A Closed-Loop Rate-based Contention Control for Optical Burst Switched Networks

Farid Farahmand§, Qiong Zhang†, and Jason P. Jue‡
§Department of Computer Electronics and Graphics Technology
Central Connecticut State University, New Britain, CT 06050
†Mathematical Sciences and Applied Computing
Arizona State University West, USA
‡Department of Computer Science
The University of Texas at Dallas, Richardson, TX 75083
email: {ffarid, qzhang77, jjue}@utdallas.edu

Abstract—In this paper we describe a rate-based OBS network architecture in which core switch nodes send explicit messages to edge nodes requesting them to reduce their transmission rate on congested links. Within this framework, we introduce a new contention avoidance mechanism called proportional control algorithm with explicit reduction request (PCwER). Through source rate control, PCwER proactively attempts to prevent the network from entering the congestion state. Basic building blocks and performance trade-offs of PCwER are the main focus of this paper. In addition, through a simple fluid model we analyze the characteristics of the algorithm. Our simulation results show that the proposed contention avoidance techniques improve the network utilization and reduce the packet loss probability.

Index Terms—Admission control, contention resolution, feedback control, optical burst switching, rate-based control.

I. INTRODUCTION

Optical burst switching (OBS) [1] has been proposed as a new paradigm to provide the flexible and dynamic bandwidth allocation required to support highly dynamic and bursty traffic. In OBS networks, incoming data is assembled into basic units, referred to as data bursts, which are then transported over the optical core network. Control signaling is performed out-of-band by control packets, which carry information such as the length, the destination address, and the QoS requirements of the optical data burst. OBS provides dynamic bandwidth allocation and statistical multiplexing of data. By aggregating packets into large sized bursts and providing out-of-band signaling, OBS eliminates the complex implementation issues of optical packet switching. For example, no optical buffers are necessary at core nodes, headers can be processed at slower speeds in the electronic domain, and synchronization requirements are relaxed in OBS. On the other hand, due to packet aggregation, OBS incurs higher end-to-end delay and higher packet loss per contention than optical packet switching.

A major concern in OBS networks is high contention and burst loss due to data channel contention, which occurs when the total number of data bursts going to the same output port at a given time is larger than the available channels on that port. Contention is aggravated when the traffic becomes bursty and when the data burst duration varies and becomes longer. Contention and loss may be reduced by implementing contention resolution policies, such as time deflection (using buffering [2]) and space deflection (using deflection routing [3]). When there is no available unscheduled channel, and a contention cannot be resolved by any one of the above techniques, one or more bursts must be dropped. The policy for selecting which bursts to drop is referred to as the dropping policy and is used to reduce the overall burst loss rate, BLR, and consequently, to enhance link utilization. Several dropping algorithms have been proposed and studied in earlier literature, including the shortest-drop policy, segmentation, and look-ahead contention resolution [4].

These dropping policies are considered as reactive approaches in the sense that they are invoked after contention occurs. An alternative approach to reduce network contention is by proactively attempting to avoid network overload through traffic management policies. Consequently, contention avoidance policies attempt to prevent a network from entering the congestion state in which burst loss occurs.

In a feedback-based network, one way to avoid contention is by dynamically varying the data burst flows at the source to match the latest status of the network and its available resources. Thus, as the available network resources are changed, a source should vary its offered load to the network, accordingly. Two critical issues in any network with feedback mechanism are determining what type of information must be conveyed to the source and interpreting the conveyed information and reacting to the current network state. We refer to these issues as signaling strategies and control strategies, respectively.

In the past two decades, numerous studies have been dedicated to designing and analyzing contention avoidance (or congestion control) mechanisms in TCP and ATM networks. Many different protocols have focused on the signaling strategies. For example, in congestion control approaches, such as DECbit [5], a single bit in the packet header explicitly notifies the source about the congestion in downstream nodes. Other feedback protocols such as eXplicit Control Protocol (XCP) [6], utilize multi-bit feedback messaging, which explicitly indicates the degree of congestion at the bottleneck.

There are also many proposals which focus on adjustment algorithms, including Binomial congestion control algo-
In this paper we propose a new rate-based congestion avoidance mechanism for bufferless OBS networks where multi-bit explicit feedback signaling is sent to each edge source node indicating the required reduction in the burst flow rate going to the congested link. We refer to such feedback-based contention avoidance as proportional control algorithm with explicit reduction request (PCwER). In this scheme, during the underload periods, the rate of transmission increases additively (AI), whereas during congestion period, the sending rate decreases multiplicatively (MD). Our proposed contention avoidance mechanism utilizes OBS network characteristics, and it differs from previous proportional rate-based algorithms based on the following four assumptions: (1) feedback information reflects the actual load level (or loss rate) at the congested link; (2) there is no queuing delay on intermediate nodes, and link propagation delays are known to all nodes; (3) the feedback signal specifically notifies the source by how much it should reduce its rate to match the targeted congestion level of the network; (4) the feedback signal is transmitted to the source from the bottleneck switch node, rather than the destination nodes as in end-to-end contention avoidance mechanisms.

The architectural details of our proposed feedback mechanism in OBS networks are also described in this paper. We explain how feedback signals can be framed within label-switched OBS networks. Through a simple fluid model we analyze convergence and evaluate the fairness of PCwER. In addition, by means of simulation, we examine the performance of the PCwER contention avoidance mechanism under specific network conditions. We compare our results to the case without source traffic control in terms of blocking probability and network throughput.

The rest of this paper is organized as follows. In Section II, we briefly describe the basic blocks and architecture of the label-switched feedback-based OBS network. In Section III we elaborate on details of our proposed contention avoidance algorithm. In section IV, we analyze behavior of PCwER. Finally, in Section V we present performance results obtained by means of simulations, followed by concluding remarks in Section VI.

II. NETWORK ASSUMPTIONS AND NODE ARCHITECTURE

The OBS network consists of \(|N|\) nodes and \(|L|\) links represented by sets \(N = \{1, 2, ..., n\}\) and \(L = \{(1, 2), \ldots (j, k)\}\), respectively, where \(j, k \in N\). Each link is characterized by the number of wavelength channels it carries, \(W\), and the capacity of each channel, \(S\). Each edge node determines the source-destination route, \(R(s, d)\), and has sufficiently large buffers in order to store incoming packets due to network congestion and transmission latency. On the other hand, switch nodes are bufferless and, hence, upon link congestion, data bursts will be dropped. Furthermore, we assume that each intermediate core node \(n\) knows the set of source nodes that are contributing to the traffic load on an egress link \((j, k)\), \(\Lambda_{(j,k)}^n\), and all nodes have full knowledge of propagation delay between each source-destination node pair, \(T(s, d)\).

Without loss of generality, we consider label-switched OBS networks using a Generalized Multi-Protocol Label Switching (GMPLS) control plane. In this model, the transmitted bursts are routed through individual Label Switch Paths (LSPs). We assume that the intermediate core nodes have no buffering capacity, and that incoming LSPs can either cut through the core nodes or be blocked. When the measured load on an egress port exceeds a predefined load threshold, the congested core node sends back a flow-rate reduction request (FRR) signal to ingress edge nodes requesting them to reduce the transmission rate of LSPs sharing the congested link. The feedback signaling to the source nodes can be implemented using the Label Distributed Protocol (LDP) employed in GMPLS.

The value of FRR, denoted by \(R_{j,k}\), indicates the actual rate reduction value required by the switch on link \((j, k)\). In the rest of this paper we use the term LSP and burst flow interchangeably.

III. OBS RATE-BASED CONTENTION AVOIDANCE ALGORITHM

The basic idea in the proportional control algorithm with explicit reduction request (PCwER) is that each core node measures the received load on each of its egress links (reflecting loss probability) and reports that to edge nodes. Based on feedback information, each source increases or decreases its transmission rate. We first describe the signalling strategy and then examine the rate adjustment (admission control) mechanism and algorithm.

A. Signalling Strategy

In the PCwER contention avoidance mechanism, each core node maintains the load information on each of its egress link, \((j, k)\), denoted by \(\rho_{j,k}\). This is calculated by measuring the duration of all incoming data bursts (unscheduled and scheduled) destined to egress link \((j, k)\), over some fixed control interval \(\Delta\). If the measured load on the egress link is greater than some predefined load threshold, \(\rho_{TH}\), then a flow-rate reduction request (FRR) will be generated. We refer to such a link as being congested. The value of FRR explicitly indicates the percentage by which edge nodes must reduce the transmission rate of all burst flows (or LSPs) sharing link \((j, k)\) in the immediate future, and it is equivalent to

\[
R_{j,k} = (\rho_{j,k} - \rho_{TH}) / \rho_{j,k}; \quad \rho_{j,k} \geq 0.
\]  

When \(R_{j,k}\) is set to zero, it indicates that no further change of transmission rate must be allowed. Whereas, \(R_{j,k} = -1\) indicates that the source can increase its rate of transmission.

Assuming that the measured traffic load on a link is around \(\rho_{TH}\), any small changes in the offered load by the source on that link can result in FRR oscillation. One way to prevent this is by setting a lower and upper threshold such that \(\rho_{UL} = \rho_{TH} - \varepsilon\) and \(\rho_{UL} = \rho_{TH} + \varepsilon\), where \(\varepsilon\) is a small percentage of \(\rho_{TH}\). Hence, the source will not be permitted to change its traffic to the near-congested link unless the measured load drops below \(\rho_{UL}\) or rises above \(\rho_{UL}\).

When a switch node \(n\) is overloaded (\(\rho_{j,k} \geq \rho_{TH}\)), the node will send a reduction request, \(R_{j,k} > 0\), and no new FRR for link \((j, k)\) will be sent until 2 RTT time units later. This is to ensure that the change has actually taken place. However, the actual control interval in which the switch measures the average load is limited to one control interval, \(\Delta\). Therefore, as long as \(R_{j,k} \leq 0\), FRR will be sent once every \(\Delta\) time units. We note that the value \(\Delta\) is determined based on the largest RTT value between node \(n\) and all other competing sources transmitting into node \(n\).
B. Rate Controller Mechanism

The basic rate adjustment mechanism in PCwER is as follows. Upon receiving a negative FRR, the source will increase its rate of transmission of data bursts on link \((j,k)\), \(\phi_{j,k}^{\Delta} \), within the next control interval \(\Delta\) by some fixed unit

\[
\phi_{j,k}^{\Delta} = \phi_{j,k}^{\Delta-1} + IR.
\]

(2)

On the other hand, if \(R_{j,k} > 0\), then the sending rate decreases as follows:

\[
\phi_{j,k}^{\Delta-1} = \phi_{j,k}^{\Delta-1}(1 - R_{j,k}).
\]

(3)

If \(R = 0\) then \(\phi_{j,k}^{\Delta-1} = \phi_{j,k}^{\Delta+1}\). In the above expressions \(IR\) is constant and is called the increase rate increment. On the other hand, \(R_{j,k}\) is a function of time and changes for each control interval \(\Delta\).

In practice, the data burst transmission rate is adjusted by the data burst interdeparture time \(\Delta\) at time \(t\) algorithm. Upon receiving the reduction rate request on link \(R\) hand, as follows:

\[
\phi_{j,k}^{\Delta} = \phi_{j,k}^{\Delta-1}(1 - R_{j,k}).
\]

On the other hand, if \(R_{j,k} > 0\), then the sending rate decreases as follows:

\[
\phi_{j,k}^{\Delta-1} = \phi_{j,k}^{\Delta-1}(1 - R_{j,k}).
\]

If \(R = 0\) then \(\phi_{j,k}^{\Delta-1} = \phi_{j,k}^{\Delta+1}\). In the above expressions \(IR\) is constant and is called the increase rate increment. On the other hand, \(R_{j,k}\) is a function of time and changes for each control interval \(\Delta\).

In practice, the data burst transmission rate is adjusted by the data burst interdeparture time \(\Delta\) at time \(t\) algorithm. Upon receiving the reduction rate request on link \(R\) hand, as follows:

\[
\phi_{j,k}^{\Delta} = \phi_{j,k}^{\Delta-1}(1 - R_{j,k}).
\]

C. Rate adjustment algorithm

Upon receiving multiple FRR messages from different links, the edge node determines the most congested link \((j,k)\) along each source-destination path and subjects all data bursts (or LSPs) passing through the congested link to a rate adjustment according to the increase/decrease functions described above. We assume that each edge node keeps track of the latest values of the following parameters: average transmission rate along each link, \(\phi_{j,k}\), the latest reduction request, \(R_{j,k}\), and its corresponding request time, \(RT_{j,k}\).

We now describe details of our proposed rate-based control algorithm. Upon receiving the reduction rate request on link \((j,k)\) at time \(t1\) \((R_{j,k}^{1})\), the edge node \(n\) takes the following actions:

(a) If \(R_{j,k}^{1} = 0\), continue transmitting at the current rate
(b) If \(R_{j,k}^{1} < 0\), increase the transmission rate of all \((s,d)\) flows where \((j,k) \in R(s,d)\), according to the increase function if:

- The most congested link on \(R(s,d)\) is \((j,k)\) and \(\phi_{j,k}^{s,d} + IR \leq \phi_{m,n}(1 - R_{m,n})\), where \((m,n)\) is the next most congested link on \(R(s,d)\)
- \(\rho_{j,k} + IR < \rho_{TH} \cdot (S \cdot W)\)

(c) If \(R_{j,k}^{1} > 0\), check the value and time of the last FRR message, \(R_{j,k}^{t0}\) and \(t0\), respectively:

- If \(R_{j,k}^{t0} < R_{j,k}^{1}\) and \(t1 < 2 \cdot RTT + t0\), ignore the incoming \(R_{j,k}^{1}\)
- If \(R_{j,k}^{t0} > R_{j,k}^{1}\) and \(t1 < 2 \cdot RTT + t0\), then decrease the transmission rate by \(R_{j,k}^{t1} - R_{j,k}^{t0}\)
- If \(t1 > 2 \cdot RTT + t0\), reduce the transmission rate of all \((s,d)\) flows to \(\rho_{s,d} = \rho_{j,k}[1 - R_{j,k}^{t1}]\) where \((j,k) \in R(s,d)\)

Clearly, each time a new FRR is received and the rate of transmission is changed, all records must be updated accordingly.

We illustrate the above concepts using the example shown in Fig. 1 where nodes S1, S2, and S3 are sending data bursts to Node S5 and S4 is the bottleneck node. We assume the system is at equilibrium and rate of transmission is constant. As shown in the timing diagram in Fig. 1, S1-S3 send their data bursts at different instances, namely \(t1, t2,\) and \(t6\), respectively. At time \(t3\), S4 detects congestion on the link between S4 and S5 and requests a reduction of 22%. Once S1 and S2 receive the new FRR, they reduce their sending rate accordingly. The average value measured at \(t9\), which is one \(RTT\) later \((t3 + RTT = t9)\), is ignored by S4. At \(t9\) a new averaging starts, and \(RTT\) time units later \((t13)\), another FRR signal is generated indicating the latest average measured load and sent back to S1-S3. After \(2RTT\), \((t13 + 2 \cdot RTT)\) all nodes are expected to reduce their transmission to meet the target load value requested by S4. At that time, the FRR is expected to be set to zero, indicating no further change in transmission rate is required. Note that after receiving and implementing the first FRR, source nodes ignore any other reduction request that involves the link between S4 and S5 until 2 \(\cdot RTT\) later.

IV. ANALYSIS

In order to analyze our proposed rate-based contention control model for OBS network, we consider a continuous-space deterministic (or fluid) model [8] as shown in Fig. 2(a). We use this model to address four important issues: (1) to determine how fast the transmission rate should increase when the system is underloaded; (2) to find the worst case instantaneous probability of loss at equilibrium when the desired system load setpoint, \(\rho_{TH}\), is given; (3) to find the convergence time for the system to approaches \(\rho_{TH}\); (4) and to determine how fairly the bandwidth
is distributed between different competing sources.

A. Algorithm Convergence

In our model we assume the bottleneck is a multi-server bufferless queueing system with \( W \) individual servers, each with a service rate of \( S \) bursts/second, and we let \( X(t) \) denote the aggregated transmission rates from all source nodes in bursts/second at time \( t \). Furthermore, we assume \( D_{FW} \) and \( D_{FB} \) represent the propagation delays from the source to the bottleneck node with the congested link and from the bottleneck node to the source, respectively. Due to the bufferless nature of our system, no queuing delay is applied to our model.

Under continuous-space assumption, we can redefine the rate increase/decrease algorithm as follows:

\[
X(t) = \begin{cases} 
X(t_0) + a \cdot (t - t_0) & \text{if increasing} \\
X(t_0) \cdot b(t - t_0) & \text{if decreasing}
\end{cases}
\]  

where \( t \) is the current time and \( t_0 \) is the time the FRR signal is sent to the source. Relating the above continuous expressions and Eqn. (2)-(3) we allow \( a = IR \cdot \Delta \) and \( b \) to be a continuous decreasing function equivalent to \( 1 - R_{j,k} \). Since we are interested in the worst case loss rate, the actual function of \( b(t) \) is not critical.

In this section we only consider the equilibrium condition when no new flow of bursts is added and no active flow is terminated. In addition, we assume that there is only one congested link.

The behavior of \( X(t) \) as a function of time and the corresponding loss rate are shown in Fig. 2(b)-(c). The maximum loss will occur when the transmission rate reaches its maximum level at \( t_1 \), as shown in Fig. 2(b). Hence, we are interested to find \( X_{max} \). From Fig. 2(b) it is clear that the elapsed time between when the feedback signal (FRR) is generated and the time the transmission rate reaches its maximum level, \( X_{max} \), is \( D_{FW} + D_{FB} \). We assume that the propagation delays between all nodes are the same, \( D = D_{FW} + D_{FB} \). Therefore, the maximum arrival rate received by the bottleneck node will be

\[
X_{max} = X_{ref} + a \cdot D.
\]

Consequently, the maximum experienced load at equilibrium state with link fluctuation around \( \rho_{TH} \), will be \( \rho_{max} = (X_{max}/S \cdot W) \)

Using the well-known Erlang-B formula, the burst loss probability can be calculated as

\[
P_{loss}(\rho) = \frac{\rho^W/W!}{\sum_{k=0}^W \rho^k/k!}.
\]

Consequently, if the arrival rate is \( X_{max} \), the maximum burst loss rate (in bursts/second) can be expressed as \( Q_{max} = P_{loss}(\rho_{max}) \cdot X_{max} \). The percentage difference of the maximum loss rate from its target value can be expressed as \( \delta_{ref} = (Q_{max} - Q_{ref})/Q_{max} \), where \( Q_{ref} = P_{loss}(\rho_{TH}) \cdot X_{ref} \) and \( X_{ref} = \rho_{TH} (S \cdot W) \).

Using the above relationships, we can see that, given the target loss rate and its maximum acceptable instantaneous fluctuation, \( Q_{ref} \) and \( \delta_{ref} \), respectively, we can determine the values of \( \rho_{TH} \) and \( IR \).

As shown in Eqn. (5), the maximum loss rate is tightly related to \( IR \) and the round-trip delay. Larger values of \( IR \) result in faster convergence to the target link load, and hence, higher throughput. The trade off, however, is a higher maximum loss rate. A closer look at Fig. 2(b) shows that, under the equilibrium condition, the system approaches the link load threshold, \( \rho_{TH} \) after \( 2 \cdot D \), where \( D \) is equivalent to one round-trip delay.

B. Algorithm Fairness

Fairness is considered to be an important issue in any rate-based contentention avoidance network with feedback. A widely adopted criterion to define fairness is known as maximum fairness criterion. In this scheme, the traffic flows from different edge nodes with the same priority must have an equal share of the congested link, \( S \cdot W/|N| \). Such a property is quantified by a fairness index defined as follow:

\[
F1(X) = \frac{(\sum X_i)^2}{|N| \cdot (\sum X_i^2)},
\]

where \( |N| \) is the number of concurrent flows into the congested link and \( X_i \) is the sending rate of the \( i \)th flow at equilibrium. \( F1 \) is a value between 0 and 1 with \( F1 = 1 \) indicating perfect fairness Based on this definition, it can be seen that, using PCwER when there is no congestion in the network and transmission rate is increasing linearly, \( F1 \) tends to increase and approach unity: \( 1 \geq F1(X + \alpha) \geq F1(X) \). On the other hand, when link \( (j,k) \) is congested and edge nodes must decrease their sending rate, the value of \( F1 \) does not change: \( F1(X - X \cdot R_{j,k}) = F1(X) \). These results indicate that our proposed rate-based adjustment algorithm stabilizes the total load around the desired target value \( (\rho_{TH}) \), and it does not change the initial ratio of offered loads by different competing edge nodes in the OBS network due to congestion.

V. PERFORMANCE RESULTS

In this section we discuss the simulation results obtained by implementing the proposed data burst admission control in a feedback-based OBS network. We consider the simple network topology, shown in Fig. 3(a), as our test network. In this network we assume the RTT delay between each node pairs is different and is between 10-50 ms. We also consider the following assumptions for the simulation environment: burst length is fixed and is equivalent to 100 \( \mu s \); containing 1250 bytes; the transmission rate is 10 Gb/s with 4 wavelengths on each link; the switching time is 10 \( \mu s \); and the burst header processing time at each node is assumed to be 2.5 \( \mu s \). The design parameters for the
PCwER congestion avoidance mechanism are as follows: $IR = 1.1$, $p_{T_H} = 0.7$ and $p_L$ and $p_H$ are 0.695 and 0.705 respectively.

We consider two traffic models for our simulation. In the first case we assume Poisson arrivals with fixed size data bursts. In the second case we assume that average arrival rate varies between $\lambda_H$ and $\lambda_L$ according to a two-state Markov modulated arrival process as shown in Fig. 3(b). We define $\alpha = \lambda_H/\lambda_L$ as the traffic persistency factor. Note that if $\alpha = 1$, we obtain a Poisson arrival model, and as $\alpha$ increases, the traffic becomes more bursty.

Fig. 4 shows the probability of burst loss for Poisson arrivals.

This figure compares the probability of data burst loss with and without the PCwER congestion avoidance mechanism. As the load threshold in the switch drops, the loss probability decreases. This occurs as a result of choking the source and lowering the maximum transmission rate allowed on the bottleneck link. When the total load at the bottleneck link reaches $p_{T_H}$, a slight increase in loss rate is experienced, which is due to the RTT. The trade-off of reducing the overall loss rate due to rate-control is lowering the link throughput, as shown in Fig. 5. This figure shows the normalized throughput for an exponentially distributed traffic model with and without the contention avoidance mechanism. The maximum achievable data link capacity in our model is 40Gb/s. The value of the load threshold of the switch directly impacts the network throughput. For example, as shown in Fig. 5, when the threshold is set to 0.7, the throughput of the bottleneck link at high loads will be 0.59 (40 Gb/s) = 23.6 Gb/s, compared to 0.81 (40 Gb/s) = 32.4 Gb/s when no contention avoidance is implemented. However, the loss at $p_{T_H} = 0.7$ is significantly lower. Note that as long as the measured load remains above the threshold, the system stays in continuous choking state. Similar results in terms of data burst loss probability and link throughput can be observed when the traffic is Poisson arriving with high and low arrivals. Fig. 6 shows that when the threshold value is low, such as 0.6, as the measured load on the bottleneck reaches the threshold, the probability of loss continues to increase until the links are overloaded and the system goes into the choke state.

VI. CONCLUSION

In this paper we proposed a rate-based contention avoidance mechanism for optical burst switching networks. Our proposed scheme, the proportional control algorithm with explicit reduction request (PCwER), significantly reduces the packet loss probability in the OBS network. The basic trade-off of PCwER is, however, the overall reduction of network utilization due to invoking admission control when the network is congested. We showed that network throughput reduction, due to rate control, is tolerable.

REFERENCES