Performance of HPC Applications over InfiniBand, 10 Gb and 1 Gb Ethernet

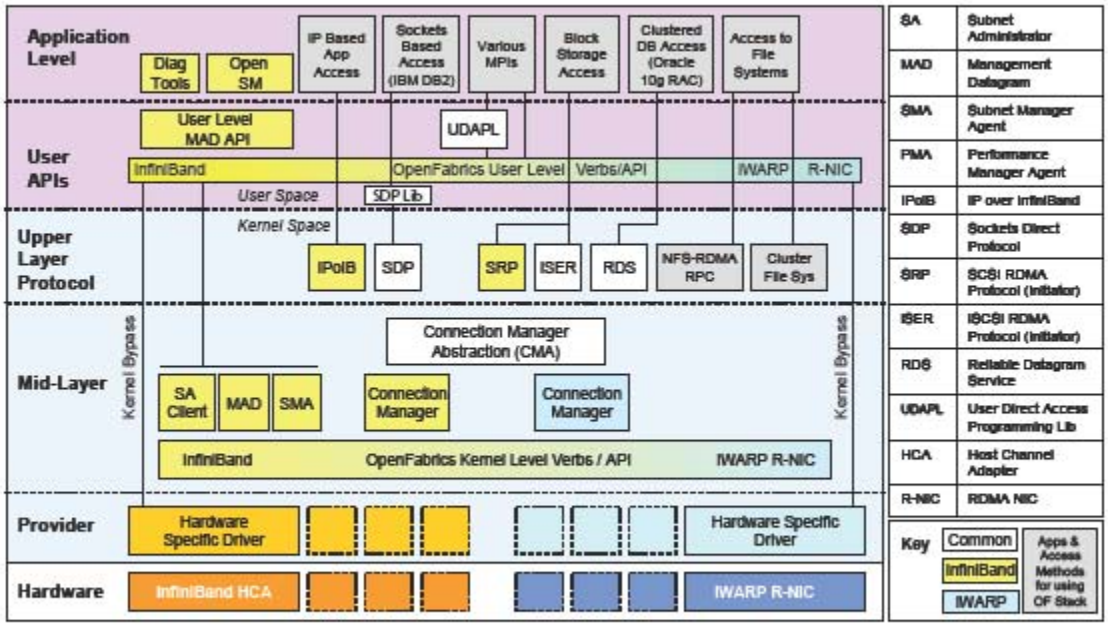
Swamy N. Kandadai and Xinghong He
swamy@us.ibm.com and xinghong@us.ibm.com

ABSTRACT:
We compare the performance of several applications in the High Performance Computing (HPC) using 4x InfiniBand DDR,10 Gigabit Ethernet using iWARP and on 1 Gigabit Ethernet using TCP/IP protocol. These applications are chosen from several areas in High Performance Computing. In addition, we also present the results of the TCP/IP and UDP performance of IP over IB, 10 Gigabit Ethernet and 1 Gigabit Ethernet for the NETPERF benchmark.

INTRODUCTION:
Adoption of InfiniBand Architecture (IBA) or iWARP (Internet Wide Area Remote Protocol) have been widely used in High Performance Computing segment. To simplify the implementation, Open Fabrics Alliance (OFA) is creating a single, open sourced interoperable software stack that supports both transports for Linux. InfiniBand is supported by the Open Fabrics Enterprise Distribution (OFED) software stack and Ethernet is supported by RDMA over Ethernet (iWARP) software stack.

Both InfiniBand and iWARP extensions of Ethernet support RDMA (Remote Direct Memory Access) and use different technologies. Figure 1 is a schematic of the software stack being developed by OFA.

Figure 1: Open Fabrics Linux software stack



Hardware:

The cluster is a 2.3 GHz quad-core Opteron processor 2356 (Barcelona). Each node has 16 GB memory and each node has two quad-core processors. Each node is configured with:

1. a 4x DDR InfiniBand connected to a 96-port Cisco SFS-7012 DDR switch. The InfiniBand HCA are the dual-port DDR ConnectX adapters from Mellanox.
2. a 10 Gigabit network using Chelsio adapters connected to a 20-port Force-10 10 Gigabit Ethernet switch. The Chelsio adapters are capable of TOE.
3. a 1 Gigabit Ethernet network connected to a Cisco switch.

Software:

The cluster is running RedHat Enterprise Linux Server release 5.2. The OFED is version 1.4.0. The 10 Gigabit TOE/NIC drivers are at level 1.2.1. We used MVAPICH2 version 1.2p1 for runs using InfiniBand and 10 Gigabit Ethernet networks. We built MVAPICH2 version 1.2p1 with PathScale compiler version 3.2 for use with 1 Gigabit Ethernet. The applications were built with PathScale compiler version 3.2 and used the same compiler flags for all three different network interfaces. In addition, we used ACML (AMD Core Math Library) version 4.1.0 and Goto BLAS version 1.26.

Applications:

We studied the following applications for the performance study described in this paper.

1. Netperf
2. Intel MPI benchmark (IMB)
3. Himeno benchmark
4. NAS parallel benchmarks (NPB)
5. NAMD
6. CPMD
7. HPCC
8. SPEC MPI

RESULTS:

NETPERF:

Netperf is a benchmark suite that can be used to measure various aspects of networking performance. Its primary focus is in bulk data transfer and request/response (RR) performance using either TCP/IP or UDP and the Berkeley Socket Interface (BSI). We used Netperf version 2.4.4. We present in Table 1 and 2 the TCP stream and RR performance for all three network interfaces: 4x DDR IB, 10 Gigabit Ethernet with TOE (TCP off-load engine) and 1 Gigabit Ethernet. In Tables 3 and 4, we present the results with UDP interface. As can be seen from these results, 10 Gigabit Ethernet network

outperforms the InfiniBand network for both TCP and UDP interfaces. As expected, the 1 Gigabit Ethernet performance is much lower compared to the other two networks.

Table 1: TCP Stream Performance

Receive socket size, bytes	Send socket size, bytes	Send message size, bytes	Throughput, Mbits/second		
			IB	10GE	1 GE
262144	262144	4096	8258	9494	949
262144	262144	8192	7903	9494	949
262144	262144	32768	7723	9494	949
114688	114688	4096	6199	9493	948
114688	114688	8192	5893	9493	948
114688	114688	32768	5872	9493	948
65536	65536	4096	4546	9493	948
65536	65536	8192	4507	9493	948
65536	65536	32768	4400	9493	948
16384	16384	4096	1538	9492	495

Table 2: TCP Request/Response performance:

Local/Remote Socket send size, bytes	Local/Remote Socket Recv size, bytes	Request size, bytes	Receive size, bytes	Transactions per second		
				IB	10GE	1GE
2048/2048	256/256	1	1	25353	55239	11712
2048/2048	256/256	64	64	25070	51383	11293
2048/2048	256/256	100	200	24723	44090	10961
2048/2048	256/256	128	8192	1580	20326	0.3

Table 3: UDP Stream performance:

Socket size, bytes	Message size,bytes	Throughput, Mbits/second		
		IB	10GE	1GE
65536/65536	64	407	420	397
65536/65536	1024	5294	5613	939
65536/65536	1472	7185	7410	957

Table 4: UDP Request/Response performance:

Local/remote send size, bytes	Local/Remote Receive size, bytes	Request/Response size, bytes	Transactions per second		
			IB	10GE	1GE
126976/126976	126976/126976	1/1	28184	52578	11997

126976/126976	126976/126976	64/64	27793	44511	11465
126976/126976	126976/126976	100/200	27475	43527	11110
126976/126976	126976/126976	1024/1024	25327	32718	7679

Intel MPI benchmark:

The Intel MPI benchmark (IMB) is a widely used benchmark to compare the performance of various computing platforms and/or MPI implementations. We present in Table 5, the performance of point-to-point and exchange bandwidth by running 1 MPI task on each node and also the latency.

Table 5: Bandwidth and Latency

	IB-DDR	10Gigabit	1 Gigabit
Ping-Pong bandwidth, MB/s	1466	1000	112.5
Exchange bandwidth, MB/s	2659	2073	157.6
Latency, us	2.01	8.23	46.52

The latency of 10 Gigabit Ethernet using iWARP layer is higher compared to InfiniBand network using Verbs layer.

We present the performance of point-to-point communication in Figures 2 and 3 and the collective communication performance in Figures 4 and 5.

Figure 2: Ping-Pong Bandwidth

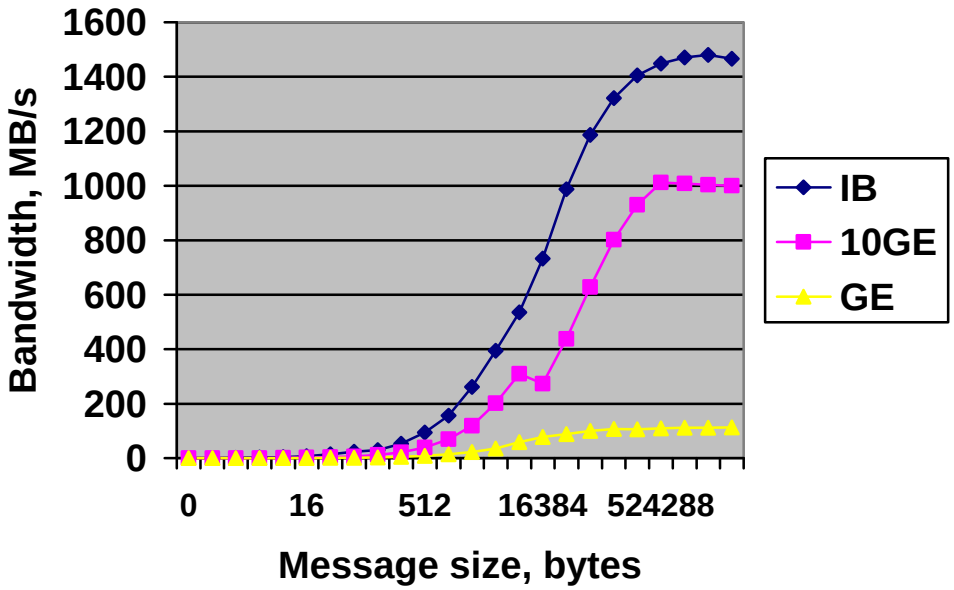


Figure 3: Exchange Bandwidth:

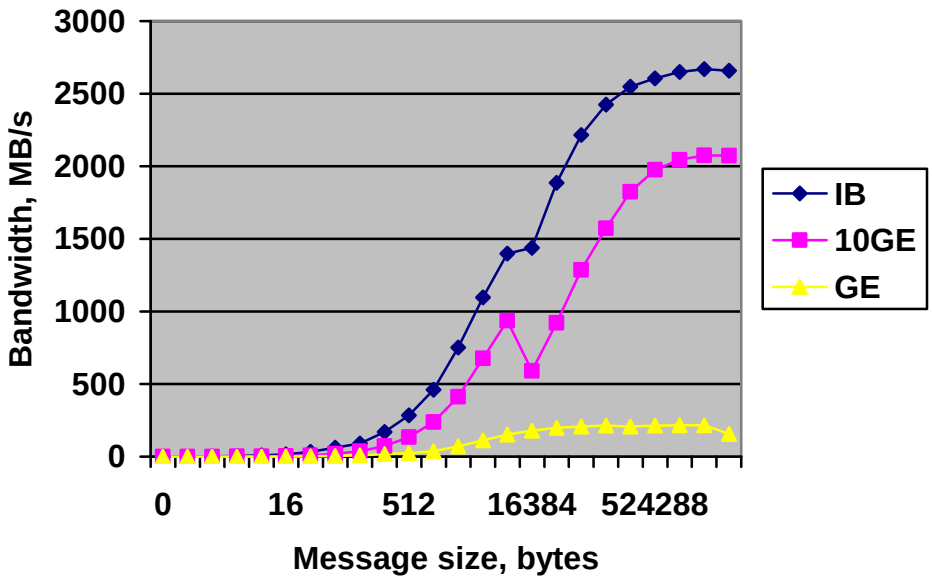


Figure 4: Broadcast performance:

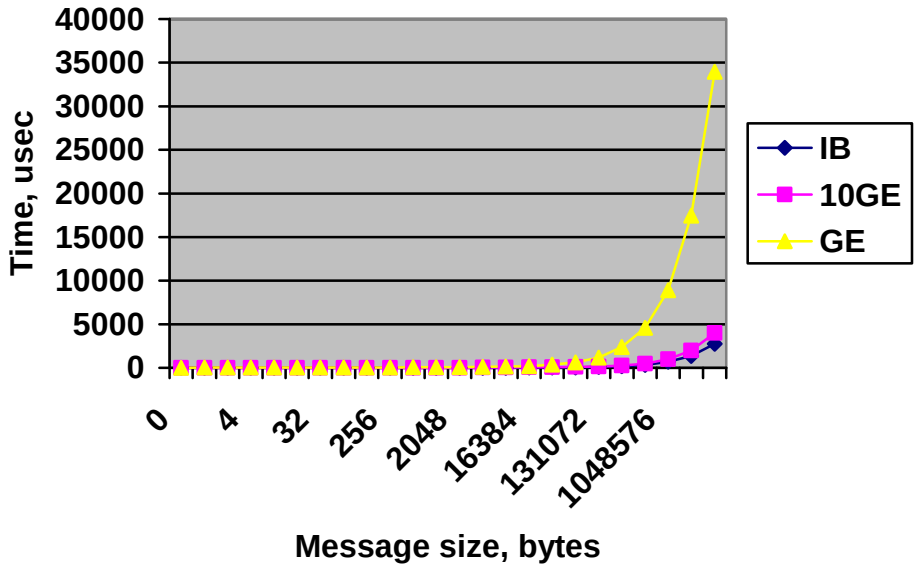
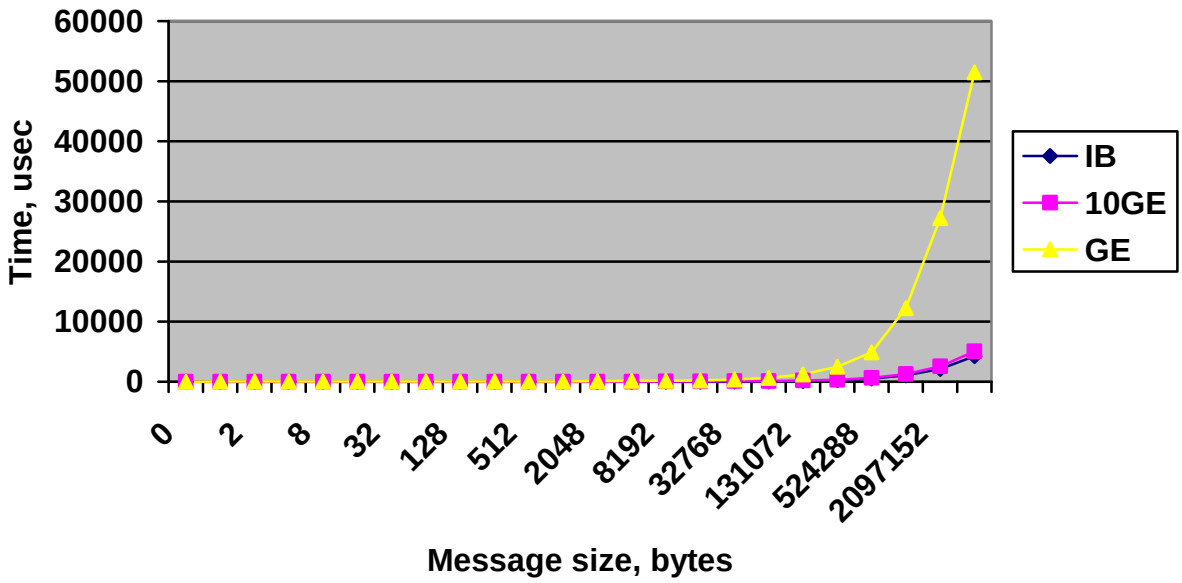


Figure 5: All-to-All Performance:



HIMENO benchmark:

This benchmark measures the performance using a kernel in a linear solver of pressure Poisson equations which appear in an incompressible Navier-Stokes solver. A point Jacobi method is employed in this solver.

We chose a larger problem size with 513 x 257 x 257 grid size in this study. We present below the performance in GFlops measured.

Table 6: HIMENO benchmark

No. of MPI tasks	Performance, GFlops		
	IB	10 Gigabit	Gigabit
128	72.8	65.6	35.4
64	46.8	36.1	26.6
32	21.8	25.7	15.1
16	11.7	12.8	7.8

From the table above, for small MPI tasks, the performance of 10 Gigabit is better compared to IB while for larger MPI tasks, IB performs better. As expected, the performance of both IB and 10 Gigabit Ethernet is better compared to 1 Gigabit Ethernet.

NAS Parallel benchmarks:

NAS (Numerical Aerodynamic Simulations) parallel benchmarks consist of eight programs. The first five (EP, FT, IS, MG and CG) are kernel benchmarks with simple data structures. The simulated application benchmarks which compute the solution to the nonlinear partial differential equations are: LU (LU decomposition), SP (Scalar Pentadiagonal) and BT (Block Tridiagonal). Five different classes (A, B, C,D and E) of problems are defined depending on the size of the problem. We selected two classes C and D in this study as classes A and B are too small to run on a parallel system and Class E is too big for this cluster. We used NPB version 3.3 in this study.

Table 7: NAS parallel benchmarks for Class C:

Application	IB		10 Gigabit		1 Gigabit	
	Time, sec	Mops	Time, sec	Mops	Time, sec	Mops
ft.C.128	8.36	47423	7.89	50249	42.76	9270
ft.C.64	14.34	27646	14.80	26777	82.32	4815
ft.C.32	27.66	14330	28.51	13903	131.77	3008
lu.C.128	21.45	95045	19.66	103717	54.08	37706
lu.C.64	34.27	59491	34.23	59571	73.58	27712
lu.C.32	70.84	28785	71.10	28676	115.05	17723
cg.C.128	12.05	11892	12.55	11420	40.96	3500
Cg.C.64	14.13	10142	15.38	9317	48.72	2942
Cg.C.32	22.06	6498	22.73	6307	63.81	2246
mg.C.128	2.02	77051	2.34	66503	6.86	22689
mg.C.64	3.71	41955	3.79	41055	9.67	16100
mg.C.32	10.15	15399	10.22	15240	25.02	6223
sp.C.121	25.52	56823	34.53	41991	92.71	15641
sp.C.64	56.13	25834	57.94	25028	134.42	10788
sp.C.25	174.90	8291	174.75	8298	358.24	4048
bt.C.121	24.27	118115	28.42	100854	60.42	47437
bt.C.64	45.34	63213	46.40	61769	85.68	33452
bt.C.25	115.16	24889	115.60	24795	187.22	15310

Table 8: NAS Parallel benchmarks Class D:

Application	IB		10 Gigabit	
	Time, sec	Mops	Time, sec	Mops
ft.D.128	246.89	36306	336.33	26651
ft.D.64	473.87	18916	496.44	18056
lu.D.128	505.34	78953	628.94	63437
lu.D.64	1147.9	34757	1194.54	33400
cg.D.128	242.86	15000	358.04	10175
cg.D.64	685.54	5314	694.17	5248
mg.D.128	86.81	35868	86.86	35850
mg.D.64	90.51	34403	93.63	35259
sp.D.121	753.1	39219	759.16	38906
sp.D.64	1361.86	21688	1365.69	21627
bt.D.121	486.45	119922	495.96	117621
bt.D.64	868.09	67200	872.24	61880

NAMD:

NAMD is a parallel molecular dynamics designed for high-performance simulation of large biomolecular systems. The benchmark is the standard apoa1 benchmark simulated for 5000 time steps. We collected the performance in Table 9.

Table 9: NAMD performance of apoa1 simulation (Time in seconds):

No. of MPI tasks	IB	10 Gigabit	1 Gigabit
128	137		215
64	220	234	279
32	398	415	470
16	806	820	881
8	1429	1432	1542

CPMD:

The CPMD (Carr-Parrinello Molecular Dynamics) code is a parallelized plane wave/pseudopotential implementation of Density Functional Theory, particularly designed for ab initio molecular dynamics.

We chose two inputs for the performance study: “c120” for 120 carbon atoms and “Si512” for 512 silicon atoms. All tests consist of two parts: a wavefunction optimization and an MD simulation. We refer to these as Step1 and Step2 respectively.

The c120 benchmark uses BLYP functional, Kleinman pseudopotential with wavefunction cutoff of 35 Ry. The number of plane waves is ~360K and the real space mesh is 154 x 96 x 96.
The si512 uses LDA Kleinman pseudopotential with a wavefunction cutoff of 20 Ry. The number of plane waves is ~ 320K, real space mesh is 108 x 108 x 108.
In addition, we studied the 60 Ry cutoff case for si512, which uses LDA Kleinman pseudopotential. The number of plane waves is ~1680K, real space mesh is 180 x 180 x 180. We present the performance comparisons in Tables 10-12.

Table 10: c120 performance. All elapsed times in seconds

No. of MPI tasks	IB		10 Gigabit		1 Gigabit	
	Step1	Step2	Step1	Step2	Step1	Step2
128	102	208	154	281	3535	4186
64	119	244	158	290	299	589
32	204	414	285	568	438	879
16	424	850	447	899	901	1811
8	810	1651	813	1657	1006	1812

Table 11: si512 performance. All times are elapsed time in seconds

No. of MPI tasks	IB		10 Gigabit		1 Gigabit	
	Step1	Step2	Step1	Step2	Step1	Step2
128	140	71	203	130	3970	1670
64	232	95	267	136	430	209
32	387	165	515	182	1351	608
16	622	236	646	245	1126	391
8	1243	488	1243	470		

Table 12: si512-60 performance. All times are elapsed time in seconds

No. of MPI tasks	IB		10 Gigabit		1 Gigabit	
	Step1	Step2	Step1	Step2	Step1	Step2
128	685	336	748	394	2455	1362
64	942	493	974	465	2836	1354
32	1723	790	1786	794	3970	1670

HPCC:

The HPC Challenge benchmark is a suite of benchmarks that measure the performance of CPU, memory subsystem and the interconnect technology used for cluster communication. The HPC Challenge benchmark consists of the following benchmarks:

1. HPL – this is the LINPACK benchmark. The test measures the floating point performance of the system (Tflops/s).

2. Stream – is a simple synthetic benchmark program that measures sustainable memory bandwidth (Gbyte/s).

3. Random Access – measures the rate of random updates of memory (Gup/s, which Giga updates per second).

4. PTRANS – measures the rate of transfer for large arrays of data from memory (Gbyte/s).

5. FFTE – embarrassingly parallel Fast Fourier transforms that measures the rate in Gflop/s.

6. DGEMM – embarrassingly parallel matrix-matrix multiplication measurement (Gflop/s).

7. Latency/Bandwidth – measures the simple ping-pong and more complicated simulation connection (μs/Gbyte/s).

We collect the performance of HPCC in Table 13.

Table 13: HPCC performance on 128-way.

	IB	10Gb	1Gb
G-HPL Tflop/s	0.591754	0.59187	0.420978
G-PTRANS, GB/s	11.3455	10.9236	1.46663
G-Random Access, Gup/s	0.278015	0.17475	0.0550967
EP-Stream Triad GB/s	1.62769	1.61664	1.6591
G-FFTE, Gflop/s	22.3621	18.1289	3.88957
EP-DGEMM, Gflop/s	8.02071	8.02034	7.98519
Random Ring BW GB/s	0.130035	0.025915	0.010774
Random Ring Lat. Us	7.30212	25.2896	129.14

SPEC MPI:

SPEC® MPI2007 is SPEC's benchmark suite for evaluating MPI-parallel, floating point, compute intensive performance across a wide range of cluster and SMP hardware. SPEC® MPI2007 continues the SPEC tradition of giving users the most objective and representative benchmark suite for measuring and comparing high-performance computer systems.

SPEC® MPI2007 focuses on performance of compute intensive applications using the Message-Passing Interface (MPI), which means these benchmarks emphasize the performance of:

- the type of computer processor (CPU),
- the number of computer processors,
- the MPI Library,
- the communication interconnect,
- the memory architecture,
- the compilers, and
- the shared file system.

It is important to remember the contribution of all these components. SPEC® MPI2007 performance depends on more than just the processor. SPEC® MPI2007 is not intended to stress other computer components such as the operating system, graphics, or the I/O system.

The SPEC® MPI 2007 performance estimates are collected in Table 14. We present comparisons between InfiniBand and 10 Gigabit interconnect network for 64-way.

Table 14: SPEC MPI performance. All times are elapsed time in seconds

Application	10 Gigabit	InfiniBand (DDR)
104.milc	275	262
107.leslie3d	987	972
113.GemsFDTD	864	832
115.fds4	392	387
121.pop2	863	585
122.tachyon	567	589
126.lammps	594	606
127.wrf2	862	831
128.GAPgeofem	293	286
129.tera_tf	580	556
130.socorro	678	433
132.zeusmp2	537	534
137.lu	768	753

ACKNOWLEDGEMENT:

We would like to thank Chelsio Communications for loaning the 10 Gigabit Ethernet adapters and Force10 Communications for loaning the 10 Gigabit Ethernet switch. We thank Barry Spielberg and Greg Geiselhart for system support.