IP over WDM dynamic link layer: challenges, open issues and comparison of files-over-lightpaths versus photonic packet switching *

M. Izal, J. Aracil
Dept. Automática y Computación
Universidad Pública de Navarra
Campus arrosadía s/n
31006 Pamplona - SPAIN
Phone: +34 948 169733 Fax: +34 948 168924
Internet: {mikel.izal,javier.aracil}@unavarra.es

April 6, 2001

Abstract

This paper addresses the suitability of WDM coarse packet switching solutions for IP traffic. Our findings show that the combination of traffic grooming at the higher layers and coarse packet switching at the optical layer provides at least the same performance as more sophisticated and difficult to realize photonic packet switching solutions. We propose a network architecture named files-over-lightpaths that not only simplifies the network optical and electronic design by making use of coarse packet switching, but also serves to the purpose of decreasing the TCP transaction latency in comparison to a flat or split Internet organization with fine grain photonic packet switching.

1 Introduction and Motivation

The optical backbones are expected to bring near-infinite bandwidth using the emerging (Dense) Wavelength Division Multiplexing (WDM) techniques [13, 1], with announced products offering 1.28 Tbps on a single fiber. We distinguish three phases in the development of the all-optical Internet: in a first stage, static WDM backbones will provide point-to-point wavelength speed channels (lightpaths). Such static lightpaths will be linking gigabit routers, as current ATM or Frame Relay permanent virtual circuits do, that will perform cell or packet forwarding in the electronic domain.

In a second generation, the WDM layer provides dynamic allocation features, by offering on-demand lightpaths or coarse packet switching solutions. The former provides a switched point-to-point connection service while the latter provides burst switching service [15]. We note that a major benefit of burst switching is higher throughput. One of the major drawbacks in optical networking is packet header processing, which is performed in the electronic domain. Thus, the maximum number of packets per second is limited by the node electronics, which become the bottleneck. Since on the contrary the optical bandwidth is abundant throughput will be maximized if the packet payload is large, so that relatively large chunks of data are transmitted in the optical domain with a single packet header processing operation in the electronic domain. That is precisely the principle of optical burst switching.

The third generation of optical networks will provide photonic packet switching, thus eliminating the electronic bottleneck. As far as the optical transmission itself, the challenge is to provide ultra-narrow optical transmitters and receivers. Even more challenging is the development of all-optical routers that perform packet header processing in the optical domain.

The deployment of an all-optical packet switching technology will not only depend on technological factors but also on cost-effectiveness issues. Indeed, the IP layer requirements should be carefully examined, so that the WDM layer is designed in a cost-effective fashion. The success of the WDM as a transfer mode for IP will basically depend

*The authors are sponsored by CICYT under contract 2FD97-0960-C05-04. Corresponding author: Mikel Izal.

on the adequate integration of both layers, so that the quality of service for the end-user is provided with a realistic network infrastructure. The lesson learned from the ATM standard, which is being questioned as the transport mode for IP today, is that over-engineering the link layer with unnecessary features becomes an inefficient solution. In fact, we are witnessing the climbing of the Packet-over-Sonet standard as the transfer mode for IP, that provides a simple and more efficient link layer alternative [14].

While it is clear that the commercially available first generation static WDM networks will not suffice for the provision of IP over WDM service, due to the inherent burstiness of Internet traffic, a second generation WDM networks wave, bringing coarse packet switching capabilities may very well satisfy the QoS requirements of the Internet. In this paper we focus on the integration of IP and WDM dynamic link layer in the forthcoming second generation optical Internet with coarse packet switching granularity. The objective of our research is to simplify bandwidth allocation at the WDM dynamic link layer thus keeping the optical network complexity at a minimum. We propose a files-over-lightpaths network model that not only provides an "optical-friendly" input to the WDM layer but also serves to the purpose of reducing the end-to-end latency for TCP connections. Furthermore, such network architecture provides a framework in which billing, pricing and differentiated QoS for IP flows can be provided in an efficient manner.

1.1 WDM Network Architecture

Figure 1 shows the WDM network architecture under analysis. The WDM wide area network serves as an Internet backbone for the different access networks, which are assumed to be large Internet administrative domains. The geographical span of the WDM ring can be metropolitan or regional, covering areas up to thousand miles. The WDM network input traffic comes from the multiplex of a large number of users (in the thousands) at each access network. Examples of access networks in our architecture are campus networks or Internet service provider networks. Access and WDM backbone will be linked by a domain border *gateway* that will perform the necessary internetworking functions. Such domain border gateway will typically consist of a high-speed IP router.

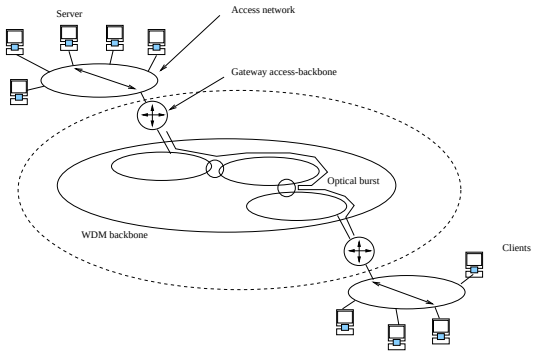


Figure 1: Network architecture

The scenario shown in figure 1 is the most likely network configuration for future all-optical backbones, that will surely have to coexist with a number of significantly different technologies in the access network such as Ethernets, wireless, HFC and xDSL networks. The deployment of fiber optics to the end-user site can be incompatible with other user requirements, such as for instance mobility, even though there is a trend towards providing high-speed access in the residential accesses. In any case, the access network will be subject to packet loss and delay due to congestion or physical layer conditions which are not likely to happen in the optical domain. On the contrary, the high speed optical backbones will provide channels with high transmission rates (in the 10 Gbps) and extremely low bit error rates (in the 10^{-15}).

2 Challenges and Open Issues for IP over WDM

The network architecture depicted in figure 1 presents a number of problems which will be taken into account in the proposal of an efficient IP over WDM transfer mode. Such problems can be classified as follows:

- Input traffic burstiness at the access node
- Access and backbone optical network adaptation
- MAC protocols adequation to IP over WDM transport

2.0.1 Input traffic burstiness at the access node

A major challenge for IP over WDM transmission from the optical network point of view comes from the fact that the input traffic has not been adequated to the particular characteristics of the IP over WDM dynamic link layer. In fact, a pure per-packet switching scheme in the source optical network access node implies that the node will be requesting access to the WDM network several times for the same IP flow. This is due to the flow packet arrival process, which is nearly random, due to the following observations: (i) since a large number of sources will be multiplexed at the gateway, different connections will be subject to interleaving at the packet level, (ii) the TCP window dynamics introduce additional randomness to the packet interarrival times, (iii) the access network may introduce significant delay jitter. The above observations are confirmed by the experimental measurements shown in figure 2 that plots the distribution of packet interarrival times within the same TCP connection from a large traffic sample at the university campus network [4], an example of access network in the scenario depicted in figure 1. We note that the order of magnitude for the packet interarrival times in the same connection is beyond the milliseconds range, several orders of magnitude larger than the packet transmission time in the optical network [1].

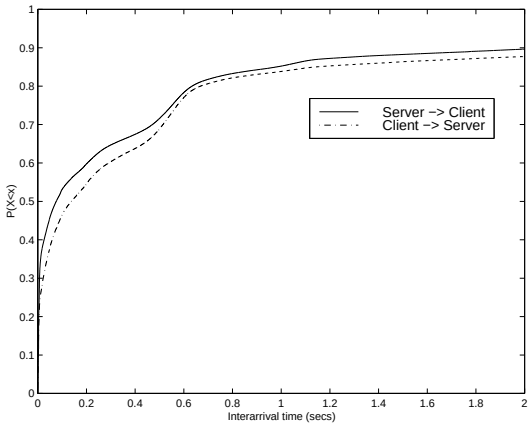


Figure 2: Packet interarrival times (in the same TCP connection)

On the other hand, the average packet length in the trace is 500 bytes in the downstream from server to client and only 150 bytes in the upstream. Short packets are mainly due to acknowledgments, since IP services are mostly asymmetric in the server to client direction. Furthermore, we note that the access network Maximum Transmission Unit (MTU) is necessarily smaller than the optical domain MTU, since for example in a wireless access networks packet size is limited due to bit error rate.

As a conclusion, we note that the transmission of IP packets with no shaping at all originates a number of packets per connection which are mostly short in size and arrive in a nearly random manner for the optical network time scale. That is a worst scenario for a twofold reason: first short packets complicate matters for the optical and electronic components whose operating frequency is increased. Secondly, the occurrence of nearly random arrivals at the backbone network routers, either electronic or optical, makes packetization delay increase heavily since the optical network minimum transmission unit may be significantly large than the packet size. We note that an adequate sorting of IP packets at the optical backbone edges results most beneficial in order to lower the packet processing burden. Indeed, for some proposals of end-to-end packet transport in optical backbones the sorting of IP packets at the optical network edges is of fundamental importance [8].

[1] The transmission time for a 1,500 bytes packet (Ethernet MTU) at 1 Gbps is equal to 12 μs.

2.0.2 Access and optical backbone network adaptation

Bridging the gap between access and backbone network also becomes an open issue. A flat architecture based on a user end-to-end TCP/IP connection, although simple and straightforward, may not be a practical solution. The TCP slow start algorithm severely constrains the use of the very large bandwidth available in the lightpath until the steady-state is reached. Even in such steady-state regime the user's socket buffer may not be large enough to provide the storage capacity needed for the huge bandwidth-delay product of the lightpath. Furthermore, an empirical study conducted by the authors in a University network [4] showed that nearly 30% of the transaction latency was due to TCP connection setup time, which poses the burden of the roundtrip time in a three-way handshake.

However, TCP provides congestion and flow control features needed in the access network, which lacks the ideal transmission conditions that are provided by the optical segment. We must notice that heterogeneous networks also exist in other scenarios, such as mobile and satellite communications. In order to adapt to the specific characteristics of each of the network segments *split TCP connection models*, an evolutionary approach of the TCP end-to-end model, have been recently proposed [7, 9]. We note that TCP splitting is not an efficient solution for optical networks, since due to the wavelength speed, in the order of Gbps, the use of TCP in the optical segment can be questioned. For example, considering a 10 Gbps wavelength bandwidth and 10 ms propagation delay in the optical backbone (2000 km) the bandwidth delay product equals 25 MBytes. File sizes in the Internet are clearly smaller than such bandwidth-delay product [11, 4]. As a result, the connection is always slow-starting, unless the initial window size is very large [10]. However, since the paths from gateway to gateway in the optical backbone have different roundtrip delays we note that the bandwidth delay product is not constant. For example, a 1 ms. deviation in roundtrip time makes the bandwidth-delay product increase in 1.25 Mbytes. Therefore, it becomes difficult to optimize TCP windows to truly achieve transmission efficiency in this scenario. Furthermore, the extremely low loss rate in the optical network makes retransmissions very unlikely to happen and since the network can operate in a burst-switched mode in the optical layer we note that there are no intermediate queues in which overflow occurs, thus making most of the TCP features not necessary.

2.0.3 Optical MAC protocols adequation to IP over WDM traffic

Current optical network architectures are based in tree, bus or ring topologies. The access gateway in figure 1 will be normally facing a feeder network, most likely an access ring as with the current SONET implementations. The ring topology has a number of advantages such as fast link restoration in case of failure. Laboratory prototypes of feeder rings are currently being developed [12] and there is an active related ongoing research [12, 6]. A number of MAC protocols have been proposed for such access rings, that can be classified as centralized or distributed. A centralized MAC protocol is based on a central resource allocator, which grants access to the network on-demand. Such access to the network consists of a wavelength allocation (lightpath) from source node to destination node for a given time, after which the lightpath is released [12]. As an alternative, several time slots in the TDM frame can be assigned. Distributed schemes are normally token-based, so that the node can grab the token in order to gain access over a certain wavelength for a given holding time [6].

We note that the fundamental problem in high-speed rings is that the access latency dominates over the transmission latency, since the transmission speed is in the order of Gbps. and beyond. In other words, the transmission slot for the access node cannot have an arbitrary duration, but should be kept within a reasonable range. If the allocated transmission slot is too small then the access latency dominates the end-to-end latency. The access node transmission slot can be viewed either as the token holding time in a token-based MAC or as the lightpath holding time in the centralized scheme. For both of them the latency is strongly related to the duration of the transmission slot as has been demonstrated in several studies [6].

2.1 Conclusions

¿From the above considerations we note that the grouping of IP packets to the same destination node in the source access node is most convenient. Since several packets are forwarded to the same destination IP address they can be encapsulated in the same container that can be switched in the optical domain, avoiding random packet arrivals from the same flow and decreasing the routing complexity. Both transmitter and receiver optical switching speed requirements are relaxed, while the electronic layer is also greatly simplified since the packet header processing operations take place at a "burst" level instead of "packet" level. Consequently, if the upper layers provide "optical-friendly" traffic to the dynamic WDM link layer a higher efficiency, throughput and cost-effectiveness can be achieved. In order to

realize this objective, we now propose a files-over-lightpaths architecture that provides traffic grooming to the WDM layer, while maintaining compatibility with the current TCP/IP protocol suite.

3 The Files-over-Lightpaths Architecture

The type of network architecture that we are seeking necessarily demands gateway functionality at the edges of the optical backbone, so that traffic grooming for the optical network can be achieved, while keeping TCP/IP compatibility and bridging the mismatch between access and backbone. A TCP splitting architecture providing concatenated TCP connections, although simple and straightforward, is not an adequate solution as noted before. In such scenario, what is clearly needed is a departure from the TCP scheme, and not an evolutionary solution providing concatenated TCP connections. Such change of paradigm in the data transmission scheme in the optical network is based on the following fundamental fact:

At wavelengths' speed of gigabits per second, files for the optical network are like packets for the access network.

In fact, the ratio between packet size (Kbytes) and transmission speed in the access network (Mbps) is in the order of thousands. In the optical network the ratio between the file size (Mbytes) and transmission speed (Gbps) is in the same order. While bit-by-bit switching in the access network would be completely inefficient the same applies to packet-per-packet switching in the optical segment. Being the optical network not limited by raw bandwidth but for packet processing at the intermediate nodes, our proposed architecture not only simplifies dramatically the optical network processing requirements, but also serves to the purpose of decreasing the overall end-to-end connection latency.

3.1 Optimizing transmission with file-switching

The concept presented in this paper goes one step further by proposing the use of a burst-switched solution to handle the end-to-end data transfer over the optical segment. At the boundaries of the optical backbone the user end-to-end connection is split into multiple segments, i.e., the user-to-gateway segments and the gateway-to-gateway segments. The server gateway retrieves the file (e.g., image, movie) from the Internet server using a TCP connection. The file is then transmitted from the server gateway to the client gateway using a burst-switched solution in the optical segment. Once the file transfer has been completed, the client gateway delivers the file to the final client using a distinct TCP connection. Figure 3 shows the three data transmission phases in the files-over-lightpaths network architecture.

The advantage of using a files-over-lightpaths architecture is many-fold. Contained end-to-end delay and efficient bandwidth utilization are achieved through the concatenation of file transfer mechanisms that are optimized for each network segment. The transfer mechanism used in the optical backbone is consistent with the emerging second generation all-optical network architectures. The end user's TCP mechanism is left unchanged, thus making it possible to transparently introduce the burst-switched transport mechanism where needed, i.e., in all-optical network islands subjected to high traffic loads. The fact that the network design relies on huge data storage devices located in the network edges is consistent with a widely accepted trend in today's telecommunications [5].

Finally, we note that the use of files-over-lightpaths provides a number of advantages other than optimization of the optical segment transmission. The first and foremost is the availability of caching space in the access gateways, as with a proxy, that decreases latency dramatically in case of cache hit. Furthermore, the TCP connection is effectively split by the cache, so that end-to-end consistency is preserved. Secondly, scheduling at the file level can be performed, allowing for differentiated quality of service and billing and pricing in the optical backbone, a most interesting feature for Internet Service Providers [16]. Providing differentiated QoS by means of scheduling at the packet level becomes a difficult task to accomplish in a very high speed environment, possibly creating bottlenecks in high load situations. Nonetheless, scheduling at the file level is more simple and efficient, since clearly the number of scheduling decisions decreases dramatically. Indeed, there is an increasing interest in flow switching solutions for Internet high-speed routers. Such solutions are based on assigning a tag to the packets belonging to the same stream prior to entering the backbone, so that the routing tasks are greatly simplified in the routers [2]. The proposed files-over-lightpaths architecture provides a truly circuit-switched solution from the optical network edges, based on burst-switching in an all-optical backbone. As a consequence, the electronic bottleneck is circumvented, the network reliability is improved and the variance of the connection latency is minimized.

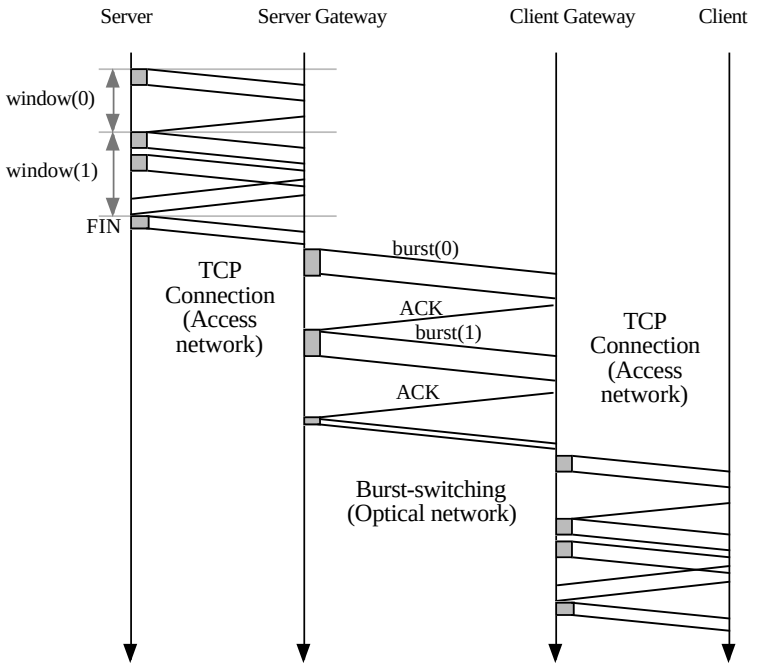


Figure 3: File transmission in the files-over-lightpaths network architecture

4 Performance evaluation of the Files Over Lightpaths architecture

4.1 Methodology

We use both analytical performance evaluation and simulation in order to study the reference WDM architecture shown in figure 3. The *ns*[2] simulator is selected as a simulation tool since an accurate TCP implementation is available. We choose a simple network topology consisting of an optical channel (1 Gbps) which connects a couple of access routers located at the boundaries of the optical network, as shown in figure 4. Regarding the access network two access links provide connectivity between the access routers and client and server respectively. We simulate a number of network conditions with varying network parameters, namely link capacities, propagation delay and loss probability. The objective is to evaluate candidate transfer modes with the performance metric being TCP connection throughput. The transfer modes under analysis are:

- End-to-End Photonic Packet Switching (EE-PPS): An end to end TCP connection is established and no splitting is performed. The optical network is a pure packet switching network at Gbps speed. This scenario resembles the current Internet, in which end-to-end connections to server/proxy are established regardless of the network segments being traversed.
- Split Photonic Packet Switching (STCP-PPS): The end-to-end TCP connection is split into two separate connections for access and backbone network. In the access network, we adopt the TCP connection parameters which are usual for most PCs in the Internet, i.e. 32 Kbytes maximum transmission window size and slow start activated. On the other hand, the backbone TCP connection uses TCP extensions for speed, as described in [10]. Such TCP extensions are larger transmission window (500 Kbytes) and no slow start, i.e. initial window size equal to 500 Kbytes. We choose not to evaluate split TCP connections with standard TCP in the optical segment since it makes no sense to use small window sizes and slow start in such optical segment, which is RTT-limited.
- Files-over-lightpath (FOL): A standard TCP connection is used between client/server and access router (32 KB window size and slow start). We consider that the optical network provides burst-switching capabilities. Optical burst transmission requires a setup time and is subject to blocking probability. Both parameters will be taken into account in the analysis.

[2]http://www-mash.cs.berkeley.edu/ns/

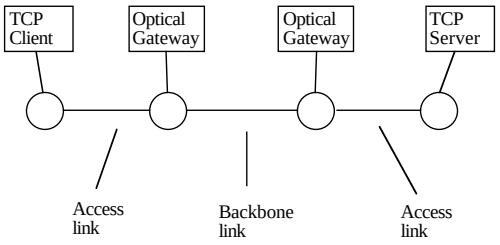


Figure 4: ns model

In order to evaluate the candidate transfer modes a number of network parameters (delay, bandwidth, loss probability) need to be selected that accurately portray the case of optical networks. We provide a summary of simulation parameters, along with their values in the different simulation runs in table 1. The simulation output is a plot showing mean throughput (i.e. total bytes divided by total duration) for file sizes in the interval $(0,5]$ MBytes. A number of 100 simulation runs is performed per each file size value and the average throughput curve is presented. In order to simulate a real situation of clients accessing a number of different destinations in the Internet propagation delay in the optical backbone is modeled as an uniform random variable in the range $(0,30]$ ms. Considering a propagation speed in the optical fiber of 5 $\mu s.$ per kilometer the backbone diameter is in the range of $(0,6000]$ Km., not taking into account intermediate router propagation. On the other hand, delay in the access network accounts for propagation and queueing delay. We note that the propagation delay in the access network depends heavily on the access technology being used. For example, a commercial ADSL network suffers a one-way delay from modem to modem header of 55 ms. aproximmately. Thus, we consider access delay values of 25, 50 and 100 ms in order to analyze a number of different access network configurations. On the other hand, we consider an access network bandwidth of 8.448 Mbps (E3), 34 Mbps (STS-1) and 100 Mbps. Lower bandwidth values make the TCP connection be bandwidth limited in the access and it makes no sense to connect such low bandwidth access networks to a Gbps optical backbone.

Parameter	Value
BW of backbone link	$1Gbps$
Backbone link propagation delay	$0-30ms$
BW of access link	$8.4Mbps, 34.4Mbps, 100Mbps$
Access link propagation delay	$25ms, 50ms, 100ms$

Table 1: Summary of simulation parameters

4.2 Results and discussion

We first evaluate split (STCP-PPS and FOL) versus non-split (EE-PPS) transfer modes. The results show that split transfer modes provide better performance. Then, we compare STCP-PPS with FOL.

4.2.1 Split versus end-to-end transfer modes

Figure 5 shows throughput versus file size for EE-PPS, STCP-PPS and FOL with the access network parameters shown in table 1. We note that throughput grows with file size towards a value which is independent of access BW, the less RTT the more steady-state throughput. For small file sizes the connection duration is dominated by setup time and slow start, which does not allow the window size to reach an steady-state value. For large files the TCP reaches steady-state and the throughput is equal to window size divided by round-trip time. Such behavior is expected in a large bandwidth-delay product network, in which connections are RTT-limited rather than bandwidth-limited.

An analytical substantiation of the above plots can be obtained by modeling the TCP connection as a series of RTT-slots. We proceed to evaluate the connection duration and then obtain the throughput as the ratio between connection size and duration. First, let *b* denote the maximum number of bytes transmitted per RTT. We note that *b* is equal the minimum of the negotiated flow control TCP window *maxcwin* and the bandwidth-delay product of the path:

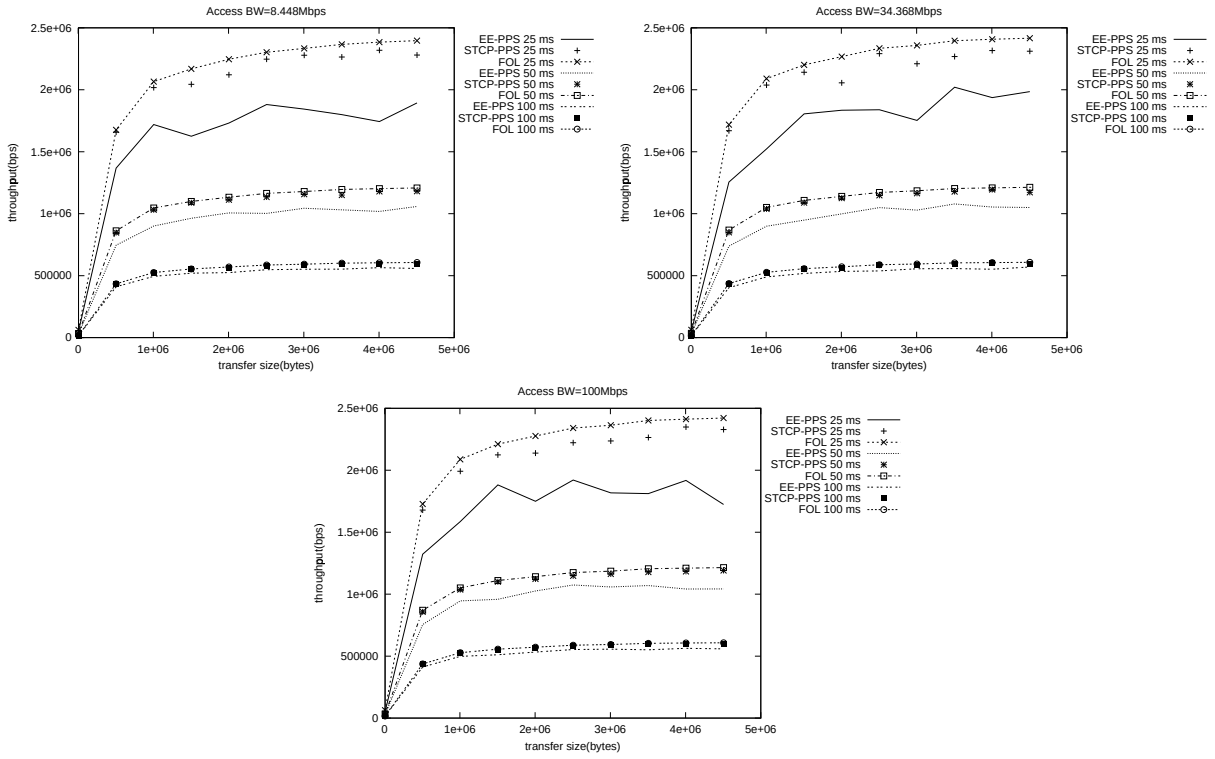


Figure 5: Average transfer throughput

$$b = min\{maxcwin, BW \times RTT\} \tag{1}$$

Since an ACK from the client serves to increase the window size in one segment in the server, an error-free TCP connection will transmit 2^i packets (with MTU bytes length) in the i-th slot during the slow-start phase. The total number of packets transmitted until slot i in the slow start phase equals $\sum_{j=1}^{i} 2^{j-1} = 2^i - 1$ Such window exponential increase phase will last until the $BW \times RTT$ product of the channel is reached, namely, until the number of packets transmitted per RTT-slot is equal to $\frac{b}{MTU}$. Thus, the i-th slot is in the slow-start phase of the connection iff:

$$2^{i-1} < \frac{b}{MTU} \quad \text{or} \quad i < log_2\left(\frac{2b}{MTU}\right) \quad \text{or} \quad i < log_2\left(\frac{b}{MTU}\right) + 1 \tag{2}$$

And the number of packets needed to complete slow start is given by:

$$2^{log_2\left(\frac{2b}{MTU}\right)} - 1 = 2\frac{b}{MTU} - 1 \tag{3}$$

Let k be the total number of packets per connection and s the total number of bytes, such that $s = k * MTU$. If $k > 2\frac{b}{MTU} - 1$ the connection is in steady state and the total number of RTT-slots until transmission of packet k is equal to the sum of two terms: $log_2\left(\frac{2b}{MTU}\right)$ slots for the slow-start phase plus the slots needed to send $k - \left(2\frac{b}{MTU} - 1\right)$ remaining packets at a rate of $\frac{b}{MTU}$ packets for slot in the steady-state phase. The total connection duration in RTT-slots for a connection size s, which we call $n(s,b)$, is equal to.

$$n(s,b) = \begin{cases} log_2(\frac{s}{MTU}) + 1 & if\ \frac{s}{MTU} \leq 2\frac{b}{MTU} - 1 \\ \frac{\frac{s}{MTU} - 2\frac{b}{MTU} + 1}{\frac{b}{MTU}} + log_2(\frac{b}{MTU}) + 1 & if\ \frac{s}{MTU} > 2\frac{s}{MTU} - 1 \end{cases} \tag{4}$$

If instead we use the high-speed TCP extensions, which avoid entering slow-start phase by using a greater initial window of *startcwin* bytes, with *startcwin* < *maxcwin*, the number of packets transmitted in slot i are equal to $2^{i-1} *$ *startcwin* and a slot i is in the slow start phase iff:

$$i < log_2\left(\frac{2b}{startcwin}\right) \tag{5}$$

The number of packets needed to complete slow start is now given by:

$$\frac{2b - startcwin}{MTU} \tag{6}$$

Thus, the time it takes for transmission of a file of size s, measured in RTT slots, which we call $n_{fs}(s,b)$ is given by:

$$n_{fs}(s,b,startcwin) = \begin{cases} log_2(\frac{s}{startcwin}) + 1 & if\ \frac{s}{MTU} \leq \frac{2b-startwin}{MTU} \\ \frac{\frac{s}{MTU} - \frac{2b-startcwin}{MTU}}{\frac{b}{MTU}} + log_2(\frac{s}{startcwin}) + 1 & if\ \frac{s}{MTU} > \frac{2b-startwin}{MTU} \end{cases} \tag{7}$$

The above formulation can be used to calculate transfer times as a function of b, *startcwin* and RTT of the network under analysis. Denoting by $RTT_{optical}$ the roundtrip time of the optical backbone path and RTT_{access} the roundtrip time of an access network, the transfer times for a file size s in the end-to-end scenario is given by:

$$t_{ee}(s) = (n(s, b_{end-to-end}) + 2)\,(2RTT_{access} + RTT_{optical}) \tag{8}$$

$$t_{stcp}(s) = (2\,(n(s, b_{access}) + 2))\,RTT_{access} + (n_{fs}(s, b_{optical}, startcwin) + 2)\,RTT_{optical} \tag{9}$$

$$t_{fol}(s) = (2\,(n(s, b_{access}) + 2))\,RTT_{access} + \left\lceil \frac{s}{l_{optical}} + 1 \right\rceil RTT_{optical} \tag{10}$$

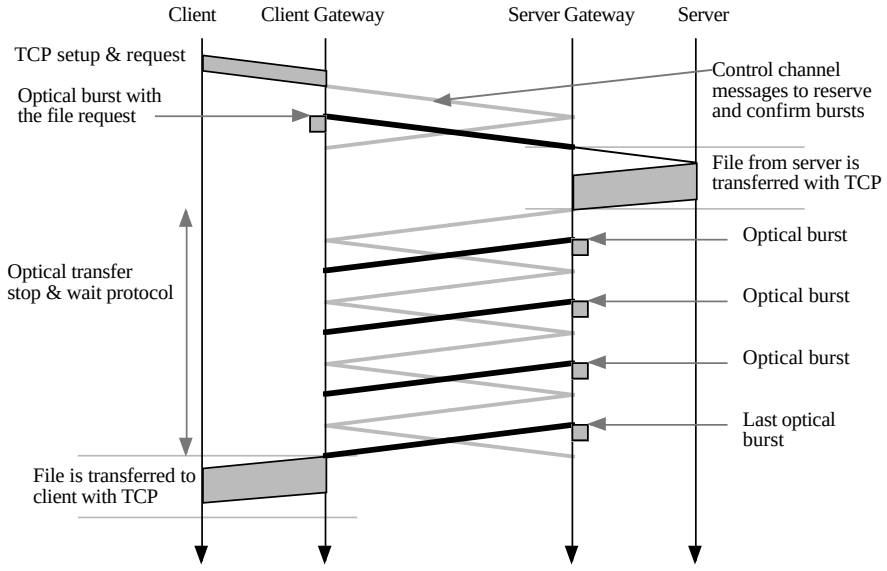


Figure 6: FOL protocol (stop and wait)

In the above expressions we take into account that two additional RTTs are consumed for the connection establishment and termination phase (SYN-SYN and FIN-FIN handshakes) and must be added to the number of minislots given by $n(s,b)$ for a given connection.

On the other hand, equation 10 accounts for the number of $RTT_{optical}$ slots using a simple stop and wait protocol and sending data in $l_{optical}$ sized bursts plus one $RTT_{optical}$ slot for the request message (figure 6). We choose a simple

stop-and-wait protocol in the optical segment as a conservative reference for comparative performance evaluation. We assume that sending an optical burst takes only one $RTT_{optical}$. Following figure 6 we consider that a burst request must be sent $\frac{RTT}{2}$ seconds before sending the burst itself. Thus, new requests are sent as soon as an ACK of the previous burst is received so that both transmission of a new burst and ACK of the previous burst are concurrent.

Besides, being b_{access}, $b_{optical}$ and $b_{end-to-end}$ are the maximum burst sizes generated by TCP in an access, optical or end-to-end path, we consider equation 1 and obtain:

$$b_{access} = min\{32Kbytes, BW_{access} \times RTT_{access}\} \tag{11}$$

$$b_{optical} = min\{500Kbytes, BW_{optical} \times RTT_{optical}\} \tag{12}$$

$$b_{end-to-end} = min\{32Kbytes, BW_{access} \times (RTT_{access} + 2RTT_{optical})\} \tag{13}$$

Finally, the connection achieved throughput can be calculated by simply dividing file size per total transfer time as follows:

$$thr_{scenario}(s) = \frac{s}{t_{scenario}(s)} \tag{14}$$

where *scenario* is in the set $\{ee, stcp, fol\}$. Figure 7 shows the connection throughput, obtained analytically (in solid lines) and by the simulation (points). We consider the following parameter values:

Parameter	Value
BW of backbone link	$1Gbps$
Backbone link propagation delay	$0-30ms$
$RTT_{optical}$	$0-60ms$
BW of access link	$8.4Mbps, 34.4Mbps, 100Mbps$
Access link propagation delay	$25ms, 50ms, 100ms$
Optical segment TCP *startcwin*	$500Kbytes$
Optical segment TCP *maxcwin*	$500Kbytes$
Access segment TCP *maccwin*	$32Kbytes$
Optical network maximum burst size (FOL)	$500Kbytes$

Table 2: Simulation parameters

First, we note that the analytical results follow closely the simulation results, thus indicating that TCP connections in such high $BW \times RTT$ product behave as a series of mini-cycles of RTT duration, in which a number of packets b (equation 1) is transmitted in the steady-state phase. ¿From figure 5 we also observe that *for a given access delay, proxy-based strategies, either STCP or FOL, get similar results and outperform PPS*. If file sizes are large enough to make TCP reach steady-state then STCP and FOL are constrained by the speed of access loops only, thus providing similar results. This can be seen in equation 10 for the FOL scenario and in equation 9 for the STCP scenario. Since $n_{fs}(s,b)$ is much lower than $n(s,b)$ and $\left\lceil \frac{s}{l_{optical}} \right\rceil$ is lower than $n_{fs}(s,b)$ then the access contribution to latency is always much higher than that of the backbone and both split method behave like

$$t_{stcp,fol}(s) = 2\left(n(s, b_{access}) + 2\right) RTT_{access} \tag{15}$$

Regarding EE-PPS, we note that an end-to-end TCP loop has a higher round trip time than that of the access network thus providing a larger value for $n(s,b)$. Equation 15 gives lower latency values than equation 8 and such values are lower the lower is $RTT_{optical}$ compared to RTT_{access}. This effect can also be appreciated in the simulation results plotted in fig 5. We note that the access bandwidth (the bottleneck bandwidth) shows no effect in the obtained throughput. Since the maximum advertised window is not larger than the $BW \times RTT$ product of the path and the data transfer rate is limited by such maximum advertised window value (equation 1), which should not be increased. While in the optical backbone it seems reasonable to use TCP with large window sizes and no slow start the same does not apply to the access network since: i) the access network is bandwidth-limited in comparison to the optical network. Large buffer sizes produce increased traffic burstiness, leading to potential congestion; ii) buffer size at both client and server is limited by memory resources.

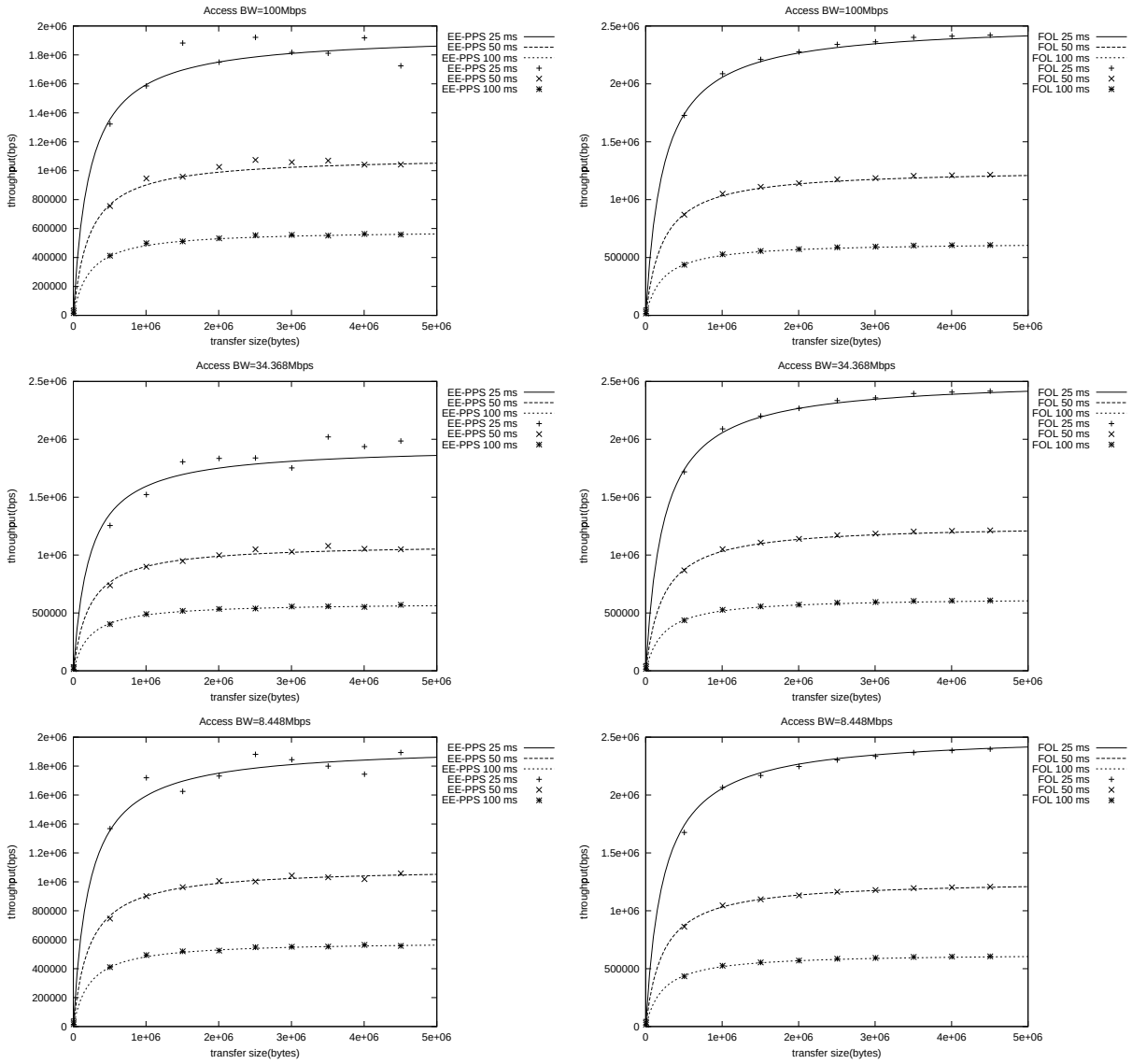


Figure 7: Theoretical vs simulation

We note that split solutions, namely STCP and FOL, provide better performance in comparison to the end-to-end PPS. Both STCP and FOL provide fast transmission in the optical backbone, with negligible transfer time, so that the TCP connection time is mainly due to the access network. Since slow start grows much faster in the access network, in comparison to the end-to-end case, it turns out that the split connections outperform end-to-end connections. While the results in this section have been obtained assuming ideal conditions (packet/burst drop probability equal zero) we note that this is actually the best possible scenario for end-to-end transfer modes. If packet loss occurs the retransmission loop for an end-to-end connection encompasses the whole network (access + backbone), thus implying severe performance drop in comparison to split methods. However, in order to compare STCP-PPS and FOL we introduce the burst blocking probability and packet blocking probability as a parameter of the optical segment in the next section.

4.2.2 STCP versus FOL

For simplicity, we assume that optical switches have no queues and, therefore, packets or bursts can be blocked due to contention for the same output port from different input ports (figure 8).

Let $(\lambda_{pack}, t_{pack})$ and $(\lambda_{burst}, t_{burst})$ be the arrival rates and transmission times for packet and burst arrival processes respectively. We assume that both the packet and burst arrival process to the switch are Poisson. By doing so we can

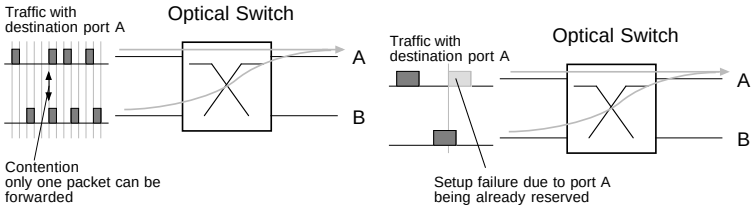


Figure 8: Backbone optical switch

make drop probability for a packet in photonic packet switching be equal to burst blocking probability in optical burst switching.

The packet drop probability is given by the probability that one or more packets arrive during the vulnerable interval of a given packet:

$$P_{drop} = 1 - e^{-\lambda_{pack} t_{pack}} \tag{16}$$

Conversely, the burst blocking probability is given by:

$$P_{block} = 1 - e^{-\lambda_{burst} t_{burst}} \tag{17}$$

Let n be the number of packets per burst. We verify that $\lambda_{burst} = \lambda_{pack}/n$ and $t_{burst} = t_{pack} n$. Thus:

$$\lambda_{pack} t_{pack} = \lambda_{burst} t_{burst} \tag{18}$$

and the packet dropping and burst blocking probabilities are equal. We now proceed to evaluate FOL and STCP under loss conditions in the optical network. The number of bursts per file (n_{burst}) is given by:

$$n_{burst} = \frac{B}{l_{optical}} \tag{19}$$

Assuming a simple stop and wait protocol (see figure 6) the average transmission time for a burst is equal to

$$E[T_{burst}] = \frac{1}{1 - P_{burst}} RTT = \frac{RTT}{1 - P_{burst}} \tag{20}$$

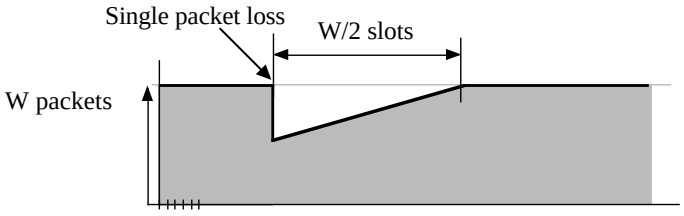
and the throughput is given by:

$$Thr_{FOL} = \frac{B}{n_{burst} T_{burst}} = \frac{B(1 - P_{burst})}{n_{burst} RTT} = \frac{L_{burst}}{RTT}(1 - P_{burst}) \tag{21}$$

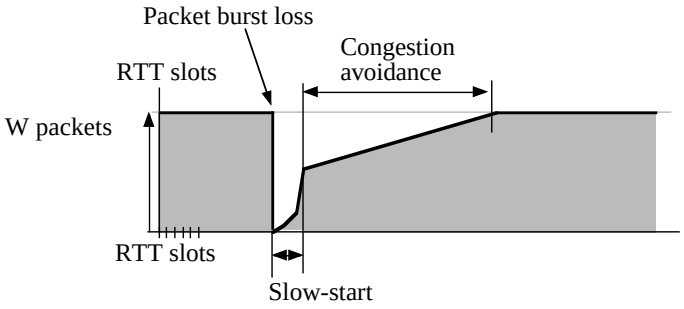
On the other hand TCP throughput will also be affected by packet drop probability. The occurrence of packet drops in an ongoing TCP flow will trigger either TCP congestion avoidance mechanisms or slow-start, as shown in figure 9. In low loss probability environment the packet loss will be detected by a duplicate ACK sent by the server. Congestion avoidance follows, which halves the window sizes and switches the TCP agent to linear growth in the congestion window. In a high loss environment the packet loss will be detected by timeout. The congestion window drops to 1 segment and the slow-start algorithm is activated until half the original rate is reached and at that point the TCP agent switches to congestion avoidance mode.

The estimation of the throughput of a TCP connection with random loss has been extensively treated in the literature. We follow [3] and obtain:

$$Thr_{TCP} = min\left(\frac{cwmax}{RTT}, \frac{1}{RTT\sqrt{\frac{P_{drop}}{3}}}\right) \tag{22}$$



a) Low loss probability



b) High loss probability

Figure 9: Errors during STCP transfer

where *cwmax* is the negotiated (flow control) window size. Figure 10 shows the achieved throughput with error probabilities lower than 0.1, for both FOL (different burst sizes) and STCP-PPS. Results for STCP-PPS have also been validated by simulation. We observe that TCP congestion avoidance severely limits transfer efficiency. *If loss probability is equal to 0.01 the throughput obtained with TCP is half the throughput obtained with a simple stop and wait protocol in FOL*. This serves to illustrate that the throughput penalty imposed by the TCP congestion control mechanisms is rather significant.

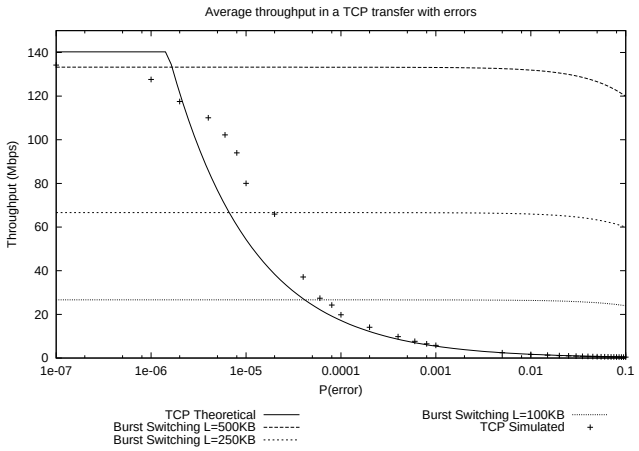


Figure 10: Throughput comparison

The main difference between a simple FOL protocol and TCP is the way both protocols interpret congestion. While TCP considers that loss is produced by queueing overflow FOL is aware that loss is due to blocking. In a loss situation, TCP will lower the transmission window, which results in *no effect at all* since congestion is due to blocking, and the more blocking probability the more number of accesses to the optical network. Furthermore, since the $BW \times RTT$ product is extremely large the TCP window size takes on a very high value. As a result, the slow start or congestion avoidance phase which follow a packet loss take the longest time to complete.

5 Conclusions

In this paper we have reported on a new architecture that efficiently tackles the challenges imposed by TCP/IP services over optical backbones. The proposed files-over-lightpaths solution provides better performance with less stringent requirements in the optical layer, which is not required to perform all-optical packet switching but coarse packet switching. We note that the grooming of traffic at the upper layers is fundamental in order to reduce complexity at the optical layer, that leads to a more efficient and cost effective network design. While the end-to-end and split model provided an effective framework for the development of the current Internet the forthcoming high-speed optical backbones claim for a different transfer paradigm, that translates the availability of gigabit bandwidth into user-perceivable quality of service.

References

[1] IEEE Communications Magazine, Optical Networks, Communications Systems and Devices, Globalization of Software Radio, February 1999.

[2] IEEE communications magazine, special issue on MPLS, December 1999.

[3] Modeling TCP reno performance: A simple model and its empirical validation. *IEEE/ACM Transactions on Networking*, 8(2), April 2000.

[4] J. Aracil, D. Morato, and M. Izal. Analysis of Internet services for IP over ATM links. *IEEE Communications Magazine*, December 1999.

[5] G. Barish and K. Obraczka. World wide web caching: Trends and techniques. *IEEE Communications Magazine*, 38(5), May 2000.

[6] A. Fumagali, J. Cai, and I. Chlamtac. The multi-token inter-arrival time (MTIT) access protocol for supporting IP over WDM ring network. In *IEEE International Conference on Communication (ICC)*, Vancouver, Canada, June 1999.

[7] E. Amir H. Balakrishnan, S. Seshan and R. Katz. Improving TCP/IP performance over wireless networks. In *ACM MOBICOM'95*, Berkeley, CA, 1995.

[8] A. Fumagalli I. Chlamtac, V. Elek and C. Szabo. Scalable WDM access network architecture based on photonic slot routing. *IEEE/ACM Transactions on Networking*, 7(1), February 1999.

[9] R. Cohen I. Minei. High-speed Internet access through unidirectional geostationary channels. *IEEE Journal on Selected Areas in Communications*, 17(2):345–359, February 1999.

[10] V. Jacobson and R.Braden. TCP extensions for long-delay paths. RFC 1072, October 1998.

[11] G. Miller K. Thompson and R. Wilder. Wide-area internet traffic patterns and characteristics. *IEEE Network*, pages 10–23, November/December 1997.

[12] K. Kuznetsov, N. M. Froberg, and Eytan Modiano et. al. Next generation optical regional access networks. *IEEE Communications Magazine*, January 2000.

[13] IEEE Communications Magazine, February 1998.

[14] J. Manchester, J.Anderson, B.Doshi, and S. Davida. IP over SONET. *IEEE Communications Magazine*, May 1998.

[15] C. Qiao and M. Yoo. Optical burst switching (OBS) - A new paradigm for an optical Internet. *Journal of High-Speed Networks*, 8(1), 1999.

[16] M. Yoo, C. Qiao, and S. Dixit. Optical burst switching for service differentiation in the next generation optical Internet. *IEEE Communications Magazine*, 39(2), February 2001.