Performance of Cray Systems – Kernels, Applications & Experiences

Mike Ashworth, Miles Deegan, Martyn Guest, Christine Kitchen, Igor Kozin and Richard Wain
Computational Science & Engineering Department, CCLRC Daresbury Laboratory

ABSTRACT: *We report of the activities of the Computational Science & Engineering Department at CCLRC Daresbury Laboratory in the evaluation of Cray high-end and mid-range systems. We examine the performance of applications from computational fluid dynamics, coastal ocean modelling and molecular dynamics as well as kernels from the HPC Challenge benchmark. We find that the CrayX1 and Cray XT3 are highly competitive with contemporary systems from IBM and SGI, the precise ranking of these systems being application dependent. We examine the performance of the Cray XD1 as a mid-range computing resource. It performs well but a Pathscale InfiniPath cluster performs equally well at some fraction of the cost. InfiniPath systems appear particularly competitive for runs on 32 and 64 processors; still considered the 'sweet spot' for the majority of applications on mid-range systems. A successor to the XD1 is required if Cray are going to provide a cost-effective solution in the mid-range cluster market.*

KEYWORDS: benchmarking, parallel computing, clusters, CFD, molecular dynamics, coastal ocean modelling

1. Introduction

The computational science and engineering activities within CCLRC provide world-class expertise and support for UK theoretical and computational science communities, in both academia and industry. Our program of work includes the development and optimisation of leading high-performance simulation codes as well as support and advice on operating systems, tools and machine performance. We also evaluate and benchmark a range of high-end and cluster architectures and can provide objective comparisons for kernel and application performance across a range of systems.

This paper highlights work of the Department which relates to the evaluation of Cray systems, including the performance of kernels and full applications as well as relating the experience of running our own Cray XD1 system. There are two main themes to the evaluation work. At the high-end we present a comparison of full application codes on the Cray XT3 and Cray X1 against other leading capability and leadership-class systems. Secondly we present data from kernels and applications which show how the Cray XD1 compares with other mid-range and cluster systems especially Pathscale's InfiniPath cluster.

We first describe the main systems used in the evaluation. Section 3 describes the application codes and the benchmark cases, giving results and analysis for their performance on high-end systems. In section 4 we examine the Cray XD1 as a mid-range computing resource and present comparative benchmark data with the Pathscale InfiniPath cluster. We close with some concluding remarks.

2. Description of the main systems

Cray XD1

The CSE Cray XD1 is a 6 chassis system with Rapid Array Interconnect. Each chassis consists of 6 nodes, where a node is a dual processor AMD 2.4 GHz Opteron 250 with 4GB local memory (2GB per processor). Three of the chasses contain Virtex IV FPGAs. Portland Group C and Fortran compilers (version 6.1-1) are installed, as are GNU C, C++ and Fortran compilers. The job scheduler in use is Sun Grid Engine.

Cray X1

Oak Ridge National Laboratory took delivery of a Cray X1 system in 2003 and have carried out an extensive evaluation of the system [1]. The Cray X1 is a scalable parallel computer with symmetric multiprocessor

(SMP) nodes where the processors are very good vector processors with weak scalar performance. The basic building block of a Cray X1 system is the Single-Streaming Processor (SSP). An SSP consists of a vector processor that has 32 vector registers of 64 elements each, implemented in two vector pipelines and operating at 800 MHz. An SSP also has a 400-MHz scalar processor. The peak performance of an SSP is 3.2 gigaflops. The two vector units in an SSP have an 800-MHz clock and can move a column (vector) of numbers from memory into high-speed registers or initiate an operation with a string of resultants obtained two per clock cycle after initial vector setup. These operations are accomplished much more efficiently than typical microprocessor operations and generate many more resultants per second and higher sustained computation rates. The runs reported here were performed in SSP mode in which each SSP is treated as a separate processor.

Cray XT3

The Cray XT3 is a productization of the architecture pioneered at Sandia National Laboratories in the "Red Storm"[1] system. The Cray XT3 system at the Swiss National Supercomputing Centre (CSCS) consists of 12 cabinets containing a total of 1100 compute processors. The processors are dual-core 2.6 GHz AMD Opterons with 2 GB of memory. Each Opteron processor is directly connected to a dedicated SeaStar chip, which contains a 6-way router and communications engine. The system runs the UNICOS operating system, with a fully-featured Linux OS on the service nodes and the Catamount lightweight kernel on the compute nodes. Codes were compiled with the Portland Group PGI Fortran compiler pgf90 version 6.1.

IBM p690+

HPCx[2] is the UK's leading and most recent national High Performance Computing service. The system has gone through three distinct technology provisions. The Phase1 system consisted of an IBM 1.3 GHz p690 cluster with the SP Switch2. In 2004 it was upgraded to 1.7 GHz p690+ together with IBM's High performance Switch (HPS), formerly known as "Federation".

The current Phase2A system, installed at the end of 2005, consists of 96 p5-575 nodes to give a total of 1536 processors. The p5-575 is a 16-way shared memory system with a three-level cache architecture. There are two POWER5 processors per chip each with its own Level 1 data and instruction caches and with a shared on-chip Level 2 cache. The sixteen processors (eight chips) share 128 MB of Level 3 cache and 32GB of main memory. As before, communication between nodes is provided by the HPS. Each node has two network adapters and there are two links per adapter, making a total of four links to the switch network shared between the 16 processors of the node. The Fortran compiler was xlf version 9.1.

Pathscale InfiniPath cluster

Although InfiniPath technology is based on InfiniBand it provides considerable performance improvements over standard InfiniBand solutions. InfiniPath's InfiniBand host channel adapter (HCA) uses a HyperTransport interface to the host processor thus also bypassing the PCI bus. However the difference is that the InfiniPath HCA connects directly to the AMD Opteron CPU via an open standard HyperTransport HTX slot. Externally, it utilizes standard InfiniBand switching fabrics. PathScale also provides an MPI library which bypasses the kernel. In contrast to the Cray XD1, the InfiniPath adapter does not have an embedded processor and relies instead on the power of the host CPU.

SGI Altix 3700

CSAR[3] at the University of Manchester operate a flagship 512 Itanium2 processor SGI Altix 3700 system known as *newton*. Of the 512 processors, 384 have a clock speed of 1.3 GHz and 128 are 1.5 GHz. By selecting different batch queues one can select which processors are used. The system has 1.5 GB of memory per processor and uses the NUMAflex interconnect. The node size (the size of a single system image) is 256 processors, though MPI jobs can span nodes, and the overall maximum job size is 358 processors. Codes were compiled with the Intel Fortran compiler. Version 8.0 was standard although version 7.1 was used for the PDNS3D code as this was found to be faster.

Streamline Cluster

CCLRC's e-Science Centre operates a cluster (known as SCARF) whose primary role is to encourage CCLRC facilities to increase their uptake of Grid computing. The cluster was supplied by Streamline Computing and consists of 128 dual CPU AMD Opteron 248 chips running at 2.2 GHz and connected by Myrinet M3F-PCIXD-2. The majority of the processors have 4GB main memory with just 16 being equipped with 8GB. We used the Portland Group PGI Fortran compiler pgf90 version 6.0.

[1] http://www.sandia.gov/ASC/redstorm.html

[2] http://www.hpcx.ac.uk/

[3] http://www.csar.cfs.ac.uk/

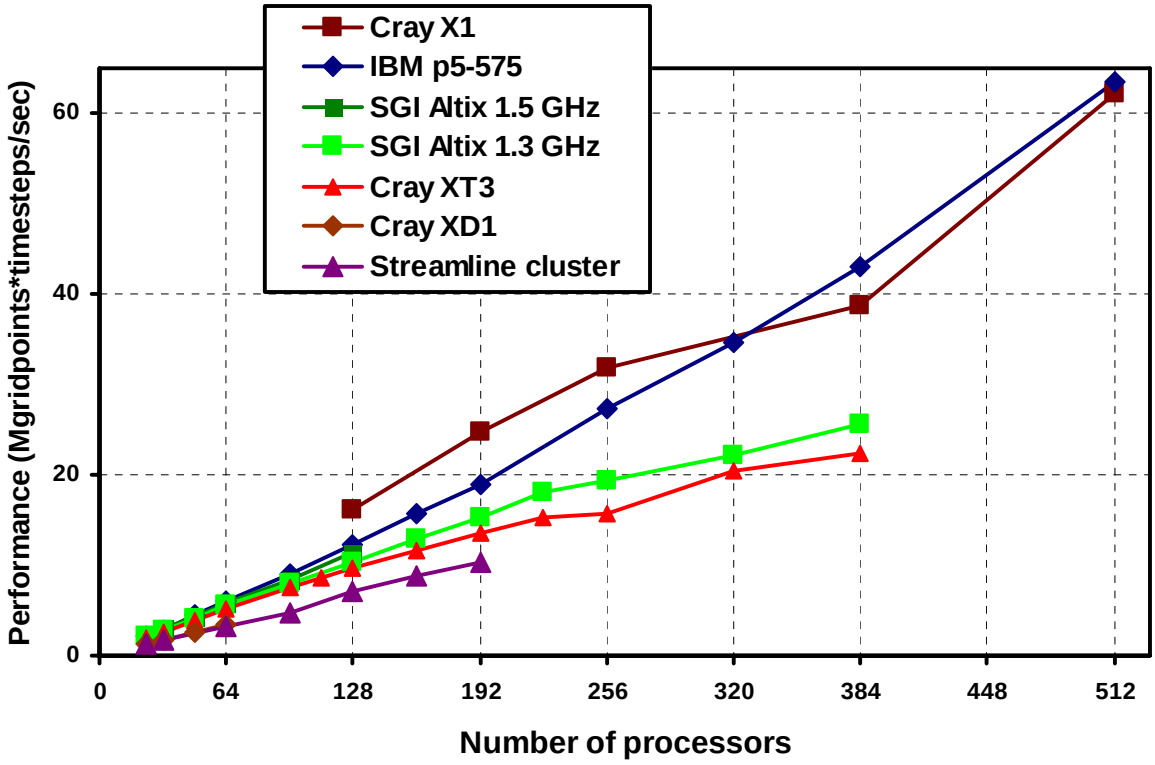

Figure 1: Performance of the PDNS3D PCHAN 360 x 360 x 360 benchmark for a range of high-end computer systems

3. Performance of Applications on High-End Systems

PDNS3D

The PDNS3D code was developed by the UK turbulence consortium (UKTC) and uses direct numerical simulation (DNS) techniques to solve the Navier-Stokes equations for turbulent fluid flow, focusing in particular on the shock/boundary-layer interaction [2]. The code was originally developed for the Cray T3E and is a sophisticated DNS code that incorporates a number of advanced features: namely high-order central differencing; a shock-preserving advection scheme from the total variation diminishing (TVD) family; entropy splitting of the Euler terms and the stable boundary scheme. The code has been written using standard Fortran 90 code together with MPI in order to be efficient, scalable and portable across a wide range of high-performance platforms. The PCHAN benchmark is a simple turbulent channel flow benchmark using the PDNS3D code. Communications is limited to nearest neighbour boundary exchange. We report on the performance of the T3 benchmark, a large case with a gridsize of 360 x 360 x 360 grid points.

Figure 1 shows the performance of the PDNS3D PCHAN T3 benchmark, reported as work divided by time where the work is estimated as the number of gridpoints times the number of timesteps. The scaling of this code is near perfect on all systems as the cost of the boundary exchange is negligible compared to the computation on each sub-domain. Indeed several systems show significant super-linear scaling as the decreasing sub-domain size starts to fit within cache limits as the number of processors increases. Absolute performance is highly dependent on cache and memory issues, and has been measured at just 4.9% of peak [3]. For this application the three Opteron systems are at the bottom of the rankings, followed by the Altix systems. The IBM p5-575 and the Cray X1 deliver the highest per processor performance and run *neck-and-neck* up to 1024 processors (not shown) reaching a performance level of around 120M gridpoint-timesteps/s. The excellent performance on the Cray X1 is due to the triple-nested loops offering easy vectorisation with sequential memory access and a vector length equal to the domain size, which in this case is equal to 360.

POLCOMS

The Proudman Oceanographic Laboratory Coastal Ocean Modelling System (POLCOMS) has been developed to tackle multi-disciplinary studies in coastal/shelf environments [4]. The central core is a

sophisticated 3-dimensional hydrodynamic model that provides realistic physical forcing to interact with, and transport, environmental parameters. The finite difference scheme is based on a latitude-longitude Arakawa B-grid in the horizontal with terrain-following S-coordinates in the vertical. Conservative monotonic PPM advection routines are used to ensure strong frontal gradients. Vertical mixing is through turbulence closure (Mellor-Yamada level 2.5).

In order to study a range of scientific and pratical problems, the hydrodynamic model has been coupled with other models, e.g. the European Seas Regional Ecosystem Model (ERSEM) [5], the WAM wave model and a sediment-transport model. Work is in progress to couple POLCOMS with the Los Alamos CICE ice model.

The performance results reported here refer to a domain covering the north-west European shelf seas at a resolution of 1/10 degree x 1/15 degree, which is known as the Medium-Resolution Continental Shelf (MRCS) model. This results in a grid size of 251 x 206 x 20. In order to improve simulation of marine processes, we need accurate representation of eddies, fronts and other regions of steep gradients. A High Resolution Continental Shelf (HRCS) model is now being run which covers a similar area at approximately 1.8km resolution.

In order to keep benchmark run times manageable, the runs were kept short (15 model days) and the initialisation and finishing times were subtracted from the total run time. The performance is reported in Figure 2 as the model days per day of execution time. This code has been extensively modified to suit cache-based processors and no longer vectorises, as discussed in [6], so the Cray X1 performance is not shown.

As with the PDNS3D code, communications are limited almost entirely to nearest neighbour boundary exchange operations but in this case the frequency of these exchanges, coupled with the relatively small grid size, causes the performance to flatten off as the number of processors is increased. The Streamline cluster shows the relatively poor performance of its Myrinet interconnect by turning over above 128 processors. The other high-end systems show much better scaling behaviour. The absolute performance is also dependent on cache and memory issues, with a measured 8.7% of peak [3], and for this application the IBM p5-575 and Altix 1.3 GHz are very close, the Cray XD1 and Altix 1.5 GHz come in at about 26% faster (than the IBM on 64 processors) and the Cray XT3 leads the pack at 40% faster. The relative performance of the three Opteron-based systems is precisely as expected from the ratio of their clock speeds.

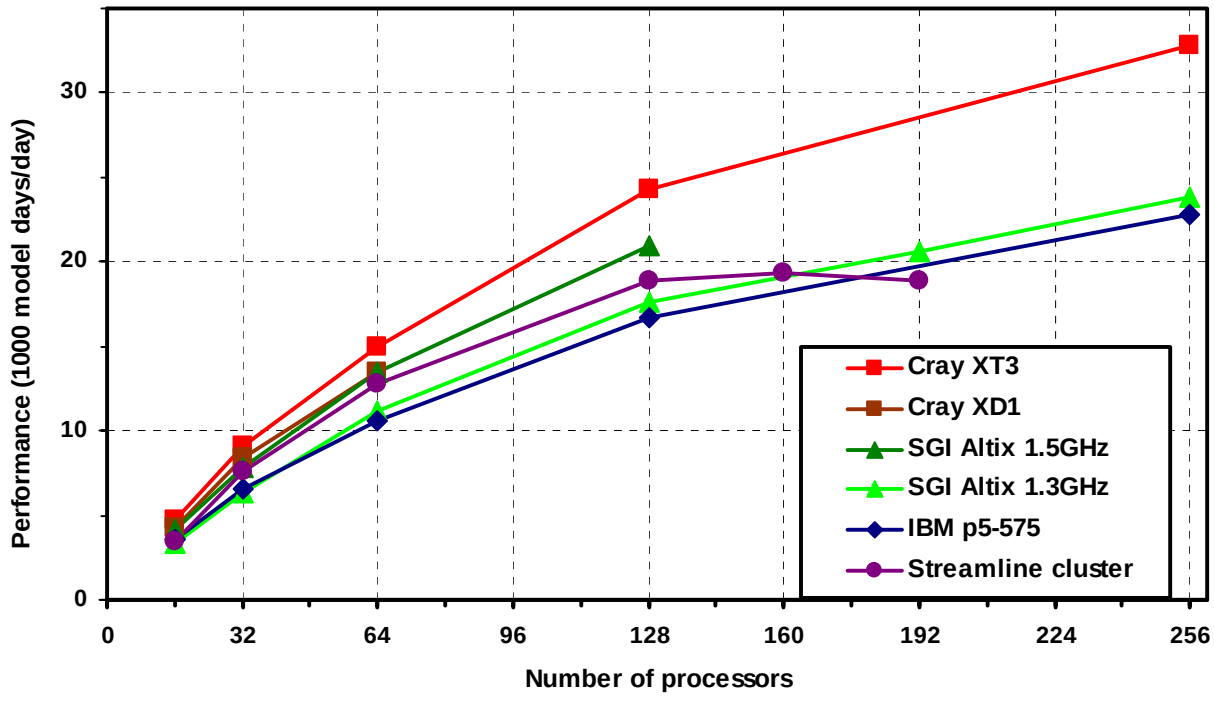

Figure 2: Performance of the POLCOMS MRCS benchmark for a range of high-end computer systems

The Cray XD1 as a mid-range computing resource

PathScale's InfiniPath technology is one of a number of emerging interconnect technologies that are able to provide latencies below two microseconds and substantially increased bandwidth over previous interconnect generations. The introduction of such low latency, low cost options means that commodity clusters can now provide a real challenge to proprietary machines such as the Cray XD1 in terms of raw performance. This paper uses data from synthetic and real-world application benchmarks to compare the performance of single and dual-core InfiniPath and Cray XD1 systems.

In this paper we focus on results from one synthetic benchmark suite (HPCC) and two real-world application benchmarks.

HPCC

The HPC Challenge benchmark[4] is a DARPA (HPCS) funded benchmark which has been designed to augment the Top500[5] list. The benchmark contains the High Performance Linpack (HPL) benchmark used in the Top500 as well as a range of other tests designed to exercise a wide range of memory access patterns. Results from the following tests are included in this report:

- HPL - Linpack TPP benchmark measures the floating point execution rate for solving a linear system of equations.
- DGEMM - measures the floating point execution rate of double precision real matrix-matrix multiplication.
- STREAM - measures the sustainable memory bandwidth (in GB/s) and the corresponding computation rate for a simple vector kernel.
- PTRANS (parallel matrix transpose) - exercises the communications where pairs of processors communicate with each other simultaneously. It is a useful test of the total communications capacity of the network.
- RandomAccess - measures the rate of integer random updates of memory (GUPS).
- FFTE - measures the floating point execution rate of a double precision complex one-dimensional Discrete Fourier Transform (DFT).
- Communication bandwidth and latency - a set of tests to measure the latency and bandwidth of a number of simultaneous communication patterns; based on the effective bandwidth benchmark (b_eff).

DL_POLY

DL_POLY [7] is a general-purpose molecular dynamics simulation package designed to cater for a wide range of possible scientific applications and computer platforms, especially parallel hardware. DL_POLY supports a wide range of application areas, including [8] ionic solids, solutions, metals, zeolites, surfaces and interfaces, complex systems (e.g. liquid crystals), minerals, bio-systems, and those in spectroscopy. Version 3 of the code is based on a distributed data scheme which can scale to a much larger number of atoms and processors than the previous replicated-data approach [9]. In our benchmarking we used the recently released version 3.04.

The benchmark case exemplifies a biological application of DL_POLY. It is a simulation of 16 Gramicidin A molecules in water with a total number of 792,960 atoms. Simulation is performed at 300 K using NVT Berendsen ensemble with SPME and RATTLE algorithm for the constrained motion.

CHARMM

Chemistry at HARvard Macromolecular Mechanics (CHARMM)[6] is a general-purpose molecular mechanics, molecular dynamics and vibrational analysis package for modelling and simulation of the structure and behaviour of molecular systems. We used version c31b1 for the work described in this report.

There are a number of benchmarks available for CHARMM. The Myoglobin in water benchmark is the most popular one introduced by Milan Hodoscek[7]. It is a MD calculation of carboxy myoglobin (MbCO) with 3830 Water Molecules (14026 atoms, 1000 steps (1 ps), 12-14 Å shift). See the URL[7] for a more detailed description and table of benchmark results obtained on a variety of platforms.

Results are presented for a number of systems whose hardware and software characteristics are described as follows and summarised in Table 1.

[4] http://icl.cs.utk.edu/hpcc/.

[5] http://www.top500.org/

[6] http://www.charmm.org.

[7] See http://www.cmm.ki.si/parallel/summary.html for a more detailed description of the benchmark and the table of results. Also some results can be found at http://arg.cmm.ki.si/vrana/amdvsintel.html and more recent at http://arg.cmm.ki.si/vrana/gccvsifort.html.

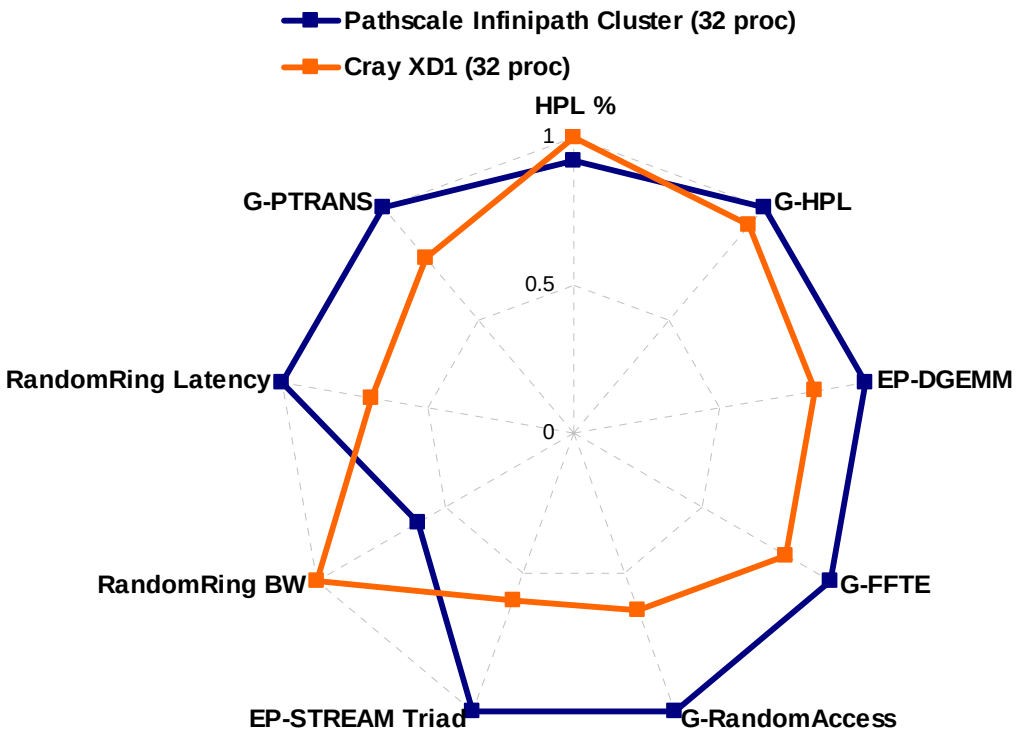

Figure 3: Kiviat diagram comparing Cray XD1 (Opt/2,4/RA) and InfiniPath (Opt/2,6/IP) clusters.

Reference name	Opt/ 2.0/ IB	Opt/ 2.4/ RA	Opt/ 2.6/ IP	Opt/ 2.2DC/ RA	Opt/ 2.2DC/ IP
Computer	IBM e325	Cray XD1	InfiniPath	Cray XD1	InfiniPath
Processor	Opteron 246	Opteron 250	Opteron 252	Opteron 275	Opteron 275
clock (GHz)	2.0	2.4	2.6	2.2	2.2
Cores	1	1	1	2	2
chips per node	2	2	2	2	2
inteconnect	InfiniBand	Rapid Array	InfiniPath	Rapid Array	InfiniPath
Compiler	pgi 5.2-1	pgi 6.02	psc 2.2	pgi 6.02	psc 2.2.1

Table 1: Mid-range systems summary

Opt/2.0/IB – reference system: IBM e325 single core

Opteron 246 2.0 GHz, IBM e325 dual CPU nodes, PGI 5.2 compilers, InfiniBand interconnect, Scali MPI library.

Opt/2.4/RA – Cray XD1 single core

Opteron 250 2.4 GHz, PGI 6.0 compilers, Rapid Arrays interconnect, Cray MPI library, software stack version 1.2.

Opt/2.6/IP – InfiniPath single core

Opteron 252 2.6 GHz, PathScale 2.2 compilers, InfiniPath interconnect, PathScale MPI library.

Opt/2.2DC/RA – Cray XD1 dual core

Opteron 275 2.2 GHz, PGI 6.0 compilers, Rapid Arrays interconnect, Cray MPI library, software stack version 1.3.

Opt/2.2DC/IP – InfiniPath dual core

Opteron 275 2.2 GHz, PathScale 2.2 compilers, InfiniPath interconnect, PathScale MPI library.

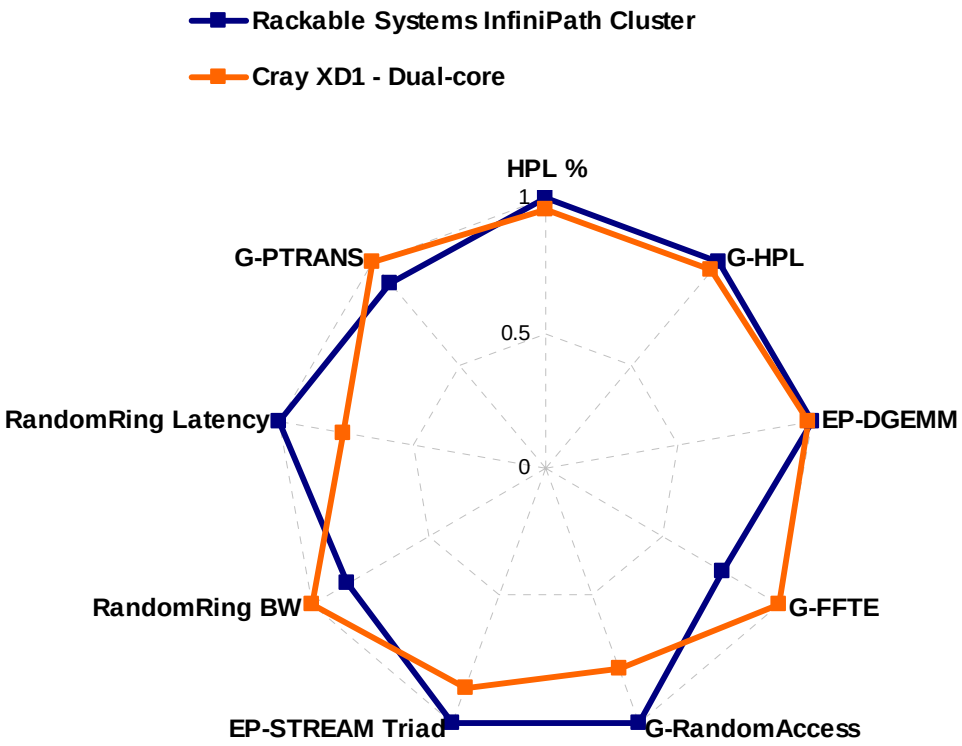

Figure 4: Kiviat diagram comparing dual-core Cray XD1 (Opt/2.2DC/RA) and InfiniPath (Opt/2.2DC/IP) clusters.

Benchmark Data and Results

HPCC

The latency and bandwidth data in table 2 is the worst case data generated by the HPCC ping-pong test which runs between each possible pair of processors. The latency and bandwidth data used in the Kiviat graphs in figures 3 and 4 were generated using the random ring communication test which is also part of the HPCC benchmark.

	Opt/ 2.0/ IB	Opt/ 2.4/ RA	Opt/ 2.6/ IP	Opt/ 2.2DC/ RA	Opt/ 2.2DC/ IP
latency (usec)	7.35	2.40	1.51	3.00	2.12
Bandwidth (GB/s)	0.42	1.26	1.55	1.40	1.63

Table 2: HPCC latency and bandwidth.

Figure 3 shows HPCC results for the Cray XD1 (Opt/2.4/RA) and InfiniPath (Opt/2.6/IP) systems using the standard Kiviat diagrams for HPCC in which the results for each test are normalized so that the best result is set to 1.0. The results shown here are for base runs of HPCC on 32 processors. Much of the difference in performance for HPL, DGEMM, FFTE and PTRANS can be put down to the difference in clock speed between the Cray (2.4GHz) and the PathScale machine (2.6GHz).

It is clear that whilst clock speed is a significant factor here we are still seeing comparable performance from the PathScale machine. Only in the bandwidth test does the XD1 significantly outscore the InfiniPath based system.

Figure 4 shows HPCC results for the dual-core Cray XD1 (Opt/2.2DC/RA) and dual-core InfiniPath (Opt/2.2DC/IP) systems. These are both dual-core systems based on the AMD Opteron 275 processor.

Results are normalized to the best result for each of the HPCC tests. The results shown here are for base runs of HPCC on 64 processing elements (32 chips, 64 cores).

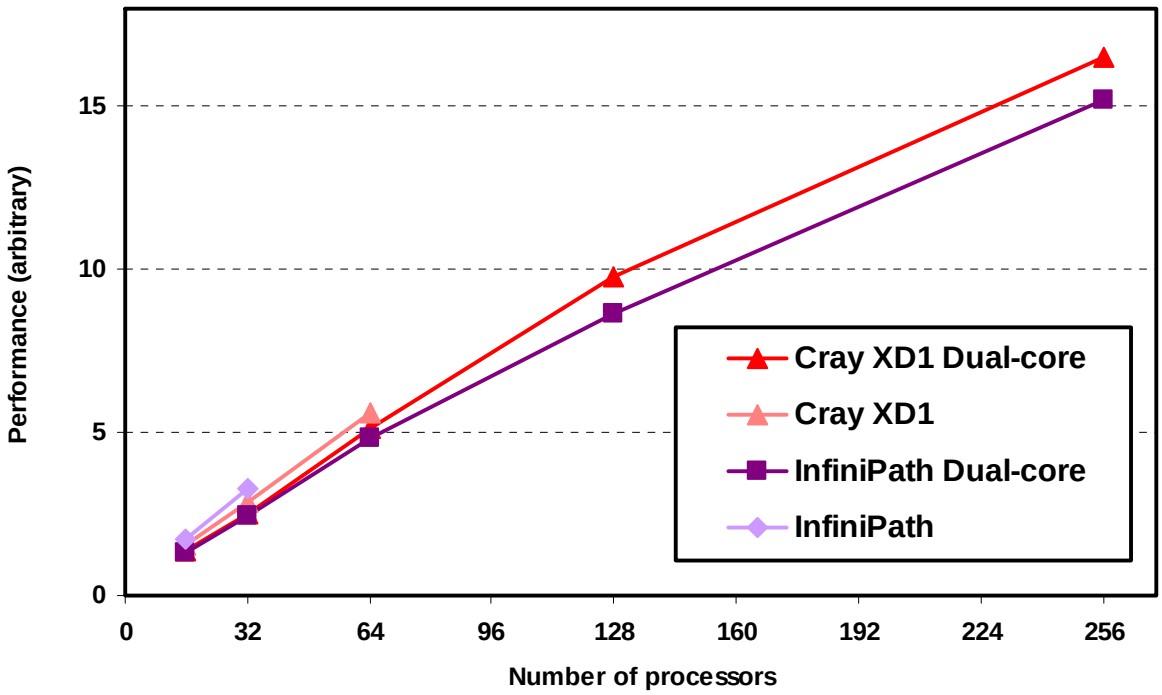

Figure 5: DL-POLY Gramicidin A performance for single and dual-core Cray XD1 and InfiniPath systems.

DL-POLY

Performance, defined as an arbitrary constant divided by the execution time, for the DL_POLY Gramicidin A benchmark is shown in Figure 5 for both single-core and dual-core versions of the Cray XD1 and the InfiniPath cluster. The single-core InfiniPath system is able to match the single-core Cray XD1 in this particular test. Similar results were observed for a range of DL-POLY benchmarks. The slight advantage held by the InfiniPath system is primarily due to its higher clock speed.

The dual-core systems are very closely matched for smaller runs. The Cray XD1 dual-core system demonstrates slightly better scaling when running on 64 processors or more.

CHARMM

The performance of the CHARMM cbenchtest, shown in Figure 6, is shown for the InfiniBand cluster, the single-core Cray XD1 and for both single-core and dual-core InfiniPath clusters.. The single-core Cray XD1 is outperformed by both the single and dual-core InfiniPath systems on this test. The single-core InfiniPath system demonstrates significantly higher performance than the Cray XD1. At least a part of this difference can be attributed to the higher clock speed of the InfiniPath system. The scalability benefits of the Cray XD1 and InfiniPath systems over InfiniBand are clearly demonstrated here.

4. Conclusions

We have examined the performance of Cray XD1, Cray XT3 and Cray X1 systems in comparison with equivalent systems from other vendors using a number of full-scale applications from a range of scientific disciplines as well as kernels from the HPC Challenge benchmark suite.

At the high-end we find that the CrayX1 and the Cray XT3 are highly competitive with contemporary systems from IBM and SGI, the precise ranking of these systems being application dependent. For the CFD code, whose triply nested loops vectorise very well, the Cray X1 delivers excellent performance, though matched by the IBM p5-575, all the way out to 1024 processors. The coastal ocean code has different characteristics. It does not vectorise well and has a greater overall communication to computation ratio. For this code the Cray XT3 outperforms systems from IBM and SGI.

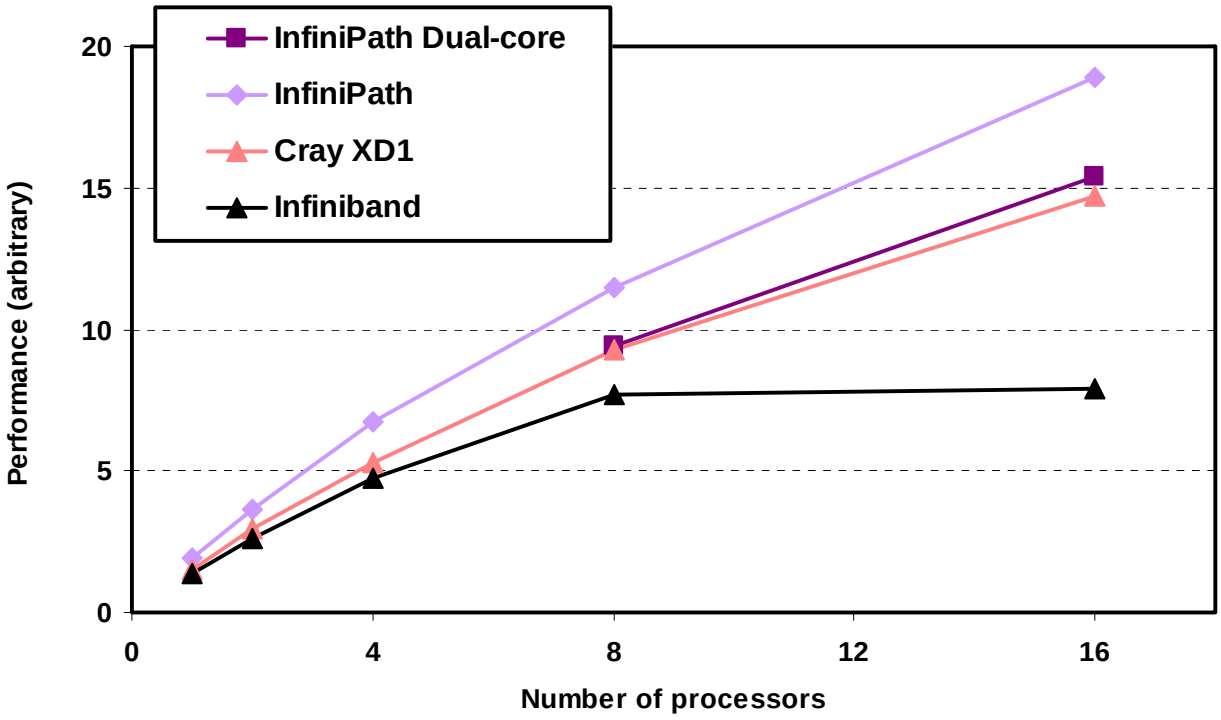

Figure 6: CHARMM cbenchtest performance for the Cray XD1, InfiniBand and InfiniPath (single- and dual-core) systems

We also examine the performance of the Cray XD1 as a mid-range computing resource. It performs well but a Pathscale InfiniPath cluster performs equally well at some fraction of the cost. InfiniPath systems appear particularly competitive for runs on 32 and 64 processors; still considered the *sweet spot* for the majority of applications on mid-range systems. With more competition on its way from the likes of Liquid Computing as well as the PCI-Express version of InfiniPath it seems that a successor to the XD1 is required if Cray are going to provide a cost-effective solution in the mid-range cluster market

5. References

[1] *Cray X1 Evaluation Status Report*, P.A. Agarwal et al (29 authors), Oak Ridge National Laboratory, Oak Ridge, TN, USA, Technical Report ORNL/TM-2004/13 http://www.ccs.ornl.gov/CRAYEvaluationTM2004-15.pdf

[2] *Direct Numerical Simulation of Shock/Boundary Layer Interaction*, N.D. Sandham, M. Ashworth and D.R. Emerson, http://www.cse.clrc.ac.uk/ceg/sbli.shtml

[3] *Single Node Performance of Applications and Benchmarks on HPCx*, M. Bull, HPCx Technical Report HPCxTR0416, (2004) http://www.hpcx.ac.uk/research/hpc/technical_reports/HPCxTR0416.pdf

[4] *Optimization of the POLCOMS Hydrodynamic Code for Terascale High-Performance Computers*, M. Ashworth, J.T. Holt and R. Proctor, HPCx Technical Report HPCxTR0415, (2004) http://www.hpcx.ac.uk/research/hpc/technical_reports/HPCxTR0415.pdf

[5] *Eddy Resolved Ecosystem Modelling in the Irish Sea*, J.T. Holt, R. Proctor, M. Ashworth, J.I. Allen, and J.C. Blackford, in Realizing Teracomputing: Proceedings of the Tenth ECMWF Workshop on the Use of High Performance Computing in

Meteorology, eds. W. Zwieflhofer and N. Kreitz, (2004), 268-278, (World Scientific).

[6] *Vector vs. Scalar Processors: A Performance Comparison Using a Set of Computational Science Benchmarks*, M. Ashworth, I.J. Bush and M.F. Guest, Cray User Group 2005.

[7] *DL_POLY: A general purpose parallel molecular dynamics simulation package*, W. Smith and T.R. Forester, J. Molec. Graphics **14** (1996) 136.

[8] *DL_POLY: Applications to Molecular Simulation*, W. Smith, C. Yong and M. Rodger, Molecular Simulation **28** (2002) 385.

[9] *The DL-POLY Molecular Simulation Package*, W. Smith, http://www.cse.clrc.ac.uk/msi/software/DL_POLY

6. Acknowledgments

The authors would like to thank the Swiss National Supercomputing Centre (CSCS), Oak Ridge National Laboratory (ORNL) and the UK Engineering and Physical Sciences Research Council (EPSRC) for access to machines.

7. About the Authors

Mike Ashworth is Head of the Advanced Research Computing Group in the Computational Science & Engineering Department (CSED) at CCLRC Daresbury Laboratory and has special interests in the optimisation of coupled environmental models for high-performance systems. **Martyn Guest** is Associate Director of CSED and leads the HPCx Terascaling Team. He is a leading researcher in the application of computational chemistry methods on high-performance computers. The Distributed Computing Group within CSED - **Miles Deegan**, **Christine Kitchen**, **Igor Kozin** and **Richard Wain** – have interests in the areas of performance evaluation of commodity based systems and novel architectures among others.

All authors can be reached at CCLRC Daresbury Laboratory, Warrington WA4 4AD, UK.

E-Mails: m.ashworth@dl.ac.uk m.j.deegan@dl.ac.uk m.f.guest@dl.ac.uk c.a.kitchen@dl.ac.uk i.kozin@dl.ac.uk r.wain@dl.ac.uk.