2

Protocol-, Operating System- and ATM Layer-Limitations in Practical Performance of UDP/IP and TCP/IP over ATM

Dipl.-Inform. André Zehl Dipl.-Ing Thomas P. Kusch

Technical University of Berlin / PRZ

Sekr. MA 073, Straße des 17. Juni 136, D-10623 Berlin, Germany
Tel.: +49.30.314-21172 FAX: +49.30.314-24590
Email: az@prz.tu-berlin.de

Abstract

This paper presents practical performance measurements with IP in a local area ATM network. A general experience running TCP/IP and UDP/IP traffic in a typical ATM LAN environment is an astonishingly poor performance. This paper shows some of the limitations we found out in analysing the reasons for the performance shortcomings. We mainly focus on operating system-, protocol- and ATM layer-limitations.
One of the key results of our work is, that the main reason for performance problems, besides the lack of QoS support in IP protocols, are current IP protocol´s implementations.

Keywords

High speed networking, practical performance analysis, IP over ATM.

1 INTRODUCTION

Asynchronous Transfer Mode (ATM) [ITU150][Prycker93] is an upcoming network technology, promising to replace current network technologies for specialized application areas (audio, video, data) in public networks as well as in private network environments. The standardization of ATM as a *Broadband Integrated Services Data Network* (B-ISDN) through the *International Telecommunication Union* (ITU) and a consortium of network vendors (*ATM Forum*) is still in progress..
Because of the ongoing standardization efforts for ATM, IP over ATM has only partly been developed as Internet Standard by the *Internet Engineering Task Force* (IETF) working group "IP over ATM". Issues like IP broadcast and multicast support, ATM signalling support for IP, "routing over large clouds" are still under study. Until now, mainly the IP encapsulation and IP to E.164 or NSAP ATM address resolution are standardized. RFC 1483 describes encapsulation mechanisms of IP over the ATM Adaption layer (AAL) 5 (Figure 1b). RFC 1577 describes the address resolution protocol (ARP) adaption to ATM, called ATMARP.
ATMARP introduces a client-server mechanism for address resolution for a logical IP subnet (LIS)(Figure 1a). The main characteristics of a LIS are, that all stations in the network belong

R. Puigjaner (ed.), *High Performance Networking*

to the same IP subnet and that different IP networks are interconnected through routers.

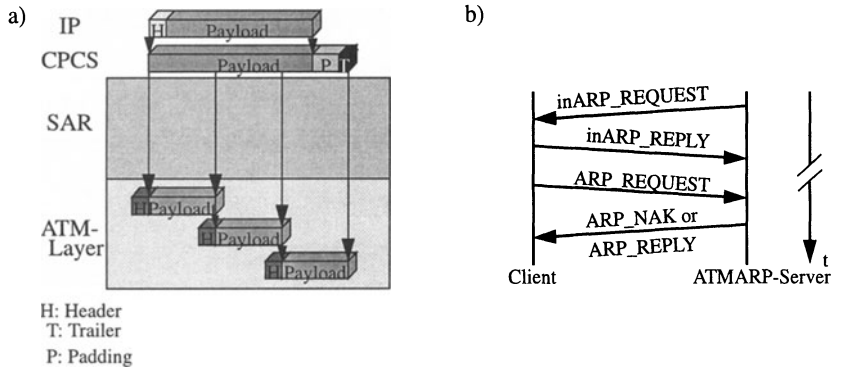

Figure 1 a) Overview on the IP encapsulation over AAL5 b) ATMARP/InATMARP Client-Server Model.

In RFC1577 and RFC1626 the *Maximum Transfer Unit* (MTU) to be used to fit the IP *Protocol Data Unit* (PDU) in the maximum 64 Kbytes sized AAL5-PDU is defined to be 9180 bytes.

2 PERFORMANCE LOSS IN THE IP/ATM STACK THROUGH PROTOCOL OVERHEAD

One of the important limitations analysing protocol performance over ATM, is the ATM layer overhead itself. Constant PDU or ATM cell size, with a 5 bytes header of a 53 bytes cell length imposes a loss of at least 9.43% on the physically available bandwidth.

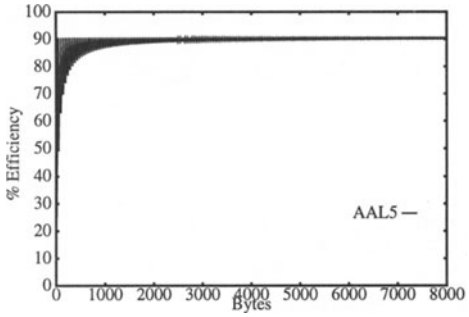

Figure 2 Efficiency of AAL5 packet usage.

Because cells are not always completely filled with user data, cells are partly filled with padding bytes. Especially for small packets (less than 512 bytes) the efficiency of cell usage ranges between 50 and 90%. The modest efficiency influences protocols on top of AAL5 as well.

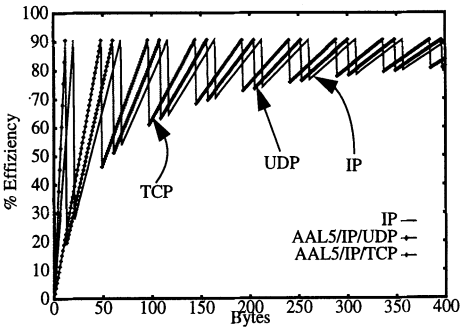

Figure 3 Efficiency of IP, TCP/IP and UDP/IP over AAL5.

Because small packet sizes are most common in typical LAN environments, the low efficiency has a strong influence on higher layer protocols like TCP/IP or UDP/IP for small packets. Measurements with NNstat [Braden91] in our local heterogeneous LAN (consisting of FDDI, Ethernet, ATM and HIPPI) shows, that about 75% of the data traffic are packets with a size of less than 320 Bytes. TCP is 50.8% and UDP 38.3% of the overall traffic. For WANs the influence is even more significant, because applications with rather big packet sizes, e.g. NFS, are less common in wide area data networks. Applications with rather small packet sizes, e.g. TELNET or RLOGIN are more common here [Cáceres91]. For large higher layer PDUs, the problem of partly filled cells is less important.

3 PRACTICAL PERFORMANCE MEASUREMENTS

3.1 Description of the Hardware and Software used in the Measurements

The performance measurements presented in the following were made on several hardware platforms [Fore02b][HP94c]. ATM switches employed are a Fore Systems ForeRunner ASX200 and a cisco LS100. The switches offer 100 Mbit/s TAXI and 155 Mbit/s OC3c as line interfaces. The drivers employed were all conforming to RFC1483, the HP cards were additionally in conformance with RFC1577.

Table 1 Hardware equipment employed

SPARCstation 10	32 MB RAM, SunOS 4.1.3, SBUS, Fore Systems SBA200 ATM adapter, 100 Mbit/s TAXI
SPARCstation 2	16 MB RAM, SunOS 4.1.3, SBUS, Fore Systems SBA100 ATM adapter, 100 Mbit/s TAXI.
HP755	64 MB RAM, HP-UX 9.01, EISA, HP J2802A, 155 Mbit/s SDH OC3c.

In practice, connections, i.e. *Virtual Paths* (VP) and *Virtual Channels* (VC), with the Fore equipment had to be setup with the *Simple Protocol for ATM Network Signalling* (SPANS) of Fore Systems [Biagioni93][Fore92c]. Connections with the HP equipment were setup with UNI3.0 signalling over the cisco switch.

In the most simple setups, pairs of equal workstations were interconnected over a single switch (Figure 4).
Most measurements were taken with the ttcp and netperf [HP94b] IP performance measurement tools. Tools to read driver statistics and transport system kernel variables were employed as well. An HP75000 ATM analyser [HP94a] was used for measurements with the switch . If not stated otherwise, all measurements were taken with socket buffers (receive and send) set to 8 Kbytes, TCP_NODELAY option on, packet sizes in the range between 64 and 40000 bytes, and the measurements running for 1 minute each. All results are average values of a 1-minute-run. Measurements were taken during *zero load condition* with no process running except the measurement tools and the regular system processes. Our main focus of the measurements was put on throughput performance, but we measured Round Trip Timing (RTT), network load, packet and byte error rate as well.

Figure 4 Most simple measurement setup.

3.2 Hardware platform´s differences

With first measurements we wanted to figure out the influence of the different hardware platforms employed. We therefore compared the networking performance of several workstations and network interface cards (NICs).

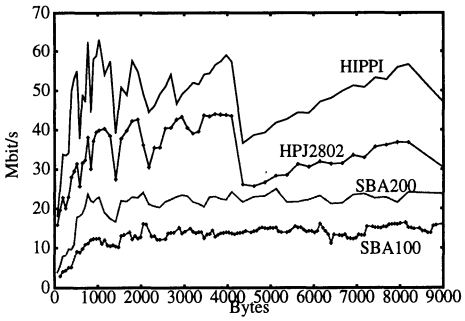

Figure 5 Comparison of TCP over SBA100, SBA200, HPJ2802A and HP HIPPI NICs.

Except for the meanwhile rather "old" Fore SBA100 adapter card, performance of the HP and Fore SBA200 showed a performance in the area of a quarter of the physical line bandwith. The maximum TCP performance reached approximately 44 Mbit/s with the HP adapter card and 26 Mbit/s with the Fore SBA200 adapter card.
For reasons of comparison, we put a measurement with two HP755 with 800 Mbit/s HP High Performance Parallel Interface (HIPPI) adapter cards into Figure 5 as well. These cards use the internal SGC bus of the HP workstation, instead of the EISA bus. The HIPPI equipment shows

up to 62 Mbit/s with TCP. The HIPPI setup was the same as the ATM scenario, two workstations interconnected through a switch, in this case a 8x8 Input Output Systems Corporation (IOSC) HIPPI switch.
The reasons for the extremely staggering shape of the graphs will be discussed in section 3.3.

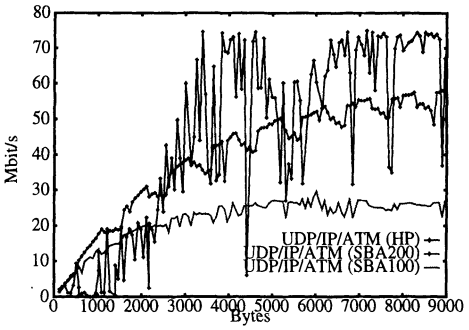

Figure 6 Comparison of UDP over SBA100, SBA200, HPJ2802A.

The UDP measurements (Figure 6) were quite ambivalent. While UDP measurements for the SBA100 card came up to approx. 25 Mbit/s, the SBA200 card came up to 60 Mbit/s and the UDP performance for the HP cards came up to over 70 Mbit/s. In contrast, HIPPI measurements show a drastic loss in throughput, keeping the bandwidth lower than 1 Kbit/s for all packet sizes (not shown in figure)! Reasons therefore will be discussed in section 3.4.

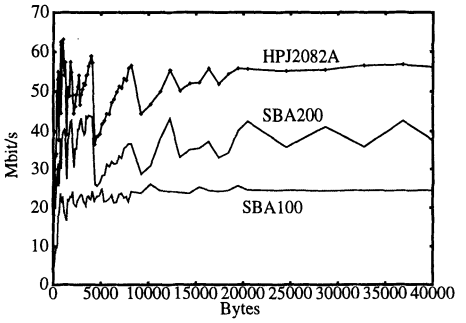

Figure 7 TCP packetsize up to 40000 bytes.

Measurements with a packet size of up to 40000 Bytes confirm the better throughput with large packet sizes (Figure 7).
Measurements with delay showed, that in the ATM LAN environment under study, the delay for TCP connections was in a range between 0.5 and 4 ms, which is no problem for voice transmission or similar (Figure 8).

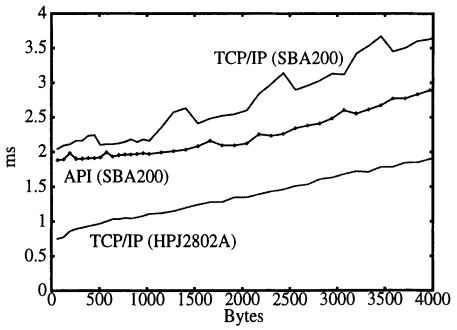

Figure 8 IP delay over ATM (HPJ802A, SBA200).

3.3 Problems with TCP/IP over ATM

One special characteristic in the TCP measurements above was the staggering shape of the graphs. The reason for this peculiar shape is the byte and packet loss at TCP level caused by damaged packets (Figure 9).

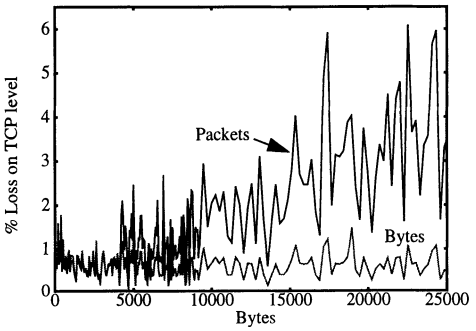

Figure 9 TCP Byte and packet loss caused by damaged packets.

Up to 6% packet loss for a single connection in a network with no other traffic is an extraordinary bad value, a value of less than 0.0025% for CSMA/CD networks is generally considered normal. The result of the high packet loss rate are frequent retransmissions with TCP's go-back-n algorithm. Because there is no guarantee, that the retransmission will succeed, the retransmissions can lead to a continuous increase of damaged packets and a decrease of successfully transmitted packets. We'll see later, that the reason for the damaged packets on the TCP level are non-deterministic cell drops on the switch and the receiving workstation.

The TCP *Congestion Avoidance* algorithm [Jacobson88] decreases the window size whenever a packet is not correctly acknowledged. The assumption is, that the situation of a network congestion will diminish as soon as a part or all connections reduce their data transmitted. After the reduction of the window size, the algorithm exponentially increases it's size, to slowly adopt to the new optimal situation.

In the ATM network examined, a continuous, non-deterministic cell drop leads to repeated window size reductions, as measurements have shown, without achieving the desired effect of

clearing the supposed network congestion. Thus, the window size and the throughput is decreased repeatedly, with no effect besides reducing the throughput.
One of the assumptions of the *Congestion Avoidance* algorithm is a packet loss rate caused by the network to be "<<1%" [Jacobson88]. That's not the case in the ATM network under study. We are presently studying possible optimizations.

3.4 UDP/IP problems

The problem with UDP in a high speed network environment is, that UDP has no flow control mechanism. This is a problem especially in environments with workstations, which are different powerful in terms of CPU performance.

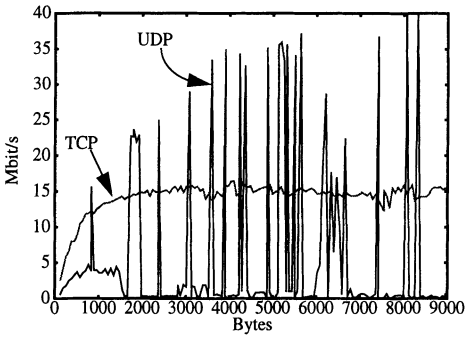

Figure 10 UDP/IP and TCP/IP between SPARC10 and SPARC 2.

In such an environment, continuous socket buffer overflow on the receiving workstation leads up to 100% packet loss (Figure 10). While the high priorised driver can write its data in the ip_input queue, the lower priorised kernel top half can write the data into the socket buffer. But the lower priorised user level application process has seldom a chance to read the data off the socket buffer. The socket buffer is repeatedly overrun, and only in rare cases, the socket buffer is empty, the kernels fills it with data, and the application reads the data completely.
In such an environment, useful high speed networking with ATM is impossible, due to lack of processing power of the receiving workstation and the double copy architecture in the kernel. For HIPPI-measurements made, this proved to be even a more serious problem, resulting in UDP throughput of less than 1 Kbit/s!

4 WAYS TO OPTIMIZE PERFORMANCE

4.1 ATM Switch Characteristics

Measurements with the HP75000 network analyser on the ForeRunner ASX200 Switch show an average 0.01% continuous cell loss on the switch during the above measurements. Here we have one of the reasons for the lost and damaged packets on higher protocol layers. This cell loss rate propagates damaged AAL5 PDUs, IP and TCP/UDP PDUs. Measurements show that the AAL5 *Convergence Sublayer* (CS) PDU loss through congestion during the measurements above comes up to 0.1% (Figure 11). Because a TCP or UDP PDU can be distributed over sev-

eral 9180 Bytes IP-PDUs, a single lost AAL5 PDU can invalidate several AAL5 PDUs, transporting a fragmented IP datagram.

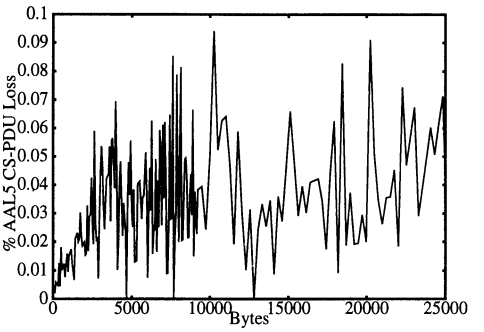

Figure 11 AAL5 CS-PDU Loss caused by congestion.

The Switch cell loss is only a partial explanation for the poor performance of IP transport protocols. Nonetheless, optimizations are necessary. Most simple optimizations against cell loss on the switch are increased buffering queues and, especially for bursty traffic like IP, *Traffic Shaping* algorithms [Hemmer91]. Switch optimizations like *Early* or *Partial Packet Discard* [Romanov94] might solve this problem as well.

4.2 Network Software

On workstations of similar processing power, UDP is able to come near the physically available bandwidth. Nontheless, optimizations in the protocol implementation are possible. One obvious optimization is the usage of large receive buffers (Figure 12).

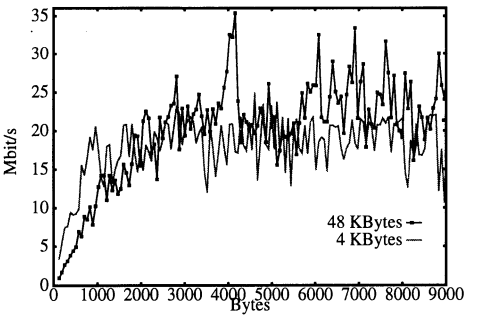

Figure 12 Optimizations through large socket receive buffers.

Up to 50% performance was gained by a socket buffer size increase. These optimizations can either be performed on a per application base or for the whole network transport system through increase of the tcp_sendspace and tcp_recvspace variables in BSD UNIX derived systems.

Another important source of performance loss is the checksum calculation in software. To figure out the influence of checksum calculations on a SPARCstation 10 we made measurements

with the optional UDP checksum calculation (Figure 13). As seen above, large packets show higher throughput. Unfortunately checksum calculations of large packets show up to 20% performance loss with checksumming on. Because checksum calculation in UDP is straightforward with no side effects like retransmissions, this performance loss occurs for TCP on a similar scale.

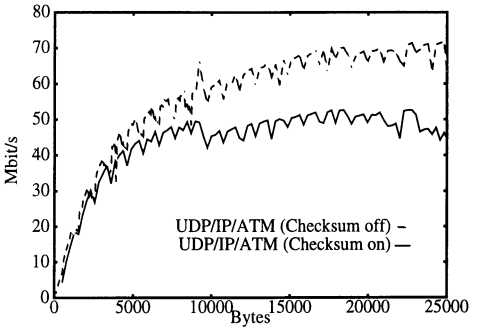

Figure 13 Influence of software checksum calculation.

Besides a hardware checksum calculation during DMA or similar, handcoded instruction optimizations have been proposed for a costless checksumming on RISC processors [Partridge93]. For future network applications which make use of large datagrams, those optimizations are necessary for better performance.

4.3 Operating System and Driver characteristics

Besides the cell loss in the ATM *switching unit*, a majority of cells is dropped in the ATM adapter card of the *receiving workstation (Figure 14)*. Measurements have shown that up to 8% of cell loss occurs in the driver for, e.g., a SBA200 ATM NIC.

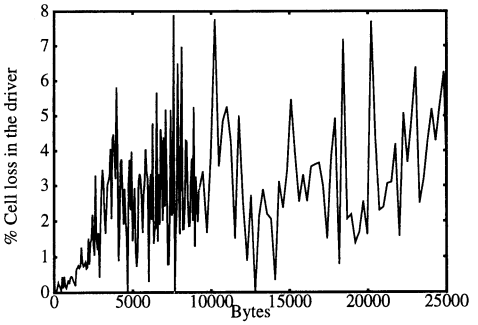

Figure 14 Cell loss in the SBA200 driver.

This high cell loss is the main reason for the high packet loss rate on the transport protocol layer measured above. Again the problem is a limited queueing capacity in the adapter cards. On the other hand the operating system itself imposes a great influence on the performance. A comparison between a provided *Application Programmers Interface* (API) for the Fore SBA200 adapter and the UDP/TCP socket API shows, that even UDP stays up to 50% below

the API performance (Figure 15).

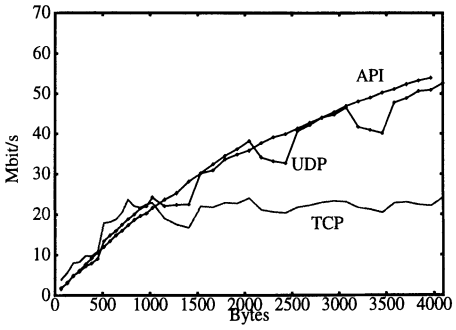

Figure 15 Comparison of UDP, TCP and Fore socket API.

Besides the driver limitations, two operating systems characteristics occurred to be influential on the performance of IP over ATM.
Especially the UDP measurements show a *regular staggering pattern*. Having a closer look at the measurements, it´s obvious that the "saw tooth" appear every 512 bytes. Up to 512 bytes, SunOS allocates 112 bytes sized mbufs. Above 512 Bytes, a 1024 bytes cluster mbuf is allocated. On HPs the cluster mbuf size is 4096 bytes, thus the breaks occur in 4096 byte steps. These mbuf sizes are ideal for MTUs like the 1500 bytes Ethernet MTUs. They are less than optimal in an environment with 9180 bytes MTU. Larger mbufs, in sizes of 8 Kbytes or even 16 Kbytes are necessary, so the mbuf allocation doesn´t impose a negative influence on the packetization delay and thus the overall throughput.

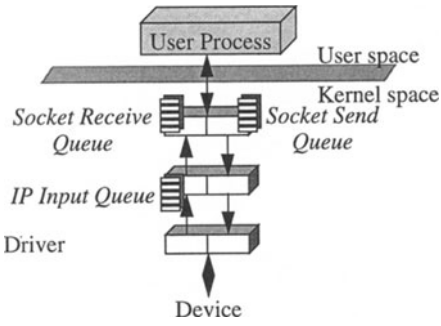

Figure 16 Double Copy Stack.

Another negative influence on the overall system performance is the *Double Copy stack* in BSD UNIX derived systems (cf. [Partridge93])(Figure 16).
The double copy stack is harmful in two ways: As described above: the double copy stack partly results in high socket buffer overflows, especially on "slow" workstations. On the other hand the process of accessing and moving data in memory twice is costly and actually not necessary. The logical design decision of the layered architecture is inefficient in the implementation.

5 CONCLUSION AND FURTHER RESEARCH

This paper presented an analysis of the performance problems that arise by using "off the shelf" UNIX workstations in a typical ATM LAN environment. We have identified several sources of problems of different impact. Some of the problems, like the approximately 10% overhead of the ATM layer protocol which are propagated to higher layers, are "generic" and cannot be solved. Other problems, like random cell drops in the switching units during higher layer, e.g. IP, packet bursts can be addressed by traffic shaping mechanisms in the end systems.
UDP/IP on workstations with a similar CPU performance are able to come close to the transmission rate of the physical medium for large datagrams, as long as the machine is powerful enough to process large amounts of incoming data without dropping packets. The high packet loss rate of TCP, due to cell loss in the ATM layer, leads to frequent retransmissions. We are presently studying optimizations in the congestion avoidance algorithm.
Measurements for TCP and UDP showed that especially for large datagrams the throughput is best. Applications therefore should use large datagrams and set socket buffer sizes high. Network protocol implementations need optimizations at several places as proposed in this paper.
The partially rather poor performance of IP over ATM is no shortcoming of the *IP protocol stack* itself, but rather of (for conventional network technologies well adapted) network protocol *implementations*.
Future work will include further performance analysis with ATM equipment of different vendors as well as development of an optimized transport layer implementation.

6 ACKNOWLEDGEMENT

We would like to thank Tanja Zseby and Krishnan Thosecan for support with the measurements and for revising this paper.

LITERATURE

[Armitage94] Armitage, G.J. 1994: *IP Multicast over UNI 3.0 based ATM Networks. Internet Draft.* 26. August 1994. Bellcore. Morristown. New Jersey.

[Biagioni93] Biagioni, Edoardo; Cooper, Eric; Sansom, Robert 1993: *Designing a Practical ATM LAN.* In: IEEE Network.Vol. 7. Nr. 2. March 1993.

[Braden91] Braden, Robert T,; DeSchon, Annette L. 1991: *NNStat: Internet statistics Collection Package. Introduction and User Guide.* USC/ Information Sciences Institute. Marina del Rey.

[Cáceres91] Cáceres, Ramón 1991: *Efficiency of Asynchronous Transfer Mode Networks in Transporting Wide-Area Data Traffic.* Computer Science Division. University of California. Berkeley.

[Fore92a] Fore Systems 1992: *ASX ATM Switch Architecture Manual. Draft.* Pittsburgh. PA.

[Fore92b] Fore Systems 1992: *SBA-100 SBUS ATM Computer Interface. User's Manual.* Pittsburgh. PA.

[Fore92c] Fore Systems 1992: *Interim Signalling Protocol for ATM Local-Area Networks. Draft.* Pittsburgh. PA.

[Fore93] Fore Systems 1993: *ForeRunner ASX-100 ATM Switch User's Manual. Draft.* Warrendale. PA.

[Hemmer91] Hemmer, Hilde; Huth, Thomas 1991: *Evaluation of Policing Functions in ATM Networks.* In: Cohen, J.W.; Pack, C.D. 1991: Queueing, Performance and Control in ATM (ITC-13). North-Holand. Amsterdam. London. New York. Tokyo. ISBN 0-444-89114-5.

[HP94a] Hewlett Packard 1994: *Broadband Series Test System. AAL User's Guide.* Edmonto, Alberta. Kanada.

[HP94b] Hewlett Packard 1994: *Netperf: A Network Performance Benchmark. Revision 1.9alpha.* Information Networks Division.

[HP94c] Hewlett Packard 1994: *J2802A. HP ATM Adapter Installation and Configuration Guide.* Grenoble. Frankreich.

[ITU150] ITU-T 1993: *B-ISDN Asynchronous Transfer Mode Functional Characteristics.* ITU-T Recommendation I.150.

[Jacobson88] Jacobson, Van 1988: *Congestion Avoidance and Control.* In: Proceedings of SIGCOMM '88. ACM. Stanford. Kalifornien. August 1988.

[Partridge93] Partridge, Craig 1993: *Gigabit Networking.* Addison-Wesley. Reading. Menlo Park. New York. etc..

[Perez94] Perez, M.; Liaw, F.; Grossman, D.; Mankin, A.; Hoffman, E.; Malis, A. 1994: *ATM Signalling Support for IP over ATM. Internet Draft.* April 1994.

[Prycker93] Prycker, Martin de 1993: *Asynchronous Transfer Mode.* Prentice Hall. New York. London. Toronto. Sydney. Tokio. Singapur. München. Mexico.

[RFC1483] Heinanen, Juha 1993: *Multiprotocol Encapsulation over ATM Adaptation Layer 5. RFC 1483.* Telecom Finnland. Tampere. Finnland.

[RFC1577] Laubach, M. 1994: *Classical IP over ATM. RFC 1577.* Hewlett Packard Laboratories. Palo Alto. California.

[Atkinson94] Atkinson, R. 1994: *Default IP MTU for use over ATM AAL5.* RFC 1626. Naval Research Laboratories. Washington. DC.

[Romanov94] Romanow, Allyn; Floyd, Sally 1994: *Dynamics of TCP Traffic over ATM Networks.* Mountain View. Berkeley. California.

[UNI3.0] ATM Forum 1993: *ATM User-Network Interface Specification. Version 3.0.* September 1993.