MPI Quick Reference: Compiling/Running

M. D. Jones, Ph.D.

Center for Computational Research
University at Buffalo
State University of New York

High Performance Computing I, 2013

Background

This document covers the essentials of compiling and running MPI applications on the CCR platforms. It does not cover MPI programming itself, nor debugging, etc. (covered more thoroughly in separate presentations).

Modules Software Management System

There are a large number of available software packages on the CCR systems, particularly the Linux clusters. To help maintain this often confusing environment, the **modules** package is used to add and remove these packages from your default environment (many of the packages conflict in terms of their names, libraries, etc., so the default is a minimally populated environment).

The module Command

module command syntax:

```
-bash-2.05b$ module help

  Modules Release 3.1.6 (Copyright GNU GPL v2 1991):
  Available Commands and Usage:
        + add|load              modulefile [modulefile ...]
        + rm|unload             modulefile [modulefile ...]
        + switch|swap           modulefile1 modulefile2
        + display|show          modulefile [modulefile ...]
        + avail                 [modulefile [modulefile ...]]
        + use [-a|--append]     dir [dir ...]
        + unuse                 dir [dir ...]
        + update
        + purge
        + list
        + clear
        + help                  [modulefile [modulefile ...]]
        + whatis                [modulefile [modulefile ...]]
        + apropos|keyword       string
        + initadd               modulefile [modulefile ...]
        + initprepend           modulefile [modulefile ...]
        + initrm                modulefile [modulefile ...]
        + initswitch            modulefile1 modulefile2
        + initlist
        + initclear
```

Using module in Batch

If you change shells in your batch script you may need to explicitly load the modules environment:

tcsh :

```
source $MODULESHOME/init/tcsh
```

bash :

```
. ${MODULESHOME}/init/bash
```

but generally you should not need to worry about this step (do a "module list" and if it works ok your environment should already be properly initialized).

Objective: Construct a very elementary MPI program to do the usual "Hello World" problem, i.e. have each process print out its rank in the communicator.

in C

```c
#include <stdio.h>
#include "mpi.h"

int main(int argc,char **argv)
{
    int myid, nprocs;
    int  namelen,mpiv,mpisubv;
    char processor_name[MPI_MAX_PROCESSOR_NAME];

    MPI_Init(&argc,&argv);
    MPI_Comm_size(MPI_COMM_WORLD,&nprocs);
    MPI_Comm_rank(MPI_COMM_WORLD,&myid);
    MPI_Get_processor_name(processor_name,&namelen);

    printf("Process %d of %d on %s\n",myid, nprocs, processor_name);
    if (myid == 0) {
       MPI_Get_version(&mpiv,&mpisubv);
       printf("MPI Version: %d.%d\n",mpiv,mpisubv);
    }
    MPI_Finalize();
    return 0;
}
```

CCR: Intel MPI

There are several commercial implementations of MPI, `Intel` and `IBM` currently being the most prominent. CCR has a license for `Intel` MPI, and it has some nice features:

- Support for multiple networks (Infiniband, Myrinet, TCP/IP)
- Part of the ScaLAPACK support in the Intel MKL
- MPI-2 features (one-sided, dynamic tasks, I/O with parallel filesystems support)
- CPU pinning/process affinity options (extensive)

Build the code with the appropriate wrappers:

```
[rush:~/d_mpi-samples]$ module load intel-mpi intel
[rush:~/d_mpi-samples]$ module list
Currently Loaded Modulefiles:
  1) null               2) modules           3) use.own
  4) intel-mpi/4.0.3    5) intel/13.1
[rush:~/d_mpi-samples]$ mpiicc -o hello.impi hello.c     # icc version
[rush:~/d_mpi-samples]$ mpicc -o hello.impi.gcc hello.c  # gcc version
```

Unfortunately Intel MPI can have somewhat flaky integration with Slurm, so you have a range of options when it comes to launching MPI codes.

Simple Example - Interactive

```
[rush:~/d_mpi-samples]$ mpirun -np 4 ./hello.impi
Process 0 of 4 on f07n05
MPI Version: 2.2
Process 1 of 4 on f07n05
Process 2 of 4 on f07n05
Process 3 of 4 on f07n05
[rush:~/d_mpi-samples]$ mpirun -np 4 ./hello.impi
[0] MPI startup(): shm data transfer mode
[1] MPI startup(): shm data transfer mode
[2] MPI startup(): shm data transfer mode
[3] MPI startup(): shm data transfer mode
[0] MPI startup(): Rank    Pid      Node name  Pin cpu
[0] MPI startup(): 0       8370     f07n05     {0,4,8,12,16,20,24,28}
Process 1 of 4 on f07n05
Process 2 of 4 on f07n05
Process 3 of 4 on f07n05
[0] MPI startup(): 1       8371     f07n05     {1,5,9,13,17,21,25,29}
[0] MPI startup(): 2       8372     f07n05     {2,6,10,14,18,22,26,30}
[0] MPI startup(): 3       8373     f07n05     {3,7,11,15,19,23,27,31}
Process 0 of 4 on f07n05
MPI Version: 2.2
```

Simple Example - Slurm/srun

```
1  #!/bin/bash
2  #SBATCH --nodes=1
3  #SBATCH --ntasks-per-node=8
4  #SBATCH --partition=debug
5  #SBATCH --time=00:10:00
6  #SBATCH --mail-type=END
7  #SBATCH --mail-user=jonesm@buffalo.edu
8  #SBATCH --output=slurmQ.out
9  #SBATCH --job-name=mpi-test
10 #
11 # Note the above directives can be commented out using an
12 # additional "#"
13 #
14 module load intel-mpi intel
15 #
16 # Intel MPI has flaky tight integration with Slurm,
17 # generally it has been safer to use Slurm's srun rather than
18 # rely on mpirun/mpiexec.
19 # You can find a description of all Intel MPI parameters in the
20 # Intel MPI Reference Manual,
21 # see <intel mpi installdir>/doc/Reference_manual.pdf
22 #
23 export I_MPI_DEBUG=4    # nice debug level, spits out useful info
24 #
25 # mpirun wrapper:
26 mpirun ./hello.impi
27 # srun:
28 export I_MPI_PMI_LIBRARY=/usr/lib64/libpmi.so
29 srun ./hello.impi
```

Output from the `mpirun` launch (note that the pinning is not what we want by default, you can likely force better behavior with `I_PIN_MODE` and its associated variables):

```
1  [7] MPI startup(): shm data transfer mode
2  [2] MPI startup(): shm data transfer mode
3  [4] MPI startup(): shm data transfer mode
4  [6] MPI startup(): shm data transfer mode
5  [3] MPI startup(): shm data transfer mode
6  [1] MPI startup(): shm data transfer mode
7  [5] MPI startup(): shm data transfer mode
8  [0] MPI startup(): shm data transfer mode
9  Process 1 of 8 on d09n29s02
10 Process 7 of 8 on d09n29s02
11 Process 5 of 8 on d09n29s02
12 Process 3 of 8 on d09n29s02
13 Process 6 of 8 on d09n29s02
14 Process 2 of 8 on d09n29s02
15 Process 4 of 8 on d09n29s02
16 [0] MPI startup(): Rank    Pid      Node name  Pin cpu
17 [0] MPI startup(): 0       15265    d09n29s02  0
18 [0] MPI startup(): 1       15266    d09n29s02  0
19 [0] MPI startup(): 2       15267    d09n29s02  0
20 [0] MPI startup(): 3       15268    d09n29s02  0
21 [0] MPI startup(): 4       15269    d09n29s02  0
22 [0] MPI startup(): 5       15270    d09n29s02  0
23 [0] MPI startup(): 6       15271    d09n29s02  0
24 [0] MPI startup(): 7       15272    d09n29s02  0
25 Process 0 of 8 on d09n29s02
26 MPI Version: 2.2
```

Output from the `srun` launch (in this case +1 indicates that pinning is turned off):

```
27  [2] MPI startup(): shm data transfer mode
28  [3] MPI startup(): shm data transfer mode
29  [5] MPI startup(): shm data transfer mode
30  [6] MPI startup(): shm data transfer mode
31  [7] MPI startup(): shm data transfer mode
32  [0] MPI startup(): shm data transfer mode
33  [1] MPI startup(): shm data transfer mode
34  [4] MPI startup(): shm data transfer mode
35  Process 1 of 8 on d09n29s02
36  Process 3 of 8 on d09n29s02
37  Process 5 of 8 on d09n29s02
38  Process 7 of 8 on d09n29s02
39  [0] MPI startup(): Rank    Pid      Node name  Pin cpu
40  [0] MPI startup(): 0       15311    d09n29s02  +1
41  [0] MPI startup(): 1       15312    d09n29s02  +1
42  [0] MPI startup(): 2       15313    d09n29s02  +1
43  [0] MPI startup(): 3       15314    d09n29s02  +1
44  [0] MPI startup(): 4       15315    d09n29s02  +1
45  [0] MPI startup(): 5       15316    d09n29s02  +1
46  [0] MPI startup(): 6       15317    d09n29s02  +1
47  Process 2 of 8 on d09n29s02
48  Process 4 of 8 on d09n29s02
49  Process 6 of 8 on d09n29s02
50  [0] MPI startup(): 7       15318    d09n29s02  +1
51  Process 0 of 8 on d09n29s02
52  MPI Version: 2.2
```

Intel MPI on Infiniband

The CCR nodes generally have Infiniband (IB) as the optimal interconnect for message-passing, running an Intel MPI job should automatically find and use IB on those machines (and they have 8, 12, 16, or 32 cores each, so adjust your script accordingly). Here we change out node count request to two, and just run with `srun`:

```
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=8
#SBATCH --partition=debug
```

```
1  [rush:~/d_mpi-samples]$ cat subQ.out
2  [1] MPI startup(): shm and ofa data transfer modes
3  [2] MPI startup(): shm and ofa data transfer modes
4  [9] MPI startup(): shm and ofa data transfer modes
5  [3] MPI startup(): shm and ofa data transfer modes
6  [10] MPI startup(): shm and ofa data transfer modes
7  [4] MPI startup(): shm and ofa data transfer modes
8  [11] MPI startup(): shm and ofa data transfer modes
9  [5] MPI startup(): shm and ofa data transfer modes
10 [12] MPI startup(): shm and ofa data transfer modes
11 [6] MPI startup(): shm and ofa data transfer modes
12 [13] MPI startup(): shm and ofa data transfer modes
13 [7] MPI startup(): shm and ofa data transfer modes
14 [0] MPI startup(): shm and ofa data transfer modes
```

```
15 Process 5 of 16 on d16n02
16 Process 10 of 16 on d16n03
17 Process 6 of 16 on d16n02
18 Process 11 of 16 on d16n03
19 Process 7 of 16 on d16n02
20 Process 13 of 16 on d16n03
21 [0] MPI startup(): Rank    Pid      Node name  Pin cpu
22 [0] MPI startup(): 0       7898     d16n02     +1
23 [0] MPI startup(): 1       7899     d16n02     +1
24 [0] MPI startup(): 2       7900     d16n02     +1
25 [0] MPI startup(): 3       7901     d16n02     +1
26 [0] MPI startup(): 4       7902     d16n02     +1
27 [0] MPI startup(): 5       7903     d16n02     +1
28 [0] MPI startup(): 6       7904     d16n02     +1
29 [0] MPI startup(): 7       7905     d16n02     +1
30 [0] MPI startup(): 8       29934    d16n03     +1
31 [0] MPI startup(): 9       29935    d16n03     +1
32 [0] MPI startup(): 10      29936    d16n03     +1
33 [0] MPI startup(): 11      29937    d16n03     +1
34 [0] MPI startup(): 12      29938    d16n03     +1
35 [0] MPI startup(): 13      29939    d16n03     +1
36 [0] MPI startup(): 14      29940    d16n03     +1
37 [0] MPI startup(): 15      29941    d16n03     +1
38 Process 14 of 16 on d16n03
39 Process 4 of 16 on d16n02
40 Process 15 of 16 on d16n03
41 Process 0 of 16 on d16n02
42 MPI Version: 2.2
43 Process 8 of 16 on d16n03
44 Process 12 of 16 on d16n03
```

Intel MPI on TCP/IP

You can force Intel MPI to run using TCP/IP (or a combination of tcp/ip and shared memory as in the example below) by setting the `I_MPI_DEVICE` variable, or equivalently `I_MPI_FABRICS`):

```
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=1
#SBATCH --partition=debug
...
export I_MPI_DEBUG=4
export I_MPI_FABRICS="ssm:tcp" # tcp/ip between nodes, shared memory within
```

```
1  Process 0 of 2 on d16n02
2  Process 1 of 2 on d16n03
3  [0] MPI startup(): shm and tcp data transfer modes
4  [1] MPI startup(): shm and tcp data transfer modes
5  [0] MPI startup(): Rank    Pid      Node name  Pin cpu
6  [0] MPI startup(): 0       9072     d16n02     +1
7  [0] MPI startup(): 1       30436    d16n03     +1
8  MPI Version: 2.2
9  Process 1 of 2 on d16n03
```

Intel MPI Summary

Intel MPI has some real advantages:

- Multi-protocol support with the same build, by default gives you the "best" network, but also gives you the flexibility to choose your protocol
- CPU/memory affinity settings
- Multiple compiler support (wrappers for GNU compilers, **mpicc, mpicxx, mpif90**, as well as Intel compilers, **mpiicc, mpicpc, mpiifort**)
- (Relatively) simple integration with Intel MKL, including ScaLAPACK
- Reference manual - on the CCR systems look at `$INTEL_MPI/doc/Reference_Manual.pdf` for a copy of the reference manual (after loading the module)

Whither Goest Thou, MPI?

MPI processes - things to keep in mind:

- You can over-subscribe the processors if you want, but that is going to under-perform (but it is often useful for debugging). Note that batch queuing systems (like those in CCR) may not let you easily over-subscribe the number of available processors
- Better MPI implementations will give you more options for the placement of MPI tasks (often through so-called "affinity" options, either for CPU or memory)
- Typically want a 1-to-1 mapping of MPI processes with available processors (cores), but there are times when that may not be desirable

Affinity

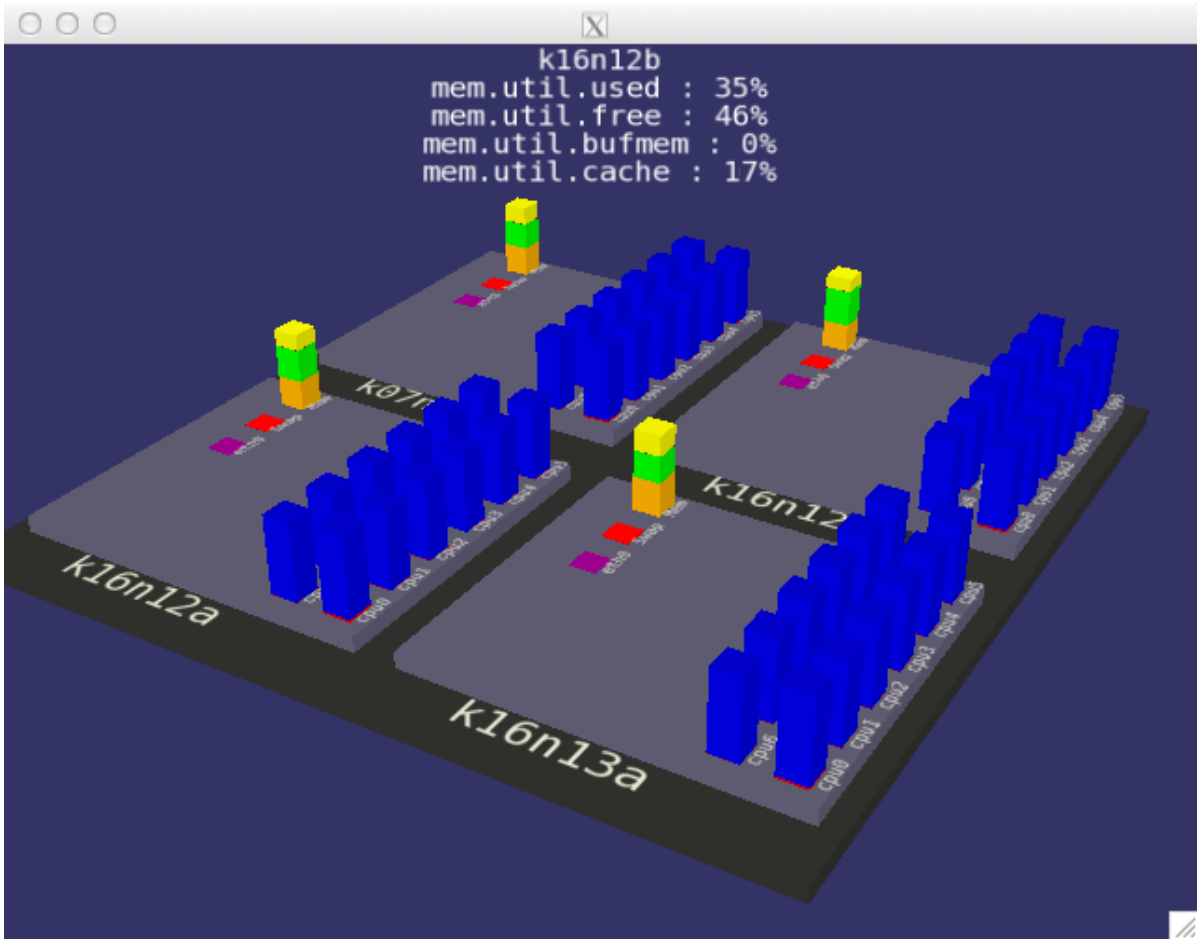
Intel MPI has options for associating MPI tasks to cores - better known as CPU-process affinity

- `I_MPI_PIN`, `I_MPI_PIN_MODE`, `I_MPI_PIN_PROCESSOR_LIST`, `I_MPI_PIN_DOMAIN` in the current version of Intel MPI (it never hurts to check the documentation for the version that you are using, these options have a tendency to change)
- Can specify core list on which to run MPI tasks, also domains of cores for hybrid MPI-OpenMP applications

Summary - MPI at CCR

- Use `modules` environment manager to choose your MPI flavor
- I recommend `Intel MPI` on the clusters, unless you need access to the source code for the implementation itself. It has a lot of nice features and is quite flexible.
- Be careful with task launching - use `srun` in Slurm jobs, or treat pinning with care
- Ensure that your MPI processes end up where you want - use `ps` and `top` to check (also use MPI_Get_processor_name in your code).
- Also use the CCR `slurmjobvis` job visualizer utility to quickly scan for expected task placement and performance issues.

`slurmjobviz.pl` Example



Sample `slurmjobvis` from a 4-node job using all 12 cores/node and roughly 50% of available memory/node.