NEMO Best Practices for Intel® Cluster Ready

HPC ADVISORY COUNCIL
NETWORK OF EXPERTISE
BEST PRACTICES

1. Introduction

The following best practices document is provided as courtesy of the HPC Advisory Council.

2. Application Description:

NEMO (Nucleus for European Modeling of the Ocean) is a state-of-the-art modeling framework for oceanographic research and operational oceanography. The core of the system is a primitive equation model applied to both regional and global ocean circulation. It is intended to be a flexible tool for studying the ocean, the sea-ice and its interactions with the others components of the earth climate system (atmosphere, land-surfaces, vegetation). NEMO is written in Fortran 90 and is parallelized with a domain decomposition using MPI library. All outputs are done with NetCDF library.

This benchmark has been built to test the performances of the OPA OGCM component (released 9.0) on various HPC platforms. It is part of DEISA benchmark suite.

NEMO is free software. It is available under the CeCILL license (public license). Benchmark version can be downloaded from http://www.deisa.eu/science/benchmarking. Complete version is under

http://www.nemo-ocean.eu

3. Version Information:

Version of this date is used: NEMO 3.2

4. Prerequisites:

The instructions from this best practice have been tested with the following configuration:

4.1 Hardware:

- HP ProLiant SL2x170z G6 16-node cluster
- Intel Xeon X5670 CPUs @ 2.93 MHz
- Memory: 24GB per node @ 1333MHz
- Mellanox ConnectX-2 QDR InfiniBand Adapters
- Mellanox QDR InfiniBand Switch

4.2 Software:

- Intel® Cluster Ready running CentOS5U5
- Mellanox OFED 1.5.3 InfiniBand Software Stack
- Application: NEMO 3.2
- Compilers: Intel compiler 11.1.064
- Math library: Intel MKL 2011.3.174, netCDF 2.122
- MPI: Intel MPI 4, Open MPI 1.7a, Platform MPI 8.0.1
- Benchmark workload: OPA (the ocean engine), confcoef=25

5. Obtaining the source code

NEMO is part of DEISA benchmark suite which can be found from this URL:

http://www.deisa.eu/science/benchmarking

6. Building NEMO

Extract DEISA_BENCH.tar.gz

NEMO source code is under DEISA_BENCH/applications/NEMO

Modify platform.xml to match compiler, netcdf, and MPI versioin on the system.

Create a new bench-intel.xml file under NEMO directory, using bench-Intel-Nehalem-JuRoPA.xml as template.

In benchmark section, pick the appropriate conf value based on your system size. We choose conf=25 for our small cluster. Node counts and tasks per node are also specified in the xml file. It is defined as below:

```
<benchmark name=”conf25_minIO” active=”1”>
  <!– version=”reuse|new” –>
  <compile    cname=”$platform”
confcoef=”$confcoef” nbproc=”$ncpus”
            version=”reuse” />
  <tasks      threadspertask=”1” taskspernode=”12”
nodes=”1,2,4,8,16” />
  <params     ncpus=”`$threadspertask*$taskspernode
*$nodes`”
            confcoef=”25” mpicommode=”I”
timesteps=”1500”
            writetimesteps=”1500”
restarttimesteps=”1500” />
  <prepare    cname=”NEMO_namelist” />
  <execution  iteration=”1” cname=”$platform” />
  <verify     cname=”NEMO” />
<analyse    cname=”NEMO” />
```

Compile NEMO

```
% perl ../../bench/jube bench-Intel.xml
```

A set of executables will be built for different number of processes. All files will be created under tmp directory.

7. Running NEMO

Cd to tmp directory, the benchmark input file "namelist" has been copied to each running folder by DEISA jube script.

Running with Intel MPI

```
    % mpdboot  --parallel-startup -r ssh -f  ~/hosts -n 16

    % mpiexec -perhost 12 -env I_MPI_DEVICE rdssm
    -np $i

    /apps/DEISA_BENCH/applications/NEMO/tmp/
    NEMO_Intel _conf25_minIO_i000003/n16p12t1_
    t001_i01/NEMO_Intel_cname_Intel_confcoef_25_
    nbproc_192.exe
```

Tune Intel MPI:

```
    % mpiexec -perhost 12 -genv I_MPI_RDMA_
    TRANSLATION_CACHE  1 -genv

    I_MPI_RDMA_RNDV_BUF_ALIGN  65536  -genv
    I_MPI_DAPL_DIRECT_COPY_THRESHOLD 65536
    -genv I_MPI_EXTRA_FILESYSTEM on -genv I_MPI_
    EXTRA_FILESYSTEM_LIST lustre -env

    I_MPI_DEVICE rdssm  -np 192  /apps/DEISA_BENCH/
    applications/NEMO/tmp/NEMO_Intel _conf25_minIO_
    i000003/n16p12t1_t001_i01/NEMO_Intel_cname_
    Intel_confcoef_25_nbproc_192.exe


     %mpdallexit
```

Running with Platform MPI

```
% mpirun -cpu_bind -hostfile ~/hosts -np 192   /apps/
DEISA_BENCH/applications/NEMO/tmp/NEMO_Intel
_conf25_minIO_i000003/n16p12t1_t001_i01/NEMO_Intel_
cname_Intel_confcoef_25_nbproc_192.exe
```

Benchmark timing information is printed out at the end of standard output.