lrz Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

MIC-MPI-lab: Simple MPI programs and Intel MPI Benchmarks in various modes on the MIC

In this lab you will gain first experiences with the various MPI execution scenarios supported by the Intel compiler for the Intel Xeon Phi architecture: Native mode, MPI ranks on hosts and/or coprocessors, MPI ranks on hosts which do offloading etc.
We use a simple hello world program to understand the various execution modes. The second part of the exercises uses the Intel MPI Benchmark suite to do some performance measurements on SuperMIC.

Appropriate Environment

- Login to SuperMIC:
  ```
  ssh supermic.smuc.lrz.de
  ```

- The right Intel module is loaded automatically on SuperMIC, on SuperMUC use:

  ```
  module unload mpi.ibm
  module load mpi.intel
  ```

- Setup important paths:
  ```
  . $ICC_BASE/bin/compilervars.sh intel64
  . $MPI_BASE/bin64/mpivars.sh
  ```
- Ensue that *I_MPI_MIC = enabled* and *I_MPI_HYDRA_BOOTSTRAP=ssh*.

Lab 1: MPI on the hosts

- Compile the simple hello world program hellompi.c using
  ```
      mpiicc hellompi.c -o hellompi-host
  ```

- Run 2 MPI ranks on a compute-host using
  ```
      mpiexec -n 2 -host i01r13a01 ./hellompi-host
  ```

- Modify the code that every MPI process outputs its rank, the total number of ranks and the hostname it is running on.

Lab 2: MPI on a single MIC in native mode

- Compile the code for the Intel MIC architecture using:

  ```
  mpiicc  -mmic hellompi.c -o hellompi-mic
  ```

- Copy the MIC-binary to a MIC coprocessor, e.g. *i01r13c01-mic0:*

  ```
  scp hellompi-mic  i01r13c01-mic0:
  ```

- Login to the MIC and run MPI natively, e.g.

  ```
  mpirun -n 10 ./hellompi-mic
  ```

Lab 3: MPI on the MICs launched from the host

- Launch many MPI tasks of the program from the host on one single MIC using

  ```
  mpiexec -n 5 -host i01r13c*-mic0 program
  ```

- Run one MPI task per MIC on the 2 MICs attached to one single host using

  ```
  mpiexec -n 2 -perhost 1 -host mic0,mic1 program
  ```

Lab 4: MPI on the host and its 2 attached MICs

- Run 2 MPI tasks on 1 host and 2 tasks per MIC on both MICs attached to the host using

  ```
  mpirun -n 2 -host hostname hostprogram: -n 2 -host hostname-mic0 micprogram
  : -n 2 -host hostname-mic1 micprogram
  ```
- If enough nodes are free, run MPI tasks on multiple hosts and MICs. *llq –l JOBID* provides you the names of the allocated hosts. When running on multiple nodes, you need to *export e.g. I_MPI_DAPL_PROVIDER_LIST=ofa-v2-mlx4_0-1u*

Lab 5: MPI on the host using Offloading

- Modify the code so that an offload region with another simple “hello world from MIC” is offloaded to the MIC.
- Run multiple MPI ranks just on 1 host and verify that the offload region runs on a MIC.
- Modify the file so that the code offloaded to a MIC tells the hostname of the coprocessor and the hostname and rank of the MPI process on the host which has done the offload.
- Offload MIC code from multiple MPI processes on the host to its 2 MICs, ensure that not only mic0 is used.

Lab 6: MPI Performance

In this Lab you use the Intel MPI Benchmark (IMB) Suite to compare the performance of various MPI communication patters.

On the MIC the benchmark binary is available under /bin/IMB-MPI1.
On the host it is available under $I_MPI_ROOT/bin64/IMB-MPI1

- What is the maximum PingPong bandwidth between the host and 1 MIC? Use a command similar to
  mpiexec -n 1 -host hostname IMB-MPI1 PingPong : -n 1 -host hostname-mic0 /bin/IMB-MPI1 PingPong
  Compare with the max. PCIe2 speed!

- How about the latency? Look at t[usec] for small #bytes.

- Also measure the PingPong Bandwidth between the 2 MICs attached to one host, between 2 hosts , and between 2 MICs attached to different hosts. If you are familiar with gnuplot (*module load gnuplot*) or other plotting programs use them to compare the performance graphically.

- The Default MPI Fabrics configured on SuperMIC is *I_MPI_FABRICS=shm:dapl*. to use DAPL via Infiniband. Change the MPI Fabrics setting to use tcp for internode communication and compare the performance with MPI using DAPL.

Beyond PingPong the Intel MPI Benchmark suite also offers the following benchmarks:

# PingPong
# PingPing
# Sendrecv
# Exchange
# Allreduce
# Reduce
# Reduce_scatter
# Allgather
# Allgatherv
# Gather
# Gatherv
# Scatter
# Scatterv
# Alltoall
# Alltoallv
# Bcast
# Barrier

Look into the Intel IMB Docu and the MPI Standard http://www.mpi-forum.org/docs/docs.html what these MPI functions do and compare the performance in native mode and using host1↔host2, host1-mic0↔host1-mic1 etc. patterns.