Photon Netw Commun (2009) 18:323–333
DOI 10.1007/s11107-009-0195-9

On transmission control protocol synchronization in optical burst switching

Oscar González de Dios · Anna Maria Guidotti · Carla Raffaelli · Kostas Ramantas · Kyriakos Vlachos

Received: 16 October 2008 / Accepted: 12 February 2009 / Published online: 9 March 2009

Abstract This article studies the transmission control protocol (TCP) synchronization effect in optical burst switched networks. Synchronization of TCP flows appears when optical bursts with segments from different flows inside are dropped in the network causing flow congestion windows decreasing simultaneously. In this article, this imminent effect is studied with different assembly schemes and network scenarios. Different metrics are applied to quantitatively assess synchronization with classical assembly schemes. A new burst assembly scheme is proposed that statically or dynamically allocates flows to multiple assembly queues to control flow aggregation within the assembly cycle. The effectiveness of the scheme has been evaluated, showing a good improvement in optical link utilization.

Keywords Transport control protocol · Synchronization · Optical burst switching

O. González de Dios
Telefónica I+D, Emilio Vargas 6, Madrid, Spain
e-mail: ogondio@tid.es

A. M. Guidotti · C. Raffaelli (✉)
DEIS – University of Bologna, Viale Risorgimento, 2-40136 Bologna, Italy
e-mail: carla.raffaelli@unibo.it

K. Ramantas · K. Vlachos
Computer Engineering and Informatics Department and Research Academic Computer Technology Institute, University of Patras, Patras, Greece

K. Vlachos
e-mail: kvlachos@ceid.upatras.gr

1 Introduction

Transmission control protocol (TCP) is the de facto standard in transport protocols, used by most of the user applications, such as web browsing, e-mail, or file transfers. TCP is also expected to be the dominant transport protocol for a long time. Thus, when studying a new networking paradigm for the future Internet, like optical burst switching (OBS) [1], it is necessary to evaluate network performance considering the characteristics of this kind of upper layers. A critical issue which can impact the TCP performance over OBS network is represented by burst loss, which can be interpreted by the TCP layer as congestion in the network and hence may, unnecessarily, reduce the transmission window, even at low loads. The probability of a burst drop depends upon the network load and/or burst contentions inside a core node. In the literature, many studies on how to reduce burst losses [2,3] have been carried out and, recently, several works have addressed some of the main OBS functions such as the burst assembly algorithm [4]. It has been shown that the burst assembly algorithm significantly affects TCP performance [5–7], since it determines how the different flows are aggregated together to form a burst. In general, burst assembly algorithms can be classified as timer based [8–10], threshold based [11], and hybrid timer/threshold based [12]. TCP performance evaluations over OBS networks have been carried for different TCP versions [13,14] and useful traffic statistics are given [9,15,16]. In this article, the synchronization effect in OBS networks is studied and, in particular, link utilization, throughput, and its standard deviation changes with the number of aggregated flows for different assembly schemes and for different network cases are analyzed. The synchronization phenomenon over OBS network has not been yet widely studied and it depends on the aggregation of the different flows in the same burst [17]. It appears when a

burst is dropped and, consequently, many segments belonging to different flows are simultaneously lost. To quantify this effect, a *synchronization index metric* is introduced in this article whose calculation is based on the number of losses per flow. Segment aggregation in bursts being the main cause of TCP synchronization in OBS networks, it can be weakened by introducing dynamic allocation of flows to different burst assembling queues. In this study, different burst assembly schemes are considered in relation to synchronization and evaluated: per flow (PF), mixed flow (MF), static multiple queue (SMQ), and dynamic multiple queue (DMQ). The PF represents an ideal scheme, where a single queue is employed per flow, while MF represents the classical scheme, employing a single burst assembly queue for all flows. The SMQ and DMQ represent new schemes, where segments from different flows are statically or dynamically divided into groups and later aggregated with MF strategy into the some burst. Both these schemes require a number of queues equal to the number of flow groups per source–destination pair. With the aim to investigate in depth the synchronization phenomenon and its impacts on the transmission, different network scenarios have been evaluated. The first analysis is carried out over a simple access network scenario with a single link, while then, a large network topology is being considered.

The rest of the article is organized as follows. Section 2 reviews the problem of flow synchronization and analyzes the effect of synchronization in OBS network. Section 3 outlines the first results on an example of access network and introduces the SMQ scheme. Section 4 studies the effect in a large-scale network under real traffic conditions and further proposes a dynamic assembly scheme, to limit flow synchronization. Section 5 presents the main achievements and the conclusions of the work.

2 TCP synchronization effect

2.1 What is TCP synchronization?

TCP synchronization is a well-known effect in packet switched networks [18]. This effect appears when several TCP connections share a link and the end-to-end control mechanisms of the different edges of the TCP flows react at the same time. To understand this phenomenon, the fundamentals of TCP end-to-end control mechanisms are reminded. Each TCP connection increases and decreases its bandwidth occupancy basically by applying the AIMD [19] principle, which aims at avoiding congestion. Ideally, if the connections decrease their window at different moments, a smooth usage of outgoing capacity of the shared link can be achieved. However, if these moments coincide, the outgoing traffic will resemble a saw-tooth profile, and a lot of bandwidth will be wasted. This effect, when multiple TCP connections increase and decrease their transmission window simultaneously, is called TCP synchronization.

2.2 Why TCP synchronization appears in Internet?

In current Internet, buffer overflowing is the prime cause of TCP synchronization [20]. In particular, Internet routers are provided with buffers to accommodate temporarily bursts of traffic. There are several queue management strategies adopted by routers. Currently, the DropTail [21] scheme is widely used in the routers. With this strategy, when the buffers are full, all incoming traffic has to be dropped. Thus, there is a high risk of dropping packets of multiple flows in a row, thus synchronizing their end-to-end congestion control mechanisms. However, more intelligent queue management schemes, like AQM [22], can help to prevent synchronization. For example, RED [23] starts dropping packets randomly before the buffer gets full. It is claimed to have several benefits, including the ability to prevent large number of consecutive packet losses by ensuring available buffer space even with bursty traffic. However, RED is still not widely deployed, and synchronization is still present in Internet.

2.3 Why TCP synchronization appears in OBS networks?

OBS networks are rather different than interconnected IP routers with buffers. In OBS networks, packets from different flows are aggregated together at optical network edge in burst containers and transmitted all-optically from source to destination. The number of flows and segments per burst varies with the assembly time as well as with the instant congestion window size of each flow. Thus, each burst contains several segments from many different flows. Core OBS routers are typically bufferless, and in case the bursts do not find available resources, usually it has to be dropped, that leading to loss of segments from different flows. As a consequence, all affected flows will trigger their end-to-end congestion mechanisms at the same time.

2.4 Effect of TCP synchronization

To better illustrate the effect, and highlight the potential waste of bandwidth, a brief introduction about the most important effects of the TCP synchronization over the OBS network is outlined.

As mentioned earlier, a very important issue of the OBS network is the aggregation of IP segments in a burst. When at the edge node, a MF strategy is implemented; after a drop event many flows suffer of simultaneous segment loss. This fact reduces the transmission rates of the TCP sources and causes an irregularity. On the contrary, a more flat traffic pro-

file is caused when a PF strategy is used, where the losses are uncorrelated. Available bandwidth is used more rationally without synchronization.

Moreover, some other experiments show that when limiting the access bandwidth, which means that the bandwidth is used more efficiently, the performance of the transmission is better without synchronization. Thus, the most important goal of this study is to reduce, in our network, synchronization as less as possible.

2.5 How can TCP synchronization be measured?

Once we have seen the effect, and shown how can it affect, we need to find metrics to measure the level of synchronization found and how dangerous is it. Synchronization can be measured by observing congestion window evolution of TCP flows. When multiple flows [24] with the same round trip time (RTT) are considered, the evolution of the aggregate throughput captures the evolution of the bandwidth usage. Two basic metrics are considered here, namely the average and the standard deviation of the aggregated throughput, which can track the variations in link usage profile induced by the synchronization.

Another way of detecting synchronization is by directly monitoring the outgoing traffic profile, i.e., at the output of the edge router or at the input of the core router after flow assembly. The latter is preferred especially in the case when flows that share the OBS edge router do not have the same RTT value (i.e., when employing different access delays). In that case, we have measured the bytes sent by the OBS router periodically, by calculating also on the average and standard deviation of the aggregated throughput.

A synchronization index is here also defined. Let us sample the window of each TCP flow at fixed intervals and count and sum the number of dropped flows per sample (N_{di}). Then, by dividing by the total number of samples related to all flows that contributed to loss, that is the product of the number of samples Ns_{di} where at least one flow had losses, and the number of flows (N_{F}), the synchronization index is defined as:

$$I = \frac{\sum^{i} N_{\mathrm{di}}}{N_{\mathrm{F}} \cdot \sum^{i} Ns_{\mathrm{di}}} \tag{1}$$

The aim of the proposed index is to define the percentage of flows that had losses in the sampled window. If at every sampling point a loss occurs, N drops are counted, I equals to one, and full synchronization is indicated. In addition, to understand how the aggregation function impacts on the synchronization phenomenon, the distribution of flows in bursts is evaluated.

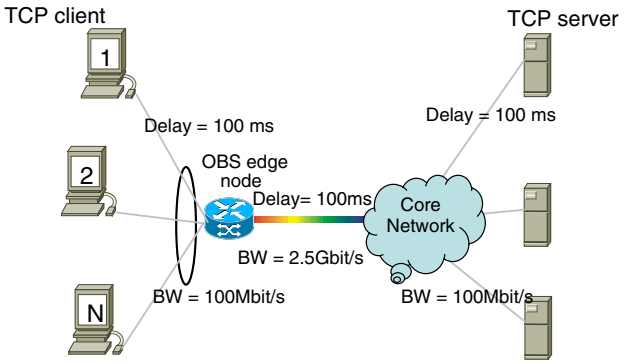

Fig. 1 Simulation scenario for an access network over a single link

3 Synchronization on access network

In this section, the intrinsic features of synchronization are studied using as an example of an access network over a single optical link, the experiment setup shown in Fig. 1. The goal of this first simulation is to study the single optical link utilization when TCP clients (or subnetwork) share the access bandwidth at the edge node ingress. The edge router supports multiple TCP agents and implements a time-based assembly algorithm. We used ns-2 [25] simulator, with dedicated tools to emulate a specific source–destination pair of edge router, each one with multiple TCP SACK agents attached, CBR traffic sources, and three different assembly schemes:

- *PF* queuing, where a different assembly queue is assigned per flow,
- *MF* scheme, where a single burst assembly queue serves all flows (normal case in OBS networks),
- *multi-queue* (MQ) scheme where more than one queue is employed per source–destination pair, and packets are assigned either statically or dynamically.

Multiple queue burst assembly schemes with static or dynamic allocation are more complex and have been mainly proposed for QoS differentiation or adaptive burst assembly [9]. We have considered the multiple queue approach, which is expected to limit the synchronization effect. The parameters of the simulation are summarized in Table 1. The delay values take into account geographical distance of some hundreds of miles and equipment delays.

This first set of simulations is carried out with an access bandwidth of 100 Mbit/s, shared by all flows and varying the number of active (N_{F}) flows. Thus, as the number of active flows increases, their share in the core bandwidth decreases. The evolution of the congestion window is monitored for each TCP agent, and it is sampled every 0.7 s during the simulation run. Finally, TCP agents start their transmission at random time between 0 and 50 s.

Table 1 Simulation parameters

Max window size (W_M)	128 MSS
Assembly timeout (T_{MAX})	3 ms
Access link delay (d)	100 ms
Burst drop probability (p)	0.001
Optical link delay (T_t)	100 ms
Optical bandwidth (B_o)	2.5 Gb/s

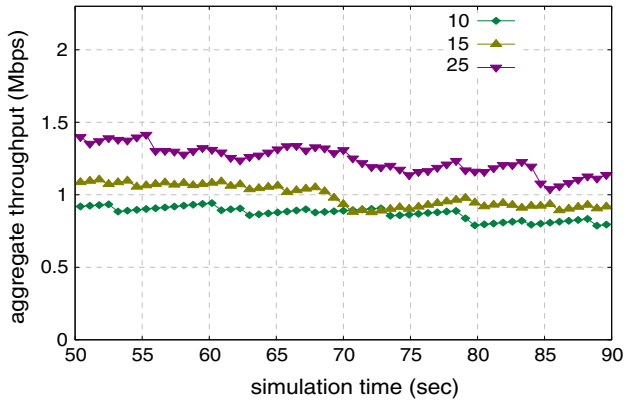

Fig. 2 Aggregated throughput of TCP flows over 600 s simulation run, plotting from 50 to 90 s for PF assembly, varying the number of TCP flows, N_f

In this section, first, simulation results showing the performance of the TCP synchronization with a PF and MF scheme are presented. Then, the results of the new assembly scheme, SMQ, is outlined to put in evidence that the multiple queue approach limits the synchronization phenomenon.

3.1 Evaluation of the TCP synchronization effect with Bernoulli losses

Figures 2 and 3 display the aggregated throughput for $N_F = 10$, 15, and 25 sources in the cases of PF and MF assembly, respectively, while Fig. 4 displays again the aggregated throughput but for a wider time span for the MF case. The aggregated throughput is calculated as the sum of all the congestion windows divided by RTT. In the case of no synchronization, the aggregated throughput would exhibit a nearly flat profile as shown in Fig. 2 for the ideal case of PF queuing. In case of synchronization, the profile is expected to have a saw-tooth profile, as shown in Figs. 3 and 4 for MF assembly, where the aggregated sending rate abruptly drops when a burst drop occurs. This instability is due to the fact that multiple flows and multiple segments per flow are present in the dropped burst and cause several flows to decrease the size of their transmission window simultaneously.

As shown in Fig. 3, when the number of flows increases the throughput dropped is more evident. In this case, more

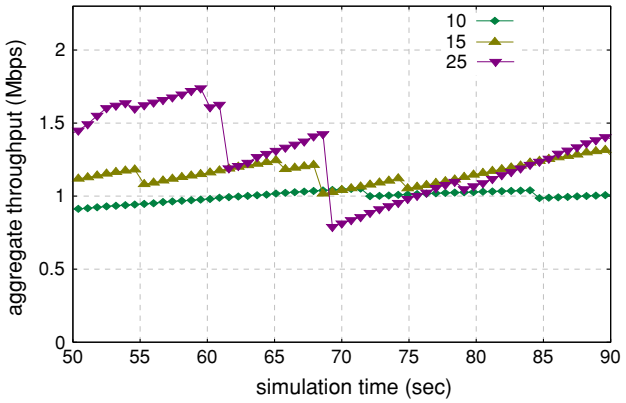

Fig. 3 Aggregated throughput of TCP flows over 600 s simulation run, plotting from 50 to 90 s for MF assembly, varying the number of TCP flows, N_f

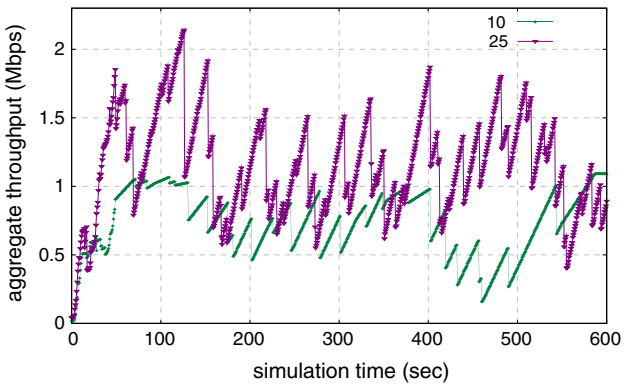

Fig. 4 Aggregated throughput of TCP flows over 600 s simulation run, for MF assembly, varying the number of TCP flows, N_f

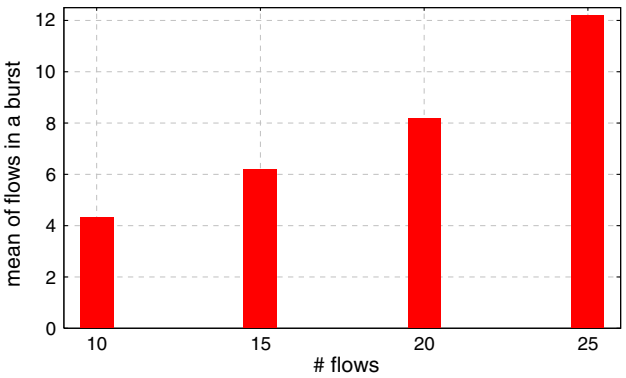

Fig. 5 Mean number of different flows in the same burst for 10, 15, 20, and 25 flows, over 600 s simulation run and MF assembly

flows are able to send at least a segment in the same burst, causing a significant slowdown of the transmission rate. At the same time, the mean number of different flows per burst transmitted is getting higher as shown in Fig. 5.

In addition, it can be seen that fluctuations become sharper, when the number of flows increases (see Fig. 4) in the case of MF strategy. In fact, due to the constant access rate selected,

the segment transmission rate of each flow decreases as the number of flows increases. In other words, each TCP agent is able to send less segments over the same burst, and thus becomes more difficult to reach the maximum window size, i.e., each agent needs a higher number of bursts to send all the window. Increasing the number of connections grows the number of bursts generated during an RTT period. As shown in Fig. 4, this impacts the aggregated throughput significantly.

These first results outline how the synchronization of multiple flows influences the stability of the throughput during transmission, and thus causing a bursty usage of the optical channel. This has been shown to be a consequence of the presence of many flows in the dropped bursts.

3.2 Influence of the burstification timeout

To understand in depth the synchronization phenomenon and evaluate its impact on the burst assembly function, simulations varying the assembly parameters were carried out. The normalized congestion window is considered in Fig. 6, defined as the sum of congestion windows of all flows normalized by the maximum value of congestion window multiplied by NF. This normalization gives an estimation of how the flows are able to reach the maximum value of the sending rate. As shown in Fig. 6, when the assembly time increases the value of the normalized congestion window gets close to 1. The figure shows also that the fluctuation of the normalized congestion window for a 5-ms timer is less frequent but more evident than for 1- and 3-ms timers. In the case of a 5-ms timer, the burst can contain multiple segments, and thus fewer bursts are needed to carry all segments of the same congestion window. On the other hand, during the drop events, multiple segments of different flows are lost simultaneously and multiple TCP agents reduce their congestion windows. Consequently, the synchronization phenomenon is more evident than in case of lower timeout values. In the cases of timeout = 1 ms and timeout = 3 ms, each burst can contain fewer segments, so more bursts are needed to carry the whole congestion window. Consequently, the synchronization phenomenon is weaker, but still present.

Figure 7 shows the distribution of flows in the bursts varying the timeout value. The percentage of bursts with the highest number of flows inside reaches its highest value, about 62%, for timeout = 3 ms. This result is related to strong synchronization because a high number of bursts contains multiple segments per flow: this causes synchronization when drop occurs. When the timeout value is smaller, the distribution of flows in bursts is more uniform, see Fig. 7. This means that the percentage of flows in a burst is often less than the highest range 80–100% and synchronization is in this case weaker. When the timeout is very high, the percentage of bursts with a high number of flows inside decreases, since

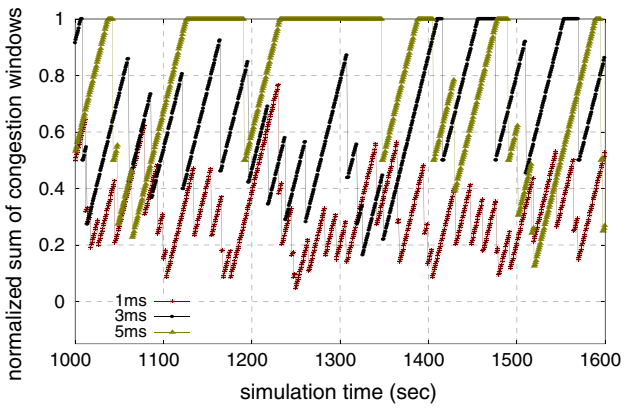

Fig. 6 Normalized sum of congestion windows measuring in segments for 15 TCP flows from $t = 1000$ to 1600 s over 2000 s run simulation for MF assembly varying the $T_{out} = 1$, 3, and 5 s

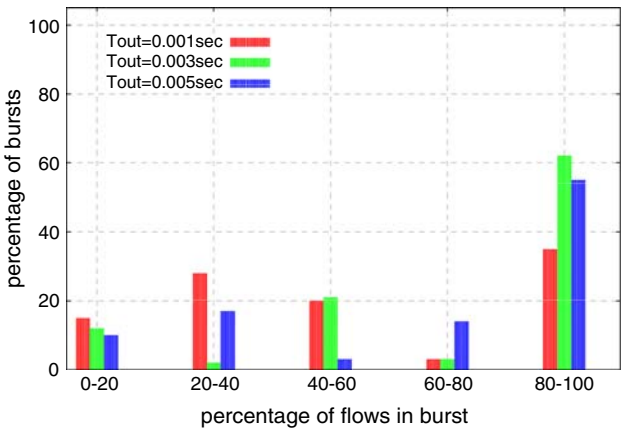

Fig. 7 Percentage of bursts with a given percentage of different flows inside, varying the $T_{out} = 1$, 3, and 5 ms, for MF assembly, $N_f = 15$ over 2000 s simulation run

more flows are awaiting longer for acknowledgments. Still, the percentage of burst with at least one segment per flow is 55%.

These results have shown how the timeout value impacts on the distribution of flows in the burst, i.e., on the synchronization. In fact, when the timeout value increases, the flows are able to carry more segments and the probability that more flows are assembled in the same burst is higher. To limit the synchronization phenomenon, it would be better to set a lower timeout value but, as shown in Fig. 6, with short assembly timers, is more difficult to reach the maximum sending rate. So a trade-off must be found.

In Fig. 8, the coefficient of variation (CoV) for the aggregate throughput is calculated as the standard deviation of the aggregate throughput divided by the mean value of the aggregate throughput, varying the loss probability to $p = 0.01$, 0.001, and 0.0001 for different values of $T_{out} = 0.001$, 0.003, and 0.005 s. This coefficient gives an estimation of the

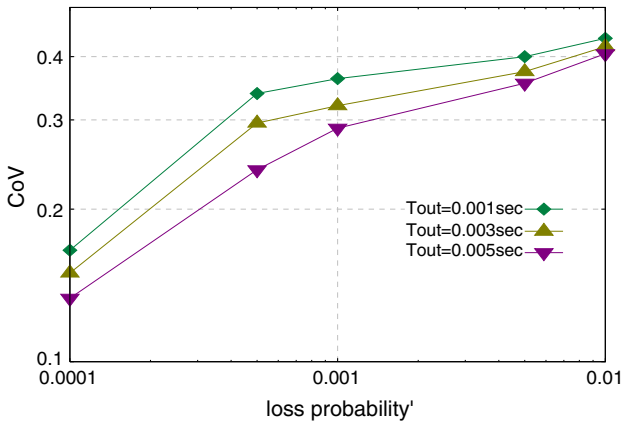

Fig. 8 Coefficient of variation (CoV) of aggregate throughput as a function of loss probability $p = 0.0001$, 0.0005, 0.001, 0.005, and 0.01 over 2000 s simulation run for MF assembly, with $N_f = 15$ for different timeout value, $T_{out} = 0.001$, 0.003, and 0.005 s

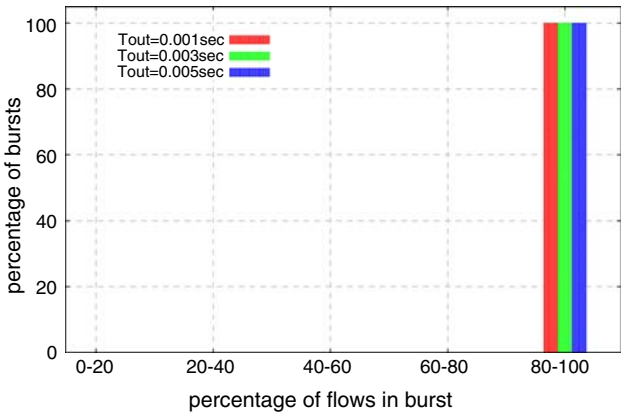

Fig. 9 Percentage of bursts with a given percentage of different flows inside with $T_{out} = 1$, 3, and 5 ms when TCP agents start their transmission at same time, for MF assembly, $N_f = 15$ over 2000 s simulation run

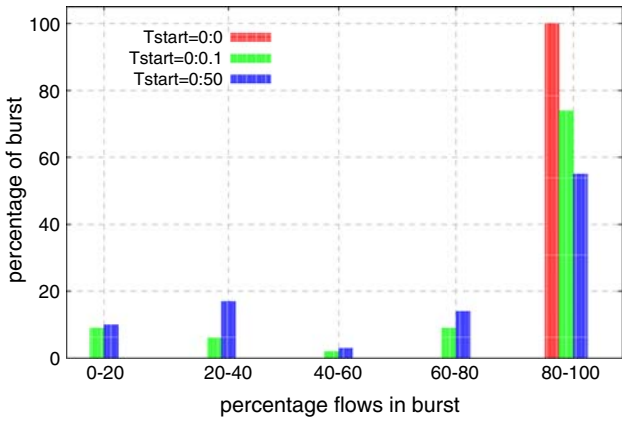

Fig. 10 Percentage of bursts with a given percentage of different flows inside with $T_{out} = 3$ ms, varying TCP agents start transmission time $T_{start} = 0{:}0$, 0:0.1, and 0:50 s, for MF assembly, $N_f = 15$ over 2000 s simulation run

fluctuation of the aggregate throughput with respect to the average throughput value.

Figure 8 shows how, given the loss probability, the CoV decreases when the timeout value increases. This phenomenon puts in evidence, simultaneously, the advantages and disadvantages of the flows correlation; in fact, when the timeout value increases, the bursts are able to aggregate more segments per flow so that the mean value of aggregate throughput grows; on the other hand, as shown in Fig. 7, by increasing the timeout value more flows are assembled in the same burst and many segments and flows are present in the dropped bursts, causing the fluctuation of the aggregate throughput. Anyway, even if many flows are present in a burst during drop events and the synchronization effect is more evident, the average of the aggregate throughput grows and the CoV value is lower. It is very interesting to see that when the loss probability is very high, the CoVs with different timeout values are close, that means that when the loss probability is very high, the transmission of the burst became very difficult .

When the loss probability increases, the number of the losses is higher and the coefficient value increases.

3.3 Influence of the start transmission time

In previous evaluations, TCP flows start their transmission at random time within a fixed range. To understand how the starting time impacts on the distribution of flows in the burst, simulations with different ranges of starting times were carried out. Figure 9 shows that when the flows start their transmission at the same time, there are 100% of bursts with at least one segment per flow, i.e., complete flow synchronization. In contrast, when flows start their transmission in a range of 0–50 s the percentage of the bursts with at least one segment per flows is 55% and the distribution of the flows is more uniform (see Fig. 10). The results put in evidence, how the distribution of flows is affected by their starting time, i.e., if the range of their arrival time grows, then the flows get less synchronized. Clearly, the synchronization phenomenon is not removed but is lower than in the case of simultaneously starting times.

3.4 Static multiple queue scheme (SMQ)

The results outlined how the synchronization of multiple flows influences the stability of the throughput during transmission. As seen in the previous sections, the synchronization depends on the burst losses and is affected by the assembly function as the timeout value, start transmission time, and assembly scheme. This has been shown to be a consequence of the presence of many flows in the dropped bursts. With the aim to reduce this instability, a new burst assembly strategy based on the multiple queue per FEC is applied to limit

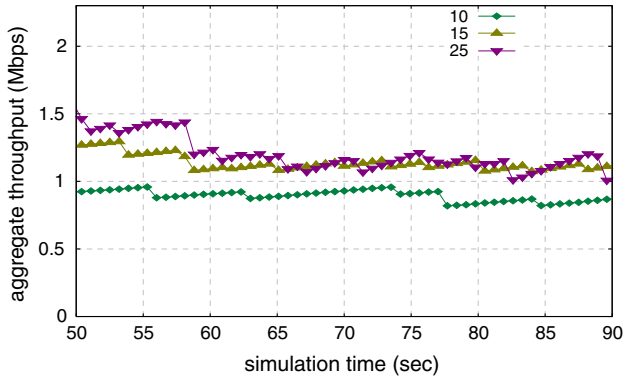

Fig. 11 Aggregated throughput of TCP flows over 600 s simulation run, plotting from 50 to 90 s for SMQ assembly, varying the number of TCP flows, N_f.

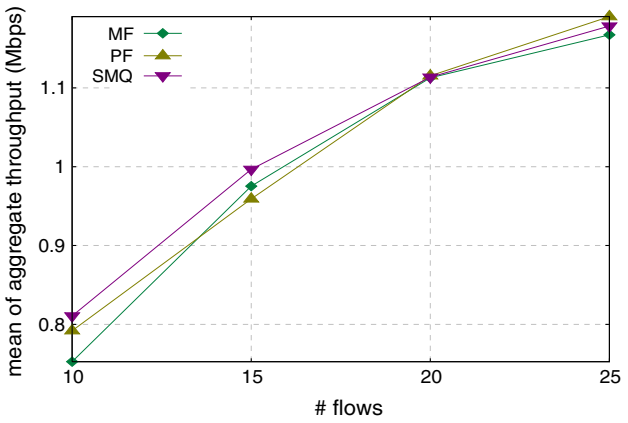

Fig. 12 Average aggregated throughput for SMQ, PF, and MF assembly schemes

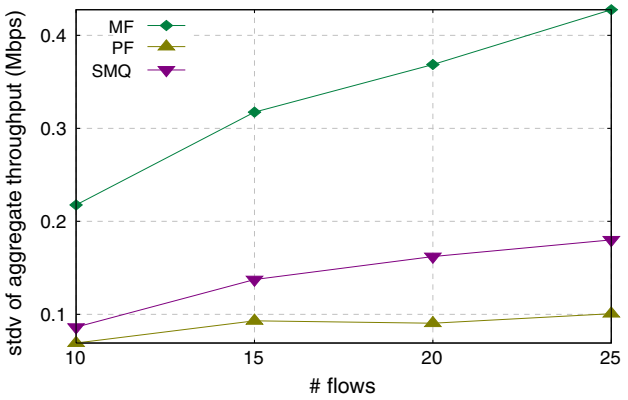

Fig. 13 Standard deviation of aggregated throughput for SMQ, PF, and MF assembly schemes

the number of flows in the same burst and, consequently, the synchronization effect. The assembly scheme with multiple burst assembly queues is here evaluated in the case of two assembly queues per FEC. Active flows are statically assigned to these and their segments are being aggregated in each queue as in the MF case. Figure 11 displays the corresponding aggregated throughput. Comparing Fig. 11 with Fig. 3, the reduction of congestion window dynamics is evident, especially for 25 flows. The aggregated behavior with smoother peaks of the multiple queue scheme is more similar to the PF behavior, shown in Fig. 2.

3.5 Average throughput and standard deviation

To reveal the effect of the synchronization phenomenon over the stability of transmission, the average value of the aggregated throughputs and the related standard deviations are here given for the three different assembly schemes, studied above. Figures12 and 13 plot the corresponding results for PF, MF, SMQ versus the number of flows. In Fig. 12, the average throughput is very similar in all the assembly schemes and increases with the number of flows, in spite of the higher synchronization.

When the number of flows increases, the bandwidth of each flows decreases accordingly. This causes in PF the generation of higher number of shortest bursts while in MF a less number of segment of each flow in a burst are assembled. In any case, a burst loss is better and faster recovered by the sack congestion control as the number of flows increases and higher values of average throughput is obtained.

Instead, when few flows are connected, a higher number of segment are aggregated in the same longer bursts which make loss recovery slower and this cause PF performance worst than MF and SMQ.

On the other hand, stability is much different among these schemes, as shown in Fig. 13. In fact, the standard deviation of the aggregated throughput well represents the dynamic of the transmission, which is much higher for MF than for PF. In the MF case, throughput stability sensibly depends on the number of aggregated flows, as can be seen by the increase of standard deviation with the number of flows. The SMQ scheme is less sensitive to the number of flows and its performance is closer to the PF case.

4 Synchronization effect in large networks with dynamic multiple queue (DMQ) burst assembly

Based on the previous analysis, it is clear that burst losses are the driving cause for flow synchronization. It was shown that the increase of assembly time results in a slow (multi-sec) but strong synchronization of a large number of flows, while shorter timeouts to weaker but fast (sec scale) synchronization effect. It is only the number of flows as well as which flows are being aggregated together per burst that matters, while burst loss ratio will determine how fast these flows will get synchronized. Thus, static allocation of flows to bursts is not enough and a dynamic process is investigated. Such a

dynamic allocation would hamper the continuous aggregation of the same sources over the same bursts. The target is to avoid same flows to appear in the dropped bursts continuously.

A DMQ burst assembly scheme is proposed by employing multiple queues to avoid the continuous aggregation of the same flows over the same bursts. Flows (and not segments) are assigned to these queues using a predefined flow allocation algorithm. The flow allocation algorithm may be *proactive* by aggregating together different flows per assembly cycle or a posteriori avoiding aggregating together flows that suffered from a segment loss in the same burst. Here performance of both will be investigated in simple two, four, and eight queue systems.

The allocation algorithm in both schemes is modeled by bounding alternate trials with n possible outcomes, equal to the number of queues. In the simple case of employing $n = 2$ burstifiers and $p_2 = 1 - p_1$, then $p_2 = p_1 = 1/2$, while the probability of k flow to be assigned as the first queue is:

$$P_0^k = \begin{cases} p_1, & k \text{ is even} \\ 1 - p_1, & k \text{ is odd} \end{cases} \quad \text{and to the second queue}$$
$$P_1^k = \begin{cases} 1 - p_1, & k \text{ is even} \\ p_1, & k \text{ is odd} \end{cases}. \tag{2}$$

In the general case of n queues, the probability P_i^j of j flow, with $j \in \{0, \ldots, N_f - 1\}$ to be assigned to queue $i \in \{0, \ldots, n-1\}$ is

$$P_i^j = \rho_i^{j \bmod (n-1)}, \tag{3}$$

where ρ_i^j is given by $\rho_{n-1}^{n-1} = 1 - \rho_{n-1}^{n-2}$. In the case of a posteriori flow allocation, the aforementioned algorithm is applied to only the flows that suffered from a burst loss. Allocation of flows to queues is initiated with the arrival of the first segment of each flow and was kept constant per aggregation cycle. In other words, when a segment of a new flow arrives, the flow allocation algorithm determines to which queue it should be forwarded. After decision is made, all segments of this flow are been forwarded to this queue without a new allocation decision, until the end of the assembly cycle. Flow allocation is reset only when the assembly period is ended and reinstated again with the arrival of new segments.

The TCP synchronization has been evaluated and compared for all assembly techniques on a large-scale network with a high number of flows. The experiments were carried out on the NSF network topology, with eight edge and six core nodes whereas each link was employing two wavelengths at 10 Gbps. Access rate was set to 100 Mbps equally for all sources. TCP arrivals were modeled with a Poisson process with a $\lambda = 50$ flows/s rate while TCP file size with a Pareto distribution process of p load and 40 KB minimum ON size. Using this set of metrics, it was possible to vary the TCP arrival rate and/or the mean file size, to obtain measurements

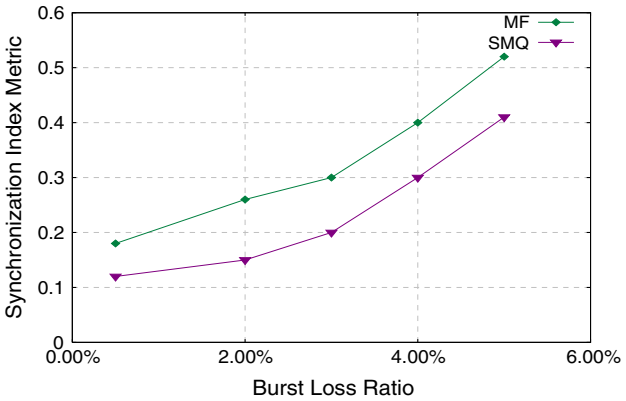

Fig. 14 Index metric change versus burst loss ratio for MF and SMQ assembly schemes

for different number of active flows. We have also employed LAUC-VF scheduling instead of the Bernoulli random process and monitored the actual burst loss ratio. In what follows we have selected a mean file size of 4 MB, which corresponded to ∼2000 active sources for an assembly time of 3 ms and to a 2% burst loss ratio.

One would expect that synchronization would be lower or even absent in such cases with a high number of active flows. However, as mentioned above, TCP synchronization in OBS networks does not depend actually on the number of flows being active but primarily on the burst losses and the distribution of flows over the transmitted bursts. To this end, we have first assessed the effect for different burst loss ratios. In particular, we have measured the yielding synchronization index metric by allowing full, partial, or no wavelength convertibility in the core, and thus obtaining different burst loss ratios. Figure 14 displays the corresponding results. As expected, TCP synchronization increases with the increase of losses and, in particular, it triples when loss increases from 0.5 to 5%.

Figures 15 and 16 display the corresponding aggregated throughputs for a specific source–destination pair for MF, SMQ schemes as well as in the case of dynamic flow allocation in two (DMQ-2) and four (DMQ-4) queues. Throughput is measured in time spans of 3 ms and it was derived by the link utilization profile that is the bytes/s received by the first core router. This is an absolute criterion of synchronization, especially when aggregated flows exhibit different round-trip-time delays. In principle, the data received within a round-trip-time frame equals to the number of all the unacknowledged segments; that is, the sum of congestion windows of all the aggregated flows:

$$\sum_{i=1,\ldots,N_f} CW_i / RTT_i \tag{4}$$

From Fig. 15, it is clear that TCP synchronization still exists in large networks with hundreds of TCP sources being simul-

Table 2 Performance summary of different assembly schemes

Assembly case	AVE	STD	MIN	MAX	Index metric
MF	35.5	61.1	0.66	661.7	0.26
SMQ	35.8	54.0	0.76	603.1	0.13
DMQ-2	38.3	34.4	0.78	258.4	0.043
DMQ-4	36.7	29.9	0.7	251.8	0.021
DMQ-8	37.5	26.2	0.65	247.2	0.011

taneously active. In particular, average aggregated throughput of all sources for a specific source–destination pair was measured to be ∼35 Mb/s in both the MF and SMQ schemes but the relative standard deviations were found to be 61.1 and 54, respectively. Similarly, the synchronization index metric was 0.26 and 0.13. With respect to DMQ scheme, from Fig. 16, it can be seen that synchronization has been significantly reduced and we may argue that the random but equal distribution of flows to different queues truly desynchronize transmission. In the results shown in Fig. 16, the standard deviation has been reduced to 34 and 29 for the DMQ-2 and DMQ-4 schemes, respectively. The higher gain in performance is obtained, however, when applying dynamic (DMQ) instead of static flow allocation (SMQ). It is therefore clear that dynamic flow allocation outperforms the other assembly schemes, weakens the synchronization effect and we may argue that also provides a notion of fairness, since performance variation between the individual TCP flows is diluted. In Table 2, we summarize the performance of all assembly schemes studied above, namely MF, SMQ, DMQ-2, DMQ-4, and DMQ-8. The performance of the DMQ scheme can be further improved upon selection of another allocation algorithm that takes into account TCP dynamics as for example retransmission timeout (RTO) or the instant flow window size. However, even the simple dynamic allocation of flows to different burst queues with equal probabilities significantly weakens TCP synchronization.

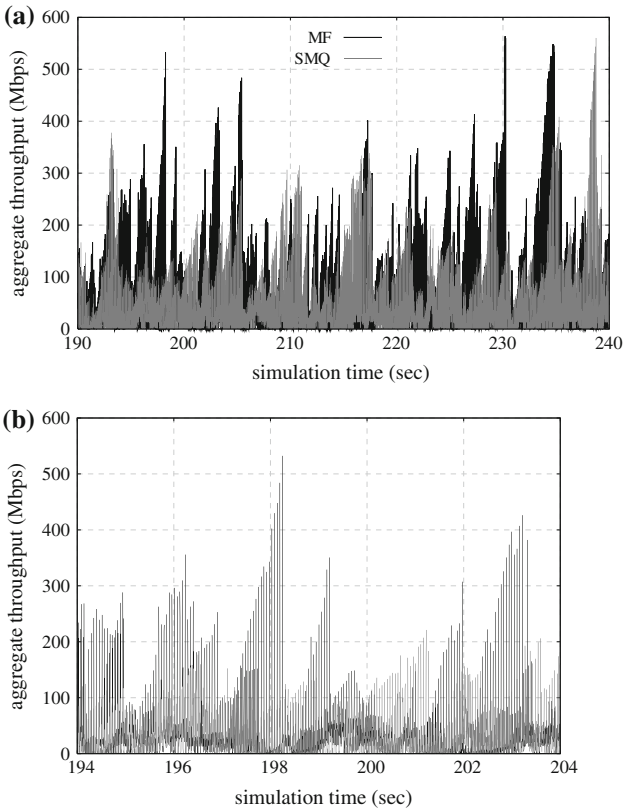

Fig. 15 **a** Aggregated throughput variation versus time of all sources of an edge router for a specific source–destination pair for the MF and SMQ assembly schemes. **b** Detailed illustration for 10 s span

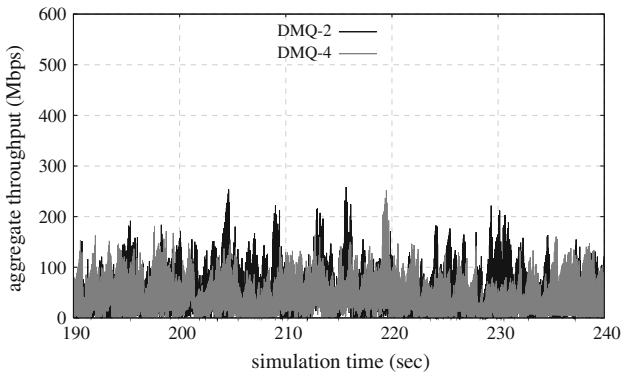

Fig. 16 Throughput variation versus time of all sources of an edge router for a specific source–destination pair for the DMQ-2 and DMQ-4 assembly cases

5 Conclusions

In this article, we have studied and analyzed the effect of TCP synchronization in OBS networks. TCP synchronization is the effect, when multiple TCP connections simultaneously increase or decrease their windows, causing a saw-tooth variation of outgoing traffic. This may result in a bad usage of available link capacity, which will not be able to accommodate the steep increases. In OBS networks, burst losses foster such an effect, which can yield an undesirable TCP performance. In this article, we have analyzed the TCP over OBS synchronization effect and studied the synchronization dynamics under different assembly scenarios and different network cases. It was shown that TCP synchronization exists even in large-scale networks with a high number of active flows. The number of flows increase does not weaken synchronization but it is only the burst loss ratio and flow-to-burst distribution that matters.

It was shown that a dynamic flow allocation to more than one assembly queue may provide a significant gain and reduce synchronization by more than 50%.

Acknowledgements The work described in this article was carried out with the support of the BONE-project ("Building the Future Optical Network in Europe"), a Network of Excellence funded by the European Commission through the 7th ICT-Framework Programme.

References

[1] Qiao, C., Yoo, M.: Optical burst switching (OBS)—a new paradigm for an optical internet. J. High Speed Netw. **8**(1), 69–84 (1999)
[2] Vokkarane, V., Jue, J.: Burst segmentation: an approach for reducing packet loss in optical burst switched networks. Opt. Netw. Mag. **4**(6), 81–89 (2003)
[3] Detti, A., Eramo, V., Listanti, M.: Optical burst switching with burst drop (OBS/BD): an easy OBS improvement. In: Proceedings of IEEE ICC 2002, New York, April–May, vol. 5, pp. 2687–2691 (2002)
[4] Cao, X., Li, J., Chen, Y., Qiao, C.: Assembling TCP/IP packets in optical burst switched networks. In: Proceedings of IEEE GLOBECOM 2002, Taipei, Taiwan, 17–21 November, vol. 3, pp. 2808–2812 (2002)
[5] Yu, X., Li, J., Cao, X., Chen, Y., Qiao, C.: Traffic statistics and performance evaluation in optical burst switched networks. J. Lightwave Technol. **22**(12), 2722–2738 (2004). doi:10.1109/JLT.2004.833527
[6] Detti, A., Listanti, M.: Impact of segments aggregation on TCP reno flows in optical burst switching networks. In: Proceedings of IEEE INFOCOM 2002, New York, June, vol. 3, pp. 1803–1812 (2002)
[7] González, O., de Miguel, I., López, V.: Performance evaluation of TCP over OBS considering background traffic. In: Proceedings of 10th Conference, ONDM 2006, Copenhagen, Denmark (2006)
[8] Zhou, J., Wu, J., Lin, J.: Improvement of TCP performance over optical burst switching networks. In Proceedings of the 11th International IFIP TC6 Conference, ONDM 2007, Athens, Greece, 29–31 May. Lecture Notes in Computer Science, vol. 4534, pp. 194–200. Springer-Verlag, Berlin (2007)
[9] Ramantas, K., Vlachos, K., González de Dios, O., Raffaelli, C.: TCP traffic analysis for timer-based burstifiers in OBS networks. Proceedings of the 11th International IFIP TC6 Conference, ONDM 2007, Athens, Greece, 29–31 May. Lecture Notes in Computer Science, vol. 4534, pp. 176–185. Springer-Verlag, Berlin (2007)
[10] Malik S., Killat U.: Impact of burst aggregation time on performance in optical burst switching networks. Opt. Switch. Netw. **2**, 230–238 (2006). doi:10.1016/j.osn.2006.01.002
[11] Vokkarane, V.M., Haridoss, K., Jue, J.P.: Threshold-based burst assembly policies for QoS support in optical burst-switched networks. In: Proceedings of the SPIE, OptiComm 2002, Boston, MA, 30–31 July, vol. 4874, pp. 125–136 (2002)
[12] Kantarci, B., Oktug, S.: Adaptive threshold based burst assembly in OBS network. In: Proceedings of the Canadian Conference on Electrical and Computer Engineering, CCECE'06, May, pp. 1419–1422 (2006).
[13] Fall, K., Floyd, S.: Simulation-based comparisons of Tahoe, Reno and SACK TCP. ACM SIGCOMM Comput. Commun. Rev. **26**(3), 5–21 (1996). doi:10.1145/235160.235162
[14] Yu, X., Qiao, C., Liu, Y.: TCP implementations and false time out detection in OBS networks. In: Proceedings of INFOCOM 2004, 7–11 March, vol. 2, pp. 774–784 (2004)
[15] Azodolmolky, S., Tzanakaki, A., Tomkos, I.: On the impact of burst assembly on self-similarity at the edge router in optical burstswitched networks. In: Proceedings of the 5th International Symposium, CSNDSP 2006, Patras, Greece, 19–21 July (2006)
[16] Raffaelli, C., Zaffoni, P.: TCP performance in optical packet-switched networks. Photon. Netw. Commun. **11**(3), 243–252 (2006). doi:10.1007/s11107-005-7351-7
[17] Guidotti, A.M., Raffaelli, C., González de Dios, O.: Effect of burst assembly on synchronization of TCP flows. In: Proceedings of the 4th International Conference on Broadband Communications, Networks, and Systems, WOBS/Broadnets 2007, Raleigh, US, 10–14 September, pp. 29–36 (2007)
[18] Qiu, L., Zhang, Y., Keshav, S.: Understanding the performance of many TCP flows. Comput. Netw. **37**(3–4), 277–306 (2001). doi:10.1016/S1389-1286(01)00203-1
[19] Baccelli, F., Hong, D.: Interaction of TCP flows as billiards. IEEE/ACM Trans. Netw. **13**(4), 841–853 (2005)
[20] Raina, G., Towsley, D., Wischik, D.: Part II: Control theory for buffer sizing. ACM/SIGCOMM Comput. Commun. Rev. **35**(3), 79–82 (2005)
[21] Han, H., Hollot, C.V., Towsley, D., Chait, Y.: Synchronization of TCP flow in networks with small droptail buffers. In: Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 12–15 December, pp. 6762–6767 (2005)
[22] Han, H., Hollot, C.V., Chait, Y., Misra, V.: TCP networks stabilized by buffer-based AQMs. In: Proceedings of IEEE INFOCOM 2004, 7–11 March, vol. 2, pp. 964–974 (2004)
[23] Bonald, T., May, M., Bolot, J.-C.: Analytic evaluation of RED performance. In: Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000), 26–30 March, vol. 3, pp. 1415–1424
[24] Casoni, M., Guidotti, A.M., Raffaelli, C.: Multiple TCP flow performance study over OBS networks. In: NOC 2007, Kista, Sweden (Invited paper) (2007)
[25] The Network Simulator ns-2. www.isi.edu/nsnam/ns.

Author Biographies

Oscar González de Dios was graduated with a Master Degree in Telecommunications Engineering (2000, University of Valladolid). In 2000, he joined Telefónica I+D, where he worked for 4 years in the development of telephonic applications and software testing. In 2005, he joined the Technology Strategy Department, where he has been participating in R&D European projects such as NOBEL II, E-photon One+, and AGAVE. He is currently working in the New Networking Technologies Department, where he is involved in IST projects, like BONE and CELTIC projects, RUBENS and BANITS 2, which he coordinates. Also, he has been involved in several internal innovation projects with the Telefónica group. In addition to his work in Telefónica, he is about to complete his Ph.D. in the topic of performance of TCP over OBS networks.

Anna Maria Guidotti has been a Ph.D. student at the University of Bologna. She received the M.S. degree in Telecommunications Engineering from the University of Bologna, Italy, in 2007. She has been working on optical networks, focusing her research on optical access and optical packet/burst switching. She was involved in the European project e-Photon/One.

Carla Raffaelli is an associate professor in Switching Systems and Telecommunication Networks at the University of Bologna. She received the M.S. and the Ph.D degrees in Electrical Engineering and Computer Science from the University of Bologna, Italy, in 1985 and 1990, respectively. Since 1985 she has been with the Department of Electronics, Computer Science and Systems of the University of Bologna, Italy. Her research interests include performance analysis of telecommunication networks, switching architectures, and protocols and broadband communication. Since 1993 she participated in European Union-funded projects on optical packet-switched networks, the RACE-ATMOS, the ACTS-KEOPS, the IST-DAVID, and e-photon/One projects. She is now active in the EU- funded BONE network of excellence. She also participated in many national research projects on telecommunication networks. She is the author of many technical papers on broadband switching and network modeling and acts as a reviewer for top international conferences and journals. She is author or co-author of more than 100 conference and journal papers mainly in the field of optical networking and performance evaluation.

Kostas Ramantas has received the Diploma of Computer Engineering from Computer Engineering and Informatics Department (CEID) of the University of Patras, Greece in 2006 and the M.Sc degree in 2008. He is currently working toward the Ph.D degree. Till now, he has been actively involved in E-Photon/One+ and BONE European projects. His research interests are in protocols, algorithms, and architectures in the area of optical communication networks.

Prof. Kyriakos Vlachos is a faculty member at the Computer Engineering and Informatics Department of University of Patras, Greece. He received his Dipl.-Ing. and Ph.D. in Electrical and Computer Engineering from the National University of Athens (NTUA), Greece, in 1998 and 2001, respectively. From 1997 to 2001 he was a senior research associate in the Photonics Communications Research Laboratory, NTUA, while in April 2001 he joined Bell Laboratories, Lucent Technologies, working on behalf of the Applied Photonics Group. Prof. Vlachos conducted research on high-speed optical networks and DWDM transmission techniques. Since 2003, he was also a member of Computer Engineering Laboratory of Technical University of Delft and scientific advisor of the National Regulation Authority of Telecommunication and Postal Service of Greece (EETT). In 2005, he became a faculty member of Computer Engineering and Informatics Department of University of Patras. His research interests are in the areas of high-speed protocols and technologies for broadband optical networks, optical packet/burst switching, and WDM transport systems. Prof. Vlachos has participated in various research projects funded by the European Commission (IST-STOLAS, IST-PRO3, ESPRIT-DOALL, e-photon/ONe+, IST-PHOSPHROUS, ICT-BONE, and ICT-DICONET). Prof. Vlachos is a member of IEEE and the Technical Chamber of Greece. He periodically acts as a scientific reviewer for the General Secretariat for Research and Technology of Greece (GSRT) as well as for the European Commission and the Netherlander Organization for Scientific Research, Technology Foundation. Prof. Vlachos is the (co)author of more than 80 journal and conference publications and holds five patents.