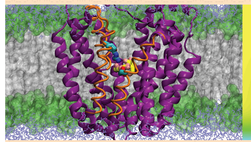
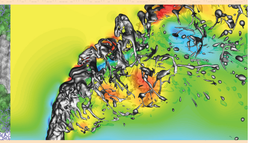
Abhinav Bhatelé, Lukasz Wesolowski, Eric Bohm, Edgar Solomonik, Laxmikant V. Kalé

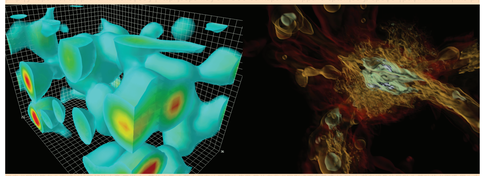
Performance Comparison of Intrepid, Jaguar and Ranger Using Scientific Applications

Introduction

This poster presents a performance comparison of three leading supercomputers: Intrepid, Jaguar and Ranger. We use architectural specifications and benchmarks to explain differences in application performance on these systems. The programs we picked are representative of the classes of applications selected by NSF for benchmarking the Track I supercomputer: NAMD, MILC, and DNS, a turbulence code.







Machines

	Intrepid	Ranger	Jaguar
Location, Year	ANL, 2007	TACC, 2008	ORNL, 2008
No. of Nodes (Cores per Node)	40,960 (4)	3,936 (16)	7,832 (4)
CPU Type (Clock Speed, MHz)	PowerPC 450 (850)	Opteron (2300)	Opteron (2100)
Peak TeraFLOPS (GFLOPS per Core)	557 (3.4)	579 (9.2)	260 (8.4)
Memory per Node (per Core), GB	2 (0.5)	32 (2)	8 (2)
Memory BW per Node (per Core), GB/s	13.6 (3.4)	21.3 (1.3)	10.6 (2.65)
Type of Network (Topology)	Custom (3D Torus)	Infiniband (full-CLOS)	SeaStar2 (3D Mesh)
Link Bandwidth GB/s	0.425	1	3.8
L1, L2, L3 Cache Sizes KB, KB, MB	64, 2, 8	128, 512, 2	128, 512, 2

Table 1: Specifications of the parallel machines used for the runs

System	Comm. Bandwidth per Core (MB/sec)	Comm. Bandwidth per FLOP (bytes)	MFLOPS per Watt
Blue Gene/P	1,275	0.375	357
Ranger	1,000	0.109	217
XT4	11,400	1.357	130

Table 2: Comparison of network bandwidth and FLOP ratings for the parallel machines

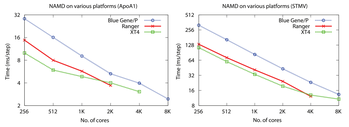
Intrepid, located at Argonne National Laboratory (ANL), is a 40-rack installation of IBM's Blue Gene/P supercomputer. Each rack contains 1,024 compute nodes consisting of four PowerPC450 cores. A midplane of 512 nodes forms a 3D torus of dimensions 8 X 8 X 8. Larger tori are formed from this basic unit. Each node is attached to additional specialized networks for collective communication, barriers/interrupts, and machine control.

Jaguar (XT4 partition) at Oak Ridge National Laboratory (ORNL), comprises 7,832 compute nodes. Each node contains a quad-core Barcelona AMD Opteron processor, and is connected through a HyperTransport (HT) link to a Cray Seastar router. The routers form a 3-dimensional mesh of dimensions 21 X 16 X 24. It is a mesh in the X dimension and a torus in the Y and Z dimensions.

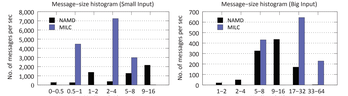
Ranger, located at Texas Advanced Computing Center (TACC), consists of 3,936 nodes connected using Infiniband technology in a full CLOS fat-tree topology. The machine comprises 328 12-node compute chassis, each with a direct connection to the two top-level switches. The nodes are Sun-Blade x6420 quad-socket blades with AMD quad-core Barcelona Opteron processors.

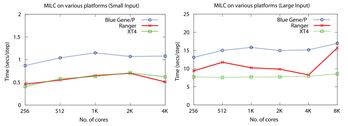
NAnoscale Molecular Dynamics

NAMD is a highly scalable MD code written in Charm++. It is latency tolerant and runs in L2 cache on most machines. Since NAMD is always run in a strong scaling mode, it stresses the machines and is good for performance benchmarking. Performance of NAMD on the three machines is in the ratio of their peak FLOP ratings, as expected.



A general trend which we notice in the performance plots of all the applications is that Blue Gene/P performs as expected, achieving a high fraction of its peak performance. This can be attributed to some of the design features and characteristics of BG/P which govern performance in many cases: 1. Highest memory bandwidth per core (3.4 GB/sec), 2. Biggest L3 cache (2 MB per core), 3. Smallest latencies for small-sized messages (4 to 1024 bytes), 4. No evidence of noise and 5. Well-implemented MPI collectives.





MIMD Lattice Computation

MILC stands for MIMD Lattice computation and is used for large scale numerical solutions to study quantum chromodynamics. For benchmarking, we used the application ks_imp_dyn, which simulates full QCD with dynamical Kogut-Susskind fermions. We used two input sizes: a 4 x 4 x 4 x 4 x 4 per core grid, which we expected to fit in cache, and a larger 8 x 8 x 8 x 8 per core grid to fit in memory.

For the small input, XT4 and Ranger have similar performance, which is roughly two times better than that of BG/P. For the bigger input, MILC is 33% faster on Ranger than on BG/P at 256 cores and the performance on XT4 is slightly better than on Ranger.

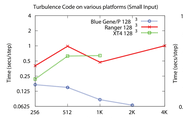
The large input stresses memory bandwidth in addition to the network. Ranger, which has the lowest memory bandwidth, is thus able to perform only slightly better than BG/P, while XT4, with reduced network contention, outperforms the other two machines. Reducing the number of cores used per node for the runs eases the stress both on the memory subsystem and the network.

	MILC (secs/step)				DNS (secs/step)			
Input	Small Input		Large Input		128^3		512^3	
#cores	256	512	256	512	256	512	256	512
XT4 4 cpn	0.41	0.58	7.70	7.57	0.219	0.626	4.96	2.74
XT4 2 cpn	0.35	0.43	5.25	5.68	115.46		121.70	
Ranger 16 cpn	0.46	0.55	13.09	15.02	0.402	0.983	10.63	5.86
Ranger 8 cpn	0.31	0.49	7.76	7.98	0.292	0.928	7.55	4.04

Table 3: Runs on XT4 and Ranger using fewer than all cores per node

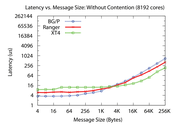
DNS Turbulence Code

DNS is a turbulence code developed at Sandia National Laboratory. This code solves viscous fluid dynamics equations in a periodic rectangular domain (2D or 3D). For benchmarking, we use the purest form of the code: Navier-Stokes with deterministic low wave number forcing. The results use strong scaling for two grid sizes: 128^3 and 512^3.



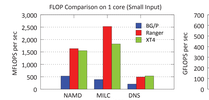
For the smaller input, DNS scales and shows good performance only on BG/P. For the larger input, the application scales well on all platforms. However, Ranger performs roughly on the same level as BG/P despite having faster processors, and XT4, though it executes well on up to 2K cores, shows no scaling past that point.

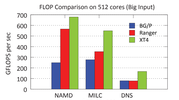
DNS does a large number of small FFTs, which lead to all-to-all communication. It sends a large number of small messages (128 to 256 bytes) - approximately 6700 MPI_Isend's per second per processor. The performance of DNS on XT4 and Ranger is hindered due to large overheads for small messages, as is apparent from the benchmark results on the right.



FLOP Comparison

The figures to the bottom right demonstrate the flop performance of all three applications on the three platforms for small and large input sizes. We can see that Ranger has the best single core performance (attributed to its high FLOPS rating). XT4 gives the best performance on 512 cores, sometimes twice that on Ranger. In fact, Ranger's performance drops quite close to BG/P for MILC and DNS. This can be attributed to the memory bandwidth and network contention issues, which are particularly detrimental on Ranger. NAMD, on the other hand, runs in cache and is latency tolerant and hence is not affected by these issues.





Conclusion

In summary, vendor specifications, benchmarking results, and application characteristics can and should be combined to form a more complete picture of application performance to guide the expectations of the supercomputer user community. We hope that the analysis techniques presented in this poster will assist application developers and users.

References:

Blue Gene/P: Downloaded from the Media Kit on the Argonne Leadership Computing Facility website: http://www.alcf.anl.gov/news/media_kit.php
Jaguar: Downloaded from the Media Center on the National Center for Computational Sciences website: http://www.nccs.gov/media-center/image-gallery/
Ranger: courtesy of Texas Advanced Computing Center and Advanced Micro Devices
Lactose Permease: Generated using VMD, Image downloaded from the Theoretical and Computational Biophysics website: http://www.ks.uiuc.edu/Research/lac/
Other Simulation Images: Downloaded from the Scientific Visualization Gallery on the Texas Advanced Computing Center website: http://cms.tacc.utexas.edu/scivis-gallery/

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
PARALLEL PROGRAMMING LAB
PPL UIUC
DEPT OF COMPUTER SCIENCE, UNIVERSITY OF ILLINOIS