# Buffer sizing in all-optical packet switches

**Neda Beheshti[*], Yashar Ganjali[*], Ramesh Rajaduray[+], Daniel Blumenthal[+], Nick McKeown[*]**
[*]*High Performance Networking Group, Stanford University, Stanford, CA 94305*
[+]*Optical Communications and Photonics Research Group, University of California, Santa Barbara, CA 93106*
*{nbehesht,, yganjali}@stanford,edu, {meshraj, danb}@ece.ucsb.edu, nickm@stanford.edu*

**Abstract:** Packet-switched routers need buffers during times of congestion. We show that a combined input-output queued router needs no more buffering than an output queued router. Using simulations, we show that 10-20 packet buffers are enough.

**OCIS codes:** (060.4250) Networks; (060.1810) Couplers, switches, and multiplexers

## 1. Introduction and Motivation

All-optical routers could offer several advantages for the core of the Internet: high capacity, reduced power consumption, and hence increased port density. Over time, this could lead to more compact, high capacity routers.

Like any router, an all-optical router needs packet buffers. Until recently, it was believed that the buffers needed to be huge: A widely used rule-of-thumb required core routers to buffer up to $C \times RTT$ where $C$ is the capacity of the bottleneck link, and $RTT$ is the effective round-trip propagation delay of packets going through the router. For a router with a link capacity of 10Gb/s and an average $RTT$ of 250ms, we would need to buffer 1,250,000 packets (of average size 250 bytes); clearly an inconceivable size for an optical buffer. At best, optical buffers based on integrated and switched optical delay lines could store up to, say, 100 packets in the next few years [4].

Recently, Appenzeller *et al.* showed that we can reduce the buffer size by a factor of $\sqrt{N}$ without any degradation in network performance; where $N$ is the number of long-lived flows going through the router [1]. Thus, if the 10Gb/s linecard has 10,000 flows, we could reduce the buffers to 10,000 packets. This has huge implications for electronic routers, but unfortunately the buffer size is still too large for optical buffers.

As part of the LASOR program[1] we were interested in how networks could be built from all-optical routers. This required understanding how a packet-switched router could work with just a handful of buffers – say 20-100 packets. Conventional wisdom said this would not work. But intuition suggests that if the traffic in the core of the network (where there are many flows) is sufficiently multiplexed and aggregated, the arrival process will be very smooth and perhaps smaller buffers will suffice. For example, in the extreme, if the arrivals were Poisson, a buffer of just 20 packets would lead to a loss rate of less than 1% in a network operating at 80% load.

Recently, it has been shown that if packets are sufficiently spaced out – either by the source, or by the network – then when they arrive at the router, the traffic is sufficiently smooth as to require only very small buffers [3,5]. Theory and simulations suggest that under certain conditions, a router needs only 20 packet buffers independent of line-rate and $RTT$ [3]. The two conditions are: (1) We are prepared to lose approximately 25% of the link capacity; i.e. a 40Gb/s link would operate like a 30Gb/s link, and (2) The access links run much slower than the core links.

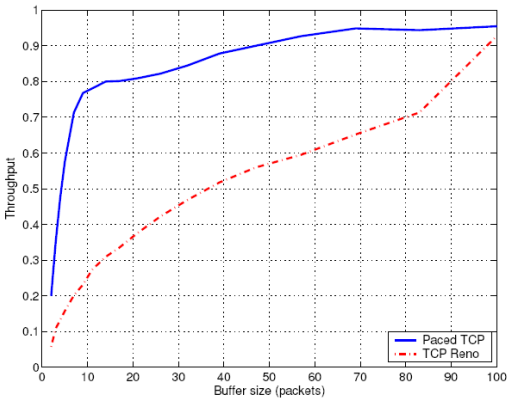


**Figure 1.** Throughput as a function of buffer size.

The basic idea is as follows. 95% of packets come from TCP sources which, today, send a burst of packets once every RTT. If instead the TCP source spaced out the packets, spreading them over the RTT, it would reduce the burstiness of (and hence smooth) the arrivals to the routers. This is called TCP Pacing [3]. Alternatively, even if the source doesn't spread the packets, we have found that the network will naturally spread packets out as the packets pass from a slow access network to the fast backbone network. Given that backbone networks typically run at 2.5Gb/s or 10Gb/s, and will run even faster in the future, almost every access link runs much slower than the backbone network. For most flows (e.g. where the user is connected to the

[1] The LASOR program is a consortium of UCSB, Stanford, Cisco, JDSU, Agility and Calient and is fanded by DARPA MO under DARPA/MTO DOD-N award no. W911NF-04-0001/KK4118 (LASOR PROJECT) and the DARPA/MTO Buffer Sizing Grant no. W911NF- 05-1-0224.

network via modem, DSL, cable or from a 10Mb/s or 100Mb/s Ethernet), packets will be naturally spaced out by the network and very small buffers will suffice. If a high-performance host is connected directly to the backbone (e.g. a supercomputer), then the access link could be too fast and the packets won't be spread by the network. In this case, we should modify the TCP sources to use TCP Pacing [3]. Figure 1 shows the throughput of a network as a function of buffer size. We can see that when with TCP pacing we can get a throughput of about 80% with only 20 packets of buffering. Without pacing however, the throughput is only 35%.

The routers studied in [3] are output-queued (OQ) routers, *i.e.* it has been assumed that any packet arriving at an ingress linecard is immediately transferred to the corresponding egress linecard. In practice, this requires a speedup of $n$ in an $n \times n$ router, which is not practical in a high performance router. Often, routers use combined input and output queueing (CIOQ) with a speedup between 1 and 2.

A CIOQ switch with speedup of $s$, has $s$ scheduling cycles per time slot. Once per cycle, a scheduling policy determines which packets to transfer from the input ports to the output ports. In a crossbar switch, the scheduling algorithm must resolve two constraints per cycle: (1) at most one packet can be removed from each input port; and (2) at most one packet can be delivered to each output port. When $1<s<N$ we need buffers at both input and output ports. Internet routers today commonly virtual output queuing (VOQ) in which each input keeps a separate FIFO queue for each output linecard.

In our work, we want to know how big the buffer should be for a CIOQ router, and we attempt to answer the question in this paper. We show that in theory, a CIOQ switch needs no more buffers than an OQ switch; i.e. 20 packets should suffice under the same constraints described above. Using simulation, we explore how much buffer is needed when we use practical scheduling algorithms.

## 2. Theoretical Bound

Consider two routers $A$ and $B$, and assume the same input traffic is fed to both routers. $B$ is said to exactly emulate $A$ if packets depart at the same time from both routers.

**Lemma.** If a CIOQ router $B$ exactly emulates an OQ router $A$, then the occupancy of any queue in $B$ is no larger than the corresponding queue in $A$.

This is because [2] shows that with speedup $s=2$, there exists a specific scheduling algorithm (stable marriage scheduling policy) such that a CIOQ router can exactly emulate an OQ router. Combining this with the results on small buffers [ref] proves the lemma.

**Theorem.** There is a scheduling algorithm, which can guarantee a high throughput (for instance, above 80% of link capacity) in a CIOQ router under the constraints stated in [3].

## 3. Simulation Results

We enhanced *ns2* [6] to include accurate router models and simulated the general topology depicted in Figure 2. The router has 32 input and output ports; link capacity is 40Gb/s. Each input port carries 500 multiplexed TCP Reno flows consisting of 1000 byte packets. The TCP flows are generated at separate source nodes on slow access links, then multiplexed together onto the backbone. Packets arriving at the switch linecards are segmented into 500 bit long cells, then reassembled before they depart. The router has speedup of two; in each cycle a simple scheduling algorithm removes either zero or one cell from every input port and sends it to the output. The scheduling algorithm selects high-priority input and output ports, i.e., those which are less recently served and distributes the load uniformly among different channels of each output port.

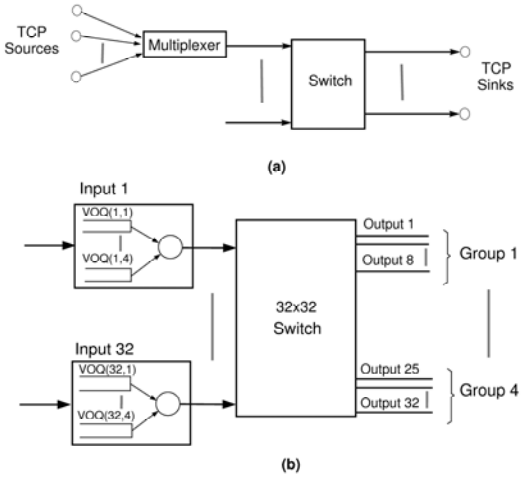


**Figure 2.** (a) General Topology of the simulated Network. (b) CIOQ with virtual output queuing.

Output ports of the switch are divided into 4 groups of 8 optical channels. All the channels in a group carry traffic to one destination. The generated traffic consists of long-lived TCP flows with an average RTT of 40ms. Different RTTs are modeled by adding random jitter to the propagation delay of the access links.

To find out how much buffering is needed at input and output ports, we first assume that the size of buffers at the input side is unlimited. Simulation results show that the tail probability of queue occupancy at input ports is a non-decreasing function of the output buffer size. But even with output buffers of size 1000 packets, the VOQ occupancy

doesn't exceed 2 packets in the duration of simulation. Therefore in the next set of simulations we assume that the size of input queues is only 2 packets per VOQ, and investigate the appropriate size of the output queues.

The amount of buffering needed in a router depends on how bursty the TCP traffic is. As argued in [3] when the access links have limited capacity compared to the core links, the router finds the arrival traffic less bursty and this reduces the size of buffers needed for achieving high throughput. Figure 3 shows how link utilization varies as the size of output buffer increases considering two scenarios: 1) Access links have the same capacity as the core links, and 2) Capacity of each access link is 2Gb/s (i.e. 20 times slower than the backbone link). The plot shows that with slow access links, we need only five packet buffers per output port to achieve 80% utilization. We would need about 100 packets if the access links run at the same speed as the backbone. In both cases there are two packets per VOQ.

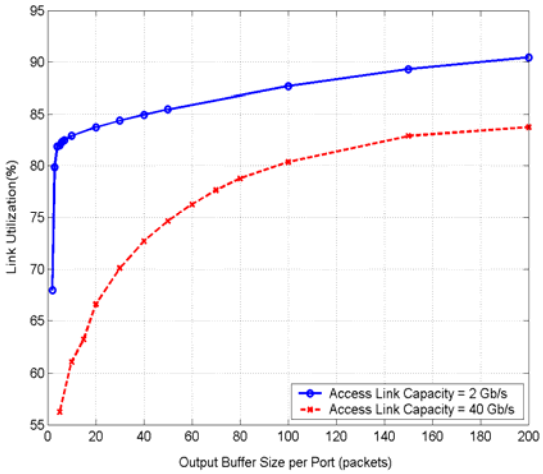


**Figure 3**. Link utilization as a function of output buffer size.

## 4. Discussion

When electronic networks were first introduced, it made sense to use large buffers so as to efficiently utilize precious long-haul links. Nowadays, long-haul capacity is abundant and backbone networks are greatly over-provisioned. On the other hand, optical buffers are expensive. Perhaps it makes sense to tradeoff the efficiency of the long-haul links so as to allow the use of small optical buffers. Early results based on theory and simulation suggest that very small buffers – small enough to build optically – might be enough. The results presented here suggest that with more practical CIOQ routers, the theory holds. Experiments on real networks are underway

## 5. Acknowledgement

The authors would like to thank Ashish Goel and Tim Roughgarden from Stanford University for their advice.

## References

[1] G. Appenzeller, I. Keslassy, and N. McKeown. Sizing router buffers. In *Proceedings of SIGCOMM '04*, pp. 281–292, New York, NY, USA, 2004.

[2] S.T. Chuang, A. Goel, N. McKeown, and B. Prabhakar. Matching Output Queueing with a Combined Input Output Queued Switch. In *Proceedings of IEEE INFOCOM '99*, 1169-1178, IEEE, 1999.

[3] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden. Part III: Routers with very small buffers. *ACM/SIGCOMM Computer Communication Review*, 35(3):83–90, July 2005.

[4] H. Park, E. F. Burmeister, S. Bjorlin, and J. E. Bowers. 40-gb/s optical buffer design and simulations. In *Numerical Simulation of Optoelectronic Devices (NUSOD)*, 2004.

[5] G. Raina, D. Towsley, and D. Wischik. Part II: Control theory for buffer sizing. *ACM/SIGCOMM Computer Communication Review*, 35(3):79–82, July 2005.

[6] The Network Simulator - *ns2*, http://www.isi.edu/nsnam/ns/