PROCEEDINGS OF SCIENCE

Application-based Network Performance Profiling

Robin Pinning*
University of Manchester
E-mail: pinning@manchester.ac.uk

The large scale modelling of many physical phenomena increasingly requires the model to be of a size that is too large for one HPC resource. MPICH-G2 is a grid-enabled MPI implementation that allows the coupling of multiple machines for the running of a single MPI-based application. This work aims to show how the use of light-switched optical networks, such as UKLight, affect codes of this class by running a series of benchmarks using the Intel MPI benchmarking suite

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

*Speaker.

PoS(ESLEA)011

1. Introduction

Traditionally networking, particularly for academic scientific researchers, has been over best effort TCP/IP packet switched networks such as SuperJANET. UKLight is an optical network comprising of a 10Gbit/s backbone connecting participating academic institutions in the UK and connecting to global optical networks such as Starlight and NetherLight.

The ESLEA [1] project aims to demonstrate the potential of circuit-switched optical networks by allowing the exploitation of the UKLight infrastructure for a range of scientific application-led projects. One of these sub-projects is RealityGrid.

The RealityGrid [2]infrastructure provides the computational scientist with a framework for computational steering and on-line visualization. The use of these interactive techniques provides some unique demands on the networking infrastructure, requiring both high bandwidth (TeraGyroid experiment [3]) and high QoS (SPICE [4]) depending on the requirement of the scientific application being used.

Recently the project has extended the scientific applications used by researchers to make use of meta-computing via MPICH-G2 middleware [5]. To overcome the inherent bottleneck of including a relatively slow, compared with the internal HPC network, wide-area network the scientific application needs both high bandwidth and high QoS. Along with careful porting of the application optical networks are essential. One such approach that has been successfully demonstrated at SC2005 across a trans-Atlantic link which included UKLight, is that taken by the Vortonics project [6]. For this application, another lattice-Boltzmann code with similarities to the LB3D code used by RealityGrid researchers, the memory requirements are typically larger than a single computational resource can provide. The code is parallelised with MPI with data distributed accross the available processors according to a scheme called geographically distributed domain decomposition or GD^3 [7].

Given the importance of the use of grid-enabled application codes using MPICH-G2 this paper presents performance figures gathered using the Intel MPI benchmarking suite run on the UKLight network. A brief analysis, some conclusions and some suggestions for future work are also presented.

2. Methodology

2.1 Equipment

The experiments involved three linux workstations situated in London and Edinburgh. The machines were connected by a dedicated UKLight link provisioned at 300Mbps. Two of the machines were configured as compute nodes with a third, running a NIST Net-instrumented linux kernel, acting as a traffic shaper connected in-between the two compute nodes. One compute node and the NIST Net box were situated in Edinburgh, connected via gigabit ethernet. The NIST Net box was then connected via the UKLight router and link to the machine at UCL.

2.2 Software

NIST Net [8] version 2.0.12 was run as a kernel module. All kernels were version 2.6.9x and all OS level TCP-stack parameters were left as standard. When using HPC machines there is

usually no way for the user to alter these parameters therefore all tuning was done at the user-level. The network parameters altered in the NIST Net module were chosen so as to emulate conditions seen in packet-switched 'production' networks were α, the packet transmission delay, to emulate different network latencies; β, the packet loss, to emulate loss due to congestion and $\delta(\alpha)$, the jitter, to emulate variability in packet delivery.

The middleware stack on each machine consisted of Globus Toolkit v4.0.3 (pre-WS components) and MPICH-G2 v1.2.6. The Intel MPI Benchmark software version 3.0 was built against MPICH-G2 with custom parameters to allow for large data sizes (to emulate the size of data that application codes such as LB3D pass around). Due to the use of only two processors only the Ping-Pong test was used. The parameter "MPICH_GLOBUS2_TCP_BUFFER_SIZE" was set, based on instructions from Brian Toonen (MPICH-G2 developer), to 524288 in the RSL used to launch each run. This was based on a measured RTT of 5ms.

3. Analysis

The data for the added latency can be seen in table 1 and plotted in figure 1. As the latency, α, is increased the time for the transfer increases. This is unsurprising as the performance of MPI-based codes is very sensitive to latency in the connection between processes. As the packet loss rate, β, is increased the transfer time also increases, again an expected result.

Results of increasing jitter, $\delta(\alpha)$, can be seen in table 2 and are plotted in figure 2. Overall the results present few surprises other than the slower than expected transfers for 16MB transfers when no jitter is present at high latency.

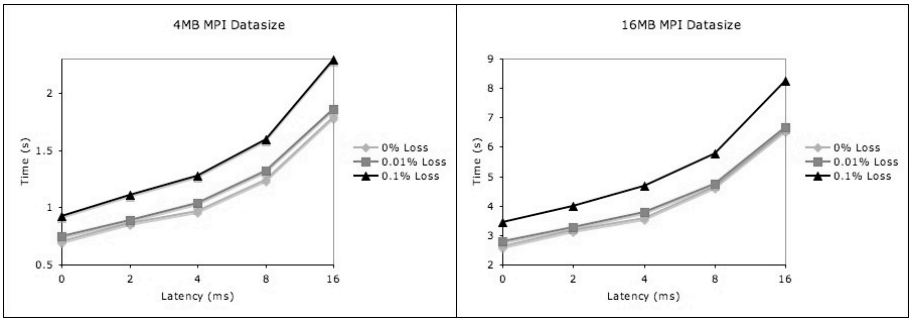

Figure 1: Plots showing transfer time in seconds for a two MPI message sizes for varying values of packet delay, α, and packet loss, β.

4. Summary

The data collected, due to time constraints, is a small snapshot of what could be achieved using these methods. In fact the difficulty faced in setting up, and maintaining, the UKLight link used illustrates how difficult it still is for application scientists to utilise these links. Especially when complex software stacks are sat on top of them, as debugging problems becomes very difficult. It is also clear that if these tests were to be repeated using HPC-class machines that some form of advance reservation and co-allocation [9] of those resources is essential for this kind of work.

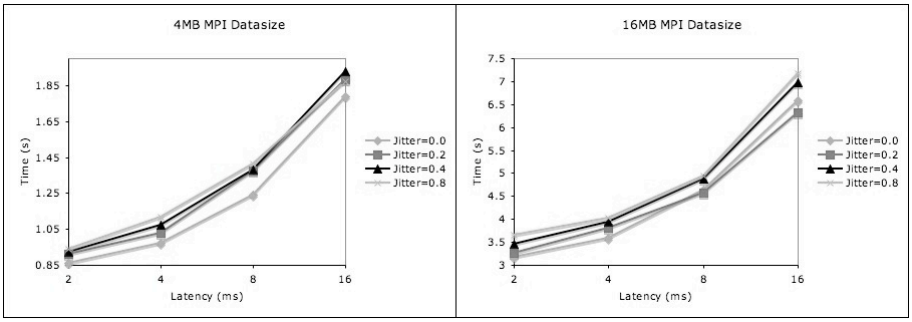

Figure 2: Plot showing transfer time in seconds for two MPI message sizes for varying values of packet delay, α, and jitter, $\delta(\alpha)$.

	Latency α (*ms*)				
Packet Loss β %	0.0	2.0	4.0	8.0	16.0
4*MB MPI Datasize*					
0.0	0.703	0.860	0.971	1.242	1.789
0.01	0.743	0.889	1.043	1.318	1.863
0.1	0.927	1.106	1.279	1.592	2.288
1.0	1.997	2.415	2.799	3.489	4.956
16*MB MPI Datasize*					
0.0	2.618	3.182	3.590	4.630	6.576
0.01	2.777	3.260	3.777	4.760	6.682
0.1	3.449	4.002	4.691	5.774	8.229
1.0	7.760	9.248	10.592	13.694	19.427

Table 1: Table showing transfer time in seconds for respective values of packet delay, α, and packet loss, β.

	Latency α (*ms*)			
Jitter $\delta(\alpha)$	2.0	4.0	8.0	16.0
4*MB MPI Datasize*				
0.2	0.910	1.027	1.372	1.881
0.4	0.922	1.071	1.379	1.928
0.8	0.940	1.117	1.416	1.872
16*MB MPI Datasize*				
0.2	3.266	3.810	4.577	6.329
0.4	3.464	3.940	4.879	6.977
0.8	3.649	4.029	4.947	7.165

Table 2: Table showing transfer time in seconds for respective values of packet delay (α) and jitter $\delta(\alpha)$.

This work points the way to further studies such as tuning of the TCP stack on host machines to get better bandwidth utilisation or investigation of the effect the underlying protocol (TCP) has on performnce compared with newer protocols such as Reliable-Blast UDP (RBUDP).

This research has been funded by ESLEA grant GR/T04465. I would like to acknowledge the invaluable help of Barney Garrett, Clive Davenhall and Nicola Pezzi in configuring the network and hardware infrastructures; and Radhika Saksena for application-specific discussions.

References

[1] ESLEA, *Exploitation of Switched Lightpaths for eScience Applications*, http://www.eslea.uklight.ac.uk

[2] S. M. Pickles, R. Haines, R. L. Pinning and A. R. Porter, *A Practical Toolkit for Computational Steering, Phil Trans R Soc*, **363**, 1833, pp. 1843-1853, 2005, http://dx.doi.org/10.1098/rsta.2005.1611

[3] M. Mc Keown, S. M. Pickles, A. R. Porter, R. L. Pinning, M. Riding and R. Haines, *The Service Architecture of the TeraGyroid Experiment*, *Phil Trans R Soc*, **363**, pp. 1743-1755, 2005.

[4] S. Jha, P. V. Coveney, M. J. Harvey, and R. Pinning, *SPICE: Simulated Pore Interactive Computing Environment, Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, p. 70, 2005, http://dx.doi.org/10.1109/SC.2005.65

[5] N. Karonis, B. Toonen, and I. Foster, *MPICH-G2: A Grid-Enabled Implementation of the Message Passing Interface*, *Journal of Parallel and Distributed Computing (JPDC)* **63**, No. 5, pp. 551-563, 2003.

[6] B. M. Boghosian, P. V. Coveney, S. Dong, L. I. Finn, S. Jha, G. Karniadakis and N. Karonis, *Nektar, SPICE and Vortonics: Using Federated Grids for Large Scale Scientific Applications*, *Proceedings of Challenges of Large Applications in Distributed Environments*, 34-42, 2006. IEEE Catalog Number: 06EX1397 ISBN 1-4244-0420-7 Library of Congress 2006925560.

[7] B. Boghosian, L. I. Finn and P. V. Coveney, *Moving the data to the computation: multi-site distributed parallel computation*, 2006, http://www.realitygrid.org/publications/GD3.pdf

[8] M. Carson, D. Santay, *NIST Net - A Linux-based Network Emulation Tool*, *Computer Communication Review*, **6**, 2003.

[9] J. MacLaren, M. McKeown and S. Pickles, *Co-Allocation, Fault Tolerance and Grid Computing*, *Proceedings of the UK e-Science All Hands Meeting 2006*, 155-162, 2006.