Switching Fabrics with Internal Backpressure using the ATLAS I Single-Chip ATM Switch

Manolis Katevenis†, Dimitrios Serpanos†, and Emmanuel Spyridakis†
Institute of Computer Science (ICS)
Foundation for Research and Technology – Hellas (FORTH)
Science and Technology Park of Crete, Vassilika Vouton,
P.O.Box 1385, Heraklion, Crete, GR 711 10 Greece
E-mail: katevenis@ics.forth.gr Tel.: +30 (81) 391664 Fax: +30 (81) 391661

Proceedings of the GLOBECOM'97 Conference, Phoenix, Arizona, USA, November 1997
ftp://ftp.ics.forth.gr/tech-reports/1997/1997.GLOBECOM.ATLAS_I_Fabrics.ps.gz

ABSTRACT: ATLAS I is a single-chip ATM switch with optional credit-based (backpressure) flow control. This 4-million-transistor 0.35-micron CMOS chip, which is currently under development, offers 20 Gbit/s aggregate I/O throughput, sub-microsecond cut-through latency, 256-cell shared buffer containing multiple logical output queues, priorities, multicasting, and load monitoring. This paper discusses the use of backpressure inside networks based on ATLAS I chips: in switching fabrics of large ATM switches, or in wormhole-style workstation cluster LANs. We explain and we show by simulation that the ATLAS I backpressure provides a switching fabric with high performance, comparable to an output queued switch, at low cost, comparable to an input buffered switch.

1. Introduction

High performance networks are made using point-to-point links and switches. The building block switches usually have relatively small fan-in/fan-out and crossbar internal organization. We are developing such a single-chip building block ATM switch, called *ATLAS I* (ATm multiLAne backpressure Switch One). Large switches, with many links, are made of small switches interconnected in *switching fabrics.*

† Institute of Computer Science, FORTH, and Dept. of Computer Science, University of Crete, Heraklion, Crete Greece.

This paper deals with flow control in the core of switching fabrics made of ATLAS I chips; the notion of fabric is broad enough to also cover entire ATLAS I subnetworks. Credit-based (backpressure) flow control is well suited for these core parts of networks, because of its accuracy (cells are never dropped), its relative ease of implementation (easier than hardware retransmission), the existing implementation experience (from wormhole routers), and the attractive properties that it offers, as discussed in this paper. ATLAS I implements optional credit-based flow control (multilane backpressure) in hardware.

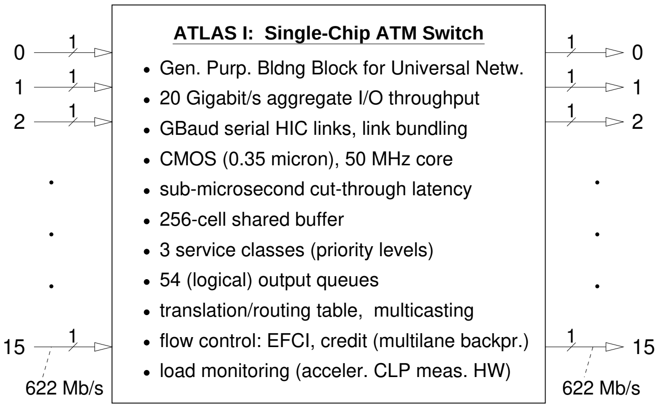

Figure 1: ATLAS I chip overview

1.1 ATLAS I: a Single-Chip ATM Switch

ATLAS I is a general-purpose, single-chip, gigabit ATM switch with advanced architectural features. It is being developed within the ASICCOM[1] project. ATLAS I is

[1] funded by the European Union *ACTS* (Advanced Communication Technologies and Services) Programme. The ASICCOM Consortium consists of industrial partners (INTRACOM,

intended for use in high-throughput and low-latency systems ranging from wide area (WAN) to local (LAN) and system/desktop (SAN/DAN) area networking, supporting a mixture of services in a range of applications, from telecom to multimedia and multiprocessor NOW (networks of workstations).

Figure 1 presents an overview of ATLAS I. It is a 16×16 switch, with point-to-point serial links running at 622 Mbits/s each. Using link bundling, ATLAS I can also be configured as 8×8 at 1.25 Gbps/link, or 4×4 at 2.5 Gbps/link, etc. The links run ATM on top of IEEE Std. 1355 ‘‘HIC/HS’’ [HIC95] as physical layer, using the BULL ‘‘STRINGS’’ GBaud serial-link transceiver. HIC was preferred over SONET because of simpler circuitry, lower latency, and the capability to encode (unbundled) credits. Internally, ATLAS I operates as a crossbar, with a 256-cell shared buffer. ATLAS I implements three levels of priority, each level having its own queues. Fifty one logical output queues (3 times 16 outputs plus 1 management port) and 3 multicast queues are maintained in the shared buffer. Owing to the all-hardware control and to the virtual cut-through provided by the crossbar, cell latency through the switch chip at low network load is much less than one microsecond. The control section of ATLAS I is shared among all incoming and outgoing links, and consists of two 5-stage pipelines. One pipeline handles cell arrivals and departures, while the second processes credit arrivals; each pipeline operates at the rate of one event per clock cycle (20 ns).

At the time of this writing, most of the switch blocks have been fully designed, at the gate or transistor level, and verification is in progress. The majority of the chip uses semi-custom logic, consisting of approximately 400,000 transistors in logic and 320,000 bits of compiled SRAM (single or dual port). A small part of the chip is laid out in full-custom; it contains 20,000 transistors in logic and 160,000 transistors in multi-port and content-addressable RAM. ATLAS I will be fabricated in a 0.35 micron CMOS process, inside a 225 mm^2 die, by SGS Thomson, Crolles, France. More details on the internal organization of the chip can be found in [KaSV96] and [Korn97].

1.2. The Backpressure Protocol of ATLAS I

Under backpressure (credit-based) flow control, cells are never dropped, because they are only transmitted when buffer space is known to exist at the receiver. Single-lane backpressure indiscriminately starts or stops the transmission of any cell according to buffer space availability; it performs poorly due to head-of-line blocking effects. Multi-lane (per-VC) backpressure fixes this problem; it has been implemented in hardware, in wormhole routers (e.g. Torus [DaSe87], iWarp [Bork90]), and in software, in older networks (e.g. Tymnet, GMDNET, Transpac) and recently in ATM networks [QFC95] [OzSV95] [KuBC94].

Figure 2 illustrates a credit-based flow control protocol that is similar to QFC [QFC95], but is adapted to hardware implementation over short and reliable links; section 4 presents simulation results for this protocol. Credits operate on the granularity of *flow groups,* because it may be undesirable or infeasible for them to operate on individual connections when the number of such connections is too large. A flow group is a set of connections over a common path through the network; their cells never overtake each other. ATLAS I links can carry up to 4096 flow groups, each; the chip can *merge* multiple incoming flow groups into each outgoing flow group.

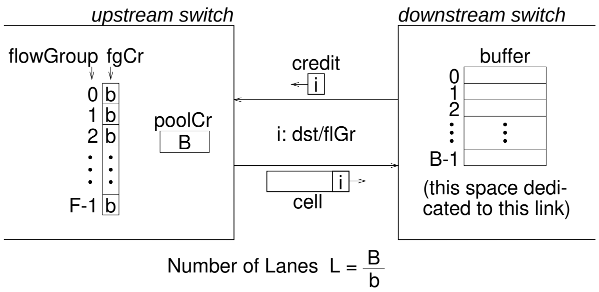

Figure 2: Credit flow control protocol for ATM switch chips (ATLAS I uses b=1, hence L=B)

For each link, the downstream switch allocates a buffer pool of size B. The upstream switch maintains a *pool credit* count per output, corresponding to the available space in that pool. This space is shared among cells according to their flow group. Separate credit counts, $fgCr$, are maintained for each flow group, initialized to b. The pool credit corresponds to QFC's $Limit_{link}$, and $fgCr[i]$ corresponds to QFC's per-connection $Limit[i]$. A cell belonging to flow group i can depart if and only if $fgCr[i] > 0$ and $poolCr > 0$; $fgCr[i]$ and $poolCr$ are then decremented by 1, each. When that cell leaves the down-stream switch, a credit carrying its flow group ID, i, is sent up-stream, where it increments $fgCr[i]$ and $poolCr$. Since the buffer pool contains B cell slots and each flow group is allowed to occupy at most b of these slots, at least $L = B/b$ of the flow groups can share this buffer pool at any given time; thus, L is the *number of lanes.*

In ATLAS I, each credit carries permission for a single cell, and is transmitted upstream as soon as the receiver forwards one cell, unlike QFC where multiple credits are bundled together for transmission. Additionally, in ATLAS I,

Greece; SGS THOMSON, France and Italy; BULL, France), telecom operators (TELENOR, Norway; TELEFONICA, Spain), and research institutes (FORTH, Greece; SINTEF, Norway; Poli. di Milano, Italy; Democritos, Greece).

flow group credits are initialized to $b = 1$, so as to maximize the number of lanes $L = B/b$; also, the hardware implementation is simplified in this way. ATLAS I is fast enough for the cell-credit round-trip-time to be less than one cell-time, for links up to a few meters in length; thus, even a single flow group is able to saturate a link by itself.

1.3 Paper Outline

This paper contributes on the following items: *(i)* we present typical ATLAS I switching fabric configurations (section 2), and explain how backpressure is used internally and how it interfaces to the external flow control mechanisms; *(ii)* we discuss why such configurations offer high performance, comparable to output queueing, at low cost, comparable to input buffering (section 3); *(iii)* we present results of simulations of such switching fabrics (section 4), that demonstrate their good throughput and delay performance, and their tolerance of bursty and hot-spot traffic; the performance of backpressured fabrics was studied in the past in the case of wormhole routing [Dally90], but no similar study has been made for hardware credit protocols (such as ATLAS I) that are appropriate for ATM; and *(iv)* we compare the ATLAS I backpressure protocol to the corresponding wormhole protocol, and show that the former performs better (sec. 4).

2. ATLAS I Switching Fabrics

Figure 3 illustrates ATLAS I chips connected directly to each other and to HIC/HS end systems, for distances up to a few meters. Switch chips are initialized through a serial ROM attached to them, or from their microprocessor port. The network is managed by software running on end system(s) or on microprocessor(s) attached to (some of) the switch chips.

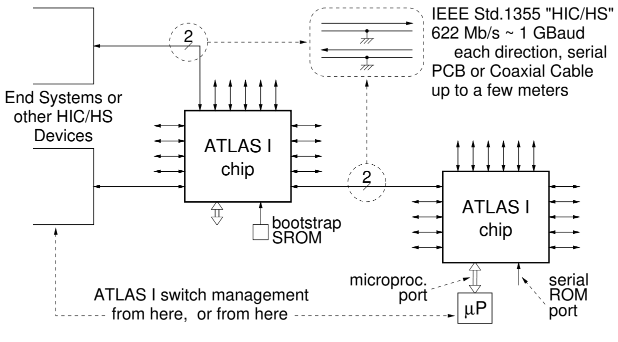

Figure 3: Basic ATLAS I connections

The large number of ATLAS I links and its link bundling capability offer important flexibility to its user. For example, the chip can be used in multiplexer/demultiplexer applications, interfacing about a dozen 622 Mbps systems to a high-speed link, or it can be a building block for high-speed

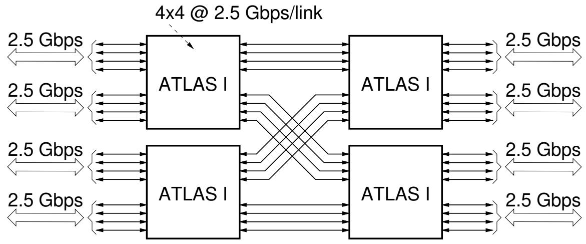

Figure 4: 8×8 fabric at 2.5 Gbps, using link bundling

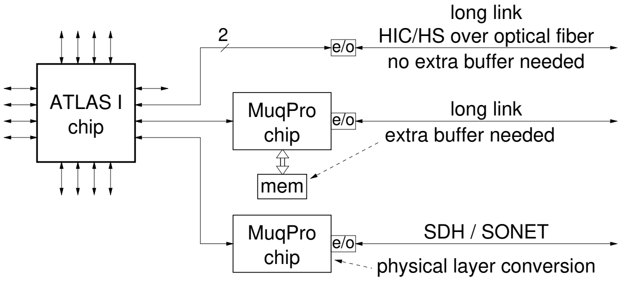

Figure 5: Multi-Queue Processor: interfacing to long links

switching fabrics like the one shown in figure 4.

When ATLAS I chips are to be connected to longer links or to different physical layers, an interface is required, as shown in figure 5; we call it *Multi-Queue Processor,* or *MuqPro* (pronounced "Mac Pro"). By placing this interface outside ATLAS I, its cost is only paid where needed. The implementation of MuqPro is outside the scope of the ASICCOM project, but a subset of its functionality is included in the (multi-chip) demonstration system currently being built within ASICCOM. Besides physical layer conversion, MuqPro provides large buffer space with an advanced queue architecture; for this purpose, external memory chips are attached to it.

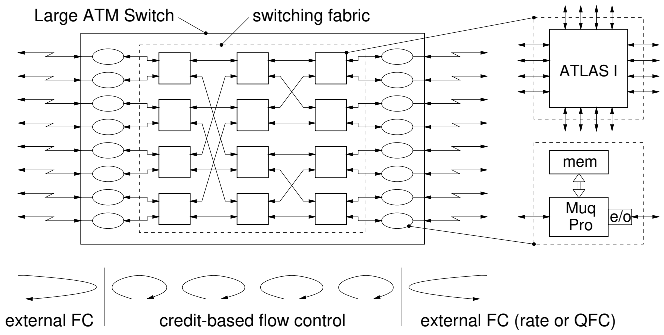

Figure 6: Typical large ATM switch using ATLAS I

Putting all the above together, figure 6 shows how to make a large ATM switch, using a switching fabric of ATLAS I chips and MuqPro interfaces for the external links. The topology of the switching fabric can be arbitrary. As explained below, in this configuration, the backpressure capability of ATLAS I is advantageously used inside the fabric, and MuqPro interfaces this internal backpressure to the external flow control protocol. This configuration is general enough to also cover arbitrary ATLAS I subnetworks, where the MuqPro's constitute either end-system interfaces or interfaces to other networks.

3. Backpressured Switching Fabrics

Figure 8 illustrates the operation of an ATLAS-MuqPro banyan switching fabric like that of figure 6. This is compared to the output queueing architecture. Output queueing, illustrated in figure 7, offers the ideal delay and throughput performance, because cells destined to a given output (e.g. G) are never blocked by cells waiting to go elsewhere (e.g. R). Simple implementations maintain a single FIFO queue per output. For improved fairness, one may maintain separate queues per incoming link in each output buffer, and serve them according to a desired scheduling policy. In the snapshot of figure 7, this would normally ensure short delays for cell entering through inputs 2 and 3 and exiting through output B.

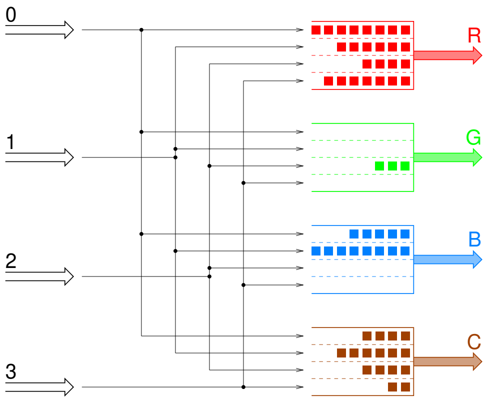

Figure 7: Ideal output queueing: impractical for large fan-in

Output queueing is impractical for large fan-in switches, because of high cost. For such large switches, the configuration of figures 6 and 8 offers comparable performance at reasonable cost. For simplicity, figure 8 shows a unilateral 4×4 fabric made of 2×2 elements. We can see now the effect of the ATLAS I backpressure on the "shape" and location of the queues. In every link, all connections going to a given output port of the fabric form one flow group. Each MuqPro maintains separate logical queues for each flow group. Backpressure effectively *pushes* most of the output queue cells *back* near the input ports. The *head* cells of the output queues, however, are still close to their respective output ports. The fabric operates like an input-buffered switch (a switch with advanced input queueing), where the ATLAS

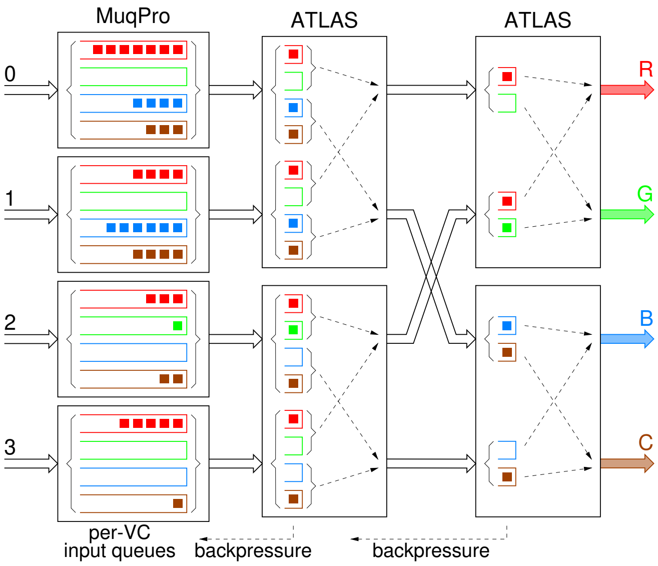

Figure 8: ATLAS/MuqPro fabric corresponding to figure 7

chips implement the arbitration and scheduling function in a distributed, pipelined fashion.

3.1 Cost and Performance

In modern VLSI technology, on-chip logic and memory are relatively inexpensive; chip-to-chip communication, however, is expensive. Off-chip buffer memory in high-speed switches is usually pin-limited, i.e. its cost is determined by its required throughput. Thus, for an $N \times N$ switch, output queueing (figure 7) is very expensive because its total memory throughput is proportional to N^2. In the ATLAS fabric (figure 8), the only external memories are those attached to the MuqPro's, whose total throughput is proportional to $2N$, i.e. the same as in input queueing or input buffering. In order for the cost to be kept at that low level, the building blocks of the switching fabric (i.e. ATLAS I) must *not* use off-chip buffer memory. Since buffer memory in the switching elements has to be restricted to what fits on a single chip, backpressure is the method to keep this small memory from overflowing all the time. Section 4 shows that about a dozen of lanes per link suffice; for short links, ATLAS I is fast enough to provide a cell-credit round-trip-time of less than one cell-time, so buffer space of 1 cell per lane is enough. Hence, the 256-cell buffer space of ATLAS I (for the entire switch) is sufficient.

Performance-wise, the fabric of figure 8 offers properties comparable to those of output queueing. Saturation throughput is close to 1.0 even with few lanes (section 4). No cell loss occurs in the fabric –cells are only dropped when the (large) MuqPro queues fill-up. Traffic to lightly loaded destination ports (e.g. G) is isolated from hot-spot outputs (e.g. R). For example, in the snapshot of figure 8, a cell entering through input 0 and destined to output G encounters only empty queues with available credits in its path. In its trip, such a cell may undergo multiplexing delays due to sharing links with other, heavy-traffic connections, but it never has to wait behind other cells in a queue, so its delay is *not* proportional to the length of other queues. This *hot-spot tolerance* was verified by simulation (section 4). *Fairness* is also provided. ATLAS I chips (e.g. the one feeding output B) service merging flow groups in round-robin fashion. Thus, when cells from inputs 2 and 3, destined to output B, arrive in that rightmost switch chip, they receive a fair share of link B, regardless of the fact that connections 0→B and 1→B are overloaded.

Compared to switching fabrics without internal backpressure, the ATLAS/MuqPro architecture offers lower cell loss than architectures with restricted internal buffers, or lower cost than architectures with large internal buffers. Compared to bufferless fabrics with sorting, trap, concentration, and recirculation stages (e.g. Starlite [HuKn84]) or to fabrics with $O(N^2)$ bufferless switching elements (e.g. Knockout [YeHA87]), the ATLAS/MuqPro architecture uses a much smaller number of significantly more complex switching elements. Given the modern VLSI technology, where on-chip logic and memory costs much less than off-chip communication, we see that the ATLAS architecture is the proper choice.

3.2 Interface to the External Flow Control

In the switching fabric of figures 6 and 8, there is one logical queue per input-output port pair in the MuqPro's. The rate of change of queue j in MuqPro i indicates whether flow group $i \rightarrow j$ is currently above or below its "fair share" of throughput utilization. This is the input required for the rate-based flow control mechanisms [Ohsa95] to operate, providing rate feedback on a per-flow-group basis –not indiscriminately to all connections passing through a link.

4. Backpressured Fabric Simulation

We have simulated the performance of switching fabrics like the one of figure 8 with internal backpressure like that of figure 2 ($b \geq 1$). The full details and results of this evaluation can be found in the Technical Report [KaSS96]; this section summarizes two of the most important results. The simulation was at the clock-cycle granularity, closely tracking the properties of switch hardware. We simulated banyan fabrics of sizes 16, 64, or 256 fan-in/fan-out, made of 2×2, 4×4, or 8×8 switching elements; the results below are for a 64×64 fabric made of 2×2 elements. The input traffic consisted of packets (bursts) of size 10 to 160 cells, with Poisson interarrival time; the results below are for 20-cell bursts. We also simulated the same fabrics operating under the traditional multilane wormhole backpressure protocol [Dally90], assuming that the flit size is equal to the cell payload size (48 bytes). This provides an interesting perspective: the flit size of wormhole routers used to build multiprocessor interconnection networks has grown with time from 1 byte originally to 16 or 32 bytes today; thus, wormhole routers will soon resemble backpressured ATM switch chips like ATLAS I, further stressing the similarity between ATM-interconnected networks (clusters) of workstations (NOW/COW) [ACPN95] and traditional multiprocessor systems.

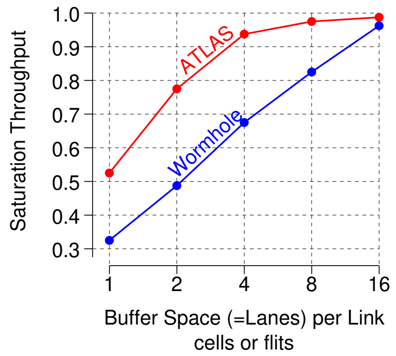

Figure 9: Saturation throughput versus lanes; b = 1

Figure 9 shows the saturation throughput of the fabric, i.e. the output throughput when all MuqPro queues are non-empty. We see that a quite modest buffer space is sufficient for ATLAS to provide a throughput of almost 1.0.

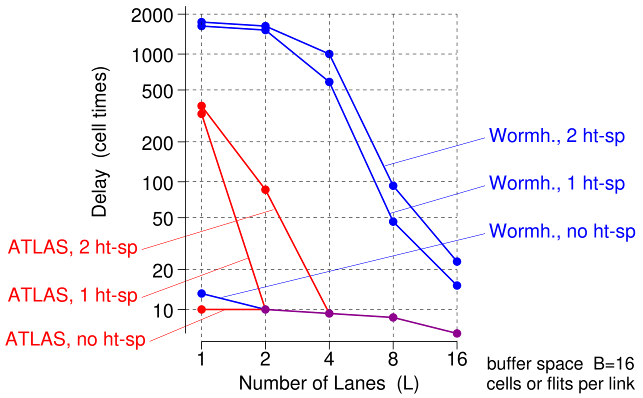

Figure 10: Delay in the presence of hot-spots

Figure 10 plots the delay throughout the fabric for cells or flits targeted to non-hot-spot outputs; here, the incoming load is 0.2, uniformly destined to all outputs, except that there is saturating traffic targeted to 1 or 2 hot-spot outputs (i.e. the corresponding 1 or 2 queues of each MuqPro are always non-empty). The key observation is that ATLAS provides *complete isolation* of well-behaved from hot-spot connections (although they all share the same buffers and priority level), as long as the number of lanes is higher than the number of hot-spot destinations.

From figures 9 and 10 it becomes obvious that the ATLAS backpressure protocol performs quite better than the corresponding wormhole protocol. There are two reasons for this: *(i)* cells that are destined to the same output are not allowed to occupy more than a single lane in ATLAS, and *(ii)* each lane can be time-shared among cells going to different destinations, unlike wormhole where lanes are dedicated to packets for the entire duration of their transmission.

Acknowledgements

This work was carried out within the "ASICCOM" project, funded by the European Union, under the ACTS (Advanced Communication Technologies and Services) Programme. The ATLAS I chip is designed by Panagiota Vatsolaki, Chara Xanthaki, George Kalokerinos, George Kornaros, Dionisios Pnevmatikatos, and George Dimitriadis. Evangelos Markatos, Christoforos Kozyrakis, Vagelis Chalkiadakis, and many others have also helped in various ways. We thank them all.

References

[ACPN95] T. Anderson, D. Culler, D. Patterson, and the NOW team: "A Case for NOW (Networks of Workstations)", *IEEE Micro Magazine,* vol. 15, no. 1, February 1995, pp. 54-64.

[Bork90] S. Borkar e.a.: "Supporting Systolic and Memory Communication in iWarp", *Proc. 17th Int. Symp. on Computer Arch.,* ACM SIGARCH 18-2, June 1990, pp. 70-81.

[DaSe87] W. Dally, C. Seitz: "Deadlock-Free Message Routing in Multiprocessor Interconnection Networks", *IEEE Trans. on Computers,* 36-5, May 1987, pp. 547-553.

[Dally90] W. Dally: "Virtual-Channel Flow Control", *Proc. of the 17th Int. Symp. on Computer Architecture,* ACM SIGARCH vol. 18, no. 2, May 1990, pp. 60-68.

[HIC95] IEEE Standard 1355-1995, ISO/IEC Standard 14575 DIS: "Standard for Heterogeneous InterConnect (HIC): low-cost, low-latency scalable serial interconnect for parallel system construction", 1995; URL: http://stdsbbs.ieee.org/ groups/1355

[HuKn84] A. Huang, S. Knauer: "Starllite: a Wideband Digital Switch", *Proc. GLOBECOM'84 Conf.,* Atlanta, GA USA, Dec. 1984, pp. 121-125.

[KaSS96] M. Katevenis, D. Serpanos, E. Spyridakis: "Credit-Flow-Controlled ATM versus Wormhole Routing", Technical Report FORTH-ICS/TR-171, *ICS, FORTH, Heraklio, Crete, Greece,* July 1996; URL: file://ftp.ics.forth.gr/tech-reports/1996/ 1996.TR171.ATM_vs_Wormhole.ps.gz

[KaSV96] M. Katevenis, D. Serpanos, P. Vatsolaki: "ATLAS I: A General-Purpose, Single-Chip ATM Switch with Credit-Based Flow Control", *Proceedings of the Hot Interconnects IV Symposium,* Stanford Univ., CA, USA, Aug. 1996, pp. 63-73. URL: file://ftp.ics.forth.gr/tech-reports/1996/ 1996.HOTI.ATLAS_I_ATMswitchChip.ps.gz

[Korn97] G. Kornaros, C. Kozyrakis, P. Vatsolaki, M. Katevenis: "Pipelined Multi-Queue Management in a VLSI ATM Switch Chip with Credit-Based Flow Control", *Proc. 17th Conference on Advanced Research in VLSI (ARVLSI'97),* Univ. of Michigan at Ann Arbor, MI USA, Sept. 1997; URL: ftp://ftp.ics.forth.gr/tech-reports/1997/ 1997.ARVLSI.Pipe_MultiQueue.ps.gz

[KuBC94] H.T. Kung, T. Blackwell, A. Chapman: "Credit-Based Flow Control for ATM Networks: Credit Update Protocol, Adaptive Credit Allocation, and Statistical Multiplexing", *Proceedings of the ACM SIGCOMM '94 Conference,* London, UK, 31 Aug. - 2 Sep. 1994, ACM Comp. Comm. Review, vol. 24, no. 4, pp. 101-114.

[Ohsa95] H. Ohsaki, e.a.: "Rate-Based Congestion Control for ATM Networks", *Computer Communication Review,* ACM SIGCOMM, vol. 25, no. 2, April 1995, pp. 60-72.

[OzSV95] C. Ozveren, R. Simcoe, G. Varghese: "Reliable and Efficient Hop-by-Hop Flow Control", *IEEE Journal on Sel. Areas in Communications,* May 1995, pp. 642-650.

[QFC95] Quantum Flow Control Alliance: "Quantum Flow Control: A cell-relay protocol supporting an Available Bit Rate Service", version 2.0, July 1995. URL: http://www.qfc.org

[YeHA87] Y. Yeh, M. Hluchyj, A. Acampora: "The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching", *IEEE Journal on Sel. Areas in Communications,* October 1987, pp. 1274-1283.