

FA 9.2: Throttled-Buffer ATM Switch Output Control Circuitry with CAM-Based Multicast Support

Kenneth J. Schultz¹, P. Glenn Gulak²

¹Nortel Semiconductors, Ottawa, Ontario, Canada

²University of Toronto, Ontario, Canada

Asynchronous transfer mode (ATM) is emerging as the predominant broadband network switching protocol. Shared buffer architectures are a superior implementation alternative for ATM switch nodes, but they suffer from inherent drawbacks that are addressed in this paper: control throughput bottleneck, inefficient multi-casting, and inflexible time-division sharing of buffer bandwidth.

Because all cells pass through a single memory, buffer access control may constitute the switch bottleneck. Previous shared buffer switch chips have used linked-list or FIFO-based output control, achieving cell rates up to 12.8Mcell/s [1, 2]. The switch shown in Figure 1 uses a CAM-access approach [3]. Cells are uniquely identified by a combination of sequence number and port for unicasts, or sequence number and multicast identifier (MCI) for multicasts. A Tag CAM is used to search for these identifiers, and a control throughput of 47Mcell/s is achieved.

The multicast hardware employs a CAM (the McCAM) to query multicast destination bit maps. Only one copy of both the tag and the cell are stored. The McCAM is searched for multicasts that are unobstructed by busy unicast ports. The McCAM also enables global multicast querying, so that multicast connections behind the head-of-line may be serviced.

Each port in a standard NxN shared buffer switch is assigned 1/N of aggregate throughput, independent of load. In a throttled-buffer switch, this constant relationship is eliminated. Buffer access is statistically multiplexed between ports, allowing higher per-port and per-connection bandwidths, and enabling different peak rates on different ports. In the 16x16 prototype, peak per-port bandwidth is 1/5 of aggregate, instead of 1/16. The shared buffer is a blockable resource, and presents the combined problem/opportunity of using it as efficiently as possible. This task is managed by the output control circuitry, implemented on our test chip, and shown shaded in the figure.

The components comprising the output control circuitry are the output arbiter and four logic-enhanced memories: the counter/semi-sorter (CSS), that keeps track of unicast queue lengths and finds the two longest queues; the read sequence register (RSR), that stores and increments sequence numbers used to identify unicast cells; the multicast RSR (McRSR), that performs the same function for multicast cells; and the McCAM.

The CSS stores unicast queue lengths in its 4-port 16x8 core (Figure 2). Ports A and B update counts stored in the core. Port C is used for background updates. Port D is a test port not shown in the figure. An 8b incrementer and an 8b decremter are associated with Ports B and A, respectively. To achieve 20ns read-modify-write performance, the dynamic network in Figure 3 is used. Each q briefly goes high at the onset of each sense, providing an interval for negative precharging. When qn is substituted for q, the same circuit implements the 8b decremter.

The RSR and McRSR perform read-increment-write, using the circuit of Figure 3, under the same speed constraints. The RSR is a 16x8 single-port SRAM. The 64x16 McRSR uses a second port to set up new multicast connections in real time.

Only a single copy of each multicast cell resides in the shared buffer, and it is read only once and sent to all destination ports simultaneously. This is advantageous in a throttled-buffer switch, and optimizes shared buffer bandwidth. The McCAM, shown in Figure 4, stores a bit map of the destination ports of each multicast connection (MCI), with one word per MCI and one column per output port. An entry is written on connection set-up with "1" in bit positions corresponding to destination ports, and "0" in other positions. Because output ports are serviced and become available again asynchronously, it would be inefficient to either (a) interrupt all unicast service to accommodate the head-of-line multicast cell, or (b) wait until all destination ports are simultaneously free for the head-of-line cell. Instead, the McCAM is continuously searched for any multicast connection with queued cells and all ports simultaneously available. This is by searching for "0" in bit positions corresponding to busy ports, and "X" (don't care) otherwise, and is implemented with a 10-T one-way-mismatch 2-port core cell. A "1" stored in the same column as a "0" in the input word results in a miss. When a hit occurs, unicast output arbitration is interrupted for one cell period to allow a multicast read from the shared buffer, and the cell goes to all of its destination ports. Multiple non-overlapping multicasts may be serviced in consecutive cycles. Each of the 64 words has 21b: 16 for the bit map, "present" and "valid" bits, and a 3b priority code that enables preferential servicing based on fanout or other criteria.

The arbiter, shown in Figure 5, combines round-robin unicast service with a dynamic priority mechanism that preferentially services the two longest queues, as selected by the CSS. McCAM hits over-ride the arbiter such that round-robin state is preserved. A binary tree structure provides $O(\log_2 N)$ delay and wrap-around capability. The input interface circuit in Figure 5(c) includes decoders for the top-two queues, and enables unicast requests only for non-empty queues with an unoccupied output port. The P2 signal disables round-robin service, and the P1 signal disables both P2 and round-robin selection. Round-robin requests and token locations are processed by the fork nodes (Figure 5(d)). The ORing of P1, P2, and RRreq signals is performed globally (Figure 5(b)).

The architecture may be extended to deal with cell age and static priorities, in addition to the length metric, enabling a variety of algorithms for arbitration between traffic classes.

The shaded components of Figure 1 are implemented on the chip shown in Figure 6. The chip functions at clock rates up to 95MHz, processing 47Mcell/s, including any mix of multicast and unicast traffic. The power dissipation is 1.0W at 95MHz. Table 1 summarizes the experimental results.

Acknowledgments:

Thanks to J. Podaima for assistance in chip testing. This work was supported by Nortel, NSERC, the Canadian Advanced Technology Association, the Walter Sumner Memorial Foundation, and the IEEE Solid-State Circuits Council.

References:

- [1] Unekawa, Y., et al., "A 5Gb/s 8x8 ATM Switch Element CMOS LSI Supporting Five Quality-of-Service Classes with 200MHz LVDS Interface," ISSCC Digest of Technical Papers, pp. 118-119, Feb., 1996.
- [2] Denzel, W. E., A. P. J. Engberson, I. Iliadis, "A Flexible Shared-Buffer Switch for ATM at Gb/s Rates," Computer Networks & ISDN Systems, vol. 27, no. 4, pp. 611-624, Jan., 1995.
- [3] Schultz, K. J., P. G. Gulak, "CAM-Based Single-Chip Shared Buffer ATM Switch," Proc. ICC, pp. 1190-1195, 1994.

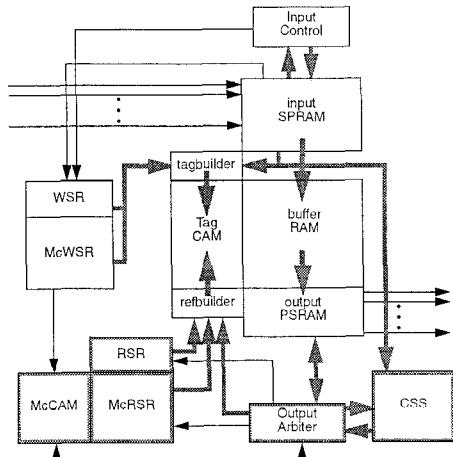


Figure 1: Complete CAM-based throttled-buffer switch (implemented components shown with shaded borders).

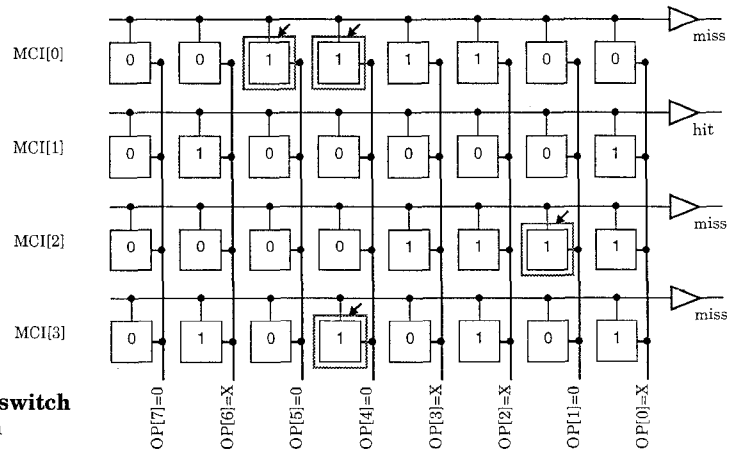


Figure 4: Multicast CAM: one of four MCIs shown is eligible for reading; OP=output port.

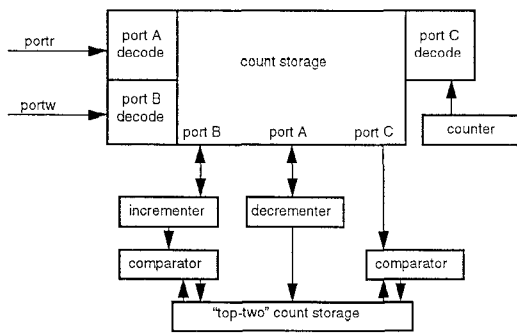


Figure 2: Counter/semi-sorter memory architecture.

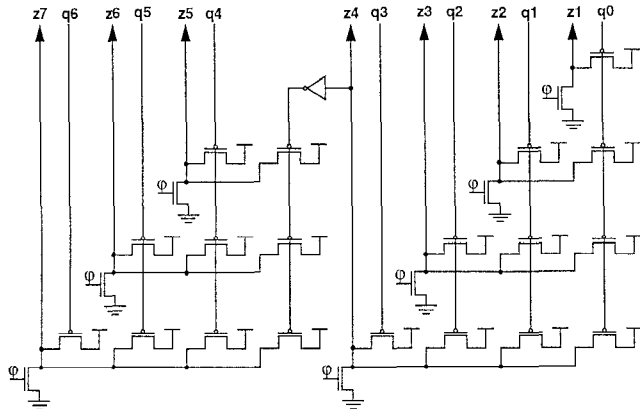


Figure 3: Flash incrementer: z outputs indicate whether a zero exists in input word q in any bit positions to the right.

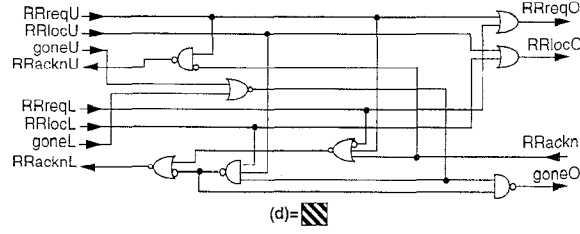
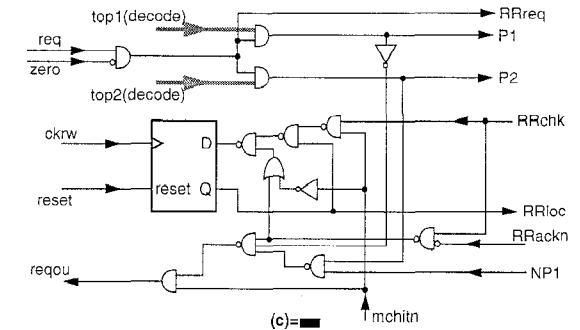
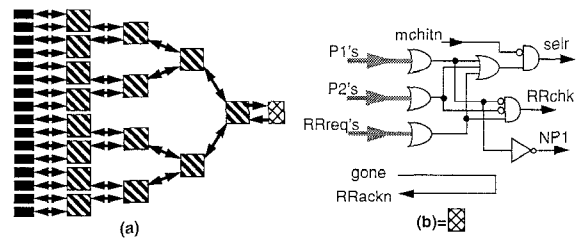


Figure 5: Binary tree arbiter: (a) over-all view, (b) end cap, (c) input interface, (d) fork node.

Figure 6, Table 1: See page 446.

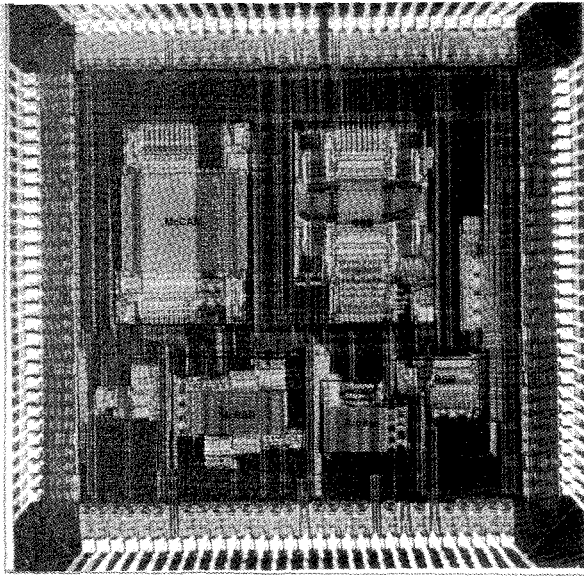


Figure 6: Output control test chip micrograph.

Function	Throttled-buffer output control
Fabrication technology	0.8 μ m BiCMOS, 3-metal
Number of transistors	69,140
Die size	4.8mmx4.8mm(23.0mm ²)
Active area	11.3mm ²
1-port SRAM core cell area	9.8 μ m \times 13.4 μ m (131 μ m ²)
2-port SRAM core cell area	15.7 μ m \times 12.6 μ m (198 μ m ²)
4-port SRAM core cell area	22.6 μ m \times 17.6 μ m (398 μ m ²)
McCAM core cell area	21.5 μ m \times 12.6 μ m (271 μ m ²)
Maximum clock frequency	95MHz at $V_{DD} = 5V, 25^{\circ}C$
ATM cell throughput	47Mcell/s
Equivalent switch bandwidth	20Gb/s
Power dissipation	1.0W at 95MHz, 5V

Table 1: Chip characteristics.

FA 9.3: A 40Gb/s 8x8 ATM Switch LSI using 0.25 μ m CMOS/SIMOX
 (Continued from page 155)

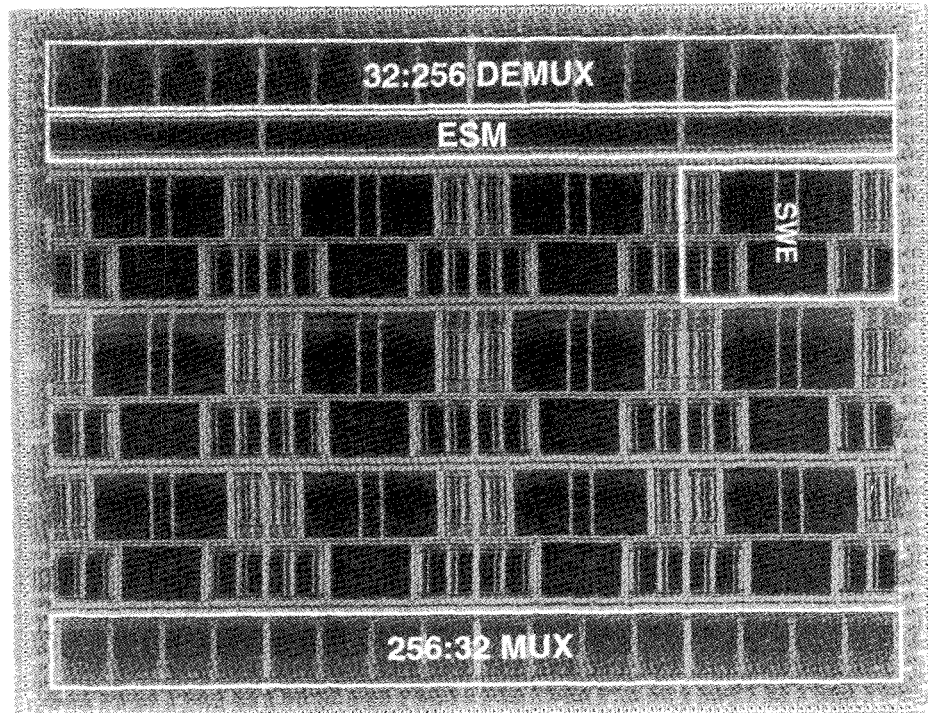


Figure 4: Chip micrograph