RAP: Protocol for Reliable Transfers in ATM Networks with Active Switches [(1)]

José Luis González-Sánchez
Department of Computer Science, University of Extremadura, Escuela Politécnica de Cáceres
Avda. Universidad S/N. (10.071)
Cáceres (Spain)

Jordi Domingo-Pascual
Department of Computer Architecture, Polytechnic University of Catalunya. Campus Nord, Modul D6, Jordi Girona 1-3 (08034)
Barcelona (Spain)

Abstract *Significant research efforts are currently centred on effective mechanisms of cell loss control and congestion control for ATM networks. We present RAP (Reliable and Active Protocol); a new protocol for ATM networks that provides reliability to a set of privileged VPI/VCI. The protocol manages the privileged connections and offers the reliability taking advantage of the idle time in the traffic sources to do the retransmissions of CSCP-PDU-AAL5. RAP is an active protocol, that achieves reliability using NACK mechanisms and which is supported by active ATM switches equipped with DMRP (Dynamic Memory to store Reliable native PDU-AAL Type 5). Through several simulations implemented in Java we demonstrate the effectiveness of the protocol that also optimises the throughput using congestion control schemes as Partial Packet Discard (PPD). The simulations over ON/OFF sources analyse point-to-point, and also point-to-multipoint connections using objects, threads, synchronizations, and distributed processes implemented in Java language.TAP is a new and distributed mechanism to offer trusted transfers over ATM networks taking advantage of silent states in ON/OFF sources.*

Keywords: Active Protocol, ATM, reliability, active ATM networks, performance analysis.

1 Introduction and related work

ATM technology is characterized by its excellent performance and by offering to the user the possibility to negotiate QoS (Quality of Service) [1] parameters such as throughput, delay, jitter and reliability. While the first three parameters above have been researched, reliability may be the least studied characteristic, although there is important research based on CRC (Cyclic Redundancy Check) that guarantees the correction of headers and payloads of the ATM cells.

The cell of 53 bytes fixed length (a header of 5 bytes of information control and a payload of 48 bytes of data user) is known [2] as the basic unit of switching and multiplexing in ATM networks.

[(1)] This work is sponsored in part by the CICYT under Grant No. TEL97-1054-C03-03 and by the Junta of Extremadura (Education and Youth Council) and the European Social Fund Register No. BRV98101.

The header has the HEC (Header Error Control) field of 8 bits used as cyclic redundancy check for header error control only. For control of errors in user data there is another CRC per CS-PDU (Common Sublayer-Protocol Data Unit). Error control in ATM networks is performed end-to-end by the terminals.

The ATM networks should experience three types of errors [3-6]: cell losses due to congestion in switches; corruption of data portions due to bit errors, and switching errors due to undetected corruption of the cell header. We note that congestion is by far the most common type of error, and here is where we want to provide reliability with RAP. The errors due to bit errors are less frequent (typical values for cell loss probability with optic fiber ranges between 10^{-8} and 10^{-12}) and less important.

ARQ [7] is a mechanism based on retransmissions of data that were not correctly received due to some of the problems cited above. ARQ offers two variants (ACK and NACK) and both require the transmitter and receiver to exchange state information.

FEC [8-12] is an important alternative to ARQ whose operation principle is to encode the packets in emitter with redundant information so that it is possible to reconstruct the original packet reducing, or even eliminating, the

retransmissions and the implosion (negative effect due to massive request of retransmissions). While ARQ adds latency (due to cost of NACK) and implosion, FEC adds overhead and thus the redundant code added by this method is useless when the network is experiencing congestion. Hence, ARQ may not be suitable for applications with requirements of low latency, and FEC has worse behaviour in networks with low bandwidth or that experiences frequent congestion. In our protocol we adopted ARQ with NACK (using RM cells) to alleviate the effect of implosion. Support for reliable multicast cannot be based on retransmissions from the source. In RAP, the intermediate active nodes do the retransmissions. These methods solve the problem of retrieval of the corrupted cells, but they cannot resolve another undesirable, and more frequent problem, such as congestion and packet fragmentation in switches due to excess of traffic, problems in buffers, etc, which causes the effective throughput to be degraded. The most commonly proposed congestion control schemes to improve throughput and fairness, while minimizing delay in ATM networks, are the Random Cell Discard (RCD), Partial Packet Discard (PPD) [13], Early Packet Discard (EPD) [14]; Early Selective Packet Discard (ESPD) [15], Fair Buffer Allocation (FBA) and Random Early Detection (RED) [16]. We implement PPD in RAP. When an ATM cell reaches a full active switch buffer, RAP applies PPD, waits for the last PDU cell, discards the PDU and requests its retransmission sending a NACK to the emitter node.

In this work we propose a mechanism to take advantage of the idle periods in the data sources to retransmit the CPCS-PDU (Common Part Convergence Sublayer) of native AAL type 5 [5] without segmentation and reassembly. In our simulation the ON-OFF [17,18] sources generate traffic and the RAP protocol manages native CPCS-PDU-AAL 5 to its destination.

Active, open and programmable networks is a new technical area [19-24] to explore ways in which network elements may be dynamically reprogrammed by network managers, network operators or general users to accomplish the required QoS and other features as customized services. This offers attractive advantages, but also important challenges in aspects such as performance, security or reliability. Hence, this is an open issue for research and development in customized routings and protocols, whether to move the service code (placed outside the transport network) to the network's switching nodes. We bring active characteristics to RAP through hardware mechanisms and software techniques.

We have used the parallelism offered by Java to simulate process-events that generate traffic ON/OFF, congestion, noises, cell loss, misinserted cells and full-duplex links of communications.

Section 2 describes RAP protocol. In section 3, we present our prototype of an active switch that support the RAP protocol. Section 4 presents the simulation network models, presents the parameter values used, gives numerical results, and compares the results obtained for RAP. Section 5 describes and outlines our work currently in progress to enhance the RAP protocol. Finally we offer some concluding remarks in section 6.

2 General description of RAP protocol

In order to transmit a packet of information to a designated destination node through the ATM network, it is known that the information is segmented into 53-byte long ATM cells. The cell is the unit of transmission, but the technology offers other options to optimize the transfers. In the next paragraphs we describe the CPCS-PDU-AAL-5 as units of transmission that RAP manages. AAL (ATM Adaptation Layer) type 5 was proposed [5] to reduce overhead introduced by the native ATM cells. AAL-5 will be applied to Variable Bit Rate sources without time relation between transmission and destination.

Fig. 1 illustrates the CPCS-PDU format of native AAL-5 with all its fields. As we can see, the tail of PDU has 4 fields. The CPCS-UU (User-to-User indication) is used for the transfer of CPCS user to user information, and we utilize this octet in RAP as sequence number between the PDUs. We noted also the CRC (Cyclic Redundancy Check) used in AAL-5 to detect bit errors in the CPCS-PDU. The value of CRC is performed over all fields of the CPCS-PDU.

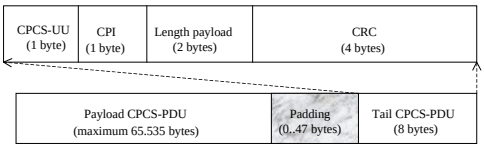

Fig. 1 CPCS-PDU AAL type 5 format

To implement NACK (Negative ACK) we use the standards RM (Resource Management) cells, without fixed frequency but generated when a switch is congested. This is to alleviate the negative overhead effect due to a fixed number of RM cells that will waste bandwidth. When congestion is detected and PPD discards a PDU, the active node generates a RM cell transmitted backwards to the upstream active switch indicating the sequence number of PDU to retransmit. The RM must also contain the VPI/VCI to identify the connection that has problems. The octet 8 of RM cell stores the sequence number of PDU requested, while the identifier VPI/VCI is stored in octets 16 to 19.

In some situations the retransmissions of PDU can be negative for network performance. This is the case when the active node requests retransmission and the next PDU in sequence is already being transmitted. To mitigate this negative effect the active switch waits before reacting to a loss, and retransmission is not requested until a certain time. This is the time needed for a RM cell to go back to the upstream active switch plus the time to initiate the retransmission of the PDU. Thus, the following contained economy time (*CET*) is defined as,

$$CET = x * \left(\frac{1}{PCR} \right)$$

in order to prevent unnecessary NACK transmissions and PDU retransmissions. The value of variable *x* may be between 1 and 2, and when the latter is reached, that is the round trip time. This approach is self-contained because the receiver gives the active switch enough prudential time to access the PDU in DMRP. Also this time is referred to as economy time because it saves or avoids retransmitting a PDU when the next PDU has already departed (if the receiver does not wait for *CET* the bandwidth can be wasted).

When the RM cell arrives at an active switch, RAP searches the solicited PDU and if this is still in DMRP then the PDU is retransmitted, as long as the idle time is sufficient. When a NACK arrives at a non-active switch this only processes the RM cell and resends it to its neighbouring switch in the direction of the active switch. The non-active switches do not have DMRP to retrieve PDUs and their function is only to send (or resend) PDUs forwards to their destination and also send NACKs (RM cells) backwards to the active nodes.

We highlight that RAP does not retrieve all PDU loss due to mechanism ARQ that can not guarantee if NACKs cells are lost. We want to compare the number of retransmitted PDU end-to-end with or without RAP.

3 Active ATM switches for RAP

The literature on this field studies several mechanisms to obtain advantage from active nodes. A network is active if there are active nodes in its multicast distribution trees with the capacity to execute the user's programs, and also if it implements mechanisms of code propagation. Many of the advantages of active protocols are achieved by installing active nodes at strategic points.

The ATM switch of our model network is an output buffered switch that just reads VPI/VCI information of arriving cells and forwards them to the corresponding output port. But we equipped this switch with active hardware and software techniques (described in next section) to achieve our objectives.

3.1 DMRP (Dynamic Memory of Reliable PDUs)

The main function of this memory is to PDU store temporally in the active switch so that they can be requested in retransmissions. The RAP protocol generates a copy of each PDU that arrives to the active switch. A copy is stored in DMRP while the original is sent to the receiver. If this second PDU experiences problems, due to congestion, the receiver can request the retransmission to the active switch, as we explain in the above section. Due to the great size of PDU-AAL-5 (up to 65,535 bytes), and the potential high number of connections (VPI/VCI), the volume of DMRP can be excessive. This is the reason that we limit the number of stored PDU for each connection, and furthermore, only support a controlled number of privileged VCI

with reliable transfers. In the following we present parameters to provide dimensions to the DMRP. We consider the maximum size of PDUs-AAL5 of 65,568 [5] octets (65,535 data user + 25 padding + 8 trailer) see ***Fig. 1***. To support all possible connections (4,096 VPI of 65,535 VCI each VPI) we will need a DMRP of 16,397 gigabytes. If we only want to offer reliability to all VCI of one VPI the DMRP will be 8 gigabytes. So, the size of memory depends on the number of reliable VCI that we want. We suppose that we dispose of a memory of fixed size to support a specific number of VCI. For example, if we dispose of a DMRP of 250 Megabytes, we will store 2 PDUs (of 65,535 bytes) of 2,000 privileged VCI. This amount is excessive for a memory and currently is not viable. In our case, we will utilize the reserved field UU (of 8 bits) as sequence number of each PDU. The octet UU allows us to differentiate up to 256 PDUs, and so the total size of DMRP will be 256 * 65,568 = 16 megabytes. This is the most unfavourable case, but if we consider the normal PDU size of 1,500 octets (MTU in segments TCP), we will need a DMRP of 256 * 1,500 = 384 Kbytes. As we have supposed that we store 2 PDUs by each VCI, we will obtain up to 128 privileged-VCI with reliable transmission guarantees. We consider that this size of memory and this number of VCI are sufficient.

RAP accesses the DMRP through index UU and VPI/VCI, and we have implemented different mechanisms to optimize the management and storage of PDUs. Also we simulate several software techniques to introduce active characteristics in switches. These mechanisms control and manage the privileged VCI and also we will offer an active mechanism to retrieve PDUs querying neighbouring active switches and to search optimized paths when a PDU is retransmitted. While a PDU from the DMRP is being retransmitted, incoming PDUs with the same VPI/VCI are discarded.

4 Simulation results and analysis

The simulation allows us to define the congestion probability in transmitters, receivers and each ATM switch. When a node is undergoing congestion, it then requests the retransmission of the corresponding PDU. We use NACKs to demand the PDU. The simulation also permits the user to introduce variable values such as ON/OFF traffic source parameters, the number of receivers or the number of non-active switches.

4.1 ON/OFF sources

In our simulation to analyse ATM cell loss we have used ON/OFF (bursty) traffic sources. The ON/OFF model [17,18] is used to characterise ATM traffic per unidirectional connection. ***Fig. 2*** shows this model as a source which either actively sends (ON state) CSCP-PDU-AAL-5 data for some time t_{on} at a traffic rate R or PCR (Peak Cell Rate) or is silent (OFF state) producing no cells for some time t_{off}.

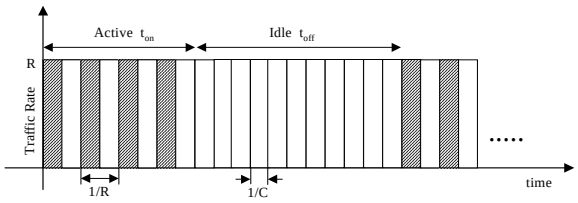

Fig. 2 Cell pattern for a single ON/OFF source

Table 1 shows the maximum and minimum source traffic descriptors used in our simulation.

Table 1 Source traffic descriptors ON/OFF

Source traffic descriptor	Parameter	Minimun	Maximum
Bandwidth Source	BS	64 kbit./s.	25 Mbit./s.
Cell arrival rate	R or PCR	167 cells/s.	65,105 c./s.
Cell inter-arrival time	1/R	6 ms.	15 μs.
Bandwidth link	BL	155.52 Mbit/s.	622 Mbit/s.
Cell slot rate	C or CSR	353,208 cell/s.	1,412,648 cell/s.
Service time per cell	1/C	2.83 μs.	0.70 μs.
Active time period	t_{on}	0.96 s.	1 s.
Mean number of cells in an active state	C_{on}	160 cells	65,105 cells
Time in idle state	t_{off}	1.69 s.	2 s.
Mean number of empty slots in idle states	C_{off}	596,921 cell slots	2,825,296 cell slots
Contained Economy Time	CET	(1/PCR)	2*(1/PCR)

We utilize a process that switches between an idle (silent) state, and the active state (sojourn time) which produces an average fixed rate of cells (between 64 Kbits/s to 25 Mbits/s) grouped

in PDUs of 1,500 bytes. During the ON states this process generates cells at a cell arrival rate R. Also periodically the source generates empty time slots. We use in all examples a C or CSR (Cell Slot Rate) of C=353,208 cell/s since our network model uses 155.52 Mbit/s links. When the cell arrival rate R is less than the cell slot rate C, there are empty slots during the active states as we can see in ***Fig. 2***.
The cell inter-arrival time 1/R is the unit of time for the ON state, and the mean duration in the active state is,

$$t_{on} = \left(\frac{1}{R}\right) * C_{on},$$

Also, the mean duration in the silent or idle state is

$$t_{off} = \left(\frac{1}{C}\right) * C_{off},$$

Empirical studies [17] demonstrate that $t_{on} = 0.96$ s. and $t_{off} = 1.69$ s. and we use these values in the simulation, although we have used other values to analyze its effect over RAP. We now report results from the simulation of the RAP protocol. This section shows several scenarios, which we have used in the simulation to analyse the performance of the RAP. Note that we have varied some of these parameters to analyse the behaviour of the RAP when it changes the scenario and the source traffic descriptors as we show in this section.

4.2 Basic scenario: point-to-point connections

Fig. 3 shows a basic network configuration consisting of 3 active ATM switches. ***Fig. 3*** illustrates the flow of PDU, and we can see the DMRP, requests of retransmissions and the source traffic descriptors and other parameters.
The next graphics show results obtained in different simulations with several sources.
Fig. 4 shows the results of varying PCR between 60 and 2,000 cells/s. As we can see, when the arrival rate is low, the number of retrieved PDUs increases. We can see now the number of NACKs not sent (not retransmitted PDUs) is greater when the PCR value increases. In this way, the network is not over-charged with useless retransmissions. In this simulation we fixed the data source that transmits 750 Kbytes; congestion probability = 10^{-3} and CET= 2(1/PCR) ms. For PCR=167 cells/s.; CET=2(1/167); t_{on}=0.96 s.; and t_{off}=1.69 s. and total PDUs discarded by congestion 56; retrieved PDU via RAP 11, and 17 PDUs are not requested.

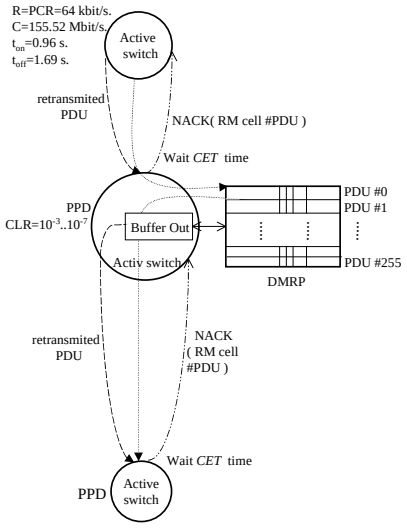

Fig 3 Simulation network for trivial case

When PCR reaches 60 cells/s. the performance is optimized (27 retrieved PDUs out of 28) since all the PDUs loss are retrieved and there are no DMRP failures (all the requested PDUs are in the DMRP).

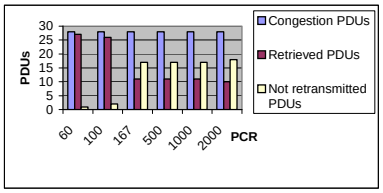

Fig. 4 Effect of PCR variation

Fig. 5 shows the effect of changing the *CET* time in the performance of protocol. As we can see we vary *CET* multiplying (1/PCR) by values from 1 to 4.

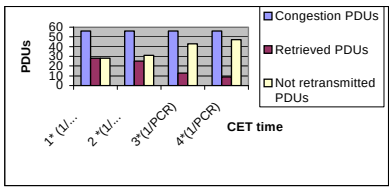

Fig. 5 Effect of CET time

The graph shows how the time (1/PCR) is the best and when we used major values the number of retrieved PDUs is less and the throughput is worse. In this case we transmit 500 PDUs with a PCR of 500 c/s and Congestion probability=10^{-3}. The graph shows how the throughput is

optimized when the delay time is well chosen to avoid the retransmissions and the DMRP hit failures.
Another scenario consists of *1* source node, *1* active ATM switch, *n* non-active switches and *1* destination node. When a NACK arrives at non-active switch, this also transfers the RM cell to the next switch. When the RM arrives at the active switch this uses the DMRP to retransmit the requested PDU. This scenario is the same as above, only the number of non-active switches varies. In this configuration we have simulated the protocol with several non-active switches and the results obtained show no changes. Only the delay in transmissions varies due to propagation times, but the index of retrieved PDUs is maintained as we have already shown.

4.3 VPN with active and non-active switches: point-to-multipoint connections

Fig. 6 presents a point-multipoint configuration consisting of *1* source node, *1* active ATM switch, *n* non-active switches and *n* destination nodes. This is equal to the above basic scenario, only the number of destination nodes in multipoint connections varies. We are currently working to achieve multipoint connections to RAP. If we consider the above results we can see intuitively that the total delay will change. Also the amount of DMRP memory required increases in active switches to manage the VPI/VCI of n connections. ***Fig. 6*** illustrates an example with 3 receivers (D1, D2 and D3) where D1 and D3 request a retransmission that is sent to the active node through the n non-active switches. We stress that there are non-negative effects of a mixed network of active and passive switch nodes, since RAP is a standard and native ATM protocol.

5 Works in progress

The above sections have presented and demonstrated the good behaviour of RAP protocol. However, we shall now describe several aspects on which we are working to achieve better *goodput*. Firstly, we will consider other source traffic descriptors such as SCR (Sustainable Cell Rate) and MBS (Maximum Burst Size). With these parameters we can characterize the traffic better. We are also working to simulate multipoint-to-multipoint connections to analyse the behaviour of RAP in all types of scenarios. We want to reinforce the reliability using hybrid techniques between ARQ and FEC when the sources are suitable. This is a new active characteristic of RAP that allows us to choose the desired reliability. Also we are implementing other congestion control schemes as Early Packet Discard and Early Selective Packet Discard [15,16].

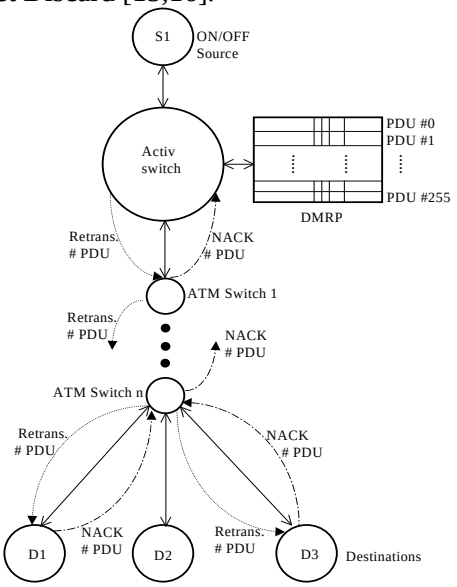

Fig. 6 VPN with point-to-multipoint transfers

6 Summary

We can consider RAP as an active protocol that can take advantage of suitably equipped active ATM switches and situated in an active network. RAP manages a set of privileged VCI to improve reliability. To achieve these active characteristics we use an active ATM switch with DMRP, a dynamic memory that temporarily stores PDUs of each privileged VCI. We demonstrate that it is possible to retrieve an important number of PDUs only with DMRP and a reasonable additional complexity of the active switches. The retransmission mechanism is based on ARQ with NACK that generates RM cells to request PDUs. Our simulations demonstrate that the intuitive idea of taking advantage of silent states in ON/OFF sources is true. Thus we can achieve better performance and QoS in ATM networks.

7 References

[1] R. Steinmetz, and L. C. Wolf, “Quality of Service: Where are We ?,” *Fifth*

International Workshop on Quality of Service IWQOS'97, pp.211-221, 1997.

[2] Martin de Prycker, "Asynchronous Transfer Mode. Solution for Broadband ISDN (3ª Ed.)," *Ed. Prentice Hall*, 1995.

[3] Recommendation I.361, "B-ISDN ATM Layer Specification," *ITU-T*, 11/1995.

[4] Recommendation I.363.1, "B-ISDN ATM Adaptation Layer Specification, Type 1," *ITU-T*, 08/1996.

[5] Recommendation I.363.5, "B-ISDN ATM Adaptation Layer Specification, Type 5," *ITU-T*, 08/1996.

[6] Recommendation I.371.1, "Traffic Control and Congestion Control in B-ISDN," *ITU-T*, 06/1997.

[7] Stephan Block, Ken Chen, Philippe Godlewski, and A. Serhrouchni, "Design and Implementation of a Transport Layer Protocol for Reliable Multicast Communication," *Université Paris, http://www.enst.fr/~block/srmtp*, 1998.

[8] Jörg Nonnenmacher, M. Lacher, M. Jung, E. Biersack, and G. Carle, "How bad is Reliable Multicast without Local Recovery?," *INFOCOM'98., Procs. IEEE*, Vol. 3 pp. 972-979, 1998.

[9] Dan Rubenstein, Sneha Kasera, Don Towsley, and J. Kurose, "Improving Reliable Multicast Using Active Parity Encoding Services (APES)*", Technical Report 98-79, Department of Computer Science, Un. of Massachusetts*, July, 1998.

[10] Luigi Rizzo, "On the feasibility of software FEC," *http://www.iet.unipi.it/~luigi*, Universita di Pisa, 1997.

[11] Ernst W. Biersack, "Performance Evaluation of Forward Error Correction in an ATM Environment," *IEEE Journal on Selected Areas in Comm.*, vol. 11, No. 4, pp. 631-640, May. 1993.

[12] H. Esaki, G. Carle, T. Dwight, A. Guha, K. Tsunoda, and K. Kanai, "Proposal for Specification of FEC-SSCS for AAL Type 5*," Contrib. ATMF/95-0326 R2, ATM Forum Technical Committee*, Oct. 1995.

[13] G. Armitage and K. Adams, "Packet Reassembly during Cell Loss," *IEEE Networks*, vol. 7, no 5, pp. 26-34, S. 1993.

[14] Maurizio Casoni and Jonathan S. Turner, "On the Performance of Early Packet Discard," *IEEE Journal on Selected Areas in Communications*, Vol. 15, no 5, pp. 892-902, Jun. 1997.

[15] Kangsik Cheon and Shivendra S. Panwar, "Early Selective Packet Discard for Alternating Resource Access of TCP over ATM-UBR," *IEEE*, pp. 306-316, 1997.

[16] Omar Elluomi and Hossam Afifi, "RED Algorithm in ATM Networks," *IEEE* pp. 312-319, 1997.

[17] J. M. Pitss, "Introduction to ATM. Design and Performance," *Ed. Wiley*, 1997.

[18] A.L. Roginsky, L.A. Tomek and K.J. Christensen, " Analysis of ATM cell loss for systems with on/off traffic sources," *IEE Proc. Commun.* Vol. 144. No. 3, pp. 129-134, June 1997.

[19] David L. Tennenhouse, Jonathan M. Smith, W. D. Sincoskie, D.J. Wetherall, and G. J. Minden, "A Survey of Active Network Research," *IEEE Communic. Magazine*, pp.80-86, Jan. 1997.

[20] David A. Halls and Sean G. Rooney, "Controlling the Tempest: Adaptive Management in Advanced ATM Control Architecture," *IEEE Journal on Selected Areas in Communications*, Vol. 16, Nº 3, pp. 414-423, April 1998.

[21] J. Chen, Y. Yu, and S. Lin, "Exploiting Multi-Agent Scheme for Traffic Management in ATM Networks," *Intelligent Control (ISIC), Proceedings, IEEE*. pp. 594-599, September 1998.

[22] J. Hardwicke and R. Davison, "Software Agents for ATM Performance Management*," NOMS 98, proceedings, IEEE*, Volume 2, pp. 313-321, 1998.

[23] German Goldszmidt, and Yechiam Yemini, "Delegated Agents for Network Management," *IEEE Communic Magazine*, Vol. 36 3, pp. 66-70, March 1998.

[24] F. Nait-Abdesselam, N. Agoulmine, and A. Kasiolas, "Agents Based Approach for QoS Adaptation In Distributed Multimedia Applications over ATM Networks*," IEEE International Conference on ATM, ICATM-98*, pp. 319-326, 1998.