BRINGING HIGH-PERFORMANCE NETWORKING TO HEP USERS

R. Hughes-Jones*, S. Dallison, Department of Physics and Astronomy, The University of Manchester, Oxford Rd., Manchester M13 9PL, UK
N. Pezzi, Y-T. Lee, Department of Physics and Astronomy, University College Gower Street, London, WC1E 6BT, UK

Abstract

In a collaboration between the BaBar experiment and the UK e-Science network project, Managed Bandwidth – Next Generation, MB-NG, we demonstrated how the new TCP/IP transport protocol stacks could be used to achieve high performance data transport in a real HEP experiment. This paper reports on investigations of the performance of these TCP stacks and their use with data transfer applications such as GridFTP, bbftp, bbcp and Apache. End-host performance was also examined in order to determine the effects of the Network Interface Card, NIC, the PCI bus, and the disk and RAID sub-systems.

INTRODUCTION

The BaBar [1] Particle Physics experiment is a large international collaboration based at the Stanford Linear Accelerator Center (SLAC), California, USA and the Tier A center for the UK is based at Rutherford Appleton Laboratory (RAL). From here data is distributed to the various UK institutes participating in BaBar using SuperJANET [2], which is the UK's academic network run by UKERNA. As many Gigabytes of data must be transferred between the disk servers at RAL and the local sites, efficient use of the network and compute resources is essential.

The tests compared the transfer performance for real BaBar data when moved between the servers purchased by the experiment and high-performance PCI-X based servers. Different data transfer applications were tested using new TCP protocol stacks. The following combinations were used

- BaBar servers on the SuperJANET network
- MB-NG servers on the SuperJANET network
- MB-NG servers on the MB-NG testbed.

The MB-NG Network

Figure 1 shows the MB-NG testbed network [3]. This comprises three "edge" domains built from CISCO 7600 series Optical Switch Routers, using 1 Gbit/s Ethernet QoS enabled line cards for the LAN connections and 2.5Gbit/s SDH interface cards to connect to the core network. These edge domains are connected via the SuperJANET development core network comprising 4 carrier class CISCO GSR 12000 series routers similarly equipped with leading edge 2.5 Gbit/s QoS enabled line cards. Under the conditions used for these tests the MB-NG network was not congested and there was no (or extremely little) packet loss.

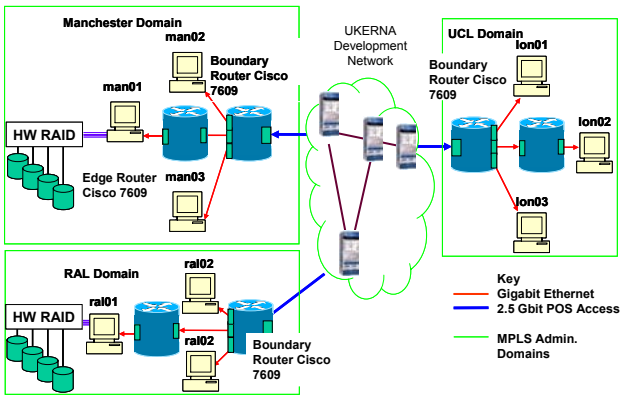

Figure 1: The MB-NG Development Network.

SYSTEM COMPONENT PERFORMANCE

TCP Stacks

Advanced network protocols implementing sender side modifications to TCP have shown highly increased bandwidth utilisation in long delay high bandwidth and multi-user network environments [4]. This allows a single stream of a modified TCP stack to transmit at rates that would otherwise require multiple streams of standard TCP. The High Speed TCP and Scalable TCP stacks were used in these tests.

TCP has two phases, Slow Start, where it exponential probes the capacity of the network, and Congestion Avoidance, where most transmission takes place. Standard TCP (sometimes called vanilla TCP) assumes that the detection of a lost packet indicates congestion, and reduces the transmission rate by half. After that Standard TCP increases the throughput buy one Maximum Transmission Unit, MTU, per Round Trip Time, RTT, on the network. It can take a considerable time to get back to the initial transmission rate for long distance connections. Both High Speed and Scalable TCP make the reduction of rate less and increase the transmission rate faster than standard TCP.

Figure 2 shows the throughput achieved by the different stacks when packets are deliberately dropped in the kernel of the receiving host. The MB-NG network (RTT 6.2ms) is shown on the left and the DataTAG [5] link from Geneva to Chicago (RTT 128 ms) on the right. The further to the right one is on the plots, the faster

*R.Hughes-Jones@man.ac.uk

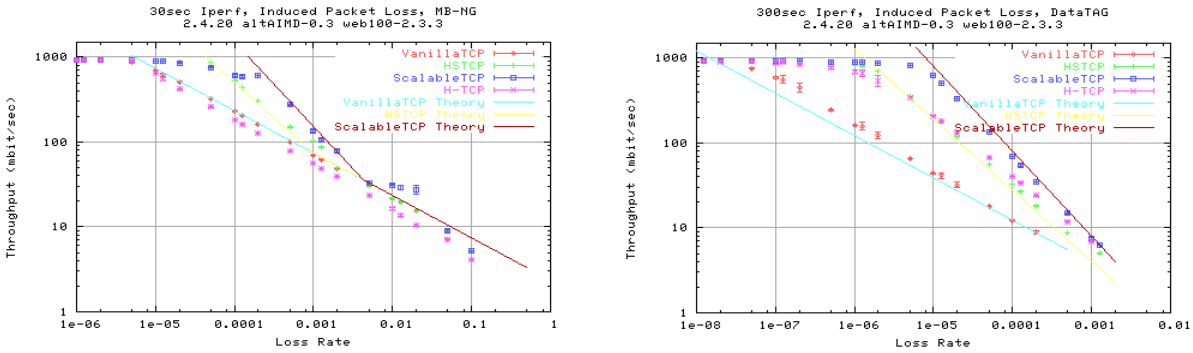

Figure 2: Experimental and theoretical results of the throughput achieved by different TCP stacks as a function of the induced packet loss. The packets were dropped in the receiving kernel. The MB-NG network (RTT 6.2 ms) is on the left and the DataTAG network (RTT 128 ms) on the right.

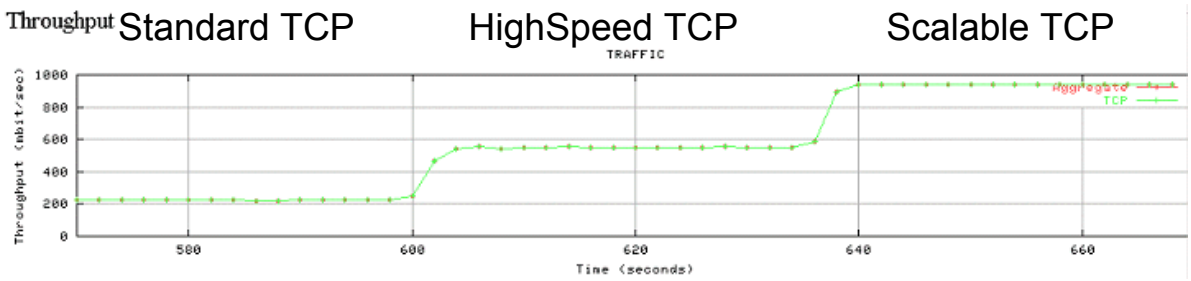

Figure 3: Average memory-memory throughput as a function of time for Standard, High Speed, and Scalable TCP stacks when operated over the MB-NG network with an induced packet loss of 1 in 25000.

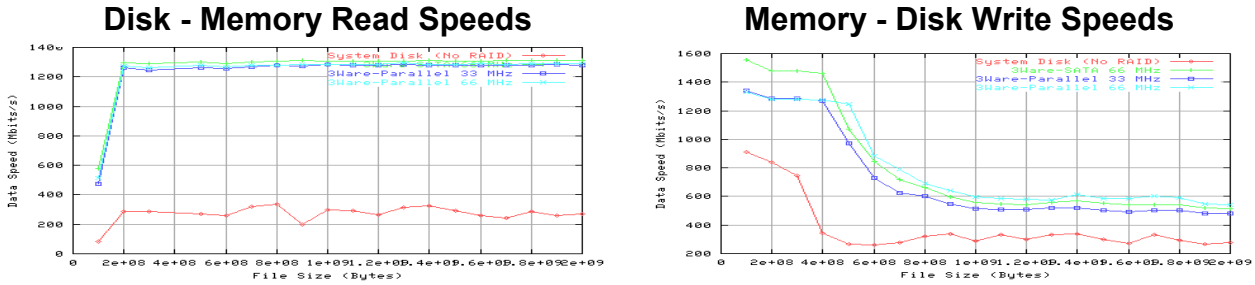

Figure 4: Memory to disk sub-system transfer rates as a function of the file size for various RAID5 controllers compared with the performance of the system disk (red line).

the recovery from packet loss. The agreement between theory and measurements is very good. Another way of visualizing the impact of the alternate TCP stacks is shown in Figure 3. This shows the throughput as a function of time for each of the stacks for memory to memory transfers over the MB-NG network when there is a packet loss of 1 in 25,000

End Hosts and NICs

It is important that the end hosts should have sufficient CPU power, memory bus bandwidth and Input/Output, I/O, capability that packets are not dropped in the end host itself. A methodology [6] for evaluating the end host network performance by using UDP packets to measure the latency, throughput, and the activity on the PCI/PCI_X buses was for these systems. Figure 5 shows the UDP achievable throughput when the UDPmon [7] tool was used to transmit streams of UDP packets at regular, carefully controlled intervals between two of the MB-NG systems connected back to back. There was no packet loss during these tests and the plot shows the system was capable of operating at line speed for packets greater that 1200 bytes.

RAID Controllers and Disks

The performance of various RAID controllers and disk sub-systems was evaluated by measuring the transfer rates between memory and the disk sub-system using a single flow of sequential reads or writes. Figure 4 shows the transfer rates as a function of the file size. The tests used RAID5 and were performed on a PC with a Supermicro P4DP6 motherboard with dual 2.0 GHz Zeon CPUs.

The red lines show the performance of the system disk. The green lines show an ICP serial-ATA controller card with a 8 disk RAID5 array. The dark blue, purple and light blue lines show results for the same disk array except using the following controllers respectively; 3Ware serial-ATA; 3Ware parallel ATA with a 33 MHz PCI bus interface; 3Ware parallel ATA with a 66 MHz

PCI bus interface. The 3Ware serial-ATA controller was used for the transfer tests reported in this paper and could read at 1200 Mbit/s and write at 500-550 Mbit/s for large files. There is a general trend that files less than 400 Mbytes can be transferred over three times faster than larger files. Note that in general the applications are optimised for transferring large files.

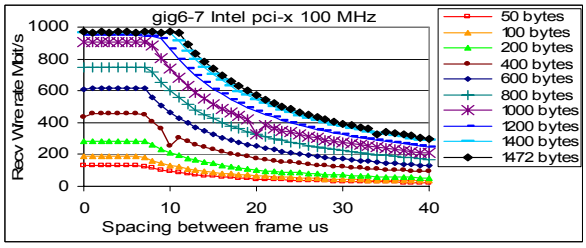

Figure 5: UDP Throughput as a function of the time between transmitting the frames for Intel Pro/1000 NIC on the SuperMicro P4DP6 motherboard.

DATA TRANSFER MEASUREMENTS

Memory to Memory Tests

iperf was used to investigate the memory to memory behaviour of the end hosts when connected to the different networks. Web100 [8] was used to instrument the TCP stacks. Figure 6 shows the behaviour of the TCP stack as a function of time when using the HighSpeed TCP. The top plot shows that the Supermicro servers connected to the MB-NG network can achieve TCP throughput of 960 Mbit/s and no packet loss occurred as expected on this network. The middle and bottom plots show the BaBar hosts connected to SuperJANET. The saw-tooth behaviour of the achievable TCP throughput is typical when there is packet loss. Which is confirmed by the number of duplicate ACKs recorded, as shown in the bottom plot.

Disk to Disk Transfers

Figure 7: Compares the disk-to-disk throughput of bbftp when used with Highspeed TCP and the BaBar hosts on SuperJANET, the MB-NG hosts on SuperJANET and the MB-NG hosts on the MB-NG Network. For the tests with MB-NG hosts there is evidence of higher throughput for the first few seconds of the transfer, and this could reflect the write behaviour of the RAID5 disks shown in Figure 4. In all cases the throughput is asymptotic to around 400 Mbit/s.This is consistent with the RAID5 performance which limits at 500-550 Mbit/s. Extra data movements or the application may account for the difference in rates.

Comparison of the behaviour of the TCP Congestion Window with the TCP throughput, as shown in Figure 8, suggests that the rapid variation in throughput is not due to the action of the TCP protocol but is more influenced by the performance of the disk sub-system, the memory bus and I/O performance, and the application design.

Figure 9 shows comparison of the performance of different data moving applications. These tests were made using Highspeed TCP to move a 2 Gbyte file between

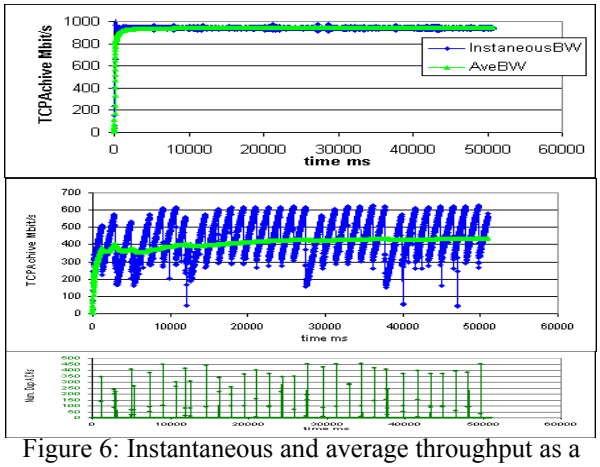

Figure 6: Instantaneous and average throughput as a function of time
Top plot MB-NG host on MB-NG Network
Middle plot: BaBar host on SuperJANET
Bottom plot: Duplicate ACKs for BaBar host

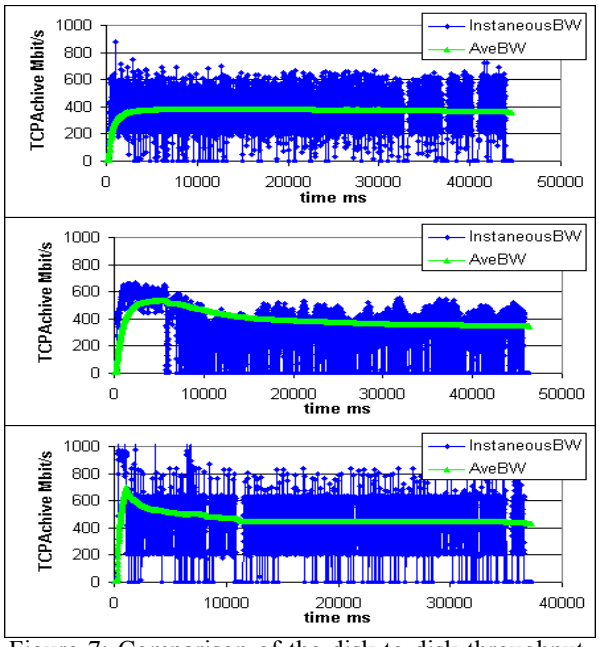

Figure 7: Comparison of the disk-to-disk throughput bbftp was used with Highspeed TCP:
Top plot BaBar host on SuperJANET
Middle plot: MB-NG host on SuperJANET
Bottom plot: MB-NG host on MB-NG Network

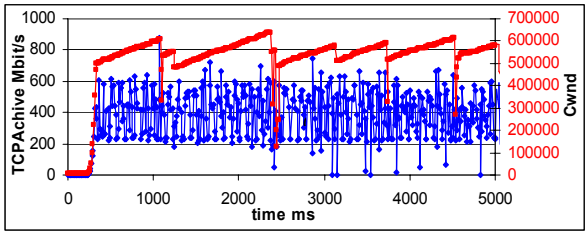

Figure 8: TCP Throughput (blue points) and the TCP Congestion Window (red line) as a function of time

Supermicro servers connected with the MB-NG network.With the possible exception of Gridftp, which is slightly lower, the transfers are limited by the speed of the disk sub-system.

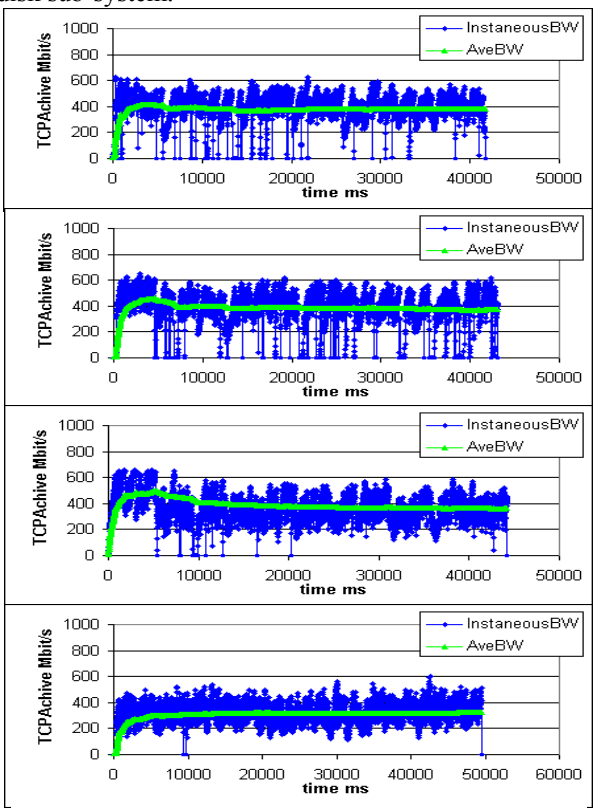

Figure 9: Comparison of the disk-to-disk throughput for bbcp, bbftp, apache, gridftp. Highspeed TCP was used on the MB-NG systems connected using SuperJANET.

App	TCP Stack	SuperMicro MB-NG	SuperMicro SuperJANET	BaBar SuperJANET
Iperf	Standard	940	350-370	425
	HighSpeed	940	510	570
	Scalable	940	580-650	605
bbcp	Standard	434	290-310	290
	HighSpeed	435	385	360
	Scalable	432	400-430	380
bbftp	Standard	400-410	325	320
	HighSpeed		370-390	380
	Scalable	430	345-532	380
apache	Standard	425	260	300-360
	HighSpeed	430	370	315
	Scalable	428	400	317
Gridftp	Standard	405	240	
	HighSpeed		320	
	Scalable		335	

Figure 10 Table of the throughputs in Mbit/s achieved for the different data transfer applications.

Figure 10 shows a table of the throughputs in Mbit/s achieved when moving 2 G byte files across the three different network - host configurations using the various applications and TCP stacks. In general the throughput decreases going from MB-NG hosts on the development network to MB-NG hosts on the production network to BaBar hosts on the SuperJANET production network.

CONCLUSIONS

It is clear that packet loss makes a major contribution to lowering the TCP throughput and effort in working with campus network engineers to reduce this is worthwhile. We have shown that the advanced TCP stacks recover much faster from the effect of packet loss and this is much more important for longer RTTs. Performance of the end hosts and the disk sub-systems is critical. Hosts should have plenty of CPU power memory bus bandwidth and Input/Output, I/O, capability. Separation of the network and disk sub-system onto different PCI-X buses is recommended. The results presented here are dominated by the performance of the particular RAID system tested and the interaction between the network and disk systems. Further work is in progress to study the behaviour of later models of RAID controller and the use of advanced RAID configurations.

ACKNOWLEDGEMENTS

The work reported in this paper has been a result of the close collaboration of members of the BaBar experiment, the MB-NG e-Science project, the Network Engineers at Manchester and RAL, and the production computing support personnel at RAL. We would like to thank all those involved in the collaboration.

REFERENCES

[1] The web site of the BaBar experiment http://www.slac.stanford.edu/BFROOT/
[2] SuperJANET topology http://www.ja.net/topology/index.html
[3] UK e-Science MB-NG project home page http://www.mb-ng.net/frontpage.html
[4] H. Bullot, R. L. Cottrell, R. Hughes-Jones, "Evaluation of Advanced TCP Stacks on Fast Long-Distance Production Networks," Journal of Grid Computing, Volume 1, Issue 4, 2003, Pages 345 - 359
[5] European DataTAG home page http://datatag.web.cern.ch/datatag
[6] R. Hughes-Jones, P. Clarke, S. Dallison, "Performance of 1 and 10 Gigabit Ethernet Cards with Server Quality Motherboards," Future Generation Computer Systems Special issue, 2004
[7] UDPmon: a Tool for Investigating Network Performance, http://www.hep.man.ac.uk/~rich/net
[8] Web100 Project home page, http://www.web100.org/