受信カウンタ RecvCNT をおく。SendCNT の値について Recv Queue[SendCNT%N] が対応するものとみなす。SendCNT、RecvCNT は整数で 0 に初期化される。

そして、次のようにメッセージ通信処理を行う。

(1) 送信する Sender は、Receiver の RecvCNT を RDMA Read により獲得する (RCNTi)。
(2) 送信する Sender は、Receiver の SendCNT に対し遠隔 Atomic Add により、SendCNT を 1 増加させる。
(3) 遠隔 Atomic Add の実行結果として加算させる前の SendCNT である SCNTi を獲得、SCNTi と RCNTi の値の違いにより以下の動作を行う。

SCNTi 値が RCNTi+N 未満の場合： Receiver の Recv Queue[NTi%N] のバッファアドレスに対し、RDMA Write によりメッセージを送信

SCNTi 値が RCNTi+N 以上の場合： Receiver の RecvCNT を RDMA Read により RCNTi を獲得し、SCNTi 値が RCNTi+N 未満になるまで待つ。

本方式においては、前述の **Rendzvous 方式による RDMA 通信**に対して，遠隔 Atomic 通信を使用することにより、ホストプロセッサによる同期処理することなく

ⓒ 2013 Information Processing Society of Japan

同期することが可能である。

遠隔 Atomic 通信を用いた受信バッファ数制限方式

図 4 に遠隔 Atomic 通信を用いた受信バッファ数制限方式のモデルを示す。

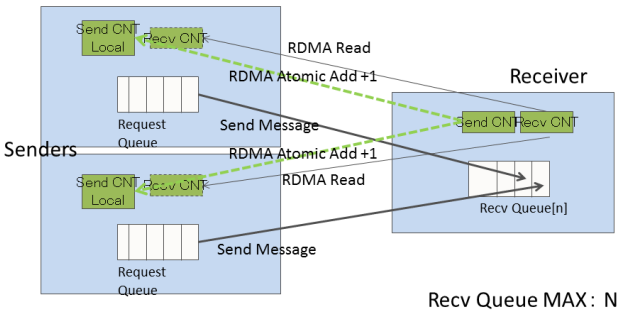

図 4 遠隔 Atomic 通信を用いた受信バッファ数制限方式

本方式においては、前述の**遠隔 Atomic 通信を用いた受信バッファ固定方式**にと動作と効果は同様で、送信側からのメッセージの送信数を N 以下に抑える効果がある。

以上のように、遠隔 Atomic 通信を利用することにより、Rendezvous 型方式による処理を受信側ホスト CPU の介在すること無く実現することができる。

4.4 遠隔 Atomic 通信を用いた局所集中緩和方式の検討

第 4.3 節で述べた方式は、ある一つのノードの有限数のバッファがオーバフローしない方式としては機能する。しかし、依然として、バッファやカウンタ獲得のための遠隔 Atomic 通信はそのノードに集中することは避けられない。これを緩和する方式として、図 3、図 4 における SendCNT、RecvCNT カウンタの分散化を検討する。

基本的な方式のアイデアは、図 3、図 4 における SendCNT、RecvCNT カウンタについて、第 4.3 節では受信バッファを持つノードと同じノードにおいたが，これを遠隔 Atomic 通信を用いて他のノードに分散しておく方式である (図 5、図 6)。

図 5 は一つのノードに分散させる方式である。本方式により、送信のための要求を他のノードで実行することにより、受信ノードへの要求処理のインタコネクトとメモリ負荷をオフロードすることが可能である。

図 6 は複数のノードに分散させる方式である。複数のノードに分散させる場合は、カウンタを分散させるノードのそれぞれのカウンタの総和が、受信側の受信バッファ数以下であれば動作可能である。本方式により、受信ノードへの要求負荷が分散ノード数分に削減することが期待される。

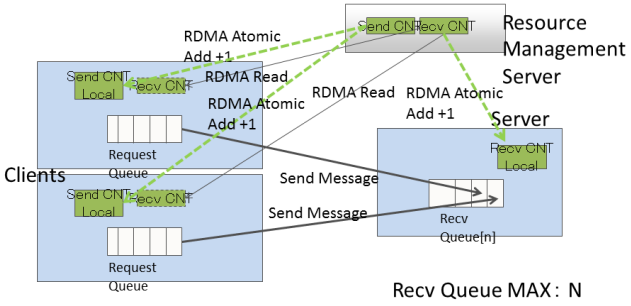

図 5 遠隔 Atomic 通信を用いた受信バッファ数制限方式:分散版

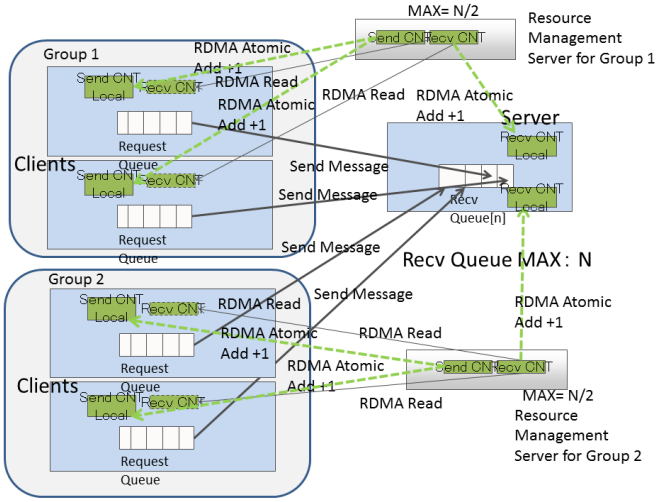

図 6 遠隔 Atomic 通信を用いた受信バッファ数制限方式:分散版 2 グループ

本手法により、受信バッファ獲得方式、送信メッセージ数制限方式についての処理分散は次のようになる。

受信バッファ獲得方式： 利用するインタコネクトと通信方式により適用方式が異なるが、受信バッファを分散するサーバ毎に分割確保すること基本とする。

送信メッセージ数制限方式： 受信側ノードの持つ受信バッファ数を各サーバ分に分割する。

SendCNT、RecvCNT カウンタノードの分散の方法であるが，集中を緩和するという観点から、ネットワークトポロジ的に分散して配置することが望ましい。例えば，均等に配置するのであれば次のように考えられる。

スイッチトポロジ： 各スイッチの同じレベルの階層毎に

ⓒ 2013 Information Processing Society of Japan

均等配置

直接結合網： 均等な形状で分割した直方体毎に均等配置

このように配置することで、各ノードからの受信バッファ獲得要求をそれぞれの要求ノードの近くで処理することができる。

以上、検討した方式により、遠隔 Atomic 通信を用いることにより、局所集中を緩和できることを示した。しかし、一般には居所的な通信集中の発生パターンと頻度は、アプリケーションにより異なるため、分散配置の方法、各分散ノード毎のカウンタ数の分配手法、中継方式については様々な最適化が可能である。

5. 関連研究

Open MPI や MVAPICH[7] における InfiniBand 実装では信頼性のあるプロトコル (RC) と信頼性のないプロトコル (UD) を採用している。RC の場合は InfiniBand のハードウェアレベルで信頼性を実現しているが、実際に RC プロトコルにおいても InfiniBand の伝送路は信頼性を保証していないので、受信バッファが枯渇した場合、メッセージは廃棄される。このため、再送を行うが，第 3.3 節で述べたように、一定時間受信バッファが供給されないと QP エラーが発生する。UD プロトコルにおいては、MPI レベルで通信の信頼性を確保する必要があるため、信頼性確保のプロトロルで、メッセージ廃棄が発生した場合は再送する必要がある。

受信側主導型 (Receiver Initiated) を用いた RDMA ベースの MPI としては、Cell Broadband Engine 上の実装がある [8]。本実装では Cell Broadband Engine というメモリ量が少ない環境下において送信側主導型 (Sender Initiated) 通信よりも高速であると述べている。しかし、遠隔 Atomic 通信は利用していない点が異なる。

6. まとめ

ポストペタスケール規模での通信ライブラリでは、省メモリ性の実現が必須となる。省メモリ性を実現するためには、少ないバッファでいかに効果的かつ安定した通信を行うかが課題となる。本論文では、ポストペタスケール規模での省メモリ性実現のために、何が問題となりえるのかを明らかにし、遠隔 Atomic 通信を用いた実現方式について、受信バッファ固定方式と受信バッファ数制限方式について述べた。

今後，プロトタイプを実装評価する予定である。

参考文献

[1] 住元真司, 安島雄一郎, 安達知也, 岡本高幸, 秋元秀行 and 三浦健一. 遠隔 Atomic 通信を用いた省メモリ性実現のための方式検討. 情報処理学会研究報告 13-HPC-138(13). 情報処理学会, Feb. 2013.
[2] The Message Passing Interface (MPI) standard: http://www.mpi-forum.org.
[3] MPICH2: http://www.mcs.anl.gov/research/projects/mpich2/.
[4] OpenMPI: http://www.open-mpi.org/.
[5] 三浦健一, 秋元秀行, 安島雄一郎, 岡本高幸, 住元真司. エクサスケールコンピューティングに向けた省メモリ通信ライブラリの検討. 情報処理学会研究報告 12-HPC-133(14). 情報処理学会, Mar. 2012.
[6] 秋元秀行, 三浦健一, 岡本高幸, 安島雄一郎, 住元真司. InfiniBand Atomic Operation の性能評価. 情報処理学会研究報告 12-HPC-133(8). 情報処理学会, Mar. 2012.
[7] MVAPICH: http://mvapich.cse.ohio-state.edu/.
[8] S. Pakin. Receiver-initiated message passing over rdma networks. In *Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on*, pp. 1 –12, april 2008.

ⓒ 2013 Information Processing Society of Japan