Fair Bandwidth Allocation in Optical Burst Switching Networks

Tananun Orawiwattanakul, Yusheng Ji, *Member, IEEE*, Yongbing Zhang, *Member, IEEE*, and
Jie Li, *Senior Member, IEEE*

Abstract—**Optical burst switching (OBS) is a promising switching technology for next-generation Internet backbone networks. One of the design challenges is how to provide fair bandwidth allocation in OBS networks; the schemes proposed for general store-and-forward IP switching networks can not be used because of the non-buffering and un-fully utilized bandwidth characteristics of OBS networks. We propose a rate fairness preemption (RFP) scheme to achieve approximately weighted max-min fair bandwidth allocation in OBS networks. We present an analysis of the burst loss probability in RFP-based OBS networks. The analysis and simulation results show that the RFP scheme provides fair bandwidth allocation in OBS networks.**

Index Terms—**Fair bandwidth allocation, optical burst switching (OBS), wavelength preemption, weighted max-min fairness.**

I. Introduction

OPTICAL BURST switching (OBS) is a promising technology for future Internet backbone networks, and it is expected to become popular in the near future. The current Internet relies on end-to-end congestion control protocols, e.g., TCP/IP. Although end-to-end congestion control mechanisms can protect a network from collapsing, most of them cannot provide service isolation and protection among flows to prevent well-behaved flows from being degraded by misbehavior from other flows. We define the term "flow" in this paper to be a connection between the same pair of ingress and egress edge switches. To provide service isolation and protection, fair bandwidth allocation (FBA) is required. An FBA scheme fairly allocates network resources to all flows. If any of the flows send traffic that is higher than its optimum bandwidth allocation, an FBA scheme can prevent quality degradation, e.g., losses and throughput, in other flows that share the same link. Many studies, e.g., [1]–[4], have tried to achieve FBA in general store-and-forward IP packet switching networks. Nevertheless, OBS networks are different from conventional store-and-forward IP networks. While most studies on store-and-forward IP networks deal with electronic queue-based mechanisms, neither electronic nor optical buffers exist for the data channels in OBS core networks. In addition, the bandwidth of OBS networks cannot be fully utilized because of the buffer-less characteristics and voids between bursts. Consequently, schemes to achieve FBA in store-and-forward IP networks cannot be directly implemented in OBS networks.

Several approaches have been proposed to ensure quality-of-service (QoS) in OBS networks [5]–[7]. However, these studies do not adequately deal with service isolation and protection issues. One way to achieve FBA in OBS networks is rate control, as in the integrated congestion control mechanism (ICCM) [8], the differentiated available bit rate (D-ABR) mechanism [9], and a scheme based on peak load [10]. The input traffic of each flow is adjusted in accordance with the allocated rate. However, rate control may decrease network utilization. Another way to achieve FBA is preemption. Preemption is effective because it does not decrease network utilization. To the best of our knowledge, only the max-min fairness preemption (MMFP) scheme [11] and our previous work, the rate fairness preemption (RFP) scheme [12], have a mechanism based on preemption to provide FBA in OBS networks. In both schemes, bandwidth is allocated to each flow based on the max-min fairness [13]. The edge switch allows each flow to send traffic even if it is over its max-min rate but misbehaved flow bursts are dropped in cases of contention. Nevertheless, MMFP imposes a high workload on the core networks because core switches need to monitor both the arrival and loss rates of each flow. In addition, MMFP lacks of global situation awareness that takes into account the state of the network. Each MMFP-based core switch independently allocates the max-min rate and performs preemption without considering the optimum max-min rate along the path, which may cause it to have inefficiency under some circumstances, e.g., bottleneck links that increase total loss.

In this paper, we extend our previous work in [12] and propose a rate fairness preemption (RFP) scheme to achieve fair bandwidth allocation (FBA) based on the weighted max-min fairness in OBS networks. We aim to achieve fairness among flows, classified by the ingress and egress edge switches' addresses. RFP uses a burst marking strategy to assign a preemptive priority to bursts, and uses a preemption technique to drop low preemptive-priority bursts only in the case of contention. We improve the adaptive max-min rate allocation method proposed in [12] to be less affected by the control message updating process. We also develop the fixed max-min allocation method in this paper in addition to the adaptive one proposed in [12]. The RFP fixed max-min rate allocation is the simplest scheme because it does not require additional control messages, as required by [8]–[11]. Our scheme does not impose a high work load on the core network because the core switches perform the max-min rate allocation process only when the input traffic

significantly changes and they do not need to monitor the arrival rate of each flow, as is needed in most of FBA schemes. Moreover, RFP allocates the max-min rate by considering the optimum rate along the path. Consequently, RFP can solve the global situation awareness problem of MMFP.

We describe the RFP scheme in Section II. Sections III and IV respectively present analysis and simulation results. Our discussion and conclusion are presented in Sections V and VI, respectively.

II. RATE FAIRNESS PREEMPTION

OBS networks consist of edge and core switches. The wavelengths are classified into two groups: control and data channels. The ingress edge switch assembles incoming traffic, e.g., IP packets, ATM cells, and Ethernet frames, with the same egress address into a burst, called a data burst (DB). The ingress edge switch sends out a burst control packet (BCP) over the control channels to set up the connection for the corresponding DB. The BCP carries the DB information, e.g., DB arrival time, burst length, and incoming wavelength. After an offset time, the ingress edge switch sends the corresponding DB over the data channels. Upon arrival at the intermediate core switches along the path, the BCP is converted into an electronic signal for the DB bandwidth reservation process. If bandwidth reservation is successful, when the corresponding DB arrives at the core switch, it will be routed over the reserved channel. In contrast, if the bandwidth reservation fails, the core switch will drop the DB. The core switch cannot delay the DB because there are neither electronic nor optical buffers for the DB channels. We assume that the route is selected based on the static routing. Moreover, it is assumed that the network does not implement fiber delay lines (FDLs) and the network has a full wavelength conversion capability.

We describe the parameters used in RFP in Table I. To support the RFP scheme, we modify the conventional OBS networks as follows.

- The field format of the BCP is modified to contain a parameter R indicating the burst type (under-rate or over-rate).
- Losses in OBS networks are high from the non-buffering characteristic, and there are voids between bursts that make the link capacity unable to be fully utilized. Therefore, well-behaved flows need to be protected even when the input rates are less than the link capacity. We modify the max-min fairness bandwidth allocation criteria originally proposed in [13]. The modification is the same as proposed in [8] and [9], whereby the core switch does not use the actual link capacity (C) to calculate the fairly allocated bandwidth F but uses the effective link capacity $(e \times C)$ instead. Network operators have to decide the value of e to achieve an acceptable burst loss level. The proper setting for e is proposed in [9], where one initially finds the maximum arrival rate $(\lambda_{\max})$ that results in an acceptable burst loss probability for well-behaved flows (P_{loss}) by using the Erlang formula shown in (6). Then, set e equal to $\lambda_{\max}/\mu$, where $1/\mu$ is the average burst service time. The modified max-min fairness of our study is called "OBS max-min fairness."

TABLE I
PARAMETERS USED IN RFP

Parameters	Description
i	Flow id
W_i	Weight max-min fairness of flow i
A_i	Arrival rate of flow i at ingress switch
F_i	Fairly allocated bandwidth of flow i
T_i	Allocated max-min rate of flow i
B_i	Progressive filling algorithm max-min rate of flow i
M_i	Adaptive max-min rate of flow i
S	Threshold for change of arrival rate
e	Effective capacity ratio
R	Rate priority parameter
C	Link capacity
$P_{i\text{-}O}$	Over-rate burst marking probability
$P_{i\text{-}U}$	Under-rate burst marking probability

The RFP scheme consists of three main parts: 1. Max-min bandwidth allocation 2. Burst marking, and 3. Preemption policy. The scheme is described as follows.

A. Max-Min Bandwidth Allocation

We propose two methods to assign the allocated max-min rate (T_i) for each flow, as follows.

1) Fixed Max-Min Rate Allocation: The allocated max-min rate (T_i) is fixed by the network operator. This method does not require any additional control messages. Therefore, it is the simplest way to achieve fair bandwidth allocation in a practical implementation. In this study, the progressive filling algorithm (PFA) is used for the fixed max-min rate allocation. The max-min rates of all flows begin at zero, and they grow together at the same pace until the total input traffic in one or several links reaches the effective link capacity $(e \times C)$. The PFA stops increasing the max-min rates for the sources that use these full-capacity links (bottleneck links) and continues increasing the rates for other flows. The PFA repeats this rate increasing process until all flows have bottleneck links. Note that the PFA allocates the max-min rate by assuming that all flows have enough traffic to send and seek to maximize throughput. However, in practical networks, some flows may transmit traffic under the max-min rate. Therefore, the fixed max-min rate allocation is not efficient enough to be used in the rate controls proposed in [8]–[10] because it may decrease bandwidth utilization. In contrast, RFP is a preemption scheme that does not limit the input traffic and the transmitted traffic over the max-min rate is dropped only in the case of contention. Consequently, fixed allocation is efficient in RFP-based OBS networks. The B_i parameter represents the max-min rate of flow i based on the PFA. Therefore,

$$T_i = B_i. \tag{1}$$

2) Adaptive Max-Min Rate Allocation: The allocated max-min rate (T_i) is allocated according to the actual traffic in the networks. We propose an RFP control plane to adaptively allocate the max-min fairness rate of flow $i(M_i)$ in accordance with the current traffic in the network. We modify the conventional OBS networks to support the RFP control plane as follows.
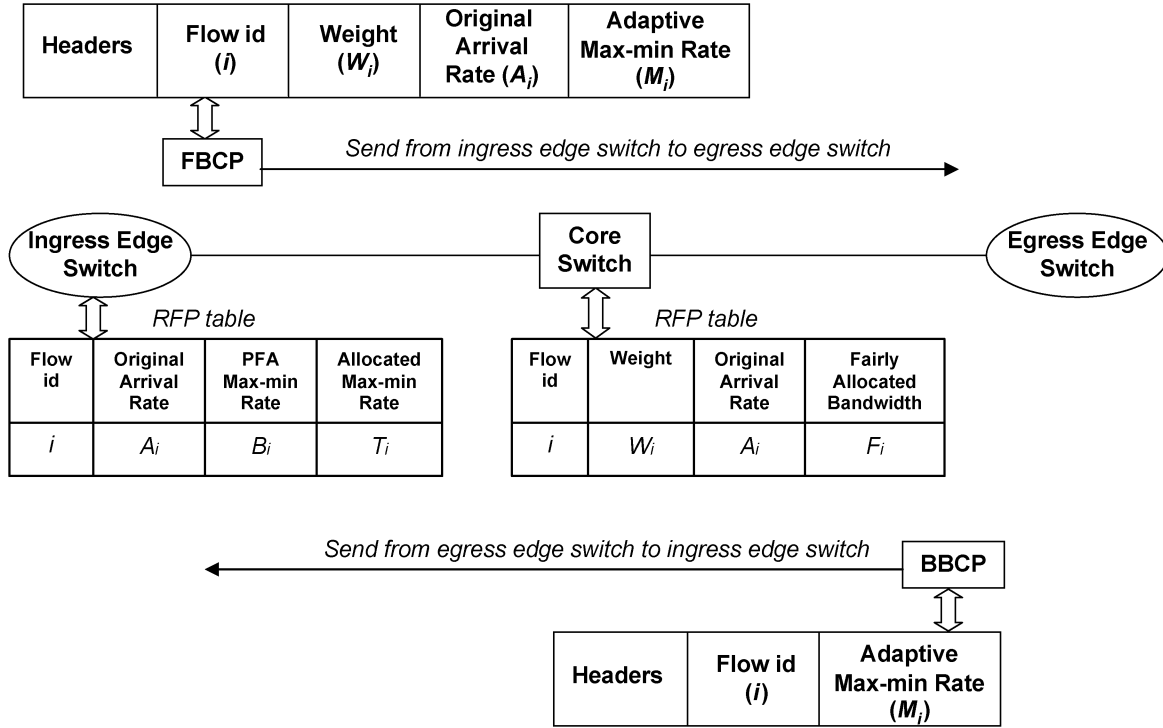
Fig. 1. Parameters contained in FBCP, BBCP, and RFP table of edge and core switches.

- Two control packets, Forward Bandwidth Control Packet (FBCP) and Backward Bandwidth Control Packet (BBCP), are created in addition to the Bandwidth Control Packet (BCP). The field formats of FBCP and BBCP are shown in Fig. 1.
- Edge and core switches maintain an RFP table. The parameters in the RFP table of edge and core switches are presented in Fig. 1. Note that B_i in the edge switch's RFP table is the max-min rate of flow i calculated from the PFA described in Section II-A-1.

First, each ingress edge switch monitors the arrival rate of each flow (A_i) by using the rate estimation scheme, e.g., the exponential moving average of the arrival rate. If the arrival rate change within a unit time is greater than S, the ingress edge switch generates and sends an FBCP to the egress edge switch over the control channel. The ingress switch sets the parameters in the FBCP packet as follows: i is the flow id, W_i equals to the max-min weight, A_i equals to the measured arrival rate of flow i, and $M_i = A_i$.

When an intermediate core switch receives the FBCP of flow i, it updates A_i and W_i in its RFP table. Then, it calculates the fairly allocated bandwidth (F) for all flows with the same outgoing link as flow i and updates the F parameter in the RFP table. The parameter F_i is the max-min rate of flow i based on the OBS max-min fairness in a single outgoing link, while the parameter M_i in FBCP is the optimum max-min rate of flow i along the path. If we assume that flow i traverses N intermediate core switches and F_i^k is the F_i calculated by core switch k, $M_i = \min\{F_i^k | 1 \leq k \leq N\}$. Therefore, if $F_i < M_i$, the core switch sets $M_i = F_i$ before forwarding the FBCP to the next switch. The switch also generates the FBCP of flow j, where j

is the set of flows whose new F parameter changes, and sends the FBCPs to the destinations.

When the egress edge switch receives the FBCP of flow i, the switch generates the BBCP for flow i's ingress edge switch. The BBCP contains the parameter i and M_i (same as M_i in the FBCP). Upon receiving the BBCP, if $M_i \geq B_i$, the switch updates T_i in its RFP table to be equal to M_i. In contrast, if $M_i < B_i$, T_i is equal to B_i. Therefore,

$$T_i = \max(M_i, B_i). \tag{2}$$

B. Burst Marking

We modify the BCP so that it contains the parameter R. The ingress edge switch sets R equal to 0 and 1 to indicate under-rate bursts and over-rate bursts, respectively. The ingress edge switch measures the arrival rate of all the flows. When $A_i > T_i$, the ingress edge switch marks flow i's bursts as over-rate or under-rate with probabilities P_{i-O} and P_{i-U}, respectively. In contrast, when $A_i \leq T_i$, the ingress edge switch marks all flow i's bursts as under-rate. P_O and P_U of flow i are calculated as follows:

$$P_{i-O} = \begin{cases} (A_i - T_i)/A_i, & \text{when } A_i > T_i \\ 0, & \text{when } A_i \leq T_i \end{cases} \tag{3}$$

$$P_{i-U} = 1 - P_{i-O}. \tag{4}$$

Note that in the adaptive max-min rate allocation, B_i is the minimum max-min allocated rate if flow i always has traffic to transmit. If $A_i < B_i$, flow i is well-behaved. Therefore, instead of relying on the parameter M_i, which is up to the FBCP and BBCP updating process depending on the S parameter and

Parameters	Description
n	Flow id
N	Total number of flows in the system
μ	Service rate
K	No. of wavelengths in output port
λ_n	Arrival rate of flow n
λ_{n-U}	Arrival rate of flow n's under-rate bursts
λ_{n-O}	Arrival rate of flow n's over-rate bursts
P_{i-O}	Over-rate burst marking probability
P_{i-U}	Under-rate burst marking probability
a_U	No. of wavelengths occupied by under-rate bursts
a_{n-O}	No. of wavelengths occupied by over-rate bursts in flow n
L_n	Burst loss probability of bursts in flow n
L_{n-U}	Burst loss probability of under-rate bursts in flow n
L_{n-O}	Burst loss probability of over-rate bursts in flow n
χ	Markov chain state
Q_χ	Probability that system is in state χ
θ	States in which all wavelengths are occupied.

end-to-end delay, the condition in (2) decreases the effect of the FBCP and BBCP updating process delay.

C. Preemption Policy

Under-rate bursts have higher priority than over-rate bursts in core networks. When contentions occur at core switches, under-rate bursts (i.e., the R in the BCP equals to 0) can preempt a channel from the originally scheduled over-rate bursts ($R = 1$). However, no preemption occurs between bursts with the same flow id.

It is known that preemption generally tends to increase the total loss probability due to inefficient use of wavelength. The BCP of the preempted burst cannot recognize that the corresponding DB is preempted and continues to reserve wavelength for the preempted DB. To minimize the total burst loss probability, the corresponding switch sends a resource-cancellation packet to release the reserved wavelengths of the preempted burst in the upstream and downstream directions. Moreover, if there are several scheduled over-rate bursts, a burst with the longest residual time (LRT) (measured from the end of the original scheduled burst to the beginning of the new arrival burst) will be preempted. It has been found that the LR wavelength selection rule is the most effective approach to reduce the burst loss probability [14].

III. ANALYSIS MODEL

Because of the difficulties related to network models, an analysis on a single congested link is a conservative OBS performance evaluation. Here, we analyze the burst loss probability of an OBS single-link system with and without the RFP implementation. Despite its simplicity, our single node analysis can possibly be adapted to a full network performance evaluation, e.g., Erlang fixed point approximation [15]. In addition, if we assume that flow n bursts travel through C core routers, we can
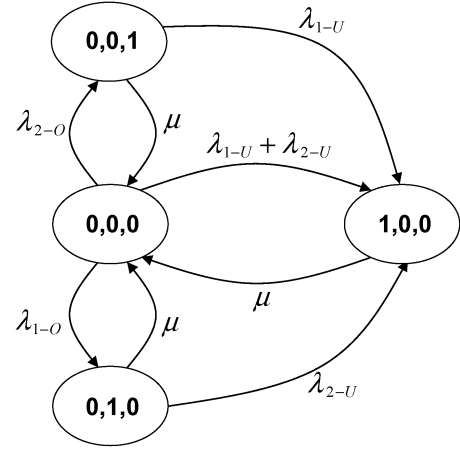

Fig. 2. Markov state transition diagram of RFP implementation for a single-link system with one wavelength and two flows.

calculate the total loss probability of flow n bursts (L_n) for a simple network topology as follows:

$$L_n = 1 - \prod_{k=1}^{C}(1 - l_k). \tag{5}$$

Here, l_k is the loss probability of flow n bursts at core router k.

Table II shows the variables used in the analysis. The system consists of N flows. The burst arrivals are Poisson, and the burst length is exponentially distributed. The average burst service time of each flow is the same as $1/\mu$.

A. Without RFP Implementation

Without the RFP implementation, the burst loss probabilities are the same for every flow in a single link system and can be calculated by using Erlang's loss formula for an $M/M/k/k$ system [8]. Each flow's burst loss probability is

$$L_1 = L_2 = \dots L_N = \frac{\rho^K/K!}{\sum_{i=0}^{K} \rho^i/i!} \tag{6}$$

where ρ is the total offered load on a link and

$$\rho = \frac{(\lambda_1 + \lambda_2 + \dots + \lambda_N)}{\mu}. $$

B. With RFP Implementation

Here, we develop an analysis model for a single link system by using a discrete-time Markov chain. The system consists of N flows competing for K data wavelengths. Each flow consists of under-rate and over-rate bursts. The definitions of under-rate and over-rate bursts are described in Section II-B. We assume that the arrivals of these bursts are Poisson. The average arrival rates of under-rate flow n and over-rate flow n can be calculated as follows:

$$\lambda_{n-U} = \lambda_n * P_{n-U}. \tag{7}$$

$$\lambda_{n-O} = \lambda_n * P_{n-O}. \tag{8}$$

The P_{n-U} and P_{n-O} calculations are shown in (3) and (4) respectively. The state of the Markov chain is expressed as $\chi = (a_U, a_{1-O}, \ldots, a_{N-O})$, where a_U is the number of wavelengths occupied by all under-rate bursts and a_{n-O} is the number of wavelengths occupied by over-rate bursts in flow n. Let Q_χ denote the probability of being in state χ, and θ denote the states in which all wavelengths are occupied, i.e.,

$$\theta \in \left\{ \chi \left| \left(a_U + \sum_{i=1}^{N} a_{i-O} \right) = K \right. \right\}. \tag{9}$$

1) One-Wavelength and Two-Flow System: We model a one-wavelength and two-flow single link system by using a discrete Markov chain as shown in Fig. 2. The state $\chi = (i,j,k)$ is the state where flow 1 and 2's under-rate bursts occupy i data channels, flow 1's over-rate bursts occupy j data channels, and flow 2's over-rate bursts occupy k data channels. According to the RFP scheme, in the case of congestion, the under-rate flow bursts can preempt a channel from over-rate flow bursts with different flow ids. Consequently, in state $(0, 0, 1)$, flow 1's under-rate arrival can preempt a channel from flow 2's over-rate burst and change the state to $(1, 0, 0)$. On the other hand, in state $(0, 1, 0)$, flow 2's under-rate arrival can preempt a channel from flow 1's over-rate burst and change the state to $(1, 0, 0)$. In the steady state, the incoming and outgoing transitions of each state are the same. The balance equations are as follows:

$$Q_{0,0,0}(\lambda_{1-U} + \lambda_{2-U} + \lambda_{1-O} + \lambda_{2-O})$$
$$= \mu(Q_{0,0,1} + Q_{0,1,0} + Q_{1,0,0}) \tag{10}$$
$$Q_{0,0,1}(\lambda_{1-U} + \mu) = \lambda_{2-O}Q_{0,0,0} \tag{11}$$
$$Q_{0,1,0}(\lambda_{2-U} + \mu) = \lambda_{1-O}Q_{0,0,0} \tag{12}$$
$$\mu Q_{1,0,0} = \lambda_{1-U}Q_{0,0,1} + (\lambda_{1-U} + \lambda_{2-U})Q_{0,0,0}$$
$$+ \lambda_{2-U}Q_{0,1,0} \tag{13}$$
$$Q_{0,0,0} + Q_{0,0,1} + Q_{0,1,0} + Q_{1,0,0} = 1. \tag{14}$$

We can use linear algebra to solve (10)–(14) to obtain the probability of being in each state (parameter Q).

According to Fig. 2, flow 1's under-rate bursts are lost in states $(1,0,0)$ and $(0,1,0)$, while flow 2's under-rate bursts are lost in states $(1,0,0)$ and $(0,0,1)$. Flow 1's over-rate bursts are lost not only when all wavelengths are occupied but also when flow 2's under-rate bursts preempt a channel in state $(0,1,0)$. On the other hand, flow 2's over-rate bursts are lost when all wavelengths are occupied and flow 1's under-rate bursts preempt a channel in state $(0,0,1)$. Consequently, the burst loss probability of flow 1 and 2's under-rate and over-rate bursts (L_{1-U}, L_{2-U}, L_{1-O}, and L_{2-O}) can be calculated as follows:

$$L_{1-U} = Q_{1,0,0} + Q_{0,1,0} \tag{15}$$
$$L_{2-U} = Q_{1,0,0} + Q_{0,0,1} \tag{16}$$
$$L_{1-O} = Q_{0,0,1} + Q_{0,1,0} + Q_{1,0,0} + \frac{\lambda_{2-U}}{\lambda_{1-O}}Q_{0,1,0} \tag{17}$$
$$L_{2-O} = Q_{0,0,1} + Q_{0,1,0} + Q_{1,0,0} + \frac{\lambda_{1-U}}{\lambda_{2-O}}Q_{0,0,1}. \tag{18}$$

Finally, we can calculate the burst loss probability of flows 1 and 2

$$L_1 = (L_{1-O} * \lambda_{1-O} + L_{1-U} * \lambda_{1-U})/(\lambda_{1-O} + \lambda_{1-U}) \tag{19}$$
$$L_2 = (L_{2-O} * \lambda_{2-O} + L_{2-U} * \lambda_{2-U})/(\lambda_{2-O} + \lambda_{2-U}). \tag{20}$$

2) K-Wavelength and Two-Flow System: Fig. 3 shows the state transition diagrams for the K-wavelength and the two-flow single-link system. The dashed lines indicate the preemption state transition. We define the unit step function, U as follows:

$$U(x) = \begin{cases} 0, & when\ x = 0 \\ 1, & when\ x > 0. \end{cases}$$

The balance equations for the state in Fig. 3 are as follows:

When $i + j + k < K$
$$Q_{i,j,k}(\mu(i+j+k) + (\lambda_1 + \lambda_2))$$
$$= Q_{i-1,j,k}(\lambda_{1-U} + \lambda_{2-U})$$
$$+ Q_{i,j-1,k}\lambda_{1-O} + Q_{i,j,k-1}\lambda_{2-O} + Q_{i+1,j,k}(i+1)\mu$$
$$+ Q_{i,j+1,k}(j+1)\mu + Q_{i,j,k+1}(k+1)\mu \tag{21}$$
and when $i + j + k = K$
$$Q_{i,j,k}(\mu(i+j+k) + U(k)\lambda_{1-U} + U(j)\lambda_{2-U})$$
$$= Q_{i-1,j,k}(\lambda_{1-U} + \lambda_{2-U}) + Q_{i,j-1,k}\lambda_{1-O}$$
$$+ Q_{i,j,k-1}\lambda_{2-O} + Q_{i-1,j+1,k}U(j+1)\lambda_{2-U}$$
$$+ Q_{i-1,j,k+1}U(k+1)\lambda_{1-U}. \tag{22}$$

$$\sum_{i,j,k} Q = 1. \tag{23}$$

For brevity, $Q_{i,j,k}$ in (21)–(23) is equal to zero if the values out of the range $0 \leq i \leq K$, $0 \leq j \leq K$, $0 \leq k \leq K$, and $(i + j + K) \leq K$. The probability of being in each state (parameter $Q_{i,j,k}$) can be calculated by using linear algebra to solve (21)–(23).

We can conclude that flow n's under-rate bursts are lost when all wavelengths are occupied and there is no originally scheduled over-rate burst with a different flow id existing in the system. Flow n's over-rate bursts are lost when all wavelengths are occupied or the under-rate burst with a different flow id successfully performs preemption. The loss probabilities of the under-rate bursts and the over-rate bursts of flows 1 and 2 are calculated as follows:

$$L_{i-U} = \sum_{Z_i} Q_{Z_i}, \text{ where } i = 1 \text{ and } 2 \tag{24}$$

$$L_{1-O} = \sum_{Z_3} Q_{Z_3} + \frac{\lambda_{2-U}}{\lambda_{1-O}} \sum_{Z_4} Q_{Z_4} \tag{25}$$

$$L_{2-O} = \sum_{Z_3} Q_{Z_3} + \frac{\lambda_{1-U}}{\lambda_{2-O}} \sum_{Z_5} Q_{Z_5} \tag{26}$$

where

$Z_1 \in \{\theta | a_{2-O} = 0\}$, $Z_2 \in \{\theta | a_{1-O} = 0\}$, $Z_3 \in \{\theta\}$
$Z_4 \in \{\theta | a_{1-O} \neq 0\}$, and $Z_5 \in \{\theta | a_{2-O} \neq 0\}$.

The burst loss probability of flows 1 and 2 can be obtained from (19) and (20).

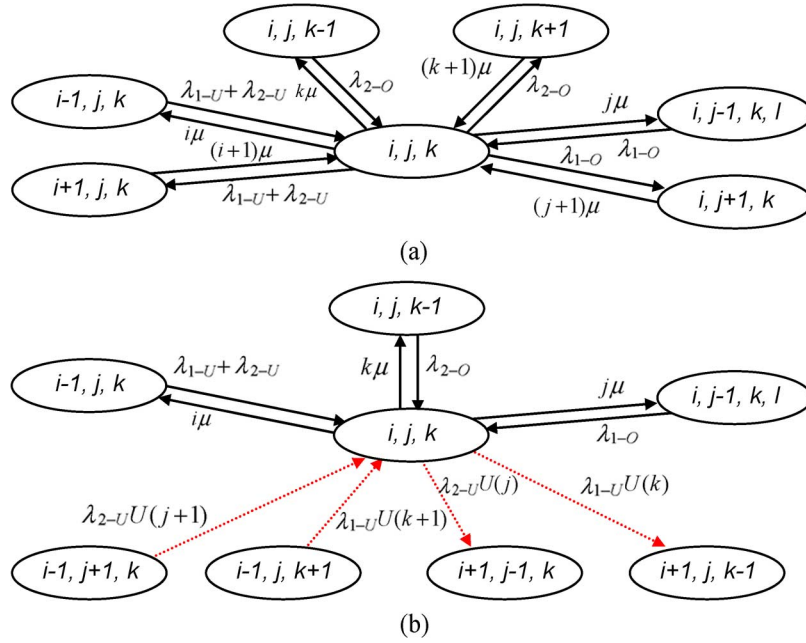Fig. 3. Markov diagram of RFP implementation for a single-link system with K wavelength and two flows. (a) When $0 < i+j+k < K$; (b) when $i+j+k = K$.

3) K-Wavelength and N-Flow System: We can model the RFP scheme in a single-link OBS system with K wavelengths and N flows by generalizing the model in Fig. 3 into an $N+1$-dimensional Markov chain. The burst loss probability of flow n, L_n, can be calculated as follows:

$$L_{n-U} = \sum_{Y1} Q_{Y1} \tag{27}$$

$$L_{n-O} = \sum_{Y2} Q_{Y2} + \frac{\sum_{i=1,i\neq n}^{N} \lambda_{i-U}}{\lambda_{n-O}} \sum_{Y3} Q_{Y3}$$
$$+ \frac{1}{\lambda_{n-O}} \sum_{i=1,i\neq n}^{N} \left(\lambda_{i-U} Q_{Y4} \left(\frac{a_{n-o}}{\sum_{j=1,j\neq i}^{N} a_{j-o}} \right) \right) \tag{28}$$

where

$Y_1 \in \{\theta | a_U + a_{n-O} = K\}$, $Y_2 \in \{\theta\}$
$Y_3 \in \{\theta | a_{n-O} = K\}$, and $Y_4 \in \{\theta | a_U \neq K \& a_{n-O} \neq K\}$.

Finally, we can calculate the burst loss probability of flow n as follows:

$$L_n = (L_{n-O} * \lambda_{n-O} + L_{n-U} * \lambda_{n-U})/(\lambda_{n-O} + \lambda_{n-U}). \tag{29}$$

IV. SIMULATION RESULTS

We conducted simulations on a modified optical burst switching simulator originally developed at the Optical Internet Research Center (OIRC) [16] on the basis of ns-2 [17]. The RFP scheme based on the fixed and adaptive max-min rate allocations are named RFP-F and RFP-A, respectively. We assume that the number of control wavelengths is large enough to ensure no losses for all control packets (BCP, FBCP, and BBCP) and that the network has full wavelength conversion capability. The capacity of each wavelength is 1 Gbps. The BCP processing time is 0.1 msec. The max-min weight of all flows is equal to one. Edge switches generate bursts with exponentially distributed inter-arrival times and burst lengths with an average burst length of 100 KB. We set the input traffic rate of flow i to the designated value by adjusting the average burst inter-arrival time. Edge switches use an exponential moving average to estimate the arrival rate of each flow. The arrival rate is calculated as follows:

$$A_{\text{new}} = (1 - e^{-T/K_\alpha})\frac{l}{T} + e^{-T/K_\alpha} A_{\text{old}} \tag{30}$$

where A_{new} is the estimated arrival rate, T is the inter-arrival time between the current and the previous burst, l is the burst size, A_{old} is the previous value of the arrival rate before updating, and $K_\alpha = 0.1$. Note that the flow's sending rate fluctuates but for simplicity the term "sending rate" in this paper refers to the average sending rate.

A. Analysis Versus Simulation

In this subsection, we validate our analysis model in a single congested link system with the two flows shown in Fig. 4. To validate the analysis, bursts are randomly preempted in the case of several preempted candidates. The transmission delay of each link is equal to 1.0 msec. The number of wavelengths for data bursts at link C1-E3 is 8 while that of other links is large enough to guarantee no losses. The normalized rate of 1 is equivalent to 8 Gbps. We set $e = 0.7$ and $S = 0.01$ Gbps. The normalized PFA max-min rate of each flow is equal to 0.35. The total simulation time is 120 sec. During the time between 0 and 60 seconds, flow 1's normalized sending rate is equal to 0.3, and after time 60 seconds, flow 1 increases its sending rate to 0.5.
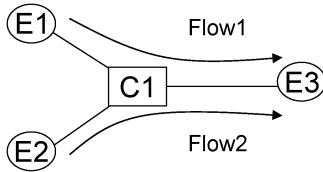
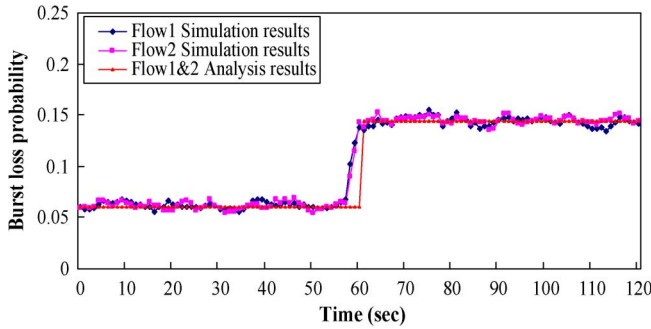Fig. 4.   Single congested link system.

Fig. 5.   Burst loss probability in single congested link system without RFP implementation.
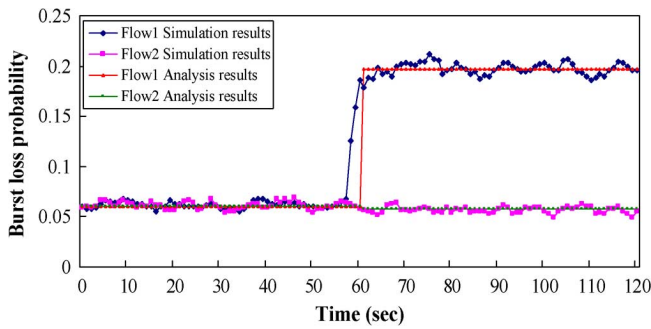

Fig. 6.   Burst loss probability in single congested link system with RFP-F implementation.
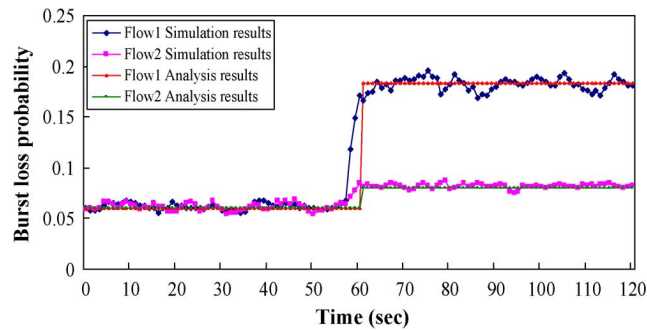

Fig. 7.   Burst loss probability in single congested link system with RFP-A implementation.

Flow 2's normalized input rate is equal to 0.3 during the whole simulation time.

The results plotted in Figs. 5–7 are based on the sliding window with a size of three seconds. The results in Fig. 5 show that without RFP when flow 1 sends a large amount of traffic (at $time > 60\,sec$), the loss probability of flow 2 also increases to the same level as that of flow 1.

In contrast, RFP efficiently provides service isolation and protection. Figs. 6 and 7 show the burst loss probability for RFP-F and RFP-A, respectively. The analytical results closely follow the simulation results. When flow 1 sends traffic under its PFA max-min rate (during 0–60 sec), the results with and without RFP implementation are almost the same, because both flows send the traffic under their PFA max-min rates. During 60–120 sec when flow 1's sending rate exceeds its max-min rate, RFP can prevent high losses in flow 2 (well-behaved flow).

Note that the loss probabilities of flow 1 and 2 during 60–120 sec in the cases of RFP-F and RFP-A are not the same because of the different allocated max-min rate. The normalized allocated max-min rate (T_i) of both flows in the case of RFP-F is equal to 0.35 whereas RFP-A allocates the un-used bandwidth of flow 2 for flow 1. Consequently, the normalized allocated max-min rate (T_i) of flow 1 and 2 in the case of RFP-A is equal to 0.4 and 0.3, respectively. We can conclude that the performance difference between RFP-F and RFP-A is that the misbehaved flows in the case of RFP-F have higher loss probabilities than those of RFP-A when the well behaved flows send traffic under their PFA max-min rate. This is because the PFA max-min rate used in RFP-F assumes that all flows have enough traffic to send and seek to maximize throughput. However, the performance difference between RFP-F and RFP-A is not much. Consequently, besides its simplicity, RFP-F is efficient to be implemented in practical networks. To validate our analysis model, a random wavelength selection rule is used for preemption; therefore, the total loss probabilities in conventional OBS, RFP-F, and RFP-A OBS networks are almost the same.

B. RFP Performance Evaluation

The objective of this subsection is to evaluate the performance of RFP in terms of service isolation and protection in a large-scale network, i.e., the NSF network. In addition, the performance of RFP in terms of total loss probability is also evaluated. The simulated NSF network consists of 14 core switches, and each core switch is attached to an edge switch (Fig. 8). The distance between adjacent edge and core switches equals to 200 km. The distance between adjacent core switches in kilometers is shown in Fig. 8. The transmission delay is 0.5 ms/100 km. We assume that there are no DB losses in links between edge and core switches while the number of wavelengths for data bursts between two core switches in each direction is equal to 16. The normalized rate of 1 is equivalent to 16 Gbps (capacity C of links between adjacent core switches). We set $e = 0.7$ and $S = 0.01$ Gbps. The route is selected based on the shortest path scheme.

Besides the ability of RFP to prevent losses in well-behaved flows (service isolation and protection), we also evaluate the total loss probability. Therefore, the number of misbehaved flows in the system must be large enough to evaluate the total burst loss probability. The number of hops between source and destination is from three to five in the NSF topology in Fig. 8. For simplicity, we classified flows into three groups according to hop counts: 3-hop, 4-hop, and 5-hop flows. The NSF network consists of 182 flows and the ratio of 3-hop, 4-hop, and 5-hop flows is 1:1.714:1.619. The normalized sending rate of each
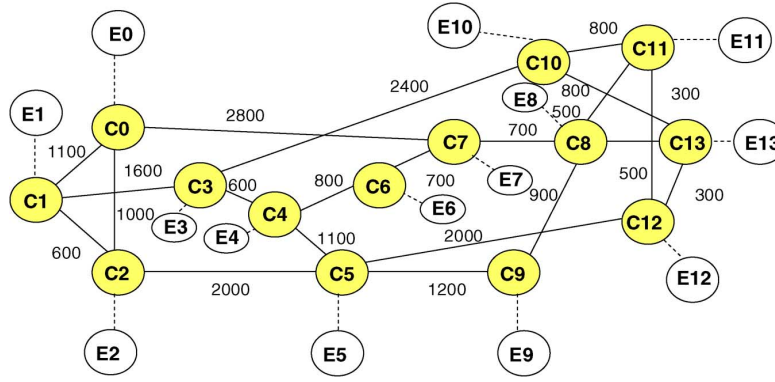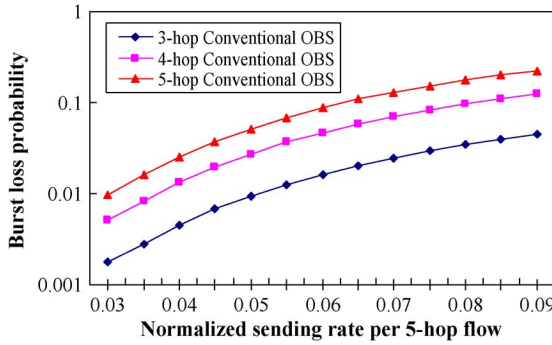
Fig. 8.   NSF network.

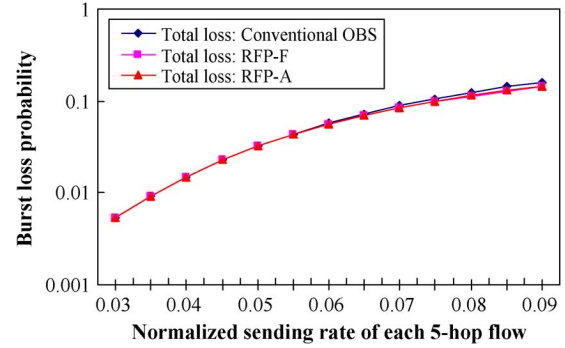Fig. 9.   Burst loss probability in conventional NSF network.

Fig. 10.   Burst loss probability in RFP-F-based and RFP-A-based NSF networks.
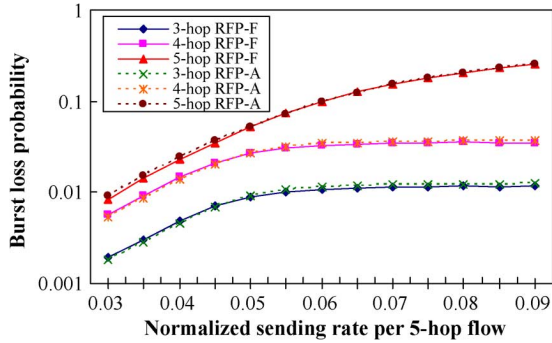

Fig. 11.   Total burst loss probability in NSF network.

3-hop and 4-hop flow is fixed at 0.05, while the sending rate of each 5-hop flow varies from 0.03 to 0.09.

In the case of conventional OBS networks (Fig. 9), we can see that when 5-hop flows increase their sending rate, the average burst loss probabilities of 3-hop and 4-hop flows rise rapidly and the loss probability is proportional to the hop counts between source and destination.

In contrast, RFP efficiently provides service isolation and protection (FBA), as shown in Fig. 10. Based on the shortest path selection, the link from C7 to C8 in the simulated NSF network (Fig. 8) has the maximum number of flows, 15 flows, competing for data wavelengths. This implies that the simulated NSF network has the bottleneck links when the normalized sending rate per flow is higher than the minimum normalized PFA max-min rate assigned to the flow, which is equal to 0.047 ($e \times \{\text{normalized rate of } C\}/\{\text{number of flows in the most congested link}\} = 0.7 \times 1/15$). In this simulation, the

average normalized sending rates of 3-hop and 4-hop flows are fixed at 0.05. Therefore, most of 3-hop and 4-hop flows are well-behaved flows whereas only a few of them are misbehaved flows. We varied the sending rate of 5-hop flows from 0.03 to 0.09. When the normalized sending rate of 5-hop flows increases from 0.03 to 0.05, we can see that the loss probabilities of 3-hop and 4-hop flows also increase as shown in Fig. 10. This is because most of 3-hop, 4-hop and 5-hop flows send traffic under the max-min rate, so they are treated equally. In case the input rate per 5-hop flow is over 0.05, some 5-hop flows send traffic over their max-min rate. We can see that RFP efficiently protects well-behaved flows (3-hop and 4-hop flows) from quality degradation. The loss probabilities of 3-hop and 4-hop flows do not increase to a high level as in the case of the conventional OBS networks. The performances of RFP-F and RFP-A are very similar. The results in Fig. 11 show that RFP does not degrade the total burst loss probability. In fact, RFP gives slightly lower burst loss probability in the high traffic load. This is because RFP drops overload traffic at the beginning of their journey before they can contend with other bursts along the path. In addition, RFP uses some strategies to prevent high losses, e.g., LRT wavelength selection rule in the case of many preempted candidates.

C. Global Situation Awareness

We investigate the performance of RFP and another existing scheme called MMFP [11] in terms of the global situation awareness in this subsection. Global situation awareness takes into account the state of the network, e.g. whether there are bottleneck links that cause inefficiency by increasing total loss. Bottleneck links occur when the offered traffic exceeds the
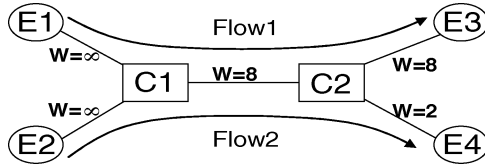
Fig. 12. Network scenario exhibiting some inefficiency in case of lacking global situation awareness.
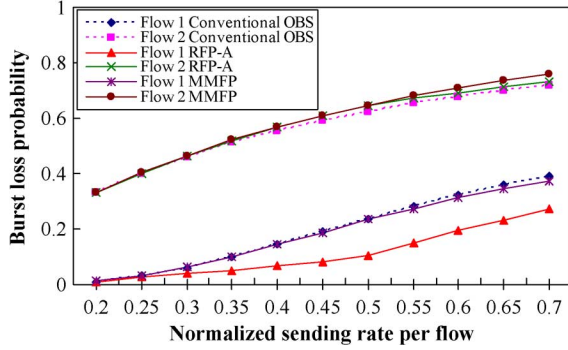

Fig. 13. Burst loss probability.

Fig. 14. Total burst loss probability.

capacity of some link. The model proposed in [2] shows that ignoring global situation awareness causes the system throughput to become zero (called congestion collapse) when the offered load goes to infinity. To illustrate the inefficiency caused by bottleneck links, consider the simple network in Fig. 12. Links C1-C2 and C2-E3 have eight wavelengths (8 Gbps) while link C2-E4 has only two wavelengths (2 Gbps). For simplicity, we assume that there is no loss between links E1-C1 and E2-C1. The max-min rate at link C1-C2 for both flows is equal to half of the effective link capacity ($e * C$). Many flow 2 bursts are dropped on link C2-E4 because of its low capacity. The performance of flow 1 depends on link C1-C2 while that of flow 2 mainly depends on link C2-E4. The FBA scheme should be aware of the global situation to avoid inefficiencies. In the rate control method, the input rate of flow 2 should not exceed 2 Gbps. By limiting flow 2's input rate, although the throughput of flow 2 decreases, the total throughput of the network increases as shown in [8]. For the preemption-based FBA schemes, e.g. MMFP and RFP, the input rate of each flow is not limited. Therefore, the preemptive priority of flow 1 should be higher than that of flow 2 in order to enhance the efficiency of the network.

Fig. 13 plots the loss probabilities of flows 1 and 2 in conventional, RFP-A, and MMFP-based OBS networks. The details of MMFP scheme are briefly described in Section V-E. We set $e = 1.0$ and $S = 0.01$ Gbps. The normalized rate of 1 is equivalent to 8 Gbps. We can see that the performance of conventional OBS and MMFP-based OBS networks are almost the same. This is because MMFP is not aware of the global situation. Each MMFP-based core switch independently allocates and performs preemption without sharing the network state among switches. In contrast, RFP gives higher preemptive priority to flow 1 on link C1-C2. Although RFP increases the loss probability of flow 2, when we consider the total loss probability of the system shown in Fig. 14, it is lower in RFP-based OBS networks than
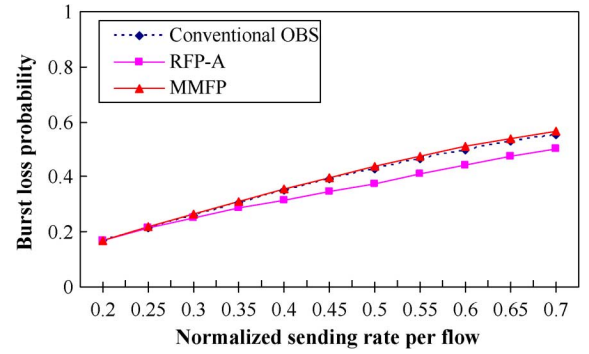
that in conventional and MMFP-based OBS networks. In addition, the total loss probability in the case of MMFP is slightly higher than that of conventional OBS. This is because the arrival rates of flow 1 and flow 2 at any particular time are not the same; therefore, some of flow 1 bursts are preempted by flow 2 bursts which are lost in the link C2-E4.

V. DISCUSSION

A. Max-Min Fairness in OBS Networks

Unlike the maximum throughput resource management (MTRM) that tries to maximize the total throughput of the network, max-min fairness allocates more bandwidth to poorly treated flows. MTRM may lead to unfairness because it tends to allocate more bandwidth for less expensive flows. Compared with MTRM, achieving max-min fairness provides rate fairness but it may lead to throughput decreasing because the network ensures max-min rate even for expensive flows such as mobile stations that are far from the base station in wireless networks or flows with a large number of hops in OBS networks. However, providing max-min fairness in OBS networks is considered feasible because the throughput will not be decreased significantly as in wireless networks. In the paper, we presented only the loss probability performance because throughput in OBS networks can be only decreased by loss, and hence the loss probability performance is the major concern in OBS networks. The results in Sections IV-B and IV-C show that RFP provides service isolation and protection without increasing the loss probability (or decreasing the throughput). In fact, RFP decreases the loss probability slightly as in Section IV-C because RFP allocates max-min rate by considering how much resource that each flow can utilize (global situation awareness).

B. Rate Control versus Preemption

Two kinds of methods can provide FBA in OBS networks. The first is rate control, which, as proposed in [8] and [9], limits the input traffic of each flow according to its optimum fairly allocated bandwidth. However, to achieve an acceptable loss level for well-behaved flows in OBS networks, the optimum fairly allocated bandwidth in [8] and [9] is calculated by assuming that the network has capacity $e * C$. If the value of e is small, rate control schemes may degrade network utilization and increase end-to-end delay. If e is too high, the burst loss probability of well-behaved flows may be too high to be acceptable for some

applications. The second method proposed in [11] and in this paper uses preemption, by which edge switches do not decrease the sending rate of each flow but the core switches drop misbehaved flow bursts only in the case of contention. This can isolate flows and provide acceptable loss level for well-behaved flows (by setting small value of e in RFP) while ensuring efficient network utilization. In addition, if the preempted bursts are selected on the basis of the burst size and reservation time, the total loss probability can be reduced [14].

C. Fairness at IP Flow Level

Since OBS switches are low-layer equipment, it is not suitable for OBS switches to be involved in issues at IP flow level. Therefore, all of the studies in [8]–[12], including this one, on providing FBA in OBS networks classify flows by the ingress and egress switches' addresses. To provide FBA at IP flow level, RFP-based edge switches can assemble misbehaved and well-behaved IP flows into different bursts and mark the misbehaved IP-flow bursts as an over-rate flow. Moreover, burst segmentation can be applied so that the misbehaved IP flows are dropped first.

D. Scalability

The scalability issue was described in [8] and [11]. For RFP, in the case of a network with N edge switches and L links, the maximum number of flows for which each edge switch needs to maintain the state information is equal to $N - 1$. The amount of state information maintained by each core switch depends on the network topology. If we assume that the average number of hops between source and destination is equal to H and all flows are uniformly distributed in all links, the estimated average number of flows in each direction per link is equal to $N(N-1)H/2L$. Consequently, scalability in an RFP-based OBS network is not a significant problem.

E. MMFP Versus RFP

To the best of our knowledge, besides our RFP scheme, only the max-min fairness preemption (MMFP) scheme [11] has a mechanism based on preemption for achieving max-min fairness in OBS networks. The details of MMFP are described as follows.

Step 1) Edge switches monitor and update the initial arrival rate of each flow (I_i) at the core network. In addition, the core switches monitor both the actual arrival (A_i) and loss rate (l_i) of each flow.

Step 2) Each core switch independently determines the fairly allocated bandwidth (F_i) based on the max-min fairness criteria in [13] for a single outgoing link for each flow.

Step 3) The core switch allocates the effective load (E_i) for each flow. E_i is calculated by using both I_i and A_i. The E_i calculation is described in [11].

Step 4) The core switch determines the theoretical loss level P_i for each flow as follows:

If $A_i \leq F_i$

$$P_i = \mathrm{Er}(E_i) \tag{31}$$

where $\mathrm{Er}(\cdot)$ is the Erlang loss formula in (6)

Else

$$P_i = \frac{A_i - F_i * (1 - Er(E_i))}{A_i} \tag{32}$$

End if.

For flow j with the largest sending rate,

$$P_j = \frac{\sum_{i=1}^{i=N} A_i l_i - \sum_{i=1, i \neq j}^{i=N} A_i P_i}{A_j}. \tag{33}$$

Step 5) The burst in the contention list which belongs to the flow whose actual measured loss rate has the largest difference from its loss level is preempted.

Note that RFP does not impose high work load on core networks as MMFP does, because an RFP-based core switch does not need to monitor the arrival rate and loss rate of each flow. In addition, since each MMFP-core switch independently allocates the max-min rate and performs preemption without considering the optimum rate along the path, this may cause inefficient network performance, as mentioned in Section IV-C. In contrast, the RFP fixed and adaptive allocation methods can allocate the optimum max-min rate along the path. Moreover, the authors of MMFP describe in [11] that the actual OBS network cannot guarantee the loss level according to (31) and (32). Therefore, MMFP needs to increase the loss level for the flow with the largest sending rate as shown in (33). This is unfair for the flow with the largest sending rate.

VI. CONCLUSION

We proposed the rate fairness preemption (RFP) scheme for achieving weighted max-min fairness in OBS networks. RFP is an alternative to rate control to provide fair bandwidth allocation (FBA) in OBS networks because of its efficient network utilization and low workload for core networks. In addition, RFP does not degrade the total burst loss probability and it can improve the loss performance in the case the misbehaved flows send high amount of traffic to networks. We proposed fixed (RFP-F) and adaptive (RFP-A) max-min rate allocation methods. RFP-F is the simplest scheme to achieve FBA in OBS networks. The analysis and simulation results show that both RFP-F and RFP-A schemes efficiently provide service isolation and protection in OBS networks.

Although many approaches have been proposed for achieving max-min fairness in OBS networks, none of them use the feedback information from the network for path selection to improve the overall burst loss probability. Therefore, our future work will develop an adaptive path selection scheme which reduces the overall burst loss probability for max-min fairness-based OBS networks. In addition, we will develop a preemption scheme for providing both rate fairness (FBA) and distance fairness (fairness among bursts with different numbers of hops between source and destination) in OBS networks.

REFERENCES

[1] I. Stoica, S. Shenker, and H. Zhang, "Core-stateless fair queueing," *IEEE/ACM Trans. Network.*, vol. 11, no. 1, pp. 33–46, Feb. 2003.

[2] J. L. Boudec, Rate Adaptation, Congestion Control and Fairness: A Tutorial [Online]. Available: http://ica1www.epfl.ch/PS_files/LEB3132.pdf

[3] Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *J. Internetworking Res. Exp.*, pp. 3–26, Oct. 1990.

[4] D. Lin and R. Morris, "Dynamics of random early detection," presented at the Proc. ACM SIGCOMM'97, Cannes, France, Set. 1997.

[5] C. Qiao, M. Yoo, and S. Dixit, "OBS for service differentiation in the next-generation optical network," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 98–104, Feb. 2001.

[6] Y. Chen, M. Hamdi, D. H. K. Tsang, and C. Qiao, "Proportional QoS over OBS networks," in *IEEE Globecom*, Nov. 2001, vol. 20, no. 1, pp. 1510–1514.

[7] J. Phuritatkul and Y. Ji, "Resource allocation algorithms for controllable service differentiation in optical burst switching networks," *IEICE Trans. Commun.*, vol. E88-B, no. 4, pp. 1424–1431, 2005.

[8] S. Kim, Y.-C. Kim, B.-Y. Yoon, and M. Kang, "An integrated congestion control mechanism for optimized performance using two-step rate controller in optical burst switching networks," *J. Comput. Telecommun. Network.*, vol. 51, no. 3, pp. 606–620, Feb. 2007.

[9] H. Boyraz and N. Akar, "Rate-controlled optical burst switching for both congestion avoidance and service differentiation," *J. Op. Switch. Network. (Elsevier)*, vol. 2, no. 4, pp. 217–229, 2005.

[10] L. Kim, S. Lee, and J. Song, "Congestion control scheme based on peak load in optical burst switching networks," in *Proc. GLOBECOM*, 2006, pp. 1–5.

[11] Y. Liu, K. C. Chua, and G. Mohan, "Max-min fairness in WDM optical burst switching networks," *J. High Speed Networks*, vol. 16, no. 4, pp. 379–398, 2007.

[12] T. Orawiwattanakul, Y. Ji, Y. Zhang, and J. Li, "Preemption scheme for providing rate fairness in optical burst switching networks," in *Proc. Int. Conf. on High Performance Switching and Routing (HPSR2008)*, 2008, pp. 39–44.

[13] D. P. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[14] T. Tachibana and S. Kasahara, "Two-way release message transmission and its wavelength selection rules for preemption in OBS networks," *IEICE Trans. Commun.*, vol. E90-B, no. 5, pp. 1079–1089, May 2007.

[15] Z. Rosberg, H. L. Vu, M. Zukerman, and J. White, "Performance analyses of optical burst switching networks," *IEEE J. Sel. Areas Commun.*, vol. 21, pp. 1187–1197, Sep. 2003.

[16] The Optical Internet Research Center Optical Burst Switching NS Simulator [Online]. Available: http://wine.icu.ac.kr/~obsns/index.php

[17] ns-2 Network Simulator 2000 [Online]. Available: http://www.isi.edu/nsnam/ns

Tananun Orawiwattanakul received the B.Eng. from Sirindhorn International Institute of Technology, Pathum Thani, Thailand, and the M.Eng. degree from the Asian Institute of Technology, Pathum Thani, Thailand, in 1997 and 2000, respectively.

She is currently working toward her Ph.D. degree at the Graduate University for Advanced Studies (in cooperation with the National Institute of Informatics), Japan. Her research interests include optical burst switching networks, fairness, traffic control, and quality-of-service provisioning.

Yusheng Ji (M'94) received the B.S., M.S., and D.Eng. degrees from the University of Tokyo, Tokyo, Japan, in 1984, 1986, and 1989, respectively, all in electrical engineering.

She joined the National Center for Science Information Systems in 1990. She is currently an Associate Professor at the National Institute of Informatics and the Graduate University for Advanced Studies, Tokyo. Her research interests include network architecture, resource management, and performance analysis for QoS provisioning in wired and wireless networks.

Dr. Ji is a member of the Institute of Electronics, Information, and Communication Engineers (IEICE) and the Information Processing Society of Japan (IPSJ).

Yongbing Zhang (M'96) received B.S. degree in electrical engineering from the Polytechnical University, Kanagawa, Japan, in 1984, and the M.S. and Dr.Eng. degrees in computer science both from the University of Electro-Communications, Tokyo, Japan, in 1989 and 1992, respectively.

From 1992 to 1996, he was a Research Associate with the Department of Computer Science at the University of Electro-Communications. In 1996, he joined the Institute of Policy and Planning Sciences, University of Tsukuba, Ibaraki, Japan. He is now an Associate Professor at the Graduate School of Systems and Information Engineering of the University of Tsukuba. His research interests include distributed/parallel computer systems, communication networks, and performance evaluation.

Jie Li (SM'04) received the B.E. degree in computer science from Zhejiang University, Hangzhou, China, in 1982, the M.E. degree in electronic engineering and communication systems from China Academy of Posts and Telecommunications, Beijing, in 1985, and the Dr. Eng. degree from the University of Electro-Communications, Tokyo, Japan, in 1993.

Since April 1993, he has been with the University of Tsukuba, Ibaraki, Japan, where he has been a Professor. His research interests are in mobile distributed multimedia computing and networking, OS, network security, modeling and performance evaluation of information systems.

Dr. Li is a senior member of ACM, and a member of the Information, and Information Processing Society of Japan (IPSJ). He received the best paper award from IEEE NAECON'97. He has served as a Secretary for the Study Group on System Evaluation of IPSJ and on several editorial boards for the *IPSJ Journal* and so on, and on Steering Committees of the SIG of system evaluation (EVA) of IPSJ, the SIG of DataBase System (DBS) of IPSJ, and the SIG of MoBiLe computing and ubiquitous communications of IPSJ. He has also served on the program committees for several international conferences such as IEEE ICDCS, IEEE INFOCOM, IEEE GLOBECOM, and IEEE MASS.