

Using TCP Congestion Control to Improve the Performances of Optical Burst Switched Networks

S.Y. Wang

shieyuan@csie.nctu.edu.tw

Department of Computer Science and Information Engineering
National Chiao Tung University
Hsinchu, Taiwan

Abstract

We propose using a modified TCP decoupling approach as a congestion control mechanism for optical burst switched networks. The TCP decoupling approach [1] is a novel way that can apply TCP congestion control to any traffic flow (which can be an aggregate) in a network. Since this approach is generic, it has found applications in several areas [2, 3]. In the optical burst switching (OBS) area, because the basic mechanism of the TCP decoupling approach matches the mechanism of the OBS very well, we propose using a modified TCP decoupling approach to congestion-control the traffic load offered to an OBS switch and regulate the timing of sending bursts. Our simulation results show that this approach enables an OBS switch to achieve a high link utilization while maintaining a very low packet (burst) drop rate.

1. Introduction

Optical burst switching (OBS) [4, 5] is a scheme for transporting traffic over a bufferless optical network. At a source node, packets are grouped into a burst and sent together as a unit before they are switched through the network all optically. Before sending a burst, the source node first sends a control packet along the burst's routing path to configure every switch on the path. The control packet is sent over an out-of-band channel and will be electronically processed at each switch to real-time allocate resources (e.g., transmission time slots.) for the burst. The time offset between sending the control packet and its burst should be large enough. This is to ensure that, at each intermediate switch, the control packet can always arrive before its burst. Figure 1 shows the mechanism of the OBS scheme.

The control packet contains information about routing, the burst length, and the offset time. The routing information is for the switch to decide the outgoing interface for the burst. The length information tells the switch how much time the burst transmission will last. The offset time information lets the switch know that after a time interval given by the offset time, a burst will arrive. With these information, the control packet will try to reserve the specified period of time at the chosen outgoing interface for the burst. If that period of time has not been allocated to any other burst, the control packet

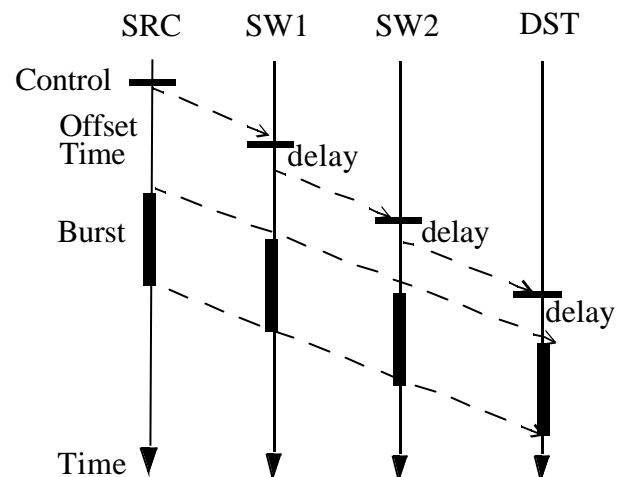


Figure 1. The source node separates the transmissions of the control packet and the burst by an offset time. At each intermediate switch, because the control packet needs to be electronically processed, it will be slightly delayed compared to the burst transmission in the optical domain.

can successfully make the reservation. Otherwise, the switch simply discards the control packet without returning any feedback to the source node. At each intermediate switch, the control packet performs the same procedure to reserve resource for its burst. This operation continues until the control packet successfully reaches its destination node or is discarded at some switch.

This control packet signaling protocol is purposely designed to be one-way rather than two-way [4, 5]. That is, the source node need not wait for the reply of the source-to-destination reservation request to come back before sending its burst. If instead a two-way signaling protocol is used, in optical networks with high link bandwidth and large RTTs, tremendous optical bandwidth would be wasted during the RTT, because no packets can be sent during this period of time. In contrast, using the proposed one-way signaling protocol, this period of waiting time can be reduced to a fixed and tiny value, which is the used time offset. References [6, 7] provide more detailed descriptions of this one-way signaling protocol and its variations.

Although the OBS scheme provides a higher link utilization than circuit (wavelength) switching scheme, its performance is still not satisfactory. Due to the optical switch's bufferless property, when multiple bursts contend for the same link at about the same time, only one burst can be successfully switched in the optical domain. All other overlapping bursts need to be dropped. This results in low link utilizations and high burst drop rates.

Some existing contention resolution schemes for photonic packet networks such as fiber-delay lines (FDLs) [4] and deflection routing [8] may be used in OBS networks. However, these methods have their own drawbacks. For FDLs, they are costly and their buffering capability is limited. For deflection routing, packets may be transported out-of-order and thus harming a TCP connection's achievable throughput. In [9], the author proposed utilizing additional capability in the form of multiple wavelengths to reduce contention. However, optical wavelength conversion is costly and not mature yet.

More recently, the authors in [10, 11] proposed a segmentation approach to reduce an OBS switch's packet drop rate. In this approach, when the desired transmission times of two contending bursts overlap partially, one burst is chosen to be switched in its entirety while the other burst is allowed to be partially switched, with the overlapping part being truncated. In [10], a simulation case shows that this approach can reduce the packet drop rate. In [11], a model is used to show the same improvement.

Although these existing approaches can reduce the degree of contention and thus the packet drop rate somewhat, they are ineffective when the offered load is excessively high. Without congestion-controlling the load offered to an OBS switch, when the load goes above 1 (measured in Erlang), the packet drop rate will eventually go up to a very large number. It is thus very important and necessary to use a congestion control mechanism to control the load offered to an OBS switch.

In this paper, we propose using a modified TCP decoupling approach [1] to control the load offered to an OBS switch. In this approach, a TCP decoupling virtual circuit (VC) is set up for each pair of burst source node and burst destination node. The VC uses TCP congestion control to control the sending rate of its burst source node. Because TCP is a well developed and refined congestion control protocol, the total sending rates of contending burst source nodes will not exceed the bottleneck link's bandwidth too much. This effectively controls the load offered to an OBS switch (and thus avoid high burst/packet drop rate) while keeping the link utilization high.

This approach also uses the arrival times of TCP decoupling VCs' acknowledgement packets to control the timing for sending bursts. Taking advantage of the TCP's famous "self-clocking" property [12], which uses the arrivals of ACK packets to trigger the transmissions of more data packets at the bottleneck link's bandwidth, source nodes now send their bursts following the timing of their returning ACK packets.

This effectively reduces the burst blocking probability and the overlapping time of contending bursts at the OBS switch. With the use of a segmentation scheme, a much reduced burst overlapping time translates to a much increased link utilization and a much decreased packet drop rate.

In the rest of the paper, Section 2 shows the burst/packet drop rate and link utilization of the OBS scheme and the OBS with segmentation scheme. These simulation results show that, although the segmentation scheme helps improve performance somewhat, the packet drop rate is still high and the link utilization is still low. Section 3 presents the operation of a modified TCP decoupling approach, pointing out how it operates in an OBS network. Section 4 presents simulation results showing that the modified TCP decoupling approach can significantly reduce the packet drop rate while keeping the link utilization high. Finally, Section 5 concludes this paper.

2. Performances of OBS Schemes

This section presents simulation results showing the performances of the OBS and the OBS with segmentation schemes. The performance metrics of interest to us is the burst/packet drop rate and the link utilization.

The topology of the simulation testbed network is shown in Figure 2. There are ten burst source nodes contending for the bandwidth of the link between the OBS switch and the burst destination node. The bandwidth of all links is set to 10 Gbps and the delay of all links is set to 5 ms.

The bursts generated by a burst source node is assumed to be a Poisson process. Each burst has the same length of 300 microseconds worth of bytes (i.e., 37,500 bytes on a 10 Gbps link, or 25 1500-byte packets). The inter-burst time is exponentially distributed with mean of X microseconds. By varying the value of X , we can vary the load (measured in Erlang) generated by each burst source node. This in turn varies the total load offered to the OBS switch. For example, if we want the total load offered to the switch to be 1, since there are ten burst source nodes, we can let each source node generate 0.1 load. In this case, the value of X should be set to 2,700 so that $300 / (300 + 2,700) = 0.1$.

Figure 3 shows the link utilization and burst drop rate performances of the OBS scheme. We see that when the offered load is 1, which is the optimal load that should be offered to a switch for it to achieve a high link utilization and a low burst drop rate, the link utilization is only about 50% and the burst drop rate is as high as 50%.

Figure 4 shows the link utilization and burst drop rate performances of the OBS with segmentation scheme. Compared to Figure 3, we see that both the link utilization and the packet drop rate are improved. The link utilization is increased from 50% to 62% and the packet drop rate is decreased from 50% to 38% when the offered load is 1. We see that, although the segmentation scheme does help improve the performances of the OBS switch, the link utilization is still low and the packet drop rate is still too high. Under such a

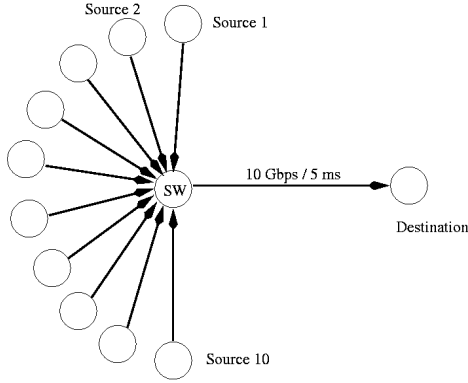


Figure 2. The topology of the simulation testbed network.

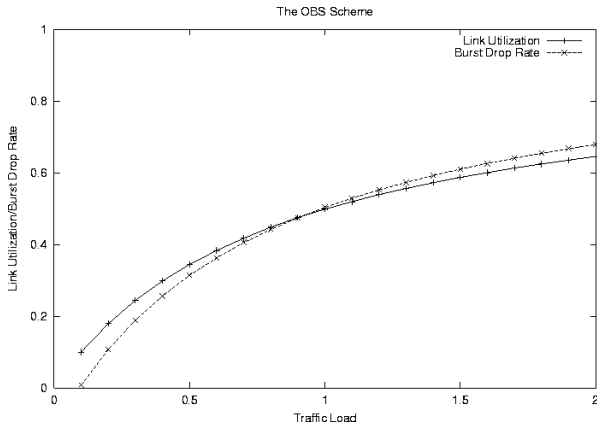


Figure 3. The link utilization and burst drop rate performance of the OBS scheme.

high packet drop rate of 38%, the quality of service perceived by end users would be intolerable.

With these simulation results, we realize that in order to fully take advantage of the benefit provided by the segmentation scheme, contending bursts should not overlap with each other too much. Otherwise, the advantage provided by the segmentation scheme would be minimal. We can think of the segmentation scheme as a passive measure. That is, a scheme that tries to do its best under an already bad situation. To further increase the link utilization and decrease the packet drop rate, we need to take an active measure to actively avoid burst collisions and, if collisions really occur, their overlapping time. This is the reason why we propose using a modified TCP decoupling approach to control the load offered to the switch and regulate the timing of sending bursts.

3. The Modified TCP Decoupling Approach

3.1. The TCP Decoupling Approach

The TCP decoupling approach has been proposed in [1] and applied to trunking [2] and wireless applications [3].

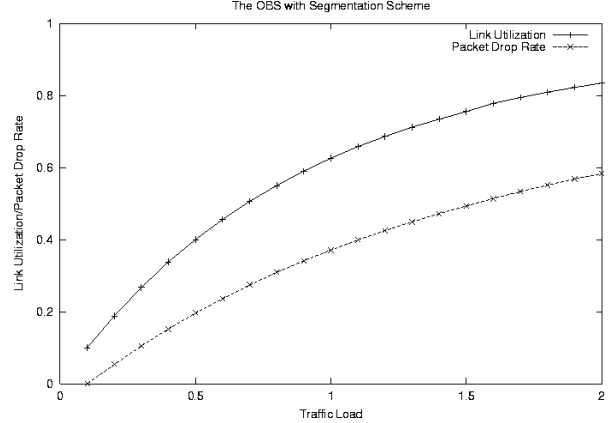


Figure 4. The link utilization and burst drop rate of the OBS with segmentation scheme.

Therefore, in this paper, we will only briefly present the operations of the TCP decoupling approach. For more detailed information, readers can refer to [1].

A TCP decoupling VC is a network path that carries traffic of multiple user flows and has its total bandwidth usage regulated by the TCP congestion control algorithm. To implement TCP congestion control for such a VC, a *management TCP connection (called management TCP for simplicity)* is set up. Packets of the management TCP are *management packets*. Since the management TCP is only for congestion control purposes, a management packet contains only a TCP/IP header and carries no TCP payload. As in a normal TCP connection, based on received ACK packets or their absence, the connection sender uses a congestion window to determine the sending rate and sending times of management packets.

Each time when the sender sends out a management packet, it is allowed to send a burst of packets to the network. The total number of bytes of these packets is a fixed number (which can be varied as a parameter for different VCs). Because there is a one-to-one relationship between a management packet and its burst, and the management packets are also fixed-sized (i.e., 40-byte TCP/IP header packets), the sending rate of management packets determines that of user packets. By this design, since the sending rate of management packets is regulated by TCP congestion control, so is that of regulated user packets. Figure 5 shows the high-level design and implementation of the TCP decoupling approach.

To make sure that management packets and user packets take the same network path, a TCP decoupling VC needs to operate on top of an MPLS [13] path or an ATM VC. This requirement is easy to meet as nowadays MPLS paths and ATM VCs are popular in the backbone networks. More operation and management issues are addressed and can be found in [1, 2].

3.2. Applied to OBS Networks

We see that the mechanism of the TCP decoupling scheme matches that of the OBS scheme very well. In both

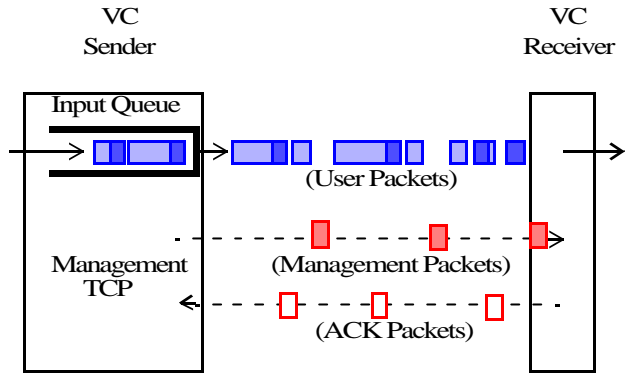


Figure 5. The high-level design and implementation of the TCP decoupling approach.

schemes, before a burst source node can send out a burst of packets, it must send out a control packet first. Although these two kinds of control packets serve different purposes, they can be combined and sent together as a single control packet in the electronic domain.

To simulate the bufferless property of an OBS switch, the maximum queue length for these control packets on their channel is set to 0. The bandwidth of this “virtual” channel is set to the optical link bandwidth divided by the packet size ratio between a burst and a control packet. For example, if a burst is 50,000 bytes and a control packet is 50 bytes and the bandwidth of the optical link is 10 Gbps, the bandwidth of the channel is set to $10 \text{ Gbps}/(50,000/50) = 10 \text{ Mbps}$. Setting such a bandwidth is to make the transmission time of a control packet on this “virtual” channel equal to its burst’s transmission time on the optical link. We will use it to simulate the bufferless property of an OBS switch and the segmentation scheme.

Suppose that the segmentation scheme is not used. When a control packet arrives at an OBS switch, the switch first checks whether the previous control packet is still being transmitted on the “virtual” channel by checking whether a transmission-time timer has expired or not. If the timer has expired, this means that the previous control packet has finished its transmission on the virtual channel (and the previous burst has finished its transmission on the optical link). In this case, this control packet’s reservation request is viewed successful and is immediately forwarded to the downstream switch as in the original OBS scheme. To simulate its transmission time on the virtual channel, the transmission-time timer now should be set up and start ticking.

Other the other hand, if the timer has not expired when the control packet arrives (this means that the previous burst has not finished its transmission on the optical link), the OBS switch will drop the control packet. Since a management packet is contained in the dropped control packet, dropping such a packet will trigger the TCP congestion control of the TCP decoupling VC, which in turn will lower the congestion level.

By setting a proper bandwidth to the virtual channel and setting the maximum queue length of the virtual queue to 0, the modified TCP decoupling approach can serve as both a signaling and a congestion control protocol for an OBS network. We note that using the modified TCP decoupling approach does not force the transmission of the first burst to be delayed by the VC’s RTT, which is avoided in the OBS scheme. The first burst can still be sent immediately because the initial congestion window size of the management TCP connection is the default one management packet.

3.3. Working with the Segmentation Scheme

To reduce the chance of burst collisions and their overlapping time, we use TCP’s famous “self-clocking” property [12] to guide the sending times of bursts. In this property, the returning rate of TCP ACK packets that return to a TCP sender reflects the bandwidth of the most congested link. This information guides the sender to choose an appropriate rate to send out its packet.

It turns out that the returning times of ACK packets also provide valuable information for senders to decide when to send out their bursts. The reason is clear. Suppose that an output port has some buffer to store packets. After being sent onto the link, two packets that arrive at the output port almost at the same time (i.e., they overlap on the time axis) will be separated by the first packet’s transmission time and not overlap any more. Since the arrival of these packets at the receiver will trigger the transmission of their corresponding ACK packets, if contending TCP decoupling VCs have the same round-trip times (RTT) and they send out their bursts only when their ACK packets come back, the newly sent bursts will not overlap with each other at the switch again. Even if there may be some processing delays that cause maintaining precise timing impossible, at least the burst overlapping time can be greatly reduced. This will greatly improve the link utilization and packet drop rate performance of an OBS switch.

To simulate the segmentation scheme, the maximum queue length of the virtual queue now is set to 1. When a control packet arrives at a switch, if the current queue length is already 1, it will be dropped. Otherwise, it will not be dropped. If the transmission-time timer has expired, the control packet is processed in the same way as in the bufferless case (i.e., sent out immediately and the timer is started). On the other hand, if the timer has not expired, the timer’s left time is stored in the control packet and the control packet is also sent out immediately. (This is because the control packet needs to configure the switches along the path. It cannot be delayed). The current queue length of the virtual queue is increased to 1 to indicate that there “should” be a control packet waiting in the virtual queue for its transmission turn. When the timer expires, if the current queue length is 1, it will decrease the length by one (to simulate dequeuing a packet

from the virtual queue) and restart itself (to simulate the transmission time of the second control packet).

When the control packet arrives at its receiver node, the receiver will delay the transmission of the corresponding ACK packet by the time stored in the control packet. Doing so will make the new bursts generated by the two contending senders no longer overlap with each other at the switch.

4. Simulation Results

Figure 6 shows the link utilization and packet drop rate performances of the modified TCP decoupling approach (working with the segmentation scheme). We modified and used the NCTUns 1.0 network simulator [14] to obtain these results. (The NCTUns 1.0 network simulator is a high-fidelity and extensible network simulator just released to the networking and communication communities. Its program code is available for download at <http://NSL.csie.nctu.edu.tw/nctuns.html>.) Each data point is an average of five runs lasting 300 seconds of simulated time.

We see that when the load is 1 the link utilization can reach about 70%. Best of all, the packet drop rate at the OBS switch now is always maintained around 1% under all load conditions.

Compared to the performances of the OBS (Figure 3) and the OBS with segmentation (Figure 4) schemes, we believe that this is a significant performance improvement. Given a 1% packet drop rate, now an OBS network is much more usable to its users.

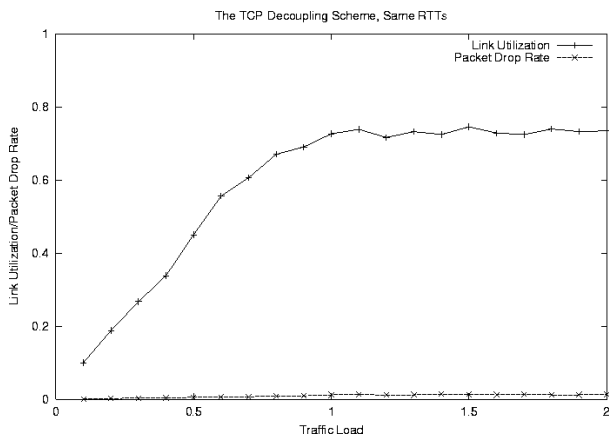


Figure 6. The link utilization and burst drop rate performance of the TCP decoupling scheme (same RTTs).

5. Conclusions

In this paper, we proposed using a modified TCP decoupling approach to improve the performances of an OBS network. This approach can congestion-control the traffic load offered to an OBS switch, thus avoiding unnecessary packet droppings. This approach also exploits TCP's famous "self-

clocking" property to regulate the sending times of contending bursts, thus resulting in a much reduced packet drop rate.

The requirement that all contending TCP decoupling VCs have the same RTTs can be easily met. For example, we can choose the maximum RTT among these VCs and let all VCs use the same RTT by artificially inserting appropriate delay before transmitting a burst.

References

- [1] S. Y. Wang, "Decoupling Control from Data for TCP Congestion Control," Ph.D. Thesis, Harvard University, September 1999. (available at <http://www.eecs.harvard.edu/networking/decoupling.html>)
- [2] H.T. Kung and S.Y. Wang, "TCP Trunking: Design, Implementation, and Performance", IEEE ICNP'99, Toronto, Canada, 1999.
- [3] S.Y. Wang and H.T. Kung, "Use of TCP Decoupling in Improving TCP Performance over Wireless Networks," Wireless Networks, Vol. 7, No. 3, pp. 221-236, 2001.
- [4] C. Qiao and M. Yoo, "Optical Burst Switching (OBS) - A New Paradigm for an Optical Internet," Journal of High Speed Networks, vol. 8, no. 1, pp. 69-84, Jan. 1999.
- [5] S. Verma, H. Chaskar, and R. Ravikanth, "Optical Burst Switching: A Viable Solution for Terabit IP Backbone," IEEE Network Magazine, Vol. 14, No. 6, pp. 48-53, November 2000.
- [6] M. Yoo and C. Qiao, "Just-Enough-Time (JET): A High Speed Protocol for Bursty Traffic in Optical Networks," IEEE/LEOS Conference on Technologies for a Global Information Infrastructure, pp. 26-27, August 1997.
- [7] I. Baldine, G. Rouskas, H. Perros, and D. Stevenson, "Jump-Start: A Just-in-Time Signaling Architecture for WDM Burst-Switched Networks," IEEE Communications Magazine, Vol. 40, No. 2, pp. 82-89, February 2002.
- [8] X. Wang, H. Morikawa, and T. Aoyama, "Burst Optical Deflection Routing Protocol for Wavelength Routing WDM Networks," Proceedings, SPIE/IEEE OPTICOM 2000, pp. 257-266, Dallas, TX, USA, October 2000.
- [9] J.S. Turner, "Terabit Burst Switching," Journal of High Speed Networks, Vol. 8, No. 1, pp. 3-16, 1999.
- [10] V. Vokkarane, J. Jue, and S. Sitaraman, "Burst Segmentation: an Approach for Reducing Packet Loss in Optical Burst Switched Networks," Proceedings IEEE, International Conference on Communication (ICC) 2002, New York, April-May, 2002.
- [11] A. Detti, V. Eramo and M. Listanti, "Optical Burst Switching with Burst Drop (OBS/BD): An Easy OBS Improvement," Proceedings IEEE, International Conference on Communication (ICC) 2002, New York, April-May, 2002
- [12] V. Jacobson, "Congestion Avoidance and Control," ACM Computer Communication Review, vol. 18, no. 4, pp 314-329.
- [13] A. Viswanathan et al., "Evolution of Multiprotocol Label Switching," IEEE Communication Magazine, Vol. 26, No. 5, pp. 165-173, May 1998.
- [14] S.Y. Wang, C.L. Chou, C.H. Huang, C.C. Hwang, Z.M. Yang, C.C. Chiou, and C.C. Lin, "The Design and Implementation of the NCTUns 1.0 Network Simulator," Computer Networks Journal (to appear), available at <http://NSL.csie.nctu.edu.tw/nctuns.html>.