Quick Reference Guide to the High Performance Computer Environments @ NeSI

New Zealand eScience Infrastructure

Center for eResearch - University of Auckland, New Zealand
Building 409, Rm G21, LG, 24 Symonds Street - (1010) Auckland
Phone : (+64) 9 373 7599 - Support list: eresearch-support@auckland.ac.nz

Contact Information

- **web**: http://www.nesi.org.nz
- **wiki**: https://wiki.auckland.ac.nz/display/CERES/
- **ganglia**: http://ganglia.uoa.nesi.org.nz
- **support** : eresearch-support@auckland.ac.nz

Useful Quick References

- VI Quick Reference
- BASH Quick Reference
- Linux Quick Reference
- OpenMP Fortran Syntax
- OpenMP 3.1 API C/C++ Syntax
- MPI Quick Reference

Hardware Information

Architecture	Gulftown	SandyBridge	LargeMemory	GPU
Model	X5660	E5-2680	X	X
Clock Speed	2.8 GHz	2.7 GHz	X	X
Cache	12MB	20MB	X	X
Intel QPI speed	6.4GT/s	8 GT/s	X	X
Cores/socket	6	8	X	X
Cores/node	12	16	X	X
Mem/node	96GB	128GB	512GB	X
GFLOPS/node	134.4	172.8	X	X

Disk Spaces + Quota

FileSystem	Space	Quota	ACL	Backup	Type	Usage
$HOME	34TB	30GB	rw	yes	GPFS	Archive
/share	134TB	-	ro	yes	GPFS	Archive
/gpfs1m	134TB	-	rw	yes	GPFS	Archive
/gpfs4m	34TB	-	rw	yes	GPFS	Archive
$TMPDIR	255GB	-	rw	NO	EXT4	io
$WORKDIR	34TB	-	rw	NO	GPFS	io

Environment Modules

Environment Modules is a very useful to manage environment variables of each application and it is very easy to use. Each module, loads all environment needed by a certain application and its dependencies automatically.

- **module avail** - lists available modules
- **module show module_name** - displays full information about the module with name *module_name*.
- **module load module_name** - loads the module with name *module_name* and its dependencies.
- **module unload module_name** - unload the module with name *module_name* and its dependencies.
- **module list** - list all modules currently loaded.

Available Compilers

Compiler	Intel	GNU	PGI
Fortran77	ifort	g77	pgf77
Fortran90	ifort	gfortran	pgf90
Fortran95	ifort	gfortran	pgf95
C	icc	gcc	pgcc
C++	icpc	g++	pgCC
Debug	idb	gdb	pgdbg
Profile	vtune	gprof	gpprof

Available MPIs

MPI	version
OpenMPI	1.4,1.6
MPICH2	1.4,1.5
PlatformMPI	08.02
MVAPICH2	1.9a2

Optimization flags

Compiler	Intel	GNU	PGI
High Opt.	-fast	-O3 -ffast-math	-fast -Mipa=fast,inline
OpenMP	-openmp	-fopenmp	-mp=nonuma
Debug	-g	-g	-g
Profile	-p	-pg	-p
Gulftown	-mtarget	-march=corei7	-tp=nehalem-64
SandyBridge	-mtarget	-march=corei7-avx	-tp=sandybridge-64
SSE	-xsse4.2	-msse4.2	-Mvect=[prefetch,sse]
AVX[1]	-xavx	-mavx	-fast

1. Advanced Vector Extension (AVX) streaming SIMD instructions. Sandy Bridge processor only.

Link MKL with OpenMPI, CDFT, ScaLAPACK, BLACS and Intel Compilers

Link line:
```
-L${MKLROOT}/lib/intel64 -lmkl_scalapack_ilp64 \
-lmkl_cdft_core -lmkl_intel_ilp64 -lmkl_sequential \
-lmkl_core -lmkl_blacs_intelmpi_ilp64 -lpthread -lm
```
Compiler options: `-DMKL_ILP64 -I${MKLROOT}/include`
More information at `http://software.intel.com/sites/products/mkl/`

Link MKL with OpenMP and Intel compilers

Link line:
```
-L${MKLROOT}/lib/intel64 -lmkl_intel_ilp64 \
-lmkl_intel_thread -lmkl_core -lpthread -lm
```
Compiler options: `-openmp -DMKL_ILP64 -I${MKLROOT}/include`

Get Involved!

We would like to encourage you to send suggestions and feedback to the NeSI Team.

OpenSSH Access

Parameters

- HOST: login.uoa.nesi.org.nz
- PORT: 22

Suggested Software

- Windows: mobaxterm, Putty, Bitvise Tunnelier
- MacOSX: Terminal, iTerm2
- Linux: Konsole, GnomeTerminal, yakuake

Remote File System Access

In order to access the file system (/home) remotely from your machine, we recommend:

- **SSHFS** (MacOSX) : `http://code.google.com/p/macfuse/`
- **SSHFS** (Linux) : `http://fuse.sourceforge.net/sshfs.html`
- **SSHFS** (Windows) : `http://code.google.com/p/win-sshfs/`
- **Konqueror** (KDE) : type fish://user@host:port
- **Nautilus** (Gnome) : type sftp://user@host:port
- **WinSCP** (Windows) : `http://winscp.net`

Remote File System Transfer with RSYNC (Unix Only)

RSYNC over SSH protocol is the best choice to transfer big data volumes.

- Transfer data from your machine to the server:
  `rsync -avHl /path/origin/* sshserver:/path/destination/`
- Transfer data from the server to your machine:
  `rsync -avHl sshserver:/path/destination/* /path/origin/`

Remote File System Transfer with scp/sftp (Unix Only)

SCP and **SFTP** are the most popular software to transfer data across the SSH protocol.

- **SCP**: `scp -pr sshserver:/path/destination/* path/destination/`
- **SFTP**: `sftp sshserver:/path/destination/* path/destination/`

LoadLeveler Commands

- **llsubmit** sends a new job to the scheduler for execution
- **llclass** show the current list of classes.
- **llq** returns queue information
  - `llq -l <job id>` gives extended job record information.
  - `llq -s <job id>` gives specific information about why the job is held.
  - `llq -u <user name>` lists only jobs initiated by the user.
- **llcancel** kills a queued job: `llcancel <job id>`

More information about LoadLeveler in User Manual :
`http://www.redbooks.ibm.com/redbooks/pdfs/sg246038.pdf`

Limits

Name	MaxJobCPU	Description
small1h	41	Jobs running for under 1 hour (high priority)
small6h	41	Jobs running for under 6 hours (high priority)
medium	12500	Jobs running for under 1 week (medium priority)
long	unlimited	Jobs running for longer than a week (low priority)

Quick Reference Guide to the High Performance Computer Environments @ NeSI

New Zealand eScience Infrastructure
Center for eResearch - University of Auckland, New Zealand
Building 409, Rm G21, LG, 24 Symonds Street - (1010) Auckland
Phone : (+64) 9 373 7599 - Support list: eresearch-support@auckland.ac.nz

SubmitScript Example : SMP job using local disk

```
#!/bin/bash
# Gaussian SubmitScript
# Optimized for run parallel job of 12 Cores in NeSI (Pandora-Gulftown)
#################################################################################
#@ job_name = Gaussian
#@ class = default
#@ notification = never
#@ group = nesi
#@ account_no = uoa
#@ wall_clock_limit = 1:00
#@ initialdir = $(home)
#@ output = $(home)/$(job_name).txt
#@ error = $(home)/$(job_name).err
#@ job_type = serial
#@ resources = ConsumableMemory(2048mb) ConsumableVirtualMemory(2048mb)
#@ parallel_threads = 12
#@ environment = COPY_ALL,OMP_NUM_THREADS=12
#@ queue
#################################################################################
###  Load the Enviroment Modules for Gromacs 4.5.4
. /etc/profile.d/modules.sh
module load g09/C.01
#################################################################################
###  Transfering the data to the local disk  ($TMPDIR)
cd $TMPDIR
cp -r $HOME/Gaussian/h2o_opt.dat .
setenv GAUSS_SCRDIR $TMPDIR
#################################################################################
###  Run the Parallel Program
smpexec g09 < ./h2o_opt.dat > h2ol_opt.log
#################################################################################
###  Transfering the results to the home directory ($HOME)
cp -pr $TMPDIR $HOME/results/
```

Most Outstanding OpenMPI (mpirun) Options

To find the full list of the available component types under the MCA architecture load the OpenMPI module and execute : `ompi_info --param all all`. More information at **OpenMPI WebSite**

- `-c, -n, --n, -np <#>` Run this many copies of the program on the given nodes.
- `-npersocket <#persocket>` Launch this processes for each sockets on the node.
- `-pernode` On each node, launch one process.
- `--app <appfile>` Provide an appfile, ignoring all other command line options.
- `-bind-to-core, -bind-to-socket` For process binding
- `-debug` Invoke the user-level debugger.
- `-debugger` Sequence of debuggers to search for when -debug is used.
- `-mca, --mca <key> <value>` Send arguments to various **MCA modules**.
  - `btl self,sm,openib` Select which network to use for MPI communications
  - `btl_sm_use_knem 1` Use KNEM for intra-node MPI.
  - `coll_sm_control_size` Buffer size (in bytes) chosen to optimize collective operations
  - `coll_fca_enable 1` Use Fabric Collective Accelerator (Only Mellanox)
  - `coll_base_verbose 1` Verbosity level for the coll framework
  - `coll_tuned_use_dynamic_rules 1` Dynamic optimization of collective operations
  - `coll_tuned_alltoall_algorithm 3` AllToAll Collective communication algorithm
  - `coll_tuned_allreduce_algorithm 4` AllReduce Collective communication algorithm
  - `mtl mxm` Use MellanoX Messaging library which provides enhancements to parallel communication libraries.

Remember that you don't need to setup this options if you are using the MPIRUN wrapper!

Submit Script Syntax

- **account_no** = (uoa, uoc, uoo, landcare or nesiXXXX)
- **blocking** = (unlimited or N). Where N is the number of tasks allocated to nodes
- **group** = Identifies what group the job submission is done under.
  - nesi - users without a particular NeSI project
  - chemistry - University of Auckland Chemistry Department
  - pd - NeSI Proposal Development allocations
  - unfunded - NeSI Research allocations without NeSI HPC funding
  - funded - NeSI Research allocations with NeSI HPC funding
- **job_name** Sets the job name and the $job_name macro
- **job_type** Specifies the job type for the submission.
  - Serial - single- or multi-threaded job, SMP, confined within a single node
  - MPI - Distributed Parallel Environment (includes OpenMPI 1.4 but not 1.6)
  - MPICH - Distributed Parallel Environment includes MPICH and OpenMPI 1.6
- **notification** Specifies under which conditions Loadleveler will send out email notifications to the specified address about the state of the job:
  - always - notifies about every change in the job status
  - start - notifies when the job is activated
  - complete - notifies when the job ends
  - error - notifies only if the job fails
  - never - never send any notifications
- **notify_user** user email
- **parallel_threads** Sets the number of cores to be allocated for each task
- **resources** Specifies what consumable resources the job requires to run per task:
  - ConsumableMemory - the amount of physical memory required.
  - ConsumableVirtualMemory - the amount of virtual memory required.
  - GPUDev - the number of GPU devices required (max 2 GPU).
  - AnsysLicenses - the number of ANSYS licenses to check out (max 40).
- **requirements** Can be (Feature=="sandybridge") or (Feature=="westmere")
- **tasks_per_node** specifies the total number of tasks (cores) to be run on each available node.
- **total_tasks** specifies how many tasks (cores) are to be run in total
- **wall_clock_limit** Sets the limit for the elapsed time for which a job can run

Useful MPIRUN Wrapper

Inside each application module file there is a useful command wrapper of mpirun that can simplify its usage. In your script you only have to write:

```
MPIRUN binary <aplication options>
```

Instead of:

```
mpirun -np 64 -mca btl self,sm,openib \
        -mca btl_sm_use_knem 1        \
        -mca mpi_paffinity_alone 1    \
        --bind-to-core                \
        binary <aplication options>
```

The **MPIRUN** Wrapper can be setup in the module file using the following syntax:

```
set-alias  MPIRUN  "mpirun -np \$LOADL_TOTAL_TASKS OPTIONS \$@"
```

Recommended Best Practices

- Use the **SubmitScripts Templates** that are in /share/SubmitScripts.
- Use the **smpexec** for SMP jobs in order to use the core binding capabilities
  `smpexec myapplication argument1 argument2`
- Ask for the minimum nodes for the specified number of cores for the very large jobs. It will take more time to enter in execution but will finish with less walltime.

SubmitScript Example : MPI job using Shared FileSystem

```
#!/bin/bash
# GROMACS SubmitScript
# Optimized for run parallel job of 64 Cores in NeSI (Pandora-SandyBridge)
#################################################################################
#@ job_name = Gromacs_9LDT-64
#@ class = default
#@ group = nesi
#@ notification = never
#@ account_no = uoa
#@ wall_clock_limit = 6:00:00
#@ resources = ConsumableMemory(4096mb) ConsumableVirtualMemory(4096mb)
#@ job_type = MPICH
#@ initialdir = $(home)
#@ output = $(job_name).$(jobid).out
#@ error = $(job_name).$(jobid).err
#@ requirements = (Feature=="sandybridge")
#@ tasks_per_node = 64
#@ node = 4
#@ queue
#################################################################################
###  Load the Enviroment Modules for Gromacs 4.5.4
. /etc/profile.d/modules.sh
module load gromacs/4.5.4
#################################################################################
###  Transfering the data to the local disk  ($TMPDIR)
cd $WORKDIR
cp $HOME/Gromacs_9LDT/input/* .
#################################################################################
###  Run the Parallel Program
export OMP_NUM_THREADS=1
grompp -f full_vdw.mdp -c 9LDT-pt-md-3.gro -p 9LDT-bu.top -o 9LDT-bu.tpr
MPIRUN  mdrun_mpi -v -s 9LDT-bu.tpr -o 9LDT-bu.trr > mdrun.out
#################################################################################
###  Transfering the results to the home directory ($HOME)
cp -pr $WORKDIR $HOME/results/
```

GriSu

- u