From: Ilia Bermous

Date: 13 August 2015

**Usage of Intel compiler and MPI library for APS3 models run on Raijin**

1. **Modules**

In my recent UM and 4DVAR jobs for PS36 UM10.1 and VAR30.0.0 sources the following Intel compiler and Intel MPI library modules have been used:

**module add intel-fc/15.0.1.133**

**module add intel-cc/15.0.1.133**

**module load intel-mpi/5.0.3.048**

Note that “-mt\_mpi” linking option has to be used for building UM executable with IO servers usage.

1. **Usage for UM jobs**
2. environment settings

**export I\_MPI\_DAPL\_TRANSLATION\_CACHE=0**

**export I\_MPI\_FABRICS=shm:dapl**

**export I\_MPI\_DAPL\_PROVIDER=ofa-v2-mlx4\_0-1u**

**export I\_MPI\_HYDRA\_BRANCH\_COUNT=$(($PBS\_NCPUS / 16))**

**export LD\_LIBRARY\_PATH=/home/548/iib548/IntelMPI/dapl-2.1.5/dapl/udapl/.libs:$LD\_LIBRARY\_PATH**

**export KMP\_AFFINITY=compact,verbose**

if hyper-threading is used by specifying

**#PBS -l other=hyperthread**

in the resources section of a corresponding PBS batch job then the KMP\_AFFINITY setting should be the following

**export KMP\_AFFINITY=scatter,verbose**

instead of using the “compact” setting mentioned above. Usage of hyper-threading can provide up to 10% performance improvements for a UM job.

I think now it is the right time to put the DAPL 2.1.5 library from my local directory to a place somewhere under ~accesss on Raijin

1. mpirun command

the following mpirun command has been used

**uniq < $PBS\_NODEFILE > hostfile.txt**

**mpirun -perhost 8 -hostfile hostfile.txt $LOADMODULE > um\_output 2>&1**

where LOADMODULE is a UM executable.

1. the recommended run configuration on Raijin using fully committed nodes with 3552 cores is

**ulimit -l 2097152**

**export UM\_THREAD\_LEVEL=${UM\_THREAD\_LEVEL:-MULTIPLE}**

**export OMP\_NUM\_THREADS=2**

**export NPERSOCKET=4**

**export UM\_ATM\_NPROCX=24**

**export UM\_ATM\_NPROCY=72**

**export OMP\_STACKSIZE=2g**

**export FLUME\_IOS\_NPROC=48**

1. IOSCNTL namelist file settings

settings for the above mentioned namelist file using PS36 UM10.1 N768L70 EG job are provided in Appendix below.

1. Usage for 4DVAR jobs
2. environment settings for N320L70 using 12x128 decomposition on 1536 cores

**export VAR\_NPROCX=12**

**export VAR\_NPROCY=128**

**export I\_MPI\_DAPL\_TRANSLATION\_CACHE=0**

**export I\_MPI\_FABRICS=shm:dapl**

**export I\_MPI\_DAPL\_PROVIDER=ofa-v2-mlx4\_0-1u**

**export I\_MPI\_HYDRA\_BRANCH\_COUNT=$(($PBS\_NCPUS / 16))**

**export LD\_LIBRARY\_PATH=/home/548/iib548/IntelMPI/dapl-2.1.5/dapl/udapl/.libs:$LD\_LIBRARY\_PATH**

**export I\_MPI\_FALLBACK=disable**

**export DAPL\_UCM\_REP\_TIME=8000**

**export DAPL\_UCM\_RTU\_TIME=8000**

**export DAPL\_UCM\_RETRY=10**

**export DAPL\_UCM\_QP\_SIZE=2000**

**export DAPL\_UCM\_CQ\_SIZE=2000**

**export DAPL\_UCM\_TX\_BURST=100**

1. mpirun command using fully committed nodes

**uniq < $PBS\_NODEFILE > hostfile.txt**

**mpirun -perhost 16 -hostfile hostfile.txt $VAR\_AP\_PROG**

or using a similar form as it is provided below for partial node usage

**export I\_MPI\_PIN\_PROCESSOR\_LIST="0-7,8-15"**

**export NP=$((VAR\_NPROCX \* VAR\_NPROCY))**

**mpirun -np $NP -f $PBS\_NODEFILE -ppn 16 $VAR\_AP\_PROG**

1. mpirun command using partially committed nodes

**export I\_MPI\_PIN\_PROCESSOR\_LIST="0-5,8-13"**

**export NP=$((VAR\_NPROCX \* VAR\_NPROCY))**

**mpirun -np $NP -f $PBS\_NODEFILE -ppn 12 $VAR\_AP\_PROG**

**Appendix. IOSCNTL namelist file settings**

**&ioscntl**

**ios\_acquire\_model\_prsts=.false.,**

**ios\_as\_concurrency=400,**

**ios\_async\_levs\_per\_pack=400,**

**ios\_async\_send\_null=.false.,**

**ios\_async\_stats=.false.,**

**ios\_backoff\_interval=400,**

**ios\_buffer\_size=20567,**

**ios\_concurrency=400,**

**ios\_concurrency\_max\_mem=400,**

**ios\_debug\_no\_packing=.false.,**

**ios\_debug\_no\_subdomaining=.false.,**

**ios\_debug\_no\_write=.false.,**

**ios\_decomp\_model=0,**

**ios\_interleave=.true.,**

**ios\_local\_ro\_files=.true.,**

**ios\_lock\_meter=.false.,**

**ios\_offset=0,**

**ios\_print\_start\_time=.false.,**

**ios\_relaytoslaves=.true.,**

**ios\_serialise\_mpi\_calls=.false.,**

**ios\_spacing=16,**

**ios\_tasks\_per\_server=12,**

**ios\_thread\_0\_calls\_mpi=.false.,**

**ios\_timeout=120,**

**ios\_unit\_alloc\_policy=3,**

**ios\_use\_async\_dump=.true.,**

**ios\_use\_async\_stash=.true.,**

**ios\_use\_helpers=.true.,**

**ios\_verbosity=3,**

**/**

**&io\_control**

**io\_data\_alignment=131072,**

**io\_external\_control=.false.,**

**io\_field\_padding=512,**

**io\_rbuffer\_count=0,**

**io\_rbuffer\_prefetch=0,**

**io\_rbuffer\_size=131072,**

**io\_rbuffer\_update=1,**

**io\_timing=1,**

**io\_wbuffer\_size=131072,**

**l\_postp=.false.,**

**print\_runtime\_info=.true.,**

**/**

**&prnt\_control**

**prnt\_force\_flush=.false.,**

**prnt\_paper\_width=80,**

**prnt\_split\_lines=.false.,**

**prnt\_src\_pref=.false.,**

**prnt\_writers=1,**

**/**