Secrets to Slikrock Success

Thursday Rabbit Diversion
March 29, 2007

Don Dazlich

Introduction

* Slikrock is a cluster running Mac OS X. It has 52 nodes each with two processors. The 'head' is slikrock and the others are sr1 through sr51.
* One runs a parallel job using mpi (lam-mpi version) and the xlf compiler. One needs /usr/local/xlf/bin in the path, or addresses the compiler directly at /usr/local/xlf/bin/mpif77. mpif77 finds all the mpi include files and libraries without explicitly specifying them
* Other combinations of different MPIs and compilers are available...

Compilation example from Bugs5

Terminal — ssh — 80x24

```
#following local systems:
#Libs     = -L/usr/local/xlf/lib -lnetcdf
#Inc      = -I/usr/local/xlf/include

#mac osx xlf/lammpi, optimized:
 Cpp      = /usr/bin/Cpp3
 Cppopts  = -P -traditional
 Comp     = /usr/local/xlf/bin/mpif77
 Opts     = -c -O3 -bmaxdata:2000000000 -qfloat=fltint:rsqrt -qzerosize
 pblOpts     = -c -O3 -bmaxdata:2000000000 -qfloat=fltint:rsqrt -qzerosize
 progOpts    = -c -O3 -bmaxdata:2000000000 -qfloat=fltint:rsqrt -qzerosize
 Load     = /usr/local/xlf/bin/mpif77 -L/usr/lib/gcc/darwin/3.3 -L/usr/lib/gcc/da
rwin -L/usr/libexec/gcc/darwin/ppc/3.3/../../..
 Fix      = -qfixed=132
 Free     = -qfree=f90
 Dp       =
 Suf      = f
#following local systems:
 Libs     = -L/usr/local/xlf/lib -lnetcdf
 Inc      = -I/usr/local/xlf/include

# MacPro Intel, optimized (with openmpi):
#Cpp      = /usr/bin/Cpp
```

Launching an mpi job

* Reserve some nodes - http://kiwi/slikrock/ displays the status of the nodes and how to reserve.
* Create a lamhostfile:

```
sr5 cpu=2
sr6 cpu=2
sr7 cpu=2
sr8 cpu=2
slikrock cpu=2
~
```

* Launch the lammpi daemon: `lamboot lamhostfile`
* `mpirun -np n executable &`
* When job is over: `lamhalt`

Issue # 1 - No Batch System

* Normally, a batch system will start your executable in the proper directory. However, launching mpi directly places you in the home directory on the remote node.
* This requires the user to do two things:
  * 1. The full path of the executable must be used when launching mpirun:

```
mpirun -np 8 /xraid2/dazlich/BUGS5/run_gcm/bugs5
```

  * 2. A full rather than relative path must be used to open files within the code.

Issue # 1 - No Batch System (cont.)

* In BUGS5 there is a character string where one can set the full path of the directory above the data directory in kinds.F.

```
!      character*40, parameter :: datapath =
!     & './'
       character*40, parameter :: datapath =
      & '/scratch/guru/run_gcm/'

       end module kinds

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
~
```

Issue # 2 - file write errors to raid

* We are strongly encouraged (as in required!) to use the raid file systems, /xraid1 and /xraid2 because the home system on slikrock is small.
* However, many i/o intensive applications run there have crashed there non-reproducibly when writing, typically saying something like 'file too large.'
* The solution is to use the local disks on the nodes. This means copying input data out there before running, and copying it back when the job is over (and cleaning up after oneself).

Issue # 2 - file write errors to raid (cont.)

* Kelley has created a /scratch directory on each node. The user can create directories within /scratch.
* Don has a template script for doing operations on the remote nodes - doremote. He keeps his in ~/bin (and has this in his path

doremote script

Terminal — ssh — 80x24

```
#
foreach node (5 6 7 8)
  echo 'node, ' ${node}

# set up remote nodes for job (after gathering input data on /xraid)
#  ssh sr${node} "mkdir /scratch/dazlich"
#  scp -r /xraid2/dazlich/BUGS5/run_gcm sr${node}:/scratch/dazlich

# monitor remote disks
#  ssh sr${node} "ls -l /scratch/dazlich/run_gcm"

# clean up after job
# scp -r sr${node}:/scratch/dazlich/run_gcm/qp_output /xraid2/dazlich/BUGS5/run_
gcm
# scp -r sr${node}:/scratch/dazlich/run_gcm/pbp_output /xraid2/dazlich/BUGS5/run
_gcm
# scp -r sr${node}:/scratch/dazlich/run_gcm/hf_output /xraid2/dazlich/BUGS5/run_
gcm
# scp -r sr${node}:/scratch/dazlich/run_gcm/restarts /xraid2/dazlich/BUGS5/run_g
cm
#  ssh sr${node} "rm -r /scratch/dazlich/run_gcm"
end
~
"~/bin/doremote" 18L, 714C written
```

Summary

* Reserve your nodes
* Set up your lamhostfile
* Export your input data and output directories to the remote nodes
* Launch lam daemon (lamboot)
* Run your job (mpirun)
* Bring your data back to xraid
* Clean up - lamhalt; remote disks (/scratch)