Reliable Group Communication in ATM Networks

Georg Carle

Institute of Telematics
University of Karlsruhe
D-76128 Karlsruhe, Germany
Telephone: ++49 / 721 / 608-4027
Fax: ++49 / 721 / 388097
E-Mail: carle@telematik.informatik.uni-karlsruhe.de

Abstract

Upcoming applications have demanding communication needs. One requirement is the provision of a reliable high performance multipoint communication service. If a reliable service in ATM networks is based on traditional transport protocols like TCP, severe performance degradations may be observed.[1] Additional problems occur for the provision of a reliable multipoint service, where errors occur more frequently, and where transmitters need to deal with many receivers. In order to meet performance requirements of demanding applications, suitable error control schemes are required which allow an efficient use of network resources and which scale well for large groups. This paper presents a novel concept for support of reliable multipoint communication in ATM networks, based on a new adaptation layer protocol, called Reliable Multicast Service Specific Convergence Sublayer (RMC-SSCS), and on a new network element, called Group Communication Server (GCS). The functionality of the adaptation layer protocol and the group communication server are described, and a basic implementation architecture for the server is proposed. For the deployment of a GCS in different communication scenarios, achievable efficiency improvements are analysed.

Introduction

In the evolution of high speed networking, two developments will be of growing importance. One issue is the fast growing deployment of ATM networks, both in local and in wide area networks. The other issue is the increasing importance of group communication scenarios. Upcoming applications, for example in the areas of computer-supported cooperative work (CSCW), distributed applications and virtual shared memory systems require point-to-multipoint (Multicast, 1:N) as well as multipoint-to-multipoint (Multipeer, M:N) communication.[2] For a growing number of applications such as multimedia collaboration systems, the provision of a multicast service with a specific quality of service (QoS) in terms of throughput, delay and reliability is crucial.

If multipoint communication is not supported by the network or by the end-to-end protocols, multiple point-to-point connections must be used for distribution of identical information to the members of a group. The support of multicasting is beneficial in various ways. It saves bandwidth, reduces processing effort of end systems, reduces mean delay and simplifies addressing and connection management.

Various issues need to be addressed in order to provide group communication services in ATM networks.[3,4] Switches need to incorporate a copy function for support of 1:N virtual channels (VCs). Signaling must be capable of managing multipoint connections, and group management functions need to be provided for administration of members joining and leaving a group. Procedures for routing and call admission control (CAC) need to be adapted for multicast communication. Another key problem that must be solved to provide a reliable multipoint service is the recovery from cell losses due to congestion in the switches.

Section 2 presents two problems that need to be overcome for the provision of a reliable multipoint service: support of multiple transmitters in a group, and cell losses due to congestion. A brief overview on existing error control mechanisms and on protocols that apply these mechanisms is given. Section 3 presents the conceptual framework for the provision of an efficient reliable multipoint service, based on a new adaptation layer protocol and the deployment of Group Communication Servers. In section 4, results of a perfomance analysis are given.

Multipoint communication in ATM networks

Multipoint bearer service in ATM networks

Applications may require multipoint communication of the types one-to-many, many-to-one and many-to-many. There are a number of ways how to support these communication types in ATM networks.[5] Virtual paths and virtual channels may be of the types point-to-point and point-to-multipoint. Many ATM switch designs are already prepared to copy incoming cells to multiple output ports, providing a basic support for multicast communication in ATM networks.

Support of multipoint connections in signaling protocols is currently under development. In the draft recommendation of the signaling protocol for B-ISDN,[6] support of multipoint connections is not yet included. In the User-Network Interface (UNI) specification version 3.0 of the ATM Forum,[7] phase 1 signaling is specified which allows the management of point-to-multipoint connections. Multipoint-to-multipoint connections are not supported by phase 1 signaling, but two techniques are proposed for multipeer communication.

According to the first proposal, each node in a group that wishes to communicate has to establish a point-to-multipoint connection to all of the other nodes of the group. N point-to-multipoint connections are required for a group with N members. This solution does not scale well for large groups. For large, long-lived groups, numerous virtual channels need to be maintained. If one receiver joins or leaves a group, every multicast tree must be modified.

According to the second proposal, each node has to establish a point-to-point connection to a 'Multicast Server'. A point-to-multipoint connection from the Multicast Server to every member of the group is used to transmit messages to the members of the group. This requires N point-to-point connections and one point-to-multipoint connection, improving the scalability significantly. If this approach is selected, mechanisms must be applied in order to distinguish cells of different senders.[8,5] One possibility is to distinguish the cells

based on an identifier in the payload of the cell. The MID-field of AAL3/4[9] may be used for this purpose. In this case, MID fields must be negotiated, and a MID demultiplexing function must be integrated into every receiver. AAL5[9] allows a simpler implementation of the adaptation layer, but it does not provide a field for demultiplexing cells. If cells of different frames are mixed, the receiver is only able to detect the collision by checksum violation and to discard the affected frames. In order to avoid these collisions, the multiplexing of different VCs onto a single VC needs to be done in a way that every receiver receives all cells of one frame before receiving cells of another frame. Such a mechanism may operate either in reassembly mode or in cut-through mode. In reassembly mode, forwarding of an incoming AAL5 frame starts after the reception of the last cell of this frame. In cut-through mode, already the first incoming cell of a frame may be forwarded if no other frame of the group is in the process of forwarding.

Cell loss in ATM networks

Two factors need to be considered which cause ATM networks to discard cells: transmission bit errors in the cell header field due to noise, and buffer overflow in multiplexing or cross connecting equipment. While fibre optic transmission technology allows to keep the bit error probability very low, the most frequent cause for cell loss is buffer overflow. In ATM networks, statistical multiplexing provides a high degree of resource sharing. Short periods of congestion may occur due to statistical correlations among variable bit rate traffic sources, resulting in buffer overflow. The probability for cell loss may vary over a wide range, depending on the strategy for usage parameter control (UPC) and call admission control which is applied. If very low cell loss probabilities are to be guaranteed even for highly bursty sources, only part of the network resources may be utilised. Utilisation may be increased on the risk of higher cell loss rates. Cell losses due to buffer overflow occur during situations of congestion, caused by superpositon of traffic bursts. Therefore, they do not occur randomly distributed, but in bursts and show a highly correlated characteristic.[10] If a reliable service has to be provided, mechanisms are required which are able to handle this type of error efficiently. For ATM multicast connections, the problem of cell losses is even more crucial than for unicast connections. Collisions of the multicast VC with independent unicast VCs may occur independently at every output port of a switch. For multicast switches with dedicated copy networks, additional collisions may occur for correlated arrivals of bursts in different multicast VCs.[11]

Error control mechanisms

For applications that cannot tolerate the cell losses of the ATM bearer service, error control mechanisms are required. Error control consists of two basic steps: error detection and error recovery. Error control is difficult in networks that offer high bandwidth over long distances. High data rates in combination with long propagation delays result in high bandwidth-delay products. A large amount of data may be in transit. For example, at a distance of 5000 km and a data rate of 622 Mbit/s, more than 2 MByte may be stored by the link. This causes problems for the following reasons:

- End-to-end control actions require a minimum of one round-trip-delay, and retransmissions require large buffers and may introduce high delays;
- Efficient error control with timer-based loss detection is difficult, because delay variations do not allow very accurate timer setting, causing deterioration of the service quality;
- Processing of error control needs to be performed at very high speeds, if no bottle-neck is to be introduced.

For the provision of a reliable service, ARQ (Automatic Repeat ReQuest) mechanisms are required. They are widely used in current data link and transport protocols. FEC (Forward Error Correction) may increase the reliability of a bearer service, but only additional ARQ mechanisms may provide a reliable service, as a residual error probability remains for pure FEC. In every retransmission based scheme, the transmitter needs to store messages upon acknowledgement. At least the data of one round-trip delay needs to be stored. For go-back-N protocols, implementation of transmitter and receiver may be very simple, and no buffering is required by the receiver. For selective repeat protocols, implementation of transmitter and receiver is more complex, and a large buffer is required for transmitter and receiver. Processing overhead of ARQ methods is proportional to the number of data and acknowledgement packets that are processed. For point-to-point communication, ARQ mechanisms are well understood, and a number of protocols for data link layer or transport layer, employing these mechanisms, are known. For multicast communication, there are still many open questions concerning acknowledgement and retransmission strategy, achievable performance and implementation. Large groups require that the transmitter stores and manages a large amount of status information of the receivers. The number of retransmissions is growing for larger group sizes, decreasing the achievable performance. Additionally, the transmitter must be capable of processing a large number of control information. If reliable communication is required to every multicast receiver, a substantial part of the transmitter complexity is growing proportionally to the group size. To overcome this problem, a scheme that provides reliable delivery of messages to K out of N receivers may be applied (K-reliable service).

Protocols for error recovery

According to the B-ISDN protocol reference model it is planned to integrate error control mechanisms into the Service Specific Convergence Sublayer (SSCS) of the adaptation layer. This is called assured mode service.[9] Up to now, only one SSCS-Protocol that offers a reliable service is subject of standardisation. This is the Service Specific Connection Oriented Protocol (SSCOP) which provides end-to-end flow control and retransmission of lost or corrupted data frames by operating in go-back-N or selective repeat mode. However, SSCOP does not support assured mode multicast connections.

There is a large number of layer 2 and layer 4 protocols that provide ARQ mechanisms for a reliable point-to-point service, but only a limited number that provides a reliable multicast service. Transport protocols that are suitable for a connectionless network layer, as for example TCP, TP4 or XTP, provide more functionality than required for a SSCS-Protocol. These transport protocols need to handle network packets that are received out of sequence, and need mechanisms for connection management. A SSCS protocol for reliable service may be simpler, as it may reject cells that are received out of sequence, and may use control plane protocols for connection management.

Implementation of communication subsystems

While transmission capacity was growing enormously over the last years, protocol processing and system functions in the transport component turned out to be a performance bot-

tleneck. High performance communication subsystems, based on parallel protocol processing[12] and hybrid architectures with hardware components for time-critical operations[13,14] are required if a service with high throughput and low latency is to be provided for the applications. For highest performance, a complete VLSI implementation of transport subsystems is envisaged.[15]

The performance bottleneck of the transport component that can be observed for point-to-point-communication is even more crucial for reliable multipoint connections. For a growing number of receivers, processing of a growing number of control packets and management of extensive status information is required. For the provision of a high performance multipoint service, components that support multicast protocol processing need to be integrated into the transport subsystem.

In order to offer a wide range of services to the applications for various network parameters, several concepts of flexible communication subsystems are under development. The parallel transport system Patroclos[14] is a parallel implementation of a high performance transport system, offering a different protocol mechanisms which may be selected according to the needs of an application. The Flexible Communication SubSystem (FCSS)[16] is a configurable, function-based transport system. It utilises a de-layered communication architecture that performs the complete transport component functionality for a specific data stream. It provides flexibility and dynamics of QoS selection and control, supporting the application-specific configuration of the protocol machines based on automatic selection of protocol mechanisms out of a protocol resource pool.

Conceptual framework for Reliable Multipoint Communication in ATM Networks

A conceptual framework was developed that allows the use of error control mechanisms best suited for a specific multipoint communication scenario at locations that allow highest performance. Figure 1 presents the ATM network scenario with multicast mechanisms in the adaptation layer of ATM end systems and in dedicated Group Communication Servers.

Reliable Multicast Service Specific Convergence Sublayer (RMC- SSCS)

The integration of error control mechanisms into the Adaptation Layer needs to be done in a way that high throughput and low latency are guaranteed. In order to offer a reliable and efficient high performance multicast service, a Reliable Multicast Service Specific Convergence Sublayer (RMC-SSCS) protocol suitable for AAL5 (ATM Adaptation Layer Type 5) was developed. Its design ideas are based on the concept of an extended ATM adaptation layer,[17] on the parallel transport system Patroclos and on the flexible communication subsystem FCSS.

It may be selected if RMC-SSCS offers a reliable service to all receivers or to a subset of K receivers (K may be 0, 1, ... up to the number of receivers). Retransmissions may be performed in selective repeat or go-back-N mode. It can be selected if retransmissions are sent by multicast or individually. Receivers send acknowledgements periodically, after reception of a frame in which an 'immediate acknowledgement' bit is set, or after detection of a missing frame. Frames carry frame sequence numbers, and receivers may acknowledge cumulative positive, sending a lower window edge, and selective positive or negative, using bit maps. For flow control, acknowledgements contain the upper window

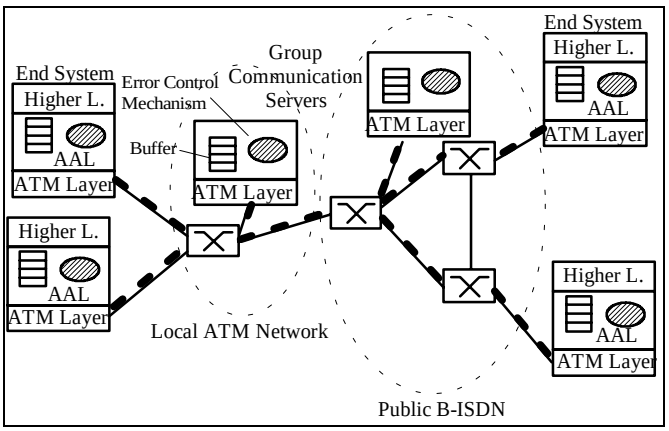

Figure 1: Group communication support in server and end systems

edge of the receiver buffer section reserved for the multipoint connection. The selection of acknowledgement mode, retransmission mode, and time-out periods is performed by the signaling protocol.

Group Communication Server (GCS)

The presented reliable multicast service specific convergence sublayer provides the required functionality for a high performance reliable multicast service. Further improvements of performance and efficiency may be achieved by the deployment of dedicated servers in the network that provide support for group communication. In many cases of multicasting, the achievable throughput degrades fast for growing group size. A significant advantage can be achieved if a hierarchical approach is chosen for multicast error control. The proposed Group Communication Server (GCS) integrates a range of mechanisms that can be grouped into the following tasks:

- Provision of a high-quality multipoint service with efficient use of network resources;
- Provision of processing support for multicast transmitters;
- Support of heterogeneous hierarchical multicasting;
- Multiplexing support for groups with multiple transmitters.

For the first task, performing error control in the server permits to increase network efficiency and to reduce delays introduced by retransmissions. Allowing retransmissions originating from the server avoids unnecessary retransmissions over common branches of a multicast tree.

For the second task, the GCS releases the protocol processing burden of a transmitter that deals with a large number of receivers, providing scalability. Instead of communicating with all receivers of a group simultaneously, it is possible for a sender to communicate with a small number of GCSs, where each of them provides reliable delivery to a subset of receivers. Integrating support for reliable high performance multipoint communication in a server allows better use of such dedicated resources.

For the third task, a GCS may use the potential of diversifying outgoing data streams, allowing conversion of error control modes and support of different qualities of service for individual servers or subgroups. The group communication server may provide sophisticated error control mechanisms that require a high implementation effort. For end systems with access to a local GCS, a simple implementation of error control will be sufficient for participation in a high performance multipoint communication over long distances.

For the fourth task, the GCS provides support for multiplexing of AAL5 frames onto a single point-to-multipoint connection. Using the signaling protocol, it may be selected if the GCS operates in reassembly or in cut-through mode. The connection structure of a multipeer scenario, based on group communication servers, is shown in figure 2.

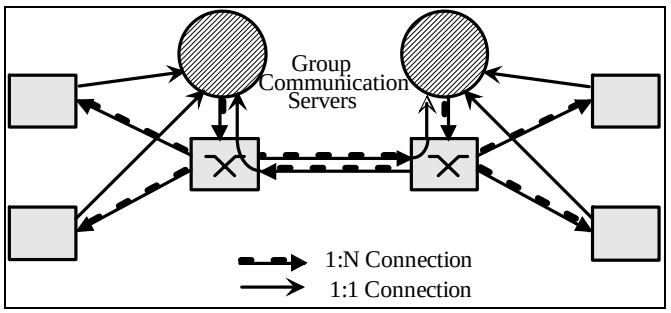

Figure 2: Connection structure

Figure 3 shows a proposed implementation architecture of the GCS. Main focus of the design was to achieve a high degree of pipelining. Acknowledgement processing for a large number of receivers is a potential bottleneck. Therefore, dedicated hardware support is provided in the *ARQ manager* unit for filtering and processing of acknowledgements, and for managing the status information of the group and of individual receivers. A component for window processing generates multicast flow control information required by the send manager. Generation of acknowledgements is also performed in the *ARQ manager* unit. The *send manger* unit schedules between ordinary transmissions, retransmissions and acknowledgements. The *connection manager* unit schedules between different connections and is also responsible for rate control and spacing. Additional hardware components are required for cyclic redundancy check (CRC), buffer management, list and timer management. For cell demultiplexing at the receiving side, a content addressable memory (CAM) is used to map the large VPI/VCI address space on smaller internal identifiers.

Signaling

For the management of multipoint connections based on RMC-SSCS and GCSs, an extended signaling protocol was developed that is based on the signaling protocols of ITU[6] and ATM Forum.[7] It allows the negotiation and selection of mechanisms used for a specific multipoint connection. Dynamic change of call participation is supported. Information of group membership is stored in a central database, administered by a group management server.

Performance Evaluation

Analytical methods were applied in order to evaluate the achievable performance of RMC-SSCS in selective repeat and go-back-N mode and to evaluate the potential gain by deployment of GCSs. Figure 6 shows the efficiency of the two retransmission modes in three different scenarios. Scenario 1 represents a basic 1:N multicast without GCS. Scenario 2 represents 1:N multicasting with a GCS that performs retransmissions as multicast. In scenario 3, the GCS uses individual VCs for retransmission. The analysis is based on the following assumptions: protocol processing times may be neglected, acknowledgements are transmitted over a reliable connection, and buffers are sufficiently large. A group of 100 receivers and a data rate of 622 Mbit/s are assumed. Two cases are distinguished. The upper diagram of figure 6 shows the efficiency for an overall distance of 1000 km (distance of 500 km from GCS to the receivers), and the lower diagram shows an overall distance of 505 km

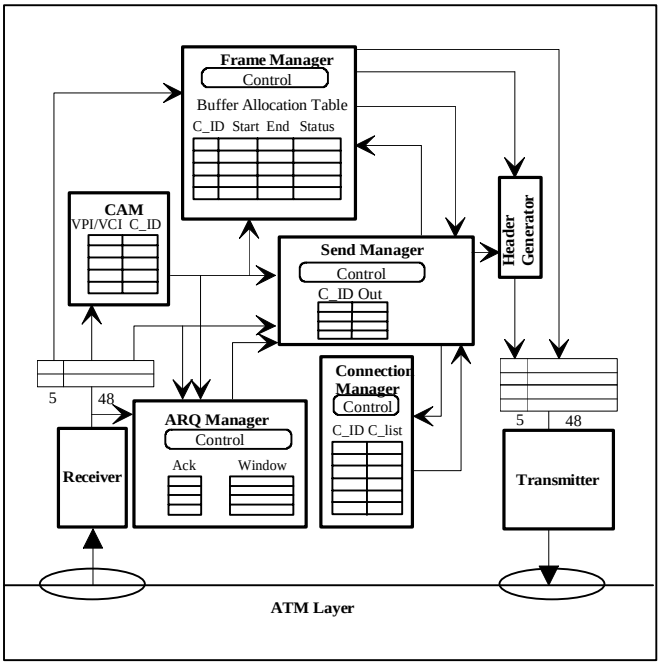

Figure 3: Architecture of the group communication server

(distance of 5 km from GCS to the receivers). The analysis shows that in all cases, the efficiency is increased significantly by the GCS. Highest efficiency may be achieved for scenario 3 and selective repeat. Scenario 2 improves significantly for a shorter distance between GCS and the receivers. Go-back-N retransmissions show acceptable performance only for moderate bandwidth-delay products. Regarding efficiency, scenario 3 and selective repeat should be selected. However, this solution requires the highest implementation complexity for end systems and GCS.

Conclusion

It was pointed out that existing strategies do not allow the provision of an efficient and reliable high performance multipoint service in ATM networks. A new concept was presented which has the potential to fulfil the requirements of upcoming distributed applications. It is based on a new service specific convergence sublayer RMC-SSCS and on a new network element called Group Communication Server (GCS). The functionality of these elements was presented,

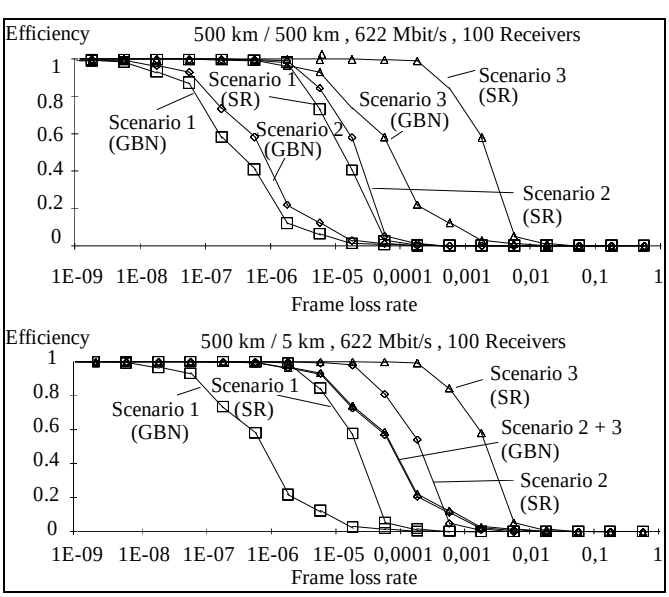

Figure 4: Efficiency analysis for go-back-N and selective repeat in scenarios with and without group communication server

and an implementation architecture for the GCS was proposed. An analytical performance evaluation was given which shows the potential improvement of multicast efficiency if GCSs are integrated into the network.
Subject of ongoing work is a more detailed evaluation of the achievable performance, including the influence of protocol processing delay. Implementation complexity will be evaluated to allow a better comparison of the presented alternatives. Additionally, integration of FEC into adaptation layer and GCS will be investigated.

Acknowledgement

The author would like to thank Martina Zitterbart and Torsten Braun for valuable discussions. The support by the Graduiertenkolleg "Controllability of Complex Systems" (DFG Vo287/5-2) is also gratefully acknowledged.

References

[1] Romanov, A.: "Some Results on the Performance of TCP over ATM," Second IEEE Workshop on the Architecture and Implementation of High Performance Communication Subsystems HPCS'93, Williamsburg, Virginia, U.S.A., September 1993

[2] Heinrichs, B.; Jakobs, K.; Carone, A.: "High performance transfer services to support multimedia group communications," Computer Communications, Volume 16, Number 9, September 1993

[3] Waters, A. G.: "Multicast Provision for High Speed Networks," 4th IFIP Conference on High Performance Networking HPN'92, Liège, Belgium, December 1992

[4] Gaddis, M.; Bubenik, R. and DeHart, J.: "A Call Model for Multipoint Communication in Switched Networks," Proceedings of International Conference on Communications ICC '92, pp. 609 - 615, June 1992

[5] Bubenik, R.; Gaddis, M.; DeHart, J.: "Communicating with virtual paths and virtual channels," Proceedings of the Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM'92, pp. 1035 - 1042, Florence, Italy, May 1992

[6] ITU-TS Draft Recommendation Q.93B: "B-ISDN User-network Interface Layer 3 Specification for Basic Call/Bearer Control," Geneva, 1993

[7] ATM Forum, UNI Specification Document Version 3, PTR Prentice Hall, Englewood Cliffs, NJ, U.S.A., 1993

[8] Wei, L.; Liaw, F.; Estrin, D.; Romanow, A.; Lyon, T.: "Analysis fo a Resequencer Model for Multicast over ATM Networks," Third International Workshop on Network and Operating Systems Support for Digital Audio and Video, San Diego, U.S.A., November 1992

[9] ITU-TS Draft Recommendation I.363: "BISDN ATM Adaptation Layer (AAL) Specification," Geneva, 1993

[10] Brochin, F., Pradel, E.: "A Call Traffic Model for Integrated Services Digital Networks," Proceedings of IEEE International Conference on Communications ICC'93, Geneva, Switzerland, May 1993

[11] Shimamoto, S., Zhong, W., Onozato, Y., Kaniyil, J.: "Recursive Copy Networks for Large Multicast ATM-Switches," IECE Trans. Commun., Vol. E75-B, No. 11, pp. 1208-1219, November 1992

[12] Zitterbart, M.; Tantawy, A.N.; Stiller, B.; Braun, T.: "On Transport Systems For ATM Networks," Proceedings of IEEE Tricomm, Raleigh, North Carolina, U.S.A., April 1993

[13] Carle, G.; Siegel, M.: "Design and Assessment of a Parallel High Performance Transport System," in Proceedings of European Informatics Congress - Computing Systems Architectures Euro-ARCH '93 (October 1993, Munich, Germany); P. P. Spies (Ed.), Springer Verlag Berlin Heidelberg, 1993

[14] Braun, T.; Zitterbart, M.: "Parallel Transport System Design," 4th IFIP Conference on High Performance Networking HPN'92, Liège, Belgium, December 1992

[15] Schiller, J.; Braun, T.: "VLSI-Implementation Architecture for Parallel Transport Protocols," IEEE Workshop on VLSI in Communications, Stanford Sierra Camp, Lake Tahoe, California, U.S.A., September 1993

[16] Zitterbart, M.; Stiller, B.; Tantawy, A.: "A Model for Flexible High Performance Communication Subsystems," IEEE Journal on Selected Areas in Communications, Volume 11, Number 4, pp. 507-517, May 1993

[17] Carle, G.; Röthig, J.: "BISDN Adaptation Layer and Logical Link Control with Resource Reservation for a Flexible Transport System," Proceedings of European Fibre Optics Communications and Networking Conference EFOC&N'93, The Hague, Netherlands, June 30 - July 2, 1993