MM5 on SGI IA32 Clusters

Elizabeth Hayes
SGI
Email: eah@sgi.com

June 21, 2000

Introduction

The fifth-generation Penn State/NCAR Mesoscale Model (MM5)[1] is the numerical weather prediction (NWP) component of the MM5 weather modeling system. This paper demonstrates the use of the same source parallel[2] version of MM5 using MPICH over Ethernet or Myrinet switches on clusters of SGI 1200L™. The SGI 1200L™ compute nodes are shared-memory multiprocessors (SMPs) built using two Intel processors, the Linux™ operating system with SGI ProPack overlay, and SGI Linux Advanced Cluster Environment[3].

SGI 1200L Cluster

SGI 1200L™ configurations include 1 or 2 Pentium® III, 256KB L2 Cache on Chip, up to 2GB SDRAM Memory (800MB/sec), 2 PCI Slots, internal Ultra2 SCSI, and SGI FibreVault. For this report, each node of the cluster consists of a two processor 700 Mhz IA32 SGI 1200L™. The hinv command on each node displays the following hardware description: two Intel Pentium III (Coppermine) processors, main memory 1024 MB, and two SCSI disks. The machines were running RedHat 6.2, SGI ProPack 1.3 and SGI Linux Advanced Cluster Environment (ACE). ACE includes MPICH[4], MPI over GM, PBS[5], Performance Co-Pilot, and an installation manager to enable cluster components to be installed and configured and to remotely reset nodes from the management node. The nodes were connected with a 100Mb Ethernet switch and a Myrinet[6] switch. The Myrinet switch on this cluster has 2 ports available for GM and 1 port available for TCP. MPICH and MPICH over GM are installed by ACE using the G77 and GCC compilers. For MM5, the Portland Group PGF77[7] compiler version 3.1-3 and GNU compiler GCC version 2.91.66 were used. The cost of the cluster configuration varies depending on what interconnect switch (Ethernet or Myrinet) is used.

Building MM5 to Run on SGI 1200L Cluster

Modifications to configure.user were made to build MM5 MPP version 3-3 for the SGI 1200L™ Linux cluster. Two configuration sections were added with the following changes:

```
7h.  Ethernet Linux PCs - pgf77, pgcc and MPICH
LINUX_MPIHOME=/shared/local/pgi/linux86
MFC = $(LINUX_MPIHOME)/bin/pgf77
MCC = $(LINUX_MPIHOME)/bin/pgcc
MLD = $(LINUX_MPIHOME)/bin/pgf77
LOCAL_LIBRARIES=-L/$(LINUX_MPIHOME)/lib -lmpich
CFLAGS=-DMPI -I$(LINUX_MPIHOME)/include

Modified /MPP/RSL/RSL
IDIR=$(LINUX_MPIHOME)/include
CC=$(LINUX_MPIHOME)/pgcc
FC=$(LINUX_MPIHOME)/pgf77 -byteswapio

7i. Myrinet Linux PCs - pgf77, pgcc and MPICH
LINUX_MPIHOME=/shared/home/eah
LINUX_MPIHOME2=/shared/local/mpich_over_gm
MFC = $(LINUX_MPIHOME)/bin/mpif77
MCC = $(LINUX_MPIHOME)/bin/mpicc
MLD=$(LINUX_MPIHOME)/bin/mpif77
LOCAL_LIBRARIES=-L($LINUX_MPIHOME2)/lib -lfmpich -
lmpich
CFLAGS =-D MPI -I$(LINUX_MPIHOME2)/include
```

For MPICH over GM, mpif77 and mpicc were built using the G77 and CC compilers. The mpicc and mpif77 scripts were copied to LINUX_MPIHOME and modified to use pgf77 and pgcc. Two undefined references were found mpi_init_ and f__xargc. These were fixed by modifying rsl_mpi_compat.c to change mpi_init_ to mpi_init__ and by setting f__xargc equal to xargc.

Run Scripts for MM5 on SGI 1200L Ethernet Cluster

Example script for run on cluster over Ethernet:

```
foreach i (2.pe.ns_2.pe.ew)
foreach j (1.pe.per.node)
date
echo run_mpp.$i.$j.output
mkdir run_mpp.$i.$j.output
```

```
cd run_mpp.$i.$j.output
../link.it
echo mm5.mpp.ethernet.$i
cp ../mm5.mpp.ethernet.$i mm5.mpp
rdist -f ./dist/distfile.$i.$j
time /shared/local/pgi/linux86/bin/mpirun -v -np 4 -machinefile machinefile mm5.mpp
cd ..
end
end
```

Example distfile for 4 processor run using 1 processor per node:

```
host= (tech3 tech4 tech5 tech6)
files=
(/data2/eah/mm5/run/run_mpp.2.pe.ns_2.pe.ew.1.pe.per.node.output /data2/eah/mm5/run/machinefile)
```

Example Machinefile for a 4 processor run using 1 processors per node

```
tech3
tech4
tech5
tech6
```

Example Machinefile for a 4 processor run using 2 processors per node:

```
tech3
tech3
tech4
tech4
```

Run Scripts for MM5 on SGI 1200L Myrinet Cluster

Example script for run on cluster over Myrinet:

```
foreach i (2.pe.ns_2.pe.ew)
foreach j (1.pe.per.node)
date
echo run_mpp.$i.$j.output
mkdir run_mpp.$i.$j.output
cd run_mpp.$i.$j.output
../link.it
echo mm5.mpp.myrinet.$i
cp ../mm5.mpp.myrinet.$i mm5.mpp
rdist -f ./dist/distfile.$i.$j
time /shared/local/mpich_over_gm/bin/mpirun.ch_gm --gm-f ./.gmpi/conf.$i.$j -np 4 mm5.mpp
cd ..
end
end
```

Example .gmpi/conf file for a 4 processor run using 1 processor per node:

```
4
tech3.houst.sgi.com 2
tech4.houst.sgi.com 2
tech5.houst.sgi.com 2
tech6.houst.sgi.com 2
```

Example .gmpi/conf file for run on 2 processors per node:

```
4
tech3.houst.sgi.com 2
tech4.houst.sgi.com 2
tech3.houst.sgi.com 4
tech4.houst.sgi.com 4
```

Run Management on Ethernet Cluster

Running MM5 on a distributed file system across Ethernet requires that the run script be launched from the head node, and that head node be used as the first processor for the parallel job. If the mpirun -nolocal flag was used, and the node from which mpirun was launched was not in the machinefile, then the run would not start successfully. The -nolocal flag specifies that the run be performed using the nodes listed in the machine file. The PI* file that is created by the p4 process and used by MPICH over Ethernet cannot be found on a remote disk, and the run will fail. The PI* file is placed in the run directory after the mm5.mpp is submitted using mpirun. Jobs launched from nodes other than the head node do not run. If MM5 is launched from an NFS-mounted file system that can be seen by all of the nodes, then this problem does not occur. The mpirun -nolocal flag will result in a successful run using the nodes specified in the machine file if executed from a shared disk.

Run Management on Myrinet Cluster

Placement of jobs across the cluster is done using the ~/.gmpi/conf file unless otherwise specified. In the example run script above, a flag is used to point to the configuration file on the distributed file system. The rdist command was used to distribute the input files, executables, and conf file to each node used by the run. On both the Myrinet and Ethernet clusters killing the mpirun job on the head node does not always result in the mm5.mpp processes being killed on the remote nodes. A remote login and killing jobs with kill -9 (process-id) are needed to kill off all remaining processes

System Administration

Situations may arise that need to be fixed on a per node basis. An example of this is when a node or nodes are hung and need to be restarted.

Occasionally NFS errors will occur on IA32 Linux clusters. NFS errors may cause the system to hang if home directories are NFS mounted. A typical error includes: "kernel: nfs: task 3120 can't get a request slot". A simple reboot is not sufficient, as kernel errors about rebooting may be generated, so a physical reset is often necessary.

Performance Comparisons of MM5 on SGI 1200L Cluster

Table 1 shows the performance of MM5 on IA32 Ethernet Cluster on shared versus distributed file system for the 36km (112x136x33) no nest benchmark.

MM5 Benchmark on Shared vs. Distributed File System on an SGI 1200L IA32 Ethernet Cluster		
	Shared (NFS)	Distributed
#CPUs	Wall Clock (Sec)	Wall Clock (Sec)
2		1463
4	872	793
8	575	473
16*	389	305

Table 2 shows the performance of MM5v3.3 on IA32 Myrinet Cluster using 1 processor per node versus 2 processors per node for the 36km (112x136x33), no nest Benchmark

.

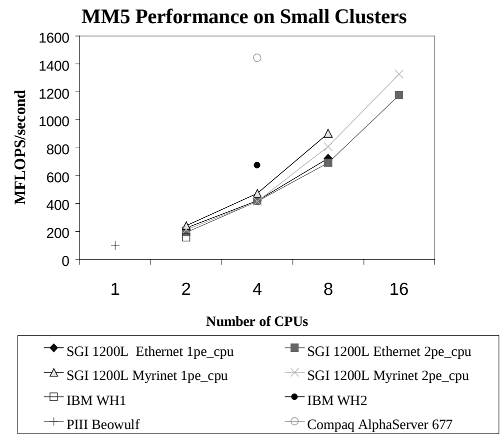
MM5 partial node versus full node benchmark on an SGI 1200L IA32 Myrinet Cluster		
	1 processor per node	2 processors per node
#CPUs	Wall Clock (Sec)	Wall Clock (Sec)
2	1375	1567
4	714	809
8	386	425
16		276

Table 3 shows the performance of MM5 on IA32 Cluster using Ethernet versus Myrinet for the 36km (112x136x33) no nest benchmark

MM5 Benchmark on an SGI 1200L IA32 Ethernet versus Myrinet Cluster		
	Ethernet	Myrinet
#CPUs	Wall Clock (Sec)	Wall Clock (Sec)
1		2697
2	1463	1375
4	793	714
8	473	386
16*	305	276

* using two processors per node

Figure 1 shows the comparitive performance of MM5 on small clusters, other vendor numbers were obtained from the MM5 MPP helpdesk web site[8]:



Conclusions

The SGI 1200L™ cluster with 16 processors has been demonstrated to run the MM5 36km benchmark using both Ethernet and Myrinet interconnects, shared and distributed file systems, and using partial and full node configurations. The increased cost of Myrinet is not justified on small clusters, as the performance improvement is not significant at low number of processors. The use of NFS-mounted file systems assists the user with file management but results in a degradation in performance. The performance of MM5 on the IA32 cluster does not degrade significantly when run using all of the processers per node as compared to partial node performance. The SGI 1200L™ cluster package includes cluster performance monitoring, job scheduling, compilers, and MPICH. The performance of the SGI 1200L™ cluster is favorable compared to the IBM WH1, and is a lower cost solution than the IBM WH2.

[1] Grell, G.A., J. Dudhia, and D. R. Stauffer, 1995: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Technical Note NCAT/TN-398+STR, 138 pp.
[2] John Michalakes, 1998: *The Same-Source Parallel MM5*. In proceeding of the *Second International Workshop on Software Engineering and Code Design in Parallel Meteorological and Oceanographic Applications.*
[3] ACE http://oss.sgi.com/ace
[4] MPICH http://www-unix.mcs.anl.gov/mpi/mpich
[5] PBS http://pbs.mrj.com
[6] Myrinet http://www.myri.com
[7] The Portland Group http://www.pgroup.com
[8] MM5 Performance Plot, Jan-June 2000 http://www.mmm.ucar.edu/mm5/mpp/helpdesk