Vol.2012-HPC-133 No.14
2012/3/26

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

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

エクサスケールスパコンの実現のためには通信ライブラリの省メモリ化が重要な課題である．本報告書では既存システムにおいてバッファ数が十分に確保できない場合の性能の調査として，InfiniBand 上で UD プロトコルを用いた通信でバッファ枯渇を前提に性能評価を実施した．また RC プロトコル使用時に受信バッファとして SRQ を使用した場合の通信性能と比較を行った. その結果再送間隔を効果的に行うことで SRQ の結果より優れた性能が得られることが確認でき, 再送制御の重要性が確認できた.

The investigation of communication library to reduce memory towards exascale computing

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

Reducing memory is a most significant problem to realize exascale computer. This report shows that performance evaluation of a protocol using UD (Unreliable Datagram) on the InfiniBand as the investigation of communication performance under insufficient receive buffer on the existing system. It is compared with the communication performance of a protocol using RC (Reliable connection) with SRQ (Shared Receive Queue) as receive buffer. We obtain good performance according to efficient interval of retransmission, and confirm importance of retransmission control.

1. はじめに

エクサフロップス級のスパコンの実現に向けて各国で検討がスタートしている．アメリカにおいては DARPA (Defense Advanced Research Projects Agency)が UHPC (Ubiquitous High Performance Computing)プロジェクトをすでに開始しており，2018 年にエクサフロップス級のスパコン実現を目指してプロジェクトが進行している[1].

エクサフロップス級のスパコンを実現するために現在の 100 倍程度，数百万～数千万程度のノード数が必要であるが，その一方で消費電力や設置面積の問題から，ノードあたりのメモリ搭載量は現在の 10 倍程度にしかならない見通しである．

一方で，並列プログラムで通信を行うためのインターフェイスとして一般に MPI が使用されている．しかし，MPI ライブラリは通信相手毎に通信用バッファや管理情報を確保することで高速通信を実現しているため，並列数が増えれば増えるほど使用メモリが増大するという問題がある．数百万～数千万規模の並列プログラムに対応するためには通信バッファを含む使用メモリ量の大幅な削減が必要であり通信性能に与える影響が非常に大きくなることが課題である．

我々は，本課題の解決を目的として，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」において,「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」の研究課題に取り組んでいる．本研究課題では，ポストペタスケール時代の省メモリ性実現について，新しいハードウェア機能，通信プロトコルなどさまざまな観点から取り組むこととしている．

具体的には，エクサフロップス級のスパコンをターゲットにした省メモリバッファ管理方法としてリモート Atomic 通信を用いたリモートメモリ操作による方式と制御メッセージを含めたデータ通信処理の最適化などを考えており，今後，実装・評価を行っていく予定である．

本研究報告では，省メモリ性を確保しながら，通信性能劣化を最小限にするための省メモリ低遅延通信プロトコルの実現について述べ，基礎データとして既存の InfiniBand 上での省メモリ化実装の評価結果について述べる. 2 章において既存の MPI ライブラリのメモリ量調査結果を踏まえてエクサフロップ数のスパコンにおける省メモリ化の検討の必要性を説明する．次に 3 章で省メモリ・低遅延通信プロトコルの実現方針，4 章で省メモリ化の検討について述べた後，5 章ではバッファ枯渇時の振る舞いの知見を得ることを目的として InfiniBand 上で UD(Unreliable Datagram)プロトコルを用いた通信(以下 UD 通信)のバッファ枯渇時の性能の調査を行った．ここでは

[†]富士通株式会社，独立行政法人科学技術振興機構,CREST
Fujitsu Limited, Japan Science and Technology Agency, CREST

InfiniBand の省メモリ化の方式として一般的に用いられている RC(Reliable Connection)プロトコルで SRQ(Shared Receive Queue)を用いた通信(以下 RC-SRQ 通信)との比較において妥当性の検証を行った．6 章で関連研究の説明を行い，7 章でまとめを行う．

2. Open MPI で使用するメモリ量

エクサフロップス規模の並列数で既存の Open MPI を使用した場合のメモリ使用量を見積もるため，全対全通信を行った後のプロセスのメモリ使用量と並列数の関係を調べ，数百万～数千万並列まで外挿してエクサフロップス規模での MPI ライブラリメモリ使用量の見積もりを行った．通信プロトコルとして RC-SRQ 通信，UD 通信を使用した通信の 2 種類について調査した．並列数とメモリ使用量の関係を図 1 に示す．

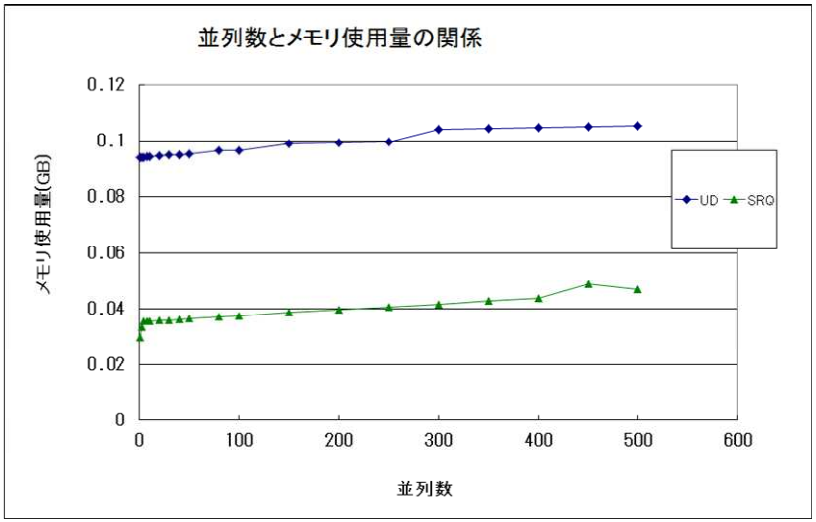

図 1. 並列数とメモリ使用量の関係

RC-SRQ 通信及び UD 通信を使用した通信のメモリ使用量は数百～数千並列程度ではメモリ使用量が小さくさほど問題にならないが，100 万並列で 25GB，1,000 万並列では 254GB に達してしまうためエクサフロップス規模でのメモリ削減は必須である(表 1).

表 1. エクサフロップス規模の並列数でのメモリ使用量見積もり（単位：GB）

並列数	UD 通信	RC-SRQ 通信
100,000	2.63	2.64
1,000,000	25.5	26.1
10,000,000	254	261
100,000,000	2,539	2,606

3. 省メモリ・低遅延通信プロトコルの実現方針

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）１）と２）について RDMA と遠隔 Atomic 操作による通信の活用による効果的，効率的な通信方式の実現

この方針に基づき，２つのステップで研究開発を進める．

1. メッセージパッシング通信（MPI）の改良
2. PGAS 言語向け通信の開発

4. エクサスケールに向けた省メモリ化の検討

２章にあげた，実現方針に基づき，エクサスケールに向けた省メモリ化通信の検討を第１ステップとしてメッセージパッシング通信（MPI）を対象に制御用メモリとデータ通信処理の省メモリ化について検討する．

4.1 制御情報の省メモリ化検討

前章でエクサスケールシステムにおいては，通信相手先の制御情報についても省メモリ化が必要であると述べた．例えば，既存の MPI 通信ライブラリでは，相手先の情報すべてを配列で格納しているが，相手先が百万ノードになると，相手先あたり１バイト必要だとしても，1MB の配列が必要になり，1KB 必要ならば，1GB のメモリが必要である．これは，すべての相手先ノードの制御情報をそもそもノードローカルに持つことができなくなることを意味している．

したがって，制御情報は，必要時に相手先ノードより動的に確保することが前提となる．この動的に確保する方式についても，データ配列を利用するのではなく，アクセス先が一意に決まる方式が必要になる．

制御情報を参照，制御する手段として我々は，RDMA と遠隔 Atomic 操作による通信の活用を検討している．これらの活用により，例えば，分散データ構造化し通信ライブラリに必要な管理構造を重複なく，かつ，相手先ノードのプロセッサの同期なく実現できるのではないかと考えている．

この基礎評価として，IB の Atomic 通信性能について，秋元[2]，ならびに安島[3]で報告する．

4.2 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

データ通信に必要なメモリ領域としては，送受信のためのバッファと信頼性を確保する場合には，信頼性を確保するための制御領域（メッセージの送受信を管理する sequence 番号管理情報など）が必要である．

MPI などメッセージパッシング通信においては，信頼性を確保した通信を前提にしている．しかし，エクサスケールシステムの各ノードで十分なメモリが確保できない状況においては，これまで予期できないことが発生しえる．

ひとつは，これまで，データの送受信時だけに発生していたバッファオーバフローの問題が制御通信（ack パケット，ランデブー制御パケット）の送受信においても，局所的に集中する場合に発生しうる．例えば，百万ノードのシステムにおいて，制御通信が一斉に特定の一つのノードに集中した場合，少なくともこれらの制御通信を受けるだけの受信 バッファが必要になる．これは，例えば 128 バイトのパケットであっても，128MB の制御用のためだけに受信バッファが必要になる．また，あふれた場合の再送方式を考えると，単純なタイムアウトによる再送だけだと，相手先の混み具合により再送パケット数は何倍にも膨れ上がり，ネットワーク網の輻輳が発生し性能劣化は避けられない．このため，制御メッセージを含めたデータ通信処理の最適化が必須である．

このように大規模並列処理において受信バッファ枯渇に伴う再送は避けられない問題であり，再送に伴う性能劣化を如何に抑えることができるかが重要な課題である．

そこで本報告では，まずはバッファ枯渇時の知見を得ることを目的として InfiniBand 上で UD 通信を用いて受信バッファ枯渇が発生した場合の性能検証を行う．またここでは比較データとして RC-SRQ 通信での性能との比較を行うことで性能面からの妥当性の検証を行う．

5. InfiniBand 上での性能評価

5.1 評価環境

性能評価に使用した環境は以下の通り．

CPU	AMD Opteron™ 8354　2.2GHz
インタコネクト	ConnectX DDR

5.2 評価方法

ここでは 31 ノードから 1 ノードに負荷を集中させた場合の評価に関して，RC-SRQ 通信と UD 通信の 2 通りの方式で検証を実施した．

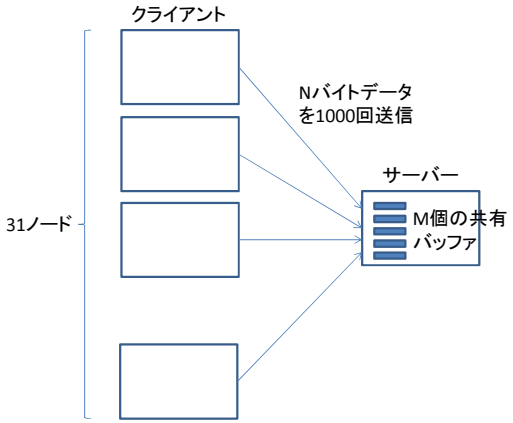

図 2. 性能評価環境

また今回の検証に使用した RC-SRQ 通信，UD 通信の通信方式は以下の通りである．

<RC-SRQ 通信>

1. クライアントからデータを転送
2. サーバは完了通知のポーリングによりデータ到着を確認
3. クライアントは送信完了通知後，次のデータを送信

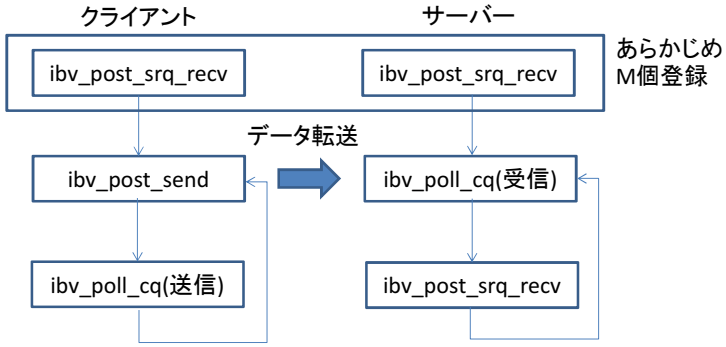

図 3. RC-SRQ 通信を用いた通信方式

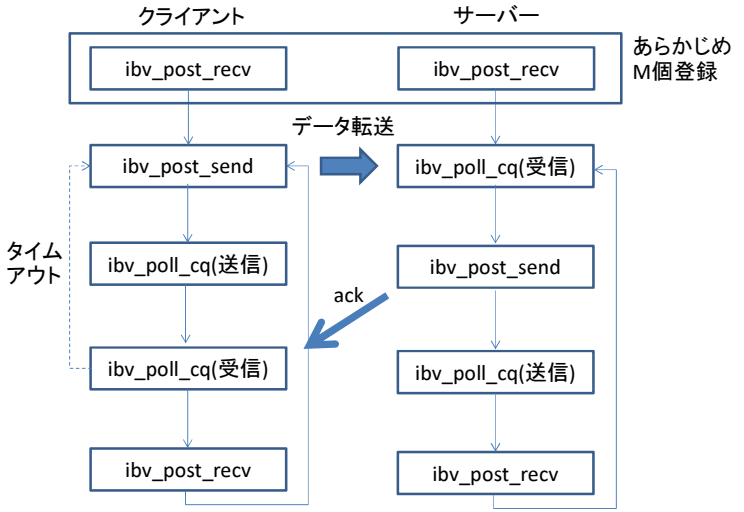

図 4. UD 通信を用いた通信方式

<UD 通信>

1. クライアントからデータを転送
2. サーバは完了通知のポーリングによりデータ到着を確認
3. サーバは完了通知を確認したらクライアントに ack を返す
4. クライアントはデータ転送後にサーバからの ack を待つ
5. 一定時間サーバからの ack が返ってこなければ再送を実施
6. ack 受信後，次のデータを送信

5.3 評価結果

RC-SRQ 通信と UD 通信の 2 つのプロトコルに関して，まずは受信バッファ数が 10, 1000, 10000 の場合の 3 通りで性能測定を実施した．結果を図 5 に示す．受信バッファ数を 100 で性能測定を実施すると RC-SRQ 通信で Retry error が発生するため今回は断念してそれ以外のバッファ数で評価を行った．結果を見ると RC-SRQ 通信を用いた方が UD 通信を用いる場合より 2 倍程度性能が良い事が分かる．これは UD 通信の方が ack 返信のためのソフトウェアオーバーヘッドが必要であるためであり，ほぼ妥当な結果と考えられる．また準備する受信バッファの数が少ない方が性能が良くなっているが，これは InfiniBand ハードウェアの受信バッファ検索時間等が影響しているのではないかと考えている．

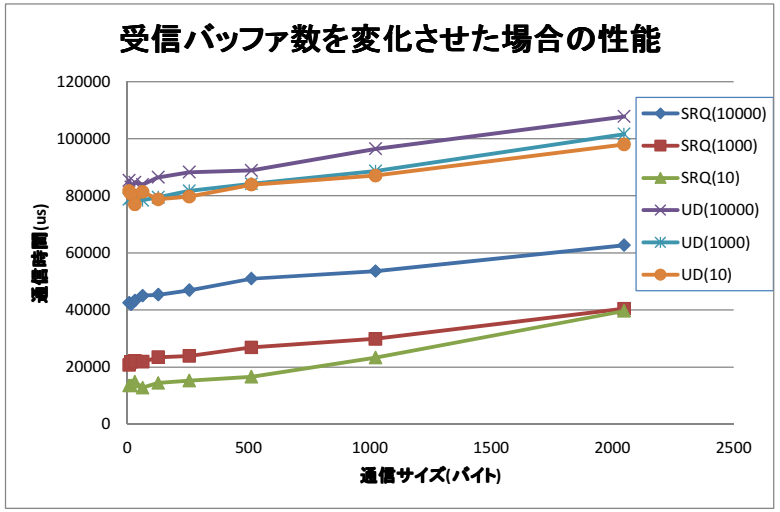

図 5. 受信バッファ数を変化させた場合の通信性能

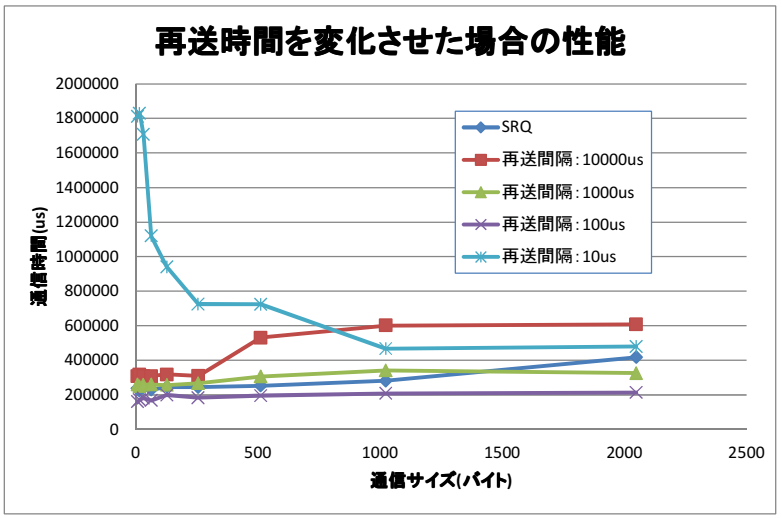

図 6. 再送時間を変化させた場合の通信性能

次に受信バッファ数を 1 に固定して UD 通信の再送のタイミングを 1us～10,000us まで変化させた場合の性能測定結果を図 6 に示す. 受信バッファ数を 1 にした場合に RC-SRQ 通信では受信バッファ数が 10 以上の場合と比較して大幅に性能劣化している. また受信バッファ数が 10 以上の場合に見られた UD 通信との性能差も見られず, むしろ再送時間如何では UD 通信の方が性能が良くなる結果になっている. この時, 内部的にどのようなことが起こっているのか検討を行うための材料として送信側における再送回数(全ランクの合計), 受信側におけるパケット破棄回数の調査を行った. 結果を表 2. に示す. 100~10,000us では送信側の再送回数は増加しているが受信側でパケット破棄がないことから受信側で重複したパケットを受信していないことがわかる. これらの中で 100us の性能が最も良いのは再送タイミングの減少に比べ再送回数の増加が緩いことが原因で, 再送タイミングを小さくすることで受信側での処理効率が改善しているのが原因であると考えている. しかしながら, 10us では受信側でのパケット破棄数が増えていることから, 受信側で受信しているにも関わらず再送が発生したため受信側で二重にパケットを受信することにより性能低下が発生しているものと考えられる.

以上の結果より理想的な通信を行うためには重複受信が発生しない範囲で再送タイミングを減らしていくことでより低遅延な通信になるものと考える.

表 2. 送信側における再送回数, 受信側におけるパケット破棄回数

	10,000us		1,000us		100us		10us	
	送信	受信	送信	受信	送信	受信	送信	受信
8	473	0	6,924	0	39,481	0	2,934,371	385,963
16	494	0	6,765	0	42,892	0	2,882,179	437,540
32	472	0	7,023	1	46,423	0	2,704,294	354,865
64	465	0	7,107	0	40,661	0	1,753,431	222,431
128	506	0	6,792	0	54,140	0	1,369,663	195,905
256	465	0	7,285	0	47,696	0	1,049,302	143,753
512	1003	0	8,381	0	53,344	0	909,866	122,541
1,024	1202	0	8,617	0	55,945	0	514,612	66,757
2,048	1277	0	8,534	0	57,728	0	344,231	60,286

6. 関連研究

通信ライブラリにおけるメモリ使用量の削減については MPI を中心として複数の

研究が行われている．RC 上で SRQ を使用して，通常の RQ 使用時よりもメモリ使用量を削減する試みは最も初期のアプローチである[4][5]．しかし，この方法では RC を使用しているため通信相手毎に QP が必要である．エクサスケールで必要となる数百万規模のノード数では QP 毎に消費される数十 KB のメモリでさえも数十 GB に相当する．Flat MPI での QP 数を削減については，一部の HCA で採用されている XRC を利用することができる[6]がノード数の増加には対応することができない．また，QP ひとつあたりのメモリ使用量を減らす試みもある[7]．根本的に QP の数を減らす提案としては，コネクションが不要な UD を使用する方法[8]や動的にコネクションの確立を行う方法がある．

本研究では IB に限らず，エクサスケール環境で必要となるであろうインターコネクトの通信ライブラリの開発を目指すが，初期の評価環境として IB を想定している．このため，本報告では IB を用いた RC 及び UD でのパケット再送の評価を行った．

7. まとめ

本研究報告では，省メモリ性を確保しながら，通信性能劣化を最小限にするための省メモリ低遅延通信プロトコルの検討について述べ，バッファ枯渇時の性能に関して UD 通信を用いて実験を行った．その結果，受信側でパケットの重複受信が起きない範囲で再送間隔を縮めることで最も効率よく転送を行うことができることが確認できた．

大規模並列においては，どのような通信方式を用いても負荷集中は避けられない問題であり，今後はより効率的な再送方式の検討を行っていく必要があると考えている．

参考文献

1) Ubiquitous High Performance Computing (UHPC), https://www.fbo.gov/download/3d7/3d7cfa30b2b1a93332a444047dea52d8/UHPC_BAA_final_3-2-10_post_(2).pdf
2) 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，第 133 回 HPC 研究会，2012 年 3 月．
3) 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，第 133 回 HPC 研究会，2012 年 3 月．
4) Shipman, G.M., Woodall, T.S., Graham, R.L., et al.: Infiniband scalability in Open MPI, IPDPS 2006, pp.10, (2006).
5) Sur, S., Chai, L., Jin, H.W., et al.: Shared receive queue based scalable MPI design for InfiniBand clusters, IPDPS 2006. 20th International, pp.10, (2006).
6) Koop, M.J., Sridhar, J.K. and Panda, D.K.: Scalable MPI design over InfiniBand using eXtended Reliable Connection, Cluster 2008, pp.203-212, (2008).
7) Koop, M.J., Jones, T. and Panda, D.K.: Reducing connection memory requirements of mpi for infiniband clusters: A message coalescing approach, CCGRID 2007, pp.495–504, (2007).
8) Koop, M.J., Sur, S., Gao, Q., et al.: High performance MPI design using unreliable datagram for ultra-scale InfiniBand clusters, ISC2007, pp.180–189, (2007).