

# A Low-Latency and Bandwidth-Efficient Distributed Optical Burst Switching Architecture for Metro Ring

Andrea Fumagalli, Prasanna Krishnamoorthy  
Optical Network Advanced Research (OpNeAR) lab.

University of Texas at Dallas

Erik Jonsson School of Engineering and Computer Science

E-mail: {andreaaf, sanna}@utdallas.edu

**Abstract**—Optical Burst Switching (OBS) provides statistical multiplexing capabilities at the optical layer with relaxed hardware requirements when compared to optical packet switching. One of the open challenges of OBS is to assemble as many packets as possible in the same burst, while at the same time ensuring low latency of the transmitted packets.

The authors propose the use of OBS to realize a geographically distributed packet switch for metro rings. High efficiency of the ring bandwidth and low packet latency are obtained at the ring node by combining a multi-token based protocol for contentionless and loss-free transmission of bursts, known as the LightRing protocol, with the creation of bursts that contain packets belonging to multiple traffic flows (classified by priority and destination).

As illustrated in the paper, the proposed solution yields throughput that is significantly higher than that one offered by a centralized packet switch connected to the ring nodes via dedicated optical circuits. Latency of real time packets is kept at few dozens of milliseconds under a variety of network scenarios. The solution scales well geographically for metro applications.

## I. INTRODUCTION

Optical Burst Switching (OBS) has been proposed as today's viable solution to close the gap between the vast bandwidth available in Wavelength Division Multiplexed (WDM) networks and the much smaller traffic load offered by individual end-users. In OBS, packets are electronically buffered at the aggregation nodes, assembled into bursts, and transmitted all-optically throughout the core network as a single burst. Generally, buffering of the burst at the intermediate nodes is not provided. The advantage of OBS is that it provides statistical multiplexing at the optical layer with relaxed switching constraints (when compared to packet switching). The main challenge in designing an OBS network is to reconcile two opposing objectives: first, increasing the burst size, which reduces the switching burden and overheads at the optical layer; second, maintaining acceptable packet latency that is mainly determined by the burst assembly time and the reservation time prior to transmitting the burst into the network.

Several solutions have been proposed to realize OBS networks. These architectures mostly concentrate on mesh networks. In [1], a control channel is used to send a burst setup message before the burst is transmitted. The control message is used to configure the optical switches at the intermediate nodes. A burst may be dropped due to contention in the network, if there are no optical buffers. In [2], a central request server is used to coordinate the burst transmissions and avoid contentions. The request is sent to the central server after the source receives the first packet for a given destination.

Reservation of a burst transmission takes a fixed delay, i.e., the round-trip propagation between the source and the central server, plus its processing time. The use of ring networks for OBS is proposed in [3]. A variation of the Just Enough Time protocol for transmitting the bursts called Only Destination Delay (ODD) is used. Each node transmits on one wavelength only to avoid contention. The assembly of the bursts itself is a critical aspect of OBS and has been studied in [4] and [5]. The methods normally proposed to assemble bursts are based on timeout, and/or burst-size threshold. By changing the timeout and threshold values, it is clearly shown that optimum values exist for both parameters, in order to provide a minimum packet loss probability.

In conclusion, in all the proposed solutions, burst switching may not always yield bandwidth-efficient solutions due to a number of reasons: transmissions are not contention-free and bursts are randomly lost, burst transmissions are scheduled by a central server without matching the burst size with the packets awaiting transmission at the source node, or wavelengths are dedicated to the transmission of individual node's bursts.

In this paper, OBS is applied to realize a geographically distributed switch with ring topology for metro areas. The distributed switch is realized by connecting a number of Optical Burst Add and Drop Multiplexing (OB-ADM) nodes to a WDM ring, known as the LightRing [6]. The LightRing multi-token protocol provides a mechanism for real-time reservation of wavelengths that is used for bandwidth-efficient and loss-free transmission of optical bursts.

Low packet latency and bandwidth-efficient burst creation and transmission are the result of a unique combination of interesting properties. Due to loss-free transmission, packets are not retransmitted and it is possible to build larger bursts for reduced overheads and better bandwidth utilization. (In solutions where bursts may be lost, increasing the burst length may cause higher packet loss.) Due to loss-free transmission, it is also possible to combine, within the same burst, packets with different priorities without affecting the performance of real-time traffic. The ordered sequence of OB-ADM nodes along the ring perimeter also allows for efficient assembly of packets intended for adjacent destination nodes: traffic flows whose destinations are in the same ring portion can be multiplexed using the same burst. By allowing packets that belong to multiple traffic flows to be transmitted using the same burst, the assembly time may be significantly reduced, without significantly affecting the ring bandwidth efficiency. The dis-



wavelength.

The token containing the updated information announces the burst transmission to the downstream OB-ADM nodes, so they can set up their on-off switches to allow the burst to reach the egress OB-ADM nodes all-optically. The burst signal is stopped at the egress OB-ADM node. This feature allows for spatial reuse — allowing multiple transmissions on the same wavelength at the same time. After the burst is completely transmitted, the first time the token returns to the ingress OB-ADM, the token is updated to reflect the fact that the wavelength is now available for other transmissions. Note that the holding time for the ring reserved resources is an integer multiple of the token round-trip time. If a burst transmission lasts less than an integer multiple of a token (or ring) round-trip time, some of the reserved bandwidth is unused.

The next two subsections describe the QoS mechanism and the burst assembly and transmission strategies that run on top of the multi-token protocol.

#### A. QoS Provisioning

In order to transmit real-time packets without excessive delay, a QoS mechanism is required that regulates the transmission of bursts. A QoS mechanism is proposed that handles *two* classes of traffic, namely, real-time and best-effort. Every effort is made by the QoS mechanism to minimize the average latency of the real-time packets at the ingress OB-ADM node, while maximizing the overall ring throughput. Fair access among the OB-ADM nodes is also provided by the proposed QoS mechanism.

Upon arrival at the ingress OB-ADM node local port(s), a packet is stored in one of the  $2 \times (N - 1)$  queues, depending on its priority and destination. A timer is associated with the first packet that enters an empty real-time queue. This timer measures the delay suffered by the packet, and it is used to force transmission of the packets in the queue, to avoid unacceptable packet latency at the ingress OB-ADM node. A counter is used to measure the total number of data bytes stored in each queue, awaiting transmission. When this latter counter exceeds a given threshold, transmission of packets in the queue is enforced.

The token traffic counter is used to advertise the total amount of real-time traffic awaiting transmission in the ring. The counter indicates the time (measured in ring round-trip delay units) required, to transmit the real-time traffic currently awaiting transmission, using one single wavelength. This value serves as an indicator to OB-ADM nodes of the amount of pending real-time traffic in the ring. When the token is received, this field is updated with the current amount of real-time traffic at the OB-ADM node. If the amount of real-time traffic indicated by the token has increased when compared to the value indicated at the previous arrival of the same token, then the OB-ADM node does not transmit best-effort bursts using this token. Only when the traffic counter has decreased or remained constant, since its last arrival, can bursts with best-effort packets alone, be transmitted.

To ensure fair access amongst OB-ADM nodes, the same token (wavelength) cannot be used by the same OB-ADM node to transmit two bursts consecutively — i.e., ring resources must be freed after transmitting each burst.

#### B. Burst Assembly and Transmission (BAT) Strategies

The use of the multi-token protocol to determine resource availability both in the network and at the egress OB-ADM nodes, enables the ingress OB-ADM node to optimally choose the burst to be transmitted at any given token arrival. In other words, once the token arrives and informs the ingress OB-ADM node of the ring available resources, it is possible to select the burst for transmission with the objective of minimizing packet latency of the real-time traffic and maximizing the ring throughput.

Upon reception of a given token, the following variables are defined at the ingress OB-ADM. Let  $N_R$  be the set of nodes to which links are free on the token wavelength, and which have available reception buffers. Let  $N_{last} \in N_R$  be the node that is farthest from the source. Let  $N_T$  be the set of nodes in  $N_R$ , whose queues have timed out.  $N_{oldest} \in N_T$  is the node which has the oldest timed out packet. Let  $N_t$  be the farthest destination in  $N_T$ .  $N_S$  is the set of nodes in  $N_R$ , for which the aggregate queue size (sum of real-time and best-effort traffic) has crossed the threshold, and  $N_s$  is the node in  $N_S$  farthest from the source. The size threshold used here is the amount of data that can be transmitted during a token round-trip time.

Three burst assembly and transmission (BAT) strategies are introduced next, which trade complexity for reduced packet latency. Consider the OB-ADM source of a packet to be  $S$ , the OB-ADM destination  $D$ , and the traffic class  $C$ .

1) *S, S/D, and S/D/C Strategies:* With the  $S$  strategy, burst consists of packets that have the same OB-ADM source,  $S$ . A burst may be created by assembling both real-time and best-effort packets, with multiple destinations. Thanks to the optical drop and continue feature of the OB-ADM node, a burst transmission reaching  $N_{last}$  will be received by any other node in  $N_R$ . Bursts can be created that contain packets intended for any node between the OB-ADM source and  $N_{last}$ . This provides a powerful means for transporting small amounts of real-time traffic, while keeping the bandwidth utilization high.

In determining the destinations whose packets may be assembled in the same burst, it is relevant to take into consideration spatial reuse. For example, reserving the resources upto  $N_{last}$ , and transmitting only to the first downstream node, causes the spatial reuse to be very limited, thereby reducing the bandwidth efficiency. A prudent approach is to assemble packets that are intended for nodes located in the same portion of the ring (i.e., physically adjacent), thus allowing spatial reuse to take place more efficiently.

In practice, to foster spatial reuse, only best-effort packets intended for  $N_{last}$  are added to the burst (if QoS is satisfied). However, when real-time packets timeout, they are added into the burst irrespective of their destinations, as in this case latency is more critical than efficiency.

The following algorithm is used.

- 1) Token is received from the control channel.
- 2) Token is checked to find  $N_R, N_T, N_t, N_S, N_s$ .
- 3) Check if there are any nodes in  $N_T$ .
  - a) If so, create a burst with all the real-time packets of nodes in  $N_T$ .
  - b) Add the best-effort packets intended for  $N_t$  to the tail of the burst.

- c) Update and release the token and transmit the burst. Notice that the wavelength is reserved between the OB-ADM source and  $N_t$ .
- 4) If  $N_T$  is empty, check if there are any nodes in  $N_S$ .
- 5) If so, create a burst from the real-time traffic for  $N_s$ , adding best-effort packets going to  $N_s$  to the tail of the burst.
- 6) Update and release the token, and transmit the burst.

Here, the QoS policy studied in the previous section, is used to decide whether best-effort packets can be transmitted alone on a burst. When the traffic counter has increased best-effort traffic is added only to fill up the final round-trip of the real-time traffic, in order to prevent any adverse effects on the real-time traffic.

The  $S$  strategy offers the best performance for real-time traffic, in terms of expected latency experienced by the packets. As already mentioned, although the wavelength is reserved till  $N_t$ , it is not always used to transmit packets to  $N_t$ , thereby slightly reducing bandwidth efficiency. If this algorithm is used, each of the intermediate nodes on a lightpath must monitor the burst, and extract those packets which are destined for it. This process is made simple with the use of HDLC encoding for each of the packets in the burst, and MPLS tags for each of the destinations. The OB-ADM node simply discards the packets (parts of the burst) that are not intended for it.

The other two BAT strategies are sub-cases of the  $S$  strategy. With the  $S/D$  strategy a burst consists of only packets that have the same pair  $(S, D)$ . Bursts are created using real-time packets and appended best-effort packets that are intended for  $N_{oldest}$ . When there are no nodes in  $N_T$ , a burst created with the real-time and best-effort packets destined for  $N_s$ , is transmitted. With the  $S/D/C$  strategy a burst consists of packets that have the same triple  $(S, D, C)$ . Bursts are created using only real-time packets for node  $N_{oldest}$ , or only best-effort packets to node  $N_s$ . The best-effort traffic burst is transmitted only when the QoS mechanism allows it, and there are no nodes in  $N_T$ .

### III. PERFORMANCE RESULTS

This section presents simulation results to assess the performance of the proposed distributed switch architecture. Results are derived using IP header traces obtained from the National Laboratory for Applied Network Research (NLNR)<sup>1</sup>. The traces used are from the Indiana University GigaPoP. These traces consist of OC-12c IP headers framed using ATM/AAL5, using LLC/SNAP encapsulation, and they are used to simulate the traffic generated by the feeder networks connected to the OB-ADM nodes.

For performance comparison, consider a centrally switched network, where the IP packets are individually switched at a central node. The topology is that of a logical star, on top of the physical ring, with an extra node serving as the central switch. Packets are transmitted from (to) the ADM node (which replaces the OB-ADM node) to (from) the central switch using dedicated circuits. Each ADM node has a fixed, equal share

of the ring transmission capacity. To provide a best bound on the performance of this architecture, zero processing time and instantaneous output buffering of all incoming packets are assumed at the central switch. To provide a fair comparison, the ring capacity is the same in both systems, the LightRing and the centralized switch architecture.

The system under consideration is a ring consisting of  $N = 16$  nodes, and  $W = 4$  data wavelengths, each with 10 Gbps rate capacity. For reduced hardware cost, each ingress OB-ADM is allowed to transmit only one burst at a time. The switch in the centrally switched architecture, is placed at a 17th node which has no local input or output traffic. The ring spans 80 km and the nodes are placed evenly around the ring. The token round-trip time is therefore approximately 400  $\mu$ s. The ingress OB-ADM node has buffers of 24 MB per class per destination. This buffer size is enough to avoid packet drops due to buffer overflow under low to moderate traffic loads. The time for scheduling the burst, reading the token, and modifying the token is assumed to be negligible compared to the ring round trip time and the burst aggregation time. Simulation statistics are collected by allowing a settling time of 0.3 seconds and collecting the statistics over the next 0.5 seconds.

The destinations of the packets in the traces are hashed to a number in  $[1, N - 1]$ , which gives the number of hops between the source and the destination node in the ring. The hash function is chosen to produce an approximately uniform distribution of the destination nodes from each trace IP address. The packets are marked as real-time based on their source address. This choice assumes that certain sources in the feeder networks are high-priority and others are best-effort. If the packet length distributions remain the same, the policy for selecting the best-effort/real-time packets may be changed. Unless otherwise specified all the simulations are done with the ratio of real-time to best-effort traffic as 20:80.

The parameters of interest are the *effective throughput*, the *expected response time*, and the  $a/D$  (average burst transmission time to the token round-trip time). Both throughput and response time of the real-time traffic, and throughput of the best-effort traffic are studied. The response time for the best-effort was found to be, on an average, 5-10 times the average response time of the real-time traffic. The effective throughput is the amount of traffic transmitted on the ring in bps normalized to the link (fiber) capacity of the ring. Due to spatial reuse this may be greater than one for the LightRing. Thus, an effective throughput of 1.3 implies that the network carries  $1.3 \times 4 \times 10\text{Gbps} = 52\text{Gbps}$ . The expected response time, is the average of the response times of all the transmitted packets. Response time is defined as the time spent in the transmission buffer awaiting transmission plus the transmission time. The propagation delay is not taken into account.

Fig. 2 shows the throughput versus response time of the real-time traffic for the three BAT (Burst Assembly and Transmission) strategies. In the figure, by varying the real-time timeout, a few curves are obtained for each strategy.

The solid lines refer to the  $S$  strategy. It is seen that they achieve high throughput, while maintaining an acceptable response time. The dashed lines refer to the  $S/D$  strategy. With less complexity, this strategy yields similar throughput

<sup>1</sup>We acknowledge the National Science Foundation Cooperative Agreement No. ANI-9807479, and the National Laboratory for Applied Network Research, for their effort in producing the IP header traces, IND-1018572878-1, IND-1018591418-1, IND-1018582160-1 and IND-1018602487-1.

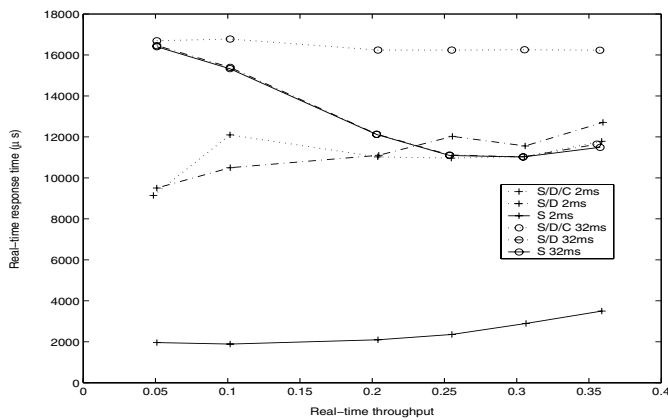


Fig. 2. Variation of real-time response time

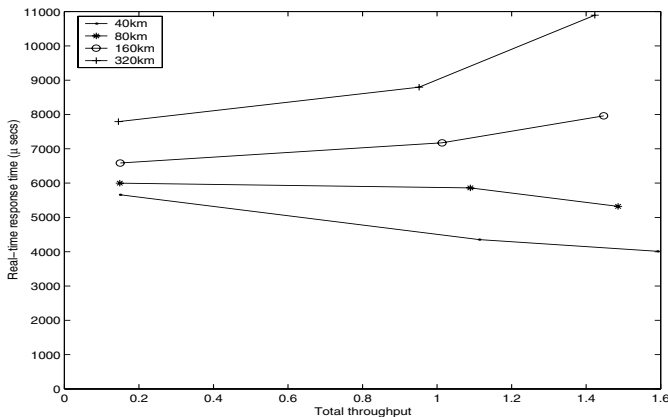


Fig. 3. Variation of response time with ring size

at the expense of an increase in the response time. The dotted lines refer to the  $S/D/C$  strategy. Being the simplest of the three, this strategy yields lower throughput and relatively large response time.

Figure 3 shows the scalability of the distributed switch ( $S$  strategy) over a wide range of ring sizes, from 40 km to 320 km. Response time of the real-time traffic is plotted against the total effective throughput. The real-time to best-effort traffic ratio is 20:80. The real-time threshold used is 10 ms.

Table I shows throughput and response time (in ms) of real-time packets when the real-time to best-effort traffic ratio is 50:50. The real-time threshold is chosen to be 2 ms and 8 ms. Obtained values seem to be acceptable when the real-time load does not exceed 0.6.

Table II, shows the throughput and response time of the centralized architecture for a real-time to best-effort traffic ratio of 50:50. Fixed multiplexing and lack of spatial reuse limit the achievable throughput to 1, as opposed to twice that for the LightRing.

#### IV. SUMMARY

The paper proposed a distributed optical burst switching architecture for efficient statistical multiplexing of traffic in metro ring networks. The paper focused on an enabling module of the proposed architecture that jointly assembles arriving packets and schedules transmission of the burst at the ring node, known as the Optical Burst Add and Drop Multiplexing (OB-ADM) node.

TABLE I

$S$ , RATIO OF REAL-TIME:BEST-EFFORT TRAFFIC IS 50:50

Total Load	RT Threshold 2 msec			RT Threshold 8 msec		
	Throughput		Response time-RT ms	Throughput		Response time-RT ms
	Total	RT		Total	RT	
0.16	0.15	0.07	1.9	0.15	0.07	4.7
1.15	0.87	0.58	1.8	1.10	0.58	5.1
2.13	1.37	1.00	12.1	1.41	1.02	14.2

TABLE II

CENTRALIZED SWITCHING CASE, RATIO OF REAL-TIME:BEST-EFFORT TRAFFIC IS 50:50

Total Load	Throughput		Response time-RT ms
	Total	RT	
0.17	0.17	0.1	0.005
1.31	0.98	0.78	0.027
2	0.99	0.99	27.2

There are many advantages with the proposed solution. Thanks to its distributed statistical multiplexing of traffic, it yields significantly higher throughput than that offered by a centralized packet switch connected to ADM nodes via dedicated circuits. Latency of real-time packets at the OB-ADM node is kept at few tens of milliseconds by 1) jointly assembling and scheduling the transmission of the burst, and 2) assembling in the same burst, packets that belong to multiple flows — assembled packets may have distinct priorities and destinations. It was demonstrated that the proposed solution scales well geographically for metro — and perhaps regional — applications thanks to the distributed multi-token access protocol. OB-ADM nodes can be gradually added to the ring as needed up to the maximum number allowed by the OB-ADM design, without requiring changes at the other active OB-ADM nodes.

#### V. ACKNOWLEDGEMENT

The work was partially supported by NSF under contract # ANI-0082085 and by CPqD. We are also grateful to the reviewers for their helpful suggestions.

#### REFERENCES

- [1] M. Yoo, M. Jeong, C. Qiao, "A High Speed Protocol for Bursty Traffic in Optical Networks", SPIE's All-Optical Communication Systems: Architecture, Control and Protocol Issues, Vol. 3230, pp. 79-90, Nov. 1997.
- [2] M. Düser, P. Bayvel, "Analysis of Wavelength-Routed Optical Burst-Switched Network Performance", Optical Communication, ECOC '01, vol.1, 2001, pp. 46-47.
- [3] L. Xu, H.G. Perros, G.N. Rouskas, "A Simulation Study of Protocols for Optical Burst-Switched Rings", Proceedings of Networking 2002, Springer, May 2002, pp. 863-874.
- [4] A. Ge, F. Callegati, L.S. Tamil, "On optical burst switching and self-similar traffic", IEEE Communications Letters, Vol. 4, No. 3, March 2000, pp. 98-100.
- [5] V. Vokkarane, K. Haridoss, J. Jue, "Threshold-Based Burst Assembly Policies for QoS Support in Optical Burst-Switched Networks", Proceedings, OptiComm'02, Boston, MA, July-Aug 2002.
- [6] J. Cai, A. Fumagalli, "LightRing: A Distributed and Contention-free Bandwidth On-Demand Architecture", IFIP 5th Working Conference on Optical Network Design and Modeling, Vienna, Austria, Feb 2001.
- [7] J. Cai, LightRing: An Integrated WDM Ring Network Solution for Optical Metropolitan Area Networks, PhD dissertation, University of Texas at Dallas, Aug 2001.
- [8] A. Fumagalli, N. Kalairajah, J. Fitchett, "Efficient Tuning Strategies for LightRing", IEEE ETTC 2002, September 23-24, 2002.