Shared Medium Based Contention Resolution in ATM Switches to Cope with Bursty Data Traffic

Andreas Kirstädter, TU München, Germany
andreas@lkn.e–technik.tu–muenchen.de

Summary:

This paper first identifies the need for input buffering and efficient contention resolution in future ATM switches that have to handle an increasing share of bursty and asymmetric data traffic. Then different types of existing contention resolution mechanisms are reviewed and a new architecture is presented based on modified LAN medium access protocols. Extensions for the integration of multicast and prioritized traffic are given together with a discussion of the scalability of this approach. Finally the performance of an adapted version of the CRMA–II MAC is compared to that of another contention resolution approach and an output buffered switch.

1 Introduction: Data Traffic in ATM Networks and Its Influence on Switch Architectures

In the first years of its life cycle ATM has primarily been regarded as a bandwidth–effective replacement of existing synchronous multiplexing hierarchies and as a low–level platform for the integration of different traffic types. But not only the recent standardization efforts for available bitrate (ABR) services show the importance of data traffic for ATM. It becomes increasingly clear that ATM will play a key role in the process of replacing and connecting the so called legacy LANs (802.x, FDDI).

This large share of data traffic puts completely new requirements on the ATM switching technology. In contrast to the first generation ATM switching architectures that primarily focused on methods for reducing the connectivity expenses within the switch core (s. [1] for an overview) now the effective handling of large cell buffers becomes the key question. Huge cell bursts will have to be managed by the switches during the reaction gaps of ABR controlled end systems (especially at large bandwidth–delay products). Also quite low cell loss guaranties have to be met in order not to disturb higher layer data communication protocols [2]. These problems are further complicated by the highly asymmetric load scenarios resulting from the popular client server applications. At the same time future ATM switches will be confronted with the task of video distribution and the handling different service classes so that multiple levels of priority and multicast traffic will have to be supported.

While it was shown that the best throughput delay performance can be achieved by pure output buffering [3] this strategy is simply not applicable to large numbers of switch ports and asymmetric load scenarios [4]: in this case a relatively small speed–up of the output buffers (in the order of 4 to 8, (s. e.g. [5]) of the input bit rate, that was found to be sufficient in the case of symmetric load scenarios with uncorrelated sources, no longer results in sufficiently low levels of cell loss so that the loss now becomes highly dependent on the applied load pattern. Shared memory buffering used by most current generation ATM switches [6] suffers not only from the fact that the necessary speed of the common buffer limits the maximum number of ports per switching module. The size of the common buffer is also severely limited since it has to be implemented on the (full custom) switch ASIC.

Input buffers can easily be implemented with standard D–RAMs up to arbitrary sizes (since they only have to work at line speed) but have to be coordinated: i.e. during each T_{cell} it has to be decided which of the contending input buffers is allowed to send a cell to a certain output. The main requirements for the contention resolution mechanism are:

- fair arbitration between contending inputs;
- waste–free operation (the max. switch throughput has to be impaired as less as possible);
- mechanisms for handling multicast and prioritized traffic;
- low implementation costs;
- independence of the special implementation of the switch core;
- scalability.

The next section of this paper reviews the principles of known contention resolution approaches before in section 3 the idea of using medium access control (MAC) protocols from the LAN domain for contention resolution is shown. The main part of this paper then discusses the usage of an adapted version of the high speed MAC protocol CRMA–II, extensions for the management of different traffic categories, and scalability aspects. In section 4 the absolute fairness even under highly asymmetric loads is demonstrated and the throughput–delay performance is compared to one of the methods of section 2 and to that of an ideal (output buffered) switch.

2 Contention Resolution Approaches for Input Buffered ATM Switches

Several approaches have been proposed for this task in the past. They can be categorized into centralized and decentral methods (Figure 1 shows their generic architectures):

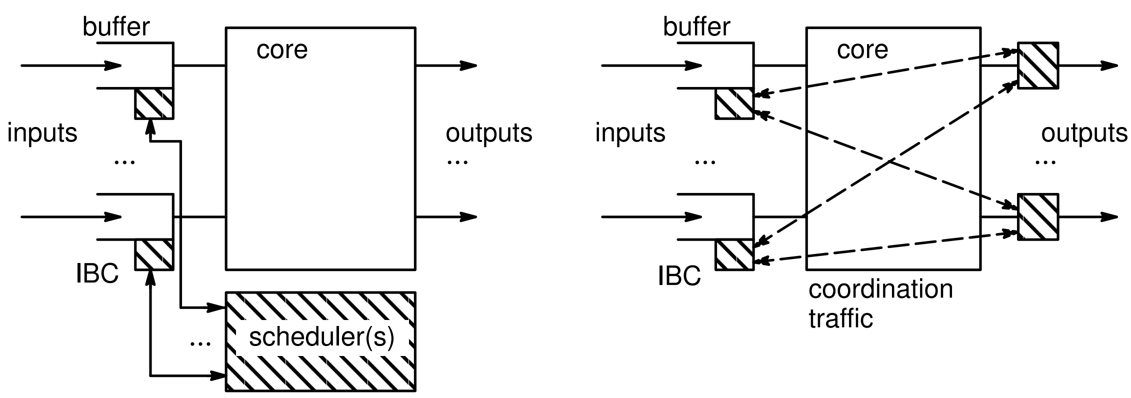

Figure 1: Generic architectures for contention resolution.

2.1 Usage of a Central Scheduler

During each T_{cell} the central scheduling engine (s. e.g. [7] and [8]) first receives information about all head of line (HOL) cells at all input buffers because otherwise its decision would be based on outdated information. Then it uses certain algorithms (e.g. round robin or neural network based etc.) to determine the winners between contending input buffers. This information is then transferred back to the input buffer control (IBC) units that then at the end will send the corresponding cells into the switch core. Thus during the same T_{cell} a large amount of information has to be transferred and processed.

The scalability of this architecture is further limited if the input buffers are split into sub–buffers (one per switch output at each input, s. e.g. [9]) in order to prevent the HOL blocking effect present in simple FIFO input buffering solutions. Then the number of cells that have to be considered by the central scheduler rises with $o(N^2)$, N being the switch size.

2.2 Decentral Approaches

2.2.1 Output Scheduling

The processing speed requirements can be reduced by introducing some parallelism: the central scheduler is divided internally into smaller decision units each of them coordinating a single switch output (s. e.g. [10]). An input having received a cell to a certain output sends a request message to the corresponding output scheduler. The output scheduler reserves a future T_{cell} as transmission time instance (as far as available) for that cell and informs the input about it. The input then will send the cell during the assigned time slot into the switch core if it has not already another cell scheduled for the same time slot. In the other case the sending opportunity would have to be dropped (leading to a large waste of possible throughput) since an input can only send a single cell during each T_{cell}. A solution of this conflict either requires additional request–acknowledge rounds if the IBC returns the unusable assignment to the output scheduler for ”recycling”, [10]. Or the IBCs have to include a bitmap showing the assignment state of a number of future T_{cell} in their request message (as in the case of the Extended Output Scheduling, EOS [11]). Both alternatives strongly increase the coordination traffic between the IBCs and the output schedulers if large switches are considered. Thus in a real systems an extra ”switch” (with its own contention resolution?) will be necessary for the transmission of the contention resolution information [11].

2.2.2 Two– / Three–Phase–Algorithms

Another family of coordination mechanisms for input buffers are the so called 2– or 3–phase–algorithms (s. e.g. [12]). But since they can only be used together with a special kind of switch core (Banyan network with a bidirectionally transmitting Batcher sorter) they are not further considered in this paper.

2.2.3 Rate Based Contention Resolution

A fully distributed approach is the dynamic bandwidth allocation (DBA, [13]) that tries to reduce the hardware expenses for contention resolution by allocating cell transmission rates instead of transmission instants for single cells. The communication between the IBCs and the output rate schedulers (placed at the core outputs) is done via regular ATM cells.

But since only cell rates are coordinated the switch core must offer high speed buffering capabilities (nearly as an output buffered core). Furthermore a low efficiency of this approach is to be expected since either a large amount of payload cells are used for the transport of coordination messages or the coordination itself will be rather slow and bandwidth wasting in the case of highly bursty traffic. These problems increase with the number of ports and if multicasts and priorities are to be considered.

2.2.4 Ring Reservation

The method proposed in [14] prevents output port conflicts by clocking a number of bit positions around a ring that connects all IBCs (figure 2). Each of the bits controls the sending of cells to a certain output (corresponding to its index within the bit frame). At the start of T_{cell} all bits have been set to ”0” and an input can reserve for its HOL cell a sending opportunity for the next T_{cell} by successfully setting (from ”0” to ”1”) the corresponding bit.

A strong reduction of the necessary coordination traffic and the corresponding hardware expenses arises from this solution since for the contention resolution only a neighbor–neighbor communication in one direction between the IBCs has to be established. Splitting the input buffers internally into sub–buffers to avoid HOL blocking does not increase the necessary flow of coordination information. Neither a central controller is required nor are the

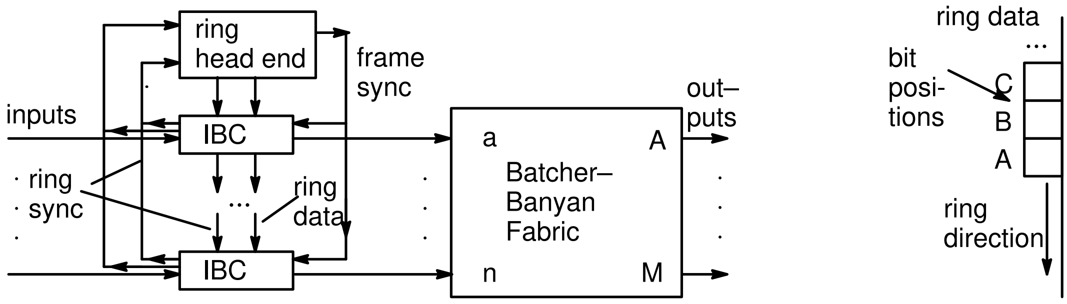

Figure 2: Ring reservation: architecture and ring data structure (control bits).

core outputs involved in this communication. Further this mechanism is completely independent of the design of the core.

But so far the fairness problem arising from the sequential probing of inputs has not been solved successfully. The upstream one of two neighboring IBCs stays favored even if (as proposed in [14]) the virtual origin of the bit frame is rotated (by a single IBC per T_{cell}) around the ring. Also up to now no satisfying native solution has been presented for the coordination of multicast and prioritized connections. Multicast cells have to be duplicated before entering the switch core (e.g. [15]) thus severely limiting the amount of multicast the switch can handle. Priorities can only be processed by rotating the bit frame a corresponding number of times per T_{cell} around the ring (each time coordinating a single priority level) thus inadequately increasing the ring bandwidth requirements.

3 Shared Medium Based Contention Resolution

3.1 Principle

A new approach has been introduced in [16] that first reduces the promising idea of ring reservation to its main principle: the authorization for sending a cell into the core is tied to the successful setting of a certain bit position on a serial line. The fairness between contending inputs is then guaranteed by adapted medium access control (MAC) protocols from the LAN domain. Especially designed for distributed fairness control they are ideally suited for this task in the above architecture. The contention resolution is now split into three layers:

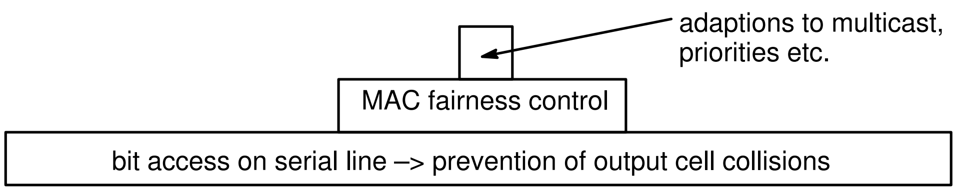

Figure 3: Layers of shared medium based contention resolution.

Most of the coordination is now already done by the basic bit access. Fairness control and the adaption to different traffic categories (s. below section 3.3) can thus operate at much lower frequencies leading to economic feasible implementations even at very large switch sizes.

The coordination part of the resulting LAN–switch hybrid system can be viewed as M virtual LANs (VLANs) existing on the same serial medium (the bit positions corresponding to the M controlled outputs) that comprise N stations (switch inputs). Each input buffer controller (IBC) then contains M virtual finite state machines (FSMs) each of them controlling the transmission of cells of a single sub–buffer to its corresponding switch output.

The requirements a MAC protocol has to fulfill (e.g. slotted transmission structure, low processing overhead, bandwidth consumption etc.) in order to be suited for this kind of coordination have been investigated in [16]. It has been shown that they can be met only by MAC protocols originally designed for high–speed bus LANs and only by those of them that are able to react properly to the fact that the single VLANs cannot operate fully independent from each other: the hardware implementations of the switch core and the input buffers severely limit the number of cells an input buffer can send into the switch core during a single T_{cell}. So called coincidence situations arise when an IBC is able to occupy more bit positions within a single T_{cell} than the maximum number of cells it can emit during the same T_{cell}. The distributed queues in the case of the well known DQDB MAC protocol, for example, get heavily disturbed by this.

3.2 Cyclic Reservation Multiple Access (CRMA–II)

The above mentioned investigations led to the usage of the fairness mechanism of CRMA–II [17]. Implemented on a dual bus topology a CRMA–II network normally consists of a fairness scheduler (located in one of the two bus headends) and two counterdirectional buses. Fairness is controlled cyclically: after the reception of a request frame that contains data about the fairness situation of the single stations the scheduler compares this information with a calculated throughput threshold and notifies the stations about the results of this comparison by emitting a confirm frame. The stations then restore the fairness in a distributed manner: if the transmission count of a station was above the threshold it defers a corresponding number of times the access to empty slots. Otherwise it is allowed to access a number of reserved slots issued by the scheduler. Finally the scheduler starts a new fairness cycle by emitting the next request frame.

In order to apply this cyclic fairness mechanism to the coordination of the IBCs the following "infrastructure" has to be provided between the IBCs (s. figure 4):

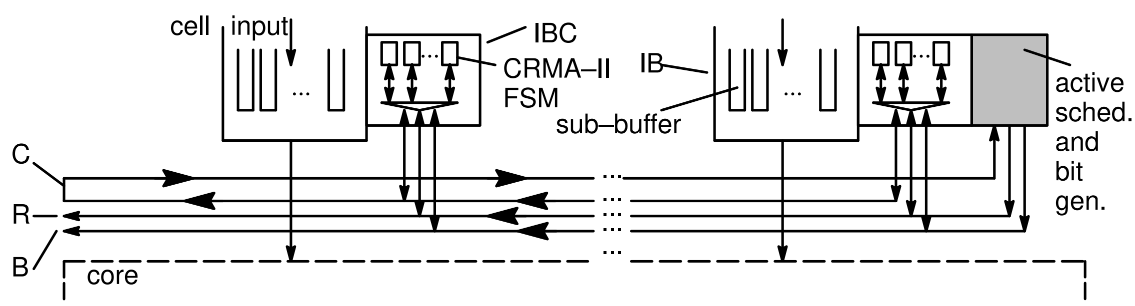

Figure 4: Architecture for the coordination of IBs by modif. CRMA–II.

- The scheduler entity within the first IBC is activated at the origin of the serial line B that transports the bit positions (that control the access to the outputs).
- A second bit on line R shows the reservation state of the corresponding bit position on B.
- Finally a line C is necessary for the transport of fairness information (transmission counters and buffer length values) between the IBCs and the scheduler. Line C is looped back to the scheduler after the last IBC to get the same transmission direction as B and R.

The duration T_{cycle} of the scheduling cycles of the single VLANs is mainly determined by the time T_t necessary for the transmission of the fairness information from the IBCs to the scheduler. At a switch size of 64x64 ports a T_t below 2* T_{cell} per VLAN has been found to be sufficient leading to a T_{cycle} in the order of 128*T_{cell} (if a single line is used for C).

Also the CRMA–II protocol itself (see [18] for details of the original protocol and [16] for a detailed discussion of the modifications) had to be adapted:

- The transmission counters are now reset each cycle after notifying the scheduler about their values.
- The deferment of stations with excess throughput is no longer used and the scheduling threshold is always set equal to the largest transmission counter value.
- At the beginning of the marking period each station pre–subtracts the number of calculated confirms from its transmission counter since they serve as the equalization of unfair throughput situations during the last cycle.

3.3 Extensions for the Management of Different Traffic Categories

Groups of differently prioritized inputs can easily be obtained by splitting the lines R and C between the groups (and activating the scheduler entities in the IBCs beneath each of those interruptions) in figure 4. Now each group is controlled by its separate fairness algorithm. Since the line B still connects all IBCs the members of downstream groups can only access the bit positions left unused by upstream groups: each group represents a different priority class of inputs.

If the IBCs of the lowest priority class are used for the input links into the switch the higher prioritized inputs may operate as servers for multicast cells and cells of different priority levels. These cells are now looped back to their corresponding servers closer to the origin of the bit stream (figure 5).

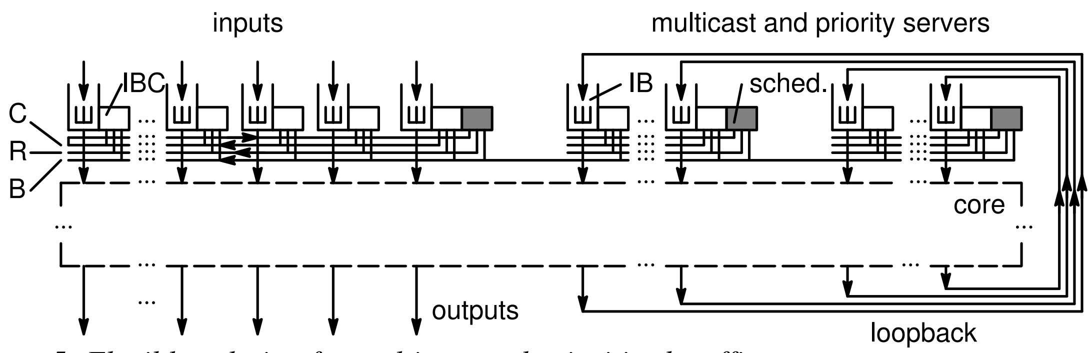

Figure 5: Flexible solution for multicast and prioritized traffic.

Among the advantages of this concept is the possibility to flexibly re–partition the server groups by simply splitting / reconnecting the lines R and C (and activating resp. deactivating the scheduler entities) even during the operation of the system [16].

3.4 Hardware Structure and Scalability

Only the basic bit access has to be implemented with high–speed logic. Just a few gates are sufficient for the setting of bits, the recognition of bit states, and the reaction to the above mentioned coincidence situations. The MAC fairness control itself operates only on the cell level so that area and power optimized standard cell design can be used nearly throughout the whole IBC.

The maximum number of outputs that can be coordinated by this architecture (and thus its total throughput TP_{max}) is mainly determined by the number H of bit positions that can be processed by all IBCs during a single T_{cell}. For large switches (above 100 input ports of 150 Mbit/s each) parallel bundles of lines can be used for the transport of the bits on lines B, R, and C. Most of the high–speed bit processing is then done in parallel. Only the reaction to coincidences reaches across all B lines. Since this high–speed logic part requires just a few gates H can reach the order of several hundred at input link speeds of 2.4 Gbit/s with today's bipolar technologies.

4 Quantitative Comparisons

In time discrete simulations at the cell level first the fairness of the adapted CRMA–II mechanism of section 3.2 has been investigated. Figure 6 shows the fairness of the throughput

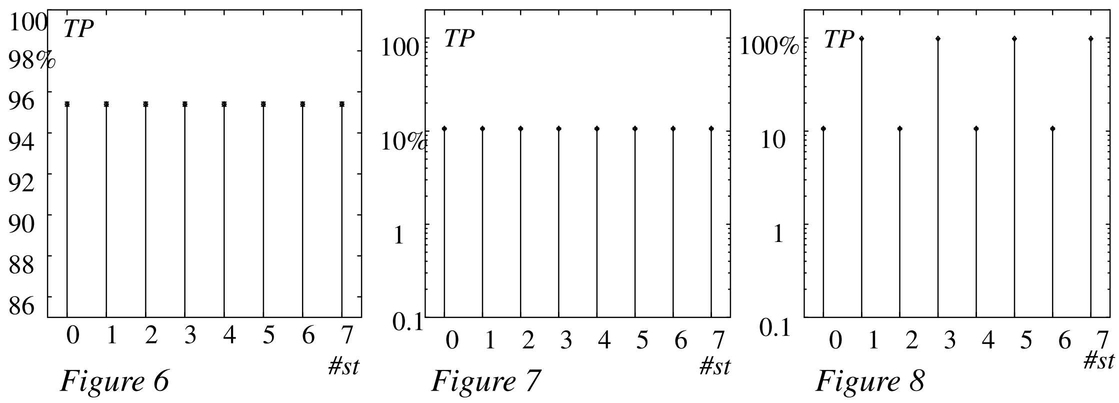

Figure 6 *Figure 7* *Figure 8*

in the case of symmetric overload (95% confidence intervals in all plots 6 to 9): the input links carry a load of 100% with equally distributed output addresses of the cells. It can also be seen that the max. achievable throughput gets very close to 100%. In the figures 7 and 8 the same model has been exercised with extremely asymmetric traffic: while inputs with an odd index get the same load as before in figure 6, the cells on even inputs are directed only to the output with index 0 (the ”hot–spot”). Figure 7 demonstrates the absolute fair sharing of this hot–spot output by the single inputs. Figure 8 shows the total throughput (as a sum) each input gets in this situation: the interesting fact is that the throughput values of the odd inputs are only neglectably impaired by the reactions of the fairness mechanism (the values get very close to 100%).

Figure 9 compares (for a 64x64 switch with a link bit rate of 150 Mbit/s) the throughput

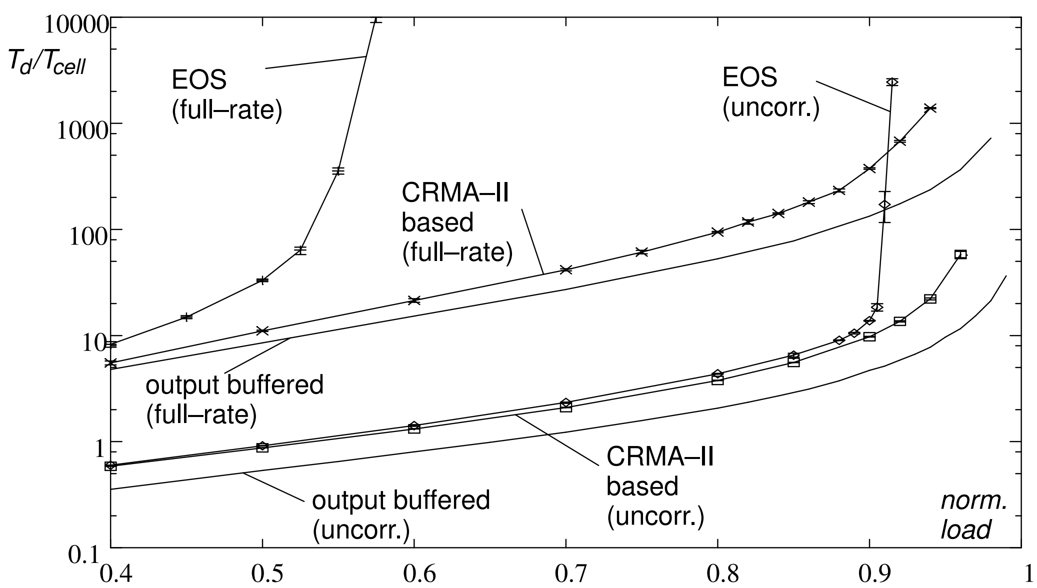

Figure 9: Different contention resolution mechanisms.

delay performance of the CRMA–II MAC coordination principle to that of EOS (with a bitmap size of 32 bit) and an ideal (output buffered) switch. Two different types of cell sources have been used for the simulation: an uncorrelated type (where the destination of each cell is independent from the destination of previous cells) and an on–off type with full–rate bursts. In the latter case the mean burst size was 16 cells and the medium cell rate was adjusted by choosing the off–time accordingly.

The comparison of the throughput delay behaviour shows that the performance of the MAC coordinated switch gets very close to that of the ideal switch. Also the maximum throughput is not impaired by bursty traffic as in the case of EOS (where the size of the bitmap restricts the range of operation).

5 Conclusions

A new approach for the resolution of output port contentions in input buffered switches has been presented that is based on the usage of MAC protocols of high–speed LANs. The requirements for the applied MAC protocol have been discussed and it has been shown that a modified version of CRMA–II provides a very high throughput and a complete fairness even under very asymmetric load scenarios. The proposed architecture can be easily extended to allow an arbitrary number of priority levels together with an effective coordination of multicast cells. Thus very large switches with throughputs of several Terabits per second can now be efficiently coordinated on the cell level.

The author wishes to thank Mr. Achim Autenrieth for his simulations of the EOS performance.

References:

[1] Ahmadi, H. and Denzel, W.: A Survey of Modern High–Performance Switching Techniques, IEEE Journal on Selected Areas in Comm., Vol. 7, No. 7, pp. 1091 – 1103

[2] ATM Forum: Reliability and Performance Considerations for ABR and VBR+ to support LAN Applications, Traffic Management Subworking Group, 94–0262

[3] Hluchyj, M. G. and Karol, M. J.: Queueing in High–Performance Packet Switching, IEEE Journal on Selected Areas in Communications Vol. 6 No. 9, pp. 1587 – 1597

[4] Simcoe, R. J. and Pei, T.–B.: Perspectives on ATM Switch Architecture and the Influence of Traffic Pattern Assumptions on Switch Design, Comp. Comm. Rv., pp. 93 – 105

[5] Karol, M. J. et al.: Input Versus Output Queueing on a Space–Division Packet Switch, IEEE Transactions on Communications Vol. 35

[6] Garcia–Haro, J. and Jajszczyk, A.: ATM Shared–Memory Switching Architectures, IEEE Network July / August 1994, pp. 18 – 26

[7] Obara, H.: "Optimum Architecture for Input Queueing ATM Switches, Electronics Letters, 28th March 1991, pp. 555 – 557

[8] Matsunaga, H. and Uematsu, H.: A 1.5 Gb/s 8x8 Cross–Connect Switch Using a Time Reservation Algorithm, IEEE Journal on Selected Areas in Comm., pp. 1308 – 1317

[9] LaMaire R. O. and Serpanos, D. N.: Two–Dimensional Round–Robin Schedulers for Packet Switches with Multiple Input Queues, IEEE/ACM Trans. on Netw., pp. 471 – 482

[10] Karol, M. J. et al.: Improving the Performance of Input–Queued ATM Packet Switches, IEEE INFOCOM 1992, pp. 110 – 115

[11] Main, J. and Sarkies, K. (1995): Cell Scheduling Using Status Arrays in Input Buffered ATM Switches, Proc. of the 1. IEEE Worksh. on Broadb. Switching Syst., Poznan, Polen

[12] Hui, J. Y. and Arthurs, E. (1987): A Broadband Packet Switch for Integrated Transport, IEEE Journal on Selected Areas in Communications Vol. 5 No. 8, pp. 1264 – 1273

[13] Worster, T. et al. (1995): Buffering and Flow Control for Statistical Multiplexing in an ATM Switch, Proceeding of the ISS 95, Berlin, Germany, April 1995, Paper no. 488

[14] Bingham, B. and Bussey, H. (1988): Reservation–Based Contention Resolution Mechanism For Batcher–Banyan Packet Switches, Electron. Letters, 23 June 88, pp. 772 – 773

[15] Xerox Corp. (1993): A Switching Network, European Patent Application, Publication Number: 0 571 166 A2, 24 Nov 1993.

[16] Kirstädter, A.: Fair and Flexible Contention Resolution for Input Buffered ATM Switches Based on LAN Medium Access Control Protocols, Proceedings of 1996 Int. IFIP IEEE Conference on Broadband Communications, Montreal

[17] van As, H. et al. (1991): CRMA–II: A Gbit/s MAC Protocol for Ring and Bus Networks with Immediate Access Capability, Proc. of EFOC/LAN 91, London, June 19–21, 1991

[18] van As, H. and Lemppenau, W. W. (1992): Performance of CRMA–II: A Reservation–Based Fair Media Access Protocol for Gbit/s LANs and MANs, Proceedings of EFOC/LAN 92, Paris, June 22–24, 1992