THE AUSTRALIAN NATIONAL UNIVERSITY

TR-CS-98-03

ACSys/RDN Experiences with Telstra's Experimental Broadband Network, First Progress Report

M. D. Wilson, S. R. Taylor, M. Rezny,
M. Buchhorn and A. L. Wendelborn

April 1998

Joint Computer Science Technical Report Series

Department of Computer Science
Faculty of Engineering and Information Technology

Computer Sciences Laboratory
Research School of Information Sciences and Engineering

This technical report series is published jointly by the Department of Computer Science, Faculty of Engineering and Information Technology, and the Computer Sciences Laboratory, Research School of Information Sciences and Engineering, The Australian National University.

Please direct correspondence regarding this series to:

Technical Reports
Department of Computer Science
Faculty of Engineering and Information Technology
The Australian National University
Canberra ACT 0200
Australia

or send email to:

`Technical.Reports@cs.anu.edu.au`

A list of technical reports, including some abstracts and copies of some full reports may be found at:

`http://cs.anu.edu.au/techreports/`

Recent reports in this series:

TR-CS-98-02 M. Manzur Murshed and Richard P. Brent. *Adaptive AT^2 optimal algorithms on reconfigurable meshes.* March 1998.

TR-CS-98-01 Scott Milton. *Thread migration in distributed memory multicomputers.* February 1998.

TR-CS-97-21 Ole Møller Nielsen and Markus Hegland. *A scalable parallel 2D wavelet transform algorithm.* December 1997.

TR-CS-97-20 M. Hegland, S. Roberts, and I. Altas. *Finite element thin plate splines for surface fitting.* November 1997.

TR-CS-97-19 Xun Qu, Jeffrey Xu Yu, and Richard P. Brent. *Implementation of a portable-IP system for mobile TCP/IP.* November 1997.

TR-CS-97-18 Richard P. Brent. *Stability of fast algorithms for structured linear systems.* September 1997.

ANU Technical Report TR-CS-98-03
DHPC Technical Report DHPC-022

ACSys/RDN Experiences with Telstra's Experimental Broadband Network

First Progress Report

M. D. Wilson, S. R. Taylor, M. Rezny, M. Buchhorn, A. L. Wendelborn

Distributed High Performance Computing Project
Research Data Networks Cooperative Research Centre,
Cooperative Research Centre for Advanced Computational Systems

Originally Released February 1997.

1. Abstract

This report summarises our experiences with the EBN and provides an indication of where we are now. We don't present a set of detailed performance measurements in this report, instead we focus primarily on bandwidth utilisation and network management. We are currently producing a more comprehensive set of performance measurements, which will be presented in a subsequent report.

2. Introduction

This report and its successor[16] describe work which was undertaken in 1997 and the second half of 1996. Initially intended as progress reports to Telstra, the provider of our experimental ATM [11,12,13,14,15] testbed, they are now being published as technical reports because they describe some of the important technical hurdles we overcame, the lessons we learned, and provide a convenient reference for discoveries which formed a basis for much other work which has been undertaken by the DHPC[7, 8].

The DHPC currently has two sites connected to Telstra's Experimental Broadband Network (EBN [1]). The Adelaide site is located in the Computer Science Department at the University of Adelaide, and the Canberra site is located at the Cooperative Research Centre for Advanced Computational Systems (ACSys), at the Australian National University. The Adelaide site has been operational since May 1996, and has taken part in numerous experiments with other EBN sites. Due to technical difficulties the ANU site did not become operational until the 11th of October 1996. Since then the two DHPC sites have made extensive use of the EBN.

Now that both of our EBN connections are functioning, we have integrated the network connection into our daily working environment. Our primary interest is data traffic running on top of the Internet Protocol suite (TCP/IP) [2, 3]. There are two common standards which implement IP over ATM: Classical IP [4, 5], and LAN Emulation [6]. We are currently using Classical IP, and all of our work to date has been concerned with optimising the performance of Classical IP over the EBN and on our local ATM LANs. In the future we may consider LAN Emulation, as it provides a more flexible network platform.

All of the DHPC's network traffic between the Adelaide and Canberra sites is routed across the EBN. The large data transfers and distributed computations, which are an integral part of the DHPC project, benefit enormously from the speed of the EBN connection between Canberra and Adelaide. We have also cross-mounted Network File Systems, for convenient file-sharing and performance testing. After some positive early experiments, we have recently purchased equipment to allow us to videoconference between the two sites, which we plan to use as a regular means of communication.

3. Site Descriptions

At the Adelaide site, 4 Digital Alpha workstations are connected via 155 Megabits/second multimode fibre to a FORE Systems ASX-1000 ATM switch. Two supercomputers, an SGI Power Challenge and a TMC CM-5

are also connected to the ASX-1000. Both of these machines belong to the South Australian Center for Parallel Computing (SACPC). The SACPC is making the supercomputers, as well as a powerful SGI graphics workstation and a high-capacity tape storage unit, available to the DHPC project. The FORE switch in Adelaide is configured with an E3 interface module, through which it is connected to the EBN.

In Canberra, 8 Digital Alpha workstations, and a Sun multi-processor 690 are connected to a Digital Equipment Corporation GIGASwitch/ATM. Because an E3 interface was not initially available for the GIGASwitch/ATM, a DECNIS router (Digital Equipment Corporation Network Integration Server) was supplied by Digital as an interim solution, equipped with an E3 interface and an OC-3c Multimode fibre interface. The DECNIS router was connected between the Digital GIGASwitch and Telstra's network termination equipment to provide a connection to the EBN. A new line card for the GIGASwitch is now available which supports an E3 interface, and the DECNIS router has been replaced with another GIGASwitch.

4. EBN Connection Experiments

4.1 Loop-back Testing

When the Adelaide site was first connected to the EBN in May 1996, we were unaware of any other EBN sites ready to participate in tests, so we arranged a loop-back test with Telstra Research Labs (TRL). We were provided with a pair of Virtual Paths, which were switched through the EBN network and back, so that we could establish our own private connections across the EBN. We used this loop-back connection to test configurations of our own equipment - setting up host interfaces and PVCs through our local switch. No throughput tests were done as connectivity, rather than performance, was our interest at the time. This exercise was extremely valuable, as we were able to configure ATM connections over the EBN in an environment which let us observe and control both end points of every connection. This test also enabled us to quickly determine the correct settings for the Adelaide switch's E3 interface.

During local tests in Adelaide, we found no compatibility problems between the DEC ATM Network Interface Cards (NICs) installed in the Alphas, the FORE ASX-1000 ATM switch or the FORE NICs installed in the SGI Power Challenge (and also several Sun and SGI workstations around the department). Problems were identified in both the FORE and Digital ATM drivers, and these problems were reported to the vendors or their representatives. Subsequent versions of the drivers from both vendors have addressed the problems we identified.

Working with Daniel Kirkham from TRL, we established that we could achieve high data transfer rates without seeing errors on either our equipment or TRL's. We also had some opportunity to observe the signalling behaviour of the ATM IP drivers installed on the Alphas. In particular we observed what we believed to be Classical-IP InATMARP address requests [4] being transmitted as the Adelaide host tried to establish a connection to a machine at TRL. This confirmed our suspicion that addresses were being exchanged automatically between hosts during PVC setup, although this was not clear from the vendor documentation available to us at the time.

4.2 ANSPAG Connection

In June 1996, we established our first full EBN connection with the Advanced Network Systems Performance and Applications Group (ANSPAG) site at Monash. At the time, they were trialing a FORE Systems ASX-200wg ATM switch. Because the ASX-200wg wasn't equipped with an E3 interface ANSPAG were using a Newbridge 36150 ATM switch as their interface to the EBN, and the ASX-200wg was connected to the Newbridge by 155 Mb/second fibre.

Initially, we found that when we established connections over PVCs between our two sites we experienced unacceptable data loss. In general we observed no data throughput at all! What this test identified was limitation of the switching equipment used to implement the EBN. Because of the design of the cell buffers used in the Alcatel switches, the EBN was extremely sensitive to variations in cell arrival rates from customer equipment. The EBN switches too readily discarded incoming data. Telstra have now upgraded the line

cards in the EBN switches to provide better cell buffering. In combination with the use of ABR flow control, this means that cell loss is now less of a problem for the DHPC (see later for performance measurements).

The throughput performance test we used at this time was to use the FTP protocol to measure file transfer speeds. Over a PVC between Adelaide and Monash an FTP client run at Adelaide appeared to transmit approximately 30 Kilobytes before any large transfer to Monash hung. This corresponds quite closely with the default TCP acknowledgment window size on our Alpha workstations. We suspect that the FTP client transmitted one 30K TCP window of data, which was never received by the server at Monash, and then hung waiting for TCP acknowledgment packets.

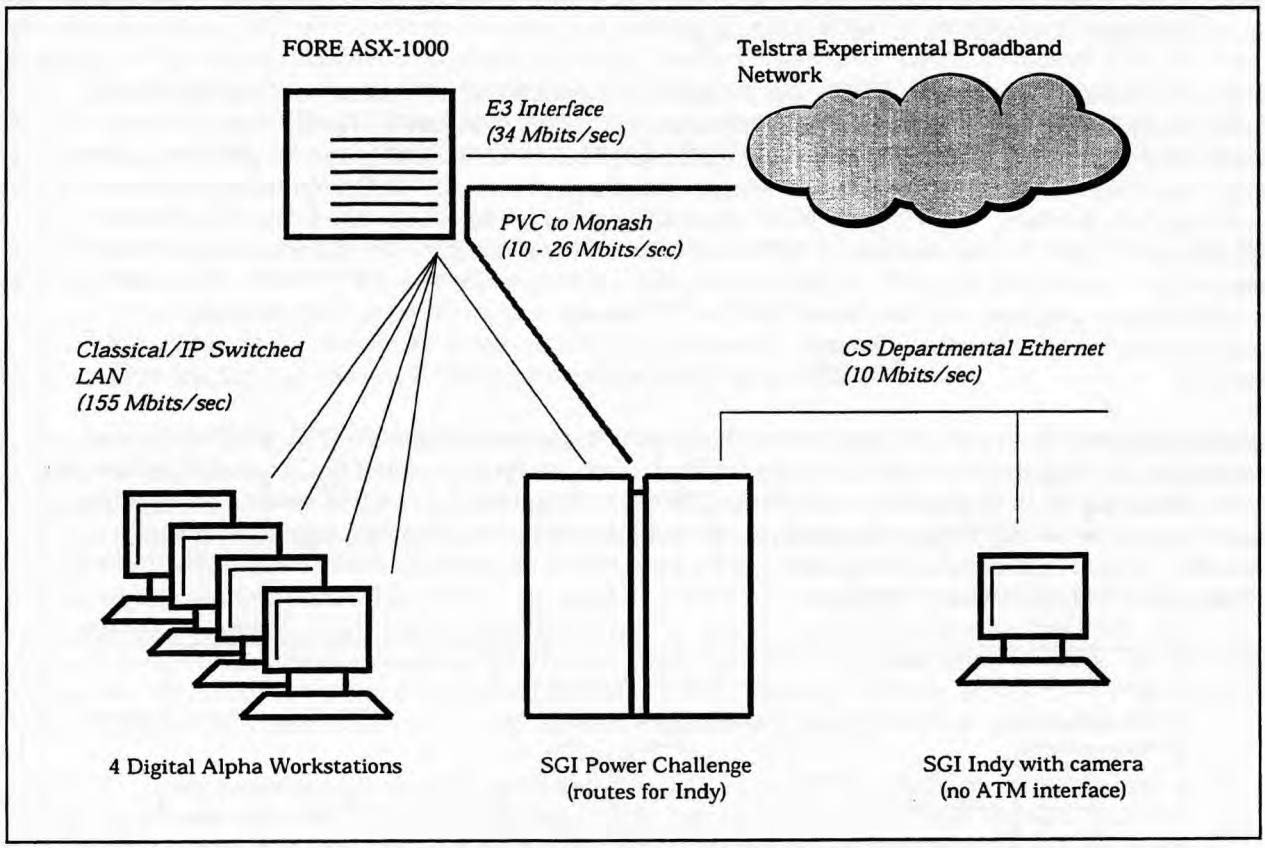

Figure 1: Topology of the Adelaide network for the DHPC-ANSPAG connection

We discovered that a partial solution to the data loss problem was to explicitly throttle the output from our workstations to just a fraction of our allocated EBN data rate. However, this was neither a scalable nor reliable solution. Because of limitations in the ATM drivers supplied with Digital Unix 3.2c we could not throttle the Alpha workstations down below the 155Mb/second rate. However, the FORE VME-200 ATM card in the SGI Power Challenge did support rate control. Consequently we implemented rate control by routing traffic through the Power Challenge. The experimental setup is described in Figure 1 below.

The only performance measurement we made of this configuration was an FTP session which achieved around 8 Megabits/second using a PVC through a Virtual Path (VP) which had an allocated data rate of 27 Megabits/second. This represented bandwidth utilisation less than 30%.

Using this configuration, with a bandwidth-throttled ATM interface, we attempted some simple videoconferencing tests using the public domain MBONE tools. The Alpha workstations, at that time, were not equipped with cameras, however we found that voice transmission performed well. In order to try video transmission, we borrowed a Silicon Graphics Indy workstation from the SACPC. This machine was equipped with a camera, but not an ATM card, so we routed its MBONE connection through the Power Challenge to the EBN.

Unfortunately the videoconferencing results were disappointing. Videoconferencing sessions were unreliable and the frame rates were always extremely poor. However, the overall results were still

encouraging, and they demonstrated that we could establish useful IP connections over the EBN. They also highlighted that cell loss was an issue that we needed to address.

4.3 DSTC Connection

In June 1996, when the Cooperative Research Centre for Distributed Systems Technology (DSTC) site at the University of Queensland, was connected to the EBN, we contacted Bob Brown to arrange a test connection. DSTC's configuration was, in a sense, similar to ANSPAG's, with their IBM 8260 ATM switch connected indirectly to the EBN via a ADC Kentrox rate converter with an E3 module. This arrangement was necessary because IBM could not supply an E3 card for the 8260.

We set up a connection between Adelaide and Brisbane, and established a PVC between a single host in Adelaide and a single host in Brisbane. This configuration is shown in Figure 2. The EBN virtual path between the two sites was allocated a bandwidth of 3.6 Megabits/second. Using an FTP client we achieved a transfer rate of approximately 370 Kilobytes/second (3.35 Megabits/sec or ~93% bandwidth utilisation) from Adelaide to Brisbane, and a slightly lower data rate in the other direction. We found that to achieve the maximum throughput, it was necessary to allow a small bandwidth margin. The Adelaide switch was configured to throttle outgoing traffic on the virtual path to 3.3 Megabits/second. We later repeated these tests with the ttcp program, and discovered that the FTP results were suspicious. Ttcp indicated a throughput much lower than the FTP result - around 270 Kilobytes/sec (2.44 Megabits/second, or ~68% utilisation).

This led us to experiment with the Early Packet Discard (EPD) feature of the ASX-1000. EPD anticipates downstream link congestion and discards AAL5 ATM frames (which correspond to Classical IP packets) to improve network performance on congested links. By discarding entire frames EPD saves link capacity, because naively removing single cells could render multiple frames useless while they still consume bandwidth. However, using the configuration in figure 2, we found that EPD didn't measurably improve our throughput from Adelaide to Brisbane.

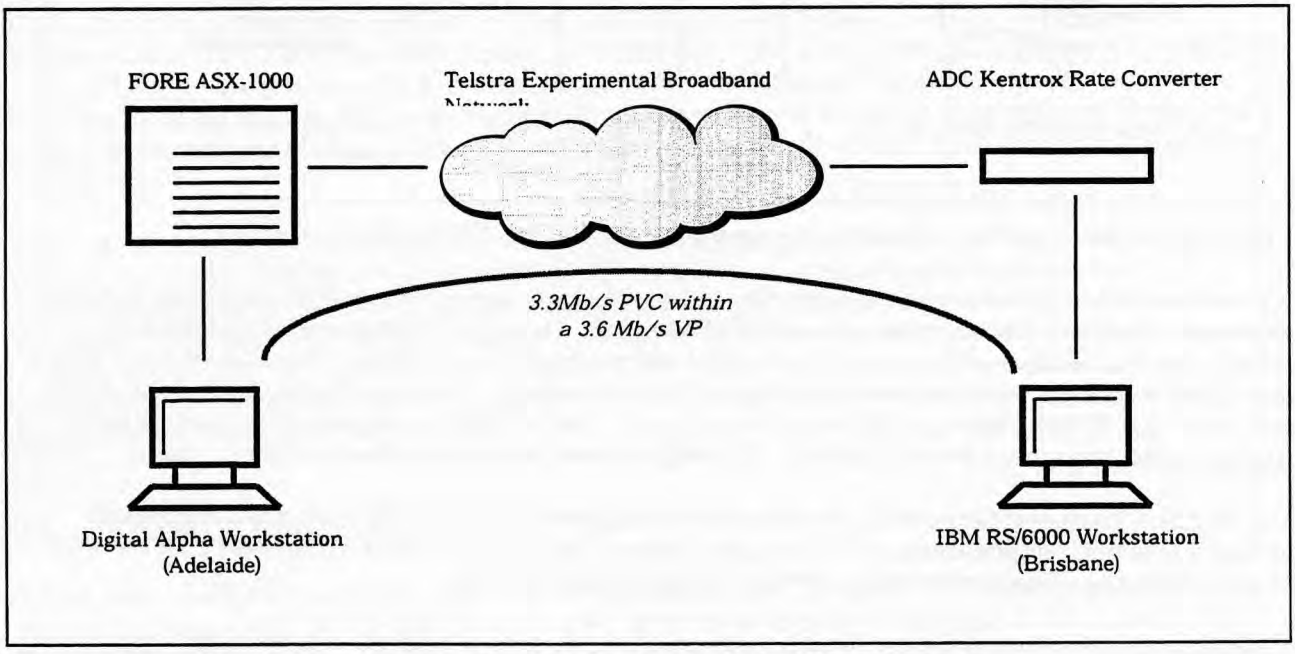

Figure 2: Configuration of the PVC used for the DHPC-DSTC connection

We found the results of the DSTC tests extremely interesting. Because the RS/6000 in Brisbane could throttle its output data rates we did not expect data loss to occur for transfers from Brisbane to Adelaide. However, because we could not throttle the output of the Alpha in Adelaide (because of the limitations already mentioned in the Digital Unix 3.2c ATM drivers) we expected the ASX-1000 to discard almost all of the outgoing cells from Adelaide. Instead, we found that we were able to achieve acceptable performance. It appeared that we were seeing the effects of some form of flow control. From documentation that we had available[9, 17], and correspondence with representatives from our equipment vendors, we believed that our

equipment was possibly using EFCI (Explicit Forward Congestion Indicator) flow control (which is one of the standards advocated by the ATM Forum [11]) which may have explained the results we observed. Another possibility was that flow control and congestion avoidance was being carried out by the TCP/IP implementations of our hosts [10].

5. The major Canberra-Adelaide connection

The EBN connection at the DHPC's Canberra site took some time to become operational. There were electrical compatibility problems between the E3 Module installed in the Digital ATM equipment and the E3 interface built into Telstra's Customer Premises Equipment. There was no obvious solution, as both companies maintained that their equipment conformed to the E3 standard. Early In October 1996, Dr Mike Rezny identified that there was a problem at the electrical level, and developed a temporary solution, using an Ethernet terminator. This problem, and the solution, were reported to Digital, who determined that there was a manufacturing error in their E3 modules. The E3 interfaces have since been replaced by Digital and are now working perfectly.

To date, two different configurations have been used for the Adelaide-Canberra EBN connection. These two configurations are described separately.

5.1 The initial Canberra EBN connection using a DECNIS 600

Our first configuration, as shown in Figure 3, necessarily made use of the DECNIS router. We arranged our EBN hosts in a star topology, where each host had its own PVC to the DECNIS. The DECNIS routed packets between the 6 PVCs to the hosts in Adelaide,and the 8 PVCs to the Canberra hosts. The PVCs to Adelaide were carried within a single EBN Virtual Path, which was allocated a bandwidth of 26 Megabits/second.

This configuration was planned before the electrical problems with the E3 interface on the DECNIS router had been solved, and we had some concerns about how it would work in practice. In particular, since we had no way to control the transmission rate of our Alphas, we were concerned about cell loss. Furthermore, as this was the first time we had tried switching multiple PVCs through a single EBN Virtual Path, we were interested to see how hosts competing for bandwidth might exacerbate the cell loss problem.

Our concern about cell loss stemmed from the fact the DECNIS supports only fairly primitive bandwidth allocation options. We anticipated two problems. First, we were forced to choose between statically allocating a fraction of the available VP bandwidth to each EBN PVC, or allocating the full available bandwidth to each of the EBN PVCs and hoping that the DECNIS would manage the congestion. Second, the DECNIS wouldn't allow us to set the maximum outgoing bandwidth for the E3 interface (implicitly assuming that the full 34 Megabits/second would be available), although we only had a 27 Megabits/second VP.

However, when the ANU EBN connection became operational in October 1996, the DECNIS based EBN configuration performed exceptionally well. We were able to allocate the full E3 data rate to each of the PVCs across the EBN and a full 155 Megabits/second for each PVC between the GIGASwitch to the DECNIS. This configuration consistently delivered EBN link utilisation for individual data transfers in excess of 95%. Simultaneous transfers involving multiple PVCs across the EBN continued to achieve high total link throughput, indicating that bandwidth contention was handled well.

We believe that ABR flow control (most likely TCP/IP or ATM Forum EFCI) was operating effectively between the Adelaide workstations and the DECNIS, across the EBN.

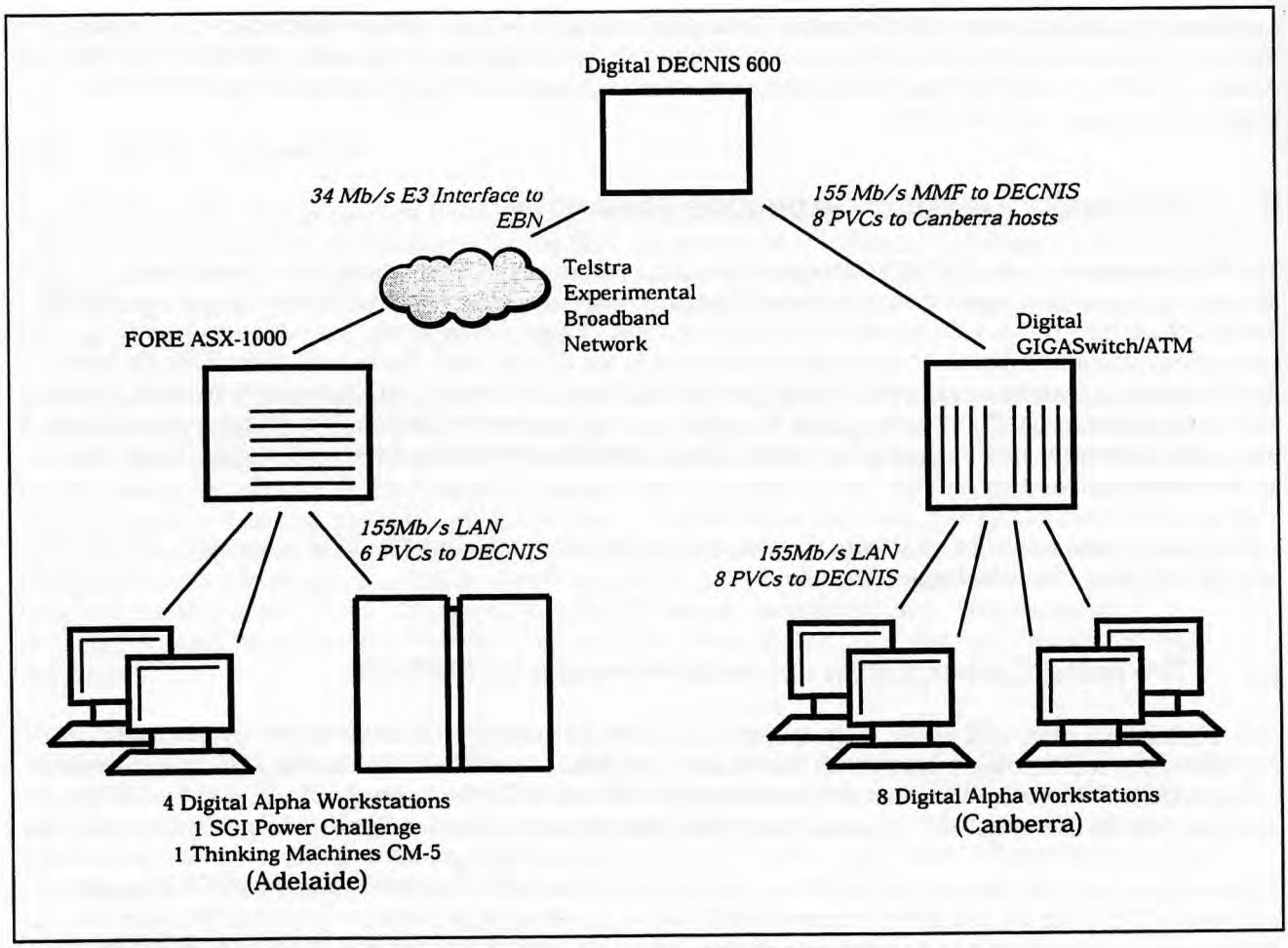

Figure 3: Topology of the DECNIS based Canberra-Adelaide connection

The DECNIS and the Canberra DHPC GIGASwitch also recognised one another as devices equipped to use FLOWMaster, Digital's proprietary flow control protocol. We aren't certain whether FLOWMaster or ATM Forum flow control was operating between the Canberra hosts and the DECNIS. Enabling and Disabling FLOWMaster support on the Alpha workstations did not apparently affect network performance. Work will continue in this area.

5.2 Performance of the DECNIS based EBN connection

We found that to achieve high performance in the EBN environment our Unix workstations needed several configuration changes. Most importantly, TCP acknowledgment windows must be increased, to allow for the long round trip times over the wide area, combined with the high throughput provided by the EBN. The round trip time between the two DHPC sites using the DECNIS configuration is around 17 ms, compared with typical LAN figures of less than 1 ms.

We also found it useful to use the Unix "route" command to statically configure a TCP Maximum Segment Size (MSS) for our EBN traffic. MSS is determined from the smallest Maximum Transfer Unit (MTU) limit of all intermediate links in a TCP connection, using the following [2]:

MSS = MTU(min) - (TCP + IP overhead)

Normal TCP behaviour is to assume that any TCP data bound for a router (in our case, the DECNIS) must be broken down into segments of length 536 bytes, to avoid fragmentation of TCP segments within the network. This is because a sending host can not make any assumptions about the Maximum Transfer Unit (MTU) size of intermediate network links. 536 bytes is considered a safe lower bound. Unfortunately, the extra overhead involved in processing these smaller segments means that we observe lower TCP performance. Because we know the MTU size of all of the links between the Canberra hosts and the Adelaide hosts, we can

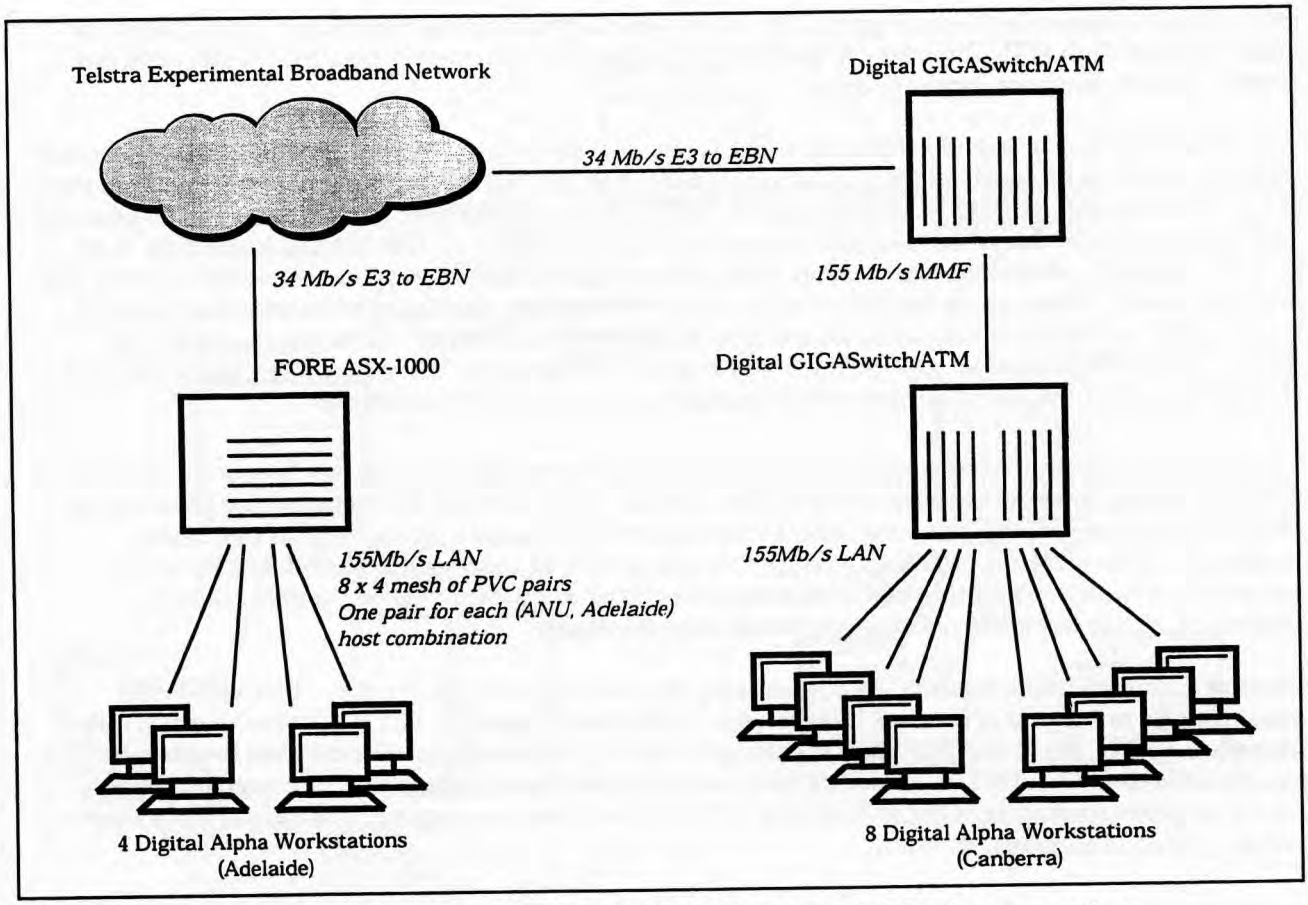

Figure 4: Topology of the GIGASwitch/ATM based Canberra-Adelaide connection

Unfortunately, we found that PVM-based applications were still affected by the new configuration, even after TCP windows were set to 64KB. We are investigating other ways to overcome the problem, including an upgrade to the GIGASwitch to increase the size of the cell buffers.

6. Future

We will continue to utilise the EBN for running distributed applications, especially distributed image processing. We are interested in both the performance of the applications themselves, and the performance of the EBN network in supporting these applications.

In the near future, the DHPC project will begin to utilise the EBN for videoconferencing as part of its routine activities, to facilitate communication between the Adelaide and Canberra project groups, as well as other RDN projects. We are very optimistic that this will be of considerable benefit to project activities.

We are also about to investigate a new system configuration, which will make use of SVCs between the two DHPC sites. Our ATM equipment is capable of managing its own SVC signalling through a designated Virtual Path, and we expect this will make our ATM resources more manageable.

Finally, we are investigating ways to incorporate the ANU Supercomputing Facility into the ATM network accessible by the DHPC, including the EBN. This will give both DHPC sites high-capacity access to resources such as the StorageTek tape storage silo. We are also interested in investigating ways to manage bandwidth between EBN sites more flexibly than is currently possible.

7. Acknowledgements

The authors wish to Acknowledge that this work was carried out within the Cooperative Research Center for Research Data Networks, established under the Australian Government's Cooperative Research Centers programme.

The Distributed High Performance Computing Project (DHPC) is a project of the Research Data Networks Cooperative Research Centre (RDN CRC), is managed by the Advanced Computation Systems CRC and is a joint activity of the Australian National University and the University of Adelaide. We thank Telstra for the provision of the Experimental Broadband Network, and we particularly would like to thank Daniel Kirkham at Telstra Research Laboratories for his invaluable help.

8. References

[1] D. Kirkham, "Telstra's Experimental Broadband Network", Telecommunications Journal of Australia, Vol 45, No 2, 1995.

[2] W. Richard Stevens, "TCP/IP Illustrated", Prentice-Hall 1994, ISBN 0-20-163346-9

[3] W. Richard Stevens, "Unix Network Programming", Prentice-Hall 1990, ISBN 0-13-949876-1

[4] M. Laubach, "Network Working Group Request for Comments: 1577. Classical IP and ARP over ATM", 1994.

[5] J Heinanen, "Network Working Group Request for Comments: 1483. Multiprotocol Encapsulation over ATM Adaptation Layer 5", 1993.

[6] "LAN Emulation Over ATM Version 1.0", Specification af-lane-0021.000, The ATM Forum Technical Committee, 1995.

[7] K.A.Hawick, H. A.James, K.J.Maciunas, F.A.Vaughan, A.L.Wendelborn, M.Buchhorn, M.Rezny, S.R.Taylor and M.D.Wilson., "An ATM-based Distributed High Performance Computing System." Technical Note DHPC-002 The RDN CRC Distributed High-Performance Computing Project.

[8] K.A.Hawick, H.A.James, K.J.Maciunas, F.A.Vaughan, A.L.Wendelborn, M.Buchhorn, M.Rezny, S.R.Taylor and M.D.Wilson, "Geographic Information Systems Applications on an ATM-Based Distributed High Performance Computing System.". Technical Note DHPC-003, The RDN CRC Distributed High-Performance Computing Project.

[9] "Frequently Asked Questions on ATM and Digital's ATM Program", Digital Equipment Corporation 1996. Available online via http://www.networks.digital.com/dr/techart/

[10] W. Stevens, "Network Working Group Request for Comments 2001: TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms", 1997.

[11] "ATM User-Network Interface Specification version 3.1", The ATM Forum Specification af-uni-0010.002, approved 1994.

[12] R. Handel, M. N. Huber, S. Schroder, "ATM Networks - Concepts, Protocols, Applications." Addison-Wesley 1994, 0-20-142274-3

[13] M. Batubara and A. J. McGregor, "An Introduction to B-ISDN and ATM", Technical Report 93/4 Faculty of Computing and Information Technology /Department of Robotics and Digital Technology.

[14] Paul Reilly, MSD Marketing Silicon Graphics Inc. "PDH, Broadband ISDN, ATM and All That: A Guide to Modern WAN Networking and How it Evolved", Silicon Graphics 1994. Available at: ftp://sgigate.sgi.com/pub/Surf/ATM.ps.Z

[15] Cisco hosts an excellent online reference for a range of networking and internetworking technologies, including ATM. http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/index.htm

[16] M. Wilson, K. Yap, "ACSys/RDN Experiences with Telstra's Experimental Broadband Network - Second Progress Report", Technical Note DHPC-023, Advanced Computational Systems CRC, 1997

[17] T. Des Jardins, S. S. Sathaye, "Traffic Management in FORE Systems' ATM Networks", (DRAFT) Doc#8.4.2, Fore Systems Inc, 1994.

[18] J. Mogul, S. Deering, "Network Working Group Request for Comments: 1191, Path MTU Discovery", 1990.