Technical Report: Benchmarking an HP DL580 cluster at Indiana University (Mason)

Craig A. Stewart
Matthew R. Link
Robert Henschel
David Hancock
Huian Li

Indiana University

Citation:

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services
Pervasive Technology Institute

This document is released under the creative commons 3.0 Unported license (http://creativecommons.org/licenses/by/3.0/).

Table of Contents

Table of Tables

1. Introduction

The Indiana University HP DL580 system was ordered with a purchase order placed on 11/23/2010. It arrived at Indiana University on 1/20/2011. This system, named Mason, has the following basic characteristics:

- The system consists of 16 HP ProLiant DL580 G7 Servers (http://h10010.www1.hp.com/wwpc/us/en/sm/WF05a/15351-15351-3328412-241644-3328422-4142916.html) with a 10 Gigabit Ethernet interconnect.
- Each node in the system contains four Intel Xeon 8-core L7555 1.866 (base frequency) GHz processors and 512 GB RAM, for a total of 32 processing cores totaling 238.85 GFLOPS per node at the base frequency, or 256 GFLOPS per node at the Turbo Boost mode.
- The overall system includes 64 CPUs, 512 processor cores, 3,822 GFLOPS peak processing capability at the base frequency (4,096 GFLOPS peak processing capability at the Turbo Boost mode), and 8,192 GB RAM.

A detailed system description follows, after which are details on the achieved performance during benchmarking.

2. System Description

2.1. Hardware

2.1.1. System topology

16 HP ProLiant DL580 G7 Servers each configured with four 8-core L7555 1.866 Ghz 95W processors, 1066MHz front-side bus (FSB), 512GB DDR3-1066 RDIMM memory, two 500 GB 7200 RPM 2.5" SAS hard drives (in RAID 1 mirror), and Intel 82546GB Dual-Port Gigabit Ethernet. Cluster interconnect is 10 Gigabit Ethernet.

2.1.2. Memory boards, sections, and/or banks

Each HP ProLiant DL580 G7 Server has 64 DIMM slots, each populated with 8 GB 2Rx4 PC3-10600R with error correcting code, running at 1066 MHz and CL9.

2.1.3. Memory size

512 GB per node, 64 x 8 GB DIMMs

2.1.4. CPU manufacturer, model, and speed

Intel Xeon L7555, 8-Core 1.866 GHz, 1066 MHz bus with 2048 KB L2 cache

2.1.5. Speed of the memory and memory bus (if applicable)

1066 MHz FSB, memory is 1066 MHz DDR3

2.1.6. I/O boards and bus interfaces

Two PCI-Express x8 Gen 2, three PCI-Express x4 Gen2

2.1.7. HBAs, Network Interface Cards and TCO Offload Engine (TOE) cards including firmware

HP NC375i 1G w/NC524SFP 10G module. Firmware version 4.0.54.

2.1.8. Network adapters, including firmware
HP NC375i Integrated Quad Port Gigabit Server Adapter. Firmware version 4.0.544.

2.1.9. All communications hardware, including private channels
One Cisco 7018, providing 16 10GbE ports via N7K-M132XP-12 line card modules. One Cisco SGE2010 48 Port Gigabit Ethernet switch.

2.1.10. RAID hardware including disks, cache, firmware, channels, GBICS and interfaces
Two 500GB 7200RPM SAS drives served as hardware RAID-1 via an HP P410 card. This provides 457GB of formatted space for local scratch (/tmp) on the compute nodes.

2.1.11. Fibre Channel switches, if used
None

2.1.12. Any other hardware used as part of the benchmark configuration
NFS mounted file system from an IBM N5500 NAS. Lustre file system mounted at /N/dc.

2.2. Software

2.2.1. Operating system, including all tunable parameters and their values
RedHat Enterprise Linux Server release 6.1 (Santiago)

2.2.2. BIOS tunable parameters and their values
HP DL580 G7 P65 05/23/2011 BIOS.

2.2.3. Network drivers
netxen_nic driver version 4.0.75.

2.2.4. Network stacks, including TOEs
Standard Linux network stack.

2.2.5. I/O drivers
None.

2.2.6. File system software and/or volume manager
Ext4 for local scratch /tmp, NFS for network mounted file system.

2.2.7. Compiler and libraries, including I/O and MPI libraries
- Fortran: Intel(R) Fortran Composer XE 2011 for Linux, version 12.0.2, Build 20110112
- C: Intel(R) C/C++ Composer XE 2011 for Linux, version 12.0.2, Build 20110112
- C++: Intel(R) C/C++ Composer XE 2011 for Linux, version 12.0.2, Build 20110112
- MPI: OpenMPI 1.4.3
- Mathematical Library: Intel MKL version 10.3.2

2.2.8. All patches and bug fixes
RHEL 6.1 with all updates as of most recent maintenance window.

2.2.9. *Any additional software used as part of the benchmark configuration*

N/A

---

3. Selected Benchmarks

The following benchmarks were selected to run on Mason.

1. HPC Challenge Benchmark 1.4.1 (HPCC-1.4.1) from University of Tennessee Knoxville (http://icl.cs.utk.edu/hpcc/). It consists of seven tests: HPL, DGEMM, STREAM, PTRANS, RandomAccess, and FFT.

2. SPEC OpenMP2001 Benchmark Suite Version 3.2 (SPEC OMP2001-3.2) from Standard Performance Evaluation Corporation (http://www.spec.org/omp2001/). This package is for evaluating performance based on OpenMP applications.

3. SPEC MPI2007 Benchmark Suite Version 2.0 (SPEC MPI2007-2.0) from Standard Performance Evaluation Corporation (http://www.spec.org/mpi2007/). This benchmark aims to evaluate MPI-parallel, floating point, compute intensive performance across a wide range of cluster and SMP hardware.

3.1. HPCC Performance

The following table lists the achieved performance for HPCC.

HPCC Benchmark Target			HPCC									
			G-HPL	G-PTRANS	G-FFTE	G-Random	G-STREAM	EP-STREAM	EP-DGEMM	Random Ring Bandwidth	Random Ring Latency	% HPL Peak
# Nodes	# CPUs	#Cores	TFLOP/s	GB/s	GLFOP/s	Gup/s	GB/s	GB/s	GFLOP/s	GB/s	μsec	Percent
16	64	512	3.3831	5.1122	17.9615	0.2447	1084.6822	2.1185	7.1575	0.0098	229.5640	82.59
8	32	256	1.6079	2.6475	8.9379	0.1538	549.1840	2.1453	7.1355	0.0106	169.0790	78.51
4	16	128	0.8465	1.5753	5.3559	0.1232	267.2973	2.0883	7.1405	0.0138	119.7360	82.66
2	8	64	0.4242	3.5449	10.0954	0.1519	137.7901	2.1530	7.1275	0.0832	71.9879	82.85
1	4	32	0.2222	6.4627	11.5422	0.2252	66.9363	2.0918	7.1568	0.3244	3.4833	86.78

Table 3-1. HPCC performance.

We noticed that HPCC jobs hung quite often during the benchmarking, most of the time in the RandomAccess component. Usually, after a job ran for a few tests, it just got stuck without any further output or activity. We suspect that this is related to our network configuration, which oversubscribes the 10G Ethernet network ports in a ratio of 4 to 1.

The HPCC results have been published at the HPC Challenge Benchmark web site maintained by the Innovative Computing Laboratory of UT Knoxville. Below are the URLs to the results on the web site:

- 512-core result: http://icl.cs.utk.edu/hpcc/hpcc_record.cgi?id=470
- 256-core result: http://icl.cs.utk.edu/hpcc/hpcc_record.cgi?id=471
- 128-core result: http://icl.cs.utk.edu/hpcc/hpcc_record.cgi?id=472
- 64-core result: http://icl.cs.utk.edu/hpcc/hpcc_record.cgi?id=473
- 32-core result: http://icl.cs.utk.edu/hpcc/hpcc_record.cgi?id=474

3.2. SPEC OMP Performance

The SPEC OMP2001-3.2 benchmark has been run on the system with two different settings, with hyperthreading turned off and on.

3.2.1. Performance with hyperthreading off

The SPEC OMP2001-3.2 benchmark was run using the medium and large data sets, respectively. The benchmark used 32 threads, specified with the environment variable, OMP_NUM_THREADS. The table below lists the performance using the medium data set when hyperthreading off.

Benchmarks	# of Threads	Base Ref Time	Base Run Time	Base Ratio
310.wupwise_m	32	6000	46.7	128583
312.swim_m	32	6000	84.1	71337
314.mgrid_m	32	7300	96.9	75332
316.applu_m	32	4000	29.9	133967
318.galgel_m	32	5100	115.0	44387
320.equake_m	32	2600	52.3	49691
324.apsi_m	32	3400	44.7	76128
326.gafort_m	32	8700	98.2	88571
328.fma3d_m	32	4600	88.3	52108
330.art_m	32	6400	31.1	205935
332.ammp_m	32	7000	152.0	45953
SPECompMbase2001				**78307**

Table 3-2. SPEC OMP performance using the medium data set with hyperthreading off.

This set of results has been published at the web site (http://www.spec.org/omp/results/) maintained by Standard Performance Evaluation Corporation (SPEC). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q3/omp2001-20110613-00427.asc
- HTML file: http://www.spec.org/omp/results/res2011q3/omp2001-20110613-00427.html
- PDF file: http://www.spec.org/omp/results/res2011q3/omp2001-20110613-00427.pdf
- Config file: http://www.spec.org/omp/results/res2011q3/omp2001-20110613-00427.cfg

The table below shows the performance using the large data set when with hyperthreading off.

Benchmarks	# of Threads	Base Ref Time	Base Run Time	Base Ratio
311.wupwise_l	32	9200	203	723729
313.swim_l	32	12500	602	331955
315.mgrid_l	32	13500	518	416695
317.applu_l	32	13500	562	384230
321.equake_l	32	13000	575	361456
325.apsi_l	32	10500	271	620871
327.gafort_l	32	11000	391	450003
329.fma3d_l	32	23500	1166	322462
331.art_l	32	25000	290	1377258
SPECompLbase2001				**493152**

Table 3-3. SPEC OMP performance using the large data set with hyperthreading off.

This set of results has also been published at the SPEC web site (http://www.spec.org/omp/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q3/omp2001-20110613-00428.asc
- HTML file: http://www.spec.org/omp/results/res2011q3/omp2001-20110613-00428.html
- PDF file: http://www.spec.org/omp/results/res2011q3/omp2001-20110613-00428.pdf
- Config file: http://www.spec.org/omp/results/res2011q3/omp2001-20110613-00428.cfg

3.2.2. Performance with hyperthreading on

The SPEC OMP2001-3.2 benchmark was run using the medium and large data sets, respectively. For the medium data set, the number of threads was set to 32 while for the large data set, 64 threads were used. The table below lists the performance using the medium data set when hyperthreading is on.

Benchmarks	# of Threads	Base Ref Time	Base Run Time	Base Ratio
310.wupwise_m	32	6000	37.3	160774
312.swim_m	32	6000	74.2	80847
314.mgrid_m	32	7300	87.7	83222
316.applu_m	32	4000	26.1	153288
318.galgel_m	32	5100	114.0	44802
320.equake_m	32	2600	47.9	54295
324.apsi_m	32	3400	46.5	73134
326.gafort_m	32	8700	109	79651
328.fma3d_m	32	4600	92.8	49543
330.art_m	32	6400	32.9	194318
332.ammp_m	32	7000	161.0	43469
SPECompMbase2001				**80989**

Table 3-4. SPEC OMP performance using the medium data set with hyperthreading on.

This set of results is available at the SPEC web site (http://www.spec.org/omp/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q4/omp2001-20111021-00432.asc
- HTML file: http://www.spec.org/omp/results/res2011q4/omp2001-20111021-00432.html
- PDF file: http://www.spec.org/omp/results/res2011q4/omp2001-20111021-00432.pdf
- Config file: http://www.spec.org/omp/results/res2011q4/omp2001-20111021-00432.cfg

The table below shows the performance using the large data set when hyperthreading is off.

Benchmarks	# of Threads	Base Ref Time	Base Run Time	Base Ratio
311.wupwise_l	32	9200	211	697727
313.swim_l	32	12500	628	318338
315.mgrid_l	32	13500	528	409378
317.applu_l	32	13500	590	366221
321.equake_l	32	13000	542	383513
325.apsi_l	32	10500	286	587380
327.gafort_l	32	11000	359	490814
329.fma3d_l	32	23500	941	399786
331.art_l	32	25000	277	1445765
SPECompLbase2001				**504788**

Table 3-5. SPEC OMP performance using the large data set with hyperthreading on.

This set of results is also available at the SPEC web site (http://www.spec.org/omp/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q4/omp2001-20111021-00431.asc
- HTML file: http://www.spec.org/omp/results/res2011q4/omp2001-20111021-00431.html
- PDF file: http://www.spec.org/omp/results/res2011q4/omp2001-20111021-00431.pdf
- Config file: http://www.spec.org/omp/results/res2011q4/omp2001-20111021-00431.cfg

The overall performance of the SPEC OMP benchmark for hyperthreading on is better than for non-hyperthreading, for both cases, medium data set and large data set. However, the difference in performance between hyperthreading and non-hyperthreading is not significant, particularly for the large data set.

3.3. SPEC MPI Performance

SPEC MPI2007 evaluates MPI-parallel, floating point, compute intensive performance across a wide range of cluster and SMP hardware. The data sets also include a medium set and a large set. Below we present the performance for both cases.

3.3.1. *Performance using the medium data set*

The tables below show the performance of SPEC MPI2007 using the medium set on 32 cores, 64 cores, 128 cores, and 256 cores, respectively.

Benchmarks	Base Ranks	Base Run Time	Base Ratio
104.milc	32	459	3.41
107.leslie3d	32	1417	3.68
113.GemsFDTD	32	968	6.51
115.fds4	32	563	3.46
121.pop2	32	904	4.57
122.tachyon	32	1005	2.78
126.lammps	32	1097	2.66
127.wrf2	32	1063	7.34
128.GAPgeofem	32	433	4.77
129.tera_tf	32	830	3.34
130.socorro	32	922	4.14
132.zeusmp2	32	743	4.18
137.lu	32	993	3.70
SPECmpiM(TM)_base2007			**4.02**

Table 3-6. SPEC MPI2007 performance using the medium data set on 32 cores.

This set of results has been published at the SPEC web site (http://www.spec.org/mpi/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00334.asc
- HTML file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00334.html
- PDF file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00334.pdf
- Config file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00334.cfg

Benchmarks	Base Ranks	Base Run Time	Base Ratio
104.milc	64	232	6.74
107.leslie3d	64	834	6.26
113.GemsFDTD	64	557	11.3
115.fds4	64	293	6.67
121.pop2	64	1424	2.90
122.tachyon	64	504	5.55
126.lammps	64	565	5.16
127.wrf2	64	646	12.1
128.GAPgeofem	64	261	7.90
129.tera_tf	64	444	6.24
130.socorro	64	441	8.65
132.zeusmp2	64	421	7.38
137.lu	64	370	9.94
SPECmpiM(TM)_base2007			**7.01**

Table 3-7. SPEC MPI2007 performance using the medium data set on 64 cores.

This set of results is available at the SPEC web site (http://www.spec.org/mpi/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00335.asc
- HTML file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00335.html
- PDF file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00335.pdf
- Config file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00335.cfg

Benchmarks	Base Ranks	Base Run Time	Base Ratio
104.milc	128	128	12.3
107.leslie3d	128	492	10.6
113.GemsFDTD	128	414	15.2
115.fds4	128	142	13.7
121.pop2	128	1971	2.09
122.tachyon	128	255	11.0
126.lammps	128	289	10.1
127.wrf2	128	350	22.3
128.GAPgeofem	128	171	12.0
129.tera_tf	128	262	10.6
130.socorro	128	284	13.4
132.zeusmp2	128	239	13.0
137.lu	128	205	17.9
SPECmpiM(TM)_base2007			**11.4**

Table 3-8. SPEC MPI2007 performance using the medium data set on 128 cores.

This set of results is available at the SPEC web site (http://www.spec.org/mpi/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00336.asc
- HTML file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00336.html
- PDF file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00336.pdf
- Config file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00336.cfg

Benchmarks	Base Ranks	Base Run Time	Base Ratio
104.milc	256	80.8	19.4
107.leslie3d	256	311	16.8
113.GemsFDTD	256	506	12.5
115.fds4	256	60.3	32.4
121.pop2	256	1991	2.07
122.tachyon	256	130	21.5
126.lammps	256	263	11.1
127.wrf2	256	278	28.0
128.GAPgeofem	256	141	14.6
129.tera_tf	256	167	16.6
130.socorro	256	309	12.4
132.zeusmp2	256	125	24.9
137.lu	256	126	29.3
SPECmpiM(TM)_base2007			**15.8**

Table 3-9. SPEC MPI2007 performance using the medium data set on 256 cores.

This set of results is also available at the SPEC web site (http://www.spec.org/mpi/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00337.asc
- HTML file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00337.html
- PDF file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00337.pdf
- Config file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00337.cfg

3.3.2. *Performance using the large data set*

The tables below show the performance of running SPEC MPI2007 with the large set on 64 cores and 128 cores, respectively.

Benchmarks	Base Ranks	Base Run Time	Base Ratio
121.pop2	64	3143	1.24
122.tachyon	64	2820	0.689
125.RAxML	64	4021	0.726
126.lammps	64	2837	0.867
128.GAPgeofem	64	4731	1.25
129.tera_tf	64	1436	0.765
132.zeusmp2	64	1877	1.13
137.lu	64	3977	1.06
142.dmilc	64	2288	1.61
143.dleslie	64	2495	1.24
145.lGemsFDTD	64	3587	1.23
147.l2wrf2	64	6610	1.24
SPECmpiL(TM)_base2007			**1.05**

Table 3-10. SPEC MPI2007 performance using the large data set on 64 cores.

This set of results has been published at the SPEC web site (http://www.spec.org/mpi/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00333.asc
- HTML file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00333.html
- PDF file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00333.pdf
- Config file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00333.cfg

Benchmarks	Base Ranks	Base Run Time	Base Ratio
121.pop2	128	1475	2.64
122.tachyon	128	1429	1.36
125.RAxML	128	2043	1.43
126.lammps	128	1436	1.71
128.GAPgeofem	128	2810	2.11
129.tera_tf	128	763	1.44
132.zeusmp2	128	962	2.20
137.lu	128	1739	2.42
142.dmilc	128	1185	3.11
143.dleslie	128	1234	2.51
145.lGemsFDTD	128	1773	2.49
147.l2wrf2	128	3291	2.49
SPECmpiL(TM)_base2007			**2.09**

Table 3-11. SPEC MPI2007 performance using the large data set on 128 cores.

This set of results is also available at the SPEC web site (http://www.spec.org/mpi/results/). The URLs to the corresponding files are:

- Text file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00332.asc
- HTML file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00332.html
- PDF file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00332.pdf
- Config file: http://www.spec.org/omp/results/res2011q4/mpi2007-20111215-00332.cfg

4. Conclusion

Based on the results presented in this report, we conclude that the system performs as expected, especially in terms of the HPL performance. However, the oversubscription of the Ethernet network needs to be further investigated.