Adaptive Burst Admission and Forwarding in OBS Networks

Sébastien Rumley[1], Oscar Pedrola[2], Miroslaw Klinkowski[2,3], Pedro Pedroso[2], Christian Gaumier[1]
Davide Careglio[2], Josep Solé-Pareta[2]
[1] TCOM - Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
[2] CCABA, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain
[3] National Institute of Telecommunications (NIT), 1 Szachowa Street, 04-894 Warsaw, Poland
Tel: (+34) 93 4016985, Fax: (+34) 93 401 7055, e-mail: mklinkow@ac.upc.edu

ABSTRACT

This paper addresses the problem of controlling the performance of optical burst switching (OBS) networks in presence of non-stationary traffic demands. We propose a joint burst admission control and forwarding mechanism that operates in core nodes. This mechanism dynamically adapts its behaviour according to the feedback messages received from other nodes. By not forwarding certain bursts not complying with given requirements, an admission control is implicitly made. Moreover, by forwarding bursts to appropriately selected nodes, traffic balancing is achieved. The advantageous effects of the proposed mechanism can additionally be amplified by granting extra offset time to the burst. The benefits provided by this mechanism are supported by numerical results.

Keywords: optical burst switching, adaptive routing, deflection routing, performance analysis.

1. INTRODUCTION

Modern communication networks are intended to deal with highly varying traffic conditions. The demands they have to carry out often imply traffic intensities varying both in the short and long term. Since the data traffic is generally statistically multiplexed, the networks that are subject to the varying traffic conditions are always susceptible to congestion states, in particular, to transient congestion due to short term traffic fluctuations, and to semi-permanent congestion due to long term traffic variation. Several approaches have been considered in order to diminish the network congestion under varying traffic demands. One consists in delaying temporarily the data traffic by means of *buffering*, which allows postponing the transmission until the outgoing link is available. This approach is effective for solving transient congestions but not for semi-permanent ones. It is also possible, when applying proper *routing***,** to take advantage of the whole network capacity rather than only the one available on a specific path. Thus, by distributing intelligently the traffic over the network, permanent congestion can be avoided. Transient congestions can also be alleviated if alternative (or deflection) paths toward destination are explored. Eventually, in critical cases, a *Connection Admission Control* (CAC) mechanism can limit the amount of traffic entering or travelling through the network. In this way, the traffic supposed to cause congestion anyway is dropped, prior to consume network resources.

On the other hand, high-capacity communication networks make an extensive use of photonic technologies. Primarily reserved to point-to-point transmission, fibre optics progressively acquired new functionalities such as switching and routing. With the advent of Wavelength Dimension Multiplexing (WDM) and optical cross-connects (OXC), Optical Circuit Switched (OCS) networks have been deployed, with nodes able to route transparently an optical signal (lightpath) to the next hop. Moreover, two upcoming technologies, Optical Packet Switching (OPS) and Optical Burst Switching (OBS) have been proposed to achieve statistical multiplexing directly at the optical level. In OPS [1], optical packets carrying their own headers are sent in the network. At each intermediate node, while these headers are processed, the payload is optically buffered. On the other hand, in OBS networks [2], a burst control packet (BCP) is sent on a dedicated channel to inform the intermediate nodes of the arrival of its associated burst. This BCP precedes the payload of a short delay (offset time) during which the latter is buffered at the network edge. While many technological obstacles are still precluding the implementation of OPS (especially the difficulty to achieve optical buffering), OBS networks can be realised with the components nowadays available. Additionally, since incoming data are buffered anyway at network edges, OBS permits to aggregate several smaller packets together within a burst. This aggregation decreases the amount of packets travelling in the network, and thus simplifies the node architecture. OBS is thus considered as a promising switching paradigm for optical networks [3].

In this paper, we propose a novel OBS operational scheme, implemented in core nodes, permitting to mitigate the congestion due to short and long term varying traffic demands in an adaptive manner. Routing and admission decisions are hence taken according to feedbacks received from other nodes. The major advantage of our approach consists in the fact that OBS core nodes do not require the full knowledge of the topology anymore at initialisation.

In the next sections, we first review how the three aforementioned approaches (buffering, routing, CAC) have been envisaged until now to improve the performance of OBS networks (Section 2). We then present our novel OBS operational scheme in Section 3 while in Section 4, we show several numerical results demonstrating the validity of our scheme. Concluding remarks and future research directions are given in Section 5.

978-1-4244-4826-5/09/$25.00 ©2009 IEEE

2. CONGESTION CONTROL IN OBS NETWORKS

Congestion in OBS networks leads to the contention of bursts which, if not resolved, results in the loss of data. Indeed, since OBS switching nodes, in the original scheme, lack of any buffering capability, a burst is automatically dropped if no resources are available to forward it. Several solutions have been proposed to diminish the contention in OBS networks.

A first solution consists in effectively balancing the flows over the network, thus making use of routing congestion countermeasures. Traffic Engineering (TE) methods have been studied and both linear [4] and non-linear [5] optimization methods have been developed for that purpose. In [4] a problem of finding a set of routes for each source-destination demand is considered with the objective to limit the maximal offered rate on the most loaded link. In [5] the objective function, which represents a network-wide burst loss probability, is non-linear and thus a proper gradient method has been applied so that to find the distribution of traffic over a given set of pre-established paths.

The TE approaches presented above allow solving permanent contention by a pro-active exploration of the knowledge of average traffic rates. Yet other approach can also help to solve transient congestion states by using deflection routing [6]. In this method, rather than dropping a burst which failed to get a reservation on a particular link, this burst can be sent on an alternative output link. For instance, in the blind or hot potato deflection scheme [7], an arbitrary link is picked to forward the contenting burst, while in the next shortest path scheme [8], the path leading to the second shortest path toward destination is selected. More variants of deflection routing exist in the literature. Main difficulty with deflection routing is caused by the problem of insufficient offset-time [6]. Namely, a burst, when deflected, might have insufficient offset to reach its destination due to the difference in the length of the primary and alternative path. This problem can be solved either by restricting the deflection to the paths with the same length as the primary one or by providing additional offset to the bursts at edge nodes.

To further avoid congestion and in particular in cases of network overload, CAC schemes can be applied. Such schemes aim at limiting the amount of traffic sent to prevent the network from blocking. CAC schemes can be based on the analysis of the incoming flow [9], or can take their decision based on feedback received from other nodes [10].

Eventually, buffering capabilities can be granted to an OBS network to mitigate the congestion. In OBS, buffering is achieved either by storing the traffic at edge nodes in electronic memories (flow smoothing) or at core nodes by the mean of Fibre Delay Lines (FDL) working as optical buffers. By using wavelength converters at the FDL entrance the buffer capacity is automatically multiplied by the number of available wavelength [11]. Thus, only a small number of FDLs might be required. The application of FDLs gives good results in the presence of transient contention but will have no effect when permanent congestion is experienced. However, FDL buffering adds a lot of complexity to OBS, whose main advantage relies precisely in its simplified architecture and operation, when comparing to OPS.

Also remark that both deflection routing and buffering lead to an unordered reception of the burst, which might have a non negligible effect on the TCP performance. Indeed, TCP will conclude to a packet drop when a packet of higher index arrives before packets with preceding indexes. The impact of OBS on TCP performance is studied in [12].

3. ADAPTIVE CONNECTION CONTROL SCHEME

Our study is based on the standard OBS model defined in [13]. The traffic is injected at edges node, connected to core nodes. One assumes no data losses between an edge and a core node. Our scheme requires no modification of the conventional OBS node architecture on the data path. On the control path, Burst Control Packet (BCP) must be able to carry, along with the habitual headers, the indexes sequence of the nodes visited so far and a unique identifier (ID). This ID should allow a disambiguation between two burst. IDs can be obtained during the aggregation process, by hashing the payload, or simply by combining their source and destination node indexes with a time stamp. The proposed scheme can be used with conventional scheduling algorithms as the FirstFit Void Filling (FFVF) and the Last Available Unscheduled Channel (LAUC) [13] and with both conventional (C-OBS) and offset time-emulated (E-OBS) OBS approaches [14]. We assume no buffering capabilities are available at core nodes.

The bursts are sent in the network with an offset time equal to the core node processing time multiplied by a hop factor. For a given source-destination (s-d) pair of nodes, this hop factor is equal to the length (in terms of hops) of the shortest path between s-d plus an extra additional offset. This offset supplement permits to give more flexibility to the intermediate nodes that will forward the bursts. In this way, burst will be allowed to take a route longer than the shortest one to reach their final destination. A larger offset also permits to deflected several times a burst before dropping it.

The scheme can be decomposed into three steps, namely *admission/emission*, *forwarding* and *feedback*. It also uses an element called Time Sliding Feedback Counter. Although in the normal node operational sequence, the admission phase takes first place, the *forwarding* and *feedback* phases are first described, in order to ease the understanding. Then, the description of the time sliding feedback counters will be given, preceding the depiction of the *admission/emission* phase.

3.1 Burst forwarding

As it will be described in subsection 3.4, a burst is declared admitted if it has to be forwarded and if a reservation has been successfully negotiated for it. Thus, after each burst admission, a core node *forwards* the updated control packet toward next node. It also stores in its local memory, associated with the burst identifier, a triplet of information (D, R, N), where D is the burst destination, R the burst remaining offset and N the index of the node toward which burst will be forwarded.

3.2 Feedback mechanism

Whenever a burst is blocked or reaches its final destination, the dropping or receiving node will look up the visited node sequence carried by the BCP, and send a feedback to each of them. Feedback messages only contain a burst identifier and a flag indicating whether the burst was delivered or not. They are supposed to travel either using the BCPs flowing in the reverse direction or via dedicated messages, sent out of the band.

An core node receiving a feedback message first retrieves from its local memory the (D, R, N) triplet associated to the burst identifier contained in the message. The contained flag will allow this core node to establish whether its previous decision of forwarding to node N a burst destined to D with R offset units remaining has been successful of not.

Core nodes store the feedbacks they receive using Time Sliding Feedback Counters (TSFCs). Each core node hence owns a TSFC for each possible (D, R, N) triplet.

3.3 Time Sliding Feedback Counters

The purpose of a TSFC is to count how many positive and negative feedback messages have been collected for a specific decision in the recent past. To achieve this goal, a TSFC includes two registers, one for positive feedbacks and one for negative ones, each composed of n elements. A TSFC also contains an index i referring to ith element of each the registers. When a feedback is reported to a TSFC, the element of the register corresponding to the actual i value and to the type (positive/negative) of the feedback is incremented. At regular time intervals T_{mem}/f, the i value is incremented and the values stored at both element ($i + 1$) are reset. If the last pair of register elements is reached, the index is reset to 0, similarly as in circular buffers. T_{mem} and f are constant parameters permitting to vary respectively the time during which feedbacks are listed and the frequency at which the counters are updated.

The register index position is modified at regular time intervals and not according to the feedback arrivals, therefore the duration of time the counter represents oscillates between ($1 - 1/f$) T_{mem} (immediately after the position pointer update) and T_{mem} (just before position update). However, since f is expected to take values typically greater than 20, this oscillation effect is neglected and a counter is assumed to represent in practice the last T_{mem} seconds.

By knowing how many positive and negative feedback indications have been collected, it is possible to compute a conditional success probability $P(\cdot|$*feedbacks received in the last* T_{mem} *seconds*). Since TSFCs are associated to (D, R, N) triplets, the probability P(*delivery of a burst destined to D with offset R forwarded to N* | *feedbacks*), noted hereafter P_{DRN}, can be estimated.

3.4 Admission at core nodes

Whenever a BCP arrives at an intermediate core node, this latter must decide if the corresponding burst should be *admitted*. Admitting a burst means granting a reservation on one of the outgoing links. The admission phase thus implies both scheduling and routing operations. Note that in our scheme an incoming burst cannot be admitted to a link directed towards the previous node.

When a core node *CN* has to decide for admission or dropping, it first refers to the knowledge of the past provided by its TSFCs. Assuming that the burst candidate for admission B has a remaining offset r and is destined to d, that node *CN* has N neighbouring nodes numbered from 1 to n, and that B has arrived on the incoming link connected to node k, *CN* will retrieve the delivery probabilities P_{drm}, with $m = 1.. n, m \neq k$.

CN will then exclude the unfavourable forwarding options. An option is considered unfavourable if 1) its delivery probability does not meet a given threshold; 2) its delivery probability is assessed by a sufficient number of feedbacks. Since excluding a forwarding option is equivalent to refuse an admission on a link, *CN* performs by this way the Channel Admission Control operation. The threshold is thus noted T_{CAC} and the number of required feedbacks F_{CAC}.

Eventually, *CN* will try to schedule a reservation on the links corresponding to the remaining forwarding options. The possibility showing the best delivery probability is considered first. If no reservation can be scheduled on the corresponding link, the next options are tried in a decreasing delivery probability order. If no reservation has been achieved after having considered all options, the burst is not admitted and a negative feedback is sent.

To exemplify the core node admission procedure, we consider the situation depicted in Fig. 2. A burst destined to node 3 ($D = 3$) with 2 units of the remaining offset ($R = 2$) arrives at node 1 from node 4. Node 1 has three forwarding options: 0, 2, 5. Due to probability $0 < T_{CAC}$ and number of feedbacks $20 \geq F_{CAC}$, forwarding to node 5 is excluded by the CAC mechanism. Node 1 tries first to schedule the burst on the link toward 0 ($P_{320} = 0.979$). If the link is congested, it will eventually try to schedule the burst toward node 2. Note that no positive feedbacks have been received for next-hop node 5, since node 3 cannot be reached from 1 via 5 in two hops.

Figure 2 depicts the different step of the core node decision process.

It is worthy to remark that the CAC mechanism is also achieved at the very first core node a burst traverses (the "access core node"). If this node notices that the large majority of the burst formerly sent to a given destination have been dropped, it will start to drop all the traffic toward this node before this traffic consumes any network resource. However, by doing this, it will not receive feedbacks for this destination anymore. The feedback received will thus eventually pass under the F_{CAC} limit, and the node will then try again forwarding this traffic.

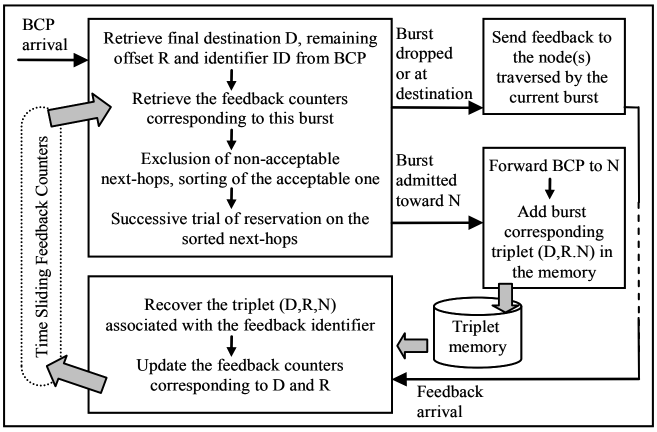
BCP arrival
Retrieve final destination D, remaining offset R and identifier ID from BCP
Retrieve the feedback counters corresponding to this burst
Exclusion of non-acceptable next-hops, sorting of the acceptable one
Successive trial of reservation on the sorted next-hops
Burst dropped or at destination
Send feedback to the node(s) traversed by the current burst
Burst admitted toward N
Forward BCP to N
Add burst corresponding triplet (D,R.N) in the memory
Triplet memory
Time Sliding Feedback Counters
Recover the triplet (D,R,N) associated with the feedback identifier
Update the feedback counters corresponding to D and R
Feedback arrival

Figure 1. Operational block diagram of the scheme.

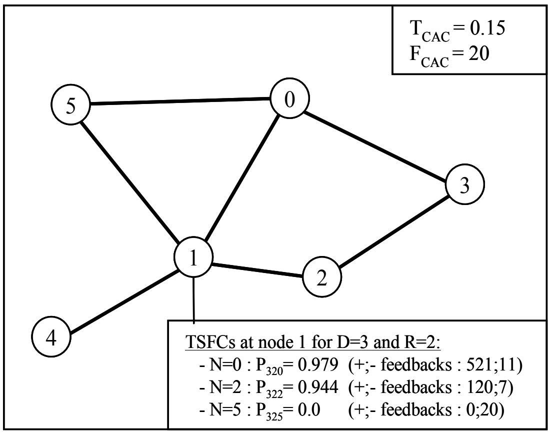
T_CAC = 0.15
F_CAC = 20
5
0
3
1
2
4
TSFCs at node 1 for D=3 and R=2:
- N=0 : P_320 = 0.979 (+;- feedbacks : 521;11)
- N=2 : P_322 = 0.944 (+;- feedbacks : 120;7)
- N=5 : P_325 = 0.0 (+;- feedbacks : 0;20)

Figure 2. Example topology and TSFCs status at node 1 for destination 3 and remaining offset 2.

4. NUMERICAL RESULTS

4.1 Simulation model

Simulations have been driven on the COST 266 topology with 27 nodes and 41 links. Additional scenarios involving shortest path routing and deflection routing have been simulated under the same conditions for comparison purposes. 32 wavelengths at 10 Gbit/s have been assumed on all links. Burst inter-arrival times and sizes have been generated according to exponential distributions. The mean burst size is fixed to 1 Mb, while the mean inter-arrival time depends on the traffic rate. Rate is normalised to the link capacity (320 Gbit/s). Simulations have been conducted using the simulator presented in [15], employing the FirstFit Void Filling (FFVF) as a scheduling algorithm.

The adaptive routing scheme has been implemented according to the description given in section 3. To simplify the implementation, the delay required for the feedback to reach their destination has not been taken into account. In other terms, feedbacks are instantaneously notified to the core nodes. The following parameter values have been used: $T_{mem} = 800$ microseconds, $f = 80$, $T_{CAC} = 0.35$ and $F_{CAC} = 500$.

The shortest path routing strategy does not provide any load-balancing or deflection routing capability. When a core node fails to schedule a reservation on the next hop corresponding to its destination, the burst is dropped. Offset times corresponding exactly to the length of their trip are granted when this strategy is used.

The deflection routing strategy has been implemented according to the two following rules. 1) A node deflects a burst to a next node only if the shortest path toward destination via this next node is shorter or equal to the burst remaining offset. 2) Next nodes closer to destination are selected in priority.

In the same way as in our adaptive approach, additional offset times are be granted to the bursts in the deflection routing strategy.

4.2 Analysis of the adaptability of the mechanism to traffic changes

At initialisation, core nodes do not have knowledge of the network size or connectivity. Therefore, in the very first moments of network activity, the bursts are forwarded in random directions, leading to a high loss rate. However, as the feedbacks are collected and taken into account, the adequate routing directions are progressively deduced. This transitory phenomenon is depicted in Fig. 3. Samples are computed each millisecond and represent the part of the traffic lost in the last ms. 3 units of additional offset time have been granted to burst.

Figure 4 depicts the performance of the mechanism in presence of sudden changes in the traffic pattern. During the first 200 ms of the simulation, each node exchanges an equal amount of traffic with all other nodes in the network. In the next 200 ms, there is a fourfold increase of the traffic generated by a quarter of the nodes, while the three other quarters stop their emissions. After the traffic pattern changes, the adaptation is performed and the amount of bursts lost is reduced. Samples are measured each 20 ms. 3 additional offset time units have been again granted to each burst

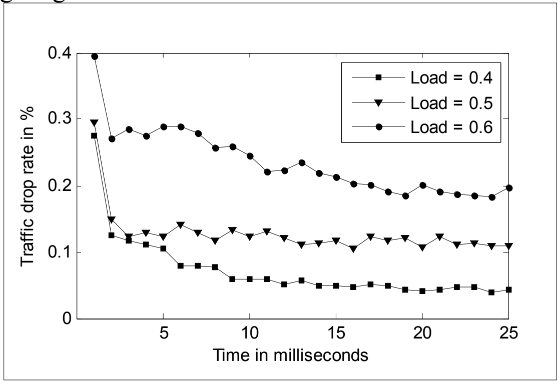

Figure 3. Transitory phase

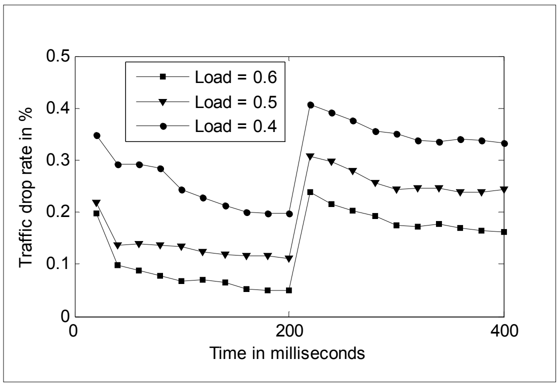

Figure 4. Reaction to the change of traffic pattern

4.3 Comparison with shortest path and deflection routing strategies

In the next step we compare the performance of our mechanism in terms of blocking rate with the performance achieved by both shortest path routing and deflection routing. Partial results are available in Fig. 5.

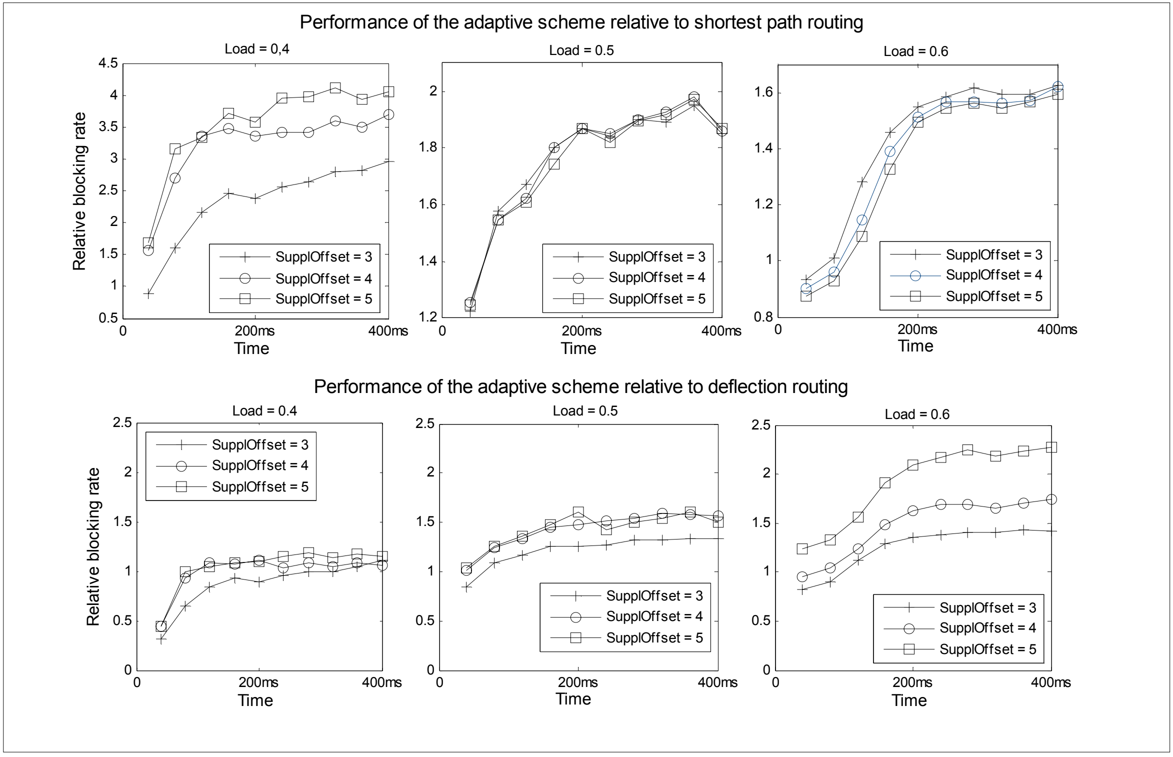

Figure 5. Relative performances of the adaptive forwarding and admission control mechanism.

Except during the transition time, our adaptive mechanism always outperforms the shortest path routing scheme. However, the improvement factor highly depends on the network load, and while blocking rate are about four times smaller for the adaptive approach at a load of 0.4, this coefficient decreases to values inferior to 2 under higher loads. This variation in the blocking gain is due to intrinsic topology properties: until a certain load, flows can be rearranged advantageously in the network; for higher loads, the capacity will inevitably lack and bottlenecks will appear. One can also note that providing larger offset time supplements leads performance improvement for a load = 0.4, but tendency is inversed for high loads. This is due to the fact that burst with larger offset are likely to consume more network resources.

When compared to the deflection routing, the adaptive approach leads to a similar performance for load = 0.4 and better performance under higher loads. This effect is due to the CAC mechanism which discards the traffic overhead earlier than the deflection routing, sparing network resources. Also, due to the ability of the CAC to drop the traffic rather than awaiting its offset time exhaustion, granting large additional offset time does not lead to performance penalties in the adaptive scheme.

5. CONCLUSIONS

An adaptive mechanism, based on feedback messages sent by network nodes, has been proposed for burst admission and routing at core nodes. According to numerical results obtained by simulation, this adaptive approach outperforms the shortest path routing in terms of burst blocking rate, while the performances of deflection routing are reached or even beaten.

The main advantage of this approach relies in the fact that core nodes do not require any a priori knowledge of the network. Unfortunately, this is not the case for edge nodes which still need to know the length of the shortest path to be able to grant the right amount of offset time to each burst.

The proposed scheme could be combined with another mechanism granting offset without network knowledge. This is definitely the next step of this study. In the future, we will also analyse the effect of particular parameter values on the scheme performances.

ACKNOWLEDGEMENTS

This work was partially funded by the European Commission under the BONE (Building the Future Optical Network in Europe) Network of Excellence of the 7th ICT Framework Program.

REFERENCES

[1] C. Guillemot *et al.*, " Transparent optical packet switching: the European ACTS KEOPS project approach", *Journal of Lightwave Technology*, vol. 16, no. 12, December 1998.

[2] C. Qiao, M. Yoo, "Optical burst switching (OBS) – A new paradigm for an optical Internet", *Journal of High Speed Networks*, vol. 8, no. 1, Jan. 1999.

[3] L. Yang, G. N. Rouskas, "Adaptive path selection in OBS networks", *IEEE Journal of Lightwave Technology*, vol. 24, no. 8, Aug. 2006.

[4] J. Zhang, H.-J. Lee, S. Wang, X. Qiu, K. Zhu,Y. Huang, D. Datta, Y.-C. Kim, B. Mukherjee "Explicit routing for traffic engineering in labeled optical burst-switched WDM networks", *Computational Science - ICCS 2004*, LNCS 3038, Springer, 2004.

[5] M. Klinkowski, M. Pióro, D. Careglio, M. Marciniak, and J. Solé-Pareta, "Non-linear optimization for multipath source-routing in OBS networks", *IEEE Communications Letters*, vol. 11, no. 12, 2007.

[6] C.-F. Hsu, T.-L. Liu, N.-F. Huang, "Performance analysis of deflection routing in optical burst-switched networks", *IEEE INFOCOM* 2002.

[7] S. Gjessin, "A novel method for re-routing in OBS networks", *7th International Symposium on Communications and Information Technologies*, 2007.

[8] A. Zalesky, H. L. Vu, Z. Rosberg, E. W. M. Wong, M. Zukerman. "Reduced load Erlang fixed point analysis of optical burst switched networks with deflection routing and wavelength reservation", *First International Workshop on Optical Burst Switching (WOBS)*, Oct. 2003.

[9] A. Lazzez, N. Boudriga, "Admission control in OBS networks: A real time QoS oriented approach", *IEEE International Conference on Signal Processing and Communications*, Nov. 2007.

[10] F. Farahmand, Q. Zhang , J. P. Jue, " A feedback-based congestion control mechanism for labeled optical burst switched networks", *Photonic Network Communication*, Springer, 2007.

[11] S. Yao, B. Mukherjee, S. J. B. Yoo, S. Dixit "A unified study of contention-resolution schemes in optical packet-switched networks", *Journal of Lightwave Technology*, vol. 21, no. 3, Mar. 2003.

[12] S. Ganguly, S. Bhatnagar, R. Izmailow, C. Qiao, "Mutli-Path Adaptive Optical Burst Forwarding", *Workshop on High Performance Switching and Routing*, May 2004

[13] J. Teng, G. N. Rouskas, "A detailed analysis and performance comparison of wavelength reservation schemes for optical burst switched networks", *Photonic Network Communication*, vol. 9, May 2005.

[14] M. Klinkowski, D. Careglio, J. Solé-Pareta, and M. Marciniak, "Performance Overview Of The Offset Time Emulated obs network architecture", to appear in *IEEE/OSA J. Lightwave Technol.*, 2009.

[15] O. Pedrola, S. Rumley, M. Klinkowski, D. Careglio, C. Gaumier, J. Solé-Pareta, "Flexible simulators for OBS network architectures", *International Conference on Transparent Optical Networks (ICTON)*, June 2008.