TCP Performance in Hybrid Multigranular OBS Networks

Maurizio Casoni[1] and Carla Raffaelli[2]
[1] Dept. of Information Engineering - University of Modena and Reggio Emilia
Via Vignolese 905, Modena, Italy
[2] D.E.I.S. - University of Bologna - Viale Risorgimento 2, Bologna, Italy
email: *maurizio.casoni@unimore.it*, *carla.raffaelli@unibo.it*

***Abstract*—This paper studies TCP performance with different offset times which arise when hybrid multi-granular technology is employed in OBS core nodes. Multi-granular switches are promising devices to match dynamic application needs while optimizing switch costs and feasibility. They foster optical burst switching in multi-service contexts where switching matrices are designed and optimized in application awareness. OBS base offset time is strictly dependent on switching matrix set up time which, in presence of hybrid technology, is different from path to path. These differences impact on end-to-end performance and in particular on TCP throughput.**

This work studies end-to-end performance for two basic classes of connections, namely slow and fast connections. Simulations results based on a careful set up of ns-2 simulations according to test bed configurations described in literature are provided to show the effects of the employment of different optical technologies on the TCP throughput.

I. Introduction

Dynamic optical networking offers flexible switching solutions to multi-service application contexts [1]. Optical Burst Switching (OBS) is one of the most flexible and fast networking techniques which can provide short term feasible migration to dynamic application support. OBS networks are designed according to optical core switches' configuration times, by properly choosing the offset times to perform switch resource reservations in advance, prior to OBS payload transmission [2], [3]. OBS payloads are prepared at the ingress edge nodes by aggregating information coming from legacy networks. Different assembly algorithms have been proposed in literature and their delay and throughput performance has been studied at the network level [4], [5]. Several works also considered the influence of OBS networks on Transmission Control Protocol (TCP) performance with different assembly techniques and algorithms [6], [7]. As a matter of facts, TCP is the dominant transport protocol in today networks and it is foreseen to be widely used also in the future.

To further enhance OBS flexibility hybrid multi-granular core nodes can be employed. Examples of this kind of switching solution are described in literature [8], [9] and testbeds have been set up to prove the feasibility of this concept [10]. The main idea behind multi-granular hybrid switching is to support multi-granular application requirements with different technologies which are characterized by different configuration times. The main target of this design concepts is to achieve short term feasibility while maintaining low switch costs, as discussed in many contexts [11]. These switches are said hybrid because they employ different technologies like MEMS and SOAs to implement slow and fast switching paths, respectively.

As a consequence of hybrid switch solutions, paths with different set up delays are available in the OBS network. One of the research activities in this field is aimed at optimizing shared usage of these paths [12]. As far as the OBS network design, hybrid solutions impact on the values of the base offset that takes into account the time needed to set up a path through the core switches. Longer offset times are needed in the slow path case than for fast paths. Different offset times differently influence upper layer performance and these effects are worth to be evaluated.

In this paper, TCP performance is considered when different offset times are applied to slow and fast paths. In particular TCP congestion control reaction to burst losses on paths with different delay characteristics is analyzed to evaluate the acceptable burst loss rate to obtain acceptable throughput. The evaluations have been obtained by simulation with set up related to a reference experimental test-bed [13].

The rest of the paper is organized as follows. Section II provides the description of the edge node and Section III the description of the core node. Performance evaluation is presented and discussed in Section IV while the concluding remarks are in Section V.

II. Edge Node Architecture

In OBS networks data bursts never leave the optical domain. Each optical burst is assembled at the network edge and a reservation request is sent in advance as a separate control packet. The main function of edge nodes which

Digital Object Identifier: 10.4108/ICST.BROADNETS2009.8020
http://dx.doi.org/10.4108/ICST.BROADNETS2009.8020

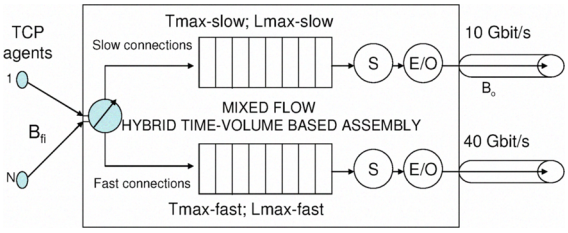

Fig. 1. The OBS edge node.

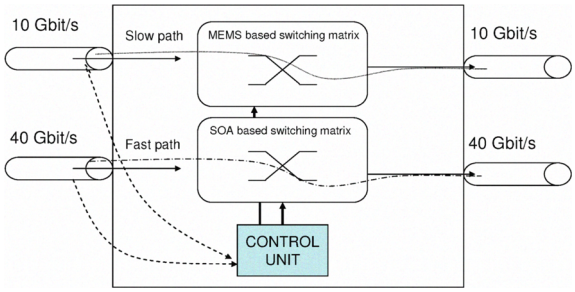

Fig. 2. The OBS core node.

impact on end to end performance is the burst assembly: IP datagrams are collected and assembled into bursts according to proper assembly algorithms. The control packet carries relevant forwarding informations, as the next hop, the burst length and the offset time. It precedes the data burst by a basic offset time, that is set to accommodate the non-zero electronic processing time inside the network, and it dinamically set up a wavelength path whenever large data flows are identified and need to traverse the network.

The edge node architecture investigated in this paper is reported in Figure 1. It can be logically divided into three main blocks. The first one encompasses the input interface cards and performs datagram classification as belonging to fast or slow connections to forward them to the proper assembly unit. The second one is the burst assembly unit which collects datagrams to form optical bursts. The third block is given by the output interfaces, equipped with a scheduler, and E/O conversion functionalities. Several assembly algorithms have been proposed and evaluated; in this paper a hybrid time-volume based assembly algorithm is employed. Since the edge-to-edge delay has to be bounded, a maximum delay T_{max} can be tolerated in the assembly phase, but after that the burst must be transmitted. Or if a maximum burst length L_{max} is reached before the timer, set to T_{max}, expiration the burst is scheduled for transmission.

Mixed flow assembly is applied, meaning that a burst may contain information from different flows of the same class. Per quality-of-service queuing is needed at the ingress of the assembly unit. The mixed flow solution leads to a simpler implementation of the assembly unit with respect to the per-flow solution.

The employment of different optical technologies in core nodes implies different set-up times for the corresponding switching matrices. This means that different offset times have also to be considered, depending on the kind of connection datagrams belong to. We assume as *fast connections* all connections whose datagrams are put in bursts which are switched in core nodes employing the fastest optical technologies, which means shortest set-up times, and as *slow connections* when the related bursts are switched employing slower optical technologies, which leads to longer set-up times.

Therefore, it is necessary that the hybrid time-volume based assembly algorithm behaves differently for fast and slow connections. It is then assumed that for slow connections $T_{max} = T_{max-slow} > T_{max} = T_{max-fast}$, used for fast connections. Also, it is assumed that slow connections carry long and medium size bursts whereas fast connections are used for short bursts. To this end, $L_{max} = L_{max-slow} > L_{max} = L_{max-fast}$

III. Core Node Architecture

In OBS networks, core nodes deal with optical data bursts and the related control packets. They have to set up on the fly internal optical paths to switch bursts and take them hop-by-hop closer to their final destination. In addition, the offset time allows the core router to be buffer-less, thus avoiding the employment of optical memories, e.g. fiber delay lines, required in optical packet switches. As mentioned above, the control packet carries relevant forwarding informations, as the next hop, the burst length and the offset time. The reference core node architecture here investigated is depicted in Figure 2. It consists of a control unit, which processes burst control packets and two sub-systems, the MEMS based switching matrix and the SOA based switching matrix, which support slow and fast paths, respectively. Optical MEMS are reliable, flexible devices at high integration and relatively low cost, which are characterized by set-up times in the range of several milliseconds. SOA gates, on the contrary, typically have lower integration properties, higher costs but also remarkably higher performance, with set-up times in the range between hundreds of nanoseconds and several microseconds.

In addition, core nodes are equipped with input and output interfaces which can operate at two different transmission rates, 10 Gbit/s and 40 Gibit/s to serve slow and fast connections, respectively.

IV. Performance evaluation

Simulation of end to end TCP connections have been obtained by means of simulations using the *ns-2 simulator* vers. 2.31 [14], which is an object-oriented tool widely

Digital Object Identifier: 10.4108/ICST.BROADNETS2009.8020
http://dx.doi.org/10.4108/ICST.BROADNETS2009.8020

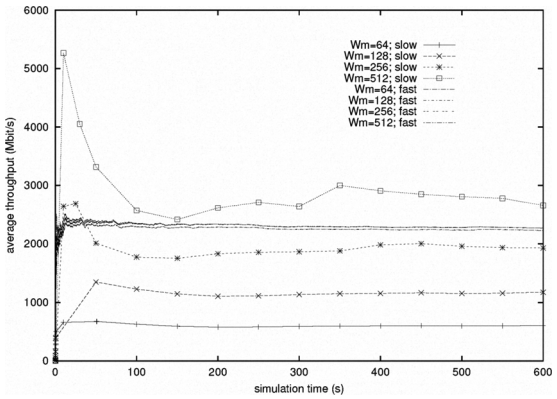

Fig. 3. Aggregate average throughput as a function of time, for $P_l = 10^{-3}$

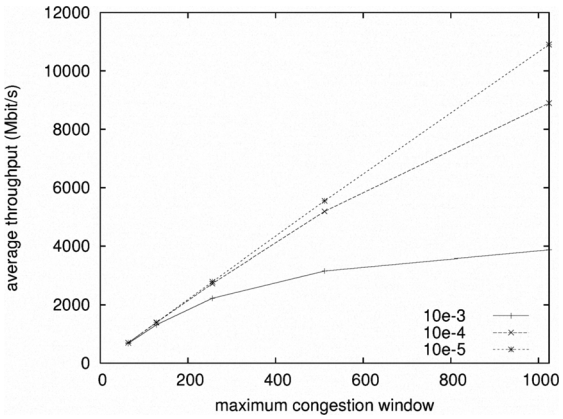

Fig. 4. Aggregate average throughput calculation as a function of W_m varying the burst loss probability, for slow connections

adopted to evaluate network performance. The numerical results presented in this section are reported for a OBS network scenario given by ingress and egress edge nodes and one core node, as in Figures 1 and 2.

As it is widely known, several TCP flavors have been proposed in literature [15], [16], [17], [18]. One of the most known and used is the TCP New Reno with the SACK option and this is the flavor here employed. TCP Sack includes a strategy which modifies the TCP behavior in case of multiple dropped segments. Using the SACK option the receiver informs the sender about the segments that are successfully received so that the sender needs to retransmit lost segments only. Users are represented by N TCP SACK agents connected to the edge node. CBR is the application assumed over the TCP agents on the customer side.

Optical bursts are formed according to mixed flow aggregation by applying hybrid time-volume assembly, as explained above.

TCP performance is studied by evaluating throughput, which is a measure of the variability of the bandwidth usage over a given time scale. The *average throughput* is the amount of successfully transmitted bytes in a given time interval, e.g. since the beginning up to t. The *aggregated average throughput* is the average throughput computed as average over all active TCP flows. All TCP flows are provided with the same access bandwidth $B_{fi} = 100$ Mbit/s. Results will be presented and discussed in the following considering two different simulation parameter settings, for slow and fast connections. This simulation scenario assumes most values according to the test-bed presented in [13].

A. TCP over slow connections

In this scenario the optical transmission rate B_o is equal to 10 Gbit/s and the number of TCP sources N is 80. As regards the edge node, the hybrid time-volume assembly algorithm is set with $T_{max-slow} = 10$ ms and $L_{max-slow} = 10$ MB. These values allow to generate bursts in the $[1 - 10]$ MB range. As far as the propagation delay, it is assumed to introduce a delay of 5 ms, representative of the propagation delay over a distance of 1000 km. Since core nodes in OBS are buffer-less, no queuing delays are considered. Furthermore, slow connections in core nodes are managed through a switching matrix based on MEMS, whose set-up times are in the range of several milliseconds. Therefore an additional 10 ms has to be considered as offset time value between the burst control packet and the burst. The TCP Retransmission Time Out (RTO) is then set to 30 ms.

Figure 3 reports the aggregate average throughput as a function of time for different values of the maximum congestion window, (W_m), and for slow and fast connections. The burst loss probability is $P_l = 10^{-3}$ and the TCP maximum segment size MSS is 512 bytes. As regards slow connections, the best performance is obtained for $W_m = 512$ segments. It is worthwhile noting that the ideal throughput for one single flow is given by $\frac{MSS \times W_m}{RTT}$ so that the ideal aggregate throughput is $\frac{MSS \times W_m}{RTT} \times N \simeq 5.592$ Gbit/s. Looking again at figure 3, this is the initial value of the curve simulated with $W_m = 512$, before burst loss occurs. The average length of the generated bursts results in this case equal to 6.32 MB, which is less of the maximum value allowed in generated bursts (10 MB). To study the aggregate throughput behavior for slow connections, as a function of W_m, the analytical formulas obtained in [19] are applied and results are reported in figure 4. Except for a slight throughput overestimation, which was discussed in that paper, the figure well captures the average behavior, by showing the effect of congestion window ($cwnd$) limitation when the burst loss is low, e.g. $P_l = 10^{-5}$. On the other side, when burst loss increases the role of $cwnd$ gets less

Digital Object Identifier: 10.4108/ICST.BROADNETS2009.8020
http://dx.doi.org/10.4108/ICST.BROADNETS2009.8020

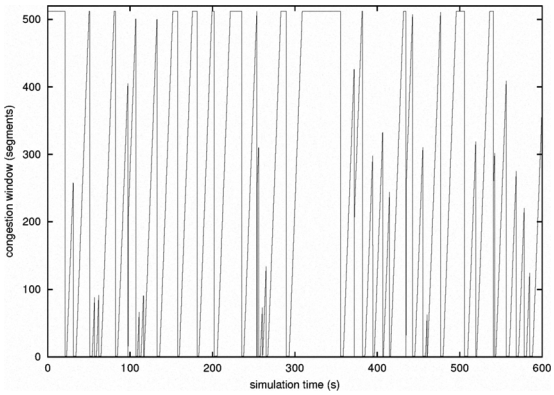

Fig. 5. Congestion window as a function of time for one TCP source over a slow connection, for $W_m = 512$ and $P_l = 10^{-3}$

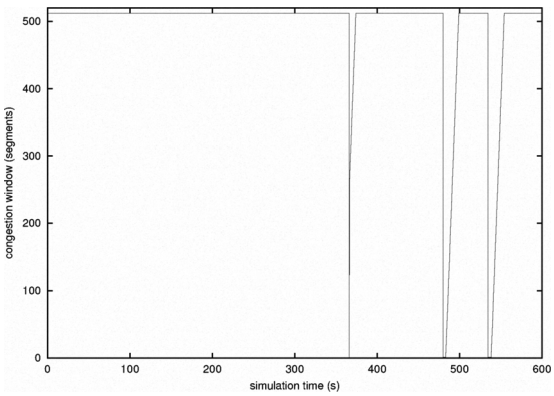

Fig. 6. Congestion window as a function of time for one TCP source over a slow connection, for $W_m = 512$ and $P_l = 10^{-4}$

important, e.g. $P_l = 10^{-3}$.

Figures 5, 6 and 7 show $cwnd$ as a function of time for $P_l = 10^{-3}$, $P_l = 10^{-4}$ and $P_l = 10^{-2}$, respectively. Losses are here always detected by means of RTO expiration, which forces $cwnd$ to drop to 1 MSS. This is explained by considering that,e.g., for $W_m = 512$ and $P_l = 10^{-3}$, the average number of TCP segments per burst is equal to 12360, which means that, being $N = 80$, each TCP flow on average looses roughly 154 segments for each burst loss. Thus, the burst loss event leads immediately to slow start. At the same time, as a consequence of the high correlation benefit, congestion window inflate is very rapid [5]. The maximum value W_m, i.e. 512, can be anyway reached, especially with low burst loss probability as shown in figure 6 for $P_l = 10^{-4}$. When burst loss is higher, e.g. $P_l = 10^{-2}$, figure 7 spike behavior is not enough to completely open the congestion window to its maximum value.

B. TCP over fast connections

In this scenario the optical transmission rate B_o is equal to 40 Gbit/s and the number of TCP sources N is 320. As regards the edge node, the hybrid time-volume assembly algorithm is set with $T_{max-slow} = 0.5$ ms and $L_{max-slow} = 25$ KB. These values allow to generate bursts in the $[10-25]$ KB range. Here, propagation delay of 5 ms is assumed and, again, no queueing delays are considered. Fast connections in core nodes are managed through a switching matrix based on SOA, whose set-up times are in the range of hundreds of nanoseconds to several microseconds. The TCP Retransmission Time Out (RTO) is here set to 10.5 ms.

Let us consider again figure 3 as far as the curves for fast connections. The aggregate average throughput basically does not depend on W_m anymore and it rapidly converges to a value close to 2.2 Gbit/s. This means that there is no gain by increasing the value of the maximum W_m. This is explained by means of figures 8, 9 and 10. First, most losses are now detected by means of triple duplicate acks (TD) rather than RTO expiration; this is due to the low value of the average number of TCP segments per burst which is equal to 44, for $W_m = 512$ and $P_l = 10^{-3}$. This means that typically each source has no more than one segment in the lost burst, which is recovered by the fast retransmit/fast recovery procedure. However, due to the very high number of short bursts generated in this configuration, burst losses are so frequent that W_m cannot completely open, i.e. it never reaches its maximum value. This is basically true for $W_m = 128$, as well (figure 9). Therefore, it seems that $W_m = 64$ is enough to optimize connection performance and figure 10 confirms this. By increasing burst loss rate, e. g. $P_l = 10^{-2}$, even $W_m = 64$ cannot be fully exploited, as shown in figure 11. In this case in fact the aggregate average throughput drops to 298 Mbit/s. On the other, if burst loss is low, e. g., $P_l = 10^{-4}$, $W_m = 128$ can be used in effective way, as shown in figure 12, which exhibits losses detected by the fast recovery/fast retransmit mechanism. The aggregate average throughput jumps to 9.2 Gbit/s.

In figure 13 the average aggregate throughput is reported for fast connections with $W_m = 64$ and slow connections with $W_m = 512$ as a function of P_l. When loss is very low fast connections give higher throughput, as expected, with respect to the slow configuration. Anyway fast connection throughput is highly sensitive to losses and it sensibly drops even for medium range burst loss rates, e.g. $P_l = 10^{-3}$.

V. Conclusions

Performance evaluation of TCP in hybrid multi-granular OBS scenario is presented. The main issue addressed is related to the availability of fast and slow paths to support

Digital Object Identifier: 10.4108/ICST.BROADNETS2009.8020
http://dx.doi.org/10.4108/ICST.BROADNETS2009.8020

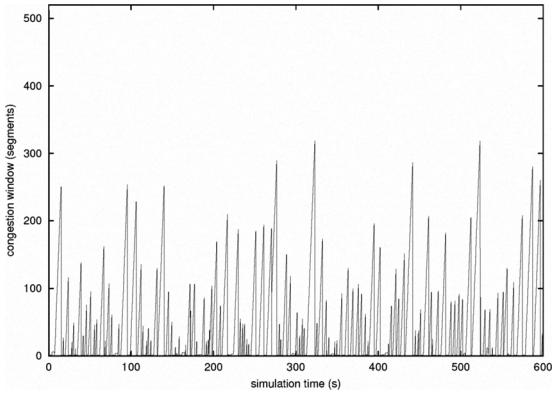

Fig. 7. Congestion window as a function of time for one TCP source over a slow connection, for $W_m = 512$ and $P_l = 10^{-2}$

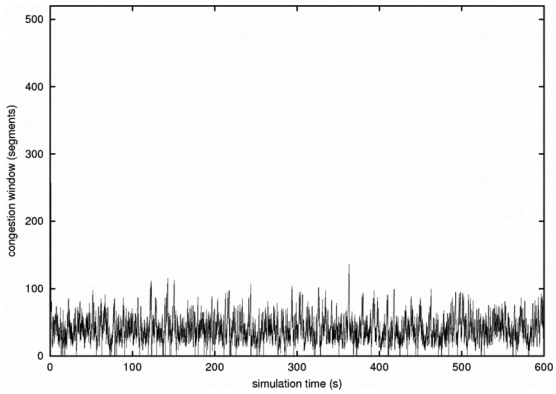

Fig. 8. Congestion window as a function of time for one TCP source over a fast connection, for $W_m = 512$ and $P_l = 10^{-3}$

multi-granular connections in optical networks. Different switching configuration times are considered, which impact on the offset time choice and on TCP congestion control mechanism. Results are obtained by setting up simulations according to experiments proposed in literature. A first interesting conclusion is that burst loss probability affects TCP throughput and the ideal maximum congestion window value in a different way, for fast and slow connections. For low burst losses, fast connections can provide remarkable higher throughputs than slow connections, whereas for burst losses greater or equal to 0.001 the average throughput is roughly the same, despite of different transmission rates. As regards the maximum value of the congestion window, for fast connections there is not any gain to increase it above 64 segments while the best for slow connections is 512. This is due to the different loss detection mechanisms which are applied in these two cases. For fast connections TD is enough to detect and recover losses while for slow connections burst losses lead to RTO expiration. Therefore, in fast connections TCP hardly can completely open the window because it experiences a new loss before reaching the maximum window value and it has to fast recover it; on the other hand, in slow connections the correlation benefit allows the window to grow faster and to reach the maximum value but, as soon as a burst loss happens, all involved TCP flows have the window closing and have to resume from slow-start.

ACKNOWLEDGMENT

The work described in this paper 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.

The authors wish to thank Dr. Michele Savi for the useful discussions on multi-granular architectures.

REFERENCES

[1] O'Mahony, C. Politi, D. Klonidis, R. Nejabati, D. Simeonidou, "Future Optical Networks", *IEEE Journal of Lightwave Technology*, vol. 24, no. 12, pp. 4684 - 4696, Dec. 2006.

[2] C. Qiao, M. Yoo, "Optical Burst Switching (OBS) - a New Paradigm for an Optical Internet", *Journal of High Speed Networks*, No.8, pp.69-84, 1999.

[3] Y. Chen, C. Qiao, X. Yu, "Optical Burst Switching: A New Area in Optical Networking Research", *IEEE Network*, May/June 2004, pp. 16-23.

[4] M. Casoni, E. Luppi, M.L. Merani, "Impact of Assembly Algorithms on End-to-End Performance in Optical Burst Switched Networks with Different QoS Classes", Proc. of 3rd International Workshop on Optical Burst Switching, October 2004, San Jose, CA, USA.

[5] A. Detti and M. Listanti, "Impact of Segments Aggregation on TCP Reno Flows in Optical Burst Switching Networks", Proc. of IEEE INFOCOM 2002, vol. 3, pp. 1803-1812, June 2002.

[6] K.Ramantas, K.Vlachos, O. Gonzalez de Dios, C. Raffaelli, "Window-based burst assembly scheme for TCP traffic over OBS", *OSA Journal of Optical Networking*, Vol. 7, Issue 5, pp. 487-495, 2008, U.S.A.

[7] A. M. Guidotti, C. Raffaelli, O. Gonzalez de Dios, "Effect of burst assembly on synchronization of TCP flows", Proc. of International Workshop on Optical Burst Switching 2007, Raleygh, USA, September 2007.

[8] G. Zervas, M. De Leenheer, L. Sadeghioon, D. Klonidis, R. Nejabati, D. Simeonidou, C. Develder, B. Dhoedt, P. Demeester, "Multi-granular Optical Cross-Connect: Design, Analysis and Demonstration", *OSA/IEEE Journal of Optical Communications and Networking*, Vol. 1, Issue 1, pp. 69-84, June 2009.

[9] Y.Y. Wang, Q. Zeng, L. Lu,Z. S., "Smart design of multigranular optical switching fabric with traffic grooming", Optical Engineering, SPIE, Vol. 45, 030505 (2006).

[10] G. Zervas,L. Sadeghioon, D. Klonidis, R. Nejabati, D. Simeonidou, "Demonstration of Novel Multi-Granular Switch Architecture on an Application-Aware End-to-End Multi-Bit Rate OBS Network Testbed", Proc. of ECOC 07, September 16-20, 2007 Berlin, Germany.

[11] A. Stavdas, C. (T) Politi, T. Orphanoudakis, A. Drakos, "'Optical packet routers: how they can efficiently and cost-effectively scale to petabits per seconds"', *OSA Journal of Optical Networking*, Vol. 7, N. 10, October 2008.

[12] M. Savi, G. Zervas, Y. Qin, V. Martini, C. Raffaelli, F. Baroncelli, B. Martini, P. Castoldi, R. Nejabati, D. Simeonidou, "Data-Plane Architectures for Multi-Granular OBS Network", Proc. of OFC 2009, San Diego, USA, March 2009.

Digital Object Identifier: 10.4108/ICST.BROADNETS2009.8020
http://dx.doi.org/10.4108/ICST.BROADNETS2009.8020

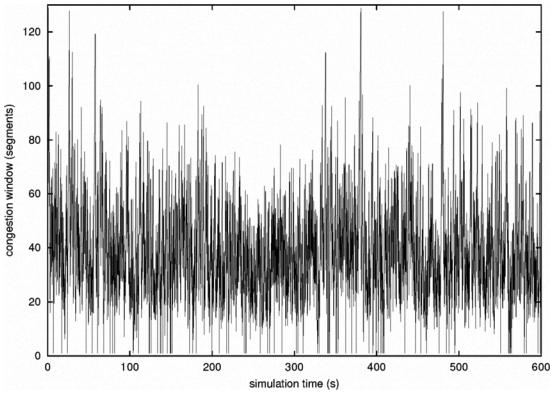

Fig. 9. Congestion window as a function of time for one TCP source over a fast connection, for $W_m = 128$ and $P_l = 10^{-3}$

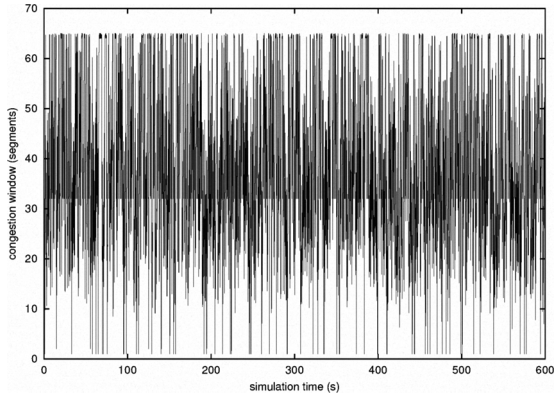

Fig. 10. Congestion window as a function of time for one TCP source over a fast connection, for $W_m = 64$ and $P_l = 10^{-3}$

[13] G. Zervas, Y. Qin, R. Nejabati, D. Simeonidou, "Service oriented optical burst switched edge and core routers for future internet", Proc. of Broadnets 2008, pp. 97-104, 8-11 September 2008, London (U.K.).

[14] NS-2 Network Simulator (Ver.2) http://www.mash.cs.berkeley.edu/ns/.

[15] S. Floyd, J.Mahdavi, M. Mathis, M. Podolsky, "An Extension to the Selective Acknowledgement (SACK) Option for TCP", RFC 2883, IETF, July 2000.

[16] C. Casetti, M. Gerla, S. Mascolo M. Y. Sanadidi, and R. Wang.,"TCP Westwood: Bandwidth estimation for enhanced transport over wireless links", Wireless Networks 2002.

[17] L. Bramko, S. O'Malley and L. Peterson, "TCP Vegas:new tecniques for congestion detection and avoidance", Proceedings of the ACM SIGCOMM, pp. 24-35, August 1994.

[18] Jin Cheng, D. Wei, S.H. Low, J. Bunn, H.D. Choe, J.C. Doylle, H. Newman, S. Ravot, S. Singh, F. Paganini, G. Buhrmaster, L. Cottrell, O. Martin, Wu-chun Feng, 'FAST TCP: from theory to experiments', *IEEE Network*, Vol. 19, Issue 1, pp.4-11, Jan.-Feb. 2005.

[19] M. Casoni, C. Raffaelli, "Analytical Framework for End-to-End Design of Optical Burst-Switched Networks", *Optical Switching and Networking*, Elsevier, Vol.4, Issue 1, pp.33-43, February 2007.

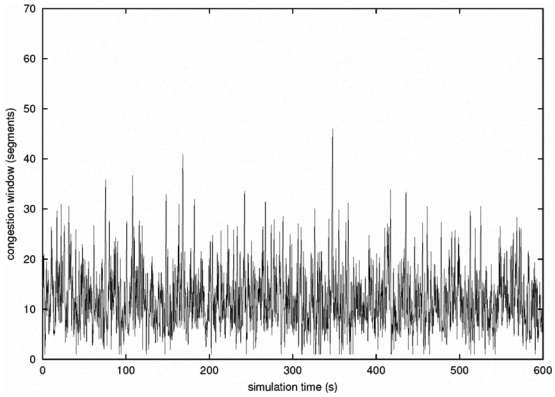

Fig. 11. Congestion window as a function of time for one TCP source over a fast connection, for $W_m = 64$ and $P_l = 10^{-2}$

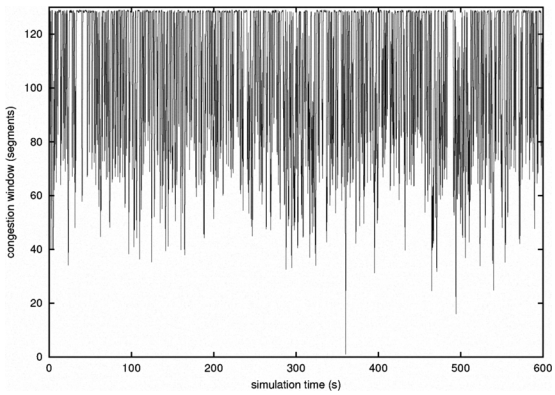

Fig. 12. Congestion window as a function of time for one TCP source over a fast connection, for $W_m = 128$ and $P_l = 10^{-4}$

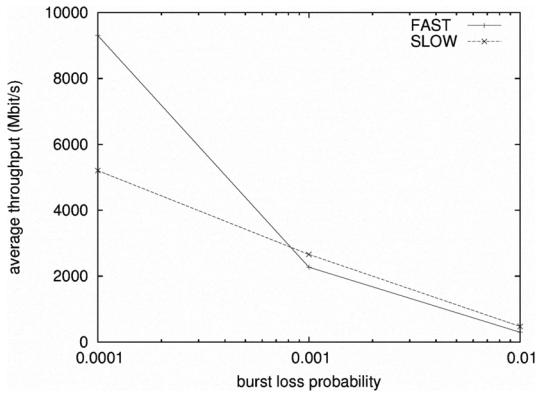

Fig. 13. Average aggregate throughput comparison for fast connections ($W_m = 64$) and slow connections ($W_m = 512$) as a function of burst loss probability

Digital Object Identifier: 10.4108/ICST.BROADNETS2009.8020
http://dx.doi.org/10.4108/ICST.BROADNETS2009.8020