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.*

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.

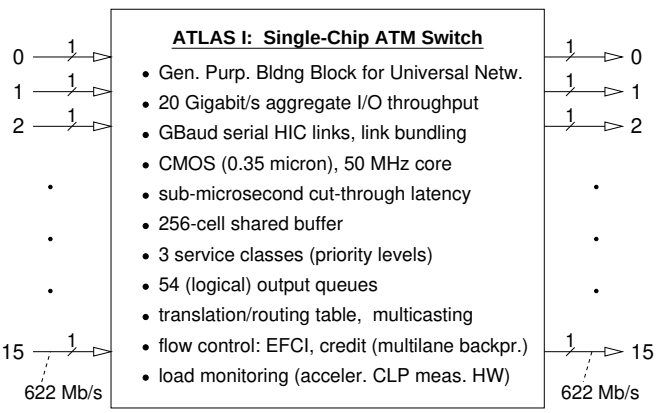

Figure 1: ATLAS I chip overview

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

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

Copyright 1997 IEEE. Published in the Proceedings of the GLOBECOM'97 Conference, November 3-8, 1997, Phoenix, Arizona, USA. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works, must be obtained from the IEEE. Contact: Manager, Copyrights and Permissions / IEEE Service Center / 445 Hoes Lane / P.O. Box 1331 / Piscataway, NJ 08855-1331, USA. Telephone: +1 (908) 562-3966.

[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.

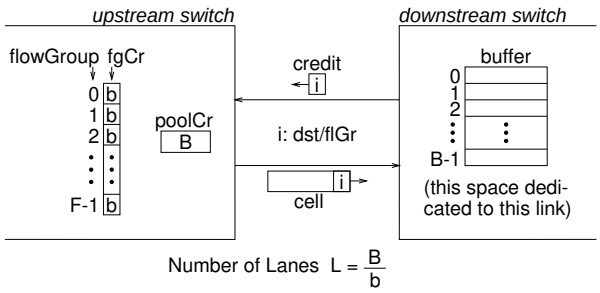

Figure 2: Credit flow control protocol for ATM switch chips (ATLAS I uses b=1, hence L=B)

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.

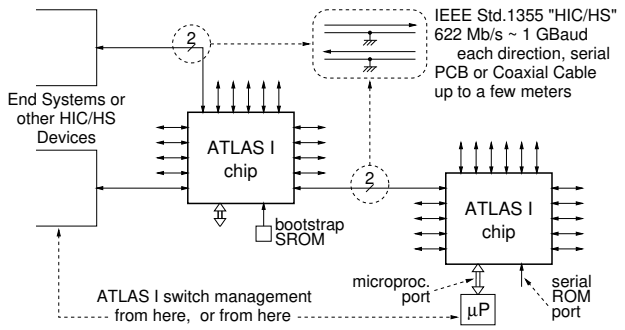

Figure 3: Basic ATLAS I connections

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 switching fabrics like the one shown in figure 4.

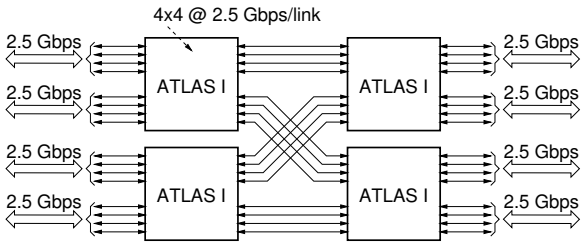

Figure 4: 8×8 fabric at 2.5 Gbps, using link bundling

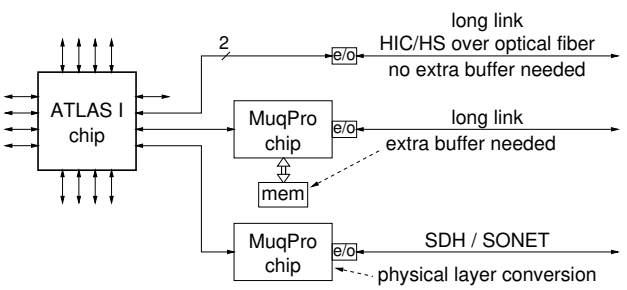

Figure 5: Multi-Queue Processor: interfacing to long links

When ATLAS I chips are to be connected to longer links or to different physical layers, an interface is required, as sho