

Experiences with CRI ATM at SARA

Jan Overweel, SARA, Amsterdam, The Netherlands

Introduction

SARA participates in a pilot project to develop a new IP Wide Area Network network infrastructure based on ATM. ATM was chosen for implementing the lower levels of the IP infrastructure because of its scalability. Furthermore, ATM offers also a smooth upgrade path for the existing IP infrastructure and it provides a single network solution for the transmission of voice, video and data. However, ATM technology is relatively new compared to existing technologies and that is the reason why an extensive test programme was started. SARA's C98 takes part in the test programme: firstly, to test its interoperability with switches and other platforms, and secondly, to gather user experience with distributed IP based network applications in an ATM Wide Area Network environment. Hence, CRI's ATM functionality tests are primarily focused on applicability in an ATM WAN environment with a limited bandwidth and, to a lesser extent, to its applicability to provide high speed link connections in a Local Area Environment. SARA shares its experience with CRI to give direction to further development of CRI's Model-E ATM interface, the ATM Bus Based Gateway.

ATM Bus Based Gateway Features

The ATM Bus Based Gateway is an external interface connected with HIPPI to the IOS. It is an Sbus based box equipped with a Sparc processor, a HIPPI interface and an SBA-200 card of FORE Systems Inc. SARA's BBG has a 100 Mb/s TAXI physical link interface.

The BBG can only be used via the Unicos IF driver, that is, the ATM protocol layer can only be accessed from user space via the IP layer of Unicos. Direct access from user space to the ATM protocol stack is not supported yet. As a consequence, an Application Programming Interface (API) to develop native ATM applications isn't available yet. Some other vendors, e.g. FORE, provide a proprietary API. However, CRI prefers to wait till final standardization of an API has been completed.

The CRI ATM implementation supports the Classical IP Encapsulation method compliant with RFC 1577. Hence, IP packets are prefixed with an LLC/SNAP header and adapted to the network using ATM Adaptation Layer 5. AAL5 provides no error correction, just error detection. Consequently, the loss of just one cell results in the retransmission of an IP packet. MTU sizes ranging from 9176 to 65536 bytes are supported. The default MTU size for IP members operating in an ATM Logical IP subnet must be 9180 bytes according to RFC 1577. However, CRI doesn't support this because this size is not a multiple of a byte (8 bytes), so the default MTU size is 9176 bytes.

Only Permanent Virtual Connections (PVCs) are supported in the current implementation. This might be sufficient in LANs to create just a few static connections between just a few platforms but it is insufficient to meet the requirements of the general application of ATM in WAN environments. Firstly, PVCs waste available bandwidth on physical links because bandwidth is reserved in a "may be needed once" fashion. Secondly, the establishment of a PVC between two ATM platforms requires the static configuration of the PVC at the two platforms and all the switches in between by administrator. The dynamic establishment and release of connections (signalling) is a prerequisite in order to use ATM in a WAN production environment. Firstly, to optimize the utilization of the available bandwidth on physical links, and secondly, to simplify the WAN administration. Switched Virtual Connections are not supported yet but in future releases the CRI implementation will be compliant with the ATM Forum UNI 3.1 standard concerning SVCs.

Early Experience with the BBG

The Bus Based Gateway, the ATM interface for CRI Model-E based systems, was installed at SARA in November '94. It was one of the very first two installations in Europe performed by CRI's network development department. The very first experience with the BBG revealed an important flaw in the software implementation. The BBG was only capable to operate at "line rate" transmission speed; the transmission speed couldn't be controlled, that is, reduced to enable interoperating with other platforms in an ATM WAN via transmission paths with limited bandwidth (< 32 Mb/s) or with other end nodes with limited ATM throughput capacity. This disabled the use of CRI ATM in real life application projects. The capability to control the transmission speed called "rate shaping" was added in a software upgrade of the BBG software in May '95. The importance of rate shaping will be illustrated by some TCP/IP performance measurements in simple test setups.

TCP/IP-ATM Throughput Measurements

TCP/IP-ATM throughput measurements were performed with "ttcp". The average TCP data rate was measured as a function of the send and receive socket buffer sizes of the sender and receiver, respectively. Their sizes and the TCP maximum segment size (MSS) have significant impact on TCP's sliding window protocol and, consequently, on TCP/IP performance. All measurements were performed with a TCP MSS of 9136 bytes in a production environment in either a back-to-back

configuration or in a host-switch-host configuration. The switch is a GDC Apex DV2 switch.

Data transfer from the C98 to a SGI Onyx

Figure 1 shows the TCP/IP-ATM throughput of data transfers from the Cray C98 to a SGI Onyx system in a back-to-back configuration. The Onyx is equipped with a FORE VMA-200 card. The C98 and Onyx have comparable transmit and receive capabilities i.e. data transfers from C98 to Onyx and vice versa at maximum transmission speed don't cause any cell loss at the receiver side. The maximum measured throughput is approximately 66 Mbits/sec (the theoretical maximum for a 100 Mbits/sec TAXI physical connection is approximately 87 Mbits/sec due to header and trailer protocol overhead in layers of the ATM protocol stack). Maximum performance is reached when large send and receive socket buffers are used. Send and receive socket buffers should be at least five times the MSS to get maximum performance. Smaller values lead to substantial reduction of the performance. For instance, the throughput is only 35 Mbits/sec with socket buffer sizes three times the MSS, that is, with 27.4 kbyte buffers. Although this socket buffer size might be sufficient to get the maximum throughput in an environment with small MTUs like for instance Ethernet (~1.5 kbyte), it is far from sufficient for an ATM environment. Receive socket buffer sizes less than two MSSs lead to TCP's "delayed ack syndrome" and consequently to poor throughput performance.

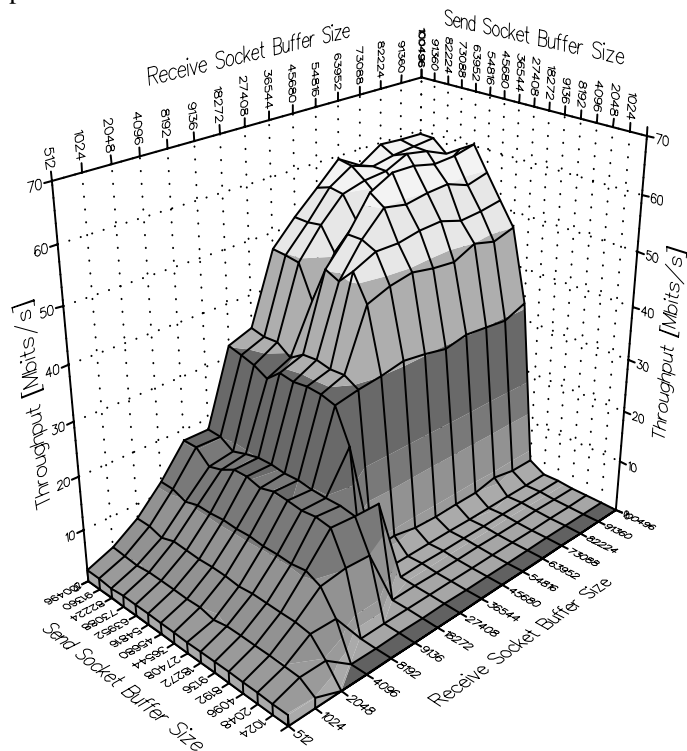


Figure 1: TCP/IP-ATM data transfer from the C98 to a SGI Onyx in a back-to-back configuration. Throughput is measured as a function of socket buffer sizes. The MSS is 9136 bytes.

Data transfer from the C98 to a SGI Indy

In cases where the receiver side is not able to keep pace with the sender's transmission speed, the sender ought to reduce its speed. As mentioned before, the very first software implementation of the BBG didn't support this feature. The result is shown in figure 2. In this case the C98 is coupled back-to-back to an SGI Indy system and data is transferred from the C98 to the Indy. The Indy is equipped with a first generation FORE GIA-100 card. Major protocol layer processing functions, like reassembly of cells into Protocol Data Units and error detection, are not implemented in the card's hardware but in system software. The Indy is simply not capable to process incoming traffic as soon as TCP gets going, that is, when the receive socket buffer of the Indy gets larger than two MSSs. Cells loss in the Indy's interface causes the throughput to collapse to almost zero. The congestion control mechanism in Unicos' TCP layer doesn't resolve the loss of TCP packets effectively. Hence, speed reduction has to be performed on the ATM layer instead.

Data transfer from the C98 to a SGI Onyx via the switch

The C98 and Onyx are logically connected by a PVC with a certain Quality of Service as defined in the switch. The sustained cell rate (SCR) is the PVC's QoS parameter on which traffic conformance enforcement is performed by the switch. Hence, the sender, the C98, has to keep its cell transmission rate below the SCR limit of the PVC, otherwise the switch will throw away incoming cells at entrance. Figure 3 shows the throughput of data transfers from the C98 to the Onyx via a PVC with a SCR of 150,000 cells per second. This corresponds with an effective sustained bit rate of approximately 55 Mbits/sec. The shape of the figure is similar to figure 1 except in the area with send and receive socket buffers equal or larger than five MSS. In this area, throughput collapses to almost zero because the C98 transmission rate exceeds the SCR of the connection and the switch starts to drop cells. Thus, the TCP throughput is in this case to a very high extent dependent on socket buffer sizes. To prevent this, the C98's transmission rate needs to be reduced or shaped according to the QoS provided by the PVC, i.e., it must remain below the SCR of the PVC.

Data transfer from the C98 to a SGI Onyx with C98 rate shaping

Rate shaping support in CRI's ATM implementation solved the problems experienced in the two previous examples. It enables the administrator to define a maximum data transmission rate per PVC in kilobits per second. Figure 4 shows the data transfer from the C98 to the Onyx in a back-to-back configuration with the C98's maximum transmission rate set to 25.4 Mbits/sec. The measured throughput values with large socket buffers match very well with the specified rate. Even when very low maximum transmission rates are defined for the PVC, like for instance 2 Mbits/sec, the defined rate matches very well with the measured rate.

Conclusion

SARA has now almost one year experience with the CRI ATM implementation for Model E systems. Hardware and software were reliable and stable during this period.

Lack of a basic feature like rate shaping disabled the use of CRI's ATM in a WAN environment with limited bandwidth up to May '95. This illustrates how ATM development progresses in general. Vendors implement the basics of ATM first and implement additional features afterwards. Standardization of ATM is still in progress and this leads to an ongoing process of hardware and software upgrades of interfaces and switches.

A major leap forward is made when Switched Virtual Connections can be established. Major vendors, like CRI, will implement SVCs compliant with the ATM Forum UNI 3.1 standard. Implementations have already been released or are expected to arrive in the very near future. The success of ATM in WAN environments will depend on the pace SVC support progresses.

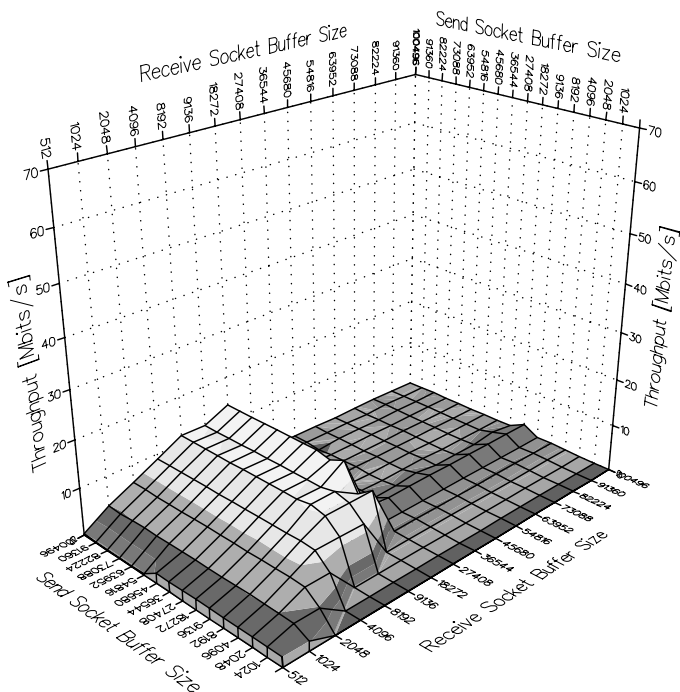


Figure 2: TCP/IP-ATM data transfer from the C98 to a SGI Indy in a back-to-back configuration. Throughput is measured as a function of socket buffer sizes. The MSS is 9136 bytes.

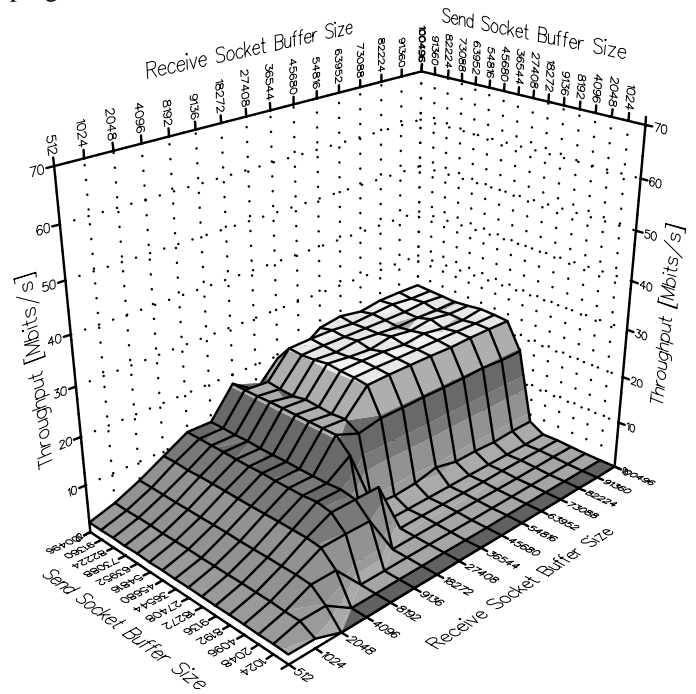


Figure 4: TCP/IP-ATM data transfer from the C98 to a SGI Onyx in a back-to-back configuration. The C98 performs rate shaping. The PVC's peak rate is 25.4 Mbits/sec.

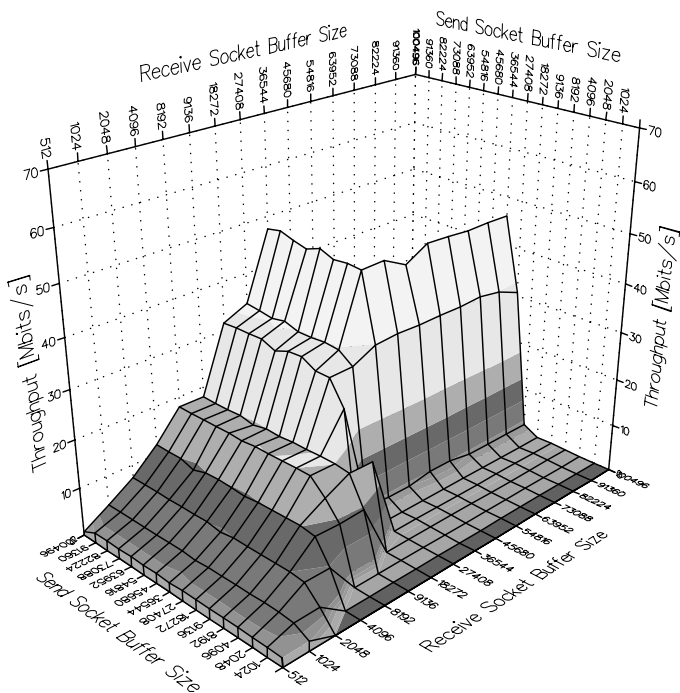


Figure 3: TCP/IP-ATM data transfer from the C98 to a SGI Onyx in a host-switch-host configuration. The QoS of the PVC is 150,000 cells per second sustained.

Acknowledgements

I would like to thank Jeff Young and Joe Golio of Cray Research Inc for the BBG installation, and Hans Nelemans and Gerben Jansen of Cray Research B.V. for their support. I would like to thank Andre Walet in particular because he did all the measurements. Andre, Rick te Lindert and Marcel van Zalen of SARA were most helpful in preparing this paper.

References

- [1] D.E. Mc Dyson, D.L. Spohn, *ATM: Theory and Application*, Prentice Hall, 1994
- [2] ATM Forum, "ATM User-Network Interface Specification Version 3.0", August 1993
- [3] J. Heinanen, "IETF RFC 1483, Multiprotocol Encapsulation over ATM Adaptation Layer 5", July 1993
- [4] M. Laubach, "IETF RFC 1577, Classical IP and ARP over ATM", January 1994
- [5] R. Jain and K-Y. Siu, "A Brief Overview of ATM: Protocol Layers, LAN Emulation and Traffic Management", *Computer Communications Review*, April 1995
- [6] B. Stiller, "A Survey of UNI Signaling Systems and Protocols for ATM Networks", *Computer Communications Review*, April 1995
- [7] Cray Research Inc., "Asynchronous Transfer Mode (ATM) Administrator's Guide, SG-2193 1.0, Draft", May 1995