

Investigation of the interaction between high-performance networking and disk sub-systems

Richard Hughes-Jones and Stephen Dallison

Abstract— In collaborations between the BaBar experiment, the UK e-Science network project, Managed Bandwidth – Next Generation (MB-NG), and UKLight, we demonstrated how the new TCP/IP transport protocol stacks can be used to achieve high performance data transport in a real HEP environment. 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 with a curl-based client utility. End-host performance was examined in order to determine the effects of the Network Interface Card, NIC, the PCI bus, and the disk and RAID sub-systems.

Index Terms— Data Communication systems, Gigabit Ethernet, Raid Disk systems, Transport Protocols

I. INTRODUCTION

In collaborations between the BaBar experiment, the UK e-Science network project, Managed Bandwidth – Next Generation (MB-NG), and UKLight, we demonstrated how the new TCP/IP transport protocol stacks can be used to achieve high performance data transport in a real HEP environment. 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 with a curl-based client utility. End-host performance was examined in order to determine the effects of the Network Interface Card, NIC, the PCI bus, and the disk and RAID sub-systems.

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 centre 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 MB-NG network[3] is a test bed comprising 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.

Measurements across Europe involve the National Research Networks in each country and the GEANT backbone that links these national networks.

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 as well as different combinations of servers and networks were used, or are being investigated. These include:

- BaBar servers on the SuperJANET network.
- MB-NG servers on the SuperJANET network.
- MB-NG servers on the MB-NG tested.
- High performance servers over European Academic Network.
- High performance servers connected on transatlantic links.

In addition, tests were made with the RAID servers loaned by Boston Ltd. [10] for the Bandwidth Challenge at Super Computing 2004 (SC2004), using the UKLight network infrastructure [11]. The UKLight Point of Access in London provides international connectivity with 10 Gbit network connections to peer facilities in Chicago (Starlight) and Amsterdam (Netherlight). Each 10 Gigabit Lambda operated at level 2 and presented independent 1 Gigabit Ethernet circuits to the user.

II. SYSTEM COMPONENT PERFORMANCE

A. 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. High Speed TCP and Scalable TCP stacks were

Manuscript received January 20, 2005. This work was supported in part by the U.K. e-science program through the MB-GN Project and in part by the UK Particle Physics and Astronomy Research Council.

R.E Hughes-Jones is with The School of Physics and Astronomy, The University of Manchester, Oxford Rd., Manchester, M13 9PL, U.K. (phone: +44 161 275 4117; fax: +44 161 273 5867; e-mail: R.Hughes-Jones@manchester.ac.uk).

S. Dallison was with The University of Manchester, Oxford Rd., Manchester, M13 9PL, U.K.. He is now with The Rutherford Appleton Lab., CCLRC, Chilton, Didcot, Oxon, OX11 0QX, UK. (e-mail: S.Dallison@rl.ac.uk).

used in these network-disk investigations; as both make the reduction of the rate less severe when detecting packet loss, whilst increasing the transmission rate more aggressively than standard TCP during the recovery. Tests performed over the MB-NG and DataTAG[6] networks show that the agreement between theory and measurements is very good [5].

B. 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 for both networking and the disk sub-system. For the network, packets should not be 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 used to evaluate these PC systems. UDPmon [8] was used to send streams of UDP packets spaced at regular, carefully controlled intervals between the server systems connected back to back. For the MB-NG and the SC2004 systems, there was no packet loss during these memory-to-memory tests and it was shown that the systems were capable of operating at line speed for packets greater than 1200 bytes. The

methodology was also applied to each of the networks tested to measure end-to-end performance and characterise packet loss.

C. 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. Both RAID 0 and RAID 5 tests were made and will be reported on. Figure 1 shows a comparison of the transfer rates as a function of the file size for different RAID Controllers; these RAID5 tests 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 an 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.

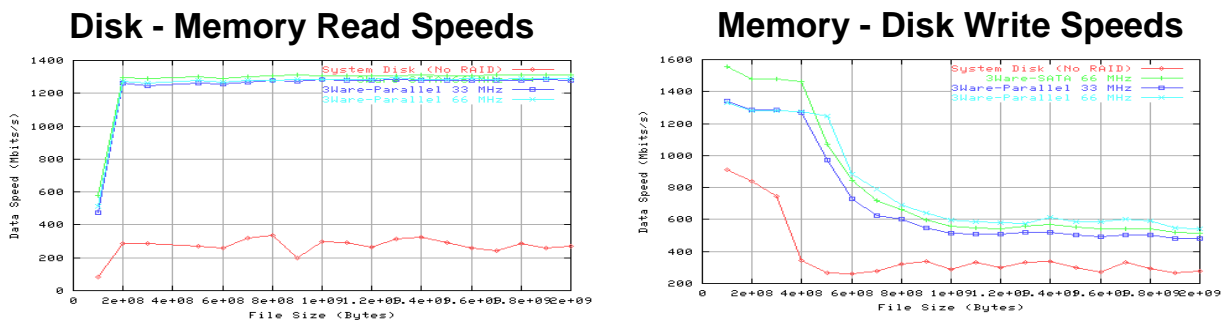


Figure 1: Memory to disk sub-system transfer rates as a function of the file size for various RAID5 controllers compared with the performance of the system disk (lowest red line).

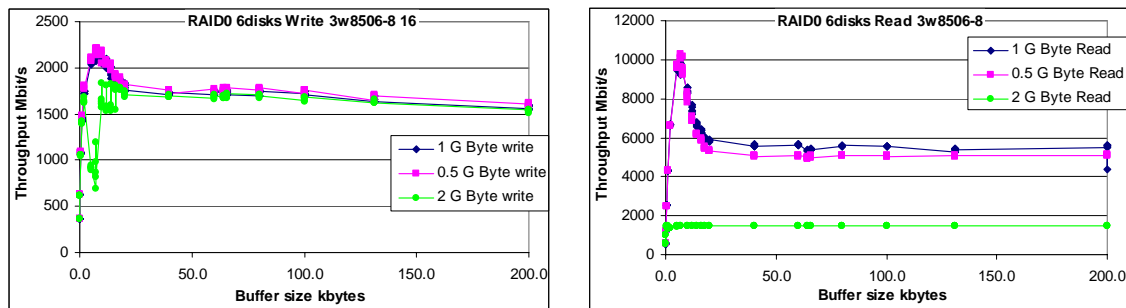


Figure 2: Memory to RAID0 subsystem transfer rates as a function of the read/write data buffer size for 0.5, 1.0 and 2 GByte files. The data was measured on a SC2004 Supermicro X5DPE-G2 server with 3Ware 8506-8 PCI-X controller with 6 disks.

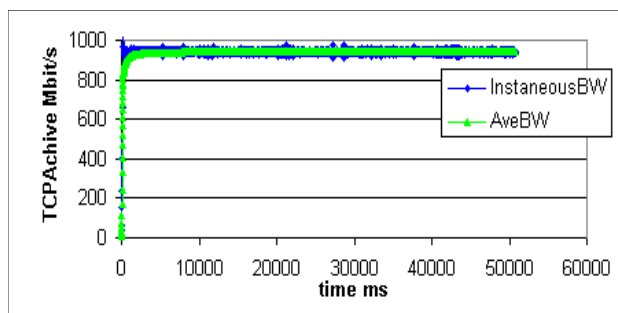
The 3Ware serial-ATA 7500 series controller was used for the BaBar transfer tests reported in this paper and could read at 1200 Mbit/s and write at 500-550 Mbit/s for large files. There is a general trend that files less than 400 Mbytes can be transferred over three times faster than larger files. Note that

in general the applications are optimized for transferring large files. The SC2004 RAID servers were built around Supermicro X5DPE-G2 motherboards with dual 2.8 GHz Zeon CPUs with 512 k byte cache and 1 M byte memory. The 3Ware 8506-8 controller was configured as a RAID0 device

with six 74.3 GByte Western Digital Raptor WD740 SATA disks and used a 64k byte stripe size. The measurements reported here were made using Scientific Linux 303 upgraded to use the 2.6.6 kernel[12], patched to allow choice of TCP stack algorithm[13]. The RAID controller and Gigabit Ethernet interfaces were on different 133 MHz PCI-X buses. Figure 2 shows the memory-RAID0 read and write transfer rates as a function of the read/write buffer size for various file sizes. These values were obtained with the 2.6.x kernel read-ahead cache settings increased by using:

```
/sbin/blockdev --setra 16384 /dev/sda
```

For writing, the plots show the throughput is independent of file size falling smoothly from 1.75Gbit/s for a buffer size of 20 k byte to 1.54Gbit/s when the buffer is 200 k bytes. With this range of buffer sizes, the read transfer rate for 2 G byte files is constant at ~1.85Gbit/s, however rates in excess of 5 Gbit/s are seen. This suggests that significant portions of the 1 Gbyte memory may be used as a cache. For both reading and writing, there is a significant peak in the transfer rates for 8 k byte buffers. Tests using Bonnie++ [14] to transfer 20 Gbytes of data gave write transfer rates of ~1.47Gbit/s and read transfer rates of ~1.45 Gbit/s, which is consistent with the



transfer rates shown for large files in Figure 2.

III. DATA TRANSFER MEASUREMENTS ON MB-NG AND SUPERJANET

Detailed measurements made on the MB-NG and SuperJANET networks are discussed in this section.

A. Memory to Memory Tests

After characterising the end to end link using UDPmon, Iperf was used to investigate the TCP memory-to-memory behaviour of the end hosts when connected to the different networks. Web100 [9] was used to instrument the TCP stacks. In Figure 3 the left 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 plots on the right show the BaBar hosts connected via SuperJANET. The saw-tooth behaviour of the achievable TCP throughput is typical when there is packet loss. This is confirmed by the number of duplicate ACKs recorded, as shown in the lower plot.

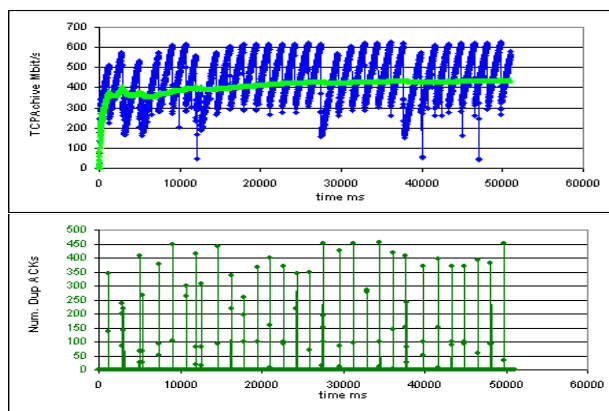


Figure 3: Instantaneous and average memory to memory throughput as a function of time. The plot on the left shows an MB-NG host on the MB-NG Network. The upper right plot shows BaBar host on the SuperJANET production network. The lower right plot duplicate ACKs for the BaBar host.

B. Disk to Disk Transfers

Figure 4 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 1. 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.

The plot on the bottom right shows the comparison of the behaviour of the TCP Congestion Window (red) with the TCP throughput (blue). In this graph the variation in the throughput occurs much more often than the decreases in the TCP congestion window, suggesting that the highly variable throughput is not due to the action of the TCP protocol or network but is more influenced by the performance of the disk sub-system, the memory bus and I/O performance, or the application design.

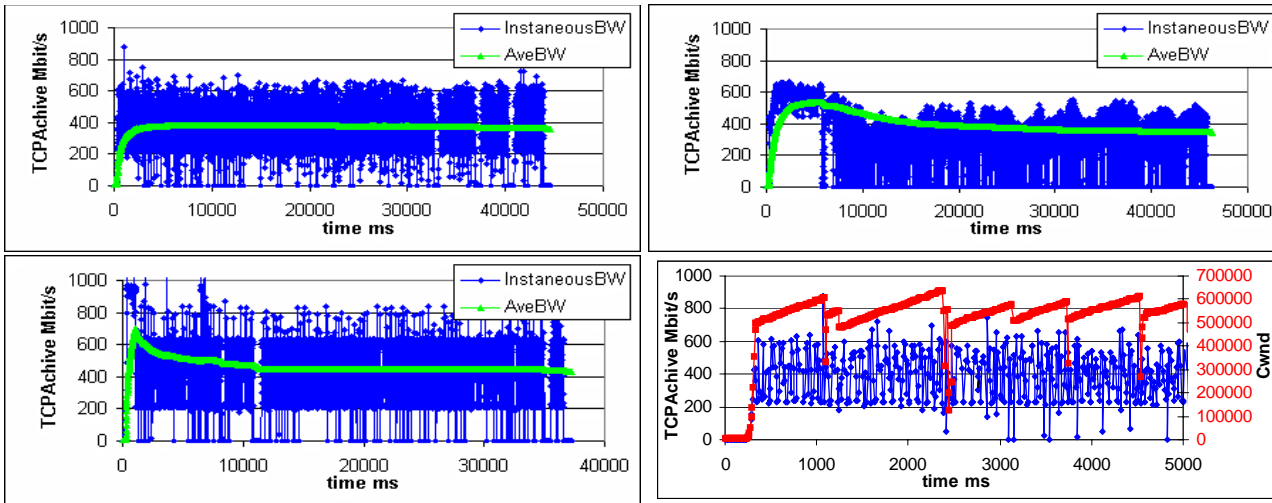


Figure 4: Comparison of the disk-to-disk throughput. bbftp was used with Highspeed TCP:
Top left: BaBar host on SuperJANET. Top right: MB-NG host on SuperJANET.
Bottom left: MB-NG host on MB-NG Network.
Bottom right: TCP Throughput (blue points) and the TCP Congestion Window (red line) as a function of time.

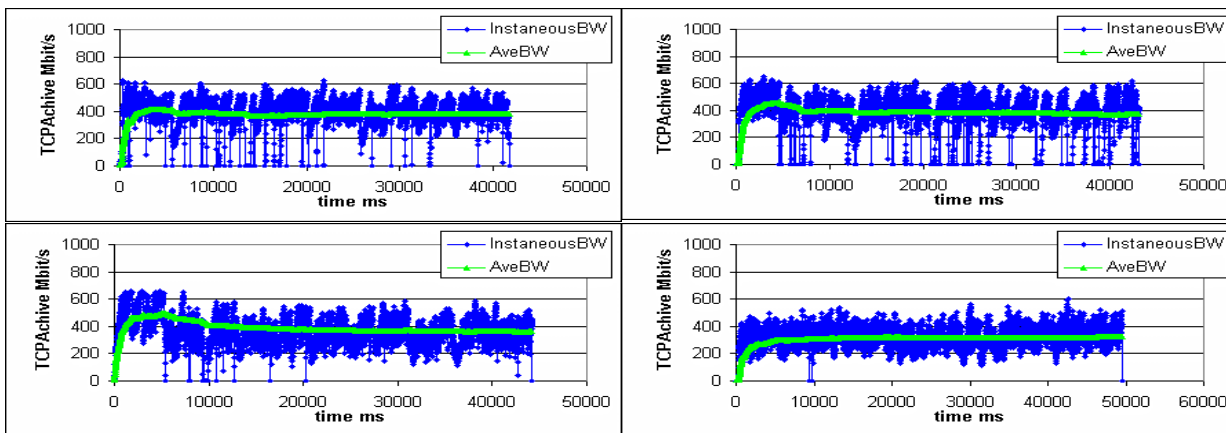


Figure 5: Comparison of the disk-to-disk throughput for bcbp, bbftp, apache, GridFTP. Highspeed TCP was used on the MB-NG systems connected using SuperJANET.

Figure 5 shows comparison of the performance of different data moving applications. These tests were made using Highspeed TCP to move a 2 Gbyte file between 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.

Table 1 shows the throughputs in Mbit/s achieved when moving 2 Gbyte 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. However transfers using the advanced TCP stacks do show an increase in throughput over standard

TCP.

Figure 6 shows the various transfers of real BaBar data that took place under different network configurations. All of the transfers used bbftp as the transfer tool and were performed using a single stream of Scalable TCP. The blue plot shows a transfer between BaBar servers over the production network and took approximately 10.5 hours. The green plot shows a transfer between Supermicro servers over the production network, this took about 8 hours. The red plot shows a transfer between Supermicro servers over the MB-NG network. This was the fastest transfer of this data set and took approximately 7 hours.

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	

Table 1: The throughputs in Mbit/s achieved for the different data transfer applications

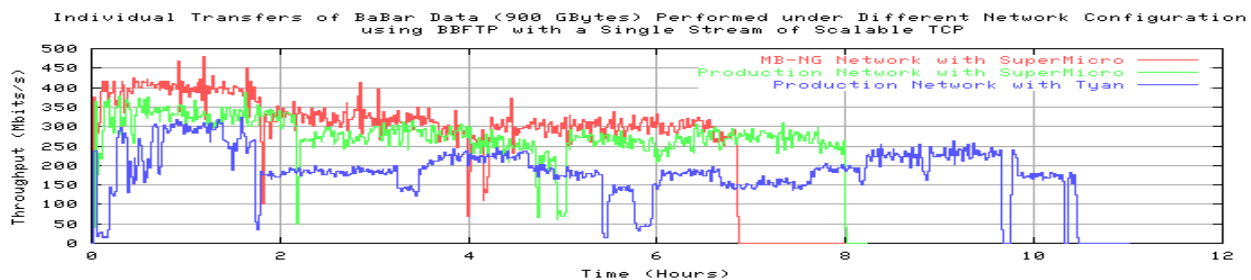


Figure 6: Transfers of BaBar data with under various network configurations using bbftp with a single stream of Scalable TCP.

The red plot (upper) shows a transfer across the MB-NG network using Supermicro servers.

The green plot (middle) shows a transfer across the SuperJANET network using Supermicro servers.

The blue plot (lower) shows a transfer across the SuperJANET network using BaBar servers.

IV. DATA TRANSFER INVESTIGATIONS AT SC2004 AND ON UKLIGHT

During SC2004, the testbed consisted of 6 Boston Supermicro RAID0 servers in the UK, 4 in London and 2 in Manchester, and 6 at the SLAC/FERMI booth in Pittsburgh. UKLight was used to connect the systems using six independent 1 Gigabit Ethernet channels. Four were sent from London to Starlight and two used UKLight to go from London to Amsterdam and then linked to two 1 Gigabit Ethernet channels from Netherlight to Starlight on the Eurolink connection. In Starlight these channels were connected to the Caltech PoP and used a 10 Gbit NLR Lambda to reach Pittsburgh. After SC2004, the UKLight SDH multiplexing equipment in Starlight was used to loop the 1 Gigabit Ethernet channels back to the UK forming a London-Chicago-London path with the rtt being 177 ms.

A. Network and Disk Sub-systems

UK-US or London-Chicago-London memory-memory TCP transfers using iperf and 1500 or 9000 byte MTUs achieved over 910Mbit, averaged over the entire 300s transfer, and 5s averages of over 950Mbit/s provided the kernel parameters allowed the socket buffer size to be set to 22 Mbytes and a TCP window scale factor of 9 was advertised. However, disk-disk transfers using bbftp were somewhat variable and typically between 450-500 Mbit/s, rather less than expected from the measured performance of individual disk and network sub-systems discussed in previous sections.

To investigate this, the memory-disk transfer rate was measured under different conditions. When run by itself on the SC2004 servers, the average RAID0 write rate for a 1 G byte file was 1735 Mbit/s, as shown in the plot of throughput vs test number at the top of Figure 7. The bottom plot shows that the RAID0 write throughput dropped by 30% to 1218

Mbit/s when a 1 Gbit/s stream of 1500 byte UDP packets was transmitted at the same time using the udpmon test program. The step up in throughput at trial number 86 corresponds to when the UDP traffic ended. When the MTU was set to 9000 bytes, the average RAID0 write rate was 1400 Mbit/s.

The scatter plots on the right hand side of Figure 7 show the % of time the two hyper-threaded CPUs (L1+L2) on Zeon chip 1 were in kernel mode vs that of chip 2 (L3+L4). Note the maximum is 200% and these percentages are averages of the length of each test – typically 6s. The negative correlation which was fitted with $y = -1.0479x + 174.44$ indicates that Linux either uses 1 chip or the other. The intercept of ~175%

(2 CPUs) shows that only 12.5% of each core is left unused on average, and the slope suggests that CPUs 3&4 take ~5% more cycles to perform the same action as CPUs 1&2. When both tests were running, one CPU chip tends to take ~80% and the other ~100% (only trials up to number 86 were included in this plot).

With this version of the 3Ware driver, all the interrupts for the RAID controller were delivered to one hyper-threaded CPU, whereas interrupts for the network were shared equally over the 4 CPUs. This data suggests that operating system scheduling and the load on each CPU core are critical.

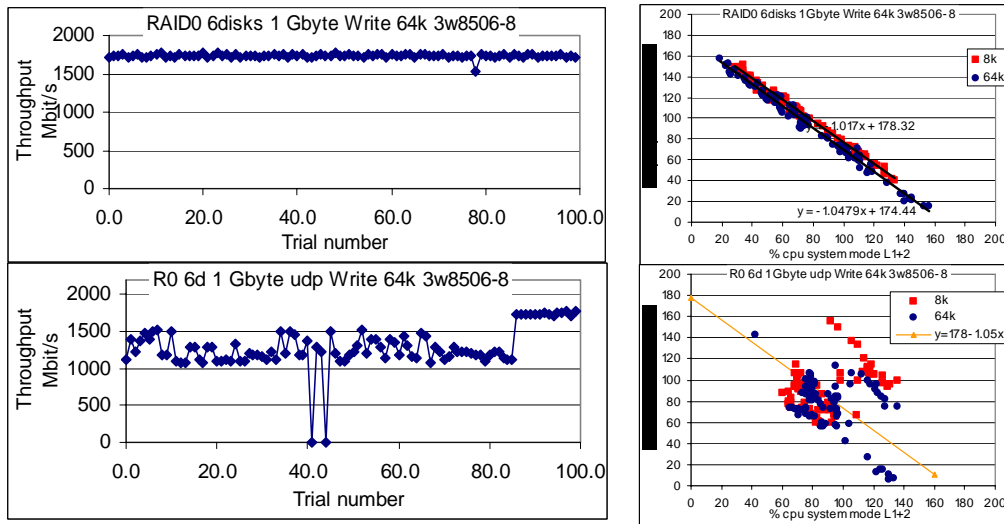


Figure 7: RAID0 memory-to-disk transfer rates for a 1 G byte file using 64 k byte RAID stripe size. The right hand plots show the correlation of the %CPU usage in kernel mode between the two Zeon CPU chips i.e. L3+L4 vs L1+L2.

Top: Only the Memory-Disk test running.

Bottom: Memory-Disk and UDP memory-memory tests running.

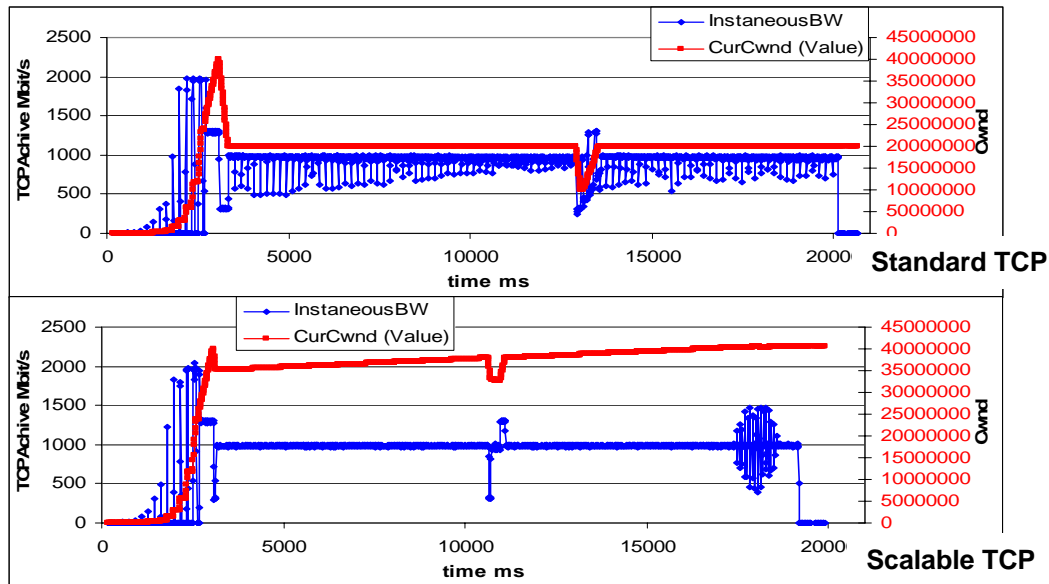


Figure 8: Comparison of the instantaneous throughput and current congestion window for disk-to-disk transfers using bftf with the Standard and Scalable TCP stacks. The SC2004 Boston Servers with RAID0 disk systems used the UKLight network London-Chicago-London.

B. *bbftp* Transfers on UKLight

Figure 8 compares the disk-to-disk throughput of *bbftp* when used with Standard and Scalable TCP to transfer a 2 Gbyte file between SC2004 hosts on the London-Chicago-London UKLight level2 Network. The TCP window size was set to 22 Mbytes, as the rtt was 177 ms., and SACK processing was turned off. The plots show the behaviour of the TCP Congestion Window (red) and the TCP throughput (blue). In both cases, at the end of slow start, after ~ 3 seconds, the transfer rate is at or close to line speed. This indicates that the end host is in fact able to move data from the RAID0 disks to the network at gigabit speeds and it is quite different from the behaviour of the MB-NG hosts using RAID5 disks shown in Figures 7,8, and 9.

It is interesting to note that the *bbftp* tool reported transfer rates of 670 Mbit/s for the Standard TCP transfer and 701 Mbit/s for Scalable whereas averages from the web100 information suggest data rates on the network of 825 and 875 Mbit/s respectively. When making comparisons it is clearly important to establish exactly what times are being measured.

V. CONCLUSION

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, and the interaction between the network hardware, protocol stack processing and the RAID disk sub-system is complex and requires further detailed study. 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 for the BaBar equipment are dominated by the performance of the particular RAID system tested. Further work is in progress to study the behaviour of later models of RAID controller and the use of advanced RAID configurations.

ACKNOWLEDGMENT

The work reported in this paper has been a result of collaborations between members of the BaBar experiment, the MB-NG e-Science project, UKLight, the Network Engineers and computing colleagues at Brunell, Manchester, RAL, ULCC and UCL, as well as colleagues from around the world at SC2004. We would like to thank all those involved in these collaborations. We would also like to thank Boston Ltd. for their tremendous and continuing support for Super Computing 2004 and the subsequent system tests.

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] Yee-Ting Li, Stephen Dallison, Richard Hughes-Jones, Peter Clarke "A Systematic Analysis of High Throughput TCP in Real Network Environments." The Second International Workshop on Protocols for Fast Long-Distance Networks (PFLDnet 2004) February 16-17, 2004 Argonne National Laboratory, Argonne, Illinois USA.
- [6] European DataTAG home page
<http://datatag.web.cern.ch/datatag>
- [7] 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
- [8] UDPmon: a Tool for Investigating Network Performance,
<http://www.hep.man.ac.uk/~rich/net>
- [9] Web100 Project home page,
<http://www.web100.org/>
- [10] Boston Ltd. Home Page: www.boston.co.uk
- [11] UKLight Home Page: www.uklight.ac.uk
- [12] Scientific_linux 303 came from <http://linuxsoft.cern.ch/> and was upgraded to 2.6.6 kernel using <http://linux.web.cern.ch/linux/updates/>.
- [13] The 2.6.6 kernel was from <http://www.kernel.org/> patched with Yee Ting Li's patch altAIMD_2-6.patch see www.hep.man.ac.uk/~rich/SC2004/patches
- [14] Bonnie++ Home Page: <http://www.coker.com.au/bonnie++/>



Richard Hughes-Jones leads the e-science and Trigger and Data Acquisition development in the Particle Physics group at Manchester University. He has a PhD in Particle Physics and has worked on Data Acquisition and Network projects. He is a member of the Trigger/DAQ group of the ATLAS experiment in the LHC programme, focusing on Gigabit Ethernet and protocol performance. He is also responsible for the High performance, High Throughput network investigations in the European Union DataGrid and DataTAG projects, the UK e-Science MB-NG project, and the UK GridPP project. He is secretary of the Particle Physics Network Coordinating Group which supports networking for UK PPARC researchers. He is a co-chair of the Network Measurements Working Group of the GGF, a co-chair of PFLDnet 2005, and a member of the UKLight Technical Committee.



Stephen Dallison has a PhD in Particle Physics and took up a postdoctoral post in the Network Team of the HEP group at the University of Manchester in May 2002. He principally works on the MB-NG project but also contributes to the EU DataTAG project. As part of the High Throughput working group Stephen has worked on network performance tests between sites in the UK, Europe and the United States, involving advanced networking hardware and novel transport layer alternatives to TCP/IP.