TCP performance in an optical link applying lightpath bundling and anycast switching techniques

Pablo Pavon-Marino, Jose-Luis Izquierdo-Zaragoza
Universidad Politécnica de Cartagena, Cuartel de Antiguones, Plaza del Hospital 1, 30202 Cartagena, Spain
Tel: (+34) 968325952, Fax: (+34) 968325973, e-mail: {pablo.pavon, josel.izquierdo}@upct.es

ABSTRACT

Lightpath bundling (LB) technique consists of grouping a set of lightpaths between two nodes so that they appear to the IP layer as a single pipe of aggregated capacity. Anycast switching (AS) technique makes a per-packet granularity balancing of the traffic among the lightpaths bundled. LB+AS combination requires seamless changes in the switching nodes and no changes in the optical infrastructure. This paper evaluates by means of simulation the improvements in TCP performances when LB+AS paradigm is applied to the network. In our results, LB+AS improved TCP throughput (up to 18%), decreased average drop rates and reduced the average queuing delay in the bottleneck link (up to 15%).

Keywords: multilayer optical networks, lightpath routing, high-performance packet switching.

1. INTRODUCTION

Optical multilayer IP over WDM networks are an enabling technology to meet the ever-increasing bandwidth requirements that the Internet is facing today. An IP over WDM multilayer network (or wavelength-routed networks) provides transparent end-to-end optical channels (lightpaths) between routers, eliminating extra electronic processing at intermediate nodes along the physical path. Thus, each lightpath corresponds to a virtual link between two routers: independently on the actual sequence of fibers traversed by the lightpath, the IP layer sees it as pipe to transmit packets.

The success of multilayer IP over WDM networks is built on the winning combination of IP/MPLS electronic packet switching for a flexible and finer traffic granularity, and optical circuit switching for a coarse granularity, by means of ROADMs (Reconfigurable Optical Add-Drop Multiplexers).

The traffic engineers at the IP layer face a trade-off between high utilization of the lightpaths and low packet delay. In particular, due to the traffic burstiness in real networks, a high lightpath utilization may result in an unacceptable degradation of the packet delay, which impacts on new delay-sensitive applications such as live streaming video or voice over IP. A recent work [1] suggests that the combined application of a control plane lightpath aggregation technique, so-called *lightpath bundling* (LB), along with a data plane per-packet traffic balance scheme, known as *anycast switching* (AS), can positively bias the utilization versus delay trade-off. The lightpath bundling (LB) paradigm is a control plane artifact, consisting of bundling together in the network those lightpaths which have a common input and output node, so that they appear to the electronic layer (IP, Ethernet or MPLS) as a single link of aggregated capacity. Lightpath bundling can be implemented in a network based on a GMPLS control plane, by making use of the GMPLS link bundling functionality as described in RFCs 4201 and 4202, being necessary to define a lightpath bundle as TE-link (Traffic Engineering link) aggregating the lightpaths, which are themselves seen as other TE-links. Anycast switching is a technique proposed in [2] to be implemented in the data plane of high-capacity electronic packet switches (e.g. IP), in conjunction with LB. Anycast switching means that the electronic switches are instructed to have a new degree of freedom to choose the output lightpath of a packet among all the lightpaths in the bundle. Electronic switches make use of this degree of freedom to implement a per-packet granularity traffic balance among the lightpaths bundled, which enhances the statistical multiplexing gain. After that, the rest of the data path remains unmodified.

The per-packet (finer) granularity of the balance, makes AS different to other traffic balancing alternatives (e.g. 802.1AX), since AS does not require wire-speed flow identification. However, precisely for this reason, out-of-sequence packet delivery may occur. Still, the order between packets within a higher layer connection traversing several bundles can be kept if: (i) the switching equipment is able to emulate output buffering, (ii) a JSQ (Join the Shortest Queue) policy is applied [1].

Simulation results in [1] shows that under *inelastic* self-similar traffic (e.g. caused by the aggregation of multimedia flows), the application of LB+AS techniques may permit, setting an end-to-end average queuing network delay target, up to 65% CAPEX/OPEX reduction (less lightpaths to carry the same traffic, assuming the number of lightpaths as the main contribution to the cost and energy consumption of the whole network) or up to 50% revenue increase (more carried traffic using the same number of lightpaths), depending on the ISP criteria.

In this work, our objective is to study how the application of LB+AS paradigm may affect to TCP flows in an optical link. We choose TCP since it is still the dominant transport protocol on the Internet. TCP traffic is *elastic* in the sense that adapts its rate to the current network state, and reacts to packet losses reducing the rate. We perform a worst-case simulation in which only long-lived TCP flows (‘elephants’ in common TCP terminology) are present in the optical link. Simulation results show a benefit in terms of a reduced queue occupancy in the IP routers and higher utilization of the bottleneck link, when LB+AS paradigm is applied.

The rest of the paper is structured as follows. In Section 2, we present our network model and assumptions. Section 3 provides the simulation results from the performance evaluation of our scheme. Finally, Section 4 concludes the paper.

2. SIMULATION SCENARIO

Fig. 1 illustrates our simulation setup. We have a three-layered hierarchical model: access layer, metro layer and core layer. Flows from TCP sources go through individual access links, and are multiplexed onto metro links before reaching the input ports of the core nodes, and finally go through backbone links to their destination on the other side of the network. The access layer consists of a set of sub-networks containing *N* TCP sources and *N* TCP sinks each one, modeling end-hosts. The core layer consists of two nodes connected by a bundle of b_D (bundling degree) lightpaths, and acts as the bottleneck of the network; the buffer size per output port follows the conservative small buffer rule $Q_{max} = C \cdot RTT / 10$ [3], where *C* is the bottleneck link capacity and *RTT* is the round-trip time. The metro layer consists of a set of multiplexing nodes, serving as a gateway between the access layer and the core layer; large buffers are used at the multiplexing nodes to prevent drops at these points. All links are assumed to be bidirectional and run at 10 Gbps. The propagation delay between each source-sink pair is random, uniformly picked from the interval 80-120 ms (with an average of 100 ms); these small variations in *RTT* are sufficient to prevent synchronization [4], which is a necessary condition to apply the small buffer rule. For the sake of simplicity, each node is assumed to implement a VOQ arbitration able to emulate output buffering behavior, and Drop-Tail buffer management scheme. When LB+AS paradigm is applied, AS is implemented following a JSQ policy which prevents packet out-of-sequence to occur.

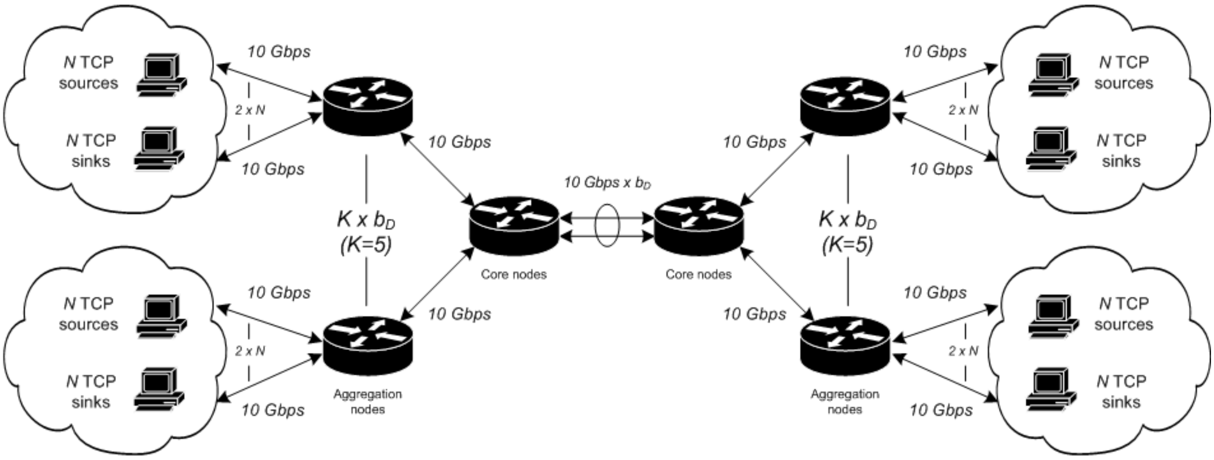

Figure 1. Simulated network topology

Each source in one side (left or right) opens a TCP connection to a sink in the other side. The data in each TCP connection goes from the source to the sink, and acknowledgements follow the opposite direction. Since we have sources and sinks in each of the $K \times b_D$ subnetworks, TCP segments with data and acknowledgements share the links. We emphasize this point, since in real networks data packets and acknowledgments occupy the same network, while in many simplified TCP studies this is not actually considered.

TCP sources transmit as much data as possible. The rate is controlled by the congestion control algorithm used by TCP NewReno, using the window scaling option [5] to set a large enough advertised receiver window to accept all data generated. Data packets have a maximum segment size of 1452 bytes, and a MTU of 1500 bytes. TCP connections generate congestion over the bottleneck link in both directions. To ensure that, $K \times b_D$ (K=5) aggregation routers are offering traffic through 10 Gbps links to inter-core router links that sum $b_D \times 10$ Gbps of capacity. We used K=5, but any value K>1 is enough for making the lightpaths between the core routers become the bottleneck.

We order the sources as follows: source 1 to *N* are the ones in the upper subnetwork in Fig. 1, sources *N*+1 to 2*N* are the ones in the subnetwork below it, etc. Same ordering applies to the sinks and sources in both left and right sides. Then, first source in the left side opens a connection with a sink in the first subnetwork of the other side, next source with a sink in the second subnetwork and so on. TCP connections from sources in the right side to sinks in the left side are arranged in the same manner. Since in our tests $N \geq K \times b_D$, it happens that every subnetwork has traffic targeted to every other subnetwork, and traffic is symmetric.

Concerning to the routing at the core, when LB+AS is not applied, the core nodes see each lightpath in the bundle as an individual link. The routers take a routing decision according to the destination subnetwork i=1,…, $K \times b_D$ of the packet, so that the traffic targeted to the *i*-th subnetwork is routed through the *j*-th lightpath, where

j=1+((i-1) modulus b_D). Then, each lightpath in the bundle transmits packets coming from every subnetwork. Also, each lightpath carried traffic targeted to K different subnetworks. In average $K \times N$ TCP flows share (and compete for) the capacity in each lightpath. When LB+AS is applied, the routers see the b_D lightpaths as a single link and, in fact, the core routers can aggregate all the IP routes in one. The traffic is spread in the lightpaths according to the per-packet granularity balance enforced by the JSQ policy. Then, $N \times K \times b_D$ TCP connections compete for the capacity of b_D lightpaths.

3. RESULTS

In order to evaluate our LB+AS approach for TCP traffic, we have used the OMNeT++-based INET framework [6]. The IP routing model in INET has been extended to support anycast switching as described in Section 1. The performance metrics considered are: (i) average queue length in the core routers, (ii) average drop rate at the core routers, and (iii) average bottleneck link utilization. The average queue length is an indicator of the average queuing delay and, therefore, of the round-trip time. The packet drop rate impacts on the congestion control mechanism of TCP. Then, as the drop rate decreases, the goodput increases. The link utilization accounts for all the traffic in the bottleneck link, including data, acknowledgements, retransmissions, etc.

Each experiment runs for 600 seconds, using for metrics computation all but the first 60 seconds (warmup period) of the experiment. We perform our simulation varying two parameters: (i) the bundling degree (b_D), and (ii) the ratio between the number of TCP flows traversing the bottleneck ($N \times K \times b_D$) and the number of lightpaths of this bottleneck (b_D). That is, this ratio of "TCP flows per lightpath" equals to $N \times K$ according to Fig. 1.

3.1 Effect on bottleneck link utilization

In Table 1 the results for average utilization of the lightpaths in the bottleneck link are shown. Results show that utilization increases in every case, when LB+AS is applied. In addition, utilization is also higher as the number of flows increases, which is expected due to a faster TCP sender-window recovery from packet drops at the queues. Another interesting result is the higher throughput as the bundling degree increases, this is due to the less bursty traffic, since packets from more subnetworks compete for the same lightpath, even though there are the same number of flows.

Table 1. Average throughput per lightpath (in Gbps)

		Bundling degree				
# Flows per lightpath	LB+AS	2	3	4	5	10
100	Yes	8.39	8.49	8.65	8.96	9.15
	No	6.92	7.68	7.97	8.10	8.39
250	Yes	8.45	8.75	8.83	9.04	9.26
	No	7.60	7.88	8.07	8.12	8.43
500	Yes	8.49	8.99	9.21	9.42	9.51
	No	7.76	8.84	8.98	9.18	9.32

3.2 Effect on average drop rate in the bottleneck

To complete the picture, Table 2 includes the drop rate information in the bottleneck link. Main observation is that when the LB+AS paradigm is applied, drop rate is always smaller in absolute numbers. This happens even though in absolute numbers, since LB+AS also implies a higher utilization. In other words, applying LB+AS permits TCP to transmit more data with also a lower drop rate. These benefits come from the traffic smoothing effect in the bottleneck that LB+AS enforces. Consistently with this, we observe that when LB+AS is applied, the drop rate is better for higher bundling degrees: the more capacity we bundle together, the better the performance.

Finally, we note that in any case, the drop rate is higher with more TCP sources. As we saw in the previous section, more TCP flows bring a higher utilization, and this brings a higher drop rate in absolute numbers.

Table 2. Average drop rate in core nodes (in Mbps)

		Bundling degree				
# Flows per lightpath	LB+AS	2	3	4	5	10
100	Yes	43.93	35.20	30.80	21.92	12.59
	No	50.64	70.40	84.54	97.56	135.81
250	Yes	53.16	49.19	48.47	46.66	45.06
	No	51.87	53.00	53.76	57.07	58.25
500	Yes	71.48	70.89	70.69	67.73	64.12
	No	78.42	81.58	82.47	83.76	85.59

3.3 Effect on average queuing delay in the bottleneck

Table 3 collects the results for the average queue length in the core nodes at network interfaces which are connected by the lightpath bundle. Results indicate that applying LB+AS always reduces the average queuing delay in the bottleneck. Reductions are in general better for a higher bundling degree and can reach the 15%. Note also that as the bundling degree increases, moderate memory savings can be achieved also when LB+AS is not applied, since the traffic is less bursty. Finally, results are consistent with [3] in the sense that the queue occupancy is lower as the number of flows in the bottleneck increases.

Table 3. Average queue length in core nodes (in MB)

		Bundling degree				
# Flows per lightpath	LB+AS	2	3	4	5	10
100	Yes	5.76	5.64	5.29	4.77	4.48
	No	5.97	5.86	5.60	5.26	5.05
250	Yes	3.47	3.41	3.37	3.35	3.14
	No	3.71	3.53	3.47	3.43	3.26
500	Yes	3.41	3.18	3.13	2.95	2.64
	No	3.44	3.31	3.24	3.13	3.05

4. CONCLUSIONS

In this paper, we study the performance of long-lived TCP flows (‘elephants’) in an optical link under the application of the novel LB+AS paradigm. In our simulation results LB+AS improved the link utilization (up to 18%), reducing the drop rate at the same time, and decreasing the queuing delay at the bottleneck link end nodes (up to 15%). We performed a worst-case simulation in which traffic is provided only by long-lived TCP flows using the classical TCP NewReno flavor, which is not recommended for long fat networks (LFN). Further research effort is necessary to study under different mix of long-lived and short-lived TCP (with up-to-date TCP flavors such as CUBIC) and UDP flows.

ACKNOWLEDGEMENTS

This research was partially supported by the Spanish project grants TEC2010-12250-E (FIERRO) and TEC2010-21405-C02-02/TCM (CALM) and the predoctoral fellowship program of Universidad Politécnica de Cartagena. It was also developed in the framework of the project “Programa de Ayudas a Grupos de Excelencia de la Región de Murcia” funded by F. Séneca (Plan Regional de Ciencia y Tecnología 2007/2010).

REFERENCES

[1] P. Pavon-Marino, J.-L. Izquierdo-Zaragoza: Lightpath bundling and anycast switching (LB+AS): a new paradigm for multilayer optical networks, to be published in *IEEE Communications Magazine*.

[2] P. Pavon-Marino: Lightpath bundling and anycast switching: a good team for multilayer optical networks, in *Proc. ONDM 2011*, pp. 1-6, Feb. 2011.

[3] Y. Ganjali, N. McKeown: Update on buffer sizing in Internet routers, *ACM SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 5, pp. 67-70, Oct. 2006.

[4] G. Iannaccone, M. May, C. Diot: Aggregate traffic performance with active queue management and drop from tail, *ACM SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 3, pp. 4-13, Jul. 2001.

[5] V. Jacobson, R. Braden, D. Borman: TCP Extensions for High Performance, RFC 1323, May 1992.

[6] A. Varga: The OMNeT++ Discrete Event Simulation System, in *Proc. ESM’2001*, Jun. 2001.