6J-06

RDMAを用いたリモートプロセスに対するメモリアクセス手法

永木 謙吾† 渡邉 和樹†† 南 豪介†† 鷲尾 元太郎†† 毛利 公一†
†立命館大学情報理工学部 ††三菱電機株式会社 情報技術総合研究所

1 はじめに

計算機において，主記憶を外部から観測し，プログラムやシステムの挙動を知ることで，信頼性向上などを実現する技術がある．具体的には，主記憶のダンプを障害時の原因究明に利用するライブフォレンジックや，動作しているプロセスのバイナリを解析して挙動を明らかにする動的バイナリ計装 (DBI)，動作しているプロセスのバイナリを書き換えるホットパッチングなどがある．本論文では，これらにネットワークを介して主記憶へのアクセスを実現する Remote Direct Memory Access (RDMA) を適用し，遠隔からの実現 [1] を検討している．そのためには，次のことが必要となる．

1. カーネルやプロセスのメモリ空間へアクセスできること
2. 取得したメモリの情報をもとに，仮想アドレス空間の推定ができること
3. メモリアクセスがアトミックに行われ，情報が一貫していること

以上の背景から，RDMA を実現するプロトコルの 1 つであり，高帯域幅での通信が可能で，広範なメモリ空間へ短時間でアクセスができる RDMA over Converged Ethernet (RoCE) [2] に着目し，上記の実現を目指す．

本論文では，特に遠隔のカーネルやプロセスのメモリ空間へアクセスし，RoCE の標準的なライブラリである libibverbs と Linux カーネルの RDMA ドライバを拡張することで，RDMA を用いたランダムアクセスを実現した．具体的には，RDMA ドライバに含まれる，アクセス可能な物理アドレス空間を登録する関数をユーザから呼び出すためのライブラリ関数とインタフェースを，libibverbs と RDMA ドライバに追加した．

2 RoCE におけるメモリ領域

RoCE では，主記憶へのアクセス時に CPU を介さず，RoCE に対応したネットワークカード (RNIC) が直接アクセスする．このために， RNIC の内部にアクセス可能となるメモリ空間を登録しておく必要があり，その単位を Memory Region (MR) と呼ぶ．MR に格納される情報を次に示す．

- メモリ空間の識別子
- メモリ空間の長さ
- ローカルや遠隔からのアクセスキー
- 所属する保護ドメイン
- MR の識別番号
- デバイスコンテキスト

RoCE において，ライブラリやカーネルの動作は InfiniBand Verbs (IB Verbs) によって規定されおり，MR はローカルの計算機が登録する．

メモリアクセスには，メモリ空間の長さや識別子，アクセスキーが利用される．メモリ空間の識別子は，RNIC がアクセスする空間を示し，IB Verbs ではメモリ空間の仮想アドレスが用いられる．これは，RNIC のメモリアクセスは物理アドレス空間を参照することに対し，RDMA で送受信したい変数等は物理アドレス空間上で不連続な場合があるためである．RDMA ドライバは，MR を登録する過程でメモリ空間の識別子に複数の物理アドレス空間を対応付け，これに対処している．

また，MR は保護ドメインに属し，保護ドメインをまたいだ空間へのアクセスは制限される．保護ドメインの設定によって，メモリ空間にアクセスできるプロセスやコネクションを制御でき，MR の識別番号はプロセス間で MR を共有する際に利用される．加えて，デバイスコンテキストは，libibverbs が提供する RNIC の情報であり，所属する RNIC の識別に用いられる．

3 ライブラリの拡張

RoCE を用いてアクセスする空間は MR に登録しておく必要があるが，IB Verbs では，MR に登録可能な空間は自プロセスの仮想アドレス空間に限定される．したがって，IB Verbs のライブラリ実装である libibverbs や RDMA ドライバといった標準的なインタフェースでは，カーネルや他プロセスのメモリ空間への RDMA が実現できない．そこで，libibverbs と Linux カーネル (Linux-signed-hwe-6.2) に含まれる RDMA ドライバを拡張し，MR にカーネルやプロセスのメモリ空間を登録するライブラリ関数を作成した．

Intel の RNIC に対応した RDMA ドライバである irdma には，MR の登録と物理アドレス空間への対応付けを行う関数があり，これを呼び出すことで，指定した

Memory Access Method for Remote Processes Using RDMA
Kengo Nagaki†, Kazuki Watanabe††, Gosuke Minami††, Gentaro Washio††, and Koichi Mouri†
†College of Information Science and Engineering, Ritsumeikan Univ.
††Information Technology R&D Center, Mitsubishi Electric Corporation.

```
426f 6f74 5072 696f 7269 7479 3a00 00f0
f3ee 00f0 54ff 00f0 8732 00f0 2f32 00f0
```

図 1　物理アドレス 0 番地 を先頭としたダンプ

```
426f6f745072696f726974793a0000f0
f3ee00f054ff00f0873200f02f3200f0
```

図 2　番地を先頭に遠隔から読み出した結果

物理アドレス空間を MR に登録できる．追加したライブラリ関数を次に示す．

```
struct ibv_mr *ibv_reg_phys_mr(struct ibv_pd *pd,
   uint64_t addr, size_t size, int access);
```

従来の MR 登録の流れから，物理アドレス空間との対応付け等の処理を変更することで，プロセス内外にかかわらず，引数で指定された物理アドレス空間を MR に登録する．

また，従来の MR 登録関数の流れをもとに，RDMA ドライバの関数を呼び出す流れを追加した．RDMA ドライバに追加したインタフェースを次に示す．

```
static int ib_uverbs_reg_phys_mr(struct
  uverbs_attr_bundle *attrs);
```

目的の関数を呼び出すことや，ユーザとの値の受け渡しを担う．

4　ライブラリの動作検証

ライブラリの動作検証として，次の 2 点を検証した．

1. 登録した物理アドレスに対して遠隔からアクセスが可能であるか
2. 物理アドレス空間全体を MR に登録し，それぞれに対して遠隔からアクセス可能であるか

検証には，Intel の RNIC である E810-CQDA1[3] を用いた．また，計算機 1，計算機 2 の主記憶はそれぞれ 16 GB， 64 GB とし，MR の範囲は 4 MB とした．

検証 1 は，拡張したインタフェースを用いてプロセス外の空間を MR に登録できるか検証するために行った．具体的には，物理アドレス 0 番地付近に対してアクセスし，ローカルでダンプした /dev/mem と比較することで検証した．

遠隔から読み出した結果とローカルで取得したダンプをそれぞれ図 1，図 2 に示す．書込みについても同様に検証し，物理アドレス空間に対してアクセス可能であることを確認した．

また，検証 2 は，物理アドレス空間のうち，デバイスに予約された空間や ROM にマップされた空間など，主記憶にマップされていない領域に対して遠隔からアクセスが可能であるかを検証するために行った．具体的な手順を図 3 に示す．計算機 1 では， 拡張したライブラリを利用して物理アドレス空間全体を MR に登録し，計算機 2 では，検証用に 0x78 で埋めたバッファと， 0x00 で埋めたバッファをそれぞれ MR に登録した．次に，計算機 1 のそれぞれの MR を 0x78 で埋め，0x00 で埋めたバッファに読み出す．その後，2 つのバッファを比較し，一致することを確認した．

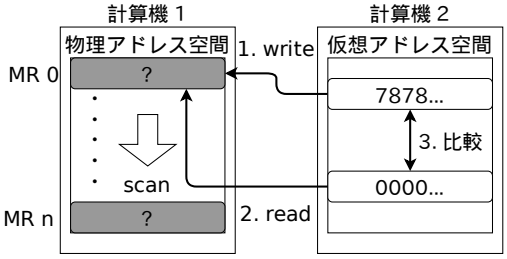

図 3　拡張したライブラリの動作確認方法

表 1　物理アドレス空間全体へのアクセス検証

結果	物理アドレスの範囲
アクセス可能	0x000000000 - 0x0C2FFFFFF
	0x100000000 - 0x436FFFFFF
読出し不可	0x0C3000000 - 0x0FFFFFFFF

物理アドレス空間全体に対してアクセスした結果を表 1 に示す．主記憶の大部分で遠隔からのメモリアクセスに成功したが，一部のアドレス空間で読出しに失敗した．読出しに失敗した空間はデバイスに予約された領域であり，タイムアウト時に返されるエラーステータス IBV_WC_RETRY_EXC_ERR が RNIC から返された．カーネルやプロセスのアドレス空間へは，遠隔からアクセスが可能であることが示された．

5　おわりに

本論文では，RDMA を用いた遠隔プロセスに対するメモリアクセスの実現を目指し，RoCE のメモリ領域作成方法について調査し，標準的なライブラリである libibverbs と Linux カーネルの RDMA ドライバを改変した．それにより，RDMA を用いた遠隔からのランダムアクセスを実現した．今後は，遠隔の仮想アドレス空間を意識したメモリアクセスの実現を目指し，必要な情報が RoCE で収集できるのかを検討する．また，メモリアクセスの一貫性について検討する．

参考文献

[1] Liu, H., Xing, J., Huang, Y., Zhuo, D., Devadas, S. and Chen, A.: Remote Direct Memory Introspection, *32nd USENIX Security Symposium (USENIX Security 23)*, USENIX Association, pp. 6043–6060.

[2] InfiniBand Trade Association: InfiniBand Architecture Specification Volume 1 Release 1.2.1.

[3] NEX Cloud Networking Group: Intel® Ethernet Controller E810 Datasheet. Revision 2.5 613875-007.