11 Traffic Management for IP over WDM Networks

JAVIER ARACIL, DANIEL MORATO † and MIKEL IZAL
Public University of Navarra, Spain

11.1 INTRODUCTION

The ever-increasing growth of the current Internet traffic volume is demanding a bandwidth which can only be satisfied by means of optical technology. However, while WDM networks will bring an unprecedented amount of available bandwidth, traffic management techniques become necessary in order to make this bandwidth readily available to the end user. In this chapter we provide an overview of traffic management techniques for IP over WDM networks. More specifically, we address the following issues: i) Why is traffic management necessary in IP over WDM networks? and ii) What are the traffic management techniques currently proposed for IP over WDM?

Concerning the first issue, we provide an intuitive, rather than mathematically concise, explanation about the specific features of Internet traffic that make IP traffic management particularly challenging. Internet traffic shows self-similarity features and there is a lack of practical network dimensioning rules for IP networks. On the other hand, the traffic demands are highly non-stationary and, therefore, the traffic demand matrix, which is the input for any traffic grooming algorithm, is time-dependent. Furthermore, traffic sources cannot be assumed to be homogeneous, since very few users produce a large fraction of traffic. In dynamic IP over WDM networks we experience the problem of the multiservice nature of IP traffic. Since a myriad of new services and protocols are being introduced on a day-by-day basis it turns out that the provision of quality of service on demand becomes complicated. We analyze the service and protocol breakdown of a real Internet link in order to show the complexity of the dynamic QoS allocation problem.

Regarding existing and proposed traffic management techniques we distinguish between static and dynamic WDM networks. Static WDM networks are dimensioned

† Presently on leave at University of California, Berkeley

beforehand and the dynamic bandwidth allocation capabilities are very restricted. Such static bandwidth is not well suited to the bursty nature of Internet traffic and, as a result, traffic peaks require buffering, which is difficult to realize in the optical domain. Deflection routing techniques, using the spare bandwidth provided by overflow lightpaths, have been proposed as a technique to minimize packet loss and reduce optical buffering to a minimum. We investigate the advantages and disadvantages of such techniques, with emphasis on the particular features of IP overflow traffic. In the dynamic WDM networks scenario we study the Optical Burst Switching paradigm [31] as a means to incorporate coarse packet switching service to the IP over WDM networks.

To end up the chapter we focus on end-to-end issues to provide QoS to the end user applications. Due to the availability of Tbps in a single fiber the TCP needs to be adapted to this high-speed scenario. Specifically, we analyze the TCP extensions for high-speed, together with split TCP connections techniques, as a solution to provide QoS to the end user in a heterogeneous access-backbone network scenario.

11.1.1 Network scenario

Along the chapter we assume the following three waves in the deployment of WDM technology: i) First generation or static lightpath networks, ii) Second generation or dynamic lightpath/Optical Burst Switching networks and iii) Third generation or Photonic Packet Switching networks.

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. Such gigabit routers perform cell or packet forwarding in the electronic domain.

In a second generation WDM network, 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. Precisely, Optical Burst Switching (OBS) [31] provides a transfer mode which is halfway between circuit and packet switching. In OBS, a reservation message is sent beforehand so that resources are reserved for the incoming burst, which carries several packets to the same destination. In doing so, a single signaling message serves to transfer several packets, thus maximizing transmission efficiency and avoiding circuit setup overhead.

The third generation of optical networks will provide photonic packet switching, thus eliminating the electronic bottleneck. As far as the optical transmission is concerned, 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. In fact, while packet header processing is very likely to remain in the electronic domain all-optical packet routers based on optical codes are currently under development [38].

11.2 TRAFFIC MANAGEMENT IN IP NETWORKS: WHY?

There is a well-established theory for dimensioning telephone networks [9] based on the hypothesis of the call arrival process being Poisson and the call duration being well-modeled as an exponential random variable. A blocking probability objective can be set and, by means of the Erlangian theory, telephone lines can be dimensioned to achieve the target degree of service.

IP over WDM networks and, in general, IP networks, lack such dimensioning rules due to a twofold reason: first the traffic is no longer Poissonian but shows self-similar features, non-stationarity and source heterogeneity. Secondly, there is very scarce network dimensioning theory for self-similar traffic. In this section we examine the specific features of self-similar traffic that are not present in other kinds of traffic, such as voice traffic. Such features provide the justification for traffic management techniques in IP over WDM networks.

11.2.1 Self-similarity

In the recent past, voice and data traffic modeling has been primarily based on processes of independent increments, such as the Poisson process. Arrivals in disjoint time intervals are assumed to be independent since, intuitively, the fact that a user makes a phone call has no influence in other users making different calls. More formally, the call arrival process has a correlation which is equal to zero in disjoint time intervals. However, the independent increments property does not hold for Internet traffic. Not only traffic in disjoint intervals is correlated but the correlation decays *slowly* [1], meaning that even if the time intervals under consideration are far apart from one another the traffic is still correlated.

The traffic process is said to have long memory or *long-range dependence*. In the specific case of Internet traffic, such long-range dependence is observed in the packet counting process which represents the number of bytes in fixed duration (δ ms) intervals [20, 29].

As a consequence of the slow decay of the autocorrelation function the overflow probability in intermediate router queues increases heavily, in comparison to a process with independent increments (Poisson). In [26] an experimental queueing analysis with long-range dependent traffic is presented, which compares an original Internet traffic trace with a shuffled version, i.e. with destroyed correlations. The results show a dramatic impact in server performance due to long-range dependence.

Figure 11.1 shows the results from trace-driven simulations of a single server infinite buffer system with self-similar traffic. The self-similar trace is shuffled in order to show the effect of dependence in queueing performance. We note that the saturation breakpoint for a queueing system with self-similar input occurs at lower utilization factor in comparison to the same system with Poissonian input.

[1] Correlation decays as a power law $\rho(k) \approx H(2H-1)k^{2H-2}$, k being the time lag.

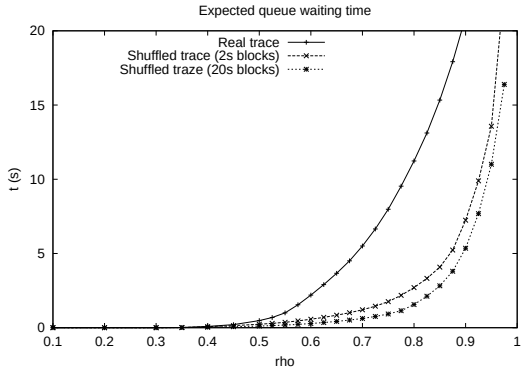

Figure 11.1 Queueing performance with self-similar process

Such performance drop is due to the presence of bursts at any time scale. In order to visually assess such phenomenon, Figure 11.2 shows traffic in several timescales (10, 100 and 1000 ms), for a real traffic trace and a Poisson process. We observe that while the Poissonian traffic smooth out with the timescale towards the rate λ the real traffic shows *burstiness at all timescales*. The self-similarity property can be explained in terms of aggregation of highly variable sources. First, the most part of Internet traffic is due to TCP connections from the WWW service [4]. Secondly, both WWW objects size and duration can be well modeled with a Pareto random variable, with finite mean but infinite variance. The aggregation (multiplex) of such connections shows self-similarity properties [13, 33]. As a result, self-similarity is an inherent property of Internet traffic, which is due to the superposition of a very large number of connections with heavy-tailed duration.

11.2.2 Demand analysis

Optical networks will provide service to a large number of users requiring voice, video and data services. In order to ensure a QoS to such users there is a need for demand estimation, so that resources in the network can be dimensioned appropriately. In the telephone network it is well-known that users are *homogeneous*, meaning that the traffic demand generated by a number of n users is simply equal to the sum of their individual demands, namely $I = I'n$, with I' equal to the demand of an individual user.

The homogeneity assumption is most convenient since the operator will dimension the network taking as a base a single user demand and extrapolating to the rest of the population. However, the traffic demand for the Internet radically differs from that of telephone networks and the homogeneity assumption is no longer valid.

We take a traffic trace from our University access link (see section 11.5) and show, in Figure 11.3 (left), the total volume of traffic generated by a group of n

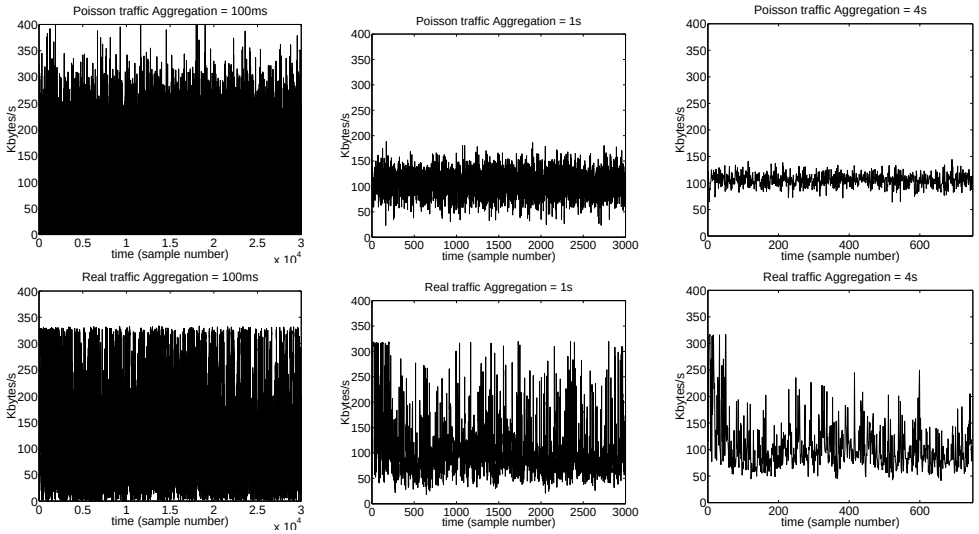

Figure 11.2 Input traffic in several timescales

users, sorted according to volume of traffic. We observe that a very few number of users is producing a large volume of traffic. Figure 11.3 (right) shows percentage of total volume of traffic in a working day versus percentage of users generating such traffic. Less than 10% of the users produce more than 90% of traffic, showing strong non-homogeneity in the traffic demand. This result is in agreement with other measurements taken at UC Berkeley campus [14].

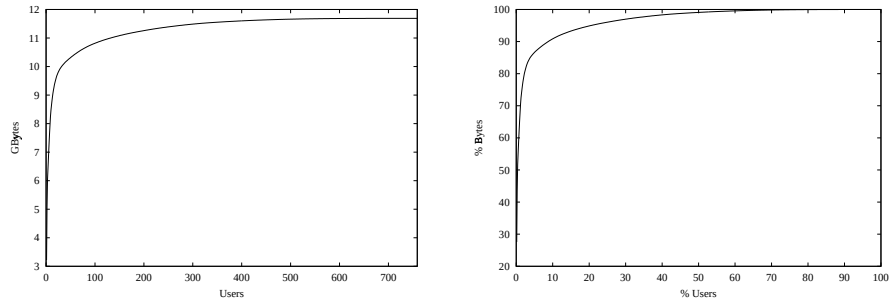

Figure 11.3 Traffic demand sorted by user. Left: Absolute value. Right: Percentage of bytes versus percentage of users

The results above show that user population can be divided into three groups: bulk users, average and light users. We note that the appearance of a sole new bulk user

can produce a sudden increase of network load. For instance, the top producer is responsible for the 27.56% of the traffic. Should her machine be disconnected the traffic load would decrease accordingly. As a conclusion, the existence of a group of top producers complicates matters for network dimensioning. For Internet traffic, the assumption that demand scales with the number of users does not hold. On the contrary, demand is highly dependent with a very few users.

Furthermore, the dynamic behavior of top producers does not follow a general law. Figure 11.4 shows the accumulated number of TCP connections and bytes from the first and seventh top producers. While the former is a highly regular user, the latter produces a burst of traffic lasting several minutes, and nearly no activity for the rest of the time.

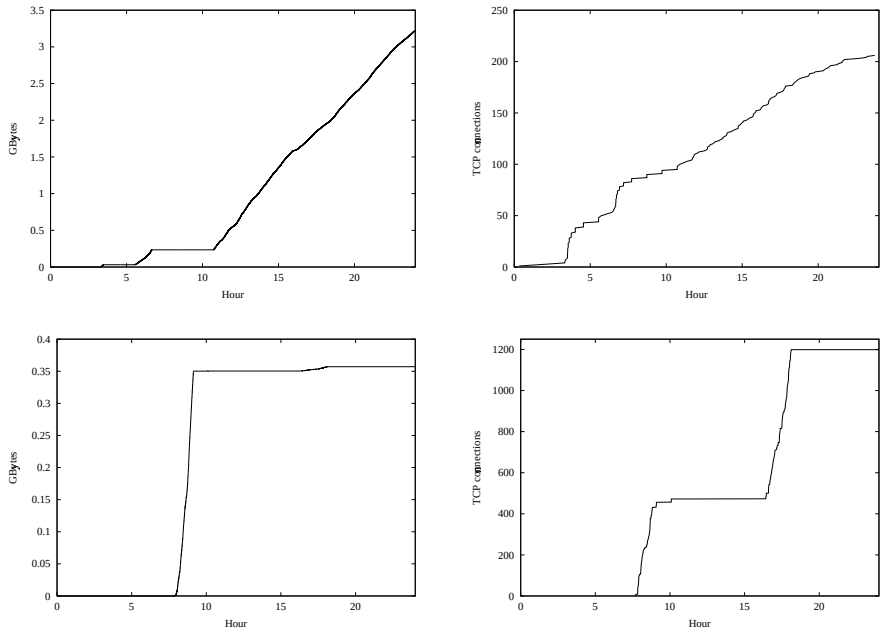

Figure 11.4 Demand versus time for the first (top) and seventh (bottom) producers. Left: TCP bytes. Right: TCP connections

11.2.3 Connection level analysis

Optical networks will surely evolve to providing QoS on demand at the optical layer, in a forthcoming photonic packet switching or lightpath-on-demand scenario. Nowadays, flow-switching mechanisms are being incorporated to IP routers in order to provide QoS at the connection level. To this end, an analysis of the Internet traffic at the connection level is mandatory. First, we observe that Internet traffic is mostly

due to TCP connections. Table 11.1 shows a breakdown per protocol of the total traffic in our University link. We note that TCP is dominant, followed at a significant distance by UDP.

Table 11.1 Traffic breakdown per protocol

Protocol	MBytes in	% Bytes in	MBytes out	% Bytes out
TCP	13718	99.03%	2835	89.38%
UDP	109	0.8%	317	10%
ICMP	13.7	0.1%	12.3	0.39%
Other	10.2	0.07%	7.4	0.23%

On the other hand, TCP traffic is highly assymetric in the outbound from server to client. Our traces show that 82.1% of the total TCP traffic corresponds to the server to client outbound. Next, we analyze what is the service breakdown for TCP traffic.

11.2.3.1 Breakdown per service Figure 11.5 shows the number of bytes per TCP service. The WWW service, either through port 80 or variants due to proxies or secure transactions (port 443), is the most popular service in the Internet. We also note that there are a number of services which cannot be classified *a-priori* since they do not use well-known ports. This is an added difficulty to per-flow discrimination in flow-switching schemes.

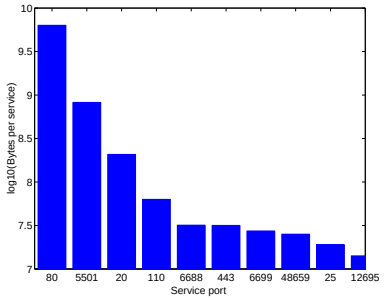

Figure 11.5 Breakdown per service

11.2.3.2 Connection size and duration Since the most part of IP traffic is due to TCP connections to WWW servers we focus our analysis in connection size and duration of such TCP connections. Figures 11.6 and 11.7 show survival functions [2] of connection size (bytes) and duration (seconds) in log-log scales. We note that

[2]A survival function provides the probability that a random variable X takes values larger than x, i. e. $P(X > x)$.

the tail of the survival function fits well with that of a Pareto distribution with α parameter. We plot the distribution tail least squared regression line in both plots and estimate values of α of 1.15 and 1.2 for duration and size respectively. Such values are in accordance with previous studies that report values of 1.1 and 1.2 [13].

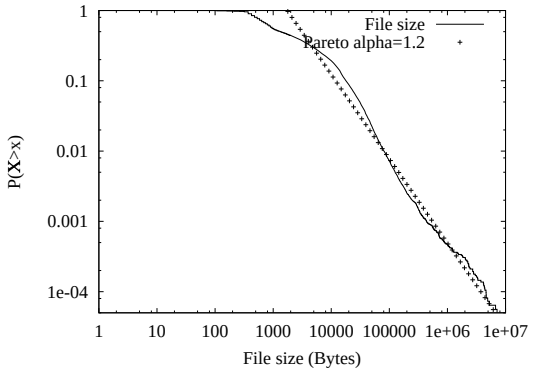

Figure 11.6 Survival function of file size

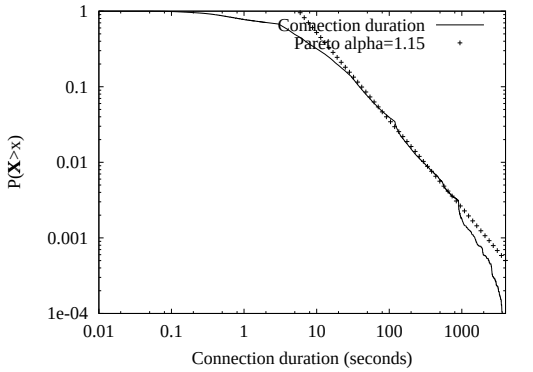

Figure 11.7 Survival function of connection duration

File sizes in the Internet are heavy-tailed due to the diverse nature of posted information which ranges from small text files to very large video files [13]. Regarding duration, we note an even larger variability (lower α), which can be due to the dynamics of the TCP in presence of congestion, which will make connection duration grow larger if packet loss occurs.

While connection size and duration are heavy-tailed, possibly causing the self-similarity features of the resulting traffic multiplex [13, 33], we note that heavy-tailedness is a property of the *tail* of the connection and size distribution. Actually,

the most part of TCP connections are short in size and duration. Figure 11.8 shows a histogram of both size and duration of TCP connections. We note that 75% of the connections have less than 6Kbytes and last less than 11 seconds. Such short lasting flows complicate matters for per-flow bandwidth allocation since the resource reservation overhead is significant. As a conclusion, not only a high flexibility and granularity in bandwidth allocation is required from the optical layer but also flow classification and discrimination capabilities, in order to distinguish which flows may be assigned separate resources.

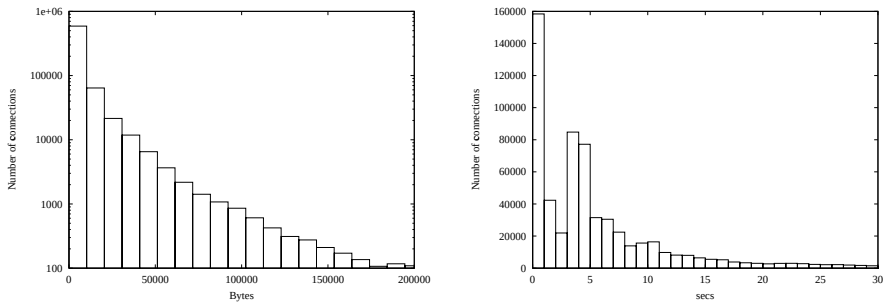

Figure 11.8 Histogram of TCP size (left) and duration (right)

11.3 IP TRAFFIC MANAGEMENT IN IP OVER WDM NETWORKS: HOW

In this section we focus on the impact of IP traffic in WDM networks in first (static lightpath) and second generation (dynamic lightpath/OBS) WDM networks. Concerning the former we explain the scenario and the traffic grooming problem. We also consider the use of overflow bandwidth to absorb traffic peaks. Concerning second generation networks we analyze the tradeoff between burstiness and long range dependence and the queueing performance implications. Then, we consider signalling aspects, IP encapsulation over WDM and label switching solutions.

11.3.1 First generation WDM networks

First generation WDM networks provide static lightpaths between network endpoints. The challenge is to provide a *virtual topology* that maximizes throughput and minimizes delay out of a physical topology consisting of a network topology with optical crossconnects linking fibers with limited number of wavelengths per fiber.

It has been shown that the general optimization problem for the virtual topology is NP-complete [16]. Such optimization problem takes as a parameter the traffic matrix, which is assumed to be constant, and, as additional assumptions, packet interarrival

times and packet service times at the nodes are independent and exponentially distributed, so that M/M/1 queueing results can be applied to each hop. We note that traffic stationarity and homogeneity become necessary conditions to assume that a traffic matrix is constant. A non-stationary traffic process provides an offered load which is time-dependent. On the other hand, a non-homogeneous demand, as shown in the previous section, may induce large fluctuations in the traffic flows, since a sole bulk user produces a significant share of traffic. If, for instance, the top producer traffic changes destinations at a given time the traffic matrix is severely affected. Furthermore the independence assumption in packet arrival times is in contrast to the long range dependence properties of Internet traffic.

A number of heuristic algorithms have been proposed to optimize the virtual topology of lightpaths (for further reference see [24, part III]), assuming a constant traffic matrix. We note that even though such algorithms provide an optimization of the physical topology chances are that traffic bursts cannot be absorbed by the static lightpaths. Being the buffering capabilities of the optical network relatively small compared to the electronic counterpart a number of proposals based on overflow or *deflection routing* have appeared recently.

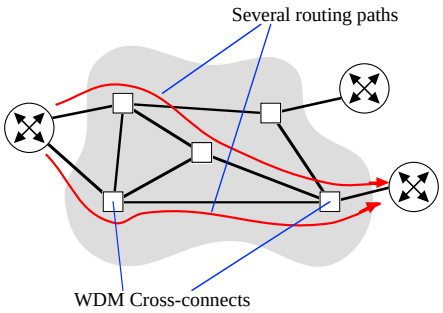

Figure 11.9 WDM network

Figure 11.9 presents a most common scenario for a first generation optical networks. IP routers use the WDM layer as a link layer with multiple parallel channels, several of those being used for protection or overflow traffic, which leads to a network design with little buffering at the routers and a number of alternate paths to absorb traffic peaks. The same scenario is normally assumed in deflection routing networks, which are based on the principle of providing nearly no buffering at the network interconnection elements but several alternate paths between source and destination, so that the high-speed network becomes a distributed buffering system. The advantage is that buffer requirements at the routers are relaxed, thus simplifying the electronic design. In the WDM case, we note that the backup channels can be used to provide an alternate path for the overflow traffic as proposed in [5]. Rather than handling the traffic burstiness via buffering, leading to delay and packet loss in the electronic bottleneck, the backup channels can be used to absorb overflow traffic

bursts. Interestingly, the availability of multiple parallel channels between source and destination routers resembles the telephone network scenario, in which a number of alternate paths (tandem switching) are available between end offices. The reason for providing multiple paths is not only for protection in case of failure of the direct link but also availability of additional bandwidth for the peak hours.

The appearance of overflow traffic in the future Optical Internet poses a new scenario which has not been studied before, since standard electronic networks are based on buffering. While the behavior of a single-server infinite queue with self-similar input has been well described [27, 37] there is little literature on the study of overflow Internet traffic. On the contrary, due to the relevance of overflow traffic in circuit switched networks there is an extensive treatment of Poissonian overflow traffic. Precisely, in the early stages of deployment of telephone networks A. K. Erlang found that the overflow traffic can no longer be regarded as Poissonian. In fact, the overflow traffic burstiness is higher than in the Poissonian model. Equivalent Erlangian models for blocking probability calculations can be established [9], that incorporate the overflow load effect. The overflow traffic can be characterized by a Poissonian input with higher intensity, in order to account for the burstiness of the latter.

In order to visually illustrate the burstiness increase of overflow traffic we plot several instance of overflow versus total traffic, for several cut-off values in Figure 11.10. The Figure shows a significant burstiness increase for overflow traffic, which, consequently, requires a different treatment in terms of network and router dimensioning [3].

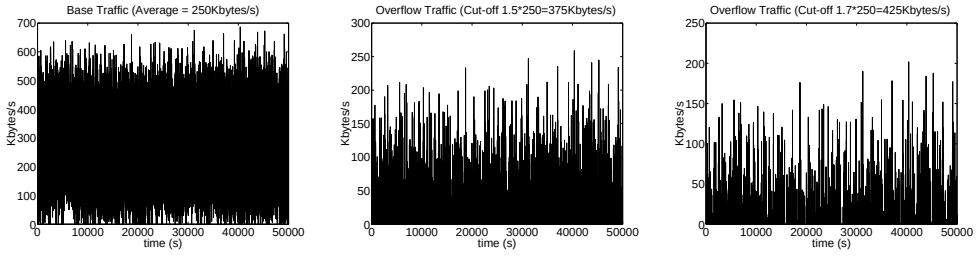

Figure 11.10 Original input and overflow traffic

11.3.2 Second generation WDM networks

Second generation WDM networks will bring a higher degree of flexibility in bandwidth allotment in comparison to first generation static networks. Prior to photonic packet switching networks, which are difficult to realize with current technology [39], dynamic lightpath networks make it possible to reconfigure the lightpaths virtual topology according to the varying traffic demand conditions. However, the provision of bandwidth on demand on a per-user or per-connection basis cannot be achieved with an optical network providing lightpaths on-demand, since they provide

a circuit switching solution with channel capacity equal to the wavelength capacity. As the next step, optical burst switching [31, 32] provides a transfer mode which is halfway between circuit switching and pure packet switching. At the edges of the optical network packets are encapsulated in an optical burst, which contains a number of IP packets to the same destination. There is a minimum burst size due to physical limitations in the optical network, which is unable to cope with packets of arbitrary size (photonic packet switching).

Optical burst switching is based on the principle of "on the fly" resource reservation. A reservation message is sent along the path from origin to destination in order to set up the resources (bandwidth and buffers) for the incoming burst. A time interval after the reservation message has been sent, and without waiting for a confirmation (circuit switching), the burst is released from the origin node. As a result of the lack of confirmation there is a dropping probability for the burst. Nevertheless, resources are statistically guaranteed and the circuit setup overhead is circumvented. Figure 11.11 shows the reservation and transmission procedure for optical burst switching.

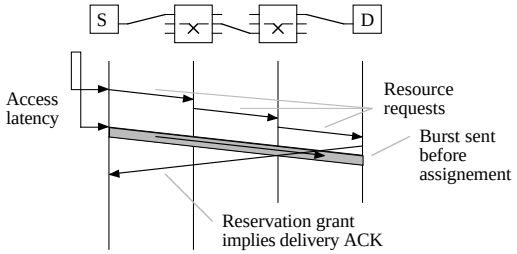

Figure 11.11 Optical Burst Switching

On the other hand, optical burst switching allows for differentiated quality of service by appropriately setting the value of the time interval between release of the resource reservation message and transmission of the optical burst [30, 40]. By doing so, some bursts are prioritized over the rest, and, thus, they are granted a better QoS.

Even though the concept of OBS has attracted considerable research attention there is scarce literature concerning practical implementations and impact in traffic engineering. A reference model for an OBS edge node is depicted in Figure 11.12. Incoming packets to the optical cloud are demultiplexed according to their destination in separate queues. A timer is started with the first packet in a queue and, upon timeout expiration, the burst is assembled and relayed to the transmission queue, possibly requiring padding to reach the minimum burst size. Alternatively, a threshold-based trigger mechanism for burst transmission can be adopted, allowing for better throughput for elastic services.

The traffic engineering implications of the reference model in Figure 11.12 can be summarized as follows [23]: First, we note that there is an increase of the traffic variability (marginal distribution variance coefficient) in short timescales, which is due to the grouping of packets in optical bursts. Furthermore, at short timescales, the

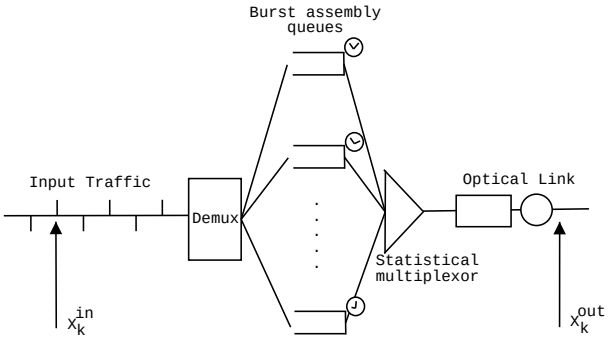

Figure 11.12 OBS reference model

process self-similarity is decreased, due to burst sequencing and shuffling at the output of the burst assembly queues. Nevertheless, at long timescale self-similarity remains the same. The beneficial effect of a self-similarity decrease at short timescales is compensated by the burstiness (traffic variability) increase at such timescales. Figure 11.13 shows an instance of the input and output traffic processes X^{in} and X^{out} in Figure 11.12, plotted in several timescales, X^{in} being a fractional gaussian noise, which proves accurate to model Internet traffic [27]. While the significant traffic bursts can be observed at short timescales we note that the process remains the same as the timescale increases, thus preserving the self-similarity features.

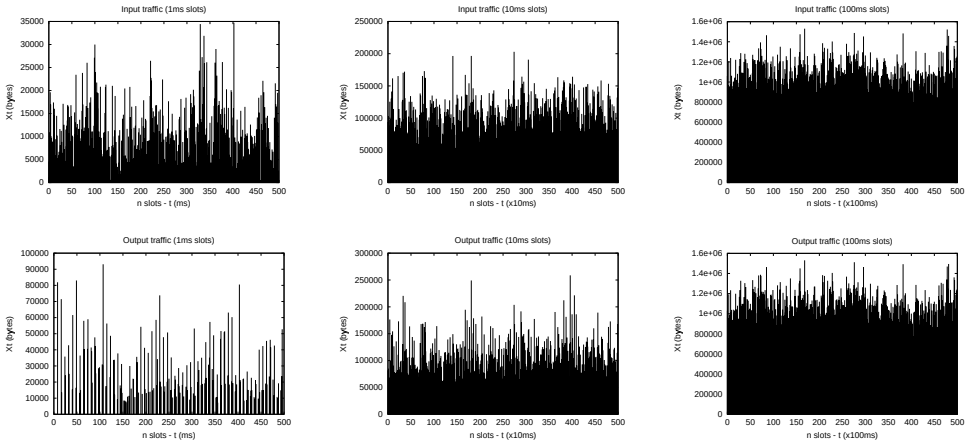

Figure 11.13 Input and output traffic to OBS shaper in several timescales (1, 10, 100 ms.)

11.3.3 Signalling

Two different paradigms are being considered for integration of IP and WDM in the forthcoming next generation Internet. The *overlay* model considers both IP and WDM networks as separate networks with different control planes (like IP over ATM). The *peer* model considers that the IP and WDM network share the same control plane, so that IP routers have a complete view of the optical network logical topology. Figure 11.14 shows the overlay model in comparison to the peer model. The routers displayed in the Figure act as *clients* from the optical network standpoint in the overlay model, since they use the optical network services in order to fulfill edge-to-edge connectivity requirements. Therefore, an Optical User-Network Interface (UNI) becomes necessary at the network edges, as shown in the same Figure. The Optical Internetworking Forum has produced a specification document containing a proposal for implementation of an Optical-UNI which interworks with existing protocols such as SONET/SDH [28].

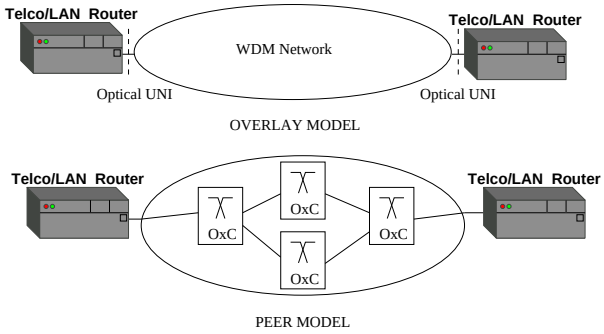

Figure 11.14 Overlay versus peer model

In order to integrate IP and WDM layers in the peer model a promising alternative is the use of Multiprotocol Label Switching (MPLS), *with enhanced capabilities for optical networks*. The aim of MPLS is to provide a higher degree of flexibility to the network manager by allowing the use of Label Switched Paths (LSPs). Labels in MPLS are like VPI/VCI identifiers in ATM in the sense that they have local significance (links between routers) and are swapped at each hop along the LSP. They are implemented as fixed length headers which are attached to the IP packet.

An LSP, thus, is functionally equivalent to a virtual circuit. On the other hand, a Forward Equivalence Class (FEC) is defined as the set of packets which are forwarded in the same way with MPLS. The MPLS label is thus a short, fixed-length value carried in the packet header to identify an FEC. Two separate functional units can be distinguished in an MPLS-capable router: control and forwarding unit. The control unit uses standard routing protocols (OSPF/IS-IS) to build and maintain a forwarding table. When a new packet arrives the forwarding unit makes a routing decision according to the forwarding table contents. However, the network operator

may change the forwarding table in order to explicitly setup an LSP from origin to destination. This explicit routing decision capability provides extensive traffic engineering features and offers scope for quality of service differentiation and virtual private networking. Figure 11.15 shows an example of LSP (routers 1-2-5-6). Even though the link-state protocol mandates, for example, that the best route is 1-4-5-6, the network operator may decide, for load balancing purposes, to divert part of the traffic through routers 2-5-6. On the other hand, MPLS can work alongside with standard IP routing (longest destination IP address prefix match). In our example, the packet may continue its way to the destination host, downstream from router 6, through a non-MPLS capable subnetwork.

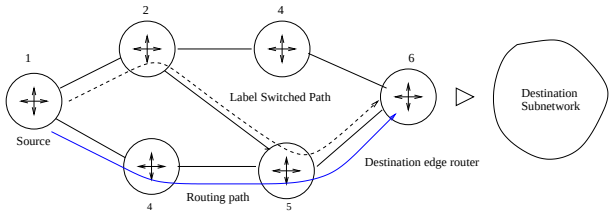

Figure 11.15 Label Switched Path example

Explicit routes are established by means of two different signalling protocols: the Resource Reservation Protocol (RSVP-TE) and Constraint-Based Routing Label-Distributed Protocol. Both allow for strict or loose explicit routing with quality of service guarantees, since resource reservation along the route can be performed. However, the protocols are different in a number of ways: for instance RSVP-TE uses TCP while CR-LDP uses IP/UDP. For an extensive discussion about MPLS signalling protocols the reader is referred to [15], and chapters 13 and 15.

MPLS has evolved into a new standard that is tailored to the specific requirements of optical networks: Generalized MPLS (GMPLS) [6]. Since optical networks may switch entire fibers, wavelengths between fibers or SONET/SDH containers a FEC may be mapped in many different ways, not necessarily packet switched. For instance, a lightpath may be setup in order to carry a FEC. In this example there is no need to attach a label to the packet since the packet is routed in an all-optical fashion end-to-end. Therefore, the (non-generalized) label concept is extended to the *generalized* label concept. The label value is implicit since the transport media identifies the LSP. Thus, the wavelength value becomes the label in a wavelength routed LSP. Labels may be provided to specify wavelengths, wavebands (sets of wavelengths), timeslots or SONET/SDH channels. A SONET/SDH label, for example, is a sequence of five numbers known as S,U,K,L, and M. A packet coming from an MPLS (packet switched) network may be transported in the next hop by a SONET/SDH channel by simply removing the incoming label in the GMPLS router and relaying the packet to the SONET/SDH channel. A generalized label is functionally equivalent to the timeslot location in a plesiochronous network where the slot location identifies a channel, with no need of additional explicit signalling (a packet header in MPLS).

GMPLS allows the data plane (in the optical domain) to perform packet forwarding with the sole information of the packet label, and thus ignoring the IP packet headers. By doing so, there is no need of packet conversion from the optical to the electronic domain and the electronic bottleneck is circumvented. The signalling for LSP setup is performed by means of extensions of both MPLS resource reservation protocols (RSVP-TE/CR-LDP) in order to allow that a LSP can be explicitly routed through the optical core [7, 8]. On the other hand, modifications have been made to GMPLS by adding a new link management protocol designed to address the specific features of the optical media and photonic switches, together with enhancements to the Open Shortest Path First protocol (OSPF/IS-IS) to provide a generalized representation of the various link types in a network (fibers, protection fibers, etc.), which was not necessary in the packet-switched Internet. Furthermore, signalling messages are carried out of band in an overlay signalling network, in order to optimize resources and in contrast to MPLS. Since signalling messages are short in size and the traffic volume is not large in comparison to the data counterpart the allocation of optical resources for signalling purposes seems wasteful. See chapters 13 and 15 for more details on GMPLS.

As an example of the enhancements provided by GMPLS, consider the number of parallel channels in a DWDM network, which is expected to be much larger than in the electronic counterpart. Consequently, assigning one IP address to each of the links seems wasteful due to the scarcity of IP address space. The solution to this problem is the use of unnumbered links as detailed in [19]. The former example illustrates the need of adapting existing signalling protocols in the Internet (MPLS) to the new peer IP over WDM paradigm.

As for third generation optical networks based on photonic packet switching there are recent proposals for Optical Code Multiprotocol Label Switching (OC-MPLS) based on optical code correlations [25]. The issue is how to read a packet header in the all-optical domain. Each bit of the packet header is mapped onto a different wavelength in a different time position, forming a sequence which can be decoded by correlation with all the entries in a code table. While the use of photonic label recognition has been demonstrated in practice the number of codes available is very scarce (8 bits/128 codes) [38], thus limiting the applicability of the technique to long haul networks.

11.3.4 Framing aspects for IP over WDM

Concerning framing techniques for IP over WDM, the IP over SONET/SDH standard is foreseen as the most popular solution in the near future [10]. The rationale of IP over SONET/SDH is to simplify the protocol stack of IP over ATM, which in turn relies over SONET/SDH as the physical layer. By encapsulating the IP datagrams on top of SONET/SDH the bandwidth efficiency is increased while the processing burden imposed by segmentation and reassembly procedures disappears. However, a link layer protocol is still needed for packet delineation. Figure 11.16 shows the IP datagram in a layer 2 PDU (HDLC-framed PPP) which, in turn, is carried in a SONET/SDH STS-1 payload.

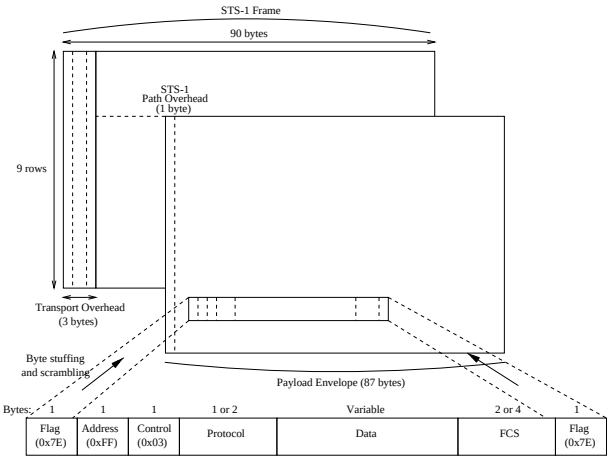

Figure 11.16 IP over SONET/SDH

The current standards for layer 2 framing propose the use of HDLC-framed PPP as described in RFC 1662/2615 [34, 35]. However, due to scrambling and reliability problems the scalability is compromised beyond OC-48 [21]. In response to the need of higher speeds other proposals for IP over SONET/SDH have appeared such as the PPP over SDL (Simplified Data Link) standard (RFC 2823) [11]. In PPP over SDL the link synchronization is achieved with an algorithm which is similar to I.432 ATM HEC delineation. Instead of searching for a flag (Ox7E), as is done in POS, the receiver calculates the CRC over a variable number of bytes until it "locks" to a frame. Then the receiver enters the SYNC state and packet delineation is thus achieved. In addition, data scrambling is performed with a self-synchronous $x^{43} + 1$ scrambler or an optional set-reset scrambler independent of user data, which makes it impossible for the malicious user to break SONET/SDH security.

We note that the performance of IP over SONET/SDH, both in POS and PPP over SDL is highly dependent on the IP packet size. Assuming no byte stuffing (escaping Ox7E flags) and 16 bits frame check sequence, the POS overhead is 7 bytes. For a SONET/SDH layer offered rate of 2404 Mbps (OC-48) the user-perceived rate on top of TCP/UDP is 2035 Mbps for an IP packet size of 300 bytes.

Finally, there are also ongoing efforts to provide lightweight framing of IP packets over DWDM using 10 Gigabit Ethernet (IEEE 802.3ae task force) [36] and Digital Wrappers (ITU G.709) [1].

11.4 END-TO-END ISSUES

The success of the next generation optical Internet will not only depend on optical technology, but also on the set of protocols that translate the availability of gigabit

bandwidth into user-perceived quality of service. Figure 11.17 shows a reference model for WDM network architecture. The WDM wide/metropolitan area network serves as an Internet backbone for the different access networks. 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.

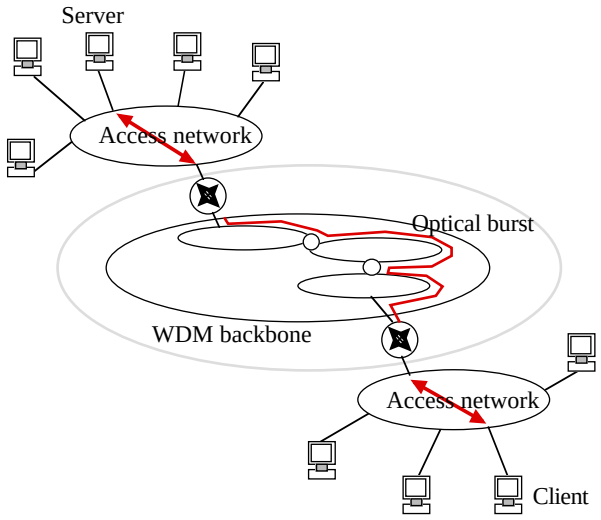

Figure 11.17 WDM network reference model

The scenario shown in Figure 11.17 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}).

Bridging the gap between access and backbone network 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 [2, 12]. 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 [4, 22]. As a result, the connection is always slow-starting, unless the initial window size is very large [18]. On the other hand, 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 to 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.

11.4.1 TCP for high-speed and split TCP connections

As a first approach to solving the adaptation problem between access and backbone the TCP connection can be *split* in the optical backbone edges. As a result, each (separate) TCP connection can be provided with TCP extensions tailored to the specific requirements of both access and backbone network. More specifically, the backbone TCP connection uses TCP extensions for speed, as described in [18]. Such TCP extensions consist of larger transmission window and no slow start.

A simulation model for the reference architecture of Figure 11.17 is shown in Figure 11.18. The ns[3] 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 11.18. Regarding the

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

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 based on TCP with the performance metric being connection throughput.

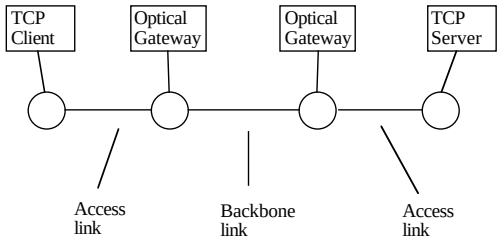

Figure 11.18 ns model

Table 11.2 Summary of simulation parameters

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	$25ms, 50ms, 100ms$

Figure 11.19 shows throughput versus file size for split-TCP (STCP-PPS) and end-to-end TCP (EE-PPS) connection with the access network parameters shown in table 11.2. On the other hand, results from a proposal of a transfer protocol using OBS [17], Files over Lightpaths (FOL), are also presented. In FOL, files are encapsulated in an optical burst in order to be transmitted across the optical network. We note that throughput grows with file size towards a value which is independent of access BW, and 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 a 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.

We note that the use of split TCP provides a significant performance improvement. Thus, we expect that the forthcoming WDM networks will incorporate a connection adaptation mechanism in the edge routers so that the WDM bandwidth can be fully exploited.

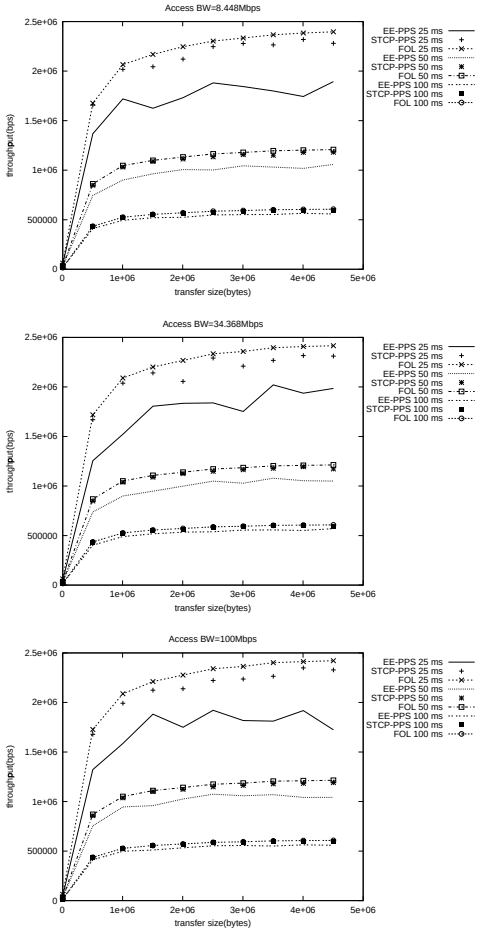

Figure 11.19 Average transfer throughput

11.4.2 Performance evaluation of file transfer (WWW) services over WDM networks

The results shown in the previous section provide a comparison in error free conditions, for instance in a first generation WDM network (static lightpath between routers). However, it turns out that second generation WDM networks suffer blocking probability as a consequence of the limited number of wavelengths and burst dropping due to limited queueing space in optical burst or photonic packet switches. In such conditions split TCP becomes inefficient and alternate protocol design become necessary.

In FOL [17], files are encapsulated in optical bursts which are released through the optical backbone using a simple stop and wait protocol for error control. Assuming that the setup of an optical burst takes RTT/2, Figure 11.20 shows the achieved throughput with error probabilities lower than 0.1, for both FOL (different burst sizes) and Split TCP (STCP). We observe that TCP congestion avoidance severely limits transfer efficiency. *If loss probabilty 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.

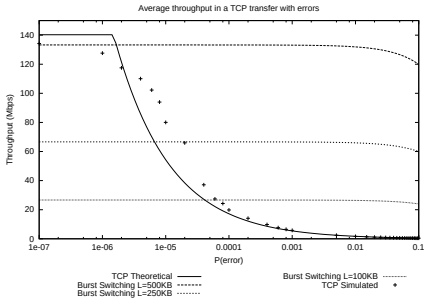

Figure 11.20 Throughput comparison

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 is the blocking probability the larger is the 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.

11.5 CONCLUSIONS

In this chapter, we have presented the motivation, solutions and open challenges for the emerging field of traffic management in all-optical networks. At the time of the writing of the chapter there are a large number of open issues, including traffic management for IP traffic in dynamic WDM networks (dynamic lightpath and burst switching), design of an efficient integrated control plane for IP and WDM and proposal of transport protocol that translate the availability of optical bandwidth into user-perceived quality of service. We believe that the research effort in the next generation optical Internet will be focused not only in the provision of a very large bandwidth but also in the flexibility, ease of use and efficiency of optical networks.

Appendix: Measurement scenario

Our traffic traces are obtained from the network configuration depicted in Figure 11.A.1. The measurements presented in this paper are performed at the ATM Permanent Virtual Circuit (PVC) that links Public University of Navarra to the core router of the Spanish academic network (*RedIris* [4]) in Madrid. Rediris topology is a star of PVCs which connect the Universities around the country to the central interconnection point in Madrid. From the central RedIris facilities in Madrid a number of international links connect the Spanish ATM academic network to the outside Internet. The measured PVC uses bridged encapsulation and it is terminated at both sides by IP routers. The Peak Cell Rate (PCR) of the circuit is limited to 4 Mbps and the transmission rate in the optical fiber is 155 Mbps.

Table 11.A.1 Trace characteristics

Start date	Mon 14/02/2000 00:00
End date	Mon 14/02/2000 24:00
TCP connections recorded	957053
IP packets analyzed	16375793

We note that the scenario under analysis is a representative example of a number of very common network configurations. For example, the most Spanish Internet Service Providers (ISPs) hire ATM PVC links to the operators in order to provide customers with access to the Internet. The same situation arises with corporate and academic networks, that are linked to the Internet through such IP over ATM links. On the other hand, measurements are not constrained by a predetermined set of destinations but represent a real example of a very large sample of users accessing random destinations in the Internet. Table 11.A.1 summarizes the main characteristics of the traffic trace presented in this chapter.

[4] http://www.rediris.es

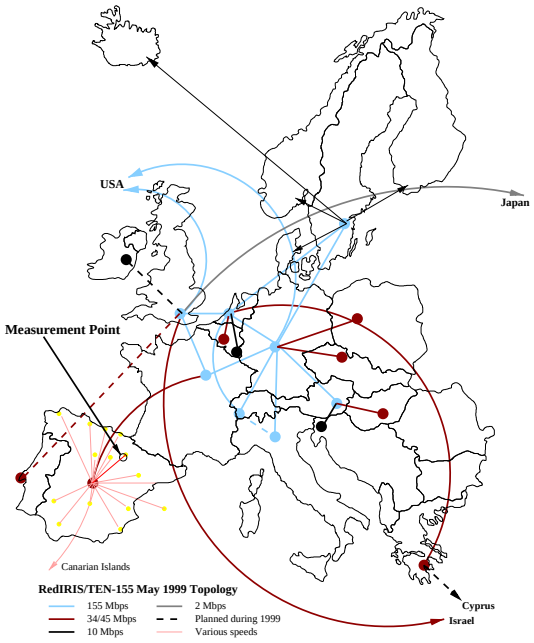

Figure 11.A.1 Network measurement scenario

REFERENCES

1. ANSIT1X1.5/2001-064 Draft ITU-T Rec. G. 709. Network node interface for the optical transport network (OTN), March 2001.

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

3. J. Aracil, M. Izal, and D. Morato. Internet traffic shaping for IP over WDM links. *Optical Networks Magazine*, 2(1), January/February 2001.

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

5. B. Arnaud. Architectural and engineering issues for building an optical internet. In *Proceedings of SPIE International Symposium on Voice, Video, and Data Communications – All-Optical Networking: Architecture, Control, and Management Issues*, Boston, MA, November 1998.

6. P. Ashwood-Smith et al. Generalized MPLS-SIgnaling Functional Description. Internet-Draft, work in progress, Nov 2000.

7. A. Banerjee, J. Drake, J. Lang, B. Turner, D. Awduche, , L. Berger, K. Kompella, and Y. Rekhter. Generalized multiprotocol label switching: An overview of signalling enhancements and recovery techniques. *IEEE Communications Magazine*, July 2001.

8. A. Banerjee, J. Drake, J. Lang, B. Turner, K. Kompella, and Y. Rekhter. Generalized multiprotocol label switching: An overview of routing and management enhancements. *IEEE Communications Magazine*, January 2001.

9. J. Bellamy. *Digital telephony*. John Wiley & Sons, New York, 1991.

10. P. Bonenfant and A. Rodriguez-Moral. Framing techniques for ip over fiber. *IEEE Network*, July/August 2001.

11. J. Carlson, P. Langner, E. Hernandez-Valencia, and J. Manchester. PPP over simple data link (SDL) using SONET/SDH with ATM-like framing. RFC 2823 (Experimental), May 2000.

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

13. M. E. Crovella and A. Bestavros. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, December 1997.

14. R. Edell, N. McKeown, and P. Varaiya. Billing users and pricing for TCP. *IEEE Journal On Selected Areas In Communication*, 13(7), September 1995.

15. E. W. Gray. *MPLS Implementing the Technology*. Addison-Wesley, 2001.

16. A. Ganz, I. Chlamtac and G. Karmi. Lightnets: Topologies for high speed optical networks. *IEEE/OSA Journal of Lightwave Technology*, 11, May/June 1993.

17. M. Izal and J. Aracil. IP over WDM dynamic link layer: challenges, open issues and comparison of files-over-lighpaths versus photonic packet switching. In *Proceedings of SPIE OptiComm 2001*, Denver, CO, August 2001.

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

19. K. Kompella and Y. Rekhter. Signalling unnumbered links in RSVP-TE. Internet-Draft, work in progress, Sept 2000.

20. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking*, 2(1):1–15, January 1994.

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

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

23. D. Morató, J. Aracil, M. Izal, E. Maga na, and L. Diez-Marca. Explaining the impact of optical burst switching in traffic self-similarity. Technical Report Public University of Navarra.

24. B. Mukherjee. *Optical Communication Netwoks*. McGrawHill, 1997.

25. M. Murata and K. Kitayama. A perspective on photonic multiprotocol label switching. *IEEE Network*, July/August 2001.

26. O. Narayan, A. Erramilli and W. Willinger. Experimental Queueing Analysis with Long-Range Dependent Packet Traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, April 1996.

27. I. Norros. On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks. *IEEE Journal on Selected Areas in Communications*, 13(6):953–962, August 1995.

28. Optical Internetworking Forum (http://www.oiforum.com). OIF UNI 1.0- controlling optical networks, 2001.

29. V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 4(2):226–244, April 1996.

30. C. Qiao. Labeled optical burst switching for ip-over-wdm integration. *IEEE Communications Magazine*, 38:104–114, September 2000.

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

32. C. Qiao and M. Yoo. features and issues in optical burst switching. *Optical Networks Magazine*, 1:36–44, April 2000.

33. R. Sherman, W. Willinger, M. S. Taqqu and Daniel V. Wilson. Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Transactions on Networking*, 5(1), February 1997.

34. W. Simpson, A. Malis. PPP in HDLC-like framing. RFC 1662, July 1994.

35. W. Simpson, A. Malis. PPP over SONET/SDH. RFC 2615, June 1999.

36. D. H. Su. Standards: The IEEE P802.3ae project for 10 Gb/s Ethernet. *Optical Networks Magazine*, 1(4), October 2000.

37. B. Tsybakov and N. D. Georganas. On self-similar traffic in ATM queues: Definitions, overflow probability bound and cell delay distribution. *IEEE/ACM Transactions on Networking*, 5(3):397–409, June 1997.

38. N. Wada and K. Kitayama. Photonic IP routing using optical codes: 10 gbit/s optical packet transfer experiment. In *Proceedings of Optical Fiber Communications Conference 2000*, Baltimore, MD, March 2000.

39. S. Yao, B. Mukherjee and S. Dixit. Advances in photonic packet switching: An overview. *IEEE Communications Magazine*, 38(2):84–94, February 2000.

40. 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.