Performance of 16 node Scali PIII based systems

using Dolphin ICS 32 bit / 33 MHz PCI-SCI cards (311/312)

Overview

ScaBench is Scali's test suite to benchmark MPI (Message Passing Interface - www-unix.mcs.anl.gov/mpi) based workstation clusters. The tests are divided into single node tests, point-to-point communication tests and collective communication tests. Since communication tests are based on user-space MPI communication, all tests indicate application-to-application performance. Scali's MPI implementation, ScaMPI™, is examined.

System description

The system consisted of 16 nodes interconnected with Dolphin SCI boards (www.dolphinics.com). The system is organized in a 4x4 2D torus topology. More details are given at the end of this document.

Point to point performance

Point to point performance is measured between two processes within the same node (memory performance) and between two nodes (network performance). *mpptest* from the MPICH test and verification suite is used to measure performance. This program is available as a part of the MPICH distribution (www-unix.mcs.anl.gov/mpi/mpich). *mpperf* is Scali's extension to mpptest

One-way communication

The one-way test uses one sender and one receiver to measure the send/receive bandwidth. The numbers reflect the expected single stream bandwidth when sending data from a process to another. This measurement is often cited as *ping-ping* performance. The numbers are produced using *mpperf.*

Bandwidth: 85.7 MByte/s (peak)
Latency: 6.5 µs (zero byte)

Node internal performance was measured to:

Bandwidth: 187.2 MByte/s (peak)
Latency: 1.5 µs (zero byte)

Concurrent two-way communication

In the two-way test both processes simultaneously send and receive equally sized blocks of data. The numbers reflect bandwidth when two processes exchange data. The numbers are produced using *mpperf*.

Bandwidth: 78.4 MByte/s (peak)
Latency: 5.4 µs (zero byte)

Ping-pong communication

In the ping-pong test two processes alternate sending and receiving data in a token passing style. The latency number is half the round-trip.

Bandwidth: 84.1 MByte/s (peak)
Latency: 6.3 µs (zero byte)
Latency: 10.5 µs (32 bytes)

Node internal performance was measured to:

Bandwidth: 133 MByte/s (peak)
Latency: 3.2 µs (zero byte)
Latency: 4.4 µs (32 bytes)

System wide performance

System wide performance is measured between all or a subset of the nodes in a system.

Bisection bandwidth

In the bisection test the complete system is logically divided into two subsystems and the aggregated bandwidth between the two subsystems is measured. An example of this is splitting the system in two and letting each node in one half communicate with a node in the other half on a one-to-one basis. The numbers are produced using *mpptest*.

The total aggregated application network bandwidth for 16 nodes was measured to:

Bandwidth: 601 MByte/s

Barrier Synchronization

A barrier makes a synchronization (rendezvous) point between all involved processes. The numbers are produced using *mpptest*:

Nodes	2	4	8	16
Time (µs)	8.9	18.1	30.2	39.9

All-to-all communication

MPI_Alltoall() exchanges unique chunks of data with all processes, itself included.The numbers are generated using a test program that performs successive calls to MPI_Alltoall(). The numbers reflect network traffic per node, hence the reduced performance for small configurations. The numbers are produced using *mpptest*:

The network bandwidth in MByte/s:

Nodes	2	4	8	16
Rate	67.7	69.9	69.5	68.2

NAS Parallel Benchmark

Nodes	2	4	8	16
Cray T3E	na	47.5	46.5	45.5
SGI O2k	53.5	43.2	41.6	39.4
Scali-1	48.1	40.8	38.4	37.8
Scali-2	74.7	72.4	63.5	59.5

The most communication intensive benchmark in NPB 2.3, is the FT kernel. The table compares the per node FT performance(Mop/s), for size A with Cray T3E-1200 and SGI Origin 2000 (195 MHz R10K). The number after the Scali entries indicate processes per node. Number for all the NPB can be found at Scali's website at: http://www.scali.com.

Contributor(s)

Lars P. Huse (lph@scali.no) ran the benchmarks on a system owned by Dolphin Interconnect Solutions AS (Nov. 2000)

Table 1: Hardware and OS

Item	Description	Details
Operating system	RedHat Linux 6.2	Kernel: 2.2.14-6.0.2smp
Network interface	Dolphin D311/D312	32 bit/33 MHz PCI-SCI card
Network topology	16 nodes in a 2D mesh	4 x 4 configuration
Mother board	Intel 440BX	82443BX AGP (rev 3)
Processor(s)	Dual Pentium III	450 MHz (Katmai)
Memory	256 Mbyte	No Bigphys area

Table 2: Installed software

Item	Description
Test source	ScaMPItst 1.9.4
Scali Software Platform	SSP 2.1 (ScaMPI 1.10 & ScaSCI 2.3.2)
Compiler	Gnu gcc (egcs-2.91.66)
Compiler switches	-O2 -mpentiumpro -malign-double -fno-omit-frame-pointer
MPI monitor switches	-init_comm_world -timeout 600