# RDMA Test Report

Zhuangdi Zhu

zhuangdizhu@yahoo.com

Abstract:

The following table summarizes test results for transferring 400MB data bi-directionally. Each value in the table is the average of 10 runs. All values are in seconds.

| | VM-VM 1 | VM-VM 4 | host-host 1 | host-host 4 |
|---|---|---|---|---|
| **TCP** | 0.982 | 1.194 | 1.097 | 1.233 |
| **RDMA** | 0.230 | 0.605 | 0.173 | 0.418 |

The details of the test steps are as follows.

**Step 1:**ssh username@raijin.nci.org.au

**Step 2:**cat transfer1.c

```
#include <stdio.h>
#include <mpi.h>
#include <stdlib.h>
#include <unistd.h>

int main(int argc, char** argv)
{
 int localID;
 int numOfPros;
 int loop = 20;
 int skip = 5;
 int i;
 MPI_Status reqstat;
 double t_start, t_end, t;
```

```
size_t Gsize = (size_t) 400 * 1024  * 1024;


char* s_buf = (char*)malloc(Gsize);

char* r_buf = (char*)malloc(Gsize);


MPI_Init(&argc, &argv);

MPI_Comm_size(MPI_COMM_WORLD, &numOfPros);

MPI_Comm_rank(MPI_COMM_WORLD, &localID);


if (localID == 0)

{

        for (i==0; i<loop+skip; i++) {

                if(i==skip) {

                        t_start = MPI_Wtime();

                }

       MPI_Send(s_buf, Gsize, MPI_CHAR, 1, 1, MPI_COMM_WORLD);

       MPI_Recv(r_buf, Gsize, MPI_CHAR, 1, 1, MPI_COMM_WORLD, &reqstat);

        }

        t_end = MPI_Wtime();

   t = t_end - t_start;

} else if (localID == 1)

{

        for(i=0; i<loop+skip; i++) {

       MPI_Recv(r_buf, Gsize, MPI_CHAR, 0, 1, MPI_COMM_WORLD, &reqstat);

       MPI_Send(s_buf, Gsize, MPI_CHAR, 0, 1, MPI_COMM_WORLD);

        }

}

if(localID == 0) {

        printf("time per 400MB: %f secs\n", 1.0 * t / loop);
```

```
 }
 if(s_buf!=NULL)
                 free(s_buf);
 if(r_buf!=NULL)
                 free(r_buf);
 MPI_Finalize();
 return 0;
 }
```

**Step 3:** module load openmpi/1.8.2

**Step 4:** mpicc transfer1.c –o transfer1

**Step 5:** mpirun –np 2 –host r11,r12 –report-bindings --mca btl self,tcp ./transfer1

**Step 6:** mpirun –np 2 -host r11,r12 –report-bindings --mca btl self,openib ./transfer1

**Step 7:** vim transfer.c

```
#include <stdio.h>
#include <mpi.h>
#include <stdlib.h>
#include <unistd.h>

int main(int argc, char** argv)
{
 int localID;
 int numOfPros;
 int loop = 20;
 int skip = 5;
 int i;
 MPI_Status reqstat;
```

```
double t_start, t_end, t;
size_t Gsize = (size_t) 400 * 1024  * 1024;

char* s_buf = (char*)malloc(Gsize);
char* r_buf = (char*)malloc(Gsize);

MPI_Init(&argc, &argv);
MPI_Comm_size(MPI_COMM_WORLD, &numOfPros);
MPI_Comm_rank(MPI_COMM_WORLD, &localID);

if (localID %2 == 0)
{
     for (i==0; i<loop+skip; i++) {
          if(i==skip) {
               t_start = MPI_Wtime();
          }
    MPI_Send(s_buf, Gsize, MPI_CHAR, localID+1, 1, MPI_COMM_WORLD);
    MPI_Recv(r_buf, Gsize, MPI_CHAR, localID+1, 1, MPI_COMM_WORLD, &reqstat);
     }
     t_end = MPI_Wtime();
  t = t_end - t_start;
} else if (localID %2 == 1)
{
     for(i=0; i<loop+skip; i++) {
    MPI_Recv(r_buf, Gsize, MPI_CHAR, localID-1, 1, MPI_COMM_WORLD, &reqstat);
    MPI_Send(s_buf, Gsize, MPI_CHAR, localID-1, 1, MPI_COMM_WORLD);
     }
}
```

```
if(localID%2 == 0) {
        printf("process %d: time per 400MB: %f secs\n",localID, 1.0 * t / loop);
}

if(s_buf!=NULL)
        free(s_buf);
if(r_buf!=NULL)
        free(r_buf);
MPI_Finalize();

return 0;
}
```

**Step 8:** mpicc transfer.c -o transfer

```
mpirun -np 8 -report-bindings -host r1,r2 --map-by node -bind-to socket --mca btl self,tcp --mca mpi_leave_pinned 1 ./transfer
```

```
mpirun -np 8 -report-bindings -host r1,r2 --map-by node -bind-to socket --mca btl self,openib --mca mpi_leave_pinned 1 ./transfer
```

Result:

| host-host | host-host | host-host | host-host | host-host | host-host | host-host | host-host | host-host | host-host | host-host | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Protocol | RDMA | TCP | TCP | TCP | TCP | TCP | RDMA | RDMA | RDMA | RDMA | |
| Num of transfer | 1 | 1 | 4 | | | | 4 | | | | |
| | | | | | | | | | | | |
| | | | Process0 | Process2 | Process4 | Process6 | Process0 | Process2 | Process4 | Process6 | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **No.1** | 0.174 | 1.021 | 1.602 | 1.002 | 1.287 | 1.010 | 0.405 | 0.438 | 0.411 | 0.414 | |
| **No.2** | 0.172 | 1.385 | 1.438 | 0.772 | 1.394 | 1.005 | 0.427 | 0.427 | 0.422 | 0.432 | |
| **No.3** | 0.174 | 0.942 | 1.548 | 0.936 | 1.367 | 0.873 | 0.338 | 0.351 | 0.339 | 0.336 | |
| **No.4** | 0.171 | 1.260 | 1.438 | 1.017 | 1.511 | 0.991 | 0.456 | 0.472 | 0.444 | 0.470 | |
| **No.5** | 0.172 | 0.913 | 1.357 | 1.024 | 1.461 | 1.156 | 0.400 | 0.410 | 0.394 | 0.408 | |
| **No.6** | 0.175 | 1.336 | 1.599 | 0.933 | 1.293 | 1.034 | 0.412 | 0.427 | 0.428 | 0.426 | |
| **No.7** | 0.174 | 0.951 | 1.502 | 1.037 | 1.487 | 0.856 | 0.452 | 0.448 | 0.440 | 0.462 | |
| **No.8** | 0.173 | 0.829 | 1.403 | 1.028 | 1.770 | 1.132 | 0.415 | 0.424 | 0.434 | 0.449 | |
| **No.9** | 0.175 | 1.286 | 1.364 | 0.774 | 1.591 | 0.934 | 0.418 | 0.417 | 0.427 | 0.427 | |
| **No.10** | 0.171 | 1.049 | 1.414 | 1.229 | 1.601 | 1.170 | 0.416 | 0.408 | 0.404 | 0.405 | |
| **Average** | 0.173 | 1.097 | 1.467 | 0.975 | 1.476 | 1.016 | 0.414 | 0.422 | 0.414 | 0.423 | |

**Step 9:** create two instances of VMs on NCI's openstack cloud.

Each VM is set with 4GB RAM, 2VCPU and 40GB Disk. We selected to use CentOS 6.5 images.

The real host is set with Intel Sandy Bridge E5-2670 processor(8-core), 2CPU, 32GB RAM. The network communication link is InfiniBand FDR(56Gb/s).

**Step 10:**

```
yum –y install openmpi-devel
```

**Step 11:** useradd test

passwd test

**Step 12:** su test

vim .bashrc (added following)

PATH=/usr/lib64/openmpi/bin:$PATH

LD_LIBRARY_PATH=/usr/lib64/openmpi/lib:$LD_LIBRARY_PATH

```
export PATH LD_LIBRARY_PATH
```

**Step 13:** ssh-keygen

```
scp .ssh/id_rsa.pub test@test12:~/.ssh/authorized_keys
chmod 600 ~/.ssh/authorized_keys
```

**Step 14:** mpicc transfer1.c –o transfer1

```
mpirun –np 2 –report-bindings –host test11,test12 --mca btl tcp,sm,self ./transfer1

[test11:10079] MCW rank 0 bound to socket 0[core 0[hwt 0]]: [B][.]
[test12:09981] MCW rank 1 bound to socket 0[core 0[hwt 0]]: [B][.]
time per 400MB:  1.021945  secs

mpirun –np 2 –report-bindings –host test11,test12 --mca btl openib,sm,self ./transfer1

[test11:10419] MCW rank 0 bound to socket 0[core 0[hwt 0]]: [B][.]
[test12:10163] MCW rank 1 bound to socket 0[core 0[hwt 0]]: [B][.]
time per 400MB:  0.230739  secs

mpicc transfer4.c –o transfer4
mpirun –np 8 –report-bindings –host test11,test12 --map-by node --mca btl tcp,sm,self ./transfer4

[test11:09830] MCW rank 0 is not bound (or bound to all available processors)
[test11:09830] MCW rank 2 is not bound (or bound to all available processors)
[test11:09830] MCW rank 4 is not bound (or bound to all available processors)
[test11:09830] MCW rank 6 is not bound (or bound to all available processors)
[test12:09915] MCW rank 5 is not bound (or bound to all available processors)
[test12:09915] MCW rank 7 is not bound (or bound to all available processors)
```

```
[test12:09915] MCW rank 1 is not bound (or bound to all available processors)
[test12:09915] MCW rank 3 is not bound (or bound to all available processors)
process 0: time per 400MB: 1.000563 secs
process 6: time per 400MB: 1.048819 secs
process 4: time per 400MB: 1.256563 secs
process 2: time per 400MB: 1.337364 secs
```

```
mpirun –np 8 –report-bindings –host test11,test12 --mca btl openib,sm,self ./transfer4
process 0: time per 400MB: 0.633935 secs
process 4: time per 400MB: 0.633760 secs
process 6: time per 400MB: 0.631587 secs
process 2: time per 400MB: 0.634126 secs
```

result:

| **VM-VM** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Protocol** | RDMA | TCP | TCP | TCP | TCP | TCP | RDMA | RDMA | RDMA | RDMA | |
| **transfer** | **1** | **1** | **4** | | | | **4** | | | | |
| | | | | | | | | | | | |
| | | | **process0** | **process2** | **process4** | **process6** | **process0** | **process2** | **process4** | **process6** | |
| **No.1** | 0.230 | 1.009 | 1.056 | 1.402 | 1.426 | 1.067 | 0.599 | 0.600 | 0.583 | 0.593 | |
| **No.2** | 0.230 | 1.022 | 1.257 | 1.074 | 1.224 | 1.085 | 0.592 | 0.594 | 0.601 | 0.615 | |
| **No.3** | 0.230 | 0.957 | 1.181 | 1.304 | 1.277 | 1.016 | 0.741 | 0.737 | 0.763 | 0.702 | |
| **No.4** | 0.230 | 0.966 | 1.131 | 1.248 | 1.323 | 1.080 | 0.568 | 0.569 | 0.579 | 0.595 | |
| **No.5** | 0.230 | 1.030 | 1.415 | 1.140 | 0.994 | 1.393 | 0.580 | 0.584 | 0.593 | 0.583 | |
| **No.6** | 0.230 | 0.992 | 1.103 | 1.291 | 1.295 | 1.117 | 0.583 | 0.552 | 0.565 | 0.560 | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **No.7** | 0.230 | 1.026 | 1.325 | 0.980 | 1.307 | 1.069 | 0.571 | 0.570 | 0.570 | 0.576 | |
| **No.8** | 0.230 | 0.899 | 1.131 | 1.356 | 1.054 | 1.363 | 0.552 | 0.555 | 0.563 | 0.558 | |
| **No.9** | 0.230 | 1.036 | 1.001 | 1.337 | 1.257 | 1.049 | 0.634 | 0.634 | 0.634 | 0.632 | |
| **No.10** | 0.230 | 0.880 | 0.983 | 1.274 | 1.004 | 1.388 | 0.615 | 0.620 | 0.618 | 0.657 | |
| **Avg** | 0.230 | 0.982 | 1.158 | 1.241 | 1.216 | 1.163 | 0.603 | 0.601 | 0.607 | 0.607 | |