Vol.2012-HPC-133 No.8
2012/3/26

InfiniBand Atomic Operation の性能評価

秋元 秀行[†,††]　三浦 健一[†,††]　岡本 高幸[†,††]
安島 雄一郎[†,††]　住元 真司[†,††]

Exa FLOPS 級スーパーコンピュータの実現には低レイテンシ・省メモリを兼ね備える通信プロトコルの開発が不可欠である．我々は本課題の解決手段の一つとして Atomic 通信を用いる通信資源の動的管理手法を検討している．本報告書は，通信手順の最適化や性能予測等に役立てるための基礎データの取得を目的に，InfiniBand の Atomic 通信を中心とした性能評価結果を報告する．メモリの不可分性を保つために自ノードメモリに対しても必要に応じて HCA を経由した Atomic Operation が必要であるが，この場合，自ノードのメモリアクセスにおいてもノード間と同程度の性能しか望めないことが分かった．一方，外乱通信による性能劣化はメモリ書き換えを伴う RDMA Write, Atomic 通信を同時に行うことによって大きくなることが分かった．

Performance Evaluation of InfiniBand Atomic Operation

Hideyuki Akimoto[†,††], Kenichi Miura[†,††],
Takayuki Okamoto[†,††], Yuichiro Ajima[†,††],
and Shinji Sumimoto[†,††]

It is necessary to establish the communication protocol which has both low latency and saving memory to realize the Exa FLOPS supercomputer. We have thought that dynamic resource management by using remote atomic memory access is one of key technologies for the solution. This article reports transportation performance of InfiniBand(IB) including the remote atomic memory operation. The data has taken to be used for optimization and estimation of the communication procedure. We found that the local memory access through IB HCA is expected almost same performance of the inter node remote atomic memory access. Because, the local memory which needs to assure atomicity should be access through the IB HCA. Also, we found that the performance deterioration is depending on combination of transfer functions. The combination of transfer functions which consists writing the memory like RDMA Write and Atomic operation brings large deterioration.

1. はじめに

コンピュータを用いるシミュレーション技術は科学技術のみならず産業界の発展にとっても，もはや不可欠な存在となっている．より高精細・高精度のシミュレーションを行うためには，高速・大容量なコンピューティング環境の継続的な発展が望まれている．こうした要望に応えるコンピューティング環境の最先端技術がスーパーコンピュータ（スパコン）である．2011 年 11 月の Top 500 List における最速スパコンは「京コンピュータ」であり，その速度は 10 Peta FLOPS（: FLoating-point Operations Per Second）強である[1]．一方より高速なコンピューティング環境の開発競争は既に Exa FLOPS 級にターゲットが移りつつある[2, 3]．Exa FLOPS 級のスパコンを現在の消費電力，設置面積を保ったまま実現するためには様々な課題がある．例えばプロセス数は数千万からそれ以上におよぶと予測され，各プロセスは互いの計算データを必要に応じて通信により交換し，協調した計算処理を進める必要がある．既存の通信プロトコルの多くは通信性能や信頼性向上のため，全ての通信相手（プロセス）毎に通信バッファを用意する等, プロセス数に応じたメモリ領域確保が必要なものとなっている．しかしながら Exa FLOPS 級スパコンにおいて，1 プロセスが使用可能なメモリ量は回路の集積度や消費電力の制約から数 GB 程度と推測される[4]．このような事情を考慮すると通信品質・性能を犠牲としない省メモリ型の通信プロトコルの開発が不可欠である．我々は本課題を重要視し，解決するための通信プロトコルの検討を九州大学，九州先端科学技術研究所と共同で JST, CREST のプロジェクトの一つとして開始している. その中で解決手段の一つとして Atomic 通信を用いたリモートノードにおける通信資源の動的管理手法による省メモリ化を検討している[5, 6].

InfiniBand（IB）は高信頼性かつ低レイテンシ・高バンド幅の優れた特長を持つことから，HPC システムにおいて現在最も広く使用されているインターコネクト規格の一つと言える．また IB は省メモリ通信プロトコルの開発において，我々が解決手段の一つと考えている Atomic 通信をサポートしており，低レイテンシ・省メモリ通信プロトコルの検討環境の一つと考えている．本報告書では前述の通信プロトコルの検討における通信手順の最適化や性能予測を行うための基礎データを取得する事を目的に，IB の Atomic 通信を含む各通信の単体性能や各通信が混在する場合の性能差を評価した結果を報告する．

本報告書の構成は第 2 章で IB の Atomic 通信の概要について述べる. 次いで第 3 章，第 4 章で性能評価環境および評価方法・項目について述べる．第 5 章では性能評価結

† 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
†† 独立行政法人科学技術振興機構 戦略的創造研究推進事業
Japan Science and Technology Agency(JST), Core Research for Evolutional Science and Technology(CREST)

ⓒ2012 Information Processing Society of Japan

果を述べ，第 6 章でまとめる．

2. InfiniBand および Atomic 通信の概要

IB の規格は複数ベンダから構成された業界団体である「InfiniBand Trade Association」により作成・提出された統一的なものであり，最新の規格書は 2007 年 11 月に発行された InfiniBand Architecture Specification Volume 1 Release 1.2.1[7]である．一般的に IB を用いた HPC システムではノード間の接続に 4 つの接続を結束した×4 が広く使用されているため，以降は×4 接続について述べる．IB は最初の製品リリース以来徐々にその性能を向上させている．IB の最初の製品である Single Data Rate（SDR）では一方向当たりの実データ転送速度は 1.0 GB/s であった. 以後 Dual Data Rate（DDR）: 2.0 GB/s, Quad Data Rate（QDR）: 4.0 GB/s 等がリリースされており，今後についても更なる性能向上が予定されている．

IB の通信サービスにはいくつかの種類が用意されているが，HPC 分野においては Reliable Connection（RC）が標準的に使用されている．一方, 通信関数には Send, Remote Direct Memory Access（RDMA）Write, RDMA Read, Atomic 等がある．Send 通信は送受信ノードそれぞれで CPU が介在した通信手順であり，送信側ノードは送信命令を，受信ノードは受信命令をそれぞれ発行する必要がある．このため送受信ノード双方で CPU リソースを消費する．一方，RDMA Write, RDMA Read, Atomic 通信は命令の発行ノード（ローカルノード）のみが CPU を介して通信命令を発行し，相手ノード（リモートノード）では CPU が介在することなく通信処理がなされる．

次に IB の Atomic 通信の仕様について述べる．IB では Atomic 通信のリモートノードにおける Operation として“Fetch and Add”と“Compare and Swap”が使用可能である．Atomic Operation の対象メモリサイズは 8 B に限定され, かつ 8 B アラインされている必要がある．Atomic 通信を使用できる IB の通信サービスは先に述べた RC と Reliable Datagram（RD）であるが，現時点で RD 通信サービスは使用できないため Atomic 通信を使用するためには, 実質的に RC 通信サービスを使用する必要がある. IB の Atomic 通信における“read, modify and write”の不可分性は，同一の HCA 内のみで保証され“read”と“write”間の処理が同時に行われないことを保証している．これは Atomic 通信として仕様上で規定されている最低限の不可分性で，将来的に拡張される可能性を持っている．Atomic 通信はリクエスト情報を含む“Atomic Command”パケットとレスポンス情報を含む“Atomic Acknowledge”パケットの 2 つから成り立っている．Atomic 通信命令発行元はレスポンスパケットの到着をもって，リモートノードにおいて Atomic Operation が完了したと見なすことができる．RDMA Read 要求と Atomic 通信リクエストを連続して発行した場合，リモートノードにおいて RDMA Read のメモリリード要求と Atomic 通信の順序性は保証されない．一方，リモートノードに連続してリクエストが到着した場合には，直前の RDMA Read を除くメモリアクセスと直後のリクエストによるメモリアクセスの間に Atomic Operation は実行される．

3. 性能評価環境

全ての測定は同一環境の計算機 2 台それぞれに同一の IB Host Channel Adapter（HCA）1 枚を接続して実施した．IB HCA に対する性能比較として Table 1 に示す Mellanox 社製の 3 種類を用いた. 3 種の HCA は全て PCI Express によって Host 計算機と接続するタイプで，設計世代・転送レートが異なる．本報告書における主な測定・比較対象は Connect X1 世代の DDR および QDR とする．また，一部については設計世代の異なる Connect X2 QDR についても測定・比較を行う．全ての HCA は PCI Express 8 レーンの同一 Slot に交換・接続した．Host/HCA 間の PCI Express（PCIe）の転送速度は，QDR HCA については設計世代によらず 5.0 GT/s ×8 width（PCIe Generation 2）であるのに対し，Connect X1 DDR HCA については HCA のサポート範囲の関係上 2.5 GT/s ×8 width（PCIe Generation 1）である．2 ノード間の接続は DDR，QDR 共に 3m 長の銅ケーブルにて Port 1 同士を直結した．

Host 計算機は Intel 社製 Z68 Express Chipset と Core i7-2600K（3.4 GHz），DDR3-1333MHz 8 GB（4 GB × 2）を用いた．同計算機に OS として CentOS 6.2（Kernel 2.6.32-220.el6.x86_64 #1 SMP）をインストールし測定に使用した．

Table 1 Measured InfiniBand HCA and Specifications

Generation Data Rate	Part. Number	Host Bus(PCIe Gen.)	Firmware
Connect X1 DDR (2.0 GB/s)	MHGH28-XTC	2.5GT/s ×8 width(1)	2.9.1000
Connect X1 QDR (4.0 GB/s)	MHQH29-XTC	5.0GT/s ×8 width(2)	2.9.1000
Connect X2 QDR (4.0 GB/s)	MHQH29B-XTR	5.0GT/s ×8 width(2)	2.9.1000

4. 性能評価方法および項目

IB の性能評価には OFED-1.5.4 に含まれる perftest-1.3.0 内の各測定プログラムを使用して行った. 測定プログラムは基本的に 2 ノード間の IB 通信性能を測定するプログラムで，通信バンド幅測定（ib_*TransFunc*_bw）とレイテンシ測定（ib_*TransFunc*_lat）の 2 種類に大別される. 更に通信関数毎に Send, RDMA Write, RDMA Read, Atomic の 4 種類を用いた．各通信関数の測定プログラムは前記の ib_*TransFunc*_bw, ib_*TransFunc*_lat の *TransFunc* 部分を send, write, read, atomic に置き換えたものである．Atomic 通信においてはオプションによりリモートノードの Operation として“Fetch and Add”と“Compare and Swap”の 2 種類を選択・評価可能であり両者について測定した．

ⓒ2012 Information Processing Society of Japan

評価に当たっては測定毎の誤差を極力抑制するため numactl コマンドを使用して，測定に使用する CPU-core, Memory を固定して測定した. 合わせて各測定のイタレーション回数をデフォルトの 1000 から通信バンド幅測定については 5000 に，通信レイテンシ測定については 2000 に変更して行った.

5. 通信性能評価結果

5.1 基本性能測定結果（通信外乱が無い場合）

Figure 1 に 2 ノード間の IB の Send, RDMA Write, RDMA Read の通信バンド幅測定結果を示す．図の横軸は各通信関数におけるメッセージサイズを，縦軸は通信バンド幅である．長メッセージサイズにおける通信バンド幅は通信関数によらず Connect X1 DDR では 1590 MB/s 程度，Connect X1/X2 QDR では 3060 MB/s 程度が得られている．これらの値は IB DDR, QDR の仕様上の通信バンド幅である 2.0, 4.0 GB/s と同様に約 2 倍の性能差が確認された．また，各絶対値は PCI Express の実質的な転送速度およびパケット構成を考慮[8]すると妥当な数値と判断している．一方，小メッセージサイズにおける通信バンド幅の立ち上がりは通信関数によって異なり，Send および RDMA Write 関数が早く RDMA Read 関数が若干遅い.

Figure 2 に IB Atomic 通信関数を含んだ(a): 通信バンド幅および(b): レイテンシの性能評価結果を示す．図の左半分は異なる 2 ノード間の通信における評価結果であり，右半分は同一ノード内のループバック通信における評価結果である．評価した各通信のメッセージサイズはいずれも 8 B である．これは IB がサポートする Atomic 通信関数の対象メッセージサイズが 8 B に限定されるためであり，他の通信関数についても同一のメッセージサイズにおける性能を図示した．Atomic 通信は2 章で述べたとおりリクエストである“Atomic Command” パケットをリモートノードに送信し，リモートノードにおいて Atomic なメモリアクセスオペレーションを実行後，その結果として“Atomic Acknowledge”パケットをローカルノードにレスポンスとして返す．このため通信動作的には RDMA Read の近く，計測プログラムも類似した作りとなっている．このような観点から両者の性能を比較すると，同一の HCA における Atomic 通信の通信バンド幅は RDMA Read の約半分に劣化している．一方，Atomic 通信のレイテンシは RDMA Read に比較して 200～300 ns の増加が認められた．Atomic 通信の Operation には“Fetch and Add”, “Compare and Swap”の 2 種をサポートしているが，Operation 種別による有意な性能差は確認できなかった. HCA の種類により長メッセージサイズの通信バンド幅には規格上の性能差とほぼ同等の差が見られたが，8 B メッセージサイズにおいては Send や RDMA Write を含め DDR に対する QDR の性能優位性は認められるものの，差分は 30%前後であった．これは 8 B メッセージの送受信では，メッセージサイズに非依存な処理，例えば送受信のためのハードウェア・ソフトウェアオーバーヘッドの占める割合が大きいためと考えられる.

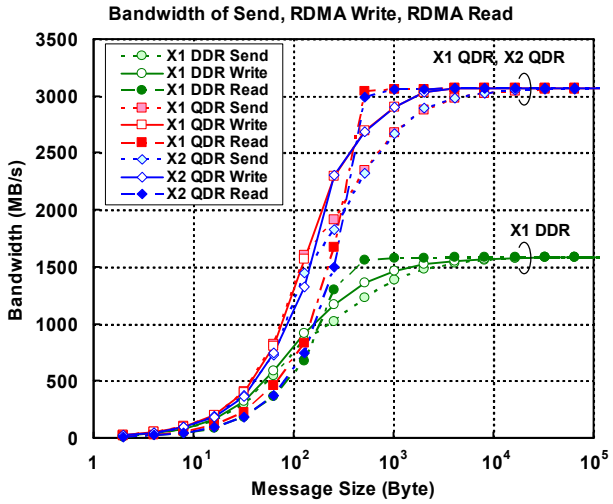
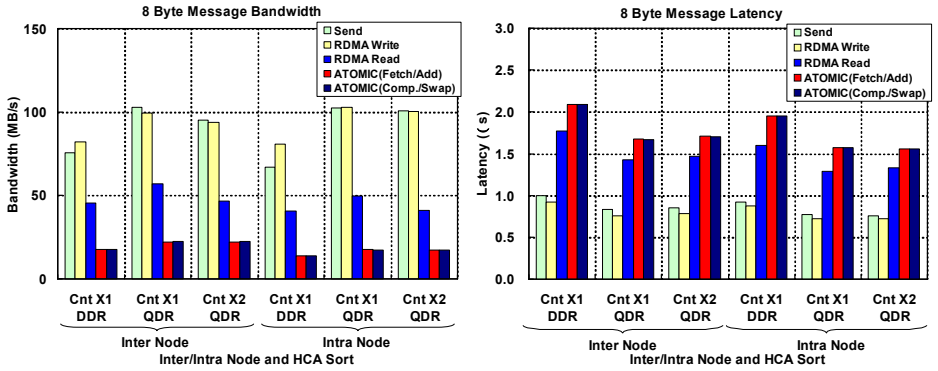


Figure 1 The bandwidth of Send, RDMA Write, and RDMA Read transportation as a function of message size



(a) Bandwidth (b) Latency

Figure 2 8B message bandwidth and Latency of transportation functions of Connect X1 DDR, X1 QDR, and X2 QDR. The right half is the communication performances of the inter node communication and the left half is that of intra node communication.

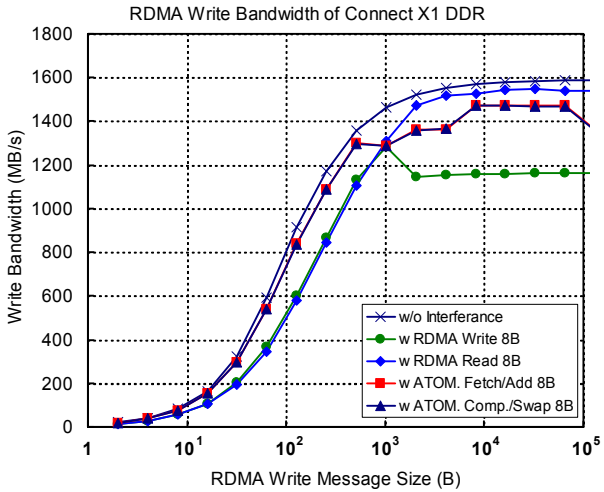
ⓒ2012 Information Processing Society of Japan

次に 2 ノード間の通信（図の左半分）と同一ノード間の通信（図の右半分）を比較する．IB の Atomic 通信の不可分性の保証範囲は2 章で述べたとおり同一 HCA によるメモリアクセスに限定される．従って，たとえ自ノードのメモリ操作であっても不可分性を確保した更新が必要な場合には HCA を経由する必要がある．HCA を経由した自ノードのメモリアクセス性能はノード間のアクセスとほぼ同等であった．IB の Atomic 通信を使用したプロトコルの開発を行う上では本特性を理解した上で使用する必要がある．なお，ノード内の測定においてはローカル，リモートプロセスの CPU リソースはそれぞれ core 3, core 2 を用いて測定を実施した．一方，ノード間の通信性能測定においては共に core 3 を使用した．

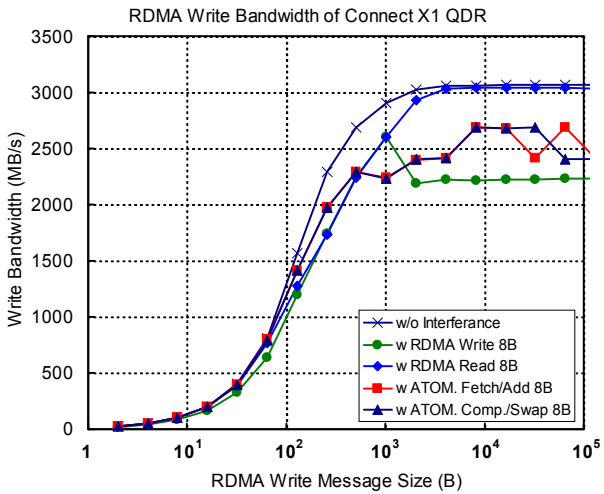
5.2 通信外乱が有る場合の性能評価結果

Figure 3 に外乱通信が有る場合の RDAM Write 通信の送信メッセージサイズ依存性を示す．本測定は 2 ノードを使用して実施し，測定対象通信と外乱通信はそれぞれ同一の HCA，同一 port を使用して行った．すなわち，性能測定対象通信と外乱通信は同一の通信資源を共有している．外乱通信は 8 B メッセージサイズのバンド幅測定を繰り返し行う事で与え，外乱通信が行われている最中に対象の通信性能測定を行った．この時，測定対象・外乱通信を行う各プロセスを core 3 および core 2 に固定することにより，CPU 資源については干渉を抑制して測定を実施した．但し，メモリバスについては共有使用されているが，測定に用いたシステムのメモリバンド幅は 20 GB/s 程度であり性能に与える影響は小さいと考えている．Figure 3 (a)は Connect X1 DDR，(b)は同 QDR における評価結果である．ここで，両者の縦軸のスケールは約 2 倍異なっている点に注意されたい．各グラフには外乱通信が無い場合および外乱として 8 B メッセージの RDMA Write, RDMA Read, Atomic 通信が有る場合の性能を示した．外乱通信が存在することにより通信性能の劣化が認められる．性能劣化の幅は外乱通信の種類によって異なり，RDMA Read が最も小さく，次いで Atomic 通信，RDMA Write であった．Atomic 通信においては Operation の種類によらず，性能劣化の程度は同等であった．これらの傾向は Connect X1 DDR, QDR で共通して見られた．図示していないが外乱通信が RDMA Write, RDMA Read とした場合には，メッセージサイズを 8 B に加え 8 KB, 8 MB についても測定を行っている．長メッセージサイズの RDMA Write を外乱として与えた場合には RDMA Write の通信性能は 50%に劣化する．これは測定対象，外乱の通信主データが共にローカルノードからリモートノードに対して送られ，通信経路の競合が生じているためである．一方，外乱通信を RDMA Read とした場合には最大でも 10%程度の性能劣化であった．外乱通信が RDAM Read の場合にはリクエストパケットはローカルノードからリモートノードへの通信であるため測定対象の通信と競合しているが，主データはリモートノードからローカルノードへの通信であり，測定対象，外乱通信で逆方向であり通信経路の競合が生じないためと考える．

Figure 4 は外乱通信が有る場合の RDMA Read 通信の測定結果で，(a)は Connect X1



(a) RDMA Write Bandwidth of X1 DDR



(b) RDMA Write Bandwidth of X1 QDR

Figure 3 The bandwidth of RDMA write transportation as a function of message size with 8B various interference communication.

ⓒ2012 Information Processing Society of Japan

DDR, (b)は同 QDR である．測定方法については外乱通信が有る場合の RDMA Write 通信評価と同様である．外乱通信がある場合に性能が劣化する点は RDMA Write と同様であり，その傾向もほぼ同様であった．また，図示していないが外乱通信を RDMA Write, RDMA Read とした場合には，メッセージサイズ 8 B に加え 8 KB, 8 MB についても測定を行った．結果は外乱通信が有る場合の RDMA Write 通信の性能とは逆に外乱通信として長メッセージサイズの RDMA Read 通信が有る場合に通信性能が約 50%に劣化した．この性能差は主データの通信経路の競合の有無によってやはり説明できる．

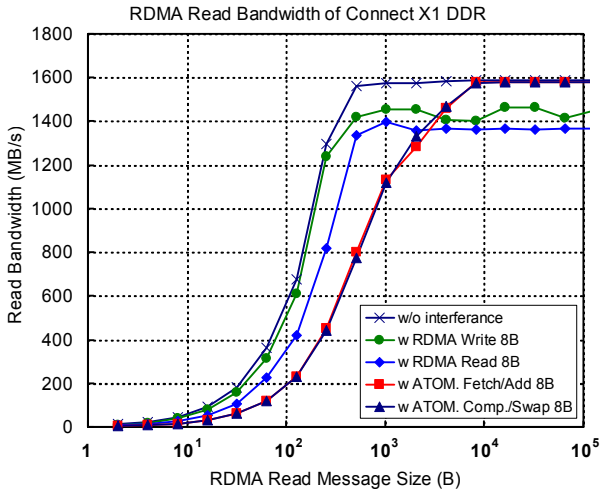
HAC として Connect X1 DDR の RDMA Write: Figure 3 (a)と RDMA Read: Figure 4(a)を詳細に比較すると，外乱通信が Atomic 通信である場合の性能劣化幅が異なり，RDMA Write に与える性能インパクトがより大きいことが分かる．これは Atomic 通信におけるリモートノードでのメモリアクセスの順序性の仕様によるものと考えている．基本的に Atomic 通信は，直前・直後のメモリアクセスの間に“read, modify and write”メモリアクセスを行う仕様で有るが，直前の RDMA Read 要求との間には順序性を保証しない仕様で有る．このため，外乱通信が RDMA Write, RDMA Read である場合を比較すると，前者において性能劣化が大きくなっているものと考える．なお，この傾向は Connect X1 QDR（RDMA Write: Figure 3 (b)と RDMA Read: Figure 4(b)の比較）についても同様であった．

Table 2，Table 3 は外乱通信の有無による 8 B Atomic “Fetch and Add”通信の通信バンド幅およびレイテンシ性能の評価結果である．性能評価対象および外乱通信のメッセージサイズは共に 8 B 固定である．外乱通信がリモートノードの書き込み動作を伴わない RDMA Read の場合に性能劣化の幅が最も小さく，次いで RDMA Write, Atomic 通信であることが分かった．なお，測定対象を Atomic “Compare and Swap”としても傾向は同様であった．また，HCA を Connect X1 QDR とした場合にも概ね同様の結果が得られた．

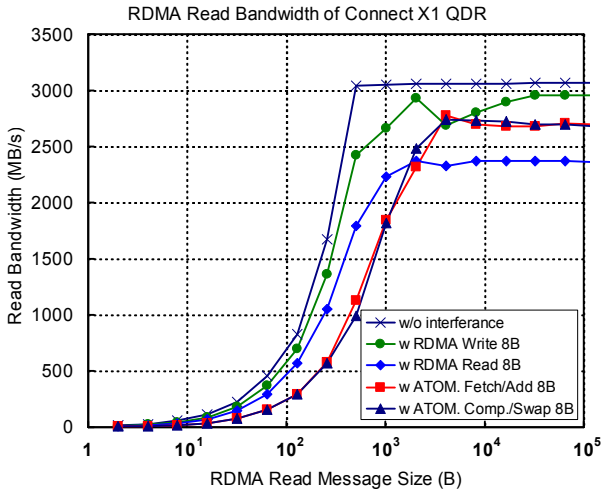
6. まとめ

Exa FLOPS 級のスーパーコンピュータの実現に向けて解決すべき課題の一つである低レイテンシ・省メモリを兼ね備える通信プロトコル検討の一部として，InfiniBand Atomic 通信を中心とした性能評価を実施し，以下の知見を得た．

1. 外乱通信のない状態において，Atomic 通信性能は RDMA Read 性能の約 1/2 程度の通信バンド幅，約 200～300 ns 程度のレイテンシ劣化が認められた．
2. メモリ内容の不可分性を確保するためには自ノードのメモリアクセスにも HCA を介した Atomic Operation の必要があるが，ノード内メモリのアクセスにおいても通信バンド幅，レイテンシ共にノード間と同程度の性能しか望めない．



(a) RDMA Read Bandwidth of X1 DDR



(b) RDMA Read Bandwidth of X1 QDR

Figure 4 The bandwidth of RDMA Read transportation as a function of message size with 8B various interference communication.

ⓒ2012 Information Processing Society of Japan

Table 2 The bandwidth of atomic “Fetch and Add” transportation with/without 8 B message interference transfer in the case of Connect X1 DDR and QDR.

HCA	without Interference	with RDMA Write	with RDMA Read	with Atomic F/A	with Atomic C/S
X1 DDR	17.6 MB/s	12.9 MB/s	17.1 MB/s	8.8 MB/s	8.7 MB/s
X1 QDR	22.4 MB/s	16.3 MB/s	19.2 MB/s	11.1 MB/s	11.0 MB/s

Table 3 The Latency of atomic “Fetch and Add” transportation with/without 8 B message interference transfer in the case of Connect X1 DDR and QDR.

HCA	without Interference	with RDMA Write	with RDMA Read	with Atomic F/A	with Atomic C/S
X1 DDR	2.08 μs	2.33 μs	2.12 μs	8.20 μs	8.20 μs
X1 QDR	1.69 μs	1.74 μs	2.66 μs	6.43 μs	6.43 μs

3. 外乱通信が有る場合の各種通信は性能劣化するが，測定対象・外乱通信の組合せによりその劣化幅が異なる．特にリモートノードのメモリ内容を書き換える RDMA Write, Atomic 通信が同時に行われる場合に性能劣化が大きい．
4. Atomic Operation の不可分性の保証範囲, 順序性についての考慮も不可欠で有る．

今後，通信プロトコルの開発において通信手順の最適化や性能予測等に役立てる予定である．

参考文献

1) Super Computer TOP500, *http://www.top500.org/*
2) P. Kogge et al.: Exascale computing study: technology challenges in achieving exascale systems, *DARPA Information Processing Technologies Office sponsored study* (2008).
3) J. Torrellas: Architectures for extreme-scale computing, *IEEE Computer*, 42(11), pp. 28-35, (2009).
4) J. Dongarra et al.: The international exascale software project roadmap, *The international journal of high performance computing applications*, 25(1) pp. 3-60 (2011).
5) 安島 雄一郎, 秋元 秀行, 岡本 高幸, 三浦 健一, 住元 真司: 片側通信による，グローバルデータ構造の効率的な操作方法の検討, to be published in *情報処理研究報告*, 2012-HPC-133 (2012).
6) 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, to be published in *情報処理研究報告,* 2012-HPC-133 (2012).
7) InfiniBand Trade Association: InfiniBand Architecture Specification, Volume 1, Release 1.2.1, (2007).
8) 松葉 浩也, 石川 裕: コモディティネットワークによる 5GB/s 通信の可能性, *情報処理研究報告*, 2007-ARC-172, pp. 169-174 (2007).

ⓒ2012 Information Processing Society of Japan