Retransmission schemes for Optical Burst Switching over star networks

Anna Agustí-Torra†, Gregor v. Bochmann*, Cristina Cervelló-Pastor†
†Escola Politècnica Superior de Castelldefels, Universitat Politècnica de Catalunya
Av. Canal Olímpic s/n, 08860 Castelldefels, Barcelona, Spain
*School of Information Technology & Engineering, University of Ottawa
800 King Edward Ave., Ottawa, Ontario, K1N 6N5, Canada
†{anna.agusti,cristina}@entel.upc.edu, *bochmann@site.uottawa.ca

***Abstract*-In this work we propose an Optical Burst Switching retransmission scheme without any loss of bursts for star topology networks that exploits the entire network capacity. The basic idea is to allow the core node to resolve contentions by itself; balancing the intelligence of the network between edge and core nodes, losing neither the advantages of optical burst switching nor the simplicity of the core switch architecture. To achieve this objective, two retransmissions schemes are defined. The former confines the average number of retransmissions per burst below 2. The later exploits the whole network capacity.**

***Keyboards*-Optical Burst Switching, Agile All-Photonic Networks.**

I. INTRODUCTION

Optical Burst Switching (OBS) [1] is a Wavelength Division Multiplexing (WDM)-based technology that exploits the large available bandwidths by reducing the electronic processing of optical packets. Data packets are aggregated into much larger bursts at the edge of the network. Before the transmission of each burst, a control packet is sent on a separate control wavelength in order to request the necessary resources at each intermediate node. If the required capacity can be reserved, then the burst can pass through the core nodes without opto-electronic conversion. After the burst transmission, the channel is released allowing for other burst transmissions to use it. This statistical multiplexing of the channel among multiple flows improves the backbone efficiency and offers excellent scalability.

Despite the benefits of the OBS paradigm, its high burst blocking probability delays its introduction in the industry. In this paper we propose two modifications of the OBS retransmission scheme presented in [2] to provide a scheme without any loss of bursts that allows the utilization of the entire network capacity for the star network architecture described in the Agile All-Photonic Network Project.

The paper is organized as follows. The remaining part of Section I provides a brief introduction to the AAPN Project and OBS technology. Section II describes the operation of the basic OBS retransmission scheme and introduces the proposed modifications. Section III includes some simulation results and performance analysis to compare the new schemes (version 2 and 3) with the original OBS retransmission scheme (version 1) and a simple-minded Time Division Multiplexing (TDM) approach. Section IV discusses some conclusions and future work.

A. AAPN Project

Current communications networks already exploit the power of light for signal transmission. Nevertheless, opto-electronic conversions are still performed at various stages in the communication process. Opto-electronic conversions are expensive, require large amounts of power, and limit the scalability of the network.

This bottleneck can be eliminated by an all-photonic network, which limits the electronic technology to network access points. To achieve this objective the network has to have the ability to perform time-domain multiplexing in order to dynamically allocate bandwidth to traffic flows as the demand varies. Moreover, the control and routing functionalities have to be concentrated at the edge switches that surround the photonic core. Furthermore, the network has to include rapidly reconfigurable all-photonic space switching in the core. These elements are the base of the Agile All-Photonic Networks (AAPN) Project [3].

AAPN is a Research Network, launched in 2003 and sponsored by NSERC- the Government of Canada's Natural Sciences and Engineering Research Council- as well as contributions from six Canadian companies and two government laboratories, with the aim of embracing all of the essential elements that are required to realize all-photonic networks in the future.

The AAPN network topology [3] is based on an overlay of several stars that provides robustness in the case of link or core node failure. The overlaid star forms a logical mesh by connecting each edge node to every other edge node. However, data traversing the network only passes through one photonic switch, resulting in a major simplification of the control problem. From the control point of view, each star can be managed independently because there is no data interaction at a single point in the network.

B. Optical Burst Switching

Optical Burst Switching (OBS) [1] is a WDM-based technology somehow between wavelength routing (circuit switching) and optical packet switching. The unit of transmission is a burst, whose length in time is arbitrary. Each burst is preceded by the transmission of a control packet, which usually takes place on a separate channel.

Using a one-way reservation process, the source node does not wait for confirmation that an end-to-end connection is set-up before sending a burst. Instead, the source starts transmitting a data burst after a small delay (called offset), following the transmission of the control packet. The purpose of this control packet is to inform each intermediate node of the upcoming data burst, allowing the intermediate node to configure its switch fabric in order to direct the burst to the appropriate output port.

The benefit of Optical Burst Switching (OBS) over conventional circuit switching is that there is no need to dedicate a wavelength to each end-to-end connection. In

addition, optical burst switching is more viable than optical packet switching as the burst data does not need to be buffered or processed at the cross connects. This leverages effectively the strengths of optical switching technologies and circumvents the problem of buffering in the optical domain.

II. Burst retransmission approach

One major concern in OBS networks is burst contention resolution. Contention occurs when two or more bursts want to use the same wavelength of a given output port of a switch at the same time. Several mechanisms have been proposed to reduce burst contention: fiber delay lines [4], wavelength conversion [4], deflecting routing [5] or burst segmentation [6]. Nevertheless, they all remain sensitive to the traffic load.

A. Basic retransmission scheme

In [2], a retransmission scheme is proposed in order to provide a transmission without any loss of bursts. With the original OBS, there is no control at the intermediate nodes; the burst is simply ignored in case of contention and the recovery is performed by higher layer protocols. However, in OBS with retransmissions, each edge node keeps a copy of every burst sent until its successful delivery. When the core node detects a collision, it should notify the incident to the concerned edge node sending a negative acknowledgement, called NACK Packet, with the identifier of the discarded burst. Hence, the edge node might try to retransmit the burst without intervention of higher layer protocols. The implementation of this scheme requires that the burst control header carry the sequence number of each burst.

The buffer size at the edge nodes is a crucial parameter for the feasibility of the retransmission scheme, especially for a very wide network where the round trip could be very significant and hence one edge node may need to store many copies of bursts transmitted. The delivery delay is another important parameter and increases linearly with the number of transmissions. A burst transmitted n times needs $n*T$ units of time (T is the round trip delay). For a wide network T maybe very significant. Therefore, n should be very small to keep the delivery delay acceptable. However, in local or metropolitan networks, where the propagation delay is relatively small, this retransmission scheme is very efficient.

B. Retransmission scheme version 2

Using the basic OBS retransmission scheme the average number of transmissions per burst increases as the load increases. In [2], the authors propose a congestion control mechanism that keeps this average number around a constant value, below 1.5, by reducing the offered load in function of the collision rate. Nevertheless, with this approach the utilization of the network remains below its potential.

In this paper, we propose a modification of the basic OBS retransmission scheme for star topology networks designed to keep the retransmission ratio below a given threshold and without modifying the amount of traffic offered by the sources. The basic idea is to allow the core node to schedule a transmission of a given burst in case of contention. Taking into account that a core node (switch) has to process each control packet, the implementation of this modification does not increase considerably the complexity of the control plane. When a core node processes a control packet and cannot reserve the necessary resources, it reschedules the burst transmission and sends a special control packet, called Core Reserve Packet, to inform the edge node for a suitable time to transmit the burst.

Edge nodes only keep a copy of the first transmission of each burst. As in the original retransmission scheme, the egresses do not send acknowledgments, so a timer controls the lifetime of a copy of a sent burst. In an AAPN star network, each burst has to pass through only one core node. Hence, the timer is set to the round trip time between the egress and the core node plus a guard band depending on the processing time of the control packet in the core node. The interval of time between the transmission of a control packet and its corresponding burst, i.e., the offset time, is usually less than the propagation delay between the edge node and the core node. Therefore, this modification does not avoid the contention of the first transmission of a burst. However, the mechanism ensures that the second transmission will be successful and guarantees that the average number of transmission per burst is less than 2, as long as the core node can always reschedule the transmission of a burst.

When the core node has to allocate the transmission of a burst, it has to notify the edge node of its allocation. The edge node will receive this information after the propagation delay between the core and this edge node. Furthermore, this allocation is referred to the transmission of a burst that will reach the core node after the propagation delay between the edge and the core node. That means that the core node has to schedule the transmission of a burst with a minimum delay equal to the sum of these propagation delays. To prevent the core node from overwriting a new allocation made by the edge node, another term has to be added. This term should be set, at least, to the maximum offset of the edge node plus the maximum burst length.

C. Retransmission scheme version 3

Although the modification of the retransmission scheme presented in the previous subsection maintains the average transmission per burst below the threshold of 2, it is easy to verify that the overall traffic offered to the network may be greater than the offered load by the sources connected to edge nodes. In the worst case, to transmit each burst successfully it is necessary to send the burst twice. Therefore, the network might become overloaded with high offered loads.

To allow the use of the total network capacity we propose another modification as follows. When the occupation of the queue where the edge node stores the transmitted burst to a given destination becomes greater than a predefined threshold, the edge node does not try to allocate the transmission of new bursts. Instead, the edge node sends a control packet, called Reservation Request Packet, to the core

node without transmitting the burst. Then, the edge node waits until the core node makes the allocation for this burst and returns a Core Reserve Packet to this edge node. In fact, the ingress node makes a request to the core node for a suitable time to transmit the concerned burst.

Therefore, in the version 3 of the retransmission scheme edge nodes have two possible modes of operation:

1) *Normal mode*: When the occupation of the queue to a particular destination is below a predefined threshold, the edge node has the behavior as defined in version 2 of the retransmission scheme.

2) *Request mode*: When the occupation of this queue is above the predefined threshold, the edge node makes a reservation request to the core node before the burst is transmitted to a given destination.

In the next section we explain how the threshold is defined to prevent the overload of the network and to allow the entire utilization of the network capacity.

The occupation of the queue, where each edge stores the copies of the bursts sent to a given egress node, is chosen to decide the operation mode of each source because it is a direct indicator of the load in the link between the core and this egress node. In fact, a copy of a given burst is stored in this queue until the edge node is sure that the transmission of this burst has been or will be successful. Thus, the occupation of each of these queues depends on the total amount of load that has to be transmitted to a given egress edge.

Fig. 1 shows the control plane operation of the OBS retransmission schemes. The dark gray shadowed part illustrates the original retransmission scheme operation (version 1). The dark gray plus the light gray shadowed parts correspond to version 2 of the retransmission scheme. The overall figure shows the operation of the control plane, using version 3 of the retransmission scheme.

III. Performance analysis

One of the major advantages of using Optical Burst Switching is its capability to keep a low average transmission delay. However, the burst loss ratio experimented in this kind of networks is usually relatively high. The retransmission schemes proposed in this paper allow a low average burst transmission delay and assure data transmission without losses, even for high variable offered load, on star topology networks.

All schemes are sensible to the number of hops between the ingress and the egress node. However, in star topology networks, data only passes through one photonic switch. Therefore, using the same offset, the probability of collision is the same for each edge node, preserving the fairness among all sources.

A. Performance of retransmission schemes

Fig. 2 shows the average burst transmission delay for each retransmission scheme as a function of the offered load per destination in a star topology network with 64 edges nodes and 1 wavelength at 10 Gbps. All edge nodes are sources and destinations and each source generates the same amount of traffic to each destination. The burst inter-arrival distribution is Poisson. All links are 100 Km in length and the burst length is 1 slot (10 microseconds in the AAPN network architecture). Fig. 2 also shows the performance of a TDM approach in order to compare the results in the next subsection. We consider in the following the additional delay introduced by queuing and retransmission in addition to propagation delay and the burst transmission time. In Fig. 2, for loads below 0.65 Er and 0.55 Er for version 2 and 3 of the retransmission scheme, respectively, this additional delay is:

$$d_1 = D_1 + z * (RTT + D_2), \quad (1)$$

where D_2 = (*maximum offset* + 1) * *slot_length* (D_2 = 50 microseconds in Fig. 2), $D_1 = r * slot\_length / (2 * (1 - r))$, z is the average number of re-transmissions per burst, *RTT* is the round trip time between the source node and the core node, r is the load for each output link of the core node and *slot_length* is the duration of a slot in time units. D_1 is the average waiting time in the M/D/1 queuing model and D_2 is an additional delay introduced when the core node schedules a burst transmission in order to prevent that the core node

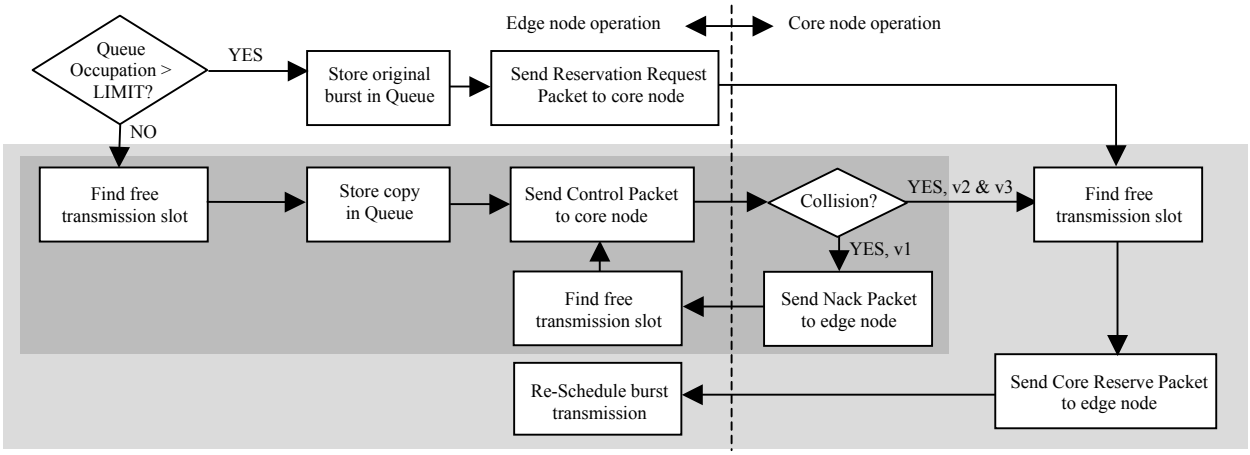

Fig. 1. Control plane operation of OBS retransmission schemes

may overwrite a new allocation made by the edge node. Note that in both versions, 2 and 3, the average number of re-transmissions per burst, *z*, is always less than 1.

Using version 1 of the retransmission scheme for loads below 0.6 Er in Fig. 2, this additional delay is independent of D_2 but the average number of re-transmissions per burst, z, is higher than using either version 2 or 3:

$$d_1 = D_1 + z * RTT. \quad (1')$$

As the load increases, the average number of re-transmissions per burst increases until the input links to the core node are almost full. After this point, the network is overloaded if no action is taken. In Fig. 2, using either version 1 or 2 of the retransmission scheme the average burst transmission delay tends to infinite when the average traffic offered per destination is bigger than 0.6 and 0.65 Er, respectively. Version 3 prevents the network overflow changing the operation mode of edge nodes from normal to request mode when the queue occupation becomes greater than a predefined threshold (in Fig. 2, the switching point between both modes is 0.55 Er). Thus, using version 3, the expression of the additional delay when sources operate in request mode (above 0.55 Er) can be defined as follows:

$$d_2 = D_1 + RTT + D_2. \quad (2)$$

In order to use the queue length to decide the mode of operation of each source it is necessary to define this occupation as a function of the traffic offered to the network. In these simulations (with uniform traffic), for version 2 below 0.65 Er and version 3 of the retransmission scheme, the expression is approximately:

$$Queue\ length = ((r / (N - 1)) / slot\_length) * RTT. \quad (3)$$

This length is independent of D_1 and D_2. In normal mode the copy of a given burst is stored in this queue for *RTT* time units when the burst transmission starts. In request mode the original burst is stored in this queue when the edge node sends the Reservation Request Packet until it processes the Core Reservation Packet, i.e., *RTT* time units.

Hence, a threshold can be derived from (3) for each value of *r*. Fig. 3 depicts the sum of the average occupation of all these queues of an ingress node as a function of the offered load per source to the network. The average occupation of these queues using version 2 below 0.65 Er and version 3 for every offered load is close to the theoretic value.

Since version 3 of the retransmission scheme avoids the overloading problem, the utilization of the network can be close to the ideal value (the simulations assume that each slot is completely full). Fig. 4 shows the utilization of each output link of the core node as a function of the utilization of each input link. Using version 1 and 2, when input links to the core node are almost full (occupation close to 1 Er.), output links are still under its potential (0.62 and 0.65 respectively). However, the curve corresponding to version 3 changes its tendency when the average carried load over input links to the core node is around 0.6 Er. This adjustment is due to the fact that the edge nodes change their behavior from normal to request mode in order to avoid overloading the network.

B. Comparison with a TDM approach

In order to make a comparison with a well-known mechanism, Fig. 2 shows the results using a simple-minded TDM approach with 64, 32, 16 and 8 edge nodes. This approach assumes the same amount of traffic per source, i.e., in each output link of the core node, 1 slot per frame (consisting of $N - 1$ consecutive slots) is reserved for a burst transmission from a given ingress node, where N is the number of input links to the core node. The average waiting time in the queue for each burst (before its transmission) can be obtained applying the M/D/1 with vacations queuing model. Then,

$$d = (0.5 * (N - 1) * slot\_length) / (1 - r), \quad (4)$$

As (4) shows, the additional burst transmission delay using the TDM approach is dependent on the number of nodes and the slot length but independent of the link lengths. In general, it is possible to modify both parameters, number of nodes and slot length, to achieve the desired average delay. On the other hand, as (1) and (2) show, the average additional delay using

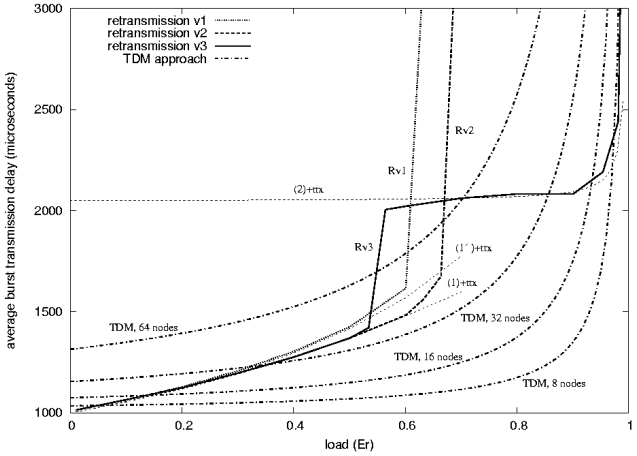

Fig. 2. Average burst transmission delay as a function of total offered load to the network per destination

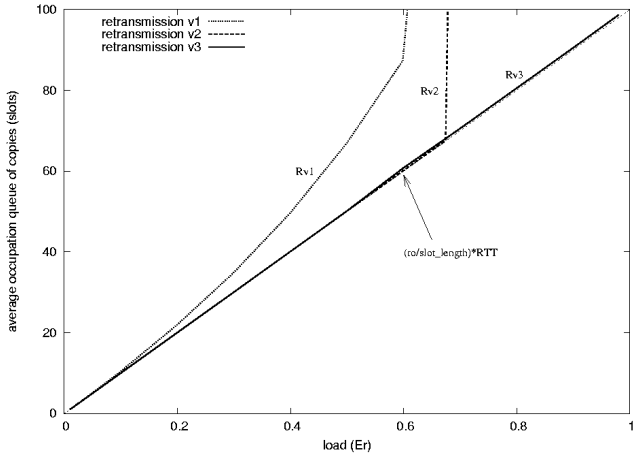

Fig. 3. Average queue length as a function of total offered load to the network per destination

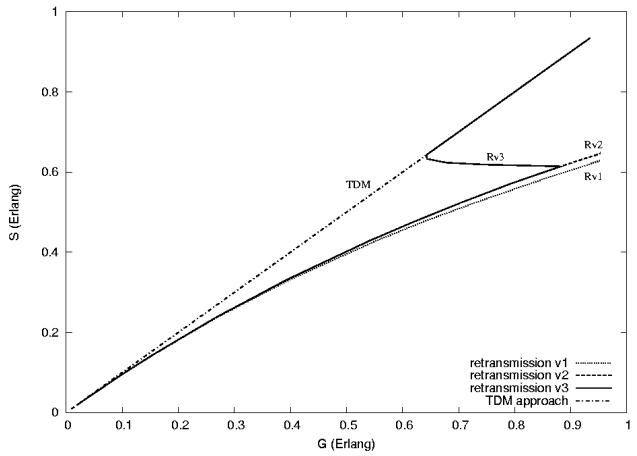

Fig. 4. Average utilization of an output link of the core node as a function of average utilization of an input link to the core node

the proposed retransmission schemes is independent of the number of nodes but dependent on the link lengths and the slot length. Modifying the link lengths is usually a difficult task. Therefore, the performance of the proposed schemes is better in local and metropolitan area networks where the propagation delay is relatively small.

By means of (1), (2) and (4) it is possible to compare the additional delay introduced by the proposed retransmission schemes (v2 and v3) with the TDM approach as a function of the number of nodes, the slot length and the link lengths. For example, using the same scenario described in the previous subsection, when r = 0.2 Er, the average number of re-transmission per bursts using either v2 or v3 is 0.11 (simulation result) and the additional delay introduced by the proposed schemes is less than the additional delay using the TDM approach when the number of edge nodes is bigger than 19, i.e., $d_1 < d$ if $N > 19$. When r = 0.5 Er, the average number of re-transmission per bursts is 0.35 and the minimum number of edge nodes increases to 39. Using version 3, when $r > 0.5$ Er the comparison has to be done using (2); e.g., when r = 0.7 Er, $d_2 < d$ if $N > 64$. The same comparison with links of a length of 500 Km provides the following results: $N > 90$, 178 and 304, respectively.

Moreover, as (3) shows, the buffer space required to implement the proposed mechanisms is also dependent on the link lengths. In fact, this storage capacity is an important design parameter and the main restriction to achieve high scalability.

It is noteworthy that the results presented in this paper show the behavior of the proposed schemes when all sources generate the same amount of traffic. In fact, this is the worst scenario for the retransmission schemes and the best for the simple-minded TDM approach because the offered load increases in a uniform way. In a generic scenario, where sources generate different amount of load, it is possible that some of the sources work in normal mode while others operate in request mode (even the same edge node may have different behaviors as a function of the destination of a given burst). An inherent feature of all the retransmission schemes is their flexibility under different traffic load. As a drawback, all the retransmission schemes might provide out-of-order delivery and it is necessary to use some kind of mechanism to reorder the burst at the egress nodes, however, this function may be left to the IP protocol layer.

IV. Conclusions and Future Work

In this paper we propose two modifications of the retransmission scheme for OBS star networks [2]. The final version provides a burst optical switching framework without any loss of bursts that exploits the entire network capacity. In contrast to the simple-minded TDM approach, using the proposed retransmission schemes the average burst transmission delay is dependent on neither the number of nodes nor the slot length. However, for high loads the average burst transmission delay and the buffer capacity as well, are highly dependent on the network diameter. Thus, better results are achieved for local and metropolitan area networks. Furthermore, the proposed schemes allow dynamic bandwidth allocation. However, they produce out-of-order delivery. Future work includes developing an analytical model of the re-transmission probability of each scheme, evaluating the fairness of the proposed schemes with different link lengths and the ability of the network to support different classes of service. As well as extending the comparison to other TDM approaches with dynamic bandwidth allocation.

Acknowledgment

This work was supported by the Spanish Ministerio de Ciencia y Tecnología (MCYT) within the framework of the TIC2003-09042-C03-02 project, as well as the Canadian Natural Sciences and Engineering Research Council (NSERC) and industrial and government partners, through the Agile All-Photonic Networks (AAPN) Research Network.

References

[1] M. Yoo, C. Qiao. “Optical Burst Switching [OBS] – A New Paradigm for an Optical Internet.” Int’l. J. High Speed Networks, vol. 8, no. 1, 1999, pp. 69-84.

[2] A. Maach, G. Bochmann and H.T. Mouftah, “Robust Optical Burst Switching.” In Proc. of Networks’2004, Vienna, Austria, June 2004, pp. 447-452.

[3] G. v. Bochmann, T. Hall, O. Yang, M. J. Coates, L. Mason, R. Vickers. “The Agile All-Photonic Network: An Architectural Outline.” In Proc. 22nd Biennial Symposium on Comm., Kingston, Canada, June 2004.

[4] J. Turner. “Terabit Burst Switching.” Int’l. J. High Speed Networks, vol. 8, no. 1, 1999, pp. 3-16.

[5] Y. Chen, H. Yu, D. Xu, C. Qiao. “Performance Analysis of Optical Burst Switched Node with Deflection Routing.” Proc. of IEEE ICC, vol.2, 2003, pp.1355-1359

[6] A. Maach, G. v. Bochmann. “Segmented Burst Switching: Enhancement of Optical Burst Switching to decrease loss rate and support quality of service.” 6th IFIP Working Conference on Optical Network design and modeling, Torino, Italy, February 2002, pp. 69-84.